Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Исследовательский проект на тему

«Кредитный скоринг. Влияние методов кодирования признаков на интерпретацию сложных моделей машинного обучения»

Выполнила студентка группы 194, 4 курса,
Смирнова Анна Романовна

Руководитель ВКР:
Старший преподаватель Титова Наталия Николаевна

Москва 2023

Содержание

Аннотация

Машинное обучение оказалось мощным инструментами для прогнозирования кредитоспособности, однако, понимание факторов, которые способствуют получению таких прогнозов, имеет решающее значение для эффективного принятия решений и управления рисками.

В данной работе исследуется, придают ли различные методы кодирования разную предсказательную способность одинаковым признакам, что потенциально может повлиять на решения по оценке кредитного риска и способствовать финансовым ошибкам.

Результаты экспериментов показывают, что методы кодирования действительно приводят к различным интерпретациям важности признаков. В заключении дается количественная оценка финансовых последствий потери одного клиента, что подчеркивает последствия, связанные с неточной оценкой рисков из-за ошибочной интерпретации модели.

Ключевые слова — кредитный скоринг, методы кодирования, SHAP, LIME

Abstract

Machine learning has proven to be a powerful tool for predicting creditworthiness, but understanding the factors that contribute to such predictions is crucial for effective decision-making and risk management.

This paper investigates whether different encoding methods impart different predictive power to the same features, potentially affecting credit risk decisions and contributing to financial errors.

The experimental results show that the encoding methods do lead to different interpretations of the importance of the attributes. Finally, the financial consequences of the loss of one customer are quantified, highlighting the consequences associated with inaccurate risk assessment due to erroneous model interpretation.

Keywords — credit scoring, encoding methods, SHAP, LIME

1 Введение

Кредитный скоринг — это процесс оценки кредитоспособности физических лиц, предприятий или других организаций на основе их кредитной истории, текущего финансового состояния и других соответствующих факторов. Он используется кредиторами, финансовыми учреждениями и другими организациями для определения вероятности погашения заемщиком кредита или выполнения других финансовых обязательств.

Исторически эта вероятность определялась кредитным экспертом, но со временем количество данных увеличилось кратно, и человеку стало невозможно учесть все закономерности.

Эпоха принятия решений на основе данных привела к тому, что модели машинного обучения стали неотъемлемой частью различных секторов, включая финансовую отрасль. В частности, кредитный скоринг значительно выиграл от появления таких моделей. Однако организациям недостаточно знать только лишь предсказание, конечной задачей для финансового бизнеса всегда является минимизация убытков или максимизация прибыли. Для достижения этих целей важно понимать, какие признаки оказывают наибольшее влияние на предсказание дефолта.

Совершенствование методов интерпретации становятся все популярнее и сложнее в последнее десятилетие, поэтому исследования в этой области являются особенно актуальными. Эта работа предоставляет новый взгляд на влияние методов кодирования на прогностическую силу признаков при оценке кредитоспособности. Влияние предобработки часто упускают из виду при анализе результатов, поэтому особенно важно изучить потенциальные ошибки, возникающие из-за различных методов кодирования.

Для достижения цели были отобраны различные методы кодирования, включая OneHot, Helmert, Frequency Encoding и Weight of Evedence преобразование, которые предстоит попробовать на различных моделях машинного обучения и сравнить интерпретации важности признаков.

Используя SHAP и LIME в качестве методов интерпретации, исследование показало, что методы кодирования признаков могут привести к расхождению в интерпретации после обучения на одной и той же модели. Кроме того, в работе проводится количественная оценка финансовых последствий потери одного клиента, что подчеркивает потенциальные последствия неточной оценки риска, вызванной неправильной интерпретацией модели.

Структура работы представлена далее. Раздел «Обзор литературы» содержит обоснование выбора моделей, метрик и способов кодирования для исследования, а также последние крупные обзорные статьи о кредитном скоринге. «Основная часть» включает в себя описание используемых данных, и полную теоретическую часть. В главе «Эксперименты» описаны результаты обучения моделей, а также комментарии к интерпретации результатов в важных для исследования случаях. «Практическая применимость» подробно раскрывает, как ложная интерпретация влияет на финансовые показатели бизнеса и включает тестовые подсчеты ущерба. Далее «Заключение», «Список литературы» и «Приложения», с необходимыми дополнениями графиков, описания данных, и дальнейших векторов работы.

2 Обзор литературы

В области влияния разных методов на интерпретацию в контексте кредитного скоринга проведено не так много исследований, поэтому опорой для обзора служат статьи по основным методам, применяемым в кредитном скоринге в целом. Классические решения — например, логистическая регрессия с WoE-преобразованием — берутся за основу во многих исследованиях, однако спектр направлений, которые помогают улучшить качество принимаемых решений, стремительно растет.

Методы машинного обучения значительно повысили точность оценки рисков в кредитном скоринге. В связи с этим особенно подчеркивается роль методов интерпретации, таких как SHAP [1] и LIME [2], в укреплении доверия к модели и в понимании особенностей, влияющих на результаты [3]. SHAP, например, помогает пользователям понимать предсказания, подчеркивается необходимость более быстрых методов оценки для конкретного типа модели и интеграции работы по оценке эффектов взаимодействия из теории игр.

Говоря о моделях, существуют достаточно широкие исследования [4], доказывающие превосходство ансамблевых моделей над одиночными, при этом выигрывая даже у сложных архитектур нейронных сетей, при этом глубокие сети с большим количеством скрытых слоев не превосходят более мелкие сети с одним скрытым слоем [5]. На основе этого, для поддержания «объективности» было решено добавить в исследование по модели каждого типа: логистическую регрессию, XGBoost и несложную архитектуру Sequential. Метрики были выбраны на основе исследования, обозревающего более 70 статей о кредитном скоринге [4].

В настоящее время определено три типа методов кодирования категориальных данных [6]: детерминированные - методы преобразования категориальных данных в векторы с низкой вычислительной сложностью, алгоритмические - более сложные, требующие объемных вычислений, автоматические - методы, использующие алгоритмы машинного обучения для кодирования

данных. Для исследования были взяты только первые, в силу особенностей используемых данных: OneHot и Helmert Encoding и Frequency. Однако, нельзя было упустить самый используемый метод для кредитного скоринга Weight Of Evedence [5] преобразование, поэтому был использован и он, в комбинации с различной предобработкой данных.

В заключение, данный обзор литературы подчеркивает формирующуюся потребность в более четком понимании интерпретируемости моделей, в частности для финансового сектора, также обоснован выбор моделей, метрик и методов кодирования для проведения экспериментов.

3 Основная часть

3.1 Данные

Все обучение проводилась на одном наборе данных: **Home Credit Default Risk** [7]. Датасет представляет собой маскированные данные о клиентах Home Credit — финансовой организации в Чешской Республике — и истории их обращений за кредитами. Набор данных содержит различные таблицы, каждая из которых содержит уникальную информацию, такую как социально-демографические данные клиентов, предыдущие заявки на кредит, истории платежей и т.д. В обучении в основном использовалась только одна из таблиц, содержащая данные из заявки на кредит.

Основная цель соревнования, в которых использовался данный набор - предсказать, будет ли одобрен или отклонен кредит заявителя, возникнут ли у него трудности с погашением или он объявит дефолт. Для достижения этой цели использовался набор данных, состоящий из обучающего и тестового множеств, содержащих около 307 000 и 47 000 наблюдений соответственно. Каждый набор содержит ряд независимых характеристик, которые могут быть использованы для прогнозирования дефолта заемщика по кредиту.

Данные также включают такие признаки как демографические характеристики, остатки по кредитам и историю транзакций. Home Credit предварительно обработал все данные для удаления информации, позволяющей установить личность, чтобы обеспечить конфиденциальность.

Этот датасет представляет собой ценную возможность для оценки вероятности невозврата кредита, тем самым предоставляя финансовым учреждениям, таким как Home Credit, важную информацию для эффективного и действенного управления рисками.

3.2 Методы кодирования

В машинном обучении кодирование — это процесс преобразования категориальных признаков в числовые значения, которые могут быть использованы в качестве входных данных для алгоритмов.

Категориальные признаки могут быть нескольких типов, рассмотрим их на примере наших данных:

- Номинальные признаки — это признаки, которые не имеют присущего им упорядочивания или ранжирования. Пример:
  NAME_FAMILY_STATUS с вариантами «Married», «Single», «Civil marriage» и другими.

- Порядковые признаки, с присущим им ранжированием. Пример:
  NAME_EDUCATION_TYPE с вариантами «Secondary», «Higher education», «Lower secondary» и другими, которые можно ранжировать.

- Бинарные признаки только с двумя значениями. Пример:
  NAME_CONTRACT_TYPE с вариантами «Cash loans» и «Revolving loans».

Выбор методов кодирования в нашем случае будет на основе того, насколько он подходит к имеющимся категориальным признакам в зависимости от их типа, и возможность корректно применять метод в задаче бинарной классификации. Перейдем к описанию методов, которые были отобраны по этим критериям.

3.2.1 Weight of evidence

Weight of Evidence (WoE) - это преобразование, отражающее связь между категориальной и целевой переменными путем сравнения вероятности дефолта (1) и выплаты (0) по каждой категории. Кодирование WoE особенно полезно, когда целевая переменная не сбалансирована, и в категориальной

переменной много уникальных значений.

$$\text{WoE}_{\,category} = \ln\left(\frac{\left(\frac{n_1}{N}\right)}{\left(\frac{n_0}{N}\right)}\right) \tag{1}$$

Где N - общее количество значений этой категории, n_0 - количество выплат в категории, n_1 - количество дефолтов. Таким образом, при кодировании каждый категориальный признак заменяется его WoE-значением. При необходимости уменьшения уникальных значений признака можно объединять значения с близким WoE, поскольку это означает их схожую предсказательную способность.

WoE-преобразование можно использовать и для численных признаков. Для этого необходимо предварительно разбить данные на бины (группы), для каждого из которых отдельно вычислить свое значение WoE. Количество записей в бинах регулируется общим числом данных и может варьироваться, важно следить за тем, чтобы в каждой группе присутствовали оба события, и объем каждого их них был более 5%.

Аналогично с преобразованием категориальных признаков, близкие по WoE значению группы могут заменять друг друга, не имея выраженной предсказательной способности. В этом случае для формулы (1): N - количество записей в бине, n_0 - количество выплат в бине, n_1 - количество дефолтов в бине.

Таким образом, существует несколько вариаций применения данного преобразования: как способ кодирования категориальных признаков, обработка пропущенных значений, способ уменьшения размерности, кодирование всех данных. Кроме того, WoE-значения можно использовать для оценки значимости признака. Для этого его применяют в вычислении Information Value.

Information Value — один из способов оценки значимости признака. Вычисляясь на основе WoE-значения, он отражает коэффициент полезности

признака и его значимость для предсказания.

$$IV = \sum_{i=1}^{N} (WoE_i) * ((\frac{n_0}{N}) - (\frac{n_1}{N})) \tag{2}$$

В формуле (2) вычисляется общая предсказательная способность: N - количество всех записей, WoE_i - значение WoE для i-ой категории или бина, n_0 и n_1 - количество выплат и дефолтов в i-ой категории или бине соответственно.

Для оценки значимости каждого отдельного признака мы смотрим на его IV отдельно. Согласно «Credit Risk Scorecards» [8], оценка IV в задаче кредитного скоринга может трактоваться следующими значениями:

Таблица 3.1: Оценка важности признака по IV значению

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive power
0.1 to 0.3	Medium predictive power
0.3 to 0.5	Strong predictive power
More than 0.5	Suspicious predictive power

Таким образом, Information value можно использовать для того, чтобы отобрать только значимые для предсказания признаки, а так же избежать переобучения и утечки целевой переменной.

3.2.2 OneHot Encoding

OneHot Encoding — это один из классических методов кодирования для представления категориальных данных в числовом формате. Принцип работы прост: каждая уникальная категория представляется в виде нового столбца в данных, состоящего из 0 и 1, где 1 означает принадлежность записи к этой категории, а 0 не принадлежность соответственно.

Кроме того, OneHot Encoding популярен в задачах кредитного скоринга, поскольку многие банковские данные о заемщиках являются категориальны-

ми. Например: статус занятости, уровень образования, цель кредита, личное положение и другие.

Пример кодирования признака, у которого 3 уникальных значения:

ID	Color
One	Green
Two	Blue
Three	Yellow
Four	Green

Таблица 3.2: До OneHot

ID	Green	Blue	Yellow
One	1	0	0
Two	0	1	0
Three	0	0	1
Four	1	0	0

Таблица 3.3: После OneHot

При работе с методом стоит учитывать, что:

- OneHot Encoding может значительно увеличить размерность данных, что может привести к переобучению, а также быть вычислительно дорогим
- Если переменная имеет много уникальных категорий с небольшим количеством наблюдений в каждой, OneHot Encoding может привести к разреженным данным, с которыми трудно эффективно работать моделям машинного обучения.

3.2.3 Helmert Encoding

Helmert Encoding (или кодирование Гельмерта) - это еще одна техника, используемая в машинном обучении для представления категориальных данных в виде числовых. Кодирование происходит следующим образом: K уникальных значений категории представляются матрицей $(K, K-1)$. На главной диагонали и над ней проставляются единицы со знаком минус, а под главной диагональю - порядковый номер значения. Все ниже порядковых номеров заполняется нулями. При этом, перед кодированием происходит сортировка уникальных значений.

Кодирование Гельмерта не так популярно в задачах кредитного скоринга, как использование OneHot, но оно может быть полезно для переменных

с упорядоченными категориями, такими как уровень образования или уровень дохода — такие категории как раз есть в нашем наборе данных. Таким образом, главным отличием является то, что Helmert-кодирование позволяет работать с категориальными переменными с упорядоченными категориями, что может быть плюсом, но плохо работать с неупорядоченными — что является недостатком.

Пример кодирования:

ID	Color
One	Green
Two	Blue
Three	Yellow
Four	Green

Таблица 3.4: До Helmert

ID	Green	Blue	Yellow
One	1	-1	-1
Two	-1	-1	-1
Three	0	2	-1
Four	1	-1	-1

Таблица 3.5: После Helert

3.2.4 FrequencyEncoder

FrequencyEncoder - это метод, используемый для кодирования категориальных переменных путем замены каждой категории частотой ее появления в наборе данных. Этот метод отражает представление каждой категории с точки зрения ее присутствия и может быть особенно полезен при работе с категориальными признаками с большим количеством уникальных значений.

Для каждой категориальной переменной частота каждой категории рассчитывается как количество вхождений данной категории, деленное на общее количество образцов. Затем каждая категория в исходном признаке заменяется соответствующей частотой.

ID	Category
1	A
2	B
3	A
4	C
5	A
6	B

Таблица 3.6: До FrequencyEncoder

ID	Encoded
1	0.50
2	0.33
3	0.50
4	0.17
5	0.50
6	0.33

Таблица 3.7: После FrequencyEncoder

3.3 Методы интерпретации

Интерпретируемость моделей - это набор методик, разработанных для расшифровки прогнозов, сделанных сложными моделями машинного обучения. Причины, побуждающие к интерпретируемости, во многом зависят от сферы применимости моделей: это и повышение прозрачности, укрепление доверия к системам искусственного интеллекта (ИИ), соблюдение нормативных требований, поиск путей оптимизации для бизнеса и многое другое.

В сфере кредитного скоринга спрос на интерпретируемость моделей особенно велик. Это облегчает выяснение решающих для банка вопросов, таких как факторы, способствующие отклонению конкретной кредитной заявки, или факторы, определяющие кредитоспособность. Полученные результаты могут повысить эффективность принятия решений, обеспечить соблюдение нормативных требований и улучшить общение с клиентами, а, самое главное, помочь банку минимизировать убытки или увеличить прибыль.

В контексте кредитного скоринга два самых известных метода интерпретации - SHAP и LIME - являются оптимальными вариантами благодаря их способности предоставлять объяснения на уровне признака.

3.3.1 SHAP

SHAP (SHapley Additive exPlanations) - это популярный метод интерпретации результатов моделей машинного обучения. Он помогает объяснить результаты с помощью единой меры важности признаков, основанной на теории кооперативных игр. Значения SHAP дают представление о процессе принятия решений моделью и обеспечивают прозрачность.

Для заданного объекта x, использует следующую формулу [1] для определения вклада i-го признака x_i в предсказание модели:

$$SHAP_i(x) = \sum_{S \subseteq \{1,\dots,p\} \setminus \{i\}} \frac{|S|!\ (p - |S| - 1)!}{p!} \left[f_S(x_{S \cup \{i\}}) - f_S(x_S) \right] \qquad (3)$$

где p - количество признаков в модели, S - подмножество признаков, f_S - функция предсказания модели для объектов, содержащих только признаки из S, x_S - объект, содержащий только признаки из S, $x_{S \cup i}$ - объект, содержащий признаки из S и i.

Эта формула определяет вклад каждого признака в предсказание модели путем сравнения предсказаний для всех возможных комбинаций признаков с предсказанием для исходного объекта x.

Метод SHAP может быть применен к различным моделям машинного обучения, в том числе к деревьям решений, линейным моделям, нейронным сетям и прочим, однако, для использования метода SHAP необходимо, чтобы модель была обучена на некоррелированных признаках или были применены методы предварительной обработки данных для удаления корреляции. Стоит также учитывать, что SHAP может быть медленным, особенно для высокоразмерных наборов данных или сложных моделей. Время вычислений SHAP увеличивается с ростом числа признаков, он лучше всего работает с наборами данных, не содержащими слишком большого числа признаков.

В целом, SHAP является универсальным методом интерпретации моделей машинного обучения и хорошо работает с широким спектром моделей и наборов данных.

3.3.2 LIME

LIME (Local Interpretable Model-agnostic Explanations) - это метод интерпретации моделей, разработанный для обеспечения понятных объяснений на уровне экземпляра для предсказаний, сделанных сложными моделями машинного обучения. LIME особенно полезен при интерпретации моделей "черного ящика где внутренние механизмы не имеют прямого доступа и не поддаются интерпретации.

LIME работает путем аппроксимации сложной модели более интерпретируемой, локально-линейной моделью для конкретного случая. Основная идея заключается в том, что даже если глобальная модель является нелинейной,

она все равно может быть точно аппроксимирована более простой линейной моделью в окрестности конкретной точки данных.

1 Выбирается объект x, который требуется объяснить.

2 Генерируется выборка X', путем случайного изменения исходных признаков объекта x.

3 Для каждого объекта x'_i из X' вычисляется вес w_i, отражающий важность объекта для объяснения исходного объекта x. В качестве веса может использоваться расстояние между x и x'_i в пространстве признаков.

4 Строится локальная модель g, используя обучающую выборку X' и соответствующие веса w_i.

5 Интерпретируются предсказания локальной модели g, чтобы объяснить принятие решения моделью на уровне отдельного объекта x.

3.4 Предобработка

В наборе данных, выбранном для проведения экспериментов, наблюдался значительный дисбаланс целевой переменной: только 8% от всего набора составляли случаи дефолта, которые необходимо предсказывать. Для упрощения процесса обучения и построения моделей датасет был отбалансирован путем удаления недефолтных заявок случайным образом, чтобы добиться распределения 30% дефолтных случаев и 70% успешных погашений.

Поскольку в данном исследовании важно попробовать различные комбинации между методами кодирования, всего были подобраны три основных варианта предобработки данных, которые позволят в дальнейшем более точно определить вклад кодирования в результат.

Базовая предобработка включает в себя:

- Удаление признаков, у которых пропущено более 60% значений
- Удаление индекса заявки, дублирующихся заявок
- Удаление заявок со значением AMT_INCOME_TOTAL выше 1.500.000 (признак содержит выбросы, но в целом имеет хорошую предсказательную способность)
- Заполнение пропусков в потенциально важных категориальных признаках NAME_TYPE_SUITE и OCCUPATION_TYPE на 'UDK'
- Заполнение пропусков в оставшихся данных с помощью методов fillna или KNNImputer в зависимости от последующего кодирования

После обработки данные содержат 307427 строк $\times$ 104 признаков.

Помимо этого, в исследовании используется два дополнительных типа подготовки данных, основанные на Weight of Evidence (WoE) преобразова-

нии.

WoE для всех данных включает в себя:

- Удаление индексов и дублирующихся заявок
- Расчет WoE-значения для всех признаков
- Удаление признаков, у которых низкое значение предсказательной силы Information Value

После обработки данные содержат 307511 строк $\times$ 61 признаков.

Кроме того, WoE можно использовать как обычное кодирование категориальных признаков, именно это и используется в третьем вариант предобработки.

WoE для категориальных признаков включает в себя:

- Выполняется базовая предобработка данных
- Расчет WoE-значения только для категориальных признаков
- Удаление категориальных признаков, у которых низкое значение предсказательной силы Information Value

После обработки данные содержат 307427 строк $\times$ 96 признаков.

Этот подход подходит, когда в наборе данных много категориальных признаков, и применение метода кодирования с большим количеством уникальных категорий не представляется возможным. Вычисляя WoE только по категориальным признакам, модель может уловить их предсказательную силу, снижая при этом сложность вычислений.

3.5 Метрики

ROC_AUC (Receiver Operating Characteristic - Area Under Curve) - это популярная метрика для задач бинарной классификации. ROC-кривая строится путем построения графика зависимости частоты истинных положительных результатов (TPR) от частоты ложных положительных результатов (FPR) при различных пороговых значениях. AUC - это площадь под ROC-кривой, и она определяет общую эффективность модели при всех пороговых значениях. AUC 1.0 означает идеальную модель, а AUC 0.5 - модель, которая работает не лучше, чем случайная.

$$\text{ROC\_AUC} = \int_0^1 \text{TPR}(t)\, dt \tag{4}$$

Precision (точность), также называемая положительной предсказательной ценностью, измеряет долю истинных положительных результатов от общего числа предсказанных положительных результатов.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5}$$

Recall, также известный как чувствительность или коэффициент истинных положительных результатов, измеряет долю истинных положительных результатов от общего числа фактических положительных результатов.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

F1 - это среднее гармоническое значение precision и recall. Он дает единый результат, который уравновешивает оба аспекта - точность и отзыв - в одном числе. Это особенно полезно в проблемах неравномерного распределения классов, где один из классов встречается значительно реже.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

3.6 Модели

Для экспериментов было принято решение взять три разных подхода: логистическую регрессию, XGBoost и несложную архитектуру нейронной сети. Небольшое описание каждой представлено ниже.

Логистическая регрессия - модель, основанная на логистической функции, также известной как сигмоидная функция, которая отображает линейную комбинацию входных характеристик на вероятность от 0 до 1. В кредитном скоринге она используется для оценки вероятности дефолта с учетом набора характеристик заемщика. Для более эффективного обучения были подобраны некоторые параметры модели: регуляризация, алгоритм оптимизации и другие.

XGBoost (eXtreme Gradient Boosting) - это алгоритм машинного обучения, использующий ансамбль деревьев решений для составления прогнозов. Он объединяет сильные стороны градиентного бустинга и деревьев решений, что приводит к улучшению эффективности прогнозирования. XGBoost строит аддитивную модель путем последовательного добавления деревьев решений, где каждое последующее дерево исправляет ошибки предыдущих. Ансамбль показал отличные результаты и без подбора параметров, но для улучшения качества было подобрано оптимальное количество моделей

Нейронная сеть с несколькими слоями позволит построить чуть более сложную модель и посмотреть на интерпретации и зависимости для совсем другой архитектуры, отличной от деревьев и линейных моделей. Для простоты был использован Sequential в библиотеке Keras - это класс, который представляет линейную структуру модели глубокого обучения, где слои последовательно соединяются друг за другом.

В итоговой архитектуре 7 слоев, с различным количеством нейронов, и функциями активации Relu и Sigmoid. В качестве функции потерь при обучении использовалась binary_crossentropy, метрики - binary accuracy, оптимизатор - adam.

3.7 Эксперименты

В этом разделе я кратко описываю полученные результаты после обучения и интерпретации каждой модели. Весь код, функции и исследования представлены на GitHub [9].

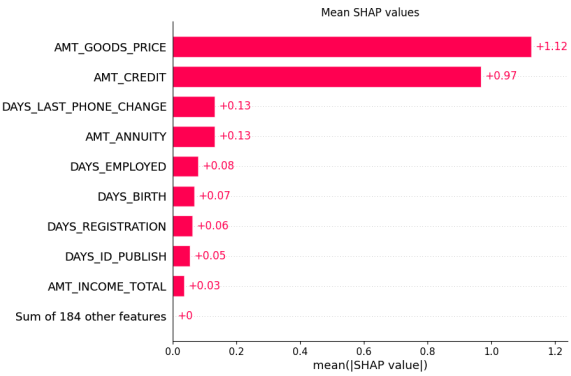
3.7.1 Логистическая регрессия

Самым классическим вариантом в кредитном скоринге является комбинация логистической регрессии и WoE-преобразования — она и показала лучший результат. Для обучения на необработанных данных были подобраны параметры, которые использовались для всех типов кодирования.

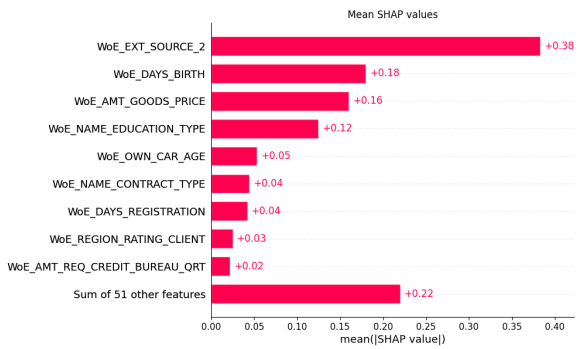
Предобработка	Кодирование	*roc_auc*	*precision*	*recall*	*F*1
Базовая	OneHot	0.628	0.526	0.025	0.049
Базовая	Helmert	0.628	0.526	0.025	0.049
Базовая	Frequency	0.623	**0.613**	0.022	0.043
-	WoE	**0.693**	0.591	**0.261**	0.362
Базовая	WoE	0.624	0.553	0.025	0.049

Таблица 3.8: Логистическая регрессия

В интерпретации не все так однозначно: для OneHot, Helmert и Frequency преобразования ведущими по значимости признаками оказались данные о сумме кредита (как обычного, так и потребительского), в то время как при обучении на WoE — один из неизвестных нам источников и возраст.



(a) OneHot Encoding на LogReg



(b) WoE for all на LogReg

Рис. 3.1: Разница набора лучших признаков в зависимости от кодирования

Другие результаты интерпретаций можно посмотреть в Приложениях: рисунки 4.1, 4.2 и другие.

Модель	Кодирование	Топ-3 признака	SHAP value
LogReg	OneHot	AMT_GOODS_PRICE	1.08
		AMT_CREDIT	0.93
		AMT_ANNUITY	0.14
	Helmert	AMT_GOODS_PRICE	1.08
		AMT_CREDIT	0.92
		AMT_ANNUITY	0.14
	Frequency	AMT_GOODS_PRICE	1.09
		AMT_CREDIT	0.93
		AMT_ANNUITY	0.14
	WoE all	EXT_SOURSE_2	0.38
		DAYS_BIRTH	0.18
		AMT_GOODS_PRICE	0.16
	WoE categ	AMT_GOODS_PRICE	1.10
		AMT_CREDIT	0.95
		AMT_ANNUITY	0.14

Таблица 3.9: Логистическая регрессия

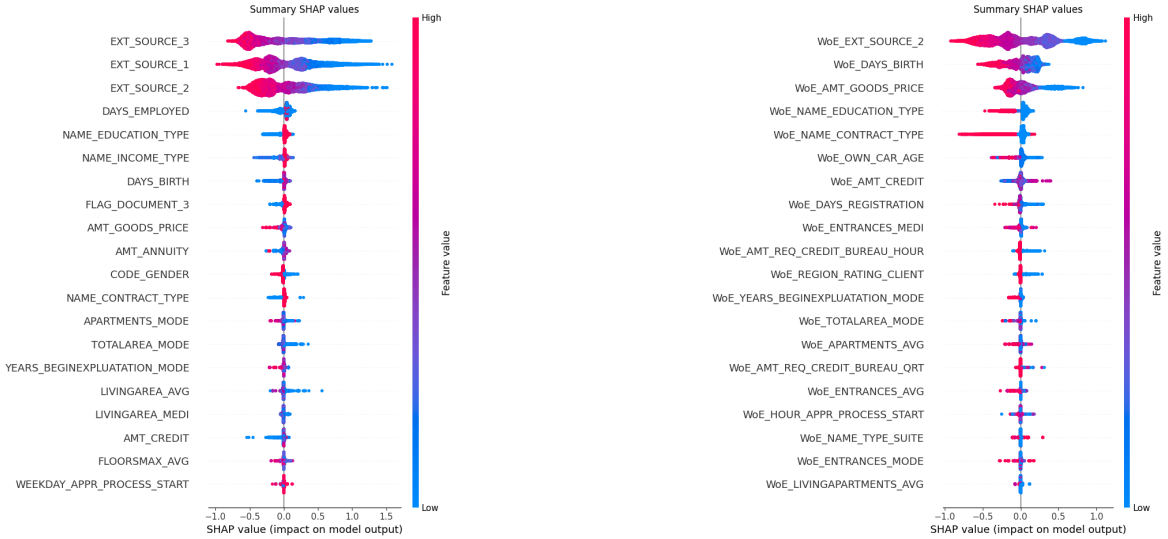
3.7.2 XGBoost

XGBoost показал наилучшие результаты обучения по выбранным метрикам, при этом комбинация с WoE-преобразованием оказалась не настолько выигрышной, как другие методы. Однако, при интерпретации результатов вновь были сильные расхождения.

Предобработка	Кодирование	*roc_auc*	*precision*	*recall*	$F1$
Базовая	OneHot	**0.758**	0.642	0.430	0.515
Базовая	Helmert	0.724	0.609	0.392	0.477
Базовая	Frequency	0.757	**0.645**	0.427	0.514
-	WoE	0.687	0.588	0.259	0.360
Базовая	WoE	**0.758**	0.640	**0.435**	**0.518**

Таблица 3.10: XGBoost

Наиболее сильно отличается интерпретация между WoE-преобразованием для всех данных и Frequency Encoding. При этом, впервые после логистической регрессии в топ-3 наиболее важных признака попали первый и второй ресурс. Интерпретация нам неизвестна, в силу конфиденциальности данных.



(a) Frequency Encoding на XGBoost

(b) WoE for all на XGBoost

Рис. 3.2: Разница набора лучших признаков в зависимости от кодирования

Модель	Кодирование	Топ-3 признака	SHAP value
XGBoost	OneHot	EXT_SOURSE_3	0.40
		EXT_SOURSE_2	0.33
		EXT_SOURSE_1	0.31
	Helmert	EXT_SOURSE_2	0.32
		EXT_SOURSE_3	0.31
		EXT_SOURSE_1	0.1
	Frequency	EXT_SOURSE_3	0.39
		EXT_SOURSE_1	0.34
		EXT_SOURSE_2	0.29
	WoE all	EXT_SOURSE_2	0.35
		DAYS_BIRTH	0.16
		AMT_GOODS_PRICE	0.15
	WoE categ	FLAG_WORK_PHONE	0.33
		FLAG_EMP_PHONE	0.26
		FLAG_PHONE	0.12

Таблица 3.11: XGBoost

Другие результаты интерпретаций можно посмотреть в Приложениях: ри-

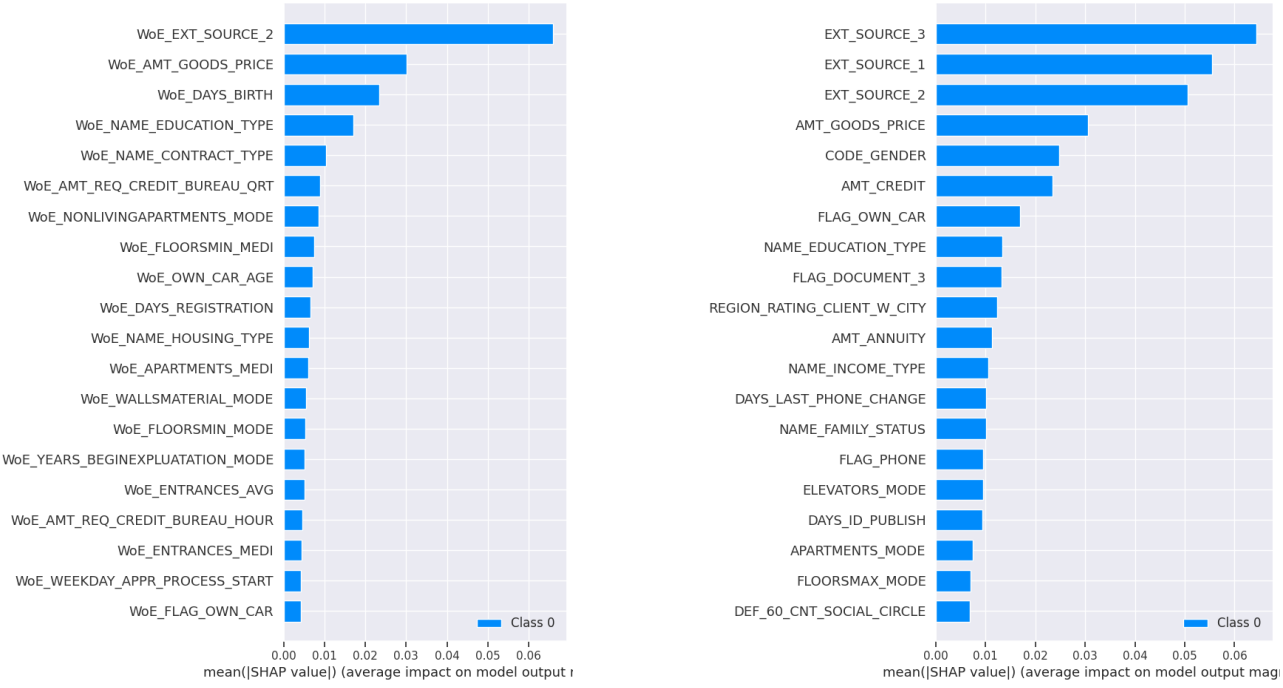
сунки 4.5, 4.6 и другие. Также в таблице 3.11 представлены результаты интерпретаций для всех типов кодирования при обучении на XGBoost.

3.7.3 Нейронная сеть

В очередной раз видим, что сильнее всего разнятся интерпретации в сравнении с WoE-преобразованием на всех данных. На данном этапе можно предварительно заключить, что свою роль вносит тот факт, что все признаки в этом случае были изменены на этапе предобработки, а не только категориальные.

Предобработка	Кодирование	*roc_auc*	*precision*	*recall*	*F*1
Базовая	OneHot	0.768	**0.291**	0.997	**0.451**
Базовая	Helmert	0.774	**0.291**	0.997	0.449
Базовая	Frequency	**0.776**	0.286	**0.999**	0.445
-	WoE	0.693	0.103	0.85	0.185
Базовая	WoE	0.774	0.290	0.998	0.450

Таблица 3.12: Sequential



(a) WoE for all на Sequential

(b) Frequency Encoding на Sequential

Рис. 3.3: Разница набора лучших признаков в зависимости от кодирования

К моменту последних экспериментов можно заметить, что, чаще всего,

Модель	Кодирование	Топ-3 признака	SHAP value
Sequential	OneHot	EXT_SOURSE_3	0.49
		EXT_SOURSE_2	0.43
		EXT_SOURSE_1	0.42
	Helmert	EXT_SOURSE_3	0.68
		EXT_SOURSE_1	0.55
		EXT_SOURSE_2	0.51
	Frequency	EXT_SOURSE_3	0.66
		EXT_SOURSE_1	0.55
		EXT_SOURSE_2	0.51
	WoE all	EXT_SOURSE_2	0.65
		AMT_GOODS_PRICE	0.31
		DAYS_BIRTH	0.22
	WoE categ	EXT_SOURSE_3	0.74
		EXT_SOURSE_1	0.59
		EXT_SOURSE_2	0.44

Таблица 3.13: Sequential

при кодировании только категориальных переменных, набор признаков с наибольшим средним SHAP-значением примерно одинаков. Больше интерпретаций обучения нейронной сети можно найти в приложении 4.7, 4.8 и далее.

3.7.4 CatBoost

В качестве контрольного эксперимента было принято решение использовать CatBoost [10], поскольку в нем втроены преобразования категориальных признаков. Обучение проводилось с учетом подобранных в прошлом исследовании [11] параметров, результаты 4.9 показали, что лидирующие в прошлых экспериментах признаки EXT_SOURSE_2 и EXT_SOURSE_3 могут косвенно создавать утечку целевой переменной из-за критически выбивающихся значений feature_importance_.

Предобработка	Кодирование	*roc_auc*
Базовая	CatBoost	0.736

Таблица 3.14: CatBoost

3.8 Практическая применимость

Существует несколько способов подсчета стоимости, которую выбранные модели могут принести бизнесу.

Минимизация ожидаемых потерь

В этом случае рассчитывается средний ожидаемый убыток, который предотвращает определенная модель, его также называют средней ошибкой финансовых потерь. Более низкие значения указывают на то, что прогнозы модели близко соответствуют фактическим результатам. Для каждого заявителя высчитывается ожидаемый убыток как вероятность дефолта, умноженная на потенциальный убыток от дефолта: $p_1 * amount\_of\_credit$. Затем подсчитывается количество убытков, предотвращаемых с помощью модели, и сравниваются ожидаемые убытки с учетом и без учета прогнозов.

Среднее значение суммы кредита по всем данным: 599k. Посмотрим на ошибки финансовых потерь у разных по типу моделей с лучшим roc_auc:

Предобработка	Кодирование	Модель	MAE
Базовая	WoE	XGBoost	13k
-	WoE	LogReg	1k
Базовая	Frequency	Sequential	172k

Таблица 3.15: Средняя ошибка финансовых потерь

Чтобы более точно определить ущерб, который возможен при ложной выдаче кредита, рассмотрим подробнее признаки, которые показали лучшую предсказательную способность на разных моделях.

1. **Снижение риска**

Если определенные характеристики определены как значимые для прогнозирования дефолтов по кредитам, банк может принять превентивные меры при работе с клиентами, которые демонстрируют характеристики высокого риска. Например, признак «NAME_EDUCATION_TYPE», или уровень образования клиента, является значимым во многих интерпретациях моде-

лей, и компания может установить политику, согласно которой кредиты не будут одобряться заявителям с определенным уровнем образования.

Можно посмотреть подробнее на убытки: издержками ложноотрицательного результата могут быть потери от выдачи кредита, который не был возвращен. В имеющихся данных заявки мы можем взять за основу сумму кредита, не учитывая издержки от его обеспечения в силу нехватки данных. Сумма убытков от всех ложновыданных кредитов в нашем случае составила: 13846851k. При этом можно посмотреть на убытки от выдачи кредитов со степенью образования «Lower secondary» — 10% от всех клиентов с этой степенью не смогли погасить кредит, подробнее в таблице **??**. Убытки клиентом с образованием ниже среднего составят: 11254263k, что является 81% от общих убытков, при том, что средний убыток по клиенту составляет 540k.

Посмотрим другую интерпретацию, где главным по предсказательной способности неоднократно стал один из закодированных источников «EXT_SOURCE_2». Как видно по распределению, значение позволяет предсказывать погашение кредита, при этом практически никак не влияя на дефолт. В случае, если у банка больше убытков несет ложноотрицательный результат, этот признак не будет нести в себе никакой информации.

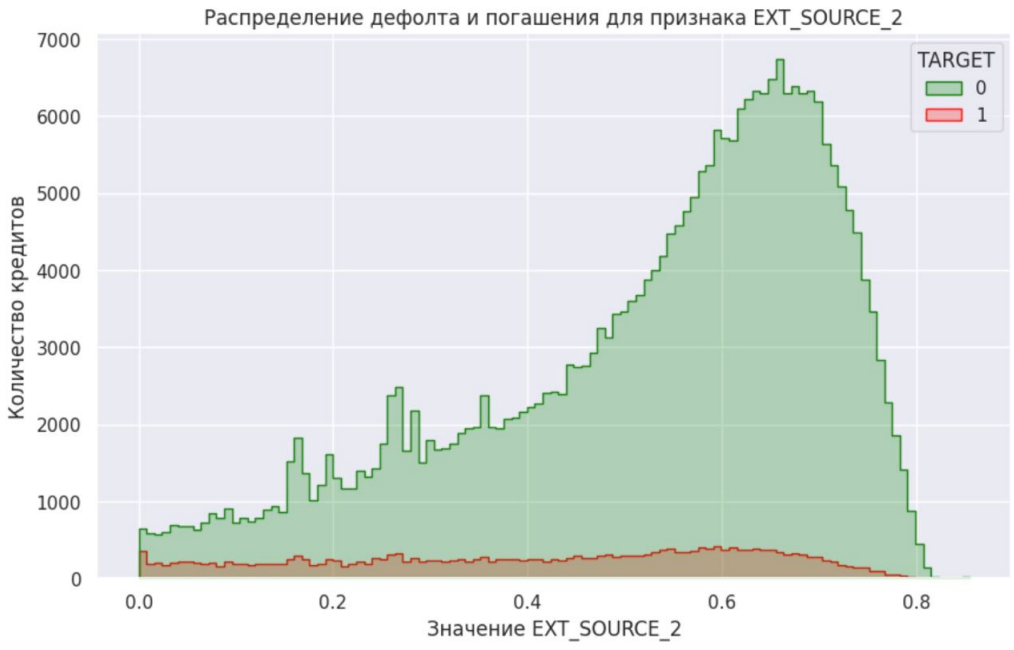


Рис. 3.4: Распределение погашения и дефолта

Если число характеристик о клиенте у банка растет, не предоставляется возможным исследовать их все с помощью банковского эксперта в силу человеческого фактора, при этом даже с помощью самых простых моделей машинного обучения возможно быстро найти самую потенциальную категорию клиентов и принять меры. Однако, как выяснилось, стоит учитывать, что не всегда предсказания только одной модели могут сразу обнаружить признаки с наибольшей предсказательной способностью, а, следовательно, и помочь с одного шага минимизировать ущерб.

NAME_EDUCATION_TYPE	TARGET	
Academic degree	0	0.981707
	1	0.018293
Higher education	0	0.946449
	1	0.053551
Incomplete higher	0	0.915150
	1	0.084850
Lower secondary	0	0.890723
	1	0.109277
Secondary / secondary special	0	0.910601
	1	0.089399

Рис. 3.5: Распределение погашения и дефолта в зависимости от степени образования

Существуют и другие методы оптимизации бизнеса, которые могут использовать интерпретации моделей:

2. **Формулирование политики**

Важность признаков также может служить основой для формулирования политики бизнеса. Например, если «CONTRACT_TYPE» является важной характеристикой, предприятие может принять решение о более строгой постановке условий, которые ассоциируются с более высоким уровнем дефолта.

3. **Маркетинговая стратегия**

Важность характеристик может также определять маркетинговые стратегии. Если определенные демографические характеристики (например, «AGE» или «GENDER», как в некоторых интерпретациях используемых моделей) являются значимыми, маркетинговые кампании могут быть направлены на

лиц, которые с большей вероятностью будут погашать кредиты, что повысит общее качество кредитного портфеля.

4. **Улучшение сбора данных**

Если определенные характеристики неизменно важны для всех моделей, это может указывать на то, что это критические точки сбора данных в процессе подачи заявки на кредит. Это также может указывать на то, что создание дополнительных функций (feature engineering), связанных с этими важными характеристиками, может еще больше улучшить производительность модели.

3.9 Заключение

В последние годы исследования в контексте кредитного скоринга направлены на попытки улучшить качество решения с помощью нейронных сетей, а также на борьбу с возникающей дискриминацией при обучении. Однако, на практике банки редко внедряют даже лучшие по результатам ансамблевые методы, из-за невозможности грамотно интерпретировать результаты сложных моделей.

Данное исследование показало, что одним из факторов, влияющих на результат интерпретации является метод кодирования признаков. При этом, удалось выяснить чуть больше, и показать, что не только категориальные признаки способны оказывать влияние, а также и кодирование и преобразование численных — это подтверждает тот факт, что сильнее всего результаты отличались при WoE-преобразовании всех данных, а не использовании его как кодировщика. Этот вывод подчеркивает необходимость более глубокого понимания влияния методов кодирования признаков на модели кредитного скоринга.

Задачи, выполненные в рамках исследования:

- Выполнен полный разведочный анализ данных для представленного набора из 122 признаков, составлена итоговая витрина.
- С помощью тестирования на безлайн модели выбрано оптимальное соотношение категорий для экспериментов и устранена несбалансированность данных.
- На основе обзорных исследований выбраны модели для экспериментов, а также подобраны параметры для логистичечской регрессии, XGBoost, протестированы различные архитектуры, метрики и другие параметры Sequential
- Для улучшения качества исследования, подобрано три варианта предобработки данных перед кодированием

- Проведена серия экспериментов для OneHot, Helmert, Frequency кодирования, а также двумя вариантами использования Weight of Evidence преобразования на каждой из трех итоговых моделей
- На основе полученных результатов посчитана средняя ошибка финансовых потерь, а также примеры интерпретации признаков с наибольшей предсказательной способностью

Кроме этого, существует широкий спект направлений для дальнейших исследований:

- Проведение экспериментов с другими типами кодирования, включая алгоритмические и автоматические, а также кодирование не только категориальных признаков
- Изучение новейших методов интерпретации, кроме SHAP и LIME
- Изучение влияния баланса между классами на интерпретируемость
- Проведение экспериментов с другим набором банковских данных, включая транзакционные, данные из бюро кредитных историй и прочие

Дальнейшие исследования в этой области могут улучшить наше понимание взаимосвязи между методами кодирования и интерпретируемостью, что приведет к совершенствованию моделей кредитного скоринга и улучшению финансовых результатов.

Список литературы

1. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.

2. M. T. Ribeiro, S. Singh, and C. Guestrin, “"why should i trust you?"explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

3. A. R. Provenzano, D. Trifirò, A. Datteo, L. Giada, N. Jean, A. Riciputi, G. L. Pera, M. Spadaccino, L. Massaron, and C. Nordio, “Machine learning approach for credit scoring,” *arXiv preprint arXiv:2008.01687*, 2020.

4. X. Dastile, T. Celik, and M. Potsane, “Statistical and machine learning models in credit scoring: A systematic literature survey,” *Applied Soft Computing*, vol. 91, p. 106263, 2020.

5. B. R. Gunnarsson, S. Vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, “Deep learning for credit scoring: Do or don’t?” *European Journal of Operational Research*, vol. 295, no. 1, pp. 292–305, 2021.

6. J. T. Hancock and T. M. Khoshgoftaar, “Survey on categorical data for neural networks,” *Journal of Big Data*, vol. 7, no. 1, pp. 1–41, 2020.

7. Kaggle, “Home credit default risk,” https://www.kaggle.com/competitions/home-credit-default-risk/overview, 2020, Дата обращения: 27/09/2022.

8. N. Siddiqi, *Credit risk scorecards: developing and implementing intelligent credit scoring*. John Wiley & Sons, 2012, vol. 3.

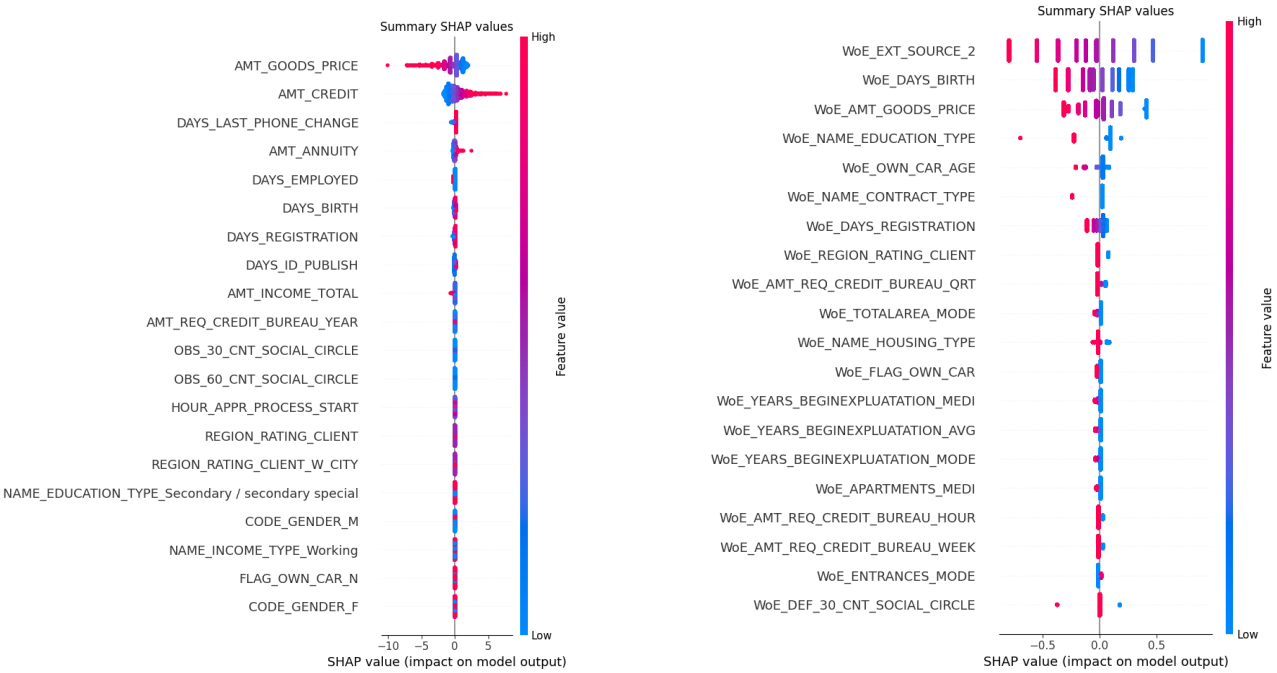
9. S. Anna, “Github,” https://github.com/AnnaSmirnova-study.

10. J. Qi, R. Yang, and P. Wang, “Application of explainable machine learning based on catboost in credit scoring,” in *Journal of Physics: Conference Series*, vol. 1955, no. 1. IOP Publishing, 2021, p. 012039.

11. А.Р.Смирнова, “Кредитный скоринг. Сравнение линейных моделей с более сложными моделями машинного обучения,” 2022. [Online]. Available: https://github.com/AnnaSmirnova-study/CreditScoring

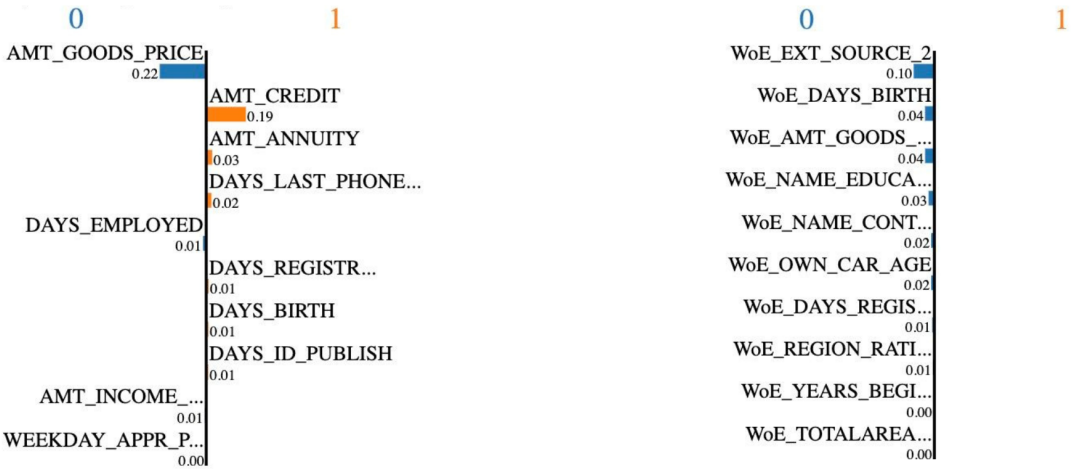
4 Приложения

4.1 Логистическая регрессия



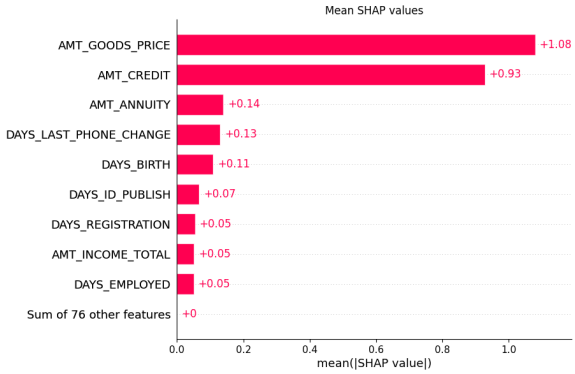
(a) OneHot Encoding (b) WoE for all

Рис. 4.1: Значения SHAP интерпретации

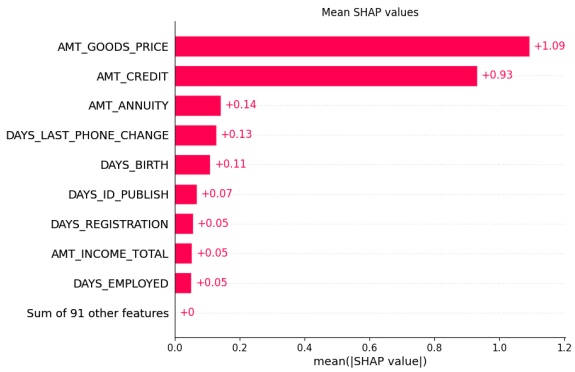


(a) OneHot Encoding (b) WoE for all

Рис. 4.2: Значения LIME интерпретации

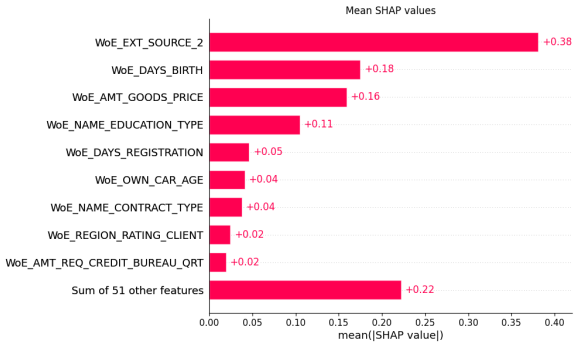


(a) Helmert Encoding

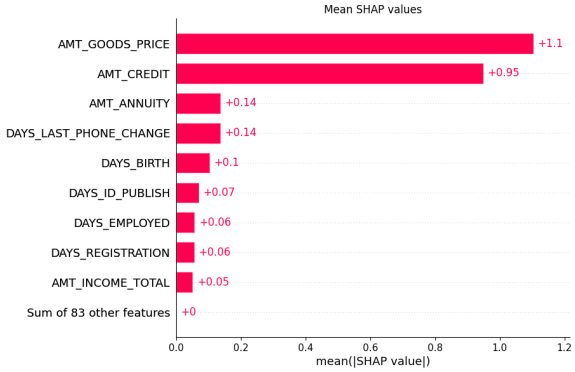


(b) Frequency Encoding

Рис. 4.3: Значения SHAP интерпретации



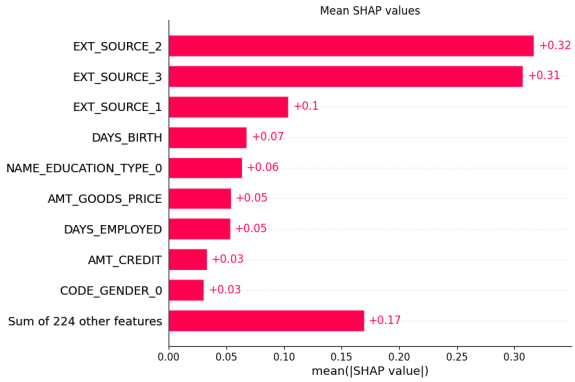
(a) WoE for all



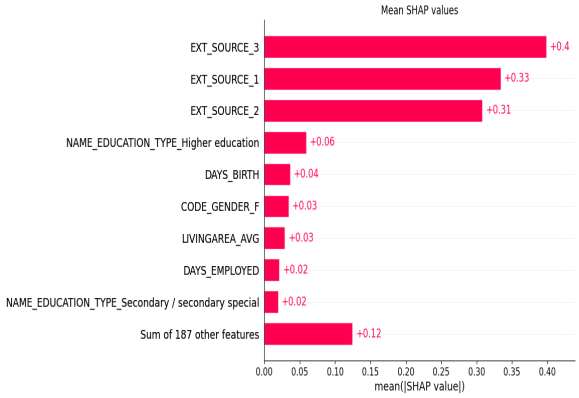
(b) WoE for categorical

Рис. 4.4: Значения LIME интерпретации

4.2 XGBoost

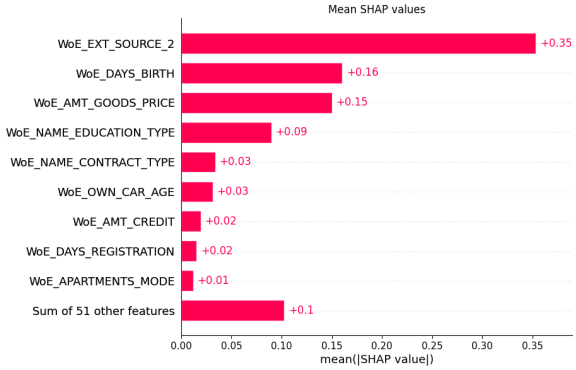


(a) Helmert Encoding

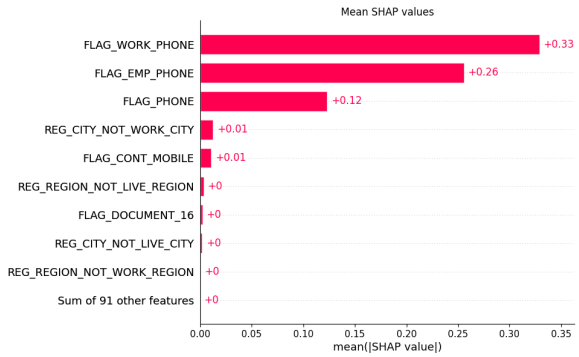


(b) OneHot Encoding

Рис. 4.5: Значения SHAP интерпретации



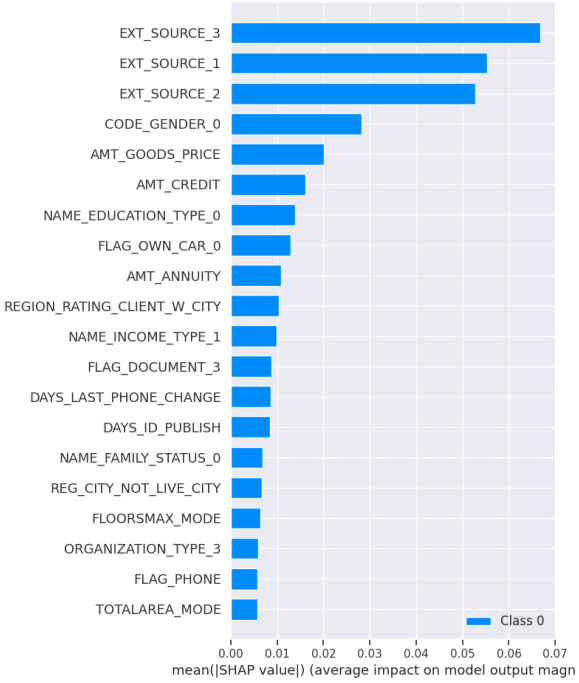
(a) WoE for all



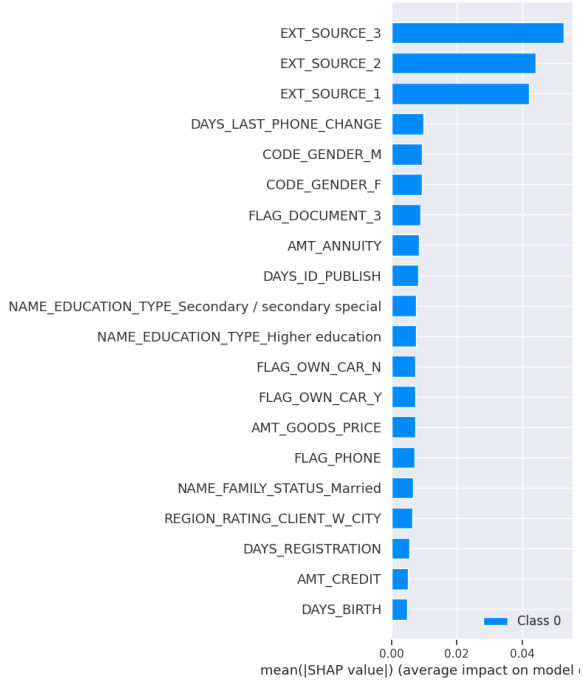
(b) WoE for categorical

Рис. 4.6: Значения SHAP интерпретации

4.3 Sequential

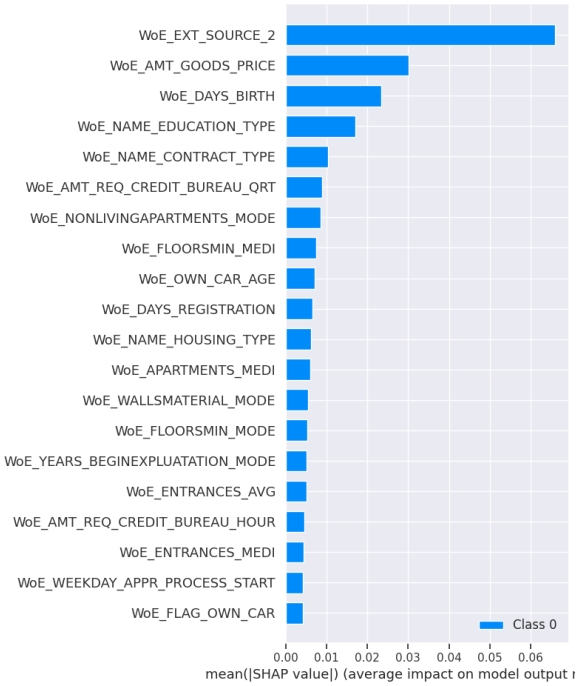


(a) Helmert Encoding

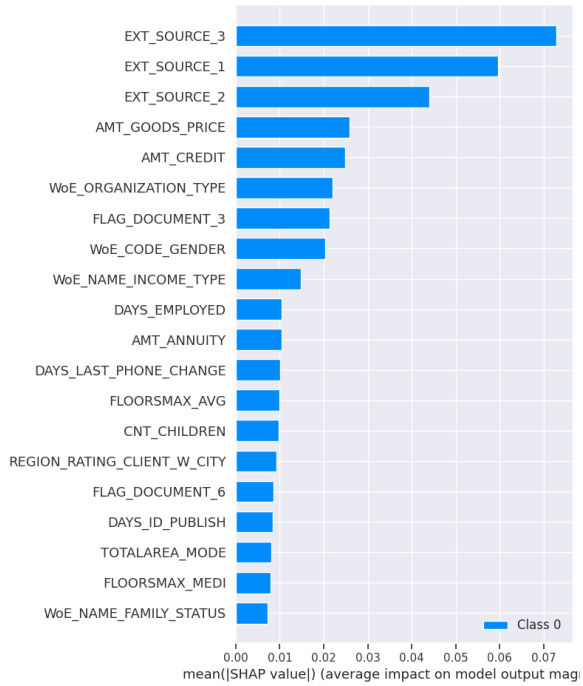


(b) OneHot Encoding

Рис. 4.7: Значения SHAP интерпретации



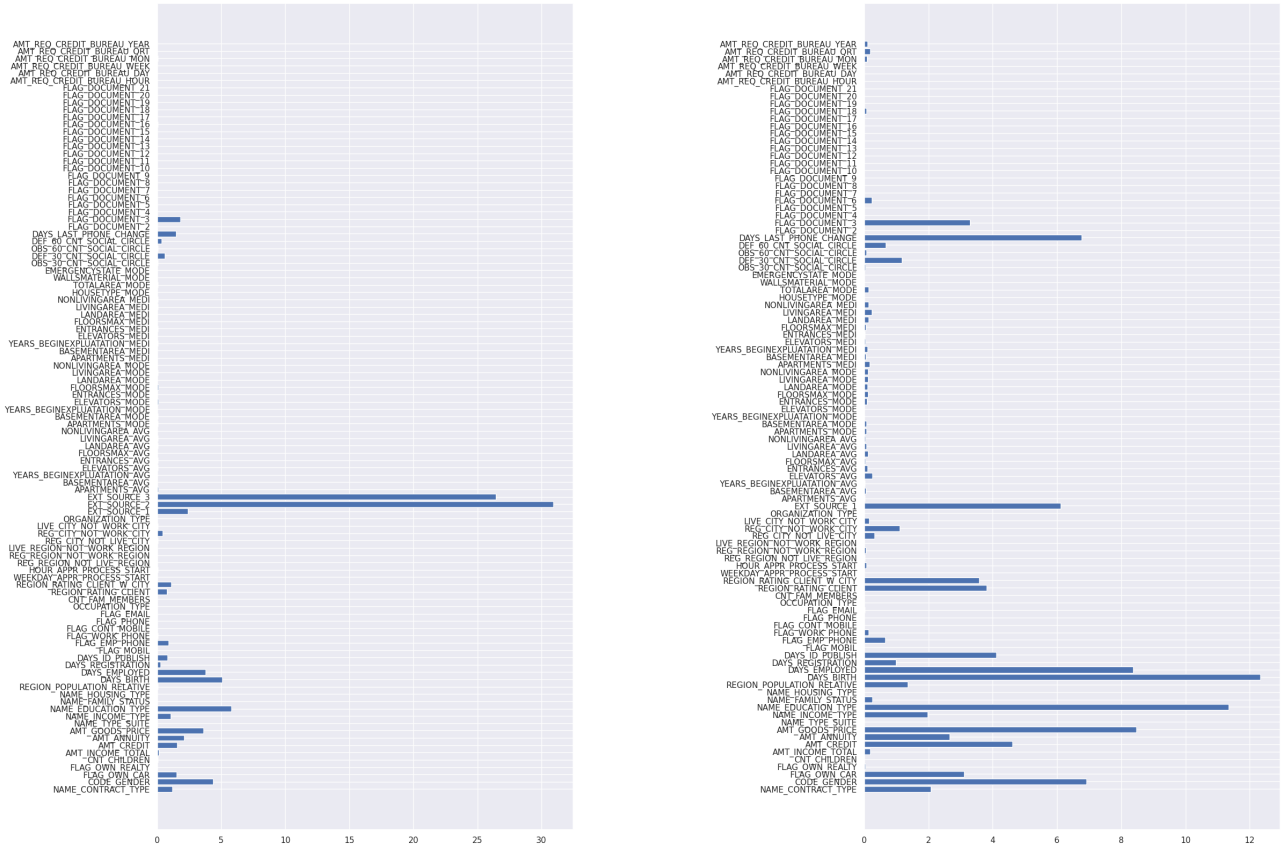
(a) WoE for all



(b) WoE for categorical

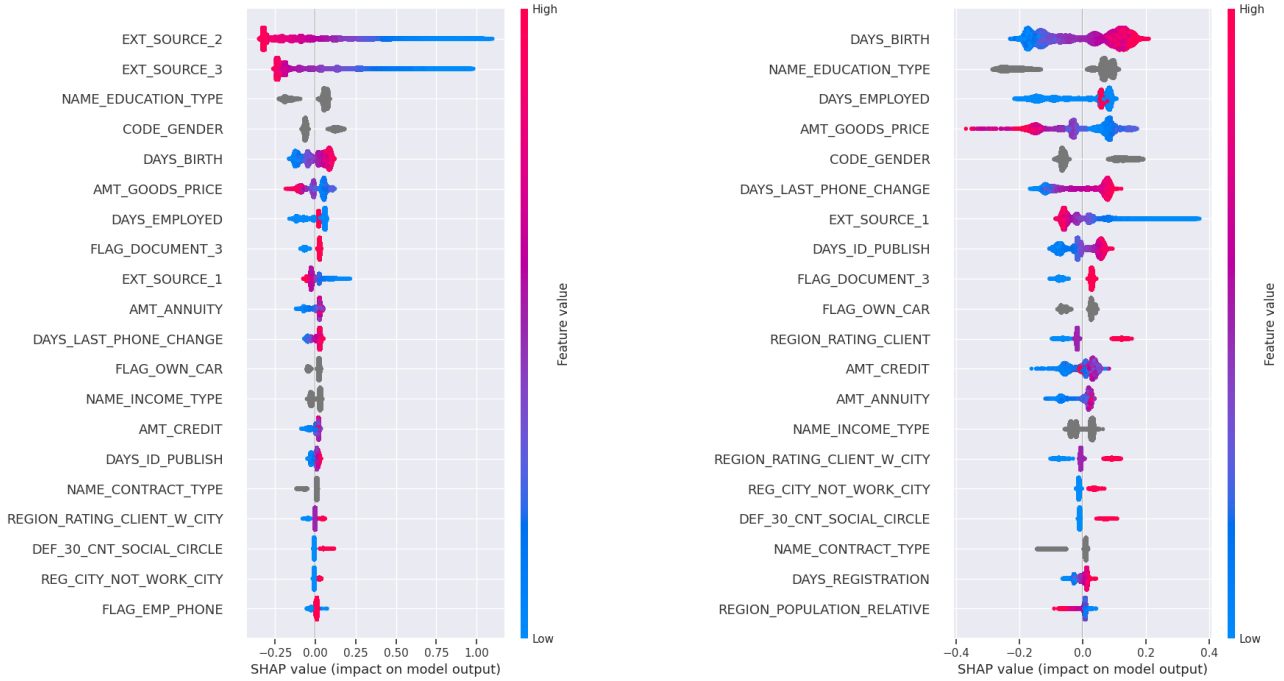
Рис. 4.8: Значения SHAP интерпретации

4.4 CatBoost



(a) На всех признаках (b) После удаления 2 и 3 источников

Рис. 4.9: Значения feature_inportance_ интерпретации



(a) На всех признаках (b) После удаления 2 и 3 источников

Рис. 4.10: Значения SHAP интерпретации