Отчёт по соревнованию «Прогнозирование вероятности невозврата кредита»

Каюмов Эмиль, 517
ММП ВМК МГУ
Курс «Прикладные задачи анализа данных»

Задача

- Задача предсказания прогнозирования вероятности невозврата кредита по кредитной истории клиентов
- Данные:
  - Кредитная история клиентов из разных БКЮ
- Метрика: ROC-AUC

Базовый подход

- Простой способ агрегации записей по клиентам — подсчёт статистик по всем записям из кредитной истории каждого клиента

- Возможно использование различных функций (среднее, сумма, минимум, максимум, среднеквадратичное отклонение, медиана)

- Можно делать взвешенную сумму, где веса будут соотноситься с датами

Дубли в истории

- Некоторые записи из разных бюро кредитных историй дублируют друг друга
- Информация может быть противоречивой
- Попытки автоматической чистки не дают выигрыша в результате, поэтому принято решение не трогать

Текстовый признак

- Один из признаков характеризует своевременность платежей и несёт много информации о платёжеспособности и дисциплинированности

- Простые признаки: частоты символов, длины строк, отношения частот, …

- Сложные признаки: логистическая регрессия на tfidf и мешке слов, суммирование индексов каждого символа, …

- Аналогичная агрегация статистиками между записями

Активные кредиты

- На текущее состоянии клиентов влияют текущие открытые кредиты — можно работать только с записями по активным кредитам

- Качество на датасете по активным кредитам немного ниже (при этом там не все клиенты), но объединение двух датасетов дало прирост

Прочее

- По нескольким категориальным признакам считались сглаженные счётчики перед агрегацией

- Несколько заведомо некорректных дат были исправлены вручную, несколько признаков были удалены

Финальная модель

- На полученных 500-600 признаках обучался LightGBM с двумя наборами параметров с разными отложенными выборками, все предсказания усреднялись

- **Private LB: 0.70502 (17 место)**