

Отчет по контексту Ticketland

Таскынов Ануар, 517 группа

Курс «Прикладные задачи анализа данных»

21 января 2018 г.

Даны:

Информация о пользователях, событиях, о том, какие события встречались данному пользователю и т.п. В обучении была дана информация с 12 января по 31 марта, в тесте с 1 апреля до 30 апреля. Также были даны афиши, которые были успешно проигнорированы в решении.

Задача:

Необходимо по этим данным предсказать вероятность клика пользователем по афишам.

Функционал качества:

Log-loss.

Обработка данных:

- Были заменены некорректные значения возраста пользователя.
- Из дат оставить только день.

Признаки:

- Были замечены накрутки (А. Попов): пользователь мог несколько раз кликать на одно и тоже объявление в течении короткого промежутка времени. Было разумно ввести примерно такой признак: количество троек (пользователь, шоу, время). Или па: количество пар (пользователь, время). Такой же логикой можно ввести признак: количество троек (пользователь, шоу, время) до данного момента.
- Добавлен признак: какое количество раз шоу в поисковой выдаче было показано первым по списку пользователю. Таким же способом, как описано выше можно добавить время: сколько раз до данного времени были замечены данные двойки-тройки.
- По данным двойкам и тройкам посчитаны ctr.

Модель:

Один LGBM со стандартными параметрами.

Итог: 0.279 на Public (на Private место утеряно, так как boosters.pro временно не доступен).