

# Закономерности в задаче "Ticketland"

Амир Мирас

МГУ имени М.В. Ломоносова

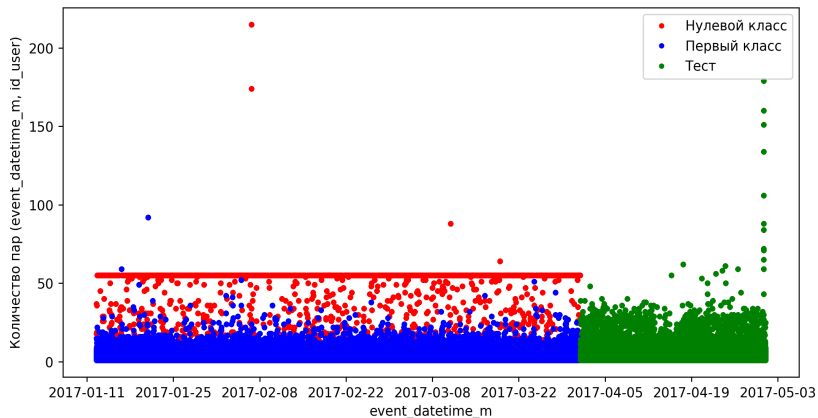
14 января 2018 г.

## Простые бенчмарки

1. F: признаки по шоу и пользователю из исходных данных
2. BOW: bag of words над жанрами и местом проведения шоу
3. OHE: OneHotEncoder над пользователями и шоу

Алгоритм	Признаки	log loss на PL
lightgbm	F	0.381234
logisticregression	F + BOW + OHE	0.383949
libffm	F + BOW + OHE	<b>0.375489</b>

## Количество пар (время события, пользователь)



Явно выделяется красная линия с количеством пар 55: для пользователя 1555205 существуют 46860 минут, где он не кликал 55 раз. Скорее всего это бот !

## Более сложные признаки

1. Количество пар признаков  $\{rank, id\_user, id\_show\}$
2. Количество пар признаков  $event\_datetime\_m + \{rank, id\_user, id\_show\}$
3. Сглаженные средние ответа по  $\{rank, id\_user, id\_show\}$  по предыдущему месяцу
4. Сглаженные средние из датасета *clicks\_no\_impressions* по  $\{rank, id\_user, id\_show\}$  по предыдущему месяцу
5. Метапризнаки, как выходы алгоритмов *LGBMClassifier*, *LGBMRegressor*, *XGBRegressor*, *ExtraTreesClassifier*, *RandomForestClassifier*
6. Будем обучаться только на марте и предсказывать апрель

## Итоговая модель

На полученных признаках обучаем 10 *LGBMClassifier* с разными *random\_state* и усредняем.

1	Максим Келаскин	53	0.24425	2017-12-15 16:53:19
2	Василий Рязанов	17	0.252409	2017-12-17 22:55:02
3	Андрей Костин	27	0.252619	2017-12-17 17:30:11
4	Эмиль Каюмов	27	0.252923	2017-12-17 23:52:58
5	Макс Щёголев	24	0.254822	2017-12-17 14:04:49
6	Мирас	40	0.255858	2017-12-17 14:19:15