Решение кейса Oil and Gas Industry

Булыгин Максим
Студент 4-го курса СПбГЭУ

30 Мая, 2022

Solution pipeline

01

Missing values, EDA

- Работа с пропусками
- EDA

02

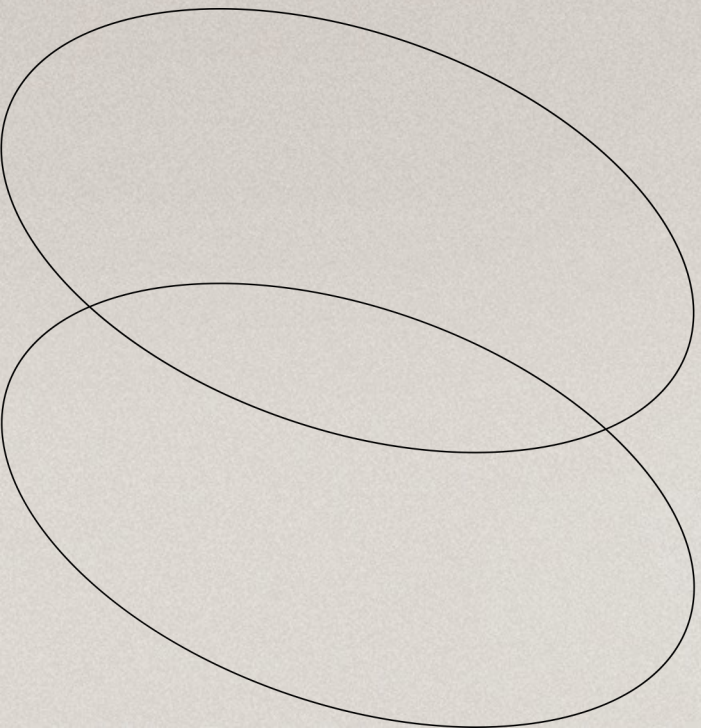
Feature engineering

- Какие фичи создавать?
- Как избежать проблем лика данных во временных рядах?

03

Training and validation

- Использованные модели
- Как валидировать?



04

Test evaluation

- Инсайты
- Факапы

05

Retrospective thougths

- Что можно было попробовать еще?
- Где были допущены ошибки?
- Идея идеального решения

EDA, missing values

Missing values

1. Фичи с большим количеством пропусков

2. Фичи с относительно небольшим количеством пропусков

Объем жидкости	63671
Объем нефти	63671
Давление буферное	56928
Газовый фактор рабочий (ТМ)	41906
Дебит газа (ТМ)	34005
Давление забойное от Рпр	24839
Дебит газа попутного	15767
Давление на входе ЭЦН (ТМ)	9056
Дебит жидкости (ТМ)	7929
Активная мощность (ТМ)	4724
Время работы (ТМ)	3661
Коэффициент мощности (ТМ)	2986
Давление забойное	2380
Давление забойное от Нд	1426
Давление линейное (ТМ)	340

Missing values

1. Фичи с большим количеством пропусков

```
['Объем жидкости',
 'Объем нефти',
 'Газовый фактор рабочий (ТМ)',
 'Давление буферное',
 'Давление забойное от Рпр',
 'Дебит газа (ТМ)']
```

Заполняем по каждой фиче по каждой скважине средним по либо нулем.

Проблема: временной лик

Missing values

2. Фичи с относительно небольшим количеством пропусков

Интерполируем пропущенные значения по каждой фиче по каждой скважине линейно.

Feature engineering

Feature engineering

Временные фичи:

1. Год
2. Месяц года
3. Неделя года
4. День года, месяца, недели, выходные
5. Сезон года

Невременные фичи:

1. Фичи исходных данных
2. Лаги n-го порядка целевой переменной

Циклические временные фичи:

1. Sin, Cos сезона года
2. Sin, Cos месяца года
3. Sin, Cos недели года, месяца
4. Sin, Cos дня недели, месяца

Feature engineering

Как создавать фичи для тестового набора данных?

1. Временные и временные циклические фичи – понятно
2. Фичи исходного набора данных: серьезное предположение о постоянности среднего фичей на тесте
3. Лаги – нужно очень аккуратно быть с ликом во времени

	datetime	Номер скважины	Дебит нефти	скважина_шифт_1	скважина_шифт_2
0	1990-08-12	1	24.5800	24.854	25.016
1	1990-08-13	1	25.2900	24.580	24.854
2	1990-08-14	1	24.9350	25.290	24.580
3	1990-08-15	1	23.8610	24.935	25.290
4	1990-08-16	1	32.2130	23.861	24.935
5	1990-08-17	1	29.8260	32.213	23.861
6	1990-08-18	1	23.8610	29.826	32.213

Training and validation

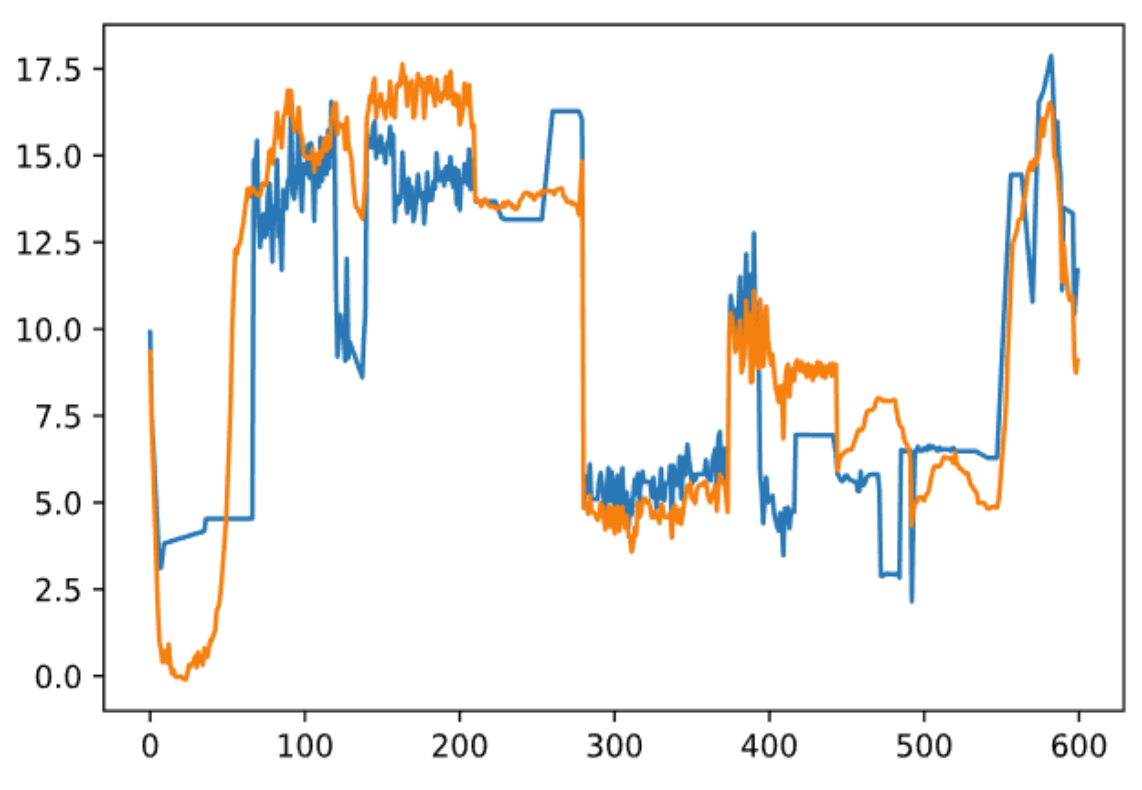
Training and validation

- Бьем на трейн и валидацию по времени (12% валидация)
- Используемые модели – бустинги в реализации CatBoost, XGBoost.
- Пробовал но не зашло: пакет pycaret.
- Пробовал feature selection по SHAP и LASSO, импрува не дало.
- Пробовал обучать 106 бустингов по отдельности и 1 бустинг на всем датасете, особой разницы не дало.
- Валидация для каждого наблюдения:

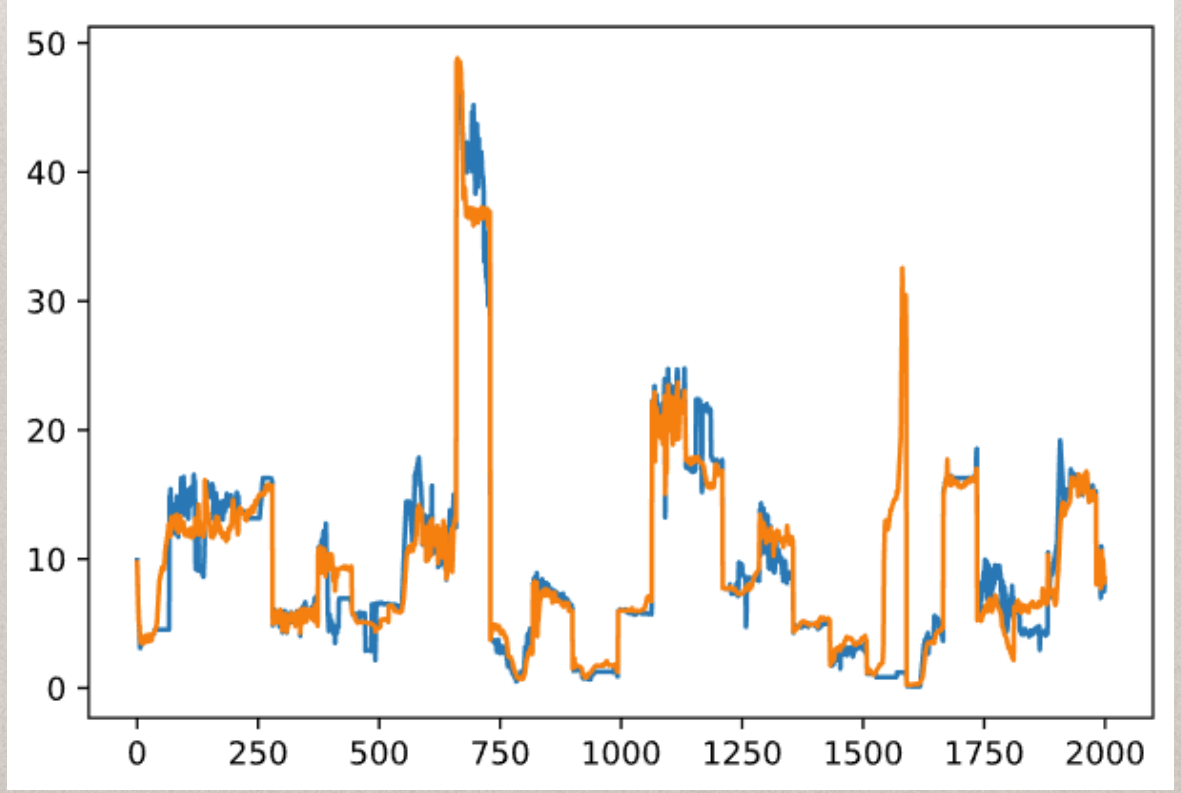
$$\hat{y}_{t,hole_j} = f(\hat{y}_{t-1,hole_j}, \cdots, \hat{y}_{t-n\_lags+,hole_j})$$

Training and validation

CatBoost с flatten’ым дебитом нефти



XGBoost с flatten’ым дебитом нефти



Test evaluation

Test evaluation

- На большом количестве сабмитов был overfit, несмотря на хорошую валидацию
- Лучший скор на катбусте с лагами
- Часть тестовых данных (177 наблюдений) была в трейне, они использовались для эвалюйта на тесте в качестве лагов и в финальном сабмите

	datetime	Номер скважины	Дебит нефти
3184	1992-04-12	4	5.624
3185	1992-04-13	4	6.142
3186	1992-04-14	4	6.331
3187	1992-04-15	4	6.545
3188	1992-04-16	4	5.279
...	...	...	...
66511	1992-04-13	104	2.658
66512	1992-04-14	104	2.380
66513	1992-04-15	104	2.274
66514	1992-04-16	104	2.277
66515	1992-04-17	104	2.244

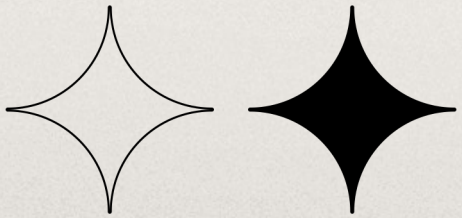
177 rows × 40 columns

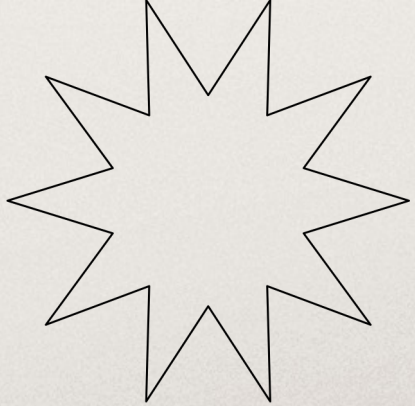
Retrospective thoughts

Retrospective thoughts

- Следовало больше времени уделить на EDA
- Хотелось попробовать прогнозировать фичи исходных данных как временные ряды и прогнозы использовать в качестве фичей для эвалюйта на тесте, но не успел
- Стоило учесть разную тенденцию у временных рядов как фичей, так и дебита нефти
- Сплит на трейн и валидацию стоило попробовать побить не только по времени, но еще и по скважинам

Retrospective thoughts





- Стоило изучить поподробнее API пакета fedot из бейзлайна и использовать пакет для сабмитов
- Стоило быть аккуратнее с сабмитами – как тщательнее валидироваться, так и аккратно смотреть на файл сабмита
- Попробовать статистический подход к прогнозированию: рассмотреть стационарные ряды и прогноз
- В формировании фичей из исходного набора данных был лик

Retrospective thoughts

Идеальный пайплайн решения:

- По EDA найти больше инсайтов, учесть различные тенденции
- Глубже исследовать физический смысл задачи и признаков
- Формировать фичи для теста как прогнозы временных рядов
- Избегать лика данных во временных рядах и признаках (!!!)
- Сплит на трейн и валидацию по времени и скважинам
- Рассмотреть большой набор регрессионных моделей: бустинги, статистические модели, fedot
- Очень тщательная валидация
- Больше внимания feature selection

Спасибо за внимание!