ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПО КУРСУ

«DATA SCIENCE»

ПО ТЕМЕ:

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель: Насыров А.И.

Характеристики анализируемого набора данных

- Файлы X_bp.xlsx и X_nup.xlsx были загружены и объединены в одну таблицу
- Объем и характеристики единого датасета: 1023 строки и 13 колонок

	1018.0	1019.0	1020.0	1021.0	1022.0
Соотношение матрица-наполнитель	2.271346	3.444022	3.280604	3.705351	3.808020
Плотность, кг/м3	1952.087902	2050.089171	1972.372865	2066.799773	1890.413468
модуль упругости, ГПа	912.855545	444.732634	416.836524	741.475517	417.316232
Количество отвердителя, м.%	86.992183	145.981978	110.533477	141.397963	129.183416
Содержание эпоксидных групп,%_2	20.123249	19.599769	23.957502	19.246945	27.474763
Температура вспышки, С_2	324.774576	254.215401	248.423047	275.779840	300.952708
Поверхностная плотность, г/м2	209.198700	350.660830	740.142791	641.468152	758.747882
Модуль упругости при растяжении, ГПа	73.090961	72.920827	74.734344	74.042708	74.309704
Прочность при растяжении, МПа	2387.292495	2360.392784	2662.906040	2071.715856	2856.328932
Потребление смолы, г/м2	125.007669	117.730099	236.606764	197.126067	194.754342
Угол нашивки, град	90.000000	90.000000	90.000000	90.000000	90.000000
Шаг нашивки	9.076380	10.565614	4.161154	6.313201	6.078902
Плотность нашивки	47.019770	53.750790	67.629684	58.261074	77.434468

Этапы выполненной работы

Разведочный анализ данных с визуализацией

- Анализ данных на наличие пропусков, дубликатов, уникальных значений
- Визуализация с использованием гистограмм распределения , “ящик с усами” и попарных графиков рассеяния точек
- Проверка признаков на нормальное распределение
- Нахождение и удаление выбросов
- Составление корреляционной тепловой карты до и после предобработки данных
- Балансировка и добавление дополнительных синтетических данных и создание 3-х новых датасетов для каждой целевой переменной с целью улучшения работы модели
- Нормализация данных
- Сохранение датасетов в отдельные файлы для каждой модели.

Обучение моделей

- Выделение целевых переменных из датасетов и преобразование в ndarray
- Разбиение на тренировочную и тестовую выборки
- Для прогнозирования «Модуль упругости при растяжении» и «Прочность при растяжении» построены регрессионные модели на классических алгоритмах машинного обучения это: гребневая регрессия, метод опорных векторов, метод к-ближайших соседей и ансамблевые методы использующие алгоритмы бэггинга и бустинга. Подбор гиперпараметров произведен с кросс-валидацией при помощи gridsearchcv и с использование библиотеки optuna. В конце была произведена оценка лучшей модели при помощи следующих метрик: R2, MAE. RMSE. MAX с визуализацией. Лучшая модель выбрана и сохранена для создания веб-приложения.
- Для рекомендации «Соотношение матрица-наполнитель» используется полносвязная нейронная сеть. Использовалась библиотека Tensorflow.keras. Количество внутренних слоев 2, количество нейронов 128 и 64 соответственно, Dropout=0.8, в качестве функции активации использовался гиперболический тангенс, оптимизатор rmsprop, метрика MAE.

Разработка веб-приложения

- Для создания использовался микрофреймворк Flask. Готовое приложение выведено в продакшн на хостинг: https://vkr-deploy.onrender.com

Разведочный анализ данных

Описательная статистика

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.930366	0.913222	0.389403	2.317887	2.906878	3.552660	5.591742
Плотность, кг/м3	1023.0	1975.734888	73.729231	1731.764635	1924.155467	1977.621657	2021.374375	2207.773481
модуль упругости, ГПа	1023.0	739.923233	330.231581	2.436909	500.047452	739.664328	961.812526	1911.536477
Количество отвердителя, м.%	1023.0	110.570769	28.295911	17.740275	92.443497	110.564840	129.730366	198.953207
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301	14.254985	20.608034	22.230744	23.961934	33.000000
Температура вспышки, С_2	1023.0	285.882151	40.943260	100.000000	259.066528	285.896812	313.002106	413.273418
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690	0.603740	266.816645	451.864365	693.225017	1399.542362
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983	64.054061	71.245018	73.268805	75.356612	82.682051
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006	1036.856605	2135.850448	2459.524526	2767.193119	3848.436732
Потребление смолы, г/м2	1023.0	218.423144	59.735931	33.803026	179.627520	219.198882	257.481724	414.590628
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1023.0	6.899222	2.563467	0.000000	5.080033	6.916144	8.586293	14.440522
Плотность нашивки	1023.0	57.153929	12.350969	0.000000	49.799212	57.341920	64.944961	103.988901

- Тип данных float64
- Дубликатов нет
- Пропуски отсутствуют

```
1 #Проверим есть ли дубликаты
2 df.duplicated().sum()
```
✓ 0.4s

```
0
```

```
1 #Сделаем проверку на пропуски
2 df.isnull().sum()
```
✓ 0.7s

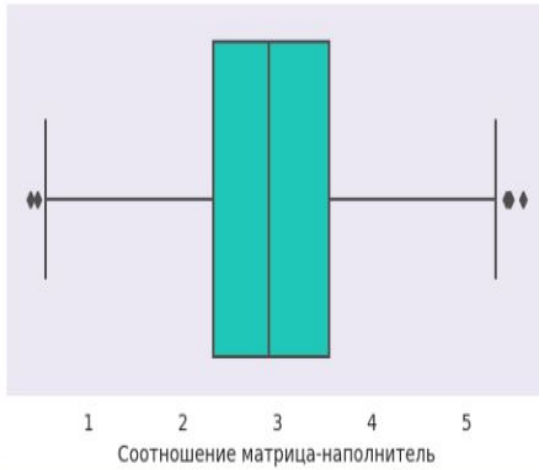
```
Соотношение матрица-наполнитель          0
Плотность, кг/м3                         0
модуль упругости, ГПа                    0
Количество отвердителя, м.%              0
Содержание эпоксидных групп,%_2          0
Температура вспышки, С_2                 0
Поверхностная плотность, г/м2            0
Модуль упругости при растяжении, ГПа     0
Прочность при растяжении, МПа            0
Потребление смолы, г/м2                  0
Угол нашивки, град                       0
Шаг нашивки                              0
Плотность нашивки                        0
dtype: int64
```

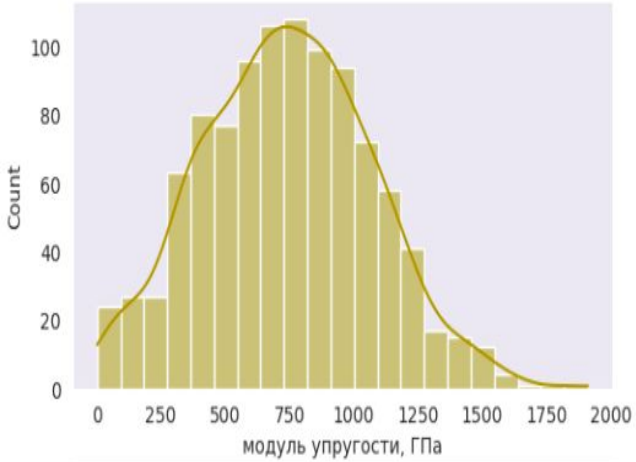
```
1 #Проверим типы
2 df.dtypes
```
✓ 1.0s

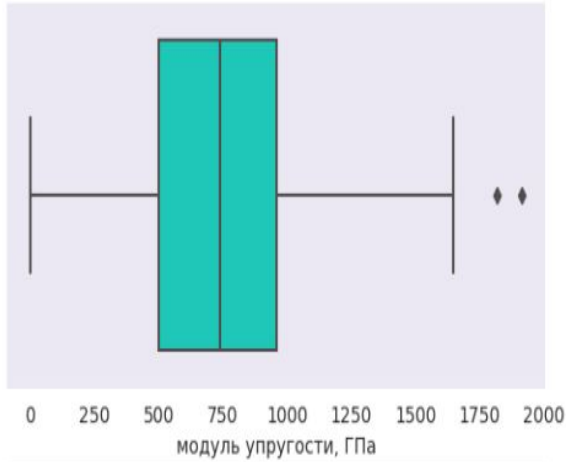
```
Соотношение матрица-наполнитель          float64
Плотность, кг/м3                         float64
модуль упругости, ГПа                    float64
Количество отвердителя, м.%              float64
Содержание эпоксидных групп,%_2          float64
Температура вспышки, С_2                 float64
Поверхностная плотность, г/м2            float64
Модуль упругости при растяжении, ГПа     float64
Прочность при растяжении, МПа            float64
Потребление смолы, г/м2                  float64
Угол нашивки, град                       float64
Шаг нашивки                              float64
Плотность нашивки                        float64
dtype: object
```

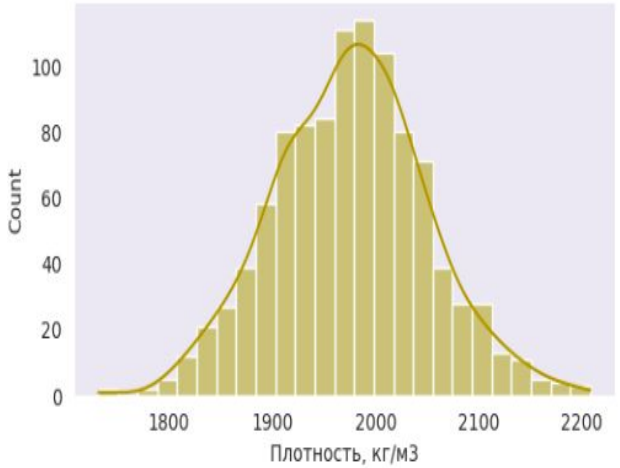
Визуализация данных

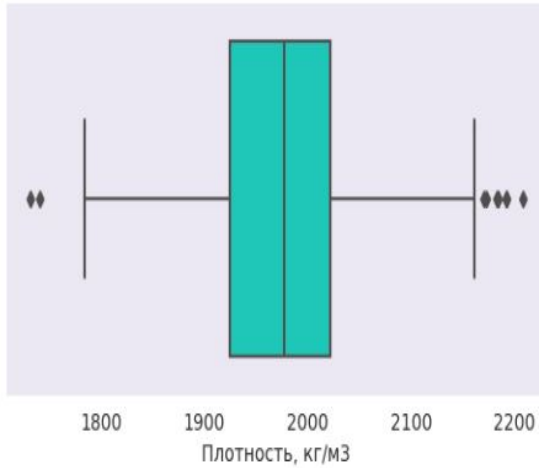
Гистограммы распределения и “ящики с усами”

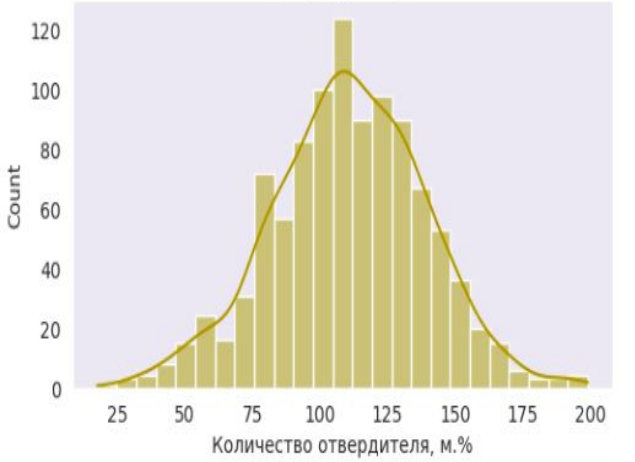


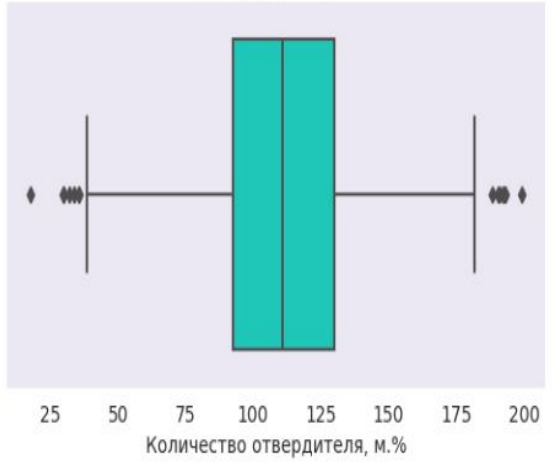


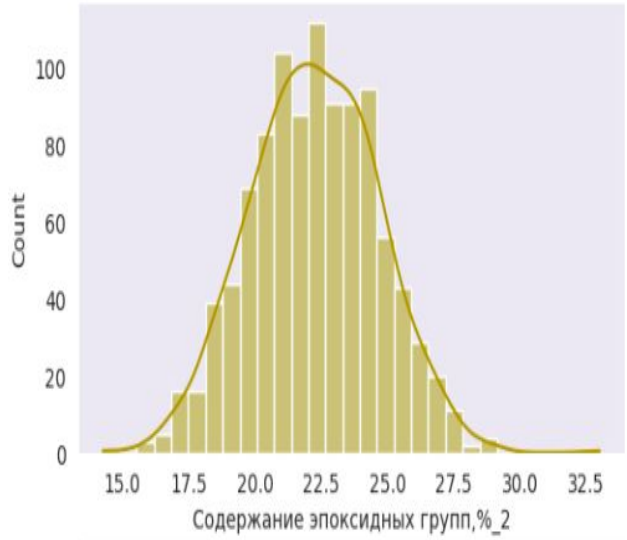


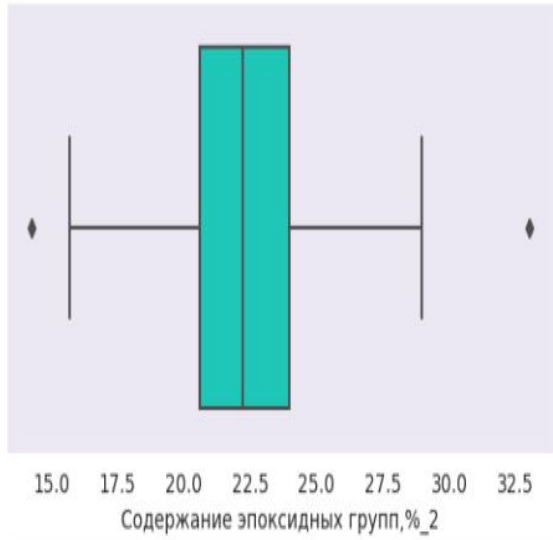


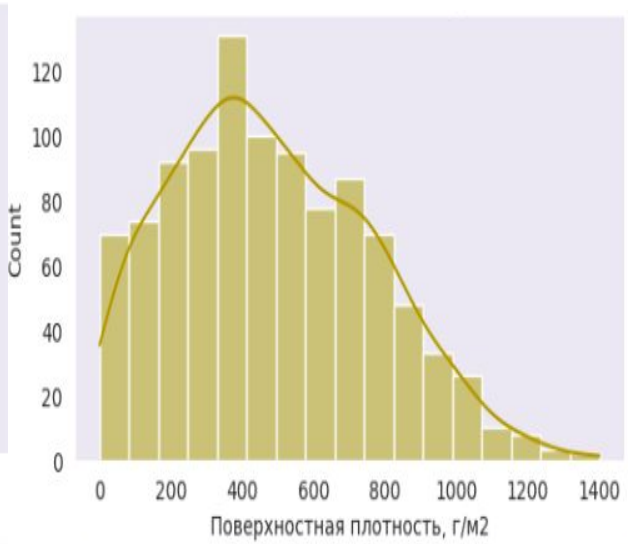


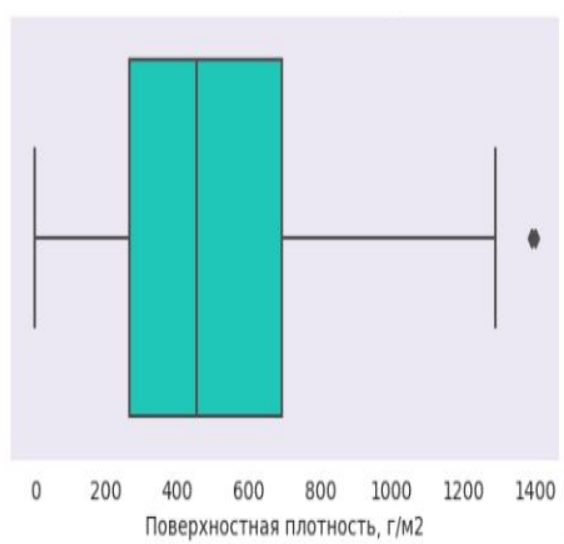


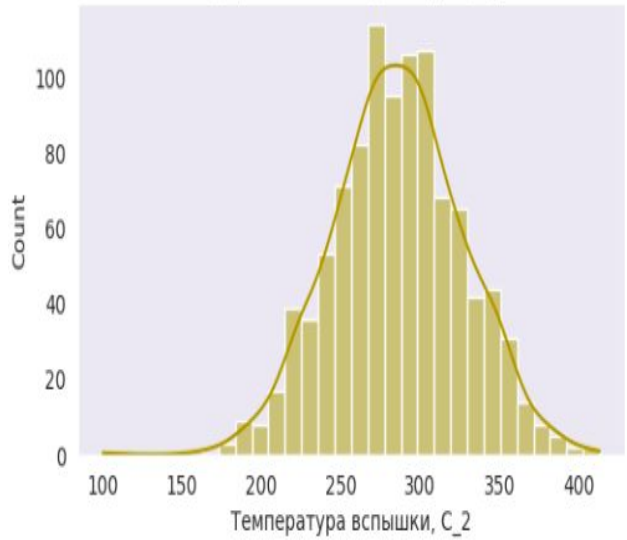


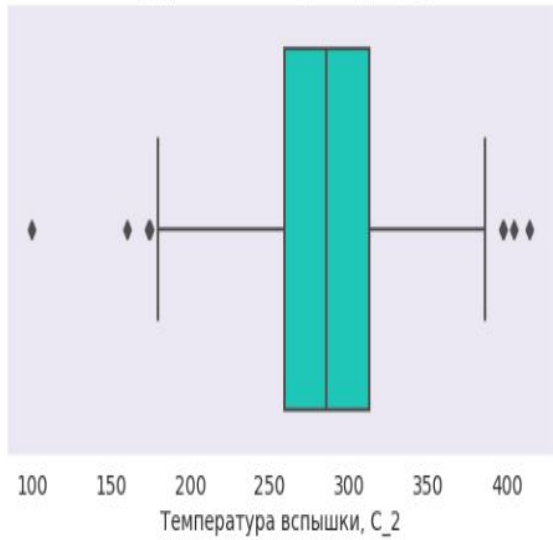

Count
Содержание эпоксидных групп,%_2

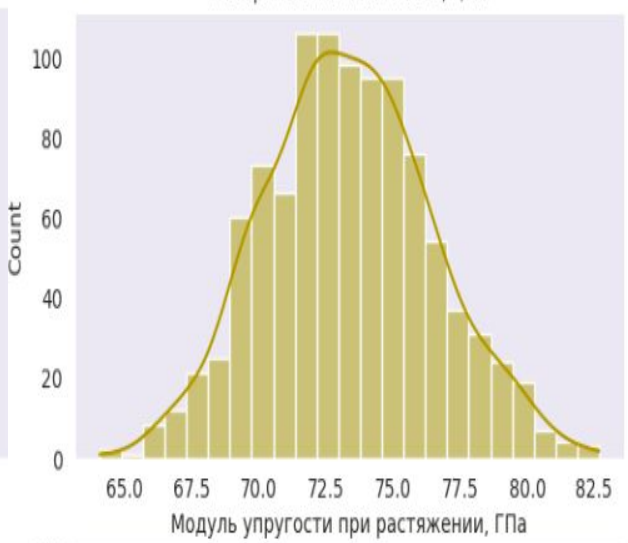

Содержание эпоксидных групп,%_2

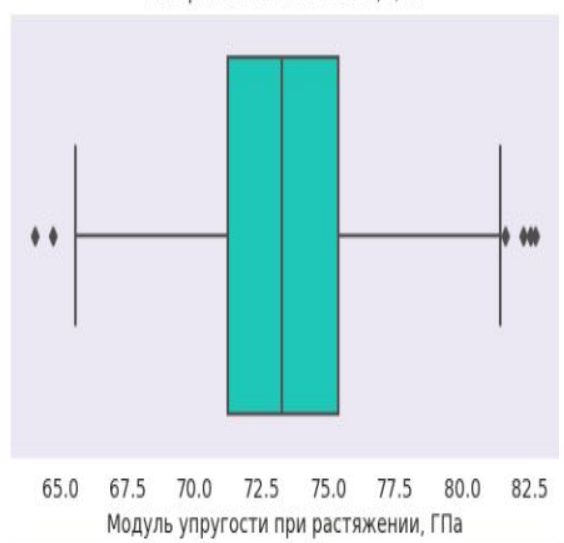

Count
Поверхностная плотность, г/м2

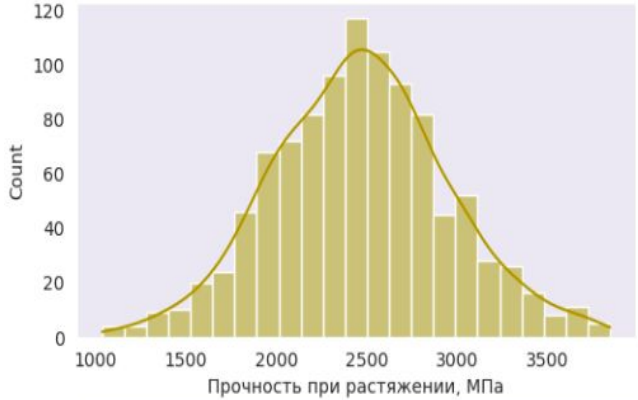

Поверхностная плотность, г/м2

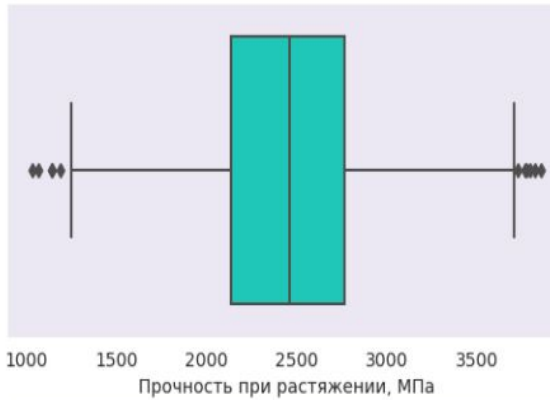

Count
Температура вспышки, С_2

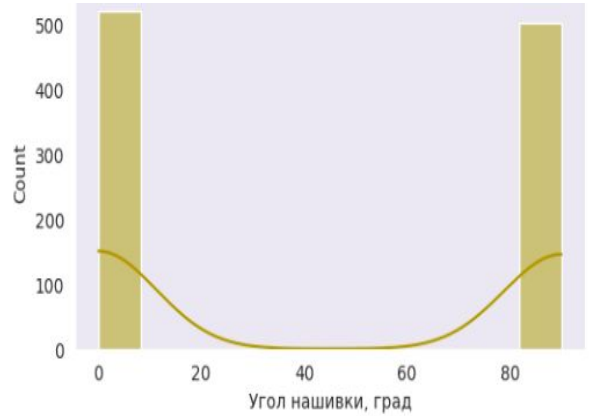

Температура вспышки, С_2

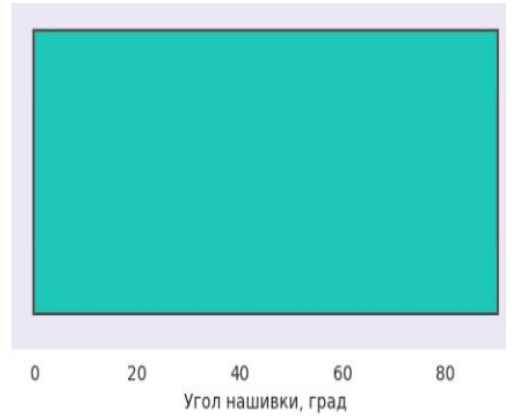

Count
Модуль упругости при растяжении, ГПа

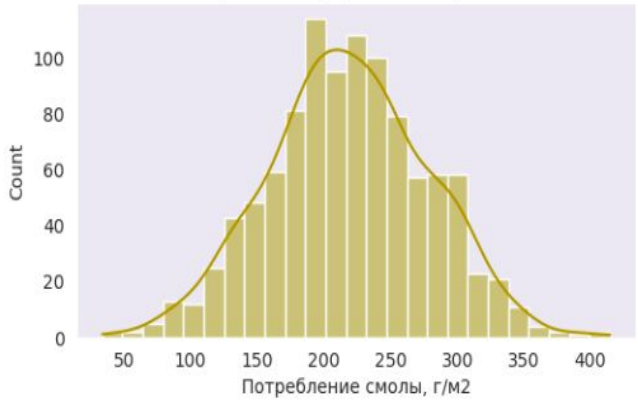

Модуль упругости при растяжении, ГПа

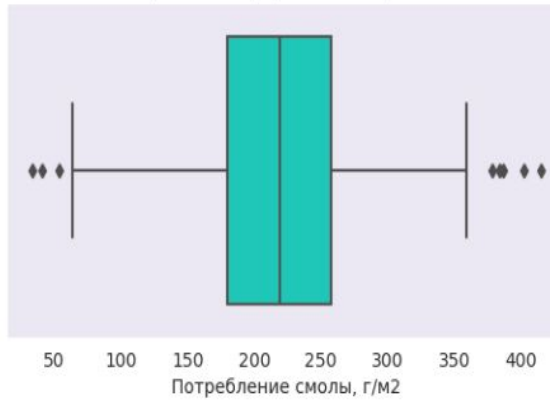

Count
Прочность при растяжении, МПа

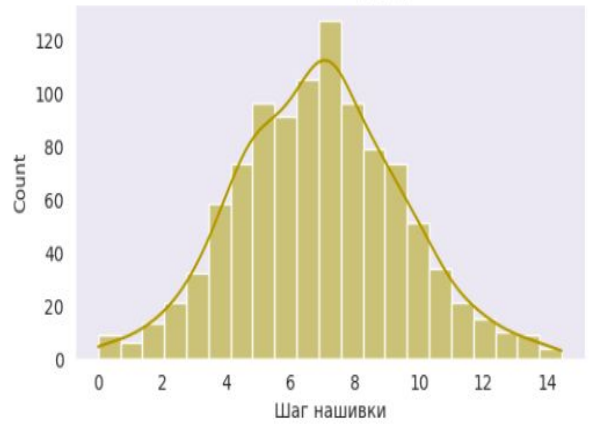

Прочность при растяжении, МПа

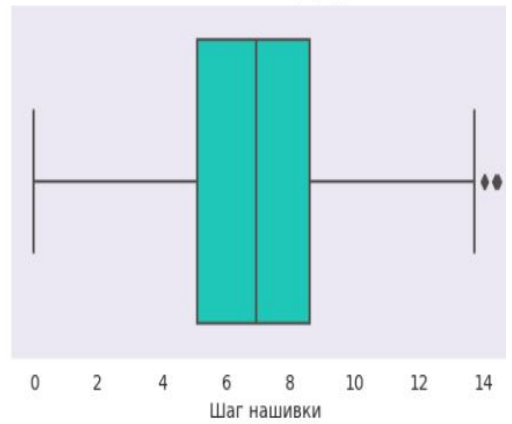

Count
Угол нашивки, град

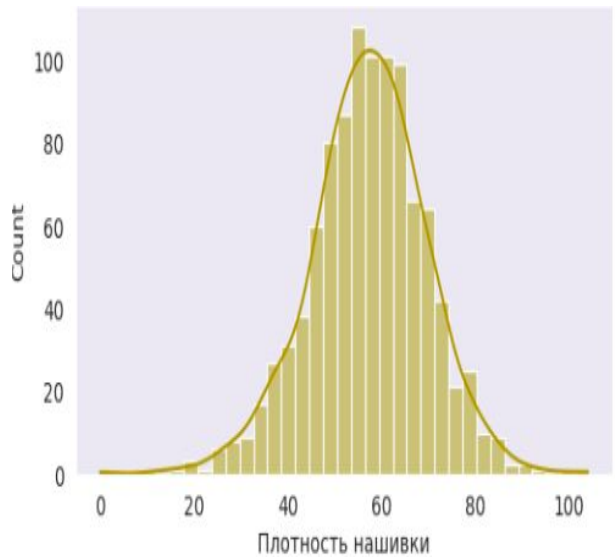

Угол нашивки, град

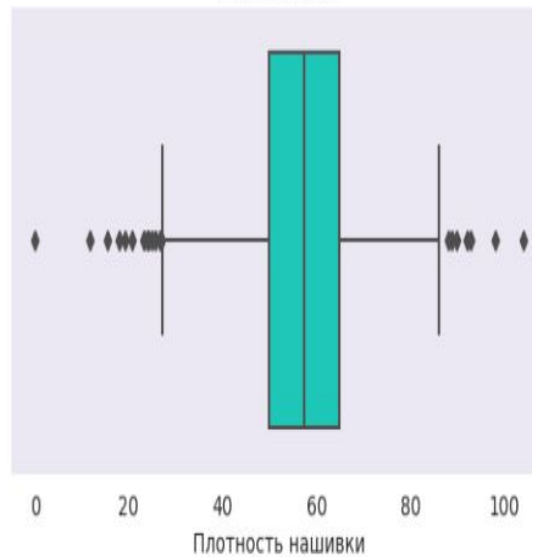

Count
Потребление смолы, г/м2


Потребление смолы, г/м2


Count
Шаг нашивки

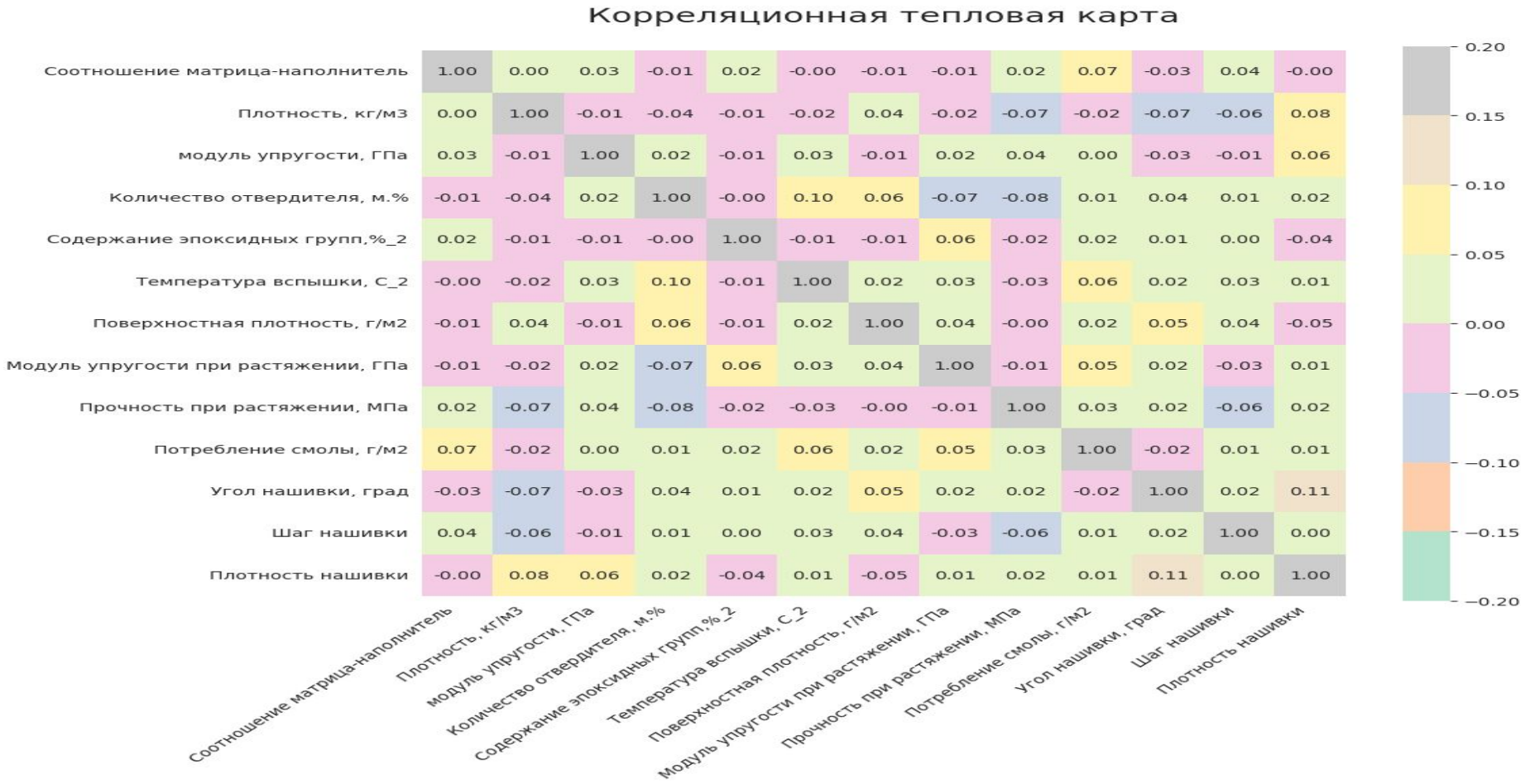

Шаг нашивки


Count
Плотность нашивки


Плотность нашивки

Корреляционная тепловая карта

Максимальная корреляция между Плотностью нашивки и углом нашивки и составляет 0.11, что говорит об отсутствии зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, что говорит об отсутствии корреляционных связей между переменными. Линейной зависимости нет



Удаление выбросов

Выбросы удаляются методом межквантильных расстояний

```
for col in df.columns:
    q75,q25 = np.quantile(df.loc[:,col],[0.75,0.25])
    iqr = q75 - q25

    min = q25-(1.5*iqr)
    max = q75+(1.5*iqr)

    # заменим на нулевые значения
    df.loc[df[col] < min,col] = np.nan
    df.loc[df[col] > max,col] = np.nan

df.isnull().sum()
```

✓ 0.2s

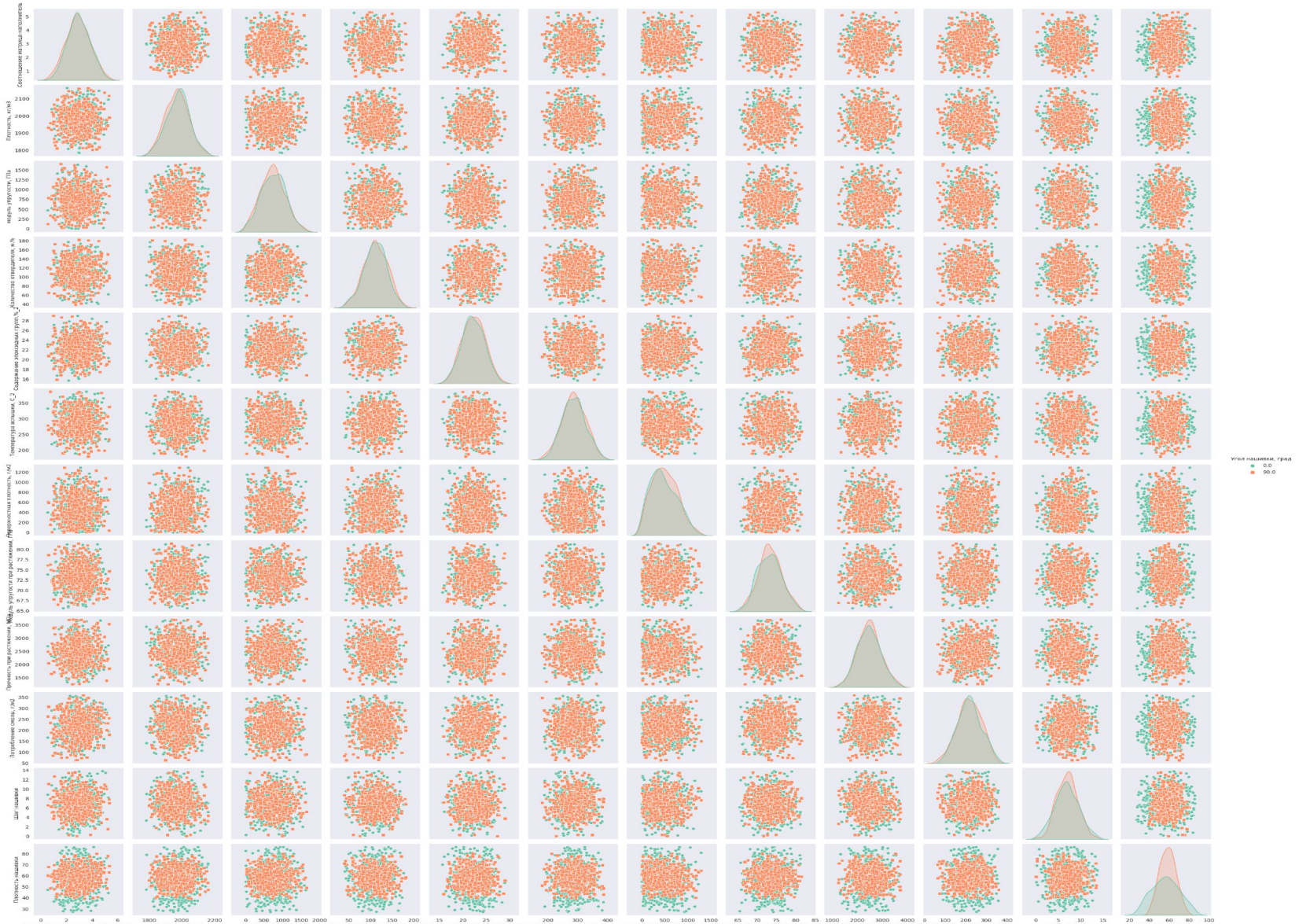
```
Соотношение матрица-наполнитель          6
Плотность, кг/м3                         9
модуль упругости, ГПа                    2
Количество отвердителя, м.%             14
Содержание эпоксидных групп,%_2          2
Температура вспышки, С_2                 8
Поверхностная плотность, г/м2            2
Модуль упругости при растяжении, ГПа     6
Прочность при растяжении, МПа           11
Потребление смолы, г/м2                  8
Угол нашивки, град                       0
Шаг нашивки                              4
Плотность нашивки                       21
dtype: int64
```

```
# удалим строки с нулевыми значениями
df = df.dropna(axis=0)
df.isnull().sum()
```

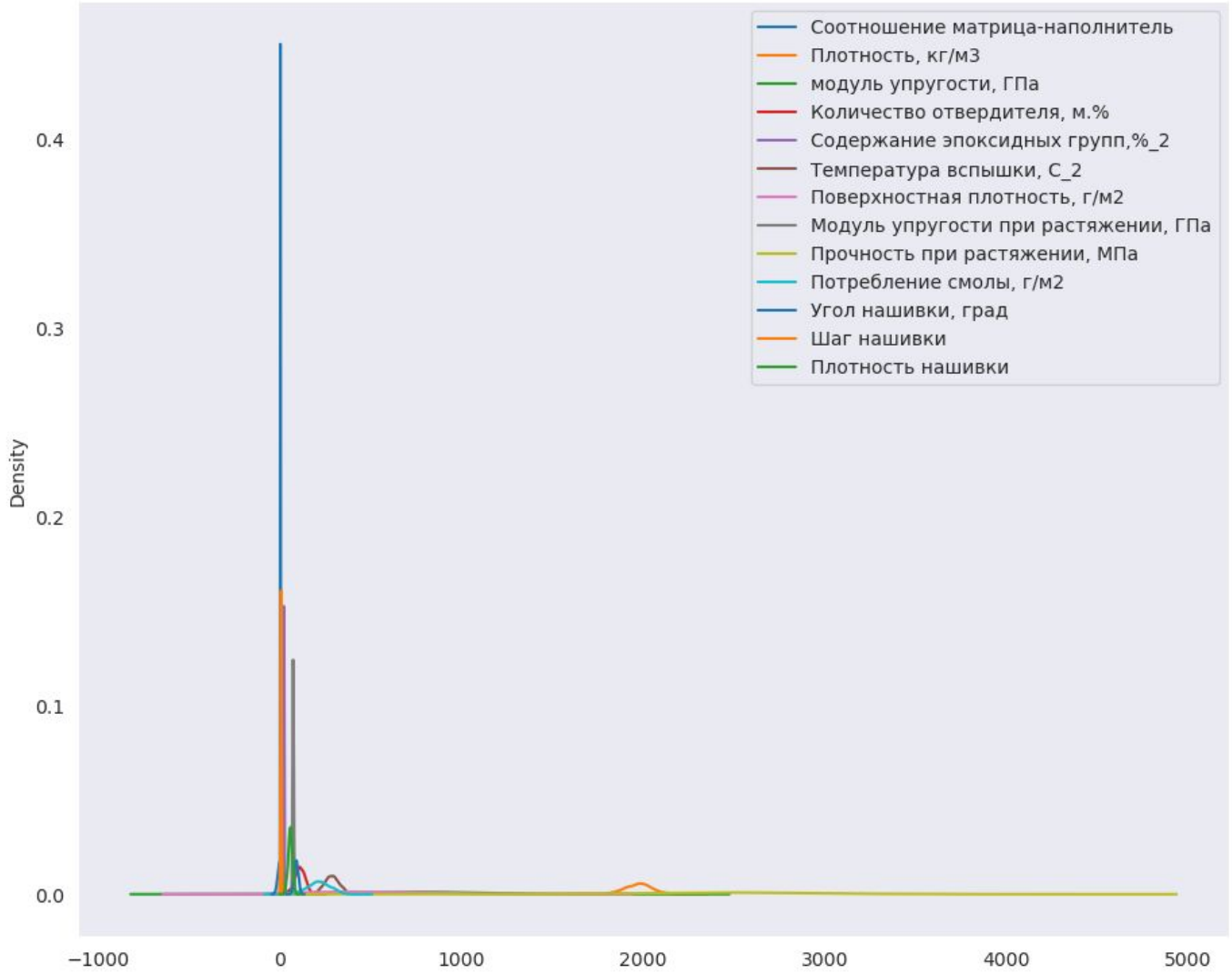
✓ 2.2s

```
Соотношение матрица-наполнитель          0
Плотность, кг/м3                         0
модуль упругости, ГПа                    0
Количество отвердителя, м.%              0
Содержание эпоксидных групп,%_2          0
Температура вспышки, С_2                 0
Поверхностная плотность, г/м2            0
Модуль упругости при растяжении, ГПа     0
Прочность при растяжении, МПа            0
Потребление смолы, г/м2                  0
Угол нашивки, град                       0
Шаг нашивки                              0
Плотность нашивки                        0
dtype: int64
```

Попарные графики рассеяния точек



Оценка плотности ядра



Синтетические данные

Разведочный анализ данных показал, что линейной связи между любыми переменными нет, корреляция равна 0, данных для выборки недостаточно, не все признаки имеют одинаковое распределение, значения необходимые для прогнозирования, редки и необычны т.е датасет не сбалансирован, то для дальнейшей задачи было решено создать 3 новые датасета для каждой модели и добавить синтетические данные, для этого была использована библиотека ImbalancedLearningRegression.

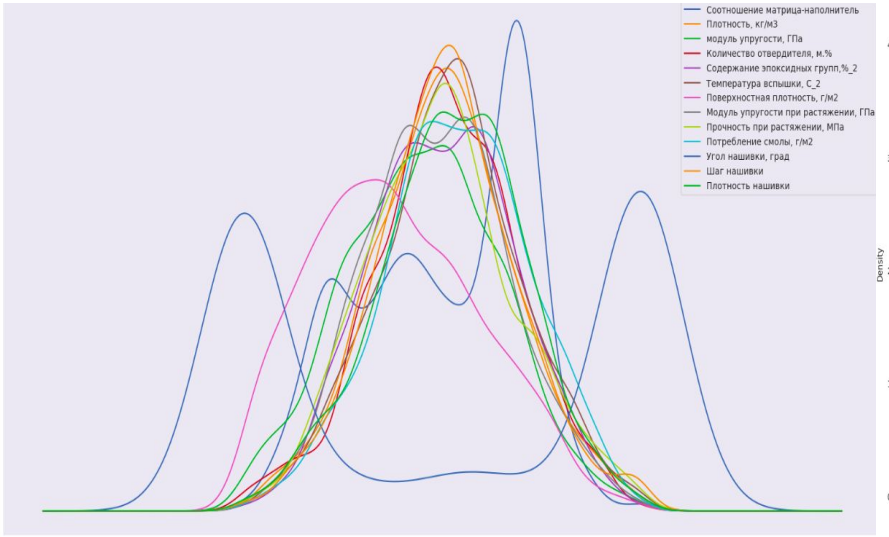
Описательная статистика после добавления синтетических данных

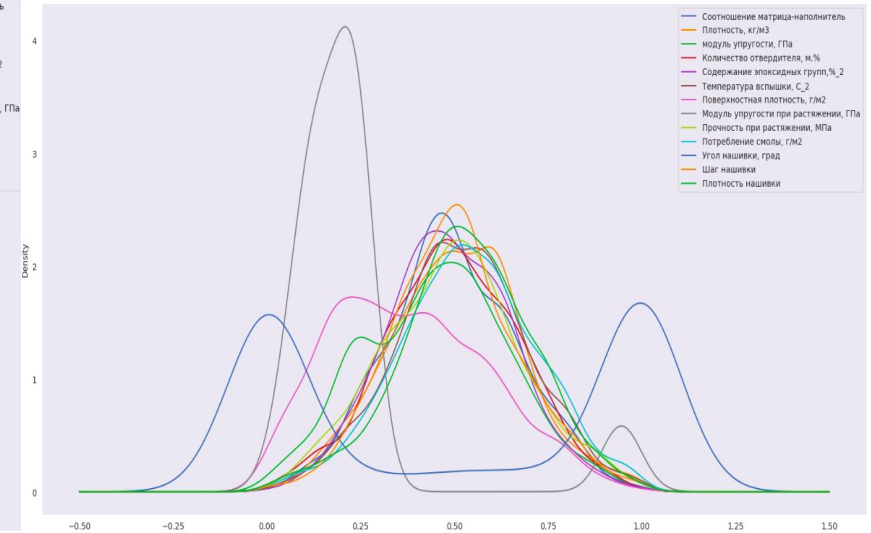
	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1573.0	0.493922	0.173935	0.0	0.382441	0.488209	0.616360	1.0
Плотность, кг/м3	1573.0	0.504886	0.177452	0.0	0.386820	0.508281	0.622042	1.0
модуль упругости, ГПа	1573.0	0.452948	0.192658	0.0	0.303617	0.459736	0.579825	1.0
Количество отвердителя, м.%	1573.0	0.494694	0.178327	0.0	0.370333	0.492540	0.624242	1.0
Содержание эпоксидных групп,%_2	1573.0	0.487531	0.166470	0.0	0.374410	0.484606	0.600314	1.0
Температура вспышки, С_2	1573.0	0.519020	0.177392	0.0	0.404402	0.520172	0.637763	1.0
Поверхностная плотность, г/м2	1573.0	0.380612	0.208636	0.0	0.213848	0.363665	0.536967	1.0
Модуль упругости при растяжении, ГПа	1573.0	0.235416	0.216609	0.0	0.126249	0.194135	0.240811	1.0
Прочность при растяжении, МПа	1573.0	0.492951	0.186782	0.0	0.364430	0.493830	0.611272	1.0
Потребление смолы, г/м2	1573.0	0.535640	0.184070	0.0	0.412876	0.531335	0.653760	1.0
Угол нашивки, град	1573.0	0.515128	0.470509	0.0	0.000000	0.579606	1.000000	1.0
Шаг нашивки	1573.0	0.501566	0.170717	0.0	0.384770	0.498796	0.604767	1.0
Плотность нашивки	1573.0	0.532122	0.177110	0.0	0.426971	0.530633	0.650741	1.0

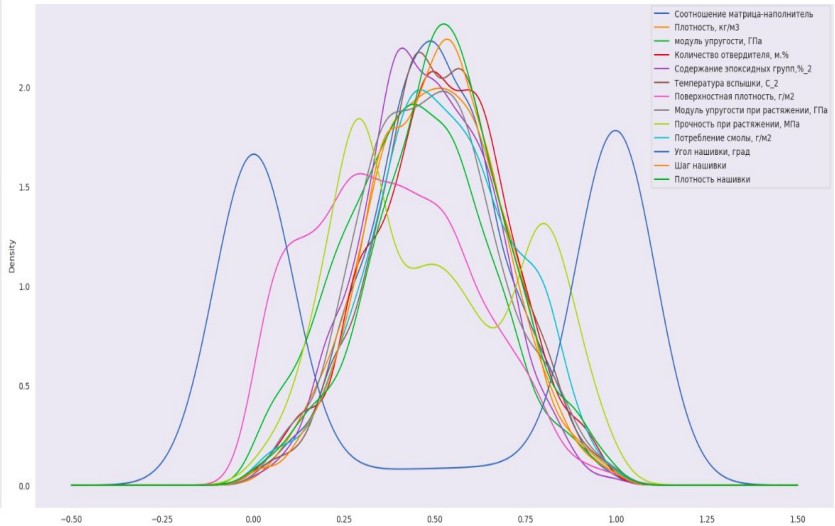
	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1773.0	0.503662	0.179237	0.0	0.388361	0.499814	0.622300	1.0
Плотность, кг/м3	1773.0	0.501413	0.181471	0.0	0.376312	0.501605	0.627761	1.0
модуль упругости, ГПа	1773.0	0.444842	0.198246	0.0	0.304218	0.444649	0.578896	1.0
Количество отвердителя, м.%	1773.0	0.514913	0.186587	0.0	0.391089	0.517263	0.643876	1.0
Содержание эпоксидных групп,%_2	1773.0	0.483747	0.172255	0.0	0.368122	0.484385	0.607554	1.0
Температура вспышки, С_2	1773.0	0.517517	0.179134	0.0	0.401684	0.515669	0.635493	1.0
Поверхностная плотность, г/м2	1773.0	0.380515	0.220123	0.0	0.206505	0.372368	0.539446	1.0
Модуль упругости при растяжении, ГПа	1773.0	0.490430	0.188307	0.0	0.354910	0.487331	0.613795	1.0
Прочность при растяжении, МПа	1773.0	0.500186	0.247821	0.0	0.289613	0.474518	0.749927	1.0
Потребление смолы, г/м2	1773.0	0.524325	0.192831	0.0	0.393951	0.521943	0.657091	1.0
Угол нашивки, град	1773.0	0.517423	0.485734	0.0	0.000000	0.707592	1.000000	1.0
Шаг нашивки	1773.0	0.506134	0.175218	0.0	0.377322	0.508925	0.629849	1.0
Плотность нашивки	1773.0	0.519450	0.184929	0.0	0.398987	0.522089	0.637637	1.0

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1666.0	0.501518	0.205900	0.0	0.335759	0.528891	0.678263	1.0
Плотность, кг/м3	1666.0	0.503850	0.177630	0.0	0.387056	0.505023	0.617055	1.0
модуль упругости, ГПа	1666.0	0.451763	0.192562	0.0	0.311698	0.451826	0.583474	1.0
Количество отвердителя, м.%	1666.0	0.508476	0.177909	0.0	0.401877	0.506532	0.630983	1.0
Содержание эпоксидных групп,%_2	1666.0	0.502778	0.177589	0.0	0.374410	0.500302	0.628215	1.0
Температура вспышки, С_2	1666.0	0.519269	0.179822	0.0	0.405498	0.524911	0.634265	1.0
Поверхностная плотность, г/м2	1666.0	0.382778	0.210044	0.0	0.221671	0.363657	0.534390	1.0
Модуль упругости при растяжении, ГПа	1666.0	0.489307	0.178169	0.0	0.367333	0.490425	0.606509	1.0
Прочность при растяжении, МПа	1666.0	0.498015	0.186914	0.0	0.369925	0.492663	0.609636	1.0
Потребление смолы, г/м2	1666.0	0.537340	0.180542	0.0	0.420546	0.534846	0.656168	1.0
Угол нашивки, град	1666.0	0.514660	0.470096	0.0	0.000000	0.569198	1.000000	1.0
Шаг нашивки	1666.0	0.511907	0.176409	0.0	0.397572	0.509787	0.625333	1.0
Плотность нашивки	1666.0	0.524474	0.180563	0.0	0.404829	0.530269	0.644932	1.0

Плотность ядра после нормализации





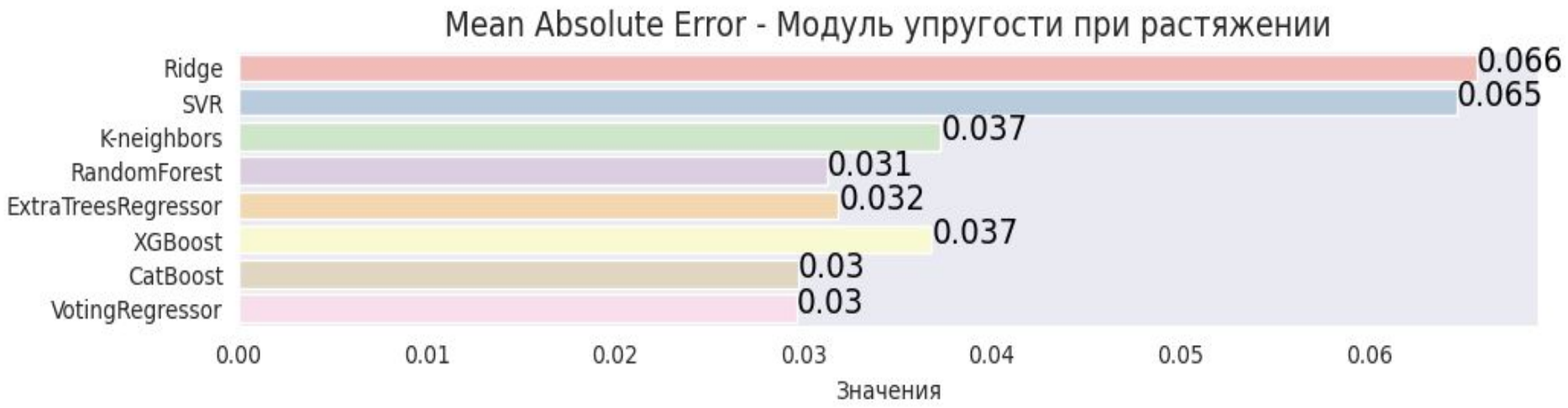


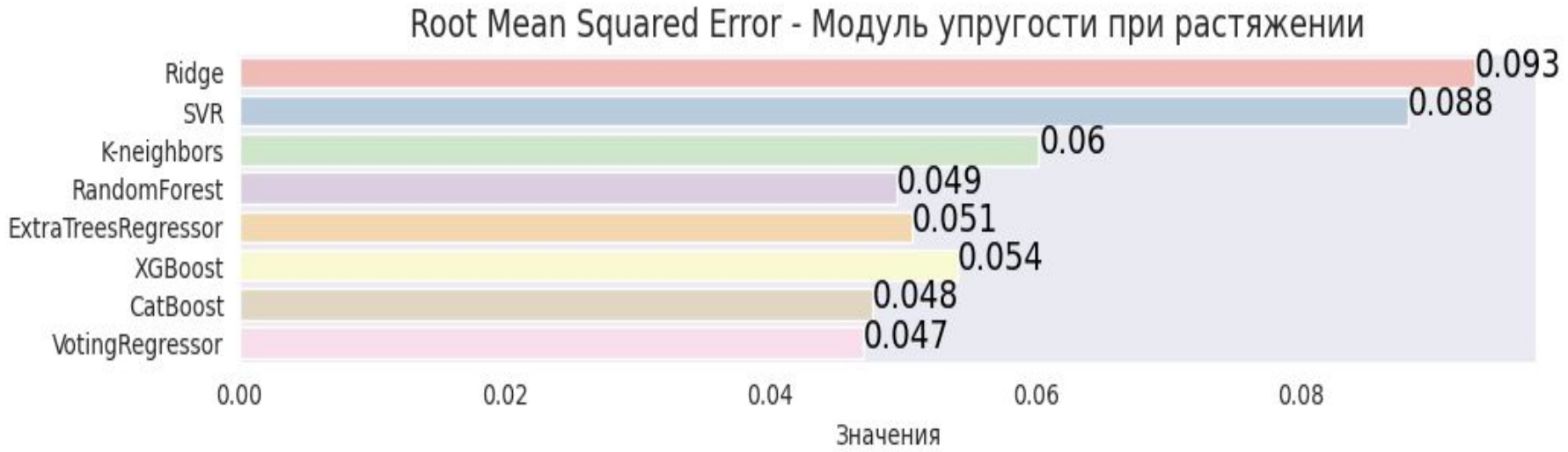
Этапы разработки и обучения моделей

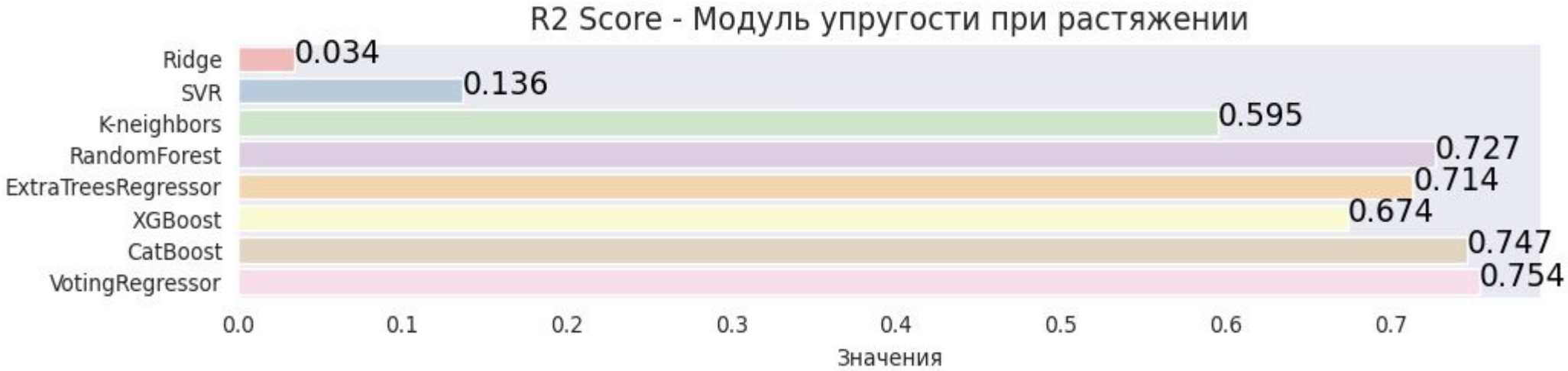
- Порядок разработки моделей для каждого параметра и для каждого выбранного метода соответствуют следующие этапы
- Разделение на обучающую тестовую выборки (соотношение 70% на 30%)
- Задание сетки гиперпараметров для оптимизации моделей с кросс-валидацией
- Подстановка гиперпараметров в модель и обучение на тренировочных данных
- Настройка нейронной сети , подбор количества нейронов,слоев, функции активации и настройка оптимизатора
- Оценка качества модели с использование метрик для регрессии и визуализация
- Сохранение модели для веб-приложения

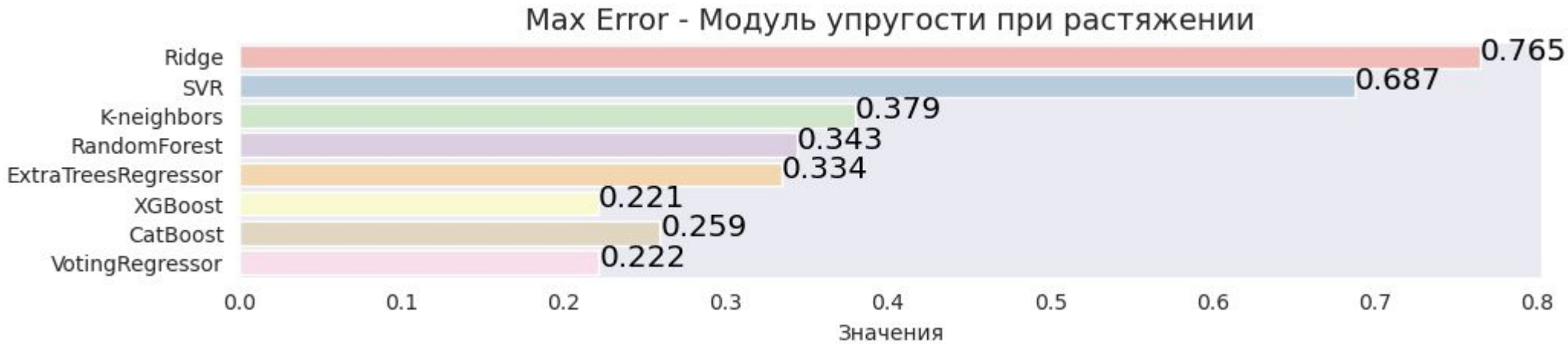
Этапы разработки и обучения моделей

Результаты моделей параметра «модуль упругости при растяжении, гпа %»

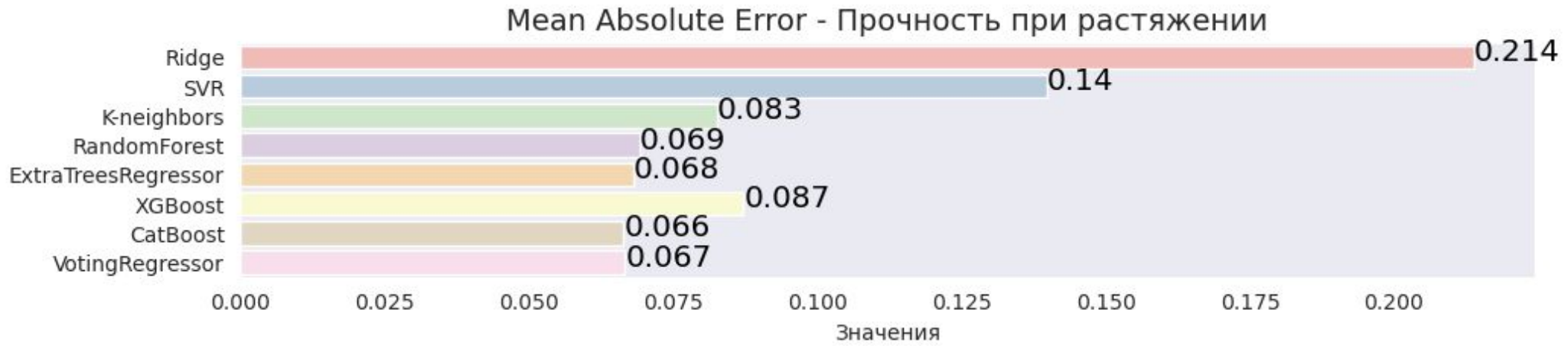


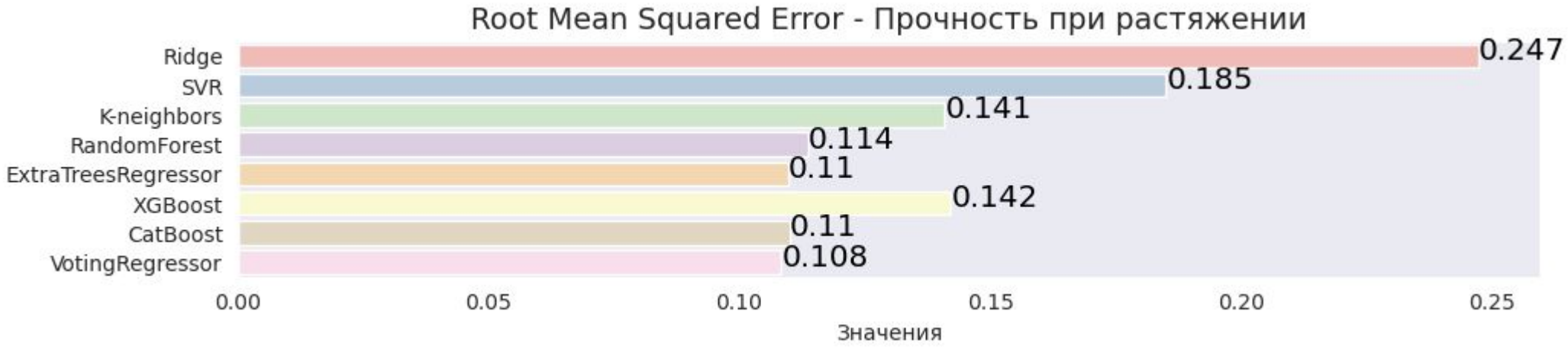


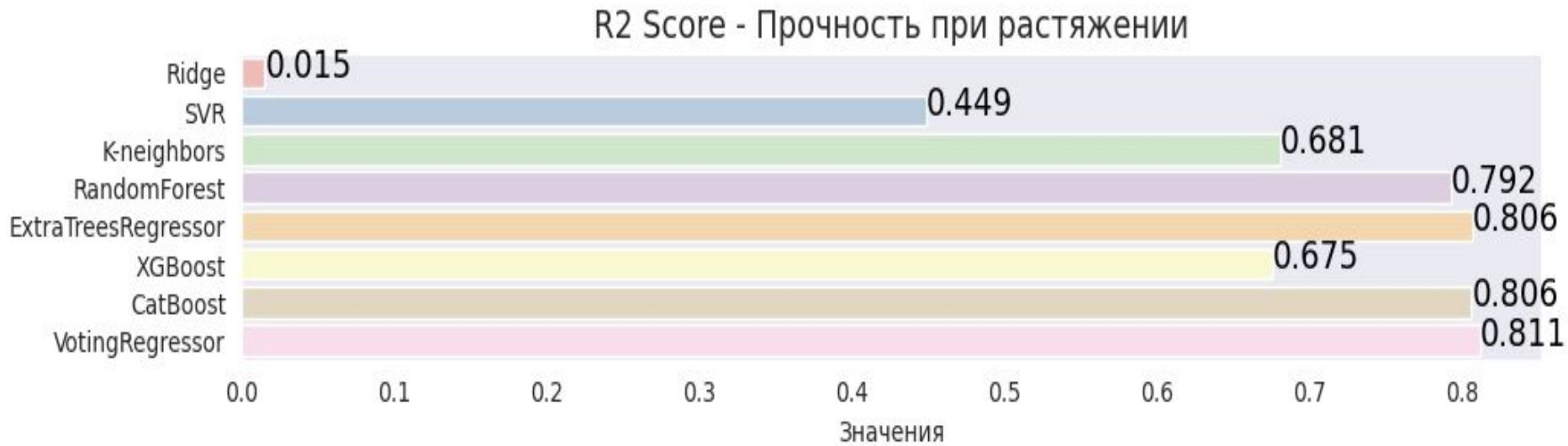

R2 Score - Модуль упругости при растяжении
Ridge
SVR
K-neighbors
RandomForest
ExtraTreesRegressor
XGBoost
CatBoost
VotingRegressor
0.034
0.136
0.595
0.727
0.714
0.674
0.747
0.754
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Значения

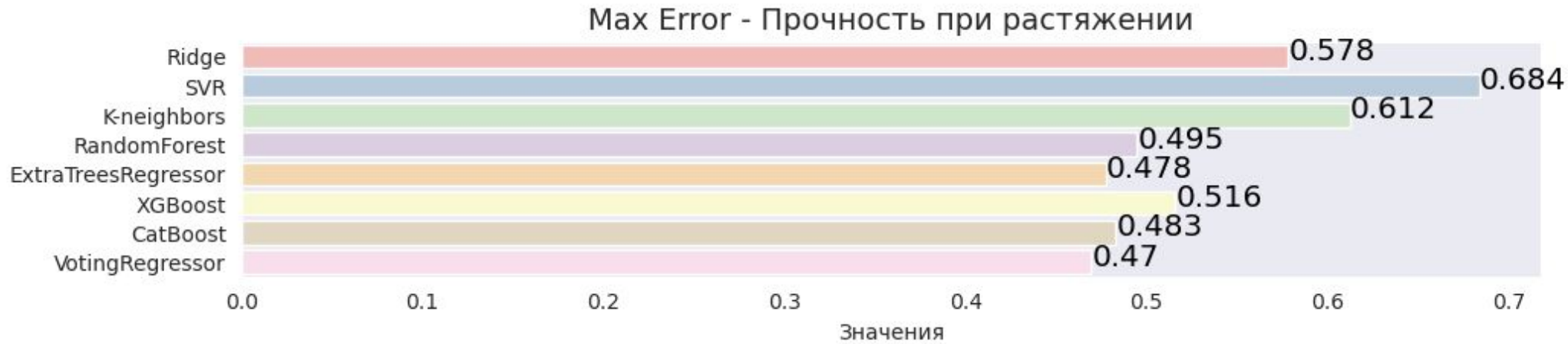

Max Error - Модуль упругости при растяжении
Ridge
SVR
K-neighbors
RandomForest
ExtraTreesRegressor
XGBoost
CatBoost
VotingRegressor
0.765
0.687
0.379
0.343
0.334
0.221
0.259
0.222
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Значения

Результаты моделей прогноза параметра «Прочность при растяжении, МПа %»

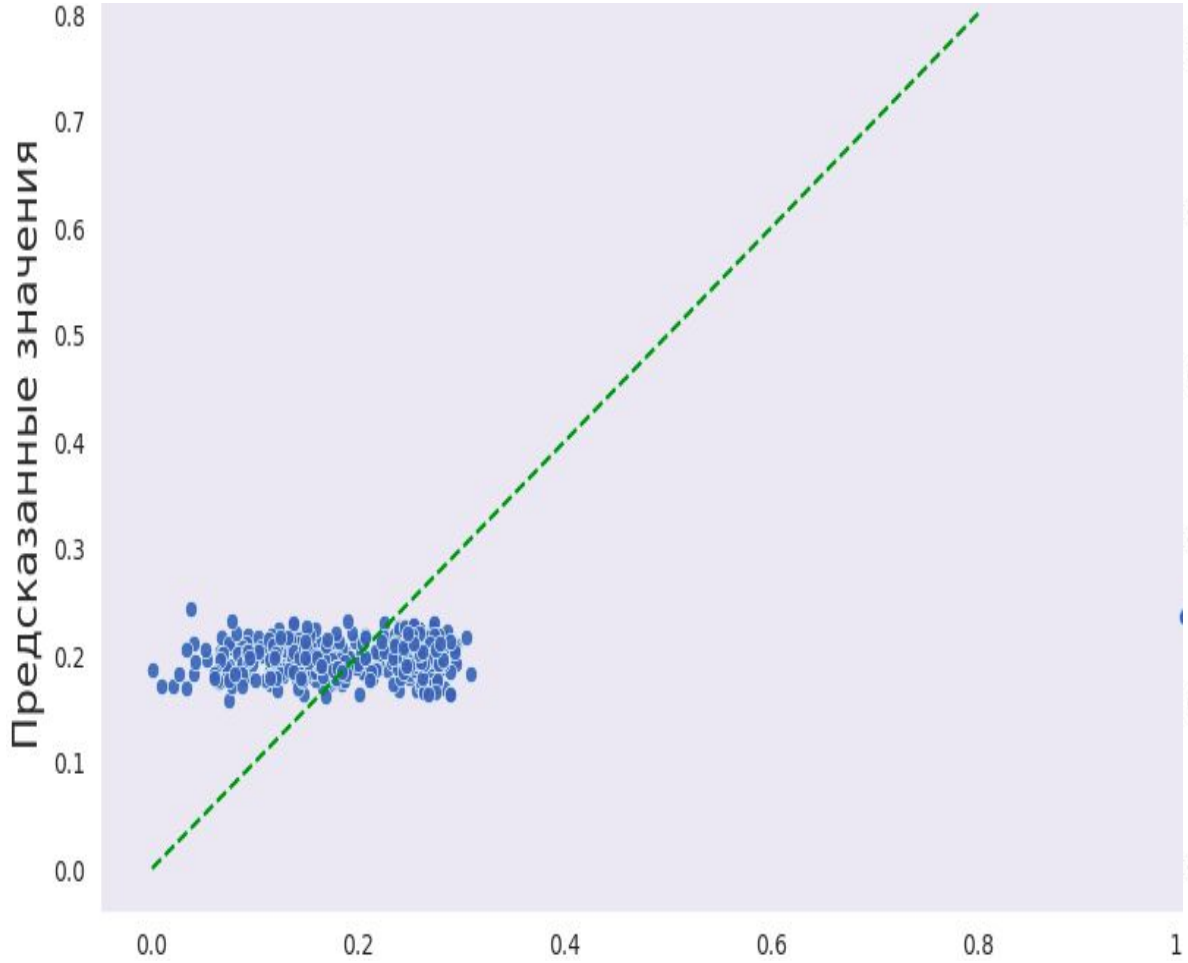


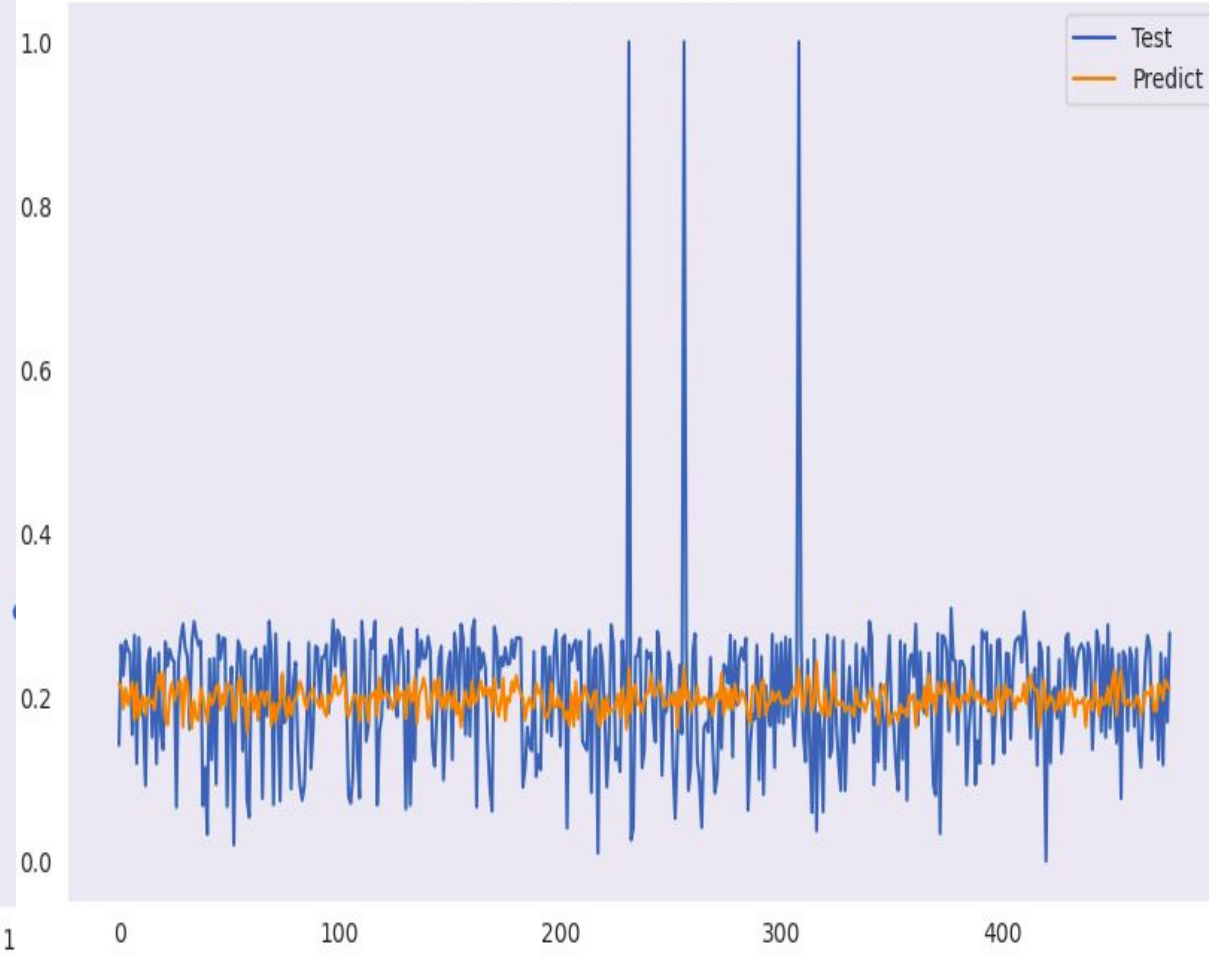



R2 Score - Прочность при растяжении
Ridge
SVR
K-neighbors
RandomForest
ExtraTreesRegressor
XGBoost
CatBoost
VotingRegressor
0.015
0.449
0.681
0.792
0.806
0.675
0.806
0.811
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Значения

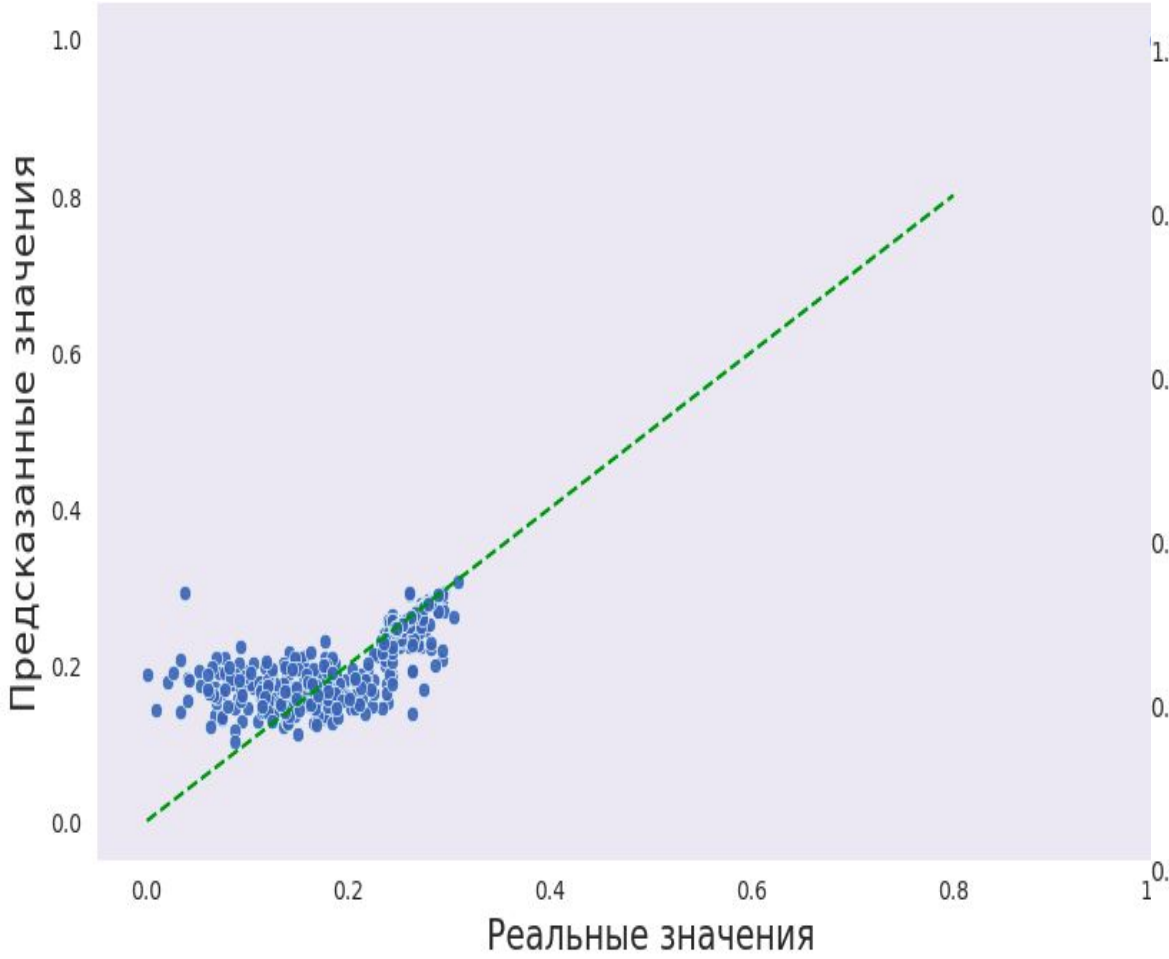

Max Error - Прочность при растяжении
Ridge
SVR
K-neighbors
RandomForest
ExtraTreesRegressor
XGBoost
CatBoost
VotingRegressor
0.578
0.684
0.612
0.495
0.478
0.516
0.483
0.47
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Значения

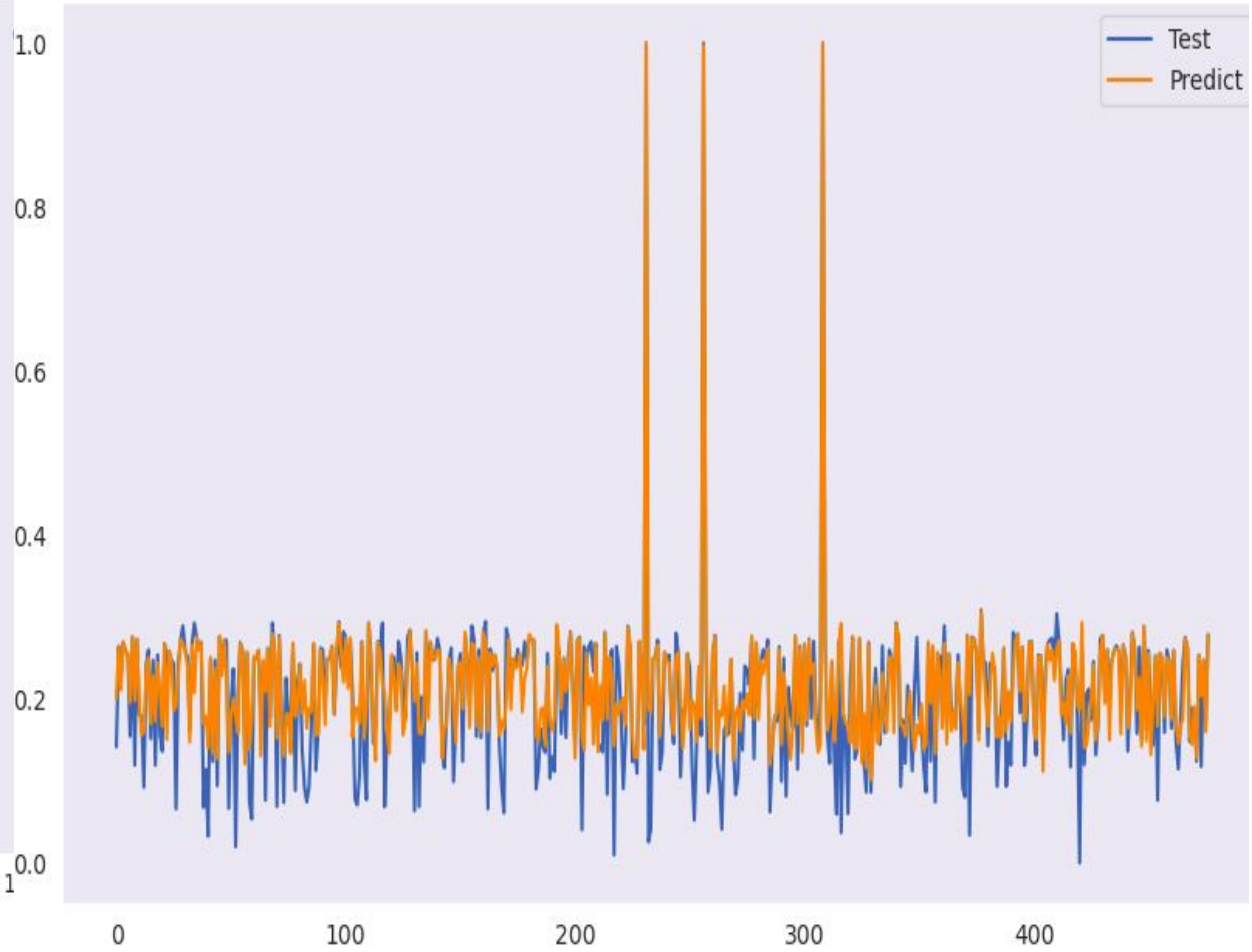
Визуализация худшей модели





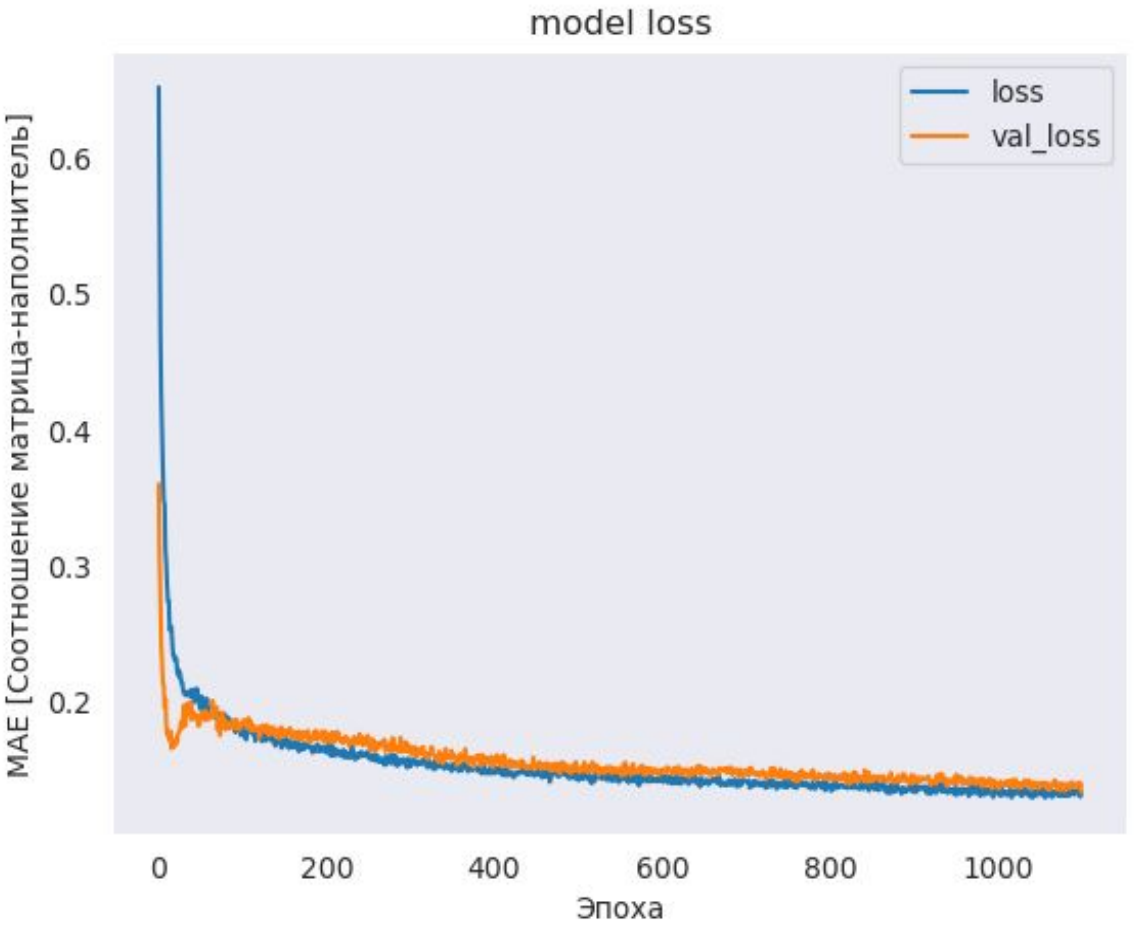
Визуализация лучшей модели



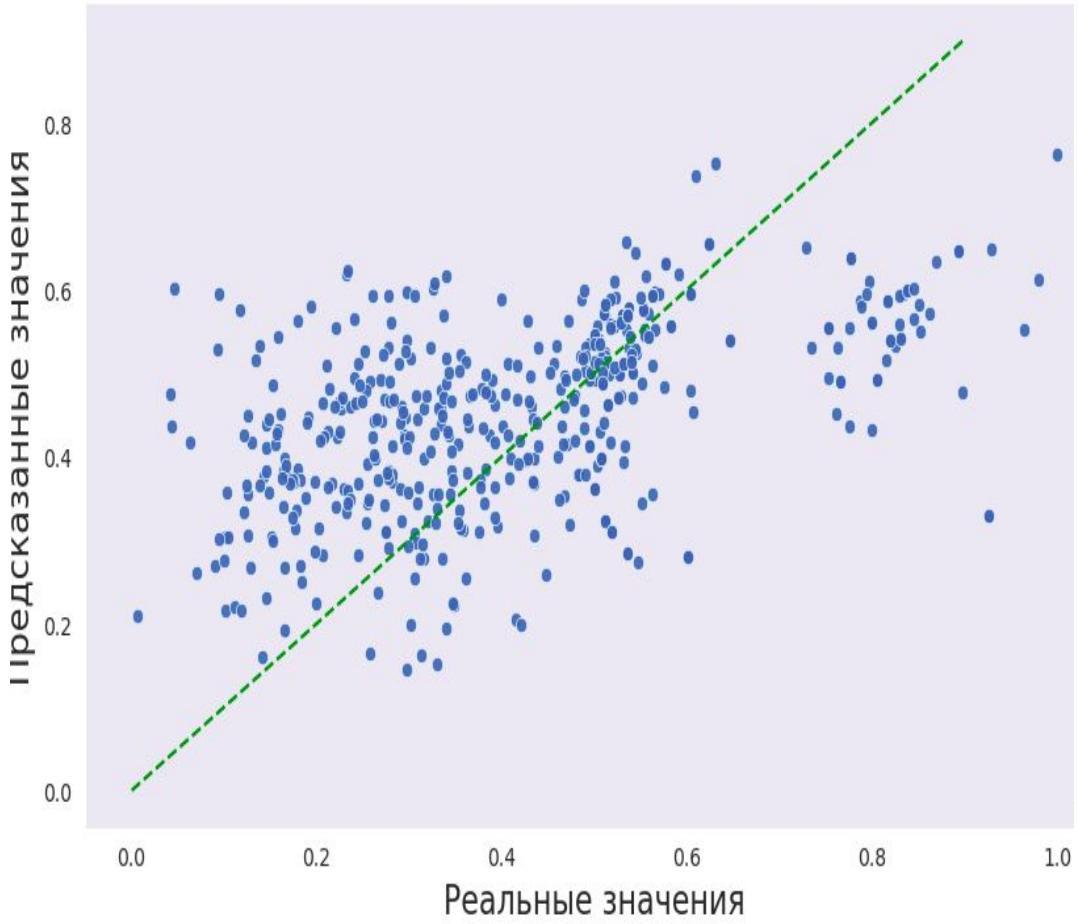


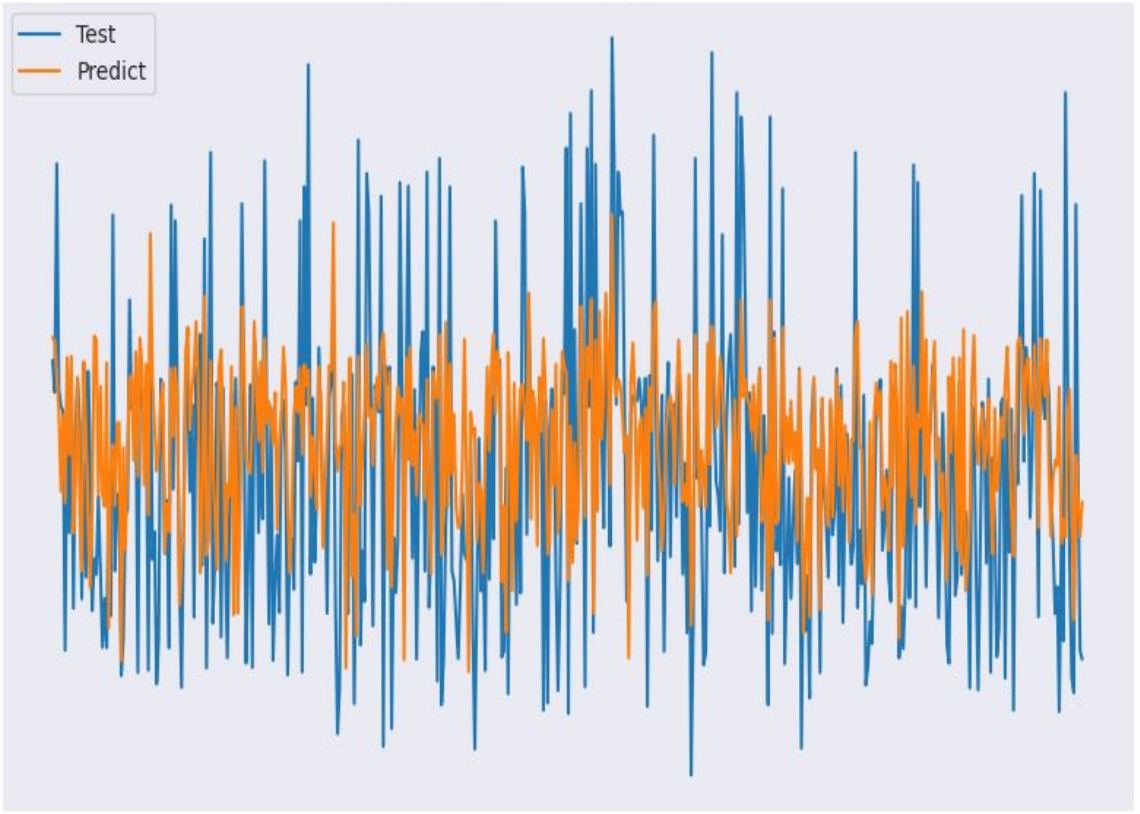
Результаты модели нейронной сети “Сооотношение матрица-наполнитель”

```
Epoch 976/1100
2/2 [==============================] - 0s 20ms/step - loss: 0.1331 - val_loss: 0.1398
Epoch 977/1100
2/2 [==============================] - 0s 20ms/step - loss: 0.1322 - val_loss: 0.1405
Epoch 978/1100
2/2 [==============================] - 0s 21ms/step - loss: 0.1382 - val_loss: 0.1412
Epoch 979/1100
2/2 [==============================] - 0s 19ms/step - loss: 0.1321 - val_loss: 0.1424
Epoch 980/1100
2/2 [==============================] - 0s 19ms/step - loss: 0.1308 - val_loss: 0.1437
Epoch 981/1100
2/2 [==============================] - 0s 19ms/step - loss: 0.1357 - val_loss: 0.1412
Epoch 982/1100
2/2 [==============================] - 0s 20ms/step - loss: 0.1339 - val_loss: 0.1377
Epoch 983/1100
2/2 [==============================] - 0s 20ms/step - loss: 0.1349 - val_loss: 0.1417
Epoch 984/1100
2/2 [==============================] - 0s 20ms/step - loss: 0.1335 - val_loss: 0.1431
Epoch 985/1100
2/2 [==============================] - 0s 19ms/step - loss: 0.1317 - val_loss: 0.1401
Epoch 986/1100
2/2 [==============================] - 0s 22ms/step - loss: 0.1366 - val_loss: 0.1385
...
Epoch 1099/1100
2/2 [==============================] - 0s 25ms/step - loss: 0.1303 - val_loss: 0.1405
Epoch 1100/1100
2/2 [==============================] - 0s 20ms/step - loss: 0.1342 - val_loss: 0.1334
```



Визуализация модели





Разработка приложения

Разработано веб-приложение которое прогнозирует конечные свойства композиционных материалов на основе введенных пользователем значений. Ансамбль машинного обучения предсказывает **"Прочность при растяжении" и "Модуль упругости при растяжении"**, а нейронная сеть **"Соотношение матрица-наполнитель"**

Прогнозирование конечных свойств композиционных материалов

Это веб-приложение прогнозирует конечные свойства композиционных материалов на основе введенных пользователем значений. Ансамбль машинного обучения предсказывает **"Прочность при растяжении"** и **"Модуль упругости при растяжении"**, а нейронная сеть **"Соотношение матрица-наполнитель"**.

Прогнозирование значения матрица-наполнитель

Прогнозирование прочности при растяжении

Прогнозирование модуля упругости при растяжении

Чтобы выбрать нужную модель, нажмите на одну из кнопок расположенных справа.

Репозиторий GitHub

Deploy модели

Предсказание модели соотношения матрица-наполнитель

Введите значения от 0 до 1

Введите Плотность, кг/м3

Введите Модуль упругости, ГПа

Введите Количество отвердителя, м.%

Введите Содержание эпоксидных групп,%_2

Введите Температура вспышки, С_2

Введите Поверхностная плотность, г/м2

Введите Модуль упругости при растяжении, ГПа

Введите Прочность при растяжении, МПа

Введите Потребление смолы, г/м2

Введите Угол нашивки, град

Введите Шаг нашивки

Введите Плотность нашивки

Предсказать Обновить

Итоги

В процессе проделанной работы было выяснено:

- Создание и добавление синтетических данных значительно улучшили работу некоторых моделей
- Для подбора гиперпараметров для ансамблевых моделей лучше использовать специализированные библиотеки: Optuna, Hyperopt
- Лучший результат для моделей “Модуль упругости при растяжении, МПа” и “Прочность при растяжении, ГПф” показала библиотека CatBoost
- “Суперансамбль” VotingRegressor в комбинации из CatBoostRegressor и RandomForestRegressor еще улучшил результат метрик.
- Для нейронной сети экспериментальным путем было определено, что лучшей функцией активации активацией это гиперболический тангенс, а оптимизатор rmsprop и что не следует добавлять большое число скрытых слоев, во избежание переобучения модели.

Мой репозиторий

- Страница создана на GitHub
- Адрес страницы: https://github.com/qalansiyah/my_vkr
- Деплой веб-приложения: https://vkr-deploy.onrender.com/
- В репозитории находятся:

✔ Файлы Jupyter Notebook
✔ Набор данных, модель
✔ Веб-приложение
✔ ВКР в тестовом формате

Спасибо за внимание