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 Ермолова Дарья Борисовна

Москва, 2023

Оглавление

[**1 Введение** 3](#_Toc133194391)

[**2 Аналитическая часть** 4](#_Toc133194392)

[**2.а Аналитическая часть. Постановка задачи** 4](#_Toc133194393)

[**2.б Аналитическая часть. Описание используемых методов** 6](#_Toc133194394)

[**2.в Аналитическая часть. Разведочный анализ данных** 9](#_Toc133194395)

[**3 Практическая часть** 11](#_Toc133194396)

[**3.а Практическая часть. Предобработка данных** 11](#_Toc133194397)

[**3.б Практическая часть. Разработка и обучение модели** 31](#_Toc133194398)

[**3.в Подбор гиперпараметров моделей** 33](#_Toc133194399)

[**3.г Рекомендательная нейронная сеть для параметра «Соотношение матрица - наполнитель»** 34](#_Toc133194400)

[**3. д Разработка приложения** 40](#_Toc133194401)

[**3.е Создание удаленного репозитория на Github и загрузка результатов работы на него** 45](#_Toc133194402)

[**4 Заключение** 46](#_Toc133194403)

[**5 Библиографический список** 47](#_Toc133194404)

**1 Введение**

Технический прогресс настоящего времени был бы немыслим без широкого распространения полимерных композиционных материалов.

Применение композитов позволяет снизить массу летательных аппаратов, автомобилей, судов, создавать новые конструкции, обладающие высокой работоспособностью и надежностью.

Под композиционными материалами понимают материалы, состоящие из двух и более разнородных по химическому составу, структуре и свойствам компонентов, нерастворимых или малорастворимых друг в друге.

В настоящее время разработаны и применяются различные композиционные материалы: волокнистые, армированные нитевидными кристаллами и непрерывными волокнами тугоплавких соединений и элементов, слоистые материалы, сплавы с упрочнением, дисперсно – упрочненные материалы.

Развитие современных композиционных материалов непосредственно связано с открытием высокой прочности нитевидных кристаллов, разработкой новых армирующих материалов – высокопрочных и высокомодульных непрерывных волокон бора, углерода и волокон других неорганических тугоплавких соединений, а также с исследованием и применением алюминидов, характеризующихся сочетанием высоких прочностных и эксплуатационных свойств.

Композиты формируются в результате объемного сочетания исходных материалов, один из которых образует матрицу, или связующее, а другой является наполнителем. Матрица и наполнитель имеют границу раздела.

Управляя объемным содержанием исходных компонентов, можно получать материалы с требуемыми значениями прочности, жаропрочности, модуля упругости, износостойкости, коррозионной стойкости и другими специальными свойствами.

# **2 Аналитическая часть**

## **2.а Аналитическая часть. Постановка задачи**

Зная 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 матрицы и наполнителя.

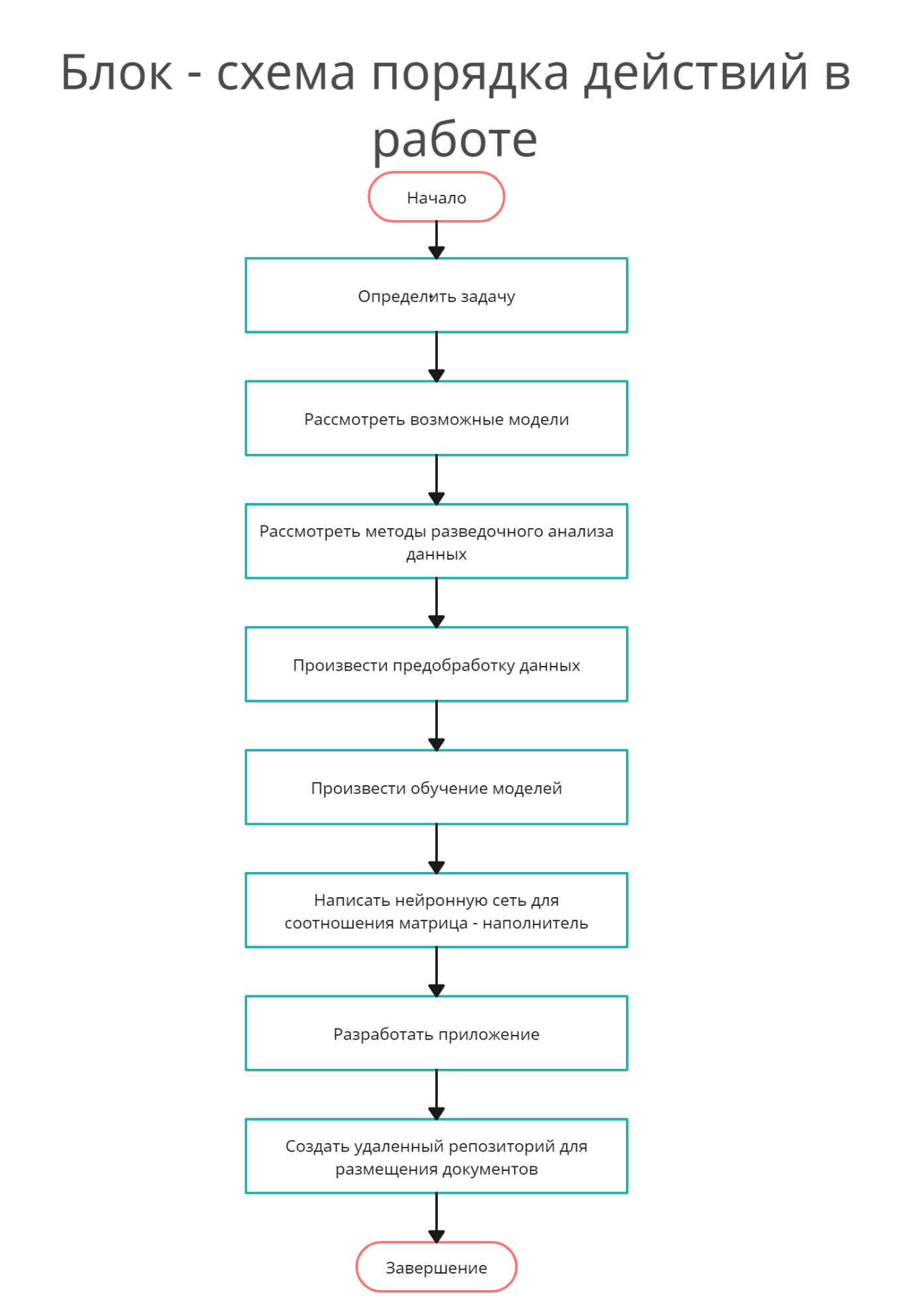
На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать модуль упругости при растяжении, прочность при растяжении и соотношение матрица-наполнитель получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов и цифровыми двойниками новых композитов.

Изначально дано два датасета с характеристиками базальтопластика и нашивки углепластика. Датасеты необходимо объединить по принципу inner.

Определим блок-схему порядка действий в работе:

Блок-схема 1



## **2.б Аналитическая часть. Описание используемых методов**

Поставленная задача является задачей регрессии.

Регрессия – способ выбрать из семейства функций ту, которая будет минимизировать функцию потерь, характеризующую насколько сильно пробная функция отклоняется от значений в заданных точках. Если точки получены в эксперименте, они неизбежно содержат ошибку измерений, шум, поэтому следует, чтобы функция передавала общую тенденцию, а не точно проходила через все точки. По сути, регрессия решает задачу, когда мы хотим провести кривую как можно ближе к точкам и при этом сохранить ее максимально простой, чтобы уловить общую тенденцию. За баланс между этими противоречивыми желаниями и отвечает функция потерь.

Для регрессионных моделей в качестве функции потерь чаще используют:

- среднеквадратичную ошибку (MSE). Это разница между прогнозируемым значением и истинным, возведенная в квадрат и усредненная по всему набору данных;

- средняя абсолютная ошибка (MAE). Это средняя абсолютная разность между предсказанием модели и целевым значением;

- коэффициент детерминации (R2) отражает объясняющую способность регрессии и определяется как доля дисперсии зависимой переменной, объяснённая регрессионной моделью с данным набором независимых переменных. Обычно определяется как единица минус доля необъяснённой дисперсии.

Рассмотрим некоторые алгоритмы машинного обучения задач регрессии:

1. Линейная регрессия (Linear regression) прогнозирует результат, основанный на непрерывных функциях. Метод универсален в том смысле, что он имеет возможность запускаться с одной входной переменной (простая линейная регрессия) или с зависимостью от нескольких переменных (множественная регрессия). Суть алгоритма заключается в назначении оптимальных весов для переменных, чтобы создать линию, которая будет использоваться для прогнозирования вывода.
2. Лассо регрессия (LASSO, Least Absolute Shrinkage and Selection Operator) — это вариация линейной регрессии, специально адаптированная для данных, которые демонстрируют сильную мультиколлинеарность (то есть сильную корреляцию признаков друг с другом).

Она автоматизирует части выбора модели, такие как выбор переменных или исключение параметров. LASSO использует сжатие коэффициентов (shrinkage), то есть процесс, в котором значения данных приближаются к центральной точке (например, к среднему значению).

1. Гребневая регрессия (Ridge regression) очень похожа на регрессию LASSO в том, что она применяет сжатие. Оба алгоритма хорошо подходят для наборов данных с большим количеством признаков, которые не являются независимыми друг от друга.

Однако самое большое различие между ними в том, что гребневая регрессия использует регуляризацию L2, то есть ни один из коэффициентов не становится нулевым, как это происходит в регрессии LASSO. Вместо этого коэффициенты всё больше приближаются к нулю, но не имеют большого стимула достичь его из-за природы регуляризации L2. Гребневая регрессия лучше подходит в ситуации, когда мы хотим сделать приоритетными большое количество переменных, каждая из которых имеет небольшой эффект. Если в модели требуется учитывать несколько переменных, каждая из которых имеет средний или большой эффект, лучшим выбором будет лассо.

1. Метод К – ближайших соседей (KNN) основан на предположении о том, что близким объектам в признаковом пространстве соответствуют похожие метки. Объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны, прогноз строится по их меткам.
2. Метод опорных векторов (Support Vector Machine) подразумевает поиск гиперплоскости, при которой риск в многомерном пространстве будет минимальным. Алгоритм оценивает коэффициенты путем минимизации квадратичных потерь. Так, если прогнозное значение попадает в область гиперплоскости, то потери равны нулю, в противном случае - разности прогнозного и фактического значений.
3. Дерево решений (Decision trees) в регрессии работают путем построения деревьев с узлами «да/нет». Конечные узлы приводят к значению в непрерывном режиме (например, 4593,49 или 10,98).

Из-за специфической и высокодисперсной природы регрессии просто как задачи машинного обучения, регрессоры дерева решений следует тщательно обрезать. Тем не менее, подход к регрессии нерегулярен — вместо того, чтобы вычислять значение в непрерывном масштабе, он приходит к заданным конечным узлам. Если регрессор обрезан слишком сильно, у него слишком мало конечных узлов, чтобы должным образом выполнить свою задачу.

Следовательно, дерево решений должно быть обрезано так, чтобы оно имело наибольшую свободу (возможные выходные значения регрессии — количество конечных узлов), но недостаточно, чтобы оно было слишком глубоким. Если его не обрезать, то и без того высокодисперсный алгоритм станет чрезмерно сложным из-за природы регрессии.

1. Случайный лес (Random forest) близкий родственник дерева решения. Регрессор случайного леса может работать лучше или не лучше, чем дерево решений в регрессии (в то время как он обычно работает лучше в классификации) из-за тонкого баланса между избыточным и недостаточным в природе алгоритмов построения дерева.
2. Градиентный бустинг (Gradient Boosting) строит предсказание в виде ансамбля слабых предсказывающих моделей, которыми в основном являются деревья решений. Из нескольких слабых моделей в итоге собирается одна, но уже эффективная. Общая идея алгоритма – последовательное применение предсказателя таким образом, что каждая последующая модель сводит ошибку предыдущей к минимуму.
3. Нейронная сеть (Neural network) обучается на тренировочном наборе данных, находя веса (коэффициенты) между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными параметрами и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения Нейросеть способна выдать верный результат на основании данных, которые отсутствовали в обучающей выборке, а также на неполных и/или «зашумленных», частично искажённых данных. Обучение сети происходит итерационно. При каждой итерации считывается весь обучающий набор данных и изменяются веса Нейросети. Этот процесс продолжается до тех пор, пока относительные изменения весов не станут меньше заданного порога или количество итераций не превысит заданной величины.

## **2.в Аналитическая часть. Разведочный анализ данных**

Разведочный анализ данных (exploratory data analysis, EDA) – это изучение данных для принятия решений по поводу их применения, очистки, преобразования и конструирования новых признаков.

В работе применены следующие средства разведочного анализа:

- гистограмма распределения — наглядное представление функции плотности вероятности некоторой случайной величины, построенное по выборке. Иногда её называют частотным распределением, так как гистограмма показывает частоту появления измеренных значений параметров объекта;

- межквартильный размах, отображающий одномерное распределение вероятности и визуализирующий аномальные значения, принимаемые параметрами. Диаграмма «ящик с усами» (box-plot) показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы;

- удаление выбросов данных производится с применением анализа стандартизированной оценки (меры относительного разброса наблюдаемого или измеренного значения, показывающей, сколько стандартных отклонений составляет разброс относительного среднего значения) и анализом межквартильного диапазона ( IQR, представляет собой разницу между 25-м процентилем (Q1) и 75-м процентилем (Q3) в наборе данных. Он измеряет разброс средних 50 процентов значений. Метод состоит в том, чтобы объявить наблюдение выбросом, если его значение в 1,5 раза больше, чем IQR, или в 1,5 раза меньше, чем IQR);

- нормализация данных приводит значения к некоторой общей шкале без потери информации о различии диапазонов, что дает возможность исключить некорректные зависимости, основанные на влиянии входных переменных на выходную. Часто для нормализации используют метод Бокса-Кокса, предполагающий работу только с положительными величинами, или метод Йео-Джонсона, позволяющий работу еще и с отрицательными величинами;

- анализ параметров на мультиколлинеарность дает понимание о наличии линейной зависимости между параметрами, если парная зависимость высокая, оба параметра не следует включать в модель регрессии. Визуально оценить коллинеарность факторов можно при помощи тепловой карты Пирсона.

# **3 Практическая часть**

## **3.а Практическая часть. Предобработка данных**

Для работы даны два файла со свойствами композитов (базальтопластик и нашивка углепластика). Загрузим данные с применением метода pandas. Ознакомимся с первыми пятью строками и просмотрим описательную статистику для оценки размеров датасетов.

Таблица 1 – Первые пять строк датасета базальтопластика

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 |
| Соотношение матрица-наполнитель | 1.857143 | 1.857143 | 1.857143 | 1.857 | 2.771 |
| Плотность, кг/м3 | 2030.0 | 2030.0 | 2030.0 | 2030.0 | 2030.0 |
| модуль упругости, ГПа | 738.736 | 738.73 | 738.736 | 738.736 | 753 |
| Количество отвердителя, м.% | 30 | 50 | 49.90 | 129 | 111.86 |
| Содержание эпоксидных групп,%\_2 | 22.267857 | 23.75 | 33 | 21.25 | 22.267 |
| Температура вспышки, С\_2 | 100 | 284.615 | 284.615 | 300 | 284.615 |
| Поверхностная плотность, г/м2 | 210.0 | 210.0 | 210.0 | 210.0 | 210.0 |
| Модуль упругости при растяжении, ГПа | 70.0 | 70.0 | 70.0 | 70.0 | 70.0 |
| Прочность при растяжении, МПа | 3000.0 | 3000.0 | 3000.0 | 3000.0 | 3000.0 |
| Потребление смолы, г/м2 | 220.0 | 220.0 | 220.0 | 220.0 | 220.0 |

Таблица 2 – Описательная статистика датасета базальтопластика

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | | Mean | std | min | 25% | 50% | 75% | max |
| Соотношение матрица-наполнитель | 1023 | | 2.493 | 0.913 | 0.389 | 2.317 | 2.906 | 3.552 | 5.591 |
| Плотность, кг/м3 | 1023 | | 1975.734 | 73.729 | 1731.764 | 1924.155 | 1977.621 | 2021.374 | 2207.77 |
| модуль упругости, ГПа | 1023 | | 739.923 | 330.231 | 2.436 | 500.047 | 739.664 | 961.812 | 1911.53 |
| Количество отвердителя, м.% | 1023 | | 110.570 | 28.295 | 17.74 | 92.443 | 110.564 | 129.73 | 198.953 |
| Содержание эпоксидных групп,%\_2 | 1023 | | 22.244 | 2.406 | 14.254 | 20.608 | 22.23 | 23.961 | 33. |
| Температура вспышки, С\_2 | 1023 | | 285.882 | 40.943 | 100 | 259.066 | 285.896 | 313.002 | 413.273 |
| Поверхностная плотность, г/м2 | 1023 | 482.731 | | 281.314 | 0.603 | 266.816 | 451.864 | 693.225 | 1399.54 |

Продолжение таблицы 2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | Mean | std | min | 25% | 50% | 75% | max |
| Модуль упругости при растяжении, ГПа | 1023 | 73.328 | 3.118 | 64.054 | 71.245 | 73.268 | 75.356 | 82.682 |
| Прочность при растяжении, МПа | 1023 | 2466.922 | 485.628 | 1036.856 | 2135.85 | 2459.524 | 2767.193 | 3848.43 |
| Потребление смолы, г/м2 | 1023 | 218.423 | 9.735 | 33.803 | 179.627 | 219.198 | 257.481 | 414.59 |

Таблица 3 - Первые пять строк датасета нашивки углепластика

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 |
| Угол нашивки, град | 0 | 0 | 0 | 0 | 0 |
| Шаг нашивки | 4.0 | 4.0 | 4.0 | 5.0 | 5.0 |
| Плотность нашивки | 57.0 | 60.0 | 70.0 | 47.0 | 57.0 |

Таблица 4 - Описательная статистика датасета нашивки углепластика

|  |  |  |  |
| --- | --- | --- | --- |
|  | Угол нашивки, град | Шаг нашивки | Плотность нашивки |
| count | 1040 | 1040 | 1040 |
| mean | 45 | 6.911 | 57.248 |
| std | 45.021 | 2.555 | 12.332 |
| min | 0 | 0 | 0 |
| 25% | 0 | 5.102 | 49.97 |
| 50% | 45 | 6.938 | 57.413 |
| 75% | 90 | 8.587 | 65.107 |
| max | 90 | 14.44 | 103.988 |

Заметим: количество строк в файлах разное, однако требуется объединить датасеты по принципу INNER, то есть число строк, которое должно получиться, будет равно 1023, число столбцов/параметров - 13.

В объединенном датасете просмотрим:

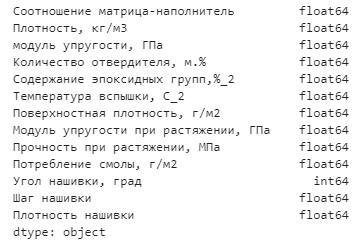


Рисунок 1 – Тип данных объединенного датасета

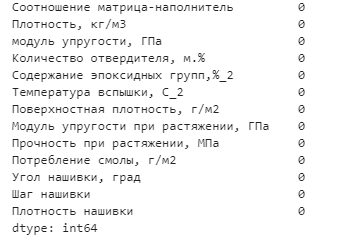


Рисунок 2 – Просмотр количества пропущенных значений

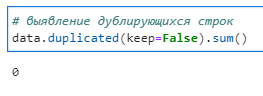


Рисунок 3 – Выявление дублирующихся строк

При анализе данных получившегося файла, можно сказать следующее:

- пропуски в данных не наблюдаются

- дубликаты записей отсутствуют

- все значения являются числовыми

- показатель "Угол нашивки" принимает только два значения "0" и "90" градусов. При исследовании предметной области выяснилось, что этот параметр может также принимать значения "-45" и "45" градусов. Возникла дилемма: следует ли идентифицировать переменную как категориальную и вводить два дополнительных класса, данными которого мы не располагаем, или не применять кодирование. Было принято решение оставить параметр вещественным, исходя из его физического смысла.

- странным фактом является наличие целочисленных данных в первых двадцати трех строках датасета, в то время как большинство измерений производится в формате "с плавающей точкой". Можно предположить, часть данных искусственно синтезирована, однако положимся на компетенции составителей датасета.

Объединенный датасет был сохранен для последующей работы.

Построим для каждого признака гистограмму для определения плотности вероятности распределения данных и диаграмму «ящик с усами» для отображения одномерного распределения вероятности и визуализации выбросов.

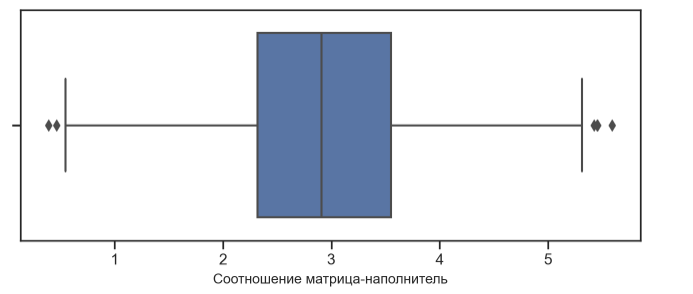


Рисунок 4 – Гистограмма и диаграмма распределения для параметра «Соотношение матрица - наполнитель»

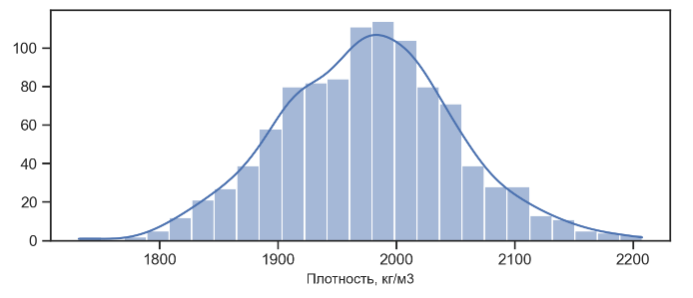
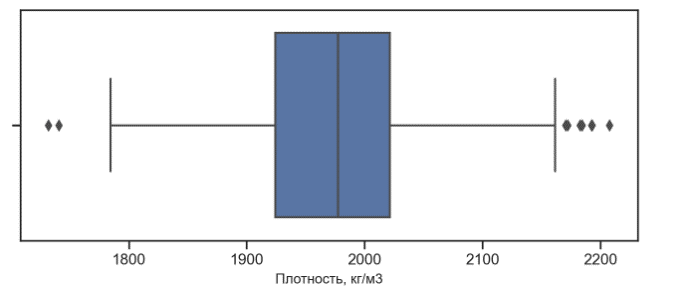
 

Рисунок 5 – Гистограмма и диаграмма распределения для параметра «Плотность, кг/м3»

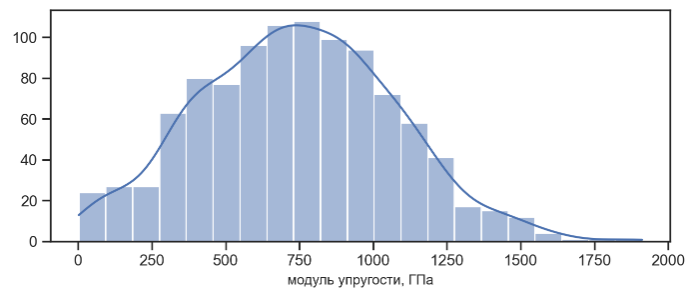
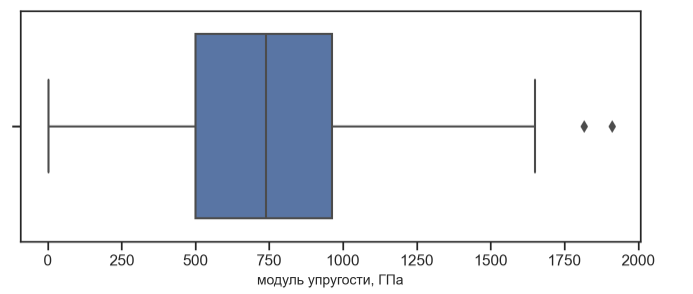
 

Рисунок 6 – Гистограмма и диаграмма распределения для параметра «модуль упругости, ГПа»

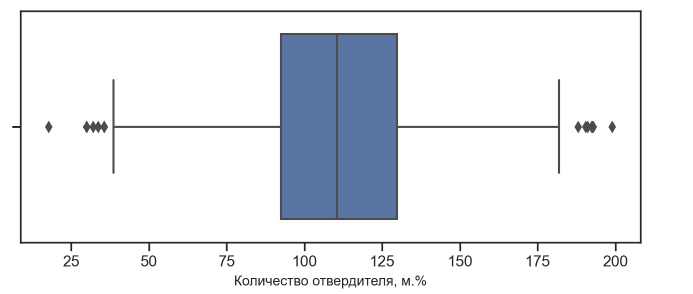
 

Рисунок 7 – Гистограмма и диаграмма распределения для параметра «Количество отвердителя, м.%»

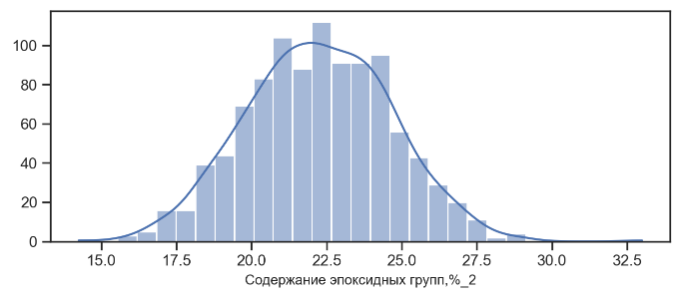
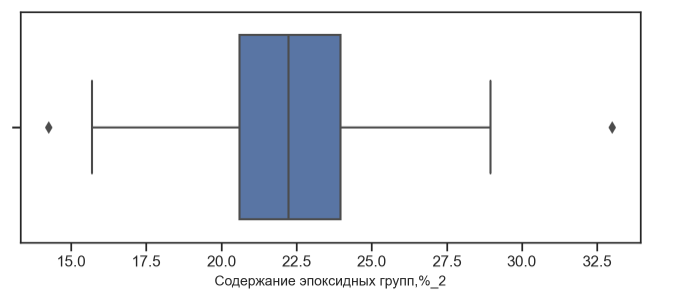
 

Рисунок 8 – Гистограмма и диаграмма распределения для параметра «Содержание эпоксидных групп, % 2»

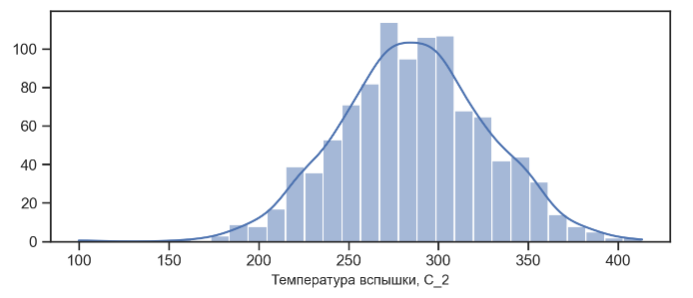
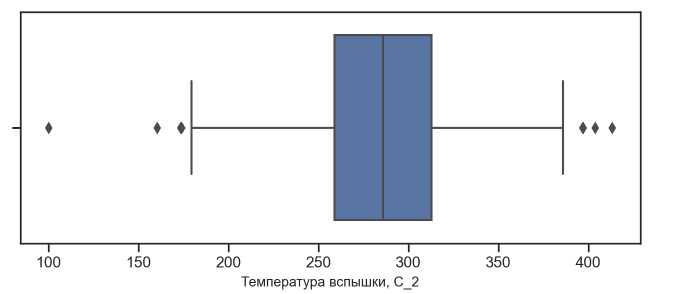
 

Рисунок 9 – Гистограмма и диаграмма распределения для параметра «Температура вспышки, С 2»

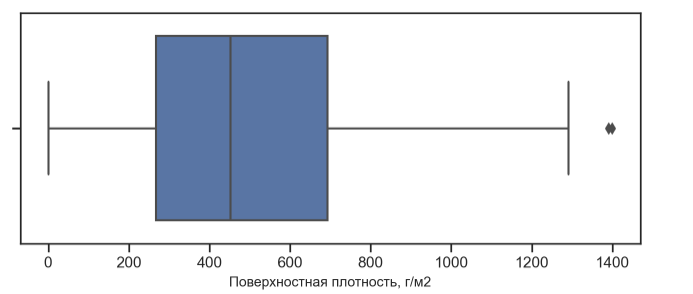
 

Рисунок 10 – Гистограмма и диаграмма распределения для параметра «Поверхностная плотность, г/м2»

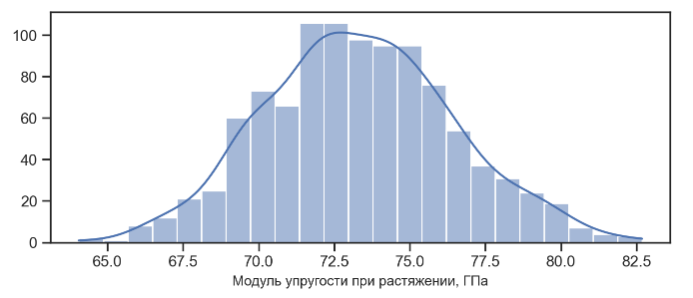
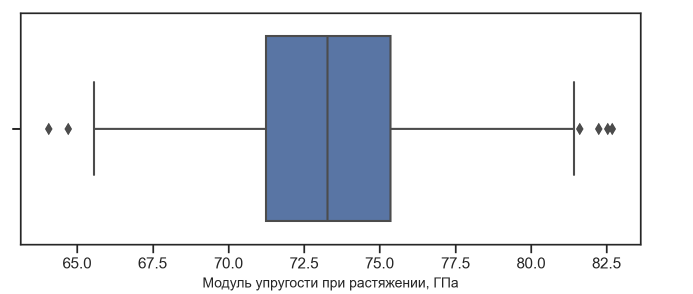
 

Рисунок 11 – Гистограмма и диаграмма распределения для параметра

«Модуль упругости, ГПа»

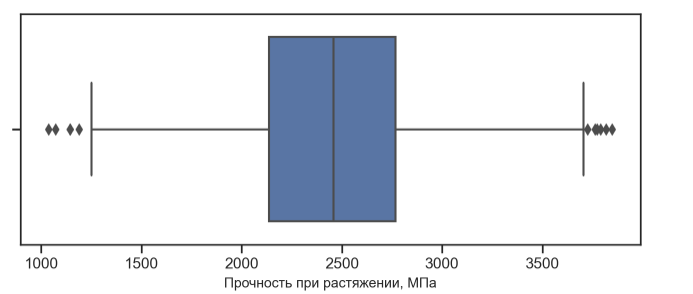
 

Рисунок 12 – Гистограмма и диаграмма распределения для параметра

«Прочность при растяжении, МПа»

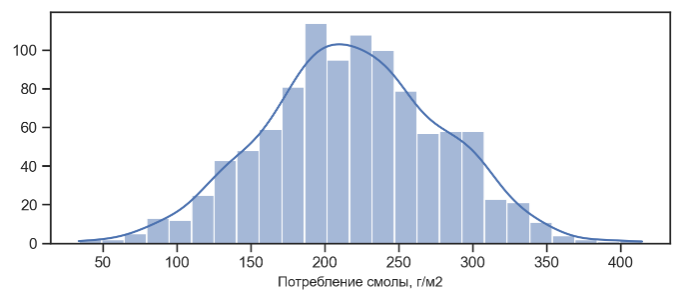
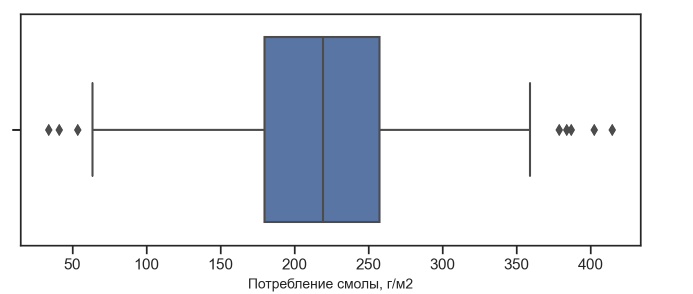
 

Рисунок 13 – Гистограмма и диаграмма распределения для параметра

«Потребление смолы, г/м2»

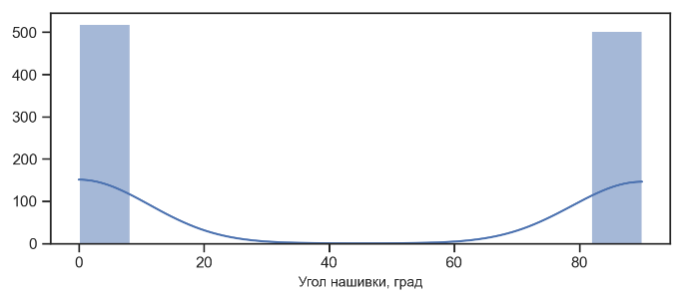
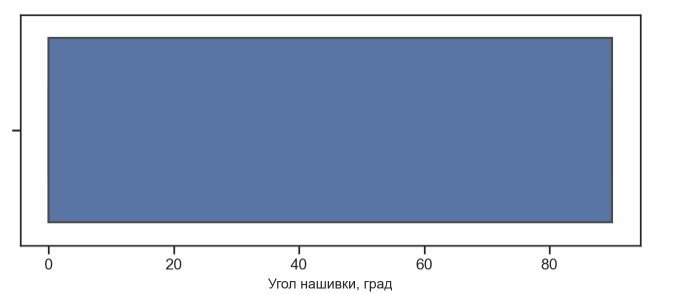
 

Рисунок 14 – Гистограмма и диаграмма распределения для параметра

«Угол нашивки, град»

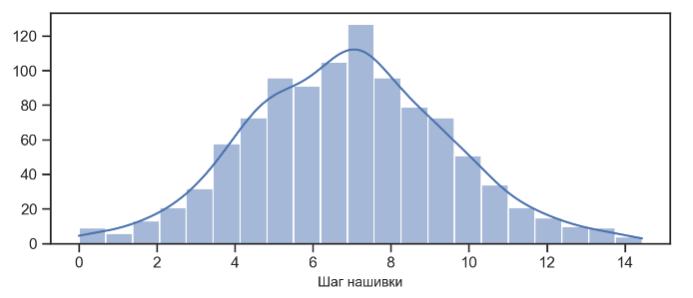
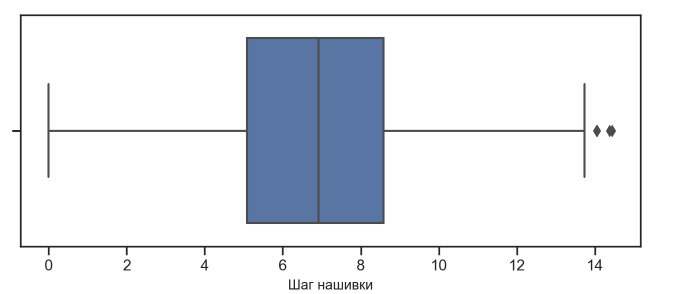
 

Рисунок 15 – Гистограмма и диаграмма распределения для параметра

«Шаг нашивки»

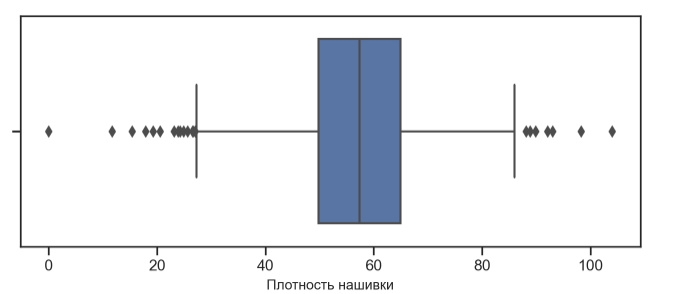
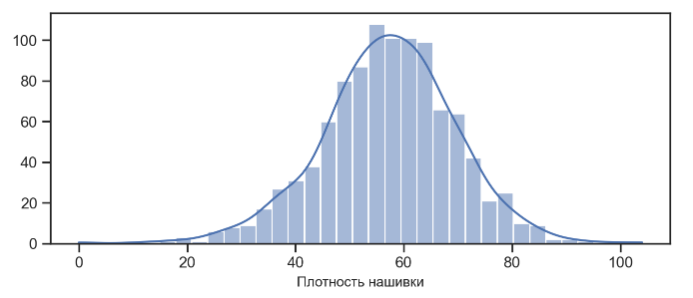


Рисунок 16 – Гистограмма и диаграмма распределения для параметра

«Плотность нашивки»

На большинстве гистограмм на первый взгляд можно наблюдать в целом нормальное распределение, однако диаграммы распределения указывают на наличие смещения и наличие аномалий/выбросов. В некоторых случаях датасет можно было бы оставить без изменений, однако мы предполагаем, что в нашем случае это неуместно. Посмотрим на количество выбросов при использовании z- оценки и методе IQR.

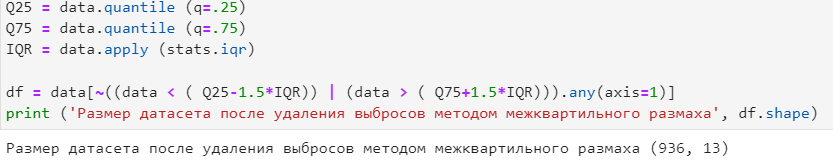
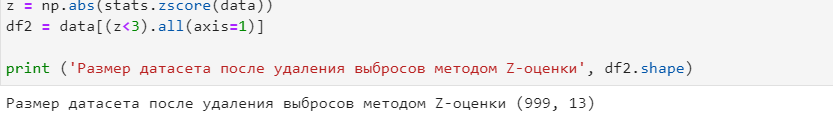


Рисунок 17 – Применение z-оценки и межквартильного размаха

При расчете z-оценки количество выбросов составляет 23 строки, при межквартильном размахе - 86, что эквивалентно ~12% всех данных. Несмотря на то, что это внушительная часть данных, применим отсечение выбросов этим параметром, основываясь на том, что в целом гистограммы распределены нормально, но местами присутствует смещение, что предполагает использование данного метода. Предположительно после исключения выбросов сформируются новые аномалии, однако после первого исключения выбросов как минимум удалены ошибки в данных и аномалии, а дальнейшая очистка может привести к искажению модели.

Нормализация данных будет проведена при помощи метода Йео-Джонсона, так как этот метод может быть применен в случае, когда имеются отрицательные значения, а при анализе сущности параметра «Угол нашивки» мы условились, что природа признака вещественная и он может принимать на вход в том числе отрицательное и нулевое значение.

Проанализируем очищенные данные на нормальность при применении преобразования Йео-Джонсона следующими инструментами:

- построение гистограмм до/после;

- построение графиков QQ (квантиль-квантиль);

- проведение теста на нормальность Шапиро – Уилка.

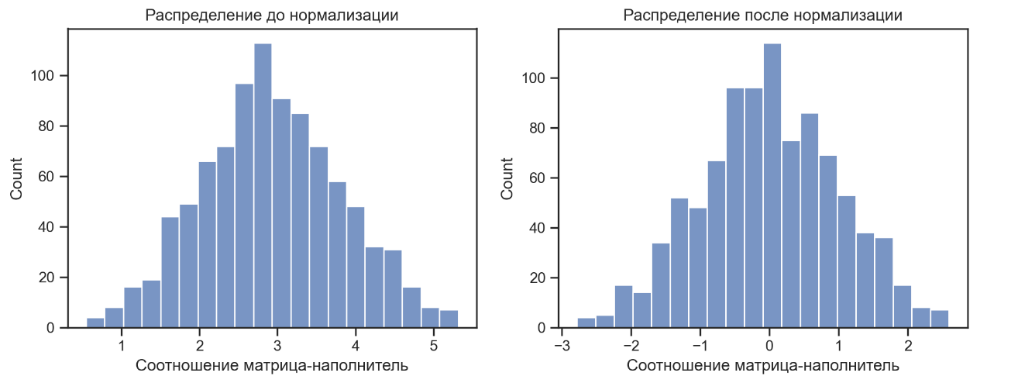


Рисунок 18 – Гистограмма до/после для параметра «Соотношение матрица - наполнитель»

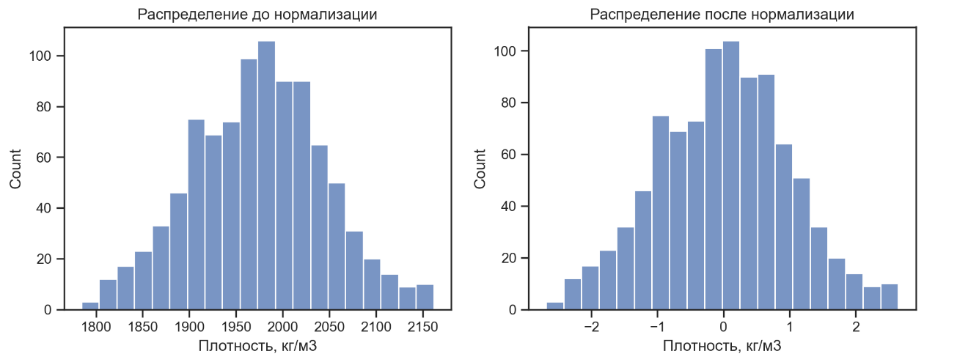


Рисунок 19 – Гистограмма до/после для параметра «Плотность, кг/м3»

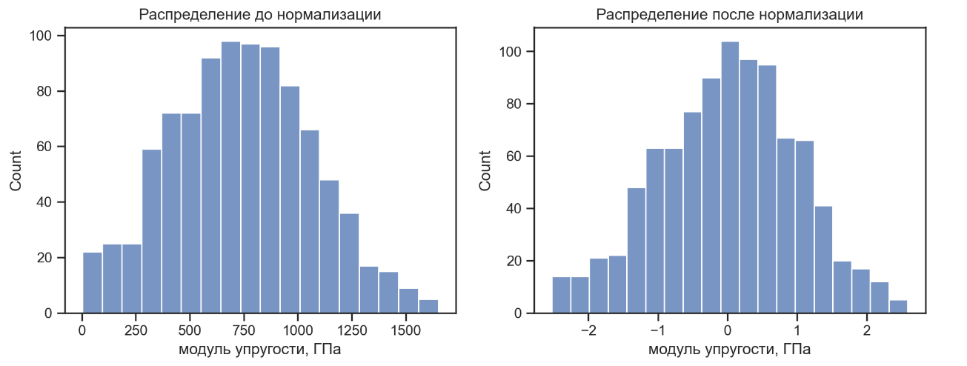


Рисунок 20 – Гистограмма до/после для параметра «модуль упругости, ГПа»

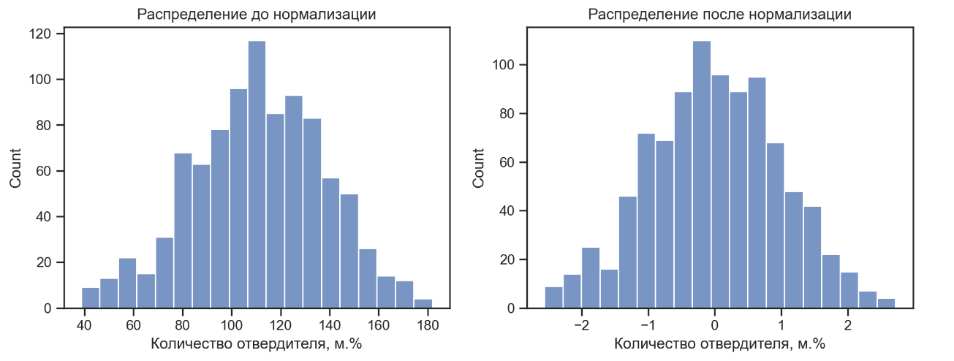


Рисунок 21 – Гистограмма до/после для параметра «Количество отвердителя, м.%»

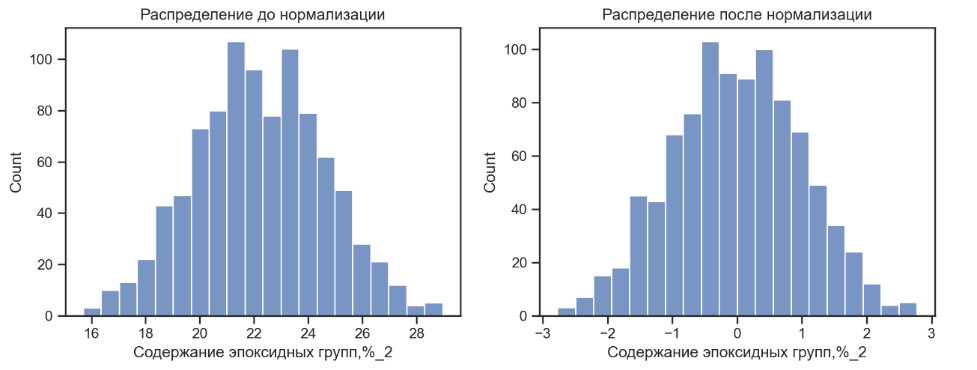


Рисунок 22 – Гистограмма до/после для параметра «Содержание эпоксидных групп, % 2»

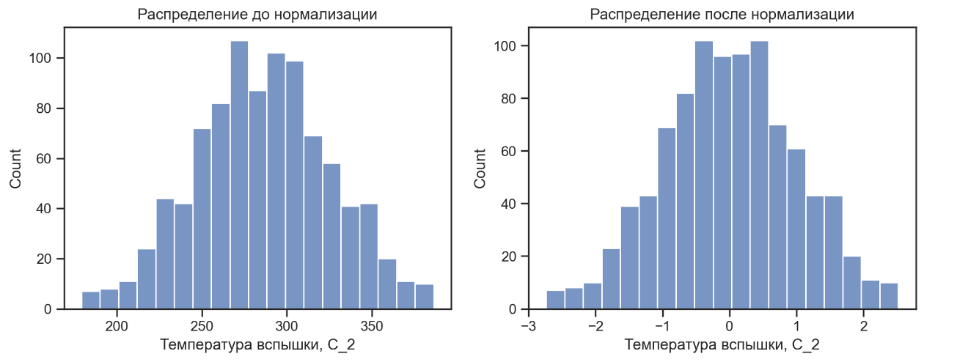


Рисунок 23 – Гистограмма до/после для параметра «Температура вспышки, С 2»

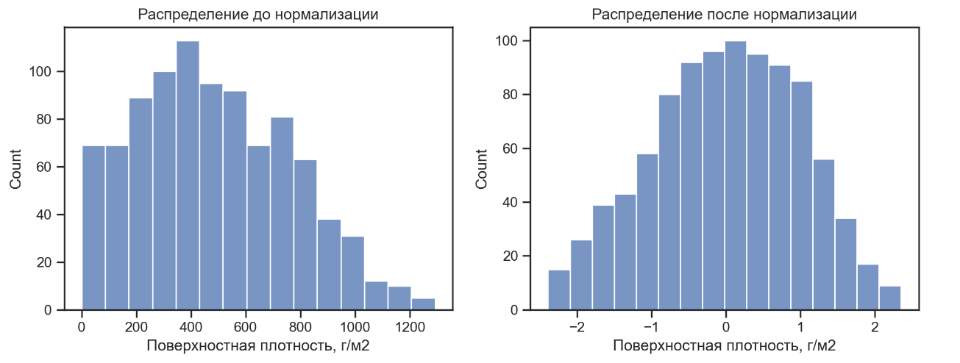


Рисунок 24 – Гистограмма до/после для параметра «Поверхностная плотность, г/м2»

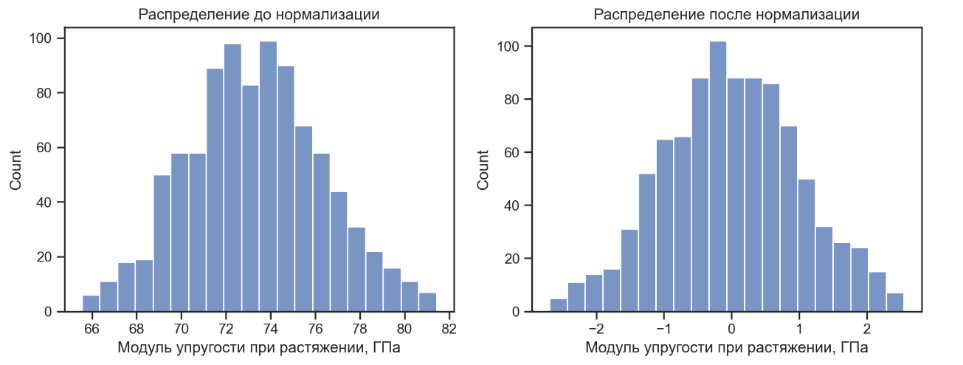


Рисунок 25 – Гистограмма до/после для параметра «Модуль упругости, ГПа»

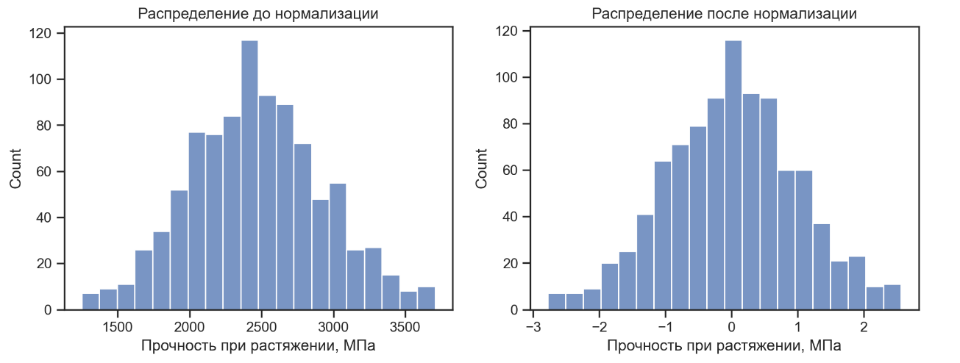


Рисунок 26 – Гистограмма до/после для параметра «Прочность при растяжении, МПа»

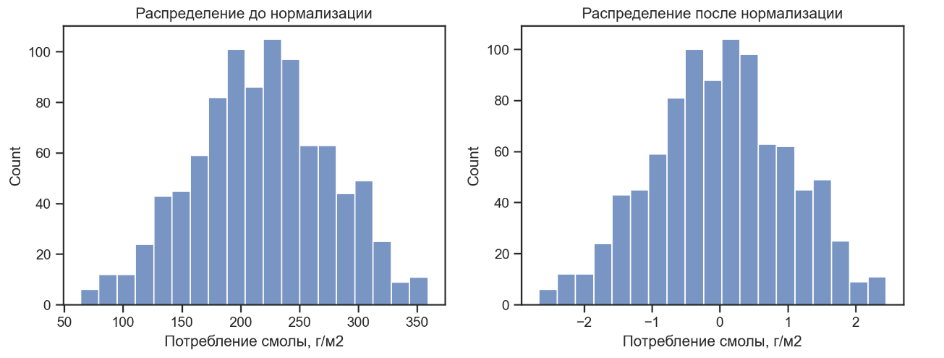


Рисунок 27 – Гистограмма до/после для параметра «Потребление смолы, г/м2»

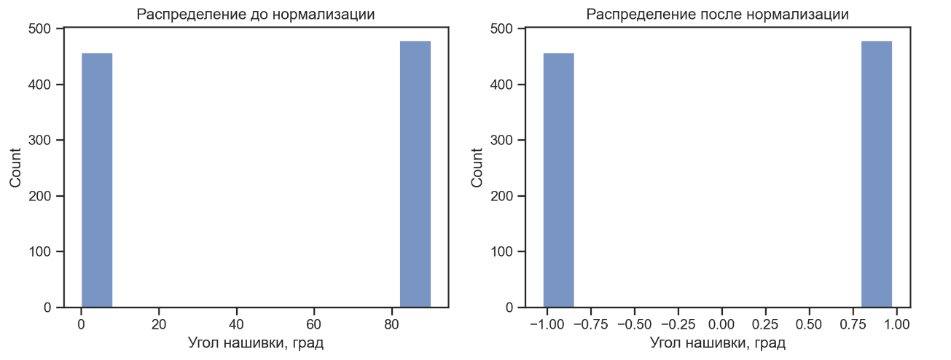


Рисунок 28 – Гистограмма до/после для параметра «Угол нашивки, град»

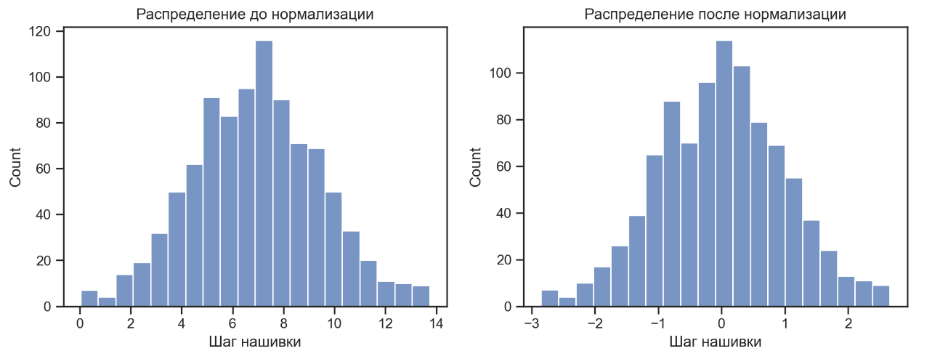


Рисунок 29 – Гистограмма до/после для параметра «Шаг нашивки»

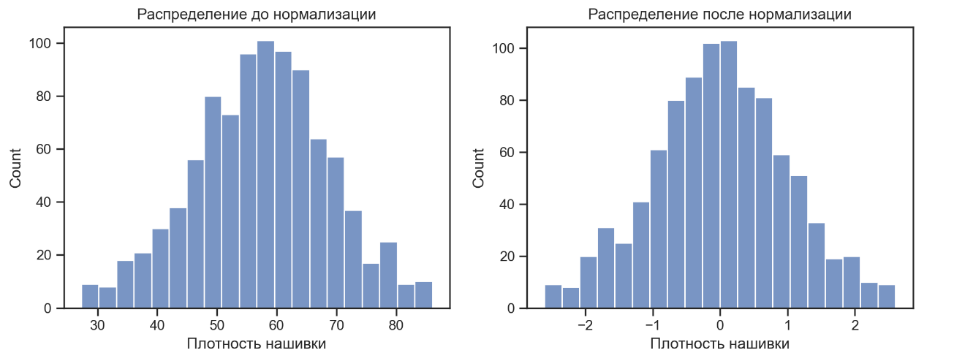


Рисунок 30 – Гистограмма до/после для параметра «Плотность нашивки»

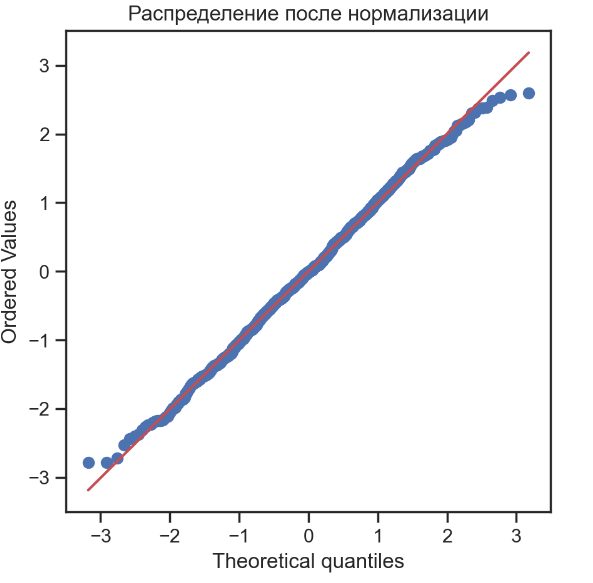
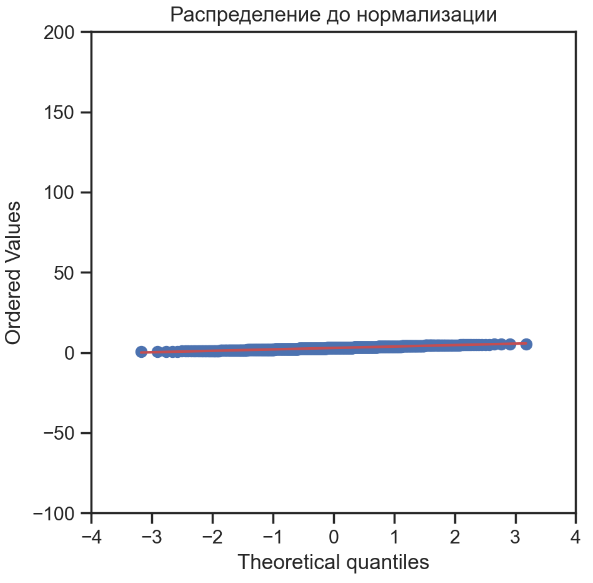


Рисунок 31 – QQ-график до и после для параметра «Соотношение матрица - наполнитель»

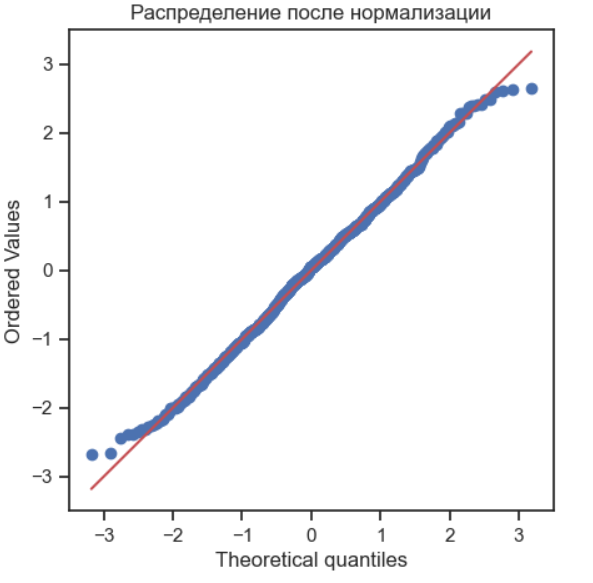
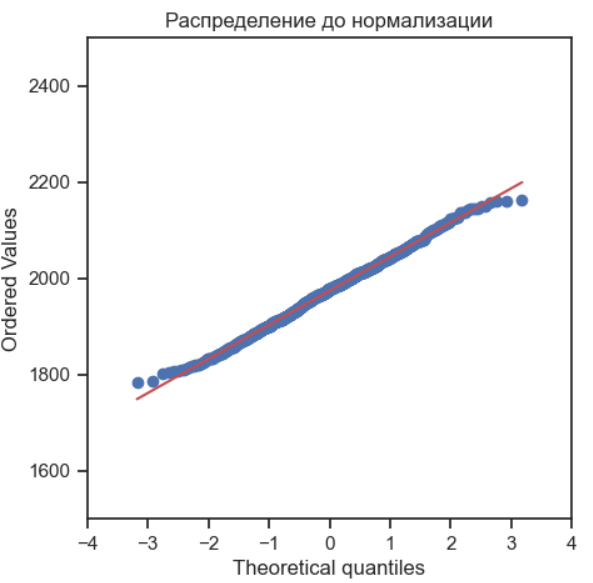


Рисунок 32 – QQ-график до и после для параметра «Плотность, кг/м3»

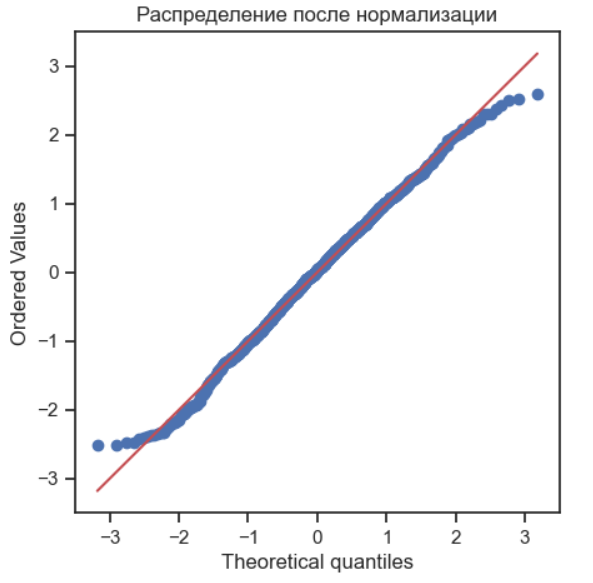
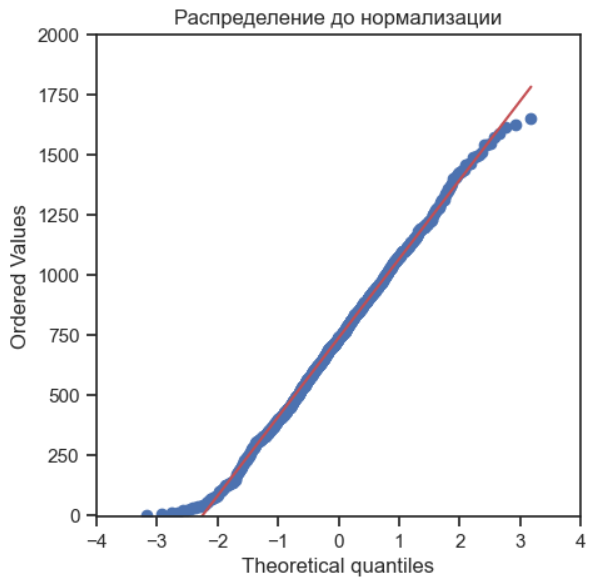


Рисунок 33 – QQ-график до и после для параметра «модуль упругости, ГПа»

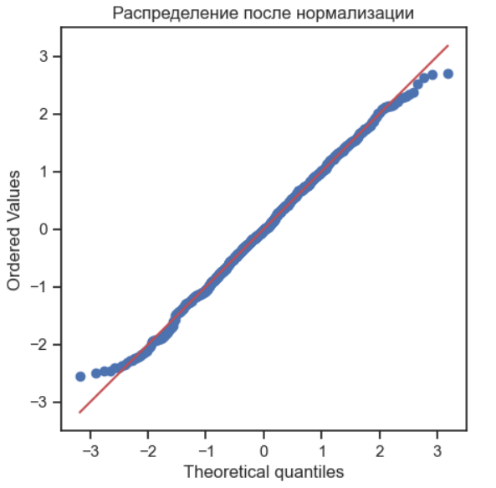
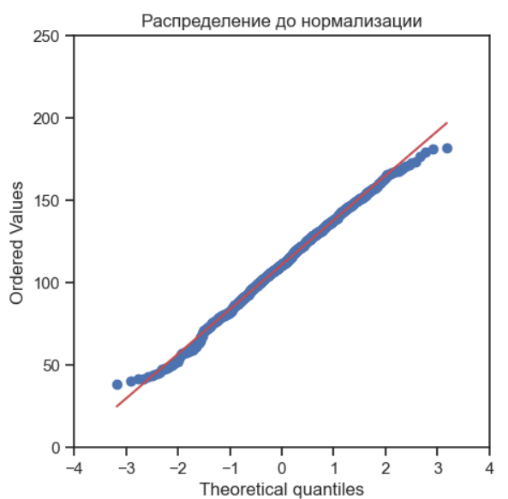


Рисунок 34 – QQ-график до и после для параметра «Количество отвердителя, м.%»

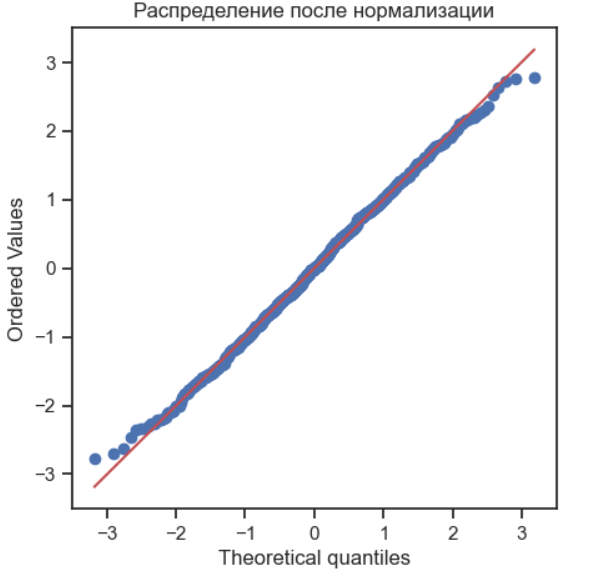
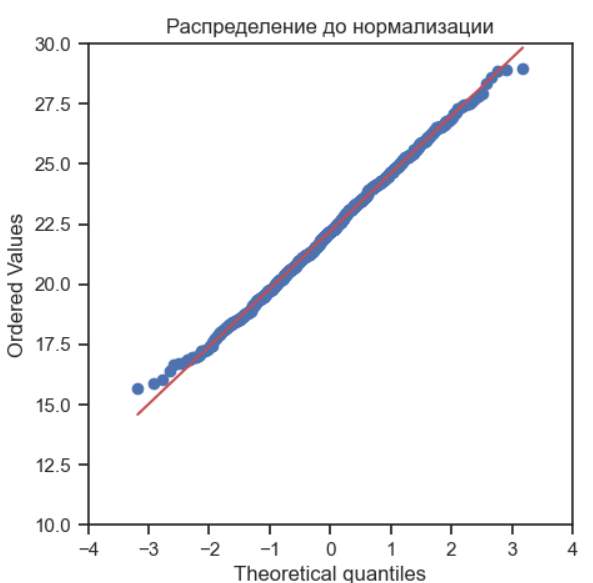


Рисунок 35 – QQ-график до и после для параметра «Содержание эпоксидных групп, % 2»

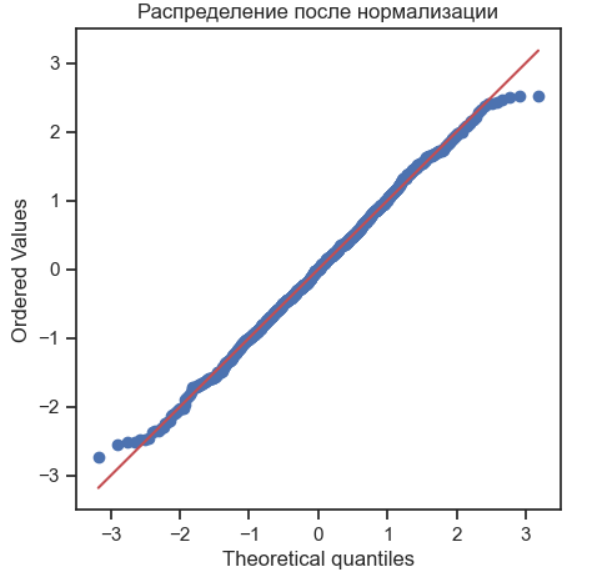
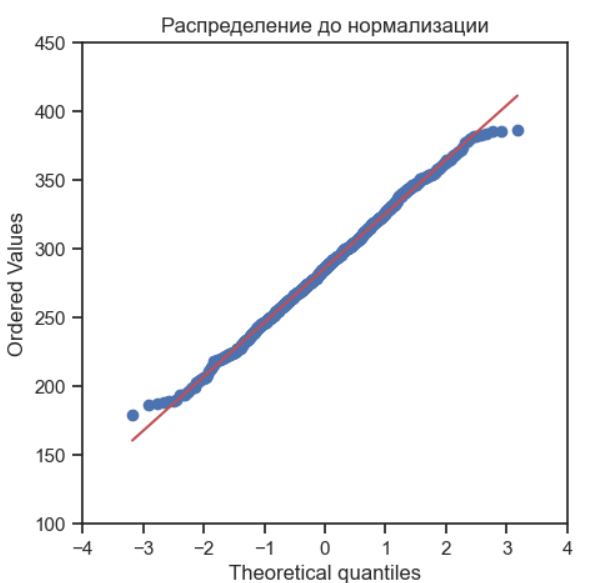


Рисунок 36 – QQ-график до и после для параметра «Температура вспышки, С2»

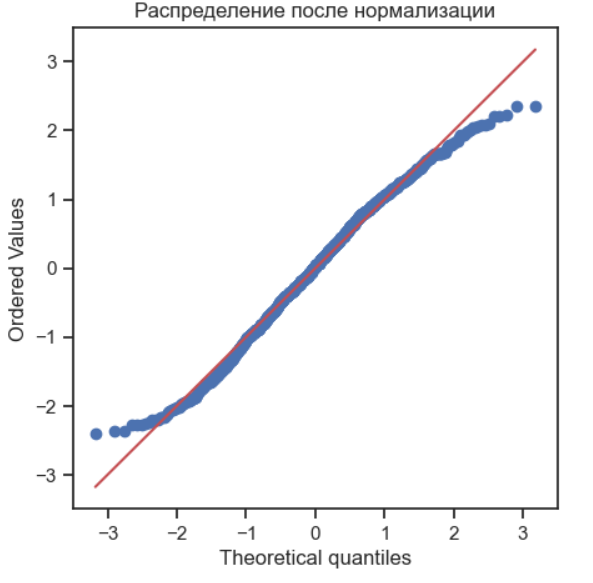
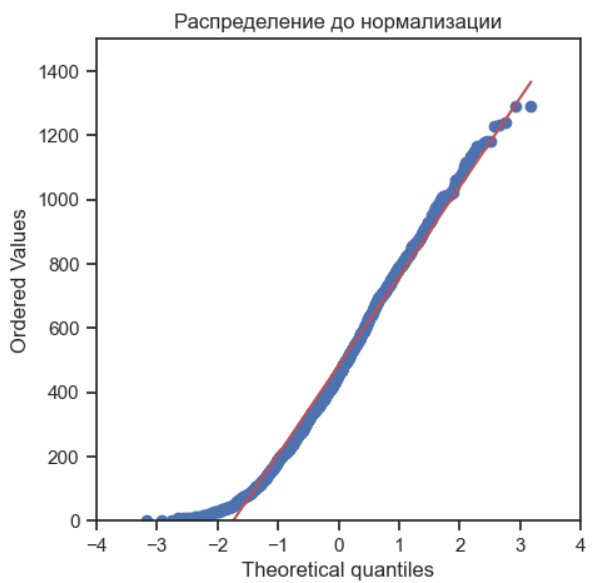


Рисунок 37 – QQ-график до и после для параметра «Поверхностная плотность, г/м2»

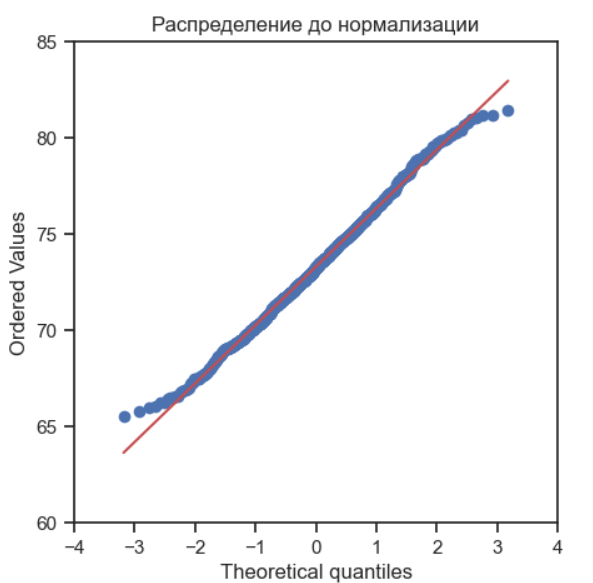


Рисунок 38 – QQ-график до и после для параметра «Модуль упругости, ГПа»

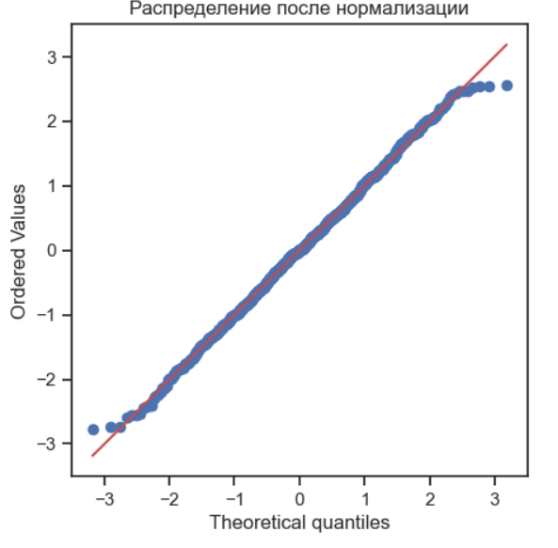
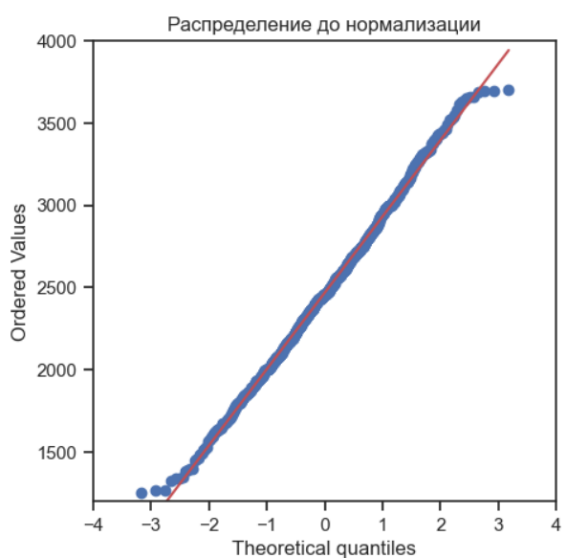


Рисунок 39 – QQ-график до и после для параметра «Прочность при растяжении, МПа»

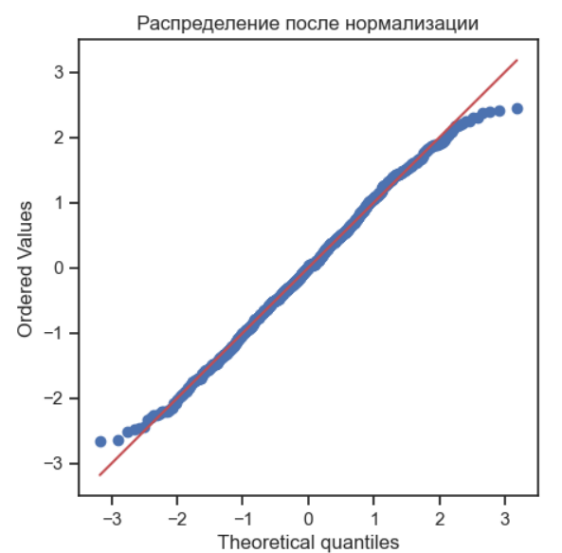
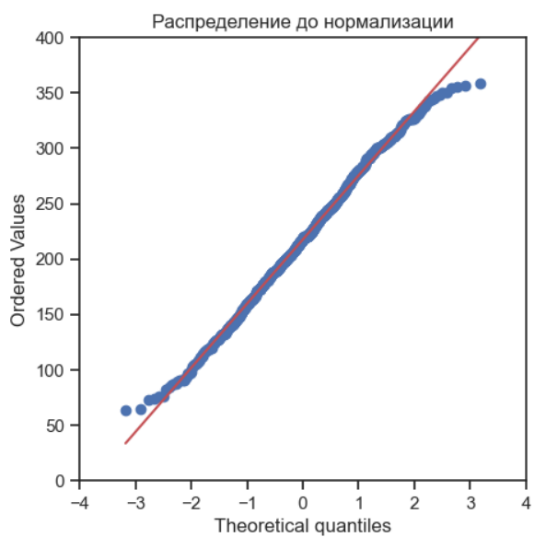


Рисунок 40 – QQ-график до и после для параметра «Потребление смолы, г/м2»

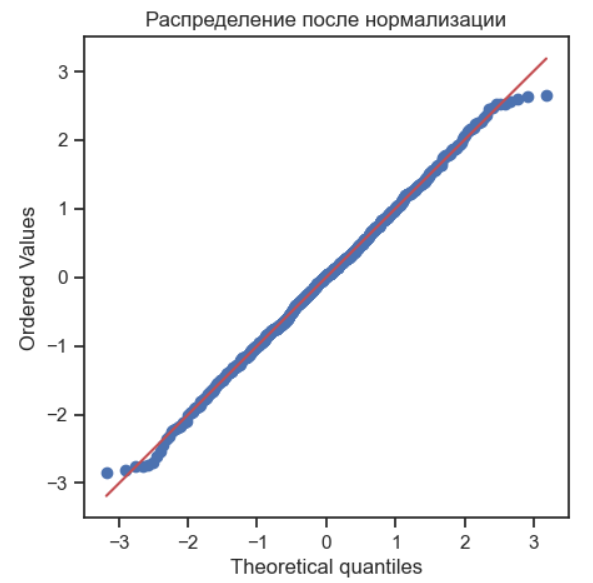
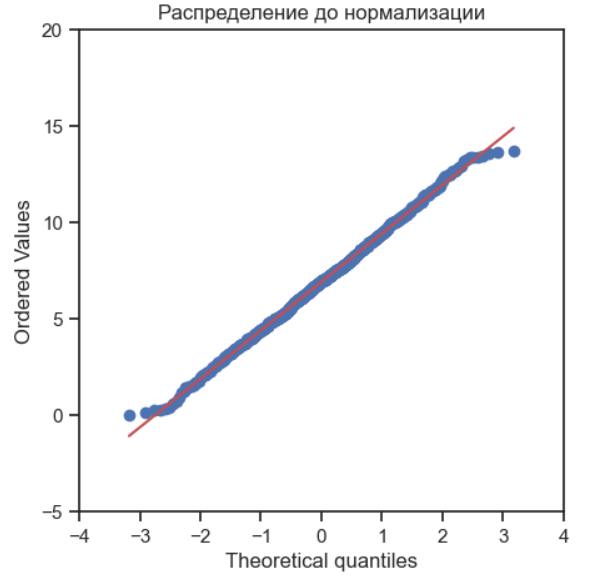


Рисунок 41 – QQ-график до и после для параметра «Шаг нашивки»

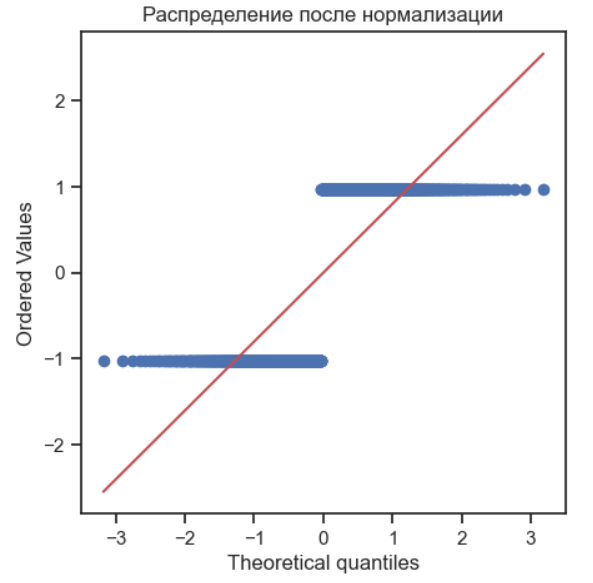
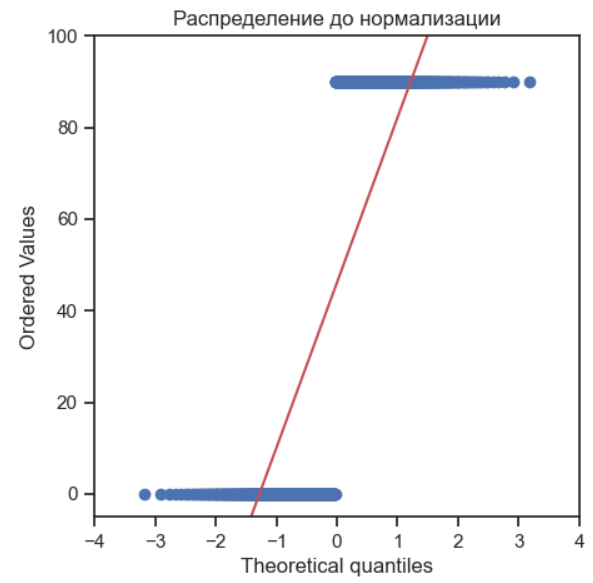


Рисунок 42 – QQ-график до и после для параметра «Угол нашивки, град»

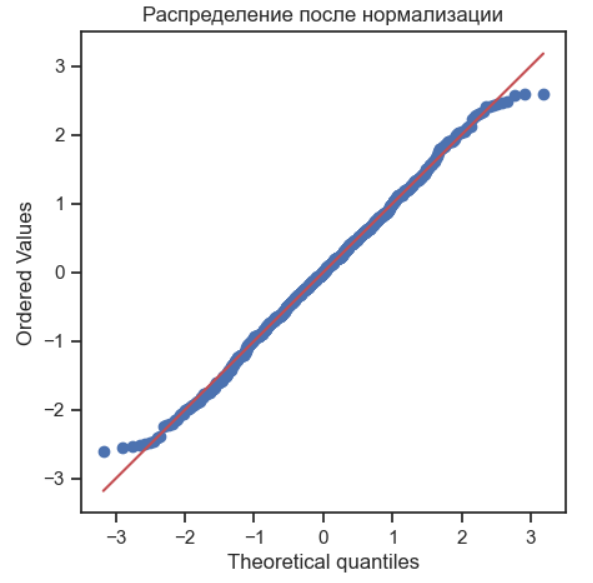
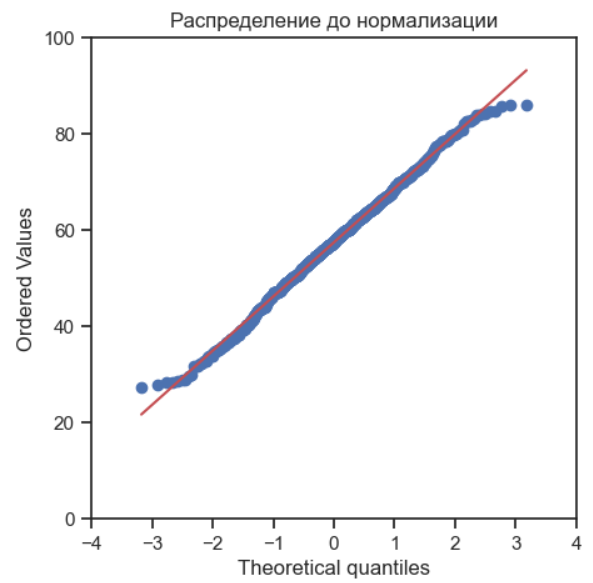


Рисунок 43 –для параметра «Плотность нашивки»

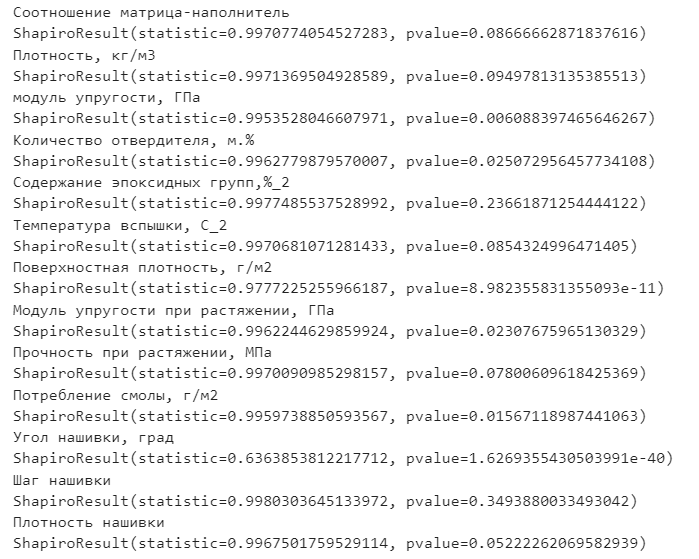


Рисунок 44 – Тест Шапиро- Уилка на нормальность датасета до нормализации

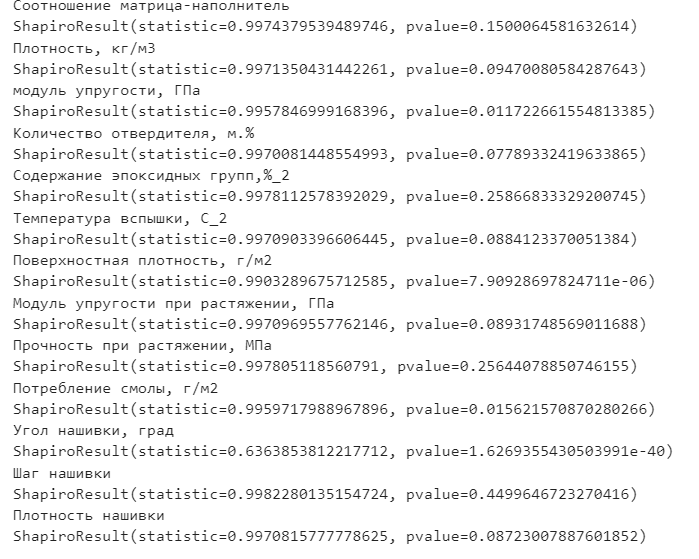


Рисунок 45 – Тест Шапиро- Уилка на нормальность датасета после нормализации Йео-Джонсона

Тест Шапиро – Уилка показывет, что после нормализации 4 из 13 показателей не проходят пороговое значение p-value < 0.05, одним из которых является "Угол нашивки", другие:

- модуль упругости, Гпа;

- поверхностная плотность, г/м2;

- потребление смолы, г/м2.

Проверка нормальности статистическими тестами является очень строгой, это связано со сравнением с идеальным распределением. Поэтому, несмотря на то что статистический тест говорит о ненормальности распределения, всегда стоит смотреть на гистограмму распределения. Гистограммы распределения «после» показывают, что нам удалось сгладить и масштабировать данные.

Посмотрим корреляционную матрицу Пирсона.

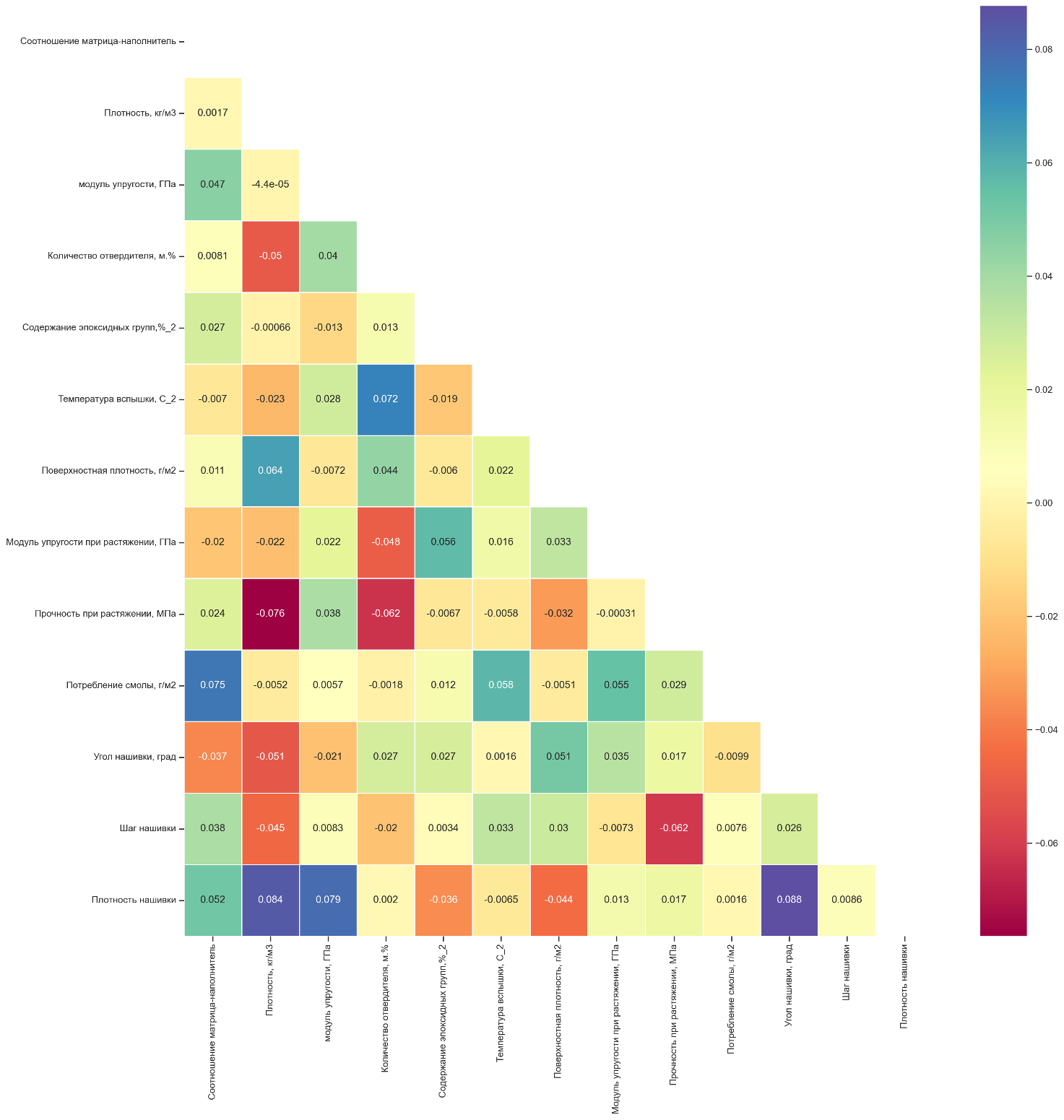


Рисунок 46 – Тепловая карта

Корреляционная матрица имеет разброс от – 0,076 до 0,88, что говорит о полном отсутствие линейно связанных параметров. Причем это касается как связи между входными переменными и прогнозируемыми, так и входными между собой. Последнее позволяет включить все входные параметры в модели регрессии, однако можно предположить, что построение достоверного прогноза будет затруднительным.

## **3.б Практическая часть. Разработка и обучение модели**

Целевыми показателями для нас являются: модуль упругости при растяжении, прочность при растяжении и соотношение матрица – наполнитель.

Чтобы спрогнозировать значение прочности при растяжении и модуля упругости при растяжении, будем использовать все параметры кроме вышеперечисленных, так как они являются результатом измерения конечного, выходного композита. На 30 процентах данных идет тестирование моделей, на остальных – обучение.

Построим базовое решение выбранных моделей без подбора гиперпараметров и произведем оценку метрик моделей.

Таблица 5 – Метрики базового решения для предсказания «Модуля упругости при растяжении, ГПа»

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | MSE | R2 |
| Линейная регрессия | 0,839 | 1,037 | -0,02 |
| Lasso регрессия | 0,83 | 1,017 | -0,001 |
| Ridge регрессия | 0,839 | 1,037 | -0,02 |
| K ближайших соседей | 0,922 | 1,27 | -0,249 |
| Метод опорных векторов | 0,897 | 1,201 | -0,182 |
| Дерево решений | 1,171 | 2,107 | -1,072 |
| Случайный лес | 0,855 | 1,119 | -0,101 |
| Градиентный спуск | 0,895 | 1,197 | -0,177 |

Таблица 6 – Метрики базового решения для предсказания «Прочности при растяжении, МПа»

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | MSE | R2 |
| Линейная регрессия | 0,796 | 0,979 | -0,006 |
| Lasso регрессия | 0,794 | 0,983 | -0,01 |
| Ridge регрессия | 0,796 | 0,979 | -0,006 |
| K ближайших соседей | 0,849 | 1,153 | -0,186 |
| Метод опорных векторов | 0,84 | 1,112 | -0,143 |
| Дерево решений | 1,041 | 1,713 | -0,76 |
| Случайный лес | 0,816 | 1,01 | -0,038 |
| Градиентный спуск | 0,831 | 1,037 | -0,066 |

Таблица 7 – Метрики базового решения для предсказания «Соотношение матрица - наполнитель»

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | MSE | R2 |
| Линейная регрессия | 0,796 | 0,979 | -0,006 |
| Lasso регрессия | 0,794 | 0,983 | -0,01 |
| Ridge регрессия | 0,796 | 0,979 | -0,006 |
| K ближайших соседей | 0,849 | 1,153 | -0,186 |
| Метод опорных векторов | 0,84 | 1,112 | -0,143 |
| Дерево решений | 1,057 | 1,759 | -0,808 |
| Случайный лес | 0,808 | 1,003 | -0,031 |
| Градиентный спуск | 0,832 | 1,038 | -0,066 |

Коэффициент детерминации во всех моделях отрицательный. Это свидетельствует о том, что построенные модели очень плохо описывают данные и зависимости в них. По входящим параметрам невозможно построить достоверный прогноз расчета модуля упругости. Модели линейной регрессии дают около нулевое значение коэффициента детерминации, на втором месте стоит модель случайного леса. Возможно, при подборе гиперпараметров, модель сможет дать лучший результат.

## **3.в Подбор гиперпараметров моделей**

Для подбора гиперпараметров используется инструмент GridSearch, помогающий найти самую эффективную настройку параметров модели.

Попробуем произвести подбор гиперпараметров для некоторых моделей и сравним метрики с базовым решением.

Таблица 8 – Сравнение метрик для параметра "Модуль упругости при растяжении, ГПа"

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | MSE | R2 |
| K ближайших соседей | 0,849 | 1,153 | -0,186 |
| K ближайших соседей GridSearch | 0,857 | 1,099 | -0,081 |
| Случайный лес | 0,808 | 1,003 | -0,031 |
| Случайный лес GridSearch | 0,843 | 1,049 | -0,032 |

Таблица 9 – Сравнение метрик для параметра «Прочность при растяжении, МПа»

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | MSE | R2 |
| K ближайших соседей | 0,849 | 1,153 | -0,186 |
| K ближайших соседей GridSearch | 0,823 | 1,051 | -0,08 |

Таблица 10 – Сравнение метрик для параметра «Соотношение матрица - наполнитель»

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | MSE | R2 |
| K ближайших соседей | 0,849 | 1,153 | -0,186 |
| K ближайших соседей GridSearch | 0,847 | 1,073 | -0,05 |
| Случайный лес | 0,808 | 1,003 | -0,031 |
| Случайный лес GridSearch | 0,837 | 1,037 | -0,015 |

Во всех моделях коэффициент детерминации увеличивается незначительно, что говорит о невозможности построения прогноза на исходных данных.

## **3.г Рекомендательная нейронная сеть для параметра «Соотношение матрица - наполнитель»**

Для создания рекомендательной нейронной сети необходимо разделить выборку на проверочную и обучающую, исключая из датасета целевой параметр Соотношение матрица – наполнитель, используя псевдослучайный генератор, равный 42, ведь галактика суровая штука, чтобы в ней выжить, надо знать, где твое полотенце.

Проверим как разделился датасет.

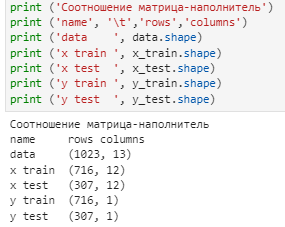


Рисунок 47 – Разделение выборки на проверочную и обучающую

Проведем стандартизацию и нормализацию выборки, а затем создадим и обучим модель. Оптимальным количеством входных слоев оказались два слоя по 8 нейронов с функцией активации relu и выходной слой с функцией активации сигмоида.

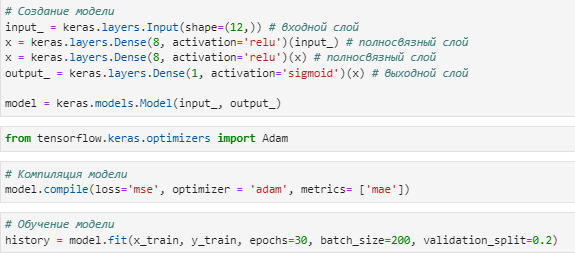


Рисунок 48 – Фрагмент кода создания модели

Приведем статистику обучения модели:

Epoch 1/30

3/3 [=========================] - 0s 47ms/step - loss: 6.6730 - mae: 2.4149 - val\_loss: 6.2378 - val\_mae: 2.3309

Epoch 2/30

3/3 [=========================] - 0s 12ms/step - loss: 6.6339 - mae: 2.4070 - val\_loss: 6.2012 - val\_mae: 2.3230

Epoch 3/30

3/3 [=========================] - 0s 12ms/step - loss: 6.5946 - mae: 2.3991 - val\_loss: 6.1645 - val\_mae: 2.3152

Epoch 4/30

3/3 [========================] - 0s 12ms/step - loss: 6.5562 - mae: 2.3912 - val\_loss: 6.1281 - val\_mae: 2.3073

Epoch 5/30

3/3 [=========================] - 0s 13ms/step - loss: 6.5174 - mae: 2.3833 - val\_loss: 6.0915 - val\_mae: 2.2994

Epoch 6/30

3/3 [=========================] - 0s 12ms/step - loss: 6.4785 - mae: 2.3754 - val\_loss: 6.0547 - val\_mae: 2.2914

Epoch 7/30

3/3 [=========================] - 0s 11ms/step - loss: 6.4406 - mae: 2.3677 - val\_loss: 6.0176 - val\_mae: 2.2833

Epoch 8/30

3/3 [=========================] - 0s 12ms/step - loss: 6.4022 - mae: 2.3597 - val\_loss: 5.9807 - val\_mae: 2.2752

Epoch 9/30

3/3 [=========================] - 0s 13ms/step - loss: 6.3640 - mae: 2.3518 - val\_loss: 5.9437 - val\_mae: 2.2671

Epoch 10/30

3/3 [=========================] - 0s 12ms/step - loss: 6.3265 - mae: 2.3440 - val\_loss: 5.9066 - val\_mae: 2.2590

Epoch 11/30

3/3 [=========================] - 0s 13ms/step - loss: 6.2886 - mae: 2.3361 - val\_loss: 5.8692 - val\_mae: 2.2507

Epoch 12/30

3/3 [=========================] - 0s 12ms/step - loss: 6.2508 - mae: 2.3283 - val\_loss: 5.8317 - val\_mae: 2.2424

Epoch 13/30

3/3 [=========================] - 0s 13ms/step - loss: 6.2132 - mae: 2.3204 - val\_loss: 5.7942 - val\_mae: 2.2341

Epoch 14/30

3/3 [=========================] - 0s 12ms/step - loss: 6.1757 - mae: 2.3126 - val\_loss: 5.7567 - val\_mae: 2.2257

Epoch 15/30

3/3 [=========================] - 0s 12ms/step - loss: 6.1391 - mae: 2.3048 - val\_loss: 5.7192 - val\_mae: 2.2173

Epoch 16/30

3/3 [=========================] - 0s 12ms/step - loss: 6.1017 - mae: 2.2969 - val\_loss: 5.6823 - val\_mae: 2.2089

Epoch 17/30

3/3 [=========================] - 0s 13ms/step - loss: 6.0652 - mae: 2.2891 - val\_loss: 5.6456 - val\_mae: 2.2005

Epoch 18/30

3/3 [=========================] - 0s 12ms/step - loss: 6.0280 - mae: 2.2811 - val\_loss: 5.6089 - val\_mae: 2.1922

Epoch 19/30

3/3 [=========================] - 0s 13ms/step - loss: 5.9903 - mae: 2.2730 - val\_loss: 5.5720 - val\_mae: 2.1837

Epoch 20/30

3/3 [=========================] - 0s 12ms/step - loss: 5.9532 - mae: 2.2649 - val\_loss: 5.5351 - val\_mae: 2.1752

Epoch 21/30

3/3 [=========================] - 0s 13ms/step - loss: 5.9154 - mae: 2.2568 - val\_loss: 5.4982 - val\_mae: 2.1666

Epoch 22/30

3/3 [=========================] - 0s 12ms/step - loss: 5.8774 - mae: 2.2485 - val\_loss: 5.4614 - val\_mae: 2.1581

Epoch 23/30

3/3 [=========================] - 0s 12ms/step - loss: 5.8394 - mae: 2.2402 - val\_loss: 5.4247 - val\_mae: 2.1495

Epoch 24/30

3/3 [=========================] - 0s 13ms/step - loss: 5.8008 - mae: 2.2316 - val\_loss: 5.3880 - val\_mae: 2.1409

Epoch 25/30

3/3 [=========================] - 0s 13ms/step - loss: 5.7617 - mae: 2.2231 - val\_loss: 5.3516 - val\_mae: 2.1324

Epoch 26/30

3/3 [=========================] - 0s 13ms/step - loss: 5.7238 - mae: 2.2145 - val\_loss: 5.3153 - val\_mae: 2.1238

Epoch 27/30

3/3 [=========================] - 0s 12ms/step - loss: 5.6846 - mae: 2.2060 - val\_loss: 5.2794 - val\_mae: 2.1153

Epoch 28/30

3/3 [=========================] - 0s 13ms/step - loss: 5.6462 - mae: 2.1974 - val\_loss: 5.2437 - val\_mae: 2.1069

Epoch 29/30

3/3 [=========================] - 0s 12ms/step - loss: 5.6079 - mae: 2.1889 - val\_loss: 5.2081 - val\_mae: 2.0984

Epoch 30/30

3/3 [=========================] - 0s 13ms/step - loss: 5.5704 - mae: 2.1803 - val\_loss: 5.1727 - val\_mae: 2.0900

Просмотрим архитектуру модели:

Model: "model\_1"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

input\_8 (InputLayer) [(None, 12)] 0

dense\_12 (Dense) (None, 8) 104

dense\_13 (Dense) (None, 8) 72

dense\_14 (Dense) (None, 1) 9

=================================================================

Total params: 185

Trainable params: 185

Non-trainable params: 0

Просмотрим данные метрики модели:

Таблица 11 – Метрики полученной модели

|  |  |
| --- | --- |
| Метрика | Значение |
| Mean Squared Error | 5.703124617794415 |
| Mean Absolute Error | 2.2030300378041465 |
| R-squared Score | -5.653689581978382 |

Построим функции loss и mae:

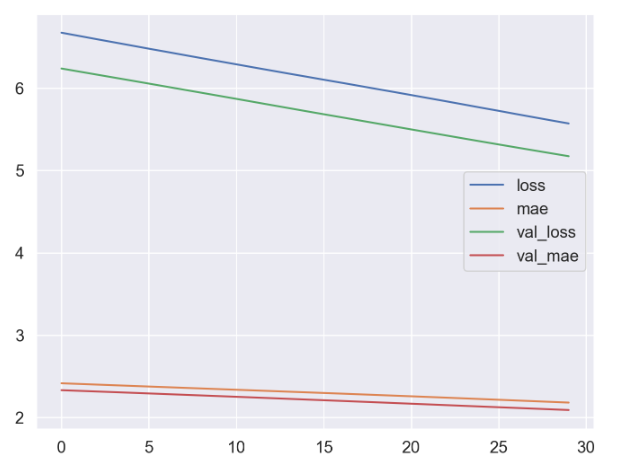


Рисунок 49 – Функции loss и mae

Построим графики тестовых и прогнозных значений.

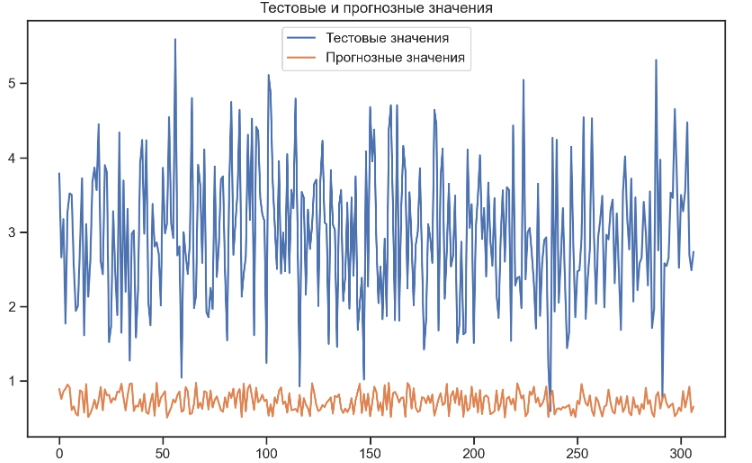


Рисунок 50 – Графики тестовых и прогнозных значений

Как видно из всех вышеперечисленных отчетов модель не справилась с предсказанием. Дальнейшая настройка и отладка иррациональны.

## **3. д Разработка приложения**

Несмотря на неудовлетворительные результаты построения модели нейронной сети, в целях обучения, попробуем создать приложение, на вход в которое подаются данные, вводимые пользователем, приложение обращается к написанной нейронной сети, и на выходе пользователь получает прогнозируемое значение соотношения матрица – наполнитель.

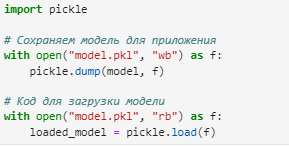


Рисунок 51 – Фрагмент кода python для сохранения модели для приложения

Приведем код приложения, написанный в среде VSCode:

import pickle

import flask

from flask import render\_template

import sklearn

import tensorflow as tf

import keras as keras

app = flask.Flask(\_\_name\_\_, template\_folder = 'templates')

@app.route('/', methods = ['POST', 'GET'])

#@app.route('/index', methods = ['POST', 'GET'])

def main():

    if flask.request.method == 'GET':

        return render\_template('main.html')

    if flask.request.method == 'POST':

        with open ('model.pkl', 'rb') as f:

            loaded\_model = pickle.load(f)

        angle = float(flask.request.form['angle'])

        step = float(flask.request.form['step'])

        density = float(flask.request.form['density'])

        elasticity\_module = float(flask.request.form['elasticity\_module'])

        hardener\_quantity = float(flask.request.form['hardener\_quantity'])

        epoxy\_group = float(flask.request.form['epoxy\_group'])

        temperature = float(flask.request.form['temperature'])

        surface\_density = float(flask.request.form['surface\_density'])

        elasticity\_module2 = float(flask.request.form['elasticity\_module2'])

        tensile\_strength = float(flask.request.form['tensile\_strength'])

        resin\_consumption = float(flask.request.form['resin\_consumption'])

        patch\_density = float(flask.request.form['patch\_density'])

        y\_pred = loaded\_model.predict([[angle, step, patch\_density, density, elasticity\_module, hardener\_quantity, epoxy\_group,

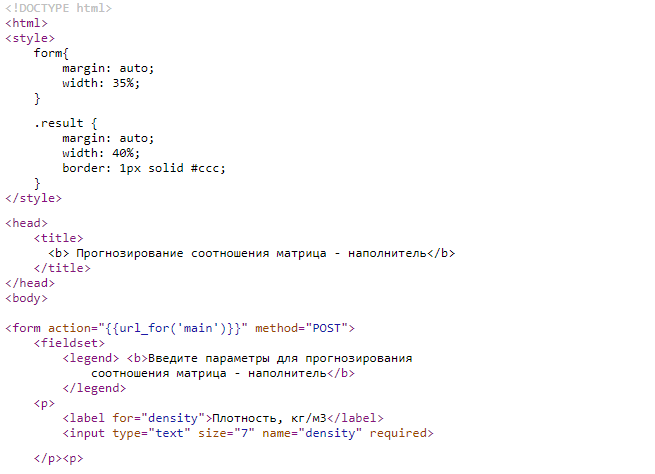
                                        temperature, surface\_density, elasticity\_module2, tensile\_strength, resin\_consumption]])

        return render\_template('main.html', result = y\_pred)

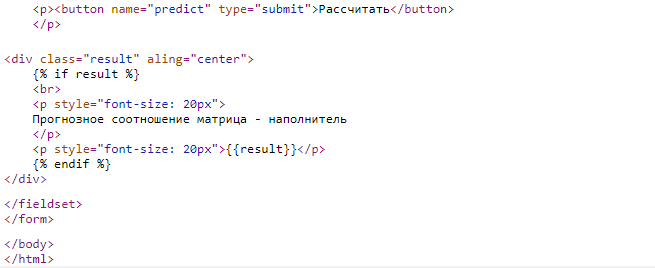
if \_\_name\_\_ == '\_\_main\_\_':

    app.run()

Приведем код веб-формы ввода, написанный в среде HTML:







Работа приложения записана на видео и доступна по ссылке <https://disk.yandex.ru/i/nZZgY1eSSG7V3Q>

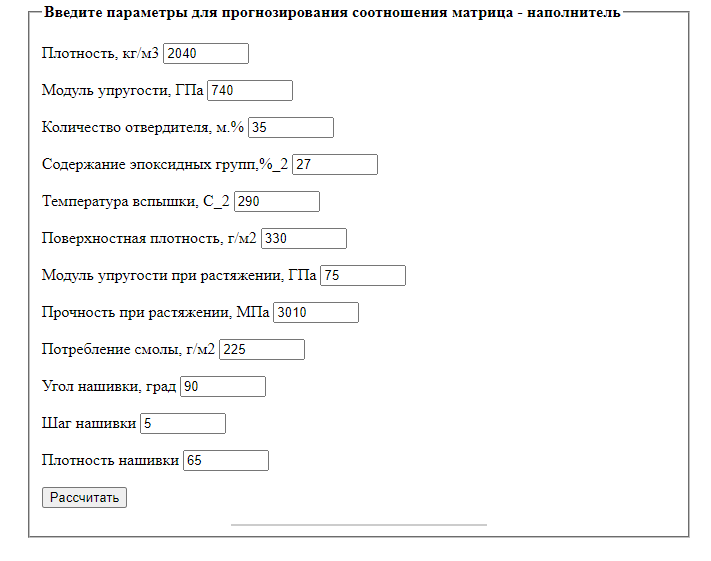


Рисунок 52 – Форма ввода данных пользователем

К сожалению, моих познаний не хватило для выявления ошибки в работе приложения.

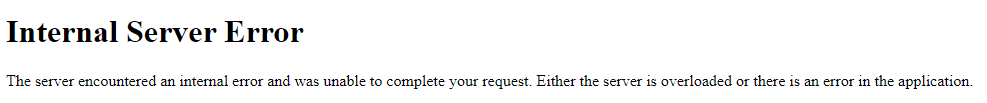


Рисунок 53 – Ошибка сервера или приложения

Сохраним установленное окружение в файл с помощью инструмента pip freeze > requirements\_main\_file.txt.

## **3.е Создание удаленного репозитория на Github и загрузка результатов работы на него**

Исходные файлы и все материалы, полученные в ходе выполнения работы, размещены по ссылке <https://github.com/Clockworkcabbage/Final-work> при помощи десктопной версии Github.

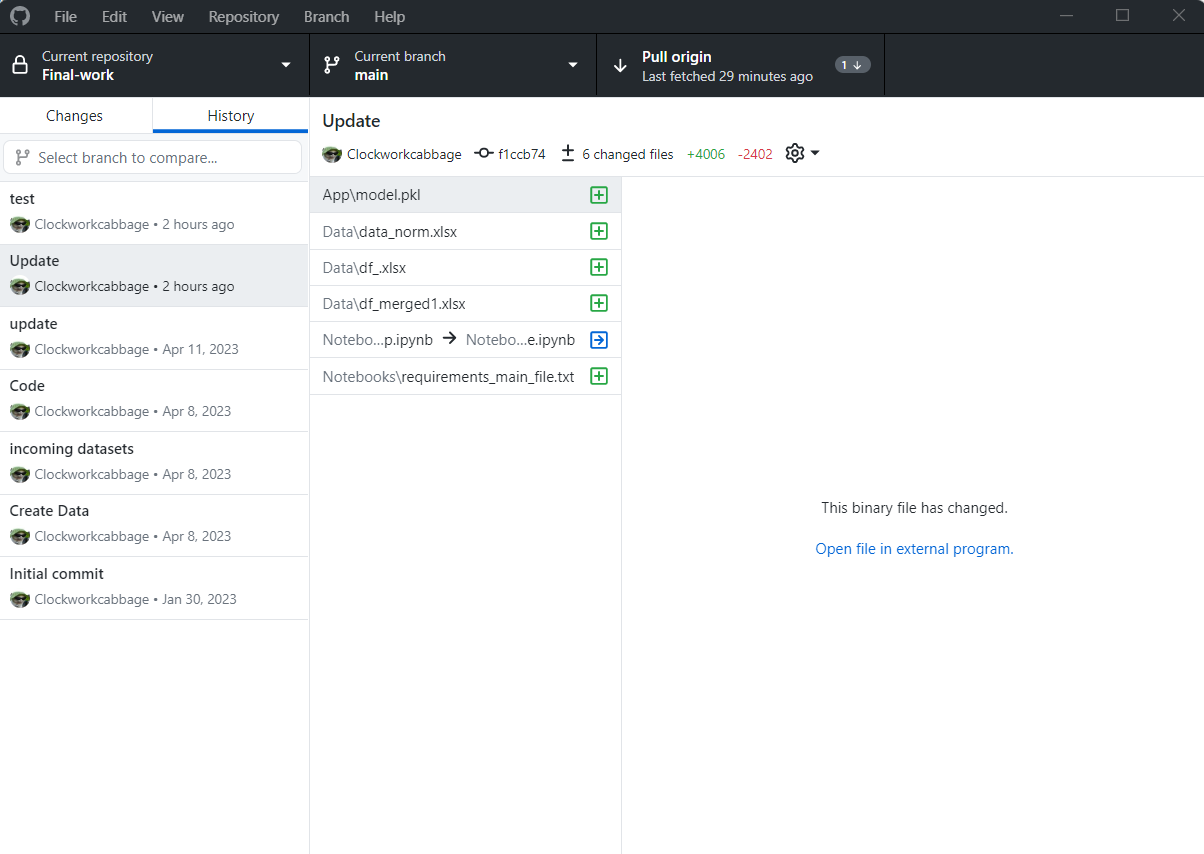


Рисунок 53 – Работа с приложением Github desktop

В файле-описании Readme указаны гиперссылки на размещенные материалы.

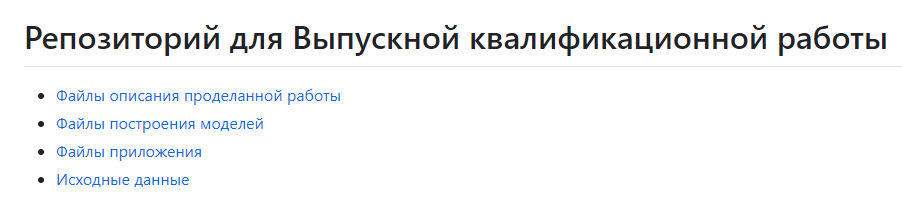


Рисунок 54 – Содержание файла readme.md

# **4 Заключение**

Поставленная цель не была мною достигнута, ни одна модель регрессии не справилась с задачей предсказания целевых признаков. На мой взгляд возможными причинами могут быть:

- недостаточный размер обучающих данных, часть из которых выглядит сгенерированными;

- нерепрезентативная выборка параметров, возможно, требуется более тщательный отбор признаков или конструирование признаков или консультация эксперта для понимания физической сути процессов;

- данные плохого качества, синтетические/ложные данные;

- недостаточная квалификация исследователя.

# **5 Библиографический список**

1. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn, Keras и TensorFlow: концепции, инструменты и техники для создания интеллек­ туальных систем , 2-е изд. : Пер. с англ. — СПб. : ООО “Диалектика”, 2020. — 1040 с .: ил. — Парал. тит. англ.
2. Ковтунов А.И., Мямин С.В., Семистенова Т.В. Слоистые композиционные материалы: электронное учебное пособие. Тольятти, Изд–во ТГУ, 2017, 75 c.
3. Рашка С., Мирджалили В. (2020). Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow 2. 3-е изд. СПб: ООО «Диалектика», 848 с.
4. Хайкин, С. Нейронные сети : полный курс / С. Хайкин. - Москва : Вильямс, 2006. 1104 с.
5. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 351 с
6. Документация по библиотеке Keras:

<https://keras.io/api/>

1. Документация по библиотеке Numpy:

<https://numpy.org/doc/1.22/user/index.html#user>

1. Документация по библиотеке Pandas:

<https://pandas.pydata.org/docs/user_guide/index.html#user-guide>

1. Документация по библиотеке Scikit-learn:

<https://scikit-learn.org/stable/user_guide.html>

1. Документация по библиотеке Tensorflow:

<https://www.tensorflow.org/guide?hl=ru>