

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Российский экономический университет имени Г. В. Плеханова»

На правах рукописи

Сорокин Александр Сергеевич

**Модели оценки и управления кредитными рисками
микрофинансовых организаций**

5.2.2. Математические, статистические и инструментальные методы в экономике

Диссертация на соискание ученой степени
доктора экономических наук

Научный консультант –
доктор экономических наук, профессор
Тихомиров Николай Петрович

Москва – 2025

Оглавление

Введение.....	5
Глава 1 Теоретико-методологические проблемы оценки и управления кредитными рисками в микрофинансовых организациях.....	20
1.1 Тенденции и закономерности развития рынка микрофинансирования в России.....	20
1.2 Особенности деятельности российских микрофинансовых организаций в условиях нестабильности рынка.....	49
1.3 Методологические подходы к оценке и управлению кредитными рисками микрофинансовых организаций.....	67
Выводы по главе 1.....	101
Глава 2 Особенности информационного обеспечения процедур оценки кредитных рисков в микрофинансовых организациях.....	102
2.1 Виды исходной информации и особенности ее выбора для оценки кредитных рисков микрофинансовых организаций.....	102
2.2 Особенности применения показателей ценности информации для оценки кредитных рисков микрофинансовых организаций.....	120
2.3 Подходы к оценке кредитного риска с учетом качества бинарной классификации заемщиков микрофинансовых организаций.....	139
Выводы по главе 2.....	158
Глава 3 Методологические подходы к оценке кредитных рисков микрофинансовых организаций.....	160
3.1 Методика разработки скоринговых моделей оценки кредитных рисков микрофинансовых организаций.....	160
3.2 Статистические модели кредитных рисков и особенности их применения в микрофинансовых организациях.....	184
3.3 Модели и алгоритмы машинного обучения оценки кредитных рисков и особенности их применения в микрофинансовых организациях.....	207
Выводы по главе 3.....	229

Глава 4 Методы управления кредитными рисками в микрофинансовых организациях.....	230
4.1 Управление кредитными рисками микрофинансовых организаций на основе регулирования лимитов.....	230
4.2 Управление кредитными рисками микрофинансовых организаций на основе модели юнит-экономики.....	251
4.3 Управление кредитными рисками микрофинансовых организаций в условиях изменчивости коэффициентов скоринговых моделей во времени.....	270
Выводы по главе 4.....	295
Глава 5 Апробация моделей оценки и управления кредитными рисками в микрофинансовых организациях.....	297
5.1 Оценка кредитных рисков на основе агрегированного признака просрочки по данным кредитной истории заемщика.....	297
5.2 Оценка кредитных рисков на основе статистических моделей и алгоритмов машинного обучения.....	319
5.3 Управление кредитными рисками на основе моделей лимитов, юнит-экономики и скоринговой модели с изменяющимися во времени параметрами.....	343
Выводы по главе 5.....	361
Заключение.....	362
Список сокращений и условных обозначений.....	367
Список литературы.....	371
Приложение А (рекомендуемое) Примеры источников данных для оценки кредитных рисков на российском рынке и их основные характеристики.....	412
Приложение Б (рекомендуемое) Пример работы алгоритма бининга на основе анализа показателей веса категорий.....	413
Приложение В (рекомендуемое) Пример работы алгоритма простой MDLP-дискретизации.....	415
Приложение Г (обязательное) Расчеты юнит-экономики первичного займа МФО.....	419

Приложение Д (обязательное) Расчеты юнит-экономики повторного займа МФО.....	427
Приложение Е (справочное) Листинг кода программы расчета признака качества кредитной истории в среде Python.....	435
Приложение Ж (справочное) Свидетельство о государственной регистрации программы для ЭВМ «Качество кредитной истории: программа для вычисления рейтинга заемщика по его кредитной истории для оценки кредитных рисков и построения скоринговых моделей».....	443
Приложение И (справочное) Справки о внедрении и апробации результатов диссертационного исследования.....	444
Приложение К (справочное) Листинг кода программы «Greeder of Data – автоматический советник для построения скоринговых моделей на базе логистической регрессии на основе больших массивов данных в среде Python».....	452
Приложение Л (справочное) Свидетельство о государственной регистрации программы для ЭВМ «Greeder of Data – автоматический советник для построения скоринговых моделей на базе логистической регрессии на основе больших массивов данных».....	468
Приложение М (справочное) Диплом Лауреата 31-го Международного конкурса научно-исследовательских работ за работу: «Моделирование оптимальных кредитных лимитов в микрофинансовых организациях».....	469
Приложение Н (справочное) Листинг кода программы для моделирования лимитов.....	470
Приложение П (справочное) Листинг кода программы в R расчетов имитационного эксперимента.....	473

Введение

Актуальность темы исследования. Успешная коммерческая деятельность в микрофинансовых организациях (МФО), как и в банковской сфере, базируется на установлении разумного компромисса между рисками и доходностью. Это связано с тем, что кредитование физических лиц в сегменте микрофинансирования является не только доходным видом деятельности, но в первую очередь сопряжено с большими рисками, обусловленными возможной некредитоспособностью заемщика.

В общей сложности за период с 2018 года по первое полугодие 2024 года рост рынка микрофинансирования в России составил 253,6 млрд руб. – почти в 2,1 раза при среднегодовом темпе прироста 21,2%. При этом в рассматриваемый период наблюдались резкие колебания темпов прироста профильных активов МФО. В частности, снижение темпов прироста в 2020 и 2022 гг. было обусловлено снижением потребительской активности после пандемии COVID-19 и из-за санкционного давления.

Увеличение количества вновь заключенных договоров кредитования между МФО и заемщиками – физическими лицами в совокупности за 2018–2023 гг. составило 45 млн договоров (увеличение в 1,56 раза при среднегодовом темпе прироста 22,1% или на 9 млн договоров ежегодно). Такая динамика сопровождалась увеличением совокупного объема выдачи микрозаймов с 298,5 млрд руб. в 2018 г. до 902,5 млрд руб. в 2023 г. Данная тенденция определяется, в первую очередь, активизацией спроса со стороны существующих и потенциальных заемщиков, в т. ч. в условиях перехода от банковского кредитования к микрофинансовому в связи со снижением уровня одобрения заявок в банковских учреждениях.

В то же время на рынке фиксируется рост объема просроченной кредитной задолженности в общем кредитном портфеле населения. Долговая нагрузка населения РФ в 2024 году достигла 57,1%, что свидетельствует о наличии

реальных финансовых трудностей у значительной части населения, низкой финансовой грамотности, зачастую агрессивной рекламной кампании банков и МФО. В связи с этим перед финансовыми институтами встает вопрос минимизации кредитных рисков на основе более адекватных сложившейся ситуации подходов к их оценке и управлению.

В настоящее время доходность микрофинансового бизнеса существенно снизилась из-за проводимой в последние три года политики регулирования этого рынка Центральным банком РФ. Тенденция такова, что Банк России все сильнее регулирует рынок и предъявляет более «жесткие» требования по управлению кредитными рисками в микрофинансовой сфере, близкие к банковским стандартам.

При этом на рынке микрофинансирования наблюдается рост конкуренции, уменьшение количества участников. В совокупности за последние 10 лет количество МФО снизилось на 3 243 организации (в относительном выражении – 78,1%, среднегодовой темп снижения составил 11,86%), составив по состоянию на начало декабря 2024 года 909 МФО.

Укрупнение рынка и уход мелких игроков обусловлен, в частности, неумением последних качественно управлять кредитными рисками. Ведущие участники рынка микрофинансирования в России имеют систему риск-менеджмента, выстроенную по банковским стандартам.

Еще одна закономерность развития рынка микрофинансирования в России – все больший переход в сферу онлайн кредитования. Такая бизнес-модель МФО требует использования в режиме реального времени более качественной модели оценки кредитного риска и идентификации надежности заемщика на основе доступных источников данных по сравнению с оффлайн кредитованием. Существенное развитие онлайн-кредитования на рынке микрофинансирования обусловлено двумя основными причинами. С одной стороны, более высокая скорость принятия решения о выдаче займа в МФО является основным критерием выбора для заемщика, с другой стороны развитие онлайн-сегмента позволяет МФО функционировать без географической привязки, а также снижать

операционные расходы. По состоянию на первое полугодие 2024 г. количество заключенных посредством онлайн-каналов договоров микрозаймов выросло до 92,77% против 43,43% в 2018 г.

Повышение уровня кредитных рисков МФО, обусловленное ростом конкуренции на рынке и проводимой регуляторной политикой Центрального Банка РФ, требует разработки и совершенствования подходов и методов к оценке и управлению кредитными рисками.

В большинстве МФО модели оценки и управления кредитными рисками являются собственными разработками с различающимися методиками, алгоритмами и методами на основе доступных на рынке данных о заемщике. Различные методики оценки кредитоспособности заемщиков МФО отличаются друг от друга составом факторов, используемых при оценке общего кредитного рейтинга заемщика, а также подходами к оценкам каждого параметра в модели. Состав источников данных для формирования независимых переменных в скоринговых моделях не универсален для всех финансовых институтов, стран и рынков, что в свою очередь требует разработки системы принятия решений на основе моделей кредитного риска, учитывающих специфику бизнес-процессов конкретной МФО. В этих условиях вопросы повышения обоснованности и достоверности оценок и управления кредитными рисками для МФО особенно актуальны.

Степень разработанности научной проблемы. Разработка подходов к моделированию и управлению кредитными рисками является популярной темой исследований ученых.

Для управления кредитными рисками финансовой организации применяются концепции, базирующиеся на моделях оценки кредитных рисков, которые можно получить на основе методов эконометрики. Значительный вклад в развитие общей теории эконометрики, на которую опирается автор в диссертационном исследовании, внесли: С.А.Айвазян, Т.А.Дуброва, П.К.Катышев, Ю.П.Лукашин, Я.Р.Магнус, В.С.Мхитарян, А.А.Пересецкий, Н.А.Садовникова, Н.П.Тихомиров, Т.М.Тихомирова, А.А.Френкель,

Е.Н.Четыркин, T.W.Anderson, G.E.Box, D.R.Cox, Ch.Dougherty, G.M.Jenkins, F.Fisher, D.G.Kendall, R.Kohn, K.D.Lewis, N.J.D.Nagelkerke, K.Pearson, A.Rencher, E.J.Snell.

Работы, посвященные особенностям моделирования кредитных рисков в банковской сфере, были выполнены следующими авторами: А.С.Бурова, Е.С.Волкова, Д.Ю.Голембиовский, А.А.Гришин, Л.В.Жукова, Д.В.Исаев, В.Е.Кислицин, М.Ковалев, В.Корженевская, Л.А.Кузнецов, А.Г.Мадера, И.Н.Мастяева, Г.И.Пеникас, А.В.Перевозчиков, А.А.Пересецкий, К.Л.Поляков, С.В.Попова, Е.Г.Снегова, М.А.Широбокова, А.Е.Яблонская, H.A.Abdou, A.Al-Aradi, K.Amzile, E.A.Angelini, R.Anderson, A.Ampountolas, T.Astebro, J.Banasik, I.Barkova, R.Bay, G.Bennouna, A.Blanco, E.Costa e Silva, R.Chamboko, J.S.Cheney, S.Chikalipah, A.Costangioară, S.Dey, D.Durand, E.B.Gonçalves, A.Husain, S.Guzmán-Castillo, E.M.Lewis, Y.Mao, B.S.Maranga, K.Miled, T.Mokheleli, J.Pointon, N.Putri, N.Siddiqi, M.Sustersic, U.Rahmani, F.Sanchez, L.C.Thomas, Y.Wang, X.Wu, R.Zhang.

Значительный вклад в развитие области общей теории риска, на которую опирается автор в диссертационном исследовании, внесли Н.П.Тихомиров, J.Cohen, C.H.Coombs, F.Knight, D.G.Pruitt, M.Rothschild, J.Stiglitz.

Особенности становления и развития рынка микрофинансирования в России и в мире представлены в работах: М.В.Ароян, Ю.Н.Беляевой, М.М.Богдановой, Е.С.Воронцовой, А.Ф.Ибрагимовой, П.А.Косенковой, В.Ю.Кузнецовой, С.В.Масалкиной, Н.В.Рубцовой, С.Б.Смелова, Н.Е.Соловьевой, Е.С.Стряпчих, Е.А.Устюговой, А.В.Фалеева, И.Е.Шакер, Н.С.Шакер, В.Ю.Шаповалова, M.Yunus.

Особенности использования статистических моделей, методов и алгоритмов машинного обучения для оценки кредитных рисков подробно описаны в трудах: А.А.Адуенко, А.В.Ветрова, М.Власенко, Е.С.Волковой, В.Б.Гисина, Е.А.Губаревой, О.А.Дьякова, Е.В.Орловой, В.И.Соловьева, С.И.Хашина, Е.С.Шемяковой, A.Albert, J.A.Anderson, M.Aniceto, C.M.Bishop, B.Boser, L.Breiman, P.J.Brockwell, W.Chen, D.Cox, T.G.Dietterich, J.H.Friedman, T.L.Fine, N.Günnemann, S.Haykin, D.Hosmer, M.Hurley, J.Jaccard, S.Lemeshow, B.Liu B,

L.Mason, E.Mays, C.Meng, T.Mokheleli, A.Montrenko, R.Rojas, N.Putri, E.Snell, Y.Tounsi, V.N.Vapnik, D.West, L.Zhou.

Определенное направление в управлении экономикой компаний, связанное с принятием решений на основе модели юнит-экономики, раскрыто в работах: А.С.Аленикова, Ю.Г.Боровко, М.В.Вахориной, М.А.Вахрушиной, В.А.Гавриловой, Я.А.Гриневич, Н.С.Ермашкевич, В.Г.Иванова, С.Г.Креновой, Т.А.Лежниной, И.Г.Рзун, А.И.Созыкиной, Е.Г.Сухих, М.Г.Треймана, О.С.Щербаковой.

Все эти разработки опираются на предпосылки экономической теории, отраженные в работах следующих авторов: П.И.Гребенникова, Л.С.Тарасевич, А.И.Леусского, К.Макконнелла, С.Брю, А.Abel, В.Bernanke, D.Croushore.

Сформулированные в работах вышеупомянутых авторов инструменты, математические модели, методы и алгоритмы машинного обучения позволили изучить существующие подходы к оценке и управлению кредитными рисками в банковской сфере. Однако, в отличие от банков в МФО существует своя специфика оценки и управления кредитными рисками, которая не отражена в полной мере в представленных выше работах. В этой связи необходимо совершенствование и уточнение моделей оценки и управления кредитными рисками, применяемых в банковской сфере, для микрофинансирования.

Предметом исследования выступают модели, методы и алгоритмы для оценки и управления кредитными рисками в микрофинансовых организациях.

Объектом исследования являются кредитные риски в микрофинансовых организациях.

Область исследования. Основные положения и выводы диссертационной работы соответствуют паспорту научных специальностей Высшей аттестационной комиссии при Минобрнауки России по специальности 5.2.2 «Математические, статистические и инструментальные методы в экономики» в пунктах: п. 1. «Теоретические и методологические вопросы применения математических, статистических, эконометрических и инструментальных методов в экономических исследованиях»; п. 2. «Типы и виды экономико-математических

и эконометрических моделей, методология их использования для анализа экономических процессов, объектов и систем»; п. 3. «Разработка и развитие математических и эконометрических моделей анализа экономических процессов (в т.ч. в исторической перспективе) и их прогнозирования»; п. 14. «Эконометрические и статистические методы анализа данных, формирования и тестирования гипотез в экономических исследованиях. Эконометрическое и экономико-статистическое моделирование».

Целью диссертационного исследования является развитие методологии оценки и управления кредитными рисками микрофинансовых организаций на основе совершенствования экономико-математических моделей, учитывающих особенности микрофинансового кредитования.

Для достижения указанной цели в работе поставлены и решены следующие **задачи**:

- 1) выявить особенности рынка микрофинансирования в России и основные тенденции и закономерности его развития;
- 2) обобщить и систематизировать методологические подходы к оценке и управлению кредитными рисками МФО;
- 3) систематизировать возможные варианты исходной информации для оценки кредитных рисков МФО и обосновать подходы к ее рациональному выбору;
- 4) предложить подходы к оценке кредитных рисков МФО с учетом особенностей показателей ценности исходной информации и финансового состояния заемщика;
- 5) усовершенствовать методы разработки скоринговых моделей оценки кредитных рисков заемщиков МФО;
- 6) разработать систему управления кредитными рисками МФО с использованием моделей юнит-экономики и лимитирования займов;
- 7) разработать подход к управлению кредитными рисками МФО в условиях изменчивости коэффициентов скоринговых моделей во времени;

8) разработать алгоритм расчета агрегированного показателя оценки кредитного риска заемщика МФО на основе данных его кредитной истории;

9) разработать программы реализации моделей, методов и алгоритмов оценки и управления кредитными рисками для ЭВМ;

10) провести апробацию разработанных моделей, алгоритмов и методик оценки и управления кредитными рисками в российских МФО и обосновать рекомендации по их применению в условиях российского рынка микрофинансирования.

Теоретической базой исследования являются труды российских и зарубежных ученых в области оценки и моделирования кредитных рисков, общей теории риска, анализа и прогнозирования временных рядов экономических процессов, теорий искусственных нейронных сетей, методов машинного обучения и интеллектуального анализа данных, экономического анализа и управления, принятия решений, статистики и эконометрики.

Методологическую базу исследования составляют статистические методы: линейный корреляционный и регрессионный анализ, логистическая регрессия, ROC-анализ; многомерные статистические методы: дискриминантный анализ, компонентный анализ, иерархический кластерный анализ, построение деревьев решений; методы анализа и моделирования временных рядов: тренд-сезонные модели, ARIMA, DCC-GARCH, пространственно-временные модели; методы и алгоритмы машинного обучения: «случайный лес», «градиентный бустинг», метод опорных векторов, нейронные сети, а также табличные и графические методы представления результатов исследования.

Информационная база исследования сформирована из официальных статистических данных Центрального Банка РФ, Федеральной службы государственной статистики РФ, аналитических отчетов ООО «Бюро кредитных историй Скоринг бюро», рейтингового агентства «ЭКСПЕРТ РА» и других печатных и электронных средств массовой информации по исследуемой тематике. В ходе подготовки диссертационного исследования изучены нормативные правовые акты, регламентирующие различные аспекты микрофинансовой

деятельности в России. Для разработки методик, алгоритмов, моделей и их апробации использовались данные компании Thomson Reuters, имитационные данные и реальные исторические данные действующих на российском рынке микрофинансовых организаций.

В диссертации был использован следующий **программный инструментарий**: пакет для анализа данных MS Excel (Microsoft), SPSS Statistics 27.0 (IBM), STATISTICA 10.0 (Dell), R 4.03, Python 3.9.4, SPSS Modeler 18.0 (IBM), Gretl 2024b.

Научная новизна. Разработаны методологические подходы, базирующиеся на системе моделей, методах и алгоритмах оценки и управления кредитными рисками, учитывающих характерные для рынка микрофинансирования особенности: высокую стоимость привлечения клиентов, значительный уровень кредитных рисков у первичных клиентов, ограниченность состава исходной информации для оценки кредитных рисков, жесткие регуляторные ограничения со стороны ЦБ и саморегулируемых организаций и другие.

На защиту выносятся следующие **результаты, определяющие новизну** диссертационного исследования:

1) Раскрыты особенности оценки и управления кредитными рисками в МФО с учетом уточненного состава факторов, влияющих на их уровень.

В качестве таких факторов выделены: стоимость привлечения клиентов, затраты на принятие решения о выдаче займа, кредитный риск первичных и повторных заемщиков, регуляторные правила и законодательные ограничения.

2) Выделены виды исходной информации для оценки кредитных рисков МФО и раскрыты особенности ее выбора.

В качестве источников информации для оценки кредитных рисков предложено учитывать данные кредитных историй и альтернативных источников (операторов сотовой связи, транзакционные данные, данные социальных сетей и другие). При обосновании их выбора предложено учитывать: доступность исходных данных, способы обработки, характер источника данных, цель

применения, временной период, возможность проведения ретро-тестирования, способ отправки информации, наличие согласия заемщика.

3) Предложена методика оценки ценности исходной информации для обоснования уровня кредитных рисков МФО.

При оценке кредитных рисков в работе предложено использовать показатели ценности исходной информации в рамках оригинальной методики, учитывающей цели ее использования в МФО: уменьшение общего уровня кредитных потерь, улучшение действующей модели оценки кредитных рисков, снижение затрат на принятие решения о выдаче займа, увеличения конверсии из заявки в выдачу.

Сформулированы и предложены критерии для оценки ценности источника данных: полнота данных, наличие влияния на показатели риска, оценка экономического эффекта, затраты на интеграцию и внедрение, надежность компании-провайдера, общая классификационная способность.

4) Систематизированы модели, методы и алгоритмы оценки кредитных рисков (логистическая регрессия, дискриминантный анализ, нейронные сети, метод опорных векторов, «градиентный бустинг», «случайный лес» деревьев решений) и определена целесообразность их применения с учетом особенностей доступной исходной информации (кредитных историй, транзакционных данных, анкетных данных и других).

5) Разработаны математические модели назначения лимитов PDL-займов для управления кредитными рисками МФО.

Модель лимитов для первичных и повторных заемщиков строится на основе логистической регрессии и взвешенного метода наименьших квадратов, учитывая дециль риска и уровень фактической доходности конкретного сегмента заемщиков в портфеле МФО по историческим данным.

Для повторных заемщиков предложен алгоритм определения лимитов с учетом уровня поведенческих факторов: количество дней просрочки по предыдущему закрытому займу в МФО; фактический срок пользования займом

(досрочное погашение, погашение в срок или пролонгация займа); сумма дохода от предыдущего закрытого займа в МФО.

6) Разработан алгоритм расчета агрегированного показателя просрочки по данным кредитной истории заемщика для оценки кредитных рисков МФО.

Показатель представляет собой вариант оценки платежеспособности в баллах от 0 до 10, рассчитанный по всем кредитам клиентам. Данный агрегированный показатель вычисляется на основании следующих параметров: даты актуальности кредита, типа кредита и платежной дисциплины.

7) Предложена модель юнит-экономики для управления кредитными рисками МФО.

Модель ориентирована на максимизацию прибыли с одного заемщика на горизонте 2 лет. Ограничением модели являются издержки по кредитным потерям МФО и другие затраты.

8) Предложен метод корректировки коэффициентов скоринговой модели логистической регрессии, учитывающий особенности их изменения в ретроспективе с использованием инструментария анализа временных рядов.

9) Разработано программное обеспечение, реализующее расчеты по моделям оценки и управления кредитными рисками на языке Python.

10) Предложены рекомендации по управлению кредитными рисками МФО на основе апробации разработанных моделей и информационных систем их реализации на примере российских МФО.

При оценке кредитных рисков целесообразно использовать адекватный постановкам задач состав исходной информации, различающийся по источникам ее сбора и обработки, полноте и достоверности, а также учитывать регуляторные ограничения (базовый стандарт микрофинансовой деятельности, макропруденциальные лимиты и другие). При управлении кредитными рисками целесообразно ориентироваться на критерии эффективности, варианты которых использованы в разработанных моделях установки лимитов, юнит-экономики, скоринговой модели с изменяющимися во времени параметрами.

Теоретическая значимость исследования заключается в совершенствовании научно-методологических основ управления кредитными рисками с использованием статистических моделей, методов и алгоритмов машинного обучения, учитывающих специфику деятельности микрофинансовой организации. Представлен экономико-математический инструментарий, обеспечивающий поддержку принятия решений в области управления кредитными рисками МФО.

Практическая значимость полученных результатов заключается в возможности уменьшения кредитных потерь и улучшении финансовых результатов микрофинансовой организации за счет лучшего понимания возможности использования различных источников данных для оценки кредитных рисков, за счет применения предлагаемых в диссертации методик, моделей и алгоритмов оценки кредитных рисков при формировании кредитной политики, за счет автоматизации процесса моделирования кредитных рисков с помощью разработанных программно-инструментальных средств. Внедрение полученных результатов работы в деятельность конкретной МФО позволят усовершенствовать систему риск-менеджмента и риск-аналитики.

Разработанные в диссертационном исследовании теоретические и практические рекомендации внедрены и используются при оценке и управлении кредитными рисками в деятельности следующих организаций ООО «Бериберу МКК», ООО «Займиго МФК», ООО «Кредиска МКК», ООО «Информационно-аналитическое агентство Финкарта», ООО «Хурма Кредит МКК», ООО «Морезаем МКК», ООО МКК «Финмолл», ООО МКК «Финфреш», что подтверждено справками о внедрении.

Степень достоверности результатов исследования. Достоверность результатов и выводов диссертационного исследования подтвердилось в ходе моделирования и расчетов на основе имитационных и реальных исторических данных микрофинансовых организаций.

Результаты исследования использовались автором при выполнении следующих прикладных научно-исследовательских работ:

1) «Разработка алгоритмов машинного обучения скоринговых моделей для микрофинансовых организаций», внутренний грант ФГБОУ ВО «РЭУ им. Г.В.Плеханова» (приказ №1172 от 12.10.2016 г.), научный руководитель;

2) «Разработка алгоритмов планирования промышленных экспериментов и статистической обработки их результатов», внутренний грант ФГБОУ ВО «РЭУ им. Г.В.Плеханова» (приказ № 1323 от 25.10.2016 г.), ответственный исполнитель;

3) «Разработка алгоритмов исследования потребительских предпочтений методом совместного (conjoint) анализа на примере московского рынка смартфонов», внутренний грант ФГБОУ ВО «РЭУ им. Г.В.Плеханова» (приказ № 835 от 09.08.2017 г.), научный руководитель;

4) «Разработка методики построения статистических моделей оценки рисков в соответствии со стандартом МСФО 9», внутренний грант РЭУ им. Г.В.Плеханова (приказ № 1047 от 12.07.2018 г.), научный руководитель;

5) «Разработка алгоритмов машинного обучения скоринговых моделей на основе различных источников данных о заемщике», внутренний грант ФГБОУ ВО «РЭУ им. Г.В.Плеханова» (приказ № 942 от 25.06.2019 г.), научный руководитель;

6) «Разработка алгоритмов применения моделей интеллектуального анализа данных для управления рисками в налоговом и таможенном администрировании», внутренний грант ФГБОУ ВО «РЭУ им. Г.В.Плеханова» (приказ № 653 от 01.06.2020, приказ № 1282 от 15.10.2020, приказ № 1468 от 13.11.2020), научный руководитель;

7) «Структурные сдвиги в экономике и обществе по результатам достижения целевых индикаторов реализации Национальных проектов, дающие возможности для организации новых сфер социально-экономической деятельности, в том числе коммерческой, как в России, так и за рубежом», государственное задание Министерства науки и высшего образования FSSW-2020-0010 (исполнитель гос. задания ФГБОУ ВО «РЭУ им. Г.В.Плеханова»), исполнитель работ, 2020-2021 гг.;

8) «Разработка алгоритмов применения моделей интеллектуального анализа данных для управления кредитными рисками микрофинансовых организаций»,

внутренний грант ФГОБУ ВО «РЭУ им. Г.В.Плеханова» (приказ № 661 от 07.06.2021), научный руководитель.

Внедрение и апробация результатов работы. Основные положения и выводы диссертационной работы изложены и обсуждены на международных, всероссийских и региональных научных и научно-практических конференциях, основными из которых являются:

- международная научно-практическая конференция Probability Theory and Mathematical Statistics, г. Казань, Казанский (Приволжский) Федеральный Университет, 6–12 ноября 2017 г.;

- международная научно-практическая конференция «Экономическое развитие России: вызовы и возможности в меняющемся мире», г. Краснодар, Кубанский государственный университет, 24–27 января 2023 г.;

- XII международная научно-практическая конференция «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками, г. Саратов, Саратовский национальный исследовательский государственный университет им. Н.Г.Чернышевского, 16–18 ноября 2023 г.;

- всероссийская (с международным участием) научно-практическая конференция «Измерение и анализ благосостояния», г. Санкт-Петербург, Санкт-Петербургский государственный экономический университет, 25–27 января 2024 г.;

- XIII международная научно-практическая конференция «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками, г. Саратов, Саратовский национальный исследовательский государственный университет им. Н.Г.Чернышевского, 14–16 ноября 2024 г.

Общественное признание. Лауреат 1 степени на 31-м международном конкурсе научно-исследовательских работ, ОНР Всероссийское общество научно-исследовательских разработок ПТСАЙНС, г. Москва, 16 апреля 2023 г.

Публикации. Автором опубликовано более 87 работ. Основные результаты диссертационного исследования нашли отражение в 37 опубликованных работах общим объемом около 46,95 печ. л. (вклад автора 34,31 авт. печ. л.), в том числе: в

20 научных работах общим объемом 28,70 печ. л. (вклад автора 23,29 авт. печ. л.), опубликованных в журналах из Перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, из них 3 статьи опубликованы в рецензируемых изданиях, отнесенных к категории К1, 11 – к категории К2, 5 – к категории К3; в 4 работах, проиндексированных в базе данных Scopus и Web of Science, из них в 2 российских журналах, индексируемых в Q3 (приравненный к категории К1) и Q4 (приравненный к категории К2), в 2 зарубежных журналах, индексируемых в Q2 (приравненный к категории К1) и Q4 (приравненный к категории К3) с общим объемом 3,57 печ. л. (вклад автора 2,25 авт. печ. л.); результаты исследования соискателя представлены в рецензированной монографии общим объемом 9,60 печ. л. (вклад автора 4,80 авт. печ. л.); в 12 других изданиях общим объемом 5,08 печ. л. (вклад автора 3,97 авт. печ. л.). Соискателем получено 3 свидетельства о государственной регистрации программ для ЭВМ.

Структура и объем работы. Диссертационная работа состоит из введения, пяти глав, заключения, списка сокращений и условных обозначений и 13 приложений, изложена на 474 страницах, содержит 89 таблиц, 45 рисунков, 160 формул. Список литературы состоит из 385 наименований.

Во введении обоснована актуальность темы, сформулированы цель и задачи диссертационной работы, определены ее научная новизна и практическая значимость, описан общий план исследования, приведены сведения о внедрении и апробации результатов работы.

В первой главе рассмотрены особенности деятельности МФО, представлены результаты исследования тенденций и закономерностей развития рынка микрофинансирования в России, что позволило сформулировать основные методологические проблемы исследования.

Во второй главе рассмотрены особенности информационных процедур оценки кредитных рисков в микрофинансовых организациях, рассмотрены критерии ценности исходной информации и выбора рационального состава

данных для оценки кредитных рисков, отражена специфика применения метрик качества бинарной классификации при оценке кредитных рисков.

В третьей главе рассмотрены основные математические модели оценки кредитных рисков на основе двух концепций: статистического моделирования и машинного обучения, отражены особенности их применения при оценке кредитных рисков МФО.

В четвертой главе представлены разработанные диссертантом модели управления кредитными рисками МФО включая: модель лимитирования, модель юнит-экономики, модель кредитного скоринга с изменяющимися во времени параметрами.

В пятой главе отражены результаты применения и апробации предложенных автором моделей, методик и алгоритмов оценки и управления кредитными рисками в российских МФО, а также на основе имитационных данных.

В заключении приведены основные результаты и выводы работы.

Глава 1 Теоретико-методологические проблемы оценки и управления кредитными рисками в микрофинансовых организациях

1.1 Тенденции и закономерности развития рынка микрофинансирования в России

Проведение исследования рынка микрофинансирования в России, отражение его особенностей, основных тенденций и закономерностей его развития в условиях систематического изменения и ужесточения регуляторных требований к МФО со стороны Банка России, актуально ввиду существенного влияния таких изменений не только на показатели рынка в целом, но и на результаты деятельности отдельных МФО, их кредитные риски и систему управления кредитными рисками.

Для анализа тенденций и закономерностей развития рынка микрофинансирования в России необходимо использование методов описательной статистики и эконометрики [128], включая анализ количественных и качественных показателей микрофинансового рынка в целом и сегмента PDL-займов, в частности, а также исследование структуры сегмента PDL-займов по основным рисковым метрикам методами многомерного статистического анализа [123] с целью выявления проблем управления кредитными рисками МФО.

Динамика количественных параметров рынка микрофинансирования прослеживается по тенденциям изменения некоторых статистических показателей изменения рынка за последние несколько лет и была подробно рассмотрена в работах автора [115, 116]. В частности, интерес представляют такие показатели, как количество МФО, объем профильных активов, количество заключенных договоров и объем выданных микрозаймов. Динамика количества МФО в России отражена на рисунке 1.1.

Представленные на графике данные показывают ежегодное сокращение количества организаций на рынке микрофинансирования в РФ. В совокупности за

последние 10 лет количество МФО снизилось на 3 243 организации (в относительном выражении – 78,1%, среднегодовой темп снижения составил 11,86%), на начало декабря 2024 года в реестре ЦБ было зарегистрировано 909 действующих МФО.

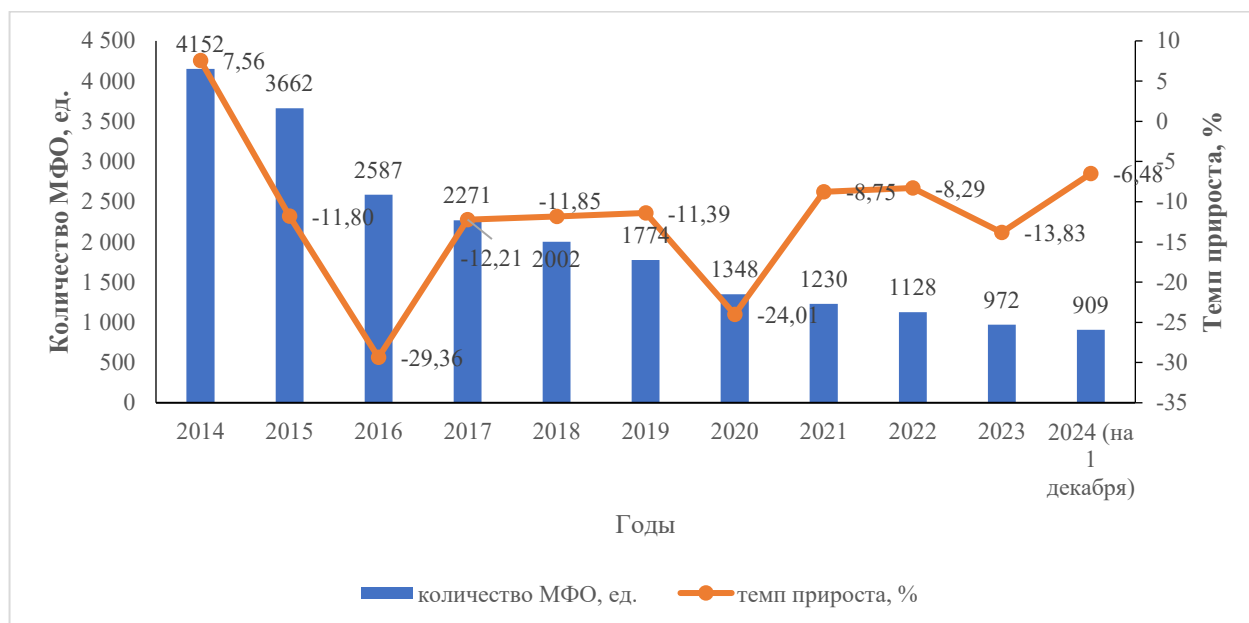


Рисунок 1.1 – Динамика количества МФО в России за 2014–2024 гг. (по состоянию на конец периода)

Источник: составлено автором по данным [23]

Тенденции существенного сокращения количества МФО обусловлены следующими детерминантами:

- во-первых, постоянное изменение регуляторных требований к МФО и ужесточение надзора со стороны Банка России, которое ведет к усложнению деятельности организаций и их нарастающей убыточности и, как следствие, добровольное прекращение деятельности МФО;
- во-вторых, ликвидация МФО по инициативе Банка России ввиду допущенных нарушений регуляторных требований;
- в-третьих, необходимость совершенствования деятельности МФО под давлением объективных внешних факторов (конкурентная среда, растущие требования регулятора, высокие кредитные и другие виды рисков) оставляет в

рыночном поле МФО, имеющие возможность незамедлительно адаптироваться к волатильности условий функционирования.

В отношении совокупного объема профильных активов МФО по физическим лицам ЦБ РФ оперирует данными, приведенными на рисунке 1.2. При этом к профильным активам МФО относится сумма задолженности по основному долгу по выданным микрозаймам на конец отчетного периода.

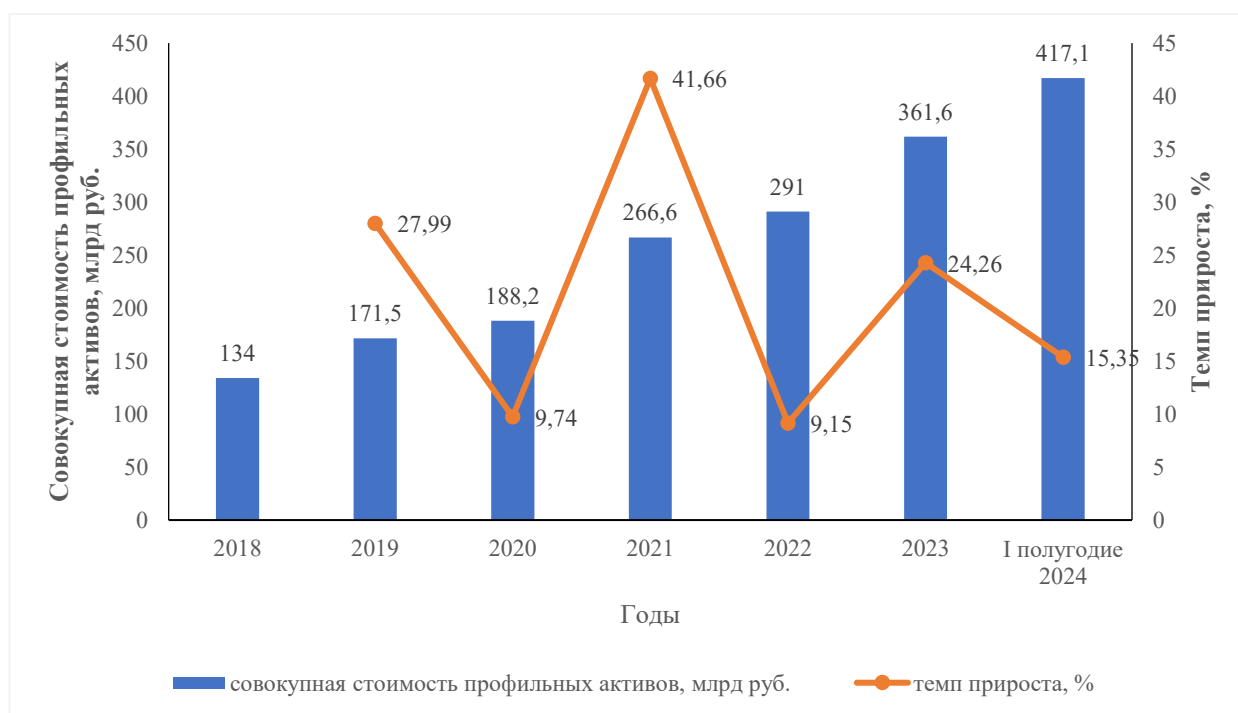


Рисунок 1.2 – Динамика совокупного объема профильных активов МФО по физическим лицам за 2018–2024 гг. (по состоянию на конец периода)

Источник: составлено автором по данным [72]

Представленные на диаграмме данные отражают ежегодное увеличение совокупной стоимости основного долга МФО по заемщикам – физическим лицам. В общей сложности за 5,5 лет рост составил 253,6 млрд руб. (почти в 2,1 раза при среднегодовом темпе прироста 21,2%). При этом темпы прироста профильных активов МФО фиксировали циклическое изменение. Снижение темпов прироста в 2020 и 2022 гг. было обусловлено снижением потребительской активности на фоне кризисных явлений в экономике.

Существенный прирост в абсолютном и относительном выражении, наблюдаемый в 2021 и 2023 гг., обусловлен увеличением количества заемщиков и ростом отложенного спроса на микрозаймы. В том числе на фоне ужесточения требований к заемщикам со стороны банковских организаций наблюдался существенный рост одобрения заявок посредством онлайн-каналов, которые начали активно развиваться после введенных на фоне пандемии в 2020 г. ограничений, что позволяет работать на рынке МФО без географической привязки.

Необходимо отметить, что в условиях санкционного давления или ужесточения требований регуляторов наблюдается снижение доли одобряемых заявок по потребительским кредитам в банковских учреждениях, что ведет к перетоку клиентов в МФО и росту количества одобренных кредитных заявок по микрозаймам (см. рисунок 1.3).



Рисунок 1.3 – Динамика выдачи кредитов МФО за 2018–2024 гг. (данные за отчетный период)

Источник: составлено автором по данным [120]

Увеличение количества вновь заключенных договоров кредитования между МФО и заемщиками – физическими лицами, в совокупности за 2018–2023 гг. составило 45 млн договоров (увеличение в 1,56 раза при среднегодовом темпе прироста 22,1% или на 9 млн договоров ежегодно). Такая динамика сопровождалась увеличением совокупного объема выдачи микрозаймов с 298,5 млрд руб. в 2018 г. до 902,5 млрд руб. в 2023 г. Данная тенденция определяется, в первую очередь, активизацией спроса со стороны существующих и потенциальных заемщиков, в т.ч. в условиях перехода от банковского кредитования к микрофинансовому в связи со снижением уровня одобрения заявок в банковских учреждениях.

Кроме того, на протяжении последних лет наблюдается активное развитие онлайн-каналов микрофинансирования, что видно по данным, представленным на рисунке 1.4.

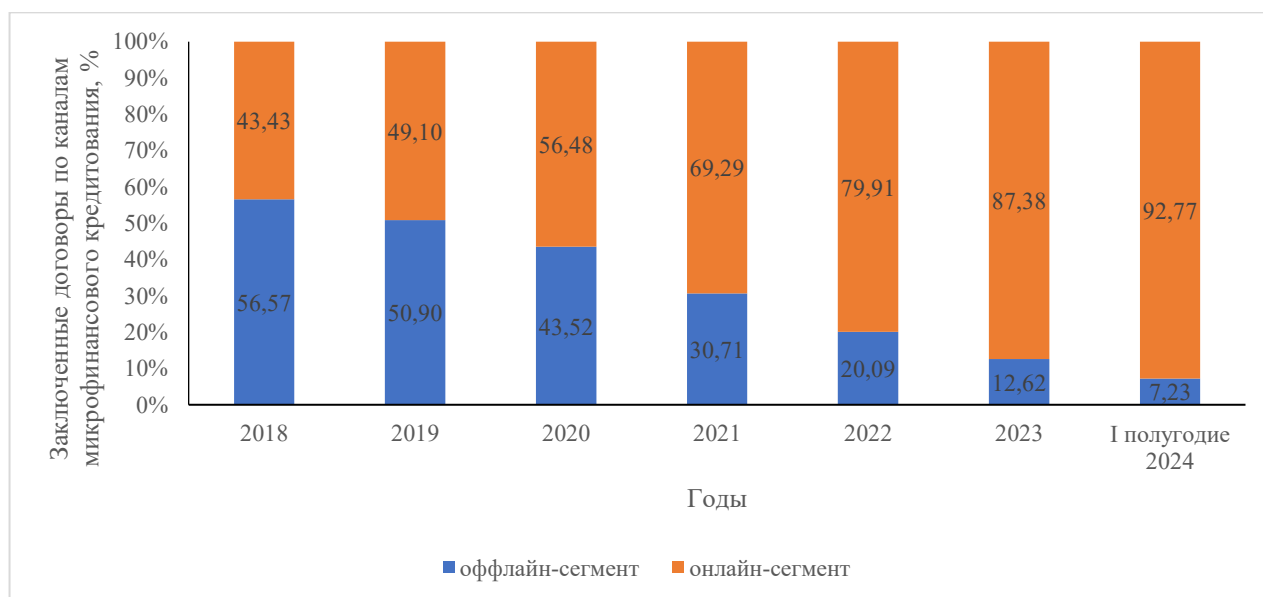


Рисунок 1.4 – Изменение структуры заключенных договоров по каналам микрофинансового кредитования в 2021–2024 гг.

Источник: составлено автором по данным [120]

Из рисунка 1.4 видно существенное развитие онлайн-кредитования на рынке микрофинансирования, которое обусловлено двумя основными причинами

– с одной стороны, более высокая скорость принятия решения о выдаче займа в МФО является основным критерием выбора для заемщика, с другой стороны развитие онлайн-сегмента позволяет МФО функционировать без географической привязки, а также снижать операционные расходы. По состоянию на первое полугодие 2024 г. количество заключенных посредством онлайн-каналов договоров микрозаймов выросло до 92,77% против 43,43% в 2018 г.

Необходимо отметить, что различные сегменты рынка микрофинансирования по-разному реагируют на изменения экономической конъюнктуры и регуляторных требований. Это обуславливает актуальность рассмотрения некоторых количественных показателей рынка в разрезе различных рыночных групп, что также необходимо для сопоставления результатов PDL-сегмента с другими группами с целью выявления особенностей управления кредитными рисками.

Динамика количественных показателей рынка микрофинансирования в разрезе видов микрозаймов представлена в таблице 1.1

Таблица 1.1 – Динамика количественных показателей рынка микрофинансирования в разрезе видов микрозаймов PDL, IL и POS за 2018–2023 гг.

Показатели	2018	2019	2020	2021	2022	2023	Изменение 2023 к 2022	
							абсол.	относит., %
PDL								
Количество активных заемщиков (на конец периода), тыс. чел.	5 133	6 575	6 972	7 345	7 538	8 126	588	7,80
Количество заключенных договоров, тыс. ед.	20 372	22 742	20 306	29 108	33 103	37 834	4 731	14,29
Объем выданных микрозаймов, млрд руб.	143,0	176,0	159,1	242,2	291,2	339,5	48,3	16,59
Портфель микрозаймов (по объему основного долга), млрд руб.	37,7	51,1	54,4	77,0	98,4	114,4	16,0	16,26
IL								
Количество активных заемщиков (на конец периода), тыс. чел.	3 684	4 814	5 408	5 832	6 329	7 129	800	12,64
Количество заключенных	8 498	11 039	11 260	13 204	18 580	32 756	14 176	76,30

Показатели	2018	2019	2020	2021	2022	2023	Изменение 2023 к 2022	
							абсол.	относит., %
договоров, тыс. ед.								
Объем выданных микрозаймов, млрд руб.	155,5	195,7	205,6	231,1	314,1	460,5	146,4	46,61
Портфель микрозаймов (по объему основного долга), млрд руб.	96,3	120,3	133,8	119,5	134,6	180,4	45,8	34,03
POS								
Количество активных заемщиков (на конец периода), тыс. чел.	н/д	н/д	н/д	693	1 373	1 728	355	25,86
Количество заключенных договоров, тыс. ед.	н/д	н/д	н/д	3 452	3 595	3 282	– 313	– 8,71
Объем выданных микрозаймов, млрд руб.	н/д	н/д	н/д	107,3	96,2	101,8	5,6	5,82
Портфель микрозаймов (по объему основного долга), млрд руб.	н/д	н/д	н/д	70,1	57,3	65,8	8,5	14,83

Источник: составлено автором по данным [120]

Из представленных в таблице 1.1 данных можно видеть существенное увеличение количественных показателей как в сегменте PDL-займов, так и в сегменте ПЛ-займов. Однако большее количество активных заемщиков PDL-займов – 8 126 тыс. чел. на конец 2023 г. против 7 129 тыс. чел. заемщиков в сегменте ПЛ-займов, сопровождалось менее значительным объемом выданных микрозаймов – 339,5 млрд руб. и 460,5 млрд руб. по PDL-займам и ПЛ-займам, соответственно.

При этом прирост количества заключенных договоров по ПЛ-займам и прирост по объему кредитования в данном сегменте существенно превышал показатели сегмента PDL-займов. Это связано с тем, что микрозаймы PDL определяются более высокими кредитными рисками, вследствие чего выдаются под более высокие процентные ставки в сравнении с другими видами микрозаймов. Также следует отметить, что данный сегмент подвержен более высоким регуляторным рискам, в частности, снижение максимальной дневной процентной ставки и снижение предельного размера переплаты от тела микрозайма негативно

отразилось на показателях сегмента PDL-займов, в частности, в сторону снижения поступлений от процентных платежей. Исходя из этих предпосылок, большинство МФО переходит из PDL-сегмента в IL-сегмент, в т. ч. с целью привлечения новых клиентов, в первую очередь, заемщиков, которые получили отказ по потребительским кредитам в банках.

На PDL-займах в большей степени акцентируют свой портфель крупные участники рынка микрофинансирования, которые имеют достаточно гибкие и эффективные в условиях регуляторных изменений скоринговые модели и имеют возможность развития онлайн-каналов. В отношении IL-займов можно отметить следующие преимущества для МФО в условиях ужесточения регуляторных требований – возможность снижения процентной ставки при возможности получения большей маржи за счет длительного срока займа, а также снижения кредитной нагрузки заемщика за счет разделения платежей на несколько периодов. POS-займы являются продуктом с более низкой маржинальностью по сравнению с другими видами микрозаймов и, как правило, применяется МФО для расширения кросс-продаж и менее дорогостоящего привлечения новых клиентов. При этом основные POS-займы выдают МФО, принадлежащие крупным банковским институтам. Изменение структуры заключенных договоров по видам микрозаймов отражено на рисунке 1.5.

Наблюдается постепенный переход от модели PDL-займов к модели IL-займов на рынке микрофинансирования. Если в 2021 году доля PDL-займов в объеме заключенных договоров составляла 63,6%, то в первом полугодии 2024 г. она снизилась до 44,92%, а доля IL-займов превысила 50,74%.

Основными причинами такой ситуации является ужесточение регуляторных требований, в частности, снижение предельной процентной ставки, сокращение максимального размера переплаты по договорам микрозаймов, а также ужесточение макропруденциальных лимитов (МПЛ) с сокращением до 25% на необеспеченные микрозаймы.

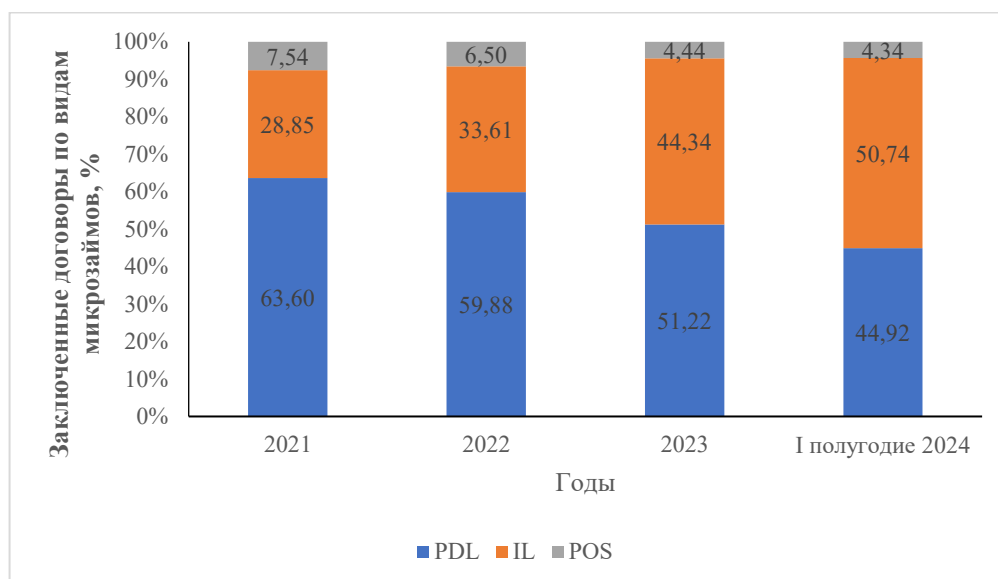


Рисунок 1.5 – Изменение структуры заключенных договоров по видам микрозаймов в 2021–2024 гг.

Источник: составлено автором по данным [120]

Однако камнем преткновения в данном случае стал переход МФО из одного сегмента в другой, что приводит к необходимости пересмотра бизнес-модели и возможного увеличения кредитных рисков.

Изменение среднего количества активных кредитов на одного заемщика по сегментам микрозаймов представлено в таблице 1.2.

Таблица 1.2 – Динамика среднего количества активных кредитов на одного заемщика по сегментам микрозаймов за 2018–2023 гг., ед.

Сегмент микрозаймов	2018	2019	2020	2021	2022	2023
PDL	3,97	3,46	2,91	3,96	4,39	4,66
IL	2,31	2,29	2,08	2,26	2,94	4,59
POS	-	-	-	4,98	2,62	1,90

Источник: составлено автором по данным [120]

Наблюдается ежегодное увеличение среднего количества активных кредитов на одного заемщика в PDL и IL сегментах микрофинансового рынка. Это увеличение на конец 2023 г. составило 4,66 и 4,59 займов соответственно. Это

свидетельствует о достаточно высокой закредитованности заемщиков на рынке «быстрых» займов.

В целом проведенный анализ количественных показателей развития рынка микрофинансирования показал положительную динамику в отношении числа и объемов выданных микрозаймов, что определяется достаточно высоким спросом со стороны заемщиков в условиях ужесточения требований к заемщикам и снижения доли одобряемых заявок по потребительским кредитам со стороны банковских организаций. При этом увеличение количества активных кредитов на одного заемщика обуславливает повышение кредитных рисков МФО.

Далее был проведен анализ тенденций изменения качественных показателей развития рынка микрофинансирования, которые подробно рассмотрены автором в статье [115]. В данном случае интерес представляют такие показатели качественного характера, как средняя сумма выданного кредита, средний срок кредита, средняя полная стоимость кредита (ПСК). Тенденции изменения средней суммы выданного кредита на российском рынке микрофинансирования в совокупности можно проследить по данным рисунка 1.6.

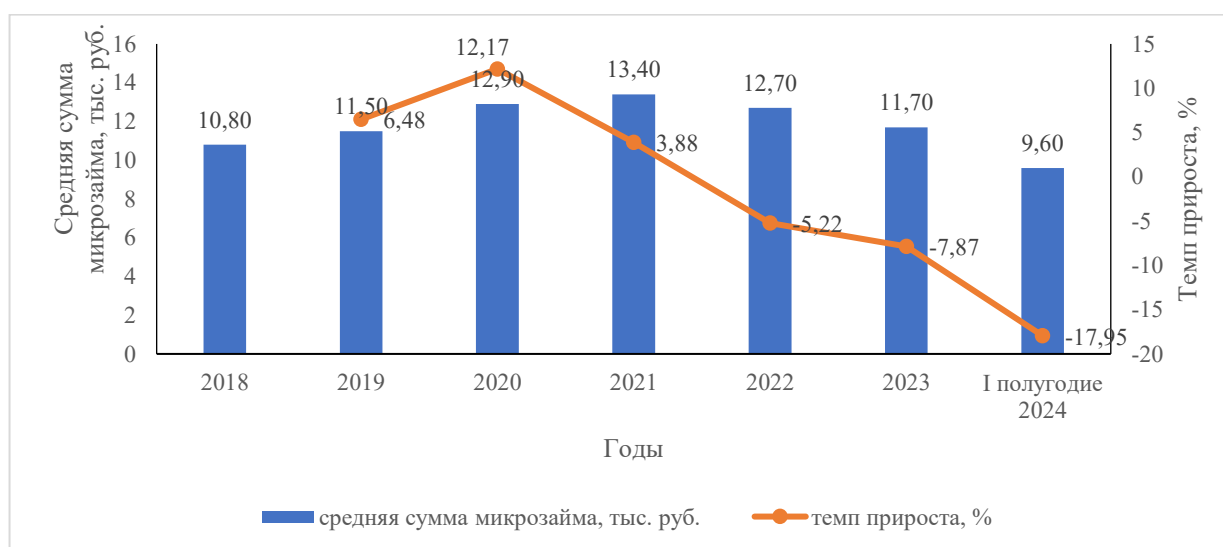


Рисунок 1.6 – Динамика средней суммы выданного кредита на российском рынке микрофинансирования за 2018–2023 гг.

Источник: составлено автором по данным [120]

Ежегодное до 2021 г. увеличение средней суммы выданного микрозайма сменилось снижением данного показателя в 2021–2024 гг. Если в 2021 г. сумма одобренных и выданных займов в МФО составляла в среднем 13,4 тыс. руб., то к концу первого полугодия 2024 г. она снизилась до 9,6 тыс. руб. Данный тренд обусловлен причинами двоякого характера: во-первых, снижением суммы одобрения кредитных заявок на фоне снижения номинальных доходов заемщиков; во-вторых, ужесточением регуляторных требований в отношении показателей долговой нагрузки заемщиков (ПДН).

Кроме того, заемщики обращаются за оформлением микрозаймов в несколько разных МФО с целью получения необходимой суммы, что позволяет перераспределить кредитные риски между МФО в условиях введения макропруденциальных лимитов.

В разрезе отдельных видов микрозаймов были получены результаты, представленные в таблице 1.3.

Таблица 1.3 – Динамика средней суммы выданного кредита на российском рынке микрофинансирования по видам микрозаймов за 2018–2024 гг., тыс. руб.

Сегмент микрозаймов	2018	2019	2020	2021	2022	2023	I квартал 2024
PDL	7,4	8,2	8,0	8,5	12,7	9,2	9,1
IL	17,9	17,4	22,5	17,1	9,0	12,3	9,4
POS	-	-	-	32,9	18,4	31,7	23,2

Источник: составлено автором по данным [120]

По данным таблицы 1.3 можно видеть цикличное изменение средней суммы кредитного договора в отношении основных видов микрозаймов. В отношении PDL-займов существенное снижение средней суммы выдачи в 2023 г. обусловлено ужесточением требований регулятора к ограничению определения средневзвешенной процентной ставки по микрозаймам, для компенсации потерь, по которым МФО вынуждены сокращать одобряемую сумму. Сокращение средней суммы выданного IL-займа определяется переходом МФО на консервативную

политику кредитования в условиях макроэкономической неопределенности и эффективного контроля рисков. Несмотря на конкуренцию со стороны маркетплейсов в отношении выдачи POS-займов, по ним наблюдается прирост средней суммы кредитования. Это обусловлено наращиванием потребительской активности в отношении приобретения более дорогостоящих товаров со стороны заемщиков МФО.

Изменение среднего срока кредитования на российском рынке микрофинансирования представлено на рисунке 1.7.

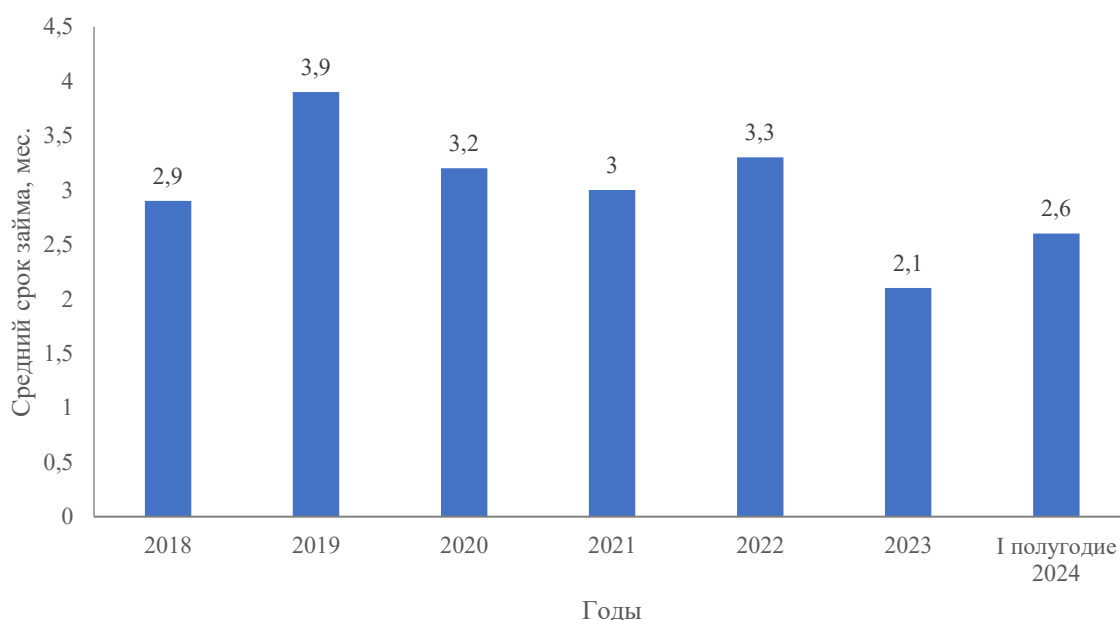


Рисунок 1.7 – Динамика среднего срока кредитования на российском рынке микрофинансирования за 2018–2024 гг.

Источник: составлено автором по данным [43]

Средний срок выданных микрозаймов варьирует в диапазоне 2–4 месяца, что обусловлено значительным количеством выданных среднесрочных ПЛ-займов и POS-займов, которые перекрывают PDL-займы, выдаваемые на срок не более 30 дней.

При этом в 2023 г. наблюдалось снижение среднего по рынку микрофинансирования срока займов до 2,1 мес., несмотря на увеличение сроков

PDL-займов до максимально возможных по требованиям регулятора 30 дней с целью удержания процентных ставок по займам на высоком уровне (в рамках установленных ограничений) при ужесточении требований к макропруденциальным лимитам.

В таблице 1.4 представлено изменение средней полной стоимости кредита на российском рынке микрофинансирования. Сокращение предельной процентной ставки за пользование и размера переплаты по микрозаймам с 1 июля 2023 г. в связи с принятием поправок к ФЗ «О потребительском кредите (займе)» привело к снижению средней полной стоимости кредита (ПСК) до 282% по состоянию на первый квартал 2024 г., что позволяет уменьшить долговую нагрузку заемщиков и соответственно кредитные риски МФО.

Таблица 1.4 – Динамика средней полной стоимости кредита на российском рынке микрофинансирования за 2019–2024 гг., %

Период	2019	2020	2021	2022	2023	2024
I квартал	483	320	312	336	340	282
II квартал	445	326	318	337	343	
III квартал	317	312	319	328	284	
IV квартал	321	302	319	341	279	

Источник: составлено автором по данным [43]

В целом проведенный анализ количественных критериев развития рынка микрофинансирования показал отрицательную динамику показателей средней суммы выданных микрозаймов за счет снижения суммы одобрения кредитных заявок на фоне снижения номинальных доходов заемщиков, а также ужесточения регуляторных требований в отношении показателей долговой нагрузки заемщиков (ПДН). Кроме того, было зафиксировано сокращение сроков выдачи микрозаймов в целом по рынку микрофинансирования, несмотря на пролонгацию PDL-займов до максимального срока. Снижение предельной процентной ставки за пользование и размера переплаты по микрозаймам привело к снижению средней полной стоимости кредита до 282%. Все выявленные тренды обуславливают необходимость снижения кредитных рисков МФО.

Тенденции изменения кредитного риска МФО можно проследить по некоторым статистическим показателям оценки, которые были подробно рассмотрены автором в работе [115].

В таблице 1.5 представлена динамика просроченной задолженности по займам на российском микрофинансовом рынке. Данные таблицы показывают существенную долю просроченной задолженности по микрозаймам по сроку более 90 дней, что обуславливает высокий кредитный риск МФО.

Таблица 1.5 – Динамика просроченной задолженности на российском рынке микрофинансирования за 2019–2024 гг.

Период		Просроченная задолженность до 90 дней (NPL 1–90), %	Просроченная задолженность свыше 90 дней (NPL 90+), %	Доля уступленной задолженности, %
2019	I квартал	7	26	5
	II квартал	8	27	4
	III квартал	8	27	3
	IV квартал	9	28	3
2020	I квартал	9	29	2
	II квартал	9	32	4
	III квартал	8	31	3
	IV квартал	7	29	2
2021	I квартал	9	29	2
	II квартал	11	29	3
	III квартал	10	29	2
	IV квартал	10	30	3
2022	I квартал	12	31	3
	II квартал	11	35	3
	III квартал	10	35	6
	IV квартал	10	35	4
2023	I квартал	10	34	4
	II квартал	10	34	5
	III квартал	10	34	4
	IV квартал	9	33	6
2024	I квартал	10	32	5
	II квартал	11	31	5

Источник: составлено автором по данным [43]

При этом со второго квартала 2022 года данный показатель варьировал в пределах 33–35%, небольшое снижение в первом и втором квартале 2024 года (до

32% и 31%, соответственно) обусловлено воздействием макропруденциальных лимитов.

В отношении доли уступок права требования по просроченной задолженности МФО наблюдается ее вариация в пределах 4–6% за последние два года. Необходимо отметить, что при заключении договора микрозаймов заемщику дается выбор, посредством которого он может отказаться от возможности переуступки прав требований по кредиту третьим лицам и сторонним организациям. В данном случае для ограничения кредитного риска перераспределяют задолженность между МФО внутри группы, это в большей степени касается крупных МФО, в т. ч. принадлежащим банкам.

Еще одним важным показателем кредитного риска МФО выступает «Индекс кредитной сознательности»¹⁾ заемщика (ИКС). По данным БКИ «Скоринг Бюро» [43] определены следующие тренды изменения ИКС (см. рисунок 1.8).

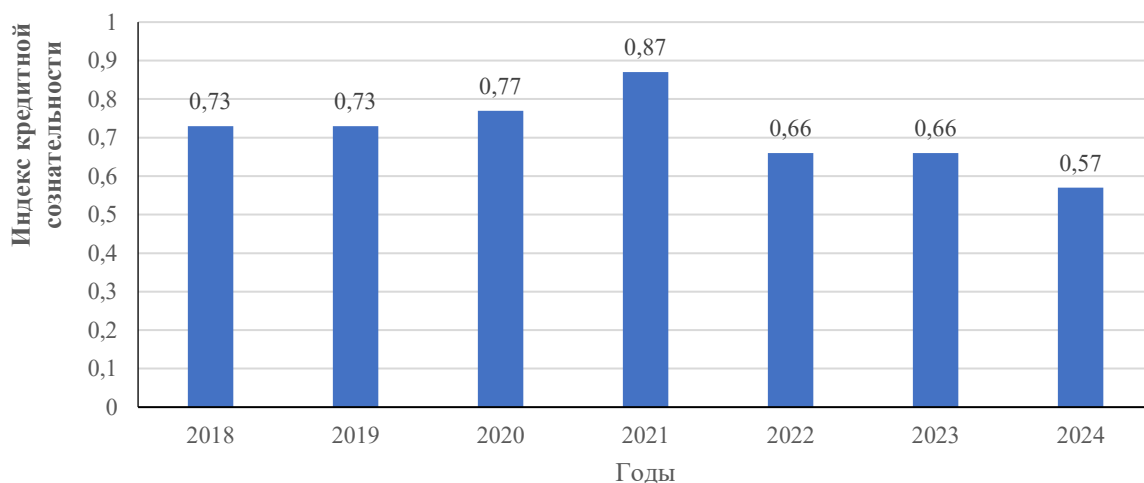


Рисунок 1.8 – Динамика индекса кредитной сознательности на российском рынке микрофинансирования за 2019–2024 гг.

Источник: составлено автором по данным [43]

¹⁾ Индекс кредитной сознательности 2.0 (Скоринг Бюро) – это математическая модель платежного поведения заемщиков, разработанная аналитиками бюро кредитных историй (БКИ) «Скоринг Бюро» на основе данных о платёжном поведении клиентов российских банков при выплате кредитов. При расчёте индекса учитываются любые виды необеспеченных ссуд. Рост индекса означает, что кредитное поведение банковских клиентов улучшается, а снижение – ухудшение платежной дисциплины.

Из представленных данных видно циклическое изменение ИКС на микрофинансовом рынке России. Максимальное значение ИКС (0,87) было достигнуто в 2021 году, с дальнейшим снижением до 0,57 в 2024 году. Данный тренд обуславливает повышение кредитного риска МФО ввиду ухудшения кредитного поведения заемщиков и снижения качества портфеля микрозаймов.

Кроме того, для оценки кредитного риска МФО используется винтажный анализ, который позволяет контролировать качество портфеля микрозаймов путем оценки доли безнадежной задолженности по поколениям выдачи микрозаймов ежемесячно. При этом винтажный анализ производится как в отношении доли по сумме (отношение просроченной задолженности более 30 дней к общей сумме выданных микрозаймов), так и по доле по количеству (отношение количества просроченных к общему количеству выданных микрозаймов).

Винтажный анализ качества портфеля микрозаймов, разработанный в исследовании [43], позволяет сделать вывод о том, что чем старше поколение выданных микрозаймов, тем выше доля просроченной задолженности по ним. При этом при оценке доли по сумме она варьирует в диапазоне от 9,0% до 16,5% в зависимости от поколения микрозаймов (от 2 до 6 месяцев), при оценке доли по количеству данный показатель варьирует от 10,5% до 19,5%, соответственно. Данный тренд также обуславливает достаточно высокий уровень кредитного риска МФО.

В целом проведенный анализ изменения кредитного риска МФО позволил определить следующие тенденции:

- во-первых, существенная доля просроченной задолженности по микрозаймам по сроку более 90 дней, небольшое снижение по которой было зафиксировано в первом квартале 2024 года под воздействием макропруденциальных лимитов;

- во-вторых, для ограничения кредитного риска перераспределяют задолженность между МФО внутри группы, это в большей степени касается крупных МФО, в т. ч. принадлежащим банкам;

- в-третьих, снижение индекса кредитной сознательности обуславливает повышение кредитного риска МФО ввиду ухудшения кредитного поведения заемщиков и снижения качества портфеля микрозаймов.

Учитывая ужесточения регулирования микрофинансового рынка и дальнейшее усиление регуляторного воздействия, можно констатировать, что снижение процентных ставок по микрозаймам ведет к дальнейшей экспансии сегмента ПЛ-займов, которые становятся более привлекательными и востребованными со стороны МФО с точки зрения выгоды как для МФО (возможность снижения процентной ставки при возможности получения большей маржи за счет длительного срока займа путем включения дополнительных сборов к сумме займа), так и для заемщика (снижения кредитной нагрузки за счет разделения платежей на несколько периодов). Однако камнем преткновения в данном случае является переход МФО из одного сегмента в другой, что приводит к необходимости пересмотра бизнес-модели и возможного увеличения кредитных рисков из-за отсутствия адекватных моделей оценки заемщиков, что определяет актуальность проводимого исследования.

Дальнейшее снижение макропруденциальных лимитов с целью уменьшения долговой нагрузки населения является одним из путей сдерживания выдачи займов сильно закредитованных заемщиков [73]. При этом меньшее по сравнению с банками лимитирование ведет к перетоку заемщиков в МФО. В данном контексте следует отметить, что в последние 3 года наблюдается тенденция открытия банками собственных структур на рынке микрофинансирования. Многие крупные банки уже имеют свои МФО: Альфа банк (МКК «А Деньги»), Росбанк (МКК «РБ Кредит»), Т-банк (МФК «Т-финанс»), Совкомбанк (МКК «Быстрые покупки»), ОТП банк (МФК «ОТП финанс»), Банк Синара (МКК «СКБ-финанс»), Уралсиб (МКК «Уралсиб финанс») и Экспобанк (МФК «Кэшдрайв»).

Выход крупных банков на рынок микрофинансирования объясняется возрастающей конкуренцией между банковскими учреждениями и МФО, что особенно обостряется в ключе ужесточения регуляторных требований. Формирование собственных МФО позволяет банкам наращивать долю на рынке

потребительского кредитования при одновременном сохранении и расширении клиентской базы. При этом преимуществом для МФО, созданных под началом крупных банковских учреждений, является сформированная база управления кредитными рисками, успешные скоринговые модели оценки заемщиков и другие рычаги и ресурсы. С другой стороны, действующие на рынке МФО, не относящиеся к банковским группам, с целью повышения своих конкурентных позиций будут вынуждены повышать качество управления кредитными рисками, в том числе за счет разработки актуальных моделей оценки заемщиков.

Несмотря на снижение маржинальности бизнеса МФО под воздействием регуляторных ограничений, существует прямая связь данных мер с повышением качества кредитного портфеля в связи с возможным повышением кредитоспособности заемщиков на фоне снижения их долговой нагрузки. Однако в данном случае нельзя упускать из виду показатель кредитной сознательности заемщиков, снижение данного индекса возможно под воздействием поведенческих предпосылок, оценка которых является актуальной при выработке моделей управления кредитными рисками в МФО.

Наиболее крупные игроки – лидеры микрофинансового рынка, которые имеют возможность развития онлайн-каналов выдачи микрозаймов, а также обладают более эффективными и гибкими системами управления кредитными рисками в условиях ужесточения регуляторных требований, формируют портфель с акцентом на PDL-займы.

Таким образом, далее автором в работе [94] была проведена оценка изменения структуры рынка в сегменте PDL-займов (преимущественно в онлайн и оффлайн) в пятилетней динамике (2019–2023 гг.) в зависимости от ухудшения рискованных метрик портфеля займов и необходимости грамотного управления кредитными рисками. Для исследования использованы методы многомерного статистического анализа: компонентный анализ и иерархический кластерный анализ [161], который позволяет разделить исследуемые компании на рынке PDL-займов России по соответствующим сегментам. Методики применения

многомерного статистического анализа для анализа социально-экономических, финансовых данных отражены и апробированы в работах автора [84, 86, 97, 108].

В настоящее время в России отсутствуют официальные статистические данные по показателям деятельности микрофинансовых организаций, а ведение статистики для МФО является добровольным, что приводит к проблеме раздробленности и неполноты данных. Подробные данные отчетности МФО в ЦБ России являются закрытыми. Сбором соответствующих статистических данных занимаются только частные рейтинговые агентства. В работе [94] для проведения анализа были использованы данные рейтингового агентства «Эксперт РА» по рэнкингам микрофинансовых организаций РФ [88].

Для проведения анализа взяты статистические данные по следующим абсолютным и относительным показателям PDL-портфеля, характеризующим масштаб и качество кредитного портфеля МФО, по состоянию на 2019–2023 гг.:

- X_1 – размер портфеля PDL на соответствующую дату (млн. руб. по состоянию на 1 января последующего за отчетным года);
- X_2 – темп роста портфеля PDL за анализируемый период (%);
- X_3 – размер портфеля PDL без NPL90+¹⁾ на соответствующую дату (млн. руб. по состоянию на 1 января последующего за отчетным года);
- X_4 – темп роста портфеля PDL без NPL90+ за анализируемый период (%);
- X_5 – объем выданных PDL за анализируемый период (млн. руб. за отчетный год);
- X_6 – темп роста объема выданных PDL за анализируемый период (%).

Размер портфеля PDL организации отражает его величину в абсолютном выражении и иллюстрирует долю, которую занимает МФО в секторе. Размер портфеля PDL без NPL90+ организации означает, что берется величина кредитного портфеля без показателя, характеризующего объем непогашенной задолженности, по которой выплата основного долга и процентов просрочена на 90 дней и более.

¹⁾ NPL90+ – показатель, который характеризует величину просроченной задолженности сроком более 90 дней в кредитном портфеле МФО. Такие кредиты относятся к повышенному уровню риска.

Объем выданных PDL-займов, как один из основных способов увеличения портфеля микрофинансовой организации, позволяет охарактеризовать активность компании в анализируемом периоде. Это позволяет зафиксировать скорость обновления портфеля и косвенно оценить его сроки и качество. Относительные показатели рассчитаны соответственно из абсолютных показателей за соответствующие периоды.

В таблице 1.6 представлены основные описательные статистики используемых исходных данных для проведения многомерного анализа.

Таблица 1.6 – Описательные статистики используемых данных

Год	2019	2020	2021	2022	2023
X ₁	1170 ± 1658 523 [123; 1368]	1044 ± 1662 390 [82; 1457]	1925 ± 2761 1142 [219; 2206]	2940 ± 3866 1783 [679; 3407]	3705 ± 4082 3053 [572; 4837]
X ₂	58 ± 121 41 [- 10; 90]	36 ± 114 7 [- 27 ;52]	122 ± 163 59 [30; 112]	26 ± 42 19 [6; 49]	49 ± 84 24 [- 3; 50]
X ₃	387 ± 488 140 [36; 806]	392 ± 532 161 [43; 584]	903 ± 1160 523 [188; 1013]	1107 ± 1242 723 [244; 1427]	1670 ± 1528 1327 [322; 2871]
X ₄	60 ± 161 21 [- 44; 80]	88 ± 265 3 [- 11; 51]	200 ± 402 90 [41; 146]	26 ± 64 10 [- 23; 63]	50 ± 66 39 [5; 77]
X ₅	2696 ± 3094 1328 [357; 4792]	2700 ± 4158 893 [262; 3990]	5996 ± 8697 2870 [997; 6648]	8311 ± 10898 4032 [2153; 14237]	11057 ± 11001 6525 [2150; 19474]
X ₆	48 ± 117 37 [- 12; 54]	43 ± 176 - 2 [- 34; 49]	124 ± 151 87 [33; 132]	27 ± 27 23 [9; 37]	54 ± 114 27 [- 1; 53]
примечание – показатели представлены в виде M ± SD – среднее значение и стандартное отклонение; Me [Q ₁ ; Q ₃] – медиана, 1 и 3 квартиль.					

Источник: составлено автором по самостоятельно выполненным расчетам

По всем показателям за весь период мы наблюдаем достаточно сильный разброс данных, коэффициент вариации, как правило, более 100%, стандартное отклонение значительно превышает среднее значение. Это говорит о значительной неоднородности данных и необходимости кластеризации исследуемых МФО на однородные группы.

При проведении исследования для каждого года применялась методика, изложенная в работе [102]. Первым этапом проводилось тестирование данных на

нормальность [301, 355] и их нормализация в случае необходимости. Исходя из описательной статистики практически по всем исходным 6 переменным медиана значительно меньше среднего значения, т. е. распределение данных экспоненциально или логарифмически нормально с длинным пологим правым «хвостом» в оценке кривой распределения плотности вероятности. Поэтому исходные данные были приближены к нормальному распределению при помощи логарифмического преобразования Бокса-Кокса со значением $\lambda=0$.

Далее по логарифмам исходных переменных была рассчитана корреляционная матрица. Выбранные показатели для анализа структуры рынка PDL-займов сильно коррелированы: объем портфеля без просрочки естественно сильно зависит от общего объема портфеля и от объема выданных займов за год (коэффициенты корреляции более 0,9), темпы роста объема выдачи влияют на темпы роста объема портфеля (корреляция 0,61). Поэтому на основе корреляционной матрицы было проведено снижение размерности исходного пространства признаков методом главных компонент. Отбирались для дальнейшего анализа компоненты с собственными значениями больше 1, информативность которых выше исходных переменных (см. таблицу 1.7).

Таблица 1.7 – Собственные значения полученных главных компонент

Главная компонента (ГК)	Собственное значение	Объясняемая дисперсия, %	Суммарная объясняемая дисперсия, %
2019 год			
ГК1	2,97	49,41	75,25
ГК2	1,55	25,84	
2020 год			
ГК1	2,84	47,40	80,34
ГК2	1,98	32,94	
2021 год			
ГК1	2,91	48,56	94,69
ГК2	2,77	46,14	
ГК2	2,80	46,74	
2022 год			
ГК1	2,77	46,11	83,35
ГК2	2,24	37,25	
2023 год			

Главная компонента (ГК)	Собственное значение	Объясняемая дисперсия, %	Суммарная объясняемая дисперсия, %
ГК1	2,86	47,70	94,44
ГК2	2,80	46,74	

Источник: составлено автором по самостоятельно выполненным расчетам

Видно, что каждый год стабильно выделялось две наиболее информативные главные компоненты. При этом в целом информативность полученных компонент со временем возросла: суммарный процент объясняемой дисперсии превышает 80–90% в 2021–2023 гг.

Для упрощения интерпретации извлеченных главных компонент проводилось вращение матрицы факторных нагрузок методом Варимакс с нормализацией Кайзера. В таблице 1.8 представлена матрица факторных нагрузок главных компонент после вращения, значения в ячейках которой представляют собой коэффициенты корреляции главных компонент и исходных логарифмированных переменных.

Практически за все года видна стабильность структуры получаемых главных компонент. Первая компонента коррелирует сильно с показателями портфеля PDL-займов на конец года, темпами роста портфеля за год и объемом качественного портфеля без просроченной задолженности. Поэтому данную компоненту можно обозначить как фактор масштаба МФО.

Вторая компонента сильно коррелирует с объемом выдачи за года и темпами его роста, а также темпом роста качественного портфеля без просроченной задолженности свыше 90 дней, что отражает качество управления кредитными рисками в МФО. Поэтому вторую главную компоненту можно обозначить как фактор динамики роста МФО за счет качественного управления кредитными рисками. В 2020 году структура матрицы нагрузок разделила компоненты на два тривиальных фактора: влияние абсолютных и относительных показателей.

На заключительном этапе исследования по полученным значениям главных компонент проводилась кластеризация МФО-лидеров рынка. Кластерный анализ проводился методом иерархических агломеративных процедур, которые из-за

небольшого числа МФО позволяют наглядно визуализировать полученные результаты в виде графика расстояний между кластерами (дендрограммы). Расстояние между наблюдениями (МФО) в кластерах рассчитывалось на основе квадрата Евклидова расстояния [286], а расстояние между кластерами – при помощи метода Варда, который позволяет максимизировать дисперсию между кластерами и минимизировать дисперсию внутри получаемых кластеров [258].

Таблица 1.8 – Матрица факторных нагрузок главных компонент после вращения

Показатели (после логарифмирования)	ГК 1: масштаб МФО	ГК2: динамика МФО
2019 год		
Портфель PDL, млн руб. на конец года	0,98	– 0,12
Темп роста портфеля PDL за год, %	0,97	– 0,20
Портфель PDL без NPL90+, млн руб. на конец год	0,94	– 0,18
Темп роста портфеля PDL без NPL90+ за год, %	0,09	0,76
Объем выданных PDL за год, млн руб.	– 0,34	0,72
Темп роста объема выдачи PDL за год, %	– 0,23	0,61
2019 год		
Портфель PDL, млн руб. на конец года	0,98	– 0,12
Темп роста портфеля PDL за год, %	0,97	– 0,20
Портфель PDL без NPL90+, млн руб. на конец год	0,94	– 0,18
Темп роста портфеля PDL без NPL90+ за год, %	0,09	0,76
Объем выданных PDL за год, млн руб.	– 0,34	0,72
Темп роста объема выдачи PDL за год, %	– 0,23	0,61
2020 год		
Портфель PDL, млн руб. на конец года	0,97	0,20
Темп роста портфеля PDL за год, %	0,01	0,91
Портфель PDL без NPL90+, млн руб. на конец год	0,97	0,20
Темп роста портфеля PDL без NPL90+ за год, %	0,30	0,56
Объем выданных PDL за год, млн руб.	0,94	0,09
Темп роста объема выдачи PDL за год, %	0,001	0,87

Показатели (после логарифмирования)	ГК 1: масштаб МФО	ГК2: динамика МФО
2021 год		
Портфель PDL, млн руб. на конец года	0,99	– 0,05
Темп роста портфеля PDL за год, %	0,96	– 0,15
Портфель PDL без NPL90+, млн руб. на конец год	0,96	– 0,21
Темп роста портфеля PDL без NPL90+ за год, %	– 0,15	0,97
Объем выданных PDL за год, млн руб.	– 0,04	0,94
Темп роста объема выдачи PDL за год, %	– 0,23	0,93
2022 год		
Портфель PDL, млн руб. на конец года	0,98	0,09
Темп роста портфеля PDL за год, %	0,95	0,26
Портфель PDL без NPL90+, млн руб. на конец год	0,90	0,34
Темп роста портфеля PDL без NPL90+ за год, %	0,17	0,88
Объем выданных PDL за год, млн руб.	0,20	0,81
Темп роста объема выдачи PDL за год, %	0,20	0,79
2023 год		
Портфель PDL, млн руб. на конец года	0,97	0,21
Темп роста портфеля PDL за год, %	0,97	0,14
Портфель PDL без NPL90+, млн руб. на конец год	0,96	0,12
Темп роста портфеля PDL без NPL90+ за год, %	0,12	0,96
Объем выданных PDL за год, млн руб.	0,15	0,95
Темп роста объема выдачи PDL за год, %	0,19	0,95

Источник: составлено автором по самостоятельно выполненным расчетам

По результатам проведенного кластерного анализа структуру рынка PDL-займов можно отразить 2–3 сегментами в зависимости от года. Количественный состав и средние значения главных компонент в каждом из кластеров представлены в таблице 1.9.

Поскольку значения главных компонент стандартизованы, мы можем на основе их средних значений в кластерах дать описание сегментов.

В 2019 году было получено по факту 2 кластера. Первый кластер представлен крупнейшими игроками на рынке (21% МФО), которые выделяются большим масштабом (ГК1 равна 1,16) и положительным ростом динамики (ГК2 равна 0,59). В следующий кластер вошли 14 компаний или 74% исследуемых МФО, у данных компаний меньший размер портфеля (ГК1 = – 0,42) и обычный для рынка динамика роста бизнеса (ГК = 0,00). Компания ГК «Финбридж» («Деньги сразу») представляет отдельное наблюдение с большим портфелем и отрицательной динамикой роста.

Таблица 1.9 – Средние значения главных компонент в кластерах

Год		2019	2020	2021	2022	2023
Кластер 1	N (%)	4 (21)	13 (68)	15 (71)	11 (58)	7 (37)
	ГК 1	1,16	0,54	0,50	0,57	0,91
	ГК 2	0,59	– 0,33	– 0,06	– 0,18	– 0,43
Кластер 2	N (%)	1 (5)	3 (16)	1 (5)	5 (26)	5 (26)
	ГК 1	1,28	– 0,44	– 0,80	– 0,47	0,24
	ГК 2	– 0,24	1,92	3,71	1,28	1,18
Кластер 3	N (%)	14 (74)	3 (16)	5 (24)	3 (16)	7 (37)
	ГК 1	– 0,42	– 1,24	– 1,33	– 1,30	– 1,08
	ГК 2	0,00	0,13	– 0,56	– 1,46	– 0,41
примечание – N (%) – количество и процент МФО в кластерах						

Источник: составлено автором по самостоятельно выполненным расчетам

В аномальный 2020 год пандемии COVID-19 структура показателей на рынке поменялась. В первый полученный кластер вошло большинство компаний (13 или 68%), которые характеризуются еще достаточно большим фактором масштаба (ГК1 = 0,62) и отрицательным значением фактора динамики (ГК2 = – 0,33). Три компании (16%) в 2020 году характеризуются пониженным значением фактора масштаба (ГК1 = – 0,44) и достаточно большим значение фактора динамики (ГК2 = 1,92), в год пандемии эти компании смогли показать хорошие темпы роста бизнеса. И оставшиеся 3 МФО имеют самый маленький масштаб и небольшую положительную динамику роста (ГК1 = – 1,24 и ГК2 = 0,13).

В 2021 компании разделились снова на 2 кластера. Первый кластер похож по составу и характеристикам на аналогичный кластер в 2019 году. В него вошли 15 (71%) МФО, которые сохранили положительное значение фактора масштаба ($ГК1 = 0,50$) и достигли около среднего уровня фактора динамики роста ($ГК = - 0,06$). Второй же кластер по численности (5 или 24% МФО) имеет, как и в 2019 году, низкое значение главной компоненты 1 ($- 1,33$) и отрицательное значение главной компоненты 2 ($- 0,56$). Одной компании «ЦФП» («VIVA Деньги») в 2021 году удалось при отрицательном значении фактора масштаба ($ГК1 = - 0,8$) достичь более 3 стандартных отклонений фактора динамики роста ($ГК2 = 3,71$).

В последующие два года эта структура рынка PDL-займов не сильно менялась, но наблюдалось изменение состава кластеров. В 2022 году в первый кластер с положительным значением $ГК1$ (0,57) и отрицательным значением $ГК2$ ($- 0,18$) вошло больше всего компаний (58%). Число компаний в депрессивном кластере с отрицательными значениями первых двух главных компонент ($- 1,3$ и $- 1,46$) сократилось. Также часть компаний (5 или 26%) перешли в кластер с отрицательным фактором масштаба ($- 0,47$), но при этом самым большим значением фактора динамики роста (1,28).

В 2023 году 1 и 2 кластер оказались похожими на аналогичные в 2022 году, но объем первого кластера стал меньше. В первый кластер со средним значением $ГК1 = 0,91$ и $ГК2 = - 0,43$ вошло уже только 7 компаний (37%). Чуть больше стало компаний (7%) в депрессивном кластере с отрицательными значениями главных компонент ($- 1,08$ и $- 0,41$). И только 5 компаний впервые после 2020 года показали положительные значения обоих факторов – масштаба и динамики роста (0,24 и 1,18 соответственно).

В таблице 1.10 представлены средние и медианные значения исходных переменных по полученным кластерам.

Одним из критериев эффективности результатов кластерного анализа является незначительная разница между средними и медианными значениями внутри кластеров и существенное различие значений показателей между кластерами.

По многим показателям внутри кластеров медианы не сильно отличаются от средних. Имеющиеся существенные различия для некоторых показателей между средним и медианой вызвано явно асимметричным распределением исходных показателей, для их интерпретации лучше использовать медиану.

Таблица 1.10 – Средние / медианные значения исходных переменных по кластерам

Показатель	Кластер 1	Кластер 2	Кластер 3
2019 год			
Портфель PDL, млн руб.	2901 / 2464	-	382 / 250
Темп роста портфеля PDL за год, %	45 / 40	-	64 / 42
Портфель PDL без NPL90+, млн руб.	1142 / 1024	-	131 / 84
Темп роста портфеля PDL без NPL90+ за год, %	36 / 18	-	67 / 4
Объем выданных PDL за год, млн руб.	5795 / 5612	-	1163 / 922
2020 год			
Портфель PDL, млн руб.	1538 / 853	235 / 290	45 / 47
Темп роста портфеля PDL за год, %	- 1 / - 8	249 / 209	5 / 7
Портфель PDL без NPL90+, млн руб.	570 / 407	126 / 149	14 / 16
Темп роста портфеля PDL без NPL90+ за год, %	93 / 3	220 / 175	- 4 / - 8
Объем выданных PDL за год, млн руб.	3969 / 1428	633 / 736	159 / 160
2021 год			
Портфель PDL, млн руб.	2634 / 1854	-	150 / 150
Темп роста портфеля PDL за год, %	109 / 59	-	48 / 20
Портфель PDL без NPL90+, млн руб.	1220 / 794	-	109 / 101
Темп роста портфеля PDL без NPL90+ за год, %	124 / 115	-	91 / 78
Объем выданных PDL за год, млн руб.	8139 / 3528	-	700 / 790
2022 год			
Портфель PDL, млн руб.	4121 / 2761	2020 / 1783	141 / 91
Темп роста портфеля PDL за год, %	22 / 19	70 / 91	- 30 / - 44
Портфель PDL без NPL90+, млн руб.	1592 / 1277	655 / 478	82 / 30
Темп роста портфеля PDL без NPL90+ за год, %	13 / 10	90 / 71	- 32 / - 31
Объем выданных PDL за год, млн руб.	12542 / 8460	3399 / 3592	983 / 1008
2023 год			
Портфель PDL, млн руб.	7031 / 4977	3461 / 3517	552 / 431
Темп роста портфеля PDL за год, %	23 / 31	143 / 103	6 / 14
Портфель PDL без NPL90+, млн руб.	2911 / 2871	1896 / 1633	268 / 311
Темп роста портфеля PDL без NPL90+ за год, %	38 / 58	124 / 106	11 / 8
Объем выданных PDL за год, млн руб.	20129 / 19474	10850 / 7778	2132 / 2134

Источник: составлено автором по самостоятельно выполненным расчетам

В то же время между кластерами для каждого года и для каждого показателя деятельности МФО получились отличные на порядок средние/медианные значения ($p < 0,001$, сравнение проводилось параметрическими и непараметрическими тестами). Это подтверждает качество разбиения МФО на разнородные кластеры.

По состоянию на конец 2023 года в России осуществляло деятельность 972 МФО, при этом в рэнкинге «Эксперт РА» по состоянию на 2023 год приведены данные по 61 МФО, ввиду того, что в участие в опросе РА «Эксперт» добровольное и не все МФО принимают в нем участие. Кроме того, не по всем участвующим в рэнкинге МФО есть полные данные по всем показателям. Соответственно по факту отсутствует возможность проанализировать показатели по всему российскому рынку PDL-займов, что может вызвать вопрос о достаточности объема выборки МФО, взятой для анализа и ее репрезентативности. Автор исследования осознает эту проблему и делает следующие допущения. Для отражение общего тренда изменения структуры рынка PDL-займов принято решение о пригодности данных по основным (присутствующим в рэнкинге РА «Эксперт») представителям на рынке микрофинансирования. То есть за каждый исследуемый период для проведения кластерного анализа брали примерно 20 МФО, специализирующихся на PDL-займах, с валидными данными по всем показателям, адекватно отражающих основные тенденции на рынке.

Предложенная методика построения агрегированных показателей для кластеризации валидна. Полученные в ходе анализа данные свидетельствуют о стабильности результатов компонентного анализа со временем, за исключением 2020 года. Отличная структура факторных нагрузок в 2020 от остальных периодов может быть объяснена пандемией COVID-19, разгар которой пришелся на 2020 год. Объявленный карантин в 2020 году повлиял на финансовый рынок России, в том числе скорректировал объемы выдачи и изменение динамики роста российских МФО.

Также следует отметить, что несмотря на стабильность структуры факторных нагрузок получившихся интерпретируемых факторов: масштаба МФО и динамики ее роста за счет качества управления кредитным портфелем, состав и описание кластеров менялся год от года. Метод кластерного анализа адекватно отражает все конъюнктурные изменения на рынке PDL-займов. После 2020 года мы можем отметить факт выделения кластера с положительной динамикой роста за счет качества управления кредитными рисками, который в наибольшей степени виден в 2023 году.

Многим компаниям удалось войти в такой кластер за счет стратегии роста объема выдачи несколькими брендами. При такой стратегии объем выдачи портфеля достигается ни одним брендом, а совокупностью брендов (группой компаний), принадлежавших одним и тем же собственникам. Такая диверсификация портфеля требует также и более качественного унифицированного подхода управления кредитными рисками. Фактическое количество компаний для кластеризации было реально больше 20 МФО каждый год, т. к. многие компании были объединены в один бренд из структуры капитала собственников.

Необходимо также отметить, собственные значения второй главной компоненты (фактора динамики роста МФО) растут после 2020 года, вес ее увеличивается. В 2023 году процент объясняемой дисперсии второй компоненты сопоставим с процентом первой компоненты. Год пандемии дал многим МФО возможность перестроить бизнес и показать динамичный рост не только объема выдачи, но и рост объема качественного портфеля без просрочки, что подчеркивает важность управления кредитными рисками – как одного из первого фактора, формирующего финансовый результат МФО.

Проведенный кластерный анализ позволяет сделать вывод о стабильности структуры полученных главных компонент, которые были обозначены как фактор масштаб МФО и фактор динамики роста МФО за счет качественного управления кредитными рисками. По результатам проведенного кластерного анализа структура рынка PDL-займов была разделена на 2–3 сегмента в зависимости от

года. Предложенная методика построения агрегированных показателей для кластеризации валидна, т. к. полученные данные свидетельствуют о стабильности результатов компонентного анализа со временем. Метод кластерного анализа адекватно отражает все конъюнктурные изменения на рынке PDL-займов – состав кластеров изменяется в зависимости от масштаба МФО и динамики ее роста за счет качества управления кредитным портфелем. Результаты описательного анализа трендов российского микрофинансового рынка и многомерного анализа изменений в структуре PDL-займов не противоречат друг другу.

Таким образом, основной костяк PDL-сегмента включают крупные участники рынка микрофинансирования, которые выделяются большим масштабом и положительным ростом динамики за счет использования достаточно гибких и эффективных моделей оценки и управления кредитными рисками в условиях регуляторных изменений.

1.2 Особенности деятельности российских микрофинансовых организаций в условиях нестабильности рынка

По результатам предыдущего параграфа выявлена закономерность, что основной костяк PDL-сегмента включают крупные участники рынка микрофинансирования, которые выделяются большим масштабом и положительным ростом динамики за счет использования достаточно гибких и эффективных моделей оценки и управления кредитными рисками в условиях регуляторных изменений. Следующим этапом исследования рассмотрим особенности деятельности российских микрофинансовых организаций в условиях нестабильности рынка.

Классический подход к микрофинансированию относится к концепции родоначальника данного вида деятельности и основателя первого в мире микрофинансового института Грамин Банка М. Юнуса, идея которого заключалась в том, чтобы при помощи микрофинансирования предоставлялся

доступ к капиталу для наиболее бедных слоев населения. М. Юнус осуществлял перераспределение полученных в традиционных финансово-кредитных институтах (банках) средств среди населения, которое нуждалось в кредитных ресурсах для развития своего дела (так называемое малое предпринимательство или самозанятость) и соответствующего генерирования доходов. В своих трудах М. Юнус определял «микрофинансирование» как «инструмент поддержки бедных слоев населения с целью улучшения их социально-экономического положения» [384].

Как указывает Д. Халм, «микрофинансирование» в классическом понимании представляет собой «один из способов раскрытия деловых возможностей бедных людей, занятых в самостоятельном ведении хозяйства с целью собственного жизнеобеспечения» [269]. При этом главной особенностью микрофинансирования на первоначальном этапе зарождения данного вида финансово-кредитных отношений являлся его групповой (общинный) характер. Займы в рамках микрофинансирования выдавались на базе коммуны, т. е. «на основе поручительства определенной группы ответственных людей из одной общины или населенного пункта» [245], что позволяло снижать риски несвоевременного возврата долга.

Таким образом, в классическом понимании микрофинансирование представляет собой деятельность по предоставлению денежных средств в виде займов краткосрочного характера, более доступных по сравнению с традиционным банковским кредитованием, для наиболее бедных слоев населения в первую очередь в регионах с недостаточно развитой финансово-кредитной инфраструктурой. По сути, в классическом понимании микрофинансовая деятельность является одним из видов социального предпринимательства, деятельность которого предполагает симбиоз реализации социальной миссии с одновременным получением коммерческого эффекта.

В соответствии с положениями Монтеррейского консенсуса, принятого в рамках «Международной конференции по финансированию развития», прошедшей 18–22 марта 2002 года под эгидой ООН, «микрофинансирование»

определяется как «предоставление кредитных средств субъектам малого и среднего предпринимательства, а также физическим лицам, в первую очередь, женщинам, имеющим ограниченный доступ к традиционным банковским услугам, и имеет важный макроэкономический и социальный эффект, который заключается в сглаживании социального неравенства в обществе, обеспечение занятости и повышения уровня жизни населения» [60]. Исходя из данного положения, четко прослеживаются две основные функции микрокредитования – социальная и экономическая.

В рамках реализации социальной функции в классическом понимании микрофинансирование выступает как инструмент повышения уровня жизни населения и сокращения бедности, расширения сектора малого предпринимательства и развития занятости населения. С точки зрения экономической функции микрофинансирование играет важную роль в развитии финансового рынка за счет расширения доступа к финансовым услугам для разных категорий заемщиков.

Реализуемая на настоящем этапе в России модель микрофинансирования далека от ее классического понимания, в большей степени микрофинансовая деятельность направлена на кредитование физических лиц, в меньшей степени – на предоставление финансовых ресурсов представителям малого бизнеса, которые не имеют доступ к традиционному банковскому кредитованию.

При этом цель микрофинансовой деятельности в России далека от первоначального ориентира данного института в мировой практике – повышение уровня жизни населения и сокращение бедности за счет расширения доступа к финансовым услугам для разных категорий заемщиков. В России микрофинансовая деятельность нацелена на извлечение коммерческой выгоды от получения процентов по микрозаймам, которые существенно превышают ставки по банковским кредитам.

Анализу современного состояния рынка микрофинансирования в России посвящены работы [5, 6, 12, 39]. Перспективы развития рынка микрофинансов и

основные тренды в условиях глобальных изменений отражены в работах [87, 90, 132, 141].

Главным назначением развития микрофинансовой деятельности в России на настоящем этапе является диверсификация финансового рынка, что подразумевает под собой: во-первых, расширение возможностей предоставления займов лицам, которые по различным причинам не имеют доступ к традиционному банковскому кредитованию; во-вторых, способствование сокращению теневого финансово-кредитного сектора.

В отношении первого утверждения реализуется разнонаправленное воздействие на экономику. С одной стороны, повышение доступности финансовых ресурсов для различных категорий граждан в рамках микрофинансирования ведет к возможному закрытию базовых потребностей населения и повышению его экономической активности. С другой стороны, упрощенная процедура получения заемных средств в рамках микрофинансирования в сравнении с банковским кредитованием ведет к дальнейшему увеличению закредитованности населения и соответствующему росту просроченной задолженности.

В данном контексте следует заключить, что вышеназванные постулаты имеют взаимообусловленную причинно-следственную связь. Низкий уровень дохода наименее защищенных в социальном плане граждан вынуждает их обращаться в микрофинансовые организации с целью получения денежных средств, необходимых в том числе для обеспечения базовых потребностей. Отсутствие необходимого объема денежных средств для погашения задолженности по имеющимся микрозаймам приводит к необходимости получения новых. Таким образом, происходит наращивание долговой нагрузки или закредитованности, что является причиной дальнейшего снижения располагаемых доходов населения.

На рисунке 1.9 представим динамику среднего показателя долговой нагрузки населения РФ за период с 2019 по 2024 год.



Рисунок 1.9 – Динамика среднего показателя долговой нагрузки населения РФ за 2019–2024 гг., %

Источник: составлено автором по данным [31, 83]

По представленной диаграмме видна существенная разница в оценках закредитованности населения по методике ЦБ РФ¹⁾ и методике Центра экономических исследований РИА Рейтинг²⁾, при этом последняя методика, по мнению автора, дает более достоверный результат.

Таким образом, наблюдается рост долговой нагрузки населения РФ, который в 2024 году достиг 57,1%, что свидетельствует о наличии реальных финансовых трудностей у значительной части населения, низкой финансовой грамотности, зачастую агрессивной рекламной кампании банков и МФО.

По данным Банка России, «около 60% всех потребительских займов входят в «цепочки займов», когда заемщик одновременно оформляет несколько займов или берет новый заем в день погашения предыдущего или с небольшим

¹⁾ По методике ЦБ РФ закредитованность определяется как отношение средней величины кредитной задолженности на одного заемщика к среднемесячному доходу.

²⁾ По методике РИА Рейтинг уровень закредитованности населения оценивается как соотношение среднедушевого долга экономически активного населения по банковским кредитам в соответствии со статистикой ЦБ РФ и годовой зарплаты за вычетом НДФЛ по данным Росстата.

интервалом» [77], более того, величина каждого последующего займа превышает сумму предыдущего, что приводит к наращиванию долга заемщика. Это определяет необходимость ужесточения регуляторных мер к микрофинансовым организациям с целью снижения закредитованности населения.

В рамках решения задачи сокращения теневого финансово-кредитного сектора реализуется планомерное развитие системы регулирования микрофинансовой деятельности наравне с банковской. В данном контексте следует отметить, что до принятия в 2010 году соответствующего законодательного акта в России довольно широко был развит сектор теневого микрофинансирования, которое подразумевало нелегальное предоставление денежных средств в долг под процент, который оценивается на необоснованно высоком уровне и существенно превышал законодательно разрешенную максимальную ставку. При этом взыскание долга по таким займам реализовывалось также с нарушениями законодательных норм, в том числе путем привлечения «черных коллекторов». Количество «нелегальных кредиторов» в 2010 году достигало 20 000. Это привело к существенному недоверию к МФО, при этом негативное отношение базировалось на следующих предпосылках:

во-первых, отождествление МФО с нелегальными или полунелегальными кредитными компаниями или частными кредиторами;

во-вторых, МФО предлагали деньги по беспредельно высоким ставкам, при этом в одностороннем порядке могли менять условия договора таким образом, что, взяв заем в размере 10 000 руб., должник мог обязан выплатить, например, 100 000 руб.;

в-третьих, МФО выдавали займы абсолютно всем обратившимся потенциальным заемщикам, включая безработных и недееспособных граждан;

в-четвертых, взыскание задолженности по микрозаймам реализовывалось, как правило, исключительно силовыми методами.

С целью полной легализации деятельности микрофинансовых организаций был принят Федеральный закон от 02.07.2010 №151-ФЗ «О микрофинансовой деятельности и микрофинансовых организациях» (далее по тексту – Закон 151-

ФЗ), в рамках которого определено, что «микрофинансирования» представляет собой «деятельность юридических лиц, имеющих статус микрофинансовой организации, а также иных юридических лиц, имеющих право на осуществление микрофинансовой деятельности по предоставлению микрозаймов» [65]. Под «микрофинансовой организации» в соответствии с Законом 151-ФЗ понимается «юридическое лицо, которое осуществляет микрофинансовую деятельность и сведения, о котором внесены в государственный реестр микрофинансовых организаций в порядке, предусмотренном настоящим Федеральным законом» [65]. К иным юридическим лицам законодатель относит «кредитные организации, кредитные кооперативы, ломбарды и другие юридические лица, ведущие микрофинансовую деятельность в соответствии с законодательством РФ» [65]. При этом заемщиком может выступать как физическое лицо, так и юридическое лицо и индивидуальный предприниматель.

В рамках настоящего диссертационного исследования объектом анализа выступают микрофинансовые организации, специализирующиеся на предоставлении займов физическим лицам. Исходя из чего, под микрофинансированием будем понимать часть финансового рынка, регулируемого Банком России и дающего доступ к финансам значительной части населения, для которых не доступны банковские услуги. Под микрофинансовой деятельностью будем понимать законодательно регулируемую активность микрофинансовых организаций, направленную на получение коммерческой выгоды и максимизации прибыли за счет получения процентов по микрозаймам, выдаваемым заемщикам – физическим лицам, не имеющим доступ к традиционному потребительскому кредитованию в банковских учреждениях.

Несмотря на принятие соответствующего законодательного акта в настоящее время теневой сектор микрофинансирования представлен следующими участниками:

- физические лица, дающие деньги в долг своему окружению;
- «бывшие» МФО, которые были исключены из реестра МФО, но при этом имеют обширную базу клиентов. Такие МФО нелегально работают либо под

другим статусом или наименованием, по состоянию на 2023 год доля исключенных из реестра Банка России, но продолжающих нелегально работать МФО составила почти 20% [29];

- «бывшие» (уволенные) сотрудники МФО или кредитных отделов банков, которые имели доступ к клиентской базе.

На рисунке 1.10 представлена динамика количества нелегальных кредиторов по данным ЦБ РФ. Из представленных данных видно существенное увеличение количества нелегальных кредиторов в 2022–2024 гг., в первую очередь в онлайн-сегменте теневой микрофинансовой деятельности, который по состоянию на конец 2024 года составил более 58% или 1 317 организаций и физических лиц.

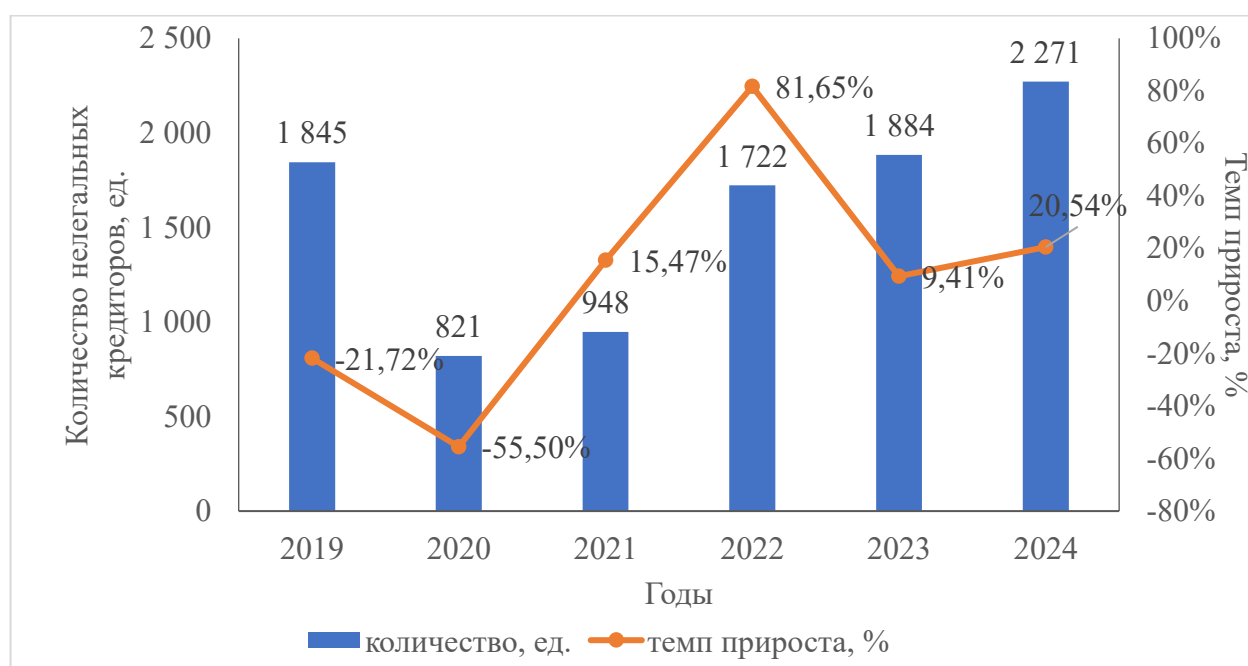


Рисунок 1.10 – Динамика количества нелегальных кредиторов по данным ЦБ РФ за 2019–2024 гг.

Источник: составлено автором по данным [82]

Рост количества кредиторов, имеющих незаконный характер, обусловлен двумя основными причинами – рост потребности населения в финансовых

средствах при одновременном снижении доступности легального кредитования на фоне ужесточения регуляторных требований.

Микрофинансовые организации (МФО), совокупность которых представлена на легальном рынке микрофинансирования в РФ, в соответствии с положениями Закона 151-ФЗ распределены на две основные группы – микрокредитные компании (МКК) и микрофинансовые компании (МФК).

В таблице 1.11 представим особенности разных видов МФО и их отличие от банков в рамках кредитования физических лиц.

Таблица 1.11 – Особенности МКК, МФК и банков в рамках кредитования физических лиц

Характеристика	МКК	МФК	Банк
Регулирование деятельности	Реестр ЦБ РФ	Реестр ЦБ РФ	Обязательное лицензирование
Минимальный размер собственных средств (капитала)	5 млн руб.	70 млн руб.	300 млн руб. (базовая лицензия) 1 млрд руб. (универсальная лицензия)
Привлечение средств физических лиц	Запрещено	Разрешено в сумме не менее 1,5 млн руб.	Разрешено
Максимальный размер займа	500 тыс. руб.	1 млн руб.	не ограничен
Валюта кредита (займа)	Российский рубль	Российский рубль	Российский рубль Иностранная валюта
Выдача кредитов (займов) под залог	Запрещено	Запрещено	Разрешено
Регулирование процентных ставок	Устанавливается ЦБ РФ	Устанавливается ЦБ РФ	Зависит от ключевой ставки ЦБ РФ
Процентные ставки	Высокие	Высокие	Ниже, соответствуют ключевой ставке
Срок кредитования	Короткий (от 7 дней до 12 мес.)	Короткий (от 7 дней до 12 мес.)	Среднесрочные (1–3 года) Долгосрочные (3–7 лет)
Диверсификация кредитного портфеля	Ограничен микрозаймами	Ограничен микрозаймами	Неограничен

Источник: составлено автором по данным [10]

По представленным характеристикам видно, что микрофинансовая деятельность имеет существенные ограничения по сравнению с банковской. В первую очередь, это определено по ограничению максимального размера валюты и срока выдачи займа, отсутствию возможности выдачи займов под залог.

Российский рынок микрофинансирования, направленный на обслуживание заемщиков – физических лиц, по видам кредитования делится на три основных сегмента (группы), представленные в таблице 1.12.

Краткосрочные микрозаймы (PDL), от англ. Pay Day Loans – «займы до зарплаты», наиболее распространенный и востребованный продукт среди заемщиков на российском рынке микрофинансирования. Основные параметры займа: срок до 30 дней, сумма до 15–30 тыс. руб. в зависимости от того первичный или повторный займ оформляет клиент.

Таблица 1.12 – Сравнительный анализ основных групп микрозаймов на российском рынке микрофинансирования

Характеристика	Краткосрочные микрозаймы (PDL)	Средне- и/или долгосрочные микрозаймы (IL)	Целевые микрозаймы (POS)
Обеспечение	Необеспеченные	Залоговые (обеспеченные) Беззалоговые (необеспеченные)	
Срок	До 30 дней	До 6 мес.	До 6–12 мес.
Сумма кредитования	До 30 тыс. руб.	Более 30 тыс. руб.	До 30 тыс. руб. (по стоимости товара)
Схема погашения	В конце срока. Единовременно	По графику	
Риски	Наиболее высокие	Высокие	
Процентные ставки	Наиболее высокие ПСК до 292%	Высокие ПСК менее 182,5%	
Маржинальность	Наиболее высокая	Высокая	

Источник: составлено автором по данным [10]

Средне- и/или долгосрочные микрозаймы (IL), от англ. Installment Loan – «кредит в рассрочку», займы от 30 тыс. со сроком погашения 10–18 недель с ежемесячным или еженедельным погашением.

Целевые микрозаймы (POS), от англ. Point of Sales – «точка продаж». Вариант целевого микрозайма, который можно получить в торговой точке или в онлайн-магазине в размере стоимости товара со сроком погашения до 12 мес.

Ввиду того, что микрофинансовая деятельность довольно обширное понятие, в рамках настоящего диссертационного исследования рассматриваются особенности деятельности, система оценки и управления кредитными рисками МФО, специализирующихся на выдаче краткосрочных PDL-займов. Необходимо отметить, что данный вид займов не только является приоритетным для МФО, но и по мнению специалистов Банка России социально значимым, при этом внутри данного сегмента сконцентрированы основные кредитные риски МФО. Таким образом, исследование моделей оценки и управления кредитными рисками МФО в сегменте PDL-займов является наиболее актуальным.

В рамках доклада о «Перспективных направлениях развития рынка микрофинансовых организаций на 2025–2027 годы» [77] с целью пересмотра регуляторных требований к PDL-сегменту микрофинансового рынка Банк России предложил категоризировать МФО следующим образом:

- компании предпринимательского финансирования;
- компании целевого финансирования;
- микрофинансовые компании.

Исключительное право на выдачу PDL-займов будут иметь микрофинансовые компании с ограничением необеспеченных займов предельным размером 100 тыс. руб.

Можно выделить следующие основные отличительные особенности микрофинансовой деятельности от банковской:

а) Стоимость услуг, т. е. процентные ставки по займам в МФО существенно превышают аналогичные банковские, как правило в 5–10 раз, величина спреда между наименьшей и наивысшей ставками имеет различие такого же порядка. Высокие процентные ставки компенсируют ограниченный размер и

краткосрочный характер выдаваемых займов с целью поддержания достаточного уровня маржинальности МФО.

б) Отсутствие у МФО возможности диверсификации кредитного портфеля, что обусловлено несколькими причинами. Во-первых, более высокий финансовый уровень заемщиков и более широкая клиентская база банков позволяет им расширять линейку кредитных продуктов под запросы клиентов, в чем ограничены МФО. Во-вторых, диверсификация кредитного портфеля определяет необходимость снижения уровня риска МФО при одновременном увеличении операционных затрат, реализация которых возможна в рамках двух направлений. В соответствии с первым направлением необходимо наращивать операционные затраты за счет увеличения расходов на ручную верификацию заемщиков и повышению оплаты труда сотрудников операционных департаментов. Второе направление предполагает увеличение расходов на автоматизацию процесса и разработку унифицированного инструмента верификации заемщиков с учетом различных видов рисков. Возможность использования обоих подходов ограничивается финансовыми ресурсами МФО.

в) Существенно более высокий уровень кредитных рисков микрофинансовой деятельности в сравнении с банковской, что обусловлено несколькими причинами. Во-первых, заемщиками МФО в большинстве случаев являются клиенты, которые не соответствуют требованиям традиционного банковского кредитования и изначально обладают более высокими рисками. Во-вторых, система оценки кредитоспособности заемщиков, которая позволяет определить «надежных» клиентов и конвертировать принятые заявки в выдачу кредитов в банках существенно развита в сравнении с аналогичными в МФО, что позволяет банкам классифицировать заемщиков по уровню риска и в соответствии с этим диверсифицировать кредитный портфель путем определения ставок и сроков по кредитам, а также выработке программ лояльности.

Также следует отметить, что экономика МФО в существенной степени отличается от экономики банков в силу следующих оснований:

- во-первых, МФО обладают менее значительными информационными, кадровыми и финансовыми ресурсами в сравнении с банками, что приводит к недостаточности данных для принятия решения о выдаче займа;
- во-вторых, меньшая сумма займа и более короткий их срок в МФО (не более 1 месяца) приводит к существенно более высоким затратам на принятие одного решения о выдаче займа, что более критично для экономики МФО;
- в-третьих, в МФО первый заем, как правило, является убыточным из-за высокой стоимости привлечения клиента. В экономике микрофинансов прибыль формируется за счет повторных займов клиентов.

Помимо вышеперечисленного различия проявляются в системе регулирования банковской и микрофинансовой деятельности.

Являясь частью финансового рынка институт микрофинансирования в настоящее время регулируется не только Банком России, но и саморегулируемыми организациями, что отличает систему регулирования микрофинансовой деятельности от банковской. В настоящее время микрофинансовая деятельность подпадает под регулирование в соответствии с положениями следующих нормативно-правовых актов:

а) Федеральный закон от 02.07.2010 №151-ФЗ «О микрофинансовой деятельности и микрофинансовых организациях» [65] (далее по тексту – Закон 151-ФЗ), в котором определены виды МФО, правовые основы, условия и порядок реализации, а также система регулирования микрофинансовой деятельности, включая обязательное лицензирование и включение в реестр МФО, обязательность предоставления открытой информации по условиям займов, предельные размеры займов и другие нормы.

б) Федеральный закон от 21.12.2013 № 353-ФЗ «О потребительском кредите (займе)» [68] (далее по тексту – Закон 353-ФЗ), в соответствии с которым МФО, предоставляющая займы физическим лицам определяется как один из видов кредитных организаций, которые имеют право вести «профессиональную деятельность по предоставлению потребительских займов» [68], в связи с этим на

микрозаймы, выдаваемые МФО, распространяются все нормы, принятые в отношении потребительских кредитов.

в) Федеральный закон от 30.12.2004 № 218-ФЗ «О кредитных историях» [64] (далее по тексту – Закон 218-ФЗ), в рамках которого зафиксировано, что МФО выступают одним из источников формирования кредитной истории заемщика и обязаны подавать сведения по всем выдаваемым займам в бюро кредитных историй (далее по тексту – БКИ).

г) Федеральный закон от 13.07.2015 № 223-ФЗ «О саморегулируемых организациях в сфере финансового рынка» [70] (далее по тексту – Закон 223-ФЗ), который определяет порядок функционирования саморегулируемых организациях (далее по тексту – СРО) в микрофинансовом сегменте, а также порядок и условия вступления и членства МФО в таких СРО.

Регулирование деятельности и надзор за соблюдением требований к деятельности МФО реализуется Банком России, что определено в ст. 14 Закона 151-ФЗ.

Необходимо отметить, что начиная с 2015 года реализуется последовательное ужесточение регуляторных мер в отношении МФО, направленное на снижение закредитованности заемщиков – физических лиц, основными мерами достижения данной цели являются следующие:

- во-первых, определение максимальной величины полной стоимости кредита;
- во-вторых, установление максимальной процентной ставки по микрозаймам;
- в-третьих, установление предельного размера переплаты от тела микрозайма;
- в-четвертых, установление минимального размера резерва на возможные потери по микрозаймам.

Обязательный расчет показателя долговой нагрузки заемщика (ПДН) был введен в 2019 году для микрозаймов свыше 10 тыс. руб., а с 1 января 2024 г. – по любым микрозаймам (вне зависимости от суммы). При этом при превышении

50% порога ПДН, МФО обязано уведомить заемщика в письменной форме о возможных кредитных рисках при получении очередного займа.

С 1 января 2023 года Банк России ввел еще одно прямое ограничение для выдачи займов МФО – макропруденциальные лимиты, направлены на снижение риска за кредитованности населения и пересматриваются ЦБ ежеквартально. Такой вид ограничений давно успешно применяется в практике банковского регулирования.

В таблице 1.13 представлено изменение регуляторных требований Банка России к микрозаймам, выдаваемым МФО.

Таблица 1.13 – Изменение регуляторных требований Банка России к микрозаймам МФО

Требование	2015–2018	2019–2022	2023–2024
Максимальная процентная ставка в день	нет	1,5% – с 28.01.2019 1% – с 01.07.2019	0,8% – с 01.07.2023
Предельный размер переплаты (раз от тела займа)	4 – с 29.03.2016 3 – с 01.01.2017	2,5 – с 28.01.2019) 2 – с 01.07.2019) 1,5 – с 01.01.2020)	1,3 – с 01.07.2023
Максимальная величина полной стоимости кредита	нет	548% – с 28.01.2019 365% – с 01.07.2019	292% – с 01.07.2023
Макропруденциальные лимиты			Введены с 01.01.2023

Источник: составлено автором по данным [10]

Снижение максимально допустимого уровня процентной ставки в день с 1 июля 2023 года до 0,8% обусловлено необходимостью снижения долговой нагрузки заемщиков МФО. Данное регуляторное требование позволило снизить максимальную величину ПСК по микрозаймам до 292%, а предельный размер переплаты по микрозаймам снижен до 1,3 от тела займа.

Изменение макропруденциальных лимитов для МФО, которые в соответствии с требованиями Банка России устанавливаются на ежеквартальной

основе и учитывают динамику закредитованности населения, представлено в таблице 1.14.

Таблица 1.14 – Изменение макропруденциальных лимитов для МФО

Период	Макропруденциальные лимиты при показателе долговой нагрузки, %	
	50–80%	более 80%
с 01.01.2023		35
с 01.07.2023		30
IV квартал 2023	30	15
I квартал 2024	25	15
II квартал 2024	25	15
III квартал 2024	20	10
IV квартал 2024	15	3

Источник: составлено автором по данным [11]

Банк России существенно снизил макропруденциальные лимиты в четвертом квартале 2024 года, определив их на уровне 15% для заемщиков МФО с ПДН 50–80% и на уровне 3% для заемщиков с ПДН более 80% на фоне роста закредитованности населения РФ. С первого квартала 2025 года данный показатель может быть снижен до 10% и 0% соответственно.

В январе 2024 года были приняты поправки к Закону 353-ФЗ, зафиксировавшие обязательство МФО включать дополнительные услуги в ПСК с целью повышения прозрачности микрофинансовой деятельности. Данное регуляторное требование привело к снижению процентных ставок по микрозаймам для того, чтобы не превысить порог максимального лимита по ПСК в размере 292% годовых, что негативно сказывается на процентных доходах и соответственно прибыли МФО.

Изменение регуляторных требований на рынке микрокредитования определяется ключевой целью Банка России по снижению закредитованности заемщиков – физических лиц и включает их постепенное ужесточение касательно ограничения максимальных лимитов по процентным ставкам, по размеру переплаты, совокупной величины ПСК, а также максимального удельного веса

необеспеченных микрозаймов. Главной целью со стороны Банка России в данном случае выступает снижение долговой нагрузки заемщиков физических лиц.

В рамках регулирования PDL-сегмента микрофинансового рынка Банк России предлагает в течение 2025–2027 гг. реализовать следующие ограничительные меры (см. таблицу 1.15).

Таблица 1.15 – Ограничительные меры для регулирования PDL-сегмента микрофинансового рынка на 2025–2027 гг.

Параметр	Ограничение	Сроки введения
Предельный размер переплаты (% от тела займа)	100%	С 1 июля 2025 года
«Новация договоров»	Полный запрет – невозможность заключения нового договора займа на погашение действующего	С 1 июля 2025 года
Количество одновременно действующих PDL-займов	2 займа с ПСК более 200%	С 1 января 2026 года
	1 займ с ПСК более 100%	С 1 января 2027 года
Период охлаждения между оформлением PDL-займов	Не менее 3 календарных дней	С 1 января 2027 года

Источник: составлено автором по данным [77]

Ужесточение регуляторных требований в отношении PDL-сегмента микрофинансового рынка оказывает двойственное влияние на деятельность МФО. Среди отрицательных сторон мер, вводимых регулятором, можно выделить следующие:

- во-первых, сокращение возможности выдачи займов наиболее обширной целевой аудитории МФО – заемщикам с высокой долговой нагрузкой, которые не могут кредитоваться в банковском секторе. Это ведет к сокращению клиентской базы МФО и вынужденному обращению заемщиков к нелегальным кредиторам;
- во-вторых, необходимость пересмотра бизнес-моделей МФО и переход в другие сегменты микрофинансового рынка, прежде всего в сегменты «дешевых» займов с более низкими дневными процентными ставками и ПСК, а также

расширением сроков предоставления займов. Это является главной причиной снижения маржинальности и рентабельности микрофинансового бизнеса;

- в-третьих, сокращение количества МФО в сегменте PDL-займов ввиду невозможности адаптации многих организаций, прежде всего небольших, к новым нормам регулирования. Это, в свою очередь, является явным толчком к развитию теневого сектора микрокредитования.

Если рассматривать вводимые ограничения со стороны положительного воздействия на МФО, то произойдет увеличение количества платежеспособных и качественных заемщиков за счет отсека клиентов с высокой долговой нагрузкой, что соответственно приведет к повышению качества кредитного портфеля и снижению кредитных рисков МФО.

Помимо регуляторных требований Банка России, некоторые меры содержатся в Перечне поручений Президента РФ:

- предоставление МФО доступа к системе Госуслуг с целью «оперативного получения актуальной официальной информации, подтверждающей доходы физического лица (в течение не более одной минуты с момента направления соответствующего запроса)» [75];

- обязанность МФО к использованию «актуальной официальной информации, подтверждающей доходы физического лица, при принятии решения о предоставлении ему потребительского займа» [75].

Данные меры приведут, с одной стороны, к более быстрому принятию решений по выдаче займов, сокращению количества «плохих» заемщиков МФО и соответствующему снижению кредитных рисков МФО. С другой стороны, снизится количество заемщиков в целом, в виду того, что более 80% клиентов МФО не имеют официально подтвержденного дохода, указываемого в анкете на получение займа.

Таким образом, проведенное исследование показало, что ужесточение требований к МФО со стороны регулятора, а также существенные ограничения в рамках системы управления кредитными рисками по сравнению с банковскими организациями ведет к необходимости не только пересмотра бизнес-моделей, но и

выработки новых, адекватных сложившейся ситуации подходов к управлению кредитными рисками, разработке стандартов и инструментов качественной оценки кредитных рисков по различным категориям заемщиков.

В связи с этим возникают методологические проблемы, решение которых позволит выстроить систему управления кредитными рисками МФО, позволяющую организациям успешно адаптироваться к волатильности экономических условий и развития микрофинансового рынка без ущерба маржинальности бизнеса. Перед формулированием данных методологических проблем рассмотрим основные тенденции и закономерности, сложившиеся в последние годы на рынке микрофинансирования в России.

1.3 Методологические подходы к оценке и управлению кредитными рисками микрофинансовых организаций

Рассмотрев особенности деятельности МФО на рынке России и основные сложившиеся тренды, перейдем к рассмотрению отличий системы управления рисками в микрофинансировании в сравнение с банковским кредитованием.

В рамках системы управления рисками реализуется надзор за деятельностью и предоставление интересов МФО, включенных в реестр саморегулируемых организаций (СРО), в Банке России. Одним из инструментов саморегулирования МФО является Базовый стандарт по управлению рисками, разработанный СРО и утвержденный Банком России, в рамках которого зафиксированы следующие положения, относящиеся к функционированию системы риск-менеджмента в МФО: цели, задачи, принципы, алгоритм, функционал и распределение полномочий в рамках системы управления рисками. Стандарт определяет несколько основных рисков МФО, включая «кредитный риск», который представляет собой «риск возникновения у МФО убытков вследствие неисполнения, несвоевременного либо неполного исполнения заемщиком или иным контрагентом договорных обязательств перед МФО» [7].

Кредитный риск также определяют как риск дефолта, который связан с «ухудшением качества кредитного портфеля в результате полученных убытков по кредитам вследствие невыполнения заемщиком условий кредитного договора» [313].

В Указании Банка России от 20.01.2020 № 5391-У «О порядке формирования микрофинансовыми организациями резервов на возможные потери по займам» регулятор определил следующие факторы возникновения кредитного риска МФО по заемщикам – физическим лицам (в тексте Указания – риски возникновения убытков): «наличие просроченной задолженности по требованиям по заключенному договору займа, использование заемщиком займа не по целевому назначению» [67]. При этом данное Указание не содержит положений относительно регламентации системы и процедур управления кредитными рисками МФО.

О.И.Лаврушин определяет систему управления кредитными рисками как «комплекс управленческих решений, направленных на идентификацию рисков, их количественную и качественную оценку, а также планирование защитных мероприятий, заключающихся в модификации организации процесса кредитования, с целью предупреждения либо минимизации потерь вследствие неисполнения или неполного исполнения заемщиками обязательств по кредитным договорам» [49]. Несмотря на то, что с позиции данного автора субъектом управления кредитными рисками выступает банковское учреждение, такая система в полной мере может быть имплицирована на деятельность МФО, которая заключается в кредитовании заемщиков.

Базовым документом, регламентирующим систему управления банковскими рисками, выступают «Рекомендации Базельского комитета по банковскому надзору «Принципы совершенствования корпоративного управления», в рамках которого не выделяются конкретные виды, но положения данного регламента относятся ко всем видам банковских рисков, включая кредитный. В рамках данного регламента определен алгоритм управления рисками банковской деятельности, включая следующие этапы: «идентификация основных рисков;

оценка рисков и степень подверженности им; отслеживание объемов рисков и расчет капитала на его покрытие; мониторинг и оценка решений по принятию определенных рисков; выработка мер, направленных на смягчение рисков; оценка соответствия принимаемых решений целевым показателям устойчивости к риску» [69].

Кроме того, система управления банковскими рисками базируется на положениях Указания Банка России от 15.04.2015 № 3624-У «О требованиях к системе управления рисками и капиталом кредитной организации и банковской группы» [71], в котором кредитному риску посвящена глава 2, включающая следующие обязательные процедуры в системе управления кредитными рисками банка: «порядок предоставления ссуд и принятия решений об их выдаче; методики определения и порядок установления лимита риска на одного заемщика; требования, предъявляемые к обеспечению исполнения обязательств заемщиков и методологию его оценки» [71].

Также в рамках регламентации банковской деятельности по предоставлению кредитов действует Положение Банка России от 28.06.2017 № 590-П «О порядке формирования кредитными организациями резервов на возможные потери по ссудам, ссудной и приравненной к ней задолженности» [77], в рамках которого определены «Общие требования по оценке кредитных рисков по ссудам», которые определяют обязательность для банков утверждать и использовать в системе управления кредитными рисками следующие инструменты: «методы и процедуры оценки финансового положения заемщика; методы получения информации о заемщике и оценки представляемой заемщиком информации по критериям полноты, актуальности, достоверности; порядок оценки кредитного риска по портфелю однородных ссуд» [77] и другие. В отношении рисков по действующим кредитам регулятор определил следующие обязательные процедуры: «оценка финансового положения заемщика, оценка качества обслуживания долга по ссуде» [77] и другие.

Для МФО в настоящее время подобных нормативно-правовых актов не разработано. В положениях «Базового стандарта по управлению рисками

микрофинансовых организаций» не определены конкретные процедуры в рамках системы управления кредитными рисками МФО, а алгоритм управления рисками включает следующие процессы: «идентификация рисков, связанных с осуществляемой деятельностью; организация управления рисками, включая установление МФО допустимых уровней потерь (лимитов) по выявленным рискам; измерение и оценка рисков; контроль за соответствием рисков установленным МФО ограничениям рисков и устранением выявленных нарушений ограничений рисков» [7].

В данном контексте следует отметить, что формально система управления кредитными рисками в МФО не отличается от банковской и включает следующие основные этапы (см. таблицу 1.16).

Таким образом, система управления кредитными рисками как банков, так и МФО представляет собой совокупность процедур и инструментов, направленных на снижение вероятности дефолта заемщика и соответственно получение убытков от деятельности, связанной с кредитованием.

Несмотря на идентичность формально установленных процедур в рамках системы управления кредитными рисками банков и МФО, в реальных условиях выявляются достаточно существенные отличия, связанные со следующими причинами:

- во-первых, отсутствие четкой регламентации процедур в рамках системы управления кредитными рисками МФО со стороны Банка России, как это сделано для банковской деятельности. Действующие нормативно-правовые акты в отношении управления кредитными рисками МФО по большому счету являются рекомендательными, а не регламентирующими;

- во-вторых, ограниченность финансово-ресурсной и информационно-ресурсной базы препятствует МФО в полной мере соблюдать процессы управления кредитными рисками на всех этапах данной системы. Ограниченность собственных информационных ресурсов и недостаточный объем средств для получения информации из сторонних источников определяет пробелы в актуальных и достоверных данных в отношении заемщиков. При этом короткий

срок займов не позволяет использовать полный объем необходимой информации по заемщикам, даже при ее наличии, в системе принятия решений;

Таблица 1.16 – Этапы управления кредитными рисками в МФО и банках

Наименование этапа	Описание этапа	Содержание этапа
Идентификация кредитного риска	Определение вероятности наступления дефолта заемщика, который может нанести ущерб МФО или банку	Выявление негативной информации о заемщике в связи с исполнением заемных обязательств, включая: - информацию о плохом финансовом положении заемщика либо об его ухудшении; - информацию о наличии просроченной или реструктурированной задолженности в других финансово-кредитных организациях и др.
Анализ и оценка кредитного риска	Определение прогнозной величины ущерба от взаимодействия с заемщиком	Проведение ретроспективного анализа кредитной надежности заемщика на основе: - анкетирования; - анализа данных бюро кредитных историй; - оценки кредитоспособности с использованием скоринговых моделей.
Мониторинг кредитных рисков	Контроль динамики показателей кредитного риска	Проведение анализа изменения: - показателей кредитоспособности заемщика; - возникновения и динамики просроченной задолженности по кредитам/займам.
Разработка стратегии управления кредитными рисками	Выработка стратегических направлений	Направления снижения, смягчения или избежания кредитных рисков могут включать: - установление лимитов на выдачу займов в зависимости от рискованного профиля заемщика; - диверсификация кредитного портфеля по срокам и условиям для каждой рискованной группы заемщиков и другие

Источник: составлено автором

- в-третьих, ограниченность финансовых ресурсов и квалифицированных специалистов в области риск-менеджмента, а также отсутствие релевантных моделей оценки кредитных рисков и специализированных программных разработок для автоматизации данного процесса в МФО, в отличие от банковской сферы, не позволяет реализовывать системный подход к процедурам управления и оценки кредитных рисков МФО.

Выявленные особенности подтверждают актуальность проводимого исследования, направленного на совершенствование методологии оценки и управления кредитными рисками МФО.

Одним из наиболее важных этапов управления кредитными рисками МФО выступает их оценка, для этого используются различные метрики. Как правило, измерение эффективности кредитной деятельности осуществляется по модели «риск-доходность». Таким образом, для оценки кредитных рисков МФО используют:

а) Показатель доходности – ROA (англ. Return Of Assets, прибыльность активов). В качестве активов в данном случае принимаются профильные активы МФО, т. е. совокупный объем предоставленных займов (основной долг по займам) за определенный период. Доход определяется как совокупный объем выплат от заемщиков по процентам за использование займа.

б) Показатели кредитных рисков включают в основном в МФО две метрики: NPL и FPD.

NPL (англ. Non-Performing Loan, проблемный кредит), данный показатель представляет собой долю заемщиков, по которым зафиксирована просроченная задолженность по займам. Как указывает [288] кредит становится проблемным (NPL), если выплата полного объема основного долга и процентов по нему не только не производится в установленный срок, но и не ожидается в будущем.

В отношении банковских организаций кредит определяется как проблемный (NPL), если по нему не осуществляется оплата процентов и основной суммы задолженности в течение как минимум 90 дней [369], то есть для оценки кредитных рисков в банках используется метрика NPL90. Вследствие коротких

сроков предоставления займов (для PDL-займов срок составляет не более 30 дней), в МФО в большинстве случаев используется показатель NPL30, но также могут использоваться и показатели со сроками просрочки менее 30 дней, например NPL1, NPL10 и т.д.

Существуют разные подходы к расчету метрики NPL:

а) по количеству клиентов, допустивших просроченную задолженность по займам. Расчет производится по формуле:

$$NPL = \frac{DCN_t}{TCN_t} \cdot 100\%, \quad (1.1)$$

где DCN_t (Default Client Number) – количество клиентов, просрочивших выплаты по займам, за определенный период;

TCN_t (Total Client Number) – общее количество клиентов, получивших займы, за определенный период.

б) по количеству займов, по которым зафиксирована просроченная задолженность. Расчет производится по формуле:

$$NPL = \frac{DL_t}{TL_t} \cdot 100\%, \quad (1.2)$$

где DL_t (Defaulted Loan) – количество займов с просрочкой выплат за определенный период;

TL_t (Total Loan) – общее количество предоставленных займов за определенный период.

Кроме того, существуют подходы, в рамках которых производится расчета «чистого» NPL, не учитывающего займы, по которым была произведена выплата процентов и основного долга после просрочки. Также часто считают NPL не по количеству займов, а по их сумме.

Большая часть эмпирических исследований направлена на оценку кредитных рисков банковских организаций. Так, в исследовании [288] проведено моделирование взаимосвязи показателя NPL с различными показателями эффективности банковской деятельности, включая ROA, которое показало отрицательную корреляционную зависимость между показателями NPL и ROA, т.

е. снижение прибыльности банков находится в прямой связи с ростом объемов проблемных кредитов.

В отношении МФО было проведено исследование [333], которое показало отрицательную корреляционную связь между показателями NPL и ROA, т. е. наращивание просроченных займов ведет к снижению маржинальности МФО. Автор пришел к выводу, что МФО с более эффективной системой оценки кредитоспособности заемщиков имеет меньшую долю проблемных займов и соответственно более высокую маржинальность. Такие же результаты были получены в исследовании [290], автор которого заключил, что существенные потери от проблемных займов не позволяют МФО получить нормальный уровень маржинальности бизнеса.

Второй показатель оценки кредитных рисков в МФО – FPD (англ. First Payment Default, дефолт первого платежа), данный показатель представляет собой выданные займы, по которым отсутствует поступление первого платежа или пролонгации займа при наступлении срока оплаты. Необходимо отметить, что данный показатель имеет специфику в отношении кредитных рисков МФО, т. к. «отсутствие платежа по займу при его продлении не является негативным событием риска» [103]. При этом сегмент клиентов, продлевающих займы, является максимально доходным для МФО.

В исследовании [293] было определено, что увеличение показателя FPD на 9–15% ведет к увеличению показателя доходности (ROA) почти на 50%, что компенсирует рост уровня кредитных потерь.

Показатель FPD, так же, как и NPL, рассчитывается по заемщикам с разным сроком просрочки выплат по займам – 1, 10, 30 дней и т. д. с целью оценки кредитных потерь и корреляции с показателями доходности портфеля займов МФО по каждой группе заемщиков.

Следует отметить, что выбор метрик для оценки кредитного риска зависит от специфики бизнес-модели и реализуемой в ней системы риск-менеджмента конкретной МФО [127].

Исследование тенденций и закономерностей развития рынка микрофинансирования в России, а также анализ особенностей деятельности и управления кредитными рисками российских микрофинансовых компаний позволил выявить ряд проблемных моментов, основными из которых являются волатильность экономических условий и развития микрофинансового рынка, ужесточение регуляторных требований к МФО, а также существенные ограничения финансового, информационного и временного характера в рамках системы управления кредитными рисками.

Решение данных проблем лежит в плоскости задач методологического характера, при этом наиболее важными для МФО являются следующие:

- выбор исходной информации для оценки кредитных рисков;
- выбор критериев ценности и качества исходной информации для оценки кредитных рисков;
- выбор концепции для оценки кредитных рисков: статистическое моделирование и/или методы и алгоритмы машинного обучения;
- определение алгоритма методики построения модели оценки кредитных рисков;
- формирование лимитной политики для управления кредитными рисками;
- учет множества бизнес-метрик через общую модель юнит-экономики при управлении кредитными рисками;
- управление кредитными рисками с учетом изменения параметров моделей оценки кредитных рисков во времени;
- оценка кредитных рисков в условиях отсутствия исторических данных для построения модели их оценки.

Рассмотрим каждую из этих методологических задач подробно.

Первая методологическая проблема заключается в *выборе исходной информации для оценки кредитных рисков*. Данная проблема возникает в условиях ограниченных ресурсов МФО, а также в необходимости системного подхода к их выбору в условиях тренда расширения источников данных.

Для решения данной задачи приведем обзор научных работ, в которых исследовалась данная методологическая проблема.

В научной литературе отмечаются две основные причины необходимости правильного выбора источников данных для задач оценки кредитных рисков в МФО:

- во-первых, недостаток или отсутствие информации по группам заемщиков [199];
- во-вторых, необходимость повышения точности моделей, самостоятельно разработанных в кредитной организации [231].

Несмотря на актуальность данного вопроса как в отечественной, так и в зарубежной научной литературе недостаточно внимания уделяется классификации источников данных, используемых для оценки кредитных рисков. При этом в большинстве научных источников проблематика выбора источников данных в рамках оценки кредитных рисков применима в основном для банковской сферы, практически отсутствуют научные исследования посвященные проблеме классификации источников данных в МФО, которые имеют свою специфику.

В большинстве случаев МФО в рамках системы оценки кредитных рисков используют три основных источника данных:

- отчеты бюро кредитных историй;
- анкетные данные заемщиков, которые подлежат самостоятельному заполнению при подаче заявке на заем (в первую очередь на сайте МФО);
- внутренняя информация о заемщиках и его кредитном поведении в случае повторных займов.

При этом еще одним важным вариантом источников информации о заемщиках выступают альтернативные источники, которые могут включать различные социально-демографические данные, данные от операторов сотовой связи, информация о платежах и другие.

В исследовании [214] приводится подробное описание работы МФО с альтернативными данными, при этом отсутствуют критерии для классификации

таких источников, а сама классификация представлена в сжатом виде. Автор выделяет ключевую проблему системы управления кредитными рисками МФО – нехватка полезных информационных ресурсов по различным источникам данных.

В работе [322] приводится обзор альтернативных источников данных для оценки кредитных рисков в кредитных организациях США. Среди основных источников альтернативных данных автор выделяет платежные транзакции по коммунальным услугам, включая платежи за потребление электроэнергии, абонентскую плату за телефон и кабельное телевидение, оплату других коммунальных услуг. Также важными источниками информации для оценки кредитоспособности заемщиков выступают данные об аренде недвижимости, данные местных органов власти по конкретному заемщику, включая начисление и оплату штрафов за нарушения правил дорожного движения, список совершенных правонарушений административного характера и сведения о судебных разбирательствах, иммиграционный и социальный статус, данные по наличию движимого и недвижимого имущества, а также других активов, сведения о банкротстве и другие данные. Кроме того, по мнению автора статьи, важными источниками данных в оценке кредитных рисков выступают данные операторов сотовой связи, движение контента в аккаунтах социальных сетей, различная информация психометрического характера.

Работе [144] посвящено исследование одного из видов систем оценки банковских кредитных рисков – социального скоринга, основой которого выступает информация из аккаунтов социальных сетей. Авторы выявляют наиболее популярные в рамках оценки поведения заемщиков социальные сети и приводят некоторые критерии отбора такой информации.

Значительный потенциал в прогнозировании кредитного риска заемщиков как в банковских организациях, так и в МФО, имеют транзакционные данные, которым также уделяется внимание в научной литературе. Так, например, в работе [42] проведено исследование возможности использования для оценки кредитных рисков исключительно данных по транзакциям потенциального заемщика. Однако применение такого подхода дает менее точные результаты в

сравнении с использованием данных БКИ. Таким образом, для повышения точности моделей оценки кредитных рисков необходимо использовать комплексный подход, включающий применение как традиционных (данные БКИ и анкетные данные), так и альтернативных данных.

Существенные возможности расширения наборов источников данных в оценке кредитных рисков появились на фоне развития технологий обработки больших данных. Например, в исследовании [374] подтверждена существенная роль в оценке кредитных рисков использования методов машинного обучения на основе больших данных, которые собираются, в первую очередь, из альтернативных источников, включая аккаунты в социальных сетях, демографическую информацию о заемщиках, данные операторов связи и др., и позволяют расширить прогностические способности моделей оценки кредитных рисков.

Работа [339] посвящена разработке метода скрининга бинарных информационных значений (BIV) и модели средневзвешенного значения на основе логистической регрессии, случайного леса и градиентного бустинга, которые работают на основе данных из нескольких источников, включая кредитную историю, транзакции по кредитной карте, анкетные данные из заявки на получение кредита и другие. В результате авторам удалось повысить качество классификации заемщиков при одновременном упрощении методики принятия решений о выдаче кредитов в банковском учреждении.

В статье [362] предложен метод оценки кредитных рисков на основе пространственного и временного анализа при помощи объединения данных из трех альтернативных источников, включая данные мобильного телефона общедоступные спутниковые и геолокационные данные. На основе эмпирического исследования было доказано повышение точности моделей оценки кредитных рисков при использовании помимо данных БКИ информации из различных альтернативных источников. При этом авторы доказали взаимосвязь между снижением количества ошибок неправильной классификации заемщиков и

сокращением расходов финансовых учреждений за счет интеграции источников данных в оценке кредитных рисков.

Некоторые авторы [246, 378] доказывают актуальность сбора данных о заемщиках ввиду появления конкурирующих с кредитными организациями платформ, которые используют инновационные разработки в кредитовании физических лиц.

В работе [267] на примере крупной китайской финансовой компании доказано преимущество применения методов машинного обучения в моделях оценки кредитных рисков в нестабильной экономической среде на основе альтернативных данных. Однако в работе отсутствует системная классификация источников потенциальных альтернативных данных для оценки кредитных рисков и рекомендаций по их отбору.

Оценка кредитных рисков в исследовании [370] осуществляется на основе комбинации наборов альтернативных данных, включая записи о телефонных разговорах, данные мобильного телефона, а также информации о кредитных и дебетовых счетах клиентов, которые можно получить посредством взаимодействия кредитно-финансовых учреждений с телекоммуникационными компаниями и кредитными бюро. Метод апробирован на кредитных картах, при этом отсутствует его экспликация на другие виды кредитных продуктов.

В работе [328] определено, что сотрудничество с телекоммуникационными компаниями позволяет получить информацию о количестве звонков и СМС по конкретному мобильному телефону, платежах за услуги мобильной связи, задолженностях по телефонным счетам. Кроме того, телекоммуникационные компании на основе внутренней информации о владельце номера предоставляют готовые скоринговые баллы.

Важную роль в оценке кредитных рисков играет наличие дополнительной информации о заемщике, связанной с его деятельностью и активностью (например, социально-демографические данные), что доказано в исследовании [176] на примере одной марокканской МФО.

В целом проведенное исследование показало, что в зарубежной и в отечественной научной литературе по вопросам оценки кредитных рисков широко распространено применение данных из различных источников данных, альтернативных данным кредитных историй. Апробация и эмпирические исследования на основе реальных данных доказывают эффективность использования такого подхода.

Однако большинство исследований проведено на банковских организациях, а проблема использования альтернативных источников данных при разработке систем принятия решений (СПР) в МФО затрагивает небольшое число исследований.

При этом на основе требований Банка России согласно базовому стандарту к использованию пяти различных источников данных для идентификации заемщика и оценки кредитных рисков при одновременной ограниченности финансовых и временных ресурсов МФО, можно сделать вывод о существенной актуальности решения задачи систематизации информационных ресурсов для риск-аналитики посредством классификации данных, имеющихся на российском информационном рынке на основе соответствующих критериев с учетом специфики функционирования МФО.

Вторая методологическая задача основана на *определении ценности данных для оценки кредитных рисков МФО*.

Как было определено ранее, одна из особенностей микрофинансового бизнеса заключается в необходимости проведения оперативного определения платежеспособности клиентов и выявления мошеннических заявок при принятии решения о выдаче займов. При этом ограниченность информационных и финансовых ресурсов МФО определяет экономическую нецелесообразность получения данных для оценки кредитного риска из множества источников. В данном контексте, в особенности в сегменте онлайн-кредитования, актуальность принимает вопрос определения и выбора наиболее ценной информации о

клиентах касательно их кредитного, транзакционного, социального поведения и других параметров.

Необходимо отметить, что проблема оценки ценности данных в рамках управления кредитными рисками рассматривалась в различных научных исследованиях, которые в основном посвящены банковскому сектору. Например, работа [55] посвящена рассмотрению методики обработки данных, основанных на кредитной истории и психологических тестах заемщиков, полученных за счет собственных ресурсов банков. Авторы проводят вероятностное моделирование и сравнение возможной прибыли банка с альтернативными издержками на получение и обработку соответствующих данных, анализируя их вероятностные законы распределения [9].

В работе [1] представлено подробное описание процесса выбора наиболее информативных параметров для проведения оценки кредитных рисков в рамках модели пошаговой логистической регрессии на основе весов отобранных переменных.

Расширение количества источников и объемов данных для оценки кредитных рисков требует совершенствования методов анализа таких данных, а также построения более продвинутых моделей на их основе. В настоящее время в зарубежной научной литературе рассматривается проблема выбора переменных, необходимых для прогнозирования поведения заемщика, на основе методов машинного обучения. Например, в исследовании [185] приведен подробный обзор методов выбора предикторов в рамках оценки кредитных рисков и предложен инновационный трехэтапный подход на основе квадратичного программирования с демонстрацией эффективности предложенных методов на данных реальных финансово-кредитных организаций.

Методы обработки больших данных используются как с целью прогнозирования поведения заемщика [156], так и с целью выбора наиболее содержательных параметров для оценки кредитных рисков [348]. В статье [1] реализована разработка двухэтапного гибридного алгоритма выбора наиболее

ценных параметров для оценки кредитного риска заемщика на основе логистической регрессии и метода опорных векторов.

Практика оценки кредитных рисков также базируется на выборе переменных с учетом влияния их использования на прибыль финансово-кредитной организации. Например, в публикации [272] разработана модель отбора данных из различных источников на базе метода опорных векторов, исходя из затраченных банком ресурсов на их получение, включая внешние затраты на приобретение данных, а также внутренние затраты на формирование и управление такими данными. В работе [148] реализуется усовершенствованный многоцелевой подход к выбору переменных для оценки кредитных рисков с учетом ожидаемой прибыли банковской организации.

Несмотря на существенную актуальность проблемы валидности данных [122] в том числе и в рамках системы управления кредитными рисками МФО, она недостаточно проработана как в зарубежной, так и отечественной научной литературе. При этом главная проблема заключается в отсутствии в имеющихся научных исследованиях описания стандартной общепринятой методики определения ценности данных на основе как теоретических предпосылок, так и в эмпирическом опыте их применения.

Таким образом, необходима разработка комплексной, обоснованной на практических данных, методики оценки ценности данных о потенциальных заемщиках с выбором соответствующих экономических критериев и параметров с учетом особенностей деятельности МФО в контексте управления кредитными рисками.

Следующая методологическая проблема определяется *выбором концепции для оценки кредитных рисков: статистическое моделирование и/или методы и алгоритмы машинного обучения.*

В большинстве случаев разработка моделей оценки кредитных рисков строится на основе классических статистических методов — линейный дискриминантный анализ и логистическая регрессия, преимуществом которых

является достаточная простота расчетов и интерпретируемость результатов. Слабой стороной данных методов является зачастую недостаточная точность оценки ввиду слабой линейной зависимости большинства наборов данных.

Для устранения данного недостатка начали разрабатывать более сложные модели оценки, базой которых выступают модели усовершенствованной логистической регрессии [249], нечетких множеств [51], нейронных сетей [27] и другие.

В исследовании [289] представлена модель нелинейного непараметрического прогнозирования риска невозвратов или просрочек по потребительским кредитам. Для выявления дефолта держателей кредитных карт авторы использовали набор данных кредитных бюро и транзакции клиентов коммерческих банков. Преимущество данной модели заключается в ее динамическом характере и возможности адаптации к кредитным циклам в зависимости от поведения заемщиков. Однако существенная сложность заключается в получении информации по всем транзакциям заемщиков.

В работе [365] проведена разработка модели кредитного рейтинга на основе нейронной сети с использованием обширного набора преимущественно бухгалтерских данных по транзакциям и остаткам на счетах клиентов для выявления проблемных потребительских кредитов. Несмотря на то, что рассмотренная модель имеют бóльшую точность в сравнении с классическими, слабым местом таких моделей является огромное количество показателей, сложность определения относительной важности потенциальных входных переменных и интерпретации, а также длительность обучения.

Большое количество исследований [149, 170, 299, 325] посвящены разработке моделей оценки кредитных рисков на основе гибридных ансамблевых систем путем двухэтапного моделирования, в рамках которого строится регрессионная модель кредитного скоринга, затем полученные значимые переменные используются в качестве входных узлов для моделей нейронных

сетей. Недостатком гибридных моделей оценки кредитного риска является их сложность и достаточно большой набор показателей.

Все рассмотренные выше модели были разработаны для проведения оценки кредитных рисков в банковских организациях. В отношении оценки кредитных рисков в микрофинансовых организациях интерес представляет работа [267], в которой была разработана модель прогнозирования погашения просроченной задолженности на базе расширенной регрессии Кокса (модели пропорциональных рисков) [48]. Измерение проводилось на базе оценки переменных по шести группам. Данная модель позволила определить, что наибольшее влияние на возможность погашения просроченной задолженности в микрофинансовых организациях оказывают поведенческие показатели. Следует отметить, что модель регрессии Кокса широко применяется при моделировании бинарных событий, как и в случае оценки кредитных рисков, с учетом фактора времени [191, 195, 209, 297, 363].

Большое количество научных исследований направлено на проведение сравнительного анализа возможностей использования различных методов статистического моделирования и алгоритмов машинного обучения для оценки кредитных рисков, хотя существенная часть таких работ апробирована на данных банковских учреждений.

Так, например, в исследовании [168] проводится сравнительный анализ эффективности моделей классификации заемщиков для оценки кредитных рисков с апробацией на данных по кредитам розничных банков Бразилии. Авторы разработали несколько моделей на основе алгоритмов машинного обучения, включая метод опорных векторов, градиентный бустинг (AdaBoost) и случайный лес, сравнили их между собой, а также с эталонной моделью логистической регрессии с точки зрения предикативной точности. По результатам исследования авторы пришли к выводу о наибольшей эффективности ансамблевых методов машинного обучения – случайный лес и градиентный бустинг (AdaBoost).

В ходе эмпирического исследования [252] проведено сравнение характеристики четырех моделей машинного обучения: дерево решений, метод опорных векторов, метод k-ближайших соседей, искусственной нейронной сети со стандартным методом статистического моделирования – логистической регрессией. В результате авторы пришли к выводу о наилучшей прогностической способности метода опорных векторов за счет полноты модели. Однако при оценке качества модели на основе параметров F2-меры и точности-полноты (AUC-PR) в наилучшая производительность была достигнута моделью нейронной сети с глубоким обучением. В ходе эксперимента авторы также провели сравнительный анализ прогнозирования на двух наборах данных: первый набор содержит информацию исключительно по текущим кредитам заемщиков, второй набор данных включает как текущую, так и ретроспективную информацию. Все анализируемые модели при их адаптации к набору, включающему ретроспективные кредитные данные, получили большую точность. Исключение составил метод дерева решений.

Исследование [20] направлено на сопоставление точности прогнозирования дефолта заемщиков по потребительским кредитам в белорусских банках в рамках использования нескольких методов машинного обучения (дерево решений, нейронная сеть, случайный лес) и модели логистической регрессии. Наилучшую предикативную способность при выявлении дефолтных заемщиков продемонстрировала нейронная сеть.

Исследователи [364] провели сравнительный анализ возможностей различных моделей оценки кредитного риска на эмпирических данных 2,5 миллионов транзакций финансовых организаций Бразилии. Авторы использовали методы статистического моделирования (логистическая регрессия) и методы машинного обучения (дерево решений, метод опорных векторов, искусственные нейронные сети, метод случайного леса, градиентный бустинг – Adaboost и XGBoost) и пришли к выводу, что ансамблевые модели машинного обучения, в частности, XGBoost существенно превосходит по точности прогноза традиционные методы статистического моделирования (логистическую

регрессию). Данное исследование было проведено на данных по банковским кредитным организациям.

Автор статьи [336] разработал модель факторизации (модель выделения множителей) для оценки кредитного риска и сравнил ее с другими моделями (логистическая регрессия, метод опорных векторов, метод k-ближайших соседей и нейронная сеть) путем проведения эмпирической оценки на реальных данных финансово-кредитных учреждений Китая. Разработанная модель показала большую точность и эффективность в вычислительном отношении по сравнению с другими моделями.

В работе [147] на основе алгоритмов машинного обучения проведена оценка эффективности скоринговых моделей (XGBoost, Adaboost, случайный лес и других) для прогнозирования кредитоспособности новых клиентов микрофинансовых организаций при отсутствии кредитной истории. Все оцениваемые модели дали примерно одинаковый результат – точность до 80%. Данное исследование не включает разработку конкретной модели для проведения оценки кредитных рисков микрофинансовых организаций.

Таким образом, авторы рассмотренных исследований доказывают эффективность тех или иных методов традиционного статистического моделирования или алгоритмов машинного обучения [327], либо их комбинаций в зависимости от выбора переменных, тестирования на различных выборках, применения данных из различных источников. При этом большая часть научных разработок и сравнительного анализа проводится в рамках оценки кредитных рисков в банковских организациях.

Однако остается открытым актуальный вопрос выбора и целесообразности использования простых интерпретируемых математических моделей статистического моделирования или сложных, не интерпретируемых моделей на основе алгоритмов и методов машинного обучения, для оценки кредитных рисков МФО в условиях ограниченности их финансовых, информационных и

технических ресурсов, что определяет данную методологическую задачу настоящего диссертационного исследования.

Четвертая методологическая проблема заключается в *определении алгоритма методики построения модели оценки кредитных рисков*.

В научной литературе [147, 217, 312, 216] по применению моделей оценки кредитных рисков в микрофинансовых организациях в большинстве случаев акцент делается на математической спецификации используемой модели, а также результатах моделирования.

В отношении собственно методики построения моделей оценки кредитных рисков можно выделить только несколько работ зарубежных авторов. При этом данные исследования направлены на рассмотрение методики построения моделей оценки кредитных рисков в банковских организациях.

В работе [358] представлен подробный алгоритм процесса разработки скоринговой карты для оценки заемщиков в банковском кредитовании. Данный алгоритм состоит из последовательных этапов, включая планирование проекта разработки скоринговой карты (выработка целей, задач, рисков проекта и т.д.); выбор данных, включая оценку их доступности и качества; формирование базы данных с выбором параметров оценки в отношении заемщиков и их кредитного поведения; непосредственное построение скоринговой карты на основе выбранной математической модели; тестирование разработанной модели; внедрение модели и проверка ее валидности.

Монография [164] посвящена инструментам разработки моделей кредитного скоринга с выработкой соответствующих этапов, включая теоретическое описание возможных к использованию в оценке кредитных рисков моделей и методик оценки их прогностической и классификационной способности, работу с данными (выбор количества, качества, источников, методов подготовки), разработку скоринговой карты (выбор параметров оценки, определение классификационных групп заемщиков), калибровку и оценку

валидности разработанной модели, внедрение и контроль эффективности использования модели в практике кредитных организаций.

В работе [373] подробно описаны различные статистические и нестатистические модели, которые могут быть использованы для оценки кредитных рисков, а также методы оценки их классификационной и прогностической способности. Также в исследовании определены возможные источники информации для оценки кредитных рисков, включая макроэкономические показатели и поведенческие характеристики заемщиков. Однако непосредственно алгоритм построения и применения моделей оценки кредитных рисков в данном исследовании отсутствует.

Руководство по кредитному скорингу [253] содержит научные работы различных авторов по отдельным этапам построения системы оценки кредитных рисков, включая сравнительный анализ моделей, методику построения и оценки валидности моделей, выбора и подготовки данных, стратегии успешного внедрения моделей в риск-менеджмент банковских организаций.

Таким образом, в настоящее время отсутствуют исследования, направленные на системное изложение методики построения таких моделей, исходя из проблем подготовки набора данных для моделирования с учетом особенностей МФО. В связи с этим актуальность приобретает решение задачи выработки подробной методики построения моделей оценки кредитных рисков в МФО.

Пятая методологическая проблема заключается в *формировании лимитной политики для управления кредитными рисками*.

Сравнительный анализ особенностей микрофинансового рынка и банковского сектора [129] показал, что разделение заемщиков на две группы: платежеспособных и потенциально дефолтных в рамках СПР в системе управления рисками в МФО в отличие от банков в большинстве случаев оказывается неэффективным. Таким образом, для повышения эффективности управления кредитными рисками, исходя из имеющихся статистических данных

по клиентам за продолжительный промежуток времени, МФО необходимо разделять заемщиков на основе модели оценки кредитных рисков на децили по соответствующему уровню риска и устанавливать для каждого дециля кредитные лимиты для получения максимальной прибыли.

Ввиду того, что для экономики МФО наряду с уровнем риска немаловажное значение имеет конверсия из заявки в одобрение, а основной доход МФО получает за счет повторных клиентов, построение СПР на основе сегментации по уровню риска должна включать определение кредитного лимита как по первичным, так и по повторным заемщикам, для обеспечения максимального уровня конверсии с приемлемым риском и максимизации доходности по портфелю. В данном контексте необходимо решение задачи поиска эффективной математической модели для оценивания кредитных лимитов в МФО.

Исследование подходов к управлению кредитными лимитами в отечественной и зарубежной научной литературе показало, что проблема управления лимитами изучена также по большей части для банковской сферы. Так, в работе [92] моделируются кредитные лимиты для экспресс-кредитования по банковским картам, при этом автор предлагает решение задачи оптимизации для установления кредитных лимитов по типам заемщиков. В статье [58] представлен алгоритм определения кредитного лимита в случае экспресс-кредитования, основанный на кредитной истории и доходах клиента. Среди зарубежных исследований можно выделить работу [222], в которой на основе статистики приводится описание эффектов установления лимитов по кредитным картам на вероятность дефолта заемщика и потенциальные убытки банка.

В большинстве случаев в банках для оценки кредитного риска заемщиков делят на два класса, выбирая точку отсечения по скоринговой модели. Такой подход можно считать самым простым вариантом использования политики лимитов: лимит равен нулю или лимит равен сумме запрашиваемого займа, размер которого сверху ограничен общей кредитной политикой. Например, в статье [361] была рассчитана наилучшая «точка отсечения» (cut-off) по скоринговому баллу с помощью методов логистической регрессии и нейронных

сетей. Авторы нашли значение, которое минимизирует общую ошибку модели. В моделировании логистической регрессии точка отсечения – это точка определения принятия или непринятия кредитной заявки. Если вероятность дефолта становится больше точки отсечения, клиент будет в классе «плохих» клиентов, в противном случае – в классе «хороших» клиентов. В нейронной сети для расчета скорости предсказания сети используется порог правильного предсказания относительно неправильного, аналогичный точке отсечения в логит-моделировании. Использование нейронной сети показывает, что наилучшая точка отсечения, полученная с помощью логит-моделирования, также подходит для нейронной сети.

Близкое к тематике данной статьи исследование описано в работе [212], где авторы предложили гибридный метод интеллектуального анализа данных из двух этапов обработки. На этапе кластеризации выборки первичных и повторных клиентов группируются в однородные кластеры, которые представляют собой рейтинговые сегменты по уровню риска. Отличие от традиционных моделей оценки дефолтности заемщиков состоит в том, что клиенты были разделены на три или четыре класса, а не на два класса («хорошие» и «плохие»). Были получены экспериментальные результаты, основанные на примере данных по кредитным операциям одного из китайских банков. При выборе правильной точки отсечения для каждого класса достигается точная идентификация «хорошей» и «плохой» кредитной истории. Далее в соответствии с характеристикой каждого класса разрабатываются стратегии управления рисками.

Что касается лимитной политики микрофинансовых организаций, то она малоизучена в научной литературе. Например, в статье [8] приводится алгоритмизация деятельности МФО с учетом кредитной истории заемщика и поступлениям от него денежных средств. В статье [200] было изучено влияние размера займа на кредитный риск микрофинансовых организаций с применением традиционных банковских методик, в частности, обобщенного метода моментов и общей линейной модели с фиксированными эффектами (Fixed-effect model). Авторы данного исследования пришли к выводу, что кредитные риски

положительно связаны с размерами займов в МФО, т. е. небольшие займы несут меньший риск по сравнению с более крупными суммами, что требует необходимости разработки соответствующей кредитной политики в МФО.

Исследование [216] направлено на разработку модели кредитного скоринга с использованием алгоритма градиентного бустинга (XGBoost) с соответствующей настройкой гиперпараметров и выбора функций для оптимизации результатов оценки рисков. Одним из параметров оценки выступает сумма займа или лимит.

В российских исследованиях отсутствуют модели и практические рекомендации по назначению лимитов для различных групп заемщиков, в зависимости от уровня их дефолтности. Так, в статье [78] разработана модель оценки вероятности просрочки выплаты займа более, чем на 30 дней для заемщиков МФО на основе методов логистической регрессии, алгоритма случайного леса и алгоритмов построения деревьев классификации. Несмотря на то, что данная модель показала хорошую валидность в оценке вероятности дефолта заемщика, авторы не дают рекомендации по установлению лимитов с целью снижения кредитных рисков в будущем.

В целом проведенное исследование показало, что проблематика управления кредитными лимитами поднималась многократно в банковской сфере. При этом предлагаемые решения в большинстве работ не связаны с математическим моделированием, а целью более углубленных исследований является выбор точки отсечения и вероятности дефолта заемщика. Кроме того, предпринимались попытки обоснования назначения лимитов в зависимости от сегмента клиентов по степени риска, т. е. применение сегментированного подхода в назначении лимитов в банковской сфере.

В отношении микрофинансовых организаций в научных работах, как правило, проводится оценка влияния суммы предоставляемого займа и кредитного лимита в доходности МФО либо рассматриваются скоринговые модели в контексте вероятности дефолта без привязки к лимитной политике

микрофинансовых организаций. Таким образом, в микрофинансовом секторе и, в частности, на рынке PDL-займов вопрос назначения лимитов на основе математического моделирования не проработан, что приводит к актуальности комплексного применения результатов оценки кредитного риска на уровне заемщика, классификацию заемщиков с последующим расчетом лимитов с помощью отдельной математической модели.

Шестая методологическая проблема заключается в учете множества бизнес-метрик через общую модель юнит-экономики при управлении кредитными рисками.

Как было определено ранее, базой системы принятия решений для управления кредитными рисками в МФО выступают либо общие скоринговые модели БКИ, либо собственные модели оценки для формирования кредитного рейтинга заемщика и для оценки кредитного риска МФО. Такие модели строятся либо на базе традиционных статистических методов, либо на основе алгоритмов машинного обучения. Однако принятие решений возможно не просто посредством оценки риска заемщика, а путем тестирования данных через общую модель юнит-экономики, которая позволяет оценивать финансовый результат на всех этапах жизненного цикла клиента МФО. При этом следует отметить, что модель юнит-экономики является не альтернативой рассмотренному традиционному подходу к оценке в рамках скоринговых моделей, а выступает дополнением для принятия решений по управлению кредитными рисками МФО. В рамках моделей юнит-экономики МФО реализуется подход моделирования риска и параметров бизнес-метрик по всему кредитному портфелю МФО. Таким образом, автором в работе [103] по отношению к особенностям ведения бизнеса МФО, было определено, что юнит-экономика представляет собой систему принятия решений в МФО при управлении кредитным риском, нацеленную на получение максимальной прибыли с одного клиента в долгосрочной перспективе на всем сроке «жизни» клиента. При этом актуальность применения юнит-экономики в МФО определяется убыточностью первого займа в связи со

значительными затратами на привлечение клиентов и возможностью выхода на достаточные объемы прибыли при выдаче повторных займов.

В данном контексте следует провести обзор научных публикаций по исследованию феномена юнит-экономики МФО в рамках теоретических и практических концепций.

В работе [130] проведено исследование возможностей использования принципов юнит-экономики в рамках управления бизнес-процессами. Автор приходит к выводу о том, что использование подхода к управлению на базе юнит-экономики позволяет выявить неэффективные бизнес-процессы в компании посредством вычисления соответствующих метрик прибыльности. Однако для проведения оценки автор не разрабатывает модели, а использует стандартные маркетинговые метрики.

В статье [16] выявлены возможности масштабирования бизнеса на основе применения соответствующих метрик юнит-экономики, выбор которых определяется при помощи сравнительного анализа влияния различных показателей деятельности на финансовые результаты предприятия вследствие его расширения по географическим и операционным сегментам.

Исследование [15] направлено на определение возможностей проведения анализа стратегий продаж в онлайн-среде посредством метрик юнит-экономики. Проведенный автором сценарный анализ предполагает последовательную оценку влияния выделенных метрик на прибыль и соответственно эффективность цифровой бизнес-модели компании.

Автор исследования [13] посредством выделения метрик юнит-экономики и построения соответствующей модели для проведения оценки ключевых параметров функционирования предприятий малого бизнеса доказывает целесообразность применения подходов к диверсификации их деятельности.

В статье [50] обоснована актуальность применения юнит-экономики с точки зрения управления прибылью предприятия, сформированы два варианта базовой формулы прибыли предприятия на основе метрик юнит-экономики во взаимосвязи основных бизнес-процессов и маркетинговой деятельности

предприятия. Разработанные модели оценки апробированы на данных предприятия пищевой промышленности.

Исследование [85] посвящено взаимосвязи клиентоориентированного подхода и юнит-экономики посредством разработки модели, учитывающей ключевые бизнес-показатели с точки зрения полученной прибыли по отношению к отдельному клиенту, а также по отношению к единице реализованной продукции. Данная модель была апробирована на примере компании IT-отрасли.

В отношении особенностей юнит-экономики в смежных с микрофинансированием областях интерес представляет статья [271], автор которой приходит к выработке направлений развития страховой деятельности на основе количественных данных. При этом в качестве юнита (единицы) выступает как страхователь (клиент), так и единичный страховой продукт (полис), затраты на реализацию которого в совокупности не должны превышать доход страховой компании.

Проведенный обзор научной литературы показал, что в настоящее время отсутствуют научные публикации по вопросам юнит-экономики как в целом в финансовой сфере (включая банковские, страховые, лизинговые организации), так и в отношении микрофинансовых организаций, в частности. При этом в специфике функционирования МФО ключевая роль юнит-экономики проявляется в контексте принятия решения при управлении кредитным риском, что обуславливает актуальность решения данной методологической задачи – построение общей модели юнит-экономики на основе множества бизнес-метрик для управления кредитными рисками МФО.

Седьмая методологическая проблема – *управление кредитными рисками с учетом изменения параметров моделей оценки кредитных рисков во времени.*

Как было определено ранее, управление кредитными рисками в микрофинансовых организациях предполагает качественную классификацию заемщиков по риску возникновения дефолта, которая позволяет снизить существенные потери от операций по кредитованию. Такая классификация

производится на базе классических методов статистического моделирования или на основе алгоритмов машинного обучения. Полученные статистические наборы данных делятся на обучающие и тестовые. Сначала используемый метод тестируется на обучающей выборке путем соответствующей подгонки параметров, а затем на тестовой выборке производится последовательная проверка точности разработанной модели. Таким образом, традиционные скоринговые модели в кредитных учреждениях подразумевают подгонку параметров к полученной статистике с последующим применением модели к потенциальным заемщикам. Однако если параметры истинной модели изменяются во времени, такой подход подразумевает применение устаревших значений параметров, что приводит к существенной неэффективности используемой модели.

Как показало ранее проведенное исследование [104] в научной литературе данная тематика не проработана и представляет собой научную новизну.

Среди отечественной литературы существуют научные исследования, которые содержат обзоры применяемых в скоринге методов. В статье [21] представлен широкий набор современных методов, применяемых в решении задач кредитного скоринга. Однако методы представлены отдельно друг от друга, не рассматривается их комбинация и сравнительный анализ. Кроме того, не поднимается вопрос динамически меняющихся коэффициентов.

В работе [143] динамика кредитного портфеля анализируется с точки зрения анализа выживаемости через случайный лес выживаемости. Такой подход имеет ряд недостатков, связанных с интерпретацией процесса принятия решений, а также базируется на анализе исторических данных, хотя и может давать точные оценки.

В отечественной научной среде также предпринималась попытка динамического отбора ансамблевых моделей в решении задач оценки кредитных рисков. Так, в работе [41] производительность ансамблевых моделей в задаче классификации была повышена с помощью методов динамического выбора.

Основной недостаток методов динамического выбора заключается в их сложности относительно методов статического выбора.

В статье [142] была проведена оценка вероятности дефолта заемщика не только для первого года жизни кредита, но и для всего срока жизни кредитного портфеля с помощью динамической модели логистической регрессии, учитывающей на каждом новом периоде данные о качестве исполнения кредитного договора заемщиком и фазе макроэкономического цикла.

Работа [26] посвящена сравнительному анализу моделей градиентного бустинга с динамическими параметрами, была получена хорошая точность классификации. Кросс-проверка позволила повысить качество построенных моделей при одновременном снижении скорости работы алгоритма.

В меняющихся условиях функционирования бизнес-структур зачастую применяют модели, находящие закономерности в данных, которые не выявляют традиционные методы. Так, например, в статье [34] описан алгоритм технологии «добычи данных», которые базируются на самообучении. Однако в исследовании не учитываются меняющиеся условия экономической среды, что является недостатком исследования.

В работе автора [104] показано, что динамические модели распространены в прогнозировании макроэкономических процессов (например, статья [323]). Динамическое моделирование риска получило широкое распространение в управлении финансовыми рисками. Так, в статье [196] показывается на примере многомерной GARCH-модели, что динамические модели, популярные в академической среде и редко используемые в бизнесе ввиду своей сложности, дают более точные оценки, чем более простые статические модели, основанные на исторических данных.

Также в исследованиях анализируется использование многомерных эконометрических моделей для прогнозирования волатильности и динамических корреляций между различными активами. Например, в исследовании [275] используется динамическая многомерная модель обобщенного авторегрессионного скор-процесса (GAS) для анализа и прогнозирования

волатильности и динамических корреляций недельных цен на сырое пальмовое масло, кокосовое масло, соевое масло и сырую нефть. Сравниваются результаты данной модели с моделью динамической условной корреляции (DCC–GARCH).

В работах [259, 352] применяется фильтрация Калмана в кажущихся несвязанными моделях регрессии для вычисления оценок, изменяющихся во времени параметров (состояний) и функции правдоподобия для ее гиперпараметров. С помощью аппроксимации сглаживания соответствующим решением фильтрации ЕМ-метод (англ. Expectation Maximization Method) используется как полностью рекурсивный метод оценки для гиперпараметров на основе теории MINQUE (англ. Minimum Norm Quadratic Unbiased Estimator). Проводится имитационное исследование для модели с одним уравнением и демонстрируется эффективность такого подхода.

Однако большинство рассмотренных работ и моделей не включают динамический характер параметров. В научной литературе по оценке кредитных рисков данная тематика практически не освещается, хотя она очень популярна в прогнозировании временных рядов. Примерами могут служить работы [181, 198, 282].

Фактор времени при моделировании бинарного события риска – наступления дефолта по микрозайму, может быть отражен через модель пропорциональных рисков Кокса [281]. Но данная модель в основном используется в анализе выживаемости, а не при оценке кредитных рисков.

В целом проведенное исследование показало, что в отечественной и зарубежной научной среде [173, 270, 284, 306, 326] неоднократно рассматривался вопрос учета динамических факторов при решении задач оценки кредитных рисков с применением методов машинного обучения. При этом отсутствует соответствующий математический подход к моделированию, который мог получить практическое применение в практике как банковских кредитных организаций, так и МФО.

Восьмая методологическая проблема заключается в *оценке кредитных рисков в условиях отсутствия исторических данных для построения соответствующей модели.*

Микрофинансовые организации с целью управления кредитными рисками разрабатывают СПР, которые позволяют в автоматизированном режиме принимать решение о выдаче кредитов и представляют собой алгоритмы оценки на базе сформированных математических моделей с использованием определенной информации о заемщиках в исторической ретроспективе. Данные кредитных историй, как правило, являются основным источником ретроспективных данных для оценки кредитного риска МФО в рамках скоринговых моделей. Достаточно высокая трудоемкость и затратность построения модели оценки кредитных рисков вынуждают МФО приобретать уже готовые агрегированные показатели скоринговых баллов, рассчитанных в рамках обобщенных моделей кредитного риска, в БКИ.

Обзор научной литературы по моделированию кредитного риска заемщиков – физических лиц на основе данных кредитной истории показал недостаточную проработку данной проблемы.

В статье [353] была предпринята попытка моделирования индивидуального платежного поведения клиентов при приобретении товаров в рассрочку через банковские и небанковские кредитно-финансовые организации, в рамках которой на основе ретроспективного анализа счетов клиентов (показатель индивидуальной ликвидности) и данных кредитной истории (показатель жизненного цикла платежей) производится оценка приемлемого периода, в течение которого клиенты могут оставаться в просрочке, прежде чем они будут идентифицированы как дефолтные. Недостаток данной модели заключается в том, что МФО не всегда имеют доступ к информации о движении средств по счетам клиентов, в отличие от банковских учреждений.

Статья [324] направлена на моделирование кредитного поведения заемщика на основе методов машинного обучения с использованием данных по финансовым

характеристикам и социальному взаимодействию заемщика. В качестве финансовых характеристик автором были определены два показателя: кредитные действия заемщика в текущий период и ретроспективные кредитные действия за последние 3 и 6 месяцев. Кредитные действия определяются набором данных совокупного количества и объема действующих кредитов заемщика с разбиением их по типу кредита, по наличию просрочек в каждом из типов кредитов, а также сумму повторных займов. Представленная в статье модель позволяет охарактеризовать каждого заемщика в соответствии с его кредитным поведением, при этом точность и эффективность оценки кредитного риска повышается в более длительном временном разрезе, т. е. при анализе кредитного поведения заемщика за 6 месяцев. Недостатком данной модели является ее достаточная сложность в интерпретации и большой массив необходимых данных для ее построения.

В статье [340] на основании данных кредитной истории заемщика, включая общую сумму и общее количество кредитов в ретроспективе за последние шесть месяцев, просроченные кредиты в процентах от общего количества кредитов, время, прошедшее от получения первого кредита до текущего времени, количество непогашенных кредитных заявок за последние 6 месяцев, разработана модель оценки кредитного риска заемщика путем определения персонального кредитного рейтинга в диапазоне от 0 до 1 на основе модели взвешенного среднего логистической регрессии. По результатам оценки было определено, что клиенты с длительной кредитной историей и показателями хорошего кредитного поведения имеют более низкий кредитный риск. Автор заключил, что ретроспективная кредитная информация, т. е. индивидуальная кредитная история и персональные характеристики заемщиков, необходимы для получения более точной оценки кредитного риска. Недостатком данной модели также является ее достаточная сложность в интерпретации и большой массив необходимых данных.

Исследование [375] посвящено оценке кредитного риска клиентов с небольшой и разреженной во времени или совсем отсутствующей кредитной историей. Автор разработал модель оценки кредитного риска на основе методов

машинного обучения, которая позволяет использовать как данные кредитной истории заемщика – количество записей в отчете бюро кредитных историй, количеств и суммы кредитов, а также альтернативные источники данных, включая ретроспективный реестр транзакций (включая поступления и расходы) и платежное поведение (различные виды задолженностей) потенциальных заемщиков, для более точной оценки кредитного риска. Разработанная автором модель использует большое количество источников данных и также достаточно сложна для интерпретации.

В статье [44] разработана модель определения процентной ставки по кредиту в зависимости от кредитного рейтинга заемщика, формируемого как агрегированный показатель кредитной истории – сумма компонент: история платежей, размер текущей задолженности, длительность кредитной истории и новые кредиты (полученные в течение последнего года).

При этом все рассмотренные модели были сформированы и апробированы на банковских учреждениях, которые могут проводить оценку на достаточно большом массиве данных.

Таким образом, проведенное исследование показало, что в настоящее время в научной литературе и практических наработках МФО отсутствует релевантная методика оценки рейтинга заемщика по кредитным историям, что является существенным препятствием для оценки кредитоспособности отдельных заемщиков и кредитных рисков микрофинансовых организаций в целом. Это определяет актуальность разработки методики измерения риска заемщика микрофинансовой организации на основе данных кредитных историй, которая может стать альтернативой обобщенным скоринговым моделям БКИ, а также сложным для расчета и интерпретации моделям с большим количеством параметров.

Выводы по главе 1

В данной главе было проведено исследование тенденций и закономерностей развития рынка микрофинансирования в России, которое показало, что динамичность экономических условий функционирования МФО, а также изменение в сторону ужесточения требований регулятора к МФО приводит к перераспределению рынка в пользу наиболее крупных игроков, главным преимуществом которых являются более гибкие системы управления кредитными рисками и возможность формировать портфель с упором на PDL-займы.

Проведенный иерархический кластерный анализ российских МФО в сегменте PDL-займов подтвердил значимость всех конъюнктурных изменений на данной рынке – состав кластеров изменяется в зависимости от масштаба МФО и динамики ее роста за счет качества управления кредитным портфелем. Исследование особенностей деятельности МФО позволило определить ряд ограничений финансового, временного и информационного характера, которые характеризуют основные проблемы системы управления кредитными рисками МФО, решение которых возможно посредством разработки соответствующих математических моделей, направленных на повышение качества проведения оценки кредитных рисков в разрезе разных категорий заемщиков.

Глава 2 Особенности информационного обеспечения процедур оценки кредитных рисков в микрофинансовых организациях

2.1 Виды исходной информации и особенности ее выбора для оценки кредитных рисков микрофинансовых организаций

Определение видов исходных данных, необходимых для оценки кредитных рисков МФО, имеет существенную актуальность в условиях ограниченности информационных и финансовых ресурсов. На российском рынке информационного обеспечения представлено множество компаний и продуктов для оценки риска мошенничества, оценки кредитных рисков. Все эти информационные продукты имеют свои особенности в зависимости от провайдера данных и специфику использования для оценки кредитных рисков МФО. Как уже было отмечено выше в параграфе 3 главы 1, перед МФО встает проблема выбора оптимального набора источников данных для оценки кредитных рисков, которую невозможно решить без системной классификации существующих источников данных.

Изучение рынка данных для оценки кредитных рисков, обзор научной литературы по данной тематике и практический опыт автора в управлении кредитными рисками в российских МФО позволили выделить несколько групп критериев классификации источников данных, представленных в таблице 2.1 и отраженных в работах автора [101, 116]. Также особенности информационного обеспечения и технологий, которые могут быть использованы при оценке кредитных рисков в МФО изложены в работах [121, 137, 138].

Далее рассмотрим подробно предложенные в таблице 2.1 критерии для классификации видов исходных данных для оценки кредитных рисков с учетом специфики МФО.

Таблица 2.1. – Критерии классификации источников данных для оценки кредитных рисков МФО

Критерий	Вид данных	Характеристика
а) Доступность данных	Открытые	Общедоступность Бесплатность Возможность получения из открытых источников
	Закрытые	Платные Получение на договорной основе
б) Способ обработки данных	Необработанные	Исходные (первичные) данные
	Обработанные (агрегированные)	Проанализированные первичные данные
в) Характер источника данных	Данные из бюро кредитных историй	Данные кредитного поведения
	Транзакционные данные	Данные о финансовых операциях заемщика
	Данные органов власти (ФССП, налоговой инспекции и других)	Данные об исполнительном производстве
	Данные операторов сотовой связи	Данные по биллинговой активности и геолокации заемщика
	Данные социальных сетей, мессенджеров	Данные о социальной активности заемщика
	Данные об устройстве заемщика	Цифровой отпечаток устройства
	Биометрические данные	Физиологические особенности заемщика
	Данные, полученные путем анкетирования заемщиков	Социально-демографические данные
г) Цели применения данных	Оценка риска мошенничества	Данные по мошенническим операциям
	Оценка платежеспособности заемщика	Данные по наличию у заемщика потенциальных средств для возврата кредита
	Оценка кредитного риска	Данные кредитного рейтинга заемщика
	Повышение уровня одобрения заявок	Альтернативные данные
д) Временной период	Моментные показатели	Данные, получаемые в определенный момент времени
	Интервальные показатели	Данные, агрегированные за определенный период времени
е) Возможность получения данных на	Источники данных с возможностью проведения ретро-тестов	Присутствует техническая возможность проведения

Критерий	Вид данных	Характеристика
ретро дату		ретро-тестирования
	Источники данных без технической возможности проведения ретро-тестов	Отсутствует техническая возможность проведения ретро-тестирования
ж) Способ отправки информации	Протоколы обмена данными	REST API GraphQL gRPC SOAP XML
з) Наличие согласия заемщика	Требующие согласия заемщика	Практически все типы данных
	Не требующие согласия заемщика	Только информация об устройстве

Источник: составлено автором

а) По доступности данные можно разделить на открытые и закрытые.

Открытые данные являются общедоступными (бесплатными) и могут быть получены из открытых информационных источников. Примером открытых данных могут быть:

- сведения о клиенте, имеющиеся в открытых профилях различных социальных сетей. Такие данные могут включать количество друзей и подписчиков, активность взаимодействия в соцсетях – комментарии, реакции, частота выкладки контента, тематика контента и другая информация;
- сведения о клиенте с порталов государственных органов – информация об уплате налогов по данным ФНС, штрафов ГИБДД, различных правонарушений, судебных разбирательствах, исполнительных производствах и другие.

Закрытые данные предоставляются различными компаниями на рынке информации и бюро кредитных историй на договорной (платной) основе. Примерами закрытых данных являются:

- кредитные истории, предоставляемые соответствующими кредитными бюро;
- агрегированные данные о платежах по банковской карте в виде скорингового балла (обезличенные транзакционные данные), такая информация

может быть запрошена на платной основе в Национальной системе платежных карт;

- данные о геолокации клиента при помощи специальных платных сервисов и баз данных и другая информация, которая может быть получена на платной основе.

б) По способу обработки данные могут быть необработанными и обработанными (агрегированными).

Необработанные данные – это исходные данные, над которыми не производилось никаких действий по их обработке. Например, это данные в формате xml-файлов по кредитным историям заемщика в кредитном отчете.

Необработанные исходные данные полезны в случае самостоятельного детального анализа риска дефолта заемщика и построения скоринговых моделей. Точность такого скоринга на исторических данных конкретной МФО, как правило, выше универсальных моделей (англ. Generic Models), предлагаемых на рынке на коммерческой основе. Однако работа с исходными «сырыми» данными требует наличие определенного штата квалифицированных специалистов в МФО и временных затрат.

Обработанные данные – это данные, полученные в результате анализа и обработки первичных данных: отдельные информативные переменные или атрибуты, агрегированные данные, скоринги или результаты работы моделей по оценке заемщика. Обработка первичных данных и расчет переменных для построения моделей или оценки кредитного риска может происходить самостоятельно в МФО, а также такие переменные могут быть предоставлены от провайдера данных. Особый тип обработанных данных – это готовые скоринговые баллы (оценки) заемщика на основе построенных провайдером данных универсальных моделей.

Использование скорингов так называемых общих моделей или готовых обработанных данных является оптимальным решением, например, при выводе на рынок нового кредитного продукта или запуске новой МФО. Существенным недостатком обработанных данных является, как правило, отсутствие сведений об

алгоритмах их предварительной обработки. У риск-аналитиков МФО в этом случае отсутствует возможность контроля процесса обработки исходных данных, получаемых от партнеров. При обновлении алгоритмов обработки данных на стороне провайдера могут произойти качественные изменения в получаемой информации, которые непременно отразятся на результатах их применения в СПР. Для этого риск-менеджерам необходимо организовывать с помощью отчетности оперативный контроль стабильности покупаемых данных. Тем не менее использование обработанных данных является достаточно распространенной практикой в российских МФО.

в) По характеру источника данные можно разделить на несколько типов. Этот критерий отражает источник формируемых данных о заемщике. Тут можно выделить множество типов данных, отметим их укрупненную группировку.

Данные кредитного поведения (кредитных историй) – отчеты из бюро кредитных историй, представляющие собой ретроспективные (за последние семь лет) данные по записям о кредитном поведении клиента, включая запросы на кредиты, выдачу кредитов, статистику погашений и просроченных задолженностей. Такие данные являются традиционными в оценке кредитного риска. Помимо кредитного отчета, бюро кредитных историй по запросу МФО могут предоставить персональные кредитные скоринги заемщиков, самостоятельно построенные на общих моделях оценки кредитного риска.

Транзакционные данные – это данные о финансовой информации по клиенту, связанные с его платежами. Это могут быть транзакции по банковским картам, данные электронных кошельков, информация о денежных переводах. В России необработанные данные о транзакциях представляют собой персональные данные и недоступны для оценки кредитных рисков. Однако на рынке существуют сервисы, которые продают эти данные в агрегированном виде или в виде скорингового балла модели, построенной по таким транзакционным данным.

К транзакционным данным также можно отнести информацию крупных интернет-магазинов, агрегаторов услуг и сервисов доставки, из которых можно получить сведения о количестве и частоте покупок, сумме покупок, адресах

доставки и купленных товарах, отзывы о заемщике, его рейтинг и геолокация покупок. Все эти данные могут быть использованы только в агрегированном виде и получены от операторов фискальных данных. В настоящее время в России только зарождаются сервисы по предоставлению таких агрегированных данных на основе фискальной информации.

Данные об исполнительном производстве – это данные из государственной информационной системы о государственных и муниципальных платежах (ГИС ГМП), банка данных Федеральной службы судебных приставов России (ФССП) и других государственных информационных ресурсах об имеющихся задолженностях заемщика. Существует множество сервисов и агрегаторов, предоставляющих такие данные. Это информация о банкротстве заемщика, о наличии правонарушений, задолженностях об уплате штрафов и налогов и др.

Данные операторов сотовой связи – это агрегированные данные о биллинговой активности заемщика, включая количественные и качественные показатели активности по звонкам, взаимодействию с интернет-сервисами, а также геолокационные данные клиента. Помимо этого операторы сотовой связи строят собственные модели скоринга клиентов, основанные на собираемых ими данных, и реализуют их на платной основе банкам и МФО, а также общие скоринги, построенные на основе этих данных. Подобную информацию на российском рынке предоставляют на коммерческой основе все крупные операторы сотовой связи, а также специализированные сервисы, обрабатывающие смс-сообщения.

Данные из социальных сетей и мессенджеров (от англ. Message, сообщение) позволяют получить информацию о социальной активности заемщика, сведения о социальном и поведенческом профиле заемщика посредством данных о количестве и тематиках публикаций, размещаемых фотографиях и видео, об участии в различных группах и подписках на каналы, наличие собственных групп и каналов. Несмотря на то, что такие необработанные данные в большинстве случаев могут быть получены на безвозмездной основе, их применение для оценки кредитных рисков в системе принятия решений вызывает сложность из-за

отсутствия структурированности и агрегирования. Оперативный сбор таких данных из множества источников и их предобработку осуществляют специальные платные сервисы мониторинга и аналитики социальных сетей.

Данные об устройстве заемщика – это данные о цифровом отпечатке и параметрах устройства пользователя. Это информация о самом устройстве – тип, модель, технические параметры, разрешение экрана и т. п.; о программном обеспечении устройства – операционная система, версия операционной системы, браузер, версия браузера, настройки браузера; о сетевом окружении устройства, например IP-адрес устройства; о часовом поясе и др. В основном такие данные востребованы в онлайн-кредитовании при оформлении заявок на кредит с сайта или с мобильного устройства для идентификации или аутентификации заемщиков и выявления случаев риска мошенничества. Но могут быть использованы в обратную сторону – как маркеры для выявления низко-рисковых заемщиков. Такие данные можно собирать самостоятельно при заполнении анкеты заемщиком на сайте. Однако ценность представляют значения переменных и атрибутов, а также скоринговых баллов, построенных по информации всего рынка. Такие данные предоставляют специализированные сервисы на платной основе.

Данные об устройстве заемщика могут собираться с помощью различных технологий. Важно различать технологии сбора данных об устройстве, например, через JavaScript или мобильную библиотеку (SDK) и технологии построения отпечатка устройства (англ. Finger Print). Таких новых информационных технологий очень много, они развиваются и совершенствуются.

Биометрические данные – это уникальные физиологические и биологические особенности заемщика, которые используют для установления его личности. Этот тип данных используется МФО для идентификации клиентов и устранения риска мошенничества в онлайн-кредитовании. Например, МФО часто запрашивают фотографию клиента с паспортом при заполнении анкеты на сайте для всех или высокорисковых в СПР сегментов заемщиков, хотя это требование воспринимается некоторыми заявителями как избыточное, что приводит к

снижению конверсии из анкеты в заявку на кредит. Практический интерес для МФО могут представлять сервисы автоматического распознавания и обработки лиц для идентификации заемщика.

Социально-демографические данные – это данные о поле, возрасте, регионе проживания или регистрации заемщика, семейном положении, наличии детей, социальном статусе, уровне образования и т. п. Как правило, эти данные получают из анкеты при заполнении заявки на кредит. Социально-демографическая информация о заемщике хорошо показывает себя с точки зрения ранжирования заемщиков по риску при самостоятельном построении скоринговых моделей при объеме выборке более сотен тысяч.

г) Следующий критерий классификации привязан непосредственно к бизнес-процессу оценки риска в МФО и целям использования данных в СПР. По целям применения данных можно выделить их следующие виды.

Для оценки риска мошенничества (англ. Fraud Scoring). Данное исследование рассматривает аспекты применения источников данных для управления кредитными рисками. Риск мошеннических операций представляет собой отдельный риск и требует разработки отдельных стоп-факторов антифродовой направленности, отдельных скоринговых моделей. Более того, при разработке скоринговых моделей оценки кредитных рисков из выборки должны исключаться мошеннические займы. Но управление кредитным риском во многих МФО включает и разработку правил идентификации заемщиков на основе оценки вероятности мошеннических операций.

Для оценки платежеспособности заемщика. СПР в МФО может быть построена на основе сегментированного подхода, когда сначала проводится верхне-уровневая сегментация по платежеспособности заемщика. Оценивается на основе этих данных наличие у заемщика потенциальных средств для возврата кредита. Затем платежеспособные клиенты ранжируются по степени кредитного риска.

Для оценки кредитного риска. Наиболее часто встречаемая цель использования данных. Это получение кредитного рейтинга заемщика на основе

данных или построенной на их основе модели оценки кредитных рисков, а также принятие решения о выдаче займа, исходя из кредитной политики финансовой организации [134].

Для повышения уровня одобрения заявок (англ. Approval Rate). Эта цель использования дополнительных данных в СПР заключается в обеспечении значимого роста конверсии из заявки в выдачу при неизменном или незначительно повышающемся риске. Альтернативные данные могут быть применены для принятия решения о выдаче займа сегменту без кредитной истории (или в случае отсутствия данных из основного источника для СПР), а также для скоринга отказных заявок. Для МФО, в отличие от банков, из-за короткого срока и ограниченных сумм кредита (речь идет в основном о PDL-займах), высокой маркетинговой стоимости привлечения первичных клиентов и получения прибыли только за счет повторных займов, в СПР более критична ошибка второго рода. Иными словами, идеальная СПР для МФО обладает высокой специфичностью (одобряет заявки как можно большему числу «хороших» клиентов). При этом чувствительность должна обеспечивать предельно допустимый риск согласно кредитной политике. Поэтому для минимизации ошибки второго рода дополнительные источники данных могут применяться для повторного скоринга отказных заявок.

д) В зависимости от временного периода можно выделить данные моментных показателей и интервальных показателей.

Моментные показатели – это данные, получаемые в определенный момент времени, как правило в момент оценки кредитного риска. Например, IP-адрес устройства в момент заполнения заявки на сайте.

Интервальные показатели – это данные, агрегированные за определенный момент времени до момента оценки заемщика. Например, количество запросов по заемщику в бюро кредитных историй за последнюю неделю.

Данный критерий влияет на технические особенности интеграции с провайдером данных и возможность проведения ретро-тестирования для выявления их ценности. При этом следует отметить, что большинство

моментальных показателей чаще всего не поддаются ретро-тестированию по своей природе.

е) В зависимости от возможности получения данных на ретро дату можно выделить источники данных *с возможностью проведения ретро-тестов* – получения значения показателей на определенную дату в прошлом, а также *без технической возможности ретро-тестирования*. Наличие технической возможности проведения ретро-тестов, а не только получения данных в «боевых» запросах СПР, позволяет риск-аналитикам МФО до внедрения нового источника данных в СПР предварительно оценить его ценность.

ж) По способу отправки информации при интеграции источника данных существует несколько возможностей их подключения в зависимости от технологии передачи. Не вдаваясь в технические подробности, кратко рассмотрим основные протоколы передачи данных, их достоинства и недостатки, которые должны учитывать риск-менеджеры при принятии решений (таблица 2.2), поскольку время на интеграцию источника данных в условиях ограниченных ресурсов и на поддержку его работы напрямую влияет на экономику МФО.

Таблица 2.2 – Сравнительные характеристики протоколов обмена данными

Характеристики	REST API	SOAP	GRPC	GRAPHQL
Формат данных	XML, JSON, POST, GET	XML	Binary	XML, JSON, POST, GET
Размер пакета данных	большой	большой	малый	большой
Сложность работы с протоколом	низкая	средняя	большая	средняя
Сложность отладки при возникновении проблем	низкая	низкая	большая	большая
Работа с интерпретируемыми языками	сервер и клиент	сервер и клиент	клиент	сервер и клиент

Источник: Составлено автором на основе открытой информации в сети интернет

REST API от англ. Representational State Transfer – наиболее часто используемый протокол. Преимуществами REST API являются большая распространенность и способность обработки различных форматов данных, масштабируемость и относительная простота.

Технология REST API имеет существенный недостаток – невозможность получения только определенной пользователем информации. Каждый запрос включает в себя набор неизменных переменных и передаваемый пакет данных имеет большой объем.

GraphQL от англ. Graph Query Language – это способ передачи информации, в котором данные представляются в виде графа. Этот протокол также поддерживает разные форматы данных. Существенные недостатки протокола – большой объем передаваемого пакета данных и сложность запросов.

GRPC от англ. Remote Procedure Calls – это система удаленного вызова процедур получения данных с открытым исходным двоичным кодом. Отсюда главное достоинство протокола – малый объем пакета передаваемых данных. Однако он более сложен в применении, по сравнению с распространенным REST.

SOAP XML от англ. Simple Object Access Protocol – данный протокол обычно используется для передачи данных Бюро кредитных историй, он позволяет передавать данные только формата xml. По своим характеристикам протокол является альтернативой REST API.

з) В зависимости от наличия согласия заемщика данные можно классифицировать *на требующие согласия заемщика и не требующие*. Данный критерий классификации регламентирует юридические аспекты получения данных. В России получение практически всех типов данных возможно только при наличии согласия заемщика. Без юридически подтвержденного согласия заемщиков можно запрашивать и собирать только информацию об устройстве при заполнении заявки на кредит на сайте.

Таким образом, предложенный системный подход к классификации источников и видов традиционных и альтернативных данных о заемщиках, их различных характеристиках, включая поведенческие, играет значимую роль в

оценке кредитных рисков МФО. При этом в контексте стремительного развития информационных ресурсов и пополнения данных приводит к необходимости использования помимо традиционных данных, зарекомендовавших свою эффективность в системе управления кредитными рисками, новых альтернативных источников данных. В Приложении А для иллюстрации приведены несколько компаний-провайдеров и классификация предоставляемых ими на коммерческой основе видов данных для оценки кредитных рисков, согласно предложенным в таблице 1.1. автором критериям.

Как было отмечено в параграфе 2 главы 1, одна из тенденций российского рынка микрофинансирования заключается в существенном смещении в сектор онлайн-кредитования. При этом при выборе вида исходных данных для построения СПР для оценки кредитных рисков МФО должна учитывать не только предложенную классификацию, но и требования базового стандарта совершения микрофинансовой организацией операций на финансовом рынке. Регламент проверок для оценки кредитных рисков МФО в России должен учитывать требования пункта 5.8. базового стандарта, согласно которому при выдаче онлайн-микрозайма микрофинансовая организация должна провести не менее пяти из десяти проверок, приведенных в базовом стандарте.

МФО вправе сократить до трех минимальное количество проверок для целей выполнения пункта 5.8 базового стандарта при условии одновременного проведение мероприятий, предусмотренных пунктами 5.8.3–5.8.5. Исходя из вышеизложенного, можно сделать вывод, что МФО должна акцентировать свое внимание на видах данных, позволяющих осуществить проверку телефона, платежной карты или счета и устройства клиента. Однако оптимальным является выбор 2–3 альтернативных друг другу видов данных по каждому набору мероприятий используемых проверок. Причем эти проверки могут быть осуществлены на основе агрегированных платных данных, а также самостоятельно агрегированных данных внутри МФО. Пример, такого выбора данных для разработки кредитной политики МФО представлен в таблице 2.3.

Таблица 2.3 – Пример выбора источников данных для оценки кредитных рисков с учетом требований базового стандарта оказания финансовых услуг

Пункт 5.8 Базового стандарта	Предмет проверки	Виды проверок	Рекомендации по использованию
1) Проверка действительности документа, удостоверяющего личность получателя финансовой услуги, с использованием сервисов органов государственной власти или иных сервисов, обеспечивающих сопоставимый с сервисами органов государственной власти уровень достоверности результатов проверки	Паспорт	Упрощенная идентификация (УПРИД) клиента, проверка действительности паспорта	Необходимо
2) Проверка сведений, предоставленные получателем финансовой услуги для получения онлайн-микрозайма, на их соответствие информации, содержащейся в кредитной истории получателя финансовой услуги	Кредитная история	Проверки на основе антифрод-сервиса БКИ	Желательно
		Анализ КИ на предмет аномалий	
3) Проверка принадлежности получателю финансовой услуги банковского счета и (или) проведение анализа использования платежного инструмента, на которые предполагается зачисление суммы онлайн-микрозайма, на предмет возможности их использования для неправомерного получения денежных средств третьими лицам	Банковский счет, карта	Проверка принадлежности карты заемщику	Необходимо
		Или проверка принадлежности счета заемщику	
4) Проверка наличия доступа получателя финансовой услуги к абонентскому номеру подвижной радиотелефонной связи и (или) проведение анализа использования абонентского номера подвижной радиотелефонной связи, информация о котором предоставлена заемщиком для получения онлайн-микрозайма, на предмет возможности его использования для неправомерного получения денежных средств третьими лицам	Телефонный номер	Проверка принадлежности номера заемщику	Необходимо
		Проверка активности телефонного номера	
		Проверка «срока жизни» телефонного номера	
		Анализ геолокации	
5) Проведение анализа использования устройства, подключенного к сети «Интернет» и применяемого при получении онлайн-микрозайма, на предмет	Устройство пользователя	Проверка аномальности технических параметров устройства	Необходимо

Пункт 5.8 Базового стандарта	Предмет проверки	Виды проверок	Рекомендации по использованию
возможности его использования для неправомерного получения денежных средств третьими лицами, в том числе анализ сетевого подключения и системной информации программного обеспечения, установленного на устройстве			
6) Проверка соответствия внешности физического лица-получателя финансовой услуги внешности физического лица, персональные данные которого предоставлены для получения онлайн-микрозайма, в том числе путем анализа фотографии физического лица-получателя финансовой услуги и фотографии в документе, удостоверяющем личность, а также организации взаимодействия с физическим лицом-получателем финансовой услуги с использованием средств видеосвязи	Внешность	Ручная сверка фотографии из заявки с фотографией в паспорте	Не обязательно
		Автоматическая идентификация по видеоконференции	
7) Получение ответов на контрольные вопросы	Ответы на контрольные вопросы		Не обязательно
8) Проверка наличия доступа получателя финансовой услуги к адресу электронной почты и (или) провести анализ использования адреса электронной почты, информация о котором представлена для получения онлайн-микрозайма, на предмет возможности его использования для неправомерного получения денежных средств третьими лицами	Адрес электронной почты		Не обязательно
9) Проверка принадлежности нескольких предоставленных для получения онлайн-микрозайма данных (в том числе идентификационных, контактных и иных данных) получателю финансовой услуги, с использованием сервисов органов государственной власти или иных сервисов, обеспечивающих сопоставимый с сервисами органов	Относимость разных данных одному лицу		Не обязательно

Пункт 5.8 Базового стандарта	Предмет проверки	Виды проверок	Рекомендации по использованию
государственной власти уровень достоверности результатов проверки			
10) Проверка сведений, предоставленных получателем финансовой услуги для получения онлайн-микрозайма, на их соответствие общедоступным данным о получателе финансовой услуги	Общедоступная информация	Проверка наличия активности заемщика в сервисах в интернете	Желательно

Источник: составлено автором

В практике МФО при онлайн кредитовании выдача первичного займа обычно ограничена лимитом 15 000 руб. В этом случае проверка действительности документа согласно пункту 5.8.1 базового стандарта в соответствии с положениями 115-ФЗ может быть при помощи упрощенной идентификации (УПРИД).

Проверка УПРИД реализуется посредством интеграции с поставщиком услуги (банк или специального провайдера), который далее по системе межведомственного взаимодействия (СМЭВ) производит упрощенную идентификацию. Для этого проводится сопоставление соответствия фамилии, имени и отчества, даты рождения, пола клиента с одним из документов СНИЛС или ИНН.

По закону заемщик должен указать номер ИНН и СНИЛС, но по в реальной практике достаточно указать один из документов в анкете-заявке. По СМЭВ поставщик услуги УПРИД получает ИНН из ФНС по номеру паспорта и СНИЛС, а СНИЛС может получить из ПФР по номеру паспорта и ИНН. УПРИД автоматически включает в себя проверку действительности паспорта.

Для проверки в рамках пункта 5.8.2 базового стандарта можно использовать готовые платные решения от БКИ. Как правило, все бюро имеют свои сервисы антифрод-проверок, результатом которых является скоринговый балл, оценивающий вероятность мошеннической заявки. Также МФО может

самостоятельно разработать правила на основе анализа кредитных отчетов на предмет аномалий кредитного поведения потенциального заемщика. Например, аномальное количество запросов по займам в последние дни, при отсутствии таковых ранее; несовпадение адресов клиента с анкеты с адресами в кредитном отчете; количество оформленных микрокредитов по данным БКИ (флаг отсутствия таковых); разница между датой заявки и последней датой появления указанного в анкете телефона заемщика в кредитной истории; количество уникальных телефонных номеров заемщика в БКИ и частота их появления и др. Следует отметить, что просто сам по себе запрос кредитного отчета и сверка анкетных данных заемщика с титульной частью кредитного отчета не является исполнением требований базового стандарта.

Касаемо пункта 5.8.3 Базового стандарта отметим, что процедура 3Ds-привязки карты¹⁾ для рекуррентных списаний по карте, проверка оператором фотография карты, ручные проверки при попытке перевести деньги на карту клиента через мобильные приложения банков или системе быстрых платежей (СБП), не является исполнением требований базового стандарта, т. к. не позволяют видеть полные ФИО потенциального заемщика. Только технология платежей по реквизитам и по СБП позволяет сверить полные ФИО получателя средств.

Многие МФО осуществляют выдачу займов только через СБП, тем самым выполняя пункт 5.8.3 требований базового стандарта и реально снижая риски мошенничества. Однако такое бизнес-решение требует соответствующей технической интеграции по подключению МФО к СБП. В рамках такой интеграции на рынке есть решения, позволяющие провести «привязку» счета клиента, а не только одной карты. В этом случае МФО видит ФИО владельца счета и может идентифицировать клиента, также привязка счета позволяет в дальнейшем в случае просрочки списывать средства рекуррентными платежами

¹⁾ 3DS (3-D Secure) – это технология двухфакторной аутентификации для обеспечения безопасности онлайн-покупок по картам. Суть технологии в следующем: после ввода реквизитов карты (номера, срока действия, защитного кода с обратной стороны карты) пользователь перенаправляется на страницу подтверждения операции. Банк, выпустивший карту, запрашивает дополнительные данные для подтверждения личности. Чаще всего это одноразовый код, который отправляется в SMS-сообщении на номер телефона, привязанный к карте.

со всех карт клиента, привязанных к счету, что повышает собираемость сборов в случае просрочки.

Анализ принадлежности карты заемщику также может быть проведен по значению скорингового балла и переменных специализированных сервисов на рынке. Особенностью проверки банковского счета или карты является то, что для ее осуществления необходимо знать реквизиты клиента, которые обычно запрашиваются уже на последнем шаге заполнения заявки на кредит в случае одобрения займа.

Проверка телефонного номера может быть произведена на основе разных источников данных на рынке по прямым и косвенным признакам принадлежности телефонного номера мошеннику. Основные направления проверок: проверка использования телефонного номера потенциальным заемщиком; оценка активности телефонного номера; анализ давности регистрации сим-карты; анализ наличия геолокации заемщика по указанному в анкете месту жительства или регистрации. Информация для этих проверок доступна на коммерческой основе у операторов сотовой связи и специализированных сервисов-агрегаторов на рынке.

Необходимо отметить, что смс код не устанавливает доступ конкретного лица к номеру телефона и может рассматриваться только как дополнительный фактор проверки, авторизация через портал госуслуг (ЕСИА) без цифровой подписи не позволяет проверить доступ конкретного лица к номеру телефона.

При анализе использования устройства заемщика недостаточно просто установить параметры устройства, интернет-соединения и прочие технические характеристики. Необходимо использование критериев аномальности этих значений в виде правил или готовых скоринговых баллов специализированных сервисов, которые присутствуют на рынке.

Проверка внешности заемщика может осуществляться оператором при помощи ручного анализа фотографии заемщика с паспортом («селфи с паспортом»), либо происходить в автоматическом режиме с помощью специализированных сервисов, способных проверять наличие внесения изменений в фотографию или использующих видеоконференции. Данная

проверка очень сильно снижает конверсию из анкеты в заявку на этапе оформления займа и при использовании специализированных сервисов создает дополнительные затраты, поэтому используется не всеми МФО.

Использование контрольных вопросов для идентификации сложно в реализации, при этом нельзя однозначно сказать об эффективности такого метода. Вопросы не должны быть простыми, должны касаться разных тематик, важно анализировать и способ предоставления ответов, регулярно необходимо менять перечень контрольных вопросов, а доступ к перечню вопросов должен быть ограничен в МФО.

Проверка наличия доступа к электронной почте малоэффективна и редко используется на рынке МФО, также, как и проверка согласно пункту 5.8.9, который был включен в базовый стандарт по рекомендации СРО, но по факту не используется на практике и ЦБ РФ планирует ее убрать из стандарта.

Суть анализа общедоступной информации заключается в проверке наличия регистрации пользователя в социальных сетях, крупных сервисах в интернете по продаже услуг и товаров (например, «Авито», «ЦИАН»), программах лояльности крупных торговых сетей и т. п. На рынке компании агрегаторы данных предоставляют информацию о более чем 30 таких сервисах, и, если не в одном из них нет регистрации заемщика, это является маркером высокого кредитного риска и риска мошенничества.

Выбор конкретного вида проверки, видов данных для ее осуществления и провайдера данных может осуществляться в каждой МФО по-разному в зависимости от особенностей бизнес-процессов и стратегии оценки кредитных рисков. В таблице 2.3 представлен один из модальных на рынке перечней минимальных проверок оптимальных по трем критериям: соблюдение федерального законодательства, идентификация клиента и снижение уровня мошеннических заявок, возможность использования данных для оценки кредитных рисков.

2.2 Особенности применения показателей ценности информации для оценки кредитных рисков микрофинансовых организаций

Как было отмечено в параграфе 3 главы 1, отсутствие в научной среде общих методик измерения ценности данных, учитывающих специфику оценки кредитных рисков МФО обосновывает актуальность рассмотрения методов выбора рационального состава данных для оценки кредитных рисков. Такие данные на рынке предоставляют МФО множество провайдеров на коммерческой основе.

Рассмотрим более подробно некоторые показатели для оценки ценности данных, подробно представленные автором в работе [101].

В рамках настоящего диссертационного исследования под данными будем понимать значения конкретных переменных, полученных от конкретного источника данных на платной или бесплатной основе с целью оценки кредитных рисков конкретного заемщика при рассмотрении его заявки. Выбор рационального состава данных для оценки кредитных рисков обосновывается при помощи проведения нескольких видов анализа, расчета статистических тестов и метрик влияния на показатели кредитного риска.

Для окончательного выбора вида данных для оценки кредитных рисков МФО необходимо проводить их системную оценку по различным критериям [106]. На выбор конкретных шагов стратегии использования для оценки кредитных рисков МФО новых данных влияет возможность проведения ретро-теста на исторических данных. В практике финансовых организаций при тестировании нового источника данных широко распространено ретро-тестирование, тестирование на «боевых» данных, А/Б тестирование.

Можно выделить четыре основных последовательных этапа (при наличии возможности проведения ретро-тестирования) стратегии ввода нового источника данных на практике в МФО (см. таблицу 2.4).

Таблица 2.4 – Основные шаги стратегии тестирования ценности данных

Шаг	Описание	Критерий основной оценки
1	Проверка источника данных на ретро-выборке	Полнота данных, наличие влияния на целевой показатель риска и оценка силы влияния
2	Запрос «боевых» данных без принятия решения об их использовании и их анализ после «вызревания» рисков	Классификационная способность моделей с новым источником данных
3	Запрос «боевых» данных на части потока клиентов с принятием решения об их использовании (A/B тест)	Бизнес метрики МФО: показатели риска и конверсии в одобрение
4	Переход на использование данных на всем потоке клиентов	Бизнес метрики МФО: доходность

Источник: составлено автором

Очевидно, ценность информации должна быть количественно измерима, поэтому в процессе тестирования нового источника данных риск-аналитики ориентируются на значения множества метрик, которые достаточно широко используются в практике оценки и управления кредитными рисками (таблица 2.5).

Таблица 2.5 – Основные критерии и показатели для оценки ценности данных

Критерий	Показатели
а) Полнота данных	Hit rate – доля покрытия клиентской базы
б) Наличие влияния на целевую (бинарную) переменную риска	Точный тест Фишера и тест независимости хи-квадрат Пирсона
в) Сила влияния на целевую переменную риска	Information value (IV) – информационное значение
	$H(x)$ – энтропия
	$H(Y X)$ – условная энтропия
	$IG(Y X)$ – прирост информации
г) Оценка экономического эффекта от нового источника данных	Cost to Income Ratio (CIR) – отношение расходов на запрашиваемые данные к доходам от их применения
	Lifetime profit (LTP) – ожидаемая прибыль на протяжении жизни кредитов
	Return on Investment (ROI) – доходность

Критерий	Показатели
	Decision Cost (DC) – стоимость принятия одного решения
д) Оценка общего экономического эффекта	Return on Investment (ROI) – доходность
	Approval Rate (AR) – конверсия из заявки в выдачу
	Non-Performing Loan (NPL), First Payment Default (FPD) – показатели уровня риска
	Стоимость затрат на 1 выдачу (в шт. и на 1 руб.)
е) Затраты на интеграцию и внедрение нового источника данных	Оценка прямых и косвенных затрат, руб.
ж) Репутация компании и ее надежность	Косвенная экспертная оценка потерь в случае сбоя получения данных
	Оценка вероятности отказа работы сервиса
з) Общая классификационная способность	Accuracy – общая точность
	Precision – прецизионность
	Sensitivity – чувствительность
	Specificity – специфичность
	F-score – F-мера
	Графики AUC
	Площадь под кривой AUC
	Gini – коэффициент Джини
	Log loss – логистическая функция потерь
	Distribution of Points – распределение баллов
	Kolmogorov-Smirnov statistics – статистика Колмогорова-Смирнова
	Divergence Coefficient – коэффициент дивергенции
	Separation Coefficient – коэффициент разделения
	UpLift – связанный прирост качества классификации

Источник: составлено автором

Представленный в таблице 2.5 перечень метрик не является исчерпывающим, но в 95% случаев практических кейсов метрик из него достаточно для принятия риск-аналитиком решения о ценности данных. В таблице 2.5 общепринятые в риск-менеджменте метрики качества моделей и бизнес-метрики приведены в рекомендуемой последовательности их применения при тестировании нового источника данных.

Ввиду того, что оценка ценности данных не является самоцелью кредитной организации, необходимо выделить несколько общих бизнес-целей внедрения новых источников данных для оценки кредитных рисков МФО, исходя из предложенных автором критериев классификации данных [101], и указать наиболее важные для измерения их достижения критерии (таблица 2.6).

Как можно заметить, выделенные цели являются комплексными и подразумевают под собой решение множества задач, с которыми подразделения риск-менеджмента в МФО сталкиваются при оценке кредитных рисков.

После выбора поставщика данных, соответствующего требованиям МФО и особенностям решаемой задачи, для оценки ценности предоставляемых данных проводится историческое тестирование (ретро-, бэк-тестирование). В рамках ретро-теста кредитная организация отправляет потенциальному для интеграции источнику запрос с выборкой клиентов, по которой может получить в ответ либо весь набор исходных данных, из которого силами штатных риск-аналитиков составит переменные для скоринга, либо уже обработанные данные, включающие готовые скоринги и переменные, полученные путем агрегирования исходных данных.

Таблица 2.6 – Возможные цели использования информации в системе оценки кредитных рисков МФО

Цели	Критерии, на которых следует сделать акцент при анализе
1) Уменьшение общего уровня кредитных потерь	Рисковые показатели
2) Улучшение действующей модели оценки кредитных рисков	Полнота данных, наличие влияния на целевой показатель и его сила, классификационная способность
3) Уменьшение затрат на принятие одного решения	Оценка эффектов от нового источника данных, затраты на внедрение
4) Увеличение конверсии из заявки в выдачу	Показатели конверсии в выдачу

Источник: составлено автором

Стоит отметить, что наибольший интерес с точки зрения анализа представляют необработанные данные, так как есть возможность их самостоятельной обработки и область применения такой информации обширнее. Однако не стоит недооценивать обработанные данные, поскольку они могут быть полезны в случаях, когда необходимо встроить в модели дополнительную оценку скорингового балла заемщика. Компании, предоставляющие обработанные данные, могут располагать большими информационными и технологическими ресурсами, что также повышает ценность таких данных. Поэтому хорошим решением для МФО может быть комбинирование обоих типов данных и определение четких требований к обработке предоставляемых извне готовых данных для их дальнейшего использования.

Получив от внешнего источника данные результатов проведенного ретро-теста, МФО может переходить к процессу оценки их ценности. Особенности применения показателей ценности информации для оценки кредитных рисков МФО рассмотрены в следующем параграфе настоящего диссертационного исследования.

а) *Метрики полноты данных.* Прежде всего, по результатам ретро-тестирования производится оценка успешности запросов: по каждой полученной переменной вычисляется процент пропущенных значений или доля покрытия (Hit Rate). Доля неуспешных запросов по новому источнику данных не должна быть ниже заранее определенного в МФО минимального порогового значения, например, 70%.

Наибольший интерес представляют источники данных, которые включают информацию обо всех запрашиваемых клиентах. Чем больше доля ответов, тем больше областей применения полученной информации. Например, валидность внешнего скорингового балла, получаемого по запросу от бюро БКИ в рамках кредитных историй заемщиков, на всем потоке клиентов составляет от 95% до 100%, то есть данная переменная является валидной практически по всей выборке и потенциально может использоваться как предиктор для моделирования

вероятности дефолта по основному потоку заявок и может быть включена в соответствующую модель в СПР.

Однако существуют источники данных, которые показывают низкий уровень ответа на всем потоке клиентов, например 70%, но при этом полностью охватывают отдельные его сегменты. Такие данные рационально применять для оценки кредитных рисков отдельных сегментов с высокой долей покрытия. Обычно они используются в сегментированных моделях в рамках СПР: например, в модели, построенной по сегменту заемщиков, которые в основном берут кредиты в банках, а не в МФО. Еще один подход – использование новых данных для оценки кредитных рисков на «серой» зоне или на «отказном» сегменте по основной СПР. Если же анализируемая переменная нового источника данных имеет очень низкий уровень валидности, например 20%, такая переменная может быть переведена в бинарную шкалу, где значение 1 данной переменной будет показывать отсутствие значения характеризуемого ею показателя для данного заемщика, а значение 0 – факт наличия значения показателя. Отсутствие информации по заемщику в новом источнике данных может быть маркером повышенного риска.

б) *Метрики наличия влияния на целевую переменную риска.* В случае успешной валидности тестируемой переменной на всей выборке или на отдельном сегменте сначала оценивают ее общую предсказательную способность по выбранному риск-аналитиком целевому показателю. В качестве показателя кредитного риска, как было отмечено в главе 1, для МФО могут выступать такие показатели, как просрочка по кредиту на определенный срок (NPL) или показатель просрочки первого платежа (FPD), рассчитываемые обычно на горизонте 10, 30 и 90 дней. В классической задаче оценки кредитных рисков принято оперировать категориальными независимыми переменными для построения моделей. Первоначальный отбор информации для оценки кредитных рисков может строиться на основе анализа таблиц сопряженности, в которых по строкам отображается бинарная зависимая переменная показателя риска, а по столбцам независимая категориальная переменная (номинальная или порядковая),

либо группировка по количественной переменной, т. е. количественные независимые переменные категоризуются. В качестве базового метода оценки предсказательной способности категориальной переменной применяют критерии независимости хи-квадрат Пирсона или точный тест Фишера [236] в случае таблиц сопряженности размерностью 2×2 . Методы анализа категориальных данных, на основе которых автором предложены метрики влияния на целевую переменную риска, подробно рассмотрены в работах [153, 154, 180, 187].

Критерий независимости хи-квадрат Пирсона – непараметрический тест, который позволяет определить наличие значимой связи между категориальными переменными. Данный критерий используют для анализа стандартных таблиц сопряженности, имеющих следующий вид (таблица 2.7).

Таблица 2.7 – Стандартная таблица сопряженности

Признаки	Y_1	Y_2	...	Y_c	C_j
X_1	f_{11}	f_{12}	...	f_{1c}	C_1
X_2	f_{21}	f_{22}	...	f_{2c}	C_2
...	...	...	...	...	...
X_r	f_{r1}	f_{r2}	...	f_{rc}	C_r
R_i	R_1	R_2	...	R_c	N

Источник: составлено автором по [202]

Критерий независимости хи-квадрат Пирсона вычисляется по формуле [331]:

$$\chi_P^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(f_{ij} - e_{ij})^2}{e_{ij}}, \quad (2.1)$$

где f_{ij} – заданные (фактические) частоты в ячейках таблицы сопряженности;

e_{ij} – ожидаемые частоты в ячейках таблицы сопряженности, которые вычисляются по формуле:

$$e_{ij} = \frac{R_i \cdot C_j}{N}, \quad (2.2)$$

где $R_i = \sum_{j=1}^c f_{ij}$ – суммы строк таблицы сопряженности;

$C_j = \sum_{i=1}^r f_{ij}$ – суммы столбцов таблицы сопряженности;

$N = \sum_{j=1}^r R_i = \sum_{j=1}^c C_j$ – общее количество наблюдений.

Статистика критерия независимости хи-квадрат Пирсона подчиняется распределению χ^2 со степенью свободы, равной $(r - 1)(c - 1)$. В рамках оценки кредитных рисков микрофинансовых организаций при использовании критерия хи-квадрат выдвигают нулевую гипотезу H_0 об одинаковом распределении «плохих» и «хороших» заемщиков по категориям независимой переменной.

Альтернативой формулы (2.1) является расчет критерия хи-квадрат на основе формулы логарифма правдоподобия:

$$\chi_{LR}^2 = 2 \sum_{i=1}^m \sum_{j=1}^k f_{ij} \ln \left(\frac{f_{ij}}{e_{ij}} \right). \quad (2.3)$$

При достаточном числе наблюдений значение по альтернативной формуле (2.3) будет мало отличаться от значения по классической формуле (2.1).

В случае таблиц сопряженности размерностью 2×2 (пример представлен в таблице 2.8) применяют точный тест Фишера [193, 236].

Таблица 2.8 – Таблица сопряженности размерностью 2×2

Признаки	Y_1	Y_2	C
X_1	f_1	f_2	$f_1 + f_2$
X_2	f_3	f_4	$f_3 + f_4$
R	$f_1 + f_3$	$f_2 + f_4$	N

Источник: составлено автором по [236]

В соответствии с точным тестом Фишера вероятность получения достоверных данных при условии независимости строк и столбцов, а также маргинальности итоговых значений в таблице сопряженности, определяется гипергеометрическим распределением, которое рассчитывается по формуле [9]:

$$p = \text{Prob}(N_1 = f_1) = \frac{(f_1 + f_2)! \cdot (f_3 + f_4)! \cdot (f_1 + f_3)! \cdot (f_2 + f_4)!}{N! \cdot f_1! \cdot f_2! \cdot f_3! \cdot f_4!}, \quad (2.4)$$

где показатель N_1 ранжируется от максимального значения $(0, f_1 - f_4)$ до минимального значения $(f_1 + f_2, f_1 + f_3)$, а $N = f_1 + f_2 + f_3 + f_4$.

Если связь между анализируемой переменной и целевым показателем окажется статистически значимой, можно переходить к дальнейшему анализу информативности переменной, воспользовавшись такими широко применяемыми критериями, как информационное значение (IV) или мера энтропии для оценки предикторов.

Предварительный анализ взаимосвязи на основе таблиц сопряженности помогает ограничить количество рассматриваемых для построения модели оценки кредитных рисков переменных.

в) *Метрики силы влияния на целевую переменную риска.* В оценке кредитных рисков с целью отбора переменных для построения скоринговых моделей одним из основных показателей для ранжирования переменных по силе влияния на кредитный риск является показатель «информативное значение» (англ. Informative Value, IV). Следовательно, остановимся на алгоритме его расчета подробно. Информационное значение рассчитывается на основе весов (WOE) по каждой категории переменной [109] по следующей формуле:

$$WOE_i = \ln \left(\frac{d_i^{(1)}}{d_i^{(2)}} \right), \quad (2.5)$$

где $d_i^{(1)}$ и $d_i^{(2)}$ – относительные частоты «плохих» и «хороших» займов в i -ой группе соответственно (относительно количества всех «плохих» и всех «хороших» займов);

$$i = 1, 2, \dots, k;$$

k – число категорий переменной.

Данный показатель применяется для построения строгой линейной зависимости вероятности дефолта заемщика с логарифмом отношения шансов. При сравнении «плохих» заемщиков с «хорошими» положительное значение WOE говорит о том, что доля «плохих» кредитов для данной категории выше доли «хороших» кредитов.

После группировки значений тестируемой переменной на категории и расчета для них WOE вычисляется показатель информационного значения для переменной в целом:

$$IV = \sum_{i=1}^k (d_i^{(1)} - d_i^{(2)}) \cdot WOE_i. \quad (2.6)$$

Чем больше показатель IV, тем больший вес имеет переменная с точки зрения полезности для будущей модели оценки кредитных рисков (см. таблицу 2.9).

Для оценки кредитных рисков обычно отбираются переменные только с информационным значением выше среднего (или энтропией ниже среднего) в сравнении с уже используемым переменным в текущих скоринговых моделях. Переменные со значением показателя IV, большим 0,5, требуют дополнительной проверки. Обычно такие переменные используются для сегментации и в дальнейшем оценка кредитного риска идет по отдельным сегментам.

Таблица 2.9 – Критерии отбора переменных по информационному значению

IV	Предсказательная способность переменной
< 0,02	Неинформативная переменная
[0,02; 0,1)	Слабая предсказательная способность
[0,1; 0,3)	Средняя предсказательная способность
[0,3; 0,5]	Сильная предсказательная способность
> 0,5	Не является информативной сама по себе, но может заменять полезную, но не наблюдаемую переменную (является прокси-переменной)

Источник: составлено автором

Второй возможный показатель силы влияния переменной на кредитный риск – энтропия (или мера неопределенности), представляет собой оценку вероятностного распределения событий по прогнозным и фактическим данным [33]. Чем больше значение энтропии, тем более случайны результаты модели оценки кредитных рисков на применяемых данных. Энтропия рассчитывается как

сумма произведений вероятности события на его логарифм с отрицательным знаком [33].

$$H(x) = -\sum_{i=1}^M p(x_i) \log_2 p(x_i), \quad (2.7)$$

где $p(x_i)$ – вероятность исхода x_i ;

x_i – разница между прогнозным и фактическим значением целевой переменной в i -й группе.

Можно сравнить энтропию с включением изучаемой переменной и без нее, рассчитав показатель прироста информации (Information Gain, IG) [329]. Для этого можно рассчитать условную энтропию (Conditional Entropy):

$$H(Y|X) = \sum_{x_i \in X} p(x_i) \cdot H(Y|X = x_i). \quad (2.8)$$

Значение показателя прироста информации вычисляется как разница между мерами неопределенности предсказания целевой переменной риска до включения анализируемого источника данных и после его включения:

$$IG(Y|X) = H(Y) - H(Y|X). \quad (2.9)$$

Таким образом, можно оценить силу связи исследуемой переменной с целевой переменной. Чем больше значение IG, тем сильнее связь между переменными и тем ценнее исследуемая переменная будет для оценки кредитных рисков. Если значение IG равно 0 или близко к этому значению, это означает, что связь между анализируемой переменной и целевым показателем либо вообще отсутствует, либо является слабой.

г) *Метрики оценки экономической эффективности.* Экономическую эффективность и техническую стоимость внедрения нового источника информации стоит рассматривать только в том случае, если по показателям информационной значимости переменных и метрикам оценки качества моделей были продемонстрированы хорошие результаты. Для сравнения нескольких моделей, использующих различные информационные данные для оценки кредитных рисков МФО, можно провести А/В тестирование [52]. Такой подход позволит на различных сегментах заемщиков оценить, как скоринговая модель с новым источником данных влияет на бизнес-метрики и экономические

показатели МФО. При проведении тестирования сравниваются две группы: экспериментальная группа – заемщики, для оценки которых используется СПР с новой информацией, и контрольная группа – заемщики, для оценки которых используется прежняя СПР. Такая стратегия поможет понять, какие бизнес-метрики улучшились после внедрения новых данных. В частности, следует отметить, что хорошим результатом применения новых данных для МФО является увеличение конверсии из заявки в выдачу (англ. Approval Rate, AR) как минимум при сохранении уровня кредитного риска (NPL, FPD).

Следует учитывать, что анализируемая переменная нового источника данных может включаться в СПР на разных этапах процесса оценки заемщика. Поэтому, кроме оценки прироста предсказательной способности модели оценки кредитных рисков, необходимо также анализировать затраты на запросы данных по заемщику, которые на каждом шаге воронки (кредитного конвейера) в рамках СПР могут различаться. Очевидно, что если данные запрашиваются из внешних источников на платной основе, то при включении новой переменной средние затраты на одобрение займа или кредита возрастут. При этом, если средние затраты на одобрение (на одну выдачу) станут превышать средний доход от кредита, использовать данную переменную нецелесообразно, поскольку МФО стремятся увеличить доходность своего кредитного портфеля.

Для проверяемой переменной можно посчитать отношение расходов на запрашиваемые данные к доходам от их применения (англ. Cost to Income Ratio, CIR). Этот показатель широко применяется в банковской сфере и показывает умение банка управлять своими операционными расходами и доходами [309]:

$$CIR = \frac{C}{I}, \quad (2.10)$$

где C – расходы на новые данные;

I – доходы от применения этих данных.

Значение показателя CIR не должно превышать 1. В целом, чем ниже показатель CIR, тем лучше для МФО.

Целесообразно также рассчитать прогнозную прибыль от применения данных на определенный промежуток времени (например, на год) на всем сроке жизни (англ. Life Time Profit, LTP) кредитов как разность между доходами и расходами на внедряемые данные:

$$LTP = LTV - C, \quad (2.11)$$

Где LTV (Life Time Value) – ожидаемый доход (за год) на протяжении жизни кредитов от внедрения источника, ден. ед.

Прибыль за все время может быть полезной метрикой с учетом всего срока жизни клиента и с учетом всех последующих повторных займов клиента МФО. На коротком временном горизонте можно рассмотреть другие метрики, например отдачу от инвестиций (Return On Investment, ROI). Показатель ROI также позволяет оценить ожидаемую прибыль от внедрения новых данных:

$$ROI = \frac{I-C}{C} * 100, \quad (2.12)$$

где I – доходы от применения новых данных, ден. ед.

Если ROI является положительным, то использование новых данных может быть целесообразным.

Чтобы принять решение о ценности данных того или иного источника, целесообразно оценить стоимость принятия одного решения (англ. Decision Cost, DC) относительно уровня одобрения. Тогда затраты следует нормировать на уровень одобрения:

$$DC = \frac{\sum_{i=1}^N P_i + F + A}{AR}, \quad (2.13)$$

где P – стоимость получения данных, ден. ед.;

F – фонд оплаты труда верификаторов, риск-менеджеров, аналитиков, ден. ед.;

A – стоимость аллокации информационных систем для хранения данных, ден. ед.;

AR – процент одобренных кредитных заявок, %.

Сравнив значения показателя DC до и после внедрения новых данных, можно сделать вывод об их ценности в рамках решаемой задачи МФО.

Уменьшение значения показателя свидетельствует о сокращении затрат относительно нового уровня одобрения.

д) *Метрики оценки общего экономического эффекта.* Метрики общей экономической эффективности связаны с метриками экономического эффекта от нового источника данных. Аналогично рассчитывают общую доходность (ROI) по всему портфелю, а также общую конверсию в одобрение (AR).

Важной задачей риск-менеджмента в МФО является повышение уровня одобрения (AR). Повышение конверсии из заявки в одобрение означает повышение количества одобренных заявок по отношению к поданным заявкам на получение займа или кредита. Данное отношение находится по формуле (в долях или процентах):

$$AR = \frac{n_A}{n_T}, \quad (2.14)$$

где n_A – количество одобренных заявок;

n_T – общее количество поступивших заявок.

Как было отмечено ранее, повышение AR непосредственно связано с внедрением новых переменных в действующие скоринговые модели. Одним из способов повышения конверсии из заявки в выдачу является построение скоринговых моделей для таких сегментов заемщиков, на которых предиктивная способность действующих в настоящий момент моделей является низкой. Например, отдельной областью для исследования является сегмент заемщиков, получающих отказ в рамках действующих СПР. Такой сегмент заемщиков требует детального анализа с применением статистических процедур кластеризации, которые могут помочь определить устойчивые профили отдельных кластеров внутри отказного сегмента, то есть получить правила определения заемщиков, принадлежащих однородным группам. Точкой приложения новых данных могут стать модели «дополнительного одобрения», построенные для выделенных групп заемщиков. Модели «дополнительного одобрения» в кредитном скоринге предназначены для принятия решений о выдаче кредитов клиентам, которые, с точки зрения основной модели скоринга, могут

быть недостаточно надежными для получения кредита (так называемая «серая» зона). Эти модели обычно используются банками и микрофинансовыми организациями для увеличения объема выданных кредитов, не увеличивая при этом риски по портфелю.

С точки зрения задачи увеличения уровня одобрения заявок в МФО судить о ценности новых данных можно по факту увеличения конверсии в выдачу и доходности. Действительно, ценные данные помогают больше зарабатывать, а не с большей точностью отказывать клиентам. Это означает, что МФО выгоднее выдать больше кредитов хорошим заемщикам, а не снижать риск до минимума и выдавать кредиты только лучшим заемщикам.

Общий экономический эффект от использования новых данных можно оценить по снижению показателей целевых метрик кредитного риска (NPL, FPD), либо, по крайней мере, их неизменности при росте других показателей эффективности.

е) *Оценка затрат на интеграцию источника данных.* Помимо анализа вышерассмотренных показателей МФО должна учитывать затраты на техническую интеграцию (прямые и косвенные) в денежном и временном выражении при оценке ценности рассматриваемого источника данных.

ж) *Оценка репутации компании-поставщика данных.* Безусловно, при выборе источника данных следует учитывать репутационные риски, связанные с легальностью их получения и правомерностью их использования. В любой сфере бизнеса репутация компании имеет большое значение при принятии решения о сотрудничестве с ней. С точки зрения репутационных рисков, следует учитывать время существования компании провайдера, предоставляющей данные, отзывы ее клиентов, специфику самих данных и т. д. При получении информации о заемщике возникает также вопрос о правомерности ее использования. С юридической точки зрения должен быть урегулирован вопрос передачи и сбора предоставляемых данных. При необходимости клиентом должны быть подписаны разрешения, позволяющие передавать информацию о нем, в соответствии со статьей 6 ФЗ № 152 «О персональных данных» [66]. Кроме того, следует также

понимать финансовые и временные затраты на интеграцию сторонних сервисов для получения данных, которые микрофинансовые организации стремятся минимизировать.

з) *Метрики общей классификационной способности.* После того, как предсказательная способность и информационная ценность анализируемой переменной были оценены с использованием описанных выше метрик, можно перейти к анализу ценности данных – изменения значений показателей качества действующей модели оценки кредитных рисков в МФО, вызванных введением новой переменной или перестройкой модели на множестве новых переменных. Особенности алгоритмов расчета и применения данных метрик связаны непосредственно с процессом построения моделей оценки кредитных рисков и рассмотрены в следующем параграфе данной главы.

Отдельно необходимо рассмотреть проблему мошеннических заявок в контексте задачи тестирования ценности новой информации для оценки кредитных рисков. Выявление мошеннических заявок (fraud), которое представляет собой присвоение определенному клиенту статуса (флага, индикатора) мошенника, исходя из данных о его поведении. Такие клиенты отсекаются и не используются при построении скоринговых моделей. По ним также нет смысла запрашивать дополнительные источники данных, например, кредитные истории.

Помимо присвоения отдельным заявкам мошеннического статуса, может вычисляться оценка риска, которая указывает на вероятность того, что заявка на получение займа или кредита является мошеннической: это позволяет проводить дополнительные шаги проверки личности заемщика приоритетно для заявок с высоким риском мошенничества.

В контексте задачи выявления мошенничества проверять на ценность можно данные, используемые как в начале кредитного конвейера, то есть до запроса по заявке данных из внешних источников и ее «прогона» по скоринговым моделям для оценки кредитного риска, так и ближе к ее концу, то есть уже после запроса внешних данных. До принятия решения о выдаче займа с целью

заблаговременного выявления любых аномалий рекомендуется проверять персональные данные, контактную информацию, анализировать данные из официальных государственных источников, поведение на сайте и данные устройства заемщика, его активность в социальных сетях или экономическую активность. На таких данных в СПР необходимо выстраивать отдельные модули с правилами против мошенников. Данные правила могут задавать пороговые значения вычисляемых параметров или проверять соответствие параметров истинному значению. Это помогает определить, были ли использованы для подачи заявки украденные учетные данные, были ли поданы несколько заявок с одного устройства или использован виртуальный номер телефона для обмана системы проверки личных данных и т. п. В случае если по заявке из БКИ была запрошена кредитная история, проверяется были ли у заемщика проблемы с выплатой задолженностей или обманом других кредиторов. В дальнейшем в случае одобрения заявки и выдачи займа могут анализироваться паттерны поведения заемщика и сравниваться с другими кредитными линиями, которые оказались мошенническими.

Следует отметить, что корректное отсеечение мошеннических заявок в начале кредитного конвейера позволяет экономить на дальнейших запросах по таким заемщикам данных, получаемых на коммерческой основе из внешних источников и используемых для скоринга заявок, а следовательно, уменьшает затраты на принятие одного решения.

Изменения рискованных метрик, вызванные введением новых антифрод правил, использующих новые данные, помогают понять их ценность. Улучшение точности отсеечения мошеннических заявок несомненно повлечет за собой улучшение качества данных, используемых для построения скоринговых моделей, за счет исключения из данных «шума». Когда качество данных становится выше, предиктивная точность построенной на них модели увеличивается, о чем может свидетельствовать улучшение метрик оценки качества модели, обсуждавшихся выше. Как следствие, рискованные метрики

улучшаются, повышается качество кредитного портфеля организации за счет снижения уровня просрочек.

Однако введение новых правил обнаружения мошеннических транзакций не всегда гарантирует улучшение качества модели. В некоторых случаях введение новых правил может привести к увеличению количества ложных срабатываний. Это может произойти в случае, если вводимые правила слишком жесткие или недостаточно точные. Введение таких правил может снизить качество используемых моделей. Кроме того, введение новых правил способно привести к увеличению времени рассмотрения заявок, поскольку необходимо будет проводить дополнительные проверки. Все это может выразиться в сильном падении конверсии заявки в одобрение и негативно повлиять на объемы кредитования и доходность компании. Так, например, адекватным является падение AR на 1–2 процентных пункта при отсечении 1% трафика в начале кредитного конвейера. Поэтому необходимо проводить тщательный анализ и оценку влияния изменений в антифрод-правилах на качество модели перед использованием новых данных.

Также необходимо отметить, что оценка показателей ценности новой информации для оценки кредитных рисков по представленной методике может проводиться в целом по портфелю, а также по сегментам клиентов. Например, заемщиков можно разделить на несколько сегментов по уровню риска. Распространенным подходом является разбиение заемщиков на десять групп по децилям риска и дальнейшая агрегация полученных групп в более крупные сегменты. Например, в сегменты с низким, средним и высоким риском дефолта. Примерами сегментов могут быть также следующие группировки заявок на заем: по числу оформленных займов в данной кредитной организации – первичные и повторные клиенты, по источнику трафика – онлайн и оффлайн заявки, комбинация сегментов по уровню риска, полученных на основе различных скоринговых моделей и др.

Таким образом, анализ ценности новых данных для оценки кредитных рисков представляет собой многошаговый процесс, требующий больших

временных затрат. Определение ценности новых данных играет важную роль в процессе оценки кредитного риска заемщика. Сама ценность данных связана с тем, насколько полезными они являются для достижения целей риск-моделирования и, как следствие, целей микрофинансовой организации. Ценность данных определяется не только их качеством, релевантностью, разнообразием, актуальностью, полнотой, доступностью, скоростью получения и легальностью, но также и их информационной ценностью в рамках конкретной решаемой задачи и влиянием на доходность кредитного портфеля микрофинансовой организации. Поэтому важно использовать эффективную методику оценки ценности данных для их использования в СПР, чтобы уменьшить риски и повысить качество кредитного портфеля.

Систематизация целевых показателей и метрик для измерения эффективности новых данных в системе оценки кредитных рисков в практической деятельности МФО, включая метрики силы влияния на показатели риска, метрики классификационной способности и метрики оценки экономического эффекта, а также представленная пошаговая методика измерения ценности и отбора наиболее подходящих данных в контексте решения задачи оценки кредитных рисков позволит фильтровать состав данных по уровню рациональности в условиях ограниченности ресурсов МФО.

После принятия окончательного решения об использовании новой информации в моделях оценки кредитных рисков или использования модели, построенной на абсолютно новых данных, оценка ее качества не заканчивается. Необходимо постоянно контролировать работоспособность модели оценки кредитных рисков с новым источником данных или всей СПР, чтобы убедиться в ее эффективности и способности улучшать бизнес-показатели МФО.

По проведенному исследованию можно заключить, что внедрение новых информационных ресурсов в СПР МФО для оценки кредитных рисков требует применения различных типов данных, как альтернативных, так и традиционных источников (данных кредитных отчетов от БКИ), которые обладают различной эффективностью. Как правило, проверенные временем источники наиболее

результативны. Однако за счет дополнительного использования альтернативных источников МФО имеет возможность уменьшить кредитные риски и повысить конверсию заявок в выдачу. Альтернативные данные могут быть использованы также для привлечения лояльных клиентов, которые принесут максимальную прибыль МФО на длительном временном интервале. Таким образом, новые источники данных становятся инструментом для повышения уровня одобряемых кредитных заявок, и, как следствие, расширения клиентской базы при одновременном снижении уровня кредитных потерь. При этом необходимо выбирать надежные источники информации для оценки кредитных рисков в МФО с использованием инструментов тестирования данных по различным критериям оценки.

2.3 Подходы к оценке кредитного риска с учетом качества бинарной классификации заемщиков микрофинансовых организаций

Финальным этапом моделирования кредитных рисков является оценка параметров выбранной модели и анализ ее классификационной способности. Особенности построения моделей от выбранной концепции (статистического моделирования или машинного обучения) будут рассмотрены в следующей главе, а в данном параграфе остановимся подробнее на метриках дискриминирующей способности выбранных источников данных или построенных на их основе моделей оценки кредитных рисков. Рассмотрим особенности алгоритмов расчета и применения данных метрик представленных в работах автора [98, 109].

Дискриминирующая способность модели оценивается на основе таблицы классификации, которая представляет собой таблицу сопряженности между фактическими и прогнозными значениями зависимой переменной размерностью 2×2 , пример которой приведен в таблице 2.10.

Результаты классификации по модели бинарной классификации иногда представляют графически в виде гистограммы оценок классификации. На рисунке

2.1 представлен пример такой гистограммы, выводимой в пакете IBM SPSS Statistics в модуле построения логистической регрессии.

Таблица 2.10 – Пример таблицы классификации

Прогноз дефолта по модели	Фактическое наличие дефолта		Процент корректных предсказаний
	Да	Нет	
Да	770	250	75,5%
Нет	224	1220	84,5%
Итого			80,8%

Источник: составлено автором

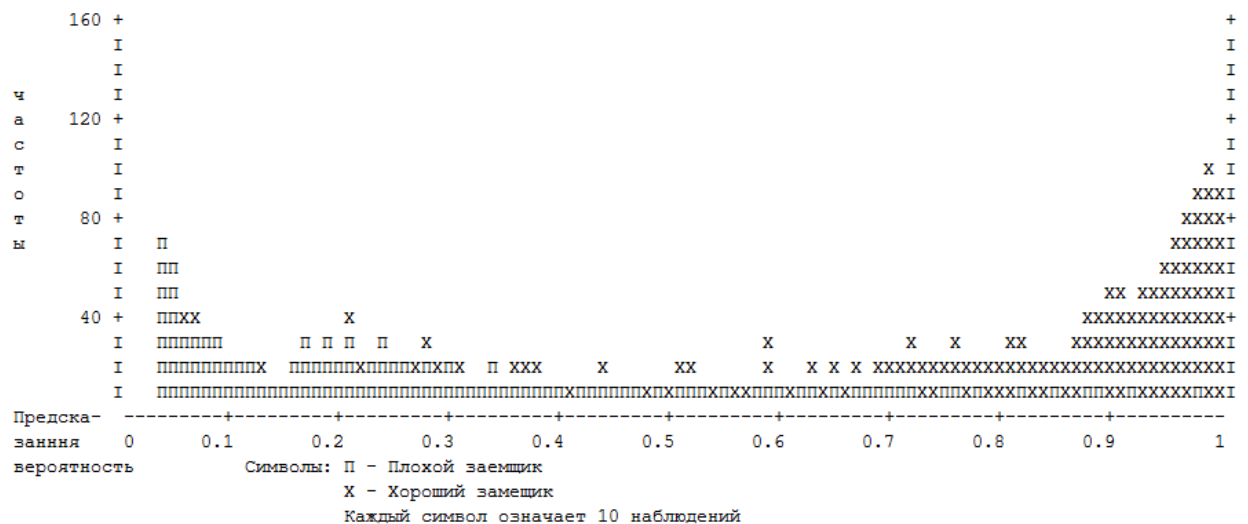


Рисунок 2.1 – Пример построения гистограммы классификации

Источник: построено автором с помощью статистического пакета IBM SPSS Statistics в модуле построения логистической регрессии

На рисунке 2.1 буквами обозначаются градации предсказанной зависимой переменной: «П» – «плохой» заемщик, если прогнозное и фактическое значение зависимой переменной не совпадают; «Х» – «хороший» заемщик, если прогнозное и фактическое значение зависимой переменной не совпадают. Каждый столбик на гистограмме соответствует определенной предсказанной вероятности, а его высота – количеству наблюдений, для которых предсказана данная вероятность. На гистограмме классификации букве «Х» в правой части, а букве «П» в левой

части соответствуют правильные предсказания. В случае идеальной классификации все буквы «Х» должны быть левее буквы «П», а разделять их будет вероятность отсечения «плохих» и «хороших» заемщиков. На представленной на рисунке 2.1 гистограмме вероятность отсечения равна 0.5. Чем лучше качество ранжирования заемщиков по уровню риска для тестируемого показателя, тем теснее наблюдения на гистограмме будут сгруппированы у соответствующих концов левой и правой оси.

В качестве зависимой переменной в модели оценки кредитных рисков микрофинансовых организаций применяется показатель дефолта заемщика, а его фактическое и прогнозное значение определяется параметрами «да» для «плохих» и «нет» для «хороших» заемщиков. Как правило, вместо суммирования показателей по строкам и столбцам в таблице сопряженности приводится процент корректных предсказаний по категориям зависимой переменной, а также общий процент всех предсказаний по модели.

Качество модели оценки кредитных рисков в рамках микрофинансовой организации определяется ее способностью корректно классифицировать заемщиков, то есть правильно определять, насколько надежными будут заемщики. Очевидно, что качество классификационной модели может существенно влиять на бизнес-метрики. Например, если модель недооценивает риски заемщиков, МФО может столкнуться с высоким процентом недоимок по кредитам и убытками от дефолтов, что негативно скажется на доходности кредитного портфеля. Существует множество метрик для оценки классификационной способности моделей оценки кредитных рисков.

Для измерения прогностической способности модели используются такие метрики, как достоверность (англ. Accuracy), точность (англ. Precision), полнота (англ. Recall), чувствительность (англ. Sensitivity), специфичность (англ. Specificity), F-мера, площадь под ROC-кривой (англ. Area Under Receiver Operating Characteristic Curve), коэффициент Джини (англ. Gini), связанный прирост качества классификации (англ. Uplift) и логистическая функция потерь (англ. Log loss). Последние пять показателей являются наиболее важными в

оценке общей точности модели оценки кредитных рисков, поскольку характеризуют ее способность находить баланс в точности предсказаний «хороших» и «плохих» исходов, а также не являются чувствительными к дисбалансу классов в имеющихся данных.

Измерение прогностической способности модели подразумевает сравнение результатов, полученных в рамках тестируемого классификатора, с реальными результатами, полученными из доверенных источников (достоверных и актуальных). На основании проведенных наблюдений строится матрица ошибок (см. таблицу 2.11).

Таблица 2.11 – Матрица ошибок модели

Общее количество наблюдений $\sum (TP, TN, FP, FN)$		Прогнозируемое состояние	
		Положительный прогноз (PP)	Отрицательный прогноз (PN)
Фактическое состояние	Положительное (P)	Истинно-положительный (TP)	Ложно-отрицательный (FN)
	Отрицательное (N)	Ложно-положительный (FP)	Истинно-отрицательный (TN)

Источник: составлено автором на основе [266]

При сравнении результатов классификации могут быть получены следующие исходы классификации:

- истинно-положительные (англ. True Positive, TP), т. е. положительное наблюдение в модели определяется как положительное, ошибка отсутствует;
- истинно-отрицательные (англ. True Negative, TN), т. е. отрицательное наблюдение в модели определяется как отрицательное, ошибка отсутствует;
- ложно-положительные (англ. False Positive, FP), т. е. отрицательное наблюдение в модели определяется как положительное, присутствует ошибка I рода, переоценка;
- ложно-отрицательные (англ. False Negative, FN), т. е. положительное наблюдение в модели определяется как отрицательное, присутствует ошибка II рода, недооценка.

Математическое описание метрик измерения прогностической способности модели оценки кредитных рисков зафиксировано, например, в работе [316].

Метрика достоверность (англ. Accuracy), полученная из тестовых данных, используется в качестве основной метрики качества прогнозирования в используемой модели оценки рисков и представляет собой долю правильных прогнозов в общем количестве реализованных наблюдений по всем классам. Метрика достоверность вычисляется по формуле:

$$Acc = \frac{\Sigma(TP, TN)}{\Sigma(TP, TN, FP, FN)}. \quad (2.15)$$

Метрика точности (англ. Precision) определяется как доля точных положительных прогнозов в общем количестве наблюдений в пределах положительного класса и вычисляется по формуле:

$$Pr = \frac{\Sigma(TP)}{\Sigma(TP, FP)}. \quad (2.16)$$

Метрика точности также называется положительной прогностической ценностью.

Метрика полноты (англ. Recall) измеряет долю правильных прогнозов положительных значений класса в тестовом наборе данных положительного фактического состояния и вычисляется по формуле:

$$Rec = \frac{\Sigma(TP)}{\Sigma(TP, FN)}. \quad (2.17)$$

Метрика полноты также называется истинной положительной частотой.

Метрики точности и полноты показали свою эффективность при наличии несбалансированных данных в модели, т. к. не зависят от соотношения классов, в отличие от метрики достоверность, которая при использовании несбалансированных данных ведет к искажению результатов.

В измерении прогностической точности модели также участвуют такие метрики, как чувствительность (англ. Sensitivity) и специфичность (англ. Specificity) Метрика чувствительность показывает, насколько хорошо в модели определяются истинно положительные результаты и вычисляется по формуле:

$$Sen = \frac{TPN}{TPN + FNN}, \quad (2.18)$$

где TPN – количество истинно-положительных результатов наблюдений;

FNN – количество ложно-отрицательных результатов наблюдений.

Метрика специфичности показывает, насколько хорошо в модели определяются истинно отрицательные результаты и вычисляется по формуле:

$$Spec = \frac{TNN}{TNN + FPN}, \quad (2.19)$$

где TNN – количество истинно-отрицательных результатов наблюдений;

FPN – количество ложно-положительных результатов наблюдений.

F-мера является метрикой, которая представляет равновесие между точностью и полнотой и рассчитывается как гармоническое среднее по формуле:

$$F = \frac{2 \cdot Pr \cdot Rec}{Pr + Rec}. \quad (2.20)$$

Максимальное значение F-меры обеспечивается при достижении метриками точности и полноты значения 1, в случае стремления одной из метрик к 0, F-мера также стремится к нулю.

Важно построить модель оценки кредитных рисков, одинаково хорошо различающей и «хороших», и «плохих» заемщиков. Для оценки качества классификации модели часто строят ROC-кривую (англ. Receiver Operator Characteristic), которая показывает зависимость доли верно классифицированных положительных исходов (чувствительности) от доли не верно классифицированных отрицательных исходов (единица минус специфичность). Пример построения ROC-кривой представлен на рисунке 2.2.

Алгоритм построения ROC-кривой подробно описан в работе [255]. Для вычисления площади под ROC-кривой (англ. Area Under Curve, AUC) используется интегральный подход. Это означает, что площадь под кривой вычисляется путем интегрирования кривой от 0 до 1 по оси абсцисс. AUC может принимать значения от 0,5 до 1, где 0,5 означает, что модель работает как случайное угадывание, а 1 означает идеальную модель (таблица 2.12). На практике анализируют не только точечную оценку площади под ROC-кривой, но и ее доверительный интервал. Следует отметить, что анализ ROC-кривой и AUC

является базовым методом оценки прогностической способности модели в задачах бинарной классификации в различных сферах [79, 100, 119, 254, 337].

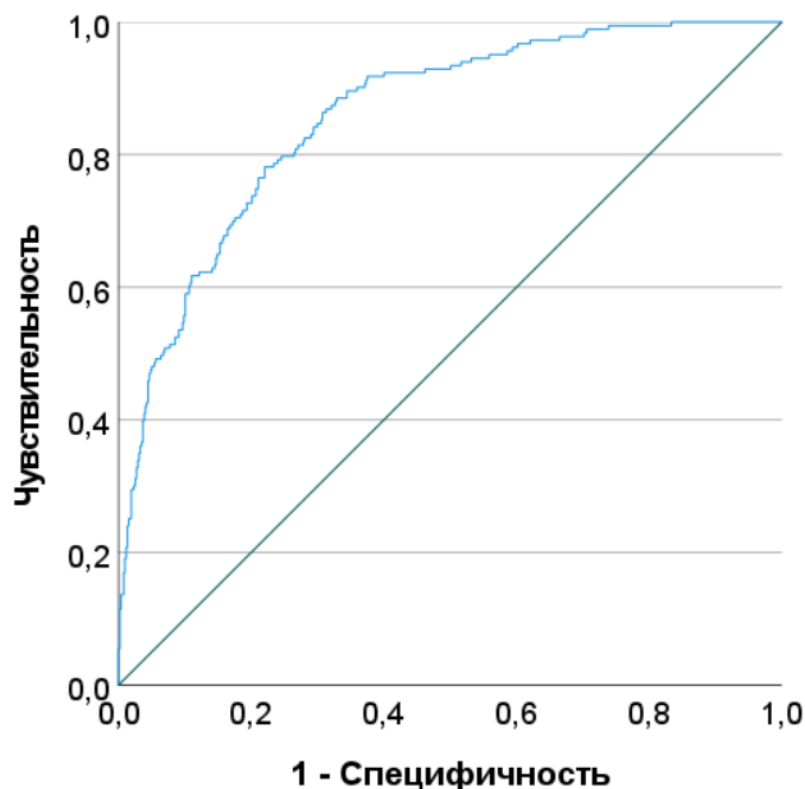


Рисунок 2.2 – Пример построения ROC-кривой

Источник: составлено автором на основе расчетов в пакете SPSS

ROC-кривая может строиться по значениям спрогнозированных вероятностей дефолта по модели оценки кредитных рисков или непосредственно по значениям тестируемого источника данных оценки кредитных рисков. В первом случае ROC-анализ позволяет выбрать порог вероятности разделения «плохих» и «хороших» заемщиков для достижения приемлемого уровня чувствительности и специфичности модели. А во втором – оптимальное значения показателя для ранжирования («отсечки») «плохих» и «хороших» заемщиков. Модель с высокой чувствительностью будет консервативной с ужесточенным отбором заемщиков микрозаймов – максимальным предотвращением пропуска «плохих» заемщиков.

Таблица 2.12 – Критерии качества модели по показателю AUC

Значение AUC	Качество модели
[0,5; 0,6)	Неудовлетворительное качество модели
[0,6; 0,7)	Среднее качество модели
[0,7; 0,8)	Хорошее качество модели
[0,8; 0,9)	Очень хорошее качество модели
[0,9; 1]	Отличное качество модели

Источник: составлено автором

Задача анализа чувствительности – минимизировать кредитный риск, связанный с выдачей микрозайма. Модель с высокой специфичностью будет менее тщательно выявлять «плохих» заемщиков и будет являться рискованной.

Задача анализа специфичности в кредитном скоринге – минимизировать риск упущенной выгоды, связанной с отказом в выдаче кредита. Для определения оптимального порога классификации существуют множество критериев, например:

- а) уровень плохих кредитов в портфеле одобренных заявок;
- б) минимизация ошибок классификации и др.

Выбор оптимального значения порога отсечения зависит от цены совершения ошибки первого и второго рода при классификации. Модель должна точнее классифицировать «плохих» заемщиков, т. к. в кредитном скоринге цена ошибки первого рода выше. При снижении порога отсечения в модели будет увеличиваться чувствительность, т. е. способность модели правильно выявлять тех заемщиков, у которых будет просрочка платежа. За оптимальный порог отсечения можно взять точку баланса между чувствительностью и специфичностью.

Коэффициент Джини (англ. Gini Coefficient) показывает степень отклонения предсказаний модели от случайного угадывания и переводит площадь под ROC-кривой в коэффициент от 0 до 1 [334]:

$$Gini = 2 \cdot AUC - 1, \quad (2.21)$$

где AUC – площадь под ROC-кривой.

ROC-кривую можно строить для полученного скорингового балла по модели оценки кредитного риска или для отдельного источника данных (переменных). Чем больше коэффициент Джини, тем больший вес имеет переменная с точки зрения полезности для будущей модели оценки кредитных рисков (см. таблицу 2.13).

Для оценки кредитных рисков, как правило, выбирают переменные, коэффициент Джини по которым составляет выше среднего в сравнении с уже используемым переменным в текущих скоринговых модели оценки кредитных рисков.

Таблица 2.13 – Критерии предсказательной способности модели по коэффициенту Джини

Коэффициент Джини	Предсказательная способность модели
[0; 0,25)	Слабая предсказательная способность
[0,25; 0,45)	Средняя предсказательная способность
[0,45; 0,6)	Сильная предсказательная способность
[0,6; 1]	Максимальная предсказательная способность

Источник: составлено автором

Площадь под ROC-кривой или коэффициент Джини часто используют для сравнения двух и более моделей между собой. Следует отметить, что показатель площади под кривой предназначен только для сравнительного анализа моделей между собой. Значение площади под кривой не содержит никакой информации о чувствительности и специфичности модели.

Значение коэффициента Джини можно рассчитать как до, так и после включения анализируемой новой переменной в модель и оценить прирост (или сокращение) данного коэффициента в процентных пунктах [296]. Такая метрика носит название *прирост качества классификации* (англ. Uplift):

$$Uplift = Gini_1 - Gini_0, \quad (2.22)$$

где $Gini_0$, $Gini_1$ – значения коэффициента Джини до и после включения новой переменной в модель.

Положительный Uplift свидетельствует о ценности анализируемой переменной для действующей модели оценки кредитных рисков.

Одной из метрик общей классификационной способности является *логистическая функция потерь* (Log loss), которая для бинарной классификации вычисляется по формуле:

$$L(\omega) = -\frac{1}{n} * \sum_{i=1}^n (y_i * \log(\hat{p}_i) + (1 - y_i) * \log(1 - \hat{p}_i)) \rightarrow \min, \quad (2.23)$$

где n – размер выборки (количество наблюдений);

y_i – истинная метка класса (1 или 0);

$\hat{p}_i$ – вероятность принадлежности к позитивному классу;

$1 - \hat{p}_i$ – вероятность принадлежности к негативному классу.

Основным преимуществом логистической функции потерь является устойчивость к выбросам данных.

В рамках модели оценки кредитных рисков микрофинансовых организаций производится ранжирование заемщиков по шансам наступления просрочки по займу. При этом дефолтные и без дефолтные займы должны иметь разные баллы, которые являются производными от вероятности дефолта по модели. А качество и эффективность модели оценки кредитных рисков определяется более явным разделением баллов между «хорошими» и «плохими» займами (рисунок 2.3).

Идеальной считается такая модель оценки кредитных рисков, при которой распределение баллов между «плохими» и «хорошими» заемщиками не пересекаются, а находятся рядом друг с другом. Поэтому качества классификационной способности модели оценки кредитных рисков можно оценить, проанализировав распределение скоринговых баллов или вероятностей дефолта «плохих» и «хороших» заемщиков.

При ранжировании заемщиков по хорошей модели оценки кредитных рисков «плохие» и «хорошие» заемщики должны иметь максимально различающиеся скоринговые баллы (вероятности дефолта). Например, «плохим» заемщикам присваивался меньший балл, чем «хорошим». В итоге группа «плохих» кредитов должна иметь в сумме меньше баллов, чем группа «хороших».

Чем более явно будут разделены распределения скоринговых баллов для «плохих» и «хороших» кредитов, тем эффективнее будет оценка кредитных рисков в МФО. Идеальной считается модель, при которой распределения баллов у «плохих» и «хороших» заемщиков не пересекаются, а находятся рядом друг с другом.

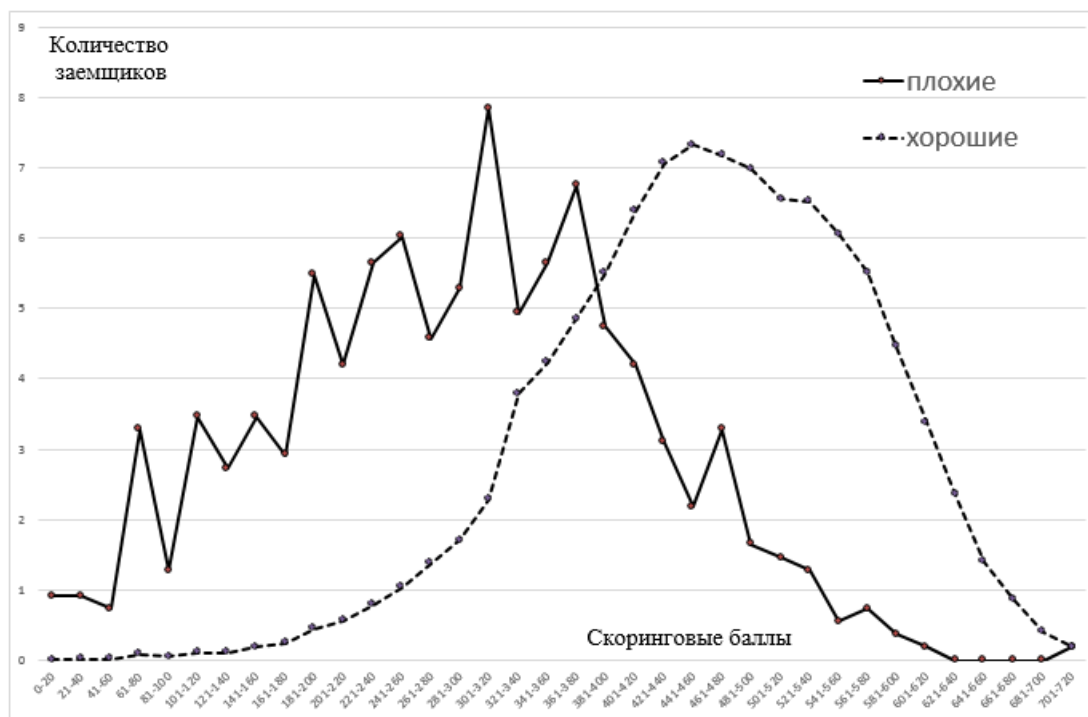


Рисунок 2.3 – Пример распределения скоринговых баллов

Источник: составлено автором по [89]

При анализе качества модели оценки кредитных рисков также анализируют значения *процентилей*, полученных по распределению баллов. Обычно анализируют значения от 10 до 20% в нижней части распределения баллов. Анализ процентилей важен для определения балла, ниже которого не одобряются заявки по займам. При анализе процентилей обращают внимание, какой процент «плохих» займов от их общего числа располагается в нижней части распределения баллов в модели. Например, если общий кумулятивный процент «плохих» займов равен 20% при интервале полученных баллов 300–320, то пороговый балл для отсеечения «плохих» заемщиков выбирают 320. Чем больше «плохих» потенциальных заемщиков будет в нижних процентах распределения балла, тем

больше «плохих» кредитов может быть исключено при помощи соответствующего данному процентиллю значения отсечения балла.

Для оценки качества прогнозирования модели оценки кредитных рисков на основе этого распределения рассчитывают тест Колмогорова-Смирнова, в рамках которого сравниваются два кумулятивных распределения баллов «хороших» и «плохих» заемщиков [46].

Статистика Колмогорова-Смирнова вычисляется как максимальная разница между кумулятивными функциями распределения «плохих» и «хороших» заемщиков по следующей формуле [91]:

$$KS = \max_x |F_m(x) - G_n(x)| \cdot 100, \quad (2.24)$$

где $F_m(x)$ и $G_n(x)$ – эмпирические кумулятивные распределения баллов для «плохих» и «хороших» заемщиков;

n, m – количество «плохих» и «хороших» заемщиков.

Алгоритм проверки гипотезы в соответствии с тестом Колмогорова-Смирнова включает следующие этапы. Ранжирование и группировка заемщиков в порядке увеличения баллов. Расчет следующих показателей в каждой полученной группе заемщиков:

- количество «хороших» заемщиков;
- количество «плохих» заемщиков;
- отношение шансов «плохих» к «хорошим» заемщикам;
- процент «плохих» и «хороших» займов;
- кумулятивная сумма «плохих» и «хороших» займов;
- кумулятивный процент «плохих» и «хороших» займов;
- общий кумулятивный процент плохих займов от их общего числа;
- разница между кумулятивными процентам плохих и хороших займов.

После чего нужно найти максимальную разность между кумулятивным процентом «хороших» и «плохих» займов и вычислить по формуле (2.24) статистику Колмогорова-Смирнова (см. рисунок 2.4).

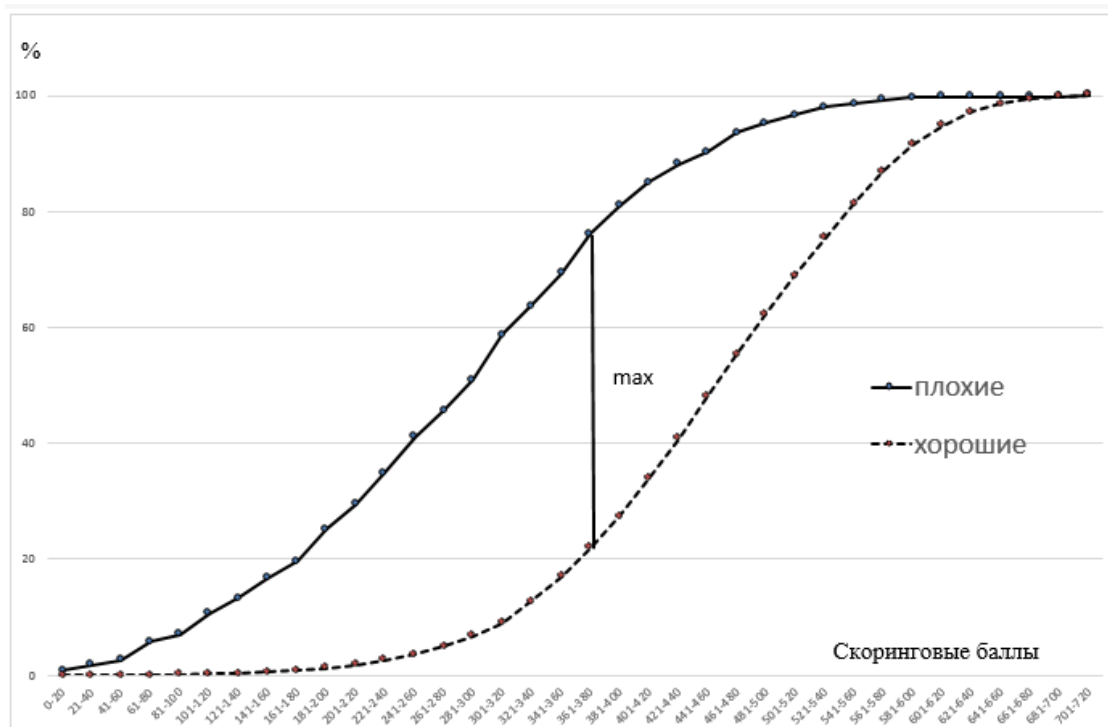


Рисунок 2.4 – Графическая иллюстрация расчета статистики Колмогорова-Смирнова

Источник: составлено автором по [89]

Полученное значение статистики необходимо сравнить с табличным значением по таблице распределения Колмогорова-Смирнова с выбранным уровнем значимости или при числе «плохих» и «хороших» заемщиков соответственно больше 80 можно взять приближенное пороговое значение, вычисляемое по формуле:

$$z(\alpha) \sqrt{\frac{m+n}{mn}}, \quad (2.25)$$

где $z(\alpha)$ – значение, соответствующее выбранному уровню значимости.

Если расчетное значение статистики по формуле (2.24) меньше порогового по таблице или по формуле (2.25), то гипотезу о равенстве двух функций распределений отвергают. Диапазон изменения значений статистики Колмогорова-Смирнова может быть от 0 до 100. Чем выше значение статистики Колмогорова-Смирнова, тем лучше классифицирует модель. Крайние значения статистика Колмогорова-Смирнова, как правило, не принимает, и ее значения сосредоточены обычно от 20–25 до 75–80.

Альтернативной мерой оценки валидации модели оценки кредитных рисков может быть коэффициент дивергенции, который представляет собой оценку разницы математических ожиданий распределений баллов для «плохих» и «хороших» заемщиков, нормализованную дисперсиями этих распределений и рассчитывается по формуле:

$$D = \frac{2 \cdot (\bar{x}_1 - \bar{x}_2)^2}{(s_1^2 - s_2^2)^2}, \quad (2.26)$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения балла для «плохих» и «хороших» заемщиков;
 s_1^2 и s_2^2 – дисперсии балла для «плохих» и «хороших» заемщиков.

Чем больше значение коэффициента дивергенции, тем лучше качество модели с точки зрения ее классификационной способности.

В таблице 2.14 приведено сравнение возможности применения критериев валидации в зависимости от формы распределения скоринговых баллов.

Таблица 2.14 – Сравнение возможности применения показателей валидности модели

Распределение баллов в модели	Статистика Колмогорова-Смирнова	Коэффициент дивергенции
Симметричны	подходит	подходит
Скошены внутрь	подходит	не подходит (будет завышен)
Скошены наружу	подходит	не подходит (будет занижен)
Вложены друг в друга	не подходит (будет завышена)	подходит

Источник: составлено автором по [89]

Значение коэффициента дивергенции чувствительно к асимметрии распределений баллов и может давать заниженную или завышенную оценку в зависимости от направления асимметрии. Статистика Колмогорова-Смирнова, наоборот, устойчива к асимметрии распределений баллов. Однако статистика Колмогорова-Смирнова может давать ошибочную оптимистичную оценку при наложении кривых распределений баллов друг на друга. При нормальном или приблизительно нормальном распределении баллов для «плохих» и «хороших» заемщиков можно использовать и статистику Колмогорова-Смирнова, и

коэффициент дивергенции. Если кривые распределения скошены внутрь – дивергенция будет очень большой и ее использование будет не корректно. Если обе кривые распределения будут скошены наружу – дивергенция будет очень маленькой и также не будет являться адекватной мерой качества валидации модели.

Еще одной мерой качества построенной модели может быть коэффициент разделения. Алгоритм расчета коэффициента разделения следующий. Сначала рассчитывают эмпирические кумулятивные распределения скорингового балла для группы «плохих» и «хороших» заемщиков. Затем для каждого значения скорингового балла или интервала скоринговых баллов принимают две кумулятивные эмпирические вероятности – координаты по оси абсцисс и оси ординат. После чего необходимо начертить координаты под прямым углом и найти площадь между начерченной кривой и прямой, расположенной под углом 45° (см. рисунок 2.5).

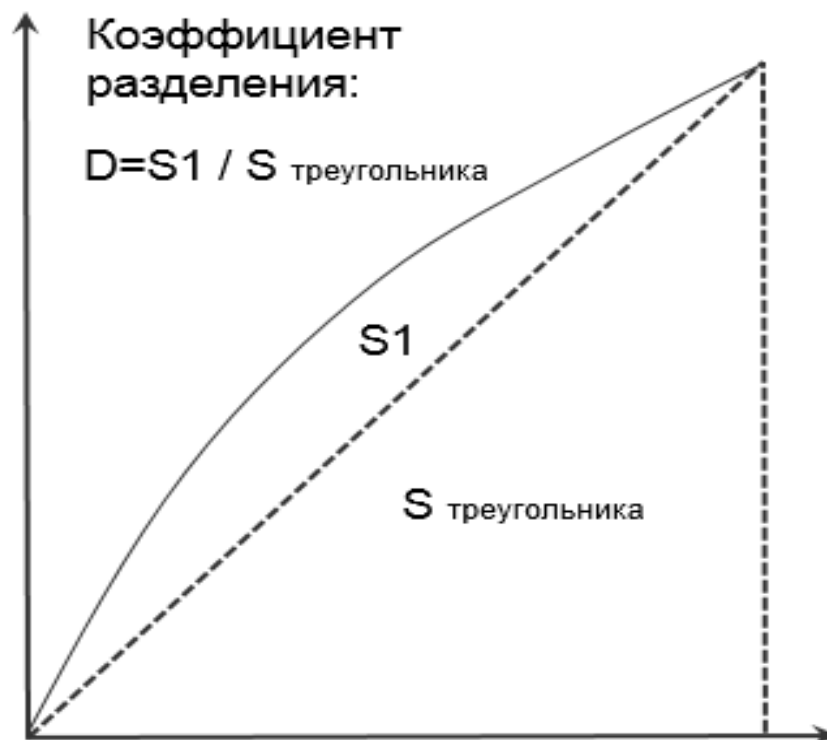


Рисунок 2.5 – Иллюстрация расчета коэффициента разделения

Источник: составлено автором по [89]

Если начерченная кривая будет лежать частично выше или частично ниже прямой, необходимо будет найти абсолютную разность между площадью выше этой прямой и площадью ниже ее. Такой случай будет наблюдаться, когда одна кривая распределения баллов будет находиться внутри другой. После чего нужно абсолютную разность площадей разделить на площадь треугольника, находящуюся под прямой, расположенной под углом 45° – полученная величина будет являться коэффициентом разделения. Данный алгоритм похож на построение кривой Лоренца и требует вычисление площадей под кривыми. Эта задача может быть выполнена, например, аппроксимации площади методом трапеций.

Перевод коэффициентов модели в скоринговую карту. Особенность применения моделей оценки кредитных рисков в банках и МФО, что часто на их основе рассчитывают скоринговые баллы. Особенно это характерно при использовании логистической регрессии в качестве математической модели (логистическая регрессия будет рассмотрена в главе 3 параграфе 2). Поэтому разберем подробнее алгоритм перевода коэффициентов логистической регрессии в скоринговые баллы, на основе которых рассчитываются метрики, рассмотренные выше.

Если взять оценки коэффициентов логистической регрессии и умножить их на значения независимых переменных, то получится итоговый скоринговый балл в шкале натуральных логарифмов:

$$\text{итоговый балл} = \hat{b}_1 x_1 + \hat{b}_2 x_2 + \dots + \hat{b}_k x_k, \quad (2.27)$$

где x_i – значение предикторов для оцениваемого заемщика, $\hat{b}_j$ – оценки коэффициентов логистической регрессии.

Для приведения скоринговых баллов в линейную шкалу используют прием масштабирования. Масштабирование не изменяет прогностическую способность модели оценки кредитных рисков, а лишь переводит скоринговые баллы в новую шкалу, удобную для использования. В большинстве случаев на практике для

различных скоринговых карт используют непересекающиеся по значениям баллов друг с другом шкалы во избежание путаницы при назначении баллов заемщику.

Скоринговый балл в линейной шкале представляет собой отношение шансов «хороших» заемщиков к «плохим» (или наоборот). Для масштабирования необходимо прежде всего задать диапазон числовой шкалы с минимальным и максимальным значением (например, от 0 до 1000).

На результат масштабирования также влияют два показателя: количество баллов, которое удваивает шансы стать «хорошим» заемщиком и значение шкалы, в котором достигается заданное отношение шансов «хороших» к «плохим».

Наиболее часто используют скоринговые карты, в которых каждые 20 баллов удваивают шансы стать «хорошим». Другой стандарт – каждые 40 баллов удваивают шансы стать «хорошим» заемщиком. Но в общем случае это значение может быть произвольной величиной.

Для приведения коэффициента логистической регрессии в скоринговый балл в линейной шкале применяют следующее преобразование:

$$\text{балл} = A + R \cdot \hat{b}_j, \quad (2.28)$$

где A – смещение;

R – множитель.

Множитель R определяют по формуле:

$$R = \frac{D}{\ln(2)}, \quad (2.29)$$

где D – количество баллов, удваивающее шансы.

Смещение A определяют по формуле:

$$A = B - R \cdot \ln(C), \quad (2.30)$$

где B – значение на шкале баллов, в которой соотношение шансов составляет $C:1$.

В таблице 2.15 для иллюстрации приведены результаты расчетов параметров логистической регрессии, используемой для оценки кредитных рисков, для одного условного атрибута скоринговой карты – стажа работы (лет).

Все параметры представленной для примера модели статистически значимые. Категориальная переменная была введена в модель с помощью нескольких бинарных фиктивных переменных. За опорную категорию была выбрана последняя категория с максимальным стажем работы.

Таблица 2.15 – Результаты расчетов коэффициентов логистической регрессии

Категории независимой переменной	Коэффициент регрессии В	Стандартная ошибка	Статистика Вальда	Значимость	Exp (В)
<= 2	2,367	0,353	44,865	0,000	10,670
3 – 4	1,780	0,385	21,313	0,000	5,928
5 – 6	1,585	0,480	10,924	0,001	4,880
7 – 14	0,944	0,349	7,311	0,007	2,570
Константа	– 2,406	0,315	58,417	0,000	0,090

Источник: составлено автором на основании самостоятельных расчетов

Предположим, что каждые 40 баллов удваивают шансы наступления дефолта по кредиту, а в точке 600 баллов отношение шансов составляет 72:1 (такие предположения являются одним из общепринятых стандартов расчета скоринговых баллов в банковской сфере).

По формуле (2.29) множитель в этом случае будет равен 57,71, а смещение по формуле (2.30) будет равно 413,20. Поскольку с увеличением стажа работы шансы дефолта по кредиту падают, множитель нужно брать с отрицательным знаком, чтобы значения скоринговых баллов возрастали с ростом стажа.

В таблице 2.16 приведен расчет скоринговых баллов по формуле (2.28). В моделях с несколькими независимыми переменными для получения общего скорингового балла необходимо сложить баллы по каждой независимой переменной.

Если уравнение логистической регрессии строиться по значениям WOE, то формула расчета скорингового балла в линейном масштабе будет следующая:

$$\text{балл} = - \left(WOE_j \cdot b_i + \frac{b_0}{n} \right) \cdot R + \frac{A}{n}, \quad (2.31)$$

где WOE_j – значение WOE для каждой j -ой категории сгруппированной переменной;

b_i – коэффициент регрессии для i -ой переменной.

b_0 – константа;

n – количество независимых переменных в уравнении регрессии.

Таблица 2.16 – Расчет скоринговых баллов

Категории независимой переменной, лет	Коэффициент регрессии b_j	Скоринговый балл в линейной шкале $b_j * R$	Скоринговый балл с учетом смещения $A - b_j * R$
≤ 2	2,367	136,6	550
3 – 4	1,780	102,7	516
5 – 6	1,585	91,48	505
7 – 14	0,944	54,5	468
15+	0	0	413

Источник: составлено автором

В таблице 2.17 приведены результаты расчета модели логистической регрессии с одной независимой переменной категорий стажа работы рассматриваемого примера после замены категорий их весами WOE.

Таблица 2.17 – Результаты расчетов коэффициентов логистической регрессии по значениям WOE

Переменные	Коэффициент регрессии В	Стандартная ошибка	Статистика Вальда	Значимость	Exp (В)
Стаж работы	1,000	0,122	67,031	0,000	2,718
Константа	– 1,039	0,093	125,814	0,000	0,354

Источник: составлено автором

В таблице 2.18 представлен расчет скоринговых баллов на основе разработанной модели. Расчет скоринговых баллов по рассмотренному алгоритму во многом зависит от того, насколько правильно была проведена категоризация количественных переменных, насколько существенна разница WOE между

категориями. Особенности категоризации независимых переменных будут рассмотрены в параграфе 1 главы 3.

Таблица 2.18 – Расчет скоринговых баллов при построении регрессии по WOE

Категории независимой переменной, лет	WOE _j	Скоринговый балл
<= 2	1,00	274
3 – 4	0,41	329
5 – 6	0,22	366
7 – 14	– 0,42	377
15+	– 1,37	411

Источник: составлено автором

Выводы по главе 2

Проведено комплексное исследование особенностей информационного обеспечения процедур оценки кредитных рисков микрофинансовых организаций. Выработан системный подход для выбора необходимой информации для оценки кредитных рисков МФО, основанный на классификации источников и видов традиционных и альтернативных данных о заемщиках МФО по различным характеристикам, в том числе поведенческим, а также на выборе критериев оптимальных проверок таких данных, включая соблюдение законодательных норм, идентификацию заемщиков и выявление мошеннических заявок.

Систематизация целевых показателей и метрик, используемых для оценки эффективности новых данных с целью выбора их рационального состава, позволила выделить базовые критерии – метрики силы влияния на показатели риска, метрики классификационной способности и метрики оценки экономического эффекта. Преимущество представленной пошаговой методики измерения ценности и отбора наиболее подходящих данных для оценки кредитных рисков заключается в возможности формировать рациональный состав данных в условиях ограниченных ресурсов МФО.

Выработаны рекомендации по применению соответствующих метрик дискриминирующей способности выбранных источников данных и построенных на их основе моделей оценки кредитных рисков, которые позволяют адекватно оценить классификационную способность модели оценки кредитных рисков посредством расчета соответствующих статистических тестов и метрик бинарной классификации заемщиков микрофинансовых организаций.

Глава 3 Методологические подходы к оценке кредитных рисков микрофинансовых организаций

3.1 Методика разработки скоринговых моделей оценки кредитных рисков микрофинансовых организаций

В данной главе будут рассмотрены часто используемые математико-статистические модели, а также модели и алгоритмы машинного обучения, которые используются для оценки кредитных рисков в микрофинансовых организациях. Модели оценки кредитных рисков на основе алгоритмов машинного обучения или классической логистической регрессии являются конечным этапом в разработке СПР МФО. Успешность работы модели в кредитном скоринге определяется во многом соблюдением общей методики ее разработки. При этом ошибочно считать, что алгоритмы машинного обучения, называемые еще методами интеллектуального анализа данных, не требуют тщательного предварительного анализа данных, подаваемых на вход алгоритма. Оба подхода, статистическое моделирование и машинное обучение, должны применяться после уже проработки всех методологических аспектов разработки модели оценки кредитных рисков, которые будут рассмотрены в данном параграфе.

Как уже было показано в параграфе 3 главы 1 в научной литературе и работах по применению моделей оценки кредитных рисков акцент делается на математической спецификации используемой модели и результатах моделирования. Работ по системному изложению методики построения таких моделей, учитывающей основные методологические проблемы подготовки набора данных для моделирования с учетом особенностей МФО, крайне мало. В связи с этим в данном параграфе остановимся подробно на методике построения моделей оценки кредитных рисков в МФО, отраженной в работах автора [98, 109].

Весь процесс моделирования можно представить в виде трех последовательных укрупненных этапов: подготовительный, исследование данных и непосредственно построение модели. Обозначим основные методологические аспекты, которые необходимо учитывать в процессе разработки моделей кредитного риска аналитику МФО на этапе подготовки выборки:

- а) выбор исходной информационной базы;
- б) определение зависимой переменной;
- в) выбор независимых переменных;
- г) формирование обучающей и тестовой выборки;
- д) определение минимального объема выборки.

а) *Выбор исходной информационной базы.* Достаточная и качественная информация о заемщиках выступает одним из гарантов результативности модели оценки кредитных рисков. Разработка модели оценки кредитных рисков, как правило, строится на анализе предыдущего кредитного опыта (исторической выборки), при этом достаточный объем информации является одной из главных предпосылок построения модели. Количество данных может варьироваться в зависимости от конкретных моделей, но в целом данные должны удовлетворять требованиям статистической значимости и случайности. Исходные данные для построения модели могут содержать внутренние данные анкет заемщиков микрофинансовых организаций, а также внешние данные. Более подробно эти виды данных рассмотрены в параграфе 1 главы 2 настоящего диссертационного исследования.

С выбором конкретных источников данных для построения модели нужно определиться заранее, т. к. для новых источников потребуется время на техническую интеграцию или проведение ретро-тестирования. Методика и критерии выбора данных для моделей оценки кредитных рисков подробно изложены во 2 главе настоящей работы.

К вопросу выбора информационной базы моделирования относится не только выбор источников данных, но и периода, за который будет формироваться обучающая выборка.

Для построения точной модели оценки кредитных рисков исходные данные должны обладать определенной исторической давностью, то есть периодом, за который собирается информация о кредитном опыте заемщиков. Исторический период данных для построения модели определяется, как правило, видом кредитования. В отношении микрофинансирования наилучшим вариантом исторической давности информации составляет 24–36 месяцев, т. е. данные не старше 3-х лет.

Из исходной информационной базы необходимо исключать информацию о нетипичных для финансово-кредитного учреждения клиентах, так для микрофинансовых организаций такими клиентами могут быть субъекты, осуществляющие мошеннические действия и другие.

Дополнительным критерием отбора исходных данных может служить регион рынка, для которого строится модель оценки кредитного риска. Также отбор исходных данных может варьироваться по виду кредитования – для МФО могут быть сформированы отдельные скоринговые модели для PDL-займов и ПЛ-займов, для первичных и повторных клиентов, которые базируются на различных исходных данных. Для реализации такого подхода перед построением модели оценки кредитных рисков в большинстве случаев исходную базу клиентов сегментируют с помощью многомерных статистических методов, например, кластерного анализа, деревьев решений или эвристическими методами.

б) *Определение зависимой переменной.* Выбор зависимой переменной определяется целью построения модели оценки кредитных рисков (см. таблицу 2.6).

Зависимая переменная может принимать как количественные, так и качественные значения. Примером количественной целевой переменной является финансовый результат по договору кредита (выданная сумма минус уплаченная сумма). Наиболее часто используемый вид измерения зависимой переменной – категориальный, который включает две категории: «плохой» заемщик, «хороший» заемщик. Как правило, в микрофинансовых организациях к категории «плохой» относят клиентов, имеющих просроченную задолженность 30 дней и более.

Микрофинансовая организация может формировать различные модели оценки кредитных рисков с вариативными значениями зависимой переменной, вводя дополнительные критерии определения «плохого» и «хорошего» заемщика, а также меняя срок просрочки платежей. Примерами зависимой переменной могут быть наличие просроченной задолженности более 10 дней, 20 дней, 30 дней и более по одному кредиту на текущий момент или худший статус за все время кредитной истории, размер просроченной задолженности, количество просрочек более заданного числа дней и другие параметры. Для правильного выбора числа дней просрочки целевой переменной необходимо построить стандартный в кредитной риск-аналитике отчет о миграции просрочки. Данный отчет показывает долю перехода заемщиков, допустивших просрочку, из одной категории просрочки в другую. Для микрофинансовых компаний, как правило, более 90% клиентов допустившие просрочку 30 и более дней, уже не погашают ее, в отличие от банков, где этот показатель исходя из практического опыта и регуляторных требований равен 90 дням.

Основные метрики кредитных рисков микрофинансовых организаций, которые могут быть использованы в качестве зависимой переменной в модели кредитного скоринга – NPL и FPD подробно рассмотрены в первой главе настоящего диссертационного исследования. Выбор между показателем NPL и FPD определяется периодом «созревания» выборки. Максимальный срок займа в PDL кредитовании 30 дней (без учета продления). Предположим, срок просрочки для определения «плохого» заемщика был выбран 30 дней. Если выборка для моделирования формируется на определенную дату, то необходимо для построения модели при целевом показателе FPD30 отобрать все договора с датой выдачи минус 60 дней от текущей даты, т. е. период созревания для FPD30 составляет 2 месяца. Показатель же NPL30 «созревает» не меньше 3 месяцев и более при высокой доле продлений займов. С этой точки зрения использование показателя FPD предпочтительнее NPL.

Однако окончательный выбор между NPL и FPD зависит от экономики и структуры доходов МФО. Значительную часть доходов компания может

зарабатывать на продлениях займов, и тогда наличие просрочки при условии продления займа не будет являться негативным событием, и наилучшим показателем измерения кредитного риска будет FPD, в противном случае можно выбрать и NPL. На практике между значениями NPL и FPD существует всегда сильная корреляция.

в) *Определение независимых переменных.* Состав независимых переменных определяется выбранной информационной базой для моделирования. В качестве независимых переменных при построении модели оценки рисков могут быть использованы данные из кредитной заявки: социально-демографические данные о заемщике (пол, семейное положение, возраст, должность, общий стаж работы и стаж работы на последнем месте, срок проживания по текущему адресу, наличие детей, уровень образования, доход заемщика и доход семьи, работает или на пенсии и др.); информация о запрашиваемом кредите (выбранный срок, сумма займа, цель займа); реже используются маркетинговые данные (источник трафика, с которого пришла заявка на займ, мотив выбора микрофинансовой организации).

Следующий тип данных для формирования независимых переменных (используется для повторных клиентов) – внутренняя кредитная история заемщика: наличие задолженностей по прошлым займам, количество закрытых займов в микрофинансовой организации и другие. Для новых клиентов МФО такая информация недоступна.

Одним из основных источников данных для формирования независимых переменных в модели оценки кредитных рисков являются данные БКИ на момент подачи заявки заемщиком: рейтинг заемщика, подробная информация об имеющихся кредитах в других микрофинансовых организациях или банках, детальная информация о просроченных или полностью погашенных прошлых кредитах и прочее.

На этапе выбора независимых переменных необходимо продумать алгоритмы расчета возможных производных показателей по исходным данным выбранных источников. Разберем пример формирования наборов независимых переменных на основе исходных данных кредитных отчетов. Исходная

информация о кредитах в xml-файлах кредитных отчетов во всех БКИ может быть описана следующими полями: сумма кредита, дата открытия и закрытия кредита, тип кредитора, тип займа, срок договора. Исходя из анализа данных полей в записях о кредитах в кредитном отчете можно выделить, например, следующие типы группировок для кредитов:

- 1) статус кредита (закрытые, текущие и всего);
- 2) тип кредитора (МФО, банки, иные кредиторы и всего);
- 3) тип кредита (микрозаем, банковский заем, кредитные карты, прочие кредиты и всего);
- 4) сумма и срок займа (кредиты с суммой до 30 тыс. руб. и сроком до 30 дней и иные кредиты и всего) – группировка представляет собой альтернативный способ определения микрозайма;
- 5) наличие просрочки по кредиту (с просрочкой, без просрочки и всего);
- 6) период, за который рассчитывают показатель (за последний месяц, квартал, полгода, год, за весь период).

Используя вышеприведенные группировки по записям о кредитах в КИ с помощью различных агрегирующих функций, вычисляются финальные независимые переменных для построения модели. В качестве агрегирующих функций можно предложить использование следующих функций: количество значений, сумма, среднее арифметическое, медианное значение, стандартное отклонение. Примеры рассчитанных показателей по приведенной методике: количество закрытых кредитов с типом микрозайм за последние 365 дней или сумма активных кредитов с просрочкой по кредитам с типом кредитора «банк» за весь период. Все возможные комбинации группировок задают стандартный набор переменных по кредитной истории с сотнями, а иногда и тысячами переменных, из которых затем на основе статистического анализа отбираются наиболее информативные для предсказания значения целевой переменной.

Аналогичным образом формируются производные независимые переменные блока информации кредитного отчета по запросам на кредиты.

Отдельный блок переменных могут быть производные показатели от дат, например, число дней между датой заявки на кредит и датой самого первого (последнего) кредита в КИ.

С точки зрения шкал измерения, независимые переменные могут быть представлены в разных шкалах измерения в зависимости от возможности объективных измерений интересующих признаков, а также применяемого подхода: статистического моделирования или машинного обучения.

На практике скоринговые модели формируются со следующими типами независимых переменных: только с количественными, только с категориальными (в этом случае количественные переменные категоризируются на основе анализа зависимости их влияния на уровень дефолтов), с категориальными и с количественными переменными одновременно.

г) *Формирование обучающей и тестовой выборки.* Доступные для построения модели оценки рисков информационные данные, как правило, называют исторической выборкой. Историческая выборка должна как можно точнее отражать исследуемую генеральную совокупность заемщиков, т. е. быть репрезентативной. Для проверки адекватности и точности предсказания модели оценки кредитных рисков на этапе ее разработки историческую выборку необходимо случайным образом разделить на две группы:

- обучающую выборку – наблюдения, по которым будет непосредственно строиться модель;

- тестовую или контрольную выборку – наблюдения, по которым будет известно значение зависимой переменной, но они не будут участвовать в построении модели, а будут использованы для проверки точности предсказания модели.

Обучающая и тестовая (контрольная) выборка должна формироваться на основе механизма случайного отбора обычно в соотношении 70–80% и 30–20% соответственно от исходного объема исторической выборки. Тестовая выборка используется после построения модели для проверки ее достоверности. Данный этап построения модели также называют валидацией (англ. Validity –

доказанность, обоснованность, пригодность). Однако сформировать стратегию валидации модели необходимо до ее построения.

Проверка достоверности модели заключается в ее применении и сравнении результатов на обучающей и тестовой выборке ввиду того, что модель должна давать корректные прогнозы не только на обучающей совокупности, но и на практике при ее применении.

Как правило, стратегию генерализации модели используют на основе двух выборок. Схожие показатели точности, полученные на обучающей и тестовой выборке выступает главным признаком того, что на практике скоринговая модель будет работать примерно также.

Более сложная стратегия генерализации модели предполагает формирование трех и более выборок:

- первая выборка используется для оценки параметров модели;
- вторая выборка используется для проверки модели, если получены значительные отклонения результатов по обучающей и тестовой выборке, то из них удаляются выбросы или переменные, влияющие на отклонения, и строится новая модель по объединенной первой и второй выборке;
- результаты новой модели тестируются на третьей выборке.

Сложные стратегии генерализации модели характерны при использовании алгоритмов машинного обучения [302, 304]. Существуют и другие стратегии валидации модели, например стратегия прогнозирования постфактум, не предполагающая задание обучающей выборки, модель тестируется на реальных данных в течение определенного периода с последующей корректировкой ее параметров.

д) *Определение минимального объема выборки.* Важным методологическим аспектом построения модели оценки кредитного риска является определение необходимого минимального объема выборки. Обучающая выборка должна формироваться из исходной статистической базы на основе рассмотренных выше критериев случайным образом. Число принятых кредитных решений за заданный период времени должно быть достаточно большим для обеспечения необходимого

объема выборки. Определение минимального объема выборки может опираться на следующие критерии: равномерность распределения зависимой переменной, число независимых переменных в модели, максимально допустимой ошибки выборки, экспертные оценки.

В литературе содержатся различные рекомендации по определению минимального объема выборки и числа предикторов в модели. Так в работе [266] определено, что выбор минимального объема выборки зависит от равномерности распределения значений зависимой переменной. При относительно равномерном распределении необходимо задавать не менее 10 наблюдений на 1 предиктор. При этом чем больше распределение зависимой бинарной переменной смещено в пользу в пользу одной из категорий, тем больше наблюдений нужно брать на один предиктор.

Другой подход, так называемое правило 20 EPV (рассмотрен в работе [257]), также связывает минимальный объем выборки с распределением зависимой переменной и количеством предикторов в модели. Согласно этому подходу, необходимо взять количество наблюдений в исторической выборке, у которых зависимая переменная имеет наименьший объем, в кредитном скоринге это «плохие» заемщики. Это число наблюдений нужно разделить на число предикторов, включенных в модель. На один предиктор должно приходиться не менее 20 наблюдений. Если это правило выполняется, то объем выборки достаточный.

В кредитном скоринге распределение зависимой переменной имеет всегда существенное смещение в пользу «хороших» заемщиков, поэтому два вышерассмотренных подхода недопустимы. Таким образом, необходимо использовать другой метод определения объема выборки, основанный на критерии мощности при задании максимально допустимой ошибки оценки соотношения «плохих» и «хороших» заемщиков в генеральной совокупности.

Обучающая выборка заемщиков для построения модели должна отражать генеральную совокупность всех потенциальных заемщиков с неизвестными долями «плохих» и «хороших» кредитов. Предположим, 95% уверенность в том,

что соотношение «плохих» и «хороших» заемщиков в обучающей выборке должно отражать генеральную популяцию заемщиков. Зная распределение зависимой переменной на тестовых данных, можно рассчитать необходимый объем выборки по формуле:

$$n = \frac{z_{\gamma}^2 w \cdot (1-w)}{\Delta_w^2}, \quad (3.1)$$

где n – минимальный объем выборки;

z_{γ} – значение стандартного нормального закона распределения, определяемое в зависимости от выбранного уровня надежности γ ;

w – доля «плохих» клиентов в тестовой выборке;

Δ_w – максимально допустимая предельная ошибка оценки доли «плохих» заемщиков.

Данная формула исходит из предположения формирования выборки простым случайным отбором из генеральной совокупности неизвестного объема.

Например, среди 700 клиентов предварительной выборки 50 оказались «плохими». Оценка доли «плохих» клиентов по имеющимся данным для построения модели составила около 0,07 или 7%. При таком значении оценки доли предположим ошибку не более чем на 5%, что будет соответствовать допустимой предельной ошибке оценки доли 0,0035. При этом необходимо получить результаты с надежностью не менее 99%. В этом случае z -значение стандартного нормального закона распределения составит около 2,58. Подставив эти значения в формулу (3.1), получим минимально необходимый объем выборки 35 260 наблюдений. Такой объем выборки не всегда доступен для микрофинансовых организаций. Если задать предельную ошибку выборки 10% или 0,007, а надежность 95%, что вполне приемлемо для построения предварительной модели на первоначальном этапе, до ее калибровки, получаем совсем другие результаты: 5 104 наблюдений – приемлемый уровень выборки для микрофинансовых организаций.

При определении минимального объема выборки можно воспользоваться и формулой интервальной оценки генеральной доли:

$$w - z_{\gamma} \sqrt{\frac{w(1-w)}{n}} \leq P \leq w + t_{\gamma} \sqrt{\frac{w(1-w)}{n}}, \quad (3.2)$$

где P – оцениваемая доля «плохих» заемщиков в генеральной совокупности.

По указанной формуле найдем верхнюю и нижнюю границу доверительного интервала для оценки генеральной доли «плохих» заемщиков для вышерассмотренного примера при уровне надежности 95%: 0,05 и 0,09. В лучшем случае с точки зрения оценки риска, но наихудшем с точки зрения для определения минимального объема выборки, будет нижняя граница доверительного интервала для доли «плохих» клиентов. В выше рассмотренном примере это около 0,05, в таком случае необходимо будет взять объем выборки примерно 7 299 клиентов при относительной ошибке оценки доли в 10%. А при доле «отрицательных» исходов в 0,09 будет достаточно 3 885 наблюдений. При повышении точности оценивания генеральной доли до относительной ошибки в 5%, необходимо провести 29 196 и 15 537 наблюдений соответственно.

Второй этап методики разработки моделей оценки кредитных рисков включает анализ и корректировку отобранных данных:

- а) описательный статистический анализ переменных;
- б) анализ пропущенных значений;
- в) корректировка распределения зависимой переменной;
- г) корректировка распределения независимых переменных;
- д) категоризация количественных независимых переменных;
- е) диагностика и устранение мультиколлинеарности.

а) *Описательный статистический анализ переменных.* Перед построением модели кредитного риска исходные данные следует подвергнуть подробному и всестороннему статистическому анализу [4]. Исследование данных решает две главные задачи:

- позволяет обнаружить возможные ошибки и пропуски в данных;
- позволяет оценить силу связи между зависимой и независимыми переменными, сделав предварительный отбор последних для построения модели.

Для категориальных независимых переменных важно выводить частотные таблицы распределений. Категории независимых переменных с малым числом наблюдений, как правило, при возможности группируют с соседними категориями. Число наблюдений в каждой категории зависимой переменной должно быть, по разным оценкам, не менее 3–5% от общего числа валидных значений по данной категориальной переменной. Также частотные таблицы позволяют обнаружить недопустимые и ошибочно введенные значения в данных.

Представление о целостности и состоятельности данных для количественных переменных могут дать такие базовые описательные статистики как среднее значение, медиана, стандартное отклонение, процентные точки, доля пропущенных значений. Визуальное представление данных отображают с помощью гистограмм, «ящичковых» диаграмм.

Распределения скоринговых переменных по выборке необходимо сравнивать с распределениями во всем кредитном портфеле для проверки репрезентативности обучающей выборки. Отдельное внимание уделяется наличию в данных допустимых значений, но имеющих экстремальный характер. При незначительном количестве таких величин их, как правило, удаляют из анализа или заменяют средними (медианными) значениями, предварительно исследовав, не являются ли они признаками мошеннических займов.

б) *Анализ пропущенных значений.* Работа с пропущенными значениями является отдельным этапом анализа данных перед построением модели. Причиной пропусков могут быть: ошибки при сборе и вводе данных, невозможность получения информации.

Наблюдения с пропусками в переменных по причине ошибок ввода данных следует исключить из анализа или заменить средними значениями переменных.

Построение модели оценки кредитного риска требует полного набора данных и одинакового числа наблюдений по каждой переменной. При наличии пропусков в данных число наблюдений для каждой переменной будет различаться. Как правило, наличие пропусков в менее 5% наблюдений можно объяснить случайностью [303]. В этом случае наблюдения с пропусками можно просто

удалить, исключив таким образом их из анализа без существенной потери информации и ухудшения качества модели. Наличие в данных наблюдений с более чем 5% пропусков может быть не случайно, а закономерно. Их удаление может привести к искажению результатов моделирования в связи с существенным сокращением количества наблюдений в выборке. В этом случае необходимо выявить причину пропусков и восстановить пропущенные значения, либо заменить их. Наблюдения или переменные, которые имеют значительное число пропущенных значений (более 50%), обычно исключают из анализа.

В случае достаточного количества пропусков следует пропущенные значения перекодировать в отдельную категорию и включить их в анализ, наравне с другими категориями. Такие категории в некоторых случаях вводят с весами, близкими к медианному или среднему значению переменной, в которой анализируются пропуски. Этот подход предполагает неслучайный характер пропусков, в ряде случаев пропуски в данных являются дополнительным индикатором «плохого» заемщика. Например, если заемщик имеет низкий уровень дохода, то он с большой вероятностью оставит пустым поле «Ваш доход» в анкете-заявке. Отдельное направление работы с пропущенными данными – их импутация на основе скрытых закономерностей в данных. В основе импутации данных лежат алгоритмы получения возможных значений для отсутствующих данных по значениям других валидных переменных на основе многомерных статистических методов [344].

в) *Корректировка распределения зависимой переменной.* В основном данная проблема характерна при использовании математической модели логистической регрессии для оценки кредитных рисков. Особенности корректировки выборки с помощью стратегий перевзвешивания или прореживания выборки рассмотрены в следующем параграфе данной главы.

г) *Корректировка распределения независимых переменных.* Независимые количественные переменные могут включаться в модель оценки кредитного риска без преобразований и на основе непрерывных преобразований для приближения их распределения к нормальному. Обычно используют следующие виды

преобразований: квадратное; кубическое; квадратный корень; натуральный или десятичный логарифм; экспоненциальное; величина, обратная квадратному корню; обратная величина. При использовании степенных преобразований ко всем значениям преобразуемой переменной могут добавлять константу для преобразования нуля или отрицательных значений. Такие преобразования количественных переменных могут привести к максимизации их связи с зависимой целевой переменной. Нужный вид преобразований можно определить на данных конкретной выборки эмпирическим путем. Наилучшее преобразование имеет наибольшую значимую тесноту связи с зависимой переменной. Окончательный выбор наиболее предпочтительного варианта преобразования происходит на основе сравнения характеристик точности полученных финальных моделей оценок кредитных рисков.

Математически оптимальный вид преобразования может быть найден с помощью метода Бокса-Кокса, который основан на степенном преобразовании переменных [188]:

$$x_i(\lambda) = \begin{cases} \frac{x_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x_i), & \lambda = 0 \end{cases}, \quad (3.3)$$

где x_i – исходные переменные с вектором значений $i = 1, 2, \dots, n$;

$x_i(\lambda)$ – преобразованные переменные с вектором значений $i = 1, 2, \dots, n$;

λ – параметр преобразования Бокса-Кокса.

Задача метода Бокса-Кокса состоит в нахождении оптимального значения параметра преобразования λ , при котором достигается максимум логарифма функции правдоподобия:

$$\ln L(\lambda) = -\frac{n}{2} \ln \left[\sum_{i=1}^n \frac{(x_i(\lambda) - \bar{x}(\lambda))^2}{n} \right] + (\lambda - 1) \cdot \sum_{i=1}^n \ln(x_i), \quad (3.4)$$

где $\sum_{i=1}^n \frac{(x_i(\lambda) - \bar{x}(\lambda))^2}{n}$ – оценка наибольшего правдоподобия;

$\bar{x}(\lambda) = \frac{1}{n} \sum_{i=1}^n x_i(\lambda)$ – среднеарифметическая преобразованных данных.

Значение параметра преобразования λ подбирают из диапазона $[-2; 2]$.

При рассмотрении количественных независимых переменных в модель также часто вводят их относительные преобразования. Примерами таких расчетных переменных могут быть отношение ежемесячных выплат по кредиту к среднему ежемесячному доходу заемщика, отношение месячного свободно располагаемого бюджета заемщика к ежемесячному доходу, отношение суммы задолженности к доходу и другие.

д) *Категоризация количественных независимых переменных.* Как было отмечено выше, при оценке кредитных рисков часто используют метод категоризации количественных независимых переменных, который позволяет добиться следующих основных преимуществ при построении модели:

- во-первых, облегчить обработку выбросов и экстремальных значений количественных переменных;
- во-вторых, упростить интерпретацию полученных результатов при использовании статистической модели логистической регрессии;
- в-третьих, отразить сложные нелинейные связи.

Расчет показателей WOE и IV по формулам (2.5) и (2.6). для количественных переменных требует их перевода в категориальную шкалу с использованием определенных методов. Эта процедура в целом получила название «биннинга» переменной (англ. Binning). Во многих пакетах прикладных программ для моделирования реализованы процедуры категоризации и оптимального биннинга количественных переменных, учитывающие распределение бинарной зависимой переменной, то есть такая категоризация является «оптимальной» по отношению к категориальной целевой переменной. Алгоритмы такого семейства носят название алгоритмов биннинга с учителем (англ. Supervised Binning). Разберем основные алгоритмы процедур категоризации количественных переменных, используемых при оценке кредитных рисков.

1) Алгоритм разбиения на основе анализа WOE. Первоначально количественная переменная разбивается на основе равных процентилей на несколько групп (может использоваться разбиение на основе от 10 до 50, а иногда и всех 100 процентилей). Затем в каждой группе просчитывается доля «плохих» и

«хороших» займов, а также показатели веса категорий предиктора WOE, рассчитываемые по формуле (2.5). Веса категорий предиктора позволяют найти по независимой переменной «границы чувствительности» к появлению моделируемого события риска и провести оптимальным образом категоризацию количественных переменных.

Далее полученные показатели весов категорий анализируются, происходит объединение соседних категорий и расчет показателей WOE повторяется. Пропущенные данные при категоризации могут кодироваться как отдельная категория и участвовать в дальнейшем анализе. При дальнейшем объединении категорий руководствуются следующими правилами:

- в каждой группе должно находиться не меньше 5% от всех валидных наблюдений переменной;
- не должно быть групп с количеством «плохих» или «хороших» займов, равным 0;
- процент «плохих» заемщиков и WOE должны в достаточной мере отличаться по получаемым группам;
- значения показателей WOE должны иметь возрастающий или убывающий тренд линейного или нелинейного характера при переходе от одной категории к другой.

При укрупнении категорий помимо статистических критериев следует руководствоваться логикой, целесообразностью и возможностью такого объединения. Также может быть проведено сравнение долей «плохих» заемщиков в соседних категориях независимой переменной z-тестом равенства долей или другими критериями для принятия решения о целесообразности укрупнения соседних категорий. В работах [89, 158] подробно описана процедура категоризации, используемая в пакете SAS.

Рассмотренная методика категоризации проиллюстрирована в Приложении Б на примере количественной переменной стажа работы при построении одной из модели оценки кредитных рисков.

2) Алгоритм MDLP-дискретизацию количественных переменных (англ. Minimal Description Length Principle, «принцип минимальной длины описания»). Метод делит переменную на небольшое число интервалов, где каждая область отображается в отдельную категорию полученной дискретной переменной. MDLP является одномерным методом дискретизации с учителем.

Для начала опишем алгоритм простой MDLP-дискретизации, представленный в работах [224, 226, 232]. Для этого введем ряд понятий, допущений и утверждений.

Пусть имеется k классов $C_1, \dots, C_k$ и $p(C_i, S)$ задает долю появлений класса C_i во множестве S всех пар наблюдаемых значений целевой переменной C и соответствующих им значений переменной-предиктора X . Тогда энтропия i -го класса $e(S)$ определяется следующим выражением:

$$e(S) = - \sum_{i=1}^k p(C_i, S) \log_2(p(C_i, S)). \quad (3.5)$$

Пусть для множества наблюдений S , непрерывной переменной-предиктора X и точки отсечения T , задающей границу между двумя интервалами-категориями (бинами), $S_1 \subset S$ является подмножеством наблюдений из множества S со значениями $X \leq T$ и $S_2 = S \setminus S_1$. Информационная энтропия $E(X, T, S)$ разбиения, вызываемого выбранным значением T , вычисляется как взвешенное среднее и задается следующим выражением:

$$E(X, T, S) = \frac{|S_1|}{|S|} e(S_1) + \frac{|S_2|}{|S|} e(S_2), \quad (3.6)$$

где $|S_i|$, $i = 1, 2$, $|S|$ – мощности соответствующих множеств.

Имея множество наблюдений S , непрерывную переменную X и точку отсечения T , можно вычислить показатель прироста информации для точки отсечения T :

$$g(X, T, S) = e(S) - E(X, T, S). \quad (3.7)$$

Разбиение множества S , $|S| = n$, вызываемое выбранным значением T , принимается тогда и только тогда, когда выполняется следующее условие:

$$g(X, T, S) > \frac{\log_2(n-1)}{n} + \frac{\Delta(X, T, S)}{n}, \quad (3.8)$$

где $\Delta(X, T, S) = \log_2(3^k - 2) - (k * e(S) - k_1 * e(S_1) - k_2 * e(S_2))$;

$k_i, i = 1, 2$ – число классов в подмножестве $S_i \subset S$.

Алгоритм бинарной дискретизации может быть описан следующим образом:

1) Вычислить $E(X, u_i, S)$ для каждого значения u_i из множества U всех уникальных пар наблюдений из S , для которых u_i и u_{i+1} не принадлежат одному и тому же классу. Пара наблюдений принадлежит данному классу, если во всех парах наблюдений с данным значением переменной-предиктора значения целевой переменной принадлежат данному классу.

2) Выбрать такую точку отсечения T , для которой значение $E(X, T, S)$ минимально среди всех возможных точек отсечения:

$$T = \underset{u_i}{\operatorname{argmin}} E(X, u_i, S). \quad (3.9)$$

Тогда алгоритм MDLP-отсечения может быть описан следующим образом:

1) для переменной-предиктора X , точки отсечения T , множеств U и S запустить алгоритм бинарной дискретизации;

2) вычислить информационный прирост $g(X, T, S)$;

3) если условие принятия разбиения выполняется, то необходимо:

- задать множество точек отсечения $T_X = T_X \cup T$;
- разбить множество U на подмножества U_1 и U_2 , множество S на подмножества S_1 и S_2 ;
- запустить алгоритм MDLP-отсечения для переменной-предиктора X , множества точек отсечения T_X , множеств U_1 и S_1 ;
- запустить алгоритм MDLP-отсечения для переменной-предиктора X , множества точек отсечения T_X , множеств U_2 и S_2 .

Здесь $S_1 \subset S$ – подмножество наблюдений в S , со значениями $X \leq T$, а $S_2 = S \setminus S_1$. U_1 и U_2 – множества всех уникальных пар наблюдений в S_1 и S_2 соответственно.

В итоге получаем алгоритм простой MDLP-дискретизации:

1) Упорядочить пары наблюдений в S , $|S| = n$, по значению переменной-предиктора X в возрастающем порядке.

2) Найти множество U всех уникальных пар наблюдений в S .

3) Задать $T_X = \emptyset$.

4) Запустить алгоритм MDLP-отсечения для переменной-предиктора X , множества точек отсечения T_X , множеств U и S .

5) Упорядочить элементы множества T_X по возрастанию их значений, получив искомое множество точек отсечения, задающих границы категорий для разбиения переменной X .

В случае если множество U является большим, вычислительные затраты на нахождение значений $E(X, u_i, S)$ для каждого u_i высоки. Для сокращения вычислительных затрат в целях уменьшения размера множества U и получения подмножества $U_r \subset U$ используется метод равномерного распределения частот без учителя. Тогда для нахождения итогового множества точек отсечения T_X запускается алгоритм MDLP-отсечения для переменной-предиктора X , множества точек отсечения T_X , множеств U_r и S .

Алгоритм равномерного распределения частот делит непрерывную переменную-предиктор X на m корзин (бинов), где каждая корзина содержит n/m наблюдений. Значение параметра m определяется исследователем, исходя из условия $1 < m < n$.

Алгоритм равномерного распределения частот может быть описан следующим образом:

1) Упорядочить элементы множества S , $|S| = n$, по значению переменной-предиктора X в возрастающем порядке.

2) Задать $U_r = \emptyset$.

3) Задать $j = 1$.

4) Эмпирическим путем выбрать значения $u_{p,i}$, которые обозначают $(\frac{i*n}{m} * 100)$ -е перцентили соответственно.

5) Задать $U_r = U_r \cup u_{p,i}; i = i + 1$.

6) Если $i \leq m$, перейти к шагу 4.

7) Из множества U_T исключить дубликаты, оставив только уникальные значения.

Тогда алгоритм гибридной MDLP-дискретизации можно представить в виде последовательности таких шагов:

1) Задать $U = \emptyset$.

2) Запустить алгоритм равномерного распределения частот для переменной-предиктора X , заданного значения параметра m , множеств U и S .

3) Задать $T_X = \emptyset$.

4) Запустить алгоритм MDLP-отсечения для переменной-предиктора X , множества точек отсечения T_X , множеств U и S .

5) Получить искомое множество точек отсечения, задающих границы категорий для разбиения переменной X .

Введем также такое понятие, как энтропия модели, которая является мерой прогностической точности переменной-предиктора X , разбитой на категории по целевой переменной C . При имеющемся множестве наблюдений S предположим, что для заданной переменной C переменная X разбивается на I категорий (бинов), где i -я категория включает множество значений предиктора X_i . Если $S_i \subset S$ – подмножество наблюдений из S с множеством значений предиктора X_i , энтропия модели задается следующим выражением:

$$E_m = \sum_{i=1}^I P(X_i) (-\sum_{j=1}^J P(C_j|A_i) \log_2 P(C_j|A_i)), \quad (3.10)$$

$$\text{где } P(A_i) = \frac{|S_i|}{|S|},$$

$$P(C_j|A_i) = \frac{P(C_j, A_i)}{P(A_i)} = P(C_j, S_i).$$

Иногда описанный алгоритм может приводить к выделению категорий с малым количеством наблюдений. Следующая процедура позволяет удалить такие псевдо-точки отсечения за один проход по всем выделенным категориям:

1) Предположим, что для заданной переменной алгоритм позволил выделить m_f точек отсечения и, как следствие, $m_f + 1$ категорию. Для категорий

$i = 2, \dots, m_f$ (со второй наименьшей по значению по вторую наибольшую по значению переменной-предиктора X), вычислить значение:

$$z_i = \frac{|b_i|}{\min(|b_{i-1}|, |b_{i+1}|)}, \quad (3.11)$$

где $|b_i|$ – число наблюдений в i -й категории.

2) Если полученное значение z_i меньше заданного исследователем порогового значения, то категория b_i считается «малонаселенной», то есть содержит недостаточное число наблюдений, чтобы считаться отдельной категорией, и объединяется с категорией b_{i-1} или b_{i+1} в зависимости от того, какая из них имеет наименьшее значение энтропии класса.

Приведем пример работы алгоритма простой MDLP-дискретизации (см. Приложение В).

е) *Диагностика и устранение мультиколлинеарности.* При использовании статистической модели оценки кредитных рисков, в частности логистической регрессии, одна из основных проблем ее построения – это наличие мультиколлинеарности. Этой проблеме при построении логистической регрессии уделяется недостаточно внимания, хотя она также актуальна при использовании количественных предикторов, как и для модели линейной множественной регрессии [133, 205, 239].

Первоначальный анализ наличия мультиколлинеарности может быть произведен на основе матрицы парных и частных корреляций между независимыми переменными. Однако парные коэффициенты корреляции не всегда могут показать наличие мультиколлинеарности. Полезен и анализ частных коэффициентов корреляции [320]. Для диагностики мультиколлинеарности часто используют показатели толерантности или допуска переменной, определяемой по формуле:

$$1 - R_i^2, \quad (3.12)$$

где R_i^2 – квадрат множественного коэффициента корреляции i -ой независимой переменной со всеми остальными предикторами.

Если толерантность переменной близка к 0, то значения данной переменной можно выразить через линейную комбинацию остальных независимых переменных.

Иногда вместо толерантности используют показатель, обратный ее величине, называемый коэффициентом или фактором увеличения дисперсии (англ. Variance Inflation Factor, VIF), вычисляемый по формуле:

$$VIF = \frac{1}{1-R_i^2}. \quad (3.13)$$

При наличии мультиколлинеарности дисперсия оценок параметров модели регрессии возрастает пропорционально данной величине, что делает их оценку нестабильной. Большое значение показателя VIF свидетельствует о наличии мультиколлинеарности. Многие исследователи указывают, что, если этот показатель больше 5, это свидетельствует о наличии мультиколлинеарности. Другое мнение, что этот показатель должен быть больше 10.

Существуют и другие менее распространенные методы анализа мультиколлинеарности: анализ собственных значений, частных коэффициентов корреляций, показателей обусловленности, определителя корреляционной матрицы и другие [175, 240, 377].

В случае наличия мультиколлинеарности необходимо найти оптимальные варианты исключения тесно коррелирующих переменных для построения модели. Другой метод борьбы с мультиколлинеарностью – это увеличение объема выборки. Следует отметить, что на больших выборках объемом в несколько тысяч и десятков тысяч заемщиков мультиколлинеарность не сильно искажает оценки параметров модели. Оценки параметров на больших выборках получаются статистически устойчивыми.

Существуют различные способы устранения мультиколлинеарности в модели логистической регрессии:

- включение в модель нелинейных преобразований от тесно связанных независимых переменных является одним из наиболее эффективных способов устранения мультиколлинеарности;

- использование модели «лассо»-регрессии [360];
- использование главных компонент как метода факторного анализа вместо исходных значений переменных является кардинальным методом устранения мультиколлинеарности [277, 279, 280, 345].

Метод главных компонент способ не часто используется при построении моделей оценки кредитных рисков из-за последующей сложности расчета скоринговых баллов на основе агрегированных показателей. Однако во многих случаях при построении моделей на малых выборках он является незаменимым инструментом, поэтому рассмотрим его более подробно.

Основоположителем метода главных компонент является К. Пирсон [332], который искал способ уменьшения размерности данных в многомерном статистическом анализе без потери существенного объема информации.

Эффективное сокращение количества переменных достигается путем получения ортогональных линейных комбинаций исходных переменных – главных компонент или ковариат, между которыми отсутствует корреляционная связь, что позволяет устранить проблему мультиколлинеарности [256].

Матрица линейной комбинации исходных переменных имеет следующий вид:

$$\begin{aligned}
 F_1 &= a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n \\
 F_2 &= a_{21}X_1 + a_{22}X_2 + \dots + a_{2n}X_n \\
 &\dots \\
 F_n &= a_{n1}X_1 + a_{n2}X_2 + \dots + a_{nn}X_n,
 \end{aligned}
 \tag{3.14}$$

где F_i – i -ая главная компонента;

X_i – i -ая исходная переменная (предиктор);

$a_{in} = \sqrt{\lambda_i} \cdot e_i$, где λ_i – собственное значение i -ой исходной переменной, e_i – собственный вектор i -ой исходной переменной ($\|e_i\| = 1$), $i = 1, 2, \dots, n$.

Таким образом, для получения главных компонент необходимо вычислить собственные значения и собственные вектора исходных переменных. Дисперсия i -ой главной компоненты равна собственному значению i -ой исходной переменной.

Дисперсия каждой линейной комбинации (компоненты X_i) имеет следующий вид [276]:

$$\text{var}(X_i) = a_i^T \cdot \Sigma a_i, \quad (3.15)$$

где $a_i = (a_{i1}, a_{i2}, \dots, a_{in})^T$.

Ковариация каждой линейной комбинации (компоненты X_i) имеет следующий вид [276]:

$$\text{cov}(X_i, X_j) = a_i^T \cdot \Sigma a_j, \quad (3.16)$$

где $a_j = (a_{1j}, a_{2j}, \dots, a_{nj})^T$.

Использование метода главных компонент гарантирует максимальную дисперсию между выбранными линейными комбинациями.

При этом «первая главная компонента имеет наибольшую возможную дисперсию (с учетом изменчивости данных), а каждая последующая компонента имеет наибольшую возможную дисперсию при ограничении ортогональности (нулевую ковариацию) с предыдущей компонентой» [287], что математически можно отобразить следующим образом:

$$\begin{cases} \text{var}(X_1) \rightarrow \max, & a_1^T \cdot a_1 = 1 \\ \text{var}(X_2) \rightarrow \max, & a_2^T \cdot a_2 = 1, a_1^T \cdot a_2 = 0 \\ \dots \\ \text{var}(X_n) \rightarrow \max, & a_n^T \cdot a_n = 1, a_{n-1}^T \cdot a_n = 0 \end{cases} \quad (3.17)$$

Для устранения проблемы потери существенного объема информации при использовании метода главных компонент кумулятивная доля дисперсии должна превышать по возможности 85%.

Заключительным этапом построения модели оценки кредитных рисков является проверка ее достоверности и апробация на реальных данных. О степени валидации модели говорит способность ее правильно классифицировать заемщиков, способность модели отличать «хороших» заемщиков от «плохих». Модель должна давать корректные прогнозы не только на обучающей

совокупности, но и на практике при ее применении. Критерии качества бинарной классификации модели подробно рассмотрены в параграфе 3 главы 2.

Таким образом, в рамках настоящего параграфа была представлена методика построения моделей оценки кредитного риска МФО с изложением методических подходов к формированию выборки данных, отбору предикторов в модель, выбора зависимой переменной и приведения алгоритмов категоризации независимых переменных, диагностики мультиколлинеарности с учетом специфики МФО.

3.2 Статистические модели кредитных рисков и особенности их применения в микрофинансовых организациях

Моделирование кредитного риска является важнейшей задачей микрофинансовой организации для ее успешного функционирования. Эффективность экономики МФО в большей степени определяют кредитные потери и затраты на привлечение клиентов (маркетинговые затраты). Как было показано в предыдущей главе для успешного построения модели оценки кредитных рисков МФО необходимо системно подойти к выбору источников информации для оценки кредитных рисков. Однако качество исходных данных и их полнота не в полной мере определяет качество получаемых моделей оценки кредитных рисков. На практике немаловажное значение играет соблюдение методологии моделирования.

Моделирование кредитных рисков МФО – это задача бинарной классификации, предсказания одного из двух состояний целевого показателя риска (NPL или FPD): выполнение обязательств по кредиту (категории «хороший» заемщик) или их невыполнение (категории «плохой» заемщик). Для решения задачи бинарной классификации могут быть использованы два принципиально разных подхода: использование классических статистических методов, уже достаточно хорошо зарекомендовавших себя в кредитном скоринге,

и использование относительно новых для финансовой сферы методов и алгоритмов машинного обучения, но имеющих в настоящее время большую популярность и распространенность их использования. Моделированию рисков, в частности кредитных рисков финансовых организаций, посвящено множество работ [203, 208, 292, 300, 343].

Выбор спецификации математической модели является частью методологии моделирования. Поэтому в данном разделе исследования будут подробно рассмотрены применяемые на практике спецификации основных математических моделей оценки кредитных рисков МФО, изложенные в работах автора [109, 113, 114]. Как было показано в параграфе 3 главы 1 перед подразделениями риск-менеджмента, риск-моделирования вопрос выбора и целесообразности использования простых интерпретируемых математических моделей или сложных, не интерпретируемых моделей на основе алгоритмов и методов машинного обучения, для оценки кредитных рисков остается открытым. Рассмотрим основные используемые статистические модели и особенности их применения в микрофинансовых компаниях.

Логистическая регрессия (англ. Logistic Regression) – это стандартный статистический метод оценки вероятности дефолта заемщика на основе исторических данных в кредитном скоринге, примененный в банковской сфере примерно с 70-х гг. XX века [150]. Данный метод моделирования дефолта используется и при оценке кредитного риска в микрофинансах.

Логистическая регрессия – это модель обобщенной линейной регрессии с бинарной зависимой переменной [266, 291]. Таким образом, логистическая регрессия является вариантом линейной регрессии [2], в рамках которой зависимая переменная является бинарной и представляет собой дихотомическую переменную, принимающую значение 0 или 1, независимо от величины предикторов, которые могут быть количественными и/или категориальными переменными, представленными в виде фиктивных (dummy) переменных [54]. Отличием логистической регрессии от линейной является возможность определения не прогнозируемого значения зависимой переменной в

количественной шкале, а вероятности нахождения в «определенной категории» (позитивный или негативный класс) зависимой переменной с учетом значений независимых переменных. Как было отмечено выше при моделировании кредитных рисков чаще всего при использовании логистической регрессии независимые количественные переменные категоризуют.

Модель логистической регрессии используется для прогнозирования вероятности наступления события путем сопоставления данных с логит-функцией и не требует в отличие от линейной регрессии соблюдение строгих предпосылок о нормальности, гетероскедастичности и автокорреляции, поскольку зависимая переменная является дихотомической (бинарной) [266]. Следует отметить, что вместо логистической регрессии может быть использована альтернативная модель бинарного выбора – пробит-регрессия [235], однако ее использование в кредитном скоринге встречается редко.

Логистическая регрессия представляет собой классификационную модель для оценки апостериорной вероятности принадлежности объекта к позитивному или негативному классу в конкретных условиях бинарной классификации. Предполагаемая вероятность наступления события (в оценке кредитного риска – вероятность дефолта заемщика, «плохие», дефолтные заемщики) определяется как $P(Y_i = 1|x_i)$, а вероятность ненаступления такого события (в оценке кредитного риска – отсутствие вероятности дефолта заемщика, «хорошие», платежеспособные заемщики) – $P(Y_i = 0|x_i) = 1 - P(Y_i = 1|x_i)$.

Основным понятием логистической регрессии является логит – натуральный логарифм шанса от вероятности p_i наступления дефолта по кредиту: $\ln\left(\frac{p_i}{1-p_i}\right)$. Модель логистической регрессии может иметь две формы записи.

а) Форма, выражающая зависимость значения натурального логарифма шанса дефолта (логита) от предикторов, является линейной функцией и имеет следующий вид:

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = z_i, \quad (3.18)$$

где $\hat{p}_i$ – апостериорная вероятность того, что элемент i имеет модальность 1 для заданных значений, принимаемых независимыми переменными (т. е. $P(Y_i = 1|x_i)$);

z_i – линейная комбинация независимых переменных и константы:

$$z_i = b_0 + \sum_{j=1}^k b_j x_i = b_0 + b_1 x_i^{(1)} + b_2 x_i^{(2)} + \dots + b_k x_i^{(k)}; \quad (3.19)$$

где $x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)})$ – значения независимых переменных;

Y_i – значения зависимой переменной (0 или 1);

b_0 – свободный коэффициент (константа).

$b_j = (b_1, b_2, \dots, b_k)$ – коэффициенты при независимых переменных.

б) Экспоненциальная форма, выражающая зависимость вероятности наступления моделируемого события (дефолта) от предикторов, имеет следующий вид:

$$\hat{p}_i = \frac{1}{1+e^{-z_i}} = \frac{1}{1+e^{-(b_0+b_1x_i^{(1)}+b_2x_i^{(2)}+\dots+b_kx_i^{(k)})}}. \quad (3.20)$$

Экспоненциальная форма записи получается из выражения оценки вероятности p_i непосредственно из уравнения (3.18). В уравнениях (3.18) и (3.19) в левой части записана именно оценка вероятности дефолта, поэтому в правой части отсутствует случайная составляющая. Форма записи модели логистической регрессии через логит используется при оценке ее параметров, а экспоненциальная форма чаще всего используется риск-аналитиками при содержательной интерпретации коэффициентов полученной модели.

Популярность использования логистической регрессии для оценки кредитных рисков в кредитных учреждениях объясняется простотой интерпретации получаемых параметров, т. е. для риск-аналитиков, андеррайтеров, риск-менеджеров модель не представляет собой «черный ящик». В логит-моделях коэффициенты при независимых переменных (предикторах) соответствуют их предельному эффекту влияния на зависимую переменную. Коэффициенты b_j логистической регрессии показывают, насколько в среднем изменится логарифм шанса при изменении независимой переменной на единицу своего измерения при

неизменности остальных предикторов. Экспоненциальная форма записи коэффициентов модели логистической регрессии показывает относительную меру риска при изменении на единицу своего измерения количественных предикторов или относительную меру риска для категории фиктивной переменной по сравнению с опорной сравниваемой категорией для качественных предикторов. При категориальных независимых переменных экспоненциальные коэффициенты будут интерпретироваться как соотношения шансов в аналогичной таблице сопряженности между зависимой категориальной переменной риска и независимой категориальной переменной [260].

Основными предположениями логистической регрессии являются независимость между последовательными наблюдениями и существование линейной связи между логитом и предикторами (x_i). Ключевым критерием успешности применения модели логистической регрессии является определение возможности изменения связи или направления между вероятностью события и независимыми переменными [317].

Для оценки параметров b_j модели логистической регрессии по обучающей выборке значений зависимой переменной Y_i и независимых переменных x_i применяют итеративный метод максимального правдоподобия [157], который предполагает, что каждый параметр выборки в наборе данных является независимым. Функцию правдоподобия можно записать в следующем виде:

$$\begin{aligned} L(y_1, y_2, \dots, y_n | x_1, x_2, \dots, x_n; b_0, b_1, \dots, b_k) &= \\ &= \prod_{i: y_i=1} \Lambda(z_i) \cdot \prod_{i: y_i=0} (1 - \Lambda(z_i)) = \prod_{i=1}^n \Lambda(z_i)^{y_i} (1 - \Lambda(z_i))^{1-y_i}, \end{aligned} \quad (3.21)$$

где $\Lambda(z_i) = \frac{1}{1+e^{-z_i}}$ – логистическая функция (сигмоида или логит-функция).

Оценка параметров определяется как точка максимума функции логарифма правдоподобия:

$$\ln(L) = \sum_{i=1}^n (y_i \ln(\Lambda(z_i)) + (1 - y_i) \ln(1 - \Lambda(z_i))). \quad (3.22)$$

Максимизация функции правдоподобия на практике происходит в статистических пакетах одним из следующих методов: метод градиентного спуска, стохастического градиентного спуска, метода Ньютона-Рафсона. Также на

практике для уменьшения эффекта переобучения модели часто используют логистическую регрессию с регуляризацией. В этом случае для нахождения параметров модели вместо метода максимального правдоподобия используется метод максимизации апостериорной оценки.

Свободный коэффициент b_0 (константа) определяет нормальную величину риска наступления дефолта по кредиту при условии равенства нулю всех независимых переменных, т. е. отражает средний базовый уровень риска по портфелю МФО. Метод максимального правдоподобия [349] предполагает, что функция L должна быть максимизирована по всем возможным значениям коэффициентов b_j . В идеальной теоретической модели, когда истинное значение $y_i = 1$, вероятность принадлежности параметра выборки к позитивному классу p_i также должна быть равна 1, то есть чем больше значение p_i для дефолтных заемщиков, тем меньше будет разница между прогнозируемым и истинным значением зависимой переменной. В случае, если истинное значение $y_i = 0$, вероятность принадлежности параметра выборки к негативному классу $(1 - p_i)$ также должна быть равна 0, то есть чем больше значение $(1 - p_i)$, тем меньше будет разница между прогнозируемым и истинным значением параметра выборки [250]. Метод максимального правдоподобия при оценке параметров логистической регрессии гарантирует наименьшую разницу между прогнозируемым и истинным значением параметра выборки.

Оценка качества модели логистической регрессии может быть проанализирована посредством различных статистических тестов и метрик. По мнению автора, минимальный анализ качества модели логистической регрессии должен включать следующие шаги: тест на значимость модели (значение функции правдоподобия, LR-тест), оценка качества подгонки модели (значение псевдо-коэффициентов детерминации, тест Хосмера-Лемешева), оценка значимости параметров модели (тест Вальда).

Качественная модель логистической регрессии определяется функцией правдоподобия близкой к 1. Для статистической значимости модели в целом

проводится тест отношений функций правдоподобий (LR-тест). Первоначально рассчитывают значение логарифма функции правдоподобия для модели только с константой:

$$\ln L(0) = n \left(\hat{p}_0 \ln \left(\frac{\hat{p}_0}{1 - \hat{p}_0} \right) + \ln(1 - \hat{p}_0) \right), \quad (3.23)$$

$$\text{где } \hat{p}_0 = \frac{\sum_{i=1}^n y_i}{n}. \quad (3.24)$$

После исходя из максимизации выражения 3.22 рассчитывается значение функции правдоподобия для модели логистической регрессии с отобранными предикторами $\ln L(1)$.

Для проверки статистической значимости отличий этих двух функции правдоподобия рассчитывают статистику, которая имеет распределение хи-квадрат с числом степеней свободы, представляющие собой разницу числа параметров сравниваемых двух моделей:

$$LR = -2 \ln \left(\frac{L(0)}{L(1)} \right) = -2(\ln L(0) - \ln L(1)). \quad (3.25)$$

Отметим, что LR-тестом можно сравнивать значимость эффекта включения или исключения одной или блока переменных при применении пошаговых процедур построения логистической регрессии.

Как и в классической линейной регрессии после оценки значимости уравнения логистической регрессии в целом переходят к анализу значимости отдельных коэффициентов. Тестирование влияния отдельных предикторов на изменение зависимой переменной по уровню значимости выполняют на основе статистики Вальда, вычисляемой по формуле [261]:

$$W_j = \frac{\hat{b}_j}{\sigma_{\hat{b}_j}^2}, \quad (3.26)$$

где $\hat{b}_j$ – оценки параметров, полученные методом максимального правдоподобия;

$\sigma_{\hat{b}_j}^2$ – значение стандартной ошибки полученных оценок параметров.

Статистика Вальда имеет распределение хи-квадрат с числом степеней свободы, равным числу оцениваемых параметров в модели. Предикторы с

незначимыми параметрами должны удаляться из моделей на основе пошаговых процедур построения регрессии [273]. Модель с включением переменных со статистически незначимыми параметрами будет не стабильна при использовании на практике для оценки кредитных рисков. Порог статистической значимости (р-значение) от 0,05 до 0,1 говорит о наличии тенденции во влиянии переменной и может быть допустим в определенных случаях. Конкретный выбор уровня статистической значимости очень сильно зависит от объема выборки ($p < 0,001$; $p < 0,01$; $p < 0,05$). Если в модели логистической регрессии присутствует в качестве независимой категориальная переменная в виде нескольких фиктивных переменных, то статистически значимыми должны быть коэффициенты при всех переменных-флагах. Если уровень значимости (р-значение) при какой-то категории выше порогового, то это означает, что уровень кредитного риска у данной категории примерно такой же, как и у опорной категории, и их можно объединить в одну группу: необходимо сделать перегруппировку значений категориальной независимой переменной.

При проверке значимости коэффициентов рекомендуется принимать в расчет не только абсолютные значения стандартных ошибок коэффициентов, но и их доверительные интервалы. Наиболее предпочтительной является модель, которая имеет наименьшее значение стандартной ошибки и более узкие доверительные интервалы коэффициентов при независимых переменных.

После получения значимой модели со всеми значимыми параметрами можно перейти к анализу качества подгонки модели. По аналогии с линейной регрессией для этой цели можно использовать коэффициент детерминации. Однако коэффициент детерминации не может быть рассмотрен как основной признак качества модели логистической регрессии в отличие от линейной регрессии ввиду невозможности выдвинуть предположение о постоянстве дисперсии из-за ее зависимости от частоты распределения значений бинарной переменной, что отличает логистическую регрессию от линейной. Коэффициент детерминации в логистической регрессии является приближенной мерой качества

подгонки модели, поэтому его часто называют псевдо-коэффициентом детерминации.

Существуют несколько альтернативных подходов определения коэффициента детерминации в логистической регрессии:

- 1) Коэффициент качества подгонки (Goodness-of-Fit) [266]:

$$GF = \sum_{i=1}^n \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)}. \quad (3.27)$$

- 2) Коэффициент детерминации Кокса и Снелла [210]:

$$R_{CS}^2 = 1 - \left(\frac{L(0)}{L(1)} \right)^{\frac{2}{n}}. \quad (3.28)$$

- 3) Коэффициент детерминации Найджелкерка [321]:

$$R_N^2 = R_{CS}^2 / \max(R_{CS}^2), \quad (3.29)$$

$$\text{где } \max(R_{CS}^2) = 1 - (L(0))^{\frac{2}{n}}. \quad (3.30)$$

Еще один метод оценки качества подгонки модели логистической регрессии – тест Хосмера-Лемешева [266]. Данный тест рассчитывается на основе таблицы сопряженности размерностью 2 на g . Число строк таблицы сопряженности определяется числом категорий зависимой переменной y_i , а число столбцов образуется значениями предсказанных вероятностей дефолта $\hat{p}_i$, разбитых на g равных по количеству наблюдений групп. Как правило, количество групп предсказания вероятности дефолта составляет 10 ($g = 10$), а данные группы называют децилями риска.

Тест Хосмера-Лемешева позволяет проверить гипотезу о соответствии наблюдаемых и ожидаемых (прогнозных) значений вероятностей дефолта, фактически представляет собой тест независимости хи-квадрат Пирсона. Основой теста является расчет следующей статистики:

$$\chi_{HL}^2 = \sum_{i=1}^g \frac{(O_{1i} - E_{1i})^2}{E_{1i}(1 - \bar{p}_i)}, \quad (3.31)$$

где g – общее количество групп;

O_{1i} – общее количество наблюдений в i -группе при $y = 1$;

E_{1i} – количество ожидаемых частот события в i -группе, вычисляемое по формуле:

$$E_{1i} = \sum_{q_1}^{q_i} m_j \cdot \bar{p}_i;$$

где $\bar{p}_i$ – средняя прогнозируемая вероятность события в i -группе;

q_1, q_n – блоки наблюдений;

m_j – количество наблюдений в q -блоке.

Статистика Хосмера-Лемешева (χ_{HL}^2) подчиняется распределению хи-квадрат Пирсона со степенью свободы, равной $(g - 2)$.

Тест Хосмера-Лемешева рекомендуется применять при построении модели логистической регрессии на основе предикторов, которые измеряются в непрерывной шкале, а также при небольшом объеме обучающей выборки.

Одна из проблем, которая возникает при применении модели логистической регрессии для оценки кредитных рисков, связана с низкой степенью точности предсказания отрицательных исходов, т. е. дефолтов, по причине недостаточного числа «плохих» заемщиков в обучающей выборке. На практике обычно строят скоринговые модели на основе выборки, измеряемой тысячами наблюдений, с равным числом «положительных» и «отрицательных» исходов зависимой переменной. Разработчики моделей оценки кредитных рисков на коммерческой основе, как правило, запрашивают выборку минимум из 4500 клиентов: 1500 «хороших», 1500 «плохих» и 1500 «отклоненных». Как правило, объем выборки составляет около 2000 «плохих» и 2000 «хороших» заемщиков. Зная истинные пропорции «плохих» и «хороших» заемщиков в генеральной совокупности и необходимый объем выборки, можно изменить распределение зависимой переменной с помощью методов перевзвешивания и прореживания выборки.

Первая стратегия корректировки распределения зависимой переменной при малом количестве «плохих» заемщиков и большом объеме выборки может быть следующей. Можно взять 100% всех «плохих» заемщиков и случайным образом отобрать часть «хороших» наблюдений. После реализации такого алгоритма доля «плохих» к «хорошим» наблюдениям может составлять от 0,1 до 0,5.

Вторая стратегия заключается в перевзвешивании данных, чтобы добиться нужного соотношения «плохих» и «хороших» наблюдений в выборке для построения модели. Такая стратегия более предпочтительна при небольшом объеме выборки. Перевзвешивание данных, в отличие от корректировки выборки за счет случайного отбора, дает, как правило, более надежные оценки параметров модели. Однако при таком подходе следует избегать двух основных ошибок, которые могут привести к смещению коэффициентов логистической регрессии:

- во-первых, если брать часть «хороших» наблюдений, а не все, то их следует отбирать на основе механизма случайного отбора, чтобы наблюдения были независимыми. Например, отобрав каждое третье «хорошее» наблюдение в данных, мы рискуем получить зависимые наблюдения;

- во-вторых, «плохие» и «хорошие» наблюдения должны быть сформированы по одной методике и на основе одинаковых критериев отбора данных, рассмотренных выше.

После корректировки выборки и оценки параметров модели логистической регрессии ее способность классифицировать редкие события улучшится. Коэффициенты при независимых переменных останутся практически неизменными. При применении логистической регрессии используются для разработки в дальнейшем скоринговой карты значения коэффициентов при независимых переменных.

Меняя пропорции «плохих» и «хороших» заемщиков в выборке при построении модели логистической регрессии, будет меняться оценка значения константы модели.

Значение константы в модели логистической регрессии зависит от распределения в данных по категориям зависимой переменной. В случае перевзвешивания выборки для изменения этого распределения для более адекватной последующей оценки качества полученной модели константу корректируют и получают следующую модель логистической регрессии:

$$\ln\left(\frac{p_i^*}{1-p_i^*}\right) = \ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right) + b_0 + b_1x_i^{(1)} + b_2x_i^{(2)} + \dots + b_kx_i^{(k)}, \quad (3.32)$$

где p_i^* – откорректированная априорная вероятность;

ρ_0 и ρ_1 – доли «хороших» и «плохих» заемщиков в выборке;

π_0 и π_1 – доли «хороших» и «плохих» заемщиков в генеральной совокупности.

Исходя из модели (3.32) могут быть откорректированы значения прогнозируемых апостериорных вероятностей дефолта и получены, таким образом априорные вероятности для генеральной совокупности [358]. Однако для построения скоринговой карты достаточно значений коэффициентов при независимых переменных, которые при преобразовании (3.32) остаются неизменными. Для оценки правильной вероятности наступления моделируемого события риска в случае перевзвешивания данных значение константы корректируется.

В таблице 3.1 приведены результаты эмпирических расчетов коэффициентов логистической регрессии для условных переменных для иллюстрации рассмотренных методик формирования обучающей выборки.

Изначально по выборке 700 наблюдений с долей «плохих» наблюдений 26% и «хороших» 74% точность предсказания «плохих» наблюдений составила всего 50%. Было реализовано два альтернативных подхода формирования выборки.

В первом были взяты все «плохие» наблюдения и случайно отобрано 50% «хороших» наблюдений. Таким образом, соотношение «плохих» и «хороших» заемщиков составило 42% и 58% соответственно. Объем откорректированной выборки составил 434 наблюдений. Это позволило увеличить процент верно предсказанных «плохих» наблюдений до 71%. При этом доля верного предсказания «хороших» наблюдений составила 82%.

Во втором подходе все данные были перевзвешены, чтобы соотношение «плохих» – «хороших» в выборке было также 42 и 58% процентов. Число наблюдений в выборке сохранилось и составило 700. Это привело практически к идентичным результатам по сравнению с первой методикой.

Таблица 3.1 – Сравнение точности классификации при различных стратегиях формирования обучающей выборки

Показатель	Исходная модель	Отобраны все «плохие» и случайным образом 50% «хороших»	Исходные данные перевзвешены, чтобы увеличить долю плохих до 42%
Количество «плохих»	183	183	295
Доля плохих	0,26	0,42	0,42
Количество «хороших»	517	251	405
Доля «хороших»	0,74	0,58	0,58
Объем выборки	700	434	700
<i>Коэффициенты модели</i>			
Переменная 1*	0,57	0,59	0,56
Переменная 2*	– 0,08	– 0,07	– 0,10
Переменная 3*	– 0,24	– 0,24	– 0,26
Переменная 4*	0,09	0,09	0,09
Константа	– 0,79	– 0,22	– 0,95
<i>Классификация</i>			
Процент корректных «хороших»	92,5	82,1	79,9
Процент корректных «плохих»	50,3	70,5	73,8
Общий процент корректных	81,4	77,2	77,3
примечание – конкретные названия переменных не указываются, данные условные			

Источник: составлено автором

Рассмотрев основные особенности применения логистической регрессии для оценки кредитных рисков в МФО, перейдем к рассмотрению основных ее преимуществ.

Применение модели логистической регрессии в оценке кредитных рисков достаточно широко распространено в научных и эмпирических исследованиях, что подтверждает представленный ниже литературный обзор.

Исследование [155] включает разработку модели логистической регрессии для прогнозирования кредитоспособности клиентов банков ФРГ с использованием 20 различных предикторов (7 количественных и 13 категориальных) с апробацией на 1000 банковских заемщиках по данным German

Credit dataset. Уровень достоверности всех используемых в модели переменных-предикторов составил 95%, что определяет их важность при оценке кредитоспособности заемщика.

В статье [146] разработана модель логистической регрессии с целью прогнозирования риска дефолта по потребительским кредитам в банковских организациях Португалии. В рамках модели в качестве предикторов были выбраны 8 количественных (включая, персональные процентные ставки, сроки кредитования, количество кредитных карт и т. д.) и 6 качественных показателей – бинарные переменные (наличие других кредитов, наличие дефолта заемщика, получение зарплаты в соответствующем банке и т. д.). Апробация модели проведена на 3221 заемщике, 319 из которых являются дефолтными. Предикативная точность модели составила 89,8%, что подтверждено тестом Хосмера-Лемешева.

Работа [160] посвящена анализу кредитного риска потенциальных заемщиков финансовых кооперативов Индонезии при помощи построения модели логистической регрессии на основе восьми факторов. Для оценки качества разработанной модели авторы провели тест Вальда, направленный на выявление значимых факторов. Данный тест показал, что шесть из восьми выбранных параметров (независимых переменных) оказывают значительное влияние на риск дефолта – возраст должников, количество иждивенцев, сумма сбережений, стоимость залога, кредитный лимит и срок кредита. Кроме того, авторы использовали тест Хосмера-Лемешева для сопоставления наблюдаемых значений выборки с прогнозными значениями параметров, который позволил сделать вывод о достаточно высокой предиктивной способности построенной модели логистической регрессии.

В исследовании [347] была проведена оценка эффективности прогнозирования рисков потребительского кредитования в финансовых организациях Перу в рамках модели логистической регрессии на выборке из пяти тысяч кредитных историй, включая следующие показатели: количество дней просрочки, сформированные резервы, количество запланированных и

непогашенных платежей. Авторы пришли к выводу о том, что построенная модель логистической регрессии обеспечила точность на 99%.

В статье [383] представлена модель кредитного скоринга, базой которой выступает логистическая регрессия. Модель позволяет выявить «хороших» и «плохих» заемщиков потребительских кредитов на основе их уникальной информации. Разработанная модель апробирована как в банках, так и в микрофинансовых организациях Китая.

В работе [312] автор использовал модель логистической регрессии для разработки методики идентификации факторов, оказывающих влияние на платежную дисциплину заемщиков микрофинансовых организаций Туниса. Автор доказал существенное влияние на уровень дефолта заемщика микрофинансовых организаций социально-экономических характеристик заемщиков, общей суммы кредитов, периода погашения и прошлый опыт в получении кредитов в микрофинансовых организациях.

Таким образом, проведенный анализ литературы показывает, что модель логистической регрессии доказала свою высокую интерпретируемость и предикативную точность для оценки кредитных рисков как в банках, так и в микрофинансовых организациях. Можно сделать выводы, что модель логистической регрессии обладает следующими достоинствами по сравнению с другими классическими методами статистического моделирования:

- во-первых, в отличие от множественной линейной регрессии и дискриминантного анализа не требуется нормальность распределения независимых переменных;
- во-вторых, в отличие от множественной линейной регрессии не требуется проверка гипотезы о нормальном распределении остатков в модели с постоянной дисперсией;
- в-третьих, может отражать нелинейные зависимости между зависимой и независимыми переменными, за счет категоризации количественных предикторов;

- в-четвертых, на основе модели логистической регрессии можно получить однозначный набор интерпретируемых предикторов, влияющих на появление моделируемого события риска, и оценку важности их влияния;

- в-пятых, алгоритмы оценки параметров логистической регрессии реализованы во всех статистических пакетах для обработки данных и моделирования;

- в-шестых, в модель можно включать в качестве предикторов и количественные и категориальные переменные, в кредитном скоринге обычно количественные переменные категоризуют, как уже было отмечено, для отражения нелинейных зависимостей.

Но модель логистической регрессии имеет и недостатки:

- во-первых, для успешного построения модели требуется достаточная статистика как с положительными исходами зависимой переменной, так и с отрицательными, хотя эта проблема обычно решается перевзвешиванием выборки;

- во-вторых, на качество логит-модели, как и при построении и линейной регрессии, влияет отсутствие или наличие явной мультиколлинеарности между независимыми переменными.

Способы решения данных проблемы мультиколлинеарности рассмотрены в параграфе 1 настоящей главы. Отметим, что существуют научные исследования, направленные на устранение основных недостатков и ограничений логистической регрессии в задачах оценки кредитных рисков. Так, например, в работе [163] была предложена модель кусочной логистической регрессии, основанной на допущении о наличии нелинейных связей между предикторами и зависимой переменной, которая показала улучшенную производительность по сравнению со стандартной логистической регрессией в оценке кредитных рисков. Исследование [357] направлено на разработку модели логистической регрессии со штрафами (поправочными коэффициентами) для повышения прогностической эффективности и интерпретируемости модели оценки кредитных рисков за счет

выявления наиболее релевантных признаков дефолта заемщиков с целью решения проблемы выбора переменных.

Следующей распространенной моделью статистической оценки кредитных рисков выступает *дискриминантный анализ* (англ. Linear Discriminant Analysis), который представляет собой многомерный метод анализа связи между одной категориальной зависимой переменной и набором метрических (нормально распределенных) независимых переменных [165, 166, 225, 351]. Дискриминантный анализ впервые применил его основоположник Р. Фишер [237], который преследовал цель найти линейную комбинацию предикторов, показывающую наибольшую разницу в групповых средних значениях относительно внутригрупповой дисперсии.

Самым первым применением модели множественного дискриминантного анализа в рамках оценки кредитных рисков является работа Д. Дюрана [227]. В данной работе автор провел анализ «плохих» и «хороших» заемщиков потребительских кредитов в банковских учреждениях США на основе сформированной модели кредитного рейтинга в рамках дискриминантного анализа, включающей переменные-предикторы по финансовым и нефинансовым показателям, извлеченным из заявок потенциальных заемщиков.

Смысл дискриминантного анализа заключается в том, чтобы на основании обучающей выборки преобразовать несколько переменных в одномерный показатель для прогнозирования (дискриминации) принадлежности наблюдений к группам. Таким образом, целью дискриминантного анализа является прогнозирование принадлежности наблюдений к группам или классам.

Зависимая переменная отклика в дискриминантном анализе может быть дихотомической, номинальной или порядковой. В частности, как и в логистической регрессии, часто в кредитном скоринге используют бинарную зависимую переменную. Независимые (дискриминантные) переменные в обязательном порядке должны быть количественными и не должны сильно коррелировать между собой. Дискриминантные переменные внутри группы

переменной отклика должны подчиняться многомерному нормальному закону распределения.

В оценке кредитных рисков дискриминантный анализ заключается в нахождении средневзвешенного значения нескольких коэффициентов (дискриминантной функции), рассчитанного для каждого заемщика, которое наилучшим образом обеспечивает различие между «хорошими» и «плохими» заемщиками, т. е. зависимую переменную рассматривают, как и в логистической регрессии, в виде бинарного показателя.

Дискриминантный анализ представляет собой сочетание методов линейной регрессии и компонентного анализа. Математически задачи дискриминантного анализа сводятся к следующему:

- 1) нахождение дискриминантной функции – определение правила отнесения наблюдения к группе;
- 2) оценка качества построенной дискриминантной функции;
- 3) оценка роли переменных, составляющих дискриминантную функцию;
- 4) классификация новых наблюдений с помощью найденной дискриминантной функции.

Дискриминантный анализ имеет две модели – линейную (линейный дискриминантный анализ, LDA) и квадратичную (квадратичный дискриминантный анализ, QDA) [221].

Линейная модель дискриминантного анализа имеет следующий вид:

$$D_j = a_{j0} + b_{j1}x_1 + b_{j2}x_2 + \dots + b_{jp}x_p, \quad (3.33)$$

где D_j – зависимая (группирующая) переменная, дискриминантная функция;

j – индикатор номера группы ($j = 1, 2, \dots, g$);

a_{j0} – свободный коэффициент (константа);

b_{ji} – коэффициент i -ой независимой переменной для j -ой группы;

i – индикатор дискриминантной переменной ($i = 1, 2, \dots, p$);

x_i – независимые (дискриминантные) переменные.

Дискриминантная функция – это линейная комбинация независимых переменных, которая обеспечивает наилучшее разделение наблюдений на группы и строится по обучающей выборке – множество наблюдений, с заданными значениями признаков и принадлежность которых к тому или иному классу уже достоверно известна [367].

Оценка качества построения дискриминантной функции заключается в правильности интерпретации возможных различий между группами. Оценка роли составляющих дискриминантной функции заключается в сравнении матриц общих дисперсий и ковариаций с помощью многомерного F-критерия.

Коэффициенты предикторов b_{ji} максимизируют расстояние между средними значениями критериальной (зависимой) переменной. При этом «хорошие» предикторы, как правило, имеют большие веса, т. е. значения коэффициентов b_{ji} . Для вычисления коэффициентов b_{ji} в рамках построения дискриминантной функции также используется метод максимального правдоподобия:

а) если размер группы одинаковый, пороговым значением является средний балл;

б) если размер группы неодинаков, пороговое значение рассчитывается на основе взвешенных средних.

В рамках построения модели дискриминантного анализа формируется уравнение, которое позволяет свести к минимуму возможность неправильной классификации случаев по соответствующим группам или категориям.

Основным недостатком линейного дискриминантного анализа Фишера является то, что, если предположение о нормальности не выполняется, теряется оптимальность решения.

Квадратичная модель дискриминантного анализа имеет следующий вид [206]:

$$\delta_i(x) = \log \pi_i - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\vec{x} - \vec{\mu}_i)^T \cdot \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i), \quad (3.34)$$

где π_i – априорная вероятность i -го класса;

Σ_i – ковариационная матрица для i -го класса;

$\vec{x}$ – множество значений признаков;

$\vec{\mu}_i$ – вектор средних значений признаков для i -го класса.

В рамках LDA данные должны быть независимыми и нормально распределенными, а ковариационная матрица должна соответствовать предположению об однородности вариаций [338]. Если ковариационные матрицы групп данных не равны, то необходимо использовать QDA, который позволяет каждой группе иметь свою собственную ковариационную матрицу. Необходимо отметить, что LDA более надежен и прост в реализации, но QDA может фиксировать более сложные связи.

Оценка качества модели дискриминантного анализа реализуется посредством различных тестов, базовым из которых является тест лямбда-статистики Уилкса.

Лямбда-статистика Уилкса выступает ключевым критерием качества дискриминационной функции, т. е. «остаточной дискриминационной способности – способности различать группы при исключении данных, полученных с помощью ранее вычисленных функций, расчет дискриминантной функции не имеет смысла при малой остаточной дискриминационной способности» [131]. Лямбда-статистика Уилкса вычисляется по формуле:

$$\Lambda = \prod_{i=k+1}^g \frac{1}{1+\lambda_i}, \quad (3.35)$$

где g – количество групп;

k – количество вычисленных дискриминантных функций;

$\lambda_i = \frac{B(d)}{W(d)}$ – собственное значение дискриминантной функции, вычисляемое

как отношение межгрупповой матрицы рассеяния наблюдаемых переменных от средних $B(d)$ к внутригрупповой матрице $W(d)$.

Для интерпретации результатов на основе лямбда-статистики Уилкса (Λ) необходимо принимать во внимание следующие допущения:

- если критерий Λ стремится к 1, то функция менее дискриминанта;

- если критерий Λ стремится к 0, то дискриминационные способности функции возрастают.

Лямбда-статистика Уилкса подчиняется распределению хи-квадрат со степенями свободы, равными $(p - k) \cdot (g - k - 1)$, где p – число дискриминантных переменных.

Лямбда-статистика Уилкса применяется для проверки гипотезы: $\mu_1 = \mu_2 = \dots = \mu_i = 0$ – средние значения признаков каждой из групп равны между собой, при альтернативной гипотезе $\mu_1 \neq \mu_2 \neq \dots \neq \mu_k \neq 0$ – средние значения признаков каждой из групп не равны между собой.

При этом переменная, не являющаяся значимой при одномерном дискриминантном анализе, может оказаться статистически значимой при многомерном дискриминантном анализе.

Отличия дискриминантного анализа от логистической регрессии заключается в том, что независимые переменные должны быть измерены в количественной шкале, в то время как при использовании логистической регрессии независимые переменные могут быть количественными и категориальными с автоматической заменой их на фиктивные переменные. Дискриминантный анализ требует выполнения более строгих предположений по независимым переменным, чем логистическая регрессия (например, предположение о многомерном нормальном распределении независимых переменных и равенстве ковариационных матриц), а применение логистической регрессии связано с меньшим количеством ограничений. В дискриминантном анализе для классификации будущих наблюдений используется дискриминантная функция – линейная комбинация взаимосвязи результативного признака и независимых переменных; а логистическая регрессия представляет собой S-образную функцию взаимосвязи результативной переменной и независимых переменных.

Преимущества дискриминантного анализа заключаются в возможности обработки как бинарных, так и многомерных групп (в отличие от классической логистической регрессии). Хотя для зависимой переменной с числом групп

больше двух может использоваться и обобщение логистической регрессии: полиномиальная логистическая регрессия. Однако при наличии трех и более категорий зависимой переменной более предпочтительным является дискриминантный анализ. При этом главный недостаток метода – чувствительность к распределению исходных данных. Более того, недостаточная точность классификации в рамках дискриминантного анализа может быть обусловлена выявлением линейных связей между переменными.

В оценке кредитных рисков индивидуальных заемщиков дискриминантный анализ на эмпирическом уровне не получил такого распространения как логистическая регрессия. В большинстве исследований проводится сравнительный анализ данных методов статистического моделирования.

Так, в исследовании [308] автор провел оценку кредитоспособности 9045 физических лиц – держателей кредитных карт Кенийских коммерческих банков, при этом на основе ретроспективных данных было определено 8340 заемщиков с полным отсутствием дефолтной информации в кредитной истории, а 705 заемщиков имели просрочку платежа свыше 90 дней. Для подтверждения результатов была построена модель логистической регрессии, уровень точности которой составил 92,4%, и дискриминантного анализа с уровнем точности 86,3%, что позволило автору сделать вывод о большей эффективности использования логистической регрессии в оценке кредитных рисков.

Статья [215] посвящена сравнительному анализу моделей логистической регрессии и дискриминантного анализа в оценке кредитных рисков по потребительским кредитам в банковских учреждениях Пакистана. Анализ проведен на основе данных по 800 действующим потребительским кредитам – 600 «хороших» и 200 «плохих». Для обеих моделей было выбрано двадцать независимых переменных по индивидуальным характеристикам заемщиков, а зависимая переменная представляет собой показатель «качество кредита», который определяется категориальными значениями – 1 («хороший» кредит, платежеспособный заемщик) и 0 («плохой» кредит, дефолтный заемщик). Точность модели логистической регрессии составила 93% при выявлении

«хороших» кредитов и только 69% при выявлении «плохих», точность дискриминантного анализа составила 85% и 85,5% соответственно, что позволило авторам констатировать большую эффективность модели дискриминантного анализа.

В статье [251] проведено бэк-тестирование классификации клиентов кредитных учреждений по степени дефолтности на основе дискриминантного анализа, которое показало точность 92,9%, проверка модели при помощи лямбда-статистики Уилкса показала высокую прогностическую ценность модели с существенной дискриминационной способностью.

Проведенный анализ позволяет сделать вывод, что модели дискриминантного анализа в большинстве эмпирических исследований используются для проверки уже применяемых на практике в банковских учреждениях моделей кредитного скоринга, т. е. строятся и интерпретируются по ретроспективным данным по действующим потребительским кредитам без апробации на потенциальных заемщиках.

В целом можно заключить, что модели кредитных рисков микрофинансовых организаций в рамках методов статического моделирования, как логистической регрессии, так и дискриминантного анализа, можно строить как на малых, так и на больших выборках, что в особенности важно в условиях ограниченности временных и финансовых ресурсов МФО.

Основным преимуществом методов статистического моделирования в оценке кредитных рисков МФО является достаточная прозрачность и интерпретируемость моделей, что безусловно важно с точки зрения регулятора и менеджмента компании. При этом методы статистического моделирования в большей степени направлены на долгосрочное прогнозирование кредитных рисков МФО, в отличие от моделей и алгоритмов машинного обучения, особенности которых будут рассмотрены в следующем параграфе.

3.3 Модели и алгоритмы машинного обучения оценки кредитных рисков и особенности их применения в микрофинансовых организациях

Как было отмечено выше, альтернативный подход к оценке кредитных рисков по сравнению со статистическим моделированием – это использование моделей, методов и алгоритмов машинного обучения. Машинное обучение стало неотъемлемой частью современной научной методологии [223]. В последние несколько лет технологии машинного обучения набирают популярность как в научной литературе, так и в эмпирических исследованиях для проведения оценки рисков кредитования [14, 19, 35, 74, 197]. В прикладных исследованиях, направленных на оценку риска дефолта заемщика как в банковской сфере, так и в микрофинансовых организациях, наиболее часто используемыми методами в рамках применяемых моделей и алгоритмов машинного обучения являются нейронные сети [278, 380, 381, 385], метод опорных векторов, ансамблевые методы: градиентный бустинг и случайный лес [167, 204, 211, 219, 234]. В связи с этим более подробно остановимся на их особенностях применения для оценки кредитного риска, а также приведем литературный обзор опыта применения технологий машинного обучения для оценки кредитных рисков в финансовой сфере.

Остановимся подробнее на основных математических аспектах построения нейронных сетей и других популярных методов машинного обучения для оценки кредитных рисков, подробно представленных в работах автора [110, 112, 114, 118].

Искусственная *нейронная сеть* (англ. Neural Network) – это математическая модель анализа данных, работающая по аналогии организации нервных клеток живых организмов. Корни разработки теории искусственных нейронных сетей уходят в такие дисциплины как нейрофизиологи, дискретная математика, статистика, методы оптимизации, физика и др.

Нейронная сеть состоит из элементов двух основных типов: нейронов и связей между нейронами. Нейроны искусственной нейронной сети осуществляют

вычисления, связи – передают результаты вычислений от одного нейрона к другому. При этом каждый нейрон выполняет простое вычисление, а в целом нейронная сеть может выполнять очень сложные вычисления за счет использования множества нейронов, объединенных в слои. В целом архитектура нейронной сети, т. е. описание входящих в нее нейронов и структура связей между ними, включает в себя следующие элементы (рисунок 3.1).

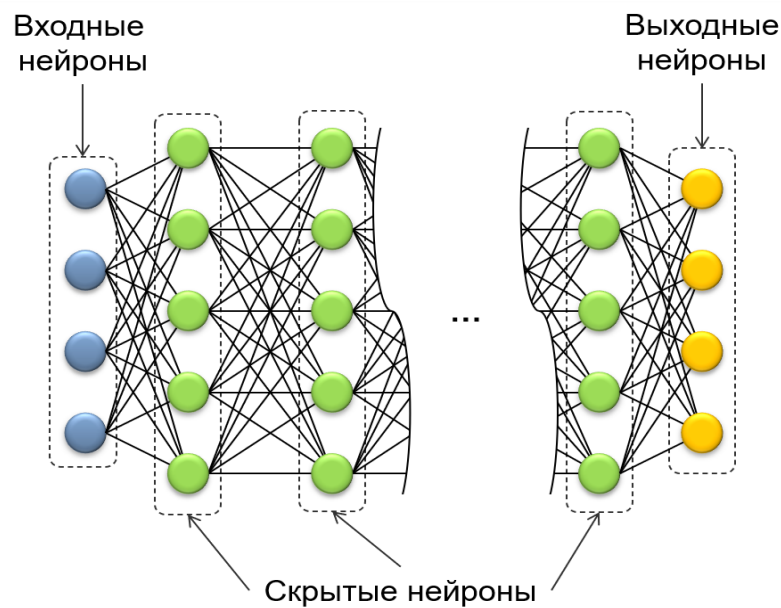


Рисунок 3.1 – Архитектура нейронных сетей

Источник: составлено автором по [341, 342]

Входные нейроны (входной слой) – служат для передачи исходной информации в нейронную сеть, не осуществляют вычислений. Скрытые нейроны (скрытый слой) – осуществляют вычисления, которые используются как промежуточные расчеты. Выходные нейроны (выходной слой) – осуществляют вычисления, переводя результаты промежуточных вычислений в финальный результат работы модели.

На практике, в том числе и области оценки кредитных рисков, наиболее распространенными являются следующие виды нейронных сетей (рисунок 3.2):

а) сети прямого распространения, в том числе многослойные сети прямого назначения – сети, в которых связи между нейронами не образуют циклов;

б) рекурсивные нейронные сети – в архитектуре сети связи образуют направленные циклы [242];

в) самоорганизующиеся сети Кохонена – нейронная сеть, обучающаяся в режиме обучения без учителя.

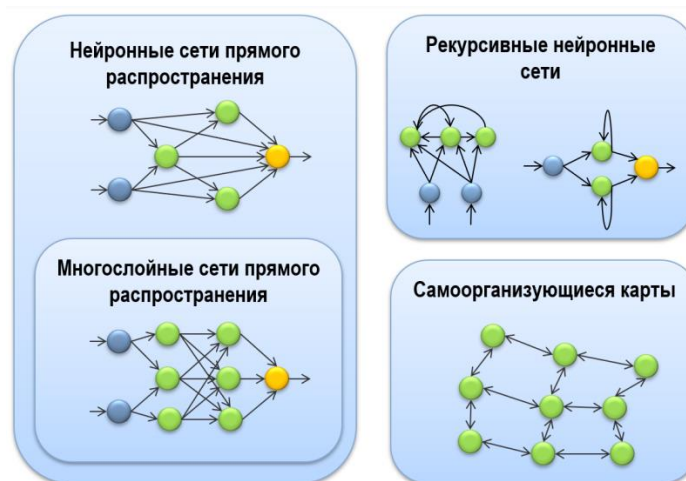


Рисунок 3.2 – Виды нейронных сетей

Источник: составлено автором по [233]

Наибольшее распространение в кредитном скоринге получили многослойные сети прямого назначения. Математическая модель нейрона, получившая название перцептрона, включает в себя следующие элементы (рисунок 3.3):

а) сумматор (s) – характеризует внешнее воздействие на нейрон входных данных или значений других нейронов, значение сумматора зависит от входных значений, весов и смещения;

1) входы (x_i) – это исходные данные или данные, вычисленные другими нейронами, число входов нейрона равно числу входящих связей и определяется архитектурой нейронной сети;

2) веса входных сигналов (w_i) – определяют вклад каждого входа в величину сумматора для расчета конечного значения нейрона;

3) смещение (w_0) – константа;

б) функция активации ($f(s)$) – определяет реакцию нейрона как отклик на значение сумматора;

в) выход (y) – значение, передаваемое другим нейронам или на выход сети.

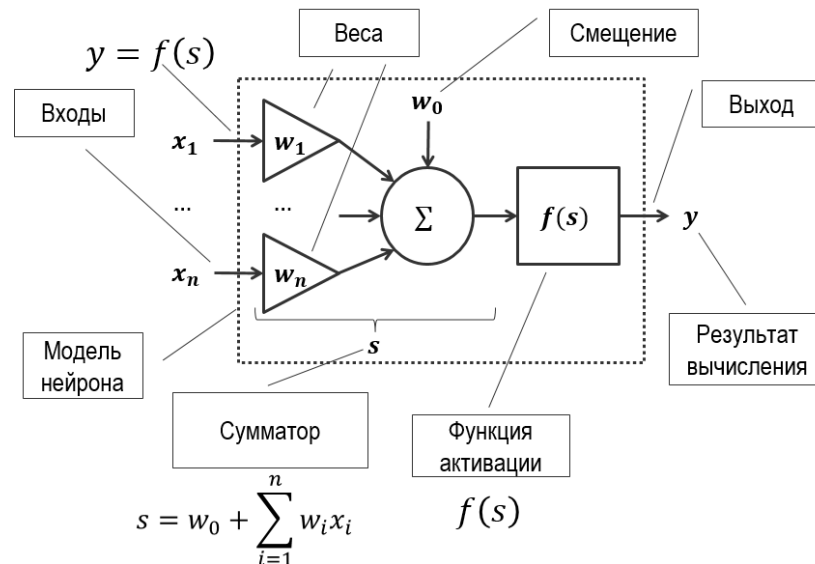


Рисунок 3.3 – Математическая модель нейрона

Источник: составлено автором по [179, 368]

Функция активации имеет немаловажное значение при задании параметров архитектуры нейронной сети. Выбор функции активации может зависеть от следующих параметров: слоя, в котором находится нейрон; решаемой задачи; используемого алгоритма обучения; предпочтений исследователя. В качестве функции активации перцептрона могут использоваться только специальные функции. Функция активации обязательно должна быть неубывающей. Использование неубывающей функции означает, что при увеличении значения сумматора, значение на выходе нейрона увеличиться или останется неизменным. Если для обучения нейронной сети используется алгоритм градиентного спуска, то все функции активации должны быть гладкими.

В скрытых слоях используют функции активации, которые удовлетворяют следующим свойствам:

- функция должна быть нелинейной;
- функция должна иметь ограниченную область выходных значений.

Использование функций с ограниченным диапазоном выходных значений обеспечивает вычислительную устойчивость нейронных сетей. В выходных слоях нейронных сетей могут использоваться и другие функции активации, не удовлетворяющие требованиям для функций в скрытых слоях. В задачах регрессии функция активации выходного слоя должна быть непрерывной, чтобы обеспечить возможность вычисления всех значений непрерывной зависимой переменной. В задачах классификации, характерных для оценки кредитных рисков, можно использовать как непрерывные, так и дискретные функции активации выходного слоя. При использовании дискретной функции активации выходное значение рассматривается как номер класса, а при использовании непрерывной функции активации выходные значения рассматриваются как вероятность принадлежности объекта различным классам.

На практике рекомендуется использовать следующие наиболее распространенные функции активации для скрытых слоев [262]:

а) сигмоида или сигмоидальная функция:

$$f_{sigmoid}(s) = \frac{1}{1+e^{-\alpha s}}; \quad (3.36)$$

б) гиперболический тангенс:

$$f_{th}(s) = th\left(\frac{\alpha s}{2}\right) = \frac{e^{\alpha s} - 1}{e^{\alpha s} + 1} = 2f_{sigmoid}(s) - 1, \quad (3.37)$$

где α – параметр.

Гиперболический тангенс и сигмоида имеют схожую форму и многие свойства. Функция гиперболического тангенса, как и сигмоида, монотонна и дифференцируема во всем диапазоне значений аргумента. Выходные значения функции гиперболического тангенса могут быть от -1 до 1 . Различие сигмоиды и гиперболического тангенса в том, что гиперболический тангенс – биполярная функция (выходные значения от -1 до 1), а сигмоида – униполярная (выходные значения от 0 до 1).

Для решения задач классификации могут использоваться нейросети с пороговыми функциями активации софтмакс (рисунок 3.4).

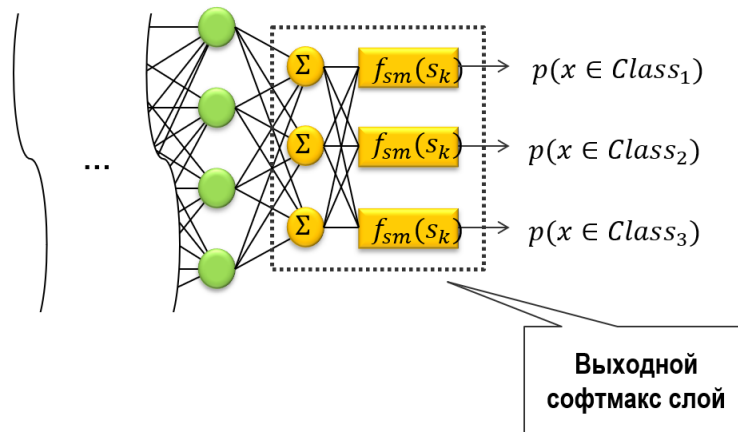


Рисунок 3.4 – Нейронная сеть с функцией активацией софтмакс

Источник: составлено автором по [135]

Архитектура нейронной сети для решения задач классификации следующая:

- число нейронов входного слоя равно числу независимых переменных в задаче;
- число нейронов выходного слоя равно числу классов зависимой переменной;
- сеть включает один или два скрытых слоя;
- значение на каждом выходе рассматривается как вероятность принадлежности объекта соответствующему классу.

К функциям активации для решения задач классификации предъявляют следующие требования: в нейронах скрытых слоев используются нелинейные функции (сигмоиду и гиперболический тангенс), а на выходном слое необходимо использовать гладкую непрерывную функцию с областью значений от 0 до 1. Для решения задач классификации в архитектуре сети используют выходной слой с функцией активации софтмакс [135]:

$$x \in Class_k = f_{sm}(s_k) = \frac{e^{s_k}}{\sum_{j=1}^m e^{s_j}}, \quad (3.38)$$

где k – индекс нейрона;

m – число нейронов софтмакс слоя (число классов);

s_k – сумматор каждого нейрона;

$f_{sm}(s_k)$ – функция активации софтмакс.

В отношении эмпирических исследований по использованию моделей нейронных сетей в оценке кредитных рисков было исследовано несколько трудов зарубежных авторов. Так, в работе [365] приведены результаты разработки модели кредитного рейтинга заемщиков потребительских кредитов в банках Словении, построенной на основе нейронной сети Кохонена. В качестве входных переменных автор использовал массив данных по транзакциям и остаткам на счетах клиентов. Точность прогноза определения проблемных потребительских кредитов в соответствии с данной моделью составила 79,3%.

Авторы статьи [217] на основе выборки 5500 заемщиков микрофинансовой организации Перу сформировали модель оценки кредитных рисков на основе нейронной сети, которая сравнивается по эффективности с классическими методами статистического моделирования – линейным дискриминантным анализом (LDA), квадратичным дискриминантным анализом (QDA) и логистической регрессией. По результатам проведенного исследования авторы пришли к выводу, что модель нейронных сетей превосходит классические методы как с точки зрения точности прогноза, так и затрат на неправильную классификацию.

Обратный результат был получен в исследовании [247], автор которого сравнил эффективность для оценки кредитных рисков крупного розничного банка Бразилии модели логистической регрессии и нейронной сети, точность прогноза дефолтности заемщиков составила 68,3% и 67,7%, соответственно. При этом в рамках нейронной сети была зафиксирована большая точность по «плохим» заемщикам – 75,2%, а в модели логистической регрессии – по «хорошим» заемщикам – 72%. Таким образом, по полученным результатам нельзя сделать достоверное заключение относительно эффективности той или иной модели.

В статье [42] проведен сравнительный анализ возможности использования для оценки кредитных рисков модели рекуррентных нейронных сетей и модели градиентного бустинга, построенных на данных по транзакциям владельцев кредитных карт. Точность обеих моделей варьировалась в пределах 59,7–66,7% с перевесом нейронных сетей.

Следующей востребованной моделью машинного обучения в рамках оценки кредитных рисков является *метод опорных векторов* (англ. Support Vector Machine, SVM) – метод классификации, основанный на статистической теории обучения, предложенный в 1995 году исследовательской группой лабораторий AT&T Bell под руководством В. Вапника [376], эксперты которой обосновали теорию двойственности в традиционной задаче оптимизации.

Базовая модель SVM представляет собой линейный классификатор с максимальным зазором, определенным в пространстве признаков. Метод опорных векторов базируется на идее построения оптимальной гиперплоскости, разделяющей данные на классы при максимальном увеличении зазора.

Алгоритм SVM определяется следующими входными параметрами обучающих данных:

$$T = \{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\}, \quad (3.39)$$

где (x_i, y_i) – точка обучающей выборки с заданным классом;

x_i – вектор признаков (элемент), который нормализуется на отрезке $[0; 1]$;

$x_i \in \mathcal{X} = R^n$;

$i = \{1, 2, \dots, N\}$;

y_i – класс вектора x_i , $y_i \in \mathcal{Y} = \{+1; -1\}$,

$\begin{cases} y_i = +1, \text{ то } x_i \text{ – положительный вектор признаков,} \\ y_i = -1, \text{ то } x_i \text{ – отрицательный вектор признаков.} \end{cases}$

Разделяющая гиперплоскость соответствует уравнению:

$$\begin{cases} w^T \cdot x_i - b = 0 \\ w^T = (w_1, w_2, \dots, w_n), \\ b = -w_0 \end{cases} \quad (3.40)$$

где w^T – вектор нормали к разделяющей гиперплоскости;

b – отсекаемый отрезок.

Разделяющая гиперплоскость делит пространство признаков на две части – положительный класс и отрицательный класс. Тот, на который указывает вектор нормали, относится к положительному классу, противоположный – к

отрицательному. В контексте оценки кредитных рисков – это классы «плохих» и «хороших» заемщиков.

Как правило, когда набор обучающих данных является линейно разделимым, существует бесконечное количество разделяющих гиперплоскостей, которые могут правильно разделить два класса данных. Линейный SVM использует принцип максимизации зазора, чтобы найти оптимальную разделяющую гиперплоскость для определения единственного решения.

Уравнения параллельных гиперплоскостей в рамках оптимального разделения принимают вид:

$$\begin{cases} w^T \cdot x - b = +1, \\ w^T \cdot x - b = -1. \end{cases} \quad (3.41)$$

В условиях линейной разделимости обучающей выборки гиперплоскости выбираются с учетом отсутствия между ними точек обучающей выборки – SVM с жестким зазором. Исключение всех точек обучающей выборки из межгиперплоскостного пространства определяется уравнением следующего вида:

$$\begin{cases} w^T \cdot x_i - b \geq 1, & y_i = 1, \\ w^T \cdot x_i - b \leq -1, & y_i = -1. \end{cases} \quad (3.42)$$

Расстояние между гиперплоскостями определяется как $\frac{2}{\|w\|} \rightarrow \max$, где $\|w\|$ – сумма абсолютных значений коэффициентов гиперплоскости, при этом $\|w\| \rightarrow \min$, т. е. должно быть минимизировано в рамках задачи квадратичной оптимизации по следующему уравнению:

$$\begin{cases} \frac{\|w\|^2}{2} \rightarrow \min, \\ y_i(w^T \cdot x_i - b) \geq 1, \quad 1 \leq i \leq n. \end{cases} \quad (3.43)$$

Оптимально разделяющий классификатор будет получен при решении данного уравнения с применением метода множителей Лагранжа.

В условиях нелинейной разделимости [184] обучающей выборки вводится набор дополнительных переменных $\varepsilon_i \geq 0$, которые определяют величину ошибки элементов x_i (слабые переменные, представляющие нарушения зазора). Далее для регулирования максимального зазора между гиперплоскостями и

минимизацией суммарной ошибки вводится штраф, определяемый уравнением вида:

$$\begin{cases} \frac{\|w\|^2}{2} + \alpha \sum \varepsilon_i \rightarrow \min \\ y_i(w^T \cdot x_i - b) \geq (1 - \varepsilon_i), \\ \varepsilon_i \geq 0 \end{cases} \quad (3.44)$$

где $(1 - \varepsilon_i)$ – минимальный требуемый зазор для каждой точки данных;

α – параметр регуляризации, определяющий баланс между шириной поля и допуском к ошибкам классификации.

Целевая функция SVM с мягким зазором имеет следующий вид:

$$\frac{\|w\|^2}{2} + \alpha \sum N_{i=1} \varepsilon_i, \quad (3.45)$$

где N – количество точек данных.

Преимущества SVM с мягким зазором определяются следующим:

- во-первых, устойчивость к выбросам – может более эффективно обрабатывать выбросы или зашумленные данные, допуская некоторые неправильные классификации. Это приводит к более надежной границе принятия решений с оптимальным обобщением невидимых данных;

- во-вторых, применимость к нелинейным данным, в отличие от SVM с жестким зазором, SVM с мягким зазором может обрабатывать нелинейно разделяемые данные, неявно помещая их в многомерное пространство с помощью функций ядра. Это позволяет SVM фиксировать сложные границы принятия решений.

В рамках метода SVM используют различные виды ядер (kernel) – линейное, полиномиальное, радиальное (гауссовское) и сигмоидальное. Преимущества использования линейного ядра в SVM заключаются в упрощении вычислений и соответственно в сокращении времени обучения и предсказания модели, что особенно предпочтительно при большом количестве признаков. Другие виды ядер используются в случае отсутствия линейной делимости данных в исходном пространстве.

Основной недостаток метода опорных векторов (SVM) заключается в возможности использования для классификации только ограниченного набора данных, находящихся в границах гиперплоскости, а также существенной чувствительности метода к стандартизированным данным.

Обзор научной литературы по применению метода опорных векторов для оценки кредитных рисков индивидуальных заемщиков позволил выделить несколько основных работ. В статье [335] проведено исследование эффективности данного метода для анализа кредитного риска розничного кредитования в банках Индонезии на обучающей и тестовой выборке данных по 610 кредитным историям заемщиков, включающих девять независимых переменных (социально-демографические, финансовые характеристики, платежная дисциплина, данные по кредитам). Тестирование алгоритма SVM в рамках данного исследования проводилось с использованием линейного, полиномиального, радиального и сигмоидального ядра, при этом модель SVM с полиномиальным ядром получила самую высокую точность – 95,08% и значение AUC (способность модели различать классы) – 0,9419. Это позволило автору прийти к выводу, что данную модель можно использовать для классификации потенциальных клиентов на классы с хорошей или плохой кредитной историей с достаточно высокой точностью для снижения кредитного риска.

В исследовании [159] представлена модель кредитного скоринга на основе метода опорных векторов с радиальным ядром для классификации заемщиков банковских учреждений Марокко и определения вероятности их дефолта. Авторы провели сравнительный анализ прогностической точности данной модели с моделью логистической регрессии, который показал большую эффективность метода опорных векторов – 98% против 71,7%. При этом авторы заключили, что разработанная модель достаточно сложна с точки зрения выбора значений параметров ядер SVM.

Необходимо отметить, что в последнее время в оценке кредитных рисков набирают популярность ансамблевые методы машинного обучения, которые представляют собой алгоритмы, в рамках которых обучают несколько отдельных

моделей для решения конкретной проблемы, а затем получают окончательный результат, объединяя их в единую прогнозную модель [186, 346]. Возможность применения ансамблевых методов в различных контекстах, включая классификацию, регрессию и обучение с подкреплением, позволяет достигать им большую эффективность в сравнении с базовыми моделями и алгоритмами. Наиболее распространенными ансамблевыми методами являются случайный лес и градиентный бустинг. В основе работы этих алгоритмов машинного обучения лежат классические методы построения деревьев решений, представленных в работах [177, 201, 248, 285, 366].

Случайный лес (англ. Random Forest) – это алгоритм машинного обучения, в основе которого построение множества независимых друг от друга «деревьев решений». Это эффективный контролируемый алгоритм машинного обучения, который можно использовать как для задач классификации, так и для регрессии. Во время обучения создается множество деревьев решения посредством использования подхода ансамблевого обучения, а окончательный результат определяется прогнозом среднего значения (регрессия) или модального класса (классификация) на основе отдельных деревьев [190]. По сравнению с одиночными деревьями решений случайные леса, как правило, имеют более высокую точность, поскольку они уменьшают переобучение и более эффективно обрабатывают зашумленные данные. Во время обучения случайные леса используют случайный набор функций, что помогает увеличить разнообразие деревьев и улучшить общую производительность работы алгоритма.

Основная идея метода случайного леса – бэггинг (бутстрэп-агрегирование). В алгоритме случайного леса определенное пользователем количество деревьев решений обучается на примерах бутстрэп (bootstrap) предоставленного набора обучающих данных [220, 354].

В задачах классификации алгоритм случайного леса обучается на наборе данных:

$$T = \{(x_i, y_i)\}_{i=1}^N, \quad (3.46)$$

где $x_i \in X = R^p$ – входные данные;

$y_i \in Y = R^k$ – соответствующие выходные данные.

Обозначим:

T – дерево решений;

B – количество деревьев решений, $0 < B \leq p$, где p – размерность вектора признака x_i (общее число входных элементов), K – количество случайно выбранных признаков,

для $b = 1, 2, \dots, B$:

$Z_b \subseteq T$ – набор данных, образец бутстрэпа из T размера N (с заменой);

$\hat{h}_b \leftarrow T(Z_b, K, \infty)$ – оценка значений зависимой переменной по построенному дереву решений (∞) для каждой из предварительно загруженных выборок со случайно выбранными признаками $K \leq p$ в каждом разделе.

Функция регрессии имеют следующий вид:

$$f^*(x) = \frac{1}{B} \sum_{j=1}^B \hat{h}_j(x). \quad (3.47)$$

Функция классификации имеют следующий вид:

$$f^*(x) = \text{majority vote} [\hat{h}_j(x)]_1^B, \quad (3.48)$$

где *majority vote* – преобладающий элемент последовательности данных, т. е. модальное значение предсказываемой зависимой переменной.

Алгоритм генерирует случайным образом B наборов данных Z_b , поскольку выборки бутстрэп, извлеченные из исходной обучающей выборки T , имеют одинаковый размер N . Каждый из этих наборов данных используется для обучения совокупности деревьев решений.

Одной из характеристик, которая отличает алгоритм случайного леса от бэггинга и способствует декорреляции деревьев, является выборочный учет только K из общего числа p входных элементов для выбора лучших точек разделения при построении каждого дерева решений. Значение K задается пользователем или устанавливается по умолчанию, где $0 < K \leq p$. Таким образом, каждое дерево значительно отличается от остальных, что делает деревья менее коррелированными.

Алгоритм повторяет эту процедуру B раз, в результате чего получается B -число деревьев решений и обучение случайного леса завершается.

В рамках метода случайного леса в качестве предикторов могут выступать входные переменные в разных шкалах: числовой, порядковой и номинальной.

Одним из преимуществ метода случайного леса является то, что пользователь может полагаться на знания, специфичные для предметной области, при проектировании входных функций. Данный метод имеет лучшую содержательную интерпретацию результатов на основе относительной важности, которую алгоритм присваивает каждому из них. Кроме того, в рамках данного метода отсутствуют ограничения на степень мультиколлинеарности зависимых переменных.

Применение метода случайного леса становится востребованным в эмпирических исследованиях, направленных на оценку кредитных рисков.

Например, целью исследования [228] является определение эффективности метода случайного леса в оценке кредитного риска на необработанном наборе данных кредитных бюро Германии с использованием 32 581 наблюдений. Исследование доказало, что метод случайного леса обеспечивает более высокую точность выводов относительно дефолтности заемщиков и классификационную способность (91%) по сравнению со стандартным методом дерева решений (83%).

В статье [143] разработана модель оценки индивидуальных кредитных рисков на основе метода случайного леса выживаемости, который позволяет оценить не только вероятность дефолта заемщиков, но и вероятность временного интервала до момента его наступления, точность модели на тестовой выборке составила 75,7%. При этом авторы пришли к выводу о достаточной интерпретируемости модели в соответствии со стандартом МСФО 9 в отношении оценки кредитных рисков на всем жизненном цикле кредита.

Исследование [311] включает разработку модели оценки вероятности дефолта заемщиков по автокредитам в рамках метода случайного леса на массиве несбалансированных данных, прогностическая способность модели составила 85%.

Работа [348] посвящена разработке модели предсказания кредитных рисков на основе метода параллельного случайного леса с уменьшенным набором параметров и исключением нерелевантных признаков. Апробация модели проведена на двух базах общедоступных данных – German Credit Dataset и Australian Credit Dataset, точность модели для которых составила 76,2% и 89,4%, соответственно при одновременном сокращении времени на принятие решений по сравнению со стандартными методами кредитного скоринга.

Следующий вид ансамблевых методов машинного обучения – *градиентный бустинг* (англ. Gradient boosting), суть которого заключается в создании модели прогнозирования в виде ансамбля более слабых моделей, как правило, деревьев решений, с целью оптимизации произвольной дифференцируемой функции потерь [183]. В работе [241] представлена стандартная версия бустинга для каждой функции потерь. Во время обучения градиентный бустинг работает путем подгонки на основе остатков моделей, представляющих собой разницу между прогнозируемым и реальным значением. Затем алгоритм корректирует прогнозируемое значение, добавляя часть прогнозируемого значения из новой менее точной модели, при этом коэффициент корректировки контролируется скоростью обучения или размером шага. Таким образом, суть метода градиентного бустинга заключается в обучении каждой последующей модели с использованием данных об ошибках предыдущих моделей и дальнейшим снижением ошибок. Процесс повторяется до тех пор, пока не будет выполнен критерий остановки алгоритма.

Алгоритм градиентного бустинга определяется следующими входными параметрами обучающих данных:

$$T = \{(x_i, y_i)\}_{i=1}^n. \quad (3.49)$$

Цель алгоритма – найти функцию $F(x)$, принимающую следующий вид:

$$F(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma), \quad (3.50)$$

где L – функция потерь, которая определяется через уравнение:

$$L = (y_i - \gamma)^2. \quad (3.51)$$

Оператор γ , включенный в данную функцию, предназначен для минимизации градиента.

Особенности градиентного бустинга заключаются в следующем:

- во-первых, алгоритм обновляет веса, вычисляя отрицательный градиент функции потерь относительно прогнозируемого результата;
- во-вторых, обновляет веса на основе градиентов, которые менее чувствительны к выбросам;
- в-третьих, может использовать широкий спектр базовых обучаемых моделей, например деревья решений и линейные модели.

Главный недостаток методов градиентного бустинга определяется существенными вычислительными затратами и лимитом минимальной обучающей выборки.

Модели градиентного бустинга находят свое применение для оценки кредитных рисков как в банках, так и микрофинансовых организациях. В настоящее время это самый популярный применяемый алгоритм машинного обучения в кредитном скоринге.

Например, статья [379] включает выработку двух моделей градиентного бустинга (XGBoost, Adaboost) для предсказания дефолта заемщиков китайских банков с апробацией на выборке из 32 500 кредитных историй. Автор провел сравнительный анализ разработанных моделей с оценкой на основе модели логистической регрессии, предикативная способность моделей XGBoost, Adaboost составила примерно одинаковую точность – около 92%, против 81,2% в логистической регрессии. При этом автор заключил, что при выборе модели оценки кредитных рисков банки должны находить баланс между эффективностью модели и получаемыми выгодами, т. к. разработка более сложных моделей связана с большими финансовыми затратами.

В работе [216] проведена разработка и соответствующее сопоставления различных алгоритмов градиентного бустинга (включая XGBoost, Adaboost) с целью оценки точности моделей в рамках прогнозирования кредитных рисков микрофинансовых организаций. Модели апробированы на реальных данных по

кредитованию заемщиков микрофинансовой организацией из Боснии и Герцеговины. Наибольшую точность (81,8%) продемонстрировал алгоритм XGBoost. Автор сделал вывод о том, что для большей точности полученных результатов необходимо проводить сравнительный анализ с другими методами машинного обучения (например, деревом решений, методом k -ближайших соседей [265] и других).

Модели градиентного бустинга (XGBoost, Adaboost) для оценки рисков возможного дефолта новых клиентов микрофинансовых организаций были представлены в статье [147]. Авторы провели сравнительный анализ эффективности и точности выработанных моделей на данных заемщиков, которые ранее не имели кредитную историю. Все модели, разработанные и принятые авторами для сравнения, получили точность в пределах 80%.

По проведенному исследованию можно заключить, что построение моделей в рамках методов машинного обучения осуществляют на больших обучающих выборках, что предполагает проведение существенного количества (более сотни тысяч) наблюдений. В отличие от крупных финансово-кредитных учреждений (например, банков) в распоряжении микрофинансовых организаций, как правило, может не быть большого количества данных для построения моделей. Однако преимущества методов машинного обучения определяются возможностью использования в качестве предикторов переменных в любых шкалах.

Несмотря на популяризацию применения методов машинного обучения для оценки кредитных рисков, есть множество противников их применения для решений задачи кредитного скоринга. И тут речь идет не об отрицании возможности применения методов машинного обучения для оценки кредитных рисков, а о понимании специфики набора исходных данных, объема выборки, особенностей работы алгоритмов машинного обучения, и, как следствие, корректности их применения для решения конкретной задачи. Рассмотрим более подробно особенности применения моделей, методов и алгоритмов машинного обучения для оценки кредитных рисков МФО.

Одним из «подводных камней» применения методов машинного обучения на практике является проблема переобучения, которая заключается в различии результатов применения сформированной модели на обучающей и тестовой выборке – увеличение сложности модели приводит к уменьшению ошибки на обучающей выборке, а в отношении тестовой выборки при усложнении модели сначала происходит уменьшение ошибки, а при достижении определенного момента ее увеличение. Среди основных причин переобучения можно выделить наличие мультиколлинеарности, наличие неинформативных признаков, ограничение размера выборки. В настоящее время используют различные стратегии решения проблемы – переобучения:

- стратегия сокращения количества переменных для оценки кредитного риска за счет отбрасывания наименее важных данных;
- стратегия расширения данных за счет более точной настройки параметров модели;
- стратегия прореживания за счет взятия для конечной оценки усредненных результатов набора моделей;
- стратегия ранней остановки базируется на определении контрольной точки, после которой начинается увеличение ошибки;
- стратегия регуляризации определяется выбором функций, которые оказывают требуемое влияние на конечный результат.

Цель стратегий решения проблемы переобучения заключается в нахождении баланса между сложностью модели и возможностью ее применения на новых данных или генеральной совокупности данных.

В отношении реализации сложных моделей машинного обучения в управлении кредитными рисками МФО также следует отметить отсутствие часто технических возможностей, недостаточно развитую информационную инфраструктуру МФО и более низкие финансовые возможности, что отличает их от крупных финансово-кредитных учреждений (банков).

Одной из существенных проблем применения методов моделирования на практике является модельный риск, который заключается в возможности

получения убытков по причине недостаточно точного описания процессов в рамках разработанной модели. Основными причинами возникновения модельного риска являются следующие:

- во-первых, некорректность определения переменных;
- во-вторых, некорректное применение модели, обученной на других данных;
- в-третьих, изменение данных и ухудшение модели при воздействии внешних факторов.

Для снижения модельного риска необходимо соблюдать методику анализа информации, используемую для оценки кредитных рисков, представленную в главе 2.

Модельные риски применяемой концепции моделирования (статистический подход или машинное обучение) определяются вероятностью ошибки прогнозирования и суммой возможного ущерба.

Вероятность и цена ошибки находится в прямой зависимости от горизонта прогнозирования при использовании модели (рисунок 3.5).

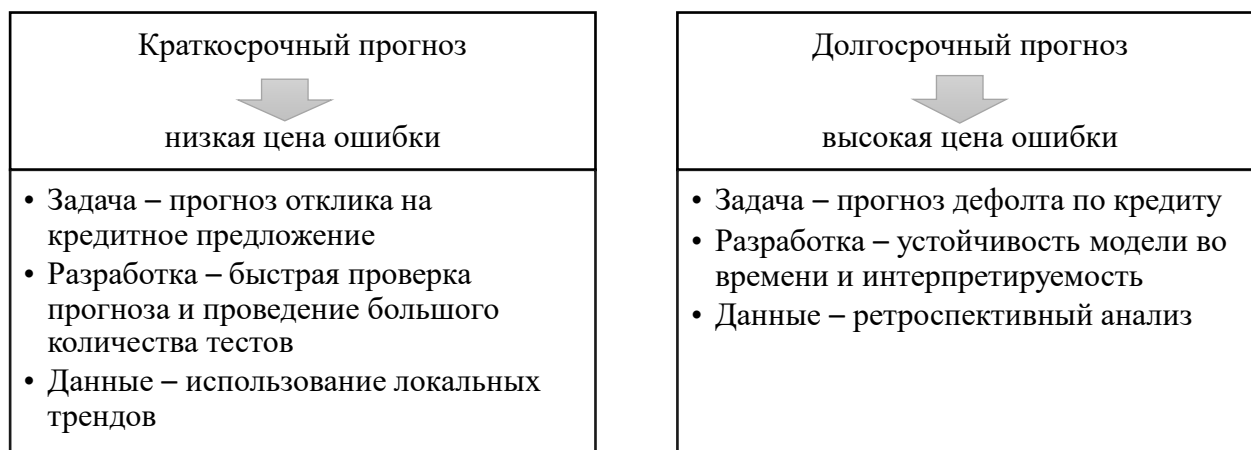


Рисунок 3.5 – Зависимость горизонта прогнозирования и ценой ошибки прогноза

Источник: Составлено автором

Модельный риск методов в рамках статистического моделирования в отношении оценки кредитных рисков микрофинансовых организаций определяется следующим:

- во-первых, статистические модели используют параметрические функции с небольшим количеством свободных параметров;
- во-вторых, на основе статистического анализа данных вычисляются как сами параметры, так и их доверительные интервалы;
- в-третьих, статистические модели не склонны к переобучению и более устойчивы при изменении внешних условий.

В отношении модельного риска методов в рамках машинного обучения при проведении оценки кредитных рисков микрофинансовых организаций можно заключить следующее:

- во-первых, используются параметрические функции с огромным количеством свободных параметров;
- во-вторых, статистическая оценка средних значений и доверительных интервалов этих параметров невозможна;
- в-третьих, модели машинного обучения склонны к переобучению и часто на практике оказываются нестабильными.

В таблице 3.2 представлен сравнительный анализ основных параметров методов статистического моделирования и машинного обучения с точки зрения возможности использования для оценки кредитных рисков в микрофинансовых организациях по основным критериям.

Методы статистического моделирования в большей степени реализуемы для оценки кредитных рисков микрофинансовых организаций в условиях имеющейся информационной инфраструктуры. Основным преимуществом методов статистического моделирования является достаточная прозрачность и интерпретируемость моделей, что безусловно важно с точки зрения регулятора. При этом методы статистического моделирования в большей степени направлены на долгосрочное прогнозирование [32] в отличие от методов машинного обучения. Соответственно методы статистического моделирования имеют более

высокую цену ошибки в сравнении с методами машинного обучения, в условиях неправильного определения входных данных, формирования модели и ее тестирования, и широко применяются в разных сферах [37, 38].

Таблица 3.2 – Сопоставление методов статистического моделирования и машинного обучения

Подход	Статистическое моделирование	Машинное обучение
Модель	Интерпретируемая Прозрачная	Не интерпретируемая Непрозрачная
Параметры	Средние и дисперсии оцениваются из данных	Средние и дисперсии не определены
Проверка модели	Тесты «goodness-of-fit» и анализ остатков	Измеряется только предсказательная сила модели
Задачи	Долгосрочное прогнозирование Ретроспективный анализ Прогноз «through the cycle»	Краткосрочное прогнозирование Локальные тренды
Цена ошибки	Высокая цена ошибки	Низкая цена ошибки

Источник: Составлено автором в рамках проведенного исследования

Проведенный сравнительный анализ использования традиционных статистических моделей и алгоритмов машинного обучения в контексте оценки кредитных рисков в микрофинансовых организациях показал, что алгоритмы машинного обучения, в первую очередь, ансамблевые методы, например метод случайного леса и градиентный бустинг, обеспечивают лучшую производительность и эффективность классификации заемщиков, чем стандартные модели дискриминантного анализа и логистической регрессии.

Тем не менее логистическая регрессия остается эталонной математической моделью оценки кредитных рисков микрофинансовых организаций, в первую очередь, по причине низкой интерпретируемости ансамблевых методов машинного обучения и их несовместимости с требованиями регулятора. Это приводит к необходимости разработки скоринговых моделей для управления кредитными рисками в микрофинансовых организациях, которые будут

отличаться с одной стороны высокой производительностью, эффективностью классификации заемщиков и точностью прогнозирования, с другой стороны – достаточной степенью интерпретируемости в соответствии с регуляторными требованиями.

Исходя из этого, в научных кругах появляются исследования, направленные на совместное использование методов традиционного статистического моделирования с алгоритмами машинного обучения для повышения эффективности моделей кредитного скоринга.

Например, как показало исследование [350] для повышения интерпретируемости и точности методов машинного обучения, используемых в банковской практике классификации заемщиков по кредитному риску, включая нейронные сети, методы случайного леса и градиентного бустинга, должны быть дополнены методом линейного дискриминантного анализа в качестве инструмента снижения размерности классов.

В статье [305] разработан высокопроизводительный и интерпретируемый метод оценки кредитных рисков – модель выбракованной логистической древовидной регрессии (англ. Penalised Logistic Tree Regression, PLTR), которая использует информацию из деревьев решений для повышения производительности логистической регрессии.

Разработка метода скрининга бинарных информационных значений (англ. Binary Informative Value, BIV), основанного на модели средневзвешенного значения на базе логистической регрессии, а также двух методов машинного обучения: случайного леса и градиентного бустинга проведена в статье [339]. Оценка проводилась на основе информации из нескольких источников (кредитной истории, транзакций по кредитной карте, анкетных данных из заявки). В результате авторам удалось улучшить качество классификации заемщиков и упростить методику принятия кредитных решений в банковских учреждениях.

Вышеуказанные исследования направлены на выработку моделей скоринга в управлении кредитными рисками в банковских учреждениях и в силу недостаточности ресурсов – финансовых, информационных, кадровых – в полной

мере не могут быть адаптированы к деятельности микрофинансовых организаций. Это ведет к необходимости формирования методики разработки скоринговых моделей с адаптацией к реалиям деятельности и соответствующего применения в управлении кредитными рисками МФО.

Выводы по главе 3

В данной главе предложена системная методика построения моделей оценки кредитного риска МФО, включающая описание соответствующих методических подходов к построению выборки данных, выбору предикторов для модели, определения зависимой переменной и методы категоризации независимых переменных, а также диагностики мультиколлинеарности с учетом особенностей микрофинансовой деятельности.

Проведен сравнительный анализ концепций статистического моделирования и машинного обучения и обоснования условий их применения для оценки кредитных рисков МФО. Определено основное преимущество методов статистического моделирования – их достаточная прозрачность и интерпретируемость моделей, а также направленность на долгосрочное прогнозирование кредитных рисков МФО.

Основное преимущество методов машинного обучения заключается в возможности использования в качестве предикторов моделей переменных в любых шкалах, а также в том, что они обеспечивают лучшую производительность и эффективность классификации заемщиков. При этом эталонной математической моделью в разработке моделей оценки кредитных рисков МФО остается логистическая регрессия, модификация которой под особенности экономики МФО позволит достичь более высокой производительности модели в рамках качественной классификации заемщиков при высоком уровне ее интерпретируемости в соответствии с регуляторными требованиями.

Глава 4 Методы управления кредитными рисками в микрофинансовых организациях

4.1 Управление кредитными рисками микрофинансовых организаций на основе регулирования лимитов

В предыдущей главе были рассмотрены математические модели, используемые при оценке кредитных рисков в микрофинансовых организациях. В данной главе будут рассмотрены уже разработанные автором модели для управления кредитными рисками микрофинансовых организаций на основе математического инструментария, представленного в предыдущей главе и авторских методик.

Модели оценки кредитных рисков, рассмотренные в предыдущей главе, на выходе дают МФО скоринговый балл или вероятность дефолта по кредиту. Дальнейшее управление кредитными рисками в классическом виде заключается в выборе оптимального значения вероятности (балла) для отсека клиентов (англ. cut off) в СПР кредитного учреждения. Клиенты, которые в результате скоринга заявки получают оценку кредитного риска выше порогового значения, – получают отказ, а ниже – одобрение (при прогнозировании вероятности дефолта). При этом базовые параметры кредита такие как, процентная ставка, срок кредита, сумма кредита, могут быть фиксированными, а могут зависеть от уровня кредитного риска заемщика.

В микрофинансах продукт PDL предполагает фиксированную процентную ставку в день, срок также фиксируется до 30 дней, а вот сумма займа может напрямую зависеть от уровня кредитного риска заемщика. В этом случае общепринятой практикой является по результатам модели оценки кредитных рисков разбить заемщиков на группы риска (на децили или квартили риска, или произвольное число групп). Если децили риска ранжируют заемщика по возрастанию степени кредитного риска, то величина предлагаемого лимита займа

в случае его успешного одобрения СПР в таком случае будет обратно пропорциональна уровню риска: чем выше дециль риска – тем ниже лимит вплоть до 0 (отказа). Обычно минимальный лимит PDL на российском рынке микрофинансирования составляет от 3000 рублей (чаще от 5000 руб.), а максимальный 30 000 рублей. Первичным клиентам МФО выдают до 15 000 рублей, а повторным клиентам лимит уже может достигать максимального значения.

Пример назначения лимитов по децилям риска полученной модели оценки кредитного риска представлен в таблице 4.1.

Таблица 4.1 – Пример установки лимитов первичным клиентам по децилям риска

Дециль	Границы	NPL30, %	ROI, %	Лимит, руб.
1	<= 24588	15	16	15 000
2	24588 – 30827	18	12	15 000
3	30827 – 35308	25	8	13 000
4	35308 – 39081	30	7	12 000
5	39081 – 42434	32	6	12 000
6	42434 – 45897	36	5	9 000
7	45897 – 49466	40	5	6 000
8	49466 – 54221	42	1	4000
9	54221 – 61706	45	-3	0
10	61706+	51	-10	0

Источник: составлено автором, данные условные

В таблице 4.1 максимальный лимит ограничен 15 тыс. руб., т. к. модель построена для первичных клиентов. В 9 и 10 дециле риска с отрицательной доходностью назначается лимит 0 рублей (отказ). В остальных же децилях лимит пропорционален кредитному риску. Анализ таблицы 4.1 можно дополнить показателями в каждом дециле: количество выданных займов в штуках и в денежном выражении, конверсией из одобрения в выдачу (TR). Доходность приведена на горизонте 6 месяцев как разница между выданными и полученными средствами без учета прочих затрат.

Использование децилей риска для назначения лимитов позволяет ранжировать заемщиков по степени кредитного риска, но не отвечает на вопрос, какой лимит является оптимальным для каждой группы заемщиков. Децили риска представляют собой порядковую шкалу оценки кредитного риска, устанавливающую между заемщиками иерархию риска, но не определяющую, насколько или во сколько раз риск одной группы отличается от другой.

Как было определено в параграфе 3 главы 1 в статьях, как отечественных, так и зарубежных авторов предпринимались попытки обоснования назначения лимитов в зависимости от сегмента клиентов по степени риска, т. е. применение сегментированного подхода в назначении лимитов в банковской сфере. Однако в микрофинансовом секторе и, в частности, на рынке PDL-займов вопрос назначения лимитов на основе математического моделирования не проработан.

В текущем параграфе диссертационного исследования автором предлагается иной подход к снижению рисков, который не был широко изучен ни в отечественной, ни в зарубежной научной среде. Новизной излагаемого подхода является комплексное применение результатов оценки риска на уровне заемщика на основе скоринговой модели и получения его дециля риска с последующим расчетом лимитов с помощью отдельной математической модели с целью оптимизации процедуры выдачи PDL-займов.

Важность правильного назначения лимитов займов для МФО, минимизирующих кредитный риск и максимизирующих прибыль, особенно актуально в контексте управления МФО на основе модели юнит-экономики, что будет показано в следующем параграфе. Как отмечалось выше, для МФО стоимость привлечения клиентов достаточно высокая, и первый займ является убыточным, в том числе за счет более высокого уровня кредитных рисков и риска мошенничества для новых клиентов. Основную прибыль МФО формирует за счет повторных займов. Соответственно в портфеле МФО должно быть оптимальное соотношение первичных и повторных займов: около 30% первичных займов и около 70% повторных, если МФО не выбирает краткосрочную стратегию кратного роста объема выдачи. Увеличение доли повторных займов, объема

выдачи заемщикам с низким уровнем кредитного риска может происходить не только в количественном выражении, но и денежном за счет лимитной политики. Заемщикам с низким уровнем риска и низкой стоимостью привлечения необходимо увеличивать лимит займа, тем самым увеличивая средний чек. Таким образом, лимитная политика является важным инструментом управления кредитными рисками для МФО.

Меняя лимиты в каждой группе риска и, как следствие, меняя средний чек в каждой группе и в целом по портфелю, может меняться уровень дефолтов в каждой группе и доходность по всему портфелю. Суть регулирования лимитов заключается в том, что заемщикам в децилях с низким уровнем риска и недостаточным лимитом его надо повышать, и, наоборот, в высокорисковых децилях в случае перегруженного лимита риск будет расти еще больше, и лимит надо понижать. Предлагаемая математическая модель установки лимитов учитывает данные закономерности. Разберем подробно методику и алгоритм ее построения.

Задача построения модели лимитов – вывести функциональный вид зависимости оптимального лимита в зависимости от уровня риска заемщика и фактического среднего чека в каждом дециле риска. При этом уровень риска влияет на средний чек, но в свою очередь средний чек и также уровень риска влияют на устанавливаемый лимит (рисунок 4.1).

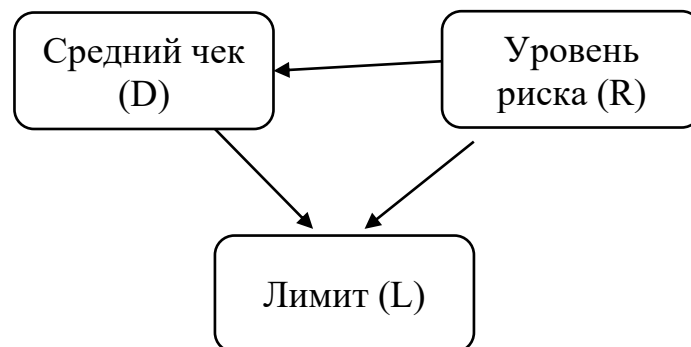


Рисунок 4.1 – Зависимости среднего чека, лимита и уровня риска

Источник: составлено автором

Между устанавливаемым лимитом в каждом дециле риска и фактически средней выданной суммой займа (средним чеком или средним основным долгом (ОД)) существует разница, измеряемая конверсией из одобрения СПР в выдачу (TR), но в целом эти показатели сильно коррелируют друг с другом. Средний чек определяется средним лимитом. Также для заемщиков с повышенным уровнем риска характерно полное использование предлагаемого лимита, в то время как заемщики с наименьшим уровнем кредитного риска часто не используют весь предлагаемый лимит, критично оценивая свои возможности по возврату долга.

Средний чек влияет на уровень риска в каждом дециле. Так же как уровень риска определяет назначаемый лимит. Поэтому оценка оптимального лимита будет функцией от среднего чека и уровня риска. А средний чек будет функцией от уровня риска. Поэтому модель оптимального лимита будет включать в себя сначала оценку оптимального среднего ОД от уровня риска на основе дополнительной модели.

Таким образом, разработанную автором методику моделирования оптимальных лимитов можно представить в виде следующих этапов:

- а) подготовка выборки;
- б) исследование данных;
- в) построение модели оптимального среднего ОД (среднего чека);
- г) построение модели оптимальных лимитов.

а) Подготовка выборки. На первоначальном этапе моделирования оптимальных значений кредитных лимитов необходимо собрать данные и подготовить исходную репрезентативную выборку. Начальной предпосылкой моделирования лимитов является наличие в МФО разработанных СПР на основе моделей оценки кредитного риска (скоринговых моделей), ранжирующих заемщиков по децилям риска, по которым есть исторические данные о выданных займах с результатами исполнения платежной дисциплины. Важные методологические аспекты подготовки исходной выборки – это выбор временного интервала, периода созревания выборки, объекта наблюдения, минимального объема выборки.

Временной интервал для формирования выборки должен быть взят за последний стабильный период работы СПР по заемщикам, чтобы все необходимые показатели для моделирования не имели экстремумов и выбросов. В данный период не должно проводиться тестирование новых СПР или экспериментов, значительно искажающих нормы уровня одобрения (AR) и риска по основным сегментам клиентов (например, новым и повторным клиентам). Если подобные эксперименты проводились, то договоры таких заемщиков желательно исключить из выборки данных.

Объектом наблюдения для решения поставленной задачи моделирования оптимального лимита должны являться группы заемщиков с агрегированием данных по ним по месяцам. Разбиение по группам необходимо формировать в разрезе риск-моделей, а также внутри моделей по децилям риска. Группы заемщиков с неприемлемым уровнем риска и лимитом 0 не должны браться для моделирования (отказные сегменты). Таким образом, наблюдением в получившейся выборке будет являться дециль риска за определенный месяц по выданным кредитам с достаточным числом заявок и приемлемым для конкретной МФО, согласно рискowej политике, уровнем риска.

Для расчета децилей риска по каждой скоринговой модели по историческим данным рассчитывают выбранные показатели оценки кредитного риска (NPL или FPD). Затем полученная выборка исходных заемщиков ранжируется по рассчитанным вероятностям дефолта (скоринговым баллам) и делится на десять частей, в каждую из которых входят примерно по 10% наблюдений. Разбиение на децили широко применяется в экономических и финансовых исследованиях [47], а также в задачах анализа больших массивов данных, например, в работе [63].

Выбор конкретного временного интервала определяется количеством сегментов клиентов, по которым принимаются решения о назначении лимитов, количеством СПР, ранжирующих заемщиков, количеством рискованных сегментов внутри СПР, и, исходя из этого, имеющимся общим объемом выборки данных. Рассмотрим пример такой группировки клиентов на примере одной МФО. При сегментировании клиентов на три группы (L1 – первый заем, L2 – второй заем,

L3+ – третий и более заем), одобрении по первичным клиентам только первым семи децилям и отказе только десятому децилю повторных клиентов при временном интервале 2 года будем иметь при месячной агрегации данных 600 наблюдений, что является достаточным объемом выборки:

$$7 \text{ децилей } L1 \cdot 24 + 9 \text{ децилей } L2 \cdot 24 + 9 \text{ децилей } L3 \cdot 24 = 600.$$

Более подробно об определении минимального объема выборки было рассмотрено в параграфе 1 главы 3.

Следует отметить, что данные последних месяцев от текущей даты нельзя брать для формирования выборочной совокупности. Как было уже рассмотрено выше, особенностью моделирования кредитных рисков в финансовой сфере является наличие периода созревания выборки, после которого фиксируются необходимые целевые показатели измерения кредитного риска (NPL, FPD). Горизонт формирования выборки 6 месяцев после даты выдачи микрозайма является оптимальным, поскольку за этот период достигается максимальный размер начисления процентов на долг, ограниченный законодательно. МФО, как правило, фиксируют доходность по займу по истечении 5–6 месяцев. За этот срок становится достоверно известно, сколько клиентов ушли в дефолт, и какая доходность получается по портфелю выданных займов.

Поскольку модель оптимального лимита формируется, исходя из уровня риска и ожидаемой доходности каждого сегмента клиентов, следующим этапом является выбор целевого показателя для измерения риска. Как было рассмотрено в первой главе работы наиболее распространенными показателями, используемыми для этой цели в риск-менеджменте МФО, являются показатели NPL и FPD.

На следующем шаге подготовки выборки необходимо рассчитать финальные показатели по каждому объекту наблюдения с агрегацией по месяцам, имеющие далее следующие обозначения в спецификациях предлагаемых моделей при выборе NPL в качестве показателя риска (при выборе альтернативного показателя риска фиксируется доля плохих клиентов по альтернативному показателю):

1) L – средний лимит (ден. ед.), т. е. среднее значение одобряемой СПР суммы займа;

2) R – процент клиентов, ушедших в просрочку более 30 дней по выданному займу, относительно общей численности клиентов за рассматриваемый месяц (%);

3) P – доходность (%);

4) D – средняя сумма займа или ОД (ден. ед.).

В таблице 4.2 приведен фрагмент примера подготовленной выборки для моделирования оптимальных лимитов на примере 600 наблюдений.

Таблица 4.2 – Пример сформированной выборки для моделирования оптимальных лимитов

Сегмент, i	Одобрено, шт.	Средний лимит, руб. (L_i)	Выдано, шт.	Средний ОД, руб. (D_i)	NPL30, % (R_i)	ROI, % (P_i)	Выдано/ Одобрено, TR, %
1	800	12000	700	8500	25%	7%	87,5
2	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
600	...	...	...	...	...	...	...

Источник: составлено автором, данные условные

б) *Исследование данных.* Следующим этапом моделирования является исследование описательных статистик рассчитанных показателей по сформированной выборке данных. Децили в формируемой выборке должны иметь достаточное количество займов для снижения доли ошибки при расчете исходного соотношения риск-доходности групп клиентов. Поэтому рекомендуется удалять из выборки сегменты с количеством займов ниже 5% квантиля распределения.

Также необходимо построение диаграммы рассеяния и исследование корреляции между показателем уровнем риска и разницы между показателями среднего лимита и среднего ОД. Наличие значимой корреляции будет

подтверждать априорную гипотезу о различии сценариев поведения заемщиков при разной одобряемой сумме займа. С помощью диаграммы рассеяния можно наглядно оценить роль лимитной политики МФО, поскольку отклонения суммы долга от заданного лимита имеют существенное значение в определении типа заемщика. Такая визуализация необходима для проведения первичного анализа по наиболее рисковым клиентам, которые склонны брать максимально возможную сумму займа с намерением отказаться от выплаты кредита.

Для создания математических правил установки лимитов в зависимости от уровня риска целесообразно ориентироваться на максимизацию средней доходности клиента в определенном дециле. Можно также исследовать зависимость доходности от уровня риска. В этих целях необходимо построить диаграмму рассеяния с линией тренда, отражающей исследуемую взаимосвязь. Такая визуализация поможет разделить всех заемщиков на кластеры по уровню риска и доходности. Кластеризация поможет таргетировать группу заемщиков для снижения уровня риска по ним за счет разрабатываемой модели лимитов.

Заключительным этапом разведочного анализа является построение диаграммы рассеяния между значениями среднего ОД и показателем риска, а также тепловой шкалы для уровня доходности в каждой точке на ней. Такая диаграмма поможет определить вид зависимости среднего ОД от уровня риска с учетом доходности сегмента клиента. В нашей конечной модели лимита оптимальный средний ОД будет являться фактором, который необходимо рассчитать отдельно, так как с ростом риска уменьшается разница между средним лимитом и средним ОД.

в) *Построение модели оптимального среднего ОД.* Оптимальный средний ОД является неизвестным параметром для построения финальной модели оптимального лимита. На данном этапе исследования задача состоит в построении соответствующей регрессии зависимости оптимального среднего ОД от уровня риска с учетом доходности по портфелю. Нужно учитывать риски и доходности сегментов клиентов, отдавая предпочтение тем, у кого доходности выше.

Прежде всего, на этапе исследования данных необходимо проанализировать форму зависимости среднего ОД от уровня риска, что позволит провести спецификацию данной регрессии. Подчеркнем, что теоретическая кривая такой зависимости должна иметь нелинейный тренд ввиду различий в выборе сумм займа разными типами клиентов. Так, заемщики с низким риском не склонны брать весь одобренный лимит, поскольку кредит им нужен для конкретных целей, они не пытаются максимизировать сумму долга, а кредитуются в соответствии с текущими потребностями. Напротив, высоко рискованные заемщики максимизируют основной долг, не намереваясь выплачивать кредит, поэтому при увеличении риска с определенного значения их средние суммы займов резко падают из-за ограничений СПР. Вследствие этого зависимость оптимального среднего ОД от уровня риска можно представить в общем случае полиномиальной моделью вида:

$$D_i = b_0 + b_1 R_i + b_2 R_i^2 + \dots + b_k R_i^k + \varepsilon_i, \quad (4.1)$$

где D_i – средняя сумма ОД в сегменте i ;

b_0 – константа модели;

b_i – числовые коэффициенты;

R_i – уровень риска в сегменте i ;

k – степень полинома;

ε_i – случайная ошибка модели.

$\hat{D}$ – оценка средней суммы ОД по уравнению (4.1) будет являться оптимальным средним лимитом. Определяя оценку показателя D , заметим, что МФО ориентируется на высокодоходных заемщиков, поэтому специфицируемая модель для оптимальной средней суммы займа должна быть взвешена по доходности (P). Это означает, что заемщики с большей доходностью получают больший вес при расчете среднего значения ОД. В связи с этим для оценки параметров модели (4.1) целесообразно воспользоваться взвешенным методом наименьших квадратов (ВМНК), т. к. он позволит «подстроить» регрессию под наиболее доходные группы.

Взвешенный метод наименьших квадратов широко используется в эконометрических исследованиях и финансовом моделировании, например, в работе [76].

Опишем алгоритм ВМНК для заданной регрессии (4.1). Каждому наблюдению присвоим свой вес, пропорциональный доходности. Предположим, что показатель доходности напрямую связан со случайной ошибкой модели (4.1), т. к. с увеличением риска по заемщику ожидается падение его доходности. Действительно, «хорошие» заемщики выплачивают кредит полностью и досрочно, а «плохие» допускают просрочки и не возвращают займы в полных объемах. Поэтому ВМНК позволяет поставить акцент на клиентах с высокой доходностью и избавиться от гетероскедастичности в модели (4.1).

Задача минимизации суммы квадратов ошибок при оценке параметров нашей модели сведется к минимизации данного выражения:

$$\sum_{i=1}^n \frac{1}{\sqrt{P_i}} (b_0 + b_1 R_i + b_2 R_i^2 + \dots + b_k R_i^k - D_i)^2 \rightarrow \min, \quad (4.2)$$

где P_i – доходность в сегменте i .

С учетом выражения (4.2) коэффициенты при R_i модели (4.1) оцениваются с помощью ВМНК. Оценку D в дальнейшем будем использовать в качестве фактора для расчета оптимальных лимитов. Следует также отметить, что в практике эконометрических исследований порядок полинома редко превышает 3-ю степень.

г) *Построение модели оптимальных лимитов.* После спецификации модели оптимальной средней суммы займа для определенного уровня риска по заемщику перейдем к спецификации модели регрессии для оптимальных лимитов, которые с одной стороны, минимизировали бы риски, а с другой – максимизировали бы доходность МФО. Для выбора спецификации модели будем руководствоваться следующими теоретическими предпосылками, эмпирически подтвержденные автором:

- Сумма среднего ОД стремится к назначаемому лимиту по мере роста уровня риска.

- Разность назначаемого лимита и суммы среднего ОД, поделенная на сумму назначаемого лимита – это величина, ограниченная снизу и сверху соответственно нулем и единицей:

$$\frac{L-D}{D} \in [0; 1]. \quad (4.3)$$

- Отношение (4.3) зависит также от оптимальной суммы ОД при одном и том же уровне риска.

Данные предпосылки позволяют сделать вывод о том, что для моделирования устанавливаемых лимитов можно воспользоваться моделью логистической регрессии. Использование логит-моделей или лог-линейных моделей достаточно распространено в разных сферах, они применяются не только в оценке кредитных рисков [218]. Часто они бывают полезны для прогнозирования вероятности банкротства компаний отдельных отраслей [30], а также нефинансовых компаний в целом [57].

Как было показано в параграфе 2 главы 3, математически модель логистической регрессии выражает зависимость логарифма шанса (логита) от линейной комбинации независимых переменных. Искомым результатом логистической регрессии будет соотношение оптимального среднего ОД к оптимальному лимиту в соответствии с нашей первой предпосылкой. Напомним, что, выбрав уровень риска, мы можем определить оптимальную сумму займа для заемщика, на которую нужно опираться при составлении лимитной политики, поскольку, как было отмечено выше, предельная разность между фактической и одобряемой суммами займов уменьшается с ростом риска. Таким образом, независимыми переменными в нашей логистической регрессии будут являться уровень риска и оптимальный средний ОД, линейная комбинация которых имеет вид:

$$z_i = b_0 + b_1 R_i + b_2 \hat{D}_i + \varepsilon_i, \quad (4.4)$$

где R_i – уровень риска в сегменте i ;

$\hat{D}_i$ – оценка оптимального среднего лимита ОД из модели (4.1) в сегменте i ;

b_0, b_1, b_2 – коэффициенты;

ε_i – случайная ошибка модели.

Тогда логит-модель будет иметь следующий вид:

$$\frac{\hat{D}_i}{L_i} = \frac{1}{1+e^{-z_i}}, \quad (4.5)$$

где L_i – оптимальный устанавливаемый лимит.

Таким образом, уровень риска и оптимальный средний ОД по модели (4.1) должны быть включены в регрессию (4.5) в качестве независимых факторов. Выразив линейную комбинацию (4.4) из уравнения (4.5), получим логистическую регрессию, записанную через логит:

$$\ln\left(\frac{L_i - \hat{D}_i}{\hat{D}_i}\right) = b_0 + b_1 R_i + b_2 \hat{D}_i + \varepsilon_i. \quad (4.6)$$

Для интерпретации коэффициентов модели логистической регрессии обычно используют экспоненциальную форму записи модели (4.6). В нашей задаче экспоненциальная форма записи поможет выразить в левой части уравнения целевую переменную (назначаемый оптимальный лимит). Таким образом, окончательная формула для расчета оптимальных лимитов примет вид:

$$\hat{L} = \frac{(1 + \exp(-(\hat{b}_0 + \hat{b}_1 R_i + \hat{b}_2 \hat{D}_i)))}{\exp(-(\hat{b}_0 + \hat{b}_1 R_i + \hat{b}_2 \hat{D}_i))} \cdot \hat{D}_i, \quad (4.7)$$

где $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$ – оценки коэффициентов модели (4.6), полученные методом наименьших квадратов (МНК).

Использование метода максимального правдоподобия для оценки параметров уравнения (4.6) не подходит для точной оценки устанавливаемого лимита, так как в нашем случае, в отличие от классической логистической регрессии, зависимая переменная не измеряется в бинарной шкале. Поэтому для оценки параметров модели (4.6) можно использовать МНК.

Предложенная методика моделирования оптимальных лимитов может использоваться отдельно для выборки первичных займов и отдельно для выборки повторных займов, а также для обобщенной выборки всех сегментов заемщиков. В этом заключается ее универсальность. Таким образом, МФО может использовать единую методику назначения лимитов для всех клиентов.

Иллюстрация расчетов по предложенной методике в результате апробации представлена в главе 5 диссертации.

Однако в большинстве МФО СПР кредитная политика для повторных клиентов в корне отличается от кредитной политики для новых клиентов. Как правило, для оценки кредитных рисков повторных клиентов используют отдельные модели, формирующие более «мягкую» кредитную политику. Уровень одобрения заявок повторных клиентов может колебаться в среднем от 85 до 100%. Фактически МФО выгоднее выдать займ с минимальным лимитом высокорисковому повторному клиенту, увеличив конверсию в выдачу и снизив стоимость привлечения одного клиента по всему портфелю. Поэтому для повторных клиентов модели оценки кредитного риска одобряют практически все заявки, а управление кредитным риском идет за счет регулирования лимитов. Сам по себе уровень кредитного риска для повторных клиентов ниже, чем у первичных, а портфель выдачи повторных займов по сумме должен превышать в среднем в 3 раза портфель первичных займов для положительной экономики МФО. Исходя из этого можно сделать вывод, что проблема назначения лимитов повторным клиентам более чем актуальна.

При формировании лимитной политики повторных клиентов МФО часто используют в отдельности или совместно два типа моделей оценки кредитных рисков:

- а) модели кредитного поведения (изменения кредитного поведения) – модели схожи с оценкой кредитных рисков первичных клиентов;
- б) поведенческие модели на основе внутренних данных МФО о платежной дисциплине по закрытым в ней займам [371, 372].

Также при назначении лимитов повторным клиентам необходимо учитывать мошеннические заявки так называемых «накрутчиков» лимитов. Суть такого поведения заемщика в следующем. Первоначально берется займ с минимальным лимитом и досрочно погашается, затем берется повторный займ и также досрочно или с минимальным сроком пользования деньгами закрывается и т. д. В большинстве МФО, чем больше в ней закрыто займов, тем больше

предлагаемый клиенту лимит. Через какое-то время при одобрении максимального лимита повторному клиенту, например, 30 000 руб., он намеренно не погашает его, уходя в дефолт.

Таким образом, учитывая факт, что в большинстве случаев повторная выдача кредитов в МФО производится без применения скоринговой модели с оценкой по поведенческой модели для соответствующего регулирования лимитов, что СПР должна «отсекать» мошеннические заявки и учитывать кредитную историю клиента в данной микрофинансовой организации.

Модель кредитного поведения по данным БКИ строится по методике, изложенной в параграфе 1 главы 3 настоящего диссертационного исследования, как и для первичных клиентов. Для повторных клиентов особенностью будет являться в наличии возможности расчета динамических переменных, например, разница между числом закрытых и открытых кредитов и т. д. Для этого расчет независимых переменных для модели ведется по кредитным отчетам, запрошенным в разные моменты времени при получении заявки на конкретный займ в этой МФО.

При использовании модели оценки кредитных рисков на основе данных БКИ назначение лимитов может происходить по децилям риска. Другой подход к определению лимитов в МФО – учитывать сумму заплаченных по всем предыдущим займам средств или по последнему займу. Чем больше заплатил клиент компании денег и чем выше его доходность на всем сроке жизни (LTV) тем агрессивнее растут лимиты для него при оформлении повторных займов.

Модель кредитного поведения в конкретной МФО может быть использована для назначения лимитов. Рассмотрим подробнее предлагаемую автором исследовании поведенческую модель. Будем предполагать стандартные условия для PDL займа повторным клиентам с минимальным лимитом 5000 руб. и максимальным 30 000 руб, максимальный срок займа составляет 30 дней. Также будем предполагать наличие стандартного правила в СПР повторных клиентов: клиент может получить новый займ только закрыв предыдущий даже с просрочкой.

Фундаментальные факторы кредитного риска при оформлении повторных займов – это наличие просрочки по предыдущему (-им) займу в этой МФО, а также минимально короткий срок пользования кредитом в случае «накрутки» лимитов. Поэтому при расчете лимита повторным клиентам будем использовать следующую бизнес-логику:

а) Наличие просрочки более 30 дней обнуляет лимит до минимального по продукту. Это предполагает возможность «реабилитации» клиента с большой просрочкой, но все-таки погасившего займ. Такие клиенты, возможно, не лучшие с точки зрения кредитного риска, но самые доходные для МФО. Поэтому МФО не выгодно их терять полностью. Снижение лимита до минимума позволяет сохранить клиента, учитывая высокую стоимость привлечения новых клиентов, снизив до минимума риск кредитных потерь.

б) При просрочке от 20 до 30 дней лимит не повышается. В этом случае сохраняется лимит, установленный в последнем закрытом займе.

в) При отсутствии просрочки или просрочке до 20 дней анализируется срок использования кредита и сумма заплаченных клиентом средств с учетом всех дополнительных продуктов. Чем больше заплатил совокупно клиент, тем больше будет увеличиваться лимит следующего займа. С другой стороны, на сумму лимита будет влиять реальный срок пользования деньгами, чем он ниже – тем меньше будет предлагаемый лимит.

В расчете необходимо использовать приведенный срок использования кредита: чем больше клиент пользуется кредитом и чем меньше разница между реальным сроком и сроком по договору, тем выше коэффициент повышения лимита. Это позволяет не повышать лимит клиентам, которые берут и быстро гасят небольшие суммы в расчете на то, чтоб потом получить много и не платить. Также такие заемщики склонны отказываться от дополнительных услуг МФО. Флаг наличие отключения дополнительных услуг может быть включен как отдельный фактор в модель, влияющий на понижение лимита.

Порог просрочки 20 дней выведен эмпирически на основе практического опыта автора, оптимальная точка числа дней просрочки может быть выбрана в

среднем в диапазоне от 10 до 20 дней в зависимости от выстроенных бизнес-процессов взыскания МФО.

Далее рассмотрим математическую логику расчета нового лимита (*Limit*). Расчет лимита по повторному займу осуществляется по следующим правилам.

а) Если просрочка по закрытому предыдущему займу больше 30 дней, то:

$$Limit = Limit_{min}, \quad (4.8)$$

где $Limit_{min}$ – минимальный лимит по продукту.

б) Если просрочка по закрытому предыдущему займу от 20 до 30 дней включительно, то:

$$Limit = Limit_{prev}, \quad (4.9)$$

где $Limit_{prev}$ – лимит по последнему закрытому займу.

в) Если просрочка по закрытому предыдущему займу от 0 до 19 дней включительно, то:

$$Limit = \min (Limit_{max}; Limit_{calc}), \quad (4.10)$$

$Limit_{max}$ – максимальный лимит по продукту;

$Limit_{calc}$ – расчетный лимит на основе кредитного поведения в конкретной МФО по формуле (4.11).

Расчетный лимит оценивается по формуле:

$$Limit_{calc} = \text{round} (SUM_{prev} \cdot (1 + K_1 \cdot K_2); 100), \quad (4.11)$$

где $\text{round} (X; 100)$ – функция математического округления числа X до значений кратных 100 рублей;

SUM_{prev} – сумма, уплаченная по закрытому предыдущему займу в конкретной МФО;

K_1 – коэффициент повышения лимита в зависимости от количества дней просрочки в предыдущем займе:

$K_1 = 0$, если просрочка составляет от 10 до 19 дней включительно;

$K_1 = 0,5$, если просрочка составляет от 3 до 9 дней включительно.

Если просрочка в предыдущем займе составляет от 0 до 2 дней включительно, коэффициент K_1 определяется по таблице 4.3. Отметим, что

округление в формуле (4.11) идет до 100 рублей, но можно выбрать округление кратко 500 или 1000 рублей. Также границы просрочек могут быть изменены в конкретной МФО. Обычно они коррелируют со сроками стадий работы отделов взыскания и взаимодействия с клиентом.

Таблица 4.3 – Пример определения коэффициента K_1 в зависимости от суммы выплаты по предыдущему закрытому займу

Показатель	Группировка					
Сумма выплаты по предыдущему закрытому займу, руб. (SUM_{prev})	до 5000	от 5000 до 7000	от 7000 до 9000	от 9000 до 11000	от 11000 до 13000	свыше 13000
Коэффициент повышения лимита (K_1)	1,5	1,2	0,9	0,5	0,3	0,2

Источник: составлено автором

В представленном примере предполагается, что просрочка до 2 дней является технической просрочкой, возникающей в базе из-за разницы часовых поясов и технических сбоев на этой основе, из-за забывчивости клиента и др. причин. До 10 дней просрочки с клиентом взаимодействуют операторы через напоминания в личном кабинете и смс, с 10 дня просрочки работу ведут коллектора по одному сценарию «мягкого» взаимодействия, а с 20 дня по «жесткому» сценарию.

Значения коэффициента K_1 в таблице 4.3. можно изменить, исходя из анализа среднего чека по повторным клиентам и зависимости уровня риска от суммы лимита для конкретной МФО.

Коэффициент K_2 – функция от соотношения срока займа по договору и реального срока займа, вычисляется по формуле:

$$K_2 = (T_{real} / T_{dogovor} * K)^2, \quad (4.12)$$

$$\text{где } K = 1 - \frac{T_{max} - T_{dogovor}}{2 \cdot T_{max}}; \quad (4.13)$$

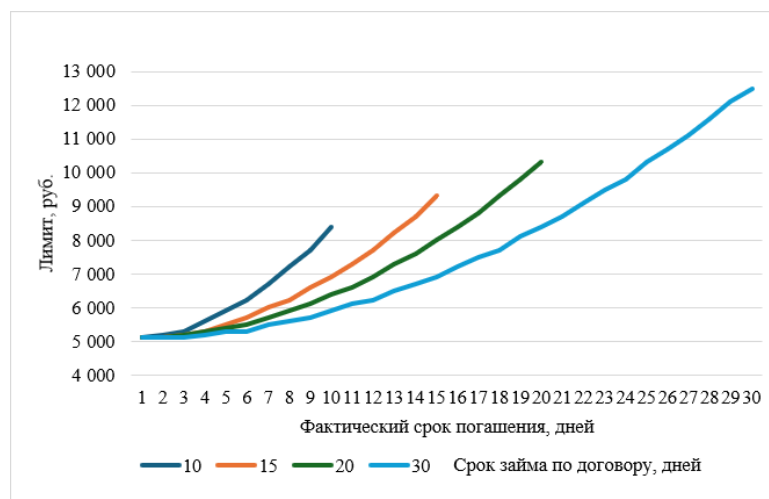
T_{real} – фактический срок займа в днях;

$T_{dogovor}$ – срок займа по договору в днях;

T_{max} – максимальный срок займа по договору в днях ($T_{max} = 30$).

Проиллюстрируем зависимость изменения лимитов повторных клиентов по предлагаемой модели в зависимости от срока пользования займом и суммой погашенных средств на условных данных. Графики изменения лимита повторного займа при разных суммах предыдущего займа, контрактных и реальных сроках займа представлены на рисунке 4.4. Во всех рассмотренных вариантах предполагается просрочка по займу не более 2-х дней.

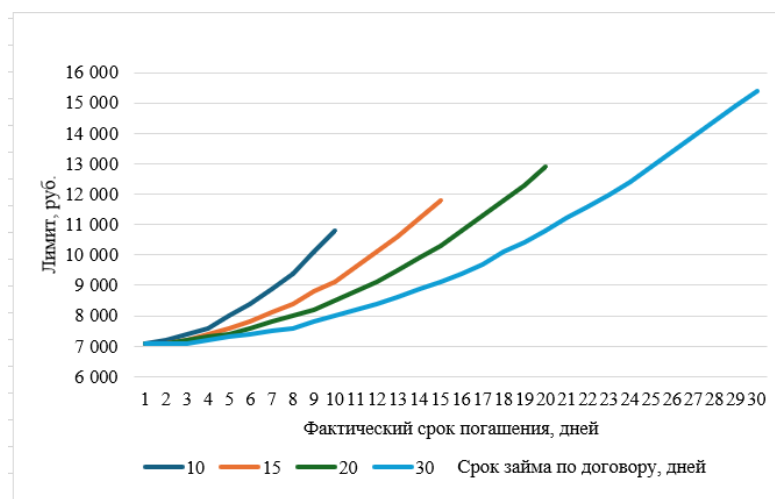
Описанный выше алгоритм за счет квадратичной функции в формуле (4.12) позволяет получить нелинейную связь изменения лимита при повторном займе в зависимости от двух факторов – реального срока займа и суммы погашенных средств (рисунок 4.2 а–е).



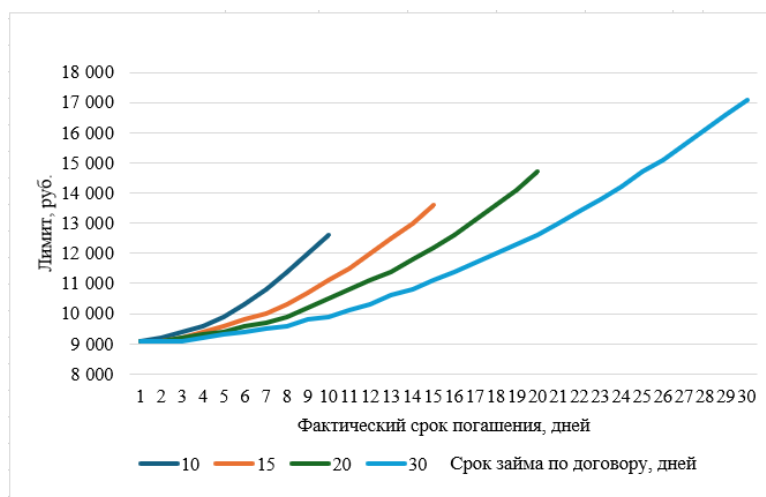
а) Сумма выплаты по предыдущему закрытому займу – 5 000 руб.

Рисунок 4.2 – Изменение лимита повторного займа при разных суммах предыдущего займа, контрактных и реальных сроках займа

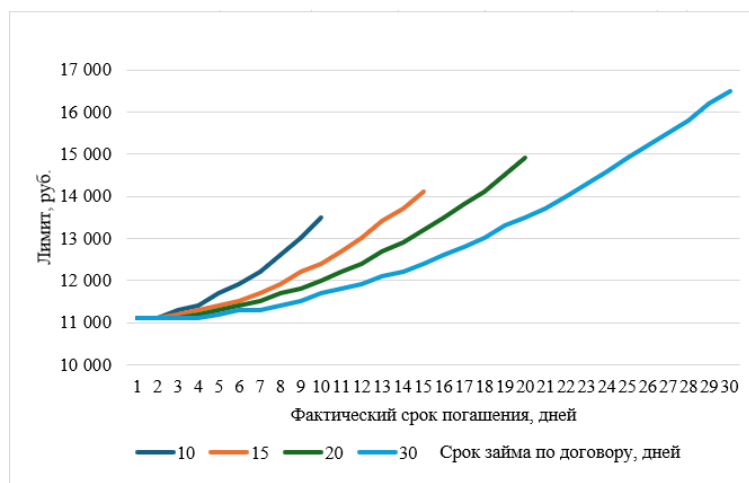
Источник: составлено автором



б) Сумма выплаты по предыдущему закрытому займу – 7 000 руб.



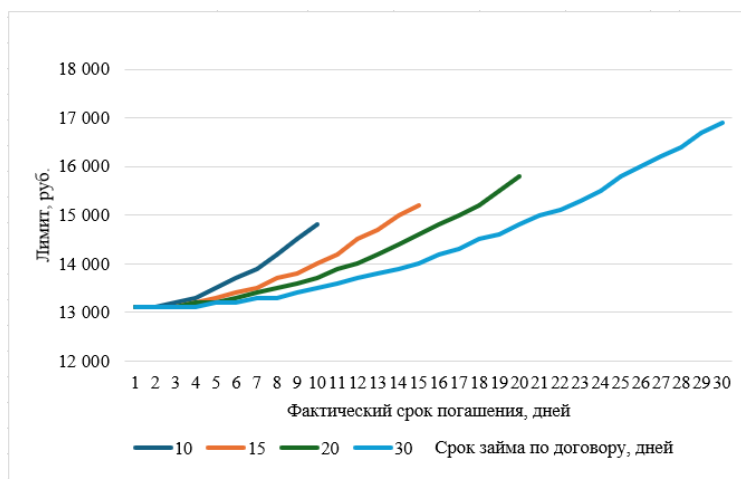
в) Сумма выплаты по предыдущему закрытому займу – 9 000 руб.



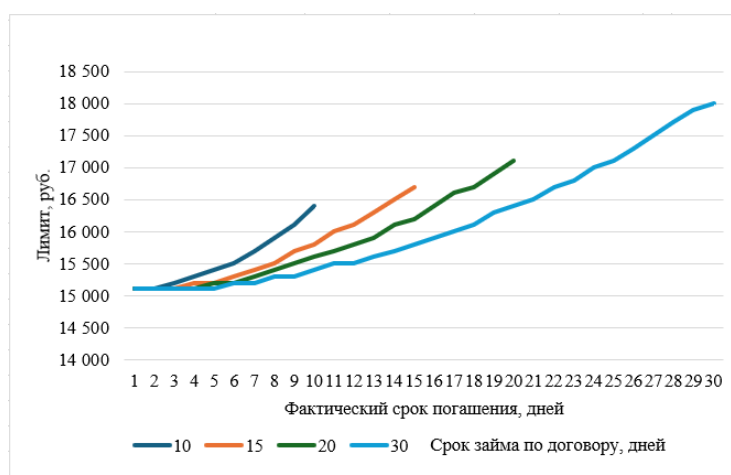
г) Сумма выплаты по предыдущему закрытому займу – 11 000 руб.

Рисунок 4.2 – Изменение лимита повторного займа при разных суммах предыдущего займа, контрактных и реальных сроках займа (Продолжение)

Источник: составлено автором



д) Сумма выплаты по предыдущему закрытому займу – 13 000 руб.



е) Сумма выплаты по предыдущему закрытому займу – 15 000 руб.

Рисунок 4.2 – Изменение лимита повторного займа при разных суммах предыдущего займа, контрактных и реальных сроках займа (Продолжение)

Источник: составлено автором

По представленным графикам рисунка 4.2 видно, что увеличение лимита займа происходит в зависимости от договорных сроков – 10 800 руб. при сроке 10 дней, 9 100 руб. при сроке 15 дней, 8 500 руб. при сроке 20 дней и 8 000 руб. при сроке 30 дней. Т. е. учитывается не абсолютная разница между фактическим и договорным сроком, а их отношение.

Таким образом, описанный алгоритм позволяет получить нелинейный максимальный рост лимита при погашении клиентом займа в срок при займе на 30 дней и нулевой или минимальный рост лимита при досрочном погашении

небольших сумм. Тем самым для доходных лояльных клиентов лимит будет расти агрессивно, а для «накрутчиков» лимита при досрочном погашении фактически не меняться, что делает данную схему мошенничества не целесообразной.

Таким образом, в рамках настоящего параграфа была представлена математическая модель установки оптимальных лимитов для новых и повторных заемщиков, основанная на комплексном применении результатов оценки кредитного риска на уровне заемщика на основе скоринговой модели и получения его дециля риска с последующим расчетом лимитов, а также поведенческая модель установки лимитов для повторных заемщиков на основе их кредитного поведения в конкретной МФО. Данные модели могут использоваться МФО для управления кредитными рисками путем регулирования лимитной политики.

4.2 Управление кредитными рисками микрофинансовых организаций на основе модели юнит-экономики

Как было определено ранее, базой системы принятия решений для управления рисками в МФО выступают либо общие скоринговые модели БКИ, либо собственные модели скоринга, разрабатываемые для формирования кредитного рейтинга заемщика и оценки кредитного риска МФО на основе классических статистических методов или алгоритмов машинного обучения. В этом случае управление кредитным риском на уровне всего портфеля происходит через управление таковым на уровне каждого заемщика.

Ключевой моделью в рамках системы управления кредитными рисками МФО остаются модели оценки вероятности дефолта заемщиков с целью снижения уровня просрочек по займам «плохих» клиентов и увеличения уровня одобрения «хороших» займов. При этом базой максимизации финансовых результатов МФО является сравнение потенциального процентного дохода по займам с ожидаемыми кредитными потерями, основным фактором которого выступает вероятность дефолта заемщика. Использование только такого подхода

имеет ограничения, которые заключаются в непринятии во внимание других критических факторов, оказывающих влияние на финансовые результаты деятельности МФО.

В современных условиях принятие решений в управлении кредитными рисками МФО возможно не только посредством стандартной оценки вероятности дефолта заемщика, а путем тестирования кредитной политики организации через общую модель юнит-экономики, которая позволяет оценивать финансовый результат на всех этапах жизненного цикла клиента МФО. В данном контексте следует отметить, что модель юнит-экономики является не альтернативой рассмотренному традиционному подходу использования скоринговых моделей, а выступает дополнением для принятия решений по управлению кредитными рисками МФО, что определяет эффективность использования комплексного подхода при принятии соответствующих решений. Концепция модели юнит-экономики реализует подход моделирования кредитного риска и параметров бизнес-метрик по всему кредитному портфелю МФО и была представлена автором в работах [105, 117].

В параграфе 3 главы 1 было определено, что в настоящее время как в российской, так и в зарубежной научной среде отсутствуют публикации по вопросам юнит-экономики как в целом в финансовой сфере (включая банковские, страховые, лизинговые организации), так и в отношении микрофинансовых организаций, в частности. Актуальность применения юнит-экономики в МФО определяется тем, что первый заем является убыточным с учетом затрат на привлечение клиентов (затрат на маркетинг), а прибыль МФО получает от повторных займов клиентов.

Многие российские МФО применяют в своей практике юнит-экономику для управления кредитными рисками на основе опыта и знаний риск-менеджеров, в том числе опыта применения юнит-экономики в других отраслях. Объем знаний о юнит-экономике в МФО сосредоточен во внутренних инструкциях, методических документах и других наработках и расчетов реальных компаний, которые закрыты для научной среды. Использование юнит-экономики на практике

предполагает множество расчетов и сбор статистики по всем бизнес-метрикам МФО за длительный период не менее 2-х лет. Фактически концепция юнит-экономики неразрывно связана с концепцией полной цифровизации бизнеса. Поэтому управление кредитными рисками на основе юнит-экономики требует определенной степени «зрелости» бизнеса и наличие соответствующей системной инфраструктуры, штата аналитиков данных и риск-аналитиков в МФО.

Таким образом, необходимо исследовать особенности юнит-экономики в разрезе микрофинансовой деятельности и сформулировать соответствующую математическую модель.

Первоначально более подробно разберем теоретические основы концепции юнит-экономики. Концепция юнит-экономики базируется на базовых законах экономической теории, рассмотренных в работах [24, 25, 56, 151].

В отношении сущности понятия «юнит-экономика» (англ. Unit Economics, Economics of One Unit) в настоящее время в научных публикациях доминирует три основных подхода:

а) Юнит-экономика как метод анализа прямых доходов и затрат бизнеса, мерой которых выступает определенная единица (юнит, unit), приносящая ценность бизнесу и измеримая в количественных показателях [40].

б) Юнит-экономика как метод экономического моделирования, целью которого является определение эффективности бизнес-модели посредством оценки прибыльности конкретной единицы [3].

в) Юнит-экономика как способ оценки получения дохода, превышающего затраты на привлечение и удержание отдельно взятого клиента [22].

Несмотря на разницу в подходах, основополагающим принципом юнит-экономики выступает соотношение успешности бизнеса с возможностью конкретной единицы приносить прибыль. Выбранной для проведения моделирования и оценки прибыльности и/или эффективности бизнеса единицей (юнитом) может выступать любой элемент, который приносит ценность бизнесу как разница между соответствующими доходами и расходами на данную единицу.

Основное преимущество юнит-экономики заключается в упрощении оценки показателей эффективности бизнеса в расчете на выбранную единицу.

В качестве таких единиц (юнитов), как правило, выделяют клиента, товар или услугу, конкретное подразделение (или сотрудника). В соответствии с этим определены три базовые модели юнит-экономики, которые могут быть использованы для любых видов бизнеса [93]:

- клиентская модель юнит-экономики, которая базируется на основных показателях взаимодействия бизнеса с клиентом – средний чек, средняя стоимость привлечения клиента (англ. Customer Acquisition Cost, CAC), пожизненная ценность клиента (англ. Life-Time Value, LTV) и другие;

- продуктовая модель юнит-экономики определяется показателями, относящимися к конкретному товару или услуге – цена товара/услуги (в т. ч. для конкретного покупателя с учетом скидок и т. д.), показатели постоянных и переменных затрат и другие;

- организационная модель юнит-экономики строится на показателях оценки качества бизнес-модели с точки зрения ее организационной структуры и рассчитываются по отдельному бизнес-подразделению, например отделу продаж - количество и объем продаж, количество обслуживаемых клиентов, скорость перехода лида в покупателя и т. д.

Наиболее распространенной моделью юнит-экономики, нашедшей применение в различных видах бизнеса, является клиентская модель, ключевыми показателями которой выступают [17]:

а) Доход от клиента или средний доход от клиента (англ. Average Revenue Per Customer, ARPC) – общий доход, полученный от клиента. Стандартная формула расчета показателя ARPC имеет следующий вид:

$$ARPC = \frac{R_{total}}{C}, \quad (4.14)$$

где R_{total} – общий объем доходов организации, ден. ед. Для МФО такой доход будет включать процентный доход, сборы и доход от дополнительных услуг;

C – количество клиентов, шт.

б) Стоимость привлечения клиента (англ. Customer Acquisition Cost, CAC) – расходы, связанные с приобретением новых клиентов (включая расходы на маркетинг и продажи).

Стандартная формула расчета показателя CAC имеет следующий вид:

$$CAC = \frac{MSC_{total}}{CA}, \quad (4.15)$$

где MSC_{total} – общий объем расходов на продажи и маркетинг, ден. ед.;

CA – количество привлеченных клиентов, шт.

в) Пожизненная ценность клиента (англ. Life-Time Value, LTV) – прогнозируемая чистая прибыль, которую получит бизнес за все время своих отношений с клиентом. Данный показатель не имеет единой формулы расчета и зависит от вида, особенностей ведения бизнеса, целей, а также имеющихся в распоряжении организации данных. Как правило, используют следующие варианты:

$$LTV = ARPC \times ACL, \quad (4.16)$$

где ACL – средний срок жизни клиента (англ. Average Customer Lifetime), мес. Для МФО это период, в течение которого клиент продолжает брать новые займы.

$$LTV = AGM \cdot PF \cdot AP \cdot ACL, \quad (4.17)$$

где AGM – средняя валовая прибыль на одного клиента (англ. Average Gross Margin), которая определяется как разница между полученным доходом и затратами организации в расчете на одного клиента, ден. ед.;

PF – частота покупок (англ. Purchase Frequency), шт. Данный показатель указывает на среднее количество покупок (для МФО это количество займов) одного клиента за определенный временной интервал;

AP – средний чек (англ. Average Purchase Amount), ден. ед.

Для МФО средний чек – это средняя сумма займа.

В рамках модели юнит-экономики МФО наибольшее влияние на LTV имеют следующие показатели:

- процентная ставка в день, %;
- средний чек, ден. ед.;
- уровень проникновения продлений, %;
- конверсия из заявки в займ, %;
- уровень кредитных потерь, %;
- уровень лояльности, %;
- доля перехода в повторный займ, %;
- средний срок жизни клиента, мес.

г) Коэффициент оттока клиентов (англ. Customer Churn Rate, CCR) – доля клиентов, которые прекращают свои отношения с бизнесом в течение определенного периода. Показатель рассчитывается по формуле:

$$CCR = \frac{C_0 - C_1}{C_0}, \quad (4.18)$$

где C_0 – количество клиентов на начало периода, шт.;

C_1 – количество клиентов на конец периода, шт.

Также в клиентской модели юнит-экономики используют и другие показатели в зависимости от вида, особенностей ведения бизнеса, целей, а также имеющихся в распоряжении организации данных.

На основе полученных соотношений и сценариев принимается решение по дальнейшему развитию бизнеса с точки зрения возможности повышения показателей. Например, оптимизация показателя LTV в МФО реализуется за счет предоставления персонализированного обслуживания, кросс-продаж и других инструментов повышения клиентской лояльности.

Эффективность бизнеса в рамках клиентской модели юнит-экономики, как правило, определяется соотношением метрик LTV/CAC, которое представляет собой совокупную прибыль, которую может получить бизнес за весь срок взаимоотношений с клиентом относительно первоначальной стоимости его привлечения [18]. При этом существует три сценария развития бизнеса в зависимости от соотношения данных показателей (см. таблицу 4.4).

Таблица 4.4 – Сценарии развития бизнеса в зависимости от соотношения метрик LTV/CAC в рамках клиентской модели юнит-экономики

Сценарий развития бизнеса	Соотношение метрик	Характеристика
Сильная (положительная) юнит-экономика	$CAC < LTV$ (наилучшее соотношение 1 : 3)	Увеличение бюджета на привлечение клиента существенно ниже полученной от клиента прибыли на всем жизненном цикле взаимоотношений с ним
Слабая (нулевая) юнит-экономика	$CAC = LTV$	Стоимость привлечения клиента равна полученной прибыли, разница между показателями равна нулю, бизнес стагнирует
Отрицательная юнит-экономика	$CAC > LTV$	Бюджет, направленный на привлечение клиентов, существенно превышает объемы полученной прибыли, что приводит к отрицательному финансовому результату и убыточности бизнеса. Необходим пересмотр бизнес-модели

Источник: Составлено автором

Соотношение показателей LTV/CAC можно определить как критически важный фактор устойчивого роста бизнеса, точная оценка которого позволяет выстраивать стратегию оптимизации данного соотношения с целью получения максимальной прибыли при одновременном регулировании уровня кредитного риска.

Один из основоположников юнит-экономики в рамках ее адаптации к компаниям сферы SaaS (Soft as a Service, предоставление программного обеспечения посредством онлайн-каналов) Дэвид Скок [359] выделил еще одну важную метрику юнит-экономики – AMPPU (средняя маржа с одного платящего пользователя), которая находится в прямой зависимости от конверсии лида в покупателя. При этом ключевым соотношением юнит-экономики выступает соотношение метрик $CAC / AMPPU$ по следующим сценариям (см. таблицу 4.5).

Вышеуказанные показатели с соответствующей адаптацией и расчетами могут быть использованы для моделирования юнит-экономики в МФО.

Главная особенность экономики МФО в отличие от банковской сферы заключается в том, что основную прибыль МФО получает от повторных клиентов, которых достаточно сложно удержать по причине короткого срока займа. Более того, политика управления кредитными рисками в рамках экономики МФО нацелена на увеличение конверсии в одобрение по клиентам с наименьшим уровнем риска.

Таблица 4.5 – Сценарии юнит-экономики в зависимости от соотношения метрик $CAC / AMPPU$ по методике Д. Скока

Соотношение метрик	Характеристика
1 : 1	Бюджет затрат на привлечение клиента существенно ниже полученной маржи
1 : 2	Низкая окупаемость затрат на привлечение клиентов
1 : 3	Оптимальное соотношение метрик. Бизнес-модель работает продуктивно
1 : 4	Высокая эффективность бизнеса. Возможность реализации агрессивных маркетинговых кампаний

Источник: Составлено автором по [359]

Использование модели юнит-экономики в МФО имеет эффективность с точки зрения управления кредитными рисками по следующим аспектам:

- своевременно выявлять проблемы, препятствующие увеличению прибыльности бизнеса;
- определять оптимальные стратегии увеличения конверсии в одобрение;
- прогнозировать затраты на привлечение заемщиков;
- оценивать потенциал конкретного кредитного продукта в рамках долгосрочной ценности для клиента;
- анализировать эффективность запуска новых кредитных продуктов и возможности масштабирования бизнеса.

Проведенное исследование позволило дать соответствующую трактовку понятия «юнит-экономики» по отношению к особенностям ведения бизнеса МФО: юнит-экономика – это система принятия решений в МФО при управлении кредитным риском, нацеленная на получение максимальной прибыли с одного клиента в долгосрочной перспективе на всем сроке «жизни» клиента. При этом главной особенностью экономики МФО является то, что немаловажное значение наряду с уровнем риска имеет конверсия из заявки в одобрение и другие бизнес-метрики.

Еще раз подчеркнем, как было определено в параграфе 3 главы 1, в настоящее время в научной среде отсутствует методика оценки юнит-экономики по отношению к специфике деятельности МФО, что определяет актуальность выработки эффективной математической модели юнит-экономики МФО, учитывающей специфические показатели деятельности МФО в рамках системы управления рисками при кредитовании заемщиков.

Математическая модель юнит-экономики МФО строится на основе клиентской модели юнит-экономики, которая базируется на основных показателях взаимодействия бизнеса микрофинансовой организации с заемщиком. Таким образом, базовым элементом (единицей) юнит-экономики МФО выступает клиент (заемщик).

Основной формулой расчета в рамках экономико-математической модели юнит-экономики МФО является:

$$UE = UE_1 + UE_2 \cdot RR, \quad (4.19)$$

где UE – итоговое значение юнит-экономики МФО, ден. ед.;

UE_1 – юнит-экономика первого займа, ден. ед.;

UE_2 – юнит-экономика второго (повторного) займа, ден. ед.;

RR – показатель возврата клиентов в повторную заявку (Retention Rate), %.

Показатель возврата клиентов в повторную заявку представляет собой прогнозное значение коэффициента удержания клиентов на следующие 24 месяца или альтернативный период.

Расчет показателей UE_1 и UE_2 базируется на следующей разнице показателей:

$$UE_i = AR_i - AC_i, \quad (4.20)$$

где индекс $i = 1, 2$ и отражает первичный или повторный заем;

AR_i – средняя выручка с займа, ден. ед.;

AC_i – средние затраты на привлечение и обслуживание займа, ден. ед.;

Показатель средней выручки рассчитывается по формуле:

$$AR_i = AIP_i + CS_i + RO_i, \quad (4.21)$$

где AIP_i – средняя комиссия, ден. ед.;

CS_i – вклад кросс-продуктов ден. ед.;

RO_i – вклад пролонгации займа ден. ед.;

Средняя комиссия (AIP_i) определяется как средняя прибыль от процентов с одного займа. В свою очередь средняя прибыль от процентов вычисляется по основным характеристикам кредитных договоров. В практике МФО основными характеристиками кредитных договоров являются: сумма займа (средний чек), срок займа и дневная процентная ставка.

Таким образом, формула вычисления средней комиссии (средней прибыли от процентов с одного займа) выглядит следующим образом:

$$AIP_i = AD_i \cdot AT_i \cdot AI_i, \quad (4.22)$$

где AD – средняя сумма займа (средний чек), ден. ед.;

AT – средний срок займа, дней;

AI – средняя процентная ставка в день, %.

Средняя процентная ставка по займам ограничивается на законодательном уровне. Процентные ставки регулируются Федеральным законом от 21.12.2013 № 353-ФЗ «О потребительском кредите (займе)». С июля 2023 года, согласно пункту 1 статьи 1 Федеральный закон от 29.12.2022 № 613-ФЗ «О внесении изменений в Федеральный закон «О потребительском кредите (займе)» [68] процентная ставка по потребительским кредитам не может превышать 0,8% в день. Для упрощения модели данный показатель взят на уровне максимального значения. При этом

ситуация с бесплатным первым займом не рассматривается (в практике МФО – лидеров-рынка принято для привлечения новых клиентов с низким профилем риска выдавать первый заем без процентов на ограниченный срок и при условии своевременного погашения займа).

Также при проведении расчетов можно использовать усложненный вариант вычисления средней комиссии – реальной средней прибыли с учетом пролонгации и досрочных погашений ($AlPr_i$), где средний срок договора займа определяется как реальный средний срок с учетом пролонгации и досрочных погашений, вычисляемый по формуле:

$$Tr_i = T_i \cdot (1 - ER_i) \cdot (1 - RN_i) + RN_i \cdot (AT_i + ATR_i), \quad (4.23)$$

где ATr_i – реальный срок займа без учета открытых займов с просрочками, но с учетом продлений и досрочных погашений, дней;

T_i – срок займа по кредитному договору, дней;

ER_i – уровень досрочного погашения, %;

RN_i – уровень проникновения продлений, %;

AT_i – средний срок займа, дней;

ATR_i – средний срок продлений, дней.

Таким образом, реальный средний срок займа с учетом пролонгации и досрочных погашений определяется как среднее между сроком займа с учетом досрочного погашения среди тех, кто не продлил займ, и срок займа с учетом продлений среди тех, кто продлил.

Вклад кросс-продуктов определяется как прибыль, полученная от реализации дополнительных продуктов/услуг при заключении договора займа. Таким образом, вклад кросс-продуктов рассчитывается по формуле:

$$CS_i = OS_i \cdot OSM_i \cdot OSV_i + OSR_i \cdot OSRM_i \cdot OSRV_i \cdot RN_i \cdot NR_i, \quad (4.24)$$

где OS_i – уровень проникновения дополнительных услуг, %;

OSM_i – маржинальность дополнительных услуг, %;

OSV_i – средняя стоимость дополнительных услуг, ден. ед.;

OSR_i – уровень проникновения дополнительных услуг при продлении, %;

$OSRM_i$ – маржинальность дополнительных услуг при продлении, %;

$OSRV_i$ – средняя стоимость дополнительных услуг при продлении займа, ден. ед.;

RN_i – уровень проникновения продлений, %;

NR_i – среднее количество пролонгаций, шт.

В качестве кросс-продуктов для модели могут быть взяты продукты страхования жизни и работы, страхования карты при продлении договора займа и другие дополнительные услуги. При этом уровень проникновения продуктов по дополнительному страхованию определяется, как количество заключенных полисов по страхованию к общему количеству полисов, действующих или продленных (для страхования карт). Средние суммы за страхование, количество полисов, пролонгаций определяются на основании фактических значений за анализируемый период.

Вклад пролонгации вычисляется по формуле:

$$RO_i = AT_i \cdot AI_i \cdot RN_i \cdot ATR_i \cdot NR_i. \quad (4.25)$$

Все показатели, кроме процентной ставки, рассчитываются на основе фактических значений бизнес-метрик за соответствующий месяц, для будущих периодов – предполагаемые (прогнозируемые целевые) значения показателей МФО.

Далее определяются средние затраты на привлечение и обслуживание займа, расчет которых осуществляется по формуле:

$$AC_i = ACC_i + CL_i + LSC_i + MTC_i + AFC_i + ADCC_i + IC_i, \quad (4.26)$$

где ACC_i – расходы на привлечение, ден. ед.;

CL_i – кредитные потери, ден. ед.;

LSC_i – затраты на скоринг одной выдачи, ден. ед.;

MTC_i – расходы на перевод денежных средств заемщику, ден. ед.;

AFC_i – средние постоянные издержки на один займ, ден. ед.;

$ADCC_i$ – средние расходы на взыскание на один займ, ден. ед.;

IC_i – процентные расходы (фондирование), ден. ед.;

Расходы на привлечение отражают среднюю стоимость привлечения клиента для выдачи одного займа. В рамках настоящей модели для упрощения расчетов данный показатель является константой.

Кредитные потери определяются как уровень потерь МФО, связанных с дефолтом заемщика или невыполнением обязательств заемщика по кредитному договору. Кредитные потери рассчитываются по формуле:

$$CL_i = (AD_i + AIP_i) \cdot NPL_i \quad (4.27)$$

где NPL_i – чистые кредитные потери по показателю NPL10 net, %. Данный показатель представляет собой долю только открытых займов с просрочкой более 10 дней (NPL10 net). Взыскание просроченной задолженности и продажа по цессии сокращают данный показатель. В рамках настоящей модели показатель чистых кредитных потерь берется именно по показателю NPL10 из предположения о том, что после 10 дней просрочки по бизнес-процессу МФО начинает работу отдел взыскания просроченной задолженности. Но в модели юнит-экономики конкретной МФО может быть взят другой показатель риска.

Затраты на скоринг одной выдачи вычисляются по формуле:

$$LSC_i = SC_i \cdot PFR_i, \quad (4.28)$$

где SC_i – средняя стоимость принятия решения по одной заявке, ден. ед.;

PFR_i – количество заявок в СПР, необходимое для выдачи одного займа (множитель воронки продаж), шт.

Средняя стоимость принятия решения по одной заявке определяется соотношением совокупных расходов на покупку данных у дата-провайдеров (бюро кредитных историй, телеком и т. д.) и количества рассмотренных заявок на выдачу займов. Для простоты расчета в рамках разработанной модели предполагается фиксированная стоимость по каждому источнику и ее неизменность от объема выдачи в течение года.

Количество заявок в системе принятия решений (СПР), необходимое для выдачи одного займа, определяется как множитель (коэффициент) воронки продаж и вычисляется по формуле:

$$PFR_i = \frac{1}{AR_i \cdot TR_i}, \quad (4.29)$$

где AR_i уровень одобрения заявок (Approval Rate), %;

TR_i – уровень выдачи займов (Take Rate), %.

В рамках рассматриваемой экономико-математической модели также необходимо учитывать структуру потока клиентов, которая определяется следующими показателями: количество заявок, поданных на выдачу займа, количество одобрений на выдачу и итоговое количество выдач займа. Уровень одобрения заявки показывает конверсию из заявки в СПР в одобрение, т. е. отношение одобренных заявок на выдачу займа ($ACRN_i$) к общему количеству полученных заявок (CRN_i). Уровень выдачи займов представляет собой конверсию из одобренной заявки в фактическую выдачу, т. е. отношение количества выданных займов (CN_i) к числу одобренных заявок на выдачу ($ACRN_i$).

Расходы на перевод денежных средств заемщику вычисляется по формуле:

$$MTC_i = (AD_i + (AD_i + AIP_i) \cdot (1 - NPL_i - RPR_i)) \cdot MTF_i, \quad (4.30)$$

где RPR_i – уровень досрочного погашения (от срока), %. Данный показатель определяется соотношением количества дней, на сколько раньше закрыт займ к сроку займа по договору;

MTF_i – комиссия за перевод денежных средств заемщику, %.

Расчет уровня досрочного погашения производится исключительно по займам, по которым было либо досрочное погашение, либо погашение точно в срок. То есть из расчета исключаются займы, если по договору была пролонгация или просрочка.

Комиссия за перевод денежных средств рассчитывается в процентах и отражает уровень комиссионного вознаграждения контрагента за перевод денежных средств от микрофинансовой организации клиенту.

Средние постоянные издержки на один займ (AFC_i) включая расходы на офис, бухгалтерию, отдел информационных технологий, колл-центр и т. д.,

является переменной величиной. Однако для упрощения расчетов в рамках настоящей модели выражается через константное значение.

Средние расходы на взыскание на один займ ($ADCC_i$) являются переменной величиной, которая складывается из комиссий агентам, фонда оплаты труда коллекторов или расходов по договорам цессии, судебные расходы и т. д. Данный показатель зависит от конкретных бизнес-процессов по взысканию в МФО. Для упрощения расчетов в рамках настоящей модели выражается также через константное значение.

Процентные расходы (фондирование) рассчитываются по формуле:

$$IC_i = \frac{(AD_i \cdot AT_i)}{365} \cdot FR_i, \quad (4.31)$$

где FR_i – ставка фондирования или инвестирования, %. Данный показатель изменяется во времени и зависит от стоимости привлечения денег в экономике, определяемой Центральным Банком. Однако для упрощения расчетов в рамках настоящей модели этот внешний фактор выражается через константное значение.

Таким образом, расчет юнит-экономики первичного займа предполагает множество последовательных расчетов по рассмотренным формулам в идеале на горизонте минимум 24 месяц назад (ретроспективные расчеты) и 24 месяца вперед (прогнозные расчеты на будущий период). Прогнозные расчеты на будущий период будут показывать, какой финансовый результат ожидать МФО в среднем на одного первичного клиента, на основе значений основных бизнес-метрик в настоящем периоде. Кредитные потери – важнейший параметр рассмотренной модели юнит-экономики.

В отношении повторного займа юнит-экономика МФО корректируется на показатель возврата клиентов в повторную заявку (англ. Retention Rate, RR), который представляет собой прогнозное значение коэффициента удержания клиентов на следующие 24 месяца или альтернативный период. Данный коэффициент может быть рассчитан по следующей формуле:

$$RR = \frac{LN_{total}}{LN_{new}} \cdot 100\%, \quad (4.32)$$

где LN_{total} – общее количества займов за весь период, шт.;

LN_{new} – количество заявок за весь период от клиентов, впервые взявших займ в этом месяце, шт.

Показатель RR достаточно сложный для расчета на практике, т. к. требует созревания, значительного времени и накопленной статистики. Следовательно, для упрощения расчетов в рамках разработанной модели данный коэффициент принимает константные значения.

Для проведения расчетов в рамках разработанной модели юнит-экономики МФО необходимо учитывать следующие особенности рынка микрофинансирования:

а) Повышенные ставки процента по микрозаймам. В рамках проведения расчетов данный показатель будет взят на уровне максимального значения (с 01.07.2019 – 1%, с 01.07.2023 – 0,8% в день).

б) Дополнительные продукты/услуги, реализуемые МФО. В рамках разработанной модели в качестве дополнительных продуктов/услуг определены услуги страхования жизни, услуги страхования карты, подписки на услуги партнеров (юридические консультации, медицинская помощь и т. д.) и другие.

в) Ввиду того, что единицей (юнитом) в разработанной модели выступает клиент, а основную прибыль МФО получает от повторных клиентов, при проведении расчетов необходимо учитывать несколько периодов жизни клиента в МФО. В рамках примера проведения расчетов данный период составляет 24 месяца. Возможная пролонгация договора займа ведет к получению дополнительной прибыли с одного юнита (заемщика).

С целью упрощения предоставления результатов для приведения примера расчетов по разработанной математической модели приняты следующие допущения:

а) расчет проводится на 24 месяца – 1 год до (отчетный период) и 1 год после (плановый период) отчетной даты;

б) в течение отчетного периода предполагается неизменность (отсутствие трендов) параметров и их колебания около среднего уровня;

в) в течение планового периода предполагается неизменность параметров во времени, берутся средние значения показателей за отчетный год;

г) в расчет не включаются резервы МФО.

Проиллюстрируем расчеты модели юнит экономики на гипотетической МФО с параметрами, близкими к рыночным. Расчет юнит-экономики МФО производится за период с апреля 2023 года по март 2025 года.

В таблице ПГ.1 Приложения Г представлены исходные данные для проведения расчетов юнит-экономики первичного займа МФО. Показатели, вычисленные по формулам модели и необходимые для дальнейшего расчета юнит-экономики первичного займа МФО представлены в таблице ПГ.2 Приложения Г. В таблице ПГ.3 Приложения Г представлен расчет юнит-экономики первичного займа МФО.

Необходимо отметить, что в рамках апробации разработанной модели было принято во внимание два варианта дневной процентной ставки по займу в соответствии с законодательными нормативами:

- 1 вариант – апрель-июнь 2023 года – 1%;
- 2 вариант – июль 2023 – март 2025 – 0,8%.

В рамках первого варианта получение отрицательного значения экономики в третий период (июнь 2023 года) был получен за счет увеличения кредитных потерь при одновременном увеличении количества выдач и средней суммы займа.

Существенное увеличение убыточности первого займа (отрицательное значение юнит-экономики) после изменения дневной процентной ставки на 0,2 п. п. до 0,8% с 1 июля 2023 года, которое привело к снижению средней комиссии, несмотря на увеличение средней суммы займа.

На дальнейшее снижение убытка по первому займу при неизменном уровне кредитных потерь оказало увеличение средней суммы займа и реального среднего срока займа с учетом пролонгаций и досрочных погашений, что позволило скомпенсировать снижение дневной процентной ставки. Также положительное влияние на изменение показателя юнит-экономики оказало увеличение вклада пролонгации.

В целом проведенные расчеты показали, что на протяжении исследуемого периода (за исключением первых двух месяцев) юнит-экономика первичного займа МФО принимает отрицательные значения. Таким образом, клиенты, обратившиеся за займом только один раз будут приносить отрицательное значение юнит-экономики МФО. Положительное значение показатель юнит-экономики МФО может быть получен только при условии наличия повторного займа. Большая часть эффекта снижения прибыльности заложена в уровне чистых кредитных потерь, т. к. при повышении кредитных потерь отрицательное значение юнит-экономики увеличивается.

Повышение продаж кросс-продуктов ведет к получению дополнительной прибыли, но при этом не выводит юнит-экономику первичного займа в положительное значение. При этом снижение расходов на привлечение новых клиентов, скоринг заявок, постоянные издержки и расходы на взыскание также не приводят к положительному значению юнит-экономики первичного займа. Таким образом, можно заключить необходимость разработки модели по управлению кредитными рисками МФО для снижения кредитных потерь, связанных с невыполнением обязательств заемщика по кредитному договору первичного займа в течение более 10 дней.

Кроме того, существует вариант осознанного увеличения уровня кредитных потерь за счет выбора нового рискованного уровня отсека клиентов по скоринговой модели первичных клиентов. При этом в рамках изменения бизнес-модели МФО будет зафиксировано повышение конверсии и снижение стоимости привлечения клиентов при одновременном увеличении прибыли от продаж дополнительных услуг и продлений, что приведет к повышению финансового результата и соответственно показателя юнит-экономики МФО.

В таблице ПД.1 Приложения Д представлены исходные данные для проведения расчетов юнит-экономики повторного займа МФО с использованием следующих допущений:

а) средняя стоимость принятия решения по одной заявке составляет 15 руб., что ниже, чем при первом займе (50 руб.) и связано с отсутствием необходимости

идентификации заемщика, а также меньшим количеством источников для получения данных для оценки кредитных рисков, что приводит к снижению стоимости оценки повторного заемщика;

б) средняя стоимость привлечения повторных заемщиков существенно ниже, чем по первичным клиентам, и в итоге стремится к нулю, что связано с отсутствием затрат на CPA (англ. Cost Per Action, цена за действие).

Показатель возврата клиентов (RR) для простоты расчетов принимает константные значения и составляет 250%.

Показатели, вычисленные по формулам модели и необходимые для дальнейшего расчета юнит-экономики повторного займа МФО представлены в таблице ПД.2 Приложения Д. Из представленных расчетных данных видно следующее отличие показателей повторного займа от показателей первого займа:

- а) средний чек повторных клиентов выше за счет повышения лимитов;
- б) увеличение среднего срока по договору;
- в) повышение уровня досрочного погашения;
- г) повышение уровня проникновения дополнительных услуг за счет повышения лояльности клиентов;
- д) снижение показателя чистых кредитных потерь по показателю NPL10 net до 12–14%;
- е) увеличение доли одобренных заявок и выдач.

В таблице ПД.3 Приложения Д представлен расчет юнит-экономики повторного займа МФО.

Получение положительных значений юнит-экономики повторного займа определяется увеличением средней выручки по повторным займам, которое обусловлено ростом среднего чека по займу ввиду увеличения лимитов для повторных клиентов на основании их кредитного поведения. Данный рост приводит к увеличению как средней комиссии по повторным займам, так и вклада пролонгации также за счет увеличения реального среднего срока займа с учетом пролонгаций и досрочных погашений.

Следует отметить, что повторные займы в МФО сопровождаются снижением средних затрат в сравнении с первичным займом за счет сокращения затрат на привлечение клиентов и скоринг одной выдачи.

В целом проведенные расчеты показали, что в отличие от юнит-экономики первичного займа юнит-экономика повторного займа в МФО принимает положительные значения, несмотря на увеличение чистых кредитных потерь. При этом за счет увеличения финансового результата МФО по повторным займам возрастает ценность клиента, т. е. общая юнит-экономика МФО.

Таким образом, в рамках настоящего параграфа была представлена математическая модель юнит-экономики МФО. Данная модель строится на базе клиентской модели юнит-экономики, основой которой выступают метрики взаимодействия бизнеса микрофинансовой организации с заемщиком. Данная модель позволяет провести оценку юнит-экономики как первичного, так и повторного займа, а также ценность клиента – общую юнит-экономику МФО на основе специфических показателей деятельности МФО в рамках системы управления рисками при кредитовании заемщиков. Модель юнит-экономики позволяет реализовать подход к моделированию кредитного риска и параметров бизнес-метрик по всему кредитному портфелю МФО.

4.3 Управление кредитными рисками микрофинансовых организаций в условиях изменчивости коэффициентов скоринговых моделей во времени

Как было определено в главе 1 при рассмотрении основных методологических проблем, возникающих при оценке и управлении рисками в МФО, в отечественной и зарубежной научной среде неоднократно рассматривался вопрос учета динамических факторов при решении задач кредитного скоринга. При этом отсутствует соответствующий математический подход к учету данного фактора, который мог получить практическое применение в кредитных организациях, в частности МФО. В данном разделе на основе работ

автора [61, 62, 104, 213, 314] представлен разработанный подход управления кредитными рисками МФО на основе модели оценки кредитного риска с изменяющимися во времени параметрами вместо традиционного подхода перепостроения скоринговой модели по новым данным по мере накопления статистики и изменений трендов на рынке.

Управление кредитными рисками в микрофинансовых организациях предполагает качественную классификацию заемщиков по риску возникновения дефолта, которая позволяет снизить существенные потери от операций по кредитованию. Качество классификации обычно тестируется на обучающей и тестовой выборке. Полученные на базе классических методов статистического моделирования или алгоритмов машинного обучения статистические наборы данных делятся на обучающие и тестовые. Сначала используемый метод тестируется на обучающей выборке путем соответствующей подгонки параметров, а затем на тестовой выборке производится последовательная проверка точности разработанной модели.

Процесс построения моделей оценки кредитного риска в МФО может происходить по двум вариантам:

- а) построение первичной модели по данным «плохих» и «хороших» клиентов без учета отклоненных заявок;
- б) построение конечной модели с учетом анализа отклоненных заявок.

При первом варианте производится использование метода стратегии «ручейка» или «чистой выборки» (от англ. open gate).

При втором варианте для определения зависимой переменной заемщиков делят на три группы: «плохие», «хорошие» и «неопределенные». Для мошенников, банкротов и безнадёжных кредитов критерий определения «плохого» заемщика однозначен. В отношении остальных заемщиков критерием определения «плохого клиента» является, как правило, количество дней просрочки платежа по кредитам. К группе «неопределенных» клиентов могут относить клиентов с недостаточной кредитной историей, имеющие небольшую допустимую просрочку платежа и другие. При построении моделей оценки

кредитных рисков используются только клиенты, определенные как «плохие» и «хорошие». Доля «неопределенных» клиентов учитывается при построении окончательного прогноза зависимой переменной в рамках модели оценки кредитного риска. Также при построении модели оценки кредитного риска часто в отдельную категорию выделяют «отклоненных» клиентов, т. е. заемщиков, которым отказали в выдаче займа. Учет «неопределенных» и «отклоненных» заемщиков позволяет учесть в обучающей выборке данных для построения модели пропорции генеральной совокупности заемщиков. Данный аспект и анализ «отклоненных» заявок глубоко не рассматриваются в настоящей работе. Подробнее об анализе отклоненных заявок, можно ознакомиться, например в работе [358].

Главный недостаток такого подхода – тестирование модели на заемщиках, отфильтрованных существующей в кредитном учреждении СПР, что не позволяет учитывать «серую зону» – клиентов, которые либо никогда не обращались за кредитом, либо были проигнорированы скоринговой системой, что ведет к сокращению числа заемщиков для моделирования [111].

Для получения более объективных результатов необходимо тестировать скоринговую модель также на заемщиках, не прошедших «фильтр» первоначальной скоринговой оценки. Кроме того, параметры любой модели не являются статичными и изменяются под воздействием развития экономической среды – некоторые параметры увеличивают степень своего влияния с течением времени, а другие – становятся незначительными.

Таким образом, если параметры истинной модели изменяются во времени, такой подход подразумевает применение устаревших значений параметров, что приводит к существенной неэффективности применяемой модели.

В системе принятия решений по стратегии «ручейка», как правило, применяются только базовые проверки, предусмотренные законодательством РФ и политикой МФО по конкретному кредитному продукту. Такая стратегия традиционно используется для сбора «чистых» данных по потоку входящих заявок клиентов для оценки качества кредитной политики и построения внутри

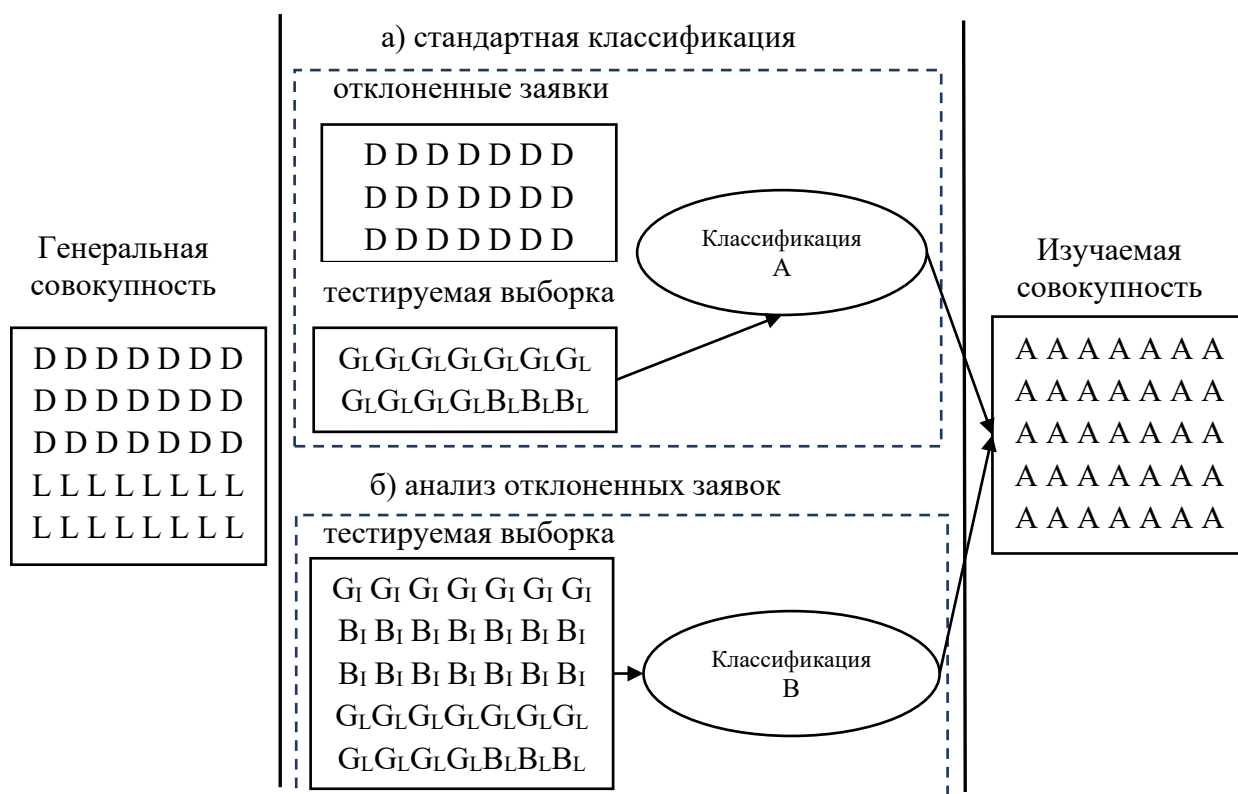
кредитной организации собственной скоринговой модели. Основным недостатком данного метода является высокая затратность финансовых и временных вопросов, а также необходимости согласования действий с маркетингом и др.

Поэтому чаще все МФО используют модели, построенные без использования стратегии «чистой выборки». При этом суть проблемы такого подхода заключается в «серой зоне», т. к. при построении модели отсутствует возможность получения данных по всему потоку клиентов, чем больше становится кредитный портфель МФО со временем, тем больше изменений и правок вносится в СПР. Это приводит к снижению точности моделей оценки кредитных рисков даже на большой по объему выборке и необходимости перестраивать и актуализировать такие модели либо внедрять новые, что не всегда возможно в условиях ограниченности финансовых и временных ресурсов МФО.

Для решения данной проблемы используют метод анализа отклоненных заявок. Данный метод представляет собой процесс, в ходе которого анализируется выборка первоначально отклоненных заявок с целью классификации заемщиков из генеральной совокупности [358].

На рисунке 4.3 представлено сравнение стандартной системы классификации кредитных заявок и системы анализа отклоненных заявок.

В стандартной системе классификации все клиенты, подавшие заявку на заем, в зависимости от параметров оценки подразделяются на потенциально «хороших» или потенциально «плохих», заявки которых сразу отклоняются. Таким образом, тестируемая выборка включает только клиентов, которым заем выдан, а дальнейшая классификация на «хороших» и «плохих» производится на основе их платежного поведения в рамках этой МФО.



примечание – D – отклоненные заявки, L – выданные займы, G_L – «хорошие» выданные займы, B_L – «плохие» выданные займы, G_I – «хорошие» отклоненные заявки, B_I – «плохие» отклоненные заявки, A – заявки на заем

Рисунок 4.3 – Сравнение стандартной системы классификации кредитных заявок и системы анализа отклоненных заявок

Источник: составлено автором по [356]

Отличие метода анализа отклоненных заявок заключается в том, что отклоненные на первоначальном этапе стандартной системы заявки клиентов включаются в тестируемую выборку, такие заявки также делятся на «хорошие» и «плохие». В данном случае используют различные способы получения недостающих данных по отклоненным клиентам:

а) метод «ручной» оценки описан в статье [318] и представляет собой метод экспертной оценки на основе опыта кредитных экспертов. Например, экспертам известно на основе исторического опыта, что 50% отклоненных заявок становятся «плохими» через определенный временной интервал, остальные 50% определяются как «хорошие». Основным недостатком данного метода является его субъективность;

б) метод получения исторических данных из БКИ является одним из наиболее распространенных в практике анализа отклоненных заявок, который был рассмотрен в следующих работах [174, 358]. В данном случае МФО для классификации заемщиков по отклоненным заявкам используют данные о платежном поведении таких заемщиков в других кредитных организациях. При этом дефолт по другим кредитным продуктам принимается эквивалентным дефолту по займам МФО. Основные недостатки данного метода заключаются в высоких финансовых и временных затратах на получение и подготовку данных, невозможности определения качества данных на этапе их получения и подготовки к анализу, дефолтность заемщика по банковским кредитным продуктам может на практике не соответствовать дефолтности по микрозаймам;

в) метод повторного взвешивания рассмотрен в статьях [172, 136] и представляет собой взвешивание принятых к рассмотрению (не отклоненных) заявок и их масштабирование до генеральной совокупности. В рамках данного метода производится оценка соотношения между количеством одобренных заявок и количеством отклоненных заявок, при этом количество весов должно быть равно общему количеству заявок на займы. По мнению автора работы [172] повторное взвешивание не повышает классификационную способность модели по следующим причинам:

- во-первых, низкие способности модели, параметризированной исключительно по одобренным заявкам;

- во-вторых, нереалистичность предположения: $P(b | x, r) = P(b | x, a)$, где $P(b | x, r)$ – вероятность того, что отклоненная заявка является «плохой», $P(b | x, a)$ – вероятность того, что принятая заявка является «плохой».

Кроме того, данный метод требует достаточный объем исторических данных и может не сработать на небольших выборках с высоким уровнем отклонений;

г) метод экстраполяции рассмотрен в работе [169] и представляет собой перенос оценок по принятым заявкам на отклоненные заявки. В работе [358]

данный метод называется «методом жесткого отсечения». По мнению [162], основным недостатком метода экстраполяции является предположение о равнозначности поведения совокупности клиентов по отклоненным заявкам и поведения клиентов по принятым заявкам, и соответственно тождественности риска дефолта по таким заемщикам;

д) метод двухшаговой коррекции Хекмана [263] представляет собой метод смещения с учетом скрытых переменных. В рамках кредитного скоринга данный метод впервые был рассмотрен в работе [315]. Модель состоит из двух механизмов: механизм выборки z_i и механизм результата y_i , которые моделируются посредством скрытых переменных z_i^* и y_i^* для наблюдения i ($i = 1, 2, \dots, N$) следующим образом:

$$\begin{aligned} z_i &= \begin{cases} 1, \text{если заявка принята,} & z_i^* \geq 0 \\ 0, \text{если заявка отклонена,} & z_i^* < 0 \end{cases} \\ y_i &= \begin{cases} 1, \text{если заявка "хорошая",} & y_i^* \geq 0 \\ 0, \text{если заявка "плохая",} & y_i^* < 0 \end{cases} \end{aligned} \quad (4.33)$$

Скрытые переменные z_i^* и y_i^* находятся в зависимости от независимых переменных w_i с коэффициентами γ и β и случайных ошибок u_i и e_i с двумерным нормальным распределением:

$$\begin{cases} z_i^* = w_i \cdot \gamma + u_i \\ y_i^* = w_i \cdot \beta + e_i \end{cases}, \begin{bmatrix} u_i \\ e_i \end{bmatrix} \sim N(\mu, \Sigma) \quad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \rho \neq 0 \quad (4.34)$$

По мнению многих исследователей [171, 194] использование данного метода на практике неработоспособно, т. к. при невыполнении предположения о нормальности распределения случайных ошибок классификационная способность модели будет ненадежной.

Таким образом, анализ отклоненных заявок с использованием любого из рассмотренных методов не всегда работоспособен на практике оценки кредитных рисков МФО, а также еще больше усложняет процесс как моделирования, так и внедрения моделей. Более того, многие эксперты в области кредитного скоринга отмечают, что анализ отклоненных заявок клиентов требует больших ресурсов и не всегда приводит к улучшению качества конечной скоринговой модели [45].

Таким образом, необходима выработка альтернативного метода для решения проблемы «серой зоны» в рамках управления кредитными рисками МФО.

Предпосылкой разработанной автором модели оценки кредитных рисков с учетом изменчивости коэффициентов во времени является наличие устойчивых трендов во времени в изменении параметров моделей оценки кредитных рисков, которые строятся последовательно за разные периоды времени. Докажем данную теоретическую предпосылку на эмпирических данных реальной МФО на российском рынке.

Для проведения расчетов в исследование была отобрана выборка о выдачах займов и их параметрах по PDL продукту первичных клиентов по месяцам за два года в период 2020–2021 гг. Для оценки вероятности дефолта заемщика была выбрана целевая переменная NPL30. Исходная выборка данных была разделена на 24 подвыборки по календарному месяцу выдачи займов.

На рисунке 4.4 представлена динамика уровня изменения во времени целевой переменной NPL30 и объема выборки – количества выданных займов первичным клиентам по продукту PDL, а также динамика их полиномиальных аппроксимаций. Виден фактически линейный рост объема выдачи МФО за рассматриваемый период, который отражает ее увеличение доли на рынке с за рассматриваемый период. С ежемесячным ростом количества выданных займов нелинейно менялся скурорень дефолтности, связанный с изменениями системы принятия решений в МФО и факторами на рынке, влияющих на кредитные риски.

Задача управления кредитными рисками для МФО в рассматриваемом примере заключается в постоянном перестроении моделей оценок кредитного риска с изменением объема выдачи. С ростом ежемесячного объема выдачи меняется конверсия в одобрение, меняется качество заявок входящих клиентов по уровню кредитного риска. Это все требует перестроения скоринговых моделей с определенной периодичностью, в среднем раз в 6–12 месяцев. При этом выборка данных для моделирования будет формироваться кумулятивным итогом с учетом

данных всех предыдущих периодов, т. е. по мере роста компании будет существенно увеличиваться историческая выборка для моделирования.

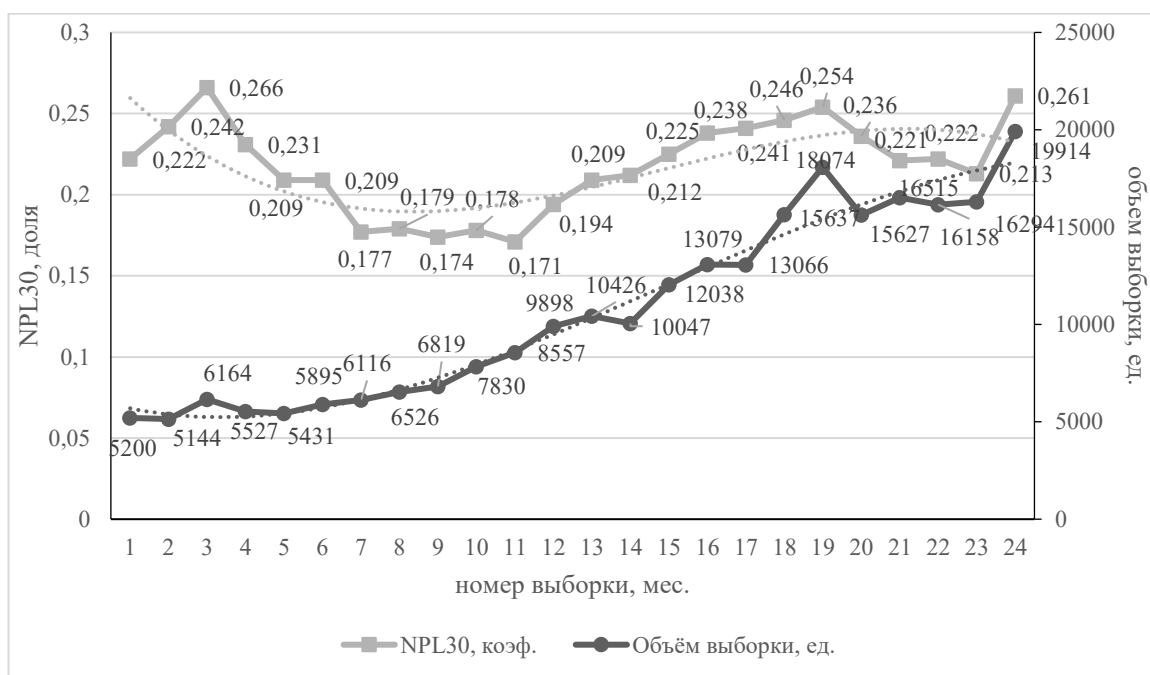


Рисунок 4.4 – Динамика уровня изменения во времени параметра NPL30 и объема подвыборки

Источник: расчеты автора

Однако на практике в данной ситуации увеличение объема выборки не приводит к улучшению качества моделей оценки кредитных рисков. Кроме того, чем больше изменений и новых скоринговых моделей вводит МФО и формирует затем после результатов выдач по ним выборку для построения новой модели, тем хуже и хуже становится качество модели оценки кредитного риска. Как было отмечено выше, причина данного эффекта – наличие «серой зоны» или отказных заявок, информация по которым не учитывается в моделировании.

Для оценки кредитных рисков будем использовать математическую модель логистической регрессии, подробно разобрannую в параграфе 2 главы 3. Предположим, что каждый месяц значения коэффициентов логистической регрессии формируются фундаментальными факторами кредитного риска. В

качестве таких для иллюстрации мы возьмем пять предикторов, рассчитанных на основе кредитных историй заемщиков в БКИ:

1) X_1 – доля долга относительно выданной суммы по микрокредитам;
2) X_2 – количество не проблемных займов с типом «кредитная карта»;
3) X_3 – количество активных кредитов с просрочкой NPL30 за последний год;

4) X_4 – доля банковских кредитов с просрочкой NPL30;

5) X_5 – количество запросов в бюро кредитных историй старше года.

Данный набор предикторов является условным, спецификация данной модели не претендует на полную скоринговую модель для оценки кредитных рисков и служит только для иллюстрации эмпирического обоснования предлагаемого автором метода. Отобранные независимые имеют сильную корреляцию с целевой переменной и не строго мультиколлинеарны между собой.

Также предположим, что на значения коэффициентов логистической регрессии влияет сезонность и долгосрочные тренды, формирующиеся на финансовом рынке.

На каждой подвыборке данных построим 24 логистических регрессии (см. таблицу 4.6). Строки в таблице 4.6 отражают результат моделирования за каждый месяц, последняя строка – результат моделирования по объединенной выборке.

При построении регрессий на данных каждого месяца была получена модель со всеми статистически значимыми коэффициентами для всех периодов (p -значение $< 0,0001$). Это говорит о статистической устойчивости влияния данных факторов во времени. Однако оценки коэффициентов для всех переменных во времени не являются случайными колебаниями от истинных значений.

На рисунке 4.5 а–д представлена динамика полученных регрессионных коэффициентов для каждой независимой переменной за два года.

Таблица 4.6 – Результаты оценки параметров логистической регрессии во времени

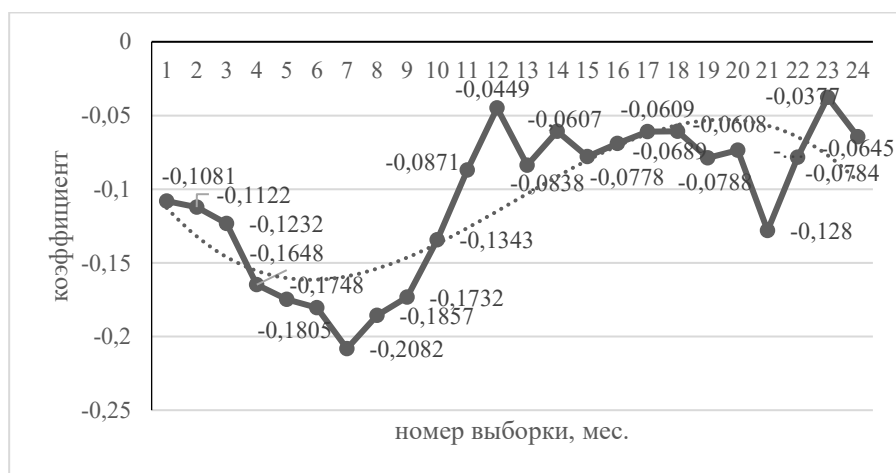
X_1		X_2		X_3		X_4		X_5	
Коэфф.	Р-значение	Коэфф.	Р-значение	Коэфф.	Р-значение	Коэфф.	Р-значение	Коэфф.	Р-значение
– 0,9623	0,0000	– 0,1081	0,0000	0,1063	0,0000	– 1,018	0,0000	– 0,0098	0,0000
– 1,1687	0,0000	– 0,1122	0,0000	0,1074	0,0000	– 0,9667	0,0000	– 0,0063	0,0000
– 0,9174	0,0000	– 0,1232	0,0000	0,1024	0,0000	– 0,7968	0,0000	– 0,0058	0,0000
– 1,3224	0,0000	– 0,1648	0,0000	0,0904	0,0000	– 0,5831	0,0000	– 0,0045	0,0000
– 1,7181	0,0000	– 0,1748	0,0000	0,1201	0,0000	– 1,0115	0,0000	– 0,0049	0,0000
– 1,5061	0,0000	– 0,1805	0,0000	0,1279	0,0000	– 0,9968	0,0000	– 0,0053	0,0000
– 1,6611	0,0000	– 0,2082	0,0000	0,1061	0,0000	– 1,2799	0,0000	– 0,004	0,0000
– 1,4246	0,0000	– 0,1857	0,0000	0,0536	0,0001	– 1,1961	0,0000	– 0,0078	0,0000
– 1,4932	0,0000	– 0,1732	0,0000	0,0842	0,0000	– 1,0794	0,0000	– 0,0105	0,0000
– 1,461	0,0000	– 0,1343	0,0000	0,0865	0,0000	– 0,9248	0,0000	– 0,0099	0,0000
– 1,2703	0,0000	– 0,0871	0,0000	0,1386	0,0000	– 0,9725	0,0000	– 0,0142	0,0000
– 1,0773	0,0000	– 0,0449	0,0000	0,106	0,0000	– 0,4246	0,0000	– 0,0147	0,0000
– 0,8263	0,0000	– 0,0838	0,0000	0,117	0,0000	– 0,5261	0,0000	– 0,0132	0,0000
– 0,9294	0,0000	– 0,0607	0,0000	0,1073	0,0000	– 0,3996	0,0001	– 0,0128	0,0000
– 0,5754	0,0000	– 0,0778	0,0000	0,1067	0,0000	– 0,4882	0,0000	– 0,0117	0,0000
– 0,4437	0,0000	– 0,0689	0,0000	0,0932	0,0000	– 0,3602	0,0000	– 0,0114	0,0000
– 0,7548	0,0000	– 0,0609	0,0000	0,088	0,0000	– 0,3366	0,0000	– 0,0103	0,0000
– 0,8112	0,0000	– 0,0608	0,0000	0,0999	0,0000	– 0,3986	0,0000	– 0,0104	0,0000
– 0,6092	0,0000	– 0,0788	0,0000	0,0699	0,0000	– 0,2808	0,0000	– 0,0088	0,0000
– 0,5998	0,0000	– 0,0735	0,0000	0,0672	0,0000	– 0,357	0,0000	– 0,0097	0,0000
– 0,8494	0,0000	– 0,128	0,0000	0,0773	0,0000	– 0,6885	0,0000	– 0,0086	0,0000
– 0,7515	0,0000	– 0,0784	0,0000	0,058	0,0000	– 0,5111	0,0000	– 0,0094	0,0000
– 0,8661	0,0000	– 0,0377	0,0000	0,0602	0,0000	– 0,3993	0,0000	– 0,011	0,0000
– 0,6213	0,0000	– 0,0645	0,0000	0,0897	0,0000	– 0,2318	0,0006	– 0,008	0,0000
– 1,0111	0,0000	– 0,0993	0,0000	0,0859	0,0000	– 0,5609	0,0000	– 0,0095	0,0000

Источник: составлено автором

По представленным графикам видно, что все коэффициенты не меняли своего знака, что говорит об их стабильности во времени в направлении влияния на уровень кредитного риска. Однако видно цикличное изменение коэффициентов при тестируемых переменных, которая доказывает временную зависимость истинных параметров в рамках применяемой модели.



а) доля долга относительно выданной суммы по микрокредитам, X_1



б) количество не проблемных займов с типом «кредитная карта», X_2



в) количество активных кредитов с просрочкой NPL30 за последний год, X_3

Рисунок 4.5 – Динамика изменения коэффициентов для независимых переменных

Источник: составлено автором

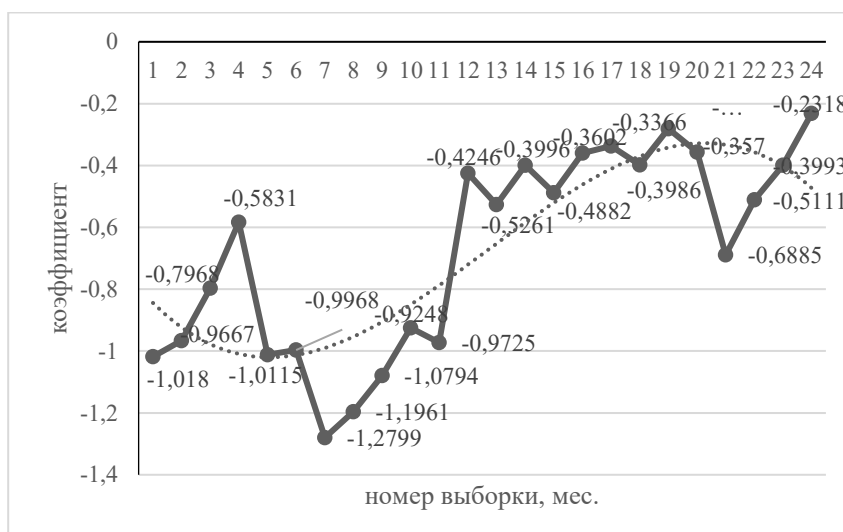
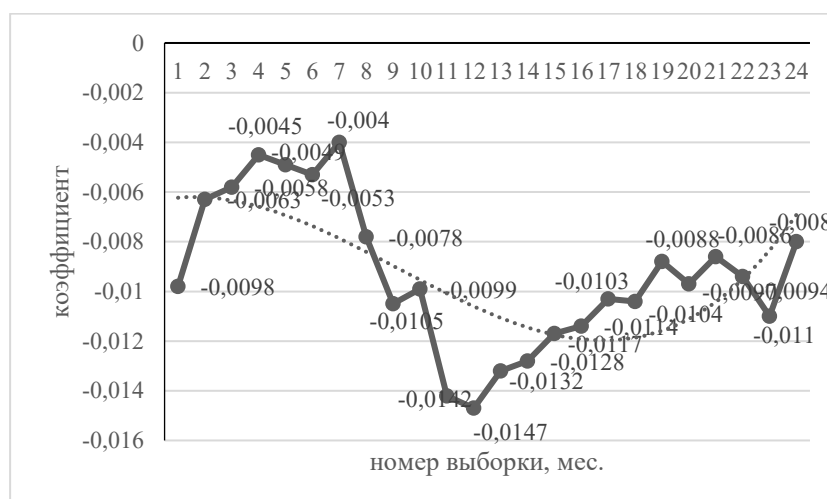
г) доля банковских кредитов с просрочкой NPL30, X_4 д) количество запросов в бюро кредитных историй старше года, X_5

Рисунок 4.5 – Динамика изменения коэффициентов для независимых переменных (продолжение)

Источник: составлено автором

Полученные временные ряды коэффициентов должны быть прогнозируемыми, т. е. они не должны представлять из себя случайное блуждание или белый шум. На рисунке 4.6 а–д представлены графики автокорреляционной (АКФ) и частной автокорреляционной функции (ЧАКФ) для динамических коэффициентов.

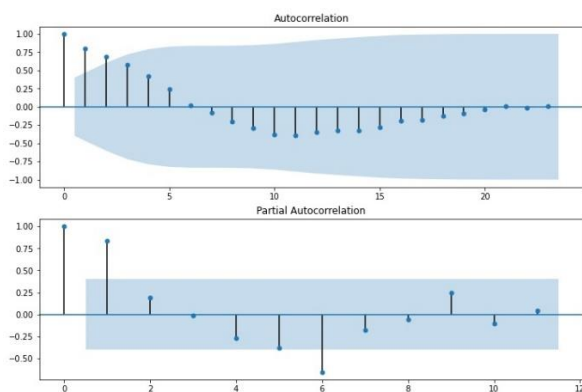
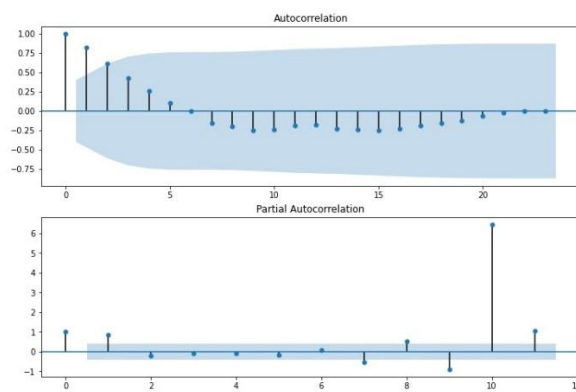
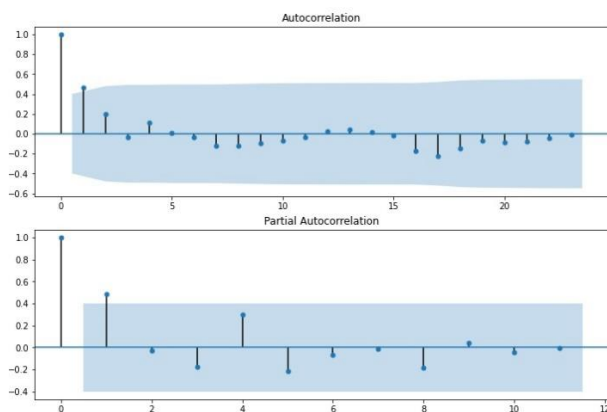
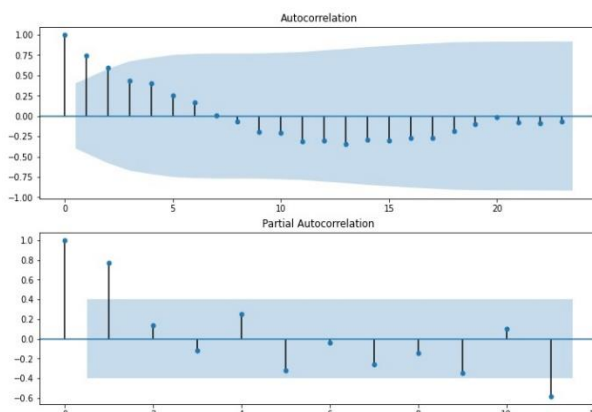
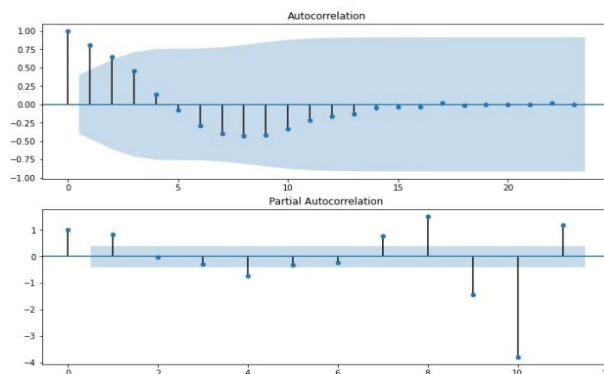
а) Графики АКФ и ЧАКФ для b_1 б) Графики АКФ и ЧАКФ для b_2 в) Графики АКФ и ЧАКФ для b_3 г) Графики АКФ и ЧАКФ для b_4 д) Графики АКФ и ЧАКФ для b_5

Рисунок 4.6 – Графики автокорреляционной (АКФ) и частной автокорреляционной функции (ЧАКФ) для динамических коэффициентов

Источник: составлено автором

Проверку значимости коэффициентов автокорреляции для доказательства отличия временных рядов от белого шума можно провести с помощью тестов Льюнга-Бокса и Бокса-Пирса. Наличие в рядах единичного корня и стационарность проверим с помощью расширенного теста Дики-Фуллера.

Наличие в полученных рядах коэффициентов тренда можно проверить с помощью критерия Манна-Уитни, а поведение дисперсии можно отследить при помощи критерия Сиджела-Тьюки. В таблицах 4.7–4.10 приведены результаты приведенных критериев для каждого ряда коэффициентов.

Таблица 4.7 – Результаты критерия Манна-Уитни для динамических рядов коэффициентов

Ряд	Расчетная статистика	p-значение	Вывод
b_1	1	0,000	тренд присутствует
b_2	15	0,001	тренд присутствует
b_3	100	0,112	тренд отсутствует
b_4	5	0,000	тренд присутствует
b_5	103	0,078	нельзя сделать определенный вывод

Источник: составлено автором

По непараметрическому тесту Манна-Уитни ни для одного коэффициента нельзя принять гипотезу об отсутствии тренда. Схожий результат мы получаем при проведении расширенного теста Дики-Фуллера на стационарность.

Таблица 4.8 – Результаты расширенного теста Дики-Фуллера для динамических рядов коэффициентов

Ряд	Расчетная статистика	p-уровень	Вывод
b_1	– 1,228	0,661	есть единичные корни, ряд не стационарен
b_2	– 1,260	0,647	есть единичные корни, ряд не стационарен
b_3	– 2,783	0,061	есть единичные корни, ряд не стационарен
b_4	– 1,396	0,584	есть единичные корни, ряд не стационарен
b_5	– 4,169	0,001	единичных корней нет, ряд стационарен

Источник: составлено автором

По комплексному анализу всех тестов, предусмотренных в базовом анализе временных рядов [139], можно сделать вывод, что данные временные ряды оценок коэффициентов не являются случайным блужданием или белым шумом, что подтверждает возможность их прогнозирования. Таким образом, эмпирически доказана возможность наличия трендов значений коэффициентов логистической регрессии во времени.

Таблица 4.9 – Результаты критериев Льюнга-Бокса и Бокса-Пирса расширенного теста Дики-Фуллера для динамических рядов коэффициентов

Ряд	Критерий Льюнга-Бокса		Критерий Бокса-Пирса		Вывод
	Расчетная статистика	р-значение	Расчетная статистика	р-значение	
b_1	100,812	0,000	63,098	0,000	коэффициенты автокорреляции значимы
b_2	71,419	0,000	44,903	0,004	коэффициенты автокорреляции значимы
b_3	22,205	0,508	10,572	0,987	коэффициенты автокорреляции незначимы
b_4	93,561	0,000	51,037	0,001	коэффициенты автокорреляции значимы
b_5	69,753	0,000	50,704	0,001	коэффициенты автокорреляции значимы

Источник: составлено автором

Таблица 4.10 – Результаты критерия расширенного теста Сиджела-Тьюки для динамических рядов коэффициентов

Ряд	Расчетная статистика	р-значение	Вывод
b_1	71	0,489	дисперсия постоянна
b_2	59	0,239	дисперсия постоянна
b_3	68	0,421	дисперсия постоянна
b_4	69	0,444	дисперсия постоянна
b_5	41	0,039	дисперсия непостоянна

Источник: составлено автором

Дальнейшая идея автора методики оценки кредитных рисков с изменяющимися во времени параметрами основывается на том, что вместо перестроения моделей каждый месяц на данных новой выборки или проведения «чистой выборки» для сбора статистики и разработки новой модели, получать оценки коэффициентов скоринговой модели с помощью отдельной прогнозной модели [314]. В качестве базовой модели временных рядов может быть использована модель ARIMA (модель Бокса-Дженкинса), часто используемая при прогнозировании в финансовой сфере, подробно рассмотренная, например, в работах [294, 295, 307, 310, 319], которая представляет собой интегрированную модель авторегрессии или расширенной версией модели ARMA для нестационарных временных рядов X_t и имеет следующий вид [189]:

$$\Delta^d X_t = c + \sum_{i=1}^p a_i \Delta^d X_{t-i} + \sum_{j=1}^q b_j \varepsilon_{t-j} + \varepsilon_t, \quad (4.35)$$

где $\Delta^d X_t$ – оператор разности временного ряда X_t порядка d ;

c, a_i, b_j – параметры модели;

d, p, q – порядок параметров модели;

ε_t – стационарный временной ряд случайных остатков.

Основные недостатки модели ARIMA заключаются в следующем [330]:

- во-первых, отсутствие возможности автоматического изменения параметров модели при получении новых данных, что ведет к необходимости постоянной переоценки и модификации модели;
- во-вторых, трудоемкость процесса оценки в связи с необходимостью постоянной калибровки модели.

Альтернативной моделью для оценки параметров тренда коэффициентов модели оценки кредитного риска во времени могут быть тренд-сезонные модели [182, 243], модели экспоненциального сглаживания [152, 244, 298] и другие модели временных рядов.

Также модель оценки кредитных рисков с изменяющимися во времени параметрами, предлагаемая автором, базируется на методах, включающих корректировку полученных коэффициентов посредством фильтра Калмана.

Фильтр Калмана, который представляет собой рекурсивную форму линейного фильтра, позволяющую оценить вектор состояния системы с изменяющимися параметрами при использовании неполных и зашумленных измерений, был разработан и впервые представлен в работе [283].

Целью фильтра Калмана является решение проблемы оценки переменной состояния x_t , которая изменяется во времени в соответствии со следующим линейным уравнением:

$$x_t = F_t \cdot x_{t-1} + B_t \cdot u_t + w_t, \quad (4.36)$$

где x_t – переменная состояния системы в t -й момент времени (последующее состояние);

F_t – матрица перехода состояний (или матрица зависимости текущего состояния системы от предыдущего);

x_{t-1} – переменная состояния системы в $(t - 1)$ -й момент времени (начальное состояние);

B_t – матрица управляющего воздействия;

u_t – управляющий вектор (воздействие);

w_t – вектор случайных ошибок процесса с нормальным распределением (шум процесса).

Уравнение наблюдаемого состояния имеет следующий вид:

$$z_t = H_t \cdot x_t + v_t, \quad (4.37)$$

где z_t – переменная наблюдаемого состояния системы в t -й момент времени;

H_t – матрица наблюдения;

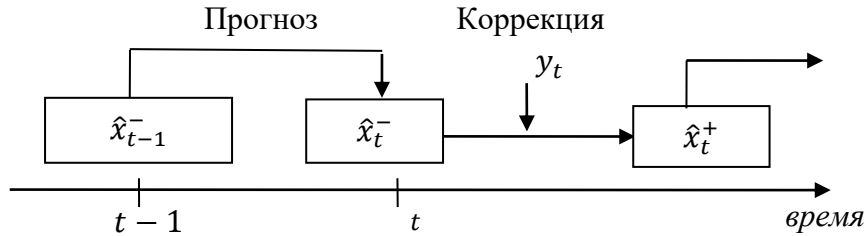
v_t – вектор случайных ошибок наблюдения – гауссов шум с нулевым средним значением (шум наблюдения).

Алгоритм фильтра Калмана состоит из двух основных этапов:

а) Этап прогнозирования, который предназначен для экстраполяции значений переменных состояния в последующий временной период с учетом неточности их измерения.

б) Этап уточнения (корректировки) результата прогнозной экстраполяции с учетом зашумленности данных.

Принцип работы фильтра Калмана представлен на рисунке 4.7.



Примечание – $\hat{x}_{t-1}^-$ – оценка вектора состояния системы в $(t - 1)$ -й (предыдущий) момент времени, $\hat{x}_t^-$ – априорная оценка вектора состояния системы в t -й момент времени, $\hat{x}_t^+$ – апостериорная оценка вектора состояния системы в t -й момент времени.

Рисунок 4.7 – Принцип работы фильтра Калмана

Источник: составлено автором по [283]

На первоначальном этапе производится оценка прогнозного состояния системы и ковариационной матрицы экстраполированного вектора состояния.

Оценка прогнозного состояния системы вычисляется по формуле:

$$\hat{x}_t^- = F_t \cdot \hat{x}_{t-1}^- + B_t \cdot u_t, \quad (4.38)$$

где $\hat{x}_{t-1}^-$ – оценка вектора состояния системы в $(t - 1)$ -й момент времени;

$\hat{x}_t^-$ – априорная оценка вектора состояния системы в t -й момент времени.

Ковариационная матрица экстраполированного вектора состояния вычисляется по формуле:

$$P_t^- = F_t \cdot P_{t-1}^- \cdot F_t^T + Q_t, \quad (4.39)$$

где P_{t-1}^- – ковариационная матрица экстраполированного вектора состояния в t -й (предыдущий) момент времени;

Q_t – ковариационная матрица процесса.

На этапе корректировки вычисляются следующие параметры:

а) ковариационная матрица вектора ошибки:

$$S_t = H_t \cdot P_t^- \cdot H_t^T + R_t, \quad (4.40)$$

где R_t – ковариационная матрица шума измерений;

б) оптимальная матрица коэффициентов усиления:

$$K_t = P_t^- \cdot H_t^T \cdot S_t^{-1}; \quad (4.41)$$

в) коррекция экстраполяции вектора состояния системы:

$$\hat{x}_t^+ = \hat{x}_t^- + K_t \cdot \hat{y}_t, \quad (4.42)$$

где $\hat{y}_t = z_t - H_t \cdot \hat{x}_t^-$ (4.42) – отклонение полученного наблюдения от ожидаемого.

Среди основных преимуществ использования фильтра Калмана в эконометрических моделях прогнозирования состояний различных систем на финансовом рынке можно выделить следующие [99]:

- во-первых, возможность прогнозирования состояния системы в трех проекциях – прошлом, настоящем и будущем в том числе при отсутствии точных данных о ее состоянии и невозможности их получения посредством прямого измерения;
- во-вторых, возможность построения линейных моделей оценки с минимальной вариацией и минимальной среднеквадратичной ошибкой, а также зависимых от времени регрессионных моделей;
- в-третьих, возможность применения для моделирования прогнозных состояний финансового рынка в рамках нестационарных временных рядов;
- в-четвертых, возможность оценки динамического характера переменных в исторической ретроспективе вне зависимости от наличия статистических выбросов в данных, а также при наличии пропусков в выборке (например, пропуски будущих значений параметра).

В отношении использования фильтра Калмана в оценке кредитных рисков можно выделить следующие исследования. В работе [178] проводится анализ преимущества использования фильтра Калмана в оценке качества скоринговых моделей в польских банках. Статья [382] посвящена оценке качества динамической модели скоринговой оценки в ипотечном кредитовании на основе использования фильтра Калмана, которая позволила автору сделать вывод об эффективности данного метода при сравнении измененных скоринговых карт с

базовыми. Что касается применения фильтра Калмана в оценке кредитных рисков микрофинансовых организаций, то таковые исследования в настоящее время отсутствуют как в отечественной, так и в зарубежной научной литературе.

В рамках настоящего диссертационного исследования фильтр Калмана используется с целью получения двух независимых оценок: первая оценка – путем прогнозирования вектора параметров, вторая оценка – с использованием подгонки параметров к накопленной статистике. Модель представлена автором в работе [107].

Рассмотрим более подробно спецификацию модели. Исходная выборка данных представляет собой массив следующего вида:

$$\{y_t, X_t: t = 1, \dots, T\}, \quad (4.43)$$

где $y_t = (y_{1t}, y_{2t}, \dots, y_{nt})^T$ – бинарная целевая переменная, значение которой равно 1 в случае дефолта заемщика и 0 в обратной ситуации,

$X_t = (1, x_{1t}, x_{2t}, \dots, x_{mt})$ – набор данных возможных объясняющих переменных, где $x_{it} = (x_{1it}, x_{2it}, \dots, x_{nit})^T$.

Индекс i означает номер наблюдения, а индекс t – временной период.

Вероятность дефолта заемщика будем моделировать с использованием экспоненциальной формы логистической функции, представленной в формуле (3.3), имеющей следующий вид:

$$p(1|X_t) = \frac{1}{1+e^{-X_t\beta_t}}, \quad (4.44)$$

где β_t – вектор изменяющихся во времени параметров.

Выбор логистической функции обоснован предпосылками и преимуществами, подробно рассмотренными в параграфе 2 главы 3.

Динамику вектора каждого из параметров β_{it} можно смоделировать методами анализа временных рядов: с помощью тренд-сезонной модели, модели экспоненциального сглаживания, либо с помощью некоторой модели класса ARIMA и др. В нашем случае будем использовать модель ARIMA, которая описывается следующей формулой:

$$\hat{\beta}_{it} = \alpha_0 + \alpha_1 \hat{\beta}_{i(t-1)} + \dots + \alpha_p \hat{\beta}_{i(t-p)} + \gamma_1 \hat{\varepsilon}_{i(t-1)} + \dots + \gamma_q \hat{\varepsilon}_{i(t-q)} + \hat{\varepsilon}_{it}, \quad (4.45)$$

где знак « $\wedge$ » соответствует оценкам, полученным из проанализированной выборки.

При этом модель будет более эффективной при использовании в формуле (4.44) прогнозируемых значений параметров (в соответствии с формулой (4.45)) вместо полученных их оценок по уравнению логистической регрессии:

$$p(1|X_t) = \frac{1}{1+e^{-x_t\hat{\beta}_t}}. \quad (4.46)$$

Кроме того, в каждый период времени при помощи используемой модели можно получить две различные оценки вектора истинных параметров β_t : первая оценка – из формулы прогнозирования (4.46), вторая оценка – из формулы подгонки модели (4.44).

Две полученные независимые оценки для получения более точного вектора параметров можно объединить при помощи фильтра Калмана. Функция плотности вероятности для таких параметров будет вычисляться по формуле:

$$f(\beta_t|\hat{\beta}_{t-1}, \dots, \hat{\beta}_{t-p}, X_t) = \frac{f_1(\beta_t|\hat{\beta}_{t-1}, \dots, \hat{\beta}_{t-p})f_2(\beta_t|X_t)}{\int_{\mathbb{R}} f_1(\beta_t|\hat{\beta}_{t-1}, \dots, \hat{\beta}_{t-p})f_2(\beta_t|X_t)d\beta_t}, \quad (4.47)$$

$$\text{где } f_1(\beta_t|\hat{\beta}_{t-1}, \dots, \hat{\beta}_{t-p}) = (2\pi)^{-\frac{m+1}{2}} \det(\Sigma_t)^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta_t - \hat{\beta}_t)^T \Sigma_t^{-1} (\beta_t - \hat{\beta}_t)}; \quad (4.48)$$

$$f_2(\beta_t|X_t) = (2\pi)^{-\frac{m+1}{2}} \det(\Omega_t)^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta_t - \tilde{\beta}_t)^T \Omega_t^{-1} (\beta_t - \tilde{\beta}_t)}. \quad (4.49)$$

Оценка ковариационной матрицы вектора параметров Ω_t из формулы (4.49) производится при помощи информационной матрицы Фишера по формуле:

$$\Omega_t = X_t^T \tilde{W} X_t, \quad (4.50)$$

$$\text{где } \tilde{W} = \text{diag} \left(\frac{e^{\sum_{j=0}^m \tilde{\beta}_{jt} x_{1jt}}}{\left(1 + e^{\sum_{j=0}^m \tilde{\beta}_{jt} x_{1jt}}\right)^2}, \dots, \frac{e^{\sum_{j=0}^m \tilde{\beta}_{jt} x_{njt}}}{\left(1 + e^{\sum_{j=0}^m \tilde{\beta}_{jt} x_{njt}}\right)^2} \right). \quad (4.51)$$

Оценка ковариационной матрицы вектора параметров Σ_t из формулы (4.48) производится при помощи незначительно модифицированной модели DCC–GARCH.

Модель динамической условной корреляции GARCH (DCC–GARCH) была впервые представлена в работе Р. Ингла, основная идея которой заключается в раздельном моделировании волатильностей и корреляций, т.е. «на первом этапе

формируются остатки одномерных моделей GARCH и оценки их условной волатильности, а на втором этапе с помощью ковариаций стандартизированных остатков определяются оценки параметров изменяющейся во времени корреляционной матрицы» [230]. Модели GARCH и их модификации широко используются в финансовой сфере [28, 53, 192].

В рамках настоящего диссертационного исследования незначительно модифицированная модель DCC–GARCH принимает вид:

$$\Sigma_t = D_t R_t D_t, \quad (4.52)$$

где D_t представляет собой матрицу вида:

$$D_t = \begin{bmatrix} \sqrt{h_{1t}} & 0 & \cdots & 0 \\ 0 & \sqrt{h_{2t}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{h_{mt}} \end{bmatrix}, \quad (4.53)$$

где каждый элемент h_{it} определяется дисперсией численно сгенерированной функции плотности вероятности ($pdf_{it}(\beta_{it})$), которая представляет собой среднее значение из N функций плотности вероятности для прогнозируемого значения β_{it} , полученное из формулы (4.45).

В данном случае необходимость численной генерации определяется отсутствием истинных значений β_{it} и получением их оценки с некоторой долей неопределенности. Таким образом, генерирование показателей $\beta_{t-1}, \dots, \beta_{t-p}$ осуществляется на основе их функции плотности вероятности, полученных на предыдущих шагах для достижения соответствия $f(\beta_{it})$, вычисляемого по формуле:

$$f_{it}(\beta_{it}) = \sum_{j=1}^N pdf_{ijt}(\beta_{ijt})/N. \quad (4.54)$$

Далее вводится соответствующая корреляционная матрица R_t для вектора β_t , которая имеет следующий вид:

$$R_t = \begin{bmatrix} 1 & \rho_{12t} & \cdots & \rho_{1nt} \\ \rho_{21t} & 1 & \ddots & \rho_{2nt} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{n1t} & \rho_{n2t} & \cdots & 1 \end{bmatrix}. \quad (4.55)$$

В соответствии с формулами (4.53) и (4.55) каждый элемент Σ_t может быть представлен следующим образом:

$$[\Sigma_t]_{ij} = \sqrt{h_{it}h_{jt}}\rho_{ijt}. \quad (4.56)$$

Модель изменения корреляционной матрицы R_t раскладывается Q_t^{*-1} и Q_t с целью получения абсолютных значений всех записей равных или меньше 1 и выглядит следующим образом:

$$R_t = Q_t^{*-1}Q_tQ_t^{*-1}, \quad (4.57)$$

$$\text{где } Q_t = (1 - a)\bar{\Omega} + a\Omega_{t-1}, \quad (4.58)$$

$$\text{где } \bar{\Omega} = \frac{1}{T}\sum_{t=1}^T \Omega_t; \quad (4.59)$$

$$Q_t^* = \begin{bmatrix} \sqrt{q_{11t}} & 0 & \dots & 0 \\ 0 & \sqrt{q_{22t}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sqrt{q_{mmt}} \end{bmatrix}. \quad (4.60)$$

В рамках настоящего исследования для оценки параметров модели DCC–GARCH, то есть для вычисления скорректированных оценок векторов β_t , на основе полученных функции плотности вероятности предлагается применять метод оценки максимального правдоподобия, при помощи формулы:

$$\hat{\beta}_t = \underset{\beta_t}{\operatorname{argmax}} f(\beta_t | \hat{\beta}_{t-1}, \dots, \hat{\beta}_{t-m}, X_t). \quad (4.61)$$

Скорректированные оценки и полученные функции плотности вероятности в дальнейшем используются в формуле прогнозирования (4.46) с целью получения прогнозов для следующих значений истинных параметров β_{t+1} .

Также следует отметить, что если каждый вектор β_{it} подчиняется нормальному распределению, а элементы Σ_t и Ω_t являются диагональными, то скорректированная плотность вероятности также определяется как нормальная и формула (4.48) может быть записана явно. Для упрощения дальнейших вычислений введем следующие обозначения:

$$E(\beta_{it} | \hat{\beta}_{it-1}, \dots, \hat{\beta}_{it-m}) = \mu_1, E(\beta_{it} | X_t) = \mu_2; \quad (4.62)$$

$$\operatorname{var}(\beta_{it} | \hat{\beta}_{it-1}, \dots, \hat{\beta}_{it-m}) = \sigma_1^2, \operatorname{var}(\beta_{it} | X_t) = \sigma_2^2. \quad (4.63)$$

Таким образом, функции плотности вероятности для β_{it} из формулы (4.47) можно записать следующим образом:

$$f(\beta_{it}) = \frac{\frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(\beta_{it}-\mu_1)^2}{2\sigma_1^2} - \frac{(\beta_{it}-\mu_2)^2}{2\sigma_2^2}}}{\int_{\mathbb{R}} \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(\beta_{it}-\mu_1)^2}{2\sigma_1^2} - \frac{(\beta_{it}-\mu_2)^2}{2\sigma_2^2}} d\beta_{it}}. \quad (4.64)$$

Для определения нормальности распределения функции плотности вероятности необходимо рассмотреть числитель дроби из формулы (4.64), упрощение которой имеет следующий вид:

$$\begin{aligned} \frac{1}{2\pi\delta_1\delta_2} e^{-\frac{(\beta_{it}-\mu_1)^2}{2\delta_1^2} - \frac{(\beta_{it}-\mu_2)^2}{2\delta_2^2}} &= \frac{1}{2\pi\delta_1\delta_2} e^{-\frac{\beta_{it}^2(\delta_1^2+\delta_2^2) - 2\beta_{it}(\mu_1\delta_2^2+\mu_2\delta_1^2) + \mu_1^2\delta_2^2 + \mu_2^2\delta_1^2}{2\delta_1^2\delta_2^2}} = \\ &= \frac{1}{2\pi\delta_1\delta_2} e^{-\frac{\left(\beta_{it} - \frac{\mu_1\delta_2^2 + \mu_2\delta_1^2}{\delta_1^2 + \delta_2^2}\right)^2 + \frac{\mu_1\delta_2^2 + \mu_2\delta_1^2}{\delta_1^2 + \delta_2^2} - \left(\frac{\mu_1\delta_2^2 + \mu_2\delta_1^2}{\delta_1^2 + \delta_2^2}\right)^2}{2\frac{\delta_1\delta_2}{\delta_1^2 + \delta_2^2}}} \end{aligned} \quad (4.65)$$

Дальнейшее интегрирование знаменателя дроби в формуле (4.64) и сокращения $\frac{\mu_1\delta_2^2 + \mu_2\delta_1^2}{\delta_1^2 + \delta_2^2} - \left(\frac{\mu_1\delta_2^2 + \mu_2\delta_1^2}{\delta_1^2 + \delta_2^2}\right)^2$, функция плотности вероятности примет следующий вид:

$$f(\beta_{it}) = \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{2\pi}\sigma_1\sigma_2} e^{-\frac{\left(\beta_{it} - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2}{2\frac{\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}}}. \quad (4.66)$$

Таким образом, скорректированная функция плотности вероятности для β_{it} также является нормальным. При этом среднее значение вычисляется по формуле:

$$\mu = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \quad (4.67)$$

а дисперсия вычисляется по формуле:

$$\sigma^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad (4.68)$$

где $\sigma^2 < \min(\sigma_1^2, \sigma_2^2)$.

Следовательно, использование вышеуказанного метода ведет к получению наилучшей оценки вектора β_{it} . Таким образом, формула (4.46) может быть

заменена более простой в вычислении формулой, учитывая справедливость вышеупомянутых предположений.

Таким образом, предложенная модель оценки кредитных рисков на базе логистической регрессии может привести к получению лучшего финансового результата для МФО по сравнению с традиционным способом подгонки модели к набору данных, доступному для периода $t-1$ и простого применения полученных параметров для периода t , что будет подтверждено в 5 главе на имитационных данных.

Подводя итоги параграфа, можно сделать вывод, что предложенная модель оценки дефолта заемщика с изменяющимися во времени параметрами базируется на методах, включающих корректировку полученных коэффициентов посредством фильтра Калмана с целью получения двух независимых оценок: первая оценка – путем прогнозирования вектора параметров, вторая оценка – с использованием подгонки параметров к накопленной статистике.

Предложенный подход к прогнозированию параметров модели оценки кредитных рисков на базе логистической регрессии может существенно повысить эффективность скоринговых моделей при внедрении системы анализа кредитных рисков в микрофинансовой организации.

Выводы по главе 4

В данной главе разработаны следующие модели управления кредитными рисками в микрофинансовых организациях:

- 1) с целью управления кредитными рисками МФО посредством регулирования лимитной политики представлена математическая модель установки оптимальных лимитов для новых и повторных заемщиков на основе применения результатов скоринговой модели с целью определения дециля риска заемщика с последующим расчетом лимитов и поведенческая модель установки

лимитов для повторных заемщиков на основе их кредитного поведения в конкретной МФО;

2) с целью моделирования кредитного риска и параметров бизнес-метрик по всему кредитному портфелю МФО представлена математическая модель юнит-экономики МФО, базовыми метриками которой выступают критерии взаимодействия МФО с заемщиком для оценки показателей по первичным и повторным займам, а также определения ценности заемщика в рамках системы управления кредитными рисками;

3) с целью повышения эффективности скоринговых моделей при внедрении системы анализа кредитных рисков в МФО предложена модель оценки дефолта заемщика на базе логистической регрессии с изменяющимися во времени параметрами и корректировкой значений коэффициентов при помощи фильтра Калмана для прогнозирования и подгонки параметров.

Глава 5 Апробация моделей оценки и управления кредитными рисками в микрофинансовых организациях

5.1 Оценка кредитных рисков на основе агрегированного признака просрочки по данным кредитной истории заемщика

Использование множества альтернативных данных наряду с традиционной информацией из кредитной истории заемщика позволяет риск-аналитикам МФО проводить более точную оценку кредитных рисков. СПР решений в МФО состоит из поэтапных процедур оценки заемщика на основе множества информации. Однако как было отмечено, процесс тестирования и внедрения новых источников данных достаточно трудоемок и требует системного подхода. В связи с этим, как было исследовано в параграфе 3 главы 1, у многих МФО возникает потребность использования комплексного интегрального показателя для оценки кредитных рисков, который они могли бы применять, не прибегая к сложному процессу построения и внедрения своих скоринговых моделей. Особенно актуальность использования такого показателя для оценки кредитных рисков возникает также при запуске нового продукта в финансовой организации или открытия МФО, когда нет исторических ретроспективных данных о выдачах кредитов, по которым можно построить скоринговую модель для оценки кредитных рисков.

По факту таким интегральным показателем для оценки кредитных рисков является скоринговый балл, полученный по математической модели. На российском рынке информационного обеспечения, как было отмечено во 2 главе, для оценки кредитных рисков практически все провайдеры данных предлагают на коммерческой основе не только исходные или агрегированные показатели, но и скоринговые баллы самостоятельно построенных моделей, в том числе на данных из разных источников. Еще раз отметим особенности применения таких общих скоринговых моделей на практике.

Во-первых, это платность, использование таких данных увеличивает значение бизнес-метрики стоимости принятия одного решения. Во-вторых, это незнание методики разработки модели, особенностей подготовки данных для модели и т. п. Скоринговый балл общей модели МФО получает в режиме «черного ящика». В случае системных и технических сбоев в процессинге данных, качества данных, на которых строится модель, динамика скорингового балла будет иметь структурный сдвиг и повлияет на качество оценки кредитных рисков. В-третьих, скоринговые баллы обобщенных моделей дают усредненную оценку риска по рынку, не учитывая специфику работы и трафика заявок конкретной МФО. Их применение на всем потоке входных заявок в кредитном конвейере может давать не эффективную оценку кредитного риска, часто их используют только для дополнительной оценки кредитного риска на определенном сегменте клиентов.

Наиболее точно отражает кредитный риск заемщика значение балла персонального кредитного рейтинга, предоставляемого БКИ на платной основе. Модели кредитного рейтинга широко распространены в финансовой сфере в работах отечественных и зарубежных ученых [36, 124, 117, 126, 274]. Альтернативой этого рейтинга может быть разработанный и предлагаемый автором в данном разделе исследования показатель качества кредитной истории (ККИ). Под качеством кредитной истории понимается отсутствие негативной информации по взятым ранее кредитам заемщика в БКИ.

Сформулируем основные требования, предъявляемые МФО к интегральному показателю для оценки кредитных рисков, который может выступать альтернативой скоринговым баллам провайдеров данных на рынке:

- а) источник информации, на основе которого он рассчитывается, должен обладать высокой долей покрытия (hit rate) входных заявок на кредит;
- б) относительно простой алгоритм расчета, который аналитик МФО мог бы реализовать самостоятельно;
- в) интерпретируемость логики расчета показателя;

г) сильная корреляция с уровнем кредитного риска заемщика.

Одним из наиболее простых в получении таких источников данных являются кредитные истории заемщики в БКИ. Кредитная история фактически обязательная информация из минимального набора необходимых данных для оценки кредитных рисков, которую используют МФО. К тому же взаимодействие МФО с БКИ регламентируется Федеральным законодательством. Кредитные отчеты, получаемые из одного или нескольких сертифицированных БКИ, удовлетворяют всем сформулированным требованиям для расчета интегрального показателя кредитного риска заемщика.

Поэтому в рамках настоящего диссертационного исследования автором предлагается использовать кредитную историю для построения интегрального показателя оценки кредитного риска, который МФО могла бы рассчитывать самостоятельно. Такой показатель может быть использован для предварительного ранжирования заемщиков по степени риска для корректировки стратегии СПР, может выступать как отдельная переменная в модели оценки кредитных рисков, разрабатываемой МФО самостоятельно, а также использоваться в СПР выходящих на рынок МФО. Наибольшую корреляцию с вероятностью дефолта по будущему займу имеет информация о просрочках по прежним кредитам заемщика. Данный раздел составлен по материалам работ автора [95, 96].

Сформируем перечень исходной информации, необходимой для реализации предлагаемого алгоритма расчетов интегрального показателя оценки кредитных рисков. При оценке кредитного риска один из основных факторов возможности его реализации – это наличие просрочки по ранее обслуживаемым кредитам или микрозаймам у заемщика. Такая информация содержится в БКИ. При этом важно учитывать давность просрочки, сумму и количество дней просрочки, вид просрочки: закрытая или текущая, а также типы кредитов, по которым возникает просроченная задолженность.

Фактор наличия просрочки по данным БКИ во многих СПР МФО на российском рынке является стоп-фактором отказа клиенту, сумма просроченной

задолженности и число дней просрочки могут варьироваться. Однако сам факт наличия просрочки не всегда является признаком 100% дефолта заемщика. Многие клиенты МФО берут займы, чтобы закрыть просрочку, например, по потребительским кредитам или кредитным картам, или даже в другом МФО. К тому же часто информация в БКИ может быть не актуальной и содержать технические ошибки. Например, при продаже по цессии и потом взыскания задолженности коллекторским агентством, запись о наличии просрочки может остаться, т. к. соответствующий статус мог быть передан с ошибкой в кредитное бюро. Или у клиента по данным кредитной истории могут быть достаточно давние просрочки, но в настоящий момент его финансовое положение поменялось. Из-за важности значения конверсии из заявки в займ (AR) из-за высокой стоимости привлечения клиентов на рынке микрофинансирования многие МФО учитывают наличие просрочки при оценке кредитного риска через скоринговые модели, вводя соответствующие переменные в модель. Однако включить в скоринговую модель множество факторов, связанных с просрочкой, одновременно без применения методов снижения размерности является проблематичным из-за наличия сильной мультиколлинеарности. Разрабатывая алгоритм расчета интегрального показателя для оценки кредитных рисков, автор учел данную особенность.

В России по состоянию на 01.12.2024 года функционировало пять БКИ [23]. Однако наиболее популярными, с которыми работают МФО, являются АО «НБКИ», АО «БКИ СБ» (ранее АО «Эквифакс») и АО «ОКБ». Предлагаемый интегральный показатель качества кредитной истории может рассчитывать по данным каждого бюро в отдельности, а также по данным объединенных кредитов по всем бюро, из которых получается информацию МФО.

В рамках предлагаемого алгоритма вычисляется агрегированный показатель ККИ, который рассчитывается как средневзвешенная категоризированная просрочка клиента и представляет собой вариант оценки платежеспособности,

рассчитанной по всем кредитам клиента их КИ. Данный агрегированный показатель определяется в зависимости от следующих параметров:

а) дата актуальности кредита – чем ближе дата оформления кредита к текущей дате заявки, тем больший вес имеет информация о таком кредите;

б) тип кредита – определяется важность определенного типа кредита, в микрофинансовой организации наибольшую ценность имеет информация о просрочке по микрокредитам;

в) платежная дисциплина – учитывается наличие просрочек по каждому из кредитов на основе записей о графике платежей в кредитном отчете.

Исходная информация для проведения расчетов о кредитах заемщика и их платежной дисциплине содержится в кредитном отчете в формате xml-файла. Формат обмена данным кредитных организаций и бюро может меняться. По состоянию на 01.12.2024 года действует единый формат формирования кредитной информации (General Unified Transmission Data Format – GUTDV). Однако структура справочников отчета, описания конкретных полей, способы хранения информации о дисциплине платежей в каждом БКИ организованы по-своему. Также для чтения файлов xml-формата необходимо иметь соответствующий навык или опыт. Поэтому следует остановиться более подробно на технических моментах извлечения необходимых данных для расчета показателя ККИ из xml-файлов. Алгоритм расчета интегрированного показателя ККИ будет рассмотрен на примере кредитного отчета АО «БКИ СБ».

Информация о своевременности платежей по договору займа (кредита) или договору поручительства отображается в кредитном отчете в атрибуте «платежная дисциплина».

Значение атрибута дисциплины платежей представляет собой строку, состоящую из некоторого количества разрядов (символов). При этом крайний левый символ в этой строке всегда соответствует месяцу даты, на которую запрашивался данный кредитный отчет. При смещении на один разряд вправо, мы

получаем значение за предыдущий месяц; при смещении на два разряда вправо – значение, которое было два месяца назад и так далее.

Пример исходных данных XML кредитного отчета АО «БКИ СБ» с информацией о платежной дисциплине по двум кредитным договорам представлен на рисунке 5.1. Для определения параметра «платежная дисциплина» (переменная v) для клиентов микрофинансовой организации необходимыми исходными данными является информация о своевременности платежей по договору микрозайма.

```

...
<attrs>
  <package id="1" name="Основной пакет">
    <section id="23" name="Кредитная часть (платежная дисциплина)">
      <period>
        <attr c_i="1" n="931" v="-24321"/>
        <attr c_i="11" n="931" v="-00000"/>
      </period>
    </section>
  </package>
</attrs>
...

```

Рисунок 5.1 – Пример XML-ответа, содержащий атрибут «платежная дисциплина» по двум договорам

Источник: составлено автором

На рисунке 5.1 параметр c_i – уникальный идентификатор кредитного договора в БКИ; n – идентификатор атрибута; v – значение атрибута, т. е. платежная дисциплина.

В таблице 5.1 приведен пример интерпретации значения атрибута «платежная дисциплина» для договора с $cred\_id=1$ из рисунка 5.1.

Как видно из таблицы 5.1 значение атрибута «платежная дисциплина» категоризируется в порядке увеличения от 1 до N (количество ежемесячных запросов кредитных историй – 1 мес., 2 мес. и т. д.) в зависимости от количества дней просрочки в каждом из запросов кредитных историй.

Таблица 5.1 – Пример интерпретации значения атрибута «платежная дисциплина»

Запрос КИ	Месяц	Значение атрибута	Расшифровка
месяц запроса КИ	май.23	-	нет данных
1 месяц назад	апр.23	2	просрочка 6–29 дней
2 месяца назад	мар.23	4	просрочка 60–89 дней
3 месяца назад	фев.23	3	просрочка 30–59 дней
4 месяца назад	январ.22	2	просрочка 6–29 дней
5 месяцев назад	дек.22	1	просрочка 1–5 дней

Источник: составлено автором

Длина атрибута (количество разрядов) зависит не только от срока жизни самого договора кредита (займа), но и от вида отображения атрибута «платежная дисциплина». Существует два вида (два различных идентификатора) атрибута «платежная дисциплина»:

- $n = 71$ – с месяца даты, на которую запрошена кредитная история, и глубиной 24 месяца назад (всегда 25 разрядов);

- $n = 931$ – с месяца даты, на которую запрошен кредитный отчет, и до месяца открытия договора, т. е. количество разрядов соответствует количеству месяцев между датой открытия кредита и датой запроса кредитной истории.

В нашем примере рассмотрен случай, когда идентификатор атрибута «платежная дисциплина» равен 931.

Состояние кредитного договора определяется по следующим параметрам:

- отсутствие просрочки по договору;
- просрочка по договору с разным временным интервалом (1–5 дней, 6–29 дней, 30–59 дней, 60–89 дней, 90–119 дней, 120–149 дней, 150–179 дней, 180–209 дней, 210–239 дней, 240 дней и более);
- безнадежный долг, который списан с баланса МФО;
- закрытый договор;
- договор продан третьей стороне в связи с переуступкой права требований;

- реструктурированный или рефинансированный договор;
- договор продан коллекторскому агентству;
- расторгнутый договор;
- субъект кредитной истории признан банкротом;
- по субъекту прекращена передача информации в бюро кредитных историй.

В таблице 5.2 приведены возможные состояния кредитного договора, в которых он может находиться в течение месяца, по параметру «платежная дисциплина» на примере кредитного отчета АО «БКИ СБ» и соответствующие им значения категорий просрочки d_{ik} , которые в виде весов далее будут использоваться в расчетах.

Каждый разряд атрибута соответствует наиболее приоритетному состоянию договора за соответствующий месяц. Приоритетность состояний следующая:

- а) наихудшая просрочка по договору, включая безнадежный долг (В, А, 9, 8, 7, 6, 5, 4, 3, 2, 1);
- б) конечное состояние (С, S, R, W, U, T, I);
- в) без просрочки (0);
- г) нет данных (-).

Отсутствие данных или прочерк в определенном разряде проставляется при отсутствии данных по следующим причинам:

- информация по договору за соответствующий месяц не выгружалась источником формирования кредитной истории, информация по договору начала передаваться в БКИ не с даты начала договора;
- договор еще не был заключен в соответствующем месяце, а вариант расчета платежной дисциплины соответствует идентификатору 71;
- информация по договору перестала выгружаться источником формирования кредитной истории.

Таблица 5.2 – Значения разрядов параметра «Платежная дисциплина» по данным АО «БКИ СБ»

Состояние кредитного договора, символ	Категория просрочки (d_{ik})
Нет данных, -	не используется в расчете
Без просрочки, 0	0
Просрочка 1–5 дней, 1	1
Просрочка 6–29 дней, 2	2
Просрочка 30–59 дней, 3	3
Просрочка 80–89 дней, 4	4
Просрочка 90–119 дней, 5	5
Просрочка 120–149 дней, 6	6
Просрочка 150–179 дней, 7	7
Просрочка 180–209 дней, 8	8
Просрочка 210–239 дней, 9	9
Просрочка ≥ 240 дней, А	10
Безнадежный долг (списан с баланса), В	10
Договор закрыт, С	не используется в расчете
Договор продан (переуступка прав требований), S	10
Договор реструктурирован / рефинансирован, R	10
Договор продан коллекторам, W	10
Договор расторгнут, U	не используется в расчете
Субъект КИ признан банкротом, Т	10
Прекращена передача информации в Бюро, I	10

Источник: составлено автором по технической документации АО «БКИ СБ»

Если договор был открыт после даты, на которую запрашивается кредитный отчет, то соответствующий договор в кредитной истории не отображается, как и его платежная дисциплина.

Отметим также несколько важных особенностей отображения платежной дисциплины. Если в течение одного календарного месяца заемщик выходил на просрочку по данному договору и закрыл ее, то она все равно будет отображена.

Например, договор был в просрочке с 17 по 30 число определенного месяца. На конец месяца просрочка была погашена. Атрибут платежной дисциплины в указанном месяце будет иметь значение «2» (13 дней просрочки).

Если в течение одного календарного месяца заемщик имел несколько просрочек разной длительности по данному договору, то отобразится просрочка с наибольшим количеством дней. Например, договор был в просрочке с 13 по 15 число определенного месяца, т. е. 15 числа просрочка была погашена (2 дня). С 20 по 27 число того же месяца договор второй раз вышел на просрочку (17 дней). Значение атрибута платежной дисциплины в указанном месяце будет равно «2» (17 дней просрочки).

Если договор заемщика находился в просрочке, и договор с источником формирования кредитной истории был расторгнут, то есть информация по данному договору перестала передаваться в БКИ, то данная просрочка будет учтена только в данном месяце. Например, договор находился в просрочке с 10 по 30 число определенного месяца (20 дней). В следующем месяце договор с источником формирования кредитной истории был расторгнут. Состояние в указанном месяце будет «2» (20 дней просрочки). Состояние в следующем месяце будет «I» (прекращена передача информации).

Если в течение одного календарного месяца договор субъекта КИ находился в просрочке и после был закрыт, то за этот месяц будет отображена данная просрочка, а факт закрытия договора отобразится в следующем месяце. Например, договор находился в просрочке с 9 числа предыдущего месяца (месяц – 31 день), а 27 числа текущего месяца просрочка была погашена, и договор был закрыт. Состояние в предыдущем месяце «2» (22 дня). Состояние в текущем месяце «3» (49 дней). Состояние в последующем месяце «С» (договор закрыт).

Статусы договоров на дату оценки кредитного риска заемщика вычисляются на основании анализа данных по кредитному договору, включая дату закрытия договора, наличие текущей просрочки и платежную дисциплину. При этом каждому типу договоров присваивается соответствующее весовое

значение (коэффициент корректировки wt_i) в зависимости от статуса, представленные в таблице 5.3.

Таблица 5.3 – Значения коэффициентов корректировки типа кредита в зависимости от статуса договора

Тип кредита	Текущий статус договора	Коэффициент, wt_i
микрозайм	закрыт	5
	активный без текущей просрочки	4
	со статусами (B, S, R, W, T, I) из таблицы 5.3	10
	активный с текущей просрочкой	10
иной кредит	закрыт	1
	активный без текущей просрочки	1
	активный с текущей просрочкой	3
	обозначение состояния договора: B, S, R, W, T, I	3

Источник: составлено автором

Представленная в таблицах 5.2–5.3 информация взята из справочников технических описаний xml файлов для БКИ «Скоринг Бюро».

Тип кредита для расчета коэффициентов в таблице 5.3 берется из соответствующего поля строки записи о кредите в xml-файле и справочника «Список типов договоров»: код 19 (микрозайм) и 20 (PDL займ) – микрозайм, все остальные коды типа договора – иной кредит. Дополнительно можно анализировать значения полей тип партнера (какая организация передает информацию о кредите в бюро) и сумму и срок займа (для микрозаймов соответственно до 30 000 руб. и до 30 дней включительно).

Рассмотрев подробно описание исходной информации из кредитного отчета в виде xml-файла для расчета агрегированного показателя ККИ перейдем непосредственно к алгоритму его расчета.

Агрегированный показатель ККИ рассчитывается по формуле:

$$CHQ = \frac{1}{W} \sum_{i=1}^n \left(\frac{wt_i}{\sqrt{M_i+1}} \cdot D_i \right), \quad (5.1)$$

где W – нормирующий коэффициент;

i – номер договора в кредитной истории;

n – общее количество кредитов в кредитной истории клиента;

wt_i – корректирующий коэффициент;

M_i – целое число полных месяцев от текущей даты до даты закрытия i -го кредита, или, если последняя отсутствует, до даты последнего обновления информации в кредитной истории заемщика.

Нормирующий коэффициент вычисляется по формуле:

$$W = \sum_{i=1}^n \frac{wt_i}{\sqrt{M_i+1}}, \quad (5.2)$$

где D_i – взвешенная просрочка i -го кредита, определяется по следующей формуле:

$$D_i = \frac{2}{P_i \cdot (P_i + 1)} \sum_{k=1}^{P_i} (P_i - k + 1) \cdot d_{ik}, \quad (5.3)$$

где d_{ik} – номер категории просрочки по k -му платежу i -го кредита;

P_i – количество платежей (длина параметра «платежная дисциплина») по i -му договору без учета значений закрытых и расторгнутых договоров. Значения платежей в параметре «платежная дисциплина» проставляются слева направо: от самого нового по дате платежа ($k = 1$) до самого старого ($k = P_i$).

Значение показателя D_i представляет собой взвешенную просрочку платежей по конкретному кредитному обязательству заемщика. Категории просрочки d_{ik} соответствуют данным таблицы 5.2. При этом наиболее актуальные платежи (просрочки) будут иметь больший вес в расчете значения показателя D_i .

Корректирующий коэффициент wt_i зависит от типа кредита и определяется по данным таблицы 5.3. Наиболее приемлемые значения коэффициентов корректировки wt_i могут быть определены посредством оптимизации некоторой целевой функции, например максимизации значения информационной величины (Information Value) исследуемого признака или значения метрики индекса Джини (Gini) рассматриваемой модели. Обозначенные коэффициенты были выведен

эмпирическим путем на основе тестируемой выборки и могут быть откорректированы с учетом особенностей конкретной выборки.

Показатель M_i рассчитывается только для закрытых договоров, для остальных договоров M_i равен 0.

На основании представленного алгоритма расчета определяют агрегированный показатель ККИ, который может принимать значение в заданном диапазоне от 0 до 10 баллов. Таким образом, агрегированный показатель ККИ равный 0 соответствует наилучшему кредитному рейтингу заемщика, а значение 10 – наихудшему.

Для иллюстрации предлагаемых в настоящем исследовании алгоритма и методики оценки кредитного риска рассмотрим пример расчета агрегированного показателя ККИ для одного заемщика на примере 5 кредитных договоров.

В таблице 5.4 представлены исходные данные по кредитным договорам.

По каждому из договоров определяются параметры «платежная дисциплина», а также в соответствии со статусом кредита присваивается корректирующий коэффициент, показатель количества месяцев до закрытия (для закрытых договоров) и взвешенная просрочка по каждому кредитному договору из таблицы 5.4.

Таблица 5.4 – Исходные данные по кредитным договорам

Номер заявки	Дата заявки	Тип кредита	Тип партнера	Дата закрытия (факт)	Дата последнего изменения записи
1	07.02.2023	Микрозаем	МФО		25.11.2022
1	07.02.2023	Заем (кредит)	МФО	05.12.2022	05.12.2022
1	07.02.2023	Заем (кредит)	Банк		04.02.2023
1	07.02.2023	Микрозаем	МФО		04.02.2023
1	07.02.2023	Микрозаем	МФО		05.02.2023

Источник: составлено автором

В таблице 5.5 представлены параметры, необходимые для расчета агрегированного показателя ККИ.

Для каждого кредитного обязательства значения средневзвешенной просрочки D_i по формуле (5.3) будут равны:

$$D_1 = \frac{2}{1 \cdot (1+1)} (1 - 1 + 1) \cdot 0 = 0,$$

$$D_2 = \frac{2}{2 \cdot (2+1)} ((2 - 1 + 1) \cdot 2 + (2 - 2 + 1) \cdot 2) = 2,$$

$$D_3 = \frac{2}{9 \cdot (9+1)} ((9 - 1 + 1) \cdot 2 + (9 - 2 + 1) \cdot 2 + (9 - 3 + 1) \cdot 0 + (9 - 4 + 1) \cdot 0 + (9 - 5 + 1) \cdot 0 + (9 - 6 + 1) \cdot 0 + (9 - 7 + 1) \cdot 2 + (9 - 8 + 1) \cdot 2 + (9 - 9 + 1) \cdot 0) = 0,98,$$

аналогично

$$D_4 = 0,70,$$

$$D_5 = 1,17.$$

Таблица 5.5 – Значения параметров для расчета агрегированного показателя ККИ

Платежная дисциплина	Статус кредита	Вес, wt_i	M_i	Взвешенная просрочка, D_i
0	Активный без просрочки	4	0	0
СС22	Закрыт	1	2	2
220000220	Активный с просрочкой	3	0	0,98
1100-	Активный с просрочкой	10	0	0,70
120	Активный с просрочкой	10	0	1,17

Источник: составлено автором

Из всех договоров только второй договор закрыт и значение M_2 для него будет равно 2 (2 месяца между датой заявки и датой закрытия договора), для остальных договоров $M_1=M_3=M_4=M_5=0$.

Значения wt_i из таблицы 5.3 будут соответственно равны:

$$wt_1 = 4;$$

$$wt_2 = 1;$$

$$wt_3 = 3;$$

$$wt_4 = 10;$$

$$wt_5 = 10.$$

Нормирующий коэффициент по формуле (5.2) будет равен:

$$W = \frac{4}{\sqrt{0+1}} + \frac{1}{\sqrt{2+1}} + \frac{3}{\sqrt{0+1}} + \frac{10}{\sqrt{0+1}} + \frac{10}{\sqrt{0+1}} = 27 + \frac{1}{\sqrt{3}}.$$

Итоговое значение показателя ККИ по формуле (5.1) будет равно:

$$CHQ = \left(\frac{1}{27 + \frac{1}{\sqrt{3}}} \right) \left(\frac{4}{\sqrt{0+1}} \cdot 0 + \frac{1}{\sqrt{2+1}} \cdot 2 + \frac{3}{\sqrt{0+1}} \cdot \frac{44}{45} + \frac{10}{\sqrt{0+1}} \cdot 0,7 + \frac{10}{\sqrt{0+1}} \cdot 1 \frac{1}{6} \right) \approx 0,83.$$

По представленному расчету значение агрегированного показателя ККИ составляет примерно 0,83, т. е. близко к 1, что соответствует хорошему кредитному рейтингу заемщика, несмотря на наличие просрочек, т. к. по текущим микрозаймам клиент платит.

Был рассмотрен подробно алгоритм и пример расчета предлагаемого показателя ККИ на примере кредитного отчета АО «БКИ СБ». Следует отметить, что сведения о платежной дисциплине в кредитных отчетах других бюро имеют иную структуру. Но в целом логика реализации алгоритмов расчетов будет аналогична. Однако необходимо будет разработать свои таблицы-справочники вычисления весов для каждого БКИ, аналогично таблицам 5.2 и 5.3. В случае использования МФО для оценки кредитных рисков информации из нескольких БКИ можно применить алгоритмы расчета показателя ККИ для каждого бюро в отдельности, а затем взять из них среднюю арифметическую простую или взвешенную. В последнем случае в качестве весов может выступать количество записей о кредитах в соответствующем бюро.

Следует отметить, что до 01.10.2023 года БКИ использовали версию TUTF единого формата передачи кредитных историй, а после этой даты стали использовать версию RUTDF. В старом формате TUTF информация о платежной дисциплине по кредитам в НБКИ, Скоринг Бюро и ОКБ содержалась в отдельном блоке. После перехода на новый формат RUTDF только «Скоринг Бюро» продолжило предоставлять этот блок информации, который стал не обязательным, в новом формате в отдельной части кредитного отчета. Именно поэтому при разработке алгоритма расчета признака ККИ и его апробации

использовались данных АО «БКИ Скоринг Бюро». Тем не менее паттерны платежной дисциплины каждого договора в кредитном отчете нового формата в АО «НБКИ» и АО «ОКБ» можно сформировать самостоятельно на основе информации о наличии просрочек в каждом месяце обслуживания договора. При этом информацию о паттернах платежной дисциплины договоров открытых и закрытых до 01.10.2023 года можно получить в старом формате TUTDF в отдельном блоке кредитного отчета. Для договоров, открытых уже после 01.10.2023 года, – вычислить самостоятельно на основе данных кредитного отчета в новом формате. А вот паттерны платежной дисциплины по договорам, открытых в период до 01.10.2023 года, но срок действия которых еще продолжился после этой даты, придется получать из кредитных отчетов двух форматов: TUTDF и RUTDF.

Апробация предложенной методики измерения качества кредитной истории заемщика для оценки кредитных рисков была проведена автором в нескольких российских МФО, что подтверждается справками. Ниже приведены результаты первичной апробации разработанной методики в одной из МФО.

В качестве выборки для апробации использовались данные по выдачам PDL займов первичным клиентам в феврале 2023 года по стратегии «ручейка» или «чистой выборки» (open gate). В системе принятия решений по данной стратегии применялись только базовые проверки, предусмотренные законодательством РФ и кредитной политикой микрофинансовой компании по данному продукту, скоринговая модель не использовалась. Все клиенты были идентифицированы, мошеннические заявки и выданные займы мошенникам были исключены. Такая стратегия выдачи традиционно используется для сбора «чистых» данных по потоку входящих клиентов для оценки качества кредитной политики и построения внутри кредитной организации своей скоринговой модели.

Одним из источников данных в реализованном эксперименте были кредитные отчеты Акционерного общества «Бюро кредитных историй «Скоринг Бюро» (АО «БКИ СБ»), скоринговый балл обобщенной модели (generic model) кредитного скоринга для продукта PDL, построенной специалистами АО «БКИ

СБ», и скоринговый балл модели, построенной самостоятельно риск-аналитиками микрофинансовой организации. Таким образом, после «вызревания» рисков тестировались три решения для оценки кредитных рисков микрофинансовой организации:

а) Скоринговый балл обобщенной модели. Объем обучающей выборки для построения модели составил более сотни тысяч выдач крупнейших МФО России за 2021–2022 гг. Используемая математическая модель – алгоритм машинного обучения XGBoost («градиентный бустинг»). В качестве проверочной выборки рассматривалась выборка «ручейка» объемом 2950 наблюдений, для которых был получен ответ от бюро с расчетом скорингового балла.

б) Скоринговый балл самостоятельно разработанной внутри МФО модели. Используемый математический аппарат: логистическая регрессия. В качестве обучающей и проверочной выборки использовалась выборка «ручейка» объемом 2868 наблюдений.

в) Рассчитанный по данным кредитных отчетов по алгоритму автора признак качества кредитной истории, представляющий собой рейтинг заемщика. Для расчетов и проверки использовалась та же выборка «ручейка» объемом 2852 наблюдений.

Объемы проверочных выборок немного различаются из-за особенностей подготовки исходных данных и состава информации, используемой для соответствующей модели оценки кредитных рисков, но в целом выборки сопоставим между собой и представляли одну и ту же совокупность наблюдений.

После «вызревания» рисков доля займов с отсутствием первого платежа через 30 дней после оформления займа или отсутствия продления (FPD30) составила на выборке «ручейка» около 60%. Для оценки качества классификационной способности в целом тестируемых моделей оценки кредитного риска использовался индекс Джини (таблица 5.6).

По представленным в таблице 5.6 данным видно, что признак качества кредитной истории обладает более низкой ранжирующей способностью по сравнению с классической статистической моделью или моделью машинного

обучения для оценки кредитных рисков. Этот результат закономерен, т. к. в скоринговых моделях в качестве независимых переменных учитывается множество признаков, и просрочка одна из них.

Таблица 5.6 – Результаты классификационной способности тестируемых решений для оценки кредитных рисков

Решение	Доля клиентов с просрочкой платежа через 30 дней в крайнем «плохом» дециле, %	Индекс Джини, %
Обобщенная модель от БКИ (n = 2950)	84,6	45
Модель логистической регрессии (n = 2868)	90,5	54
Признак качества кредитной истории (n = 2852)	77,5	29

Источник: составлено автором

Однако коэффициент Джини около 30% это достаточно хороший результат качества классификации для одного признака. Порог 30% обычно рассматривается риск-аналитиками как минимальное значение коэффициента Джини, при котором скоринговая модель может быть внедрена в СПР. При этом трудозатраты на реализацию алгоритма расчета признака качества кредитной истории и его интеграцию в правила оценки кредитных рисков в СПР намного меньше по сравнению с разработкой и внедрению полноценной скоринговой модели. Более того, источник относительно бесплатный, как и калькулирование балла по внутренней скоринговой модели.

Полученные скоринговые баллы и значения признака качества кредитной истории были для апробируемых выборок разбиты на децили риска (таблица 5.7).

Необходимо отметить, что тестируемая обобщенная модель БКИ ранжирует заемщиков от худшего к лучшему: минимальный балл по модели характерен для наихудшего заемщика, а максимальный для наилучшего, выборка была разбита на 10 групп по уровню риска. Ранжирование риска идет достаточно хорошо, а

распределение балла близко к нормальному, т. к. изначально модель строилась на выборке более сотни тысяч наблюдений по всем рынку на основе данных БКИ, а тестировалась уже на данных «ручейка» конкретной МФО объемом 2950 наблюдений (рисунок 5.2).

Таблица 5.7 – Децили риска обобщенной модели БКИ на выборке «ручейка» и значения показателя FPD30

Децили риска	Обобщенная модель БКИ		Модель логистической регрессии		Признак ККИ	
	FPD30		FPD30		FPD30	
	нет	да	нет	да	нет	да
1	44 (15%)	242 (85%)	202 (77%)	60 (23%)	308 (54%)	262 (46%)
2	54 (18%)	242 (82%)	238 (77%)	72 (23%)		
3	78 (27%)	213 (73%)	111 (57%)	85 (43%)	168 (59%)	117 (41%)
4	74 (25%)	217 (75%)	190 (50%)	189 (50%)	133 (47%)	152 (53%)
5	107 (36%)	192 (64%)	123 (37%)	206 (63%)	129 (45%)	156 (55%)
6	119 (41%)	172 (59%)	71 (30%)	166 (70%)	99 (34%)	188 (66%)
7	150 (49%)	159 (51%)	76 (26%)	218 (74%)	108 (38%)	176 (62%)
8	144 (50%)	143 (50%)	64 (23%)	215 (77%)	75 (26%)	210 (74%)
9	191 (65%)	103 (35%)	46 (18%)	208 (82%)	64 (22%)	222 (78%)
10	231 (75%)	75 (25%)	31 (10%)	297 (90%)	64 (22%)	221 (78%)
Всего	1192 (40%)	1758 (60%)	1152 (40%)	1716 (60%)	1148 (40%)	1704 (60%)

Источник: составлено автором на основе выполненных расчетов

Модель логистической регрессии выдает вероятность дефолта заемщика от 0 до 1 и ранжирует риски в таблице 3.13 от наилучшей до наихудшей группы заемщиков. Модель также достаточно хорошо ранжирует кредитный риск, но распределение балла имеет более равномерное распределение, что объясняется тем, что модель строилась и тестировалась на выборке ручейка объемом 2868 наблюдений (рисунок 5.3).

Признак ККИ также ранжируют заемщиков от наилучшего к худшему. При группировке заемщиков из-за особенностей распределения было образовано 9 групп (1 и 2 дециль были объединены). Распределение признака ККИ имеет экспоненциальный вид, как и показатели количества дней просрочки по кредитам, сумма просрочки по кредитам, что является логичным. Наибольшее количество заемщиков без просрочек, поэтому имеем модальное значение в точке ноль, а далее значения показателя экспоненциально убывают (рисунок 5.4).

Диаграмма

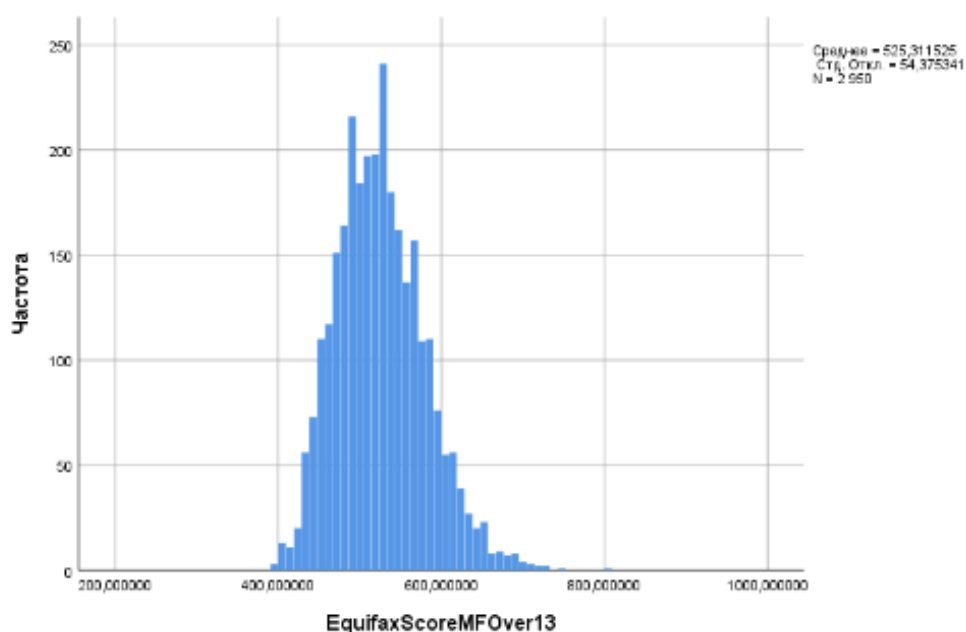


Рисунок 5.2 – Распределение скорингового балла обобщенной модели БКИ

Источник: составлено автором на основании самостоятельных расчетов в SPSS

Для полученных распределений оценок кредитного риска по трем апробируемым моделями были рассчитаны значения 90% процентиля. В крайнем 10 «плохом» с точки зрения кредитного риска дециле была рассчитана доля заемщиков с просрочкой первого платежа через 30 дней (FPD30). Признак ККИ по сравнению с двумя тестируемыми моделями в меньше степени ранжирует риски, как было отмечено из-за особенностей экспоненциального распределения, а также потому что один показатель будет хуже предсказывать кредитный риск по сравнению с моделью, построенной на множестве признаков. Однако у

агрегированного показателя качества кредитной истории доля FPD30 в последнем высоко рисковом дециле составила 77,5%, что является уже достаточной величиной для применения данного признака для оценки кредитных рисков. Информационное значение (IV) для признака ККИ по тестируемой выборке «ручейка» составило 0,31. В кредитном скоринге переменные для построения модели с информационным значением от 0,3 до 0,5 обычно определяют как предикторы с высокой прогностической способностью. Поэтому в практической деятельности МФО признак качества кредитной истории может быть использован и как независимая переменная в самостоятельно разрабатываемой скоринговой модели.



Рисунок 5.3 – Распределение скорингового балла модели логистической регрессии

Источник: составлено автором на основании самостоятельных расчетов в SPSS

Кроме того, применение признака ККИ подразумевает решение проблемы мультиколлинеарности для переменных видов просрочек по кредитам, которые являются всегда одними из самых сильных предикторов в моделях оценки кредитных рисков.

Разработанная автором методика расчета признака качества кредитной истории была реализована в виде программы на объектно-ориентированном языке программирования Python 3.9.7 и зарегистрирована как результат интеллектуальной собственности в качестве программы для ЭВМ [81].

Диаграмма

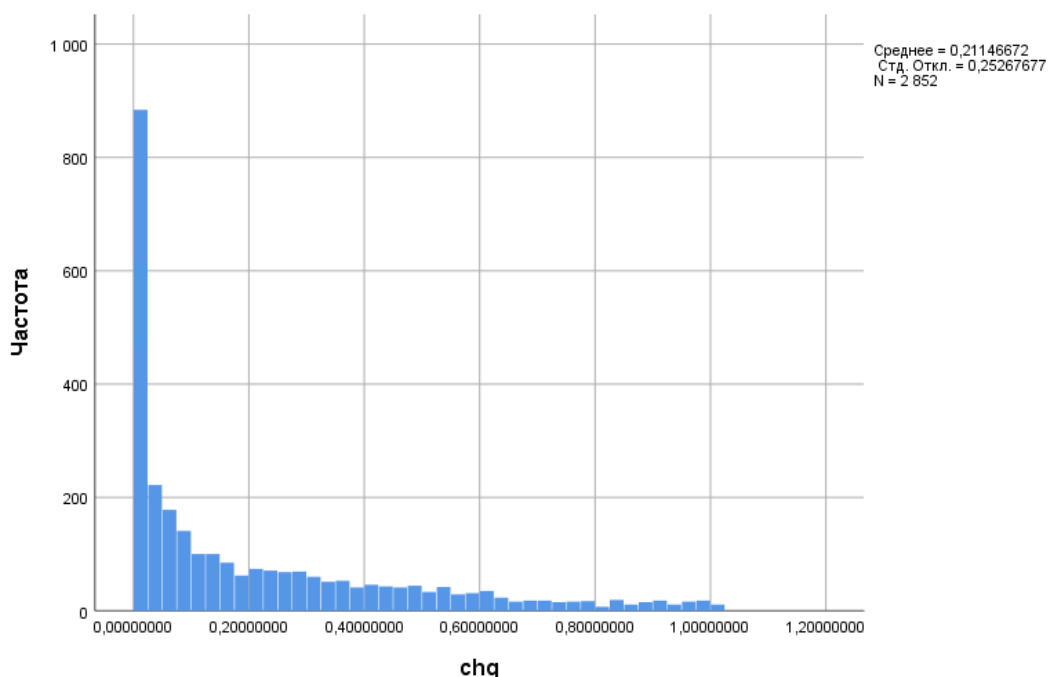


Рисунок 5.4 – Распределение значения признака ККИ

Источник: составлено автором на основании самостоятельных расчетов в SPSS

Программа актуальна к применению для оценки кредитных рисков как в микрофинансовых организациях, так и в банковских и прочих кредитно-финансовых организациях. Листинг кода программы приведен в Приложении Е, а полученное свидетельство программы для ЭВМ в Приложении Ж. Результаты применения показателя качества кредитной истории, на основе алгоритма данной программы используется во нескольких российских МФО, что подтверждено справками о внедрении (см. Приложение И).

Подводя итоги данного параграфа, можно отметить, что разработанная методика оценки кредитного риска заемщика микрофинансовой организации на

основе агрегированного показателя «качество кредитной истории» представляет собой вариант оценки платежеспособности, рассчитанной по всем кредитам клиента, и является альтернативой обобщенным скоринговым моделям или персональному кредитному рейтингу БКИ, а также альтернативой сложным для расчета и интерпретации моделям оценки рисков МФО с большим количеством параметров. Данный агрегированный показатель может использоваться как отдельный показатель для ранжирования заемщиков по шкале от 0 до 10 баллов для определения платежеспособности заемщика, а также в качестве предиктора в скоринговых моделях. Оценка рейтинга заемщика производится с учетом типов кредитов, давности кредитов в кредитной истории и платежной дисциплины заемщика.

5.2 Оценка кредитных рисков на основе статистических моделей и алгоритмов машинного обучения

В настоящем параграфе будут приведены результаты апробации разработанных и предлагаемых автором методики анализа источников данных для оценки кредитных рисков, методики построения моделей оценки кредитных рисков с помощью автоматического советника, сравнительный анализ результатов практического применения статистических моделей и моделей на основе алгоритмов машинного обучения для оценки кредитных рисков на реальных данных российских МФО.

Для иллюстрации, рассматриваемой в главе 2 методики оценки пригодности нового источника данных, приведем пример результатов тестирования для российской МФО данных аналитического агентства «Финкарта» – альтернатива традиционным данным кредитных историй заемщика. Данный источник представляет собой вектор из более 80 исходных переменных с информацией о транзакциях по карте заемщика и о самом заемщике, а также скоринговый балл

обобщенной модели. Новые данные тестировались для решения бизнес-задачи увеличения уровня одобрения и снижения кредитных потерь для сегмента первичных клиентов. Ниже приведены только результаты анализа ценности данных для оценки кредитных рисков общего скорингового балла аналитического агентства «Финкарта».

На первом этапе было проведено ретро-тестирование нового источника данных на выборке заемщиков, представляющей собой выдачу всему потоку входящих клиентов без скоринговой модели (open gate). В выборку попали клиенты, которые получали отказы только по первичным проверкам в СПР и антифрод-правилам. Следует отметить наличие корреляции между флагом успешности запроса о заемщике в сервисе нового источника данных и флагом одобрения заявки (в случае успешно пройденных первичных проверок) (см. таблицу 5.8).

Таблица 5.8 – Пример анализа показателей ценности данных при ретро-тестировании

		Успешность запроса		Метрики
		Нет	Да	
Одобрение заявки	Нет	810 (22,7%)	566 (7,8%)	p < 0,000 тест хи-квадрат Пирсона
	Да	2752 (77,3%)	6724 (92,2%)	
Скоринговый балл	NPL30			IV=0,154
	1 категория	701 (24%)	2009 (41%)	
	2 категория	1242 (43%)	1778 (37%)	
	3 категория	978 (33%)	1054 (22%)	

Источник: составлено автором

Наличие данной корреляции может служить косвенным признаком возможности использования источника данных в начале кредитного конвейера для отсева высокорисковых мошеннических заявок. Однако основной целью применения новых данных было повышение уровня одобрения заявок при

сохранении общего уровня кредитного риска по портфелю. Показатель полноты данных составил около 92%, что является достаточным значением. Отсутствие информации о заемщике в тестируемом новом сервисе рассматривалась как отдельная категория заемщиков с соответствующей степенью риска.

Для ретро-тестируемого скорингового балла была проведена оптимальная категоризация по целевой переменной NPL30 и рассчитан показатель информационного значения (IV). Количество выданных займов оказалось меньше одобренных, т. к. не все заемщики согласились взять заем с одобренным лимитом, который был ограничен 10000 руб. при стратегии выдачи по «чистой выборке». Показатель информационного значения отражает среднюю предсказательную способность. Кроме того, была проведена оценка увеличения коэффициента Джини при включении данного скорингового балла в качестве дополнительной независимой переменной в действующую скоринговую модель МФО для всего потока клиентов. Прирост (Uplift) составил незначительную величину 0,6 п. п. Незначительный прирост коэффициента Джини при использовании в скоринговой модели нового источника можно объяснить достаточно высокой предсказательной способностью уже используемых в ней переменных.

Поэтому следующим шагом стало продолжение тестирования нового источника на отказных заявках в режиме АБ тестирования в «боевом» режиме в течение недели. Для сегмента «серой» зоны аналитически было установлено значение тестируемого скорингового балла (cut off) для одобрения отказных заявок. В таблице 5.9 приведены значения среднего уровня риска в кредитном портфеле организации (NPL), коэффициента конверсии из заявки в выдачу (AR), затрат на выдачу одного кредита (CPI), отношения расходов на получение данных к доходам от их применения (CIR) и коэффициента Джини (Gini) для сравниваемых групп.

Как видно из данных таблицы 5.9, уровень невозвратных кредитов снизился на 1 п. п., коэффициент конверсии увеличился на 2 п. п., а затраты МФО на выдачу одного кредита увеличились на 300 рублей, однако по увеличению показателя CIR можно сделать вывод, что дополнительный доход от внедрения новой

переменной оказался выше затрат на ее использование. При этом коэффициент Джини действующей модели за счет введения новой переменной удалось поднять до значения 0,43. Это свидетельствует о том, что введение новой переменной улучшило точность и эффективность оценки кредитного риска и позволило МФО увеличить свою доходность. При этом следует отметить, что эффективным является применение нового источника данных не на всем потоке заявок, а на определенном сегменте первичных клиентов.

Таблица 5.9 – Пример анализа изменения бизнес-метрик после введения в СПР нового источника данных

Метрика	Контрольная группа А	Тестируемая группа В
NPL, %*	15	14
AR, %*	25	27
CPI, тыс. руб.	2,0	2,3
CIR	0,92	0,90
Gini	0,42	0,43
примечание – различия NPL и AR для контрольной и тестируемой группы статистически значимы при $p < 0,001$ (z-тест равенства долей)		

Источник: составлено автором

Как было показано в главе 3 статистические модели оценки кредитных рисков наиболее часто используются на практике в финансовых организациях и обладают рядом преимуществ. Однако алгоритмы и методы машинного обучения при их корректном применении с соблюдением методики построения скоринговых моделей, рассмотренной в параграфе 1 главы 3, дают в отдельности или в сочетании со статистическим моделированием неплохой результат снижения кредитных рисков.

В рамках исследования была проведена оценка стабильность во времени предсказательной способности четырех типов моделей оценки кредитных рисков, построенных на данных микрофинансовой организации за период с 01.11.2022 по 01.07.2023. Для тестирования статистической модели была выбрана

логистическая регрессия. Для тестирования алгоритмов машинного обучения были выбраны: алгоритм построения простого дерева решений, метод градиентного бустинга и случайного леса. Полная выборка данных представляла 254 756 PDL займа, выданных повторным клиентам, сроком от 5 до 30 дней и суммой ОД от 5 000 до 30 000 руб. В качестве целевой переменной использовался показатель риска NPL30. На момент формирования выборки все статусы категорий показателя NPL30 уже были известны, то есть не было пропущенных значений, и вся выборка представляла собой «созревшие» наблюдения.

Важность правильности расчета целевого показателя оценки кредитного риска при построении скоринговых моделей была показана при рассмотрении методики их построения в параграфе 1 главы 3. Некорректность расчета показателя риска FPD и NPL с учетом пролонгации – одна из типичных ошибок при моделировании кредитных рисков в МФО. Поэтому разберем более подробно на конкретном примере логику расчета данных показателей риска (таблица 5.10).

Таблица 5.10 – Исходные данные для примера расчета показателей FPD и NPL

Дата выдачи	Дата погашения по договору	Дата последнего продления	Фактическая дата закрытия	Всего продлений	Дней просрочки
27.01.2023	06.01.2024	07.01.2024	12.01.2024	1	1
16.01.2024	23.01.2024	22.01.2024	-	1	0
28.12.2023	07.01.2024	17.01.2024	-	2	17

Источник: составлено автором

При расчете показателей риска при вычислении разницы между переменными дат необходимо все даты оформления и платежей по договору привести к часовому поясу заемщика. Исходной информацией для расчета являются даты оформления займа, дата планового погашения, дата продления и фактического закрытия займа. В рассматриваемом примере предполагаем, что продление может быть на срок точно такой же, как и в основном договоре. На расчет показателей риска будет влиять текущая дата расчета, а точнее разница

между количеством дней между ней и датой выхода в просрочку или закрытия договора.

Предположим, что расчет показателей риска производится на дату 24.01.2024. По первому договору клиент взял займ 27.12.2023 года на 10 дней, но не погасил его 6 января, оформив продление также на 10 дней только 7 января. 12 января клиент сделал досрочное погашение и полностью закрыл займ. В этом случае показатель NPL1 и FPD1 будут равны 1, т. к. клиент пользовался деньгами незаконно один день. Поскольку между датой расчета 24 января и датой закрытия по договору 6 января уже прошло 10 дней и клиент совершил продление и закрыл займ, то NPL10 и FPD10 будут равны нулю. В случае закрытия займа мы можем утверждать, что показатели NPL30 и FPD30 также будут равны нулю. Однако для корректности анализа, чтобы не контролировать постоянно «окно созревания» выборки, корректнее определить эти показатели на дату 24.01.2024 как пропущенные, т. к. между 24 и 6 января еще не прошло 30 дней (показатели риска «не вызрели») (таблица 5.11). Таким образом, если между NPLX или FPDX = системному пропуску, если (Дата расчета показателя – Дата закрытия по договору) $\leq X$.

По второму договору клиент брал займ на 7 дней и за один день до его закрытия совершил продление также на 7 дней с датой погашения 02.02.2024, поэтому FPD1 и NPL1 для него на дату 24.01.2024 равны 0, а вот показатели риска на горизонте 10 и 30 дней будут еще не заполнены.

По аналогичной логике выполняются расчеты показателей риска для третьего договора, по которому было два продления. В случае нескольких продлений для определения FPD в силу сущности показателя берется только первый платеж, а вот для показателя NPL расчет ведется по всем периодам между продлениями, т. е. $NPL X = 1$, если хотя бы один из периодов просрочки $\geq X$ (рисунок 5.5).

По третьему договору клиент совершил два продления в срок окончания договора на 10 дней каждое, поэтому показатели FPD не равны 1, однако займ не

закрит, поэтому показатели NPL равны 1. Показатели риска на горизонте 30 дней не заполнены, т. к. еще «не созрели».

Таблица 5.11 – Значения показателей FPD и NPL

Срок договора	Сумма	Погашено	FPD1	FPD10	FPD30	NPL1	NPL10	NPL30
10	4000	4682	1	0	-	1	1	-
7	4000	200	0	-	-	0	-	-
10	10000	1600	0	0	-	1	1	-

Источник: составлено автором

На рисунке 5.5. клиент два раза просрочил оплату займа на 2 и 5 дней, но затем сделал продление каждое по 10 дней.

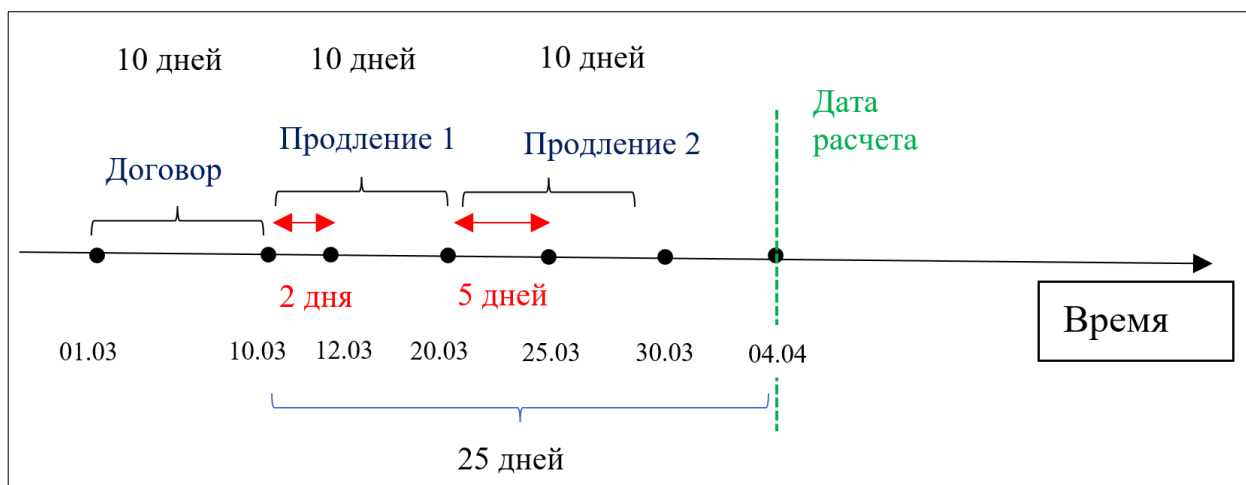


Рисунок 5.5 – Пример расчета показателя NPL5 в случае нескольких продлений (NPL5=1)

Источник: составлено автором

На дату расчета 4 апреля показатель риска NPL5 является «созревшим» и составляет 1, но поскольку займ был закрыт 30 марта, то, например, показатель NPL10 будет равен 0, а показатель NPL30 примет пропущенной значение, т. к. прошло только 25 дней.

В качестве независимых переменных в тестируемых моделях были использованы переменные, вычисленные на основе данных кредитных историй клиентов, полученных из двух БКИ. Для расчетов переменных использовались как данные каждого из БКИ в отдельности, так и данные объединенной выборки обеих БКИ, с произведенной очисткой такой выборки от дубликатов записей о кредитах. Полученный в итоге набор данных насчитывал в итоге порядка 6 500 переменных-агрегатов, описывающих количество кредитов, суммы их основного долга, суммы остатков срочной и просроченной задолженности, количество и долю кредитов с просрочками и другие параметры в разрезах различных типов кредитов, типов кредитных или микрофинансовых организаций, их выдавших, периодов выдачи или наступления события по кредиту, других разрезах рассмотренных в параграфе 1 главы 3 диссертационной работы.

Для проверки гипотезы о большей устойчивости моделей кредитного риска в условиях наличия у кредитной организации сравнительно небольших выборок, из основной выборки была сначала сформирована выборка для тестирования моделей (далее – тестовая выборка), которую составляли данные о займах, выданных за последние 3 месяца (период с 01.05.2023 по 31.07.2023, $N = 73\,582$, $NPL30 = 29\%$). Из выборки займов за предшествующих период были случайным образом, но со стратификацией по распределению целевой переменной, извлечены еще 3 подвыборки для обучения моделей:

а) Обучающая выборка 1: $N = 2\,000$, $NPL30 = 25\%$, период с 01.11.2022 по 30.04.2023;

б) Обучающая выборка 2: $N = 20\,000$, $NPL30 = 25\%$, период с 01.11.2022 по 30.04.2023;

в) Обучающая выборка 3: $N = 181\,174$, $NPL30 = 25\%$, период с 01.11.2022 по 30.04.2023.

Затем на каждой из обучающих подвыборок все независимые переменные были категоризированы методом оптимальной категоризации. Для оптимальной категоризации использовался алгоритм MDLP, выполняющий поиск такой

трансформации переменной в категориальную, при котором максимизируется значение информационного значения (IV) относительно целевой переменной NPL30. При моделировании использовался подход замены значений категорий независимых переменных после MDLP-дискретизации числовыми значениями WOE для полученных категорий.

Далее между независимыми переменными до преобразования были рассчитаны коэффициенты корреляции и показатели VIF. Из коррелирующих между собой переменными с коэффициентом корреляции более 0,6 по модулю были оставлены те переменные, у которых значение показателя IV с целевой переменной было максимально. Оставшиеся в выборке независимые переменные были проранжированы по значению IV. В итоге для каждой тренировочной подвыборки было подобрано около 20 слабо коррелирующих между собой независимых переменных с наибольшими значениями IV для предсказания целевого показателя риска.

По каждой из трех подвыборок были обучены модели кредитного риска четырех разных типов и замерены такие их параметры, как коэффициент Джини, мера достоверности (Ассигасу), F-мера и доля займов с предсказанной вероятностью дефолта (PD) NPL30 меньше 15% (таблица 5.12).

Спецификация моделей оценки кредитного риска была неслучайна, как было показано в 3 главе, представленные в таблице 5.12 модели – наиболее популярные на практике из классического моделирования и алгоритмов машинного обучения.

Следующим шагом анализа, для каждой из 12 полученных моделей были получены оценки скоринга займов для тестовой выборки, и для каждой из модели на тестовой выборке были рассчитаны вышеописанные метрики качества моделей (таблицы 5.13–5.15).

При расчете показателей качества классификации использовался порог отсечения по вероятности дефолта по умолчанию 0,5. На первой обучающей малого объема и тестовой выборке мы видим, что модель логистической регрессии была сильно переобучена: коэффициент Джини на обучающей выборке

составил 0,41, а на тестовой он упал почти в 2 раза до 0,23. Дерево решений или построение случайного леса из дерева решений дали не самый лучший результат на тестовой выборке.

Таблица 5.12 – Список моделей и их основных гиперпараметров, выбранных при обучении модели в пакете Python

Выборка для обучения	Тип модели	Основные гиперпараметры модели
Обучающая выборка 1	Дерево решений	<code>sklearn.tree.DecisionTreeClassifier(max_depth=2)</code>
	Случайный лес	<code>sklearn.ensemble.RandomForestClassifier(n_estimators=2, max_depth=2)</code>
	Логистическая регрессия	<code>sklearn.linear_model.LogisticRegression(max_iter=100)</code>
	Градиентный бустинг	<code>xgboost.XGBClassifier(max_depth=5, learning_rate=0.01, n_estimators=100)</code>
Обучающая выборка 2	Дерево решений	<code>sklearn.tree.DecisionTreeClassifier(max_depth=3)</code>
	Случайный лес	<code>sklearn.ensemble.RandomForestClassifier(n_estimators=8, max_depth=5)</code>
	Логистическая регрессия	<code>sklearn.linear_model.LogisticRegression(max_iter=100)</code>
	Градиентный бустинг	<code>xgboost.XGBClassifier(max_depth=5, learning_rate=0.02, n_estimators=100)</code>
Обучающая выборка 3	Дерево решений	<code>sklearn.tree.DecisionTreeClassifier(max_depth=5)</code>
	Случайный лес	<code>sklearn.ensemble.RandomForestClassifier(n_estimators=9, max_depth=9)</code>
	Логистическая регрессия	<code>sklearn.linear_model.LogisticRegression(max_iter=100)</code>
	Градиентный бустинг	<code>xgboost.XGBClassifier(max_depth=5, learning_rate=0.03, n_estimators=200)</code>

Источник: составлено автором

Стабильный результат на обучающей и тестовой выборке дал алгоритм градиентного бустинга, однако данный алгоритм на первой обучающей и тестовой выборке склонен давать завышенную оценку вероятности дефолта: доля займов с предсказанной вероятностью дефолта (PD) NPL30 меньше 15% получилась равной 0.

Таблица 5.13 – Сравнение метрик качества построенных моделей на обучающей выборке 1 и контрольной выборке

Тип модели	Обучающая выборка 1				Тестовая выборка			
	Джини	Достовер- ность	F- мера	Доля PD < 0,15, % (шт)	Джини	Достовер- ность	F-мера	Доля PD < 0,15, % (шт)
Дерево решений	0,20	0,60	0,43	18 (354)	0,17	0,58	0,44	16 (12115)
Случайный лес	0,24	0,59	0,46	42 (843)	0,17	0,57	0,46	44 (32087)
Логистическая регрессия	0,41	0,70	0,54	37 (739)	0,23	0,67	0,46	36 (26745)
Градиентный бустинг	0,22	0,42	0,46	0 (0)	0,20	0,49	0,49	0 (0)

Источник: составлено автором

На второй обучающей выборке среднего объема и тестовой подвыборке по-прежнему алгоритмы построения модели оценки кредитных рисков на основе деревьев решений не являются наиболее подходящими. Логистическая регрессия и градиентный бустинг дали примерно сопоставимые между собой результаты. Аналогичные результаты получены и для третьей обучающей самого большого объема и тестовой выборки.

Таким образом, на эмпирических расчетах показана статистическая устойчивость и стабильность работы модели логистической регрессии и алгоритма градиентного бустинга. В целом эти две модели оценки кредитного риска дают примерно схожий результат, учитывая возможный модельный риск при машинном обучении и простоту интерпретации логистической регрессии. Но при большом потоке входящих заявок на займ алгоритм машинного обучения на основе градиентного бустинга может дать больший экономический эффект.

Таблица 5.14 – Сравнение метрик качества построенных моделей на обучающей выборке 2 и контрольной выборке 2

Тип модели	Обучающая выборка 2				Тестовая выборка 2			
	Джини	Достоверность	F-мера	Доля PD < 0,15, % (шт)	Джини	Достоверность	F-мера	Доля PD < 0,15, % (шт)
Дерево решений	0,25	0,57	0,46	11 (2118)	0,21	0,62	0,46	10 (7372)
Случайный лес	0,29	0,61	0,48	19 (3718)	0,25	0,65	0,48	18 (13448)
Логистическая регрессия	0,29	0,63	0,48	26 (5284)	0,26	0,66	0,49	25 (18578)
Градиентный бустинг	0,31	0,57	0,49	11 (2140)	0,28	0,62	0,51	10 (7282)

Источник: составлено автором

Таблица 5.15 – Сравнение метрик качества построенных моделей на обучающей выборке 3 и контрольной выборке 3

Тип модели	Обучающая выборка 3				Тестовая выборка 3			
	Джини	Достоверность	F-мера	Доля PD < 0,15, % (шт)	Джини	Достоверность	F-мера	Доля PD < 0,15, % (шт)
Дерево решений	0,25	0,59	0,46	22 (39852)	0,22	0,66	0,45	21 (15778)
Случайный лес	0,31	0,61	0,49	21 (38033)	0,26	0,65	0,49	20 (14383)
Логистическая регрессия	0,27	0,61	0,47	24 (43529)	0,25	0,66	0,48	23 (16979)
Градиентный бустинг	0,31	0,62	0,49	24 (42888)	0,28	0,67	0,5	22 (16594)

Источник: составлено автором

Несмотря на популярность применения алгоритма градиентного бустинга при оценки кредитных рисков модель логистической регрессии является неким стандартом в кредитном скоринге. Выше в главе 3 была подробно разобрана проблема мультиколлинеарности, которая, также как и для линейной регрессии,

актуальная для модели логистической регрессии, особенно в условиях ограниченности объема выборки. Приведем пример практической реализации метода главных компонент, который позволил оценить кредитные риски для потока первичных клиентов МФО в условиях малой выборки в условиях мультиколлинеарности.

Объем обучающей выборки составил 306 наблюдений с периодом выдачи PDL займов с ноября по декабрь 2023 года по стратегии максимально приближенной к «чистой выборке»: в СПР работали только базовые проверки по законодательству, антифрод-правила для идентификации клиента и кредитные правила, но без скоринговой модели.

В качестве целевого показателя риска использовался FPD30. Уровень дефолтности составил около 55%. Займы выдавались на срок от 7 до 30 дней, средний срок займа составил $26,5 \pm 6,5$ дней. Сумма займа варьировала от 2 000 до 15 000 руб., средний чек составил $12\,397 \pm 3\,819$ руб. В качестве исходных данных для моделирования использовались кредитные отчеты из ОАО «Скоринг Бюро» и другие альтернативные источники данных.

Анализ независимых переменных позволил отобрать 8 предикторов, сильно влияющих на зависимую переменную, а также вычислить еще одну переменную – признак качества кредитной истории, алгоритм расчета которой представлен в параграфе 1 данной главы.

Среди отобранных предикторов для дальнейшего моделирования были взяты переменные, образованные различными группировками по типу кредитора или типу кредита, по разному периоду агрегации, разной агрегирующей функции – медианы или средней арифметической, на основе следующих показателей: количество и сумма открытых кредитов, количество и сумма погашенных кредитов, количество и сумма просроченных кредитов, дней просрочки по всем кредитам.

При объеме выборке 306 наблюдение применение методов и алгоритмов машинного обучения некорректно, поэтому для оценки кредитного риска была выбрана математическая модель логистической регрессии. Попытка построить

статистически устойчивую модель со значимыми коэффициентами и приемлемым качеством классификации была неудачна. Причина заключалась в сильной корреляции между независимыми переменными и невозможности их включения одновременно в модель при таком малом количестве наблюдений. Многие коэффициенты корреляции между независимыми переменными были выше 0,8, а многие значения показателей VIF превышали 5. Поэтому было принято решение провести компонентный анализ на основе исходных 9 независимых переменных ($X_1 - X_9$).

Компонентный анализ проводился на основе матрицы парных корреляций Пирсона [229], отбирались компоненты с собственными значениями больше 1, для улучшения интерпретации полученных компонент использовали вращение матрицы нагрузок методом Варимакс [264]. Извлеченные компоненты по большинству исходных переменных получили общности выше 0,8, что является удовлетворительным результатом (таблица 5.16).

Таблица 5.16 – Общности исходных переменных

Независимая переменная	До извлечения	После извлечения
X_1	1,000	0,846
X_2	1,000	0,813
X_3	1,000	0,624
X_4	1,000	0,225
X_5	1,000	0,938
X_6	1,000	0,920
X_7	1,000	0,915
X_8	1,000	0,790
X_9	1,000	0,750
примечание – точные названия переменных не раскрываются из-за конфиденциальности		

Источник: составлено автором на основании расчетов

Были отобраны 3 главные компоненты с собственными значениями больше 1, которые в совокупности объясняют около 76% дисперсии исходных независимых переменных (таблица 5.17).

Таблица 5.17 – Собственные значения главных компонент

Главная компонента	Собственное значение	% объясняемой дисперсии	Кумулятивный % объясняемой дисперсии
1	2,977	33,078	33,078
2	2,597	28,853	61,931
3	1,277	14,191	76,122
4	0,873	9,699	85,821
5	0,528	5,864	91,686
6	0,358	3,975	95,660
7	0,181	2,015	97,675
8	0,121	1,346	99,022
9	0,088	0,978	100,000

Источник: составлено автором на основании расчетов

По матрице нагрузок (коэффициентов корреляций исходных показателей и главных компонент) после вращения была дана интерпретация полученных главных компонент (таблица 5.18).

Таблица 5.18 – Матрица нагрузок после вращения

Переменная	ГК1	ГК2	ГК3
1	0,954	0,014	– 0,097
2	0,941	0,104	0,137
3	0,937	0,033	0,243
4	0,054	0,917	– 0,059
5	0,039	0,889	– 0,142
6	0,037	0,770	– 0,170
7	0,084	0,028	0,884
8	0,260	– 0,183	0,805
9	0,045	0,188	– 0,467
примечание – точные названия переменных не раскрываются из-за конфиденциальности			

Источник: составлено автором на основании расчетов

Результаты компонентного анализа получены удовлетворительные, интерпретация полученных компонент логичная и хорошо интерпретируема. Первая главная компонента (ГК1) очень сильно коррелирует с переменными просрочек по микрокредитам, чем она выше, тем выше кредитный риск у заемщика. Она отражает историческую просрочку по кредитам. Вторая главная

компонента (ГК2) коррелирует в большей степени с переменными платежами по микрокредитам, чем она выше, тем ниже кредитный риск у заемщика. Третья главная компонента (ГК3) отражает текущую просрочку по микрокредитам, чем она больше, тем больше количество дней просрочки открытых кредитов с учетом давности просрочки и типа кредита.

В дальнейшем на основе линейных комбинаций исходных независимых переменных (ГК) строилась логистическая регрессия, результаты которой приведены в таблице 5.19.

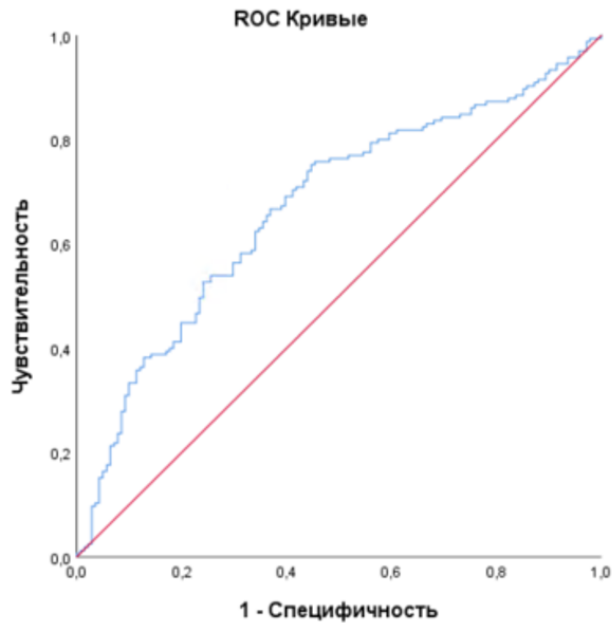
Уравнение получилось статистически значимым со значимыми коэффициентами. Р-значение статистики Вальда при проверке значимости коэффициента при ГК3 получено 0,09, что формально меньше 0,05. Однако значение р-уровня от 0,05 до 0,1 можно считать в зоне неопределенности, как статистическую тенденцию, и с учетом малой выборки можно данную компоненту оставить в уравнении логистической регрессии. Константа модели статистически незначима ($p > 0,1$), но ее все равно можно оставить в модели, т. к. главные компоненты стандартизованы и их среднее равно нулю.

Таблица 5.19 – Коэффициент логистической регрессии

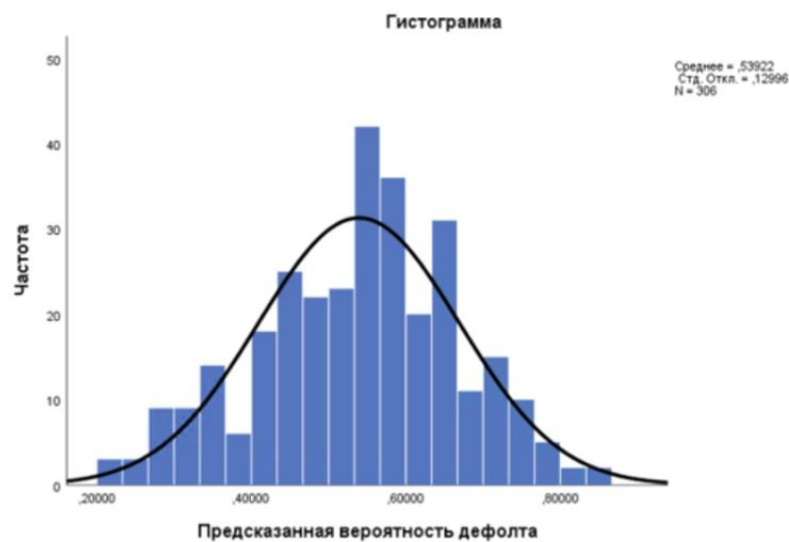
Переменная	Коэффициент	Стандартная ошибка	Статистика Вальда	р-значение	Экспонента коэффициента
константа	0,167	0,119	1,974	0,160	1,182
ГК1	0,321	0,133	5,801	0,016	1,379
ГК2	– 0,417	0,123	11,525	0,001	0,659
ГК3	0,187	0,101	1,873	0,090	1,206

Источник: составлено автором на основании расчетов

Полученная модель имеет индекс Джини 0,34 (95%-доверительный интервал площади под ROC-кривой от 0,610 до 0,732), распределение баллов близко к нормальному, по квартилям риска распределение дефолтов ранжирует риски (рисунок 5.6).



а) ROC-кривая



б) гистограмма распределения скорингового балла

Рисунок 5.6 – Качество классификации построенной модели

Источник: составлено автором

Таким образом, на основании эмпирических расчетов реальной МФО была показана эффективность модели логистической регрессии на малой выборке. Даже при объеме выборке 306 единиц была получена модель оценки кредитного риска с коэффициентом Джини выше минимального порога принятия обобщенных скоринговых моделей (0,3) в СПР для оценки кредитным риском.

В результате апробации изложенных выше в главах 2–4 моделей, методов и алгоритмов автором была разработана и успешно внедрена в практику работы нескольких российских МФО программа «Greeder of Data –автоматический советник построения скоринговых моделей на базе логистической регрессии» (см. Приложение К и Приложение Л) [80]. Программа написана на объектно-ориентированном языке программирования Python 3.9.7 и предназначена для автоматизации и ускорения времени разработки скоринговых моделей для риск-аналитиков и модельеров банков, микрофинансовых организаций на основе логистической регрессии.

Следует отметить, что код программы является универсальным и может работать с любым количеством наблюдений или переменных в качестве исходных данных, а также применяться при построении скоринговых моделей в других сферах. Однако разработанный автором программный код учитывает специфику использования логистической регрессии именно для скоринга в микрофинансовой сфере, подробно изложенную в параграфе 1 главы 3 настоящего исследования.

На вход работы алгоритма программы должен подаваться уже подготовленный массив данных с предикторами в виде объекта DataFrame и бинарный вектор со значениями целевой переменной. Объект DataFrame – это двумерная размеченная структура данных со столбцами, типы которых могут быть разными, является входной информацией для работы программы.

Массив данных с предикторами может содержать количественные и категориальные данные в числовом формате, в значениях переменных могут быть системные пропуски. Вектор зависимой переменной должен иметь числовой формат (значения переменной 0 и 1) и не содержать пропущенных значений.

Укрупненный алгоритм построения модели логистической регрессии, реализованный в программе, можно отразить следующим образом (рисунок 5.7):

- а) категоризация количественных предикторов;
- б) обработка пропущенных значений в предикторах;
- в) расчет информационного значения и веса категорий предикторов;

г) итерационный подбор оптимального сочетания предикторов с учетом мультиколлинеарности и построения на их основе модели логистической регрессии;

д) валидация полученной модели логистической регрессии и сохранение полученных результатов.



Рисунок 5.7 – Общий алгоритм работы программы Greeder of Data

Источник: составлено автором

Полученные результаты построения модели логистической регрессии с помощью разработанной автором программы не обязательно могут выступать в качестве финального решения. В этом аспекте программа выступает в роли советника для модельера или риск-аналитика. Результаты расчетов программы могут быть «фундаментом» для окончательной доработки скоринговой модели в «ручном» режиме или могут быть использованы для быстрой предварительной

оценки возможности построения адекватной скоринговой модели на основе имеющегося массива данных.

Для запуска расчетов программой-советником Greeder of Data для подготовленного массива данных в объекте DataFrame необходимо в среде Python загрузить в оперативную память код программы (см. Приложение К). Вначале отработки кода программы загружаются необходимые для работы библиотеки, а затем необходимые самостоятельно написанные функции для реализации работы алгоритма действий по моделированию, отраженных на рисунке 5.7.

Ниже приведен пример листинга кода для запуска программы в среде Python (рисунок 5.8).

```
>>> model = Modeler(d, 'Target')
>>> model.bin_data(min_hit_rate = 0.1)
>>> model.feature_greeder(max_feature_allow = 10, rounds = 5,
max_pvalue_allow = 0.015)
>>> best_features = model.fr.sort_values('gini', ascending =
False).Features.iloc[0]
>>> model.logit_model(best_features)
>>> model.save_model('./model.json')
```

Рисунок 5.8 – Листинг кода для запуска программы

Источник: составлено автором

В первой строке кода листинга 1 создается объект класса, в который передается набор предикторов (d) и наименование целевой переменной ('Target'). Во второй строке кода запускается функция bin_data, осуществляющая оптимальную категоризацию количественных предикторов. В третьей строке кода запускается функция feature_greeder для подбора наилучшей модели логистической регрессии. В четвертой строке кода лучшие результаты отработки функции feature_greeder сохраняются в отдельные переменные. Предпоследняя строка кода выводит сводку по наилучшей модели, подобранной советником.

При отработке кода последней строки параметры финальной модели сохраняются в формате json для экспорта и использования в других приложениях.

Дадим более подробное описание параметров встроенных функций разработанной автором программы, представленной в Приложении К.

Функция `Bin_Data` предназначена для автоматической категоризации количественных предикторов. В основе категоризации лежит оптимизационный алгоритм, минимизирующий меру энтропии при предсказании значений целевой переменной по информации о распределении значений предиктора по образуемым категориям. Процедура оптимальной категоризации выполняет MDLP дискретизацию (категоризацию) переменных.

Полученные в результате работы функции `Bin_Data` новые категоризованные значения предикторов сохраняются в виде отдельных объектов. Для вывода итоговой таблицы результатов разбиения (категоризования) по всем предикторам нужно запустить команду:

```
>>> model.summ.
```

Для просмотра результатов категоризования конкретной переменной нужно задать команду:

```
>>> optd['Feature1'],
```

где `Feature1` – имя переменной.

Функция `Bin_Data` имеет следующие параметры:

- `cols_to_bin` – список из названия предикторов, которые нужно разбить на категории. Если параметр не задан, то по умолчанию категоризируются все предикторы.

- `min_n_bins` – минимально допустимое количество категорий в разбиваемой переменной. По умолчанию функция разбивает на основе заданной функции оптимизации внутри алгоритма.

- `max_n_bins` – максимальное допустимое количество категорий в разбиваемой переменной. По умолчанию функция разбивает на основе заданной функции оптимизации внутри алгоритма.

- `min_bin_size` – минимальная доля категории в разбиваемой переменной. По умолчанию данный параметр равен 10%.

- `min_hit_rate` – минимальная доля валидных (непропущенных) значений в исходной количественной переменной. Если доля валидных значений меньше заданного значения, переменная исключается из итоговой выборки, не категорируется и не рассматривается для построения модели. При этом пропущенные значения кодируются в отдельную категорию и для них отдельно анализируется риск наступления целевого события. По умолчанию этот параметр установлен на уровне 20%.

- `must_event_rate` – необходимая доля категории события риска в целевой переменной (в долях единицы). Данный параметр позволяет изменить распределение по категориям целевой переменной, по которой идет оптимизация при разбиении, т.е. перевзвесить выборку данных. Если данный параметр не задан (настройка по умолчанию), то балансировка распределения по категориям целевой переменной не производится.

- `missing_merge_perc` – доля пропущенных значений в предикторе меньше которой, категория пропусков объединяется с ближайшей категорией по значению WOE. По умолчанию данный параметр равен 0, т. е. все пропущенные значения рассматриваются как отдельная категория и для них анализируется риск наступления целевого события.

Функция `Feature_Greeder` предназначена для построения группы моделей, подбора параметров этих моделей и их валидации. С использованием категоризованных оптимально предикторов последовательно строятся модели логистической регрессии с включением различных параметров. Каждая модель подлежит валидации и сравнению с другими построенными моделями на основе ROC-анализа для выявления наилучших параметров по критериям, регулируемым исследователем.

Функцию `Feature_Greeder` можно запустить следующим образом:

```
>>> model.feature_greeder(max_feature_allow = 10, rounds = 5, by_score = ['iv', 'gini'], max_pvalue_allow = 0.015, constant = False, top = 100).
```

Полученную в результате работы функции `Feature_Greeder` итоговую таблицу разбитых переменных можно вывести, запустив команду:

```
>>> model.fr.
```

Для вывода переменных, использованных в лучшей модели, необходимо запустить следующую команду:

```
>>> model.fr.sort_values('gini', ascending = False).Features.iloc[0].
```

Функция `Feature_Greeder` имеет следующие параметры:

а) `max_pvalue_allow` – максимальное р-значение для коэффициентов, при котором переменная включается в модель логистической регрессии. Значение по умолчанию равно 0,015.

б) `by_score` – задание критериев для добавления предикторов в модель при использовании пошагового алгоритма включения. Возможные варианты значений данного параметра:

- 1) 'iv' – информационное значение, первыми в модель включаются предикторы с наибольшим значением IV;
- 2) 'js' – мера дивергенции Йенсена-Шеннона, предиктор включается в модель, если дивергенция увеличивается.
- 3) 'gini' – индекс Джини, предиктор включается в модель, если его добавление увеличивает индекс Джини;
- 4) 'quality_score' – индекс качества модели (QS), предиктор включается в модель, если его добавление увеличивает индекс качества модели;
- 5) 'all' – использование всех критериев.

Можно задавать несколько параметров для включения переменных, указав их в списке (пример: ['js', 'gini']). По умолчанию установлен параметр 'all' для использования всех критериев.

Также доступны следующие настройки работы советника:

- `max_feature_allow` – максимально разрешенное количество переменных в модели. Заданное по умолчанию значение 5.

- `max_vif_allow` – максимально разрешенное значение коэффициента «разбухания» дисперсии (VIF) для предикторов внутри модели, рассчитываемое

по исходным количественным переменным до их категоризации. Заданное по умолчанию значение равно 1,5.

- `constant` – включение в модель логистической регрессии константы. Параметр принимает логические значения `True` или `False`. По умолчанию модель строится с константой.

- `must_feature` – список предикторов, которые обязательно должны быть включены в модель. По умолчанию параметр не задан.

- `top` – задает количество лучших предикторов, предсказывающих целевую переменную по критерию, заданному в параметре `by_score`, которые используются для построения моделей. Заданное по умолчанию значение 5000.

- `rounds` – количество итераций построения модели с предикторами, отсортированными по величине параметра `by_score`. Значение параметра по умолчанию равно 5. Например, при первой попытке в регрессионную модель первой будет включена переменная с максимальной предсказательной способностью. При второй попытке первой в регрессионную модель будет включена вторая по силе предсказательной способности переменная и т. д. до 5. Такой алгоритм позволяет избежать ситуации, когда самая сильная по предсказательной силе независимая переменная не дает включить в модель другие переменные с намного меньшей предсказательной способностью. Например, в модель может быть включены всего две переменные: самая сильная и одна более слабая с коэффициентом Джини модели 0,35. А если первой включить в регрессионную модель вторую по предсказательной силе переменную, то с ней вместе уже будет включено, например, еще три предиктора, и коэффициент Джини получится 0,43. Второй набор предикторов будет являться более предпочтительным.

Подводя итоги данного параграфа, можно сделать вывод об успешном применении рассмотренных и предложенных автором методик оценки ценности данных в главе 2, моделей оценки кредитного риска, рассмотренных в главе 3, на исторических данных российских МФО, что подтверждается соответствующими справками о внедрении (см. Приложение И).

5.3 Управление кредитными рисками на основе моделей лимитов, юнит-экономики и скоринговой модели с изменяющимися во времени параметрами

В главе 4 автором были предложены разработанные модели и методы управления кредитными рисками МФО. В данном параграфе будут рассмотрены основные результаты их апробации на исторических данных российских МФО и на имитационных данных.

Для эмпирической апробации разработанной модели назначения лимитов для управления кредитными рисками были взяты реальные данные микрофинансовой компании на рынке России за период с 01.01.2021 по 01.06.2021 с разбивкой по двум сегментам: первичных и повторных клиентов.

Результаты апробации были представлены в работе [103], а также получили положительную оценку на 31-м Международном конкурсе научно-исследовательских работ в номинации «научные статьи по экономическим наукам и менеджменту» (см. Приложение М).

Подготовка данных и все расчеты проводились с помощью языка программирования Python с использованием библиотек Numpy и Pandas, а также MS Excel и SPSS. В полученной выборке показатели были рассчитаны по выданным займам, агрегированным по месяцам, сегментам клиентов и децилям риска. В таблице 5.20 приведен фрагмент исходных данных для моделирования.

Таблица 5.20 – Фрагмент подготовленной таблицы данных для моделирования

Месяц	Сегмент	Дециль	Кол-во займов	NPL30, %	Средний лимит, руб.	Средний ОД, руб.	ROI, %
4	первичный	1	1191	6,9	17 100	7 052	22,3
1	повторный	1	1158	11,9	15 799	9 177	23,1
5	первичный	2	1272	14,5	15 576	10 729	26,6
...	...	...	...	...	...	...	...
3	первичный	4	1199	22,6	15 967	12 622	26,9

Источник: расчеты автора

Первоначально были исследованы описательные статистики исходных данных (таблица 5.21). В целом данные достаточно однородны без явных выбросов и экстремумов: средние значения не сильно отличаются от медиан.

Таблица 5.21 – Статистические характеристики показателей для моделирования

Показатель	Обозначение в моделях	Минимум; Максимум	Среднее $\pm$ стандартное отклонение	Медиана [1 квартиль; 3 квартиль]
NPL30, %	R	4,6; 43,5	26,00 $\pm$ 10,30	28,5 [18,5; 35,1]
ROI, %	P	0,5; 30,9	18,12 $\pm$ 7,03	19,7 [13,5; 23,1]
Средний ОД, руб.	D	5 619; 14 571	10 570 $\pm$ 2 318	10 452 [9 181; 12 428]
Средний лимит, руб.	L	5 664; 18 445	14 224 $\pm$ 3 103	15 441 [12 440; 16 061]

Источник: расчеты автора

Период «созревания» выборки был взят 6 месяцев, т. е. показатели риска и доходности рассчитывались через 6 месяцев после даты выдачи займа. Для оценки риска был выбран показатель NPL30, поскольку он использовался в риск-менеджменте данной МФО в рассматриваемый период времени. В качестве показателя доходности взят ROI, равный разности между суммой полученных от заемщиков и суммой выданных средств, поделенную на объем выданных средств. Объектом наблюдения по подготовленной выборке являлись одобренные и выданные заявки клиентов по действующим в МФО скоринговым моделям, разбитые на децили по уровню риска за каждый месяц.

При подготовке данных были удалены наблюдения ниже 5% нижнего квантиля распределения по количеству наблюдений. Итоговый объем выборки составил 73 наблюдения. Следует отметить, что в реальном исследовании для принятия решений следовало бы взять больший временной охват и объем выборки. Однако в нашем случае в целях неразглашения коммерческой информации осознанно взят меньший объем данных. В рассматриваемом примере 73 наблюдения достаточно для иллюстрации предлагаемой методики моделирования лимитов.

Построение диаграммы рассеяния между уровнем риска и разницей между средним лимитом и средним ОД дало следующие результаты (рисунок 5.9).

Из полученной диаграммы рассеяния видно, что средний ОД приближается к среднему лимиту по мере роста NPL30, то есть более рискованные заемщики склонны практически полностью выбирать предоставленный лимит, в то время как заемщики с низким уровнем риска кредитуются гораздо ниже предоставленного лимита.

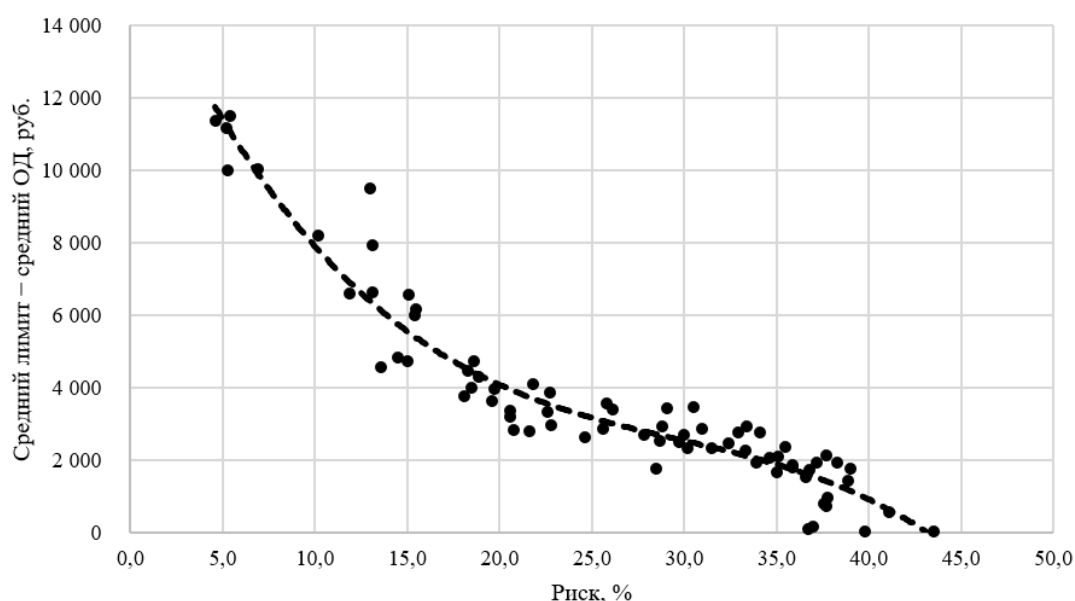


Рисунок 5.9 – График зависимости уровня риска (NPL30) и разницы между средним лимитом и средним ОД

Источник: составлено автором

Таким образом, мы не можем с помощью установки лимитов существенно влиять на ОД «хороших» заемщиков, однако можем значительно повлиять на «плохих». Далее проанализируем зависимость между доходностью и уровнем риска (см. рисунок 5.10).

Из представленного рисунка 5.10 видно, что целесообразно разделить всех заемщиков на три группы:

а) Сегмент заемщиков с низким уровнем риска и минимальной доходностью. Такие клиенты закрывают заем, как правило, досрочно, избегают штрафов, просрочек и отключают дополнительные услуги.

б) Сегмент со средним риском, самый прибыльный сегмент, несмотря на более высокий риск по сравнению с первым сегментом. Заемщики допускают просрочки, но тем не менее в большинстве успешно закрывают заем при этом оплачивая услугу за пролонгацию займа, штрафы за просрочку и проценты за более длительный срок займа.

в) Сегмент заемщиков с высоким риском и низкой доходностью за счет больших кредитных потерь.

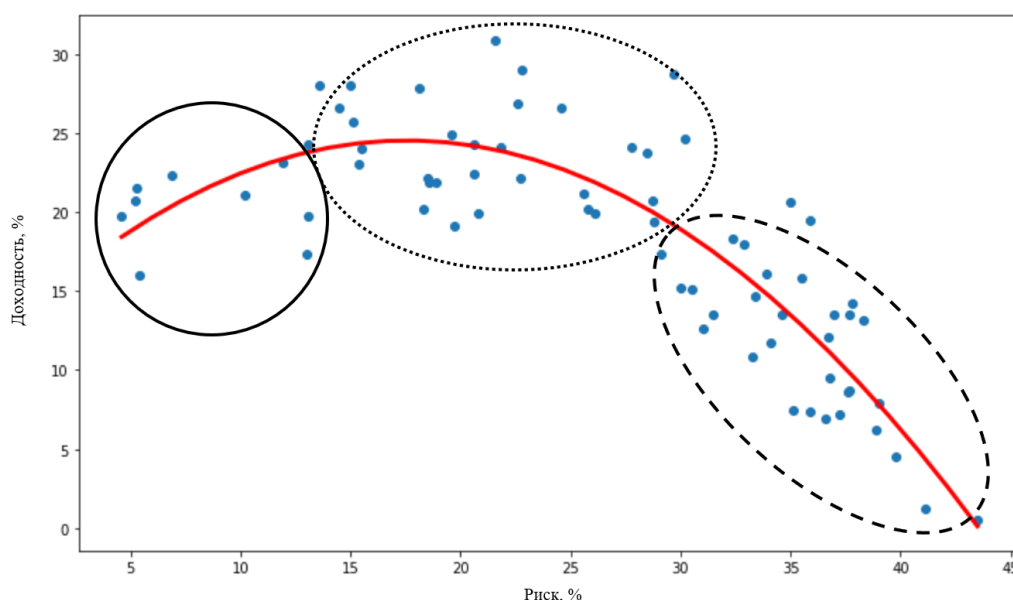


Рисунок 5.10 – График зависимости доходности (ROI) от уровня риска (NPL30)

Источник: составлено автором

Благодаря сегменту заемщиков со средним риском, которые склонны брать практически всю одобренную сумму займа, доходность в этом сегменте в среднем растет с ростом среднего ОД. Таким образом, задача разрабатываемой модели лимитов видится в том, чтобы в большей степени воздействовать на третий сегмент, повышая его доходность за счет ограничения лимитов пропорционально риску.

Далее построим тепловую диаграмму рассеяния, проанализировав зависимость фактического среднего ОД от риска. Приведем полученную диаграмму с наложением на нее графика оцененной полиномиальной регрессии по формуле (4.1) (рисунок 5.11).

Каждая точка имеет свою яркость, обозначающую уровень доходности. Параболическая форма графика обусловлена недоиспользованием своих лимитов заемщиками из 1 группы и ограничения лимитов со стороны МФО заемщикам из 3 группы.

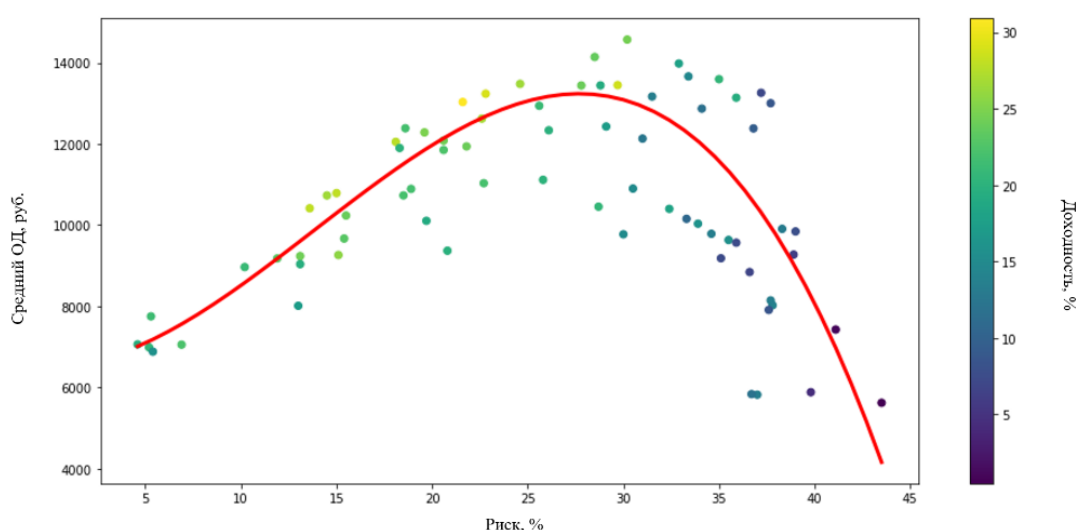


Рисунок 5.11 – Тепловая диаграмма рассеяния зависимости среднего ОД от NPL30 с наложением полиномиальной регрессии

Источник: составлено автором

На следующем шаге были найдены оценки параметров полиномиальной регрессии (4.1) на основе ВМНК по формуле (4.2). Итоговая сводка результатов расчета и код в программе Python приведена в Приложении Н. Полученную модель зависимости оптимального среднего займа от уровня риска можно записать следующим образом:

$$\hat{D} = 6514 + 26,3 \times R^2 - 0,63 \times R^3, \quad (5.4)$$

где $\hat{D}$ – оценка оптимального среднего ОД по модели.

Построенное уравнение регрессии получилось статистически значимым (F-статистика равна 93.9, $p < 0,001$) со всеми значимыми коэффициентами ($p < 0,001$) (см. таблицу 5.22).

Скорректированный коэффициент детерминации, в целом удовлетворительный и равен 0,72. Ввиду того, что в модели использовался взвешенный МНК, который априори дает высокую степень положительной автокорреляции ошибок модели, не следует обращать внимание на статистику Дарбина-Уотсона в приведенной сводки модели.

Таблица 5.22 – Оценка модели полиномиальной регрессии

Фактор	Оценка коэффициента	Стандартная ошибка оценки	t-статистика	p-значение
Константа	6 514,0	386,69	16,85	< 0,001
R^2	26,3	2,27	11,61	< 0,001
R^3	– 0,63	0,06	– 10,22	< 0,001

Источник: расчеты автора

Сглаженная линия на рисунке 5.11 представляет собой расчетные значения оптимального среднего ОД по формуле (4.6). Насыщенность точек на диаграмме рассеяния отражает уровень доходности: чем светлее точки, тем выше доходность. Таким образом, оптимальные лимиты находятся в районе сглаженной кривой. Следовательно, для тех групп заемщиков, которые практически полностью выбирают одобренный лимит, мы можем постараться снизить или повысить сумму среднего ОД в зависимости от того, находится ли данная группа соответственно выше или ниже рассчитанной кривой по нашей модели регрессии.

Заключительным этапом наших расчетов была оценка параметров логистической регрессии (4.7). Итоговая сводка результатов расчета в программе Python приведена на рисунках ПН.1 и ПН.2 в Приложении Н.

Построенная логистическая регрессия с константой имеет скорректированный коэффициент детерминации 0,69. В нашем случае константа оказалась незначимой, и была исключена из уравнения регрессии. Финальная

модель без константы получилась статистически значимой (F-статистика равна 314,2, $p < 0,001$) со значимыми коэффициентами ($p < 0,001$) таблица 5.23).

Таблица 5.23 – Оценка модели логистической регрессии

Фактор	Оценка коэффициента	Стандартная ошибка оценки	t-статистика	p-значение
R	8,106400	0,5770000	14,04	< 0,001
D	– 0,000068	0,0000149	– 4,59	< 0,001

Источник: составлено автором

Окончательная формула для расчета назначаемых лимитов согласно (4.7) после оценки параметров принимает вид:

$$\hat{L} = \frac{(1 + \exp(-(8,106 \times R - 0,000068 \times \hat{D})))}{\exp(-(8,106 \times R - 0,000068 \times \hat{D}))} \cdot \hat{D}. \quad (5.5)$$

Модель апробировалась для PDL-займов сроком до 30 дней и ограничением по сумме выдачи 30 тыс. руб. для новых клиентов МФО. После корректировки лимитов и «созревания» рисков по всем сегментам клиентов были получены приемлемые значения риска и доходности, незначительно отклоняющиеся от прогнозных значений.

На основе рассчитанной модели приведем пример ее использования на конкретном примере¹⁾. В целом задача определялась необходимостью повышения средней суммы займа (среднего чека) по первичным клиентам за счет увеличения лимитов во всех децилях на 2000–3000 руб. при сохранении текущего уровня риска (показателей NPL, FPD). Для проведения анализа использовались данные по займам первичных клиентов за 9 месяцев с 1 января по 1 октября.

В качестве целевой метрики использовался показатель NPL30, статус которого фиксировался на 9 января следующего года (период вызревания составил около 3 месяцев). По всем выданным займам были рассчитаны и сохранены значения скорингового балла по действующей в СПР МФО скоринговой модели: заемщики в 9 и 10 дециле получали отказ.

¹⁾ Данные скорректированы и изменены из-за конфиденциальности, но отражают реальную постановку задачи и ее решение для реальной МФО.

Все заемщики были разделены на два сегмента:

- а) клиенты, которые полностью взяли одобренный лимит;
- б) клиенты, которые взяли займ на сумму, меньшую одобренного лимита.

Такое деление было проведено в связи с базовой предпосылкой, что уровень риска у этих двух сегментов клиентов разный (таблица 5.24), и на основе этой разницы можно вывести формулу зависимости риска от изменения лимита. Клиенты, полностью утилизирующие предлагаемый скоринговой моделью лимит, обладают повышенным уровнем кредитного риска.

Таблица 5.24 – Уровень риска в зависимости от степени использования одобренного лимита

Степень использования одобренного лимита, %	NPL30, %
100	39
75–100	36
45–75	29
<45	19

Источник: составлено автором

Таким образом, из таблицы 5.24 видно, что с ростом использования (утилизации) лимита заемщиком растет кредитный риск.

Далее по каждому сегменту в разрезе децилей были рассчитаны и проанализированы следующие показатели: количество выданных займов, средняя сумма займа, одобренный лимит и NPL30.

На следующем этапе была проведена линейная экстраполяция риска относительно изменения суммы займа по следующей формуле:

$$\frac{NPL_{1,i} - NPL_{2,i}}{LoanSum_{1,i} - LoanSum_{2,i}} \cdot S, \quad (5.6)$$

где 1 и 2 – номер сегмента;

i – номер дециля риска;

NPL – уровень риска;

S – размер увеличения лимита;

LoanSum – средняя сумма займа.

Таким образом, были получены ожидаемые риски при увеличении лимита для 1-го сегмента. Для 2-го сегмента лимит был сохранен. Затем были рассчитаны взвешенные значения показателя $NPL30$ для двух сегментов. Далее по разработанной автором модели расчета оптимального лимита были получены и проанализированы новые рекомендуемые лимиты для каждого дециля риска в зависимости от ожидаемых рисков.

Объем общей выборки при анализе составил 57 498 займов. Объем 1-го и 2-го сегмента равен соответственно 35 267 и 22 228, а их соотношение – 61,3% и 38,7% соответственно (таблица 5.25).

Таким образом, в 1 сегменте показатель риска $NPL30$ во всех децилях выше. Далее, были найдены оптимальные размеры увеличения лимита для каждого из децилей. Расчеты представлены в таблицах 5.26 и 5.27 в столбце «Новые лимиты, руб.» полужирным шрифтом выделены лимиты, которые не превосходят оптимальные более чем на 2000 руб. в связи с исходной погрешностью модели.

Таблица 5.25 – Основные характеристики сегментов

Децили	1 и 2 сегмент		1 сегмент		2 сегмент	
	Лимит, руб. /Фактическая сумма займа, руб.	$NPL30$, шт. (%)	Лимит, руб.	$NPL30$, шт. (%)	Лимит, руб. /Фактическая сумма займа, руб.	$NPL30$, шт. (%)
1	17536 / 12308	6610 (19)	17404	3036 (24)	17649 / 7980	3574 (15)
2	14930 / 11530	7865 (28)	14873	4060 (32)	14991 / 7963	3805 (24)
3	12965 / 10574	8752 (32)	12934	4665 (35)	13000 / 7879	4086 (29)
4	11964 / 10051	9844 (36)	11937	5399 (38)	11996 / 7759	4445 (33)
5	11970 / 10147	7615 (37)	11948	4270 (40)	11996 / 7846	3345 (33)
6	8984 / 8148	7807 (39)	8979	6049 (40)	8997 / 5286	1758 (34)
7	5995 / 5752	8018 (39)	5994	6801 (40)	6000 / 4386	1215 (33)
8	4000 / 4000	987 (34)	4000	987 (34)	-	-

Источник: составлено автором

Таблица 5.26 – Ожидаемый уровень риска при увеличении лимита на 1000 руб. по разработанной модели назначения лимитов

Дециль	NPL30, %	Оптимальный средний ОД, руб.	Оптимальный лимит, руб.	Новые лимиты, руб.	Текущие лимиты, руб.
1	20,9	12263	17190	18536	17536
2	29,4	13237	16069	15930	14930
3	33,5	12355	14153	13965	12965
4	36,9	10702	11773	12964	11964
5	38,6	9499	10268	12970	11970
6	40,5	7810	8299	9984	8984
7	42,1	6105	6407	6995	5995

Источник: составлено автором

По новой модели предполагается выдача займов только в 1–7 дециле риска. Для остальных децилей повышение лимитов превышает оптимальное значение, предсказанное моделью более чем на 2000 руб., следовательно, лимиты следует повысить в первых 7 децилях (модель линейной регрессии прогнозирует оптимальный ОД и лимит, используя уровень риска и доходности). При этом при реализации такого предложения прогноз общего NPL30 по портфелю первичных клиентов должен увеличиться в среднем на 1,4 п. п. до 34,5%. Это повышение уровня кредитного риска не настолько значительно и критично, согласно юнит-экономике, при достижении решения задачи увеличения среднего чека.

Таблица 5.27 – Ожидаемый уровень риска при увеличении лимита на 2000 руб. по разработанной модели назначения лимитов

Дециль	NPL30, %	Оптимальный средний ОД, руб.	Оптимальный лимит, руб.	Новые лимиты, руб.	Текущие лимиты, руб.
1	21,6	12424	17220	19536	17536
2	30,1	13161	15806	16930	14930
3	34,3	12037	13644	14965	12965
4	37,5	10251	11195	13964	11964
5	39,6	8642	9254	13970	11970
6	42,0	6278	6595	10984	8984
7	45,5	1691	1740	7995	5995

Источник: составлено автором

Итоговые предложенные оптимальные лимиты на основе выполненных расчетов представлены в таблице 5.28. В целом задача повышения среднего чека при сохранении уровня кредитных потерь решена на примере реальных данных с использованием разработанной автором модели назначения лимитов в зависимости от уровня кредитного риска и оптимального среднего чека.

Таблица 5.28 – Сформулированные предложения по изменению лимитов

Дециль	Прогнозируемый NPL30, %	Оптимальный средний ОД, руб.	Оптимальный средний лимит, руб.	Увеличение лимита, руб.	Новый максимальный лимит, руб.
1	20	11 978	17 111	2 000	22 000
2	29	13 259	16 162	2 000	17 000
3	34	12 323	14 100	2 000	15 000
4	36	10 961	12 114	1 000	13 000
5	38	9 891	10 746	1 000	13 000
6	40	8 091	8 618	1 000	10 000
7	42	6 313	6 634	1 000	7 000

Источник: составлено автором

Апробация модели юнит-экономики для управления кредитными рисками проводилась автором в нескольких МФО, что подтверждено соответствующими справками о внедрении (Приложение И). Реальные расчеты не приводятся в данном параграфе из-за конфиденциальности данных и коммерческой тайны, поскольку показатели юнит-экономики полностью отражают всю бизнес-модель конкретной МФО. Однако в параграфе 2 главы 4 и Приложениях Г и Д были приведены все расчеты на гипотетических данных, приближенных к реальным показателям работы МФО для иллюстрации расчетов модели юнит-экономики, а также рассмотрены кейсы влияния на финансовый результат компании изменения процентной ставки и уровня кредитных потерь.

Для апробации, представленной в параграфе 3 главы 4 модели кредитного риска с изменяющимися во времени параметрами, мы не можем воспользоваться

историческими данными МФО, т. к. уже было отмечено выше столкнемся с проблемой «отказных заявок». Со временем по мере накопления исторических данных точность скоринговой модели для оценки кредитного риска падает из-за постоянного «отсева» заемщиков действующей кредитной политикой в МФО. Поэтому для апробации модели мы воспользуемся имитационными данными. Метод имитационного моделирования часто используется при оценке кредитных рисков [59]. Результаты апробации на основе имитационных данных представлены автором в работе [213].

Для простоты расчетов предположим, что на вероятность дефолта влияют три фундаментальных фактора риска. В качестве математической модели возьмем логистическую регрессию. В нашем имитационном эксперименте мы рассматриваем трехфакторную логит-модель вида:

$$p(1|X_t) = \frac{1}{1 + e^{-\beta_{t0} - \beta_{t1}x_{t1} - \beta_{t2}x_{t2} - \beta_{t3}x_{t3}}} , \quad (5.7)$$

где x_{tj} – независимые переменные ($j = 1, 2, 3$);

β_{tj} – вектор изменяющихся во времени параметров ($j = 0, 1, 2, 3$).

Для каждого параметра модели (5.7) генерируется авторегрессионный циклический процесс с различной периодичностью. Затем были сгенерированы выборки данных для каждого временного периода t в соответствии с моделью уравнения (5.7). Всего было сгенерировано 30 наборов данных (30 месяцев). Для тестирования предложенных методов автор использовал оценку максимального правдоподобия методом `generic`, который заключается в единственном построении модели, в целях подгонки значений параметров к сгенерированной выборке и функцию `auto.arima` в среде R для построения моделей класса ARIMA и прогнозирования параметров.

На рисунках 5.12–5.15 проиллюстрированы оценки параметров по методу максимального правдоподобия (отображены в виде кружков) и их прогнозные значения, полученные по модели ARIMA (отображены сплошной линией).

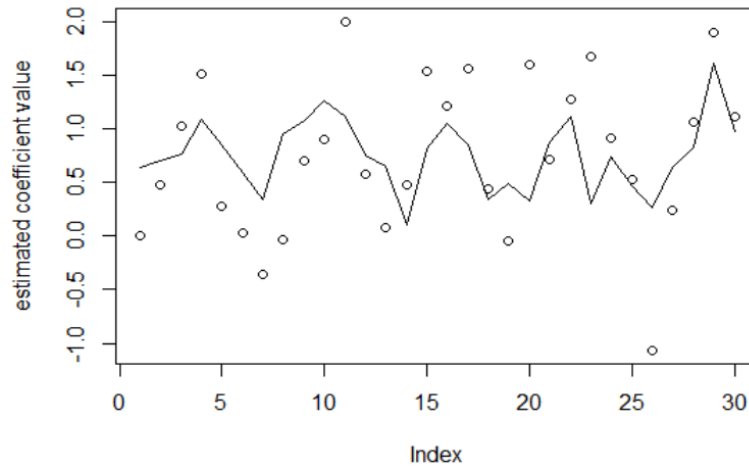


Рисунок 5.12 – Динамика коэффициента β_{t0}

Источник: составлено автором на основе расчетов в R

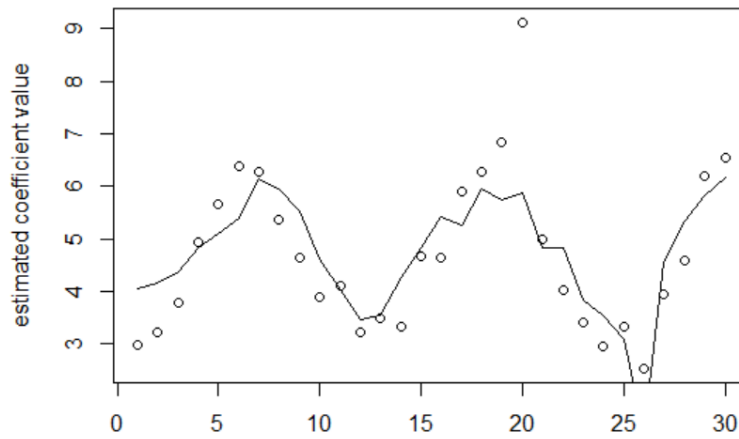


Рисунок 5.13 – Динамика коэффициента β_{t1}

Источник: составлено автором на основе расчетов в R

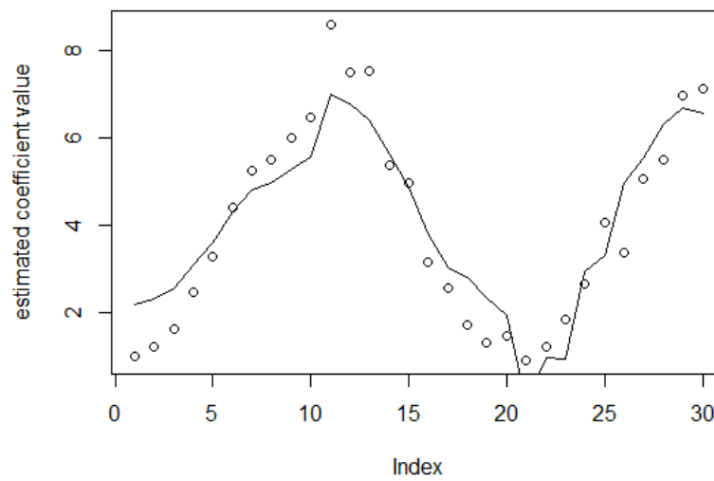


Рисунок 5.14 – Динамика коэффициента β_{t2}

Источник: составлено автором на основе расчетов в R

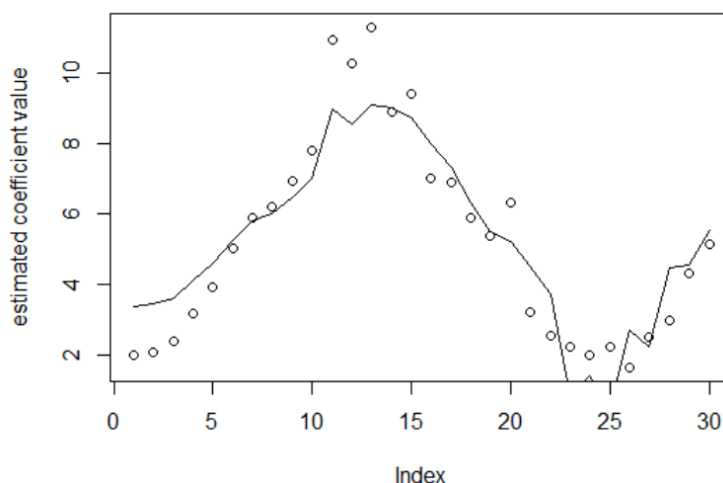


Рисунок 5.15 – Динамика коэффициента β_{t3}

Источник: составлено автором на основе расчетов в R

Чтобы доказать эффективность предложенной модели оценки кредитного риска с изменяющимися во времени параметрами, сравним результаты расчетов по ней с несколькими другими популярными методами классификации, рассмотренными в главе 3. Сравнения выполняются по производительности вне обучающей выборки по аналогии с работой [238] по следующим моделям:

- а) логистическая регрессия, где полученные оценки коэффициентов применялись в период $t-1$ к заемщикам, поступающим в период t ;
- б) алгоритм случайного леса из библиотеки RandomForest в пакете R с заданными по умолчанию настройками гиперпараметров;
- в) алгоритм градиентного бустинга из библиотеки gbm в пакете R с заданными по умолчанию настройками гиперпараметров;
- г) предложенная автором модель (4.46) с изменяющимися во времени параметрами, в которой значения параметров прогнозируются с помощью модели ARIMA на период t и применяются к заемщикам, приходящим на период t с оценкой параметров методом максимального правдоподобия [330];
- д) прогнозируемые коэффициенты с помощью комбинации подхода пространства состояний и модели DCC-GARCH (4.52) с использованием фильтра Калмана.

Условная процентная ставка принята на уровне 10% годовых. Рассмотренные методы тестируются на временном интервале из 30 периодов, каждый из которых включает 20 000 потенциальных заемщиков, занимаемых условно по одной денежной единице. Следует отметить, что в нашем имитационном эксперименте рассматривается условный пример, универсальный для банковского сектора или микрофинансирования, поэтому величина процентной ставки не имеет значения.

Полученная прибыль затем рассчитывается следующим образом:

$$profit = (nl - nd) \times r - nd, \quad (5.8)$$

где nl – количество выданных кредитов;

nd – количество просроченных кредитов;

r – процентная ставка.

Результаты расчетов на наших условных данных имитационного эксперимента для разных моделей представлены в таблице 5.29 и на рисунке 5.16

Таблица 5.29 – Сравнение результатов имитационного эксперимента

Метод	Полученная прибыль, у.е.	Количество выданных займов, шт.	Количество дефолтных займов, шт.	Процент дефолтных займов, %
Логит-регрессия	19 414	301 291	9 741	3,23
Случайный лес	19 627	301 232	9 542	3,17
Градиентный бустинг	19 735	301 168	9 438	3,13
ARIMA	24 538	300 904	5 048	1,68
Модель пространства состояний DCC-GARCH	28 532	300 831	1 405	0,47

Источник: составлено автором на основании расчетов в R

Как видно из таблицы 5.29, количество выданных кредитов практически одинаково для всех рассмотренных методов. Было гипотетически выдано немного меньше кредитов для модели с изменяющимися во времени параметрами, прогнозируемыми по ARIMA, и модели пространства-состояний с наложением

фильтра Калмана, чем для модели логит-регрессии и двух методов машинного обучения. Однако разница в количестве полученных плохих кредитов значительна: около девяти с половиной тысяч для логит-модели, случайного леса и градиентного бустинга и пяти тысяч для модели прогнозируемых коэффициентов ARIMA. При этом только полторы тысячи плохих кредитов было при использовании модели DCC-GARCH в пространстве состояний.

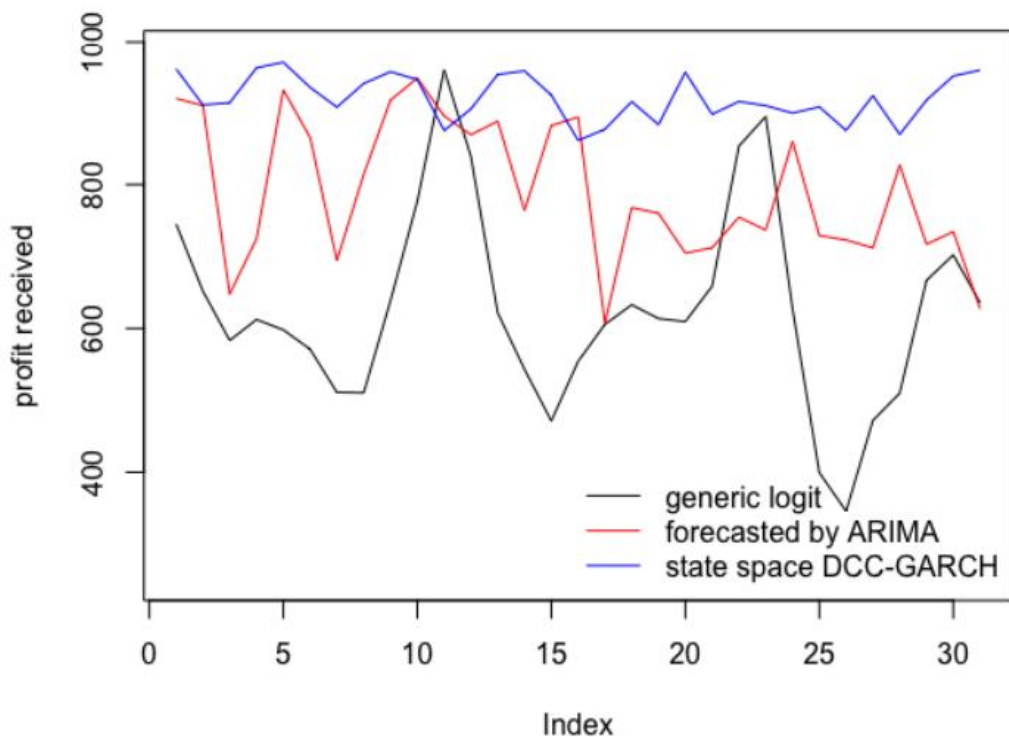


Рисунок 5.16 – Сумма полученной прибыли для протестированных методов по периодам времени

Источник: составлено автором на основании расчетов в R

Следовательно, процент просроченных кредитов для предлагаемых методов значительно ниже, что приводит к тому, что прибыль, полученная в соответствии с предлагаемыми методами, на 26% превышает прибыль при использовании для оценки кредитных рисков классической логистической регрессии по сравнению с моделью, прогнозирующие изменяющиеся во времени коэффициенты по модели

ARIMA, и на 47% для модели DCC–GARCH в пространстве состояний с фильтром Калмана.

Рисунок 5.16 наглядно иллюстрирует абсолютное превосходство предлагаемой модели с изменяющимся во времени параметрами над логистической регрессией почти для всех 30 рассмотренных периодов времени. Прибыль, полученная с помощью тривиальных методов, имеет отчетливую сезонность с резкими падениями, в то время как прибыль, получаемая предлагаемыми методами, остается стабильной с течением времени и, таким образом, может рассматриваться как доминирующая стратегия в управлении кредитными рисками МФО.

На рисунке 5.17 показана динамика эффективности методов машинного обучения по сравнению с предложенной моделью DCC–GARCH в пространстве состояний.

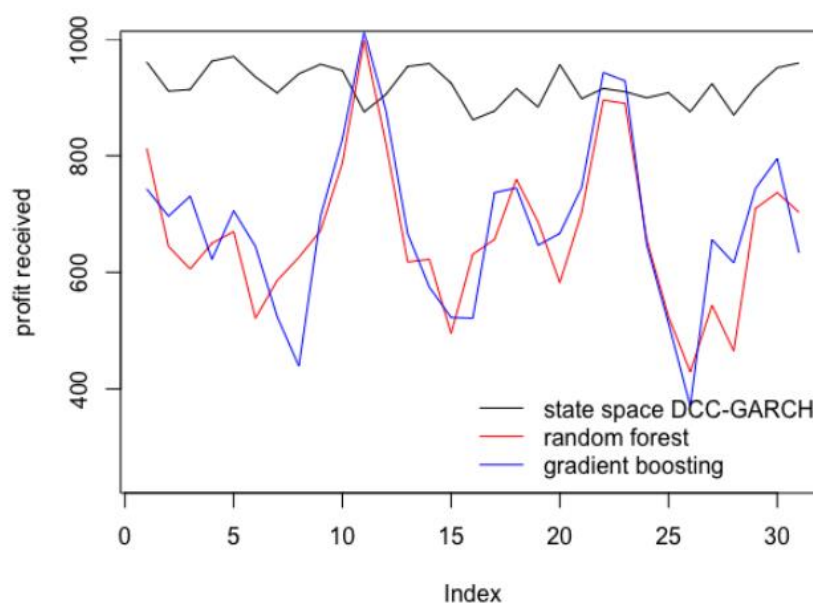


Рисунок 5.17 – Сумма полученной прибыли для протестированных методов по периодам времени

Источник: составлено автором на основании расчетов в R

Подходы к машинному обучению демонстрируют сходную картину с моделью логистической регрессии, хотя и несколько более высокую

эффективность. Метод градиентного бустинга с небольшим отрывом превосходит случайный лес. Предлагаемый автором метод использования модели оценки кредитными рисками с учетом изменения параметров во времени практически на всех рассмотренных периодах значительно превосходит проверенные методы машинного обучения и, таким образом, может рассматриваться как предпочтительная альтернатива.

В приложении II приведен код в программе R проведенных расчетов имитационного эксперимента.

Обобщая результаты проведенных экспериментов, можно констатировать, что в экономических условиях, когда нарушается предположение о постоянстве во времени истинных параметров моделей оценки кредитных рисков, более рационально применять оценки прогнозируемых коэффициентов вместо оценок, просто вычисленных по накопленному историческому набору данных, который уже отражает прошлые взаимосвязи внутри данных.

Таким образом, в данном диссертационном исследовании автором предложен новый метод работы с изменяющимися во времени параметрами скоринговых моделей, который направлен на вычисление вероятности дефолта заемщика. Эмпирически было показано, что в постоянно меняющейся экономической среде влияние факторов на целевую переменную также меняется. Прогнозирование коэффициентов логистической регрессии дает лучший финансовый результат, чем простое применение параметров, полученных на основе накопленной статистики за прошлые периоды времени.

В работе восполняется пробел в методах, включающих корректировку полученных коэффициентов, выполненные с помощью фильтра Калмана, поскольку мы получаем две независимые оценки: первую – путем прогнозирования вектора параметров, вторую – путем подгонки параметров к накопленной статистике. После проведения обширного имитационного эксперимента было показано, что предлагаемый метод значительно лучше традиционного подхода и, таким образом, может рассматриваться в качестве

предпочтительной альтернативы при работе с моделями оценки кредитного риска в нестабильной экономической среде.

Выводы по главе 5

В данной главе автором представлена альтернативная обобщенным скоринговым моделям методика оценки кредитного риска заемщика микрофинансовой организации на основе оценки платежеспособности заемщика по всем кредитам с учетом их типа, давности и платежной дисциплины – агрегированного показателя «качество кредитной истории», который может использоваться как отдельный показатель для ранжирования заемщиков по степени кредитного риска, так и в качестве предиктора в скоринговых моделях.

Проведена апробация разработанных моделей, алгоритмов и методик оценки и управления кредитными рисками на российских МФО, в частности успешное применение на ретроспективных данных российских МФО разработанных автором методик оценки ценности данных, а также моделей оценки кредитного риска – модели установления лимитов, модели юнит-экономики и модели оценки дефолта заемщика с изменяющимися во времени параметрами, что подтверждено соответствующими справками о внедрении в Приложении И.

Заключение

В рамках достижения поставленной цели настоящего диссертационного исследования – развития методологии оценки и управления кредитными рисками в микрофинансовых организациях на основе разработок и совершенствования экономико-математических моделей, учитывающих особенности микрофинансового кредитования, с учетом поставленных методологических задач были получены соответствующие результаты.

Исследование тенденций и закономерностей развития рынка микрофинансирования в России показало, что под воздействием нестабильности условий экономической среды функционирования микрофинансового рынка, а также ужесточения регуляторных требований к МФО на рынке остаются только наиболее крупные игроки, которые располагают более эффективными и гибкими системами управления кредитными рисками в сравнении с небольшими МФО и имеют возможность формировать портфель с упором на PDL-займы.

Проведенный кластерный анализ структуры рынка PDL-займов позволил разделить все анализируемые МФО на 2–3 основных сегмента. При этом было определено, что состав кластеров изменяется по периодам проведения анализа в зависимости от масштаба МФО и динамики ее роста за счет качества управления кредитным портфелем. Это позволило сделать вывод о том, что метод кластерного анализа адекватно отражает все конъюнктурные изменения на рынке PDL-займов и его результаты не противоречат результатам описательного анализа трендов российского микрофинансового рынка. При этом валидность предложенной методики построения агрегированных показателей для кластеризации доказана стабильностью результатов компонентного анализа во времени.

Рассмотрение особенностей деятельности МФО позволило выявить существенные отличия их от банков и соответствующие проблемы, связанные с финансовыми, информационными и временными ограничениями как в целом в рамках функционирования МФО, так и в контексте системы управления

кредитными рисками, которые приводят к необходимости выработки адекватных рыночным реалиям подходов, включающих разработку соответствующих регламентов и инструментов качественной оценки кредитных рисков в разрезе разных категорий заемщиков.

Решение задачи систематизации возможных вариантов исходной информации для оценки кредитных рисков МФО было достигнуто посредством выработки системного подхода, базой которого выступают два компонента: во-первых, классификация источников и видов традиционных и альтернативных данных о заемщиках МФО по различным характеристикам, в том числе поведенческим, во-вторых выбор критериев оптимальных проверок таких данных, включая соблюдение законодательных норм, идентификацию заемщиков и выявление мошеннических заявок. Применение системного подхода в практике оценки кредитных рисков играет значимую роль в эффективном функционировании МФО в условиях ограничений.

С целью обоснования критериев ценности исходной информации и выбора рационального состава данных для оценки кредитных рисков МФО была проведена систематизация целевых показателей и метрик, используемых для оценки эффективности новых данных, которые являются инструментом повышения уровня одобряемых кредитных заявок при одновременном снижении уровня кредитного риска. Метрики включают критерии силы влияния на показатели риска, критерии классификационной способности и критерии оценки экономического эффекта. Также была представлена пошаговая методика измерения ценности и отбора наиболее подходящих данных на основе расчета статистических тестов в контексте решения задачи оценки кредитных рисков, которая обеспечит возможность фильтровать состав данных по уровню рациональности в условиях ограниченности ресурсов МФО.

Для адекватной оценки классификационной способности моделей оценки кредитных рисков выработан алгоритм расчета и применения соответствующих метрик дискриминирующей способности, а также расчета соответствующего

статистического тестирования выбранных источников данных и построенных на их основе моделей оценки кредитных рисков.

Решение задачи выбора концепции статистического моделирования и/или методов и алгоритмов машинного обучения для оценки кредитных рисков МФО реализовано посредством сравнительного анализа, который показал, что методы статистического моделирования в большей степени реализуемы для оценки кредитных рисков в условиях имеющейся информационной инфраструктуры МФО и регуляторных требований по сравнению с алгоритмами машинного обучения ввиду достаточно высокой прозрачности и интерпретируемости моделей, а также их направленности на долгосрочное прогнозирование дефолта заемщиков. При этом определено, что несмотря на сложность построения и интерпретации моделей на основе алгоритмов машинного обучения, они обеспечивают лучшую производительность и эффективность классификации заемщиков в сравнении с традиционными статистическими методами.

В целом проведенный сравнительный анализ позволил сделать вывод, что эталонным методом в разработке моделей оценки кредитных рисков МФО остается логистическая регрессия, модификация которой под особенности функционирования МФО приведет к достижению более высокой производительности и эффективности классификации заемщиков в рамках разрабатываемой модели с одновременно более высоким уровнем ее интерпретируемости в соответствии с требованиями регулятора и в условиях ресурсной ограниченности МФО.

Проведено системное изложение методики построения модели оценки кредитных рисков МФО, включая подробное описание методических подходов к формированию выборки данных, отбору предикторов в модель, выбора зависимой переменной и приведения алгоритмов категоризации независимых переменных, а также диагностики мультиколлинеарности с учетом особенностей микрофинансировой деятельности. Данная методика обеспечит построение производительной, эффективной модели оценки кредитных рисков МФО с высокой классификационной и прогностической способностью.

Решение задачи комплексного применения результатов оценки кредитного риска на уровне заемщика реализовано путем разработки отдельной математической модели классификации заемщиков с последующим расчетом лимитов по займам в системе управления кредитными рисками МФО, новизна которой заключается в комплексном применении результатов оценки риска на уровне заемщика на основе скоринговой модели с получением его дециля риска, а также оценки кредитного поведения заемщика в конкретной МФО в рамках поведенческой модели установки лимитов для повторных заемщиков с последующим расчетом лимитов с целью оптимизации процедуры выдачи PDL-займов.

Для использования комплексного подхода в рамках системы принятия решений по управлению кредитными рисками МФО в качестве дополнения к традиционным скоринговым моделям разработана общая модель юнит-экономики, позволяющая оценивать финансовый результат МФО на всех этапах жизненного цикла заемщика. Разработанная модель сформирована на базе клиентской модели юнит-экономики, основой которой выступают метрики взаимодействия бизнеса МФО с заемщиком, обеспечивает возможность проведения оценки юнит-экономики первичного и повторного займа, а также общую юнит-экономику МФО – ценность клиента на основе специфических показателей деятельности МФО в рамках системы управления кредитными рисками и позволяет реализовать подход к моделированию кредитного риска и параметров соответствующих бизнес-метрик по всему кредитному портфелю МФО.

С целью возможности управления кредитными рисками микрофинансовых организаций с учетом изменчивости коэффициентов скоринговых моделей во времени разработана математическая модель оценки кредитных рисков МФО на базе логистической регрессии с корректировкой полученных коэффициентов фильтром Калмана для получения оценок как путем прогнозирования вектора изменения параметров, так и использованием их подгонки к накопленной статистике, что существенно повышает производительность стандартных

скоринговых моделей и позволяет получить более высокий финансовый результат МФО по сравнению с традиционным способом подгонки модели к набору данных.

Решение задачи оценки кредитных рисков МФО в условиях отсутствия исторических данных реализована посредством разработки альтернативной традиционным скоринговым моделям методики оценки кредитного риска заемщика МФО на основе оценки платежеспособности заемщика по всем кредитам с учетом их типа, давности и платежной дисциплины – агрегированного показателя «качество кредитной истории», который может использоваться как отдельный показатель для ранжирования заемщиков, так и в качестве предиктора в скоринговых моделях.

В заключительной части диссертационного исследования проведена комплексная апробация разработанных моделей, алгоритмов и методик оценки и управления кредитными рисками на российских МФО. Успешное применение на ретроспективных данных российских МФО разработанных автором методик оценки ценности данных, а также моделей оценки кредитного риска – модели установления лимитов, модели юнит-экономики и модели оценки дефолта заемщика с изменяющимися во времени параметрами, подтверждено соответствующими справками о внедрении.

Таким образом, полученные результаты проведенного в диссертационной работе исследования, свидетельствуют о полном достижении поставленной цели – развития методологии оценки и управления кредитными рисками в микрофинансовых организациях на основе разработок и совершенствования экономико-математических моделей, учитывающих особенности микрофинансового кредитования.

Список сокращений и условных обозначений

Условное обозначение	Наименование составной части
АКФ	автокорреляционная функция
БКИ	бюро кредитных историй
ВМНК	взвешенный метод наименьших квадратов
ГИБДД	Государственная инспекция безопасности дорожного движения
ГИС ГМП	Государственная информационная система о государственных и муниципальных платежах
ГК	главные компоненты
ЕСИА	единая система идентификации и аутентификации
ИКС	индекс кредитной сознательности
ИНН	идентификационный номер налогоплательщика
КИ	кредитная история
ККИ	качество кредитной истории
МКК	микрокредитная компания
ММП	метод максимального правдоподобия
МНК	метод наименьших квадратов
МПЛ	макропруденциальные лимиты
МФК	микрофинансовая компания
МФО	микрофинансовая организация
НБКИ	АО «Национальное бюро кредитных историй»
ОД	основной долг
ОКБ	АО «Объединенное кредитное бюро»
ПДН	предельная долговая нагрузка
ПСК	полная стоимость кредита
ПФР	пенсионный фонд России
РФ	Российская Федерация

СБ	АО «Скоринг Бюро»
СБП	система быстрых платежей
СМЭВ	система межведомственного электронного взаимодействия
СНИЛС	страховой номер индивидуального лицевого счета
СПР	система принятия решений
СРО	саморегулируемая организация
УПРИД	удаленная упрощенная идентификация клиентов
ФЗ	Федеральный закон
ФИО	фамилия, имя, отчество
ФНС	Федеральная налоговая служба
ФССП	Федеральная служба судебных приставов
ЦБ	Центральный банк России
ЧАКФ	частная автокорреляционная функция
ЭВМ	электронно-вычислительная машина
ADABOOST	Adaptive Boosting, адаптивный бустинг, алгоритм машинного обучения
AR	Approval Rate, уровень одобрения заявок на кредит
ARIMA	Autoregressive Integrated Moving Average, интегрированная модель авторегрессии-скользящего среднего
ARMA	Autoregressive Moving Average, модель авторегрессии-скользящего среднего
ARPC	Average Revenue Per Customer, общий доход, полученный от клиента
AUC	Area Under the Curve, площадь под кривой
CAC	Customer Acquisition Cost, средняя стоимость привлечения клиента
CIR	Cost to Income Ratio, отношение расходов на запрашиваемые данные к доходам от их применения
DC	Decision Cost, стоимость принятия одного решения
DCC-GARH	Dinamic Conditional Correlations-Generalized Autoregressive

	Conditional Heteroscedastic, обобщенная авторегрессионная модель гетероскедастичности с динамическими условными корреляциями
FN	False Negative, ложно-отрицательные наблюдения
FP	False Positive, ложно-положительные наблюдения
FPD	First Payment Default, дефолт первого платежа
GARH	Generalized Autoregressive Conditional Heteroscedastic, обобщенная авторегрессионная модель гетероскедастичности
GraphQL	Graph Query Language, протокол передачи данных
GRPC	Remote Procedure Calls, система удаленного вызова процедур получения данных с открытым исходным двоичным кодом
GUTDV	General Unified Transmission Data Format, единый формат формирования кредитной информации
IG	Information Gain, показатель прироста информации
IL	Installment Loan, «кредит в рассрочку», микрозайм со сроком более 30 дней
IV	Informative Value, информационное значение
L1	Сегмент первичных клиентов
L2+	Сегмент повторных клиентов от 2 и более займа
L3+	Сегмент повторных клиентов от 3 и более займа
LDA	Linear Discriminant Analysis, линейный дискриминантный анализ
LT	Life Time, срок «жизни» клиента
LTP	Life Time Profit, ожидаемая прибыль на протяжении срока «жизни» кредита или клиента
LTP	Lifetime Profit, ожидаемая прибыль на всем сроке «жизни» кредита
LTV	Life Time Value, ожидаемый доход на протяжении срока «жизни» кредита или клиента
MDLP	Minimal Description Length Principle, «принцип минимальной длины описания», алгоритм категоризации зависимой переменной

NPL	Non-Performing Loan, проблемный кредит
PDL	Pay Day Loans, «займ до зарплаты», микрозайм со сроком до 30 дней
POS	Point of Sales, «точка продаж», целевой микрозайм
QDA	Quadratic Discriminant Analysis, квадратичный дискриминантный анализ
QS	Quality Score, индекс качества модели
REST API	Representational State Transfer Application Programming Interface, протокол передачи данных
ROA	Return Of Assets, показатель рентабельности активов
ROI	Return on Investment, доходность
RR	Retantion Rate, показатель возврата клиентов в повторную заявку
RUTDF	формат передачи данных кредитных историй
SOAP XML	SOAP XML Simple Object Access Protocol eXtensible Markup Language, протокол передачи данных
SVP	SVP Support Vector Machine, метод опорных векторов
TN	True Negative, истинно-отрицательные наблюдения
TP	True Positive, истинно-положительная наблюдения
TR	Take Rate, конверсией из одобрения заявки на кредит в выдачу
TUTF	формат передачи данных кредитных историй
VIF	Variance Inflation Factor, показатель увеличения дисперсии из-за мультиколлинеарности
XGBoost	eXtreme Gradient Boosting, экстремальный градиентный бустинг, алгоритм машинного обучения
WOE	Weight of Evidence, веса категорий независимой переменной

Список литературы

1. Адуенко, А. А. Выбор признаков и шаговая логистическая регрессия для задачи кредитного скоринга / А. А. Адуенко // Машинное обучение и анализ данных. – 2012. – Т. 1. – № 3. – С. 279–291.
2. Айвазян, С. А. Прикладная статистика и основы эконометрики / С. А. Айвазян, В. С. Мхитарян. – М. : ЮНИТИ, 1998. – 1030 с.
3. Алеников, А. С. Сущностные аспекты юнит-экономики и ее применение в проектной деятельности / А. С. Алеников // Экономический анализ: теория и практика. – 2023. – Т. 22. – № 5 (536). – С. 933–953.
4. Анализ данных : учебник для вузов / В. С. Мхитарян, М. Ю. Архипова, Т. А. Дуброва; под ред. В. С. Мхитаряна. – М. : Юрайт, 2025. – 448 с.
5. Анализ современного состояния и тенденций развития микрокредитных организаций на макро- и мезо- уровнях / И. Е. Устюгова, Е. А. Беляева, Ю. Н. Воронцова, Е. С. Стряпчих // Экономика: вчера, сегодня, завтра. – 2024. – Т. 14. – № 3А. – С. 287–298.
6. Ароян, М. В. Особенности деятельности микрофинансовых организаций в России / М. В. Ароян // Тенденции развития науки и образования. – 2023. – № 104–5. – С. 30–34.
7. Базовый стандарт по управлению рисками микрофинансовых организаций (утв. Банком России, Протокол от 27.07.2017 № КФНП-26) [Электронный ресурс] // Банк России : [официальный сайт]. – Режим доступа: <https://cbr.ru/Queries/XsltBlock/File/90010/14> (дата обращения: 13.04.2025).
8. Байрамов, О. Б. Алгоритмизация процесса микрофинансирования и методы оптимизации / О. Б. Байрамов // Управление развитием крупномасштабных систем (MLSD'2021) : труды Четырнадцатой международной конференции (Москва, 27–29 сентября 2021 г.). – М. : Институт проблем управления им. В. А. Трапезникова РАН, 2021. – С. 724–730.
9. Бамбаева, Н. Я. Применение законов распределения случайных величин для моделирования экономических явлений и процессов : монография / Н. Я. Бамбаева, А. С. Сорокин. – М. : МЭСИ, 2010. – 154 с.

10. Банк России : официальный сайт. – Режим доступа: <https://cbr.ru/> (дата обращения: 13.04.2025).
11. Банк России установил более строгие ограничения на кредитование заемщиков с высокой долговой нагрузкой, а также повысил макропруденциальные надбавки по нецелевым потребительским кредитам с залогом транспортного средства [Электронный ресурс] // Банк России : [официальный сайт]. – Режим доступа: <https://cbr.ru/press/pr/?file=638606197703353528DSD.htm> (дата обращения: 21.04.2025).
12. Богданова, М. М. Микрофинансовые организации: современное положение на финансовом рынке / М. М. Богданова, П. А. Косенкова // Экономика, предпринимательство и право. – 2023. – Т. 13. – № 5. – С. 1703–1716.
13. Боровко, Ю. Г. Возможности применения модели юнит-экономики к анализу деятельности субъектов малого предпринимательства и обоснованию целесообразности диверсификации бизнеса / Ю. Г. Боровко, Н. С. Ермашкевич // Вектор экономики. – 2019. – № 11 (41). – С. 92.
14. Бурова, А. Б. Применение модели вероятности дефолта для оценки прогнозируемого кредитного риска / А. Б. Бурова, Г. И. Пеникас, С. В. Попова // Деньги и кредит. – 2021. – Т. 80. – № 3. – С. 49–72.
15. Васильева, Е. В. Воронка онлайн-продаж как аналитический инструмент управления эффективностью бизнеса / Е. В. Васильева, В. В. Лосев // Управление. – 2019. – № 3. – С. 63–74.
16. Вахорина, М. В. Практика применения системы показателей юнит-экономики с целью масштабирования бизнеса / М. В. Вахорина // Научные исследования и разработки. Экономика. – 2024. – Т. 12. – № 1. – С. 41–44.
17. Вахорина, М. В. Юнит-экономика как инструмент управления бизнес-рисками / М. В. Вахорина // Научные исследования и разработки. Экономика. – 2023. – Т. 11. – № 3. – С. 46–49.

18. Вахрушина, М. А. Юнит-экономика как инструмент современного управленческого учета / М. А. Вахрушина // Экономический анализ: теория и практика. – 2022. – Т. 21. – № 5 (524). – С. 972–990.
19. Ветров, А. В. Классификация и методика оценки рисков в интегрированных структурах / А. В. Ветров // Проблемы экономики и юридической практики. – 2020. – Т. 16. – № 5. – С. 142–147.
20. Власенко, М. Определение кредитоспособности розничных клиентов на основе алгоритмов машинного обучения / М. Власенко // Банкаўскі веснік. – 2022. – № 11 (712). – С. 30–44.
21. Волкова, Е. С. Современные подходы к применению методов интеллектуального анализа данных в задаче кредитного скоринга / Е. С. Волкова, В. Б. Гисин, В. И. Соловьев // Финансы и кредит. – 2017. – Т. 23. – № 34 (754). – С. 2044–2060.
22. Гаврилова, В. А. Юнит-экономика как метод экономического анализа / В. А. Гаврилова // Фундаментальные основы инновационного развития науки и образования : сборник статей IX Международной научно-практической конференции (Пенза, 25 декабря 2020 г.). – Пенза : Наука и Просвещение, 2020. – С. 91–93.
23. Государственный реестр бюро кредитных историй [Электронный ресурс] // Банк России : [официальный сайт]. – 2024. – Режим доступа: https://cbr.ru/ckki/gosreestr_ckki/ (дата обращения: 11.04.2025).
24. Гребенников, П. И. Макроэкономика в 2 т. Т. 1 : учебник и практикум для вузов / П. И. Гребенников, Л. С. Тарасевич, А. И. Леусский. – 11-е изд., перераб. и доп. – М. : Юрайт, 2021. – 300 с.
25. Гребенников, П. И. Экономика / П. И. Гребенников, Л. С. Тарасевич. – 6-е изд., перераб. и доп. – М. : Юрайт, 2025. – 348 с.
26. Гришин, А. А. Разработка модели поведенческого скоринга с использованием методов градиентного бустинга / А. А. Гришин, С. П. Строев // Научно–технический вестник Поволжья. – 2018. – № 9. – С. 93–98.

27. Губарева, Е. А. Сравнение эффективности методов минимизации нулевого и первого порядка в нейронных сетях / Е. А. Губарева, С. И. Хашин, Е. С. Шемякова // Вестник университета. – 2022. – № 11. – С. 48–55.
28. Данилишин, А. Р. Риск-нейтральная динамика для модели ARIMA-GARCH с ошибками, распределенными по закону Су Джонсона / А. Р. Данилишин, Д. Ю. Голембиовский // Информатика и ее применения. – 2020. – Т. 14. – № 1. – С. 48–55.
29. Данные аналитических отчетов СРО «МиР» [Электронный ресурс] // Саморегулируемая организация «МиР» : [сайт]. – Режим доступа: <https://npmir.ru/vzaimodeystvie-s-regulyatorami/> (дата обращения: 15.04.2025).
30. Демешев, Б. Б. Прогнозирование банкротства российских компаний: межотраслевое сравнение / Б. Б. Демешев, А. С. Тихонова // Экономический журнал ВШЭ. – 2014. – Т. 18. – № 3. – С. 359–386.
31. Долговая нагрузка граждан снизилась: итоги микрофинансовой отрасли за 2023 год [Электронный ресурс] // Банк России : [официальный сайт]. – Режим доступа: <https://cbr.ru/press/event/?id=18576> (дата обращения: 13.04.2025).
32. Дуброва, Т. А. Прогнозирование социально-экономических процессов : учебное пособие / Т. А. Дуброва. – М. : Market DS, 2007. – 192 с.
33. Дулесов, А. С. Определение энтропии как меры информации при сопоставлении прогнозных и фактических показателей предприятия / А. С. Дулесов, В. И. Хрусталева // Современные проблемы науки и образования. – 2012. – № 1. – С. 151.
34. Дьяков, О. А. Особенности применения методов Data Mining в скоринговых решениях для коммерческих банков / О. А. Дьяков // Научные записки молодых исследователей. – 2017. – № 3. – С. 5–11.
35. Егоров, В. В. Моделирование динамики двухфакторных социально-экономических состояний посредством отображений, близких к растяжению / В. В. Егоров // Вестник университета. – 2022. – № 8. – С. 104–110.

36. Живайкина, А. Д. Кредитные рейтинги российских банков и отзывы банковских лицензий 2012-2016 гг. / А. Д. Живайкина, А. А. Пересецкий // Журнал Новой экономической ассоциации. – 2017. – Т. 4. – № 36. – С. 49–80.
37. Звездина, Н. В. Исследование потребительских предпочтений на Московском рынке смартфонов / Н. В. Звездина, А. С. Сорокин // Вопросы статистики. – 2017. – № 7. – С. 41–51.
38. Звездина, Н. В. Моделирование потребительских предпочтений на московском рынке смартфонов методом совместного анализа / Н. В. Звездина, А. С. Сорокин // Вопросы статистики. – 2018. – Т. 25. – № 12. – С. 28–39.
39. Ибрагимова, А. Ф. Влияние банка России на развитие микрофинансовых организаций / А. Ф. Ибрагимова, Н. И. Морозко // Самоуправление. – 2023. – № 2 (135). – С. 573–576.
40. Иванов, В. Г. Unit-экономика российских политических партий: политический консьюмеризм и идеологический фактор / В. Г. Иванов // Вестник Российского университета дружбы народов. Серия: Государственное и муниципальное управление. – 2021. – Т. 8. – № 3. – С. 223–228.
41. Исаев, Д. В. Динамическое ансамблевое обучение для оценки кредитоспособности / Д. В. Исаев // Инновации и инвестиции. – 2022. – № 3. – С. 74–79.
42. Исаев, Д. В. Оценка кредитного скоринга на основе карточных транзакций / Д. В. Исаев // Инновации и инвестиции. – 2021. – № 5. – С. 105–109.
43. Квартальные отчеты за 2018–2024 гг. [Электронный ресурс] // БКИ «Скоринг бюро» : [сайт]. – Режим доступа: https://scoring.ru/legal_persons/info_sources/?source=kvartalnye-otchety (дата обращения: 14.04.2025).
44. Кислицин, В. Е. Имитационное моделирование процесса кредитования физических лиц с использованием кредитного рейтинга / В. Е. Кислицин // Вестник Воронежского государственного университета. Серия экономика и управление. – 2018. – № 3. – С. 112–118.

45. Ковалев, М. Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц / М. Ковалев, В. Корженевская // Банки Казахстана. – 2008. – № 1. – С. 43–48.
46. Колмогоров, А. Н. О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения / А. Н. Колмогоров // Доклады Академии наук СССР. – 1957. – Т. 114. – № 5. – С. 953–956.
47. Коротаев, А. В. Безработица как предиктор социально-политической дестабилизации в странах Западной и Восточной Европы / А. В. Коротаев, А. А. Хохлова, С. В. Цирель // Экономическая социология. – 2018. – Т. 19. – № 2. – С. 118–166.
48. Крамаренко, И. В. Особенности применения регрессии Кокса в различных инструментальных средах / И. В. Крамаренко, Л. А. Константинова // Вестник университета. – 2022. – № 10. – С. 80–88.
49. Кредитные отношения в современной экономике : монография / О. И. Лаврушин, Р. К. Нурмухаметов, Ю. И. Меликов [и др.]; под ред. О. И. Лаврушина, Е. В. Травкиной. – М. : КноРус, 2024. – 352 с.
50. Кренева, С. Г. UNIT-экономика как инструмент принятия решений / С. Г. Кренева, Т. А. Лежнина // Инновационное развитие экономики. – 2019. – № 6 (54). – С. 120–129.
51. Кузнецов, Л. А. Оценка кредитной истории физических лиц на основе нечетких моделей / Л. А. Кузнецов, А. В. Перевозчиков // Финансы и кредит. – 2008. – № 14 (302). – С. 19–26.
52. Кузнецова, В. Ю. Информационная технология принятия решений в микрофинансовой организации / В. Ю. Кузнецова // Системная инженерия и информационные технологии. – 2023. – Т. 5. – № 3 (12). – С. 27–41.
53. Лукашин, Ю. П. Статистические методы изучения фондового рынка / Ю. П. Лукашин // Вопросы статистики. – 1995. – № 7. – С. 14–21.
54. Магнус, Я. Р. Эконометрика. Начальный курс / Я. Р. Магнус, П. К. Катыхев, А. А. Пересецкий. – 5-е изд., испр. – М. : Дело, 2001. – 400 с.

55. Мадера, А. Г. Прогнозирование кредитной благонадежности заемщика / А. Г. Мадера // Финансы и кредит. – 2013. – № 12 (540). – С. 2–10.
56. Макконнелл, К. Р. Экономикс: принципы, проблемы и политика : учебник / К. Р. Макконнелл, С. Л. Брю, Ш. М. Флинн. – 21-е изд. – М. : Инфра-М, 2022. – 1152 с.
57. Макушина, Е. Ю. Моделирование вероятности банкротства российских нефинансовых компаний / Е. Ю. Макушина, И. А. Шихлярова // Финансы и кредит. – 2018. – Т. 24. – № 1 (769). – С. 95–110.
58. Мастяева, И. Н. Алгоритм вычисления оптимального кредитного лимита для заемщика для случая экспресс-кредитования / И. Н. Мастяева, Е. Г. Воловатова // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6. – № 2. – С. 157–167.
59. Мишин, Ю. В. Прогнозирование результатов деятельности кредитных учреждений на базе имитационного моделирования / Ю. В. Мишин, А. В. Оглоблина // Микроэкономика. – 2021. – № 3. – С. 5–14.
60. Монтеррейский консенсус Международной конференции по финансированию развития [Электронный ресурс] // Организация Объединенных Наций : [официальный сайт]. – Режим доступа: https://www.un.org/ru/documents/decl_conv/declarations/monterrey.shtml (дата обращения: 15.04.2025).
61. Мусин, А. Р. Возможности построения экономико-математических моделей динамики финансового рынка с использованием компонент локального уровня / А. Р. Мусин, А. С. Сорокин // Вестник Тверского государственного университета. Серия: Прикладная математика. – 2018. – № 4. – С. 76–86.
62. Мусин, А. Р. Применение математического аппарата гидрогазодинамики для моделирования динамики временных рядов финансового рынка / А. Р. Мусин, А. С. Сорокин // Ученые записки Казанского университета. Серия: Физико-математические науки. – 2018. – Т. 160. – № 4. – С. 709–717.

63. О задаче калибровки и взвешивания данных выборочных обследований большого объема / В. С. Жаромский, В. В. Сопцов, А. М. Рудберг, С. А. Тер-Акопов // Народонаселение. – 2011. – № 3 (53). – С. 69–83.

64. О кредитных историях : Федеральный закон от 30.12.2004 № 218-ФЗ [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_51043/ (дата обращения: 15.04.2025).

65. О микрофинансовой деятельности и микрофинансовых организациях : Федеральный закон от 02.07.2010 № 151-ФЗ [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_102112/ (дата обращения: 15.04.2025).

66. О персональных данных : Федеральный закон от 27.07.2006 № 152-ФЗ [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 15.04.2025).

67. О порядке формирования микрофинансовыми организациями резервов на возможные потери по займам : Указание Банка России от 20.01.2020 № 5391-У [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_347191/ (дата обращения: 13.04.2025).

68. О потребительском кредите (займе) : Федеральный закон от 21.12.2013 № 353-ФЗ [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_155986/ (дата обращения: 15.04.2025).

69. О рекомендациях Базельского комитета по банковскому надзору «Принципы совершенствования корпоративного управления : Письмо Банка России от 06.02.2012 №14-Т [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_126196/ (дата обращения: 07.04.2025).

70. О саморегулируемых организациях в сфере финансового рынка : Федеральный закон от 13.07.2015 № 223-ФЗ [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_182662/ (дата обращения: 15.04.2025).

71. О требованиях к системе управления рисками и капиталом кредитной организации и банковской группы : Указание Банка России от 15.04.2015 № 3624-У [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_180268/ (дата обращения: 07.04.2025).

72. Обзор ключевых показателей микрофинансовых институтов [Электронный ресурс] // Банк России : [официальный сайт]. – Режим доступа: https://cbr.ru/analytics/microfinance/reveiw_mfo/ (дата обращения: 18.04.2025).

73. Пеникас, Г. И. Оценка эффективности макропруденциальной политики Банка России по ограничению необеспеченного потребительского кредитования модифицированным методом разность разностей / Г. И. Пеникас // Финансы и бизнес. – 2021. – Т. 17. – № 2. – С. 3–35.

74. Пересецкий, А. А. Экономические методы в дистанционном анализе деятельности российских банков / А. А. Пересецкий. – М. : ИД ВШЭ, 2012. – 235 с.

75. Перечень поручений по итогам совещаний по экономическим вопросам (утв. Президентом РФ 29.09.2024 № Пр-1999) [Электронный ресурс] // КонсультантПлюс : [сайт]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_486954/ (дата обращения: 13.04.2025).

76. Перова, В. И. Анализ влияния сферы здравоохранения на развитие человеческого капитала как фактора экономического роста стран ЕАЭС / В. И. Перова, П. В. Корчемный // Вестник Нижегородского университета им. Н. И. Лобачевского. – 2021. – № 1 (61). – С. 22–32.

77. Перспективные направления развития рынка микрофинансовых организаций на 2025–2027 годы : доклад для общественных консультаций [Электронный ресурс] // Банк России : [официальный сайт]. – 2024. – Режим доступа:

https://cbr.ru/Content/Document/File/165304/Consultation_Paper_19082024.pdf (дата обращения: 10.04.2025).

78. Поляков, К. Л. Опыт моделирования вероятности кредитного дефолта клиентов микрофинансовых организаций (на примере одной МФО) / К. Л. Поляков, Л. В. Жукова // Экономический журнал ВШЭ. – 2019. – Т. 23. – № 4. – С. 497–523.

79. Применение ROC-анализа для оценки функциональных характеристик цифровых детекторов / Д. И. Галкин, Е. А. Иванайский, А. С. Сорокин, А. Е. Шубочкин // SIBTEST – 2019 : сборник тезисов докладов V Международной конференции по инновациям в неразрушающем контроле (Екатеринбург, 26-28 июня 2019 г.). – Томск : НИ ТПУ, 2019. – С. 19.

80. Программа для ЭВМ № RU 2023612592. Greeder of Data - автоматический советник для построения скоринговых моделей на базе логистической регрессии на основе больших массивов данных : № 2022686296 : заявл. 26.12.2022 : опубл. 06.02.2023 / автор и правообладатель А. С. Сорокин.

81. Программа для ЭВМ № RU 2023682976. Качество кредитной истории : программа для вычисления рейтинга заемщика по его кредитной истории для оценки кредитных рисков и построения скоринговых моделей : № 2023682030 : заявл. 24.10.2023 : опубл. 01.11.2023 / автор и правообладатель А. С. Сорокин.

82. Противодействие нелегальной деятельности на финансовом рынке [Электронный ресурс] // Банк России : [официальный сайт]. – Режим доступа: https://cbr.ru/analytics/inside/2024_1/ (дата обращения: 13.04.2025).

83. Рейтинг регионов по закредитованности населения – 2024 // РИА Рейтинг : [сайт]. – Режим доступа: <https://riarating.ru/infografika/20240311/630259198.html> (дата обращения: 16.04.2025).

84. Ретроспективный статистический анализ рынка труда на постсоветском пространстве / А. П. Цыпин, М. М. Шайлиева, А. С. Сорокин, И. Б. Хмелев // Вестник Евразийской науки. – 2019. – Т. 11. – № 6. – С. 1–8.
85. Рзун, И. Г. Проактивное развитие организации: Unit-экономика в разрезе клиентоориентированного подхода / И. Г. Рзун, О. С. Щербакова, Я. А. Гриневич // Вестник Академии знаний. – 2022. – № 3 (50). – С. 265–269.
86. Розинская, Н. А. От количественных к цифровым методам исследований в экономической истории / Н. А. Розинская, А. С. Сорокин // Journal of Institutional Studies (Журнал институциональных исследований). – 2019. – Т. 11. – № 1. – С. 115–128.
87. Рубцова, Н. В. Микрофинансирование в Российской Федерации: изменение отраслевых показателей в условиях глобальных вызовов / Н. В. Рубцова // Baikal Research Journal. – 2024. – Т. 15. – № 1. – С. 13–24.
88. Рэнкинги микрофинансовых организаций [Электронный ресурс] // Рейтинговое агентство «Эксперт РА» : [сайт]. – Режим доступа: <https://raexpert.ru/rankings/mfi/> (дата обращения: 15.04.2025).
89. Сиддики, Н. Скоринговые карты для оценки кредитных рисков. Разработка и внедрение интеллектуальных методов кредитного скоринга / Н. Сиддики; пер. с англ. Е. Ильичева. – М. : Манн, Иванов и Фербер, 2014. – 268 с.
90. Смелов, С. Б. Перспективы развития рынка микрофинансирования в кризис / С. Б. Смелов // Московский экономический журнал. – 2023. – Т. 8. – № 1. – С. 520–531.
91. Смирнов, Н. В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках / Н. В. Смирнов. – М. : МГУ, 1939. – 16 с.
92. Снегова, Е. Г. Модель управления лимитом кредитования при заданном уровне потерь / Е. Г. Снегова // Экономика, статистика и информатика. Вестник УМО. – 2013. – Т. 9. – № 5. – С. 181–184.
93. Созыкина, А. И. Юнит-экономика: оценка жизнеспособности бизнес-модели / А. И. Созыкина, Е. Г. Сухих // Умные технологии в современном мире :

материалы V Всероссийской научно-практической конференции (Челябинск, 22-23 ноября 2022 г.). – Челябинск : ИЦ ЮУрГУ, 2022. – Т. 1. – С. 123–128.

94. Сорокин, А. С. Анализ изменения структуры рынка PDL-займов в России за 2019–2023 гг. на основе многомерного статистического анализа / А. С. Сорокин // Мягкие измерения и вычисления. – 2024. – Т. 84. – № 11–1. – С. 112–124.

95. Сорокин, А. С. Измерение кредитного риска заемщика микрофинансовой организации на основе агрегированного показателя просрочки / А. С. Сорокин // Мягкие измерения и вычисления. – 2023. – Т. 72. – № 11–2. – С. 5–23.

96. Сорокин, А. С. Использование агрегированного признака качества кредитной истории для оценки закредитованности заемщика на российском рынке микрофинансирования / А. С. Сорокин // Измерение и анализ благосостояния : тезисы докладов Всероссийской (с международным участием) научно-практической конференции (Санкт-Петербург, 25-27 января 2024 г.). – СПб. : СПбГЭУ, 2024. – С. 336–339.

97. Сорокин, А. С. Исследование зависимости доходов региональных бюджетов Республики Казахстан от развития малого бизнеса [Электронный ресурс] / А. С. Сорокин // Интернет-журнал «Науковедение». – 2015. – Т. 7. – № 2. – С. 1–23. – Режим доступа: <http://naukovedenie.ru/PDF/147EVN215.pdf> (дата обращения: 14.04.2025).

98. Сорокин, А. С. К вопросу валидации модели логистической регрессии в кредитном скоринге [Электронный ресурс] / А. С. Сорокин // Интернет-журнал «Науковедение». – 2014. – № 2. – С. 1–14. – Режим доступа: <https://naukovedenie.ru/PDF/173EVN214.pdf> (дата обращения: 14.04.2025).

99. Сорокин, А. С. К вопросу применения фильтра Калмана в эконометрических моделях / А. С. Сорокин, А. Р. Мусин // Научно-аналитический журнал Наука и практика Российского экономического университета им. Г. В. Плеханова. – 2017. – № 1 (25). – С. 71–76.

100. Сорокин, А. С. Количественная оценка информативности радиографического контроля с помощью ROC-анализа / А. С. Сорокин, Д. И. Галкин, Е. А. Иванайский // Контроль. Диагностика. – 2019. – № 5. – С. 4–12.
101. Сорокин, А. С. Методика анализа ценности данных для построения скоринговых моделей кредитного риска в микрофинансовых организациях / А. С. Сорокин // Социальные и экономические системы. – 2023. – № 6–1 (49). – С. 171–199.
102. Сорокин, А. С. Многомерный статистический анализ структуры рынка микрофинансирования в России [Электронный ресурс] / А. С. Сорокин, В. А. Шилов // Интернет-журнал «Науковедение». – 2016. – Т. 8. – № 1. – С. 1–18. – Режим доступа: <http://naukovedenie.ru/PDF/10EVN116.pdf> (дата обращения: 14.04.2025).
103. Сорокин, А. С. Моделирование оптимальных кредитных лимитов в микрофинансовых организациях / А. С. Сорокин // Экономический журнал ВШЭ. – 2022. – Т. 26. – № 2. – С. 285–306.
104. Сорокин, А. С. Модель кредитного риска на основе логистической регрессии с изменяющимися во времени параметрами / А. С. Сорокин // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками : материалы XII Международной научно-практической конференции (Саратов, 16–18 ноября 2023 г.). – Саратов : СГУ, 2023. – № 8. – С. 141–146.
105. Сорокин, А. С. Модель юнит-экономики для управления кредитными рисками в микрофинансовых компаниях / А. С. Сорокин // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками : материалы XIII Международной научно-практической конференции (Саратов, 14–16 ноября 2024 г.). – Саратов : СГУ, 2024. – № 9. – С. 230–236.
106. Сорокин, А. С. Особенности выбора источников данных для оценки кредитного риска в микрофинансовых организациях / А. С. Сорокин // Мягкие измерения и вычисления. – 2023. – Т. 71. – № 10–2. – С. 92–104.

107. Сорокин, А. С. Оценка кредитных рисков с помощью скоринговой модели с изменяющимися во времени параметрами / А. С. Сорокин // Дайджест-финансы. – 2024. – Т. 29. – № 4 (272). – С. 396–420.

108. Сорокин, А. С. Планирование и статистическая обработка данных экспериментов в пакете R / А. С. Сорокин, Н. А. Моисеев, В. И. Митрофанов // Интеллект. Инновации. Инвестиции. – 2017. – № 1. – С. 58–64.

109. Сорокин, А. С. Построение скоринговых карт с использованием модели логистической регрессии [Электронный ресурс] / А. С. Сорокин // Интернет-журнал «Науковедение». – 2014. – № 2. – С. 1–29. – Режим доступа: <https://naukovedenie.ru/PDF/180EVN214.pdf> (дата обращения: 14.04.2025).

110. Сорокин, А. С. Разработка алгоритмов машинного обучения скоринговых моделей на основе различных источников данных о заемщике / А. С. Сорокин, Н. А. Моисеев // Плехановский научный бюллетень. – 2019. – № 2 (16). – С. 157–165.

111. Сорокин, А. С. Разработка алгоритмов применения моделей интеллектуального анализа данных для управления кредитными рисками микрофинансовых организаций / А. С. Сорокин // Плехановский научный бюллетень. – 2022. – № 2 (22). – С. 99–108.

112. Сорокин, А. С. Разработка алгоритмов применения моделей интеллектуального анализа данных для управления рисками в налоговом и таможенном администрировании / А. С. Сорокин, Н. А. Моисеев // Плехановский научный бюллетень. – 2021. – № 2 (20). – С. 77–82.

113. Сорокин, А. С. Разработка методики построения статистических моделей оценки рисков в соответствии со стандартом МСФО 9 / А. С. Сорокин, Н. А. Моисеев // Плехановский научный бюллетень. – 2019. – № 1 (15). – С. 138–143.

114. Сорокин, А. С. Сравнительный анализ использования статистического моделирования и машинного обучения для оценки кредитного риска в микрофинансовых компаниях / А. С. Сорокин // Экономический вестник. – 2024. – Т. 3. – № 2. – С. 51–65.

115. Сорокин, А. С. Тенденции и закономерности развития рынка микрофинансирования в России / А. С. Сорокин // Социальные и экономические системы. – 2024. – № 7 (57). – С. 237–262.

116. Сорокин, А. С. Тенденции развития российского рынка данных для оценки кредитных рисков / А. С. Сорокин // Экономическое развитие России: вызовы и возможности в меняющемся мире : материалы Международной научно-практической конференции (Краснодар, 24–27 января). – Краснодар : КубГУ, 2023. – Т. 4. – С. 332–337.

117. Сорокин, А. С. Управление кредитными рисками в микрофинансовой организации на основе модели юнит-экономики / А. С. Сорокин // Финансовая аналитика: проблемы и решения. – 2025 – № 1. – С. 101–124.

118. Сорокин, С. В. Использование нейросетевых моделей в поведенческом скоринге / С. В. Сорокин, А. С. Сорокин // Прикладная информатика. – 2015. – Т. 10. – № 2 (56). – С. 92–109.

119. Статистическая оценка результатов расшифровки радиографических снимков сварных соединений / Н. П. Алешин, Д. И. Галкин, А. С. Сорокин, О. И. Колесников // Сварка и диагностика. – 2015. – № 1. – С. 11–14.

120. Статистические данные к обзору ключевых показателей микрофинансовых институтов [Электронный ресурс] // Банк России : [официальный сайт]. – Режим доступа: <https://www.cbr.ru/microfinance/statistics/> (дата обращения: 18.04.2025).

121. Суглобов, А. Е. Развитие и проблемы цифровизации банковских услуг / А. Е. Суглобов, А. В. Минаков // Modern Economy Success. – 2021. – № 4. – С. 88–102.

122. Тихомиров, Н. П. Верификация прогнозов на основе анализа рисков их ошибок / Н. П. Тихомиров, Д. А. Максимов, А. В. Щербаков // Вестник Российского экономического университета им. Г. В. Плеханова. – 2011. – № 3 (39). – С. 103–113.

123. Тихомиров, Н. П. Методы и многомерного статистического анализа : учебник для вузов / Н. П. Тихомиров, Т. М. Тихомирова, О. С. Ушмаев. – М. : Экономика, 2011. – 647 с.

124. Тихомиров, Н. П. Методы оценки риска снижения эффективности инвестиционного проекта в условиях статистической неопределенности / Н. П. Тихомиров, Т. М. Тихомирова // Вестник Российского экономического университета им. Г. В. Плеханова. – 2024. – Т. 21. – № 2 (134). – С. 28–35.

125. Тихомиров, Н. П. Методы оценки эффективности инвестиционных проектов в реальном секторе экономики в условиях неопределенности исходной информации / Н. П. Тихомиров // Вестник Российского экономического университета им. Г. В. Плеханова. – 2011. – № 6 (42). – С. 86–94.

126. Тихомиров, Н. П. Методы прогнозной оценки критерия NPV инвестиционного проекта при неопределенности исходной информации / Н. П. Тихомиров // Экономика природопользования. – 2011. – № 6. – С. 3–13.

127. Тихомиров, Н. П. Теория риска : учебное пособие / Н. П. Тихомиров, Т. М. Тихомирова. – М. : Юнити-Дана, 2020. – 308 с.

128. Тихомиров, Н. П. Эконометрика : учебник / Н. П. Тихомиров, Е. Ю. Дорохина. – М. : Экзамен, 2003. – 512 с.

129. Травкина, Е. В. Устойчивость российского банковского сектора: тенденции и риски / Е. В. Травкина // Интеллект. Инновации. Инвестиции. – 2021. – № 2. – С. 10–19.

130. Трейман, М. Г. Исследование особенностей управления бизнес-процессами и внедрение принципов юнит-экономики в деятельность компаний / М. Г. Трейман // Научный журнал НИУ ИТМО. Серия: Экономика и экологический менеджмент. – 2024. – № 1. – С. 69–75.

131. Тюрин, В. В. Дискриминантный анализ в биологии : монография / В. В. Тюрин, С. Н. Щеглов. – Краснодар : КубГУ, 2015. – 126 с.

132. Фалеев, А. В. Современное состояние и перспективы рынка микрофинансирования России / А. В. Фалеев, С. В. Масалкина // Столыпинский вестник. – 2023. – Т. 5. – № 5. – С. 2332–2342.

133. Френкель, А. А. Корреляционный и регрессионный анализ в экономических приложениях : учебное пособие / А. А. Френкель, Е. В. Адамова. – М. : МЭСИ, 1987. – 96 с.
134. Хабибулин, А. А. Применение интернет-технологий в формировании кредитного рейтинга физических лиц / А. А. Хабибулин, О. Б. Хабибулина, Э. И. Булатова // Инновации и инвестиции. – 2020. – № 1. – С. 178–180.
135. Хайкин, С. Нейронные сети: полный курс / С. Хайкин; пер. с. англ. – 2-е изд. – М. : Вильямс, 2006. – 1104 с.
136. Хенд, Д. Анализ отклоненных заявок в кредитных операциях: теория и практика / Д. Хенд // Руководство по кредитному скорингу / под ред. Э. Мэйз; пер. с англ. И. М. Тикота. – Минск : Гревцов Пабlishер, 2008. – С. 295–314.
137. Цыпин, А. П. Информационные технологии при проведении экономико-статистических исследований на основе исторических временных рядов / А. П. Цыпин, А. С. Сорокин // Азимут научных исследований: экономика и управление. – 2017. – Т. 6. – № 2 (19). – С. 299–302.
138. Цыпин, А. П. Статистические пакеты программ в социально-экономических исследованиях / А. П. Цыпин, А. С. Сорокин // Азимут научных исследований: экономика и управление. – 2016. – Т. 5. – № 4 (17). – С. 379–384.
139. Четыркин, Е. М. Статистические методы прогнозирования / Е. М. Четыркин. – 2-е изд., перераб. и доп. – М. : Статистика, 1977. – 199 с.
140. Шакер, Н. С. Регулирование рисков деятельности микрофинансовых организаций в Российской Федерации : монография / Н. С. Шакер, И. Е. Шакер. – М. : КноРус, 2021. – 179 с.
141. Шаповалов, В. Ю. Современные тенденции развития микрофинансирования в России / В. Ю. Шаповалов, Н. Е. Соловьева // Журнал прикладных исследований. – 2024. – № 7. – С. 124–130.
142. Широбокова, М. А. Модель оценки риска дефолта на всем протяжении жизни кредита / М. А. Широбокова // Вестник Удмуртского университета. Серия: Экономика и право. – 2018. – Т. 28. – № 2. – С. 228–233.

143. Широбокова, М. А. Применение случайного леса выживаемости к динамической оценке кредитного риска / М. А. Широбокова, А. В. Летчиков // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками. – 2019. – № 4. – С. 113–118.

144. Яблонская, А. Е. Социальный скоринг как перспективная банковская технология в системе потребительского кредитования / А. Е. Яблонская, М. Ю. Костыкова // Российское предпринимательство. – 2018. – № 4. – С. 1143–1150.

145. A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring / D. Wang, Z. Zhang, R. Bai, Y. Mao // Journal of Computational and Applied Mathematics. – 2018. – Vol. 329. – P. 307–321.

146. A logistic regression model for consumer default risk / E. C. E. Silva, C. Lopes, A. Correia, S. Faria // Journal of Applied Statistics. – 2020. – Vol. 47. – № 13–15. – P. 2879–2894.

147. A Machine Learning Approach for Micro-Credit Scoring / A. Ampountolas, T. N. Nde, P. Date, C. Constantinescu // Risks. – 2021. – Vol. 9. – № 3. – Art. 50. – P. 1–20.

148. A multi-objective approach for profit-driven feature selection in credit scoring / N. Kozodoi, S. Lessmann, K. Papakonstantinou [et al.] // Decision Support Systems. – 2019. – Vol. 120. – P. 106–117.

149. A new hybrid credit scoring ensemble model with feature enhancement and soft voting weight optimization / D. Yang, B. Xiao, M. Cao, H. Shen // Expert Systems with Applications. – 2024. – Vol. 238. – № 2. – P. 12–21.

150. Abdou, H. A. Credit scoring, statistical techniques and evaluation criteria: A review of the literature / H. A. Abdou, J. Pointon // Intelligent Systems in Accounting Finance & Management. – 2011. – Vol. 18. – № 2–3. – P. 59–88.

151. Abel, A. B. Macroeconomics / A. B. Abel, B. S. Bernanke, D. Croushore. – 11-th ed. – London : Pearson, 2023. – 641 p.

152. Abraham, B. Statistical methods of forecasting / B. Abraham, J. Ledolter. – New York : Wiley, 1983. – 472 p.

153. Agresti, A. An Introduction to categorical data analysis / A. Agresti. – New York : Wiley, 1996. – 320 p.
154. Agresti, A. Categorical data analysis / A. Agresti. – New York : Wiley, 1990. – 584 p.
155. Al-Aradi, A. Credit Scoring via Logistic Regression [Электронный ресурс] / A. Al-Aradi // Department of Statistical Sciences, University of Toronto, Canada. – 2014. – Режим доступа: <https://pdf4pro.com/view/credit-scoring-via-logistic-regression-6e2e16.html> (дата обращения: 14.04.2025).
156. Albanesi, S. Predicting Consumer Default: A Deep Learning Approach: Working Paper № 26165 [Электронный ресурс] / S. Albanesi, D. F. Vamossy // National Bureau of Economic Research : [сайт]. – 2019. – Режим доступа: https://www.nber.org/system/files/working_papers/w26165/w26165.pdf (дата обращения: 14.04.2025).
157. Albert, A. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models / A. Albert, J. A. Anderson // Biometrika. – 1984. – Vol. 71. – № 1. – P. 1–10.
158. Allison, P. D. Logistic regression using the SAS system: theory and application / P. D. Allison. – Cary : SAS Institute, 1999. – 303 p.
159. Amzile, K. Assessment of Support Vector Machine performance for default prediction and credit rating / K. Amzile, M. Habachi // Banks and Bank Systems. – 2022. – Vol. 17. – № 1. – P. 161–175.
160. Analysis of Credit Scoring Using Particle Swarm Optimization Algorithm under Logistic Regression Model [Электронный ресурс] / U. Rahmani, S. Riaman, S. Subiyanto, A. T. Bon // Proceedings of the International Conference on Industrial Engineering and Operations Management Pilsen. – 2019. – Режим доступа: <https://ieomsociety.org/pilsen2019/papers/340.pdf> (дата обращения: 14.04.2025).
161. Anderberg, M. R. Cluster analysis for applications / M. R. Anderberg. – New York : Academic Press, 1973. – 384 p.

162. Anderson, R. A. Credit Intelligence & Modelling: Many Paths through the Forest of Credit Rating and Scoring / R. A. Anderson. – Oxford : Oxford University Press, 2022. – 944 p.

163. Anderson, R. Piecewise Logistic Regression: an Application in Credit Scoring [Электронный ресурс] / R. Anderson // Credit Scoring and Control Conference XIV, Edinburgh, 26-28 August 2015. – Режим доступа: https://www.researchgate.net/publication/305701103_Piecewise_Logistic_Regression_and_Application_in_Credit_Scoring (дата обращения: 14.04.2025).

164. Anderson, R. The credit scoring toolkit: theory and practice for retail credit risk management and decision automation / R. Anderson. – Oxford : Oxford University press, 2007. – 790 p.

165. Anderson, T. W. An Introduction to Multivariate Statistical Analysis / T. W. Anderson. – 2-nd ed. – New York : Wiley, 1984. – 710 p.

166. Anderson, T. W. Introduction to multivariate statistical analysis / T. W. Anderson. – New York : Wiley, 1958. – 396 p.

167. Angelini, E. A neural network approach for credit risk evaluation / E. Angelini, G. Di Tollo, A. Roli // The Quarterly Review of Economics and Finance. – 2008. – Vol. 48. – № 4. – P. 733–755.

168. Aniceto, M. Machine learning predictivity applied to consumer creditworthiness / M. Aniceto, F. Barboza, H. Kimura // Future Business Journal. – 2020. – Vol. 6. – № 1. – Art. 37. – P. 1–14.

169. Ash, D. Best practices in reject inferencing / D. Ash, S. Meester // Conference on Credit Risk Modeling and Decisioning. – Philadelphia : Wharton Financial Institutions Center, 2002. – P. 35–43.

170. Aslan, A. Are sustainable companies more likely to default? Evidence from the dynamics between credit and ESG ratings / A. Aslan, L. Poppe, P. Posch // Sustainability. – 2021. – Vol. 13. – № 15. – Art. 8568. – P. 1–16.

171. Astebro, T. Bound and collapse Bayesian reject inference for credit scoring / T. Astebro, G. Chen // Journal of the Operational Research Society. – 2012. – Vol. 63. – № 10. – P. 1374–1387.

172. Banasik, J. Does reject inference really improve the performance of application scoring models? / J. Banasik, J. Crook // *Journal of Banking & Finance*. – 2004. – Vol. 28. – № 4. – P. 857–874.

173. Bansal, G. Tuning data mining methods for cost-sensitive regression: A study in loan charge-off forecasting / G. Bansal, A. P. Sinha, H. Zhao // *Journal of Management Information Systems*. – 2008. – Vol. 25. – P. 315–336.

174. Barkova, I. Sample selection bias in acquisition credit scoring models: an evaluation of the supplemental-data approach / I. Barkova, D. Glennon, A. Palvia // *Journal of Credit Risk*. – 2013. – Vol. 9. – № 3. – P. 77–117.

175. Belsley, D. A. Regression diagnostics: Identifying influential data and sources of collinearity / D. A. Belsley, E. Kuh, R. E. Welsch. – New York : Wiley, 1980. – 320 p.

176. Bennouna, G. Scoring in microfinance: credit risk management tool – Case of Morocco / G. Bennouna, M. Tkiouat // *Procedia Computer Science*. – 2019. – Vol. 148. – P. 522–531.

177. Biggs, D. A method of choosing multiway partitions for classification and decision trees / D. Biggs, B. De Ville, E. Suen // *Journal of Applied Statistics*. – 1991. – Vol. 18. – № 1. – P. 49–62.

178. Bijak, K. Kalman filtering as a performance monitoring technique for a propensity scorecard / K. Bijak // *Journal of the Operational Research Society*. – 2011. – Vol. 62. – № 1. – P. 29–37.

179. Bishop, C. M. *Neural Networks for Pattern Recognition* / C. M. Bishop. – 3rd ed. – Oxford : Oxford University Press, 1995. – 512 p.

180. Bishop, Y. M. *Discrete multivariate analysis: Theory and practice* / Y. M. Bishop, S. E. Feinberg, P. W. Holland. – Cambridge : MIT Press, 1975. – 568 p.

181. Bitto, A. Achieving shrinkage in a time-varying parameter model framework / A. Bitto, S. Frühwirth-Schnatter // *Journal of Economic Literature*. – 2019. – Vol. 210. – P. 75–97.

182. Bloomfield, P. *Fourier analysis of time series: an introduction* / P. Bloomfield. – New York : Wiley, 1976. – 284 p.

183. Boosting Algorithms as Gradient Descent in Function Space / L. Mason, J. Baxter, P. Bartlett, M. Frean // *Advances in Neural Information Processing System*. – 1999. – Vol. 8. – P. 479–485.
184. Boser, B. E. A training algorithm for optimal margin classifiers / B. E. Boser, I. Guyon, V. N. Vapnik // *Proceedings of the Fifth Annual Workshop on Computational Learning Theory – COLT '92*. – Pittsburgh : ACM Press, 1992. – P. 144–152.
185. Bouaguel, W. On Feature Selection for Credit Scoring [Электронный ресурс] / W. Bouaguel // *ResearchGate* : [сайт]. – 2015. – Режим доступа: https://www.researchgate.net/publication/271513838_On_Feature_Selection_for_Credit_Scoring (дата обращения: 13.04.2025).
186. Boughaci, D. Three local search-based methods for feature selection in credit scoring / D. Boughaci, A. Alkhawaldeh // *Vietnam Journal of Computer Science*. – 2018. – Vol. 5. – № 36. – P. 107–121.
187. Bowker, A. H. A test for symmetry in contingency tables / A. H. Bowker // *Journal of the American Statistical Association*. – 1948. – Vol. 43. – № 244. – P. 572–574.
188. Box, G. An analysis of transformations (with discussion) / G. Box, D. Cox // *Journal of the Royal Statistical Society. Series B (Methodological)*. – 1964. – Vol. 26. – № 2. – P. 211–252.
189. Box, G. E. P. *Time Series Analysis: Forecasting and Control* / G. E. P. Box, G. M. Jenkins. – 3rd ed. – Englewood Cliffs : Prentice Hall, 1994. – 598 p.
190. Breiman, L. Random Forests / L. Breiman // *Machine Learn.* – 2001. – Vol. 45. – № 1. – P. 5–32.
191. Breslow, N. E. Covariance analysis of censored survival data / N. E. Breslow // *Biometrics*. – 1974. – Vol. 30. – № 1. – P. 89–99.
192. Brockwell, P. J. *Time Series: Theory and Methods* / P. J. Brockwell, R. A. Davis. – 2nd ed. – New York : Springer-Verlag, 1991. – 596 p.

193. Brown, M. B. Sampling behavior of tests for correlation in two-way contingency tables / M. B. Brown, J. K. Benedetti // *Journal of the American Statistical Association*. – 1977. – Vol. 72. – P. 309–315.

194. Bücker, M. Reject inference in consumer credit scoring with nonignorable missing data / M. Bücker, M. van Kampen, W. Krämer // *Journal of Banking and Finance*. – 2013. – Vol. 37. – P. 1040–1045.

195. Cain, K. C. Approximate case influence for the proportional hazards regression model with censored data / K. C. Cain, N. T. Lange // *Biometrics*. – 1984. – Vol. 40. – № 1. – P. 493–499.

196. Carol, A. Static and dynamic models for multivariate distribution forecasts: Proper scoring rule tests of factor–quantile versus multivariate GARCH models / A. Carol, H. Yang, M. Xiaochun // *International Journal of Forecasting*. – 2022. – Vol. 39. – № 3. – P. 1078–1096.

197. Chamboko, R. Advanced Survival Modelling for Consumer Credit Risk Assessment: Addressing recurrent events, multiple outcomes and frailty : Doctoral Thesis [Электронный ресурс] / R. Chamboko // NOVA Information Management School : [сайт]. – Режим доступа: <http://hdl.handle.net/10362/87634> (дата обращения: 12.11.2024).

198. Chan, J. C. Bayesian model comparison for time-varying parameter VARs with stochastic volatility / J. C. Chan, E. Eisenstat // *Journal of Applied Economics*. – 2018. – Vol. 33. – P. 509–532.

199. Cheney, J. S. Alternative Data and its Use in Credit Scoring Thin- and No-File Consumers. FRB of Philadelphia – Payment Cards Center Discussion Paper, 2008 – № 08-01 [Электронный ресурс] / J. S. Cheney // Social Science Research Network : [сайт]. – Режим доступа: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1160283 (дата обращения: 12.04.2025).

200. Chikalipah, S. Credit risk in microfinance industry: Evidence from sub-Saharan Africa / S. Chikalipah // *Review of Development Finance*. – 2018. – Vol. 8. – № 1. – P. 38–48.

201. Classification and Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – New York : Chapman & Hall/CRC, 1984. – 368 p.
202. Cochran, W. G. Some methods for strengthening the common chi-square tests / W. G. Cochran // Biometrics. – 1954. – Vol. 10. – P. 417–451.
203. Cohen, J. Risk and gambling: the study of subjective probability / J. Cohen, M. Hansel. – London : Philosophical Library, 1956. – 153 p.
204. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance / Y. Zhu, C. Xie, G. J. Wang, X. G. Yan // Neural Computing and Applications. – 2017. – Vol. 28. – P. 41–50.
205. Cook, R. D. Detection of influential observations in linear regression / R. D. Cook // Technometrics. – 1977. – Vol. 19. – P. 15–18.
206. Cooley, W. W. Multivariate data analysis / W. W. Cooley, P. R. Lohnes. – New York : Wiley, 1971. – 392 p.
207. Coombs, C. H. Components of risk in decision making: Probability and variance preferences / C. H. Coombs, D. G. Pruitt // Journal of Experimental Psychology. – 1960. – Vol. 60. – № 5. – P. 265–277.
208. Costangioară, A. Consumer Credit Scoring / A. Costangioară // Romanian Journal of Economic Forecasting. – 2011. – №. 3. – P. 162–177.
209. Cox, D. R. Regression models and life tables / D. R. Cox // Journal of the Royal Statistical Society: Series B (Methodological). – 1972. – Vol. 34. – № 2. – P. 187–202.
210. Cox, D. The Analysis of Binary Data / D. Cox, E. Snell. – 2-nd ed. – London : Chapman and Hall, 1989. – 160 p.
211. Credit rating analysis with support vector machines and neural networks: A market comparative study / Z. Huang, H. Chen, C. J. Hsu [et al.] // Decision Support Systems. – 2004. – Vol. 37. – № 4. – P. 543–558.
212. Credit risk Evaluation by hybrid data mining technique / W. Chen, G. Xiang, Y. Liu, K. Wang // Systems Engineering Procedia. – 2012. – Vol. 3. – P. 194–200.

213. Credit risk theoretical model on the base of DCC-GARCH in time-varying parameters framework / N. Moiseev, A. Sorokin, N. Zvezdina [et al.] // *Mathematics*. – 2021. – Vol. 9. – № 19. – P. 1–12.
214. Credit Scoring in Microfinance Using Non-traditional Data / S. Ruiz, P. Gomez, L. Rodrigues, J. Gama // *Lecture Notes in Computer Science*. – 2017. – Vol. 10423. – P. 447–458.
215. Credit Scoring Model for Retail Banking Sector in Pakistan / A. Hussain, M. Khan, S. U. Rehman, A. Khattak // *Journal of Managerial Sciences*. – 2019. – Vol. 14. – № 4. – P. 153–161.
216. Credit Scoring Model Implementation in a Microfinance Context / A. Terko, E. Zunic, D. Donko, A. Dzelihodzic // *XXVII International Conference on Information, Communication and Automation Technologies*. – New York : Springer, 2019. – P. 38–44.
217. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru / A. Blanco, R. Pino-Mejias, J. Lara-Rubio, S. S. Rayo // *Expert Systems with Applications*. – 2013. – Vol. 40. – № 1. – P. 356–364.
218. Darroch, J. N. Generalized iterative scaling for log-linear models / J. N. Darroch, D. Ratcliff // *Annals of Mathematical Statistics*. – 1972. – Vol. 43. – P. 1470–1480.
219. Davis, R. H. Machine-learning algorithms for credit-card applications / R. H. Davis, D. B. Edelman, A. J. Gammerman // *IMA Journal of Management Mathematics*. – 1992. – Vol. 4. – № 1. – P. 43–51.
220. Davison, A. C. Bootstrap Methods and their Application / A. C. Davison, D. V. Hinkley. – Cambridge : Cambridge University Press, 2006. – 594 p.
221. Dempster, A. P. Elements of Continuous Multivariate Analysis / A. P. Dempster. – Reading : Addison-Wesley, 1969. – 408 p.
222. Dey, S. Modeling the Combined Effects of Credit Limit Management and Pricing Actions on Profitability of Credit Card Operations / S. Dey // *International Journal of Business and Management*. – 2010. – Vol. 5. – № 4. – P. 168–177.

223. Dietterich, T. G. Machine-learning research / T. G. Dietterich // *AI Magazine*. – 1997. – Vol. 18. – № 4. – P. 97–136.
224. Discretization: An Enabling Technique / H. Liu, F. Hussain, C. L. Tan, M. Dash // *Data Mining and Knowledge Discovery*. – 2002. – Vol. 6. – № 4. – P. 393–423.
225. Dixon, W. J. BMD Biomedical computer programs / W. J. Dixon. – Los Angeles : University of California Press, 1971. – 662 p.
226. Dougherty, J. Supervised and unsupervised discretization of continuous features / J. Dougherty, R. Kohavi, M. Sahami // *Proceedings of the Twelfth International Conference on Machine Learning*. – Los Altos : Morgan Kaufmann, 1995. – P. 194–202.
227. Durand, D. Risk Elements in Consumer Instalment Financing, *Studies in Consumer Instalment Financing* / D. Durand. – New York : National Bureau of Economic Research, 1941. – 163 p.
228. Dynamic Effectiveness of Random Forest Algorithm in Financial Credit Risk Management for Improving Output Accuracy and Loan Classification Prediction / Sh. Kuyoro, O. Ogunyolu, T. Ayanwola, F. Y. Ayankoya // *Ingénierie des Systèmes d'Information*. – 2022. – Vol. 27. – № 5. – P. 815–821.
229. Dziuban, C. D. When is a correlation matrix appropriate for factor analysis? / C. D. Dziuban, E. C. Shirkey // *Psychological Bulletin*. – 1974. – Vol. 81. – № 6. – P. 358–361.
230. Engle, R. Dynamic conditional correlation / R. Engle // *Journal of Business & Economic Statistics*. – 2002. – Vol. 20. – № 3. – P. 339–350.
231. Enhancing Credit Scoring with Alternative Data / V. D. Biatat, J. Crook, R. Calabrese, M. Hamid // *Expert Systems with Applications*. – 2020. – Vol. 163. – P. 1–30.
232. Fayyad, U. Multi-interval discretization of continuous-value attributes for classification learning / U. Fayyad, K. Irani // *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*. – San Mateo : Morgan Kaufmann, 1993. – Vol. 2. – P. 1022–1027.

233. Fine, T. L. Feedforward Neural Network Methodology / T. L. Fine. – 3-rd ed. – New York : Springer-Verlag, 1999. – 340 p.
234. Finlay, S. Are we modelling the right thing? The impact of incorrect problem specification in credit scoring / S. Finlay // Expert Systems with Applications. – 2009. – Vol. 36. – № 5. – P. 9065–9071.
235. Finney, D. J. Probit analysis / D. J. Finney. – Cambridge : Cambridge University Press, 1971. – 272 p.
236. Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P / R. A. Fisher // Journal of the Royal Statistical Society. – 1922. – Vol. 85. – № 1. – P. 87–94.
237. Fisher, R. A. The use of multiple measurements in taxonomic problems / R. A. Fisher // Annals of Eugenics. – 1936. – Vol. 7. – P. 179–188.
238. Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach / Y. Zhu, L. Zhou, C. Xie [et al.] // International Journal of Production Economics. – 2019. – Vol. 211. – P. 22–33.
239. Fox, J. Applied Regression Analysis, Linear Models, and Related Methods / J. Fox. – Thousand Oaks : SAGE Publications, 1997. – 632 p.
240. Fox, J. Generalized collinearity diagnostics / J. Fox, G. Monette // Journal of the American Statistical Association. – 1992. – Vol. 87. – P. 178–183.
241. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine / J. H. Friedman // The Annals of Statistics. – 2001. – Vol. 29. – № 5. – P. 1189–1232.
242. Frydman, H. Introducing recursive partitioning for financial classification: The case of financial distress / H. Frydman, E. I. Altman, D. L. Kao // Journal of Finance. – 1985. – Vol. 40. – № 1. – P. 269–291.
243. Fuller, W. A. Introduction to statistical time series / W. A. Fuller. – New York : Wiley, 1976. – 488 p.
244. Gardner, E. S. Exponential smoothing: The state of the art / E. S. Gardner // Journal of Forecasting. – 1985. – Vol. 4. – № 1. – P. 1–28.

245. Ghatak, M. The economics of lending with joint liability: theory and practice / M. Ghatak, T. Guinnane // *Journal of Development Economics*. – 1999. – Vol. 60. – № 1. – P. 195–228.
246. Goharshady, A. K. Secure Credit Reporting on the Blockchain / A. K. Goharshady, A. Behrouz, K. Chatterjee // *IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. – Halifax : IEEE Xplore, 2018. – P. 1343–1348.
247. Gonçalves, E. B. Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models / E. B. Gonçalves, M. A. Gouvêa // *International Journal of Advanced Engineering Research and Science*. – 2021. – Vol. 8. – № 9. – P. 198–209.
248. Goodman, L. A. Simple models for the analysis of association in cross-classifications having ordered categories / L. A. Goodman // *Journal of the American Statistical Association*. – 1979. – Vol. 74. – № 367. – P. 537–552.
249. Greenidge, K. Forecasting non-performing loans in Barbados / K. Greenidge, T. Grosvenor // *Journal of Business, Finance and Economics in Emerging Economies*. – 2010. – Vol. 5. – P. 80–108.
250. Günnemann, N. Cost matters: a new example-dependent cost-sensitive logistic regression model / N. Günnemann, J. Pfeffer // *Pacific-Asia conference on knowledge discovery and data mining*. – New York : Springer, 2017. – P. 210–222.
251. Guzmán-Castillo, S. Credit Risk Scoring Model Based on The Discriminant Analysis Technique / S. Guzmán-Castillo, C. Garizabalo-Davila, G. Gatica // *Procedia Computer Science*. – 2023. – Vol. 220. – P. 928–933.
252. Ha, S. Credit scoring with a feature selection approach based deep learning / S. Ha, H.-N. Nguyen // *MATEC Web of Conferences*. – 2016. – Vol. 54. – P. 50–64.
253. Handbook of credit scoring / ed. E. Mays. – Chicago : Glenlake Publishing, 2001. – 392 p.

254. Hanley, J. A. A method of comparing the areas under receiver operating characteristic curves derived from the same cases / J. A. Hanley, B. J. McNeil // *Radiology*. – 1983. – Vol. 148. – № 3. – P. 839–843.

255. Hanley, J. A. The meaning and use of the area under a receiver operating characteristic (ROC) curve / J. A. Hanley, B. J. McNeil // *Radiology*. – 1982. – Vol. 143. – № 1. – P. 29–36.

256. Harman, H. H. *Modern Factor Analysis* / H. H. Harman. – 3-rd ed. – Chicago : University of Chicago Press, 1976. – 492 p.

257. Harrell, F. *Regression modeling strategies* / F. Harrell. – New York : Springer, 2001. – 608 p.

258. Hartigan, J. A. *Clustering algorithms* / J. A. Hartigan. – New York : Wiley, 1975. – 376 p.

259. Harvey, A. C. *Forecasting, structural time series models and the Kalman filter* / A. C. Harvey. – Cambridge : Cambridge University Press, 1989. – 580 p.

260. Hauck, W. Odds ratio inference from stratified samples / W. Hauck // *Communications in Statistics – Theory and Methods*. – 1989. – Vol. 18. – P. 767–800.

261. Hauck, W. Wald's test as applied to hypotheses in logit analysis / W. Hauck, A. Donner // *Journal of the American Statistical Association*. – 1977. – Vol. 72. – № 360. – P. 851–853.

262. Haykin, S. *Neural Networks: A Comprehensive Foundation* / S. Haykin. – 2-nd ed. – New York : Macmillan College Publishing, 1998. – 872 p.

263. Heckman, J. Sample selection bias as a specification error / J. Heckman // *Econometrica*. – 1979. – Vol. 47. – P. 153–161.

264. Hendrickson, A. E. Promax: a quick method for rotation to oblique simple structure / A. E. Hendrickson, P. O. White // *British Journal of Statistical Psychology*. – 1964. – Vol. 17. – № 1. – P. 65–70.

265. Henley, W. A k -nearest-neighbour classifier for assessing consumer credit risk / W. Henley, D. J. Hand // *The statistician*. – 1996. – Vol. 45. – № 1. – P. 77–95.

266. Hosmer, D. *Applied Logistic Regression* / D. Hosmer, S. Lemeshow. – 2-nd ed. – New York : Wiley, 2000. – 408 p.

267. How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm. BIS Working Paper, 2019 – № 834 [Электронный ресурс] / L. Gambacorta, Y. Huang, H. Qiu, J. Wang // Social Science Research Network : [сайт]. – Режим доступа: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3506945 (дата обращения: 12.04.2025).

268. Hsieh, N. C. Hybrid mining approach in the design of credit scoring models / N. C. Hsieh // Expert Systems with Applications. – 2005. – Vol. 28. – № 4. – P. 655–665.

269. Hulme, D. Finance against Poverty / D. Hulme, P. Mosley. – London : Routledge. – 1996. – 240 p.

270. Hurley, M. Credit scoring in the era of big data / M. Hurley, J. Adebayo // Yale Journal of Law & Technology. – 2016. – Vol. 18. – P. 148–216.

271. Influence of unit economics on the development of insurance marketing strategies / O. V. Zhukova, A. S. Ermolaeva, N. B. Shchegolev, V. D. Sekerin // Academy of Strategic Management Journal. – 2021. – Vol. 20. – № 4. – P. 1–8.

272. Integrated framework for profit-based feature selection and SVM classification in credit scoring / S. Maldonado, C. Bravo, J. López, J. Pérez // Decision Support Systems. – 2017. – Vol. 104. – № 1. – P. 113–121.

273. Jaccard, J. Interaction effects in logistic regression / J. Jaccard. – Thousand Oaks : Sage Publications, 2001. – 70 p.

274. Jadhav, S. Information gain directed genetic algorithm wrapper feature selection for credit rating / S. Jadhav, H. He, K. Jenkins // Applied Soft Computing. – 2018. – Vol. 69. – P. 541–553.

275. Jayanti, D. Multivariate generalized autoregressive score model (case study: vegetable oils and crude oil price data) / D. Jayanti, K. Sadik, F. M. Afendi // Journal of Physics: Conference Series. IOP Publishing. – 2021. – Vol. 1863. – № 1. – P. 1–18.

276. Jeffers, J. N. R. Two case studies in the application of principal component analysis / J. N. R. Jeffers // Applied Statistics. – 1967. – Vol. 16. – № 3. – P. 225–236.

277. Jennrich, R. I. Rotation for simple loadings / R. I. Jennrich, P. F. Sampson // *Psychometrika*. – 1966. – Vol. 31. – № 3. – P. 313–323.
278. Jensen, H. L. Using neural networks for credit scoring / H. L. Jensen // *Managerial Finance*. – 1992. – Vol. 18. – P. 15–26.
279. Jöreskog, K. G. Factor analysis by least-square and maximum-likelihood method / K. G. Jöreskog // *Statistical Methods for Digital Computers* / K. Enslein, A. Ralston, R. S. Wilf eds. – New York : Wiley, 1977. – Vol. 3. – P. 108–120.
280. Kaiser, H. F. Image analysis / H. F. Kaiser // *Problems in Measuring Change* / C. W. Harris, ed. – Madison : University of Wisconsin Press, 1963. – P. 161–187.
281. Kalbfleisch, J. D. The statistical analysis of failure time data / J. D. Kalbfleisch, R. L. Prentice. – 2-nd ed. – New York : Wiley, 2002. – 344 p.
282. Kalli, M. Time-varying sparsity in dynamic regression models / M. Kalli, J. E. Griffin // *Journal of Economic Literature*. – 2014. – Vol. 178. – № 2. – P. 779–793.
283. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems / R. E. Kalman // *Transactions of ASME, Journal of Basic Engineering*. – 1960. – Vol. 82. – № 1. – P. 35–45.
284. Kamaloo, E. Credit risk prediction using fuzzy immune learning / E. Kamaloo, M. Saniee Abadeh // *Advances in Fuzzy Systems*. – 2014. – Vol. 16. – P. 1–11.
285. Kass, G. V. An exploratory technique for investigating large quantities of categorical data / G. V. Kass // *Applied Statistics*. – 1980. – Vol. – 29. – № 2. – P. 119–127.
286. Kaufman, L. Finding groups in data: An introduction to cluster analysis / L. Kaufman, P. J. Rousseeuw. – New York : Wiley, 2005. – 370 p.
287. Kendall, M. G. A course in multivariate analysis / M. G. Kendall. – London : Griffin, 1957. – 196 p.
288. Khan, M. A. Determinants of non-performing loans in the banking sector in developing state / M. A. Khan, A. Siddique, Z. Sarwar // *Asian Journal of Accounting Research*. – 2020. – Vol. 5. – № 1. – P. 135–145.

289. Khandani, A. Consumer credit risk models via machine-learning algorithms / A. Khandani, K. Adlar, L. Andrew // *Journal of Banking and Finance*. – 2010. – Vol. 34. – № 11. – P. 2767–2787.
290. Kitonyi, J. M. Non-Performing Loans and Financial Performance of Microfinance Institution in Kenya / J. M. Kitonyi, W. Sang, D. Muriithi // *The Strategic Journal of Business and Change Management*. – 2019. – Vol. 6. – № 3. – P. 840–848.
291. Kleinbaum, D. G. Logistic regression: A Self-Learning Text / D. G. Kleinbaum. – New York : Springer-Verlag, 1994. – 282 p.
292. Knight, F. H. Risk, uncertainty and profit / F. H. Knight. – New York : Houghton Mifflin Company, 1921. – 404 p.
293. Koc, U. Consumer loans' first payment default detection: a predictive model / U. Koc, T. Sevgili // *Turkish Journal of Electrical Engineering and Computer Sciences*. – 2020. – Vol. 28. – № 1. – P. 167–181.
294. Kohn, R. Efficient estimation and prediction in time series regression models / R. Kohn, C. Ansley // *Biometrika*. – 1985. – Vol. 72. – № 3. – P. 694–697.
295. Kohn, R. Estimation, prediction, and interpolation for ARIMA models with missing data / R. Kohn, C. Ansley // *Journal of the American Statistical Association*. – 1986. – Vol. 81. – P. 751–761.
296. Kritzinger, N. An optimized credit scorecard to enhance cut-off score determination / N. Kritzinger, G. Vuuren // *South African Journal of Economic and Management Sciences*. – 2018. – Vol. 21. – № 1. – P. 15–71.
297. Lawless, R. F. Statistical models and methods for lifetime data / R. F. Lawless. – New York : Wiley, 1982. – 600 p.
298. Ledolter, J. Some comments on the initialization of exponential smoothing / J. Ledolter, B. Abraham // *Journal of Forecasting*. – 1984. – Vol. 3. – P. 79–84.
299. Lee, T.-S. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines / T.-S. Lee, I.-F. Chen // *Expert Systems with Applications*. – 2005. – Vol. 28. – P. 743–752.
300. Lewis, E. M. An introduction to credit scoring / E. M. Lewis. – San Rafael : The Athena Press, 1992. – 172 p.

301. Lilliefors, H. W. On the Kolmogorov-Smirnov tests for normality with mean and variance unknown / H. W. Lilliefors // *Journal of the American Statistical Association*. – 1967. – Vol. 62. – № 318. – P. 399–402.
302. Lim, T. S. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms / T. S. Lim, W.-Y. Loh, Y.-S. Shih // *Machine Learning*. – 2000. – Vol. 40. – № 3. – P. 203–228.
303. Little, R. J. A test of missing completely at random for multivariate data with missing values / R. J. Little // *Journal of the American Statistical Association*. – 1998. – Vol. 83. – № 404. – P. 1198–1202.
304. Loh, W.-Y. Split selection methods for classification trees / W.-Y. Loh, Y. S. Shih // *Statistica Sinica*. – 1997. – Vol. 7. – P. 815–840.
305. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects / E. Dumitrescu, S. Hué, Ch. Hurlinb, S. Tokpavi // *European Journal of Operational Research*. – 2022. – Vol. 297. – P. 1178–1192.
306. Makowski, P. Credit scoring branches out / P. Makowski // *The Credit World*. – 1985. – Vol. 75. – № 1. – P. 30–37.
307. Makridakis, S. G. Forecasting: Methods and applications / S. G. Makridakis, S. C. Wheelwright, R. J. Hyndman. – 3-rd ed. – New York : Wiley, 1997. – 744 p.
308. Maranga, B. S. Application of Linear Logistic and Discriminant Analysis on Forecasting Creditworthiness of Individual Borrowers / B. S. Maranga. – Nairobi : University of Nairobi, 2013. – 60 p.
309. Mathuva, D. Capital Adequacy, Cost Income Ratio and the Performance of Commercial Banks: The Kenyan Scenario / D. Mathuva // *The International Journal of Applied Economics and Finance*. – 2009. – Vol. 3. – № 2. – P. 35–47.
310. Melard, G. A fast algorithm for the exact likelihood of autoregressive-moving average models / G. Melard // *Applied Statistics*. – 1984. – Vol. 33. – № 1. – P. 104–114.

311. Meng, C. The Practice Study of Consumer Credit Risk Based on Random Forest / C. Meng, B. Liu, L. Zhou // *Advances in Intelligent Systems Research*. – 2019. – Vol. 168. – P. 101–106.
312. Miled, K. Risk Assessment for Reimbursement of Microfinance Institutions / K. Miled, M. Landolsi // *International Journal of Professional Business Review*. – 2023. – Vol. 8. – № 10. – P. 1–23.
313. Mohamed, W. Risk Management of Banking Sector: A Critical Review / W. Mohamed // *American Journal of Social Sciences*. – 2016. – Vol. 5. – P. 159–167.
314. Moiseev, N. A. Interval forecast for model averaging methods / N. A. Moiseev, A. S. Sorokin // *Model Assisted Statistics and Applications*. – 2018. – Vol. 13. – № 2. – P. 161–172.
315. Mok, J.-M. Reject inference in credit Scoring [Электронный ресурс] / J.-M. Mok // Vrije Universiteit Amsterdam : [сайт]. – 2009. – Режим доступа: http://www.few.vu.nl/en/Images/werkstuk-mok_tcm244-91398.pdf (дата обращения: 14.02.2025).
316. Mokheleli, T. Machine Learning Approach for Credit Score Predictions / T. Mokheleli, T. Museba // *Journal of Information Systems and Informatics*. – 2023. – Vol. 5. – № 2. – P. 497–517.
317. Montrenko, A. Sample Size Determination for Logistic Regression / A. Montrenko, V. Strijov, G.-W. Weber // *Journal of Computational and Applied Mathematics*. – 2014. – Vol. 255. – № 2. – P. 743–752.
318. Montrichard, D. Reject inference methodologies in credit risk modeling / D. Montrichard // *The Proceedings of the South-East SAS Users Group*. – 2007. – Vol. 3. – P. 12566–12572.
319. Morf, M. Some new algorithms for recursive estimation in constant, linear, discrete-time systems / M. Morf, G. S. Sidhu, T. Kailath // *IEEE Transactions on Automatic Control*. – 1974. – Vol. 19. – № 4. – P. 315–323.
320. Morrison, D. F. Multivariate statistical methods / D. F. Morrison. – New York : McGraw-Hill, 1976. – 496 p.

321. Nagelkerke, N. J. D. A note on the general definition of the coefficient of determination / N. J. D. Nagelkerke // *Biometrika*. – 1991. – Vol. 78. – № 3. – P. 691–692.
322. Njuguna, R. Poster: A Scoping Review of Alternative Credit Scoring Literature / R. Njuguna, K. Sowon // *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '21)*. – New York : Association for Computing Machinery, 2021. – P. 437–444.
323. Nymoen, R. Introduction to Dynamic Macroeconometrics / R. Nymoen // *Dynamic Econometrics for Empirical Macroeconomic Modelling*. – New York : WSPC, 2019. – P. 4–32.
324. On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance / R. Munoz-Cancino, C. Bravo, S. A. Ríos, M. Graña // *Expert Systems with Applications*. – 2022. – Vol. 218. – P. 1–18.
325. Opitz, D. Popular ensemble methods: An empirical study / D. Opitz, R. Maclin // *Journal of Artificial Intelligence*. – 1999. – Vol. 11. – P. 169–198.
326. Orlando, G. Non-performing loans for Italian companies: When time matters: An empirical research on estimating probability to default and loss given default / G. Orlando, R. Pelosi // *International Journal of Financial Studies*. – 2020. – Vol. 8. – № 4. – Art. 68. – P. 1–22.
327. Orlova, E. V. Methodology and models for individuals' creditworthiness management using digital footprint data and machine learning methods / E. V. Orlova // *Mathematics*. – 2021. – Vol. 9. – № 15. – Art. 1820. – P. 1–28.
328. Ots, H. Mobile Phone Usage Data for Credit Scoring / H. Ots, I. Liiv, D. Tur // *Communications in Computer and Information Science*. – 2020. – Vol. 1243. – P. 113–121.
329. Pal, R. A Study on Credit Scoring Models with different Feature Selection and Machine Learning Approaches / R. Pal, Sh. Kapali, Sh. Trivedi // *Technology in Society*. – 2020. – Vol. 63. – P. 1–9.

330. Pearlman, J. G. An algorithm for the exact likelihood of a high-order autoregressive-moving average process / J. G. Pearlman // *Biometrika*. – 1980. – Vol. 67. – № 1. – P. 232–233.
331. Pearson, K. On lines and planes of closest fit to systems of points in space / K. Pearson // *Philosophical Magazine*. – 1901. – Vol. 2. – P. 559–572.
332. Pearson, K. Problem of the Random Walk / K. Pearson // *Nature*. – 1905. – Vol. 72. – № 1. – P. 294–342.
333. Preap, S. Determinants of Non-Performing Loan of Micro-Finance Sector in Cambodia [Электронный ресурс] / S. Preap, R. Phok // 3rd Annual National Bank of Cambodia Macroeconomic Conference. – 2016. – Режим доступа: https://nbc.gov.kh/download_files/research_papers/english/4908DeterminantsofNon-PerformingLoan-ofMicro-FinanceSectorinCambodia_Sreynich_Rathak.pdf (дата обращения: 14.04.2025).
334. Predicting creditworthiness in retail banking with limited scoring data / H. A. Abdou, M. Tsafack, C. G. Ntim, R. D. Baker // *Knowledge-Based Systems*. – 2016. – Vol. 103. – P. 89–103.
335. Putri, N. Credit risk analysis using support vector machines algorithm / N. Putri, M. Fatekurohman, I. Tirta // *Journal of Physics: Conference Series*. – 2021. – Vol. 1836. – P. 1–11.
336. Quan, J. Credit risk assessment using the factorization machine model with feature interactions / J. Quan, X. Sun // *Humanities & Social Sciences Communications*. – 2024. – Vol. 11. – № 234. – P. 1–10.
337. Quantitative assessment of radiographic control informativeness using ROC analysis / A. S. Sorokin, D. I. Galkin, A. E. Shubochkin, E. A. Ivanayskiy // *Journal of Physics: Conference Series*. – 2019. – Vol. 1327. – Art. 12013. – P. 1–8.
338. Rencher, A. *Methods of Multivariate Analysis* / A. Rencher. – Hoboken : Wiley, 2002. – 656 p.
339. Research on personal credit scoring model based on multi-source data / Z. Haichao, Z. Ruishuang, C. Linling, Z. Shangfeng // *Journal of Physics: Conference Series*. – 2020. – Vol. 1437. – № 1. – P. 466–473.

340. Research on personal credit scoring model based on multi-source data / H. Zhang, R. Zeng, L. Chen, S. Zhang // *Journal of Physics: Conference Series*. – 2020. – Vol. 1437. – P. 2–14.
341. Ripley, B. D. *Pattern Recognition and Neural Networks* / B. D. Ripley. – Cambridge : Cambridge University Press, 1996. – 403 p.
342. Rojas, R. *Neural Networks: A Systematic Introduction* / R. Rojas. – Berlin : Springer-Verlag, 1996. – 502 p.
343. Rothschild, M. Increasing risk: I. A definition / M. Rothschild, J. E. Stiglitz // *Journal of Economic Theory*. – 1970. – Vol. 2. – № 3. – P. 225–243.
344. Rubin, R. B. *Multiple Imputation for Nonresponse in Surveys* / R. B. Rubin. – New York : Wiley, 1987. – 320 p.
345. Rummel, R. J. *Applied factor analysis* / R. J. Rummel. – Evanston : Northwestern University Press, 1970. – 617 p.
346. Runchi, Zh. An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects / Zh. Runchi, L. Xue, W. Qin // *Expert Systems with Applications*. – 2023. – Vol. 212. – № 1. – Art. 118732. – P. 1–12.
347. Sanchez, F. Aplicación del modelo credit scoring y regresión logística en la predicción del crédito, en una entidad financiera de la ciudad del Cusco 2022 / F. Sanchez, S. Yeli. – Cusco : Universidad Nacional de San Antonio Abad del Cusco, 2023. – 106 p.
348. Sang, H. A Novel Credit Scoring Prediction Model based on Feature Selection Approach and Parallel Random Forest / H. Sang, N. Nam, N. Nguyen // *Indian Journal of Science and Technology*. – 2016. – Vol. 9. – № 20. – P. 1–6.
349. Santner, T. J. A Note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models / T. J. Santner, E. D. Duffy // *Biometrika*. – 1986. – Vol. 73. – P. 755–758.
350. Saxena, A. Credit Risk Assessment Model for UAE's Commercial Banks: A Machine Learning Approach / A. Saxena, P. Dungore // Reviewed and Submitted by Undergraduate Research Committee. – Abu Dhabi : Abu Dhabi University, 2021. P. 1–9.

351. Schafer, J. L. Analysis of Incomplete Multivariate Data / J. L. Schafer. – London : Chapman and Hall, 1997. – 444 p.
352. Schneider, W. Systems of seemingly unrelated regression equations with time varying coefficients – An interplay of Kalman filtering, scoring, EM- and MINQUE-method / W. Schneider // Computers & Mathematics with Applications. – 1992. – Vol. 24. – № 8–9. – P. 1–16.
353. Schwarz, A. Measurement, monitoring, and forecasting of consumer credit default risk: An indicator approach based on individual payment histories / A. Schwarz // Schumpeter Discussion Papers. – 2011. – № 2011-004. – P. 1–28.
354. Shao, J. The Jackknife and Bootstrap / J. Shao, D. Tu. – New York : Springer, 1995. – 517 p.
355. Shapiro, S. S. An analysis of variance test for normality / S. S. Shapiro, M. B. Wilk // Biometrika. – 1965. – Vol. 52. – № 3. – P. 591–599.
356. Shen, F. Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory / F. Shen, X. Zhao, G. Kou // Decision Support Systems. – 2020. – Vol. 137. – Art. 113366. – P. 1–10.
357. Shofiyah, F. Split and Conquer Method in Penalized Logistic Regression with Lasso (Application on Credit Scoring Data) / F. Shofiyah, A. Sofro // Journal of Physics: Conference Series. – 2018. – Vol. 1108. – Art. 012107. – P. 1–7.
358. Siddiqi, N. Credit risk scorecards: developing and implementing intelligent credit scoring / N. Siddiqi. – New Jersey : John Wiley and Sons, 2012. – 208 p.
359. Skok, D. SaaS Metrics 2.0 – A Guide to Measuring and Improving what Matters / D. Skok // Matrix Management Corporation : [сайт]. – Режим доступа: <https://www.forentrepreneurs.com/saas-metrics-2/> (дата обращения: 14.04.2025).
360. Smith, L. D. Forecasting losses on a liquidating long-term loan portfolio / L. D. Smith, E. C. Lawrence // Journal of Banking & Finance. – 1995. – Vol. 19. – № 6. – P. 959–985.
361. Soureshjani, M. H. Calculating the best cut off point using logistic regression and neural network on credit scoring problem – A case study of a

commercial bank / M. H. Soureshjani, A. M. Kimiagari // *African Journal of Business Management*. – 2013. – Vol. 7. – № 16. – P. 1414–1421.

362. Spatiotemporal Integration of Mobile, Satellite, and Public Geospatial Data for Enhanced Credit Scoring / N. Simumba, S. Okami, A. Kodaka, N. Kohtake // *Symmetry*. – 2021. – Vol. 13. – Art. 575. – P. 1–17.

363. Storer, B. E. A diagnostic for Cox regression and general conditional likelihoods / B. E. Storer, J. Crowley // *Journal of the American Statistical Association*. – 1985. – Vol. 80. – P. 139–147.

364. Suhadolnik, N. Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach / N. Suhadolnik, J. Ueyama, S. Da Silva // *Journal of Risk and Financial Management*. – 2023. – Vol. 16. – № 12. – P. 1–21.

365. Sustersic, M. Consumer credit scoring models with limited data / M. Sustersic, D. Mramor, J. Zupan // *Expert Systems with Applications*. – 2009. – Vol. 36. – № 3. – P. 4736–4744.

366. Tan, P. Introduction to Data Mining / P.-N. Tan, M. Steinbach, V. Kumar. – Reading : Addison-Wesley, 2006. – 719 p.

367. Tatsuoka, M. M. Multivariate analysis / M. M. Tatsuoka. – New York : Wiley, 1971. – 357 p.

368. The handbook of brain theory and neural networks / ed. M. A. Arbib. – London : MIT Press, 1998. – 1134 p.

369. The Treatment of Nonperforming Loans. Clarification and Elaboration of Issues Raised by the December 2004 Meeting of the Advisory Expert Group of the Intersecretariat Working Group on National Accounts [Электронный ресурс] // International Monetary Fund : [сайт]. – Режим доступа: <https://www.imf.org/external/pubs/ft/bop/2005/05-29.pdf> (дата обращения 11.04.2025).

370. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics / M. Óskarsdóttir, C. Bravo, C. Sarraute [et al.] // *Applied Soft Computing*. – 2018. – Vol. 74. – № 2. – P. 26–39.

371. Thomas, L. C. A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers / L. C. Thomas // *International Journal of Forecasting*. – 2000. – Vol. 16. – P. 149–172.
372. Thomas, L. C. Consumer credit models: pricing, profit, and portfolios / L. C. Thomas. – New York : Oxford University press, 2009. – 400 p.
373. Thomas, L. Credit Scoring and Its Applications / L. Thomas, D. Edelman, J. Crook. – Philadelphia: Society for Industrial and Applied Mathematics, 2002. – 392 p.
374. Tounsi, Y. Credit scoring in the age of Big Data -A State-of-the-Art / Y. Tounsi, L. Hassouni, H. Anoun // *International Journal of Computer Science and Information Security*. – 2017. – Vol. 15. – № 7. – P. 134–145.
375. Van Braak, B. How can consumers without credit history benefit from the use of information processing and machine learning tools by financial institutions? / B. Van Braak, J. R. Osterrieder, M. R. Machado // *Information Processing and Management*. – 2024. – Vol. 62. – № 2. – P. 1–22.
376. Vapnik, V. N. The nature of statistical learning theory / V. N. Vapnik. – Berlin : Springer-Verlag, 1995. – 314 p.
377. Velleman, P. F. Efficient computing of regression diagnostics / P. F. Velleman, R. E. Welsch // *American Statistician*. – 1981. – Vol. 35. – P. 234–242.
378. Wang, Y. Improving Investment Suggestions for Peer-to-Peer Lending via Integrating Credit Scoring into Profit Scoring / Y. Wang, X. S. Ni // *Proceedings of the 2020 ACM Southeast Conference (ACM SE '20)*. – New York : Association for Computing Machinery, 2020. – P. 141–148.
379. Wen, H. Study on Effect of Consumer Information in Personal Credit Risk Evaluation / H. Wen, X. Sui, Sh. Lu // *Complexity*. – 2022. – Art. 7340010. – P. 1–11.
380. West, D. Neural network credit scoring models / D. West // *Computers & Operations Research*. – 2000. – Vol. 27. – № 11. – P. 1131–1152.
381. West, D. Neural network ensemble strategies for financial decision applications / D. West, S. Dellana, J. Qian // *Computers and Operations Research*. – 2005. – Vol. 32. – № 10. – P. 2543–2559.

382. Whittaker, J. A dynamic mortgage scorecard using Kalman filtering [Электронный ресурс] / J. Whittaker, C. Whitehead, M. Somers // Credit Research Centre : [сайт]. – Режим доступа: <https://cer.business-school.ed.ac.uk/wp-content/uploads/sites/55/2017/03/whittaker-whitehead-somers-1.pdf> (дата обращения: 12.04.2025).

383. Wu, X. Research on Consumer Credit Rating Model / X. Wu, J. Shang // Proceedings of the 2023 International Conference on Finance, Trade and Business Management (FTBM 2023). – Paris : Atlantis Press, 2023. – P. 157–163.

384. Yunus, M. Banker to the Poor / M. Yunus. – New York : Public Affairs, 1999. – 288 p.

385. Zhou, S. R. A nearly neutral model of biodiversity / S. R. Zhou, D. Y. Zhang // Ecology. – 2008. – Vol. 89. – № 1. – P. 248–258.

Приложение А

(рекомендуемое)

Примеры источников данных для оценки кредитных рисков на российском рынке и их основные характеристики

Таблица А.1 – Примеры источников данных для оценки кредитных рисков на российском рынке

Дата-провайдер	Данные (продукт)	Способ обработки	Тип данных	Цель применения	Возможность ретро-теста
«Бюро кредитных историй «Скоринг бюро»	кредитный отчет	необработанные исходные данные xml	данные кредитного поведения и кредитных заявок	для оценки кредитных рисков	да
	скоринг для МФО.PDL	скоринговый балл	данные кредитных заявок		да
	сервис по противодействию мошенничеству Fraud Prevention Service (FPS)	скоринговый балл и агрегированные переменные		для оценки риска мошенничества	да
Juicy Labs	скоринговый отчет Juicy Score	скоринговый балл и агрегированные переменные	данные об устройстве	для оценки риска мошенничества	нет
Аналитическое агентство «Финкарта»	аналитический отчет «Финкарта»	скоринговый балл, агрегированные переменные	транзакционные данные	для повышения конверсии	да
			данные об устройстве	для оценки риска мошенничества	нет
			данные заявок на получение займа	для верификации	да
One Factor	сервис оценки кредитных рисков и мошенничества на телеком-данных Smart Machine	скоринговый балл и агрегированные переменные	данные сотовых операторов	для оценки кредитных рисков для оценки риска мошенничества	да

Источник: составлено автором на основе открытой информации в сети интернет

Приложение Б

(рекомендуемое)

Пример работы алгоритма бининга на основе анализа показателей веса категорий

Работа алгоритма бининга на основе анализа показателей WOE и IV реализуется на примере количественной переменной стажа работы при построении одной из условной модели оценки кредитных рисков.

В таблице Б.1 приведены рассчитанные показатели веса категорий (WOE). В данном примере при расчете WOE сравниваются «плохие» заемщики с «хорошими».

Таблица Б.1 – Рассчитанные показатели WOE на основе децилей

Группы переменной стаж работы на основе децилей		Наличие дефолта по кредиту		Итого	Доля дефолтов	Доля не дефолтов	WOE	IV
		Да	Нет					
Стаж работы на последнем месте, лет	<= 1	51	60	111	0,28	0,12	0,88	0,14
	2	25	19	44	0,14	0,04	1,31	0,13
	3 – 4	31	58	89	0,17	0,11	0,41	0,02
	5 – 6	21	61	82	0,11	0,12	–0,03	0,00
	7 – 7	8	30	38	0,04	0,06	–0,28	0,00
	8 – 9	14	62	76	0,08	0,12	–0,45	0,02
	10 – 12	8	48	56	0,04	0,09	–0,75	0,04
	13 – 14	14	57	71	0,08	0,11	–0,37	0,01
	15 – 18	7	66	73	0,04	0,13	–1,21	0,11
	19+	4	56	60	0,02	0,11	–1,60	0,14
Итого		183	517	700	1	1		0,62

Источник: составлено автором

По результатам анализа представленных в таблице Б.1 значений WOE можно объединить группы с близкими значениями WOE в одну, получив новую переменную стажа работы с 5 категориями вместо 10. Для новой переменной значения WOE повторно рассчитаны в таблице Б.2.

Таблица Б.2 – Рассчитанные показатели WOE после укрупнения категорий

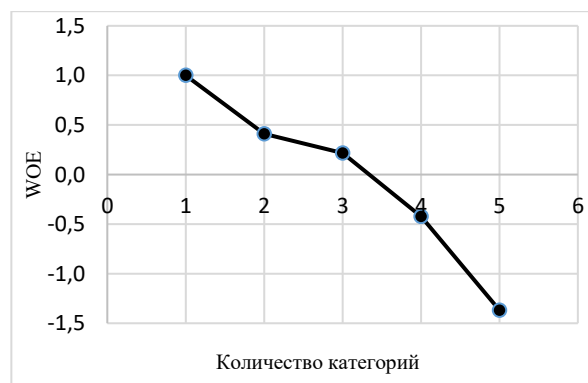
Группы переменной стаж работы на основе 5 категорий		Наличие дефолта по кредиту		Итого	% дефолтов	% не дефолтов	WOE	IV
		Да	Нет					
Стаж работы на последнем месте, лет	<= 2	76	79	155	0,42	0,15	1,00	0,26
	3 – 4	31	58	89	0,17	0,11	0,41	0,02
	5 – 6	21	61	82	0,11	0,12	–0,03	0,00
	7 – 14	44	197	241	0,24	0,38	–0,46	0,06
	15 +	11	122	133	0,06	0,24	–1,37	0,24
Итого		183	517	700	1,00	1,00		0,59

Источник: составлено автором

На рисунке Б.1 представлено изменение предсказательной способности категорий в зависимости от их числа. Видно, что благодаря процедуре биннинга удалось получить новую переменную с меньшим числом категорий, для каждой из которых значение WOE уменьшается с ростом значения категории.



а) Значения WOE до объединения категорий



б) Значения WOE после объединения категорий

Рисунок Б.1– Изменение предсказательной способности категорий в зависимости от их числа

Источник: составлено автором

Приложение В

(рекомендуемое)

Пример работы алгоритма простой MDLP-дискретизации

В данном примере расчета используется искусственный набор данных из 250 наблюдений. Множество наблюдений S не представлено, т. к. оно может быть восстановлено путем упорядочивания записей по возрастанию значений переменной X .

На первоначальном этапе были упорядочены наблюдения из S по значению X в порядке возрастания. Далее находится множество всех уникальных пар наблюдений U . В данном примере $|U| = 46$.

Множество значений переменной X для всех пар наблюдений из U имеет вид:

$$\begin{aligned}
 X_U = & -2,6; -2,4; -2,1; -2; -1,9; -1,8; -1,7; -1,6; -1,5; -1,4; -1,3; -1,2; -1,1; -1; \\
 & -0,9; -0,8; -0,7; -0,6; -0,5; -0,4; -0,3; -0,2; -0,1; \\
 & 0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; \\
 & 1; 1,1; 1,2; 1,3; 1,4; 1,5; 1,6; 1,7; 1,8; 1,9; 2; 2,1; 2,3\}.
 \end{aligned}$$

Далее для каждого класса вычислена частота встречаемости каждого уникального значения $u_i \in U$ и получена матрица V , элемент v_{ij} которой обозначает общее количество наблюдений, которые соответствуют значению u_i и принадлежат j -му классу (см. таблицу В.1).

Таблица В.1 – Матрица частоты встречаемости V

U	Класс				U	Класс			
	0	1	2	3		0	1	2	3
-2,6	0	1	0	0	0	0	0	0	14
-2,4	0	1	0	0	0,1	0	0	0	6
-2,1	0	2	0	0	0,2	0	0	0	10
-2,0	0	0	2	0	0,3	0	0	0	14
-1,9	0	0	2	0	0,4	0	0	0	12

U	Класс				U	Класс			
	0	1	2	3		0	1	2	3
-1,8	0	0	3	0	0,5	0	0	0	4
-1,7	0	0	2	0	0,6	0	0	0	9
-1,6	0	0	3	0	0,7	0	0	0	5
-1,5	0	0	3	0	0,8	0	0	0	3
-1,4	0	0	2	0	0,9	0	0	0	10
-1,3	0	0	4	0	1	3	0	0	0
-1,2	0	0	3	0	1,1	8	0	0	0
-1,1	0	0	3	0	1,2	5	0	0	0
-1,0	0	0	8	0	1,3	7	0	0	0
-0,9	0	0	6	0	1,4	2	0	0	0
-0,8	0	0	7	0	1,5	2	0	0	0
-0,7	0	0	13	0	1,6	3	0	0	0
-0,6	0	0	8	0	1,7	3	0	0	0
-0,5	0	0	4	0	1,8	4	0	0	0
-0,4	0	0	6	0	1,9	4	0	0	0
-0,3	0	0	7	0	2	2	0	0	0
-0,2	0	0	13	0	2,1	2	0	0	0
-0,1	0	0	14	0	2,3	1	0	0	0

Источник: составлено автором по [216, 218, 224].

Далее был запущен алгоритм MLDP-отсечения. На первом шаге алгоритма вычислены значения $E(X, u_i, S)$ для каждого элемента $u_i \in U$ и $u_{i+1} \in U$, не принадлежащих одному и тому же классу (см. таблицу В.2).

Таблица В.2 – Промежуточные вычисления на шаге 1

u_i	-2,1	-0,1	0,9
$E(X, u_i, S)$	1,4742	0,5955	0,9038

$$T_A = \{-0,1\}.$$

$$U_1 = \{-2,6; -2,4; -2,1; -2; -1,9; -1,8; -1,7; -1,6; -1,5; -1,4; -1,3; -1,2; -1,1; -1; -0,9; -0,8; -0,7; -0,6; -0,5; -0,4; -0,3; -0,2; -0,1\};$$

$$S_1 = \{S | X \leq -0,1\}.$$

$$U_2 = \{0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1; 1,1; 1,2; 1,3; 1,4; 1,5; 1,6; 1,7; 1,8; 1,9; 2; 2,1; 2,2; 2,3\};$$

$$S_2 = \{S|X > -0,1\}.$$

На втором шаге алгоритма вычислены значения $E(X, u_i, S_1)$ для каждого элемента $u_i \in U_1$ и $u_{i+1} \in U_1$, не принадлежащих одному и тому же классу (см. таблицу В.3).

Таблица В.3 – Промежуточные вычисления на шаге 2

u_i	-2,1
$E(X, u_i, S_1)$	0

$$T_A = \{-0,1; -2,1\}.$$

$$U_{1,1} = \{-2,6; -2,4; -2,1\}; S_{1,1} = \{S_1|X \leq -2,1\}.$$

$$U_{1,2} = \{-2; -1,9; -1,8; -1,7; -1,6; -1,5; -1,4; -1,3; -1,2; -1,1; \\ -1; -0,9; -0,8; -0,7; -0,6; -0,5; -0,4; -0,3; -0,2; -0,1\};$$

$$S_{1,2} = \{S_1|X > -2,1\}.$$

Все наблюдения в $S_{1,1}$ принадлежат одному классу, следовательно $S_{1,1}$ более делить невозможно. Все наблюдения в $S_{1,2}$ также принадлежат одному классу, следовательно $S_{1,2}$ аналогично более делить невозможно.

На третьем шаге алгоритма вычислены значения $E(X, u_i, S_2)$ для каждого элемента $u_i \in U_2$ и $u_{i+1} \in U_2$, не принадлежащих одному и тому же классу (см. таблицу В.4).

Таблица В.4 – Промежуточные вычисления на шаге 3

u_i	0,9
$E(X, u_i, S_2)$	0

$$T_A = \{-0,1; -2,1; 0,9\}.$$

$$U_{2,1} = \{0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9\}; S_{2,1} = \{S_2|X \leq 0,9\}.$$

$$U_{2,2} = \{1; 1,1; 1,2; 1,3; 1,4; 1,5; 1,6; 1,7; 1,8; 1,9; 2; 2,1; 2,3\}; S_{2,2} = \{S_2|X > 0,9\}.$$

Все наблюдения в $S_{2,1}$ принадлежат одному классу, следовательно $S_{2,1}$ более делить невозможно. Все наблюдения в $S_{2,2}$ также принадлежат одному классу, следовательно $S_{2,2}$ аналогично более делить невозможно. Таким образом, определено, что множество T_A является множеством границ «оптимальных» категорий для данной числовой переменной X .

Приложение Г

(обязательное)

Расчеты юнит-экономики первичного займа МФО

Таблица Г.1 – Исходные данные для проведения расчетов

	Номер когорты, i		1	2	3	4	5	6
	Период		апр.23	май.23	июн.23	июл.23	авг.23	сен.23
	Показатель	Обозначение						
Константные значения бизнес-метрик за соответствующий месяц	Средняя ставка в день, %	AI_i	1,0	1,0	1,0	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	$OSRV_i$	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF_i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявке, руб.	SC_i	50,0	50,0	50,0	50,0	50,0	50,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC_i	2 000	2 000	2 000	2 000	2 000	2 00
	Показатель возврата клиентов из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC_i	5	5	5	5	5	5
Фактические значения бизнес-метрик за соответствующий месяц	Средний чек, руб.	AD_i	10 000	10 100	10 200	10 300	10 400	10 500
	Средний срок по договору, дней	AT_i	18,6	18,8	20,3	22,0	22,6	23,0
	Уровень досрочного погашения (от срока), %	ER_i	34,5	36,8	34,4	36,6	34,5	37,0
	Проникновения дополнительных услуг, %	OS_i	80,2	81,5	79,3	80,1	82,1	80,3
	Средняя сумма за дополнительные услуги, руб.	OSV_i	1 536	1 666	1 589	1 477	1 603	1 578
	Проникновение дополнительных услуг при продлении, %	OSR_i	83,8	83,8	83,8	83,8	83,8	83,8
	Проникновения продления, %	RN_i	40,0	39,1	41,1	39,5	39,8	40,2
	Средний срок пролонгаций у продливших, дней	ATR_i	29,1	30,0	28,5	25,6	29,3	29,9
	Среднее количество пролонгаций у продливших, ед.	NR_i	2,1	2,0	1,9	1,9	2,2	1,9
	Чистые кредитные потери по показателю $NPL_{10\ net}$, %	NPL_i	25,1	26,5	27,3	26,4	26,8	27,0
	Количество заявок в СПР, ед.	CRN_i	105 345	101 344	98 456	101 345	100 235	105 254
	Одобрено заявок в СПР, ед.	$ACRN_i$	30 051	29 568	31 456	30 623	30 532	30 943
	Количество выдач, ед.	CN_i	21 459	22 563	23 745	22 478	20 365	22 505

	Номер когорты, i		7	8	9	10	11	12
	Период		окт.23	ноя.23	дек.23	янв.24	фев.24	мар.24
	Показатель	Обозначение						
Константные значения бизнес-метрик за соответствующий месяц	Средняя ставка в день, %	AI _i	0,8	0,8	0,8	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	OSRV _i	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF _i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявки, руб.	SC _i	50,0	50,0	50,0	50,0	50,0	50,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC _i	2 000	2 000	2 000	2 000	2 000	2 000
	Показатель возврата клиентов (Retention Rate) из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC _i	5	5	5	5	5	5
Фактические значения бизнес-метрик за соответствующий месяц	Средний чек, руб.	AD _i	10 600	10 000	10 100	10 200	10 300	10 400
	Средний срок по договору, дней	AT _i	24,2	22,6	23,7	22,7	23,0	21,2
	Уровень досрочного погашения (от срока), %	ER _i	36,2	36,6	34,7	35,5	37,1	35,2
	Проникновения дополнительных услуг, %	OS _i	80,5	80,2	79,3	81,6	80,4	79,8
	Средняя сумма за дополнительные услуги, руб.	OSV _i	1 589	1 789	1 692	1 568	1 680	1 703
	Проникновение дополнительных услуг при продлении, %	OSR _i	83,8	83,8	83,8	83,8	83,8	83,8
	Проникновения продления, %	RN _i	40,1	38,6	38,8	39,9	40,4	40,1
	Средний срок пролонгаций у продливших, дней	ATR _i	28,0	28,3	28,9	29,0	29,5	28,8
	Среднее количество пролонгаций у продливших, ед.	NR _i	2,1	2,1	2,0	2,2	2,1	1,9
	Чистые кредитные потери по показателю NPL10 net, %	NPL _i	27,1	26,4	26,9	25,9	26,1	26,8
	Количество заявок в СПР, ед.	CRN _i	104 325	110 243	115 527	98 321	101 924	103 294
	Одобрено заявок в СПР, ед.	ACRN _i	32 451	35 623	40 114	29 341	30 654	31 239
	Количество выдач, ед.	CN _i	23 011	26 934	29 830	20 927	23 423	23 347

	Номер когорты, i		13	14	15	16	17	18
	Период		апр.24	май.24	июн.24	июл.24	авг.24	сен.24
	Показатель	Обозначение						
Константные значения бизнес-метрик за соответствующий месяц	Средняя ставка в день, %	AI _i	0,8	0,8	0,8	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	OSRV _i	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF _i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявке, руб.	SC _i	50,0	50,0	50,0	50,0	50,0	50,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC _i	2 000	2 000	2 000	2 000	2 000	2 000
	Показатель возврата клиентов (Retention Rate) из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC _i	5	5	5	5	5	5
Фактические значения бизнес-метрик за соответствующий месяц	Средний чек, руб.	AD _i	10 300	10 300	10 300	10 300	10 300	10 300
	Средний срок по договору, дней	AT _i	21,9	21,9	21,9	21,9	21,9	21,9
	Уровень досрочного погашения (от срока), %	ER _i	35,8	35,8	35,8	35,8	35,8	35,8
	Проникновения дополнительных услуг, %	OS _i	80,4	80,4	80,4	80,4	80,4	80,4
	Средняя сумма за дополнительные услуги, руб.	OSV _i	1 623	1 623	1 623	1 623	1 623	1 623
	Проникновение дополнительных услуг при продлении, %	OSR _i	83,8	83,8	83,8	83,8	83,8	83,8
	Проникновения продления, %	RN _i	39,8	39,8	39,8	39,8	39,8	39,8
	Средний срок пролонгаций у продливших, дней	ATR _i	28,7	28,7	28,7	28,7	28,7	28,7
	Среднее количество пролонгаций у продливших, ед.	NR _i	2,0	2,0	2,0	2,0	2,0	2,0
	Чистые кредитные потери по показателю NPL10 net, %	NPL _i	26,5	26,5	26,5	26,5	26,5	26,5
	Количество заявок в СПР, ед.	CRN _i	103 801	103 801	103 801	103 801	103 801	103 801
	Одобрено заявок в СПР, ед.	ACRN _i	31 883	31 883	31 883	31 883	31 883	31 883
	Количество выдач, ед.	CN _i	23 347	23 347	23 347	23 347	23 347	23 347

	Номер когорты, i		19	20	21	22	23	24
	Период		окт.24	ноя.24	дек.24	янв.25	фев.25	мар.25
	Показатель	Обозначение						
Константные значения бизнес-метрик за соответствующий месяц	Средняя ставка в день, %	AI_i	0,8	0,8	0,8	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	$OSRV_i$	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF_i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявке, руб.	SC_i	50,0	50,0	50,0	50,0	50,0	50,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC_i	2 000	2 000	2 000	2 000	2 000	2 000
	Показатель возврата клиентов (Retention Rate) из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC_i	5	5	5	5	5	5
Фактические значения бизнес-метрик за соответствующий месяц	Средний чек, руб.	AD_i	10 300	10 300	10 300	10 300	10 300	10 300
	Средний срок по договору, дней	AT_i	21,9	21,9	21,9	21,9	21,9	21,9
	Уровень досрочного погашения (от срока), %	ER_i	35,8	35,8	35,8	35,8	35,8	35,8
	Проникновения дополнительных услуг, %	OS_i	80,4	80,4	80,4	80,4	80,4	80,4
	Средняя сумма за дополнительные услуги, руб.	OSV_i	1 623	1 623	1 623	1 623	1 623	1 623
	Проникновение дополнительных услуг при продлении, %	OSR_i	83,8	83,8	83,8	83,8	83,8	83,8
	Проникновения продления, %	RN_i	39,8	39,8	39,8	39,8	39,8	39,8
	Средний срок пролонгаций у продливших, дней	ATR_i	28,7	28,7	28,7	28,7	28,7	28,7
	Среднее количество пролонгаций у продливших, ед.	NR_i	2,0	2,0	2,0	2,0	2,0	2,0
	Чистые кредитные потери по показателю $NPL_{10\text{net}}$, %	NPL_i	26,5	26,5	26,5	26,5	26,5	26,5
	Количество заявок в СПР, ед.	CRN_i	103 801	103 801	103 801	103 801	103 801	103 801
	Одобрено заявок в СПР, ед.	$ACRN_i$	31 883	31 883	31 883	31 883	31 883	31 883
	Количество выдач, ед.	CN_i	23 347	23 347	23 347	23 347	23 347	23 347

Источник: составлено автором

Таблица Г.2 – Расчетные показатели, необходимые для дальнейшей оценки юнит-экономики первичного займа МФО

Номер когорты, i		1	2	3	4	5	6
Период		апр.23	май.23	июн.23	июл.23	авг.23	сен.23
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP _i	1 860	1 899	2 071	1 813	1 880	1 932
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	AIP _Г	2 639	2 658	2 846	2 245	2 460	2 514
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T _Г	26,4	26,3	27,9	27,2	29,6	29,9
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR _i	28,5	29,2	31,9	30,2	30,5	29,4
Конверсия из одобрения в выдачу (Take Rate), %	TR _i	71,4	76,3	75,5	73,4	66,7	72,7
Множитель воронки	PFR _i	4,9	4,5	4,1	4,5	4,9	4,7
Затраты на скоринг одной выдачи, руб.	LSC _i	245	225	207	225	246	234
Номер когорты, i		7	8	9	10	11	12
Период		окт.23	ноя.23	дек.23	январ.24	фев.24	мар.24
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP _i	2 052	1 808	1 915	1 852	1 895	1 764
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	AIP _Г	2 559	2 275	2 414	2 401	2 458	2 353
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T _Г	30,2	28,4	29,9	29,4	29,8	28,3
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR _i	31,1	32,3	34,7	29,8	30,1	30,2
Конверсия из одобрения в выдачу (Take Rate), %	TR _i	70,9	75,6	74,4	71,3	76,4	74,7
Множитель воронки	PFR _i	4,5	4,1	3,9	4,7	4,4	4,4
Затраты на скоринг одной выдачи, руб.	LSC _i	227	205	194	235	218	221
Номер когорты, i		13	14	15	16	17	18
Период		апр.24	май.24	июн.24	июл.24	авг.24	сен.24
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP _i	1 805	1 805	1 805	1 805	1 805	1 805
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	AIP _Г	2 357	2 357	2 357	2 357	2 357	2 357
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T _Г	28,6	28,6	28,6	28,6	28,6	28,6
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR _i	30,7	30,7	30,7	30,7	30,7	30,7
Конверсия из одобрения в выдачу (Take Rate), %	TR _i	73,2	73,2	73,2	73,2	73,2	73,2
Множитель воронки	PFR _i	4,4	4,4	4,4	4,4	4,4	4,4
Затраты на скоринг одной выдачи, руб.	LSC _i	222	222	222	222	222	222

Номер когорты, i		19	20	21	22	23	24
Период		окт.24	ноя.24	дек.24	январ.25	фев.25	мар.25
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP _i	1 805	1 805	1 805	1 805	1 805	805
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	AIP _{Ti}	2 357	2 357	2 357	2 357	2 357	2 357
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T _{Ti}	28,6	28,6	28,6	28,6	28,6	28,6
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR _i	30,7	30,7	30,7	30,7	30,7	30,7
Конверсия из одобрения в выдачу (Take Rate), %	TR _i	73,2	73,2	73,2	73,2	73,2	73,2
Множитель воронки	PFR _i	4,4	4,4	4,4	4,4	4,4	4,4
Затраты на скоринг одной выдачи, руб.	LSC _i	222	222	222	222	222	222

Источник: составлено автором

Таблица Г.3 – Расчет юнит-экономики первичного займа МФО

Номер когорты, i		1	2	3	4	5	6	7	8
Период		апр.23	май.23	июн.23	июл.23	авг.23	сен.23	окт.23	ноя.23
Показатель	Обознач.								
Средняя выручка с первого займа, руб.	AR _i	5 994	6 031	6 003	4 971	5 801	5 504	5 780	5 470
Средняя комиссия, руб.	API _i	1 860	1 899	2 071	1 813	1 880	1 932	2 052	1 808
Вклад кросс-продуктов, руб.	CS _i	1 689	1 746	1 662	1 575	1 786	1 654	1 729	1 827
Вклад пролонгации, руб.	Roi	2 444	2 386	2 270	1 583	2 135	1 918	1 999	1 835
Средние затраты на привлечение и обслуживание первого займа, руб. в т. ч.	AC _i	– 5 846	– 6 025	– 6 185	– 6 052	– 6 171	– 6 223	– 6 293	– 5 947
Расходы на привлечение, руб.	ACC _i	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000
Кредитные потери, руб.	CL _i	– 2 977	– 3 180	– 3 350	– 3 198	– 3 291	– 3 357	– 3 429	– 3 117
Затраты на скоринг, руб.	LSC _i	– 245	– 225	– 207	– 225	– 246	– 234	– 227	– 205
Расход на перевод денег, руб.	MTC _i	– 148	– 145	– 149	– 148	– 152	– 150	– 152	– 144
Средние постоянные издержки на 1 займ, руб.	AFC _i	– 200	– 200	– 200	– 200	– 200	– 200	– 200	– 200
Средние расходы на взыскание на 1 займ, руб.	ADCC _i	– 250	– 250	– 250	– 250	– 250	– 250	– 250	– 250
Процентные расходы (фондирование), руб.	IC _i	– 25	– 26	– 28	– 31	– 32	– 33	– 35	– 31
Юнит-экономика первого займа	UE _i	148	5	– 182	– 1 081	– 370	– 719	– 513	– 477
Номер когорты, i		9	10	11	12	13	14	15	16
Период		дек.23	январ.24	фев.24	мар.24	апр.24	май.24	июн.24	июл.24
Показатель	Обознач.								
Средняя выручка с первого займа, руб.	AR _i	5 468	5 688	5 749	5 315	5 416	5 416	5 416	5 416
Средняя комиссия, руб.	API _i	1 915	1 852	1 895	1 764	1 805	1 805	1 805	1 805
Вклад кросс-продуктов, руб.	CS _i	1 728	1 759	1 791	1 725	1 715	1 715	1 715	1 715
Вклад пролонгации, руб.	Roi	1 825	2 077	2 062	1 826	1 896	1 896	1 896	1 896
Средние затраты на привлечение и обслуживание первого займа, руб. в т. ч.	AC _i	– 6 056	– 5 987	– 6 031	– 6 112	– 6 060	– 6 060	– 6 060	– 6 060
Расходы на привлечение, руб.	ACC _i	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000	– 2 000
Кредитные потери, руб.	CL _i	– 3 232	– 3 122	– 3 183	– 3 260	– 3 208	– 3 208	– 3 208	– 3 208
Затраты на скоринг, руб.	LSC _i	– 194	– 235	– 218	– 221	– 222	– 222	– 222	– 222
Расход на перевод денег, руб.	MTC _i	– 147	– 149	– 148	– 150	– 149	– 149	– 149	– 149
Средние постоянные издержки на 1 займ, руб.	AFC _i	– 200	– 200	– 200	– 200	– 200	– 200	– 200	– 200
Средние расходы на взыскание на 1 займ, руб.	ADCC _i	– 250	– 250	– 250	– 250	– 250	– 250	– 250	– 250
Процентные расходы (фондирование), руб.	IC _i	– 33	– 32	– 32	– 30	– 31	– 31	– 31	– 31
Юнит-экономика первого займа	UE _i	– 588	– 298	– 282	– 797	– 644	– 644	– 644	– 644

Источник: составлено автором

[illegible]

Приложение Д

(обязательное)

Расчеты юнит-экономики повторного займа МФО

Таблица Д.1 – Исходные данные для проведения расчетов

	Номер когорты, i		1	2	3	4	5	6
	Период		апр.23	май.23	июн.23	июл.23	авг.23	сен.23
	Показатель	Обозначение						
Константные значения бизнес– метрик за соответствующий месяц	Средняя ставка в день, %	AI _i	1,0	1,0	1,0	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	OSRV _i	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF _i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявке, руб.	SC _i	15,0	15,0	15,0	15,0	15,0	15,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC _i	40	40	40	40	40	40
	Показатель возврата клиентов (Retention Rate) из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC _i	5	5	5	5	5	5
Фактические значения бизнес– метрик за соответствующий месяц	Средний чек, руб.	AD _i	15 000	15 100	15 200	15 300	15 400	15 500
	Средний срок по договору, дней	AT _i	21,5	21,7	22,4	23,1	23,5	24,0
	Уровень досрочного погашения (от срока), %	ER _i	40,1	40,4	40,7	41,1	40,7	41,0
	Проникновения дополнительных услуг, %	OS _i	90,2	91,5	89,3	90,1	92,1	90,3
	Средняя сумма за дополнительные услуги, руб.	OSV _i	1 555	1 686	1 609	1 495	1 623	1 597
	Проникновение дополнительных услуг при продлении, %	OSR _i	82,4	82,4	82,4	82,4	82,4	82,4
	Проникновения продления, %	RN _i	30,8	30,8	30,8	30,8	30,8	30,8
	Средний срок пролонгаций у продливших, дней	ATR _i	31,6	32,6	30,9	32,9	31,8	32,5
	Среднее количество пролонгаций у продливших, ед.	NR _i	1,9	1,8	1,7	1,7	2,0	1,7
	Чистые кредитные потери по показателю NPL10 net, %	NPL _i	12,7	13,4	13,8	13,3	13,5	13,7
	Количество заявок в СПР, ед.	CRN _i	42 454	40 842	39 678	40 842	40 395	42 417
	Одобрено заявок в СПР, ед.	ACRN _i	37 240	35 826	34 805	35 826	35 434	37 208

	Количество выдач, ед.	CN _i	34 376	32 959	32 347	33 234	32 809	34 452
	Номер когорты, i		7	8	9	10	11	12
	Период		окт.23	ноя.23	дек.23	янв.24	фев.24	мар.24
	Показатель	Обозначение						
Константные значения бизнес-метрики за соответствующий месяц	Средняя ставка в день, %	AI _i	0,8	0,8	0,8	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	OSRV _i	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF _i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявке, руб.	SC _i	15,0	15,0	15,0	15,0	15,0	15,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC _i	40	40	40	40	40	40
	Показатель возврата клиентов (Retention Rate) из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC _i	5	5	5	5	5	5
Фактические значения бизнес-метрики за соответствующий месяц	Средний чек, руб.	AD _i	15 600	15 000	15 100	15 200	15 300	15 400
	Средний срок по договору, дней	AT _i	24,7	23,6	24,5	23,9	24,0	22,3
	Уровень досрочного погашения (от срока), %	ER _i	41,8	40,2	41,0	40,8	43,2	40,6
	Проникновения дополнительных услуг, %	OS _i	90,5	90,2	89,3	91,6	90,4	89,8
	Средняя сумма за дополнительные услуги, руб.	OSV _i	1 609	1 811	1 713	1 587	1 701	1 724
	Проникновение дополнительных услуг при продлении, %	OSR _i	82,4	82,4	82,4	82,4	82,4	82,4
	Проникновения продления, %	RN _i	30,8	30,8	30,8	30,8	30,8	30,8
	Средний срок пролонгаций у продливших, дней	ATR _i	30,4	30,7	31,4	31,5	32,0	31,3
	Среднее количество пролонгаций у продливших, ед.	NR _i	1,9	1,9	1,8	2,0	1,9	1,7
	Чистые кредитные потери по показателю NPL10 net, %	NPL _i	13,7	13,3	13,6	13,1	13,2	13,5
	Количество заявок в СПР, ед.	CRN _i	42 043	44 428	46 557	39 623	41 075	41 627
	Одобрено заявок в СПР, ед.	ACRN _i	36 880	38 972	40 840	34 757	36 031	36 515
	Количество выдач, ед.	CN _i	34 148	36 085	37 815	32 183	33 362	33 810

	Номер когорты, i		13	14	15	16	17	18
	Период		апр.24	май.24	июн.24	июл.24	авг.24	сен.24
	Показатель	Обозначение						
Константные значения бизнес– метрик за соответствующий месяц	Средняя ставка в день, %	AI _i	0,8	0,8	0,8	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	OSRV _i	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF _i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявке, руб.	SC _i	15,0	15,0	15,0	15,0	15,0	15,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC _i	40	40	40		40	40
	Показатель возврата клиентов (Retention Rate) из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC _i	5	5	5	5	5	5
Фактические значения бизнес– метрик за соответствующий месяц	Средний чек, руб.	AD _i	15 500	15 600	15 000	15 100	15 200	15 300
	Средний срок по договору, дней	AT _i	22,9	23,0	22,8	22,9	23,0	23,1
	Уровень досрочного погашения (от срока), %	ER _i	41,3	40,2	42,2	40,6	41,3	40,3
	Проникновения дополнительных услуг, %	OS _i	90,4	90,4	90,4	90,4	90,4	90,4
	Средняя сумма за дополнительные услуги, руб.	OSV _i	1 643	1 643	1 643	1 643	1 643	1 643
	Проникновение дополнительных услуг при продлении, %	OSR _i	82,4	82,4	82,4	82,4	82,4	82,4
	Проникновения продления, %	RN _i	30,8	30,8	30,8	30,8	30,8	30,8
	Средний срок пролонгаций у продливших, дней	ATR _i	31,2	31,2	31,2	31,2	31,2	31,2
	Среднее количество пролонгаций у продливших, ед.	NR _i	1,8	1,8	1,8	1,8	1,8	1,8
	Чистые кредитные потери по показателю NPL10 net, %	NPL _i	13,4	13,4	13,4	13,4	13,4	13,4
	Количество заявок в СПР, ед.	CRN _i	41 832	41 832	41 832	41 832	41 832	41 832
	Одобрено заявок в СПР, ед.	ACRN _i	36 695	36 695	36 695	36 695	36 695	36 695
	Количество выдач, ед.	CN _i	33 976	33 976	33 976	33 976	33 976	33 976

	Номер когорты, i		19	20	21	22	23	24
	Период		окт.24	ноя.24	дек.24	янв.25	фев.25	мар.25
	Показатель	Обозначение						
Константные значения бизнес– метрик за соответствующий месяц	Средняя ставка в день, %	AI _i	0,8	0,8	0,8	0,8	0,8	0,8
	Сумма дополнительных услуг при продлении займа, руб.	OSRV _i	1 000	1 000	1 000	1 000	1 000	1 000
	Комиссия за перевод денег и банковские расходы, %	MTF _i	1	1	1	1	1	1
	Средняя стоимость принятия решения по одной заявке, руб.	SC _i	15,0	15,0	15,0	15,0	15,0	15,0
	Средняя стоимость привлечения одной выдачи, руб.	ACC _i	40	40	40	40	40	40
	Показатель возврата клиентов (Retention Rate) из выдачи в выдачу на горизонте 24 мес., %	RR	250	250	250	250	250	250
	Ставки за фондирование или инвестирование, %	IC _i	5	5	5	5	5	5
Фактические значения бизнес– метрик за соответствующий месяц	Средний чек, руб.	AD _i	15 400	15 500	15 600	15 600	15 600	15 600
	Средний срок по договору, дней	AT _i	23,2	23,4	23,5	23,5	23,5	23,5
	Уровень досрочного погашения (от срока), %	ER _i	42,3	40,2	41,7	41,7	41,7	41,7
	Проникновения дополнительных услуг, %	OS _i	90,4	90,4	90,4	90,4	90,4	90,4
	Средняя сумма за дополнительные услуги, руб.	OSV _i	1 643	1 643	1 643	1 643	1 643	1 643
	Проникновение дополнительных услуг при продлении, %	OSR _i	82,4	82,4	82,4	82,4	82,4	82,4
	Проникновения продления, %	RN _i	30,8	30,8	30,8	30,8	30,8	30,8
	Средний срок пролонгаций у продливших, дней	ATR _i	31,2	31,2	31,2	31,2	31,2	31,2
	Среднее количество пролонгаций у продливших, ед.	NR _i	1,8	1,8	1,8	1,8	1,8	1,8
	Чистые кредитные потери по показателю NPL10 net, %	NPL _i	13,4	13,4	13,4	13,4	13,4	13,4
	Количество заявок в СПР, ед.	CRN _i	41 832	41 832	41 832	41 832	41 832	41 832
	Одобрено заявок в СПР, ед.	ACRN _i	36 695	36 695	36 695	36 695	36 695	36 695
	Количество выдач, ед.	CN _i	33 976	33 976	33 976	33 976	33 976	33 976

Источник: составлено автором

Таблица Д.2 – Расчетные показатели, необходимые для дальнейшей оценки юнит-экономики повторного займа МФО

Номер когорты, i		1	2	3	4	5	6
Период		апр.23	май.23	июн.23	июл.23	авг.23	сен.23
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP _i	3 225	3 277	3 405	2 827	2 895	2 976
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	AIP _{гi}	3 789	3 876	3 895	3 264	3 287	3 372
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T _{гi}	25,3	25,7	25,6	26,7	26,7	27,2
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR _i	87,7	87,7	87,7	87,7	87,7	87,7
Конверсия из одобрения в выдачу (Take Rate), %	TR _i	92,3	92,0	92,9	92,8	92,6	92,6
Множитель воронки	PFR _i	1,2	1,2	1,2	1,2	1,2	1,2
Затраты на скоринг одной выдачи, руб.	LSC _i	19	19	18	18	18	18
Номер когорты, i		7	8	9	10	11	12
Период		окт.23	ноя.23	дек.23	январ.24	февр.24	мар.24
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP _i	3 083	2 832	2 960	2 906	2 938	2 747
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	AIP _{гi}	3 360	3 180	3 288	3 264	3 267	3 161
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T _{гi}	26,9	26,5	27,2	26,8	26,7	25,7
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR _i	87,7	87,7	87,7	87,7	87,7	87,7
Конверсия из одобрения в выдачу (Take Rate), %	TR _i	92,6	92,6	92,6	92,6	92,6	92,6
Множитель воронки	PFR _i	1,2	1,2	1,2	1,2	1,2	1,2
Затраты на скоринг одной выдачи, руб.	LSC _i	18	18	18	18	18	18
Номер когорты, i		13	14	15	16	17	18
Период		апр.24	май.24	июн.24	июл.24	авг.24	сен.24
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP _i	2 840	2 870	2 736	2 766	2 797	2 827
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	AIP _{гi}	3 219	3 269	3 088	3 148	3 164	3 215
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T _{гi}	26,0	26,2	25,7	26,1	26,0	26,3
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR _i	87,7	87,7	87,7	87,7	87,7	87,7
Конверсия из одобрения в выдачу (Take Rate), %	TR _i	92,6	92,6	92,6	92,6	92,6	92,6
Множитель воронки	PFR _i	1,2	1,2	1,2	1,2	1,2	1,2
Затраты на скоринг одной выдачи, руб.	LSC _i	18	18	18	18	18	18

Номер когорты, i		19	20	21	22	23	24
Период		окт.24	ноя.24	дек.24	январ.25	фев.25	мар.25
Показатель	Обозначение						
Средняя прибыль от процентов с одного займа, руб.	AIP_i	2 858	2 902	2 933	2 933	2 933	2 933
Реальная средняя прибыль с учетом пролонгации и досрочных погашений, руб.	$AIPr_i$	3 204	3 285	3 285	3 285	3 285	3 285
Реальный средний срок с учетом пролонгации и досрочных погашений, дней	T_{Ti}	26,0	26,5	26,3	26,3	26,3	26,3
Конверсия из заявки в СПР в одобрение (Approval Rate), %	AR_i	87,7	87,7	87,7	87,7	87,7	87,7
Конверсия из одобрения в выдачу (Take Rate), %	TR_i	92,6	92,6	92,6	92,6	92,6	92,6
Множитель воронки	PFR_i	1,2	1,2	1,2	1,2	1,2	1,2
Затраты на скоринг одной выдачи, руб.	LSC_i	18	18	18	18	18	18

Источник: составлено автором

Таблица Д.3 – Расчет юнит-экономики повторного займа МФО

Номер когорты, i		1	2	3	4	5	6	7	8
Период		апр.23	май.23	июн.23	июл.23	авг.23	сен.23	окт.23	ноя.23
Показатель	Обознач.								
Средняя выручка с повторного займа, руб.	AR_i	7 602	7 719	7 448	6 446	7 012	6 669	6 949	6 779
Средняя комиссия, руб.	AIP _i	3 225	3 277	3 405	2 827	2 895	2 976	3 083	2 832
Вклад кросс-продуктов, руб.	CS _i	1 604	1 695	1 580	1 509	1 703	1 585	1 647	1 789
Вклад пролонгации, руб.	Roi	2 773	2 748	2 463	2 109	2 414	2 108	2 220	2 158
Средние затраты на привлечение и обслуживание повторного займа, руб. в т. ч.	AC_i	– 3 101	– 3 251	– 3 360	– 3 212	– 3 275	– 3 320	– 3 360	– 3 170
Расходы на привлечение, руб.	ACC _i	– 40	– 40	– 40	– 40	– 40	– 40	– 40	– 40
Кредитные потери, руб.	CL _i	– 2 313	– 2 462	– 2 568	– 2 419	– 2 479	– 2 522	– 2 560	– 2 380
Затраты на скоринг, руб.	LSC _i	– 19	– 19	– 18	– 18	– 18	– 18	– 18	– 18
Расход на перевод денег, руб.	MTC _i	– 236	– 236	– 237	– 236	– 238	– 239	– 239	– 233
Средние постоянные издержки на 1 займ, руб.	AFC _i	– 200	– 200	– 200	– 200	– 200	– 200	– 200	– 200
Средние расходы на взыскание на 1 займ, руб.	ADCC _i	– 250	– 250	– 250	– 250	– 250	– 250	– 250	– 250
Процентные расходы (фондирование), руб.	IC _i	– 44	– 45	– 47	– 48	– 50	– 51	– 53	– 48
Юнит-экономика повторного займа	UE_i	4 501	4 468	4 088	3 234	3 738	3 349	3 589	3 609
Номер когорты, i		9	10	11	12	13	14	15	16
Период		дек.23	январ.24	фев.24	мар.24	апр.24	май.24	июн.24	июл.24
Показатель	Обознач.								
Средняя выручка с повторного займа, руб.	AR_i	6 761	6 935	6 944	6 434	6 647	6 691	6 473	6 518
Средняя комиссия, руб.	AIP _i	2 960	2 906	2 938	2 747	2 840	2 870	2 736	2 766
Вклад кросс-продуктов, руб.	CS _i	1 684	1 671	1 712	1 670	1 648	1 648	1 648	1 648
Вклад пролонгации, руб.	Roi	2 118	2 358	2 294	2 017	2 159	2 173	2 089	2 103
Средние затраты на привлечение и обслуживание повторного займа, руб. в т. ч.	AC_i	– 3 248	– 3 164	– 3 198	– 3 251	– 3 252	– 3 274	– 3 160	– 3 183
Расходы на привлечение, руб.	ACC _i	– 40	– 40	– 40	– 40	– 40	– 40	– 40	– 40
Кредитные потери, руб.	CL _i	– 2 456	– 2 371	– 2 406	– 2 459	– 2 457	– 2 475	– 2 376	– 2 394
Затраты на скоринг, руб.	LSC _i	– 18	– 18	– 18	– 18	– 18	– 18	– 18	– 18
Расход на перевод денег, руб.	MTC _i	– 233	– 235	– 233	– 237	– 238	– 242	– 229	– 233
Средние постоянные издержки на 1 займ, руб.	AFC _i	– 200	– 200	– 200	– 200	– 200	– 200	– 200	– 200
Средние расходы на взыскание на 1 займ, руб.	ADCC _i	– 250	– 250	– 250	– 250	– 250	– 250	– 250	– 250
Процентные расходы (фондирование), руб.	IC _i	– 51	– 50	– 50	– 47	– 49	– 49	– 47	– 47
Юнит-экономика повторного займа	UE_i	3 513	3 771	3 746	3 183	3 394	3 418	3 313	3 335

Номер когорты, i		17	18	19	20	21	22	23	24
Период		авг.24	сен.24	окт.24	ноя.24	дек.24	январ.25	фев.25	мар.25
Показатель	Обознач.								
Средняя выручка с повторного займа, руб.	AR _i	6 562	6 607	6 651	6 709	6 754	6 754	6 754	6 754
Средняя комиссия, руб.	AIP _i	2 797	2 827	2 858	2 902	2 933	2 933	2 933	2 933
Вклад кросс-продуктов, руб.	CS _i	1 648	1 648	1 648	1 648	1 648	1 648	1 648	1 648
Вклад пролонгации, руб.	Roi	2 117	2 131	2 145	2 159	2 173	2 173	2 173	2 173
Средние затраты на привлечение и обслуживание повторного займа, руб. в т. ч.	AC _i	– 3 201	– 3 223	– 3 238	– 3 264	– 3 281	– 3 281	– 3 281	– 3 281
Расходы на привлечение, руб.	ACC _i	– 40	– 40	– 40	– 40	– 40	– 40	– 40	– 40
Кредитные потери, руб.	CL _i	– 2 411	– 2 429	– 2 446	– 2 465	– 2 483	– 2 483	– 2 483	– 2 483
Затраты на скоринг, руб.	LSC _i	– 18	– 18	– 18	– 18	– 18	– 18	– 18	– 18
Расход на перевод денег, руб.	MTC _i	– 234	– 237	– 235	– 240	– 239	– 239	– 239	– 239
Средние постоянные издержки на 1 заем, руб.	AFC _i	– 200	– 200	– 200	– 200	– 200	– 200	– 200	– 200
Средние расходы на взыскание на 1 заем, руб.	ADCC _i	– 250	– 250	– 250	– 250	– 250	– 250	– 250	– 250
Процентные расходы (фондирование), руб.	IC _i	– 48	– 48	– 49	– 50	– 50	– 50	– 50	– 50
Юнит-экономика повторного займа	UE _i	3 361	3 384	3 413	3 445	3 473	3 473	3 473	3 473

Источник: составлено автором

Приложение Е

(справочное)

Листинг кода программы расчета признака качества кредитной истории в среде Python

```
import pandas as pd
import re
import numpy as np
from dateutil.relativedelta import relativedelta
```

```
def get_chq(data: pd.DataFrame) -> pd.Series:
```

```
    """Функция для расчета значений признака ККИ для заданного набора
    заемщиков
```

```
    Parameters
```

```
    -----
```

```
    data : pd.DataFrame
```

```
        Исходные данные, содержащие поля:
```

1. ApplicationID - внутренний уникальный идентификатор клиента (заемщика);
2. Платежная_дисциплина - строка "платежной дисциплины";
3. Код_вида_займа - код вида займа, специфичный для формата КО Эквивалент;
4. Тип_партнера - тип партнера БКИ, у которого заемщиком было открыто кредитное обязательство;
5. Статус_кредита - статус договора обязательства на текущий момент;
6. Дата_закрытия_факт - фактическая дата закрытия договора кредитного обязательства;
7. Дата_заявки - внутренняя дата заявки на открытие займа;
8. Дата_последнего_изменения_записи - дата последнего изменения записи о кредитном обязательстве;
9. УИД_сделки - уникальный универсальный идентификатор сделки.

```
    Returns
```

```
    -----
```

```
    pd.Series
```

```
        Значения признака
```

```
    """
```

```
# класс названий используемых в вычислениях полей кредитной истории
```

```

class ChqCols:
    uid = 'УИД_сделки'
    period = 'Платежная_дисциплина' # "платежная дисциплина", строка
платежей
    period_cleansed = 'period_cleansed' # очищенная от символов "-" и "С"
строка платежей
    deal_type = 'Код_вида_займа'
    partner_type = 'Тип_партнера'
    loan_status = 'Статус_кредита'
    date_of_factual_closure = 'Дата_закрытия_факт'
    ch_relevance_date = 'Дата_заявки'
    loan_relevance_date = 'Дата_последнего_изменения_записи'

# класс возможных нецифровых символов "платежной дисциплины"
формата КО БКИ "Эквивалент"
class PaymentDisciplineSymbols:
    FINAL = 'BCIRSTW'
    FINAL_W_OVERDUE = 'BIRSTW'

# класс возможных статусов кредитных обязательств
class LoanStatus:
    closed = 'Закрыт'
    info_stopped = 'Передача информации прекращена'
    active_wo_overdue = 'Активный без просрочки'
    active_w_overdue = 'Активный с просрочкой'

COLS = ChqCols
PMT_STATUSES = PaymentDisciplineSymbols
LOAN_STATUSES = LoanStatus

def clear_pmt_line(line: str) -> str:
    """Функция для очистки строки "платежной дисциплины" от
неиспользуемых при
    вычислении значения признака символов и дублированных конечных
статусов

    Parameters
    -----
    line : str

```

Строка "платежной дисциплины"

Returns

str

Очищенная от ненужных символов строка "платежной дисциплины"
 """

```
if (not line) or pd.isnull(line):
    return "
```

```
line = re.sub('[-]+' , "", line)
```

```
for status in PMT_STATUSES.FINAL:
    line = re.sub(f'^[{status}-]+' , f'{status}', line.upper())
```

```
return line
```

```
def weights_by_loan_type_and_state(d: pd.DataFrame, microloans_only: bool =
False) -> pd.Series:
```

"""Функция для задания корректирующих коэффициентов, зависящих от
 типа кредита,
 кредитным обязательствам заемщика

Parameters

d : pd.DataFrame

Набор данных, содержащий поля кредитного отчета, необходимые для
 расчета признака,

на уровне одного заемщика

microloans_only : bool, optional

Флаг, принимающий значение True в случае если в расчете значения
 признака учитываются

только микрокредиты, by default False

Returns

pd.Series

Вектор значений корректирующих весов для каждого кредитного
 обязательства клиента

"""

```

# создание фильтра по типу кредитного обязательства (используются 2
типа: микрозайм, иной)
microloans = (d[COLS.deal_type] == 3) \
    | ((d[COLS.partner_type] == 3) &
        ((d[COLS.deal_type] == 99) | (d[COLS.deal_type] == 8)))

# создание фильтров по статусам договоров кредитных обязательств на
текущий момент
# активные без просрочки
active_wo_overdue = (d[COLS.loan_status] ==
LOAN_STATUSES.active_wo_overdue)
# закрытые
closed = (d[COLS.loan_status] == LOAN_STATUSES.closed)
# активные с просрочкой
active_w_overdue = (d[COLS.loan_status] ==
LOAN_STATUSES.active_w_overdue)
# иные состояния, соответствующие кодам BIRSTW, согласно
обозначениям БКИ "Эквифакс"
finished_badly = d[COLS.period_cleansed].str.translate(
    str.maketrans({' ' for ch in
PMT_STATUSES.FINAL_W_OVERDUE})).str.contains('_')

# создание маппера, ставящего в соответствие каждому возможному
статусу договора
# кредитного обязательства значение корректирующего коэффициента
choices = {
    4: microloans & active_wo_overdue,
    5: microloans & closed,
    10: microloans & (active_w_overdue | (finished_badly
        & ~closed
        & ~active_wo_overdue))
}
if not microloans_only:
    choices.update({
        1: ~microloans & (active_wo_overdue | closed),
        3: ~microloans & (active_w_overdue | (finished_badly
            & ~closed
            & ~active_wo_overdue)),
    })

return pd.Series(np.select(list(choices.values()), list(choices.keys())),
    index=d.index, name='loan_weights', dtype='float64').fillna(0)

```

```

def loan_age_in_months(s: pd.Series) -> int:
    """Функция для расчета количества месяцев от даты запроса КИ до даты
    закрытия или
    даты последнего обновления информации (в случае если обязательство
    закрыто, а дата
    закрытия отсутствует)

    Parameters
    -----
    s : pd.Series
        Набор данных, содержащий поля кредитного отчета, необходимые для
        расчета признака,
        на уровне одного кредитного обязательства заемщика

    Returns
    -----
    int
        "Возраст" кредитного обязательства в виде целого числа месяцев
    """

    # если статус обязательства "Активный", возвращаем 0
    if s[COLS.loan_status] in {
        LOAN_STATUSES.active_wo_overdue,
        LOAN_STATUSES.active_w_overdue
    }:
        return 0
    # иначе считаем количество месяцев
    else:
        last_date = s[COLS.date_of_factual_closure] or
s[COLS.loan_relevance_date]
        dates_are_not_none = (pd.notnull(s[COLS.ch_relevance_date]) and
pd.notnull(last_date))
        # заполнение пропусков по причине отсутствия дат медианой медиан
        количества месяцев от
        # текущей даты до даты закрытия или даты последнего обновления
        информации
        return relativedelta(s[COLS.ch_relevance_date], last_date).months if
dates_are_not_none else 2

def specific_mean_overdue_by_ovd_line(line: str) -> float:

```

"""Функция для расчета взвешенной просрочки платежей по каждому конкретному кредитному обязательству заемщика

Parameters

line : str

Строка "платежной дисциплины", почищенная от дублированных конечных статусов

Returns

float

Взвешенная просрочка платежей по кредиту

"""

```
def calc_summand(ndx: int, pmt_cat_symbol: str) -> int:
    return (
        (num_of_payments - (ndx + 1) + 1) # чем старше платеж, тем менее
он значим
        * (int(pmt_cat_symbol) if pmt_cat_symbol.isdigit() else 0)
    )

    # удаление из строки "платежной дисциплины" символов "-", "C" и "U"
    line_ = re.sub('[CU-]+', '', line)

    # вычисление длины строки "платежной дисциплины" (количества
платежей по обязательству)
    num_of_payments = len(line_)

    # замена нечисловых разрядов строки "платежной дисциплины" на
значение 10
    line_ = re.sub(f[A{PMT_STATUSES.FINAL_W_OVERDUE}], '10', line_)

    # вычисление значения взвешенной просрочки платежей
    return ((2.0 / (num_of_payments * (num_of_payments + 1)))
            * sum([calc_summand(ndx, symbol) for ndx, symbol in
enumerate(line_)])
            ) if num_of_payments > 0 else np.nan
```

```
def chq(d: pd.DataFrame, microloans_only: bool, col: str) -> pd.DataFrame:
```

"""Функция для расчета значения признака ККИ

Parameters

`d : pd.DataFrame`

Набор данных, содержащий поля кредитного отчета, необходимые для расчета признака,

на уровне одного заемщика

`microloans_only : bool`

Флаг, принимающий значение True в случае если в расчете значения признака учитываются

только микрокредиты, by default False

`col : str`

Название колонки набора данных, которая будет содержать значения признака

Returns

`pd.DataFrame`

Исходный набор данных, содержащий рассчитанное значение признака

"""

```
d[COLS.period_cleansed] = d[COLS.period].apply(clear_pmt_line)
```

```
# вычисление значений весов для каждого кредитного обязательства
# клиента
```

```
loan_weights = weights_by_loan_type_and_state(d,
microloans_only=microloans_only)
```

```
# вычисление "возраста" каждого кредитного обязательства клиента
```

```
loan_ages = d.apply(loan_age_in_months, axis=1)
```

```
# вычисление взвешенной просрочки платежей для каждого кредитного
# обязательства клиента
```

```
loan_arrears =
d[COLS.period_cleansed].apply(specific_mean_overdue_by_ovd_line)
```

```
# если выполняются следующие условия:
```

```
# 1) нет обязательств, по которым можно проставить веса в признак
```

```
# 2) нет платежей, по которым можно посчитать просрочку
```

```
# то возвращаем "пропуск"
```

```
if (loan_weights.sum() == 0) or all(loan_arrears.isna()):
```

```
    d[col] = np.nan
```

```

    return d

    # вычисление значения числителя
    numerator = (loan_weights * loan_arrears / (loan_ages + 1) ** (1 / 2)).sum()

    # вычисление значения знаменателя
    normalizing_denominator = (loan_weights / (loan_ages + 1) ** (1 / 2)).sum()

    # вычисление итогового значения признака
    d[col] = numerator / normalizing_denominator if normalizing_denominator >
0 else np.nan

    return d

columns_to_calc = [
    'ApplicationID', # внутренний уникальный идентификатор клиента
    'Платежная_дисциплина',
    'Код_вида_займа',
    'Тип_партнера',
    'Статус_кредита',
    'Дата_закрытия_факт',
    'Дата_заявки',
    'Дата_последнего_изменения_записи',
    'УИД_сделки'
]

# производим агрегацию данных на уровне одного клиента и рассчитываем
значение признака
data_w_chq = data.groupby(by='ApplicationID')[columns_to_calc].apply(chq,
microloans_only=False, col='chq')
agg_chq = {'chq': lambda s: s.values[0]}
data_collapsed = data_w_chq.groupby(by='ApplicationID').agg(agg_chq)
return data_collapsed

```

Приложение Ж

(справочное)

**Свидетельство о государственной регистрации программы для ЭВМ
«Качество кредитной истории: программа для вычисления рейтинга
заемщика по его кредитной истории для оценки кредитных рисков и
построения скоринговых моделей»**

РОССИЙСКАЯ ФЕДЕРАЦИЯ



СВИДЕТЕЛЬСТВО
о государственной регистрации программы для ЭВМ
№ 2023682976

**Качество кредитной истории: программа для
вычисления рейтинга заемщика по его кредитной
истории для оценки кредитных рисков и построения
скоринговых моделей**

Правообладатель: *Сорокин Александр Сергеевич (RU)*

Автор(ы): *Сорокин Александр Сергеевич (RU)*



Заявка № **2023682030**
Дата поступления **24 октября 2023 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ **01 ноября 2023 г.**

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов

Приложение И

(справочное)

Справки о внедрении и апробации результатов диссертационного исследования

ООО МКК "МОРЕЗАЙМ"
123112, г. Москва, вн.тер.г. муниципальный
округ Пресненский, наб. Пресненская, д.12,
помещ. 6/64
Тел. 8 800 200 50 73
e-mail: adm@s-cash.ru
ОКПО 74373633;
ОГРН 1227700061270;
ИНН/КПП 9701196100/770301001

Исх. №Б\Н от 15.04.2025г.
На Запрос № _____ от _____ г

СПРАВКА О ВНЕДРЕНИИ*результатов диссертационного исследования*

ООО МКК «МОРЕЗАЙМ» подтверждает, что научные исследования Сорокина Александра Сергеевича по теме докторской диссертации «Модели оценки и управления кредитными рисками микрофинансовых организаций» были рассмотрены, одобрены и приняты для внедрения в кредитном конвейере компании, в частности:

- методики построения скоринговых моделей;
- методики управления лимитами на основе разработанных диссертантом моделей;
- предложения по построению системы принятия решений для оценки кредитных рисков;
- методики оценки ценности данных для построения моделей оценки кредитных рисков.

Генеральный директор
ООО МКК «Морезайм»





инко И.С./



СПРАВКА

13 ноября 2023 года

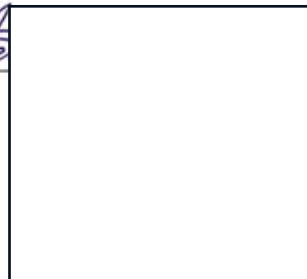
г. Москва

Общество с ограниченной ответственностью «Информационно-аналитическое агентство Финкарта» (ИНН 7714382580) подтверждает, что результаты диссертационной работы Сорокина Александра Сергеевича на тему «Модели оценки и управления кредитными рисками микрофинансовых организаций» на соискание ученой степени доктора экономических наук, использовались в 2021-2022 годах специалистами нашей компании для разработки и внедрения скоринговой модели оценки вероятности будущих платежей заемщика для участников микрофинансового рынка.

Генеральный директор _____

А. Сергеевич

розов)





Общество с ограниченной ответственностью Микрокредитная компания «ФИНМОЛЛ» (ООО МКК «ФИНМОЛЛ»), Р/о, 630112, г. Новосибирск, ул. Фрунзе, д. 242, пом. 406, тел.: +7(383) 210-50-04, ОГРН 1117746509638, ИНН 7705954143, КПП 785250001, р/с 40702810444050004792, к/с 30101810500000000641, БИК 045004641, Сибирский банк ПАО Сбербанк. www.finmoll.com

СПРАВКА

О внедрении результатов исследования, полученных в диссертации Сорокина А.С.

«Модели оценки и управления кредитными рисками микрофинансовых организаций»

Результаты, полученные Сорокиным Александром Сергеевичем в рамках диссертационного исследования на тему «Модели оценки и управления кредитными рисками микрофинансовых организаций», имеют реальное практическое значение для ООО МКК «ФИНМОЛЛ».

К наиболее существенным результатам исследования, реализованным на практике в ООО МКК «ФИНМОЛЛ», относятся:

1. Разработка методики оценки кредитного риска заемщика на основе агрегированного показателя просрочки, рассчитываемого по данным кредитной истории.
2. Разработка методики построения скоринговых моделей на малом объеме выборки с применением метода главных компонент, позволяющей разработать устойчивую скоринговую модель от 300 наблюдений.
3. Разработка методики отбора факторов при построении скоринговых моделей и ее автоматизация на основе «Greeder of Data» – автоматического советника для построения скоринговых моделей, – программы, написанной Сорокиным А.С.

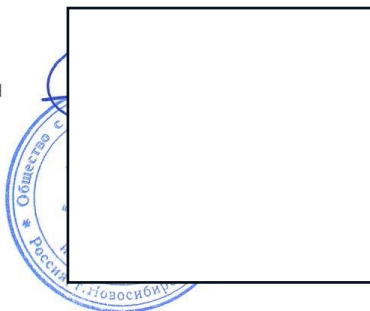
В соответствии с моделями оценки и управления кредитными рисками, изложенными в диссертации Сорокина А.С., были построены две скоринговые модели оценки кредитных рисков первичных и повторных клиентов по продукту PDL.

Внедрение построенных моделей позволило добиться ООО «ФИНМОЛ» положительного экономического эффекта.

Самородов В.Е.

Директор по продажам

13.11.2024





ООО «Займиго МФК», ОГРН 1135260005363, ИНН / КПП 5260355389 / 526001001, Юридический адрес: 603093, г.Н.Новгород, ул.Ковровская, д.21А, пом. 603; e-mail: support@zaymigo.ru; телефон 8-800-700-72-86

Справка

о внедрении результатов диссертационного исследования

Настоящим подтверждаю, что результаты диссертационного исследования Сорокина Александра Сергеевича на тему «Модели оценки и управления кредитными рисками микрофинансовых организаций» обладают актуальностью, представляют практический интерес и были использованы при создании и построении системы риск-менеджмента в 2015-2023 гг. в микрофинансовой компании ООО «Займиго МФК».

Директор
25.  Чураков А.В.


Krediska

ООО «Кредиска МКК», ОГРН 1215200013170, ИНН / КПП 5262378712 / 526001001,
Юридический адрес: 603093, г. Нижний Новгород, ул. Ковровская, д. 21А, пом. 404; mail: support@krediska.ru

СПРАВКА

о внедрении результатов исследования,
полученных в диссертации А.С. Сорокина

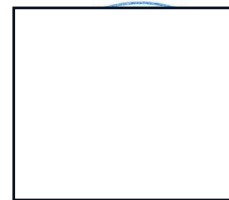
Полученные результаты лично Александром Сергеевичем Сорокиным в рамках диссертационного исследования на тему «Модели оценки и управления кредитными рисками микрофинансовых организаций» имеют реальное практическое значения для компании. Работа выполнялась в период с 2021 по 2023 год и была направлена на построение и мониторинг работы скоринговых моделей. Разработанные модели Сорокиным А.С. внедрены в систему управления кредитными рисками ООО «Кредиска МФК».

К наиболее существенным результатам исследования Сорокина А.С., обладающими научной новизной, можно отнести:

1. Внедрение политики установления кредитных лимитов на основе разработанной математической модели;
2. Внедрение в систему принятия решений компании скоринговых моделей с динамически изменяющимися параметрами.

Директор ООО «Кредиска МКК»

23.01.2024



М.В. Зотов

ОБЛАСТНАЯ ФЕДЕРАЦИЯ

**Hurmacredit**

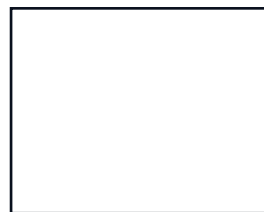
ООО «Хурма Кредит МКК», ОГРН 1235200008240, ИНН/КПП 5262390614/526201001
Юридический адрес 603022, г. Нижний Новгород, ул. Окский Съезд, д. 8, пом. ПЗ, офис 103,
mail: support@hurmacredit.ru

СПРАВКА О ВНЕДРЕНИИ

Настоящая справка подтверждает, что результаты диссертационного исследования Сорокина А.С. «Модели оценки и управления кредитными рисками микрофинансовых организаций» на соискание ученой степени доктора экономических наук, использованы в практике деятельности микрофинансовой компании ООО «Хурма Кредит МКК» в виде:

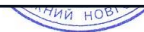
- 1) Использована разработанная в рамках диссертационного исследования классификация источников данных для оценки кредитных рисков и проведено их подключение для оценки кредитных рисков;
- 2) Для оценки кредитных рисков по данным бюро кредитных историй использовался при построении скоринговых моделей агрегированный признак «качество кредитной истории», алгоритм расчета которого представлен в диссертации.

Использование результатов диссертационной работы Сорокина А.С. позволило сформировать ООО «Хурма Кредит МКК» оптимальную систему принятия решений для оценки кредитных рисков в условиях трансформации экономики и изменения регуляторных правил на рынке.



Директор
Катышев А.Н.

25.01.2024





Общество с ограниченной ответственностью
Микрокредитная компания «Бериберу»

ООО «Бериберу МКК», ОГРН 1211600072342

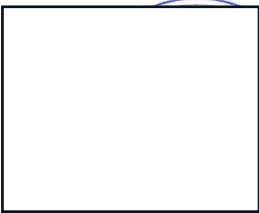
ИНН / КПП 1657270975 / 526001001

Юридический адрес: 603093, г. Н.Новгород, ул. Ковровская, д.21А, пом.401
e-mail: mkk@beriberu.ru

**Справка о внедрении результатов диссертационной работы Сорокина А.С. на тему
«Модели оценки и управления кредитными рисками
микрофинансовых организаций»**

ООО «Бериберу МКК» подтверждает, что научные исследования Сорокина А.С. по теме «Модели оценки и управления кредитными рисками микрофинансовых организаций» были рассмотрены, одобрены и приняты к использованию менеджментом компании, в частности, рекомендованы для внедрения изложенные в диссертационной работе методики:

1. построения скоринговых моделей;
2. установки кредитных лимитов;
3. анализа ценности источников данных для оценки кредитных рисков.



«Бериберу»
Звонкова Г.И.
29.01.2024

ИНН 9724097843 ОГРН 1227700562517
115230, г. Москва, Хлебозаводский пр-д, дом 7, стр. 9,
помещение 33/Н, комната 8, офис 511



СПРАВКА

о внедрении результатов диссертационного исследования

Сорокина Александра Сергеевича

Результаты диссертационного исследования Сорокина А.С. на тему «Модели оценки и управления кредитными рисками микрофинансовых организаций», представленного на соискание ученой степени доктора экономических наук, использованы в практической деятельности ООО МКК «ФинФреш» в виде:

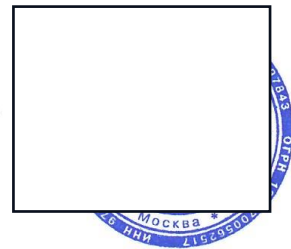
- 1) практических рекомендаций по построению системы принятия решений для оценки кредитных рисков;
- 2) практического применения алгоритмов, моделей, стратегий принятия решений при оценке заемщиков;
- 3) предложений по оптимальному построению системы принятия решений для оценки кредитных рисков и организации бизнес-процессов работы компании.

Использование результатов диссертационной работы Сорокина А.С. позволило выстроить систему управления рисками в ООО МКК «ФинФреш» при выходе компании на рынок микрофинансирования.

Генеральный директор
ООО МКК «ФинФреш»



Бакулина Е. А.



Приложение К

(справочное)

Листинг кода программы «Greeder of Data – автоматический советник для построения скоринговых моделей на базе логистической регрессии на основе больших массивов данных в среде Python»

```
# Импорт необходимых библиотек
import numpy as np
import pandas as pd
import math as mt
import os
import datetime

from datetime import datetime as dt
from datetime import timedelta
import warnings
import shutil
import time
import seaborn as sns

warnings.filterwarnings("ignore", category = FutureWarning)
pd.set_option('display.max_columns', 100)

from optbinning import BinningProcess

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import classification_report
from sklearn.metrics import auc, roc_auc_score, roc_curve
```

```

from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.feature_selection import RFE

```

```

from statsmodels.stats.outliers_influence import variance_inflation_factor
from matplotlib import pyplot as plt
from tqdm import tqdm
from patsy import dmatrices
from random import shuffle

```

```

import statsmodels.api as sm
import openpyxl
import json
from itertools import chain
import gc
import lightgbm as lgb
import seaborn as sns

```

```

class Modeler(object):

```

```

    # Мастер объект модельера. Позволяет осуществить полный цикл
    построения модели,
    # начиная с нахождения оптимальных категорий переменных, заканчивая
    сохранением модели.

```

```

    def __init__(self, data: pd.DataFrame, target: str):

```

Инициализация класса модельера. Необходимо передать DataFrame и название целевой переменной.

В целевой переменной не должно быть пропусков. Признаки должны быть числового формата.

```

self.d = data.replace([-np.inf, np.inf], np.nan).select_dtypes(include =
'number').astype(float)
self.d = self.d.reset_index(drop = True)
if type(target) == str:
    self.target = data[target]
    self.target_name = target
else:
    self.target = target
    self.target_name = target.name
try:
    self.d = self.d.join(self.target, how = 'left')
except:
    pass
self.d = self.d[self.d[self.target_name].notna()]
self.max_pvalue_allow = 0.015
self.max_feature_allow = 5
self.max_vif_allow = 1.5
self.constant = True
self.must_feature = []
pass

```

```
def bin_data(self,
```

Инициализация параметров для оптимальной категоризации количественных предикторов и ее проведение.

```

cols_to_bin: list = None,
min_n_bins: int = None,
max_n_bins: int = None,
min_bin_size: float = 0.1,
min_hit_rate: float = 0.2,
must_event_rate: float = False,
missing_merge_perc: float = 0.0) -> None:

```

```

def get_test_sample(data, target, must_event_rate = False):
    # Изменения доли целевого значения в случае необходимости
    балансировки данных.
    if must_event_rate:
        target_name = target.name
        target_count = target.sum()
        current_event_rate = target_count / data.shape[0]
        if current_event_rate < must_event_rate:
            to_add_sample = int(target_count / must_event_rate) -
            int(target_count)

            t = data.join(target)
            return t[t[target_name] == 1].append(t[t[target_name] ==
0].sample(to_add_sample)).reset_index(drop = True)
        else:
            to_add_sample = int((data.shape[0] - target_count) * must_event_rate
/ (1 - must_event_rate))

            t = data.join(target)
            return t[t[target_name] == 0].append(t[t[target_name] ==
1].sample(to_add_sample)).reset_index(drop = True)

```

```

else:
    return data.join(target).reset_index(drop = True)

def get_new_missing_cat(col):
    # Объединение категорий переменных в зависимости от заданного
    параметра обработки пропусков в данных
    subject = self.optd[col]
    subject = subject[subject.Count != 0]
    woes = subject[subject.Bin.str.contains(',')].WoE
    if len(woes) == 1:
        return 'pass'
    missing_woe = subject[subject.Bin == 'Missing'].WoE.values

    abss = list(abs(woes - missing_woe))
    return abss.index(min(abss))

self.cols_to_bin = cols_to_bin
self.min_n_bins = min_n_bins
self.max_n_bins = max_n_bins
self.min_bin_size = min_bin_size
self.min_hit_rate = min_hit_rate
self.must_event_rate = must_event_rate
self.missing_merge_perc = missing_merge_perc
self.summ, self.optd, self.cat_df, self.bp = self.get_bins(cols_to_bin,
min_n_bins, max_n_bins, min_bin_size, min_hit_rate)
self.test = self.bp.transform(self.d, metric = 'indices', metric_missing = -1)
if self.missing_merge_perc != 0:
    to_miss_merge = (self.d.isna().sum() / self.d.shape[0])

```



```

to_miss_merge_list      =      list(to_miss_merge[(to_miss_merge      <=
missing_merge_perc) & (to_miss_merge != 0)].index)
for col in to_miss_merge_list:
    to_replace_missting = get_new_missing_cat(col)
    if to_replace_missting == 'pass':
        continue
    self.test[col] = self.test[col].replace(-1, to_replace_missting)
self.t = get_test_sample(self.test, self.d[self.target_name], must_event_rate)
self.target = self.t[self.target_name]

```

```
def get_params(self, args):
```

Функция вывода заданных параметров на всех этапах работы модуля
(настройки категоризации, таблицы категоризации и т.д.).

```

to_return = []
if 'opt_info' == args:
    print('Optimal Binning Info')
    print(' cols_to_bin: ', self.cols_to_bin, '\n',
          'min_n_bins: ', self.min_n_bins, '\n',
          'max_n_bins: ', self.max_n_bins, '\n',
          'min_bin_size: ', self.min_bin_size, '\n',
          'min_hit_rate: ', self.min_hit_rate)
    return
if 'best_feature' == args:
    return self.fr
if 'summury' in args:
    to_return.append(self.summ)
if 'optd' in args:
    to_return.append(self.optd)
if 'cat_df' in args:

```

```

    to_return.append(self.cat_df)
    if 'bp' in args:
        to_return.append(self.bp)
    if len(to_return) == 0:
        print('Args not found')
        return
    if type(args) == str:
        return to_return[0]
    return to_return

```

```

def get_bins(self, cols_to_bin, min_n_bins, max_n_bins, min_bin_size,
min_hit_rate):

```

Функция оптимальной категоризации. Убираются переменные в которых количество валидных значений (не пропусков) не советует заданным параметрам. Создаются две таблицы: 1. с общим описание полученных переменных, разбитых на категории. 2. С описанием переменных на разбитые категории (их границы, количество, р-значения и т.д.)

```

    if min_hit_rate:
        hit_table = (self.d.count() / self.d.shape[0]).to_frame()
        d = self.d[hit_table[hit_table[0] >= min_hit_rate].index]
        hit_table = hit_table.reset_index()
        hit_table.columns = ['name', 'hit_rate']
    else:
        hit_table = (self.d.count() / self.d.shape[0]).to_frame()
        d = self.d
        hit_table = hit_table.reset_index()
        hit_table.columns = ['name', 'hit_rate']

```

```

if cols_to_bin:
    x = d[cols_to_bin].values
else:

    x = d.drop(columns = [self.target_name]).values
    cols_to_bin = list(d.columns)
    cols_to_bin.remove(self.target_name)

y = d[self.target_name].values
binning_process = BinningProcess(variable_names = cols_to_bin,
                                  min_n_bins = min_n_bins, max_n_bins = max_n_bins,
                                  min_bin_size = min_bin_size, n_jobs = -1,
max_pvalue_policy = 'all')
    binning_process.fit(x, y)

summary_res = binning_process.summary().merge(hit_table, how = 'left', on
= 'name')
    cols_to_corr = summary_res.name.values
    opt_dtc = {}
    for col in cols_to_corr:
        opt_dtc[col] =
binning_process.get_binned_variable(col).binning_table.build()

new_x = binning_process.transform(x)
new_d = pd.DataFrame(data = new_x, columns = cols_to_bin)
return summary_res, opt_dtc, new_d, binning_process

```

```
def feature_greeder(self, max_pvalue_allow = 0.015, by_score = 'all',
max_feature_allow = 5, max_vif_allow = 1.5, constant = True, must_feature = [],
top = 5000, rounds = 5):
```

Перебор предикторов, удовлетворяющим условиям, для нахождения лучших для модели логистической регрессии. Проверки параметров и их категорий по заданным критериям (VIF, BIC, p-значение). Сохранение набора лучших переменных и оценки модели.

```
self.max_pvalue_allow = max_pvalue_allow
```

```
self.max_feature_allow = max_feature_allow
```

```
self.max_vif_allow = max_vif_allow
```

```
self.by_score = by_score
```

```
self.constant = constant
```

```
self.must_feature = must_feature
```

```
warnings.filterwarnings("ignore", message = 'Maximum Likelihood
optimization failed to ')
```

```
def get_params(best_feature):
```

Функция подбора по заданным критериям признаков. Перебор разных наборов предикторов в заданной очереди (от наилучшего к наихудшему).

```
def get_to_drop_col(col):
```

```
return to_fit.filter(like = col).sum().to_frame().sort_values(by =
0).index[-1]
```

```
y = self.target
```

```
target_name = self.target_name
```

```
res_feature = self.must_feature
```

```
best_bic = 0
```

```
max_pvalue_allow = self.max_pvalue_allow
```

```

max_feature_allow = self.max_feature_allow
max_vif_allow = self.max_vif_allow
constant = self.constant
first_best_feat = True

for feature in best_feature:
    col_to_train = []
    col_to_train = res_feature.copy()
    col_to_train.append(feature)
    to_fit = pd.get_dummies(self.t[col_to_train], columns = col_to_train,
prefix = col_to_train)
    to_drop = list(map(get_to_drop_col, col_to_train))
    to_fit.drop(columns = to_drop, inplace = True)

    if constant:
        to_fit = sm.add_constant(to_fit)
        log_model = sm.Logit(y, to_fit, maxiter = 35)

    try:
        res_model = log_model.fit(dis=0)
    except:
        continue

    form = target_name+' ~ ' + '+'.join(col_to_train)
    y, X = dmatrices(form, data =
self.t[col_to_train].join(self.t[target_name]), return_type = 'dataframe')
    vifs = [variance_inflation_factor(X.values, i) for i in
range(X.shape[1])][1:]

```

```

        if (res_model.pvalues.max() <= max_pvalue_allow) &
        (~res_model.pvalues.isna().any()) & (best_bic < res_model.bic) & (max(vifs) <=
        max_vif_allow):

```

```

            res_feature = col_to_train.copy()

```

```

            if first_best_feat == True:

```

```

                first_best_feat = col_to_train[-1]

```

```

            if len(res_feature) == max_feature_allow:

```

```

                break

```

```

        to_fit = pd.get_dummies(self.t[res_feature], columns = res_feature, prefix =
        res_feature)

```

```

        to_drop = list(map(get_to_drop_col, res_feature))

```

```

        to_fit.drop(columns = to_drop, inplace = True)

```

```

        if constant:

```

```

            to_fit = sm.add_constant(to_fit)

```

```

        log_model = sm.Logit(y, to_fit, maxiter = 35)

```

```

        log_model = log_model.fit(dis=0)

```

```

        probs = log_model.predict(to_fit)

```

```

        fpr1, tpr1, threshold = roc_curve(y, probs)

```

```

        roc_auc = auc(fpr1, tpr1)

```

```

        return {'Features':res_feature, 'roc':roc_auc, 'gini':roc_auc*2-1},
        first_best_feat

```

```

def load_to_model(best_feature):

```

Передача разных списков переменных для нахождения лучшего набора (комбинации) переменных в зависимости от критерия оценки ('iv', 'js', 'gini', 'quality_score').

```

    feature_splitter = 0
    best_feature = list(best_feature)
    for i in tqdm(range(rounds)):

        try:
            dict_to_ap, prev_best_feature =
get_params(best_feature[feature_splitter:])
            feature_splitter = best_feature.index(prev_best_feature) + 1
            dict_to_ap.update({'by_score':score_type})
            best_res.append(dict_to_ap)
        except Exception as e:
            print(e)

    if by_score == 'all':
        all_score = ['iv', 'js', 'gini', 'quality_score']
    else:
        if type(by_score) == str:
            all_score = [by_score]
        else:
            all_score = by_score

    best_res = []
    for score_type in all_score:
        load_to_model(self.summ.sort_values(score_type, ascending =
False).head(top).name.values)

```

```

self.fr = pd.DataFrame(best_res)
self.fr['n_feature'] = self.fr.Features.apply(lambda x: len(x))
self.fr = self.fr.sort_values("gini", ascending = False).reset_index(drop =
True)

def logit_model(self, col_to_train, constant = True, print_info = True):
    # Построение модели логистической регрессии с заданными
переменными, предварительно разбитых на категории.
    def get_to_drop_col(col):
        return to_fit.filter(like = col).sum().to_frame().sort_values(by = 0).index[-
1]
    self.col_to_train = col_to_train
    to_fit = pd.get_dummies(self.t[col_to_train], columns = col_to_train, prefix =
col_to_train)
    to_drop = list(map(get_to_drop_col, col_to_train))
    to_fit.drop(columns = to_drop, inplace = True)
    if constant:
        to_fit = sm.add_constant(to_fit)

    log_model = sm.Logit(self.target, to_fit, maxiter = 35)
    self.res_model = log_model.fit()
    if print_info:
        print(self.res_model.summary())

    probs = self.res_model.predict(to_fit)
    fpr1, tpr1, threshold = roc_curve(self.target, probs)
    self.roc_auc = auc(fpr1, tpr1)

    if print_info:

```



```

print('Roc: ',self.roc_auc)
self.gini = self.roc_auc * 2 - 1
print('Gini: ',self.gini)

plt.plot(fpr1, tpr1)
plt.plot([0,1],[0,1],'k--')

def save_model(self, path):
    # Сохранение параметров наилучшей построенной модели
    логистической регрессии.
    def get_bins_cat(col_to_train):
        if col_to_train != 'const':
            bins = [-np.inf] + list(self.bp.get_binned_variable(col_to_train).splits) +
            [np.inf]

            to_fit = pd.get_dummies(self.t[col_to_train], columns = col_to_train,
            prefix = col_to_train)
            counts = to_fit.sum().to_frame()
            counts = counts.reset_index()
            counts.columns = ['categories','counts']
            na_from_cat = 'pass'
            if counts.categories.values[0][-2:] == '-1':
                nan_ex = True
            else:
                nan_ex = False
            def get_new_missing_cat(col):
                subject = self.optd[col]
                subject = subject[subject.Count != 0]
                woese = subject[subject.Bin.str.contains(',')].WoE

```

```

if len(woes) == 1:
    return 'pass'
missing_woe = subject[subject.Bin == 'Missing'].WoE.values

try:
    abss = list(abs(woes - missing_woe))
    return abss.index(min(abss))
except:
    return 'pass'
na_miss = get_new_missing_cat(col_to_train)
if na_miss != 'pass':
    nan_ex = True
    na_from_cat = na_miss

params = self.res_model.params.to_frame()
params = params.reset_index()
params.columns = ['categries', 'params']
params = counts.merge(params, on = 'categries', how = 'left')
params['params'] = params['params'].fillna(0)
if nan_ex:
    if na_from_cat != 'pass':
        nan_value = params.iloc[na_from_cat].params
        return {col_to_train : {'nan':nan_value, 'bins':bins, 'coefs':
list(params.iloc[0:].params.values)}}
    else:
        nan_value = params.iloc[0].params
        return {col_to_train : {'nan':nan_value, 'bins':bins, 'coefs':
list(params.iloc[1:].params.values)}}

```

```

        return {col_to_train : {'nan':0.0, 'bins':bins, 'coefs':
list(params.params.values)}}
    else:
        to_const = pd.DataFrame( [{'categories':'const', 'counts':'1'}])
        params = self.res_model.params.to_frame()
        params = params.reset_index()
        params.columns = ['categories', 'params']
        params = to_const.merge(params, on = 'categories', how = 'left')
        params['params'] = params['params'].fillna(0)
        return {col_to_train : params.iloc[0].params}

const_exist = False
if 'const' in self.res_model.params.index:
    const_exist = True

cols = self.col_to_train
model_info = {}
for col in cols:
    model_info.update(get_bins_cat(col))
if const_exist:
    model_info.update(get_bins_cat('const'))
with open(path, 'w') as fp:
    json.dump(model_info)

```

Приложение Л

(справочное)

Свидетельство о государственной регистрации программы для ЭВМ

**«Greeder of Data – автоматический советник для построения
скоринговых моделей на базе логистической регрессии на основе
больших массивов данных»**

РОССИЙСКАЯ ФЕДЕРАЦИЯ



СВИДЕТЕЛЬСТВО
о государственной регистрации программы для ЭВМ
№ 2023612592

**Greeder of Data - автоматический советник для
построения скоринговых моделей на базе логистической
регрессии на основе больших массивов данных**

Правообладатель: *Сорокин Александр Сергеевич (RU)*

Автор(ы): *Сорокин Александр Сергеевич (RU)*

Заявка № 2022686296
Дата поступления 26 декабря 2022 г.
Дата государственной регистрации
в Реестре программ для ЭВМ 06 февраля 2023 г.

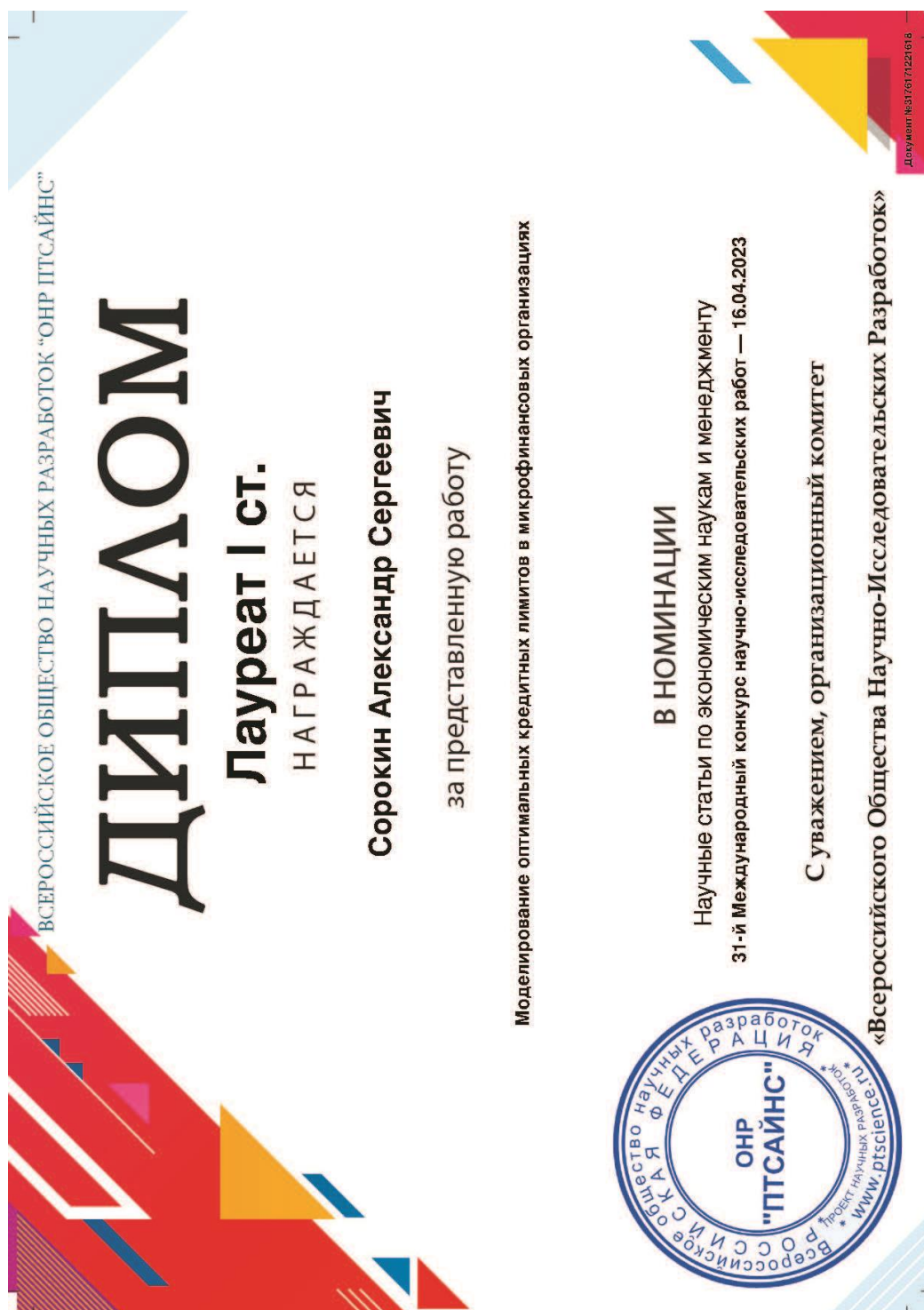
Руководитель Федеральной службы
по интеллектуальной собственности
Ю.С. Зубов



Приложение М

(справочное)

Диплом Лауреата 31-го Международного конкурса научно-исследовательских работ за работу: «Моделирование оптимальных кредитных лимитов в микрофинансовых организациях»



Приложение Н

(справочное)

Листинг кода программы для моделирования лимитов

Листинг Н.1 Код в программе Python для реализации расчетов для полиномиальной регрессии с использованием ВМНК

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from datetime import datetime

df=pd.read_excel('Путь к очищенным данным\limits.xlsx',header=None)
df

import statsmodels.api as sm
plt.title('limits and risk')
plt.scatter(df[1],df[0],c=df[2],cmap='viridis')
X=np.concatenate((np.array(df[1]).reshape(len(df),1)**2,
                    np.array(df[1]).reshape(len(df),1)**3),1)
X=sm.add_constant(X)
Y=np.array(df[0])
model=sm.WLS(Y,X,weights=np.array(df[2])**4).fit()
print(model.summary())
```

Листинг Н.2 Код в программе Python для реализации расчетов логистической регрессии

```
path = r'/Путь к файлу /limits.xlsx'
df=pd.read_excel(path)
df
df['OD_diff_Limit'] = (df['OD'] / df['Limit']).round(3)
df['Limit_extract_OD'] = (df['Limit'] - df['OD'])
```

```

df['LN_OD_div_Limit_extract_OD'] = - np.log(df['Limit_extract_OD']/df['OD'])
df
X=np.concatenate((np.array(df['NPL30']/100).reshape(len(df),1),
                    np.array(df['OD']).reshape(len(df),1)),1)
# X=add_constant(X)
Y=np.array(df['LN_OD_div_Limit_extract_OD'])
log_reg = LinearRegression()
model=log_reg.fit(X,Y)
# print(model.summary())
model.coef_
X=np.concatenate((np.array(df['NPL30']/100).reshape(len(df),1),
                    np.array(df['OD']).reshape(len(df),1)),1)
# X=add_constant(X)
Y=np.array(df['LN_OD_div_Limit_extract_OD'])
log_reg = LinearRegression()
X = sm.add_constant(X)
model = sm.OLS(Y,X)
fitted_model = model.fit()
print(fitted_model.summary())

```

```

=====
WLS Regression Results
=====
Dep. Variable:          y      R-squared:          0.728
Model:                  WLS    Adj. R-squared:       0.721
Method:                 Least Squares    F-statistic:      93.87
Date:                   Mon, 15 Nov 2021    Prob (F-statistic): 1.54e-20
Time:                   11:37:23    Log-Likelihood:   -649.99
No. Observations:      73    AIC:              1306.
Df Residuals:          70    BIC:              1313.
Df Model:               2
Covariance Type:       nonrobust
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
const          6514.0751    386.693     16.846     0.000     5742.841     7285.310
x1^2           26.3041      2.266     11.608     0.000      21.785      30.824
x2^3           -0.6333      0.062    -10.222     0.000      -0.757      -0.510
=====
Omnibus:              0.473    Durbin-Watson:      0.780
Prob(Omnibus):        0.789    Jarque-Bera (JB):    0.386
Skew:                 -0.174    Prob(JB):            0.825
Kurtosis:              2.924    Cond. No.            5.42e+04
=====

```

Рисунок Н.1 Статистическая сводка результатов расчетов модели оптимального ОД в программе Python

```

=====
                        OLS Regression Results
=====
Dep. Variable:          y      R-squared (uncentered):      0.898
Model:                  OLS    Adj. R-squared (uncentered):    0.896
Method:                  Least Squares    F-statistic:          314.2
Date:                    Wed, 13 Apr 2022    Prob (F-statistic):    5.39e-36
Time:                    15:46:01    Log-Likelihood:        -58.954
No. Observations:        73      AIC:                   121.9
Df Residuals:            71      BIC:                   126.5
Df Model:                 2
Covariance Type:         nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
x1	8.1064	0.577	14.038	0.000	6.955	9.258
x2	-6.848e-05	1.49e-05	-4.589	0.000	-9.82e-05	-3.87e-05

```

=====
Omnibus:                28.572    Durbin-Watson:          1.781
Prob(Omnibus):           0.000    Jarque-Bera (JB):       55.802
Skew:                    1.381    Prob(JB):               7.63e-13
Kurtosis:                6.274    Cond. No.:              9.70e+04
=====

```

Рисунок Н.2 – Статистическая сводка результатов расчетов модели оптимального лимита в программе Python

Приложение П

(справочное)

Листинг кода программы в R расчетов имитационного эксперимента

```

N<-100 #the number of samples
set.seed(1000)
a0<-1 #regression coefficients (as an example)
a1<-1
a2<-1
j<-1
w<-1
n<-300 #the number of observations
eb<-0
H<-0
for (i in 1:n)
{z1<-rnorm(N,1,1) #normally distributed
z2<-rnorm(N,2,2) #exogenous
u1<-rnorm(N,0,2) #variables,
u2<-rnorm(N,0,1) #parameters as an example
x1<-j*z1+z2+u1 #auxiliary regression equations
x2<-w*u1+u2 #for factors
e<-rnorm(N,0,3) #epsilon with normal distribution where 0 - the mathematical
expectation, 3 - standard deviation
y<-a0+a1*x1+a2*x2+e #regression equation
model<-lm(x1~z1+z2) #linear model x of instrumental variables z1 and z2
model1<-lm(y~predict(model))
model2<-lm(y~x1) #linear model y of x1 variable
b1<-model1$coefficients

```

```

b2<-model2$coefficients #calculating OLS coefficients
X1<-cbind(rep(1,N),x1)
Z<-cbind(rep(1,N),z1,z2) #auxiliary calculations
P<-Z%*%solve(t(Z)%*%Z)%*%t(Z)
e1<-sum((y-X1%*%b1)^2)/(N-2)
s1<-solve(t(X1)%*%P%*%X1)*e1
X1<-cbind(rep(1,N),x1)
e2<-sum((y-X1%*%b2)^2)/(N-2)
s2<-solve(t(X1)%*%X1)*e2
hausman<-c(b1-b2)%*%solve(s1-s2)%*%c(b1-b2) #calculation of Hausman
statistic
eb<-eb+(a1*j-summary(model)$coefficients[2,1])^2
if (hausman<qchisq(0.95,2))
{H<-H+1}
}
eb/n
H/n
power<- 1- H/n #calculating the power of the test
power
data<- cbind(y, x1, x2,z1,z2,u1,u2,e)
data<- as.data.frame(data,columns=c('y','x1','x2','z1','z2','u1','u2','epsilon'))
write.csv(data,'3.csv')

```