МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное 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 Калашникова Елена Александровна

Москва, 2026

Содержание

Введение

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. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита — железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. 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.

Физические испытания бывают дорогостоящими, длительными и трудозатратными. Процесс изготовления опытных образцов требует специализированного оборудования, соблюдения строгих температурных режимов и значительного расхода дорогостоящего сырья.

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

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

1. Аналитическая часть

1.1 Постановка задачи

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.

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

Для решения поставленной задачи использовано два датасета.

Первый датасет содержит 10 признаков, объем выборки – 1023 строки. Второй датасет содержит 3 признака, объем выборки – 1040 строк.

Общий исходный датасет состоит из 1023 строк и 13 признаков (столбцов). В качестве входных данных использовано 11 признаков, целевыми переменными являются столбцы Модуль упругости при растяжении и Прочность при растяжении.

	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
0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0	0	4.0	57.0
1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0	0	4.0	60.0
2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0	0	4.0	70.0
3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0	0	5.0	47.0
4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0	0	5.0	57.0
5	2.767918	2000.0	748.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0	0	5.0	60.0
6	2.569620	1910.0	807.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0	0	5.0	70.0
7	2.561475	1900.0	535.000000	111.86	22.267857	284.615385	380.0	75.0	1800.0	120.0	0	7.0	47.0
8	3.557018	1930.0	889.000000	129.00	21.250000	300.000000	380.0	75.0	1800.0	120.0	0	7.0	57.0
9	3.532338	2100.0	1421.000000	129.00	21.250000	300.000000	1010.0	78.0	2000.0	300.0	0	7.0	60.0

Рисунок 1 – Исходный датасет

Все данные являются числовыми, пропуски в датасете отсутствуют. Каждый из признаков, а таже целевые переменные содержат выбросы.

Очевидные корреляционные зависимости между целевыми переменными и признаками отсутствуют, линейных зависимостей нет.

1.2. Описание используемых методов

Для решения поставленной задачи использованы следующие методы:

- линейная регрессия;
- KNN-регрессия;
- регрессия на основе дерева решений.

В качестве первоначальной модели предсказания целевых переменных была выбрана линейная регрессия. Несмотря на отсутствие явных линейных зависимостей, линейная регрессия является самой простой базовой моделью регрессии, поэтому была выбрана первой для предсказания целевых переменных.

Это математическая модель, предполагающая, что зависимость между переменными можно описать линейной функцией. В основе ее работы лежит поиск таких весов и смещений, при которых предсказания, опирающиеся на обучающие данные, будут максимально точными.

Линейная регрессия легко интерпретируется и обладает высокой скоростью работы. При этом в случае отсутствия прямых линейных зависимостей точность модели будет низкой.

Второй моделью была выбрана KNN-регрессия – метод, основанный на идее близости объектов в пространстве признаков.

KNN является непараметрическим методом машинного обучения, что означает отсутствие жестко заданных параметров модели, таких как веса и смещение, присущие линейной регрессии. Вместо этого модель «запоминает» обучающие данные и применяет их напрямую к новым точкам данных для принятия решений.

Из минусов выбранной модели модно отметить следующие: для работы модели требуется большой объем памяти для выполнения расчета расстояний между всеми объектами выборки, чувствительность к масштабу данных и выбросам.

Следующей моделью была выбрана регрессия на основе дерева решений.

Главный принцип построения дерева решений для регрессии заключается в разбиении пространства признаков на области, где значение целевой переменной примерно одинаково. Каждый внутренний узел дерева представляет собой условие проверки определенного признака, а каждый листовой узел содержит среднее значение целевой переменной для всех наблюдений, попавших в эту область.

Данная модель также не требует наличия линейных зависимостей, при этом не требуется и масштабирование данных.

Вместе с тем модель имеет склонность к «переобучению», прогноз, полученный моделью, является дискретным, что для регрессии может повлечь уменьшение точности.

Помимо простых моделей в работе были использованы ансамблевые методы:

- градиентный бустинг;
- случайный лес.

Основной принцип работы ансамблевых методов заключается в обучении нескольких слабых моделей для решения одной и той же задачи и объединения их результатов для получения итогового решения. Совокупность слабых моделей позволяет получить более точные и надежные модели.

Первым ансамблевым методом был выбран градиентный бустинг.

Градиентный бустинг – это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Обучение ансамбля проводится последовательно. На каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке. Слабые алгоритмы обучаются на остатках (ошибках) предыдущих, корректируя их с помощью градиентного спуска.

Данная модель обладает высокой точностью, не требует масштабирования признаков.

Среди минусов можно отметить склонность к «переобучению», а также чувствительность к выбросам.

Вторым ансамблевым методом был выбран случайный лес – алгоритм машинного обучения, который состоит из множества отдельных решающих деревьев.

В данном алгоритме каждое дерево строится независимо друг от друга на разных подвыборках обучающих данных. При этом при обучении каждого дерева используются разные комбинации признаков (характеристик) объектов, для которых делается предсказание. Комбинация независимых деревьев повышает точность модели.

Модель устойчива к выбросам и гораздо стабильнее одиночного дерева решений за счет предсказания целевой переменной на основании голосования отдельных деревьев, входящих в ансамбль.

Из минусов можно отметить дискретность значений, как и у одиночного дерева решений.

Работоспособность и точность каждой из моделей зависит от предподготовки исходного датасета, в том числе наличия/отсутствия линейных зависимостей, выбросов, масштабирования.

Основные требования к датасету для оптимизации работы каждого из алгоритмов приведены в таблице 1.

Таблица 1 – Требования к датасету

	Линейная регрессия	KNN-регрессия	Дерево решений	Градиентный бустинг	Случайный лес
Линейная зависимость	да	нет	нет	нет	нет
Масштабирование	не обязательно	да	нет	нет	нет
Устойчивость к выбросам	низкая	низкая	высокая	средняя	высокая
Мульти-коллинеарность	недопустима	средне	не влияет	не влияет	не влияет
Большое количество признаков	плохо	плохо	хорошо	хорошо	хорошо
Интерпрети-руемость	высокая	средняя	высокая	высокая	низкая
Склонность к переобучению	низкая	средняя	высокая	высокая	низкая
Дискретность прогноза	непрерывный	гладкий	дискретный	дискретный	дискретный

Учитывая требования и свойства каждого из алгоритмов, можно предположить, что на анализируемом датасете линейная регрессия и метод ближайших соседей не дадут хороших результатов, при этом из-за дискретности прогнозных значений точность остальных алгоритмов тоже могут быть достаточно небольшой.

1.3. Разведочный анализ данных

Как уже отмечалось ранее, общий исходный датасет состоит из 1023 строк и 13 признаков (столбцов). В качестве входных данных использовано 11 признаков, целевыми переменными являются столбцы Модуль упругости при растяжении и Прочность при растяжении.

Все данные являются числовыми, пропуски в датасете отсутствуют. Каждый из признаков, а таже целевые переменные содержат выбросы.

Очевидные корреляционные зависимости между целевыми переменными и признаками отсутствуют, линейных зависимостей нет.

Разберем датасет более подробно.

С помощью функции describe библиотеки pandas проанализируем данные датасета. На рисунке 2 представлен датасет с описательной статистикой.

	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
count	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000
mean	2.930366	1975.734888	739.923233	110.570769	22.244390	285.882151	482.731833	73.328571	2466.922843	218.423144	44.252199	6.899222	57.153929
std	0.913222	73.729231	330.231581	28.295911	2.406301	40.943260	281.314690	3.118983	485.628006	59.735931	45.015793	2.563467	12.350969
min	0.389403	1731.764635	2.436909	17.740275	14.254985	100.000000	0.603740	64.054061	1036.856605	33.803026	0.000000	0.000000	0.000000
25%	2.317887	1924.155467	500.047452	92.443497	20.608034	259.066528	266.816645	71.245018	2135.850448	179.627520	0.000000	5.080033	49.799212
50%	2.906878	1977.621657	739.664328	110.564840	22.230744	285.896812	451.864365	73.268805	2459.524526	219.198882	0.000000	6.916144	57.341920
75%	3.552660	2021.374375	961.812526	129.730366	23.961934	313.002106	693.225017	75.356612	2767.193119	257.481724	90.000000	8.586293	64.944961
max	5.591742	2207.773481	1911.536477	198.953207	33.000000	413.273418	1399.542362	82.682051	3848.436732	414.590628	90.000000	14.440522	103.988901

Рисунок 2 – Описательная статистика датасета

Как видно из таблицы, для большинства данных среднее значение и медиана лежат достаточно близко, значит аномально высокие выбросы отсутствуют, только для признаков Поверхностная плотность и Угол нашивки эти значения отличаются достаточно сильно. Можно сделать вывод, что у признака Поверхностная плотность будет достаточно большой «хвост» у

распределения. Для признака Угол нашивки такой вывод сделать нельзя, так как значения признака содержат всего 2 значения (0 и 90 градусов). Данный признак можно рассматривать как категориальный.

Для более детального анализа распределений были построены гистограммы распределения, приведенные на рисунке 3.

Как видно из построенных гистограмм, многие признаки имеют распределение, близкое к нормальному, например Соотношение матрица-наполнитель, Плотность, Температура вспышки, Содержание эпоксидных групп. Признак Поверхностная плотность, как и предполагалось, имеет достаточно большой «хвост», Угол нашивки – категориальный признак.

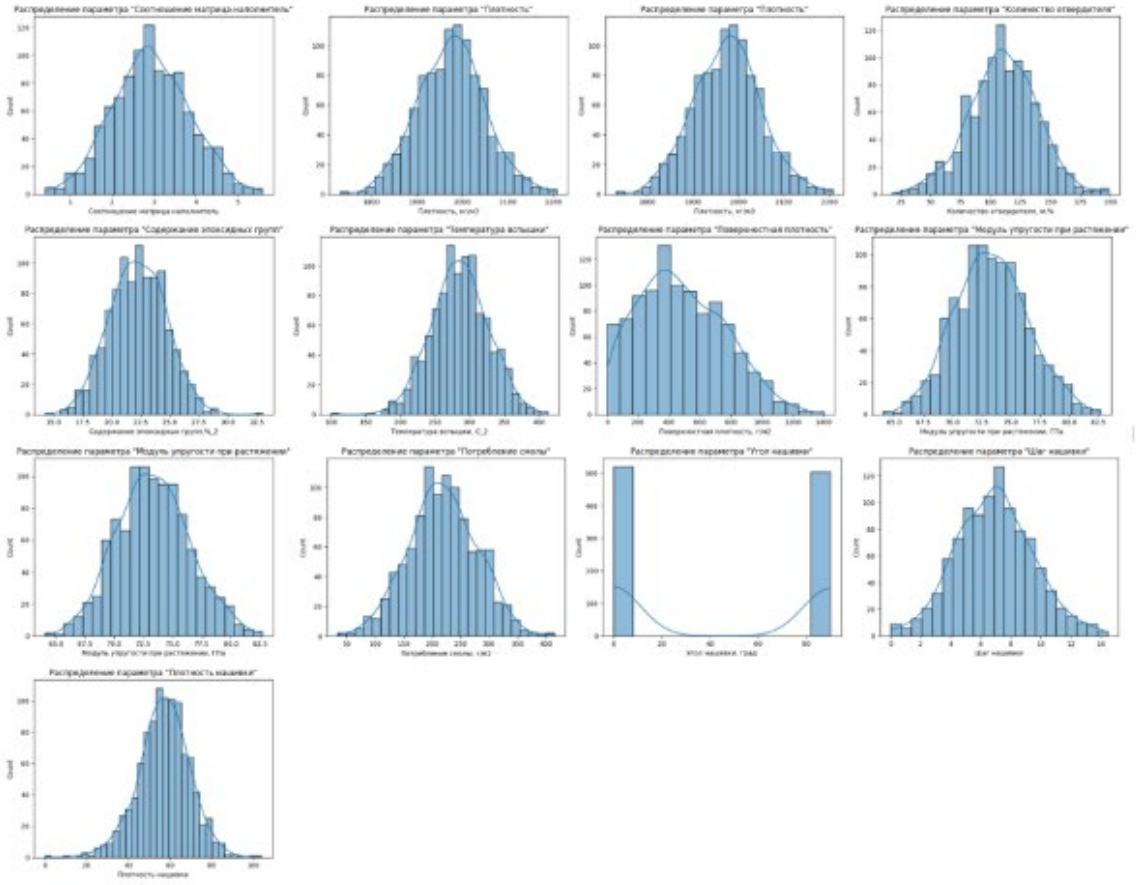

Рисунок 3 – Гистограммы распределения переменных

Также для более детального анализа выбросов построены диаграммы «ящик с усами», приведенные на рисунке 4.

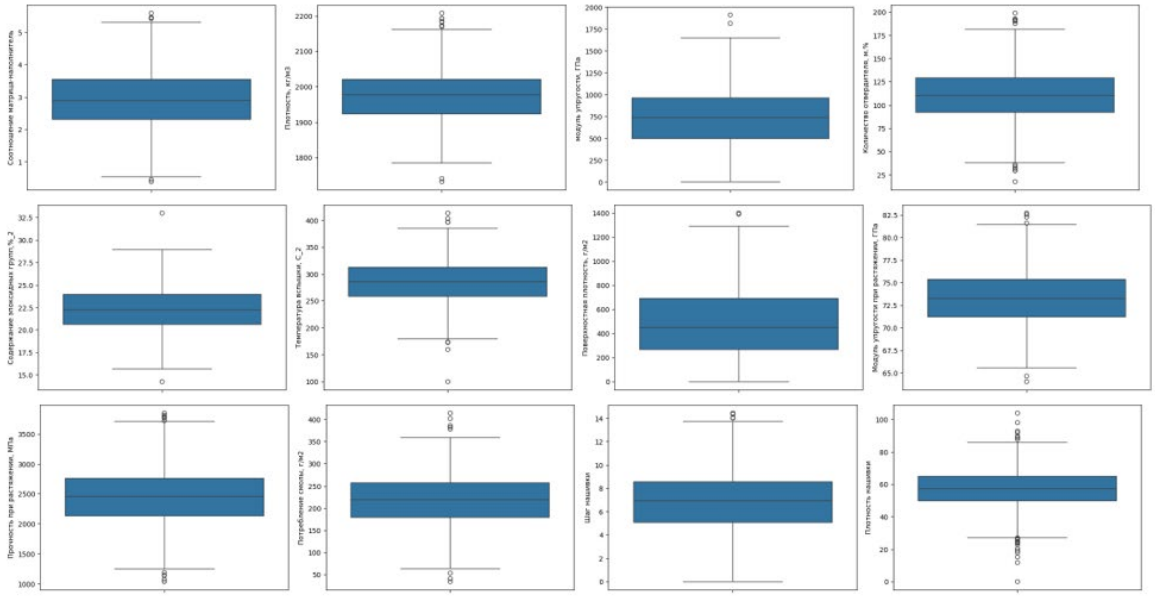

Рисунок 4 – Диаграммы «ящик с усами»

Как видно из диаграмм, у всех признаков датасета и целевых переменных есть выбросы. Диаграмма для признака Угол нашивки не строилась, так как признак фактически категориальный.

Для анализа датасета на наличие линейных зависимостей между переменными была построена тепловая карта корреляции (рисунок 5).

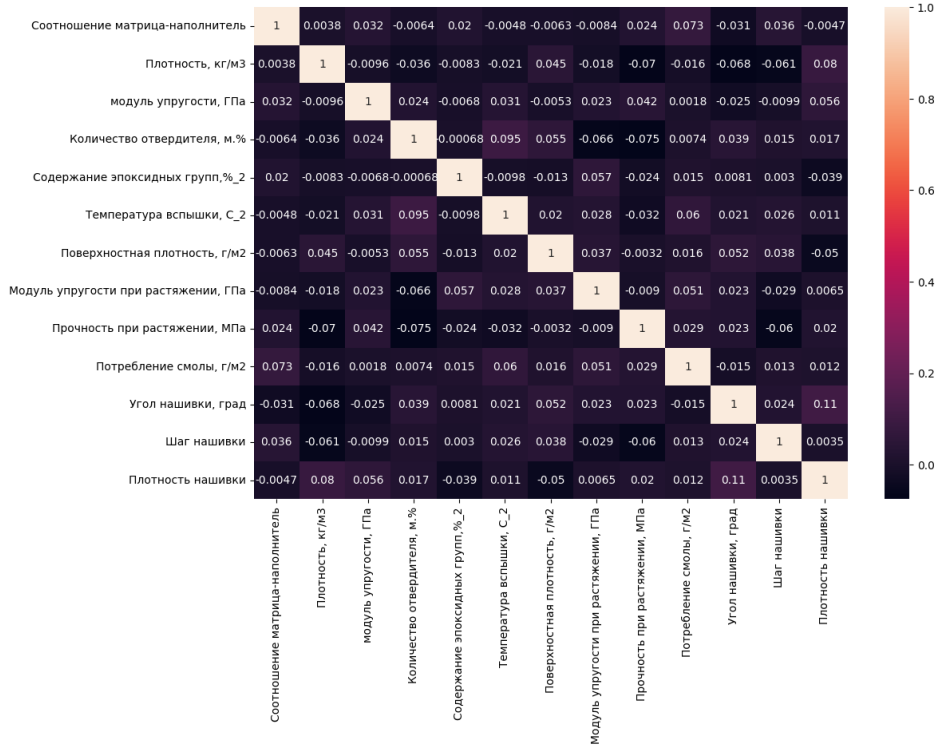

Рисунок 5 – Тепловая карта корреляции

Как видно из построенной карты корреляции, все коэффициенты корреляции между переменными очень малы, из чего следует что прямая линейная зависимость между переменными практически отсутствует.

Рассмотренная карта корреляции была построена на основании корреляции Пирсона, которая ищет линейную зависимость. Для дополнительного анализа была построена тепловая карта корреляции по методу Спирмена, которая проверяет, изменяются ли переменные в одном направлении (рисунок 6).

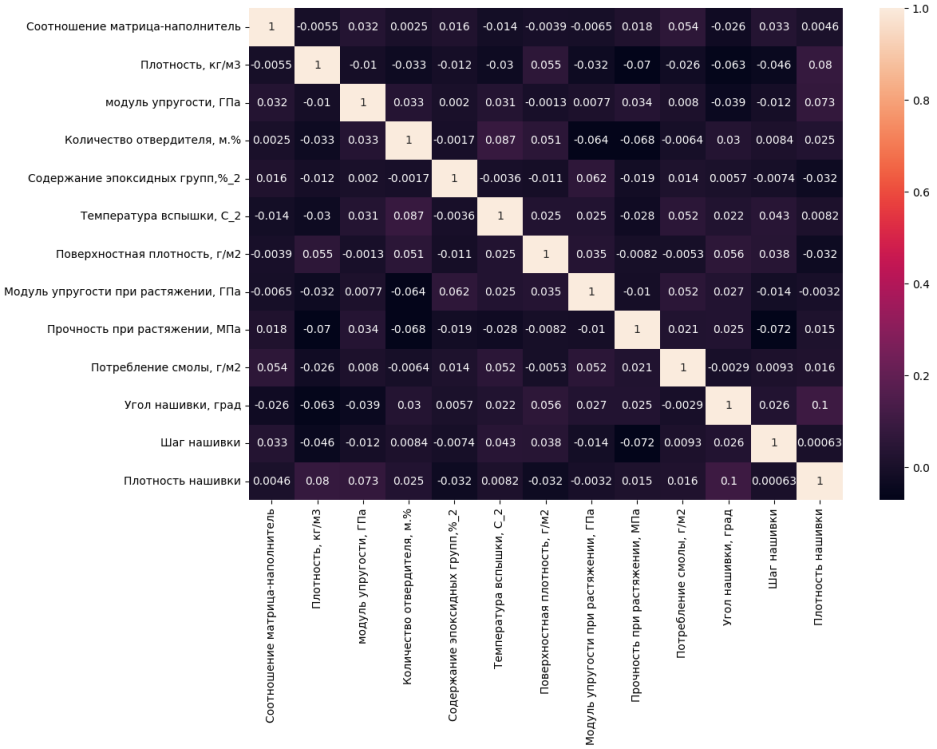

Рисунок 6 – Тепловая карта корреляции по Спирмену

На основании данной тепловой карты видно, что между переменными отсутствует даже нелинейная постоянная зависимость.

Для более детальной оценки наличия зависимостей целевых переменных от признаков были построены попарные графики рассеяния.

На рисунке 7 представлены графики рассеяния для целевой переменной Модуль упругости при растяжении. Как видно из графиков, все графики рассеяния выглядят как облака случайных точек, очевидные зависимости между переменными отсутствуют.

Аналогичная ситуация наблюдается для попарных графиков рассеяния для Прочности при растяжении, представленных на рисунке 8.

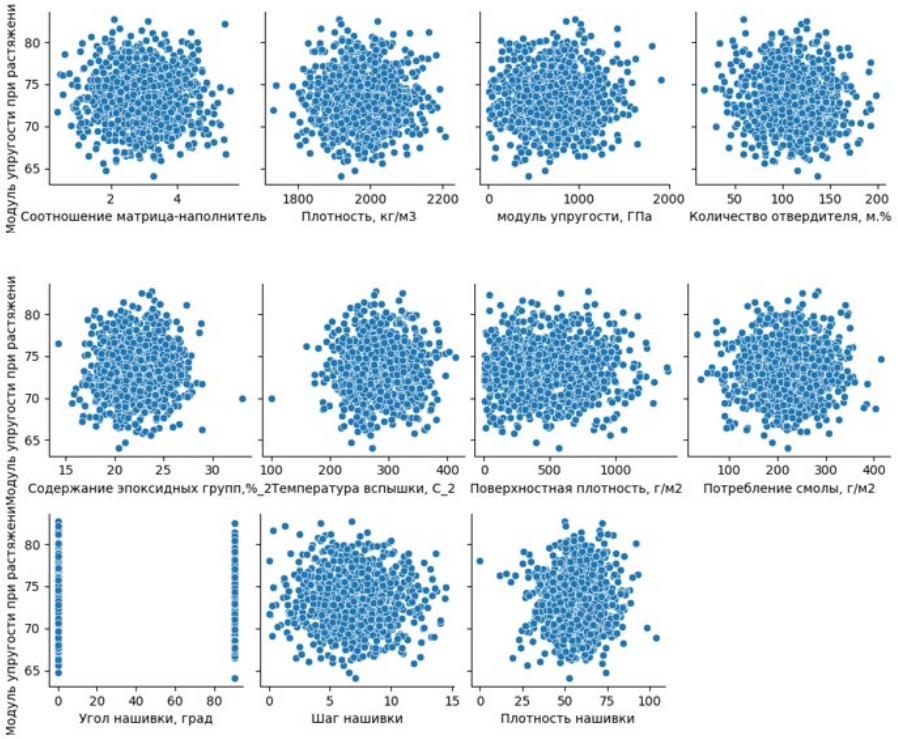

Рисунок 7 – Попарные графики рассеяния для Модуля упругости при растяжении

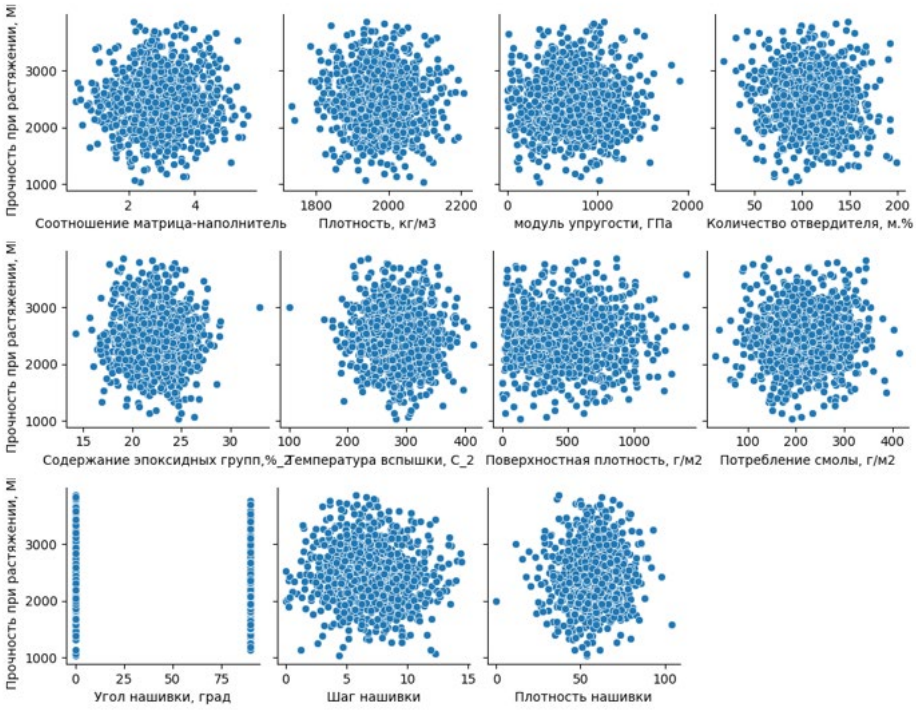

Рисунок 8 – Попарные графики рассеяния для Прочности при растяжении

Таким образом, на основании проведенного анализа можно сделать вывод, что прямые зависимости между целевыми переменными и признаками отсутствуют, в связи с чем можно сделать предварительный вывод о том, что линейная регрессия не даст высоких результатов. Кроме того, метод ближайших соседей также, скорее всего, не даст высоких результатов в связи с тем, что ближайшие по признакам объекты имеют разные значения целевых переменных, в связи с чем модели будет трудно определить «ближайших соседей».

2. Практическая часть

2.1. Предобработка данных

По результатам разведочного анализа данных было установлено, что все переменные имеют выбросы. Так как часть выбранных методов регрессии чувствительны к выбросам, была проведена работы по выявлению выбросов. Несмотря на то, что часть переменных имеет распределение, близкое к нормальному, в целях единообразия подхода для выявления выбросов использовался метод межквартильного размаха.

Для начала была проведена работа с выбросами для целевых переменных. Было определено, что для Модуля упругости при растяжении количество выбросов составляет 6 значений, для Прочности при растяжении – 11.

С учетом общего объема выборки 1023, указанные выбросы можно просто удалить. После удаления выбросов общая выборка составляет 1006 строк.

Далее было определено количество выбросов по каждому из признаков. В таблице 2 представлена сводная информация о выбросах.

Таблица 2 – Сводная информация о выбросах

Признак	Количество выбросов
Соотношение матрица-наполнитель	5
Плотность	9
Модуль упругости	3
Количество отвердителя	14
Содержание эпоксидных групп	2
Температура вспышки	3
Поверхностная плотность	2
Потребление смолы	8
Шаг нашивки	4
Плотность нашивки	21
ИТОГО:	71

Общее количество выбросов достаточно большое, после удаления всех выбросов выборка составляет 935 строк. Чтобы сохранить объем выборки

также был использован метод замены выбросов на граничные значения. В таком случае объем выборки составил 1006 строк.

Так как объем выбросов достаточно большой, также был использован метод IsolationForest, чтобы комплексно оценить многомерные выбросы.

Указанный метод выявил 51 выброс, после удаления которых общий объем выборки составил 955 строк.

Также в рамках предобработки данных были проведены стандартизация и нормализация всех полученных датасетов. Оба метода были использованы для оценки в дальнейшем работы моделей, чтобы оценить, какой из методов предобработки дадут более точные предсказания.

2.2. Разработка и обучение моделей

Несмотря на отсутствие явных линейных зависимостей, первым методом регрессии была выбрана линейная регрессия.

Для первой модели была выбрана выборка, полученная путем удаления всех одномерных выбросов, которая была разделена на обучающую и тестовую в соотношении 70% на 30%.

Несмотря на достаточно небольшие значения абсолютных ошибок при расчете Модуля упругости при растяжении (MAE – 2,268, MAPE – 3,1 %) и достаточно близкие значения средней абсолютной ошибки и корня из среднеквадратичной ошибки (RMSE – 2,86), значение коэффициента детерминации – -0.0234, что свидетельствует о том, что средняя ошибка выше, чем дисперсия, значит модель не смогла найти зависимостей.

Аналогичная ситуация для Прочности при растяжении, коэффициент детерминации также меньше 0.

Чтобы проанализировать возможность применения метода линейной регрессии на анализируемых данных, модель была обучена также на нормализованных данных.

Учитывая, что метрики не улучшились, были предприняты попытки выделить наиболее значимые признаки и построить модель на них.

Значимые признаки отбирались двумя способами: на основании тепловой карты корреляции, а также на основании полученных коэффициентов при обучении модели линейной регрессии на нормализованных данных.

Таким образом, одна модель линейная регрессия была построена для датасета, содержащего признаки Количество отвердителя, Содержание эпоксидных групп, Потребление смолы, другая модель – для датасета, содержащего признаки Содержание эпоксидных групп, Поверхностная плотность, Потребление смолы.

Сводная таблица полученных метрик приведена в таблице 3.

Таблица 3 – Сводная таблица полученных метрик модели линейной регрессии

	MAE	MAPE	MSE	RMSE	R^2
Модуль упругости при растяжении					
Исходный датасет, полученный путем удаления всех одномерных выбросов	2,268	0,031	8,2	2,86	-0,023
Нормализованный датасет, полученный путем удаления всех одномерных выбросов	0,14	0,53	0,032	0,18	-0,023
Исходный датасет с признаками на основе тепловой карты корреляции	2,235	0,03	8,00	2,83	0,001
Исходный датасет с признаками на основе коэффициентов	2,267	0,031	8,23	2,87	-0,028
Прочность при растяжении					
Исходный датасет, полученный путем удаления всех одномерных выбросов	391,23	0,17	238112,7	487,97	-0,024
Нормализованный датасет, полученный путем удаления всех одномерных выбросов	0,159	0,58	0,039	0,198	-0,024

Полученные метрики для модели линейной регрессии оказались достаточно низкими. Для лучшей из полученных моделей коэффициент детерминации равен примерно 0, то есть модель предсказывает значение просто на уровне среднего.

С учетом достаточно близких значений метрик, целесообразность сравнения данной модели на других датасетах, полученных на стадии предобработки данных, отсутствует.

Следующей моделью для предсказания целевых переменных была выбран метод ближайших соседей. Также с учетом полученных распределений переменных, высокой точности от данной модели не ожидается.

Для модели был выбран параметр «Количество ближайших соседей», равный 5, а также для всех точек задан одинаковый вес.

Модель была обучена на необработанном датасете, полученном путем удаления всех одномерных выбросов, на нормализованном датасете, а также на датасете с отдельно выбранными параметрами.

Сводная таблица полученных результатов приведена в таблице 4.

Таблица 4 – Сводная таблица полученных метрик модели KNN-регрессии

	MAE	MAPE	MSE	RMSE	R^2
Модуль упругости при растяжении					
Исходный датасет, полученный путем удаления всех одномерных выбросов	2,583	0,035	10,72	3,27	-0,339
Нормализованный датасет, полученный путем удаления всех одномерных выбросов	0,17	1,0	0,043	0,21	-0,132
Исходный датасет с признаками на основе тепловой карты корреляции	2,453	0,033	9,82	3,13	-0,226
Прочность при растяжении					
Исходный датасет, полученный путем удаления всех одномерных выбросов	414,21	0,18	265670,5	515,43	-0,14

Метрики всех моделей хуже, чем у модели линейной регрессии, изменение количества ближайших соседей на полученные результаты не влияет.

Метрики по оставшимся моделям приведены в таблице 5.

Таблица 5 – Сводная таблица полученных метрик моделей регрессии Дерево решений, Градиентный бустинг, Случайный лес

	MAE	MAPE	MSE	RMSE	R^2
Модуль упругости при растяжении					
Дерево решений Исходный датасет, полученный путем удаления всех одномерных выбросов	2,52	0,034	9,83	3,13	-0,227
Градиентный бустинг Исходный датасет, полученный путем удаления всех одномерных выбросов	2,39	0,03	9,01	3,01	-0,13
Градиентный бустинг Нормализованный датасет, полученный путем удаления всех одномерных выбросов	0,15	0,56	0,036	0,19	-0,125
Случайный лес Исходный датасет с признаками на основе тепловой карты корреляции	2,27	0,031	8,38	2,9	-0,05
Прочность при растяжении					
Дерево решений Исходный датасет, полученный путем удаления всех одномерных выбросов	420,26	0,18	281805,9	530,85	-0,212

Ни одна из обученных моделей не дала результатов, пригодных для дальнейшего использования моделей, каждая из моделей «угадывает» значение целевой переменной.

Учитывая, что ни одна из моделей не дала нормальных результатов, необходимо вернуться к стадии разведочного анализа данных и предобработки данных.

Датасет был разделен на два по целевым переменным, а также были проведены стандартизация и нормализация полученных датасетов.

Чтобы выявить зависимости целевых переменных было решено провести кластеризацию датасетов для осуществления дальнейшей работы с каждым из кластеров.

Так как для кластеризации важен масштаб, все работа проводилась на нормализованных датасетах.

Для кластеризации был применен метод K-means, методом «локтя» было определено, что оптимальное количество кластеров и для датасета с Модулем упругости при растяжении и для датасета с Прочностью при растяжении является 2 кластера.

Далее работа осуществлялась с датасетом с Модулем упругости при растяжении.

После применения метода K-means датасет был разделен на 2 с количеством объектов 520 и 503 соответственно.

Для оценки целесообразности разделения на кластеры были обучены модели линейной регрессии, регрессии на основе дерева решений и случайный лес на первом кластере. Полученные результаты приведены в таблице 6.

Таблица 6 – Сводная таблица полученных метрик после кластеризации

	MAE	MAPE	MSE	RMSE	R^2
Модуль упругости при растяжении					
Линейная регрессия	0,14	0,42	0,03	0,18	-0,053
Дерево решений	0,16	0,44	0,04	0,21	-0,043
Случайный лес	0,14	0,42	0,033	0,18	-0,114

Полученные метрики аналогичны метрикам ранее обученных моделей, таким образом кластеризация не дала результата.

С учетом большого количества признаков, моделям сложно построить зависимости. Перед обучением моделей попробуем тактику уменьшения размерности.

Для этого используем метод главных компонент, обучение будем проводить на стандартизованном датасете. Количество признаков определим как 4.

Для оценки целесообразности уменьшения размерности были также как и при кластеризации обучены модели линейной регрессии, регрессии на основе дерева решений и случайный лес. Полученные результаты приведены в таблице 7.

Таблица 7 – Сводная таблица полученных метрик после уменьшения размерности

	MAE	MAPE	MSE	RMSE	R^2
Модуль упругости при растяжении					
Линейная регрессия	0,79	1,05	0,99	0,99	-0,008
Дерево решений	0,83	1,87	1,06	1,03	-0,1
Случайный лес	0,81	1,4	1,05	1,03	-0,09

Полученные метрики хуже метрик ранее обученных моделей, таким образом уменьшение размерности также не дало результата.

Попробуем оценить датасет визуально. При изучении датасета можно отметить, что первые 23 строки содержат большое количество целочисленных и повторяющихся значений. Можно предположить, что данные параметры получены расчетным путем. Попробуем взять первые 23 строки за расчетные и построить модели на них.

Тепловая карта корреляции для датасета из первых 23 строк представлена на рисунке 9.

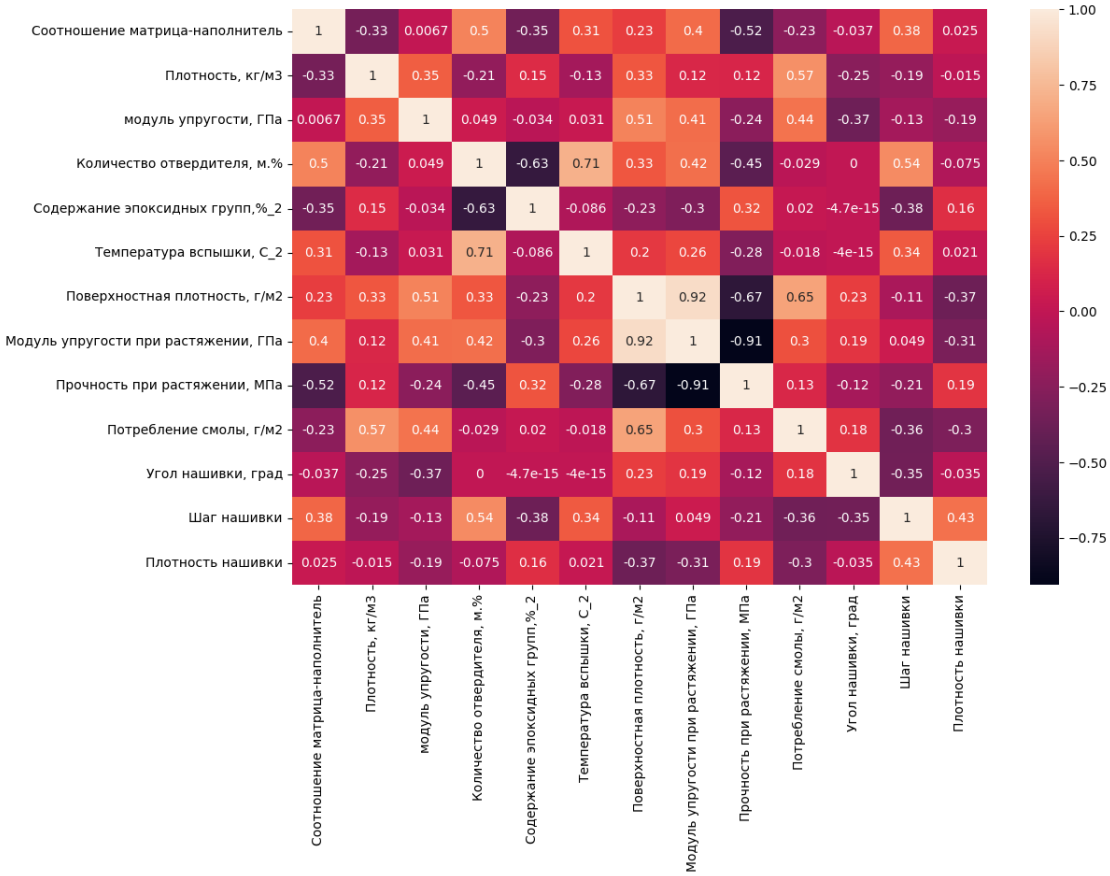

Рисунок 9 – Тепловая карта корреляции датасета из первых 23 строк

Тепловая карта корреляции для датасета из оставшихся строк представлена на рисунке 10.

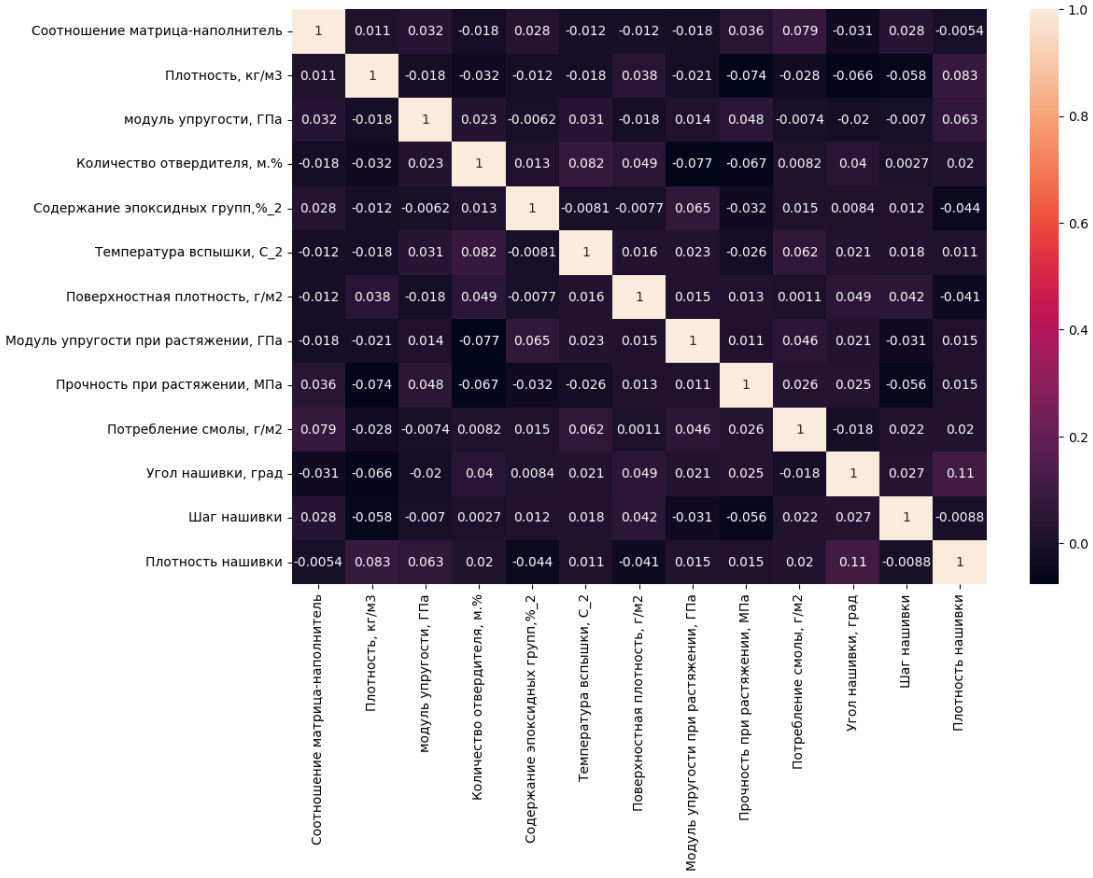

Рисунок 10 – Тепловая карта корреляции датасета из оставшихся строк

Как видно из построенных тепловых карт корреляции, для первого датасета есть явные зависимости, для второго датасета зависимости отсутствуют.

Обучим все ранее рассматриваемые модели на датасете из первых 23 строк. Так как данных мало, будем использовать метод кросс-валидации Leave-One-Out. В этом методе на каждой итерации в качестве тестового набора используется ровно один объект, а все остальные — для обучения.

Сравнительные результаты работы моделей приведены в таблице 8.

Таблица 8 – Сводная таблица полученных метрик моделей для выборки из 23 строк

	MAE	MAPE	MSE	RMSE	R^2
Модуль упругости при растяжении					
Линейная регрессия	0,0	0,0	0,0	0,0	1,0
KNN-регрессия	1,28	0,017	3,44	1,86	0,64

	MAE	MAPE	MSE	RMSE	R^2
KNN-регрессия. Нормализация	0,17	-	0,04	0,21	0,7
KNN-регрессия. Подбор гиперпараметров	0,05	-	0,019	0,14	0,87
Дерево решений	0,0	0,0	0,0	0,0	1,0
Градиентный бустинг	0,25	0,003	0,47	0,69	0,95
Случайный лес	0,33	0,004	0,29	0,53	0,97
Прочность при растяжении					
Линейная регрессия	0,0	0,0	0,0	0,0	1
KNN-регрессия	251,98	0,11	146229,9	382,4	0,38
KNN-регрессия. Нормализация	0,2	-	0,055	0,24	0,66
KNN-регрессия. Подбор гиперпараметров	0,04	0,13	0,01	0,1	0,94
Дерево решений	0,0	0,0	0,0	0,0	1,0
Градиентный бустинг	15,16	0,007	1830,18	42,78	0,99
Случайный лес	103,3	0,05	19172,8	138,47	0,92

Все модели, кроме KNN-регрессии дали высоких результат, поэтому при работе с данным методом был проведен подбор гиперпараметров, после которого модель значительно улучшила свои результаты.

2.3. Тестирование модели

Все модели на ограниченной выборке из 23 строк дали неплохие результаты с применением метода кросс-валидации Leave-One-Out.

Применим все эти модели, используя выборку из 23 строк как обучающую, а оставшуюся часть – как тестовую.

Сводные результаты тестирования моделей приведены в таблице 9.

Таблица 9 – Сводная таблица итоговых метрик моделей

	MAE	MAPE	MSE	RMSE	R^2
Модуль упругости при растяжении					
Линейная регрессия	4,07	0,06	25,7	5,01	-1,64
KNN-регрессия. Нормализация	0,32	-	0,15	0,39	-4,37
Дерево решений	3,36	0,05	17,38	4,17	-0,79
Градиентный бустинг	3,26	0,04	16,13	4,01	-0,66
Случайный лес	3,2	0,04	15,4	3,92	-0,58
Прочность при растяжении					
Линейная регрессия	729,18	0,31	832440,2	912,38	-2,53
KNN-регрессия. Нормализация	0,32	-	0,15	0,39	-4,06
Дерево решений	525,92	0,23	425619,3	652,4	-0,81
Градиентный бустинг	509,9	0,22	399590,1	632,13	-0,7
Случайный лес	505,94	0,21	394156,2	627,8	-0,67

Для оценки полученных результатов более подробно построим графики распределения предсказаний относительно реальных значений и распределения ошибок.

Для модели линейной регрессии графики представлены на рисунке 11.

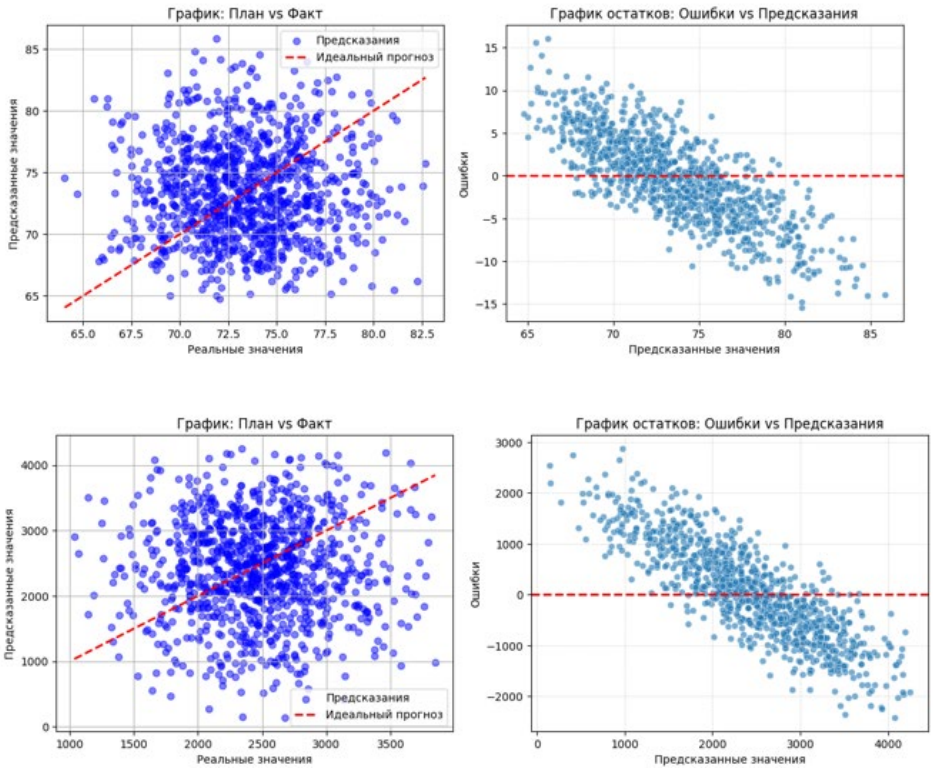

Рисунок 11 – График оценки результата модели линейной регрессии

Исходя из полученных графиков предсказанные значения распределены хаотично, то есть модель «угадывает» целевую переменную, анализ второго графика показывает, что модель завышает маленькие значения и занижает большие. Это подтверждает то, что модель предсказывает среднее значение.

Для модели KNN-регрессии графики представлены на рисунке 12.

Исходя из полученных графиков, модель предсказывает среднее исходя из ближайших кластеров, модель предсказывает одно и тоже значение для диапазона значений.

Для модели регрессии на основе дерева решений графики представлены на рисунке 13.

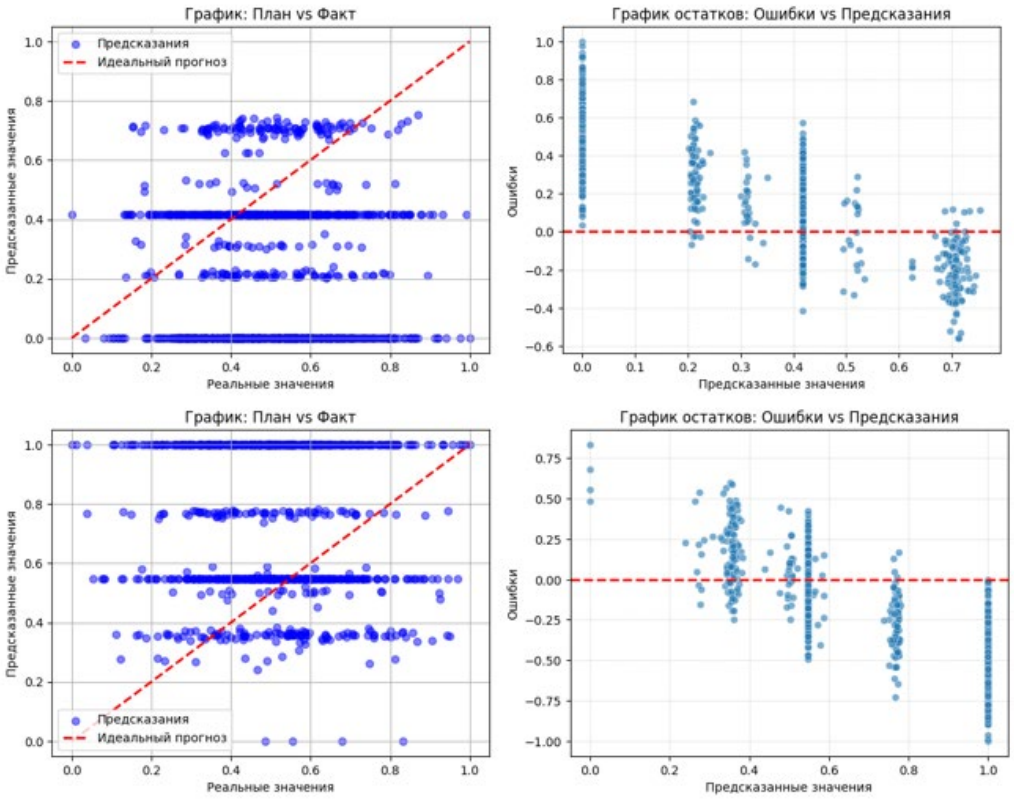

Рисунок 12 – График оценки результата модели KNN-регрессии

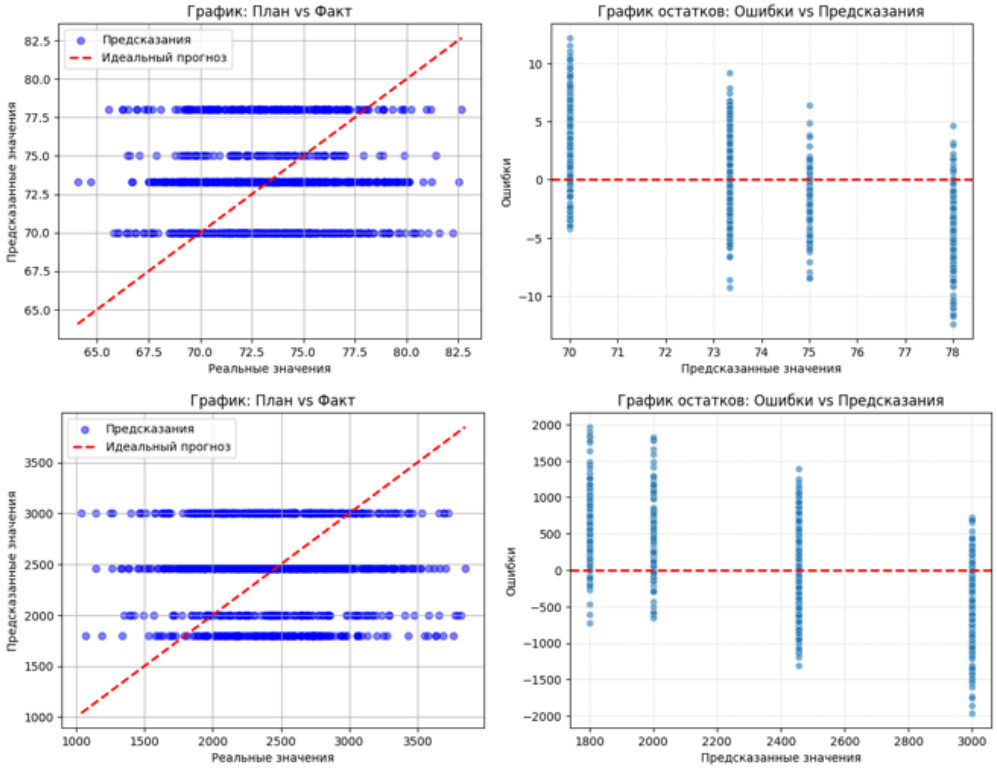

Рисунок 13 – График оценки результата модели регрессии на основе дерева решений

Исходя из графиков, дерево решений также строит дискретные предсказания, при этом выделяя небольшое количество ступеней, значит модель достаточно сильно обобщает данные.

Для модели градиентный бустинг графики представлены на рисунке 14.

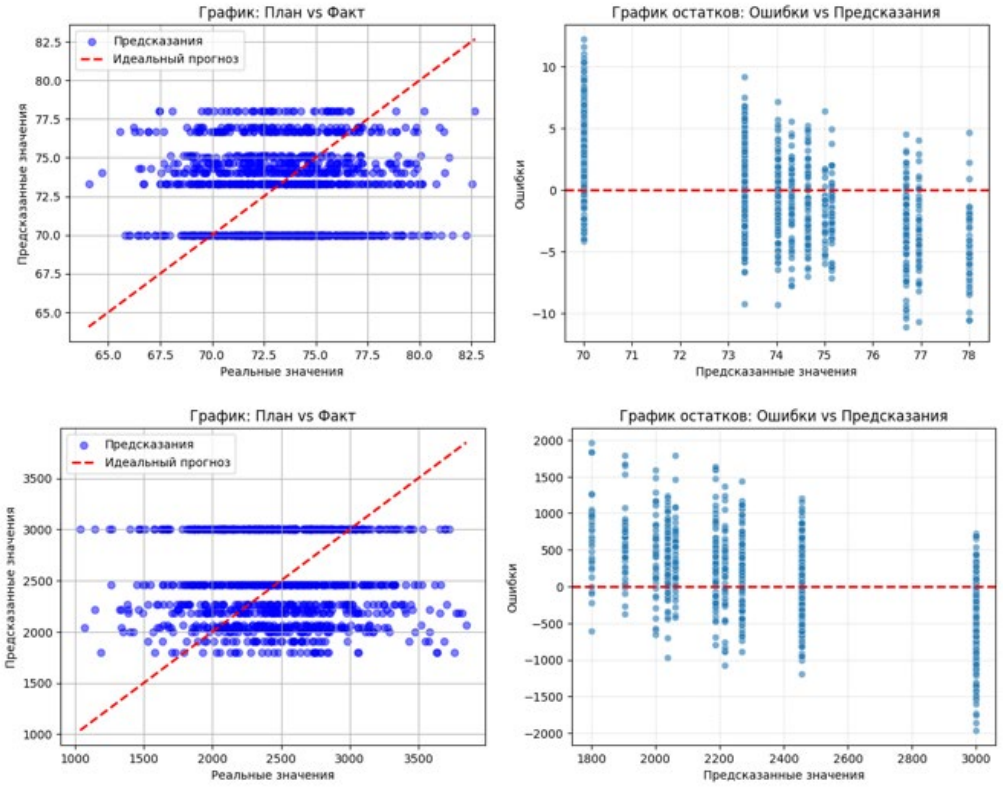

Рисунок 14 – График оценки результата модели регрессии градиентный бустинг

Данные графики имеют аналогичную проблему, как и у одиночного дерева решений, модель усредняет данные, разбивая их на классы и предсказывая единое значение внутри диапазонов.

Для модели случайный лес графики представлены на рисунке 15.

На данных графиках можно увидеть, что наблюдается четкое разделение графика. Можно предположить, что для ряда значений набор одиночных деревьев принимает одинаковые решения. При этом в оставшихся частях графиков значения разбросаны хаотично, то есть модель также не выявила очевидных зависимостей.

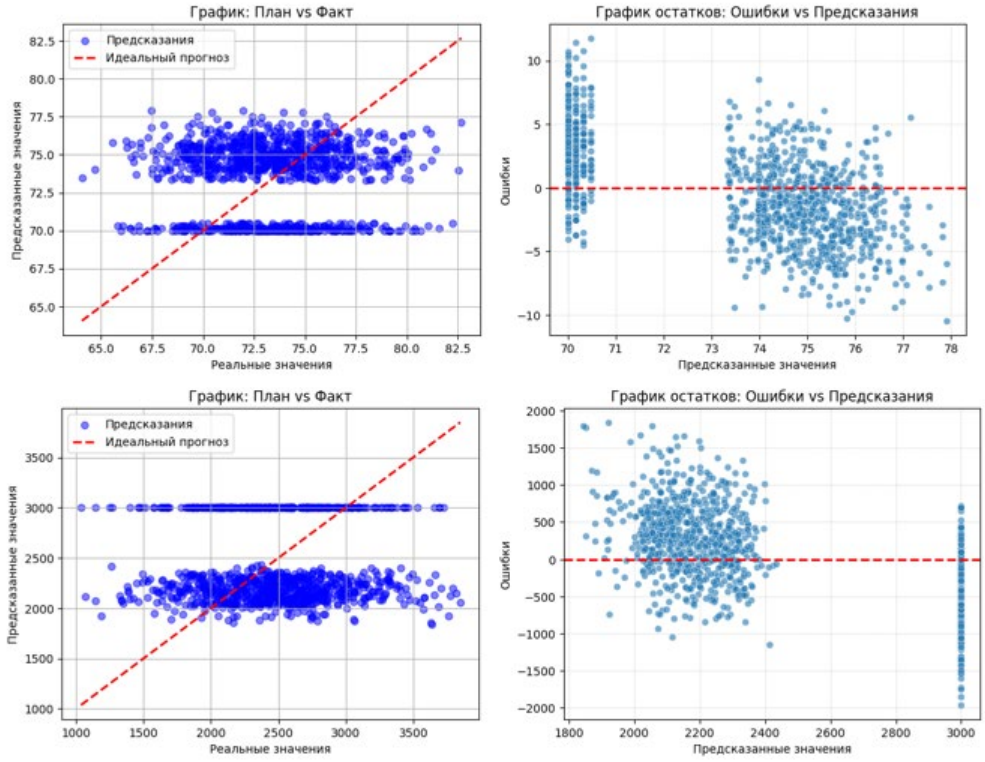

Рисунок 15 – График оценки результата модели регрессии случайный лес

Исходя из вышеизложенного, все модели показали низкие значения метрик, в связи с чем можно сделать вывод, что ни одна из рассматриваемых моделей не может быть применима для расчета Модуля упругости при растяжении и Прочности при растяжении.

2.4. Нейронная сеть по рекомендации соотношения матрица-наполнитель

Так как для нейронной сети важно масштабирование, будем использовать нормализованный датасет после удаления выбросов с использованием IsalationForest.

Для построения нейронной сети будем применять библиотеку keras. Для первоначальной оценки работы нейронной сети зададим следующие параметры:

1 скрытый слой с 32 нейронами;

алгоритм оптимизации – Adaptive Moment Estimation;

функция потерь – среднеквадратичная ошибка;

метрика – средняя абсолютная ошибка;

количество эпох – 100;

пакет из 32 строк.

На рисунке 16 представлены графики оценки нейронной сети, из которых видно «недообучение» модели. Модель предсказывает среднее значение, не подходя к экстремумам, чтобы не ошибиться.

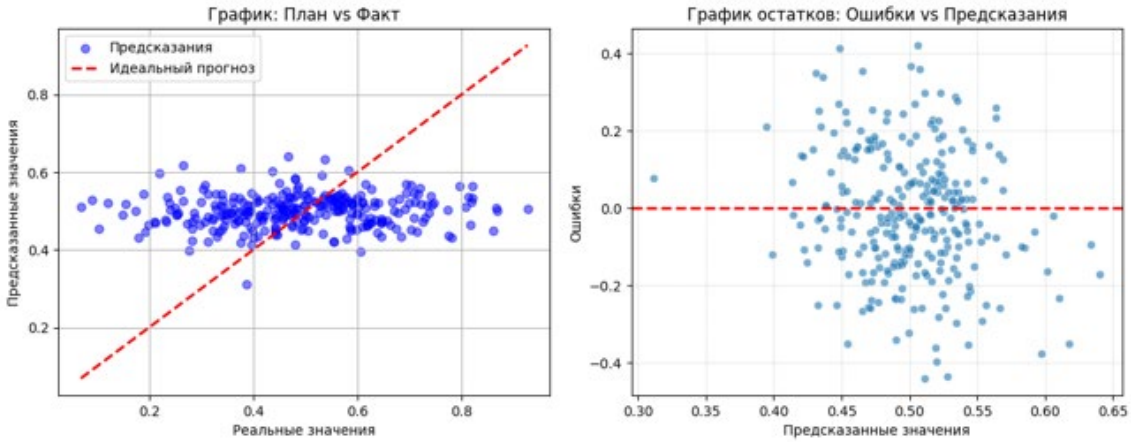

Рисунок 15 – График оценки результата нейронной сети 1

Изменим количество нейронов в нейронной сети. Зададим 2 скрытых слоя, в первом 64 нейрона, во втором 32. Рассмотрим графики оценки результата для новой нейросети (рисунок 16).

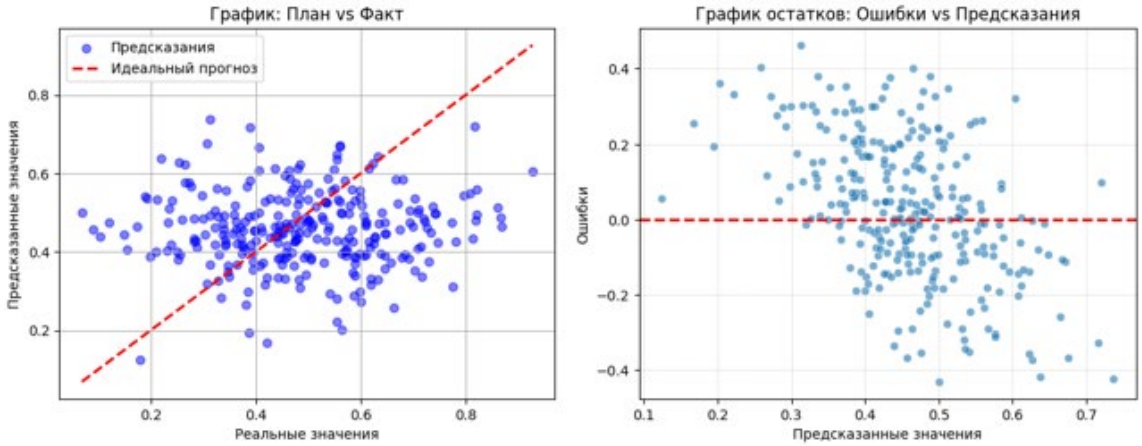

Рисунок 16 – График оценки результата нейронной сети 2

Модель стала завышать маленькие значения и занижать большие.

Вернемся к модели с 1 скрытым слоем в 32 нейрона, заменим алгоритм оптимизации на градиентный спуск. Рассмотрим графики оценки результата для нейросети (рисунок 17).

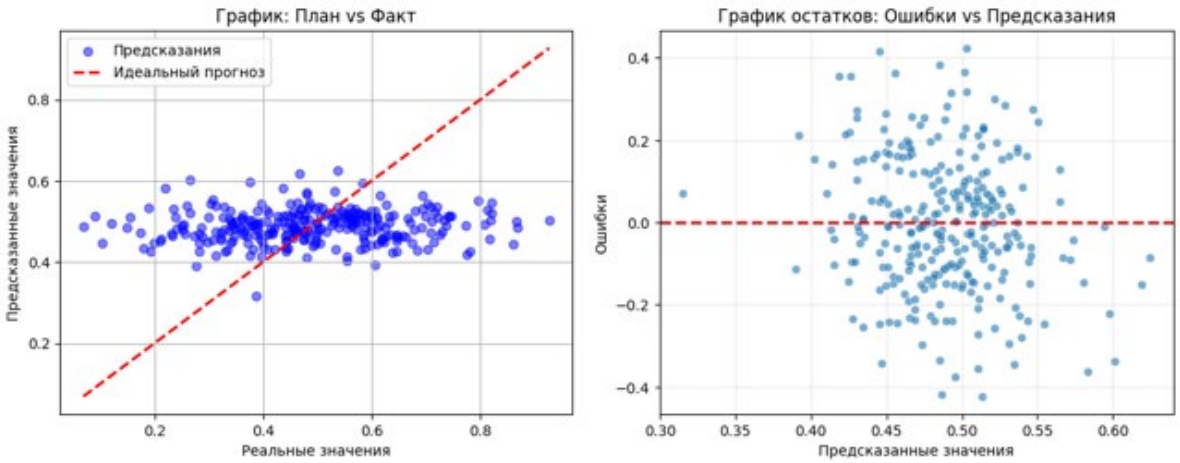

Рисунок 17 – График оценки результата нейронной сети 3

Модель, как и прежде, предсказывает среднее значение.

Попробуем построить нейронную сеть в 2 скрытых слоя с 32 и 16 нейронами, алгоритм оптимизации – градиентный спуск. Результаты работы модели приведены на рисунке 18.

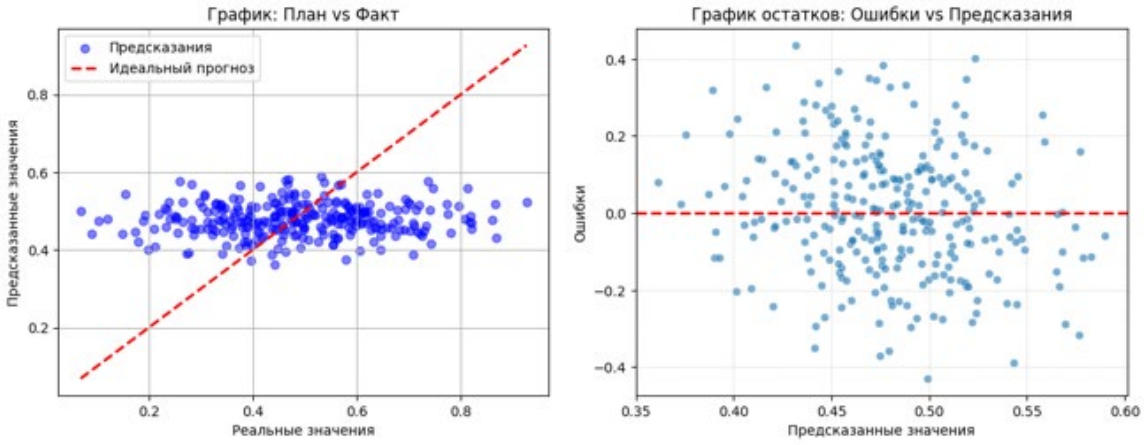

Рисунок 18 – График оценки результата нейронной сети 4

Модель, как и прежде, предсказывает среднее значение.

Попробуем построить нейронную сеть для датасета, состоящего из 23 первых строк. Как и в классическом машинном обучении будем использовать кросс-валидацию Leave-One-Out.

Зададим следующие параметры:

1 скрытый слов в 16 нейронов;

алгоритм оптимизации - Adaptive Moment Estimation;

функция потерь – среднеквадратичная ошибка;

метрика – средняя абсолютная ошибка;

количество эпох – 100;

пакет из 4 строк.

Коэффициент детерминации полученной модели отрицательный, модель работает не стабильно, средняя ошибка превышает дисперсию.

Для улучшения нейронной сети попробуем исключить лишние признаки. На основе тепловой карты корреляции оставим только признаки, коэффициент корреляции с целевой переменной выше 0,1, таким образом останется 9 признаков.

Так как улучшений метрик не произошло, сократим еще количество признаков до 5, убрав те признаки, коэффициент корреляции с целевой переменной у которых меньше 0,35.

Произошло незначительное улучшение метрик. Попробуем оценить работу нейронной сети с таким количеством метрик. Обучим нейронную сеть на датасете из 23 строк, взяв за тестовую выборку оставшуюся часть датасета.

Оценим полученный результат на рисунке 19.

Как видно из графиков, нейронной сети недостаточно 23 строк для обучения. Данные искажаются достаточно сильно, применение такой нейронной сети невозможно. Все предсказания завышены относительно реальных. Модель ошибается в среднем в 3 раза.

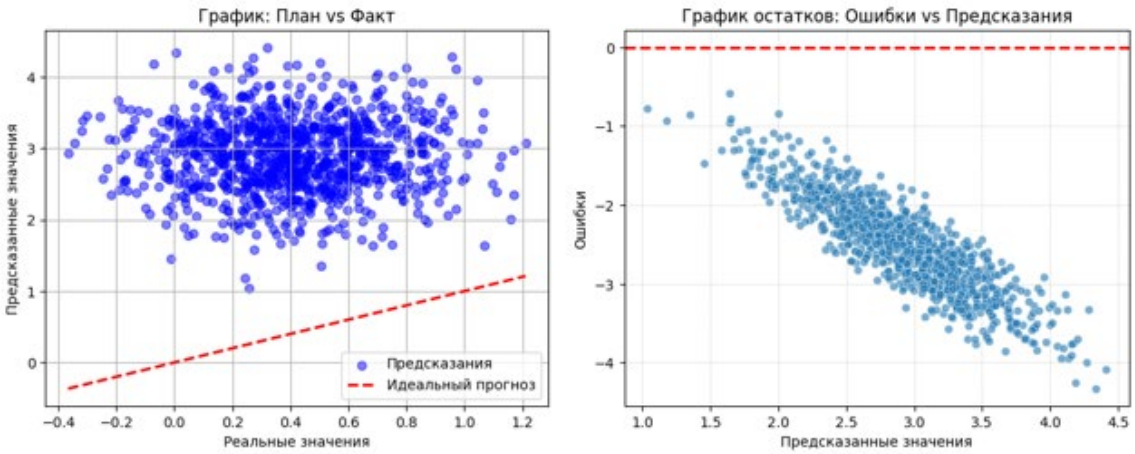

Рисунок 18 – График оценки результата нейронной сети

Сводные результаты тестирования нейронных сетей приведены в таблице 10.

Таблица 10 – Сводная таблица метрик нейронных сетей

	MAE	MAPE	MSE	RMSE	R^2
1 слой, 32 нейрона, adam	0,13	0,37	0,02	0,16	-0,03
2 слоя, 64 и 32 нейрона, adam	0,14	0,39	0,03	0,19	-0,35
1 слой, 32 нейрона, SGD	0,13	0,36	0,026	0,16	-0,02
2 слоя, 32 и 16 нейронов, SGD	0,13	0,36	0,027	0,16	-0,04
loo, 12 признаков	0,78	0,3	1,01	1,01	-0,3
loo, 9 признаков	0,78	0,31	1,00	1,00	-0,29
loo, 5 признаков	0,81	0,31	0,97	0,98	-0,24
train-23	2,5	28, 6	6,57	2,56	-84,8

Применение ни одной из обученных нейронных сетей для рекомендации соотношения матрица-наполнитель не представляется возможным.

2.5. Разработка приложения

Для разработки приложения была выбрана модель регрессии Случайный лес. После обучения модели на датасете из 23 строк (на расчетных данных), данная модель была сохранена и загружена в код приложения.

Приложение позволяет на основе параметров композиционного материала рассчитать значение Модуля упругости при растяжении.

Для расчета необходимо заполнить все данные, соответствующие признакам исходного датасета, на выходе программа дает предсказанное значение Модуля упругости при растяжении.

Пример работы приложения приведен на рисунках 19, 20.

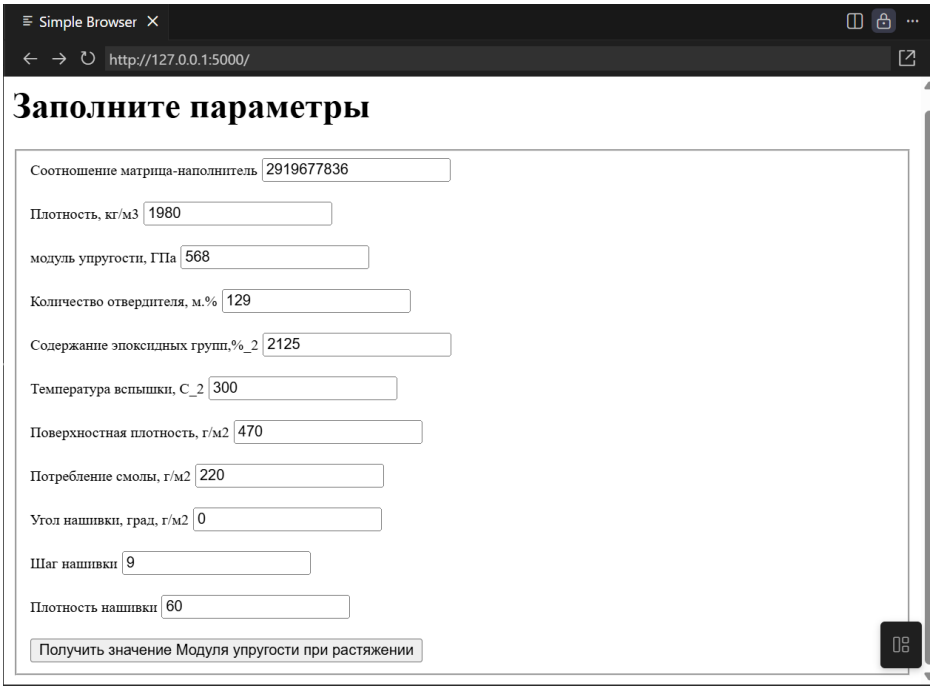

Рисунок 19 – Интерфейс страницы приложения по расчету Модуля упругости при растяжении

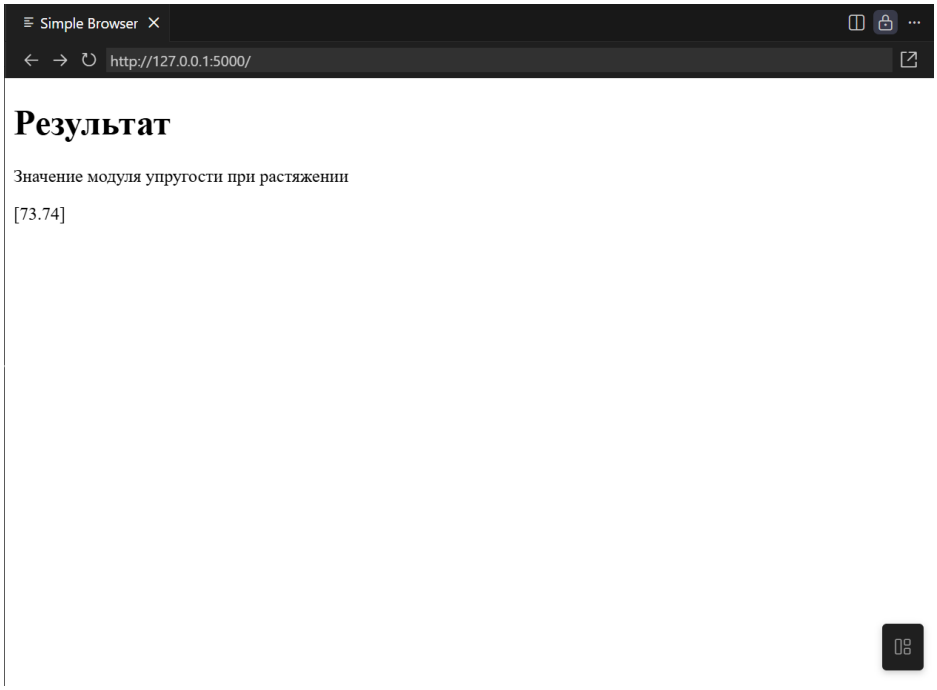

Рисунок 19 – Интерфейс результирующей страницы приложения по выводу результата Модуля упругости при растяжении

2.6. Создание удаленного репозитория и загрузка результатов работы на него

Страница на github.com:

https://github.com/Elena2410ell.

Страница репозитория:

https://github.com/Elena2410ell/Proect.2026_02_15.

Заключение

В рамках работы была проведен и анализ датасета, содержащего свойства композиционных материалов, обучение моделей машинного обучения по предсказанию Модуля упругости при растяжении и Прочности при растяжении, построена нейронная сеть по рекомендации соотношения матрица – наполнитель.

По результатам проведенной работы было отмечено четкое разделение датасета на две части: выборка, содержащая большое количество целочисленных значений, для которых прослеживалась явная линейная зависимость целевых переменных от признаков, и выборка, где практически отсутствуют взаимосвязи между переменными.

На основании гипотезы, что первая выборка является расчетной, для обучения моделей были использованы именно она. При обучении и тестировании моделей с использованием кросс-валидации модели показали высокую точность предсказания, однако при тестировании таких моделей на оставшейся выборке, точность модели упала до очень низких значений.

Таким образом, применение данных моделей для предсказания значений Модуля упругости при растяжении и Прочности при растяжении не целесообразно. Все модели прогнозируют средние значения целевых переменных.

Аналогичная ситуация наблюдается при построении нейронной сети. Из-за отсутствия явных зависимостей между переменными, нейронной сети сложно их выявить и рекомендовать точные значения. Ни одна из построенных моделей нейронной сети не дала высокой точности.

Таким образом, применение ее для рекомендации соотношения матрица-наполнитель не целесообразно.

Возможным путем решения проблемы низкой точности моделей будет дополнения выборки, а также удаление шумов из второй ее части, где отсутствуют явные зависимости между переменными, увеличение выборки за счет большего количества расчетных значений.

Библиографический список

1. ГОСТ Р ИСО 16269-4-2017. Статистические методы. Статистическое представление данных. Часть 4. Выявление и обработка выбросов. – Введ. 2017-08-10. – М. : Стандартинформ, 2017. – 53 с.

2. Жерон, О. Прикладное машинное обучение с помощью Scikit-Learn, Keras и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем – Москва: Диалектика, 2024. – 1056 с.

3. Шолле, Ф. Глубокое обучение на Python – Санкт-Петербург: Питер, 2024. – 576 с.

4. Вандер Плас, Дж. Python для сложных задач: наука о данных и машинное обучение – Санкт-Петербург: Питер, 2024. – 576 с.

5. Брюс, П., Брюс, Э., Гедек, П. Практическая статистика для специалистов Data Science – Санкт-Петербург: БХВ-Петербург, 2024. – 352 с.

6. Рашка, С., Мирджалили, В. Python и машинное обучение. Машинное и глубокое обучение с использованием Python, Scikit-Learn и TensorFlow 2 – Москва: Диалектика, 2024. – 848 с.

7. Мхитарян, В. С. Анализ данных: учебник для вузов – Москва: Издательство Юрайт, 2024. – 490 с.

8. Клементьев, И. П. Введение в методы анализа данных и машинного обучения – Москва: Альпина Паблишер, 2024. – 280 с.

9. Нилд, Т. Математика для data science. Управляем данными с помощью линейной алгебры, теории вероятностей и статистики – Астана: «Спринт Бук», 2025. – 352 с.

10. Грас, Д. Data science. Наука о данных с нуля: Пер. с англ. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2025. – 416 с.