

PIK Digital Day

Хакатон в Kaggle-стиле

Финал - 14.07.2018

Команда



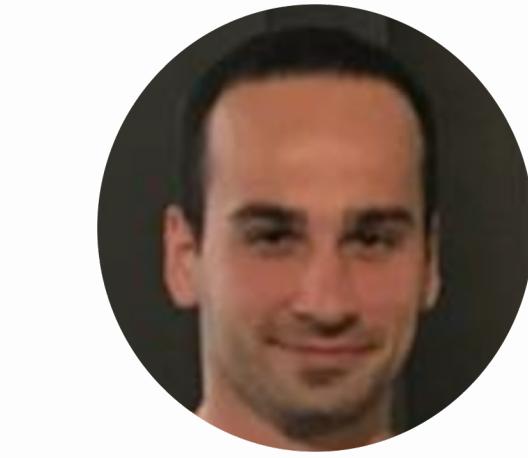
Сергей Белов
–МФТИ, Skoltech
–Венчурный фонд Techsensor,
финансовый аналитик



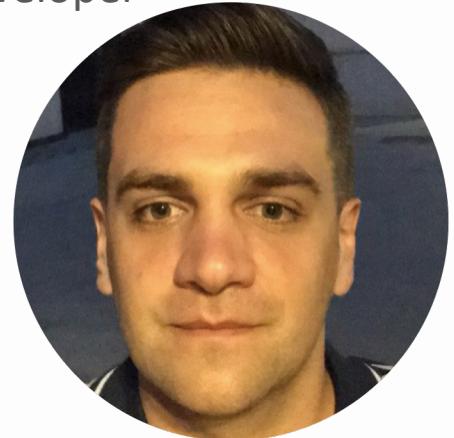
**BBDO
GROUP**



Александр Дроботов
–РЭУ им. Г.В. Плеханова,
экономист-математик
–BBDO Group, эконометрист



Алексей Смирнов
–СГАУ
–Maxifier, Software developer



Алексей Каюченко
–Quantitative Finance, SBS-EM
–PMI, Data Science manager

Задача



Предсказание темпов продаж квартир для компании ПИК

В качестве обучающего датасета в первом этапе предоставляется история продаж квартир за 2,5 года по 30 проектам в Москве и Московской области.

Online - этап

255 участников

Наши места

-53	-56
-54	-115

Offline - этап

43 команды – модель без ограничений

25 команд – линейная регрессия

Результат

2-е место в конкурсе на лучшую модель без ограничений

1-е место в конкурсе на лучшую модель линейной регрессии



Данные

TRAIN

shape = (9244, 55)

Target value

Количество проданных квадратных метров в месяце month_cnt

Фичи в сетах

Пешком до метро (км)	Класс объекта
flat_mean	date_start
flat_mean	price
flat_mean	month
flat_mean	month_cnt
flat_mean	Do TTK (км)
flat_mean	Машиномест
flat_mean	date
flat_mean	Поликлиника
flat_mean	Ставка по ипотеке
flat_mean	Детский сад
flat_mean	До кремля (км)
flat_mean	Количество объектов

TEST

shape = (1817, 47)

Предсказать value

вперед на

- 1 месяц
- 2 месяц
- 3 месяц

Метрика оценки качества решения

Root Mean Square Error (RMSE)

Дополнительные файлы для генерации фич

flat.csv [45 mb]

(исходный набор данных, где 1 строка = 1 квартира), Содержит данные по каждой квартире из корпуса

status.csv [72 mb]

(история изменения статуса по каждой квартире)

- Когда квартира введена в реализацию
- Когда продана
- Когда находится в резерве

price.csv [73 mb]

(история изменения цены по каждой квартире)

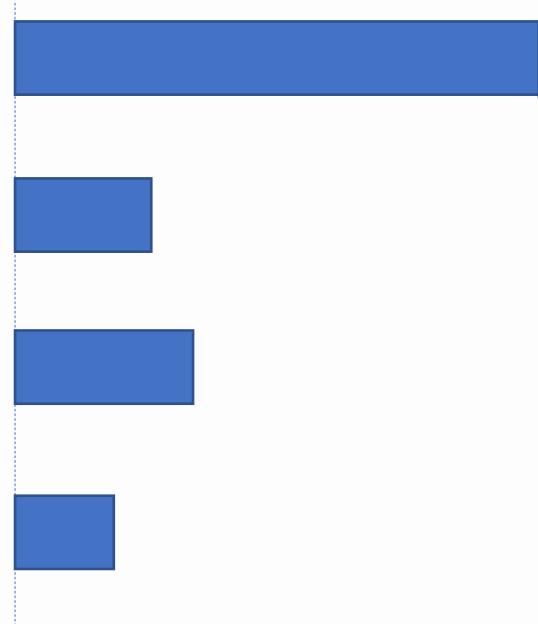
Организация работы

Что стало причиной нашего успеха?

Feature

1. Командная работа
2. Батончики Corny 
3. Coca-cola 
4. Победа Бельгии против Англии на ЧМ 

Feature_importance



Разделение ролей и слаженная работа

Работа над данными, создание фич

Получение предсказаний на различных моделях

Лучшие модель линейной регрессии

Лучшая модель без ограничений

Работа над данными

Сравнение test и train



Преобразование данных

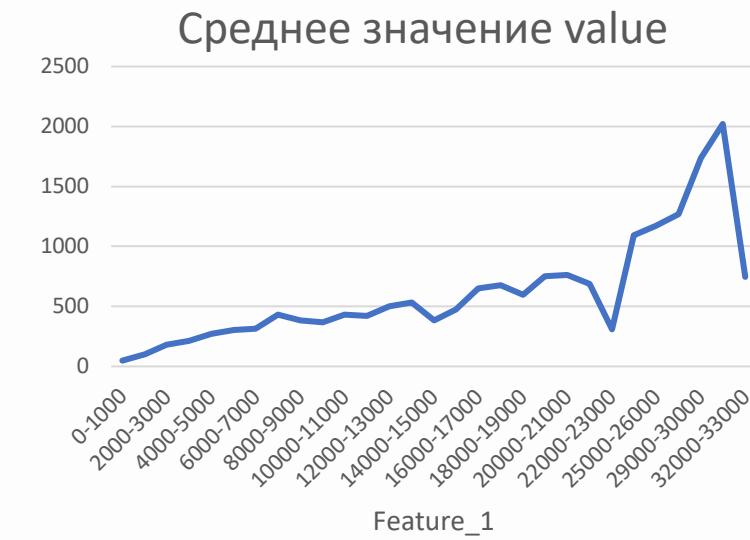
Для линеаризации фичей

- Log
- Exp
- Обратное гаусс. преобразование



Генерация фич

Придуманы 3 новые фичи из файла flat.csv



GBR + RF

(без новых фич)

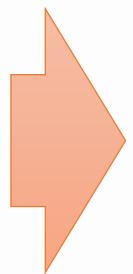
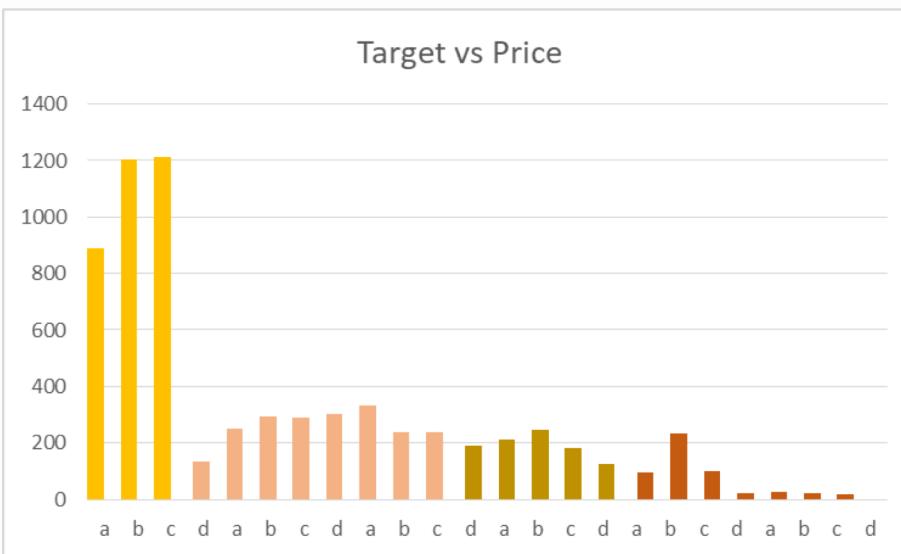
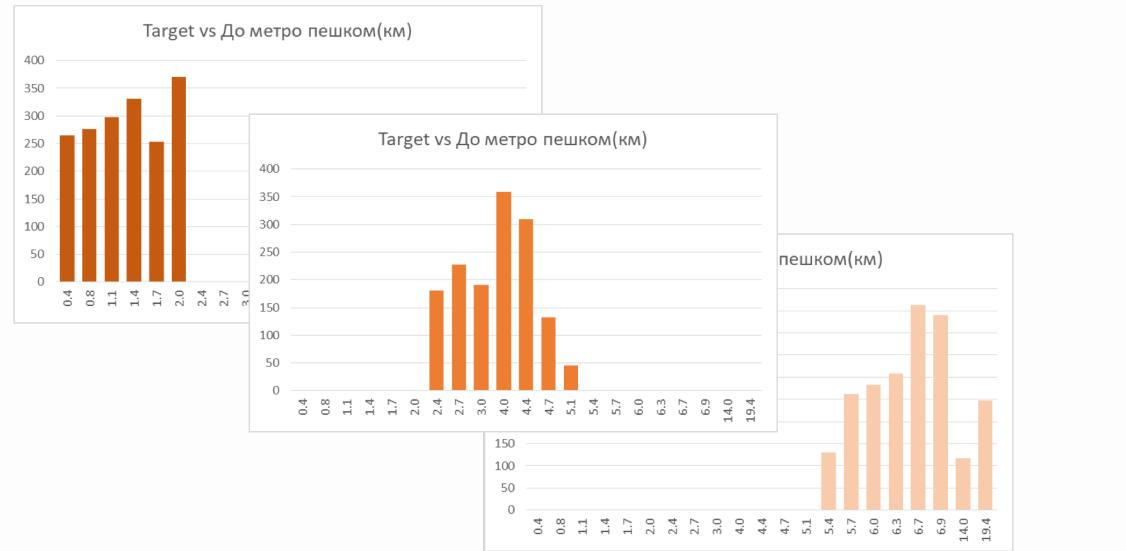
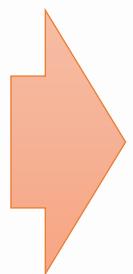
RMSE = 239

GBR + RF

(с новыми фичами)

RMSE = 184

Работа над данными

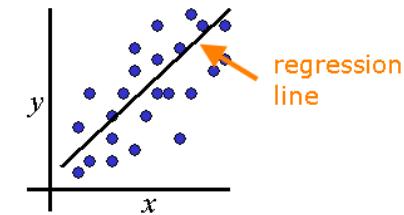


Подходы

Microsoft
LightGBM

dmlc
XGBoost

 Yandex
CatBoost

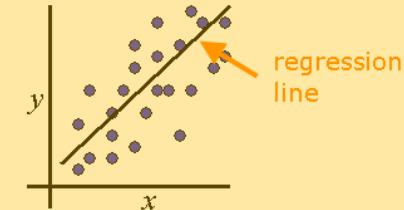


PROPHET

Подходы

Microsoft
LightGBM

dmlc
XGBoost



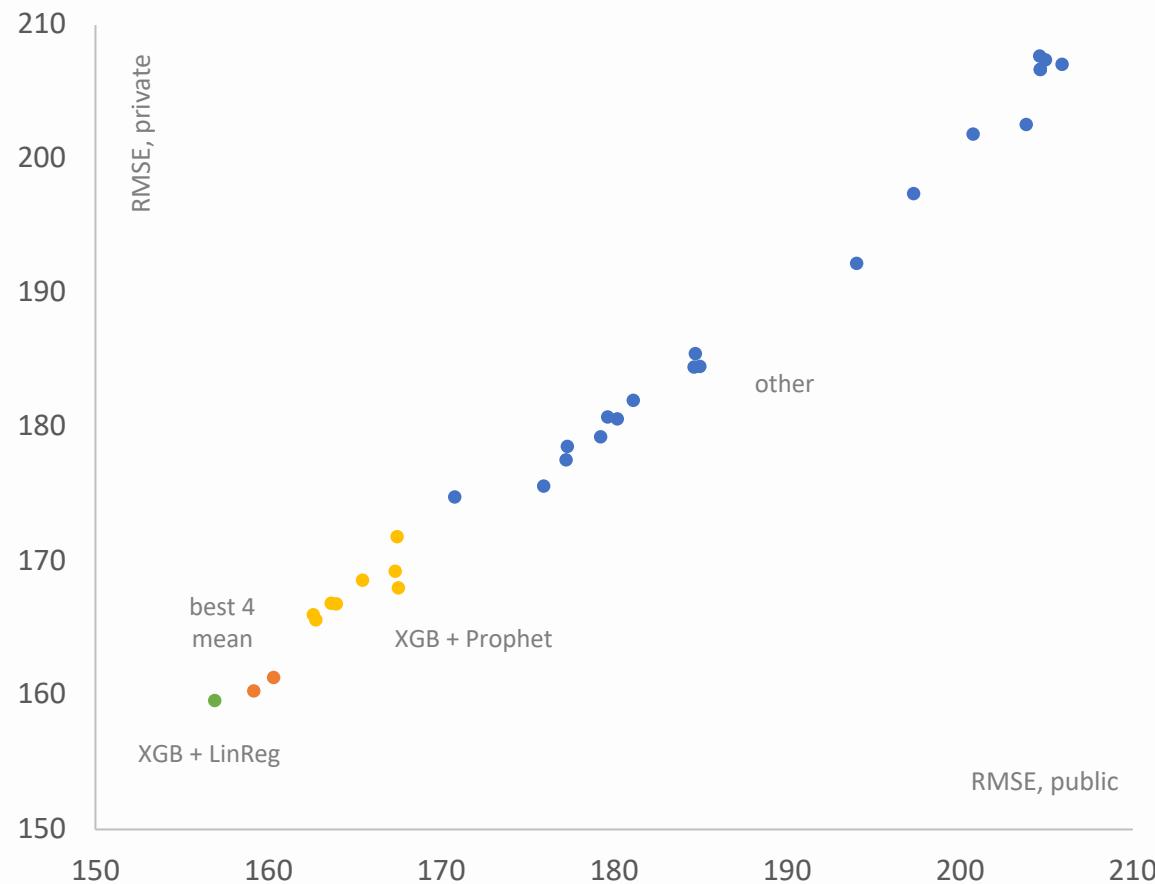
Yandex
CatBoost



Random
Forest

PROPHET

Отправка решений и работа команды



PIK Digital Day, 14.07.2018

- 1) обновление имеющихся индивидуальных моделей, проверка результатов
- 2) обмен идеями и признаками, общий брейнштурминг
- 3) финальный блендинг результатов

Финальный лидерборд

Final

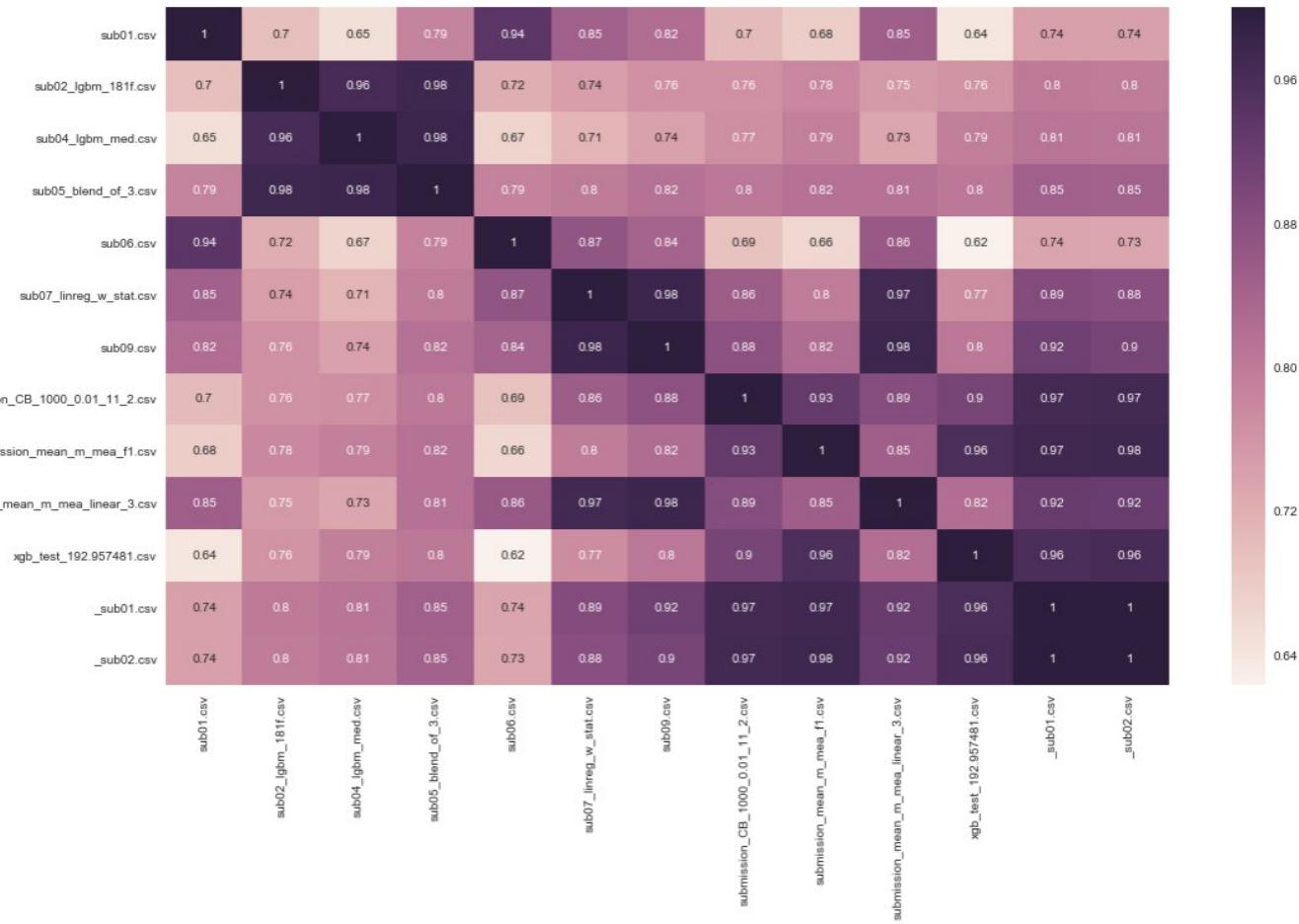
#	Участник	Счет	Попыток
1	ezavialov	158.54885	11
2	Techsensor	159.63089	47
3	Team7	159.8004	32
4	ppleskov	161.04154	37
5	Kernel Trick	163.59102	43

Final Linear

1	Techsensor	185.67501	10
2	AlexNich	190.69348	6
3	Стасямба	190.81364	5
4	YuryMarinskiy	195.39393	3
5	kvartirka s pikami	195.45945	32

Стекинг и блэндинг

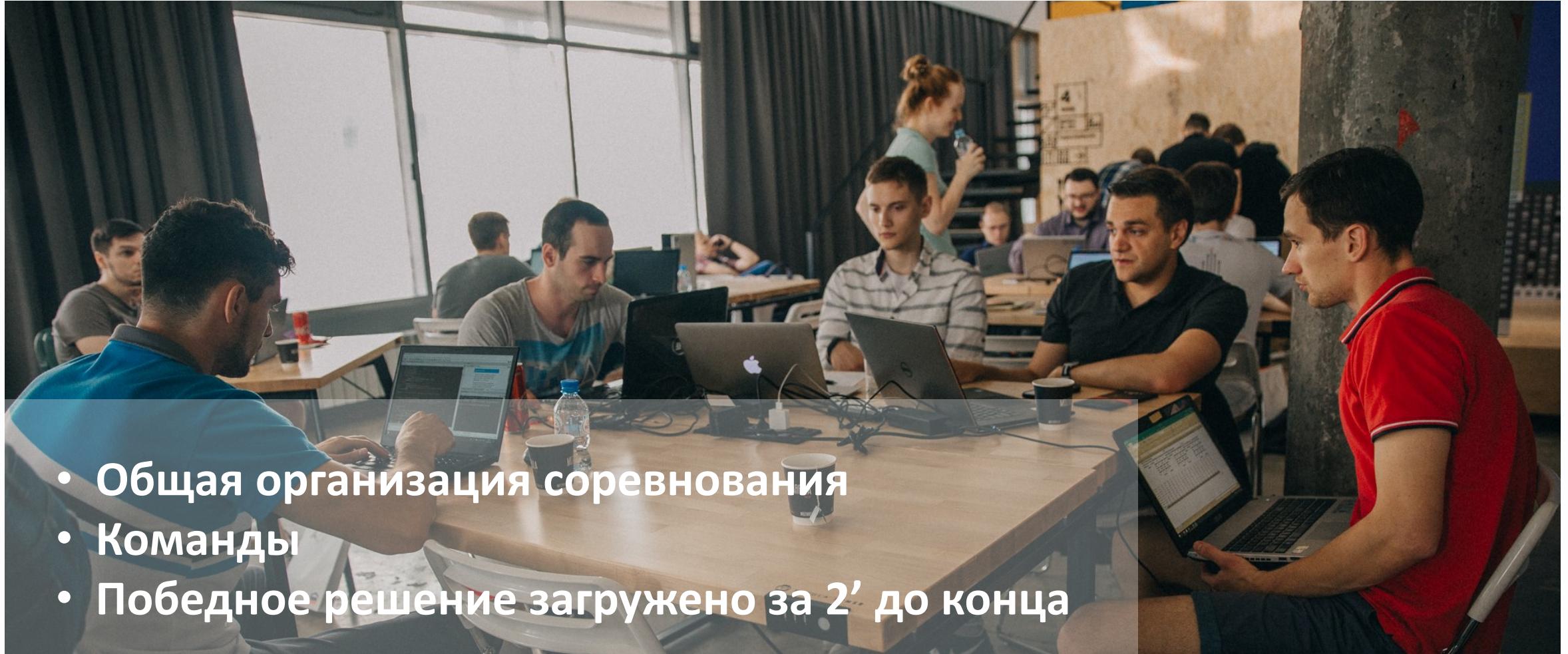
- Использован блэндинг на основе величин корреляции лучших сабмитов
- Итоговый бленд состоял из Xgboost (80%) и бленда линейных регрессии (20%)



Линейная регрессия

- Были использованы трансформации: `exp`, `log`,
обр.гаус.преобразование, `dummy`

Процесс соревнования



- Общая организация соревнования
- Команды
- Победное решение загружено за 2' до конца



Спасибо за внимание!