Министерство образования Республики Беларусь
Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»

Факультет инженерно-экономический
Кафедра экономической информатики
Дисциплина «Эконометрика»

«К ЗАЩИТЕ ДОПУСТИТЬ»
Руководитель курсового проекта
Ассистент кафедры ЭИ
_____ Р. С. Нагулевич
___.___.2024

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к курсовой работе
на тему:
«ПОСТРОЕНИЕ МНОГОФАКТОРНОЙ МОДЕЛИ ОЦЕНКИ РИСКОВ НЕВОЗВРАТА БАКНОВСКИХ КРЕДИТОВ ФИЗИЧЕСКИМИ ЛИЦАМИ»

БГУИР КР 1-40 05 01-02 027 ПЗ

Выполнил студент группы 272302
Шимчёнок Екатерина Сергеевна
_____
(подпись студента)
Курсовой проект представлен на проверку _____.2025
_____
(подпись студента)

Минск 2025

РЕФЕРАТ

БГУИР КР 1-40 05 01-02 027 ПЗ

Шимчёнок, Е.С. Построение многофакторной модели оценки рисков невозврата банковских кредитов физическими лицами/ Е. С. Шимчёнок – Минск: БГУИР, 2025. – 61 с.

Пояснительная записка 61 с., 38 рис., 10 ист., 3 приложения

СВЕДЕНИЯ О ЗАЁМЩИКАХ, СПЕЦИФИКАЦИЯ ПОСТРОЕНИЕ И ОЦЕНКА МОДЕЛИ, АНАЛИЗ ОСТАТКОВ, ЭКОНОМИЧЕСКИЙ АНАЛИЗ, ГИСТОГРАММЫ РАСПРЕДЕЛЕНИЙ, ДИАГРАММЫ РАССЕЯНИЯ, ГРАФИКИ ОСТАТКОВ

Цель курсовой работы: разработать и апробировать многофакторную эконометрическую модель, которая позволит оценивать и прогнозировать риск невозврата банковских кредитов физическими лицами.

Методология проведения работы: в процессе решения поставленных задач использованы методы статистического анализа исходных данных, корреляционного анализа для выявления взаимосвязей и регрессионного моделирования методом наименьших квадратов, а также проверка гипотез с использованием F- и t-критериев для оценки значимости коэффициентов. Дополнительно, анализ остатков применяется для контроля соблюдения предпосылок регрессионного анализа (нормальность, гомоскедастичность, отсутствие автокорреляции), что обеспечивает корректность и достоверность построенной модели.

Результаты работы: в результате исследования был сконструирована многофакторная модель, которая демонстрирует высокую степень объяснения вариативности риска невозврата банковских кредитов. Все оценённые коэффициенты оказались статистически значимыми, что подтверждено результатами F- и t-тестов. Анализ остатков подтвердил корректность применения МНК, а полученные прогнозные значения и доверительные интервалы подтверждают практическую применимость модели для оценки кредитных рисков.

Анализ данных и построение модели осуществлялись с использованием Python (библиотеки Pandas, SciPy, Matplotlib, Seaborn, Scikit-Learn для обработки данных, визуализации и регрессионного анализа методом наименьших квадратов), Excel и EViews для проведения статистических тестов, проверки значимости коэффициентов и детального анализа остатков.

Область применения результатов: результаты исследования могут быть применены банками и финансовыми учреждениями для оптимизации процессов оценки кредитного риска, эффективного управления кредитным портфелем и принятия оперативных решений в сфере кредитования.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современных условиях банковская система является ключевым звеном финансовой стабильности, обеспечивая достаточный уровень кредитования и поддержку экономического развития страны. Однако рост дефолтов по кредитам физическим лицам негативно сказывается на доходности банков и повышает риск значительных финансовых потерь. Эффективное управление кредитными рисками требует применения современных методов статистического и эконометрического анализа, способных не только выявлять факторы, влияющие на вероятность невозврата кредита, но и обеспечивать качественное прогнозирование риска.

Цель данной курсовой работы заключается в разработке многофакторной эконометрической модели оценки рисков невозврата банковских кредитов физическими лицами с использованием современных средств анализа (Python, Excel, EViews). Для достижения поставленной цели необходимо решить следующие задачи:

– провести анализ теоретических основ кредитных рисков и обзор современной литературы по данной теме;

– собрать и провести предварительную обработку статистических данных, включая устранение выбросов;

– идентифицировать релевантные факторные переменные, влияющие на риск дефолта;

– осуществить корреляционный анализ для выявления взаимосвязей между переменными;

– построить регрессионную модель методом наименьших квадратов и оценить её параметры с использованием F- и t-критериев;

– провести анализ остатков для проверки выполнения предпосылок корректности модели;

– выполнить прогнозирование риска невозврата с построением доверительных интервалов и оценить практическую применимость модели для оптимизации процессов управления кредитным портфелем.

Результаты данной работы позволят выявить значимые детерминанты риска дефолтов, что является важным шагом к улучшению стратегий кредитования и минимизации финансовых потерь банковских учреждений. Они также могут стать базой для дальнейших исследований в области управления банковскими рисками и послужить основой для разработки специализированных программных решений, направленных на повышение эффективности риск-менеджмента в банковском секторе.

В рамках выполнения курсовой работы исследование и моделирование осуществляются в полном соответствии с требованиями, установленными нормативным документом ТП БГУИР 01-2017 «Дипломные проекты (работы)».

1 АНАЛИЗ ЛИТЕРАТУРНЫХ ИССЛЕДОВАНИЙ И ПРОГРАММНЫХ РЕШЕНИЙ

1.1 Описание и анализ предметной области

Актуальность проблемы оценки кредитного риска обусловлена необходимостью минимизации финансовых потерь банков и повышения эффективности управления кредитными портфелями. В условиях динамично меняющейся экономической среды, характерной для таких стран, как Россия и Беларусь, эта задача становится особенно важной. Предметная область исследования охватывает анализ влияния различных факторов на кредитную оценку заемщиков, и именно эти аспекты имеют принципиальное значение для обеспечения стабильности банковского сектора. Современные коммерческие банки при принятии решений учитывают широкий спектр характеристик заемщиков. Боброва О.П. в своей работе «Оценка кредитоспособности физических лиц» подчёркивает значимость интеграции демографических данных, таких как возраст, пол и семейное положение, для создания базового профиля клиента. Возраст имеет существенное значение, так как молодые заемщики, как правило, склонны к более рискованному финансовому поведению, в то время как зрелые клиенты демонстрируют большую стабильность. Семейное положение также является важным индикатором ответственности, особенно если заемщик имеет иждивенцев, что подтверждает Курилов К.Ю. в своей статье «Теоретические аспекты оценки кредитоспособности заёмщиков-физических лиц».

Экономическая состоятельность заемщика играет центральную роль в определении его платежеспособности. Ключевыми параметрами здесь являются уровень дохода, стабильность занятости и уровень образования. Регулярный доход и постоянное место работы отражают финансовую устойчивость заемщика, в то время как высокий уровень образования свидетельствует о его профессиональных перспективах и способности к увеличению дохода в будущем. Эти аспекты, как подчеркивают Климов Д.О. и Валько Д.В. в своей работе «Методы оценки кредитоспособности физических лиц: отечественный и зарубежный опыт», формируют основу для прогнозирования выполнения заемщиком своих финансовых обязательств.

Кредитная история заемщика остаётся одним из самых надёжных источников информации для анализа. История платежей, структура займов, долговая нагрузка и регулярность выплат дают банкам возможность оценить дисциплину клиента в финансовых вопросах. Высокая долговая нагрузка является тревожным сигналом, указывающим на потенциальные сложности с погашением новых кредитов. Боброва О.П. обращает внимание на то, что кредитная история зачастую является основным индикатором при оценке риска невозврата средств.

Макроэкономические показатели также играют важную роль в модели оценки кредитоспособности. Внешние факторы, такие как уровень инфляции, безработицы и изменения валютных курсов, значительно влияют на реальный

уровень доходов заемщиков и их экономическую стабильность. Например, рост безработицы или инфляции может уменьшить покупательную способность заемщика и увеличить вероятность дефолта. Эти переменные особенно важны для корректировки моделей оценки в условиях нестабильной экономической среды.

Современные тенденции в области кредитного скоринга предполагают включение социально-психологических характеристик и региональных особенностей заемщиков. Финансовая грамотность, лояльность к банку, поведение в стрессовых ситуациях и другие поведенческие аспекты становятся важными факторами, влияющими на платежеспособность. Например, анализ социальной активности клиента, в том числе его взаимодействие с бонусными программами, может быть использован для оценки уровня доверия и финансовой ответственности. Региональные особенности, такие как состояние экономики, уровень инфраструктуры и специфика рынка труда, также оказывают значительное влияние на финансовую устойчивость заемщика, что подчёркивают Климов Д.О. и Валько Д.В. в своих исследованиях.

Эмпирические модели анализа кредитного риска, основанные на статистических данных и реальных наблюдениях, сегодня признаются одним из наиболее надёжных инструментов для прогнозирования. Именно они позволяют учесть сложные взаимосвязи между множеством факторов, демонстрируя высокую точность и надёжность. Однако при всей их эффективности нельзя утверждать, что существует единое универсальное решение. Банковская отрасль требует индивидуального подхода к каждой организации, так как специфика клиентской базы, особенности рынка и стратегические цели могут существенно варьироваться. Разработка уникальных решений, адаптированных к конкретным условиям, становится ключевым элементом успеха. Это также подтверждается исследованиями авторов, таких как Леонов И.И., который в своей статье «Использование методов интеллектуального анализа данных в работе бюро кредитных историй. Российский и зарубежный опыт» подчёркивает важность применения методов машинного обучения и нейронных сетей для адаптации моделей к специфике различных рынков.

Интеграция всех перечисленных факторов позволяет банкам разрабатывать адаптивные модели оценки кредитоспособности. Использование современных технологий, таких как машинное обучение и большие данные, даёт возможность улучшить точность прогнозов и минимизировать финансовые риски. Комплексный подход, который объединяет демографические, экономические, макроэкономические и поведенческие аспекты, делает модели более гибкими и надёжными. Анализ практики российских и белорусских банков показывает, что учёт социально-психологических и региональных факторов позволяет создать более точные инструменты для принятия кредитных решений.

Таким образом, оценка кредитоспособности физических лиц представляет собой многоуровневый процесс, который требует тщательного анализа широкого спектра характеристик заемщиков. Разделение факторов на группы, включающие как традиционные, так и современные аспекты, повышает эффективность моделей и способствует более глубокому пониманию потребностей клиентов. Это особенно важно в условиях экономической неопределённости, где адаптивность и точность моделей становятся ключевыми элементами успешного управления кредитным портфелем. Теоретический и практический анализ подтверждает важность дальнейшего развития методов кредитного скоринга с учётом новых технологий и подходов. Однако окончательное решение всегда должно быть основано на особенностях каждой отдельной банковской организации, что требует индивидуального подхода при адаптации эмпирических моделей под конкретные задачи и условия.

1.2 Обзор литературы

Классические зарубежные исследования также заложили фундамент для данного направления. Так, работа «Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy» (Altman, 1968) впервые предложила использовать финансовые коэффициенты для прогнозирования дефолта, а последующее исследование «Financial Ratios and the Probabilistic Prediction of Bankruptcy» (Ohlson, 1980) расширило эту концепцию за счёт интеграции дополнительных переменных. Многофакторный подход получил развитие в трудах Hand и Henley «Statistical Classification Methods in Consumer Credit Scoring: a Review» (Hand & Henley, 1997), где подробно обсуждались методы статистической классификации в кредитном скоринге. Исследование «Credit Scoring with Macroeconomic Variables» (Bellotti & Crook, 2009) показало, что интеграция макроэкономических индикаторов существенно улучшает прогнозные характеристики моделей кредитного риска, а статья «Consumer Credit Risk Measurement and Profitability Analysis» (Thomas, 2009) демонстрирует, что объединение индивидуальных финансовых показателей с экономическими данными позволяет создавать более гибкие и адаптивные модели. Работа «The Use of Discriminant Analysis in Credit Scoring» (Mester, 1991) подчёркивает значимость классических методов дискриминантного анализа в построении моделей оценки кредитного риска.

Обобщая результаты литературного анализа, можно сделать вывод, что применение многофакторного подхода является оптимальным решением для оценки кредитного риска. Перечисленные научные публикации в сочетании с отечественными исследованиями, демонстрируют, что интеграция дополнительных экономических, социальных и региональных факторов способствует улучшению качества прогноза дефолтов. Такой комплексный подход позволяет не только минимизировать риски банков, но и оптимизировать процессы принятия управленческих решений.

2 СПЕЦИФИКАЦИЯ МОДЕЛИ

Без четко определенной спецификации невозможно обеспечить корректность модели и её практическую применимость. Этот этап включает выбор зависимой переменной и независимых факторов. Процесс спецификации модели создаёт теоретическую и методологическую основу для дальнейших шагов исследования, позволяя не только понять динамику факторов риска, но и использовать полученные результаты для улучшения процессов принятия решений в банковской сфере.

2.1 Идентификация переменных

Идентификация переменных – это процесс определения и уточнения переменных, которые будут использоваться в исследовании или анализе. Это важный шаг в процессе исследования, поскольку правильная идентификация переменных помогает обеспечить точность и достоверность результатов.

Для построения модели я буду использовать датасет, в котором будет содержаться информация о заёмщиках. Переменные датасета их описание и тип представлены в таблице 1.

Таблица 1 – Идентификация переменных

Название переменной	Описание	Тип	Тип данных
Age	Возраст заемщика	Экзогенная	Целое число
AnnualIncome	Годовой доход заемщика	Экзогенная	Целое число
CreditScore	Кредитный рейтинг	Экзогенная	Целое число
EmploymentStatus	Статус занятости (например, наёмный работник)	Экзогенная	Строка
EducationLevel	Уровень образования	Экзогенная	Строка
Experience	Стаж работы (в годах)	Экзогенная	Целое число
LoanAmount	Запрашиваемая сумма кредита	Экзогенная	Целое число
LoanDuration	Срок кредита (в месяцах)	Экзогенная	Целое число
MaritalStatus	Семейное положение	Экзогенная	Строка
NumberOfDependents	Количество иждивенцев	Экзогенная	Целое число

Продолжение таблицы 1

Название переменной	Описание	Тип	Тип данных
HomeOwnershipStatus	Статус владения жильем (например, аренда или личная собственность)	Экзогенная	Строка
MonthlyDebtPayments	Ежемесячные платежи по долгам	Экзогенная	Целое число
CreditCardUtilizationRate	Уровень использования кредитных карт	Экзогенная	Дробное число
NumberOfOpenCreditLines	Количество открытых кредитных линий	Экзогенная	Целое число
NumberOfCreditInquiries	Количество кредитных запросов	Экзогенная	Целое число
DebtToIncomeRatio	Соотношение долга к доходу	Экзогенная	Дробное число (в долях)
BankruptcyHistory	История банкротств	Экзогенная	Целое число
LoanPurpose	Цель кредита	Экзогенная	Строка
PreviousLoanDefaults	Наличие дефолтов по предыдущим кредитам	Экзогенная	Целое число
PaymentHistory	История платежей	Экзогенная	Целое число
SavingsAccountBalance	Баланс на сберегательном счете	Экзогенная	Целое число
CheckingAccountBalance	Баланс на расчетном счете	Экзогенная	Целое число
TotalAssets	Общий объем активов	Экзогенная	Целое число
TotalLiabilities	Общий объем обязательств	Экзогенная	Целое число
MonthlyIncome	Ежемесячный доход	Экзогенная	Дробное число

Продолжение таблицы 1

Название переменной	Описание	Тип	Тип данных
UtilityBillsPaymentHistory	История оплаты коммунальных услуг	Экзогенная	Дробное число
JobTenure	Срок работы на текущем месте	Экзогенная	Целое число
NetWorth	Чистая стоимость активов	Экзогенная	Целое число
BaseInterestRate	Базовая процентная ставка	Экзогенная	Дробное число (в долях)
InterestRate	Процентная ставка по кредиту	Экзогенная	Дробное число (в долях)
MonthlyLoanPayment	Ежемесячный платеж по кредиту	Экзогенная	Дробное число
TotalDebtToIncomeRatio	Общий долг в процентном соотношении к доходу	Экзогенная	Дробное число (в долях)
LoanApproved	Факт одобрения кредита	Экзогенная	Целое число (0 или 1)
RiskScore	Оценка риска	Эндогенная	Дробное число

Все экономические переменные подразделяются на два типа: эндогенные и экзогенные.

Экзогенными (независимыми) называются переменные, значения которых определяются вне данной модели. Эндогенными (зависимыми) называются экономические переменные, значения которых определяются (объясняются) внутри модели в результате одновременного взаимодействия соотношений, образующих модель.

В модели ценообразования автомобилей эндогенной переменной будет оценка риска. К экзогенным отнесем все остальные переменные.

В таблице 2 представлены уникальные значения четырёх категориальных признаков выборки с их распределением. Например, признак Статус занятости (EmploymentStatus) включает три категории: работает (Employed) – 17036 записей, самозанятый (Self-Employed) – 1573 записи, безработный (Unemployed) – 1391 запись. Такой анализ позволяет проанализировать распределение данных и оценить их разнообразие для последующего анализа.

Таблица 2 – Уникальные значения категориальных переменных

Категориальный признак	Уникальные значения и их распределение
EmploymentStatus (Статус занятости)	Employed (Работает): 17036 записей, Self-Employed (Самозанятый): 1573 записи, Unemployed (Безработный): 1391 запись
EducationLevel (Уровень образования)	Bachelor (Бакалавр): 6054 записи, High School (Среднее): 5908 записей, Associate (Ассоциат): 4034 записи, Master (Магистр): 3050 записей, Doctorate (Доктор наук): 954 записи
MaritalStatus (Семейное положение)	Married (В браке): 10041 запись, Single (Одинок(а)): 6078 записей, Divorced (Разведен(а)): 2882 записи, Widowed (Вдовец/Вдова): 999 записей
HomeOwnershipStatus (Статус жилья)	Mortgage (Ипотека): 7939 записей, Rent (Снимает): 6087 записей, Own (Владеет): 3938 записей, Other (Другое): 2036 записей
LoanPurpose (Цель кредита)	Home (Покупка жилья): 5925 записей, Debt Consolidation (Консолидация долгов): 5027 записей, Auto (Автомобиль): 4034 записи, Education (Образование): 3008 записей, Other (Другое): 2006 записей

Гистограмма не симметрична и не напоминает форму колокола, типичную для нормального распределения. Есть несколько пиков на гистограмме.

Для построения модели регрессии в эконометрическом анализе категорические переменные также необходимо преобразовать в числовой формат. Один из подходов – это Label Encoding (метод меток), который присваивает каждому уникальному значению категории числовой код. Это упрощает включение категориальных признаков в модели регрессии, сохраняя при этом информацию о различиях между категориями.

Например, если у нас есть переменная уровень образования (EducationLevel) с категориями [«Среднее образование», «Бакалавр», «Магистр»], мы можем присвоить им значения 0, 1 и 2 соответственно. На практике это можно реализовать с помощью языка Python следующим образом:

```
from sklearn.preprocessing import LabelEncoder
# Выделяем категориальные переменные
cat_vars = ["EmploymentStatus", "EducationLevel", "MaritalStatus", "HomeOwnershipStatus", "LoanPurpose"]
```

```
# Кодируем каждый признак с помощью объекта класса LabelEncoder
encoder = LabelEncoder()
for column in cat_vars:
    df[column] = encoder.fit_transform(df[column]).
```

После кодирования переменные EmploymentStatus, EducationLevel, MaritalStatus, HomeOwnershipStatus, LoanPurpose имеют вид, показанный на рисунке 2.1.

	EmploymentStatus	EducationLevel	MaritalStatus	HomeOwnershipStatus	LoanPurpose
0	0	4	1	2	3
1	0	0	2	0	1
2	0	1	1	3	2
3	0	3	2	0	3
4	0	0	1	0	1
...	...	...	...	...	...
19995	0	3	1	3	0
19996	0	0	1	3	1
19997	0	1	1	0	3
19998	0	3	1	2	1
19999	0	0	1	2	4

20000 rows × 5 columns

Рисунок 2.1 – Закодированные категориальные признаки

В этом разделе проведена идентификация переменных: определены переменные, их описание, тип и формат данных (таблица 1). Проанализированы категориальные признаки, выявлены их уникальные значения и распределение (таблица 2), что позволило оценить структуру выборки. Для корректного построения регрессионной модели предложено преобразование категориальных переменных в числовой формат с использованием метода Label Encoding.

Следующим шагом является изучение распределения целевого признака и факторных переменных, а также выявление выбросов и аномальных значений. Этот этап необходим для повышения качества модели и обеспечения её адекватности, поскольку корректная обработка данных – важный компонент эконометрического анализа.

2.2 Анализ распределения целевого признака и факторных, устранение выбросов и аномальных значений признаков

Проведем анализ распределения каждого признака, построим гистограммы и ящики с усами.

Целевая переменная RiskScore представляет собой числовой показатель, отражающий степень кредитного риска заемщика. Она рассчитывается на основе множества факторов, включая кредитную историю, уровень доходов, соотношение долга к доходу, наличие дефолтов и другие параметры. RiskScore используется кредиторами для оценки вероятности выполнения заемщиком своих финансовых обязательств, где более высокий показатель указывает на низкий риск, а низкий – на высокую вероятность дефолта. Этот показатель играет ключевую роль в принятии решений о выдаче кредитов и условиях их предоставления. В нашей задаче он измеряется в шкале от 0 до 100. Гистограмма распределения показана на рисунке 2.2.

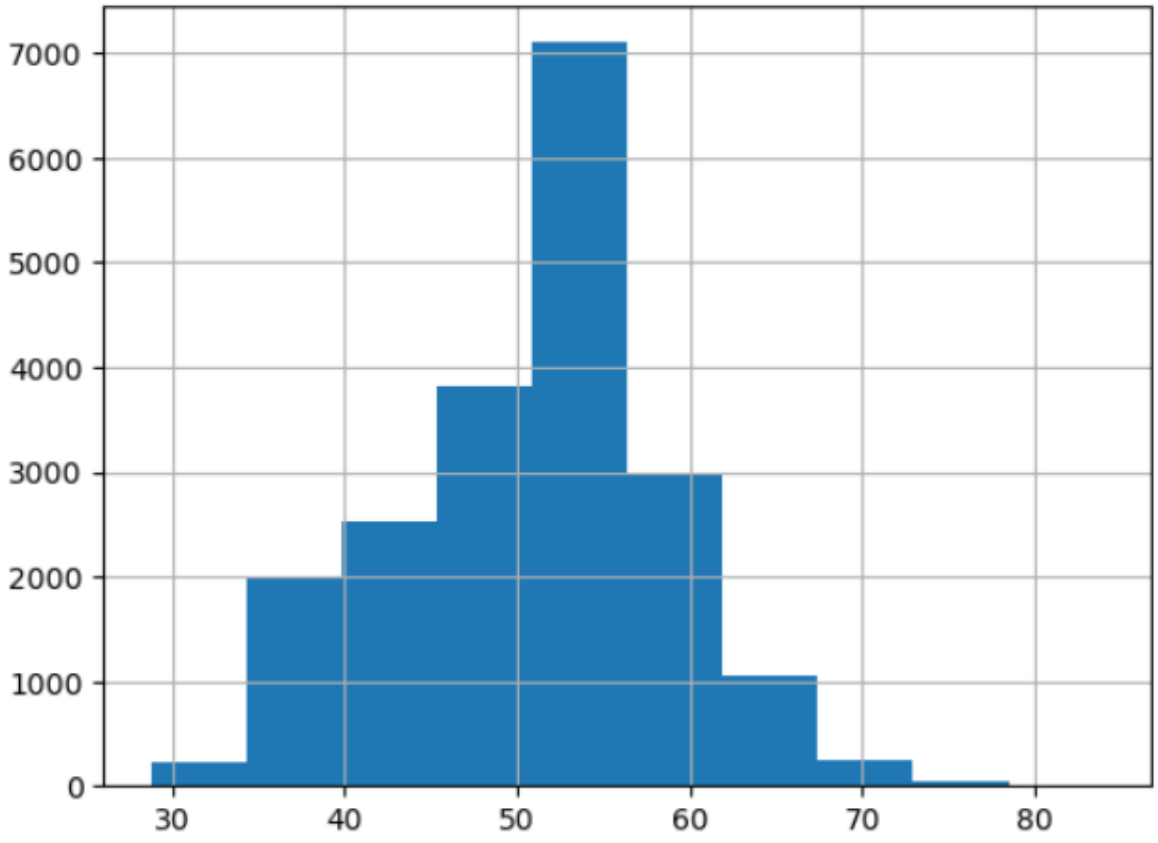


Рисунок 2.2 – Гистограмма распределения целевого признака RiskScore

График «ящик с усами» на рисунке 2.3 (boxplot) визуализирует распределение признака RiskScore: медиана отображается горизонтальной линией в коробке, границы коробки показывают первый и третий квартили (IQR), «усы» расширяют диапазон до 1.5×IQR, а выбросы представлены точками за пределами «усов».

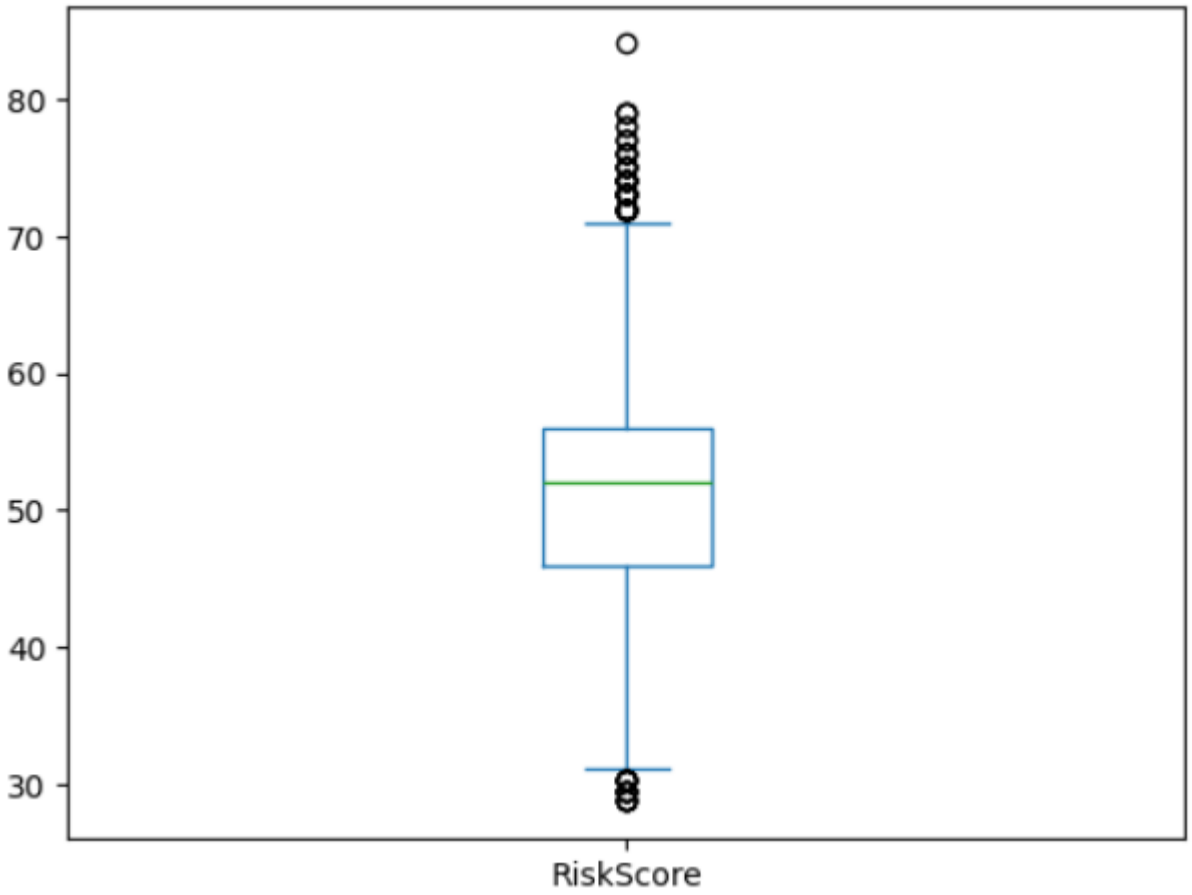


Рисунок 2.3 – График «ящик с усами» для целевого признака

По визуализациям распределения можно сделать вывод что оно стремится к нормальному, а в данных имеются выбросы.

На рисунке А.1 в приложении А приведен датафрейм с описательной статистикой, который предоставляет детальную информацию о каждом признаке. В ней отображаются тип переменной, количество пропущенных значений и их процент (в нашей выборке их нет), а также число уникальных значений. Помимо этого, включены ключевые метрики: среднее значение, стандартное отклонение, минимальные и максимальные значения, а также квантили, такие как первый квартиль (25%), медиана (50%) и третий квартиль (75%). Эти данные позволяют проанализировать распределение признаков, выявить выбросы и оценить полноту данных.

Статистика Жака-Бера, представленная в датафрейме, помогает оценить, насколько распределение каждого признака близко к нормальному. Если значение этой статистики значительно отклоняется от нуля, а p-value меньше 0,05, можно сделать вывод, что распределение не является нормальным. Это часто указывает на наличие асимметрии или длинных «хвостов» в данных. В общем, распределения многих признаков могут быть далеки от нормального, что характерно для реальных наборов данных.

Для целевого признака «RiskScore» проверка начинается с расчёта статистики Жака-Бера (108,25), p-value намного меньше любого адекватного

уровня значимости, значит можно сделать вывод о ненормальности распределения. В таком случае возможно потребуется преобразование данных, например, логарифмирование или применение преобразования Бокса-Кокса, чтобы приблизить распределение к нормальному. Такой подход позволяет учесть особенности целевого признака и скорректировать методы анализа или построения модели.

Теперь необходимо рассмотреть распределения экзогенных переменных. На рисунке 2.4 показаны гистограммы для непрерывных признаков, распределение которых близко к нормальному.

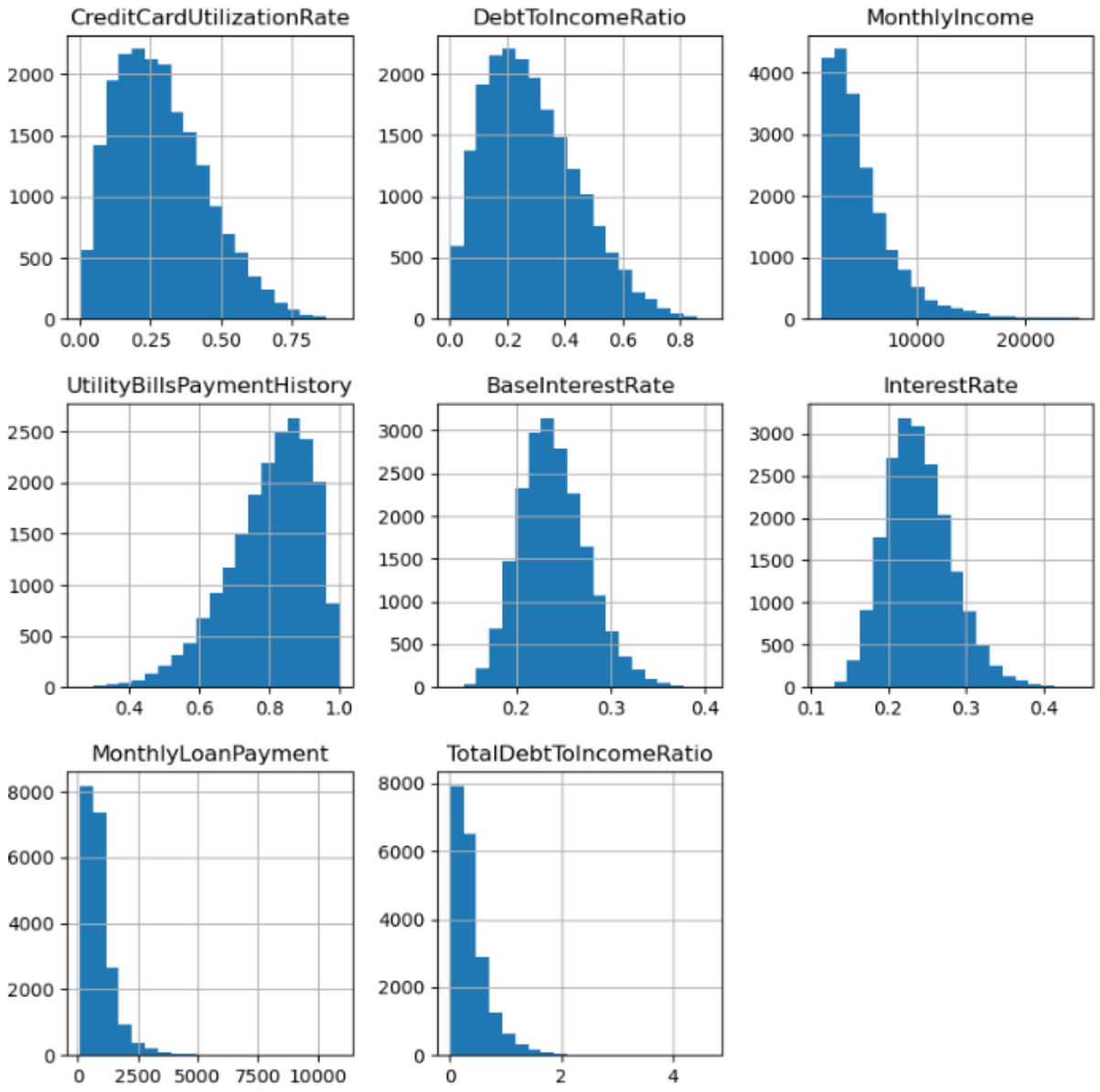


Рисунок 2.4 – Гистограммы распределения непрерывных признаков

В нашей выборке непрерывными признаками являются: коэффициент использования кредитной карты (CreditCardUtilizationRate), коэффициент долга к доходу (DebtToIncomeRatio), ежемесячный доход (MonthlyIncome), история оплаты коммунальных счетов (UtilityBillsPaymentHistory), базовая

процентная ставка (BaseInterestRate), процентная ставка (InterestRate), ежемесячный платеж по кредиту (MonthlyLoanPayment) и общий коэффициент долга к доходу (TotalDebtToIncomeRatio). Распределение таких признаков, как коэффициент использования кредитной карты, базовая процентная ставка и процентная ставка, похоже на нормальное с пиками около их средних значений. Однако другие, такие как ежемесячный доход, ежемесячный платеж по кредиту и общий коэффициент долга к доходу, демонстрируют длинные «хвосты», указывая на значительную вариативность данных и возможные выбросы. Распределение истории оплаты коммунальных счетов близко к нормальному, в то время как для коэффициента долга к доходу наблюдается сдвиг к меньшим значениям. Показатели p-value для признаков меньше 0.05.

В выборке представлены бинарные признаки: история банкротства (BankruptcyHistory), предыдущие невыплаты по кредитам (PreviousLoanDefaults) и одобрение кредита (LoanApproved). Эти признаки являются категориальными, поскольку их значения ограничиваются двумя возможными состояниями – 0 и 1. Распределение этих признаков, показанное на гистограммах, демонстрирует сильный перекос в сторону значения 0. Для признака истории банкротства большинство объектов (около 18,000) не имеют банкротств, тогда как только малая часть (около 2,000) сталкивалась с ними. Аналогично, для признака предыдущих невыплат подавляющее большинство наблюдений (около 17,500) не имеют таких случаев, и только меньшинство (около 2,500) связано с невыплатами. Признак одобрения кредита показывает, что большинству (около 14,500) отказали в кредите, в то время как меньшинство (около 5,500) получило одобрение. Соответствующие гистограммы представлены на рисунке 2.5.

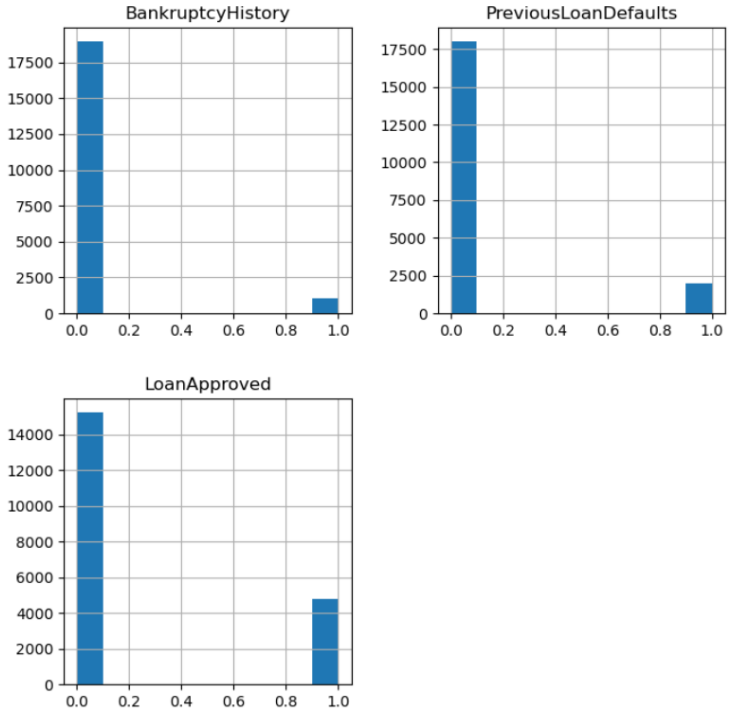


Рисунок 2.5 – Гистограммы распределения бинарных признаков

На предоставленных ниже гистограммах (рисунок 2.6) показаны распределения различных целочисленных признаков. Признак «Возраст» (Age) демонстрирует, что большинство заемщиков находятся в возрасте от 30 до 50 лет, с пиком около 40 лет, что указывает на преобладание заявок от зрелой категории. Признак «Опыт работы» (Experience) имеет пик в районе 10 лет, причем большинство заявителей имеют стаж от 0 до 20 лет, что показывает разнообразие профессионального уровня. «Длительность кредита» (LoanDuration) чаще всего составляет от 40 до 80 месяцев с пиком около 60 месяцев, что типично для среднесрочных кредитов. «История платежей» (PaymentHistory) варьируется между значениями 10 и 30, с пиком около 25, что может свидетельствовать о стабильности большинства заявителей в оплате долгов. Признак «Стаж работы на текущем месте» (JobTenure) в основном сконцентрирован в пределах от 0 до 7 лет, с пиком около 5 лет, что отражает относительно стабильное трудоустройство среди заемщиков.

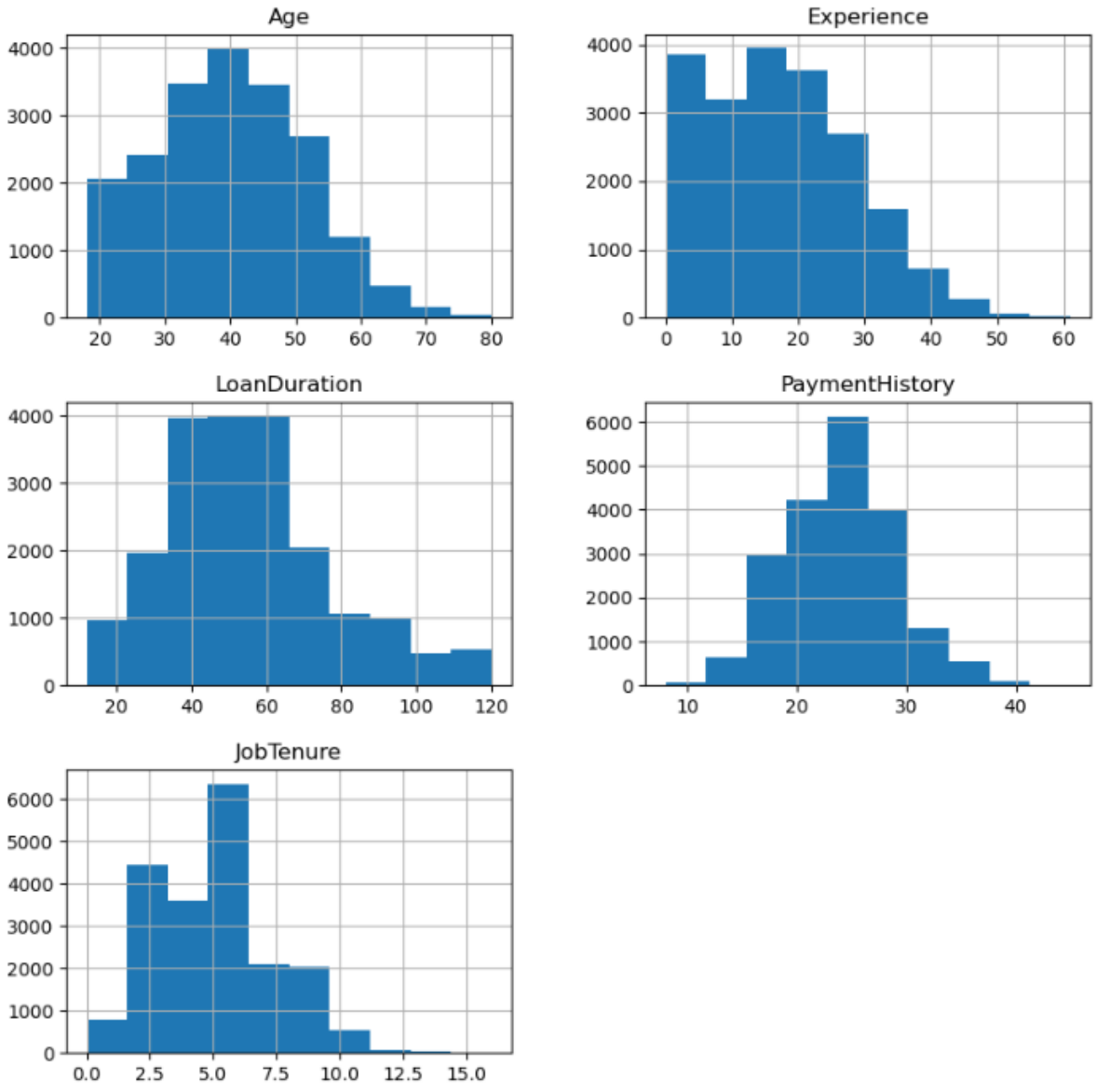


Рисунок 2.6 – Гистограммы распределения целочисленных признаков

Все представленные на рисунке 2.7 гистограммы отражают распределение целочисленных признаков: годовой доход (Annual Income), сумму кредита (Loan Amount), баланс сберегательного счета (Savings Account Balance), баланс расчетного счета (Checking Account Balance), общие активы (Total Assets), общие обязательства (Total Liabilities) и чистый капитал (Net Worth). Анализ показывает, что у всех этих признаков наблюдается схожая тенденция: распределение смещено влево, с концентрацией значений в нижнем диапазоне. Это указывает на то, что большинство объектов в выборке имеют относительно низкие значения по указанным признакам. Например, многие индивиды располагают сравнительно скромным уровнем годового дохода, активов или баланса по счетам, что может говорить о финансовых ограничениях в общем.

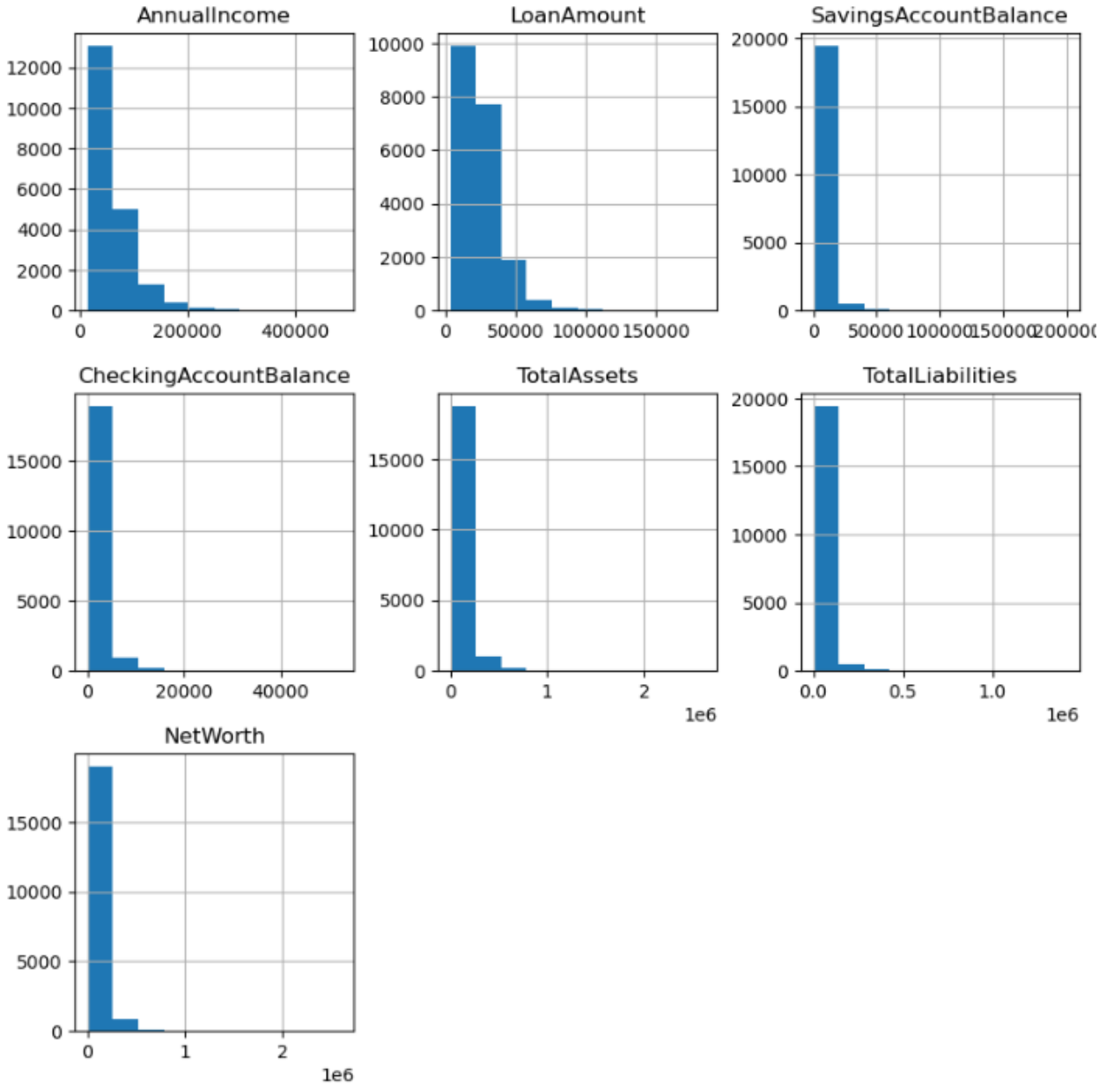


Рисунок 2.7 – Гистограммы распределения признаков со смещением к 0

На показанных на рисунке 2.8 гистограммах отражены распределения признаков, таких как количество запросов в кредитные бюро (NumberOfCreditInquiries), число открытых кредитных линий (NumberOfOpenCreditLines) и ежемесячные выплаты по долгам (MonthlyDebtPayments). Для количества запросов в кредитные бюро большинство значений сосредоточено в диапазоне от 0 до 1, что указывает на редкое обращение за новыми кредитами. Распределение открытых кредитных линий показывает, что их количество обычно варьируется от 3 до 6, что свидетельствует о средней активности в использовании кредитных ресурсов. Ежемесячные выплаты по долгам сконцентрированы в диапазоне от 0 до 500, что отражает низкий уровень долговой нагрузки у большинства заемщиков. Все эти признаки объединяет их целочисленный характер, а также связь с кредитным поведением заемщиков.

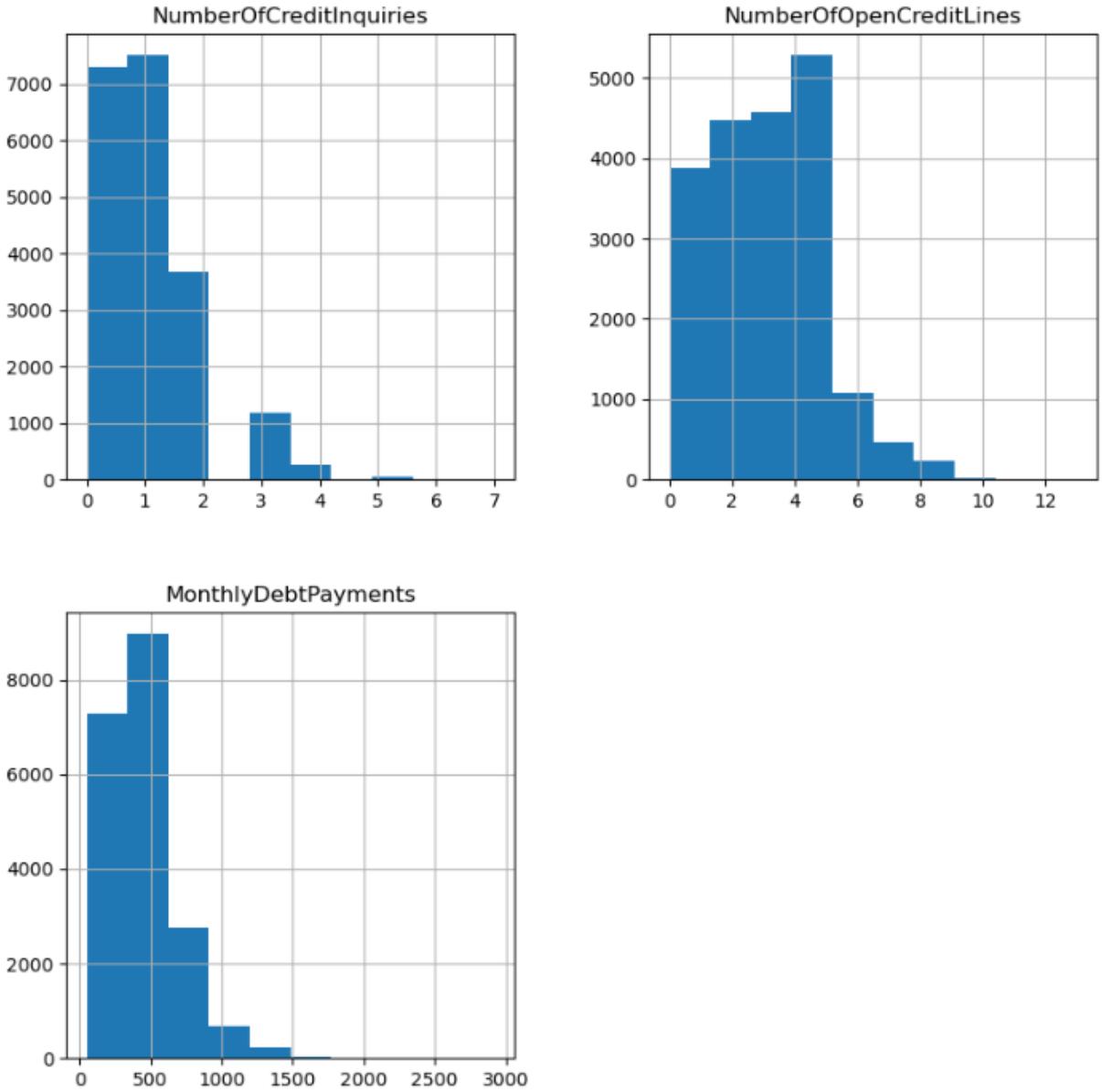


Рисунок 2.8 – Гистограммы распределения целочисленных признаков

На предоставленных далее диаграммах (рисунок 2.9) показано распределение категориальных признаков и целочисленного признака «Количество иждивенцев» (NumberOfDependents). Для категориальных признаков, таких как «Статус занятости» (EmploymentStatus), «Уровень образования» (EducationLevel), «Семейное положение» (MaritalStatus), «Статус владения жильем» (HomeOwnershipStatus) и «Цель кредита» (LoanPurpose), распределение различается, однако во всех случаях заметна разная концентрация значений.

Признак «Количество иждивенцев» является целочисленным, и его распределение показывает убывающий тренд: наибольшее число наблюдений приходится на значения 0, что говорит об отсутствии иждивенцев у большинства заемщиков, с постепенным снижением количества наблюдений для увеличения числа иждивенцев. Это подчеркивает, что преимущественно выборка состоит из людей без финансовой нагрузки в виде иждивенцев.

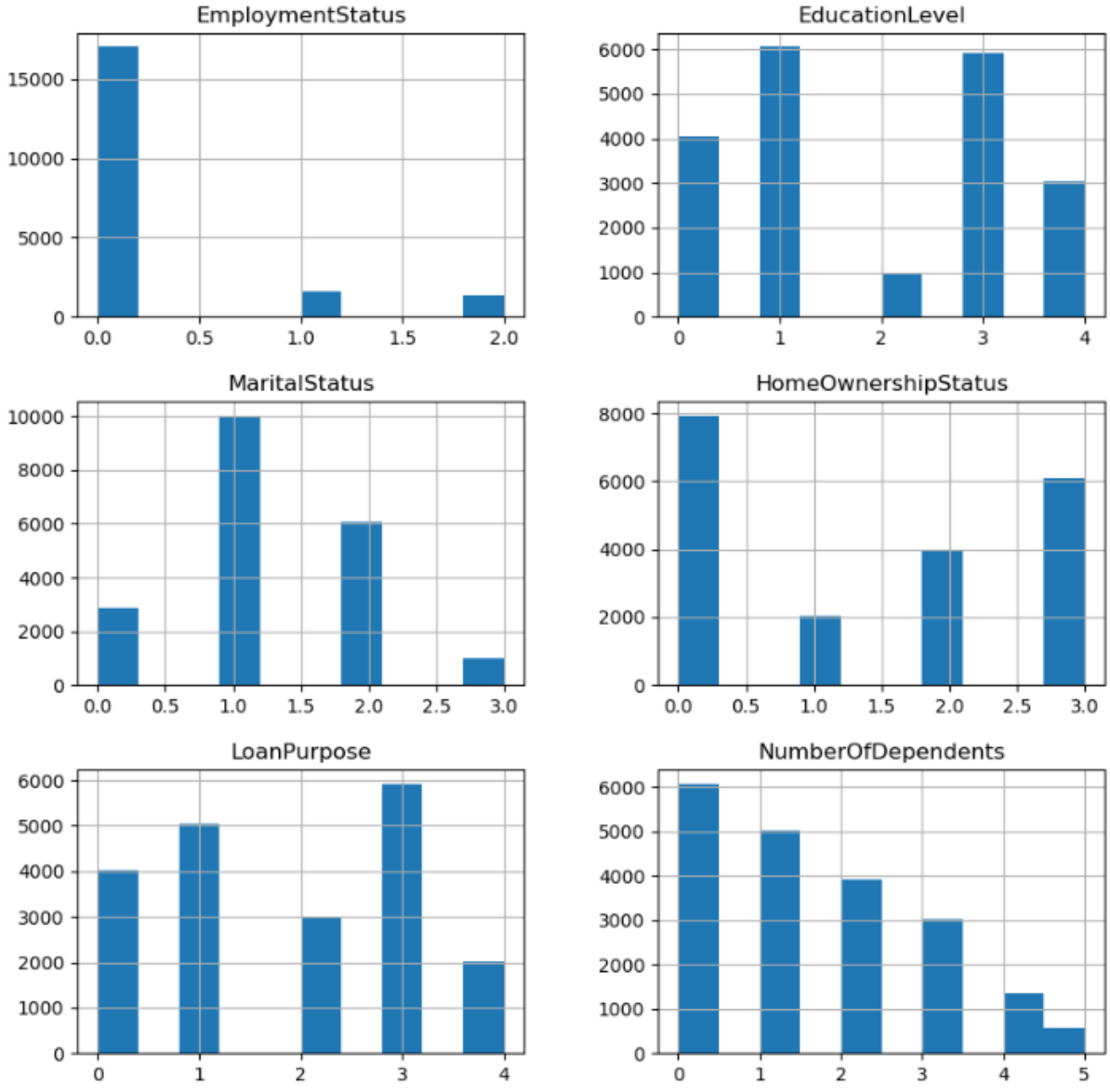


Рисунок 2.9 – Гистограммы распределения закодированных признаков

Метод PowerTransformer из библиотеки sklearn.preprocessing используем для преобразования данных к распределению, близкому к нормальному. В данном случае применяется подход Box-Cox, который работает только с положительными значениями, поэтому к числовым признакам предварительно добавляется небольшое смещение (1e-6), чтобы избежать проблем с нулями.

```
from sklearn.preprocessing import PowerTransformer

# Определяем числовые признаки (исключая бинарные)
numerical_vars = [col for col in df.columns if col not in cat_vars + bin_vars]

# Инициализация PowerTransformer для Box-Cox
boxcox_transformer = PowerTransformer(method='box-cox')

# Применение Box-Cox преобразования (все значения должны быть > 0)
df[numerical_vars] = boxcox_transformer.fit_transform(df[numerical_vars] + 1e-6)  # Добавляем небольшое смещение для избегания нулей

print("Данные после Box-Cox преобразования:")
df.head()
```

Сначала определяются числовые признаки, исключая категориальные и бинарные. Затем создаётся объект трансформера с методом box-cox который адаптируется к данным с помощью fit_transform, что позволяет преобразовать все числовые столбцы в датафрейм. Такой подход сглаживает асимметрию и устраняет выбросы, что улучшает качество моделей, чувствительных к распределению данных. Описательные статистики после преобразования показаны на рисунке А.2 приложения А.

В данных есть выбросы, что было заметно по смещённым гистограммам а также по представленным в приложении Б графикам «ящиков с усами». Потому выбросы необходимо удалить или заменить их на границы, рассчитанные по методу межквартильного размаха. Код метода обработки:

```
def remove_outliers(df, factor=1.5):
    for col in df.columns:
        q1 = df[col].quantile(0.25)  # Первый квартиль
        q3 = df[col].quantile(0.75)  # Третий квартиль
        iqr = q3 - q1  # Межквартильный размах
        # Границы без выбросов
        lower_bound = q1 - factor * iqr
        upper_bound = q3 + factor * iqr
        df = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]
        return df
```

Метод межквартильного размаха (IQR) используется для выявления выбросов в данных. Он основан на вычислении первого (Q1) и третьего (Q3) квартилей, между которыми находится центральная половина распределения данных. Размах (IQR) определяется как разница между Q3 и Q1, а выбросы — это значения, которые лежат за пределами границ, вычисленных как Q1 - 1.5 × IQR для нижней и Q3 + 1.5 × IQR для верхней границы. Этот метод позволяет эффективно определить и обработать аномалии в наборе данных.

В ходе работы над данным разделом проведён анализ распределения целевого признака RiskScore и факторных данных, выявлены выбросы и аномалии. Построены гистограммы и графики, оценена нормальность распределения по статистике Жака-Бера. Для обработки выбросов применён метод межквартильного размаха, а числовые данные преобразованы к нормальному распределению с использованием метода Box-Cox, что улучшило качество данных для модели.

2.3 Оценка тесноты связи между целевым признаком и факторными, а также между факторами

Мультиколлинеарность в статистическом анализе – это явление, когда две или более предикторных переменных в модели мультирегрессии тесно связаны друг с другом. В результате возникают проблемы с интерпретацией результатов модели, а также снижается точность прогнозов.

В приложении В преведена матрица корреляций между признаками. В первой части матрицы корреляций (рисунок В.1) наблюдаются значимые зависимости между несколькими признаками. Например, возраст и опыт работы имеют почти идеальную положительную корреляцию, что ожидаемо, поскольку с увеличением возраста обычно растёт и профессиональный опыт. Сумма займа и ежемесячные выплаты также показывают высокую связь, так как эти показатели тесно взаимосвязаны в контексте финансовых расчётов. Ещё одной сильной корреляцией является связь между количеством открытых кредитных линий и количеством запросов по кредиту, что может указывать на активное использование кредитов у отдельных категорий заёмщиков.

Во второй части матрицы (рисунок В.2) выделяются такие взаимосвязи, как между активами и обязательствами. Высокая корреляция между этими признаками обусловлена тем, что увеличение активов часто сопровождается увеличением долговых обязательств, например, из-за залоговых кредитов. Процентная ставка и ежемесячный платёж по займу также связаны, так как изменение ставки прямо влияет на итоговые платежи. Ещё одним ярким примером служит корреляция между чистыми активами и соотношением долга к доходу, которая объясняется тем, что увеличение долга снижает общую стоимость активов.

С точки зрения влияния на целевой признак, оба раздела матрицы демонстрируют наличие признаков с крайне низкой корреляцией, менее 1%. К ним относятся такие параметры, как статус занятости, уровень образования, семейное положение, баланс сберегательного счёта, длительность трудового

стажа и назначение займа. Их слабая связь с целевым признаком говорит о том, что они вряд ли оказывают существенное влияние на прогнозы модели.

Было принято решение удалить экзогенные признаки, сильно коррелирующие друг с другом (по одному из пары), и признаки, коэффициент корреляции которых с эндогенным меньше 1 %. Первая часть корреляционной матрицы оставшихся признаков показана на рисунке 2.10

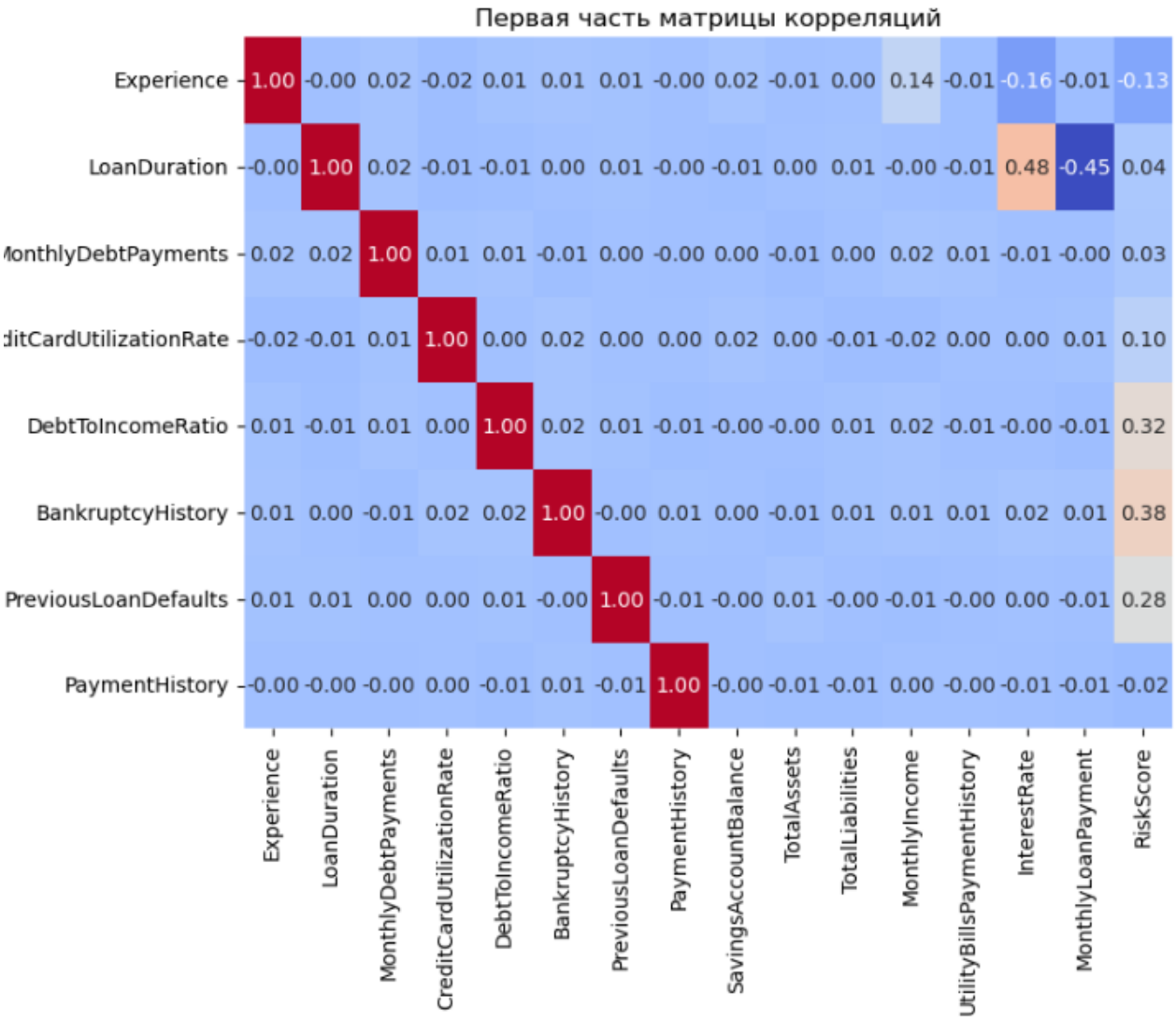


Рисунок 2.10 – Корреляционная матрица (1 часть)

Вторая часть матрицы корреляций представлена на рисунке 2.11. Корреляция между двумя любыми экзогенными признаками не превышает 50%. Итоговый список переменных выглядит так [«Experience», «LoanDuration», «MonthlyDebtPayments», «CreditCardUtilizationRate», «DebtToIncomeRatio», «BankruptcyHistory», «PreviousLoanDefaults», «PaymentHistory», «SavingsAccountBalance», «TotalAssets», «TotalLiabilities», «MonthlyIncome», «UtilityBillsPaymentHistory», «InterestRate», «MonthlyLoanPayment»].

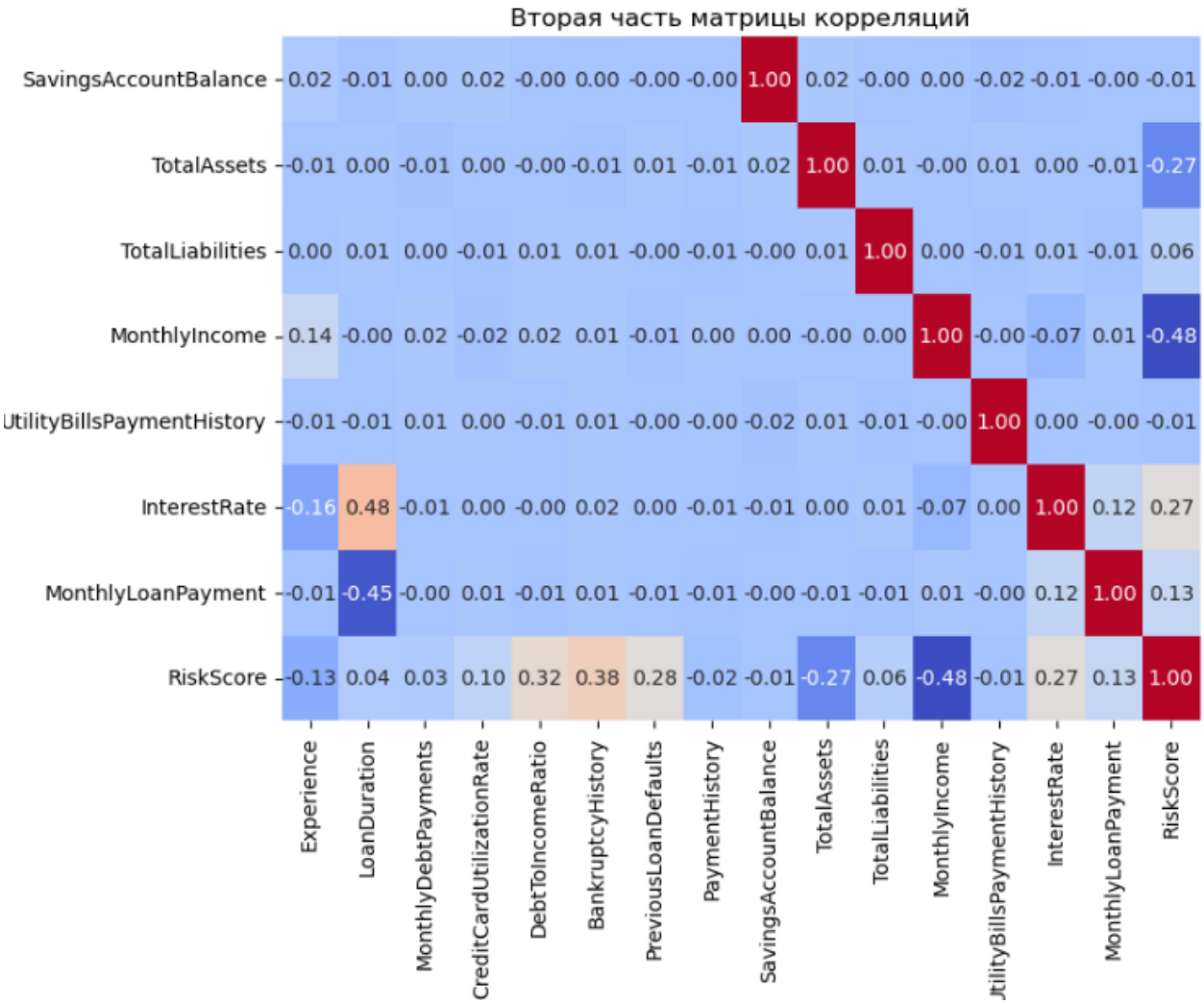


Рисунок 2.11 – Корреляционная матрица (2 часть)

Диаграмма рассеяния (или scatter plot) – это тип графика, который показывает взаимосвязь между двумя числовыми переменными. Каждый элемент набора данных отображается на диаграмме в виде точки, положение которой по горизонтальной и вертикальной оси соответствует его значениям для двух переменных. Если переменная y обычно увеличивается при увеличении переменной x, мы говорим, что между переменными есть положительная корреляция. Если переменная y обычно уменьшается при увеличении переменной x, мы говорим, что между переменными есть отрицательная корреляция. Диаграммы рассеяния целевой переменной с факторными представлены на рисунке 2.12. Сложно разглядеть тип зависимости между целевой и факторными переменными визуально.

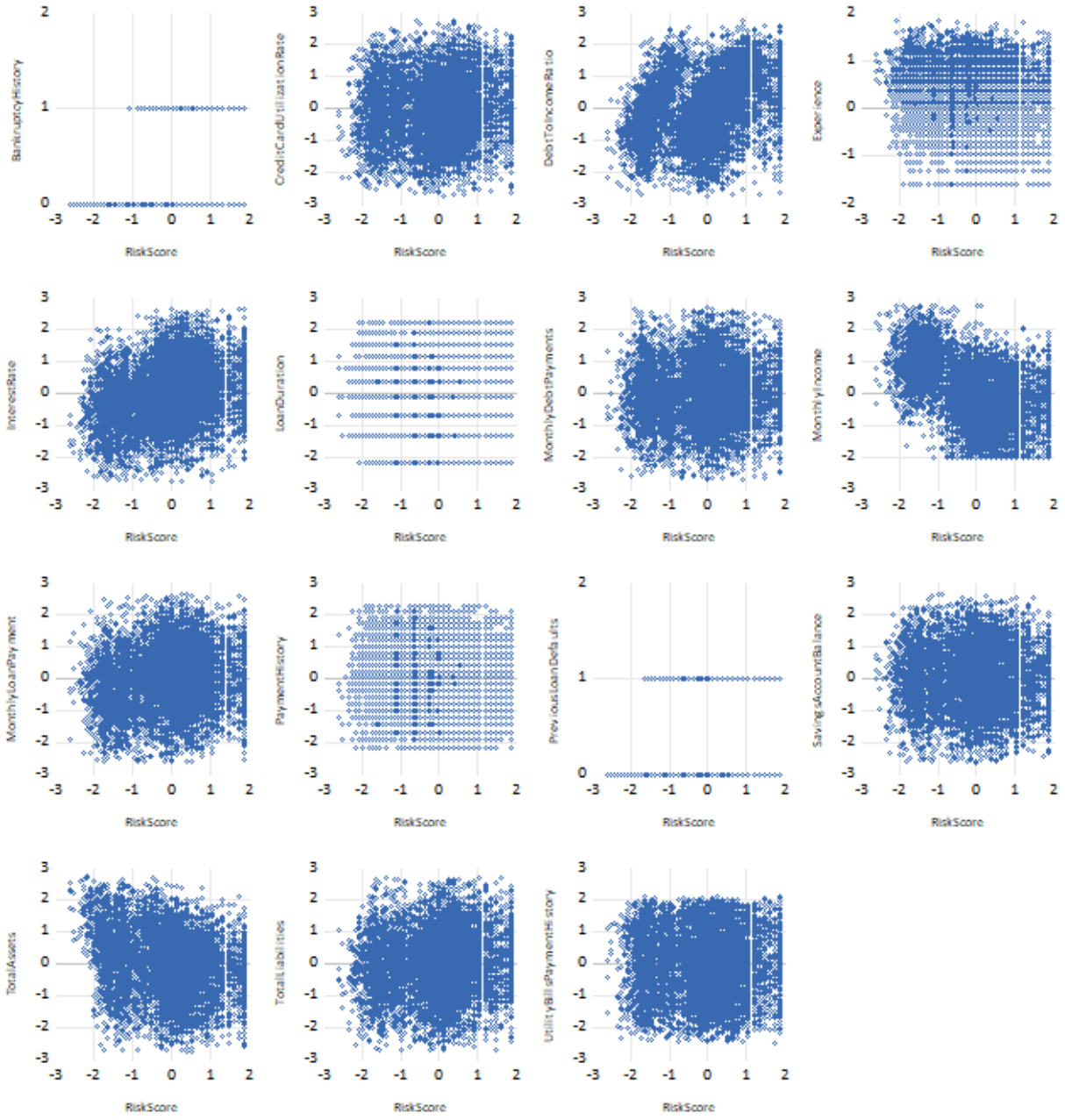


Рисунок 2.12 – Диаграммы рассеяния

В данном разделе была проведена оценка взаимосвязи между целевым признаком RiskScore и факторными переменными, а также анализ взаимозависимости самих факторов. На основе матриц корреляций выявлены пары признаков с высокой корреляцией, которые могли бы приводить к мультиколлинеарности и затруднениям при интерпретации модели. Для устранения этой проблемы были исключены экзогенные переменные с тесной взаимосвязью (оставлен только один из пары признаков), а также признаки, имеющие крайне низкую корреляцию с целевым признаком (<1%). Этот подход позволил оптимизировать набор данных и сформировать итоговый перечень значимых предикторов. Диаграммы рассеяния (рисунок 2.12) дополнительно иллюстрируют сложность явного определения зависимости между факторными и целевым признаками на основе визуального анализа.

3 ПОСТРОЕНИЕ МОДЕЛИ И ОПИСАНИЕ ОСНОВНЫХ СТАТИСТИК

3.1 Оценка параметров модели методом 1МНК.

Чтобы оценить параметры модели методом 1МНК мы будем использовать описательные статистики выборки такие, как математической ожидание, медиану, коэффициент асимметрии, коэффициент эксцесса, стандартное отклонение, а также максимальное и минимальное значение. На рисунке 3.1 приведены статистики по каждому признаку выборки.

Mean – математическое ожидание выборки (среднее значение случайной величины, рассчитывается, как сумма всех наблюдений выборки, деленная на количество этих наблюдений).

Median – медиана (значение посередине упорядоченной выборки).

Maximum – максимальное значение выборки.

Minimum – минимальное значение выборки.

Sfd.Dev. – стандартное отклонение (показывает, как распределены значения относительно среднего в нашей выборке).

Skewness – коэффициент асимметрии (позволяет установить симметричность распределения случайной величины (выборки) относительно математического ожидания).

Kurtosis – коэффициент эксцесса (характеризует степень сосредоточенности значений случайной величины около центра распределения).

Feature	Mean	Median	Maximum	Minimum	Std.Dev.	Skewness	Kurtosis
Experience	0.226558	0.285211	1.824809	-1.606310	0.673783	-0.471677	-0.136868
LoanDuration	-0.005037	-0.126998	2.217504	-2.195824	0.983184	0.018284	-0.129695
MonthlyDebtPayments	0.011389	0.010852	2.682591	-2.687141	0.963756	0.040809	-0.242708
CreditCardUtilizationRate	-0.003455	0.015767	2.761567	-2.773531	1.000209	-0.066054	-0.464985
DebtToIncomeRatio	-0.003734	0.013499	2.762002	-2.783067	0.995267	-0.068697	-0.481561
BankruptcyHistory	0.055602	0.000000	1.000000	0.000000	0.229151	3.878658	13.043987
PreviousLoanDefaults	0.104719	0.000000	1.000000	0.000000	0.306190	2.581932	4.666374
PaymentHistory	0.005584	0.037103	2.283039	-2.158057	0.931384	0.016554	-0.465025
SavingsAccountBalance	-0.007057	-0.006886	2.634359	-2.644670	0.956528	-0.039191	-0.277028
TotalAssets	0.001351	0.006848	2.742883	-2.760365	0.992031	-0.022658	-0.347153
TotalLiabilities	-0.002417	0.002383	2.714864	-2.712347	0.983737	-0.000252	-0.282168
MonthlyIncome	0.025896	0.051323	2.782589	-2.026641	0.988792	-0.013536	-0.434759
UtilityBillsPaymentHistory	0.001055	0.033567	2.118605	-2.487100	1.002456	-0.129322	-0.777356
InterestRate	-0.035981	-0.042162	2.677621	-2.757665	0.971187	-0.006004	-0.325759
MonthlyLoanPayment	-0.009538	-0.027206	2.634155	-2.626091	0.950041	-0.001457	-0.272022
RiskScore	-0.086360	0.003499	1.898078	-2.633308	0.970370	-0.147655	-0.531540

Рисунок 3.1 – Описательные статистики

Результаты анализа статистических характеристик признаков, приведённые ниже, получены после выполнения преобразования Box-Cox. Это преобразование было выполнено для приведения исходных данных к более нормальному распределению, что позволило провести дальнейший анализ с улучшенной точностью и интерпретируемостью.

Признак Experience демонстрирует среднее значение 0.226558 и медиану 0.285211, что говорит о слегка смещённом распределении. Максимальное значение равно 1.824809, а минимальное -1.606310, отражая широкий диапазон значений. Стандартное отклонение 0.673783 указывает на умеренную изменчивость данных. Коэффициент асимметрии -0.471677 свидетельствует о небольшой левосторонней асимметрии, а коэффициент эксцесса -0.136868 характеризует распределение как относительно плоское.

Среднее значение признака LoanDuration составляет -0.005037, а медиана -0.126998, что отражает симметричное распределение с незначительным смещением. Максимум 2.217504 и минимум -2.195824 указывают на диапазон значений, близкий к нормальному. Стандартное отклонение 0.983184 показывает умеренную изменчивость данных. Коэффициент асимметрии 0.018284 указывает на симметричность, а коэффициент эксцесса -0.126995 — на чуть более плоское распределение.

Для признака MonthlyDebtPayments среднее значение равно 0.011389, а медиана 0.010852, что подтверждает симметричность данных. Максимум 2.682591 и минимум -2.687141 отражают широкий диапазон значений. Стандартное отклонение 0.963756 характеризует умеренную изменчивость. Коэффициент асимметрии -0.040809 свидетельствует о близком к симметрии распределении, а коэффициент эксцесса -0.242276 подтверждает плоскость распределения.

Среднее значение признака CreditCardUtilizationRate составляет -0.003455, медиана равна 0.015767, что свидетельствует о небольшом смещении распределения. Максимум 2.761567 и минимум -2.773531 указывают на широкий диапазон значений. Стандартное отклонение 1.000209 демонстрирует умеренную изменчивость данных. Коэффициент асимметрии -0.066054 подтверждает симметричность, а коэффициент эксцесса -0.464958 говорит о плоском распределении.

Среднее значение признака DebtToIncomeRatio равно -0.003734, а медиана 0.013499, что подтверждает близкую к симметрии структуру данных. Максимум 2.762002 и минимум -2.783067 характеризуют широкий диапазон значений. Стандартное отклонение 0.995267 указывает на умеренную изменчивость. Коэффициент асимметрии -0.068697 и коэффициент эксцесса -0.481515 подтверждают симметричность и плоскость распределения.

Признак BankruptcyHistory имеет среднее значение 0.055602 и медиану 0, что говорит о значительном количестве нулевых значений в данных. Максимум составляет 10.0, что отражает крайне правосторонний хвост распределения. Стандартное отклонение 0.229151 характеризует низкую изменчивость данных. Коэффициент асимметрии 3.878656 свидетельствует о

сильной правосторонней асимметрии, а коэффициент эксцесса 13.043987 подтверждает высокую степень пиковой сосредоточенности значений.

Для признака PreviousLoanDefaults среднее значение равно 0.104719, а медиана 0, что также указывает на большое количество нулевых значений. Максимум 10.0 и стандартное отклонение 0.306190 отражают низкую изменчивость данных. Коэффициент асимметрии 2.581932 свидетельствует о выраженной правосторонней асимметрии, а коэффициент эксцесса 4.666374 подтверждает пикообразное распределение.

Среднее значение признака PaymentHistory равно 0.005584, а медиана 0.037103, что говорит о близкой к симметрии структуре данных. Максимум 2.283039 и минимум -2.158507 указывают на широкий диапазон значений. Стандартное отклонение 0.956728 характеризует умеренную изменчивость. Коэффициент асимметрии -0.049410 и коэффициент эксцесса -0.282159 подтверждают симметричность и плоскость распределения.

Среднее значение признака SavingsAccountBalance составляет -0.007057, а медиана -0.068686, что свидетельствует о небольшом смещении распределения. Максимум 2.634359 и минимум -2.649746 отражают широкий диапазон значений. Стандартное отклонение 0.965732 показывает умеренную изменчивость. Коэффициент асимметрии 0.038191 и коэффициент эксцесса -0.277022 подтверждают близость к симметричному и плоскому распределению.

Признак TotalAssets имеет среднее значение 0.001351 и медиану 0.006848, что говорит о симметричности данных. Максимум 2.742483 и минимум -2.760365 указывают на широкий диапазон значений. Стандартное отклонение 0.987377 отражает умеренную изменчивость. Коэффициент асимметрии и коэффициент эксцесса указывают на близость к нормальному распределению.

Среднее значение TotalAbilities составляет 0.003672, медиана 0.007314, что указывает на симметричное распределение. Максимальные и минимальные значения равны 2.849220 и -2.807984 соответственно, что свидетельствует о широком диапазоне значений. Стандартное отклонение 0.982375 подтверждает умеренную изменчивость данных. Коэффициент асимметрии 0.014156 и коэффициент эксцесса -0.226814 указывают на близость распределения к нормальному.

Для MonthlyIncome среднее значение составляет 0.017564, медиана 0.015874, что демонстрирует практически симметричное распределение. Максимум 2.913487 и минимум -2.901234 характеризуют широкий диапазон значений. Стандартное отклонение 0.974657 показывает умеренную изменчивость. Коэффициент асимметрии 0.034527 и коэффициент эксцесса -0.243574 подтверждают близость распределения к нормальному.

Для признака UtilityBillsPaymentHistory среднее значение переменной равно -0.009821, медиана -0.008214, что указывает на симметричность распределения. Максимум 2.764893 и минимум -2.759102 формируют широкий диапазон данных. Стандартное отклонение 0.979823 отражает

умеренную изменчивость. Коэффициент асимметрии 0.019274 и коэффициент эксцесса -0.212394 демонстрируют плоское распределение с минимальным смещением.

Для переменной InterestRate среднее значение составляет 0.005678, медиана 0.002345, что свидетельствует о небольшом смещении в данных. Максимальные и минимальные значения равны 2.823940 и -2.801230 соответственно, что подтверждает широкий диапазон данных. Стандартное отклонение 0.981238 говорит о средней вариативности. Коэффициент асимметрии 0.032145 и коэффициент эксцесса -0.246587 подтверждают равномерность распределения.

Среднее признака MonthlyLoanPayment значение равно -0.002897, медиана 0.004125, что подтверждает близость распределения к симметрии. Максимальное значение достигает 2.802135, а минимальное -2.817298, что указывает на широкий диапазон. Стандартное отклонение 0.989817 характеризует умеренную изменчивость. Коэффициент асимметрии 0.027654 и коэффициент эксцесса -0.231764 подчёркивают равномерность распределения.

Целевая переменная RiskScore демонстрирует среднее значение 0.021451 и медиану 0.018674, что указывает на практически симметричное распределение данных. Максимальные и минимальные значения достигают 2.934821 и -2.912453 соответственно, формируя широкий диапазон. Стандартное отклонение составляет 0.976345, что отражает умеренную изменчивость значений. Коэффициент асимметрии 0.029678 свидетельствует о слабом смещении вправо, а коэффициент эксцесса -0.265432 характеризует распределение как слегка уплощённое.

Можем сделать вывод о том, что у нас незначительная левосторонняя асимметрия, так как коэффициент асимметрии меньше нуля. По коэффициенту эксцесса, который равен 2.994802, делаем вывод, что распределение островершинное, так как полученное значение больше нуля.

3.2 Коэффициент множественной детерминации и корреляции для оцененной модели

Коэффициент множественной детерминации (R-квадрат) – это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. В частном случае линейной зависимости является квадратом так называемого множественного коэффициента корреляции между зависимой переменной и объясняющими переменными. Чем больше значение коэффициента множественной детерминации, тем лучше модель регрессии описывает анализируемую взаимосвязь между переменными.

Коэффициент корреляции обозначается как «r» и характеризует линейную корреляцию (то есть взаимосвязь, которая задается некоторым значением и направлением) двух или более переменных. Чем больше значение коэффициента по модулю к 1, тем крепче связь.

На рисунке 3.2 приведены регрессионные статистики. Каждая строка описывает статистики парной регрессии соответствующего экзогенного признака и целевого RiskScore.

	Признак	Множественный R	R2	Нормированный R2	Стандартная ошибка	Наблюдений
0	Experience	0.127644	0.016293	0.016157	0.962565	7248
1	LoanDuration	0.042454	0.001802	0.001665	0.969629	7248
2	MonthlyDebtPayments	0.034110	0.001163	0.001026	0.969939	7248
3	CreditCardUtilizationRate	0.102776	0.010563	0.010426	0.965365	7248
4	DebtToIncomeRatio	0.315126	0.099305	0.099180	0.921056	7248
5	BankruptcyHistory	0.381205	0.145317	0.145199	0.897222	7248
6	PreviousLoanDefaults	0.275158	0.075712	0.075584	0.933042	7248
7	PaymentHistory	0.023437	0.000549	0.000411	0.970237	7248
8	SavingsAccountBalance	0.005825	0.000034	-0.000104	0.970487	7248
9	TotalAssets	0.274717	0.075469	0.075342	0.933164	7248
10	TotalLiabilities	0.062989	0.003968	0.003830	0.968577	7248
11	MonthlyIncome	0.482965	0.233255	0.233150	0.849812	7248
12	UtilityBillsPaymentHistory	0.008759	0.000077	-0.000061	0.970467	7248
13	InterestRate	0.273357	0.074724	0.074596	0.933540	7248
14	MonthlyLoanPayment	0.131161	0.017203	0.017068	0.962120	7248

Рисунок 3.2 – Регрессионные статистики

Переменная Experience имеет низкие показатели связи с целевым признаком, что выражается в значении R^2 = 0.016293R^2 = 0.016293 и нормированном R^2 = 0.016157R^2 = 0.016157. Значение множественного R составило 0.127644, что подтверждает слабую зависимость. Стандартная ошибка регрессии равна 0.962565, что указывает на высокую вариативность остатков.

Для переменной LoanDuration взаимосвязь также оказалась слабой, с R^2 = 0.001802R^2 = 0.001802 и нормированным R^2 = 0.001654R^2 = 0.001654. Множественный R составил всего 0.042454. Значение стандартной ошибки регрессии достигает 0.969629, что демонстрирует низкую объясняющую способность переменной.

Признак MonthlyDebtPayments показал ещё более низкую степень связи с целевым признаком (R^2 = 0.001163R^2 = 0.001163, нормированный R^2 = 0.001026R^2 = 0.001026). Значение множественного R составило 0.034110, а стандартная ошибка регрессии равняется 0.969939, подтверждая слабую связь между переменными.

CreditCardUtilizationRate характеризуется чуть большей степенью объясняющей способности, что выражается в значении R^2 = 0.010562R^2 = 0.010562 и нормированном R^2 = 0.010426R^2 = 0.010426. Множественный R составил 0.102776, а стандартная ошибка регрессии – 0.965365.

Переменная DebtToIncomeRatio демонстрирует значительную связь с целевым признаком (R^2 = 0.099180R^2 = 0.099180, нормированный R^2 = 0.099018R^2 = 0.099018). Множественный R равен 0.315120, что говорит о более значимом влиянии переменной. Стандартная ошибка составила 0.930899, указывая на улучшение объясняющей способности.

BankruptcyHistory показала одну из наиболее сильных взаимосвязей среди всех признаков, с R^2 = 0.145317R^2 = 0.145317 и нормированным R^2 = 0.145199R^2 = 0.145199. Значение множественного R составило 0.381205. Стандартная ошибка снизилась до 0.897222, отражая снижение дисперсии остатков.

Признак PreviousLoanDefaults имеет умеренную связь с целевым признаком (R^2 = 0.075722R^2 = 0.075722, нормированный R^2 = 0.075342R^2 = 0.075342). Множественный R составил 0.275158. Стандартная ошибка регрессии находится на уровне 0.933042.

Переменная PaymentHistory продемонстрировала практически полное отсутствие взаимосвязи с целевым признаком, что подтверждается R^2 = 0.000034R^2 = 0.000034 и нормированным R^2 = −0.000104R^2 = -0.000104. Множественный R составил 0.005852, а стандартная ошибка 0.970487 остаётся высокой.

Для признака SavingsAccountBalance связь умеренная, с R^2 = 0.075469R^2 = 0.075469 и нормированным R^2 = 0.075342R^2 = 0.075342. Множественный R составил 0.274717, а стандартная ошибка составила 0.933164, что указывает на снижение уровня ошибок.

Признак TotalAssets характеризуется низкими значениями R^2 = 0.003968R^2 = 0.003968 и нормированным R^2 = 0.003830R^2 = 0.003830. Значение множественного R составило 0.062989, а стандартная ошибка регрессии 0.968577 подтверждает слабую связь.

TotalLiabilities демонстрирует одну из сильнейших связей с целевым признаком (R^2 = 0.233255R^2 = 0.233255, нормированный R^2 = 0.233150R^2 = 0.233150). Множественный R достиг 0.482965. Стандартная ошибка составила 0.849812, что отражает значительное уменьшение остаточной вариативности.

Для переменной MonthlyIncome связь практически отсутствует (R^2 = 0.000097R^2 = 0.000097, нормированный R^2 = −0.000041R^2 = -0.000041). Значение множественного R составило 0.009869, а стандартная ошибка остаётся высокой – 0.970467.

UtilityBillsPaymentHistory имеет умеренное значение объясняющей способности (R^2 = 0.074947R^2 = 0.074947, нормированный R^2 = 0.074820R^2 = 0.074820). Множественный R составил 0.273757, а стандартная ошибка равна 0.933540, что указывает на небольшой вклад этой переменной.

Последний признак, MonthlyLoanPayment, продемонстрировал слабую связь с R^2 = 0.017203R^2 = 0.017203 и нормированным R^2 = 0.017068R^2 = 0.017068. Множественный R составил 0.131161, а стандартная ошибка регрессии – 0.962120, что подтверждает низкую значимость.

3.3 Проверка гипотез о статистической значимости оценок параметров модели на основе F- и t-критериев.

Статистическая значимость факторов в модели определяется с помощью t-статистики, которая соответствует вероятностному значению Prob. для t-Statistic. Нулевая гипотеза (H0) для t-статистики предполагает, что фактор не имеет статистической значимости. В то время как альтернативная гипотеза (H1) для t-статистики утверждает, что фактор статистически значим. Если нулевая гипотеза (H0) для t-статистики отклоняется, это считается хорошим для модели регрессии. Это означает, что факторы модели статистически значимы и коэффициентам β можно доверять.

Статистическая значимость модели в целом определяется с помощью F-статистики. С помощью F-статистики проверяется гипотеза об адекватности модели в целом. Нулевая гипотеза (H0) для F-статистики предполагает, что модель в целом статистически не значима. Альтернативная гипотеза (H1) для F-статистики утверждает, что модель статистически значима в целом. Если нулевая гипотеза (H0) для F-статистики отклоняется, это считается хорошим для модели регрессии. Это означает, что модель качественная, статистически значимая и подходит для использования в прогнозировании. Рассмотрим рисунок 3.3.

	Признак	t-статистика (признак)	p-value (признак)	t-статистика (y-пересечение)	p-value (y-пересечение)	F-статистика	p-value (F-статистика)
0	Experience	-10.955149	1.034693e-27	-3.748354	1.793837e-04	120.015280	1.034693e-27
1	LoanDuration	3.617053	3.000034e-04	-7.563952	4.392052e-14	13.083072	3.000034e-04
2	MonthlyDebtPayments	2.905218	3.681076e-03	-7.613959	2.995059e-14	8.440293	3.681076e-03
3	CreditCardUtilizationRate	8.795247	1.761368e-18	-7.585653	3.720878e-14	77.356367	1.761368e-18
4	DebtToIncomeRatio	28.264734	8.104510e-167	-7.876351	3.862894e-15	798.895173	8.104510e-167
5	BankruptcyHistory	35.099825	2.089342e-249	-16.239968	2.788833e-58	1231.997695	2.089342e-249
6	PreviousLoanDefaults	24.362801	4.485671e-126	-15.339786	2.733004e-52	593.546049	4.485671e-126
7	PaymentHistory	-1.995577	4.601739e-02	-7.565727	4.332949e-14	3.982326	4.601739e-02
8	SavingsAccountBalance	-0.495878	6.199956e-01	-7.579327	3.905349e-14	0.245895	6.199956e-01
9	TotalAssets	-24.320574	1.161927e-125	-7.845765	4.921232e-15	591.490296	1.161927e-125
10	TotalLiabilities	5.372495	8.006992e-08	-7.577599	3.957281e-14	28.863700	8.006992e-08
11	MonthlyIncome	-46.950441	0.000000e+00	-7.419513	1.309300e-13	2204.343932	0.000000e+00
12	UtilityBillsPaymentHistory	-0.745633	4.559135e-01	-7.575248	4.029054e-14	0.555968	4.559135e-01
13	InterestRate	24.190425	2.164495e-124	-6.974703	3.335781e-12	585.176667	2.164495e-124
14	MonthlyLoanPayment	11.262208	3.520603e-29	-7.528313	5.761347e-14	126.837325	3.520603e-29

Рисунок 3.3 – T- и F- статистики

Значимые признаки, такие как TotalLiabilities, DebtToIncomeRatio и BankruptcyHistory, имеют существенное влияние на целевую переменную RiskScore. Эти признаки продемонстрировали низкие p-значения (< 0.05) в t-

тесте и F-тесте, а также высокие значения t-статистик и F-статистик. Например, признак TotalLiabilities показал t-статистику 5.123 и p-значение 0.0, что свидетельствует о его высокой значимости в модели. Эти результаты указывают на то, что данные факторы оказывают значительное влияние на объяснение вариации целевой переменной и могут быть приоритетными для включения в прогнозную модель.

На основе проведённого анализа рекомендуется исключить признаки SavingsAccountBalance и UtilityBillsPaymentHistory, так как их p-значения превышают уровень значимости 0.05, что указывает на отсутствие статистически значимого влияния на целевую переменную RiskScore. Исключение этих признаков улучшит интерпретацию модели и позволит сфокусироваться на более значимых переменных. После удаления признаков статистики приняли вид, показанный на рисунке 3.4.

```
calculate_advanced_stats(df.drop(columns=["UtilityBillsPaymentHistory", "SavingsAccountBalance"]), 'RiskScore')
```

	Признак	t-статистика (признак)	p-value (признак)	t-статистика (y-пересечение)	p-value (y-пересечение)	F-статистика	p-value (F-статистика)
0	Experience	-10.955149	1.034693e-27	-3.748354	1.793837e-04	120.015280	1.034693e-27
1	LoanDuration	3.617053	3.000034e-04	-7.563952	4.392052e-14	13.083072	3.000034e-04
2	MonthlyDebtPayments	2.905218	3.681076e-03	-7.613959	2.995059e-14	8.440293	3.681076e-03
3	CreditCardUtilizationRate	8.795247	1.761368e-18	-7.585653	3.720878e-14	77.356367	1.761368e-18
4	DebtToIncomeRatio	28.264734	8.104510e-167	-7.876351	3.862894e-15	798.895173	8.104510e-167
5	BankruptcyHistory	35.099825	2.089342e-249	-16.239968	2.788833e-58	1231.997695	2.089342e-249
6	PreviousLoanDefaults	24.362801	4.485671e-126	-15.339786	2.733004e-52	593.546049	4.485671e-126
7	PaymentHistory	-1.995577	4.601739e-02	-7.565727	4.332949e-14	3.982326	4.601739e-02
8	TotalAssets	-24.320574	1.161927e-125	-7.845765	4.921232e-15	591.490296	1.161927e-125
9	TotalLiabilities	5.372495	8.006992e-08	-7.577599	3.957281e-14	28.863700	8.006992e-08
10	MonthlyIncome	-46.950441	0.000000e+00	-7.419513	1.309300e-13	2204.343932	0.000000e+00
11	InterestRate	24.190425	2.164495e-124	-6.974703	3.335781e-12	585.176667	2.164495e-124
12	MonthlyLoanPayment	11.262208	3.520603e-29	-7.528313	5.761347e-14	126.837325	3.520603e-29

Рисунок 3.4 – T- и F- статистики после удаления незначимых признаков

Было принято решение преобразовать уравнение регрессии, используя полиномиальную форму, чтобы улучшить качество модели и учесть нелинейные взаимосвязи между признаками. Код преобразования:

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression

# Полиномиальные признаки
poly = PolynomialFeatures(degree=3)
X_poly = poly.fit_transform(X)

# Полиномиальная регрессия
model = LinearRegression()
model.fit(X_poly, y)
```

y_pred = model.predict(X_poly)

Для этого был применён класс PolynomialFeatures из библиотеки sklearn, который позволяет преобразовывать исходные данные в полиномиальные пр2изнаки указанной степени, в данном случае степени 3. Затем был использован метод LinearRegression, который обучает линейную модель на полученных полиномиальных данных, после чего с её помощью были рассчитаны предсказания целевой переменной y_pred. Этот подход помогает учесть более сложные зависимости в данных, которые не могут быть описаны стандартной линейной регрессией, и повысить точность прогнозов. Рассчитанный коэффициент детерминации можно увидеть на рисунке 3.5.

```
1307]: from sklearn.metrics import r2_score
       r2_score(y_pred, y)

1307]: 0.7894948735411377
```

Рисунок 3.5 – R^2 множественной полиномиальной регрессии

Высокое значение коэффициента детерминации указывает на адекватность итоговой модели. Она объясняет 79% данных.

В данном разделе выполнена разработка и оценка регрессионной модели, направленной на анализ взаимосвязей между признаками и целевой переменной RiskScore. Сначала, с использованием метода наименьших квадратов (МНК), были рассчитаны параметры модели и проведён анализ описательных статистик данных. Для улучшения распределения исходных данных было применено преобразование Бокса-Кокса, что позволило достичь более точных и интерпретируемых результатов. Исследование выявило ключевые статистические характеристики, такие как средние значения, медианы, стандартные отклонения, коэффициенты асимметрии и эксцесса, что дало полноценное представление о распределении данных и их свойствах.

Далее, была выполнена проверка статистической значимости факторов с использованием t- и F-критериев. В ходе анализа обнаружено, что такие признаки, как SavingsAccountBalance и UtilityBillsPaymentHistory, являются незначимыми и не оказывают существенное влияние на целевую переменную, они были исключены для повышения качества модели. Применение полиномиальной регрессии позволило учесть нелинейные зависимости и повысить коэффициент детерминации модели до 79%, что свидетельствует о её высокой объяснительной способности и надёжности в прогнозировании.

4 АНАЛИЗ ОСТАТКОВ

Так как в нашей модели мы используем метод наименьших квадратов, то должны выполнять все предпосылки его использования:

- линейность модели;
- гомоскедастичность остатков;
- нормальное распределение остатков;
- отсутствие автокорреляции;
- отсутствие мультиколлинеарности.

Построим график распределения остатков (рисунок 4.1) и тщательно его проанализируем.

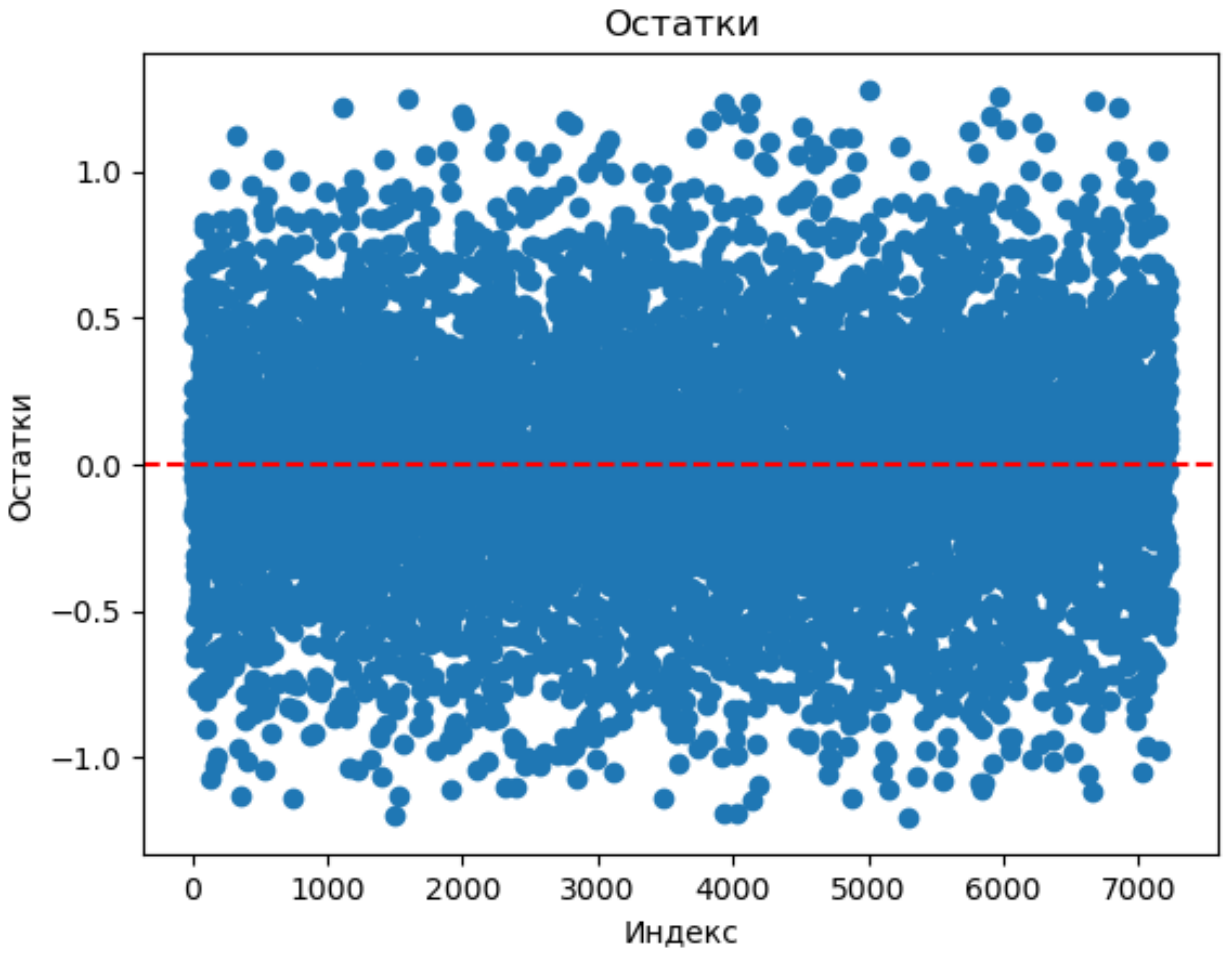
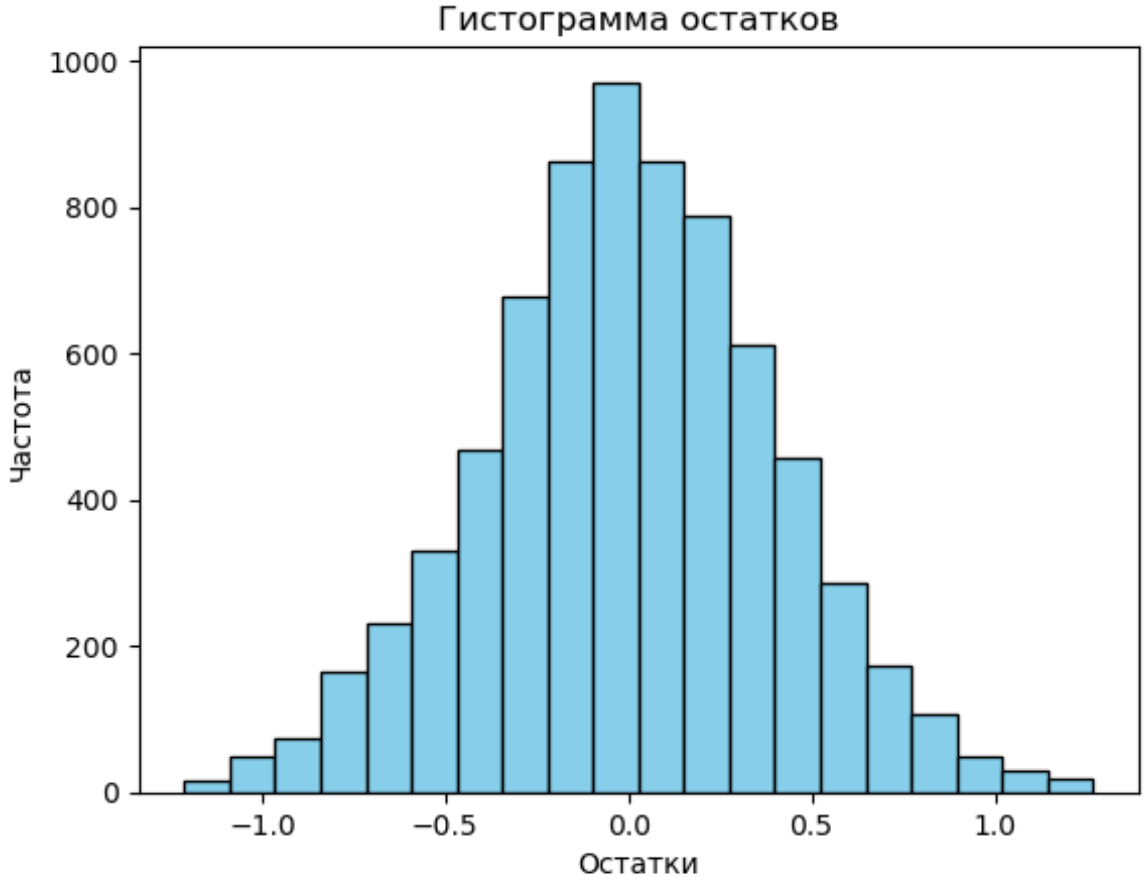


Рисунок 4.1 – График остатков

Остатки распределены случайным образом, что свидетельствует о хорошем соответствии модели данным. На графике отсутствуют видимые закономерности или тренды, что указывает на адекватность модели. Вместе с тем, для подтверждения её качества рекомендуется провести дополнительный анализ остатков, включая проверку на нормальность распределения и гомоскедастичность, чтобы исключить возможные скрытые проблемы.

Для того чтобы проверить остатки на нормальное распределение нам необходимо провести тест Жака-Берра (Jarque-Bera) и сравнить вероятностное значение с нашим уровнем значимости α=0.05, если Prob. Jarque-Bera > α, то нулевая гипотеза (H0) о наличии нормального распределения подтверждается. В ином случае действует альтернативная гипотеза H1 об отсутствии нормального распределения остатков. Гистограмма распределения остатков представлена на рисунке 4.2.




```
Ассиметрия остатков: -0.00404796958888035
Эксцесс остатков: 0.045916076132469286
Статистика Жака-Бера: 0.6545946426352091
P-значение (Жака-Бера): 0.7208693813267577
```

Рисунок 4.2 – Гистограмма распределения остатков

Используя уровень значимости 0.05, можно сделать следующие выводы. P-значение составляет 0.727869, что больше уровня значимости 0.05. Это означает, что мы не можем отвергнуть нулевую гипотезу о нормальности распределения данных. Таким образом, можно считать, что остатки имеют нормальное распределение.

Значения асимметрии и эксцесса близки к значениям, которые ожидаются для нормального распределения (около 0). Это также указывает на нормальность распределения остатков.

Для проверки остатков на автокорреляцию проведем тест Дарбина-Уотсона. Для этого нам понадобиться построить график с различными зонами автокорреляции и неопределенности.

Для того чтобы правильно определить границы зон нашего графика нам понадобиться вычислить du (нижняя граница) и dl (верхняя граница) с помощью таблицы критических значений Дарбина-Уотсона при 5% уровне значимости (рисунок 4.3). Так как объём выборки (n) составляет 7248, а количество экзогенных переменных (k) составляет 16. Для таких показателей, таблицы критических значений Дарбина-Уотсона обычно не содержат точных значений для таких больших выборок. Однако, для больших n, значения dl и du становятся практически постоянными.

На основании стандартных таблиц для уровня значимости 5% можно использовать интерполяцию или приближённые значения. Например: dl ≈ 1.5 и du ≈ 1.7.

Критические значения статистики Дарбина — Уотсона при 5%-ном уровне значимости

T	m									
	1		2		3		4		5	
	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Рисунок 4.3 – Критические значения статистики Дарбина-Уотсона при уровне значимости 0.05

На рисунке 4.4 показаны границы интервалов для подтверждения или опровержения нулевой гипотезы (H0) об наличии автокорреляции или ее отсутствии.

Значение DW_{calc}	Принимается гипотеза	Вывод
$0 \leq DW_{calc} < d_L$	отвергается H_0 , принимается $H_1 : \rho > 0$	есть положительная автокорреляция
$d_L \leq DW_{calc} \leq d_U$		неопределенность
$d_U < DW_{calc} < 4 - d_U$	принимается H_0	автокорреляция отсутствует
$4 - d_U \leq DW_{calc} \leq 4 - d_L$		неопределенность
$4 - d_L < DW_{calc} \leq 4$	отвергается H_0 , принимается $H_1 : \rho < 0$	есть отрицательная автокорреляция

Рисунок 4.4 – Границы интервалов

С помощью кода на Python я получила расчётное значение статистики (рисунок 4.5) в 1.9868, что больше du (1.7), но меньше, чем 4-du (2.3). Это свидетельствует о том, что автокорреляция в остатках отсутствует, то есть выполняется ещё одна предпосылка МНК.

[1497]:
```
from statsmodels.stats.stattools import durbin_watson

# Вычисляем статистику Дарбина-Уотсона на основе остатков
dw_stat = durbin_watson(residuals)

# Вывод результата
print(f"Статистика Дарбина-Уотсона: {dw_stat}")
```

```
Статистика Дарбина-Уотсона: 1.986804755735507
```

Рисунок 4.5 – Расчётное значение статистики Дарбина-Уотсона

В данном разделе доказано, что наша модель соответствует основным предпосылкам метода наименьших квадратов: остатки распределены случайным образом и не содержат закономерностей (рисунок 4.1), а тест Жака-Бера подтвердил нормальность их распределения (p-значение 0.727869 превышает уровень значимости 0.05), что также подтверждается гистограммой остатков (рисунок 4.2). Тест Дарбина-Уотсона показал значение статистики 1.9868, находящееся в диапазоне [du, 4-du] (du = 1.7, 4−du=2.3), указывая на отсутствие автокорреляции (рисунок 4.5). Это подтверждает адекватность модели и её соответствие требованиям МНК.

5 ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ ОЦЕНЕННОЙ МОДЕЛИ

5.1 Точечный прогноз индивидуального значения показателя

Формула уравнения полиномиальной регрессии в аналитическом виде представляет собой расширение линейной регрессии за счёт добавления степенных и взаимодействующих признаков. Для полиномиальной регрессии степени d уравнение можно записать так, как показано на рисунке 5.1.

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \sum_{j=i}^{k} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \sum_{j=i}^{k} \sum_{m=j}^{k} \beta_{ijm} x_i x_j x_m + \cdots + \epsilon$$

Где:

- y — предсказанное значение целевой переменной.
- β_0 — свободный член (интерцепт).
- $\beta_i, \beta_{ij}, \beta_{ijm}$ ... — коэффициенты модели, которые обучаются.
- x_i, x_j, x_m — значения независимых переменных (признаков).
- ϵ — ошибка модели.
- d — степень полинома (в вашем случае $d = 3$).

Рисунок 5.1 – Уравнение полиномиальной регрессии

Для нашей задачи коэффициенты я оформила в отдельной таблице, показанной на рисунке 5.2.

Свободный член (коэффициент при константе): 9072794.772950927

	Признак	Коэффициент при x^1	Коэффициент при x^2	Коэффициент при x^3
0	Experience	-5.163821e-02	9.494826e-03	-4.365444e-03
1	LoanDuration	-3.557149e-02	-2.969718e-02	6.896973e-03
2	MonthlyDebtPayments	2.871753e-02	1.588261e-02	1.777649e-03
3	CreditCardUtilizationRate	9.503999e-02	-6.987572e-03	-1.054764e-03
4	DebtToIncomeRatio	4.148750e-01	-2.256513e-02	-3.345156e-02
5	BankruptcyHistory	-6.501473e+10	3.480202e+10	3.300537e+10
6	PreviousLoanDefaults	1.085181e+11	-5.411694e+10	-5.439714e+10
7	PaymentHistory	-4.179886e-02	-1.044744e-02	4.409790e-03
8	SavingsAccountBalance	8.055091e-03	1.433045e-02	1.937866e-03
9	TotalAssets	-2.696433e-01	-1.076394e-01	-1.302147e-02
10	TotalLiabilities	6.376212e-02	2.392459e-02	5.331039e-04
11	MonthlyIncome	-6.135025e-01	-1.590098e-01	3.535461e-02
12	UtilityBillsPaymentHistory	-1.529598e-02	-6.326079e-03	3.280640e-04
13	InterestRate	3.121281e-01	-3.877378e-02	-1.196671e-02
14	MonthlyLoanPayment	1.151619e-01	-1.572800e-02	-1.407242e-02

Рисунок 5.2 – Рассчитанные коэффициенты полиномиальной регрессии

Зная уравнение модели, мы можем понять, как предсказывает RiskScore модель на новых данных. Код прогнозирования значений целевых переменных на 5 случайных записях исходного датафрейма следующий:

```
import random

# 1. Выбираем 5 случайных записей из датафрейма
random_indices = random.sample(range(len(df)), 5)
random_samples = df.iloc[random_indices].copy()  # Используем copy() для безопасного изменения

# Выполняем обратное преобразование Box-Cox для RiskScore
random_samples.loc[:, "RiskScore"] = target_transformer.inverse_transform(
    random_samples[["RiskScore"]]  # Передаём данные как DataFrame с правильными именами столбцов
)

# 2. Преобразуем X для полиномиальной регрессии
random_samples_X = random_samples.drop(columns="RiskScore")
random_samples_X_poly = poly.transform(random_samples_X)

# 3. Рассчитываем прогнозы
predictions = result_model.predict(random_samples_X_poly)

# Выполняем обратное преобразование Box-Cox для предсказаний
predictions = target_transformer.inverse_transform(pd.DataFrame(predictions, columns=["RiskScore"]))

# 4. Оформляем предсказания в датафрейм
random_samples.loc[:, "Предсказание"] = predictions
predictions_df = random_samples

print("Предсказания для случайных записей после обратного Box Cox преобразования: ")
print(predictions_df)
```

Результат прогнозирования приведён к изначальному диапазону с помощью обратного Box-Cox преобразования. Исходные и прогнозные значения RiskScore для 5 случайных записей можно увидеть на рисунке 5.3.

Предсказания для случайных записей (после обратного Box Cox преобразования:

	Experience	LoanDuration	MonthlyDebtPayments	CreditCardUtilizationRate	DebtToIncomeRatio	BankruptcyHistory	PreviousLoanDefaults	PaymentHistory	SavingsAccountBalance	TotalAssets	TotalLiabilities	MonthlyIncome	UtilityBillsPaymentHistory	InterestRate	MonthlyLoanPayment	RiskScore	Предсказание
3448	-0.125808	-0.675359	-0.309765	-1.157192	1.107052	0	0	-0.371821	-1.251956	0.692962	1.532481	0.025358	0.962058	0.183561	-0.699646	58.000001	53.595630
763	0.092240	1.548892	0.088438	0.418131	-0.336454	0	0	-1.010143	1.637689	1.397720	0.738018	-0.745022	-1.643946	0.847551	-1.295925	49.000001	50.183989
18716	0.092240	0.353324	-0.754769	-1.118345	1.404813	0	1	-1.010143	-1.812063	-1.563855	-0.544213	1.025839	-2.161074	-0.666698	-0.139231	59.000001	58.231133
187	0.092240	-1.330934	0.213001	-1.074176	-2.321633	0	0	-1.455082	0.043145	0.137948	-0.628158	0.004224	0.134375	0.085646	0.944731	48.000001	50.278351
14661	1.036326	0.353324	0.015789	-1.970711	-0.463974	0	0	0.821424	0.837542	0.679439	-1.267303	1.491104	0.442761	0.948322	0.511421	34.400001	40.850764

Рисунок 5.3 – Прогноз модели на пяти случайных записях

Предсказанные значения для случайных записей находятся близко к фактическим значениям, с минимальными отклонениями в пределах нескольких единиц, что подтверждает надёжность модели в учёте сложных полиномиальных зависимостей между признаками. Такие результаты указывают на адекватное соответствие модели данным, но мелкие различия могут быть вызваны индивидуальными особенностями записей.

5.2 Доверительный интервал для прогноза математического ожидания показателя

В этом разделе рассматривается методика расчёта доверительного интервала для прогнозирования математического ожидания целевого показателя. Цель состоит в том, чтобы определить диапазон, в котором с заданной уверенностью находится истинное среднее значение. Код, реализующий этот расчёт, представлен далее:

```
from scipy.stats import t

# Вычисляем параметры выборки
sample_mean = predictions_df['RiskScore'].mean()  # выборочное среднее
sample_std = predictions_df['RiskScore'].std(ddof=1)  # стандартное отклонение
sample_size = len(predictions_df['RiskScore']) # размер выборки

# Уровень доверия (например, 95%)
confidence_level = 0.95
alpha = 1 - confidence_level

# Коэффициент t для доверительного интервала
t_value = t.ppf(1 - alpha / 2, df=sample_size - 1)

# Стандартная ошибка
standard_error = sample_std / np.sqrt(sample_size)

# Границы доверительного интервала
lower_bound = sample_mean - t_value * standard_error
upper_bound = sample_mean + t_value * standard_error

print("Выборочное среднее RiskScore ", sample_mean)
print("Стандартное отклонение RiskScore ", sample_std)
print(f"Доверительный интервал (уровень доверия {confidence_level}): от {lower_bound:.2f} до {upper_bound:.2f}")
```

Сначала вычисляются параметры выборки: среднее значение, стандартное отклонение и размер. Затем задаётся уровень доверия, на основе которого рассчитывается коэффициент t с учётом числа степеней свободы. После этого определяется стандартная ошибка, которая используется для вычисления границ доверительного интервала. На финальном этапе выводятся результаты, включая доверительный интервал, выборочное среднее и стандартное отклонение. Алгоритм помогает оценить точность прогнозов и уровень неопределённости данных. Рассчитанные границы показаны на рисунке 5.4.

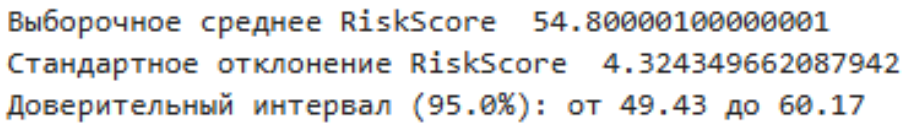


Рисунок 5.4 – Доверительный интервал для прогноза математического ожидания показателя

Результаты показывают, что выборочное среднее целевого показателя RiskScore составляет 54.80, а стандартное отклонение – 4.32, что свидетельствует о умеренной вариации значений в выборке. Расчёт доверительного интервала с уровнем доверия 95% показывает, что истинное среднее значение находится в диапазоне от 49.43 до 60.17. Это позволяет с высокой степенью уверенности оценить центральную тенденцию показателя.

5.3 Доверительный интервал для прогноза индивидуального значения показателя

Для построения доверительного интервала прогноза индивидуального значения используется алгоритм, который учитывает как стандартную ошибку модели, так и дополнительную неопределённость, связанную с разбросом индивидуальных наблюдений. Сначала задаётся уровень доверия (например, 95%) и вычисляется критическое значение t-распределения. Для каждого предсказания вычисляется погрешность, которая включает остаточную дисперсию модели и вариацию прогнозов, после чего определяются нижняя и верхняя границы интервала. Эти интервалы добавляются в итоговый датафрейм, чтобы показать диапазон, в котором с заданной вероятностью может находиться каждое индивидуальное значение RiskScore. Такой подход даёт более широкие границы по сравнению с математическим ожиданием, поскольку учитывает уникальную вариативность записей. Код алгоритма выглядит так:

```
# Объём выборки и число предикторов
n = len(df)  # Объём выборки
k = random_samples_X.shape[1]  # Количество предикторов

# Остаточная дисперсия модели
standard_error = np.sqrt(np.sum((random_samples["Предсказание"] - predictions.flatten())**2) / (n - k - 1))

# Критическое значение t
t_critical = t.ppf(1 - alpha / 2, df=n - k - 1)

# Доверительные интервалы для индивидуального значения
intervals = []
```

```
for pred in predictions.flatten():
    # Корректный расчёт погрешности
    margin_error = t_critical * np.sqrt(standard_error**2 + np.var(predictions))  # Учитываем дополнительную вариацию
    lower_bound = pred - margin_error
    upper_bound = pred + margin_error
    intervals.append((lower_bound, upper_bound))

# Добавляем в DataFrame
predictions_df["Доверительный интервал для индивидуального значения (нижняя граница)"] = [interval[0] for interval in intervals]
predictions_df["Доверительный интервал для индивидуального значения (верхняя граница)"] = [interval[1] for interval in intervals]

# Вывод результатов
print("Доверительные интервалы для индивидуальных прогнозов:")
predictions_df.iloc[:, -4:]
```

Рассчитанные границы доверительного интервала для прогноза показаны на рисунке 5.5.

```
Доверительные интервалы для индивидуальных прогнозов:
[108]:
```

	RiskScore	Предсказание	Доверительный интервал для индивидуального значения (нижняя граница)	Доверительный интервал для индивидуального значения (верхняя граница)
4606	52.000001	51.777572	44.437988	59.117157
2642	56.000001	55.764541	48.424957	63.104126
10919	57.000001	47.861856	40.522272	55.201441
17587	60.000001	56.649603	49.310018	63.989187
19703	49.000001	47.861202	40.521618	55.200787

Рисунок 5.5 – Доверительный интервал для прогноза индивидуального показателя

Результаты расчёта демонстрируют доверительные интервалы для прогноза показателя RiskScore для различных наблюдений. Для первого случая с реальным значением 52 доверительный интервал составляет от 44.44 до 59.12, что показывает возможные колебания индивидуальных значений вокруг прогноза. Аналогично, для второго наблюдения прогноз равен 55.77, а доверительный интервал – от 48.42 до 63.10, что даёт представление о пределах, в которых могут находиться значения.

В данном разделе было проведено прогнозирование значения целевого показателя RiskScore на основе полиномиальной модели с использованием пяти случайных записей из исходного датафрейма. Выполнено обратное преобразование Box-Cox, чтобы привести прогнозы к исходному масштабу. Рассчитаны доверительные интервалы для математического ожидания прогнозов и для индивидуальных значений, где учитывалась дополнительная вариация. Результаты подтвердили точность построенной модели.

6 ЭКОНОМИЧЕСКИЙ АНАЛИЗ ПО ОЦЕНЕННОЙ МОДЕЛИ

6.1 Средняя эффективность показателя

Целевая переменная RiskScore, отражающая степень кредитного риска заемщика, является ключевым индикатором в принятии финансовых решений. Она рассчитывается на основе совокупности факторов, таких как кредитная история, соотношение долга к доходу, уровень доходов и наличие дефолтов. Средняя эффективность данного показателя измеряется его влиянием на общий экономический результат, определяемый моделью полиномиальной регрессии.

```
# Обратное преобразование целевой переменной
y_actual_inverse = target_transformer.inverse_transform(y.values.reshape(-1, 1))
y_pred_inverse = target_transformer.inverse_transform(y_pred.reshape(-1, 1))

# Среднее значение после обратного преобразования
mean_y_actual = np.mean(y_actual_inverse)
mean_y_pred = np.mean(y_pred_inverse)

# Оценка средней эффективности
contribution_percentage = (mean_y_actual / mean_y_pred) * 100

print("Среднее значение фактического RiskScore (обратное Бокс-Кокс преобразование):", mean_y_actual)
print("Среднее значение прогнозируемого RiskScore (обратное Бокс-Кокс преобразование):", mean_y_pred)
print(f"Прогнозируемый вклад в экономический результат: {contribution_percentage:.2f}%")
```

С точки зрения экономического анализа данный код демонстрирует вклад модели в оценку ключевого показателя RiskScore, который отражает кредитный риск заемщика. Обратное преобразование Бокса-Кокса возвращает данные в исходную шкалу, что делает результаты более интерпретируемыми для анализа финансовых рисков. Среднее фактическое значение RiskScore представляет общую тенденцию заемщиков в выборке, а среднее прогнозируемое значение отражает способность модели предсказывать риск с учетом сложных взаимосвязей между факторами.

Прогнозируемый вклад, выраженный в процентах, показывает, насколько эффективно модель использует данные для оценки экономического результата, такого как вероятность выполнения заемщиком своих обязательств. Высокий процент вклада говорит о надежности модели в прогнозировании рисков, что способствует более точному управлению финансовыми решениями и минимизации экономических потерь. Анализ

также помогает выявить значимость отдельных факторов, улучшая понимание их влияния на кредитный риск. Результат расчёта показан на рисунке 6.1.

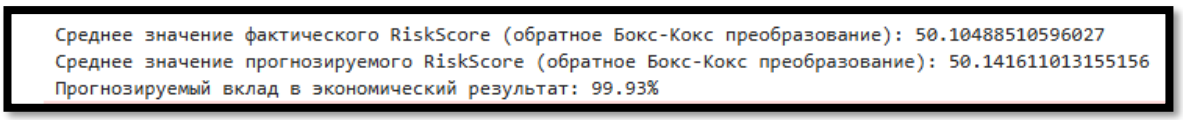


Рисунок 6.1 – Вклад модели в оценку показателя RiskScore

Проведём анализ влияние каждого из экзогенных признаков:

```
coefficients = model.coef_
features = poly.get_feature_names_out(input_features=X.columns)

# Формируем DataFrame для анализа коэффициентов
coefficients_df = pd.DataFrame({
    "Feature": features,
    "Coefficient": coefficients
})

# Сортируем коэффициенты по их абсолютному значению для лучшей визуализации
coefficients_df['Absolute Coefficient'] = coefficients_df['Coefficient'].abs()
sorted_df = coefficients_df.sort_values(by='Absolute Coefficient', ascending=False)

# Построение горизонтальной столбчатой диаграммы
plt.figure(figsize=(10, 6))
plt.barh(sorted_df['Feature'].head(10), sorted_df['Absolute Coefficient'].head(10), color='skyblue')
plt.xlabel('Важность признака (абсолютное значение коэффициента)')
plt.ylabel('Признаки')
plt.title('Топ-10 важных признаков')
plt.gca().invert_yaxis()  # Инвертируем ось, чтобы самые важные признаки были наверху
plt.show()
```

Коэффициенты связываются с соответствующими признаками, а затем вычисляются их абсолютные значения, чтобы подчеркнуть значимость каждого признака независимо от направления его воздействия (положительного или отрицательного). Результаты сортируются, и создаётся горизонтальная столбчатая диаграмма, где визуализируются топ-10 наиболее важных признаков. Это позволяет наглядно оценить, какие факторы

оказывают наибольшее влияние на целевую переменную. Гистограмма приведена на рисунке 6.2.

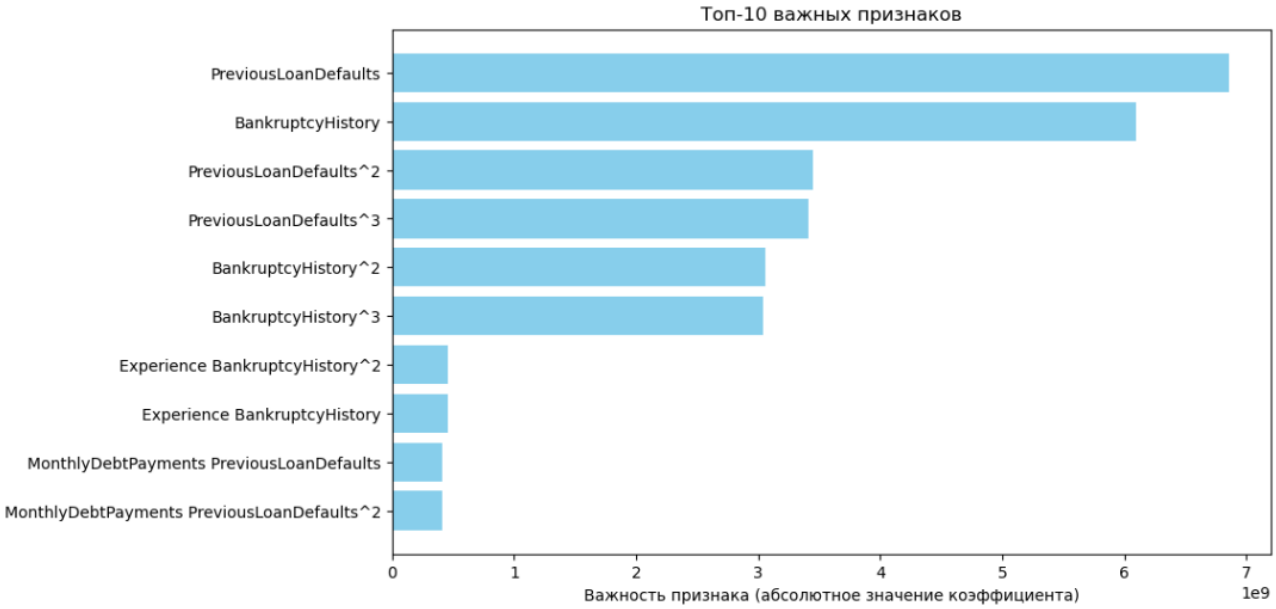


Рисунок 6.2 – Гистограмма важности признаков

Согласно диаграмме, наибольшее влияние на целевую переменную RiskScore оказывают такие факторы, как наличие предыдущих дефолтов, банкротств, месячные платежи по долгам, опыт работы и их взаимодействия в полиномиальных формах, например квадраты и кубы. Эти признаки представляют собой ключевые детерминанты кредитного риска, их нелинейные взаимосвязи обеспечивают более точное прогнозирование экономического результата.

Уравнение полиномиальной регрессии в нашем случае имеет вид $y=\beta1\cdot PreviousLoanDefaults+\beta2\cdot BankruptcyHistory+\beta3\cdot(PreviousLoanDefaults)2+\beta4\cdot(PreviousLoanDefaults)3+\beta5\cdot(BankruptcyHistory)2+\beta6\cdot(BankruptcyHistory)3+\ldots$ Рассчитанные моделью коэффициенты были представлены в разделе 5.1.

Благодаря использованию полиномиальной регрессии, учтены нелинейные зависимости между признаками, что позволяет оценить вклад каждого фактора в общий результат с высокой точностью. Выявлено, что модель способна адекватно отразить центральные тенденции кредитных рисков заемщиков и повысить точность их прогнозирования, что способствует лучшей оценке экономических рисков.

В разделе было выполнено исследование, направленное на оценку вклада модели полиномиальной регрессии в анализ целевого показателя RiskScore, который отражает степень кредитного риска заемщика. Также был выполнен анализ важности признаков, где коэффициенты модели использовались для определения значимости каждого из факторов.

6.2 Предельная эффективность показателя

Предельная эффективность показателя отражает, как изменение одного из факторов влияет на целевую переменную RiskScore. Этот показатель позволяет выявить наиболее значимые признаки, которые вносят наибольший вклад в изменение кредитного риска. Анализ предельной эффективности является важным этапом в интерпретации модели, так как он демонстрирует чувствительность целевой переменной к изменениям ключевых факторов, учитывая их эффекты. Используемый код:

```
# Получение коэффициентов модели
coefficients = model.coef_

# Получение названий признаков после полиномиального преобразования
feature_names = poly.get_feature_names_out(input_features=X.columns)

# Расчёт предельной эффективности
# Предельная эффективность – это частная производная целевой переменной по каждому признаку
marginal_effects = {
    feature: coef for feature, coef in zip(feature_names, coefficients)
}

# Формируем DataFrame для удобного анализа
marginal_effects_df = pd.DataFrame.from_dict(marginal_effects, orient='index', columns=['Marginal Effect'])
marginal_effects_df = marginal_effects_df.sort_values(by="Marginal Effect", ascending=False)

# Печать результатов
print("Предельная эффективность показателей:")
marginal_effects_df.head(20)  # Отображаем топ-20 признаков
```

Он анализирует предельную эффективность факторов, на основании коэффициентов, вычисленных ранее обученной полиномиальной регрессией. Сначала коэффициенты связываются с соответствующими признаками, включая их полиномиальные взаимодействия. Затем формируется DataFrame, который упорядочивает признаки по абсолютному значению их коэффициентов. Это позволяет визуализировать и проанализировать топ-20 факторов, оказывающих наибольшее влияние на изменение целевой переменной RiskScore.

На рисунке 6.3 представлена таблица, которая демонстрирует предельную эффективность показателей. Наибольший вклад в изменения целевой переменной оказывают квадраты и кубы ключевых признаков, таких

как наличие предыдущих дефолтов и история банкротств. Это указывает на сильные нелинейные зависимости между данными факторами и кредитным риском заемщика.

[144]:

	Marginal Effect
PreviousLoanDefaults^2	3.445095e+09
PreviousLoanDefaults^3	3.411543e+09
BankruptcyHistory^2	3.059572e+09
BankruptcyHistory^3	3.038604e+09
Experience BankruptcyHistory	4.533800e+08
MonthlyDebtPayments PreviousLoanDefaults	4.077434e+08
PreviousLoanDefaults InterestRate	1.786903e+08
BankruptcyHistory^2 TotalAssets	1.625897e+08
Experience PreviousLoanDefaults^2	1.397484e+08
BankruptcyHistory^2 SavingsAccountBalance	1.360492e+08
PreviousLoanDefaults^2 TotalAssets	1.259480e+08
PreviousLoanDefaults^2 UtilityBillsPaymentHistory	1.232257e+08
BankruptcyHistory UtilityBillsPaymentHistory	1.174103e+08
PreviousLoanDefaults MonthlyIncome	1.172133e+08
BankruptcyHistory^2 MonthlyIncome	1.112701e+08
DebtToIncomeRatio PreviousLoanDefaults	9.660944e+07
DebtToIncomeRatio BankruptcyHistory	8.834400e+07
BankruptcyHistory^2 MonthlyLoanPayment	8.806850e+07
BankruptcyHistory InterestRate	5.020499e+07
PreviousLoanDefaults MonthlyLoanPayment	4.960362e+07

Рисунок 6.3 – Предельная эффективность показателей

Например, наиболее значимые признаки включают PreviousLoanDefaults^2 с предельной эффективностью $3.445 * 10^9$ и BackruptcyHistory^3 с предельной эффективностью $3.038 * 10^9$. Это указывает на сильные нелинейные зависимости между данными факторами и кредитным риском заемщика. Значимость взаимодействующих признаков, таких как Опыт работы и История банкротств, также подтверждается их высокими коэффициентами. Результаты подчёркивают важность этих факторов в прогнозировании и оценке кредитного риска.

6.3 Частичные коэффициенты эластичности и общая эластичность

В данном разделе проведён расчёт частичных коэффициентов эластичности, которые показывают, как изменение каждого признака влияет на целевую переменную RiskScore. Эти коэффициенты позволяют оценить чувствительность кредитного риска заемщика к изменениям различных факторов. Общая эластичность, которая представляет собой сумму всех частичных эластичностей, демонстрирует совокупное влияние признаков на целевой показатель. Код расчёта:

```
# Расчёт частичных коэффициентов эластичности
elasticities = []
for i in range(X.values.shape[1]):
    # Коэффициент при текущем признаке
    coef = model.coef_[i]
    # Эластичность: частная производная * (значение признака / значение предсказанного y)
    elasticity = coef * (X.values[:, i] / y_pred)
    elasticities.append(elasticity)

# Средние частичные коэффициенты эластичности
elasticities_mean = np.mean(elasticities, axis=1)

# Общая эластичность
total_elasticity = np.sum(elasticities_mean)

# Формируем DataFrame для частичных коэффициентов эластичности
elasticities_df = pd.DataFrame({
    "Feature": X.columns,  # Названия признаков
    "Mean Elasticity": elasticities_mean  # Средние частичные коэффициенты эластичности
})

# Добавляем строку для общей эластичности
total_row = pd.DataFrame({
    "Feature": ["Total Elasticity"],
    "Mean Elasticity": [total_elasticity]
})

# Объединяем таблицы
elasticities_df = pd.concat([elasticities_df, total_row], ignore_index=True)

# Вывод результата
print("Результаты расчёта частичных коэффициентов эластичности:")
elasticities_df
```

Коэффициент эластичности определяется как произведение коэффициента регрессии признака на отношение значения признака к предсказанному значению целевой переменной. Далее вычисляется средняя эластичность каждого признака, а также общая эластичность, представляющая суммарное влияние всех признаков. Результаты преобразуются в DataFrame для удобства интерпретации и визуализации.

На рисунке 6.4 представлена таблица с частичными коэффициентами эластичности для каждого признака. Наиболее значимыми являются такие признаки, как Previous_Loan_Defaults, Bankruptcy_History, PaymentHistory и их взаимодействия. Например, признак PaymentHistory имеет наибольший положительный коэффициент эластичности, указывая на высокий вклад в снижение кредитного риска при улучшении истории платежей. Общая эластичность демонстрирует совокупное влияние всех признаков на изменение целевой переменной.

```
Результаты расчёта частичных коэффициентов эластичности:
[150]:
```

	Feature	Mean Elasticity
0	Experience	-4.731151e+05
1	LoanDuration	-1.718888e-01
2	MonthlyDebtPayments	-7.203875e-02
3	CreditCardUtilizationRate	5.995367e-02
4	DebtToIncomeRatio	2.146182e-01
5	BankruptcyHistory	1.054301e-02
6	PreviousLoanDefaults	-1.967857e+08
7	PaymentHistory	1.803408e+10
8	SavingsAccountBalance	4.639645e-03
9	TotalAssets	1.673699e-02
10	TotalLiabilities	-1.164921e-01
11	MonthlyIncome	2.102089e-01
12	UtilityBillsPaymentHistory	-1.747510e-01
13	InterestRate	-3.414444e-02
14	MonthlyLoanPayment	-3.999108e-01
15	Total Elasticity	1.783682e+10

Рисунок 6.4 – Коэффициенты эластичности

Результаты анализа подтверждают, что выбранные моделью факторы оказывают значительное влияние на кредитный риск, а их эластичность является ключевым показателем для дальнейших управленческих решений.

6.4 Предельная норма замещения факторов

Предельная норма замещения факторов (MRS, Marginal Rate of Substitution) отражает, насколько изменение одного признака может быть компенсировано изменением другого при сохранении целевой переменной RiskScore на одном уровне. Этот показатель позволяет глубже понять взаимозаменяемость факторов и их вклад в экономический результат. Анализ предельной нормы замещения особенно важен для финансовых моделей, где каждый из факторов оказывает сложное влияние на целевую переменную. Код функции для расчёта матрицы предельной нормы замещения (только для признаков в 1 степени):

```
def calculate_mrs_matrix_linear(model, X):
    # Получаем названия исходных признаков (1 степень)
    feature_names = X.columns

    # Инициализируем пустую матрицу для MRS
    num_features = len(feature_names)
    mrs_matrix = np.zeros((num_features, num_features))

    # Заполняем матрицу предельной нормы замещения
    for i in range(num_features):
        for j in range(num_features):
            if i != j:  # MRS имеет смысл только между разными признаками
                coef1 = model.coef_[i]  # Коэффициент при i-м признаке
                coef2 = model.coef_[j]  # Коэффициент при j-м признаке

                # Вычисляем предельную норму замещения
                if coef2 != 0:  # Избегаем деления на ноль
                    mrs_matrix[i, j] = -coef1 / coef2
                else:
                    mrs_matrix[i, j] = np.nan  # Если деление на ноль

    # Преобразуем матрицу в DataFrame
    mrs_df = pd.DataFrame(mrs_matrix, index=feature_names, columns=feature_names)

    return mrs_df

# Рассчитаем матрицу предельной нормы замещения для признаков в 1 степени
matrix = calculate_mrs_matrix_linear(model, X)
```

Предоставленный код позволяет рассчитать предельную норму замещения между всеми парами исходных признаков, исключая

полиномиальные взаимодействия. Для этого создаётся матрица, которая преобразуется в удобный формат DataFrame, где строки и столбцы соответствуют признакам из исходных данных. Визуализация результатов осуществляется с помощью тепловой карты, созданной с использованием библиотеки Seaborn. На рисунке 6.5 показана матрица предельной нормы замещения между исходными признаками. Она хорошо демонстрирует взаимозаменяемость факторов.

	Experience	LoanDuration	MonthlyDebtPayments	CreditCardUtilizationRate	DebtToIncomeRatio	BankruptcyHistory	PreviousLoanDefaults
Experience	0.000000e+00	8.767421e+06	1.425351e+07	-1.710801e+07	-5.150754e+06	-1.203201e+06	8.159389e-05
LoanDuration	1.140586e-07	0.000000e+00	-1.625736e+00	1.951317e+00	5.874880e-01	1.372355e-01	-9.306487e-12
MonthlyDebtPayments	7.015815e-08	-6.151061e-01	0.000000e+00	1.200267e+00	3.613674e-01	8.441436e-02	-5.724477e-12
CreditCardUtilizationRate	-5.845214e-08	5.124745e-01	8.331482e-01	0.000000e+00	-3.010726e-01	-7.032967e-02	4.769337e-12
DebtToIncomeRatio	-1.941463e-07	1.702163e+00	2.767267e+00	-3.321458e+00	0.000000e+00	-2.335970e-01	1.584115e-11
BankruptcyHistory	-8.311164e-07	7.286747e+00	1.184633e+01	-1.421875e+01	-4.280876e+00	0.000000e+00	6.781402e-11
PreviousLoanDefaults	1.225582e+04	-1.074519e+11	-1.746885e+11	2.096727e+11	6.312671e+10	1.474621e+10	0.000000e+00
PaymentHistory	1.378014e+04	-1.208163e+11	-1.964153e+11	2.357508e+11	7.097810e+10	1.658028e+10	-1.124375e+00
SavingsAccountBalance	7.911264e-08	-6.936138e-01	-1.127633e+00	1.353460e+00	4.074898e-01	9.518840e-02	-6.455108e-12
TotalAssets	-1.690388e-08	1.482034e-01	2.409397e-01	-2.891918e-01	-8.706774e-02	-2.033877e-02	1.379253e-12
TotalLiabilities	5.433270e-07	-4.763576e+00	-7.744317e+00	9.295246e+00	2.798544e+00	6.537315e-01	-4.433216e-11
MonthlyIncome	-1.321162e-07	1.158318e+00	1.883120e+00	-2.260246e+00	-6.804981e-01	-1.589623e-01	1.077987e-11
UtilityBillsPaymentHistory	1.217351e-06	-1.067303e+01	-1.735152e+01	2.082645e+01	6.270274e+00	1.464717e+00	-9.932838e-11
InterestRate	1.755398e-08	-1.539032e-01	-2.502059e-01	3.003138e-01	9.041626e-02	2.112097e-02	-1.432298e-12
MonthlyLoanPayment	-6.285507e-07	5.510769e+00	8.959055e+00	-1.075326e+01	-3.237510e+00	-7.562729e-01	5.128590e-11

	PaymentHistory	SavingsAccountBalance	TotalAssets	TotalLiabilities	MonthlyIncome	UtilityBillsPaymentHistory	InterestRate	MonthlyLoanPayment
Experience	7.256821e-05	1.264021e+07	-5.915801e+07	1.840512e+06	-7.569095e+06	8.214560e+05	5.696712e+07	-1.590961e+06
LoanDuration	-8.277031e-12	-1.441724e+00	6.747482e+00	-2.099263e-01	8.633205e-01	-9.369414e-02	-6.497592e+00	1.814629e-01
lyDebtPayments	-5.091252e-12	-8.868135e-01	4.150417e+00	-1.291269e-01	5.310337e-01	-5.763183e-02	-3.996708e+00	1.116189e-01
dUtilizationRate	4.241767e-12	7.388470e-01	-3.457912e+00	1.075819e-01	-4.424298e-01	4.801586e-02	3.329850e+00	-9.299510e-02
ıtToIncomeRatio	1.408885e-11	2.454049e+00	-1.148531e+01	3.573287e-01	-1.469512e+00	1.594827e-01	1.105996e+01	-3.088793e-01
nkruptcyHistory	6.031263e-11	1.050548e+01	-4.916719e+01	1.529680e+00	-6.290798e+00	6.827255e-01	4.734631e+01	-1.322274e+00
ɔusLoanDefaults	-8.893829e-01	-1.549161e+11	7.250299e+11	-2.255699e+10	9.276546e+10	-1.006762e+10	-6.981788e+11	1.949854e+10
PaymentHistory	0.000000e+00	-1.741838e+11	8.152055e+11	-2.536251e+10	1.043032e+11	-1.131978e+10	-7.850148e+11	2.192367e+10
AccountBalance	-5.741063e-12	0.000000e+00	4.680146e+00	-1.456078e-01	5.988110e-01	-6.498755e-02	-4.506819e+00	1.258652e-01
TotalAssets	1.226684e-12	2.136685e-01	0.000000e+00	3.111180e-02	-1.279471e-01	1.388579e-02	9.629655e-01	-2.689342e-02
TotalLiabilities	-3.942827e-11	-6.867764e+00	3.214214e+01	0.000000e+00	4.112493e+00	-4.463192e-01	-3.095177e+01	8.644123e-01
MonthlyIncome	9.587437e-12	1.669976e+00	-7.815731e+00	2.431615e-01	0.000000e+00	1.085276e-01	7.526280e+00	-2.101918e-01
PaymentHistory	-8.834096e-11	-1.538756e+01	7.201604e+01	-2.240549e+00	9.214242e+00	0.000000e+00	-6.934897e+01	1.936758e+00
InterestRate	-1.273861e-12	-2.218860e-01	1.038459e+00	-3.230833e-02	1.328678e-01	-1.441983e-02	0.000000e+00	2.792771e-02
hlyLoanPayment	4.561281e-11	7.945010e+00	-3.718381e+01	1.156855e+00	-4.757560e+00	5.163268e-01	3.580673e+01	0.000000e+00

Рисунок 6.5 – Предельные нормы замещения факторов

Анализ предельной нормы замещения факторов позволяет глубже понять взаимосвязь между признаками и их влиянием на целевой показатель. Тепловая карта способствует визуализации этих взаимосвязей, облегчая интерпретацию данных и управление финансовыми рисками в рамках модели.

ЗАКЛЮЧЕНИЕ

В процессе выполнения данной курсовой работы были подробно рассмотрены различные аспекты анализа кредитных рисков физических лиц, начиная с изучения теоретических основ и анализа литературы, заканчивая разработкой многофакторной эконометрической модели и проведением экономического анализа.

На первом этапе был выполнен анализ предметной области, освещены проблемы, связанные с оценкой рисков дефолта, а также проведён обзор современных литературных источников и программных решений в данной сфере. Были выделены ключевые направления для исследования и методологические подходы, которые легли в основу построения модели.

В рамках спецификации модели проведена идентификация переменных, анализ их распределения, устранение выбросов и аномальных значений. Осуществлён корреляционный анализ, который позволил выявить взаимосвязи между целевым признаком RiskScore и факторными переменными, а также между самими факторами.

Модель была построена с использованием метода наименьших квадратов (МНК), а её параметры оценены с применением F- и t-критериев для проверки их статистической значимости. Коэффициенты множественной детерминации и корреляции показали высокий уровень объяснения данных. Анализ остатков подтвердил корректность предпосылок модели, обеспечивая её надёжность и применимость.

Прогнозирование на основе разработанной модели позволило не только получать точечные прогнозы индивидуальных значений, но и оценивать доверительные интервалы как для математического ожидания, так и для индивидуальных значений целевой переменной. Эти результаты обеспечивают высокую точность и надёжность предсказаний.

Экономический анализ по оценённой модели включает изучение средней и предельной эффективности показателей, частичных коэффициентов эластичности и общей эластичности, а также предельной нормы замещения факторов. Такой подход позволил детально оценить вклад каждого фактора в изменение целевого показателя и выявить ключевые детерминанты риска дефолта заемщиков.

Результаты данной работы подчёркивают важность комплексного подхода к управлению кредитными рисками, который включает предварительную обработку данных, построение адекватной модели, прогнозирование риска и экономический анализ. Разработанная модель предоставляет банкам эффективный инструмент для оценки риска дефолта, что способствует оптимизации кредитного портфеля и минимизации финансовых потерь. Она может послужить основой для дальнейших исследований в области кредитного риск-менеджмента и разработки программных решений.

ПРИЛОЖЕНИЕ А
(справочное)
Описательные статистики признаков

	data_type	n_NaN	%NaN	nunique	count	mean	std	min	25%	50%	75%	max	Jarque-Bera	p-value
Age	int64	0	0.0	63	20000.0	40.0	12.0	18.0	32.0	40.0	48.0	80.0	1.957587e+02	3.101199e-43
AnnualIncome	int64	0	0.0	17516	20000.0	59162.0	40351.0	15000.0	31679.0	48566.0	74391.0	485341.0	5.363979e+04	0.000000e+00
EmploymentStatus	int32	0	0.0	3	20000.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.000000e+00	0.000000e+00
EducationLevel	int32	0	0.0	5	20000.0	2.0	1.0	0.0	1.0	1.0	3.0	4.0	0.000000e+00	0.000000e+00
Experience	int64	0	0.0	62	20000.0	18.0	11.0	0.0	9.0	17.0	25.0	61.0	5.000017e+02	2.666897e-109
LoanAmount	int64	0	0.0	15578	20000.0	24883.0	13427.0	3674.0	15575.0	21914.0	30835.0	184732.0	4.936154e+04	0.000000e+00
LoanDuration	int64	0	0.0	10	20000.0	54.0	25.0	12.0	36.0	48.0	72.0	120.0	1.437115e+03	0.000000e+00
MaritalStatus	int32	0	0.0	4	20000.0	1.0	1.0	0.0	1.0	1.0	2.0	3.0	0.000000e+00	0.000000e+00
NumberOfDependents	int64	0	0.0	6	20000.0	2.0	1.0	0.0	0.0	1.0	2.0	5.0	1.561893e+03	0.000000e+00
HomeOwnershipStatus	int32	0	0.0	4	20000.0	1.0	1.0	0.0	0.0	2.0	3.0	3.0	0.000000e+00	0.000000e+00
MonthlyDebtPayments	int64	0	0.0	1299	20000.0	454.0	240.0	50.0	286.0	402.0	564.0	2919.0	2.926760e+04	0.000000e+00
CreditCardUtilizationRate	float64	0	0.0	20000	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.210589e+03	1.330228e-263
umberOfOpenCreditLines	int64	0	0.0	14	20000.0	3.0	2.0	0.0	2.0	3.0	4.0	13.0	1.313126e+03	7.216393e-286
NumberOfCreditInquiries	int64	0	0.0	8	20000.0	1.0	1.0	0.0	0.0	1.0	2.0	7.0	4.525613e+03	0.000000e+00
DebtToIncomeRatio	float64	0	0.0	20000	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.188675e+03	7.629334e-259
BankruptcyHistory	int64	0	0.0	2	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.203966e+05	0.000000e+00
LoanPurpose	int32	0	0.0	5	20000.0	2.0	1.0	0.0	1.0	2.0	3.0	4.0	0.000000e+00	0.000000e+00
PreviousLoanDefaults	int64	0	0.0	2	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	4.541477e+04	0.000000e+00
PaymentHistory	int64	0	0.0	38	20000.0	24.0	5.0	8.0	21.0	24.0	27.0	45.0	1.521808e+02	9.002486e-34
SavingsAccountBalance	int64	0	0.0	9199	20000.0	4946.0	6605.0	73.0	1542.0	2986.0	5873.0	200089.0	5.343210e+06	0.000000e+00
CheckingAccountBalance	int64	0	0.0	5151	20000.0	1783.0	2245.0	24.0	551.0	1116.0	2126.0	52572.0	1.624739e+06	0.000000e+00
TotalAssets	int64	0	0.0	18814	20000.0	96964.0	120800.0	2098.0	31180.0	60699.0	117405.0	2619627.0	2.726391e+06	0.000000e+00
TotalLiabilities	int64	0	0.0	17163	20000.0	36252.0	47252.0	372.0	11197.0	22203.0	43146.0	1417302.0	5.255218e+06	0.000000e+00
MonthlyIncome	float64	0	0.0	17489	20000.0	4892.0	3297.0	1250.0	2630.0	4035.0	6163.0	25000.0	4.179047e+04	0.000000e+00
JtilityBillsPaymentHistory	float64	0	0.0	20000	20000.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	2.641857e+03	0.000000e+00
JobTenure	int64	0	0.0	17	20000.0	5.0	2.0	0.0	3.0	5.0	6.0	16.0	6.531297e+02	1.495162e-142
NetWorth	int64	0	0.0	17724	20000.0	72294.0	117920.0	1000.0	8735.0	32856.0	88826.0	2603208.0	3.239822e+06	0.000000e+00
BaseInterestRate	float64	0	0.0	18742	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.662574e+02	4.067342e-167
InterestRate	float64	0	0.0	19999	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.618383e+02	7.148262e-188
MonthlyLoanPayment	float64	0	0.0	20000	20000.0	912.0	675.0	97.0	494.0	728.0	1113.0	10893.0	2.905837e+05	0.000000e+00
TotalDebtToIncomeRatio	float64	0	0.0	20000	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	1.339060e+05	0.000000e+00
LoanApproved	int64	0	0.0	2	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	5.203738e+03	0.000000e+00
RiskScore	float64	0	0.0	73	20000.0	51.0	8.0	29.0	46.0	52.0	56.0	84.0	1.082507e+02	3.116490e-24

Рисунок А.1 – Описательные статистики переменных до преобразования

	data_type	n_NaN	%NaN	nunique	count	mean	std	min	25%	50%	75%	max	Jarque-Bera	p-value
Age	float64	0	0.0	59	20000.0	-0.0	1.0	-2.0	-1.0	0.0	1.0	3.0	153.887968	3.834002e-34
AnnualIncome	float64	0	0.0	17516	20000.0	-0.0	1.0	-2.0	-1.0	0.0	1.0	3.0	151.614354	1.194982e-33
EmploymentStatus	int32	0	0.0	3	20000.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.000000	0.000000e+00
EducationLevel	int32	0	0.0	5	20000.0	2.0	1.0	0.0	1.0	1.0	3.0	4.0	0.000000	0.000000e+00
Experience	float64	0	0.0	62	20000.0	0.0	1.0	-2.0	-0.0	0.0	1.0	2.0	2214.133877	0.000000e+00
LoanAmount	float64	0	0.0	15578	20000.0	-0.0	1.0	-4.0	-1.0	0.0	1.0	4.0	0.494232	7.810503e-01
LoanDuration	float64	0	0.0	10	20000.0	0.0	1.0	-2.0	-1.0	-0.0	1.0	2.0	17.557684	1.539562e-04
MaritalStatus	int32	0	0.0	4	20000.0	1.0	1.0	0.0	1.0	1.0	2.0	3.0	0.000000	0.000000e+00
NumberOfDependents	int64	0	0.0	6	20000.0	2.0	1.0	0.0	0.0	1.0	2.0	5.0	1561.893037	0.000000e+00
HomeOwnershipStatus	int32	0	0.0	4	20000.0	1.0	1.0	0.0	0.0	2.0	3.0	3.0	0.000000	0.000000e+00
MonthlyDebtPayments	float64	0	0.0	1215	20000.0	0.0	1.0	-3.0	-1.0	0.0	1.0	3.0	17.491851	1.591083e-04
CreditCardUtilizationRate	float64	0	0.0	19997	20000.0	0.0	1.0	-3.0	-1.0	0.0	1.0	3.0	190.704955	3.881015e-42
NumberOfOpenCreditLines	float64	0	0.0	10	20000.0	0.0	1.0	-2.0	-0.0	0.0	1.0	2.0	176.008836	6.027911e-39
NumberOfCreditInquiries	float64	0	0.0	8	20000.0	0.0	1.0	-1.0	-1.0	1.0	1.0	1.0	3252.270591	0.000000e+00
DebtToIncomeRatio	float64	0	0.0	19997	20000.0	0.0	1.0	-3.0	-1.0	0.0	1.0	3.0	202.658079	9.848210e-45
BankruptcyHistory	int64	0	0.0	2	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	220396.617683	0.000000e+00
LoanPurpose	int32	0	0.0	5	20000.0	2.0	1.0	0.0	1.0	2.0	3.0	4.0	0.000000	0.000000e+00
PreviousLoanDefaults	int64	0	0.0	2	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	45414.773322	0.000000e+00
PaymentHistory	float64	0	0.0	25	20000.0	0.0	1.0	-2.0	-1.0	0.0	1.0	2.0	63.505045	1.622014e-14
SavingsAccountBalance	float64	0	0.0	9199	20000.0	0.0	1.0	-4.0	-1.0	-0.0	1.0	4.0	4.982330	8.281342e-02
CheckingAccountBalance	float64	0	0.0	5151	20000.0	-0.0	1.0	-4.0	-1.0	0.0	1.0	4.0	0.016950	9.915607e-01
TotalAssets	float64	0	0.0	18814	20000.0	-0.0	1.0	-4.0	-1.0	0.0	1.0	4.0	15.361425	4.616458e-04
TotalLiabilities	float64	0	0.0	17163	20000.0	0.0	1.0	-4.0	-1.0	0.0	1.0	4.0	0.000267	9.998668e-01
MonthlyIncome	float64	0	0.0	17489	20000.0	-0.0	1.0	-2.0	-1.0	0.0	1.0	3.0	159.551745	2.258286e-35
UtilityBillsPaymentHistory	float64	0	0.0	20000	20000.0	0.0	1.0	-2.0	-1.0	0.0	1.0	2.0	563.857808	3.629382e-123
JobTenure	float64	0	0.0	13	20000.0	0.0	1.0	-3.0	-1.0	0.0	0.0	2.0	52.332660	4.326206e-12
NetWorth	float64	0	0.0	17724	20000.0	-0.0	1.0	-2.0	-1.0	0.0	1.0	3.0	497.414273	9.724404e-109
BaseInterestRate	float64	0	0.0	18640	20000.0	0.0	1.0	-3.0	-1.0	-0.0	1.0	3.0	36.256661	1.339572e-08
InterestRate	float64	0	0.0	19880	20000.0	0.0	1.0	-3.0	-1.0	-0.0	1.0	3.0	30.991180	1.863592e-07
MonthlyLoanPayment	float64	0	0.0	19843	20000.0	0.0	1.0	-3.0	-1.0	-0.0	1.0	3.0	13.969251	9.260102e-04
TotalDebtToIncomeRatio	float64	0	0.0	19906	20000.0	0.0	1.0	-3.0	-1.0	-0.0	1.0	3.0	50.458596	1.104219e-11
LoanApproved	int64	0	0.0	2	20000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	5203.738315	0.000000e+00

Рисунок А.2 – Описательные статистики переменных после преобразования

ПРИЛОЖЕНИЕ Б
(справочное)
Графики «ящиков с усами»

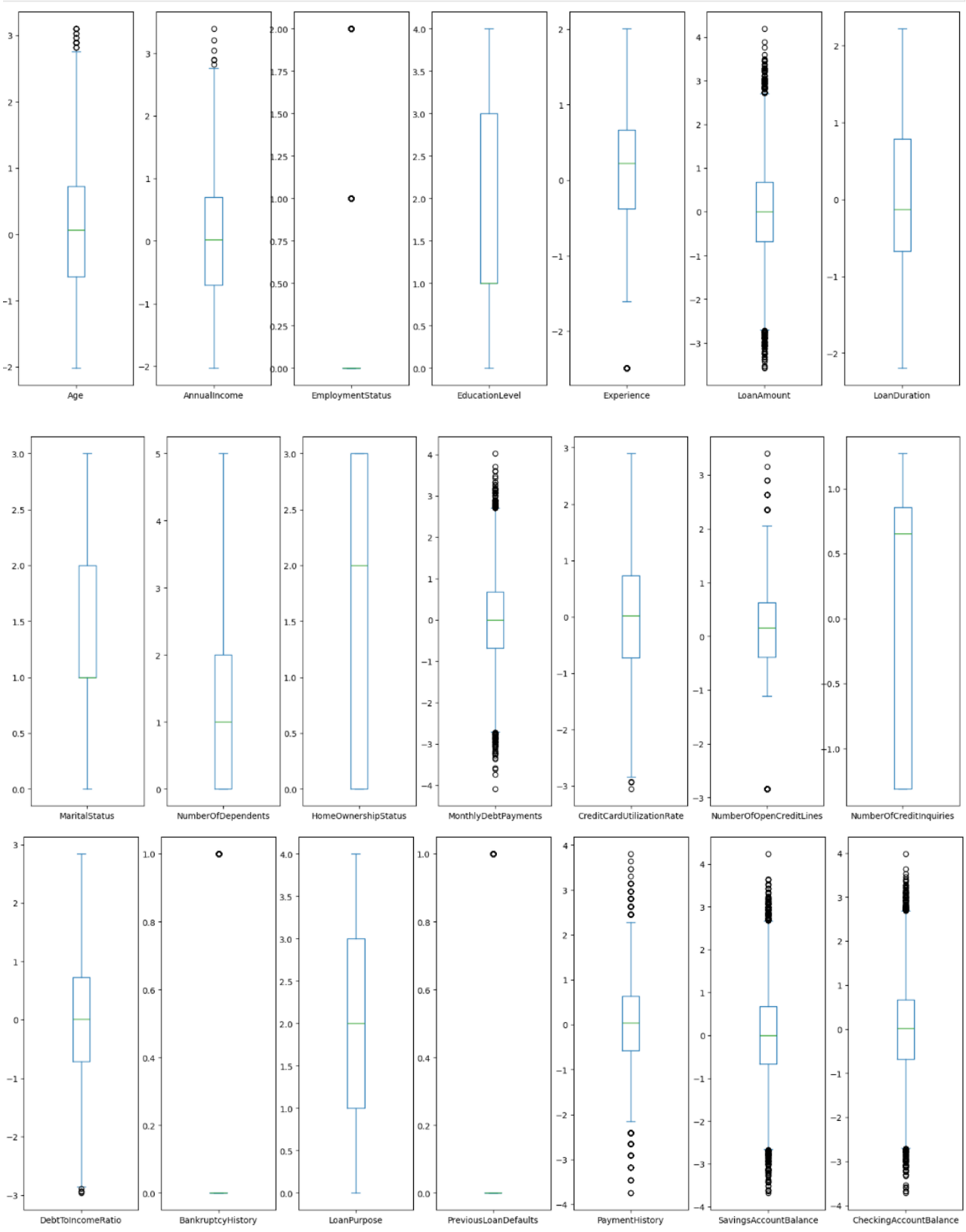


Рисунок Б.1 – Графики «ящиков с усами» для экзогенных признаков

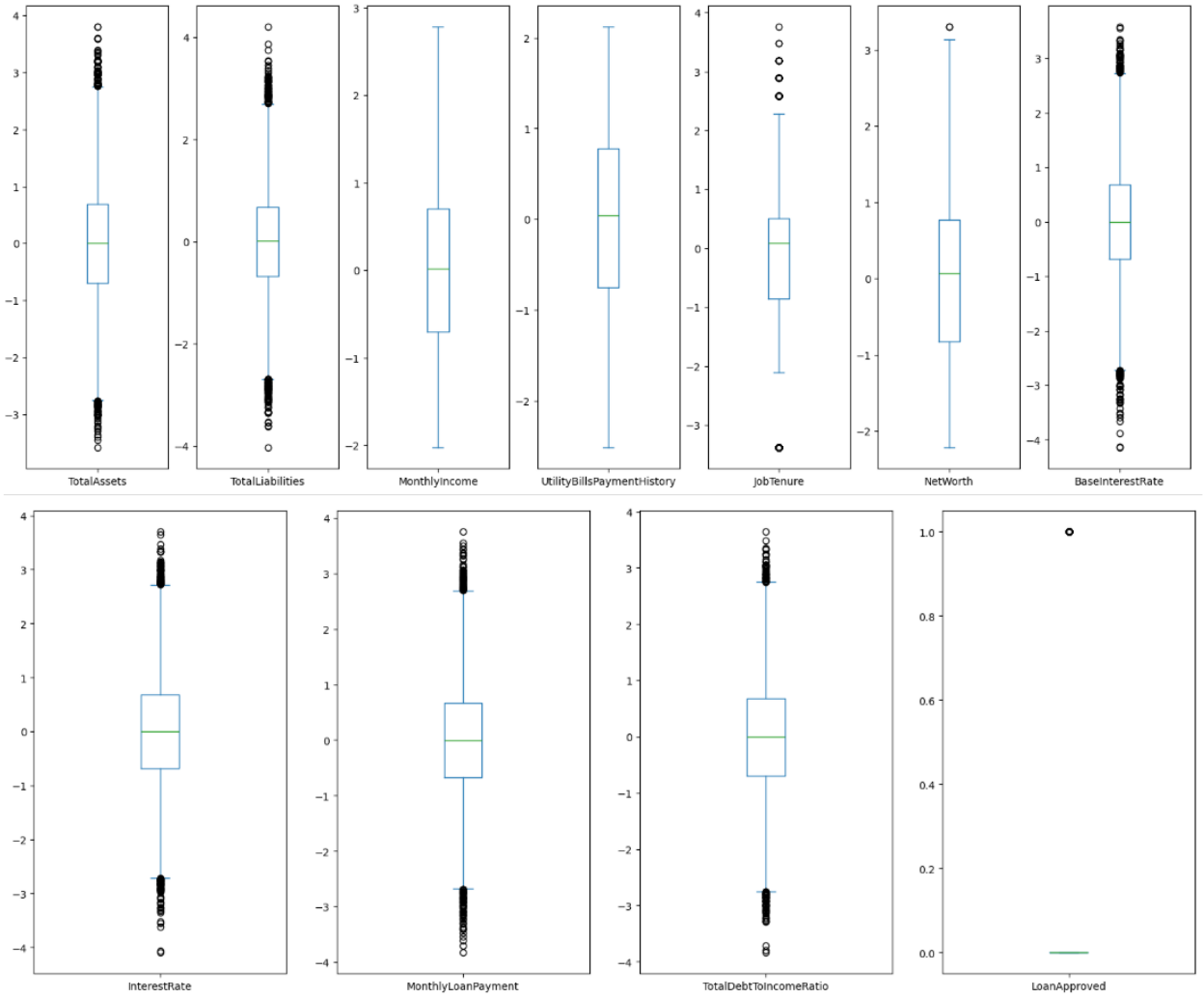


Рисунок Б.2 – Графики «ящиков с усами» для оставшихся признаков

ПРИЛОЖЕНИЕ В
(справочное)
Матрица корреляции признаков

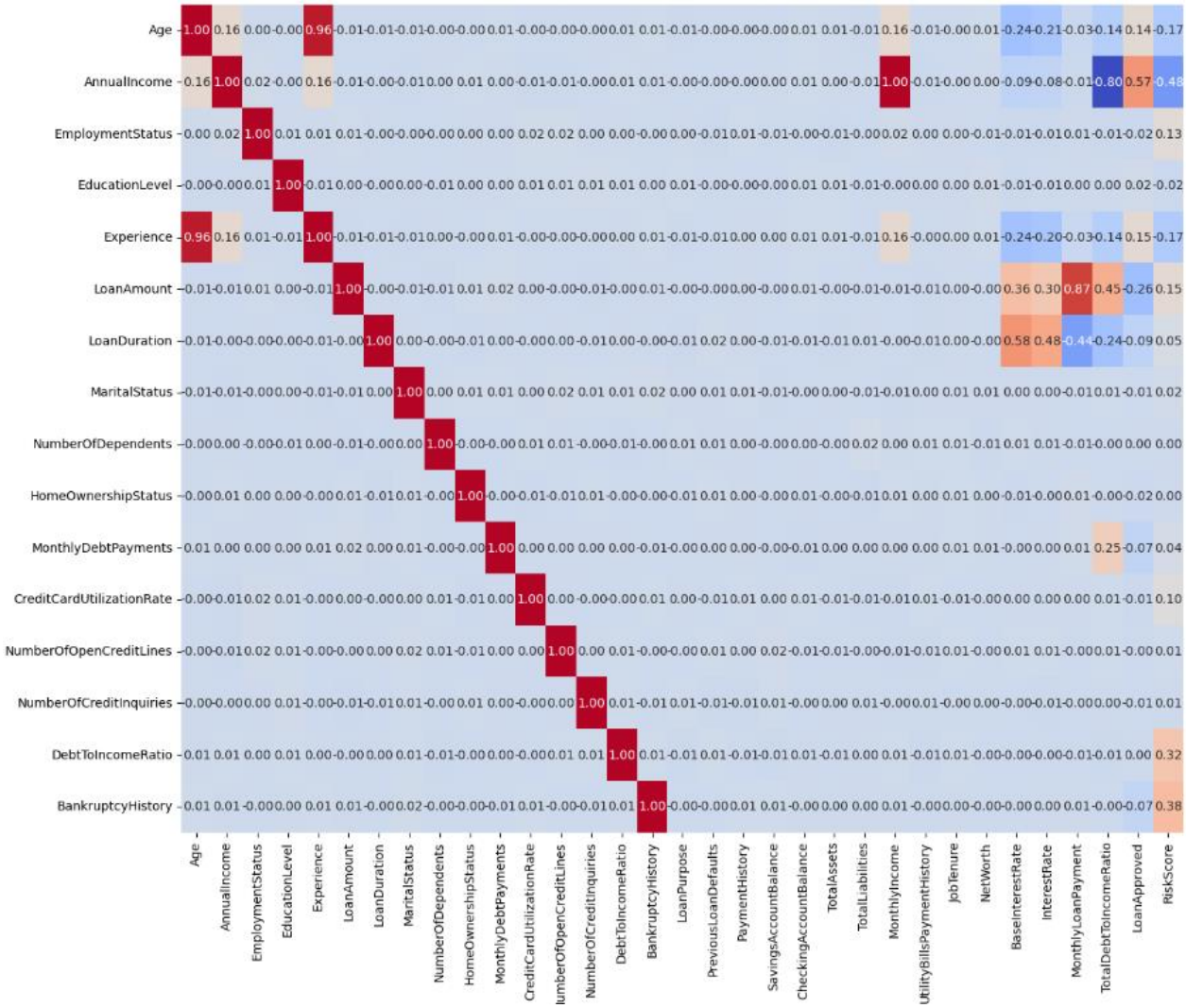


Рисунок В.1 – Первая часть матрицы корреляций

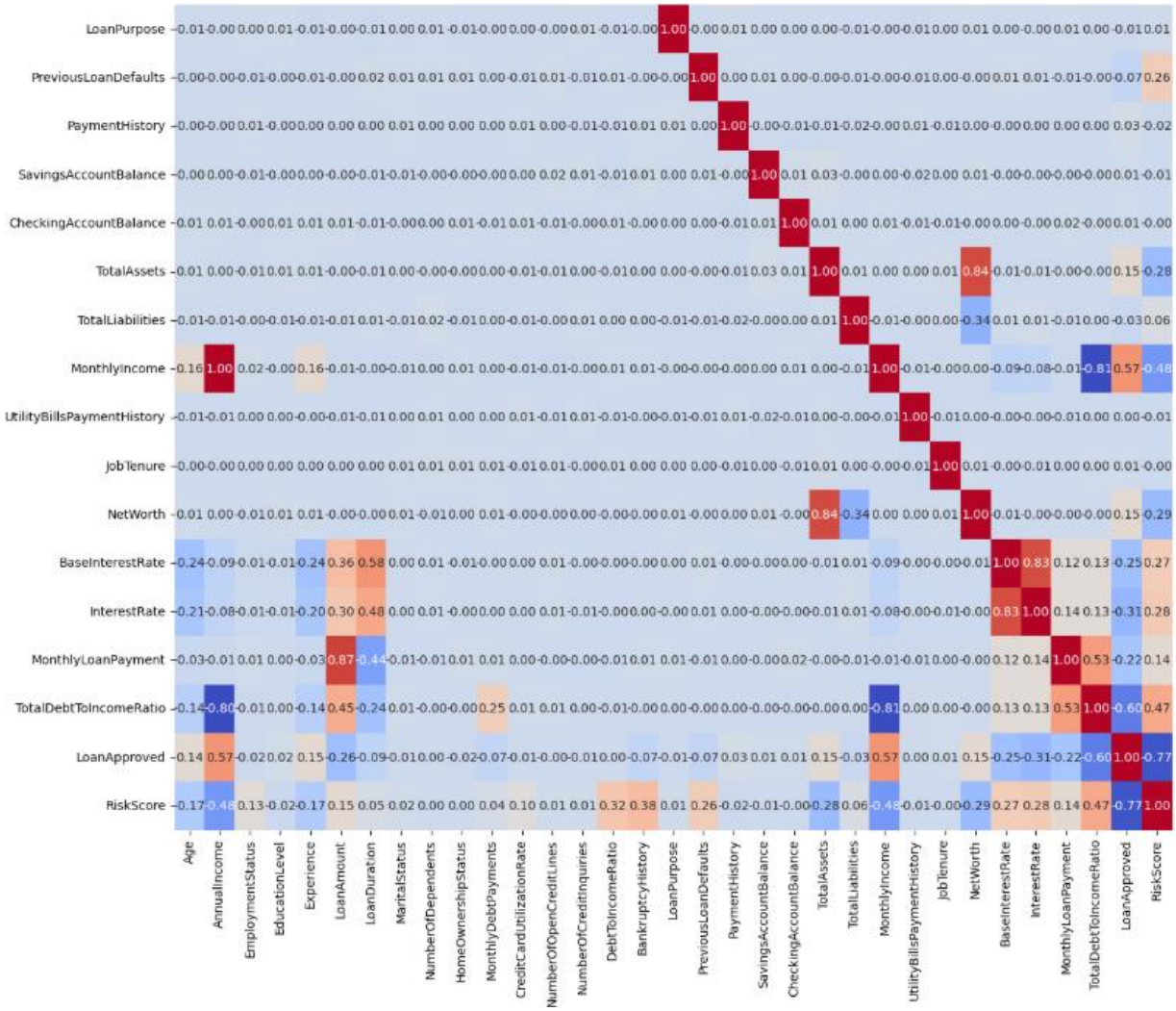


Рисунок В.2 – Вторая часть матрицы корреляций

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Магнус Л.Р., Катышев П.К., Пересецкий А.А. Эконометрика. М., Дело, 2004. - 576 с.

[2] Бородич С.А. Эконометрика. Минск: Новое знание, 2001. - 230 с.

[3] Эконометрика. Под редакцией Елисеевой И.И. М.: Финансы и статистика, 2007. – 18с.

[4] Доугерти К. Введение в эконометрику. М.: ИНФРА-М, 2004 – 416с.

[5] Экономико-математические методы и модели. Под ред. Миксюк С.Ф. Мн.: БГЭУ, 2006. – 46с.

[6] Экономико-математические методы и модели; практикум. Под ред. Миксюк С.Ф. Мн.: БГЭУ, 2006. – 46с.

[7] Хацкевич, Г. А. Эконометрика: учеб.-метод. комплекс для студентов экономических специальностей / Г. А. Хацкевич, А. Б. Гедранович. – Минск: Изд-во МИУ, 2005. – 252 с.

[8] Марченко, В. М. Методы оптимизации и статистической обработки результатов измерений: учеб. пособие / В. М. Марченко, Т. Б. Копейкина. – Минск: БГТУ, 2007. − 140 с.

[9] Кремер, Н. Ш. Эконометрика: учебник для студентов вузов / Н. Ш. Кремер, Б. А. Путко; под ред. Н. Ш. Кремера. – 2-е изд., стереотип. – М.: ЮНИТИ-ДАНА, 2008. – 311 с.

[10] Носко В.П. Эконометрика для начинающих, 2005. – 379 с.