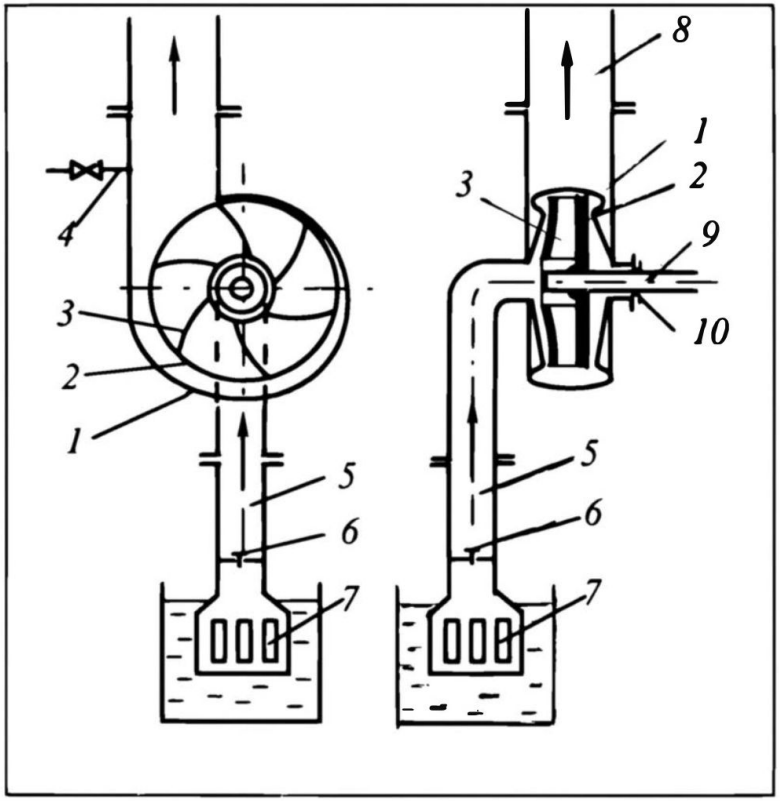
Прогнозирование градиента давления на приеме центробежного насоса для нефтегазовой промышленности

Студент: Ф. Гарбар
Научный руководитель: С. Абдуракипов

HSE, 2023

Введение

- Жидкость, поступает в колесо (2);
- При быстром вращении колеса жидкость между лопатками (3) быстро отбрасывается под действием центробежной силы: *передача механической энергии лопаток жидкости*;
- Согласно уравнению Бернулли сумма параметров системы постоянная, поэтому рост скорости жидкости влечет снижение давления в системе.



$$z + \frac{p}{\rho g} + \frac{w^2}{2g} = const$$

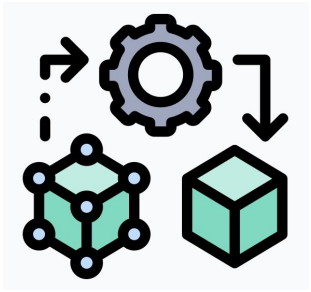
- *Ожидаемо, что для исправной работы подобной системы необходимы современные способы прогнозирования и анализа данных.*

Источник: Айнштейн, Захаров, Носов: Процессы и аппараты химической технологии. Общий курс. В 2-х книгах. Книга 1 ; Издательство Лань, 2022 г.

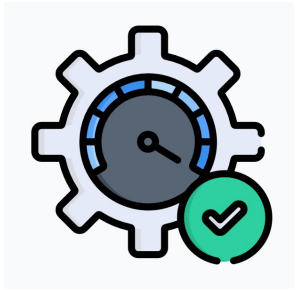
Методология

- Для подбора оптимальных параметров были использованы:

1. `LightGBM` - градиентный спуск на основе дерева решений (метрика L2);
2. `RandomForestRegressor` - дерево решений, основанное на разбиении на случайные подвыборки с дальнейшим усреднением (метрика AUC);
3. `Ridge Regression` - линейная регрессия с L2 регуляризацией.

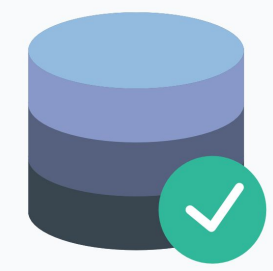


- Библиотека `Hyperopt` для поиска гиперпараметров модели град. бустинга `LightGBM` и `RFRegressor`.



- Вместо прямого поиска перебором по сетке (`GridSearch`) можно использовать поиск на основе модели и перебирать нужные области пространства параметров (сетки) и быстрее сходиться к минимуму

Датасет и его обработка 1/3



- Период наблюдений: **Июль 2019**
- Разбиение по номеру скважины: **WELL_ID**
- Кол-во скважин: **17**
- Моделирование осуществляется на 3-х time series
- Целевая переменная: **DSHORTT1138P2300058***
- Частота наблюдений 5 минут
- Данные нормализованы, используя min-max scaler:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Наибольшая связь (по абсолютной величине) у целевой переменной **DSHORTT1138P2300058**

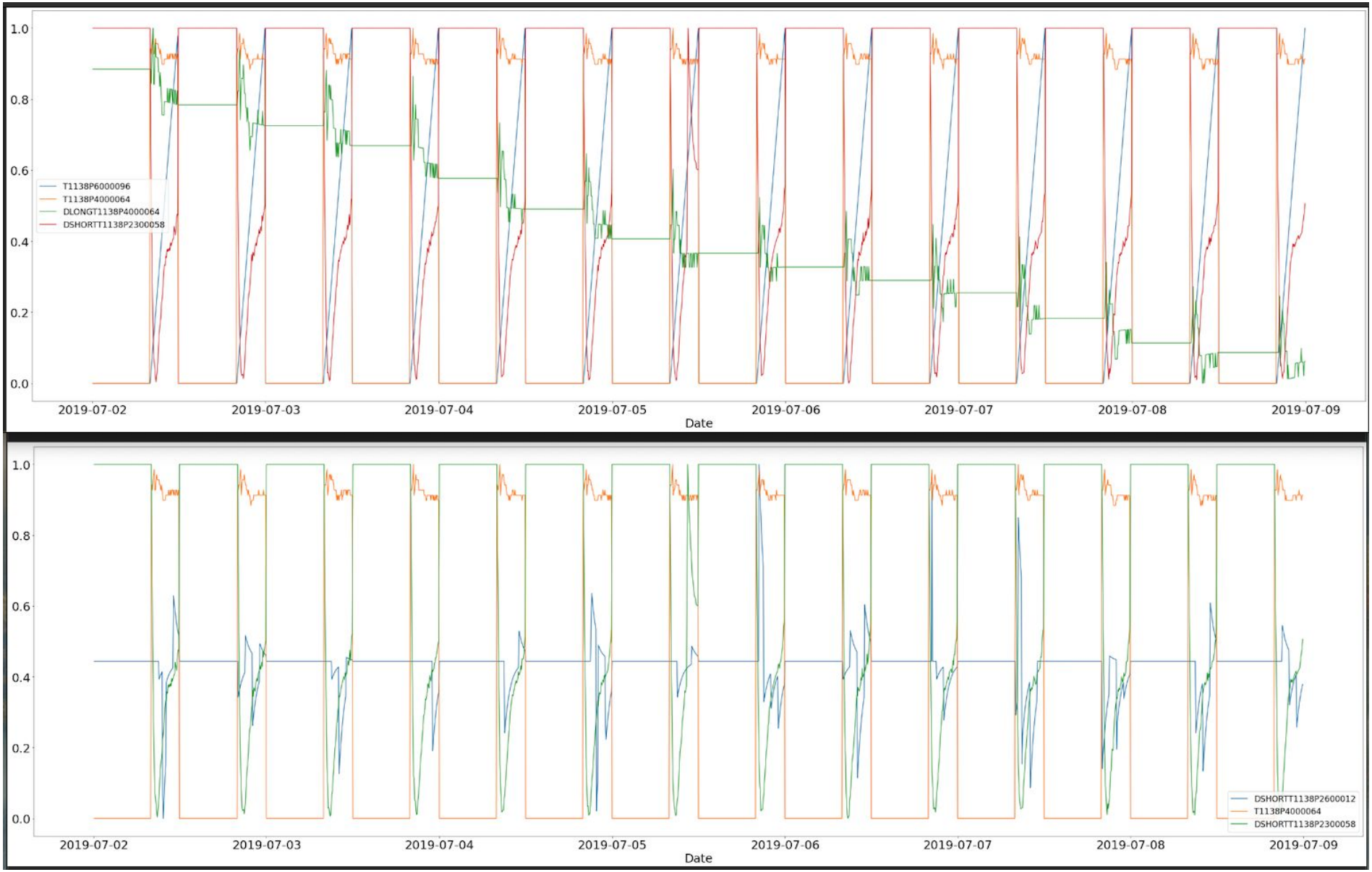
1) **T1138P6000096**: Наработка двигателя с момента последнего включения, сек
2) **T1138P6000315**: Время простоя двигателя с момента последнего выключения, сек
3) **T1138P4000064**: Загрузка двигателя, %
4) **T1138P2600012**: Ток фазы А двигателя, А

*Средняя скорость изменения давления на приеме насоса в ЧАС, МПа/час

Датасет и его обработка 2/3

Признак	Описание	Интервал значений
T1138P6000096	Наработка двигателя с момента последнего включения, с	0-17100
T1138P6000315	Время простоя двигателя с момента последнего выключения, с	0-86100
DSHORTT1138P4000064	Средняя скорость изменения загрузки двигателя ЧАС, %/час	-49.2 - 79.2
DSHORTT1138P2600012	Средняя скорость изменения тока фазы А двигателя в ЧАС, А/час	-21.12 - 21.6
DSHORTT1205P2300000	Средняя скорость изменения давления в коллекторе ИУ в ЧАС, МПа/час	-0.38 - 3.45
T1138P4000064	Загрузка двигателя, %	0 - 108
T1138P2600012	Ток фазы А двигателя, А	0 - 37.60
T1205P2300000	Давление в коллекторе измерительной установки, МПа	0.69 - 3.99
DSHORTT1138P2300058	**Средняя скорость изменения давления на приеме насоса в ЧАС, МПа/час**	**0.30 - 0.31**

Датасет и его обработка 3/3

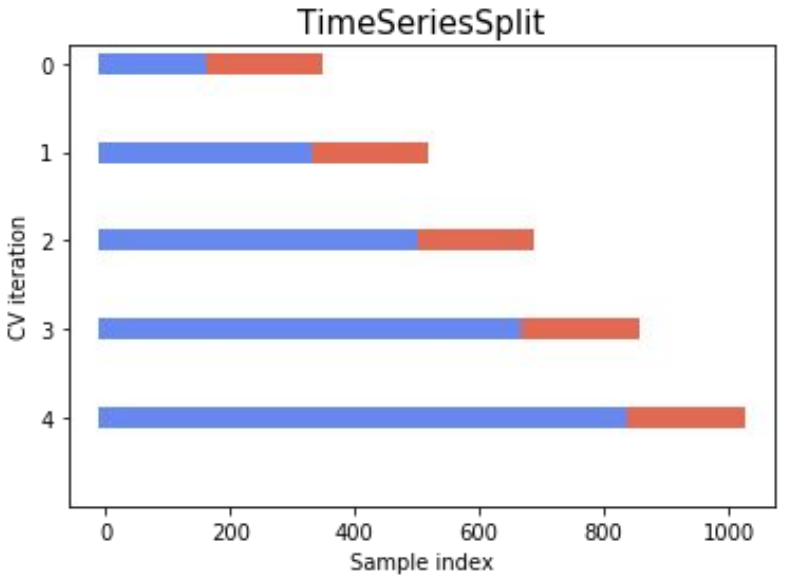


Классическая схема кросс-валидации для TS

Равномерное добавление новых наблюдений к набору данных по мере “течения времени”.

Недостатки:

- Первый синий фолд очень короткий по времени
- Большая дисперсия по метрикам в test фолдах (из-за сильно разного количества данных в train)
- Когда у вас много WELL_ID – нужно сначала группировать по отдельным WELL_ID, а потом делить по времени – это вне стандартного функционала **sklearn.TimeSeriesSplit**

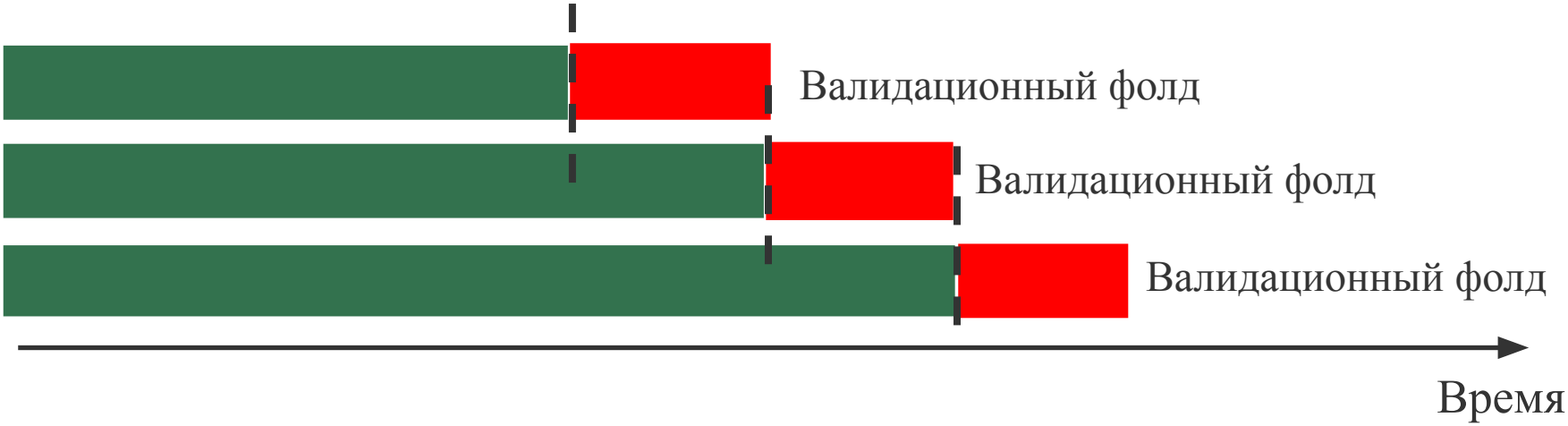


Кастомная схема кросс-валидации для TS

Вся обучающая часть = **Июль месяц** для каждой **WELL_ID**

Сразу отсекаем 50% для первого тренировочного фолда

Остальные 50% делим на равные 3 части (или N частей)

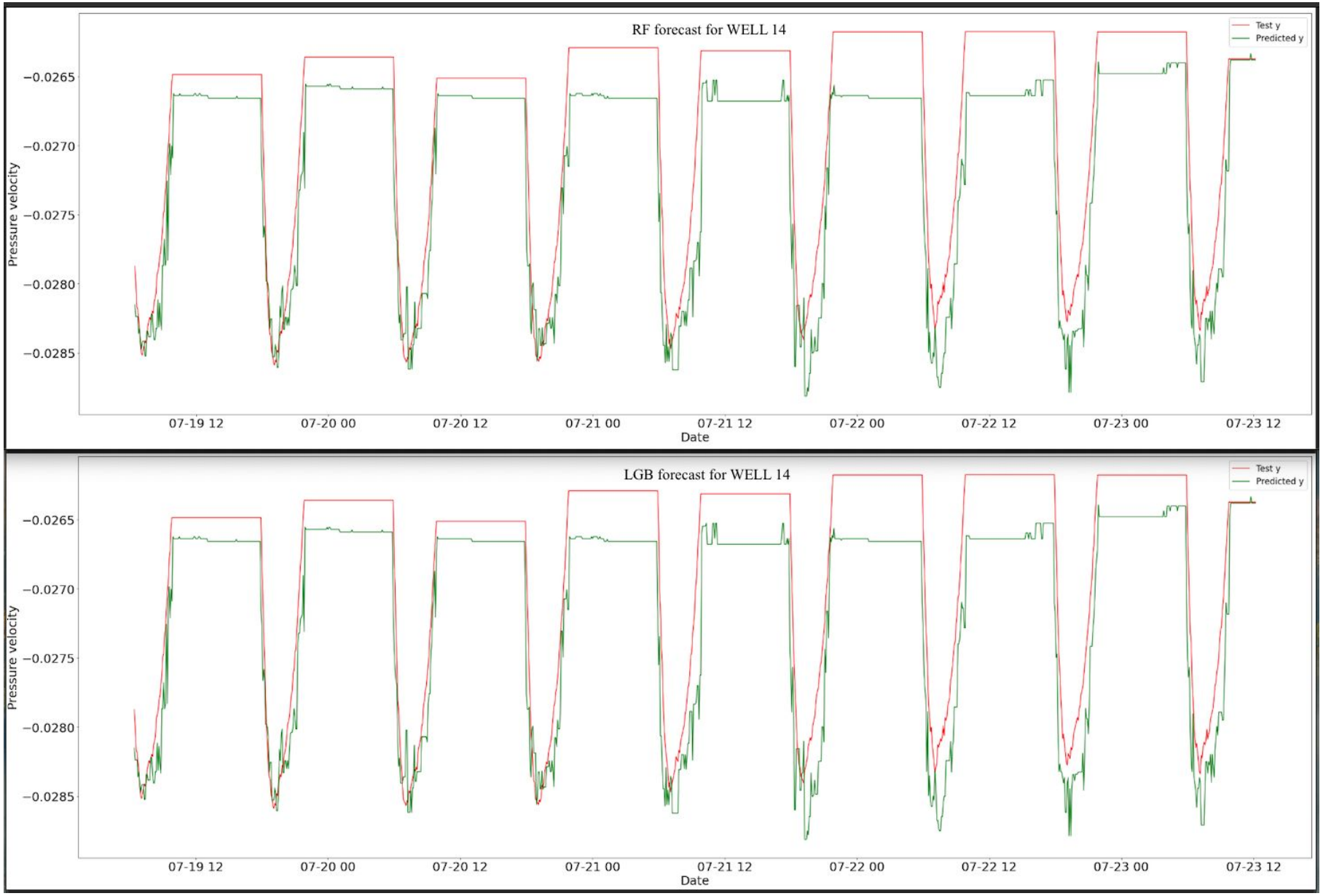


Результаты 1/3

Модель	MAPE average, %
LightGBM	0.0803
RandomForestRegressor	0.0781
Ridge Regression	0.0855

Модель	Best parameters (Скважина 14)		
LightGBM	colsample_bytree	доля параметров для trian каждого дерева	0.9
	learning_rate	скорость обучения при град. спуске	0.09
	max_depth	глубина дерева решения	23
	min_child_weight	кол-во элементов нужных для образования “листа”	2
	subsample	доля данных, отбираемая для построения ветки дерева	0.81
RandomForestRegressor	n_estimators	общее число деревьев	4
	max_depth	макс. глубина дерева	10
	max_features	макс. число features для наилучшего разделения	0.10
	min_samples_leaf	мин. число наблюдений в ноде	1
	min_samples_split	мин. число наблюдений для образования ноды	5

Результаты 2/3



Результаты 3/3

- Ridge Regression. Наибольшее влияние имеют следующие переменные:
  1. **T1205P2300000** - Давление в установке, МПа (Знак +)
  2. **T1138P4000064** - Загрузка двигателя, % (Знак -)
  3. **DSHORTT1205P2300000** - Средняя скорость изменения давления в коллекторе, МПа/час (Знак -)
  4. **DSHORTT1138P2600012** - Средняя скорость изменения тока фазы А, А/час (Знак -)
- *Отрицательное значение коэффициента говорит в пользу увеличения скорости прокачки (создание движущей силы).*

- Наименьшее влияние имеют следующие переменные:
  1. **T1138P6000096** - Время простоя двигателя с момента последнего выключения, с
  2. **T1138P6000096**: Наработка двигателя с момента последнего включения, с
- *Вклад в предсказание скорости изменения давления может быть мал ввиду не информативности переменных после выхода насоса на рабочие параметры добычи.*

Выводы

Что уже сделано?

- Разработана модель на основе LGBoost, Random Forest и Ridge regression;
- Полученная ошибка значительно меньше условно принятой (5%);
- Результаты интерпретируемой модели согласуются с физическим смыслом.

Что можно улучшить?

- Применение Нейронной сети;
- Генерация новых параметров из имеющихся (библиотека tsfresh);
- Рассмотрение влияния новых внешних факторов.

Appendix 1/4

- Light GB

`'learning_rate'` - скорость обучения при град. спуске;

`'max_depth'` - глубина дерева решения (во избежании чрезмерного роста дерева);

`'min_child_weight'` - мин. вес Гессиана/кол-во элементов нужных для образования “листа”; [1]

`'colsample_bytree'` - доля параметров (случайно выбранных), которые будут использоваться для trian каждого дерева; [2]

`'subsample'` - `bagging_fraction` - доля данных, отбираемая для построения ветки дерева [3]

`'n_estimators'` - количество деревьев для fit [4]

`'eval_metric': 'l2'` - Евклидова метрика (среднеквадратичная ошибка) [5]

- Random Forest

`'min_samples_leaf'` - мин. число наблюдений в ноде; [6]

`'min_samples_split'` - мин. число наблюдений для разбиения внутренней ноды;

`'max_depth'` - макс. глубина дерева;

`'max_features'` - макс. число features, которые следует учитывать при поиске наилучшего разделения

`'n_estimators'` - общее число деревьев;

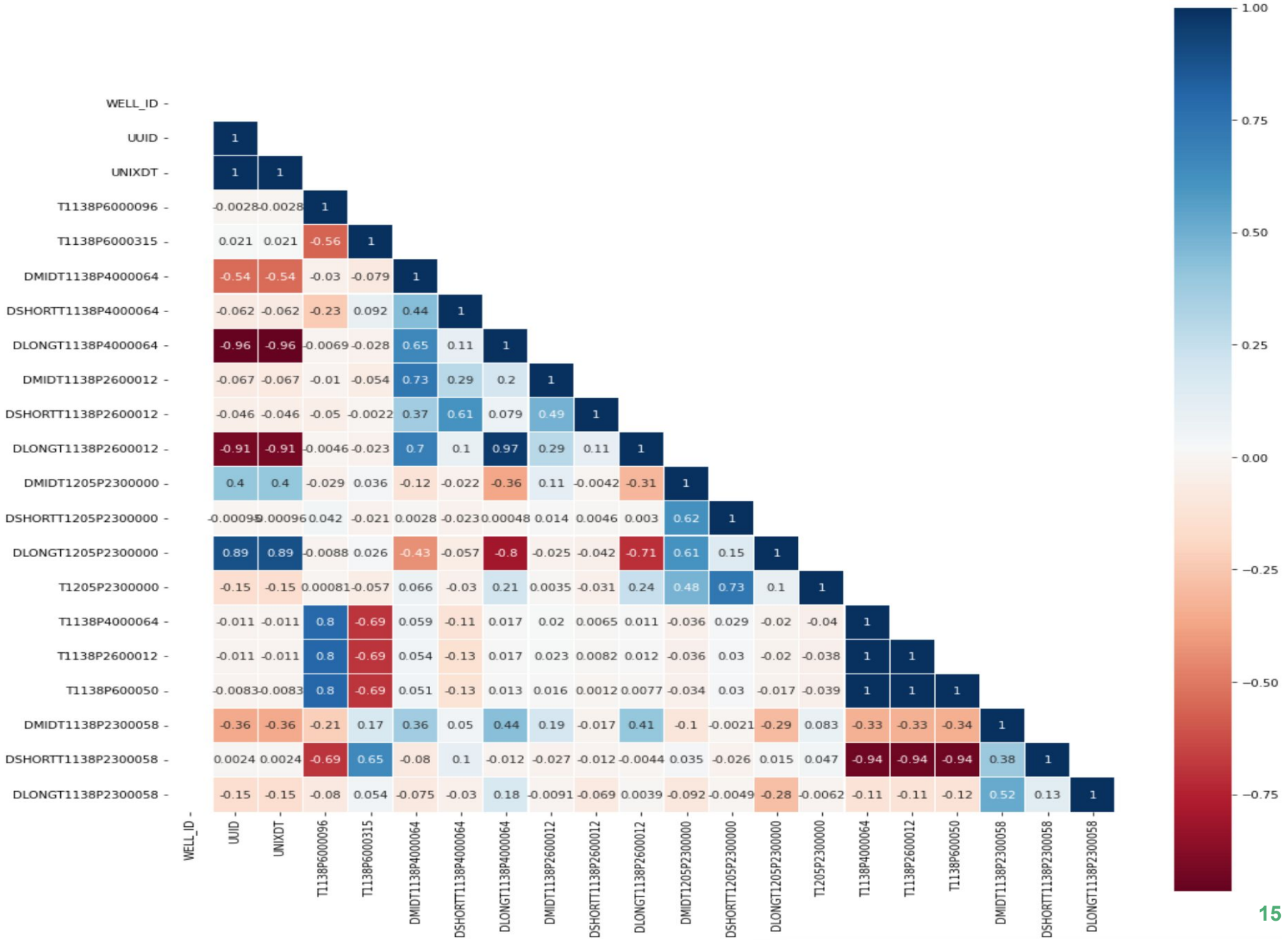
`'eval_metric': 'auc'` - площадь под кривой ROC (TPR vs. FPR) [7]

Appendix 2/4

- Библиотека tsfresh для генерации признаков из временного ряда
- Функция EfficientFCParameters() позволяет сгенерировать большое количество новых объясняющих признаков:
  1. Квантили
  2. Линейные тренды и агрегирующие функции от них
  3. Коэффициенты преобразования Фурье и их агрегаты
  4. Коэффициенты Вейвлет преобразования и их агрегаты
  5. Минимумы и максимумы функций их положения во времени

- Требует очень больших вычислительных мощностей
- Сгенерированные признаки слабоинтерпретируемы

Appendix 3/4


WELL_ID
UUID
UNIXDT
T1138P6000096
T1138P6000315
DMIDT1138P4000064
DSHORTT1138P4000064
DLONGT1138P4000064
DMIDT1138P2600012
DSHORTT1138P2600012
DLONGT1138P2600012
DMIDT1205P2300000
DSHORTT1205P2300000
DLONGT1205P2300000
T1205P2300000
T1138P4000064
T1138P2600012
T1138P600050
DMIDT1138P2300058
DSHORTT1138P2300058
DLONGT1138P2300058
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Appendix 4/4

- Метод определяет такие оси (PCA компоненты) в пространстве признаков, относительно которых дисперсия (информативность) максимальная.
- Оси должны быть ортогональны друг другу.
- Берем N компонент, которые описывают кумулятивную (например, 95%) дисперсию выборки.