1 слайд

1. Роли в команде:
2. Анализ входных данных
   1. Удаление столбцов, заведомо не влияющих на результат – например, ‘account\_uid’
   2. Аггрегация данных – важный шаг, который во многом окажет влияние на выбор столбцов для анализа. Здесь мы должны будем выбрать аггрегационные функции для каждого из столбцов.
   3. После аггрегации необходимо произвести анализ столбцов на миссинги, при недостаточном количестве информации в столбцах (много миссингов) – рассмотреть возможность удаления столбца из датасета
3. Создание датасетов
   1. Генерация фичей: **осмысленные** арифметические действия с уже существующими данными и сохранение их в новые столбцы (увеличение признакового пространства) – сложение/вычитание/умножение/деление, максимумы/минимумы, подсчёт количества дней до какого-либо события итд.
   2. Составление описания фичей в виде комментариев в коде
   3. Объединение аггрегированного датасета с таргетами – по ключу (application\_id, client\_id)
   4. Датасет является несбалансированным в сторону клиентов без дефолта (23605 против 6395 клиентов). Может возникнуть необходимость в сбалансировании данных (например, Oversampling – вручную путем увеличения количества клиентов с дефолтом в 3,69 раза до достижения 23600 клиентов, либо с использованием метода RandomOversampler библиотеки imbalanced-learn (imblearn))

2 слайд

1. WOE-преобразование / биннинг фичей – вручную, либо с использованием библиотеки Optbinning
2. Однофакторный анализ датасетов (IV, Gini, PSI), Gini считаем во времени с использованием столбца 'reporting\_dt'
3. Предварительный отбор фичей с соблюдением условий заказчика:
   1. IV> 0.01;
   2. Gini > 0.03;
   3. Gini во времени стабилен и не падает до 0 ни в одном из периодов;
   4. PSI <= 0.10.
4. Многофакторный анализ и отбор фичей в модель (не более 10 штук)
   1. Планируем сравнить работу классических моделей МО, ансамблей (bagging, blending, stacking), бустингов, нейронных сетей, с различными вариантами гиперпараметров (GridSearch)
5. Оценить итоговой моделью все 3 выборки (трейн/тест и неразмеченную выборку) и посчитать тесты по модели (PSI, Gini, Gini во времени, Хосмер-Лемешов)
   1. Запаковка обученной модели, показавшей наилучший результат, в файл pickle

3 слайд

1. Оформление подробного word-отчета с указанием всех шагов, графиками, таблицами со статистиками и документированием всех использованных скриптов.
2. Финализация презентации
3. Передача документации и презентации заказчику