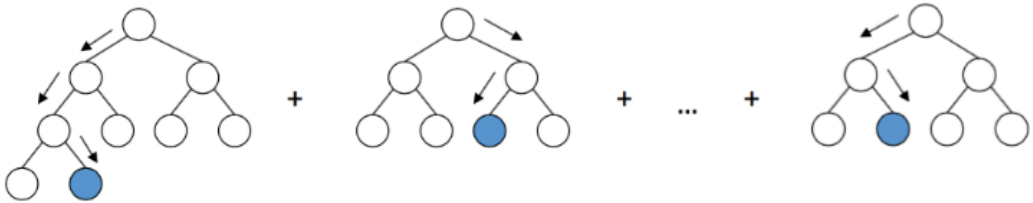
Решение задачи многомерной регрессии методом Gradient Boosting (CatBoost)



Матвеев И. Ю.

Постановка задачи

Необходимо выполнить следующие пункты:

- Провести предварительный анализ данных;
- Построить регрессионную модель. Обосновать выбор модели;
- Сравнить ошибки на тестовых и тренировочных данных. Использовать метрику R^2;
- Построить зависимости ошибок на тестовых и тренировочных данных.

Предварительный анализ данных

Dataset состоит из матриц признаков X и значений Y:

$X = \{s_{mt}, s_{mq}, d, h_p\}$ и $Y = \{QW, DP\}$

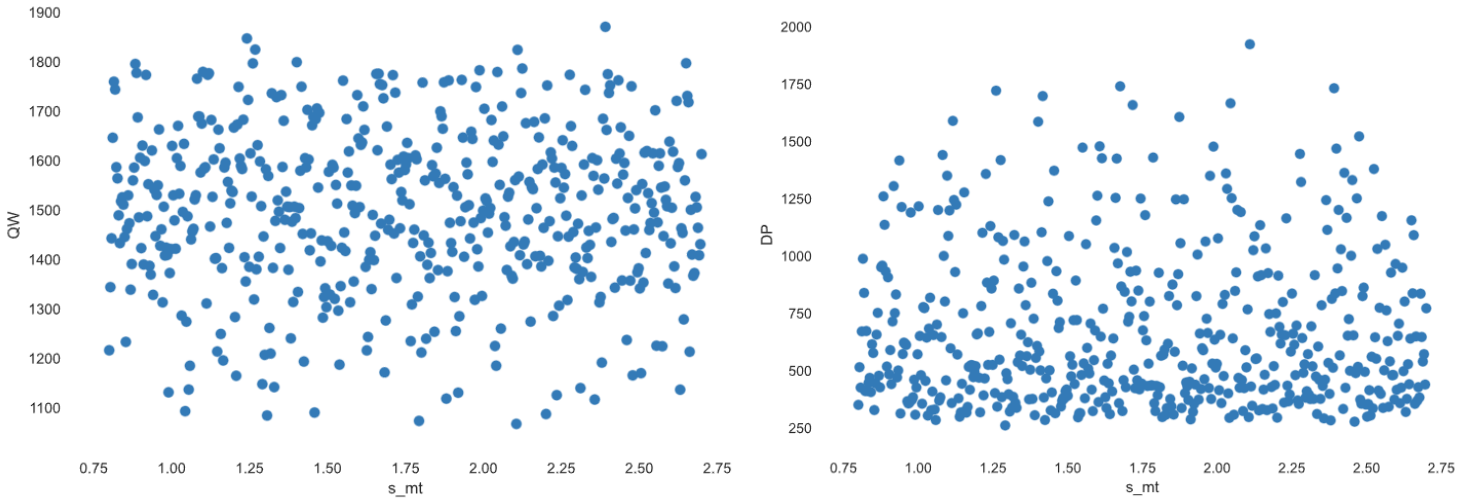
- где каждый объект матриц это вектор $\bar{v}_i, \in \mathbb{R}^{500\times1}$

Диапазон значений входных данных представлен в таблице:

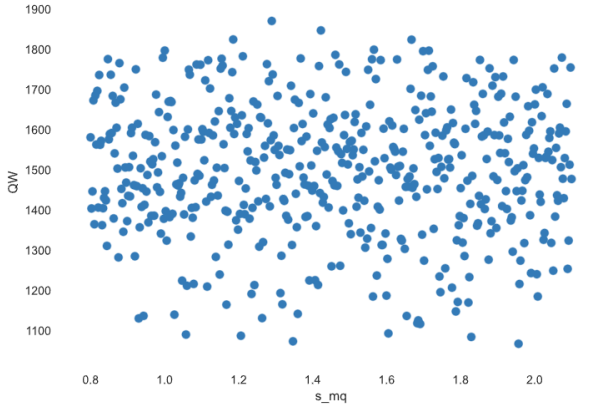
	s_{mt}	s_{mq}	d	h_p
Min	0.8	0.8	1.0	4.0
Max	2.7	2.1	3.0	10.0

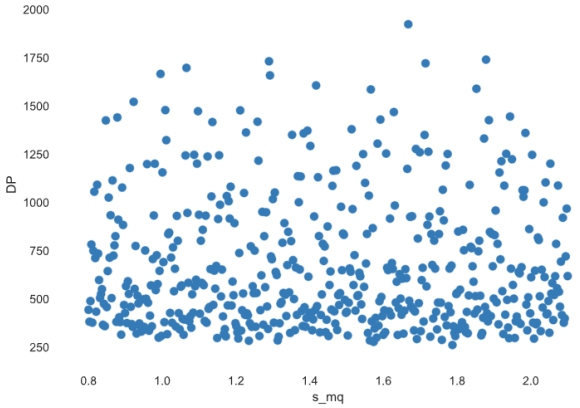
Предварительный анализ данных

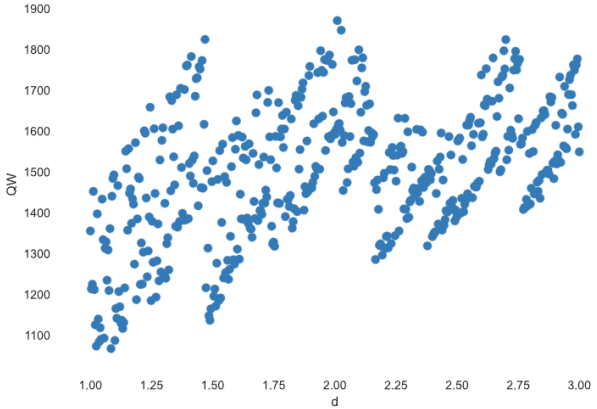
- Среди генеральной совокупности не обнаружены пустые ячейки значений переменных и ложные типы данных;
- Ниже продемонстрированы зависимости целевых переменных от признаков $X = \{s_{mt}, s_{mq}, d, h_p\}$ и $Y = \{QW, DP\}$:

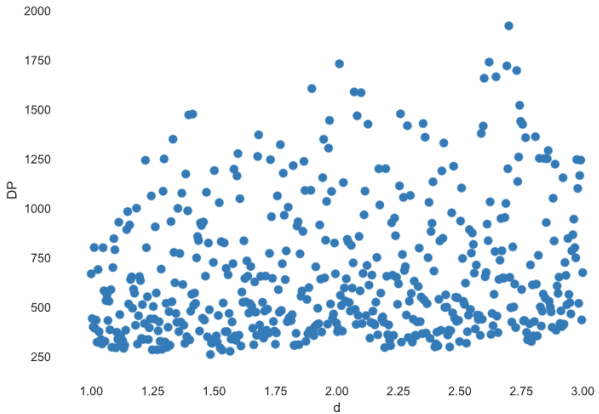


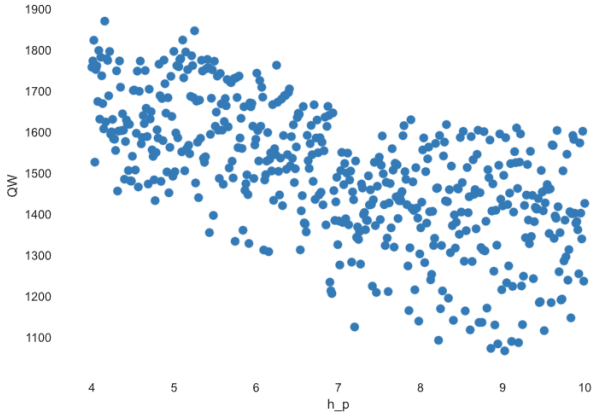
Предварительный анализ данных

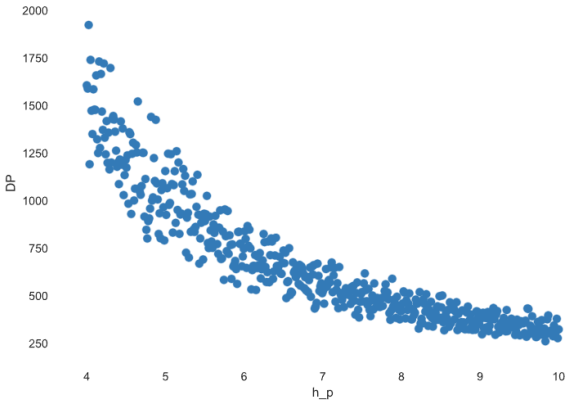




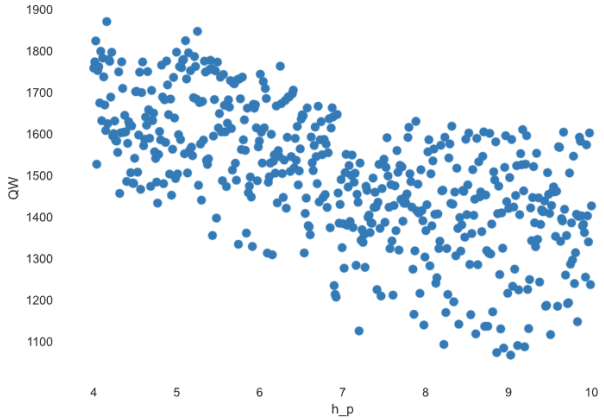


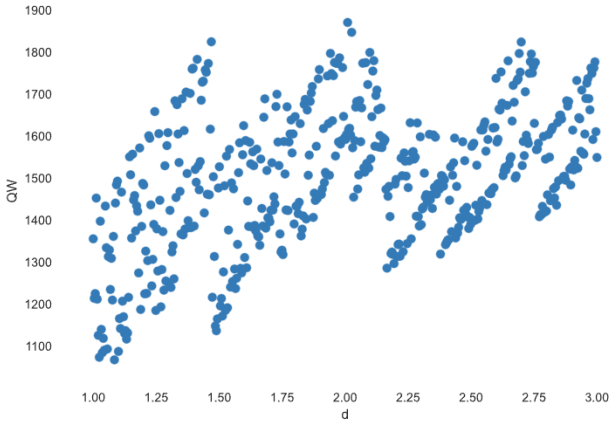


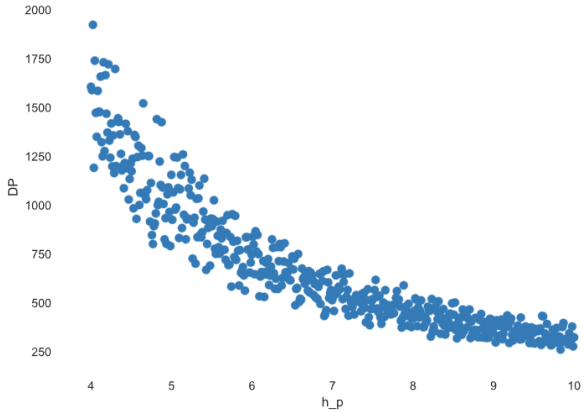




Предварительный анализ данных

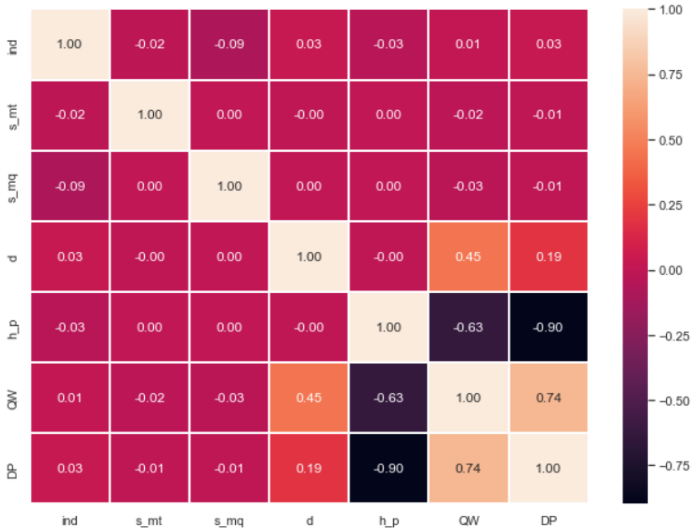






- Зависимости наблюдается у пар переменных: $(h_p, QW), (h_p, DP), (d, QW)$;
- Далее необходимо построить матрицу корреляций переменных для вычисления коэффициентов Пирсона.

Предварительный анализ данных



- Как видно из матрицы корреляций линейная зависимость наблюдается у группы признаков: $(h_p, QW), (h_p, DP), (QW, DP)$. При этом влияние признака h_p уменьшает значения целевых переменных QW, DP. Признак d имеет слабую линейную зависимость с целевой переменной QW;
- Входные признаки между собой никак не коррелирует и следовательно они либо зависят нелинейно, либо связь полностью отсутствует.

Выбор и построение модели

На основе предварительного анализа данных можно сделать вывод о применимости регрессионных моделей. Примеры моделей представлены ниже:

- Линейная регрессия
- Регрессия LASSO

Только для данных, которые сильно коррелируют

- Гребневая регрессия
- Нейросетевая модель

Необходимо много данных для качественного обучения

- Деревья решений/**Gradient Boosting**

Отлично могут описать любую форму зависимости данных

Выбор и построение модели

Перед обучением модели **Gradient Boosting** необходимо задать следующие гиперпараметры: **iterations = 100, learning rate = 0.1, Loss function = 'MultiRMSE'**

Оптимизация функционала ошибки на train, test

Анализ результатов алгоритма

- В результате рассчитана метрика $\boldsymbol{R^2}$ для тренировочного и тестового набора данных: $\boldsymbol{R^2_{train} = 0.975}$, $\boldsymbol{R^2_{test} = 0.924}$;
- Определены максимальные и минимальные относительные ошибки предсказаний $\widehat{QW}, \widehat{DP}$ для каждой выходной переменной QW, DP:

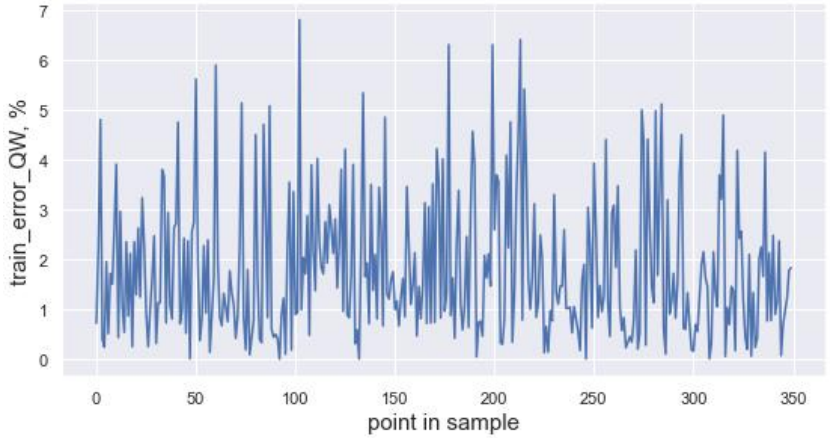
Max Relative Error (%)	Train	Test
QW	6.798	11.428
DP	14.618	18.750

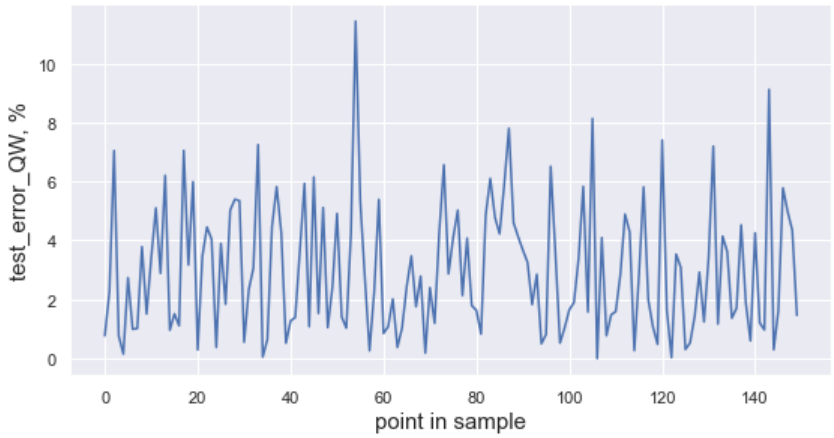
Min Relative Error (%)	Train	Test
QW	0.004	0.001
DP	0.011	0.009

Анализ результатов алгоритма

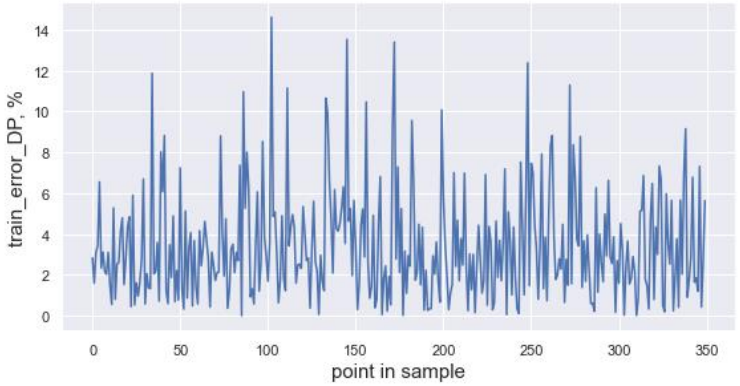
Характер изменения относительных ошибок для величин QW, DP на тренировочных и тестовых данных.

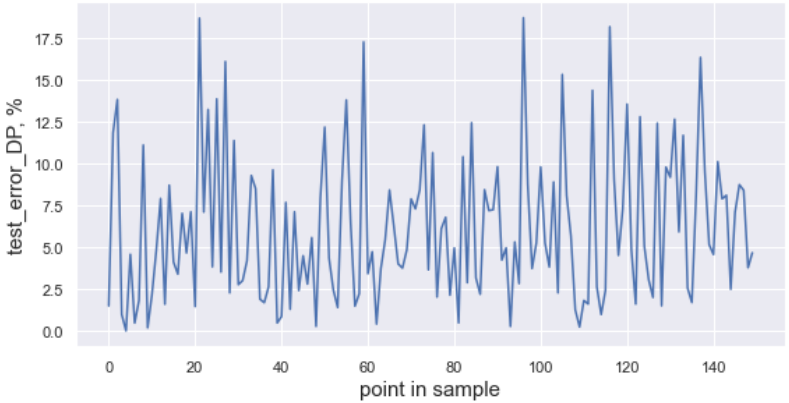
- По оси абсцисс указаны номера объектов в train/test data;
- По оси ординат распределена относительная ошибка (%) для каждого объекта в указанной выборке.





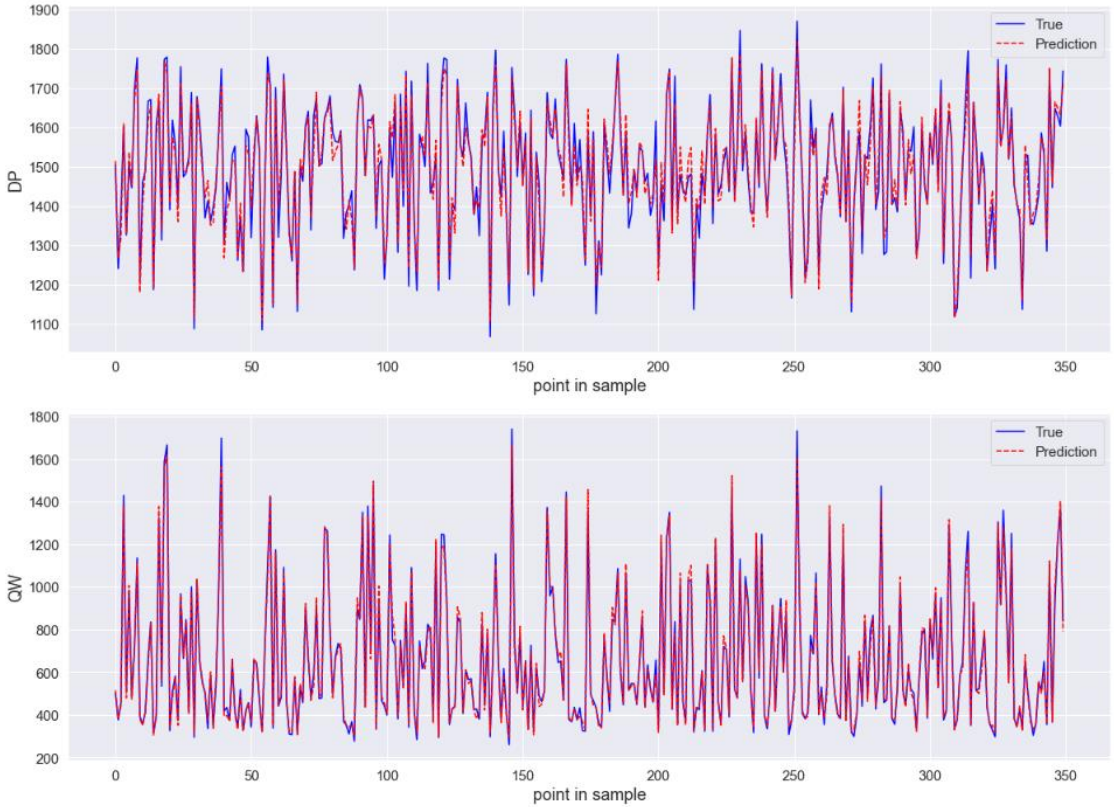
Анализ результатов алгоритма





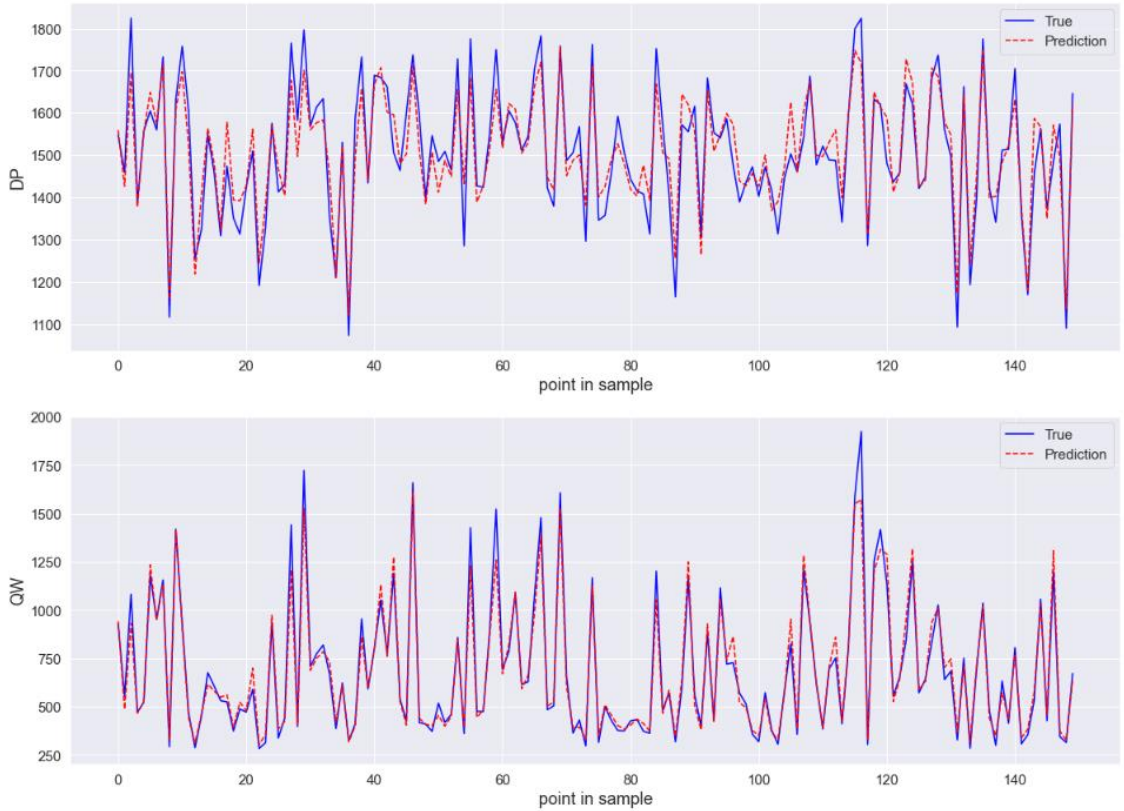
Анализ результатов алгоритма

Аппроксимация на тренировочных данных для величин QW, DP



Анализ результатов алгоритма

Аппроксимация на тестовых данных для величин QW, DP



Вывод

- Выполнен предварительный анализ данных;
- Построена модель регрессии **Gradient Boosting**;
- Определены и построены относительные ошибки на тестовых и тренировочных данных;
- Рассчитаны метрики $\boldsymbol{R^2_{train} = 0.975}$, $\boldsymbol{R^2_{test} = 0.924}$;
- Построены графики аппроксимации данных на базе **Gradient Boosting**.