ГАЗПРОМ НЕФТЬ | НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

Иннопрактика

Gazprom Neft Smartoil Contest

Александр Дроботов

Выпускник РЭУ им. Г.В. Плеханова, экономист-математик

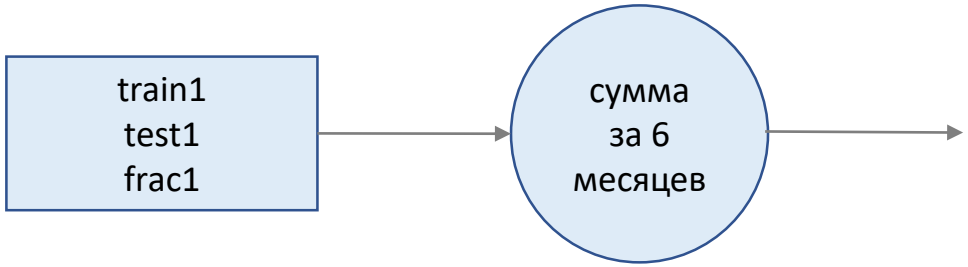
BBDO Group, эконометрист

email: sashadrbtv@gmail.com

FB: fb.com/alexander.drobotov

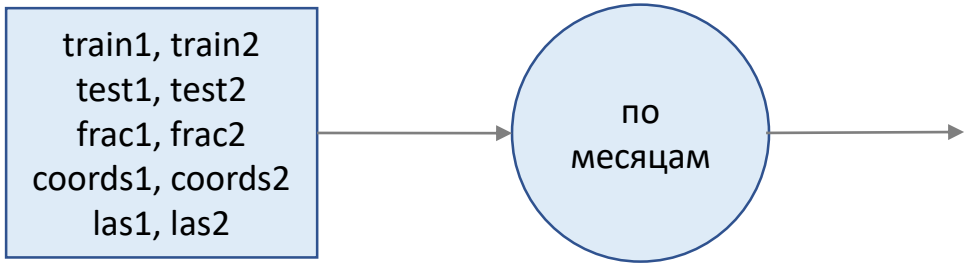
Результаты

базовое решение: линейная регрессия



Public MAE
≈360

итоговая модель: XGBoost



Public MAE
≈320

OLS

(никнейм)

МЕСТО

	Public	Privat
Задача 1	9	9
Задача 2	50	41

MAE

	Public	Privat
Задача 1	316.371	394.097
Задача 2	318.263	210.695

Ключевые моменты

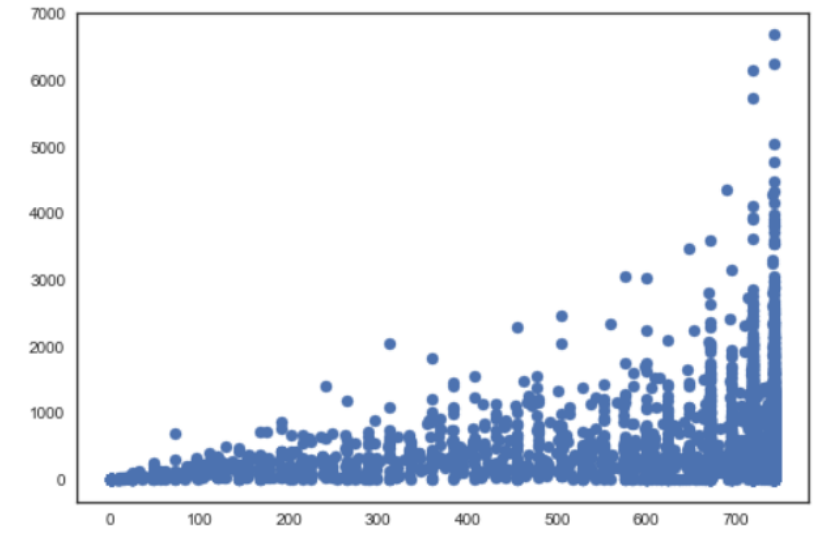
- Борьба с выбросами в данных
- Использование данных из 1-ой и 2-ой задачи
- Сильные признаки: las-признаки, линейный тренд, кол-во дней с последних показаний, время работы
- Построение моделей на разных признаках
- Понижение размерности: PCA, FA, SVD; кластеризация
- 5-Fold кросс-валидация


```
def is_outlier(points, thresh=3.5):
    if len(points.shape) == 1:
        points = points[:,None]
    median = np.median(points, axis=0)
    diff = np.sum((points - median)**2, axis=-1)
    diff = np.sqrt(diff)
    med_abs_deviation = np.median(diff)

    modified_z_score = 0.6745 * diff / med_abs_deviation

    return modified_z_score > thresh
```

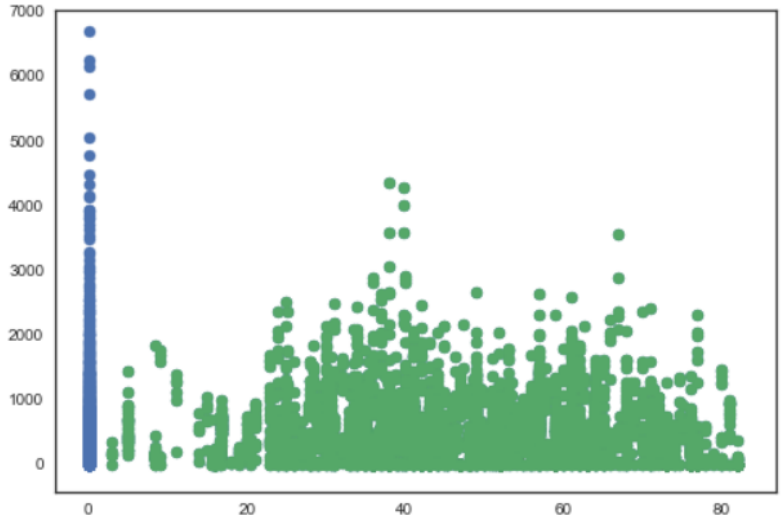
```
plt.scatter(trainset['vrema_raboty_c'], trainset['target']);
```

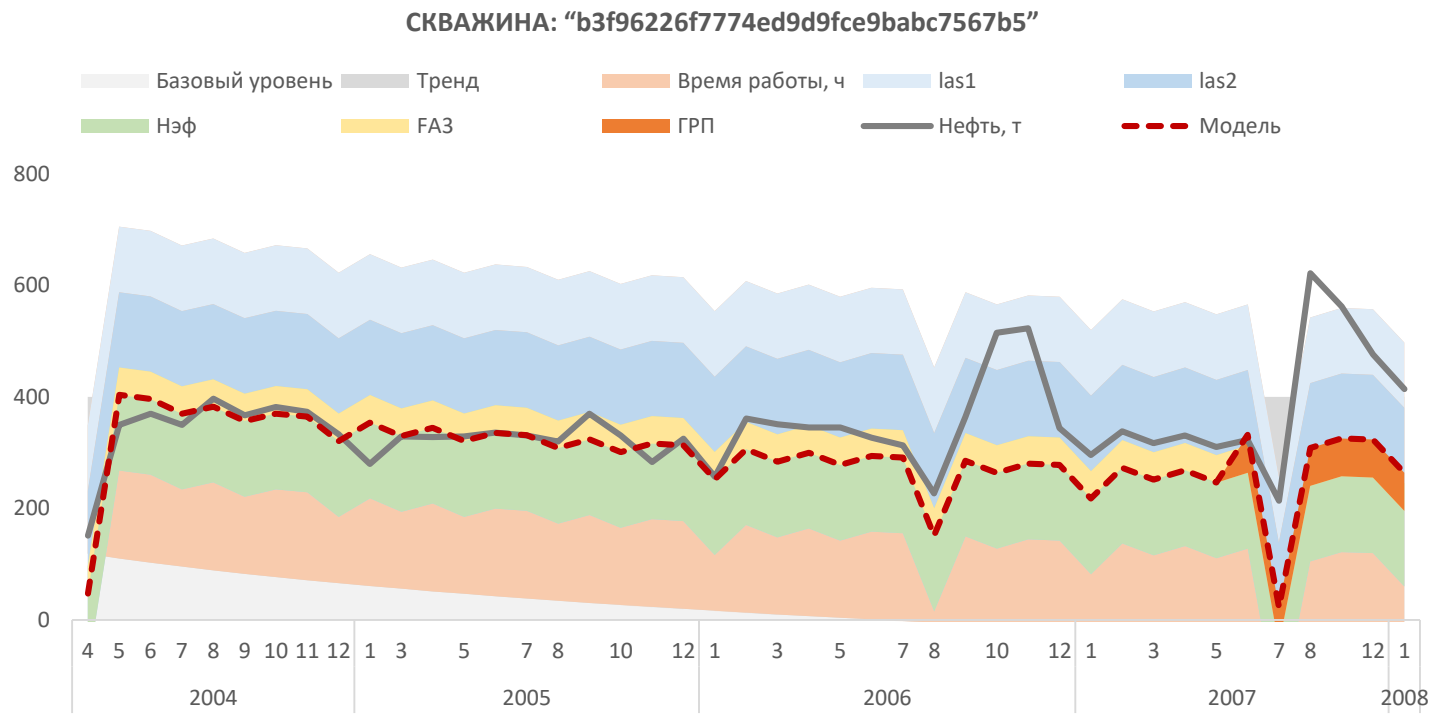


Что ещё хотелось бы сделать

- Smart-подход к заполнению пропусков
- Дополнительный анализ специфики бизнеса и физических взаимосвязей признаков
- Проработка линейной регрессии с полным пулом данных, возможность интерпретации

```
plt.scatter(trainset['eff'].fillna(0), trainset['target']);
plt.scatter(trainset['eff'], trainset['target']);
```





ГАЗПРОМ НЕФТЬ | НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

Иннопрактика

СПАСИБО!

Буду рад ответить на ваши вопросы

email: sashadrbtv@gmail.com
FB: fb.com/alexander.drobotov