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 Никифоров Денис Николаевич

Москва, 2023

Оглавление

Введение

Тема данной выпускной квалификационной работы – прогнозирование свойств новых композиционных материалов.

Композиционные материалы — это материалы, состоящие из двух или более различных компонентов, которые объединены вместе для создания новых свойств и характеристик. Они широко используются в различных отраслях, таких как авиация, автомобильная промышленность, строительство, спортивные товары и даже медицина.

Одним из основных преимуществ композиционных материалов является их высокая прочность при небольшом весе. Они обладают высокой относительной прочностью и жесткостью, что делает их идеальными для использования в конструкциях, где требуется легкий, но прочный материал. Например, в авиационной промышленности композиты широко применяются для создания крыльев и фюзеляжей самолетов.

Композиционные материалы также обладают высокой устойчивостью к коррозии и химическому воздействию, что делает их долговечными и надежными в различных условиях эксплуатации. Они также обладают отличными тепло- и звукоизоляционными свойствами.

Одним из наиболее распространенных типов композиционных материалов являются стеклопластиковые композиты. Они состоят из стекловолокна, которое встроено в матрицу из полимерного связующего вещества, такого как эпоксидная смола. Стеклопластиковые композиты обладают высокой прочностью, устойчивостью к ударам и хорошей электроизоляцией.

Еще одним примером композиционных материалов являются углепластиковые композиты. Они состоят из углеродного волокна, которое также встроено в полимерную матрицу. Углепластиковые композиты обладают очень высокой прочностью и жесткостью, а также низким коэффициентом теплового расширения.

Композиционные материалы имеют огромный потенциал для развития и применения в различных областях. Исследования и разработки в этой области продолжаются, и ожидается, что в будущем мы увидим еще более инновационные и эффективные композиционные материалы.

Машинное обучение — это область искусственного интеллекта, которая изучает разработку алгоритмов и моделей, которые позволяют компьютерам обучаться и делать прогнозы на основе данных. В последние годы машинное обучение стало все более популярным и широко используется в различных отраслях, включая материаловедение.

Прогнозирование свойств композитных материалов является сложной задачей, поскольку они зависят от множества факторов, таких как состав материала, структура, процессы изготовления и условия эксплуатации. Традиционные методы прогнозирования, такие как эксперименты и моделирование, могут быть дорогостоящими и трудоемкими.

Машинное обучение предлагает новый подход к прогнозированию свойств композитных материалов. С его помощью можно создать модели, которые могут анализировать большие объемы данных и выявлять скрытые закономерности и связи между различными параметрами материала. Эти модели могут использоваться для прогнозирования свойств композитных материалов на основе имеющихся данных.

Одним из примеров применения машинного обучения в прогнозировании свойств композитных материалов является использование нейронных сетей. Нейронные сети — это модели, которые имитируют работу человеческого мозга и состоят из множества связанных нейронов. Они могут быть обучены на основе данных о свойствах композитных материалов и использоваться для прогнозирования этих свойств для новых материалов.

Другой подход — это использование алгоритмов машинного обучения, таких как случайные леса или градиентный бустинг, для создания моделей прогнозирования свойств композитных материалов. Эти алгоритмы могут

анализировать множество факторов, включая состав материала, структуру и процессы изготовления, и предсказывать свойства материала на основе этих данных.

Преимущества использования машинного обучения в прогнозировании свойств композитных материалов включают более быструю и эффективную оценку свойств материалов, снижение затрат на эксперименты и моделирование, а также возможность прогнозирования свойств для новых материалов, основываясь на имеющихся данных.

Однако, важно отметить, что машинное обучение не является универсальным решением для всех проблем прогнозирования свойств композитных материалов. Оно требует качественных данных для обучения моделей и может столкнуться с ограничениями в случае отсутствия достаточного объема данных или сложных зависимостей между параметрами материала.

В целом, машинное обучение представляет собой мощный инструмент, который может помочь в прогнозировании свойств композитных материалов. Оно может быть использовано для создания моделей, которые могут анализировать данные и предсказывать свойства материалов на основе имеющихся данных. Это может ускорить процесс разработки новых материалов и повысить их производительность и надежность. Однако, для достижения оптимальных результатов необходимо провести дальнейшие исследования и разработки в этой области.

1. Аналитическая часть

1.1. Постановка задачи

Цель данной работы заключается в использовании методов машинного обучения для прогнозирования характеристик композитных материалов. Требуется дать прогноз зависимости модуля упругости при растяжении и прочности при растяжении, а также рекомендацию оптимального соотношения матрицы и наполнителя для этих материалов. В рамках работы будут построены модели, основанные на данных о свойствах материалов, с использованием методов машинного обучения. Это позволит ускорить процесс прогнозирования свойств материалов и предложить оптимальные соотношения матрицы и наполнителя для достижения желаемых характеристик.

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

Файл X_bp.xlsx содержит 1023 строки, индекс и 10 признаков.

Файл X_nup.xlsx содержит 1040 строк индекс и 3 признака.

1.2. Описание используемых методов

Для решения данной задачи регрессии было использовано несколько методов машинного обучения:

- DummyRegressor – регрессия по простым правилам
- LinearRegression - Линейная регрессия;
- Ridge - Гребневая регрессия;
- Lasso - Лассо (Лассо-регрессия);
- SVR - Метод опорных векторов для регрессии;
- KNeighborsRegressor – Метод K- ближайших соседей
- DecisionTreeRegressor - Регрессионное дерево решений.
- RandomForestRegressor - Случайный лес регрессии
- GradientBoostingRegressor - Градиентный бустинг

DummyRegressor. Этот метод используется как базовая модель для сравнения с другими моделями. Он предсказывает среднее значение целевой переменной в обучающем наборе данных.

Линейная регрессия (LinearRegression). Этот метод использует линейную регрессию для построения модели прогнозирования. Он предполагает линейную зависимость между признаками и целевой переменной.

Гребневая регрессия (Ridge) Этот метод также использует линейную регрессию, но добавляет штраф к большим значениям коэффициентов модели. Это помогает уменьшить переобучение и улучшить обобщающую способность модели.

Лассо (Лассо-регрессия) (Lasso). Этот метод также использует линейную регрессию, но добавляет L1-регуляризацию, которая приводит к разреженным моделям (некоторые коэффициенты становятся нулевыми). Это может быть полезно для отбора признаков.

Метод опорных векторов для регрессии (SVR). Этот метод использует метод опорных векторов для регрессии. Он строит гиперплоскость, которая

максимально близка к наибольшему количеству образцов, при этом допускается некоторое количество ошибок.

Метод K- ближайших соседей (KNeighborsRegressor). Этот метод использует метод ближайших соседей для регрессии. Он предсказывает значение целевой переменной на основе значений ближайших образцов.

Регрессионное дерево решений (DecisionTreeRegressor). Этот метод использует решающее дерево для регрессии. Он делит данные на подмножества на основе значений признаков и строит дерево принятия решений для прогнозирования значений целевой переменной.

Случайный лес регрессии (RandomForestRegressor). Этот метод использует ансамбль случайных деревьев для регрессии. Он комбинирует прогнозы нескольких деревьев, чтобы получить более точные прогнозы.

Градиентный бустинг (GradientBoostingRegressor) это метод машинного обучения, который использует ансамбль деревьев решений для предсказания значений целевой переменной. Он работает путем последовательного обучения слабых моделей (деревьев решений) на остатках предыдущих моделей. Каждая новая модель пытается улучшить ошибки предыдущих моделей, взвешивая их остатки. В результате получается сильная модель, способная предсказывать целевую переменную с высокой точностью.

1.3. Разведочный анализ данных

Цель разведочного анализа данных — выявить закономерности в данных. Для корректной работы большинства моделей желательна сильная зависимость выходных переменных от входных и отсутствие зависимости между входными переменными.

В данном проекте были использованы следующие методы разведочного анализа данных:

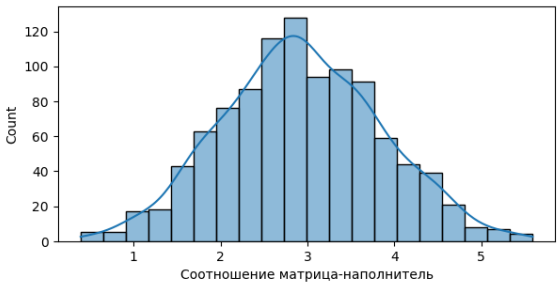
- Описательная статистика данного датасета.
- Визуальный анализ гистограмм12
- Визуальный анализ диаграмм размаха («ящик с усами»)
- Проверка нормальности распределения по критерию Пирсона
- Анализ попарных графиков рассеяния переменных
- Корреляционный анализ с целью поиска коэффициентов

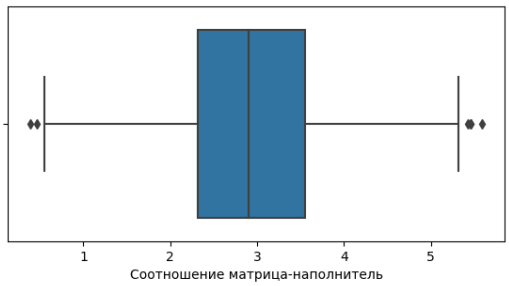
Для обработки данных объединим два файла по методу INNER.

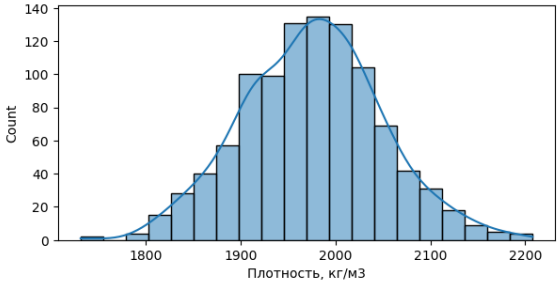
	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град
0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0	0
1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0	0
2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0	0
3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0	0
4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0	0
5	2.767918	2000.0	748.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0	0
6	2.569620	1910.0	807.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0	0
7	2.561475	1900.0	535.000000	111.86	22.267857	284.615385	380.0	75.0	1800.0	120.0	0
8	3.557018	1930.0	889.000000	129.00	21.250000	300.000000	380.0	75.0	1800.0	120.0	0
9	3.532338	2100.0	1421.000000	129.00	21.250000	300.000000	1010.0	78.0	2000.0	300.0	0

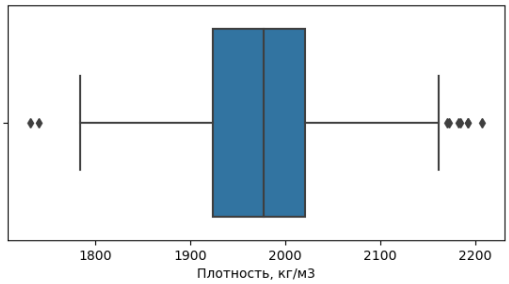
Рис1. Данные датасета после объединения

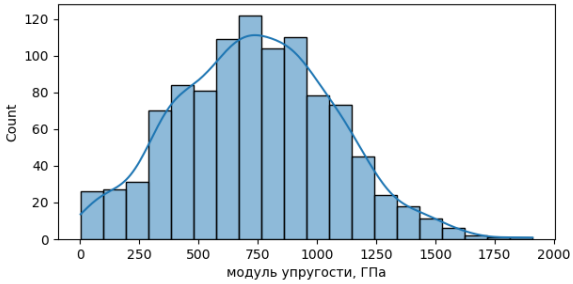
Построим гистограмму для каждой переменной и диаграммы ящика с усами.

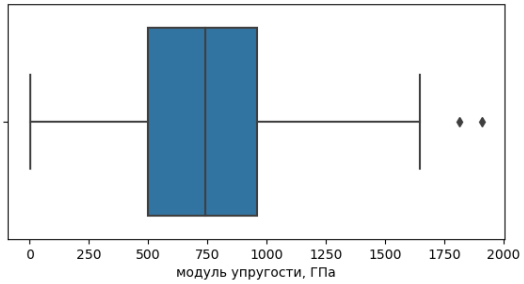

Count
Соотношение матрица-наполнитель

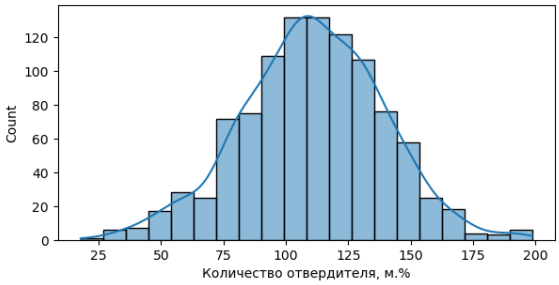

Соотношение матрица-наполнитель

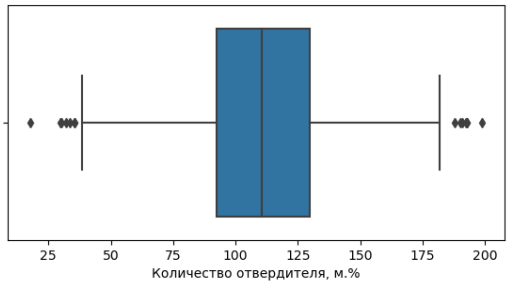

Count
Плотность, кг/м3

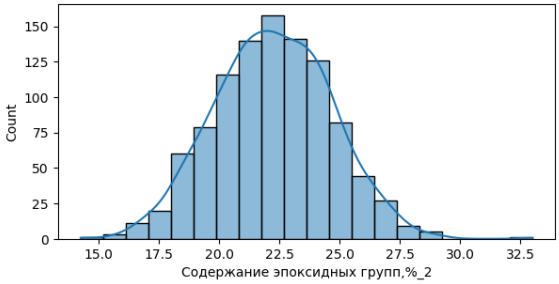

Плотность, кг/м3

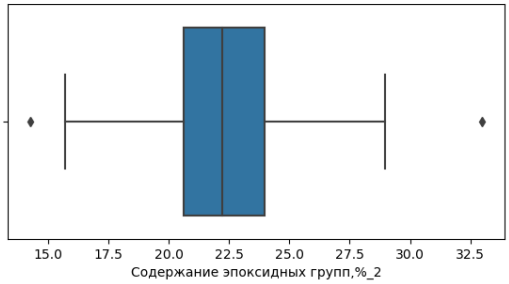

Count
модуль упругости, ГПа

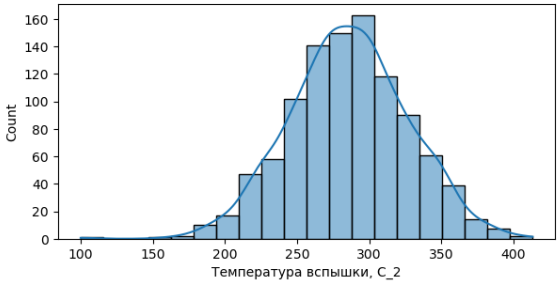

модуль упругости, ГПа

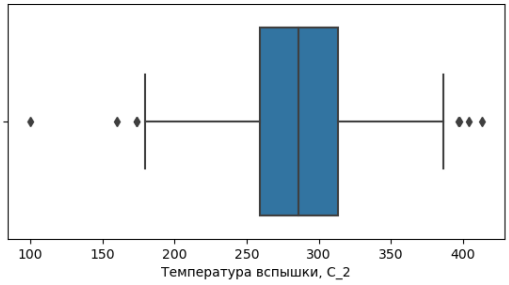

Count
Количество отвердителя, м.%

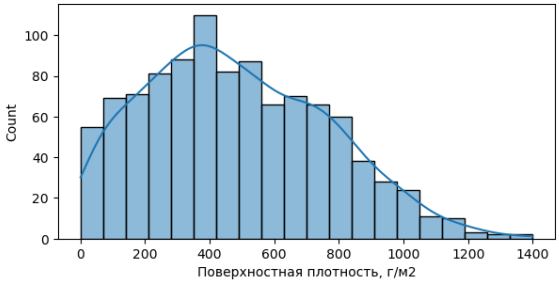

Количество отвердителя, м.%

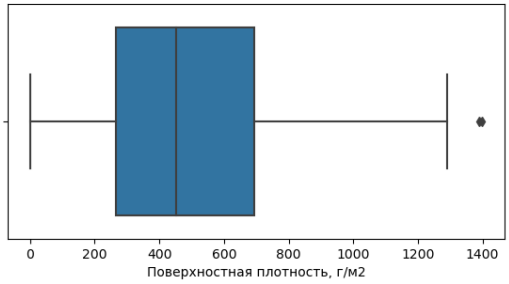

Count
Содержание эпоксидных групп,%_2


Содержание эпоксидных групп,%_2


Count
Температура вспышки, С_2


Температура вспышки, С_2


Count
Поверхностная плотность, г/м2


Поверхностная плотность, г/м2

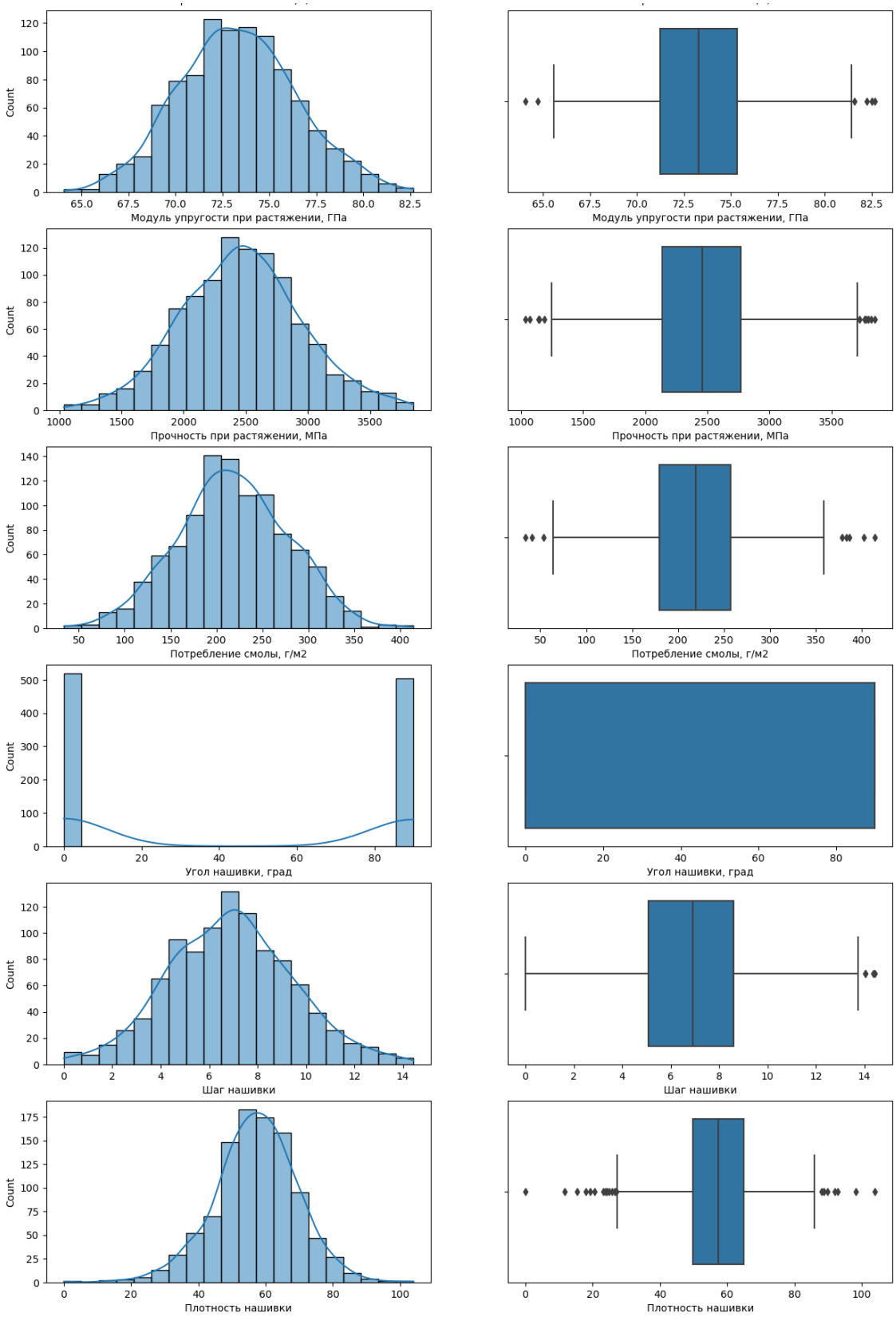


Рис2 Гистограмма распределения и «ящики с усами»

Выведем основную статистическую информацию.

	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

Рис3. Статистическая информация

Построим попарные графики рассеяния точек.

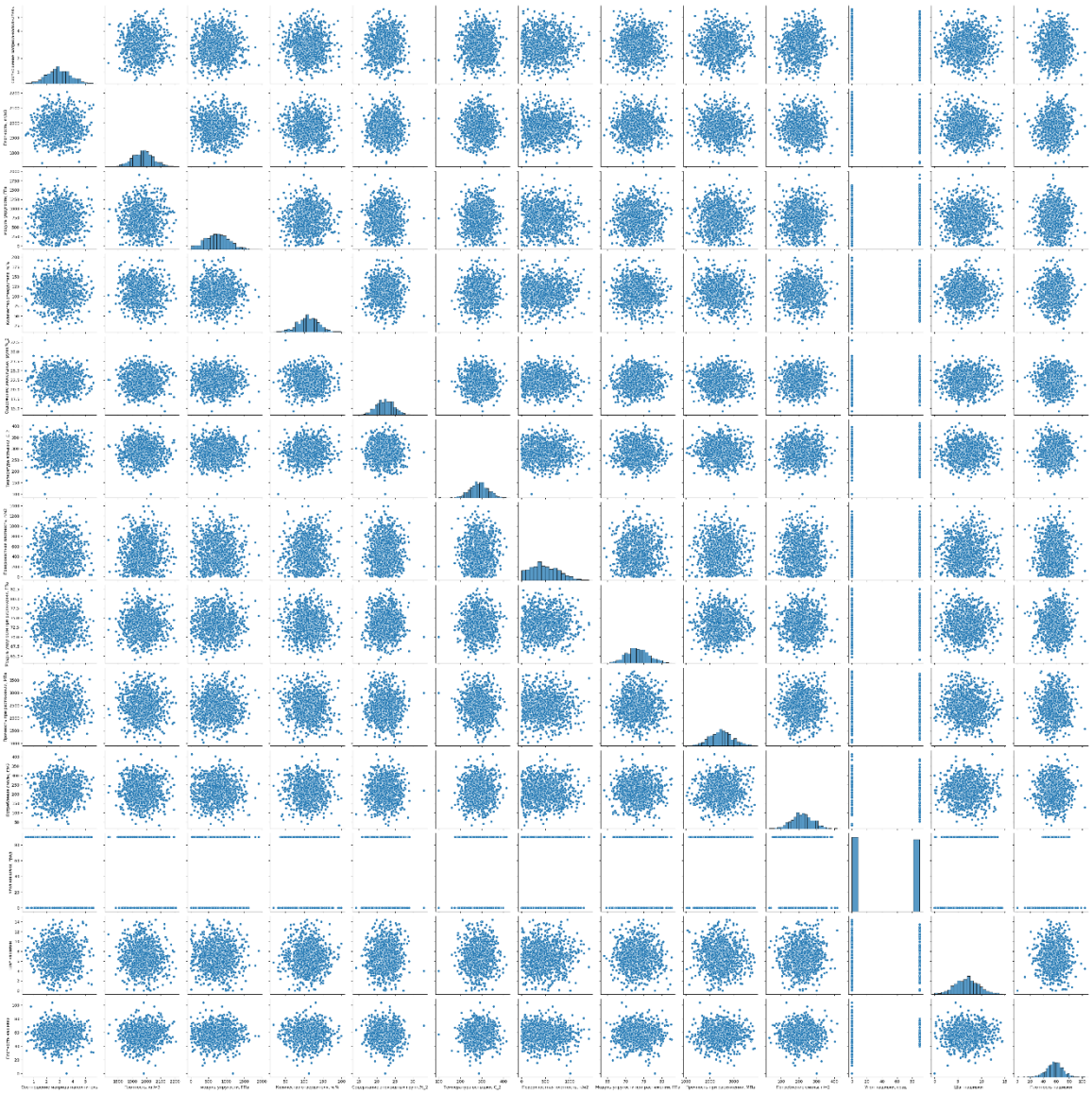


Рис4. Графики рассеяния точек.

Проведем анализ и исключение выбросов. Применим два метода для работы с выбросами:

- Метод 3-х сигм. основан на стандартном отклонении данных от среднего значения. Согласно правилу трех сигм, большинство данных должны находиться в пределах трех стандартных отклонений от среднего значения.

- Метод межквартильных расстояний — это разница между 75-м процентилем (Q3) и 25-м процентилем (Q1) в наборе данных. Он измеряет разброс средних 50% значений. Выброс есть, если он в 1,5 раза превышает межквартильный размах, превышающий третий квартиль (Q3), или в 1,5 раза превышает межквартильный размах, меньше первого квартиля (Q1).

Метод межквартильных расстояний следует использовать если распределение данных отличается от нормального, есть перекос в правую или левую сторону.

Изначально метод 3-х сигм нашел 24 выброса по всем столбцам датасета, а метод межквартильных расстояний - 93 выброса. Применяя метод межквартильных расстояний несколько раз количество выбросов сведено к 0.

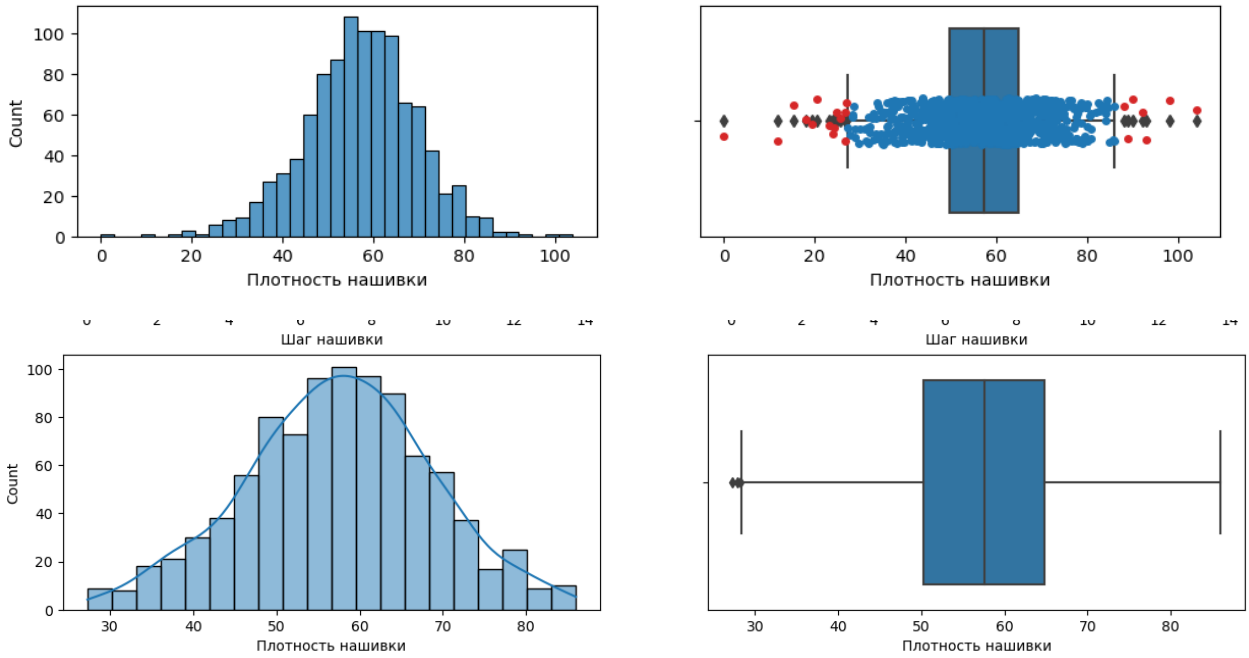


Рис5. Гистограмма и «ящик с усами» до и после очистки для Плотности нашивки.

Теперь проверим датасет на пропуски и убедимся, что их нет.

```
Соотношение матрица-наполнитель          0
Плотность, кг/м3                         0
модуль упругости, ГПа                    0
Количество отвердителя, м.%              0
Содержание эпоксидных групп,%_2          0
Температура вспышки, С_2                 0
Поверхностная плотность, г/м2            0
Модуль упругости при растяжении, ГПа     0
Прочность при растяжении, МПа            0
Потребление смолы, г/м2                  0
Угол нашивки, град                       0
Шаг нашивки                              0
Плотность нашивки                        0
dtype: int64
```

Рис6. Проверка на пропуски

Для визуализации коэффициентов корреляции и определения наличия между переменными зависимости, была построена тепловая карта коэффициентов корреляции методом Пирсона.

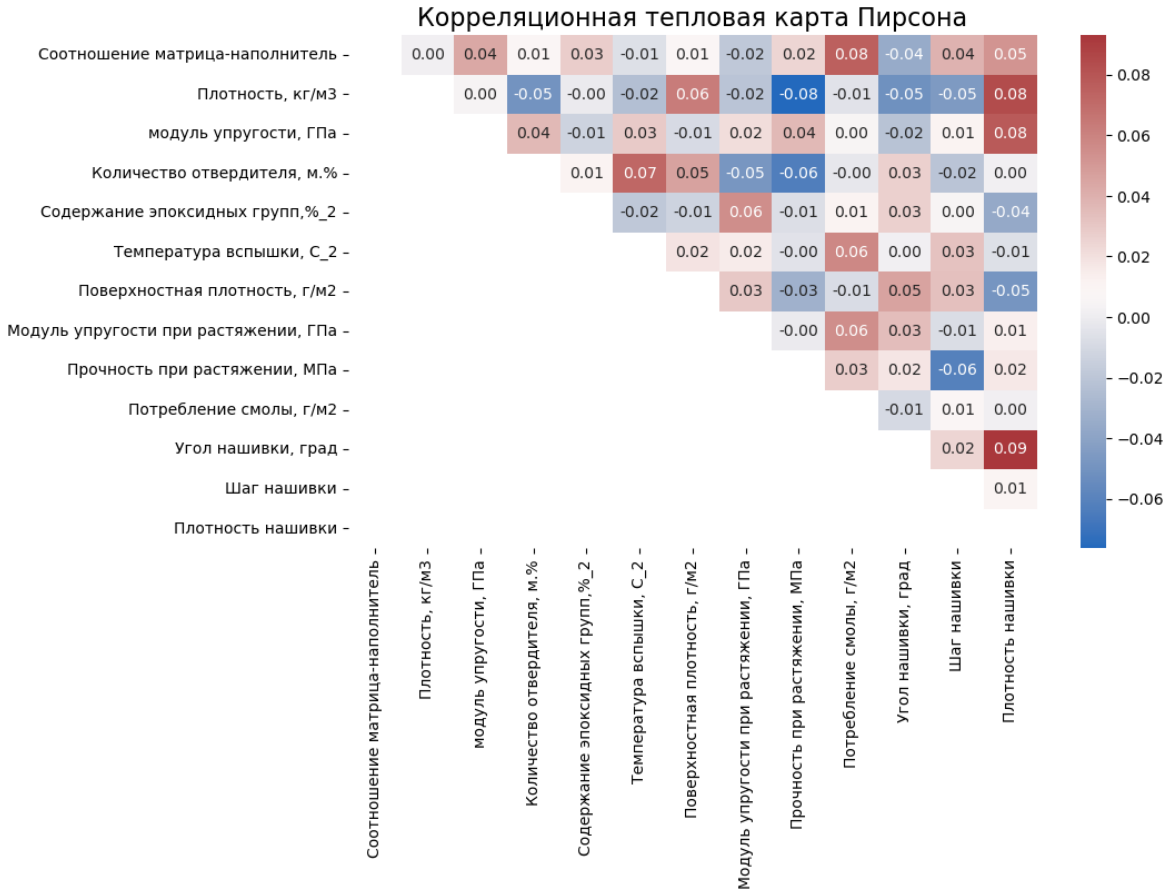


Рис6. Корреляция Пирсона

Корреляция Пирсона — это статистический показатель, который используется для измерения степени линейной связи между двумя переменными. Она измеряет силу и направление линейной связи между двумя непрерывными переменными.

Коэффициент корреляции Пирсона принимает значения от -1 до 1.

Значение -1 указывает на полную обратную линейную связь, значение 1 указывает на полную прямую линейную связь, а значение 0 указывает на отсутствие линейной связи между переменными.

Корреляции между переменными отсутствует.

– «Модуль упругости при растяжении, ГПа»;

– «Прочность при растяжении, МПа»;

– «Соотношение матрица-наполнитель».

2. Практическая часть

2.1. Предобработка данных

Распределение исходных данных близко к нормальному распределению кроме Поверхностной плотности.

Также применим нормализацию данных.

Нормализация — это метод предварительной обработки данных, используемый для стандартизации или масштабирования числовых значений в определенном диапазоне. Это гарантирует, что разные переменные имеют сопоставимые масштабы, предотвращая доминирование одних переменных над другими во время анализа или моделирования.

MinMaxScaler — это метод предварительной обработки, используемый для масштабирования характеристик набора данных в указанном диапазоне (обычно от 0 до 1).

Формула для MinMax-нормализации следующая:

$x\_scaled = (x - x\_min) / (x\_max - x\_min)$

где:

x_scaled - отмасштабированное значение признака;

x - оригинальное значение признака;

x_min - минимальное значение признака в наборе данных;

x_max - максимальное значение признака в наборе данных.

После предварительной обработки данные приведены к состоянию удобному для дальнейшей обработки.

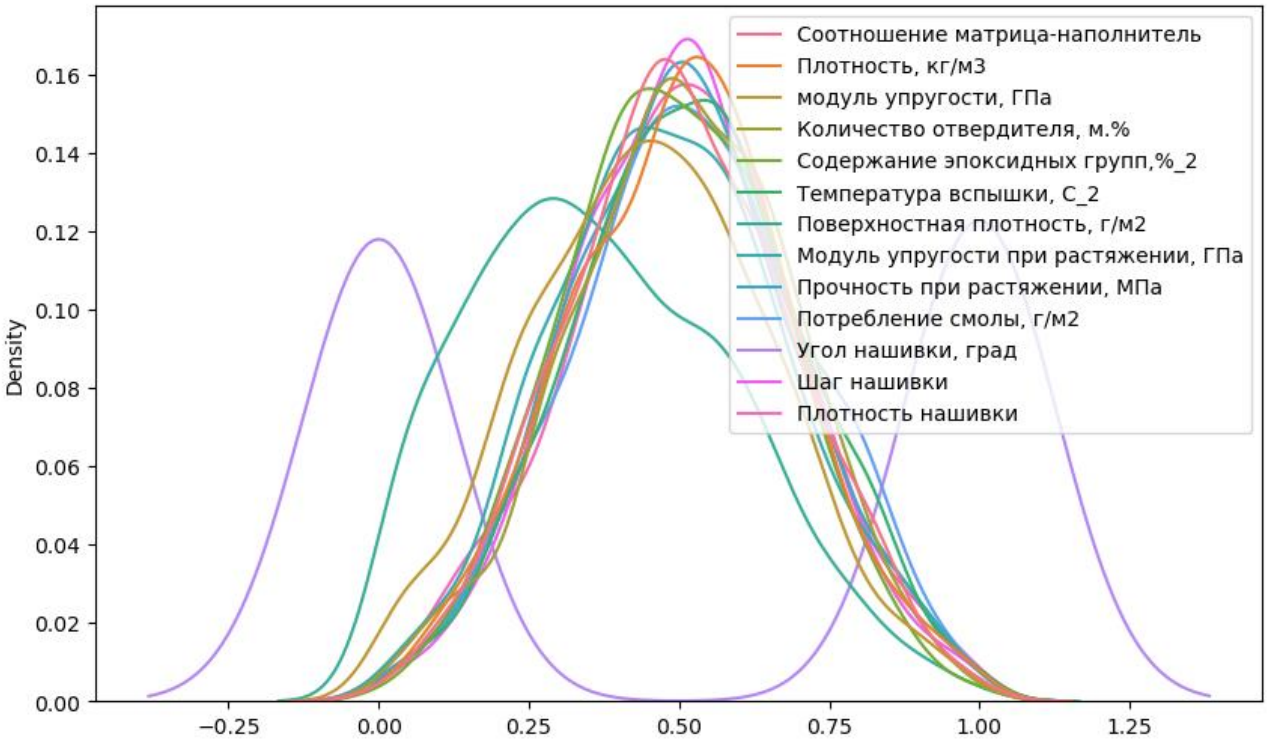


Рис7. Датасет после нормализации

Также проверим параметры датасета.

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	926.0	0.499687	0.187841	0.0	0.373077	0.494936	0.629774	1.0
Плотность, кг/м3	926.0	0.503041	0.188261	0.0	0.368184	0.511740	0.625266	1.0
модуль упругости, ГПа	926.0	0.451888	0.201469	0.0	0.305669	0.452599	0.587461	1.0
Количество отвердителя, м.%	926.0	0.506559	0.187611	0.0	0.378514	0.506532	0.639120	1.0
Содержание эпоксидных групп,%_2	926.0	0.491063	0.180438	0.0	0.367101	0.488912	0.623296	1.0
Температура вспышки, С_2	926.0	0.516443	0.190857	0.0	0.385988	0.516931	0.646553	1.0
Поверхностная плотность, г/м2	926.0	0.373626	0.216945	0.0	0.204863	0.356181	0.538397	1.0
Модуль упругости при растяжении, ГПа	926.0	0.487330	0.196140	0.0	0.353512	0.483718	0.617568	1.0
Прочность при растяжении, МПа	926.0	0.504018	0.189451	0.0	0.373350	0.501053	0.623037	1.0
Потребление смолы, г/м2	926.0	0.521132	0.194829	0.0	0.391647	0.523459	0.652734	1.0
Угол нашивки, град	926.0	0.509719	0.500176	0.0	0.000000	1.000000	1.000000	1.0
Шаг нашивки	926.0	0.503077	0.183709	0.0	0.372377	0.506414	0.626112	1.0
Плотность нашивки	926.0	0.505927	0.193315	0.0	0.379888	0.507146	0.631896	1.0

Рис8. Статистика датасета

2.2 Разработка, обучение и тестирование десяти моделей (включая нейросеть) для прогнозирования трёх целевых признаков по отдельности.

1. Прогнозирования признака «Модуль упругости при растяжении, ГПа»

Для решения задачи прогнозирования модуль упругости при растяжении, был использован метод машинного обучения, а именно обучение моделей на обучающей выборке и оценка их качества на тестовой выборке.

Перед обучением моделей данные были разделены на обучающую и тестовую выборки в соотношении 70/30. Это позволяет проверить качество работы моделей на независимых данных и избежать переобучения.

Для каждой модели был создан словарь с гиперпараметрами. Гиперпараметры - это настройки модели, которые не могут быть изучены из данных и должны быть установлены вручную. Для поиска лучших гиперпараметров был использован метод поиска по сетке с перекрестной проверкой. Этот метод позволяет оценить качество модели с различными комбинациями гиперпараметров и выбрать наилучшую комбинацию.

Качество работы каждой модели было оценено с помощью нескольких метрик. Одной из них является коэффициент детерминации (R^2), который показывает, насколько хорошо модель объясняет вариацию в данных. Чем ближе значение R^2 к 1, тем лучше модель объясняет данные.

Другой метрикой является среднеквадратическая ошибка (RMSE), которая измеряет среднее отклонение прогнозов модели от фактических значений. Чем меньше значение RMSE, тем лучше модель предсказывает данные.

Третьей метрикой является средняя абсолютная ошибка (MAE), которая измеряет среднее абсолютное отклонение прогнозов модели от фактических значений. Чем меньше значение MAE, тем лучше модель предсказывает данные.

Используя эти метрики, можно сравнить качество работы различных моделей и выбрать наилучшую модель для решения задачи.

Для прогнозирования модуля упругости при растяжении и прочности при растяжении были использованы следующие методы машинного обучения:

- DummyRegressor – регрессия по простым правилам
- LinearRegression - Линейная регрессия;
- Ridge - Гребневая регрессия;
- Lasso - Лассо (Лассо-регрессия);
- SVR - Метод опорных векторов для регрессии;
- KNeighborsRegressor – Метод K- ближайших соседей
- DecisionTreeRegressor - Регрессионное дерево решений.
- RandomForestRegressor - Случайный лес регрессии
- GradientBoostingRegressor - Градиентный бустинг

Результаты тестирования моделей с параметрами для прогнозирования модуля упругости при растяжении по умолчанию показывают неудовлетворительные результаты.

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.014369	-0.191387	-0.156417	-0.609281	-0.451870
LinearRegression	-0.032590	-0.193068	-0.158108	-0.610380	-0.453773
Ridge	-0.030940	-0.192918	-0.158001	-0.610288	-0.453167
Lasso	-0.014369	-0.191387	-0.156417	-0.609281	-0.451870
SVR	-0.358672	-0.220606	-0.178406	-0.676160	-0.572981
KNeighborsRegressor	-0.310141	-0.216813	-0.177070	-0.681401	-0.522064
DecisionTreeRegressor	-1.138144	-0.274442	-0.219973	-0.753242	-0.696741
RandomForestRegressor	-0.083913	-0.197681	-0.161187	-0.633771	-0.481651

Рис9 Модели с параметрами по умолчанию

Лучший из худших результат показал DummyRegressor и Lasso.

Поиск по сетке также показал неинформативность используемых моделей.

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=1500, positive=True, solver='lbfgs')	-0.014343	-0.191386	-0.156439	-0.609219	-0.451881
Lasso(alpha=0.005)	-0.014369	-0.191387	-0.156417	-0.609281	-0.451870
SVR(C=0.04, kernel='sigmoid')	-0.013755	-0.191319	-0.156381	-0.604445	-0.452127
KNeighborsRegressor(n_neighbors=29)	-0.059535	-0.195578	-0.159547	-0.622749	-0.472972
DecisionTreeRegressor(criterion='poisson', max_depth=2, max_features=5, random_state=3000)	-0.019018	-0.191589	-0.155039	-0.607660	-0.460805

Рис10 Модели с лучшими параметрами

Лучший из худших результат показал SVR.

Сравнение моделей и исходных данных. Базовая модель – DummyRegressor, лучшая модель- SVR.

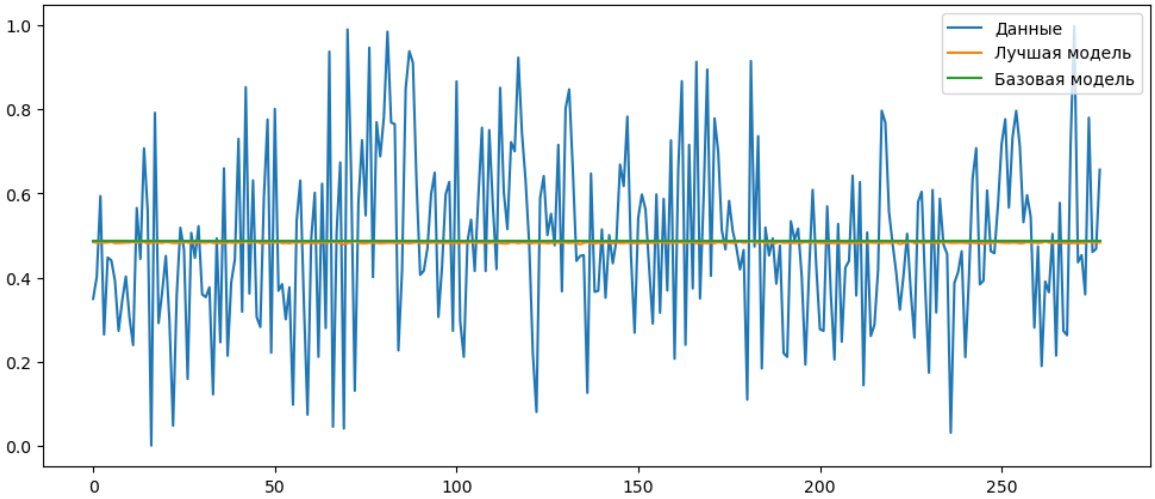


Рис11 Исходные данные по сравнению с моделями

2. Прогнозирования признака «Прочность при растяжении, МПа»

Результаты тестирования моделей с параметрами по умолчанию для предсказания прочности при растяжении показывают неудовлетворительные результаты.

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.019621	-0.190403	-0.151815	-3521152612319.554688	-0.469354
LinearRegression	-0.033334	-0.191600	-0.153487	-3927564128168.418945	-0.469799
Ridge	-0.031568	-0.191445	-0.153329	-3908985374430.334473	-0.469586
Lasso	-0.019621	-0.190403	-0.151815	-3521152612319.554688	-0.469354
SVR	-0.299595	-0.213581	-0.171366	-3621535248060.870117	-0.510943
DecisionTreeRegressor	-1.055530	-0.269714	-0.216114	-4088487506854.995117	-0.689723
GradientBoostingRegressor	-0.140626	-0.200972	-0.161459	-4855337034318.583984	-0.504240

Рис12 Модели с параметрами по умолчанию

Лучший из худших результат показал DummyRegressor и Lasso.

Поиск по сетке также показал отсутствие предсказательной силы используемых моделей.

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=990, positive=True, solver='lbfgs')	-0.020010	-0.190438	-0.151845	-3520303507813.850586	-0.469313
Lasso(alpha=0.1)	-0.019621	-0.190403	-0.151815	-3521152612319.554688	-0.469354
SVR(C=0.001, kernel='sigmoid')	-0.020827	-0.190505	-0.151952	-3504814566659.317383	-0.469531
DecisionTreeRegressor(criterion='absolute_error', max_depth=2, max_features=2, random_state=3000, splitter='random')	-0.014374	-0.189994	-0.151191	-3480337457527.190430	-0.468691

Рис13 Модели с лучшими параметрами

Лучший из худших результат показал DecisionTreeRegressor.

Сравнение моделей и исходных данных. Базовая модель – DummyRegressor, лучшая модель- DecisionTreeRegressor.

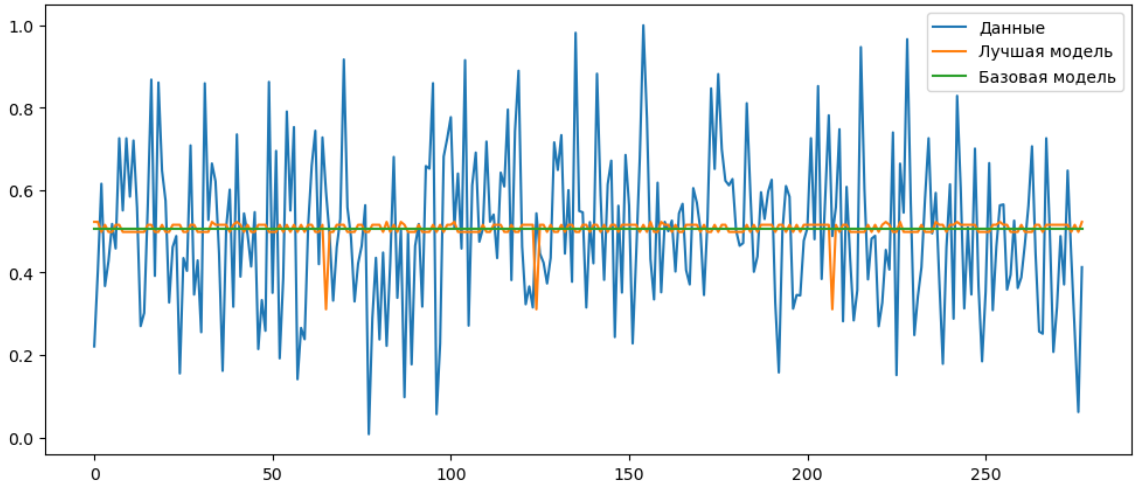


Рис14 Исходные данные по сравнению с моделями

2.3 Анализ полученных данных

Отсутствие результата использования регрессионных моделей показывает недостаточную предобработку данных.

Воспользуемся методом главных компонентов (principal component analysis, PCA) для понижения размерности.

Метод главных компонентов (Principal Component Analysis, PCA) является одним из наиболее распространенных методов снижения размерности данных. Он используется для извлечения наиболее информативных признаков из исходного набора данных путем проекции их на новое пространство меньшей размерности.

Основная идея метода главных компонентов заключается в том, чтобы найти линейную комбинацию исходных признаков, которая максимально сохраняет дисперсию данных. Эти комбинации называются главными компонентами. Первая главная компонента объясняет наибольшую долю дисперсии, вторая - следующую по величине долю, и так далее.

Для кластеризации воспользуемся методом k средних. Метод k-средних (k-means) является одним из наиболее популярных методов кластеризации данных. Он используется для разделения набора данных на заранее определенное количество кластеров.

Основная идея метода k-средних заключается в том, чтобы найти k центроидов, которые представляют собой точки в пространстве данных. Каждая точка данных присваивается к ближайшему центроиду, и это определяет принадлежность кластера. Центроиды обновляются путем вычисления среднего значения всех точек данных, принадлежащих кластеру. Этот процесс повторяется до тех пор, пока центроиды не стабилизируются или не будет достигнуто максимальное количество итераций.

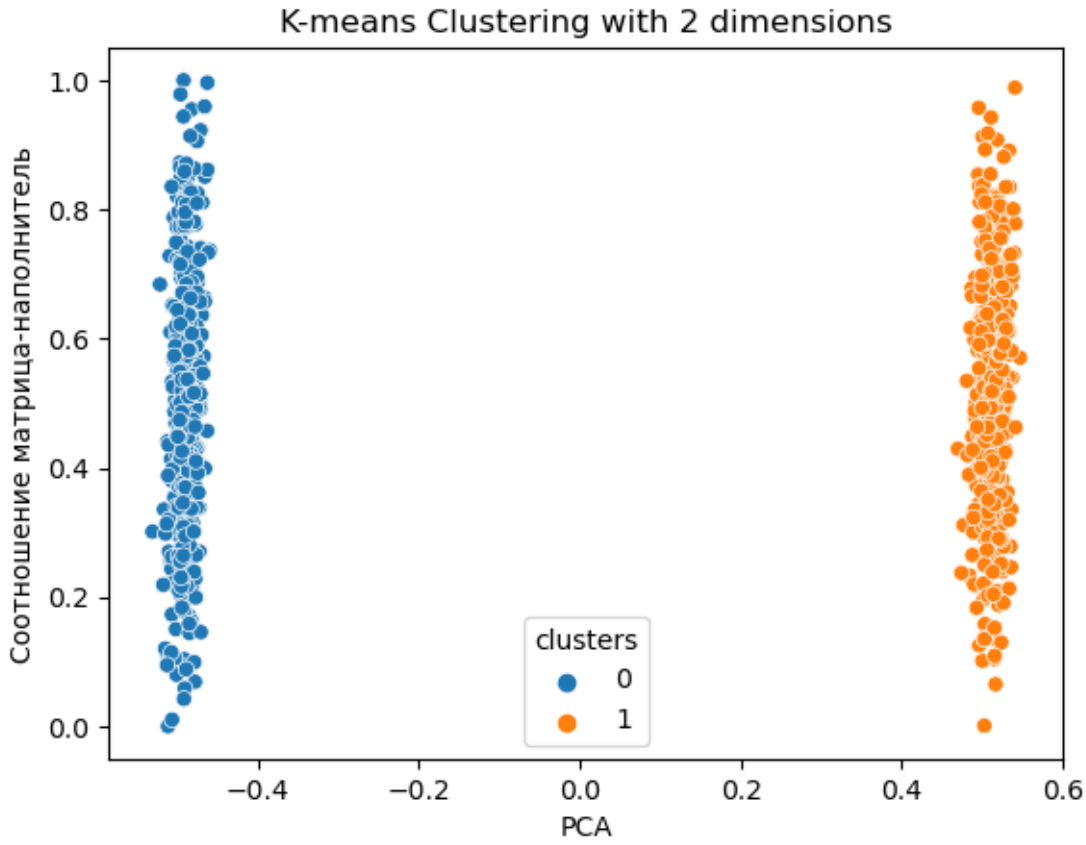


Рис16 Кластеризация данных

Кластеризация данных нашла 2 больших кластера со средним значением ±0.5 для PCA со средним числом элементов ~ 450.

Данные сильно зашумлены.

Вернемся к исходному датасету и выполним фильтрацию данных для удаления шумов.

Применим 2 фильтра: Медианный фильтр и фильтр Кальмана.

Медианный фильтр может быть применен для обработки значений в датасете. В этом случае, каждый элемент датасета заменяется медианой значений в его окрестности.

Медианный фильтр для обработки значений в датасете может быть полезным для удаления выбросов или экстремальных значений, которые могут исказить статистические показатели или анализ данных. Он также может помочь сохранить общую структуру данных и уменьшить влияние выбросов на результаты анализа.

Фильтр Кальмана - это рекурсивный алгоритм, который используется для оценки состояния динамической системы на основе наблюдаемых данных. Он может быть применен для обработки значений в датасете, чтобы улучшить точность и качество этих данных.

Фильтр Кальмана обладает рядом преимуществ при обработке значений в датасете. Он способен учитывать неопределенность и шум в данных, что позволяет получать более точные оценки состояния системы. Он также может работать с большими объемами данных и обрабатывать данные в реальном времени.

Используя методы, описанные в разделе разведочного анализа данных получим следующие скорректированные значения

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.925079	0.231656	1.857143	2.786194	2.917217	3.074718	3.555675
Плотность, кг/м3	1023.0	1976.031790	20.641883	1915.730538	1961.248202	1976.596851	1989.398771	2032.848359
модуль упругости, ГПа	1023.0	740.487319	87.664301	512.300769	679.282435	736.971162	799.781368	1068.524569
Количество отвердителя, м.%	1023.0	110.128929	9.113668	30.000000	105.605022	110.715337	115.714253	127.859051
Содержание эпоксидных групп,%_2	1023.0	22.250829	0.669969	20.562701	21.798977	22.219439	22.640956	24.493499
Температура вспышки, С_2	1023.0	284.954342	15.056054	100.000000	277.769424	285.937802	292.726004	319.156969
Поверхностная плотность, г/м2	1023.0	481.429707	79.731006	210.000000	430.896370	480.681307	537.575352	677.173335
Модуль упругости при растяжении, ГПа	1023.0	73.316299	0.953378	70.000000	72.662669	73.372630	73.994907	75.518638
Прочность при растяжении, МПа	1023.0	2470.724597	133.469877	2107.062231	2375.576141	2471.756760	2554.717679	3000.000000
Потребление смолы, г/м2	1023.0	218.522323	16.472395	170.365963	206.458787	218.130116	230.117764	255.496370
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1023.0	6.885627	0.688419	4.000000	6.478305	6.912483	7.320709	8.726775
Плотность нашивки	1023.0	57.117400	3.088990	45.234922	55.104290	57.370752	59.228941	65.513004

Рис17. Статистическая информация

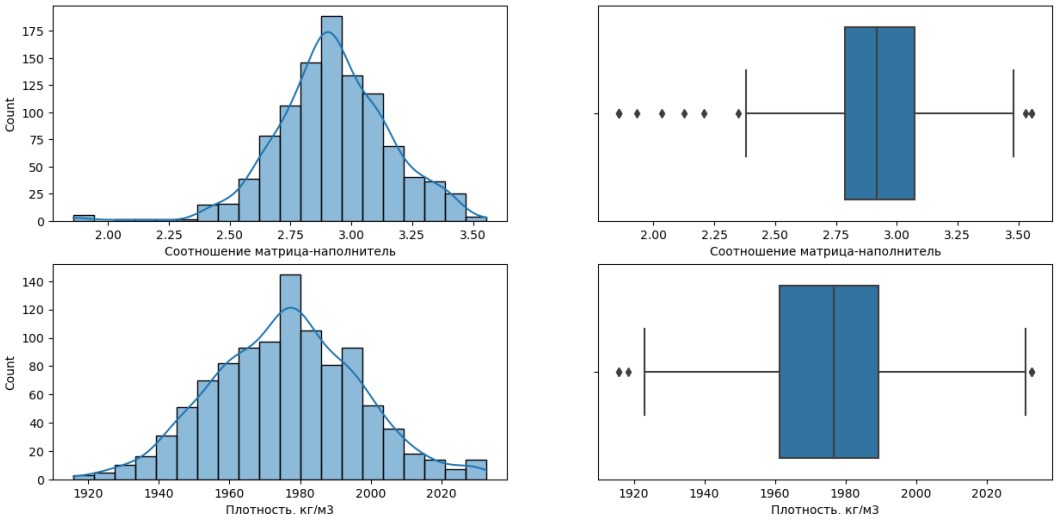


Рис18 Гистограмма распределения и «ящики с усами»

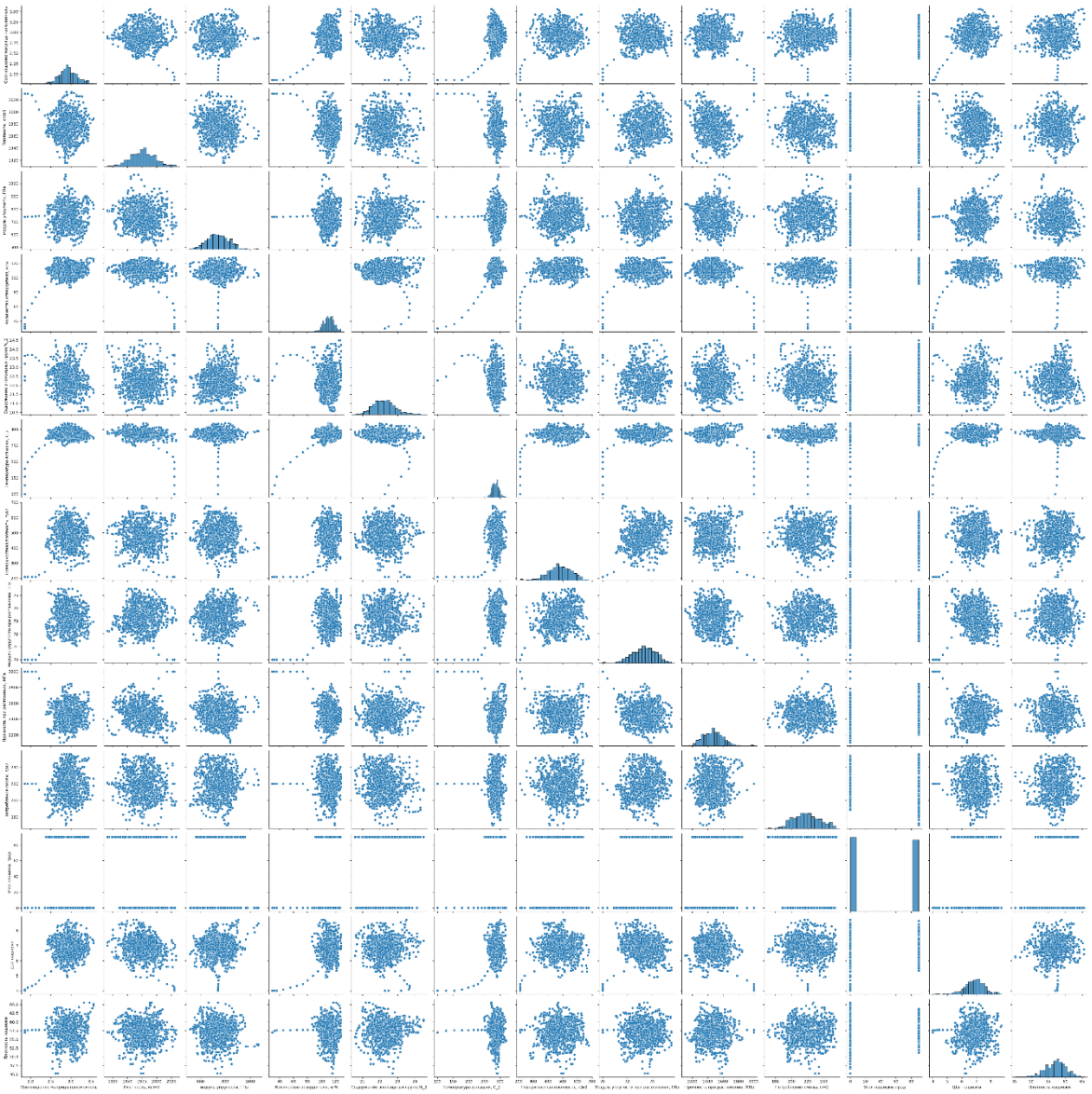
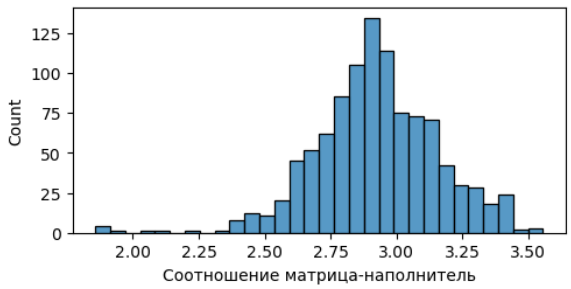
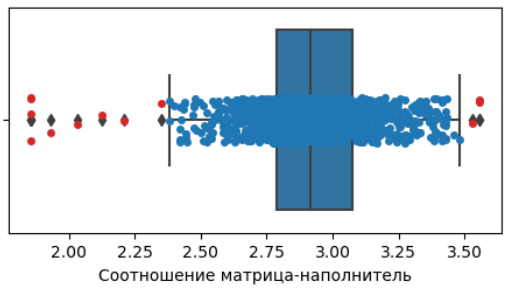


Рис19. Графики рассеяния точек.

Как видно из графика рассеяния точек после фильтрации данных появилось больше выбросов.

После удаления выбросов данные пригодны для дальнейшей обработки.





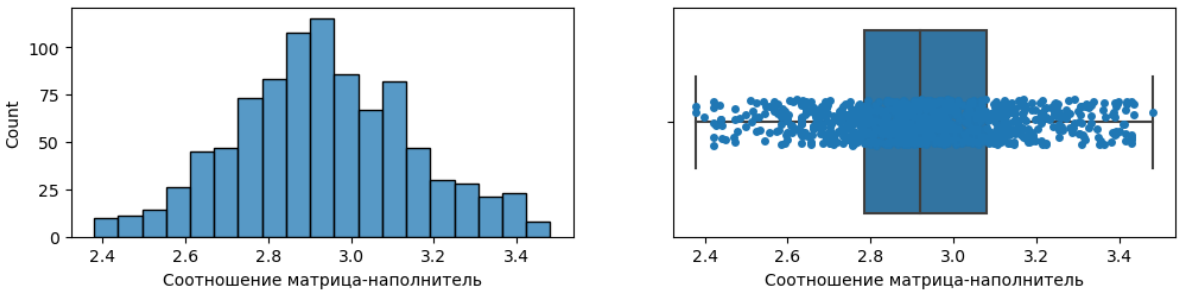


Рис20. Гистограмма и «ящик с усами» до и после очистки

Тепловая карта Пирсона показывает более сильную корреляцию некоторых параметров

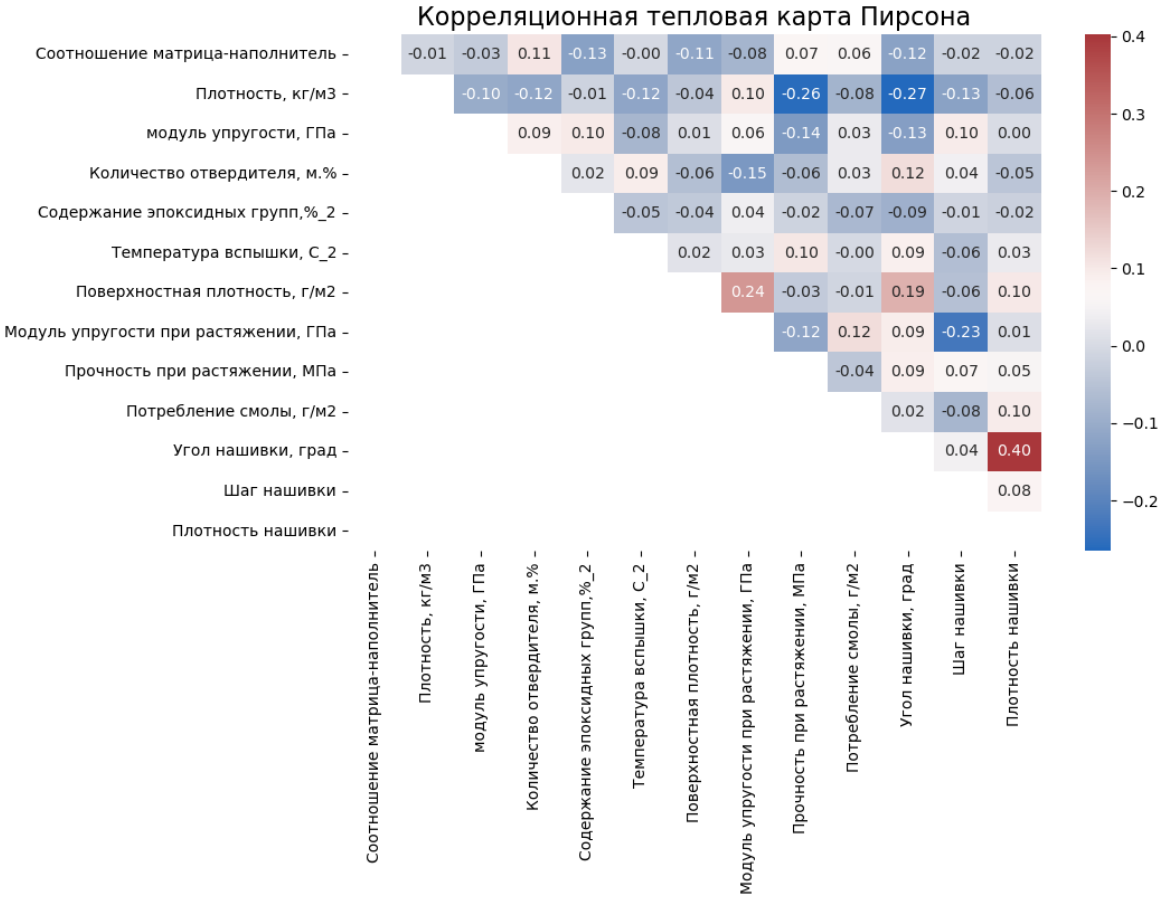


Рис21. Корреляция Пирсона

Регрессионные модели после удаления шума показывают результат лучше чем до обработки.

Лучший результат для модели модуль упругости у KNeighborsRegressor.

Лучший результат для модели прочность при растяжении у SVR.

Теперь можно переходить к построению нейронных сетей.

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.023790	-0.206570	-0.168089	-0.663950	-0.482383
LinearRegression	0.127503	-0.190622	-0.156021	-0.548095	-0.446032
Ridge	0.128356	-0.190538	-0.155920	-0.551555	-0.443519
Lasso	-0.023790	-0.206570	-0.168089	-0.663950	-0.482383
SVR	0.702189	-0.111048	-0.089952	-0.289343	-0.387441
KNeighborsRegressor	0.664991	-0.117629	-0.087594	-0.281993	-0.377483
DecisionTreeRegressor	0.131435	-0.189391	-0.120376	-0.363056	-0.687412
RandomForestRegressor	0.608477	-0.127583	-0.098812	-0.360453	-0.377803

Рис22 Модели модуль упругости с параметрами по умолчанию

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=10, solver='lsqr')	0.117797	-0.191737	-0.156796	-0.575752	-0.438060
Lasso(alpha=0.001)	0.117717	-0.191701	-0.156220	-0.564403	-0.449054
SVR(C=0.02, kernel='poly')	0.405342	-0.156877	-0.125619	-0.421175	-0.430320
KNeighborsRegressor(n_neighbors=3)	0.798118	-0.091141	-0.062416	-0.190646	-0.330939
DecisionTreeRegressor(criterion='absolute_error', max_depth=3, max_features=8, random_state=3000)	0.042181	-0.199153	-0.159513	-0.583590	-0.492044

Рис23 Модели модуль упругости с лучшими параметрами

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.014977	-0.184895	-0.150164	-3521621176373.428223	-0.439286
LinearRegression	0.084826	-0.175164	-0.141750	-2540445240983.733887	-0.433362
Ridge	0.086096	-0.175068	-0.141686	-2594856255449.057617	-0.430118
Lasso	-0.014977	-0.184895	-0.150164	-3521621176373.428223	-0.439286
SVR	0.647987	-0.108660	-0.085982	-1360307764636.201904	-0.330648
DecisionTreeRegressor	0.064370	-0.177365	-0.119034	-3047049950003.668945	-0.576177
GradientBoostingRegressor	0.396265	-0.142353	-0.112267	-2514647420319.952637	-0.389055

Рис24 Модели прочность при растяжении с параметрами по умолчанию

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=10, solver='lsqr')	0.078749	-0.175455	-0.141922	-2896413966050.510254	-0.408503
Lasso(alpha=0.1)	-0.016847	-0.184674	-0.150025	-3536279030688.050781	-0.425564
SVR(C=0.3)	0.512793	-0.127459	-0.100090	-1602416386036.175293	-0.348134
DecisionTreeRegressor(criterion='absolute_error', max_depth=3, max_features=5, random_state=3000)	0.080539	-0.175126	-0.138605	-2083710653417.593750	-0.421304
GradientBoostingRegressor(max_features=5, n_estimators=150, random_state=3000)	0.461828	-0.134253	-0.105258	-2106608649504.983643	-0.391175

Рис25 Модели прочность при растяжении с лучшими параметрами

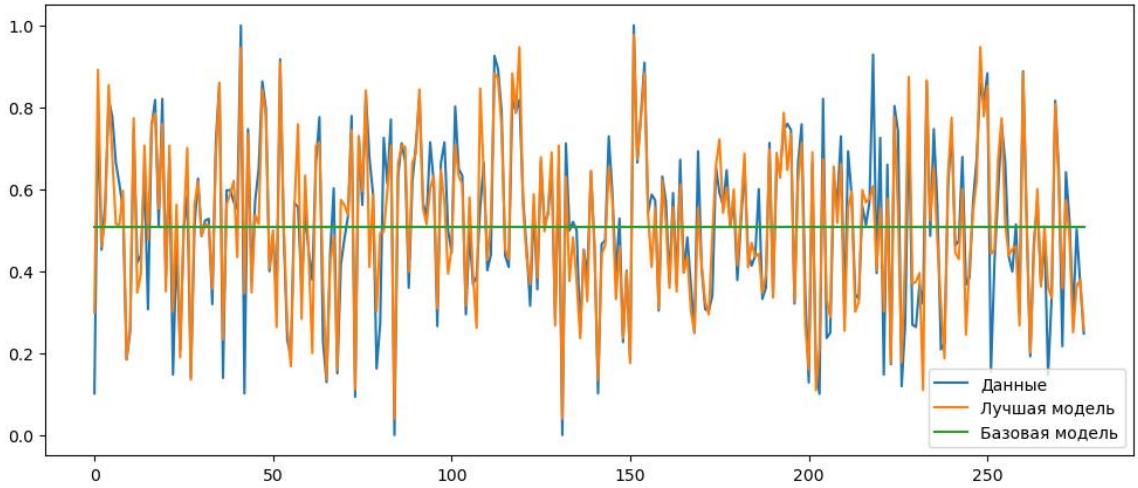


Рис25 Исходные данные по сравнению с лучшей моделью.

Для решения задачи по прогнозированию значения параметра “Соотношение матрица-наполнитель” создадим нейросети с помощью MLPRegressor из библиотеки sklearn и с помощью Sequential из библиотеки Keras.

MLPRegressor (Многослойный перцептрон для регрессии) является моделью нейронной сети, которая используется для решения задач регрессии. Он основан на архитектуре многослойного перцептрона (MLP), который состоит из нескольких слоев нейронов, включая входной слой, скрытые слои и выходной слой.

Основная идея MLPRegressor заключается в том, чтобы обучить модель на основе входных данных и соответствующих целевых значений. Во время обучения модель настраивает веса и смещения нейронов, чтобы минимизировать ошибку между предсказанными и фактическими значениями.

Процесс обучения MLPRegressor включает следующие шаги:

1. Инициализация весов и смещений: Начальные значения весов и смещений выбираются случайным образом или на основе предварительных знаний о данных.

2. Прямое распространение (forward propagation): Входные данные передаются через сеть от входного слоя к выходному слою. Каждый нейрон вычисляет свою активацию на основе взвешенной суммы входных значений и функции активации.

3. Вычисление ошибки: Разница между предсказанными и фактическими значениями вычисляется с использованием функции потерь, такой как среднеквадратичная ошибка (MSE) или абсолютная ошибка.

4. Обратное распространение (backpropagation): Ошибка распространяется обратно через сеть, и веса и смещения нейронов обновляются с использованием градиентного спуска. Это позволяет модели корректировать свои предсказания и уменьшать ошибку.

5. Повторение шагов 2-4: Процесс прямого и обратного распространения повторяется до тех пор, пока модель не достигнет заданного критерия остановки, например, определенного числа эпох или достижения минимальной ошибки.

MLPRegressor может быть настроен с помощью различных параметров, таких как количество скрытых слоев, количество нейронов в каждом слое, функции активации, скорость обучения и другие. Эти параметры могут влиять на производительность модели и ее способность обобщать на новые данные.

MLPRegressor широко используется для решения задач регрессии в различных областях, включая финансы, прогнозирование, медицину и другие. Он может быть эффективным инструментом для аппроксимации сложных нелинейных функций и построения моделей, способных предсказывать непрерывные значения на основе входных данных.

Параметры нейросети MLPRegressor:

- Последовательная модель нейронной сети

- Модель состоит из четырёх скрытых слоев, в каждом слое по 12 нейронов и выходного слоя с одним нейроном.

- Функция активации слоев выбран tanh.

- В качестве оптимизатора нейронной сети используется ADAM.

- В нейронной сети используется функция ранней остановки обучения.

- Максимальное количество итераций для обучения модели – 5000.

- Для валидации будет использовано 30% обучающих данных.

Модель нейронной сети, созданной с помощью MLPRegressor показала неудовлетворительный результат по сравнению с регрессионными моделями.

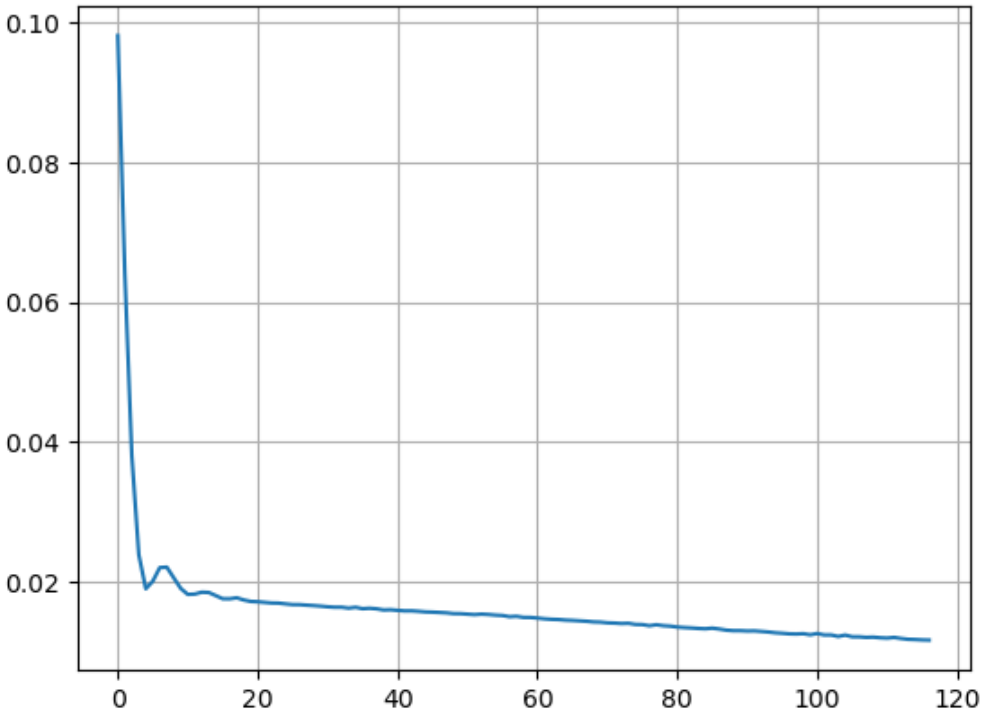


Рис26 - График ошибки

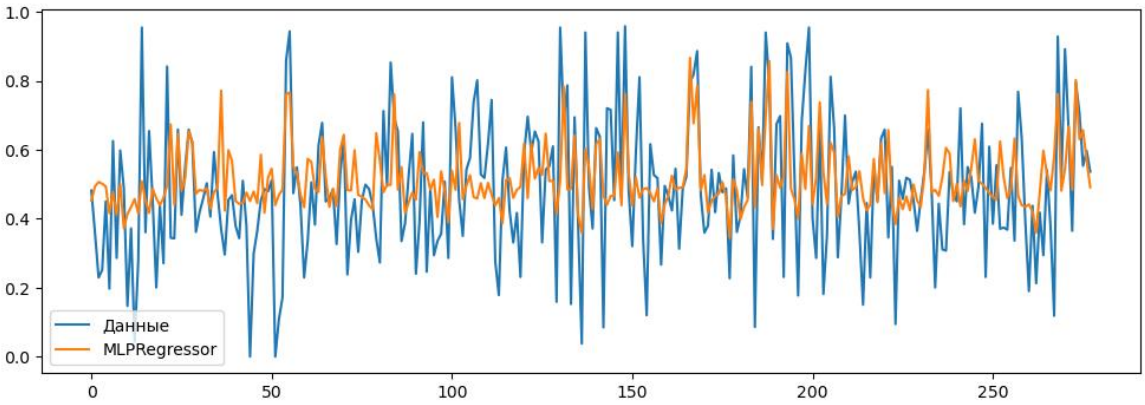


Рис25 Исходные данные по сравнению с MLPRegressor

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.014993	0.200563	0.157719	1.650086e+13	0.509286
MLPRegressor	0.218544	0.175983	0.135013	1.442920e+13	0.462553

Рис25 Метрика MLPRegressor

Архитектура и параметры нейронной сети из библиотеки Keras:

- Последовательная модель (Sequential) нейронной сети.

- Модель состоит из четырёх скрытых слоев (Dense), с количеством нейронов, в которых равно 24 и выходного слоя с одним нейроном, так же в нейронной сети есть три слоя регуляризации (Dropout) со значениями 0.05

- Функция активации слоев выбран гиперболический тангенс (relu).

- В качестве оптимизатора нейронной сети используется ADAM (стохастический градиентный спуск) функцией потерь "среднеквадратическая ошибка" и метрикой оценки качества "среднеквадратичная ошибка".

- В нейронной сети используется функция ранней остановки обучения, если в течение пяти эпох не наблюдается улучшения потерь на валидационной выборке.

- Количество эпох обучения равно 100

- Для валидации будет использовано 30% обучающих данных

Результаты использования моделей нейросетей. Лучший результат у нейросети с ранней остановкой.

	R2	**RMSE**	**MAE**	**MAPE**	**max_error**
Нейросеть переобученная	0.118761	0.186881	0.14111	5.037199e+12	0.520695
Нейросеть с ранней остановкой	0.208985	0.177056	0.135422	5.625733e+12	0.666826
Нейросеть dropout	-0.349995	0.231305	0.173923	3.329039e+12	0.791174

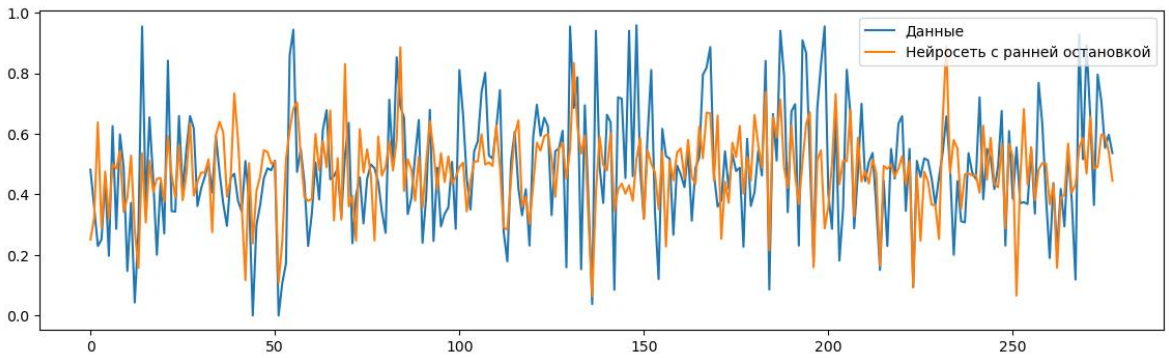


Рис25 Исходные данные по сравнению с нейросетью

	R2	RMSE	MAE	MAPE	max_error
Модуль упругости, тренировочный	0.963301	0.039467	0.022721	6.819143e-02	0.297971
Модуль упругости, тестовый	0.853040	0.080487	0.054977	1.280103e+12	0.320850
Прочность при растяжении, тренировочный	0.685419	0.103562	0.084067	1.251893e+12	0.386806
Прочность при растяжении, тестовый	0.588939	0.116025	0.091157	2.465326e-01	0.395783
Соотношение матрица-наполнитель, тренировочный	0.209754	0.172197	0.126207	2.950356e-01	0.615073
Соотношение матрица-наполнитель, тестовый	0.208985	0.177056	0.135422	5.625733e+12	0.666826

2.4. Разработка приложения Модуль упругости при растяжении

Простое консольное приложение было разработано для прогноза модуля упругости при растяжении

.

2.5. Создание удаленного репозитория

В дополнение к данной работе, был создан удаленный репозиторий на GitHub, который находится по адресу https://github.com/klerik87/DSPro. На него были загружены все необходимые материалы по нашему заданию. А именно: исследовательский ноутбук, пояснительная записка, файлы приложения, а также необходимые файлы для работы с ноутбуком

.

Заключение

По итогу данной работы удалось выполнить все поставленные задачи. Исходные данные были обработаны и несколько моделей были созданы на их основе.

Данные модели, конечно, непригодны для использования в промышленной среде, прогнозированные значения дают большую погрешность. Требуется дополнительная корректировка параметров обучения моделей

В ходе данной работы было применено множеством различных статистических методов, использующихся для прогноза и предобработки данных. Полученные данные в виде графиков и таблиц позволяют сравнить методы и выбрать самй эффективный.

Было освоено множество методов и библиотек на языке программирования Python для решения различных задач и написания нейросетей, подборки архитектуры и создания прикладных приложений.

Список использованных источников и литературы

Список использованной литературы

1) Композиционные материалы: Справочник /Под. ред. В.В. Васильева, Ю.М.Тарнопольского. –М.: Машиностроение, 1990. –512 с.

2) Библиотека Keras - инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow / пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2018. - 294 с.

3) Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.

4) Платформа scikit-learn [Электронный ресурс]: https://scikit-learn.org/stable/

5) Библиотека Seaborn-: https://seaborn.pydata.org/.

6) Язык программирования Python- https://www.python.org

7) Библиотека Pandas https://pandas.pydata.org/

8) Библиотека Sklearn https://scikit-learn.org/stable/

9) Библиотека Pandas- https://pandas.pydata.org/.

10) Библиотека Matplotlib- https://matplotlib.org

11) Библиотека Tensorflow: https://www.tensorflow.org/.

12) https://habr.com/ru/companies/otus/articles/4429

13) https://chat.openai.com/