ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science»

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель:

Цели и задачи

Цель исследования состоит в прогнозировании ряда конечных свойств получаемых композиционных материалов на основе имеющихся входящих параметров.

Задачи исследования:

- Изучить теоретические основы и методы решения поставленной задачи;
- Провести разведочный анализ данных;
- Провести предобработку данных;
- Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении;
- Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель;
- Разработать приложение;
- Оценить точность модели на тренировочном и тестовом датасете;
- Создать удаленный репозиторий и разместить там код исследования. Оформить файл readme.

Разведочный анализ данных

Объединённый датасет по индексу и типу объединения INNER состоит из 13 столбцов и 1023 строк.

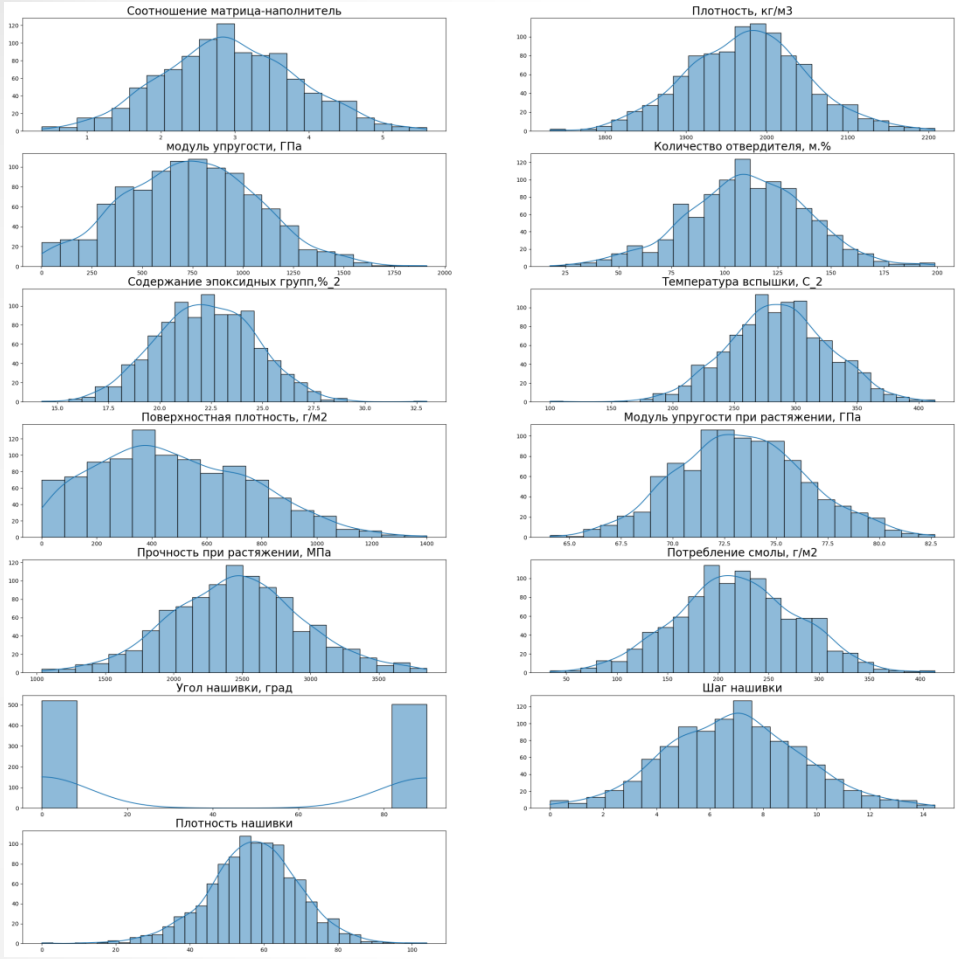
- Дубликаты и пропущенные значения в датасете отсутствуют.
- В основном в каждом столбце содержатся только уникальные значения, кроме столбца “Угол нашивки” всего 2 значения.

Целевые переменные:
- Модуль упругости при растяжении, Гпа
- Прочность при растяжении, МПа
- Соотношение матрица-наполнитель

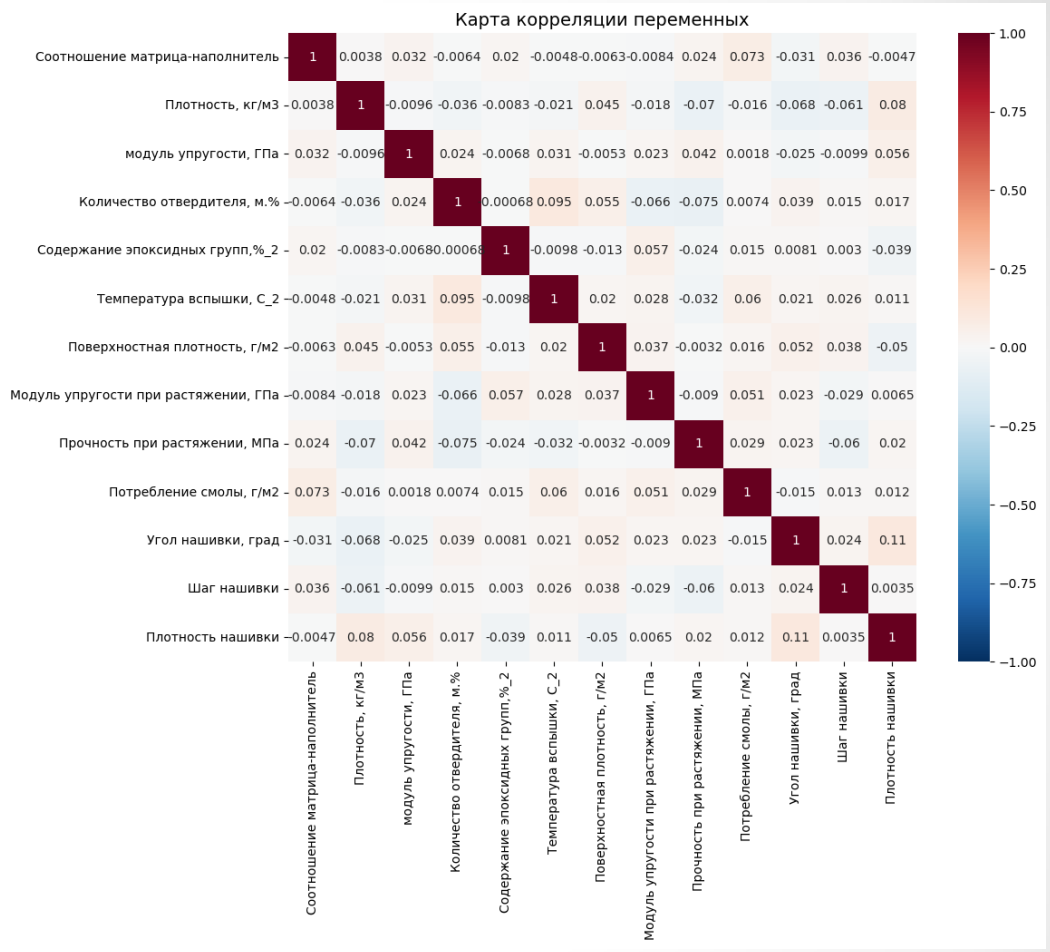
```
# Выведем описательную статистику
data.describe().round(2).T
```

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.93	0.91	0.39	2.32	2.91	3.55	5.59
Плотность, кг/м3	1023.0	1975.73	73.73	1731.76	1924.16	1977.62	2021.37	2207.77
модуль упругости, ГПа	1023.0	739.92	330.23	2.44	500.05	739.66	961.81	1911.54
Количество отвердителя, м.%	1023.0	110.57	28.30	17.74	92.44	110.56	129.73	198.95
Содержание эпоксидных групп,%_2	1023.0	22.24	2.41	14.25	20.61	22.23	23.96	33.00
Температура вспышки, С_2	1023.0	285.88	40.94	100.00	259.07	285.90	313.00	413.27
Поверхностная плотность, г/м2	1023.0	482.73	281.31	0.60	266.82	451.86	693.23	1399.54
Модуль упругости при растяжении, ГПа	1023.0	73.33	3.12	64.05	71.25	73.27	75.36	82.68
Прочность при растяжении, МПа	1023.0	2466.92	485.63	1036.86	2135.85	2459.52	2767.19	3848.44
Потребление смолы, г/м2	1023.0	218.42	59.74	33.80	179.63	219.20	257.48	414.59
Угол нашивки, град	1023.0	44.25	45.02	0.00	0.00	0.00	90.00	90.00
Шаг нашивки	1023.0	6.90	2.56	0.00	5.08	6.92	8.59	14.44
Плотность нашивки	1023.0	57.15	12.35	0.00	49.80	57.34	64.94	103.99

Разведочный анализ данных

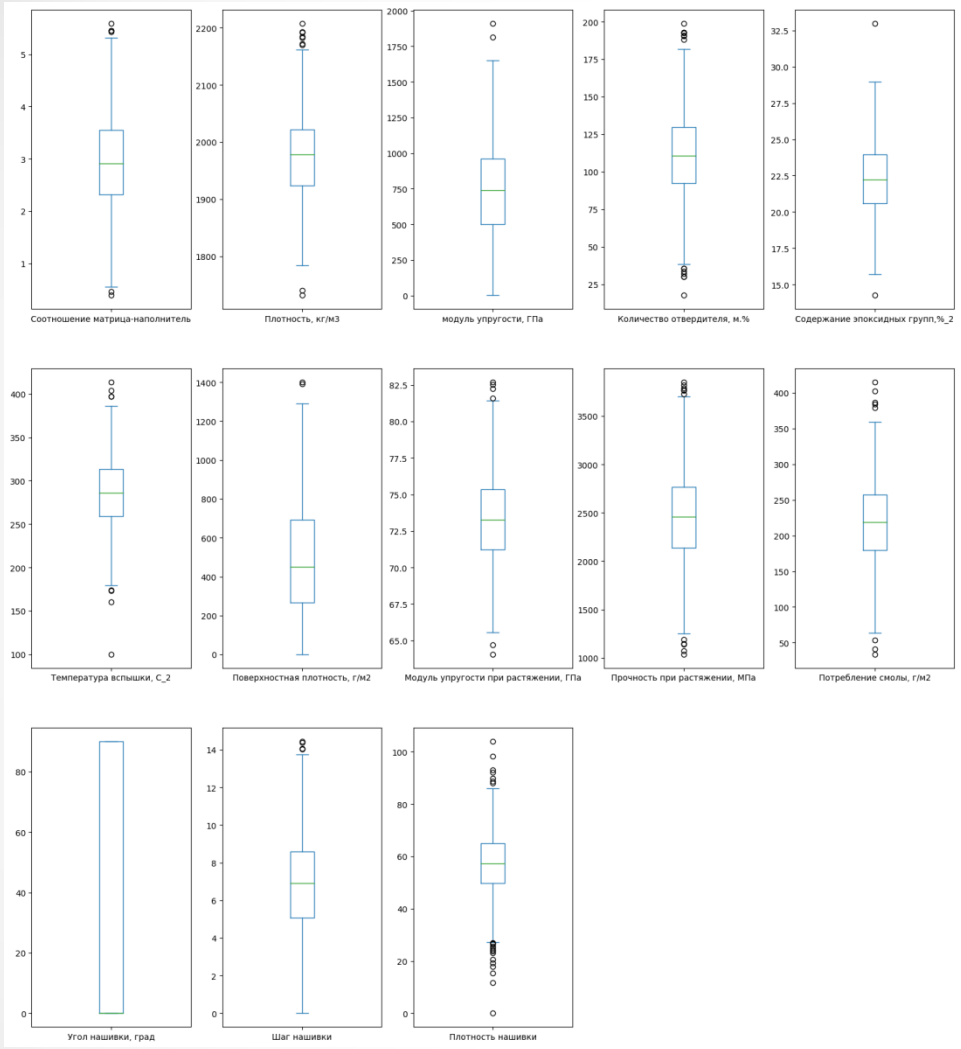


По гистограммам распределения видно, что распределение величин близко к нормальному для большей части переменных, за исключением поверхностной плотности – смещением влево и угла нашивки – дискретная величина, график оказался не показателен.

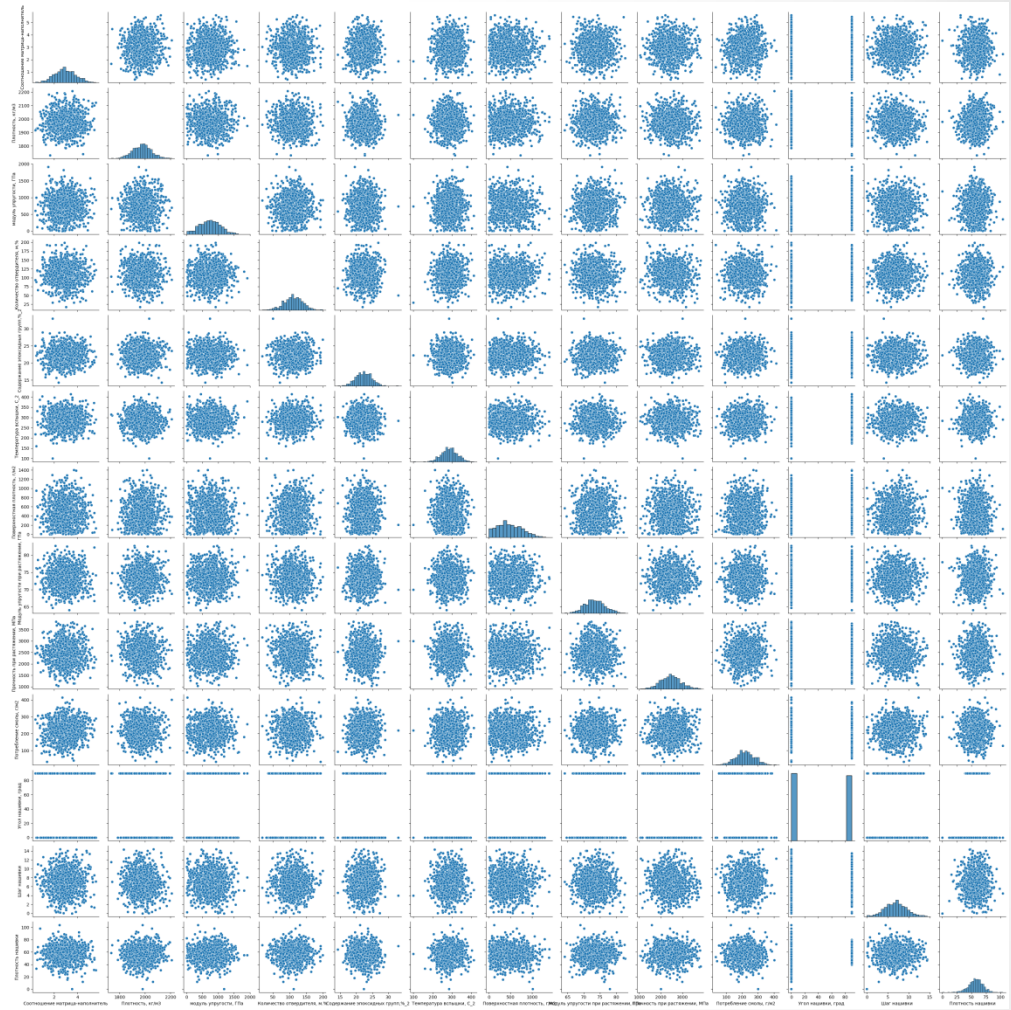


В целях выявления зависимостей между переменными построим тепловую карту коэффициентов корреляции. Коэффициенты корреляции показывают слабую зависимость между переменными.

Разведочный анализ данных

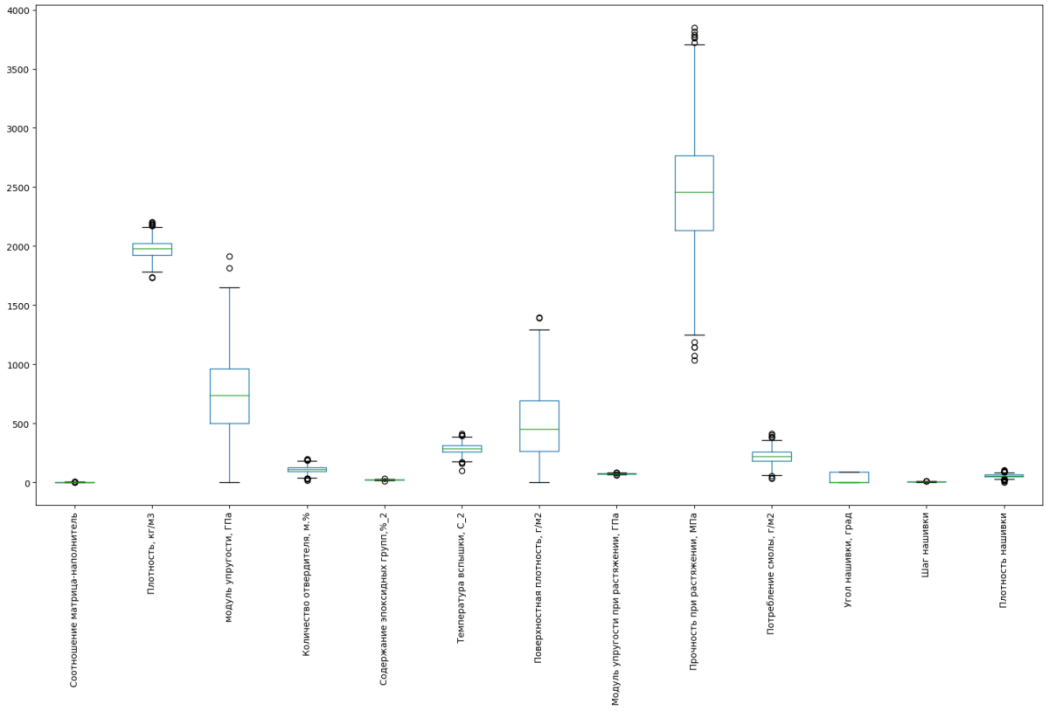


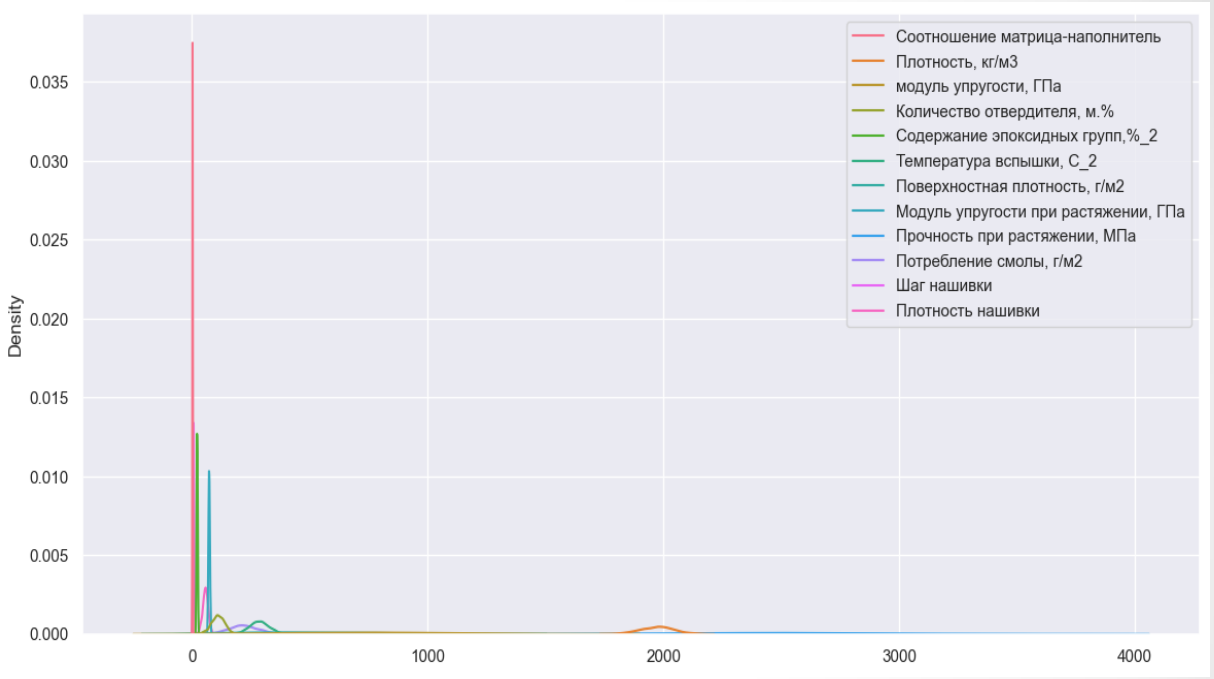
Диаграммы размаха показывают на наличие выбросов во всех признаках кроме “Угол нашивки, град”.



Попарные графики рассеяния также указывают на наличие выбросов и отсутствие каких-либо выраженных зависимостей между переменными.

Разведочный анализ данных



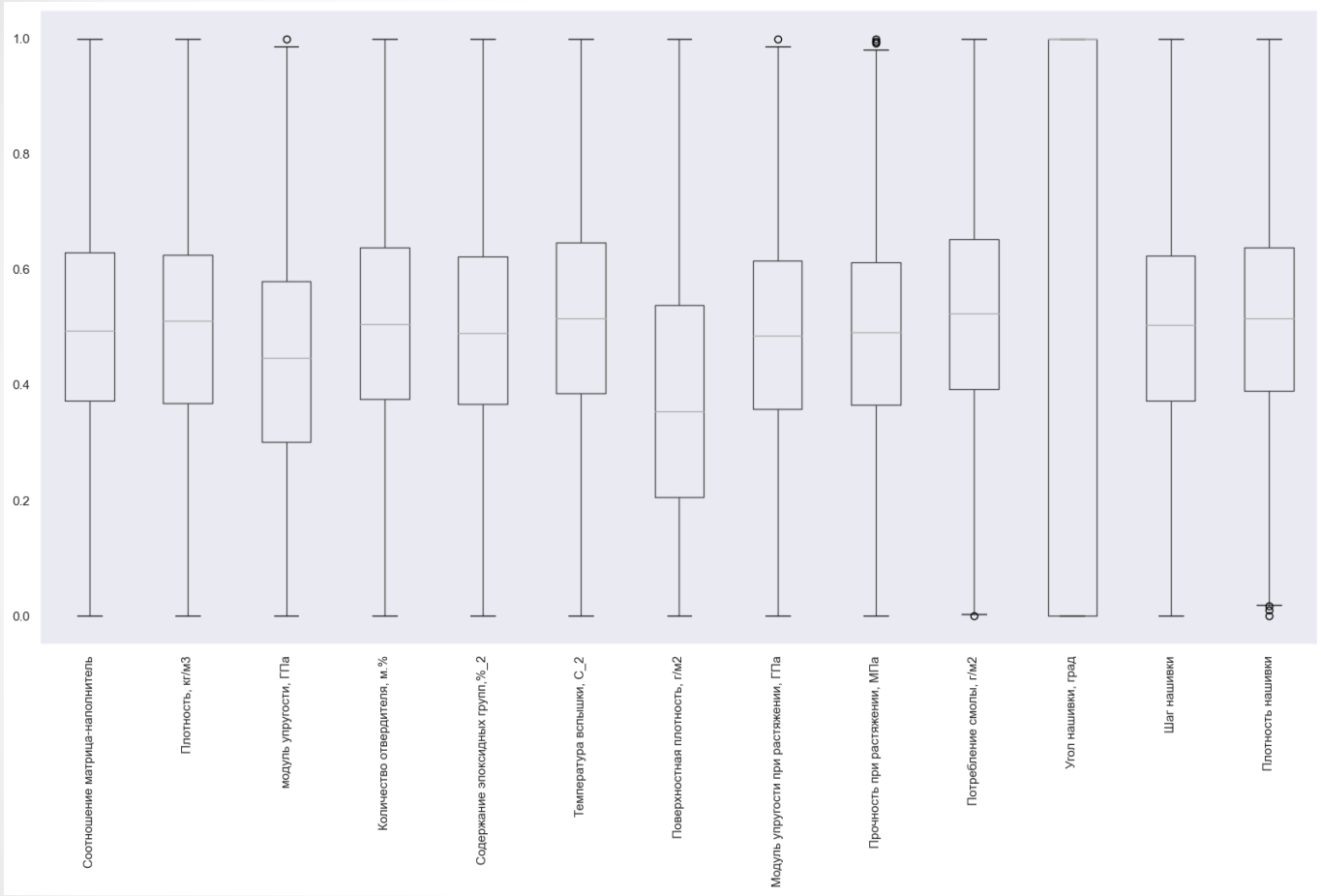


Из диаграммы boxplot («Ящик с усами») видно, что порядок значений переменных различается в разы.

Оценка плотности ядра также показывает, что значения переменных находятся в разных диапазонах поэтому требуется нормализация данных.

Предобработка данных

На стадии предобработки данных произведено удаление выбросов и нормализация данных.



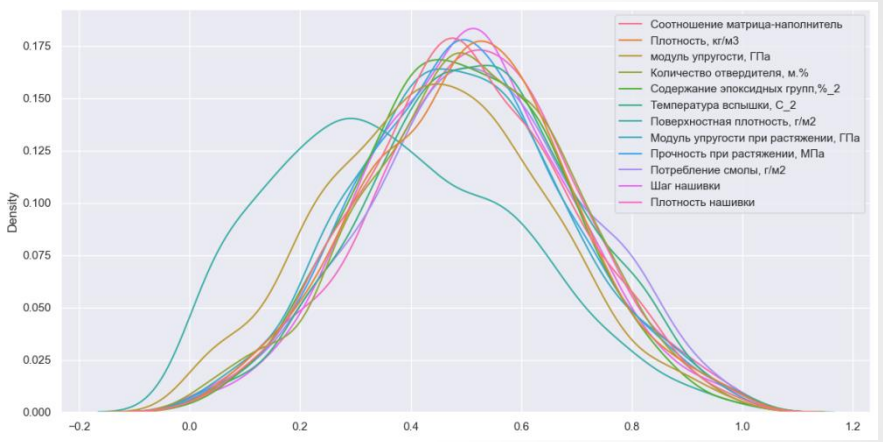


График оценки плотности после нормализации

- В результате самые сильные выбросы удалены, размерность очищенного датасета составляет 936 на 13.
- Нормализация данных выполнена с помощью метода MinMaxScaler.

Разработка, обучение и тестирование моделей

В данной работе применялись наиболее часто используемые методы и модели машинного обучения из библиотеки scikit-learn. При разработке и обучении моделей был проведен поиск оптимальных гиперпараметров моделей с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10, для чего был применен метод GridSearchCV()

Перед обучением моделей датасеты были разделены на обучающую и тестовую выборки, в соответствии с условием задачи 70% на обучение и 30% на тестирование

Модуль упругости при растяжении				
Метод машинного обучения	**MAE**	**MSE**	**R^2**	**R^2 (train)**
LinearRegression (Линейная регрессия)	0.151824	0.034828	-0.005818	0.009843
DecisionTreeRegressor (Регрессионное дерево решений)	0.152394	0.035123	-0.014331	0.004219
RandomForestRegressor (Случайный лес регрессии)	0.150918	0.034644	-0.000508	0.015890
KneighborsRegressor (Метод К ближайших соседей)	0.158387	0.038174	-0.102440	0.070250
SVR (Метод опорных векторов)	0.150919	0.034270	0.010285	0.023628

Прочность при растяжении				
Метод машинного обучения	**MAE**	**MSE**	**R^2**	**R^2 (train)**
LinearRegression (Линейная регрессия)	0.147422	0.035332	-0.043577	0.004432
DecisionTreeRegressor (Регрессионное дерево решений)	0.148127	0.035708	-0.054684	0.024868
RandomForestRegressor (Случайный лес регрессии)	0.147014	0.034962	-0.032641	0.027347
KneighborsRegressor (Метод К ближайших соседей)	0.153163	0.038854	-0.147594	0.999999
SVR (Метод опорных векторов)	0.147598	0.035166	-0.038648	0.031339

Метрики для оценки качества работы модели:

MAE - измеряет среднюю абсолютную ошибку прогнозов. Для каждой точки вычисляется разница между прогнозами и целью, а затем усредняются эти значения.

MSE - измеряет средний квадрат ошибок прогнозов. Для каждой точки вычисляется квадратная разница между прогнозами и целью, а затем усредняются эти значения.

R^2 - коэффициент детерминации, показывает насколько хорошо регрессионная модель описывает данные. R2 равный 1, означает что функция идеально ложится на все точки – данные идеально описаны моделью. Метрика помогает понять, какую долю данных модель смогла объяснить.

Из таблиц видно, что при предсказаниям не удалось приблизится к идеальному результату более, чем до предсказания среднего значения. Коэффициент детерминации близкий к нулю, а тем более отрицательное его значение свидетельствует о низком качестве модели и отсутствии линейных связей. MAE и MSE показывает высокие показатели, что так же подтверждает отсутствие линейных связей.

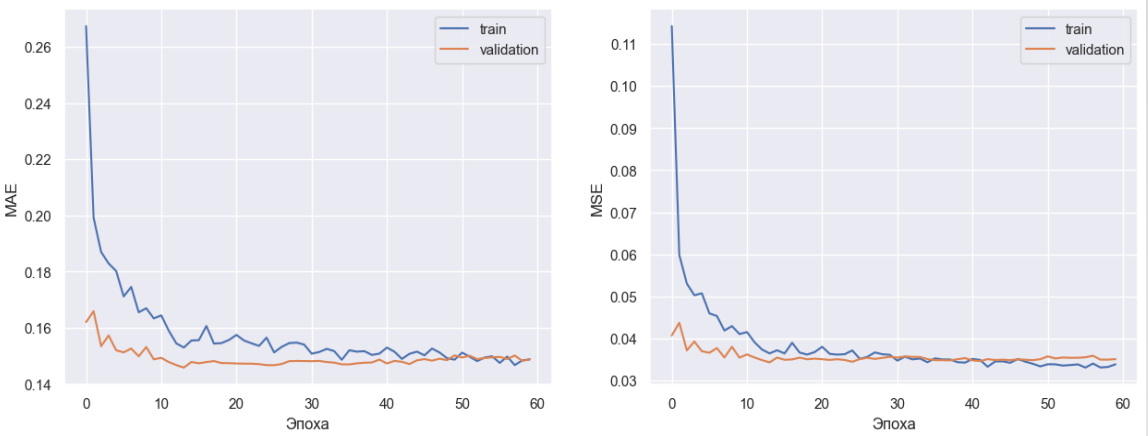
Нейронная сеть для прогнозирования соотношения матрица-наполнитель

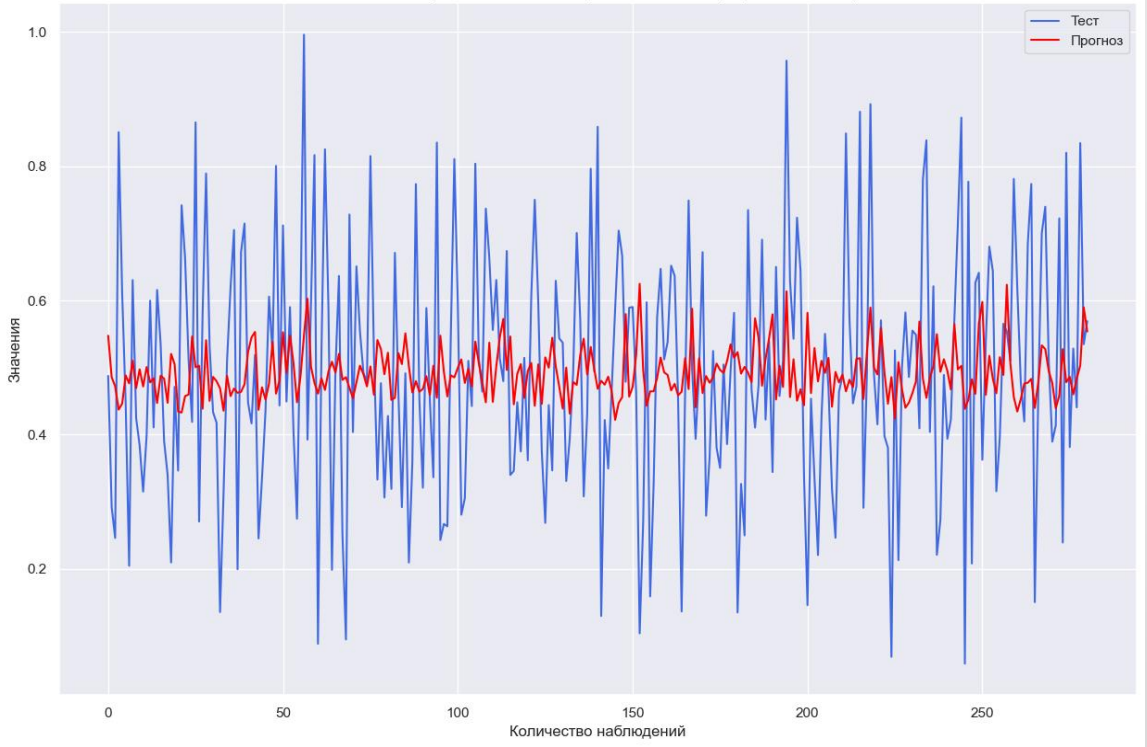
```
# Создадим нейронную сеть для предсказания параметра "Соотношение матрица-наполнитель", добавим прореживание Dropout
# Создадим модель помощью класса Sequential, который предполагает, что выходы одного слоя идут на следующий слой
model2 = Sequential()
# Добавим входной слой
model2.add(Dense(64, input_dim=X_matr_train.shape[1], activation='relu'))
model2.add(Dropout(0.12))
model2.add(Dense(32, activation='relu'))
model2.add(Dropout(0.12))
model2.add(Dense(16, activation='relu'))
model2.add(Dropout(0.12))
model2.add(Dense(8, activation='tanh'))
# Добавим выходной слой
model2.add(Dense(1))
# Компилируем модель
model2.compile(loss='mse', optimizer='adam', metrics=['mae'])
# Выведем информацию о модели
model2.summary()
```

```
# Обучим модель, отправляя 20% данных
# на валидационную выборку
history2 = model2.fit(X_matr_train,
                      y_matr_train,
                      epochs=60,
                      batch_size=32,
                      validation_split=0.2,
                      verbose=1)
```

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_8 (Dense)             (None, 64)                832

 dropout (Dropout)           (None, 64)                0

 dense_9 (Dense)             (None, 32)                2080

 dropout_1 (Dropout)         (None, 32)                0

 dense_10 (Dense)            (None, 16)                528

 dropout_2 (Dropout)         (None, 16)                0

 dense_11 (Dense)            (None, 8)                 136

 dense_12 (Dense)            (None, 1)                 9

=================================================================
Total params: 3,585
Trainable params: 3,585
Non-trainable params: 0
_________________________________________________________________
```

Нейронная сеть	MAE	R^2
Последовательная нейросеть с Dropout (Keras)	0.150222	-0.051625





Модель нейронной сети показала неудовлетворительный результат. Коэффициент детерминации, имеет значение близкое к нулю, это говорит о том, что результат использования нейронной сети не точнее использования для прогноза среднего значения прогнозируемого параметра

Спасибо за внимание!