Kaggle Santander Value Prediction Challenge

Предсказание размера транзакций потенциальных клиентов

Анатолий Ильенков

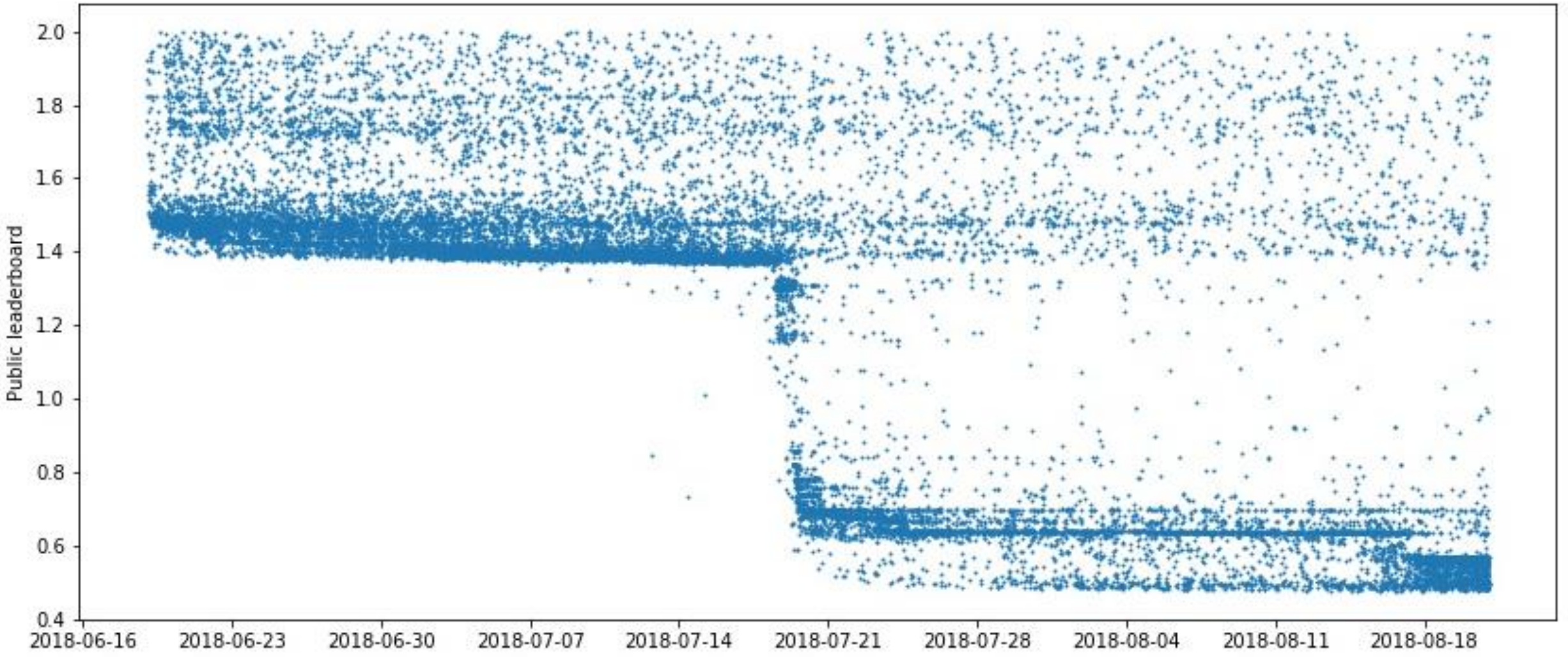
О себе и о ~~моей новой книге~~ соревновании

- ✓ Экономический факультет МГУ
- ✓ Специализация по ML на Coursera от Яндекса и МФТИ
- ✓ Места public/private leaderboard – 39/3 из 4 484

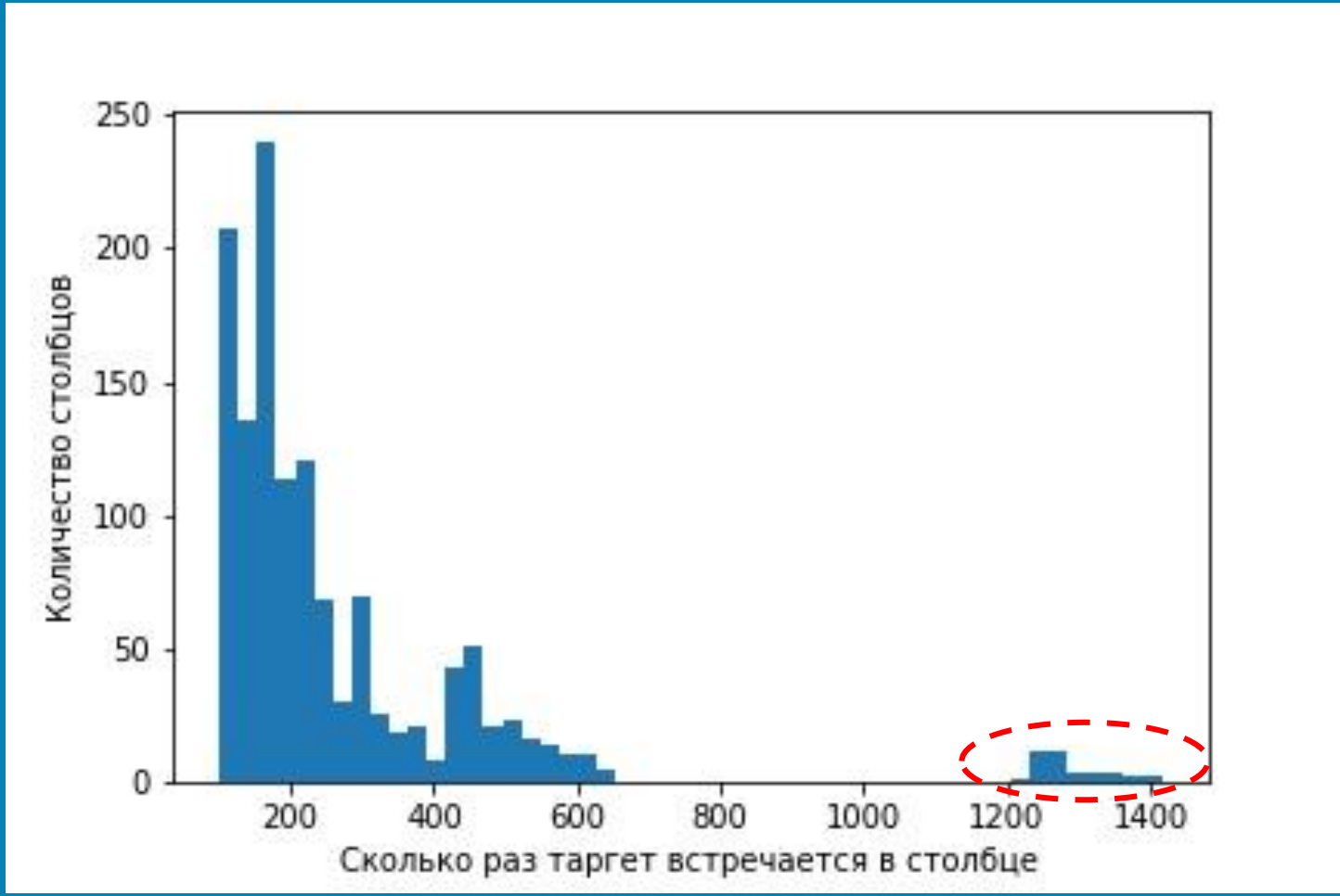
Данные

- ✓ Train 4 459 строк × 4 993 столбца
- ✓ Test 49 342 строки × 4 992 столбца
- ✓ Public / Private - 49/51
- ✓ Данные анонимизированы
- ✓ Видимо, около 80% теста не участвовала в метрике
- ✓ Задача регрессии
- ✓ Метрика RMSLE

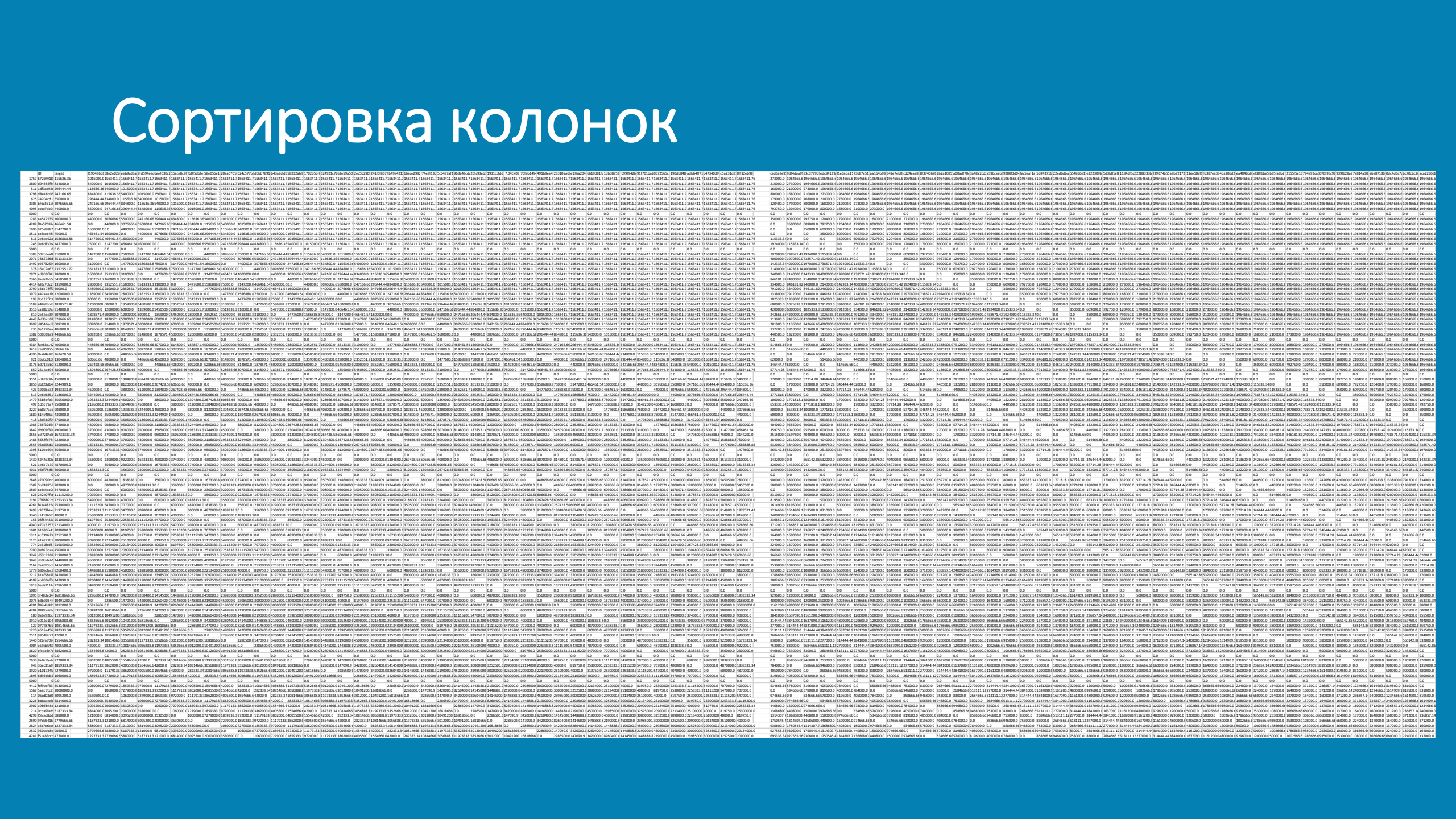
Динамика соревнования



Лик


250
200
150
100
50
0
Количество столбцов
200
400
600
800
1000
1200
1400
Сколько раз таргет встречается в столбце

Сортировка колонок



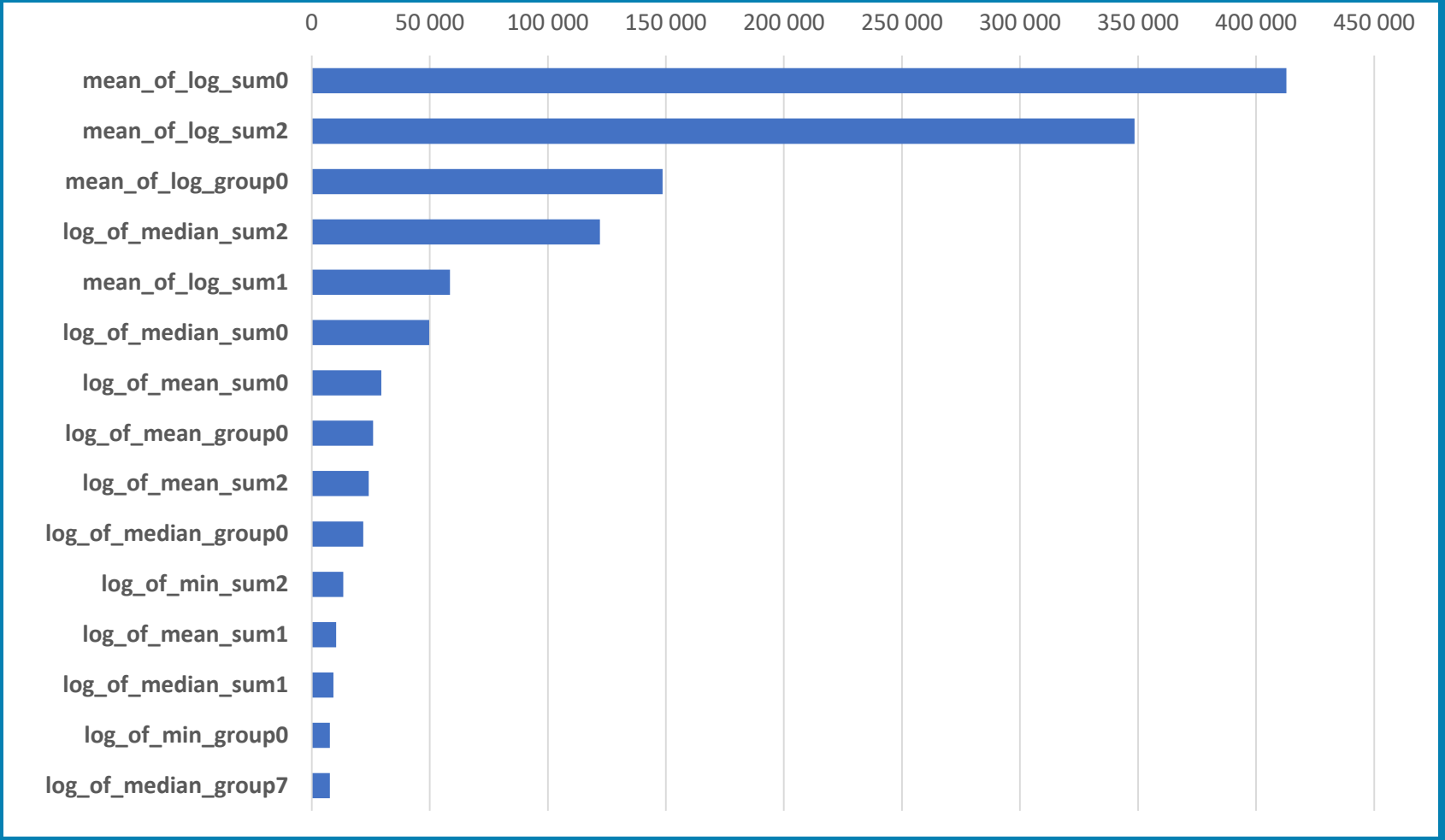
Результаты лика по 110 группам колонок

- ✓ lag 35
- ✓ Количество ликов в трейне и тесте: 3895 7844
- ✓ % правильных ликов в тесте: 99.846 %
- ✓ Local result / Public / Private : 0.564 / 0.553 / 0.606

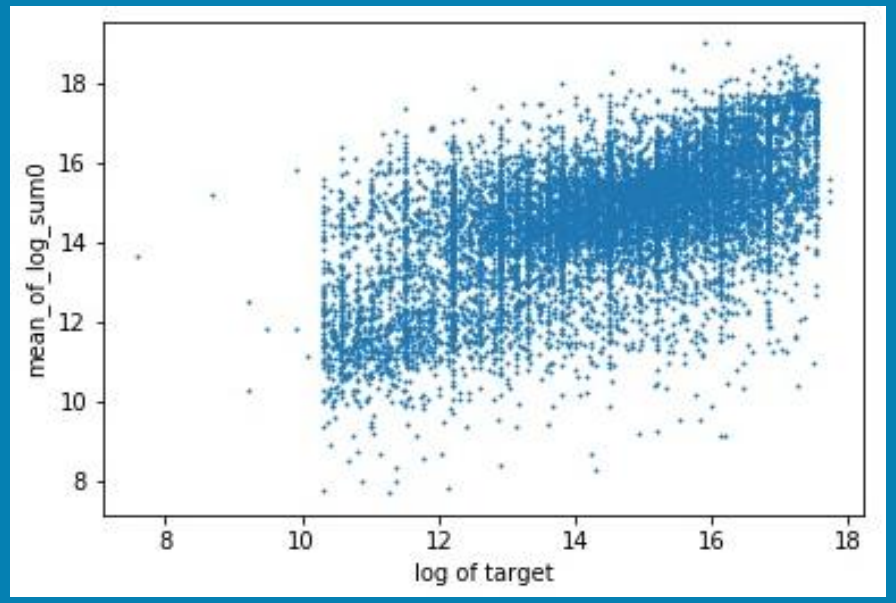
LightGBM и фичи

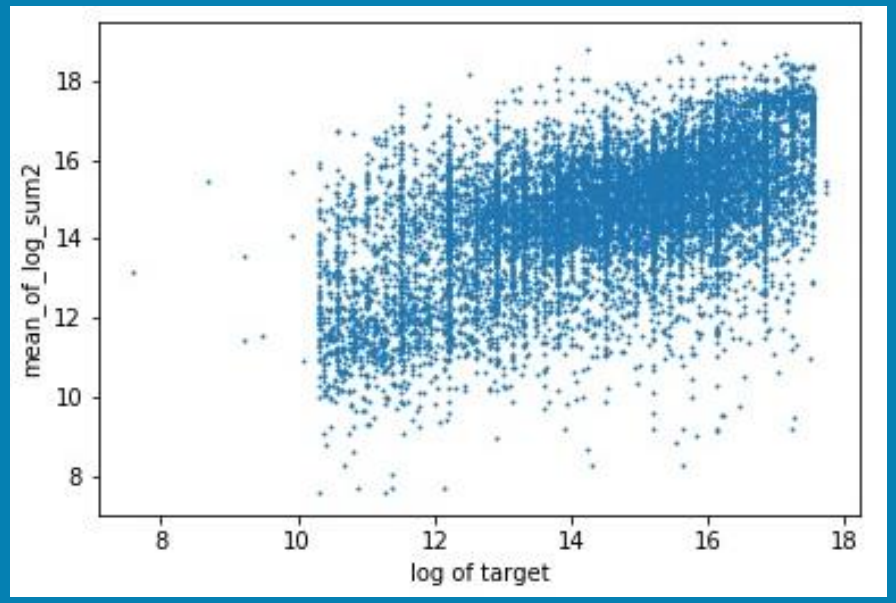
- ✓ Аугментация трейна ликом
- ✓ Оригинальные колонки
- ✓ Статистические фичи (логарифм среднего, среднее логарифмов, логарифм медианы, логарифм максимума и минимума, количество нулей, стандартное отклонение, сумма, коэффициенты асимметрии и эксцесса)
- ✓ По всем колонкам, по группам колонок и по некоторым суммам пар и троек групп колонок
- ✓ Отбор фич по значимости в LightGBM

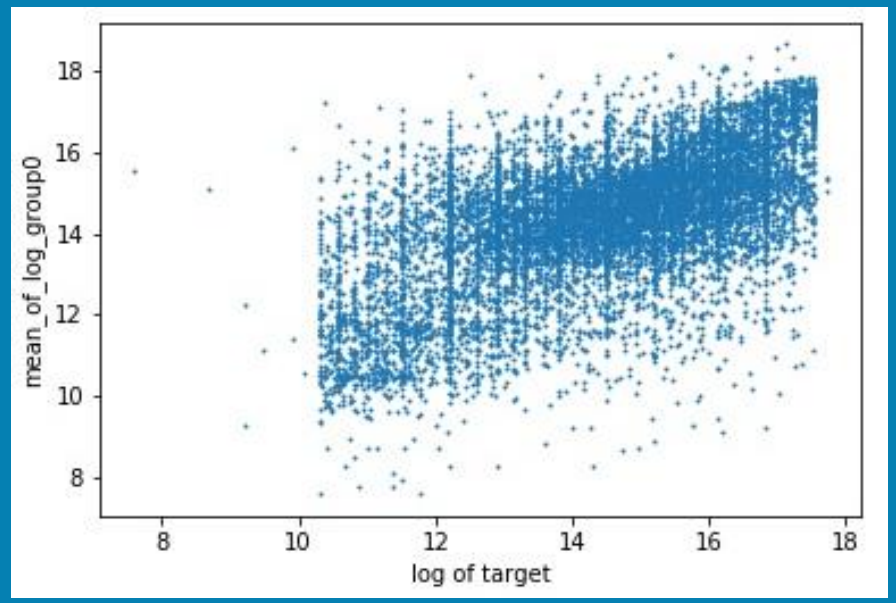
Feature importance

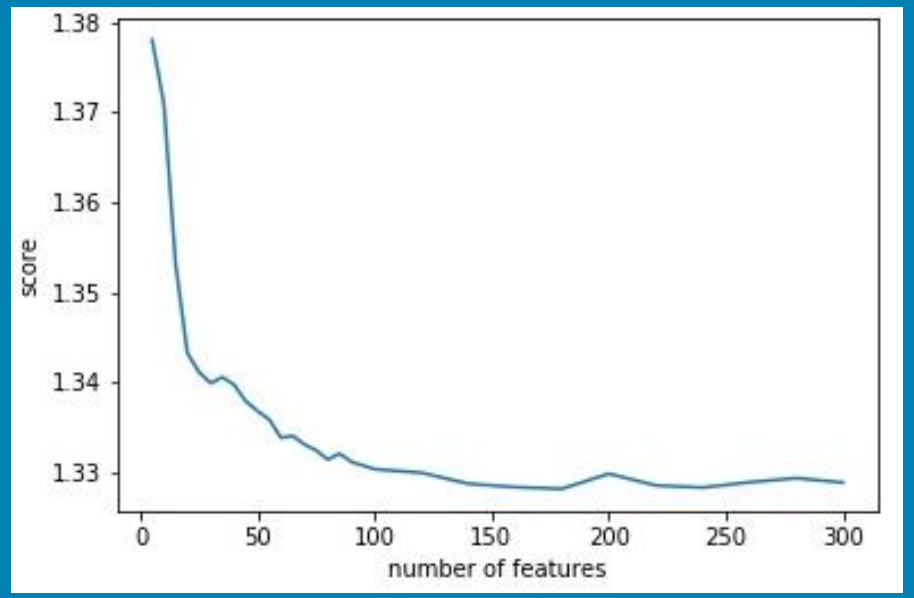


Scatter plots for most important features









Блендинг

- ✓ 23 модели LightGBM и XGBoost
- ✓ Подбор весов по out-of-fold валидации на 4 fold’ах

Добавление ML моделей и блендинг:

- ✓ Public 0.553 -> 0.487
- ✓ Private 0.606 -> 0.524

Прочее

- ✓ В ходе соревнования был исправлен баг, позволявший снимать галочку выбора решений после окончания соревнований

Спасибо!