Отчёт по соревнованию Ticketland ML Contest

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

Задача

- Задача предсказания вероятности клика на шоу из поисковой выдачи ticketland.ru

- Данные:
  - История кликов по выдаче за 3 месяца
  - История кликов вне выдачи
  - Базовые характеристики шоу, мест проведений, пользователей
  - Картинки, соответствующие шоу
- Метрика: LogLoss

Базовый подход

- Join всех таблиц
- Сглаженное кодирование категориальных переменных
- Несколько признаков из кликов вне выдачи
- Lightgbm
- **Private: 0.389**

Особенность в данных

- Некоторые пары и тройки по user_id, show_id, datetime встречаются в выборке слишком часто

- Добавление признаков на их основе (count после groupby, метапризнаки через логистическую регрессию) ведут к переобучению и ломают валидацию

- **Private: 0.357**

Честная + переобученная модели

- Несмотря на особенность в данных, честная модель позволяет делать хорошие предсказания для обычных объектов

- Смешивание двух моделей с весами позволяет заметно улучшить результат за счёт баланса между необычными и обычными объектами

- Подход с исправлением вероятностей у необычных объектов не привёл к успеху

- **Private: 0.274**

Докрутка результатов

- Исключение отдельных переобученных признаков, смешивание разных моделей и ручной подбор весов поднимали результат

- На этот момент результат перестал быть воспроизводимым…

- **Private: 0.254**

Финальная модель

- Также полезным оказался признак, учитывающий длину поисковой выдачи (но приводил к искажению распределения вероятностей)

- На 3 датасетах (базовый, с переученными признаками и с длиной выдачи) обучались LightGBM и логистическая регрессия

- Предсказания смешивались с весами из головы на основе опыта предыдущих сабмитов (больший вес у второго датасета и небольшой у 1 и 3)

- **Private: 0.253 (4 место)**