

Mercedes-Benz Competition: 11 место



Данила Савенков

#danila_savenkov

<https://www.kaggle.com/daniel89>

- <https://www.kaggle.com/c/mercedes-benz-greener-manufacturing/discussion/36242#202443>
- https://github.com/Danila89/kaggle_mercedes

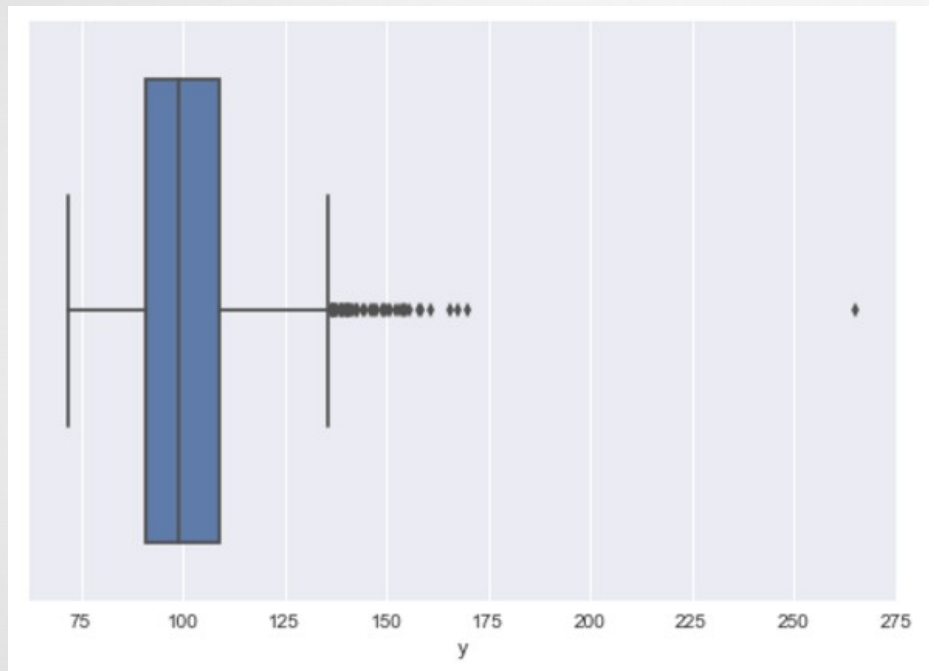
План

- Задача: данные, LB probing, LB shake-up
- Public Kernel “Stacked and Averaged”
- Бустинг и категориальные признаки
- Кросс-валидация и важность отложенной выборки
- Моя модель
- Подход обладателя второго места

Задача и данные

- Целевая переменная – время тестирования автомобиля (сек)
- Метрика R2
- Train 4029 строк
- Test 4029 строк: 81% private, 19% public
- Признаки (378 колонок):
 - Бинарные – характеристики тестирования (369 колонок)
 - Категориальные – характеристики машины (8 колонок)
 - ID – порядковый номер















Задача: public leaderboard



- Train **4029** строк
- Test **4029** строк:
81% private, **19%** public
- Std кросс-валидации по 5 фолдам **0.068**

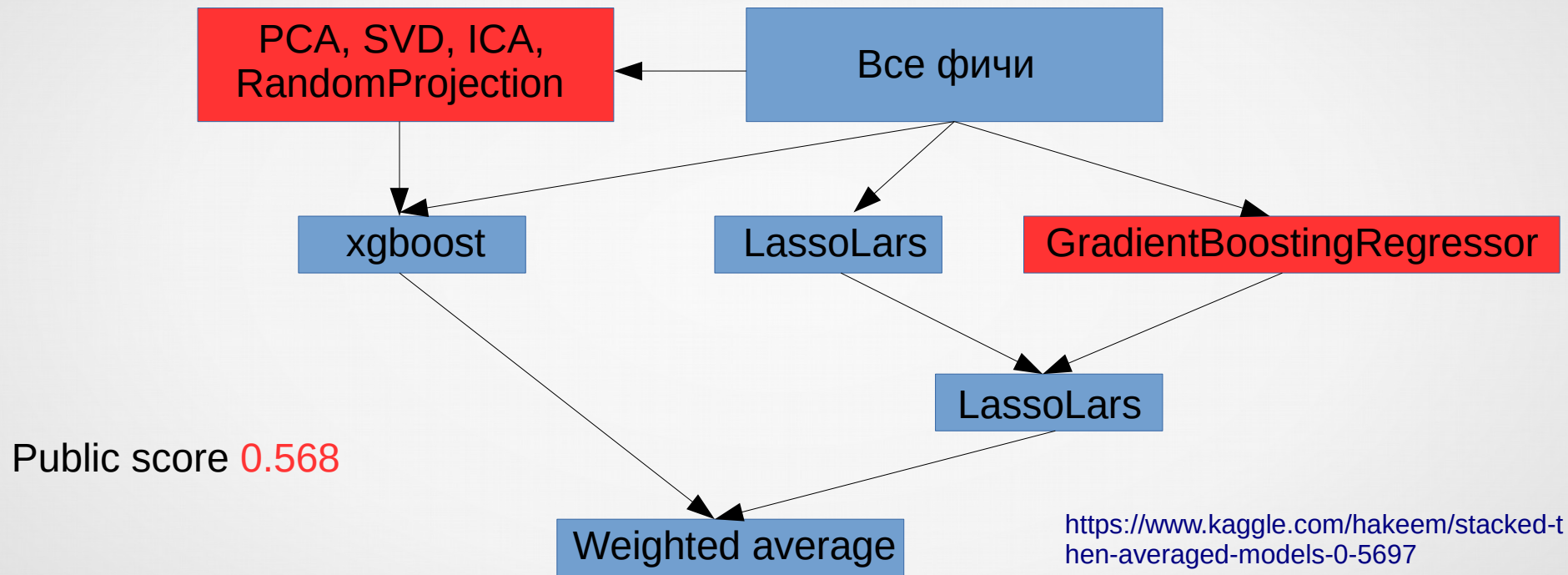
```
390 leaks = {  
391     1:71.34112,  
392     12:109.30903,  
393     23:115.21953,  
394     28:92.00675,  
395     42:87.73572,  
396     43:129.79876,  
397     45:99.55671,  
398     57:116.02167
```

Задача: public leaderboard

#	Δpriv	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	▼ 2666	Cro-Magnon			0.63045	126	15d
2	▼ 748	Zidmie			0.60409	157	15d
3	▼ 328	gavrand			0.60103	80	15d
4	▼ 2795	Alabsi: Creationline			0.59693	160	15d
5	▼ 924	doua69			0.59676	156	18d
6	▼ 935	Ivanhoe			0.59517	114	15d
7	▼ 1039	steubk			0.59508	114	15d
8	▼ 1035	boomboom			0.59488	151	15d
9	▼ 1072	Victor S D			0.59438	59	15d
10	▼ 1175	InvisiblePower			0.59437	56	15d
11	▼ 1672	Ziv Cohen			0.59434	24	15d
12	▼ 1613	julienec			0.59427	75	15d
13	▼ 16	x0x0w1			0.59423	77	15d
14	▼ 71	olegpolivin			0.59419	167	15d

Private score победителя 0.55551

Public Kernel “Stacked and Averaged”



Public Kernel “Stacked and Averaged”: PCA, SVD...

PCA, SVD, ICA,
RandomProjection

- Понижения делались по сырым данным без масштабирования
- Масштаб ID отличается от масштаба категориальных признаков
- Бинарные признаки практически не представлены в проекциях
- Удаление проекций из паблик скрипта статистически значимо повышает cv, повышает private score, но public score падает на 0.01
- Public Leaderboard Overfitting

6

▲ 1132

DDgg

<> stacked then average...



0.55425

41

1mo

Public Kernel “Stacked and Averaged”: GradientBoostingRegressor

GradientBoostingRegressor

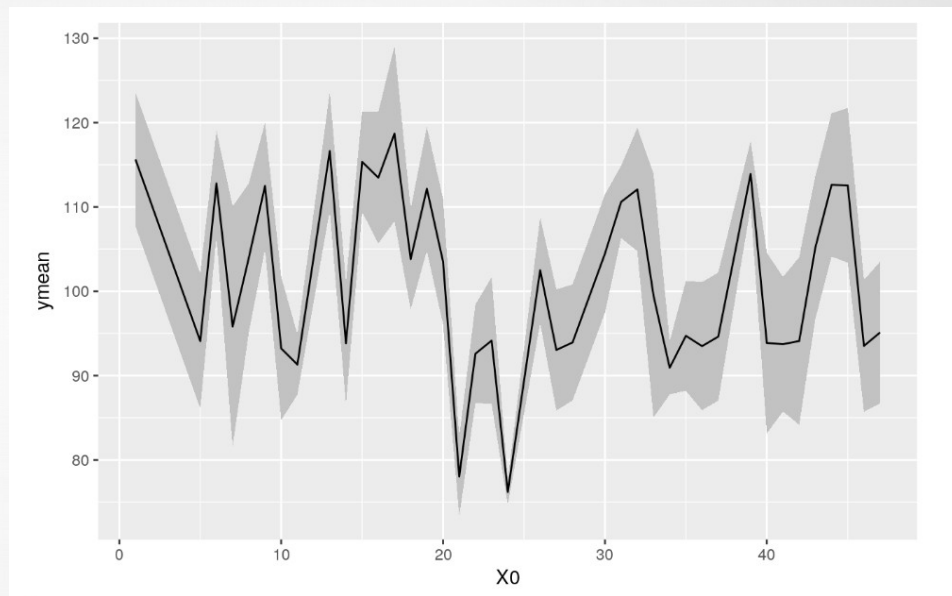
- GradientBoostingRegressor(learning_rate=0.001, loss="huber", max_depth=3, max_features=0.55, min_samples_leaf=18, min_samples_split=14, subsample=0.7)
- **Random_state** не фиксирован
- У скрипта 503 forks, DDgg – обладатель самого счастливого сида

Public Kernel “Stacked and Averaged”: лучший результат

- Лучший скрипт на основе этого керна дает в среднем 0.554 на private, что соответствует 10 месту:
<https://www.kaggle.com/adityakumarsinha/stacked-then-averaged-models-private-lb-0-554/code>
- Здесь есть ряд изменений по сравнению с оригинальной версией:
 - Удалены понижения размерности
 - Каждому объекту добавлены фичи предыдущего и последующего – развитие использования зависимости от ID

Бустинг и категориальные признаки

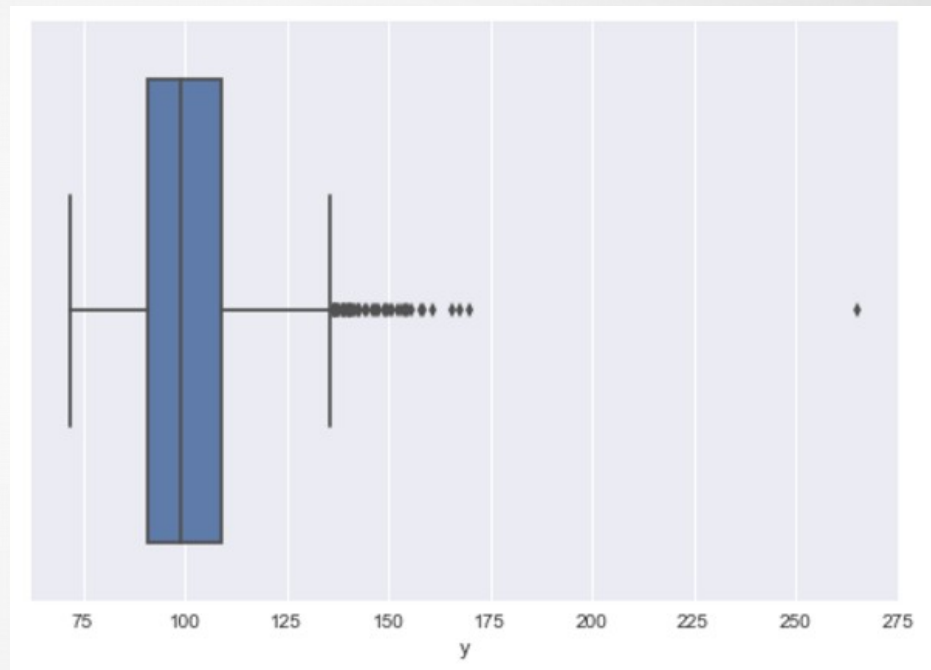
- Можно ли приблизить такую зависимость кусочно-постоянной функцией?
- LightGBM реализует деревья с условием `==` в сплитах (categorical_features)
- CatBoost предлагает несколько настраиваемых вариантов работы с категориями
- В h2o встроены разные варианты кодирования категориальных признаков



<https://www.kaggle.com/headsortails/mercedes-2-feature-interactions>

Кросс-валидация v1

- Много выбросов – std на кросс-валидации порядка **0.05**
- Убрать выбросы или стратифицировать фолды по ним не помогает



Кросс-валидация v1

- 10 разбиений на 5 фолдов – 50 скоров по каждому фолду
- Для сравнения двух алгоритмов используется t-критерий Стьюдента для связанных выборок (`scipy.stats.ttest_rel`)
- Оцениваем насколько значимы отличия качества алгоритмов

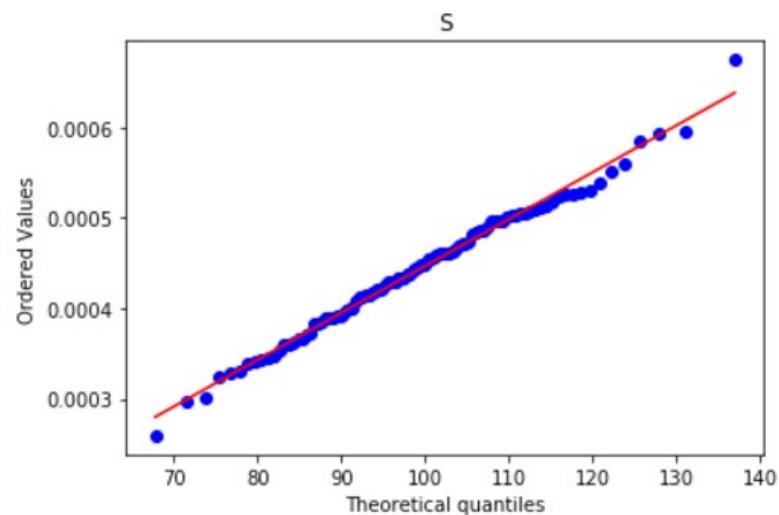
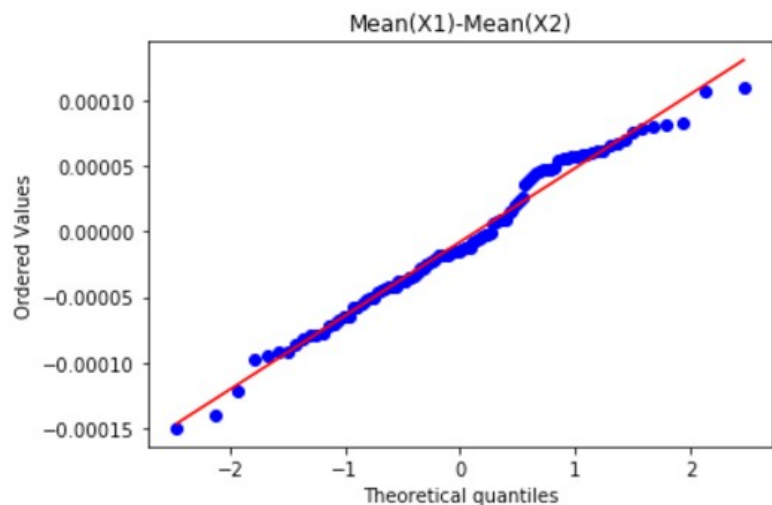
$$T(X_1^n, X_2^n) = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}}$$

X_1, X_2 – значения R^2 по соответствующим тестовым фолдам, S – дисперсия попарных разностей, n – число фолдов

Кросс-валидация v1: нормальность

$$T(X_1^n, X_2^n) = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}}$$

X_1, X_2 – значения R2 по соответствующим тестовым фолдам, S – дисперсия попарных разностей, n – число фолдов

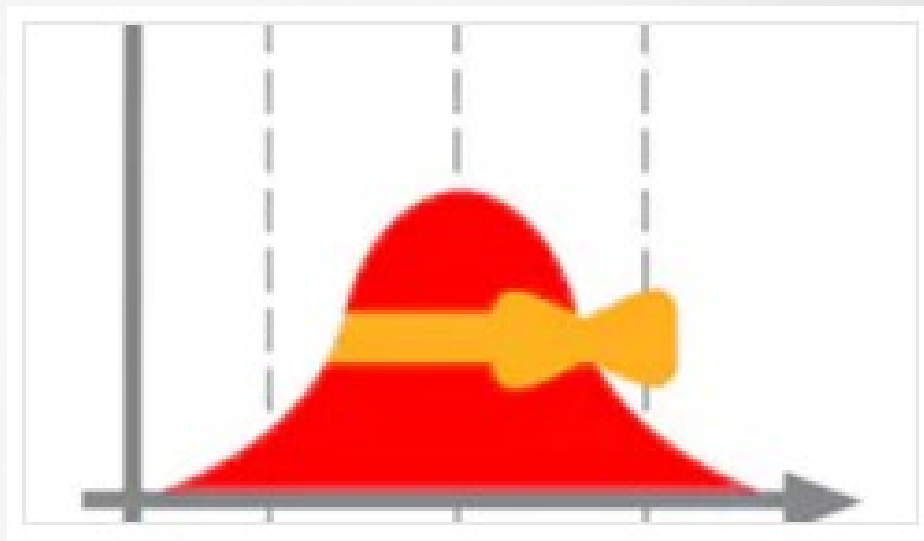


Кросс-валидация v1: множественная проверка

- Главная стратегия – проверять меньше гипотез
- При изменении гиперпараметра я следил за тенденцией изменения t -статистики. Если изменение плавное и имеет один локальный минимум, в котором $p\text{-value} < 0.05$, то я считаю, что статистически значимые отличия получены
- Если t -статистика меняется резко при малых изменениях гиперпараметра, я считаю, что отличий нет

Что все это было?

- Если для вас это ново - превосходный курс по статистике:
- <https://ru.coursera.org/learn/stats-for-data-analysis>



Кросс-валидация v1: что-то пошло не так

Важность отложенной выборки

- Валидируясь и улучшая модель, я добился локального R2 **0.58**
- Скор победителя на private **0.55551**
- Получив на public **0.54** после локального 0.58, при том, что бейзлайн на public – это 0.55, я задумался
- Public leaderboard сыграл роль отложенной выборки

Кросс-валидация v1: что пошло не так?

Важность отложенной выборки

- Произошло переобучение на выбросах в обучающей выборке
- Кросс-валидация – это не оценка итогового перформанса модели, это способ сравнения моделей
- Мы хотим чтобы наша модель хорошо предсказывала “нормальные” объекты и признаем невозможность предсказывать выбросы
- Можно не только оценивать, но и обучать модель только на “нормальных” объектах, но в этой задаче это сместит мат ожидание предсказаний

Кросс-валидация v2

- Получаем out-of-fold предсказания для всей обучающей выборки (cross_val_predict)
- Объекты с большой ошибкой предсказания считаем выбросами
- При кросс-валидации алгоритм **обучается на всех объектах**, но **качество оценивается только на “нормальных” объектах** (не выбросах)
- Функция cross_validation_score_statement отсюда:
https://github.com/Danila89/cross_validation_custom

Кросс-валидация v2

- Мой отбор признаков оказался переобучением
- Количество деревьев в бустинге оказалось слишком велико

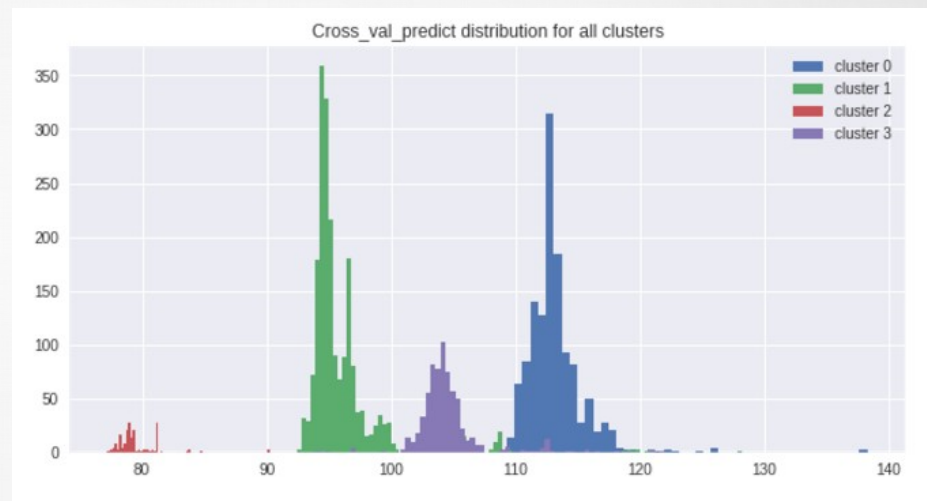
Мое решение: кластеризация

- По категориальной фиче X0 легко было выделить четыре типа машин:
<https://www.kaggle.com/daniel89/mercedes-cars-clustering>



Мое решение: кластеризация

- Однако бустинг очень легко и без нашей помощи разделяет эти кластеры
- Проблема в том, что внутри каждого кластера `xgb` показывает очень низкий, а в некоторых случаях **отрицательный R2**



Мое решение: кластеризация

- В красном кластере `xgb` дает стабильно отрицательный R^2 – всем объектам этого кластера предсказываем средний таргет по кластеру
- Каждому из оставшихся кластеров – свой `xgb` со своими параметрами и своими фичами
- Тем не менее в `fit` передаются все объекты



Мое решение: признаки

- Использовал разные техники отбора фич, пытался строить понижения размерности, случайные проекции. В итоге переобучился
- В итоговой модели участвуют все бинарные фичи, ID и номер кластера (фича, производная от X0 и принимающая 4 значения)

Мое решение: параметры xgb

- Использовались деревья глубины 2
- Количество деревьев в пределах 100
- Активно использовал subsampling признаков (colsample_bytree, colsample_bylevel)
- Параметр gamma очень помогал не ветвиться слишком много

Мое решение: резюме

- По X_0 разделили объекты на 4 кластера
- По одному из кластеров ничего предсказать не получилось – предсказали кластерное среднее
- Для каждого из оставшихся настроили отдельный `xgb`, который тем не менее обучался по всем объектам

Подход владельца второго места

- Обрезать выбросы на $y=155$
- Feature Learner: фичи выбираются случайно, причем вероятность выбора пропорциональна скору предыдущей итерации с этой фичей
- <https://www.kaggle.com/c/mercedes-benz-greener-manufacturing/discussion/36390>



Вопросы из слака



stasg7 Jul 15th at 11:09 PM
in #kaggle_crackers

Немного вброшу... А зачем народ вообще участвует в млбуткампах (2х последних), мерседесах и т.д. если сразу по постановке задачи, данным, метрике понятно что будет дикий рандом в финальных результатах?



stasg7 13 days ago

@bearstrikesback Я не говорил что задача дно 😞 Хотелось услышать людей которые решали - зачем и почему 😊 Мне показалось что нельзя было извлечь сильно больше информации из данных признаков по сравнению с бейзлайнами



- После Сбера хотелось еще медальку :)
- Действительно, разница с бейзлайнами ~ 0.015 , но для себя я нашел много нового, решая эту задачу
 - Кросс-валидация с учетом статистической значимости
 - Кросс-валидация без выбросов
 - Необходимость отложенной выборки
 - Опасность переобучения на выбросах
 - Разделение объектов на кластеры и дифференцированный подход при построении моделей
 - Важность параметра γ в xgb
 - Feature Learner от обладателя второго места
 - ...

Спасибо за внимание

- Вопросы
- Комментарии
- Критика
- ...

