ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ МГТУ им. Н.Э. Баумана

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

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 Pro»

«Прогнозирование конечных свойств новых материалов (композиционных материалов).»

Слушатель Петров Александр Евгеньевич (ФИО)

Москва, 2024

Содержание

Введение

Композитные материалы – материалы, созданные из двух или более составляющих материалов со значительно различными физическими и химическими свойствами, которые в комбинации придают итоговому материалу характеристики отличные от характеристик отдельных компонентов и не являющиеся простой их суперпозицией.

Ключевые компоненты композитных материалов:

Матрицу материала – это основной материал, окружающий и связывающий вместе армирующие материалы. Матрица обеспечивает защиту и поддержку армирующего материала. Часто, в качестве матрицы используются полимеры.

Армирующий материал – материалы, внедряющиеся в матрицу материалов для улучшения определенных свойств, таких как, например, прочность, жесткость или ударная вязкость. В качестве наполнителей композитов как правило выступают углеродные или стеклянные волокна.

Сочетание разных компонентов позволяет улучшить характеристики материала и делает его одновременно лёгким и прочным. При этом отдельные компоненты остаются таковыми в структуре композитов, что отличает их от смесей и затвердевших растворов. Варьируя состав матрицы и наполнителя, их соотношение, ориентацию наполнителя, получают широкий спектр материалов с требуемым набором свойств. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик.

1. Аналитическая часть

1.1 Постановка задачи

В данной работе исследуется композит с матрицей из базальтопластика и нашивками из углепластика. От специалистов в предметной области был получен датасет, содержащий данные о свойствах матрицы и наполнителя, производственных параметрах и свойствах готового композита. От нас, как специалистов в машинном обучении, требуется разработать модели, прогнозирующие значения некоторых свойств в зависимости от остальных. Так же требуется разработать приложение, делающее удобным использование данных моделей специалистом предметной области.

Датасет состоит из двух файлов: X_bp (составляющая из базальтопластика) и X_nup (составляющая из углепластика).

Файл X_bp содержит:

- признаков: 10 и индекс;
- строк: 1023.

Файл X_nup содержит:

- признаков: 3 и индекс;
- строк: 1040.

Известно, что файлы требуют объединения с типом INNER по индексу. После объединения часть строк из файла X_nup была отброшена. И дальнейшие исследования проводим с объединенным датасетом, содержащим 13 признаков и 1023 строк или объектов.

Описание признаков объединенного датасета приведено в таблице 1. Все признаки имеют тип float64, то есть вещественный. Пропусков в данных нет. Все признаки, кроме «Угол нашивки», являются непрерывными, количественными. «Угол нашивки» принимает только два значения и будет рассматриваться как категориальный признак.

Таблица 1 — Описание признаков датасета

Название	Файл	Тип данных	Непустых значений	Уникальных значений
Соотношение матрица-наполнител	X_bp	float64	1023	1014
Плотность, кг/м3	X_bp	float64	1023	1013
модуль упругости, ГПа	X_bp	float64	1023	1020
Количество отвердителя, м.%	X_bp	float64	1023	1005
Содержание эпоксидных групп,%_2	X_bp	float64	1023	1004
Температура вспышки, С_2	X_bp	float64	1023	1003
Поверхностная плотность, г/м2	X_bp	float64	1023	1004
Модуль упругости при растяжении, ГПа	X_bp	float64	1023	1004
Прочность при растяжении, МПа	X_bp	float64	1023	1004
Потребление смолы, г/м2	X_bp	float64	1023	1003
Угол нашивки, град	X_nup	float64	1023	2
Шаг нашивки	X_nup	float64	1023	989
Плотность нашивки	X_nup	float64	1023	988

1.1.1. Анализ распределения данных

Гистограммы распределения переменных и диаграммы «ящик с усами» приведены на рисунках 1-3. По ним видно, что все признаки, кроме «Угол нашивки», имеют нормальное распределение и принимают неотрицательные значения. «Угол нашивки» принимает значения: 0, 90.

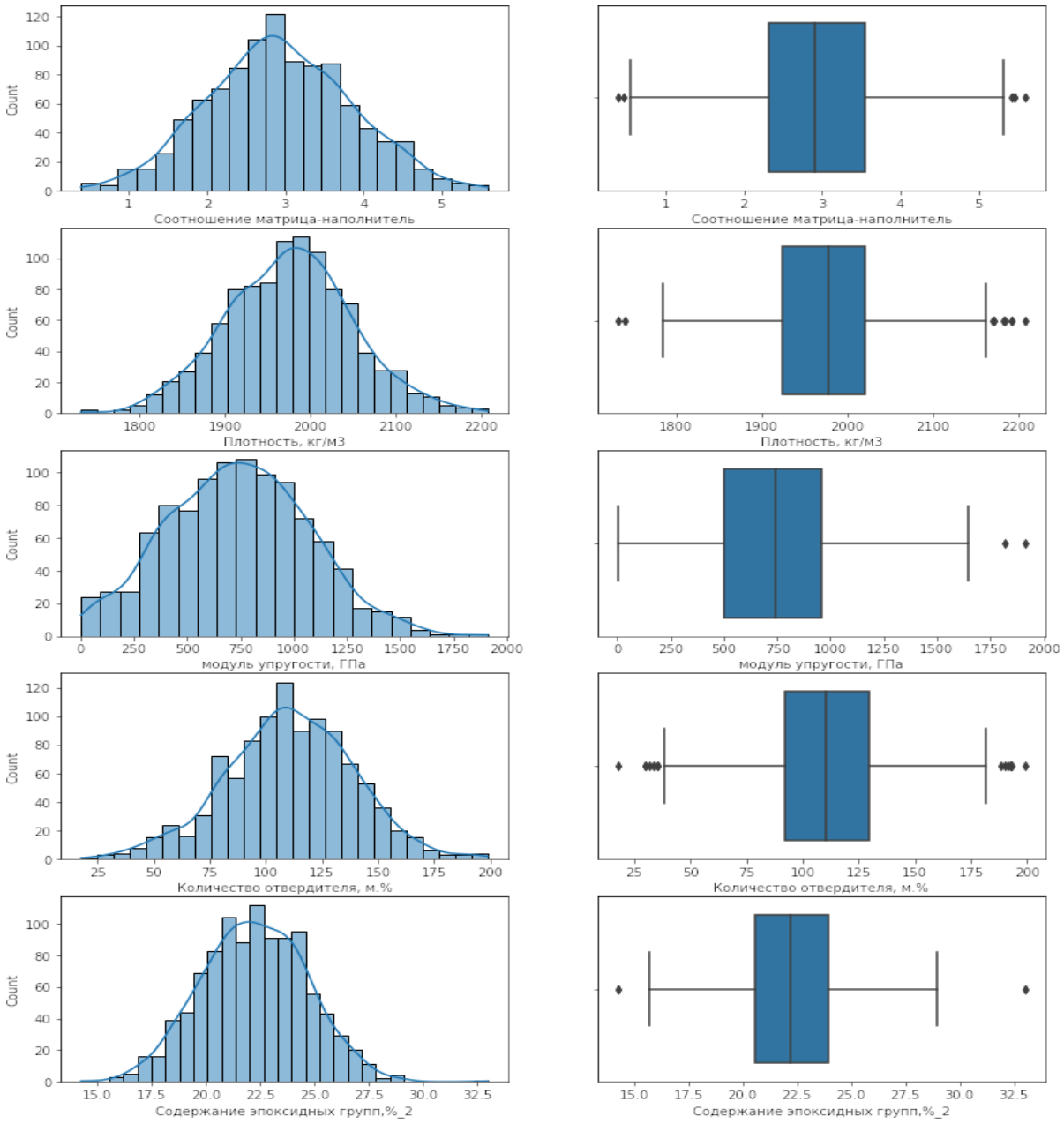

Рисунок 1 – Гистограммы распределения переменных и диаграммы «ящик с усами»

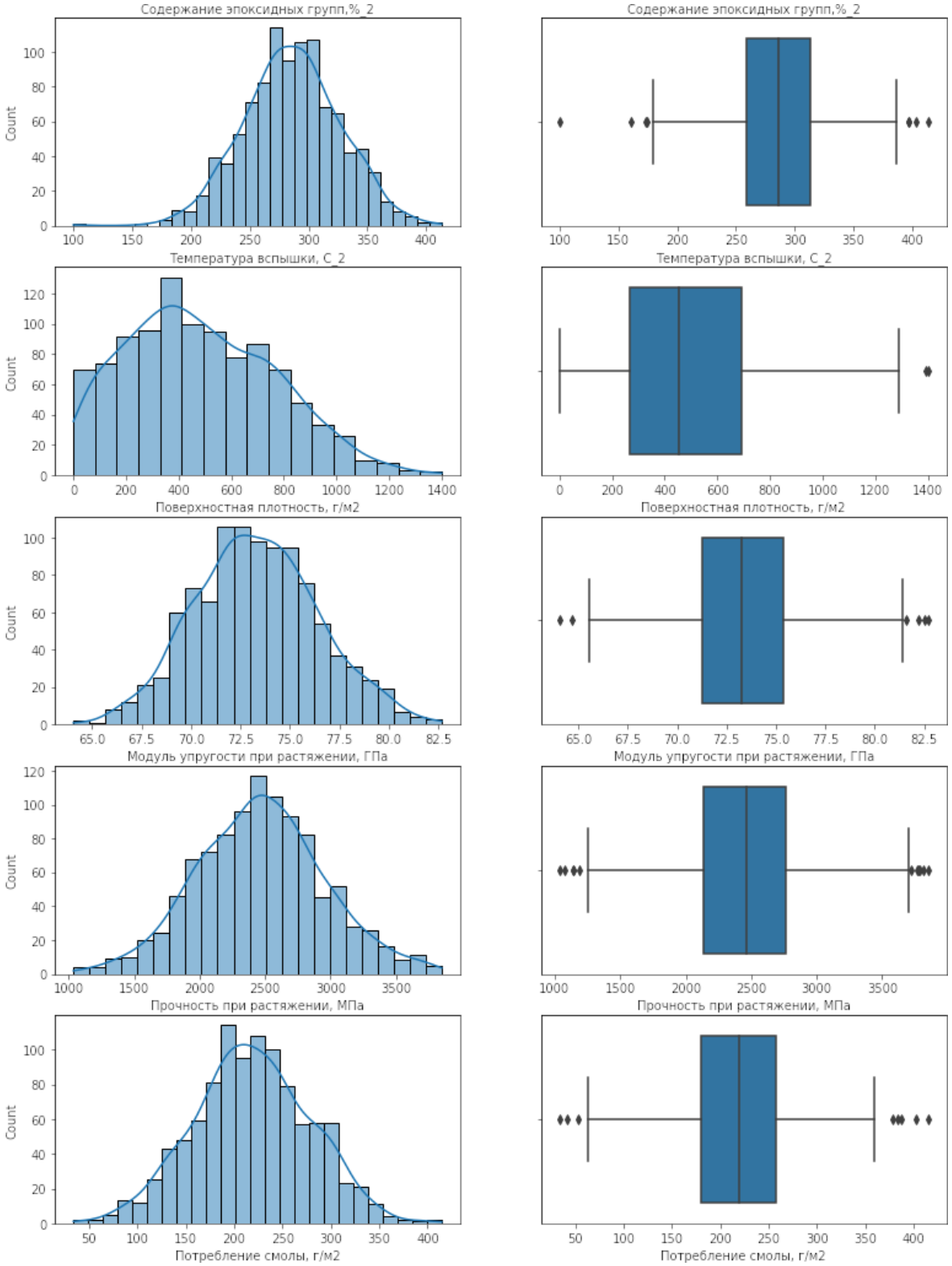

Рисунок 2 – Гистограммы распределения переменных и диаграммы «ящик с усами»

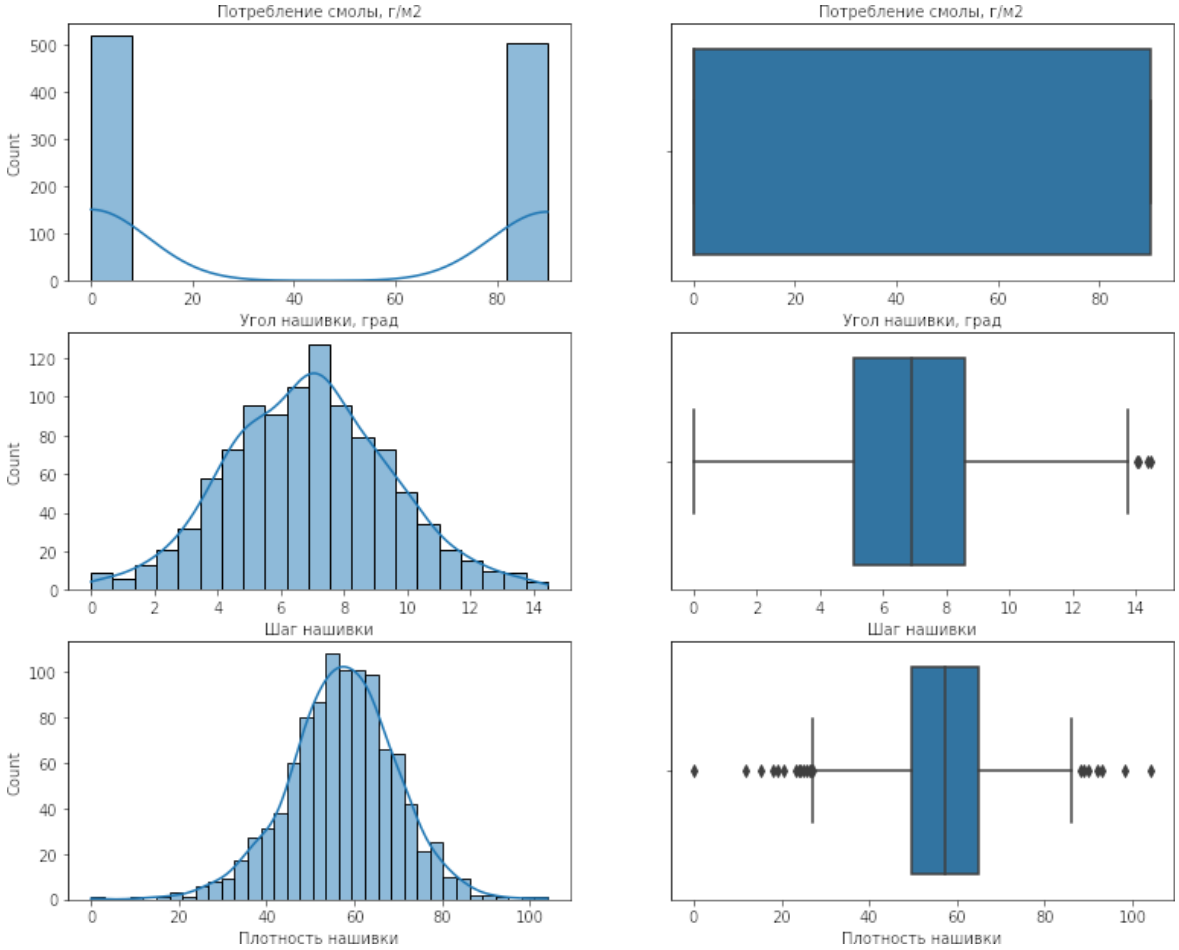

Рисунок 3 – Гистограммы распределения переменных и диаграммы «ящик с усами»

Нам известно, что датасет был предварительно подготовлен, поэтому отсутствие пропусков не удивило. В сырых данных пропуски и значения некорректных типов как правило присутствуют.

Так же нас интересует описательная статистика датасета. Она представлена в таблице 2. Она в численном виде отражает то, что мы видим на гистограммах.

Попарные графики рассеяния точек приведены на рисунке 4.

По графикам рассеяния мы видим, что некоторые точки отстоят далеко от общего облака. Так визуально выглядят выбросы — аномальные, некорректные значения данных, выходящие за пределы допустимых значений признака.

Таблица 2 — Описательная статистика признаков датасета

	Среднее	Стандартное отклонение	Минимум	Максимум	Медиана
Соотношение матрица-наполнитель	2.9304	0.9132	0.3894	5.5917	2.9069
Плотность, кг/м3	1975.7349	73.7292	1731.7646	2207.7735	1977.6217
модуль упругости, ГПа	739.9232	330.2316	2.4369	1911.5365	739.6643
Количество отвердителя, м.%	110.5708	28.2959	17.7403	198.9532	110.5648
Содержание эпоксидных групп, %_2	22.2444	2.4063	14.2550	33.0000	22.2307
Температура вспышки, С_2	285.8822	40.9433	100.0000	413.2734	285.8968
Поверхностная плотность, г/м2	482.7318	281.3147	0.6037	1399.5424	451.8644
Модуль упругости при растяжении, ГПа	73.3286	3.1190	64.0541	82.6821	73.2688
Прочность при растяжении, МПа	2466.9228	485.6280	1036.8566	3848.4367	2459.5245
Потребление смолы, г/м2	218.4231	59.7359	33.8030	414.5906	219.1989
Угол нашивки, град	44.2522	45.0158	0.0000	90.0000	0.0000
Шаг нашивки	6.8992	2.5635	0.0000	14.4405	6.9161
Плотность нашивки	57.1539	12.3510	0.0000	103.9889	57.3419

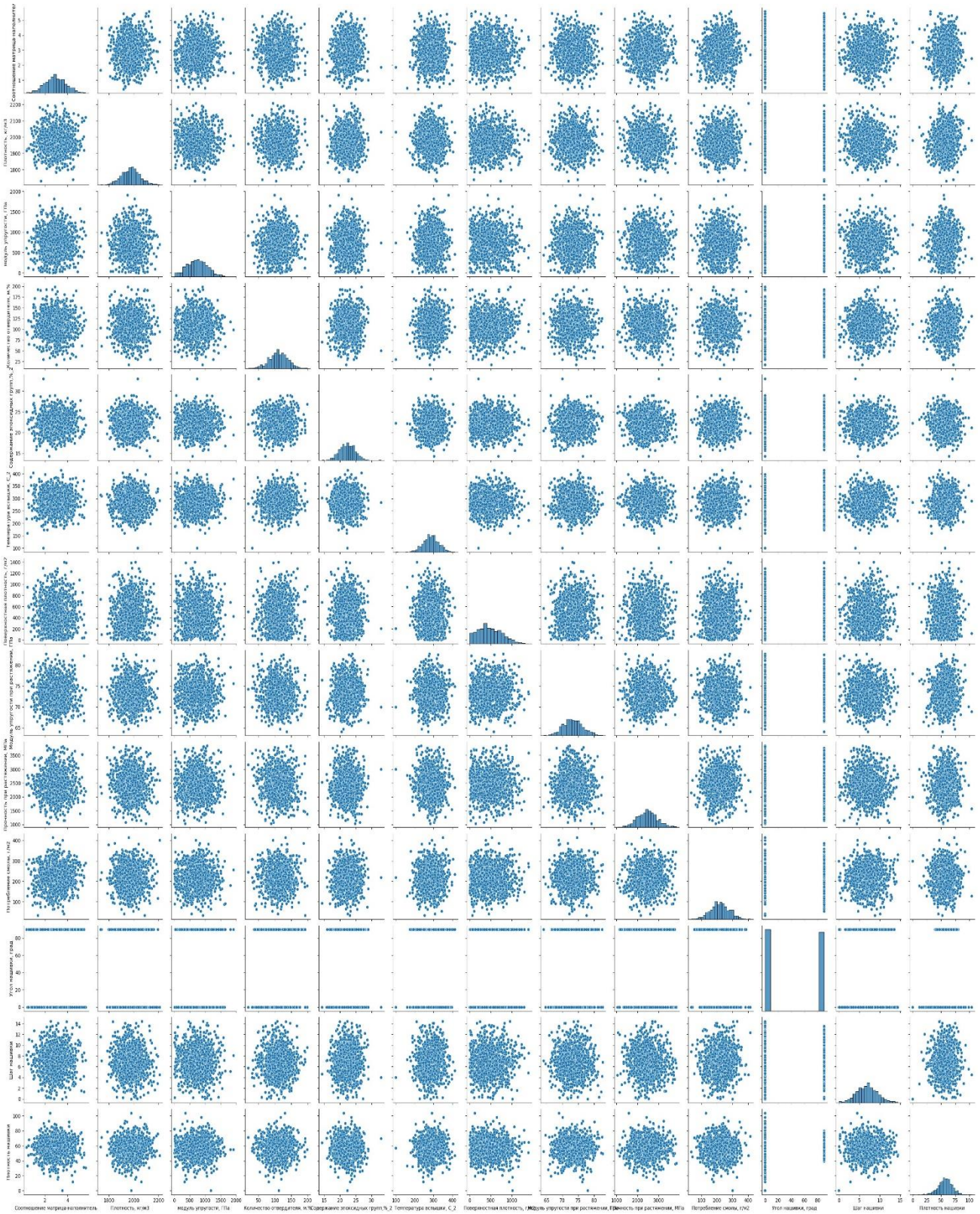

Рисунок 4 – Зависимость распределения параметров относительно друг друга

1.1.2 Нахождение выбросов

Один из распространенных способов найти выбросы в наборе данных — использовать межквартильный диапазон.

Межквартильный диапазон, часто сокращенно IQR, представляет собой разницу между 25-м процентилем (Q1) и 75-м процентилем (Q3) в наборе данных. Он измеряет разброс средних 50% значений.

Один из популярных методов 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.

Доля выбросов, найденная этим способом указана на рисунке 7

	доля выбросов
Соотношение матрица-наполнитель	0.59%
Плотность, кг/м3	0.88%
модуль упругости, ГПа	0.20%
Количество отвердителя, м.%	1.37%
Содержание эпоксидных групп,%_2	0.20%
Температура вспышки, С_2	0.78%
Поверхностная плотность, г/м2	0.20%
Модуль упругости при растяжении, ГПа	0.59%
Прочность при растяжении, МПа	1.08%
Потребление смолы, г/м2	0.78%
Угол нашивки, град	0.00%
Шаг нашивки	0.39%
Плотность нашивки	2.05%

Рисунок 7 – Доля выбросов

Значения, определенные как выбросы, удаляем. После этого осталось в датасете осталось 936 строк и 13 признаков-переменных.

1.2 Описание используемых методов

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

В настоящее время разработано много методов регрессионного анализа. Например, простая и множественная линейная регрессия. Эти модели являются параметрическими в том смысле, что функция регрессии определяется конечным числом неизвестных параметров, которые оцениваются на основе данных.

1.2.1. Линейная регрессия

Простая линейная регрессия имеет место, если рассматривается зависимость между одной входной и одной выходной переменными. Для этого определяется уравнение регрессии (1) и строится соответствующая прямая, известная как линия регрессии.

$$y = ax + b \quad , \quad (1)$$

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

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

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \quad , \quad (2)$$

где n - число входных переменных.

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

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

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

1.2.2Гребневая (Ridge) регрессия

Гребневая регрессия или ридж-регрессия — так же вариация линейной регрессии. Она так же применяет сжатие и хорошо работает для данных, которые демонстрируют сильную мультиколлинеарность.

Самое большое различие между ними в том, что гребневая регрессия использует регуляризацию L2, которая взвешивает ошибки по их квадрату, чтобы сильнее наказывать за более значительные ошибки.

Когда мы делаем линейную регрессию, то функция потерь выглядела так(4):

$$L(X, \vec{y}, \vec{w}) = \frac{1}{2n}\sum_{i-1}^{n}(\vec{x}_i^T - y_i)^2 \ , \ (3)$$

Теперь формула выглядит так:

$$L(X, \vec{y}, \vec{w}) = \frac{1}{2n}\sum_{i-1}^{n}(\vec{x}_i^T - y_i)^2 - \lambda \sum_{j-1}^{m} \vec{w}_j^2 \ \ , \ \ (4)$$

Регуляризация позволяет интерпретировать модели. Если коэффициент стал близким к 0 (для Ridge), значит данный входной признак не является значимым.

Эти методы реализованы в sklearn.linear_model.Lasso и sklearn.linear_model.Ridge.

1.2.3. Случайный лес

Случайный лес (RandomForest) — представитель ансамблевых методов.

Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив. Формула итогового решателя (5) — это усреднение предсказаний отдельных деревьев.

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i\,(x)\ ,\ (5)$$

где

N – количество деревьев;

i – счетчик для деревьев; b – решающее дерево;

x – сгенерированная нами на основе данных выборка.

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

Преимущества случайного леса:

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

Из недостатков можно отметить, что его построение занимает больше

времени. Так же теряется интерпретируемость.

Метод реализован в sklearn.ensemble.RandomForestRegressor.

1.2.4. Метод K-ближайших соседей (KNN)

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

Метод k-ближайших соседей (k Nearest Neighbors, или kNN) – популярный алгоритм классификации, который используется в разных типах задач машинного обучения. Наравне с деревом решений это один из самых понятных подходов к классификации. Метод kNN - это простой алгоритм машинного обучения с учителем, который можно использовать для решения задач классификации и регрессии. Он прост в реализации и понимании, но имеет существенный недостаток – значительное замедление работы, когда объем данных растет.

На интуитивном уровне суть метода проста: посмотри на соседей вокруг, какие из них преобладают, таковым ты и являешься. Формально основой метода является гипотеза компактности: если метрика расстояния между примерами введена удачно, то схожие примеры гораздо чаще лежат в одном классе, чем в разных.

1.2.5. Нейронная сеть

Нейронная сеть — это последовательность нейронов, соединенных между собой связями. Структура нейронной сети пришла в мир программирования из биологии. Вычислительная единица нейронной сети — нейрон или персептрон.

У каждого нейрона есть определённое количество входов, куда поступают сигналы, которые суммируются с учётом значимости (веса) каждого входа.

Смещение – это дополнительный вход для нейрона, который всегда равен 1 и, следовательно, имеет собственный вес соединения.

Так же у нейрона есть функция активации, которая определяет выходное значение нейрона. Она используется для того, чтобы ввести нелинейность в нейронную сеть. Примеры активационных функций: relu, сигмоида.

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

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

Прямое распространение – это процесс передачи входных значений в нейронную сеть и получения выходных данных, которые называются прогнозиру- емым значением.

Прогнозируемое значение сравниваем с фактическим с помощью функции потери. В методе обратного распространения ошибки градиенты (производные значений ошибок) вычисляются по значениям весов в направлении, обратном прямому распространению сигналов. Значение градиента вычитают из значения веса, чтобы уменьшить значение ошибки. Таким образом происходит процесс обучения. Обновляются веса каждого соединения, чтобы функция потерь минимизировалась.

Для обновления весов в модели используются различные оптимизаторы.

Количество эпох показывает, сколько раз выполнялся проход для всех примеров обучения.

Нейронные сети применяются для решения задач регрессии, классификации, распознавания образов и речи, компьютерного зрения и других. На настоящий момент это самый мощный, гибкий и широко применяемый инструмент в машинном обучении.

1.3. Разведочный анализ данных

1.3.1. Анализ корреляционных составляющих

При вычислении корреляции между параметрами явной зависимости между параметрами не было выявлено (см. рисунок 8). Выявлены единичные случаи корреляции около 0.1

	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
Соотношение матрица-наполнитель	1.000000	0.003841	0.031700	-0.006445	0.019766	-0.004776	-0.006272	-0.008411	0.024148	0.072531	-0.031073	0.036437	-0.004652
Плотность, кг/м3	0.003841	1.000000	-0.009647	-0.035911	-0.008278	-0.020695	0.044930	-0.017602	-0.069981	-0.015937	-0.068474	-0.061015	0.080304
модуль упругости, ГПа	0.031700	-0.009647	1.000000	0.024049	-0.006804	0.031174	-0.005306	0.023267	0.041868	0.001840	-0.025417	-0.009875	0.056346
Количество отвердителя, м.%	-0.006445	-0.035911	0.024049	1.000000	-0.000684	0.095193	0.055198	-0.065929	-0.075375	0.007446	0.038570	0.014887	0.017248
Содержание эпоксидных групп,%_2	0.019766	-0.008278	-0.006804	-0.000684	1.000000	-0.009769	-0.012940	0.056828	-0.023899	0.015165	0.008052	0.003022	-0.039073
Температура вспышки, С_2	-0.004776	-0.020695	0.031174	0.095193	-0.009769	1.000000	0.020121	0.028414	-0.031763	0.059954	0.020695	0.025795	0.011391
Поверхностная плотность, г/м2	-0.006272	0.044930	-0.005306	0.055198	-0.012940	0.020121	1.000000	0.036702	-0.003210	0.015692	0.052299	0.038332	-0.049923
Модуль упругости при растяжении, ГПа	-0.008411	-0.017602	0.023267	-0.065929	0.056828	0.028414	0.036702	1.000000	-0.009009	0.050938	0.023003	-0.029468	0.006476
Прочность при растяжении, МПа	0.024148	-0.069981	0.041868	-0.075375	-0.023899	-0.031763	-0.003210	-0.009009	1.000000	0.028602	0.023398	-0.059547	0.019604
Потребление смолы, г/м2	0.072531	-0.015937	0.001840	0.007446	0.015165	0.059954	0.015692	0.050938	0.028602	1.000000	-0.015334	0.013394	0.012239
Угол нашивки, град	-0.031073	-0.068474	-0.025417	0.038570	0.008052	0.020695	0.052299	0.023003	0.023398	-0.015334	1.000000	0.023616	0.107947
Шаг нашивки	0.036437	-0.061015	-0.009875	0.014887	0.003022	0.025795	0.038332	-0.029468	-0.059547	0.013394	0.023616	1.000000	0.003487
Плотность нашивки	-0.004652	0.080304	0.056346	0.017248	-0.039073	0.011391	-0.049923	0.006476	0.019604	0.012239	0.107947	0.003487	1.000000

Рисунок 8 – Вычисление корреляции между параметрами

Для наглядности данная таблица представлена в виде тепловой карты на рисунке 9.

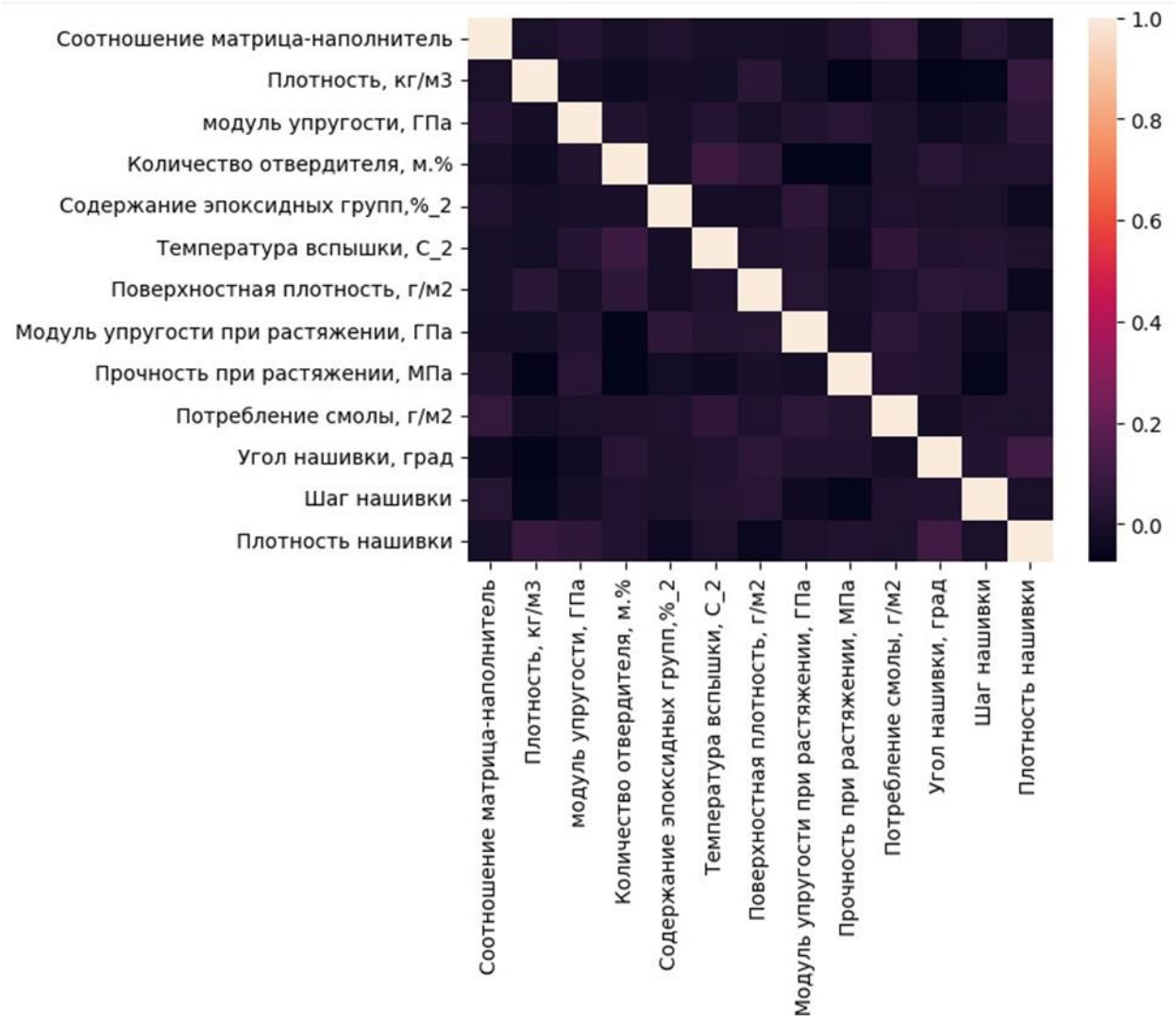

Рисунок 9 – Тепловая карта

1.3.2. Ход решения задачи

Ход решения каждой из задач и построения оптимальной модели будет следующим:

- разделить данные на тренировочную и тестовую выборки. В задании указано, что на тестирование оставить 30% данных;
- выполнить препроцессинг, то есть подготовку исходных данных;
- выбрать базовую модель для определения нижней границы качества

предсказания. Использую базовую модель, возвращающую среднее значение целевого признака. Лучшая модель по своим характеристикам должна быть лучше базовой;

- взять несколько моделей с гиперпараметрами по умолчанию, и используя перекрестную проверку, посмотреть их метрики на тренировочной выборке;

- подобрать для этих моделей гиперпараметры с помощью с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10;

- получить предсказания лучшей и базовой моделей на тестовой выборке, сделать выводы;

- сравнить качество работы лучшей модели на тренировочной и тестовой выборке.

1.3.3. Препроцессинг

Цель препроцессинга, или предварительной обработки данных — обеспечить корректную работу моделей.

Его необходимо выполнять после разделения на тренировочную и тестовую выборку, как будто мы не знаем параметров тестовой выборки (минимум, максимум, матожидание, стандартное отклонение).

Препроцессинг для категориальных и количественных признаков выполняется по-разному.

Категориальный признак один - 'Угол нашивки, град'. Он принимает значения 0 и 90. Модели отработают лучше, если мы превратим эти значения в 0 и 1 с помощью LabelEncoder или OrdinalEncoder.

Вещественных количественных признаков у нас большинство. Проблема вещественных признаков в том, что их значения лежат в разных диапазонах, в разных масштабах. Это видно в таблице 2. Необходимо провести одно из двух возможных преобразований:

- нормализацию — приведение в диапазон от 0 до 1 с помощью MinMaxScaler;
- стандартизацию — приведение к матожиданию 0, стандартному отклонению 1 с помощью StandartScaler.

Буду использую стандартизацию и StandardScaler.

А для метода KNN нормализацию MinMaxScaler.

Препроцессинг необходимо повторить в приложении для введенных данных. Поэтому я буду встраивать стандартизацию или нормализацию в pipline моделей машинного обучения.

1.3.4. Перекрестная проверка

Для обеспечения статистической устойчивости метрик модели используем перекрестную проверку или кросс-валидацию. Чтобы ее реализовать, выборка разбивается необходимое количество раз на тестовую и валидационную. Модель обучается на тестовой выбоке, затем выполняется расчет метрик качества на валидационной. В качестве результата мы получаем средние метрики качества для всех валидационных выборок. Перекрестную проверку реализует функция cross_validate из sklearn.

1.3.5. Поиск гиперпараметров по сетке

Поиск гиперпараметров по сетке реализует класс GridSearchCV из sklearn. Он получает модель и набор гиперпараметров, поочередно передает их в модель, выполняет обучение и определяет лучшие комбинации гиперпараметры. Перкрестная проверка уже встроена в этот класс.

1.3.6 Метрики качества моделей

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

- R2 или коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то прогнозы сопоставимы по качеству с константным предсказанием;

- MSE (Mean Squared Error) или средняя квадратичная ошибка между прогнозируемыми значениями и фактическими значениями в наборе данных. Чем ниже MSE, тем лучше модель соответствует набору данных.

- MAE (Mean Absolute Error) - средняя абсолютная ошибка так же принимает значениях в тех же единицах, что и целевая переменная;

RMSE, MAE, MAPE и max error принимают положительные значения. Но отображать я их буду со знаком «-». Так корректно отработает выделение цветом лучших моделей — эти метрики надо минимизировать.

R2 в норме принимает положительные значения. Эту метрику надо максимизировать. Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

2. Практическая часть

2.1. Разбиение и предобработка данных

2.1.1. Для прогнозирования модуля упругости при растяжении

Признаки датасета были разделены на входные и выходные, а строки - на тренировочное и тестовое множество. Описательная статистика входных признаков до и после предобработки показана на рисунке 10. Описательная статистика выходного признака показана на рисунке 11.

```
#Описательная входных статистика до обработки
```

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	936.0	-3.985416e-16	1.000535	-2.662564	-0.675917	-0.023457	0.695189	2.673947
Плотность, кг/м3	936.0	-1.040004e-15	1.000535	-2.678494	-0.714937	0.045471	0.651668	2.649773
модуль упругости, ГПа	936.0	1.480297e-16	1.000535	-2.239686	-0.729517	0.001489	0.670165	2.773444
Количество отвердителя, м.%	936.0	-4.934325e-17	1.000535	-2.673520	-0.680609	0.007288	0.706247	2.624101
Содержание эпоксидных групп,%_2	936.0	-8.264994e-16	1.000535	-2.721073	-0.684125	-0.010160	0.732284	2.818386
Температура вспышки, С_2	936.0	2.827748e-16	1.000535	-2.708660	-0.681975	-0.000414	0.684390	2.540085
Поверхностная плотность, г/м2	936.0	1.537232e-16	1.000535	-1.722572	-0.774851	-0.090207	0.760272	2.886532
Модуль упругости при растяжении, ГПа	936.0	-4.531987e-15	1.000535	-2.553495	-0.677361	-0.015119	0.660679	2.672150
Прочность при растяжении, МПа	936.0	7.591269e-16	1.000535	-2.625366	-0.691456	-0.020555	0.620548	2.670853
Потребление смолы, г/м2	936.0	-1.622634e-16	1.000535	-2.663276	-0.659631	0.013415	0.671035	2.447193
Угол нашивки, град	936.0	1.214603e-16	1.000535	-1.023787	-1.023787	0.976766	0.976766	0.976766
Шаг нашивки	936.0	-2.068621e-16	1.000535	-2.742040	-0.709873	0.011064	0.668120	2.717671
Плотность нашивки	936.0	-1.821904e-16	1.000535	-2.686557	-0.644710	0.011780	0.653975	2.542482

```
#Описательная статистика входных после обработки
```

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	936.0	2.925683	0.893712	0.547391	2.321931	2.904731	3.546650	5.314144
Плотность, кг/м3	936.0	1974.040023	70.808120	1784.482245	1923.443748	1977.258043	2020.158764	2161.565216
модуль упругости, ГПа	936.0	738.247627	328.708665	2.436909	498.577158	738.736842	958.418993	1649.415706
Количество отвердителя, м.%	936.0	110.916216	27.037891	38.668500	92.523816	111.113175	130.001450	181.828448
Содержание эпоксидных групп,%_2	936.0	22.209030	2.394871	15.695894	20.571516	22.184713	23.961818	28.955094
Температура вспышки, С_2	936.0	286.040414	39.400677	179.374391	259.184486	286.024118	312.991425	386.067992
Поверхностная плотность, г/м2	936.0	482.993901	280.190377	0.603740	266.004099	457.732246	695.900862	1291.340115
Модуль упругости при растяжении, ГПа	936.0	73.305127	3.037381	65.553336	71.248823	73.259230	75.310788	81.417126
Прочность при растяжении, МПа	936.0	2467.488822	463.838911	1250.392802	2146.936034	2457.959767	2755.169485	3705.672523
Потребление смолы, г/м2	936.0	217.613374	57.827255	63.685698	179.489091	218.388715	256.396777	359.052220
Угол нашивки, град	936.0	46.057692	45.011619	0.000000	0.000000	90.000000	90.000000	90.000000
Шаг нашивки	936.0	6.915585	2.509672	0.037639	5.134988	6.943337	8.591450	13.732404
Плотность нашивки	936.0	57.451895	11.239331	27.272928	50.209656	57.584225	64.798211	86.012427

Рисунок 10 – Описательная статистика входных признаков до и после предобработки для 1-й задачи

	Прочность при растяжении, МПа
count	936.000000
mean	2467.488822
std	463.838911
min	1250.392802
25%	2146.936034
50%	2457.959767
75%	2755.169485
max	3705.672523

Рисунок 11 – Описательная статистика выходного признака для 1-й задачи

2.1.2. Для прогнозирования модуля упругости при растяжении

Признаки датасета были разделены на входные и выходные, а строки - на тренировочное и тестовое множество. Описательная статистика входных признаков до и после предобработки показана на рисунке 12. Описательная статистика выходного признака показана на рисунке 13.

```
#Описательная входных статистика до обработки
```

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	936.0	-3.985416e-16	1.000535	-2.662564	-0.675917	-0.023457	0.695189	2.673947
Плотность, кг/м3	936.0	-1.040004e-15	1.000535	-2.678494	-0.714937	0.045471	0.651668	2.649773
модуль упругости, ГПа	936.0	1.480297e-16	1.000535	-2.239686	-0.729517	0.001489	0.670165	2.773444
Количество отвердителя, м.%	936.0	-4.934325e-17	1.000535	-2.673520	-0.680609	0.007288	0.706247	2.624101
Содержание эпоксидных групп,%_2	936.0	-8.264994e-16	1.000535	-2.721073	-0.684125	-0.010160	0.732284	2.818386
Температура вспышки, С_2	936.0	2.827748e-16	1.000535	-2.708660	-0.681975	-0.000414	0.684390	2.540085
Поверхностная плотность, г/м2	936.0	1.537232e-16	1.000535	-1.722572	-0.774851	-0.090207	0.760272	2.886532
Модуль упругости при растяжении, ГПа	936.0	-4.531987e-15	1.000535	-2.553495	-0.677361	-0.015119	0.660679	2.672150
Прочность при растяжении, МПа	936.0	7.591269e-16	1.000535	-2.625366	-0.691456	-0.020555	0.620548	2.670853
Потребление смолы, г/м2	936.0	-1.622634e-16	1.000535	-2.663276	-0.659631	0.013415	0.671035	2.447193
Угол нашивки, град	936.0	1.214603e-16	1.000535	-1.023787	-1.023787	0.976766	0.976766	0.976766
Шаг нашивки	936.0	-2.068621e-16	1.000535	-2.742040	-0.709873	0.011064	0.668120	2.717671
Плотность нашивки	936.0	-1.821904e-16	1.000535	-2.686557	-0.644710	0.011780	0.653975	2.542482

```
#Описательная статистика входных после обработки
```

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	936.0	2.925683	0.893712	0.547391	2.321931	2.904731	3.546650	5.314144
Плотность, кг/м3	936.0	1974.040023	70.808120	1784.482245	1923.443748	1977.258043	2020.158764	2161.565216
модуль упругости, ГПа	936.0	738.247627	328.708665	2.436909	498.577158	738.736842	958.418993	1649.415706
Количество отвердителя, м.%	936.0	110.916216	27.037891	38.668500	92.523816	111.113175	130.001450	181.828448
Содержание эпоксидных групп,%_2	936.0	22.209030	2.394871	15.695894	20.571516	22.184713	23.961818	28.955094
Температура вспышки, С_2	936.0	286.040414	39.400677	179.374391	259.184486	286.024118	312.991425	386.067992
Поверхностная плотность, г/м2	936.0	482.993901	280.190377	0.603740	266.004099	457.732246	695.900862	1291.340115
Модуль упругости при растяжении, ГПа	936.0	73.305127	3.037381	65.553336	71.248823	73.259230	75.310788	81.417126
Прочность при растяжении, МПа	936.0	2467.488822	463.838911	1250.392802	2146.936034	2457.959767	2755.169485	3705.672523
Потребление смолы, г/м2	936.0	217.613374	57.827255	63.685698	179.489091	218.388715	256.396777	359.052220
Угол нашивки, град	936.0	46.057692	45.011619	0.000000	0.000000	90.000000	90.000000	90.000000
Шаг нашивки	936.0	6.915585	2.509672	0.037639	5.134988	6.943337	8.591450	13.732404
Плотность нашивки	936.0	57.451895	11.239331	27.272928	50.209656	57.584225	64.798211	86.012427

Рисунок 12 – Описательная статистика входных признаков до и после

предобработки для 2-й задачи

	Модуль упругости при растяжении, ГПа
count	936.000000
mean	73.305127
std	3.037381
min	65.553336
25%	71.248823
50%	73.259230
75%	75.310788
max	81.417126

Рисунок 13 – Описательная статистика выходного признака для 2-й задачи

2.2. Разработка и обучение моделей для прогнозирования модуля упругости при растяжении

Для подбора лучшей модели для этой задачи я взял следующие модели:

- LinearRegression — линейная регрессия (раздел 1.2.1);
- Ridge — гребневая регрессия (раздел 1.2.2);
- KneighborsRegressor — метод ближайших соседей (раздел 1.2.4);
- RandomForestRegressor — случайный лес (раздел 1.2.6).

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

Метрики работы выбранных моделей с подобранными гиперпараметрами полученные с помощью перекрестной проверки на тестовом множестве, приведены на рисунке 14.

	Model	MSE	MAE	R2 score
Модуль упругости при растяжение, ГПа	KNeighborsRegressor	8.678704	2.392482	0.00408
Модуль упругости при растяжении, ГПа	DummyRegressor	8.765902	2.398634	-0.00600
Модуль упругости при растяжение, ГПа	RandomForestRegressor	8.735981	2.404383	-0.00200
Модуль упругости при растяжение, ГПа	Rige	8.732845	2.412596	-0.00200
Модуль упругости при растяжении, ГПа	LinearRegression	8.745602	2.416002	-0.00400

Рисунок 14 – Результаты моделей после подбора гиперпараметров

Ни одна из выбранных мной моделей не оказалась подходящей для наших данных.

Коэффициент детерминации R2 чуть больше 0 для метода ближайших соседей. Это чуть лучше базовой модели. У всех моделей остальные метрики примерно совпадают с базовой моделью, в соответствии с рисунком 15

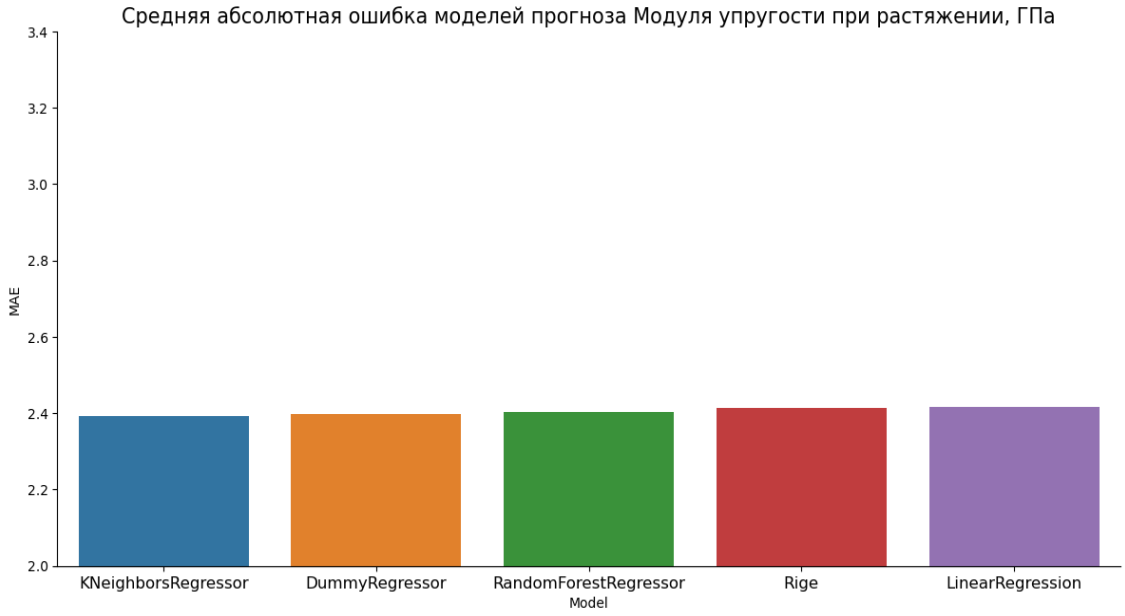

Рисунок 15 - MSE для модуля упругости

Поэтому в качестве лучшей модели выбираю метод ближайших соседей. На рисунке 16 приведена визуализация работы лучшей модели на тестовом множестве.

	Model	MSE	MAE	R2 score
Модуль упругости при растяжении, ГПа	DummyRegressor	8.765902	2.398634	-0.00600
Модуль упругости при растяжение, ГПа	KNeighborsRegressor	8.678704	2.392482	0.00408

Рисунок 16 – Метрики работы лучшей модели на тестовом множестве

Метрики работы лучшей модели на тестовом множестве и сравнение с базовой отражены на рисунке 20. Они подтверждают: полученная модель хуже базовой. Результат исследования отрицательный. Не удалось получить модели, которая могла бы оказать помощь в принятии решений специалисту предметной области.

2.3. Для прогнозирования прочности при растяжении

Для подбора лучшей модели для этой задачи я взяла следующие модели:

- LinearRegression — линейная регрессия (раздел 1.2.1);
- Ridge — гребневая регрессия (раздел 1.2.2);
- RandomForestRegressor — случайный лес (раздел 1.2.3);
- KneighborsRegressor — (раздел 1.2.5);

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

Метрики работы выбранных моделей с подобранными гиперпараметрами полученные с помощью перекрестной проверки на тестовом множестве, приведены на рисунке 17.

	Model	MSE	MAE	R2 score
Прочность при растяжении, МПа	KNeighborsRegressor	209750.489526	360.097838	-0.028
Прочность при растяжении, МПа	DummyRegressor	210162.565630	360.306251	-0.030
Прочность при растяжении, МПа	Rige	212349.035083	361.903011	-0.040
Прочность при растяжении, МПа	LinearRegression	213381.584134	362.854584	-0.045
Прочность при растяжении, МПа	RandomForestRegressor	212325.620988	363.774842	-0.040

Рисунок 17 – Результаты моделей после подбора гиперпараметров

R2 близок к 0 для линейных моделей и метода ближайших соседей. Значит, они не лучше базовой модели. И остальные метрики у них примерно совпадают с базовой моделью, в соответствии с рисунком 18

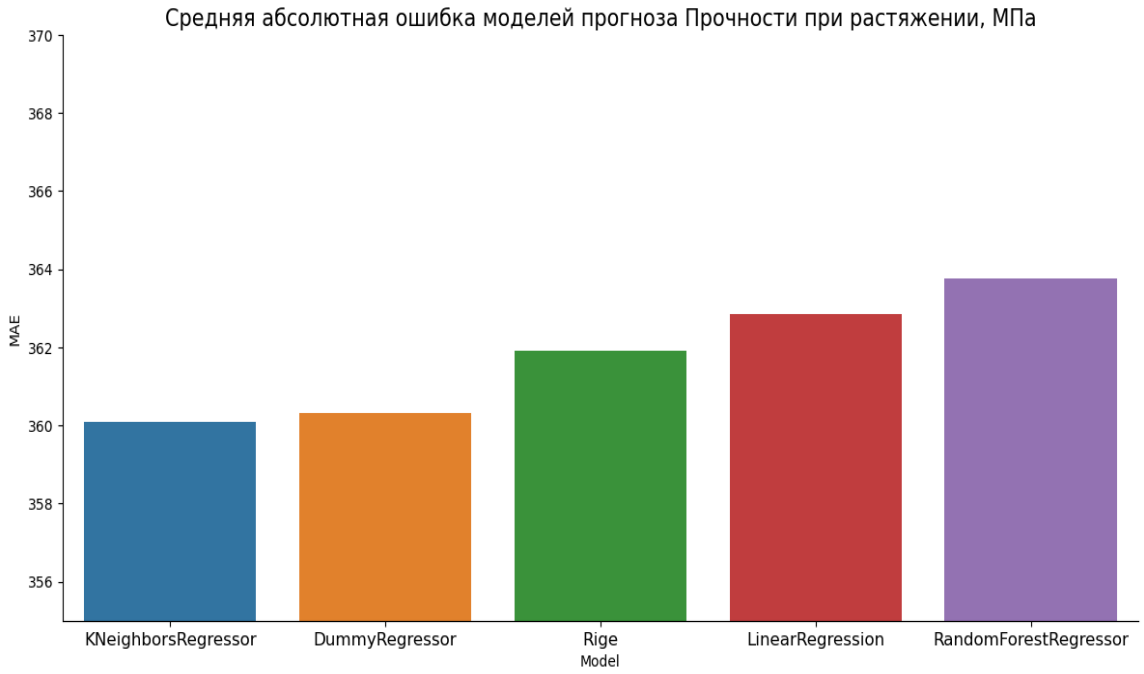

Рисунок 18 – MSE для прочности при растяжение

Гораздо хуже линейных моделей умолчанию отработал случайный лес.

Но лучший результат дает метод ближайших соседей. Значения ошибок примерно такие же, как у базовой модели. Но коэффициент детерминации немного больше, что показывает чуть лучшую объясняющую способность модели.

Метрики работы лучшей модели на тестовом множестве и сравнение с базовой отражены на рисунке 19. Несмотря на то, что метод ближайших соседий показывает результаты чуть-чуть лучше базовой, результат исследования отрицательный. Не удалось получить модели, которая могла бы оказать помощь в принятии решений специалисту предметной области.

	Model	MSE	MAE	R2 score
Прочность при растяжении, МПа	DummyRegressor	210162.565630	360.306251	-0.030
Прочность при растяжении, МПа	KNeighborsRegressor	209750.489526	360.097838	-0.028

Рисунок 19 – Метрики работы лучшей модели на тестовом множестве

2.4 Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель

По заданию для соотношения матрица-наполнитель необходимо построить нейросеть. Но для сравнения нам также понадобится базовая модель DummyRegressor, возвращающая среднее целевого признака.

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

- входной слой для 12 признаков;
- выходной слой для 1 признака;
- скрытых слоев: 1;
- нейронов на скрытом слое: 4 ;
- активационная функция скрытых слоев: relu;
- оптимизатор: Adam;
- loss-функция: Mean.
- Архитектура нейросети приведена на рисунках 28

Layer (type)	Output Shape	Param #
dense_112 (Dense)	(None, 12)	156
batch_normalization_57 (BatchNormalization)	(None, 12)	48
dense_113 (Dense)	(None, 4)	52
batch_normalization_58 (BatchNormalization)	(None, 4)	16
dense_114 (Dense)	(None, 1)	5

Total params: 277 (1.08 KB)

Trainable params: 245 (980.00 B)

Non-trainable params: 32 (128.00 B)

Рисунок 20 – Архитектура нейросети

Запускаю обучение нейросети со следующими параметрами:

- пропорция разбиения данных на тестовые и валидационные: 30%;
- количество эпох: 100.
- используя раннюю остановку.

График потерь модели на тренировочной и тестовой выборках приведен на рисунке 21

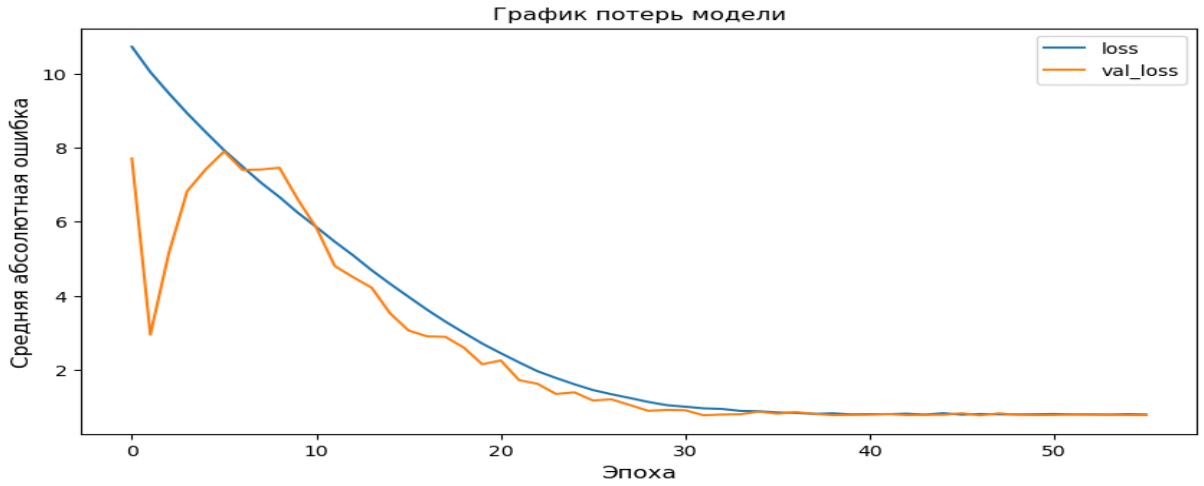

Рисунок 21 – График потерь модели на тренировочной и тестовой выборках

Визуализация результата работы нейросети отображена на рисунке 22, а её метрики относительно базовой модели на рисунке 23.

Рисунок 21 – Результат работы нейросети

	Model	MSE	MAE	R2 score
Матрица-наполнитель	DummyRegressor	0.762885	0.694175	-0.002
Матрица-наполнитель	Нейронная сеть	0.781559	0.714372	-0.026

Рисунок 21 – Метрики нейросети

Визуализация результатов показывает, что нейросеть из библиотеки Keras старалась подстроиться к данным. Выглядят результаты «похоже», но метрики чуть хуже чем у базовой модели.

2.5. Тестирование модели

Согласно заданию, необходимо сравнить ошибку каждой модели на тренировочной и тестирующей части выборки.

Модель для предсказания модуля упругости при растяжении и прочности при растяжении– KneighborsRegressor. Этот метод, который показал положительный, хоть и близкий к 0 коэффициент детерминации. Сравнение ее ошибок показано на рисунке 22, 23.

	Model	MSE	MAE	R2 score
Модуль упругости при растяжение, Train	KNeighborsRegressor_train	9.366098	2.468172	0.00500
Модуль упругости при растяжение, Test	KNeighborsRegressor test	8.690824	2.392496	0.00269

Рисунок 22 – Сравнение ошибок модели для модуля упругости при растяжении на тренировочном и тестовом датасете.

	Model	MSE	MAE	R2 score
Прочность при растяжении, Train	KNeighborsRegressor_train	215666.156117	371.301244	0.010
Прочность при растяжении, Test	KNeighborsRegressor_test	209750.489526	360.097838	-0.028

Рисунок 23 – Сравнение ошибок модели для прочности при растяжении на тренировочном и тестовом датасете.

Метод ближайших соседей имеет ошибку на тренировочном датасете больше, чем на тестовом. Это означает что у тестовой выборки репрезентативность выше, чем у обучающей. Если обучающая и тестовая выборка были независимы, то оценка, сделанная по тестовой выборке, является несмещенной (математическое ожидание равно истинному значению оцениваемого параметра). Оценку качества, сделанную по тестовой выборке, можно применить для выбора лучшей модели.

Модели работает чуть точнее среднего, и бесполезны для применения в реальных условиях.

Модель для предсказания соотношения матрица-наполнитель — нейро-сеть из keras, обученная с ранней остановкой. Сравнение ее ошибок показано на рисунке 24.

	Model	MSE	MAE	R2 score
Матрица-наполнитель trein	Нейронная сеть	0.862030	0.755862	-0.06
Матрица-наполнитель test	Нейронная сеть	0.876235	0.765448	-0.15

Рисунок 25 - Сравнение ошибок модели для соотношения матрица-наполнитель на тренировочном и тестовом датасете.

У нейросети показатели для тестовой выборки отличаются в худшую сторону от показателей тренировочной. Это говорит о том, что она не на шла закономерностей, а стала учить данные из тестовой выборки. Возможно, требуется более тщательное и грамотное построение архитектуры нейронной сети, чтобы получить лучший результат. Но сейчас задача далека от решения.

Модель работает не точнее среднего, и бесполезна для применения в реальных условиях.

2.6. Разработка приложения

Несмотря на то, что пригодных к внедрению моделей получить не удалось, можно разработать функционал приложения. Возможно, дальнейшие исследования позволят построить качественную модель и внедрить ее в готовое приложение.

Решено разработать веб-приложение с помощью языка Python, фреймворка Streamlit.

Слева виджеты для ввода входных параметров, после ввода которых на экране отображаются значения:

- Модуль упругости при растяжении, ГПа;
- Прочность при растяжении, МПа.

Подробнее с приложением можно ознакомиться в репозитории на git- hab: https://github.com/rewcom/BMSTU_DS_project/tree/main/streamlit_app

Заключение

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 транформацию данных. Здесь нам был предоставлен гото-

вый 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 обу- чающую, 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 реше- ний специалистом предметной области, которое использовало бы найденную модель;
- внедрение решения и приложения в эксплуатацию. Этот блок задач мы тоже пока не затронули.

В этой работе мы имели дело не с учебными наборами данных, которые дают хорошо изученные решения, а с реальной производственной задачей. И к сожалению, не смогли поставленную задачу решить — не получили моделей, которые бы описывали закономерности предметной области. Я проделала мак- симум исследований, которые в моей компетенции как начинающего дата-сай- ентиста, применила большую часть знаний, полученных в ходе прохождения курса.

Возможные причины неудачи:

- нечеткая постановка задачи, отсутствие дополнительной информации о зависимости признаков с точки зрения физики процесса.

Незначимые при- знаки являются для модели шумом, и мешают найти зависимость целевых от значимых входных признаков;

- исследование предварительно обработанных данных. Возможно, на "сырых", не предобработанных данных можно было бы получить более каче- ственные модели, воспользовавшись другими методами очистки и подготовки;

- мой недостаток знаний и опыта. Нейросети являются самым современ- ным подходам к решению такого рода задач. Они способны находить скрытые и нелинейные зависимости в данных. Но выбор оптимальной архитектуры нейросети является неочевидной задачей.

Дальнейшие возможные пути решения этой задачи могли бы быть:

- углубиться в изучение нейросетей, попробовать различные архитекту- ры, параметры обучения и т.д.;

- провести отбор признаков разными методами. Испробовать методы уменьшения размерности, например метод главных компонент;

- после уменьшения размерности градиентный бустинг может улучшить свои результаты. Так же есть большой простор для подбора гиперпараметров для этого метода;

- проконсультироваться у экспертов в предметной области. Возможно, они могли бы поделиться знаниями, необходимыми для решения задачи.

Библиографический список

1 Композиционные материалы : учебное пособие для вузов / Д. А. Иванов, А. И. Ситников, С. Д. Шляпин ; под редакцией А. А. Ильина. — Москва : Из- дательство Юрайт, 2019 — 253 с. — (Высшее образование). — Текст : непо- средственный.

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., пере- раб. 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 Руководство по быстрому старту в flask: – Режим доступа: https://flask-

russian-docs.readthedocs.io/ru/latest/quickstart.html.

12 Loginom Вики. Алгоритмы: – Режим доступа: https://wiki.loginom.ru/ algorithms.html.