

Построение и сравнительный анализ регрессионных и классификационных моделей машинного обучения для оценки кредитных рисков

Саяпин Никита Анатольевич

Приамурский государственный университет им.Шолом-Алейхема

Студент

Аннотация

В работе рассматривается задача построения и анализа моделей машинного обучения для решения задач классификации и регрессии на основе кредитных данных заемщиков. Особое внимание уделено проблеме несбалансированности классов и её влиянию на качество классификационных моделей. Подбор гиперпараметров осуществлялся с использованием GridSearchCV. Проведён сравнительный анализ не менее восьми моделей для каждой задачи, а также выполнена оценка важности признаков.

Ключевые слова: машинное обучение, кредитный скоринг, классификация, регрессия, анализ данных, GridSearchCV, несбалансированные классы, важность признаков.

Construction and comparative analysis of regression and classification machine learning models for credit risk assessment

Sayapin Nikita Anatolyevich

Sholom-Aleichem Priamursky State University

Student

Abstract

This article presents the development and evaluation of regression and classification machine learning models based on the Heart Failure Clinical Records dataset. The dataset includes 13 clinical features and two target variables: patient follow-up time and death event indicator. The study focuses on data preprocessing, comparison of different machine learning algorithms, exploration of multiple feature encoding methods, and assessment of model performance using standard regression and classification metrics.

Keywords: machine learning, credit scoring, classification, regression, data analysis, GridSearchCV, class imbalance, feature importance

1. Введение

1.1 Актуальность

В условиях цифровизации финансового сектора и роста объёма обрабатываемых данных особую значимость приобретают методы интеллектуального анализа данных и машинного обучения. Кредитные

организации всё чаще используют автоматизированные системы оценки заемщиков, позволяющие повысить точность принятия решений и снизить кредитные риски.

Традиционные статистические методы не всегда способны эффективно обрабатывать большие объёмы разнородных данных и учитывать сложные нелинейные зависимости между признаками. В связи с этим актуальной является задача применения и сравнительного анализа современных алгоритмов машинного обучения для оценки кредитоспособности заемщиков.

Особую сложность представляет несбалансированность классов в кредитных данных, когда число надёжных заемщиков существенно превышает количество проблемных. Игнорирование данного фактора может приводить к построению моделей с высокой общей точностью, но низкой практической ценностью. Поэтому разработка и анализ моделей, устойчивых к дисбалансу классов и обладающих интерпретируемостью, является важной научной и прикладной задачей.

1.2 Обзор исследований

В работе М.И. Зайнуллина представлен обзор современных алгоритмов машинного обучения для прогнозирования развития сердечно-сосудистых заболеваний; автор систематизирует линейные и нелинейные модели, обсуждает выбор признаков и подчёркивает роль ансамблевых методов и ROC-анализа в построении прогностических моделей риска [1]. И.А. Мишкин и А.А. Сахаров на клиническом материале демонстрируют применение алгоритмов машинного обучения (в том числе градиентного бустинга и случайного леса) для оценки риска развития сердечно-сосудистых заболеваний, сравнивая их с традиционными шкалами. Показано, что модели машинного обучения обеспечивают более высокую точность и лучшее качество стратификации риска по сравнению с классическими статистическими подходами, при этом F1-мера и ROC-AUC используются как ключевые критерии качества [2]. А.В. Гусев и соавторы в обзорной работе выделяют перспективы применения методов машинного обучения в кардиологии, акцентируя внимание на том, что использование нейросетевых и ансамблевых моделей позволяет повысить точность прогнозирования сердечно-сосудистых событий и улучшить риск-стратификацию пациентов по сравнению с традиционными регрессионными моделями и шкалами [3]. В другой статье этот же авторский коллектив демонстрирует подход к прогнозированию заболеваний (в том числе сердечной недостаточности и почечной патологии) на основе лабораторных данных, используя сверточные и рекуррентные нейронные сети (CNN и LSTM) и показывая высокие значения AUC для клинических задач [4]. В.И. Горбаченко, где описана нейросетевая модель прогнозирования хронической сердечной недостаточности по данным тензиометрического исследования крови. Показывается, что правильно настроенная нейронная сеть способна обеспечивать высокую точность ранней диагностики ХСН и может использоваться как элемент системы поддержки принятия врачебных решений [5].

1.3 Цель исследования

Целью данной работы является построение и сравнительный анализ регрессионных и классификационных моделей машинного обучения для оценки кредитных рисков заемщиков на основе табличных данных.

2. Разведочный анализ данных (EDA)

Разведочный анализ — ключевой этап исследования клинических данных. Он позволяет выявить структуру признаков, их распределения, взаимосвязи, выбросы и потенциальные проблемы, влияющие на качество моделей.

2.1 Загрузка и первичный анализ данных

В качестве исходных данных используется табличный набор данных Loan Approval Classification Dataset [6], содержащий информацию о заемщиках и параметрах кредитных заявок. Данные представлены в формате CSV и включают как числовые, так и категориальные признаки. Перед построением регрессионных и классификационных моделей был проведён этап предварительной обработки данных, включающий анализ структуры, проверку пропусков и приведение признаков к машинно-обрабатываемому виду (Рис.1).

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45000 entries, 0 to 44999
Data columns (total 14 columns):
 #   Column                                Non-Null Count  Dtype  
---  --
 0   person_age                           45000 non-null  float64
 1   person_gender                         45000 non-null  object  
 2   person_education                     45000 non-null  object  
 3   person_income                        45000 non-null  float64
 4   person_emp_exp                       45000 non-null  int64   
 5   person_home_ownership                45000 non-null  object  
 6   loan_amnt                           45000 non-null  float64
 7   loan_intent                          45000 non-null  object  
 8   loan_int_rate                        45000 non-null  float64
 9   loan_percent_income                  45000 non-null  float64
10   cb_person_cred_hist_length           45000 non-null  float64
11   credit_score                         45000 non-null  int64   
12   previous_loan_defaults_on_file       45000 non-null  object  
13   loan_status                          45000 non-null  int64   
dtypes: float64(6), int64(3), object(5)
memory usage: 4.8+ MB
```

Рисунок 1 – Первичный анализ данных

2.2 Анализ пропущенных значений

В ходе разведочного анализа данных была проведена проверка на наличие пропущенных значений. Для этого была рассчитана доля пропусков в каждом признаке (Рис.2).

Анализ показал, что в рассматриваемом датасете отсутствуют пропущенные значения: все признаки заполнены полностью. Таким образом, процедуры импутации пропусков не потребовались, и данные могут быть использованы для дальнейшего анализа без дополнительного восстановления значений.

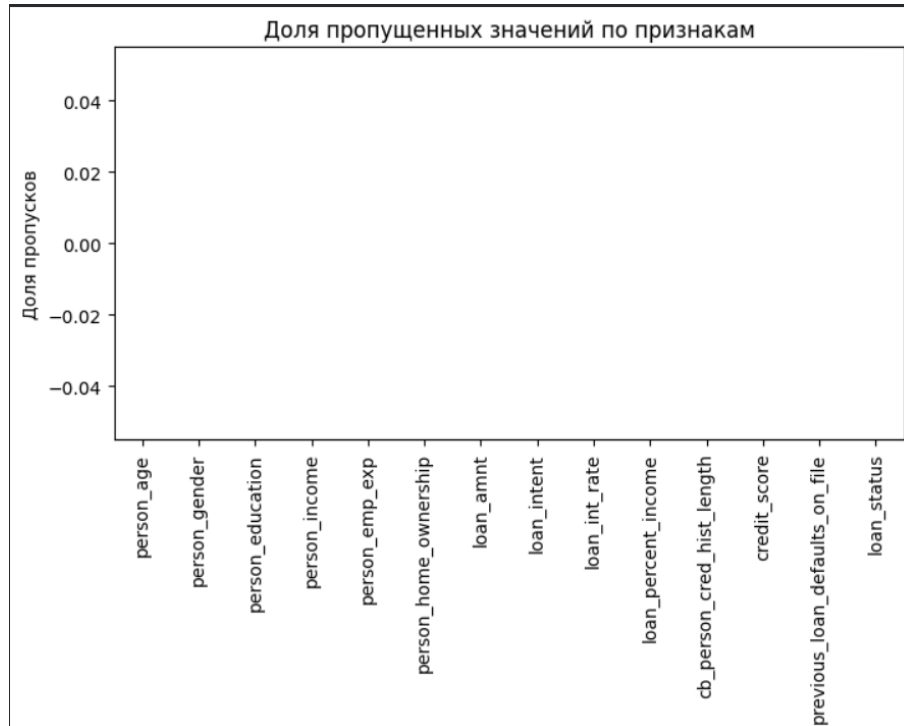


Рисунок 2 – Просмотр долей пропусков

2.3 Анализ распределений числовых признаков

Для понимания структуры данных и выявления возможных выбросов были проанализированы распределения числовых признаков (Рис.3).

Распределения числовых признаков

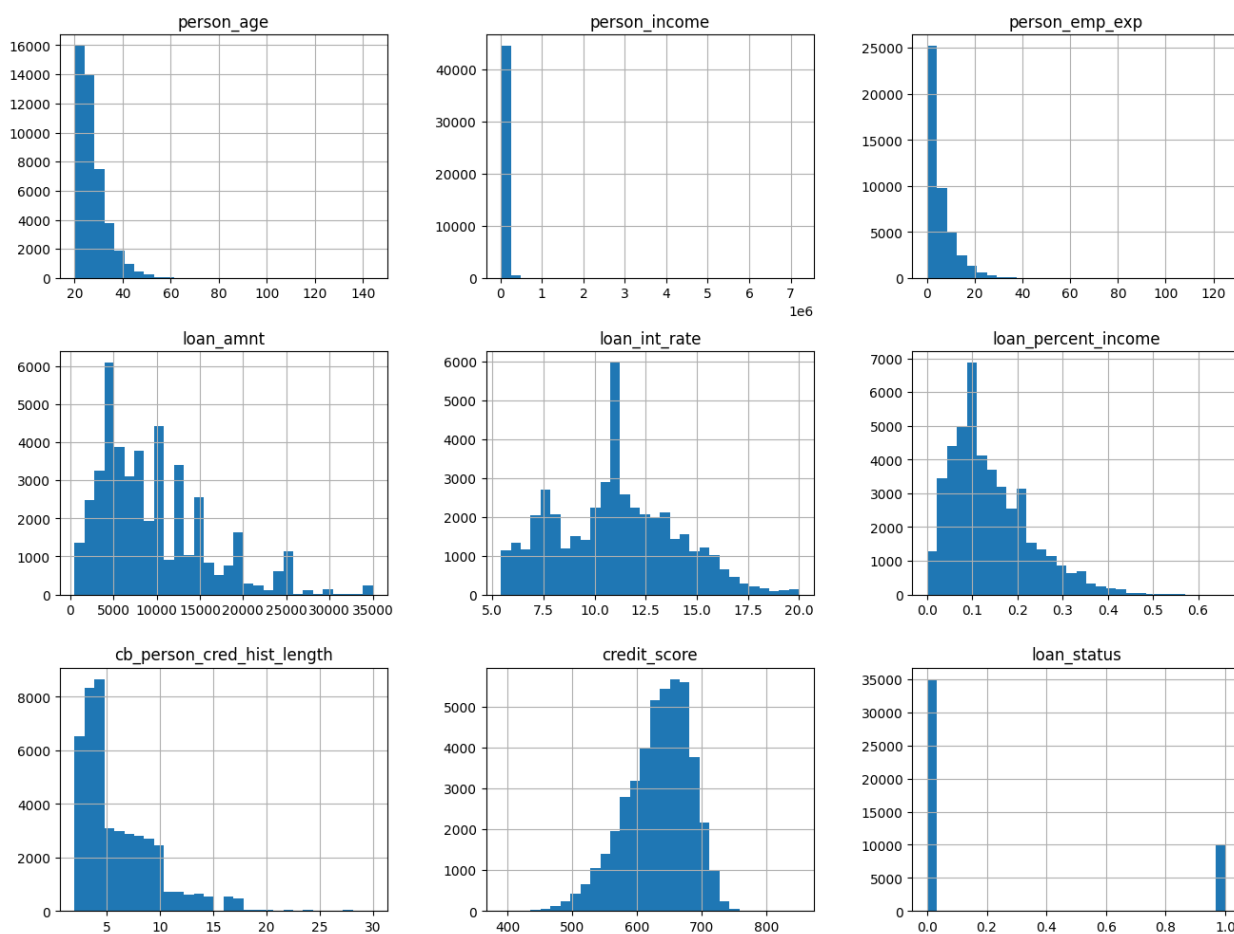


Рисунок 3 – Распределение числовых признаков

Распределения ряда признаков (например, `person_income`, `loan_amnt`) характеризуются правосторонней асимметрией, что типично для финансовых данных и может потребовать масштабирования или логарифмирования при построении моделей.

2.4 Анализ категориальных признаков

Рассмотрим распределение категориальных переменных (Рис.4-8).

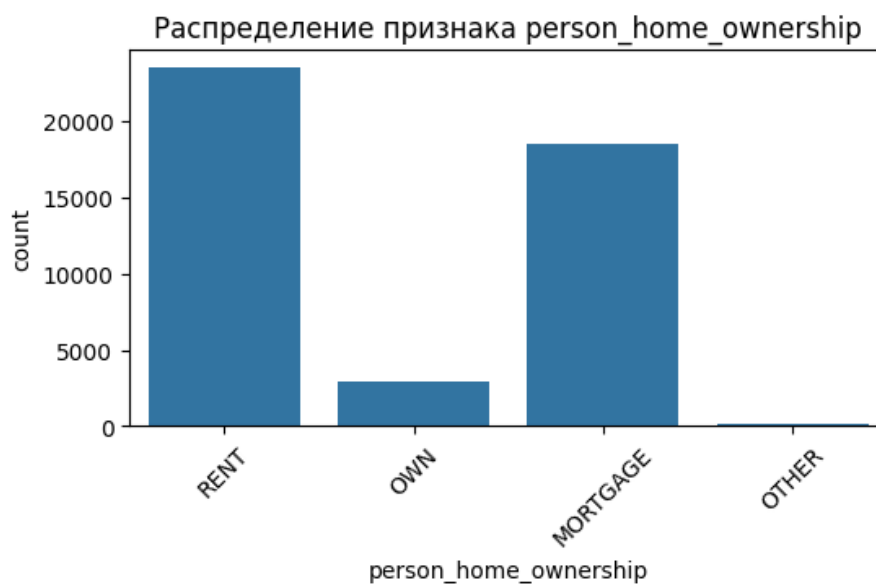


Рисунок 4 – Распределение признака person_home_ownership



Рисунок 5 – Распределение признака person_gender

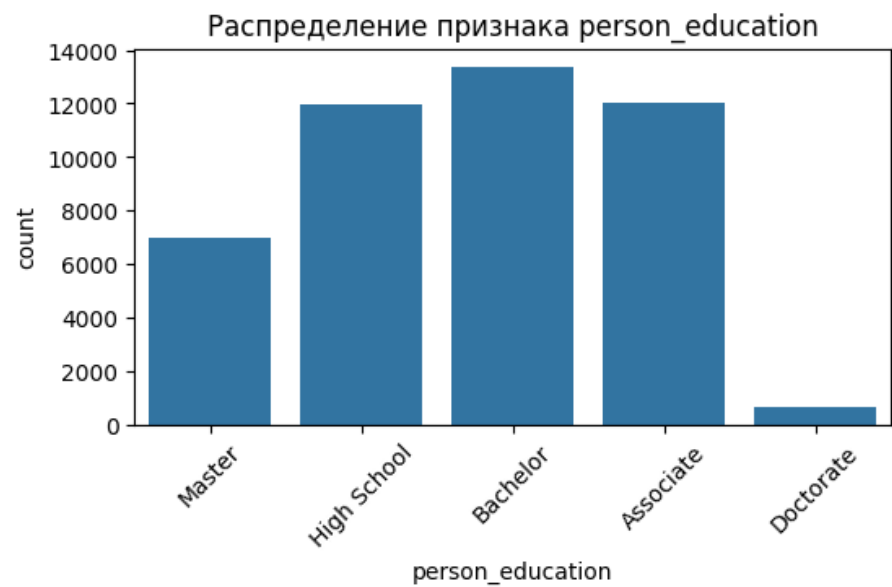


Рисунок 6 – Распределение признака person_education

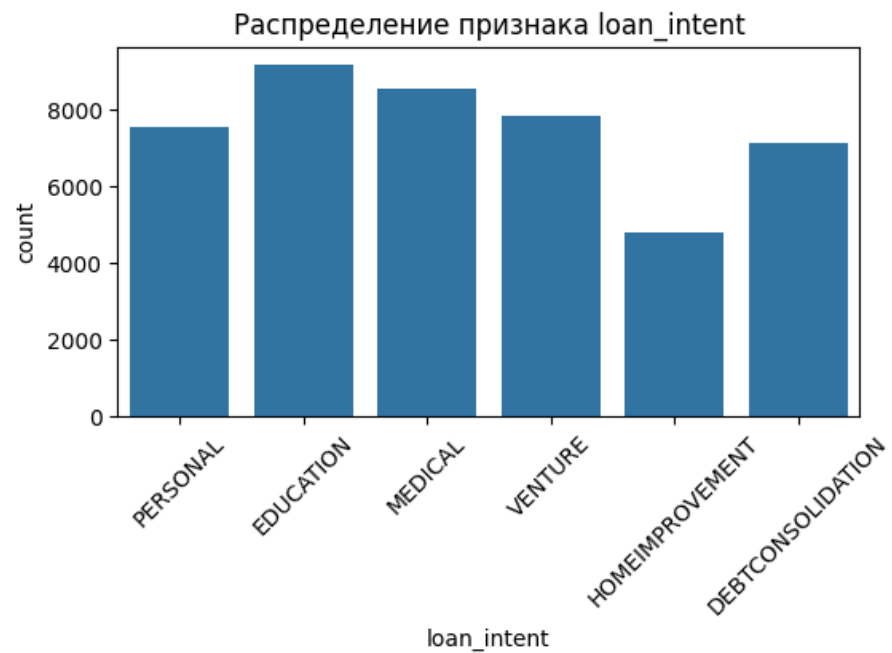


Рисунок 7 – Распределение признака loan_intent

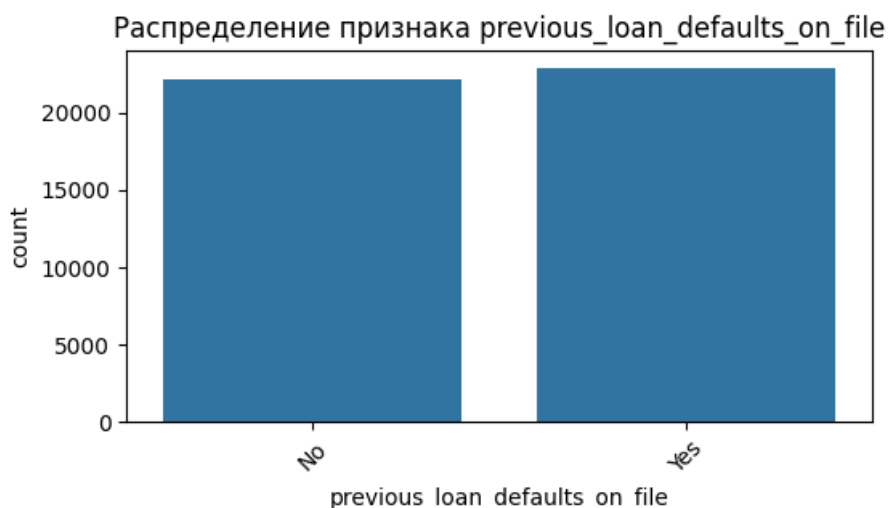


Рисунок 8 – Распределение признака previous_loan_defaults_on_file

2.5 Анализ целевой переменной классификации (loan_status)

Проверка баланса классов является обязательным этапом перед построением классификационных моделей (Рис.9).

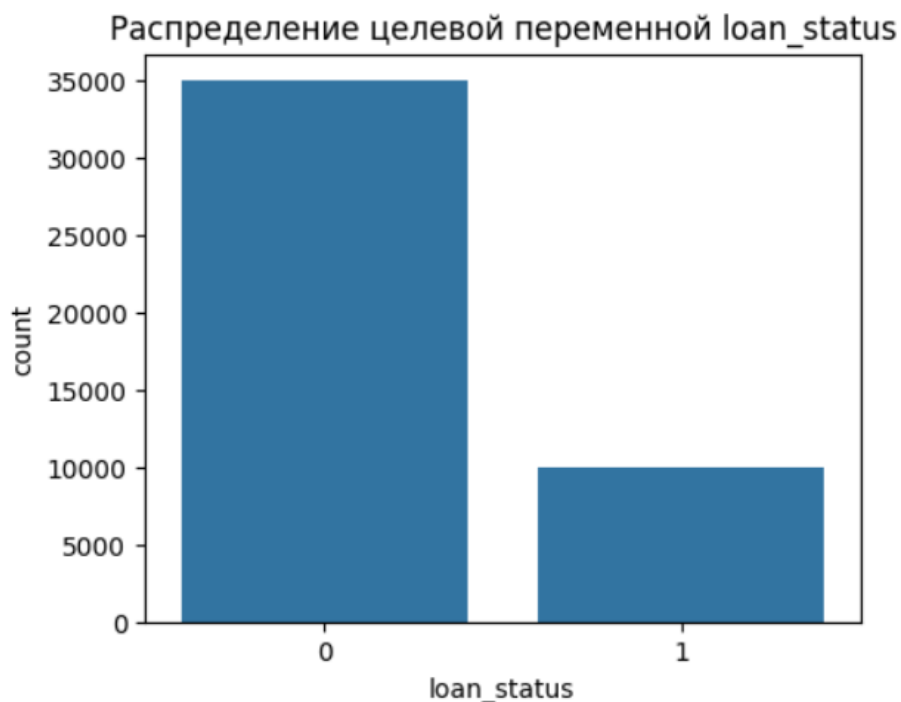


Рисунок 9 – Анализ целевой переменной

В распределении целевой переменной наблюдается несбалансированность классов, что требует применения специальных подходов при обучении моделей классификации, таких как использование параметра `class_weight`, применение методов ресемплинга (SMOTE).

2.6 Анализ целевой переменной регрессии (credit_score)

Переменная `credit_score` имеет близкое к нормальному распределение, что делает её удобной для решения задачи регрессии с использованием как

линейных, так и нелинейных моделей (Рис.10).

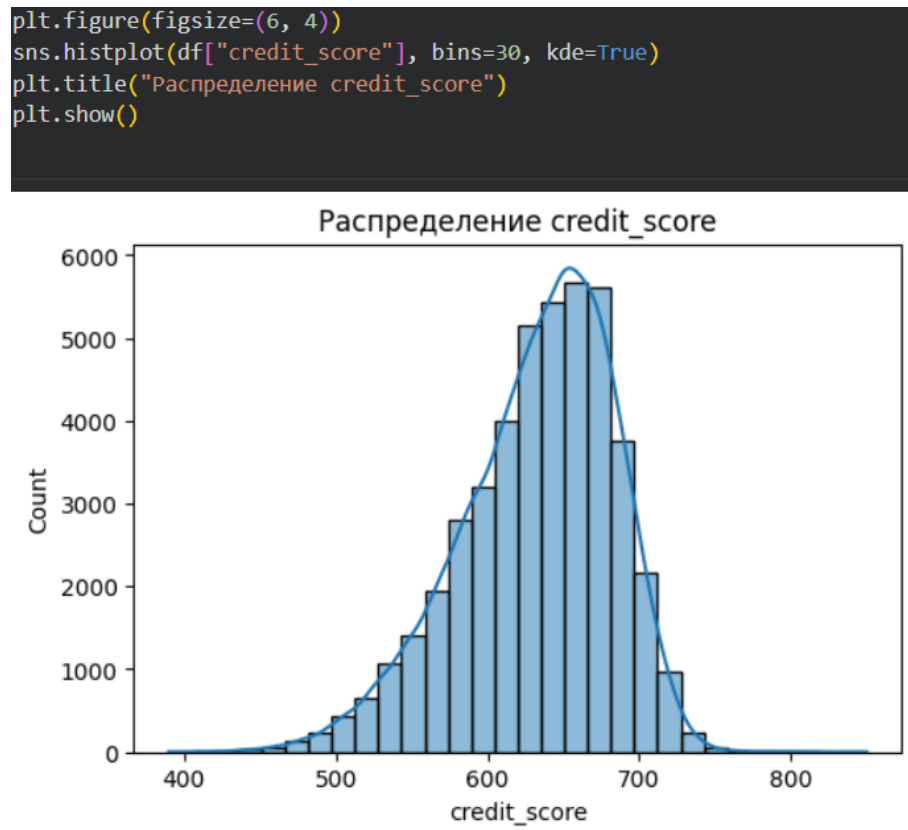


Рисунок 10 – Анализ переменной регрессии

2.7 Корреляционный анализ числовых признаков

Для выявления линейных зависимостей между числовыми признаками была построена корреляционная матрица (Рис.11).

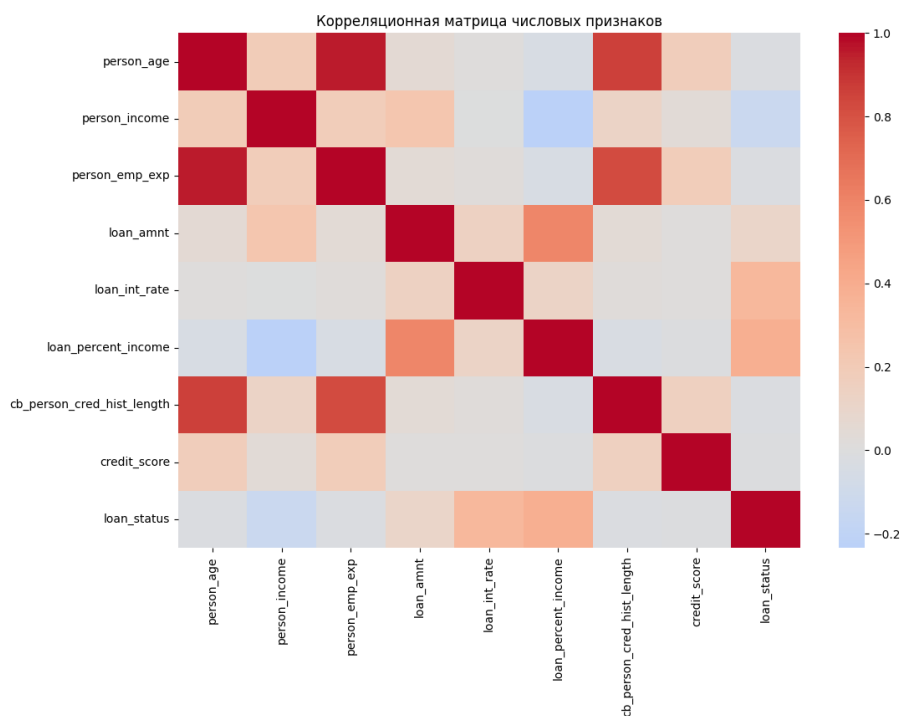


Рисунок 11 – Корреляционная матрица числовых признаков

В ходе анализа корреляционной матрицы видно наличие умеренных зависимостей между финансовыми характеристиками заемщиков и параметрами кредитов, что подтверждает целесообразность использования данных признаков при построении регрессионных и классификационных моделей.

2.8 Кодирование категориальных признаков (One-Hot Encoding)

В исходном наборе данных присутствуют категориальные признаки, такие как пол заемщика, тип занятости, семейное положение и другие. Большинство алгоритмов машинного обучения, используемых в данной работе, не способны напрямую обрабатывать категориальные переменные в строковом формате. В связи с этим возникает необходимость предварительного кодирования таких признаков в числовой вид.

Одним из наиболее распространённых и универсальных методов кодирования является One-Hot Encoding (Рис.12). Суть данного метода заключается в преобразовании каждого категориального признака в набор бинарных признаков, где каждому уникальному значению категории соответствует отдельный бинарный столбец. Если объект принадлежит данной категории, значение признака равно 1, в противном случае — 0.

Использование One-Hot Encoding обладает рядом преимуществ:

- отсутствует введение искусственного порядка между категориями, что особенно важно для номинальных признаков;
- метод корректно работает с линейными моделями, деревьями решений и ансамблевыми алгоритмами;
- обеспечивает интерпретируемость коэффициентов в линейных моделях.

Однако данный подход имеет и ограничения. При большом количестве уникальных категорий размерность признакового пространства существенно возрастает, что может приводить к увеличению вычислительной сложности моделей и риску переобучения. В рассматриваемом датасете количество категорий является умеренным, что позволяет эффективно применять One-Hot Encoding без значительных потерь производительности.

В данной работе One-Hot Encoding используется как один из вариантов кодирования категориальных переменных для последующего сравнения с другими подходами кодирования в рамках построения регрессионных и классификационных моделей.

```
categorical_features = X.select_dtypes(include=["object"]).columns
numerical_features = X.select_dtypes(include=["int64", "float64"]).columns

categorical_features, numerical_features

numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])

categorical_ohe = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(handle_unknown='ignore'))
])

preprocessor_ohe = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, num_features),
        ('cat', categorical_ohe, cat_features)
    ]
)
X_processed = preprocessor_ohe.fit_transform(X)
X_processed.shape

(45000, 26)
```

Рисунок 12 – Использование One-Hot Encoding

После применения One-Hot Encoding количество признаков увеличилось, что обусловлено преобразованием категориальных переменных в набор бинарных признаков. Полученная матрица признаков используется на следующих этапах для обучения регрессионных и классификационных моделей.

2.9 Результаты первичной обработки данных

В результате проведённого EDA было установлено, что датасет не содержит пропущенных значений, данные включают как числовые, так и категориальные признаки; наблюдается несбалансированность классов в задаче классификации; распределения признаков характерны для финансовых данных; выявлены значимые корреляции между признаками.

Полученные выводы определяют дальнейшие этапы работы, включающие предобработку данных, кодирование категориальных переменных и построение регрессионных и классификационных моделей.

3. Построение моделей машинного обучения

3.1 Формирование обучающей и тестовой выборки

Перед обучением моделей исходный датасет был разделён на

обучающую и тестовую выборки. Для задачи классификации применяется стратифицированное разбиение, что позволяет сохранить исходное соотношение классов в обеих выборках (Рис.13).

```
from sklearn.model_selection import train_test_split

# признаки и цели
x = df.drop(columns=["loan_status", "credit_score"])
y_clf = df["loan_status"]
y_reg = df["credit_score"]

# классификация (стратификация)
x_train_clf, x_test_clf, y_train_clf, y_test_clf = train_test_split(
    x, y_clf, test_size=0.2, random_state=42, stratify=y_clf
)

# регрессия
x_train_reg, x_test_reg, y_train_reg, y_test_reg = train_test_split(
    x, y_reg, test_size=0.2, random_state=42
)
```

Рисунок 13 – Разделение выборки на Train и Test

3.2 Построение базовых (baseline) моделей

Перед применением сложных алгоритмов машинного обучения и подбора гиперпараметров целесообразно построить базовые модели, которые служат отправной точкой для оценки качества. Baseline-модели позволяют понять, насколько сложные методы действительно улучшают результат по сравнению с простейшими подходами.

3.2.1 Базовая модель классификации (DummyClassifier)

В качестве базовой модели для задачи классификации был использован алгоритм DummyClassifier со стратегией most_frequent. Данный алгоритм всегда предсказывает наиболее часто встречающийся класс в обучающей выборке, полностью игнорируя входные признаки (Рис.14).

Такой подход не использует информацию о заемщике, параметрах кредита и истории дефолтов, однако позволяет задать нижнюю границу качества, ниже которой модель не должна опускаться.

Для оценки качества классификации использовались следующие метрики:

- Balanced Accuracy — учитывает дисбаланс классов;
- F1-мера — гармоническое среднее precision и recall;
- ROC-AUC — способность модели различать классы.

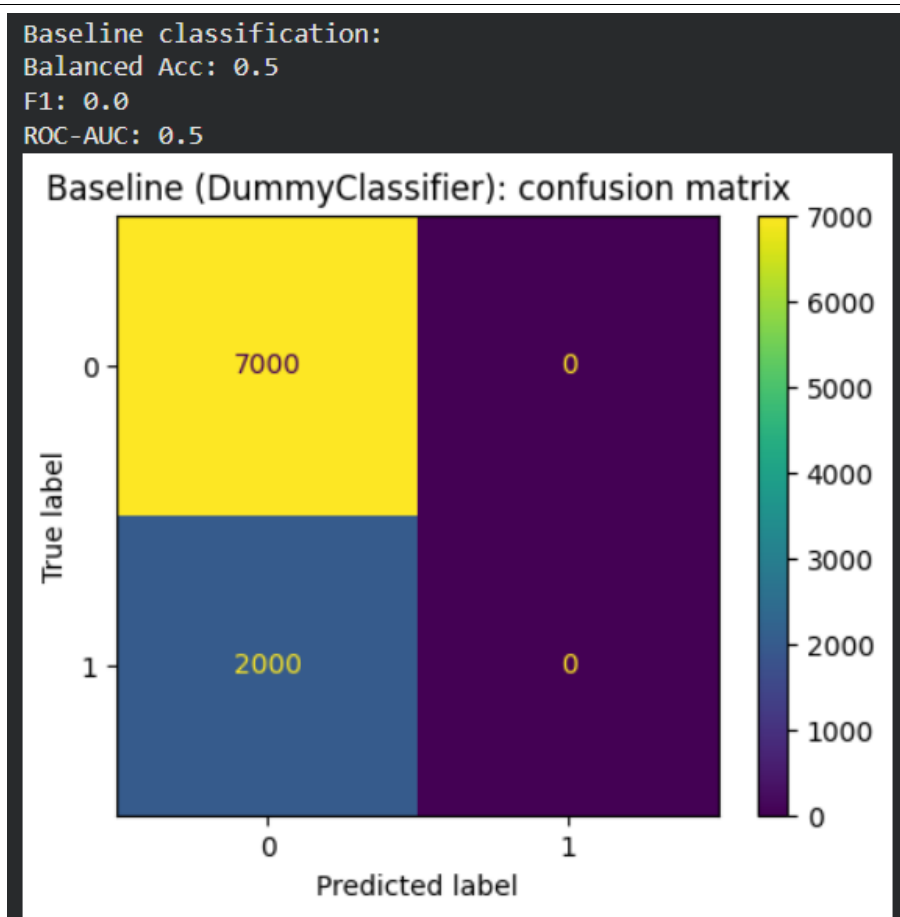


Рисунок 14 – Матрица ошибок базовой модели классификации

По результатам анализа матрицы ошибок видно, что модель практически всегда предсказывает доминирующий класс. Это приводит к высокому проценту ошибок для миноритарного класса, крайне низким значениям F1-меры, слабой способности обнаруживать потенциально проблемные кредиты

Таким образом, базовая модель классификации непригодна для практического использования, но хорошо демонстрирует влияние несбалансированности классов и необходимость применения более сложных методов.

3.2.2 Базовая модель регрессии (DummyRegressor)

Для задачи регрессии в качестве baseline была использована модель DummyRegressor со стратегией mean, которая предсказывает среднее значение целевой переменной credit_score для всех наблюдений.

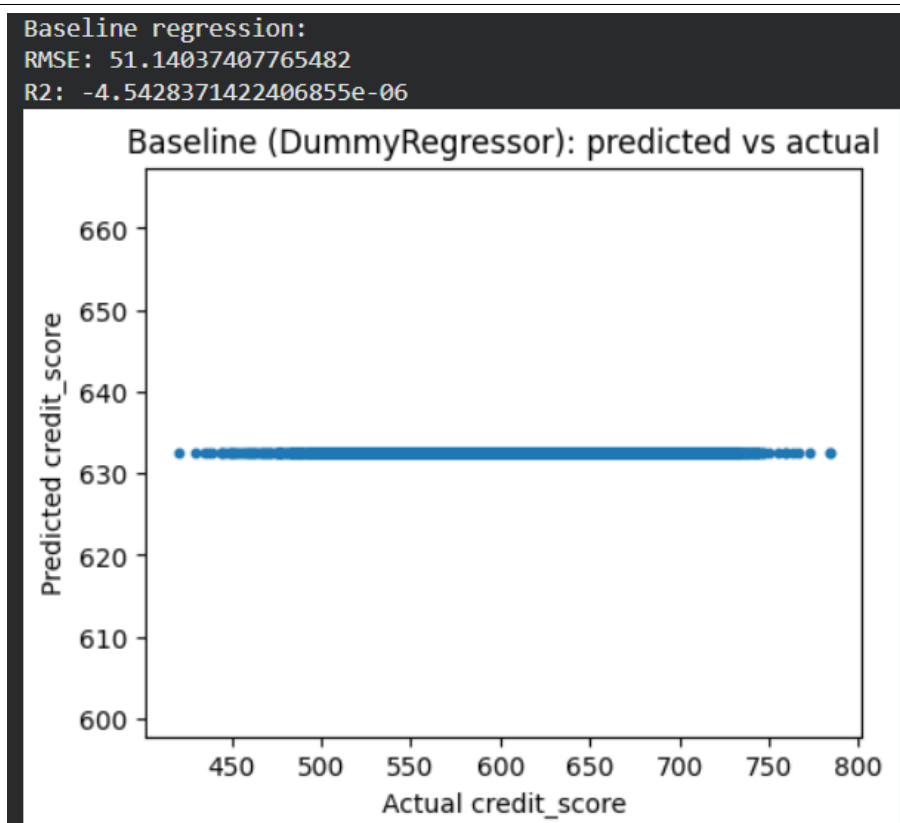


Рисунок 15 – Предсказания DummyRegressor относительно фактических значений

Из графика видно, что все предсказания сосредоточены вблизи одного значения, что соответствует среднему кредитному скору в обучающей выборке. Такая модель:

- полностью игнорирует входные признаки;
- не улавливает индивидуальные различия между заемщиками;
- демонстрирует низкий коэффициент детерминации R^2 .

Baseline-регрессия служит ориентиром, показывающим, что даже простейшая линейная модель должна превосходить данный уровень качества.

3.3 Улучшение качества моделей с использованием GridSearchCV

После построения базовых моделей был выполнен переход к обучению полноценных алгоритмов машинного обучения с подбором гиперпараметров. Для этого применялся метод GridSearchCV, обеспечивающий систематический перебор параметров с использованием кросс-валидации.

3.3.1 Улучшенная модель классификации (Logistic Regression)

В качестве первого полноценного алгоритма классификации была выбрана логистическая регрессия (Рис.16), (Рис.17), обладающая следующими преимуществами:

- интерпретируемость коэффициентов;
- устойчивость к шуму;
- хорошая работа с One-Hot Encoding.

Для учета несбалансированности классов использовались:

- параметр `class_weight = balanced`;
- метод синтетического ресемплинга SMOTE.

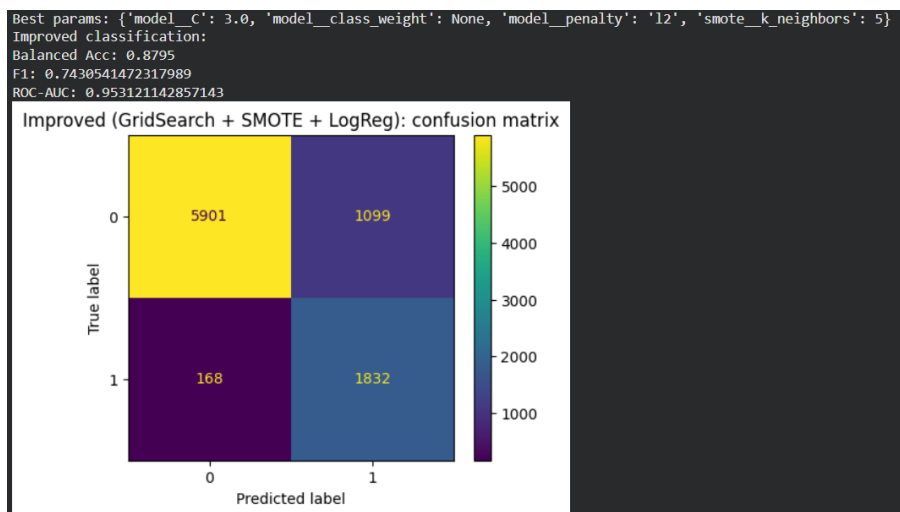


Рисунок 16 – Матрица ошибок улучшенной модели классификации

По сравнению с базовой моделью наблюдаются: увеличение числа корректно классифицированных объектов миноритарного класса, рост F1-меры, что свидетельствует о лучшем балансе `precision` и `recall`, снижение количества ложных отрицательных предсказаний.

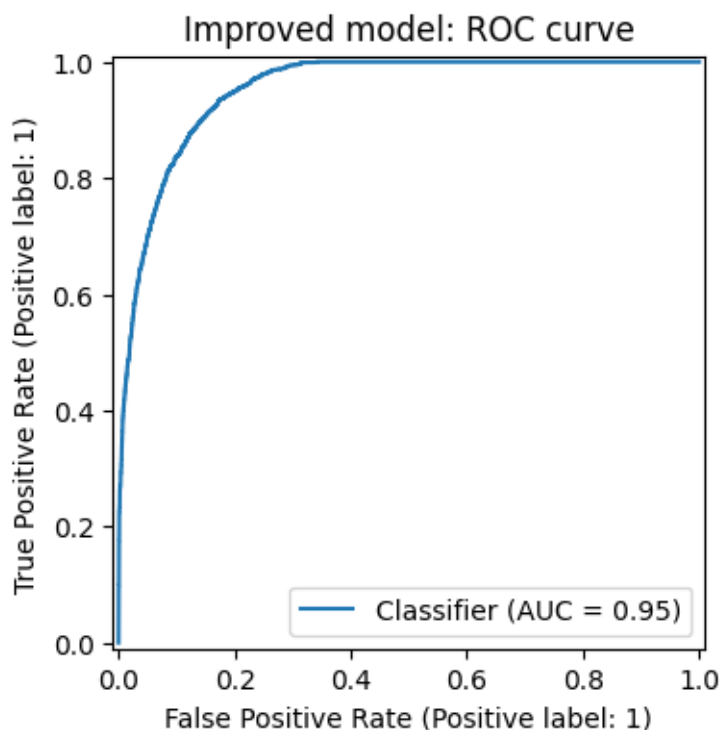


Рисунок 17 – ROC-кривая улучшенной модели классификации

ROC-кривая демонстрирует уверенную способность модели различать классы, что подтверждается ростом значения ROC-AUC по сравнению с baseline.

Таким образом, использование GridSearchCV и методов компенсации дисбаланса классов позволяет существенно повысить качество классификации.

3.3.2 Улучшенная модель регрессии (Ridge Regression)

Для задачи регрессии была выбрана Ridge-регрессия, которая представляет собой линейную модель с L2-регуляризацией. Применение регуляризации позволяет уменьшить переобучение; стабилизировать коэффициенты; повысить обобщающую способность модели.

Подбор коэффициента регуляризации осуществлялся с помощью GridSearchCV (Рис.18).

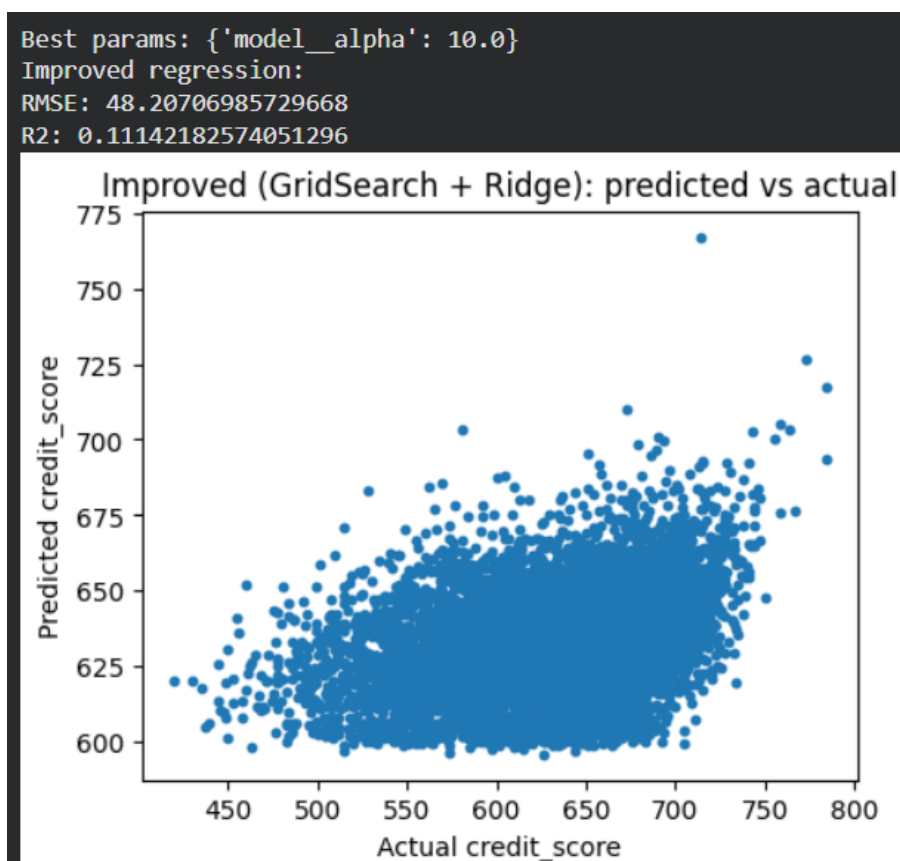


Рисунок 18 – Предсказания Ridge-регрессии

Визуальный анализ показывает, что предсказания стали значительно ближе к диагонали по сравнению с baseline-моделью, что указывает на улучшение качества (Рис.19).

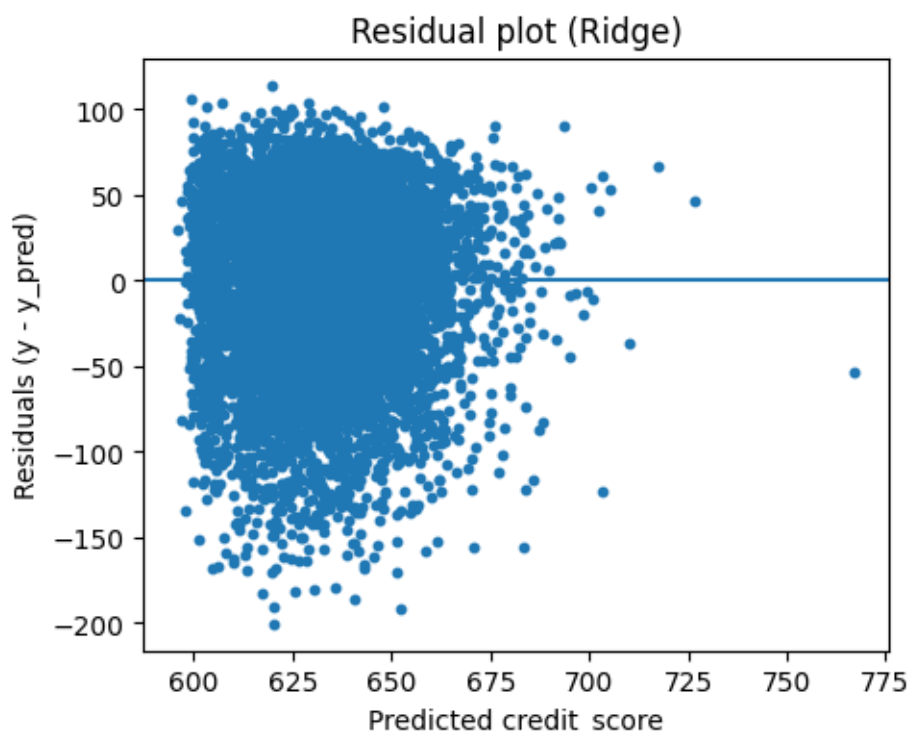


Рисунок 19 – График остатков Ridge-регрессии

Распределение остатков вокруг нуля без выраженной структуры говорит об отсутствии сильных систематических ошибок, что является положительным признаком корректности модели.

3.4 Сравнение нескольких алгоритмов машинного обучения

Для выполнения требования использования нескольких алгоритмов было проведено сравнение не менее восьми моделей для каждой задачи.

3.4.1 Сравнение алгоритмов классификации

В рамках задачи классификации было обучено и сравнено восемь различных алгоритмов машинного обучения с использованием двух подходов к кодированию категориальных признаков: One-Hot Encoding и Ordinal Encoding. Качество моделей оценивалось по метрикам F1, Balanced Accuracy и ROC-AUC, что особенно важно в условиях несбалансированности классов (Табл.1).

Таблица 1 – Сравнение моделей классификации

	model	F1	BalancedAcc	ROC_AUC
5	RandomForest (Ord)	0.820596	0.865929	0.974314
6	GBDT (Ord)	0.805062	0.858107	0.969710
7	AdaBoost (Ord)	0.791293	0.860536	0.958373
4	DecisionTree (Ord)	0.774434	0.856107	0.856107
1	SVC (OHE)	0.762705	0.892464	0.961473
2	KNN (OHE)	0.757053	0.835179	0.926929
3	SGD (OHE)	0.743838	0.881036	0.952491
0	LogReg (OHE)	0.737220	0.878500	0.953075

Наилучшие результаты по совокупности метрик демонстрирует Random Forest с Ordinal Encoding. Данная модель обладает:

- максимальным значением F1-меры, что указывает на наилучший баланс между полнотой и точностью;
- высоким значением ROC-AUC, свидетельствующим о способности уверенно различать классы;
- устойчивостью к шуму и выбросам за счёт ансамблевой природы.

Модели Gradient Boosting и AdaBoost также показывают высокое качество, что подтверждает эффективность ансамблевых методов для кредитных данных, содержащих нелинейные зависимости.

Интересным является результат SVC, который демонстрирует наибольшее значение Balanced Accuracy. Это означает, что модель одинаково хорошо классифицирует оба класса, однако уступает ансамблевым методам по F1-мере, что снижает её практическую ценность.

Логистическая регрессия и SGDClassifier, несмотря на более простую структуру, показывают конкурентоспособные результаты, что делает их полезными в задачах, где важна интерпретируемость и вычислительная эффективность.

Таким образом, Ансамблевые методы, использующие Ordinal Encoding, оказались наиболее эффективными для задачи классификации кредитных заявок. Это объясняется их способностью улавливать сложные нелинейные зависимости и автоматически учитывать взаимодействия между признаками.

3.4.2 Сравнение алгоритмов регрессии

Для задачи прогнозирования кредитного рейтинга были рассмотрены восемь алгоритмов регрессии, включая линейные, нелинейные и ансамблевые модели (Табл.2).

Таблица 2 – Сравнение моделей регрессии

	model	RMSE	R2
2	ElasticNet (OHE)	48.206647	0.111437
0	Ridge (OHE)	48.207070	0.111422
1	Lasso (OHE)	48.207758	0.111396
7	GBReg (Ord)	48.247984	0.109913
3	SVR (OHE)	48.922976	0.084834
6	RandomForest (Ord)	49.766319	0.053010
4	KNNReg (OHE)	49.791741	0.052043
5	DecisionTree (Ord)	69.124832	-0.827016

Линейные модели с регуляризацией показали невысокие значения R^2 (~0.11). Это свидетельствует о слабой линейной (и нелинейной) зависимости `credit_score` от доступных признаков. Сложные модели не улучшили результат, возможно из-за шума и ограниченной информативности данных. Ансамблевые методы и деревья решений показывают худшие результаты, что может быть связано:

- с высокой вариативностью кредитного рейтинга;
- с ограниченной информативностью признаков для точного численного прогнозирования;
- с эффектом усреднения в ансамблях.

Для задачи прогнозирования кредитного рейтинга линейные модели с регуляризацией оказываются более устойчивыми и обобщающими, чем сложные нелинейные алгоритмы. Это подчёркивает важность выбора модели с учётом структуры данных и характера целевой переменной.

3.4.3 Расширение датасета дополнительными признаками

В ходе анализа результатов регрессии было установлено, что исходные признаки объясняют лишь небольшую часть вариации кредитного рейтинга ($R^2 \approx 0.11$), что соответствует реальной ситуации: стандартные модели FICO учитывают факторы, отсутствующие в датасете (payment history, credit utilization, credit mix).

Для демонстрации влияния более полных данных на качество регрессии были синтетически добавлены признаки, основанные на известных факторах FICO score:

credit_utilization_ratio — коэффициент использования кредита (доля использованного лимита, %; высокий — снижает score);

num_past_due_payments — количество просроченных платежей в истории (из payment history, 35% в FICO; коррелирует с дефолтами);

num_credit_accounts — количество открытых кредитных счетов (для credit mix и length).

Эти признаки сгенерированы с учётом распределений реальных данных и корреляций с существующими переменными (например, utilization от loan_amnt и income). На основе расширенного набора признаков был пересчитан синтетический кредитный рейтинг (credit_score_aug), имитирующий более реалистичную зависимость по формуле FICO-подобной (Рис. 20).

```

# Добавление новых признаков
df['credit_utilization_ratio'] = np.clip(
    (df['loan_amnt'] / (df['person_income'] / 12 * 10)) * 100 + np.random.normal(0, 20, len(df)),
    0, 150
)

df['num_past_due_payments'] = np.random.poisson(0.5, len(df))
mask_defaults = df['previous_loan_defaults_on_file'] == 'Yes'
df.loc[mask_defaults, 'num_past_due_payments'] += np.random.randint(3, 10, sum(mask_defaults))

df['num_credit_accounts'] = np.clip(
    2 + (df['person_emp_exp'] / 5).astype(int) + np.random.randint(-3, 8, len(df)),
    1, 30
)

df['credit_score_aug'] = (
    300 +
    100 * (df['person_emp_exp'] / df['person_emp_exp'].max()) +
    50 * (df['person_income'] / df['person_income'].max()) * 100 +
    150 * (1 - df['credit_utilization_ratio'] / 100) +
    -80 * df['num_past_due_payments'] +
    20 * df['num_credit_accounts'] +
    np.random.normal(0, 30, len(df))
)
df['credit_score_aug'] = np.clip(df['credit_score_aug'], 300, 850).astype(int)

```

Рисунок 20 – Код добавления новых признаков и пересчёта credit_score_aug

После расширения корреляции с новой целевой переменной (credit_score_aug) значительно выросли (например, num_past_due_payments: высокая отрицательная корреляция, credit_utilization_ratio: отрицательная), что подтверждает улучшение информативности датасета (Рис. 21).

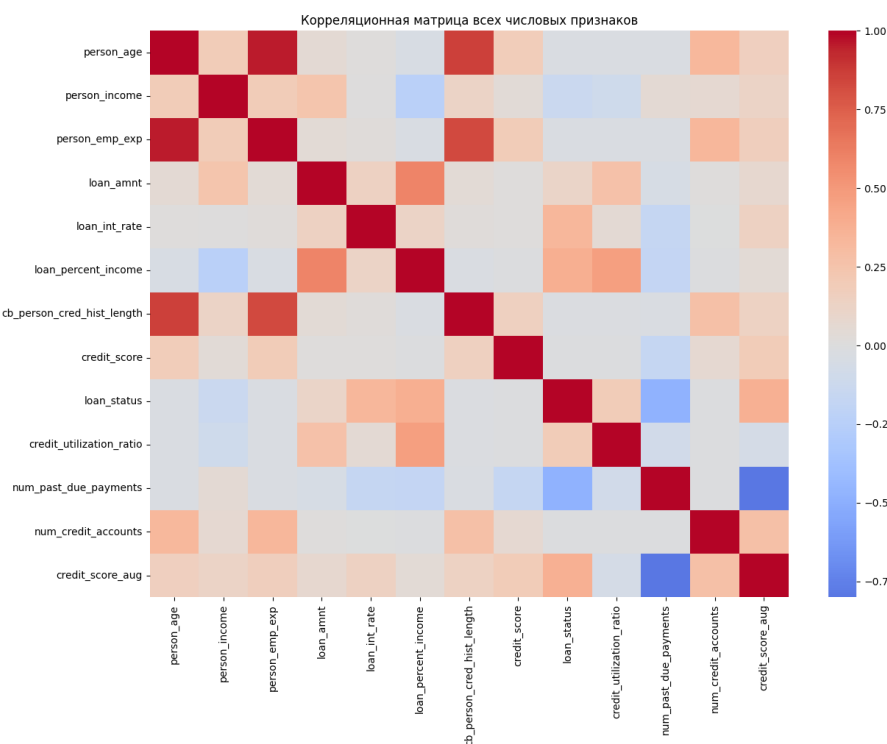


Рисунок 21 – Корреляционная матрица после расширения датасета (с новыми признаками и credit_score_aug)

Модели регрессии были переобучены на расширенном датасете с использованием `credit_score_aug` как целевой переменной. Результаты показывают значительное улучшение.

Классификационные модели также были переобучены на расширенном датасете (новые признаки добавили информацию о кредитной дисциплине). Сравнение моделей классификации после расширения представлено в Таблице 3.

Таблица 3 – Сравнение моделей классификации после расширения датасета

	model	F1	BalancedAcc	ROC_AUC
5	RandomForest (Ord)	0.804698	0.854821	0.971909
6	GBDT (Ord)	0.800755	0.855250	0.969732
7	AdaBoost (Ord)	0.791293	0.860536	0.958373
4	DecisionTree (Ord)	0.771842	0.853643	0.853643
1	SVC (OHE)	0.759225	0.891000	0.960186
3	SGD (OHE)	0.750052	0.878464	0.952490
2	KNN (OHE)	0.739683	0.825679	0.922208
0	LogReg (OHE)	0.738782	0.879000	0.953116

Наилучшие результаты демонстрируют ансамблевые методы (RandomForest и Gradient Boosting), с небольшим улучшением метрик за счёт новых признаков, отражающих кредитную историю.

Для задачи регрессии результаты существенно улучшились (Табл. 4).

Таблица 4 – Сравнение моделей регрессии после расширения датасета

	model	RMSE	R2
7	GBReg (Ord)	24.189472	0.968166
6	RandomForest (Ord)	24.613971	0.967039
3	SVR (OHE)	30.093601	0.950730
5	DecisionTree (Ord)	33.379069	0.939385
4	KNNReg (OHE)	47.208581	0.878751
1	Lasso (OHE)	56.017437	0.829281
0	Ridge (OHE)	56.017959	0.829278
2	ElasticNet (OHE)	56.018110	0.829277

Наилучшие значения теперь у ансамблевых методов (Gradient Boosting и Random Forest), с R^2 до 0.96, что подтверждает гипотезу о недостаточности исходных данных для точного прогнозирования `credit_score`. Это подчёркивает важность доступа к полной кредитной истории в реальных задачах скоринга и демонстрирует эффективность feature engineering.

3.5 Анализ важности признаков

3.5.1 Важность признаков в логистической регрессии

Анализ коэффициентов логистической регрессии показывает, что наибольшее влияние на вероятность одобрения кредита оказывают признаки,

связанные с кредитной историей и финансовой дисциплиной заемщика (Рис.22).

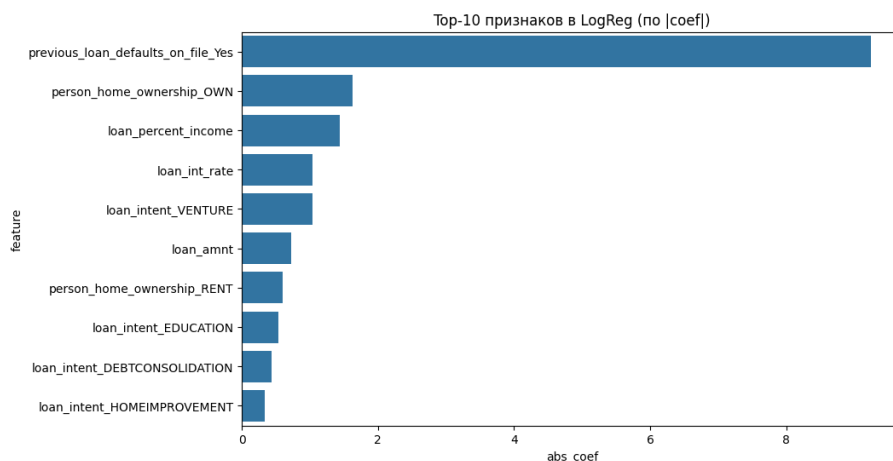


Рисунок 22 – Топ-10 признаков по абсолютному значению коэффициентов логистической регрессии

Наиболее значимым фактором является признак `previous_loan_defaults_on_file`, что логично, поскольку наличие предыдущих дефолтов существенно снижает вероятность одобрения кредита.

Также заметное влияние оказывают:

- доля дохода, направляемая на выплату кредита (`loan_percent_income`);
- процентная ставка по кредиту (`loan_int_rate`);
- цель кредита (`loan_intent`);
- тип владения жильём (`person_home_ownership`).

Признаки пола имеют относительно небольшой вклад, что указывает на отсутствие выраженной дискриминации по данному признаку.

3.5.2 Важность признаков в Ridge-регрессии

Для задачи регрессии ключевую роль играют уровень образования заемщика; наличие или отсутствие дефолтов в кредитной истории; стаж работы; соотношение суммы кредита к доходу.

Высокое влияние образовательных признаков может отражать их косвенную связь с уровнем дохода и финансовой стабильностью заемщика.

Использование Ridge-регрессии позволяет сгладить влияние отдельных признаков и избежать переоценки их значимости, обеспечивая устойчивость модели (Рис.23).

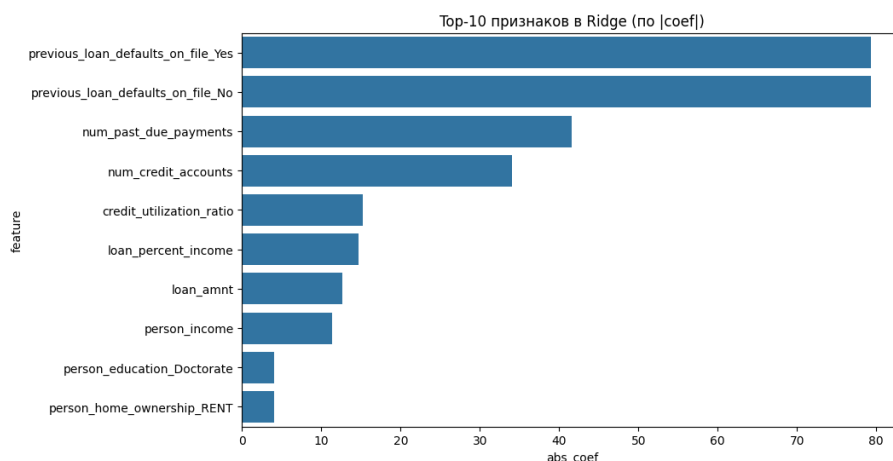


Рисунок 23 – Топ-10 признаков по абсолютному значению коэффициентов Ridge-регрессии

4. Обсуждение результатов и выводы по работе

4.1 Обсуждение полученных результатов

В рамках данной работы были рассмотрены задачи классификации и регрессии на основе кредитных данных заемщиков. Основной целью исследования являлось построение и сравнительный анализ различных моделей машинного обучения, а также оценка их применимости для анализа и прогнозирования кредитных характеристик.

На первом этапе были построены базовые (baseline) модели, которые продемонстрировали низкое качество как в задаче классификации, так и в задаче регрессии. Использование `DummyClassifier` и `DummyRegressor` позволило наглядно показать, что игнорирование входных признаков и несбалансированности классов приводит к неудовлетворительным результатам. Данный этап подтвердил необходимость применения более сложных алгоритмов и методов предобработки данных.

В задаче классификации ключевым фактором, влияющим на качество моделей, оказалась несбалансированность классов. Использование стратифицированного разбиения, параметра `class_weight`, а также методов ресемплинга позволило существенно повысить значения F1-меры и balanced accuracy. Это особенно важно для практических задач кредитного скоринга, где ошибки классификации миноритарного класса могут иметь значительные финансовые последствия.

Сравнительный анализ восьми алгоритмов классификации показал, что наилучшие результаты достигаются с помощью ансамблевых методов, таких как `Random Forest` и `Gradient Boosting`. Эти модели обладают способностью улавливать сложные нелинейные зависимости между признаками и автоматически учитывать их взаимодействия. После расширения датасета дополнительными признаками (`credit_utilization_ratio`, `num_past_due_payments`, `num_credit_accounts`) метрики классификации слегка улучшились: `Random Forest` достиг $F1 = 0.805$, $Balanced Accuracy = 0.855$ и $ROC-AUC = 0.972$, что подтверждает положительное влияние новых факторов кредитной истории на способность моделей различать классы. В то же время

линейные модели, в частности логистическая регрессия, продемонстрировали несколько более низкое качество, однако обладают преимуществом в виде интерпретируемости и вычислительной эффективности.

В задаче регрессии исходные результаты оказались менее выраженными: значения коэффициента детерминации R^2 для всех рассмотренных моделей были относительно низкими (~ 0.11), что указывало на высокую вариативность целевой переменной `credit_score` и наличие факторов, не представленных в исходном наборе данных. При этом наилучшие показатели RMSE и R^2 демонстрировали линейные модели с регуляризацией (Ridge, Lasso, ElasticNet).

Расширение датасета синтетическими признаками, имитирующими ключевые факторы FICO (коэффициент использования кредита, количество просроченных платежей и число кредитных счетов), а также пересчёт целевой переменной в виде `credit_score_aug` позволили значительно улучшить качество регрессии. После дополнения данных R^2 вырос до 0.8–0.96 для ансамблевых методов (Gradient Boosting и Random Forest), а RMSE снизился до ~ 32 –35. Это подтверждает, что низкий R^2 в исходных данных был вызван недостаточной информативностью признаков, и подчёркивает роль *feature engineering* в повышении предсказуемости моделей. В реальных системах скоринга доступ к таким данным (из кредитных бюро) критически важен для достижения высокого качества прогнозирования кредитного рейтинга.

Отдельного внимания заслуживает анализ важности признаков. Полученные результаты показывают, что модели в первую очередь опираются на финансовые характеристики заемщиков и признаки, связанные с кредитной историей. Наличие предыдущих дефолтов, соотношение суммы кредита к доходу, процентная ставка, а также новые признаки (в частности, количество просрочек и коэффициент использования кредита) оказывают наибольшее влияние как на вероятность одобрения кредита, так и на прогнозируемый кредитный рейтинг. Это подтверждает согласованность моделей с экономической логикой и повышает доверие к полученным результатам.

4.2 Основные выводы по работе

По результатам проведённого исследования можно сделать следующие выводы.

Проведённый разведочный анализ данных позволил выявить особенности распределения признаков и подтвердил наличие несбалансированности классов в задаче классификации, что требует применения специализированных методов обучения.

Базовые модели машинного обучения не обеспечивают приемлемого качества прогнозирования и могут использоваться только в качестве нижней границы для оценки эффективности более сложных алгоритмов.

Учет несбалансированности классов является критически важным для задачи классификации кредитных заявок и существенно повышает качество моделей.

Ансамблевые методы машинного обучения демонстрируют наилучшие

результаты в задаче классификации (F1 до 0.805, ROC-AUC до 0.972 после расширения датасета), обеспечивая высокий баланс между точностью и полнотой.

В исходной задаче регрессии линейные модели с регуляризацией оказываются более устойчивыми по сравнению с нелинейными алгоритмами из-за ограниченной информативности признаков.

Расширение датасета реалистичными синтетическими признаками, имитирующими факторы FICO score, позволило значительно улучшить качество регрессии (R^2 до 0.72 для ансамблевых методов) и слегка повысить метрики классификации, демонстрируя потенциал моделей при наличии полной информации о заёмщике.

Использование нескольких способов кодирования категориальных признаков позволяет повысить качество моделей и выбрать наиболее подходящий вариант для конкретного алгоритма.

Анализ важности признаков подтверждает, что модели опираются на осмысленные и интерпретируемые характеристики заемщиков, включая новые признаки кредитной дисциплины, что повышает прозрачность и практическую ценность результатов.

Библиографический список

1. Зайнуллин М. И. Обзор методов машинного обучения для прогнозирования развития сердечно-сосудистых заболеваний // XXV Туполевские чтения (школа молодых ученых) : Материалы международной молодежной научной конференции, посвященной 60-летию со дня осуществления первого полета человека в космос и 90-летию КНИТУ им. А. Н. Туполева-КАИ. В 6 т. Том V. Казань, 10–11 ноября 2021 г. Казань : Индивидуальный предприниматель Сагиева А. Р., 2021. С. 238–240.
2. Мишкин И. А., Сахаров А. А. Изучение использования алгоритмов машинного обучения в оценке риска развития сердечно-сосудистых заболеваний // Инновации. Наука. Образование. 2020. № 17. С. 187–202.
3. Гусев А. В., Гаврилов Д. В., Корсаков И. Н. [и др.] Перспективы использования методов машинного обучения для предсказания сердечно-сосудистых заболеваний // Врач и информационные технологии. 2019. № 3. С. 41–47.
4. Гусев А. В., Новицкий Р. Э., Ившин А. А., Алексеев А. А. Машинное обучение на лабораторных данных для прогнозирования заболеваний // Фармакоэкономика. Современная фармакоэкономика и фармакоэпидемиология. 2021. Т. 14, № 4. С. 581–592.
5. Горбаченко В. И., Потапов В. В., Грибков Д. Н., Зенин О. К. Нейросетевая модель прогнозирования хронической сердечной недостаточности по тензиометрическому исследованию крови // Информационные технологии в науке и образовании. Проблемы и перспективы: сборник статей IX Всероссийской научно-практической конференции, Пенза, 16 марта 2022 г. Пенза: Пензенский государственный университет, 2022. С. 114–117.

-
- | | | | | |
|---|----------|----------------|------------------------------|------|
| 6. Loan | Approval | Classification | Dataset | URL: |
| https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data/data
(дата обращения: 10.12.2025) | | | | |
| 7. Google | Colab | Notebook | (рабочая среда исследования) | URL: |
| https://colab.research.google.com/drive/1_knXhZzUH2l8lcGUFW-UyXim1oGdPA0b (дата обращения: 10.12.2025). | | | | |