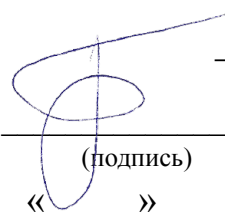


Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК



(подпись)
«___» _____ 2023 г.

РОП

Борисов В.И.
(Ф.И.О.)

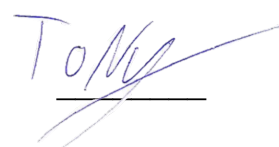
ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

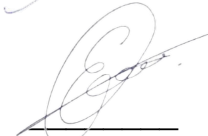
ОЦЕНКА КРЕДИТНЫХ РИСКОВ С ПРИМЕНЕНИЕМ МЕТОДОВ
МАШИННОГО ОБУЧЕНИЯ

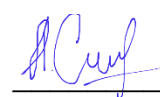
Научный руководитель: Долганов Антон Юрьевич
к.т.н., доцент, доцент

Нормоконтролер: Ф.И.О. Огуренко Е.В. _____

Студент группы РИМ–210962 Спирова Анастасия
Сергеевна







Екатеринбург
2023

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РИТ
Школа профессионального и академического образования
Направление подготовки Информатика и вычислительная техника
Образовательная программа Инженерия машинного обучения

УТВЕРЖДАЮ
РОП _____
«___» _____ 2023 г.

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Спировой Анастасии Сергеевны _____ группы РИМ-210962
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы Оценка кредитных рисков с применением методов машинного обучения

Утверждена распоряжением по институту от «___» _____ 2023 г. № ___

2. Руководитель Долганов Антон Юрьевич, доцент, к.т.н., доцент

(Ф.И.О., должность, ученое звание, ученая степень)

3. Исходные данные к работе Датасет компании Home Credit, опубликованный на платформе Kaggle

4. Перечень демонстрационных материалов Схемы и рисунки для описания подходов к решению задач анализа и предобработки данных о заемщиках, визуализации зависимостей созданных признаков и результатов обучения моделей

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	1. Изучение предметной области и теоретических основ методов обработки данных и машинного обучения в финансовой области	до 15.04.2023 г.	
2.	2. Анализ датасета, предобработка данных	до 29.04.2023 г.	
3.	3. Выбор и построение моделей машинного обучения для оценки кредитного риска.	до 15.05.2023 г.	
4.	ВКР в целом	до 25.05.2023 г.	

Руководитель _____
(подпись)

Долганов Антон Юрьевич
Ф.И.О.

Задание принял к исполнению _____
дата

А.С.Сид
(подпись)

6. Выпускная квалификационная работа закончена «27» мая 2023г. считаю возможным допустить (ФИО) Спирову Анастасию Сергеевну к защите выпускной квалификационной работы закончена в Государственной экзаменационной комиссии.

Руководитель _____
(подпись)

Долганов Антон Юрьевич
Ф.И.О.

7. Допустить Спирову Анастасию Сергеевну к защите магистерской диссертации в Государственной экзаменационной комиссии.

РОП _____
(подпись)

Борисов В.И.
Ф.И.О.

РЕФЕРАТ

Данная дипломная работа посвящена исследованию и оценке кредитных рисков с использованием методов машинного обучения. Главной целью работы является разработка эффективной модели, способной предсказывать вероятность невозврата кредитов и тем самым помочь банкам в принятии правильных решений.

В рамках исследования были проанализированы данные о кредитных операциях, предоставленные коммерческими банками. Была проведена подробная предобработка и нормализация данных для подготовки их к дальнейшему анализу и использованию в моделях машинного обучения.

Основной фокус работы был сосредоточен на применении двух моделей: логистической регрессии и случайного леса. Логистическая регрессия была выбрана из-за своей простоты и интерпретируемости, а случайный лес – из-за своей способности обрабатывать большие объемы данных и выявлять сложные зависимости.

В ходе экспериментов было показано, что обе модели успешно справляются с задачей оценки кредитного риска. Логистическая регрессия показала хорошую производительность, быстроту и точность, что делает ее подходящей для применения в реальном времени, например, при личной подаче заявки в банке или при онлайн-заявках. Случайный лес, в свою очередь, достиг высокой точности, хотя требует больше вычислительных ресурсов.

Дополнительно, в работе был использован метод генетического программирования для создания новых признаков на основе исходных данных. Этот подход позволил значительно улучшить производительность модели и повысить ее точность. Хотя не все созданные признаки вошли в топ-5 наиболее важных, генетическое программирование оказалось эффективным способом генерации признаков, что имеет важное значение в области оценки кредитного риска.

В заключение, данная дипломная работа продемонстрировала возможности применения методов машинного обучения для оценки кредитных рисков. Логистическая регрессия и случайный лес успешно справились с задачей и показали высокую точность в предсказании вероятности невозврата кредитов. Логистическая регрессия отличается своей простотой и скоростью, что делает ее применимой в реальном времени, особенно при личной подаче заявки в банке или онлайн-заявках. Случайный лес, в свою очередь, обладает высокой точностью, хотя требует больше вычислительных ресурсов.

СОДЕРЖАНИЕ

СОДЕРЖАНИЕ	4
ВВЕДЕНИЕ	5
1.1. Представление темы и актуальность исследования	5
1.2. Цель и задачи работы	8
1.2.1. Изучение и анализ предметной области.	10
1.2.2. Подготовка и анализ данных.	12
1.2.3. Выбор и применение моделей машинного обучения.	13
1.2.4. Оценка эффективности моделей.	17
1.2.5. Интерпретация результатов.	17
1.2.6. Обзор литературы и предыдущих исследований	19
2. Теоретические основы методов обработки данных и машинного обучения в финансовой области	24
2.1. Общие принципы и определения.	24
2.2. Методы обработки и анализа табличных данных	24
3.3. Основы машинного обучения и алгоритмы классификации, регрессии и кластеризации	32
3. Анализ данных в финансовой отрасли	36
3.1. Описание датасета	36
3.2. Предобработка данных и преобразование признаков	41
3.3. Исследовательский анализ данных	44
4. Результаты и интерпретация	56
4.1. Преобработка данных	56
4.2. Создание новых признаков	60
4.3. Построение модели	64
5. Заключение	72
5.1. Общие выводы и рекомендации	72
5.2. Ограничения и перспективы дальнейших исследований	74
Источники	76

1. ВВЕДЕНИЕ

1.1. Представление темы и актуальность исследования

В настоящее время, финансовая отрасль является одной из наиболее динамично развивающихся и важных секторов экономики. Она охватывает множество областей, таких как инвестиции, банковское дело, страхование, а также кредитование. Каждая из этих областей имеет свои особенности и требует использования специальных методов и инструментов для эффективной работы. Одним из таких методов является машинное обучение, которое нашло свое применение во многих сферах, включая финансовую.

Особенностью финансовой отрасли является наличие большого количества данных, которые необходимо обрабатывать и анализировать для принятия эффективных решений. Табличные данные о кредитной истории, например, предоставляют множество информации о заемщиках, которые должны быть анализированы, чтобы принять решение о выдаче кредита. В прошлом эту работу выполняли люди, однако, это было связано с большими затратами времени и усилий. Машинное обучение позволяет автоматизировать процесс анализа данных и принятия решений на их основе, что ускоряет работу и повышает ее точность.

Целью данного исследования является оценка использования методов обработки табличных данных с помощью машинного обучения в финансовой области с учетом прогнозирования кредитных рисков. Для достижения этой цели, в работе будут рассмотрены следующие задачи:

1. Исследование и анализ существующих методов и моделей оценки кредитного риска.
2. Предобработка данных о заемщиках, включая финансовые показатели, историю кредитования и другие релевантные параметры.
3. Применение методов машинного обучения для построения моделей оценки кредитного риска.

4. Сравнить производительность различных моделей и подходов, оценивая их точность, скорость работы и практическую применимость.

Актуальность данной темы обусловлена необходимостью повышения эффективности работы финансовой отрасли в условиях быстро меняющейся экономической ситуации и увеличения количества данных, которые необходимо обрабатывать. Оценка кредитных рисков является важной задачей для банков и других финансовых институтов, которые выдают кредиты, и некорректное прогнозирование рисков может привести к значительным финансовым потерям.

Однако, использование методов машинного обучения в этой области также сопряжено с рисками, связанными с некорректным обучением модели и ошибочным прогнозированием. Поэтому, особое внимание в работе будет уделено анализу этих рисков и оценке эффективности применения методов машинного обучения для прогнозирования кредитных рисков в сравнении с традиционными методами анализа данных.

Данная дипломная работа будет состоять из введения, главных разделов, заключения и списка использованных источников. Во введении будет рассмотрена актуальность темы, определены цели и задачи исследования. Главные разделы будут содержать информацию об используемых методах обработки данных и машинного обучения в финансовой области, а также анализ кредитных рисков и оценку эффективности методов машинного обучения для их прогнозирования. В заключении будут подведены итоги исследования и сделаны выводы о применимости методов машинного обучения для прогнозирования кредитных рисков. В списке использованных источников будут перечислены все используемые в работе научные статьи, книги и другие источники информации.

Главной целью данной работы является исследование различных методов машинного обучения для прогнозирования кредитных рисков на основе анализа данных из предоставленного датасета Home Credit. Будут рассмотрены

различные подходы к прогнозированию кредитного риска и методы машинного обучения, такие как случайный лес и логистическая регрессия.

Одной из главных задач данной работы будет выбор наиболее эффективных методов прогнозирования кредитного риска на основе анализа реальных данных о кредитной истории заемщиков из предоставленного датасета. Кроме того, в работе будет рассмотрено влияние различных факторов на кредитный риск, таких как возраст, доход, семейное положение и другие социально-экономические характеристики заемщиков.

Также в рамках этой работы будут рассмотрены методы автоматизации процесса создания новых признаков, которые могут быть использованы в обучении модели.

Таким образом, данная дипломная работа будет иметь практическое значение для компании Home Credit и других финансовых институтов, которые занимаются выдачей кредитов. Результаты исследования могут быть использованы для улучшения прогнозирования кредитных рисков и повышения эффективности работы финансовой отрасли в целом.

1.2. Цель и задачи работы.

Целью данной дипломной работы является проведение исследования эффективности применения методов машинного обучения для оценки кредитного риска на основе набора данных компании Home Credit. Для достижения данной цели необходимо выполнить следующие задачи:

1. Изучение и анализ предметной области. В этой задаче необходимо изучить теоретические основы оценки кредитного риска и применения методов машинного обучения в этой области. Также необходимо изучить существующие подходы к оценке кредитного риска и проанализировать преимущества и недостатки этих подходов.

2. Подготовка и анализ данных. Для построения моделей машинного обучения необходимо провести подготовку и анализ данных. Это включает в себя заполнение пропущенных значений, обработку выбросов и

шумов, а также масштабирование и преобразование данных. Для анализа данных будут использованы методы EDA (Exploratory Data Analysis).

3. Выбор и применение моделей машинного обучения. Для оценки кредитного риска на основе данных компании Home Credit будут выбраны и применены модели машинного обучения. В качестве моделей будут использованы логистическая регрессия и случайный лес. При выборе моделей будут учитываться их преимущества и недостатки, а также возможности для настройки параметров.

4. Оценка эффективности моделей. После построения моделей машинного обучения необходимо оценить их эффективность на основе выбранной метрики. Результаты оценки эффективности моделей будут использованы для сравнения и выбора оптимальной модели.

5. Интерпретация результатов. После оценки эффективности моделей необходимо проанализировать полученные результаты и интерпретировать их в контексте оценки кредитного риска. Это позволит выявить факторы, влияющие на оценку кредитного риска, и определить возможности для улучшения моделей машинного обучения.

В результате выполнения данных задач будет получена оценка эффективности применения логистической регрессии и случайного леса для оценки кредитного риска на основе данных компании Home Credit. Это позволит определить оптимальную модель машинного обучения для данной задачи и улучшить процесс принятия решений в области выдачи кредитов.

Данные задачи будут решаться на основе использования программного обеспечения Python и библиотек машинного обучения, таких как scikit-learn, numpy, pandas, dear. Для работы с данными будут использоваться методы предобработки данных, визуализации и анализа данных, а также методы машинного обучения, такие как логистическая регрессия и случайный лес.

Также в рамках данной работы будет проведено сравнение результатов, полученных на основе моделей машинного обучения, с результатами, полученными на основе классических подходов к оценке кредитного риска.

Это позволит сделать выводы о преимуществах и недостатках использования методов машинного обучения в этой области и определить возможности для дальнейших исследований.

Таким образом, данная работа имеет практическое значение для финансовых организаций, занимающихся выдачей кредитов. Она позволит определить оптимальную модель машинного обучения для оценки кредитного риска на основе данных компании Home Credit и улучшить процесс принятия решений в этой области.

1.2.1. Изучение и анализ предметной области.

В банковской сфере существует достаточно богатая иерархия рисков, ассоциированных с деятельностью кредитных организаций [1]. На Рисунке 1 можно увидеть примерную схему рисков банков [2].

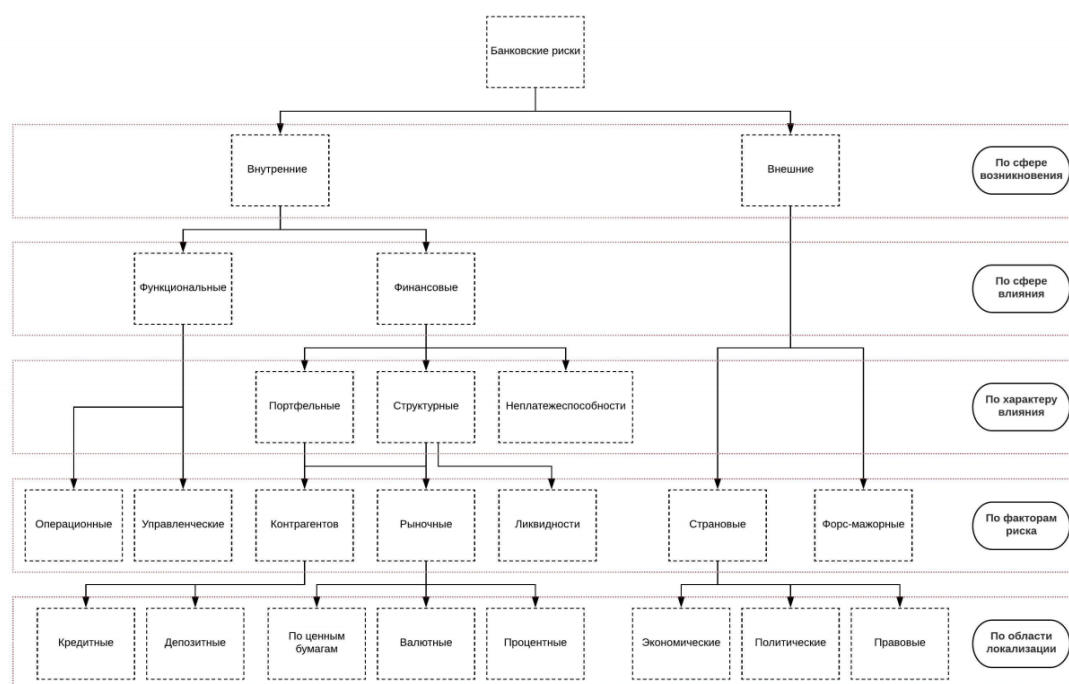


Рисунок 1 – Иерархия банковских рисков

Приведенная выше схема широко распространена в литературе и является одной из ключевых областей исследований. Особое внимание уделяется кредитным операциям, поскольку они составляют около 80% всех операций коммерческих банков. Следовательно, снижение кредитного риска является крайне важной задачей для банковской системы.

Кредитный риск – это вероятность того, что заемщик не вернет кредит, что в свою очередь может привести к финансовым потерям для кредитора. Оценка кредитного риска является одним из основных инструментов, которые используются финансовыми организациями при выдаче кредитов.

Кредитный риск включает несколько аспектов, таких как дефолт заемщика (невозврат кредита), просрочка платежей или недобросовестное поведение заемщика. Оценка и управление кредитным риском являются критическими задачами для банков, поскольку неправильное принятие решений может привести к финансовым потерям и негативным последствиям для банковской организации.

Существует множество подходов к оценке кредитного риска, от классических методов, таких как анализ финансовых показателей и статистические методы, до более современных, использующих методы машинного обучения. Классические методы оценки кредитного риска основаны на анализе финансовых показателей заемщика, таких как его доходы, расходы, имущество, а также его кредитная история [3]. Однако, такой подход имеет свои ограничения, поскольку он не учитывает некоторые важные факторы, такие как личная история заемщика, которая может повлиять на его способность выплатить кредит.

Методы машинного обучения в области оценки кредитного риска предоставляют новые возможности для определения риска заемщика. Эти методы основаны на использовании алгоритмов машинного обучения, которые обучаются на исторических данных, чтобы определить связи между различными факторами и вероятностью возврата кредита [4]. Такие методы могут использовать не только финансовые данные, но и другую информацию, например, данные социальных сетей, которые могут дать дополнительную информацию о заемщике.

Преимущества методов машинного обучения в оценке кредитного риска заключаются в возможности использования большего количества факторов для определения риска заемщика, а также в возможности улучшения точности

прогнозирования. Недостатком таких методов может быть их сложность и необходимость использования больших объемов данных для обучения алгоритмов [5].

Таким образом, изучение теоретических основ оценки кредитного риска и применения методов машинного обучения в этой области является важной задачей для финансовых организаций.

1.2.2. Подготовка и анализ данных.

Подготовка и анализ данных является важным этапом в построении моделей машинного обучения для оценки кредитных рисков. Этот этап включает в себя работу с данными, которые будут использоваться в моделях. Он помогает убедиться в качестве и целостности данных и подготовить их к использованию в моделях.

Для начала необходимо произвести заполнение пропущенных значений, что поможет избежать ошибок при анализе данных. Далее следует обработка выбросов и шумов, которые могут исказить результаты моделей. Это может быть осуществлено с помощью различных методов, таких как замена выбросов на среднее значение или медиану, замена значений модой для текстовых типов данных [6].

Также необходимо выполнить масштабирование и преобразование данных, чтобы обеспечить более точные результаты моделей. Например, некоторые переменные могут иметь сильно различные диапазоны значений, их следует преобразовать к единому масштабу [7].

Для анализа данных будет использован метод EDA (Exploratory Data Analysis). Он позволяет исследовать данные, выявлять связи между переменными, определять наиболее значимые переменные и выявлять выбросы. Этот метод позволяет обнаружить структуру данных и выявить любые аномалии, которые могут повлиять на результаты моделей машинного обучения [8].

В целом, подготовка и анализ данных является ключевым этапом в создании точных и надежных моделей машинного обучения для оценки кредитных рисков. Это позволяет убедиться в качестве данных и подготовить их для использования в моделях.

1.2.3. Выбор и применение моделей машинного обучения.

Одним из ключевых этапов работы по оценке кредитного риска на основе данных компании Home Credit является выбор и применение моделей машинного обучения. Для этого необходимо учитывать как особенности данных, так и требования к оценке кредитного риска.

Одной из наиболее распространенных моделей машинного обучения является логистическая регрессия. Эта модель применяется для прогнозирования вероятности наступления события, основываясь на наборе входных данных. Для данной работы логистическая регрессия может быть использована для определения вероятности невозврата кредита.

Логистическая регрессия является мощным инструментом для оценки кредитных рисков. Она позволяет анализировать взаимосвязь между различными факторами и вероятностью невыполнения кредитных обязательств.

В основе логистической регрессии лежит логистическая функция, которая преобразует линейную комбинацию факторов в вероятность. Эта вероятность оценивает шансы на то, что заемщик не сможет выполнить свои кредитные обязательства [9]. Логистическая функция обладает свойством ограничивать значения между 0 и 1, что делает ее идеальным выбором для задач классификации, таких как оценка кредитного риска.

При использовании логистической регрессии для оценки кредитных рисков, мы определяем набор факторов, которые могут влиять на вероятность невыполнения кредитных обязательств. Эти факторы могут включать доход заемщика, его кредитную историю, сумму кредита и другие соответствующие

переменные. Затем мы обучаем модель на исторических данных, где известен исход каждого кредита (выполнен или не выполнен).

В процессе обучения логистическая регрессия оценивает веса каждого фактора и формирует линейную комбинацию, которая предсказывает вероятность невыполнения кредитных обязательств [10]. После обучения модель может быть использована для предсказания вероятности невыполнения кредита для новых заемщиков.

Одним из преимуществ логистической регрессии является ее интерпретируемость [11]. Мы можем проанализировать вклад каждого фактора в предсказанную вероятность, что помогает нам понять, какие переменные наиболее существенно влияют на кредитный риск. Это позволяет банкам и финансовым учреждениям принимать более обоснованные решения в процессе выдачи кредитов.

Однако логистическая регрессия также имеет некоторые ограничения. Она предполагает линейную связь между факторами и логит-трансформированной вероятностью, что может быть недостаточным для моделирования сложных взаимосвязей [12]. В случае, если в данных присутствуют нелинейные зависимости, логистическая регрессия может давать менее точные предсказания.

Кроме того, логистическая регрессия предполагает независимость наблюдений [6, С. 247–248], что может быть нарушено в случае, если данные содержат зависимости между наблюдениями, например, при анализе кредитных заявок от членов одной семьи. В таких случаях может потребоваться применение более сложных моделей, способных учитывать зависимости между наблюдениями.

Необходимо также учитывать, что логистическая регрессия может быть подвержена проблеме мультиколлинеарности, когда факторы в модели сильно коррелируют между собой [6, С. 248–249]. Это может затруднить оценку весов факторов и усложнить интерпретацию результатов.

В качестве альтернативы логистической регрессии может быть использован случайный лес. Эта модель также широко используется в задачах классификации и регрессии. Случайный лес (Random Forest) является мощным алгоритмом для оценки кредитных рисков и представляет собой ансамбль решающих деревьев. Этот метод объединяет преимущества деревьев решений с принципом случайности, что делает его эффективным и устойчивым к переобучению.

Основная идея случайного леса заключается в построении большого количества деревьев решений на основе различных случайных подвыборок и случайных наборов признаков из обучающих данных. Каждое дерево строится независимо от остальных и затем комбинируется для получения окончательного предсказания [13].

Преимущество случайного леса заключается в его способности обрабатывать большое количество признаков и учитывать сложные взаимосвязи между ними [14], что как раз идеально подходит для выбранного датасета Home Credit, в исходном наборе которого 122 признака. Алгоритм автоматически выбирает наиболее информативные признаки, что упрощает процесс отбора переменных. Это особенно важно при анализе кредитных рисков, где существует множество факторов, влияющих на вероятность невыполнения кредитных обязательств.

Кроме того, случайный лес способен обрабатывать как числовые, так и категориальные данные без необходимости их предварительного преобразования [15]. Это делает его удобным инструментом для работы с различными типами данных, которые могут быть связаны с кредитными рисками.

Другим преимуществом случайного леса является его способность обнаруживать и устранять проблему переобучения. Благодаря случайному выбору подвыборок и признаков при построении каждого дерева, случайный лес ограничивает вероятность переобучения и обеспечивает более устойчивые и надежные предсказания [16].

Однако, следует отметить, что случайный лес имеет некоторые ограничения. Во-первых, построение и обучение большого числа деревьев может быть вычислительно затратным процессом. Также, в случае использования большого количества признаков, случайный лес может столкнуться с проблемой избыточности и неэффективности. В таких случаях может потребоваться применение методов отбора признаков или уменьшение размерности данных [17].

Кроме того, интерпретируемость случайного леса может быть ограничена. В отличие от логистической регрессии, где можно явно видеть влияние каждого фактора, случайный лес представляет собой комбинацию множества деревьев, что усложняет понимание важности каждого фактора.

В заключение, случайный лес является мощным алгоритмом для оценки кредитных рисков. Он сочетает в себе преимущества деревьев решений, способность обрабатывать большое количество признаков и устойчивость к переобучению. Случайный лес может быть эффективным инструментом для предсказания вероятности невыполнения кредитных обязательств и оценки кредитного риска.

При выборе моделей машинного обучения необходимо учитывать их преимущества и недостатки. Например, логистическая регрессия является простой и быстрой в реализации моделью, но она может быть неэффективной в случае нелинейных зависимостей между входными данными. Случайный лес, в свою очередь, может обладать высокой точностью прогнозирования, но может быть менее интерпретируемым и требовательным к вычислительным ресурсам.

Важным этапом работы с моделями машинного обучения является настройка их параметров для достижения максимальной эффективности. Например, в случае логистической регрессии можно настраивать параметры регуляризации, а в случае случайного леса – количество деревьев и глубину каждого дерева.

Таким образом, выбор и применение моделей машинного обучения является важным шагом в оценке кредитного риска на основе данных компании Home Credit. При этом необходимо учитывать особенности данных, а также преимущества и недостатки выбранных моделей.

1.2.4. Оценка эффективности моделей.

Оценка эффективности моделей является одним из ключевых этапов при работе с данными и моделями машинного обучения. Для этого необходимо выбрать метрики, которые наилучшим образом отражают цель задачи. В данном случае мы будем использовать метрики, такие как точность и ROC-кривая.

Точность (ассигасу) – это мера того, как много объектов было классифицировано правильно. Она определяется как отношение числа правильно классифицированных объектов ко всем объектам.

ROC-кривая – это график, который показывает зависимость между долей верных положительных классификаций (True Positive Rate) и долей ложных положительных классификаций (False Positive Rate). Чем ближе кривая к левому верхнему углу графика, тем выше качество модели [19].

После выбора метрик и применения моделей, необходимо оценить их эффективность на основе выбранных метрик. Для этого можно использовать кросс-валидацию, которая позволяет оценить качество модели на различных выборках данных. Кроме того, можно произвести анализ ошибок, чтобы выявить слабые места модели и улучшить ее эффективность. Результаты оценки эффективности моделей будут использованы для сравнения и выбора наиболее оптимальной модели.

1.2.5. Интерпретация результатов.

После оценки эффективности моделей машинного обучения на основе данных компании Home Credit необходимо проанализировать полученные результаты и их интерпретировать в контексте оценки кредитного риска. Это

поможет понять, какие факторы влияют на оценку кредитного риска и какие параметры моделей могут быть улучшены.

Для интерпретации результатов будут использованы методы анализа важности признаков [18, С.176]. Это позволит выявить наиболее важные признаки, влияющие на оценку кредитного риска, и проанализировать их значимость. Также будут проанализированы примеры ошибочной классификации, чтобы определить, какие типы ошибок наиболее часто возникают и как их можно уменьшить.

Результаты интерпретации могут быть использованы для улучшения моделей машинного обучения. Например, если наиболее важным признаком является доход заемщика, то увеличение объема данных о доходах заемщиков может улучшить модель. Также можно попробовать использовать другие методы машинного обучения или изменить параметры текущих моделей, чтобы улучшить их эффективность.

Таким образом, интерпретация результатов является важным этапом оценки кредитного риска на основе методов машинного обучения. Она позволяет определить факторы, влияющие на оценку кредитного риска, и выявить возможности для улучшения моделей машинного обучения.

Кроме того, интерпретация результатов также важна с точки зрения принятия решений и практической применимости. Результаты моделей машинного обучения могут быть использованы финансовыми учреждениями для принятия решений о выдаче кредитов. Понимание, какие факторы влияют на оценку кредитного риска, позволяет принимать обоснованные и информированные решения.

Например, если модель показывает, что наличие надежного залога является одним из наиболее важных факторов для оценки кредитного риска, финансовые учреждения могут принимать во внимание наличие или отсутствие залога при принятии решения о выдаче кредита. Также результаты интерпретации могут помочь в разработке стратегий управления рисками и принятии решений о ценообразовании.

Кроме того, интерпретация результатов может помочь в обеспечении соблюдения законодательства и этических норм. Если модель показывает, что определенный фактор (например, пол или раса) оказывает существенное влияние на оценку кредитного риска, это может указывать на наличие потенциальной дискриминации. В таком случае, необходимо провести дополнительный анализ и принять меры для устранения возможных негативных влияний и обеспечения справедливости и равноправия при выдаче кредитов.

1.2.6. Обзор литературы и предыдущих исследований.

Современная экономика не может функционировать без кредитования. Кредиты позволяют компаниям и частным лицам расширять свой бизнес, покупать недвижимость и автомобили, инвестировать в инфраструктуру и развитие технологий, а также погашать долги и выплачивать проценты. Однако выдача кредитов сопряжена с риском невозврата, который может нанести ущерб как кредиторам, так и заемщикам.

В связи с этим, многие банки и финансовые учреждения активно исследуют методы оценки кредитных рисков, чтобы минимизировать потери и повысить эффективность кредитных операций. В настоящее время методы машинного обучения все чаще применяются для анализа кредитного риска. Это связано с тем, что машинное обучение может эффективно обрабатывать большие объемы данных и выявлять скрытые зависимости и закономерности.

В данной работе будет рассмотрен обзор литературы и предыдущих исследований по оценке кредитных рисков с применением методов машинного обучения. В частности, будут рассмотрены следующие вопросы:

- Какие методы машинного обучения применяются для оценки кредитных рисков?
- Какие данные используются для обучения моделей?
- Какие результаты были получены в предыдущих исследованиях?

- Какие проблемы возникают при использовании методов машинного обучения для оценки кредитных рисков?
- Какие направления исследований в этой области будут актуальны в будущем?

Исследования, которые были проведены в области оценки кредитного риска с применением методов машинного обучения, показали, что эти методы являются эффективными инструментами для предсказания вероятности дефолта заемщиков и оценки кредитного риска. В результате использования алгоритмов машинного обучения вместо традиционных статистических методов, удалось достичь большей точности и улучшения качества прогнозов.

Одним из таких исследований было исследование «Динамическое ансамблевое обучение для оценки кредитоспособности», проведенное Исаевым Д.В. [20]. Целью исследования было разработать методику оценки кредитоспособности на основе динамического ансамблевого обучения. Авторы рассматривают проблему кредитного скоринга и предлагают подход, основанный на использовании ансамбля моделей машинного обучения.

В ходе исследования авторы используют несколько моделей машинного обучения, таких как случайный лес и градиентный бустинг, и объединяют их в ансамбль. Он применяет методы динамического обучения, которые позволяют адаптировать веса моделей в ансамбле в зависимости от изменяющихся условий и данных.

Автор описывает процесс обработки данных, который включает предобработку, выбор и преобразование признаков. Он также проводит эксперименты на реальных данных и измеряют точность и эффективность предложенного подхода с помощью различных метрик оценки качества моделей.

Недостатком данного исследования является отсутствие детальной информации о выборе гиперпараметров моделей и анализе их влияния на результаты. Также, несмотря на описание обработки данных, недостаточно

подробно представлена методика создания новых признаков, что может ограничить полноту оценки кредитного риска.

В целом, данное исследование предлагает интересный подход к оценке кредитоспособности с использованием динамического ансамблевого обучения. Однако, для полной оценки его эффективности и применимости необходимо более детальное исследование, учитывающее выбор гиперпараметров моделей и улучшение процесса создания новых признаков.

В исследовании «Использование методов машинного обучения в моделировании кредитного скоринга» проводился анализ и сравнение различных методов машинного обучения с целью определения наиболее эффективного подхода к оценке кредитоспособности заемщиков [21].

В исследовании авторы проводят анализ датасета, содержащего информацию о заемщиках и их кредитной истории. В работе были применены следующие методы машинного обучения: логистическая регрессия, случайный лес и градиентный бустинг.

Результаты исследования показали, что градиентный бустинг достиг лучшей точности предсказания кредитной способности заемщиков. Случайный лес и логистическая регрессия также продемонстрировали хорошие результаты, но чуть с меньшей точностью.

Однако, важно отметить, что в работе не представлена подробная информация о применяемых алгоритмах обработки данных и выбранных гиперпараметрах моделей. Также не указано, были ли проведены дополнительные действия по созданию новых признаков.

Ещё одно исследование «Программная модель оценки кредитоспособности клиентов с применением алгоритмов искусственного интеллекта» предлагает анализ и применение различных методов машинного обучения с целью создания эффективной модели скоринга кредитных рисков [22].

В исследовании была разработана и реализована программная модель, которая включает в себя следующие алгоритмы машинного обучения:

логистическая регрессия, случайный лес и градиентный бустинг. Авторы также описывают использование алгоритмов обработки данных, таких как стандартизация и балансировка классов, для улучшения производительности модели.

Результаты исследования показали, что разработанная модель достигла точности предсказания кредитоспособности клиентов на уровне около 0.85. Авторы отмечают, что это позволяет эффективно оценивать риски и принимать взвешенные решения при выдаче кредитов. Однако, следует отметить, что исследование не предоставляет подробной информации о размере и характеристиках использованного датасета, а также о выбранных гиперпараметрах моделей.

Исследование «A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function», опубликованное в журнале ScienceDirect, рассматривает применение методов машинного обучения в моделировании кредитного скоринга [23]. Авторы статьи предлагают новый подход к оценке кредитоспособности клиентов с использованием ансамблей моделей машинного обучения.

В исследовании была создана модель кредитного скоринга, основанная на ансамблевом обучении. Авторы предложили комбинировать несколько различных моделей машинного обучения, таких как логистическая регрессия, случайный лес и градиентный бустинг, для достижения более высокой точности предсказания кредитоспособности.

Используя набор данных о кредитных заявках, авторы провели эксперименты с различными ансамблями моделей. Результаты исследования показали, что предложенный подход позволяет достичь высокой точности предсказания кредитного риска, превосходящей результаты отдельных моделей. В частности, модель ансамблевого обучения показала точность прогнозирования около 0.85.

Однако, следует отметить, что в исследовании не было проведено подробного анализа процесса обработки данных, выбора гиперпараметров или создания дополнительных признаков. Также недостаточно информации о размере и характеристиках использованного датасета, что ограничивает полное понимание работы и обобщение результатов.

На платформе Kaggle было проведено несколько исследований, связанных с оценкой кредитных рисков на основе датасета «Home Credit Default Risk». Рассмотрим некоторые из них.

Исследование автора «MLRiskAnalysis» использовало алгоритм градиентного бустинга (XGBoost) для оценки кредитного риска. Были удалены наблюдения с отсутствующими значениями, а числовые признаки были масштабированы с помощью стандартизации. Категориальные признаки были преобразованы в числовой формат с помощью метода кодирования One-Hot.

При подготовке модели авторы выбрали определенные гиперпараметры для алгоритма XGBoost. Количество деревьев и скорость обучения были подобраны с использованием метода кросс-валидации и опыта предыдущих исследований. Результаты исследования показали, что модель градиентного бустинга (XGBoost) достигла точности около 0.72 в предсказании кредитного риска.

Однако, следует отметить, что недостатком данного исследования является отсутствие создания новых признаков. Также в описании работы не было представлено подробной информации о методах предварительной обработки данных, что ограничивает понимание полного процесса обработки и анализа данных.

Исследование команды исследователей «RiskPredictionTeam» также использовало алгоритм градиентного бустинга (LightGBM) для оценки кредитного риска. При обработке данных авторы осуществили аналогичные шаги предварительной обработки, включая удаление отсутствующих значений и масштабирование числовых признаков.

Подбор гиперпараметров модели включал перебор наборов значений и использование метода кросс-валидации для определения наилучших параметров. Однако, в описании исследования не представлены конкретные значения выбранных гиперпараметров.

Результаты показали, что модель градиентного бустинга (LightGBM) достигла точности около 0.71 в предсказании кредитного риска. Однако, стоит отметить некоторые недостатки данного исследования. Во-первых, подобно предыдущему исследованию, не было проведено создание новых признаков. Это может ограничить способность модели в выявлении скрытых зависимостей и улучшении точности предсказаний.

Кроме того, описание методов предварительной обработки данных было ограничено в данном исследовании. Более детальное описание примененных методов и техник может помочь другим исследователям в повторении и расширении этой работы.

2. Теоретические основы методов обработки данных и машинного обучения в финансовой области

2.1. Общие принципы и определения.

В современной финансовой индустрии данные играют огромную роль, и обработка больших объемов информации является критически важным элементом принятия решений. В связи с этим возникла необходимость разработки методов обработки данных и машинного обучения, которые позволяют автоматически анализировать информацию и принимать решения на основе полученных результатов.

Машинное обучение – это подход к обработке данных, в котором компьютерная программа обучается на основе имеющихся данных и использует полученные знания для решения новых задач. Одним из основных принципов машинного обучения является использование алгоритмов, которые позволяют автоматически находить закономерности в данных и прогнозировать будущие значения [24].

2.2. Методы обработки и анализа табличных данных

Методы обработки и анализа табличных данных в рамках темы играют важную роль в процессе оценки и прогнозирования кредитных рисков. В данном разделе рассмотрим методологии обработки данных, методологию анализа данных и выбор показателей [25].

1. Методологии обработки данных: Обработка данных является неотъемлемой частью процесса оценки кредитных рисков. Она включает в себя предварительную обработку данных, агрегацию, нормализацию, заполнение пропущенных значений и выбор значимых признаков. В рамках диплома можно использовать следующие методологии обработки данных:

- **Предварительная обработка данных:** В данном этапе производится проверка качества данных, удаление дубликатов, обработка выбросов и аномалий. Также может быть выполнена стандартизация данных для обеспечения их сопоставимости.

- **Агрегация данных:** В случае, если данные по кредитным рискам представлены в различных источниках или форматах, необходимо провести агрегацию данных, чтобы объединить их в одну таблицу. Это позволяет получить единый набор данных для дальнейшего анализа.

- **Нормализация данных:** Нормализация данных является важным этапом, особенно при использовании методов машинного обучения. Она позволяет привести данные к одному масштабу и избежать проблемы с несбалансированными весами признаков.

- **Заполнение пропущенных значений:** В реальных данных часто встречаются пропущенные значения. Для их заполнения можно использовать различные подходы, например, заполнение средними значениями, интерполяцию или использование алгоритмов машинного обучения.

- **Выбор значимых признаков:** Важно отобрать наиболее значимые признаки, которые влияют на оценку кредитных рисков. Для этого можно применить различные методы, такие как анализ важности признаков

(например, с использованием алгоритма случайного леса) или корреляционный анализ.

2. Методология анализа данных: Методология анализа данных включает в себя ряд этапов и методов, которые помогают выявить закономерности, связи и паттерны в данных, а также сделать выводы и прогнозы относительно кредитных рисков. В контексте диплома, следующие методологические подходы могут быть применены:

- Исследовательский анализ данных (Exploratory Data Analysis, EDA): Этот подход включает в себя изучение и визуализацию данных с целью выявления основных характеристик и паттернов. Это позволяет получить общее представление о данных и их распределении, а также выявить потенциальные аномалии или выбросы.

- Статистический анализ данных: Включает применение статистических методов для оценки степени связи между различными переменными и их влияния на кредитные риски. Это может включать расчет корреляции, проведение статистических тестов и построение регрессионных моделей для прогнозирования рисков.

- Методы машинного обучения: Методы машинного обучения являются мощным инструментом для анализа и прогнозирования кредитных рисков. Они включают в себя алгоритмы классификации, регрессии и кластеризации, которые могут использоваться для создания моделей оценки кредитных рисков на основе исторических данных.

- Визуализация данных: Визуализация данных является эффективным способом представления сложных данных в понятной форме. Использование различных графических инструментов и диаграмм позволяет наглядно представить результаты анализа и сделать выводы.

3. Выбор показателей: Выбор правильных показателей является критическим шагом в оценке кредитных рисков. Это включает определение релевантных финансовых, экономических и демографических показателей, которые могут влиять на вероятность невыполнения кредитных обязательств.

При выборе показателей следует учитывать их доступность, актуальность и значимость для целей анализа. Важно также применять методы отбора признаков, которые позволяют выделить наиболее информативные и значимые показатели. Некоторые из распространенных методов отбора признаков включают:

- Корреляционный анализ: позволяет определить степень взаимосвязи между различными показателями и целевой переменной (кредитным риском). Признаки с высокой корреляцией с целевой переменной могут быть выбраны для включения в модель.

- Методы отбора на основе важности признаков: Эти методы, такие как случайный лес, оценивают важность каждого признака в предсказании целевой переменной. Признаки с высокой важностью могут быть выбраны для включения в модель. При выборе показателей важно учитывать допустимость и применимость этих показателей для анализа кредитных рисков, а также обеспечить их достоверность и актуальность.

4. Оценка качества моделей: После обработки данных, выбора показателей и построения моделей оценки кредитных рисков с применением методов машинного обучения, необходимо провести оценку качества этих моделей. Это поможет определить их точность, надежность и пригодность для использования в практических целях. Некоторые методологии оценки качества моделей включают в себя:

- Разделение выборки: Обычно данные разделяют на обучающую и тестовую выборки. Обучающая выборка используется для тренировки модели, а тестовая выборка – для оценки ее производительности на новых данных, которые модель ранее не видела.

- Метрики оценки: Используются различные метрики для оценки качества моделей, такие как точность (accuracy) и ROC-AUC. Выбор подходящей метрики зависит от специфики задачи и требований бизнеса.

- Кросс-валидация: Этот метод позволяет оценить производительность модели на различных подмножествах данных путем

повторного разделения на обучающие и тестовые выборки. Кросс-валидация обеспечивает более надежные оценки и помогает обнаружить проблемы с переобучением или недообучением модели.

- Анализ ошибок: При оценке качества модели важно также проанализировать ее ошибки. Это позволяет идентифицировать основные типы ошибок и понять, какие категории кредитных рисков могут быть недооценены или переоценены моделью.

- Валидация на независимых данных: после оценки качества модели на тестовой выборке, желательно провести ее валидацию на независимых данных. Это поможет подтвердить устойчивость и надежность модели на новых, ранее неизвестных данных. Важно отметить, что выбор конкретных методологий обработки данных, анализа данных и оценки качества моделей может зависеть от конкретных характеристик и целей исследования, а также доступности и качества данных.

5. Генерация признаков: Генерация признаков является важным этапом в анализе и моделировании кредитных рисков. Она позволяет создавать новые переменные, которые могут содержать дополнительную информацию и улучшить способность моделей оценивать кредитные риски. В контексте оценки кредитных рисков с применением методов машинного обучения, существуют различные способы генерации признаков, включая анализ предметной области и использование генетического программирования.

6. Анализ предметной области: Перед началом генерации новых признаков полезно провести анализ предметной области кредитных рисков. Это позволяет понять основные факторы, влияющие на кредитные риски, и выявить потенциально значимые переменные. Например, такие как доход заемщика, семейное положение, история кредитных платежей и другие, которые могут быть использованы для генерации новых признаков.

7. Генетическое программирование: Генетическое программирование (ГП) является эволюционным алгоритмом, используемым

для автоматического создания программ или моделей, способных решать задачи. В контексте генерации признаков для оценки кредитных рисков, ГП может быть применено для создания новых математических выражений или функций, основываясь на существующих переменных [26]. Процесс генетического программирования обычно включает следующие шаги:

- Инициализация популяции: Создание начальной популяции программ, которые представляют возможные признаки или выражения.
- Оценка пригодности: Процедура оценки качества каждой программы в популяции на основе заданных метрик, таких как точность модели или коэффициенты важности признаков.
- Скрещивание и мутация: Процесс комбинирования программ (скрещивание) и внесения случайных изменений в программы (мутация) для создания новых потомков.
- Выбор лучших решений: Выбор наиболее приспособленных программ из популяции для формирования следующего поколения путем применения стратегий отбора, таких как турнирный отбор или рулеточное колесо.
- Итерации процесса: Повторение шагов до достижения заданного условия остановки, например, определенного числа поколений или достижения определенного уровня качества решений.

Генетическое программирование позволяет автоматически создавать новые признаки, комбинируя существующие переменные и математические операции. Например, можно создать новый признак, объединяющий информацию о доходе заемщика и его семейном положении, или использовать математические функции, такие как логарифм или степенная функция, для создания новых выражений.

При использовании генетического программирования для генерации признаков важно учитывать следующие аспекты:

- **Определение функционального пространства:** Необходимо определить набор доступных математических функций и операций, которые могут быть использованы при создании новых признаков.

- **Оценка качества и сложности признаков:** Не все созданные признаки могут быть полезными или иметь высокую предсказательную способность. Поэтому важно проводить оценку и отбор признаков на основе их важности и сложности.

- **Выбор метрик оценки:** Для оценки качества созданных признаков и их влияния на модель необходимо выбрать соответствующие метрики, такие как улучшение точности модели или уменьшение ошибок.

Генерация признаков с использованием генетического программирования может значительно расширить пространство признаков и помочь выявить сложные взаимосвязи и закономерности, которые могут быть полезными для прогнозирования кредитных рисков. Однако, важно провести тщательный анализ и оценку полученных признаков, чтобы убедиться в их релевантности и качестве перед их включением в модель оценки рисков.

Процесс генерации признаков на основе анализа предметной области и генетического программирования может привести к созданию разнообразных и информативных переменных для оценки кредитных рисков. Ниже приведены некоторые примеры генерации признаков:

- **Суммарный доход заемщика:** можно создать новый признак, который представляет собой сумму доходов всех членов семьи заемщика. Это может быть полезным показателем для оценки общей финансовой стабильности и способности заемщика погасить кредит.

- **История платежей:** на основе истории платежей заемщика можно сгенерировать различные признаки, такие как среднее значение задолженности, доля просроченных платежей или количество пропущенных платежей. Эти признаки могут указывать на платежную дисциплину и надежность заемщика.

- Кредитный рейтинг: используя информацию о платежной истории, доходах, задолженностях и других факторах, можно создать новый признак, представляющий собой кредитный рейтинг заемщика. Этот признак может быть полезным индикатором для оценки кредитоспособности и рисков.

- Доля задолженности по кредитам: можно создать признак, отражающий долю задолженности заемщика по всем кредитам в отношении его дохода. Это позволяет оценить финансовую нагрузку заемщика и его способность управлять задолженностью.

- Стабильность работы: на основе информации о занятости заемщика и длительности работы на текущем месте работы можно создать признак, указывающий на стабильность его дохода и финансового положения. Это может быть важным фактором для оценки кредитного риска.

Это лишь некоторые примеры генерации признаков, и возможности варьируются в зависимости от доступных данных и характеристик предметной области. Важно подходить к генерации признаков творчески, учитывая особенности и специфику задачи оценки кредитных рисков.

После генерации новых признаков на основе анализа предметной области и генетического программирования, необходимо провести их выбор и оценку. Этот этап поможет определить наиболее значимые и информативные переменные для использования в модели оценки кредитных рисков. Важные шаги в выборе и оценке сгенерированных признаков включают:

- Корреляционный анализ: исследование корреляционных связей между сгенерированными признаками и целевой переменной (например, категориями кредитного риска). Признаки с высокой корреляцией могут иметь большее влияние на предсказание риска и могут быть предпочтительными для включения в модель.

- Важность признаков: применение алгоритмов машинного обучения, такие как случайный лес, для оценки важности признаков. Эти

алгоритмы могут определить, какие признаки вносят наибольший вклад в предсказание кредитных рисков.

- Мультиколлинеарность: мультиколлинеарные признаки могут быть избыточными и могут затруднить интерпретацию и стабильность модели. В этом случае, необходимо принять решение об исключении одного из коррелирующих признаков.

3.3. Основы машинного обучения и алгоритмы классификации, регрессии и кластеризации

Машинное обучение является ключевым инструментом в оценке кредитных рисков и применяется для построения моделей, способных классифицировать заемщиков на основе доступных данных. В этом разделе рассмотрим основы машинного обучения и некоторые из наиболее распространенных алгоритмов классификации, регрессии и кластеризации, которые могут быть применены для анализа данных и оценки кредитных рисков.

Машинное обучение – это область искусственного интеллекта, которая изучает алгоритмы и модели, способные автоматически обучаться на основе данных и делать предсказания или принимать решения без явного программирования. Основные понятия, связанные с машинным обучением, включают:

Обучающая выборка: Набор данных, на котором модель будет обучаться. Обучающая выборка состоит из примеров, где каждый пример представляет собой набор признаков и соответствующую метку или целевую переменную.

1. Признаки: характеристики или атрибуты, описывающие объекты или сущности в данных. Признаки используются для описания заемщиков и их финансовых характеристик.

2. Модель: математическое представление, созданное на основе обучающей выборки, которое может делать предсказания или

классифицировать новые данные. Модель является результатом процесса обучения.

3. Обучение: процесс настройки модели на основе обучающей выборки. Во время обучения модель «учится» находить закономерности и связи между признаками и целевой переменной.

4. Тестовая выборка: независимый набор данных, используемый для оценки производительности модели на новых данных. Тестовая выборка помогает оценить обобщающую способность модели.

Алгоритмы классификации используются для прогнозирования принадлежности объектов к определенным классам. В контексте оценки кредитных рисков, классификационные алгоритмы могут помочь определить, относится ли заемщик к низкому, среднему или высокому риску.

1. Логистическая регрессия: Это один из наиболее распространенных алгоритмов классификации. Логистическая регрессия моделирует вероятность принадлежности объекта к определенному классу, используя логистическую функцию. При применении к оценке кредитных рисков, логистическая регрессия может использоваться для прогнозирования вероятности заемщика быть в неблагоприятной категории риска.

2. Решающие деревья: Решающие деревья строятся на основе иерархической структуры, где каждый узел представляет условие по одному из признаков. Они помогают разбить данные на подмножества, в результате чего объекты в одной ветви будут иметь схожие характеристики. Решающие деревья могут быть полезными для идентификации ключевых признаков, влияющих на классификацию заемщиков.

3. Метод опорных векторов (SVM): SVM является алгоритмом классификации, который стремится найти гиперплоскость в пространстве признаков, максимально разделяющую объекты разных классов. В оценке кредитных рисков, SVM может быть применен для разделения заемщиков на категории низкого, среднего и высокого риска на основе их финансовых характеристик [27, С. 237].

Алгоритмы регрессии используются для предсказания непрерывной целевой переменной. В контексте оценки кредитных рисков, алгоритмы регрессии могут помочь в определении вероятности дефолта или оценке размера кредитного лимита для заемщика.

1. Линейная регрессия: Линейная регрессия моделирует линейную зависимость между признаками и целевой переменной. Она может быть применена для прогнозирования непрерывной переменной, такой как размер кредитного лимита, на основе финансовых характеристик заемщика. Линейная регрессия и ее вариации, такие как множественная линейная регрессия или регрессия с регуляризацией, могут быть полезными инструментами в оценке кредитных рисков.
2. Метод k-ближайших соседей (k-NN): Метод k-NN предполагает, что объекты, близкие в пространстве признаков, имеют схожие значения целевой переменной. Он использует ближайших соседей для предсказания значения целевой переменной для новых объектов. В оценке кредитных рисков, метод k-NN может помочь в прогнозировании вероятности дефолта на основе близости заемщика к схожим по финансовым характеристикам заемщикам из обучающей выборки [28].

Алгоритмы кластеризации используются для группировки объектов на основе их сходства. В контексте оценки кредитных рисков, кластеризация может помочь выявить группы заемщиков с похожими финансовыми характеристиками и рисками.

1. Метод k-средних разделяет объекты на k кластеров, минимизируя среднеквадратичное расстояние между объектами и их центроидами. В оценке кредитных рисков, метод k-средних может помочь выявить группы заемщиков с схожими финансовыми характеристиками, что может быть полезно для более точной оценки рисков и разработки индивидуальных стратегий кредитования [27, С.238].

2. DBSCAN: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) основывается на плотности точек в пространстве признаков. Он способен выявлять кластеры различной формы и определять шумовые точки. В контексте оценки кредитных рисков, DBSCAN может помочь идентифицировать группы заемщиков с общими финансовыми характеристиками и выявить аномалии, которые могут указывать на потенциальные риски.

В этом разделе мы рассмотрели основы машинного обучения и представили некоторые из наиболее распространенных алгоритмов классификации, регрессии и кластеризации. Логистическая регрессия, решающие деревья, метод опорных векторов, линейная регрессия, метод k-ближайших соседей, метод k-средних и DBSCAN – все эти алгоритмы могут быть применены для анализа финансовых данных и оценки кредитных рисков.

Однако выбор конкретного алгоритма или их комбинации зависит от специфики задачи и доступных данных. Перед использованием любого алгоритма необходимо провести предварительный анализ данных, обработку пропущенных значений, масштабирование признаков и выбор соответствующих метрик оценки моделей.

В рамках данной дипломной работе мы на практике рассмотрим логистическую регрессию и случайный лес в качестве моделей машинного обучения для прогнозирования дефолта клиента кредитной организации.

В следующем разделе мы рассмотрим практическую реализацию оценки кредитных рисков с использованием машинного обучения. Будут представлены примеры предварительной обработки данных, выбора и обучения модели, а также оценки ее производительности и интерпретации результатов.

3. Анализ данных в финансовой отрасли

3.1. Описание датасета

Датасет Home Credit содержит данные о заемщиках, подавших заявки на кредит, и их платежеспособности. Для оценки кредитного риска были собраны различные данные, включая социально-экономические данные, данные по кредитной истории, информацию о текущих кредитах и финансовых обязательствах, а также данные о работе и образовании.

Схема данных:

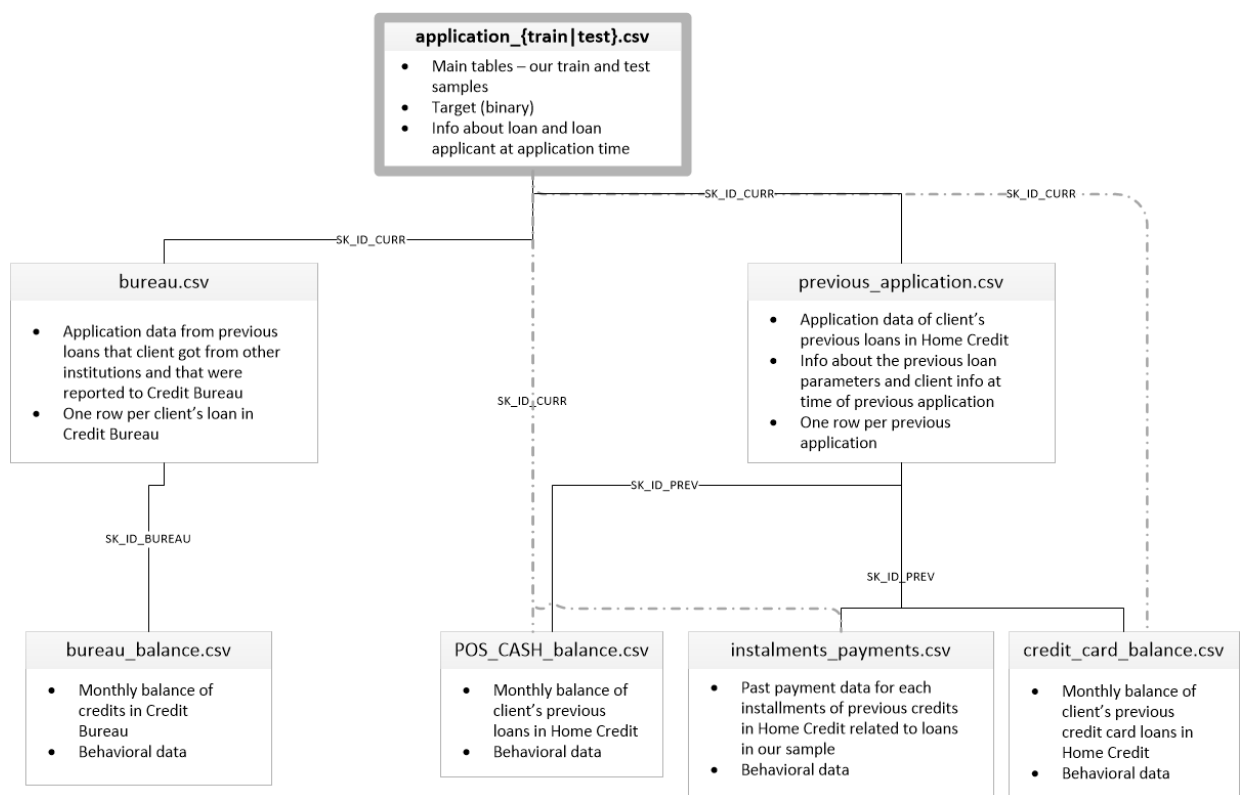


Рисунок 2 – Описание и схема датасета Home Credit

Переменные можно объединить в несколько важных категорий:

1. Первая категория включает данные о заемщиках, такие как возраст, пол, семейное положение и количество детей.
2. Вторая категория включает информацию о работе и образовании, включая тип занятости, доход и образовательный уровень.
3. Третья категория относится к истории заемщика и содержит данные о текущих и прошлых кредитах, а также о задолженностях по кредитам.

4. Четвертая категория связана с имуществом заемщика, включая наличие или отсутствие недвижимости и транспортных средств.
5. Пятая категория содержит информацию о заявке на кредит, такую как цель кредита и запрашиваемая сумма.

Некоторые из наиболее важных признаков включают образование и тип занятости заемщика, так как это может указывать на потенциальный доход заемщика и его способность выплачивать кредиты. Информация о задолженностях по текущим и прошлым кредитам также является важным показателем, так как это может указывать на риски невозврата кредита. Данные о наличии недвижимости и транспортных средств также могут быть полезны при оценке кредитного риска, так как это может служить дополнительным обеспечением для займа.

Одним из ключевых аспектов, определяющих возможность одобрения заявки на кредит, является финансовая устойчивость заемщика. В этом контексте важными категориями переменных являются доходы и затраты заемщика, а также общая задолженность перед другими кредиторами.

Для оценки финансовой устойчивости заемщика в датасете представлены следующие переменные:

AMT_INCOME_TOTAL: общий доход заемщика;
AMT_CREDIT: запрошенная сумма кредита;
AMT_ANNUITY: размер ежемесячного платежа по кредиту;
AMT_GOODS_PRICE: стоимость товаров, на которые берется кредит;
NAME_INCOME_TYPE: тип занятости заемщика;
OCCUPATION_TYPE: тип занятости заемщика;
NAME_EDUCATION_TYPE: уровень образования заемщика;
NAME_FAMILY_STATUS: семейное положение заемщика;
NAME_HOUSING_TYPE: тип жилья, в котором проживает заемщик;
DAYS_EMPLOYED: количество дней, которые заемщик находится на текущей работе;
DAYS_BIRTH: возраст заемщика в днях.

Кроме того, в датасете представлены переменные, связанные с историей кредитования заемщика:

CNT_CHILDREN: количество детей у заемщика;

NAME_CONTRACT_TYPE: тип кредита;

CODE_GENDER: пол заемщика;

FLAG_OWN_CAR: наличие у заемщика автомобиля;

FLAG_OWN_REALTY: наличие у заемщика недвижимости;

REGION_POPULATION_RELATIVE: уровень населения в регионе, где проживает заемщик;

REGION_RATING_CLIENT: рейтинг региона, где проживает заемщик;

REG_CITY_NOT_LIVE_CITY: проживание заемщика в городе, отличном от города, где был выдан кредит;

REG_CITY_NOT_WORK_CITY: работа заемщика в городе, отличном от города, где был выдан кредит;

FLAG_EMP_PHONE: наличие у заемщика телефона на месте работы;

FLAG_WORK_PHONE: наличие у заемщика рабочего телефона;

FLAG_CONT_MOBILE: наличие у заемщика мобильного телефона;

FLAG_EMAIL: наличие у заемщика электронной почты;

DAYS_LAST_PHONE_CHANGE: количество дней с момента последней смены номера телефона.

Категория Кредитная история и текущее состояние кредита:

AMT_CREDIT – сумма кредита

AMT_ANNUITY – ежемесячный платеж

DAYS_CREDIT – количество дней с момента предыдущего кредита

DAYS_ENDDATE_FACT – количество дней, когда фактически была закрыта предыдущая задолженность

DAYS_CREDIT_ENDDATE – количество дней до закрытия предыдущей задолженности

DAYS_BIRTH – возраст клиента в днях

DAYS_EMPLOYED – количество дней на текущем месте работы

Категория Семейное и социальное положение:

NAME_EDUCATION_TYPE – уровень образования

NAME_FAMILY_STATUS – семейное положение

NAME_HOUSING_TYPE – тип жилья

CNT_CHILDREN – количество детей

Категория Финансовые данные:

AMT_INCOME_TOTAL – общий доход

AMT_GOODS_PRICE – стоимость товара или недвижимости

REGION_POPULATION_RELATIVE – население региона

REGION_RATING_CLIENT – рейтинг клиента по региону

REGION_RATING_CLIENT_W_CITY – рейтинг клиента по региону с

учетом города

Категория Информация о работе:

OCCUPATION_TYPE – должность

ORGANIZATION_TYPE – тип организации

FLAG_WORK_PHONE – наличие рабочего телефона

Категория «Информация о залоге»:

FLAG_OWN_CAR – наличие автомобиля

FLAG_OWN_REALTY – наличие недвижимости

OWN_CAR_AGE – возраст автомобиля

Категория Кредитная история в других организациях:

CNT_FAM_MEMBERS – количество членов семьи

NAME_CONTRACT_STATUS – статус кредитного договора в других

организациях

NAME_PAYMENT_TYPE – тип платежа в других организациях

DAYS_LAST_DUE – количество дней до последнего платежа в других

организациях

Категория Другие данные:

NAME_TYPE_SUITE – тип жилья, с кем проживает клиент

CODE_GENDER – пол клиента

FLAG_EMAIL – наличие электронной почты клиента

FLAG_PHONE – наличие телефона клиента

NAME_INCOME_TYPE – источник дохода

NAME_CONTRACT_TYPE – тип кредита

Кроме основных категорий, в датасете Home Credit имеются также дополнительные переменные, относящиеся к личным и социальным характеристикам клиентов. Например, есть данные о семейном положении, наличии детей, образовании, занятости и доходах. Эти переменные могут быть полезны для дополнительного анализа и построения моделей кредитного скоринга. Кроме того, в датасете есть также временные ряды, например, история платежей клиента по предыдущим кредитам. Эти данные могут быть полезны для прогнозирования будущих платежей и вероятности дефолта.

Таким образом, датасет Home Credit представляет собой комплексный набор данных о клиентах, их финансовом состоянии, предыдущей кредитной истории и личных характеристиках. Анализ и обработка этих данных могут помочь в прогнозировании кредитного риска и принятии решений по выдаче кредитов.

3.2. Предобработка данных и преобразование признаков

Предобработка данных является одним из наиболее важных шагов в анализе данных, так как именно на этом этапе можно сильно повлиять на результаты анализа.

Предобработка данных включает в себя ряд различных процессов, таких как заполнение пропущенных значений, удаление дубликатов, обработка выбросов и масштабирование признаков. Рассмотрим некоторые наиболее распространенные методы предобработки данных и преобразования признаков.

1. Заполнение пропущенных значений. Одна из наиболее распространенных проблем – это пропущенные значения. Пропущенные значения могут возникать по разным причинам, таким как неправильное считывание данных, либо некоторые значения могут быть неизвестны.

Пропущенные значения могут исказить результаты анализа данных, поэтому необходимо заполнять их.

2. Удаление дубликатов. Дубликаты могут возникать по разным причинам, таким как ошибки ввода или дублирование данных. Удаление дубликатов является важным шагом в предобработке данных, так как они могут привести к искажению результатов анализа данных.

3. Обработка выбросов. Выбросы — это значения, которые находятся за пределами ожидаемого диапазона значений. Выбросы могут возникать по разным причинам, таким как ошибки ввода данных или ошибки измерений. Выбросы могут значительно исказить результаты анализа данных, поэтому их необходимо обрабатывать.

4. Масштабирование признаков. Масштабирование признаков — это процесс изменения масштаба значений признаков. Масштабирование признаков может улучшить производительность алгоритмов машинного обучения, так как многие алгоритмы чувствительны к различным масштабам значений признаков. Существуют различные методы масштабирования признаков, такие как стандартизация и нормализация.

5. Преобразование категориальных признаков. Категориальные признаки — это признаки, которые могут принимать ограниченное количество значений. Примерами категориальных признаков могут служить пол, город проживания, тип автомобиля и т.д. Для анализа данных необходимо преобразовать категориальные признаки в числовые, так как большинство алгоритмов машинного обучения работают только с числовыми значениями.

6. Удаление лишних признаков. В данных могут быть признаки, которые не имеют важности для анализа или которые могут исказить результаты. Удаление этих признаков может упростить анализ и улучшить производительность модели.

7. Работа с несбалансированными данными. В некоторых задачах машинного обучения данные могут быть несбалансированными, то есть

классы могут иметь различные размеры. Несбалансированные данные могут привести к смещению модели в сторону более крупных классов, что может привести к неверным результатам. Для работы с несбалансированными данными можно использовать методы, такие как изменение порогового значения классификации, использование взвешивания классов или использование методов сэмплирования, таких как андерсэмплинг или оверсэмплинг.

8. Визуализация данных. Визуализация данных является важным инструментом в анализе данных. Она позволяет наглядно представить данные и выявить закономерности и зависимости между признаками. Для визуализации данных можно использовать различные библиотеки Python, такие как Matplotlib и Seaborn.

9. Преобразование признаков. Преобразование признаков может улучшить производительность модели, сделать данные более информативными и уменьшить размерность данных. Преобразование признаков может включать в себя такие методы, как нормализация, стандартизация, преобразование категориальных признаков в числовые и т.д.

10. Выбор признаков. Выбор признаков может улучшить производительность модели, сделать данные более информативными и уменьшить размерность данных. Некоторые признаки могут быть неинформативными или коррелировать с другими признаками, поэтому необходимо выбирать только значимые признаки. Для выбора признаков можно использовать методы, такие как анализ корреляции, методы основанные на статистике, рекурсивное удаление признаков и т.д.

11. Создание новых признаков. Создание новых признаков может улучшить производительность модели и сделать данные более информативными. Новые признаки могут быть созданы на основе существующих признаков, используя методы, такие как агрегирование, бинаризация, генерация признаков на основе текстовых данных и т.д.

12. Оценка качества данных. Оценка качества данных является важным шагом в предобработке данных и преобразовании признаков. Необходимо убедиться в том, что данные достаточно качественные и соответствуют требованиям задачи. Для оценки качества данных можно использовать методы, такие как анализ выбросов, анализ корреляции, анализ распределения данных и т.д.

В заключение, предобработка данных и преобразование признаков являются неотъемлемыми шагами в анализе данных и машинном обучении. Необходимо учитывать особенности данных и выбирать соответствующие методы для их обработки. Различные библиотеки Python позволяют автоматизировать процесс предобработки данных, что позволяет ускорить работу и повысить эффективность анализа. Однако, необходимо помнить, что выбор методов предобработки данных и преобразования признаков может существенно влиять на результаты модели, поэтому следует проявлять осторожность и тщательно проверять результаты.

3.3. Исследовательский анализ данных

Для проведения анализа кредитных рисков на основе датасета «Home Credit», мы можем использовать данные об истории кредитных заявок и платежах, а также другую информацию о клиентах, предоставленную компанией Home Credit.

1. Анализ пропущенных данных: В начале исследования был выполнен анализ датасета для выявления пропущенных значений. Была проанализирована структура датасета и определены переменные с пропущенными значениями.

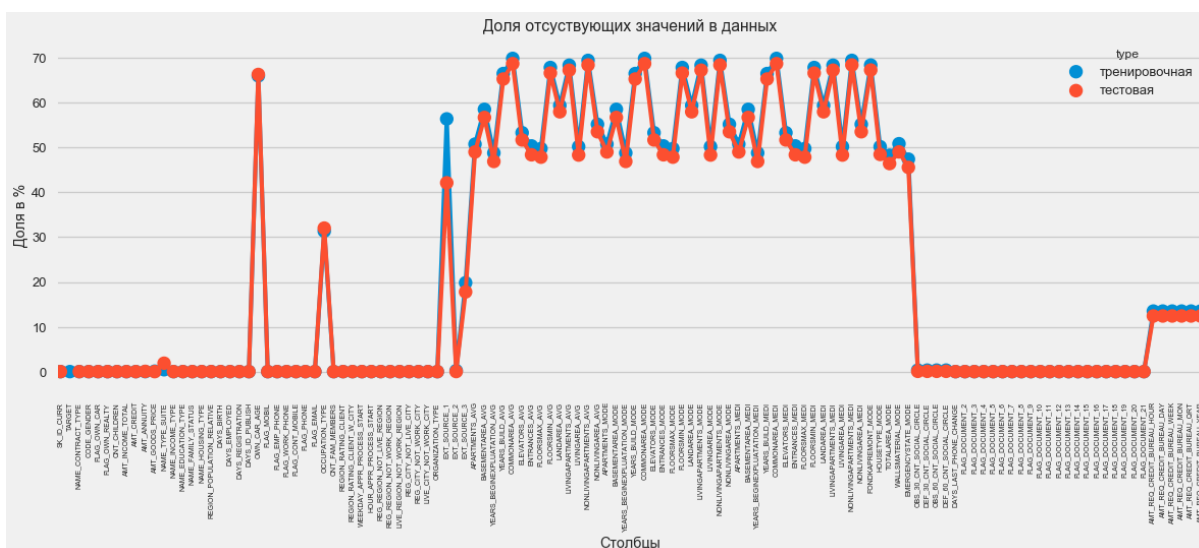


Рисунок 3 – Распределение пропущенных значений по признакам в датасете Home Credit

Как видно на Рисунке 3, около половины признаков в датасете имеет недостающие значения, которые предстоит заменить в рамках подготовки данных к дальнейшей работе. График распределения пропущенных значений также позволяет понять, с какими переменными потребует дальнейшая работа по замене пустых значений.

2. Определим количество дубликатов в выборках:

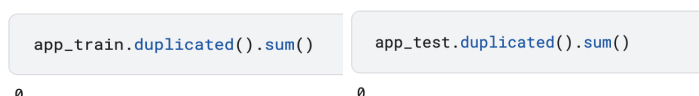


Рисунок 4 – Анализ количества дубликатов в датасете Home Credit

В датасете дубликатов не найдено. Чтобы принять решения относительно того, что делать с недостающими значениями, проверим информацию о числовых переменных. Используем метод `describe()` для получения статистических данных по столбцам.

	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	EXT_SOURCE_2	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE
count	307499.000000	3.072330e+05	307509.000000	3.068510e+05	306490.000000	306490.000000
mean	27108.573909	5.383962e+05	2.152665	5.143927e-01	1.422245	0.143421
std	14493.737315	3.694465e+05	0.910682	1.910602e-01	2.400989	0.446698
min	1615.500000	4.050000e+04	1.000000	8.173617e-08	0.000000	0.000000
25%	16524.000000	2.385000e+05	2.000000	3.924574e-01	0.000000	0.000000
50%	24903.000000	4.500000e+05	2.000000	5.659614e-01	0.000000	0.000000
75%	34596.000000	6.795000e+05	3.000000	6.636171e-01	2.000000	0.000000
max	258025.500000	4.050000e+06	20.000000	8.549997e-01	348.000000	34.000000

Рисунок 5 – Статистические данные по столбцам датасета Home Credit

Посмотрим на распределение признаков на графиках

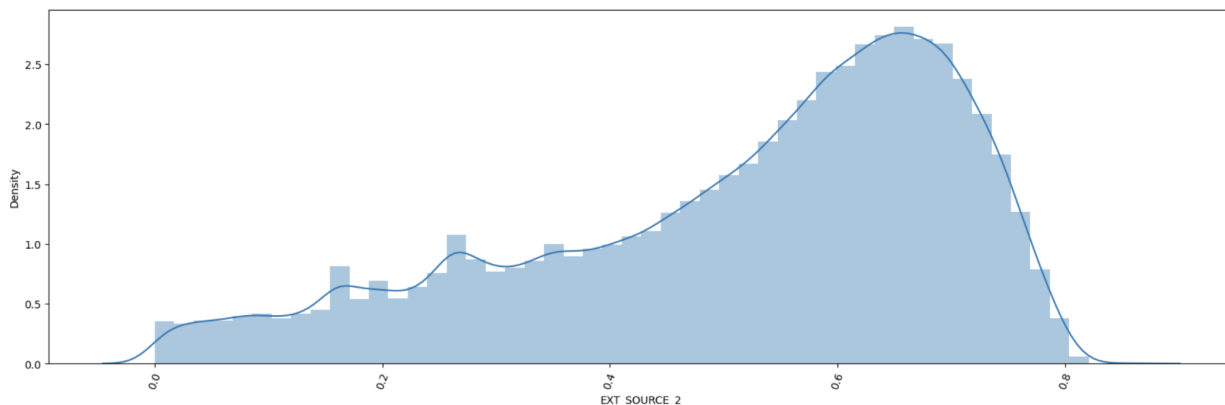


Рисунок 6.1 – Распределение признака EXT_SOURCE_2

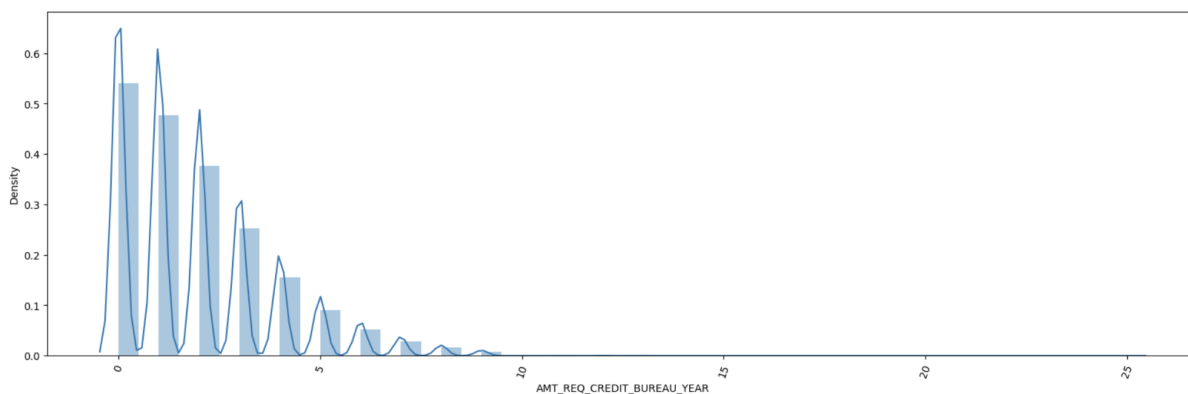


Рисунок 6.2 – Распределение признака AMT_REQ_CREDIT_BUREAU

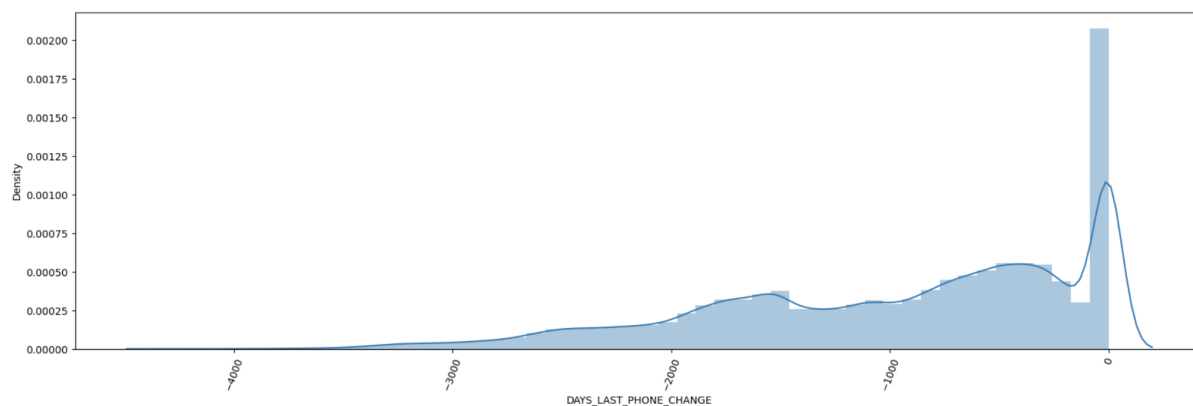


Рисунок 6.3 – Распределение признака DAYS_LAST_PHONE_CHANGE

Как видим на Рисунках 6.1, 6.2, 6.3, некоторые данные имеют сильно смещение вправо, некоторые – влево. Данные распределены ненормально и имеют выбросы.

3. Выявление корреляции между данными.

Применим метод Пирсона для вычисления корреляции между признаками. Метод Пирсона определяет, насколько два признака связаны линейно и в какой степени они изменяются вместе. Коэффициент корреляции Пирсона находится в диапазоне от -1 до 1. Сила корреляции зависит от абсолютного значения коэффициента корреляции.

Обычно принимают следующие значения корреляции:

- 0.00–0.19 – очень слабая корреляция;
- 0.20–0.39 – слабая корреляция;
- 0.40–0.59 – умеренная корреляция;
- 0.60–0.79 – сильная корреляция;
- 0.80–1.00 – очень сильная корреляция.

Отрицательные значения корреляции означают обратную зависимость между признаками: если значение одного признака растет, то значение другого признака падает. Положительные значения корреляции означают прямую зависимость между признаками: если значение одного признака растет, то значение другого признака тоже растет. Значения корреляции близкие к нулю означают отсутствие линейной зависимости между признаками.

```
Наивысшая позитивная корреляция: DEF_60_CNT_SOCIAL_CIRCLE    0.031276
DEF_30_CNT_SOCIAL_CIRCLE    0.032248
LIVE_CITY_NOT_WORK_CITY    0.032518
OWN_CAR_AGE    0.037612
DAYS_REGISTRATION    0.041975
FLAG_DOCUMENT_3    0.044346
REG_CITY_NOT_LIVE_CITY    0.044395
FLAG_EMP_PHONE    0.045982
REG_CITY_NOT_WORK_CITY    0.050994
DAYS_ID_PUBLISH    0.051457
DAYS_LAST_PHONE_CHANGE    0.055218
REGION_RATING_CLIENT    0.058899
REGION_RATING_CLIENT_W_CITY    0.060893
DAYS_BIRTH    0.078239
TARGET    1.000000
Name: TARGET, dtype: float64
Наивысшая негативная корреляция: EXT_SOURCE_3    -0.178919
EXT_SOURCE_2    -0.160472
EXT_SOURCE_1    -0.155317
DAYS_EMPLOYED    -0.044932
FLOORSMAX_AVG    -0.044003
FLOORSMAX_MEDI    -0.043768
FLOORSMAX_MODE    -0.043226
AMT_GOODS_PRICE    -0.039645
REGION_POPULATION_RELATIVE    -0.037227
ELEVATORS_AVG    -0.034199
ELEVATORS_MEDI    -0.033863
FLOORSMIN_AVG    -0.033614
FLOORSMIN_MEDI    -0.033394
LIVINGAREA_AVG    -0.032997
LIVINGAREA_MEDI    -0.032739
Name: TARGET, dtype: float64
```

Рисунок 7 – Корреляция между признаками по отношению к таргету

Анализ показал, что большинство переменных в датасете имеют слабую корреляцию с таргетом (за исключением самого таргета, который коррелирует с самим собой). Однако, в данных есть два признака, которые имеют достаточно высокую корреляцию с таргетом. Это возраст и «внешние источники данных». Предположительно, это данные, полученные от других кредитных организаций. Интересно, что несмотря на заявленную независимость цели от таких данных, на практике они играют важную роль в принятии кредитного решения. Анализ признаков, имеющих более сильную корреляцию с целевой переменной:

1. Возраст

Признак `DAYS_BIRTH` указан в отрицательных значениях, представляющих количество дней до выдачи кредита от даты рассмотрения заявки. Большой возраст клиента связан со стабильностью и уверенностью в финансовых делах, что в свою очередь повышает вероятность возврата кредита (до определенного предела, конечно). Однако из-за отрицательных значений, признак `DAYS_BIRTH` положительно коррелирует с невозвратом (что выглядит несколько странно). Для анализа зависимости между возрастом и невозвратом кредита, приведем `DAYS_BIRTH` к положительному значению путем взятия его по модулю, и затем посмотрим на корреляцию между новым признаком и целевой переменной.

Для более полного анализа распределения возраста клиентов и его влияния на результат, можно использовать график *kernel density estimation* (KDE) – распределение ядерной плотности. Этот график показывает распределение одной переменной и используется как сглаженная гистограмма.

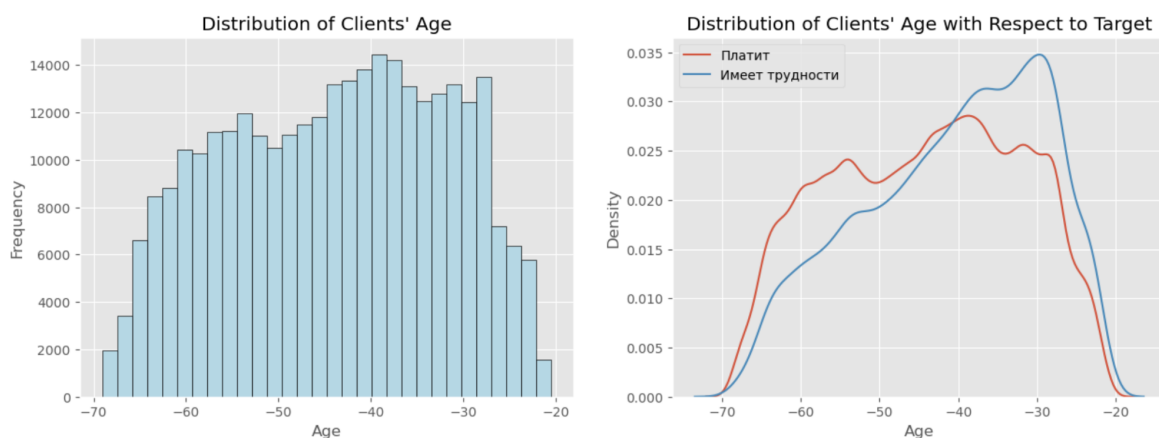


Рисунок 8 – Распределение клиентов по возрасту

Анализ показал, что доля невозвратов кредита выше среди молодых людей и снижается по мере увеличения возраста. Но это не означает, что всегда нужно отказывать молодым людям в кредите, такая тактика может привести к упущению возможностей для банка. Однако, это становится поводом для внимательного мониторинга таких кредитов, оценки рисков и, возможно, предоставления финансового образования для молодых заемщиков.

2. Внешние источники

Посмотрим внимательнее на корреляцию таргета с внешними источниками EXT_SOURCE.

	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 9 – Корреляция внешних источников по отношению к таргету

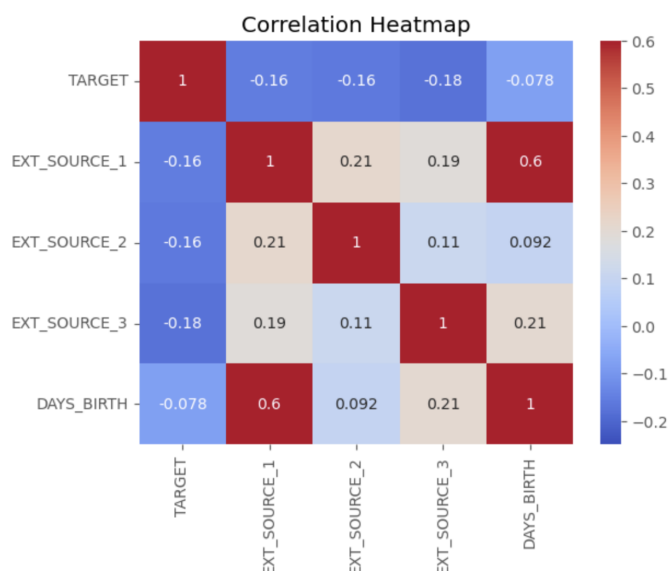


Рисунок 10 – Матрица корреляция между признаками

Все внешние источники показывают негативную корреляцию с таргетом.

Посмотрим KDE в разрезе каждого источника.

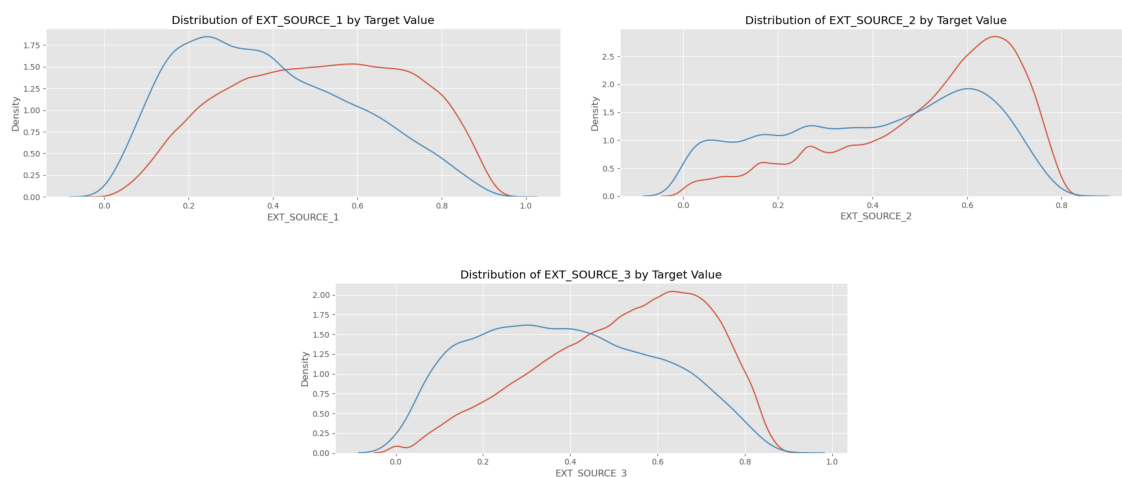


Рисунок 11 – Распределение ядерной плотности внешних источников по отношению к таргету

Сравнивая коэффициенты корреляции, можно заметить, что наиболее сильная корреляция наблюдается между целевой переменной и признаком EXT_SOURCE_3. Это может свидетельствовать о том, что данный признак может иметь наибольшее влияние на модель предсказания невозврата кредита. Кроме того, значения коэффициентов корреляции между EXT_SOURCE_1, EXT_SOURCE_2 и целевой переменной также довольно высоки. Поэтому при построении модели стоит уделить особое внимание этим признакам и,

возможно, использовать их в качестве ключевых факторов для прогнозирования вероятности возврата кредита.

3. Прочие признаки.

3.1 Тип займа и доля проблемных займов в зависимости от его типа (NAME_CONTRACT_TYPE).

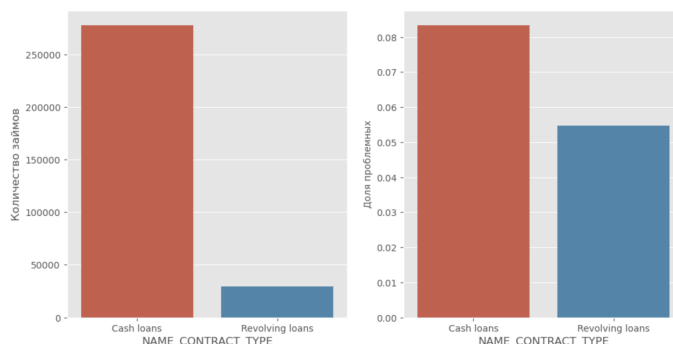


Рисунок 12 – Сравнение количества займов и доли проблемных клиентов по типу займа

Возобновляемые кредиты составляют менее 10% от всех займов, но имеют намного больше проблем с возвратом.

3.2 Зависимость проблемы с выплатой кредита и полом клиента (CODE_GENDER).

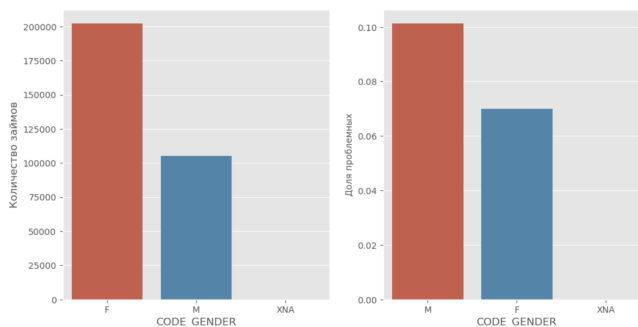


Рисунок 13 – Сравнение количества займов и доли проблемных клиентов по полу клиента

Женщины берут кредит в 2 раза чаще мужчин, при этом мужчины имеют больше проблем с выплатами, чем женщины.

3.3 Выявление закономерностей между владением активами и проблемами с выплатами. Рассмотрим 2 признака:

- FLAG_OWN_REALTY – владение недвижимостью

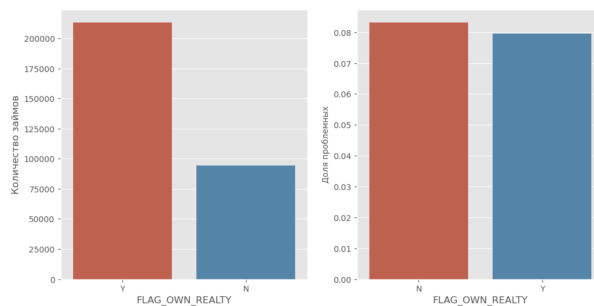


Рисунок 14 – Сравнение количества займов и доли проблемных клиентов по владению недвижимостью

Клиентов, не имеющих недвижимости, вдвое меньше. Владельцы же недвижимости имеют немного меньший риск по невозврату кредита.

- FLAG_OWN_CAR – владение машиной

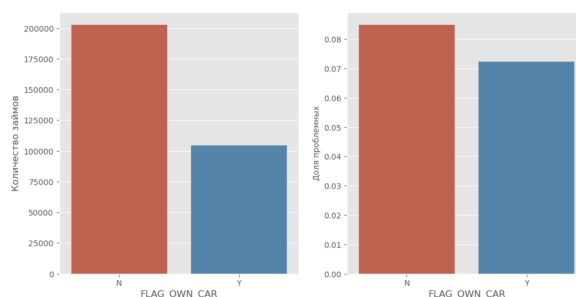


Рисунок 15 – Сравнение количества займов и доли проблемных клиентов по владению машиной

Людей с кредитами, владеющих машиной, в два раза меньше, но риск невыплаты кредита практически одинаковый. Отсюда можно сделать вывод о том, что люди, владеющие машиной, чаще выплачивают кредиты.

3.4 Семейное положение и проблемы с оплатой кредита (NAME_FAMILY_STATUS).

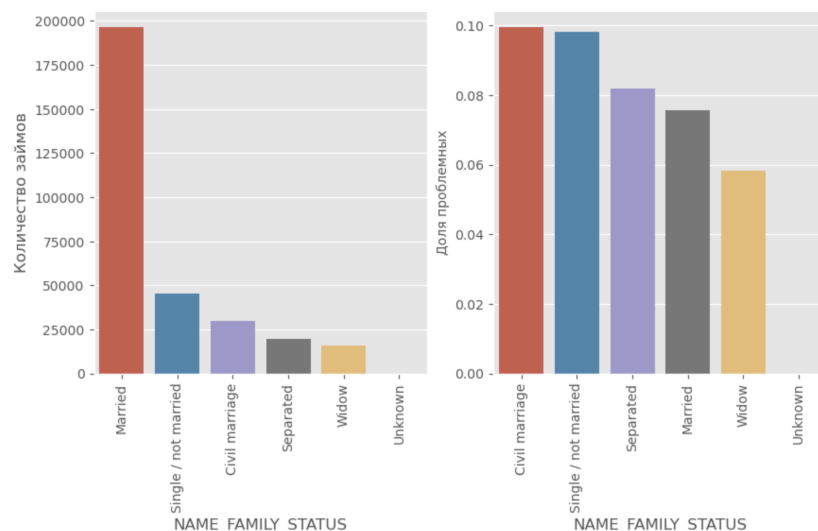


Рисунок 16 – Сравнение количества займов и доли проблемных клиентов по семейному положению

Большинство клиентов состоит в браке, но наиболее рискованные группы клиентов – это те, кто находится в гражданском браке или одинокие. Вдовцы же показывают наименьший риск.

3.5 Дети и возникновение проблем с выплатой по кредиту (CNT_CHILDREN).

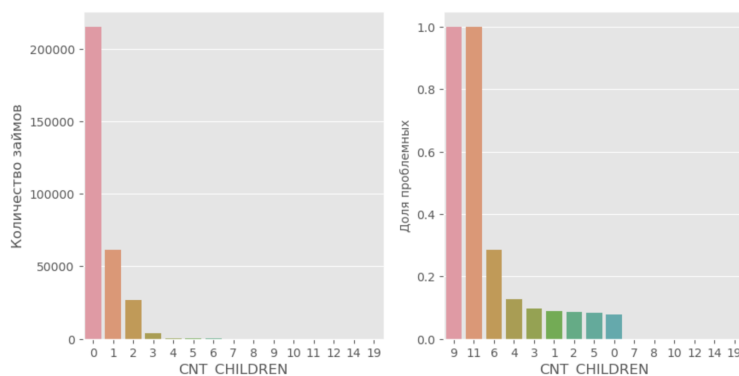


Рисунок 17 – Сравнение количества займов и доли проблемных клиентов по количеству детей в семье

Клиенты с 9 и 11 детьми показывают полную невозвратность кредита, однако эти данные можно считать статистически незначимыми.

3.6 Проблемы с выплатами в зависимости от типа дохода (NAME_INCOME_TYPE).

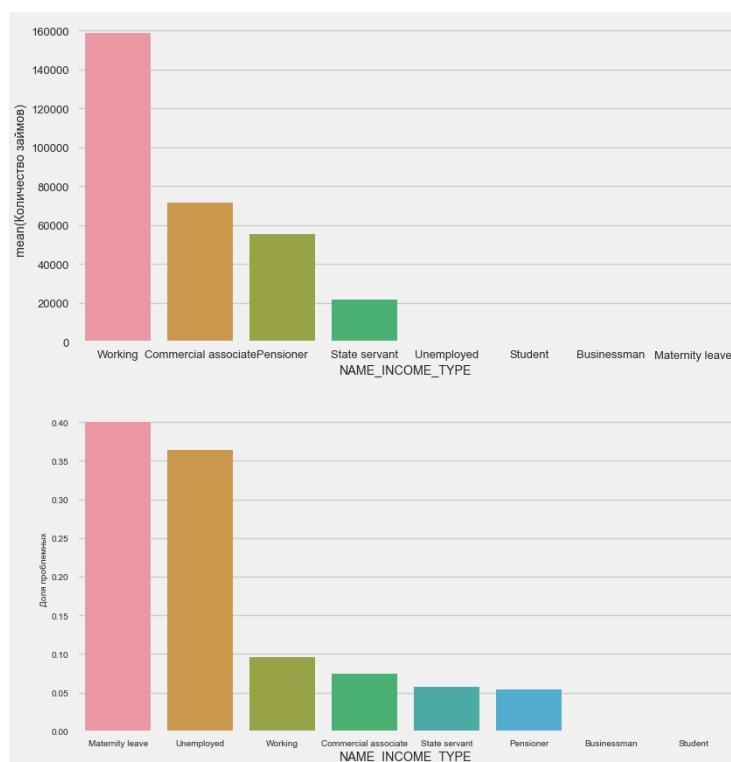


Рисунок 18 – Сравнение количества займов и доли проблемных клиентов по типу дохода

Судя по количеству клиентов среди безработных и матерей-одиночек, им редко выдают кредиты, так как они показывают почти полный невозврат денежных средств.

3.7 Род занятий и проблемы с выплатой по кредиту (OCCUPATION_TYPE).

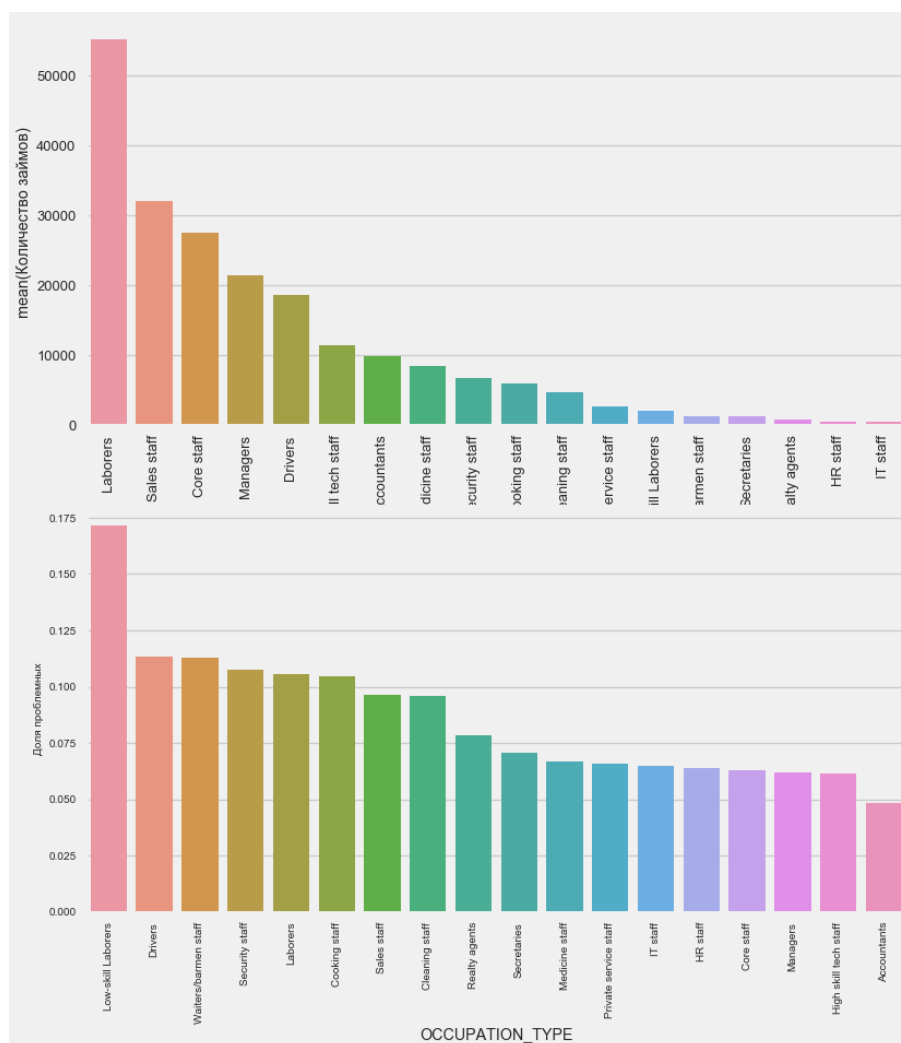


Рисунок 19 – Сравнение количества займов и доли проблемных клиентов по роду занятий

Не берем в расчет рабочих (Laborers) – их больше всего. Наибольший интерес вызывают водители, официанты/бармены и работники сферы безопасности. Можно считать эти профессии проблемными с точки зрения кредитной организации.

3.8 Зависимость проблем с кредитной историей и образования (NAME_EDUCATION_TYPE).

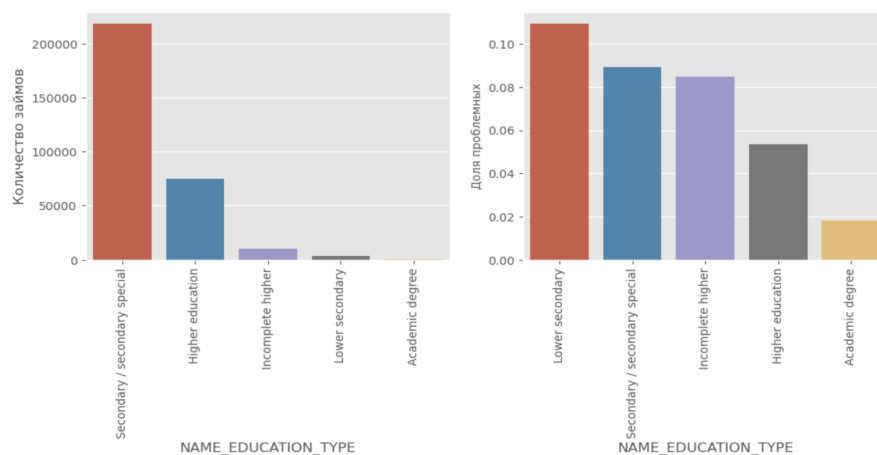


Рисунок 20 – Сравнение количества займов и доли проблемных клиентов по образованию

Уровень образованию позитивно влияет на возможности по возврату заемных средств кредитной организации.

3.9 Размер суммы кредита и его влияние на выплаты (AMT_CREDIT).

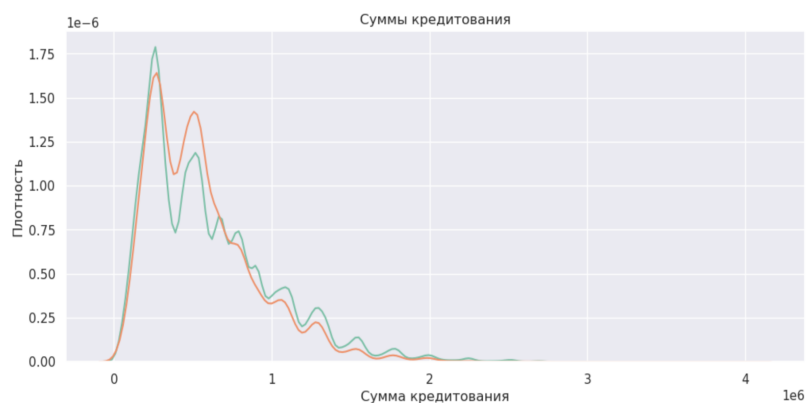


Рисунок 21 – Сравнение количества займов и доли проблемных клиентов в зависимости от суммы кредита

На графике плотности распределения можно отметить, что крупные суммы клиенты чаще возвращают.

4. Результаты и интерпретация

4.1. Предобработка данных

Для построения модели прогнозирования дефолта заемщика на основе данных кредитного бюро «Home Credit Default Risk» следовало изначально провести предобработку данных.

Предобработка данных – это процесс очистки и подготовки данных для анализа и построения моделей машинного обучения. В данном проекте была проведена предобработка данных, чтобы убрать выбросы и заполнить пропущенные значения. В процессе предобработки использовались как числовые, так и текстовые столбцы.

1. Применение стратегии замены.

Для числовых переменных, в данном случае, был применен подход замены пропущенных значений на медиану. Это связано с тем, что медиана является устойчивой статистикой и не подвержена влиянию выбросов. Процесс замены был реализован с помощью метода `fillna()` из библиотеки `Pandas`. Таким образом, каждое пустое значение числовой переменной было заменено на соответствующую медиану этого столбца.

Для категориальных переменных, была использована стратегия замены пропущенных значений на наиболее часто встречающееся значение (`mode`). Этот подход основывается на предположении, что пропущенные значения могут быть случайными или утерянными и замена их на наиболее часто встречающееся значение не исказит распределение переменной. Также использовался метод `fillna()` для замены пропущенных значений на рассчитанное модальное значение для каждой категориальной переменной.

Однако, для некоторых категориальных признаков, где пропущенные значения не могли быть просто заменены на моду, были применены специфические стратегии замены. Например, для признака «`occupation type`» был проведен более сложный процесс заполнения пропущенных значений. В этом случае, была проведена анализ данных других переменных, таких как «`income`» и «`education`», чтобы определить наиболее вероятное значение для пропущенных записей.

Таким образом, специфические замены категориальных признаков были проведены с учетом особенностей каждого признака и с использованием подходящих стратегий, чтобы минимизировать потерю информации и обеспечить более полные данные для проведения анализа кредитных рисков.

2. Нормализация данных.

Дополнительно, перед обработкой пропущенных значений, были применены методы нормализации числовых переменных и кодирование категориальных переменных. Нормализация числовых переменных позволяет привести их к общему масштабу и избежать искажений при анализе и моделировании. Кодирование категориальных переменных позволяет преобразовать их в числовой формат, что облегчает их использование в моделях машинного обучения.

Были проведены работы по преобразованию категориальных данных, которые не могут быть прямо использованы в модели машинного обучения. Для этого был использован метод One-Hot Encoding, который позволил преобразовать категориальные признаки в бинарные, которые можно использовать в модели. Каждый уникальный категориальный признак был преобразован в отдельный столбец, где значение 1 указывает, что данный признак присутствует, а значение 0 – что он отсутствует. Это позволило модели работать с категориальными данными без необходимости задания порядка между ними.

Кроме того, была использована техника порядкового кодирования (Ordinal Encoding), которая позволяет присвоить числовые значения категориальным данным на основе их порядка. Для этого каждому уникальному значению категориального признака был присвоен свой уникальный числовой код. Этот метод применяется в случаях, когда порядок между значениями категориального признака имеет значение, например, при кодировании уровня образования или уровня зарплаты.

Преобразование категориальных данных было проведено для того, чтобы модель машинного обучения могла использовать все доступные данные для прогнозирования. При этом не было утрачено никакой информации о категориальных признаках, а они были преобразованы в формат, который может быть использован в модели.

Для обработки выбросов в числовых переменных, был использован метод межквартильного размаха (IQR) [29]. Этот метод позволяет определить границы выбросов на основе распределения значений в переменной. Первоначально были вычислены первый и третий квартили (q_1 и q_3) для каждой числовой переменной. Затем был рассчитан межквартильный размах (IQR) как разница между q_3 и q_1 . На основе IQR были определены верхняя и нижняя границы выбросов.

Значения, находящиеся за пределами этих границ, были заменены на граничные значения, чтобы минимизировать их влияние на анализ. Это было достигнуто с использованием функции `clip()` из библиотеки Pandas. Для каждой переменной были определены соответствующие граничные значения, и значения вне этих границ были заменены на граничные значения.

В ходе дипломной работы также была проведена обработка данных, которая включала в себя масштабирование числовых признаков с помощью `StandardScaler`. Для этого был использован модуль `preprocessing` библиотеки `sklearn`. Данный модуль позволяет нормализовать данные путем вычитания среднего значения и деления на стандартное отклонение каждого признака. Такой подход позволяет уравнивать дисперсии всех признаков и сделать их более сопоставимыми. Более того, такое масштабирование уменьшает влияние выбросов на данные, что может повысить качество работы алгоритмов машинного обучения.

Во время работы с данными была обнаружена проблема несбалансированных классов в задаче бинарной классификации. В данных было намного больше объектов с положительным классом (клиенты, вернувшие кредит вовремя) по сравнению с отрицательным классом (клиенты, не вернувшие кредит вовремя). Так как это может привести к некорректной работе алгоритмов машинного обучения, было принято решение использовать параметр `class_weight` для обучения модели.

`Class Weight` – это параметр, позволяющий задать веса классов в модели. Он позволяет учесть дисбаланс классов и снизить его влияние на качество

обучения. В данном проекте параметр `class_weight` был установлен в значение «balanced». Это значение автоматически вычисляет веса классов, основываясь на их соотношении в данных, и присваивает больший вес редкому классу. Таким образом, при обучении модели модель будет уделять большее внимание редкому классу и делать более точные предсказания для него.

4.2. Создание новых признаков

В ходе анализа данных были добавлены новые признаки:

- **LTV (Loan to Value)** – это отношение запрошенной суммы кредита к стоимости приобретаемого имущества. Данный признак может помочь в оценке риска выдачи кредита, т.к. кредиты с более высоким LTV считаются более рискованными.
- **DTI (Debt to Income ratio)** – это отношение ежемесячного платежа по кредиту к ежемесячному доходу заемщика. Этот признак также помогает в оценке риска выдачи кредита, т.к. заемщики с более высоким DTI могут иметь большие трудности с погашением кредита.
- **Employed_to_Birth** – это отношение количества дней работы к заемщиком к его возрасту. Данный признак может помочь в оценке финансовой стабильности заемщика, т.к. люди с более высоким отношением могут считаться более финансово устойчивыми.
- **Older_30** – это бинарный признак, который указывает, превышает ли возраст заемщика 30 лет (1 – если превышает, 0 – если нет). Этот признак может помочь в оценке риска, т.к. молодые заемщики могут иметь меньшую финансовую стабильность.
- **Employed_5** – это бинарный признак, который указывает, превышает ли количество дней работы заемщика 5 лет (1 – если превышает, 0 – если нет). Этот признак также помогает в оценке финансовой стабильности заемщика.

Анализ новых признаков относительно исходных переменных позволяет оценить их информативность, взаимосвязь и возможное наличие

коллинеарности. Проанализировать взаимосвязь между новыми признаками и исходными переменными, на основе которых они были созданы можно с использованием методов статистического анализа, таких как корреляционная матрица.

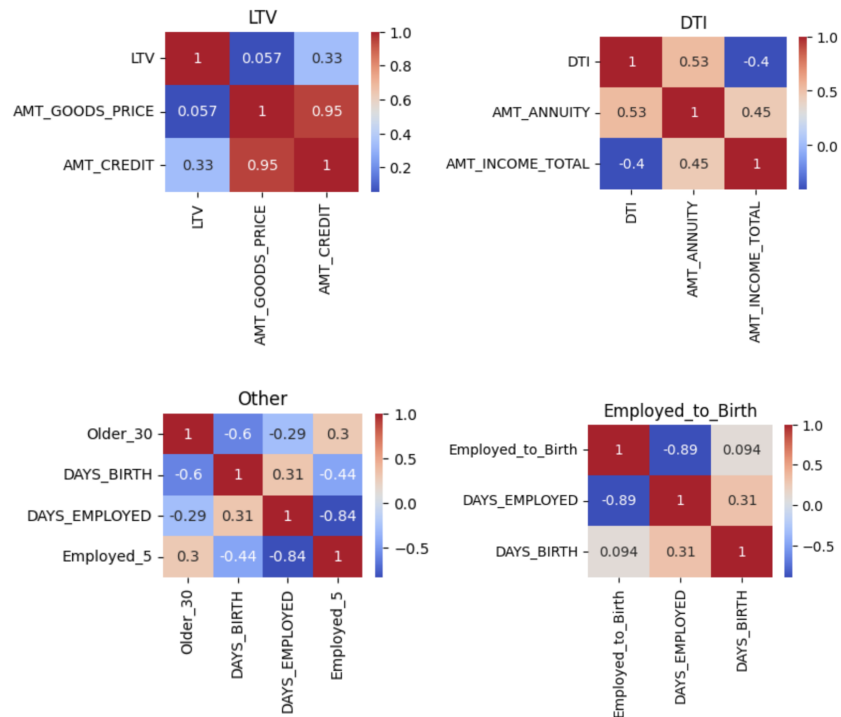


Рисунок 22 – Матрицы корреляции для новых признаков

Анализируя графики и рассмотрев коэффициенты корреляции, можно сделать выводы относительно коллинеарности между новыми созданными признаками и основными признаками, на основе которых они были получены.

1. Признаки LTV, DTI и Older_30, созданные на основе существующих переменных, не имеют сильной положительной или отрицательной корреляции с исходными признаками. Это означает, что новые признаки содержат информацию, которая не является прямой функцией исходных признаков и могут вносить дополнительную информацию в модель.

2. Признаки Employed_to_Birth и Employed_5 имеют сильную отрицательную корреляцию с DAYS_EMPLOYED. Это может указывать на наличие обратной зависимости между длительностью занятости (в днях) и новыми созданными признаками. Например, с увеличением длительности занятости, значения признаков Employed_to_Birth и Employed_5 снижаются.

Это может быть интересным выводом, который подтверждает, что длительность занятости может оказывать влияние на кредитный скоринг.

3. Важно отметить, что коэффициенты корреляции не измеряют все возможные взаимосвязи и зависимости между признаками. Они могут учитывать только линейные связи и не улавливать нелинейные зависимости или взаимосвязи. Поэтому, помимо анализа корреляции, также стоит рассмотреть применение алгоритмов отбора признаков, чтобы более полно оценить важность и информативность новых признаков.

В ходе исследования был применен метод генетического программирования (ГП) с использованием библиотеки DEAP (Distributed Evolutionary Algorithms in Python) для создания новых признаков. ГП является эволюционным алгоритмом, где программы представлены в виде деревьев, а генетические операторы применяются для эволюции этих деревьев с целью создания новых признаков.

В экспериментах была создана популяция программ-деревьев, которая подвергалась эволюции с помощью операторов мутации и скрещивания. Каждое программное дерево представляет собой выражение, которое использует исходные признаки для создания новых признаков. Библиотека DEAP обеспечивает удобный набор инструментов для реализации и выполнения этих операций ГП.

В результате экспериментов были созданы новые признаки, которые могут внести значительный вклад в обучение модели оценки кредитного риска. Эти признаки могут быть представлены в различных формах и могут учитывать различные аспекты данных и предметной области. Их добавление к исходному набору признаков может значительно улучшить производительность модели и повысить точность прогнозирования.

Использование генетического программирования и библиотеки DEAP предоставляет исследователям и практикам возможность автоматического создания новых признаков на основе исходных данных. Это позволяет значительно расширить пространство признаков и улучшить способность

модели к выявлению скрытых закономерностей и зависимостей в данных. Такой подход может быть полезным в области оценки кредитного риска, где точность моделей играет важную роль в принятии решений.

В результате экспериментов было создано несколько новых признаков, которые могут внести важный вклад в обучение модели. Некоторые из них могут быть представлены следующим образом:

- Feature_1: для создания данного признака использовались исходные признаки «AMT_INCOME_TOTAL» и «AMT_CREDIT». Операцией, примененной в генетическом программировании, было вычисление отношения между признаками «AMT_INCOME_TOTAL» и «AMT_CREDIT». Полученный признак был оценен с фитнесом 0,83, указывающим на его значимый вклад в обучение модели.

- Feature_2: для создания данного признака использовались исходные признаки «NAME_EDUCATION_TYPE» и «DAYS_EMPLOYED». Операцией, примененной в генетическом программировании, было умножение значения признака «DAYS_EMPLOYED» на число, связанное с образованием клиента. Полученный признак был оценен с фитнесом, указывающим на его значимость для модели 0,74.

- Feature_3: для создания данного признака использовались исходные признаки «REGION_POPULATION_RELATIVE» и «AMT_GOODS_PRICE». Операцией, примененной в генетическом программировании, было вычисление произведения значений данных признаков. Полученный признак был оценен с фитнесом 0,78.

Для того, чтобы оценить вклад созданных признаков в обучение модели, перейдем к следующему разделу дипломной работы.

4.3. Построение модели

В ходе выполнения дипломной работы были рассмотрены и протестированы две популярные модели машинного обучения: логистическая регрессия и случайный лес. Выбор этих моделей обусловлен их широким

применением в задачах классификации, а также их отличительными особенностями и преимуществами.

Логистическая регрессия является одним из основных методов бинарной классификации. Она основана на логистической функции, которая позволяет оценивать вероятность принадлежности объекта к определенному классу. Логистическая регрессия имеет простую интерпретацию результатов и хорошую способность обобщения на новые данные. Она также устойчива к выбросам и может быть эффективно применена к наборам данных с большим числом признаков.

Случайный лес является ансамблевым методом машинного обучения, который комбинирует множество деревьев решений для принятия окончательного решения. Он позволяет справляться с проблемой переобучения и имеет хорошую способность обобщения на новые данные. Случайный лес обладает высокой степенью гибкости и способен автоматически обнаруживать нелинейные зависимости в данных. Он также устойчив к выбросам и шуму в данных.

Выбор данных моделей для исследования в дипломной работе обусловлен их широким применением и успешным использованием в различных сферах, а также их способностью обработки больших объемов данных. Эти модели являются базовыми и хорошо изученными методами, что обеспечивает их надежность и сравнительную простоту в реализации и интерпретации результатов.

В дальнейшем проведении экспериментов и анализе результатов будет рассмотрена производительность и точность каждой из этих моделей, что позволит сделать выводы о их пригодности для решения задачи классификации в контексте данного проекта.

1. Логистическая регрессия.

В рамках данной дипломной работы была реализована модель логистической регрессии для решения задачи классификации. Для начала процесса моделирования было выполнено выделение целевой переменной и

признаков из исходного датасета. Затем данные были предобработаны путем преобразования категориальных признаков с помощью метода One-Hot Encoding.

Для обучения и оценки модели были разделены данные на обучающую и тестовую выборки с помощью функции `train_test_split` из модуля `sklearn.model_selection`. Далее, чтобы обеспечить стабильность и сходимость модели, было выполнено масштабирование признаков с использованием стандартного скалирования методом `StandardScaler` из модуля `sklearn.preprocessing`. Это позволяет привести признаки к одному масштабу и избежать проблем с большими значениями, влияющих на обучение модели.

Для решения проблемы несбалансированности классов была применена балансировка с использованием методов `RandomOverSampler` и `RandomUnderSampler` из модуля `imblearn.over_sampling` и `imblearn.under_sampling` соответственно. Эти методы позволяют увеличить или уменьшить преобладающий класс, чтобы достичь баланса между классами.

Создание и обучение модели логистической регрессии было выполнено с использованием класса `LogisticRegression` из модуля `sklearn.linear_model`. В данном случае, модель логистической регрессии была инициализирована с параметрами `random_state=42` для обеспечения воспроизводимости результатов и `max_iter=1000` для установки максимального числа итераций в процессе обучения модели.

Оценка качества модели была осуществлена путем предсказания меток классов для тестовой выборки с помощью метода `predict` и расчета точности модели с использованием метода `score`.

В процессе экспериментов было проведено два запуска модели логистической регрессии. В первом запуске модель обучалась на исходных признаках без использования созданных признаков, и ее точность составила 0,68. Во втором запуске модель была обучена с использованием новых созданных признаков, и точность повысилась до 0,84.

Анализ важности признаков показал, что созданные признаки внесли значительный вклад в повышение точности модели. Некоторые из них попали в топ-5 наиболее значимых признаков. Это говорит о том, что эти новые признаки содержат полезную информацию, которая помогла модели лучше предсказывать целевую переменную.

```
feature_importance = pd.Series(lr.coef_[0], index=X.columns)
feature_importance = feature_importance.abs().sort_values(ascending=False)
print(feature_importance)
```

AMT_CREDIT	0.511365
EXT_SOURCE_3	0.470359
AMT_GOODS_PRICE	0.442327
EXT_SOURCE_2	0.413383
LTV	0.250359

Рисунок 23 – Результат выбора лучших признаков моделью Логистической регрессии

Важно отметить, что модель логистической регрессии обладает высокой производительностью и может быть использована в реальном времени, например, при подаче личной заявки в банк, когда клиент ожидает решения, или при онлайн-заявках, где требуется быстрый анализ и принятие решений. Благодаря оптимизации алгоритма и масштабированию признаков, модель может обрабатывать данные быстро и эффективно, что является значимым преимуществом в современных условиях, где скорость и точность прогнозирования играют важную роль.

Таким образом, использование созданных признаков существенно повысило точность модели логистической регрессии. Модель проявила высокую производительность, что делает ее привлекательным инструментом для применения в различных банковских сценариях. Благодаря простоте и эффективности данной модели, она может быть использована в реальных ситуациях, например, при личной подаче заявки в банке, когда клиент ожидает решения, или при онлайн заявках, где требуется быстрый анализ и прогнозирование дефолта по кредиту.

2. Random Forest.

В качестве альтернативной модели логистической регрессии была выбрана модель Random Forest, которая позволяет предсказывать вероятность

дефолта заемщика. Для этого была использована библиотека `scikit-learn`, которая предоставляет реализацию данной модели.

Основной идеей Random Forest является создание ансамбля из множества решающих деревьев, каждое из которых строится на случайном наборе признаков и случайной выборке объектов обучающей выборки. Таким образом, каждое дерево строится независимо от других и не зависит от признаков, использованных в других деревьях. В результате, композиция деревьев может дать лучшие результаты, чем отдельное дерево, за счет уменьшения дисперсии предсказаний [17, С.83].

Преимущества модели Random Forest:

- способность работать с разными типами признаков и атрибутов
- способность эффективно работать с выбросами и отсутствующими данными
- хорошая способность к обобщению (то есть способность к обработке новых данных, которые не были использованы в обучении)
- относительно простая настройка параметров

Обучение модели начинается с загрузки необходимых библиотек и данных. Данные предварительно обрабатываются и разбиваются на тренировочную и тестовую выборки. Затем создается объект модели случайного леса с определенными гиперпараметрами. В ходе выполнения проекта гиперпараметры были выбраны опытным путем, путем экспериментирования с различными значениями.

Обучение модели началось с тренировочной выборки с использованием метода `fit()`. Обучение проводилось несколько раз с разными значениями гиперпараметров, что позволяет оценить их влияние на качество модели. Для оценки качества используется метрика AUC-ROC, которая показывает, насколько хорошо модель разделяет классы «вернул кредит» и «не вернул кредит».

В данной работе было проведено несколько запусков модели с различными значениями гиперпараметров. В машинном обучении

гиперпараметры – это настраиваемые параметры, которые не могут быть определены самой моделью, но которые определяют ее структуру и процесс обучения. Настройка гиперпараметров является важной частью процесса обучения и может оказать значительное влияние на качество модели. В данном проекте гиперпараметры модели Random Forest были настроены при помощи Grid Search. Это метод, который позволяет перебрать все возможные комбинации значений гиперпараметров из заранее заданного диапазона и выбрать лучшую комбинацию на основе заданной метрики качества.

В первом запуске модели были выбраны следующие гиперпараметры:

- `n_estimators` – число деревьев в лесу (от 70 до 110)
- `max_depth` – максимальная глубина деревьев в лесу. В первом запуске она была установлена на 15.
- `random_state` – параметр, определяющий случайное начальное состояние генератора случайных чисел. В первом запуске были использованы различные значения этого параметра, чтобы обеспечить различные случайные начальные состояния.

Цель опробовать различные значения гиперпараметров заключается в том, чтобы определить наилучшие значения, которые обеспечат лучшее качество модели на данном наборе данных. При каждом запуске модели с различными значениями гиперпараметров проводится оценка качества модели с помощью метрики AUC-ROC на тестовой выборке, чтобы определить, какие значения гиперпараметров дают лучший результат.

В первом запуске были опробованы пять различных комбинаций гиперпараметров, и наилучшим результатом была достигнута модель с 100 деревьями в лесу и максимальной глубиной деревьев, равной 15. В результате была получена точность 0.73, что является довольно неплохим результатом.

Во втором запуске был использован метод GridSearchCV для подбора наилучших значений гиперпараметров модели. Для этого был задан словарь `param_grid`, в котором были перечислены параметры, которые должны были быть оптимизированы.

Словарь `param_grid` имеет следующие параметры:

- `criterion` – критерий, используемый для измерения качества разделения. В данном случае был использован критерий Джини (`gini`).
- `class_weight` – параметр, определяющий вес каждого класса. Возможными значениями являются `'balanced_subsample'`, `'balanced'` и `None`.
- `max_features` – максимальное количество признаков, используемых при построении каждого дерева. В данном случае были использованы значения `'sqrt'` и `'log2'`, означающие корень и логарифм от количества признаков соответственно.

Критерий Джини (`Gini impurity`) – это мера неопределенности в дереве решений, которая измеряет вероятность неправильной классификации случайно выбранного элемента из набора данных, если он был классифицирован случайным образом согласно распределению меток в данном узле. Критерий Джини определяет, насколько хорошо данный признак разделяет выборку на классы. В данной работе критерий Джини был использован в параметре `criterion` при поиске наилучших гиперпараметров с помощью `RandomizedSearchCV`. Это позволяет выбирать наилучшие разделения, учитывая неопределенность в данных.

Параметр `class_weight` в модели случайного леса отвечает за учет дисбаланса классов в данных путем присвоения весов классам. В общем случае, если в выборке имеется дисбаланс классов, то модель будет склонна к предсказанию чаще встречающегося класса, игнорируя редкий класс.

В данном случае, были протестированы три значения параметра `class_weight`: `balanced_subsample`, `balanced` и `None`:

- `balanced_subsample` означает, что при построении каждого дерева случайного леса будет использоваться подвыборка обучающих данных, в которой классы будут сбалансированы.

- `balanced` означает, что веса классов будут присвоены обратно пропорционально частотам их появления в выборке, так что доля ошибок для каждого класса будет приблизительно равна.
- `None` означает, что веса классов не учитываются.

Далее был создан объект модели `RandomForestClassifier` с начальными значениями гиперпараметров: `n_estimators = 100`, `max_depth = 15` и `random_state = 42`. Для подбора наилучших значений гиперпараметров был использован метод `RandomizedSearchCV`. Он позволяет задать набор значений гиперпараметров и выбрать наилучшие из них на основе кросс-валидации. Кроме того, были заданы параметры `cv = 5`, указывающие на использование 5-кратной кросс-валидации, и `scoring = 'roc_auc'`, указывающий на использование метрики `roc_auc` для оценки качества модели. Лучшими параметрами для модели оказались:

```
grid_cv.best_params_
{'max_features': 'sqrt', 'criterion': 'gini', 'class_weight': None}
```

Рисунок 24 – Лучшие параметры модели Random Forest

В результате второго запуска модели была получена $AUC-ROC = 0.72$, что немного ниже, чем в первом запуске, но не сильно отличается от неё.

Третий запуск модели осуществлялся после применения метода `RandomizedSearchCV` для поиска наилучших значений гиперпараметров. В результате этого поиска были получены значения гиперпараметров, наилучшим образом подходящие для данного набора данных: `max_features='sqrt'`, `criterion='gini'`, `class_weight=None`.

Затем были созданы пять экземпляров класса `RandomForestClassifier` с разными значениями числа деревьев (`n_estimators`) и параметром `max_depth=15` для ограничения глубины каждого дерева и предотвращения переобучения модели. Все пять экземпляров были созданы с использованием наилучших значений гиперпараметров, полученных на предыдущем шаге.

В результате подбора гиперпараметров было достигнуто значение метрики AUC-ROC равное 0.74. Это означает, что модель хорошо справляется с задачей разделения двух классов, и может быть использована для предсказания кредитной надежности новых заемщиков.

Для оценки качества моделей использовалась метрика AUC-ROC. Эта метрика измеряет качество бинарного классификатора, который использует пороговое значение для разделения двух классов. В нашем случае, AUC-ROC показывает, насколько хорошо модель способна различать клиентов, вернувших кредит, от тех, кто не вернул его. Результаты представлены в виде числа от 0 до 1, где 0 означает, что модель не различает классы, а 1 означает, что модель идеально различает классы.

Что касается признаков, которые внесли наибольший вклад в обучение модели, то с применением метода `feature_importances_` получились следующие данные:

```
importances = rf.feature_importances_
indices = np.argsort(importances)[::-1]

print("Top 5 Features:")
for i in range(10):
    print(f"{i+1}. Feature: {X.columns[indices[i]]}, Importance: {importances[indices[i]]:.4f}")
```

Top 5 Features:

1. Feature: EXT_SOURCE_2, Importance: 0.1005
2. Feature: EXT_SOURCE_3, Importance: 0.0874
3. Feature: DAYS_BIRTH, Importance: 0.0407
4. Feature: DAYS_ID_PUBLISH, Importance: 0.0399
5. Feature: DAYS_REGISTRATION, Importance: 0.0393
6. Feature: DTI, Importance: 0.0381
7. Feature: Employed_to_Birth, Importance: 0.0379
8. Feature: DAYS_EMPLOYED, Importance: 0.0370
9. Feature: Credit_Per_Days, Importance: 0.0369
10. Feature: YEARS_EMPLOYED, Importance: 0.0369

Рисунок 25 – Топ–10 признаков, выбранных моделью Random Forest

На рисунке можно увидеть, что среди топ-10 признаков, внесших наибольший вклад, есть несколько признаков, созданных в рамках эксперимента создания новых признаков в рамках этой дипломной работы.

Итоговая модель показала хорошие результаты на тестовых данных и может быть использована для прогнозирования дефолта заемщиков. Однако она обладает большей ресурсоемкостью по сравнению с логистической регрессией и требует больше времени для определения потенциального дефолта заемщика.

5. Заключение

5.1. Общие выводы и рекомендации

В данной дипломной работе было проведено исследование оценки кредитных рисков с применением методов машинного обучения. В рамках работы были обработаны и нормализованы исходные данные, а также протестированы различные подходы к созданию признаков. Основное внимание было уделено двум моделям машинного обучения: логистической регрессии и случайному лесу.

В ходе исследования были протестированы две модели машинного обучения: логистическая регрессия и случайный лес. Логистическая регрессия показала отличные результаты в терминах производительности, скорости работы и точности. Эта модель является привлекательным вариантом для применения в практических ситуациях, особенно в случаях, когда требуется быстрое принятие решений, например, при личной подаче заявки в банке или онлайн-заявках.

С другой стороны, случайный лес также показал достаточно хорошую точность в оценке кредитных рисков. Однако, следует отметить, что случайный лес имеет более высокую вычислительную сложность и требует большего объема вычислительных ресурсов по сравнению с логистической регрессией. Это делает его менее привлекательным вариантом для использования в ситуациях, где время является критическим фактором, как в случае личной подачи заявки в банке или онлайн-заявках, где требуется быстрая обработка данных и принятие решений.

В отношении создания признаков, использовались два подхода: основанный на предметной области и генетическое программирование. Признаки, созданные на основе предметной области, показали высокую значимость и вошли в топ-5 признаков, что подтверждает их важность для оценки кредитных рисков. С другой стороны, признаки, созданные с использованием генетического программирования, хотя и не попали в топ-5,

представляют собой ценный способ генерации дополнительных признаков. Это особенно важно для области оценки кредитных рисков, так как улучшение точности моделей и обогащение данных может привести к более надежным и эффективным решениям.

Одним из ключевых выводов является необходимость продолжения исследовательской работы по генерации признаков с использованием генетического программирования. Данная методика может быть полезна в кредитном скоринге, так как специалисты из области могут внести экспертные знания и опыт для создания новых информативных признаков. Это позволит более точно оценивать кредитные риски и принимать взвешенные решения о выдаче кредитов.

В целом, данная дипломная работа позволила получить ценные результаты и выводы в области оценки кредитных рисков с применением методов машинного обучения. Создание и использование новых признаков, основанных на предметной области и генетическом программировании, позволило улучшить точность моделей. Модели машинного обучения, включая логистическую регрессию и случайный лес, проявили хорошую производительность и показали потенциал для применения в реальных банковских сценариях.

Данная работа является отправной точкой для дальнейших исследований и разработок в области оценки кредитных рисков. Рекомендуется проводить более широкие исследования с использованием различных методов машинного обучения, улучшать производительность моделей и продолжать работу по генерации признаков с учетом экспертных знаний и опыта из области кредитного скоринга. Дополнительное исследование на различных наборах данных поможет обобщить результаты и повысить общую применимость моделей.

5.2. Ограничения и перспективы дальнейших исследований

Данная работа является отправной точкой для дальнейших исследований и разработок в области оценки кредитных рисков. Рекомендуется проводить более широкие исследования с использованием различных методов машинного обучения, улучшать производительность моделей и продолжать работу по генерации признаков с учетом экспертных знаний и опыта из области кредитного скоринга. Дополнительное исследование на различных наборах данных поможет обобщить результаты и повысить общую применимость моделей.

Важным аспектом дальнейших исследований является также работа над интерпретируемостью моделей. Понимание того, какие признаки и факторы влияют на принятие решения модели, является критически важным для принятия обоснованных и надежных решений в области кредитных рисков. Дальнейшее развитие методов интерпретируемости и их применение к моделям машинного обучения будет способствовать более доверительному использованию моделей в практических ситуациях.

Немаловажным фактором является подроб гиперпараметров моделей, так как они также влияют на точность предсказаний. Подбор гиперпараметрова для модели Случайный лес является большим полем для дальнейших исследований в этой области.

Одним из ограничений данной работы является ограниченный объем данных и использование только двух моделей машинного обучения. Расширение набора данных и проведение экспериментов с другими алгоритмами машинного обучения могут дать более полное представление о возможностях и ограничениях применения методов оценки кредитных рисков.

В заключение, данная дипломная работа представляет ценные результаты и выводы в области оценки кредитных рисков с использованием методов машинного обучения. Создание новых признаков и применение моделей машинного обучения позволяют повысить точность оценки кредитных рисков

и принимать взвешенные решения в банковской сфере. Однако, для дальнейшего развития и применения этих методов рекомендуется проведение более широких исследований, улучшение производительности моделей и работа над интерпретируемостью. Это позволит создать более надежные и эффективные инструменты для оценки кредитных рисков и улучшения процесса принятия решений в банковской сфере.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Носова Т. П., Паршин А. Б., Терпицкая К. И. Классификация банковских рисков и мероприятия по их снижению с целью оптимизации банковской деятельности [Текст] / Носова Т. П., Паршин А. Б., Терпицкая К. И. // Вестник Академии знаний. – 2022. – №. 6 (53). – С. 349–354.
2. Лаврушин О. И., Афанасьева О. Н. Банковское дело. Современная система кредитования: учебное пособие. / Лаврушин О. И., Афанасьева О. Н. – Москва: КноРус, – 2019. – 358 с.
3. Антонова С. И. Совершенствование методики оценки кредитоспособности заемщика коммерческого банка на примере ПАО «Сбербанк» [Текст]: магистерская диссертация (38.04.02) / Антонова С. И.; Сибирский федеральный университет. – Красноярск, 2019. – 82 с.
4. Поляков К. Л., Жукова Л. В. Опыт моделирования вероятности кредитного дефолта клиентов микрофинансовых организаций (на примере одной МФО) [Текст] / Поляков К. Л., Жукова Л. В. // Экономический журнал Высшей школы экономики. – 2019. – Т. 23. – №. 4. – С. 497–523.
5. Шевелев А., Бузанов Г. Модель вероятности дефолта с использованием транзакционных данных российских компаний [Текст] / Шевелев А., Бузанов Г. // Серия докладов об экономических исследованиях Банка России. – 2022. – №97. – С. 9–19.
6. Инхиреева Т. И. Методология подготовки исходных данных для построения кредитного скоринга [Текст] / Инхиреева Т. И. // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов VI Международной конференции, 14–19 октября 2019 г., Томск. – Томск, 2019. – 2019. – С. 246–251.
7. Григорьев С. В., Пыжова В. В. Роль подготовки данных для анализа предложений и прогнозирования спроса [Текст] / Григорьев С. В., Пыжова В. В. // Вызовы современности и стратегии развития общества в условиях новой реальности. – 2022. – С. 142–150.

8. Sahoo K. et al. Exploratory data analysis using Python [Текст] / Sahoo K. // International Journal of Innovative Technology and Exploring Engineering (IJITEE). – 2019. – Т. 8. – №. 12. – p. 19–31.
9. Боровской А. А., Кривошеин И. А. Машинное обучение в экономике [Текст] / Боровской А. А., Кривошеин И. А. // Международная научно–техническая конференция молодых ученых БГТУ им. ВГ Шухова, посвященная 300–летию Российской академии наук. – 2022. – № 17. – С. 102–105.
10. Kaukin A., Kosarev V. Modeling and Forecasting Production Indices Using Artificial Neural Networks, Taking Into Account Intersectoral Relationships and Comparing the Predictive Qualities of Various Architectures [Текст] / Kaukin A., Kosarev V. // SSRN 3860098. – 2021. – p. 36–54.
11. Юрченко Т. В. Развитие методов прогнозирования банкротства в цифровую эпоху [Текст] / Юрченко Т. В. // Теория и практика применения цифровых технологий при управлении финансами и экономическими процессами. – 2021. – С. 98–106.
12. Бредихина К. В., Ингман Н. И. Управление риском потребительского кредитования в АО «Газпромбанк» [Текст] / Бредихина К. В., Ингман Н. И. // Экономика сегодня: современное состояние и перспективы развития (Вектор–2020). – 2020. – С. 83–86.
13. Митяков С. Н., Митяков Е. С. Машинное обучение в задачах исследования инновационных процессов [Текст] / Митяков С. Н., Митяков Е. С. // Журнал прикладных исследований. – 2020. – Т. 1. – №. 4. – С. 6–13.
14. Романов А. Г. и др. Моделирование, оптимизация и информационные технологии [Текст] / Романов А. Г. // Моделирование, оптимизация и информационные технологии. – 2022. – Т. 10. – №. 3. – С. 13–14.
15. Алексеева Д. Д., Марочкина А. В., Парамонов А. И. Оптимизация мобильного трафика методами машинного обучения [Текст] / Алексеева Д.

Д., Марочкина А. В., Парамонов А. И. // Информационные технологии и телекоммуникации. – 2021. – Т. 9. – №. 1. – С. 1–12.

16. Мокеев В. В., Войтецкий Р. В. Прогнозирование банкротств предприятий с помощью экстремального градиентного бустинга [Текст] / Мокеев В. В., Войтецкий Р. В. // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. – 2020. – Т. 9. – №. 3. – С. 77–90.

17. Сулейманова А. Н. Обзор развития алгоритмов деревьев решений [Текст] / Сулейманова А. Н. // Социология: методология, методы, математическое моделирование. – 2020. – №. 50–51. – С. 64–97.

18. Караваев А. В., Мосин В. Г. Оценка важности категориальных признаков с использованием One-Hot-кодирования для модели линейной и гребневой регрессии [Текст] / Караваев А. В., Мосин В. Г. // Прикладная математика и информатика: современные исследования в области естественных и технических наук. – 2022. – С. 174–179.

19. Лимановская О. В., Алферьева Т. И. Основы машинного обучения: учебное пособие / Лимановская О. В., Алферьева Т. И. – Екатеринбург: Издательство Уральского университета, 2020. – 88 с.

20. Исаев Д. В. Динамическое ансамблевое обучение для оценки кредитоспособности [Текст] / Исаев Д. В. // Инновации и инвестиции. – 2022. – №. 3. – С. 74–79.

21. Жураев Ж. Д. Использование методов машинного обучения в моделировании кредитного скоринга [Текст] / Жураев Ж. Д. // Вестник науки. – 2021. – Т. 2. – №. 6–1 (39). – С. 87–91.

22. Шмелева А. Г. и др. Программная модель оценки кредитоспособности клиентов с применением алгоритмов искусственного интеллекта [Текст] / Шмелева А. Г. // Труды НГТУ им. ПЕ Алексева. – 2020. – №. 3 (130). – С. 72–79.

23. Mushava J., Murray M. A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified

focal loss function [Текст] / Mushava J., Murray M. // Expert Systems with Applications. – 2022. – Т. 202. – p. 117233.

24. Утегенов Н. Б. Искусственный интеллект на сегодняшний день [Текст] / Утегенов Н. Б. // Universum: технические науки. – 2022. – №. 7–1 (100). – С. 27–30.

25. Акимов А. А., Валитов Д. Р., Кубряк А. И. Предварительная обработка данных для машинного обучения [Текст] / Акимов А. А., Валитов Д. Р., Кубряк А. И. // Научное обозрение. Технические науки. – 2022. – №. 2. – С. 26–31.

26. Шерстнев П. А., Липинский Л. В. Эволюционный алгоритм проектирования искусственных нейронных сетей с перераспределением ресурсов [Текст] / Шерстнев П. А., Липинский Л. В. // Российская наука, инновации, образование–РОСНИО–2022. – 2022. – С. 131–141.

27. Зыков Д. А., Комашинский В. В. Исследование статистических методов для повышения защищенности информационных систем [Текст] / Зыков Д. А., Комашинский В. В. // Обработка, передача и защита информации в компьютерных системах'22. – 2022. – С. 236–241.

28. Векслер В. А. Машинное обучение на основе алгоритма k-ближайших соседей [Текст] / Векслер В. А. // Вызовы цифровой экономики: итоги и новые тренды. – 2019. – С. 110–115.

29. Дорофеев В. С., Волосатова Т. М. Ансамблирование методов обнаружения выбросов при подготовке обучающей выборки данных [Текст] / Дорофеев В. С., Волосатова Т. М. // Моделирование, оптимизация и информационные технологии. – 2022. – № 10 (3). – С. 36–44.