МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Г.В. ПЛЕХАНОВА»

Институт Цифровой экономики и информационных технологий
Базовая кафедра цифровой экономики института развития информационного общества

«Допустить к защите»
Заведующий кафедрой
Управления информационными
системами и программирования
д.э.н., профессор Уринцов А.И.
«___»_____ 2020 г.

Выпускная квалификационная работа

Направление 09.04.01 «Информатика и вычислительная техника»

магистерская программа «Управление ИТ-инфраструктурой цифровой экономики»

ТЕМА «Методика проектирования ИТ-инфраструктуры системы выявления мошенничества с банковскими кредитами среди юридических лиц»

Выполнил студент Толстяков Никита Юрьевич
Группа 291М-08/18м

Научный руководитель выпускной
квалификационной работы
к.э.н., доцент
Мамедова Наталья Александровна

_____
(Подпись)
Автор _____
(Подпись)

Москва - 2020

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....6

ГЛАВА 1. АНАЛИЗ МЕТОДОВ ВЫЯВЛЕНИЯ МОШЕННИЧЕСТВА С БАНКОВСКИМИ КРЕДИТАМИ СРЕДИ ЮРИДИЧЕСКИХ ЛИЦ.....11

1.1 Анализ проблемы мошенничества с банковскими кредитами среди юридических лиц.....11

1.1.1 Основные понятия и статистика.....11

1.1.2 Проблема мошенничества в сфере кредитования.....13

1.1.3 Проблема незаконного получения кредита.....18

1.1.4 Проблема преднамеренного банкротства.....19

1.2 Общие рекомендации по выявления мошенников в сфере банковского кредитования среди юридических лиц.....24

1.3 Финансовые методы по выявления мошенников в сфере банковского кредитования среди юридических лиц.....27

1.4 Вероятность дефолта и уровень потерь при дефолте как методы по выявления мошенников в сфере банковского кредитования среди юридических лиц 34

1.5 Использование алгоритмов машинного обучения для выявления мошенников в сфере банковского кредитования среди юридических лиц.....41

1.6 Постановка задачи на разработку методики создания ИТ-инфраструктуры для выявления мошенничества с банковскими кредитами среди юридических лиц 53

ГЛАВА 2. РАЗРАБОТКА МЕТОДИКИ ПРОЕКТИРОВАНИЯ ИТ-ИНФРАСТРУКТУРЫ СИСТЕМЫ ВЫЯВЛЕНИЯ МОШЕННИЧЕСТВА С БАНКОВСКИМИ КРЕДИТАМИ СРЕДИ ЮРЕДИЧЕСКИХ ЛИЦ.....56

2.1 Выбор источника данных для нахождения мошенников среди юридических лиц 56
2.2 Процесс построения алгоритмов для нахождения мошенников среди юридических лиц 59
2.2.1 Алгоритм выявления мошенничества на основе назначения платежа в транзакции 63
2.2.2 Алгоритм выявления мошенничества на основе отчетности РСБУ 79
2.2.3 Алгоритм графового анализа связи клиента для выявления мошенничества 89
2.2.4 Общие выводы 96
2.3 Процесс построения архитектуры системы выполнения алгоритмов для нахождения мошенников среди юридических лиц 97
ГЛАВА 3. ОЦЕНКА ТЕХНИКО-ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ РАЗРАБОТАННОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ПРЕДЛОЖЕННОЙ МЕТОДИКИ 115
3.1 Разработка пользовательской части системы поиска мошенников с банковскими кредитами среди юридических лиц 115
3.2 Оценка качественных показателей разработанного программного средства 120
3.3 Оценка экономической эффективности 125
ЗАКЛЮЧЕНИЕ 135
СПИСОК ЛИТЕРАТУРЫ 139

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

АС – автоматизированная система

БД – база данных

ВКР – выпускная квалификационная работа

ГСЗ – группа связанных заемщиков

ГОСТ – межгосударственный стандарт

ЕГРИП – единый государственный реестр индивидуальных предпринимателей

ЕГРН – единый государственный реестр недвижимости

ЕГРЮЛ – единый государственный реестр юридических лиц

ЕФРСБ – единый федеральный реестр сведений о банкротстве

ИИ – искусственный интелект

ИНН – идентификационный номер налогоплательщика

ИП – индивидуальный предприниматель

ИТ – информационные технологии

МСФО – международный стандарт финансовой отчетности

МЭК – международная техническая комиссия

НБКИ – национальное бюро кредитной истории

НКРЯ – национальный корпус русского языка

ПО – программное обеспечение

ПС – программное средство

РСБУ – российский стандарт бухгалтерской отчетности

РФ – Российская Федерация

СУБД – система управления базами данных

УК – уголовный кодекс

УПА – управление проблемными активами

ФНС – федеральная налоговая служба

ЭВМ – электронно-вычислительная машина

ЮЛ – юридическое лицо

AFS (anti-fraud system) – система против мошенничества

AI (artificial intelligence) – искусственный интеллект

API (application programming interface) – интерфейс прикладного программирования

EAD (exposure at default) - требования под риском дефолта

EL (expected loss) – ожидаемые потери

ENN (Edited Nearest Neighbor) – редактирование методом ближайшего соседа

GRU (gated recurrent units) – управляемый рекуррентный блок

HDFS (Hadoop distributed filesystem) – распределённая файловая система Hadoop

IDF (inverse document frequency) – обратная частота документа

IT (information technology) – информационные технологии

IV (information value) – ценность информации

kNN (k nearest neighbors) – k ближайших соседей

LGD (loss given in default) - доля потерь в случае дефолта

LSTM (Long short-term memory) - Долгая краткосрочная память

PD – (probability of default) – вероятность дефолта

RBF (Radial basis function) – функция радиального базиса

RELU (Rectified linear unit) – полулинейный элемент

RNN (Recurrent neural network) - Рекуррентные нейронные сети

SMOTE (Synthetic Minority Oversampling Technique) - синтетическая техника передискретизации

SVM (support vector machine) - Метод опорных векторов

TF (term frequency) - частота слова

UL (unexpected loss) – неожиданные потери

WoE (weight of evidence) – вес значимости

ВВЕДЕНИЕ

В 2018 году компания ООО «Уралэлектрострой» подала заявление в суд о признание себя банкротом и данное заявление было удовлетворено [46]. В рамках процедуры банкротства было введено внешнее управление. Общий требования только ПАО «Сбербанк» к данной компании составляют 6,2 миллиардов рублей, при этом залоговое имущество оценивается в 1,1 миллиардов рублей. 11 марта 2019 года арбитражный управляющий подал заключение о нахождение признаков преднамеренного банкротства [45]. Более того, компания ПАО «Сбербанк» выдала ООО «Уралэлектрострой» новый кредит в размере 200 миллионов рублей за несколько месяцев до подачи заявления о банкротстве. Самые ранние сделки, оспоренные внешним управляющим, датируются 2016 годом [20].

Таким образом, если бы ПАО «Сбербанк» зафиксировал подозрительную активность в 2018 году, то были бы сохранены 200 миллионов рублей. В случае фиксирования таких действий в 2016 году, банк смог бы сохранить всю сумму кредита подав заявление о банкротстве на начальном этапе совершения незаконных действий должника [50].

Только за 2019 год преднамеренное банкротство было выявлено в 5,5% всех закрытых дел о банкротстве, что по оценкам составляет 181 миллиард рублей требований кредиторов [51].

Помимо преднамеренного банкротства большой ущерб банкам наносят мошенничество с кредитами и незаконное получение кредита, нижняя граница ущерба от которых в 2017 году оценивается в 1076 миллионов рублей.

При этом, все приведенные правонарушения фиксируются только после выдачи кредита банком. В следствие чего банк теряет большую часть этих денег без возможности их вернуть.

В данной работе под термином «мошенничество» будет пониматься совокупность трех видов преступлений: мошенничество с кредитами, незаконное получение кредита, преднамеренное банкротство.

Как видно из приведенной статистики банки не имеют специализированных автоматизированных инструментов для поиска мошенничества или эти инструменты являются недостаточно эффективными. При этом, ключевыми точками во времени для максимально эффективной фиксации мошенничества являются: момент принятия решения о выдачи кредита и момент принятия решения о стратегии урегулирования проблемной задолженности.

При этом, может показаться, что момент принятия решения о выдачи кредита покрыт высокоэффективными инструментами во всех современных банках, которые поддерживают условия Базель II. Такими алгоритмами является являются модели LGD и PD. Более того, одним из самых популярных способов применения искусственного интеллекта и машинного обучения является задача кредитного скоринга. Но как видно из приведенных данных этих методов недостаточно.

Одну из ключевых проблем в решение данной задачи отлично сформулировал автор статьи на сайте Habr.com: «Первое правило антифрода — никому не рассказывать про антифрод» [38]. На сегодняшний день крайне мало научных работ, в которых предложена методика и алгоритмы выявление мошенников среди юридических лиц с банковскими кредитами [23].

Таким образом в данной работе будет представлена методика на основе которой любой банк сможет построить систему поиска мошенников с банковскими кредитами среди юридических лиц.

Актуальность работы как было показано ранее состоит в необходимости создания данного инструмента для банковской сферы. Были показаны большие потери, которые несут банки. Более того, как следствие таких ущербов банки будут становится все более осторожными, их аппетит к риску будет уменьшаться. Что в свою очередь приведет к уменьшению количества выданных кредитов и реструктуризации добросовестным клиентам, которые способны выполнить обязательства.

Цель и задачи исследования.

Целью исследование является разработка методики построения ИТ-инфраструктуры системы поиска мошенников с банковскими кредитами среди юридических лиц. Предлагаемая методика позволяет повысить качество проверки заемщика и обоснованность принятия решения по вопросам кредитования, таким образом уменьшает риск выдачи мошеннику кредита или выдачи реструктуризации.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести подробный анализ методов, которые применяют мошенники при совершении правонарушений.
2. Проанализировать существующие методы, которые могут использоваться для выявления мошенников.
3. Изучить общие подходы к построению взаимодействия различных подразделений банка и подходы к построению взаимодействия подразделений в рамках поиска мошенников.
4. Провести анализ видов данных для решения данной задачи и ранжировать их по эффективности.
5. Разработать методику построения алгоритмов для поиска мошенников.
6. Разработать методику построения архитектуры системы применения алгоритмов для поиска мошенников.
7. Проанализировать технические показатели системы, разработанной по предложенной методики.
8. Проанализировать экономическую эффективность разработанной системы на основе предложенной методики.

Объектом исследования является вопросы обеспечения безопасности банка в сфере кредитования юридических лиц.

Предметом исследования является методы выявления мошенников с банковскими кредитами среди юридических лиц.

Новизна работы заключается в разработке уникальных алгоритмов по выявлению мошенников и архитектуры системы для применения алгоритмов, что позволит снизить потери банка от мошенничества.

Теоретической основной исследования стали:

1) Отечественные и зарубежные исследования по методам выявления мошенничества в различных сферах деятельности.
2) Отечественные и зарубежные исследования по современным методам машинного обучения, графам, временным рядам.
3) Публикации на сайтах о методах построения распределенных систем вычисления.

Теоретическая и практическая значимость работы заключается в том, что на основе полученных знаний:

1) Можно в короткий срок ознакомиться с необходимой теорией по выявлению мошенничества на основе, которой можно разработать собственную методику построения ИТ-инфраструктуры систем выявления мошенничества с банковскими кредитам среди юридических лиц, исходя из требований и возможностей компании.
2) На основе данного материала можно построить систему по выявлению мошенничества с банковскими кредитами среди юридических лиц, которая будет иметь высокие показатели качества и высокую эффективность.
3) собранный материал представляет собой уникальное методическое пособие по изучению методов мошенничества и методов выявления мошенничества в банковской сфере на территории Российской Федерации.

Основные положения, выносимые на защиту:

1) Обоснование целесообразности построения системы выявления мошенничества с банковскими кредитами среди юридических лиц.

2) Описание разработанной методики построения ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц.
3) Порядок применения описанной методики ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц на примере системы «АнтиФрод».
4) Оценка технических параметров применения методики ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц.
5) Оценка эффективности применения методики построения ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц.

ВКР состоит из введения, трех глав, заключения и списка литературы. Общий объем основного текста содержит 139 страниц, включая 29 рисунков и графиков и 12 таблиц.

ГЛАВА 1. АНАЛИЗ МЕТОДОВ ВЫЯВЛЕНИЯ МОШЕННИЧЕСТВА С БАНКОВСКИМИ КРЕДИТАМИ СРЕДИ ЮРИДИЧЕСКИХ ЛИЦ

1.1 Анализ проблемы мошенничества с банковскими кредитами среди юридических лиц

1.1.1 Основные понятия и статистика

Для выстраивания процесса и методологии борьбы с различными видами мошенничества с банковскими кредитами необходимо провести анализ самой проблемы. Это необходимо для понимания методов, которыми пользуются злоумышленники для построения наиболее эффективных систем, и для наиболее эффективного встраивания IT-процесса в процессы работы банка и правоохранительных органов. Данному анализу и будет посвящен данный параграф.

Как уже упоминалось ранее в данной работе под понятием мошенничества будут пониматься три статьи уголовного кодекса Российской Федерации:

- Мошенничество в сфере кредитования. Статья 159.1 УК РФ.
- Незаконное получение кредита. Статья 176 УК РФ.
- Преднамеренное банкротство. Статья 196 УК РФ.

Далее в работе под термином мошенничество будет пониматься совокупность всех трех факторов, а когда речь будет идти о статье 159.1 УК РФ будет использоваться термин «Мошенничество в сфере кредитования».

В данной работе не будет рассматриваться статья 197 УК РФ «Фиктивное банкротство», хотя ее включение выглядит логичным. Данное решение было принято по двум причинам:

1. За 2019 год из 36354 дел о банкротстве только в 57 были выявлены признаки фиктивного банкротства.

2. Фиктивное банкротство наносит меньший ущерб банку, нежели остальные пункты, так как юридическое лицо в данном случае не теряет своей возможности погашать долги.

Только в 2019 году общая сумма требований, включенных в реестр требований кредиторов составила 3301 миллиард рублей, из которых были удовлетворены требования только на 146 миллиардов рублей. Если взять для анализа только кредиторов первой очереди, чьи требования удовлетворяются в первую очередь, из 809 миллионов рублей были выплачены только 321 миллион рублей, что составляет 39%. При этом из 36354 дел о банкротстве признаки преднамеренного банкротства были выявлены в 2 023 случаях (5.5%), 5349 (14%) случаях недостаточно информации о вынесение заключения [51].

За 2017 год было вынесено 96 приговоров по статье 159.1 часть 4 УК РФ, которая характеризуется особо крупным размером, более 6000000 рублей, то есть минимальный нанесенный ущерб составляет 360000000 рублей. По остальным частям данной статьи было вынесено 2284 приговора за 2017 год. В таблице 1 представлены данные о количестве приговоров по части 4 статьи 159.1 УК РФ, а также минимальный ущерб, который могли нанести злоумышленники [12].

Таблица 1

Статистика приговоров по статье 159.1 часть 4 УК РФ

	2013	2014	2015	2016	2017
Количество приговоров	65	95	137	92	96
Минимальный нанесенный ущерб в млн. руб.	390	570	822	552	576

По статье 176 УК РФ за 2017 год было раскрыто 216 преступлений [55].

После анализа приведенной статистики становится очевидно, что основной ущерб от дынных преступлений несет банковская система.

На данный момент банки могут защитить себя от мошенничества в банковской сфере или от незаконного получения кредита в момент его выдачи, но не могут защитить себя от преднамеренного банкротства или от мошенничества, решение о котором недобросовестный предприниматель принял уже после получения кредита. Такие дела расследуются уже после нанесения ущерба банку и в большинстве случаев не могут быть возвращены, о чем говорит приведенная ранее статистика.

Как следствие страдает от данных незаконных действий и отсутствия у банков точных и качественных инструментов определения потенциального или состоявшегося, но еще не найденного правонарушителя, страдает предпринимательство. Из-за больших потерь от приведенных преступлений и при отсутствии уверенности в возможности ранней диагностики этих преступлений, банки, у которых низок аппетит к рискам будут выдавать меньше кредитов рискованным или изначально несколько подозрительным заемщикам.

1.1.2 Проблема мошенничества в сфере кредитования

В рамках данной работы в качестве потенциальных злоумышленников будут рассматриваться юридические лица, в том числе и индивидуальные предприниматели. Когда совершается мошенничество в сфере кредитования завладевает денежными средствами юридическое лицо (индивидуальный предприниматель), что позволяет заключать кредитные договора на достаточно крупные суммы и достаточно весомо затрудняет раскрытие подлинных намерений злоумышленников и привлечение их к уголовной ответственности. Это связанно с тем, что такой подход позволяет злоумышленнику выдать свои действия за обычную деятельность хозяйствующего субъекта [6].

Обязательным признаком для данного вида преступления является предоставление банку заведомо ложных и недостоверных сведений. Стоит отметить, что с точки зрения закона только умышленное предоставление таких данных является преступлением, но с точки зрения банка нет разницы в мотивах и осознанности действий лица, предоставившего такие документы, имеет значение только будет ли нанесен ущерб или нет.

Относиться предоставленные ложные и не достоверные сведенья могут к самым разным обстоятельствам и могут характеризовать финансовое состояние заемщика, качество и ликвидность предлагаемого заемщиком обеспечения, его кредито- и платежеспособность. Например, в том случае, если заемщик является руководителем организации или индивидуальным предпринимателем, то к таким сведеньям могут относиться две большие группы сведенья о хозяйственном положение и сведения о финансовом состоянии [12].

Пример сведений о хозяйственном положении

1) Сведения о хозяйственном положении: недостоверные данные об руководстве, учредителях, акционерах, партнерах предприятия, связях.
2) Фиктивные гарантийные письма, поручительства, предоставление в залог имущества, на которое нельзя обратить взыскание, не находящегося в собственности, с неверно указанной стоимостью.
3) Технико-экономическое обоснование, в котором неверно указаны основные направления использования кредита.
4) Поддельные договоры и другие документы, свидетельствующие о фиктивной конкурентоспособности заемщика, его положение на рынке и т. п.
5) Поддельные договоры, платежные, транспортные и иные документы, касающиеся хозяйственной операции, на которую испрашивается кредит.
6) Искаженные данные складского и бухгалтерского учета.

Примеры ложных сведений о финансовом состоянии:

1) Сфальсифицированные бухгалтерские документы о регистрации в налоговой инспекции, в которых финансовое состояние представлено лучше, чем это имеет место в действительности.
2) Справки о дебиторской и кредиторской задолженности, о полученных кредитах и займах в других банках.
3) Выписки из расчетных и текущих счетов.

Крайне часто как способ фальсификации используется создание фиктивных юридических лиц. В данном случае, фиктивные компании создаются как в роли заемщика, так и в роли контрагента заемщика. Случаи, когда фиктивные юридические лица сами являются заемщиками в настоящий момент идут на спад, что связанно с простотой проверки времени создания компании и ее владельца. С другой стороны, развивается метод создания фиктивных компаний контрагентов заемщика, они создаются с целью увеличения оборотов компании в бухгалтерской отчетности, хотя на самом деле бенефициаром всех сделок с данными контрагентами является мошенник. Такая компания не получает реальной прибыли от сделок. В случае, когда для мошенничества используются фиктивные компании контрагенты, возникает проблема их выявления «ручным» способом, то есть работником банка при помощи анализа открытых источников информации. Трудности связаны с необходимостью ручного анализа большого количества информации. На рисунке 1 приведено соотношение созданных и ликвидированных юридических лиц в разные года, в количестве созданных юридических не учитывались те, которые были ликвидированы в году создания [13].

Большая часть преступлений по статье 159.1 УК РФ совершается с участием сотрудника банка. На одном из этапов проверки потенциального заемщика появляется возможность разоблачить мошенника, но злоумышленник решает эту проблему находя сообщника из службы безопасности, кредитного инспектора или аналитика и других сотрудников банка, которые участвуют в выдачи кредита. На рисунке 2 приведена схема данного вида мошенничества.

Рисунок 1. Сравнение количества созданных и ликвидированных юридических лиц в РФ (Автор. MS Excel)

В независимости от конкретной фальсификации на основание ложных документов и недостоверных сведений кредитор (банк) делает не верное заключение о возможности возврата полученного кредита. Но если бы заемщик предоставил кредитору подлинные сведенья, то он мог получить кредит меньшего размера или не получить его совсем, кредитор мог бы применить средства более активного мониторинга клиента или применить другие средства минимизации риска.

Рисунок 2. Схема мошенничества путем получения кредита с участием сотрудников банка [12]

В большинстве работ рассматривается только один критичный момент времени, в который можно предотвратить преступление – это момент выдачи кредита. Но рассмотрим ситуацию, при которой заемщик при получении кредита не нарушал закон и действительно намеревался вернуть деньги [54]. Но со временем у юридического лица начались проблемы, и заемщик начал выходить на просрочку и скоро должен выйти в дефолт. В такой ситуации кредитор стоит перед выбором: применять ли ему «кредитную стратеги», например, реструктуризацию или применять «дефолтную стратегию», например, начинать процедуру банкротства заемщика. В этот момент заемщик может совершить мошенничество и передать банку не верные сведенья в попытке склонить банк к принятию «кредитной» стратегии.

Мотивация правонарушителя в описанной выше ситуации может быть двух видов:

1) Заемщик может надеется на восстановление своей платежеспособности, но не может доказать банку свою способность к этому с помощью реальных данных.
2) Заемщик, осознавая, что не сможет предотвратить банкротство, возможно полную ликвидацию юридического лица, склоняет кредитора к принятию решения о применении «кредитной стратегии» для получения времени, в которое он сможет незаконно использовать средства компании в личных целях.

Кроме финансовых убытков мошенничество в сфере кредитования может привести к репутационным потерям, например, при совершении цессии на основе ложной информации предоставленной заемщиком, была согласована цены сильно выше реальной.

1.1.3 Проблема незаконного получения кредита

Данная проблема крайне похожа на проблему мошенничества с кредитами и имеет много общего по объективному составу преступления. Как и в случае с мошенничеством, незаконное получение кредита предполагает предоставление банку заведомо ложных сведений с целью получения кредита или льготных условий кредита. Но в отличии от мошенничества, данная статья предполагает, что правонарушитель собирается вернуть кредит и выполнить все прописанные в контракте условия. Под льготными условиями кредитования обычно понимается более выгодная по сравнению с обычной сделкой получения или возврата кредита [31].

Во второй части рассматриваемой статьи Уголовного кодекса предметом преступления является нецелевое использование государственного целевого кредита, в случае если это причинило крупный ущерб гражданам, государству или организациям.

Нецелевое использование кредита – это умышленная реализация полученных денежных средств с нарушением условий, которые были прописаны в нормативных документах о предоставлении государственного целевого кредита и в кредитном договоре [55].

Приведенная в данной работе статистика данного правонарушения может выглядеть не значительной по сравнению с преднамеренным банкротством, но как утверждает В. И. Гладких, данный вид преступлений является высоко латентным. Как указанно в данной работе количество преступлений данного вида выше регистрируемого числа в 300-400 раз. Банки достаточно часто не заявляют о совершении преступления по статье 176 УК РФ, так как это наносит репутационный вред банку, у которого могут быть проблемы устойчивости вследствие подобных преступлений.

В данном случае способы незаконного получения кредита или льготных условий кредитования совпадает с методами, которые применяют при мошенни-

честве, такие как предоставление фиктивной информации о финансовом состоянии компании и ложные данные о хозяйственном положение. Также крайне активно используются и методы создания фиктивных юридических лиц. И также совпадают два ключевых момента времени, в которые банк имеет возможность предотвратить преступление или серьезно уменьшить свои денежные потери.

1.1.4 Проблема преднамеренного банкротства

Большую угрозу экономической безопасности государства представляют собой преднамеренные банкротства. В Стратегии экономической безопасности Российской Федерации на период до 2030 г., которая была утверждена Президентом Российской Федерации от 13.05.2017 г. № 208, прямо говориться, что одной из основных задач по реализации этого направления является предотвращение преднамеренного банкротства.

Сам институт несостоятельности (банкротства) юридического лица включили в российский право Федеральный закон от 26.19.2002 № 123-ФЗ «О несостоятельности (банкротстве)». В законодательстве есть существенная юридическая разница между терминами: «должник», «несостоятельность», «банкротство» [13]. На рисунке 3 приведена схема классификации различных видов должников.

Для того, чтобы разобраться в принципах и методах совершения рассматриваемого вида преступление необходимо понять, что такое банкротство. Основное, что необходимо знать о банкротстве – это критерии банкротства:

1) Невозможность удовлетворить требования кредиторов, как по выплате заработной платы и выходных пособий, так и уплате обязательных платежей. Данный пункт возможен только в случае, если соответствующие обязательства не исполнены в течении девяноста дней с той даты, в которую необходимо было удовлетворить требования.

2) Совокупные требования к должнику должны составлять не менее 300000 рублей, в случае юридического лица и 500000 рублей, если рассматриваемое лицо является индивидуальным предпринимателем или физическим лицом.

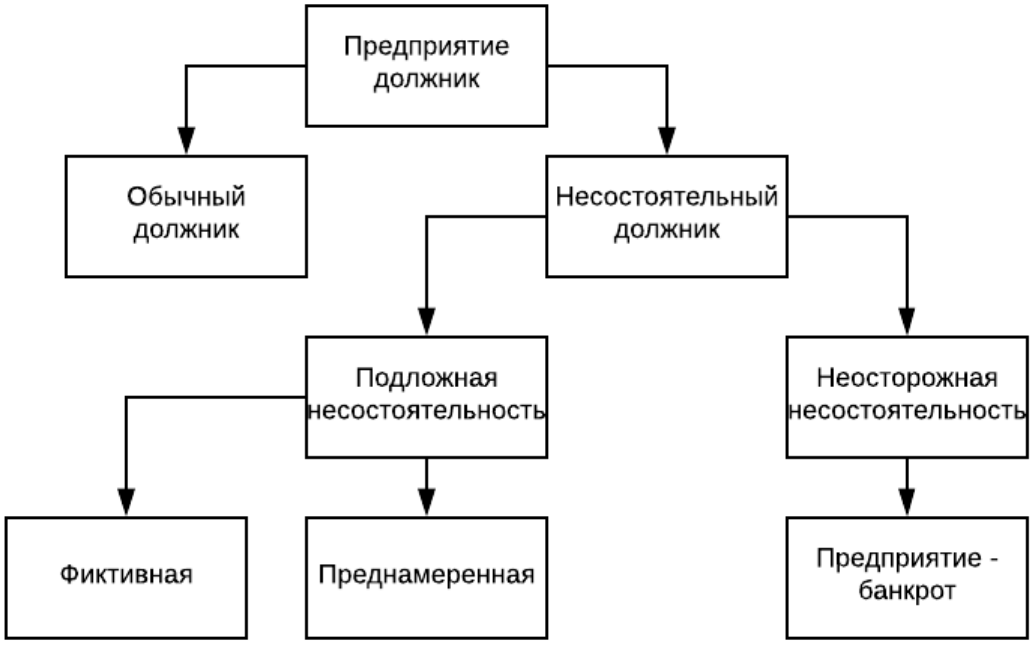

Рисунок 3. Виды несостоятельности предприятий (Автор, creately.com)

Само преднамеренное банкротство – это умышленные действия, которые привели к увеличению или создали неплатежеспособность, или другими словами, это намеренное введение лица в состояние банкротства. Юридические лица различно относятся к банкротству: некоторые считают, что банкротство приводит к разрушению социально-экономической системы, другие, считают банкротство путем к новым возможностям и развитию компании. Как пример можно привести президента Соединенных Штатов Америки Дональда Трампа, компании которого суммарно были банкротами 6 раз [22].

Как можно судить, исходя из приведенного определения преднамеренного банкротства и обязательных условий банкротства, преступными в данном случае являются действия по увеличению кредитной задолженности, заведомо зная о

невозможности ее выплатить и действия по уменьшению своей платежеспособности, например, очевидно невыгодные сделки, как продажа своей продукции сильно ниже рыночной цены.

Чаще всего руководители организаций и индивидуальные предприниматели используют инструмент преднамеренного банкротства для получения личной выгоды. Так преднамеренному банкротству может предшествовать продажа всего большей части имущества компании и вывод этих денег со счетов компании на счета бенефициара. Или продажа имущества (продукции) по ценам сильно ниже рыночных предприятиям, где владелец компании правонарушителя является также владельцем или бенефициаром [41].

В качестве методов, которые приведут к преднамеренному банкротству могут использоваться не только продажи по цене ниже рыночной, но и покупка мало ликвидных активов по завышенным ценам, которые также могут покупаться у связанных компаний для получения выгоды управляющими лицами.

Вопрос о наличии признаков преднамеренного банкротства решается арбитражным управляющим на стадии наблюдения. В этот период арбитражный управляющий должен проверить:

1) Соответствие всех решений лицами, которые управляют организацией действующему законодательству.
2) Факты исполнения обязательств, не соответствующие рыночным условиям.
3) Сделки, результатом которых стали признаки неплатежеспособности.
4) Сделки с высоколиквидными активами, результатом, которых был ущерб организации.
5) Сделки, которые были заключены на заранее невыгодных условиях.

Текущие законодательство дает возможность только арбитражному управляющему открывать дело о преднамеренном банкротстве, причем только после начала самого банкротства, когда уже начата процедура банкротства и банк уже понес ущерб.

Основным риском для банка является не только само банкротство, которое приведет к потере денег, о чем можно судить из приведенной ранее статистики, но и потери из-за незаконных действий при создании преднамеренного банкротства: продажи имущества, вывод денег со счетов предприятия и так далее.

Для дальнейшего понимания процесса применения инструментов для выявления преднамеренного банкротства банком необходимо понять время совершения преступления относительно времени взаимоотношения кредитора и заемщика по конкретному кредитному договору.

Если заемщик берет кредит и планирует его отдавать, результатом чего будет процедура банкротства кредитора, то данное деяние, скорее всего, будет расценено как мошенничество с кредитом. Преднамеренное банкротство чаще всего возникает после того как заемщик совершил некоторое количество платежей по кредиту. В обоих случаях, если заемщику были выданы деньги и банком были обнаружены признаки преднамеренного банкротства, то необходимо максимально быстро подать заявление о начале процедуры банкротства. Но возможность подачи заявления о банкротстве ограничена по времени. Следовательно, как и в случае с мошенничеством и незаконным получением кредита, преднамеренное банкротство необходимо «ловить» на стадии выдачи кредита и стадии принятия решения о дальнейшем взаимодействие с проблемным задолжником.

Проблема преднамеренного банкротства тесно связана с проблемой мошенничества в кредитовании. Так заемщик, который решил получить выгоду за счет преднамеренного банкротства могут предоставлять в банк не действительные данные о своем финансовом состоянии, о состоянии заложенного имущества и так далее. Делается это как с целью «отвести взгляд» кредитора от подготовки себя к банкротству и получения выгоды от этой процедуры, так и для получения дополнительного времени на продажу имущества или вывод денег со счетов юридического лица на счета физических лиц, злоумышленников, за счет получения реструктуризации [45].

Как один из самых распространённых методов мошенничества в сфере кредитования и преднамеренного банкротства одновременно – это совершение скрытой продажи. Данный метод заключается в том, что заемщик убеждает кредитора дать максимально возможную оценку имущества, затем заемщик выходит в дефолт, а банк продает заложенное имущество гораздо дешевле оценочной стоимости. Это связано с тем, что ликвидационная стоимость меньше рыночной более чем на 30%. Данная схема представлена на рисунке 4.

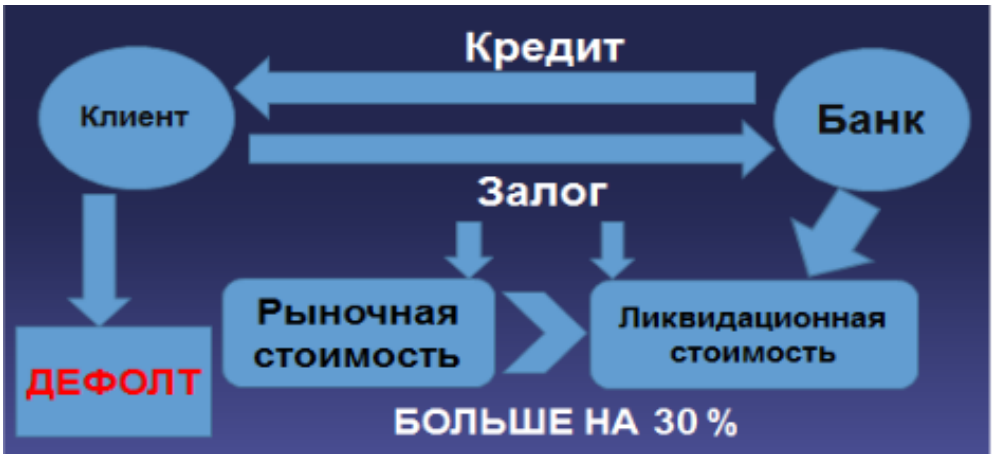

Рисунок 4. Схема мошенничества путем получения кредита с целью совершения скрытой сделки [12]

Данная схема может использоваться, как и в момент получения кредита, так и в момент принятия решения по проблемному клиенту, в качестве аргумента для выдачи реструктуризации заемщик может предложить банку дополнительный залог по схеме мошенничества «совершения скрытой продажи». Предложив дополнительный залог, который на самом деле не будет ликвидным заемщик может добиться не только реструктуризации, но и выдачи дополнительного финансирования.

1.2 Общие рекомендации по выявления мошенников в сфере банковского кредитования среди юридических лиц

Ранее в данной работе было выявлено два основных момента времени, когда необходимо выявить мошенничество – это момент принятия решения о выдачи кредита и момент принятия решения о дальнейшем взаимодействии с проблемным должником. Приведенные далее методы могут использоваться, как и на моменте выдачи кредита так и в процессе мониторинга состояния клиента, и в процессе принятия решения о дальнейшем взаимодействии с проблемным клиентом.

Основой любой методики по выявлению мошенников является построение слаженного процесса и контроль за безопасностью самого банка. Для предотвращения мошенничества различными путями необходимо соблюдать следующие правила.

1) Децентрализация функций и многоуровневый контроль со стороны банка за процессом выдачи кредита и мониторингом.
2) Максимизировать выдачу кредитов под залог в виде недвижимости.
3) Эффективно брать поручительство владельца бизнеса.

Большую часть работы по предупреждению мошенничества осуществляет кредитный отдел. Одним из важнейших этапов работы этого отдела является проверка потенциального заемщика перед заключением контракта. После заключения договора данные функции переходят к отделу мониторинга клиентов. Данная проверка заключается в:

1) Выяснение платежеспособности клиента.
2) Оценка финансового положения.
3) Качество управления компанией.
4) Состояние отрасли и региона.
5) Конкурентоспособность компании.
6) Положение заемщика в отрасли или сфере деятельности.
7) Наличие активов и пассивов, товарных запасов.

При мошенничестве предоставленные банку документы могут быть подложными, для проверки целесообразно выехать на предприятие для сравнения фактических бухгалтерских данных с представленными банку и сверить полученные документы с документами, которые были предоставлены заемщиком в налоговую инспекцию.

При взятии имущества в залог необходимо провести независимую оценку имущества или оценку имущества экспертами банка. Минимально необходимо получить свидетельство о наличие этого имущества и подтверждения его состояния.

Идеальная система противодействия мошенническим действиям, которая применяется при рассмотрении заявки и мониторинга клиента должна учитывать максимально доступный объем информации. Такая система должна предусматривать построение различных связей на основе гигантского количества данных о клиенте, при этом работать быстро и не приводить к потере качества обслуживания.

Необходимо организовать строго регламентируемый процесс по выявлению мошеннических действий со стороны клиента. В рамках этого процесса должны быть разработаны различные критерии мошенничества, и строго формализованы действия сотрудника если какие-то из критериев сработали. Также необходимо наладить совместную работу сотрудников отдела безопасности банка и правоохранительные органы.

Одним из ключевых методов проверки кредитной заявки является система андеррайтинга. Она может быть по-разному внедрена в процесс работы банка, но наиболее эффективным является внедрение заключение андеррайтера как финального согласования выдачи кредита на основание самой заявки, данным из внешних источников и результатам проверки клиентов.

Каждому банку рекомендуется использовать «Национальное бюро кредитных историй» (НБКИ) для поиска кредитов, оформленных на потенциального заемщика и его кредитного рейтинга. Также рекомендуется использовать систему

«НБКИ-AFS» (Национальное бюро кредитных историй – Anti-Fraud Service). Данная система предоставляет услугу как сервис и позволяет не только прямые, но и опосредованные связи с мошенническими заявками. По заявлению Александра Викулина, сервис «НБКИ-AFS» является серьезным подспорьем в борьбе с кредитным мошенничеством для любого типа кредитов [27].

Данный сервис построен на системе правил и использует в своей основе базу знаний, он использует более 200 правил, которые были созданы при помощи различных кредитных организаций, банков и МФО. Данный сервис работает с высокой скоростью и обладает хорошей устойчивостью, он способен обрабатывать более 200 заявок в секунду без потери качества обслуживания.

Крайне важным в выявление мошенников является процесс обработки заявки и честность сотрудников, без этого любые другие инструменты станут бесполезны. Так же необходимо максимизировать количество документов, получаемых из альтернативных источников и возможность проверить документы, которые предоставил клиент, ведь именно на их основе строится вся методология принятия решения о выдачи кредита.

Большой проблемой в предотвращение мошенничества является затрудненная передача данных между различными кредитными организациями, частично это связано с Федеральный закон «О персональных данных» от 27.07.2006 №152-ФЗ. Обмен информацией может быть расценен как преступление, а внесение обмена информацией между организациями напрямую, минуя НБКИ, в согласия на обработку персональных данных может оттолкнуть клиентов. Вторым фактором в проблеме передачи данных является высокая конкуренция в отрасли.

Еще одним эффективным методом является построение группы связанных заемщиков (ГСЗ). Под такой группой понимают несколько хозяйствующих субъектов, экономическое состояние которых связанно между собой. Так при дефолте или банкротстве одного из членов группы резкое ухудшение финансового

состояния всей группы ухудшится. Формирование ГСЗ необходимо для выполнения норматива Н6 Банка России. Так, если один из участников ГСЗ совершит попытку методом предоставления ложных документов создать видимость хорошего финансового состояния, его реальное состояние можно выяснить благодаря анализу добросовестных членов ГСЗ.

1.3 Финансовые методы по выявления мошенников в сфере банковского кредитования среди юридических лиц

Одной из центральных групп методов являются финансовые методы. Они позволяют на основе документов спрогнозировать состояние заемщика на несколько месяцев вперед и своевременно среагировать на ухудшающиеся состояние или принять правильное решение о выдачи кредита. Данные методы должны учитывать факторы из разных источников и быть проверенными. Иначе результаты предсказания могут быть не верными. Наилучшим вариантом будет основывать расчеты на документах, которые были собраны из других организаций, таких как НБКИ, ФНС, ЕГРЮЛ, Спарк-Интерфакс или внутренние данные банка собранные на основе транзакций клиента. Их основным преимуществом является полная интерпретируемость результатов и относительная простота реализации и автоматизации.

В мировой финансовой практике наибольшее распространение получили следующие методы:

1) Выявление неудовлетворительной структуры баланса на основе системы критериев оценки возможного банкротства.
2) Модель Э. Альтмана «Z-счет», модели, основанные на факторах.

Система критериев основывается на анализе ликвидности, обеспеченности собственными средствами, способности восстановить или полностью утратить платежеспособность. Данный подход прописан в постановлении Правительства Российской Федерации от 20.05.1994 № 498. В данном постановление приведены

два ключевых признака для определение финансовой состоятельности заемщика [21]:

1) Коэффициент текущей ликвидности, который определяет общую обеспеченность оборотными средствами для ведения предпринимательской деятельности и для возможности погасить необходимую часть кредитной задолженности в установленный срок. Данный коэффициент рассчитывается по формуле 1.3.1, где IIA – итог II раздела актива баланса «Оборотные активы», VП – итог V раздела пассива баланса «Краткосрочные обязательства», VIC – сумма статей V раздела пассива баланса: «Доходы будущих периодов».

$$Кл = IIA / VП – VIC \quad (1.3.1)$$

2) Коэффициент, который показывает обеспеченность организации собственными средствами и вычисляется по формуле 1.3.2, где IIIП – итог III раздела баланса «Капитал и резервы», IA – итог I актива баланса «Внеоборотные активы».

$$Ко = (IIIП - IA) / ПА \quad (1.3.2)$$

Данный подход позволяет выявить неплатежеспособность организации, признать структуру баланса неудовлетворительной. Такой вывод можно сделать, если коэффициент текущей ликвидности был меньше 2, а коэффициент обеспеченности собственными средствами меньше 0,1. Проблема данного подхода заключается в отсутствие предсказательной возможности, он позволяет только определить текущее состояние клиента, но не позволяет его предсказать.

Для решение указанной проблемы используется метод предсказания возможности восстановления (утраты) платежеспособности на основе коэффициентов восстановления платежеспособности и коэффициентов утраты платежеспособности, которые рассчитываются по формулам 1.3.3 и 1.3.4 соответственно, где Клф – коэффициент ликвидности на конец отчетного периода, Клн – коэф-

фициент ликвидности на начало отчетного периода, Клнорм – нормативное значение коэффициента ликвидности, в данном случае равно 2, Т – отчетный период, в месяцах.

$$Кв = (Клф + 6 / Т * (Клф - Клн)) / Клнорм \quad 1.3.3$$

$$Ку = (Клф + 3 / Т * (Клф – Клн)) / Клнорм \quad 1.3.4$$

В случае, когда за шесть месяцев Кв> 1, то признается, что организация имеет может восстановить свою платежеспособность. Если коэффициент текущей ликвидности больше или равен 2 и при это коэффициент обеспеченности собственными средствами больше или равен 0,1, а Ку>1 в течении трех месяцев, то с высокой долей вероятности он не утратит платежеспособность.

В соответствие с законодательством, если структура баланса неудовлетворительна и не существует возможности восстановление платежеспособности, то организация должна быть признана банкротом. Данная методика позволяет определить реальное состояние клиента и предсказать его состояние на некоторый период. Данный метод позволяет регулярно оценивать клиента.

Основной проблемой является необходимость иметь для расчета настоящие данные и получать их своевременно, что часто невозможно, так отчетность РСБУ за квартал публикуется в середине следующего за отчетным кварталом.

Данные критерии в законодательстве были сформулированы на основание зарубежного опыта, что связанно с отсутствием опыта в данном вопросе из-за продолжительного нахождения страны в состоянии плановой экономики.

Эволюцией данных критериев становятся методы, которые основываются на показателе Аргента (А-счете) [27]. Появление данной группы методов вносит огромный вклад в развитие финансового анализа состоятельности хозяйствующих субъектов и прогнозирования риска несостоятельности действующих субъектов в условиях рыночной экономики. Метод, предложенный Джорджом Аргенти, исходит из того, что процесс разорения компании начинается за несколько лет и имеет определенные признаки.

Данный метод заключается в анализе трех групп факторов:

1) Анализ слабостей, в этой группе находятся различные затруднения и недостатки фирмы, которые она может испытывать долгое время, ниже приведены некоторые из них:
    a. Авторитарный директор.
    b. Пассивность совета директоров.
    c. Отсутствие прогноза денежных потоков.
2) Анализ ошибок. Сами слабости не приводят к неплатежеспособности компании, но они приводят к ошибкам, которые приводят к разорению напрямую. Среди явных ошибок, можно выбелить:
    a. Наличие крупного проекта «All-In».
    b. Слишком быстрый рост бизнеса, приводящий к недостатку оборотных средств.
    c. Чрезмерно высокая доля заемных средств.
3) Анализ симптомов. Иногда ошибки могут проходить без последствий, но чаще они приводят к разорению. В этом случае у фирмы появляются «симптомы» неплатежеспособности, которые указывают, что в течение 2-3 лет компания потеряет платежеспособность, вот пример таких симптомов:
    a. Появление публичных скандалов, судебные иски, отставки сотрудников (массово или с высоких должностей).
    b. Снижение финансовых показателей.

Каждому из перечисленных выше критериев ставится оценка, причем строго определенная. Сумма этих оценок и будет А-счетом, если он больше 35, то компания находится на грани разорения, если больше 25, то можно прогнозировать разорение в течение 5 лет, у идеальных компаний показатель колеблется в диапазоне от 5 до 18.

А-счет является крайне точным и высокоэффективным показателем, но обладает рядом проблем: некоторые показатели субъективны, некоторые показа-

тели скрыты от стороннего наблюдателя. Данный метод трудно автоматизировать, что связанно с необходимостью, например, строить систему, которая будет просматривать различные новостные ресурсы и сайты судов, на предмет скандалов и судебных дел.

С другой стороны, если переработать данный метод, то можно основываясь только на публичной информации предсказывать состояние клиента, возможно с несколько меньшей точностью, чем оригинальный метод. Такой подход позволит обойти такие виды мошенничества как предоставление в кредитную организацию не действительную информацию заемщиком.

Для оценки вероятности банкротства и неплатежеспособности компании стали активно применять модели Э. Альтмана, Ж. Конана, Ж. Лего, Р. Лиса, Г. Спрингейта, Р.Тафлера, Г. Тишоу и других. Все эти модели основываются на стохастическом факторном анализе и на детерминированном факторном анализе.

Построенные на основе стохастического факторного анализа, к которым принадлежат крайне известные Z-модели, первую из которых предложил Э. Альтман, основываются на идеи разделения всех исследуемых предприятий на два класса: подлежащие банкротству и способные его избежать [31]. Данное разделение происходит на основе моделирования классифицирующей функции в виде корреляционной модели. При построение данных моделей ставится задача максимизации точности предсказания неплатежеспособности клиента за счет эмпирического уравнения определения дискриминантной границы, разделяющей в пространстве признаков исследуемые фирмы.

Такая функция называется дискриминантной или Z-счет. Дискриминантная функция обычно представляется в линейном виде, как показано в формуле 1.3.5:

$$Z = a_1X_1 + a_2X_2 + \cdots + a_nX_n \qquad 1.3.5$$

- Z – дифференциальный индекс (Z-счет);
- X_i – независимая переменная (i = 1, …, n);

- a_i – коэффициент при независимой переменной.

Z-модели получили широкое распространение из-за ряда из достоинств, самыми важными из которых стали:

1) Простота применения, в следствии того, что чаще всего они рассчитываются на основе бухгалтерской отчетности.
2) Достаточно высокая точность прогноза.
3) На оценку влияют сразу несколько факторов.
4) Возможность охватить широкий круг разнообразных факторов.

К основным недостаткам можно отнести:

1) Неустойчивость к изменениям в исходной информации.
2) Отсутствие статистической однородности выборки событий.

Модели, основанные на детерминированном факторном анализе призваны устранить приведенные выше недостатки стохастического факторного анализа. Данные модели делятся на два больших семейства: однофакторные и многофакторные.

Однофакторные модели призваны количественно предсказать платежеспособность компаний на основе одного частного показателя. Однофакторные модели не стали популярными и сейчас почти не используются из-за невысокой предсказательной силы. Куда более широкое распространение получили многокритериальные модели, что связанно с их возможностью охватывать многие сферы деятельности предприятия и учитывают отраслевые особенности. Данные методы основываются на скоринговом и комплексном анализе.

Одной из первых моделей, которая основывалась на интегральной модели стал американский ученый Э. Альтман. Интегральная модель – совокупность коэффициентов с весовыми значениями, которая рассчитывает интегральный показатель, который позволяет оценить финансовое состояние предприятия. На рисунке 5 представлены различные модели прогнозирования вероятности неплатежеспособности.

В качестве примера будет приведена пятифакторная модель Альтмана. Данная модель была разработана для анализа организаций с акционерной формой капитала, при этом акции представлены на фондовом рынке. Несколько позже была создана Э. Альтманом была создана вторая пятифакторная модель для компаний с акциями, которые не котируются на бирже. Уравнение данной модели приведена в формуле 1.3.6.

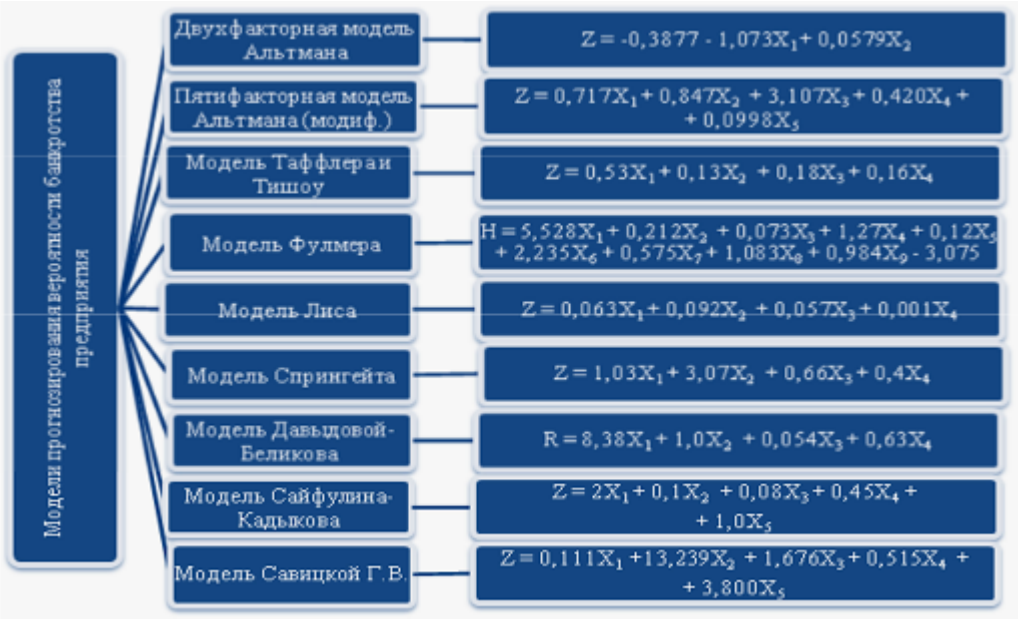

Рисунок 5. Количественные модели прогнозирования вероятности банкротства предприятия [31]

Она основывается на следующих параметрах: X_1 – оборотный капитал к сумме активов предприятия, показатель оценивающий сумму чистых ликвидных активов компании по отношению к совокупным активам; X_2 – не распределенная прибыль к сумме активов предприятия, который отражает уровень финансового рычага компании; X_3 - прибыль до налогообложения к общей стоимости активов, отражает эффективность операционной деятельности компании; X_4 - рыночная стоимость собственного капитала / бухгалтерская (балансовая) стоимость всех обязательств; X_5 - объем продаж к общей величине активов предприятия, характеризует рентабельность активов предприятия.

$$Z = 1{,}2X_1 + 1{,}4X_2 + 3{,}3X_2 + 0{,}6X_4 + X_5 \qquad 1.3.6$$

Для интерпретации результатов расчета их сравнивают с пороговыми значениями:

1) Зона финансового риска, при Z < 1,8.
2) Зона неопределённости, при Z <= 2,9 и Z > 1,8.
3) Зона финансовой устойчивости, при Z > 2,9.

Данная модель обладает крайне высокой точностью, она может предсказать банкротство на горизонте года с точностью 95%, на горизонте два года с точностью 83%. Все модели, основанные на Z-счете, используют в своей основе идею Альтмана, в том числе и разработки российских экономистов.

Несмотря на то, что данные модели с высокой точностью могут предсказать банкротство компании, они не могут помочь в определение мошенников в сфере кредитования и незаконное получение кредита. Это связанно с тем, что данные модели используют в своей основе финансовую отчетность клиента, которая с высокой долей вероятности будет подделана. Но при этом, если заемщик намеренно ухудшает свое финансовое состояние для совершения преднамеренного банкротства, но при этом не занимается подделкой документов, модели на основе Z-счета легко предскажут ухудшение состояния.

1.4 Вероятность дефолта и уровень потерь при дефолте как методы по выявления мошенников в сфере банковского кредитования среди юридических лиц

PD (probability of default) – вероятность дефолта. LGD (loss given default) – уровень потерь при дефолте. Оба показателя были внесены в «Базель II» - документ, который был принят в 2004 году Базельским комитетом по банковскому надзору. Оба включены в рамках подхода Internal Rated Based Approach для измерения кредитного риска и представляют собой математические модели. А в 2014 году Банк России представил проект Положение «О порядке расчета величины кредитного риска на основе внутренних рейтингов», в котором также фигурируют PD и LGD [18].

Устаревший метод оценки кредитного риска строится на основание фиксированных весовых коэффициентов, которые были предложены Банком России. Внедрение новой системы позволяет резко увеличить качество оценки рисков. В представленном Банком России документе предлагается три подхода:

1) Использовать коэффициенты.
2) Использовать только PD.
3) Использовать PD, LGD, EAD.

В данной статье представлен сравнительный анализ представленных выше подходов, где показано, что первый подход самый простой, но и наименее выгодны, второй подход обладает средними характеристиками, третий подход труднореализуемый, ресурсозатратный, но при этом обладает крайне высокой эффективностью.

Стоит отметить, что Exposure at default (EAD) – сумма под риском, данный показатель используется в методике Базель II и означает сумму, которую может потерять банк в случае дефолта заемщика.

Построение самих моделей PD и LGD является крайне трудной задачей, и банки, которые используют данную технологию тщательно скрывают их реализацию даже от своих сотрудников, предотвращая таким образом мошенничество совместно с сотрудниками банка.

Есть несколько работ, которые предлагают реализацию моделей LGD и PD. Так в данной работе предлагаются 4 подхода на основе регрессии к построению модели LGD, в ней используются линейная регрессия, бета-регрессия, бинарная трансформация, бета-трансформация. Все четыре модели показали крайне близкие результаты, значение критерия Gini находится в ε-окрестности равной 0.4% от 58%. Из-за сильного сходства моделей по качеству, автор приводит сравнение их преимуществ и недостатков. Также в данной работе предлагается метод расчета PD и двустороннее преобразование PD к LGD.

При этом в данной работе не указывается метод предварительной обработки данных, что несколько уменьшает ее актуальность. Но автор решает данную проблему в данной статье, предлагая методы обработки данных, их отбор и кодирование.

Автор предлагает использовать Information Value (IV) и Weight of Evidence (WoE), ниже приведены формулы их расчета 1.4.1 и 1.4.2 соответственно. Данный подход является классическим для построения скоринговых моделей на основе линейной и логистической регрессии. Это обосновывается тем, что закодированные при помощи WoE признаки представляют собой значение логистической регрессии, построенной на только на данном факторе. А IV является влияние признака, закодированного при помощи WoE на целевую переменную.

$$WoE = \ln \frac{Event\%}{Non\ Event\%} \quad 1.4.1$$

$$IV = \sum(Event\% - Non\ Event\%) * (WoE) \quad 1.4.2$$

Есть и более сложные подходы к построению модели LGD, так модель, предложенная ПАО Сбербанк, подразумевает совмещение результатов сразу шести различных моделей. Так ими была предложена модель, являющаяся линейной комбинацией $LGD_р$ – LGD «реализации», потери при дефолте в случае если необходимо реализовать имущество, $LGD_с$ – LGD «списания», потери если придется списывать долг при дефолте и $LGD_в$ – LGD «выздоровление», потери если после дефолта заемщик смог войти в новый, предложенный при реструктуризации график платежей с умноженные на свои вероятности, уравнение приведено в формуле 1.4.3.

$$LGD = LGD_в * p_в + LGD_р * p_р + LGD_с * p_с \quad 1.4.3$$

Безусловно каждый из этих показателей может рассчитываться индивидуально для каждого клиента с помощью машинного обучения, финансовых формул, эконометрических подходов, или же могут быть рассчитаны статистически по для всего портфеля или его среза.

Описаны и системы расчета значения PD и LGD на основе сложных моделей машинного обучения. В данной работе приведен метод расчета основанный на градиентном бустинге, в статье используются реализация от Microsoft XGBoost и реализация от Яндекс CatBoost. При этом, несмотря на то, что в статье использовались более сложные методы, чем линейная регрессия, которая рассматривалась ранее, их качество заметно ниже, Gini равняется 0,42.

Стоит отдельно отметить факторы, которые используются для расчета моделей LGD и PD, частично они совпадают с факторами, которые используются для расчета Z-счета, приведенного в предыдущем параграфе. Данное совпадение не случайно и говорит о том, что существует не так много значимых для предсказания состояния клиента. Но список используемых факторов не ограничивается пятью, как в модели Альтмана. Далее приведен пример набора факторов:

1) Прибыль до вычета налогов и процентов по кредитам на выручку.
2) Чистый оборотный капитал на активы.
3) Прибыль до вычета налогов и процентов по кредитам минус дебиторская задолженность на выручку.
4) Чистая прибыль на собственный капитал.
5) Коэффициент абсолютной ликвидности.
6) Чистая прибыль на выручку.
7) Нераспределенная прибыль на активы.
8) Коэффициент текущей ликвидности.
9) Прибыль до вычета налогов и процентов по кредитам на активы.
10) Коэффициент срочной ликвидности.
11) Прибыль до вычета налогов и процентов по кредитам минус дебиторская задолженность на активы.
12) Чистый оборотный капитал на активы.
13) Общий долг на активы.
14) Долг на собственный капитал.
15) Чистый оборотный капитал на краткосрочные активы.

Приведенные факторы не являются наилучшим набором из всех возможных. Выбор наилучшего набора зависит от конкретного реализуемого алгоритма и среза портфеля, на котором планируется рассчитывать показатели [9].

Сам подход в расчете кредитного риска используя методику предложенную Базель II, подразумевает формирование кредитного портфеля, расчет вероятности дефолта потенциальных заемщиков, которым должны быть выданы займы в следующий отчетный период, расчет финансовых показателей потенциальных заемщиков на основе вероятности дефолта.

На основе рассчитанных показателей и кредитной политики банка может быть упрощена процедура выдачи кредитов некоторым заемщикам. В основу данного метода положен расчет стоимости, выданной суды для банка на основе фондирования. Далее происходит корректировка стоимости данной ссуды на риск дефолта. После чего стоимость ссуды рассматривается как изменяемая во времени и зависящий от стоимости фондирования и изменения сроков погашения кредита. Дальнейшие предположение, которое делается в расчетах, не является тривиальным, для определения волатильности внутренней стоимости кредита предлагается рассмотреть выдачу банком займа как реальный опцион колл на облигацию заемщика. В таком случае цена этого опциона будет отражать издержки банка по кредиту.

Рассчитав цену этого опциона, можно решить систему уравнений относительно волатильности методом Ньютон-Рафсона и вывести волатильность изменений внутренней стоимости долга для банка. Далее предлагается рассчитать дюрацию кредитного портфеля исходя из задачи оптимизации кредитного портфеля. Предполагается, что дюрация портфеля не будет превышать дюрацию пассивов банка. Двумя условиями выдачи кредита являются: полная иммунизация портфеля и размер выданных кредитов не превышает размер привлеченных пассивов.

Задача оптимизации портфеля в начальный момент времени, с учетом всех условий, сводится к решению задачи максимизации уровня доходности при желаемом уровне волатильности, при этом на систему уравнений накладывается условие иммунизации. Для решения этой системы уравнения необходимо найти подходящее значение EAD.

После этих расчетов предлагается метод переформирования портфеля к следующему отчетному периоду, за счет погашения старых займов и выдачи новых. План переформирования портфеля подготавливается в момент формирования предыдущего портфеля, что позволяет рассчитать показатели идеального клиента и выдать ему кредит по упрощённой процедуре.

При расчете переформирования портфеля уже известны показатели PD и финансовые показатели заемщиков, так как уже были выданы кредиты в начальный момент времени. Так же уже известен план по росту кредитного портфеля.

Для нахождения финансовых показателей желаемых заемщиков, требуется найти целевой показатель PD для новых заемщиков, для чего необходимо решить задачу оптимизации относительно переменных вероятности дефолта.

После того, как такая вероятность найдена, банку необходимо решить обратную задачу, где на основе внутренних моделей PD вычисляются значения финансовых показателей желаемых заемщиков.

В литературе встречается еще один подход к использованию LGD и PD, для расчета кредитного риска. Данный метод реализует методологию Value-at-Risk, которая является оценкой убытков, выраженной в валюте с некоторым заданным не превышающей потери портфеля в заданный период времени. Данный подход выражен в формуле 1.4.4, где p – заданный доверительный уровень, $Loss_p$ – величина убытков по портфелю. VaR – это статистический подход, который основывается на распределение вероятностей величин рыночных факторов [51].

$$P\{Loss_p < VaR\} = p \qquad 1.1.4$$

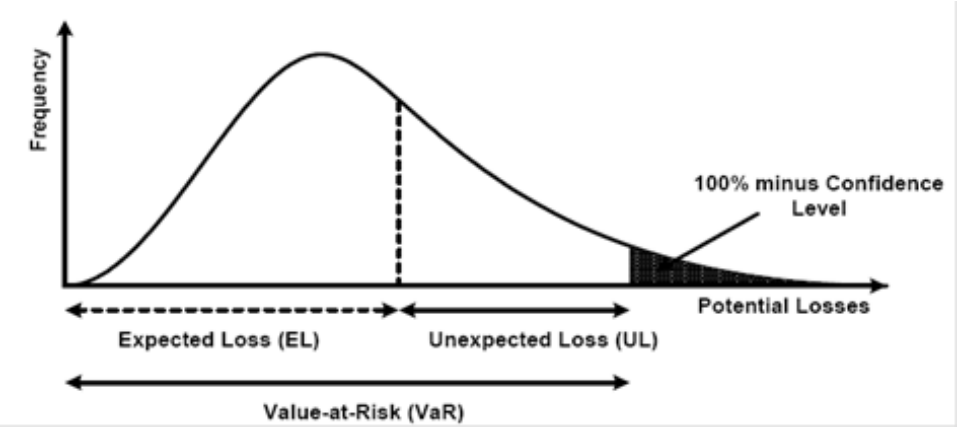

Рисунок 6. Распределение потерь по кредитному портфелю [51]

В рамках данной методологи кредитный риск разбивается на два компонента: это классический для подходов расчета кредитного риска при помощи LGD и PD показатель ожидаемых потерь (Expected Loss, EL) и неожиданные потери (Unexpected Loss, UL), данное распределение представлено на рисунке 6.

Если существуют принятые в Базель II методы расчета EL, приведены в формуле 1.1.5, то UL рассчитывается только как разница между всеми потерями и ожидаемыми потерями. В следствии данного, алгебраического определения, UL можно сказать, что с точки зрения теории вероятность UL – это отклонение потерь от их среднего ожидаемого уровня.

$$EL = LGD * PD * EAD \qquad 1.1.5$$

В связи с тем, что представленное на рисунке 6 распределение не является ни одним из стандартных, и, хотя оно похоже нормальное, таковым оно не является, что можно выяснить используя, к примеру, t-тест авторами предлагается использовать метод Монте-Карло для вычисления общего риска. На основании этого значения и вычисляется UL.

Таким образом модели PD и LGD, могут участвовать в поиске мошенников различными способами. В случае, если заемщик не подделывает документы или не подделывает показатели, на которых считается PD, то данная модель сможет показать движение заемщика к банкротству, даже преднамеренному. В случае подделки документов, которые используются для расчета PD, необходимо рас-

сматривать еще и LGD, значение которого основывается не только на бухгалтерских данных, но и на данных о залоге и других показателей, которые трудно подделать.

Кроме использования самих моделей, можно использовать и метод расчета показателей по портфелю, в который может учитывать возможность мошенничества и, хотя не предотвратит его, но скомпенсировать потери за счет правильного составления портфеля. На основании UL можно балансировать портфель уменьшая данный показатель, что уменьшит риск выдачи кредита мошеннику.

1.5 Использование алгоритмов машинного обучения для выявления мошенников в сфере банковского кредитования среди юридических лиц

Необходимо уточнить, что в данной работе алгоритмы глубокого обучения являются одним из видов алгоритмов машинного обучения, их соотношение продемонстрированно на рисунке 7.

Алгоритмы машинного обучения рассматривались и ранее в данной работе как напрямую в обзоре методов построения LGD и PD, так и неявно, при рассмотрении моделей, основанных на Z-счете, которые представляют собой линейную регрессию с уже подсчитанными коэффициентами.

В литературе описывается различные методы применения систем машинного обучения в банкинге, большая часть из них описывает системы банковского скоринга, а системы поиска мошенников в подавляющем большинстве описывают задачу для физических лиц.

Массовое внедрение систем на основе машинного обучения в процессы банков связанно с высокой эффективностью данных методов. Так в исследование Feedzai, в котором банкам предоставлялось программное обеспечение для фильтрации новых заявок на открытие счета и утверждения новых клиентов и оценивая риск. По результатам исследования, все банки, которые установили ИИ

Feedzai увеличили новые установки приложений и уменьшили ложно положительных срабатываний систем поиска мошенников, при этом потери от мошенничества не увеличились [8].

На сегодняшний день машинное обучения является одним из главных трендов в ИТ-трансформации банков, которую возглавляют банки Тинькофф и Сбербанк. В рамках проекта AI-First Сбербанк в 2018 году сосредоточился на подготовке инфраструктуры, процессов, моделей, удобных платформ для разработки. Данный отчет был представлен Сбербанком в начале 2018 года [19].

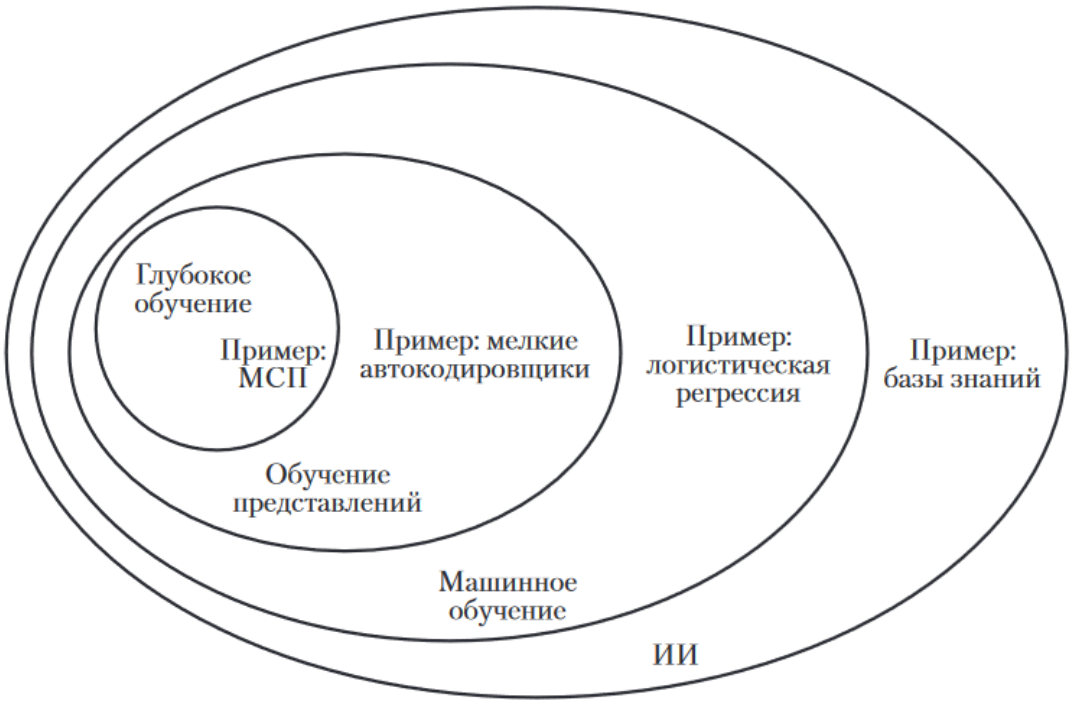

Рисунок 7. Диаграмма Венна на которой представлены взаимоотношение разных видов ИИ [11]

В различных статьях приводятся положительные и отрицательные стороны использования машинного обучения в банковской сфере, в том числе и в сфере безопасности. Обычно приводят следующие положительные стороны применения машинного обучения [23]:

1) Быстрота и эффективность. Вместо ручной проверки, используется компьютерный анализ, при этом обычно обладающий высокой точностью

работы. При если точность не сильно ниже человеческой, то эта дельта может компенсироваться скорость.

2) Экономическая эффективность. Она следует из предыдущего пункта, более того, массовое внедрение ИИ уменьшает затраты на сотрудников или перенаправляет\концентрирует их на другие задачи.

В качестве недостатков выделяют следующие пункты:

1) Не прибыльные системы. Иногда, гипотеза о том, что алгоритм машинного обучения поможет в конкретной задаче не подтверждается, что приводит к пустой трате ресурсов на разработку. Данный минус в большей части нейтрализуется внедрением методологии циклической разработки, или Agile в процесс.
2) Невозможность интерпретации. На сегодняшний день можно интерпретировать небольшое количество семейств моделей машинного обучения, в частности: линейные модели, деревья решений. Невозможность интерпретации ведет к невозможности принятия решения. Данный минус решается новыми алгоритмами интерпретации более сложных моделей, таких как бустинги и нейронные сети.
3) Ошибки при внедрении. Часто при внедрении модели происходят ошибки, которые могут заметить не сразу, и они приведут к большим финансовым и репутационным потерям. Данная проблема устраняется внедрением систем валидации самой модели и ИТ-валидации модели после внедрения, но до запуска.

В основном системы борьбы с мошенничеством на основе машинного обучения описываются только для розничного кредитования, что связано с ограниченным набором способов мошенничества. Методы для юридических лиц описаны кратко и не являются основной темой статей. К примеру, в данной статье предлагается просмотр конверсионных операций для поиска мошенников среди юридических лиц [57].

Одной из самых больших проблем является сбор выборки, банк обладает большим количеством примеров хорошего поведения клиента, и крайне малым количеством примеров мошенничества. Такая проблема в машинном обучение называется несбалансированной выборкой.

В работе Анны Бучневой предлагается сравнение семи моделей машинного обучения в задаче поиска мошенников. Была взята обучающая выборка из 10000 прецедентов, из которых 9932 были не мошеннических, а тестовая выборка составила 1000 прецедентов и из них 996 были не мошеннических. В качестве данных были взяты:

- Сумма операций.
- Тип операций (выдача наличных, платеж).
- Состояние баланса отправителя до операции.
- Состояние баланса получателя до операции.
- Состояние баланса отправителя после операции.
- Состояние баланса получателя после операции.

В качестве меры качества для сравнения моделей была выбрана accuracy в разрезе каждого класса. В результатах исследования наблюдается большой разброс по качеству модели, и в таких условиях банк должен выбрать что ему важнее: точно определить всех хороший клиентов, но пропустить часть мошенников или пропустить сильно меньше мошенников, но заставить сотрудников в пустую дополнительно проверять являются ли данные клиенты мошенниками. В таблице 2 приведены результаты тестирования различных типов моделей машинного обучения.

Достаточно часто стали применяться системы визуального скоринга, который основан на классификации изображений. Его используют для оценки объектов залога, на случай, если для этой задачи привлекают стороннего оценщика и боятся, что он даст заведомо ложную оценку или в целях экономии, для минимизации человеческой оценки. Могут оцениваться внешний вид и лицо управляющего лица заемщика, лицо может исследоваться на факт фигурирования других

делах о мошенничестве, так и на определенные черты благонадежности, что является не очень точным методом. Гораздо лучше работает система способная распознавать пульс и эмоции во время переговоров для последующего анализа.

Таблица 2

Сравнение методов для обнаружения мошенничества

Метод	Обучающее (обыч.)	Тестовое (обыч.)	Обучающее (мошен.)	Тестовое (мошен.)
Линейная регрессия	1.00	0.03	1.00	0.50
Логистическая регрессия	0.98	0.93	0.97	0.50
Нейросеть	0.99	0.06	1.00	0.00
Дерево решений	1.00	1.00	1.00	0.75
Случайный лес	1.00	0.51	1.00	0.75
Метод опорных векторов	1.00	1.00	1.00	0.00
Наивный байесовский классификатор	0.99	0.99	0.98	0.50

Предлагается использовать технологию генеративно-состязательных нейронных сетей для поиска мошенников. Данная технология заключается в создание двух нейронных сетей, одна из которых должна учиться распознавать является ли клиент мошенником, а вторая должна учится создавать мошенническое

поведение и нормальные поведение клиента. В результате должны две нейронные сети должны соревноваться и увеличивать свою эффективность. Данный подход не представляет интереса, так как подразумевает обучение сети-генератора с помощью сети-классификатора, а сеть-классификатор уже должна быть обучена.

Стоит отдельно обратить внимание на архитектурные решения, которые используют при создании систем машинного обучения. В отличие от других методов, например, финансовых, они имеют определенную особенность. В данной статье представлена крайне интересная система кредитного скоринга и предсказания платежеспособности клиента [56]. Общая структура данной системы представлена на рисунке 8.

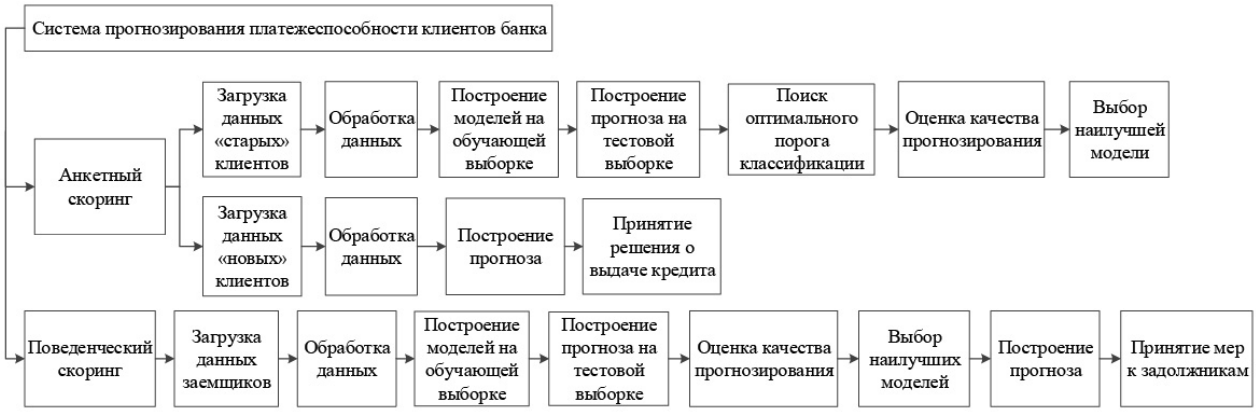

Рисунок 8. Схема системы скоринга клиента и прогнозирования его платежеспособности [56]

Данная система основана на двух моделях, одна из которых определяет стоит ли выдавать кредит заемщику, а вторая предсказывает платежеспособность заемщика, которому уже был выдан кредит, на рисунке 8 эти системы называются «анкетный скоринг» и «поведенческий скоринг» соответственно.

Для реализации анкетного скоринга используется агрегация 7 классификаторов: нейронная сеть, логистическая регрессия, дискриминантный анализ, наивный байесовский классификатор, метод опорных векторов, деревья решений, случайный лес.

Результат скоринга формируется по одной из следующих процедур: среднее значение классификаторов, медианное значение классификаторов, голосование.

Агрегация по среднему значению подразумевает усреднение вероятности принадлежности объекта к данному классу по формуле 1.5.1, где $P(Y^{AK_{mean}})$ – вероятность кредитоспособности клиента, $P(Y^i)$ – вероятность полученная i-ым классификатором, H – количество классификаторов. Затем на основе данной вероятности принимается решение о принадлежности объекта классу 1.

$$P(Y^{AK_{mean}}) = \frac{\sum_{i=1}^{H} P(Y^i)}{H} \quad 1.5.1$$

Для агрегации по медиане необходимо сначала провести ранжирование результатов всех моделей, выбрать средний элемент ряда, если количество элементов нечетное, то выбирается среднее между двумя элементами посередине. Далее на основание выбранной вероятности формируется класс клиента.

В случае, если используется голосование, то сначала для каждого алгоритма предсказанная вероятность переводится в класс, а финальный класс формируется как мода классов всех алгоритмов. Данный алгоритм является наилучшим из всех алгоритмов агитаторов и активно используется, в отличие от других алгоритмов он позволяет выбрать порог отсечения вероятности максимально эффективный для каждого алгоритма.

Агрегация результатов нескольких алгоритмов способно нивелировать проблемы каждого из них, например, ограниченность значений дерева, возможность нейронной сети попасть в локальный минимум при градиентном спуске или плохое качество линейных алгоритмов в зоне где классы разделяет нелинейная поверхность.

Предсказание дальнейшей платежеспособности клиента происходит на основе марковских цепей первого и второго порядка, которые позволяют учитывать предыдущие состояния кредитной истории клиента, цепи представлены в

формулах 1.5.2 и 1.5.3 соответственно, где $v_i(t)$ – вероятность перехода кредитного договора в состояние S_iв момент времени t; w – количество состояний; $p_{ij}(t)$ – вероятность перехода договора в состояние S_j из состояния S_i за один переход; $\varphi_k(t+1)$ – вероятность перехода счета в состояние S_kв момент времени t+1, если предыдущими состояниями были S_i и S_j.

$$v_i(t+1) = \sum_{i=1}^{w} p_{ij}(t) * v_i(t) \quad 1.5.2$$

$$\varphi_k(t+1) = \sum_{i=1}^{w} \sum_{j=1}^{w} p_{ijk}(v_i(t-1)\, v_j(t)) \quad 1.5.3$$

В качестве состояний используется информация о просроченной задолженности по кредитным договорам. А для оценки вероятностей используются методы машинного обучения. То есть данная система прогнозирует платежеспособность клиента на основе его предыдущего состояния и данных о самом клиенте. В таблице 3 представлены результаты тестирования модели кредитного скоринга, а в таблице 4 представлены результаты тестирования модели предсказания следующего состояния.

Таблица 3

Значения дисперсии ошибок прогнозирования

Классификатор	Полная исходная выборка	Информативные признаки	Дискретизация и информативные признаки
НС	0,2523	0,2469	0,2513
ДА	0,2533	0,2435	0,2445
БК	0,3671	0,3121	0,2955
МОВ	0,2466	0,2449	0,2448
ДР	0,3796	0,3544	0,3290
ЛР	0,2457	0,2431	0,2439
БДР	0,2278	0,2206	0,2595
АК	0,22721	0,21872	0,24363

Такой подход к предсказанию платежеспособности клиента основывается в том числе и на его предыдущем состояние и внутренние данные банка о клиенте, а не только данные, полученные от клиента. Это позволяет нам предположить вероятность возврата долга банку, что и является основой целью.

Таблица 4

Процент верных прогнозов по всем возможным переходам из состояния S_i в другие состояния

Состояние	**Марковская цепь 1-го порядка**						**Марковская цепь 2-го порядка**						**Макс.**
S_i	ЛР	ДА	НБК	ДР	МОВ	НС	ЛР	ДА	НБК	ДР	МОВ	НС	
S_1	33,3	0	0	33,3	33,3	33,3	33,3	33,3	33,3	33,3	33,3	33,3	33,3
S_2	94,1	92,1	17	95,9	11,9	94,5	96,1	94,3	13,6	95,5	29,4	96,4	96,4
S_3	4,2	1,4	1,4	25,4	0	0	5,6	11,3	50,7	28,2	2,8	2,8	50,7
S_4	79,1	68,6	1,2	50	20,9	86	63,9	61,6	8,1	38,4	20,9	74,4	86
S_5	53,3	43,3	10	31,7	3,3	56,7	43,3	46,7	16,7	25	33,3	51,7	56,7
S_6	98,2	97,1	30,6	92,6	55	98,5	97,1	95,9	52,6	89	10,6	98,2	98,5
S_7	94,7	94,7	5,3	31,6	0	73,7	89,5	89,5	0	31,6	0	68,4	94,7
S_8	94,9	93,8	3,6	94,9	0,3	94,9	94,9	93,8	5,4	94,6	10,7	94,9	94,9
S_9	36,9	41,1	20,1	19,6	50	38,3	17,3	24,8	15,9	18,2	38,8	20,1	50
Ср.	65,4	59,1	9,9	52,8	19,4	64	60,1	61,2	21,8	50,4	20	60	**73,5**

В задачах нахождения мошенников в розничном кредитовании хорошо зарекомендовал себя метод SMOTE и ENN [41]. В данном случае исследователь исходит из предположения, что классы мошенник и не мошенник хорошо разделены в пространстве признаков, но задача не решается из-за того, что классы не сбалансированы. Так как клиентов с мошенническим поведением обычно менее 1% от общей выборки обычные методы работы с несбалансированными выборками не подходят. В таблице 5 представлены результаты выравнивания выборки методом SMOTE и ENN, в качестве классификатора используется случайный лес, в выборке 28 признаков, выборка анонимизирована.

Таблица 5

Результаты применения методов SMOTE и ENN к задаче поиска мошенников

Класс	Precision	Recall	F1-score	Выборка
0	1	1	1	56878
1	0,85	0,87	0,86	84
Среднее	0,925	0,935	0,93	56962

Метод SMOTE (Synthetic minority oversampling technique) является методом генерации новых объектов меньшего класса. В отличие от обычного Oversampling, который, просто, дублирует уже имеющиеся объекты, метод SMOTE использует гипотезу о нахождение объектов одного класса на одном многообразие в пространстве признаков. Данный метод создает новые объекты, они похожи на существующие, но не являются их копиями. Процесс создания нового объект крайне прост, на рисунке 9 визуализирован результат работы данного метода:

1) Для случайного объекта X_1 находится его ближайший сосед X_2, методом kNN.
2) Находится расстояние между двумя объектами X_1 и X_2 в пространстве признаков, получают вектор d.
3) Получившийся вектор d умножается на случайное число от 0 до 1, получаем вектор d^*.
4) Вектор d^* складывают с вектором X_1, полученный объект является новым сгенерированным признаком.
5) Далее связка из объектов X_1 и X_2 удаляется из рассмотрения и алгоритм повторяется до получения необходимого количества объектов.

Метод ENN (Edited Nearest Neighbor) является методом для уменьшения (Undersampling) количества примеров превосходствующего (мажоритарного) класса. Как и все семейство методов undersampling nearest neighbor, ENN основан на идеи, что объекты миноритарного находятся в разных кластерах с большей частью объектов мажоритарного класса. Используя это предположение происходит отсечение части выборки мажоритарного класса, которая линейна разделима со всеми объектами миноритарного класса. Алгоритм ENN, так же, как и SMOTE использует kNN, результаты его работы представлены на рисунке 10:

1) Пусть L – это вся выборка, тогда создается S подмножество L, которое формируется как все объекты миноритарного класс и 1 объект мажоритарного класса.

2) Для каждого объекта из S ищется k его ближайших соседей, каждый сосед другого класса помещается в S, которая является урезанной выборкой.

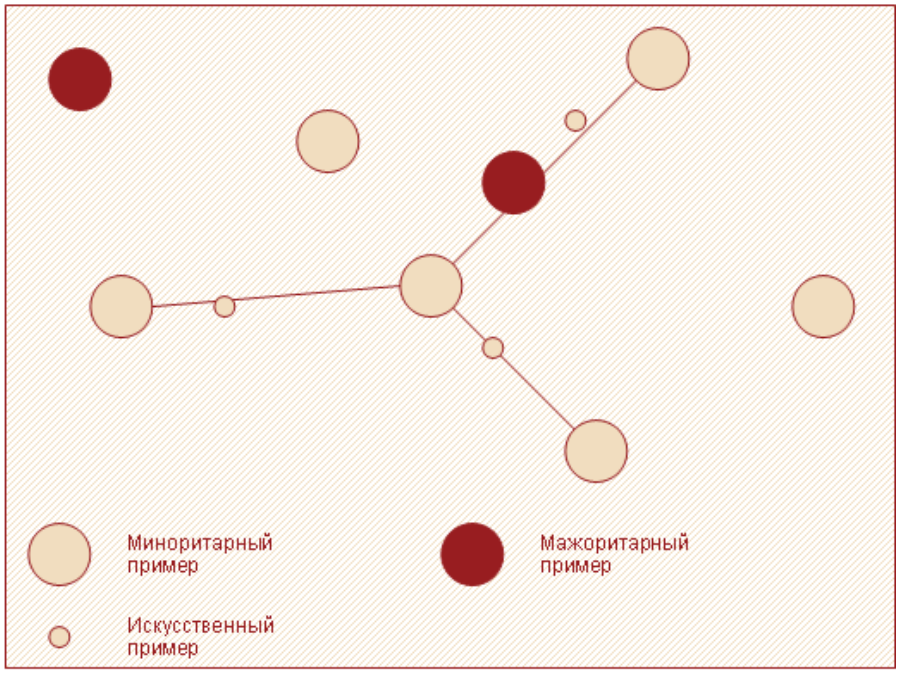

Рисунок 9. Пример работы алгоритма SMOTE [41]

Еще один подход к решению данной задачи в розничном кредитование строится на предположение, что если объектов с мошенническим поведением сильно меньше, чем с нормальным, то объекты с классом мошенником являются выбросами в общей выборке. Следовательно, стоит использовать методы детекции аномалий для нахождения мошенников [39]. Как классические алгоритмы для поиска аномалий используют, визуализация работы алгоритмов представлена на рисунке 11:

1) Метод опорных векторов для одного класса (OneClassSVM).
2) Изолирующий случайный лес (IsilationForest).
3) Эллипсоидальная аппроксимация данных (EllipticEnvelope).

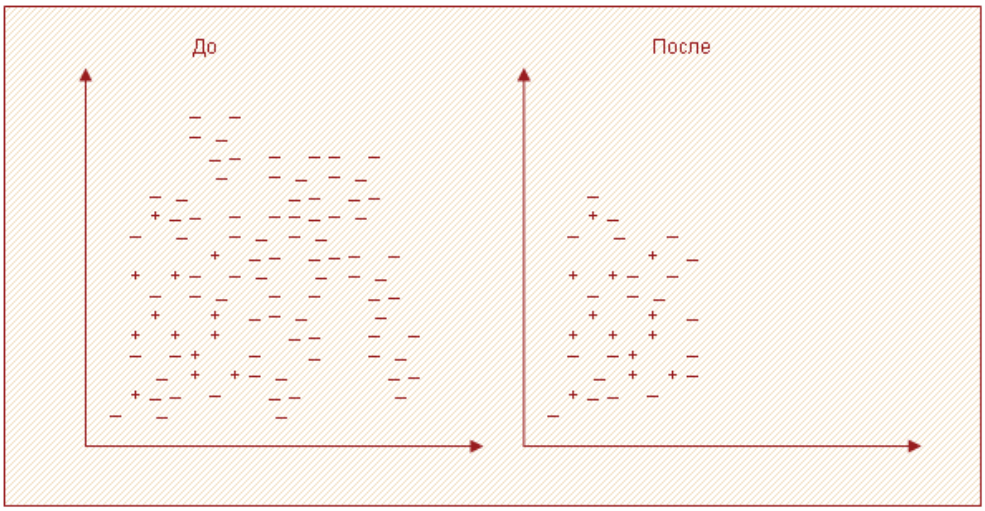

Рисунок 10. Пример работы алгоритма ENN [41]

OneClassSVM – это обычный SVM задача которого построить область с центром в начале координат и отделить ей максимально большое количество объектов выборки. В основном используется только ядра RBF так, как только они приводят к качественному результату, что можно увидеть на рисунке 11 сравнив результаты OneClassSVM ядром RBF и полиномиальным ядром. Данный алгоритм, скорее предназначен для поиска новшеств в поступающих данных, чем для поиска аномалий в текущей выборке, но на практике данный метод отлично справляется с обеими задачами.

IsolationForest – является одним из видов случайного леса, но в этом случае делается предположение, что выбросы будут попадут в листья решающих деревьев при небольшой глубине. Данный алгоритм лучше всех представленных справляется с задачей нахождения аномалий.

EllipticEnvelope – создает разделяющую плоскость в виде гиперэллипсоида вокруг выборки минимизируя расстояние Махалонобиса от начала координат, но при этом максимизируя количество точек, которое попадет в гиперэллипсоид.

Как можно заметить, алгоритмы OneClassSVM и EllipticEnvelope требуют, чтобы выборки были нормализованы, а IsolationForest не накладывает такие требования.

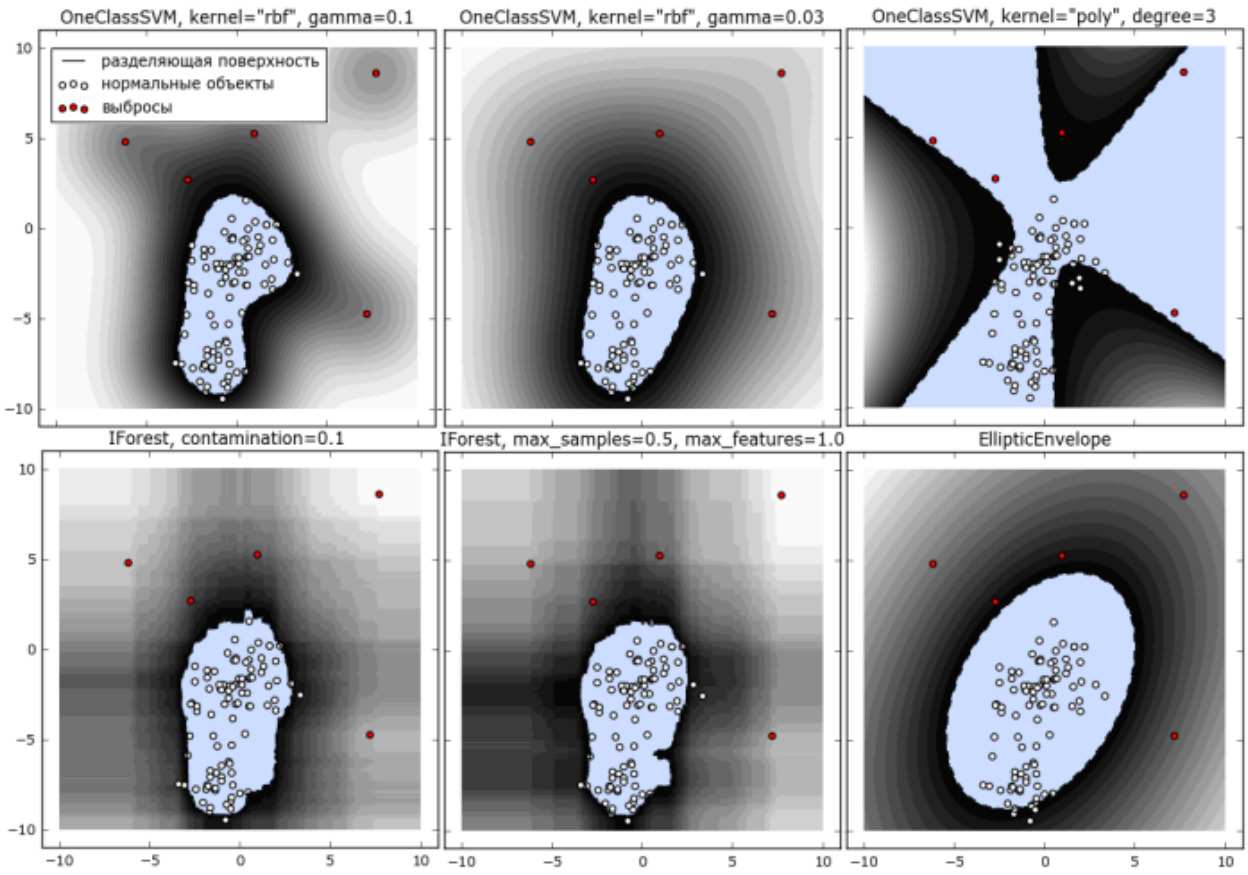

Рисунок 11. Визуализация работы разных алгоритмов поиска аномалий [39]

Как можно увидеть, машинное обучение используется и в других методах, на основе LGD и PD, финансовых методах, хотя, казалось бы, они должны основываться линейной алгебре и статистике.

1.6 Постановка задачи на разработку методики создания ИТ-инфраструктуры для выявления мошенничества с банковскими кредитами среди юридических лиц

Для правильной постановки задачи необходимо выбрать точный сегмент, для которого будет строится данная инфраструктура. В рамках данной работы будет предложена методика создания ИТ-инфраструктуры для крупного банка, которому необходимо использовать технологии работы с Big Data для хранения и обработки информации о клиентах.

Также стоит оговориться, что в данной работе делается предположение о соблюдение общих рекомендаций, указанных в параграфе 2 данной главы. Самым важным пунктом является контроль сотрудников на факт состояния в сговоре с мошенником, если он не выполняется, то самая лучшая система не сможет предотвратить потерю денег, хотя и поможет выявить злоумышленника.

Учитывая данные замечания, необходимо разработать методику построения системы для выявления мошенников на этапе анализа кредитной заявки, на этапе мониторинга, пока внутренних данных о клиенте мало и на этапе мониторинга, в случае, когда внутренних данных о клиенте достаточно для построения вывода на их основе.

Для построения данной системы необходимо использовать методы машинного обучения и другие современные и эффективные математические алгоритмы, что обосновывается их эффективностью в решении задач, основанных на данных, популярности в обществе, что приведет к росту инвестиций в компанию кредитора, и успешного опыта решения подобных задач. Необходимо проработать способ хранения и передачи результатов анализа непосредственно сотруднику для дальнейшей работы с ними. Взаимодействие данной системы с внешними и внутренними источниками данных. А также дополнительная система контроля за своевременным выполнением задач по анализу мошеннического поведения сотрудниками банка. Стоит отметить, что в данном случае действие является ненамеренным и совершается по причине халатности сотрудника, а не по злому умыслу.

1.7 Выводы по первой главе

В первой главе были подробно рассмотрены сами проблемы, решение которых предложено в данной выпускной квалификационной работе. Были рассмотрены основные способы мошенничества и экономический урон от них, а также особенности совершения. Далее рассматривались различные методы

борьбы с мошенничеством, которые описаны в литературе. В частности, были рассмотрены общие рекомендации, без которых не сможет функционировать ни одна, предложенная система. Также были рассмотрены экономические методы, методы на основе методики Базель II и методы на основе машинного обучения.

Было показано, что поиск мошенников является не просто задачей классификации, которую можно решить простым алгоритмом, но и напрямую связан с общей задачей кредитного риска. Что дает возможность не просто строить систему, которая не пропустит ни одного мошенника, но и будет выдавать ложноположительный результат, а строить систему, которая не будет показывать ложно положительных результатов, но у нее будут ложноотрицательные. Такой риск можно принять на себя, благодаря технологии формирования портфеля, в который будет заложены потери от мошенничества.

Были рассмотрены хорошо интерпретируемые модели, такие как Z-счет, что позволит не только выявить злоумышленника, но и одновременно предсказывать ухудшение состояния у «хорошего» заемщика, который допустил ошибку. Данный подход позволит иметь одну систему для обеих задач за счет чего сократит затраты, но не позволит помочь «хорошему» клиенту исправить ситуацию.

На сегодня описаны самые разные подходы по затратам и качеству, сложности реализации и принять решение об эффективности и необходимости внедрения той или иной системы должен сам кредитор. Далее будет предложена методология построения высокоэффективной системы определения мошенников.

ГЛАВА 2. РАЗРАБОТКА МЕТОДИКИ ПРОЕКТИРОВАНИЯ ИТ-ИНФРАСТРУКТУРЫ СИСТЕМЫ ВЫЯВЛЕНИЯ МОШЕННИЧЕСТВА С БАНКОВСКИМИ КРЕДИТАМИ СРЕДИ ЮРЕДИЧЕСКИХ ЛИЦ

2.1 Выбор источника данных для нахождения мошенников среди юридических лиц

В данной работе предлагается строить алгоритмы поиска мошенников, в максимально возможной степени основываясь на внутренних данных банка, затем основываясь на данных сторонних организаций, и уже в последнюю очередь основывая свои заключения на данных, которые предоставил сам клиент. Такой подход позволит создать максимально защищенную от внешнего вмешательства систему.

В качестве внутренних данных банка ключевыми данными стали транзакции клиента и информация об изменение его счетов. Такой подход поможет выявить контрагентов клиента и ранжировать их по важности. Он поможет найти скрытые связи между различными клиентами банка и другими компаниями, которые формально не связаны.

В качестве данных о клиенте из сторонних источников выступили ФНС, НБКИ, ЕГРЮЛ, ЕГРИП, ЕГРН

1) НБКИ [34]. Данные оттуда позволят узнать кредитную историю клиента и его текущее кредитное состояние, что является более интересным, чем его история. Если информация о кредитной истории уже используется в кредитном скоринге, то бессмысленно дублировать ее анализ в других системах. С другой стороны, зная обороты клиента и данные отчетности МСФО [33] и РСБУ [52], количестве кредитов на заемщике сейчас можно судить о дальнейших возможностях его платежеспособности.
2) ФНС [50]. Налоговая служба может передавать банкам отчетность клиента, которую тот предоставил. При этом данная отчетность уже будет

проверена ФНС, что уменьшает риск получения неправильной информации о клиенте.

3) ЕГРЮЛ [17] и ЕГРИП [14]. Может предоставить информацию о юридическом устройстве компании в том числе и исторические, информацию о владельце компании. Особенно интересно на основе данной информации строить связь заемщика с другими компаниями, например, через владельца.
4) ЕГРН [15]. Здесь можно получать информацию о различном недвижимом имуществе, которое принадлежит заемщику, на основе которой можно принимать решение о риске выдачи кредита, даже если вероятность дефолта достаточно высока, то при большом количестве имущества можно будет успешно вернуть деньги в банкротстве. Данный источник интересен и на стадии мониторинга, для оценки вероятности мошенничества: если заемщик стал продавать свое имущество, но это не сильно отражается в финансовой отчетности.
5) Спарк-Интерфакс [46]. Система, которая выступает как агрегатор нескольких предыдущих источников, при этом уже проведя некоторый анализ, но это будет рассматриваться далее в данной работе при выборе внешних источников для системы. Помимо функций агрегатора, данная система предоставляет различную оценку риска и состояния хозяйствующих субъектов, построенную на собственных алгоритмах.
6) ЕФРСБ [16]. Предоставляет данные по делам о банкротстве и заявления на начало процедуры банкротства.
7) Картотека арбитражных дел [25]. Это источник данных обо всех арбитражных делах внутри страны, что представляет собой большой вес в модели Аргенти.
8) Новостные агрегаторы и издания. Иногда работник банка может даже не предполагать о реальном состояние дел заемщика из-за сокрытия информации, при этом об этом может узнать журналист и выпустить об

этом статью. Из новостей можно узнать о происходящих проблемах отросли, в регионе, где действует заемщик, о скандалах, связанных с заемщиком.

9) Социальные сети. При помощи социальных сетей можно составить портрет глав компаний, узнать инсайды о внутренних делах компании, так как работники не всегда добросовестно сохраняют коммерческую тайну.

Основными показателями, которые банк получает от заемщика является финансовая отчетность: МСФО, РСБУ, Оборотно-сальдовая ведомость или другая бухгалтерская отчетность, которую предоставляет клиент.

Дополнительно, в зависимости от подписанного договора, заемщик может предоставлять: отчетность из различных систем ведения документации (1С, SAP), копии договоров с контрагентами.

Далее в данной работе будут использоваться транзакции клиента как пример внутренних данных банка, данные из ЕГРЮЛ и ЕГРИП как пример информации из внешних, независимых источников и отчетность РСБУ как пример данных, предоставляемых самим заемщиком.

С точки зрения информации о залогах, клиент обязуется предоставлять информацию о них, также, в зависимости от заключенного договора. В том числе, он может предоставлять отчет о состояние залога, может предоставлять фото и видео доказательства целостности имущества, предоставлять документы, которые подтверждают владение этим имуществом или принимать проверки эксперта оценщика от банка.

Максимизация безопасности источников данных должна стать первым приоритетом при построении систем поиска мошенников. Но не стоит отказываться совсем от анализа «ненадежных» данных. В них можно выявить закономерности, которые можно использовать для поиска мошенников.

Выводы. Банк обладает большим количеством независимых источников данных, которые можно использовать. Достаточно большая часть этих источников являются открытыми, что резко уменьшает затраты на разработку различных систем.

Первая критика источника должна быть направлена на его достоверность и правильность данных в нем, как правильно ранжировать источники по данного критерию и конкретные примеры были приведены в данной главе.

Стоит сделать вывод, что если кредитная организация по каким-то причинам не хранит историю внутренних данных о пользователе, то ей необходимо начать это делать, так как именно они являются самым ценным и достоверным источником информации.

2.2 Процесс построения алгоритмов для нахождения мошенников среди юридических лиц

В данном параграфе будет предложено три различных алгоритма поиска мошенников с банковскими кредитами среди юридических лиц:

1) Алгоритм выявления мошенничества на основе назначения платежа в транзакции.
2) Алгоритм выявления мошенничества на основе отчетности РСБУ.
3) Алгоритм графового анализа связи клиента для выявления мошенничества.

Алгоритм выявления мошенничества на основе назначения платежа в транзакциях использует внутренние данные банка, а, следовательно, является самым надежным, с точки зрения используемых данных. В его основу легли идеи поиска мошенников на основе определения аномалий в данных.

Как объект для анализа были выбраны именно назначения платежа, так как сумма и период проведения транзакций могут колебаться, что характерно для

закупки расходных материалов, например, канцелярии или при появление нового контрагента, который заместил старого поставщика или покупателя. При этом, в приведенных примерах не изменится семантика назначение платежа. В следствии чего, наиболее устойчивым к статистическим ошибкам первого рода будет именно анализ назначения платежа.

Данные алгоритмы строятся на основе двух базовых видах нейронной сети: RNN (recurrent neural network), которая является общепринятым инструментом для анализа текста и Autoencoder, вид нейронной сети, которая на сегодня активно используется для нахождения аномалий в данных.

Алгоритм на основе назначений платежа возможно применить только на этапе мониторинга клиента или принятия решения по клиенту в рамках процесса урегулирования проблемной задолженности, так как на момент принятия решения о выдачи кредита у банка не будет достаточного количества транзакций клиента для анализа. Данный алгоритм хорошо подходит для выявления преднамеренного банкротства или мошенничества с кредитами, так как сможет показать не типичные для данного заемщика сделки, при помощи которых он может совершать преступление, как показано в параграфе один первой главы.

Алгоритм выявления мошенничества на основе анализа РСБУ строится на основе анализа изменения отчетности клиента во времени. Данный алгоритм строится для выявления мошенничества с кредитами, преднамеренного банкротства и незаконного получения кредита. Это обосновывается тем, что данный алгоритм выявляет основной способ мошенничества заемщика, то есть подделку финансовой отчетности.

Алгоритм базируется на исследованиях Э. Альтмана и его модели Z-счет. Происходит анализ изменения во времени основных показателей, выделенных Э. Альтманом, и использует гипотезу, подтвержденную наблюдениями о том, что мошенники подделывают свою отчетность и приводят ее значение к среднему по отрасли, что характеризуется продолжительным отсутствием тренда в изменение параметра. Для анализа наличия тренда используются методы работы

с временными рядами, такие как: поиск автокорреляции, разложение временного ряда на составляющие и оценка наличия тренда методом Фостера-Стюарта.

Данный алгоритм подходит для анализа клиента как на стадии мониторинга и принятия решения об урегулирование проблемного актива, так и на стадии принятии решения о выдачи кредита.

Алгоритм графового анализа связи клиента для выявления мошенничества использует сторонние данные о компании из ЕГРЮЛ и ЕГРИП. Данный алгоритм, как и предыдущий может использоваться на любой анализ клиента, так как в нем используется информация о клиенте, которую юридическое лицо или индивидуальный предприниматель подает своем создании.

Идея данного алгоритма основывается на подходе к поиску мошенников на различных торговых онлайн площадках и использует марковские сети для классификации клиента. Общая идея данного алгоритма состоит в классификации клиента на основе его связей с другими компаниями. Такой подход обоснован приведенной в первой главе спецификой совершения мошенничества: чаще всего как пособники выступают связанные компании, что понижает для мошенника риск быть обманутым своим пособниками.

Для правильного понимания работы алгоритма, стоит уточнить, что для алгоритмов все виды мошенничества будут переведены в один общий класс «мошенник». Как следствие будет решаться задача не многоклассовой классификации, а задача бинарной классификации. Такой подход обосновывается двумя факторами. Первый, близостью способов мошенничества в случаях преднамеренного банкротства, мошенничества в сфере кредитования и незаконного получения кредита, это делает данные правонарушения трудноразделимыми без глубокого анализа причин и последствий правонарушения. Вторым фактором является факт несбалансированной выборки. В подобных случаях принято применять подход, называемый «один против всех». При этом подходе все классы выборки кроме одного объединяются в один, и задача сводится к бинарной классифика-

ции. После того, как была решена новая задача классификации, сгруппированный класс разбивается обратно и опять необходимо решить задачу мультиклассовой классификации, но в выборке на один класс меньше.

Еще одним аргументом такого подхода является следующее: в случае выявления любого из приведенных видов мошенничества наиболее эффективным для банка решением будет или не выдавать кредит или начать процедуру банкротства.

Так как алгоритм классификации видов мошенничества между собой на момент написания данной работы находится на стадии разработки, он не был включен в данную работу. Но несмотря на это, как будет показано далее приведенные в данной главе, алгоритмы показывают высокую эффективность в поиске мошенников. Для выявления конкретного способа совершения мошенничества сотруднику необходимо провести более детальный, самостоятельный анализ.

Так как далее совместно с описанием работы алгоритма будут приведены машинные эксперименты, целью которых является установить качество работы алгоритмов, необходимо указать основные правила при создании набора данных:

1) Для каждого эксперимента будет собираться набор данных из: признанных мошенников, клиентов с высоким кредитным рейтингом, клиенты без признаков мошенничества, по которым вышил на просрочку в 90 дней или была начата процедура банкротства.
2) Данные собираются из временного промежутка с 01.06.2014 по 01.01.2020.
3) При создании выборки из мошенников, берутся все три вида мошенничества в равных долях.
4) Компании, принадлежащие сегментам микро и малый бизнес, и средний и крупный бизнес берутся в пропорции 80% к 20% соответственно.
5) Данные по клиенту мошеннику собираются в состояние на самый ранний известный период мошенничества.

6) В случае, если по данным клиента, собранным на выбранный период невозможно рассчитать алгоритм, то данный клиент игнорируется.
7) Для клиента с хорошим кредитным рейтингом выбираются данные в состояние на случайный момент из выбранного периода.
8) Для клиентов, которые вышли в дефолт или по которым началась процедура банкротства, данные собираются на момент выхода на просрочку или банкротства.

2.2.1 Алгоритм выявления мошенничества на основе назначения платежа в транзакции

Данный алгоритм основывается на поиске аномалий в данных, а не задачу классификации с расширением выборки.

Трудно выделить несколько паттернов мошенничества, но при этом легко выделить паттерны поведения клиентов. Обычно заемщик имеет ограниченного количество контрагентов, на основание договоров, с которыми совершает денежные переводы. При этом, чаще всего, клиент совершает переводы за одни и те же услуги и товары.

В связи с чем, в данной работе предлагается реализовать алгоритм выявления подозрительных клиентов на основе назначения платежа. Логично предположить гипотезу о том, что должны совпадать суммы и период времени. Данная гипотеза имеет ряд недостатков, в действительности, заемщик может покупать различные расходные материалы по необходимости, это касается и услуг. Также изменение цены и периода после кредита могут быть связаны с желанием заемщика увеличить обороты, расширить бизнес.

Необходимо построить алгоритм, который будет выявлять аномалии в транзакциях заемщика. При этом, если была найдена одна или две аномалии, то вероятность, того, что это ошибка или появился новый контрагент высока. Следовательно, необходимо наложить на данный алгоритм дополнительное условие,

которое уменьшит ошибку второго рода. Данное ограничение необходимо по двум причинам:

1) Репутационный риск. При ложном определение как мошенника «хорошего» клиента может возникнуть ухудшение отношений с ним и потерю репутации у других клиентов.
2) Финансовый риск. При отказе в выдачи кредита или принятия решения о банкротстве, когда можно было выдать реструктуризацию, приведет к финансовым потерям.
3) Риск нецелесообразных трудозатрат. Сотрудник, который будет реагировать на ложное сообщение о мошенничестве, не сможет потратить это время на другие задачи.

С другой стороны, необходимо минимизировать количество пропущенных мошенников, то есть ошибку первого рода. В терминах машинного обучения, ошибка первого рода называется «False positive», а ошибка второго рода «False Negative». Но само количество таких ошибок не дает точного понимания о качестве, для этого используются такие показатели как точность (precision) – показывает насколько много ошибок первого рода совершается, полнота (recall) – показывает насколько много совершается ошибок второго рода. Для удобства, обычно, используют показатель F-мера, который объединяет в себе precision и recall. Все эти метрики принимают значение из промежутка [0, 1]. Чем выше значение, тем лучше работает классификация. Несмотря на то, что при построении классификаторов для финансовых задач используют метрику Gini, так как ее значение можно интерпретировать в доход, в данной работе как основная метрика будет использоваться F-мера, так как оно лучше отражает качество и хорошо подходит для задач с несбалансированными классами, в отличие от Gini [7].

Для поиска аномалий в назначениях платежей, необходимо сначала привести тексты к форме, которая будут удобна для чтения компьютером. Необходимо закодировать неким числом, а текст представить в виде вектора или матрицы. Самым простым подходом к данной задачи является Bag-of-words и TF-IDF [38].

Они являются простыми в реализации и их результатом является вектор, каждый элемент которого является словом, а значение, в случае Bag-of-words - входит ли данное слово в текст или нет, в случае TF-IDF вероятностью его появления в тексте. Данные подходы просты в реализации, но у них есть ряд недостатков:

1) Вектор для каждого текста имеет длину равную количеству слов во всем корпусе текстов обучающей выборки.
2) Вектор является крайне разряженным, то есть содержит большое количество нулей.
3) Теряется связь слов в предложение.

В связи с этими минусами в данной работе будет использоваться более современный подход: кодирование слов при помощи Word2Vec [24], а точнее одна из новейших реализаций – Negative Sampling [67]. Данный подход заключается в замене слова в предложении на соответствующий ему вектор из 255-мерного пространства. Данный вектор строится на основе слов соседей, при этом вычисляется вероятность данного контекста слова на каждом примере. Сам подход при обучении модели Word2Vec основан на минимизации дивергенции Кульбака-Лейблера, для функции восстановления вероятности того, что данное слово имеет данный контекст, под контекстом подразумевается k слов слева и справа от данного в тексте. При использовании Negative Sampling, одновременно происходит оптимизация и для функции вероятности того, что данное слово не встречается в данном контексте. Сама функция вероятности для Negative Sampling представлена в формуле 2.2.1

$$NegS(w_0) = \sum_{i=1,x_i\sim D}^{i=k} -\log\left(1 + e^{s(x_i,w_0)}\right) + \sum_{i=1,x_j\sim D'}^{j=k} -\log\left(1 + e^{-s(x_j,w_0)}\right) \quad 2.2.1$$

Где w_0- исследуемое слово, x_i-i слово из контекста, $s(x_i, w_0)$ – сопоставляющая двум векторам одно число, D - это распределение совместной встречаемости слова w_0 и остальных слов корпуса, D'- равномерное распределение слов в корпусе.

Данный подход устраняет все недостатки методов Word2Vec и TF-IDF, кроме необходимости в ограничение длины вектора и при этом задачи с текстом, которые используют данный подход показывают резкое улучшение качества. У данного подхода есть свой недостаток – необходимо сначала обучить словарь, а уже потом уже кодировать им текст, при это для обучения требуется большой массив данных. Есть два решения данной проблемы:

1) В открытом доступе содержится большое количество уже обученных словарей на различных языках.
2) Банк, которых хранит историю всех транзакций может обучить свой корпус на нем.

Проблема ограничения по длине текста остается актуальной при любом типе кодирования, так как она связана с жесткой фиксацией размера входного вектора в классические модели машинного обучения. Один из лучших подходов, который позволит потерять наименьшее количество информации состоит в выборе длины входного вектора в модель, а затем заполнить его значение усреднёнными значением n стоящих рядом слов. Таким образом будет получен вектор, состоящий из усредненных значений слов, что не отразиться на качестве модели, так как эти слова имеют похожие значения векторов по построению. Значение n выбирается экспериментально исходя из задачи оптимизации финальной функции качества.

При кодирование текстовой информации принято разделять текст на предложения и формировать не одни вектор, а матрицу, где в столбцах будет находиться слова, а в строках предложения текста. В данном случае имеет смысл построить матрицу, где каждому столбцу также соответствует слово, а каждой строке соответствует одно из значений вектора в 255 мерном пространстве. Соответственно, в качестве входного элемента для обучения модели будет трехмерный тензор, а при использовании матрица.

Для того, чтобы перевести каждое слово в некоторый вектор, сначала необходимо проделать несколько подготовительных шагов. Данные шаги являются стандартными для работы с текстами на ЭВМ:

1) Необходимо провести разбиение текста на отдельные лексемы, такой процесс называется токенизация. Необходимо разбить текст на слова или n-граммы. Так как метод Word2Vec подразумевает работу с каждым отдельным словом, токенизация должна проходить по словам.
2) Необходимо удалить стоп-слова. Данная процедура подразумевает удаления слов, которые очень часто встречаются в тексте и не несут особой смысловой нагрузки. Такими словами могут выступать союзы и междометия. Их необходимо удалить, так как они являются шумовыми и могут значительно сказаться на результате.
3) Проведение лемматизации Данный шаг не является обязательным и зависит от объема словаря при обучении Word2Vec, а при использовании уже готового словаря необходимо провести проверку обучался ли он на лемматизированной выборке. Сама процедура лемматизации подразумевает приведение слов к их начальной форма.

После получения закодированного текста необходимо построить модель машинного обучения для нахождения аномалий в тексте. Классическим подходом к работе с текстами является использование нейронной сети типа Рекуррентная нейронная сеть RNN [11], а для задачи поиска аномалий применение нейронной сети типа Autoencoder [11] показывает наилучшие результаты. Следовательно, для решения поставленной задачи необходимо спроектировать RNN-Autoencoder.

RNN является нейронной сетью с обратной связью, то есть из каждого нейрона после вычисления результата он поступает на вход следующему нейрону и обратно в сам нейрон и участвует в вычислении значения для следующего поступившего в него значения. То есть, для каждая строка в матрицы бу-

дет анализироваться не как отдельное значение, а будет участвовать в вычислениях на предыдущем, что пытается повторить алгоритм, по которому человек читает тексты.

Autoencoder – является нейронной сетью, которая обучается без учителя, ее основная задача предсказать набора данных, который пришел ей на вход. При поиске аномалий используется разница в значениях, которые поступили на вход и были получены на выходе, чем выше ошибка, тем больше данная запись не похожа на остальные.

Общая архитектура данной нейронной сети формулируется так: первый слой является RNN, затем идут линейные слои, построенные по архитектуре Autoencoder, затем идет финальный декодирующий слой RNN. На рисунке 12 представлена архитектура слоев RNN, а линейные слои скрыты. Задача, которые решаются нейронные сити с похожими архитектурами называется «Последовательность в Последовательность» (Seq-to-Seq).

S_i – это i элемент вектора, которым закодировано слово в данных, а $\hat{S_i}$ - элемент вектора слова, который будет возвращен, вычисляется по формуле 2.2.2. Значением RNN сети кодировщика на предыдущем шаге является $h_i^{(E)}$, которые будут вычисляться по формуле 2.2.3, а $h_i^{(D)}$ – предыдущие значение сети RNN декодировщика, рассчитывается по формуле 2.2.4, T – это последовательность слов, а $\hat{T}$ - предсказанная последовательность слов в обратном порядке [65].

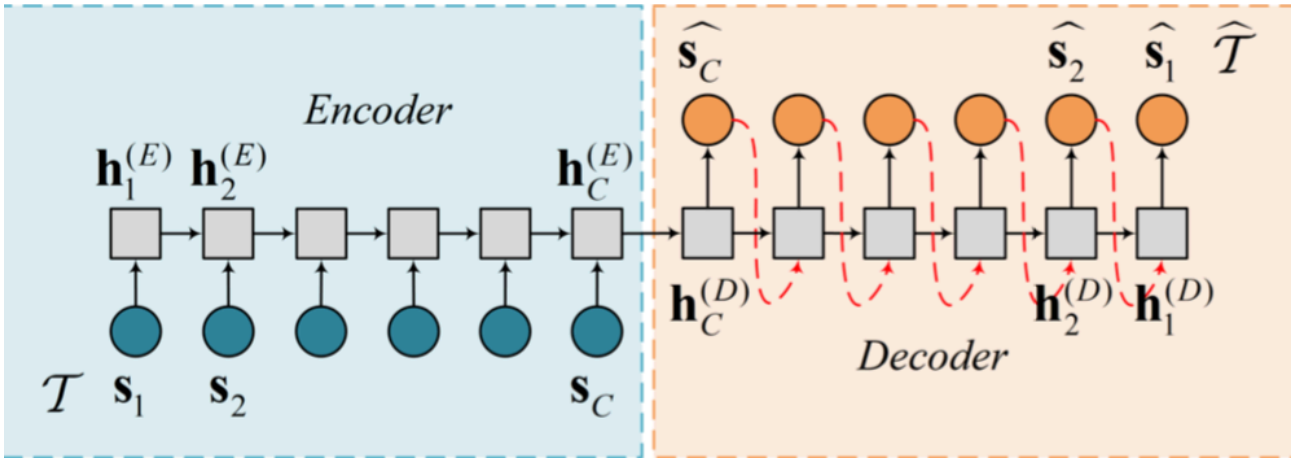

Рисунок 12. Seq-to-Sec RNN Autoencoder [65]

$$h_i^{(E)} = f(S_i, h_{i-1}^{(E)}) \quad 2.2.2$$

$$\hat{S_i} = g(S_{i+1}, h_{i+1}^{(E)}) \quad 2.2.3$$

$$h_i^{(D)} = f(S_i, h_{i+1}^{(D)}) \quad 2.2.4$$

Где f- это некоторая нелинейная функция, которая может быть sigmoid, тангенсом или одной из более сложных реализаций RNN нейронов Long-Short Term Memory (LSTM) или Gated Recurrent Unit (GRU). Принципы работы данных реализаций выходят за рамки данной работы. И g так же нелинейная функция, которая обычно является тангенсом, sigmoid или relu.

Следующим шагом, который необходимо выполнить является внесение регуляризации по средствам случайного выключения связи между нейронами (Dropout). Данный метод показал высокую эффективность при разработке нейронных сетей, он позволяет увеличить качество модели и предотвратить переобучение.

Если для Autoencoder функционал, который позволяет реализовать Dropout есть в любой библиотеке для глубокого обучения, то реализация данного функционала для RNN сетей является не тривиальной задачей. Необходимо не просто прервать связи между двумя нейронами, что так же имеет множество реализаций, но также необходимо случайно выключать и рекуррентную связь нейрона.

Для решения этой задачи необходимо ввести коэффициент разрежённости, который является вектором $w_t = (w_t^f, w_t^{f'})$ определяющим какое соединение должно быть прервано на шаге t. При этом $w_t^f \in \{0,1\}$ и $w_t^{f'} \in \{0,1\}$, а на w_tнакладывается ограничение $||w_t|| > 0$. На основание данного коэффициента необходимо построить новую модель вычисления $h_i^{(E)}$, где вычисления каждого значения будут происходить в соответствие с формулой 2.2.5.

$$h_t = \frac{f\left(S_i, h_{i-1}^{(E)}\right) w_t^f + f'\left(S_i, h_{i-L}^{(E)}\right) w_t^{f'}}{\|w_t\|} \quad 2.2.5$$

Здесь f и f' могут быть как различными нелинейными функциями, так и одно нелинейной функцией. $\|w_t\|$ – в качестве нормы используется l_1. Эффективность такого подхода к построению не только подтверждается математически и экспериментально, но и логически, он позволяет учитывать в понимание текста связку из не стоящих рядом слов. Такая ситуация часто встречается в русском языке, когда логически связанные слова находятся в разных частях предложения.

Следующем шагом, который рекомендуется выполнить в рамках данной работы, является построение ансамбля RNN-Autoencoder, общая архитектура такой системы представлена на рисунке 13. Применение ансамбля нейросетей не является обязательным, но его использование повысит качество работы всей системы, хотя и увеличит количество необходимых вычислительных мощностей.

В данном случае ансамбль содержит N нейронных сетей, при этом E_i- i кодировщик, D_i – i декодировщик, $1 \leq i \leq N$. Обычно ансамбли в моделях строятся отдельна, а затем продолжают оптимизироваться на основе некоторой объединяющей функции. В данном случае предлагается исходить из идеи, что если все модели ансамбля получают один и тот же набор данных и должны получить один и тот же набор данных на выходе, то стоит их обучать одновременно [68].

Для реализации такого подхода необходимо ввести слой $h_c^{(E)}$, который является линейной комбинацией всех выходов скрытых слоев кодировщика, что представлено в формуле 2.2.6:

$$h_c^{(E)} = concatinate(h_c^{(E_i)} * W^{(E_i)}, \dots, h_c^{(E_c)} * W^{(E_c)}) \qquad 2.2.6$$

Каждый декодер D_i связан с $h_c^{(E)}$ как с инициализатором скрытого состояния при реконструкции назначения платежа. В данной архитектуре все нейронные сети тренируются совместно для минимизации функции Δ, являющуюся суммой ошибок восстановления каждой нейронной сети и L1 регуляризатора для скрытых слоев нейронной сети, формула функции оптимизации 2.2.7:

$$\Delta = \sum_{i=1}^{N} \Delta_i + \lambda \left\| h_c^{(E)} \right\|_1 = \sum_{i=1}^{N} \sum_{t=1}^{C} \left\| s_t - \hat{s}_t^{(D_i)} \right\|_2^2 + \lambda \left\| h_c^{(E)} \right\|_1 \qquad 2.2.7$$

Здесь λ – это вес, который силу регуляризатора, а сам регуляризатор отвечает за отбор наиболее эффективных связей между скрытыми слоями. Это позволяет избежать ситуации, когда некоторые кодировщики переобучаются на исходных данных и помогают стать декодировщикам более устойчивыми к выбросам. Следовательно, когда RNN-Autoencoder встретит выбросы разница исходной последовательности и полученной станет более выраженной.

Далее, следуя принципам кодировщиков для данных, не представленных последовательностями необходимо вычислить ошибку предсказания последовательности ансамблем RNN-Autoencoder. На данном этапе мы имеем N RNN-Autoencoder каждый из которых пытается восстановить назначение платежа $T = (s_1, s_2, \dots, s_c)$ и соответственно N восстановленных назначений платежа $\hat{T}^{(i)} = (\hat{s}_c^{(i)} \dots, \hat{s}_2^{(i)}, \hat{s}_1^{(i)})$. Для каждого вектора s_k необходимо вычислить N ошибок восстановления этого вектора по формуле 2.2.8, а затем рассчитать обобщенную ошибку на данном примере. В качестве функции обобщения можно использовать среднее или медианное значение, что показано в функции 2.2.9. Хотя можно использовать и среднее значение, рекомендуется использовать медиану, так как это поможет избежать внесения в результат ошибки от RNN-Autoencoder, которые переобучились [67].

Следующим шагом после построения нейронной сети мы получаем значение ошибки на каждом примере, теперь необходимо найти пороговое Θ значение такой что $OS(s_k) > \Theta$, то данный пример является выбросом. Метод нахождения Θ вычислительно крайне прост. Необходимо взять некоторый тестовый набор данных, который будет содержать «хорошие» назначения платежа и мошеннические с распределение сходным с реальным распределением таких назначений платежа, при этом весь тестовый набор данных не должен участвовать в обучение нейронной сети. Затем необходимо вычислить ошибку предсказания на тестовом наборе и начать изменять значение Θ, от самой маленького значения ошибки до самого большого каждый раз вычисляя F-меру. После работы данного

алгоритма необходимо выбрать то значение Θ при котором F-мера будет наилучшей.

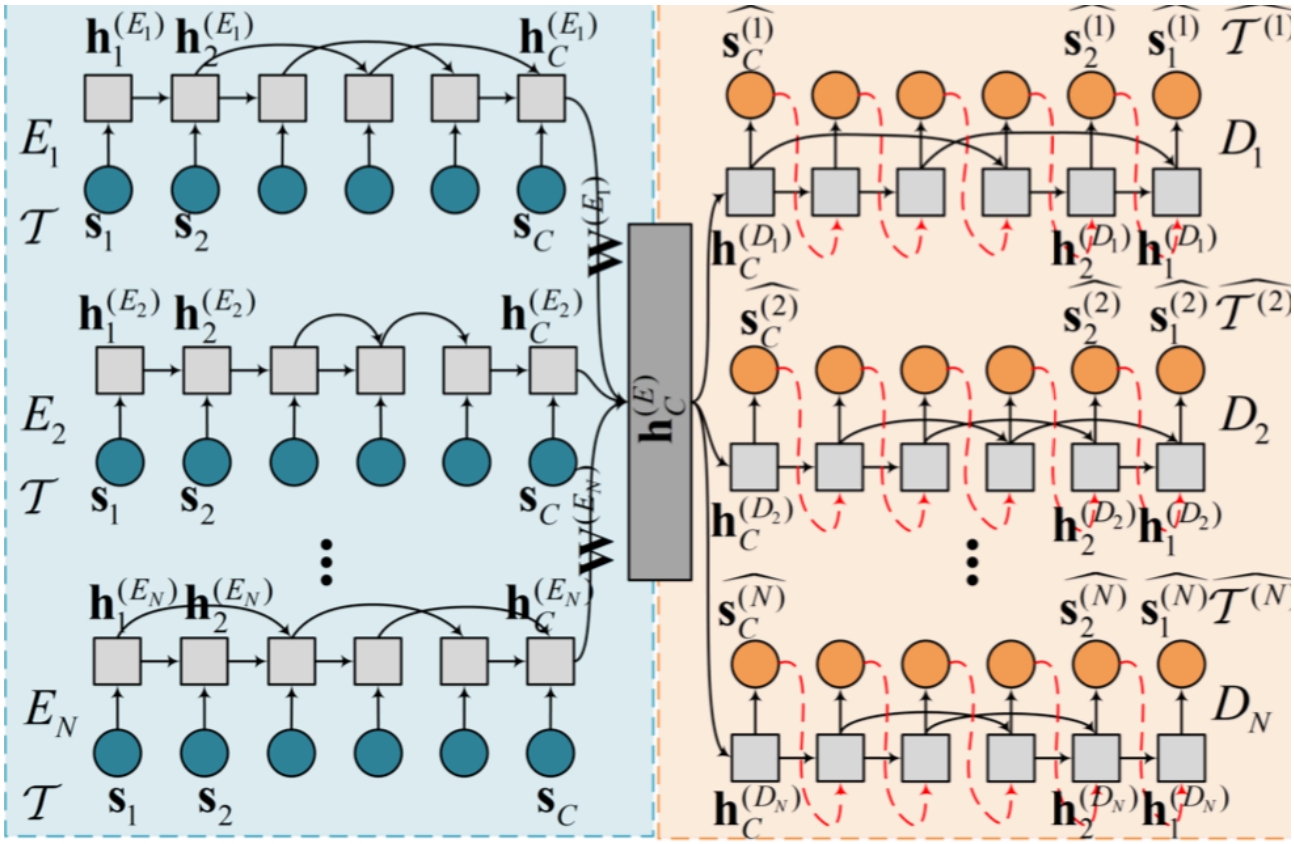

Рисунок 13. Ансамбль RNN-Autoencoder [65]

$$\left\{\left\|s_k-\hat{s}_k^{(1)}\right\|_2^2,\left\|s_k-\hat{s}_k^{(2)}\right\|_2^2,\ldots,\left\|s_k-\hat{s}_k^{(N)}\right\|_2^2\right\} \quad 2.2.8$$

$$OS(s_k)=median\left(\left\|s_k-\hat{s}_k^{(1)}\right\|_2^2,\left\|s_k-\hat{s}_k^{(2)}\right\|_2^2,\ldots,\left\|s_k-\hat{s}_k^{(N)}\right\|_2^2\right) \quad 2.2.9$$

Финальным шагом данного алгоритма является классификация конкретного заемщика как мошенника. Для этого необходимо повторить алгоритм для классификации транзакции как мошеннической, только в качестве значения взять не набор значений для поиска порогового значения Ω, а количество транзакций, помеченных как мошеннические.

После построения ансамбля RNN-Autoencoder необходимо подобрать наилучшие гиперпараметры данной системы, в их перечень входит: значение веса при регуляризаторе λ, количество нейронных сетей в ансамбле N, количе-

ство слоев нейронной сети и размерность входной матрицы. Существует множество способов подбора гиперпараметров. Все способы сводятся к перебору различных множеств гиперпараметров и нахождению наилучшего такого множества с точки зрения качества. В качестве наиболее простых и популярных способов можно привести: «поиск по сетки» и «случайный поиск по сетке». Как более продвинутые способы используются: генетические алгоритмы и подбор параметров при помощи нейронной сети, они не сильно отличаются от простых способов по результирующему качеству, но способны существенно увеличить скорость нахождения оптимального набора.

Еще одним важным гиперпараметром модели является количество эпох при обучении. В данной задаче нет необходимости подбирать данный гиперпараметр, так как целью обучения является максимизация способности нейронной сети предсказывать обучающую выборку. Следовательно, можно выбрать некоторое большое число эпох, которое гарантированно приведет к переобучению нейронной сети. Можно попробовать подобрать данный параметр, но только с целью ускорения работы алгоритма.

В качестве общих рекомендация для поиска гиперпараметров стоит выделить следующие: чем меньше данных предоставлено для обучения тем больше должно значение N и λ, и тем меньше должно быть количество слоев нейронной сети. При этом обратное влияние данных на гиперпараметры нет.

Общий алгоритм построения word2vec, который был предложен в данном подпункте приведен на рисунке 14, стоит отметить, что данная схема строилась из предположения о том, что уже есть собранная выборка для обучения. Результатом данного алгоритма является наличие готового словаря word2vec, который можно использовать для обучения и применения алгоритма выявления мошенников на основе назначения платежа. Стоит отметить, что словарь word2vec предлагается обучать самостоятельно, так как большинство русскоязычных словарей word2vec обучены на НКРЯ и Wikipedia, что не позволяет им учитывать специфику семантической связи слов в назначениях платежа.

На рисунки 15 приведен алгоритм выявления мошенников, он так же исходит из предположения о том, что уже готовы выборки для тестирования и обучения. Результатом данного алгоритма является готовый ансамбль RNN-Autoencoder, который можно переводить в тестирование, а затем применять.

Для проверки качества алгоритма были собраны транзакции 1000 клиентов, у которых была выявлена просрочка платежа более 90 дней, 1000 клиентов с высоким кредитным рейтингом и 198 клиентов мошенников преступление которых было доказано в суде.

Для проверки качества алгоритма был выбран язык python, так как разработка на нем происходит быстрее, чем на большинстве других языков, которые поддерживают библиотеки для глубокого обучения. В качестве библиотеки для реализации алгоритма была выбрана TensorFlow. Такой выбор обосновывается предоставляемым библиотекой модулем Keras, который позволяет быстро строить нейронные сети из стандартных слоев, что отлично подходит для построения одного RNN-Autoencoder. Так и тем, что данная библиотека позволяет создавать не стандартные архитектуры нейронных сетей и даже собственные нейроны за счет API для эффективных математических вычислений, что отлично подходит для построения ансамбля нейронных сетей по предложенному алгоритму.

Для проверки качества алгоритма был выбран язык python, так как разработка на нем происходит быстрее, чем на большинстве других языков, которые поддерживают библиотеки для глубокого обучения. В качестве библиотеки для реализации алгоритма была выбрана TensorFlow. Такой выбор обосновывается предоставляемым библиотекой модулем Keras, который позволяет быстро строить нейронные сети из стандартных слоев, что отлично подходит для построения одного RNN-Autoencoder. Так и тем, что данная библиотека позволяет создавать не стандартные архитектуры нейронных сетей и даже собственные нейроны за счет API для эффективных математических вычислений, что отлично подходит для построения ансамбля нейронных сетей по предложенному алгоритму.

Так как данная система машинного обучения является обучением без учителя, то для нее невозможно использовать стандартные методы тестирования, например, кросс-валидация. При этом остается возможность проводить исследование типа Out-of-Time, которое заключается в разбиение выборки на тестовую и обучающую в соответствие со временем получения данных, что позволит определить устойчивость алгоритма во времени.

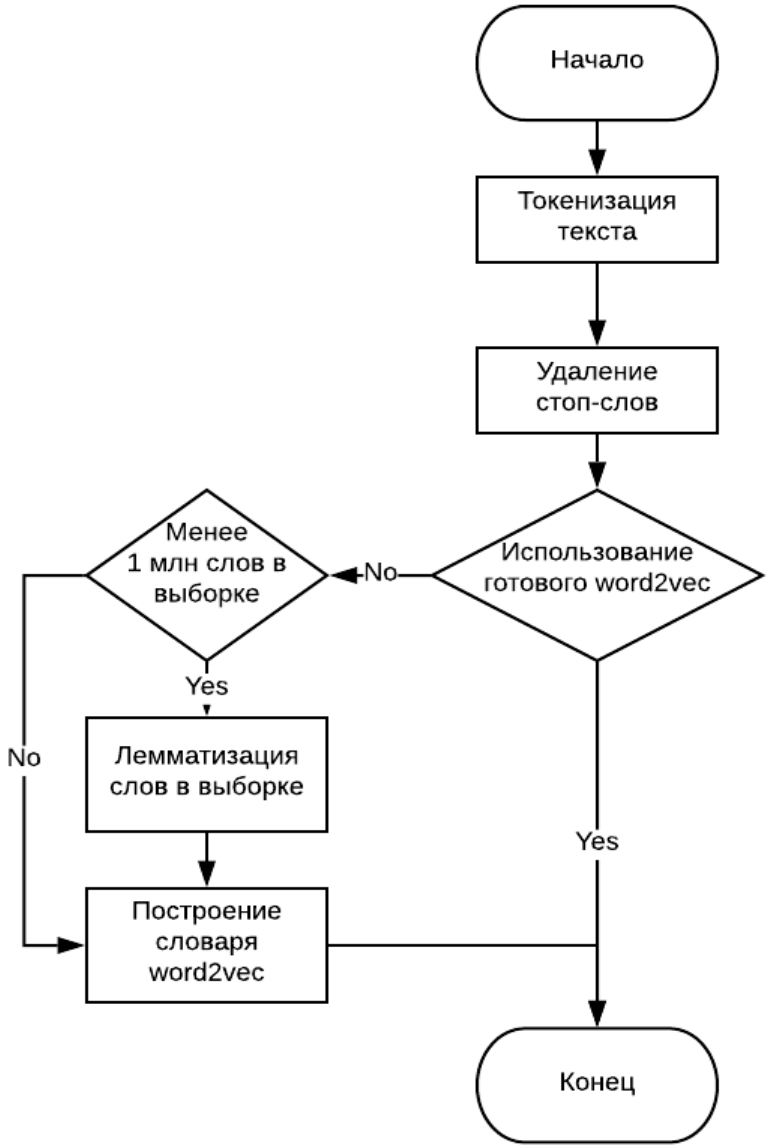

Рисунок 14. Алгоритм построения Word2Vec (Автор, creately.com)

Таким образом, на основе приведенного алгоритма был, сначала был обучен словарь Word2Vec, так же используя библиотеку TensorFlow.

Для тестирования применялся стандартный метод разбиения выборки на тестовую и обучающую случайно. Для каждой из них было проведено тестирование Out-of-Time.

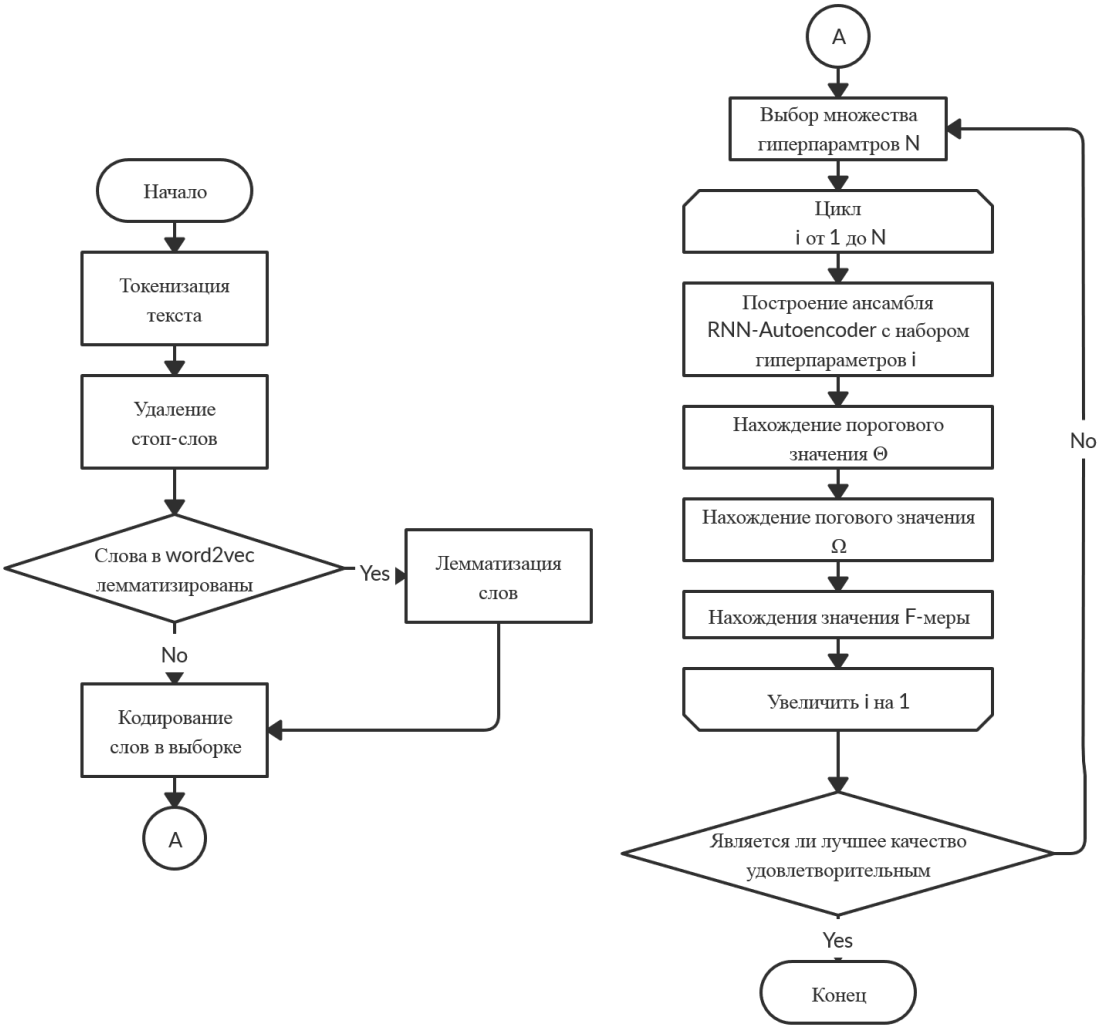

Рисунок 15. Алгоритм построения ансамбля RNN-Autoencoder (Автор, cre-ately.com)

Первым шагом были соединены в одну выборку 500 клиентов с просрочкой и 99 мошенников. Для каждого все его транзакции были разбиты на 5 равных временных промежутка и тестирование происходило следующим образом:

1) Ансамбль нейронных сетей обучается на k временных отрезках и применяется на k+1 отрезке для каждого клиента, где $k \in [1,4]$.
2) Рассчитывается $OS(s_k)$ для каждого объекта из выборки.

3) На основе выбранного значения Θ, каждая транзакция классифицируется как мошенническая или нет.
4) На основе выбранного значения Ω, происходит классификация клиента.
5) Рассчитывается значение F-меры на основе всех объектов выборки.

Далее рассчитывается среднее значение F-меры на основе результатов F-меры для каждого значения k. Такое среднее значение и буте является финальным качеством для тестирования Out-of-Time для данного набора гиперпараметров.

Данный алгоритм повторялся для каждого набора значений гиперпараметров из заданного множества значений, при этом использовался ранее описанный метод поиска по сетке. В рассматриваемые значения гиперпараметров приведены в таблице 7. И дополнительно в качестве гиперпараметров были использованы рассмотрены различные виды самих нейронов: GRU и LSTM.

Таблица 6

Значения гиперпараметров

Параметр	Минимальное значение	Максимальное значение	Шаг
Количество слоев	4	12	2
λ	0,01	1,5	0,02
Θ	0,01	1	0,01
Ω	1	1000	10

Стоит отметить, что возможные интервалы значения гиперпараметров выбирались экспертно. В случае, если оптимальным значением окажется максимальное или минимальное, то необходимо изменить значение интервала.

Вторым шагом является проверка качества на оставшихся 500 клиентов с просрочкой платежа и 99 мошенниках. При этом так же была применено тестирование методом Out-of-Time по описанному ранее методу. Полученное в результате тестирования значение F-меры и является финальной оценкой качества

алгоритма. В случае, если значение было бы слишком мало, то необходимо было отказаться от данной архитектуры модели, сами значения на тестовой и обучающей выборке приведены в таблице 7.

Последним шагом были протестированы 1000 клиентов с хорошей кредитной историей. Так как в этой выборке присутствуют только не мошенники, то есть только один класс, следовательно, невозможно рассчитать F-меру. Поэтому для этого оценки качества будет использоваться показатель Accuracy, которая показывает отношение правильно предсказанных объектов ко всем объектам. Тест с «хорошими» клиентами необходим для полной оценки качества модели. Он покажет, сколько добропорядочных клиентов было классифицировано как мошенники, если это число окажется не приемлемым, то необходимо отказаться от данного алгоритма. Результаты данного теста так же приведены в таблице 7.

Приведенный в данном подпункте алгоритм показывает высокое качество при тестировании, краткая сводка приведена в таблице 7, где $OS(s_k)_{good}$- средняя ошибка на «хороших» примерах, $OS(s_k)_{bad}$-средняя ошибка для мошенников, A - accuracy. Стоит отметить, что значение Accuracy не стоит рассчитывать для тестовой и обучающей выборки, так как из-за несбалансированных классов оно заведомо окажется близким к 1 или к 0.

Выводы. Был предложен алгоритм поиска мошенников на основе транзакций, а точнее, назначения платежа из транзакций. Использование именно назначения платежа обосновывается меньшей вероятность ложноположительной классификации по сравнению с использованием сумм и дат переводов. Данный алгоритм обладает высоким качеством предсказания, но при этом достаточно труден в построении и требует наличие у организации специалистов с навыками разработки программного обеспечения и хорошей математической подготовкой.

Был описан не только сам алгоритм, но и даны подробные пояснения по всем шагам, обоснование выбора конкретного инструмента и представлено общее описание работы данных инструментов с ссылками для более подробного изучения в случае необходимости.

Таблица 7

Качество алгоритма выявления мошенников на основе назначения платежа

	$OS(s_k)_{good}$	$OS(s_k)_{bad}$	Preci-sion	Recall	F-мера	Θ	Ω	A
Train	0.02	0.9	0.92	0.90	0.91	0.42	43	-
Test	0.3	0.8	0.82	0.76	0.79	-	-	-
Good	-	-	-	-	-	-	-	0,89

2.2.2 Алгоритм выявления мошенничества на основе отчетности РСБУ

Как уже упоминалось в параграфе 2.1, не стоит полностью игнорировать информацию, предоставленную клиентом при поиске мошенничества. Несмотря на то, что она может быть не достоверной, при правильном анализе данной отчетности можно выявить достаточно «странное» изменение показателей во времени.

Данный алгоритм основывается на идеях алгоритма Z-счета предложенного Альтмоном, который рассматривался в параграфе 1.3 данной работы. Следовательно, для анализа будут использоваться переменные схожие с переменными Z-счета:

1) X_1 - оборотный капитал.
2) X_2 – не распределенная прибыль.
3) X_3 - прибыль до налогообложения.
4) X_4 - рыночная стоимость собственного капитала, а для частных компаний собственный капитал.
5) X_5 - объем продаж.

При исследовании данных показателей выяснилось, что большинство мошенников, которые подделывают отчетность РСБУ делают это по одному и тому же алгоритму. Визуализация данного эффекта приведена на рисунках 16 и 17, на

рисунке 16 приведена типичное изменение данных показателей во времени заемщика, который не подделывает отчетность, а на рисунке 17 приведено типичное изменение показателей для клиента, который подделывает свою отчетность. Для удобства данные для графиков прошли нормализацию методом вычитания среднего значения и делением на дисперсию.

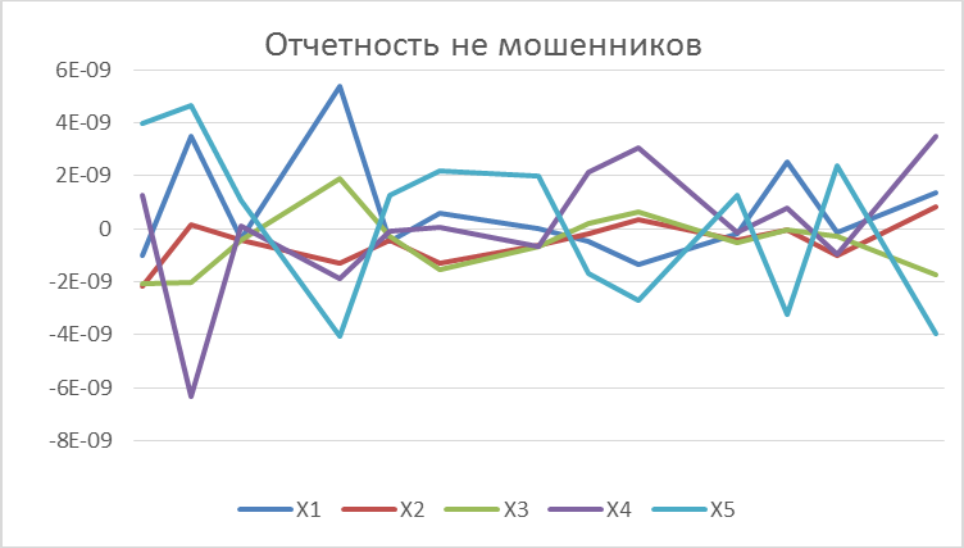

Рисунок 16. Поведение показателей у не мошенников (Автор, MS Excel)

Как можно увидеть из графиков, показатели отчетности у не мошенников имеют сильную амплитуду колебания относительно нуля на протяжение всего времени наблюдения. С другой стороны, можно выявить постепенное уменьшение амплитуды изменения показателей у не мошенников. То есть можно точно выявить точку, где происходит сглаживание кривой. Учитывая, тот факт, что такое поведение является типичным для обеих групп клиентов, то можно выполнить задачу поиска тренда во временном ряду для каждого показателя. У заемщиков, которые не производят мошеннической деятельности большую часть времени тренд будет возрастающим или убывающим, а у мошенников на протяжение определенного времени будет фиксироваться отсутствие тренда.

Сам временной ряд может состоять из четырех различных компонент:

1) Тренд u_t, который показывает общую тенденцию временного ряда.
2) Сезонная составляющая s_t, которая показывает характерные для сезона изменения в данных. Примером может служить туристическая отрасли,

где мы ожидаем, что на пляжные курорты у черного моря летом будет приезжать больше туристов, чем зимой. Следовательно, обороты в летнее время будут сильно больше, чем зимой. Период в сезонной не должен превышать 1 год.

3) Циклическая составляющая v_t, которая похожа на сезонную составляющую, но ее период превышает год, обычно они оцениваются в десятилетие.
4) Случайная составляющая e_t. Характеризуется как то, что остается во временном ряду, в случае, если убрать из временного ряда все составляющие. На самом деле она не является случайно, но зависит от некоторых неизвестных в момент анализа переменных.

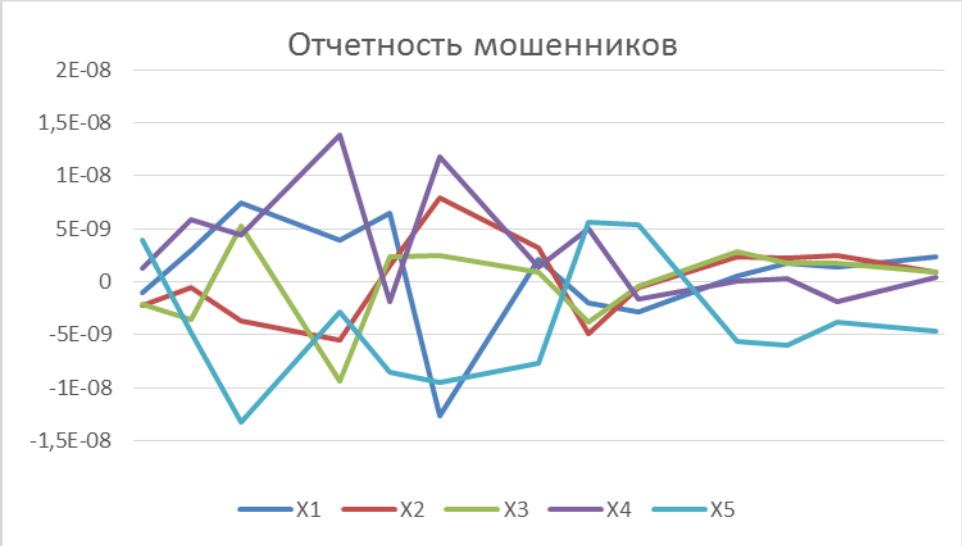

Рисунок 17. Поведение показателей у мошенников (Автор, MS Excel)

Для выявления тренда необходимо устранить сезонную составляющую и циклическую составляющую из данных. Убрать из временного ряда случайную составляющую невозможно, но учитывая характер задачи – это не требуется, в случае если клиент начинает подделывать свою отчетность РСБУ, то случайная составляющая исключается автоматически. Для устранения данной проблемы у клиентов, которые предоставляют реальную отчетность и из-за случайной со-

ставляющей у них наблюдается отсутствие тренда, необходимо выделить некоторое окно во временном промежутке внутри которого такое поведение является нормальным. Как видно на рисунке 16, даже у клиентов, которые не являются мошенниками может наблюдаться период отсутствия тренда. Такое окно необходимо для уменьшения ложноположительных результатов поиска мошенников.

Стоит также отметить, что временные ряды обычно делятся на три типа по признаку комбинации составляющих [47]:

1) Мультипликативные. $y_t = u_t * s_t * v_t * e_t$
2) Аддитивные. $y_t = u_t + s_t + v_t + e_t$
3) Смешанные $y_t = u_t * s_t * v_t + e_t$

Для применения большинства моделей, которые выделяют сезонность необходимо привести временной ряд к аддитивной форме, для этого необходимо прологарифмировать все значения временного ряда. Есть ряд тестов, которые позволят выяснить тип по комбинации составляющих, но их применение является избыточным в данной задаче, простое логарифмирование не приведет к потере информации в аддитивном ряду.

Для выделения сезонной составляющей предлагается использовать уравнение сначала необходимо определить ее наличие при помощи автокорреляционной функции. Пусть X_t – значение временного ряда в момент времени $0< t < n$, n – количество доступных измерений, рассчитывается среднее значение μ и дисперсия σ^2 по всем значениям временного ряда. Затем по формуле 2.2.10 рассчитывается значение функции автокорреляции для всего дискретного временного ряда [55].

$$\hat{R}(k) = \frac{1}{(n-k)} \sum_{t=1}^{n-k} [X_t - \mu][X_{t+k} - \mu] \qquad 2.2.10$$

Где k – это период в котором необходимо найти автокорреляцию. После того как было найдено окно необходимого размера необходимо избавится от ав-

токорреляции во временном ряду. Для этого необходимо найти коэффициент автокорреляции по формуле 2.2.11 для каждого объекта во временном ряду и скорректировать значение всех объектов по формулам 2.2.12 и 2.2.13.

$$a_t = \frac{x_t}{f_t} \quad 2.2.12$$

$$f_t = \frac{x_t}{a_{t-k}} \quad 2.2.13$$

Где f_t – коэффициент сезонной составляющей в момент времени t, а a_t новое значение временного ряда без сезонной компоненты.

После того была убрана сезонная составляющая можно применить один из методов выявления наличия тренда во временном ряду. Было разработано большое количество данных методов, в том числе: критерий квадратов последовательных разностей [29], метод проверки разностей средних уровней, критерий серий [32]. Все эти критерии отличаются друг от друга мощностью и математической сложностью. Но так как во временных рядах изменения отчетности мошенников наблюдается резкий переход к отсутствию тренда, то можно ограничится не очень сложным и в меру эффективным методом: критерием Фостера-Стюарта.

Данный алгоритм проверяет отрезок временного ряда длиной h, что является гиперпараметром данного метода и его необходимо подобрать, основываясь на метриках качества классификации. В качестве метрики предлагается использовать F-меру аналогично предыдущему подпункту.

На первом шаге алгоритма необходимо найти вспомогательные характеристики m_t и l_t, при этом $1 < t < h$, значение данных характеристик вычисляются по формулам 2.2.14 и 2.2.15.

$$m_t = \begin{cases} 1, \text{если } a_t > a_{t-1}, a_{t-2}, \dots, a_{t-h} \\ 0, \text{в противном случае} \end{cases} \quad 2.2.14$$

$$l_t = \begin{cases} 1, \text{если } a_t < a_{t-1}, a_{t-2}, \dots, a_{t-h} \\ 0, \text{в противном случае} \end{cases} \quad 2.2.15$$

Следующим шагом алгоритма является вычисление вспомогательного параметра $d_t = m_t - l_t$ для всех t. Откуда по формуле 2.2.16 вычисляется характеристика D.

$$D = \sum_{t=2}^{h} d_t \qquad 2.2.16$$

Далее при помощи критерия Стьюдента была проверена гипотезу о том, что выбранный временной ряд не содержит тренда. Перед этим была вычислена вспомогательный коэффициент σ_D, значение которого рассчитывается исходя из величины h по формуле 2.2.17. Для проверки гипотезы был вычислен критерий $t_{\text{набл}}$ по формуле 2.2.18 и сравнивается с пороговым значением критерия Стьюдента $t_{\text{кр}}$ со степенями свободы $v = h - 1$ и заданным уровнем значимости α, который принято брать равным 0,05. В случае, если, то отвергается гипотеза об отсутствие тренда.

$$\sigma_D = \sqrt{2\sum_{t=2}^{n} \frac{1}{t}} \approx \sqrt{\ln(n) - 0.8456} \qquad 2.2.17$$

$$t_{\text{набл}} = \frac{D}{\sigma_D} \qquad 2.2.18$$

Теперь, после необходимо проверить данный критерий на всех возможных окнах длинной h во временном ряду, если хотя бы у одного окна будет подтверждена гипотеза об отсутствие тренда для трех из пяти признаков X, то можно признать такого клиента мошенником.

Для тестирования алгоритма использовалась выборка, аналогичная выборке для алгоритма на основе назначения платежа в транзакции, но для половины половина всех мошенников была изменена на клиентов, у которых было выявлено мошенничество на стадии выдачи кредита. Несмотря на то, данный алгоритм является статистическим и его не требуется вычислять, необходимо подобрать значение параметра окна h.

Как и для тестирования предыдущего алгоритма, для тестирования этого использовался язык python и его библиотеки. Основными библиотеками стали: scikit learn, которая хотя и предназначена для машинного обучения, содержит в

себе мощные статистические инструменты и statsmodels, которая предоставляет реализацию популярных статистических тестов.

В данном случае, также необходимо использовать метод тестирования Out-of-Time, который является стандартным методом для временных рядов. Стоит уточнить, что отчетность РСБУ подается компанией раз в квартал, следовательно, один период измерения равен кварталу.

Для подбора гиперпараметра невозможно использовать метод Out-of-Time, как это было использовано для предыдущего алгоритма. Эта проблема связана непосредственно с методом выявления отсутствия тренда. В данном алгоритме поиск тренда ведется во временном окне, следовательно, если будет обрезан период, то больший вес начнут получать временные окна маленького размера.

Стоит отметить, что максимальный размер окна равный 5 был получен статистически, это минимальное количество кварталов, которые клиент из обучающей выборки подает отчетность РСБУ.

Для получения оптимального размера временного окна были сделаны расчеты для каждого клиента из обучающей выборки с одним из значений h, но для всех одинаковым. Затем была рассчитана F-мера для всей обучающей выборки. Данная последовательность была повторена для всех возможных значений h.

На основе эксперимента было выяснено, что оптимальным значение для h является 4 периода измерения, что демонстрирует график на рисунке 19. Под K на данном рисунке подразумевается процент клиентов, по которым началась процедура банкротства в квартале анализа. Такое значение позволяет с достаточной точность выявить мошенника, при этом успев среагировать на его действия. В случае если h будет выбран меньше, то есть риск получения большого количества ложноположительных результатов. При h большем качество предсказания растет, но уменьшается время реагирования на действия мошенника, как видно из рисунка 19 резко возрастает количество клиентов, по которым была начата

процедура банкротства в квартале, когда было выявлено мошенничество. Учитывая, что отчетность РСБУ подается организацией раз в квартал, то h = 4 подразумевает отсутствие тренда в течение года.

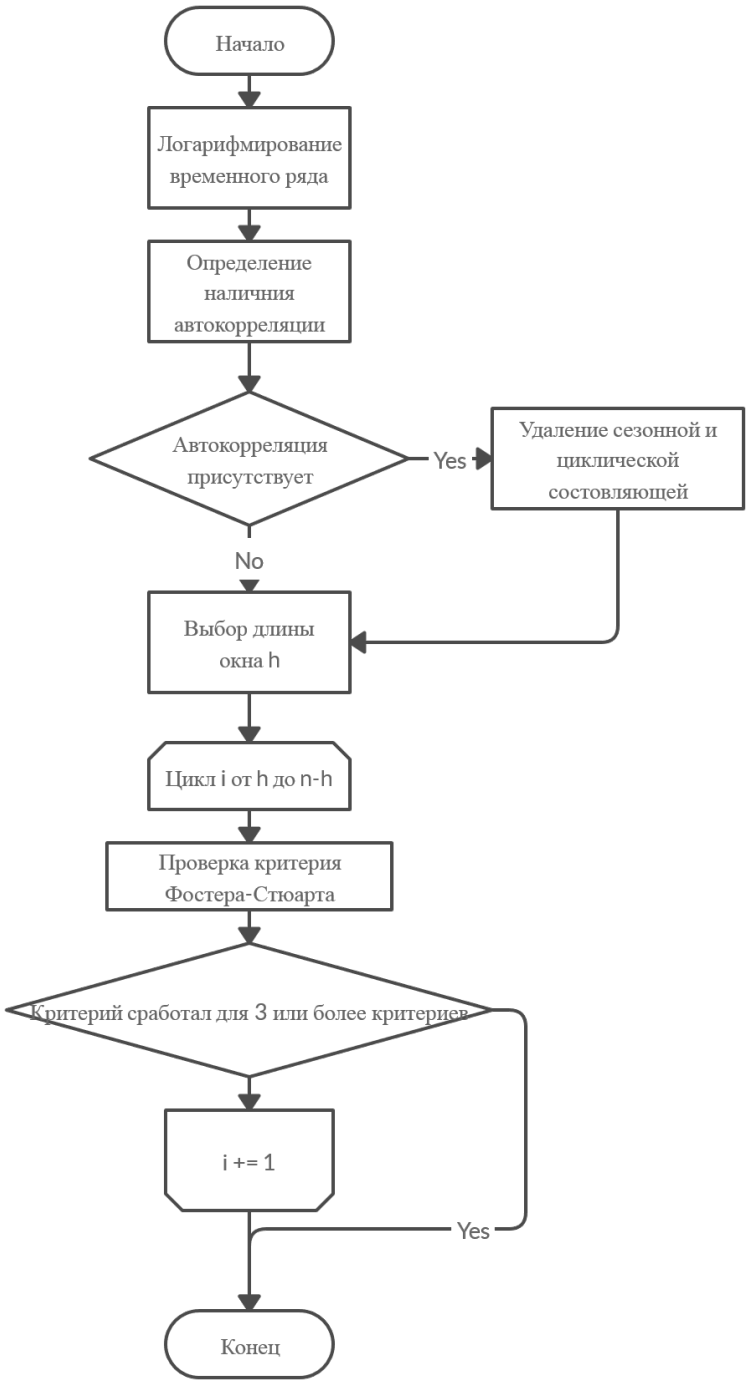

Рисунок 18. Схема работы алгоритма поиска мошенников на основе отчетности РСБУ (Автор, creately.com)

Для получения качества алгоритма был использован подход Out-of-Time, как и для предыдущего алгоритма каждый временной ряд был разбит на 5 равных временных промежутков. Причем на первый промежуток накладывается ограничение по размеру по размеру, он не должен содержать менее 4 измерений. В случае, если нельзя разделить объект на 5 равных частей, то создавался первый промежуток, в который входило 4 измерения, весь остальной временной ряд разбивался на n равных промежутков, где $n < 5$. В данном случае была собрана выборка, которая не содержит клиентов, для которых невозможно выделить хоть один такой временной промежуток. Так как в данном случае отсутствует шаг обучения, то сразу происходила классификация клиента методом Out-of-Time. Для каждого шага Out-of-Time производился расчет F-меры на всей выборке, в случае, если для клиента нельзя сделать шаг, то он исключался из расчета.

После было рассчитано значение F-меры как среднее по всем шагам Out-of-Time, что и является финальным качеством алгоритма.

На рисунке 18 представлена общая схема алгоритма, которая подразумевает, что данные уже получены. Результаты испытаний для различных значений h приведены в таблице 8.

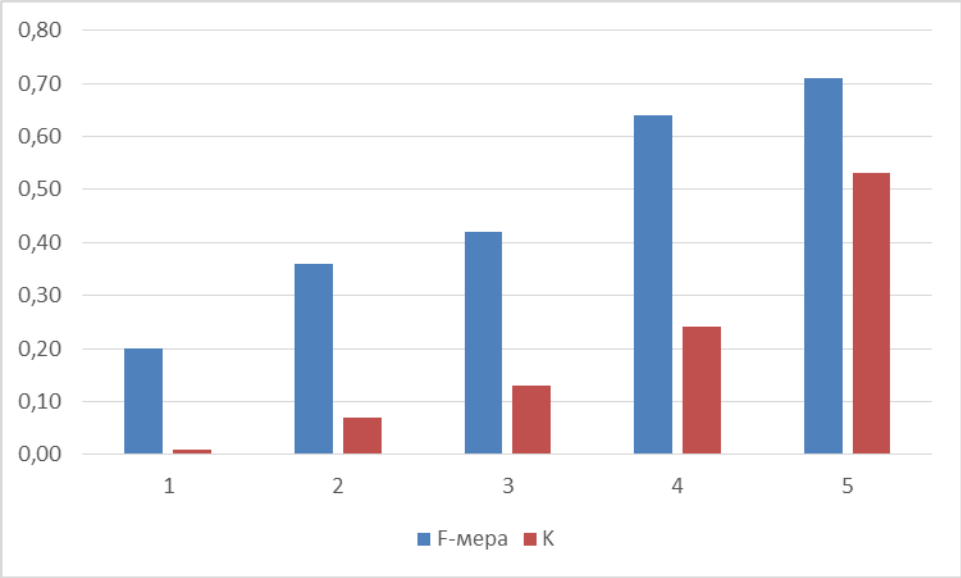

Рисунок 19. Качество алгоритма на обучающей выборке с разными значениями h (Автор, MS Excel)

Вывод. В данном подпункте был представлен алгоритм поиска мошенников на основе отчетности РСБУ. Алгоритм был представлен крайне подробно, и на его основе можно строить программное обеспечение без дополнительного изучения литературы. При этом построение алгоритма было основано на поведение клиента, который пытается совершить мошенничество, подделывая свою отчетность, что также было раскрыто в данной части работы.

Таблица 8

Качество работы алгоритма нахождения мошенников на основе РСБУ

	Precision	Recall	F-мера	K
h = 2	0,23	0,91	0,36	0,07
h = 4	0,57	0,73	0,64	0,24
h = 5	0,63	0,78	0,69	0,53
Test	0.54	0.69	0,6	0,29

Хотя качество данного алгоритма хуже, чем у представленного в предыдущем параграфе, он обладает рядом весомых преимуществ:

1) Простота вычислений. Алгоритм не требует сложных вычислений, и программное обеспечение на его основе можно выполнять непосредственно на компьютерах пользователей.
2) Алгоритм не требует внутренних данных банка. Данный алгоритм строится на отчетности РСБУ, которая является открытой информацией и доступна в сети интернет.
3) Алгоритм можно применять в момент принятия решения о выдачи кредита. Если предыдущий алгоритм возможно было применять только в период мониторинга или принятия решения об урегулировании проблемной задолженности, то данный алгоритм имеет ограничения в период размером h.

Хотя данный алгоритм и имеет ограничение по времени начала использования в временное окно размером h, при использовании рекомендаций в один

год, многие банки не выдают кредиты юридическим лицам, которые существуют менее года, а если выдают, то ставят на жесткий контроль.

2.2.3 Алгоритм графового анализа связи клиента для выявления мошенничества

Данный алгоритм рассчитан на использование сторонних данных в первую очередь. Идея этого алгоритма основывается на нескольких выявленных закономерностях:

1) Чаще всего заемщик при мошенничестве пытается вывести кредитные деньги через связанных с ним контрагентов, которые являются соучастниками.
2) Мошенник может улучшать свои финансовые показатели, при помощи связанных лиц. Могут совершаться фиктивные сделки, при которых одна компания переводит деньги другой, но не получает ничего за это.
3) При продаже имущества по заниженной цене с целью совершить преднамеренное банкротство, с большей вероятностью будет происходить через связанные компании.

При этом связь между компанией мошенником и компанией, которая помогает ему может основываться на: близкой, родственной связи владельцев компаний; владелец компании мошенника может быть бенефициаром контрагента; они могут входить в одну группу компаний; компания мошенник может владеть частью компании контрагента. Реже, в сговор вступают компании, которые слабо связаны, а выступают как обычные соучастники.

Необходимо построить алгоритм, который сможет строить связи различного типа между компаниями. Классическим способом решения подобных задач является использование алгоритмов на основе графа. В основу представленного далее алгоритма легли идеи анализа мошеннических учетных записей в социальных сетях и на электронных торговых площадках.

Алгоритмы по выявлению мошеннических аккаунтов предполагают разделение всех аккаунтов на три категории: мошенники, пособники, обычные пользователи. Таким образом необходимо разбить всех заемщиков и связанных с ними лиц на эти три категории. Очевидно, что в категорию мошенников необходимо добавить непосредственно компании, которые совершают правонарушение. Тогда пособниками будет считать те компании, с помощью которых совершается мошенничество. Все остальные компании будут переведены в класс обычных клиентов или класс «неизвестных», если по ним слишком мало информации.

Следующий шаг, который необходимо совершить – это выделить непосредственно типы связи, по которым будет строиться граф. Исходя из приведенных ранее закономерностей и алгоритмов в данной работе предлагается выделить следующие типы связи, отсортированные по их силе и каждой такой связи, присвоим некоторый вес w от 0 до 1:

1) Выявленные мошеннические транзакции на основе алгоритма, представленного в подпункте 1 данного параграфа или найденные «вручную» сотрудниками банка. Вес w равен 1.
2) Одна компания владеет частью другой компании. Вес w равен 0,8.
3) Бенефициар одной компании является бенефициаром другой компании или родственником бенефициара. Вес w равен 0,6
4) Юридические, почтовые адреса или адреса офисов компаний совпадают. Вес w равен 0,3.
5) Существует информация о том, что осуществляются денежные переводы между компаниями. Вес w равен 0,05.

После выбора видов связи на их основе необходимо построить неориентированный граф, на котором, предварительно можно разметить компании по классам, что увеличит качество предсказания. Стоит отметить, что разметка не должна исходить из предположений, а должна исходить из фактической информации о категории заемщика.

Сам алгоритм строится на основе марковской сети [28]. Марковская сеть – это модель графа, в которой множество случайных величин обладает Марковским свойством. Если более подробно, то суть алгоритма заключается в определение категории вершины графа (категории компании) на основе информации о категории своих соседей и собственного текущего состояния. Для категоризации компании необходимо обновить информацию о состояние вершин при помощи метода распространения доверия (Belief Propagation), который представлен в формулах 2.2.19 и 2.2.20.

$$m_{ij}(\sigma) \leftarrow \sum_{\sigma'} \psi(\sigma', \sigma) * \sum_{l \in L} w_l * \prod_{n \in N(i)/j} m_{nj}(\sigma') \qquad 2.2.19$$

$$b_i(\sigma) \leftarrow k \prod_{j \in N(i)} m_{ij}(\sigma) \qquad 2.2.20$$

Здесь $m_{ij}(\sigma)$- сообщение, отправленное узлом i узлу j; N(i) – совокупность узлов, соседних i; $\psi(\sigma', \sigma)$ – вход в матрицу распространения, дающий вероятность нахождения в состоянии σ' при наличии соседнего узла в состояние σ; w_l- вес l типа связи между вершинами; L – набор всех связей между двумя вершинами; k – константа нормировки; $b_i(\sigma)$ – уровень доверия узлу i в состояние σ.

Матрица распространения доверия – это матрица, в которой описаны вероятности перехода одного состояния в другое на основе состояний соседних вершин, она приведена в таблице 9. Теперь на основе уровня доверия к узлу $b_i(\sigma)$ необходимо присвоить ему одну из категорий, для этого на основе максимизации функции качества классификации F-меры необходимо выбрать пороговое значение r.

Стоит отметить, что значения функции $\psi(\sigma', \sigma)$ и весов w составлялись экспертно и требуют исправления для каждого пользователя алгоритма. Данные веса можно получить, построив граф по заданным правилам, но с достоверно известными категориями компаний и провести вычисление вероятности на основе Байесовской статистики или на основе задачи оптимизации.

Таблица 9

Матрица распространения доверия

Соседнее соединение	Состояние узла		
	Мошенник	Пособник	Честный
Мошенник	0,4	0,9	0,2
Пособник	0,8	0,4	0,3
Честный	0,2	0,3	0,6

Для определения качества предложенного алгоритма и построения матрицы распространения доверия был проведен машинный эксперимент. Для проведения эксперимента, как и ранее, был выбран язык python. В частности, как основной инструмент была использована библиотека SciPy, созданная для решения задач алгебры и оптимизации.

Была собрана выборка аналогичная собранной для тестирования предыдущих алгоритма, представленного в данном параграфе. Далее, аналогичным способом была разбита на обучающую и тестовую выборки.

Для нахождения оптимальных значений матрицы распространения доверия было решено использовать задачу оптимизации. Первым этапом было построено N графов из клиентов обучающей выборки и были размечены все классы компаний на этих графах, где N – число компаний в обучающей выборке. Для построения были указаны все указанные ранее виды связи между клиентами. Так как количество связей клиента при помощи транзакций может быть крайне большим, то было принято ограничить размеры графа. Для этого отсекались все связи рассматриваемого в данном графе клиента удаленные от него более чем на 3 шага.

Далее было построено N таких же графов, но на них не отмечался класс клиента. После чего была построена задача оптимизации, при решение которой должны быть подобраны такие значения матрицы распространения доверия, при которых ошибка в предсказание класса клиента на заранее размеченном графе и графе, который размечает алгоритм была минимальной. В задачах оптимизации

принято минимизировать значение некоторой функции качества. В данном случае для нахождения оптимальных значений матрицы распространения доверия использовалась метрика LogLoss [7], которая является стандартной метрикой оптимизации для задач классификации.

Для окончательной проверки качества алгоритма на тестовой выборке было проведено тестирование Out-of-Time. Для этого у хороших клиентов использовались 5 различных состояний связей за различные периоды времени существования компании. Для мошенников происходило аналогичное разбиение, но как период времени для разбиения использовался промежуток, в который совершалось мошенничество.

Для тестирования возможности алгоритма предсказывать мошенничество на период, в который не хватает транзакций для построения алгоритма поиска мошенничества на основе назначения платежа, было проведено тестирование аналогичное проведенному ранее, но была убрана связь по транзакциям, которые признали мошенническими.

Для тестирования возможности алгоритма работать без связи по обычным транзакциям было проведено еще одно тестирование, аналогичное предыдущему, но в данном случае была убрана связь по обычным транзакциям. Такое тестирование необходимо. В случае, результаты данного тестирования не будут отличаться от результатов со всеми связями, то можно удалить связь по обычным транзакциям. Что в свою очередь сделает алгоритм вычислительно более простым.

Последим проведенным тестированием было тестирование с удалением всех видов транзакций. Это необходимо для проверки качества алгоритма на моменте принятия решения о выдачи кредита, когда отсутствует информация о транзакциях. Результаты тестирования приведены в таблице 10.

В результате выполнения данного алгоритма было получено знание о категории компании в данный момент времени, при построение которого, в отли-

чие от обычных алгоритмов на основе марковской сети, было внесено очень важное знание о силе типа связи двух компаний. Очевидно, что наличие транзакций между мошенником А и некоторой компанией Б не делает компанию Б пособником, если между ними существует только транзакционная связь, но вероятность того, что компания Б является пособником резко возрастает, если часть транзакций были признаны мошенническими.

Таблица 10

Качество классификации клиентов на основе графового анализа

	Precision	Recall	F-мера
Все виды связи	0,61	0,64	0,69
Исключены транзакции признанные мошенническими	0,52	0,49	0,6
Исключена связь по обычным транзакциям	0,57	0,51	0,64
Исключены все виды транзакций	0,46	0,44	0,45

В таблице 10 приведено качество данного алгоритма, а на рисунке 20 приведена визуализация работы алгоритма в двух состояниях: начальное состояние без внесения априорной информации о категории компании (а) и результатом работы алгоритма (б), где красными размечены мошенники, желтым пособники, белым «хорошие» компании, а серым компании, по которым не известен результат.

Выводы. В данной части выпускной квалификационной работы был представлен алгоритм поиска мошенников и тех компаний, с помощью которых происходит правонарушение.

Представленный алгоритм обладает меньшей эффективность, алгоритм на основе нейронной сети типа RNN-Autoencoder, но схожим с алгоритмом на основе анализа временных рядов. При этом, алгоритм отчетности РСБУ и приведенный в данной части алгоритм выполняют различные задачи, первый пытается

найти мошенников, которые подделывают документацию, второй пытается найти мошенников, которые используют в своих действиях другие компании.

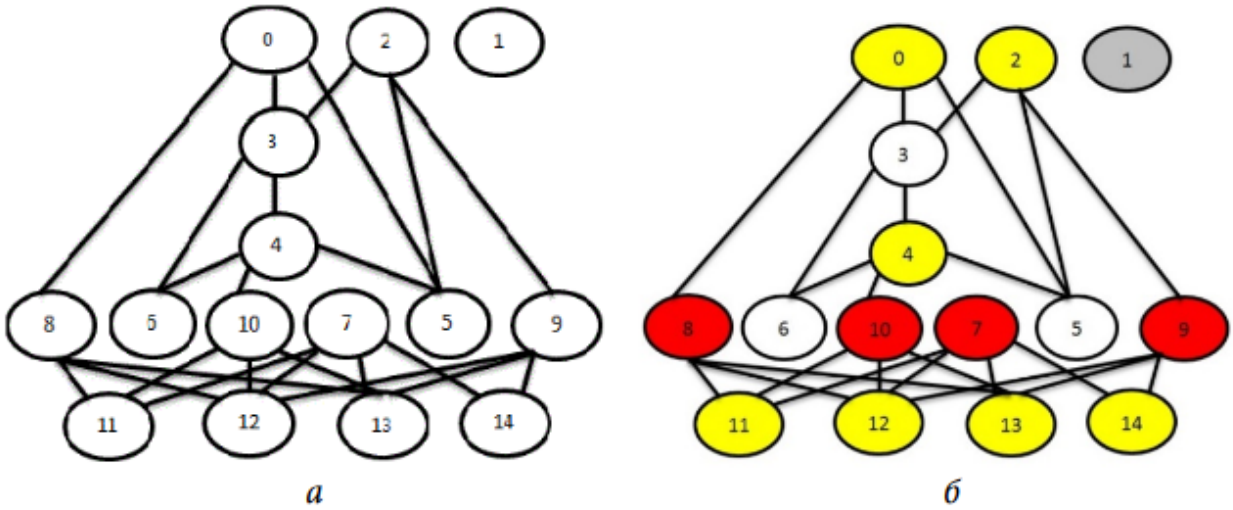

Рисунок 20. Модель графа до (а) и после применения алгоритма (б) (Автор, cre-ately.com)

Описание данного алгоритма было приведено достаточно подробно для реализации программного обеспечения на его основе без изучения дополнительной литературы. Стоит отметить, что данный метод обладает большой гибкостью с точки зрения выбора основных влияющих факторов, каждый пользователь данного алгоритма сможет сам настроить типы связей и внести коэффициенты, которые ему кажутся правильными. За счет такой гибкости, данный алгоритм показывает гибкость и в необходимых для его использования ресурсов, ведь юридических связей у одной компании обычно не очень много, также можно ограничить на количество вершин, которые будут учтены в расчете.

При этом, данный алгоритм не ограничивается во времени, как предыдущие. Его можно использовать и для компаний, которые только что открылись, так как большая часть признаков основывается на информации, которая существует в открытом доступе с самого начала существования юридического лица.

2.2.4 Общие выводы

В данном параграфе были предложены три различных алгоритма которые можно использовать, как совместно, так и раздельно друг от друга. При этом совместно они покрывают все время жизни юридического лица.

Для их построения использовались данные всех категорий надежности, приведенные в первом параграфе данной главы, что позволяет учесть максимально много информации о клиенте при поиске мошенников.

Приведенные алгоритмы, в совокупности, покрывают все виды мошенничества, которые рассматриваются в данной работе. Все три метода способны выявлять преднамеренное банкротство, как самое опасное с финансовой точки зрения действие для банка. Они способны выявить и мошенничество в сфере кредитование, и незаконное получение кредита, особенно хорошо это удается алгоритму на основе анализа РСБУ, о чем можно судить исходя из его построения и методов мошенничества, приведенных в параграфе 1 первой главы.

Разработанные методы основываются на математических концепциях, которые показывают высокую эффективность в других схожих задачах. При этом были эти алгоритмы были адаптированы для решения поставленной задачи.

Стоит особенно выделить тот факт, что предложенные алгоритмы не являются зависимыми друг от друга и могут использоваться отдельно. К примеру, если банк уже анализирует отчетность РСБУ и транзакции клиента, но не анализирует его связи, или процесс анализа происходит вручную, то банк может реализовать только алгоритм на основе графа связи клиентов без потери общего качества.

Стоит отметить тот факт, что все представленные методы являются интерпретируемыми, что позволяет сотруднику, который будет проводить анализ не только узнать факт срабатывания критерия мошенничества для заемщика, но и более подробно и детально изучить проблему, что также снизит вероятность ложноположительных срабатываний. В частности, не составляет труда вывести

на экран личного компьютера сотрудника график изменения показателей отчетности РСБУ и указать период, где отсутствует тренд или вывести конкретные транзакции, которые были размечены как мошеннические. С точки зрения графового анализа, сам граф с размеченными компаниями станет хорошим пояснением к выводам алгоритма.

2.3 Процесс построения архитектуры системы выполнения алгоритмов для нахождения мошенников среди юридических лиц

В данном параграфе будет предложена методика построения IT-инфраструктуры для выполнения методов поиска мошенников, приведенных в предыдущем параграфе. В частности, будут представлены: базы данных для хранения исходных данных и результата анализа, среды для работы алгоритмов и способы передачи результатов конечному пользователю.

Стоит отметить, что предложенная IT-инфраструктура является рекомендуемой, но не обязательной. Если предложенные инструменты не используются в банке, то стоит сделать расчет эффективности их внедрения, если срок окупаемости будет слишком большой, то рекомендуется исходить из текущей реализации IT-инфраструктуры банка.

Программная архитектура. Задачу построения хранилища данных необходимо разделить на две части. Первая часть – это построения системы хранения большого количества реляционных данных. Под такими данными понимается информация о транзакциях клиента, отчетность РСБУ, и результатов по каждому клиенту банка. Второй задачей является построение графовой базы данных для хранения информации о графах связи заемщика.

Первая часть сводится к выбору наиболее оптимальной системы хранения и обработки Больших данных.

Несмотря на то, что обычные базы данных не предназначены для хранения таких объемов данных, какие порождает транзакционная деятельность клиентов

(могут доходить до нескольких зеттабайт за десятилетие), при предоставлении достаточных вычислительных мощностей отлично решают эту задачу.

С другой стороны, специализированные системы уже достаточно давно разрабатываются совместно с дополнительными инструментами для обработки и передачи данных, чего лишены обычные СУБД.

Наиболее оптимальным выбором системы для хранения данных в данном случае является Apache Hadoop (далее hadoop) [48]. Сам Hadoop представляет собой набор библиотек, фреймворков и утилит для распределенного хранения о обработки данных. Первым и самым весомым преимуществом Hadoop является то, что эта система разрабатывается под лицензией Apache, а значит является системой с открыты исходным кодом и является бесплатным в отличие от своего главного конкурента Oracle Big Data. С другой стороны, при использовании Hadoop будет недоступна техническая поддержка со стороны разработчика. В отличие от MongoDB, Hadoop позволяет хранить данные в виде реляционных таблиц, а не только как файлы, что делает работу с реляционными данными проще и удобнее. Также в отличие от Oracle Big Data, Hadoop крайне гибко настраивается и позволяет развернуть только те системы, которые необходимы предприятию.

В рамках данной работы предлагается использовать систему Hadoop для хранения реляционных данных.

Для решения второй задачи, а конкретно выбора системы хранения и обработки данных в виде графа. Для таких задач существует специализированные базы данных в их перечень входят: ArangoDB, FlockDB, InfiniteGraph. Отдельно стоит выделить Neo4j, которая стала современным стандартом для хранения и обработки данных в виде графов. Данная система поддерживает обработку транзакций в реальном времени.

Однако, у всех данных СУБД имеется существенный недостаток: большая часть из них не являются распределенными и выполняют все вычисления в оперативной памяти одного сервера, данная архитектура представлена на рисунке

20. Такой подход противоречит стандартам по обработке Big Data, и делает невозможным горизонтальное расширение аппаратного обеспечения для проведения вычислений. Данную проблему можно обойти, передавая единовременно обрабатывая лишь ограниченный в размерах граф для одного клиента по запросу пользователей. Что в свою очередь приведет к замедлению процесса получения данных пользователем и возможному ухудшению качества получаемых результатов.

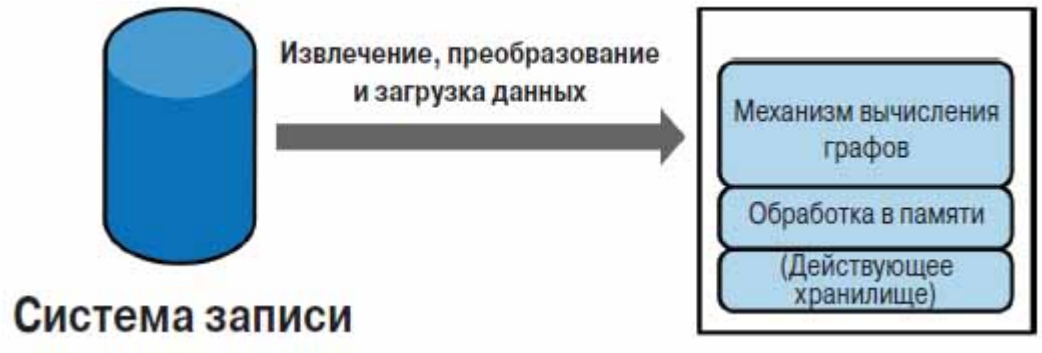

Рисунок 20. Укрупненная схема типичной среды расчета графовых СУБД [10]

Существуют графовые СУБД, которые решают данную проблему, но учитывая тот факт, что для обработки реляционных данных предлагается использовать Hadoop, то стоит обратить внимание на реализации систем для хранения и обработки графовой информации созданных для него. Таких систем существует две [10]:

1. Giraph – распределенная отказоустойчивая система, основанная на модели параллельных вычислений (Bulk Synchronous Parallel) для обработки больших объемов данных с применением алгоритмов параллелизаци.
2. GraphLab – основанная на графах высокопроизводительная инфраструктура распределенных вычислений, написанная на C++.

Обе системы являются распределенными, отказоустойчивыми и предназначены для работы с графовыми данными. Но при этом разработка на основе

системы Giraph ведется на языке программирования Java, как и для системы Hadoop, когда разработка на основе системы GraphLab ведется на основе языка C++. Следовательно, при выборе GraphLab присутствует необходимость в компетенции разработчика на на C++.

С другой стороны, GraphLab позиционируется как Real-Time система вычислений, она хранит централизованно метаинформацию о данных и ходе вычисления. В то время, когда Giraph необходимо сначала перенести все отобранные по определённому условию данные на кластер для вычисления.

Существенной разницы в выборе между двумя этими инструментами нет, она зависит от способа решения поставленной задачи: в случае, если вычисление планируется производить в real-time режиме, то необходимо использовать GraphLab. В случае, если планируются пакетные вычисления по расписанию, а затем предоставление данных пользователю, то рекомендуется использовать систему Giraph.

В данной работе было принято решение об использовании системы Apache Hadoop для хранения как реляционных данных, так и графовых данных, для реализации доступа к данным, вычислений и передачи данных пользователю можно использовать инструменты из экосистемы Hadoop.

Перед решением задачи о выборе инструментов и реализации методики построения инфраструктуры необходимо принять решение о том, будут ли выполняться расчеты в real-time режиме или будут происходить пакетно с определенным периодом, а пользователю будет предоставляться только результат.

К данному вопросу можно подойти как с технической точки зрения, так и с точки зрения бизнеса.

С технической точки зрения расчет по запросу невозможно реализовать в реальном времени, что связано с необходимостью найти нужные транзакции среди всех транзакций, сохраненных в банке, затем по выбранным транзакциям необходимо провести расчет. Все эти действия занимают, достаточно большое

количество времени. То же касается и остальных алгоритмов, всегда будет необходимость реализовать поиск среди всех данных банка, а затем реализовать расчет по ним.

С точки зрения бизнеса также стоит вести расчет одновременно по всему портфелю. Если проводить поиск только по отдельным клиентам, для которых был сделан запрос, то есть шанс упустить мошенника, о котором даже не подозревали.

Следовательно, для реализации необходимо выбрать пакетную выгрузку данных. После того, как был выбран тип расчета, можно приступать к реализации архитектуры система.

Сам Hadoop представляет собой распределенную файловую систему – HDFS (Hadoop Distributed File System). Данная файловая система подразумевает наличие двух типов серверов. Первый тип сервера называется NameNode – является менеджером системы, а второй DataNode – выступаю непосредственными хранилищами данных и выполняет операции.

Для хранения графовых данных, как уже указывалось ранее, будет использована система Giraph вследствие ее приспособленности для пакетных расчетов. Для хранения реляционных данных будет использована СУБД Hive, самая популярная на сегодняшний день реляционная СУБД для Hadoop. Обе данные платформы используют алгоритм MapReduce для быстрого доступа к данным [59]. Кратко данный подход можно выразить так: работа с данными делится на два шага: map – параллельно собирает данные подходящие под заданное условие по всей БД и reduce – агрегирует данные собранные на шаге map и возвращает их пользователю.

Первым шагом необходимо собрать из внешних для данной системы источников данные в реляционные таблицы. Так как алгоритм графового анализа связей клиентов может использовать результаты алгоритма поиска мошенников на основе транзакций, а операция обновления данные в Hadoop стоит дороже

операции записи и чтения, то данные для графового алгоритма будут собираться позже. Схема исходных данных приведена на рисунке 21.

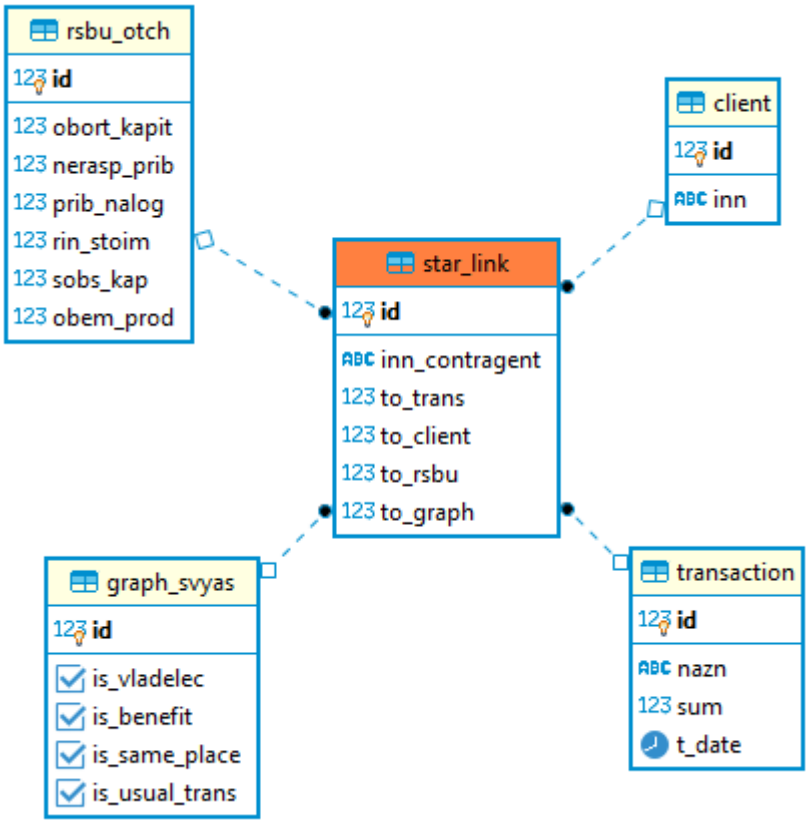

Рисунок 21. Схема БД для исходных данных (Автор. Dbeaver)

Такой тип построения схемы базы данных позволяет без проблем добавить новые таблицы в связь, не создавая при этом множество внешних ключей в других таблицах связи, а добавив только один в таблицу «star_link».

После получения исходных данных необходимо провести вычисления алгоритма. Для этого в Hadoop предусмотрено множество инструментов для реализации алгоритмов. Необходимо использовать следующие три:

1) Apache Spark – позволяет производить распределенные вычисления на нескольких DataNode.
2) MLlib – является одной из библиотек Spark, предназначенной для машинного обучения и статистических расчетов.
3) Deeplearning4j – это библиотека для глубокого обучения.

На основе данных библиотек достаточно просто реализовать приведенные ранее алгоритмы. При этом их выполнение будет проходить распараллелено на различных DataNode, что существенно увеличит скорость работы системы. После того как будет получен результат алгоритма анализа назначений платежа и алгоритма анализа отчетности РСБУ их необходимо сохранить в результирующие таблицы. Эту процедуру легко можно проделать при помощи Spark.

После сохранения первых двух алгоритмов необходимо провести расчет алгоритма на графах. Так как Giraph входит в экосистему Hadoop в нем присутствуют инструменты для конвертации данных представленных в Hive в виде реляционных таблиц во внутреннее представление. Следовательно, при помощи MapReduce необходимо перенести данные из исходных таблиц и результаты работы алгоритма поиска мошенников на основе назначения платежа в систему Giraph, где будет произведен расчет, а данные будут записаны в результирующие таблицы.

Таким образом, на данном этапе будут готовы таблицы с результатами работы алгоритмов, и необходимо передать их конечному пользователю. Для этого в экосистеме Hadoop присутствует инструмент под название Apache Kafka [58]. Данная система основана на технологии очередей и может передавать данные одновременно в несколько источников. Кроме того, в ней реализован инструмент, позволяющий передавать некоторый срез данных по запросу. Общая схема работы Kafka представлена на рисунке 22, на данном рисунке схематично изображена передача данных от нескольких источников нескольким получателем. Такой подход к передаче данных позволит выполнять ее мгновенно по запросу пользователя, а не хранить данные в системе, где они будут использоваться.

На данный момент был описан процесс расчета, при этом был выбран способ расчета пакетами в некоторый период времени. Таким периодом предлагается выбрать минимальный период обновления данных в источниках данных. Но также, до этого момента был упущен непосредственный запуск расчета по рас-

писанию. Для этого в Hadoop предусмотрена система Oozie, она является менеджером задач в Hadoop и позволяет объединять в одном процессе несколько различных задач. В данную систему включена функция выполнения таких процессов по расписанию, при этом одна задача начинается после того, как пришел сигнал от предыдущей задачи о том, что выполнение задачи закончено.

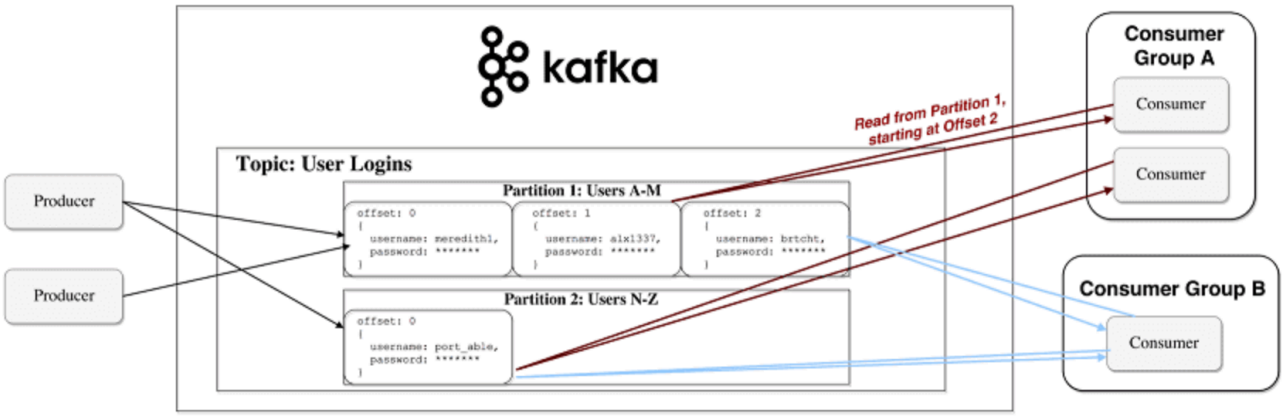

Рисунок 22. Принцип действия распределенной системы обработки сообщений Apache Kafka [58]

Остался один не затронутый участок архитектуры построенной системы – это не посредственный источник данных. Наилучшим вариантом, с точки зрения оптимизации скорости и ресурсов будет ситуация, когда у банка уже есть инфраструктура хранения данных на системе Hadoop, в таком случае необходимо реализовать предложенную архитектуру в уже существующей системе.

Если подходящей архитектуры нет, то необходимо организовать выгрузку дынных из целевых систем. В подавляющем большинстве банков уже реализовано хранение транзакций, тогда можно организовать передачу файлов с необходимыми данными в Hadoop через Kafka или уже существующую в организации систему передачи данных.

С точки зрения источника внешних данных рекомендуется использовать Spark-Interfax, так как данная система агрегирует данные из всех необходимых внешних источников. В ней содержится информация и об юридическом устройстве компании, и о бенефициарах, а также собирается отчетность РСБУ.

Последним необходимым шагом является проектирование системы сигналов для конечного пользователя, а также контроль за выполнением проверки по клиенту, который был размечен системами как мошенник. В данной работе предлагается следующий механизм:

1) Создать дубликат результирующих данных.
2) Сравнить обновленные данные с дубликатом и выявить расхождение по конкретным клиентам.
3) Передать данные об этих клиентах в систему пользователя.
4) Обновить данные в дублирующих таблицах.

Данный подход предлагается в данной работе в случае, если система пользователя не хранит данные с результатами на своей стороне, а получает по запросу из системы расчета. В случае, если система пользователя хранит все результаты расчетов на своей стороне, то она сама должна контролировать поиск новых клиентов, размеченных как мошенники.

Для контроля проверки сотрудником данных «плохих» клиентах необходимо производить логирование запросов на выгрузку результатов работы алгоритма сотрудником. Как и в предыдущем варианте данную процедуру можно реализовать, как и на стороне системы расчета, так и на стороне системы пользователя. В обоих случаях необходимо создать таблицу в БД для хранения идентификатора компании (в данном случае ИНН) и информацию о том был ли проверен клиент сотрудником и временем, когда данный клиент был размечен как мошенник первый раз, так же возможно хранить время обновления класса данного клиента. На основе этих данных уже в целевой системе пользователя необходимо осуществлять передачу уведомлений о том, что время на проверку клиента истекло. Такой функционал уже может быть реализован в самой системе, также можно использовать почтовые уведомления.

О заемщиках, которые не были проверены своевременно необходимо сообщать в контролирующую структуру, которой может выступать: подразделение

безопасности, подразделение комплаенс или другое контролирующее подразделение банка.

Процесс вычисления алгоритмов. Для формирования точных требований к аппаратной части архитектуры необходимо сформулировать процесс вычисления алгоритмов на базе предложенной программной архитектуры.

Предлагается рассчитывать алгоритм выявления мошенников на основе назначения платежа для всех клиентов банка одновременно. Тогда на вход алгоритму передаются все исходящие транзакции клиента, за весь срок хранения транзакций клиента, по всему портфелю юридических лиц. Такой расчет является крайне трудоемким, к примеру, количество клиентов, юридических лиц Сбербанка оценивается в 2,5 миллиона, а портфель Альфа Банка оценивается в 543 тысячи. Среднее количество транзакций одного корпоративного клиента оценивается в 5000 в год. Даже предположив, что расчет будет вестись только на горизонте года, то за один раз необходимо сделать расчет на основе 12,5 миллиардов транзакций для Сбербанка и 2.7 миллиардов транзакций для Альфа Банка [43].

Для выполнения алгоритма на основе отчетности РСБУ предлагается совершать расчет для всех юридических лиц, которые существуют год или более, так как для расчета необходимо иметь хотя бы 4 точки измерения, как было показано в предыдущем параграфе. Количество таких клиентов оценивается в 3919046. Учитывая, что количество параметров, для которых необходимо сделать расчет равняется 6, и тот факт, что в современном виде отчетность РСБУ существует с 2016 года, то расчет необходимо провести 70542828 раза.

Для выполнения алгоритма графового анализа связи клиента предлагается совершить расчет одновременно по всем юридическим лицам, зарегистрированным в России, что оценивается в 3648715 объектов. Таким образом, учитывая среднее количество транзакций можно оценить количество связей для всех клиентов в 18 миллиардов. В данном случае были опущены все остальные типы связей, так как их заведомо меньше чем транзакций.

Скорость обновления данных у источников разная, так РСБУ обновляется раз в квартал, а данные из ЕГРЮЛ и транзакции обновляются в течение дня. Но для статистических значимых расчетов нет необходимость совершать перерасчет сразу после появления новых данных. Также, чаще всего источником данных для алгоритма будет не база данных самой системы, а некоторая реплика данной базы. Для системы хранения транзакций такой подход обосновывается высокими нагрузками на саму систему, так как транзакции происходят с высокой частотой. Данные ЕГРЮЛ необходимо получать из сторонней для банка системы, такое взаимодействие происходит по API, предоставляемому ФНС.

Таким образом рекомендуется совершать расчет каждого алгоритма по факту обновления данных для него в непосредственном источнике, но не чаще одного раза в день.

Аппаратное обеспечение. Hadoop является распределенной системой и подразумевает масштабирование системы по горизонтали. Такой подход является крайне удобным и не ограничивает возможности системы максимальными возможностями вычислительной техники.

Общее количество памяти необходимое для хранения исходных данных для вычисления алгоритмов и хранения их результатов, учитывая приведенное ранее количество объектов оценивается в 30 терабайт, для расчета были взяты данные для Сбребанка.

Для формирования аппаратных требований необходимо уточнить, что подход используемы в Hadoop и позволяющий получить быстрый доступ к данным требует дублирование каждого файла несколько раз на различных жестких дисках. Каждый из таких жестких дисков должен находится на своем сервере. Такой подход называется реплицированным. Принято создавать три реплики для каждого файла.

Дополнительно необходимо выделять сервер для NameNode, но данный сервер не требователен к аппаратному обеспечению. Так как NameNode является критичной точкой для системы, поэтому в современной реализации Hadoop

предлагается выделять еще один сервер, который выполняет роль резервной копии NameNode.

При развертывании платформы Hadoop, классическим решением является использование нескольких серверных стоек и, хотя бы одну из реплик хранить на серверной стойке отличной от серверных стоек, на которых хранятся другие реплики. Такой подход увеличивает надежность системы, так как в случае выхода из строя одной серверной стойки (например, из-за неисправности ее блока питания), то останется реплика на другой серверной стойки [63].

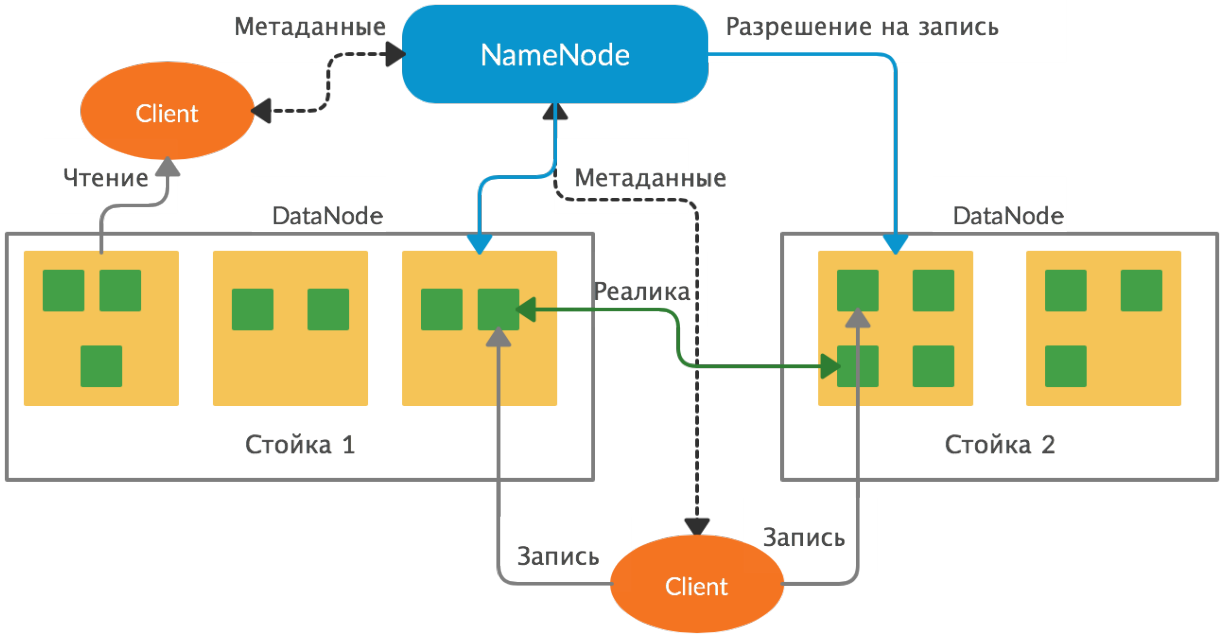

Рисунок 23. Архитектура системы на основе Hadoop (Автор, creately.com)

Общая архитектура системы представлена на рисунке 23. На данную схему было добавлено взаимодействие системы с пользователем для лучшего понимания принципа работы.

Для точного понимания необходимых требований необходимо ознакомится с двумя видами задач, решаемыми на Hadoop. Первой задачей является хранение данных, для наиболее эффективного решения данной задачи принято выделять большое количество жестких дисков небольшого размера. При этом один файл разбивается на несколько небольших частей, каждая из которых хранится на своем носителе, что ускоряет доступ к данным, так как чтение файла

идет параллельно. Второй тип задачи – это проведение операций над хранящимися данными. Для ее выполнения принято уменьшать количество жестких дисков и увеличивать их объем. Общий принцип показан на рисунке 24.

Рисунок 24. Принцип формирования аппаратных требований для Hadoop [63]

Как показано на рисунке 24, при формировании требований в зависимости от задачи не выставляются требования к количеству процессоров и их качественным характеристикам.

Но не смотря на все приведенные требования, стоит учитывать, что Hadoop – это распределенная система и ее общие вычислительные мощности зависят от количества машин. Именно количество машин принесет наибольший прирост к эффективности работы, а не качественные характеристики данных серверов.

Таким образом, предлагается использовать 10 жестких дисков размером по 3 терабайта для хранения информации. Учитывая необходимость создания 3 реплик, общее количество жестких дисков оценивается в 30 штук.

При этом предлагается выделить для каждого жесткого диска отдельный сервер. Данные сервера предлагается разделить на две серверные стойки, в одной из которых будет 20 серверов. Такое расположение уменьшит сложность постро-

ения процесса репликации данных. Более того, данный подход позволит сократить затраты, так как минимизирует количества необходимого аппаратного обеспечения.

Представленное распределение жестких дисков между вычислительными машинами и расположение самих машин, необходимо только при использование собственных мощностей. Альтернативным решением может стать использование облачных вычислений, которые предоставляет сторонняя компания. Такие облачные решения, зачастую, предлагают уже готовую и настроенную под требования арендатора архитектуру. Примером подобного сервиса может стать компания Amazon EMR, которая предоставляет гибкую возможность выбора аппаратной и программной архитектуры.

Для проведения эффективных вычислений на предложенном количестве серверов необходимо выделить достаточное количество оперативной памяти и процессоры с высокой производительностью. В данном случае использование Hadoop позволяет выставлять достаточно низкие минимальные требования к оперативной памяти и производительности процессора для каждой DataNode. В основном, от данных показателей зависит скорость вычисления. Таким образом за минимальные требования можно принять средние показатели для персональных компьютеров.

Для формирования оптимальных требований был проведен эксперимент. Данный эксперимент заключался в произведение расчетов по всей имеющейся выборке, размером в 30 терабайт с различным количеством выделенных ресурсов.

В результате данного эксперимента было показано, что при выделении 50 гигабайт оперативной памяти и двух восьмиядерных процессоров с частотой 3 ГГц расчет совершается за 3 часа, что является удовлетворительным. В случае необходимости увеличить количество серверов.

Следующим важным пунктом в аппаратной архитектуре являются сетевые карты. Скорость доступа к данным и их передачи зависит не только от жестких

дисков, но и от скорости работы сети. От качества и способа распределения жестких дисков зависит скорость доступа к данным на конкретной DataNode, но скорость передачи данных между DataNode и пользователем зависит от скорости сети. Таким образом, необходимо сформировать требования для сетевой архитектуры.

В качестве механизмы передачи данных в похожих системах принято использовать Ethernet. Данный выбор обосновывается большей надежностью, чем беспроводные методы передачи данных. С другой стороны, в отличие от своих аналогов, например, оптоволоконных технологий, можно более точно выбирать скорость передачи данных.

Существуют стандартные рекомендации по выбору скорости передачи данных между NameNode, DataNode и пользователем [62]. В случае ранее представленных требований к оперативной памяти, процессорам и жестким дискам рекомендуется использование сетевых интерфейсов со скоростью 1 гигабит в секунду. При увеличении пространства на жестких дисках на каждом сервере в два раза, предлагается увеличить скорость передачи данных в два раза. В случае увеличения оперативной памяти и производительности процессоров в три раза, так же рекомендуется увеличить скорость передачи данных по сети в два раза.

Представленная аппаратная архитектура, по мнению автора, обладает возможностью к масштабированию как горизонтально, так и вертикально. В случае, ели нет возможности добавить новые сервера, можно увеличить качественные характеристики текущих серверов. Обычно, вертикальное масштабирование приводит к повышению производительности менее чем с линейной зависимостью. В данном случае, было предложено использовать больше серверов, чем необходимо учитывая рекомендации для архитектуры, нацеленной на вычисление.

Выводы. В данном параграфе была изложена наиболее оптимальная с точки зрения архитектуры, скорости работы и финансовых затрат архитектура системы по анализу клиентов банка с целью поиска мошенников. Данная система

способна проводить анализ всех клиентов банка одновременно, что связано с выбранной платформой, Hadoop. Данная платформа является распределенной системой, которая легко масштабируется по горизонтали.

Безусловно, предложенная архитектура не отталкивается от потребностей, возможностей и уже реализованных инструментов конкретного пользователя и при принятии решения о ее реализации банк должен опираться на имеющиеся ресурсы и требования.

При этом было дано достаточно подробное описания шагов, которые необходимо выполнить с предложениями по конкретным инструментам, но любой реализующий данную архитектуру может заменить предложенные инструменты уже существующими. Так, не составит труда найти необходимые для построения системы на основе предложенной архитектуры в том, что предлагает Oracle Big Data.

Было предложен способ выполнения алгоритма при промышленном применение, вместе с этим были представлены требования к размеру хранилища для исходных данных и результатов работы алгоритмов.

Далее была предложена аппаратная архитектура и сформированы оптимальные требования к качественным характеристикам аппаратного обеспечения. Безусловно существует множество различных конфигураций системы, которые будут одинаково эффективны для решения поставленной задачи. Исходя из этого, были предложены общие подходы и правила при формировании аппаратной архитектуры, которыми можно воспользоваться при реализации системы на основе данной методики.

Неоспоримыми качествами предложенной системы является возможность использования инструментов одной платформы, которые используют легко совместимые интерфейсы и нет необходимости в разработке собственных оберток для взаимодействия между компонентами.

2.4 Выводы по второй главе

В данной главе была предложена методика проектирования ИТ-инфраструктуры систем для поиска мошенничества с банковскими кредитами среди юридических лиц. При этом были рассмотрены три ключевых элемента: источники данных, алгоритмы и архитектура самой системы.

Первыми был рассмотрен вопрос источников данных. Была предложена классификация источников по их надежности, а также предложены конкретные варианты для использования. Такой подход необходим, любая система по поиску мошенников не будет эффективной, если будет использовать данные, которые может подделать мошенник или требуется специальный анализ этих данных как было предложено в подпараграфе 2.2.2 данной главы.

Далее были предложены три алгоритма для выявления мошеннических действий клиента, что является наиболее важной частью данной работы, так как описанных в научных или достоверных источниках аналогов практически нет или они плохо раскрыты. Предложенные алгоритмы хорошо описаны, а на недостаточно раскрытые темы, в связи с выходом их за рамки данной работы, приведены ссылки на подробное описание. Это позволяет без проблем реализовать программное обеспечение на основе приведенного описания. Данные алгоритмы показали высокое качество на тестовых данных. Кроме того, они являются независимыми друг от друга и могут применяться отдельно, в зависимости от требований.

Далее была приведена архитектура системы на основе предложенных алгоритмов. Она была спроектирована специально для них и является наиболее приемлемой с точки зрения финансовых затрат и эффективности. Архитектура построена на свободном программном обеспечение, что делает ее не затратной в реализации. При этом, данное программное обеспечение является одним из лучших в мире для своих задач, таки образом не было потеряно в качестве при выборе данного инструмента. Также были предложены различные альтернативы в

зависимости от уже реализованной в компании архитектуры и используемых инструментов.

Таким образом в данной главе была предложена полная методика разработки ИТ-инфраструктуры системы поиска мошенников с банковскими кредитами среди юридических лиц. Была полностью выполнена поставленная в первой главе задача.

Как еще один плюс данной методики стоит отметить возможность использования ее частично по потребностям заказчика или реализатора. Возможно использовать только алгоритмы или какое-то их подмножество, можно использовать только классификацию данных или можно использовать только построение архитектуры приложения.

ГЛАВА 3. ОЦЕНКА ТЕХНИКО-ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ РАЗРАБОТАННОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ПРЕДЛОЖЕННОЙ МЕТОДИКИ

3.1 Разработка пользовательской части системы поиска мошенников с банковскими кредитами среди юридических лиц

В предыдущей главе была рассмотрена только архитектура системы для исполнения алгоритмов и хранения результатов анализа, а пользовательская часть рассматривалась как черный ящик. Для лучшего понимания финального результата и расчета экономической эффективности предложенной методики необходимо рассмотреть пользовательскую часть подробнее.

В качестве целевой системы пользователя использовалась автоматизированная система «Управление проблемными активами» (УПА). Точнее, ее функциональная подсистема «Модели». УПА предназначена для управления всем процессом урегулирования клиента, который был переведен в группу проблемных активов. Функциональная подсистема «Модели» предоставляет пользователям возможность воспользоваться, разработанными на основе сложных математических моделей, инструментами, помогающими принять решение по методу урегулирования проблемного актива.

Выбор данной АС как целевой был принят обосновывается тем, что любой клиент банка, который был заподозрен в мошенничестве становится проблемным активом и дальнейшая работа с ним видеться в соответствующем подразделение.

Несмотря на то, что доступ к самой АС УПА имеют ограниченное число сотрудников, к ФП «Модели» может получить доступ любой сотрудник. Такой подход делает возможным использовать результаты системы, разработанной на основе приведенной методики, и подразделению мониторинга, и кредитному подразделению при рассмотрении заявки на кредит.

Целевыми технологиями данной системы являются:

- Spring Framework – фреймворк для разработки серверной части приложений на языке программирования Java.
- React – библиотека языка JavaScript для разработки веб-интерфейса.
- Oracle Database – СУБД для хранения данных.

Следовательно, для разработки пользовательской части системы необходимо было использовать данный стек технологий.

Было принято решение о хранение результатов работы алгоритмов еще и на стороне системы, при этом накапливая изменения результатов работы. Это решение объясняется более быстрым доступом к данным для пользователей, проводить различный анализ изменений для в результатах по конкретному клиенту, собирать статистику о результатах по всему портфелю. Вместе с этим, УПА, на момент начала разработки системы на основе предложенной методики, уже имела интеграцию с контролирующей качество работы сотрудников АС.

Для передачи данных из Hadoop в АС использовался инструмент Kafka, как и было предложено в предыдущей главе. При этом выгрузка процесс передачи данных начинается по сигналу окончания расчета раз в день. На стороне АС пришедшие данные сравниваются с текущими, если по какому-то клиенту было найдено несоответствие, то пришедшие по нему данные до записываются в таблицы на стороне АС. Далее при сверке данных будет использоваться самая свежая информация о клиенте.

Само приложение, которое предоставляет результаты работы системы разработанной по данной методике получило название «Модель АнтиФрод». При разработке была использована идея одностраничных приложений. Стартовое состояние страницы показано на рисунке 25. Пользователю необходимо ввести ИНН компании и нажать кнопку «Поиск», после чего произойдет поиск результатов работы алгоритмов в базе данных, их переформатирование в соответствие с протоколом передачи данных между серверной частью приложения и пользовательской, передача данных в новом формате и их отображение пользователю.

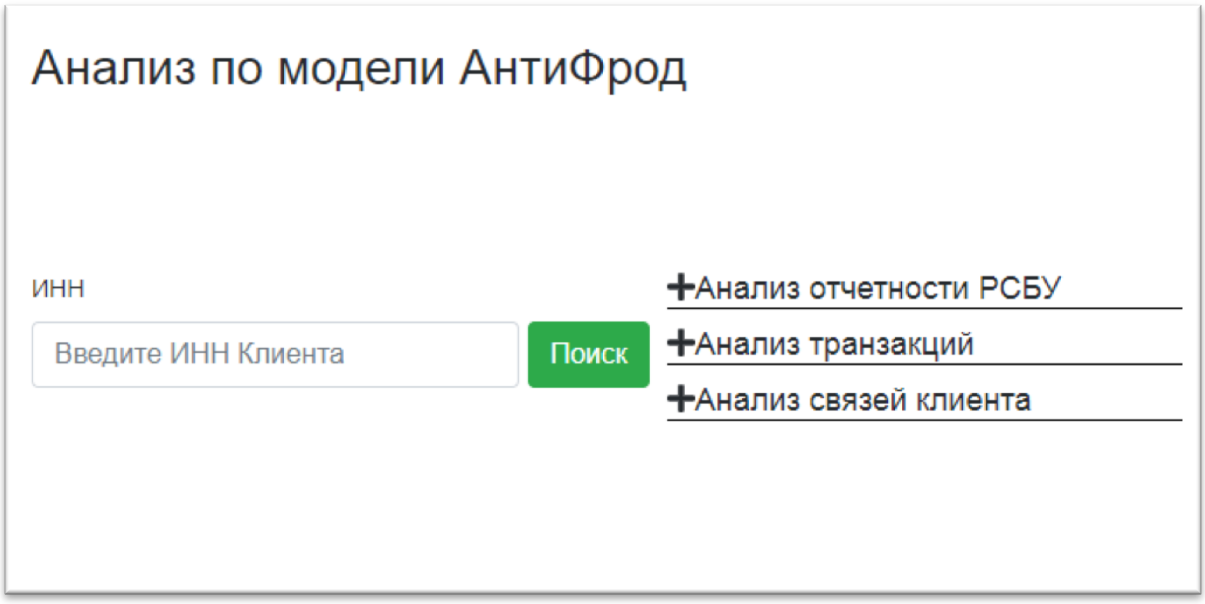

Рисунок 25. Базовое состояние страницы приложения «Модель АнтиФрод» (Автор, ФП «Модель АнтиФрод»)

После совершения запроса сотруднику будет показан результат работы алгоритмов: «Хороший», если алгоритмы не обнаружили у клиента признаков мошенничества, «Подозрительный», если у клиента были обнаружены признаки мошенничества.

Будут представлены результаты работы алгоритмов. На рисунке 26 представлен результат работы алгоритма на основе назначений платежа в транзакциях. Пользователю выводится 4 контрагента, с которыми было совершено наибольшее количество мошеннических транзакций.

По ним указываются дополнительные данные:

1) Общая сумма транзакций подозрительных транзакций.
2) Количество подозрительных транзакций.
3) Средняя сумма подозрительных транзакций.
4) Период совершения подозрительных транзакций.
5) Выведено назначение платежа с самой большой ошибкой у нейронной сети.

Для сравнения предлагается информация из первых четырех пунктов в разрезе всех транзакций клиента. При нажатии на кнопку «i» рядом с названием

компании контрагента пользователь получит список всех уникальных транзакций, которые были помещены как мошеннические с данным контрагентом.

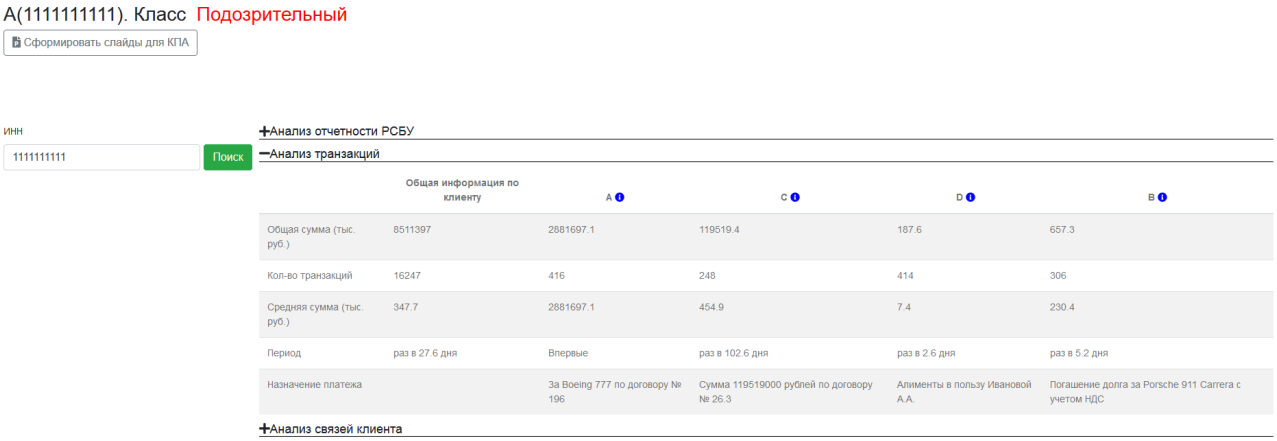

Рисунок 26. Состояние страницы приложения после выполнения поиска клиента (Автор, ФП «Модель АнтиФрод»)

При выборе вкладки «Анализ отчетности РСБУ» пользователь увидит графики изменения во времени пяти анализируемых в алгоритме показателей отчетности РСБУ. Каждый показатель вынесен на отдельный график для удобства анализа показателей сотрудником. При этом участки во времени, которые алгоритм разметил как признак мошенничества отмечаются красным цветом.

Пользователь может посмотреть точные значения по оси абсцисс и ординат наведя курсор мыши на конкретное измерение. Также, присутствует возможность более подробно рассмотреть выбранный на графике участок.

На рисунке 27 представлены два из пяти графиков, которые может увидеть пользователь. На графике изображающем изменение показателя «Оборотный капитал» представлен участок, который был выделен алгоритмом как мошенничество.

Отображение результатов графового анализа связи клиента на момент написания данной работы находятся в разработке, поэтому на рисунке 28 представлен прототип. Отображение связей клиента ограничивается глубиной равной

двум для всех типов связи кроме транзакций. Транзакции ограничиваются глубиной равной одному. Такое ограничение было выбрано из-за недостаточных для отображения более глубокого графа связей клиента.

Рисунок 27. Анализ отчетности РСБУ на странице приложения (Автор, ФП «Модель АнтиФрод»)

Компании, которые были помечены как мошенники красятся на графе в красный цвет, компании, помеченные алгоритмом как пособники красятся в желтый цвет, компании, помеченные как хорошие, красятся в зеленый цвет.

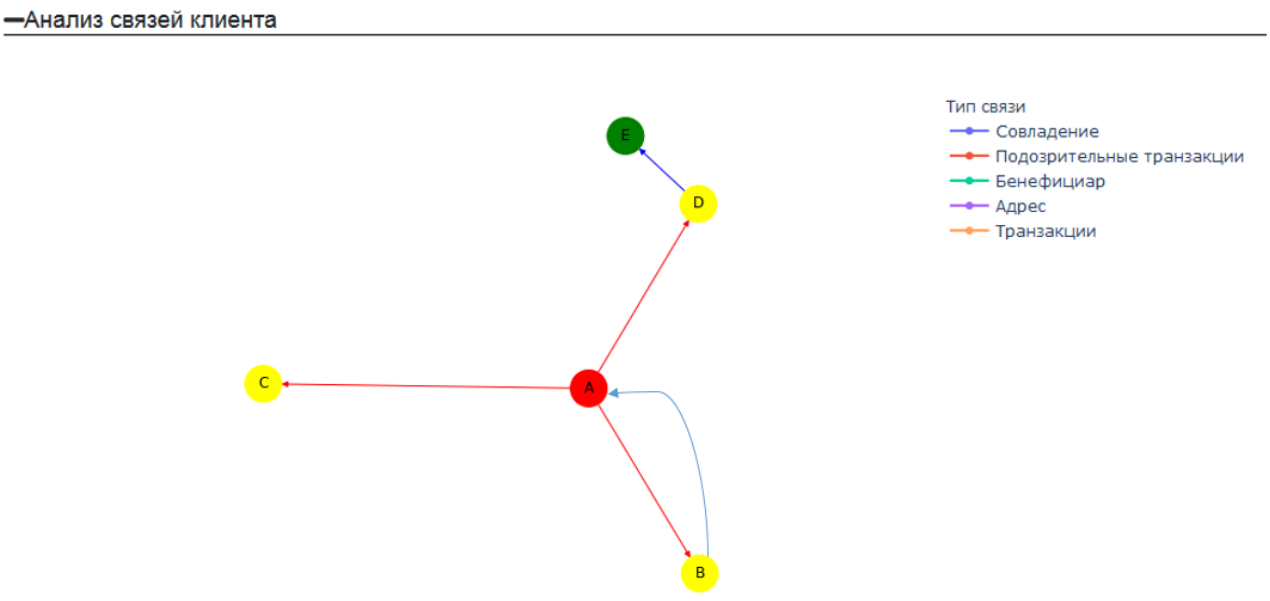

Рисунок 28. Анализ связей клиента на странице приложения (Автор, ФП «Модель АнтиФрод»)

Присутствует возможность выбора отображаемого типа связи графа из предложенных во второй главе данной работы для реализации. Также присутствует возможность менять положение вершины графа в пространстве, что особенно удобно при отображении связи по обычным транзакциям между компаниями

3.2 Оценка качественных показателей разработанного программного средства

Оценка качественных показателей программного средства позволяет определить удовлетворяет ли программное средство заданным потребностям в соответствии с его назначением. Для определения оценочных характеристик качества используется международный стандарт ISO 9126 (ГОСТ Р ИСО / МЭК 9126-93) – «Информационная технология. Оценка программного продукта. Характеристики качества и руководство по их применению». В данном стандарте выделяются шесть качественных характеристик программного обеспечения:

1. Эффективность – соотношение между уровнем качества функциональности программного средства и используемыми ресурсами в установленных условиях.
2. Надежность – свойство программного средства обеспечивать работоспособность за определенный период времени.
3. Функциональность – характеризует соответствие функциональных возможностей программного обеспечения требуемой пользовательской функциональности.
4. Практичность – характеристика определяющая сложность понимания, изучения и использования программного средства.
5. Мобильность – способность программного средства адаптироваться при изменении аппаратно-операционной среды.

6. Сопровождаемость –расположенность программного средства к изменениям (модификациям).

Основной характеристикой качества программного средства является надежность, так как к ней наблюдается устойчивый рост требований со стороны клиента. Несмотря на то, что программное средство проходило через несколько этапов тестирования, согласно требованиям международной системы контроля качества, возможно появление недочетов в надежности продукта.

Для определения надежности применяются модели, которые можно разделить на две группы: аналитические и эмпирические модели.

Эмпирические модели основываются на анализе структурных особенностей программного средства. Модели, которые относятся к данной группе не всегда дают конечных результатов показателей надежности, однако использование этих моделей на этапе проектирования программного средства полезно с точки зрения прогнозирования требуемых ресурсов и сроков завершения проекта.

Аналитические модели позволяют рассчитать качественные показатели надежности, основываясь на данных определенных в процессе тестирования. Аналитические модели разделяются на статические и динамические. Статические модели отличаются от динамических тем, что в них не учитывается время появления отказов в процессе тестирования. В свою очередь, в динамических моделях поведение программы рассматривается во времени.

Для определения показателя надежности системы поиска мошенников с банковскими кредитами среди юридических лиц была выбрана модель Коркорэна [35]. При этом в расчетах надежности участвовала и пользовательская часть системы. Применение модели предполагает знание следующих ее показателей:

- Модель содержит изменяющуюся вероятность отказов для различных источников ошибок разную вероятность их исправления.
- В модели используются только N испытаний, в которых наблюдается N_i ошибок i-го типа.

- Выявление ошибки i-го типа происходит с вероятностью a_i в ходе N испытаний.

Показатель уровня надежности R вычисляется по формуле (3.2.1), где $N0$-число безотказных испытаний, которые были выполнены в серии из N испытаний, k–известное число типов ошибок, Y_i–вероятность появления ошибок, причем при $N_i > 0$, $Y_i = a_i$, при $N_i = 0$, $Y_i = 0$.

$$R = \frac{N_0}{N} + \sum_{i=1}^{k} \frac{Y_i * (N_i - 1)}{N} \quad (3.2.1)$$

Таблица 10

Данные для расчета надежности разработанного ПС

Тип ошибки	Вероятность появления ошибки a_i	Число появления ошибок N_i при испытании	Y_i
Ошибки вычисления	0,09	7	0,09
Логические ошибки	0,26	25	0,26
Ошибки ввода/вывода	0,16	20	0,16
Ошибки манипулирования данными	0,18	19	0,18
Ошибки сопряжения	0,17	13	0,17
Ошибки определения данных	0,08	8	0,08
Ошибки БД	0,06	0	0

Оценка надежности разработанного программного средства в рамках выпускной квалификационной работы. В ходе тестирования было проведено 100 испытаний, 18 из которых прошли безуспешно, следовательно, N_0=82, N=100. В этой модели вероятность, a_i должна оцениваться на основании априорной информации или данных предшествующего периода функционирования однотипных ПС. Наиболее часто встречающиеся ошибки и вероятности их выявлений при тестировании ПС прикладного назначения приводятся в таблице 10 [35].

Соответственно, производя вычисления по формуле (3.2.1) используя данные из таблицы 10 получаем:

$$R = \frac{82}{100} + \frac{16,6}{100} = 0,985.$$

В итоге, после проведения 100 испытаний была получена вероятность безотказной работы равная 98,6%, что является высоким показателем надежности.

При оценки качественных показателей программного средства выбираются каждому из показателей качества устанавливаются веса w_i, такие что $\sum_{i=0}^{6} w(i) = 1$. Для каждого показателя устанавливается конкретная численная оценка riот 0 до 1, исходя из следующей классификации:

- 0 –свойство в программном средстве присутствует, но качество его не приемлемо.
- 0,5-1 –свойство в программном средстве присутствует и обладает приемлемым качеством.
- 1 –свойство в программном средстве присутствует и обладает очень высоким качеством.

В таблице 11 представлены качественные показатели в соответствии с их весами и оценками, установленными в ходе испытаний.

После получения данных и выставления для них весов можно получить взвешенную сумму всех весов $ПК_{пс}$, которая равна:

$$ПК_{пс} = \sum_{i=0}^{6} w(i) * r(i)$$

$$= 0.2 * 0.95 + 0.1 * 0.9 + 0.3 * 0.985 + 0.1 * 0.93 + 0.2 * 0.84$$

$$+ 0.01 * 1 = 0.9365$$

Таблица 11

Качественные показатели ПС

Показатели качества	Экспертная оценка (вес) w_i	Оценка, установленная экспериментом r_i
Портируемость	0,2	0,95
Сопровождаемость	0,1	0,9
Надежность	0,3	0,985
Эффективность	0,1	0,93
Функциональность	0,2	0,84
Практичность	0,1	1

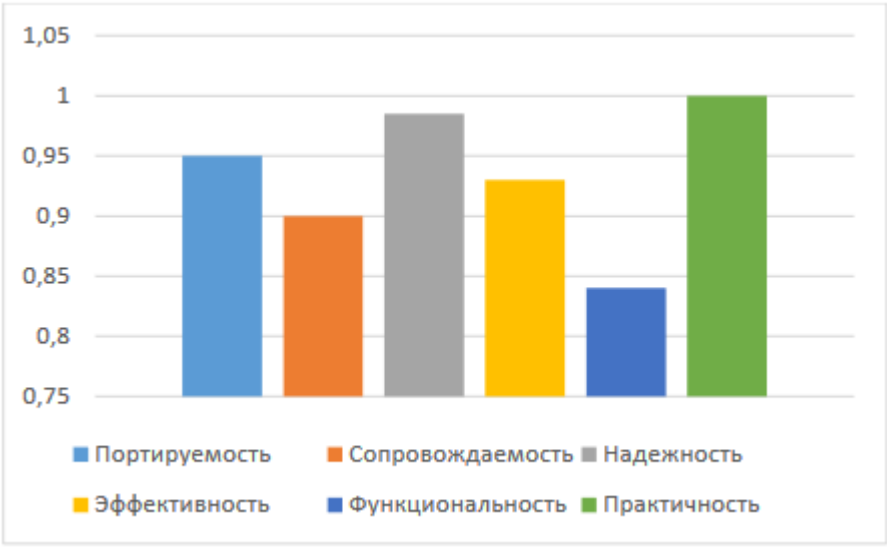

Рисунок 29. Диаграмма качественных показателей (Автор, MS Excel)

Подводя итог данного раздела, можно судить, что испытания программного средства в соответствие со стандартом ISO 9126 показали хорошие результаты качественных показателей, в частности, высокая надежность подтверждает

стабильность и отказоустойчивость программного средства. На рисунке 29 представлена диаграмма качественных характеристик.

3.3 Оценка экономической эффективности

Расчет экономической эффективности разработки программного средства заключается в сопоставление эффекта от внедрения ПС с затратами на его разработку, внедрение и сопровождение. Экономический эффект относится к основным показателям экономической эффективности. Экономический эффект – это результат после внедрения программного продукта, который выражается в стоимостной форме, в виде экономии от его осуществления.

Для расчета эффекта необходимо поставить дополнительный машинный эксперимент. В предыдущей главе данной работы проводилась общая оценка качества предложенных алгоритмов. На данном этапе необходимо провести оценку качества алгоритмов на моменте принятия решения о выдачи кредита и моменте принятия решения о способе урегулирования проблемной задолженности.

Ранее используемая метрика качества F-мера не подходит, так как она не интерпретируема с точки зрения финансовых показателей. При расчете необходимо использовать показатели:

- Positive Accuracy – показывает долю правильно определенных мошенников среди всех мошенников в выборке.
- Negative Accuracy – показывает долю правильно определенных «хороших» клиентов среди всех «хороших» клиентов.

Для проведения машинного эксперимента были собраны 150 клиентов, взявших кредит до 2019 года и по которым не было выявлено признаков мошенничества внутри банка. Также были взяты 26 клиентов, по которым были выявлены признаки мошенничества и которым было отказано в кредите, и 20 клиентов, у которых были выявлены признаки мошенничества после выдачи кредита,

но не более чем через один квартал. Данная выборка будет использоваться для оценки качества алгоритмов в моменте принятия решения о выдачи кредита.

Было собрано 150 клиентов, которые вышли на просрочку более 90 дней и им были приняты реструктуризации или другие кредитные стратегии до 2019 года, при этом данные заемщики не выходили на просрочку долее 5 дней в течение года. Были собраны 26 клиентов, по которым были выявлены признаки мошенничества и была введена процедура банкротства или исполнительного производства. А также были собраны 20 клиентов, по которым была принята кредитная стратегия, а после решения были выявлены признаки мошенничества в течении полугода. Данная выборка будет использоваться для определения качества на моменте принятия решения о стратегии урегулирования проблемной задолженности.

Стоит отметить, что в выборку не попали кредиты, выданные на государственные заказы и оборонно-промышленным комплексам. Так же соблюдалась пропорция по сегментам бизнеса: 20% крупный, средний бизнес, 80% микро-малый бизнес. Данные о клиенте были собраны в состояние на момент принятия решения о выдачи кредита и на момент принятия решения о способе урегулирования в соответствие с методом сбора выборки.

Далее, для каждого объекта из выборки были получены результаты алгоритмом, если хотя бы один из них показал, что клиент является мошенником, то данному объекту проставляется класс «мошенник». Если ни один алгоритм не классифицировал объект как мошенника, то данному объекту проставлялся класс «хороший». Стоит отметить, что для выборки, собранной для проверки качества алгоритмов в момент выдачи кредита, алгоритм анализа назначения платежа в транзакциях не применим, поэтому для него использовались только оставшиеся два алгоритма. А для выборки на момент принятия решение о стратегии урегулирования проблемной задолженности использовались все 3 алгоритма. В таблице 12 представлены результаты тестирования.

Таблица 12

Результаты тестирования алгоритмов для определения эффекта

Тип выборки	Класс	Positive Accuracy	Negative Accuracy
Выдача кредита	Хороший	-	0,93
	Мошенник пойманный	0,77	-
	Мошенник не пойманный	0,72	-
Урегулирование проблемной задолженности	Хороший	-	0,88
	Мошенник пойманный	0,85	-
	Мошенник не пойманный	0,75	-

Таким образом было получено качество алгоритмов, которое можно использовать для расчета эффекта. Такое качество представляет собой вероятность того, что алгоритмы правильно определят класс клиента. Так как банку скорее интересен эффект по портфелю в общем, а не по конкретным клиентам расчет эффекта будет ввестись по формуле 3.3.1.

$$E = \sum_{i=0}^{N_b} K_{bi} * w_{pi} - \sum_{i=0}^{N_g} P_{gi} * w_{ni} - \sum_{i=0}^{N_b} K_{bi} * (1 - w_{pi}) \quad (3.3.1)$$

Где E – экономический эффект, N_{bi}- количество мошенников, N_{gi}- количество «хороших» клиентов, K_{bi}- сумма, выданная мошеннику или реструктурированная сумма, P_{gi}- сумма процентов по кредиту, которую должен выплатить «хороший» клиент, w_{pi}- вероятность правильного определения мошенника, w_{ni} – вероятность правильно определить «хорошего» клиента.

Данный подход предполагает, что не будет происходить дополнительная проверка клиента сотрудниками банка, то есть решение будет опираться только

на результат алгоритмов. Соответственно, как прибыль банка от алгоритма берется сумма выданного мошеннику кредита или реструктурированного кредита в случае, если компания будет определена как мошенник правильно. Как потери от модели принимается сумма процентов (доход банка от выданного кредита), которую банк мог бы получить, если бы выдал кредит «хорошему» клиента и сумма реструктурированного кредита, если классификация мошенника прошла неправильно.

Отдельно был протестирован клиент ООО «УЭС». Для этого клиента в 2018 году была начата процедура банкротства и было введено внешнее управление, а в 2019 году внешний управляющий заявил о нахождение признаков преднамеренного банкротства. Общий долг ООО «УЭС» ПАО «Сбербанк» составляет 1 миллиард рублей, а процедура банкротства началась через 2 недели после выдачи нового кредита в 200 миллионов рублей.

Для расчета экономической эффективности разработанного программного средства, требуется определить затраты по базовому и проектному варианту. Расчет производится с помощью формул (3.3.2 –3.3.12).

Абсолютная прибыль от внедрения ΔЕ:

$$\Delta \mathrm{E} = E_a - E_b \tag{3.3.2}$$

где, E_a – прибыль по проектному варианту, E_b – прибыль по базовому варианту.

Коэффициент относительного повышения прибыли затрат K_{E}:

$$\mathrm{K_E} = \frac{\Delta \mathrm{E}}{E_b} * 100\% \tag{3.3.3}$$

Индекс повышения прибыли при повышении качества поиска мошенников Y_{E}:

$$\mathrm{Y}_E = \frac{E_a}{E_b} \tag{3.3.4}$$

Абсолютное снижение стоимостных затрат ΔC:

$$\Delta \mathrm{C} = C_b - C_a \tag{3.3.5}$$

где C_a– стоимостные затраты по проектному варианту; C_b– стоимостные затраты по базовому варианту;

Коэффициент относительного повышения стоимости затрат К$_C$:

$$\text{К}_C = \frac{\Delta \text{С}}{C_b} * 100\% \quad (3.3.6)$$

Индекс повышения производительности труда:

$$\text{Y}_C = \frac{C_a}{C_b} \quad (3.3.7)$$

Годовой экономический эффект:

$$\text{Э} = \Delta \text{Е} - \Delta \text{С} - T_H * \Delta \text{К} \quad (3.3.8)$$

где T_H – нормативный коэффициент эффективности капитальных вложений, значение которого принято брать равным 0.15, ΔK–капитальные вложения при переходе от базового варианта к проектному.

Капитальные вложения при переходе от базового варианта к проектному ΔK:

$$\Delta K = K_a - K_b \quad (3.3.9)$$

где K_a–капитальные затраты по базовому варианту; K_b–капитальные затраты по проектному варианту;

Капитальные затраты по проектному варианту K_b:

$$K_b = Z_p * N_p * t_p \quad (3.3.10)$$

где Zр–часовая оплата программиста; Nр–число программистов; tр–время на разработку программного средства в часах.

Коэффициент эффективности Т$_H$:

$$\text{Т}_H = \frac{\Delta E}{\Delta K} \quad (3.3.11)$$

Срок окупаемости затрат на внедрение программного средства:

$$\text{Т}_{OK} = \frac{1}{\text{Т}_H} \quad (3.3.12)$$

Для расчета трудозатрат были проигнорированы затраты на проверку клиента при принятии решения о выдачи кредита, так как банком было принято ре-

шение о необходимость сохранить методы проверки клиента до внедрения системы разработанной на основе приведенной методики. Данная система выступает только дополнительным инструментом для проверки.

В результате пилота было выяснено, что сотруднику необходимо потратить около 3 часа на проверку клиента, которого система определила, как мошенника, один час сотрудника оценивается в 568 рублей, при этом в год необходимо проверять около 300 клиентов в год. Таким образом, трудозатраты оцениваются в:

$$C'_a = 568 * 300 * 3 = 511200 \text{ рублей}$$

Трудозатраты по базовому варианту можно приравнять 0, так как выявление мошенника происходила в рамках других процессов, которые остались после внедрения.

В соответствии с формулой 3.3.1, был произведен расчет прибыли по базовому и проектному варианту. Для этого расчета были собраны все клиенты, подходящие под условия, которые были приведены ранее в данном пункте за 2018 год:

$$E_a = 2700000000 - 200000000 - 15000000 = 1350000000 \text{ рублей}$$

$$E_b = 1100000000 - 0 - 62000000 = 480000000 \text{ рублей}$$

Абсолютная прибыль от внедрения, в миллионах рублей:

$$\Delta E = 1350 - 480 = 870 \text{ млн. руб.}$$

Коэффициент относительного повышения прибыли затрат:

$$К_E = \frac{870}{480} * 100\% = 181\%$$

Индекс повышения прибыли при повышении качества поиска мошенников:

$$Y_E = \frac{1350}{480} = 2.8 \text{ раза}$$

Для расчета стоимостных показателей, рассчитаны стоимостные затраты за год. Пусть стоимость часа работы серверов составляет 1000 рублей. Во время

пилота было показано, что время, необходимое для расчета по всем клиентам, равно 3 часам:

$$C_a = C_a' + 1000 * 365 * 3 = 1146120 \text{ рублей}$$

Данные расчеты отражают стоимостные затраты на проведение требуемой работы по выполнению основных процессов.

Абсолютное снижение стоимостных затрат:

$$\Delta C = 1146120 - 0 = 1146120 \text{ рублей}$$

Коэффициент относительного снижения трудовых затрат:

$$К_C = \frac{1146120}{1146120} * 100\% = 100\%$$

Тогда повышение производительности труда:

$$Y_C = \frac{0}{1146120} = 0 \text{ раз}$$

Исходя из полученных данных, можно провести расчёт показателей эффективности программного средства. Для расчета годового экономического эффекта необходимо рассчитать капитальные затраты на внедрение. Трудоемкость разработки программного средства включает:

1. Изучение поставленной задачи: 40 часов.
2. Изучение имеющихся данных: 100 часов.
3. Разработка алгоритмов: 124 часа.
4. Создание прототипа: 172 часа.
5. Разработка архитектуры системы: 20 часов.
6. Разработка системы вычисления на основе Hadoop: 260 часов.
7. Разработка приложения в целевой системе: 132 часа.
8. Тестирование программы: 100 часов.
9. Документирование программы: 12 часов.

$$T = \sum_{i=1}^{9} t_i = 40 + 100 + 124 + 172 + 20 + 260 + 132 + 100 + 12$$
$$= 1220 \text{ часов}$$

Количество разработчиков 6, но разные пункты выполняли различные специалисты из команды:

1) С пункта 1 по пункт 4 задачи выполняли два специалиста по машинному обучению и анализу данных.
2) В пункт 5 задачи выполняли специалист по машинному обучению и анализу данных и специалист по BigData.
3) В пункт 6 задачи выполняли два специалиста по BigData.
4) С 7 по 9 пункт задачи выполняли специалисты по разработке программного обеспечения.

Для большей точности расчета предлагается внести дополнительные 40 часов на человека, отражающие консультацию специалистов одного профиля специалистами другого профиля.

Почасовая оплата специалиста оценивается в 2200 рублей в час, тогда затраты на внедрение равны:

$$K = 1220 * 2200 * 2 + 40 * 2200 * 6 = 5896000 \text{ рублей}$$

Годовой экономический эффект после внедрения программного обеспечения равен:

$$Э = 870000000 - 1146120 - \ 0{,}15 * 5896000 = 868765440 \text{рублей}$$

Расчетный коэффициент эффективности:

$$T_H = \frac{870000000}{5896000} = 147{,}5$$

Срок окупаемости затрат на внедрение проекта:

$$T_{OK} = \frac{1}{147{,}5} = 0{,}007 \text{ года} = 3 \text{ дня}$$

Действительно, учитывая, что даже клиенты из сегмента микро и малого бизнеса берут кредиты на несколько миллионов, трудозатраты на проект могут окупиться с первым пойманным мошенником. Но данный показатель не отображает реальную ситуацию, так как он рассчитан на некоторый непрерывный процесс, а мошенники возникают не регулярно и их количество не стабильно.

Исходя из полученных данных, внедрение программного средства позволит получить прибыль примерно равную миллиарду рублей в год. При этом, окупаемость проекта произойдет после первого пойманного с помощью данной системы мошенника.

3.4 Выводы по третьей главе

В данной главе была показана пользовательское приложение, которое использует результаты работы системы, разработанной по методике, которая представлена во второй главе данной работы. Данный пункт дает полное представление о работе системы, за счет приведенной в нем визуальной информации и описания работы приложения. Разработка такого приложения не была включена в методику, так как его функционал сильно зависит от требований конечного пользователя. Например, при урегулирование проблемной задолженности и принятие решения о выдаче кредита крупному и среднему бизнесу необходимо подробное описание, а при работе с микро и малым бизнесом может понадобиться только класс клиента, определенный системой поиска мошенников.

Далее был рассмотрено качество системы построенной на основе методики, которая была представлена во второй главе. При этом было показано высокое качество программного средства, в частности было показана высокая отказоустойчивость. Данный показатель является одним из самых важных, так как по действующей в банке методологии проведения релизов и восстановления работоспособности систем, развернутых на Hadoop может занимать от нескольких дней до месяца. Так, одна ошибка в процессе может остановить весь процесс расчета, тогда как неправильная работа одного из алгоритмов или некорректная обработка его результатов не приведет к полной остановке системы. Отображение результатов алгоритма, в котором была выявлена ошибка может быть приостановлено на стороне приложения пользователя, при это остальные алгоритмы продолжат свою работу.

Затем был произведен расчет экономического эффекта от внедрения данной системы. Несмотря на то, что разработка системы требует достаточно весомых трудозатрат, равно как и проверка результатов работы алгоритма требует введения дополнительных трудозатрат для пользователей, прибыль, которую будет приносить данная система является крайне весомой. Если совокупные годовые затраты в первый год, учитывая и затраты на разработку составят около 6 миллионов рублей, то прибыль оценивается примерно в миллиард рублей.

Таким образом была доказана надежность системы и был доказан существенный экономический эффект от внедрения системы.

ЗАКЛЮЧЕНИЕ

Данная работа посвящена разработке методики построения ИТ-инфраструктуры системы выявления мошенничества с банковским кредитами среди юридических лиц.

В итоге проведенного исследования получены следующие результаты:

1) Была обоснована целесообразность построения системы выявления мошенничества с банковскими кредитами среди юридических лиц.

На данный момент банки несут большие потери из-за мошенничества с кредитами. Такие потери являются как финансовыми и оцениваются миллиардами рублей, так и репутационными. Кроме того, от отсутствия систем по выявлению мошенничества страдает и рынок в целом из-за уменьшения аппетита к рискам у банков, как у крупнейших кредиторов. При этом в открытом доступе практически невозможно найти информацию о методах выявления мошенничества среди юридических лиц, хотя присутствует большое количество информации по поиску мошенников среди физических лиц.

Также на рынке отсутствуют компании, которые предлагают уже готовые решения для выявления мошенничества среди юридических лиц, хотя достаточно много компаний, которые предлагают подобные системы для физических лиц.

Таким образом, данная методика является актуальной на сегодняшний день. Учитывая финальные показатели качества и финансовый эффект системы, построенной на основе приведенной методики она может применяться банками для построения на ее основе собственных систем выявления мошенничества.

2) Была описана разработанная методика построения ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц.

Была проведен анализ источников данных, которые можно использовать для построения данной системы. На основе данного анализа были выделены три

класса источников на основе их надежности. Такой подход позволяет в дальнейшем строить систему с учетом надежности источника, что является критически важным для получения высококачественной системы. Результатом стало разделение источников данных на: «надежные», «внешние», «не надежные».

Была предложена методика построения трех алгоритмов для выявления мошенничества. Данные алгоритмы совместно покрывают все критические для выявления мошенничества точки во времени. Совместно разработанные алгоритмы используют все классы надежности данных, при этом учитывая к какому классу относится конкретный источник. Был проведен машинный эксперимент по результатам которого было показано, что предложенные алгоритмы обладают высокими показателями качества и могут использоваться в промышленном решение.

Далее была разработана методика построения архитектуры системы для выполнения предложенных алгоритмов. Данная архитектура базируется на распределенных вычислениях, что позволяет быстро производить расчеты на больших объемах данных. Также архитектура предлагает использование экосистемы Hadoop, что сильно уменьшает стоимость ее реализации, так как инструменты Hadoop являются программным обеспечением с открытым исходным кодом. Следовательно, из затрат на разработку исключается покупка лицензии на использование ПО. Еще одним плюсом предложенной архитектуры является использование ПО только из экосистемы Hadoop, что облегчает разработку, так как все интерфейсы взаимодействия между различными ПО хорошо проработаны и совместимы друг с другом.

3) Был показан порядок применения описанной методики ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц на примере системы «АнтиФрод».

Был показан способ применения результатов работы системы, построенной на основе приведенной во второй главе методики. Показан способ представ-

ления результатов пользователю, учитывающий особенности визуализации результатов каждого алгоритма. Представленная система является, по мнению автора, наиболее удобной для пользователя, и при этом, отражающей всю необходимую для анализа информаци.

4) Была проведена оценка технических параметров применения методики ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц.

Было показано, что данная система, разработанная на основе данной методики, обладает очень хорошими техническими показателями. На основе приведенного анализа можно судить, что испытания программного средства в соответствие со стандартом ISO9126 показали хорошие результаты качественных показателей, в частности, высокая надежность подтверждает стабильность и отказоустойчивость программного средства.

Как было показано, надежность программного средства при разработке системы на основе данной методики является критически важным показателем.

5) Была проведена оценка эффективности применения методики построения ИТ-инфраструктуры системы выявления мошенников с банковскими кредитами среди юридических лиц.

В рамках данной оценки был произведен дополнительный машинный эксперимент, который должен показать качество работы системы построенной на основе приведенной методики и использовался для дальнейшего расчета экономического эффекта.

На основе данного эксперимента было показано, что данная система может приносить около 1 миллиарда рублей в год, при затратах на разработку и первый год эксплуатации оцениваемых в 2 миллиона рублей. Но при этом эффект данной системы является не стабильным, так как появление мошенников среди клиента банка является случайным процессом. Таким образом, в один год система может принести только убытки, а в другой год может вовремя выявить клиента похо-

жего на ООО «Уралэлектрострой» и предупредить убытки на несколько миллиардов рублей. Таким образом, необходимо признать разработанную методику эффективной и рекомендовать к использованию в банковском секторе.

Была достигнута цель и выполнены все задачи, поставленные во введение данной работы.

СПИСОК ЛИТЕРАТУРЫ

1. ГОСТ Р ИСО/МЭК 9126-93 Информационная технология. Оценка программной продукции. Характеристики качества и руководства по их применению.
2. ГОСТ 28806-90. Качество программных средств. Термины и определения
3. ГОСТ 19.301-79 Программа и методика испытаний. Требования к содержанию и оформлению.
4. ГОСТ 28195-89. Оценка качества программных средств. Общие положения.
5. Архитектура современных распределённых систем [Электронный ресурс]. - Режим доступа - https://vk.cc/a4jhLG (дата обращения: 15.03.2020).
6. Байко С. Я. Мошенничество в сфере кредитования (статья 159.1 Уголовного кодекса Российский Федерации): вопросы квалификации / Байко С. Я. // Вестник Краснодарского университета МВД России – 2016 г. - № 2 (32) – 70-74 с.
7. Бринк Х., Ричардс Д., Феверолф М. Машинное обучение / Бринк Х., Ричардс Д., Феверолф М. // Издательский дом «Питер» 2017 г. – 61 - 64 с.
8. Бутенко Е. Д. Искусственный интеллект в банках сегодня: опыт и перспективы / Бутенко Е. Д. // Финансы и кредит – 2018 г. - № 1-24 – 143-153 с.
9. Власова Г. А. Анализ и оценка кредитоспособности заемщика, как получателя инвестиционного кредита на примере ОАО Сбербанк России, с использованием модели LGD / Власова Г. А. // Управление инвестициями и инновациями – 2017 г. - № 1 – 30-38 с.
10. Графовые базы данных [Электронный ресурс] – Режим доступа https://oracle-patches.com/db/3680-%D0%B3%D1%80%D0%B0%D1%84%D0%BE%D0%B2%D1%8B%D0%B5-%D0%B1%D0%B0%D0%B7%D1%8B-%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85 (Дата обращения 05.05.2020)

11. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / Гудфеллоу Я., Бенджио И., Курвилль А. // ДМК-Пресс 2017 г. – 1-652 с.
12. Дубов Е. И., Дубова М. Е. Мошенничество в сфере розничного и корпоративного кредитования: способы, противодействия и проблемы квалификации / Дубов Е. И., Дубова М. Е. // Вестник Санкт-Петербургского военного института войск национальной гвардии – 2018 г. - № 4 (5) – 84-89 с.
13. Дужий Т. А. Фиктивное или преднамеренное банкротство как способ выйти из кризиса или мошенничество «в рамках закона» / Дужий Т. А. // Шаг в науку – 2019 г. - № 3 – 84-86 с.
14. Единый государственный реестр индивидуальных предпринимателей [Электронный ресурс] – Режим доступа https://egrul.nalog.ru/index.html (Дата обращения 24.05.2020).
15. Единый государственный реестр недвижимости [Электронный ресурс] – Режим доступа https://rosreestr.ru/site/fiz/poluchit-svedeniya-iz-egrn/ (Дата обращения 25.05.2020).
16. Единый государственный реестр сведений о банкротстве [Электронный ресурс] – Режим доступа https://bankrot.fedresurs.ru/?attempt=1 (Дата обращения 21.05.2020).
17. Единый государственный реестр юридических лиц [Электронный ресурс] – Режим доступа https://egrul.nalog.ru/index.html (Дата обращения 24.05.2020).
18. Ермолова М. Д., Пеникас Г. И. PD-LGD correlation study: evidence from the Russian corporate bond market / Ермолова М. Д., Пеникас Г. И. // Model assisted statistics and applications– 2017 г. - № 4-12 – 335-358 с.
19. Замятина Е. В., Луценко А. В. Анализ значимости алгоритмов machine learning в антифрод – системах коммерческого банка / Замятина Е. В., Луценко А. В. // Материалы и меточы инновационных иссоедований и разработок – Уфа, 2018 г. – 129-131 с.

20. Заявление о признании сделки должника недействительной [Электронный ресурс] // Единый федеральный реестр сведений о банкротстве - Режим доступа https://bankrot.fedresurs.ru/MessageWindow.aspx?ID=9DD75CECD797C67B8B94B229494555FE&attempt=1 (Дата обращения 13.05.2020).
21. Земсков В. В., Соловьев А. И., Соловьев С. А. Модели оценки риска несостоятельности (Банкротства): история и современность / Земсков В. В., Соловьев А. И., Соловьев С. А. // Экономика, налоги, право – 2017 г. - № 6 - 40 – 91-100 с.
22. Иванова А. Б. Криминологическая оценка преднамеренного и фиктивного банкротства / Иванова А. Б. // Вестник экономики, права и социологии – 2018 г. - № 3 – 83-87 с.
23. Иванова Ю. К. Применение инструментов планирования и прогнозирования в деятельности коммерческого банка с использованием машинного обучения/ Иванова Ю. К. // Бенефициар – 2019 г. - № 46 – 22-26 с.
24. Кайбасова Д. Ж. Извлечение статистических данных для определения уникальности документов на основе анализ контента учебных программ дисциплин / Кайбасова Д. Ж. // The Scientific Heritage – 2020 г. – № 44-1 – 57 – 62 с.
25. Картотека арбитражных дел [Электронный ресурс] – Режим доступа https://kad.arbitr.ru/ (Дата обращения 21.05.2020).
26. Ковальчук А. В., Мальцев Е. Г. Международный опыт микрофинансовых организаций в сфере онлайн-кредитования с возможностью его применения в России / Ковальчук А. В., Мальцев Е. Г. // Актуальные вопросы современной экономики – 2019 г. - № 4 – 479-483 с.
27. Компания «Деньги взаймы» внедрила межбанковскую систему «НБКИ-AFS» для борьбы с кредитным мошенничеством [Электронный ресурс] / Национальное бюро кредитных историй - Режим доступа https://www.nbki.ru/company/news/?id=21599 (Дата обращения 21.05.2020).

28. Кондорова Т. И., Татаровский Ю. А. Перспективы использования нефинансовой информации в диагностике банкротства бизнеса / Кондорова Т. И., Татаровский Ю. А. // Проблемы развития предприятий: теория и практика – 2018 г. - № 3 – 105-109 с.
29. Копать Д. Я., Маталыцкий М. А. Анализ ожидаемых доходов в открытых марковских сетях с различными особенностями / Копать Д. Я., Маталыцкий М. А. // Вестник Томского государственного университета. Управление, Вычислительная техника и информатика. – 2020 г. - № 50 – 31-38 с.
30. Критерии для исключения систематических погрешностей [Электронный ресурс] – Режим доступа https://studme.org/16581005/tovarovedenie/kriterii_dlya_isklyucheniya_sistematicheskih_pogreshnostey (Дата обращения 05.06.2020).
31. Малахова А. А., Светличная А. В. Классификация преступлений о незаконном получении кредита / Малахова А. А., Светличная А. В. // Управление в условиях глобальных мировых трансформаций: экономика, политика, право – 2019 г. – 299-301 с.
32. Малышенко В. А., Малышенко К. А., Кругликова Е. Ю. Модели комплексной оценки финансовой устойчивости предприятия / Малышенко В. А., Малышенко К. А., Кругликова Е. Ю. // Финансово-экономическое и информационное обеспечение инновационного развития региона – 2020 г. - 114-116 с.
33. Мансурова А. С. Сравнительный анализ методов проверки гипотезы об отсутствии тренда во временном ряду. / Мансурова А. С. // Актуальные проблемы экономики современной России – 2016 г. - №3 – 516-522 с.
34. Международные стандарты финансовой отчетности и разъяснение к ним [Электронный ресурс] – Режим доступа http://www.consultant.ru/document/cons_doc_LAW_140000/ (Дата обращения 05.05.2020).

35. Модель Коркорэна [Электронный ресурс] - Режим доступа https://mydocx.ru/2-105752.html (Дата обращения 10.06.2020).
36. Национальное бюро кредитных историй [Электронный ресурс] – Режим доступа https://www.nbki.ru/poleznaya-informatsiya/ (Дата обращения 22.05.2020).
37. Нидхем М., Ходлер Э. Графовые алгоритмы. / Нидхем М., Ходлер Э. // ДМК-Пресс 2020 г. – 31-44 с.
38. Орлов С. Н., Татаринцев А. В. Экономическая безопасность бизнес-процессов банковского кредитования в условиях цифровой экономики/ Орлов С. Н., Татаринцев А. В. // Разработка стратегии социальной и экономической безопасности государства – 2019 г. – 441-444 с.
39. Первое правило антифрода – никому не рассказывай про антифрод [Электронный ресурс] // Хабр - Режим доступа https://habr.com/ru/company/rbkmoney/blog/477950/ (Дата обращения 02.04.2020).
40. Пилипенко А. С., Коломойцева И. А. Определение тональности текста на основе модели «Bag-Of-Words» / Пилипенко А. С., Коломойцева И. А. // Сборник материалов XI Международной научно-технической конференции в рамках VI (Международного Научного форума Донецкой Народной Республики) – Донецк, 2020 г. – 77 – 81 с.
41. Поиск аномалий [Электронный ресурс] / Анализ малых данных – Режим доступа https://dyakonov.org/2017/04/19/%D0%BF%D0%BE%D0%B8%D1%81%D0%BA-%D0%B0%D0%BD%D0%BE%D0%BC%D0%B0%D0%BB%D0%B8%D0%B9-anomaly-detection/ (дата обращения 17.05.2020)
42. Преднамеренное и фиктивное банкротство [Электронный ресурс] / Банкротство в России – Режим доступа http://dolgnikov.net/%D0%BF%D1%80%D0%B5%D0%B4%D0%BD%D0%

B0%D0%BC%D0%B5%D1%80%D0%B5%D0%BD%D0%BD%D0%BE%D0%B5-%D0%B1%D0%B0%D0%BD%D0%BA%D1%80%D0%BE%D1%82%D1%81%D1%82%D0%B2%D0%BE/ (дата обращения 17.04.2020).

43. Различные стратегии сэмплинга в условиях несбалансированности классов [Электронный ресурс] / BaseGroup Lab – Режим доступа https://basegroup.ru/community/articles/imbalance-datasets (Дата обращения 04.05.2020)
44. Результаты процедур в делах о банкротстве за 2019 год [Электронный ресурс] // Федресурс – Режим доступа https://fedresurs.ru/news/d9263eb1-10a9-43db-8755-3dc0add94bd3 (Дата обращения 05.04.2020).
45. Сбербанк для корпоративных клиентов [Электронный ресурс] – Режим доступа https://www.sberbank.ru/ru/legal/credits_new (Дата обращения 04.05.2020).
46. Сообщение о наличии или об отсутствии признаков преднамеренного или фиктивного банкротства [Электронный ресурс] // Единый федеральный реестр сведений о банкротстве - Режим доступа https://bankrot.fedresurs.ru/MessageWindow.aspx?ID=3E65F1368DCE435A1684A7E410627B33 (Дата обращения 13.05.2020).
47. Сообщение о судебном акте [Электронный ресурс] // Единый федеральный реестр сведений о банкротстве - Режим доступа https://bankrot.fedresurs.ru/MessageWindow.aspx?ID=511B8333AB271199EED44BA09FC7B3BB (Дата обращения 13.05.2020).
48. СПАРК [Электронный ресурс] – Режим доступа https://www.spark-interfax.ru/ru/integration (Дата обращения 25.05.2020).
49. Тамер О. С. Аддитивные и мультипликативные модели временных рядов в инвестиционном проектировании / Тамер О. С. // Вестник Волжского университета им. В. Н. Татищева – 2020 г. - № 1-2 – 227-233 с.

50. Технологии Big Data – [Электронный ресурс] / Digital-агенство Uplab. – 2006–2019. – Режим доступа https://www.uplab.ru/blog/big - data-technologies/ (дата обращения 17.04.2020).
51. «Уралэлектрострой» рано объявлять банкротом [Электронный ресурс] // Комсомольская правда - Режим доступа https://www.msk.kp.ru/daily/26958/4011896/ (Дата обращения 13.04.2020).
52. Федеральная налоговая служба [Электронный ресурс] – Режим доступа https://www.nalog.ru/rn50/ (Дата обращения 23.05.2020).
53. Чистяков М. Е. Метод оценки кредитного риска / Чистяков М. Е. // Системный анализ в науке и образовании – 2012 г. - № 4 – 148-157 с.
54. Что такое РСБУ в бухгалтерии [Электронный ресурс] – Режим доступа https://ppt.ru/art/buh-uchet/rsbu (Дата обращения 05.05.2020).
55. Щетинина Н. В., Кокорин Д. Л. Особенности квалификации мошенничества в сфере кредитования / Щетинина Н. В., Кокорин Д. Л. // Вестник Воронежского государственного университета – 2018 г. - № 2 (32) – 191-196 с.
56. Шимеева Ж. Ш., Токторова А. Э. Криминалистический анализ обстановки совершения незаконного получения и нецелевого использования кредита / Шимеева Ж. Ш., Токторова А. Э. // Наука, новые технологии и инновации – 2017 г. - № 3 – 129-132 с.
57. Шумилина В. Е., Цвиль М. М. Построение модели регрессии по временным рядам с целью прогнозирования индекса производительности труда в российской федерации / Шумилина В. Е., Цвиль М. М. // Вестник евразийской науки – 2020 г. - № 1-12 – 73-72 с.
58. Шунина Ю. С., Клячкин В. Н. Прогнозирование платежеспособности клиентов банка на основе методов машинного обучения и марковских цепей / Шунина Ю. С., Клячкин В. Н // Программные продукты и системыт – 2016 г. - № 2 – 105-112 с.

59. Якунина А. В., Якунин С. В. Некоторые аспекты цифровизации кредитного риска / Якунина А. В., Якунин С. В. // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками – 2018 г. - № 3 – 277 – 281 с.
60. Apache Kafka: обзор [Электронный ресурс] / Хабр – Режим доступа https://habr.com/ru/company/piter/blog/352978/ (дата обращения: 10.03.2020)
61. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce [Электронный ресурс]// Хабрахабр – Режим доступа https://habrahabr.ru/company/dca/blog/267361/ (дата обращения: 10.03.2020).
62. Daniel Cohen, Scott M. Jordan, W. Bruce Croft. Learning a better negative sampling policy with deep neural networks for search / Daniel Cohen, Scott M. Jordan, W. Bruce Croft // Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval - Santa Clara CA USA, 2019 г. – 19–26 с.
63. Jinghui Chen, Saket Sathe, Charu C. Aggarwal. Outlier detection with autoencoder ensembles / Jinghui Chen, Saket Sathe, Charu C. Aggarwal, Deepak S. Turaga // SDM – 2017 г. – 90-98 с.
64. Hadoop для сетевых инженеров. [Электронный ресурс]// Хабр. – Режим доступа https://habr.com/ru/company/cisco/blog/245339/ (Дата обращения 23.06.2020)
65. Hadoop, часть 1: развертывание кластера [Электронный ресурс]// Selectel – Режим доступа https://selectel.ru/blog/hadoop-chast-1-razvertyvanie-klastera/ Дата обращения 23.06.2020)
66. Medetov, A. A. Term Big Data and methods of its application // Young Scientist. – 2017 г. – № 11. – 207–210 с.
67. Mingsheng Long, Zhangjie Cao, Jianmin Wang. Learning multiple tasks with multilinear relationship networks. / Mingsheng Long, Zhangjie Cao, Jianmin

Wang, Philip S. Yu. // NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems – 2017 г. – 1593-1602 с.

68. Razvan-Gabriel Cirstea, Darius-Valer Micu, Gabriel-Marcel Muresan. Correlated time series forecasting using multi-task deep neural networks. / Razvan-Gabriel Cirstea, Darius-Valer Micu, Gabriel-Marcel Muresan, Chenjuan Guo, and Bin Yang. // CIKM17: Conference on Information and Knowledge Management – 2018 г. – 1527-1530 с.
69. Tretyak E. Measuring similarity of texts base on distributional semantic models (case study of the Russian original text and English translations of M. Bulgakov's Novel "The Master and Margarita") / Tretyak E. // International journal of open information technologies – 2020 г. - № 8-1 – 17-26 с.
70. Tung Kieu, Bin Yang, Christian S. Jensen. Outlier detection for multidimensional time series using deep neural networks. / 5. Tung Kieu, Bin Yang, Christian S. Jensen // 19th IEEE International Conference on Mobile Data Management – 2018 г. – 125-134 с.