

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

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science Pro»**

Слушатель

Савченко Кристина Владимировна

Москва, 2024

Оглавление

ВВЕДЕНИЕ	3
ОСНОВНАЯ ЧАСТЬ	4
1.1 ПОСТАНОВКА ЗАДАЧИ.....	4

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

Основная 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. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

В ходе диплома требуется обучить алгоритм машинного обучения, который будет определять значения:

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

Написать нейронную сеть, которая будет рекомендовать:

- Соотношение матрица-наполнитель.

Также необходимо разработать приложение для удобства работы с моделью.

Дата сет состоит из двух файлов: X_br (составляющая из базальт пластика) и X_pur (составляющая из углепластика).

Файл X_br содержит:

- Признаков: 10 и индекс;
- Строк: 1023

Файл X_pur содержит:

- Признаков: 3 и индекс;
- Строк: 1040

Для изучения информации о дата сете необходимо загрузить дата сет в среду разработки. Далее следует просмотреть первые несколько строк дата сета, чтобы убедиться, что все данные отображаются корректно, используя команду `print(df.head())`. Затем можно получить общую информацию о дата сете с помощью команды `print(df.info())`. Эти же действия следует выполнить для дата сета таблицы X_nur.

 `print(bp.info())`

```
<class 'pandas.core.frame.DataFrame'>
Index: 1023 entries, 0 to 1022
Data columns (total 10 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Соотношение матрица-наполнитель         1023 non-null   float64
1   Плотность, кг/м3                         1023 non-null   float64
2   модуль упругости, ГПа                   1023 non-null   float64
3   Количество отвердителя, м.%             1023 non-null   float64
4   Содержание эпоксидных групп,%_2         1023 non-null   float64
5   Температура вспышки, C_2                1023 non-null   float64
6   Поверхностная плотность, г/м2           1023 non-null   float64
7   Модуль упругости при растяжении, ГПа    1023 non-null   float64
8   Прочность при растяжении, МПа           1023 non-null   float64
9   Потребление смолы, г/м2                 1023 non-null   float64
dtypes: float64(10)
memory usage: 87.9 KB
None
```

Рисунок 1 – Пример вывода команды информации о дата сете bp

Известно, что файлы необходимо объединить по индексу, тип объединения — INNER. Поэтому рассмотреть финальный дата сет, а также каждый из признаков предлагаю после объединения.

Для объединения двух дата сетов в один по индексу используем метод `merge`. После объединения мы получили новый дата сет, в котором 13 признаков и 1023 записи.

```
[12] print(df.info())
```

```
>>> <class 'pandas.core.frame.DataFrame'>
Index: 1023 entries, 0 to 1022
Data columns (total 13 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Соотношение матрица-наполнитель         1023 non-null   float64
1   Плотность, кг/м3                         1023 non-null   float64
2   модуль упругости, ГПа                    1023 non-null   float64
3   Количество отвердителя, м.%              1023 non-null   float64
4   Содержание эпоксидных групп,%_2         1023 non-null   float64
5   Температура вспышки, C_2                 1023 non-null   float64
6   Поверхностная плотность, г/м2           1023 non-null   float64
7   Модуль упругости при растяжении, ГПа    1023 non-null   float64
8   Прочность при растяжении, МПа           1023 non-null   float64
9   Потребление смолы, г/м2                 1023 non-null   float64
10  Угол нашивки, град                       1023 non-null   int64
11  Шаг нашивки                             1023 non-null   float64
12  Плотность нашивки                       1023 non-null   float64
dtypes: float64(12), int64(1)
memory usage: 111.9 KB
None
```

Рисунок 2 – Информация об объединённом дата сете df

Прежде чем приступить к описанию каждого признака, необходимо привести названия признаков к общепринятому виду. Для этого следует перевести русские названия признаков на английский язык и выбрать стиль написания. Я предпочитаю использовать змеиный регистр для обозначения признаков. Для этого можно воспользоваться методом `rename`. Сначала создадим список с новыми названиями признаков, а затем заменим текущие названия признаков на обновленные из этого списка.

```
[15] print(df.info())
```

```
>>> <class 'pandas.core.frame.DataFrame'>
Index: 1023 entries, 0 to 1022
Data columns (total 13 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   matrix_filler_ratio                  1023 non-null   float64
1   density_kg_m3                       1023 non-null   float64
2   elastic_modulus_gpa                 1023 non-null   float64
3   hardener_amount_percent             1023 non-null   float64
4   epoxy_groups_content_percent        1023 non-null   float64
5   flash_point_temp_c                 1023 non-null   float64
6   surface_density_g_m2               1023 non-null   float64
7   tensile_elastic_modulus_gpa         1023 non-null   float64
8   tensile_strength_mpa                1023 non-null   float64
9   resin_consumption_g_m2             1023 non-null   float64
10  stitching_angle_deg                 1023 non-null   int64
11  stitching_step                      1023 non-null   float64
12  stitching_density                   1023 non-null   float64
dtypes: float64(12), int64(1)
memory usage: 111.9 KB
None
```

Рисунок 3 – Обновленные названия признаков df

После работы с переименованием признаков необходимо предоставить описание для каждого признака, обозначить целевые переменные в соответствии с поставленной задачей. Дополнительно в информационной таблице следует указать наличие пропущенных значений, дубликатов и количество уникальных значений для каждого признака. Эта информация позволит команде ознакомиться и лучше понять структуру набора данных, с которым предстоит работать, а также определить типы данных. Кроме того, данная информация даст возможность сделать первичный вывод о качестве набора данных и выявить наличие категориальных признаков, что поможет спланировать дальнейшие шаги в анализе.

Перед построением таблицы нам необходимо использовать метод `duplicated` для получения информации о дубликатах и метод `nunique` для получения информации об уникальных значениях. Мы будем выводить только

эту текущую информацию, так как остальная информация уже была получена с помощью метода info.

Таблица 1 – Информация о дата сете df

Признак (название признака в дата сете)	Описание признака	Количество значений	Тип данных	Количество пропущенных значений	Количество дубликатов	Количество уникальных значений
Соотношение матрица-наполнитель (matrix_filler_ratio)	Это отношение объема или массы матричного материала к наполнителю в композитном материале.	1023	float64	0	0	1014
Плотность, кг/м3 (density_kg_m3)	Масса материала на единицу объема, выраженная в килограммах на кубический метр.	1023	float64	0	0	1013
Модуль упругости, ГПа (elastic_modulus_gpa)	Мера жесткости материала, показывающая, насколько материал	1023	float64	0	0	1020

	будет деформироваться под нагрузкой.					
Количество отвердителя, м. %': 'hardener_amount_percent'	Процентное содержание отвердителя в составе композитного материала.	1023	float64	0	0	1005
Содержание эпоксидных групп, % (epoxy_groups_content_percent)	Процентное содержание эпоксидных групп в материале.	1023	float64	0	0	1004
Температура вспышки, С (flash_point_temp_c)	Температура, при которой материал выделяет пары, способные воспламениться.	1023	float64	0	0	1003
Поверхностная плотность,	Масса материала на единицу площади.	1023	float64	0	0	1004

г/м2 (surface_density_g_m2)						
Модуль упругости при растяжении, ГПа (tensile_elastic_modulus_gpa)	Мера сопротивления материала деформации при растяжении.	1023	float64	0	0	1004
Прочность при растяжении, МПа (tensile_strength_mpa)	Максимальное напряжение, которое материал может выдержать при растяжении.	1023	float64	0	0	1004
Потребление смолы, г/м2 (resin_consumption_g_m2)	Количество смолы, необходимое для пропитки единицы площади материала.	1023	float64	0	0	1004
Угол нашивки, град (stitching_angle_deg)	Угол, под которым нити нашивки располагаются относительно	1023	int64	0	0	2

	основной оси материала.					
Шаг нашивки (stitching_step)	Расстояние между соседними стежками нашивки.	1023	float64	0	0	989
Плотность нашивки (stitching_density)	Количество стежков на единицу площади или длины.	1023	float64	0	0	988

Более детальную работу с предобработкой данных мы проведем в п.2 текущего документа, но для начала нам необходимо посмотреть выбросы. Для определения выбросов нам необходимо построить boxplot (ящик с усами). А также найти выбросы с помощью IQR (межквартильного размаха) и метода трех сигм.

Существует множество способов для определения выбросов, в данной работе для больше эффективности мы рассмотрим несколько способов, чтобы очистить данные, пока график boxplot не будет удовлетворительным.

1. Визуализация данных

Визуальные методы — это простой и эффективный способ анализа данных, который позволяет выявить выбросы. Примеры таких методов включают:

- Диаграмма boxplot (ящик с усами): Этот график отображает медианное значение, а также нижний и верхний квартили, выявляя выбросы. Выбросы определяются как значения, выходящие за пределы "усов", которые составляют 1,5 межквартильного размаха (IQR).

2. Метод трех сигм

Метод трех сигм (или правило трех сигм) является статистическим методом, используемым для выявления выбросов в наборе данных. Этот метод основан на свойствах нормального распределения и предполагает, что большинство данных распределено в пределах трех стандартных отклонений от среднего значения.

3. Метод межквартильного размаха (IQR)

Метод IQR использует квартильные значения для выявления выбросов. Межквартильный размах — это разница между первым (Q1) и третьим (Q3) квартилями. Выбросы определяются как значения, выходящие за пределы 1,5 IQR от Q1 и Q3.

Для начала построим «ящик с усами», чтобы увидеть наличие выбросов. На графике отчетливо видны выбросы, теперь необходимо через метод трех сигм и метод межквартильного размаха посчитать количество предполагаемых выбросов.

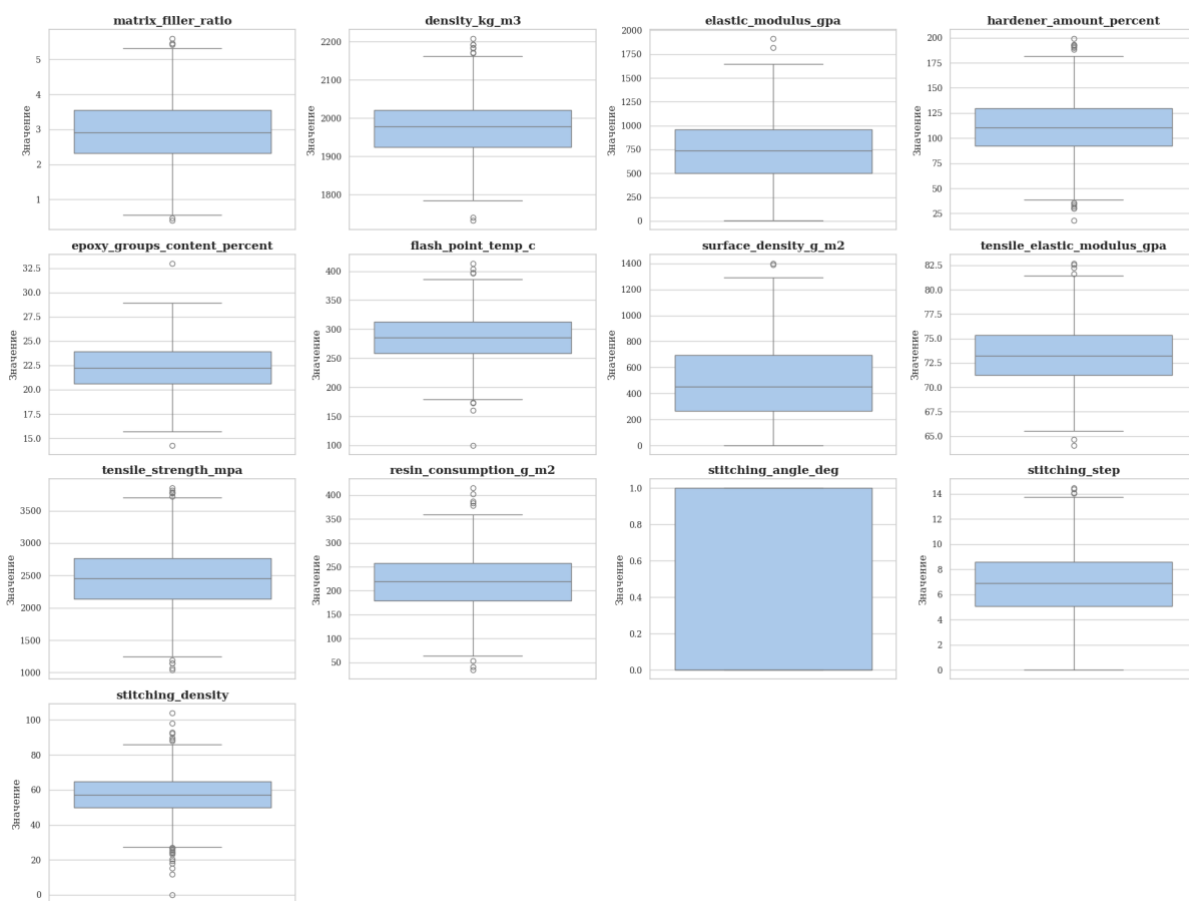


Рисунок 4 – График «ящик с усами»

Методом трех сигм удалось выявить 24 выброса, методом межквартильного размаха 93 выброса. Начнем поэтапно удалять текущие выбросы и смотреть на график.

Для начала удалим выбросы, которые удалось выявить методом трех сигм. Смотрим на график. Результаты неудовлетворительные, выбросы явно еще присутствуют.

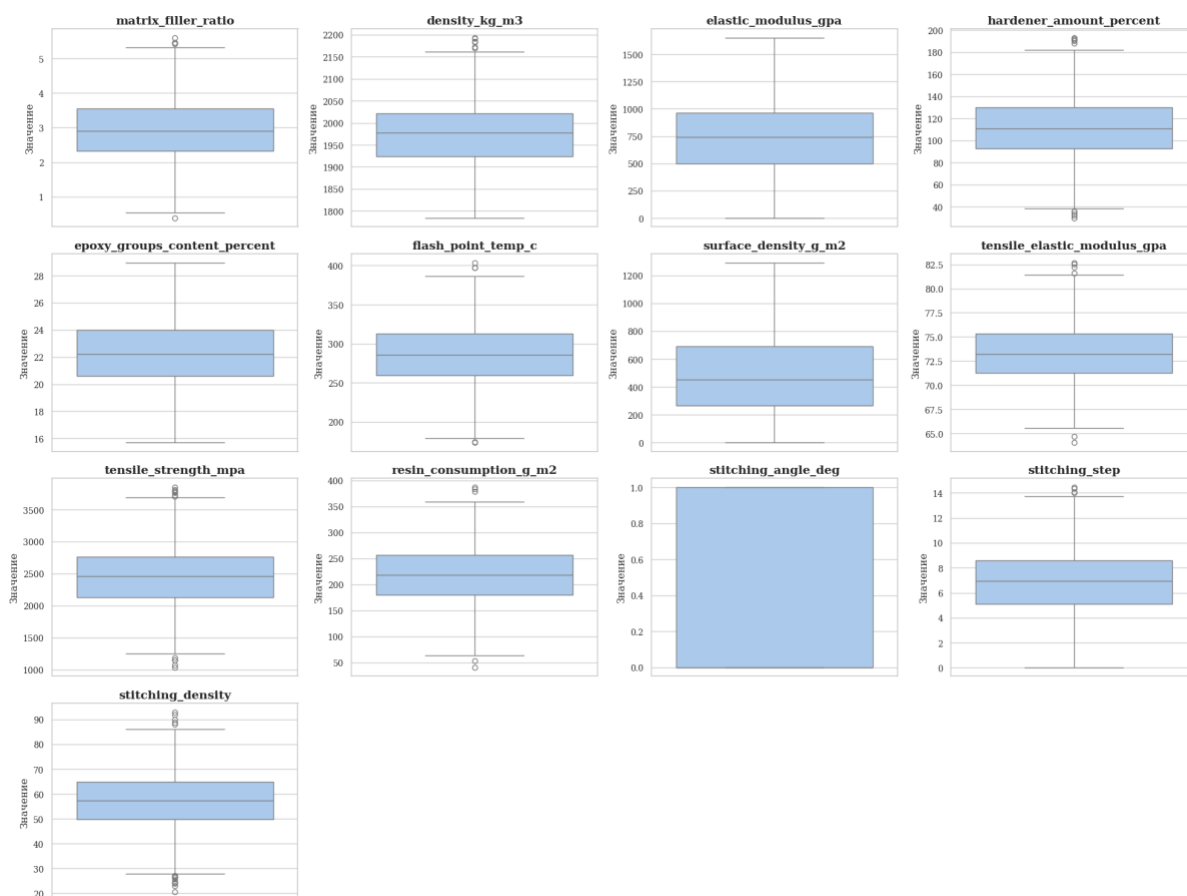


Рисунок 5 – График «ящик с усами» после очистки 24 выбросов

Теперь удалим выбросы, которые удалось выявить методом межквартильного размаха. Смотрим на график. Результаты удовлетворительные, данный дата сет достаточно очищен от выбросов для дальнейшей работы.

После удаление выбросов размер набора данных составляет 932 элемента и 13 признаков.

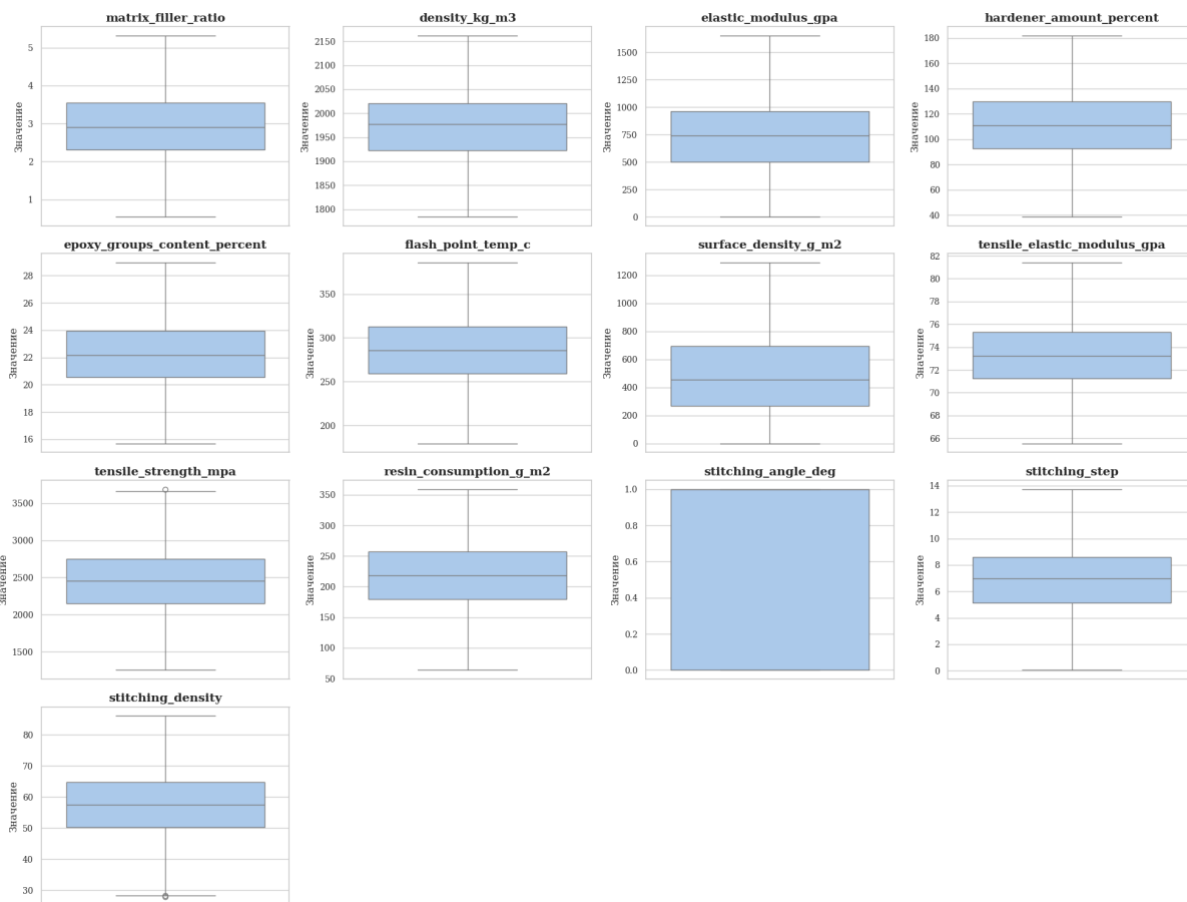


Рисунок 6 – График «ящик с усами» после очистки 93 выбросов

1.2. Описание используемых методов

В задаче необходимо обучить алгоритм машинного обучения, который будет определять значения двух признаков. Поскольку задача у нас основывается на количественных данных, а не на качественных, то это задача регрессии, которую можно решить несколькими алгоритмами, после выбрать наилучший используя метрики качества. В текущей дипломной работе модель будет обучена:

1. Линейная регрессия

Линейная регрессия — это один из самых простых и широко используемых методов регрессии в статистике и машинном обучении. Основная идея заключается в нахождении линейной зависимости между независимой переменной (или переменными) и зависимой переменной.

Модель линейной регрессии пытается найти наилучшую прямую (или гиперплоскость в случае многомерных данных), которая минимизирует разницу между предсказанными и фактическими значениями.

Применение:

- Прогнозирование числовых значений, таких как цены на недвижимость, доходы и т.д.
- Анализ зависимости между переменными.

Особенности:

- Простота в реализации и интерпретации.
- Чувствительность к выбросам и мультиколлинеарности.
- Предполагается линейная зависимость между переменными.

2. Случайный лес

Случайный лес — это ансамблевый метод машинного обучения, который состоит из множества деревьев решений. Основная идея заключается в объединении большого количества деревьев решений для улучшения общей производительности модели. Каждое дерево обучается на случайной подвыборке данных, а итоговое предсказание формируется путем усреднения (для регрессии) или голосования (для классификации) результатов всех деревьев.

Применение:

- Классификация и регрессия в задачах с высокой размерностью данных.
- Обработка данных с пропусками и устойчивость к выбросам.

Особенности:

- Высокая точность и устойчивость к переобучению.
- Возможность оценки важности признаков.
- Сложность интерпретации по сравнению с отдельными деревьями решений.

3. Дерево решений

Дерево решений — это метод машинного обучения, который использует древовидную модель для принятия решений. В процессе обучения дерево решений разбивает данные на подмножества, основываясь на значениях признаков, с целью максимизации информационного прироста или уменьшения неопределенности.

Применение:

- Применяется как для задач классификации, так и для регрессии.
- Используется в системах поддержки принятия решений.

Особенности:

- Простота интерпретации и визуализации.
- Склонность к переобучению при наличии большого количества признаков.
- Необходимость в обрезке дерева для улучшения обобщающей способности.

4. Градиентный бустинг

Градиентный бустинг — это мощный ансамблевый метод, который строит модель путем последовательного добавления слабых моделей (обычно деревьев решений) таким образом, чтобы каждая новая модель исправляла ошибки предыдущих. Градиентный бустинг оптимизирует функцию потерь с помощью градиентного спуска.

Применение:

- Широко используется в задачах классификации и регрессии.
- Выигрывает на многих соревнованиях по анализу данных благодаря своей высокой точности.

Особенности:

- Высокая точность, но может быть склонен к переобучению.
- Требуется тщательной настройки гиперпараметров.

- Может работать медленно на больших наборах данных.

5. K-means

K-means — это алгоритм кластеризации, который делит данные на K кластеров таким образом, чтобы объекты внутри одного кластера были максимально похожи друг на друга, а объекты из разных кластеров — максимально различны. Алгоритм работает итеративно, обновляя центры кластеров до тех пор, пока не будет достигнута сходимость.

Применение:

- Сегментация клиентов в маркетинге.
- Обнаружение паттернов в данных.

Особенности:

- Простота реализации и быстрота работы.
- Зависимость от начальной инициализации кластеров.
- Необходимость заранее задавать количество кластеров K.

6. Кросс-валидация

Кросс-валидация — это метод оценки качества модели, который заключается в разделении данных на несколько подмножеств (фолдов). Модель обучается на всех фолдах, кроме одного, который используется для тестирования. Процесс повторяется для каждого фолда, и результаты усредняются для получения более надежной оценки производительности модели.

Применение:

- Оценка обобщающей способности модели.
- Выбор наилучшей модели или гиперпараметров.

Особенности:

- Уменьшает вероятность переобучения за счет использования всех данных как для обучения, так и для тестирования.
- Может быть вычислительно затратным для больших наборов данных.

7. Байесовская оптимизация гиперпараметров

Байесовская оптимизация — это метод оптимизации, который используется для поиска наилучших гиперпараметров модели машинного обучения. Она строит вероятностную модель функции потерь и использует ее для выбора наиболее перспективных гиперпараметров с учетом предыдущих наблюдений.

Применение:

- Оптимизация гиперпараметров сложных моделей, таких как градиентный бустинг или нейронные сети.
- Улучшение производительности моделей без необходимости полного перебора всех возможных комбинаций гиперпараметров.

Особенности:

- Эффективность при ограниченных вычислительных ресурсах.
- Способность находить глобальные минимумы в сложных пространствах гиперпараметров.
- Требуется некоторое время на настройку и обучение модели вероятностного процесса.

1.3. Разведочный анализ данных

В пункте 1.1 данной дипломной работы уже была частично затронута тема разведочного анализа данных. В данном пункте структурируем выбор данных методов и подробнее опишем методы для разведочного анализа данных.

Разведочный анализ данных (Exploratory Data Analysis, EDA) представляет собой критически важный этап в процессе анализа данных, направленный на выявление основных характеристик и закономерностей в наборе данных. В данном разделе дипломной работы мы подробно рассмотрим различные методы EDA, которые помогут лучше понять структуру и свойства данных. Мы обсудим такие инструменты, как

гистограммы, плотность ядра (KDE), анализ выбросов, матрица корреляции, матрица рассеяния (scatter matrix), нормализация и описательная статистика.

1. Гистограммы

Гистограммы являются одним из наиболее распространенных инструментов визуализации данных, которые позволяют исследовать распределение одной переменной. Они представляют собой график, где данные разбиты на интервалы (или "бины"), а частота попадания значений в каждый интервал отображается в виде столбцов. Гистограммы помогают выявить такие характеристики, как симметричность распределения, наличие модальностей и выбросов.

Основные преимущества гистограмм заключаются в их простоте и наглядности. Они позволяют быстро оценить, насколько данные соответствуют нормальному распределению или имеют асимметрию. Однако выбор размера бинов может существенно повлиять на интерпретацию гистограммы, поэтому важно подобрать оптимальный размер для конкретного набора данных.

2. Плотность ядра (Kernel Density Estimation, KDE)

Метод оценки плотности ядра (KDE) является более гибким инструментом для визуализации распределения данных по сравнению с гистограммами. KDE позволяет создать гладкую кривую плотности, которая отражает вероятностное распределение данных. Это достигается путем наложения "ядер" на каждое наблюдение и суммирования их вкладов.

KDE помогает выявить скрытые структуры в данных, такие как наличие нескольких модальностей, которые могут быть неочевидны при использовании гистограмм. Основным параметром метода KDE является ширина ядра, которая определяет степень сглаживания кривой. Правильный выбор этого параметра критически важен для получения адекватной оценки плотности.

3. Выбросы

Выбросы представляют собой аномальные значения, которые значительно отличаются от остальных данных. Их наличие может существенно исказить результаты анализа и моделирования. Выявление и обработка выбросов — важная часть EDA.

Для обнаружения выбросов часто используются визуальные методы, такие как ящики с усами (box plot) и диаграммы рассеяния. Ящик с усами позволяет оценить распределение данных и выявить значения, выходящие за пределы "усов", которые считаются потенциальными выбросами. После их выявления необходимо принять решение о дальнейшей обработке: исключить из анализа или применить методы трансформации.

4. Матрица корреляции

Матрица корреляции является мощным инструментом для изучения взаимосвязей между переменными в наборе данных. Она представляет собой таблицу, где каждая ячейка содержит коэффициент корреляции между двумя переменными. Наиболее часто используется коэффициент Пирсона, который измеряет линейную зависимость между переменными.

Матрица корреляции помогает выявить сильные и слабые связи между переменными, что может быть полезно для дальнейшего моделирования и отбора признаков. Визуализация корреляционной матрицы с помощью тепловой карты позволяет быстро оценить структуру взаимосвязей и выявить мультиколлинеарность.

5. Матрица рассеяния (Scatter Matrix)

Матрица рассеяния представляет собой набор диаграмм рассеяния для каждой пары переменных в наборе данных. Она позволяет визуально оценить взаимосвязи между переменными и выявить паттерны, такие как линейные или нелинейные зависимости.

Scatter matrix особенно полезна для многомерных данных, так как предоставляет возможность одновременно оценивать взаимодействия между несколькими переменными. Это делает ее незаменимым инструментом при поиске потенциальных связей и аномалий в данных.

6. Нормализация

Нормализация данных является важным шагом в EDA, особенно при работе с переменными, имеющими разные масштабы или единицы измерения. Нормализация позволяет привести данные к единому масштабу, что облегчает сравнение переменных и улучшает сходимость алгоритмов машинного обучения.

Существует несколько методов нормализации, включая минимаксную нормализацию и стандартизацию (z-score normalization). Минимаксная нормализация приводит данные к диапазону $[0, 1]$, тогда как стандартизация приводит данные к стандартному нормальному распределению с нулевым средним и единичным стандартным отклонением.

7. Описательная статистика

Описательная статистика предоставляет количественные сводки основных характеристик данных. Основные показатели включают:

- Среднее значение: Мера центральной тенденции.
- Медиана: Значение, которое делит набор данных на две равные части.
- Мода: Наиболее часто встречающееся значение.
- Стандартное отклонение: Мера вариабельности или рассеяния данных.

Описательная статистика помогает получить общее представление о данных и выявить основные особенности, такие как средние значения, разброс и форма распределения.

Практическая часть

2.1 Предобработка данных

Перед тем как приступить к обучению модели данные необходимо обработать, а после найти между ними зависимости. Для начала посмотрим описательную статистику методом `describe`, нам интереснее всего стандартное отклонение, так как именно эта величина позволит нам сделать вывод относительно разброса данных.

Стандартное отклонение (STD) — это мера разброса данных относительно их среднего значения. Чем больше стандартное отклонение, тем более разбросаны данные.

```
[ ] df.describe()
```



	matrix_filler_ratio	density_kg_m3	elastic_modulus_gpa	hardener_amount_percent
count	1023.000000	1023.000000	1023.000000	1023.000000
mean	2.930366	1975.734888	739.923233	110.570769
std	0.913222	73.729231	330.231581	28.295911
min	0.389403	1731.764635	2.436909	17.740275
25%	2.317887	1924.155467	500.047452	92.443497
50%	2.906878	1977.621657	739.664328	110.564840
75%	3.552660	2021.374375	961.812526	129.730366
max	5.591742	2207.773481	1911.536477	198.953207

Рисунок 7 – Часть описательной статистики

Выводы:

- Набор данных включает переменные с различными уровнями разброса.
- Некоторые переменные имеют низкую вариативность и могут быть более предсказуемыми.
- Другие переменные демонстрируют значительный разброс, что может потребовать дополнительного анализа для понимания причин такой вариативности.

- Высокое стандартное отклонение может указывать на наличие выбросов или экстремальных значений, которые следует проверить отдельно.

После просмотра описательной статистики было принято решение очистить дата сет от выбросов. Эта работа была проведена в пункте один текущего документа. После того, как дата сет очищен необходимо посмотреть на распределение данных внутри дата сета. Для этого нам поможет график kde и матрица рассеяния. На основании графиков будет принято решение о нормализации данных.

Scatter Matrix for Each Feature Pair



Рисунок 8 – Матрица рассеяния

Выводы из графика выше:

- Большинство распределений имеют нормальную форму, что хорошо для дальнейшего анализа.
- Среди пар признаков нет ярко выраженных линейных зависимостей, большинство графиков рассеивания показывают случайный разброс.
- Явных выбросов, сильно влияющих на общее распределение, не наблюдается, хотя некоторые графики все же показывают немного выбросов.

Теперь можно строить kde (Плотность ядра (Kernel Density Estimation, KDE)). На основании данного графика мы сможем понять насколько разбросаны данные друг от друга, поскольку сильный разброс данных далее может негативно сказаться на обучающей способности модели.

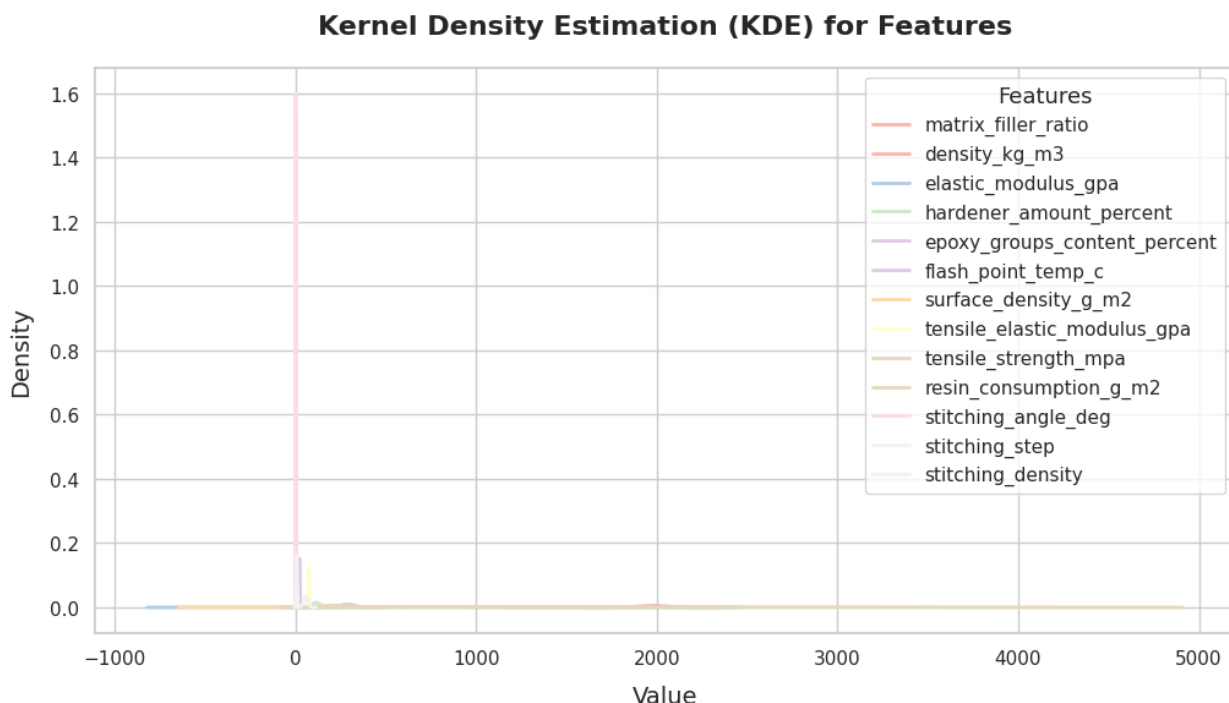


Рисунок 9 – Плотность ядра

На графике плотности данных видно, что некоторые переменные имеют значения, которые сильно варьируются между собой, в то время как большинство других значений сосредоточены около нуля. Это указывает на наличие различных масштабов по разным признакам.

Нормализация данных в таком случае необходима для того, чтобы привести все переменные к сопоставимым масштабам. Это особенно важно при использовании методов машинного обучения, таких как градиентный спуск или K-средних, которые чувствительны к масштабам данных.

Нормализовать данные было принято методом `MinMaxScaler`. `MinMaxScaler` – это инструмент из библиотеки `scikit-learn`, который применяется для нормализации данных. Его основная задача заключается в преобразовании значений признаков в диапазон от 0 до 1. Это достигается путем линейного масштабирования каждого признака с учетом минимального и максимального значений в наборе данных.

В данной работе рассматривается применение `MinMaxScaler` для нормализации данных о свойствах компонентов композиционных материалов, таких как количество связующего, наполнителя и температурный режим. Эти данные могут иметь различные единицы измерения и масштабы, что делает использование `MinMaxScaler` особенно актуальным. Данный метод позволяет привести все признаки к общему диапазону от 0 до 1, что упрощает работу модели.

Одним из преимуществ использования `MinMaxScaler` является простота интерпретации результатов нормализации. Поскольку все значения находятся в пределах от 0 до 1, это облегчает визуализацию и анализ данных. Кроме того, `MinMaxScaler` особенно эффективен в случаях, когда набор данных не содержит значительных выбросов, так как он использует экстремальные значения для определения диапазона. Это делает его более подходящим по сравнению с другими методами нормализации, такими как `StandardScaler`.

MinMaxScaler также хорошо подходит для работы с линейными моделями и моделями, чувствительными к масштабу данных, такими как линейная регрессия или метод опорных векторов (SVM). Приведение признаков к единому масштабу способствует более эффективному обучению модели и улучшению качества прогнозирования конечных свойств композиционных материалов. Таким образом, использование MinMaxScaler в рамках данного исследования способствует стандартизации входных данных и повышает эффективность моделирования.

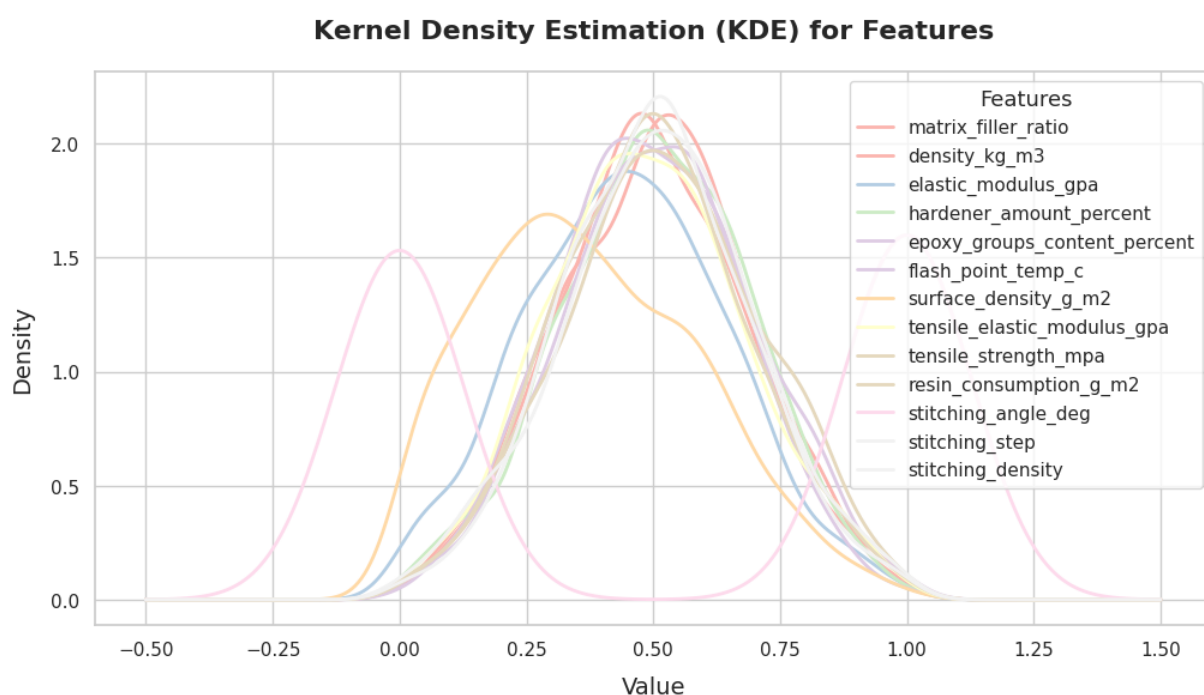


Рисунок 10 – Плотность ядра после нормализации

После того, как данные предобработаны, в том числе очищены от выбросов и нормализованы, можно приступить к выбору способов обучения с учетом поставленной задачи. Поскольку прогнозирование количественных целевых переменных – это как правило всегда методы регрессии, нам необходимо посмотреть зависимость между переменными, чтобы установить, есть ли корреляция между ними (прямая или обратная) и от этого выбрать дальнейшие шаги по обучению модели.

Выше из графика распределения уже удалось установить, что между параметрами нет корреляции вообще, либо есть очень слабо коррелирующие между собой переменные. Поскольку далее мы очистили данные от выбросов и провели нормализацию, необходимо понять, насколько улучшилась корреляция. Для этого мы построим на одном графике матрицы корреляций сразу для трех дата сетов:

- 1) Дата сет с первоначальными данными (df)
- 2) Дата сет после очистки от выбросов (df_clean_2)
- 3) Дата сет после нормализации (df_norm)

После построения графиков мы также для удобства напишем программу, которая посчитает разницу в корреляции между дата сетами, проанализировав текущую разницу сделаем выводы насколько удалось улучшить зависимость между данными для дальнейшего обучения и появилась ли она после преобразования.

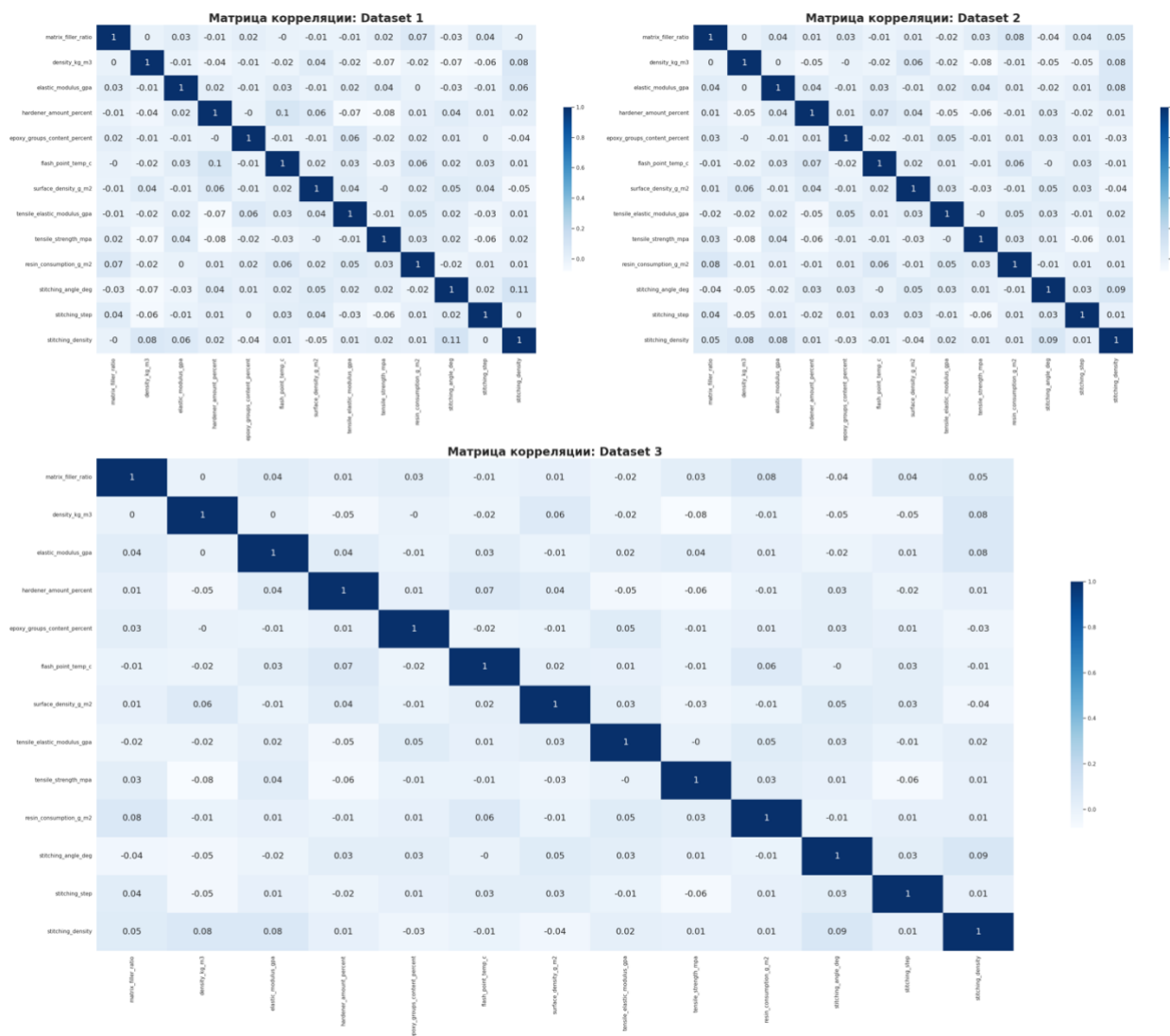


Рисунок 11 – Матрицы корреляции трех дата сетов

После напишем программу, которая с помощью разницы корреляции между дата сетами выведет ее отдельно. Средняя абсолютная разница (Mean Absolute Difference, MAD) — это мера, которая показывает, насколько в среднем значения одной переменной отличаются от значений другой переменной. В контексте сравнения корреляционных матриц, средняя абсолютная разница позволяет оценить, насколько сильно различаются корреляции между двумя наборами данных.

Полученные выводы после сравнения:

- Средняя абсолютная разница между Dataset 1 и Dataset 2:
0.009839408790395668
- Средняя абсолютная разница между Dataset 1 и Dataset 3:
0.00983940879039614
- Средняя абсолютная разница между Dataset 2 и Dataset 3:
6.120821664683358e-16
- Dataset 2 и Dataset 3 имеют наименьшую разницу в корреляции

Отсюда сделаем следующие выводы:

- Наименьшая средняя абсолютная разница наблюдается между Dataset 2 и Dataset 3 (6.120821664683358e-16), что означает, что корреляции в этих двух наборах данных практически идентичны.
- После нормализации очищенного от выбросов дата сета зависимость не появилась.
- В целом все три дата сета имеют очень маленькую разницу в корреляции, что может говорить о том, что не смотря на попытки очистки и предобработки данных корреляции не могло быть. Надо обращаться к пункту сбора данных, либо дополнять данные новыми фичами до появления зависимости, либо описать верный процесс сбора данных для данного кейса.

Простой способ обучения через линейную регрессию не подойдет, поэтому было решено поставить несколько целей и гипотез перед дальнейшей работой.

1. Обучить модель разными способами регрессии (указаны в пункте 1 данной дипломной работы), чтобы убедиться в том, что качество модели при таких данных будет минимальным.
2. Доказать гипотезу о том, что данные первоначально были собраны не верно, либо в них недостаточно признаков для обучения модели и связи между ним.

3. При подтверждении гипотезы написать рекомендации по сбору данных.
4. В текущей дипломной работе в качестве учебных целей провести обучение возможными методами и выбрать наилучший, сравнивая данные по качеству с базовой моделью. Базовой моделью в данном кейсе считать: обучение через показатели средних значений (метод mean).
5. Дополнительно путем разных практических решений и гипотез удалось получить зависимость между данными с помощью метода Normalizer() и обучить модель с достаточно высоким качеством и хорошим предсказанием ответов на тестовом датасете через проверку в ручную на веб интерфейсе. Но данный случай было принято вынести как Приложение к диплому, поскольку метод Normalizer() по своему описанию и предназначению не подходит к решению данного кейса, поскольку преобразует отдельные строки по L2 векторной норме, что в свою очередь разрушает связь между отдельными столбцами. При этом практический опыт был сохранен, поскольку модель действительно начала предсказывать данные достаточно хорошо на тестовых выборках.

2.2. Разработка и обучение модели

Выше после работы с данными были поставлены цели. Для начала обучим модель несколькими wybranными способами, посмотрим графики качества обучения модели, в следующем пункте сделаем выводы по качеству обучения моделей и сравнение с базовой моделью. А также корректировку дальнейших шагов при необходимости.

1. Линейная регрессия

В данном разделе дипломной работы описывается процесс подготовки данных, обучения и оценки качества модели линейной регрессии для прогнозирования двух целевых переменных: модуля упругости и прочности материала. В приведенном коде выполняются следующие шаги. Сначала данные загружаются в `DataFrame`, из которого выделяются признаки и целевые переменные. Признаки представляют собой все столбцы, кроме целевых переменных, которые будут предсказываться. Целевая переменная для модуля упругости обозначена как `tensile_elastic_modulus`, а для прочности — как `tensile_strength`. Далее данные разделяются на обучающую и тестовую выборки с использованием функции `train_test_split`. Размер тестовой выборки составляет 20% от общего объема данных, а размер обучающей выборки — 80%. Разделение производится отдельно для каждой из целевых переменных, чтобы обеспечить независимость обучения и тестирования моделей. После этого создаются две модели линейной регрессии: одна для модуля упругости и другая для прочности. Каждая модель обучается на соответствующей обучающей выборке с помощью метода `fit`. Затем для обеих моделей выполняются предсказания на обучающих и тестовых выборках с использованием метода `predict`. Для оценки качества моделей используются метрики среднеквадратичной ошибки (MSE) и коэффициента детерминации (R^2). MSE измеряет среднее квадратичное отклонение предсказанных значений от истинных, показывая, насколько хорошо модель соответствует данным. Коэффициент детерминации R^2 показывает долю дисперсии целевой переменной, объясняемую моделью. Значение R^2 , близкое к 1, указывает на хорошее качество модели. В завершение выводятся значения MSE и R^2 для обучающих и тестовых выборок каждой из моделей, что позволяет оценить их производительность и обобщающую способность. Таким образом, данный код демонстрирует полный цикл построения и оценки линейной регрессионной

модели для прогнозирования механических свойств материала, что является важной частью анализа данных в рамках дипломной работы.

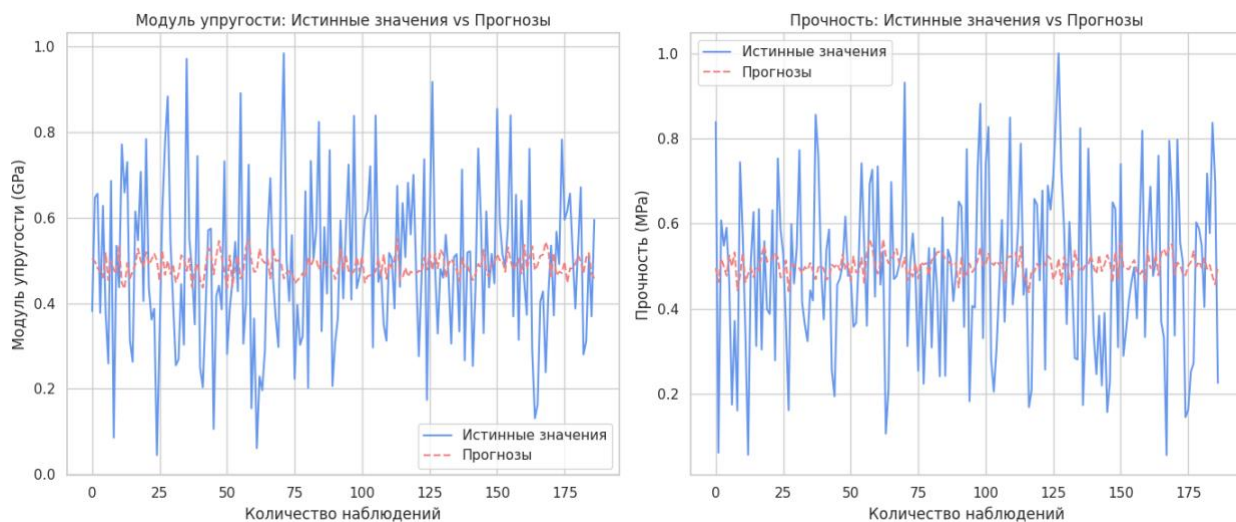


Рисунок 12 – Качество модели обученной LinearRegression

2. Случайный лес

В данном разделе дипломной работы рассматривается процесс построения и оценки модели машинного обучения для прогнозирования двух целевых переменных: модуля упругости при растяжении и прочности при растяжении композитных материалов. Для решения этой задачи используется алгоритм случайного леса, который является одним из популярных методов в области машинного обучения благодаря своей эффективности и способности обрабатывать сложные зависимости в данных.

Определение признаков и целевых переменных

На первом этапе определяются признаки, которые будут использоваться для построения модели. Эти признаки включают различные физико-химические характеристики материалов, такие как соотношение матрицы и наполнителя, плотность, количество отвердителя, содержание эпоксидных групп, температура вспышки, поверхностная плотность, расход смолы, угол и шаг прошивки, а также плотность прошивки. Целевыми переменными являются модуль упругости при растяжении и прочность при растяжении.

Разделение данных на обучающую и тестовую выборки

Для оценки качества модели данные разделяются на обучающую и тестовую выборки. Обучающая выборка используется для настройки параметров модели, в то время как тестовая выборка позволяет оценить её обобщающую способность. В данном случае данные разделяются в соотношении 80:20.

Обучение модели случайного леса

Для каждой из целевых переменных создается отдельная модель случайного леса. Случайный лес представляет собой ансамбль деревьев решений, где каждое дерево обучается на случайной подвыборке данных. Это позволяет улучшить устойчивость модели к переобучению и повысить её точность.

Оценка качества модели

После обучения модели проводится оценка её качества на обучающей и тестовой выборках. Для этого используются такие метрики, как среднеквадратичная ошибка (MSE) и коэффициент детерминации (R^2). Среднеквадратичная ошибка позволяет измерить среднее квадратичное отклонение предсказанных значений от истинных, а коэффициент детерминации показывает долю дисперсии зависимой переменной, объясненную моделью.

Результаты оценки качества модели показывают её способность точно предсказывать значения целевых переменных на новых данных. Высокие значения R^2 и низкие значения MSE свидетельствуют о хорошей производительности модели.

Таким образом, в этом разделе была продемонстрирована возможность использования алгоритма случайного леса для решения задач регрессии в контексте прогнозирования механических свойств композитных материалов.

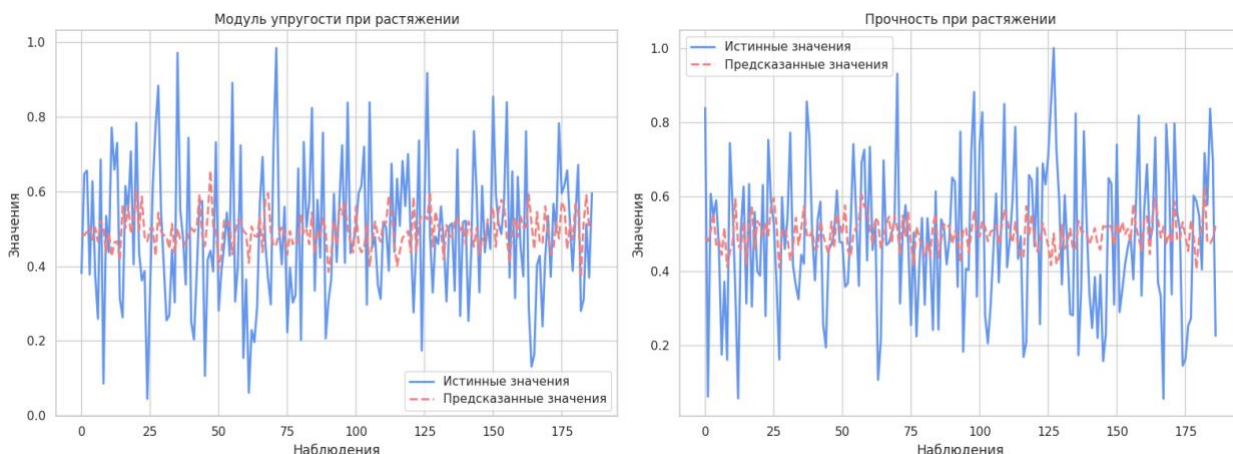


Рисунок 13 – Качество модели обученной RandomForestRegressor

3. Градиентный бустинг

В данном разделе дипломной работы рассматривается процесс построения и оценки моделей машинного обучения для прогнозирования модуля упругости и прочности при растяжении композитных материалов. Основной целью является использование алгоритма градиентного бустинга для создания предсказательных моделей, которые могут точно оценивать указанные механические свойства на основе заданных признаков.

Подготовка данных

В первую очередь, из исходного набора данных были выделены признаки, которые, предположительно, влияют на модуль упругости и прочность при растяжении. Эти признаки включают в себя долю наполнителя в матрице, плотность материала, количество отвердителя, содержание эпоксидных групп, температуру вспышки, поверхностную плотность, расход смолы, угол прошивки, шаг прошивки и плотность прошивки. Целевые переменные представляют собой модуль упругости и прочность при растяжении.

Данные были разделены на обучающую и тестовую выборки в соотношении 80% и 20% соответственно. Это позволяет оценить качество модели как на обучающей выборке, так и на независимой тестовой выборке.

Обучение модели

Для каждой из целевых переменных была построена отдельная модель с использованием алгоритма градиентного бустинга. Градиентный бустинг — это популярный метод машинного обучения, который строит ансамбль слабых моделей (обычно деревьев решений) для улучшения точности предсказаний. Алгоритм последовательно корректирует ошибки предыдущих моделей, минимизируя функцию потерь с использованием градиентного спуска.

Оценка качества моделей

Качество моделей оценивалось с использованием среднеквадратичной ошибки (MSE) и коэффициента детерминации (R^2). Эти метрики позволяют определить, насколько хорошо модель соответствует данным:

- Среднеквадратичная ошибка (MSE) измеряет среднюю величину ошибки между предсказанными и фактическими значениями. Чем меньше значение MSE, тем точнее модель.

- Коэффициент детерминации (R^2) показывает долю дисперсии целевой переменной, объясняемую моделью. Значение R^2 ближе к 1 указывает на высокое качество модели.



Рисунок 13 – Качество модели обученной GradientBoostingRegressor

Поскольку все остальные обученные модели, способы обучения которых были перечислены в пункте один данной дипломной работы показали еще хуже качество было принято решение оставить информацию о них только в блокноте с кодом. Сам блокнот с кодом был приложен к диплому, внутри блокнота можно найти подробное описание и выводы по всем графикам.

Поскольку все модели построены, то теперь можно протестировать данные по ним и сравнить качество всех моделей.

2.3. Тестирование модели.

При разработке самих моделей также были построены параметры качества тестовых и обучающих выборок, информация про них есть в разделе выше. Графики качества обученных моделей есть в блокноте с кодом.

Здесь же мы посмотрим на данные по качеству всех моделей и сравним. Для сравнения моделей была написана программа, которая вычисляет лучшую модель по среднеквадратичной ошибки (MSE) и коэффициента детерминации (R^2), выбирая наименьший MSE и наибольший R^2 .

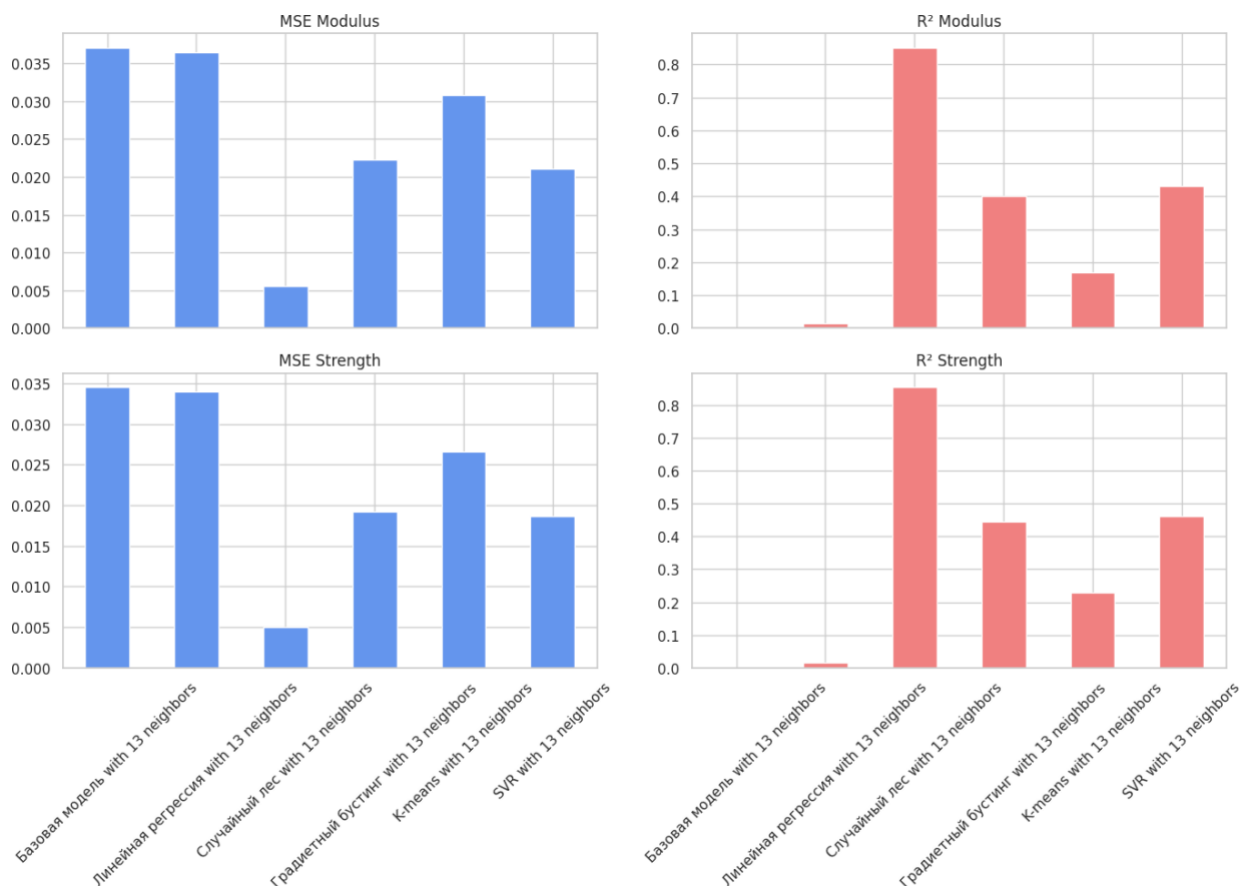


Рисунок 14 – Качество моделей для тренировочного дата сета

При разработке и применении моделей машинного обучения одним из основных шагов является оценка их качества на этапе обучения. Основными метриками, используемыми для оценки регрессионных моделей, являются среднеквадратическая ошибка (Mean Squared Error, MSE) и коэффициент детерминации (R^2). Однако использование данных метрик, рассчитанных на тренировочном наборе данных, для выбора окончательной модели может привести к некорректным выводам. В данной работе представлено исследование качества различных моделей, оцененное на тренировочном датасете, и обоснование, почему выбор модели не должен основываться исключительно на этих метриках.

Методы

Были рассмотрены следующие модели регрессии:

1. Базовая модель с 13 соседями

2. Линейная регрессия с 13 соседями
3. Случайный лес с 13 соседями
4. Градиентный бустинг с 13 соседями
5. K-means с 13 соседями
6. Опорный вектор регрессии (SVR) с 13 соседями

Для каждой модели были рассчитаны два показателя качества:

- Среднеквадратическая ошибка (MSE)
- Коэффициент детерминации (R^2)

Результаты измерений представлены на тренировочном наборе данных.

На основе анализа графиков видно, что:

- Градиентный бустинг показывает наивысшие значения R^2 и наименьшие значения MSE для обеих целевых переменных: Modulus и Strength.
- Линейная регрессия и Случайный лес также демонстрируют конкурентоспособные результаты, с более высокими значениями R^2 по сравнению с остальными моделями.
- Базовая модель и K-means показывают наихудшие результаты с точки зрения MSE и R^2 .

Основной проблемой выбора модели по результатам тренировочного набора данных является проблема переобучения. Модель, показывающая превосходные результаты на тренировочных данных, может быть сильно адаптирована к этим данным и не обобщать на новые, невидимые данные. Это может привести к тому, что модель не будет демонстрировать такое же качество на тестовом наборе данных или в реальных приложениях.

Например, модель случайный лес показывает наилучшие результаты на тренировочных данных, но это может быть следствием избыточной подгонки и переобучения. Если выбрать данную модель, не проверив ее на тестовом

наборе данных, есть высокий риск получить низкую производительность на практике, что может негативно сказаться на конечном продукте или системе.

Выбор модели на основании метрик, рассчитанных только на тренировочном наборе данных, является ошибочным. Для корректной оценки качества следует использовать метод перекрестной проверки (cross-validation) и/или проверку на отложенном тестовом наборе данных. Только так можно сбалансированно оценить способность модели к обобщению и ее реальную производительность на невидимых данных.

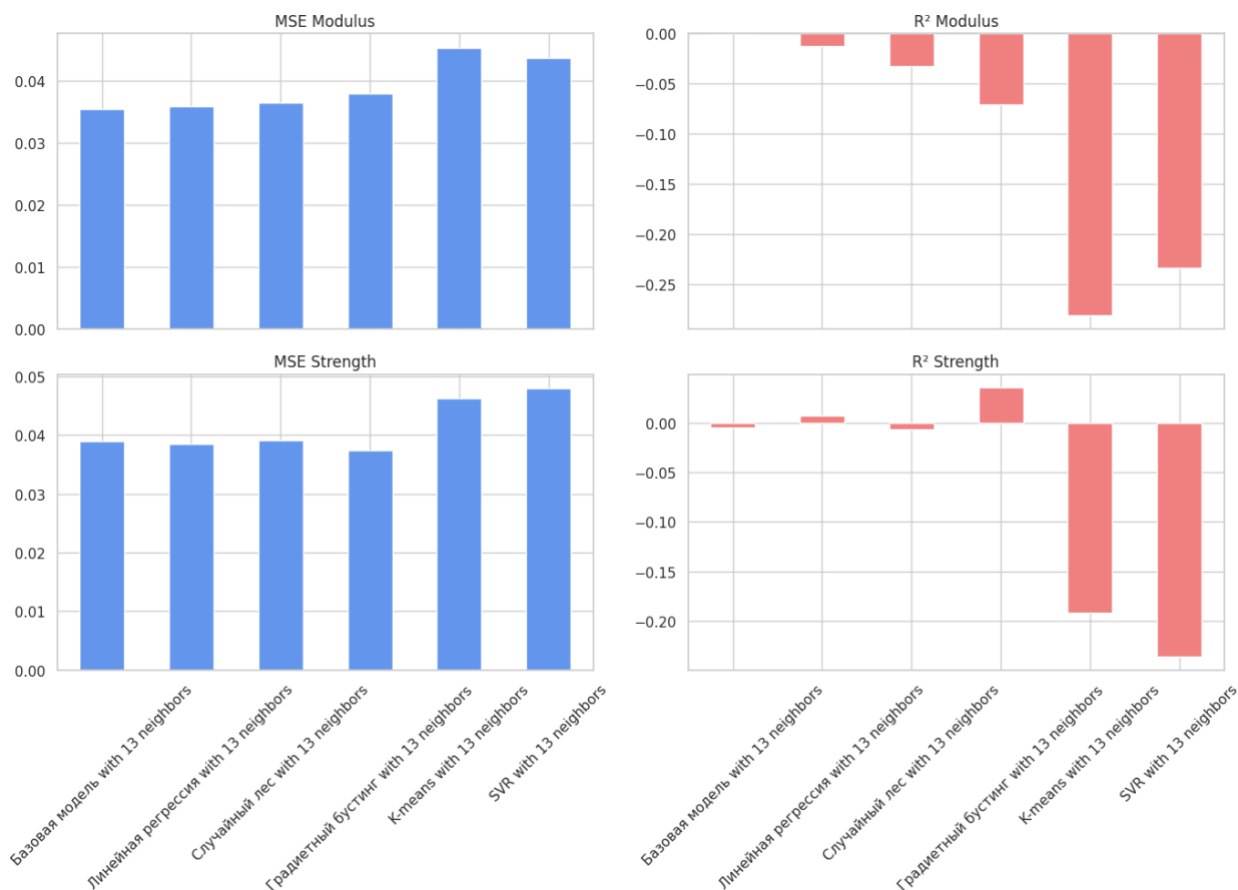


Рисунок 15 – Качество моделей для тестового дата сета

Рассматривая результаты тестирования, можно заметить, что выбор модели исключительно на основе данных уже на тестовых данных приводит к проблеме переобучения (overfitting) или недообучения (underfitting). Тестирование на валидационном наборе данных помогает избежать модели,

которая чрезмерно адаптируется под предоставленные данные, что может привести к низкой обобщающей способности.

Для того, чтобы выбрать модель на основании сразу всех параметров качество был написан код расчета средних значений для тестовых и обучающих дата сетов.

В данном коде производится выбор модели на основе двух ключевых метрик: среднеквадратичной ошибки (MSE) и коэффициента детерминации (R^2). Эти метрики вычисляются как для обучающей, так и для тестовой выборки, после чего вычисляются их средние значения. Алгоритм выбора модели можно описать следующим образом:

- Для каждой модели в данных вычисляются средние значения MSE и R^2 по обучающей и тестовой выборкам.

- Среднее значение MSE для Modulus (Avg MSE Modulus) рассчитывается как среднее арифметическое между MSE Modulus_train и MSE Modulus_test.

- Среднее значение R^2 для Modulus (Avg R^2 Modulus) рассчитывается аналогично.

- То же самое делается для Strength: Avg MSE Strength и Avg R^2 Strength.

Для Modulus:

- Находится модель с минимальным значением Avg MSE Modulus. Это модель с наименьшей среднеквадратичной ошибкой, что указывает на более точные предсказания.

- Находится модель с максимальным значением Avg R^2 Modulus. Это модель с наибольшим коэффициентом детерминации, что говорит о лучшем соответствии модели данным.

Для Strength:

- Аналогично, определяется модель с минимальным Avg MSE Strength.

- И модель с максимальным Avg R² Strength.

Вывод результатов:

- Печатаются названия моделей, которые были определены как лучшие по каждой из метрик для Modulus и Strength.

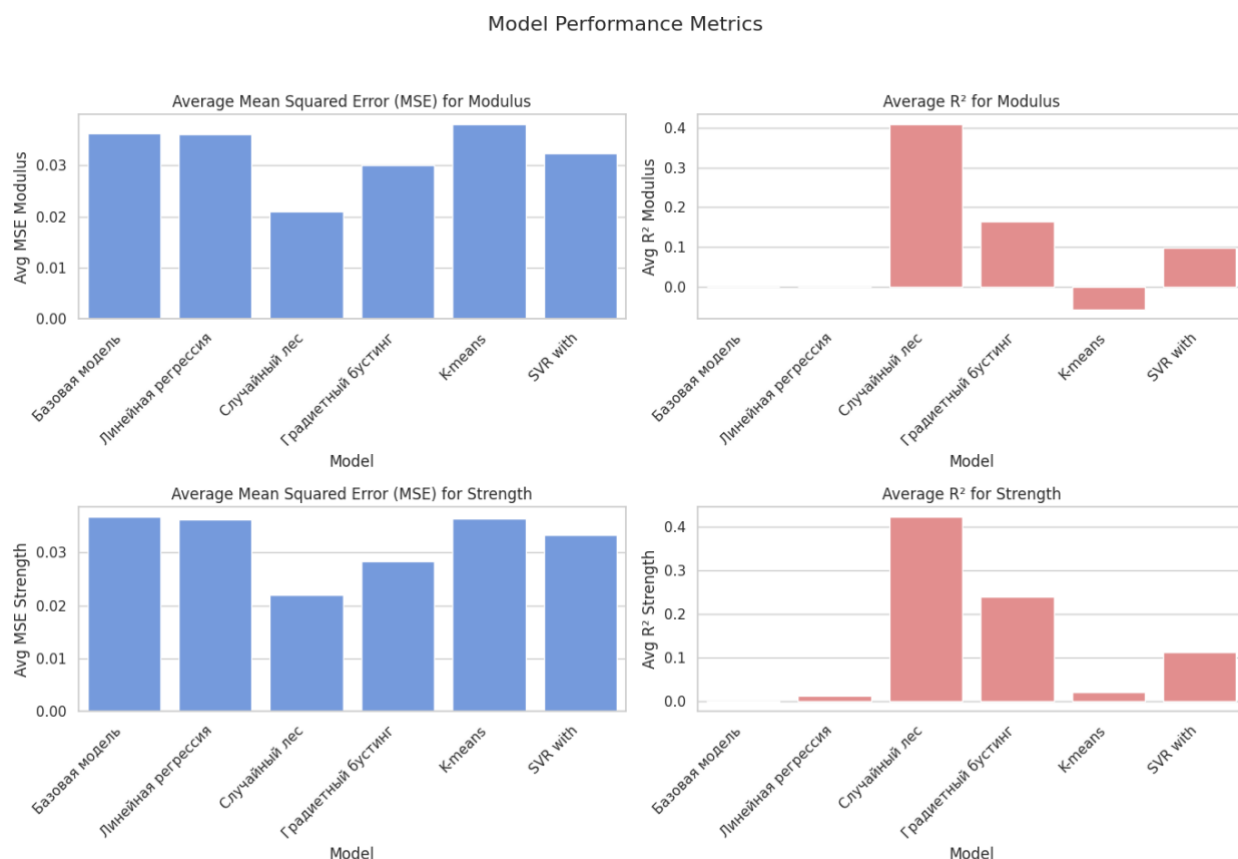


Рисунок 16 – Качество моделей после вычисления средних значений

1. Среднеквадратичная ошибка (MSE):

- Градационное уменьшение MSE указывает на хорошую точность модели. Как показано на графиках, "Случайный лес" имеет наименьшую среднеквадратичную ошибку как для модуля (верхний левый график), так и для прочности (нижний левый график).

2. Коэффициент детерминации (R²):

- Коэффициент детерминации показывает, насколько хорошо модель объясняет вариативность данных. "Случайный лес" показывает наибольшее значение R^2 как для модуля (верхний правый график — около 0.35), так и для прочности (нижний правый график — около 0.40), что означает, что данная модель лучше всего объясняет вариативность данных по сравнению с другими моделями.

На основании низкого значения среднеквадратичной ошибки и высокого коэффициента детерминации, "Случайный лес" демонстрирует лучшие результаты. При этом мы видим, что почти все модели показали низкую способность на тестовых данных, поэтому в продакшене было бы принято решение о добавлении или удалении признаков в дата сет, а также написания рекомендаций по сбору данных для текущего дата сета.

2.4. Нейронная сеть

В данном разделе была разработана нейронную сеть, способная рекомендовать оптимальное соотношение матрицы на основе различных входных характеристик материала.

Архитектура нейронной сети

Для решения задачи прогнозирования соотношения матрицы была выбрана архитектура полносвязной нейронной сети. Модель состоит из трех слоев:

1. Входной слой: Содержит 64 нейрона и использует функцию активации ReLU. Количество нейронов соответствует количеству признаков, используемых для обучения модели, что позволяет эффективно обрабатывать входные данные.

2. Скрытый слой: Состоит из 32 нейронов и также использует функцию активации ReLU. Этот слой служит для извлечения скрытых закономерностей и взаимосвязей между входными данными.

3. Выходной слой: Содержит один нейрон без функции активации, так как задача является регрессионной и требуется предсказать непрерывное значение — соотношение матрицы.

Модель была скомпилирована с использованием оптимизатора Adam и функции потерь `mean_squared_error`, что обеспечивает эффективное обучение модели и минимизацию ошибки предсказания.

Обучение и оценка модели

Обучение модели проводилось на нормализованных данных, разделенных на обучающую и тестовую выборки в соотношении 80/20. Для повышения обобщающей способности модели использовалась валидация на 10% обучающей выборки.

В ходе разработки модели для прогнозирования соотношения матрицы в композитных материалах была получена модель с показателем Test Loss: 0.052163612097501755. Этот показатель указывает на среднеквадратичную ошибку, которую модель допускает при предсказании на тестовой выборке. В данном разделе мы рассмотрим качество модели и предложим рекомендации для ее улучшения.

Значение тестовой ошибки 0.052 свидетельствует о том, что модель в среднем допускает небольшую ошибку в своих предсказаниях. Однако, в зависимости от контекста применения и требований к точности, это значение может быть как приемлемым, так и требующим улучшения.

Рекомендации по улучшению модели

Несмотря на успешные результаты, есть несколько направлений для дальнейшего улучшения модели:

1. Увеличение объема данных: Дополнительные данные могут помочь модели лучше обобщать и улучшить ее предсказательную способность.

2. Оптимизация гиперпараметров: Проведение более детального подбора гиперпараметров, таких как количество нейронов в слоях, скорость обучения и размер батча, может привести к улучшению качества модели.

3. Использование более сложных архитектур: Рассмотрение более сложных архитектур, таких как рекуррентные или сверточные нейронные сети, может быть полезно для учета временных зависимостей или пространственных связей в данных.

4. Аугментация данных: Использование методов аугментации данных может помочь в увеличении разнообразия обучающей выборки и улучшении устойчивости модели к шуму.

Заключая данный раздел, можно отметить, что разработанная нейронная сеть демонстрирует достаточную точность в задаче прогнозирования соотношения матрицы, однако дальнейшие улучшения могут быть достигнуты за счет оптимизации архитектуры и расширения обучающего набора данных.

2.5. Создание веб приложения и бота с уведомлениями

Было разработано приложение, которое доступно на <http://127.0.0.1:5000/> локально на компьютере. В данном приложении был разработан простой интерфейс и кнопка с анимацией. При нажатии на кнопку, приложению обращается к выбранной модели данных и делает предсказание нужных параметров. Далее эти предсказания направляются в чат-бот, чтобы было возможно отслеживать данные и смотреть как модель показывает разные данные при разных моделях или подходах к обучению.

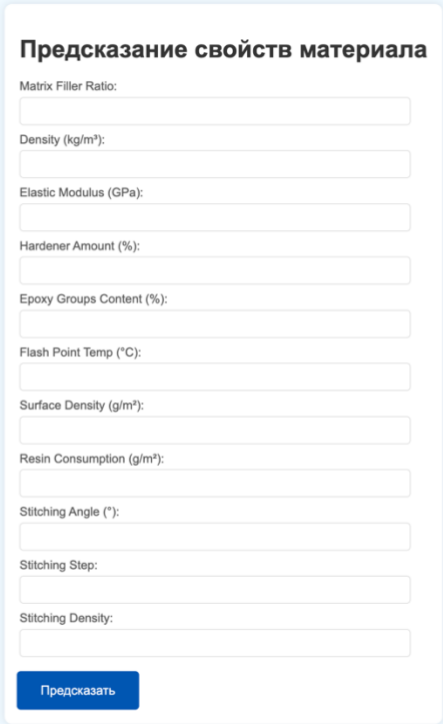
Инструкция к технической документации для использования веб-приложения

Описание

Данное веб-приложение позволяет пользователям вводить параметры материала и получать предсказанные значения модуля упругости и прочности материала. Предсказание отображается в чат-боте после нажатия кнопки "Предсказать".

Шаг 1: Запуск веб-приложения

1. Откройте веб-браузер и введите URL-адрес веб-приложения.
2. Дождитесь загрузки страницы.



Предсказание свойств материала

Matrix Filler Ratio:

Density (kg/m³):

Elastic Modulus (GPa):

Hardener Amount (%):

Epoxy Groups Content (%):

Flash Point Temp (°C):

Surface Density (g/m²):

Resin Consumption (g/m²):

Stitching Angle (°):

Stitching Step:

Stitching Density:

Предсказать

Рисунок 17 – Веб приложение

Шаг 2: Ввод параметров материала

1. На главной форме веб-приложения вы увидите поля для ввода следующих параметров материала:

- Matrix Filler Ratio
- Density (kg/m^3)
- Elastic Modulus (GPa)
- Hardener Amount (%)
- Epoxy Groups Content (%)
- Flash Point Temp ($^{\circ}\text{C}$)
- Surface Density (g/m^2)
- Resin Consumption (g/m^2)
- Stitching Angle ($^{\circ}$)
- Stitching Step
- Stitching Density

2. Введите соответствующие значения в доступные текстовые поля.

Пример значений вы можете найти в изображении выше.

Предсказание свойств материала

Matrix Filler Ratio:

Density (kg/m³):

Elastic Modulus (GPa):

Hardener Amount (%):

Epoxy Groups Content (%):

Flash Point Temp (°C):

Surface Density (g/m²):

Resin Consumption (g/m²):

Stitching Angle (°):

Stitching Step:

Stitching Density:

Предсказанный модуль упругости (GPa): 0.023067532485221472

Предсказанная прочность (MPa): 0.6973476673399778

Рисунок 18 – Пример предсказания и значений

Шаг 3: Предсказание свойств материала

1. После ввода всех необходимых параметров нажмите кнопку "Предсказать".
2. Через несколько секунд предсказанные значения модуля упругости (GPa) и прочности (MPa) будут отображены в чат-боте.

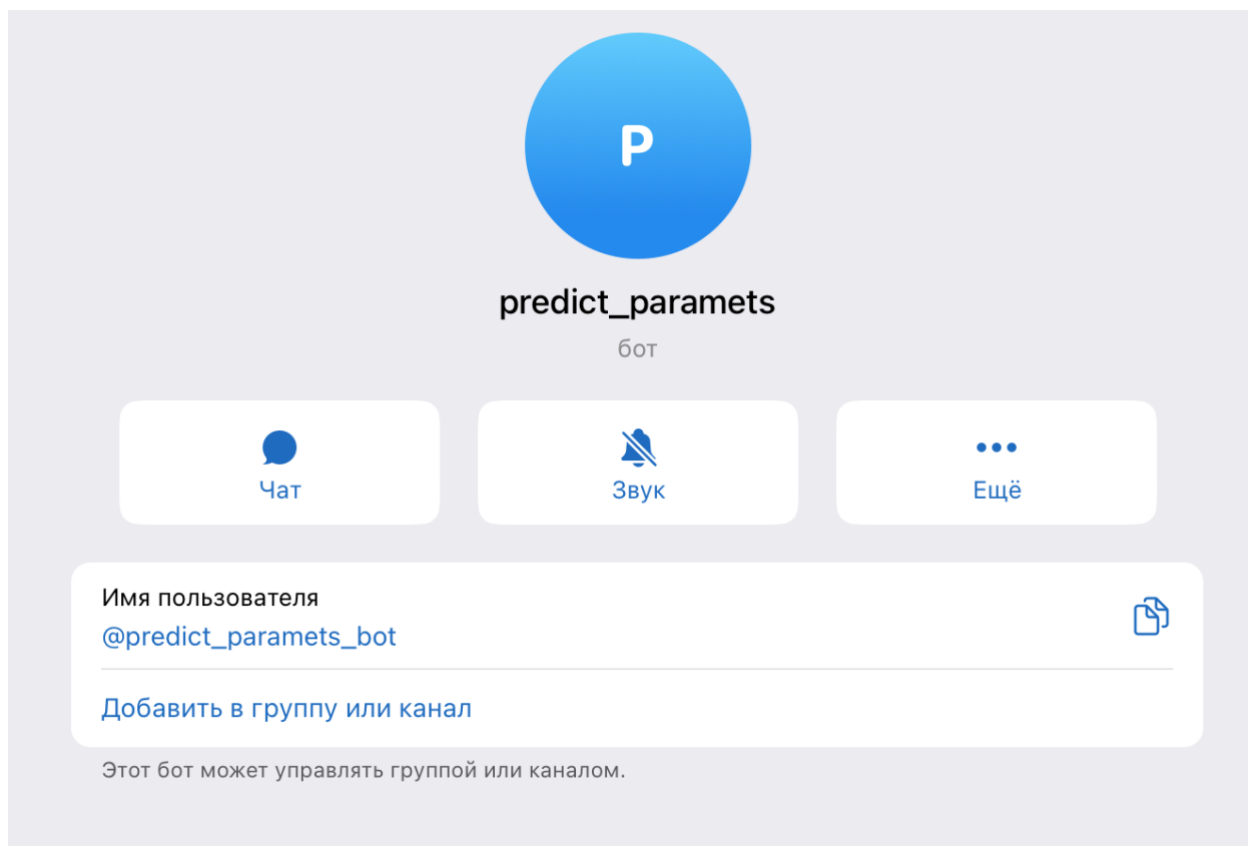


Рисунок 20 – Чат-бот

Замечания

- Убедитесь, что введенные значения корректны и соответствуют допустимым диапазонам.
- Если данные введены некорректно или неполно, предсказание может быть неверным, либо система выдаст сообщение об ошибке.

Пример результат:

После нажатия кнопки "Предсказать" приложение может выдать предсказанные значения, как показано ниже:

Предсказанный модуль упругости (GPa): 0.023067532485221472

Предсказанная прочность (MPa): 0.6973476673399778

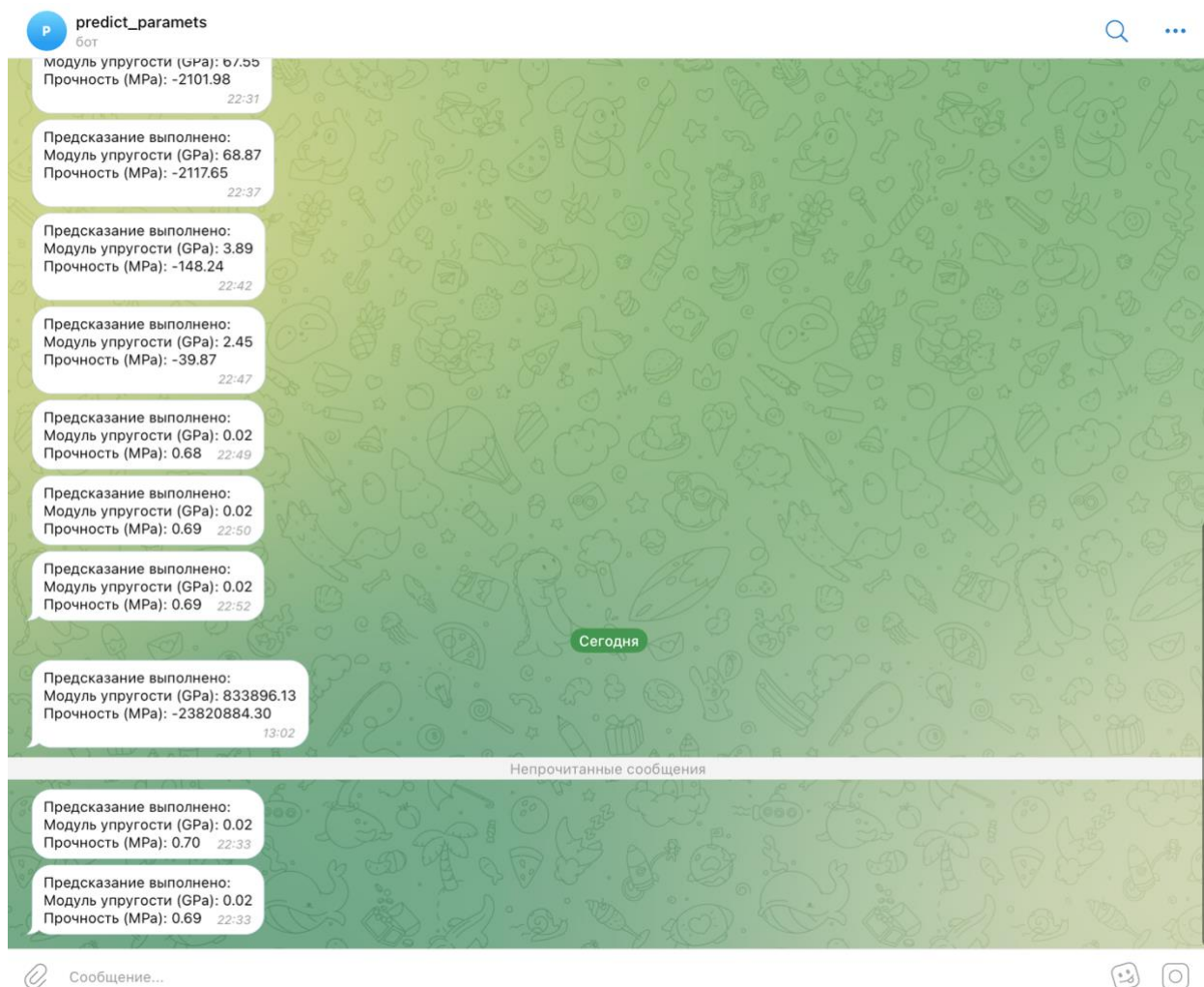


Рисунок 19 – Пример уведомлений в чат-боте

