Образовательная платформа



Диплом по теме: “Прогнозирование кредитного дефолта”

|  |  |
| --- | --- |
| Выполнил работу: | Воронков А.С. |
| Наименование программы: | Машинное обучение: фундаментальные инструменты и практики |
| Код программы: | AML-18 |

Москва 2022

# Содержание

Введение. Проблема прогнозирования кредитного дефолта 3

1. Описание данных и используемые метрики 4

2. Первичная подготовка данных 5

3. Анализ данных и выбор наиболее подходящих для обучения 6

4. Модели машинного обучения 7

4.1 Дерево решений 7

4.2 Нейронные сети 7

4.3 Градиентный бустинг 8

Заключение 9

Список литературы 10

# Введение. Проблема прогнозирования кредитного дефолта

## Исходная задача и актуальность.

Будь то в ресторане или покупая билеты на концерт, современная жизнь рассчитывает на удобство кредитной карты для совершения ежедневных покупок. Это избавляет нас от необходимости носить с собой большие суммы наличных, а также может ускорить полную покупку, которая может быть оплачена со временем. Как эмитенты карт узнают, что мы вернем то, что мы взимаем? Это сложная проблема со многими существующими решениями и еще большим количеством потенциальных улучшений.

Прогнозирование дефолта по кредиту занимает центральное место в управлении рисками в бизнесе потребительского кредитования. Кредиторы оптимизируют кредитные решения, что приводит к улучшению качества обслуживания клиентов и надежной экономики бизнеса.

Необходимо спрогнозировать дефолт по кредиту. В данной задаче используется набор данных промышленного масштаба для построения модели машинного обучения.

Главными задачами являются предсказание вероятности того, что клиент не вернет остаток на своей кредитной карте в будущем, на основе его ежемесячного профиля клиента, улучшение качества обслуживания клиентов для владельцев карт, упрощение процедуры получения одобрения кредитной карты.

# Описание Данных и используемые метрики

## Описание исходных данных

Целевая двоичная переменная рассчитывается путем наблюдения за 18-месячным периодом выполнения после последней выписки по кредитной карте, и, если клиент не выплачивает причитающуюся сумму в течение 120 дней после даты последней выписки, это считается событием по умолчанию.

Всего признаков 190.

Набор данных содержит агрегированные характеристики профиля для каждого клиента на каждую дату выписки. Функции анонимизированы и нормализованы и подразделяются на следующие общие категории:

* D\_\* = переменные просрочки
* S\_\* = Переменные расходов
* P\_\* = Переменные платежа
* B\_\* = переменные баланса
* R\_\* = переменные риска

при этом следующие функции являются категориальными:

['B\_30', 'B\_38', 'D\_114', 'D\_116', 'D\_117', 'D\_120', 'D\_126', 'D\_63', 'D\_64', 'D\_66', 'D\_68']

Задача - предсказать для каждого customer\_ID из них вероятность дефолта по платежу в будущем (target = 1).

## Используемые метрики

Показатель оценки, *M* для этого соревнования является средним значением двух показателей рангового порядка: нормализованный коэффициент Джини, G, и показатель по умолчанию зафиксирован на уровне 4%, *D*.

*M=0.5(G+D)*

Показатель по умолчанию, зафиксированный на уровне 4%, представляет собой процент положительных меток (по умолчанию), зафиксированных в 4% прогнозов с наивысшим рейтингом, и представляет статистику чувствительности / отзыва.

Максимальное значение этого показателя равно 1.0.

# Первичная подготовка данных

Исходный тестовый набор весит 16 Gb, а тестовый порядка 32 Gb. Воспользуемся работой человека с ником raddar с сайта Kaggle (<https://www.kaggle.com/code/raddar/amex-data-int-types-train>). Автор этой работы заметил, что есть данные, у которых значения переменных очень близки к 0 и 1. Он приблизил их к этим значения, а также сделал labelencoding для текстовых данных. Затем преобразовал данные в формат parquet. В итоге получилось, что тестовый набор весит 1,7 Gb, а тестовый порядка 3 Gb. С этими наборами данных и будем работать.

Наборы данных для обучения, проверки и тестирования включают поведенческие данные временных рядов и анонимизированную информацию о профиле клиента. Для построения моделей машинного обучения нам необходимо обработать данные и избавиться от тех, которые нам не подходят. Будем рассматривать следующие категории данных:

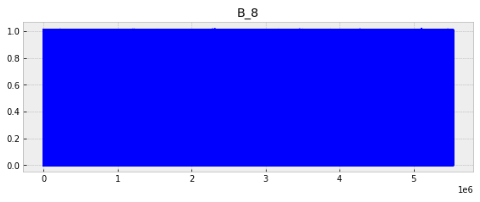
* с большим количеством пропусков более 25 процентов
* количество повторяющихся элементов больше 75 процентов
* низкие значения отклонений
* большое количество выбросов
* высоко коррелированные данные



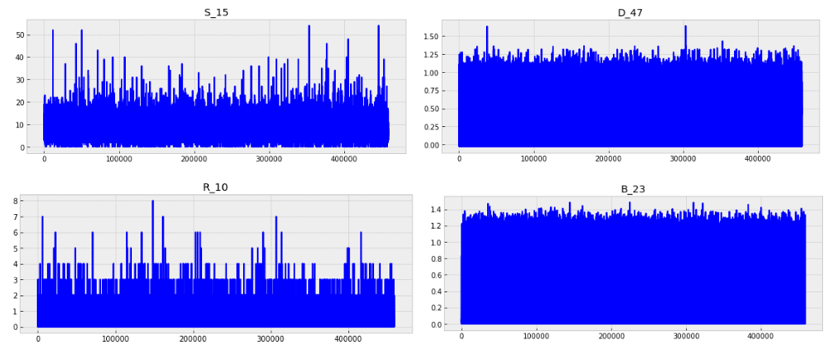
# Анализ данных и выбор наиболее подходящих для обучения

Проанализируем графики оставшихся данных и выберем наиболее подходящие для обучения.

Атрибуты 'B\_8', 'D\_52', 'D\_60', 'D\_64', 'B\_16', 'S\_13', 'D\_114' имеют подобный вид графика и не будут использованы



Наиболее подходящие для обучения являются параметры 'B\_23', 'D\_47', 'R\_10', 'S\_15'. Их и будем использовать.



# Модели машинного обучения

## Дерево решений

Мы будем использовать самое обычное дерево решений с глубиной 10. На обучающей выборке получились следующие значения:



На тестовой выборке и загрузке данных на сайт Kaggle получились следующие результаты:



## Нейронные сети

Второй выбранной мною моделью будет нейронная сеть.

На обучающей выборке получаются следующие результаты:



На тестовой выборке и загрузке данных на сайт Kaggle получились следующие результаты:



## Градиентный бустинг

Используем классический градиентный бустинг.

Качество на обучающей выборке:



На тестовой выборке и загрузке данных на сайт Kaggle получились следующие результаты:



# Заключение

Как видно из используемых моделей самой эффективной является градиентный бустинг.

Для улучшения качества модели можно попробовать использовать другие параметры, улучшить качество анализа данных, их обработку и использовать более совершенные модели.

В примерах наилучшего решения этой задачи зачастую используют градиентный бустинг. Например XGBoost.

# Список литературы

1. <https://www.kaggle.com/competitions/amex-default-prediction/overview>
2. <https://www.kaggle.com/code/raddar/amex-data-int-types-train>
3. Введение в машинное обучение с помощью Python. Андреас Мюллер, Сара Гвидо.
4. Python и машинное обучение. Себастьян Рашка, Вахид Мирджалили
5. Практическая статистика для специалистов Data Science. Питер Брюс и Эндрю Брюс