Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа экономики и управления
Кафедра «Информационные технологии в экономике»

ПРОЕКТ ПРОВЕРЕН
Рецензент, ген. директор
ООО «ТРАНС-ЛАЙН»
_____(В.Ф. Бурдин)
«____» _____ 2020 г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой, д.т.н., с.н.с,
_____(Б.М. Суховилов)
«____» _____ 2020 г.

Разработка математических моделей для решения задачи классификации потенциальных кредитополучателей в целях уменьшения рисков для банка

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ–38.04.05.2020.456.ПЗ ВКР

Руководитель работы, д.т.н., профессор
_____(В.В. Мокеев)
«____» _____2020г.

Автор работы,
студент группы ЭУ-226
_____(А.Р. Магсумов)
«____» _____2020 г.

Нормоконтролер, к.т.н., доцент
_____(Е.В. Бунова)
«____» _____2020 г.

Челябинск 2020

АННОТАЦИЯ

Магсумов А. Р. Разработка математических моделей для решения задачи классификации потенциальных кредитополучателей в целях уменьшения рисков для банка – Челябинск: ЮУрГУ, ЭУ-226, 2020, 89 с., 37 ил., 17 табл., библиогр. список – 81 наим., 0 прил.

Сфера машинного обучения из дня в день занимает всё большее место в нашей жизни ввиду широких возможностей для его применения. Анализ дорожного трафика, спам-фильтры, самоуправляемые автомобили, анализ массивов данных и т. д. – все больше задач перекладывается на

Порой, мы даже не замечаем «присутствие» машинного обучения в нашей жизни, но тем не менее, решения, которые предлагает данная сфера – ежедневно вокруг нас. Как уже было сказано выше, такое понятие, как спам-фильтр, уже стало для нас чем-то привычным, обыденным; при этом именно методы машинного обучения решают, какое письмо отнести к спаму, а какое – нет.

Основная идея машинного обучения заключается в том, чтобы компьютер, без специфичного усиленного кодинга с нашей стороны, сам обучился решению поставленной перед ним задачи, находя зависимости, закономерности, взаимосвязи и т. д. среди предоставленных ему данных.

Машинное обучение является ветвью искуственного интеллекта.

В формате решения задачи от Home Credit Bank, предложено использовать методы машинного обучения для получения результата.

Необходимо, через анализ данных, создать решение, которое могло бы спрогнозировать кредитный риск у отдельного кредито-получателя.

Основные задачи работы:

1) описание процесса предкредитного анализа клиента банком;

2) теоретическое описание машинного обучения, его виды и алгоритмы;

3) анализ и сравнение имеющихся прогнозных методов машинного обучения;

4) формулировка требований к прогнозной модели;

5) построение прогнозной модели;

6) сравнительный анализ результатов прогнозирования;

7) формирование плана коммерциализации.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы.

В настоящее время существует много рисков, связанных с банковскими кредитами, особенно для банков, связанных с потерей капитала. В связи с этим анализ рисков и оценка дефолта становится критически важными. Банки хранят огромные объемы данных о поведении клиентов, из которых они не могут прийти к суждению, может ли заявитель быть неплательщиком или нет, будут ли проблемы с возвратом кредита или нет.

Для банковской сферы применение методов машинного обучения показало высокую эффективность. В частности, применительно к данной работе, анализ больших массивов данных и разработка модели прогнозирования существенно снижают нагрузку на персонал банка, уменьшает время обработки информации и вынесения решения, в целом – уменьшает издержки и практически исключает фактор человеческой ошибки.

В общем – значительно упрощает и автоматизирует процесс анализа и вынесения решения, что в текущей мировой конъюнктуре, в которой популярность банковской сферы, и в целом клиентооборот банков, значительно увеличились, может сыграть серьезную роль в плане повышения конкурентоспособности банка и повышения экономической стабильности через уменьшение потенциальных рисков.

Оценка кредитного риска является критически важной проблемой, с которой сталкиваются банки в настоящее время, и которая помогает им оценить, может ли соискатель кредита быть неплательщиком на более позднем этапе, чтобы они могли принять решение – предоставить кредит или нет. Это помогает банкам минимизировать возможные убытки и может увеличить объем кредитов. Результатом этой оценки кредитного риска будет прогноз категории заявителя – 0 или 1. Следовательно, становится важным создать модель, которая будет учитывать различные аспекты заявителя. Это поможет банку решить, могут ли они предложить кредит заявителю или нет.

В таком сценарии анализируемые данные огромны и сложны, и использование методов интеллектуального анализа данных для получения результата является наиболее подходящим вариантом при условии его эффективной аналитической методологии, которая находит полезные знания.

Целью данной работы является предложение модели анализа данных с использованием методов машинного обучения для прогнозирования категорий для новых заявителей на получение кредита в банке.

Данные, используемые для анализа, содержат много несоответствий, таких как отсутствующие значения, выбросы и несоответствия, и их необходимо обработать перед использованием для построения модели. Лишь немногие из параметров клиента действительно способствуют прогнозированию неплательщика. Таким образом, эти параметры или функции должны быть определены до применения модели.

В ходе выполнения работы, будут опробованы разные методики машинного обучения. В результате – будет определена лучшая.

Теоретической и методологической основой магистерской диссертации являются труды зарубежных и отечественных ученых в области машинного обучения. Так, например в российской литературе известны такие авторы как: Вьюгин В.В.[1], Николенко С.И.[2], Барскир А.Б.[3], Матвеев А.С.[4], Ручкин В.Н.[5], В.А. Фулин[5], Аксенов С.В.[6], Новосельцев В.Б.[6], Воронцов К.В.[7], Лепский А.Е.[8], Броневич А.Г.[8]

В зарубежной литературе: Я.Гудфеллоу[9], И.Бенджио[9], А.Курвилль [9], А. Джули[10], Суджит Пол[10], Мохамед Али,[11] Арно Мейсман[11], Дэви Силен[11], Андреас Мюллер[12], Сара Гвидо[12], Себастьян Рашка[13].Джоэл Грас[13], Дж. Вандер Плас[14],

Рассмотрев различные точки зрения в отечественной и зарубежной литературе, можно прийти к выводу, что существуют различные методы прогнозирования, каждый из которых имеет свои преимущества и недостатки.

Объект исследования – процесс предкредитной оценки и анализа потенциальных кредитополучателей банка.

Предмет исследования – методы прогнозирования вероятности возврата кредита клиентами банка.

Цель исследования – совершенствование методов предкредитной оценки кредитополучателя на основе анализа данных о предшествующих кредитах других клиентов банка, для минимизации рисков и уменьшения объема не возвращенных денежных средств.

Для достижения цели были поставлены **следующие задачи**:

1) описание понятия предкредитного анализа.

2) провести исследование и анализ современных подходов к проведению предкредитного анализа.

3) анализ имеющихся прогнозных методов машинного обучения.

4) формулировка требований к прогнозной модели.

5) построение прогнозной модели.

6) сравнительный анализ результатов прогнозирования методами машинного обучения.

7) разработка плана коммерциализации.

Новизна работы заключается в том, что на основе комплексного анализа:

1) рассмотрены и проанализированы методы прогнозирования.

2) проведен сравнительный анализ методов.

3) разработана модель для решения задачи классификации потенциальных кредитополучателей в целях уменьшения рисков для банка.

Практическая значимость работы обусловлена применением результатов исследования на практике, для решения задачи классификации потенциальных кредитополучателей в целях уменьшения рисков для банка.

ГЛАВА 1. БИЗНЕС ПРОЦЕСС КРЕДИТОВАНИЯ КЛИЕНТОВ БАНКАМИ

1.1 Понятие кредитного риска

Кредитный риск – это риск неуплаты долга, который может возникнуть из-за того, что заемщик не может произвести требуемые платежи.

В первую очередь, риск несет кредитор и включает потерю основного долга и процентов, нарушение денежных потоков и увеличение затрат на организацию возврата долга.[78] Потеря может быть полной или частичной. На эффективном рынке более высокий уровень кредитного риска будет связан с более высокой стоимостью заимствования.

В целях уменьшения кредитного риска для кредитора, он может выполнить проверку кредитоспособности потенциального заемщика.[78]

Предсказать многие факторы, которые могут повлиять на способность возврата кредита заемщиком, невозможно, но можно на основе анализа уже имеющейся, постоянно пополняющейся информации, попытаться выявить закономерности, особенности, зависимости, между некоторыми «показателями» кредитозаемщика, например, возраст, размер заработной платы, трудовой стаж и др.

На основе такой информации, банк сможет на раннем этапе определить существует ли потенциальный риск невозврата кредита конкретным заемщиком, что позволит, уменьшить объем невозвращенных средств и невыплаченных по ним процентов.

Кредитные риски являются наиболее частой причиной банкротств банков, в связи с чем все регулирующие органы устанавливают стандарты по управлению кредитными рисками. Несмотря на инновации в секторе финансовых услуг, кредитный риск до сих пор остаётся основной причиной банковских проблем. Более 80 % содержания балансовых отчётов банков посвящено именно этому аспекту управления рисками.

В основе процедур оценки кредитных рисков лежат следующие понятия:

1) вероятность дефолта – вероятность, с которой дебитор в течение некоторого срока может оказаться в состоянии неплатёжеспособности;

2) кредитный рейтинг – классификация дебиторов организации, контрагентов эмитентов ценных бумаг или операций с точки зрения их кредитной надёжности;

3) кредитная миграция – изменение кредитного рейтинга дебитора, контрагента, эмитента, операции;

4) сумма, подверженная кредитному риску – общий объём обязательств дебитора, контрагента перед организацией, сумма вложений в ценные бумаги эмитента и т. д.;

5) уровень потерь в случае дефолта – доля от суммы, подверженной кредитному риску, которая может быть потеряна в случае дефолта.

Собственно оценка кредитного риска может производиться с двух позиций: оценка кредитного риска отдельной операции и портфеля операций.

Базовая оценка (без учёта миграции) кредитного риска отдельной операции может производиться с различным уровнем детализации:

1) оценка суммы, подверженной риску;
2) оценка вероятности дефолта;
3) оценка уровня потерь в случае дефолта;
4) оценка ожидаемых и неожиданных потерь.

1.2 Организация процесса кредитования в банках

Этапы процесса кредитования

Кредитование условно можно разделить на несколько этапов, на каждом из которых уточняются характеристики ссуды, способы ее выдачи и погашения:

1) рассмотрение кредитной заявки и собеседование с клиентом;
2) изучение кредитоспособности клиента;
3) подготовка и заключение кредитного договора, выдача кредита;
4) формирование резерва на возможные потери по ссудам;

5) контроль банка за выполнением условий договора и погашением кредита (сопровождение кредита);

6) работа банка с проблемными ссудами.

Подготовка и заключение кредитного договора

Решение о целесообразности выдачи кредита принимается либо уполномоченным должностным лицом, либо соответствующим органом управления банка на основе имеющейся информации.

Цель кредита

Первый вопрос, который интересует банк, – для чего берется ссуда. Цель кредита служит важным индикатором степени риска.

Сумма кредита

Банк должен проверить обоснованность заявки в отношении суммы кредита. Важно с самого начала правильно определить требуемую сумму кредита.

Погашение кредита

При выдаче кредита должен быть ясно определен его источник погашения. Есть два главных источника: за счет поступления доходов или от продажи активов. Банк должен проверить, соответствуют ли условия, предложенные клиентом, его реальным возможностям.

Срок ссуды

Чем более продолжителен срок ссуды, тем выше риск, тем больше вероятность того, что возникнут непредвиденные трудности и клиент не сможет погасить долг в соответствии с договором.

Обеспечение

Важным элементом кредитной сделки является то, какие активы заемщик сможет заложить в качестве обеспечения, кто владелец обеспечения, место нахождения обеспечения, издержки на хранение, как оценено имущество, предлагаемое в качестве обеспечения.

Процентная ставка

Процентная ставка определяется конкретным банком и зависит от множества разных параметров.

1.3 Управление кредитными рисками

Кредитный анализ – это метод, с помощью которого можно рассчитать кредитоспособность бизнеса или организации. Другими словами, это оценка способности компании или человека выполнять свои финансовые обязательства.

Целью кредитного анализа является рассмотрение как заемщика, так и предлагаемой кредитной линии, и присвоение рейтинга риска. Рейтинг риска определяется путем оценки вероятности дефолта заемщика с заданным уровнем достоверности в течение срока службы объекта и путем оценки суммы убытков, которые кредитор понесет в случае дефолта.

Кредитный анализ включает в себя широкий спектр методов финансового анализа, включая анализ коэффициентов и тенденций, а также составление прогнозов и подробный анализ потоков денежных средств.

1.4 Подходы к анализу и оценке кредитоспособности клиента

Проблема определения кредитоспособности кредитополучателя актуальна с момента возникновения банков. В разные периоды развития и в разных странах к данной проблеме подходили по-разному. Кредитоспособность (англ. creditabllity) – способность кредитополучателя получить кредит, а также своевременно и полностью рассчитаться по долгам. В широкой банковской практике кредитоспособность потенциального кредитополучателя – главный критерий при определении целесообразности и формы кредитных отношений.

Одним из самых важных этапов в организации процесса кредитования является оценка кредитоспособности клиента. Неправильная оценка кредитоспособности может привести к не своевременному возврату кредита, что в свою очередь способно нарушить ликвидность банка и, в конечном счете, привести к банкротству кредитной организации. Банки придают

огромное значение разработке современной методологической базы оценки кредитоспособности, тестированию кредитных работников, а также совершенствованию система контроля и оценки кредитных рисков. Проблема выбора показателей для оценки способности кредитополучателя выполнять свои обязательства была актуальна во все периоды развития банковского дела, и вошла в экономическую литературу как проблема определения кредитоспособности. Основные цели и задачи оценки кредитоспособности потенциального кредитополучателя представлены на рисунке.

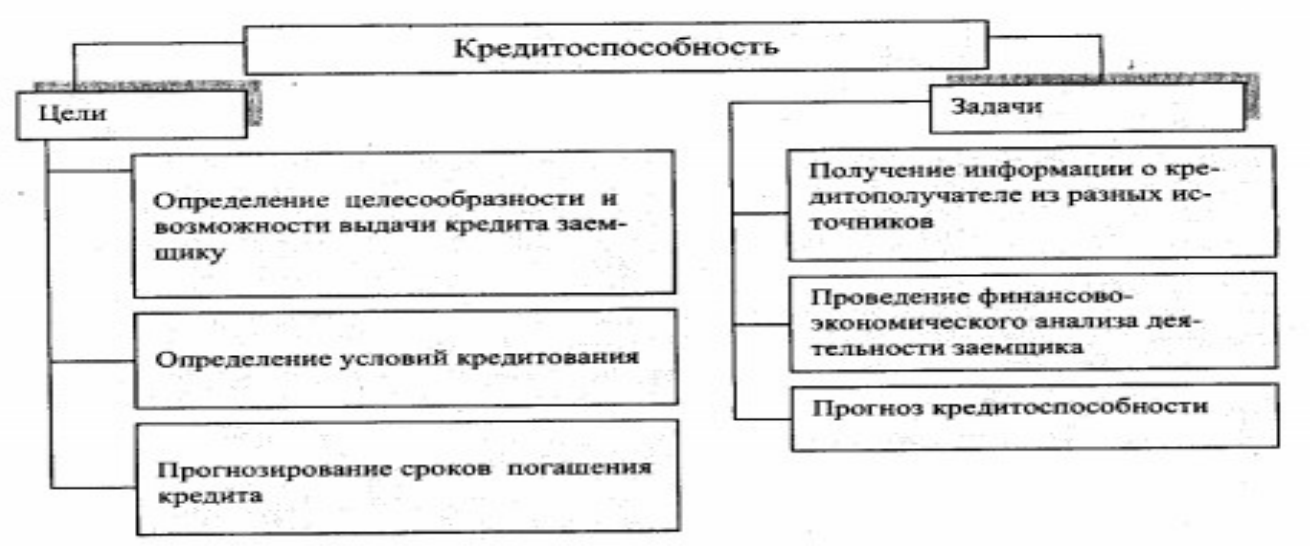

Рисунок 1 – Основные цели и задачи оценки кредитоспособности

Способами оценки кредитоспособности клиента банка являются:

1) способ коэффициентов;
2) рейтинговая оценка деятельности;
3) кредитный скоринг;
4) анализ денежного потока;
5) расчет комплексного коэффициента кредитоспособности;

6) оценка эффективности проекта и др.

Рассмотрим классификацию методов и моделей оценки кредитоспособности кредитополучателей банков (рисунок 2).

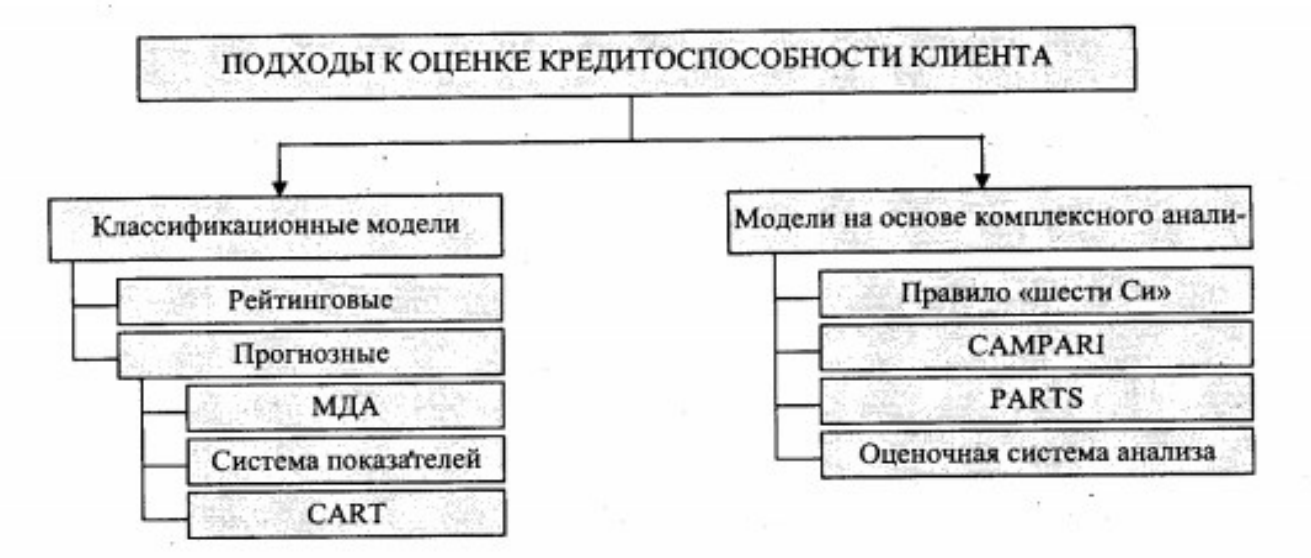

Рисунок 2 – Методы определения кредитоспособности клиента

1.5 Скоринговые модели как средство управления кредитными рисками в банках

Кредитный скоринг — система оценки кредитоспособности (кредитных рисков) лица, основанная на численных статистических методах. Скоринг заключается в присвоении баллов по заполнению некой анкеты, разработанной оценщиками кредитных рисков андеррайтерами. По результатам набранных баллов системой принимается решение об одобрении или отказе в выдаче кредита. Данные для скоринговых систем получаются из вероятностей возвратов кредитов отдельными группами заёмщиков, полученными из анализа кредитной истории тысяч людей. Считается, что

существует корреляция между определенными социальными данными (наличие детей, отношение к браку, наличие высшего образования) и добросовестностью заемщика.

Скоринговые модели используются кредитными организациями для определения кредитоспособности потенциального заемщика. На вход такой модели подаются определенные характеристики клиента (возраст, доход, стаж работы и т.д.), а на выходе формируется интегрированный показатель (score), который определяет вероятность возврата или невозврата кредита. Скоринговая модель является главным инструментом кредитного скоринга. Фактически она связывает параметры клиента с суммой, которую можно выдать ему, или степенью кредитного риска в конкретных условиях через систему скоринговых баллов.

Скоринговые системы делятся на несколько видов, самые популярные из них:

• Скоринг на основании кредитной истории учитывает кредитную историю потенциального заемщика (ссуды различных банков, просрочки (если они есть), попытки взять ссуду у банка, наличие кредитных карт). Главный недостаток такого подхода очевиден. Выборка классифицируется только по клиентам, которым уже давали кредит. Оставалось неизвестным, как повели бы себя те клиенты, которым было отказано в кредите или которые за ним даже не обращались.

• Социо-демографический скоринг – это оценка заемщика на основании таких показателей как возрастной и половой признак, семейное положение, наличие иждивенцев, стаж работы, профессия. Также скоринг учитывает уровень заработной платы, обычно, за последний год.[47] Кредитный менеджер проводит собеседование и анкетирование с потенциальным заемщиком, после чего вносит данные в программу. На основании этих данных скоринговая система присваивает баллы за каждый фактор, а в конце процедуры относит заемщика к определенной «группе риска» и дает

заключение о возможности предоставления кредита. Помимо этого, сотрудники могу визуально оценивать потенциального контрагента, чье девиантное или даже неадекватное поведение или неприемлемый внешний вид может стать преградой к получению ссуды. После всех, вышеизложенных методов оценки, заявка должна пройти этап андеррайтинга и получить одобрение у департамента анализа рисков и службы безопасности. В случае, если все этапы пройдены, это еще не гарантия получения кредита.

Управление кредитными рисками занимает отдельное место в эффективном менеджменте любого банка. Под кредитным риском подразумевается невыполнение контрагентом своих кредитных обязательств по тем или иным причинам. Наиболее распространенный кредитный риск – дефолт заемщика, когда контрагент не выполняет обязательства перед банком по возврату денежных средств согласно условиям кредитного соглашения в силу экономической неспособности или нежелания.[41]

Таблица 1 – Внутренние факторы кредитного риска [41]

Внутренние факторы	Характеристика факторов кредитного риска
Факторы, связанные с деятельностью заемщика	• Содержание и условия коммерческой деятельности заемщика • Кредитоспособность заемщика • Уровень менеджмента заемщика • Репутация заемщика • Банкротство заемщика • Мошенничество со стороны заемщика

Окончание таблицы 1

Внутренние факторы	Характеристика факторов кредитного риска
Факторы, связанные с деятельностью банка-кредитора	• Адекватность выбора кредитной политики • Структура кредитного портфеля • Квалификация персонала • Ошибочные действия кредитных работников • Качество технологий • Тип рыночной стратегии • Способность разрабатывать и продвигать новые кредитные продукты

Виды кредитного скоринга:

- **Apllication scoring**

Данная система является наиболее распространенной и применяется для оценки платежеспособности оставившего заявку потенциального заемщика. По результатам apllication scoring банк может выдать или не выдать кредит, а также предложить клиенту с невысокой благонадежностью другие условия: меньшую сумму или более высокий процент. Скоринг осуществляется на основе анализа кредитной истории. Данные из нее преобразуются в скоринговый балл, который может варьироваться от 300 до 850. Наиболее высокие значения получают благонадежные потенциальные клиенты. Низкий балл соответствует большому риску невозврата кредита, поэтому получившие его заемщики считаются недобросовестными. Таким образом программа ранжирует клиентов по уровню относительного риска невозврата займа.

- **Fraud scoring**

Данная система применяется для определения вероятности мошенничества со стороны потенциального заемщика. Fraud scoring отличается высокой прогностической точностью, особенно при применении в совокупности с другими способами оценки рисков, связанных с

кредитованием. При использовании данной системы скоринговый балл может варьироваться в диапазоне от 1 до 999. Причем чем выше полученное значение, тем больше риск мошенничества потенциального заемщика. Использование Fraud scoring совместно с другими системами скоринга дает возможность значительно улучшить эффективность управления кредитными рисками.

- **Collection scoring**

Эту скоринговую оценку клиентов применяют на стадии работы с невозвращенными кредитами. Collection scoring помогает определить приоритетные действия кредитора для возврата непогашенных займов. По факту система предлагает предпринять определенные меры с целью воздействия на недобросовестных клиентов – от первичного предупреждения до привлечения коллекторского агентства. Интересно, что до 40 % таких заемщиков возвращают средства после напоминания, ссылаясь на забывчивость.

- **Behavioral scoring**

Данная скоринговая оценка применяется для прогнозирования финансовых действий потенциального клиента. Система позволяет предсказывать, как будет меняться платежеспособность заемщика, и корректировать установленные под него лимиты. В качестве основы для анализа программой может использоваться статистика по финансовым действиям в течение определенного промежутка времени (например, по операциям по банковской карте).

- **Расширенный скоринг**

Данная система используется для оценки благонадежности тех заемщиков, у которых еще нет кредитной истории. При этом в качестве критериев принимаются социально-демографические данные. В процессе расчета скорингового балла программа может учитывать такие параметры, как семейное положение, возраст, место и стаж работы, размер заработной

платы. Итоговая сумма варьируется в диапазоне от 50 до 250. Чем выше балл, тем ниже риск невозврата. Расширенный скоринг часто применяется в дополнение к другим методам анализа платежеспособности заемщика.

1.6 Анализ работ, посвященных прогнозированию вероятности возврата кредита

В настоящее время методы машинного обучения набирают большую популярность в сфере кредитного скоринга. Благодаря их скорости и точности, данные модели позволяют сэкономить время, требуемое для принятия решения, и уменьшить фактор человеческой ошибки, что позволит снизить объем проблемных кредитов.

В данной главе рассматриваются методы машинного обучения, которые в настоящее время снискали популярность в области кредитного скоринга.

Традиционно финансовые учреждения используют логистическую регрессию для оценки заемщиков. Выбор использования логистической регрессии обусловлен простотой и прозрачностью данной модели.

В литературе можно найти сложные модели машинного обучения (ML), которые могут заменить модель логистической регрессии. Несмотря на высокую точность моделей ML, они, как правило, не могут объяснить свои прогнозы. Финансовые учреждения являются регулируемыми субъектами и должны быть прозрачными в своих решениях при использовании методов машинного обучения.

Логистическая регрессия

При поиске связей между набором входных переменных и категориальной выходной переменной получила распространение логистическая регрессия.

Логистическая регрессия – это метод построения линейных классификаторов. Ее хорошо использовать для задач бинарной классификации, так как она позволяет предсказывать значения непрерывной зависимой переменной на интервале от 0 до 1.

Логистическая функция выглядит как большая буква S и преобразовывает любое значение в число в пределах от 0 до 1. Это весьма полезно, так как мы можем применить правило к выходу логистической функции для привязки к 0 и 1 (например, если результат функции меньше 0.5, то на выходе получаем 1) и предсказания класса.

Благодаря тому, как обучается модель, предсказания логистической регрессии можно использовать для отображения вероятности принадлежности образца к классу 0 или 1. Это полезно в тех случаях, когда нужно иметь больше обоснований для прогнозирования.

Логистическая регрессия выполняет свою задачу лучше, если убрать лишние и похожие переменные. Модель логистической регрессии быстро обучается и хорошо подходит для задач бинарной классификации.

Деревья решений

Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение.

Random Forest

Random Forest (случайный лес) – является композицией множества решающих деревьев, что позволяет снизить проблему переобучения и повысить точность в сравнении с одним деревом.

Градиентный бустинг деревьев

Градиентный бустинг также является ансамблевым методом, но в отличие от случайного леса который реализует парадигму бегинга т.е., усреднения результатов по множеству деревьев, он реализует парадигму бустинга, отсюда и название метода, т.е. каждое следующее дерево пытается исправить ошибки предыдущего дерева.

Extra Trees

Метод Extra Tree (стоящий для extremely randomized trees) был предложен с главной целью дальнейшего построения дерева рандомизации в контексте

числовых признаков ввода, где выбор оптимальной точки пересечения отвечает для значительной части дисперсии индуцированного дерева. Что касается случайных лесов, метод исключает идею использования загрузочных копий учебного образца, и вместо того, чтобы пытаться найти оптимальную точку пересечения для каждой из случайно выбранных признаков K на каждом узле, он выбирает точку пересечения наугад.

Эта идея довольно продуктивна в контексте многих проблем, характеризующихся большим числом числовых признаков, изменяющихся более или менее непрерывно: она часто приводит к повышенной точности благодаря ее сглаживанию и в то же время значительно снижает вычислительное бремя, связанное с определением оптимальных срезы в стандартных деревьях и в случайных лесах.

Этот метод позволил получить самые современные результаты в нескольких многомерных сложных задачах.

Extreme Gradient Boosting

Экстремальный Градиентный Бустинг (Extreme Gradient Boosting) – это продвинутая реализация Градиентного Бустинга. Этот алгоритм обладает высокой предсказательной способностью и в десять раз быстрее любых других методов градиентного бустинга. Кроме того, включает в себя различные регуляризации, что уменьшает переобучение и улучшает общую производительность.

1.7 Постановка задачи

В качестве задачи был выбран конкурс с сайта kaggle.com. А в частности – конкурс от компании Home Credit Bank.

Целью данного соревнования является создание методики оценки кредитоспособности заемщиков, которые не имеют кредитной истории на основе данных о предыдущих клиентах банка.

Обучающая выборка включает 307511 записей, 122 признака, среди которых присутствуют категориальные. Признаков много, и они довольно

подробно описывают заемщика. Некоторая часть данных разнесена по 6 дополнительным таблицам, которые в дальнейшем нужно также обработать и загрузить к основным.

Конкурс представляет из себя задачу классификации (имеется поле TARGET где 0 – означает отсутствие проблем с платежами, а 1 – их наличие). Но следует понимать, что нам нужно предсказывать не 0/1, а вероятность возникновения проблем.

Бинарная классификация

Необходимо классифицировать элементы определенного множества в две группы по правилам классификации.

Основные методы, которые используются в двоичной классификации:

1. Деревья решений.
2. Случайные леса.
3. Байесовские сети.
4. Методы опорных векторов.
5. Искусственные нейронные сети.
6. Логистическая регрессия.
7. Пробит-регрессия.

Существует много метрик, которые можно использовать для измерения производительности классификатора или предсказателя. Различные поля имеют различные преимущества для конкретных метрик ввиду различных целей.

Если дана классификация множества данных, существует четыре базовые комбинации действительной категории и назначенной категории:

Таблица 2 – Категории-комбинации

	Condition positive (CP)	Condition negative (CN)
Test outcome positive (OP)	True positive	False positive
Test outcome negative (ON)	False negative	True negative

1) правильно назначенные положительные классификации TP;
2) правильно назначенные отрицательные классификации TN;
3) ложно назначенные положительные классификации FP;
4) ложно назначенные отрицательные классификации FN.

Выводы по главе 1

Кредитный риск – это риск неуплаты долга, который может возникнуть из-за того, что заемщик не может произвести требуемые платежи. В первую очередь, риск несет кредитор и включает потерю основного долга и процентов, нарушение денежных потоков и увеличение затрат на организацию возврата долга. Потеря может быть полной или частичной. На эффективном рынке более высокий уровень кредитного риска будет связан с более высокой стоимостью заимствования.

Определение категории для обратившегося клиента методами машинного обучения позволит сократить время для анализа и принятия быстрого решения с достаточно высокой точностью. Так же определение категории клиента будет эффективным с помощью методов машинного обучения – это единый подох ко всем клиентам с использованием одних и тех же однородных показателей.

ГЛАВА 2 МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ КАК ЭФФЕКТИВНОЕ СРЕДСТВО ПРОГНОЗИРОВАНИЯ

Описание машинного обучения, сфер его применения, методов и т.д. можно найти в интернете. Например в литературном источнике [70].

2.1 Логистическая регрессия

Описание данного метода можно найти в источнике [70].

Следует добавить, что в данной работе мы уменьшим показатель регуляризации – C, что позволит, в теории, избежать проблем с «переобучением» модели.

Данная функция выглядит как вытянутая большая буква S. Она преобразовывает любое значение в число в отрезке [0;1]. Это весьма полезно, так как мы можем применить правило к выходу логистической функции для привязки к 0 и 1 (например, если результат функции меньше 0.5, то на выходе получаем 1) и предсказания класса. Благодаря тому, как обучается модель, предсказания логистической регрессии можно использовать для отображения вероятности принадлежности образца к классу 0 или 1. Это полезно в тех случаях, когда нужно иметь больше обоснований для прогнозирования.

Логистическая регрессия выполняет свою задачу лучше, если убрать лишние и похожие переменные. Модель логистической регрессии быстро обучается и хорошо подходит для задач бинарной классификации.

2.2 Деревья решений

Описание данного метода можно найти в источнике [71].

2.3 Ансабли

На сегодняшний день ансамбли используются повсеместно, в связи с их высокой точностью. Идея ансамблей очень просто, достаточно взять несколько не очень эффективных методов обучения и обучить их, так что бы они исправляли ошибки друг друга, при таком подходе качество будет намного выше, нежели использовать каждый метод по отдельности.

2.3.1 Стекинг

Стекинг (рисунок 3) подразумевает под собой обучение различных алгоритмов, их результаты передаются на вход последнему, в качестве последнего часто применяют регрессию.

Так же отмечается, что стекинг применяется крайне редко, так как Беггинг и Бустинг работают точнее.

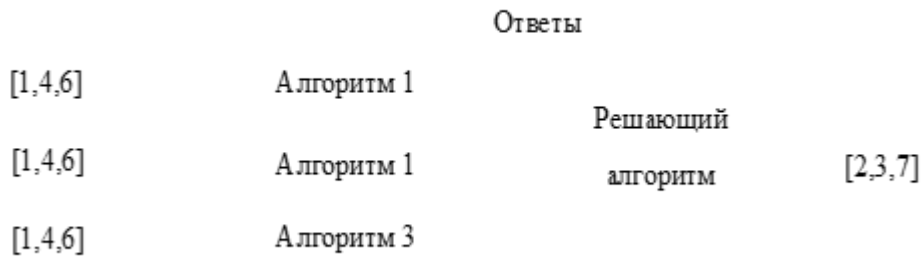

Рисунок 3 – Пример работы «Стекинга»

2.3.2 Бэггинг

Суть метода Bootstrap AGGregatING (рисунок 4) заключается в обучение одного алгоритма на разных выборках из исходных данных, а в конце производится усреднение результатов.

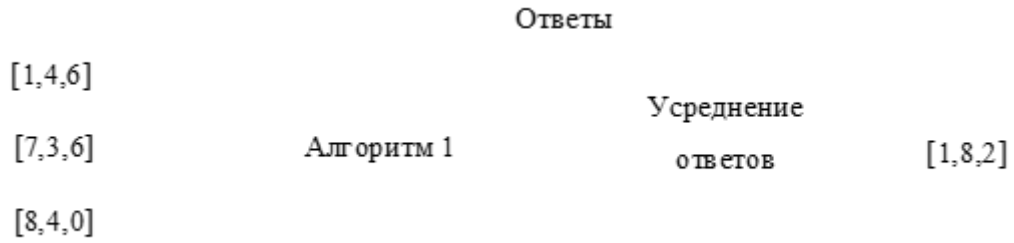

Рисунок 4 – Пример работы «Беггинга»

2.3.3 Бустинг

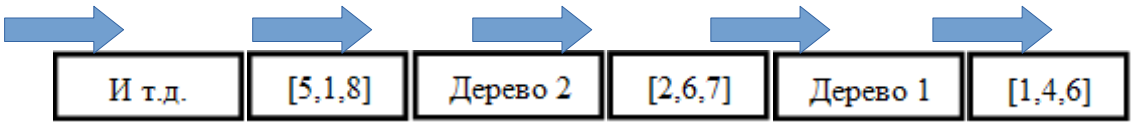

Рисунок 5 – Пример работы «Бустинга»

Обучение алгоритмов происходит последовательно, каждый следующий алгоритм уделяет особое внимание тем случаям, на которых ошибся прошлый.

В качестве применения Бустинга можно привести Яндекс поиск, для ранжирования результата как раз-таки и применяется Бустинг.

2.4 Ансабли деревьев решений

Сегодня существует множество моделей машинного обучения, которые принадлежат к ансамблям, но есть две ансамблевых модели, которые доказали свою эффективность в разных реальных задачах. В обоих случая за основу взяты деревья решений, это:

- случайный лес деревьев решений
- градиентный бустинг деревьев решений

2.4.1 Random Forest

Описание данного метода можно найти в источнике [72].

2.4.2 Градиентный бустинг деревьев

Градиентный бустинг также является ансамблевым методом, но в отличие от случайного леса который реализует парадигму бегинга т.е., усреднения результатов по множеству деревьев, он реализует парадигму бустинга, отсюда и название метода, т.е. каждое следующее дерево пытается исправить ошибки предыдущего дерева. Еще одним отличием градиентного бустинга от предыдущих методов является то, что, в градиентом бустинге используются деревья с маленькой глубиной, в основном от одного до пяти, это благоприятно сказывается на скорости обучения и использования памяти.

В основе этого метода положено правило объединения множества простых моделей, так называемых «слабых учеников», каждое дерево может дать хорошие прогнозы только для части данных, и таким образом для итеративного улучшения качества добавляется все большее количество деревьев.

Градиентный бустинг чувствительный к настройкам, но при правильно подобранных параметрах может дать существенный прирост качества модели.

В градиентном бустинге важен такой параметр как «learning_rate», с помощью которого появляется возможность контролировать скорость обучения.

Под скоростью обучения понимается то, насколько сильно дерево будет исправлять ошибки предыдущего дерева.

Основным недостатком данного алгоритма является чувствительность к параметрам модели, а также то, что для обучения может понадобиться время. Так же, как и другие алгоритмы, которые базируются на дереве решений, алгоритм отлично работает на данных сочетающие в себе непрерывные и бинарные признаки.

Так же стоит выделить основные параметры градиентного бустинга, это learning_rate и n_estimators. Эти два параметра тесно связанны между собой, так как при низком значение learning_rate требуется большое количество деревьев, в отличие от вышеописанных методов большое количество деревьев в градиентом бустинге делает модель более сложной, что может привести к переобучению.

Существует общепринятая рекомендация, которая заключается в том, чтобы настраивать n_estimators в зависимости от возможности вычислительной машины, а затем подгонять learning_rate.

2.4.3 Extra Trees

Метод Extra Tree (стоящий для extremely randomized trees) был предложен с главной целью дальнейшего построения дерева рандомизации в контексте числовых признаков ввода, где выбор оптимальной точки пересечения отвечает для значительной части дисперсии индуцированного дерева. Что касается случайных лесов, метод исключает идею использования загрузочных копий учебного образца, и вместо того, чтобы пытаться найти

оптимальную точку пересечения для каждой из случайно выбранных признаков K на каждом узле, он выбирает точку пересечения наугад.

Эта идея довольно продуктивна в контексте многих проблем, характеризующихся большим числом числовых признаков, изменяющихся более или менее непрерывно: она часто приводит к повышенной точности благодаря ее сглаживанию и в то же время значительно снижает вычислительное бремя, связанное с определением оптимальных срезы в стандартных деревьях и в случайных лесах.

Этот метод позволил получить самые современные результаты в нескольких многомерных сложных задачах.

2.5 Extreme Gradient Boosting

Описание данного метода можно найти в источнике [73].

2.6 Гиперпараметры

Гиперпараметры – это настраиваемые параметры для обучения модели, которая самостоятельно регулирует процесс обучения.[74]

2.6.1 Настраиваемые гиперпараметры

Для проведения исследования в выбранных методах, для построения модели прогнозирования критического финансового состояния предприятия, были настроены следующие гиперпараметры (таблица 3):

Таблица 3 – Гиперпараметры

Наименование гиперпараметра	Настройка гиперпараметра
n_estimators	Количество деревьев в лесу
criterionstring	Функция для измерения качества разделения. Поддерживаемыми критериями являются “mse” для среднеквадратичной ошибки, которая равна уменьшению дисперсии в качестве критерия выбора признаков, и “mae” для средней абсолютной ошибки.
max_depthinteger	Максимальная глубина дерева.

Окончание таблицы 3

Наименование гиперпараметра	Настройка гиперпараметра
max_featuresint	Количество функций, которые следует учитывать при поиске лучшего разделения: • Если int, то рассмотрим функции max_features при каждом разбиении. • Если float, то max_features–это дробь, и при каждом разбиении учитываются функции int(max_features * n_features). • Если “auto", то max_features=n_features. • Если “корень”, затем max_features=квадратный_корень (n_features). • Если "log2", то max_features=log2 (n_features). • Если нет, то max_features=n_features. Примечание: поиск разделения не прекращается до тех пор, пока не будет найден хотя бы один допустимый раздел образцов узлов, даже если для этого требуется эффективно проверить больше, чем функции max_features. (только для Extra Tree, Random Forest)
learning_rate	Скорость обучения (только для Extreme Gradient Boosting) – с его помощью определяют порядок того, как мы будем корректировать наши весы с учётом функции потерь в градиентном спуске. Чем ниже величина, тем медленнее мы движемся по наклонной. Хотя при использовании низкого коэффициента скорости обучения мы можем получить положительный эффект в том смысле, чтобы не пропустить ни одного локального минимума, – это также может означать, что нам придётся затратить много времени на сходимость.

Выводы по главе 2

Существует множество различных методов машинного обучения. Мы рассмотрели самые популярные и подходящие методы для решения поставленной задачи:

- Extra Trees;
- Random forest;

- Extreme Gradient Boosting.

Каждый из представленных методов имеют свои достоинства и недостатки, связанные с их спецификой. Так же были обозначены настраиваемые гиперпараметры для каждой модели.

Проанализировав различные алгоритмы, можно сделать вывод что каждый алгоритм имеет свои достоинства и недостатки. Выбор же алгоритмов производится исходя из поставленной задачи, описанной в начале главы. Составим таблицу для выбора подходящих методов обучения (таблица 4).

Для решения нашей задачи повышение эффективности поиска мошеннических транзакций по банковской карте будем строить модели на следующих алгоритмах: логистическая регрессия, деревья решений и градиентный бустинг.

В целом можно сказать, что для решения задачи классификации подходят множество алгоритмов, каждый алгоритм имеет свои достоинства и недостатки, связанные с их спецификой.

Таблица 4 – Сравнение алгоритмов

Алгоритм	Точность	Время обучения	Тип данных	Решаемая задача
Логистическая регрессия	Средняя	Зависит от параметров модели	Смесь бинарных и непрерывных признаков; Не требует масштабируемости данных.	Классификация
Деревья решений	Средняя	Зависит от параметров модели	Смесь бинарных и непрерывных признаков; Не требует масштабируемости данных.	Классификация; Регрессия.

Окончание таблицы 4

Алгоритм	Точность	Время обучения	Тип данных	Решаемая задача
Случайный лес	Средняя	Зависит от параметров модели	Смесь бинарных и непрерывных признаков; Не требует масштабируемости данных.	Классификация; Регрессия.
Градиентный бустинг	Высокая	Зависит от параметров модели	Смесь бинарных и непрерывных признаков; Не требует масштабируемости данных.	Классификация; Регрессия.

ГЛАВА 3. ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ВЕРОЯТНОСТИ ВОЗВРАТА КРЕДИТА НА ПРИМЕРЕ ДАННЫХ БАНКА HOME CREDIT

В качестве задачи был выбран конкурс с сайта kaggle.com. А в частности – конкурс от компании Home Credit Bank.

Целью данного соревнования является создание методики оценки кредитоспособности заемщиков, которые не имеют кредитной истории на основе данных о предыдущих клиентах банка.

Обучающая выборка включает 307511 записей, 122 признака, среди которых присутствуют категориальные. Признаков много, и они довольно подробно описывают заемщика. Некоторая часть данных разнесена по 6 дополнительным таблицам, которые в дальнейшем нужно также обработать и загрузить к основным.

Конкурс представляет из себя задачу классификации (имеется поле TARGET где 0 – означает отсутствие проблем с платежами, а 1 – их наличие). Но следует понимать, что нам нужно предсказывать не 0/1, а вероятность возникновения проблем.

Основываясь на результатах предыдущей главы, мы сделали вывод что, для решения данной задачи нам подходят такие методы машинного обучения как, логистическая регрессия, деревья решений и градиентный бустинг.

Для работы с данными составим следующий план:

1. Понимание проблемы и ознакомление с данными.
2. Чистка данных и форматирование.
3. EDA.
4. Базовая модель.
5. Улучшение модели.
6. Интерпретация модели.

3.1 Понимание проблемы и ознакомление с данными

Начнем с импорта библиотек, которые нам понадобятся в анализе для работы с данными в виде таблиц, построения графиков и для работы с матрицами.

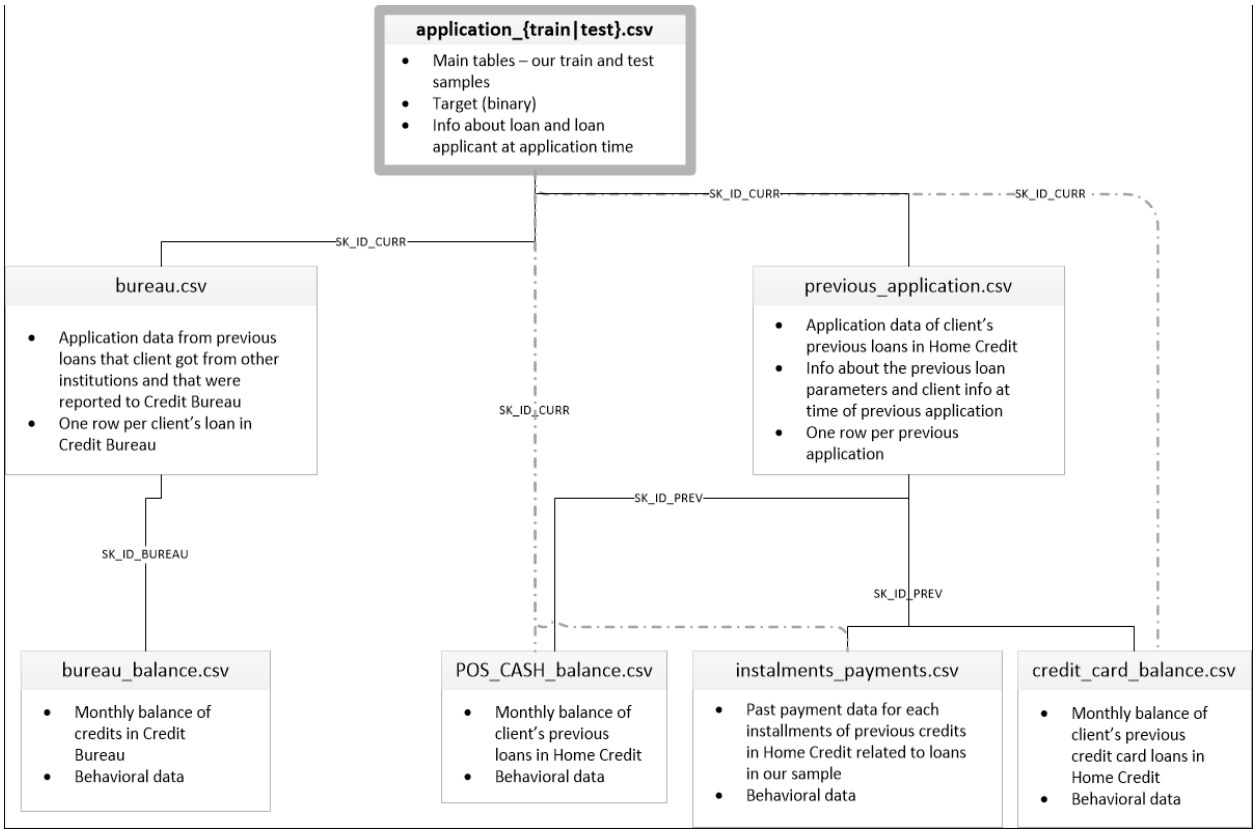

Рисунок 6 – Таблицы с данными

Существует 7 различных источников данных:

- application_train / application_test: основные данные обучения и тестирования с информацией о каждой кредитной заявке в Home Credit. Каждый кредит имеет собственную строку и идентифицируется функцией SK_ID_CURR. Данные заявки на обучение поставляются с ЦЕЛЕВЫМ указанием 0: ссуда была погашена или 1: ссуда не была погашена.
- bureau: данные о предыдущих кредитах клиента из других финансовых учреждений. Каждый предыдущий кредит имеет собственную строку в бюро, но один кредит в данных application может иметь несколько предыдущих кредитов.

● bureau_balance: ежемесячные данные о предыдущих кредитах в бюро. Каждая строка представляет собой один месяц предыдущего кредита, и один предыдущий кредит может иметь несколько строк, по одной на каждый месяц длины кредита.

● previous_application: предыдущие заявки на кредиты в Home Credit клиентов, у которых есть кредиты в данных приложения. Каждый текущий кредит в данных приложения может иметь несколько предыдущих кредитов. Каждое предыдущее приложение имеет одну строку и идентифицируется функцией SK_ID_PREV.

● POS_CASH_BALANCE: ежемесячные данные о предыдущих точках продаж или кредитах наличными, которые клиенты имели с Home Credit. Каждая строка является одним месяцем предыдущего пункта продажи или ссуды наличными, и у одного предыдущего ссуды может быть много строк.

● credit_card_balance: ежемесячные данные о предыдущих кредитных картах, которые клиенты имели с Home Credit. Каждая строка представляет собой один месяц баланса кредитной карты, и одна кредитная карта может иметь много строк.

● installments_payment: история платежей по предыдущим кредитам в Home Credit. Существует один ряд для каждого совершенного платежа и один ряд для каждого пропущенного платежа.

Обучающие данные имеют 307511 наблюдений (каждое – отдельный кредит) и 122 характеристики (переменные), включая TARGET (метка, которую мы хотим предсказать).

Данные test: (48744, 121)

Подробная информация о типах данных:

307511 наблюдений

122 признака

float64 65

int64 41

object 16

dtype: int64

3.2 Exploratory Data Analysis (первичное исследование данных)

Исследовательский анализ данных (EDA) – это открытый процесс, в котором мы рассчитываем статистику и создаем цифры для поиска тенденций, аномалий, закономерностей или взаимосвязей в данных.

Цель EDA – узнать, что наши данные могут сказать нам. Обычно он начинается с обзора высокого уровня, а затем сужается до конкретных областей, поскольку мы находим интригующие области данных. Результаты могут быть интересны сами по себе, или они могут быть использованы для информирования наших вариантов моделирования, чтобы, например, помочь нам решить, какие признаки использовать.

Рассмотрим распределение целевой переменной

Цель – это то, что нас просят предсказать:

либо 0 для кредита было погашено вовремя,

либо 1 означает, что у клиента возникли трудности с оплатой.

Мы можем сначала изучить количество кредитов, попадающих в каждую категорию.

1) *0 – 282686;*
2) *1 – 24825.*

Name: TARGET, dtype: int64

График выглядит следующим образом:

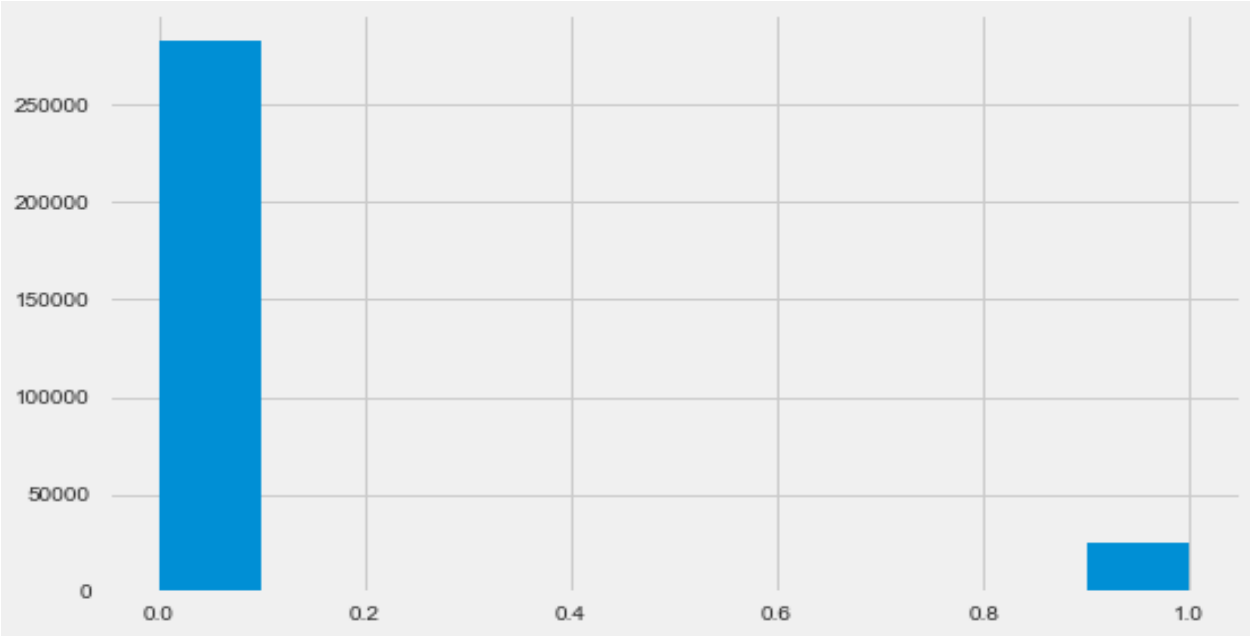

Рисунок 7 – Распределение целевой переменной

Мы видим, что данные непропорционально распределены:

1) 91.92% – проблем нет – 0;
2) 8.08% – проблемы есть – 1;

Классы несбалансированы. Существует гораздо больше кредитов, которые были погашены вовремя, чем кредитов, которые не были погашены. Как только мы перейдем к более сложным моделям машинного обучения, мы можем взвесить классы по их представлению в данных, чтобы отразить этот дисбаланс.

Работа с недостающими данными

В наборе данных 122 колонки, в 67 из них есть недостающие данные.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4

Рисунок 8 – Неполные данные

Типы столбцов и кодирование категориальных данных

Прежде чем идти дальше, нам нужно разобраться с категориальными переменными. К сожалению, некоторые модели машинного обучения не могут работать с категориальными переменными (за исключением некоторых моделей, таких как LightGBM). Поэтому мы должны найти способ кодировать (представлять) эти переменные в виде чисел, прежде чем передать их в модель. Есть два основных способа выполнить этот процесс.

Кодирование категориальных переменных

Прежде чем идти дальше, нам нужно разобраться с категориальными переменными. К сожалению, модель машинного обучения не может работать с категориальными переменными (за исключением некоторых моделей, таких как LightGBM). Поэтому мы должны найти способ кодировать (представлять) эти переменные в виде чисел, прежде чем передать их в модель. Есть два основных способа выполнить этот процесс:

Label Encoding: присвоение каждой уникальной категории в категориальной переменной целого числа. Новые столбцы не создаются.

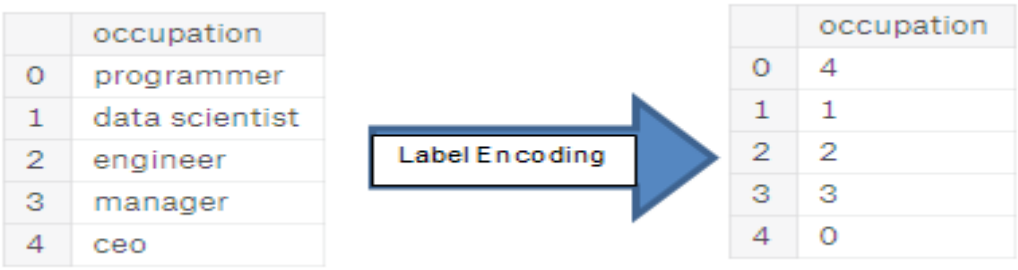

Рисунок 9 – Label Encoding

One-hot encoding: создание нового столбца для каждой уникальной категории в категориальной переменной. Каждое наблюдение получает 1 в столбце для соответствующей категории и 0 во всех других новых столбцах.

Рисунок 10 – One-hot Encoding

Проблема с Label Encoding состоит в том, что она дает категориям произвольный порядок. Значение, присвоенное каждой из категорий, является случайным и не отражает какой-либо неотъемлемый аспект категории. В приведенном выше примере программист получает 4, а ученый – 1, но если мы повторим тот же процесс снова, метки могут быть расставлены в обратном порядке или полностью изменены. Фактическое назначение целых чисел является произвольным. Поэтому, когда мы выполняем Label Encoding, модель может использовать относительное значение функции (например, programmer = 4 и data scientist = 1) для присвоения весов, а это не то, что нам нужно. Если у нас есть только два уникальных значения для категориальной переменной (например, Мужской / Женский), то кодирование меток – это хорошо, но для более чем 2 уникальных категорий безопасным является кодирование One-hot Encoding.

Есть некоторые споры об относительных достоинствах этих подходов, и некоторые модели могут иметь дело с категориальными переменными, закодированными по Label Encoding, без проблем. Единственный недостаток One-hot Encoding заключается в том, что количество признаков (измерений данных) может резко увеличиться с категориальными переменными со большим количеством категорий. Чтобы справиться с этим, мы можем выполнить One-hot Encoding с последующим PCA или другими методами уменьшения размерности, чтобы уменьшить количество измерений (при этом пытаясь сохранить информацию).

В этой работе мы будем использовать Label Encoding для любых категориальных переменных ровно с 2 категориями и One-hot Encoding для любых категориальных переменных с более чем 2 категориями. Этот процесс, возможно, должен измениться, когда мы углубимся в работу, но сейчас мы посмотрим, к чему это нас приведет.

Label Encoding и One-hot Encoding

Реализуем план, описанный выше: для любой категориальной переменной (dtype == object) с 2 уникальными категориями мы будем использовать Label Encoding, а для любой категориальной переменной с более чем 2 уникальными категориями мы будем использовать One-hot Encoding.

Для Label Encoding мы используем Scikit-Learn LabelEncoder, а для One-hot Encoding – функцию pandas get_dummies (df).

Выравнивание test и train

В данных train и test должны быть одинаковые признаки (столбцы). One-hot Encoding создало больше столбцов в train данных, потому что были некоторые категориальные переменные с категориями, не представленными в данных test. Чтобы удалить столбцы в train данных, которых нет в данных test, нам нужно выровнять форматы данных. Сначала мы извлекаем целевой столбец из train (потому что этого нет в данных test, но нам нужно сохранить эту информацию). Когда мы выполняем выравнивание, мы должны

убедиться, что axis = 1, чтобы выровнять форматы данных на основе столбцов, а не строк.

Training Features shape: (307511, 240)

Testing Features shape: (48744, 239)

Теперь train и test наборы данных имеют одинаковый набор признаков, что и требуется для машинного обучения. Количество признаков сильно возрасло, в следствие применения One-hot Encoding. В будущем, нам скорее всего придется использовать уменьшение размерности – dimensionality reduction – для уменьшения размеров наборов данных.

Корреляция в данных

Один из способов понять данные – найти корреляции между признаками и target. Мы можем рассчитать коэффициент корреляции Пирсона между каждой переменной и целью, используя метод данных .corr.

Коэффициент корреляции – не лучший метод для представления «релевантности» признака, но он дает нам представление о возможных отношениях внутри данных. Некоторые общие интерпретации абсолютного значения коэффициента корреляции:

1) 0.00–0.19 «очень слабый»;
2) 0.20–0.39 «слабый»;
3) 0.40–0.59 «умеренный»;
4) 0.60–0.79 «сильный»;
5) 0.80–1.0 «очень сильный».

Most Positive Correlations:

- *NAME_INCOME_TYPE_Working = 0.057481;*
- REGION_RATING_CLIENT = 0.058899;
- REGION_RATING_CLIENT_W_CITY = 0.060893;
- DAYS_EMPLOYED = 0.074958;
- DAYS_BIRTH = 0.078239

Most Negative Correlations:

- EXT_SOURCE_3 = -0.178919;
- EXT_SOURCE_2 = -0.160472;
- EXT_SOURCE_1 = -0.155317;
- NAME_EDUCATION_TYPE_Higher education = -0.056593;
- CODE_GENDER_F = -0.054704

Все данные относительно слабо коррелируют с target, но среди них выделяется возраст и EXT_SOURCE_x. Рассмотрим данные признаки подробнее.

Возраст

Чем старше клиент, тем больше вероятность того, что он вернет кредит и с ним не будет никаких проблем. Изначально, наш признак age имел отрицательные значения, так как данные были записаны в формате «дней до текущей заявки», поэтому умножим колонку на -1 и разделим на количество дней в году – 365.

Корреляция в абсолютных значениях дней с момента рождения и target:

-0.078239308309827088

Построим гистограмму возраста клиента

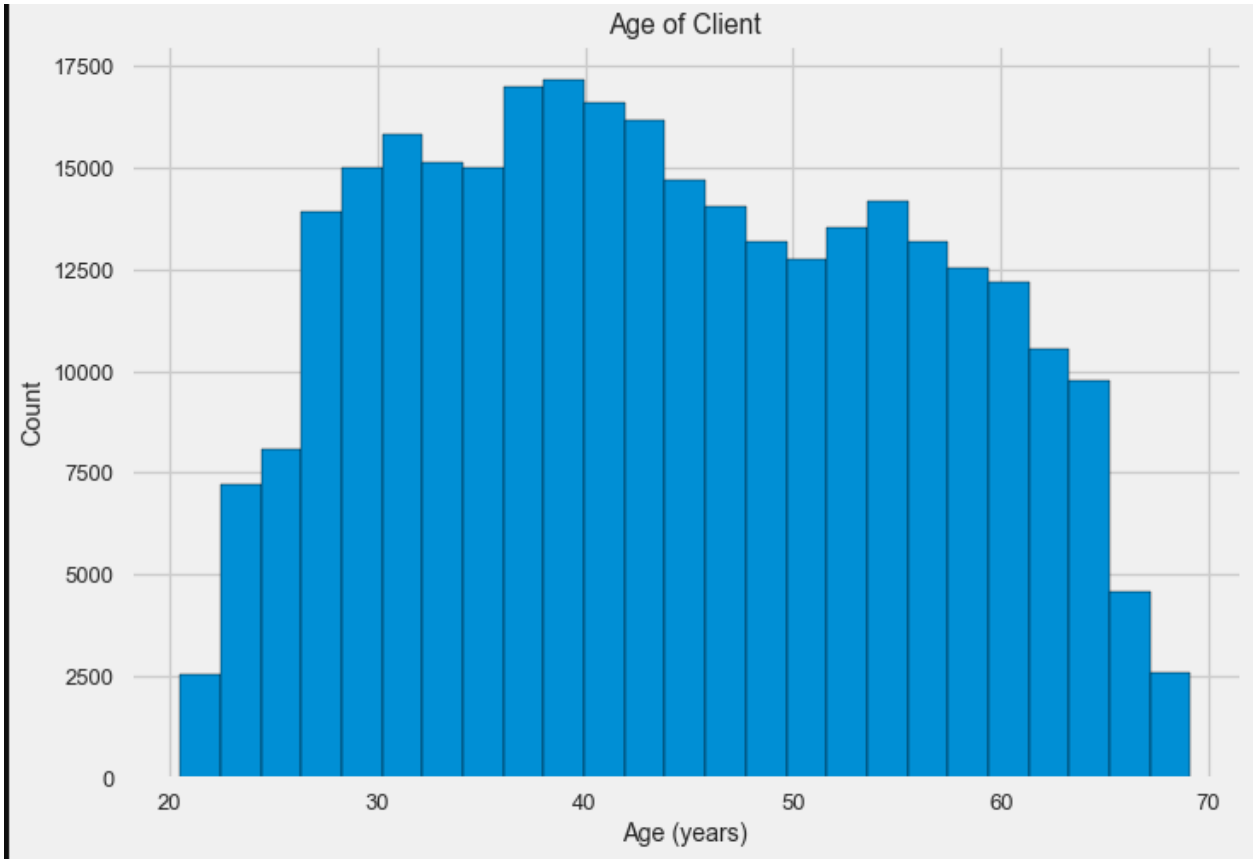

Рисунок 11 – Гистограмма возраста

Само по себе распределение возраста не говорит нам о многом, кроме того, что нет никаких выбросов, так как все возрасты адекватны. Чтобы визуализировать влияние возраста на цель, мы создадим график оценки плотности ядра (KDE), окрашенный по значению цели.

График оценки плотности ядра показывает распределение одной переменной и может рассматриваться как сглаженная гистограмма (она создается путем вычисления ядра, обычно гауссовского, в каждой точке данных, а затем усреднения всех отдельных ядер для создания единой сглаженной кривой). Для этого графика мы будем использовать Kdeplot.

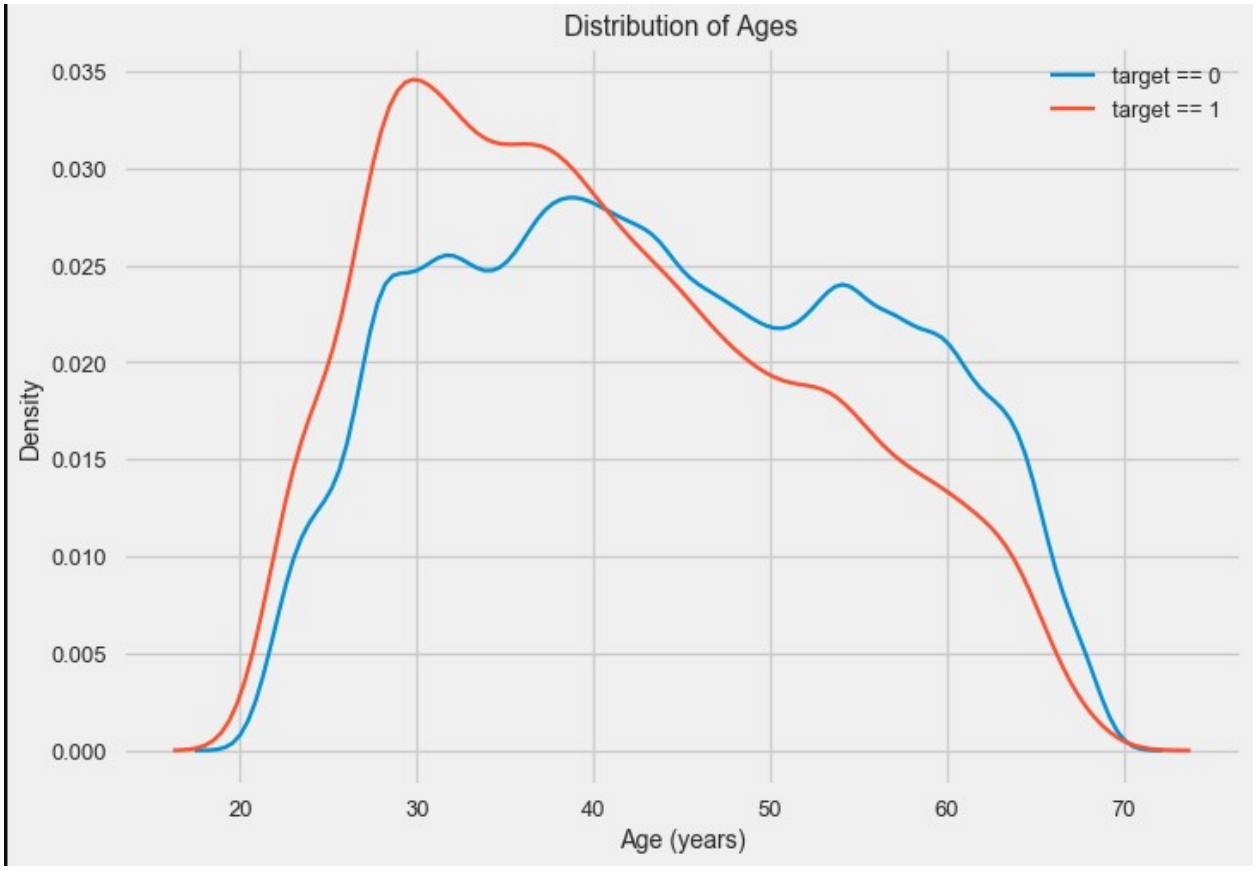

Рисунок 12 – KDE «Возраст»

Кривая target == 1 отклоняется к младшему концу диапазона. Хотя это не является существенной корреляцией (-0,07 коэффициент корреляции), эта переменная, вероятно, будет полезна в модели машинного обучения, потому что она действительно влияет на target. Посмотрим на эти отношения по-другому: средняя неспособность погасить кредиты по возрастным группам.

Чтобы составить этот график, сначала мы разрезали возрастную категорию на группы по 5 лет каждая. Затем для каждой группы мы рассчитываем среднее значение целевого показателя, который сообщает нам соотношение кредитов, которые не были погашены в каждой возрастной категории.

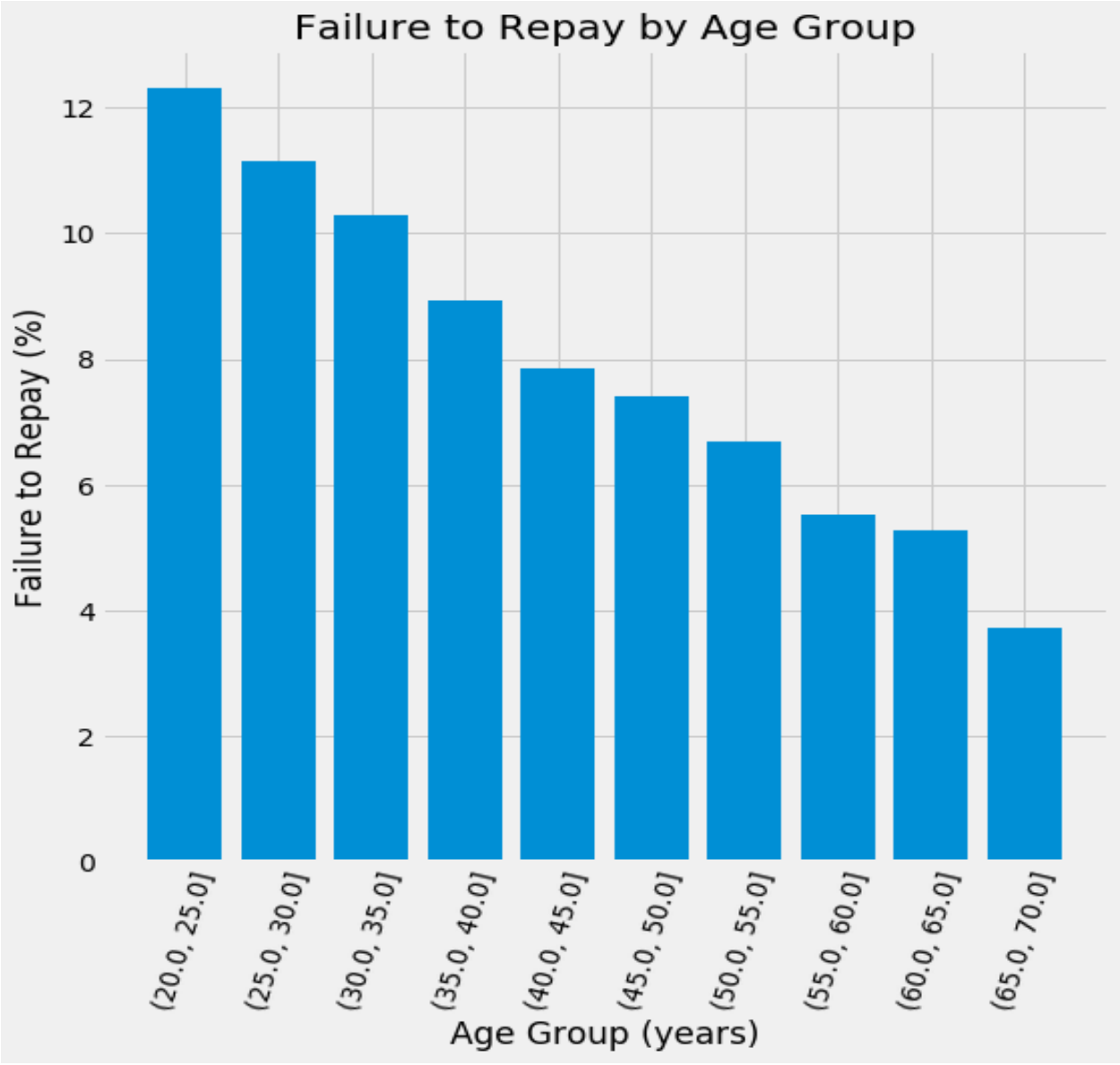

Рисунок 13 – Неспособность выплатить кредит по возрастным группам

Наблюдается четкая тенденция: молодые заявители с большей вероятностью не погасят кредит. Уровень невозврата составляет более 10% для младших трех возрастных групп и ниже 5% для самой старшей возрастной группы.

Эта информация может напрямую использоваться банком: поскольку молодые клиенты с меньшей вероятностью погасят кредит, возможно, им следует предоставить больше рекомендаций или советов по финансовому планированию. Это не означает, что банк должен дискриминировать

молодых клиентов, но было бы разумно принять меры предосторожности, чтобы помочь молодым клиентам своевременно платить.

Внешние источники

3 переменные с наиболее сильными отрицательными корреляциями с target: EXT_SOURCE_1, EXT_SOURCE_2 и EXT_SOURCE_3. Согласно документации, эти функции представляют собой «нормализованную оценку из внешнего источника данных». Возможно, это кумулятивная оценка кредитоспособности с использованием многочисленных источников данных.

Давайте посмотрим на эти переменные.

Во-первых, мы можем показать корреляции функций EXT_SOURCE с target и друг с другом.

	TARGET	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	DAYS_BIRTH
TARGET	1.000000	-0.155317	-0.160472	-0.178919	-0.078239
EXT_SOURCE_1	-0.155317	1.000000	0.213982	0.186846	0.600610
EXT_SOURCE_2	-0.160472	0.213982	1.000000	0.109167	0.091996
EXT_SOURCE_3	-0.178919	0.186846	0.109167	1.000000	0.205478
DAYS_BIRTH	-0.078239	0.600610	0.091996	0.205478	1.000000

Рисунок 14 – Внешние источники данных

Также корреляцию удобно отображать при помощи heatmap:

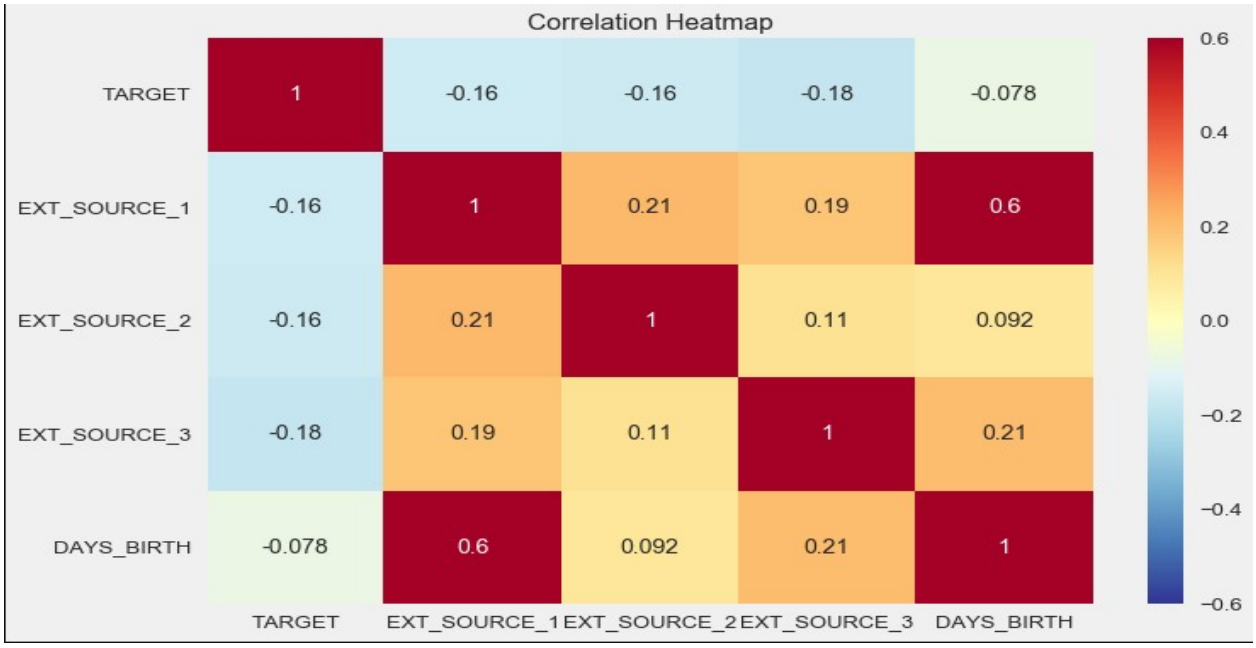

Рисунок 15 – HeatMap

Все три функции EXT_SOURCE имеют отрицательную корреляцию с целью, что указывает на то, что с увеличением значения EXT_SOURCE клиент с большей вероятностью погасит кредит. Мы также видим, что DAYS_BIRTH положительно коррелирует с EXT_SOURCE_1, указывая на то, что, возможно, одним из факторов в этой оценке является возраст клиента.

Далее мы можем посмотреть на распределение каждой из этих функций, окрашенных по значению цели. Это позволит нам визуализировать влияние этой переменной на цель.

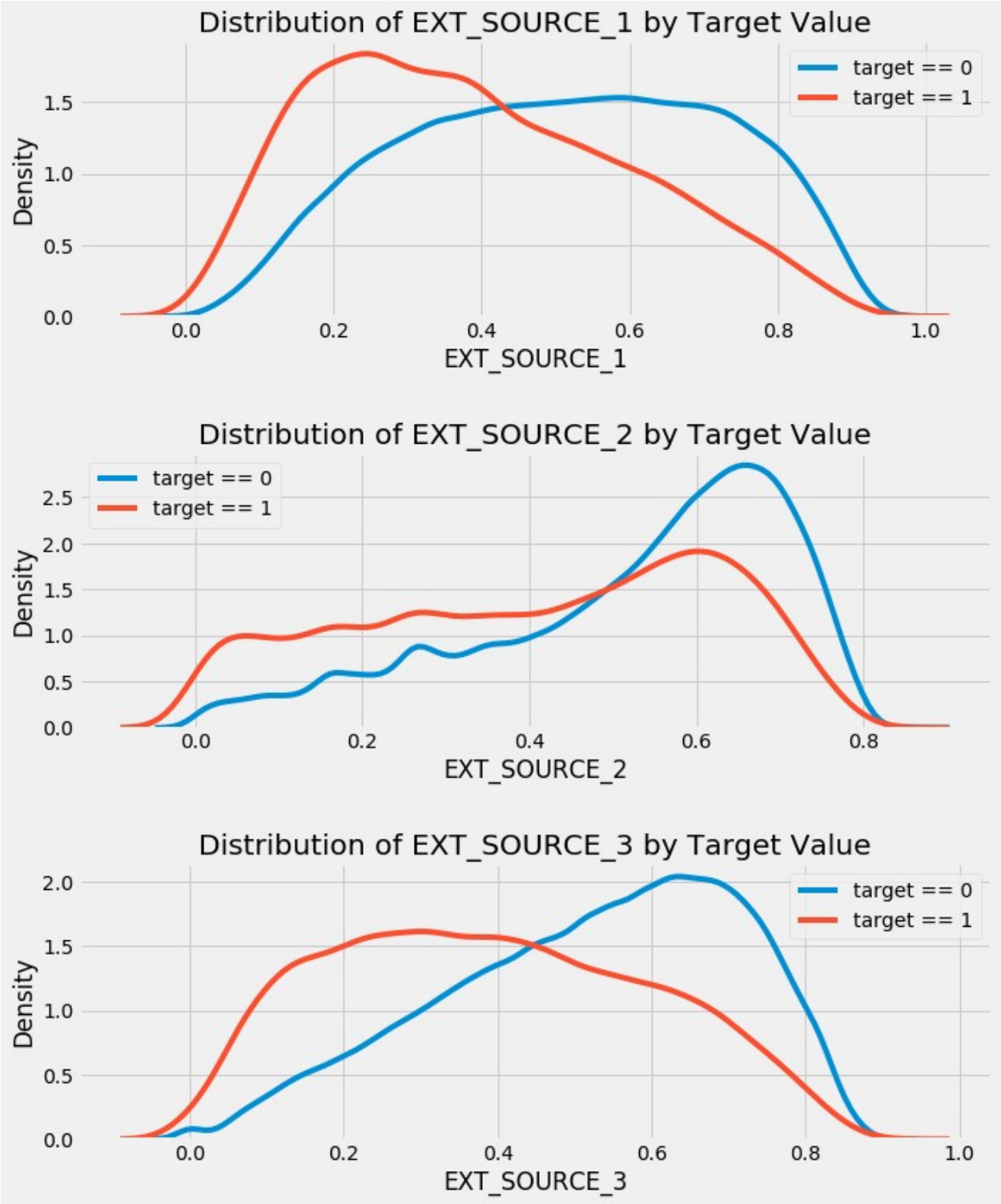

Рисунок 16 – KDE EXT_SOURCE_X

EXT_SOURCE_3 отображает наибольшую разницу между значениями target. Мы ясно видим, что эта функция имеет некоторое отношение к вероятности того, что заявитель погасит кредит. Отношения не очень сильны (на самом деле все они считаются очень слабыми, но эти переменные все

равно будут полезны для модели машинного обучения, чтобы предсказать, вернет ли заявитель кредит вовремя).

Исследование прочих признаков

Рассмотрим остальные признаки и их корреляцию с переменной target.

Отобразим данные на графиках:

Тип займа

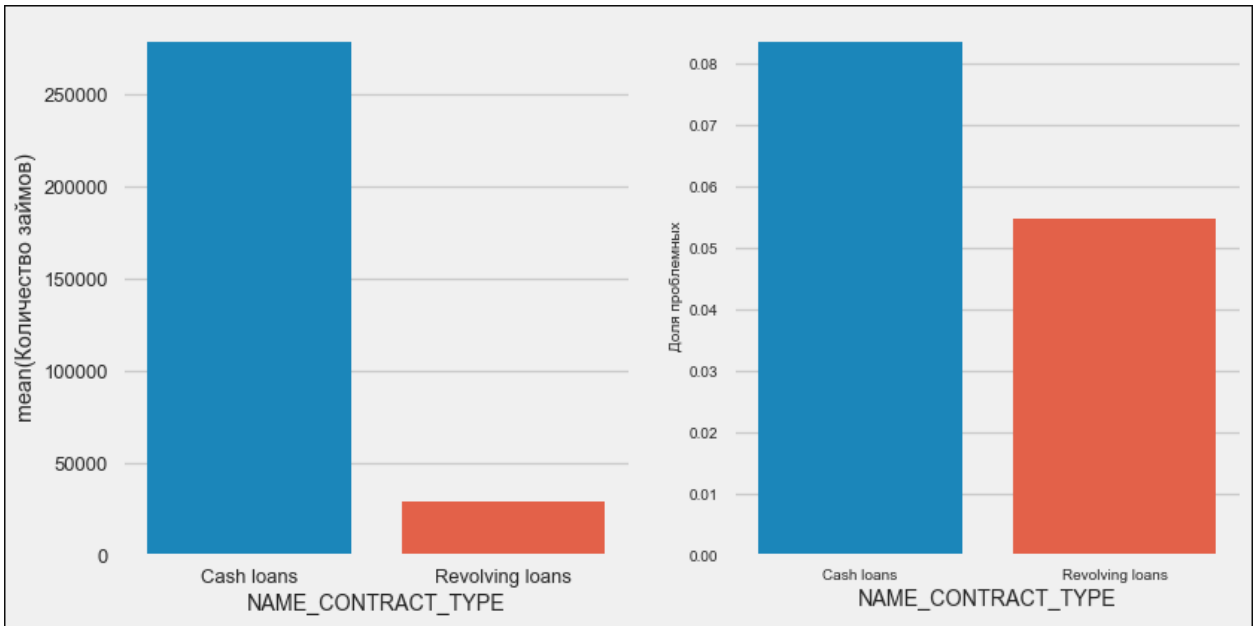

Рисунок 17 – Тип займа

Возобновляемые кредиты имеют малую долю в общем количестве займов (порядка 10%), но при этом процент невозврата по ним очень высок.

Возможно банку стоит пересмотреть свою стратегию работы с данным типом займа.

Пол

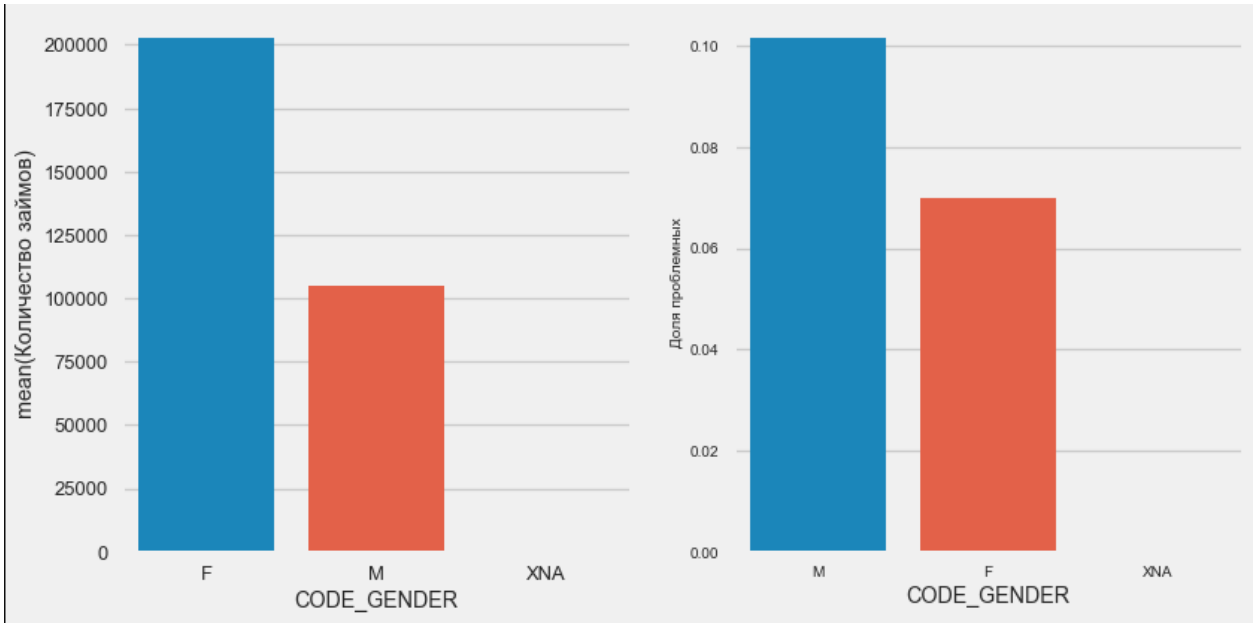

Рисунок 18 – Пол клиента

Количество клиентов-женщин почти в два раза больше клиентов-мужчин. Также, клиенты-мужчины имеют больший риск.

Наличие автомобиля и недвижимости

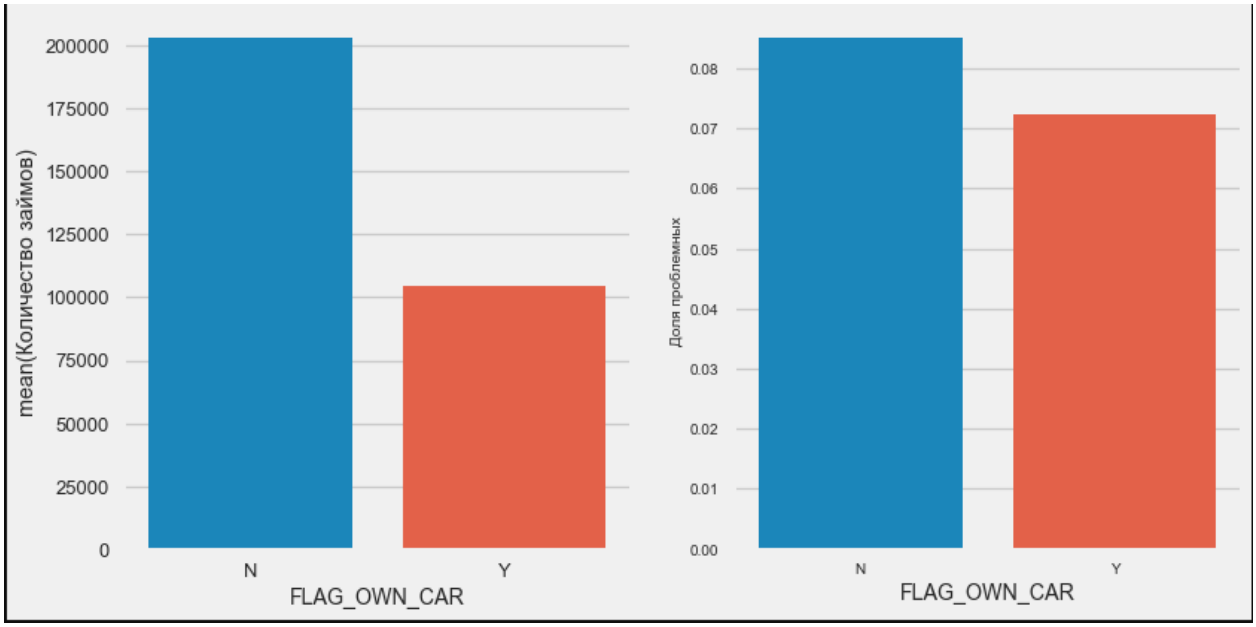

Рисунок 19 – Наличие автомобиля

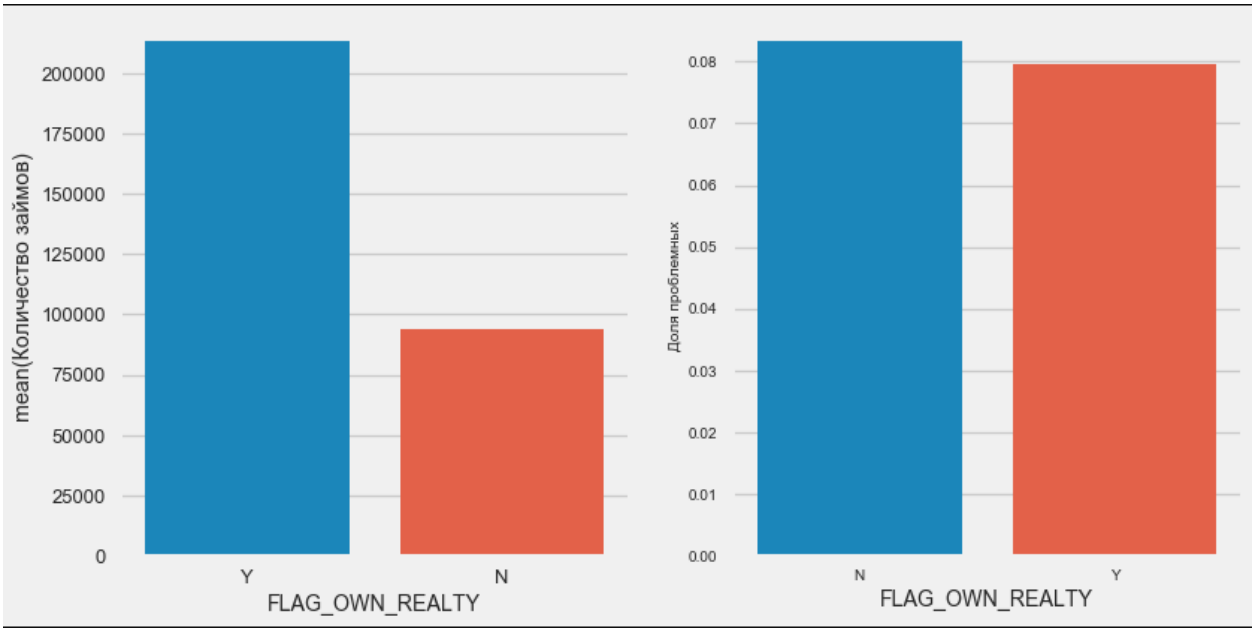

Рисунок 20 – Наличие недвижимости

Клиентов, имеющих автомобиль почти в два раза меньше, чем клиентов без автомобиля. Риск по этим группам практически одинаковый. Клиенты, имеющие автомобиль платят чуть лучше.

По наличию недвижимости картина обратная – клиентов, не имеющих недвижимости вдвое меньше. Риск, по клиентам с недвижимостью чуть меньше.

Семейный статус

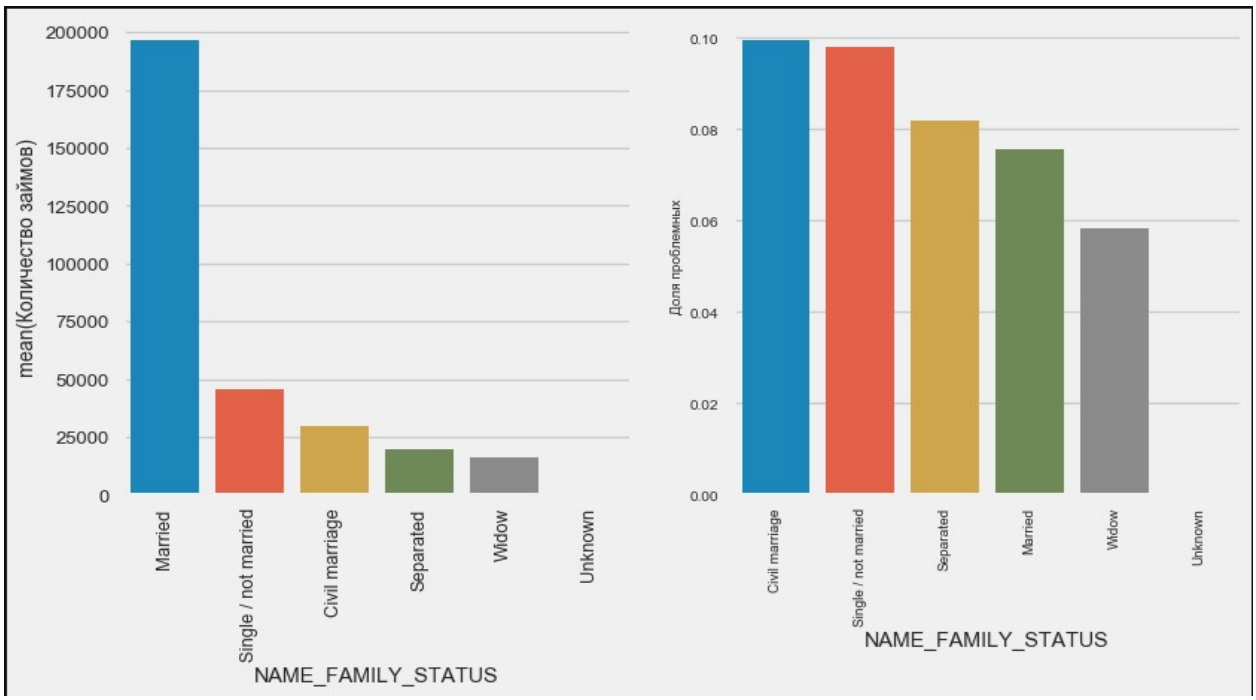

Рисунок 21 – Семейный статус

Большинство клиентов состоят в браке. Наибольший риск показывают клиенты, состоящие в гражданском браке или одинокие лица. Вдовцы имеют наименьший из всех категорий риск.

Количество детей

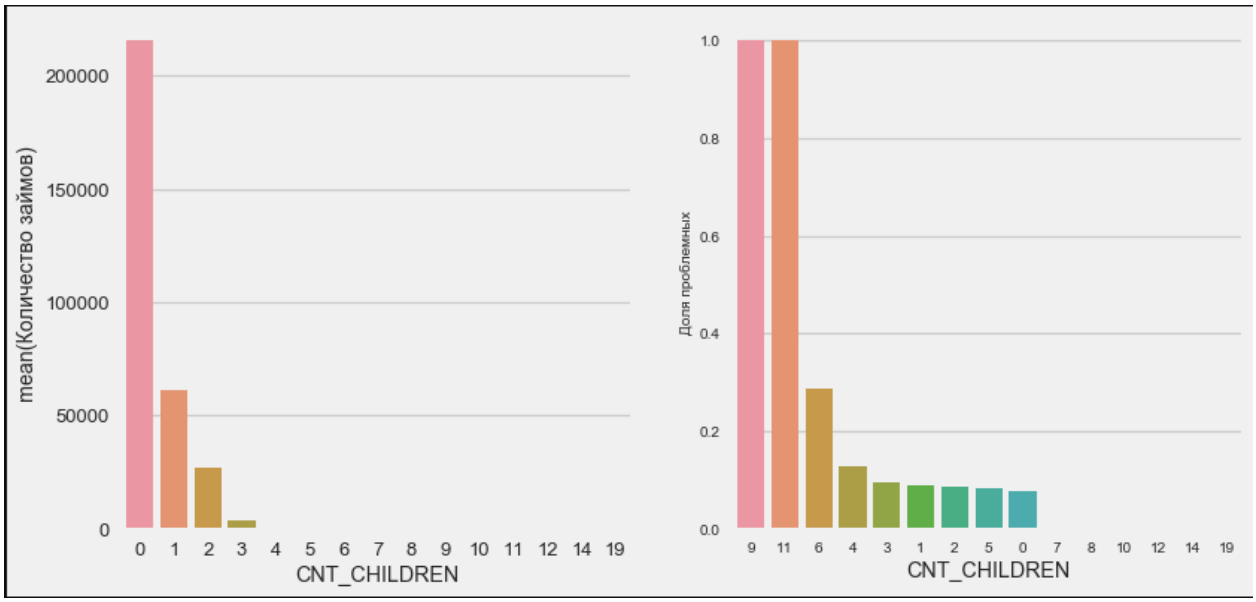

Рисунок 22 – Количество детей

Большинство клиентов не имеют детей. Клиенты с 9 и 11 детьми показывают высокий риск при минимальном количестве займов из всех категорий.

Количество членов семьи

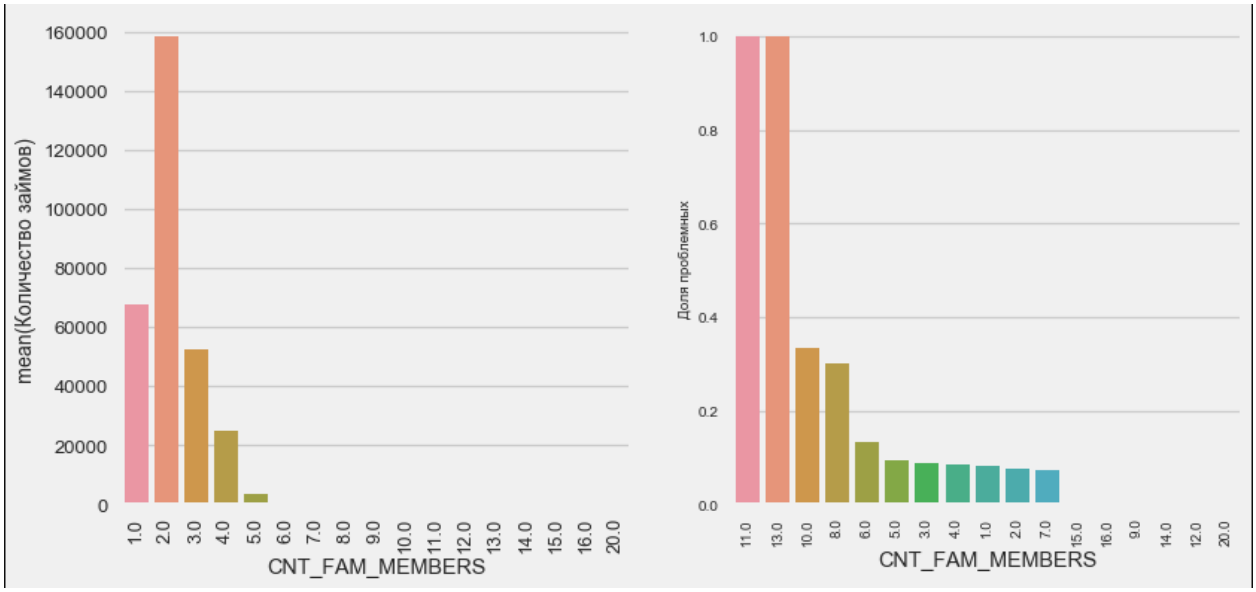

Рисунок 23 – Количество членов семьи

Меньше членов семьи – больше возвратность.

Тип дохода

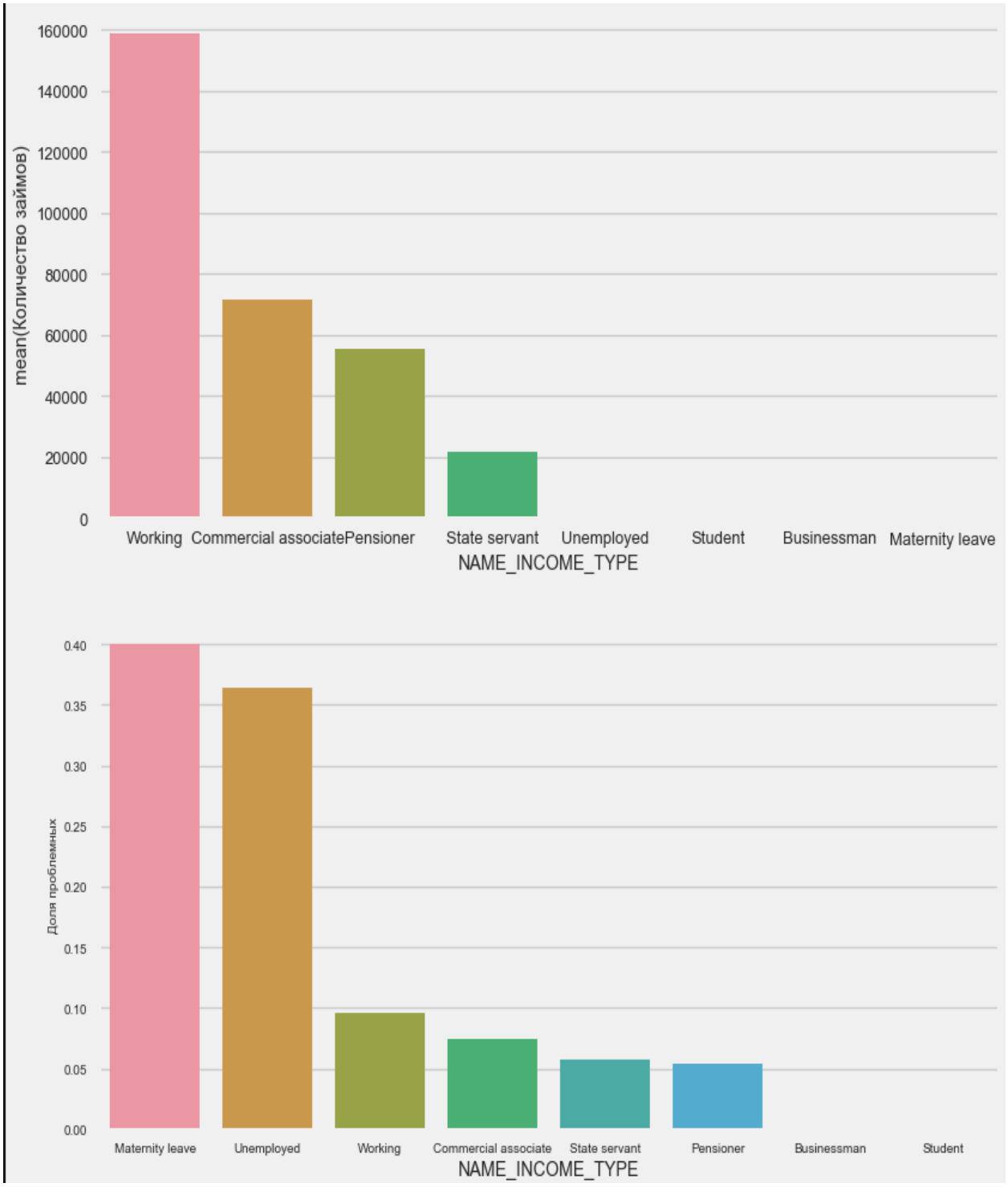

Рисунок 24 – Тип дохода

Безработные клиенты и матери одиночки, вероятно, получают отказ на ранних этапах подачи заявки на кредит – по ним мало данных в выборке.

Деятельность

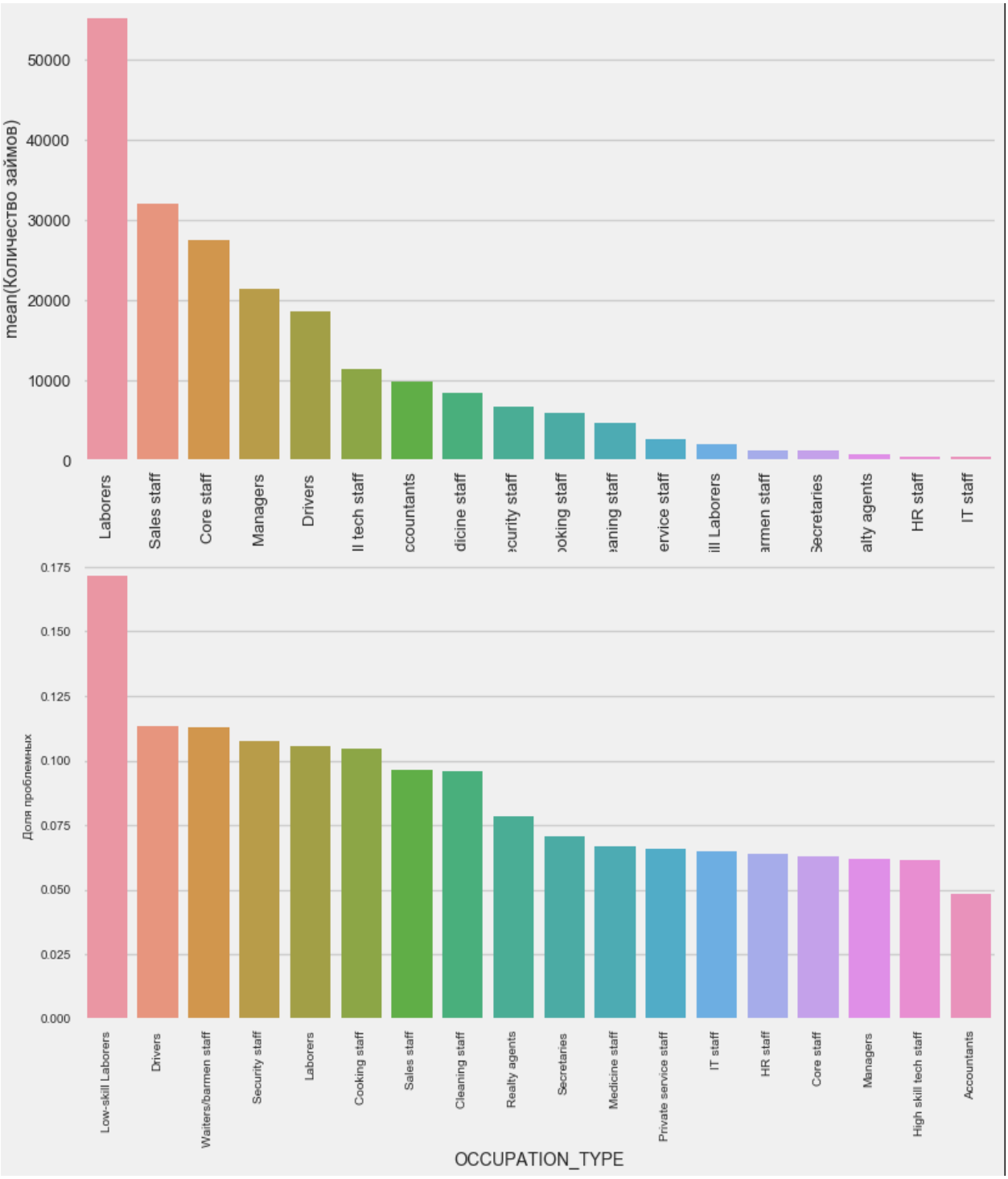

Рисунок 25 – Вид деятельности

Водители и сотрудники безопасности довольно широко представлены в выборке и чаще показывают проблемы, нежели другие категории клиентов.

Образование

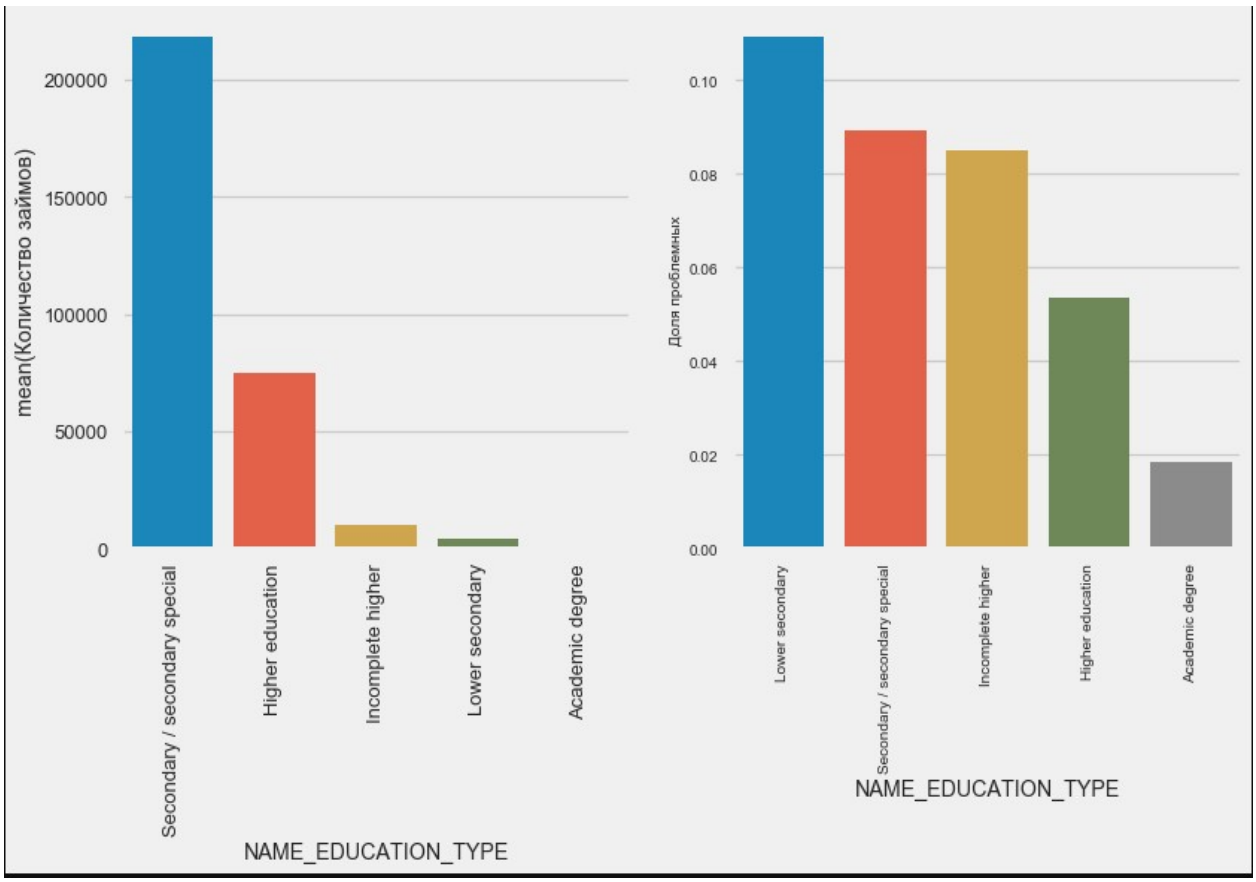

Рисунок 26 – Образование

Выше уровень образования – выше возвратность

Организация-работодатель

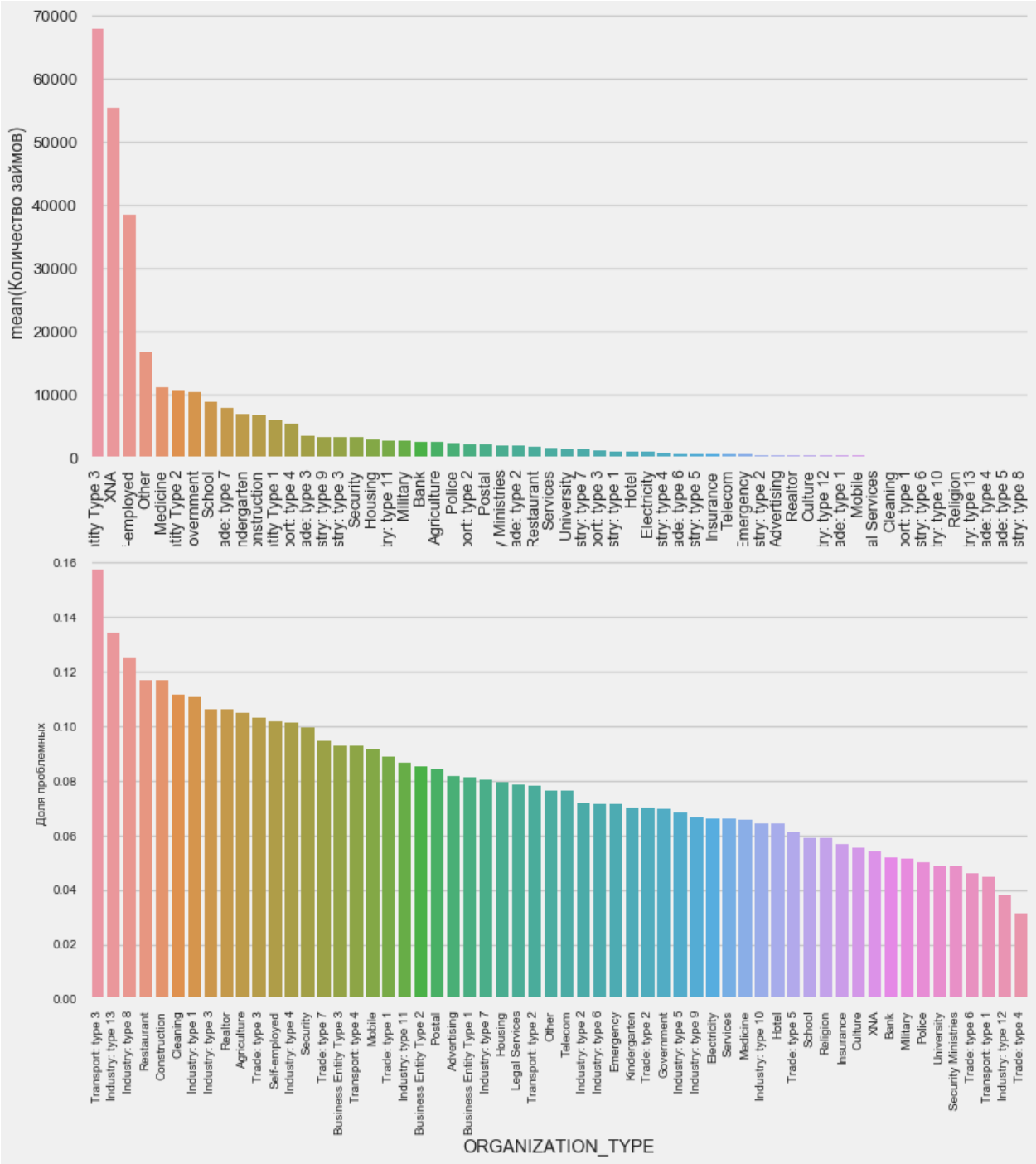

Рисунок 27 – Тип организации

Таблица 5 – Результаты по «тип организации»

Тип	Процент невозврата
Transport:type3	16
Industry:type3	13.5
Industry:type8	12.5
Restaurant	<12

Сумма кредитования (распределение)

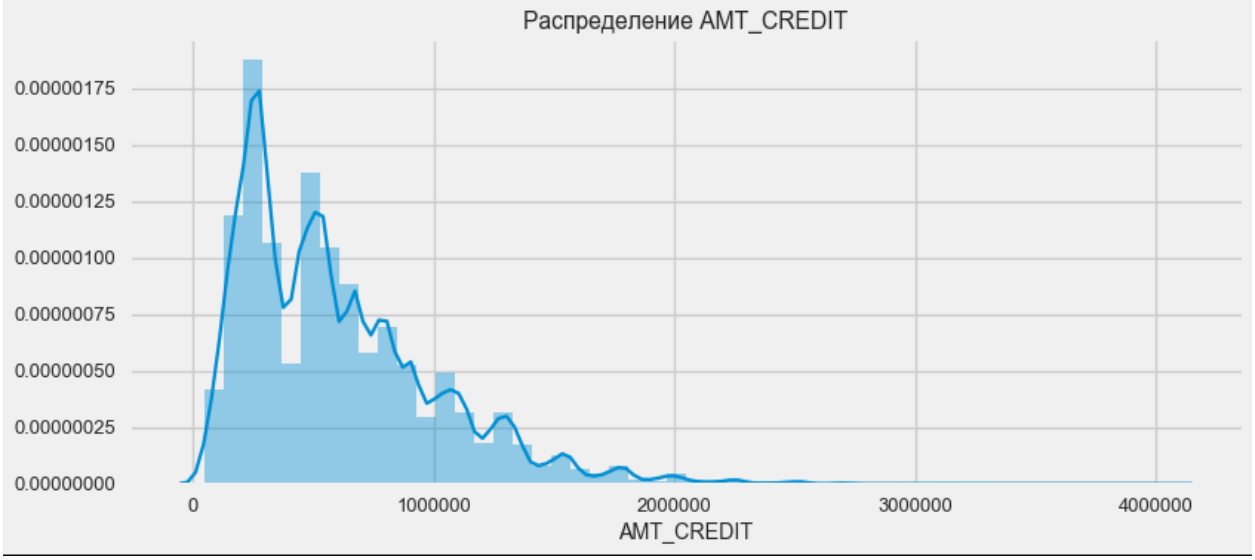

Рисунок 28 – Распределение AMT_CREDIT

Рисунок 29 – Плотность - Сумма кред.

Крупные суммы, судя по графику, возвращаются клиентами чаще.

Плотность проживания (распределение)

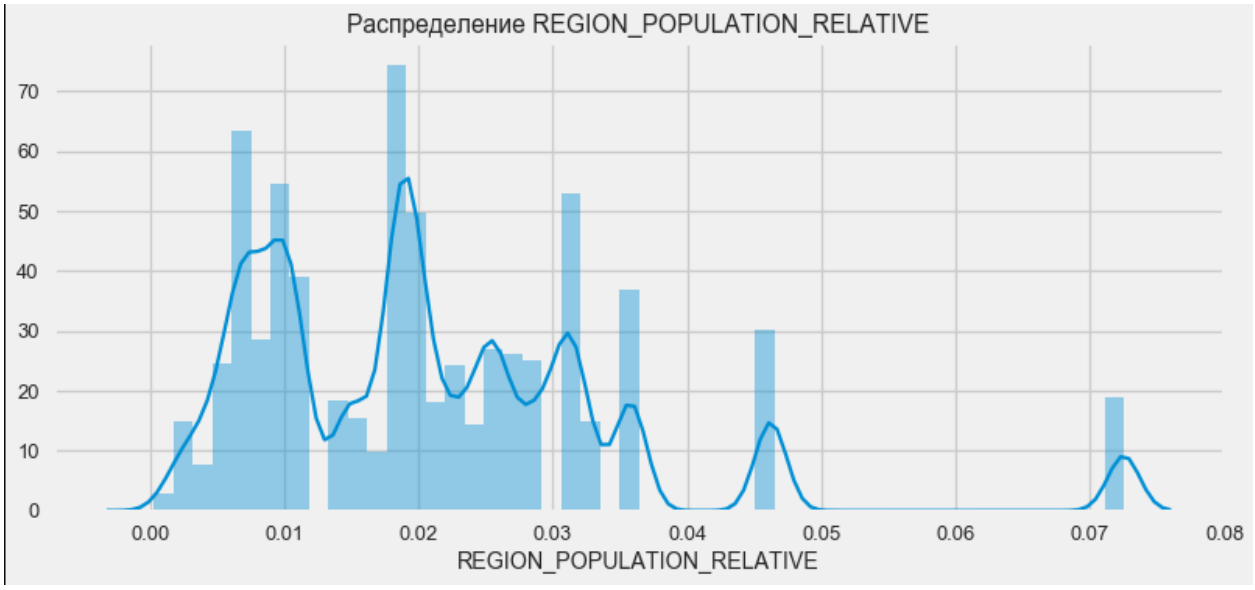

Рисунок 30 – Распределение плотности населения

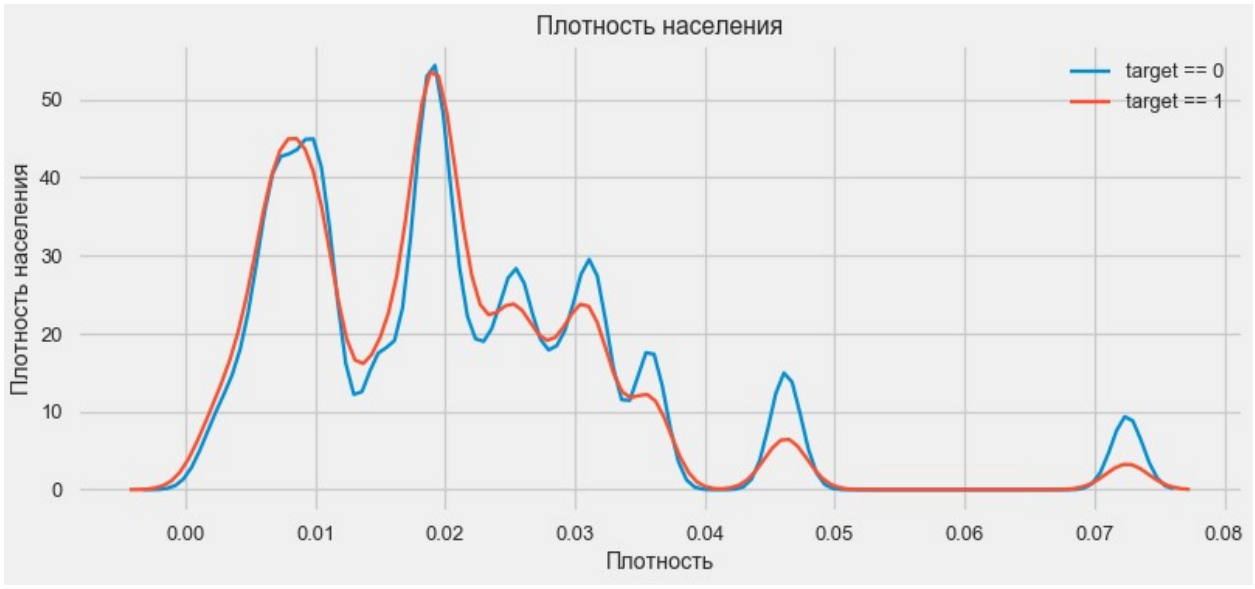

Рисунок 31 – Плотн. Населения - плотность

Клиенты, проживающие в более плотно населенных регионах, показывают лучший процент возврата кредитов.

3.3 Тренировка модели

Будем отталкиваться от базовой модели с результатом 0.5 – случайное угадывание по типу «да – нет».

3.3.1 Логистическая регрессия

Перед началом работы возьмем все наши данные после кодирования. Затем подготовим данные, добавив в них недостающие переменные и нормализовав признаки (feature scaling).

Тренировочная выборка: (307511, 242)

Тестовая выборка: (48744, 242)

Мы будем использовать LogisticRegression от Scikit-Learn для нашей первой модели. Единственное изменение, которое мы сделаем из настроек модели по умолчанию, – это понизим параметр регуляризации, C, который контролирует степень переоснащения (более низкое значение должно уменьшить степень «переобучения»). Это даст нам немного лучшие результаты, нежели стандартная LogisticRegression, но все равно будет устанавливать нижнюю границу для любых будущих моделей.

Теперь, когда модель обучена, мы можем использовать ее для прогнозирования. Мы хотим предсказать вероятности невыплаты кредита, поэтому мы используем метод model predict.proba. Это возвращает массив m x 2, где m – количество наблюдений.

Первый столбец – это вероятность того, что цель равна 0, а второй столбец – это вероятность того, что цель равна 1 (поэтому для одной строки два столбца должны быть равны 1). Мы хотим получить вероятность невозврта кредита, поэтому мы выбираем второй столбец.

Пример полученного результата:

Прогнозы представляют вероятность от 0 до 1, что кредит не будет погашен. Если бы мы использовали эти прогнозы для классификации кандидатов, мы могли бы установить вероятностный порог для определения того, является ли кредит рискованным.

Результат = 0.671

3.3.2 Random Forest

Чтобы попытаться справиться с низкой производительностью нашей базовой модели, мы можем обновить алгоритм. Мы попробуем использовать

Random Forest (Случайный лес) на тех же данных обучения, чтобы увидеть, как это влияет на производительность. Случайный лес – гораздо более мощная и продвинутая модель, особенно когда мы используем большое количество деревьев. Главное – определиться с верными гиперпараметрами, о них было сказано ранее в главе 2. В данной работе мы будем использовать 100 деревьев в случайном лесу.

Настройки выглядят следующим образом:

random_forest = RandomForestClassifier(n_estimators = 100, random_state = 50)

Результат Random Forest – 0.678.

В качестве простого метода, чтобы увидеть, какие переменные являются наиболее релевантными, мы можем посмотреть на важность функций случайного леса. Учитывая корреляции, которые мы увидели в предварительном анализе данных, следует ожидать, что наиболее важными функциями являются EXT_SOURCE и DAYS_BIRTH. Мы можем использовать эти важные особенности как метод уменьшения размерности в будущей работе.

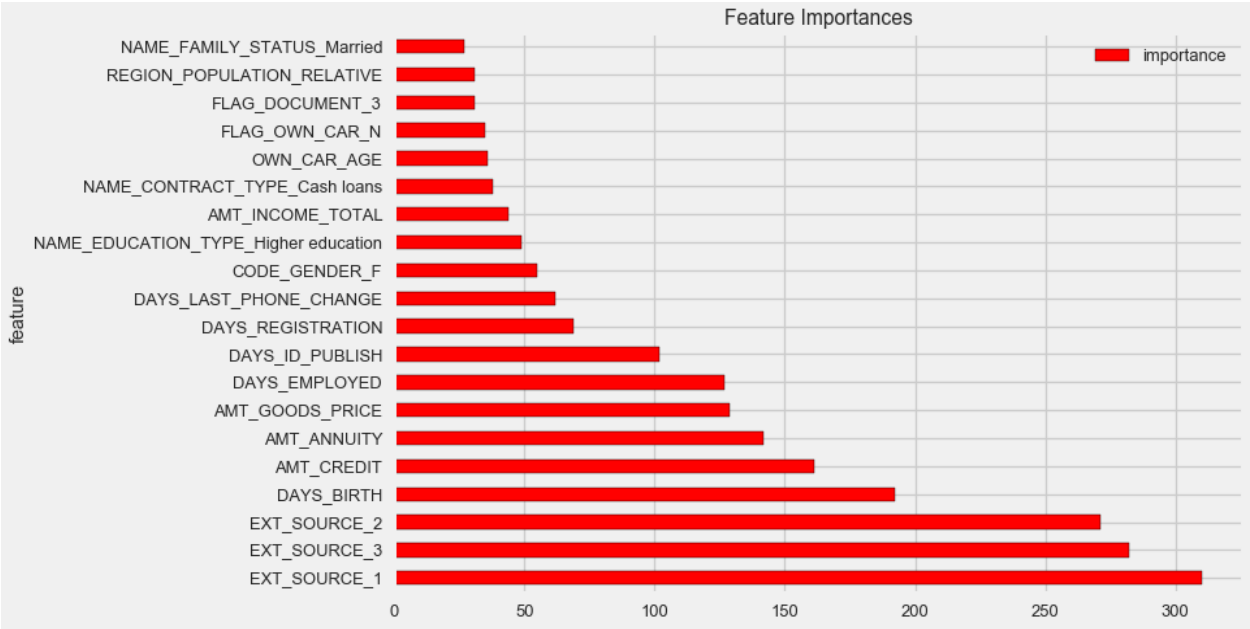

Рисунок 32 – Важность признаков

Ожидаемо, наиболее важными функциями являются те, которые так или иначе связаны с EXT_SOURCE и DAYS_BIRTH. Мы видим, что есть только несколько функций, имеющих важное значение для модели, что говорит о том, что мы можем отбросить многие функции без снижения производительности (возможно даже увидеть увеличение производительности). Важность признаков – не самый сложный метод для интерпретации модели или уменьшения размерности, но они позволяет нам начать понимать, какие признаки учитывает наша модель при прогнозировании.

Рассмотрение данных из прочих таблиц

Данные кредитного бюро по ежемесячному балансу кредитов.

MONTHS_BALANCE – месяцы до даты подачи заявления на выдачу кредита. Имеем следующие данные-статусы:

1) *C 13646993;*
2) *0 7499507;*
3) *X 5810482;*
4) *1 242347;*
5) *5 62406;*
6) *2 23419;*
7) *3 8924;*
8) *4 5847.*

Пояснения по статусам:

- C – closed, кредит погашен;
- X – статус неизвестен;
- 0 – открытый кредит, при этом по нему отсутствуют просрочки;
- 1 – 2 –3 –4 – просрочки с разным количеством дней;
- 5 – кредит продан или списан.

Выделим следующие признаки buro:

1) _grouped_size – количество записей в базе;

2) _grouped_max – максимальный баланс по кредиту;

3) _grouped_min – минимальный баланс по кредиту

Общие данные по кредитным бюро buro.head()

Закодируем данные методом One-Hot-Encoding, сгруппируем по SK_ID_CURR, чтобы в дальнейшем объединить с основной таблицей.

Предыдущие заявки

Поступим аналогичным образом.

Баланс по кредитной карте

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	CNT_INSTALMENT	CNT_INSTALMENT_FUTURE	NAME_CONTRACT_STATUS	SK_DPD	SK_DPD_DEF
0	1803195	182943	-31	48.0	45.0	Active	0	0
1	1715348	367990	-33	36.0	35.0	Active	0	0
2	1784872	397406	-32	12.0	9.0	Active	0	0
3	1903291	269225	-35	48.0	42.0	Active	0	0
4	2341044	334279	-35	36.0	35.0	Active	0	0

Рисунок 33 – Баланс по кредитной карте

Закодируем категориальные признаки и подготовим таблицу для объединения

Данные по картам

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	AMT_BALANCE	AMT_CREDIT_LIMIT_ACTUAL	AMT_DRAWINGS_ATM_CURRENT	AMT_DRAWINGS_CURRENT
0	2562384	378907	-6	56.970	135000	0.0	877.5
1	2582071	363914	-1	63975.555	45000	2250.0	2250.0
2	1740877	371185	-7	31815.225	450000	0.0	0.0
3	1389973	337855	-4	236572.110	225000	2250.0	2250.0
4	1891521	126868	-1	453919.455	450000	0.0	11547.0

Рисунок 34 – Данные по картам

Данные по платежам

	SK_ID_PREV	SK_ID_CURR	NUM_INSTALMENT_VERSION	NUM_INSTALMENT_NUMBER	DAYS_INSTALMENT	DAYS_ENTRY_PAYMENT	AMT_INSTALMENT
0	1054186	161674	1.0	6	-1180.0	-1187.0	6948.360
1	1330831	151639	0.0	34	-2156.0	-2156.0	1716.525
2	2085231	193053	2.0	1	-63.0	-63.0	25425.000
3	2452527	199697	1.0	3	-2418.0	-2426.0	24350.130
4	2714724	167756	1.0	2	-1383.0	-1366.0	2165.040

Рисунок 35 – Данные по платежам

Используем минимальные, средние и максимальные значения из данной таблицы для создания трех таблиц.

Объединение таблиц

Тренировочная выборка (307511, 504)

Тестовая выборка (48744, 504)

Целевой столбец (307511,)

3.3.4 Кросс-валидация

Применим сложную методику с разделением на фолды, кросс-валидацией и выбором лучшей итерации.

Параметры выберем следующие:

k_fold = KFold(n_splits = n_folds, shuffle = True, random_state = 50)

model = lgb.LGBMClassifier(n_estimators=10000, objective = 'binary', class_weight = 'balanced', learning_rate = 0.05, reg_alpha = 0.1, reg_lambda = 0.1, subsample = 0.8, n_jobs = -1, random_state = 50)

model.fit(train_features, train_labels, eval_metric = 'auc', eval_set = [(valid_features, valid_labels), (train_features, train_labels)], eval_names = ['valid', 'train'], categorical_feature = cat_indices, early_stopping_rounds = 100, verbose = 200)

Получаем:

	fold	train	valid
0	0	0.810671	0.762858
1	1	0.808665	0.765861
2	2	0.820401	0.770629
3	3	0.815066	0.766318
4	4	0.810422	0.764517
5	overall	0.813045	0.766050

Full AUC score 0.754

Сравнение моделей

Таблица 6 – Сравнение моделей

Модель	Результат
Logistic regression	0.681
Random forest	0.683
LightGBM+Kfolds	0.754

Выводы по главе 3

Наилучший результат показала модель LightGBM+Kfolds – 0.754

ГЛАВА 4. КОММЕРЦИАЛИЗАЦИЯ РАЗРАБОТАННОГО ПРОЕКТА

Данная продукция, а точнее программный продукт реализован на языке программирования python, с использование методов машинного обучения.

4.1 Участники процесса коммерциализации

Коммерциализацию можно представить как процесс вывода инновационного продукта на рынок.

Рисунок 36 – Участники процесса коммерциализации.

Разработчики инноваций:

- научно-исследовательские институты – в настоящее время один из наиболее успешных и быстро развивающихся участников процесса коммерциализации, имеющий значительное количество перспективных разработок. Процесс коммерциализации здесь осуществляется не самим институтом, а его владельцем (заказчиком разработки) – государством, крупной фирмой, частным инвестором;
- малые и средние предприятия – также быстро развивающийся участник процесса коммерциализации, который, в отличие от научно-

исследовательских институтов, реализует самостоятельно (либо через посредников);

- коллективы изобретателей и изобретатели-одиночки – состоят в основном из молодых ученых, по каким-либо причинам «отделившихся» от научно-исследовательских институтов или предприятий. Часто имеют большое количество разработок, но неспособны довести их до рыночного применения.

Покупатели инноваций (инвесторы):

- государственные фонды и программы – используются во всех развитых странах мира, и предназначены для обеспечения разработчиков инноваций финансовыми, информационными и другими ресурсами, а также оказания помощи при коммерциализации разработок;
- негосударственные фонды, гранты и программы – оказывают такой же спектр услуг, что и государственные;
- венчурные фонды и «бизнес – ангелы» – предоставляют значительную финансовую помощь разработчикам инноваций, в обмен на возврат вложений или долю в капитале, либо передачу прав на созданную инновацию;
- крупные и средние фирмы – полностью финансируют создание и продвижение инноваций с целью их дальнейшего выпуска или внедрения в собственное производство.

Можно выделить еще одного участника процесса коммерциализации инновационных продуктов, который выступает посредником между разработчиками и покупателями инноваций – это центры трансферта и коммерциализации инноваций, консалтинговые компании, инновационные центры и бизнес-инкубаторы, оказывающие разнообразные брокерские, консультационные или юридические услуги, включая защиту и продвижение на рынок интеллектуальной собственности разработчиков.[80]

4.2 Выбор способа коммерциализации

В настоящее время существует несколько основных подходов, в рамках которых реализуются проекты коммерциализации результатов инновационных научных исследований.

В целом все это можно представить в виде следующей схемы (рисунок 42).

Рисунок 37 – Способы коммерциализации

Таблица 7 – Достоинства и недостатки способов коммерциализации инноваций

Способы коммерциализации	Достоинства	Недостатки
Самостоятельное использование	Захват части рынка, при условии организованным должным образом производством, позволит получить высокие доходы; Контроль и анализ как управленческих, так и производственных процессов; Владение правами на инновацию.	Высокая вероятность возникновения рисков; Большой срок окупаемости; Необходимо значительно финансирование проекта.
Переуступка части прав на инновацию	Низкая вероятность возникновения рисков; Низкие затраты; Короткий срок окупаемости; Освоение новых направлений за счет других компаний; Собственный стиль; финансирования от заказчика	Низкие доходы; Риск нарушения патентных прав; Риск появления подделок.

Окончание таблицы 7

Способы коммерциализации	Достоинства	Недостатки
Полная передача прав на инновацию	Низкий срок окупаемости; Доход пропорционально зависит от работы; Минимальные риски и затраты	Возможность недополучить доход; Частая смена области разработок.

На сегодняшний день главной целью предприятия является получение прибыли. Соответственно для того, чтобы выбрать тот или иной способ коммерциализации, необходимо проанализировать расходы и доходы, которые понесет предприятие, выбрав его.

Таблица 8 – Доходы и расходы предприятия при коммерциализации инноваций

Способы коммерциализации	Доходы предприятия	Расходы предприятия
Самостоятельное использование	Выручка от продажи инновационной продукции; Выручка от сдачи оборудования в лизинг; Выручка от оказания инжиниринговых услуг.	Затраты на организацию и поддержание производства; Затраты на маркетинговые исследования и рекламную кампанию; Затраты на модификацию или доработку продукции; Затраты на привлечение клиентов.
Переуступка части прав на инновацию	Выручка от продажи лицензии (паушальный платеж); Платежи от использования лицензиатом патента (роялти).	Затраты на модификацию или доработку продукции, в случае если ее не проводит лицензиат; Затраты на привлечение клиентов (лицензиатов); Затраты на оказание помощи и консультационных услуг лицензиату; Затраты на поддержание и защиту патентных прав.

Окончание таблицы 8

Способы коммерциализации	Доходы предприятия	Расходы предприятия
Полная передача прав на инновацию	Выручка от продажи патентных прав (паушальный платеж).	Затраты на привлечение клиента (покупателя прав); Затраты на оказание помощи и консультационных услуг покупателю прав.

Проанализировав различные аспекты коммерциализации инновационных проектов, можно сделать вывод о том, что:

- самостоятельное использование инновации позволит предприятию максимизировать свою прибыль, но при этом данный метод является и самым затратным;
- частичная продажа, позволит освоить новые рынки за счет лицензиата, а также вернуть часть потраченных средств;
- при полной передачи прав доход может быть сопоставим с доходом от самостоятельного использования, но при этом компании придется менять область деятельности с каждым проектом.

Подводя итог, можно сказать, что первый способ коммерциализации не подходит нам по той причине, что данные необходимые для анализа попросту отсутствуют.

Второй способ коммерциализации, не подходит так как при частичной передачи прав мы теряем часть прибыли, при этом увеличивая свои затраты.

Выберем третий способ коммерциализации, по следующим причинам:

- низкие затраты по сравнению с двумя предыдущими способами;
- возможность решения различного рода задач;
- высокие доходы.

Таким образом, являясь важнейшим элементом инновационного процесса, коммерциализация служит одним из основных условий успешного внедрения результатов инновационной деятельности в любой стране.[81]

Поэтому для эффективной коммерциализации инноваций предприятиям необходимо уделять особое внимание выбору способа коммерциализации. [81]

4.3 Описание продукта

Предлагаемое нами решение, в частности модель для решения задачи классификации потенциальных кредитополучателей в целях уменьшения рисков для банка., реализованное на языке программирования python, с использованием библиотек машинного обучения, позволит финансовым предприятиям минимизировать объем рисков и не возвращенных кредитов, и как следствие, повысит уровень конкурентоспособности и финансовой устойчивости банка.

4.4 Решаемая проблема

Решаемая задача – задача классификации потенциальных кредитополучателей в целях уменьшения рисков для банка

Предмет исследования разработка математических моделей для решения задачи классификации потенциальных кредитополучателей в целях уменьшения рисков для банка.

Цель исследования – совершенствование методов предкредитной оценки кредитополучателя на основе анализа данных о предшествующих кредитах других клиентов банка, для минимизации рисков и уменьшения объема не возвращенных денежных средств.

Практическая значимость работы обусловлена применением результатов исследования на практике, для решения задачи классификации потенциальных кредитополучателей в целях уменьшения рисков для банка.

Эффективность процесса прогнозирования можно измерять следующими способами:

- минимальное количество ошибок первого рода, связанных с ложными срабатываниями системы.

4.5 Объем рынка

В компании планируется работа 2 разработчиков, 1 из которых разрабатывают прогнозные модели, 1 разрабатывает интерфейсы, при условии выпуска 2 моделей каждые месяц, объем реально достижимого объема рынка – 24 программных продукта за год.

4.6 Дорожная карта коммерциализации проекта

Дорожная карта

В таблице 9 представлена дорожная карта проекта на 2020 г.

Таблица 9 – Дорожная карта проекта на 2020 год

	ГОД			
	1 квартал	**2 квартал**	**3 квартал**	**4 квартал**
Исследования и разработки	Исследование рынка банковских услуг, в частности – кредитования. Анализ текущего положения дел. Просмотр применяющихся в настоящее время технологий.	Составление плана разработки модели, с учетом требований и потребностей заказчика. Создание базового прототипа модели и его испытание на имеющихся данных.	Подготовка усовершенствованной модели, с учетом выявленных недостатков и пожеланий заказчика. Тестирование почти-финальной модели. Продолжение работы над ней.	Подготовка финального варианта модели. Демонстрация заказчику. Финальные тесты. Подготовка к продаже продукта.
Создание продукта	Анализ рынка. Просмотр уже имеющихся технологий. Рассмотрение существующих аналогов. Получение задач и требований от заказчика.	Составление плана создания продукта. Начало его разработки. Создание прототипа. Его испытания.	Устранение недостатков. Усовершенствование модели. Получение дополнительных пожеланий и требований от заказчика. Работа над моделью.	К данному этапу должен быть готов финальный продукт, который можно продемонстрировать заказчику. Финализация всех формальностей. Продажа продукта.

Окончание таблицы 9

	ГОД			
	1 квартал	**2 квартал**	**3 квартал**	**4 квартал**
Общее организационное развитие и план по найму	Анализ рынка. Просмотр штата сотрудников. Найм необходимых специалистов. Расширение организации (если необходимо).	Просмотр штата сотрудников. Внесение коррективов по необходимости. Возможен найм специалистов, которые будут работать удаленно	Просмотр штата сотрудников. Внесение коррективов по необходимости.	Просмотр штата сотрудников. Внесение коррективов по необходимости. Расторжение договоров с частью сотрудников на удаленной работе или же их найм в штат.
Защита интеллектуальной собственности и лицензирование	Анализ рынка. Получение информации. Составление всех необходимых документов.	Получение информации. Проверка документации. Внесение коррективов.	Получение информации. Готовый пакет документов.	Наличие всех необходимых документов.
Маркетинг, внедрение продвижение	Анализ рынка. Получение информации. Составление плана маркетинга.	Внесение коррективов в план по необходимости. Продвижение товара на рынке.	Продвижение товара на рынке.	Анализ ошибок. Формирование планов маркетинга на будущее.

4.7 Бизнес-Модель

Таблица 10 – Бизнес-Модель

<table>
<tr><th>Ключевые партнеры</th><th>Ключевые виды деятельности</th><th>Предлагаемые преимущества</th><th>Отношение с клиентами</th><th>Сегменты клиентов</th></tr>
<tr><td rowspan="3">Дом.ру – основной поставщик Интернета;

Bluehost – хостинг;

Новостные сайты и блоги.</td><td>Проектирование и разработка предиктивных систем;
Консультирования заказчиков;
Внедрение разработанных систем;
Поддержка разработанных систем.</td><td rowspan="3">Основу анализа составляют данные находящиеся во внутренне среде предприятия;
Продукт позволяет снизить количество мошеннических транзакций по банковской карте;

Невысокая цена по сравнению с конкурентами.</td><td>Индивидуальный подход к каждому клиенту, за счет специфики данных;</td><td rowspan="3">Финансовые учреждения.</td></tr>
<tr><td>Ключевые ресурсы</td><td>Каналы сбыта</td></tr>
<tr><td>Данные предприятия;
Трудовые ресурсы;
Финансовые ресурсы;
Информационные ресурсы;
Материальные ресурсы.</td><td>Web-представительство компании.</td></tr>
<tr><td colspan="3">Структура расходов</td><td colspan="2">Структура доходов</td></tr>
<tr><td colspan="3">Постоянные издержки:
Налоги; Арендная плата; Оборудование;
Оплата труда административного персонала.
Переменные издержки:
Маркетинг;
Затраты на консультирования;
Оплата труда производственного персонала.</td><td colspan="2">Продажа прав на пользование программным продуктом;
Поддержка программного продукта.</td></tr>
</table>

4.8 Команда проекта

Таблица 11 – Команда проекта

Необходимые роли в проекте	**Обоснование, краткое описание функций**
Руководитель проекта	Менеджер проекта выполняет огромное количество работ, начиная от разработки плана проекта, оценки рисков, контроля функциональных и стоимостных рамок и заканчивая ежедневной работой с командой на проекте.
Разработчики (2 человека)	Это ключевые люди в любой ИТ команде, именно они занимаются непосредственным созданием программного продукта или сложным конфигурированием базового коробочного решения.
Бухгалтер	Специалист по бухгалтерскому учёту, работающий по системе учёта в соответствии с действующим законодательством.

4.9 Ценообразование

Для расчета себестоимости продукта необходимо знать объем затрат на разработку ПО. Группировку затрат будем производить по экономическим элементам, а именно:

1. Материальные затраты;
2. Затраты на оплату труда;
3. Амортизация основных средств;
4. Прочие затраты.

Материальные затраты рассчитываются по формуле 16.

$$З_м = \sum Q_i \cdot З_i, \quad (16)$$

где:

- Зм – затраты на материалы;
- Qi – количество;
- Зi – затраты на единицу.

Затраты на материалы представлены в таблице.

Таблица 12 – Затраты на материалы

Наименование	Единица измерения	Стоимость за единицу, руб.	Количество, шт.	Сумма, руб.
Бумага для принтера	Пачка	220	2	440
USB – флэш накопитель	Штук	800	3	2 400
Ручка	Штук	10	10	100
Стикеры	Штук	170	2	340
Картридж	Штук	880	1	880
Итого				4 160

Затраты на оплату труда будем рассчитывать следующим образом:

$$З_п = \sum(О_i + О_i \cdot С) \cdot G, \quad (17)$$

где:

- Зп – месячный фонд оплаты труда;
- Oi – оклад;
- С – страховые сборы, С=0,34;
- G – занятость.

Таблица 13 – Затраты на оплату труда

Наименование	Оклад (без страховых взносов), руб.	Страховые сборы, руб.	Занятость, %	Сумма, руб.
Руководитель проектов	75 000	25 500	50	50 250
Ведущий Python – разработчик	65 000	22 100	50	43 550
Python – разработчик	50 000	17 000	40	26 800
Разработчик интерфейсов	60 000	20 400	70	56 280
Итого	176 880	Итого	176 880	Итого

Расчет затрат на амортизацию будем производить по следующий формуле:

$$А_{мес} = \sum (С_i / (С_с \cdot Т)) \cdot З_i, \quad (18)$$

где:

- Амес – амортизация за месяц;
- Ci –первоначальная стоимость;
- Сс – срок службы (год);

- T – количество месяцев в году (12);
- Зi – загруженость.

Таблица 14 – Затраты на амортизацию

Наименование	**Кол-во**	**Цена, руб.**	**Сумма, руб.**	**Срок службы, месяцев**	**Амортизация в месяц, руб.**	**Загруженность, %**	**Сумма, руб.**
Ноутбук Asus E203MA-FD017T	1	16999	16999	36	473	90	424
Ультрабук Asus ZenBook UX434FAC-A5147T	2	79999	159998	48	3334	85	2834
Ноутбук Asus FX705DD-AU048T	1	61999	61999	48	1292	85	1098
Windows 10 корпорат-я	4	15000	60000	48	1250	90	1125
MS Office	4	5000	20000	36	556	80	445
Антивирус Kaspersky Security	4	1600	6400	12	533	90	480
Принтер лазерный Samsung SL-M2020W	1	5950	5950	36	166	30	50
ИТОГО	6 456	ИТОГО	6 456	ИТОГО	6 456	ИТОГО	6 456

Также стоит отразить прочие затраты. В состав арендных платежей входит стоимость аренды и обслуживания помещения.

Таблица 15 – Прочие затраты

Наименование	Затраты в месяц, руб.	Количество, шт.	Сумма, руб.
Аренда помещения	12000 за 26 м2	1	12000
Хостинг	849	1	849
Интернет	6500	1	6500
Итого			19 349

Суммарные затраты на разработку рассчитываются по формуле:

$$З = \sum З_{мес} \cdot t_р, \quad (19)$$

где:

- З – суммарные затраты;
- Змес – затраты за месяц;
- tр – время на разработку.

Таблица 16 – Суммарные затраты

Наименование	**Затраты в месяц, руб.**
Материальные затраты	4 160
Затраты на оплату труда	176 880
Амортизация основных средств	6 456
Прочие затраты	19 349
Итого	206 845

Для расчета себестоимости 1 единицы продукта, разделим затраты на переменные и постоянные, а также воспользуемся формулой:

$$Полная\ себестоимость = (З_{пер} \cdot Q + З_{пост}) / Q \quad (20)$$

где:

- Зпер – переменные затраты;
- Q – целевой объем продаж;
- Зпост – постоянные затраты.

Таблица 17 – Переменные и постоянные издержки

Наименование	Затраты на разработку	Время на разработку, месяцев	Сумма, руб.
Переменные издержки			
Затраты на оплату труда	176 880	1	176 880
Итого			176 880
Постоянные издержки			
Материальные затраты	4 160	1	4 160
Амортизация основных средств	6 456	1	6 456
Прочие затраты	19 349	1	19 349
Итого			29 965

Себестоимость 1 единицы продукции составляет 206 845 руб.

Установим целевой объем продаж в месяц равным 2. Зная себестоимость единицы продукции, мы можем рассчитать оптимальную цену и прибыль. Рассчитывать будем по формуле 28:

$$\text{Отпускная цена за 1 единицу} = C+(C*q) \quad (21)$$

где:

- C – себестоимость единицы продукции;
- q – процент прибыли от продаж (25 %).

$$Пр = \text{Отпускная цена за 1 ед} - C \quad (22)$$

где:

- C – себестоимость единицы продукции;
- ПР – прибыль.

$$S = \text{Отпускная цена за 1 ед} \cdot r \quad (23)$$

где:

- S – цена для потребителя;
- r – НДС (20 %).

Из этого следует что, при наценке компании в 25%, отпускная цена за единицу продукции составит 258 556 рублей, а так как у нас еще есть НДС, то цена для потребителя составит 310 267 рублей, соответственно прибыль составит 51 711 рублей с одного проекта, а так как планируется выполнять два проекта в месяц, то сумма составит 103 422 рублей в месяц.

Выводы по главе 4

В данной главе были рассмотрен проект создания сервиса классификации клиента, проанализированы и выбраны возможные методы коммерциализации, были рассчитаны приблизительные затраты, которые составили 206 845 рублей, произведена оценка рынка, выбрана форма распоряжения авторским правом, а также составлена дорожная карта проекта, с описанием процедур и действий на этапе внедрения.

ЗАКЛЮЧЕНИЕ

В данной работе был проведен полноценный анализ исходных данных, были построены графики, таблицы и т.д.

В главе 1 было рассмотрено понятие кредитного риска, были проанализированы текущие практики, также приведен анализ работ, посвященных данной тематике.

В главе 2 было рассмотрено понятие машинного обучения, были разобраны методы, которые могут быть применены к текущей задаче.

В главе 3 были проанализированы исходные данные, выполнен первичный анализ данных, проведена работа с недостающими данными, построены модели с использованием различных методов.

В главе 4 был описан потенциальный план коммерциализации, были приведены дорожная карта, таблицы затрат, план по персоналу и т.д.

Лучший результат – **Full AUC score 0.754** – был получен с применением методики разделения на фолды и использования методики кросс-валидации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ ЛИТЕРАТУРЫ

1) Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1983.

2) Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: исследование зависимостей. – М.: Финансы и статистика, 1985.

3) Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989.

4) Вапник В. Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979.

5) Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. – М.: Фазис, 2006.

6) Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999.

7) Флах П. Машинное обучение. – М.: ДМК Пресс, 2015. – 400 с.

8) Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – Киев: Наукова думка, 2004.

9) Hastie, T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – Springer-Verlag, 2009. – 746 p.

10) Mitchell T. Machine Learning. – McGraw-Hill Science/Engineering/Math, 1997.

11) Ryszard S. Michalski, Jaime G. Carbonell, Tom M. Mitchell, Machine Learning: An Artificial Intelligence Approach, Tioga Publishing Company, (Machine Learning: An Artificial Intelligence Approach в «Книгах Google»), 1983.

12) Vapnik V. N. Statistical learning theory. – N.Y.: John Wiley & Sons, Inc., 1998.

13) Bernhard Schölkopf, Alexander J. Smola Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond. – MIT Press, Cambridge, MA, 2002.

14) I. H. Witten, E. Frank Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). – Morgan Kaufmann, 2005.

15) Liang Wang, Li Cheng, Guoying Zhao. Machine Learning for Human Motion Analysis. – IGI Global, 2009. – 318 p.

16) http://edoc.bseu.by:8080/bitstream/edoc/10164/

17) Lysyuk_R_S_Metodiki_analiza_i_otsenki_s_31-32_ocr.pdf

18) https://scienceproblems.ru/analiz-kreditnyh-riskov-v-kommercheskom-banke/2.html

19) https://en.wikipedia.org/wiki/Credit_risk

20) https://ru.wikipedia.org/wiki/Кредитный_риск

21) https://en.wikipedia.org/wiki/Credit_analysis

22) https://en.wikipedia.org/wiki/Machine_learning

23) https://ru.wikipedia.org/wiki/Машинное_обучение

24) http://www.machinelearning.ru/wiki/index.php?

25) Bluhm, Christian; Ludger Overbeck & Christoph Wagner. An Introduction to Credit Risk Modeling. Chapman & Hall/CRC, 2002.

26) Damiano Brigo and Massimo Masetti. Risk Neutral Pricing of Counterparty Risk, in: Pykhtin, M. (Editor), Counterparty Credit Risk Modeling: Risk Management, Pricing and Regulation. Risk Books, 2006.

27) de Servigny, Arnaud; Olivier Renault. The Standard & Poor's Guide to Measuring and Managing Credit Risk. McGraw-Hill, 2004.

28) Darrell Duffie and Kenneth J. Singleton. Credit Risk: Pricing, Measurement, and Management. Princeton University Press, 2003.

29) Principles for the management of credit risk from the Bank for International Settlements

30) Nils J. Nilsson, Introduction to Machine Learning.

31) Trevor Hastie, Robert Tibshirani and Jerome H. Friedman. The Elements of Statistical Learning. - Springer, 2001.

32) Pedro Domingos, The Master Algorithm, Basic Books. 2015.

33) Ian H. Witten, Eibe Frank. Data Mining: Practical machine learning tools and techniques. - Morgan Kaufmann, 2011. - 664 p.

34) Ethem Alpaydin. Introduction to Machine Learning. - MIT Press, 2004.

35) David J. C. MacKay. Information Theory, Inference, and Learning Algorithms Cambridge: Cambridge University Press, 2003.

36) Richard O. Duda, Peter E. Hart, David G. Stork. Pattern classification (2nd edition), Wiley, New York, 2001.

37) Christopher Bishop. Neural Networks for Pattern Recognition, Oxford University Press. 1995.

38) Stuart Russell & Peter Norvig. Artificial Intelligence – A Modern Approach. Pearson, 2009.

39) Ray Solomonoff, An Inductive Inference Machine, IRE Convention Record, Section on Information Theory, Part 2, 1957, pp., 56–62.

40) Голубев А. А. Финансы и кредит: Учеб. Пособие / А. А. Голубев, Н. П. Гаврилов, – СПб.: СПб ГУИТМО, 2006. – 95 с.

41) Ермаков С. Л. Работа коммерческого банка по кредитованию заёмщиков: Методические рекомендации / С. Л. Ермаков, – М.: Алес, 2005. – 145 с.

42) Митрофанова К. Б. Понятие кредитного риска и факторы, на него влияющие / К. Б. Митрофанова // Молодой ученый. – 2015. – № 2. – С. 284–288.

43) Тен В. В. Проблемы анализа кредитоспособности заемщика / В. В. Тен //Банковское дело. – 2006. – № 3.

44) Черкашенко В. Н. Этот «загадочный» скоринг / В. Н. Черкашенко // Банковское дело– 2006. – № 3.

45) Кредитный скоринг, оценка заемщика, балы, рейтинги [Электронный ресурс] // – справ.-информ. портал. – Электрон. дан. - М., 2016. - URL: http://allcred.ru/articles/kreditnyj_skoring.html (Дата обращения: 18.05.2020)

46) Кредитный скоринг: реальные возможности [Электронный ресурс] / А. Коптелов // - Статья: справ.-информ. портал. - Электрон. дан. - М., 2015. - URL: http://www.cnews.ru/articles/kreditnyy_skoring_realnye_vozmozhnosti (Дата обращения: 18.05.2020)

47) Национальные особенности кредитного скоринга [Электронный ресурс] / В. А. Клапчук // Журнал - Электрон. дан. - М., 2014. - URL: http://www.factoringpro.ru/index.php/credit-scoringstatya/407-skoring-vibor (Дата обращения: 18.05.2020)

48) Скоринг как метод оценки кредитного риска [Электронный ресурс] // Статья - Электрон. дан. - М., 2002. - URL: http://www.cfin.ru/finanalysis/banks/ scoring.shtml (Дата обращения: 18.05.2020)

49) Управление кредитными рисками [Электронный ресурс] // Статья - Электрон. дан. - М., 2005. - URL: http://www.cfin.ru/finanalysis/banks/kreditrisks_mana gement.shtml (Дата обращения: 18.05.2020)

50) What is a Good Credit Score Rating? [Электронный ресурс] // - Электрон. дан. - М., 2016. - URL: http://www.moolanomy.com/1805/what-is-agood-credit-score/ (Дата обращения: 18.05.2020)

51) Statistical and machine learning models in credit scoring: A systematic literature survey [Электронный ресурс] // - Электрон. дан. - М., 2020. - URL: https://www.sciencedirect.com/science/article/abs/pii/S1568494620302039 (Дата обращения: 18.05.2020)

52) Методика оценки кредитоспособности заемщика, используемая Банками Франции Модель Г. Чонаевой [Электронный ресурс]. – Режим доступа: https://studfile.net/preview/1496252/page:2/

53) Банкротство предприятий [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/bankrotstvo-predpriyatiy-1

54) Кдин В.В. Использование методик многомерного рейтингового анализа в диагностике риска банкротства сельскохозяйственных организаций ошмянского района гродненской области. [Электронный ресурс]. – Режим доступа: https://elibrary.ru/item.asp?id=37136949

55) Модель прогнозирования финансового состояния предприятий агропромышленного комплекса. [Электронный ресурс]. – Режим доступа: http://masters.donntu.org/2013/fknt/vovk/library/art5.htm

56) И. Дежина, Б. Салтыков. Механизмы стимулирования коммерциализации исследований и разработок//Общество и экономика, № 7-8, 2004.

57) Министерство образования и науки Российской Федерации. Основные методологические подходы по разработке дорожных карт по приоритетным направлениям научно-технологического и инновационного развития 2011, г. Москва, 86 с.

58) Дорожная карта: как реализовать стратегию интернет-маркетинга [Электронный ресурс]. – Режим доступа: https://www.uplab.ru/blog/ dorozhnaya-karta/

59) Симаранов С., Шох Х. Как создать совместную лабораторию по научно-техническому сотрудничеству: методические рекомендации – Проект EuropeAid «Наука и коммерциализация технологий», 2006.

60) Цуканова О.А., Шашкова Е.В. Разработка комплекса мер для повышения результативности инновационной системы спектральной оптической когерентной микроскопии // Фундаментальные исследования. – 2014. – Вып. 6. – Ч. 2. – С. 340–344.

61) Яновский А. Как финансировать проекты по коммерциализации технологий. – Проект EuropeAid «Наука и коммерциализация технологий», 2006.

62) Ляшин А. Стратегии коммерциализации инноваций – мост между инноватором и бизнесом // Экономика и жизнь. – 2011. – № 36(9402). – URL: http://www.eg-online.ru/ (дата обращения: 27.01.2012).

63) Фёдорова Е.А., Фёдор Ю.Ф., Хрустова Л.Е. Прогнозирование банкротства предприятий на примере отраслей строительства, промышленности, транспорта, сельского хозяйства и торговли

64) Министерство образования и науки Российской Федерации. Основные методологические подходы по разработке дорожных карт по приоритетным направлениям научно-технологического и инновационного развития 2011, г. Москва, 86 с.

65) Дорожная карта: как реализовать стратегию интернет-маркетинга [Электронный ресурс]. – Режим доступа: https://www.uplab.ru/blog/ dorozhnaya-karta/

66) 4 Главных показателя для оценки вашего рынка [Электронный ресурс]. – Режим доступа: https://rb.ru/opinion/market-capacity/

67) Шамаева Д. Р. Деревья решения для задач построения рейтинга коммерческих банков [Текст] // Технические науки: проблемы и перспективы: материалы V Междунар. науч. конф. (г. Санкт-Петербург, июль 2017 г.). – СПб.: Свое издательство, 2017.

68) https://habr.com/ru/post/414613/

69) https://ru.wikipedia.org/wiki/Двоичная_классификация

70) https://vas3k.ru/blog/machine_learning/

71) https://cyberleninka.ru/article/n/evolyutsionnyy-algoritm-postroeniya-dereva-resheniy/

72) https://habr.com/ru/post/320726/

73) https://lambda-it.ru/post/busting-s-pomoshchiu-adaboost-i-gradient-boosting

74) https://docs.microsoft.com/ru-ru/azure/machine-learning/how-to-tune-hyperparameters

75) http://elibrary.ru/item.asp?id=35339610

76) https://scienceproblems.ru/images/PDF/Academy-2-2.pdf

77) http://elibrary.ru/item.asp?id=24413691

78) https://documents.tips/documents/credit-risk-prob-of-default.html

79) Зарубежный опыт коммерциализации инновационных технологий. [электронный ресурс].- Режим доступа: http://elibrary.ru/item.asp?id=29441985

80) Коммерциализация инноваций в регионах. [электронный ресурс].- Режим доступа: http://elibrary.ru/item.asp?id=24124516

81) Инвестиционная деятельность и инновации в Российской Федерации: публично-правовой аспект. [электронный ресурс].- Режим доступа: https://www.book.ru/book/920771