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»

Тема: «Прогнозирование конечных свойств новых материалов (композиционных материалов)»

Слушатель Мезенцева Екатерина Михайловна

Москва, 2023

Содержание

Содержание 2
Введение 3
1 Аналитическая часть 4
1.1 Постановка задачи 4
1.2 Описание используемых методов 8
1.1.1 Линейная регрессия (Linear regression) 8
1.1.2 Гребневая регрессия (Ridge) 9
1.1.3 Лассо регрессия (Lasso) 9
1.1.4 Метод опорных векторов (SVR) 10
1.1.5 Деревья решений (DecisionTreeRegressor) 10
1.1.6 Случайный лес (RandomForest) 11
1.1.7 Градиентный бустинг (GradientBoostingRegressor) 12
1.2 Разведочный анализ данных 13
2 Практическая часть 17
2.1 Предобработка данных 17
2.2 Разработка и обучение модели 21
2.2.1 Оценка работы моделей и подбор гиперпараметров применительно к задаче прогнозирования модуля упругости при растяжении 22
2.2.2 Оценка работы моделей и подбор гиперпараметров применительно к задаче прогнозирования прочности при растяжении 24
2.3 Тестирование модели 25
2.4 Написание нейронной сети, рекомендующей соотношение матрица-наполнитель 26
2.4.1 MLPRegressor из библиотеки sklearn 26
2.4.2 Нейросеть из библиотеки tensorflow 27
2.5 Разработка приложения 31
2.6 Создание удаленного репозитория и загрузка результатов работы на него.33
Заключение 33
Библиографический список: 35

Введение

Тема данной работы – прогнозирование конечных свойств новых материалов (композиционных материалов).

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. Композитам свойственна монолитность – компоненты, из которых состоит композит, не могут быть отделены друг от друга без разрушения композиционного материала.

Ярким примером такого материала является железобетон: удачно комбинируя свойства бетона и стальной арматуры для получения новых, уникальных свойств, данный материал успел зарекомендовать себя как прекрасное решение в случаях, когда речь идёт о строительстве.

Другими примерами композиционных материалов можно назвать органопластики, древесно-композиционные материалы, стеклопластики, текстолиты, композиционные материалы с металлической матрицей и композиционные материалы на основе керамики.

Структура композиционных материалов представляет собой матрицу (основной компонент), содержащую в своем объеме или армирующие элементы, часто называемые наполнителем. Матрица и наполнитель разделены границей (поверхностью) раздела. Наполнитель равномерно распределен в матрице и имеет заданную пространственную ориентацию.

Композиционные материалы характеризуются совокупностью свойств, не присущих каждому в отдельности взятому компоненту. За счет выбора армирующих элементов, варьирования их объемной доли в матричном материале, а также размеров, формы, ориентации и прочности связи по границе «матрица-наполнитель», свойства композиционных материалов можно регулировать в значительных пределах.

У такого подхода 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).

1 Аналитическая часть

1.1 Постановка задачи

В данной работе исследуется композит с матрицей из базальтопластика и нашивками из углепластика. На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов.

Цель исследования: решение актуальной производственной задачи по прогнозированию свойств получаемых композиционных материалов.

Объект исследования: композиционные материалы, которые означают искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними.

Предмет исследования: прогнозные данные трех свойств композитов: соотношение матрица-наполнитель, модуль упругости при растяжении, прочность при растяжении.

В ходе исследовательской работы были поставлены следующие задачи:

1) Изучить теоретические основы и методы решения поставленной задачи;

2) Провести разведочный анализ предложенных данных. Необходимо нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек. Необходимо также для каждой колонке получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков;
3) Провести предобработку данных (удаление шумов, нормализация и т.д.);
4) Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении модели необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10;
5) Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель;
6) Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании 4 или 5 (один или два прогноза, на выбор учащегося);
7) Оценить точность модели на тренировочном и тестовом датасете;
8) Создать репозиторий в GitHub / GitLab и разместить там код исследования. Оформить файл README.

Для исследовательской работы в качестве датасета были даны 2 файла: X_bp.xlsx (с данными о параметрах базальтопластика, состоящий из 1023 строк и 10 столбцов данных) и X_nup.xlsx (данными нашивок углепластика, состоящий из 1040 строк и 3 столбцов данных).

```
# Загрузка исходных данных из файла X_bp
x_bp = pd.read_excel('X_bp.xlsx', index_col = 0)
x_bp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1023 entries, 0 to 1022
Data columns (total 10 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
dtypes: float64(10)
memory usage: 87.9 KB
```

Рисунок 1 – Загрузка исходных данных из файла X_bp.xlsx

```
# Загрузка исходных данных из файла X_nup
x_nup = pd.read_excel('X_nup.xlsx', index_col = 0)
x_nup.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1040 entries, 0 to 1039
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Угол нашивки, град  1040 non-null   int64  
 1   Шаг нашивки         1040 non-null   float64
 2   Плотность нашивки   1040 non-null   float64
dtypes: float64(2), int64(1)
memory usage: 32.5 KB
```

Рисунок 2 – Загрузка исходных данных из файла X_nup.xlsx

В качестве входных данных приняты данные о начальных свойствах компонентов композиционных материалов:

– Соотношение матрица-наполнитель;

– Плотность;

– Модуль упругости;

– Количество отвердителя;

– Содержание эпоксидных групп;

– Температура вспышки;

– Поверхностная плотность;
– Модуль упругости при растяжении;
– Прочность при растяжении;
– Потребление смолы;
– Угол нашивки;
– Шаг нашивки;
– Плотность нашивки.

Общее количество параметров для анализа – 13.

Датасеты объединены, тип объединения INNER. Пропуски отсутствуют. Элементы массива соответствуют типу float64. За исключением признака «Угол нашивки», все признаки в датасете являются непрерывными, количественными. Признак «Угол нашивки» принимает только два значения и потому будет рассматриваться как категориальный признак.

Для каждой колонки получены среднее, стандартное отклонение, минимальное, максимальное, первый квартиль, медианное значение, третий квартиль.

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.930	0.913	0.389	2.318	2.907	3.553	5.592
Плотность, кг/м3	1023.0	1975.735	73.729	1731.765	1924.155	1977.622	2021.374	2207.773
модуль упругости, ГПа	1023.0	739.923	330.232	2.437	500.047	739.664	961.813	1911.536
Количество отвердителя, м.%	1023.0	110.571	28.296	17.740	92.443	110.565	129.730	198.953
Содержание эпоксидных групп,%_2	1023.0	22.244	2.406	14.255	20.608	22.231	23.962	33.000
Температура вспышки, С_2	1023.0	285.882	40.943	100.000	259.067	285.897	313.002	413.273
Поверхностная плотность, г/м2	1023.0	482.732	281.315	0.604	266.817	451.864	693.225	1399.542
Модуль упругости при растяжении, ГПа	1023.0	73.329	3.119	64.054	71.245	73.269	75.357	82.682
Прочность при растяжении, МПа	1023.0	2466.923	485.628	1036.857	2135.850	2459.525	2767.193	3848.437
Потребление смолы, г/м2	1023.0	218.423	59.736	33.803	179.628	219.199	257.482	414.591
Угол нашивки, град	1023.0	0.492	0.500	0.000	0.000	0.000	1.000	1.000
Шаг нашивки	1023.0	6.899	2.563	0.000	5.080	6.916	8.586	14.441
Плотность нашивки	1023.0	57.154	12.351	0.000	49.799	57.342	64.945	103.989

Рисунок 3 – Среднее, стандартное отклонение, минимальное, максимальное, первый квартиль, медианное значение, третий квартиль для каждой колонки

1.2 Описание используемых методов

Данная задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем и традиционно является задачей регрессии. Говоря о методах, что были применены в ходе данной исследовательской работы, можно выделить следующие:

- Линейная регрессия (Linear regression);
- Гребневая регрессия (Ridge);
- Лассо регрессия (Lasso);
- Метод опорных векторов (SVR);
- Деревья решений (DecisionTreeRegressor);
- Случайный лес (RandomForest);
- Градиентный бустинг (GradientBoostingRegressor).

1.1.1 Линейная регрессия (Linear regression)

Линейная регрессия (Linear regression) – алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Являясь одним из наиболее простых и эффективных инструментов статистического моделирования, этот алгоритм определяет зависимость переменных с помощью линии наилучшего соответствия. Модель регрессии создаёт несколько метрик: R2, или коэффициент детерминации, позволяет измерить, насколько модель может объяснить дисперсию данных. Если R-квадрат равен 1, это значит, что модель описывает все данные. Если же R-квадрат равен 0,5, модель объясняет лишь 50 процентов дисперсии данных. Оставшиеся отклонения не имеют объяснения. Чем ближе R2 к единице, тем лучше.

Достоинства метода:

- Простота в реализации;

– Интерпретируемость, меньшая сложность в сравнении с прочими алгоритмами.

Недостатки метода:

– Моделирует только прямые линейные зависимости;
– Необходима прямая связь между зависимыми и независимыми переменными;
– Значительное влияние выбросов, линейность границ.

1.1.2 Гребневая регрессия (Ridge)

Гребневая регрессия (Ridge) – это регрессия, которая добавляет дополнительный штраф к функции стоимости, но вместо этого суммирует квадраты значений коэффициентов (норма L-2) и умножает их на некоторую постоянную лямбду. Ридж-регрессию лучше применять, когда предсказательная способность набора данных распределена между различными характеристиками. Ридж-регрессия не обнуляет характеристики, которые могут быть полезны при составлении прогнозов, а просто уменьшает вес большинства переменных в модели.

1.1.3 Лассо регрессия (Lasso)

Лассо регрессия (Lasso) – это линейная модель, которая оценивает разреженные коэффициенты, позволяя таким образом уменьшить сложность модели и предотвратить переопределение, что может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое регуляризации в оптимизацию модели.

Достоинства метода:

– Легкость в полном избавлении от шумов в данных;
– Скорость работы;
– Низкая энергоёмкость;
– Способность полностью убрать признак из датасета;

– Доступное обнуление значений коэффициентов.

Недостатки метода:

– Часто страдает качество прогнозирования;
– Возможно ложное срабатывание результата;
– Случайным образом выбирает одну из коллинеарных переменных;
– Отсутствие оценки правильности формы взаимосвязи между независимой и зависимой переменными;
– Данный метод не всегда лучше, чем пошаговая регрессия.

1.1.4 Метод опорных векторов (SVR)

Метод опорных векторов (support vector machine, SVM) создает гиперплоскость или набор гиперплоскостей в многомерном пространстве, которые могут быть использованы для решения задач классификации и регрессии. Чаще всего данный метод применяется в постановке бинарной классификации. Вариация метода для регрессии называется SVR (Support Vector Regression).

Достоинства метода:

– Хорошая изученность метода;
– Для классификации достаточно небольшого набора данных.

Недостатки метода:

– Чувствительность к выбросам;
– Отсутствие интерпретируемости.

1.1.5 Деревья решений (DecisionTreeRegressor)

Дерево решений (DecisionTreeRegressor) – метод автоматического анализа больших массивов данных, в котором используется древовидная структура, подобная блок-схеме, или модель решений и всех их возможных результатов, включая результаты, затраты и полезность. Дерево состоит из элементов двух типов: узлов (node) и листьев (leaf).

Достоинства метода:

- Простота в применении и интерпретации;
- Работа с разными переменными;
- Помощь в визуализации процесса принятия решения и совершения правильного выбора в ситуациях, при которых результаты одного решения влияют на результаты следующих решений;
- Выделение наиболее важных полей для прогнозирования;
- Заполнение пропусков в данных наиболее вероятным решением.

Недостатки метода:

- Нестабильность процесса (изменение в одном узле может привести к построению совсем другого дерева);
- Затратность вычислений;
- Ограниченное число вариантов решения проблемы;
- Необходимость обращать внимание на размер;
- Ошибки при классификации с большим количеством классов и небольшой обучающей выборкой.

1.1.6 Случайный лес (RandomForest)

Случайный лес (RandomForest) — это множество решающих деревьев. Если точность дерева решений оказывается недостаточной, есть возможность собрать множество моделей в коллектив.

Достоинства метода:

- Высокая точность предсказания и внутренней оценки обобщающей способности модели, высокая масштабируемость и параллелизуемость.
- Эффективная обработка пропущенных данных, данных с большим числом классов и признаков;
- Не переобучается;
- Не требует предобработки входных данных.

Недостатки метода:

- Трудоёмкая прогнозируемость;
- Сложность интерпретации;
- Отсутствие возможности экстраполяции;
- Иногда работает хуже, чем линейные методы;
- Построение занимает много времени;
- Может недообучаться.

1.1.7 Градиентный бустинг (GradientBoostingRegressor)

Градиентный бустинг (AdaBoost) – это алгоритм, работающий по принципу перевзвешивания результатов. В основе данного алгоритма лежит итеративное обучение деревьев решений с целью минимизировать функцию потерь. Основная идея градиентного бустинга заключается в строительстве последовательно нескольких базовых классификаторов, каждый из которых как можно лучше компенсирует недостатки предыдущих. Финальный классификатор является линейной композицией этих базовых классификаторов.

Достоинства метода:

- Наблюдения выбираются на основе ошибки;
- Новые алгоритмы учатся на ошибках предыдущих;
- Требуется меньше итераций, чтобы приблизиться к фактическим прогнозам;
- Простота в настройке темпа обучения и применения;
- Легко интерпретируем.

Недостатки метода:

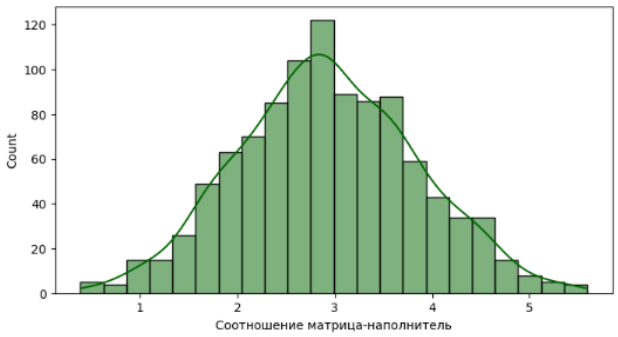
- Необходимость тщательно выбирать критерии остановки, иначе это может привести к переобучению;
- Наблюдения с наибольшей ошибкой появляются чаще;
- Слабее и менее гибок чем нейронные сети.

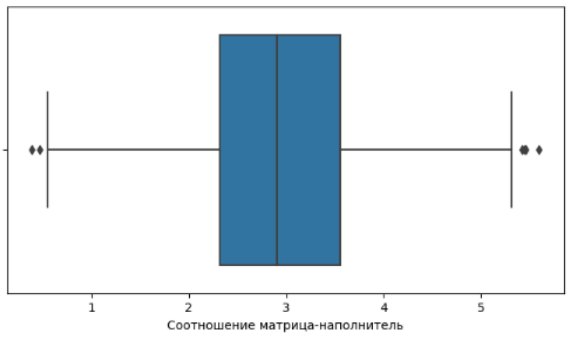
1.2 Разведочный анализ данных

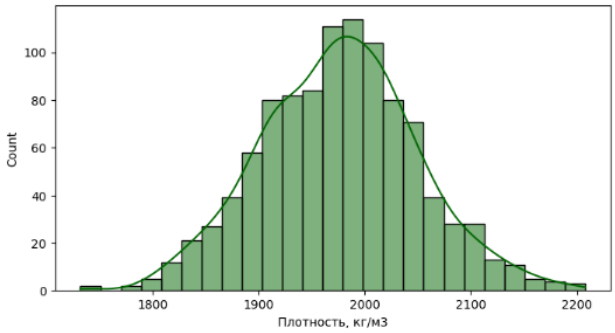
Для того, чтобы наиболее эффективно работать с имеющимися данными, необходимо произвести их разведочный анализ, что позволит получить первоначальные представления о характерах распределений переменных исходного набора данных, сформировать оценку их качества и выявить характер взаимосвязи между переменными.

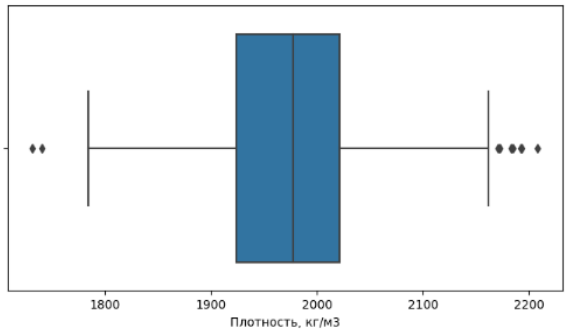
В качестве инструментов разведочного анализа используются гистограммы распределения, ящики с усами, попарные графики рассеяния точек, тепловые карты.

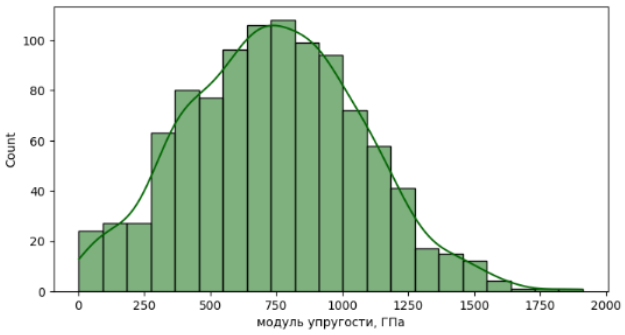
Построим гистограммы распределения и ящики с усами до нормализации данных, исключая угол нашивки как категориальный признак.

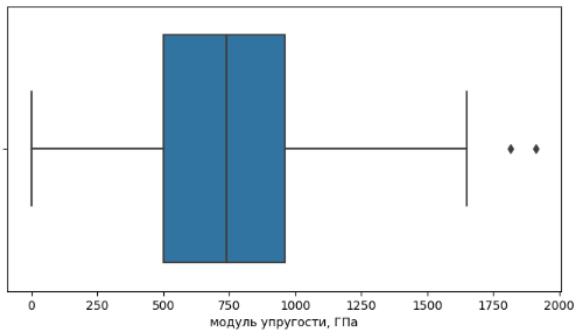


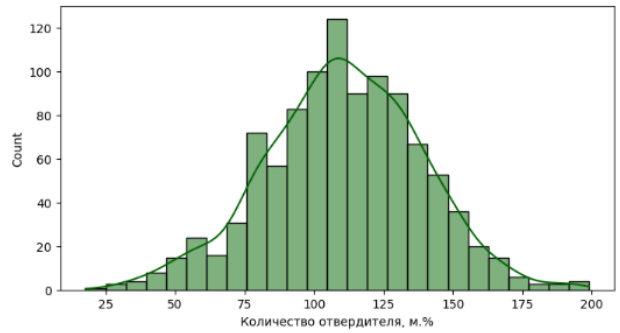


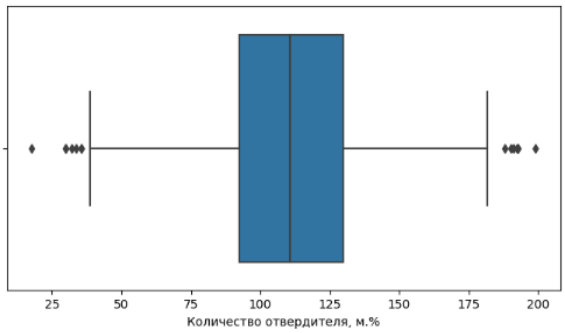


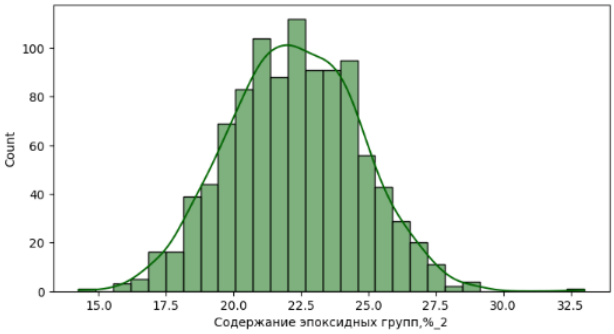


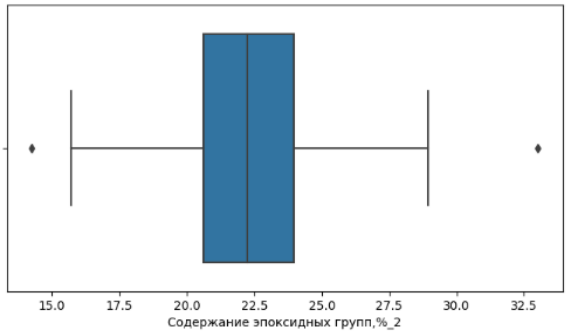


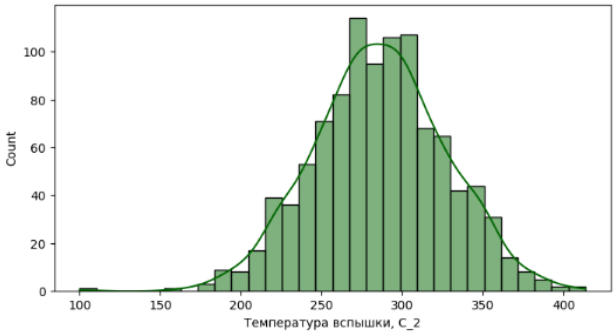


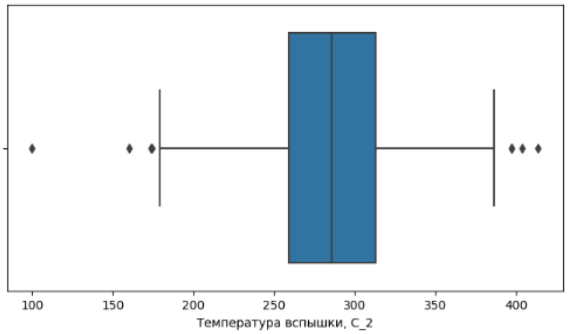

Count
Количество отвердителя, м.%

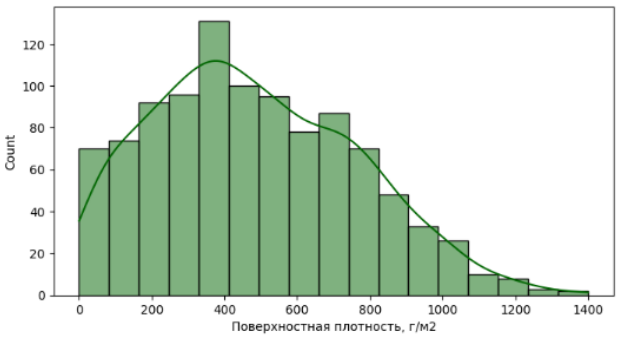

Количество отвердителя, м.%

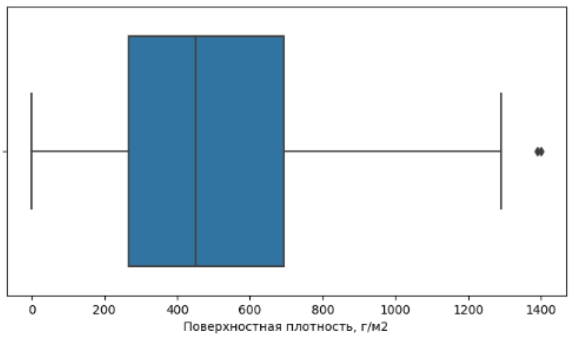

Count
Содержание эпоксидных групп,%_2

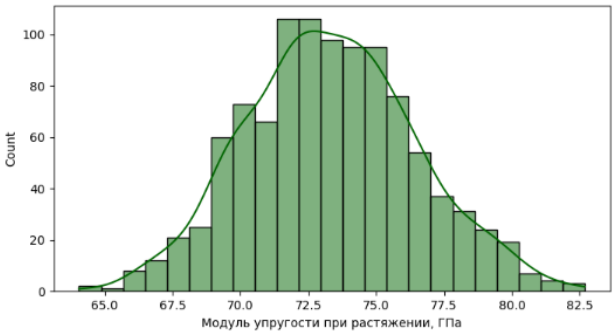

Содержание эпоксидных групп,%_2

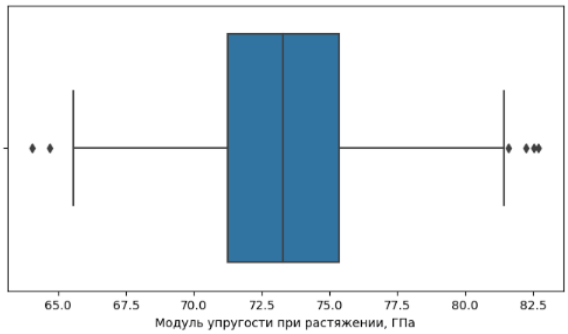

Count
Температура вспышки, С_2


Температура вспышки, С_2


Count
Поверхностная плотность, г/м2


Поверхностная плотность, г/м2


Count
Модуль упругости при растяжении, ГПа


Модуль упругости при растяжении, ГПа

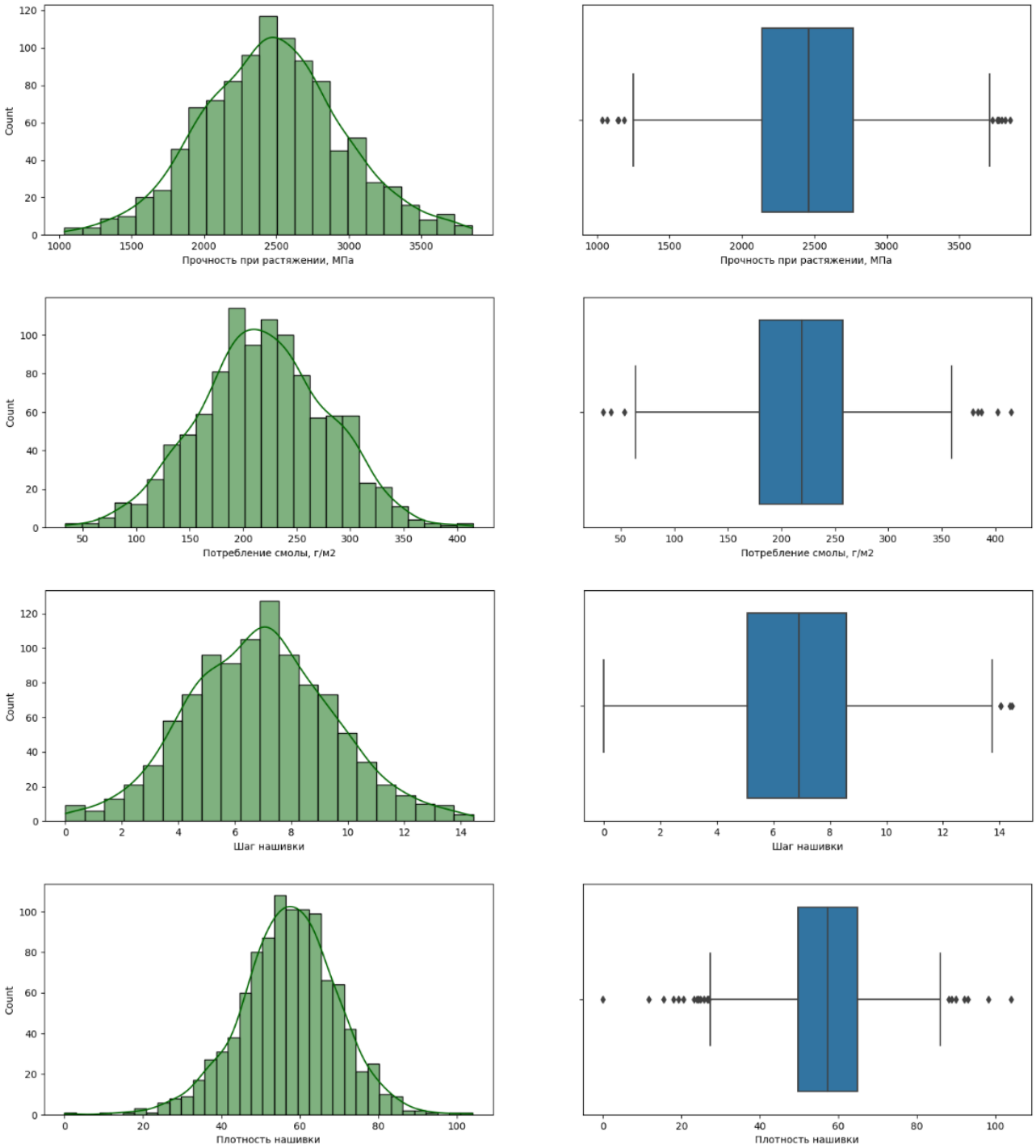


Рисунок 4 – Гистограммы распределения и ящики с усами до нормализации данных, исключая угол нашивки как категориальный признак

Выбросы наблюдаются по всем параметрам, кроме угла нашивки, т.к. данный параметр принимает дискретные значения и диаграмма «ящик с усами» для него не показательна.

Помимо этого, можно отметить наличие выбросов с двух сторон (например, потребление смолы), а также со стороны наибольших значений (например, поверхностная плотность).

Построим попарные графики рассеяния точек.

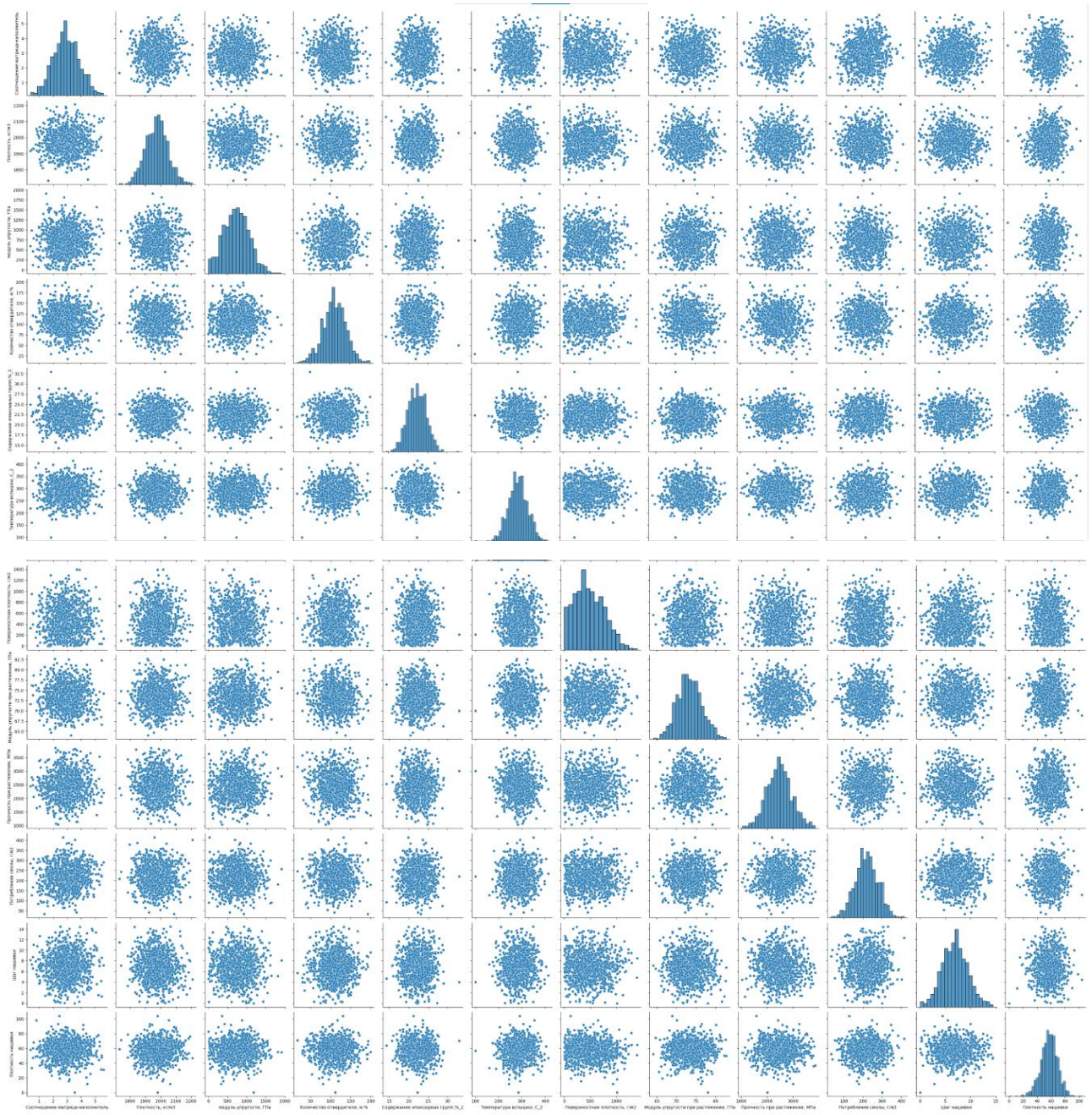


Рисунок 5 – Попарные графики рассеяния точек

Попарные графики рассеяния точек в нашем случае малоинформативны. Зависимости не линейные. Найдём коэффициенты корреляции.

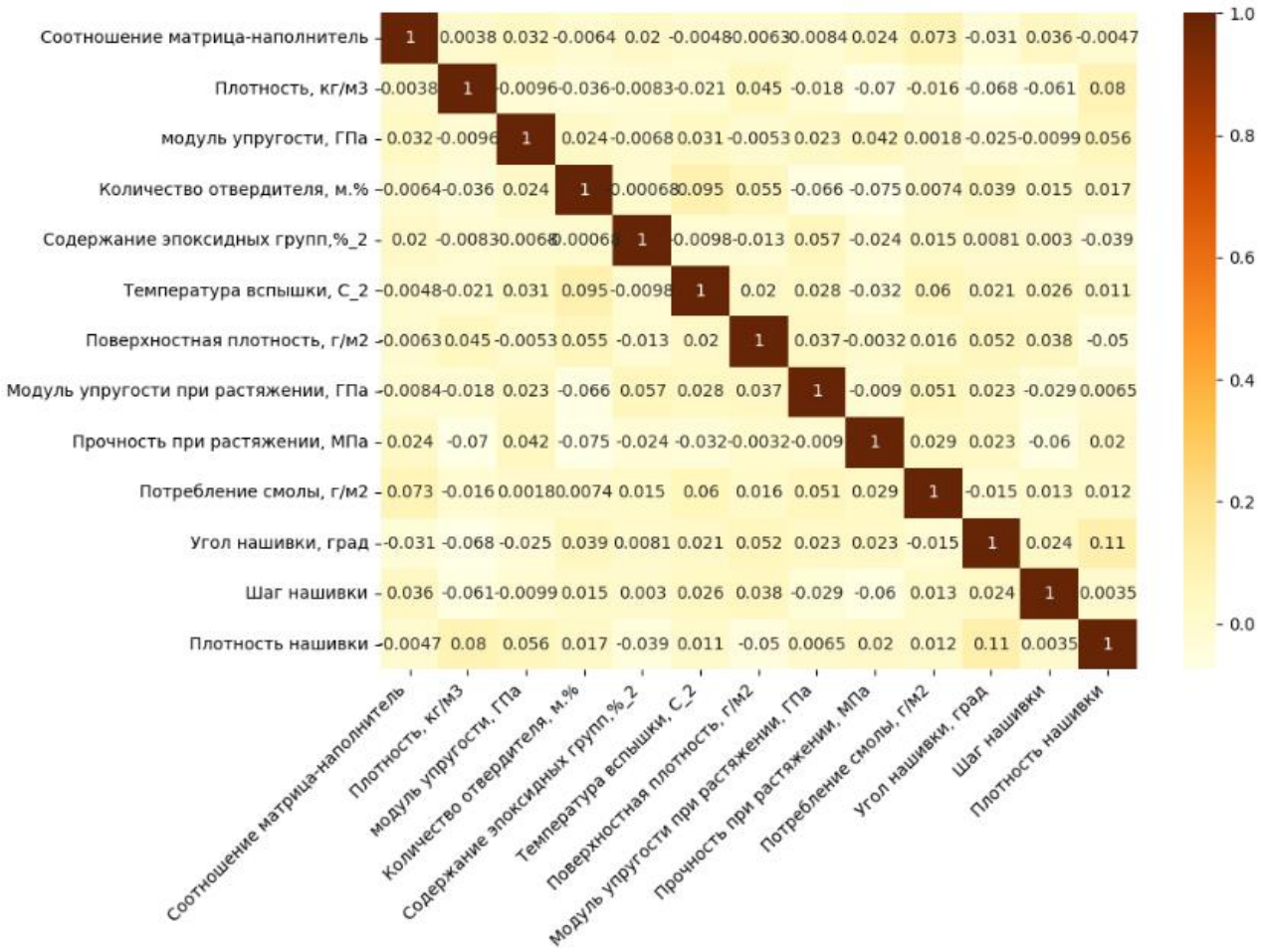


Рисунок 6 – Тепловая карта коэффициентов корреляции

На тепловой карте видно, что корреляция между данными близка к нулю: можно сделать вывод, что исходный датасет был предварительно обработан (либо сгенерирован) и переменные являются независимыми.

2 Практическая часть

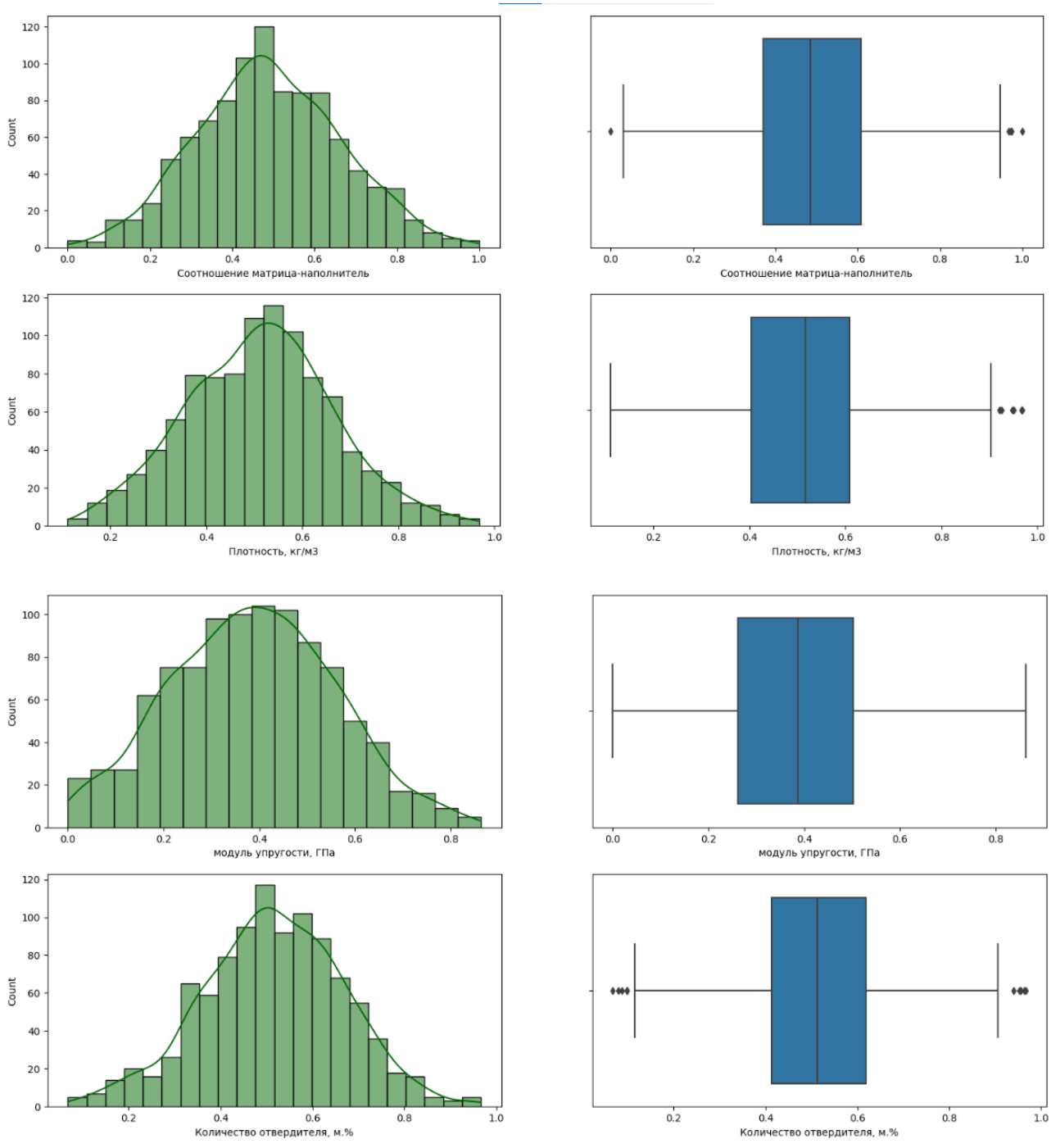
2.1 Предобработка данных

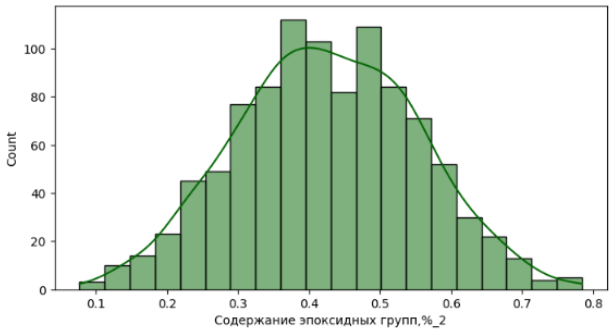
Перед построением моделей для прогноза необходимо произвести предобработку данных: нормализацию (приведение в диапазон от 0 до 1 с помощью MinMaxScaler) и стандартизацию (приведение к матожиданию 0, стандартному отклонению 1 с помощью StandartScaler).

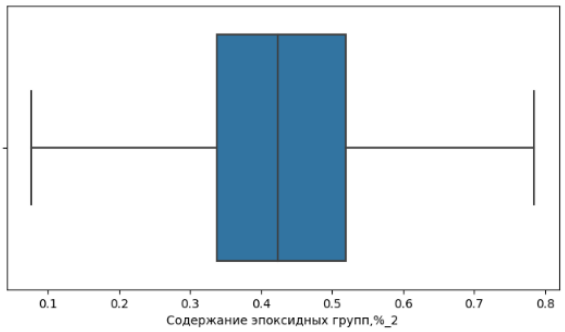
Сделаем масштабирование параметров таким образом, чтобы они находились между 0 и 1, и максимальное абсолютное значение каждой функции

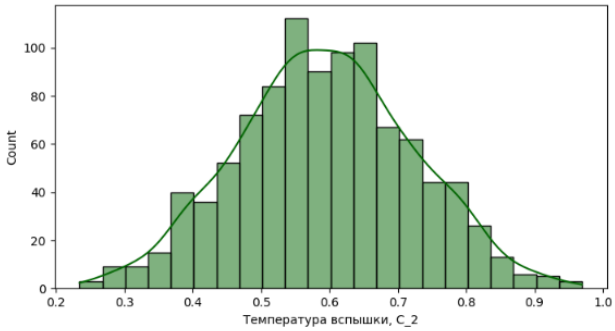
масштабировалось до размера единицы. Это позволит повысить устойчивость к небольшим стандартным отклонениям.

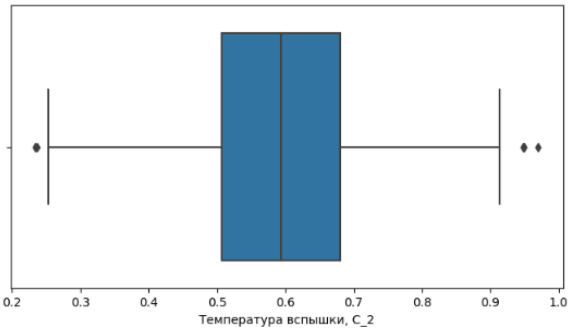
Для удаления выбросов будет использоваться межквартильный диапазон набора данных, который представляет собой разницу между первым квартилем (25-й процентиль) и третьим квартилем (75-й процентиль) от медианы. Это позволит исключить все данные, которые более чем в f раз превышают межквартильный диапазон от медианы данных.

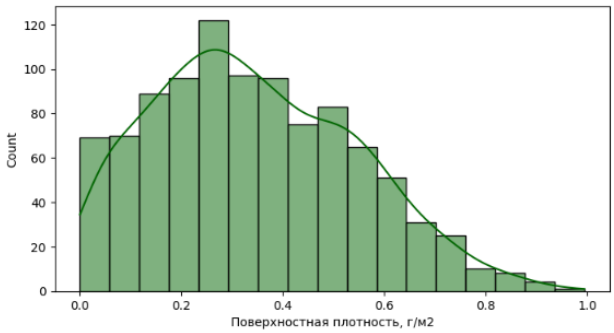


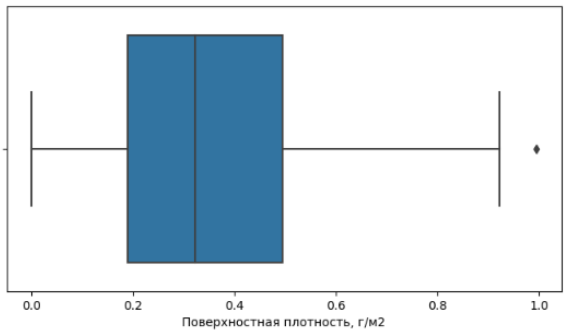

Count
Содержание эпоксидных групп,%_2

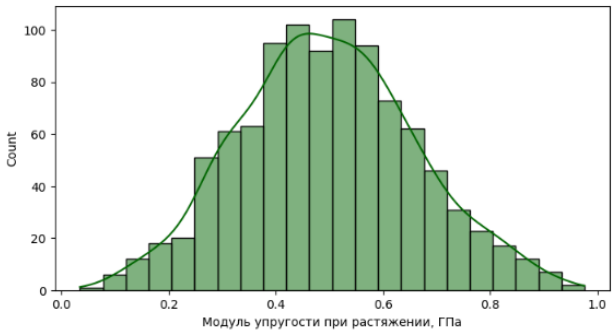

Содержание эпоксидных групп,%_2

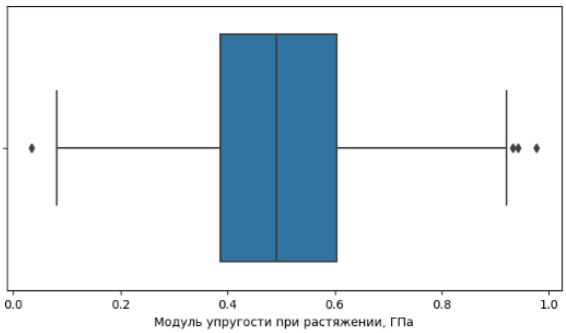

Count
Температура вспышки, С_2


Температура вспышки, С_2


Count
Поверхностная плотность, г/м2


Поверхностная плотность, г/м2


Count
Модуль упругости при растяжении, ГПа


Модуль упругости при растяжении, ГПа

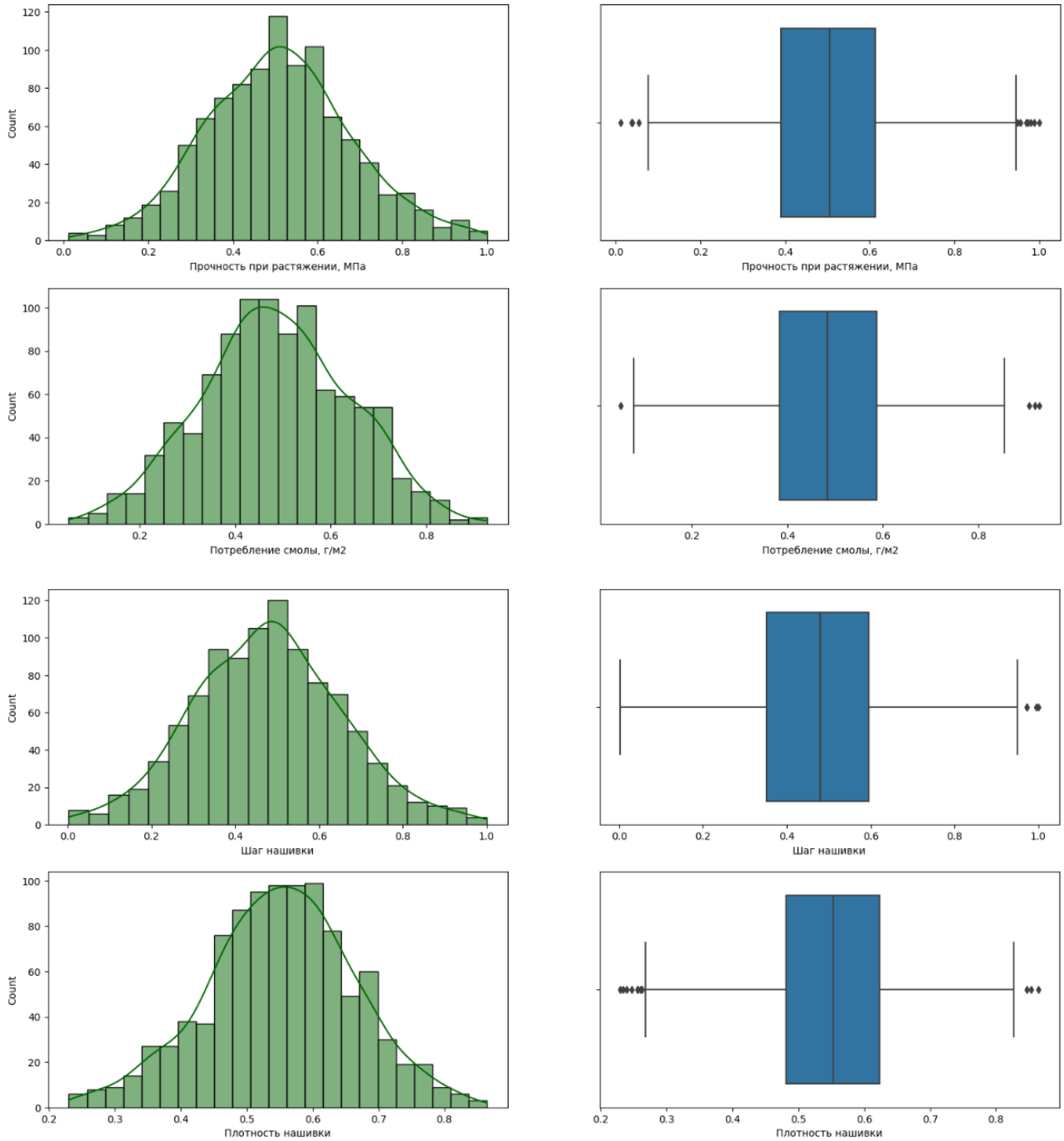


Рисунок 7 – Гистограммы распределения и ящики с усами после нормализации данных и удаления выбросов, исключая угол нашивки как категориальный признак

Гистограммы показывают нормальное распределение, за исключением признака Угол нашивки, который имеет всего два значения: 0 и 90 градусов.

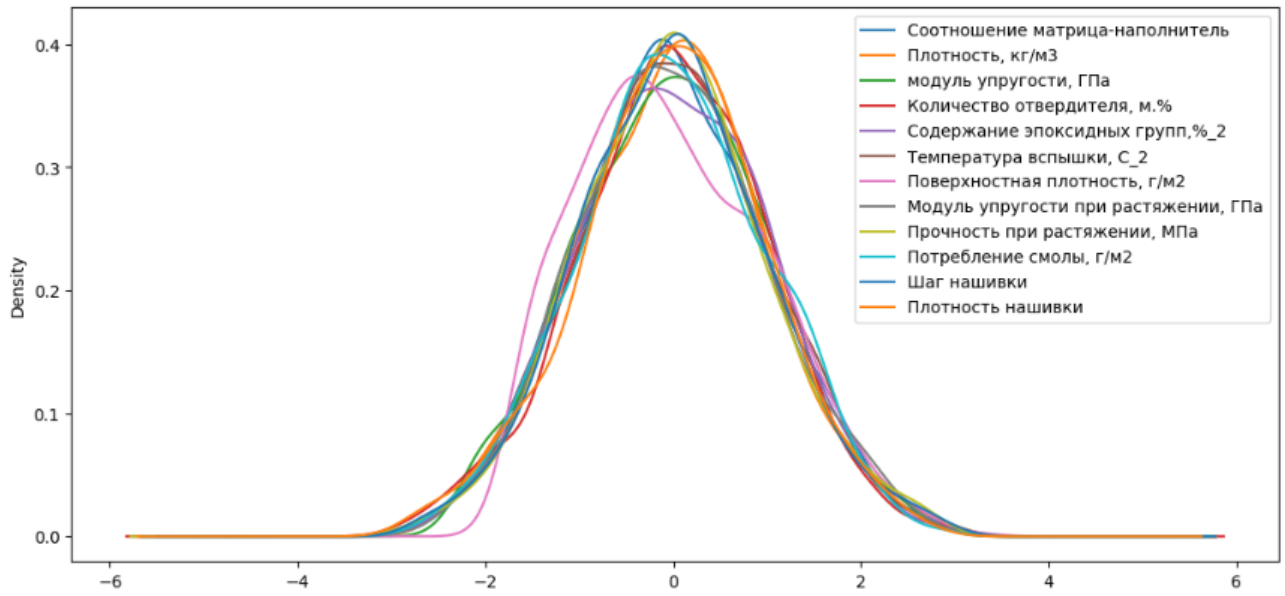


Рисунок 8 – Визуализированные данные после нормализации

На рисунке 8 изображена визуализация данных после стандартизации и нормализации, исключая угол нашивки как категориальный признак.

2.2 Разработка и обучение модели

Для подбора лучшей модели для прогноза модуля упругости при растяжении и прочности при растяжении были взяты следующие модели:

– LinearRegression — линейная регрессия (раздел 1.2.1);

– Ridge — гребневая регрессия (раздел 1.2.2);

– Lasso — лассо-регрессия (раздел 1.2.3);

– SVR — метод опорных векторов (раздел 1.2.4);

– DecisionTreeRegressor — дерево решений (раздел 1.2.5);

– RandomForestRegressor — случайный лес (раздел 1.2.6);

– GradientBoostingRegressor – градиентный бустинг (раздел 1.2.7).

В качестве базовой модели взят DummyRegressor, возвращающий среднее значение целевого признака.

Порядок разработки модели для каждого параметра и для каждого выбранного метода можно разделить на следующие этапы:

- разделение нормализованных данных на обучающую и тестовую выборки (в соотношении 70 на 30%);
- проверка моделей при стандартных значениях;
- сравнение с результатами модели, выдающей среднее значение;
- создание графика;
- сравнение моделей по метрике MAE;
- поиск сетки гиперпараметров, по которым будет происходить оптимизация модели;
- оптимизация подбора гиперпараметров модели с помощью выбора по сетке и перекрёстной проверки;
- подстановка оптимальных гиперпараметров в модель и обучение модели на тренировочных данных;
- оценка полученных данных;
- сравнение со стандартными значениями.

В качестве параметра оценки выбран коэффициент детерминации (R2).

2.2.1 Оценка работы моделей и подбор гиперпараметров применительно к задаче прогнозирования модуля упругости при растяжении

Метрики работы выбранных моделей с гиперпараметрами по умолчанию, полученные с помощью перекрестной проверки на тестовом множестве, приведены на рисунке 9.

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.025801	-3.013795	-2.408483	-0.032937	-7.702150
LinearRegression	-0.040280	-3.033795	-2.419878	-0.033087	-7.785673
Ridge	-0.040201	-3.033684	-2.419806	-0.033087	-7.785518
Lasso	-0.025801	-3.013795	-2.408483	-0.032937	-7.702150
SVR	-0.064043	-3.067918	-2.446795	-0.033437	-7.896264
KNeighborsRegressor	-0.218550	-3.279741	-2.632233	-0.035947	-8.352322
DecisionTreeRegressor	-1.228511	-4.423601	-3.583321	-0.048992	-11.451281
RandomForestRegressor	-0.087091	-3.100472	-2.472667	-0.033824	-8.005775
GradientBoostingRegressor	-0.155287	-3.197943	-2.524645	-0.034507	-8.531914

Рисунок 9 – Метрики работы выбранных моделей с гиперпараметрами по умолчанию

Далее были произведены поиск сетки гиперпараметров, по которым будет происходить оптимизация модели, и оптимизация подбора гиперпараметров модели с помощью выбора по сетке (GridSearchCV) и перекрёстной проверки (cross validation K-fold). После подбора гиперпараметров были получены метрики, приведенные на рисунке 10.

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=410, positive=True, solver='lbfgs')	-0.028041	-3.016480	-2.402560	-0.032854	-7.739427
Lasso(alpha=0.1)	-0.028228	-3.017201	-2.405317	-0.032891	-7.738947
SVR(C=0.05, kernel='poly')	-0.021899	-3.007710	-2.397686	-0.032754	-7.738811
KNeighborsRegressor(n_neighbors=53)	-0.037535	-3.030675	-2.416598	-0.033056	-7.720174
DecisionTreeRegressor(max_depth=2, max_features=5, random_state=0, splitter='random')	-0.011775	-2.991498	-2.387448	-0.032646	-7.755630
RandomForestRegressor(bootstrap=False, criterion='absolute_error', max_depth=4, max_features=1, n_estimators=50, random_state=0)	-0.035630	-3.028397	-2.411007	-0.032948	-7.862964
GradientBoostingRegressor(max_depth=1, max_features=1, n_estimators=50, random_state=0)	-0.046504	-3.044165	-2.420581	-0.033099	-7.861759

Рисунок 10 – Метрики работы выбранных моделей с настроенными гиперпараметрами

Модель после настройки гиперпараметров показала результат немного лучше. Однако, не удалось добиться положительного значения R2. Согласно метрикам, на рисунке 10, лучшим оказался метод DecisionTreeRegressor (параметры: max_depth=2, max_features=5, random_state=0, splitter='random'). Самая лучшая модель дает коэффициент детерминации близкий к нулю, что соответствует базовой модели.

2.2.2 Оценка работы моделей и подбор гиперпараметров применительно к задаче прогнозирования прочности при растяжении

Метрики работы выбранных моделей с гиперпараметрами по умолчанию, полученные с помощью перекрестной проверки на тестовом множестве, приведены на рисунке 11.

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.026342	-484.807293	-380.185919	-0.167277	-1253.431489
LinearRegression	-0.033167	-486.474961	-385.238945	-0.168730	-1258.639354
Ridge	-0.033108	-486.460907	-385.220677	-0.168723	-1258.607692
Lasso	-0.032865	-486.404296	-384.988615	-0.168664	-1258.645065
SVR	-0.026718	-484.947303	-380.218028	-0.166329	-1262.072451
DecisionTreeRegressor	-1.069986	-682.369273	-545.666323	-0.234120	-1773.372125
RandomForestRegressor	-0.078820	-496.382732	-399.731376	-0.174450	-1243.269911
GradientBoostingRegressor	-0.148984	-511.332319	-409.803025	-0.178062	-1291.511611

Рисунок 11 – Метрики работы выбранных моделей с гиперпараметрами по умолчанию

Далее были произведены поиск сетки гиперпараметров, по которым будет происходить оптимизация модели, и оптимизация подбора гиперпараметров модели с помощью выбора по сетке (GridSearchCV) и перекрёстной проверки (cross validation K-fold). После подбора гиперпараметров были получены метрики, приведенные на рисунке 12.

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=990, positive=True, solver='lbfgs')	-0.023267	-484.113037	-380.018387	-0.167120	-1257.022616
Lasso(alpha=50)	-0.026342	-484.807293	-380.185919	-0.167277	-1253.431489
SVR(C=0.001, kernel='sigmoid')	-0.026311	-484.853433	-379.936358	-0.166191	-1261.977540
DecisionTreeRegressor(criterion='absolute_error', max_depth=1, max_features=5, random_state=0)	-0.026397	-484.768578	-378.716225	-0.166234	-1255.193290
RandomForestRegressor(bootstrap=False, criterion='absolute_error', max_depth=2, max_features=1, n_estimators=50, random_state=0)	-0.020263	-483.477346	-380.126241	-0.166222	-1248.234411
GradientBoostingRegressor(max_depth=1, max_features=1, n_estimators=50, random_state=0)	-0.032345	-486.308546	-382.902477	-0.168049	-1244.425207

Рисунок 12 – Метрики работы выбранных моделей с настроенными гиперпараметрами

Модель после настройки гиперпараметров показала результат немного лучше. Однако, не удалось добиться положительного значения R2. Согласно

метрикам, на рисунке 12, лучшим оказался метод RandomForestRegressor (параметры: bootstrap=False, criterion='absolute_error', max_depth=2, max_features=1, n_estimators=50, random_state=0). Самая лучшая модель дает коэффициент детерминации близкий к нулю, что соответствует базовой модели.

Прочность при растяжении и модуль упругости не имеет линейной зависимости. Все использованные модели не справились с задачей. Результат неудовлетворительный. Свойства композитных материалов в первую очередь зависят от используемых материалов.

2.3 Тестирование модели

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовых выборках. Результат неудовлетворительный.

	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.004581	-3.184357	-2.599556	-0.035346	-9.006040
Лучшая модель (дерево решений)	-0.016276	-3.202838	-2.601556	-0.035307	-9.247423

Рисунок 13 – Результат оценки точности применительно к задаче прогнозирования модуля упругости при растяжении

	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.002151	-474.683874	-382.633231	-0.164142	-1317.160974
Лучшая модель (Случайный лес)	-0.013125	-477.275842	-385.630487	-0.164716	-1332.911750

Рисунок 14 – Результат оценки точности применительно к задаче прогнозирования прочности при растяжении

Метрики работы лучших моделей на тестовом множестве и сравнение с базовой отражены на рисунках 13 и 14. Они подтверждают: полученные модели хуже базовой. Результат исследования отрицательный. Не удалось получить модели, которая могла бы оказать помощь в принятии решений специалисту предметной области.

2.4 Написание нейронной сети, рекомендующей соотношение матрица-наполнитель

2.4.1 MLPRegressor из библиотеки sklearn

С помощью класса MLPRegressor построена нейронная сеть следующей архитектуры:

– слоев: 8;

– нейронов на каждом слое: 24;

– активационная функция: relu;

– оптимизатор: adam;

– пропорция разбиения данных на тестовые и валидационные: 30%;

– ранняя остановка, если метрики на валидационной выборке не улучшаются;

– количество итераций: 5000.

Обучим и оценим модель, посмотрим на потери, зададим функцию для визуализации факт/прогноз для результатов моделей.

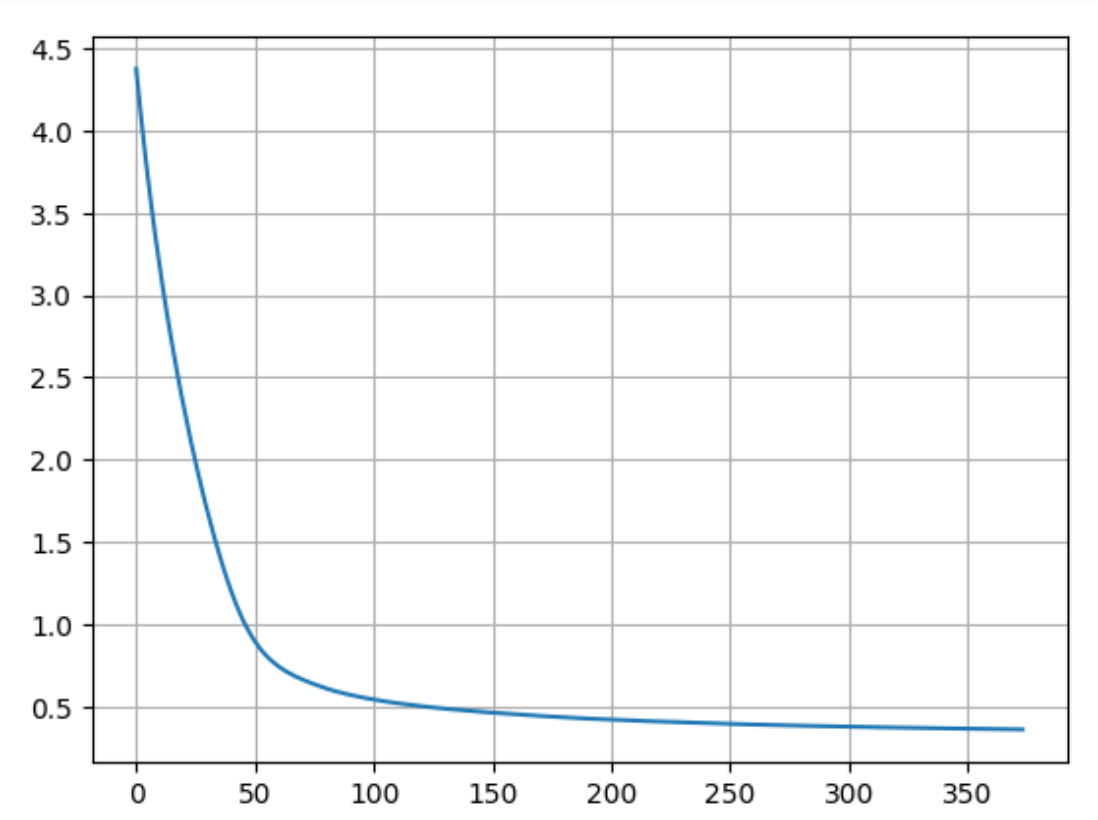


Рисунок 15 – График ошибки модели

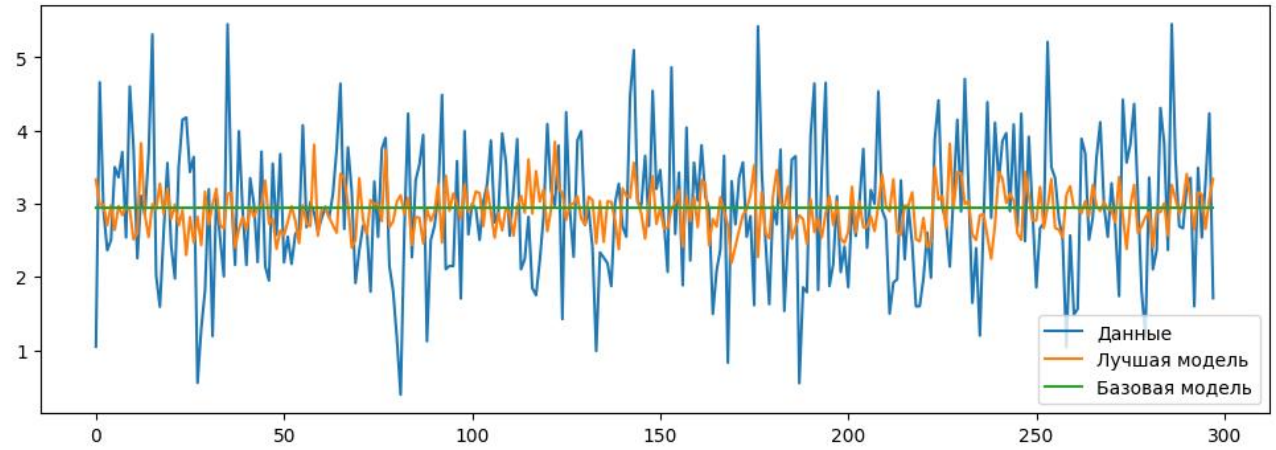


Рисунок 16 – Визуализация работы модели

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.001195	-0.938472	-0.747204	-0.348824	-2.556776
MLPRegressor	-0.053299	-0.962583	-0.763131	-0.348466	-3.153270

Рисунок 17 – Сравнение предсказаний базовой модели и лучшей модели на тестовом множестве

Несмотря на красивый график с рисунка 16, метрики говорят об отсутствии результата, который можно внедрить. Не удовлетворившись таким результатом, создадим другую модель глубокого обучения с другой архитектурой.

2.4.2 Нейросеть из библиотеки tensorflow

С помощью класса keras.Sequential построена нейронная сеть со следующими параметрами:

– входной слой для 12 признаков;

– выходной слой для 1 признака;

– скрытых слоев: 8;

– нейронов на каждом скрытом слое: 24;

– активационная функция скрытых слоев: relu;

– оптимизатор: Adam;

– loss-функция: MeanAbsolutePercentageError.

Обучим и оценим модель, посмотрим на потери, зададим функцию для визуализации факт/прогноз для результатов моделей.

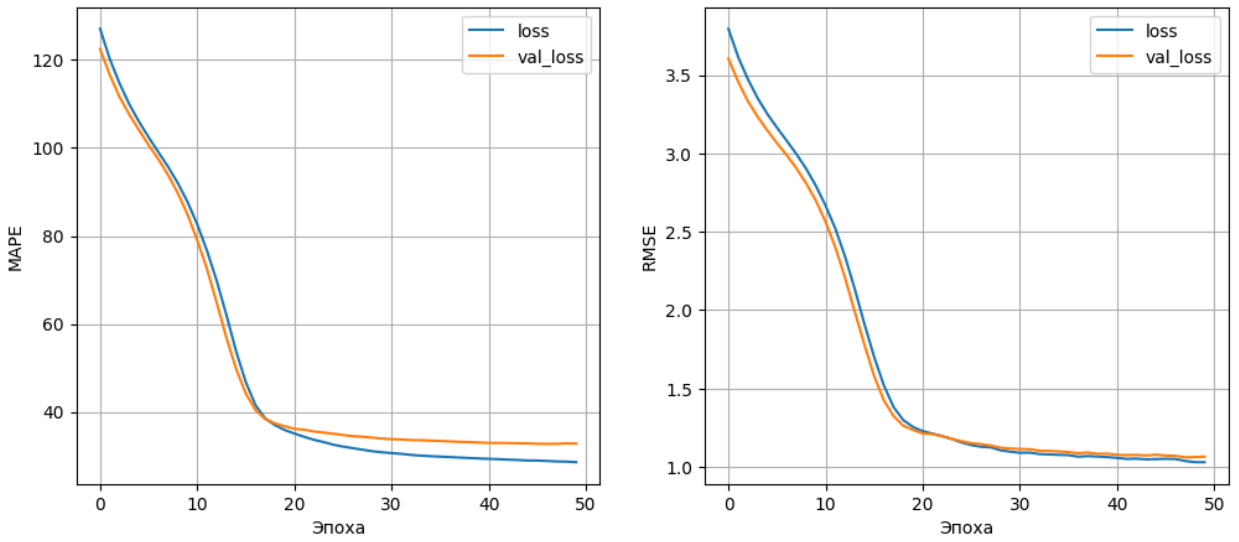


Рисунок 18 – График ошибки модели

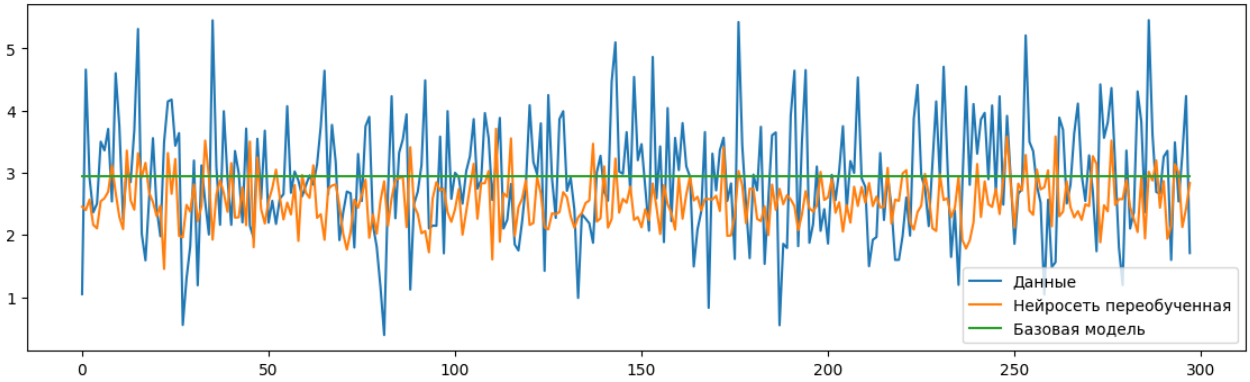


Рисунок 19 – Визуализация работы переобученной модели

Одним из способов борьбы с переобучением может быть ранняя остановка обучения, если val_loss начинает расти. Для этого в tensorflow используются callbacks. Попробуем взять нейросеть с той же архитектурой и запустить обучение с ранней остановкой. Очевидно, что решение проблемы переобучения повышает точность модели на новых данных.

Создаем модель с той же архитектурой. Обучим и оценим модель, посмотрим на потери, зададим функцию для визуализации факт/прогноз для результатов моделей.

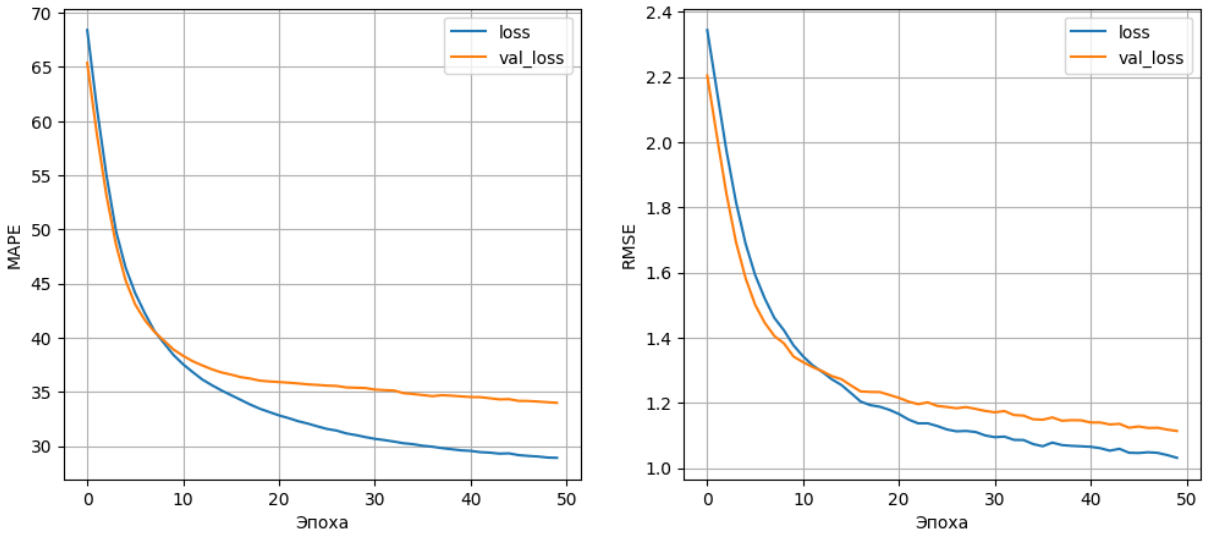


Рисунок 20 – График ошибки модели

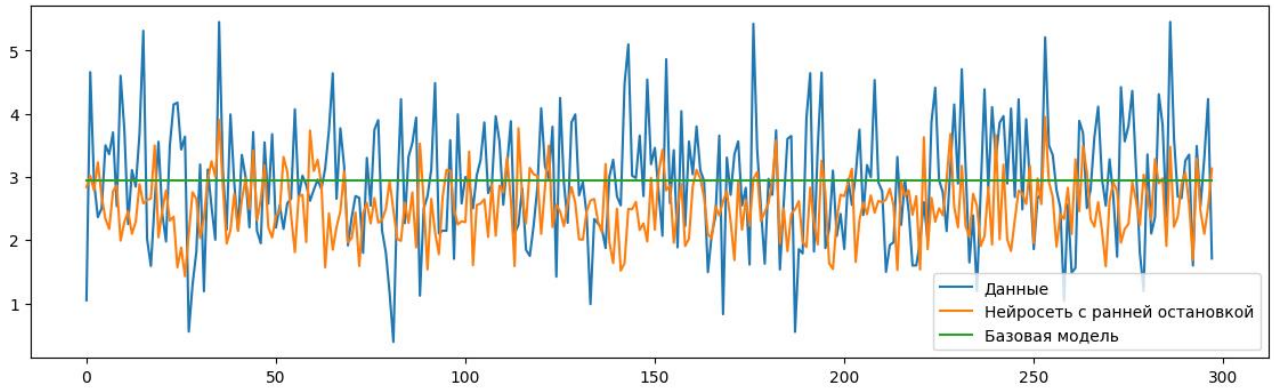


Рисунок 21 – Визуализация работы модели с ранней остановкой

Еще одним методом борьбы с переобучением является добавление Dropout-слоев. Построим модель аналогичной архитектуры, только после каждого скрытого слоя добавим слой Dropout с параметром 0.05. Такой слои слои выключат 5% случайных нейронов на каждом слое.

Построение аналогичной модели с Dropout слоем. Обучим и оценим модель, посмотрим на потери, зададим функцию для визуализации факт/прогноз для результатов моделей.

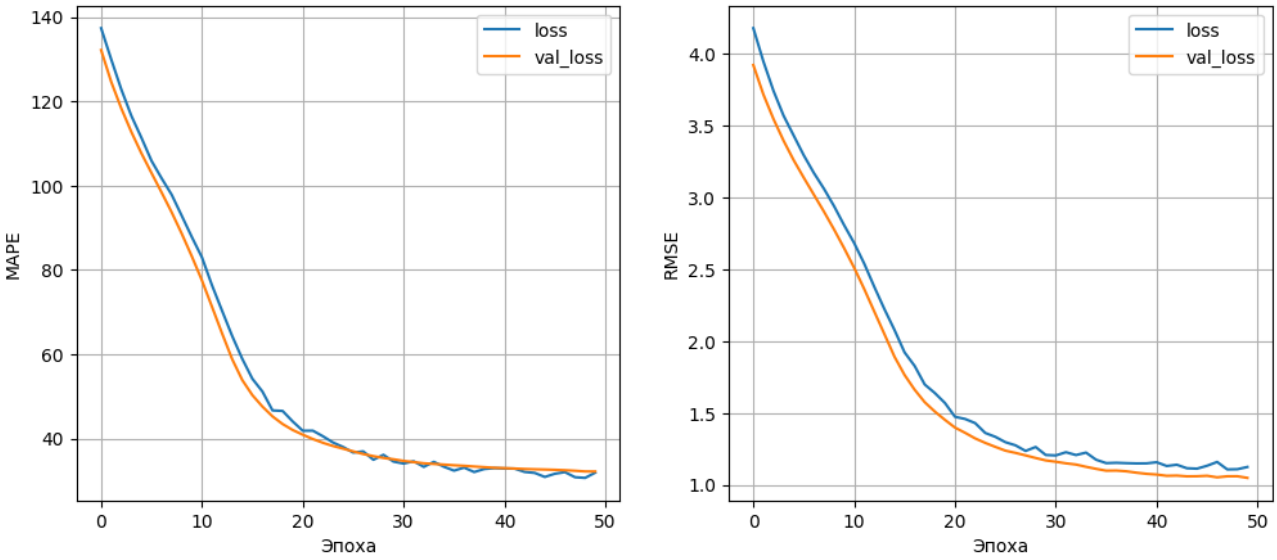


Рисунок 22 – График ошибки модели

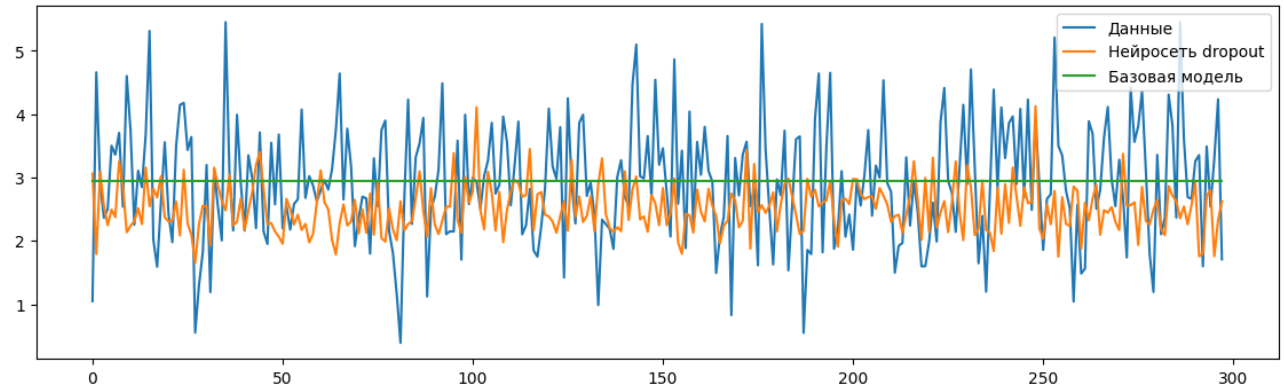


Рисунок 23 – Визуализация работы нейросети Dropout

Использование ранней остановки сокращает время на обучение модели, а использование Dropout увеличивает. Но уменьшается риск, что мы остановились слишком рано.

На рисунке 24 можно увидеть сравнение предсказаний на тестовом множестве.

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.001195	-0.938472	-0.747204	-0.348824	-2.556776
Нейросеть переобученная	-0.349450	-1.089533	-0.863589	-0.347206	-3.528243
Нейросеть с ранней остановкой	-0.424751	-1.119519	-0.891926	-0.354251	-2.832593
Нейросеть dropout	-0.310950	-1.073879	-0.846670	-0.327881	-3.098548

Рисунок 24 – Сравнение предсказаний на тестовом множестве

Визуализация результатов показывает, что нейросеть из библиотеки tensorflow старалась подстроиться к данным. Выглядят результаты «похоже», но метрики разочаровывают. Лучшая обобщающая способность и меньшие значения ошибок на тестовом множестве оказались у нейросети, обученной с использованием метода Dropout. Но и она предсказывает гораздо хуже базовой модели.

2.5 Разработка приложения

В приложении необходимо реализовать следующие функции:

– выбор целевой переменной для предсказания;

– ввод входных параметров;

– проверка введенных параметров;

– загрузка сохраненной модели, получение и отображение прогноза выходных параметров.

Эту задачу получилось решить. Скриншоты разработанного веб-приложения приведены на рисунках 25, 26 и 27.

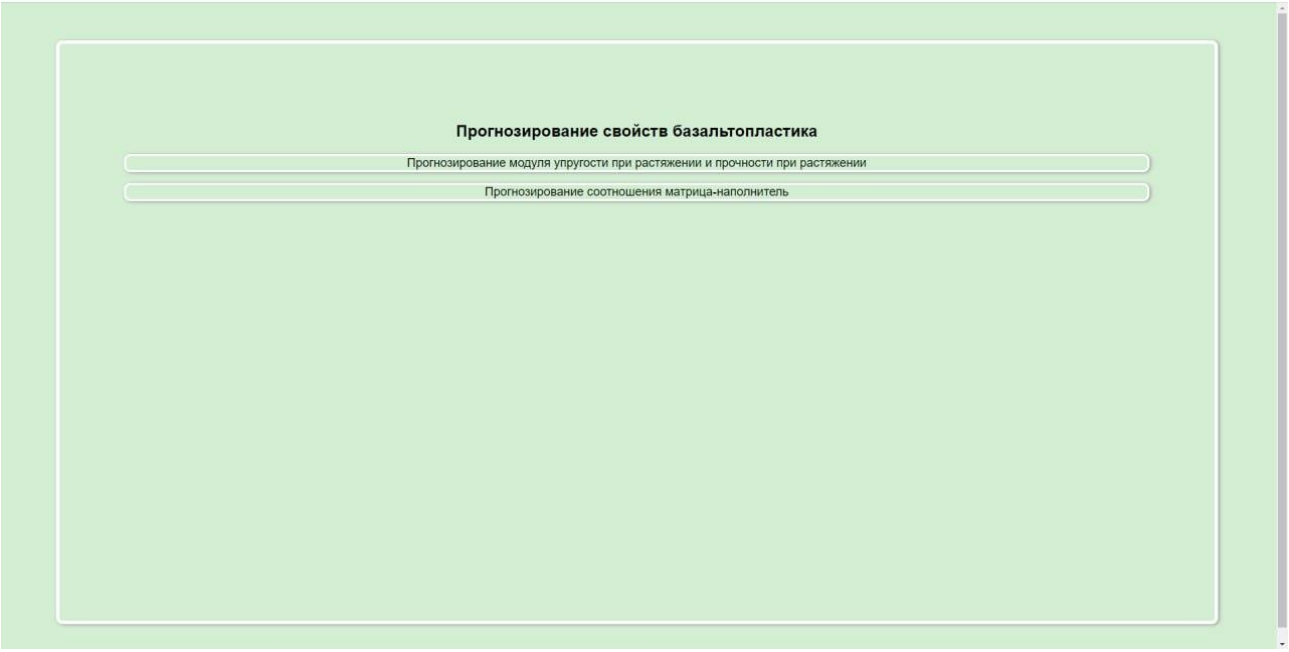


Рисунок 25 – Скриншот приложения, выбор целевой переменной для предсказания

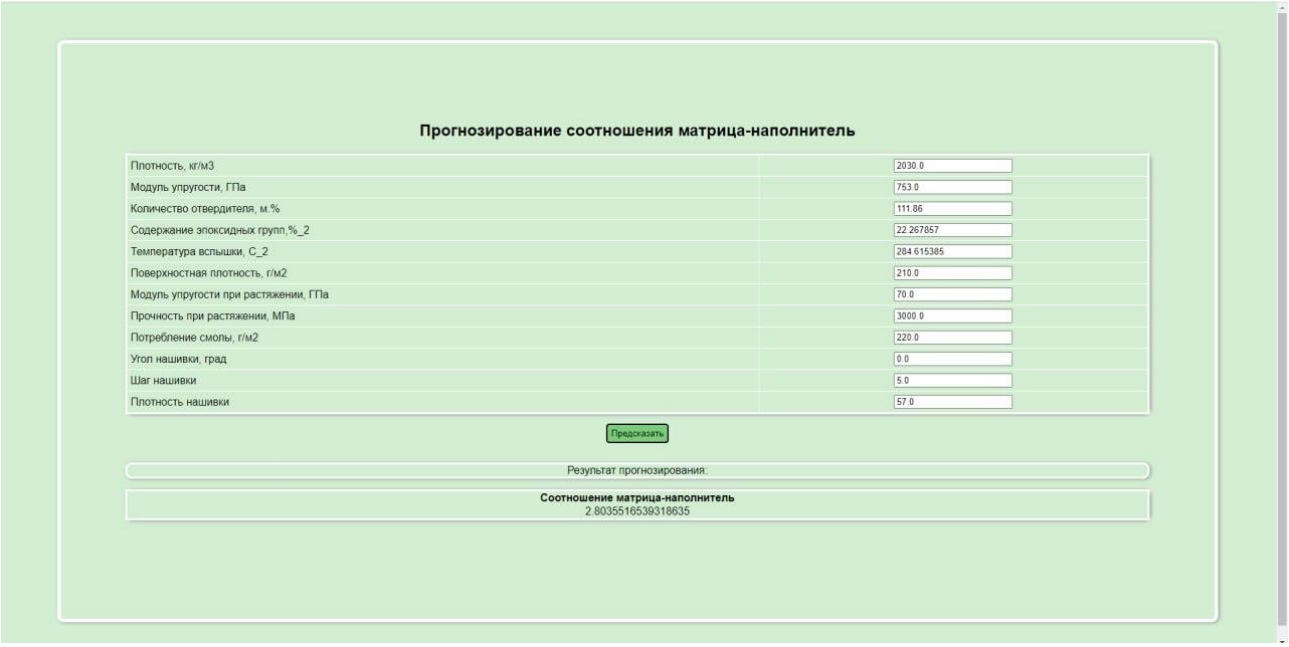


Рисунок 26 – Скриншот приложения: ввод параметров и результат: соотношение матрица-наполнитель

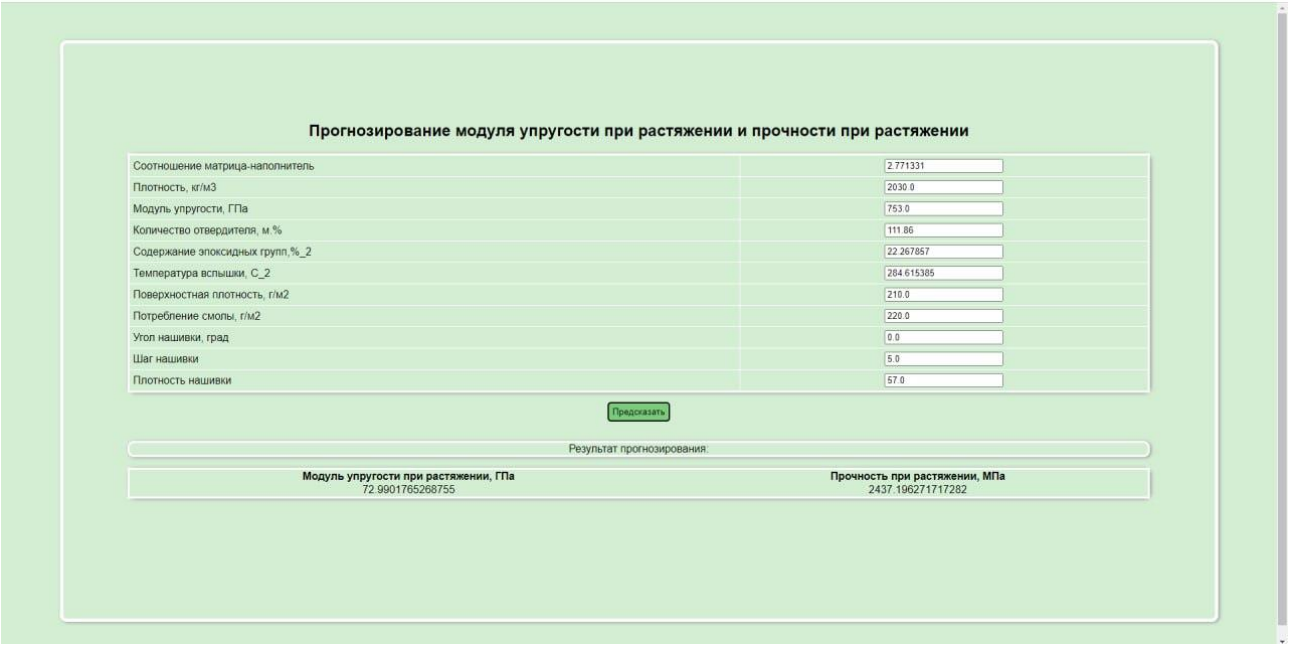


Рисунок 27 Скриншот приложения, ввод параметров и результаты: модуль упругости при растяжении и прочность при растяжении

2.6 Создание удаленного репозитория и загрузка результатов работы на него.

Для данного исследования был создан удаленный репозиторий на GitHub, который находится по адресу https://github.com/Katya-mem/KompositProject. На него были загружены результаты работы, ноутбук с решением и приложением.

Заключение

Данная исследовательская работа позволяет сделать некоторые основные выводы по теме. Распределение полученных данных в объединённом датасете близко к нормальному, но коэффициенты корреляции между парами признаков стремятся к нулю. Использованные при разработке моделей подходы не позволили получить сколько-нибудь достоверных прогнозов. Применённые модели регрессии не показали высокой эффективности в прогнозировании свойств композитов.

В ходе выполнения данной работы мы прошли практически весь Dataflow pipeline, рассмотрели большую часть операций и задач, которые приходится выполнять специалисту по работе с данными.

Этот поток операций и задач включает:

– изучение теоретических методов анализа данных и машинного обучения;

– изучение основ предметной области, в которой решается задача;

– извлечение и трансформацию данных. Здесь нам был предоставлен готовый набор данных, поэтому через трудности работы с разными источниками и парсингом данных мы еще не соприкоснулись;

– проведение разведочного анализа данных статистическими методами;

– DataMining — извлечение признаков из датасета и их анализ;

– разделение имеющихся, в нашем случае размеченных, данных на обучающую, валидационную, тестовую выборки;

– выполнение предобработки (препроцессинга) данных для обеспечения корректной работы моделей;

– построение аналитического решения. Это включает выбор алгоритма решения и модели, сравнение различных моделей, подбор гиперпараметров модели;

– визуализация модели и оценка качества аналитического решения;

– сохранение моделей;

– разработка и тестирование приложения для поддержки принятия решений специалистом предметной области, которое использовало бы найденную модель;

– внедрение решения и приложения в эксплуатацию. Этот блок задач мы пока не затронули.

Был сделан вывод, что невозможно определить из свойств материалов соотношение «матрица – наполнитель». Данный факт не указывает на то, что прогнозирование характеристик композитных материалов на основании предоставленного набора данных невозможно, но может указывать на недостатки базы данных, подходов, использованных при прогнозе, необходимости пересмотра инструментов для прогнозирования.

Необходимы дополнительные вводные данные, получение новых результирующих признаков в результате математических преобразований, релевантных доменной области, консультации экспертов предметной области, новые исследования, работа эффективной команды, состоящей из различных учёных.

В целом прогнозирование конечных свойств/характеристик композитных материалов без изучения материаловедения, погружения в вопрос экспериментального анализа характеристик композитных материалов не демонстрирует сколько-нибудь удовлетворительных результатов. Проработка моделей и построение прогнозов требует внедрения в процесс производных от имеющихся показателей для выявления иного уровня взаимосвязей. Отсюда,

также учитывая отсутствие корреляции между признаками, делаем вывод, что текущим набором алгоритмов задача не решается, возможно, решается трудно или не решается совсем.

Библиографический список:

1. Композиционные материалы : учебное пособие для вузов / Д. А. Иванов, А. И. Ситников, С. Д. Шляпин ; под редакцией А. А. Ильина. — Москва : Издательство Юрайт, 2019 — 253 с. — (Высшее образование). — Текст : непосредственный.
2. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
3. ГрасД. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2021. - 416 с.: ил.
4. Документация по языку программирования python: – Режим доступа: https://docs.python.org/3.8/index.html.
5. Документация по библиотеке numpy: – Режим доступа: https://numpy.org/doc/1.22/user/index.html#user.
6. Документация по библиотеке pandas: – Режим доступа: https://pandas.pydata.org/docs/user_guide/index.html#user-guide.
7. Документация по библиотеке matplotlib: – Режим доступа: https://matplotlib.org/stable/users/index.html.
8. Документация по библиотеке seaborn: – Режим доступа: https://seaborn.pydata.org/tutorial.html.
9. Документация по библиотеке sklearn: – Режим доступа: https://scikit-learn.org/stable/user_guide.html.
10. Документация по библиотеке keras: – Режим доступа: https://keras.io/api/.
11. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-

neighbour-2021-07-19. (дата обращения: 07.06.2022)
12. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам сле-дует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/(дата обращения: 01.06.2022).
13. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
14. Абросимов Н.А.: Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010
15. Абу-Хасан Махмуд, Масленникова Л. Л.: Прогнозирование свойств композиционных материалов с учётом наноразмера частиц и акцепторных свойств катионов твёрдых фаз, статья 2006 год
16. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.
17. Гафаров, Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учеб. пособие /Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Издательство Казанского университета, 2018. – 121 с.
18. Yury Kashnitsky. Открытый курс машинного обучения. Тема 3. Классификация, деревья решений и метод ближайших соседей: – Режим доступа: https://habr.com/ru/company/ods/blog/322534/.
19. Yury Kashnitsky. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес: – Режим доступа: https://habr.com/ru/company/ods/blog/324402/.
20. Alex Maszański. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest): – Режим доступа: https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12.
21. Alex Maszański. Решаем задачи машинного обучения с помощью алгоритма градиентного бустинга: – Режим доступа: https://proglib.io/p/reshaem-

zadachi-mashinnogo-obucheniya-s-pomoshchyu-algoritma-gradientnogo-bustinga-2021-11-25.