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 Курцева Лариса Юрьевна

Москва, 2023

Содержание

Введение

Задача данной работы заключается в прогнозировании характеристик компонентов композиционных материалов на основе данных о составе композитов с использованием подхода, ориентированного на данные.

Композитный материал или просто композит – это материал, состоящий из двух или более компонентов, каждый из которых обладает различными физическими и химическими свойствами. При этом в сочетании друг с другом они создают новый материал или улучшают характеристики одного из них.

В настоящее время композиты являются неотъемлемой частью нашей жизни и широко используются в различных направлениях промышленности, например, в строительной, как в гражданской, так и в индустриальной, в судостроении, авиастроении, на железной дороге и так далее.

Исходя из этого и обладая знаниями, накопленными в течение многих лет, значимость композитов и необходимость в них постоянно растут. Современные композиты также демонстрируют некоторые преимущества перед традиционными материалами, в том числе в долговечности, прочности и легковесности. Необходимо также и снижение стоимости производства композитов, отсюда мы делаем вывод, что данная работа актуальна и значима.

1. Аналитическая часть

1.1. Постановка задачи

Данная работа направлена на разработку моделей прогнозирования значений:

- Модуля упругости при растяжении, Гпа;

- Прочности при растяжении, Мпа;

- Соотношения матрица-наполнитель.

А также на разработку приложения для практического использования при прогнозировании значений.

1.2. Характеристика датасета

Датасет представлен в виде двух файлов формата excel:

- X_bp;
- X_nup.

Таблица 1 – Описание файлов датасета

Файл	Признаки	Индексы	Строки
X_bp	10	1	1023
X_nup	3	1	1040

После объединения двух файлов по индексу тип INNER, и удаления неинформативного индекса, мы получаем единый датасет с 13 признаками и 1023 строками, а именно:

Таблица 2 – Описание датасета

Столбец	Тип данных	Пропуски	Уникальные значения
1	2	3	4
Соотношение матрица-наполнитель	float64	нет	1014
Плотность, кг/м3	float64	нет	1013

Продолжение таблицы 2

1	2	3	4
модуль упругости, ГПа	float64	нет	1020
Количество отвердителя, м.%	float64	нет	1005
Содержание эпоксидных групп,%_2	float64	нет	1004
Температура вспышки, С_2	float64	нет	1003
Поверхностная плотность, г/м2	float64	нет	1004
Модуль упругости при растяжении, ГПа	float64	нет	1004
Прочность при растяжении, МПа	float64	нет	1004
Потребление смолы, г/м2	float64	нет	1003
Угол нашивки, град	float64	нет	2
Шаг нашивки	float64	нет	989
Плотность нашивки	float64	нет	988

Признак «Угол нашивки, град» считается категориальным и представлен только двумя значениями 0^o и 90^o, который при дальнейшем препроцессинге и с помощью кодировщика LabelEncoder будет преобразован в числовой.

1.3. Статистический анализ датасета

Таблица 3 – Статистическое описание датасета

Столбец	Среднее значение	Медианное значение	Минимальное значение	Максимальное значение
1	2	3	4	5
Соотношение матрица-наполнитель	2.9304	2.9069	0.3894	5.5917
Плотность, кг/м3	1975.7349	1977.6217	1731.7646	2207.7735
модуль упругости, ГПа	739.9232	739.6643	2.4369	1911.5365
Количество отвердителя, м.%	110.5708	110.5648	17.7403	198.9532
Содержание эпоксидных групп,%_2	22.2444	22.2307	14.2550	33.0000
Температура вспышки, С_2	285.8822	285.8968	100.0000	413.2734

Продолжение таблицы 3

1	2	3	4	5
Поверхностная плотность, г/м2	482.7318	451.8644	0.6037	1399.5424
Модуль упругости при растяжении, ГПа	73.3286	73.2688	64.0541	82.6821
Прочность при растяжении, МПа	2466.9228	2459.5245	1036.8566	3848.4367
Потребление смолы, г/м2	218.4231	219.1989	33.8030	414.5906
Угол нашивки, град	44.2522	0.0000	0.0000	90.0000
Шаг нашивки	6.8992	6.9161	0.0000	14.4405
Плотность нашивки	57.1539	57.3419	0.0000	103.9889

1.3.1. Корреляция

Корреляция – взаимозависимость двух или нескольких случайных величин. При изменении значения одной переменной происходит закономерное изменение другой (-их) переменной (-ых). Тепловая карта, на рисунке 1, показывает, что взаимосвязь между двумя или несколькими переменными нашего датасета очень слабая. Надо понимать, что корреляционная зависимость отражает только взаимосвязь между переменными, но не говорит о причинах и следствиях.

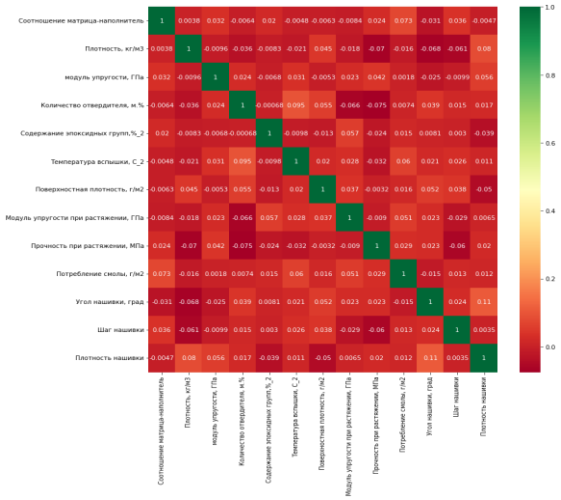


Рисунок 1 - Тепловая карта корреляции

1.3.2. Гистограммы распределения, диаграммы ящика с усами, попарные графики рассеяния точек

Для демонстрации распределения значений, выбросов и потенциальных взаимосвязей между переменными, построены следующие диаграммы, гистограммы и графики:

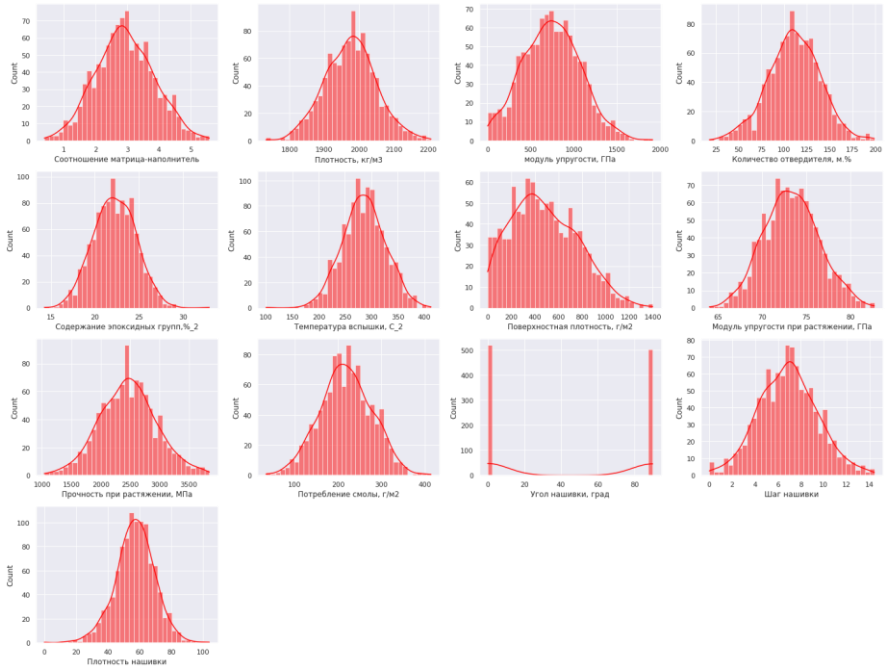


Рисунок 2 - Гистограммы распределения

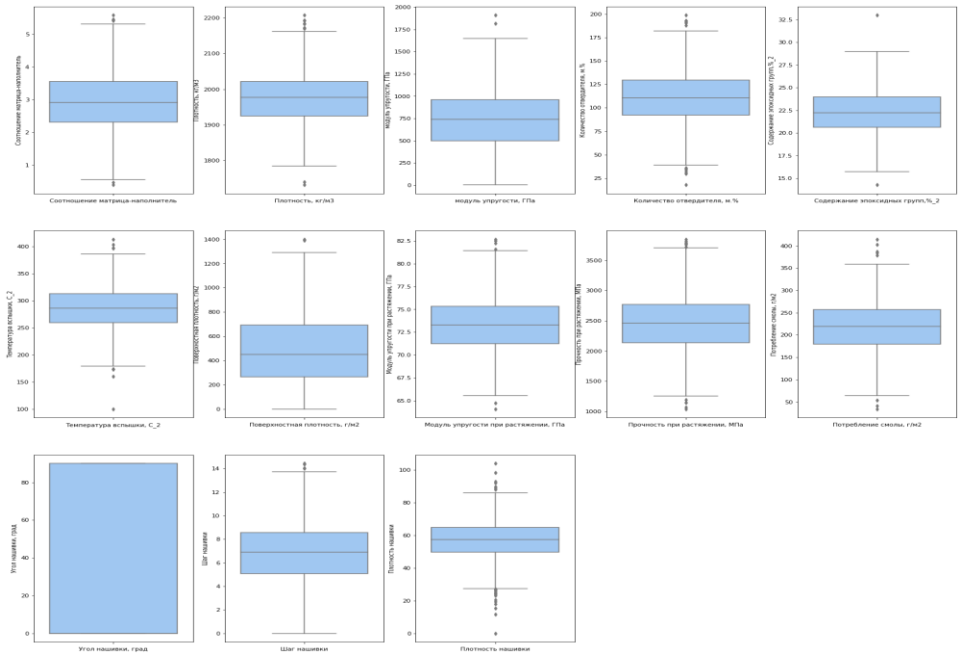


Рисунок 3 - Диаграммы ящиков с усами

Диаграммы ящиков с усами показывают наличие выбросов по признакам, гистограммы распределения показывают, что распределение данных равномерное. Опять же плотность ядер на диаграммах рассеяния точек на Рисунке 3 показывает равномерное распределение переменных.

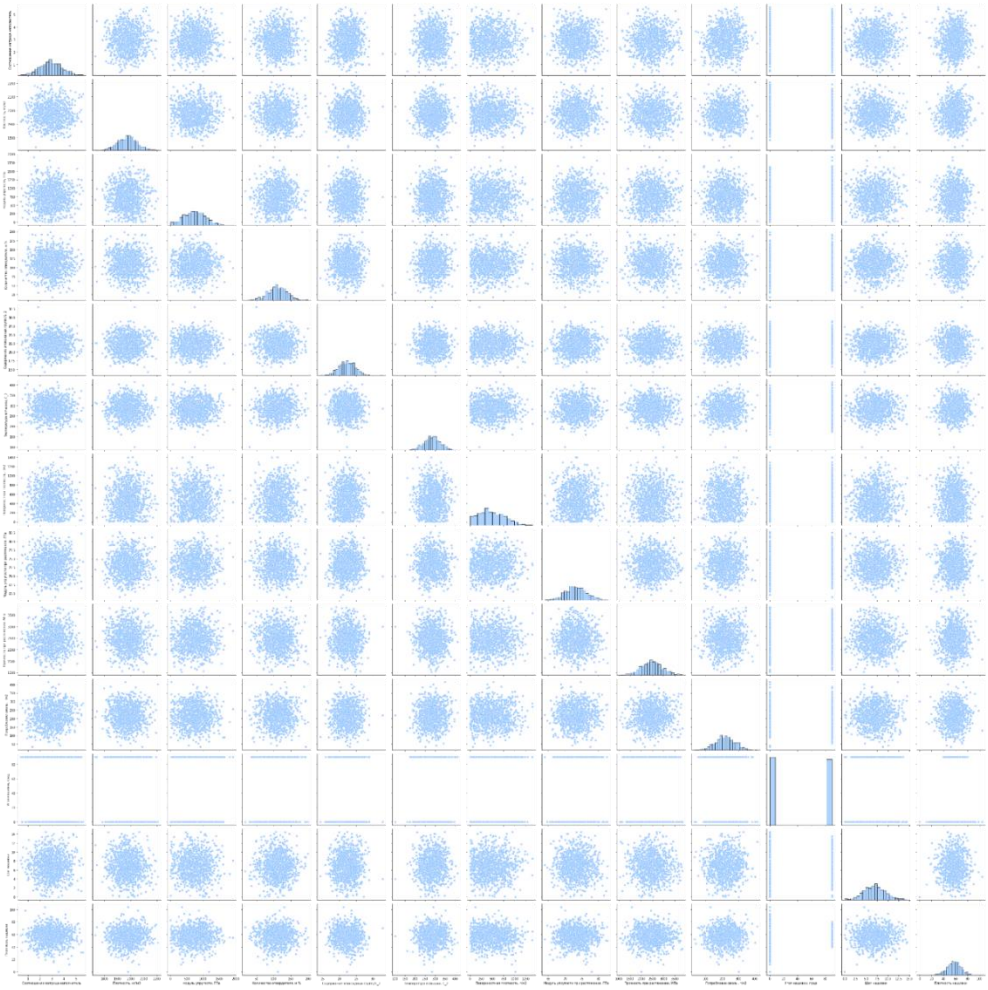


Рисунок 4 – Попарные диаграммы рассеяния точек

1.3.3. Выбросы

Как уже показано выше на рисунке 3, в некоторых признаках датасета присутствуют выбросы. Для дальнейшей работы с датасетом рекомендуется удалить выбросы, чтобы избежать их влияния на анализ данных.

Для расчета выбросов возьмем межквартильный размах IQR, т.е. средние 50% значений и рассчитаем выбросы по следующей формуле:

IQR = Q3 - Q1,

где Q1 - первый квартиль (0.25), а Q3 – третий квартиль (0.75),

IQR –разница между Q3 и Q1.

После удаления выбросов размер датасета составляет 936 строк и 13 признаков.

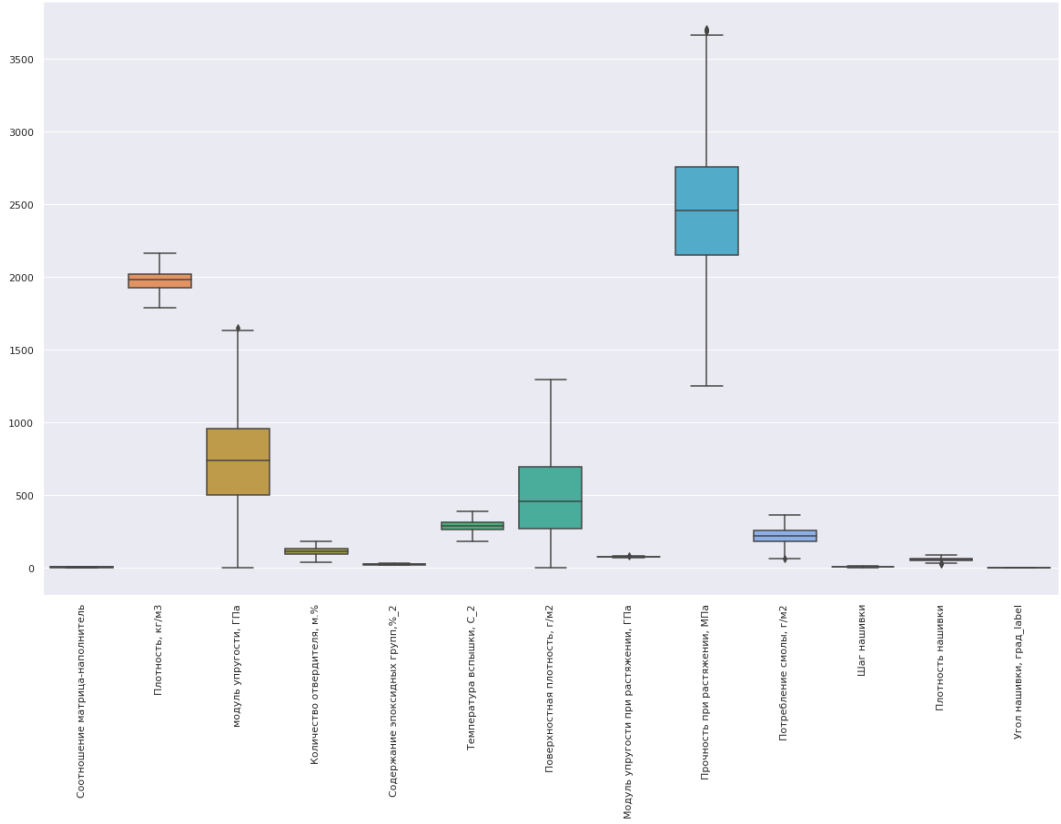


Рисунок 5 – Ящики с усами после удаления выбросов

1.4. Описание методов

Предсказание значения целевой переменной — это задача регрессии. Регрессия — это зависимость среднего значения какой-либо величины от некоторой другой величины или нескольких величин. Таким образом, задача регрессии заключается в получении вещественного числа.

Существует несколько методов регрессии. При работе с нашим датасетом будут использоваться следующие модели регрессии, для определения, какая из них более работоспособна, дает меньшие ошибки и т.д.:

- линейная регрессия;

- ридж (гребневая) регрессия;

- регрессия по методу «лассо»

- регрессор случайного леса;

- градиентный бустинг;

- нейронная сеть.

1.4.1. Линейная регрессия

Регрессия—это метод, используемый для моделирования и анализа отношений между переменными, а также для того, чтобы увидеть, как эти переменные вместе влияют на получение определенного результата. Линейная регрессия (Linear Regression) относится к такому виду регрессионной модели, который состоит из взаимосвязанных переменных.

Достоинства линейной регрессии:

- скорость и простота получения модели;

- интерпретируемость модели. Линейная модель является прозрачной и понятной. По полученным коэффициентам регрессии можно судить о том, как тот или иной фактор влияет на результат, сделать на этой основе дополнительные полезные выводы;

- широкая применимость.

- изученность данного подхода.

Главный недостаток линейной регрессии заключается в том, что она может моделировать только прямые линейные зависимости, в то время как часто возникает необходимость создания модели других типов отношений между данными.

1.4.2. Ридж регрессия

Ридж (гребневая) регрессия (Ridge Regression) – это усовершенствованная версия линейной регрессии. Она заставляет алгоритм обучения не только соответствовать данным, но и сохранять веса модели как можно меньшими. Методы регуляризации зачастую позволяют добиться уменьшения дисперсии прогноза за счет незначительного увеличения его смещенности. В результате точность прогноза растет.

1.4.3. Регрессия по методу «лассо»

Регрессия по методу «лассо» (LASSO, Least Absolute Shrinkage and Selection Operator) также является усовершенствованной версией линейной регрессии. Сходна с гребневой регрессией, за исключением того, что коэффициенты регрессии могут равняться нулю (часть признаков при этом из модели исключается).

Метод заключается во введении дополнительного слагаемого регуляризации в функционал оптимизации модели, что часто позволяет получать более устойчивое решение.

1.4.4. Регрессор случайного леса

Алгоритм случайного леса (Random Forest Regressor) – один из самых универсальных алгоритмов машинного обучения. Его универсальность заключается в том, что он используется для решения задач классификации, кластеризации, поиска аномалий, регрессии и т.д. Случайный лес – это ансамблевый алгоритм, т.е. это множество решающих деревьев. В задаче регрессии ответы усредняются.

Все деревья строятся независимо по следующей схеме:

- выбирается подвыборка обучающей выборки размера, по которой строится дерево;

- для построения каждого расщепления в дереве просматриваются max_features случайных признаков;

- выбираем наилучшие признак и расщепление по нему. Дерево строится, как правило, до исчерпания выборки.

Такая схема построения соответствует главному принципу ансамблирования: базовые алгоритмы должны быть хорошими и разнообразными (поэтому каждое дерево строится на своей обучающей выборке и при выборе расщеплений есть элемент случайности).

Преимущества алгоритма заключаются в быстроте анализа, в независимости деревьев друг от друга. Хорошо подходит для анализа сложных структур, т.е. он одинаково хорошо обрабатываются как непрерывные, так и дискретные признаки. Существуют методы построения деревьев по данным с пропущенными значениями признаков. Нечувствительность к монотонным преобразованиям значений признаков.

Недостатки у алгоритма тоже есть, например, наилучшие результаты можно получить только если деревья вырастут до очень больших размеров.

1.4.5. Градиентный бустинг

Градиентный бустинг (Gradient Tree Boosting) это также ансамблевый метод, в котором алгоритмы применяются последовательно. Этот метод использует логику, в которой последующие модели учатся на ошибках предыдущих. Следовательно, наблюдения имеют неодинаковую вероятность появления в последующих моделях, а наблюдения с наибольшей ошибкой появляются чаще. (Таким образом, наблюдения выбираются не на основе процесса начальной загрузки, а на основе ошибки). Из используемых методов перечислим деревья решений, регрессоры, классификаторы и т.д. Поскольку новые алгоритмы учатся на ошибках, совершенных предшественниками, требуется меньше времени / итераций, чтобы приблизиться к фактическим прогнозам. Но мы должны тщательно выбирать критерии остановки, иначе это может привести к переобучению.

1.4.6. Нейронная сеть

Нейронная сеть это система нейронов, которые взаимодействуют между собой. Каждый нейрон принимает сигналы или же отправляет их другим процессорам (нейронам). Объединённые в одну большую сеть, нейроны, обучаясь, могут выполнять сложные задачи.

Входной слой – это первый слой в нейронной сети, который принимает входящие сигналы и передает их на последующие уровни.

Скрытый слой применяет различные преобразования ко входным данным. Все нейроны в скрытом слое связаны с каждым нейроном в следующем слое.

Выходной слой – последний слой в сети, который получает данные от последнего скрытого слоя. С его помощью мы сможем получить нужное количество значений в желаемом диапазоне.

Вес представляет силу связи между нейронами. Вес определяет влияние ввода на вывод. В модели могут использоваться разные оптимизаторы.

Функция активации используется для того, чтобы ввести нелинейность в нейронную сеть. Она определяет выходное значение нейрона, которое будет зависеть от суммарного значения входов и порогового значения.

Также эта функция определяет, какие нейроны нужно активировать, и, следовательно, какая информация будет передана следующему слою. Благодаря функции активации глубокие сети могут обучаться.

2. Практическая часть

2.1. Разведочный анализ данных

Разведочный анализ данных (Exploratory Data Analysis) – предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками, а также сужения набора методов, используемых для создания модели машинного обучения.

При проведении разведочного анализа датасета была получена следующая информация:

- пропусков значений в датасете нет;

- дубликатов нет;

- тип данных float64;

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1023 entries, 0 to 1022
Data columns (total 13 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Соотношение матрица-наполнитель        1023 non-null   float64
 1   Плотность, кг/м3                       1023 non-null   float64
 2   модуль упругости, ГПа                  1023 non-null   float64
 3   Количество отвердителя, м.%            1023 non-null   float64
 4   Содержание эпоксидных групп,%_2        1023 non-null   float64
 5   Температура вспышки, С_2               1023 non-null   float64
 6   Поверхностная плотность, г/м2          1023 non-null   float64
 7   Модуль упругости при растяжении, ГПа   1023 non-null   float64
 8   Прочность при растяжении, МПа          1023 non-null   float64
 9   Потребление смолы, г/м2                1023 non-null   float64
 10  Угол нашивки, град                     1023 non-null   float64
 11  Шаг нашивки                            1023 non-null   float64
 12  Плотность нашивки                      1023 non-null   float64
dtypes: float64(13)
memory usage: 111.9 KB
```

Рисунок 6 – Краткий отчет по датасету

- количество уникальных значений согласно рисунку 7:

```
df.nunique ()
```

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

Рисунок 7 – Уникальные значения

- выполнен статистический анализ по датасету:

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.9304	0.9132	0.3894	2.3179	2.9069	3.5527	5.5917
Плотность, кг/м3	1023.0	1975.7349	73.7292	1731.7646	1924.1555	1977.6217	2021.3744	2207.7735
модуль упругости, ГПа	1023.0	739.9232	330.2316	2.4369	500.0475	739.6643	961.8125	1911.5365
Количество отвердителя, м.%	1023.0	110.5708	28.2959	17.7403	92.4435	110.5648	129.7304	198.9532
Содержание эпоксидных групп,%_2	1023.0	22.2444	2.4063	14.2550	20.6080	22.2307	23.9619	33.0000
Температура вспышки, С_2	1023.0	285.8822	40.9433	100.0000	259.0665	285.8968	313.0021	413.2734
Поверхностная плотность, г/м2	1023.0	482.7318	281.3147	0.6037	266.8166	451.8644	693.2250	1399.5424
Модуль упругости при растяжении, ГПа	1023.0	73.3286	3.1190	64.0541	71.2450	73.2688	75.3566	82.6821
Прочность при растяжении, МПа	1023.0	2466.9228	485.6280	1036.8566	2135.8504	2459.5245	2767.1931	3848.4367
Потребление смолы, г/м2	1023.0	218.4231	59.7359	33.8030	179.6275	219.1989	257.4817	414.5906
Угол нашивки, град	1023.0	44.2522	45.0158	0.0000	0.0000	0.0000	90.0000	90.0000
Шаг нашивки	1023.0	6.8992	2.5635	0.0000	5.0800	6.9161	8.5863	14.4405
Плотность нашивки	1023.0	57.1539	12.3510	0.0000	49.7992	57.3419	64.9450	103.9889

Рисунок 8 – Статистика по датасету до нормализации

- показано распределение значений и выполнен поиск аномалий, показанных на рисунке 2;

- корреляция, согласно рисунку 1, показывает, что коэффициенты корреляции близки к нулю, что показывает отсутствие линейной зависимости между признаками.

2.2. Препроцессинг

Предварительная обработка данных (препроцессинг) является важным шагом в процессе интеллектуального анализа данных. Фраза «мусор на входе — мусор на выходе» применима, в частности, и для проектов интеллектуального анализа данных и машинного обучения. Здесь имеется в виду то, что даже самый изощренный анализ не принесет пользы, если за основу взяты сомнительные данные.

Статистический анализ нашего датасета показал, что его данные распределены равномерно, есть выбросы по некоторым признакам, имеется один категориальный признак. Согласно данной информации была проведена кодировка признака «Угол нашивки, град» при помощи LabelEncoder поскольку

данный признак является категориальным и имеет только 2 уникальных значения 0 и 90, что может существенно повлиять на результаты моделей. Отброшены выбросы по признакам. Выполнена нормализация датасета при помощи MinMaxScaler.

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	936.0	0.4989	0.1875	0.0	0.3723	0.4945	0.6292	1.0
Плотность, кг/м3	936.0	0.5027	0.1878	0.0	0.3685	0.5112	0.6250	1.0
модуль упругости, ГПа	936.0	0.4468	0.1996	0.0	0.3012	0.4471	0.5804	1.0
Количество отвердителя, м.%	936.0	0.5047	0.1889	0.0	0.3762	0.5060	0.6380	1.0
Содержание эпоксидных групп,%_2	936.0	0.4912	0.1806	0.0	0.3677	0.4894	0.6234	1.0
Температура вспышки, С_2	936.0	0.5161	0.1906	0.0	0.3861	0.5160	0.6464	1.0
Поверхностная плотность, г/м2	936.0	0.3737	0.2171	0.0	0.2056	0.3542	0.5387	1.0
Модуль упругости при растяжении, ГПа	936.0	0.4886	0.1915	0.0	0.3590	0.4858	0.6151	1.0
Прочность при растяжении, МПа	936.0	0.4957	0.1889	0.0	0.3651	0.4918	0.6129	1.0
Потребление смолы, г/м2	936.0	0.5211	0.1958	0.0	0.3921	0.5238	0.6524	1.0
Угол нашивки, град_label	936.0	0.5118	0.5001	0.0	0.0000	1.0000	1.0000	1.0
Шаг нашивки	936.0	0.5022	0.1833	0.0	0.3722	0.5043	0.6246	1.0
Плотность нашивки	936.0	0.5138	0.1913	0.0	0.3905	0.5160	0.6388	1.0

Рисунок 9 – Статистический анализ датасета после нормализации

Таблица 4 – Сводная информация по датасету до и после нормализации:

Столбец	Среднее значение		Медианное значение		Минимальное значение		Максимальное значение	
1	2	3	4	5	6	7	8	9
	до	после	до	после	до	после	до	после
Соотношение матрица-наполнитель	2.9304	0.4989	2.9069	0.4945	0.3894	0.0	5.5917	1.0
Плотность, кг/м3	1975.7349	0.5027	1977.6217	0.5112	1731.7646	0.0	2207.7735	1.0

Продолжение таблицы 4

1	2	3	4	5	6	7	8	9
модуль упругости, ГПа	739.9232	0.4468	739.6643	0.447	2.4369	0.0	1911.5365	1.0
Количество отвердителя, м.%	110.5708	0.5047	110.5648	0.5060	17.7403	0.0	198.9532	1.0
Содержание эпоксидных групп,%_2	22.2444	0.4912	22.2307	0.4894	14.2550	0.0	33.0000	1.0
Температура вспышки, С_2	285.8822	0.5161	285.8968	0.5160	100.0000	0.0	413.2734	1.0
Поверхностная плотность, г/м2	482.7318	0.3737	451.8644	0.3542	0.6037	0.0	1399.5424	1.0
Модуль упругости при растяжении, ГПа	73.3286	0.4886	73.2688	0.4858	64.0541	0.0	82.6821	1.0
Прочность при растяжении, МПа	2466.9228	0.4957	2459.5245	0.4918	1036.8566	0.0	3848.4367	1.0
Потребление смолы, г/м2	218.4231	0.5211	219.1989	0.528	33.8030	0.0	414.5906	1.0
Угол нашивки, град	44.2522	0.5118	0.0000	1.0000	0.0000	0.0	90.0000	1.0
Шаг нашивки	6.8992	0.5022	6.9161	0.5043	0.0000	0.0	14.4405	1.0
Плотность нашивки	57.1539	0.5138	57.3419	0.5160	0.0000	0.0	103.9889	1.0

Видно, что и после нормализации признаков датасета плотность ядер осталась без изменений

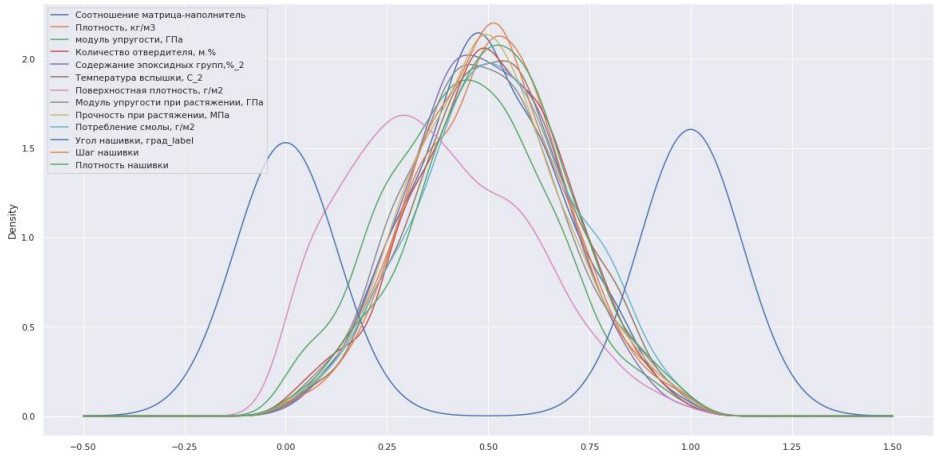


Рисунок 10 - Распределение переменных после нормализации датасета

2.3. Разбивка данных

По заданию известно, что необходимо построить модели регрессии и нейронную сеть для прогнозирования значений:

- «Прочности при растяжении, Мпа»;

- «Модуля упругости при растяжении, Гпа»;

- «Соотношения матрица-наполнитель».

2.3.1. Для прогнозирования значений «Модуль упругости при растяжении, Гпа»

Нормализованный датасет необходимо разделить на 2 подвыборки: обучающую и тестовую, 0,7 и 0,3 от датасета соответственно, и определить целевую переменную y.

```
y = df_norm_df['Модуль упругости при растяжении, ГПа']
x = df_norm_df.drop(('Модуль упругости при растяжении, ГПа'), axis = 1)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)

print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(655, 12)
(281, 12)
(655,)
(281,)
```

Рисунок 11 - Размерности тренировочной и тестовой подвыборок

```
df_norm_df['Модуль упругости при растяжении, ГПа'].describe().round(4).T
```

```
count    936.0000
mean       0.4886
std        0.1915
min        0.0000
25%        0.3590
50%        0.4858
75%        0.6151
max        1.0000
Name: Модуль упругости при растяжении, ГПа, dtype: float64
```

Рисунок 12 - Статистика по «Модуль упругости при растяжении, ГПа»

2.3.2. Для прогнозирования значений «Прочность при растяжении, Мпа»

Разбиваем датасет, как и на предыдущем этапе, только назначаем другую целевую переменную y.

```
y = df_norm_df['Прочность при растяжении, МПа']
x = df_norm_df.drop(('Прочность при растяжении, МПа'), axis = 1)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)

print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(655, 12)
(281, 12)
(655,)
(281,)
```

Рисунок 13 – Размерности тренировочной и тестовой подвыборок

```
df_norm_df['Прочность при растяжении, МПа'].describe().round(4).T
```

```
count    936.0000
mean       0.4957
std        0.1889
min        0.0000
25%        0.3651
50%        0.4918
75%        0.6129
max        1.0000
Name: Прочность при растяжении, МПа, dtype: float64
```

Рисунок 14 – Статистика по признаку «Прочность при растяжении, Мпа»

2.3.3. Для прогнозирования «Соотношения матрица-наполнитель»

Нормализованный датасет снова делим на 2 подвыборки: обучающую и тестовую, 0,7 и 0,3 от датасета соответственно, и определяем целевую переменную.

```
y = df_norm_df['Соотношение матрица-наполнитель']
x = df_norm_df.drop(('Соотношение матрица-наполнитель'), axis = 1)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)

print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(655, 12)
(281, 12)
(655,)
(281,)
```

Рисунок 15 - Размерности тренировочной и тестовой подвыборок

```
df_norm_df['Соотношение матрица-наполнитель'].describe().round(4).T
```

```
count    936.0000
mean       0.4989
std        0.1875
min        0.0000
25%        0.3723
50%        0.4945
75%        0.6292
max        1.0000
Name: Соотношение матрица-наполнитель, dtype: float64
```

Рисунок 16 - Статистика по признаку «Соотношение матрица-наполнитель»

2.4. Подготовка и обучение моделей для «Модуля упругости при растяжении, Мпа»

- линейная регрессия;

- ридж регрессия;

- регрессия по методу «лассо»

- регрессор случайного леса;

- градиентный бустинг.

Метрики обученных моделей, полученные на тестовой выборке, и кросс-валидация показаны в таблице 5.

2.5. Подготовка и обучение моделей для «Прочности при растяжении, Мпа»

- линейная регрессия;

- ридж регрессия;

- регрессия по методу «лассо»

- регрессор случайного леса;

- градиентный бустинг.

Метрики обученных моделей, полученные на тестовой выборке, и кросс-валидация показаны в таблице 6.

2.6. Подготовка и обучение моделей нейронной сети для признака «Соотношение матрица-наполнитель»

Построены две нейронные сети с помощью keras.Sequential.

Параметры первой нейросети:

- входной слой 12 признаков;

- выходной слой 1 признак;

- 3 скрытых слоя;

- первый скрытый слой содержит 8 нейронов;

- второй скрытый слой содержит 8 нейронов;

- третий скрытый слой содержит 8 нейронов;

- функция активации –relu;

- на последнем скрытом слое функция активации – linear;

- оптимизатор – adam;

- функция потери - mean_absolute_error;

- количество эпох – 20.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 8)                 104

 dense_1 (Dense)             (None, 8)                 72

 dense_2 (Dense)             (None, 8)                 72

 dense_3 (Dense)             (None, 1)                 9

=================================================================
Total params: 257
Trainable params: 257
Non-trainable params: 0
_________________________________________________________________
None
```

Рисунок 17 - Архитектура первой нейронной сети

	loss	mae	val_loss	val_mae	epoch
15	0.156793	0.156793	0.139715	0.139715	15
16	0.156753	0.156753	0.139119	0.139119	16
17	0.156166	0.156166	0.138575	0.138575	17
18	0.155659	0.155659	0.137963	0.137963	18
19	0.154861	0.154861	0.137469	0.137469	19

Рисунок 18 – Результаты работы нейронной сети

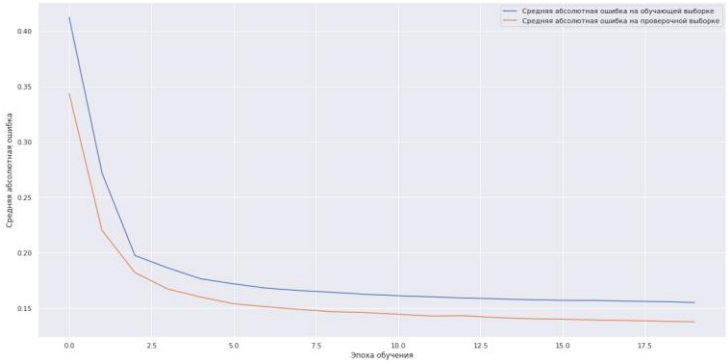


Рисунок 19 – Качество обучения нейронной сети

```
21/21 [==============================] - 0s 1ms/step
Коэффициент детерминации на обучающей выборке:  -0.029
9/9 [==============================] - 0s 1ms/step
Коэффициент детерминации на тестовой выборке (R2):  -0.09
MSE для нейронной сети: 0.037
MAE для нейронной сети: 0.155
MAPE для нейронной сети: 0.466
Max error для нейронной сети: 0.469
```

Рисунок 20 – Метрики

Вторая нейронная сеть

Параметры второй нейросети:

- входной слой 12 признаков;

- выходной слой 1 признак;

- 3 скрытых слоя;

- первый скрытый слой содержит 32 нейрона;

- второй скрытый слой содержит 16 нейронов;

- третий скрытый слой содержит 8 нейронов;

- на выходном слое 1 нейрон;

- функция активации –tanh;

- на последнем скрытом слое функция активации – linear;

- оптимизатор – adam;

- функция потери - mean_absolute_error;

- количество эпох – 20.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_4 (Dense)             (None, 32)                416

 dense_5 (Dense)             (None, 16)                528

 dense_6 (Dense)             (None, 8)                 136

 dense_7 (Dense)             (None, 1)                 9

=================================================================
Total params: 1,089
Trainable params: 1,089
Non-trainable params: 0
_________________________________________________________________
None
```

Рисунок 21 – Архитектура второй нейронной сети

	loss	mae	val_loss	val_mae	epoch
15	0.037117	0.156406	0.030753	0.134106	15
16	0.036764	0.154751	0.030406	0.134755	16
17	0.036540	0.154963	0.030695	0.136426	17
18	0.036384	0.154519	0.030528	0.133568	18
19	0.036397	0.154197	0.030371	0.134333	19

Рисунок 22 – Результаты работы нейронной сети

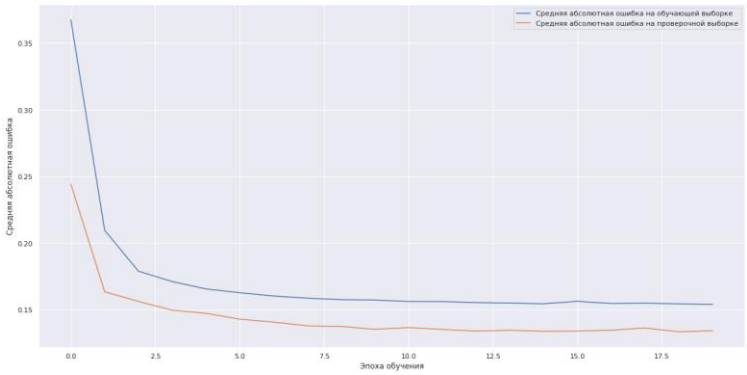


Рисунок 23 – Качество обучения нейронной сети

```
21/21 [==============================] - 0s 1ms/step
Коэффициент детерминации на обучающей выборке:  0.008
9/9 [==============================] - 0s 1ms/step
Коэффициент детерминации на тестовой выборке (R2):  -0.09
MSE для нейронной сети: 0.037
MAE для нейронной сети: 0.155
MAPE для нейронной сети: 0.466
Max error для нейронной сети: 0.469
```

Рисунок 24 – Метрики нейронной сети

2.7. Тестирование моделей

Тестирование моделей показало следующие ошибки, приведенные в таблицах 5, 6, 7.

Таблица 5 – Сводная информация по метрикам моделей для прогнозирования «Модуль упругости при растяжении, Гпа»

Модель	Коэф. детерм. на обуч. выборке	Коэф. детерм. на тест. выборке	MSE	MAE	MAPE	Max Error	Cross val score (mean)	Стандартное отклонение
Линейная регрессия	0.015	-0.005	0.035	0.152	0.453	0.483	-0.061	0.057
Ридж регрессия	0.015	-0.005	0.035	0.152	0.453	0.484	-0.059	0.056
«Лассо» регрессия	0.0	-0.006	0.035	0.151	0.450	0.500	-0.031	0.040
Регрессор случайного леса	0.848	-0.019	0.035	0.152	0.452	0.504	-0.114	0.078
Градиентный бустинг	0.456	-0.099	0.038	0.157	0.474	0.569	-0.133	0.125

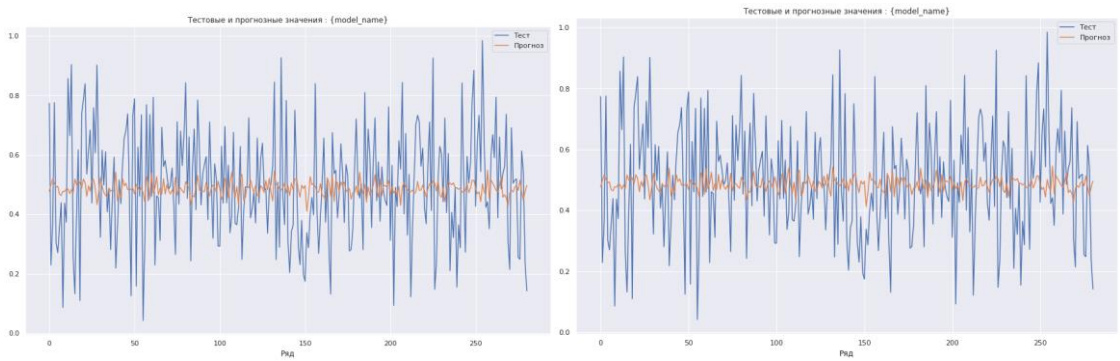
Таблица 6 – Сводная информация по метрикам моделей для прогнозирования «Прочности при растяжении, Мпа»

Модель	Коэф. детерм. на обуч. выборке	Коэф. детерм. на тест. выборке	MSE	MAE	MAPE	Max Error	Cross val score (mean)	Стандартное отклонение
Линейная регрессия	0.025	-0.047	0.035	0.148	0.473	0.507	-0.065	0.083
Ридж регрессия	0.025	-0.045	0.035	0.148	0.473	0.507	-0.063	0.079
«Лассо» регрессия	0.0	-0.03	0.035	0.147	0.473	0.507	-0.031	0.042
Регрессор случайного леса	0.854	-0.097	0.037	0.154	0.484	0.539	-0.087	0.085
Градиентный бустинг	0.539	-0.123	0.038	0.156	0.475	0.536	-0.154	0.126

Таблица 7 – Сводная информация по метрикам модели нейронной сети для прогнозирования «Соотношения матрица-наполнитель»

Модель	Коэф. детерм. на обуч. выборке	Коэф. детерм. на тест. выборке	MSE	MAE	MAPE	Max Error
Нейронная сеть 1	-0.029	0.009	0.037	0.155	0.466	0.469
Нейронная сеть 2	0.008	-0.09	0.037	0.155	0.466	0.469

Метрики качества моделей, показанные выше, дают нам представление об их работе. Вообще, основным показателем эффективности работы моделей регрессии является метрика R2 или коэффициент детерминации на тестовой выборке, т.е. то, как целевая переменная прогнозируется моделью. Его значения, как правило, варьируются от 0 до 1, чем выше, тем лучше, т.е. чем выше значение, тем лучше модель предсказывает значения. Одновременно с этим очень высокий коэффициент показывает переобученность модели. В нашем случае коэффициент отрицательный, т.е. модель вообще не понимает признаков. Ошибка не в данных, а в обработке датасета. Перекрестная проверка и поиск гиперапарметров по сетке наши модели практически не улучшили, что еще раз говорит о вероятно неверном подходе к препроцессингу данных.



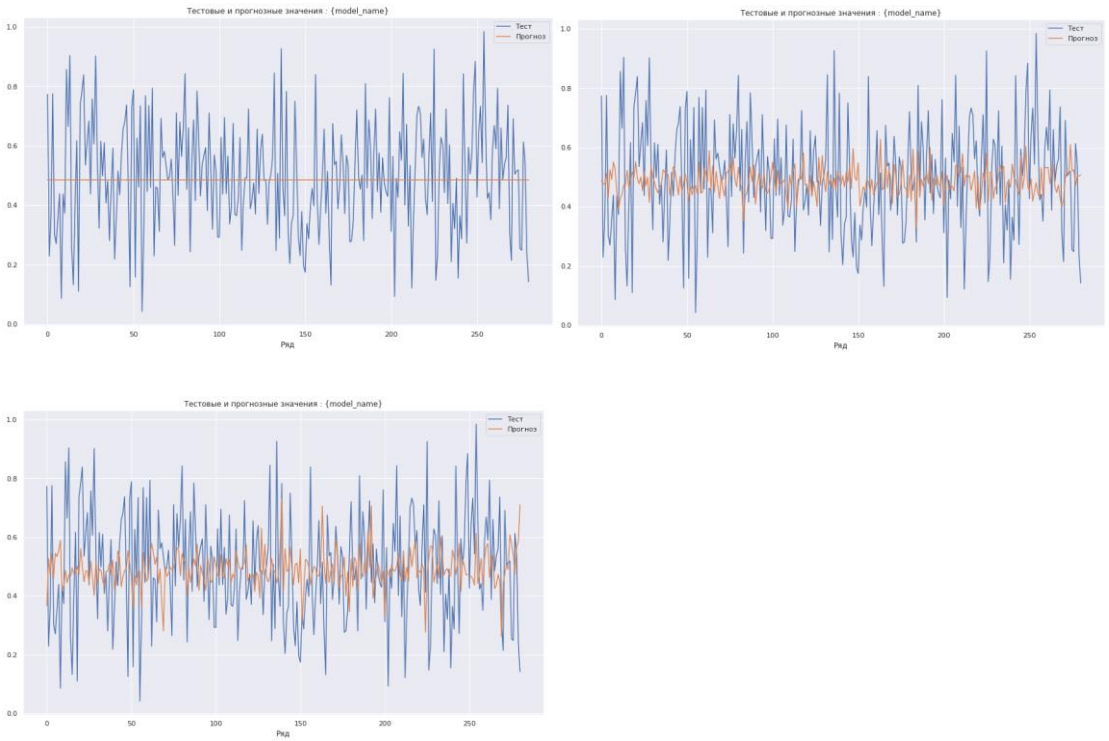
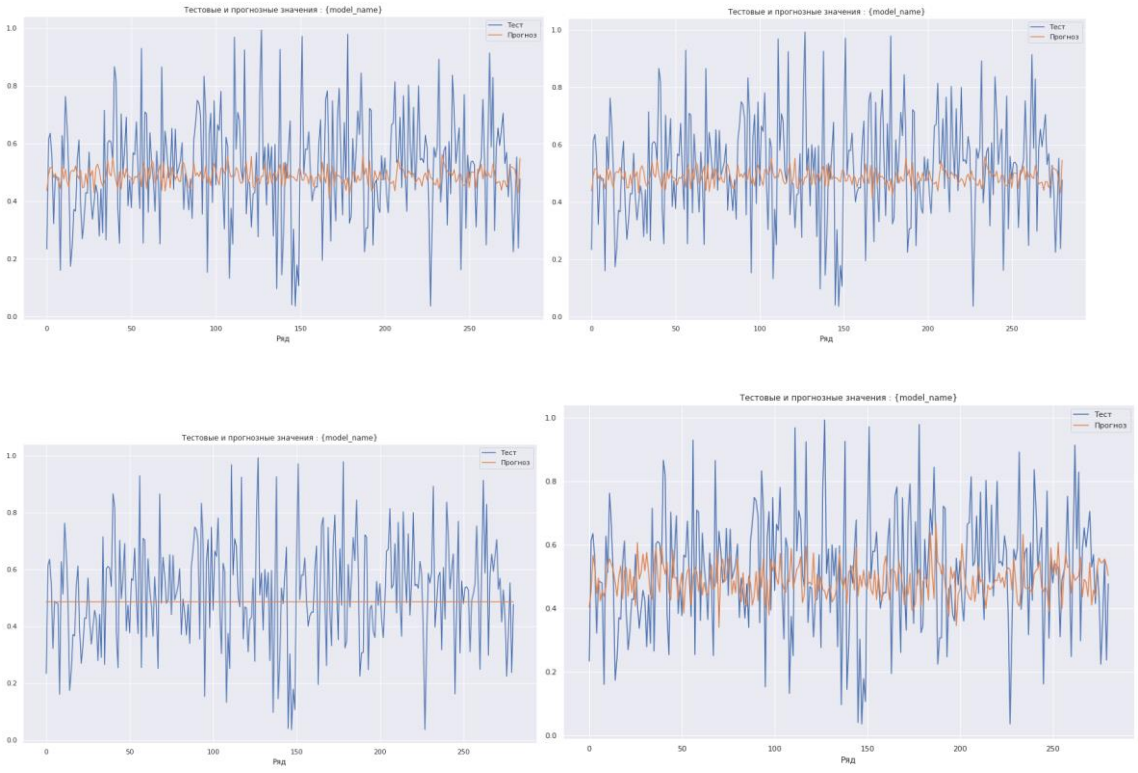


Рисунок 25 - Визуализация работы моделей для «Модуля упругости при растяжении, Гпа»



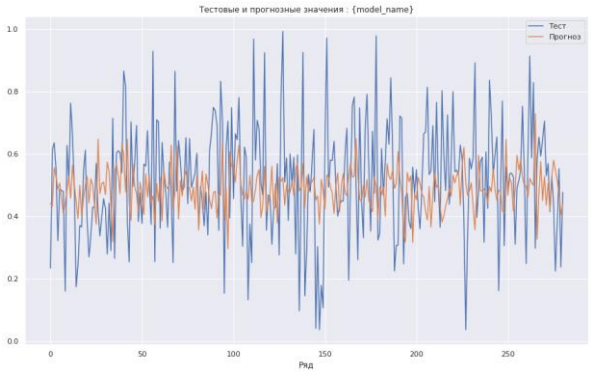


Рисунок 26 - Визуализация работы моделей для «Прочности при растяжении, Мпа»

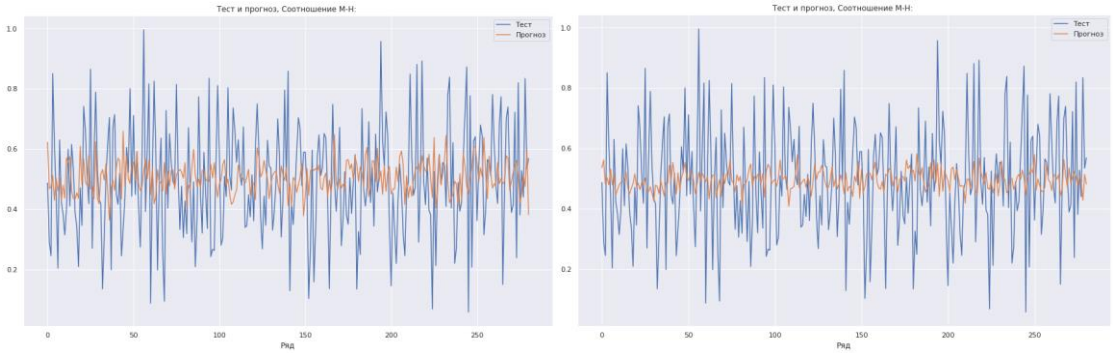


Рисунок 27 - Визуализация работы моделей для «Соотношения матрица-наполнитель»

2.8 Разработка приложения

Разработка приложения проводилась на фреймворке Flask на языке Python.

Работа приложения подразумевает следующее:

- выбор целевой переменной;

- ввод параметров;

- получение прогноза по целевой переменной.

Скриншоты о работе приложения в Приложении А.

2.9. Создание репозитория

Репозиторий создан на Github, куда были загружены файлы проекта. Ссылка на репозиторий: https://github.com/LarisaK2023/Prediction-of-composites-final-features

Заключение

При разработке данного проекта был пройден практически весь Pipeline, начиная с поиска закономерностей, важных признаков в данных и первичного анализа, на основе которых разбит датасет на тестовую и обучающую выборки, выполнены задачи регрессии с использованием алгоритмов машинного обучения. В нашем случае алгоритмы и их результаты описаны в разделе 1.4. данного проекта. Выполнена визуализация обучения, получены необходимые метрики, ошибки. Выполнена валидация моделей на данных, которые она не видела, т.е. я убедилась в их точности. В нашем случае, к сожалению, модели не показали хороших результатов. Сохранены модели обучения. Создано приложение.

Возможно, мною были допущены ошибки при анализе данных, что является результатом недостатка опыта и знаний. При дальнейшей практике с использованием других способов препроцессинга, алгоритмов и т.д., вероятно, можно получить более качественные результаты.

Список литературы

1. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: - Режим доступа: https://habr.com/ru/company/vk/blog/513842/
2. Документация по Flask: – Режим доступа: https://flask.palletsprojects.com/en/2.2.x/
3. Документация по Keras: – Режим доступа: https://keras.io/getting_started/
4. Документация по numpy: – Режим доступа: https://numpy.org/doc/stable/
5. Документация по pandas: – Режим доступа: https://pandas.pydata.org/docs/
6. Документация по Python: – Режим доступа: https://docs.python.org/3.10/
7. Документация по Matplotlib: – Режим доступа: https://matplotlib.org/
8. Документация по Seaborn: – Режим доступа: https://seaborn.pydata.org/tutorial/introduction
9. Документация по sklearn: – Режим доступа: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
10. Документация по Tensorflow: – Режим доступа: https://www.tensorflow.org/?hl=ru
11. Композитные материалы. Режим доступа: https://qw-russia.ru/kompozitnye-materialy.
12. Регрессия: – Режим доступа: https://www.helenkapatsa.ru/rieghriessiia/
13. Nuances of programming. 5 видов регрессии и их свойства: - Режим доступа: https://dzen.ru/media/nuancesprog.ru/5-vidov-regressii-i-ih-svoistva-5e31b6235561a65d78dee6c5
14. Rostec. Композитная история. Режим доступа: https://rostec.ru/news/kompozitnaya-istoriya/.
15. Tara Boyle, Linear Regression Models: - Режим доступа: https://towardsdatascience.com/linear-regression-models-4a3d14b8d368
16. Ye, Andre Modern Deep Learning Design and Application Development. Versatile Tools to Solve Deep Learning Problems. eBook 2022.

Приложение А

Скриншоты работы приложения.