МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э.
Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science»

Слушатель Грицай Александр Николаевич

Москва, 2022

Содержание

Введение

Первым создателем композиционных материалов была сама природа. Множество природных конструкций (стволы деревьев, кости человека и животных, зубы) имеют характерную волокнистую структуру, т. е. это природные волокнистые композиты.

Одним из способов разработки новых композиционных материалов является структурное модифицирование, то есть изменение его физических свойств без изменения химического состава.

Композиционные материалы - материалы, в которых имеет место сочетание двух (или более) химически разнородных компонентов (фаз) с четкой границей раздела между ними. Это неоднородные (гетерогенные) по химическому составу и структуре материалы. Они характеризуются совокупностью свойств, не присущих каждому в отдельности взятому компоненту, входящему в состав данного композита.

Композиционные материалы нужны при производстве деталей для космических аппаратов, атомных станций, спортивного инвентаря (например, легких и прочных велосипедов). Применяются для изготовления элементов приборов и оборудования, эксплуатирующихся в агрессивных средах и при высоких температурах.

В композиционном материале достигается сочетание разнородных компонентов для получения качественно новых свойств. При этом очень важно, что в нем проявляются достоинства составляющих его компонентов, а не их недостатки. В большинстве композитов компоненты можно разделить на матрицу и включённые в нее элементы

Варьируя состав матрицы и наполнителя, их соотношения, ориентацию наполнителя, можно получить материалы с требуемым сочетанием эксплуатационных и технологических свойств. Многие

композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик.

В настоящее время композиционные материалы на полимерных и металлических матрицах находят всё более широкое применение в различных отраслях промышленности в качестве конструкционных материалов. Внедрение обусловлено стремлением использовать их преимущества по сравнению с традиционно используемыми металлами и сплавами. Уникальность композиционных материалов проявляется в их высоких значениях удельной жесткости (отношения модуля упругости к плотности) и удельной прочности (отношения предела прочности к плотности), химической и коррозионной стойкости к агрессивным средам, анизотропии свойств и возможности их варьирования для наилучшего восприятия действующих нагрузок. Внедрение композитных материалов в конструкцию различных агрегатов и узлов позволяет снизить массовые характеристики изделия, увеличить ресурс и срок службы, уменьшить издержки, связанные с обслуживанием композитных конструкций в эксплуатации.

Расширение использования композитов в различных отраслях связано с возможностью реализации таких свойств композитных материалов, как:

Повышенная вибрационная стойкость, что позволяет использовать композитные материалы в зонах действия повышенных вибрационных нагрузок;

Высокий коэффициент затухания волн в композитных материалах, что обеспечивает надежное гашение вибраций, особенно высокого значения декремента затухания колебаний возможно достичь при

применении в вибропоглощающих конструкциях органопластиков – материалов на основе пара- и метаарамидных волокон и полимерных связующих;

Хорошие демпфирующие свойства стеклопластиков, базальтопластиков и органопластиков, что позволяет применять композитные материалы в качестве материала демпферов, защитных кожухов, корпусов и гасителей ударных динамических воздействий на узлы;

Высокие значения шумопоглощения, что позволяет снизить вредное акустическое воздействие на обслуживающий персонал;

Высокие прочностные и жесткостные свойства конструкционных углепластиков, что дает возможность применять композитные материалы в средне- и высоконагруженных узлах и агрегатах;

Химическая и коррозионная стойкость композитных материалов, что позволяет внедрять подобные материалы в эксплуатацию в агрессивных средах;

Высокая усталостная прочность, обусловленная анизотропией свойств композиционных пластиков, в том числе слоистых, что приводит к наличию у композитных материалов высоких коэффициентов трещиностойкости, и как следствие, к высоким параметрам усталостной прочности.

Современные композиты изготавливаются из материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. Но есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования

заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

Прогнозирование модели может существенно сократить количество проводимых испытаний, а также пополнить базу данных материалов новыми свойствами материалов, и цифровыми двойниками новых композитов. Актуальность решения задачи обусловлена широким использованием композитных материалов практически во всех областях производства.

1. Аналитическая часть

1.1. Постановка задачи

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов).

Предметом выпускной квалификационной работы являются построение моделей для прогнозирования таких характеристик композиционных материалов, как модуль упругости при растяжении, прочность при растяжении и создание нейронной сети для рекомендации соотношения матрица-наполнитель.

Для решения поставленной задачи потребуется:

- Описать методы, которые используются для решений
- Провести разведочный анализ предложенных датасетов:
    1. построить гистограммы распределения каждой из переменных
    2. построить диаграммы «ящики с усами»
    3. построить попарные графики рассеяния точек
    4. получить среднее и медианное значения
    5. исключить выбросы, проверить отсутствие пропусков.
- Провести предобработку данных: удаление шумов, нормализацию
- Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении.
- Написать нейронную сеть, предназначенную для рекомендаций соотношения матрица-наполнитель.
- Разработать приложение с графическим интерфейсом, которое будет выдавать прогноз.
- Оценить точность модели на тренировочном и тестовом датасете.
- Создать репозиторий в GitHub и разместить там код исследования.

Исходные датасеты о свойствах композиционных материалов получены МГТУ им. Н.Э. Баумана – Центр НТИ «Цифровое материаловедение: новые материалы и вещества».

Датасет состоит из двух файлов:

1. файл X_bp.xlsx с данными о параметрах базальтопластика;
2. файл X_nup.xlsx с данными о нашивках из углепластика.

index	характиристики нашивок
0	Угол нашивки, град
1	Шаг нашивки
2	Плотность нашивки

index	характиристики базальтопластика
0	Соотношение матрица-наполнитель
1	Плотность, кг/м3
2	модуль упругости, ГПа
3	Количество отвердителя, м.%
4	Содержание эпоксидных групп,%_2
5	Температура вспышки, С_2
6	Поверхностная плотность, г/м2
7	Модуль упругости при растяжении, ГПа
8	Прочность при растяжении, МПа
9	Потребление смолы, г/м2

Рисунок 1. Характеристики нашивок из углепластика и характеристики базальтопластика

Далее проводится объединение двух файлов X_bp.xlsx и X_nup.xlsx по индексу, используя тип объединения INNER.

Количество строк в файле X_bp.xlsx было 1023, столбцов 10. А количество строк в файле X_nup.xlsx – 1040, столбцов 3.

```
xbp_dataFrame.shape
(1023, 10)
xnup_dataFrame.shape
(1040, 3)
```

Рисунок 2. Размерность датасетов до объединения

После объединения таблиц, 17 строк из файла X_nup.xlsx было отброшено. Дальнейшие исследования проводились с датасетом join_dataFrame, содержащим 1023 строк и 13 столбцов.

index	характиристики композиционных материалов в join_dataFrame
0	Соотношение матрица-наполнитель
1	Плотность, кг/м3
2	модуль упругости, ГПа
3	Количество отвердителя, м.%
4	Содержание эпоксидных групп,%_2
5	Температура вспышки, С_2
6	Поверхностная плотность, г/м2
7	Модуль упругости при растяжении, ГПа
8	Прочность при растяжении, МПа
9	Потребление смолы, г/м2
10	Угол нашивки, град
11	Шаг нашивки
12	Плотность нашивки

Рисунок 3. Характеристики после объединения

Проведя анализ объединенного датасета join_dataFrame, получены следующие характеристики, приведённые в Таблице 1.

Таблица 1 - Характеристики датасета

Характеристика	Тип	Уникальных	Непустых
1	2	3	4
Соотношение матрица-наполнитель	float64	1014	1023
Плотность, кг/м3	float64	1013	1023
модуль упругости, ГПа	float64	1020	1023
Количество отвердителя, м.%	float64	1005	1023
Содержание эпоксидных групп,%_2	float64	1004	1023
Температура вспышки, С_2	float64	1003	1023

1	2	3	4
Поверхностная плотность, г/м2	float64	1004	1023
Модуль упругости при растяжении, ГПа	float64	1004	1023
Прочность при растяжении, МПа	float64	1004	1023
Потребление смолы, г/м2	float64	1003	1023
Угол нашивки, град	float64	2	1023
Шаг нашивки	float64	989	1023
Плотность нашивки	float64	988	1023

Итоговый датасет join_dataFrame является начальным для дальнейших исследований, требует предобработки, которая будет выполнена дальше.

1.2. Описание используемых методов

Предсказание значений вещественной, непрерывной переменной — это задача регрессии. Эта зависимая переменная должна иметь связь с одной или несколькими независимыми переменными, называемых также предикторами или регрессорами. Регрессионный анализ помогает понять, как типичное значение зависимой переменной изменяется при изменении независимых переменных.

В настоящее время разработано много методов регрессионного анализа. Например, простая и множественная линейная регрессия. Эти модели являются параметрическими в том смысле, что функция регрессии

определяется конечным числом неизвестных параметров, которые оцениваются на основе данных.

1.2.1 Линейная регрессия

За базовую модель для прогнозирования всех искомых параметров принята линейная регрессия. LinearRegression соответствует линейной модели с коэффициентами w = (w1, …, wp), чтобы минимизировать остаточную сумму квадратов между наблюдаемыми целями в наборе данных и целями, предсказанными линейным приближением.

Простая линейная регрессия имеет место, если рассматривается зависимость между одной входной и одной выходной переменными. Для этого определяется уравнение регрессии (1.1) и строится соответствующая прямая, известная как линия регрессии.

$$y = ax + b \tag{1.1}$$

Коэффициенты a и b, называемые также параметрами модели, определяются таким образом, чтобы сумма квадратов отклонений точек, соответствующих реальным наблюдениям данных, от линии регрессии была бы минимальной. Коэффициенты обычно оцениваются методом наименьших квадратов.

Если ищется зависимость между несколькими входными и одной выходной переменными, то имеет место множественная линейная регрессия. Соответствующее уравнение имеет вид (2).

$$\left[Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n\right] \tag{1.2}$$

где n - число входных переменных.

Очевидно, что в данном случае модель будет описываться не прямой, а гиперплоскостью. Коэффициенты уравнения множественной линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек данных от этой гиперплоскости.

Линейная регрессия — первый тщательно изученный метод регрессионного анализа. Его главное достоинство — простота. Такую модель можно построить и рассчитать даже без мощных вычислительных средств. Простота является и главным недостатком этого метода. Тем не менее, именно с линейной регрессии целесообразно начать подбор подходящей модели.

На языке python линейная регрессия реализована в sklearn.linear_model.LinearRegression.

1.2.2 Регрессия k-ближайших соседей

Метод ближайших соседей (kNN - kNearestNeighbours) - метод решения задач классификации и задач регрессии, основанный на поиске ближайших объектов с известными значения целевой переменной.

Для целевой переменной метод предполагает найти ближайшие к нему объекты x1, x2…xk и построить прогноз по их меткам, то есть определить границы классов и выстроить гиперплоскость регрессии. Метка, назначенная целевой переменной, вычисляется на основе среднего значения меток ее ближайших соседей.

Алгоритм находит расстояния между запросом и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии).

Преимущества данного метода – простая реализация, низкая чувствительность к выбросам, отсутствие необходимости строить модель, настраивать несколько параметров или делать дополнительные допущения, универсальность.

Метод k-ближайших соседей (k-nearest neighbors) – это простой алгоритм машинного обучения с учителем, который можно использовать для решения задач классификации и регрессии. Он прост в реализации и понимании, но имеет существенный недостаток – значительное замедление работы, когда объем данных растет.

1.2.3 Случайный лес

Случайный лес (RandomForest) — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

– Выбирается подвыборка обучающей выборки размера samplesize (м.б. с возвращением) – по ней строится дерево (для каждого дерева — своя подвыборка).

– Для построения каждого расщепления в дереве просматриваем max_features случайных признаков (для каждого нового расщепления — свои случайные признаки).

– Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив. Формула итогового решателя (1.3) — это усреднение предсказаний отдельных деревьев.

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x) \qquad (1.3)$$

где

N – количество деревьев;

i – счетчик для деревьев;

b – решающее дерево;

x – сгенерированная нами на основе данных выборка.

Для определения входных данных каждому дереву используется метод случайных подпространств. Базовые алгоритмы обучаются на различных подмножествах признаков, которые выделяются случайным образом.

Преимущества случайного леса:

- высокая точность предсказания;
- редко переобучается;
- практически не чувствителен к выбросам в данных;
- одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки, данные с большим числом признаков;
- высокая параллелизуемость и масштабируемость.

Из недостатков можно отметить, что его построение занимает больше времени. Так же теряется интерпретируемость.

Метод реализован в sklearn.ensemble.RandomForestRegressor.

1.2.4 Многослойный перцептрон

Нейронная сеть — это последовательность нейронов, соединенных между собой связями. Нейронные сети применяются для решения задач регрессии, классификации, распознавания образов и речи, компьютерного зрения и других. На настоящий момент это самый мощный, гибкий и широко применяемый инструмент в машинном обучении. Персептроны пытаются имитировать функциональность человеческого мозга для решения задач. Вычислительная единица нейронной сети — нейрон или персептрон.

У каждого нейрона есть определённое количество входов, куда поступают сигналы, которые суммируются с учётом значимости (веса) каждого входа. Смещение – это дополнительный вход для нейрона, который всегда равен 1 и, следовательно, имеет собственный вес соединения. Так же у нейрона есть функция активации, которая определяет выходное значение нейрона. Она используется для того, чтобы ввести нелинейность в нейронную сеть. Примеры активационных функций: relu, сигмоида.

У полносвязной нейросети выход каждого нейрона подается на вход всем нейронам следующего слоя. У нейросети имеется:

- входной слой — его размер соответствует входным параметрам;
- скрытые слои — их количество и размерность определяем специалист;
- выходной слой — его размер соответствует выходным параметрам.

Прямое распространение – это процесс передачи входных значений в нейронную сеть и получения выходных данных, которые называются прогнозируемым значением. Прогнозируемое значение сравниваем с фактическим с помощью функции потери. В методе обратного распространения ошибки градиенты (производные значений ошибок)

вычисляются по значениям весов в направлении, обратном прямому распространению сигналов. Значение градиента вычитают из значения веса, чтобы уменьшить значение ошибки. Таким образом происходит процесс обучения. Обновляются веса каждого соединения, чтобы функция потерь минимизировалась. Для обновления весов в модели используются различные оптимизаторы. Количество эпох показывает, сколько раз выполнялся проход для всех примеров обучения.

1.2.5 Лассо регрессия

Lasso (Least absolute shrinkage and selection operator) - метод оценивания коэффициентов линейной регрессионной модели.

Метод заключается во введении ограничения на норму вектора коэффициентов модели, что приводит к обращению в 0 некоторых коэффициентов модели. Метод приводит к повышению устойчивости модели в случае большого числа обусловленности матрицы признаков X, позволяет получить интерпретируемые модели - отбираются признаки, оказывающие наибольшее влияние на вектор ответов.

1.2.6 Метрики качества моделей

Существует множество различных метрик качества, применимых для регрессии. В этой работе используются:

R2 или коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой переменной (т.е. очень низкое). Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

MSE (Mean Squared Error) или средняя квадратичная ошибка принимает значениях в тех же единицах, что и целевая переменная. Чем ближе к нулю MSE, тем лучше работают предсказательные качества модели.

1.3. Разведочный анализ данных

Для получения представления о характере распределения переменных в датасете, формирования оценки качества исходных данных (наличия пропусков, выбросов), выявления характера взаимосвязи между переменными с целью последующего выдвижения гипотез о наиболее подходящих для решения задачи моделях машинного обучения проведем разведочный анализ данных.

В качестве инструментов разведочного анализа используется оценка статистических характеристик данных, а также матрица попарной корреляции, тепловая карта корреляции, гистограммы нормального распределения, поиск выбросов через ящик с усами.

	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

Рисунок 4. Описательная статистика

Проверку пропусков выполняли с метода join_dataFrame.info(), который показывает количество ненулевых значений и тип данных:

```
<class 'pandas.core.frame.DataFrame'>
Float64Index: 1023 entries, 0.0 to 1022.0
Data columns (total 13 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   Соотношение матрица-наполнитель   1023 non-null   float64
 1   Плотность, кг/м3                  1023 non-null   float64
 2   модуль упругости, ГПа             1023 non-null   float64
 3   Количество отвердителя, м.%       1023 non-null   float64
 4   Содержание эпоксидных групп,%_2   1023 non-null   float64
 5   Температура вспышки, С_2          1023 non-null   float64
 6   Поверхностная плотность, г/м2     1023 non-null   float64
 7   Модуль упругости при растяжении, ГПа  1023 non-null   float64
 8   Прочность при растяжении, МПа     1023 non-null   float64
 9   Потребление смолы, г/м2           1023 non-null   float64
 10  Угол нашивки, град                1023 non-null   float64
 11  Шаг нашивки                       1023 non-null   float64
 12  Плотность нашивки                 1023 non-null   float64
dtypes: float64(13)
memory usage: 111.9 KB
```

Рисунок 5. Проверка количества пропусков

Метод join_dataFrame.nunique() возвращает количество уникальных значений для каждого столбца:

```
Соотношение матрица-наполнитель          1014
Плотность, кг/м3                         1013
модуль упругости, ГПа                    1020
Количество отвердителя, м.%              1005
Содержание эпоксидных групп,%_2          1004
Температура вспышки, С_2                 1003
Поверхностная плотность, г/м2            1004
Модуль упругости при растяжении, ГПа     1004
Прочность при растяжении, МПа            1004
Потребление смолы, г/м2                  1003
Угол нашивки, град                          2
Шаг нашивки                               989
Плотность нашивки                         988
dtype: int64
```

Рисунок 6. Количество уникальных значений

Для проведения разведочного анализа использовали язык программирования Python и библиотеки Numpy, Pandas, Matplotlib, Seaborn и Sklearn.

Для визуализации распределения значений по каждому столбцу и взаимосвязи между данными используем sns.histplot, sns.boxplot, sns.pairplot.

2. Практическая часть

2.1. Предобработка данных

Гистограммы распределения переменных показаны на Рисунке 7. Видно, что все признаки, кроме «Угол нашивки», имеют нормальное распределение и принимают неотрицательные значения. «Угол нашивки» принимает два значения: 0, 90.

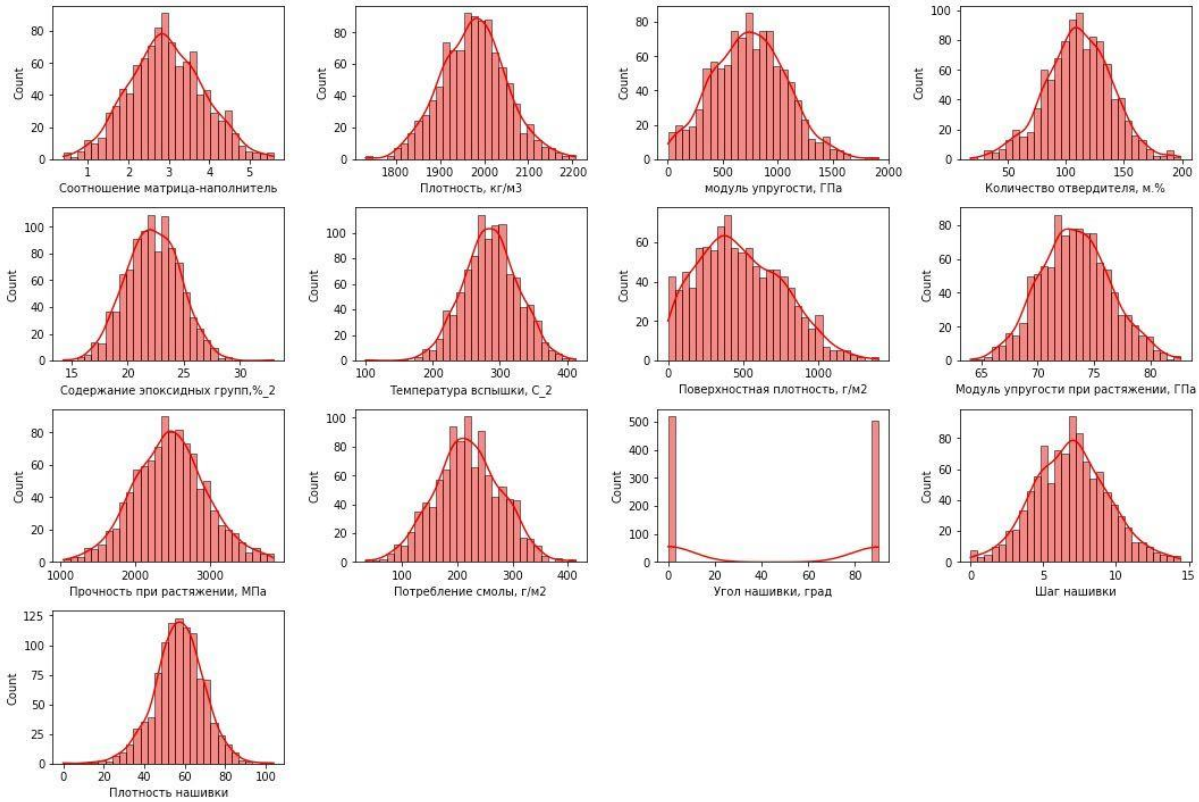

Рисунок 7. Гистограммы распределения переменных

Характеристику «Угол нашивки, град» закодируем с помощью LabelEncoder. Класс LabelEncoder используется для кодирования данных, имеющих два варианта значений, одно из которых будет закодировано 0, а второе 1.

Изучим графики попарного рассеяния (Рисунок 8).

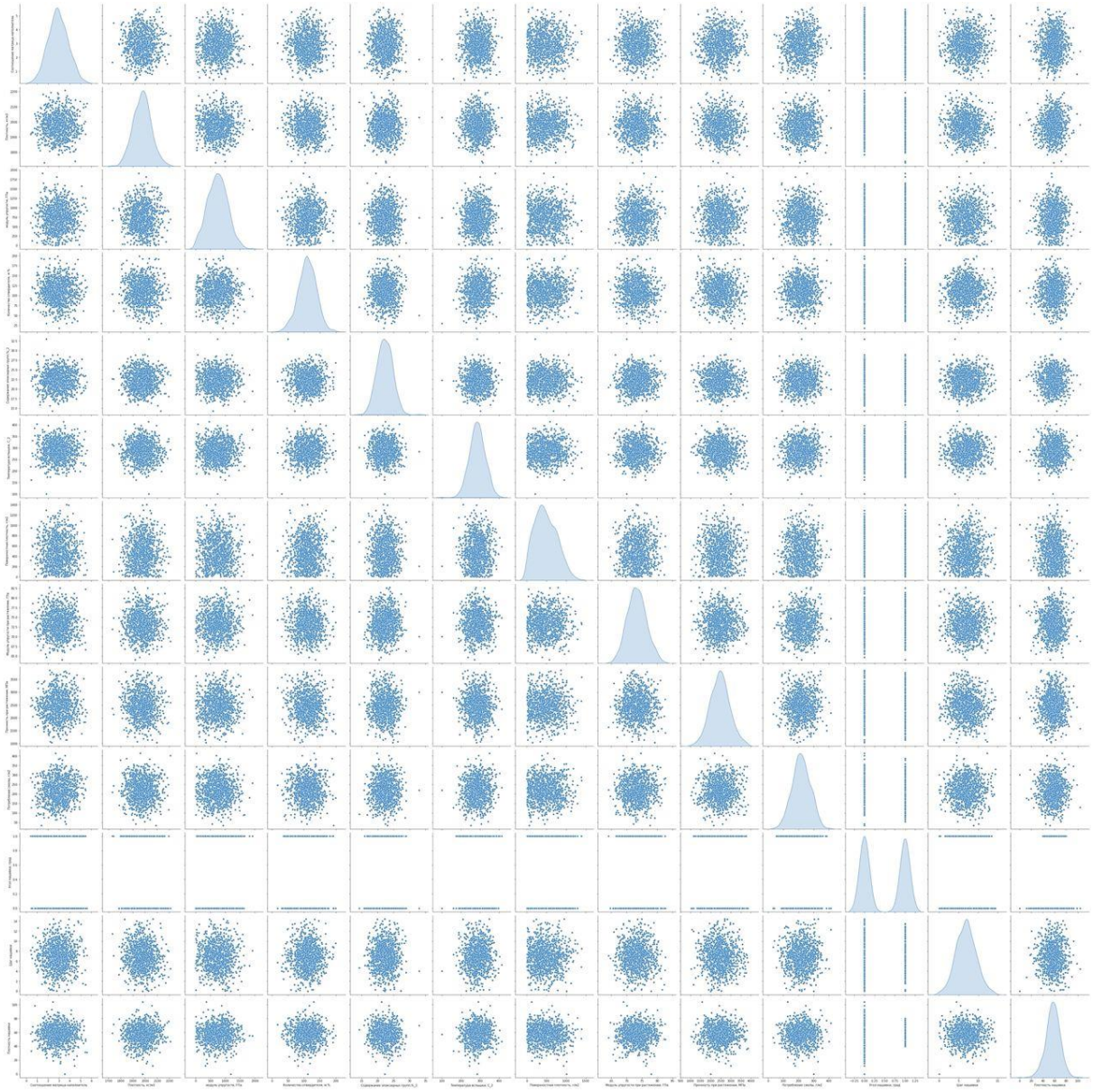

Рисунок 8. Графики попарного рассеяния

По форме разброса точек, в виде облаков, становится понятно, что зависимости между переменными, на которых будет основываться работа модели, не обнаруживаются.

Ящик с усами, диаграмма размаха (англ. box-and-whiskers diagram or plot, box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

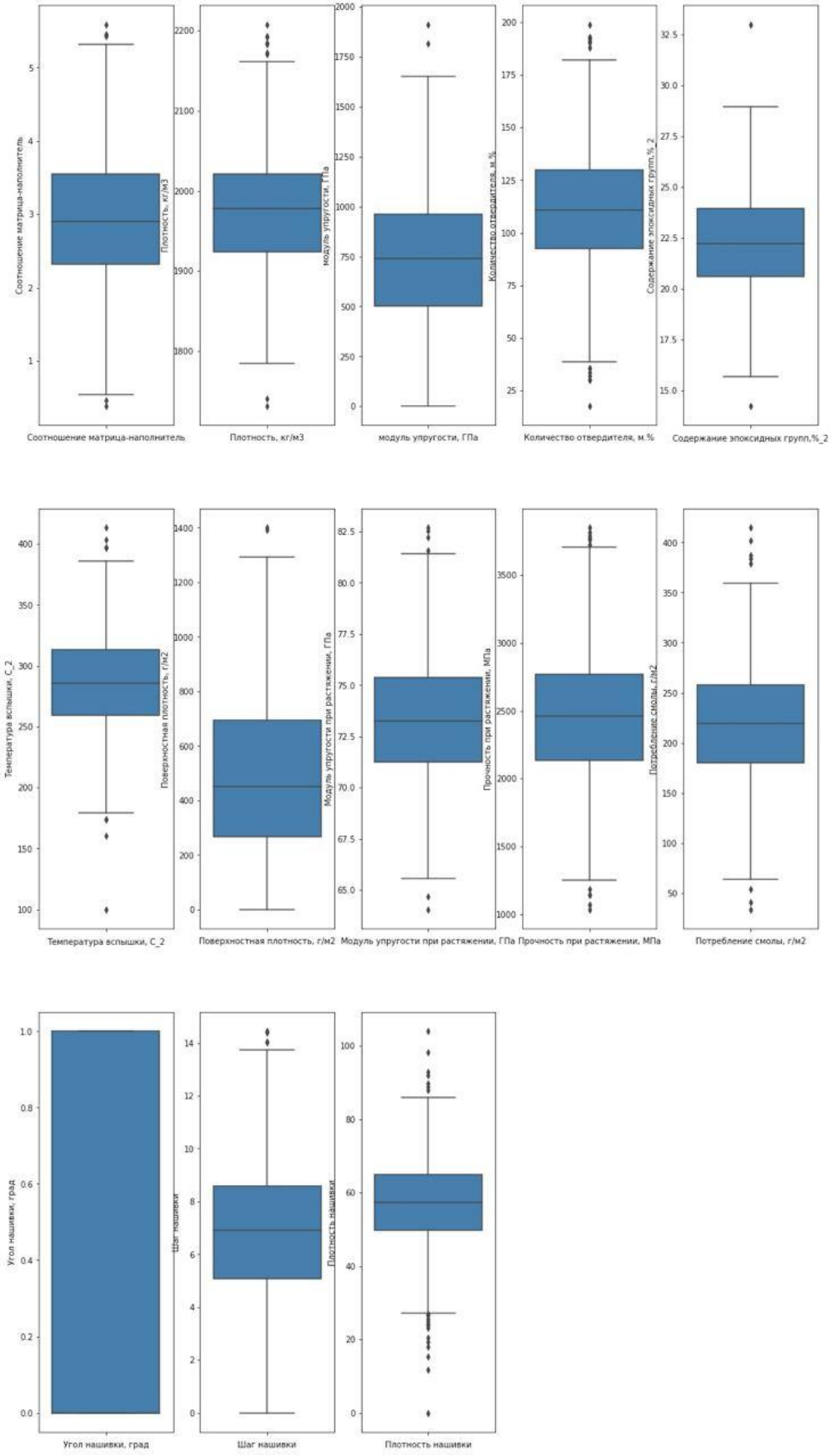

Рисунок 9. График «ящик с усами» до удаления выбросов

Попробуем выявить связь между характеристиками с помощью матрицы корреляции.

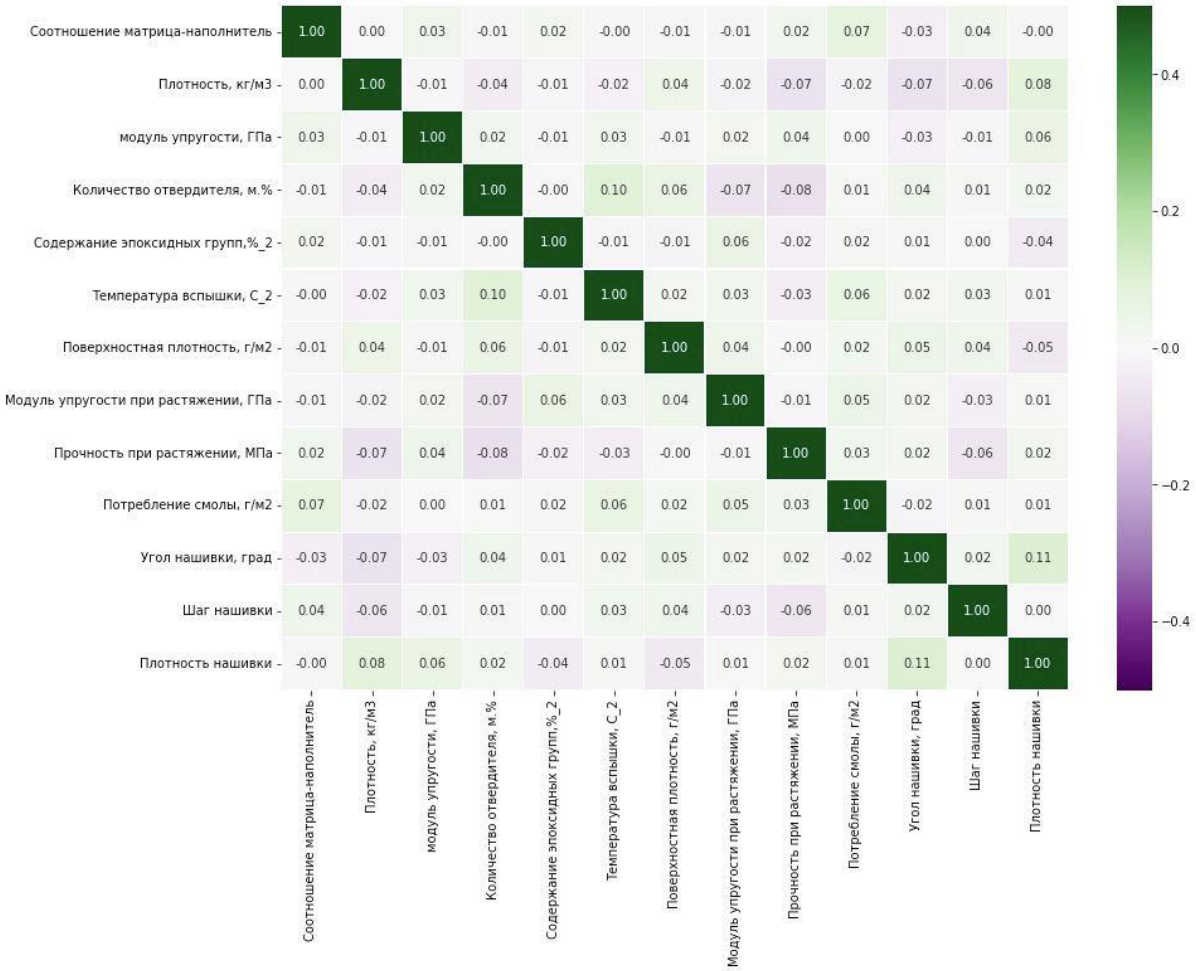

Рисунок 10. Корреляционная матрица до удаления выбросов

Визуализация коэффициентов корреляции выполнена с помощью тепловой карты sns.heatmap.

Видно, что все коэффициенты корреляции близки к нулю. Это означает отсутствие линейной зависимости между признаками.

Проанализировав диаграммы ящика с усами, делаем вывод, что пригодным является параметр "Угол нашивки". В остальных данных необходима работа с выбросами.

Выброс — это экстремальные значения во входных данных, которые находятся далеко за пределами других наблюдений.

Многие алгоритмы машинного обучения чувствительны к разбросу и распределению значений признаков обрабатываемых объектов. Соответственно, выбросы во входных данных могут исказить и ввести в заблуждение процесс обучения алгоритмов машинного обучения, что приводит к увеличению времени обучения, снижению точности моделей и, в конечном итоге, к снижению результатов.

Причины возникновения выбросов:

- сбой работы оборудования;
- человеческий фактор;
- случайность;
- уникальные явления и др.

В ходе предобработки выявлено следующее число выбросов:

24 -- Выбросы методом 3-х сигм

93 -- Выбросы методом межквартильных расстояний

Метод 3-х сигм найдено меньше выбросов. График «ящик с усами» показывает небольшой размах. Посмотрим на распределение выбросов по разным характеристикам.

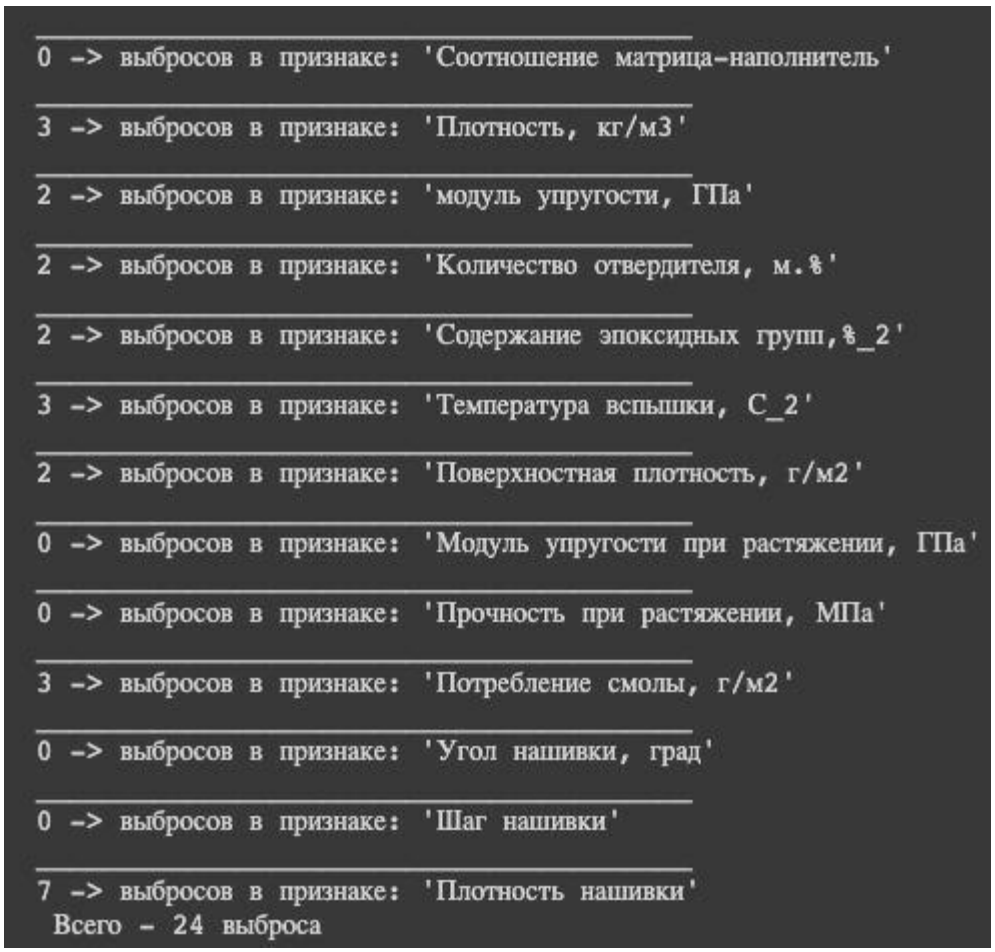

Рисунок 11. Распределение выбросов по характеристикам

Выбросы достаточно хорошо распределены по характеристикам, поэтому удалим те данные, которые выявили, как выбросы методом 3-х сигм, тем самым, сохраним большее количество данных в датасете: 999 rows × 13 columns, количество строк стало на 24 меньше.

Построим ящики с усами после удаления выбросов (Рисунок 12).

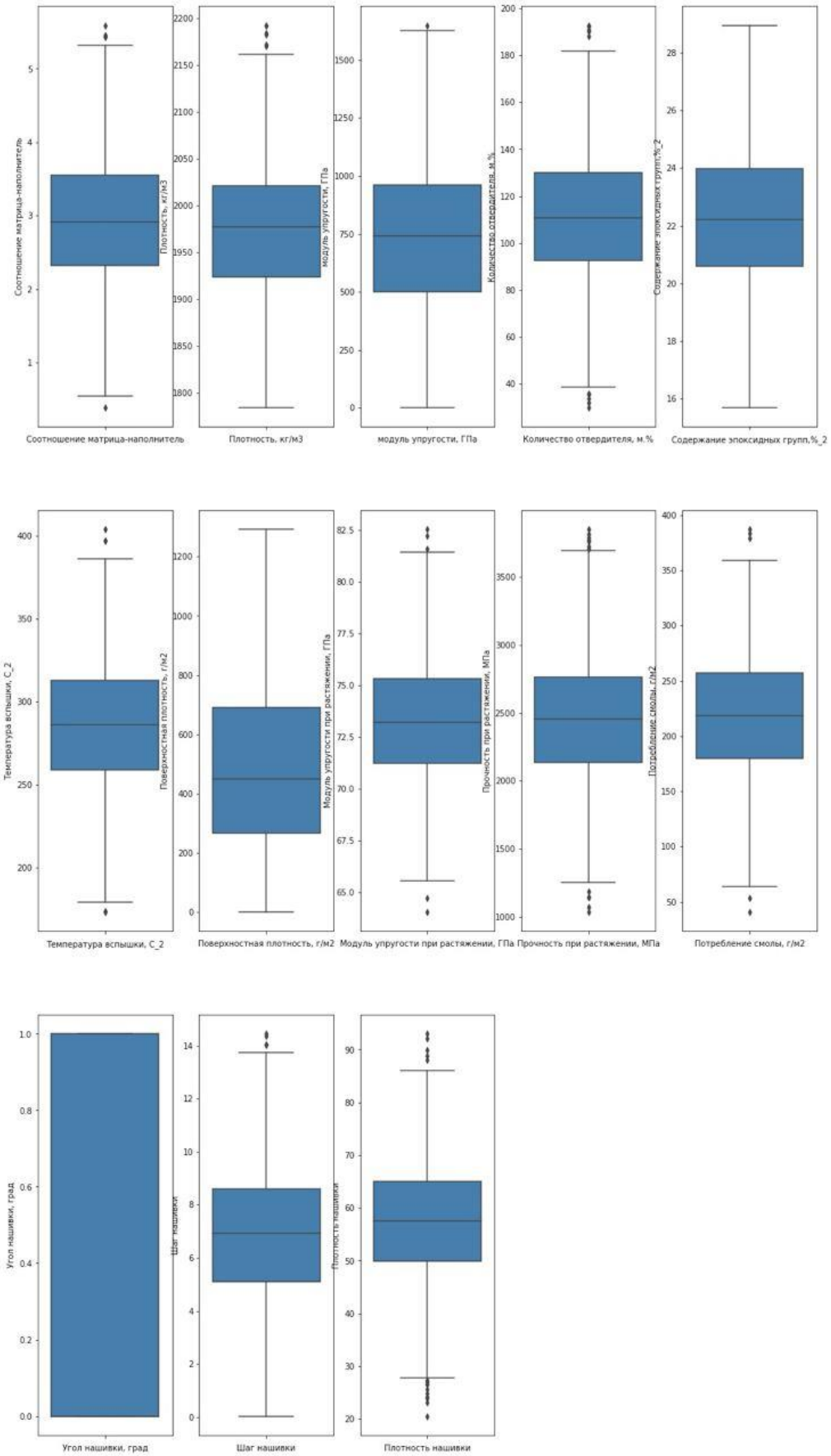

Рисунок 12. График «ящик с усами» после удаления выбросов

Построим матрицу корреляции после удаления выбросов (Рисунок 13), чтобы посмотреть, как изменились зависимости.

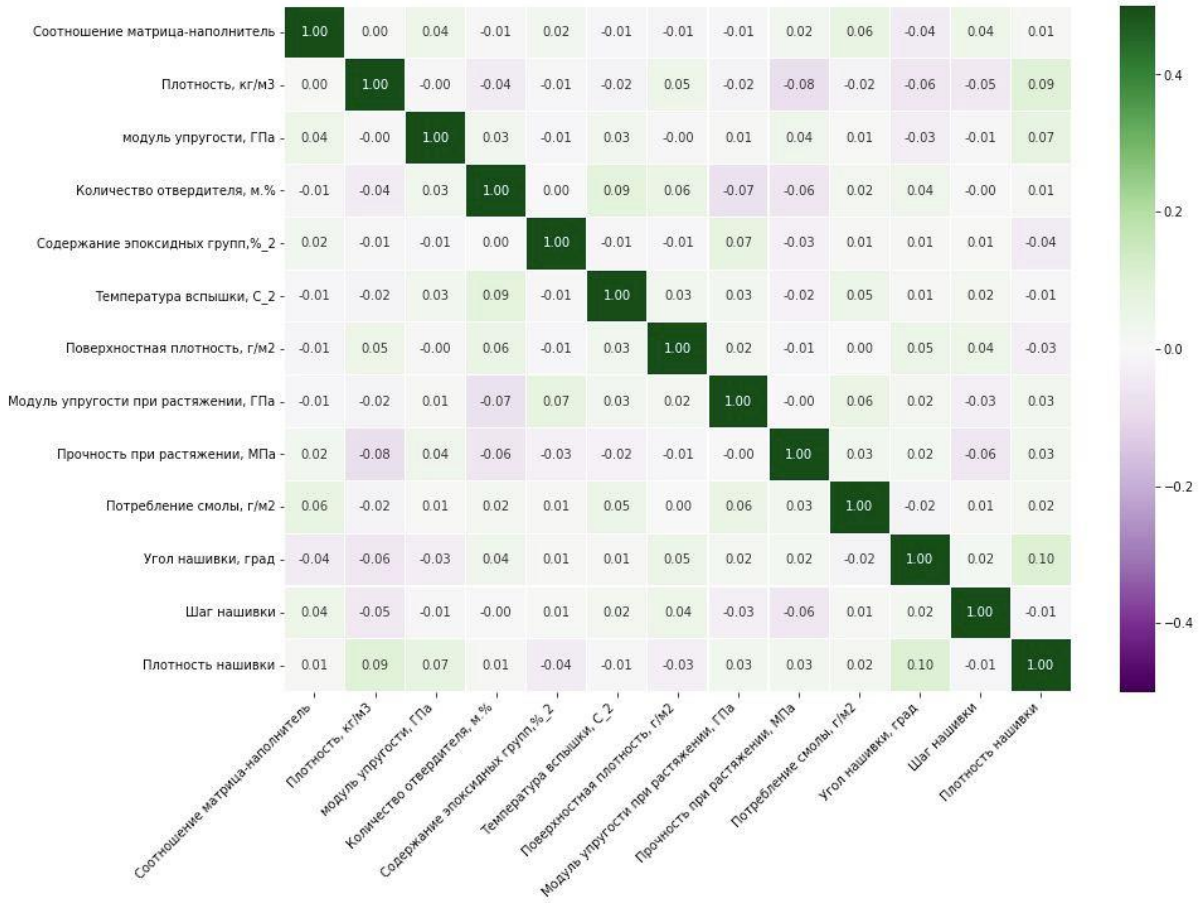

Рисунок 13. Матрица корреляции после удаления выбросов

Как видим по графику, в результате удаления выбросов, корреляция выросла незначительно, и кардинальных изменений нет, корреляция между признаками по-прежнему не выявляется.

Видно, что данные находятся в разных диапазонах. Оценка плотности ядра показывает, что данные нужно нормализовать.

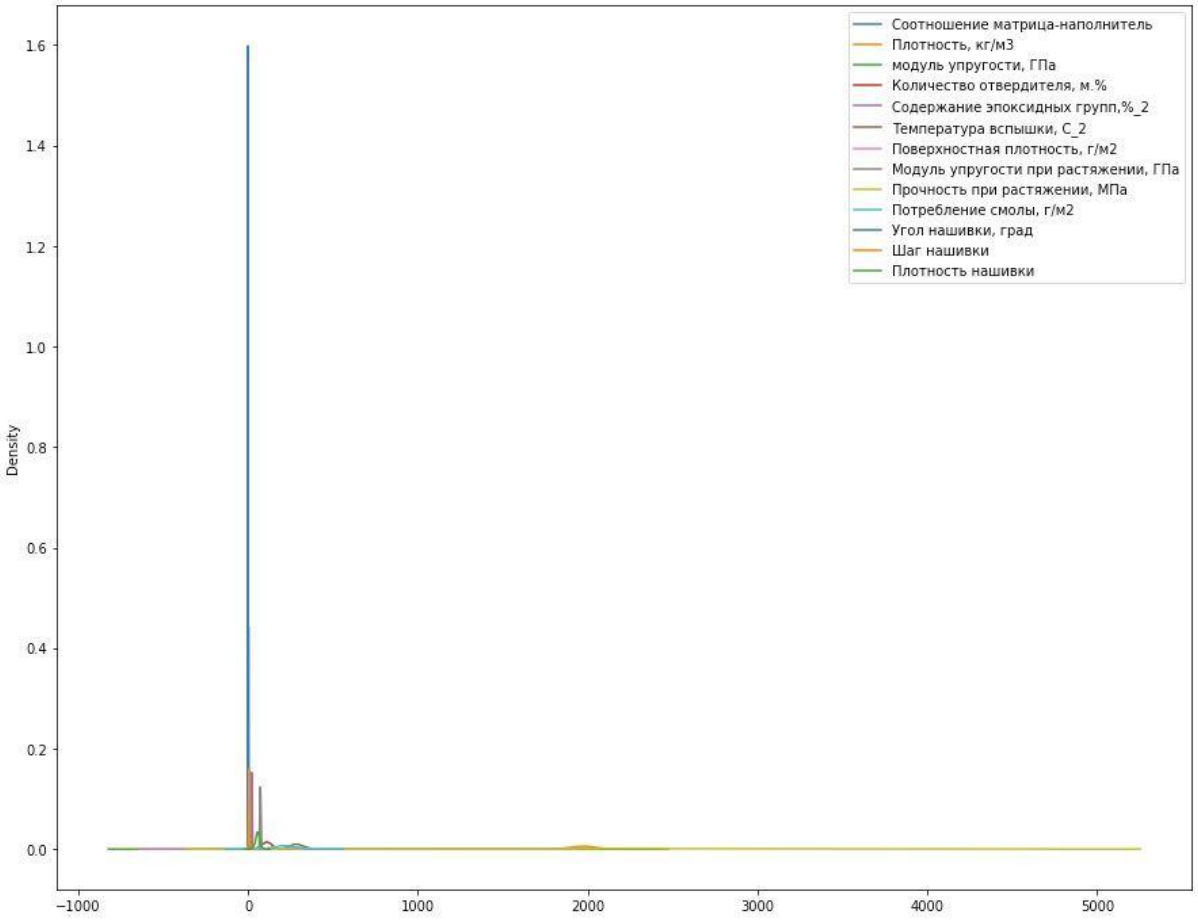

Рисунок 14. Оценка плотности ядра

Масштабировать будем с помощью приведения каждого признака к диапазону от 0 до 1 с помощью метода MinMaxScaler.

Описательная статистика характеристика после нормализации (Рисунок 15).

index	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	999.0	0.49	0.17	0.0	0.37	0.48	0.61	1.0
Плотность, кг/м3	999.0	0.47	0.18	0.0	0.34	0.47	0.58	1.0
модуль упругости, ГПа	999.0	0.45	0.2	0.0	0.3	0.45	0.58	1.0
Количество отвердителя, м.%	999.0	0.5	0.17	0.0	0.38	0.5	0.61	1.0
Содержание эпоксидных групп,%_2	999.0	0.49	0.18	0.0	0.37	0.49	0.62	1.0
Температура вспышки, С_2	999.0	0.49	0.17	0.0	0.37	0.49	0.61	1.0
Поверхностная плотность, г/м2	999.0	0.37	0.22	0.0	0.21	0.35	0.54	1.0
Модуль упругости при растяжении, ГПа	999.0	0.5	0.17	0.0	0.39	0.5	0.61	1.0
Прочность при растяжении, МПа	999.0	0.51	0.17	0.0	0.39	0.5	0.61	1.0
Потребление смолы, г/м2	999.0	0.51	0.17	0.0	0.4	0.51	0.62	1.0
Угол нашивки, град	999.0	0.5	0.5	0.0	0.0	0.0	1.0	1.0
Шаг нашивки	999.0	0.48	0.18	0.0	0.35	0.48	0.59	1.0
Плотность нашивки	999.0	0.51	0.16	0.0	0.41	0.51	0.61	1.0

Рисунок 15. Описательная статистика характеристика после нормализации

Оценим «Ящики с усами», гистограммы и корреляционную матрицу после нормализации (Рисунок 16, 17, 18)

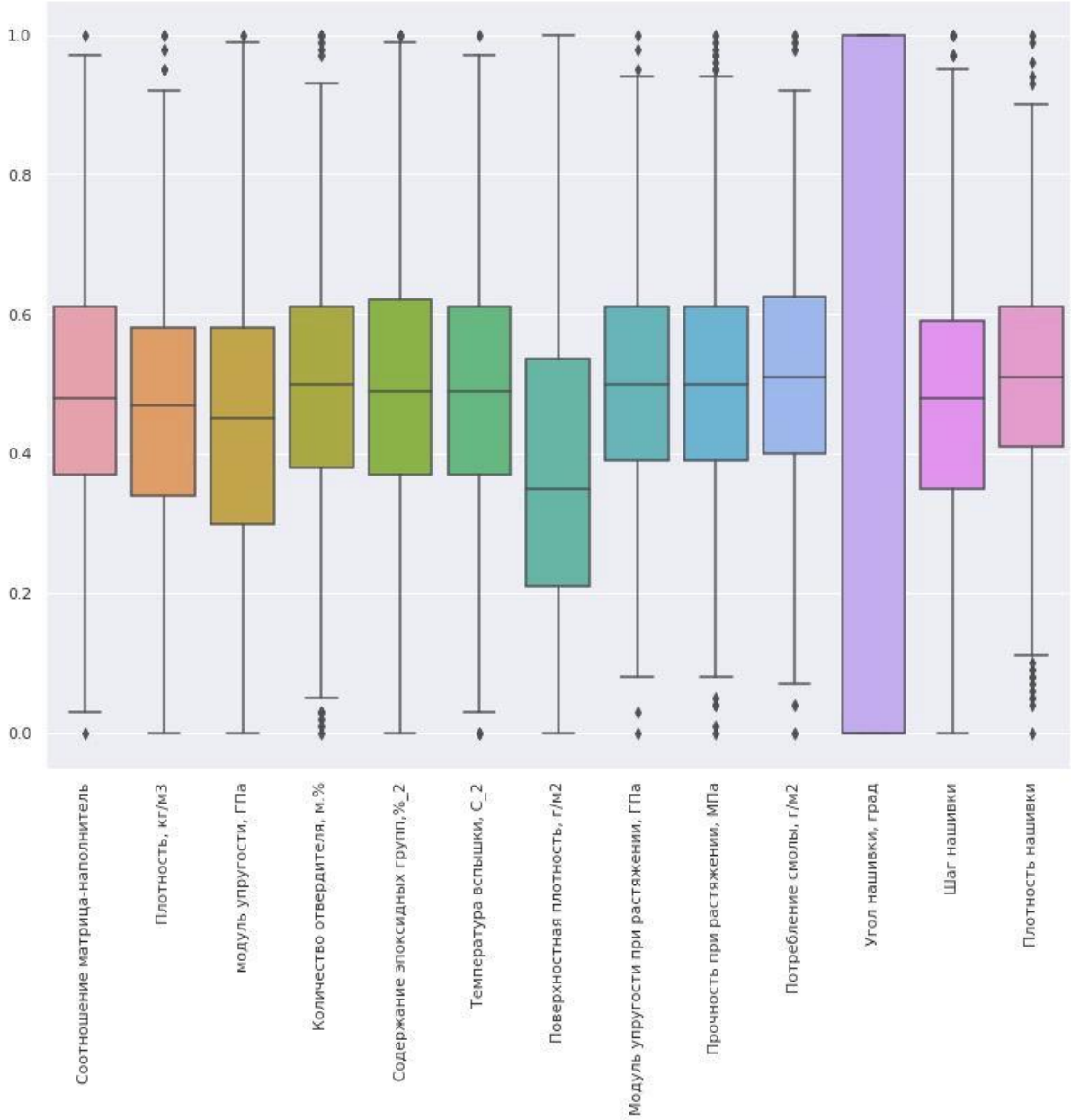

Рисунок 16. «Ящики с усами после нормализации

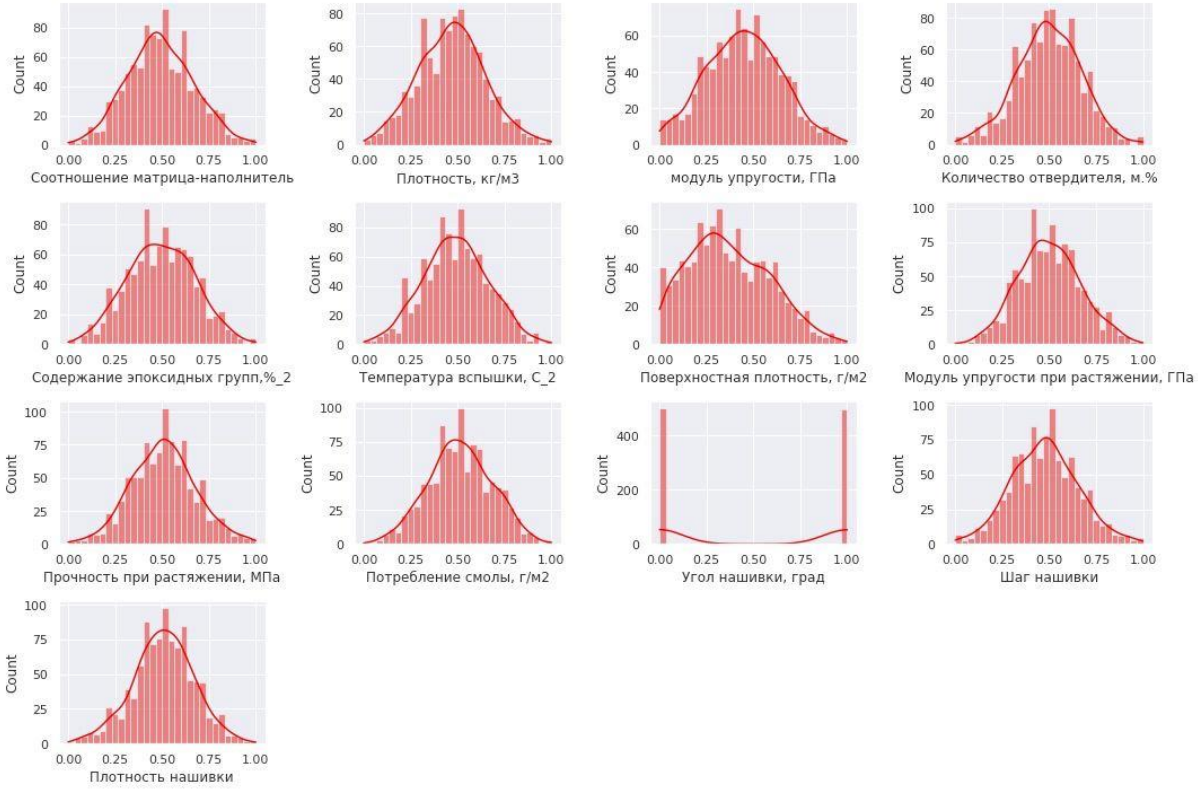

Рисунок 17. Гистограммы после нормализации

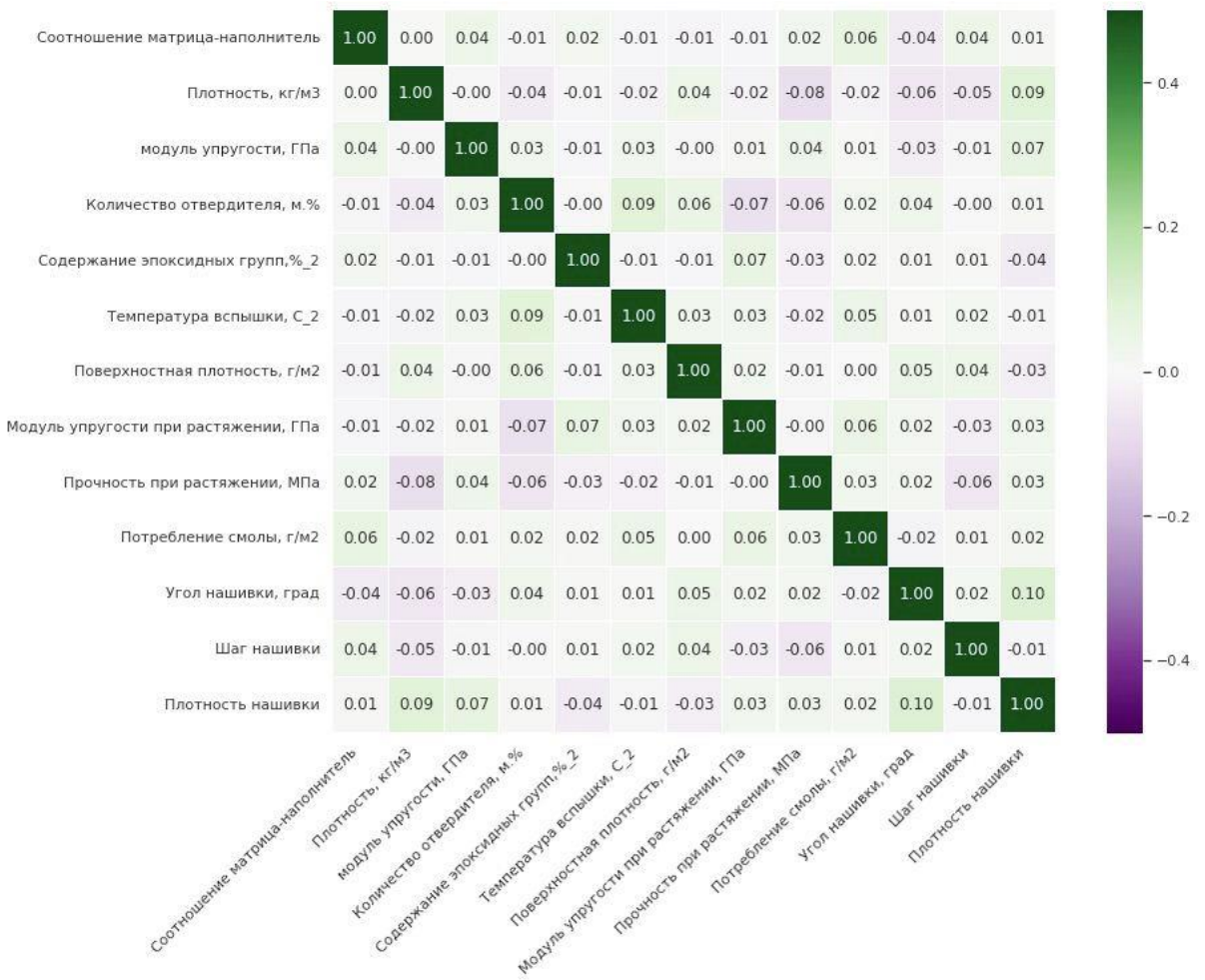

Рисунок 18. Корреляционная матрица после нормализации

Можно предположить, что качество прогноза линейных моделей будет невысоким, т.к. корреляции близки к нулю.

2.2. Разработка и обучение моделей

Для прогноза модуля упругости при растяжении и прочности при растяжении использованы модели:

- Линейной регрессии;
- Метод k-ближайших соседей;
- «Случайный лес»;
- Лассо регрессия;
- Многослойный персептроный регрессор.

Для обучения используем 70 % данных, а для тестирования — 30 %.

Зерно генератора случайных чисел зададим постоянным для воспроизводимости результатов обучения.

2.3. Тестирование моделей

Ниже представлены результаты работы моделей для двух целевых переменных в виде соотношения тест/прогноз, а также итоговый датасет ошибок.

2.3.1 Линейная регрессия

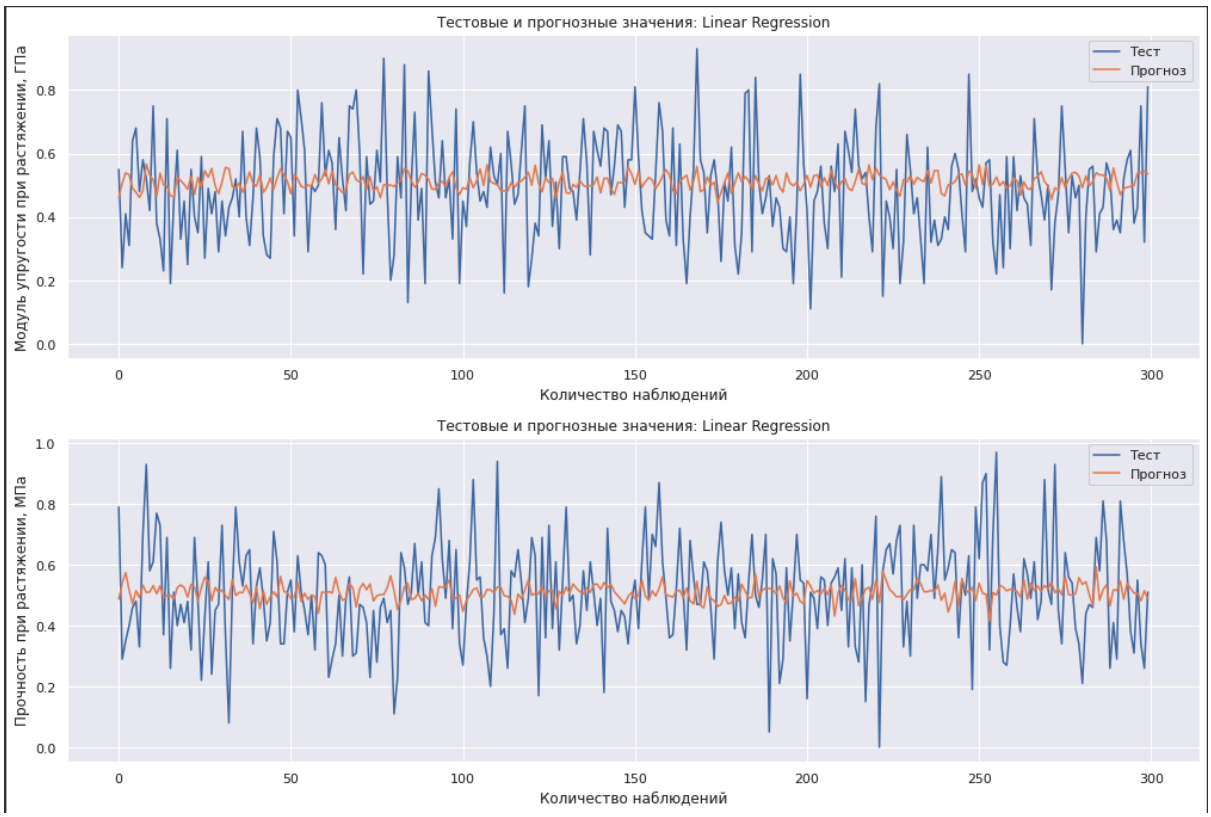

Рисунок 19. Визуализация линейной регрессии

2.3.2 Регрессия k-ближайших соседей

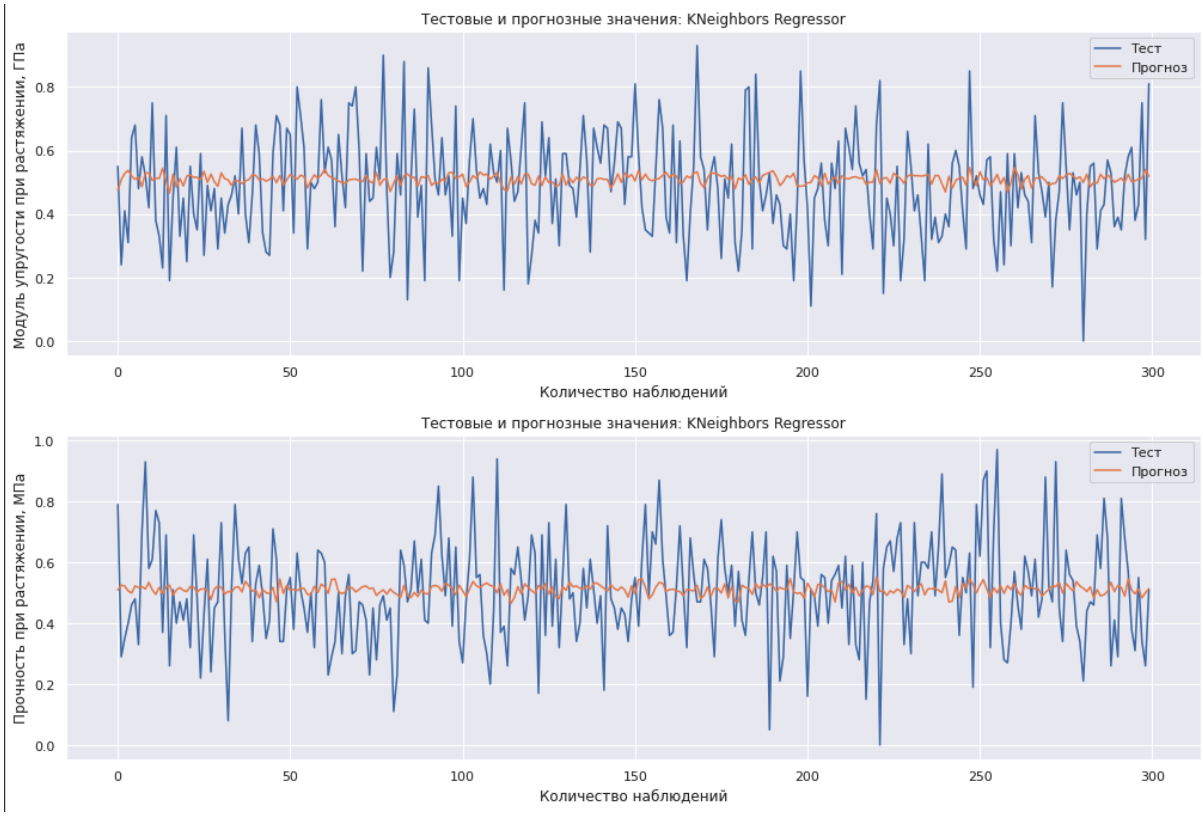

Рисунок 20. Визуализация k-ближайших соседей

2.3.3 Случайный лес

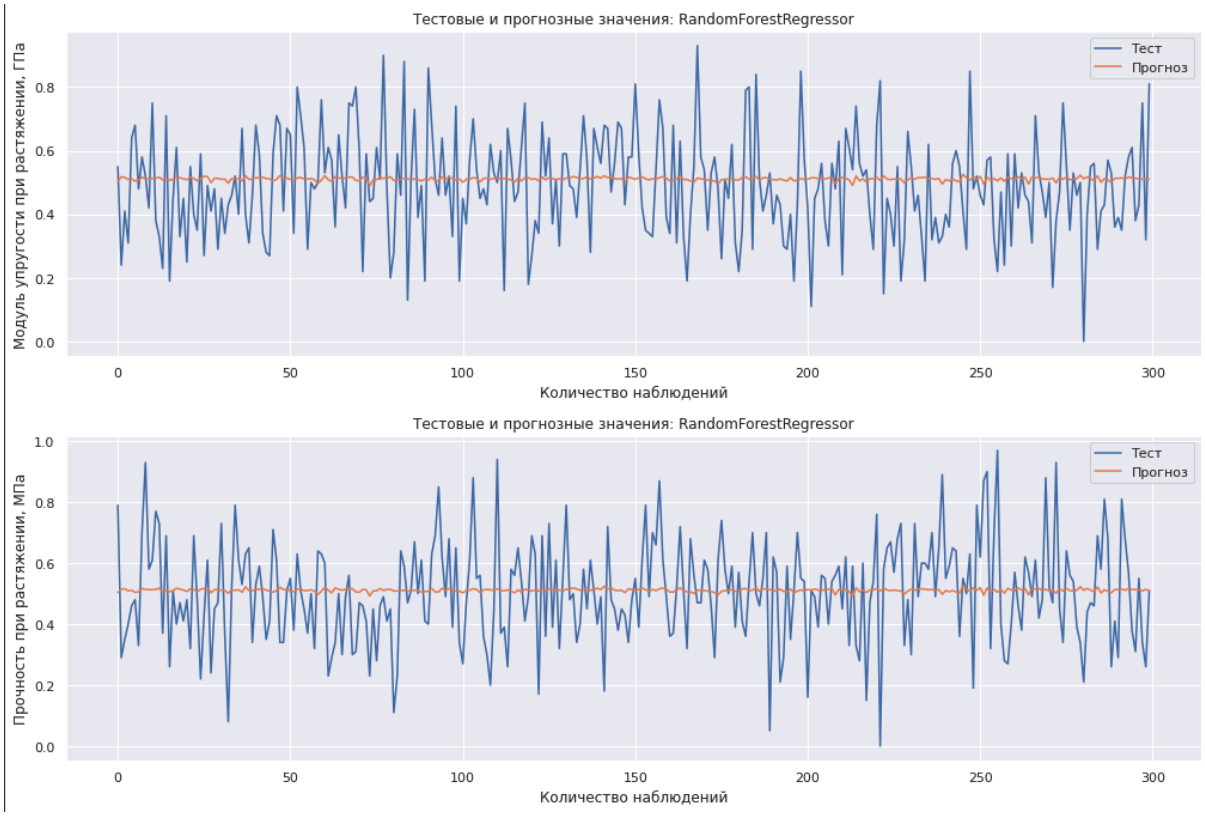

Рисунок 21. Визуализация «случайный лес»

2.3.4 Многослойный перцептрон

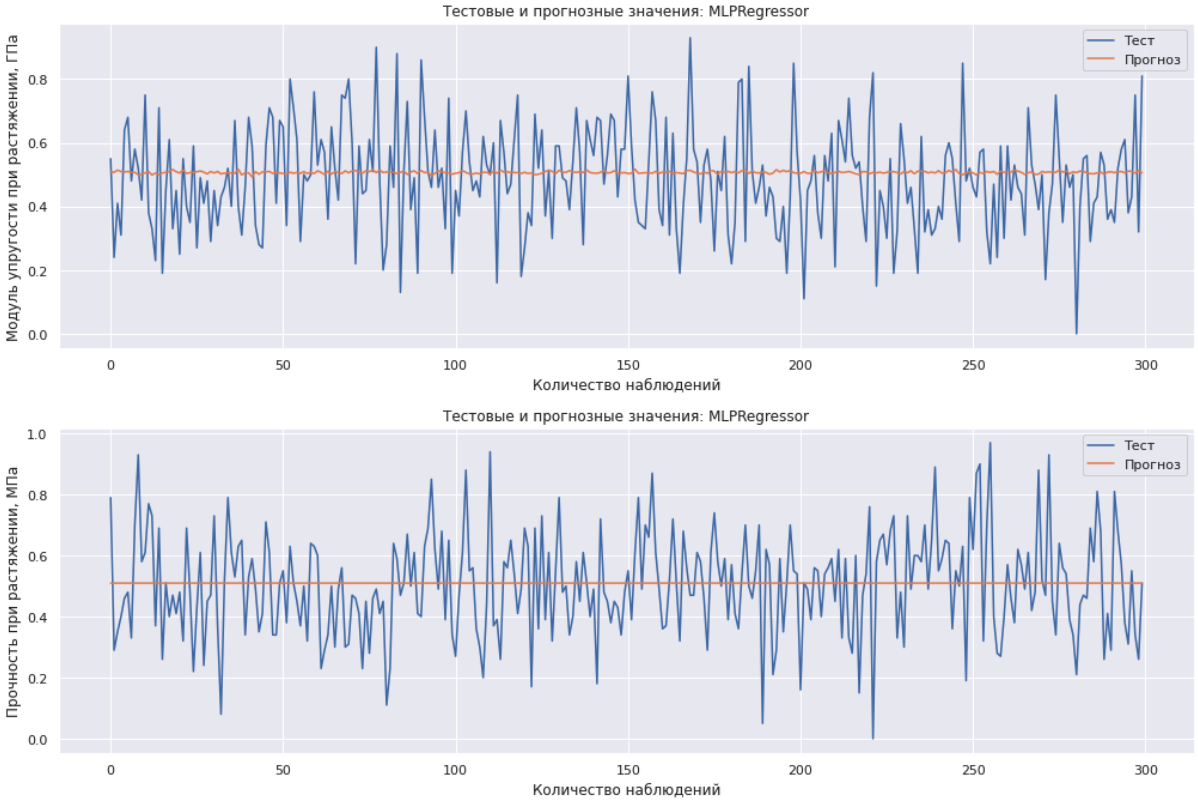

Рисунок 22. Визуализация Многослойный персептрон

2.3.5 Лассо регрессия

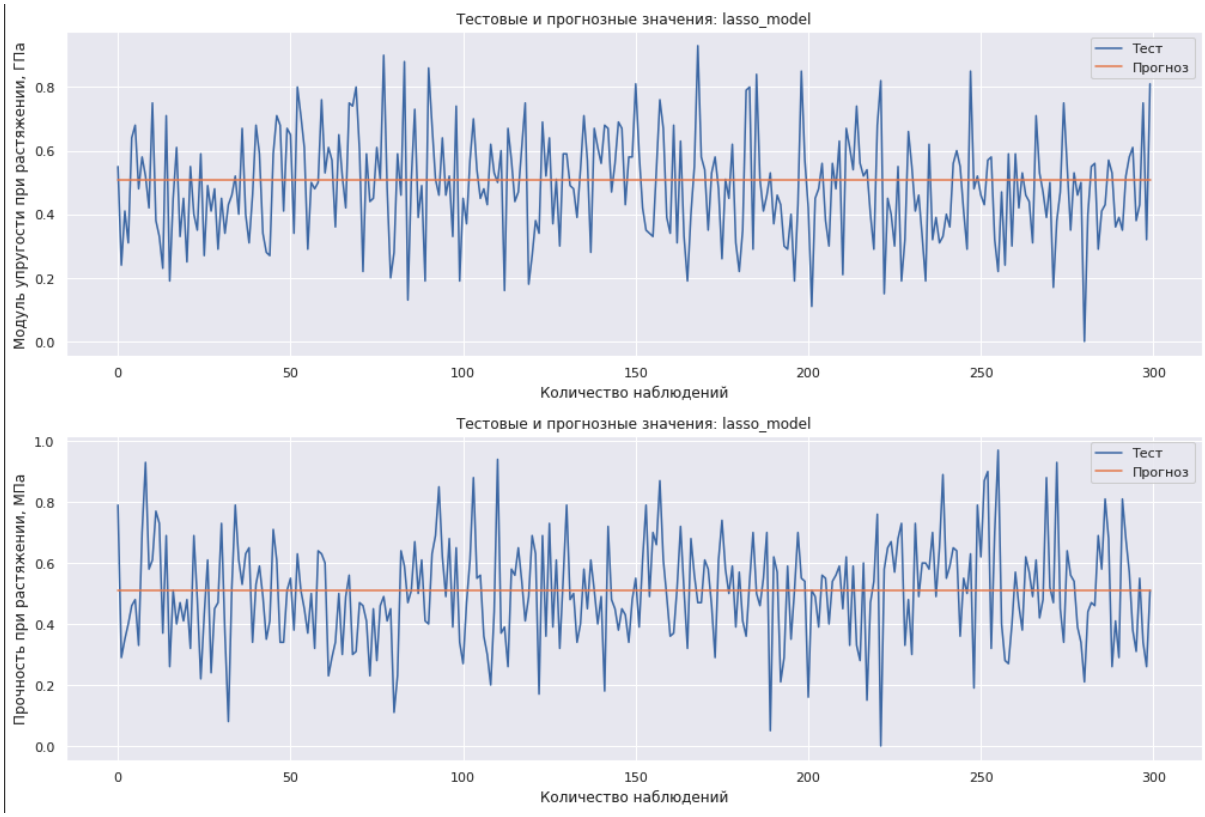

Рисунок 23. Визуализация Лассо регрессия

index	target_var	model_name	MSE	R2
0	Модуль упругости при растяжении, ГПа	Linear Regression	0.02798367226727433	-0.027799467409495682
1	Прочность при растяжении, МПа	Linear Regression	0.02882712765798571	-0.012278080179114692
2	Модуль упругости при растяжении, ГПа	KNeighborsRegressor	0.027689151925078044	-0.016982164806770284
3	Прочность при растяжении, МПа	KNeighborsRegressor	0.02888706698854738	-0.014382877826249807
4	Модуль упругости при растяжении, ГПа	RandomForestRegressor	0.027616678542793847	-0.014320323179923822
5	Прочность при растяжении, МПа	RandomForestRegressor	0.028468337706038273	0.00032099691430820254
6	Модуль упругости при растяжении, ГПа	MLPRegressor	0.027745752151894395	-0.01906100858472337
7	Прочность при растяжении, МПа	MLPRegressor	0.028503449010552653	-0.0009119529774772595
8	Модуль упругости при растяжении, ГПа	lasso_model	0.027706436085886032	-0.017616986823816516
9	Прочность при растяжении, МПа	lasso_model	0.028506689898301473	-0.00102575826766932

Рисунок 24. Датасет с ошибками

Модели показали неудовлетворительный результат.

Если результат отрицательный, наша модель не так хороша, как догадки.

2.4. Нейронная сеть, которая будет рекомендовать соотношение матрица-наполнитель.

Постоим нейронную сеть с помощью класса keras.Sequential со следующими параметрами:

– входной слой нормализации 12 признаков;
– выходной слой для 1 признака;
– скрытых слоев: 5;
– нейронов в скрытом слое: 128, 128, 128, 64, 32, 16;
– активационная функция скрытых слоев: relu;
– оптимизатор: Adam;
– loss-функция: MeanSquaredError.

```
[160] model = tf.keras.Sequential([x_train_norm, layers.Dense(128, activation='relu'),
                                                 layers.Dense(128, activation='relu'),
                                                 layers.Dense(128, activation='relu'),
                                                 layers.Dense(64, activation='relu'),
                                                 layers.Dense(32, activation='relu'),
                                                 layers.Dense(16, activation='relu'),
                                                 layers.Dense(1)
                                                ])

      model.compile(optimizer=tf.keras.optimizers.Adam(0.001), loss='mean_squared_error')

[161] model.summary()

      Model: "sequential"
      _________________________________________________________________
       Layer (type)                Output Shape              Param #
      =================================================================
       normalization (Normalizatio  (None, 12)              25
       n)

       dense (Dense)               (None, 128)               1664

       dense_1 (Dense)             (None, 128)               16512

       dense_2 (Dense)             (None, 128)               16512

       dense_3 (Dense)             (None, 64)                8256

       dense_4 (Dense)             (None, 32)                2080

       dense_5 (Dense)             (None, 16)                528

       dense_6 (Dense)             (None, 1)                 17

      =================================================================
      Total params: 45,594
      Trainable params: 45,569
      Non-trainable params: 25
      _________________________________________________________________
```

Рисунок 25. Слои и конфигурация нейросети

Параметры нейросети следующие:

– разбиение данных на тестовые и проверочных: 30%

– количество эпох: 100

Визуализация тест/прогноз и график потерь модели (MSE) показаны ниже.

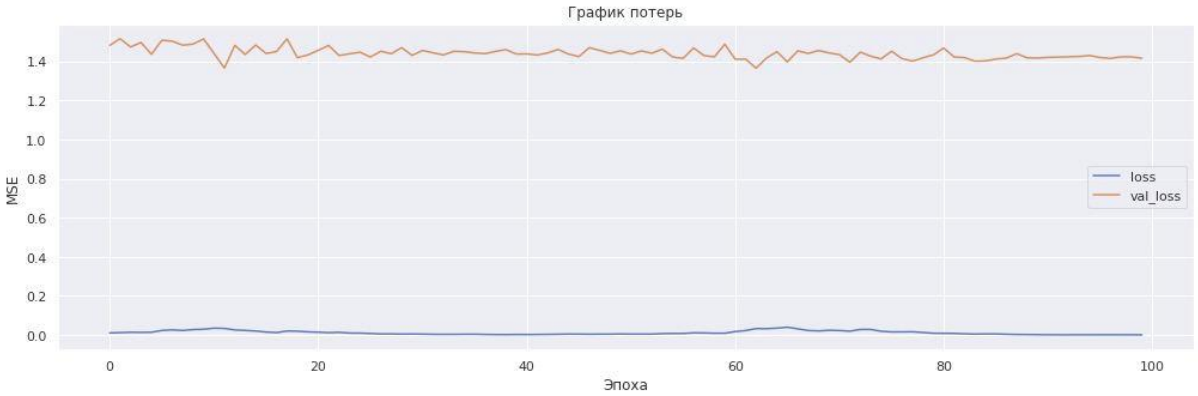

Рисунок 26. График потерь

Рисунок 27. Визуализация тест/прогноз соотношения матрицы-наполнитель

Ошибки модели: MSE= 1.1775, R^2 = -0.5459. Результаты неудовлетворительны.

2.6. Создание удаленного репозитория и загрузка результатов работы на него.

В процессе выполнения выпускной квалификационной работы был создан репозиторий на GitHub, который находится по адресу:

pprf01/VKR-Bauman (github.com)

В репозитории доступны следующие результаты: код анализа данных и обучения моделей, презентация, пояснительная записка к выпускной квалификационной работе.

Заключение

В ходе выполнения данной работы было выполнено:

- изучение теоретических методов анализа данных и машинного обучения;
- разведочный анализ данных;
- предобработка данных;
- построение регрессионных моделей;
- визуализация модели и оценка качества прогноза;

Использованные при разработке моделей подходы не позволили получить достоверных прогнозов. Возможные причины неудовлетворительной работы моделей и пути решения:

- Необходима дополнительная информации о зависимости признаков с точки зрения физики процесса.
- Возможно, исследование предварительно обработанных данных, не позволяет построить качественные модели на этом датасете.
- Надо использовать и другие методы прогноза, но мой опыт сейчас не достаточен для системного подхода к сложной задаче.

На основании проведенного исследования можно сделать следующие основные выводы по теме:

- распределение полученных данных близко к нормальному;
- коэффициенты корреляции между парами признаков стремятся к нулю.

Библиографический список

1 Язык программирования Python- Режим доступа: https://www.python.org/.

2 Библиотека Pandas- Режим доступа: https://pandas.pydata.org/.

3 Библиотека Matplotlib- Режим доступа: https://matplotlib.org/.

4 Библиотека Seaborn- Режим доступа: https://seaborn.pydata.org/.

5 Библиотека Sklearn- Режим доступа: https://scikit-learn.org.

6 Библиотека Tensorflow: Режим доступа: https://www.tensorflow.org/.