МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет им Н.Э. Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу

«Data Science Pro»

Слушатель Нурукулиев Егор Игоревич

Москва, 2025

Содержание

Введение

Целью данной работы является создание методов прогнозирования характеристик новых композитов. Композитные материалы представляют собой сложные системы, состоящие из нескольких компонентов, которые четко разграничены между собой, сохраняя свои индивидуальные свойства. Такие материалы обладают неоднородной структурой, что позволяет достигать уникальных эксплуатационных характеристик, недостижимых для каждого компонента по отдельности.

Основу композитов составляет матрица — базовый элемент, в который добавляются армирующие компоненты, называемые наполнителями. Наполнитель распределяется внутри матрицы равномерно или в определённой ориентации, образуя сложную пространственную структуру. Эта структура определяет механические и физические свойства материала.

Ключевая особенность композитов заключается в их адаптируемости. Изменяя долю наполнителя, его геометрию, пространственное распределение и характеристики межфазной границы, можно получать материалы с заданными свойствами. Это делает их востребованными в высокотехнологичных отраслях:

- авиация и космическая техника;
- металлургия и добывающая промышленность;
- автомобилестроение и машиностроение;
- электроника и ядерная энергетика;
- медицина и строительство.

Высокая стоимость разработки композитов обусловлена сложностью их состава и необходимости многочисленных испытаний для получения требуемых характеристик. Поскольку предсказать свойства композитов на основе исходных компонентов затруднительно, производителям приходится проводить экспериментальные исследования, что увеличивает затраты времени и ресурсов.

Использование методов машинного обучения позволяет ускорить процесс создания новых материалов. Системы прогнозирования, основанные на анализе данных, могут минимизировать количество физических экспериментов, одновременно повышая точность и скорость проектирования. Это делает предложенный подход крайне актуальным в условиях растущего спроса на композитные материалы с уникальными свойствами.

1 Аналитическая часть

1.1 Постановка задачи

Целью данной работы является исследование композитного материала, состоящего из базальтопластиковой матрицы и углепластиковых нашивок. Основной задачей является разработка моделей машинного обучения для прогнозирования конечных свойств данного материала на основе входных данных. Кроме того, предполагается создание удобного приложения, которое позволит использовать построенные модели специалистам в предметной области.

Исходные данные: для решения задачи был предоставлен датасет, содержащий информацию о свойствах матрицы и наполнителя, производственных параметрах и характеристиках готового композита. Данные представлены в виде двух отдельных файлов:

1. Файл X_bp (свойства базальтопластика):
   - Количество признаков: 10 (плюс индекс);
   - Объем данных: 1023 строки.
2. Файл X_nup (свойства углепластика):
   - Количество признаков: 3 (плюс индекс);
   - Объем данных: 1040 строк.

Файлы были объединены по индексу с использованием операции INNER JOIN. В результате объединения часть строк из файла X_nup была исключена, поскольку для них отсутствовали соответствующие данные в файле X_bp. Итоговый датасет содержит:

- Количество признаков: 13 (плюс индекс);
- Объем данных: 1023 строки.

Характеристики датасета:

- Тип данных: Все признаки, кроме одного, представлены в формате float64 (вещественные числа), и один признак в формате int64.

- Пропуски: Отсутствуют.

Типы признаков:

- Все признаки, за исключением "Угол нашивки", являются количественными и принимают непрерывные значения.

- Признак "Угол нашивки" принимает два значения и является категориальным.

Объединенный датасет представляет собой достаточно чистый набор данных, не требующий предварительной обработки, связанной с устранением пропусков или преобразованием типов. Эти данные станут основой для построения прогнозных моделей. Характеристики датасета указаны в таблице 1.

Таблица 1 - Характеристики датасета.

№	Параметр	Количество строк	Тип данных
1	2	3	4
1	Соотношение матрица-наполнитель	1023	float64
2	Плотность, кг/м³	1023	float64
3	модуль упругости, ГПа	1023	float64
4	Количество отвердителя, м.%	1023	float64
5	Содержание эпоксидных групп,%_2	1023	float64
6	Температура вспышки, С_2	1023	float64
7	Поверхностная плотность, г/м²	1023	float64
8	Модуль упругости при растяжении, ГПа	1023	float64
9	Прочность при растяжении, МПа	1023	float64
10	Потребление смолы, г/м²	1023	float64
11	Угол нашивки, град	1023	int64
12	Шаг нашивки	1023	float64
13	Плотность нашивки	1023	float64

На основе статистических данных, которые приведены на рисунке 1, по столбцам таблицы можно сделать следующие выводы:

1. Соотношение матрица-наполнитель:

- Среднее значение составляет 2.93, с отклонением (стандартное отклонение) 0.91.

- Минимальное значение 0.39, максимальное — 5.59, что указывает на широкое распределение значений.

- Большинство значений сосредоточено между 2.32 и 3.55.

2. Плотность, кг/м³:

- Среднее значение плотности — 1975.73 кг/м³, с небольшим стандартным отклонением (73.73).

- Плотность варьируется от 1731.76 до 2207.77 кг/м³, с 50% значений, лежащих в диапазоне от 1924.16 до 1977.62 кг/м³.

3. Модуль упругости, ГПа:

- Среднее значение модуля упругости — 739.92 ГПа с очень высоким стандартным отклонением (330.23 ГПа).

- Диапазон значений от 2.44 до 1911.54 ГПа, что свидетельствует о большом разнообразии в материалах или измерениях.

- 50% значений сосредоточено между 500.05 и 961.81 ГПа.

- Количество отвердителя, м.%:

- Среднее количество отвердителя — 110.57%, с отклонением 28.30%.

- Диапазон значений от 17.74% до 198.95%, что может свидетельствовать о различной концентрации отвердителя в материалах.

4. Содержание эпоксидных групп, %:

- Среднее значение — 22.24%, с отклонением 2.41%.

- Значения варьируются от 14.25% до 33%, что говорит о разнообразии химического состава.

5. Температура вспышки, °С:

- Среднее значение температуры вспышки — 285.88°C, с отклонением 40.94°C.

- Диапазон температур от 100°C до 413.27°C, что свидетельствует о возможном широком спектре химических веществ или различных условиях тестирования.

6. Поверхностная плотность, г/м²:

- Среднее значение поверхностной плотности — 482.73 г/м², но есть значительное отклонение (281.31 г/м²).

- Большинство значений сосредоточено в диапазоне от 266.82 до 693.23 г/м².

7. Модуль упругости при растяжении, ГПа:

- Среднее значение — 73.33 ГПа, с отклонением 3.12 ГПа.

- Диапазон значений от 64.05 до 82.68 ГПа, что говорит о сравнительно стабильных материалах в этом параметре.

8. Прочность при растяжении, МПа:

- Среднее значение прочности — 2466.92 МПа, с отклонением 485.63 МПа.

- Прочность варьируется от 1036.86 МПа до 3848.44 МПа, что указывает на значительные различия в прочности материалов.

9. Потребление смолы, г/м²:

- Среднее значение потребления смолы — 218.42 г/м², с отклонением 59.74 г/м².

- Потребление смолы варьируется от 33.80 г/м² до 414.59 г/м².

10. Угол нашивки, град:

- Среднее значение — 44.25°, с высоким стандартным отклонением (45.02°).

- Углы нашивки варьируются от 0° до 90°, при этом 50% значений имеют угол 0°.

11. Шаг нашивки:

- Среднее значение шага нашивки — 6.90, с отклонением 2.56.

- Шаг нашивки варьируется от 0 до 14.44, что может свидетельствовать о разных типах тканей или изделий.

12. Плотность нашивки:

- Среднее значение плотности нашивки — 57.15 г/м², с отклонением 12.35.

- Плотность варьируется от 0 до 103.99 г/м², что также указывает на разнообразие в характеристиках материалов.

Рисунок 1 – Статистические данные датасета

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.930366	0.913222	0.389403	2.317887	2.906878	3.552660	5.591742
Плотность, кг/м3	1023.0	1975.734888	73.729231	1731.764635	1924.155467	1977.621657	2021.374375	2207.773481
модуль упругости, ГПа	1023.0	739.923233	330.231581	2.436909	500.047452	739.664328	961.812526	1911.536477
Количество отвердителя, м.%	1023.0	110.570769	28.295911	17.740275	92.443497	110.564840	129.730366	198.953207
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301	14.254985	20.608034	22.230744	23.961934	33.000000
Температура вспышки, С_2	1023.0	285.882151	40.943260	100.000000	259.066528	285.896812	313.002106	413.273418
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690	0.603740	266.816645	451.864365	693.225017	1399.542362
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983	64.054061	71.245018	73.268805	75.356612	82.682051
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006	1036.856605	2135.850448	2459.524526	2767.193119	3848.436732
Потребление смолы, г/м2	1023.0	218.423144	59.735931	33.803026	179.627520	219.198882	257.481724	414.590628
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1023.0	6.899222	2.563467	0.000000	5.080033	6.916144	8.586293	14.440522
Плотность нашивки	1023.0	57.153929	12.350969	0.000000	49.799212	57.341920	64.944961	103.988901

Общие выводы:

Данные имеют широкий разброс, особенно для таких параметров, как модуль упругости, прочность при растяжении и плотность нашивки.

Для большинства параметров характерна нормальная или близкая к нормальной дисперсия, что может свидетельствовать о стабильности данных с небольшими вариациями.

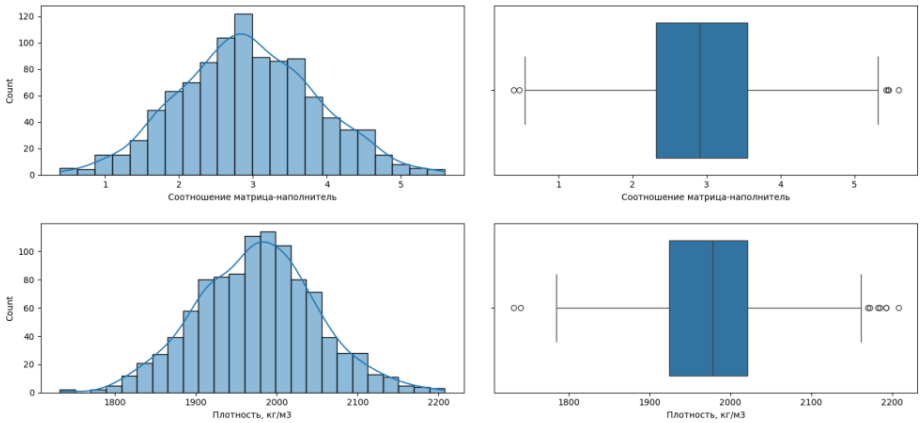
Некоторые параметры (например, "Угол нашивки") могут требовать дополнительной проверки на наличие аномальных значений (например, нулевые углы или слишком большие значения)

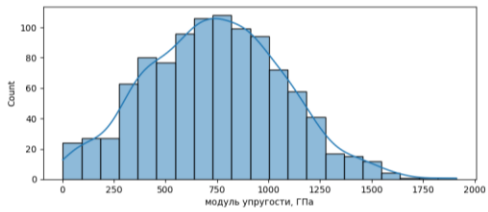
Некоторые параметры демонстрируют широкий диапазон значений, такие как модуль упругости и поверхностная плотность, что может свидетельствовать о гетерогенности материалов. В то же время в данных могут присутствовать выбросы, особенно в параметрах, связанных с температурой вспышки, модулем упругости и поверхностной плотностью.

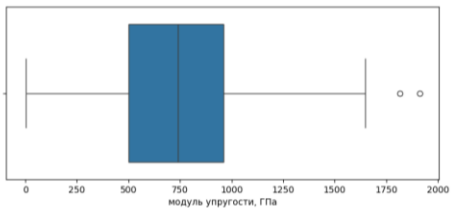
Параметры, такие как модуль упругости при растяжении и прочность при растяжении, имеют относительно узкие диапазоны значений и, вероятно, могут быть хорошо предсказаны с использованием методов машинного обучения.

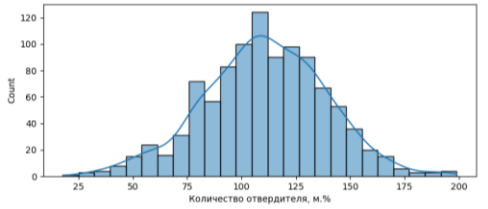
Для более детального анализа распределений и выявления возможных выбросов выполнена визуализация данных, включающая построение гистограмм и диаграмм типа «ящик с усами». Эти графики, представленные на рисунках 2–4, показали, что все признаки, кроме «Угол нашивки», имеют распределение, близкое к нормальному, и принимают только неотрицательные значения. Признак «Угол нашивки» является категориальным и принимает строго два значения: 0 и 90.

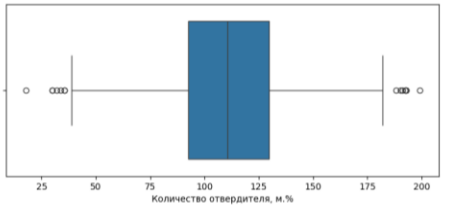
Рисунок 2 – Графики распределения всех типов столбцов

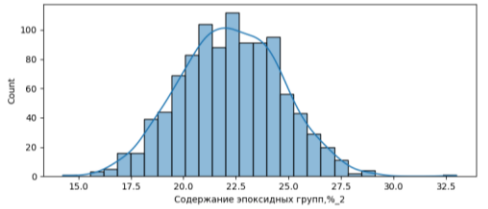


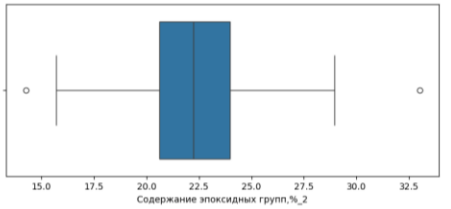

Count
модуль упругости, ГПа

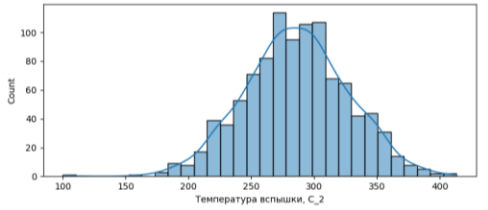

модуль упругости, ГПа

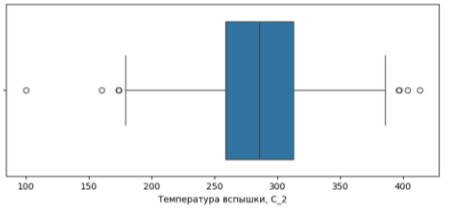

Count
Количество отвердителя, м.%

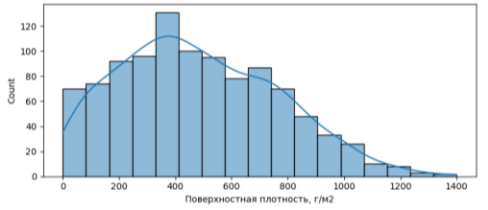

Количество отвердителя, м.%

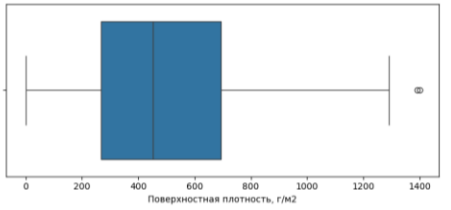

Count
Содержание эпоксидных групп,%_2

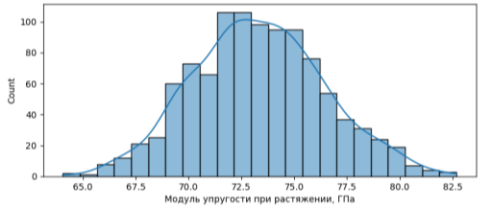

Содержание эпоксидных групп,%_2

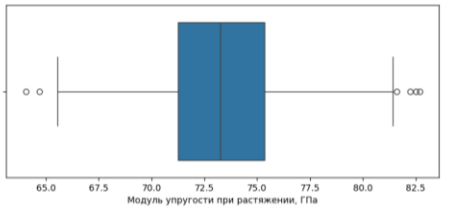

Count
Температура вспышки, С_2

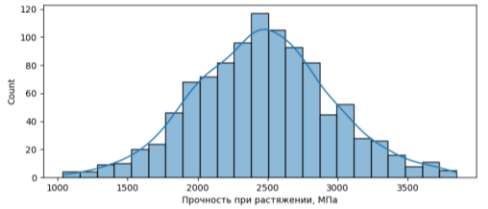

Температура вспышки, С_2

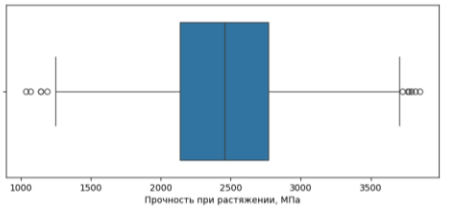

Count
Поверхностная плотность, г/м2

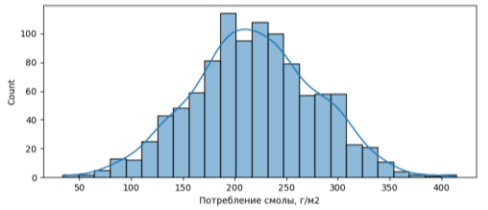

Поверхностная плотность, г/м2

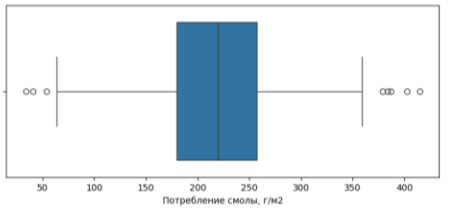

Count
Модуль упругости при растяжении, ГПа

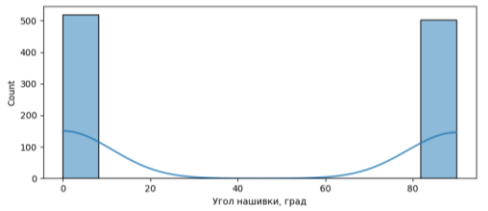

Модуль упругости при растяжении, ГПа

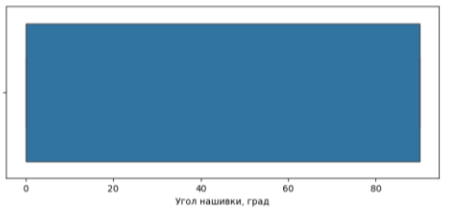

Count
Прочность при растяжении, МПа

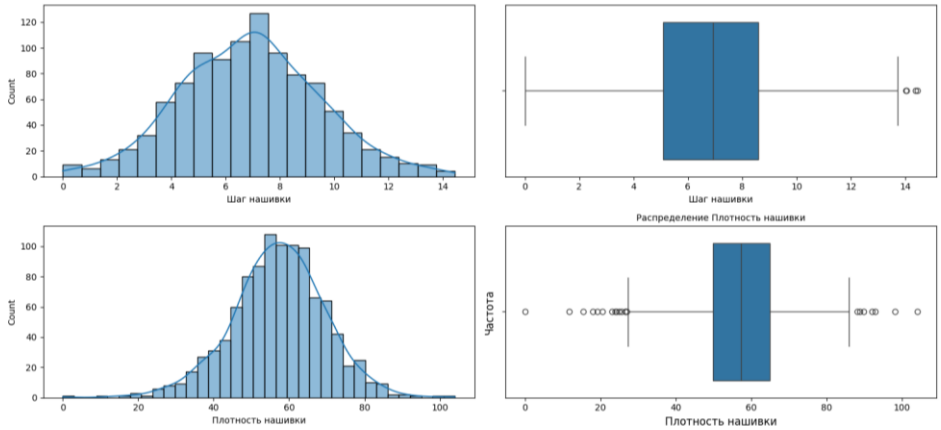

Прочность при растяжении, МПа


Count
Потребление смолы, г/м2


Потребление смолы, г/м2


Count
Угол нашивки, град

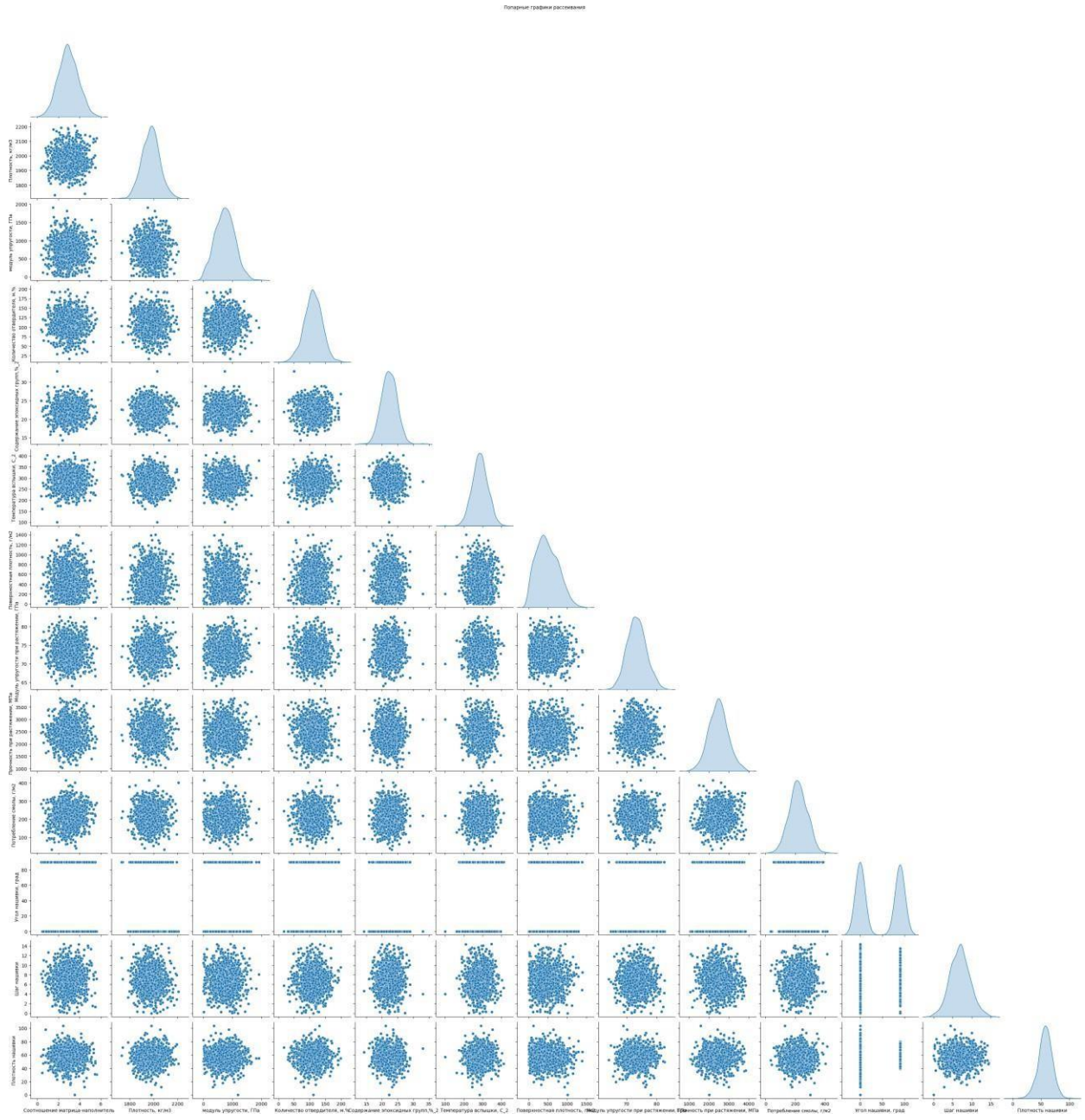

Угол нашивки, град



Датасет был предварительно обработан, что объясняет отсутствие пропусков. В исходных необработанных данных обычно встречаются пропуски и значения некорректных типов.

Попарные графики рассеяния точек, приведенные на рисунке 3, демонстрируют взаимосвязи между признаками. На этих графиках видно, что некоторые точки значительно удалены от основного облака данных. Такие точки являются потенциальными выбросами — аномальными или некорректными значениями, которые выходят за пределы допустимых диапазонов признаков. По графику видно, что некоторые точки стоят далеко от общего скопления.

Рисунок 3 – Попарный график рассеивания



Попробуем найти выбросы и подобрать для этой цели метод.

Метод IQR (межквартильного размаха) устойчив к выбросам, подходит для данных с неравномерным распределением. Не учитывает общую структуру распределения. Хорош, когда данные асимметричны.

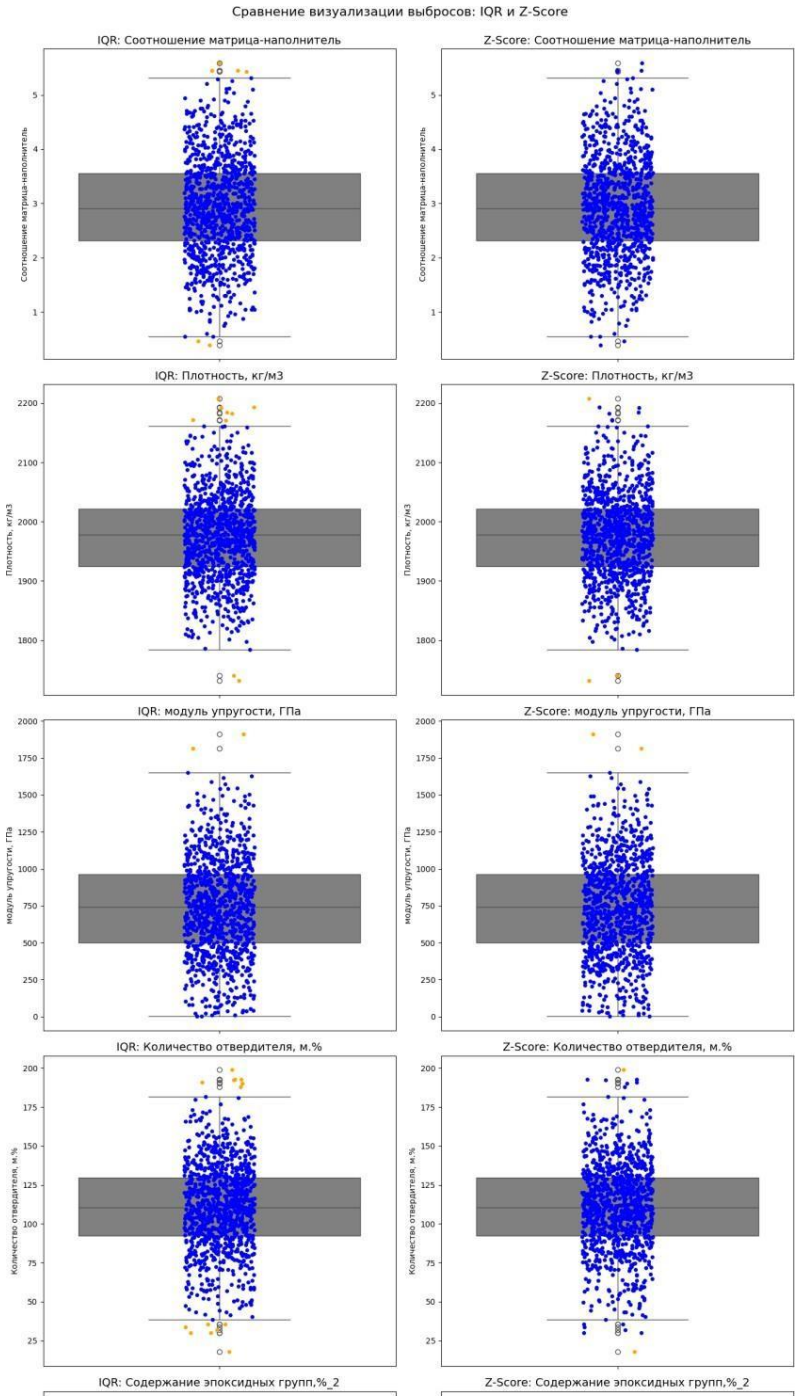
Метод Z-Score - основан на статистике выборки. Чувствителен к выбросам, плохо работает с асимметричными данными. Хорош, когда данные нормальны.

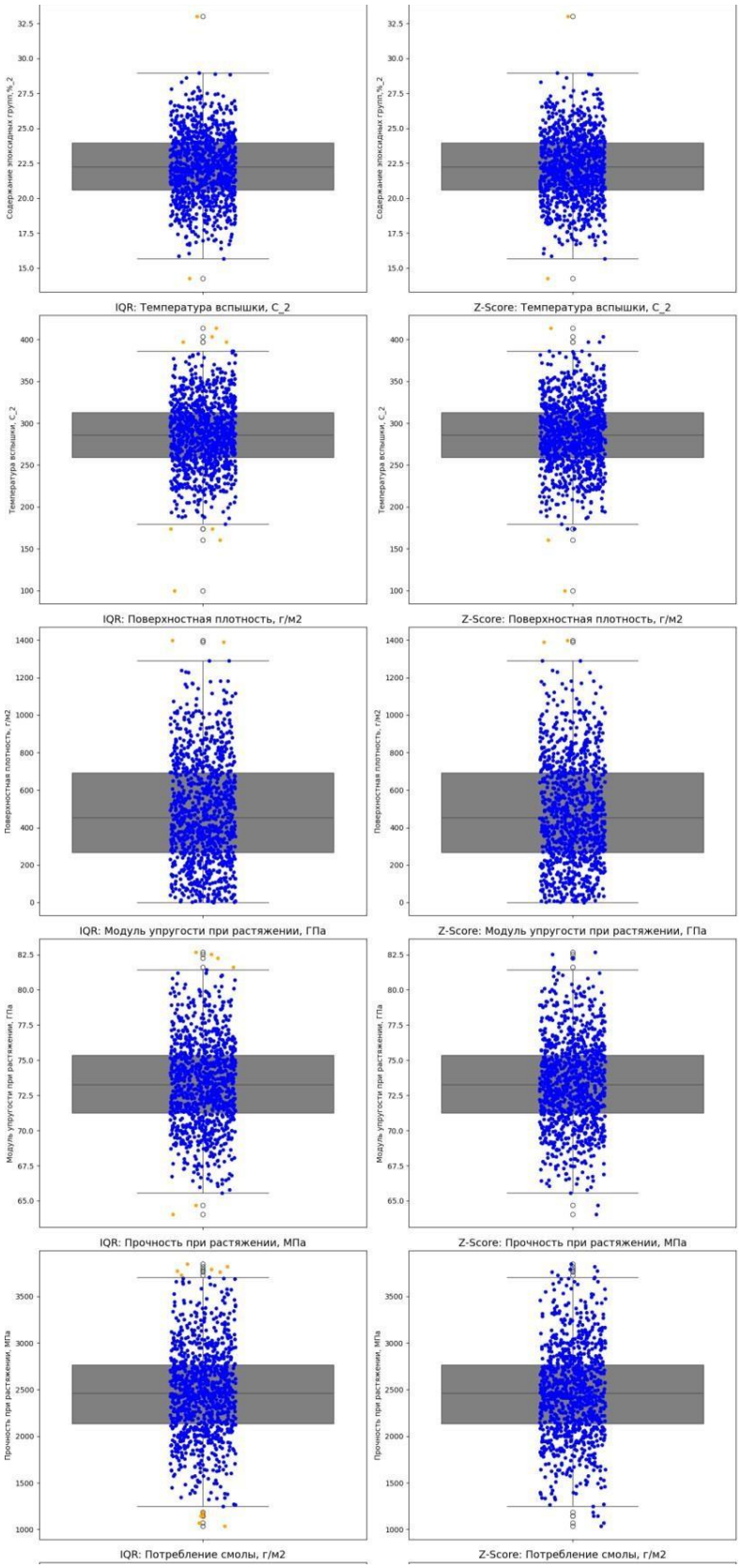
Метод DBSCAN - считывает многомерные данные. Сложность настройки параметров. Для сложных, многомерных данных.

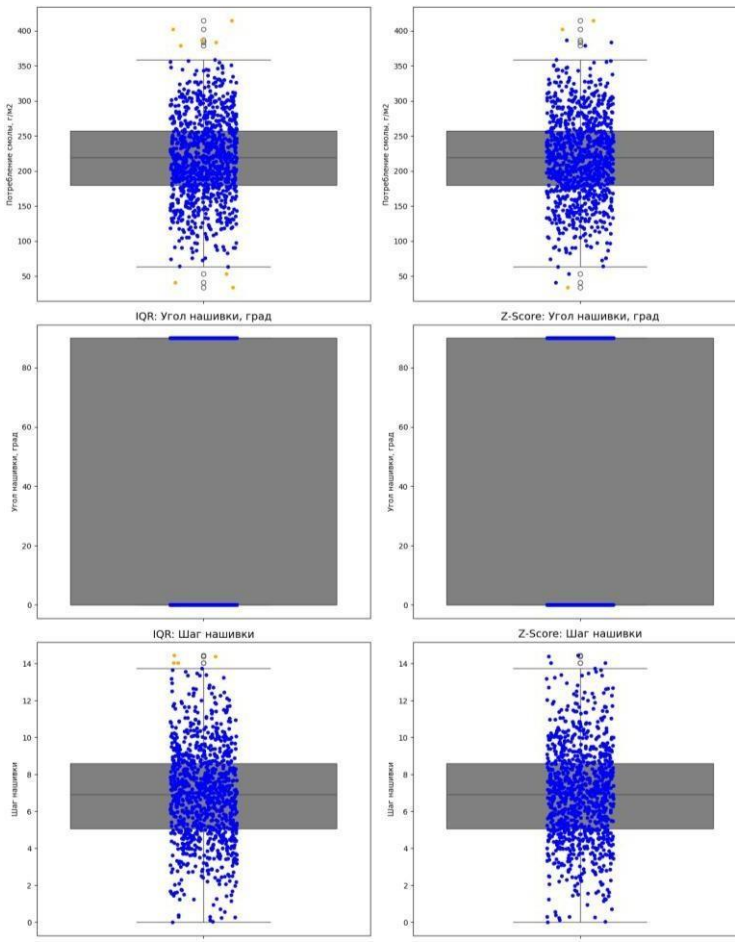
Метод пороговых значений - Требует ручной настройки, подходит не для всех данных.

Мы используем метод Z-Score, так как наши данные нормальны, но сравним их с методом IQR. График представлен на рисунке 4.

Рисунок 4 – Визуализация выбросов



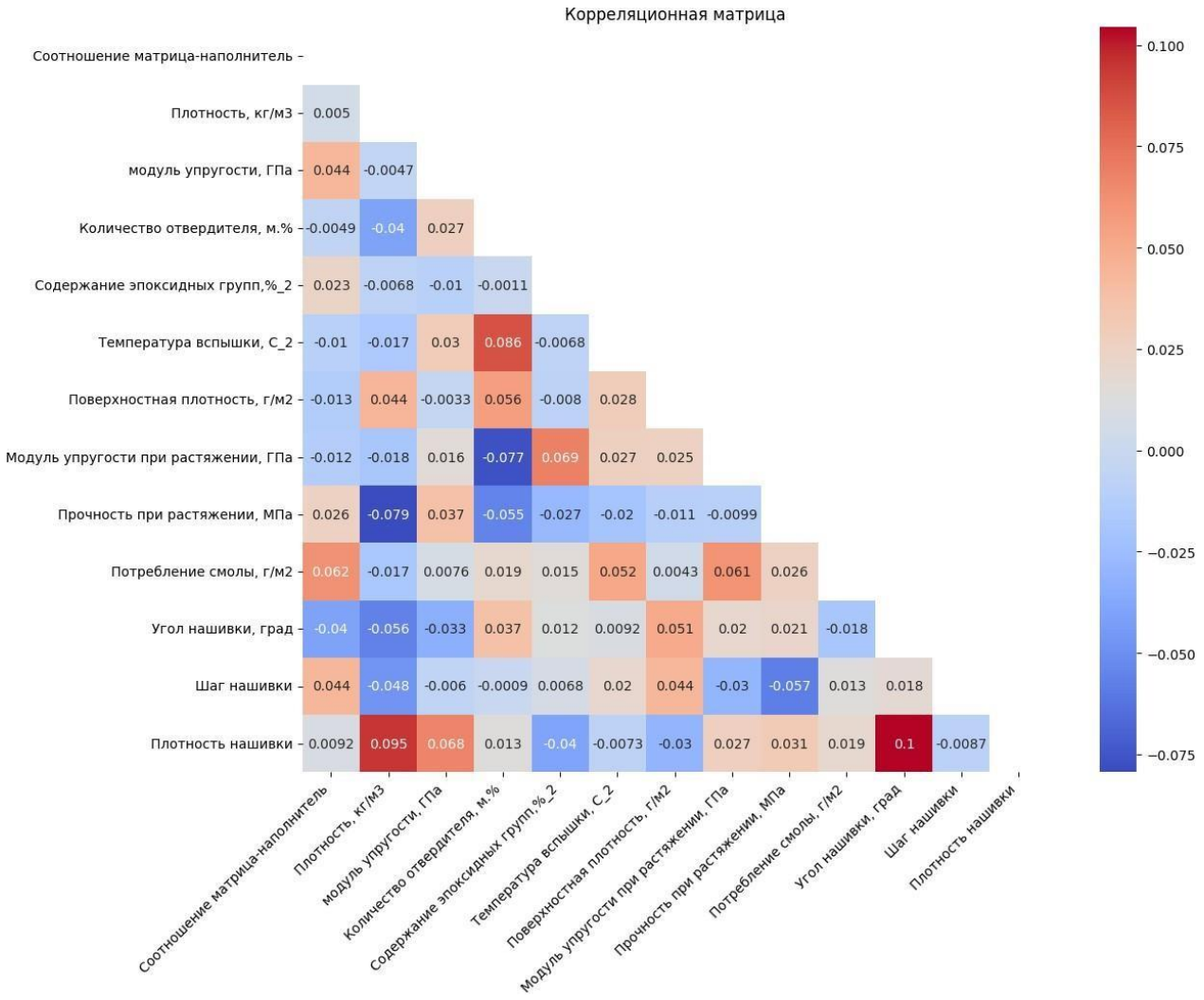

IQR: Температура вспышки, С_2
Z-Score: Температура вспышки, С_2
IQR: Поверхностная плотность, г/м2
Z-Score: Поверхностная плотность, г/м2
IQR: Модуль упругости при растяжении, ГПа
Z-Score: Модуль упругости при растяжении, ГПа
IQR: Прочность при растяжении, МПа
Z-Score: Прочность при растяжении, МПа
IQR: Потребление смолы, г/м2
Z-Score: Потребление смолы, г/м2



Построим корреляционную матрицу, которая приведена на рисунке 5. Она помогает найти взаимосвязи между переменными. Анализ взаимосвязей помогает понять, насколько сильно и в каком направлении связаны переменные. Это важно при построении моделей, чтобы избежать избыточности данных, которые дают высокие корреляции между независимыми переменными и могут ухудшить работу моделей, особенно линейных. Если две переменные сильно коррелируют, одну из них можно исключить из модели для упрощения анализа. Также, матрица может выявить интересные взаимосвязи, которые требуют дополнительного изучения.

Но необходимо учитывать, что коэффициенты корреляции показывают только линейные зависимости. Если связь нелинейная, она может быть не выявлена, а также корреляция не указывает на причинно-следственные связи. Высокая корреляция не означает, что одна переменная влияет на другую.

Рисунок 5 – Корреляционная матрица в виде тепловой карты



Из тепловой карты мы видим, что максимальная корреляция между плотностью нашивки и углом нашивки равна 0,1, корреляция между остальными параметрами близка к нулю. Это значит, что линейные зависимости между ними отсутствуют. И линейная модель регрессии не даст приемлемого результата.

1.2 Описание используемых методов

Поскольку линейные зависимости мы не нашли, мы рассмотрим другие методы регрессии. Прогнозирование значений непрерывной числовой переменной относится к задаче регрессии, зависимая переменная связана с одной или несколькими независимыми переменными, которые также называют

предикторами или регрессорами. Регрессионный анализ позволяет изучить, как среднее значение зависимой переменной изменяется в зависимости от изменений независимых переменных.

Нормализуем данные методом MinMaxScaler — это метод масштабирования данных, который преобразует значения признаков таким образом, чтобы они попадали в заданный диапазон, обычно от нуля до единицы. Далее рассмотрим выбранные модели машинного обучения.

1.2.1 Ансамбль из множества деревьев решений

RandomForestRegressor — это метод машинного обучения, который использует ансамбль из множества деревьев для решения задачи регрессии (предсказания числовых значений). Каждый из этих деревьев делает своё предсказание, а итоговый результат вычисляется как среднее значение всех деревьев.

Проще говоря:

1. Данные делятся на множество небольших частей.
2. Для каждой части строится своё дерево решений.
3. Все деревья "голосуют" за предсказание, и берётся их среднее значение.

Метод работает следующим образом. Данные рандомно делятся на обучающие наборы для каждого дерева. Каждое дерево строится независимо от других. Деревья обучаются на случайных подмножествах признаков, чтобы избежать переобучения и сделать модель более устойчивой. На этапе предсказания все деревья делают свои прогнозы, а результат усредняется.

Рассмотрим плюсы метода:

1. Высокая точность: метод часто работает лучше, чем отдельное дерево решений, так как сглаживает ошибки отдельных деревьев.
2. Устойчивость к переобучению: благодаря усреднению, модель менее склонна запоминать шумы в данных.

3. Обработка нелинейных зависимостей: хорошо работает даже с сложными данными.

4. Может работать с разнородными данными: хорошо справляется как с числовыми, так и с категориальными признаками.

5. Простота настройки: не требует много тонкой настройки гиперпараметров.

6. Оценка важности признаков: можно увидеть, какие признаки вносят наибольший вклад в предсказание.

Рассмотрим минусы метода:

1. Медленная работа на больших данных: построение множества деревьев требует времени и ресурсов, особенно для больших наборов данных.

2. Сложность интерпретации: итоговый результат создаётся на основе множества деревьев, что делает модель менее интерпретируемой, чем простые методы (например, линейная регрессия).

3. Неэффективность на высокоразмерных данных: если у данных очень много признаков, модель может становиться менее точной из-за случайного выбора признаков для построения деревьев.

4. Большое потребление памяти: хранение большого количества деревьев требует значительных ресурсов.

Данный метод хорошо подходит для сложных задач регрессии, где линейные модели не дают хороших результатов, ситуаций, где важна устойчивость к переобучению и для анализа данных, где важна оценка важности признаков.

1.2.2 Метод K-ближайших соседей (K-Nearest Neighbors, KNN)

Это алгоритм машинного обучения, который используется для решения задач регрессии (предсказания числовых значений). Он основывается на идее, что похожие объекты должны иметь схожие результаты.

Принцип работы достаточно прост, для каждой новой точки, которую нужно предсказать, алгоритм находит K ближайших соседей в обучающих данных. После этого делает предсказание, усредняя значения целевой переменной этих соседей (или используя другие способы, такие как взвешенное среднее).

Плюсами данного метода являются простота, KNeighborsRegressor легко понять и реализовать, он не делает предположений о форме данных, таких как линейность или нормальность, алгоритм может работать с различными метриками расстояния (например, евклидово расстояние, манхэттенское расстояние).

У метода есть и минусы. Для больших данных или многомерных данных вычисления расстояний для каждого предсказания могут быть очень медленными. Результаты сильно зависят от выбранного числа соседей (K). Если K слишком маленькое, модель может быть чувствительна к шуму. Если K слишком большое, она может быть слишком сглаженной и упускать важные детали. В высоких измерениях (многомерные данные) становится сложнее найти «ближайших соседей», и метод может терять свою эффективность из-за так называемого парадокса измерения.

1.2.3 Метод Gradient Boosting

GradientBoostingRegressor — это мощный алгоритм машинного обучения для решения задач регрессии, основанный на деревьях решений. Он работает путем построения ансамбля слабых моделей (обычно деревьев решений), которые объединяются, чтобы дать сильное предсказание.

Рассмотрим принцип работы метода.

Алгоритм начинает с простой модели, например, предсказывая среднее значение целевой переменной. На каждом шаге алгоритм вычисляет ошибки (остатки) текущей модели. Строится новое дерево решений, которое учится предсказывать ошибки предыдущей модели. Новое дерево добавляется к

ансамблю, чтобы уменьшить ошибки. Этот процесс повторяется заданное число раз (или до достижения нужной точности). Итоговое предсказание — это сумма предсказаний всех деревьев с учетом их весов.

Плюсами метода являются высокая точность, благодаря своей способности хорошо улавливать сложные зависимости в данных. Он подходит для работы с различными типами данных и не требует их масштабирования. Использование регуляризации (например, ограничение глубины деревьев, темп обучения) позволяет избежать переобучения.

Минусы метода сводятся к построению большого числа деревьев, что может замедлять его работу, особенно на больших наборах данных. Результат обработки данных сильно зависит от таких параметров, как глубина деревьев, число деревьев, темп обучения. Самый большой минус модели в том, что он чувствителен к шуму и выбросам, поэтому данные перед использованием должны быть тщательно обработаны.

GradientBoostingRegressor подходит для задач, где требуется высокая точность, и есть время на настройку гиперпараметров. Если данные сложные или имеют нелинейные зависимости, этот метод может стать отличным выбором.

1.2.4 Метод с поддерживающими векторами (Support Vector Regression, SVR)

Это алгоритм машинного обучения, основанный на методе опорных векторов (SVM), который применяется для решения задач регрессии.

Принцип работы метода заключается в том, что SVR пытается найти гиперплоскость (в многомерном пространстве), которая максимально точно описывает зависимость между входными и выходными данными. При этом допускаются небольшие ошибки (в пределах заданного порога ε) для большей гибкости. Вместо минимизации ошибок для всех точек SVR фокусируется только на ключевых точках данных (опорных векторах), которые определяют положение гиперплоскости.

К плюсам метода можно отнести гибкость, он может моделировать как линейные, так и нелинейные зависимости (с помощью ядер, например, RBF или полиномиальных), робустность, т.е. устойчивость к выбросам, так как они не оказывают большого влияния на гиперплоскость, и контроль точности, так как метод настраиваемый и в нем можно задавать допустимую погрешность через параметр ε.

К минусам метода можно отнести высокую вычислительную сложность, так как он требует больших ресурсов на больших или высокоразмерных данных. Так же он чувствительность к гиперпараметрам несмотря на то, что в нем можно тщательно настраивать параметры, такие как C, ε, и тип ядра. Так же модель менее интуитивно понятна, чем простые методы.

SVR хорошо подходит для задач с небольшим числом признаков и выборкой, особенно если есть сложные зависимости между данными, и требуется высокая точность.

1.3 Разведочный анализ данных

Разведочный анализ данных представляет собой важный этап при работе с данными, поскольку его цель заключается в выявлении скрытых закономерностей и взаимосвязей внутри набора данных. Этот процесс помогает лучше понять структуру данных и определить, какие признаки могут оказывать влияние на результаты модели.

При построении большинства моделей машинного обучения требуется выполнение двух ключевых условий:

1. Зависимость выходных переменных от входных: чем сильнее эта зависимость, тем точнее модель сможет предсказывать результат на основе входных признаков.

2. Независимость между самими входными переменными: наличие сильной корреляции между входными признаками может привести к переобучению модели и снижению её точности.

Ранее, на примере графика попарного рассеяния точек (рисунок 3), мы могли наблюдать распределение данных. Визуальный анализ показал, что форма «облаков точек» не позволяет сразу заметить явные зависимости, которые были бы полезны для построения моделей. Однако визуализация — лишь первый шаг в анализе данных. Для выявления более тонких связей между признаками часто используется такой инструмент, как матрица корреляции.

Матрица корреляции, представленная ранее на рисунке 5, даёт количественную оценку степени линейной связи между различными признаками. Это позволяет оценить, насколько сильно одни признаки влияют на другие, а также обнаружить возможные мультиколлинеарности, когда несколько признаков оказываются тесно связаны друг с другом. Такой подход помогает сделать вывод о том, какие признаки стоит включить в модель, а какие, возможно, следует исключить для повышения качества прогнозирования.

Таким образом, разведочный анализ данных играет ключевую роль в подготовке данных для дальнейшего моделирования, позволяя исследователю получить представление об основных характеристиках и взаимосвязях в наборе данных.

1.3.1 Выбор признаков

Анализ корреляции показал, что все коэффициенты корреляции между признаками близки к нулю, что указывает на отсутствие линейной зависимости между ними. Это может свидетельствовать о том, что признаки не связаны напрямую или существуют сложные нелинейные зависимости, которые невозможно выявить простыми статистическими методами. Важно подчеркнуть, что не всегда низкая корреляция означает отсутствие взаимосвязи, так как могут быть скрытые зависимости, которые требуют более сложных подходов, например, методов машинного обучения. В контексте этого исследования мы можем предположить, что признаки делятся на несколько групп:

1. Свойства матрицы - характеристики материала, такие как модуль упругости, плотность и другие, которые определяют основные физико-химические свойства.

2. Свойства наполнителя - параметры, которые влияют на прочность и другие механические характеристики композита.

3. Свойства смеси и производственного процесса - параметры, связанные с технологией создания композита, такие как температура, время отверждения и другие условия.

4. Свойства готового композита - итоговые характеристики готового продукта, такие как прочность, эластичность и другие эксплуатационные качества.

Таким образом, группы признаков могут влиять на конечные характеристики композита, и каждый из этих факторов стоит учитывать при построении модели. Также важно отметить, что точное распределение признаков по этим категориям может требовать уточнения на основе более глубокого знания предметной области.

На основе зависимостей будут построены отдельные модели для каждого целевого признака, что позволяет решать три независимые задачи.

1.3.2 Препроцессинг

Препроцессинг данных играет ключевую роль в обеспечении корректной работы моделей. Важно отметить, что предварительная обработка должна проводиться после разделения данных на тренировочную и тестовую выборки, чтобы избежать утечек информации.

Категориальный признак в этом исследовании — "Угол нашивки, град", который принимает значения 0 и 90.

Для вещественных признаков возникает проблема различия в диапазонах значений, что может привести к искажению результатов при обучении модели.

Чтобы исправить это, применяются методы нормализации или стандартизации. В данном случае выбрана стандартизация, поскольку она более предпочтительна для алгоритмов, чувствительных к масштабу, таких как линейные модели или методы на основе градиентного спуска. Стандартизация приводит признаки к нулевому среднему и единичному стандартному отклонению, что помогает моделям работать более эффективно.

1.3.3 Поиск гиперпараметров по сетке

Поиск гиперпараметров по сетке (GridSearchCV) позволяет автоматически оптимизировать параметры модели, пробуя различные их комбинации и оценивая результат с помощью перекрестной проверки. Это помогает найти наилучшие параметры, которые улучшают производительность модели.

1.3.4 Метрики качества моделей

Для оценки эффективности моделей в задаче регрессии используются следующие метрики:

- R^2 (коэффициент детерминации): показывает, какая доля дисперсии целевой переменной объясняется моделью. Чем ближе значение R^2 к единице, тем лучше модель объясняет данные. Если R^2 близок к нулю, модель не объясняет данные и ее прогнозы хуже, чем простое предсказание среднего значения.

- MSE (Mean Squared Error) — это метод оценки ошибки модели, который вычисляет среднее значение квадратов разностей между реальными значениями и предсказанными моделью. Чем меньше MSE, тем точнее модель.

- MAE (средняя абсолютная ошибка): метрика, которая оценивает среднюю абсолютную разницу между предсказанными и истинными значениями. Она более устойчива к выбросам по сравнению с MSE.

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. Размерности полученных наборов данных показаны на рисунке 6. Описательная статистика входных признаков после предобработки показана на рисунке 7.

```
x1_train: (700, 12) y1_train: (700,)
x1_test: (300, 12) y1_test: (300,)
```

Рисунок 6 – размерности набора данных

	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
1	0.282131	0.601381	0.447061	0.123047	0.607435	0.482823	0.162230	0.319194	0.698235	0.517418	0.0	0.275109	0.544652
3	0.282131	0.601381	0.447061	0.608021	0.418887	0.549664	0.162230	0.319194	0.698235	0.517418	0.0	0.344539	0.365074
4	0.457857	0.601381	0.455721	0.502800	0.495653	0.482823	0.162230	0.319194	0.698235	0.517418	0.0	0.344539	0.503211
5	0.457201	0.527898	0.452685	0.502800	0.495653	0.482823	0.162230	0.319194	0.698235	0.517418	0.0	0.344539	0.544652
6	0.419084	0.307448	0.488508	0.502800	0.495653	0.482823	0.162230	0.319194	0.698235	0.517418	0.0	0.344539	0.682789
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1018	0.361750	0.410540	0.552781	0.350139	0.333908	0.657301	0.161609	0.485125	0.480312	0.242759	1.0	0.627565	0.365347
1019	0.587163	0.650588	0.268550	0.712271	0.294428	0.350746	0.271207	0.475992	0.470745	0.221717	1.0	0.730963	0.458327
1020	0.555750	0.460227	0.251612	0.494656	0.623085	0.325580	0.572959	0.573346	0.578340	0.565435	1.0	0.286298	0.650046
1021	0.637396	0.691520	0.448724	0.684130	0.267818	0.444436	0.496511	0.536217	0.368070	0.451281	1.0	0.435716	0.520631
1022	0.657131	0.259472	0.251903	0.609147	0.888354	0.553803	0.587373	0.550550	0.647135	0.444423	1.0	0.419448	0.785487

1000 rows × 13 columns

Рисунок 7 – Описательная статистика признаков

2.1.2 Для прогнозирования прочности при растяжении

Признаки датасета были разделены на входные и выходные, а строки – на тренировочное и тестовое множество. Размерности полученных наборов данных показаны на рисунке 8.

```
x2_train: (700, 12) y2_train: (700,)
x2_test: (300, 12) y2_test: (300,)
```

Рисунок 8 – размерности набора данных

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

Для подбора лучшей модели для этой задачи были подобраны модели, описанные в разделе 1.2.

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

	MAE	MSE	R2
Случайный лес (без Модуль упругости)	0.138948	0.030037	-0.105227
Случайный лес (без Прочности)	0.134053	0.028761	-0.054115
GradientBoostingRegressor (без Модуль упругости)	0.141053	0.031211	-0.148409
GradientBoostingRegressor (без Прочность)	0.137829	0.029994	-0.099316
SVR (без Модуль упругости)	0.154055	0.037427	-0.377129
SVR (без Прочность)	0.155312	0.037279	-0.366320
KNN (без Модуль упругости)	0.144251	0.032005	-0.177618
KNN (без Прочность)	0.144932	0.033238	-0.218219

Рисунок 9 – Сравнение моделей базовых моделей

Ни одна из выбранных мной моделей не оказалась подходящей для наших данных. Коэффициент детерминации R^2 близок к нулю. Значит, они не лучше базовой модели. И остальные метрики у них примерно совпадают с базовой моделью.

После выполнения подбора гиперпарметров по сетке с перекрестной проверкой, можно сделать вывод, что, подбирая гиперпараметры, можно значительно улучшить предсказание выбранной модели.

Все модели крайне плохо описывают исходные данные - не удалось добиться положительного значения R^2. Самая лучшая модель дает коэффициент детерминации близкий к нулю, что соответствует базовой модели.

2.3 Нейросеть, рекомендующаяв соотношение матрицы при помощи библиотеки TensorFlow

Мною была построена нейронная сеть, которая представляет собой простую последовательную модель с двумя скрытыми слоями и одним выходным слоем. Она предназначена для решения задачи регрессии, так как выходной слой имеет одну единицу без активации, а функция ошибки – среднеквадратичная ошибка (MSE).

Рассмотрим ее ключевые аспекты:

1. Архитектура модели

- Входной слой определяется автоматически через input_shape, который равен количеству признаков в обучающих данных.

- Скрытые слои представлены в виде двух полносвязанных слоя по 64 нейрона каждый с активацией ReLU. Это стандартные параметры для задач машинного обучения средней сложности. ReLU является одной из самых популярных активаций благодаря своей вычислительной эффективности и способности предотвращать проблему исчезающего градиента.

- Выходной слой – это дин нейрон без активации, поскольку модель решает задачу регрессии. Если бы это была классификация, то количество нейронов соответствовало бы числу классов, и использовалась бы соответствующая активация (например, softmax для многоклассовой классификации).

2. Параметры компиляции

- Оптимизатор используется Adam с фиксированным значением learning_rate=0.001. Adam – популярный оптимизатор, который хорошо справляется со стабилизацией процесса обучения и позволяет быстрее находить оптимальное решение. Значение learning_rate выбрано небольшое, чтобы избежать резких скачков при обучении и сделать процесс более плавным.

- Функция потерь – это среднеквадратическая ошибка (MSE) – стандартная метрика для оценки качества моделей регрессии. Она устойчива к выбросам и подходит для большинства случаев.

- Помимо MSE, добавлены две дополнительные метрики:

- MeanAbsoluteError (MAE) – средняя абсолютная ошибка, которая показывает среднее отклонение предсказаний от истинных значений.

- RootMeanSquaredError (RMSE) – корень из среднеквадратичной ошибки, которая также широко применяется для оценки точности регрессионных моделей.

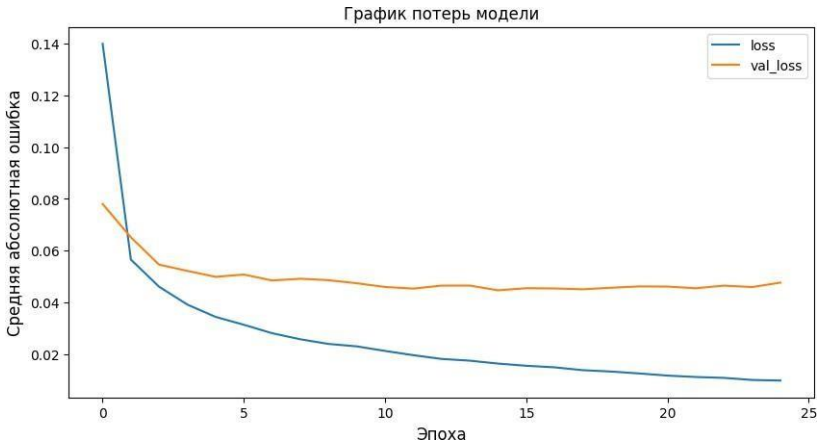
3. Обучение модели (График обучения приведен на рисунке 10).

- Применяется механизм ранней остановки, который отслеживает значение функции потерь на валидационном наборе данных ('val_loss') и останавливает обучение, если улучшение не происходит в течение 10 эпох (patience=10). Это помогает предотвратить переобучение и сохранить лучшие веса модели.

- Параметры обучения. Максимальное число эпох установлено равным 50. При этом, благодаря ранней остановке, обучение может завершиться раньше. 30% данных отводится под валидацию, что позволяет контролировать процесс обучения и вовремя остановить его при признаках переобучения.

- Размер батча установлен равным 32. Это стандартный выбор, позволяющий балансировать между скоростью обучения и точностью.

Рисунок 10 – график обучения модели НС.



Видно, что с самого начала сеть начала переобучаться. Значение loss на тестовых выборках продолжило уменьшаться, а на валидационной начало расти.

Поскольку такая модель не справилась с задачей, мною была выбрана другая модель с DROPOUT слоем. Рассмотрим ее аспекты (График обучения приведен на рисунке 11).

Эта нейросеть представляет собой последовательную модель с тремя скрытыми слоями и одним выходным слоем.

1. Архитектура модели.

Модель принимает данные через входной слой с количеством признаков, соответствующим размеру обучающего набора данных (x3_train.shape).

Три полносвязанных слоя с уменьшающимся числом нейронов: 64, 32 и 16 соответственно. Каждый слой использует активацию ReLU, которая является популярной и эффективной для предотвращения проблемы исчезающего градиента.

Между каждым скрытым слоем добавлен слой Dropout с вероятностью 0.05. Это снижает риск переобучения, случайным образом отключая часть нейронов во время обучения.

Выходной слой представлен одним нейроном без активации, так как модель решает задачу регрессии.

Оптимизатор Adam с фиксированной скоростью обучения хорошо стабилизирует процесс обучения и позволяет быстро находить оптимальное решение.

Функция потерь - среднеквадратичная ошибка (MSE) – стандартная метрика для оценки качества моделей регрессии.

Дополнительно используются MeanAbsoluteError (MAE) и RootMeanSquaredError (RMSE) для более полной оценки точности модели.

Процесс обучения:

Максимально возможное количество эпох равно 50, но процесс обучения может остановиться раньше, если модель перестанет улучшаться. 30% данных

резервируется для валидации (validation_split=0.3), что позволяет отслеживать процесс обучения и предотвращать переобучение.

Плюсы:

1. Три скрытых слоя позволяют модели улавливать сложные зависимости в данных, что потенциально повышает точность.

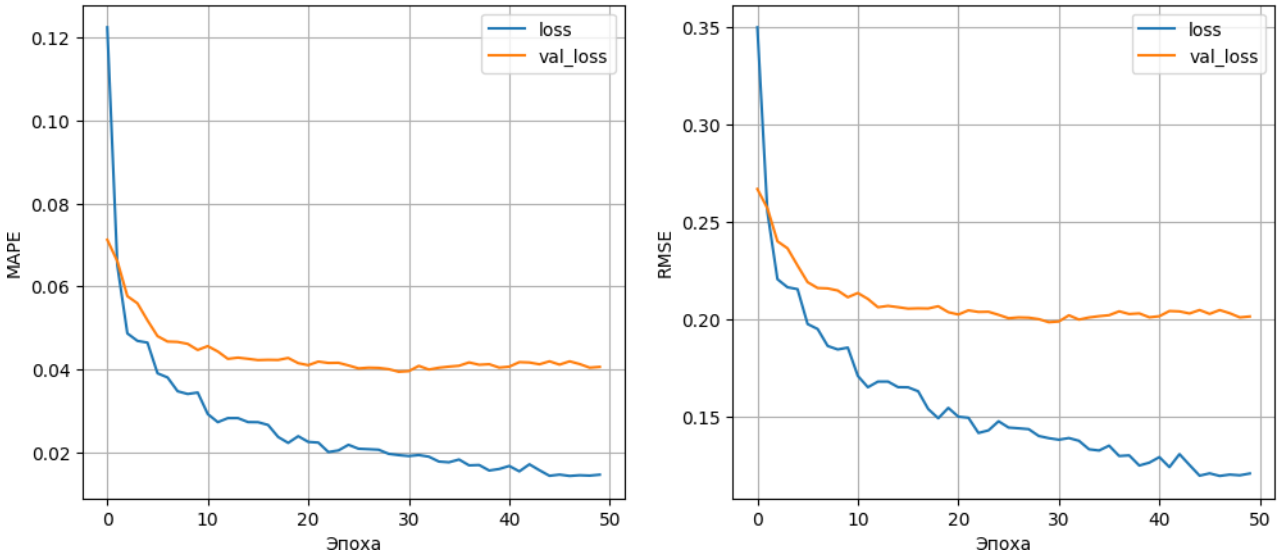
2. Регуляризация с помощью Dropout уменьшает вероятность переобучения, что способствует улучшению обобщающей способности модели.

3. Оптимизатор Adam эффективно справляется с задачей минимизации функции потерь.

4. Помимо основной метрики MSE, используются MAE и RMSE, что дает более полную картину качества модели.

В отличие от предыдущей версии, здесь отсутствует механизм ранней остановки, что может привести к продолжительному обучению даже после достижения оптимальной точки.

Рисунок 11 - График обучения нейронной сети с Dropout.



По графику видно, что данная модель так же не справилась с работой и начала переобучаться примерно на 10 эпохе.

2.4. Разработка приложения

Хотя на данном этапе я не смогл получить модели, готовые к внедрению, разработка функционала приложения все же возможна. Важно отметить, что этот шаг необходим для создания основы, которую впоследствии можно будет дополнить рабочими моделями. Дальнейшие исследования и эксперименты могут позволить мне создать качественные модели, которые будут интегрированы в уже готовое приложение.

При разработке приложения следует учесть несколько ключевых функций, обеспечивающих удобство использования и надежность работы системы, таких как ввод входных параметров. Необходимо предусмотреть удобный интерфейс для ввода данных пользователем. Здесь важно учитывать типы данных и возможные ограничения, чтобы минимизировать ошибки ввода.

Так же была реализована проверка введенных параметров. Перед тем как передать данные на обработку, они должны пройти проверку на корректность. Это поможет избежать ошибок и неверных результатов.

В приложение возможно загружать предварительно обученную модель, когда она станет доступна. Это обеспечит гибкость и возможность обновления модели без необходимости изменять весь код приложения.

После обработки данных моделью результат должен быть представлен пользователю в удобном формате. Например, это может быть числовое значение, график или таблица.

Для реализации этих функций решено использовать язык программирования Python и фреймворк Flask. Python предоставляет широкий спектр библиотек и инструментов для работы с данными и машинным обучением, а Flask — легкий и мощный фреймворк для разработки веб-приложений. Приложение и рабочий ноутбук размещены в GitHub и доступны по адресу: https://github.com/SamsungPipidon/-DS.

Заключение

В ходе выполнения данной работы мы прошли через основные этапы Dataflow pipeline, который включает в себя множество операций и задач, с которыми сталкивается специалист по работе с данными. Мы рассмотрели и применили теоретические методы анализа данных и машинного обучения, а также познакомились с основами предметной области, что позволило нам глубже понять задачу и организовать дальнейшие исследования.

Этот процесс включал несколько ключевых этапов:

1. Изучение теоретических методов анализа данных и машинного обучения, что стало основой для выбора правильных алгоритмов и моделей.
2. Ознакомление с предметной областью, что важно для правильной интерпретации данных и построения осмысленных моделей.
3. Извлечение и трансформация данных, хотя в нашем случае набор данных был уже готов, что упростило этот этап, но и не позволило столкнуться с трудностями работы с разными источниками данных.
4. Проведение разведочного анализа данных (EDA), который дал представление о характере данных и возможных зависимостях между признаками.
5. Разделение данных на обучающую, валидационную и тестовую выборки, что является стандартной практикой для корректной оценки модели.
6. Выполнение предобработки данных для обеспечения корректной работы моделей, включая стандартизацию и обработку категориальных признаков.
7. Построение аналитического решения, выбор и оценка различных моделей, а также подбор гиперпараметров для достижения наилучших результатов.

8. Визуализация результатов работы моделей, что позволило понять их поведение и оценить качество аналитических решений.
9. Разработка и тестирование приложения для поддержки принятия решений, что является важной частью внедрения моделей в реальное использование.

Работа с нейросетями требует значительных усилий и глубокого понимания принципов их функционирования. Несмотря на использование современных подходов и значительное вложение времени, иногда результаты оказываются ниже ожидаемых. Эта ситуация, безусловно, вызывает разочарование, однако она также открывает новые возможности для роста и углубленного изучения предмета.

Почему ни одна из выбранных моделей не смогла дать удовлетворительных результатов? Причины могут быть разнообразными, начиная от недостатка опыта и заканчивая особенностями самих данных. Вот некоторые возможные объяснения:

1. Выбор правильной модели и ее параметров — это сложный процесс, который требует глубокого знания теории и практики. Возможно, недостаточная теоретическая подготовка привела к тому, что были упущены важные моменты в настройке модели.

2. Возможно, что в данных присутствуют нелинейные зависимости, которые сложно уловить простыми моделями.

3. Избыточные или нерелевантные признаки могут усложнить процесс обучения модели и снизить ее точность. Методы отбора признаков, такие как PCA, могут помочь уменьшить размерность данных и сосредоточиться на наиболее важных характеристиках.

4. Некоторые методы машинного обучения работают лучше на определенных типах данных. Возможно, выбранные модели просто не подходят для решения поставленной задачи, и требуется поиск альтернативных подходов.

Каковы дальнейшие шаги для улучшения результатов?

Понимание основ нейросетевых архитектур и методов обучения позволит более точно настраивать модели и выбирать подходящие подходы для конкретных задач. Эксперименты с различными архитектурами и методами оптимизации могут выявить скрытые закономерности в данных.

Применение методов уменьшения размерности, таких как PCA, может помочь избавиться от шума и сконцентрироваться на наиболее информативных признаках. Это упростит задачу для модели и повысит ее точность.

Алгоритмы градиентного бустинга, такие как XGBoost или LightGBM, показали свою эффективность в решении многих задач машинного обучения. Они позволяют строить сложные модели, учитывая взаимодействия между признаками, и предоставляют широкие возможности для тонкой настройки гиперпараметров.

Обсуждение результатов с коллегами или специалистами в предметной области может открыть новые перспективы и предложить идеи для улучшения подхода.

Таким образом, несмотря на текущее разочарование, полученные знания и опыт представляют собой ценный ресурс для будущих исследований. Анализ причин неудач и планирование дальнейших шагов помогут добиться успеха в будущем и развить навыки работы с нейросетями и анализом данных.

Библиографический список

1. Иванов Д. А., Ситников А. И., Шляпин С. Д. Композиционные материалы: учебное пособие для вузов / под ред. А. А. Ильина. — Москва: Издательство Юрайт, 2019. — 253 с. — (Высшее образование). — Текст: непосредственный.
2. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. — СПб.: Питер, 2017. — 336 с.: ил.
3. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. — 2-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2021. — 416 с.: ил.
4. Документация по языку программирования Python. — Режим доступа: https://docs.python.org/3.11/index.html.
5. Документация по библиотеке NumPy. — Режим доступа: https://numpy.org/doc/1.23/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 Scikit-learn. — 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://flaskrussian-docs.readthedocs.io/ru/latest/quickstart.html.
12. Loginom Вики. Алгоритмы. — Режим доступа: https://wiki.loginom.ru/algorithms.html.

13. Ye, A. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать. — Режим доступа: https://habr.com/ru/company/vk/blog/513842/.

14. Maszański, A. Метод k-ближайших соседей (k-nearest neighbour). — Режим доступа: https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearestneighbour-2021-07-19.

15. Kashnitsky, Y. Открытый курс машинного обучения. Тема 3. Классификация, деревья решений и метод ближайших соседей. — Режим доступа: https://habr.com/ru/company/ods/blog/322534/.

16. Maszański, A. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest). — Режим доступа: https://proglib.io/p/mashinnoeobuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12.

17. Maszański, A. Решаем задачи машинного обучения с помощью алгоритма градиентного бустинга. — Режим доступа: https://proglib.io/p/reshaem-zadachi-mashinnogo-obucheniya-s-pomoshchyu-algoritma-gradientnogo-bustinga-2021-11-25.