Работа не волк. Работа - это work

Ян Будакян
ML-engineer @
Cinimex DataLab
Ph.D. student @ MSU

Дмитрий Борисов
ML-engineer @
Looking for a job
MIPT, Grenoble INP

Александр Сидоренко
NLP-engineer @ Sber

https://github.com/astromid/pandemicdatahack-track3

Предобработка данных

1. В категориальных полях пропуски заполнили значением “NA_category”
2. Текст в категориальных перевели в нижний регистр
3. Выкинули некоторые поля (locality, ...)
4. schedule, driver_licence разделили по уникальным значениям, ohe
5. В числовых полях выбросы заполнили константами
6. salary и desired_salary логарифмировали
7. Добавили внешние данные (об этом позже)
8. В employments.csv для текстовых полей брался усредненный вектор fasttext (модели от deeppavov) (100-dim) по словам в предложении. Текст предобрабатывался с помощью удаления html-тэгов и лемматизация с pymorphy

Внешние данные

1. Использовали дополнительные данные с 2015 по 2020 год
2. ВВП России по годам в рублях и долларах, уровень безработицы, инфляции
3. Ежедневные котировки евро, доллара, нефти, газа, золота
4. Количество больных COVID-19 по дням и дням-регионам

Настройка валидации

- Настроили 5-Fold валидацию со стратификацией по году публикации резюме, потому что в train и test у них схожие распределения
- Локальная валидация коррелировала с public leaderboard’ом
- Итоговое предсказание: усреднение предсказаний по фолдам

```
[92]: train['publish_date'].apply(lambda x: x[:4]).value_counts(True)
```

```
[92]: 2020    0.800258
      2019    0.120204
      2018    0.049293
      2017    0.023035
      2016    0.006099
      2015    0.001110
      Name: publish_date, dtype: float64
```

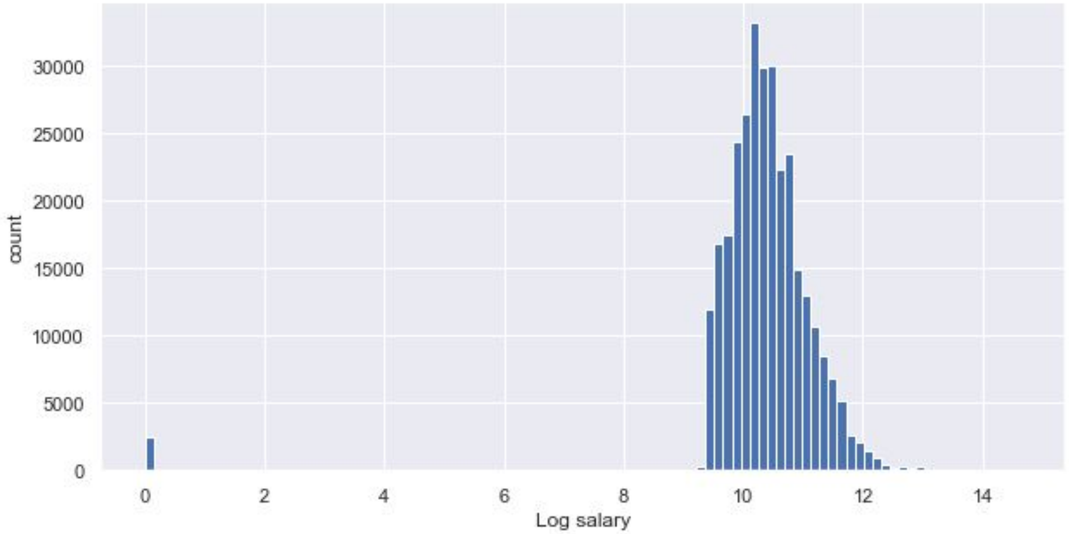
```
[93]: test['publish_date'].apply(lambda x: x[:4]).value_counts(True)
```

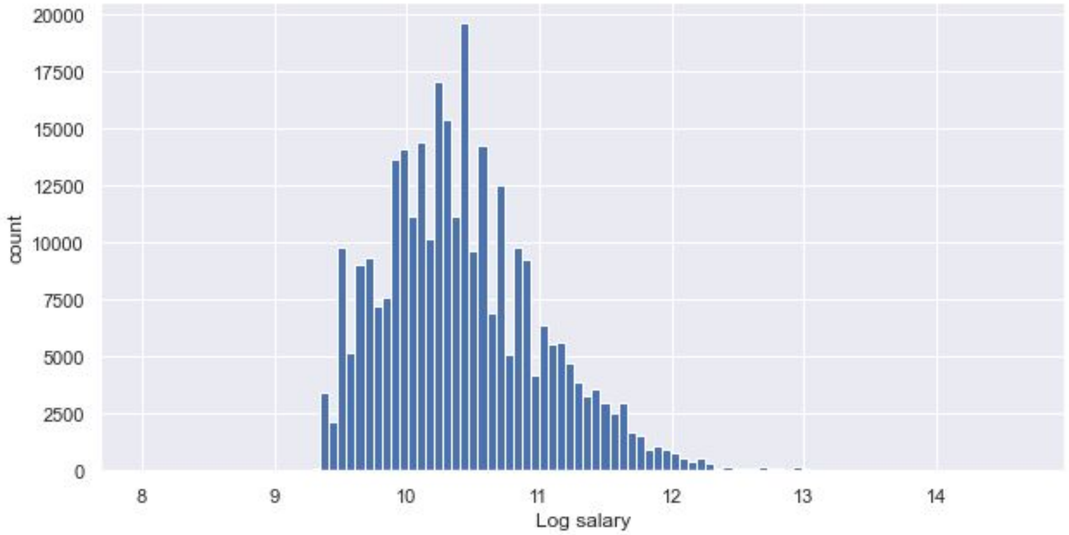
```
[93]: 2020    0.798787
      2019    0.120472
      2018    0.049909
      2017    0.023084
      2016    0.006506
      2015    0.001242
      Name: publish_date, dtype: float64
```

Выбор моделей

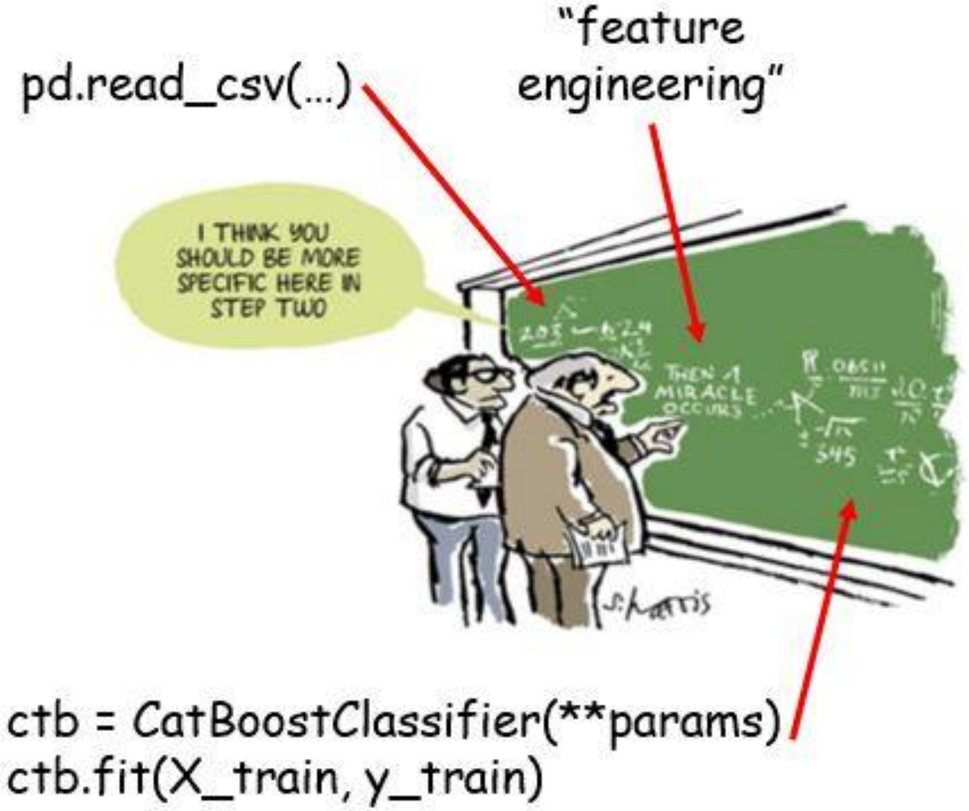
- CatBoost (GPU)
- LinearRegression
- RandomForest
- XGBoost
- LightGBM

Выбор таргета



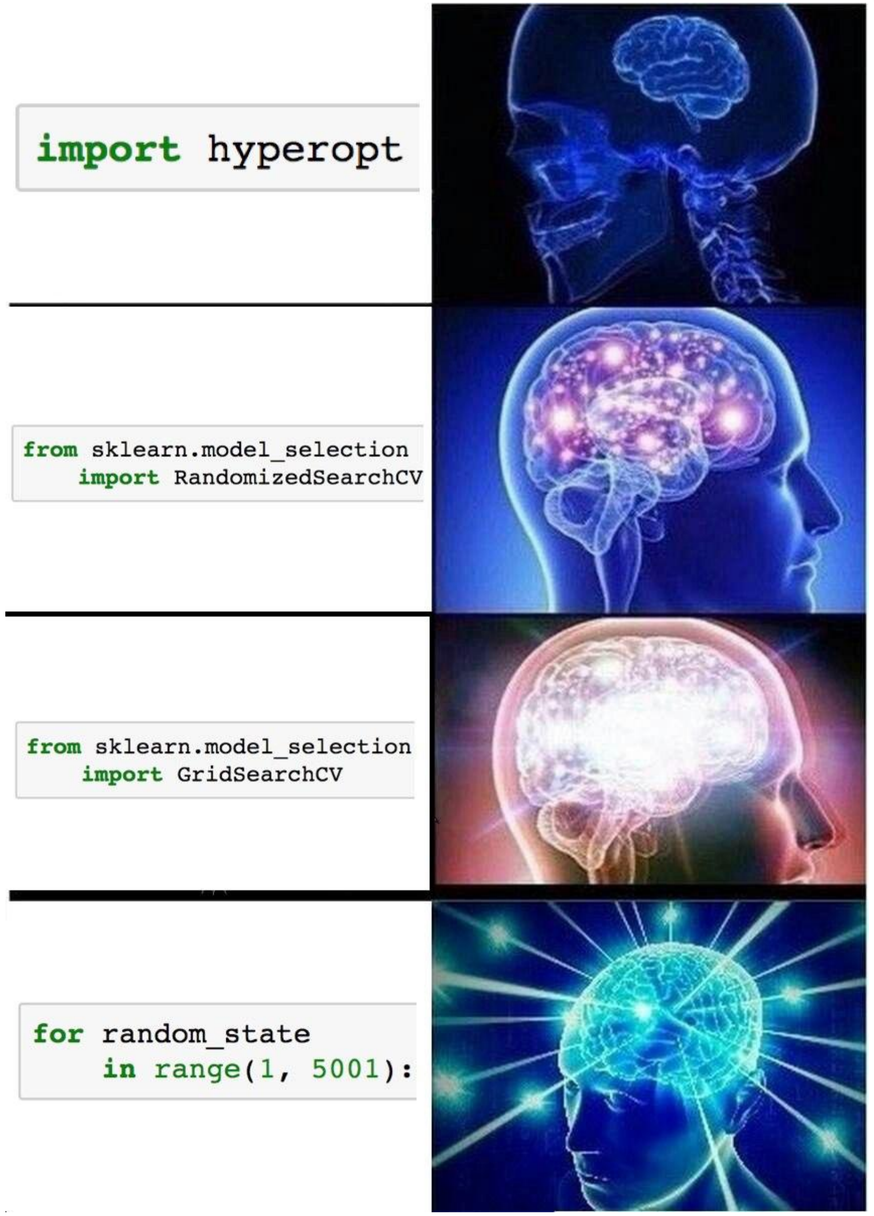


- RMSLE -> log(target) + MSE
- Classification + Regression
- result = regression_score * (1 - outlier_probability)



Optuna: подбор гиперпараметров

```
params = {
    "n_estimators": 1000,
    "verbosity": -1,
    "lambda_l1": trial.suggest_float("lambda_l1", 1e-8, 10.0, log=True),
    "lambda_l2": trial.suggest_float("lambda_l2", 1e-8, 10.0, log=True),
    "num_leaves": trial.suggest_int("num_leaves", 2, 256),
    "feature_fraction": trial.suggest_float("feature_fraction", 0.4, 1.0),
    "bagging_fraction": trial.suggest_float("bagging_fraction", 0.4, 1.0),
    "bagging_freq": trial.suggest_int("bagging_freq", 1, 7),
    "min_child_samples": trial.suggest_int("min_child_samples", 5, 100),
}
```



Ансамблирование моделей

1. Усреднение предсказаний по фолдам
2. Усреднение предсказаний различных моделей

Эволюция скора

День 1

- CatBoost: 0.21 CV -> 1.02103 LB
- RandomForest 0.19584 CV -> 1.01514 LB
- RandomForest 1.0262 CV -> 1.00888 LB
- Linear Regression 1.0236 CV -> 1.01838 LB

День 2

- CatBoost: 1.023 CV -> 1.01296 LB
- CatBoost + classification: 1.0187 CV
- CatBoost + classification (salary < 300): 1.0120 CV
- -||- + COVID + Currency: 1.0099 CV
- -||- + COVID + Currency + external: 1.01002 CV
- RF + classification + COVID + Currency: 1.00958 CV
- XGB + classification + COVID + Currency: 1.01299 CV
- LinReg + classification + COVID + Currency: 1.02516 CV
- CatBoost + classification + COVID + Currency + Text embeddings: **1.0036 CV**

Результаты:

- Усреднение (геометр.) всех результатов <1 LB 0.99017 Privat
- RandomForest + COVID + Currency 0.99181 Privat

Спасибо за внимание!

Ян Будакян
ML-engineer @
Cinimex DataLab
Ph.D. student @ MSU

Дмитрий Борисов
ML-engineer @
Looking for a job
MIPT, Grenoble INP

Александр Сидоренко
NLP-engineer @ Sber

https://github.com/astromid/pandemicdatahack-track3