ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Экономика и Анализ Данных»

**Отчет о проекте на тему:**

# Построение скоринговой модели с использованием WOE-преобразований

**Выполнил студент:**

группы №БЭАД223, 2 курс Шевченко Артём Эдуардович

**Руководитель проекта:**

Васильева Наталья Васильевна
Приглашенный преподаватель
Факультет экономических наук НИУ ВШЭ

Москва 2024

# Содержание

# 1 Введение

## 1.1 Аннотация

Скоринговая модель – это модель бинарной классификации, которая используется для оценки вероятности наступления события дефолт/недефолт. В работе используется различные модели машинного обучения поверх WOE (weight-of-evidence) трансформации для повышения их интерпретируемости.

## 1.2 Цель работы

Цель работы - изучение различных подходов машинного обучения, построение полноценной и интерпретируемой модели кредитного скоринга для оценки вероятности дефолта заёмщика, в том числе для прохождения ей предложенных тестов валидации.

## 1.3 Методология

- Провести предварительный анализ данных.
- Сгенерировать новые признаки и провести WOE-преобразования.
- Построить модели и оптимизировать гиперпараметры.
- Провести валидацию модели.
- Интерпретировать результаты модели.

## 1.4 Данные

### 1.4.1 Описание

Датасет содержит данные о кредитах, выданных компанией LendingClub. Оригинальный набор данных представлен на Kaggle - анализ я проводил на его обрезанной по признакам версии. В тренировочной выборке содержится 61169 записей, а в тестовой - 60334 записи.

В датасете 23 фичи, их описание представлено в таблице ниже:

### 1.4.2 Фичи

| Field | Description | |
|---|---|---|
| 1 | issue_d | The month which the loan was funded |

| | | |
|---|---|---|
| 2 | purpose | A category provided by the borrower for the loan request. |
| 3 | addr_state | The state provided by the borrower in the loan application |
| 4 | sub_grade | External assigned loan subgrade |
| 5 | home_ownership | The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER |
| 6 | emp_title | The job title supplied by the Borrower when applying for the loan. |
| 8 | installment | The monthly payment owed by the borrower if the loan originates. |
| 9 | dti | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income. |
| 10 | funded_amnt | The total amount committed to that loan at that point in time. |
| 11 | annual_inc | The self-reported annual income provided by the borrower during registration. |
| 12 | emp_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| 13 | term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| 14 | inq_last_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| 15 | mths_since_recent_inq | Months since most recent inquiry. |
| 16 | delinq_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| 17 | chargeoff_within_12_mths | Number of charge-offs within 12 months |
| 18 | num_accts_ever_120_pd | Number of accounts ever 120 or more days past due |
| 19 | num_tl_90g_dpd_24m | Number of accounts 90 or more days past due in last 24 months |
| 20 | acc_open_past_24mths | Number of trades opened in past 24 months. |
| 21 | avg_cur_bal | Average current balance of all accounts |

| | | |
|---|---|---|
| 22 | tot_hi_cred_lim | Total high credit/credit limit |
| 23 | delinq_amnt | The past-due amount owed for the accounts on which the borrower is now delinquent. |

# 2 Explaratory Data Analysis

Начну работу с первичного анализа - буду рассматривать отдельные признаки, что-то говорить про них, обрабатывать их и генерировать на их основе новые.

Целевой переменной в модели является def. В ней лежит бинарное число (1 или 0) - решение о выдаче или невыдаче кредита. Если быть точнее, число показывает, будет ли у клиента дефолт или не будет (где 1 - будет, 0 - нет).

## 2.1 Категориальные признаки

Часть признаков на этом этапе была удалена - например, emp_title, который содержал места работы, выгруженные из анкет клиентов. При анализе признака было обнаружено большое количество уникальных значений (63% на трейне, 39% на тесте), составляющее значительную часть от выборки, в связи с чем признак просто вносил бы лишний шум в модель.

Кроме того, был удален признак installment - использование ежемесячного платежа в модели может приводить к лику данных. Так, в installment входит ставка по кредиту, которая формируется исходя из личных данных клиента.

Наконец, был удалён признак issue_d - то есть дата заявки. Так, конкретные даты подачи создают лишний шум для модели и спобствуют переобучению, а появившиеся в датасете даты заведомо не появятся для новых заявок.

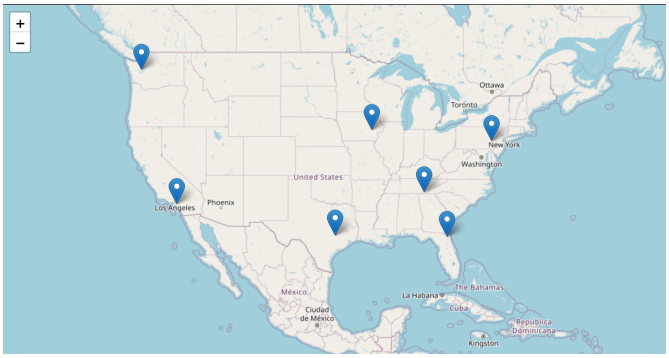
### 2.1.1 Распределения

В работе было рассмотрено распределение категориальных переменных по дискретным значениям (например purpose принимает значения debt_consolidation, medical и тд). Так, я построил распределения в тренировочной и тестовой выборках, сравнил их и графически показал отсутствие выбросов.

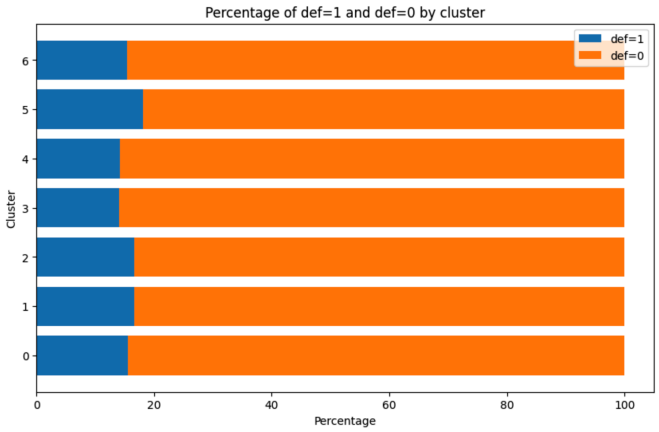
### 2.1.2 Генерация признаков

Далее я обработал addr_state - FIPS код штата США заявителя на кредит. Так, в интернете я нашёл датасет, с помощью которого составил эмбеддинг для перевода сокращения штата в

широту и долготу его местонахождения. На этой основе можно генерировать новые признаки - воспользуемся алгоритмом k-средних для кластеризации значений. Алгоритм будет учиться только на тренировочной выборке во избежание data leak. Так, алгоритм делит все addr_state по кластерам (количество которых задано мануально), и каждому заявителю генерируется признак по отношению к тому или иному кластеру. Идея почти аналогична делению на бины, но при этом учитывает метрики дальности городов заявителей друг от друга, соответственно распределение значений целевой переменной могут отличаться в зависимости от кластера. В рамках построения модели я проводил эксперименты с количеством кластеров и получил, что 6 - оптимальное число для равномерного распределения заявителей и улучшения качества модели.




В результате, гипотеза о том, что для разных кластеров может значительно отличаться соотношение дефолтов, оказалась отвергнута:




В дальнейшем я буду относиться к этому признаку как к категориальному с натуральным числом - номером кластера.

### 2.1.3 Преобразование в числовые типы

Так как в общем случае модель не может воспринимать категориальные признаки для обучения, нам нужно преобразовать их в числовые типы, то есть, по сути, закодировать по некоторой логике. Эта логика может варьироваться в зависимости от задачи, но в нашей задаче я воспользовался ce.OrdinalEncoder, который реализует подход one-hot encoding, то есть присвоение каждому

возможному значению категориального признака некоторго натурального числа. Этот подход особенно эффективен в случае, если область значения признака не слишком велика (а у большинства наших признаков область значений ограничена 15-20-ю).

В процессе подобного преобразования важно не допустить неочевидную ошибку. Если признак имеет ранговый смысл (как, например категориальный признак subgrage, отражающий кредитный рейтинг клиента, принимающий значения $(A_1, ..., A_n, B_1, ...)$ и являющийся ранговым, поскольку каждый рейтинг с буквой, идущей раньше в алфавите, лучше любого рейтинга с буквой позже в алфавите, а в рамках одной буквы лучше рейтинг с меньшей цифрой), то его нельзя кодирвать натуральными числами по порядку встречи нового значения в выборке, как это делает классический энкодер. Если наше преобразование окажется не линейным $(A_1 = 1, ..., A_n = n, B_1 = n + 1, ...))$, а беспорядочным $(A_1 = n, ..., A_n = k, B_1 = m, ...))$, то создаётся риск в будущем не суметь поделить значения признака на последовательные промежутки, в рамках которых процент дефолтов по кредитам схож (в чем, как окажется позже, и заключается смысл WOE-преобразований). Это приведёт к неизбежному критическому падению качества модели, поскольку тот же subgrade, как выяснится позже - один из ключевых признаков на обучении. Во избежание этой ошибки в работе разработан собственный энкодер для проведения преобразования, описанного выше.

## 2.2 Непрерывные признаки

### 2.2.1 Заполнение пропущенных значений

В отличие от категориальных признаков, многие непрерывные содержали пропущенные значения. В работе с данными существует множество подходов к их заполнению, но для начала важно лучше понять распределение и область значений наших признаков.

```
dti                          0
funded_amnt                  0
annual_inc                   0
emp_length                5809
term                         0
inq_last_6mths               0
mths_since_recent_inq    19518
delinq_2yrs                  0
chargeoff_within_12_mths     0
num_accts_ever_120_pd    11941
num_tl_90g_dpd_24m       11941
acc_open_past_24mths      7886
avg_cur_bal              11945
tot_hi_cred_lim          11941
delinq_amnt                  0
dtype: int64
```

Рис. 2.1: Количество пропущенных значений для каждого признака

Для этого в работе реализованы две функции: одна рассматривает распределения значений признака на тренировочной и тестовой выборки для анализа возможных выбросов и различий в распределении, а другая ищет сильную корреляцию с другими признаками с помощью заданного погороа корреляции, который в работе я установил равным 0.7. В ноутбуке можно ознакомиться с графиками соответствующих распределений.

По результатам экспериментов, была обнаружена сильная корреляция признаков avg_cur_bal и tot_hi_cred_lim. Её я опишу чуть дальше.

Выборчочные распределения признаков в тренировочной и тестовой выборках не отличались, в связи с чем соответствующие пропущенные значения заполнены средними значениями для каждого отдельного признака.

### 2.2.2 Стандартизация

Стандартизация - это процесс приведения непрерывных признаков к одному масштабу путем вычитания среднего значения и деления на стандартное отклонение. Этот метод помогает сделать данные более интерпретируемыми для модели и обеспечивает лучшую сходимость алгоритмов машинного обучения. В работе я также воспользовался этим методом обработки данных.

Важно отметить, что скейлер применим только к действительно непрерывным признакам, которые изначально не являлись категориальными и позже оказались закодированные некоторыми числами, ведь иначе стандартизация бы потеряла смысл.

### 2.2.3 Корреляционный анализ

Рассмотрена корреляцию признаков в рамках анализа. Для этого, я воспользовался корреляционной матрицей из numpy. В ней, конечно, рассмотрены только непрерывные признаки.


Видно, что немногие признаки демонстрируют высокую связь - среди таковых можно выделить только признак среднего текущего баланса и кредитного лимита (0,81). Так как оба признака могут иметь собственную интерпретацию в WOE-анализе, я пожертвовал некоторой разгрузкой модели в виде удаления одного из коррелируемых признаков из датасета.

# 3 Обучение

## 3.1 Подбор моделей обучения

Для задачи кредитного скоринга я воспользовался тремя моделями машинного обучения: логистической регрессией, случайным лесом и градиентным бустингом. Всего в работе 4 этапа обучения моделей - на исходном обработанном датасете, на таковом с неизбирательными WOE-преобразованиями (то есть делением всех признаков на бины и преобразованием датасета, вне зависимости от монотонности и WOE получившихся признаков), с избирательными WOE-преобразованиями (то есть с использованием значимых WOE-признаков, где значимость определяется монотонностью, или смыслом, хорошо укладывающимся в совокупную логику подбора признаков для задачи), и с WOE на основе IV (Information Value) - максимизации (где останутся только монотонные признаки с IV выше некоторого порога, который в работе установлен на уровне 0.02). Для краткости обращения, в дальнейшем я буду обращаться к каждому подходу по его порядковму номеру.

На каждом этапе работы обучена логистическая регрессия - будучи важнейшеймоделью классификации и классической для задачи скоринга, она выбрана за её интерпретируемость, простоту в реализации и хорошую работу с монотонными признаками (в контексте WOE-задачи), что позволяет легко объяснить полученные результаты и влияние различных факторов на вероятность дефолта.

Случайный лес используется благодаря его способности работать с данными без необходимости тщательной предобработки, а также за его высокую устойчивость к переобучению и возможность оценивать важность признаков. Он добавлен в качестве дополнительной модели для сравнения метрик на каждом этапе.

Градиентный бустинг применяется для достижения высокой точности предсказаний за счёт объединения слабых моделей, что позволяет эффективно справляться с более сложными паттернами в данных. Он обучен только на этапе 1 в качестве бонуса, когда данные наиболее сложны в интерпретации и выявлении паттернов.

В самой работе можно подробно ознакомиться с кодом и небольшими особенностями обучения для каждой модели.

## 3.2 WOE-преобразования

### 3.2.1 Теория

WOE (Weight of Evidence) преобразования - полезный инструмент предобработки данных в задаче кредитного скоринга. Метод используется для преобразовании признаков в значения, которые лучше отражают их взаимосвязь с вероятностью дефолта.

Логика WOE-преобразований заключается в том, чтобы для каждого признака сгруппировать значения (сформировать бины), имеющие схожие характеристики по таргету. После формиро-

вания бинов рассчитывается WOE для каждого из них по следующей формуле:

$$WOE_i = \ln\left(\frac{P(\text{good}|\text{bin}_i)}{P(\text{bad}|\text{bin}_i)}\right)$$

где $P(\text{good}|\text{bin}_i)$ и $P(\text{bad}|\text{bin}_i)$ - доли хороших и плохих заемщиков в данном бине соответственно.

С помощью таких преобразований можно, например, лучше выявить теоретическую линейную зависимость возраста от дефолта. Пусть у нас есть гипотеза, что пенсионеры более надежные заемщики - тогда поделим возраст на некоторые промежутки (18-25, 26-36, ..., 60-80), для каждого из которого подсчитаем долю дефолтов. Будем называть такую связь монотонной, если у признака для каждого бина доля дефолтов не меньше (не больше), чем у следующего бина.

WOE-признаки позволяют линейно разделить классы и получить более стабильные коэффициенты модели, что способствует улучшению метрик качества обучения моделей.

### 3.2.2 Применение и обучение

Преобразования осуществлены с помощью библиотеки scorecardpy. Она умеет автоматически делить признаки на бины для максимальной интерпретируемости, но так как в задаче важна и монотонность признаков - я самостоятельно дивагл границы вычисленных интервалов.

После обработки в датасете осталось 19 признаков. Ниде описаны изменения данных на каждом из четырех этапов.

- **Этап 1:** WOE-преобразования не применены
- **Этап 2:** Применение ко всем 19 признакам
- **Этап 3:** Применение к 13 признакам; для 5 признаков изменены границы бинов
- **Этап 4:** Применение к 9 признакам; для 5 признаков изменены границы бинов

Введём некоторые метрики для интерпретируемости результатов обучения модели:

- **Коэффициент Джини**

  Измеряет степень неравенства в распределении значений целевой переменной. Чем ближе к 1, тем лучше разделение классов моделью.

- **Точность (Accuracy)**

  Оценивает долю правильных предсказаний модели среди всех предсказаний. Подходит для сбалансированных классов.

- **Полнота (Recall)**

  Измеряет способность модели обнаруживать все положительные примеры. Важна для минимизации пропущенных дефолтов.

- **Точность (Precision)**

  Определяет долю правильно предсказанных положительных примеров среди всех положительных предсказаний модели. Важна для минимизации ложных срабатываний.

# 4 Таблица результатов

Отмечу что на всех этапах обучения я использовал подбор гиперпараметров с использованием GridSearchCV. Единственный шаг, на котором я это не сделал - на этапе 1, для сравнения результатов с оптимизацией гиперпараметров и без неё на одном и том же наборе данных.

Рассмотрим сначала метрики, которые дала эта модель на этапе 1:

| Метрика | Значение |
|---|---|
| Джини | 0.21705 |
| Accuracy | 0.823 |
| Recall | 0.000 |
| Precision | 0.227 |

Приведём результаты всех прочих экспериментов

| Джини | Этап | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| Лог | 0.34931 | 0.35729 | 0.35297 | 0.36444 |
| Дерево | 0.35259 | 0.33867 | 0.33233 | 0.35307 |
| Бустинг | 0.30518 | - | - | - |

Видно, что качество лорегрессии последовательно улучшалось при проведении преобразований. Случайный лес вёл себя по разному наборах данных. Бустинг, несмотря на выдвинутую гипотезу, оказался хуже в проведении предсказаний на самых сырых данных.

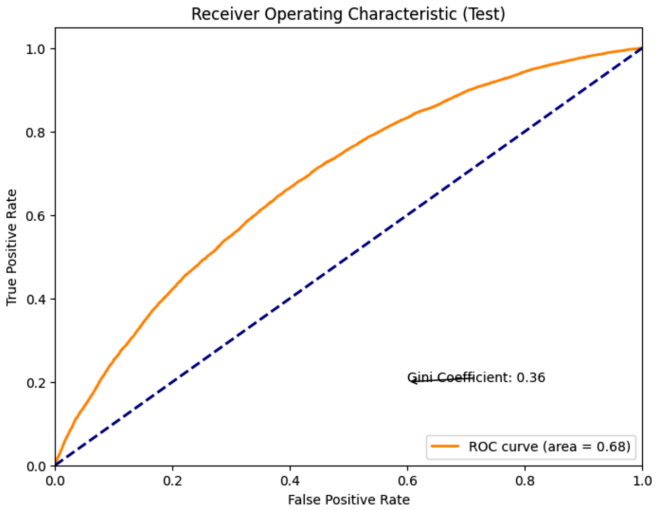
Для каждого из проведённых экспериментов в работе приведена матрица ошибок, а также Accuracy, Recall и Precision.

# 5 Валидация

Валидация проводилась на модели, лучшей по коэффициенту Джини - логистической регрессии этапа 4. Функции для построения представленных визуализаций, а также функции-светофоры, определяющие результат каждого теста, можно найти в ноутбуке.

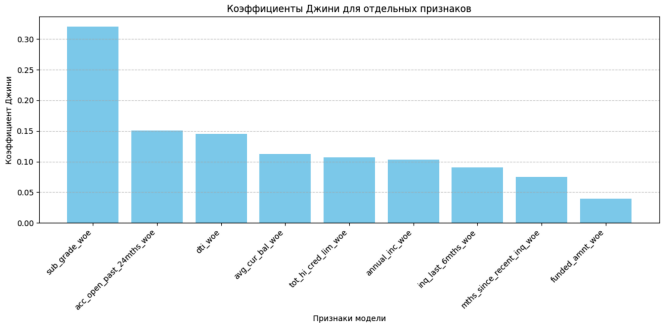
## 5.1 Эффективность ранжирования всей модели

Тест пройден - коэффициент Джини в итоговой модели составил 0.36444



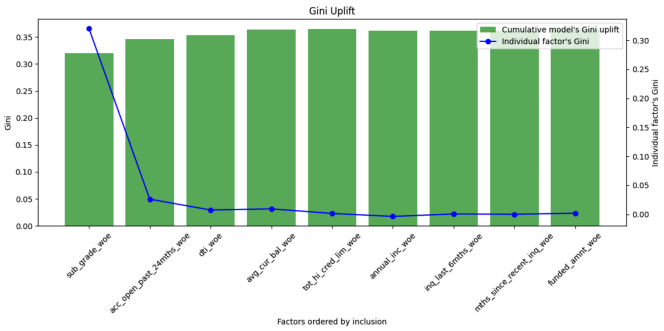

## 5.2 Эффективность ранжирования отдельных факторов

Тест пройден частично - поскольку funded_amnt_woe дал прибавку к Джини модели чуть меньшую, чем 0.05, а так признаков в лучшей модели немного - получилось так, что он один дал долю более 0.1 к желтым признакам, что, по критериям, является желтым цветом светофора валидации. Несмотря на это, я решил не удалять признак из итоговой модели - очень странно удалять информацию об объёме заёма из модели кредитного скоринга.



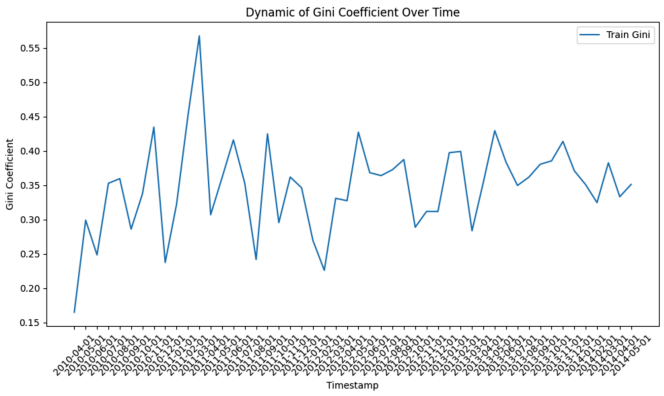
## 5.3 Анализ вкладов факторов в формирование Джини модели

Информационный тест - без светофора



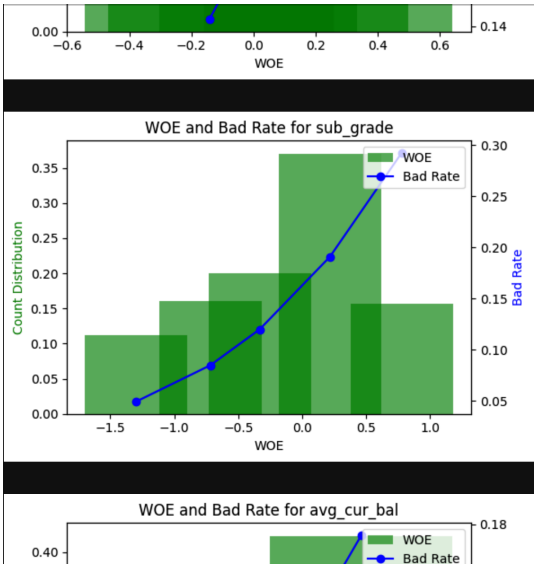

## 5.4 Динамика коэффициента Джини

Информационный тест - без светофора



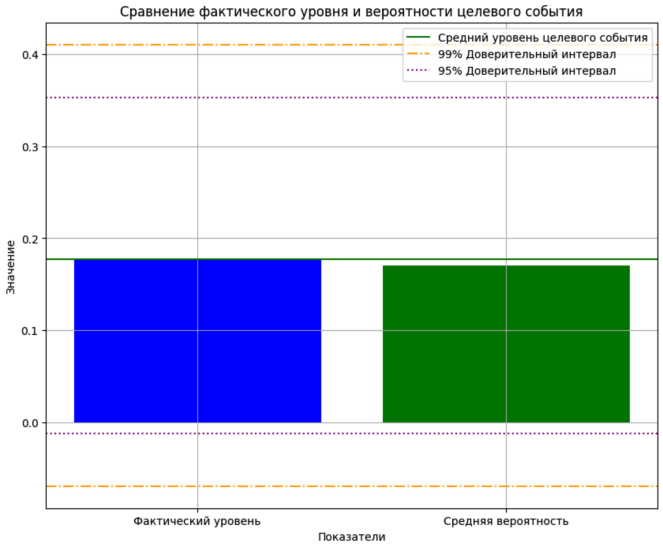
## 5.5 Анализ корректности дискретного преобразования факторов

Тест пройден - в ноутбуке можно ознакомиться с визуализациями для каждого графика (ниже приведу только пример), а также с датафреймом с разностями для каждого признака



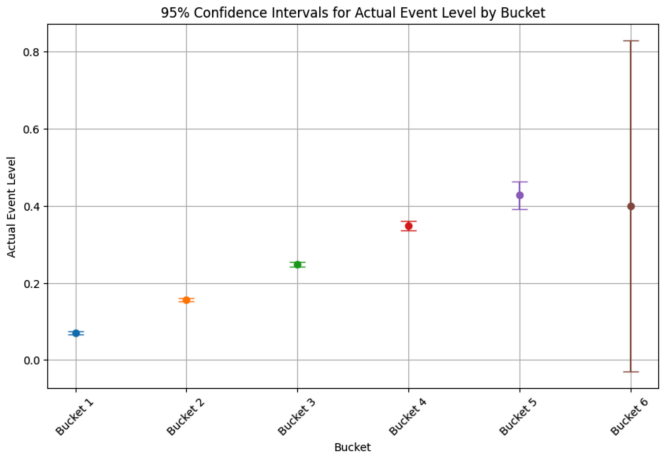

## 5.6 Сравнение прогнозного и фактического TR (Target Rate) на уровне выборки

Тест пройден - фактическая разница между фактическим уровнем и прогнозной вероятностью составила 0.00622




## 5.7 Тест формы калибровочной кривой

Тест пройден в соответствии с критериям светофора




## 5.8 Сравнение эффективности ранжирования модели на разработке и валидации

Тест пройден - абсолютная разница в Джини составила 0.00979, относительная - 0.02619

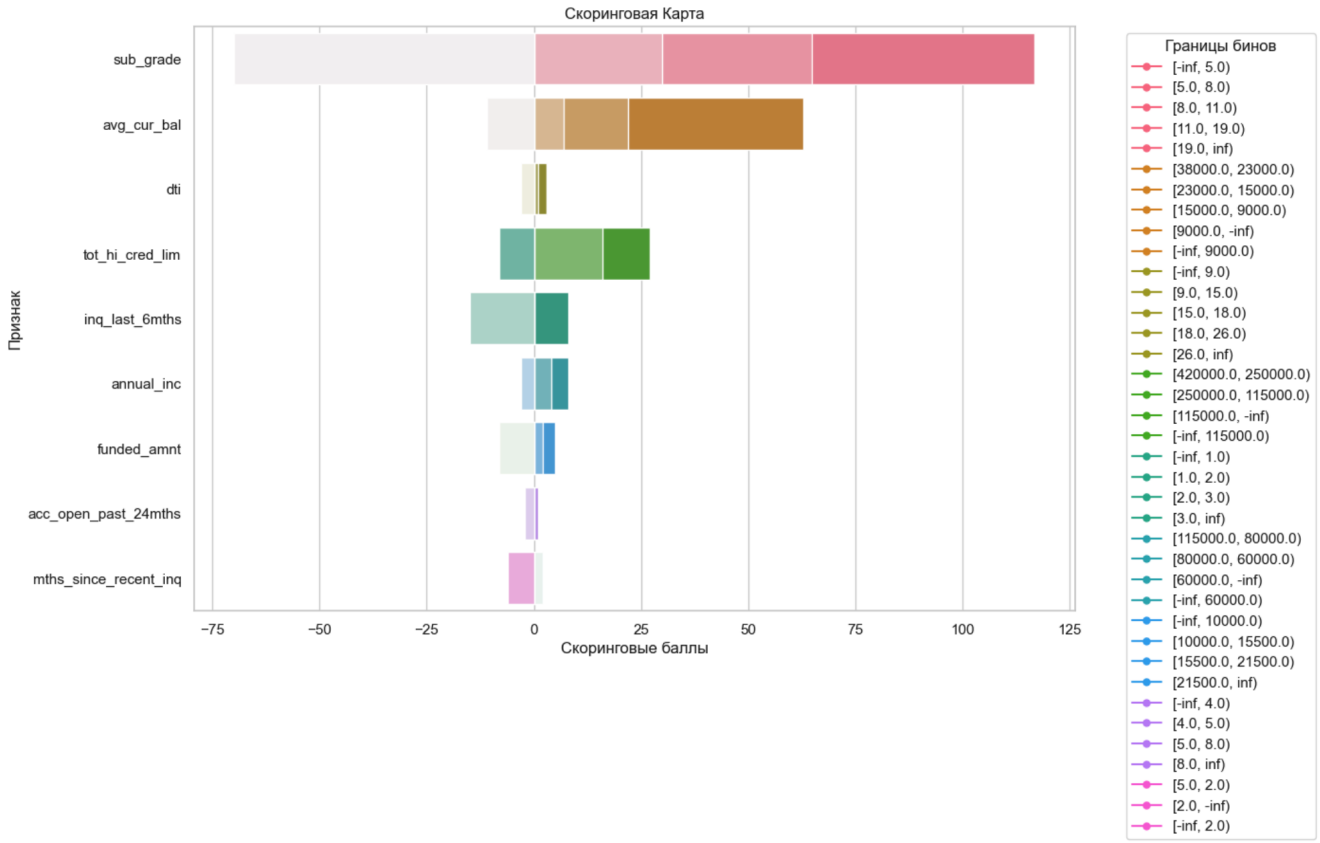

## 5.9 Сравнение эффективности ранжирования модели на разработке и валидации

Тест пройден - все факторы оказались зелёными, кроме acc_open_past_24mths

# 6 Скоринговая карта

Скоринговая карта - инструмент в скоринге, который помогает оценить кредитный риск заемщика на основе его характеристик. Она представляет собой таблицу или график, в котором каждый признак (например, возраст, доход, тд) разбивается на несколько бинов. Каждому бину соответствует определенное количество скоринговых баллов. Скоринговая карта позволяет оценить, какие характеристики влияют на решение о выдаче кредита, и как их значения соотносятся с рискованными или безрисковыми заемщиками

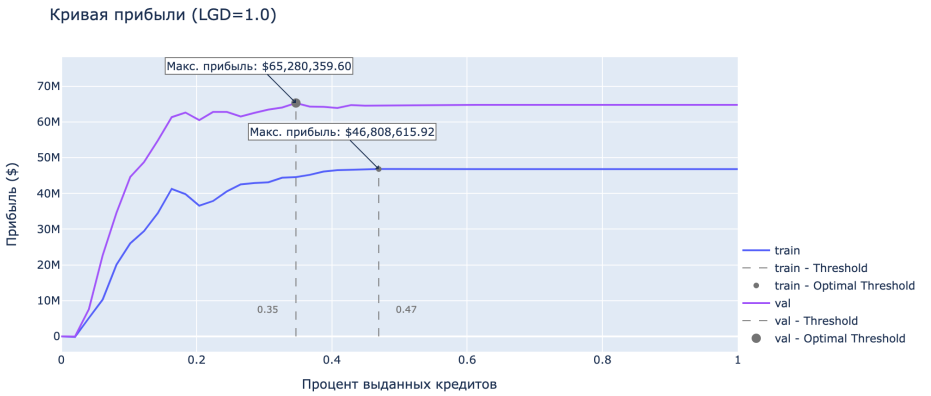
Ниже представлена скоринговая карта для модели, в которой каждый признак разбит на бины (в соответствии с WOE-преобразованием), а каждый бин имеет свой диапазон скоринговых баллов. Там же редставлены бины для каждого признака с их границами и соответствующими скоринговыми баллами.

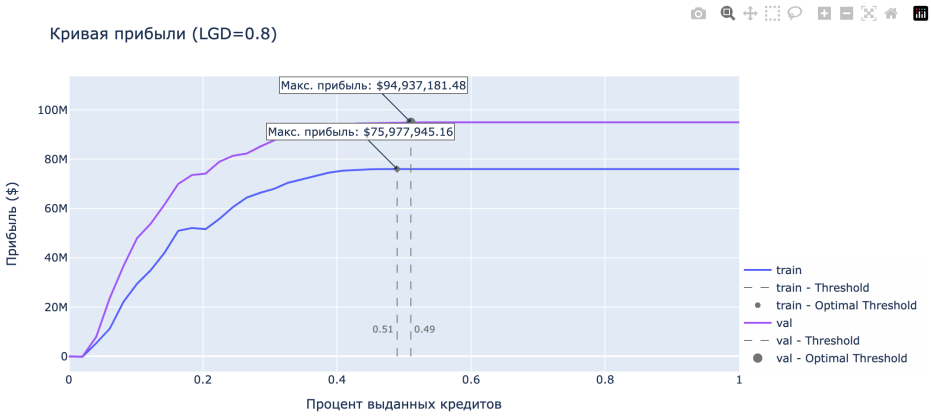



Видно, что наибольший вклад внсит признак subgrade, который определяется кредитным рейтингом заявщика. Далее идут текущий средний баланс и прочие признаки, для каждого из которых в легенде можно посмотреть тепловую карту бинов.

# 7 Подсчёт прибыли

Основываясь на предсказаниях, подсчитаны прибыль банка, который получает приведённые в датасете заявки. Выведено две кривых - одну для валидационных данных, другую - для тестовых. Рассчитан оптимальный порог выдачи кредитов для максимизации прибыли. Логика подсказывает, что валюта рассчётов - доллары США.




Видно, что по итогам анализа нашлась точка с максимальной прибылью как на валидационной, так и на тестовой выборке. При изменении LGD (то есть доли потерь банка в каждом кейсе дефолта с1 до 0,8), прибыль банка в каждой точке увеличилась, кривая сдвинулась, из-за чего и увеличилась прибыль.

Точка оптимального порога выдачи кредитов и суммарной прибыли представлена на каждом графике.