Санкт-Петербургский государственный университет
Кафедра математического моделирования энергетических систем

Прокопьева Анна Анатольевна

Магистерская диссертация

Применение информационных технологий и математического моделирования в управлении банковскими рисками

Направление 01.04.02
Прикладная математика и информатика
Магистерская программа «Математическое и информационное обеспечение экономической деятельности»

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Свиркин М.В.

Санкт-Петербург
2018

Содержание

Введение

Мировая экономика развивается стремительными темпами. В связи с этим растет количество субъектов, влияющих на экономику, которые привносят неопределенности в процессы развития. Следовательно, требуется разработка и использование эффективного средства или механизма для защиты от таких неопределенностей — механизм управления рисками.

Данный термин используется около шестидесяти лет. Основной задачей управления рисками является максимизация эффективности работы предприятия. Этапы этого процесса:

1. Определение риска
2. Измерение риска
3. Оценка различных методов обработки риска
4. Выбор метода уменьшения риска
5. Мониторинг результатов

В 1974 году был основан Базельский комитет по банковскому надзору, который призван внедрять единые стандарты в банковское регулирование. В декабре 2010 года был принят Базель III — документ, содержащий методические рекомендации, главной целью которого является повышение уровня управления рисками в банковском деле. Базель IV — документ, содержащий правки в рекомендациях, был принят в 2017 году.

В 1994 и 1997 годах банком JP Morgan были разработаны модели для управления рисками — RiskMetrics (рыночные риски) и CreditMetrics (кредитные риски).

В 2009 году Международная Организация по Стандартизации опубликовала стандарт ISO-31000, который призван установить общую терминологию и концепции. В связи с этим в РФ на государственном уровне был введен стандарт по управлению рисками (ГОСТ Р 51897–2002) [1].

Определение термина «риск», принятое в международном стандарте ISO-31000, звучит следующим образом. *Риск* — влияние неопределенности на цели (как негативно, так и позитивно). Формальное определение «риска» выглядит следующим образом:

$$R = F(u, p), \tag{1}$$

где p — вероятность возникновения неблагоприятного результата; u — количественная оценка потерь; F — функция, характеризующая риск.

В настоящее время многие российские компании развивают методы управления рисками. Можно сказать, что на российском рынке эти методы появились благодаря требованиям зарубежных компаний-партнеров. Управление рисками присутствует во многих отраслях, например, нефтегазовой, строительной, в банковском и финансовом деле. Большинство компаний используют внутренние стандарты по управлению рисками, основанные на международных документах. Такие стандарты должны часто обновляться и актуализироваться, описывать широко используемые методики и пополняться инновационными методами. Актуальность данной работы обуславливается следующими факторами:

- наличие большого количества современной тематической литературы и научных исследований
- слабо развитая система оценки и анализа рисков в российских банках (и компаниях в целом) по сравнению с зарубежными
- использование в практике субъективных методов оценки рисков
- необходимость внедрения современных методик, информационных технологий и способов обработки большого количества данных для принятия объективных решений

Постановка задачи

Коммерческие банки и банковское дело в целом являются сложноорганизованной системой с огромным количеством факторов, сущностей, параметров и связей. Целью данной работы является оценка кредитоспособности заемщика и кредитного риска банка (и возможная адаптация под другие предприятия) при принятии решения о выдаче кредита. В свою очередь для достижения цели требуется решить ряд задач:

1. Провести сравнительный анализ существующих количественных оценок банковских рисков для выбора подходящего способа;
2. Подготовить и проанализировать данные кредитных заявок;
3. Провести исследование выделенных переменных, исключить коррелирующие переменные, провести категоризацию количественных переменных;
4. Построить логическую схему данных для создаваемого программного комплекса;
5. Определить необходимый объем выборки;
6. На основе построенной логической схемы создать математическую модель и физическую модель БД программного комплекса;
7. С помощью выделенных выборок обучить модель и получить набор коэффициентов. Оценить качество модели по тестовому набору данных;
8. Масштабировать полученные коэффициенты модели в скоринговые баллы;
9. Зафиксировать полученную рейтинговую систему с определением числа заемщиков и количества дефолтов в каждой группе;
10. Определить ожидаемые потери кредитного портфеля;
11. С помощью имитационного моделирования вычислить VaR и оценить неожиданные потери;
12. Привести интерпретацию результатов, проанализировать их и принять решение по управлению оцененным риском.

Обзор литературы

Задача оценки кредитоспособности и кредитных рисков предприятия актуальна в виду ее большой распространенности. Работы, посвященные этой теме, активно публикуются по сей день. Ниже отмечены некоторые из них, опубликованные за последние несколько лет.

Основной работой, на которую опирался автор данного исследования, является статья, описывающая методологию построения скоринговых карт с использованием модели логистической регрессии [2]. Автор статьи приводит описание различных методик и подходов к построению скоринговых карт, и акцентирует внимание на основных проблемах построения модели. Статья [3] посвящена задаче оценки качества модели логистической регрессии. Также в исследовании применялся метод кросс-валидации, описанный в работе [4], для проверки модели на независимых наборах данных. В книге [5] представлены некоторые способы масштабирования коэффициентов логистической регрессии в скоринговые баллы.

Источником для изучения видов риска и их классификации послужила книга [6]. Авторы приводят основные определения, методы оценки и управления различными видами риска. Также подробно освещен подход в классификации финансовых рисков в труде [7]. Разбор методов статистического анализа приводится в работе [8]. Рассмотрение подходов к управлению рисками и практические рекомендации представлены в работе [9].

По теме машинного обучения в кредитном скоринге стоит отметить статью [10], в которой приводится описание инструмента оценки кредитоспособности контрагента. В работе используются модели логистической регрессии и случайного леса, а также показана аналитика результатов предсказательной способности моделей. Кроме того, в статье [11] представлено применение методов машинного обучения в исследованиях кредитного риска.

Работы [12] и [13] описывают подход к исследованию кредитного риска с помощью статистических методов оценки максимальных убытков на основании рейтинговых систем предприятия.

Глава 1. Исследование предметной области

В текущей главе автором проводится знакомство с предметной областью риск-менеджмента в банковской сфере. В дальнейших разделах последовательно вводится определение риска, классификация различных видов рисков, методы их оценки и управления. Результатом знакомства с предметной областью является выбор вида риска для дальнейшего исследования и метода его оценки. Кроме того, в конце главы приводится сравнительный анализ основных количественных методов оценки риска.

1.1. Определение понятия банковского риска

Банковское дело является по существу предпринимательством и как следствие подвержено рискам. Необходимо ввести понятие «банковский риск». В литературе можно найти множество разных определений, например, в работе [14] оно введено следующим образом: «*банковский риск* — вероятность того, что произойдет событие, которое неблагоприятно скажется на прибыли или капитале банка»

В данной работе принимается следующее более общее определение [15]: *Банковский риск — вероятность понесения кредитной организацией потерь вследствие неблагоприятного исхода операций, проводимых организацией, или наступления непредвиденных ситуаций.*

Все банки подвержены рискам, как самым распространенным в виде невозврата заемщиком средств, так и редким, таким как стихийные бедствия и катастрофические события. Управление рисками считается одним из самых важных аспектов банковской деятельности.

Банковские риски можно разделить по разным критериям. Далее рассматривается подход к их классификации.

1.2. Классификация банковских рисков

В качестве признаков, по которым можно выделить определенные группы банковским рисков, используются [7]:

- По сфере влияния (возникновения);
- По времени;
- По степени (уровню);

- По принадлежности к группе системы отношений;
- По степени постоянства;
- В зависимости от возможного результата;
- По возможности страхования;
- По степени охвата;
- По характеру банковских операций;
- По виду операций;
- По виду клиентов банка.

Далее рассматривается иерархическая (многоуровневая) классификация (Рис. 1) и даются определения некоторым рискам [6].

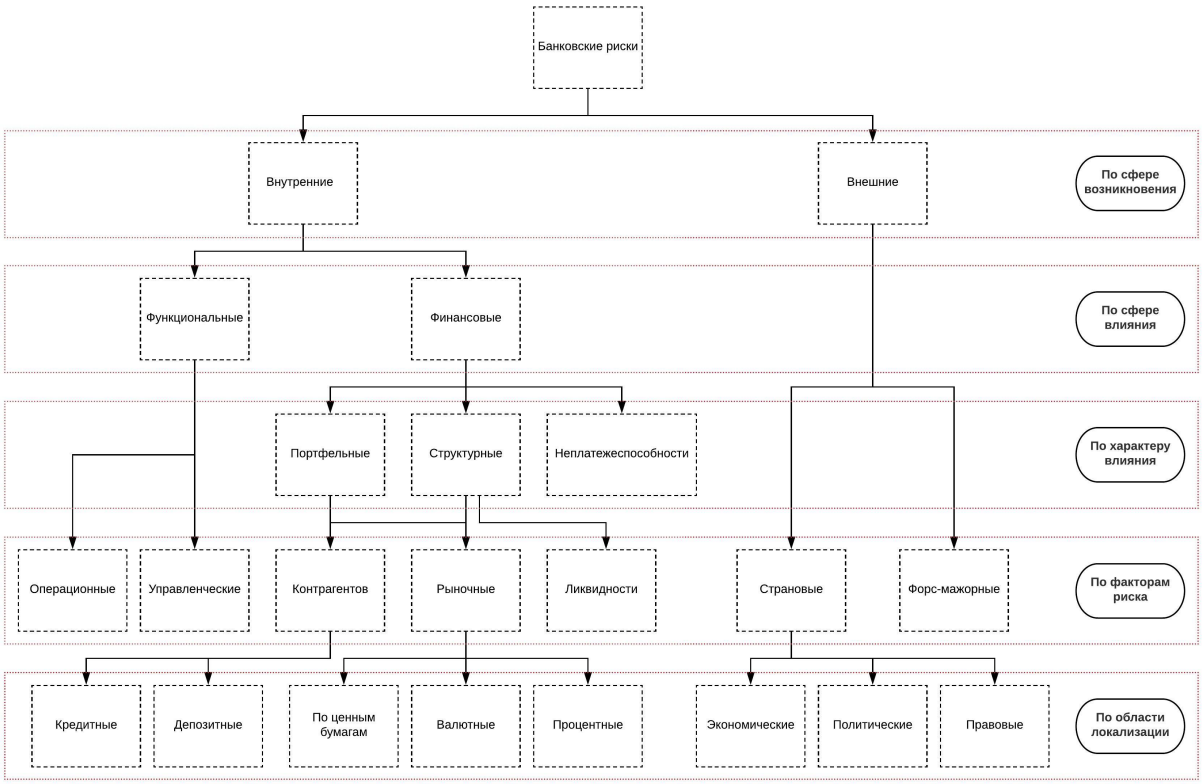

Рис. 1: Иерархическая классификация банковских рисков

Самым широким критерием является «сфера влияния (возникновения)», по которой риски делятся на внешние и внутренние. Первые относятся к воздействию политических, социальных и прочих изменении окружающей среды. Внутренние риски непосредственно связаны с деятельностью банка.

Далее внутренние риски можно разделить по «сфере влияния» на функциональные и финансовые. Функциональные — это операционные и управленческие риски, а финансовые — портфельные и структурные. Внешние риски делятся по «факторам риска» на страновые (сюда относятся экономические, политические и правовые) и форс-мажорные.

Портфельные риски — это риски, которые влияют на объем, стоимость и доходность обязательств банка (отражаются в активах и пассивах). Они подразделяются на риски контрагентов и рыночные.

В свою очередь структурные риски — риски, влияющие на структуру, стоимость и доходность обязательств, делятся на рыночные и риски ликвидности (вероятность потерь в связи с трудностями в реализации активов).

Необходимо различать структурные рыночные риски и портфельные. В первом случае они относятся к изменению рынка (сюда относятся валютные — вероятность потерь в связи с изменением курсов валют; и процентные риски — вероятность потерь в связи с изменением процентных ставок), а во втором к рискам по ценным бумагам (вероятность потерь в связи с изменением стоимости ценных бумаг в портфеле банка).

К рискам контрагентов относятся кредитные и депозитные риски. Кредитный риск — вероятность потерь банка при невозврате заемщиком основной суммы долга и процентов. Депозитный риск — вероятность потерь в связи с досрочным изъятием депозитных средств.

Операционный риск — вероятность потерь в связи с ошибками и мошенничеством банковского персонала, а также с техническими сбоями.

Стоит упомянуть, что чем больше степень риска, тем выше должна быть прибыль. Таким образом, главная цель состоит в достижении баланса прибыли и рисков. Среди способов достижения такой цели можно выделить формирование резервов в ЦБ (центральном банке), хеджирование и диверсификация, формирование «подушки безопасности» путем перечисления процента с каждой сделки на счет в национальном банке. Эти подходы призваны к снижению уровня риска каждого банка, так как банкротство одного может стать толчком к падению всей банковской системы и кризисной ситуации.

В итоге, стандартного подхода к классификации банковских рисков не существует. Приведенная выше схема имеет наибольшее распростране-

ние в литературе. Важно отметить, что 80% всех операций коммерческих банков — это кредитные операции, соответственно снижение кредитного риска является крайне важной и актуальной задачей любого банка. Исследование именно этого риска будет проведено в данной работе.

1.3. Количественная оценка риска

Перед тем, как перейти к количественной оценке, необходимо рассмотреть задачу качественного анализа риска. Целью качественного анализа является выделение причин риска, другими словами в данное понятие входит определение потенциальных зон появления риска, формирование перечня рисков, присущих деятельности предприятия (например, банк должен рассматривать риски, которые несут их клиенты, т. к. это автоматически является и риском банка), предсказание негативных последствий (или позитивных) вследствие риска. Такой подход позволяет руководителю на поверхностном уровне увидеть бесперспективные или имеющие серьезные негативные последствия решения и отказаться от них. В свою очередь, количественный анализ применяется для оценки тех рисков, которые могут возникнуть, если решение все-таки было принято.

С помощью количественного анализа [16] вычисляются значения мер рисков и возможная величина ущерба. Данная оценка использует аппарат таких дисциплин, как теории вероятностей, математической статистики и исследования операций. В литературе встречаются следующие методы количественного анализа:

1. Статистические методы;
2. Аналитические методы;
3. Метод экспертных оценок;
4. Метод аналогий;
5. Машинное обучение.

В подразделах ниже изучаются каждый из приведенных методов.

Статистические методы

Основа статистических методов состоит в определении вероятности реализации риска на основании статистических данных прошлых периодов

и фиксировании области риска [17]. Преимуществом таких методов является их способность анализировать различные события (сценарии) и учитывать разные факторы рисков. Недостаток метода состоит в использовании вероятностных характеристик.

Требуется ввести дополнительные определения [8]. *Среднее значение* — обобщенная количественная характеристика, является средневзвешенным для всех возможных результатов. *Вариативность* — это степень отклонения ожидаемого значения результата от его средней величины. Основные расчетные показатели [6], [8]:

1. Уровень финансового риска;
2. Дисперсия;
3. Среднеквадратическое (стандартное) отклонение;
4. Коэффициент вариации;
5. Бета-коэффициент.

Далее рассматриваются перечисленные показатели.

Уровень финансового риска представляет собой общий алгоритм оценки:

$$LR = PR \times PD, \tag{2}$$

где LR — уровень риска; PR — вероятность риска; PD — размер возможных потерь.

Обычно в качестве вероятности риска используется коэффициент вариации или частотная вероятность, а размер потерь — абсолютное значение. Таким образом, уровень риска будет также абсолютным значением, что уменьшает способы его сравнения с другими вариантами.

Дисперсия характеризует степень вариативности ожидаемого дохода от среднего значения. Чем выше колебания, тем выше степень риска. Показатель рассчитывается следующим образом:

$$\sigma^2 = \sum_{i=1}^{n} (R_i - \overline{R})^2 \times P_i \tag{3}$$

где σ^2 — дисперсия; R_i — значение возможных вариантов ожидаемого дохода; $\overline{R}$ — среднее ожидаемое значение дохода; P_i — вероятность получения отдельных вариантов ожидаемого дохода; n — количество наблюдений.

Из недостатков следует отметить, что дисперсия не дает представление о линейных отклонениях $\Delta X = X - \overline{R}$, которые более наглядны для анализа рисков.

Среднеквадратическое (стандартное) отклонение является одним из самых популярных показателей при оценке финансового риска:

$$\sigma = \sqrt{\sum_{i=1}^{n} (R_i - \overline{R})^2 \times P_i}, \tag{4}$$

где σ — среднеквадратическое отклонение.

Особенностью показателя является следующий факт: при близости наблюдаемого распределения к нормальному, данный показатель можно использовать для определения границ, в которые с заданной вероятностью попадет значение случайной величины. Однако, особенность выступает одновременно и недостатком, т.к. показатель не даст корректной оценки риска, если распределение перестанет быть нормальным. К тому же, риск-менеджерам предпочтительнее получать результаты оценки в виде реальных финансовых потерь.

Коэффициент вариации используется в случае различия средних ожидаемых доходов и рассчитывается по формуле:

$$CV = \pm \frac{\sigma}{\overline{R}} \times 100\%, \tag{5}$$

где CV — коэффициент вариации.

Коэффициент измеряется от 0% до 100% и показывает степень вариативности (до 10% — слабая; 10-25% — умеренная; свыше 25% — сильная). Преимущество метода состоит в том, что с его помощью можно сравнивать вариативность признаков с разными единицами измерения.

Бета-коэффициент применяется для оценки рисков инвестирования в ценные бумаги и является мерой рыночного риска. Для дальнейшей работы показатель не представляет интереса и далее рассматриваться не будет.

Другими статистическим методами являются:

1. Метод «дерево решений»
2. Метод имитационного моделирования (Монте-Карло)
3. Методология RiskMetrics

4. VaR-метод
5. Модель Чессера
6. Z-модель Альтмана

Далее рассматриваются некоторые из методов.

Метод «дерево решений» используется [8] при анализе рисковых событий, имеющих обозримое число ветвей развития. Достоинством метода является его ценность при рассмотрении события, которое зависело от предыдущих.

Суть метода заключается в построении картины возможных вариантов развития событий после принятия решений и вычислении ожидаемой стоимостной оценки (EMV) как максимальной из сумм оценок выигрышей, умноженных на вероятность реализации этих выигрышей для всех возможных альтернатив.

Метод имитационного моделирования (Монте-Карло) представляет собой мощный инструмент для анализа [18]. Суть метода состоит в проведении экспериментов (на вычислительной машине) с математическими моделями систем, когда реальные испытания являются неосуществимыми или очень затратными. Кроме того, при нехватке информации метод способен генерировать необходимые величины. Рассмотрим метод поэтапно:

1. На первом этапе создается математическая модель. Основные принципы заключаются в последовательном определении переменных (факторов), входящих в модель; задании каждой выделенной переменной функции распределения; определении связей (зависимости) между переменными. Общий вид модели:

$$F = f(x_1, \ldots, x_i, \ldots, x_n; a_1, \ldots, a_j, \ldots, a_m), \tag{6}$$

где x_i — случайные величины (риск-переменные); n — количество случайных величин; a_j — фиксированные параметры модели; m — количество фиксированных параметров. При выделении переменных можно проводить анализ чувствительности с помощью расчета рейтинга эластичности, чтобы выбрать наиболее рисковые (важные) факторы.

Далее для каждой случайной величины требуется выбрать закон распределения. Наиболее часто используемые законы: нормальный, тре-

угольный, равномерный, дискретный. Основные рекомендации для подбора закона распределения: определить границы диапазона изменения случайной величины; взять общий закон распределения; оценить характеристики закона на основании предыдущих шагов (для непрерывных величин) или составить таблицу вероятности для каждого значения дискретной случайной величины. Также рекомендуется провести проверку на наличие корреляции между переменными и отразить это в математической модели, в противном случае, результаты моделирования могут быть серьезно искажены.

2. На втором этапе реализуется алгоритм проведения имитации (метод Монте-Карло). Шаги алгоритма:

   (a) Генерирование независимых и равномерно распределенных на $[0; 1]$ величин. Такие величины интерпретируются как значение функции распределения соответствующей случайной величины;

   (b) С помощью полученных вероятностей вычисляется соответствующее значение случайной величины;

   (c) Вычисленное значение подставляется в математическую модель, рассчитываются параметры имитационной модели;

   (d) Алгоритм повторяется N раз (количество имитаций). При этом после каждой имитации вычисленное на Шаге (c) значение сохраняется.

   Значение N влияет на репрезентативность выборки и должно быть достаточно большим.

3. Заключительным этапом выступают анализ и интерпретация результатов моделирования. Анализ может быть основан на графической информации и на количественных показателях.

   Количественно: вероятность каждого из N случайных сценариев равна $P(i) = \frac{1}{N}$; таким образом, вероятность того, что оценка риска будет в пределах определенного уровня, равна количеству экспериментов, в которых значение модели было в тех же пределах, умноженному на вероятность одного сценария.

   Графически: построение гистограммы результирующих значений модели позволяет определить закон распределения итогового показателя.

Такое построение осуществляется обычным способом через построение вариационного ряда и его разбиение на интервалы. Для того, чтобы убедиться, что подобранный закон согласуется с наблюдаемым, используется критерий согласия χ^2.

Методология RiskMetrics разработана компанией J. P. Morgan для оценки риска рынка ценных бумаг. Представляет собой набор средств, позволяющих оценить уровень влияния рыночного риска с помощью вычисления VaR (Value-at-Risk). Останавливаться на подробном рассмотрении данной методики не будем.

VaR-метод. VaR (Value-at-Risk) — это оценка величины (в денежном эквиваленте), которую не превысят максимальные потери в течение заданного периода времени с определенной вероятностью [8]. Параметры, характеризующие данную величину:

- Временной горизонт (в Базельском соглашении принято 10 дней);
- Доверительный интервал — уровень допустимого риска (в Базельском соглашении принято 99%);
- Валюта, в которой измеряется величина.

Методы вычисления данной оценки:

1. Аналитический (или параметрический). Расчеты проводятся в предположении об известности функции распределения доходов (часто явно задается нормальный закон). Однако предположение о нормальности является достаточно сомнительным, что не позволяет полностью доверять такому методу.
2. Исторический. Вид функции распределения находится из наблюдаемого ранее временного ряда. Такой метод не требует слишком упрощающих предположений и достаточно эффективен, однако всецело зависит от выбора истории.
3. Статистический. Метод основан на моделировании значений случайных величин. Основное отличие от исторического метода в том, что данный метод сам генерирует требуемые значения и не зависит от предыстории. Данный метод достаточно точен, но не всегда прост в применении. В дальнейшей работе (Глава 3) рассматривается именно статистический метод вычисления VaR.

Другие статистические методы. Модель Чессера рассматривается в работе [19], Z-модель Альтмана подробно описана в [20]. Данные модели не будут рассматриваться далее.

Аналитические методы

С помощью аналитических методов можно определить вероятность реализации риска на основании математических моделей. Они применяются, когда организация располагает ограниченной информацией, и необходимо провести количественную оценку. Эти методы в основном применяются для анализа риска инвестиционных проектов. Используются такие аналитические методы, как:

1. Анализ чувствительности;
2. Метод корректировки нормы дисконта с учетом риска;
3. Метод эквивалентов;
4. Метод сценариев;
5. Дюрация;
6. Стресс-тестирование;
7. GAP-анализ.

Подробнее рассматриваются некоторые из методов.

Анализ чувствительности заключается в исследовании зависимости итогового показателя от изменения значений переменных модели [8]. Основные этапы метода:

1. Выбор показателя, чувствительность которого исследуется;
2. Выбор факторов и задание их взаимосвязи с итоговым показателем;
3. Расчет значения показателя с помощью изменения значений факторов;
4. Построение диаграмм зависимости выбранных показателей (в динамике) и интерпретация результатов.

Основные преимущества метода заключаются в его простоте и в возможности определить какие именно факторы наиболее сильно влияют на показатель и относятся к рисковым. Таким образом, проект с меньшей чувствительностью считается менее рискованным и наоборот.

Коэффициент эластичности — это число, показывающее изменение функции относительно изменения коэффициента, вычисляется по формуле:

$$\epsilon_i = \frac{\Delta y \times x_i}{\Delta x_i \times y}, i = 1, \ldots, n, \quad (7)$$

где x_i — переменная, относительно которой исследуется эластичность.

Чем больше абсолютное значение коэффициента эластичности, тем выше степень чувствительности.

Однако у метода есть и недостатки. При изменении одной переменной, остальные фиксируются, но в реальности между переменными существует связь, и изменение одной влечет изменение другой.

Метод сценариев продолжает собой метод анализа чувствительности, изменяя одновременно несколько факторов.

Стресс-тестирование. Главной задачей метода является определение устойчивости системы при превышении установленных предельных показателей. Позволяет оценить способность капитала организации покрыть возможные убытки, появившиеся в следствие установленного сценария.

GAP-анализ позволяет увидеть, существуют ли разрывы между целями организации и ее реальными возможностями. Метод в основном используется аудиторами.

Другие аналитические методы. Метод корректировки нормы дисконта с учетом риска и метод эквивалентов представлены в работе [21], а метод дюрации рассматривается в [22].

Метод экспертных оценок

Метод включает в себя логические и статистические методы, а также способы обработки результатов опросов группы экспертов. Методика позволяет использовать интуицию и опыт экспертов. Достоинством метода является его применимость в случае недостатка или полного отсутствия информации. Недостатком — сложность в формировании компетентной и профессиональной группе экспертов.

Суть метода состоит в проведении опросов, например, для выявления степени влияния факторов на уровень риска. После опроса собранная информация обрабатывается, анализируется и используется в дальнейшей оценке или в принятии решений.

К экспертным оценкам относят такие методы, как: метод Дельфи, деревья решений, рейтинговые методы (scoring-модели).

Метод Дельфи обладает рядом преимуществ: эксперты независимы, присутствует анонимность (нет влияния авторитета), заочность (опросы можно проходить удаленно), многоуровневость (знакомство с оценками других экспертов, обоснование своих оценок, пересмотр своего мнения). Цель метода — с помощью последовательных опросов и интервью достигнуть согласованности в оценках экспертов. В качестве меры согласованности используется коэффициент конкордации Кендалла W, который вычисляется по формуле

$$W = \frac{12S}{n^2(m^3 - m)} \tag{8}$$

где S — сумма квадратов отклонений всех оценок рангов каждого объекта экспертизы от среднего значения; n — число экспертов; m — число объектов экспертизы (факторов).

Если коэффициент равен 1, то это означает полную согласованность мнений экспертов. Если равен 0 — имеет место полная несогласованность. Соответственно сами оценки могут проходить несколькими способами: попарные сравнения, ранжирование и др.

Рейтинговые методы работают аналогично дискриминантному анализу в scoring-системах и заключаются в экспертной оценке индивидуальных характеристик заемщика. Параметрам присваиваются весовые коэффициенты, которые подставляются в scoring-карту, с помощью которой вычисляется рейтинг заемщика. По значению рейтинга принимается решение о предоставлении кредита.

Очевидным недостатком всех экспертных оценок является их субъективность.

Метод аналогий

Метод используется при анализе новых ситуаций или продуктов. Суть метода состоит в переносе аналогичного случая на исследуемый объект. Главным недостатком метода является крайняя сложность в воссоздании условий, при которых прошлый опыт повторится. В виду этого недостатка далее данный метод не рассматривается.

Машинное обучение

Машинное обучение широко применяется в кредитном скоринге. Строго говоря, машинное обучение можно отнести к статистическим методам [11], [10], [23], [24], однако следует рассмотреть данную группу методов отдельно, в виду активного развития и широкого применения машинного обучения в настоящее время. К методам машинного обучения по оценке рисков относятся:

1. Логистическая регрессия;
2. Случайный лес (деревья решений);
3. Нейронные сети;
4. Генетический алгоритм.

Требуется рассмотреть подробнее каждый метод.

Логистическая регрессия строит линейный классификатор для оценки апостериорных вероятностей принадлежности объектов классам [25], [26]. Касательно процесса кредитования, задачу можно сформулировать следующим образом. Пусть заемщики описываются n признаками $f_j : X \to \mathcal{R}, j = 1, \ldots, n$. Тогда пространство признаков $X = \mathcal{R}^n$. Пусть Y — конечное множество номеров классов ($Y = \{-1; +1\}$). Пусть также задана обучающая выборка пар (заемщик, исполнение обязательств по кредиту): $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Тогда строим линейный классификатор вида

$$a(x, w) = sign\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = sign\left\langle x, w\right\rangle, \tag{9}$$

где w_j — вес j-го признака; w_0 — порог принятия решения; $w = (w_0, \ldots, w_n)$ — вектор весов; $\langle x, w\rangle$ — скалярное произведение признакового описания объекта на вектор весов.

Задача обучения линейного классификатора состоит в том, чтобы по обучающей выборке X^m настроить вектор весов w характеристик заемщиков. Для этого решается следующая задача минимизации

$$Q(w) = \sum_{i=1}^{m} \ln\left(1 + exp(-y_i\left\langle x_i, w\right\rangle)\right) \to \min_w, \tag{10}$$

Когда решение задачи найдено, можно не только проводить классификацию заемщиков $a(x) = sign\langle x, w\rangle$, но и определять их апостериорные вероятности принадлежности тому или иному классу по формуле

$$\mathcal{P}\{y|x\} = \sigma(y\langle x, w\rangle), y \in Y, \tag{11}$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция.

Таким образом можно классифицировать заемщиков на основе обученной модели, как «хороших» (погашающих долг вовремя), так и «плохих» (просрочивших или совсем не заплативших). В дальнейшей работе данный метод применяется для классификации заемщиков и построении скоринговой карты. Его основным преимуществом является возможность интерпретации коэффициентов для каждого признака в отдельности.

Случайный лес. Суть метода заключается в использовании множества решающих деревьев [27]. В работе стоит задача классификации заемщиков, для решения такой задачи используется метод голосования в множестве деревьев. Схема построения решающего дерева:

1. Берется подвыборка из обучающей выборки с повторениями, именно по ней строится дерево;
2. Для построения каждого ветвления берем случайных признаков (по умолчанию, но его следует настраивать самостоятельно подбором);
3. Выбираем наилучший признак и ветвление по нему по заданному критерию (например, критерий Джини [27]). Дерево строится до исчерпания выборки или до предела по высоте.

Чем больше деревьев построено, тем качественнее модель, однако время решения задачи существенно увеличивается с ростом количества деревьев.

Нейронные сети. Достаточно широко используемым методом машинного обучения считаются нейронные сети. Они представляют собой мощный инструмент моделирования, с помощью которого можно рассматривать сложные (в том числе нелинейные) зависимости.

Искусственная нейронная сеть — это модель, которая состоит из множества сгруппированных в слои нейронов. Каждый нейрон представляет собой некоторый элемент с заданной функцией, который обрабатывает входные данные. Нейроны и связи между ними в совокупности представляют нейронную сеть, которая, обучаясь, может решать задачи дискри-

минантного анализа и кластеризации. При построении метод оценивает каждый признак, его вес, взаимосвязи между признаками, что позволяет ему строить нетривиальные модели. Метод нейронных сетей в кредитном скоринге используется, однако построить высокоточную сеть достаточно сложно. Существенный недостаток использования нейронных сетей именно в кредитовании — невозможность интерпретирования полученных весов связей. Также для хорошего обучения нейронной сети требуется большой объем размеченных данных, которые обычно находятся в закрытом доступе. Все обезличенные выборки в открытом доступе имеют малый объем, поэтому на этом этапе работы данный метод не рассматривался.

Генетический алгоритм представляет собой метод многомерной оптимизации, который базируется на моделировании эволюционного процесса [28]. Особенность алгоритма заключается в «скрещивании» (комбинировании параметров). Алгоритм делится на несколько этапов:

- Скрещивание;
- Селекция;
- Формирование нового поколения.

Шаги повторяются до тех пор, пока результат не достигнет нужного уровня или количество поколение не достигнет установленного максимума.

Создание новой популяции. На этом шаге формируется стартовая (начальная) популяция.

Размножение. Для получения потомка нужны два родителя. Главное, чтобы потомок (ребенок) мог унаследовать у родителей их черты. При этом накладывается условие, что размножаются все, а не только выжившие, иначе выделится одна модель, «гены» которой перекроют всех остальных.

Мутации. Из потомков выбирают некоторое количество особей и изменяют их в соответствии с заранее определенными операциями.

Отбор. Выбор «выживших», их количество (доля) задается в самом начале алгоритма. После мутации начинаем выбирать из популяции долю тех, кто «пойдет дальше». Остальные особи должны «погибнуть».

В кредитовании это представляется как классификация моделей, подвергающихся «мутациям», которые еще могут скрещиваться, в итоге выбирается наилучшая модель, дающая наиболее точный ответ. У генетиче-

ских алгоритмов достаточно высокая вычислительная сложность и, как следствие, долгое время обучения. В связи с этим далее данный метод не рассматривается.

1.4. Сравнительный анализ основных количественных методов

Сравнительная характеристика основных количественных методов оценки рисков сведена в Табл.1.

По результатам сравнения в данной работе будут использоваться методы машинного обучения для скоринга кредитных заявок, как наиболее востребованный на рынке инструмент, а также имитационное моделирование для оценки кредитного риска в виду его широко признанной эффективности.

В качестве инструмента реализации программного комплекса был выбран язык Python, т. к. данный язык программирования содержит большое количество свободно распространяемых библиотек и является главным инструментом аналитика. Для реализации методов машинного обучения используется библиотека scikit-learn [29], для построения интерфейса — HTML, CSS и фреймворк Flask. Инструментами для имитационного моделирования использовались распространенные математические пакеты — NumPy [30], SciPy [31] и Pandas [32], предлагающие удобный формат для манипуляции с данными.

Таблица 1: Сравнение основных количественных методов

№	Методы	Вид риска	Преимущества метода	Недостатки метода
1.	**Статистические**			
1.1.	Расчетные показатели	Финансовый	Простота вычисления, широкое распространение	Ограниченное применение в сравнительном анализе
1.2.	Метод «дерево решений»	Кредитный, инвестиционный	Позволяет увидеть картину в целом	Трудоемкость
1.3.	Метод имитационного моделирования (Монте-Карло)	Кредитный	Высокое определение размера потерь и вероятность реализации риска в обычных условиях	Необходимость обработки большого количества статистической информации. Малая эффективность оценки в условиях кризиса
1.4.	Методология RiskMetrics	Рыночный	Возможность сравнительного анализа потерь и соответствующих им рисков	Использование исторических данных, трудность адаптации к изменениям рынка
1.5.	VaR-метод	Кредитный, рыночный	Высокое определение размера потерь и вероятность реализации риска в обычных условиях	Необходимость обработки большого количества статистической информации. Малая эффективность оценки в условиях кризиса
1.6.	Модель Чессера	Кредитный		
1.7.	Z-модель Альтмана	Кредитный		

Таблица 1: Сравнение основных количественных методов

№	Методы	Вид риска	Преимущества метода	Недостатки метода
2.	**Аналитические**			
2.1.	Анализ чувствительности	Кредитный	Возможность факторного анализа	Переменные фиксируются, нельзя выделить взаимосвязи
2.2.	Метод корректировки нормы дисконта с учетом риска	Инвестиционный	Простота расчетов	Производит увеличение риска во времени с постоянным коэффициентом
2.3.	Метод эквивалентов	Инвестиционный	Дает комплексную рыночную оценку	Сложность определения достоверных эквивалентов и показателя дисконтирования с поправкой на риск
2.4.	Метод сценариев	Инвестиционный	Возможность факторного анализа	Не всегда точен, т.к. не учитывает сложные взаимосвязи
2.5.	Дюрация	Процентный, инвестиционный	Включает в себя возможности факторного анализа параметров. Высокая эффективность оценки в условиях кризиса	Трудоемкость
2.6.	Стресс-тестирование	Валютный, инвестиционный		
2.7.	GAP-анализ	Процентный		

Таблица 1: Сравнение основных количественных методов

№	Методы	Вид риска	Преимущества метода	Недостатки метода
3.	**Метод экспертных оценок**			
3.1.	Метод Дельфи	Кредитный, валютный, процентный	Результативен в условиях недостатка или отсутствия достоверной информации	Субъективный характер
3.2.	Деревья решений	Инвестиционный		
3.3.	Рейтинговый метод	Кредитный	Эффективен в условии кризиса	
4.	**Метод аналогий**	Кредитный, ликвидности, процентный, валютный, инвестиционный	Высокая эффективность оценки в обычных условиях	Трудно создать аналогичные условия
5.	**Машинное обучение**			
5.1.	Логистическая регрессия	Кредитный	Высокая эффективность, удобство включения новых параметров	Требуются размеченные данные для обучения
5.2.	Случайный лес	Кредитный		
5.3.	Нейронные сети	Кредитный		
5.4.	Генетический алгоритм	Кредитный		

Глава 2. Построение скоринговой карты

2.1. Работа с исходными данными

Для формирования скоринг-карты необходимо обладать качественными данными о клиентах-заемщиках банка. Данные определяют, насколько точно будет оценивать потенциальных заемщиков будущая модель.

Как правило скоринг-карты строятся на основании исторических данных, важно, чтобы их объем был существенен — этому посвящен отдельный раздел в данной работе. Среди исходных данных могут быть как характеристики заемщика, так и его внутренние (внешние) кредитные истории. На рынке существует услуга по предоставлению данных о кредитной истории заемщика.

Также достаточно важным является использование исторических данных из той же области, для которой строится скоринговая карта. Например, следует использовать данные о потребительских кредитах для построения карты по тому же виду кредитования.

Дифференцирование карт считается хорошей практикой для более адекватной оценки риска. Например, можно построить несколько карт в зависимости от региона или вида кредитования.

В свободном доступе находится мало наборов размеченных кредитных заявок. Большинство из них не подходят по причине небольшого объема или малого количества дефолтных заявок. Для корректной работы модели в тренировочной выборке должны присутствовать размеченные данные обоих классов. В целях исследования важно как можно точнее выделять «плохих» заемщиков, соответственно, выборка должна содержать большое количество «плохих» заявок. В результате анализа доступных датасетов в качестве исходной выборки в работе используется набор обезличенных заявок, загруженный с ресурса [33]. Известно, что это выборка по заемщикам (физическим лицам) из Испании в период 2011-2012 гг., однако более подробной информации о происхождении набора данных нет. Выборка содержит 4454 записи о заемщиках, из которых 3200 «хороших» заявок и 1254 «плохих». Объем дефолтных заявок составляет около 30% от общего количества заявок и является приемлемым для целей работы. Каждого заемщика характеризует 14 переменных. Все переменные пере-

числены в Приложении А, в Табл. 10. В работу взяты все характеристики заемщиков, предоставленные выборкой, т.к. в дальнейших разделах проводится подробный анализ набора данных, в котором все переменные представляют практический интерес. Также далее продемонстрировано, что все характеристики (за исключением одной) являются значимыми.

Кроме того, для развития Системы предусмотрено сохранение данных о заемщике и принятое по нему решение для включения этих результатов в модель при повторном обращении заемщика за услугой кредитования. Таким образом в Системе можно реализовать дифференцирование карт для оценки «новых» и «старых» клиентов. Далее в работе все клиенты будут считаться «новыми».

Выбор зависимой и независимых переменных модели

Целью построения скоринг-карты в данной работе является уменьшение числа просроченных платежей или невыплат в будущих кредитах. В соответствии с этим в качестве зависимой переменной принимается категориальная величина с двумя категориями «хороший» и «плохой». Чаще всего к «плохому» относят заемщика, который имеет «просрочку» по платежу свыше 90 дней. Определение зависимой переменной позволяет оценивать требуемую банком характеристику.

Далее требуется определить набор независимых (или скоринговых) переменных. Скоринговыми переменными для физического лица могут выступать личная информация, финансовые показатели, кредитная история и т.д., получаемые из анкет или ранее имеющихся данных. Внешняя кредитная история Заемщика доступна по запросу в Бюро кредитных историй [34]. Как упоминалось в предыдущем разделе, для работы используется предоставленная выборкой информация по 14 переменным, полный перечень которых приведен в Приложении А, в Табл. 10.

Для юридического лица за основу берутся финансовые показатели деятельности, получаемые из бухгалтерского баланса, отчета о прибылях и убытках, а также из отчета об изменении собственного капитала, заверенных независимым аудитором.

В следующем разделе проводится анализ независимых переменных и формирование окончательного набора — исключение переменных, улучшение качества данных, категоризация.

Общие принципы подготовки данных

При работе с реальными данными обычно возникает проблема пропусков или некорректных значений. Чаще всего это происходит по причине человеческого фактора: ошибка оператора, умышленный отказ от предоставления информации заемщиком и т.д.

На практике рекомендуют исключать из выборки строки с пропусками по некоторым параметрам, если их не более 5%. В противном случае пропущенные параметры нужно изучить, т.к. например, умышленный отказ от заполнения поля может серьезно повлиять на рейтинг заемщика. Чаще всего в таких случаях пропуск заменяют заведомо уникальным значением и оставляют в выборке. Однако, большое количество пропусков может также говорить о том, что наблюдаемая переменная не внесет большого вклада в результирующую модель, и ее можно исключить.

В рассматриваемой выборке пропущено следующее число параметров: в 34 записях не указаны доходы, в 47 — активы, еще в 18 — обязательства (долги). Для всех трех переменных количество пропусков менее 5%, влияние на результат скорее всего будет незаметно. Однако, смысл этих переменных достаточно интересен с точки зрения оценки кредитоспособности, поэтому найденные записи с пропусками не были исключены.

Далее нужно исследовать корреляции переменных. В результате исследования должны быть исключены переменные, между которыми обнаружится зависимость, для того, чтобы результаты предсказания модели не были искажены. Рекомендуется [2] использовать коэффициент толерантности (для определения наличия мультиколлинеарности), который находится как $1 - R^2$, где R — значение коэффициента корреляции i-ой переменной с остальными.

Для рассматриваемой выборки проведено исследование корреляции переменных, результаты сведены в Табл. 2.

В качестве порогового значения коэффициента толерантности обычно принимают $0,2$. Как видно из результатов, анализ не выявил проявления мультиколлинеарности между зависимыми переменными. Самым близким к пороговому значению является коэффициент для переменной «Запрашиваемая сумма» и переменной «Цена товара». Это объясняется тем, что заемщик запрашивает сумму исходя из конкретных целей приобрести тре-

Таблица 2: Результат вычисления коэффициентов толерантности

	Опыт работы	**Тип жилья**	**Срок кредита**	**Возраст**	**Семейное положение**	**Судимости**	**Тип работы**	**Расходы**	**Доходы**	**Активы**	**Обязательства**	**Запрашиваемая сумма**	**Цена товара**
Опыт работы	—												
Тип жилья	0,98	—											
Срок кредита	1	1	—										
Возраст	0,74	0,93	1	—									
Семейное положение	0,97	0,93	1	0,89	—								
Судимости	1	1	1	1	0,99	—							
Тип работы	0,99	1	0,98	0,97	1	1	—						
Расходы	0,98	0,89	1	0,94	0,96	1	1	—					
Доходы	1	1	1	1	1	1	1	0,99	—				
Активы	1	1	1	1	1	1	1	1	1	—			
Обязательства	1	1	1	1	1	1	1	1	0,99	0,62	—		
Запрашиваемая сумма	1	1	0,81	1	1	0,99	1	1	1	1	1	—	
Цена товара	1	1	0,98	1	1	1	1	1	1	1	1	0,47	—

буемый товар. По итогам исследования корреляции переменных принято обоснованное решение их не исключать.

Необходимо провести исследование на значимость переменных. Для этого можно опираться на показатель IV — Information Value — в модели следует оставить те переменные, для которых значение этого показателя будет наибольшим. IV характеризует степень влияния независимой переменной на зависимую. Вычисление данного параметра рассмотрено в последующем разделе о взаимосвязи независимых переменных с зависимой.

Формирование выборок и определение их объема

Для корректного обучения модели требуется разделить выборку случайно на две части: обучающую и тестовую. По первой проходит обучение модели, т.е. непосредственное ее формирование, а по второй проходит контроль качества модели — известно значение зависимой переменной, ему сопоставляется предсказанное моделью. Обычно на практике соотношение объема выборок $7:3$ или $8:2$.

Для целей работы требуется оценить предсказательную способность модели отличать «плохих» и «хороших» заемщиков. Объем выборки определяется на основании критерия мощности [2] при задании максимально допустимой ошибки оценки соотношения «плохих» и «хороших» заемщиков в генеральной совокупности по формуле:

$$n = \frac{Z_y^2 w(1-w)}{\Delta_w^2}, \tag{12}$$

где n — минимальный объем выборки, Z_y — значение стандартного нормального закона распределения для уровня надежности y, w — доля «плохих» клиентов в тестовой выборке, Δ_w — максимально допустимая предельная ошибка оценки доли «плохих» заемщиков.

В рамках данной работы в качестве цели принято, что отношение «плохих» и «хороших» заемщиков в тренировочной выборке должно совпадать с их соотношением во всей генеральной совокупности с долей уверенности 95%. Для значений $w = 0,28$ (соотношение из исходной выборки), $Z_y = 1,96$ (для уровня 95%), $\Delta_w = 10\%$ или $0,028$ значение n будет порядка 1000 записей. Рассматриваемая выборка удовлетворяет этому требованию. Разбиение на тренировочную и тестовую выполнено в пропорции 7:3 случайным образом.

Разделение значений количественных переменных по категориям

На практике часто используются модели с категориальными переменными. Это имеет свои преимущества: облегчается обработка шумов и выбросов, упрощается чтение скоринг-карты. Процесс категоризации описывается следующим алгоритмом [2]:

1. Разбить переменную (количественную) на группы на основании равных процентилей;
2. Посчитать в каждой группе доли «хороших» и «плохих» заемщиков;
3. Рассчитать для каждой категории показатель WOE (Weight of Evidence)

$$WOE_i = \ln \frac{d_i(1)}{d_i(2)}, \tag{13}$$

   где $d_i(1)$ и $d_i(2)$ — относительные частоты «плохих» и «хороших» заемщиков, соответственно, в i-ой группе категоризованной переменной; $i = 1, 2, \ldots, k$, k — число категорий переменной.

4. Проанализировать показатели WOE групп и объединить соседние (те, для которых эти значения наиболее близки). При объединении следует руководствоваться следующими принципами: не должно быть групп с нулевым количеством «плохих» или «хороших» заемщиков; значения WOE должны возрастать или убывать при переходе между группами;
5. Произвести пересчет показателей WOE для новых групп.

Так например, для характеристики «Возраст» без процедуры категоризации нельзя учесть тот факт, что молодые и пожилые заемщики возвращают кредиты реже, чем заемщики среднего возраста. Показатель WOE описывает вес категории переменной и определяет границы, чувствительные к событию дефолта, помогая точнее выделять категории.

Далее приводится пример категоризации переменной «Опыт работы». На Рис. 2 представлена взаимосвязь между показателем WOE и переменной до процедуры биннинга [2], в Табл. 3 приведены численные показатели графика. Из графика видно, что, например, для категорий 6 и 7 веса практически одинаковы, соответственно эти категории одинаково чувствительны к риску и их можно объединить в одну.

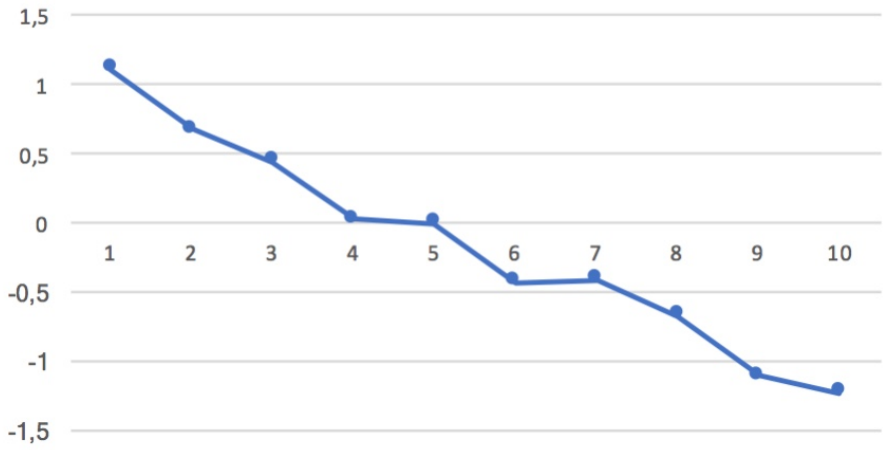

Рис. 2: До процедуры биннинга

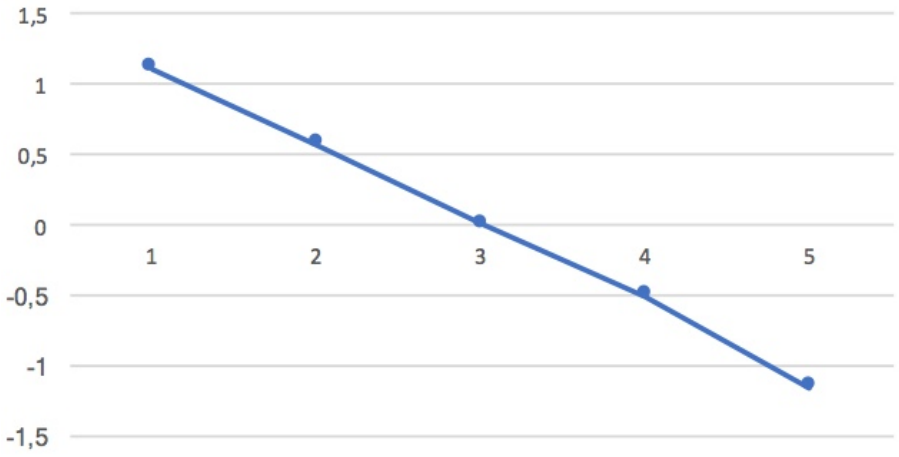

Рис. 3: После процедуры биннинга

После категоризации получена переменная, разбитая на 5 категорий. На Рис. 3 представлен график значений WOE после проведения категоризации, Табл. 4 содержит численные показатели графика. График показывает, что категории явно разделяются, соответственно при категоризации параметра заемщика можно будет четко наблюдать, насколько он влияет на событие дефолта. Например, 5-ая категория (опыт работы больше 10 лет) обладает низкой чувствительностью к рисковому событию, а 1-ая (без опыта) — высокой.

Процедура была проведена над всеми переменными, в сумме по всем характеристикам было выделено 60 категорий. Разбиение переменных на категории отражено в Приложении А, в Табл. 10.

Таблица 3: Численные показатели до процедуры биннинга

	«Хорошие»	**«Плохие»**	**WOE**
0	245	290	1,105
1	228	222	0,677
2	282	172	0,442
3	240	96	0,02
4-5	359	140	-0,005
6-8	383	98	-0,426
9-10	250	65	-0,410
11-14	369	74	-0,67
15-20	471	61	-1,107
21+	313	36	-1,226

Таблица 4: Численные показатели после процедуры биннинга

	«Хорошие»	**«Плохие»**	**WOE**
0	245	290	1,105
1-2	570	394	0,568
3-5	599	236	0,005
6-15	1002	237	-0,505
15+	784	97	-1,153

Таблица 5: Значение показателя Information Value

	Опыт работы	**Тип жилья**	**Срок кредита**	**Возраст**	**Семейное положение**	**Судимости**	**Тип работы**	**Расходы**	**Доходы**	**Активы**	**Обязательства**	**Запрашиваемая сумма**	**Цена товара**
IV	0,51	0,25	0,1	0,12	0,15	0,34	0,33	0,1	0,4	0,24	0,18	0,3	0,04

Взаимосвязь независимых и зависимой переменных

Для сокращения числа переменных в модели (для избежания переобучения) и расчета степени взаимозависимости независимых и зависимой переменной используется показатель IV, о котором упоминалось ранее. Данный показатель вычисляется по формуле [2]

$$IV = \sum_{i=1}^{k} (d_i(1) - d_i(2))WOE_i, \tag{14}$$

где k — число категорий независимой переменной.

В Табл. 5 представлены значения показателя IV для каждой переменной выборки.

Чем выше IV для переменной, тем более существенный вклад она вносит в модель. В литературе [2] рекомендуется использовать следующую градацию при выборе переменных: если значение IV менее 0.02 — независимая переменная не обладает прогностической способностью; от 0.02 до 0.1 — низкая прогностическая способность; от 0.1 до 0.3 — средняя прогностическая способность; от 0.3 до 0.5 — высокая прогностическая способность; более 0.5 — превосходная прогностическая способность. Как видно из приведенной Табл. 5 переменная «Цена товара» имеет низкую прогностическую способность — этот параметр исключен из дальнейшего рассмотрения. Наиболее весомыми являются параметры «Опыт работы» и «Судимость».

По итогам всех действий исходная выборка была полностью исследована на качество, каждая операция над данными и результаты обоснованы

и проанализированы. Полученная выборка разбита на тестовую и обучающую выборки случайным образом в пропорции 7:3.

2.2. Модель предметной области

На основании проведенных исследований построена модель предметной области в виде диаграммы с изображением выделенных концептуальных классов, их атрибутов и ассоциаций между ними (Рис. 4).

Класс « <название параметра> » описывает множество классов категоризированных параметров Заемщика, одинаковых по структуре (объединены для лучшей читаемости схемы). По этим же причинам на схему не попали некоторые служебные объекты, которые в целом не влияют на логическую модель системы.

На основании полученной схемы спроектирована база данных, в которую первично была занесена информация по текущей выборке (класс «Заемщик (физ.лицо)»). Объекты рейтингов и параметров заданы по результатам решения основной задачи обучения модели. В остальные сущности информация заносилась уже в процессе тестирования Системы.

2.3. Логистическая регрессия

Логистическая регрессия чаще всего используется в кредитном скоринге, если в качестве зависимой принята бинарная переменная. Математически модель логистической регрессии выражает зависимость логарифма шанса (логита) от линейной комбинации независимых переменных [2]:

$$\ln \frac{pi}{1-pi} = b_0 + b_1 x_i(1) + b_2 x_i(2) + \cdots + b_n x_i(n) + e_i, \qquad (15)$$

где

p_i — вероятность наступления дефолта по кредиту для i-го заемщика;

$x(j)$ — i значение j-ой независимой переменной;

b_0 — независимая константа модели;

b_j — параметры модели;

e_i — компонент случайной ошибки.

Уравнение (15) показывает линейную зависимость вероятности наступления дефолта и независимыми переменными. Константа означает уровень риска при нулевом значении всех скоринговых переменных. Значения

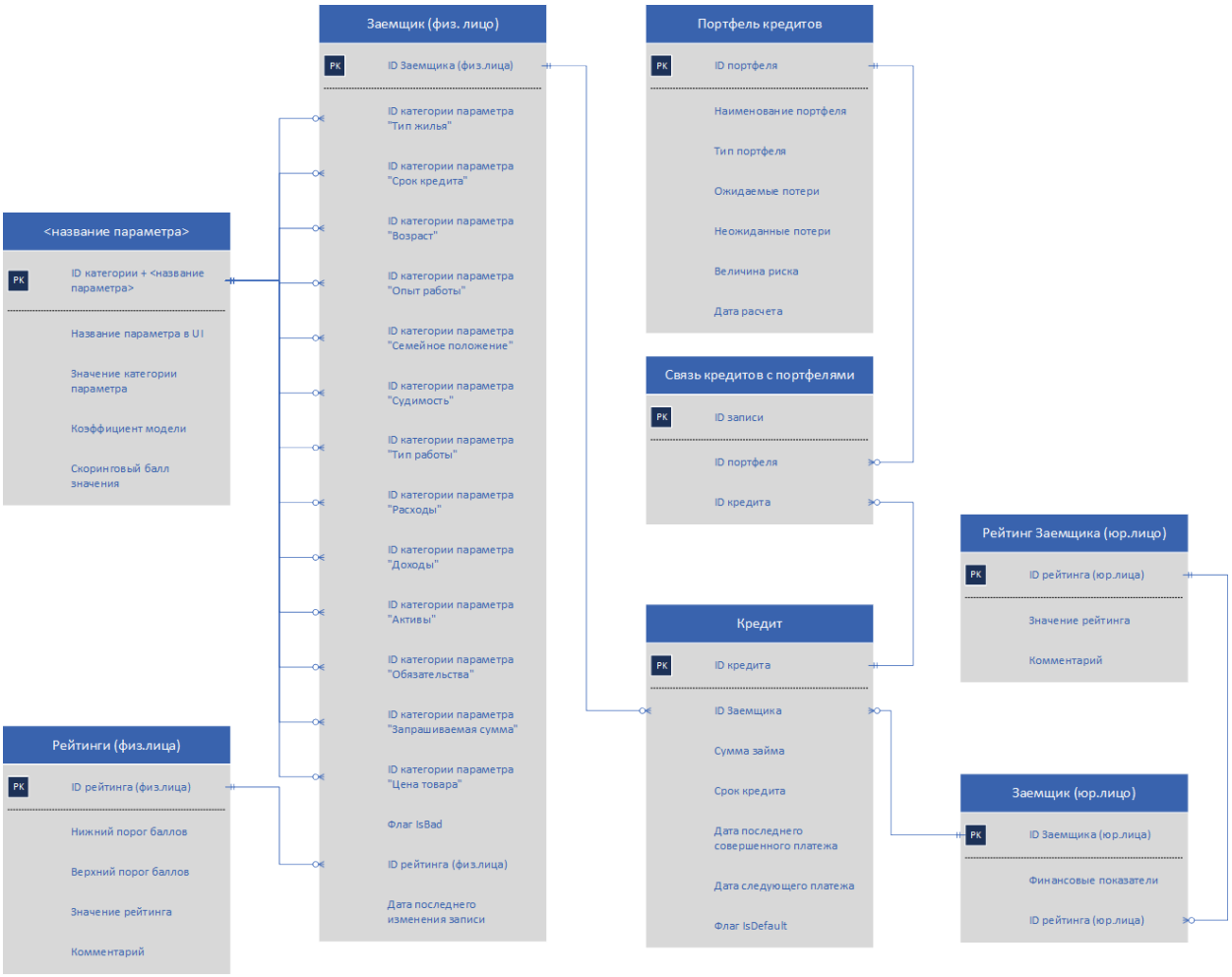

Рис. 4: Модель предметной области

коэффициентов при независимых переменных в дальнейшем участвуют при вычислении скоринговых баллов.

Прежде чем перейти к обучению модели, были изучены методы исключения независимых признаков для улучшения предсказательных качеств модели [35]. Статистические методы исследования каждого признака в отдельности уже применялись в разделе выше — так была исключена одна независимая переменная. Однако после проведения категоризации всех количественных переменных возник вопрос о переобучении модели и требовалось рассмотреть признаки в совокупности. Для этих целей выбран метод RFE (recursive feature elimination) — это метод последовательного исключения признаков из модели. Однако, при большом количестве признаков этот метод может не решить проблему переобучения модели. Поэтому в целях исследования были проверены обе реализации: с использованием метода исключения признаков и без его использования. В результате качество моделей обеих реализаций не имело существенного различия, поэтому в даль-

нейшей работе используется модель с полным набором признаков. Однако применение метода RFE позволило выделить наиболее влиятельные категории характеристик, например, для характеристики «опыт работы» в модель вошла всего одна категория — «опыт работы более 10 лет», «Судимость» важна в обоих случаях («нет»/«есть»), а характеристика «Возраст» вообще не имела влияния на результат. Полученные результаты по категориям подтверждают вычисленные значения IV в целом для характеристик.

Значения коэффициентов модели (в обеих реализациях) представлены в Приложении А в Табл. 10, константа $b_0 = 0.25743371$. Модель присваивает «хорошим» заемщикам значение 0, а «плохим» — 1. Анализ результатов показывает, что по умолчанию (при пустой скоринговой анкете) заемщик будет считаться «плохим», т.к. константа положительна и модель присвоит такому набору характеристик значение 1. Такое решение логично, т.к. нельзя стать «хорошим» заемщиком, не предоставив данные о себе. Знак коэффициентов при переменных показывает их вклад в «положительность» или в «отрицательность» заемщика. Например, характеристика «Срок кредита» разделена на 5 категорий, из них «самой положительной» (вносящей наибольший вклад в то, что заемщик «хороший») является категория «срок от 6 до 18 месяцев», т.к. значение коэффициента при ней наименьшее, а при переменной категории «срок от 42 месяцев и выше» коэффициент наибольший — заемщик с таким параметром расценивается как более «плохой». Подобным образом характеризуются все категории. В результате анализа полученных коэффициентов можно сказать, что они отвечают реальной статистике дефолтов.

Кроме того, в целях исследования была использована модель принятия решений, опирающаяся на метод случайного леса. Однако время обучения по этому методу было существенно выше при схожих результатах, поэтому далее он не рассматривается.

Оценка качества модели

Оценка качества предсказательной возможности модели — это один из важнейших пунктов при построении модели [3]. Важно построить такую модель, которая будет хорошо определять и «хороших» и «плохих» кредиторов.

Как правило, для оценки качества модели в машинном обучении ис-

пользуют ROC-кривую, отражающую соотношение долей верно найденных несущих признак (положительных) исходов и неверно найденных не несущих признак (отрицательных) исходов. По-другому ROC-кривая называется кривой ошибок, чем ближе кривая к левому верхнему углу графика, тем лучше предсказательная способность модели (Рис. 5). Площадь под кривой ошибок — показатель AUC — отражает качество классификации модели. Чем больше значение AUC, тем лучше предсказывает модель.

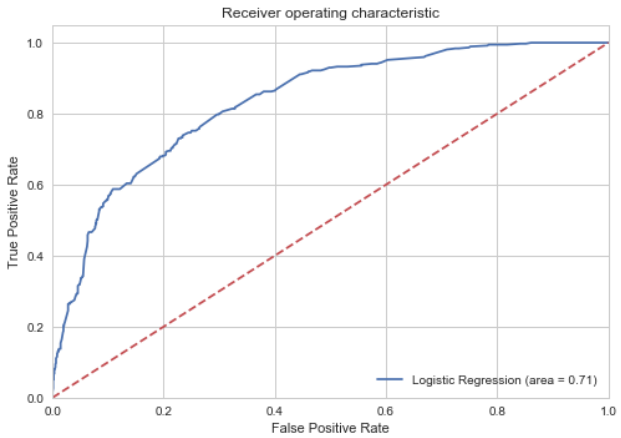

Рис. 5: Кривая ошибок

По значению AUC вычисляется такой показатель, как индекс Джини [27]. Этот показатель переводит значение площади под кривой в диапазон от 0 до 1 и вычисляется по формуле

$$G = 2(AUC - 0.5) \tag{16}$$

Также для оценки качества модели может использоваться тест Колмогорова-Смирнова [3]. Он показывает, насколько хорошо происходит разделение «хороших» и «плохих» заемщиков. В тесте Колмогорова–Смирнова сравниваются два кумулятивных распределения скоринговых баллов «хороших» и «плохих» заемщиков. Статистика Колмогорова–Смирнова вычисляется как максимальная разница между кумулятивными функциями этих распределении. Диапазон значении статистики от 0 до 100, и чем выше ее значение, тем лучше работает модель. Этот тест проведен в следующем разделе после построения итоговой скоринговой карты.

Точность модели равна 0,8 на тестовой выборке. Показатель $AUC = 0.71$ (отражено на Рис. 5), коэффициент Джини $G = 0.42$. Также был рассчитан показатель успешных предсказаний «хороших» и «плохих» заемщиков по отдельности (Рис. 6). Из графика видно, что около 82% и

72% «хороших» и «плохих» заемщиков, соотвественно, были определены верно. Кроме того, успешность обучения модели доказывает тест кроссвалидации [4]: $CV = 0.79$, что близко к вычисленной точности, дальнейшее улучшение при тех же параметрах не принесет результатов. Для выборки такого объема это приемлемые показатели качества — модель обладает хорошей прогностической способностью, следовательно можно продолжать исследование с полученными коэффициентами при переменных.

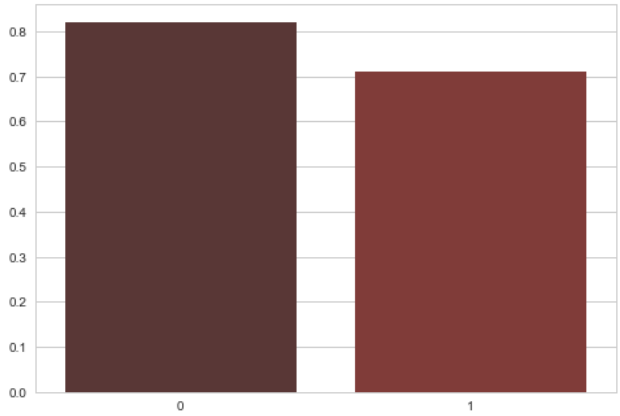

Рис. 6: Показатель успешных предсказаний «хороших» и «плохих» заемщиков

2.4. Построение рейтинговой системы

Перевод коэффициентов модели в скоринговые баллы

Перевод коэффициентов модели в баллы является завершающим этапом в построении скоринг-модели [2].

Скоринговый балл в логарифмической шкале считается следующим образом: коэффициенты модели умножаются на значения соответствующих независимых переменных и полученные произведения суммируются. Однако для чтения результатов удобнее масштабировать баллы, т.е. перевести их в линейную шкалу.

За основу взят общепринятый стандарт FICO [36]. В первую очередь задается диапазон, например, от 0 от 1000. Затем вводится такой показатель D, характеризующий удвоение шанса стать «хорошим» заемщиком, чаще всего $D = 40$. Т. е. каждые 40 баллов у наблюдаемого заемщика удваиваются его шансы стать «хорошим» заемщиком.

Для преобразования коэффициентов модели в скоринговый балл (B) используется следующий подход к масштабированию [5]:

$$B = A + Rb_j, \tag{17}$$

где A — смещение; R — множитель; b_j — коэффициент модели.

Множитель R рассчитывается следующим образом:

$$R = \frac{D}{\ln 2}, \tag{18}$$

а смещение A по формуле

$$A = B_0 + R \ln C, \tag{19}$$

где B_0 — значение в баллах, в котором шанс стать хорошим заемщиком составляет C:1. Согласно FICO рекомендуется брать $C = 72$ и $B = 660$, как общепринятый стандарт расчета скоринговых баллов [36]. Однако в целом нет ограничений для выбора масштаба, поэтому в данной работе для удобства чтения все баллы «сдвигаются влево» на минимальный среди всех балл.

Выражение (19) переводит баллы в «восходящие». Например, для характеристики «Расходы» коэффициенты регрессии b_i и баллы $score_i$, полученные по формуле (19), следующие:

- > 80 у.е.; $b_1 = 0,775$; $score_1 = 458$;
- $35 - 80$ у.е.; $b_2 = -0,153$; $score_2 = 404$;
- ≤ 35 у.е.; $b_3 = -0,37$; $score_3 = 392$.

Чем меньше коэффициент, тем больше вклад он делает в «положительность» заемщика, т.к. 0 — «хороший», 1 — «плохой» заемщик. Однако после масштабирования должно быть наоборот — чем выше балл, тем лучше. Следовательно, на примере характеристики «Расходы», множитель нужно брать с обратным знаком.

Для получения скоринг-балла по заемщику необходимо сложить полученные баллы по всем независимым переменным модели.

После расчета скоринговых баллов по каждому Заемщику из тестовой выборки был проведен тест Колмогорова-Смирнова (Рис. 7), описанный в разделе оценки качества. Значение статистики $0,56$, что превышает критическое значение для уровня достоверности $0,01$. Следовательно, полученная модель достаточно хорошо разделяет «хороших» и «плохих» заемщиков, значит на основании модели можно построить эффективную рейтинговую систему.

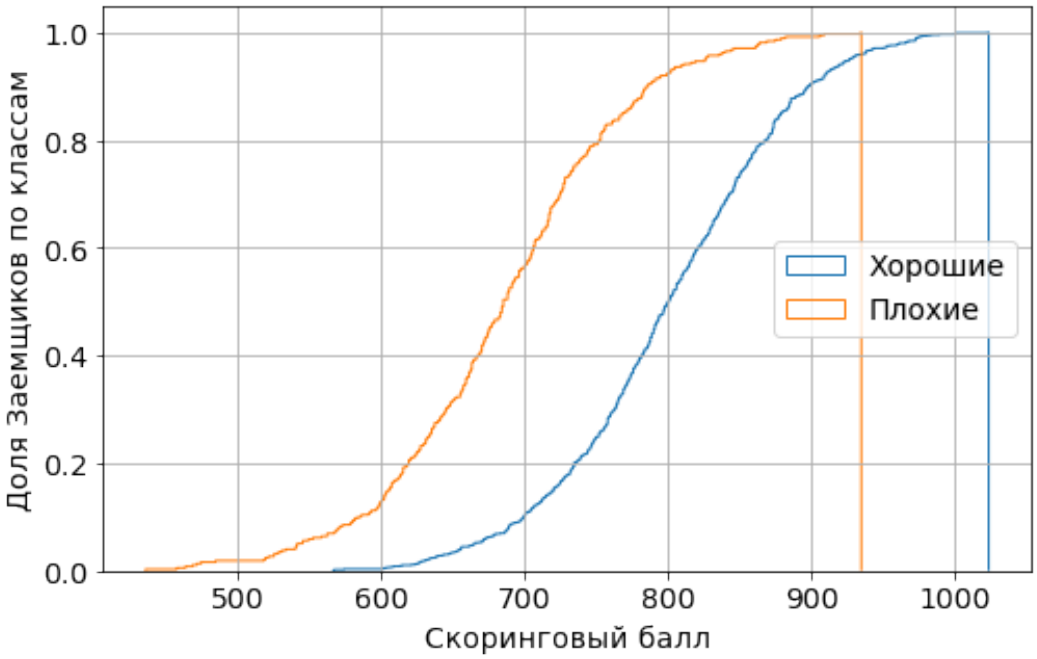

Рис. 7: Кумулятивные распределения скоринговых баллов

Рейтинговая система

1. Физические лица. Рейтинговая система физических лиц основа на полученной скоринг-карте. Общепринятой [36] градацией баллов является следующее распределение:

- 690 – 850 баллов — отличный балл. Можно предложить заемщику более выгодные условия.
- 650 – 690 баллов — стандартный балл. заемщик проходит по общим условиям.
- 600 – 650 баллов — удовлетворительный балл. Рекомендуется пересмотреть запрос Заемщика (уменьшить одобренную сумму).
- 500 – 600 баллов — балл ниже среднего. Следует внимательно проанализировать цели кредита, срок и ставку и одобрить только на определенные категории товаров/услуг и на небольшой срок.
- 300 – 500 баллов — плохая оценка. Такому заемщику следует отказать.

Так как масштабирование коэффициентов производилось со смещением, то использовать разбиение рейтингов в таком же виде нельзя. В данной работе для выделения рейтинговых групп использовался алгоритм

Таблица 6: Соотношение «хороших» и «плохих» Заемщиков

Рейтинг	**«Хорошие»**	**«Плохие»**
>904	87	3
845-904	201	10
795-844	226	19
745-794	239	52
<745	130	94

классификации k-means, количество групп принято равным 5, как и в рекомендованной рейтинговой системе. Диапазоны распределения баллов и количественное соотношение «хороших» и «плохих» заемщиков в тестовой выборке по каждому рейтингу представлено в Табл. 6.

Результаты показывают, что большая часть «плохих» заемщиков попала в последние две группы. Кроме того, некоторое количество «хороших» заемщиков также попало в последнюю группу. По рекомендованной рейтинговой системе последней группе следует отказать, значит эти заемщики составляют так называемый *коммерческий риск* — риск отказать потенциально «хорошему» клиенту. Однако, как будет показано далее, заемщики из этой группы имеют достаточно высокую вероятность дефолта.

2. Юридические лица. В разработанном программном комплексе также предусмотрена работа с юридическими лицами. Рейтинговая система юридических лиц основана на финансовых показателях и рекомендациях нормативных актов (аналитический расчет коэффициентов). Источниками финансовых показателей является предоставляемая заемщиком документация и отчетность РСБУ и МСФО. Также на рейтинг может влиять динамика изменения показателей отчетности по годам или внутренняя и внешняя история по заемщику (например, опыт работы с наблюдаемым контрагентом).

Для определения внутреннего рейтинга среди показателей рассчитываются:

1. Коэффициент покрытия;
2. Коэффициент быстрой ликвидности;
3. Оборачиваемость оборотного капитала;

4. Коэффициент финансовой устойчивости;

5. Рентабельность продаж;

6. Рентабельность активов;

7. Репутация (опыт работы, дисциплина.

Всего выделено 12 рейтинговых классов для юридических лиц (см. Табл. 7), определение рейтинга происходит в два этапа: уровень контрагента определяется экспертным путем, а скоринговый балл вычисляется по перечисленным выше коэффициентам. Далее в работе продолжается исследование анкет физических лиц, так как в Системе отсутствует информация по историческим данным юридических лиц. По мере накопления данных в Системе появится возможность работы с юридическими лицами.

Таблица 7: Рейтинговая система для юридических лиц

Рейтинг	**Описание**
A1	Высокий уровень контрагента с высоким скоринговым баллом
A2	Высокий уровень контрагента с хорошим скоринговым баллом
A3	Высокий уровень контрагента со средним скоринговым баллом
A4	Высокий уровень контрагента с низким скоринговым баллом
B1	Средний уровень контрагента с высоким скоринговым баллом
B2	Средний уровень контрагента с хорошим скоринговым баллом
B3	Средний уровень контрагента со средним скоринговым баллом
B4	Средний уровень контрагента с низким скоринговым баллом
C1	Низкий уровень контрагента с высоким скоринговым баллом
C2	Низкий уровень контрагента с хорошим скоринговым баллом
C3	Низкий уровень контрагента с со средним скоринговым баллом
C4	Низкий уровень контрагента с низким скоринговым баллом

Глава 3. Моделирование кредитного риска

3.1. Метод VaR

В этой главе рассматривается построение модели для определения кредитного риска с помощью оценки VaR [12].

VaR (Value-at-Risk) — выраженная в заданной валюте оценка убытков, которую с определенным уровнем доверия не превысит кредитный портфель в течение заданного времени (20).

$$P(L_p \leq VaR) \geq p, \tag{20}$$

где L_p — убытки по кредитному портфелю, а p — уровень доверия.

Для оценки кредитного риска строится эмпирическая функция распределения убытков по текущему портфелю (далее будем рассматривать оценку риска для описанной ранее выборки заемщиков — физических лиц). Для построения функции выбран метод имитационного моделирования Монте Карло (см. Главу 1).

Оценку VaR делят на две составляющие: ожидаемые потери (EL_p) и неожиданные потери (UL_p) [13]. Величина EL_p — среднее значение убытков, наступивших вследствие дефолтов по кредитным договорам. UL_p представляют собой отклонение убытков от их ожидаемого значения.

В качестве временного горизонта принят обычно задаваемый банками один год, а уровень доверия Базельским комитетом рекомендуется устанавливать 99%. Как было показано в Главе 2, модель с достаточной точностью определяет «плохих» и «хороших» заемщиков, следовательно для каждой выделенной рейтинговой группы можно рассчитать оценку вероятности наступления дефолта по кредитным обязательствам [13]. В выражении (21) представлен пример такого расчета для рейтинговой группы I:

$$PDefault_I = \frac{NDefault_I}{N_I}, \tag{21}$$

где

$PDefault_I$ — оценка вероятности дефолта заемщика с рейтингом I;

$NDefault_I$ — количество дефолтов среди заемщиков с рейтингом I;

N_I — количество заемщиков с рейтингом I.

Расчет оценок по всем выделенным рейтинговым группам представлен в Табл. 8.

Таблица 8: Оценка вероятности дефолта в каждой группе

Рейтинг	**Оценка**
>904	0,03
845-904	0,04
795-844	0,078
745-794	0,18
<745	0,42

По полученным результатам видно, что самой надежной группой является первая — дефолты в ней наступают в 3 случаях из 100. Последняя группа заемщиков имеют высокую вероятность дефолта по кредитным обязательствам, максимальные потери по этой группе будут большими. Как и рекомендовалось в универсальной рейтинговой системе, таким заемщикам будет отказано — соответствующие записи были исключены из тестовой выборки.

Ожидаемые потери

Далее необходимо оценить ожидаемые потери EL_p по портфелю [12]:

$$EL_p = \sum_{i=1}^{N} PDefault_i C_i (1 - E_i), \tag{22}$$

где

$PDefault_i$ — вероятность дефолта i-го заемщика;

C_i — величина, которую банк потеряет в случае дефолта заемщика. Обычно эта величина принимается равной задолженности и начисленным процентам (а также сопутствующие расходы) на момент дефолта. В связи с отсутствием размеченных данных такого рода за C_i принята сумма задолженности i-го Заемщика;

E_i — характеристика возмещения потерь при дефолте i-го Заемщика. В данной работе для упрощения примем данный параметр равным нулю. Характеристику легко ввести, присвоив каждому кредиту определенный признак обеспеченности с установленной величиной E_i, например, экспертным путем;

N — количество кредитов в портфеле.

Итоговые расчетные значения ожидаемых потерь по каждой рейтин-

Таблица 9: Ожидаемые потери

Рейтинг	**Ожидаемые потери**
>904	2143
845-904	9042
795-844	19101
745-794	52952
<745	102450
Итого	83238

говой группе и в целом по портфелю представлены в Табл. 9. Для сравнения в таблице приведены результаты по последней исключенной группе, но в итоговом расчете они не участвуют.

Значение EL_p составило 10% от всего портфеля. Это средний уровень риска. Если исключить 4-ую группу, заемщики которой обладают вероятностью 0,18 дефолта, то риски снизятся до хорошего уровня в 6%. Однако, такое исключение не учитывает коммерческий риск. Для понижения кредитного риска можно наложить на 4-ую группу дополнительные ограничения, например, задать понижающий коэффициент и отказывать заемщикам, «выходящим» за нижнюю границу группы после его применения. К примеру, 5%-ый понижающий коэффициент к 4-ой группе позволил снизить риск до 8,5%.

Имитационное моделирование

Следующим шагом необходимо оценить неожиданные потери с помощью метода VaR.

Для этого строится эмпирическая функция потерь с помощью метода имитационного моделирования Монте-Карло [12], [13]:

1. Генерируются равномерно распределенные на диапазоне $[0, 1]$ случайные величины $D_i^k \sim R_{[0,1]}$, где $i = 1, \ldots, N$, N — количество кредитов в портфеле. На этом этапе можно ввести матрицу связности для учета зависимостей Заемщиков между собой. В работах [38], [37] предлагаются подходы к учету корреляции дефолтов. В связи с отсутствием исторических данных такого характера на этапе проектирования системы статистические взаимосвязи между заемщиками не учитывают-

ся. Однако в процессе развития системы рекомендуется внести данное уточнение в модель.

2. Рассчитывается величина потерь по каждому заемщику с помощью обратной функции распределения [12]:

$$L_i^k = \begin{cases} C_i, & \text{если } D_i^k < PDefault_i \\ 0, & \text{если } D_i^k > PDefault_i, \end{cases} \tag{23}$$

где C_i — сумма задолженности i-го контрагента, $PDefault_i$ — вероятность дефолта i-го контрагента.

3. Рассчитывается сумма максимальных убытков по портфелю:

$$L_p^k = \sum_{i=1}^{N} L_i^k \tag{24}$$

4. Шаги 1-3 повторяются K раз. Есть несколько подходов к вычислению количества итераций [12], в данной работе величина K принята по умолчанию равной 10^4. Система позволяет задать пользовательское значение количества необходимых итераций. По полученной выборе максимальных убытков строится эмпирическая функция распределения (Рис. 8).

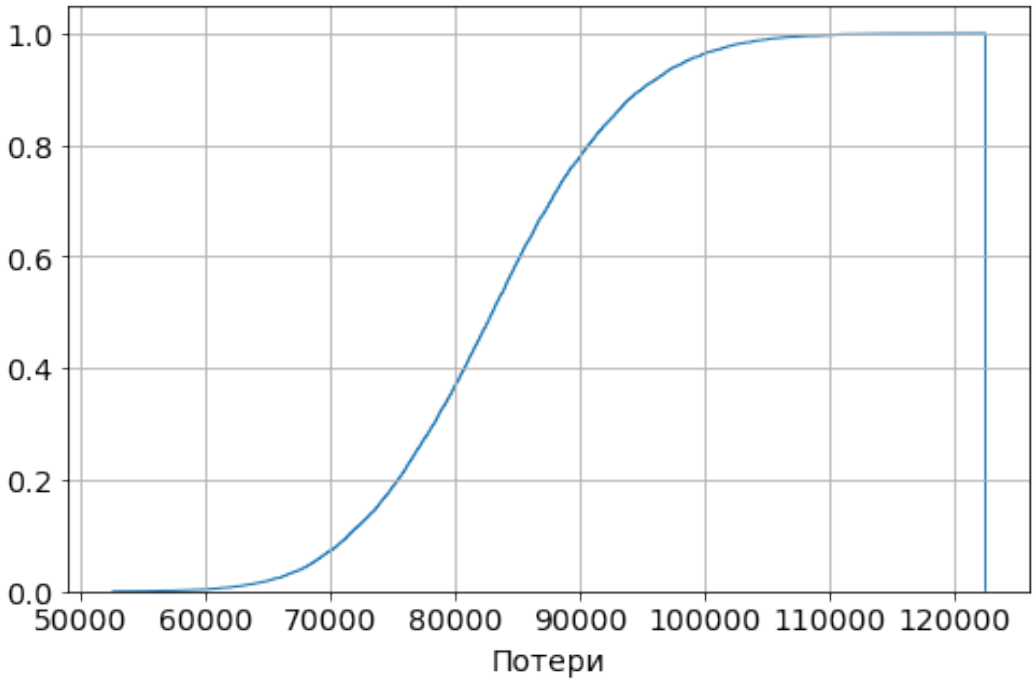

Рис. 8: Функция распределения максимальных убытков

Ранее был определен уровень доверия $0,99$. Для данного уровня доверия находится $P(L \leq VaR) \geq 0,99$. Полученное значение максимальных убытков $VaR_{99\%}$ составило 105157 у.е..

Далее рассчитываются неожиданные потери:

$$UL_p = VaR_{99\%} - EL_p = 105157 - 83238 = 21919. \quad (25)$$

Величина максимальных убытков составила 13% от общего объема портфеля. Анализ результатов показывает, что уровень риска по максимальным потерям находится на среднем уровне. Как уже упоминалось, можно воспользоваться понижающим коэффициентом для более гибкого выделения потенциально «плохих» клиентов. Также стоит отметить, что в приведенной выборке заемщиков нет информации по обеспеченности кредита, по процентной ставке и уже выплаченной части ссуды на момент дефолта. Полученные результаты показывают уровень риска при наихудших условиях, когда все выданные кредиты не обеспечены, и при дефолте кредитор теряет всю величину ссуды по сделке. На практике стремятся к обеспеченности кредита. В реализованную в данной работе модель достаточно легко включить эти параметры — при расчете ожидаемых потерь по заемщику следует брать не всю сумму кредита, а только его необеспеченную часть.

В заключение раздела следует отметить следующее:

1. Оценка кредитного риска портфеля должна пересчитываться при изменении состава портфеля;
2. Рейтинговая система должна обновляться при внесении новых данных о дефолтности заемщиков — отдельные признаки могут стать более значимыми для общей оценки кредитоспособности контрагента;
3. Рассчитанные ожидаемые потери используются для расчета объема резервов по каждому кредиту как способ понижения риска;
4. Неожиданные потери отражают уровень надежности кредитного портфеля.

3.2. Управление кредитным риском

Предпоследним этапом в управлении рисками является выбор способа разрешения риска. Чаще всего наблюдается стремление к понижению

уровня риска. Далее рассмотрены некоторые наиболее распространенные методы минимизации риска [9]:

1. Страхование позволяет снизить кредитный риск за счет его передачи третьему лицу Страховщику за некоторую страховую премию. Сумма страхования рассчитывается по общей сумме долга. Чаще всего таким способом покрывается не весь риск — в таком случае на остаток можно воспользоваться, например, методом резервированием средств.

2. Залоги (обеспечение обязательств). Метод позволяет снизить риск по ссудам за счет их обеспеченности залогом.

3. Лимитирование заключается в установке определенного ограничения на размер выдаваемых ссуд. Таким образом подобные ограничения на выдаваемые кредиты позволяют уменьшить убытки в случае дефолта.

4. Диверсификация — процесс распределения кредитного риска по различным направлениям, не связанным между собой. Например, по отраслям, по регионам, по категориям заемщиков (предприятия малого и среднего бизнеса). Однако диверсификация должна сопровождаться подробным анализом и прогнозированием, иначе это может привести к обратному росту кредитного риска.

5. Резервирование или самострахование состоит в создании фондов для возмещения потерь при дефолтах за счет собственных средств. Важнейшая характеристика резерва — это оптимальный объем запасов. Задача поиска оптимального объема запасов решается в теории управления запасами, которая в данной работе не рассматривалась. Однако, как упоминалось в предыдущем разделе, рассчитанные ожидаемые потери могут использоваться для расчета объема резервов по наблюдаемому кредиту. Для банковского сектора ЦБ РФ предусмотрены нормативные рекомендации по вычислению резервов на возможные потери.

В данной работе предполагается, что в каждом случае решение по управлению риском принимается непосредственно руководителем. Создание единой автоматизированной системы для принятия таких решений является отдельной обширной задачей и рассматривается автором как перспективы расширения функциональности разработанного программного комплекса.

Заключение

В данной работе был проведен анализ количественных методов оценки рисков, применяющихся в риск-менеджменте. Среди них были выделены методы машинного обучения, имитационного моделирования и метод VaR. Далее в работе был проведен подробный анализ выборки данных, их очистка и приведение к требуемому виду. В процессе работы автором было продемонстрировано, что выбранные данные качественны. Подготовленные данные использовались для обучения и тестирования модели логистической регрессии. Рассчитанные метрики качества показали высокий уровень прогностической способности модели. Полученная модель позволила сформировать скоринговую карту, хорошо разделяющую «хороших» и «плохих» заемщиков. Эффективное разделение прослеживается в низких вероятностях дефолта «хороших» рейтинговых групп (1-3). Четвертая рейтинговая группа — это группа с повышенным риском, по ней автором предложено формировать дополнительные ограничения и обеспечение кредитов с таким рейтингов залогами для понижения общего уровня кредитного риска портфеля.

Скоринговая карта применялась в расчете максимальных убытков с помощью метода имитационного моделирования Монте-Карло. В работе было показано, что имитационная модель позволяет сформировать величину максимальных убытков, характеризующую кредитный риск. При выдаче новых кредитов (и в целом при изменении) портфеля можно оперативно пересчитать уровень риска в соответствии с введенными изменениями. Оцененный уровень риска помогает принять решение по его дальнейшему управлению.

Разработанный программный комплекс может использоваться не только банковскими структурами, но и предприятиями, предоставляющими услугу кредитования без попечительства банка.

Список литературы

[1] Гост Р 51897-2011. Менеджмент риска. Термины и определения. // http://docs.cntd.ru/ URL: http://docs.cntd.ru/document/gost-r-51897-2011 (дата обращения: 15.12.2017).

[2] Сорокин А.С. // Построение скоринговых карт с использованием модели логистической регрессии. Науковедение №2, 2014

[3] Сорокин А.С. // К вопросу валидации модели логистической регрессии в кредитном скоринге. Науковедение, выпуск 2, 2014.

[4] Golub, G. H., Heath, M. and Wahba G. // Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21, p. 215–223, 1979.

[5] Naeem Siddiqi // Intelligent Credit Scoring. Building and Implementing Better Credit Risk Scorecards. John Wiley Sons Limited, 2017

[6] Лаврушина О.И., Валенцева Н.И. // Банковские риски. КНОРУС, Москва, 2013.

[7] Е. В. Иода, Л. Л. Мешкова, Е. Н. Болотина // Классификация банковских рисков и их оптимизация. ТГТУ, Тамбов, 2002.

[8] Балабаев В.Е. // Статистический анализ финансовых рисков. Ярославль, ЯрГУ, 2015. 60 с.

[9] Шапкин А. С., Шапкина В. А. // Теория рисков и моделирование рисковых ситуаций. М.: Издательско-торговая корпорация «Дашков и Ко», 2005. 880 с.

[10] Marie-Laure Charpignon, Enguerrand Horel, Flora Tixier // Prediction of consumer credit risk, 2014.

[11] Amir E. Khandani, Adlar J. Kim, and Andrew W. LoConsumer. Credit Risk Models via Machine-Learning Algorithms // Journal of Banking Finance 34, p. 2767-2787, 2010.

[12] Ивлиев С.В. // Исследование кредитного риска методом Монте-Карло. [Электронный ресурс] URL: https://www.cbr.ru/ckki/?PrtId=restr (дата обращения 17.04.2018)

[13] Никулина О.В., Коваленко А.И. // Управление кредитными рисками коммерческих банков в условиях нестабильности финансовой системы. Финансы и кредит 30, с.2-17, 2015.

[14] Бабичева Ю.А. // Банковское дело. Справочное пособие. М.: Экономика, 1993. 297 с.

[15] [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Банковский_риск (дата обращения 12.12.2017)

[16] Alexander J. McNeil, Rüdiger Frey, Paul Embrechts // Quantitative Risk Management. Princeton University Press, 2005.

[17] Aijun Zhang // Statistical Methods in Credit Risk Modeling. Dis. PhD, 2009.

[18] Кудрявцев А. А., Радионов А. В // Введение в количественный риск-менеджмент. Изд-во С.-Петерб. ун-та, СПб, 2016. 192 с.

[19] Chesser D. // Predicting loan noncompliance. The Journal of Commercial Bank Lending, August, p. 28–38, 1974.

[20] Altman E.I. // Financial rations. Discriminent analysis, and the prediction of corporate bankruptcy. Journal of Finance, September, 1968.

[21] Виленский П.Л., Лившиц В.Н., Смоляк С.А. // Оценка эффективности инвестиционных проектов. Теория и практика. М.: «Дело», 2004. 888 с.

[22] Дирочка А.А., Меньшиков И.С. // Доходность и дюрация портфеля облигаций. URL: http://www.fast.ane.ru/menshikov/yd.pdf

[23] Galindo J., Tamayo P. // Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics 15, p. 107-143, 2000.

[24] Лукашевич, Н.С. // Сравнение нейросетевых и статистических методов оценки кредитного риска. Финансы и кредит 1 (433), с. 32-41, 2011.

[25] Hosmer, D., Lemeshow, S. //Applied logistic regression. John Wiley and Sons, 2000.

[26] [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=Логистическая_регрессия (дата обращения 27.01.2018)

[27] Соколов Е. // Семинары по решающим деревьям, 2013. URL: http://www.machinelearning.ru/wiki/images/8/89/Sem3_trees.pdf

[28] [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Генетический _алгоритм (дата обращения 27.01.2018)

[29] [Электронный ресурс] URL: http://scikit-learn.org/stable/documentation.html (дата обращения 19.03.2018)

[30] [Электронный ресурс] URL: http://www.numpy.org (дата обращения 19.03.2018)

[31] [Электронный ресурс] URL: https://www.scipy.org (дата обращения 19.03.2018)

[32] [Электронный ресурс] URL: https://pandas.pydata.org (дата обращения 19.03.2018)

[33] [Электронный ресурс] URL: http://www.lsi.upc.edu/%7Ebelanche/Docencia/mineria/mineria.html (дата обращения 23.03.2018)

[34] [Электронный ресурс] URL: https://www.cbr.ru/ckki/?PrtId=restr (дата обращения 23.03.2018)

[35] [Электронный ресурс] URL: https://habr.com/post/264915/ (дата обращения 27.03.2018)

[36] [Электронный ресурс] URL: http://www.fico.com/en/latest-thinking/white-papers/introduction-to-model-builder-scorecard (дата обращения 03.04.2018)

[37] Арис Е.Т. // Модели оценки кредитных рисков. Проблемы анализа риска, выпуск 4, с.68-75, 2017.

[38] John C. Hull // Options, Futures, and Other Derivatives (9th Edition). Pearson, 2017.

Приложение А

В Табл. 10 представлены итоги категоризации всех переменных, коэффициенты моделей и скоринговый балл каждого параметра.

Таблица 10: Результат обучения и построения скоринг-карты

	Категория	Коэффициент LogReg без RFE	Коэффициент LogReg с RFE	Скоринг балл
Признак	1	—	—	—
	2	—	—	—
Опыт работы	0	0.74749946	—	13
	1	0.46276693	—	29
	2	-0.03638661	—	58
	3	-0.21546164	—	69
	4	-0.70098443	-0.94744964	97
Тип жилья	0	-0.26145858	—	71
	1	-0.05155725	—	59
	2	-0.11196951	-0.57721191	63
	3	0.41785872	—	32
	4	0.48075663	—	28
	5	-0.50397805	-0.76564317	85
	6	0.28778174	—	40
Срок кредита	0	-0.21627265	-0.91009993	69
	1	-0.75270405	-1.08760304	100
	2	0.12589047	—	49
	3	0.53481578	—	25
	4	0.56570416	—	20
Возраст	0	-0.08639894	—	61
	1	-0.18631345	—	67
	2	0.17588148	—	46
	3	-0.03487524	—	58
	4	0.38913986	—	34

Таблица 10: Результат обучения и построения скоринг-карты

	Категория	Коэффициент LogReg без RFE	Коэффициент LogReg с RFE	Скоринг балл
Семейное положение	0	-0.36256266	—	77
	1	-0.01205658	—	57
	2	-0.34505418	—	76
	3	0.13391696	—	48
	4	0.77033409	0.98890528	12
	5	0.07285608	—	52
Судимости	1	-0.68943725	-0.78297126	96
	2	0.94687096	0.94162131	2
Тип работы	0	0.0136451	—	56
	1	-0.45445408	—	82
	2	0.89311776	1.42852407	5
	3	-0.13254117	—	64
	4	-0.04868881	—	59
Расходы	0	-0.36500992	—	77
	1	-0.15286834	—	65
	2	0.77531198	0.82414935	11
Доходы	0	0.97099937	0.94093045	0
	1	0.94506789	0.88420722	2
	2	0.36329011	—	35
	3	0.14097052	—	40
	4	0.15927639	—	48
	5	-0.50781994	-0.68614675	86
	6	-0.69870774	-0.89980788	97
	7	-1.11564287	-1.1603125	121

Таблица 10: Результат обучения и построения скоринг-карты

	Категория	Коэффициент LogReg без RFE	Коэффициент LogReg с RFE	Скоринг балл
Активы	0	0.2689551	—	41
	1	0.78498635	—	11
	2	0.71155318	—	15
	3	0.18839209	—	45
	4	-0.31138697	—	74
	5	-0.59065999	—	90
	6	-0.79440606	-0.56843276	102
Обязательства	0	0.88761478	1.3029031	5
	1	-0.6209723	—	92
	2	-0.16678264	—	66
	3	0.15757386	—	47
Запрашиваемая сумма	0	-0.59103131	—	90
	1	0.02393463	—	55
	2	0.82453039	1.13485859	9
Цена товара	—	—	—	—

Приложение В

Блок программного кода, который реализует обучение модели, представлен ниже.

```
1       from sklearn import datasets
2       from sklearn.feature_selection import RFE
3       from sklearn.metrics import roc_auc_score
4       from sklearn.grid_search import GridSearchCV
5       from sklearn.cross_validation import train_test_split
6       from sklearn import metrics
7       from sklearn import preprocessing
8       from sklearn.linear_model import LogisticRegression
9       import numpy as np
10      import pandas as pd
11
12      data = pd.read_csv(data.csv,header=0,sep=;)
13
14      #вырезан фрагмент категоризации и других действий над
15      #данными. Далее выделены финальные выборки
16
17      X=data_final[cols]
18      y=data_final[Flag]
19      X_train, X_test, y_train, y_test = train_test_split(X, y,
20                                              test_size=0.3)
21      t = np.logspace(-5, 0, 100)
22      params = {"C": t}
23      grid_searcher = GridSearchCV(linear_model.
24                      LogisticRegreson(), params, cv = 5,
25                      scoring = "roc_auc")
26      grid_searcher.fit(X_train, y_train)
27
28      clf = grid_searcher.best_estimator_.fit(X_train, y_train)
29      results = clf.predict_proba(X_test)[:1]
```

Приложение С

В Приложении С представлены скриншоты интерфейса разработанного программного комплекса.

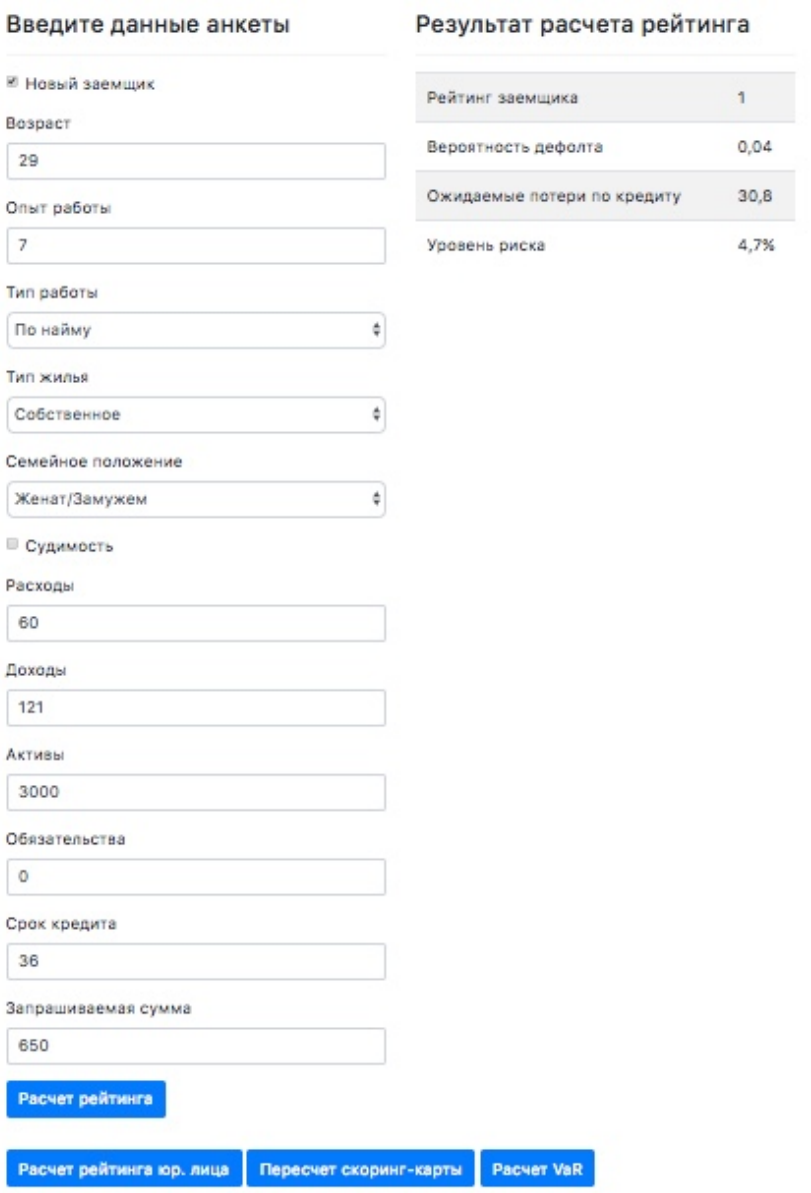

Рис. 9: Страница с анкетой потенциального клиента

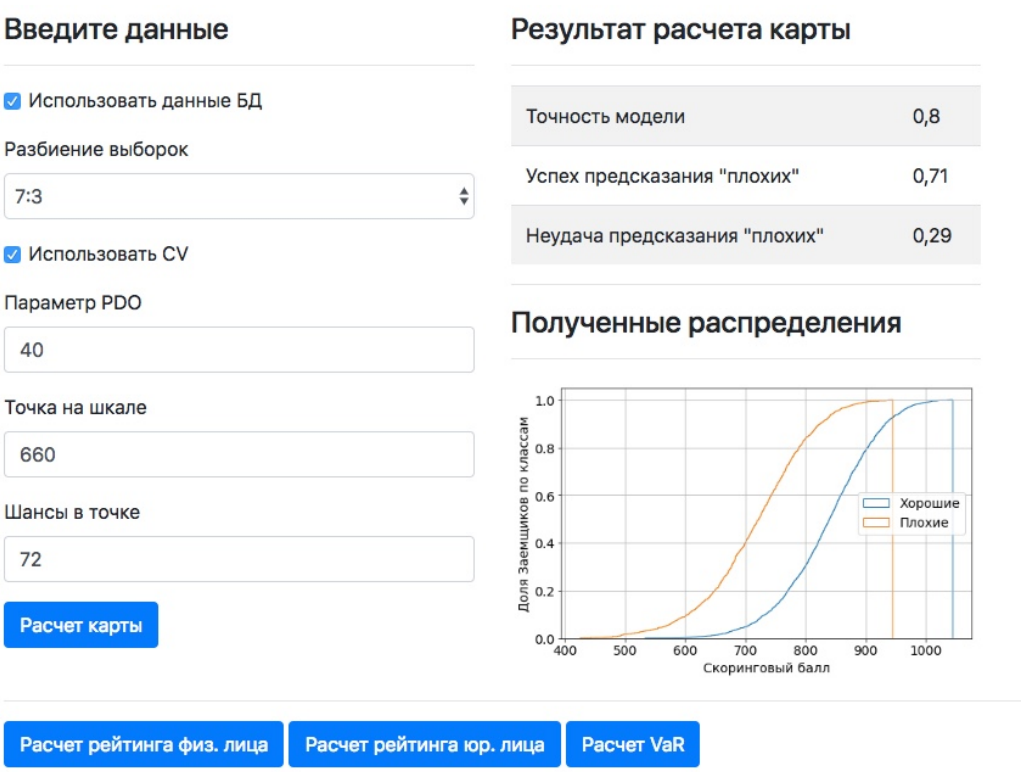

Рис. 10: Страница пересчета скоринговой карты

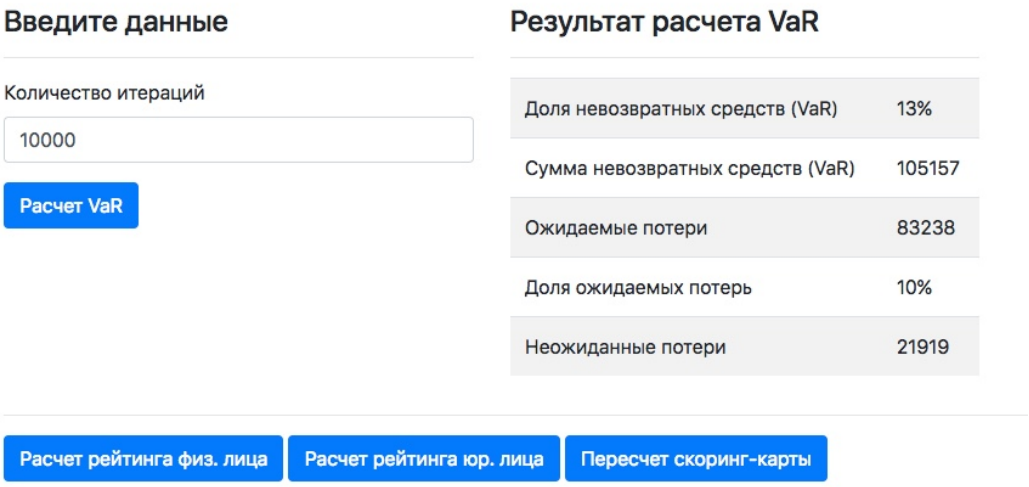

Рис. 11: Страница расчета показателя VaR