Альфа Банк

Задача кредитного скоринга клиентов

Предскажите, какие клиенты выйдут в дефолт по кредитному продукту, основываясь на собранных о них данных. Вам необходимо разработать модель, которая решит задачу кредитного скоринга на основании данных последовательных действий клиента.

Постановка задачи

Разработайте модель кредитного скоринга. Модель должна определять вероятность выхода клиента в дефолт по кредитному продукту. Для этого:

01 Изучите особенности проведения кредитного скоринга в Альфа-Банке.

02 Изучите предоставленные данные о клиентах и проанализируйте взаимосвязи между признаками и целевой переменной.

03 Постройте модель на тренировочных данных и оцените ее качество.

04 Сделайте предсказание для тестовой выборки и сохраните его в формате CSV.

05 Отправьте решение на платформу в виде файла-предсказания и узнайте его качество в виде оценки с помощью метрики ROC-AUC.

06 Доработайте модель, отправьте улучшенные предсказания для повышения результата.

ВАЖНО!

Вы имеете право загружать ваше предсказание на платформу в формате CSV не более 3 раз в течение одного дня (в промежутке между 00:00 и 23:59 мск).

Бизнес-контекст

Перед выдачей клиенту кредита банк проводит оценку вероятности возврата выданных средств. Для оценки вероятности возврата проводится кредитный скоринг клиента. Задача кредитного скоринга состоит в том, чтобы оценить вероятность ухода клиента в дефолт по кредитному продукту. Дефолт клиента — это ситуация, когда заемщик прекращает платить по кредиту в течение определенного периода.

Специфика сбора данных для проведения кредитного скоринга состоит в том, что целевая переменная становится известной только спустя год. По прошествии 12 месяцев становится однозначно возможно определить, вышел клиент в дефолт или нет.

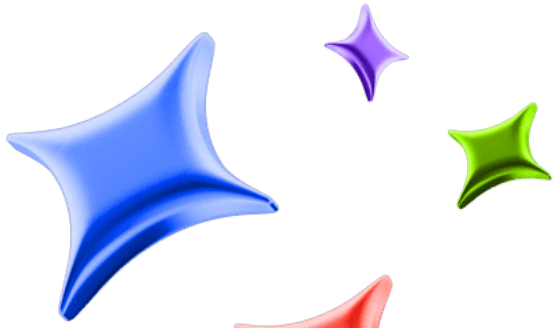
В качестве последовательных данных для обучения модели банку доступны:

- По умолчанию:
  - Данные банковских транзакций.
  - Данные транзакций расчетного счета.
  - Данные об истории коммуникации.
  - Логи приложения и сайта.
- С согласия клиента:
  - Данные Бюро кредитный историй (БКИ).
  - Чеки по произведенным оплатам.

Классическое определение дефолта: 90@12 (просрочка на 90 дней в течение 12 месяцев).

Факторы, влияющие на кредитный скоринг клиента:

- Финансовое поведение и структура расходов.
- Данные о транзакциях и движениях средств на счетах.
- Кредитная история.
- Финансовая оценка окружения клиента.
- Персональная информация о клиенте.



Для оценки бизнес-эффективности моделей кредитного скоринга в Альфа-Банке используется метрика Джинни.

Метрика Джинни (или коэффициент Джинни) — это показатель качества модели кредитного скоринга, который используется для оценки ее дискриминационной способности, то есть способности различать «хороших» и «плохих» заемщиков.

$$Gini = (ROC\ AUC - 0.5) \times 2$$

Gini принимает значения от 0 до 100:

100 — идеальная модель, которая полностью разделяет заемщиков по уровню риска.
50–70 — хорошее качество модели, обычно встречается в банковских скоринговых системах.
0 — модель ранжирует случайным образом, не обладая предсказательной силой.

Метрика Джинни применяется в банках для оценки качества скоринговых моделей. Она показывает, насколько уверенно модель отделяет надежных заемщиков от тех, кто с высокой вероятностью уйдет в дефолт. Чем выше значение Джинни, тем лучше модель позволяет банку принимать обоснованные решения о выдаче кредитов.

Бизнес мыслит в терминах Джинни, поскольку каждый пункт Джинни имеет прямую конвертацию в прибыль. По сути, специалисты по Data Science и Machine Learning решают задачу бинарной классификации, оценивают качество моделей с помощью ROC-AUC, а демонстрируют результаты бизнесу с помощью метрики Джинни.

Критерии успеха

Основным критерием успешной модели считается умение качественно предсказать вероятность выхода клиента в дефолт. Для оценки качества модели применяется метрика ROC-AUC, а само предсказание реализуется на тестовых данных, что позволяет независимо оценить качество модели, а также понять, является модель недо- или переобученной.

Ограничения

Для создания и настройки модели вы можете использовать язык Python версии 3.10 и выше. Также вы можете использовать любую библиотеку для машинного обучения и предобработки данных, являющуюся OpenSource-ресурсом. Использование закрытых библиотек, частных API и чужого кода, не подпадающего под разрешение о свободном распространении и использовании, запрещено.

Источники информации

(Ссылка на данные на Яндекс Диск)

Полный датасет был разделен на несколько датасетов одинаковой структуры и разного содержания:

- 12 тренировочных датасетов, пронумерованных в формате **train_data_номер.parquet**.
- 2 тестовых датасета, пронумерованных в формате **test_data_номер.parquet**.
- Каждый из датасетов содержит следующие признаки:
  - id — уникальный анонимный идентификатор клиента;
  - rn — номер строки из выгруженных данных;
  - enc_col_0 - enc_col_56 — закодированные признаки данных о клиенте.
- Файл **train_target.csv** — файл с конечными классами клиентов, информация о которых описана в тренировочных данных. Содержит признаки:
  - id — уникальный анонимный идентификатор клиента;
  - target — целевая переменная (1 — факт ухода клиента в дефолт).
- Файл **test_target.csv** — файл о клиентах, значение целевой переменной для которых только предстоит определить. Содержит признаки:
  - id — уникальный анонимный идентификатор клиента.
- Файл **sample_submission.csv** — пример сабмита для отправки на платформу. Содержит признаки:
  - id — уникальный анонимный идентификатор клиента;
  - target — предсказание модели.

Мы не предоставляем baseline-модель напрямую, однако идеи решений вы можете почерпнуть из уже проведенных соревнований и сохраненных baseline в GitHub — здесь и здесь.

Для участия в отборе необходимо получить качество в виде метрики ROC-AUC выше 77.2 пунктов на приватной части лидерборда.

Пространство решений

Для обучения модели вы можете использовать только данные, предоставленные в тренировочной и тестовой выборке. В своем решении вы можете:

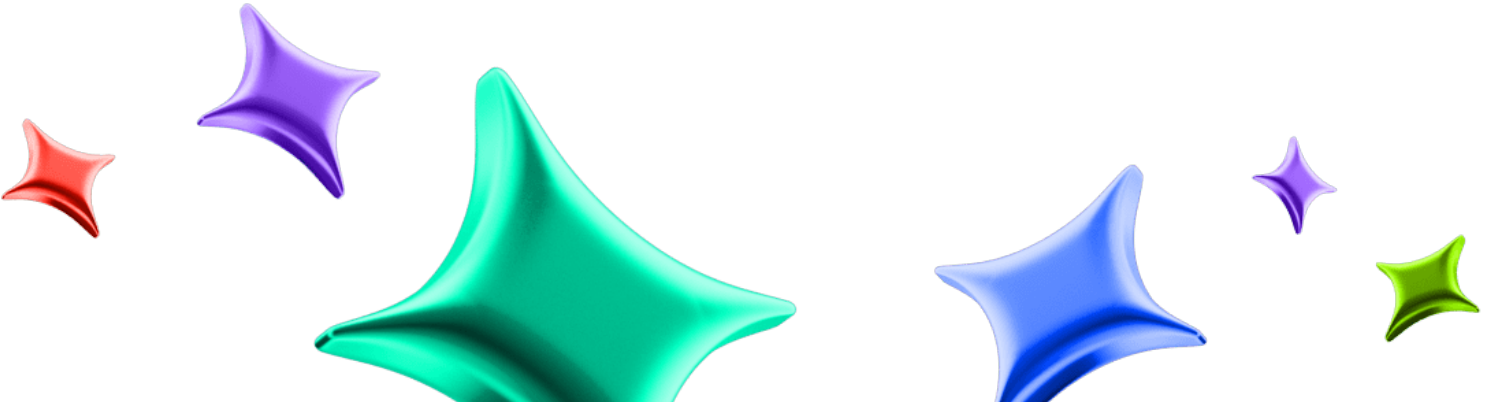
01 Обучить модель на части датасета.

02 Использовать для обучения модели только часть признаков.

03 Использовать сгенерированные в ходе предобработки признаки.

04 Проводить селекцию и отбор признаков.

Вы имеете право использовать любую модель из любой Open-Source-библиотеки для предсказания. Выбор и настройка модели не ограничены организаторами.



Приложение 1. Метрика ROC-AUC

Результат метрики roc-auc

Метрика *ROC-AUC* отражает качество модели бинарной классификации. Кривая ROC иллюстрирует производительность классификационной модели при всех порогах классификации. Для построения кривой ROC в качестве оси X берутся значения ложноположительной частоты FPR, которые рассчитываются по формуле:

$$FPR = \frac{FP}{(TN+FP)},$$

где *FP* — ложноположительные результаты, а *TN* — истинно отрицательные. Значение по оси Y рассчитывается как истинно положительная частота — TRP, или recall, которая рассчитывается по формуле:

$$TPR = \frac{TP}{(TP + FN)},$$

где *TP* — истинно положительные результаты, а *FN* — ложно отрицательные.

Показатель *AUC* — это сумма производительности модели, представленная в виде одного числа, т. е. измеренная площадь под кривой *ROC*, которая колеблется в диапазоне от 0 до 1, где с увеличением значения увеличивается и производительность модели.

CHANGELLENGE >>

Задание написано и опубликовано Changellenge >> — ведущей организацией по кейсам в России.

www.changellenge.com

Альфа Банк

Задание создано по заказу АО «Альфа-Банк»

www.alfabank.ru