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»

ТЕМА:

«Прогнозирование конечных свойств новых материалов (композиционных материалов)»

Слушатель Стоянов Павел Александрович

Москва, 2023

Содержание

Введение

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними, что схематично отображено на рисунке 1. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом.

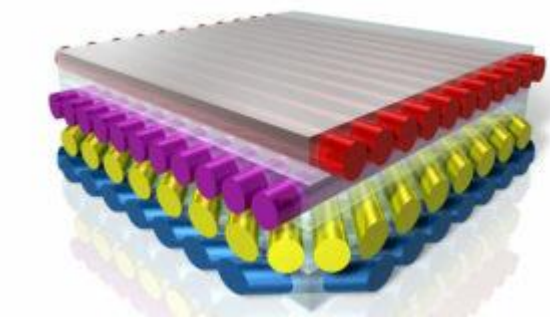


Рисунок 1—Структура композиционного материала

5 интересных фактов о композитах:

1) одни из самых первых рукотворных композитных материалов - высушенные на солнце глиняные кирпичи с добавлением рубленной соломы. Первое использование этого метода относится к 1500 году до нашей эры. Древние Египтяне оставили на стенах пирамид изображения этой технологии;
2) 1200 год нашей эры, постарались монголы: они создали первый композиционный лук из таких материалов, как древесина, кость и животный клей;
3) самый известный искусственный композитный материал – бетон;
4) при пошиве спортивной одежды и обуви используется материал Gore-Tex, который является композитом из слоев различных материалов. Он одновременно водонепроницаемый и пропускающий молекулы воздуха;
5) почти половина деталей современного самолета произведены из композитов;

6) существует самовосстанавливающийся полимер. Этот композит содержит химические вещества, которые образуют новый слой при повреждении поверхности изделия.

Современные композиты изготавливаются из материалов: полимеры, керамика, стеклянные и углеродные волокна.

Сейчас мировой рынок композитов составляет 80 млрд долл. На нем лидируют Китай 32 процента (25,6 млрд долл.) и США с 26 процентами (21,6 млрд долл.). Сегмент России – 1 процент (1,1 млрд долл.). В структуре российского рынка композитов преобладает, в частности, строительная индустрия – 35 процентов (22,9 млрд руб.). Это водоотводные лотки, изолирующие накладки, композитные материалы и т.д. На втором месте – гражданское авиа- и судостроение: 19 процентов (12,4 млрд руб.).

По структуре композиты делятся на несколько основных классов: волокнистые, дисперсно-упрочнённые, упрочнённые частицами и нанокомпозиты.

Композиционные материалы классифицируют по следующим основным признакам: типу матрицы, виду армирующего элемента, особенностям макростроения и методам получения. Сначала осуществляют ориентировочный выбор материала матрицы, основных наполнителей и арматуры, а также технологии формирования изделий.

В составе композита принято выделять матрицу/матрицы и наполнитель/наполнители, последние выполняют функцию армирования (по аналогии с арматурой). В качестве наполнителей композитов как правило выступают углеродные или стеклянные волокна, а роль матрицы играет полимер. Определение содержания и относительного расположения различных армирующих элементов в матрице, прежде всего, зависит от таких исходных требований, как прочность и жесткость, тепло- и электропроводность, технологичность, стоимость материала и т. д. Часто процессы формирования изделия и композиционного материала совмещаются.

Композиты, в которых матрицей служит полимерный материал, являются одним из самых многочисленных и разнообразных видов материалов. В качестве

наполнителей ПКМ используется множество различных веществ: Стеклопластики, Углепластики, Боропластики , Органопластики, Полимеры, наполненные порошками, Текстолиты.

При создании композитов на основе металлов в качестве матрицы применяют алюминий, магний, никель, медь и так далее. Наполнителем служат или высокопрочные волокна, или тугоплавкие, не растворяющиеся в основном металле частицы различной дисперсности.

Армирование керамических материалов волокнами, а также металлическими и керамическими дисперсными частицами позволяет получать высокопрочные композиты, однако, ассортимент волокон, пригодных для армирования керамики, ограничен свойствами исходного материала.

Для решения проблемы моделирования композитов 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). У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично.

1 Аналитическая часть

1.1 Постановка задачи

На основе набора данных из файлов X_bp.xlsx и X_nup.xlsx (объединение делать по индексу тип объединения INNER) необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов, а именно:

1) обучить нескольких моделей для определения значений «Модуль упругости при растяжении, ГПа» и «Прочность при растяжении, МПа». При построении моделей необходимо 30% данных оставить на тестирование моделей, на остальных провести обучение моделей. При построении моделей провести поиск гиперпараметров моделей с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10;
2) написать нейронную сеть, которая будет рекомендовать характеристику «Соотношение матрица-наполнитель»;
3) разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз характеристик «Модуль упругости при растяжении, ГПа» и «Прочность при растяжении, МПа» или характеристику «Соотношение матрица-наполнитель».

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

1.2 Характеристика датасета

Начнем с описательного анализа данных, в большинстве случаев он используется для первичного определения типов информации.

Он включает:

1) проверку типа данных;
2) удаление нерелевантных столбцов (не используем, не имеют для анализа никакого смысла);
3) переименование столбцов (улучшает читаемость набора данных);
4) удаление повторяющихся, дубликатов строк (сокращает размер датасета);

5) удаление отсутствующих или нулевых значений либо замена их средним/медианой или модой для этого столбца;
6) обнаружение выбросов и идентификация их таковыми, удаление (чтобы не искажали результаты).

На входе имеются набор данных (файлы X_bp.xlsx и X_nup.xlsx) с начальными свойствами компонентов композиционных материалов, таблица 1.

Файл X_bp.xlsx содержит 1023 записи и одиннадцать столбцов с признаками (без названия, соотношение матрица-наполнитель, плотность, модуль упругости, количество отвердителя, содержание эпоксидных групп, температура вспышки, поверхностная плотность, модуль упругости при растяжении, прочность при растяжении, потребление смолы), в том числе три выходные переменные, которые нас интересуют (соотношение матрица-наполнитель, модуль упругости при растяжении, прочность при растяжении).

Таблица 1— Наборы данных (файлы X_bp.xlsx и X_nup.xlsx)

X_bp.head(5)

	Unnamed: 0	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
0	0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0
1	1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0
2	2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0
3	3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0
4	4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0

X_nup.head(5)

	Unnamed: 0	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	0	0	4.0	57.0
1	1	0	4.0	60.0
2	2	0	4.0	70.0
3	3	0	5.0	47.0
4	4	0	5.0	57.0

Файл X_nup.xlsx содержит 1040 записей и четыре столбца с признаками (без названия, угол нашивки, шаг нашивки, плотность нашивки).

Колличество записей в файлах отличается на 17 строчек. Учитывая условие задачи «объединение делать по индексу тип объединения INNER», после объединения файлов в один, теряем эти 17 строчек. Общий размер набора данных и отсутствие основной и важной инфо для этих строчек в объединяемом файле позволяют нам так поступить. В объединенном файле присутствует первый столбец без названия с индексом, который не несет никакой информации, его удаляем, для упрощения дальнейшей работы. Получаем датасет в 1023 строчки и 13 столбцов.

Тип данных всех столбцов - float64 (числа с плавающей точкой), кроме столбца "Угол нашивки, град", он имеет тип int64 (целые числа). Пропусков в данных нет, нулевых значений также нет, дубликатов нет, все значения имеют вещественный тип данных, что видно по таблице 2. Т.о. заполнять пропуски, чистить датасет, преобразовывать тип данных не требуется.

Практически все значения признаков по столбцам уникальны, как видно по таблице 3.

Проанализировав наиболее часто встречающиеся значения по всем 13 столбцам, отражено в таблице 4, выясняем, что первые 23 строки в объединенном датасете полностью сгенерированы усредненными значениями, таблица 5. Причина этого пока не ясна, т.к. при размерности в 1000 строк полная генерация усредненных дополнительных 23 строк ни на что не влияет.

Таблица 2—Описание объединенного датасета

#	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, С_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: 104.0 KB

Таблица 3—Наличие уникальных значений

Соотношение матрица-наполнитель	1014
Плотность, кг/м3	1013
модуль упругости, ГПа	1020
Количество отвердителя, м.%	1005
Содержание эпоксидных групп,%_2	1004
Температура вспышки, С_2	1003
Поверхностная плотность, г/м2	1004
Модуль упругости при растяжении, ГПа	1004
Прочность при растяжении, МПа	1004
Потребление смолы, г/м2	1003
Угол нашивки, град	2
Шаг нашивки	989
Плотность нашивки	988

dtype: int64

Таблица—4 Наиболее часто встречающиеся значения

300.000000	12
284.615385	10
278.457686	1
382.759808	1
236.284743	1
..	
248.849293	1
237.911236	1
364.872558	1
273.857159	1
300.952708	1

Name: Температура вспышки,

70.000000	9
73.333333	5
78.000000	5
75.000000	4
74.042708	1
..	
71.758909	1
76.398875	1
66.522175	1
72.124020	1
74.309704	1

Name: Модуль упругости при растяжении,

Таблица—5 Сгенерированные строки в датасете с 0 по 22

оотношение матр	Плотность, кг/м3	модуль упругости, Г	Количество отверди	Содержание эпокси	Температура вспыш	Поверхностная пло	Модуль упругости г	Прочность при рас	Потребление смол	Угол нашивки, град	Шаг нашивки	Плотность нашивки
1.857142857	2030	736.7368421	30	22.26785714	100	210	70	3000	220	0	4	57
1.857142857	2030	736.7368421	50	23.75	284.6153846	210	70	3000	220	0	4	60
1.857142857	2030	736.7368421	49.9	33	284.6153846	210	70	3000	220	0	4	70
1.857142857	2030	736.7368421	129	21.25	300	210	70	3000	220	0	5	47
2.771331058	2030	753	111.86	22.26785714	284.6153846	210	70	3000	220	0	5	57
2.767918089	2000	748	111.86	22.26785714	284.6153846	210	70	3000	220	0	5	60
2.569620253	1910	807	111.86	22.26785714	284.6153846	210	70	3000	220	0	5	70
2.56147541	1900	535	111.86	22.26785714	284.6153846	380	75	1800	120	0	7	47
3.557017544	1930	889	129	21.25	300	380	75	1800	120	0	7	57
3.532338308	2100	1421	129	21.25	300	1010	78	2000	300	0	7	60
2.919677836	2160	933	129	21.25	300	1010	78	2000	300	0	7	70
2.877358491	1990	1628	129	21.25	300	1010	78	2000	300	0	9	47
1.598173516	1950	827	129	21.25	300	470	73.33333333	2455.555556	220	0	9	57
2.919677836	1980	568	129	21.25	300	470	73.33333333	2455.555556	220	0	9	60
4.029126214	1910	800	129	21.25	300	470	73.33333333	2455.555556	220	0	9	70
2.934782609	2030	302	129	21.25	300	210	70	3000	220	0	10	47
3.557017544	1880	313	129	21.25	300	210	70	3000	220	0	10	57
4.193548387	1950	506	129	21.25	300	380	75	1800	120	0	10	60
4.897959184	1890	540	129	21.25	300	380	75	1800	120	0	10	70
3.532338308	1980	1183	111.86	22.26785714	284.6153846	1010	78	2000	300	0	0	0
2.877358491	2000	205	111.86	22.26785714	284.6153846	1010	78	2000	300	90	4	47
1.598173516	1920	456	111.86	22.26785714	284.6153846	470	73.33333333	2455.555556	220	90	4	57

Значений похожих на выбросы согласно таблице 6 в датасете небольшое колличество.

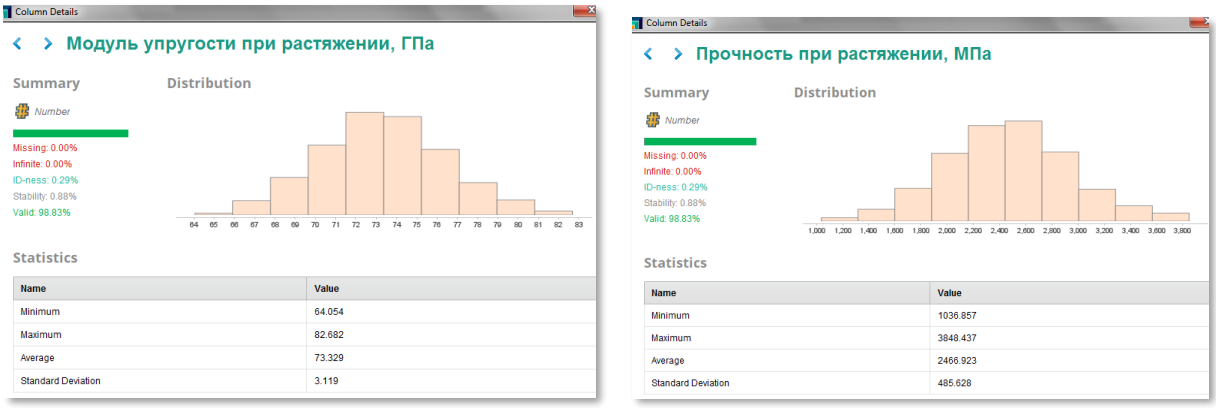
Таблица—6 Количество предполагаемых выбросов в датасете

```
Количество выбросов в столбце  Соотношение матрица-наполнитель :  6
Количество выбросов в столбце  Плотность, кг/м3 :  9
Количество выбросов в столбце  модуль упругости, ГПа :  2
Количество выбросов в столбце  Количество отвердителя, м.% :  14
Количество выбросов в столбце  Содержание эпоксидных групп,%_2 :  2
Количество выбросов в столбце  Температура вспышки, С_2 :  8
Количество выбросов в столбце  Поверхностная плотность, г/м2 :  2
Количество выбросов в столбце  Модуль упругости при растяжении, ГПа :  6
Количество выбросов в столбце  Прочность при растяжении, МПа :  11
Количество выбросов в столбце  Потребление смолы, г/м2 :  8
Количество выбросов в столбце  Угол нашивки, град :  0
Количество выбросов в столбце  Шаг нашивки :  4
Количество выбросов в столбце  Плотность нашивки :  21
Общее число ошибок: 93
```

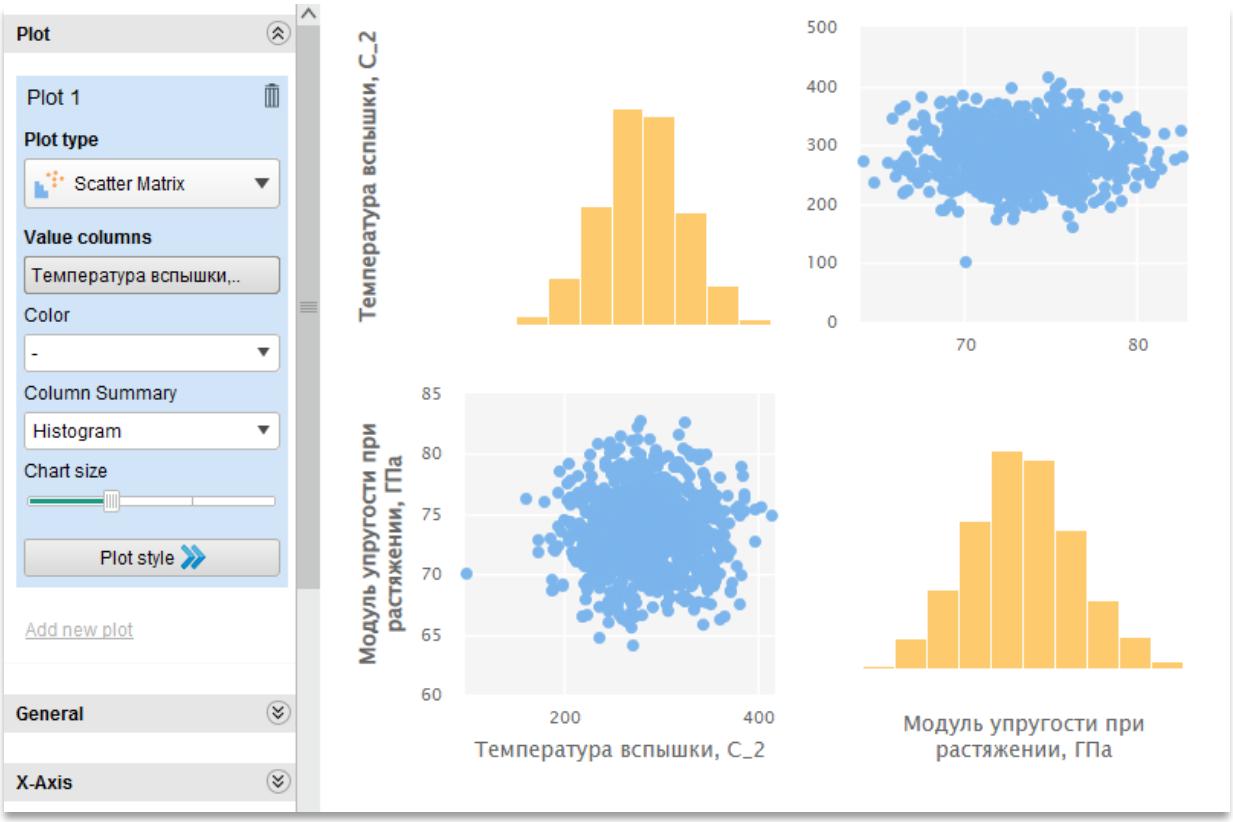
Чтобы идентифицировать их таковыми и принять решение удалить, нужно выбросы оценить, иначе информация может быть упущена (выполним это в подразделе 1.4). В сочетании с фактическими данными выбросы можно разделить на "истинные аномалии" и "ложные аномалии". Экспертная оценка выбросов нам не доступна (не являемся экспертами), т.к. опирается на вопросы:

1) является ли выброс результатом ошибки ввода данных?
2) влияет ли выброс на результаты анализа?
3) влияет ли выброс на предположения?

RapidMiner—аналитическая платформа, по сути конструктор с набором готовых «кубиков-решений» для работы с данными без написания кода и без специфических знаний. Минусом аналитических платформ является ограниченное количество компонентов. Посмотрим характеристику объединенного датасета в RapidMiner, рисунки 2-3.



Рисунок—2 Статистический модуль в RapidMiner



Рисунок—3 Визуализация в RapidMiner

Описательную статистику смотрим по таблице 7, выполнив транспонирование для удобства. Расшифруем отдельные характеристики:

1) Std — стандартное отклонение значения;
2) 25% — первый квартиль, 25% значений в столбце ниже этого значения;
3) 50% — медиана, половина значений в столбце ниже этого значения;
4) 75% — третий квартиль, 75% значений в столбце ниже этого значения;

Таблица—7 Описательная статистика

	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

1.3 Описание используемых методов

При выборе методов будем опираться на рекомендации экспертов, уже опробовавших разные методы при работе с композитными материалами и отметивших точность отдельных их них. Ниже приведены выдержки из работ по этой теме.

Реутов Ю.А.: «Решение задачи «структура – свойство» в общем случае может выполняться различными методами машинного обучения, такими как: классическая регуляризация, статистический анализ, нейронные сети, метод опорных векторов, кластеризация, алгоритмические композиции и другие».

Чун-Те Чен и Грейс Х. Гу в своей статье дают краткий обзор некоторых основных алгоритмов машинного обучения и обзор недавних исследований с использованием моделей машинного обучения для прогнозирования механических свойств композитов:

«Рассмотрели линейную регрессию, логистическую регрессию, нейронные сети (NN), сверточные нейронные сети (CNN) и гауссовский процесс (GP) в контексте проектирования материалов, ссылаясь на исследования (как экспериментальные, так и вычислительные) по применению этих алгоритмов машинного обучения к композитным исследованиям (включая нанокомпозиты).

При линейной регрессии как только функция ошибок (MSE) модели определена, веса модели могут быть рассчитаны с помощью алгоритма оптимизации, такого как классический стохастический градиентный спуск или алгоритм оптимизации Адама. Более сложные модели машинного обучения (нейронные сети), в целом дают более точные прогнозы. Однако в исследованиях по композитам линейные модели давали ценную информацию, например, какие входные переменные были более важными (с более сильным влиянием) для прогноза».

Авторы показали, что использование машинного обучения для прогнозирования механических свойств композитов на порядки быстрее, чем обычный анализ методом конечных элементов.

Итак, в своей работе будем использовать метод обучения с учителем через задачу регрессии (наш датасет состоит из размеченных числовых данных). Для целевых признаков «Модуль упругости при растяжении», «Прочность при растяжении» будут применены следующие методы:

1) LinearRegression;
2) Lasso;
3) RandomForestRegressor;
4) KNeighborsRegressor;
5) SVR;
6) DecisionTreeRegressor;
7) AdaBoostRegressor.

Для целевого признака «Соотношение матрица-наполнитель» используем нейронную сеть прямого распространения (FF, класс Sequential).

Линейная регрессия (Linear regression) – это метод машинного обучения с учителем, который используется для предсказания непрерывной целевой переменной от одного или нескольких независимых признаков. В основе метода лежит предположение о том, что существует линейная связь между признаками и целевой переменной. Модель линейной регрессии пытается найти лучшую прямую, которая может описывать зависимость между независимыми признаками

и зависимой переменной. Эта модель может быть использована для предсказания, для анализа влияния признаков на целевую переменную.

Достоинства:

- простота и удобство в использовании;
- эффективность при линейных зависимостях;
- интерпретируемость.

Недостатки:

- ограниченная эффективность при нелинейных зависимостях;
- необходимость преподготовки данных: она чувствительна к выбросам.

Регрессия «Лассо» (LASSO, Least Absolute Shrinkage and Selection Operator). Регрессия по методу наименьших квадратов часто может стать неустойчивой, то есть сильно зависящей от обучающих данных, что обычно является проявлением тенденции к переобучению. Избежать такого переобучения помогает регуляризация – общий метод, заключающийся в наложении дополнительных ограничений на искомые параметры, которые могут предотвратить излишнюю сложность модели. Смысл процедуры заключается в «стягивании» в ходе настройки вектора коэффициентов таким образом, чтобы они в среднем оказались несколько меньше по абсолютной величине, чем это было бы при оптимизации по методу наименьших квадратов.

Лассо регрессия заключается во введении дополнительного слагаемого регуляризации в функционал оптимизации модели, что позволяет получать более устойчивое решение.

Достоинства:

- более точные и стабильные оценки истинных параметров;
- уменьшение ошибок выборки и отсутствия выборки.

Недостатки:

- трудно интерпретировать коэффициенты в конечной модели, поскольку они уменьшаются до нуля. Лассо следует использовать, когда вы заинтересованы в оптимизации для способности к прогнозированию, а не для вывода.

RandomForestRegressor: Метод является универсальным для большинства моделей машинного обучения как «с учителем», так и без него. Этот представитель ансамблевых методов успешно применяется для решения задач кластеризации, классификации, регрессии. Random Forest Regressor представляет собой множество (лес) независимых деревьев решений. Случайный лес случайным образом выбирает наблюдения (строки) и конкретные объекты (переменные) для построения нескольких деревьев решений, а затем усредняет результаты.

Достоинства:

- делает достаточно точные предсказания;
- обрабатывает пропуски в наборе данных;
- не переобучается;
- не требует предварительной обработки входных данных;
- хорошо масштабирует.

Недостатки:

- требуется значительный объем вычислительных ресурсов, отнимает больше времени, чем деревья решений или линейные алгоритмы;
- склонен к переобучению на зашумленных данных;
- большой размер моделей.

KNeighborsRegressor: объекту присваивается среднее значение по (k) ближайшим к нему объектам, значения которых уже известны. Перед применением алгоритма нужно определить функцию расстояния; классический вариант такой функции — евклидова метрика как расстояние между двумя точками евклидова пространства.

Достоинства:

- позволяет настроить несколько параметров одновременно;
- не чувствителен к выбросам;
- является универсальным для решения задач как с» учителем», так и без него.

Недостатки:

- замедляется с ростом объёма данных;
- не создаёт правил; не обобщает предыдущий опыт;
- сложно интерпритировать;
- требуется значительный объем вычислительных ресурсов.

SVR: Это один из наиболее широко используемых алгоритмов машинного обучения, применяемый для решения задач классификации, регрессии и обнаружения выбросов. Алгоритм для решения задач классификации строит гиперплоскость в n-мерном пространстве для разделения объектов двух или более классов. Гиперплоскость выбирается таким образом, чтобы максимизировать расстояние между гиперплоскостью и ближайшими объектами разных классов (зазор). Объекты, которые расположены ближе всего к гиперплоскости, называются опорными векторами.

Достоинства:

- является одним из наиболее точных алгоритмов машинного обучения, которые могут обучаться на больших наборах данных;
- хорошо работает с данными, которые имеют большое количество признаков;
- работает с небольшими выборками данных;
- имеет несколько гиперпараметров, что делает его относительно простым для настройки.

Недостатки:

- чувствительность к шуму;
- вычислительная сложность для обучения на больших наборах данных;

DecisionTreeRegressor: Состоит из «узлов», «листьев» и «веток». «Ветки» содержат записи атрибутов, от которых зависит целевая функция, «листья» – значения целевой функции, а «узлы» – остальные атрибуты, по которым происходит классификация. Два типа деревьев: для классификации и для регрессии.

Достоинства:

- простота понимания и интерпретации;

- минимальные требования к подготовке данных, способность работы с большими объемами данными;
- одинаково хорошо работает с разными видами признаков;
- позволяет оценить модель статистическими тестами.

Недостатки:

- подверженность переобучению.

AdaBoostRegressor: Алгоритм машинного обучения, в котором каждый следующий weak learner фокусирует внимание на тех примерах, на которых предыдущие weak learner'ы дали неверные ответы. При этом он не знал, какие именно ответы даны предыдущими weak learner'ами - было лишь известно, что ответы неверны. При этом не используется валидационный датасет. Используется обучающий датасет, на нем оценивается точность предыдущих weak learner'ов. Нужно использовать решающие деревья небольшой глубины: weak learner должен быть "слабым", не переобучаясь слишком сильно.

Достоинста:

- можно быстро и просто запрограммировать;
- гибкий, чтобы комбинировать его с любым алгоритмом машинного обучения без настройки параметров;
- универсален, можно использовать с числовыми или текстовыми данными.

Недостатки:

- алгоритм доказывается эмпирически и очень уязвим к равномерно распределенному шуму;
- слабые классификаторы могут привести к плохим результатам и переобучению.

Сеть прямого распространения (Feed Forward) — это искусственная нейронная сеть, в которой нейроны никогда не образуют цикла. В этой нейронной сети все нейроны расположены в слоях, где входной слой принимает исходные данные, а выходной слой генерирует результат в заданном виде. По-

мимо входного и выходного слоев, есть еще **скрытые слои** — это слои, которые не имеют связи с внешним миром. В нейронной сети прямого распространения каждый нейрон одного слоя связан с каждым нейроном на следующем слое. Слои с такими нейронами называются полносвязными. Цель обучения нейронной сети — найти такие параметры сети, при которых нейронная сеть будет ошибаться наименьшее количество раз. Ошибка нейронной сети — отличие между предсказанным значением и правильным.

Применение:

- сжатие данных;
- распознавание образов;
- компьютерное зрение;
- распознавание речи.

Для сравнения моделей и оценки точности их работы используем метрики:

- Absolute Error (*MAE*)—абсолютная ошибка (ошибка в 10 долларов должна интерпретироваться как в два раза худшая, чем ошибка в 5 долларов);
- Squared Error (*MSE*)—среднеквадратичная ошибка (евклидово расстояние). Применяется в случаях, когда требуется подчеркнуть большие ошибки и выбрать модель, которая дает меньше именно больших ошибок. Большие значения ошибок становятся заметнее за счет квадратичной зависимости. Модель, которая обеспечивает меньшее значение *MSE* допускает меньше именно больших ошибок;
- Mean Absolute Percentage Error (MAPE) –средняя абсолютная ошибка в процентах, эта ошибка не имеет размерности и очень проста в интерпретации. Её можно выражать как в долях, так и в процентах. Если получилось, например, что *MAPE* равно11.4, то это говорит о том, что ошибка составила 11.4 процента от фактического значения;
- Root mean Squared Error (*RMSE*)—корень из среднеквадратичной ошибки. Сравнение моделей с помощью *RMSE* даст такой же результат, что и

для *MSE*. Однако с *MSE* работать несколько проще. Большие ошибки оказывают непропорционально большое влияние на *RMSE*. Следовательно, *RMSE* можно считать чувствительной к аномальным значениям;

- Coefficient of determination *(R*-квадрат)—коэффициент детерминации *R*2 равный 0 показывает, что между независимой и зависимой переменными модели имеет место функциональная зависимость. Коэффициент *R*2 принимает отрицательные значения в случае, если ошибка модели среднего становится меньше ошибки модели с переменной. Модель, для которой *R*2 больше 0.5, является удовлетворительной. Если *R*2 больше 0.8, то модель рассматривается как очень хорошая. Значения, меньшие 0.5 говорят о том, что модель плохая;
- Test score—метод обеспечивающий критерий оценки по умолчанию, показывает качество прогнозов моделей на обучающих и тестовых выборках.

Основная цель алгоритма обучения - подобрать значения параметров таким образом, чтобы для объектов обучающей выборки, для которых мы уже знаем правильные ответы, предсказанные значения были как можно ближе к тем, которые есть в датасете, истинным значением.

1.4 Разведочный анализ данных

Разведочный (EDA) или, как его еще принято называть, исследовательский анализ применяется для выявления тенденций. При его использовании выявляют закономерности, обобщаются основные характеристики.

К основным методам разведочного анализа относится:

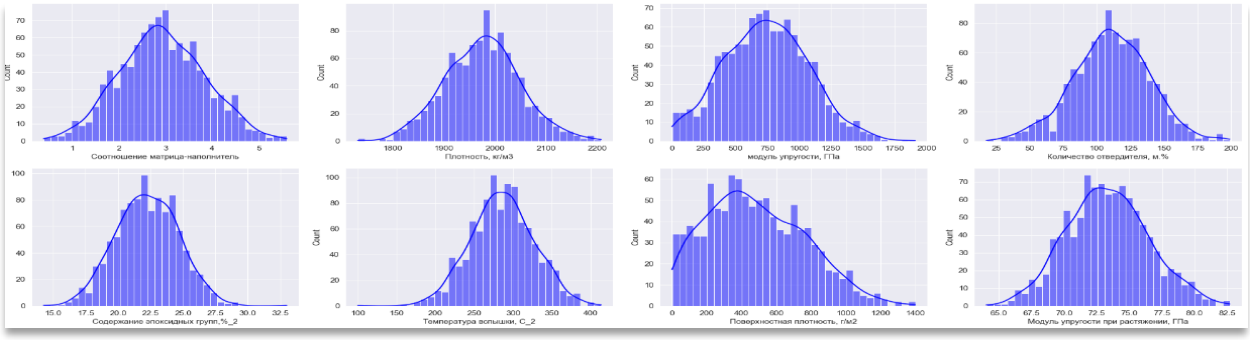
1) процедура анализа распределений переменных (чтобы выявить переменные с несимметричным или негауссовым распределением, в том числе и бимодальные). Гистограмма позволяет "на глаз" оценить нормальность эмпирического распределения. На гистограмму также накладывается кривая нормального распределения.

2) анализ корреляционных матриц с целью поиска коэффициентов, превосходящих по величине определенные пороговые значения. Корреляция представляет собой меру зависимости переменных. Происходит проверка значимых (ожидаемых и неожиданных) корреляций, попытка понять общую природу обнаруженной статистической значимости;
3) анализ многовходовых таблиц частот ("послойный" последовательный просмотр комбинаций уровней управляющих переменных). Это метод анализа категориальных переменных, показывающий каким образом различные группы данных распределены в выборке.
4) многомерный разведочный анализ (для поиска закономерностей в многомерных данных или последовательностях одномерных данных). К ним относятся: кластерный анализ, факторный анализ, анализ дискриминантных функций, многомерное шкалирование, анализ временных рядов;
5) статистические методы - среднее значение, медиана, мода, стандартное отклонение, дисперсия, квантили и др.

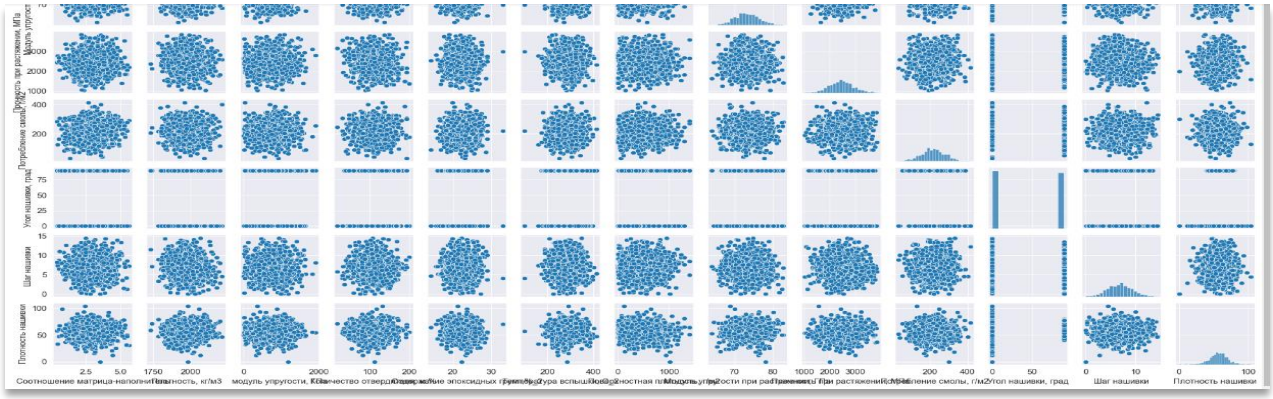
Визуализации выделяют взаимосвязи в данных и раскрывают информацию, видимую человеческому глазу, которую нельзя передать только числами и цифрами. Используем для этого библиотеки Pandas, Seaborn, Matplotlib.

Гистограмма отображает частоту появления переменных в определенном интервале и характер распределения. Посмотрим на данные на рисунках 4 и 5. Все переменные имеют нормальное распределение. «Угол нашивки» имеет только два значения. «Прочность при растяжении», «Поверхностная плотность»: медианы отличаются от среднего значения, по остальным признакам очень близки. Все попарные графики рассеяния точек показывают отсутствие зависимости между признаками.

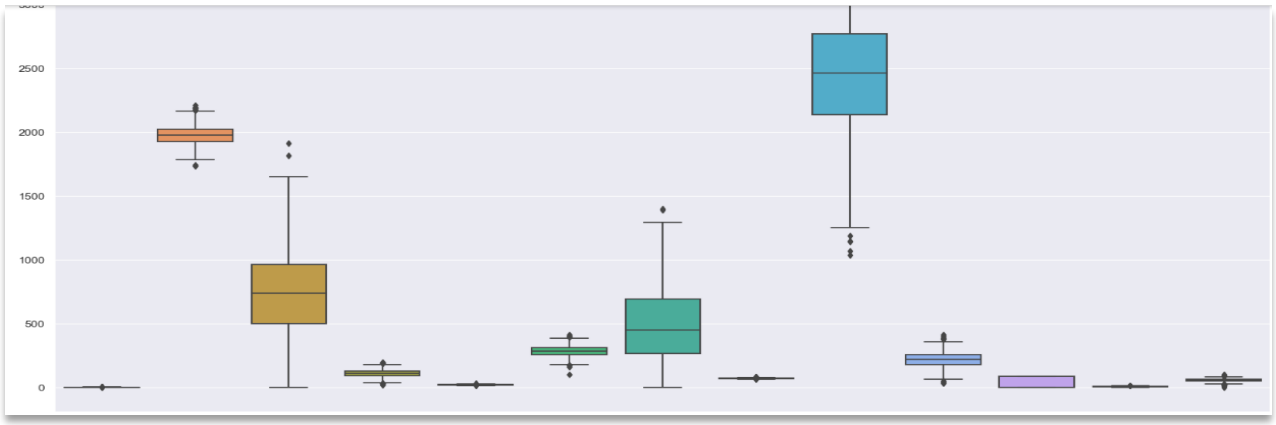
Проведем визуальный поиск наличия аномалий и выбросов с помощью метода Boxplot (ящик с усами), рисунок 6. Точки за границами "усов" (полутора межквартильных расстояний от первого и третьего квантиля), не что иное, как выбросы.



Рисунок—4 Гистограммы частот и характера распределения признаков



Рисунок—5 Попарные графики рассеяния точек (скаттерплоты)



Рисунок—6 Выбросы методом Boxplot «Ящика с усами»

Посмотрим также зависимость целевых признаков от признака «Соотношение матрица-наполнитель». Для этого разобьем значения каждого из целевых признаков на десять групп с шагом 10 процентов.

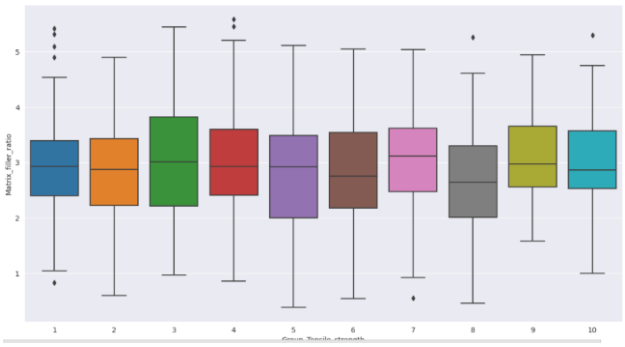


Рисунок 7 Group Tensile strength

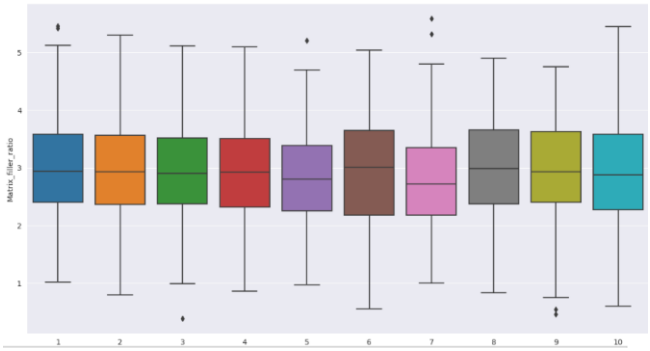


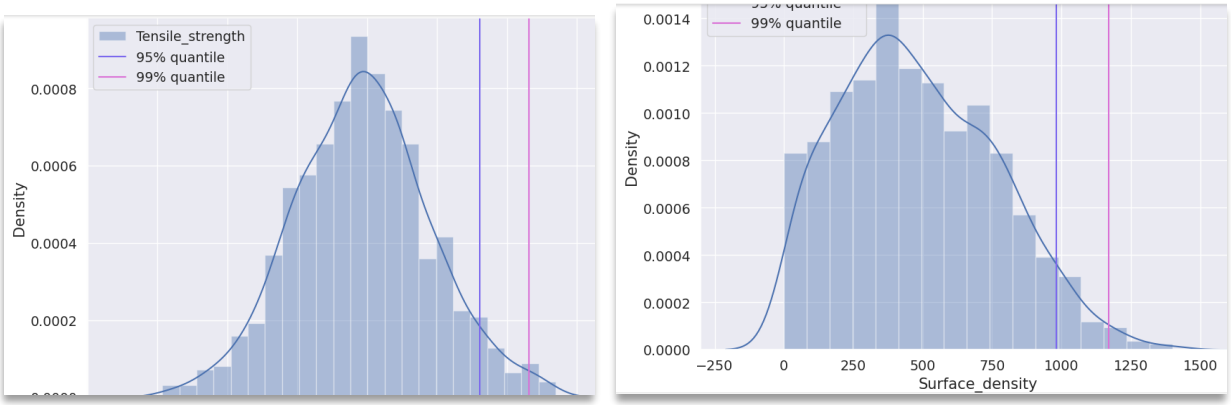
Рисунок 8 Group Tensile modulus strengt

Во всех группах признака «Модуль упругости при растяжении» среднее значение медианы находится в диапазоне от 2,2 до 3,6. Выбросы расположены не во всех группах, а в 1, 3, 5, 7, 9 группах (рисунок 8).

Во всех группах признака «Прочность при растяжении» среднее значение медианы находится в диапазоне от 2 до 2,9. Выбросы также расположены не во всех группах, а в 1, 4, 7, 8, 10 группах. Больше всего выбросов в1 группе (рисунок 7).

Переименуем столбцы на английские названия, т.к. не весь код запускается. Используем в таком виде в нужных нам местах.

Посмотрим на значения находящиеся в соответствии с рисунком 9 за 95 и 99 процентными квантилями.

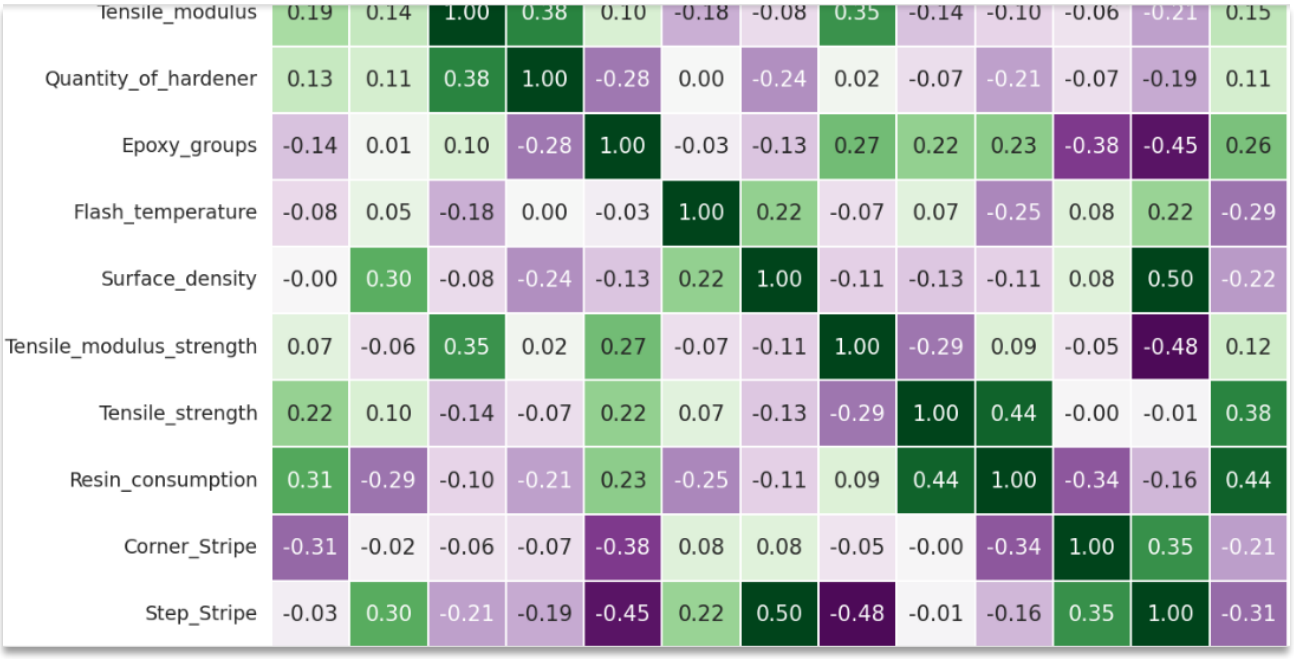


Рисунок—9 Значения, находящиеся за 95 и 99 квантили.

Посмотрим самые большие 25 значений (100 значений) в столбце «Модуль упругости при растяжении» таблицы—7, также посмотрим какая у них корреляция на рисунке 10.

Таблица—7 Самые большие 25 значений «Модуль упругости при растяжении»

Matrix_filler_ratio	Density	Tensile_modulus	Quantity_of_hardener	Epoxy_groups	Flash_temperature	Surface_density	Tensile_modulus_strength
2.106888	1914.089369	963.748397	58.710809	23.813708	278.612240	793.402805	82.682051
3.109130	2019.939791	855.877168	125.654249	22.788022	323.766318	585.456532	82.525773
5.452959	1918.688919	1016.975202	119.048341	23.461800	275.341538	41.129428	82.237600
2.271032	1862.432065	1428.870308	125.557042	24.023624	317.571591	293.085220	81.594750
2.921002	1935.790960	1027.545257	114.251686	20.919742	258.760946	358.878702	81.417126
4.414416	1901.586009	1000.289288	98.244089	24.481028	288.401849	399.870134	81.203147
2.051710	2161.565216	1012.659478	145.167640	24.603621	274.893657	837.480364	81.175231
3.267523	2107.549113	712.148822	60.994261	27.312167	268.485007	121.565802	81.053293
3.625686	2063.129132	1004.574573	89.075581	21.844375	247.156750	779.849453	80.970959
2.587348	1953.274926	1136.596135	137.627420	22.344534	234.716883	555.893453	80.803222
1.397467	2005.502050	755.383225	88.536497	23.864561	245.543354	264.281505	80.691499
1.828754	1920.564566	379.333489	103.509806	17.998559	257.285562	37.629351	80.419862
3.326891	2011.609966	1056.778632	117.355195	20.608609	322.136138	282.614376	80.384924
2.015363	2028.596291	634.827025	115.034794	23.703777	293.849227	245.590610	80.254453
3.052355	1922.447078	136.421878	47.226653	21.001922	294.906351	1058.858770	80.239609
1.810788	1939.888139	871.084608	136.163670	21.389096	273.631768	443.959193	80.147703

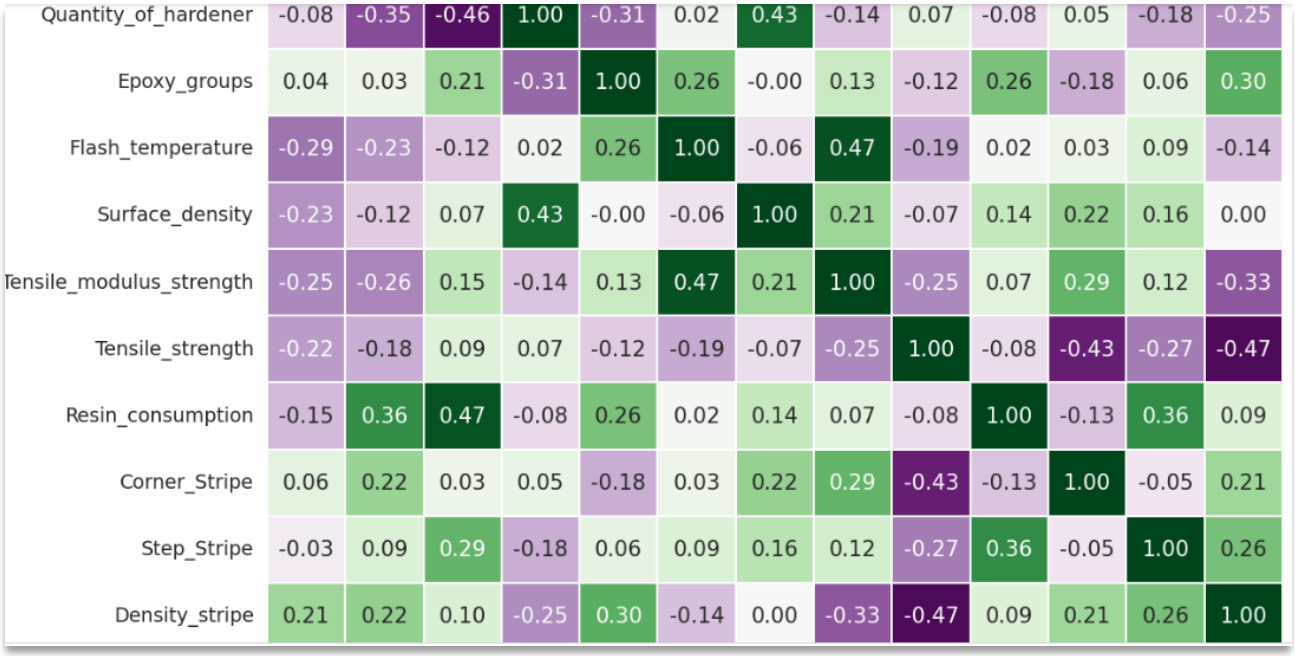


Рисунок—10 Корреляция топ 25 значений «Модуль упругости при растяжении»

Посмотрим самые большие 25 значений (100 значений) в столбце «Прочность при растяжении» таблицы—8, также посмотрим какая у них корреляция на рисунке 11. Изучив корреляции для 25, 50, 100 максимальных значений целевых признаков и сравнив их с корреляцией всего датасета видим, что она полностью отличается во всех группах (кроме «Шаг нашивки» для «Прочность при растяжении») и в сравнении с полным датасетом.

Таблица—8 Самые большие 25 значений «Прочность при растяжении»

Density	Tensile_modulus	Quantity_of_hardener	Epoxy_groups	Flash_temperature	Surface_density	Tensile_modulus_strength	Tensile_strength	Resin_consumption
1938.282144	1065.625743	148.104233	19.093392	234.856713	865.629036	72.299462	3848.436732	143.123640
1988.596676	948.981345	135.707494	22.012106	221.204295	609.350655	71.801782	3817.269484	344.718381
1839.864649	935.511792	99.079859	20.715839	330.371637	556.153263	76.632138	3791.072810	284.852950
2020.976475	1013.513596	109.505135	24.330905	266.258622	719.870002	68.166534	3773.151949	264.835584
2052.863693	607.307517	109.654355	17.638147	267.369398	322.090391	71.446958	3763.445179	110.801517
1931.687894	348.354645	108.403571	22.951690	208.041941	97.545358	71.807955	3725.190760	134.392555
1942.595777	901.519947	146.252208	23.081757	351.231874	864.725484	76.178075	3705.672523	226.222760
1962.815063	450.162904	96.774928	24.017798	321.824505	105.594833	73.196463	3694.298044	224.098213
1952.477882	875.447875	41.429139	22.458023	288.312124	155.638229	72.686248	3693.676531	90.546147
2035.926322	1432.754588	79.491398	18.888497	277.001756	394.155471	73.268805	3689.223681	345.065258
1896.194305	581.094072	155.314446	24.628883	309.775503	520.511064	70.266007	3660.450210	183.274296
2030.510185	1024.012541	54.316672	25.347284	319.968751	494.514149	73.719703	3656.158363	328.960979
2013.083059	989.043793	108.013185	22.370657	258.991905	32.863455	71.146135	3654.434359	298.092880
1932.046092	11.312943	122.388881	18.518832	327.516557	384.207155	70.916879	3636.892992	89.825568
1884.778813	392.869593	124.743925	18.726909	263.059631	319.200820	72.279862	3628.877193	165.402757
2099.645536	1511.681841	112.750541	20.989594	238.532446	709.878674	72.776358	3616.428136	238.933700



Рисунок—11 Корреляция топ 25 значений «Прочность при растяжении»

Отдельно остановимся на топ 25 максимальных значений целевых признаков. Предположим, что это не случайные аномальные значения. Тогда можно опираться на данные корреляции, которая показывает определенную зависимость (положительную и отрицательную), рисунки 10-11. Также посмотрим на гистограммы признаков внутри топ 25 максимальных значений целевых признаков, рисунки 14-15. Там также прослеживаются интересные закономерности.

Разобьем на три равные группы целевые признаки в каждом из топ 25 максимальных значений целевых признаков. И посмотрим на средние значения наиболее коррелированных признаков, рисунки 12-13.

	Tensile_strength_category	avg_Step_Stripe	avg_Density_stripe_qty	avg_Tensile_modulus_strength_qty	avg_Matrix_filler_ratio_qty	avg_Corner_Stripe_qty
0	1	7.298747	63.432605	73.936213	3.567944	70.00
1	2	5.875649	60.122720	72.132500	3.181264	33.75
2	3	5.462695	51.481903	72.691171	3.052428	33.75

Рисунок 12—Средние значения по группам в топ 25 «Прочность при растяжен, категории»

	Tensile_modulus_strength_category	avg_Step_Stripe	avg_Tensile_modulus_qty	avg_Epoxy_groups_qty	avg_Tensile_strength_qty
0	1	6.975364	763.087913	22.309330	2565.714280
1	2	6.648227	746.874946	21.594429	2305.846474
2	3	4.096014	1002.264240	23.925464	2255.656811

Рисунок 13—Средние значения по группам в топ 25 «Модуль упругости при растяжен, категории»

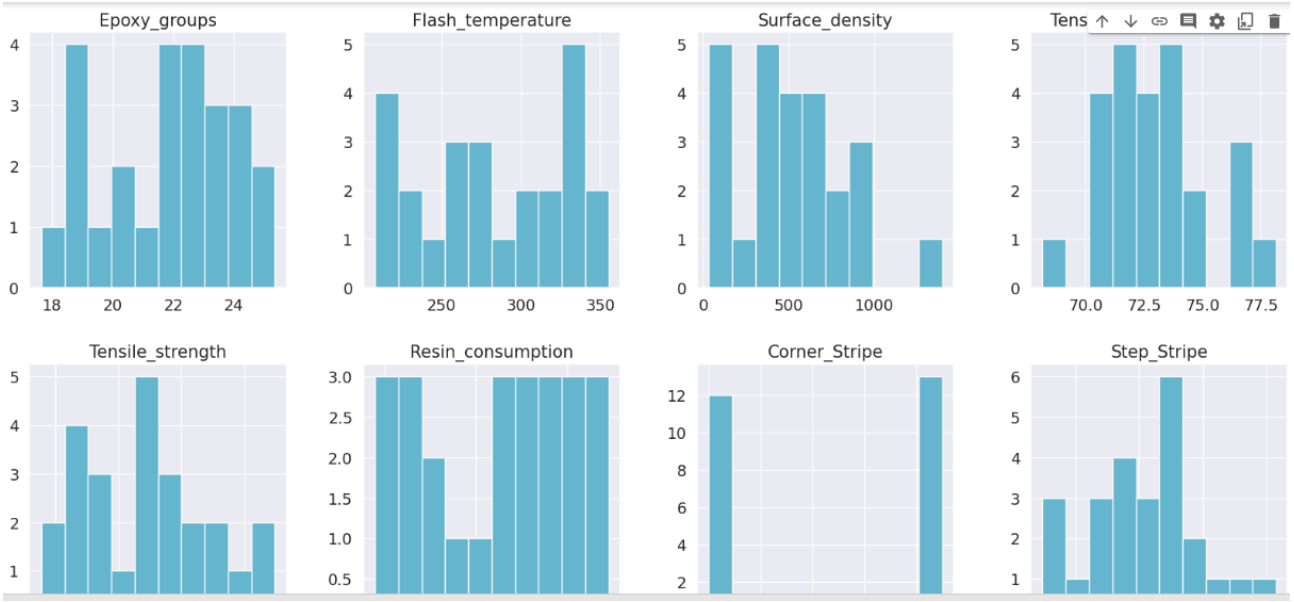


Рисунок 14—Гистограммы топ 25 «Прочность при растяжении»

Т.о. можно продолжить изучение этих топ 25 (рассматривать плюсом категориальные зависимости, задачу классификации), но сгенерировать на их основе дополнительные строки для корректной работы моделей.

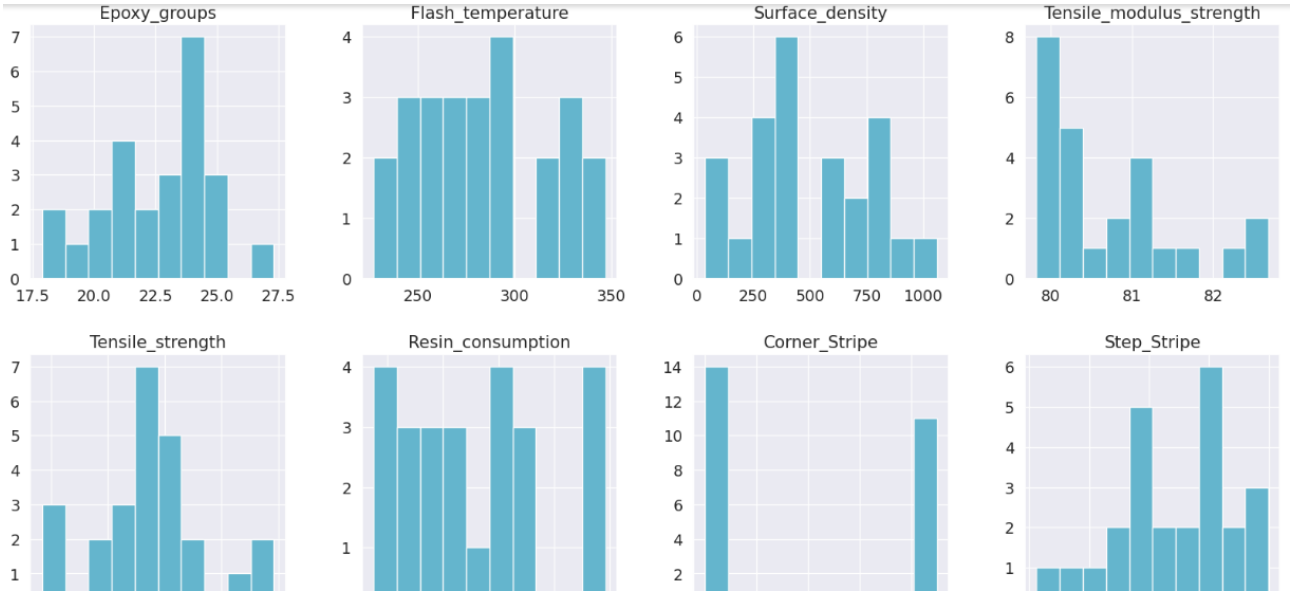


Рисунок15—Гистограммы топ 25 «Модуль упругости при растяжении»

По такому же принципу продолжим работу с полным датасетом. Создадим дополнительный столбец «Модуль упругости при растяжении, категории»: 3категория включает 150 самых больших значений данного признака, 2 и 1 категории создадим как равные половины оставшихся значений.

Создадим дополнительный столбец «Соотношение матрица-наполнитель, категории»:

1) категория 3—значения от 4;
2) категория 2—значения от 2.5 до 4;
3) категория 1—значения от 0 до 2.5.

В этом столбце деление по принципу сбалансированного количества в категориях. Теперь для наших целевых переменных посмотрим агрегированные значения по некоторым признакам, таблицы 9-11:

Таблица—9 Средние значен для «Модуль упругости при растяжен, категории»

	Tensile_modulus_strength_category	avg_Step_Stripe	avg_Tensile_modulus_qty	avg_Epoxy_groups_qty	avg_Tensile_strength_qty	avg_Density_qty
0	1	6.802948	739.093026	22.149070	2477.269302	1977.635080
1	2	7.102289	731.553204	22.283301	2453.874039	1974.652213
2	3	6.595390	761.054827	22.324077	2478.983674	1974.866751

Таблица—10 Средние значен для «Соотношение матрица-наполнит, категории»

	Matrix_filler_ratio_category	avg_Step_Stripe	avg_Density_stripe_qty	avg_Tensile_modulus_strength_qty	avg_Tensile_strength_qty
0	1	6.831130	57.217242	73.349276	2467.626659
1	2	6.977522	57.167344	73.341375	2471.227106
2	3	7.200890	53.935410	71.989191	2289.549552

Таблица—11 Средние значения для «Прочность при растяжении, категории»

nsile_strength_category	avg_Step_Stripe	avg_Density_stripe_qty	avg_Tensile_modulus_qty	avg_Epoxy_groups_qty	avg_Flash_temperature_qty	avg_Surface_density_qty
1	6.980294	56.964980	740.787254	22.194186	286.904074	485.608320
2	6.678774	58.885056	683.023655	22.516057	282.456520	473.277917
3	6.769157	56.970131	760.357120	22.270870	284.501389	478.769895

1.5 Выводы к разделу:

1) датасет после объединения двух исходных таблиц включает 1023 строки и 13 столбцов, два из которых являются целевыми;
2) значения признаков числовые, пропусков в данных нет, нулевых значений нет, дубликатов строк нет, все значения имеют вещественный тип данных, присутствует один неинформативный столбец «Unnamed» (удалили его), дублирующий столбец с индексом;
3) признак «Угол нашивки» имеет два уникальных значения (возможен перевод признака в категориальный), остальные признаки имеют только уникальные значения (за исключением нескольких видимо искусственно сгенерированных строк с 0 по 22, выявленных нами в ходе изучения данных); причины генерации строк не смогли выяснить в процессе работы;
4) все признаки (кроме «Угол нашивки») имеют незначительное колличество выбросов, хотя по определенным признакам однозначно идентифицировать их как выбросы нельзя; необходимо в дальнейшем параллельно работать с датасетом как с выбросами, так и без них, а после сравнить результаты;
5) все признаки имеют очень низкую корреляцию между собой и с целевыми признаками;

6) по характеристикам датасета и основываясь на рекомендуемых в научных публикациях моделях, хорошо показавших себя на композитах, определили перечень моделей для дальнейшей работы: статистический анализ, LinearRegression, Lasso, RandomForestRegressor, KNeighborsRegressor, SVR, DecisionTreeRegressor, AdaBoostRegressor.
7) признаки имеют нормальное распределение, кроме «Поверхностная плотность» и «Модуль упругости» (имеют небольшую ассиметрию), признак «Угол нашивки» имеет бимодальное распределение, все распределения признаков неравномерные;
8) определили среднее значение, медиану, моду, стандартное отклонение, дисперсию;
9) определили квантили 95 и 99 процентные;
10) рассмотрели внимательно статистику по выборкам топ 25 наибольших значений целевых признаков, заметили закономерности;
11) выполнили визуализацию описательного анализа датасета и EDA.

2.Практическая часть

2.1 Предобработка данных

Выбросы не явные, но дальнейшую работу выполним с их удалением. Удалим выбросы, идентифицированные по методу межквартильных расстояний (значения выше верхней границы и ниже нижней границы являются выбросами), посмотрим что получилось, таблица 17.

Тепловая карта Пирсона —это тип визуализации, применяемый, когда нам нужно найти зависимые переменные. Зависимость между признаками так мала, что в соответствии с рисунком 14 можно говорить о ее практическом отсутствии. Максимальная положительная корреляция 0.11 между «Плотность нашивки» и «Угол нашивки», Максимальная отрицательная корреляция минус 0.08 между «Количество отвердителя» и «Прочность при растяжении».

Таблица—12 Датасет без выбросов, метод межквартильных расстояний

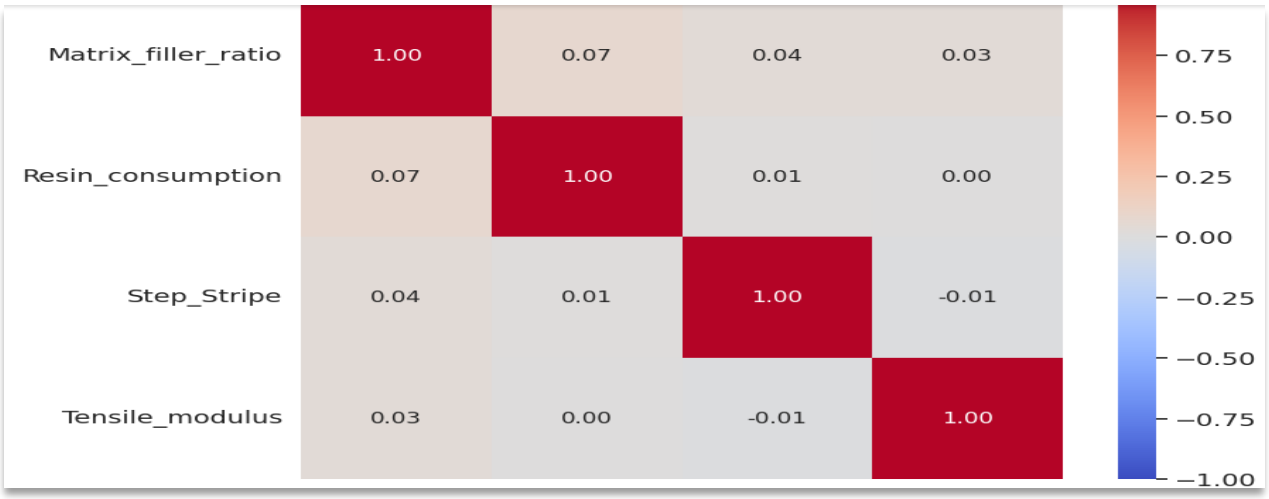
```
Количество выбросов в столбце  Соотношение матрица-наполнитель :  0
Количество выбросов в столбце  Плотность, кг/м3 :  0
Количество выбросов в столбце  модуль упругости, ГПа :  1
Количество выбросов в столбце  Количество отвердителя, м.% :  0
Количество выбросов в столбце  Содержание эпоксидных групп,%_2 :  0
Количество выбросов в столбце  Температура вспышки, С_2 :  0
Количество выбросов в столбце  Поверхностная плотность, г/м2 :  0
Количество выбросов в столбце  Модуль упругости при растяжении, ГПа :  1
Количество выбросов в столбце  Прочность при растяжении, МПа :  4
Количество выбросов в столбце  Потребление смолы, г/м2 :  1
Количество выбросов в столбце  Угол нашивки, град :  0
Количество выбросов в столбце  Шаг нашивки :  0
Количество выбросов в столбце  Плотность нашивки :  3
Общее число ошибок: 10
```

Рисунок—14 Корреляционная тепловая карта Пирсона

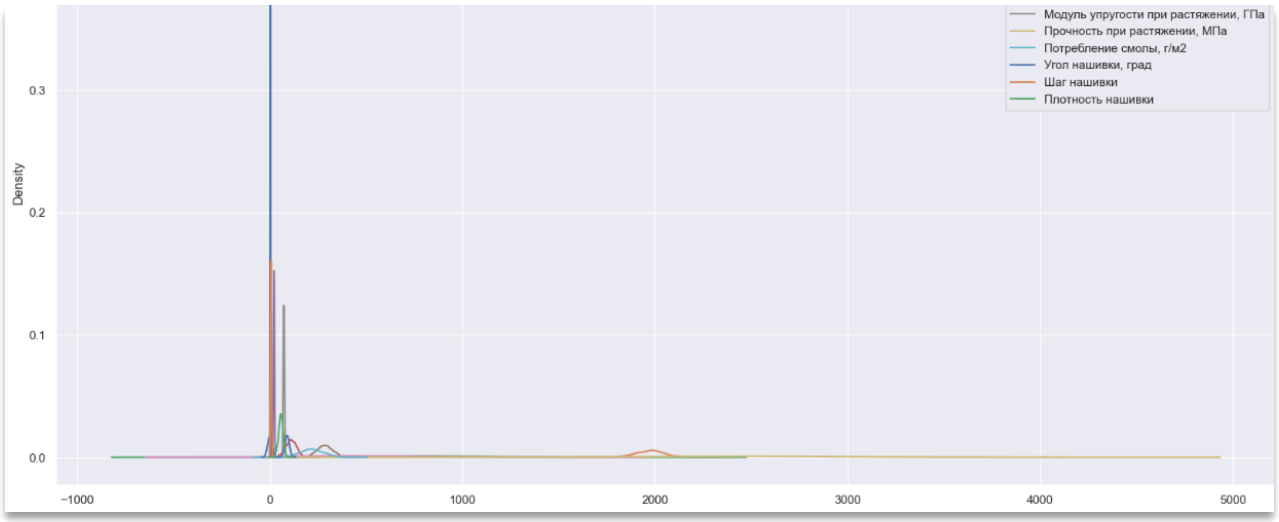
Рассмотрим топ зависимостей:

1) «Соотношение матрица-наполнитель» зависит от «Потребление смолы», «Шаг нашивки», «Модуль упругости» (плюс «Угол нашивки» с отрицательной корреляцией) в соответствии с рисунком 15;
2) «Модуль упругости при растяжении» зависит от «Количество отвердителя» (отрицательная корреляция), «Содержание эпоксидных групп», «Потребление смолы» в соответствии с рисунком 14;

3) «Прочность при растяжении» зависит от «Количество отвердителя» (отрицательная корреляция), «Плотность» (отрицательная корреляция), «Шаг нашивки» (отрицательная корреляция) в соответствии с рисунко 14.



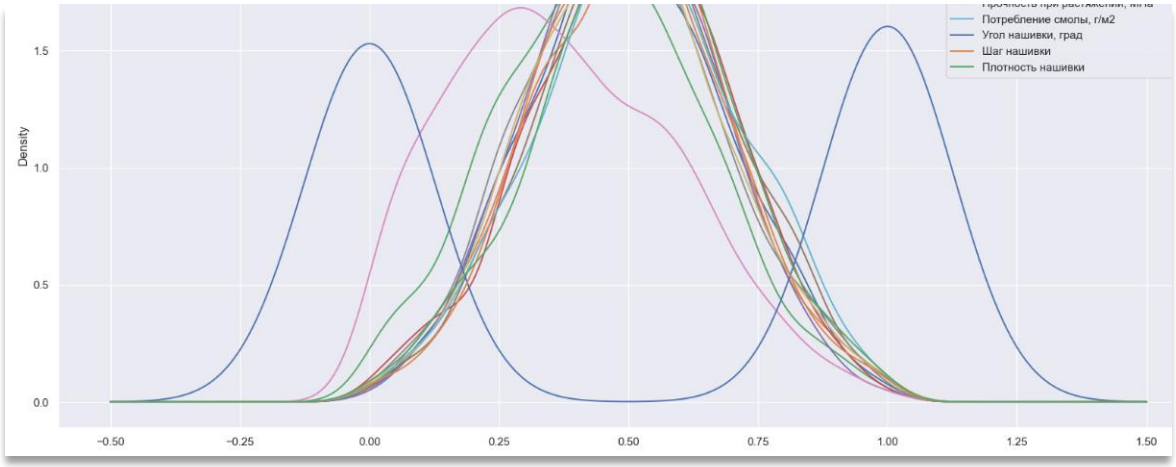
Рисунок—15 Топ 4 по корреляции для «Соотношение матрица-наполнитель»



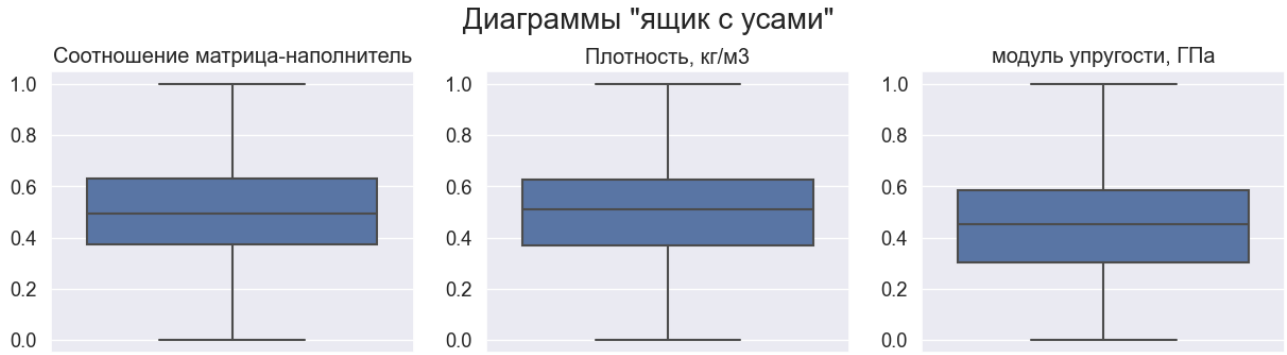
Рисунок—16 Оценка плотности ядра до нормализации

Многие алгоритмы чувствительны к функциям, находящимся в разных масштабах (алгоритмы на основе метрик (KNN, K Means) и алгоритмы на основе градиентного спуска (регрессия, нейронные сети). Древовидные алгоритмы (деревья решений, случайные леса) не имеют этой проблемы. Выполним нормализацию данных, используя метод MinMaxScaler. Графики на рисунке 16

показывают, что численные признаки датасета разных масштабов. Т.к. выбросов очень немного, скорее всего они не повлияют на анализ данных даже если их оставить. Т.к. решили удалить выбросы, то после полного удаления осталось 922 строки. Визуализируем распределение признаков после нормализации (значения приведены в диапазон от 0 до 1), показано на рисунке 17. Смотрим как изменились характеристики признаков после нормализации, таблица 13.



Рисунок—17 Оценка плотности ядра после нормализации



Рисунок—1 «Ящик с усами» после нормализации

Медианное и среднее значение, в соответствии с таблицей 13, в пределах предыдущих значений после удаления выбросов.

Создадим переменную для названия всех столбцов. Это нам пригодится при построении моделей и перейдем к визуализации данных. Для «Модуль упругости при растяжении» и «Прочность при растяжении» топ зависимостей

не изменился, признаки слегка поменялись местами. Для «Соотношение матрица-наполнитель» топ зависимостей также не поменялся, но в топ добавился признак «Плотность нашивки».

Таблица—13 Статистическое описание признаков после нормализации и удаления выбросов

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	936.0	0.498933	0.187489	0.0	0.372274	0.494538	0.629204	1.0
Плотность, кг/м3	936.0	0.502695	0.187779	0.0	0.368517	0.511229	0.624999	1.0
модуль упругости, ГПа	936.0	0.446764	0.199583	0.0	0.301243	0.447061	0.580446	1.0
Количество отвердителя, м.%	936.0	0.504664	0.188865	0.0	0.376190	0.506040	0.637978	1.0
Содержание эпоксидных групп,%_2	936.0	0.491216	0.180620	0.0	0.367716	0.489382	0.623410	1.0
Температура вспышки, С_2	936.0	0.516059	0.190624	0.0	0.386128	0.515980	0.646450	1.0
Поверхностная плотность, г/м2	936.0	0.373733	0.217078	0.0	0.205619	0.354161	0.538683	1.0
Модуль упругости при растяжении, ГПа	936.0	0.488647	0.191466	0.0	0.359024	0.485754	0.615077	1.0
Прочность при растяжении, МПа	936.0	0.495706	0.188915	0.0	0.365149	0.491825	0.612874	1.0
Потребление смолы, г/м2	936.0	0.521141	0.195781	0.0	0.392067	0.523766	0.652447	1.0
Угол нашивки, град	936.0	0.511752	0.500129	0.0	0.000000	1.000000	1.000000	1.0
Шаг нашивки	936.0	0.502232	0.183258	0.0	0.372211	0.504258	0.624604	1.0
Плотность нашивки	936.0	0.513776	0.191342	0.0	0.390482	0.516029	0.638842	1.0

Рисунок—18 Корреляция после нормализации

При выделении из датасета топ 25, 50, 150, 250 самых высоких значений целевых признаков корреляция признаков менялась как по величине, так и по составу топ 4 наиболее зависимых признаков. Не менялся только знак корреляции (положительная или отрицательная);

2.2 Разработка и обучение моделей

Используем следующие модели для анализа:

1) LinearRegression;
2) Lasso;
3) RandomForestRegressor;
4) KNeighborsRegressor;
5) SVR;
6) DecisionTreeRegressor;
7) AdaBoostRegressor.

Для сравнения моделей и оценки точности их работы используем метрики MAE, MSE, MAPE, RMSE, *R*2, Score. Основными будут метрики MAE и R2.

Сначала построим все модели, прогнозирующие «Модуль упругости при растяжении» и «Прочность при растяжении» при стандартных параметрах.

Создаем DataFrame, куда в дальнейшем будем добавлять метрики моделей. Пишем функцию для вывода метрик эффективности, для моделей, переданных в качестве аргумента. Объявим модели и составим из них список, а так же сделаем список их названий. Выделяем целевые переменные. Сначала выделим «Модуль упругости при растяжении».

Разделим датасет на обучающую и тестовую выборки по условию 30% данных оставить на тестирование моделей, на остальных провести обучение моделей. Проверим правильность разделения: видим, что условие деления выполняется, целевой признак отсутствует.

Запускаем функцию для рассчета метрик качества работы моделей «Модуль упругости при растяжении». Оценим их эффективность по сводной таблице, таблица 14. Визуализируем полученные результаты работы моделей, рисунок 19.

Таблица 14—Метрики моделей, прогнозирующих «Модуль упругости при растяжении» при стандартных параметрах

	target	model	MAE	MSE	MAPE	RMSE	R2	score
0	Модуль упругости при растяжении	LinearRegression	0.165	0.039	0.697	0.198	-0.014	0.019
1	Модуль упругости при растяжении	Lasso	0.163	0.039	0.699	0.197	-0.000	0.000
2	Модуль упругости при растяжении	RandomForestRegressor	0.163	0.039	0.692	0.198	-0.013	0.847
3	Модуль упругости при растяжении	KNeighborsRegressor	0.179	0.047	0.741	0.217	-0.215	0.171
4	Модуль упругости при растяжении	SVR	0.177	0.047	0.735	0.216	-0.203	0.451
5	Модуль упругости при растяжении	DecisionTreeRegressor	0.227	0.081	0.800	0.285	-1.096	1.000
6	Модуль упругости при растяжении	AdaBoostRegressor	0.164	0.039	0.695	0.197	-0.002	0.094

Рисунок 19—Визуализация результатов работы одной из моделей, прогнозирующих «Модуль упругости при растяжении» при стандартных параметрах.

Теперь выделим «Прочность при растяжении», разделим датасет на обучающую и тестовую выборки по прежнему условию. Запускаем функцию для рассчета метрик качества работы моделей «Прочность при растяжении». Оценим их эффективность по сводной таблице, таблица 15. Визуализируем полученные результаты работы моделей, рисунок 20.

Таблица 15—Метрики моделей, прогнозирующих «Прочность при растяжении» при стандартных параметрах

7	Прочность при растяжении	LinearRegression	0.147	0.034	0.661	0.184	0.009	-0.014
8	Прочность при растяжении	Lasso	0.148	0.034	0.676	0.185	-0.003	-0.009
9	Прочность при растяжении	RandomForestRegressor	0.150	0.035	0.651	0.187	-0.025	-0.417
10	Прочность при растяжении	KNeighborsRegressor	0.164	0.041	0.725	0.202	-0.199	-0.231
11	Прочность при растяжении	SVR	0.159	0.040	0.726	0.199	-0.164	-0.271
12	Прочность при растяжении	DecisionTreeRegressor	0.217	0.073	0.838	0.270	-1.147	-0.964
13	Прочность при растяжении	AdaBoostRegressor	0.148	0.034	0.673	0.183	0.010	-0.054

Рисунок 20—Визуализация результатов работы одной из моделей, прогнозирующей «Прочность при растяжении» при стандартных параметрах.

По каждой из моделей был проведен поиск сетки гиперпараметров для оптимизации, сравнение работы всех моделей.

2.3 Тестирование модели

После определения лучших параметров для каждой модели произведем тестирование моделей на тренировочном и тестовом наборе данных.

Теперь построим все модели, прогнозирующие «Модуль упругости при растяжении» и «Прочность при растяжении» с использованием лучших гиперпараметров определенных с помощью поиска по сетке с перекрестной

проверкой (метод GridSearchCV, количество блоков равно 10). Оценивать будем с помощью коэффициента детерминации R2.

Составим словари с параметрами для моделей, а так же составляем список из этих словарей. Пишем функцию вывода лучших параметров моделей. Определяем лучшие параметры для моделей, рисунок 21.

```
LinearRegression() - {'fit_intercept': True}
Lasso() - {'alpha': 0.01}
RandomForestRegressor() - {'max_depth': 1, 'max_features': 'log2', 'n_estimators': 64}
KNeighborsRegressor() - {'algorithm': 'auto', 'n_neighbors': 173, 'weights': 'uniform'}
SVR() - {'C': 0.1, 'kernel': 'sigmoid'}
DecisionTreeRegressor() - {'criterion': 'friedman_mse', 'max_depth': 5, 'max_features': 'log2', 'min_samples_leaf': 40,
AdaBoostRegressor() - {'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 8}
```

Рисунок 21—Лучшие гиперпараметры работы моделей

Создадим модели с подобранными для них лучшими гиперпараметрами для «Модуль упругости при растяжении», и составим список из этих моделей. Пишем функцию и выводим метрики эффективности моделей в сводной таблице, таблица 16. Визуализируем полученные результаты работы моделей, рисунок 22.

Таблица 16—Метрики моделей, прогнозирующих «Модуль упругости при растяжении» с лучшими гиперпараметрами

14	Модуль упругости при растяжении (с подбором па...	LinearRegression	0.165	0.039	0.697	0.198	-0.014	0.019
15	Модуль упругости при растяжении (с подбором па...	Lasso	0.163	0.039	0.699	0.197	-0.000	0.000
16	Модуль упругости при растяжении (с подбором па...	RandomForestRegressor	0.163	0.039	0.695	0.197	-0.001	0.018
17	Модуль упругости при растяжении (с подбором па...	KNeighborsRegressor	0.165	0.039	0.700	0.198	-0.008	0.011
18	Модуль упругости при растяжении (с подбором па...	SVR	0.163	0.039	0.693	0.197	-0.001	0.002
19	Модуль упругости при растяжении (с подбором па...	DecisionTreeRegressor	0.164	0.039	0.705	0.198	-0.010	0.010
20	Модуль упругости при растяжении (с подбором па...	AdaBoostRegressor	0.163	0.039	0.706	0.197	0.001	0.043

Рисунок 22—Визуализация результатов работы одной из моделей, прогнозирующих «Модуль упругости при растяжении» с лучшими гиперпараметрами

Создадим модели с подобранными для них лучшими гиперпараметрами для «Прочность при растяжении», и составим список из этих моделей. Пишем функцию и выводим метрики эффективности моделей в сводной таблице, таблица 17. Визуализируем полученные результаты работы моделей, рисунок 23.

Таблица 17—Метрики моделей, прогнозирующих «Прочность при растяжении» с лучшими гиперпараметрами

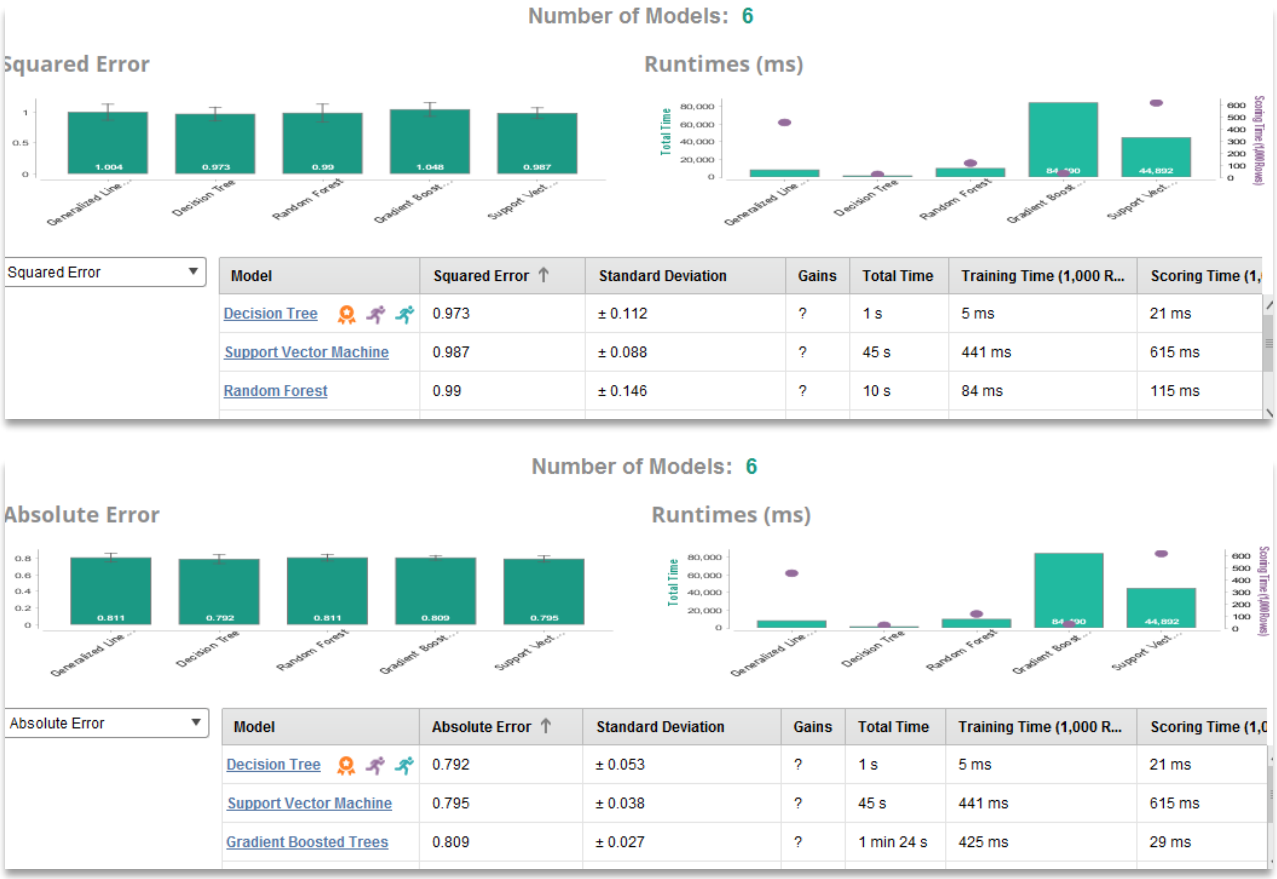
21	Прочность при растяжении (с подбором параметров)	LinearRegression	0.147	0.034	0.661	0.184	0.009	-0.014
22	Прочность при растяжении (с подбором параметров)	Lasso	0.148	0.034	0.676	0.185	-0.003	-0.009
23	Прочность при растяжении (с подбором параметров)	RandomForestRegressor	0.147	0.034	0.666	0.184	0.005	-0.011
24	Прочность при растяжении (с подбором параметров)	KNeighborsRegressor	0.147	0.034	0.678	0.185	-0.002	-0.006
25	Прочность при растяжении (с подбором параметров)	SVR	0.147	0.034	0.669	0.184	-0.000	-0.005
26	Прочность при растяжении (с подбором параметров)	DecisionTreeRegressor	0.148	0.034	0.640	0.184	0.001	-0.014
27	Прочность при растяжении (с подбором параметров)	AdaBoostRegressor	0.147	0.034	0.641	0.184	0.008	-0.013

Рисунок 23—Визуализация результатов работы одной из моделей, прогнозирующих «Прочность при растяжении» с лучшими гиперпараметрами

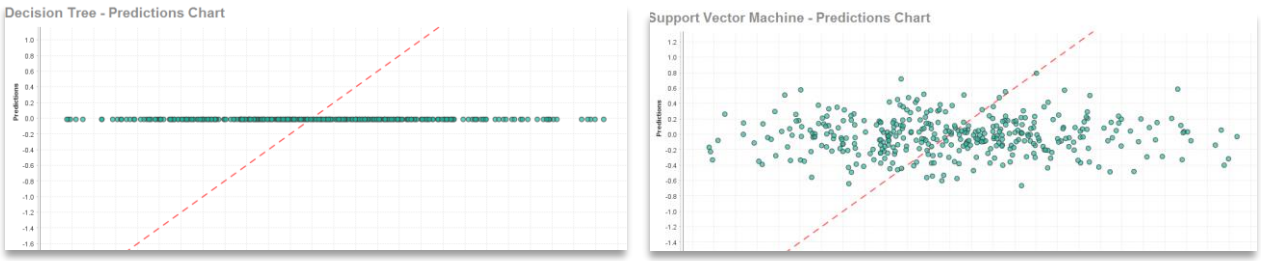
Для сравнения используем построение моделей в программной многопользовательской платформе RapidMiner. Для целевых переменных (для каждой по отдельности) построим модели (Generalised Linear Model, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine) в RapidMiner. Изначально, на вход подали датасет без выбросов, стандартизированный: 1) соглашаемся с

автоматическим выбором и ранжированием признаков для построения моделей; 2) соглашаемся с авто настройками параметров моделей.

Все оставляем по умолчанию, т.к. не разбираемся в датасете. Смотрим на результаты на рисунках 24-25:



Рисунок—24 Обобщенная таблица результатов обучения моделей в RapidMiner



Рисунок—25 Предсказание модели в RapidMiner

По итогу обучения все модели показали плохой результат, лучшими из них стали Decision Tree по обоим целевым переменным, таблица 18.

Выполнили также вариант с настройкой параметров моделей Random Forest, Gradient Boosted Trees: изменили колличество деревьев до 100 (по умолчанию

было 20), а глубину сократили до 5 (по умолчанию было 20). Результат улучшился совсем незначительно, нужно еще пробовать разные настройки.

Таблица—18 Оценка точности моделей по метрикам в RapidMiner

«Модуль упругости при растяжении»					
Модели	**Оценка точности моделей по метрикам**				
	MSE (среднеквадратичная ошибка)	**RMSE** (корень из среднеквадратичной ошибки)	**MAE** (абсолютная ошибка)	**MAPE** (средняя абсолютная ошибка в процентах)	**R-квадрат** (коэффициент детерминации)
Decision Tree	0,973	0,985	0,792		
Support Vector Machine	0,987	0,993	0,795		
Random Forest	0,99	0,993	0,811		
Generalised Linear Model	1,004	1	0,811		
Gradient Boosted Trees	1,048	1,023	0,809		
«Прочность при растяжении»					
Decision Tree	1,028	1,012	0,806		
Support Vector Machine	65,95	256,08	250,73		
Random Forest	1,058	1,025	0,82		
Generalised Linear Model	1,052	1,023	0,806		
Gradient Boosted Trees	1,05	1,022	0,812		

Для «Прочность при растяжении» метрики модели SVM показали результат, похожий на аномальный. Поэтому выполнили вариант с удалением части признаков (на вход обучения подали 9 признаков), в итоге метрики получились сопоставимые с другими.

Выполнили также вариант с подачей на вход датасета стандартизированного, но с выбросами. Метрики результата отличались от варианта с полностью обработанным датасетом на 10%.

Выполнили также вариант с подачей на вход датасета не стандартизированного, с выбросами. Метрики результата отличались незначительно.

Лучшая модель для прогноза «Модуль упругости при растяжении» по метрике коэффициента детерминации R2 AdaBoostRegressor с лучшими гиперпараметрами.

Лучшая модель для прогноза «Прочность при растяжении» по метрике коэффициента детерминации R2 также AdaBoostRegressor со стандартными параметрами. В целом предсказания моделей по обоим целевым переменным показали плохой результат как со стандартными параметрами, так и с найденными лучшими гиперпараметрами. Среднеквадратичная ошибка (MSE) сравнительно небольшая, абсолютная ошибка (MAE) средняя. Коэффициент детерминации для

«Модуля упругости при растяжении» в лучшей модели 0,001, а для» Прочности при растяжении» в лучшей модели 0,010.

Таблица—19 Оценка точности моделей по метрикам

Оценка точности моделей по метрикам

«Модуль упругости при растяжении»

Модели	MAE (абсолютная ошибка)		MSE (среднеквадратичная ошибка)		MAPE (средняя абсолютная ошибка в процентах)		RMSE (корень из среднеквадратичной ошибки)		R-квадрат (коэффициент детерминации)		SCORE	
	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)
LinearRegression	0.165	0.165	0.039	0.039	0.697	0.697	0.198	0.198	-0.014	-0.014	0.019	0.019
Lasso	0.163	0.163	0.039	0.039	0.699	0.699	0.197	0.197	-0.000	-0.000	0.000	0.000
RandomForestRegres	0.163	0.163	0.039	0.039	0.692	0.695	0.198	0.197	-0.013	-0.001	0.847	0.018
KNeighborsRegresso	0.179	0.165	0.047	0.039	0.741	0.700	0.217	0.198	-0.215	-0.008	0.171	0.011
SVR	0.177	0.163	0.047	0.039	0.735	0.693	0.216	0.197	-0.203	-0.001	0.451	0.002
DecisionTreeRegress	0.227	0.164	0.081	0.039	0.800	0.705	0.285	0.198	-1.096	-0.010	1.000	0.010
AdaBoostRegressor	0.164	0.163	0.039	0.039	0.695	0.706	0.197	0.197	-0.002	0.001	0.094	0.043
Лучшее значение	0.163	0.163	0.039	0.039	0.692	0.693	0.197	0.197	-0.000	0.001	0.000	0.000

Оценка точности моделей по метрикам

«Прочность при растяжении»												
LinearRegression	0.147	0.147	0.034	0.034	0.661	0.661	0.184	0.184	0.009	0.009	-0.014	-0.014
Lasso	0.148	0.148	0.034	0.034	0.676	0.676	0.185	0.185	-0.003	-0.003	-0.009	-0.009
RandomForestRegres	0.150	0.147	0.035	0.034	0.651	0.666	0.187	0.184	-0.025	0.005	-0.417	-0.011
KNeighborsRegresso	0.164	0.147	0.041	0.034	0.725	0.678	0.202	0.185	-0.199	-0.002	-0.231	-0.006
SVR	0.159	0.147	0.040	0.034	0.726	0.669	0.199	0.184	-0.164	-0.000	-0.271	-0.005
DecisionTreeRegress	0.217	0.148	0.073	0.034	0.838	0.640	0.270	0.184	-1.147	0.001	-0.964	-0.014
AdaBoostRegressor	0.148	0.147	0.034	0.034	0.673	0.641	0.183	0.184	0.010	0.008	-0.054	-0.013
Лучшее значение	0.147	0.147	0.034	0.034	0.651	0.640	0.183	0.184	0.010	0.009	-0.009	-0.005

2.4 Написание нейронной сети, рекомендующей соотношение «матрица-наполнитель»

Построим полносвязную нейронную сеть (FF), с помощью класса keras.Sequential. На входе у нейронной сети датасет, на выходе—«Соотношение матрица-наполнитель». Выделим целевую переменную и удалим ее из DataFrame. Разделим выборку на обучающую и тестовую.

Входной слой, функция активации relu, полносвязный слой с 5 нейронами, функция активации relu, выходной слой с одним линейным нейроном, функция активации tanh, рисунок 26.

```
model = Sequential()
model.add(Dense(10, activation='relu', input_shape=(X_train.shape[1],)))
model.add(Dense(5, activation='relu'))
model.add(Dense(1, activation='tanh'))
```

Рисунок—26 Архитектура модели нейросети

```
Model: "sequential_20"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_66 (Dense)            (None, 10)                130

 dense_67 (Dense)            (None, 5)                 55

 dense_68 (Dense)            (None, 1)                 6

=================================================================
Total params: 191
Trainable params: 191
Non-trainable params: 0
```

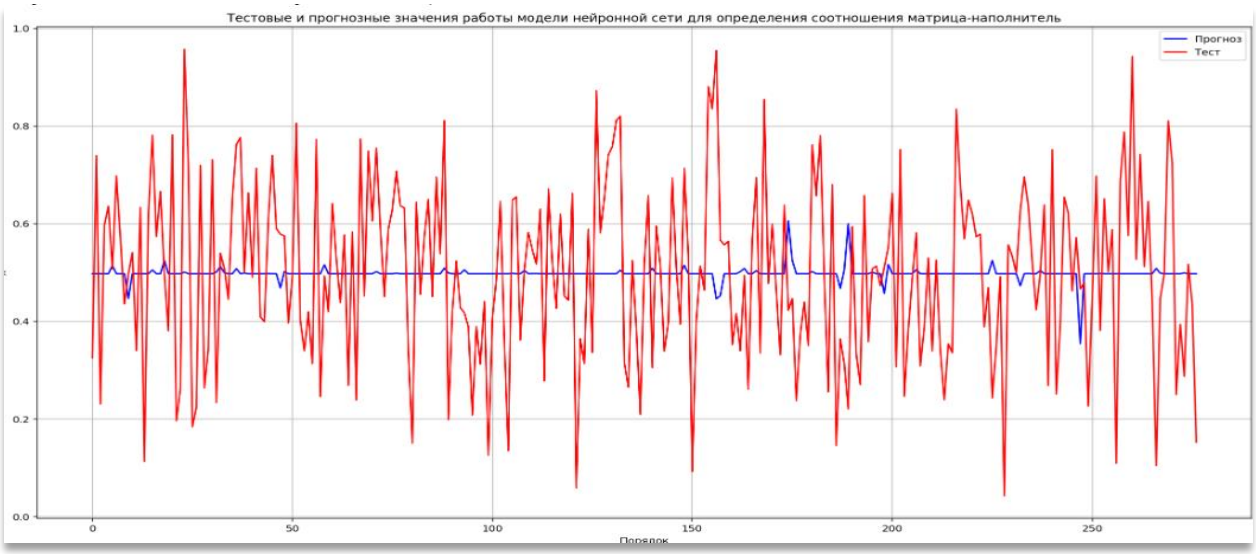
Рисунок—26 Архитектура модели нейросети

Компилируем модель (определяем метрики, алгоритм оптимизации) и обучаем нейронную сеть, рисунок 27. Выбран оптимизатор Adam, функция ошибки MSE, метрика MAE.

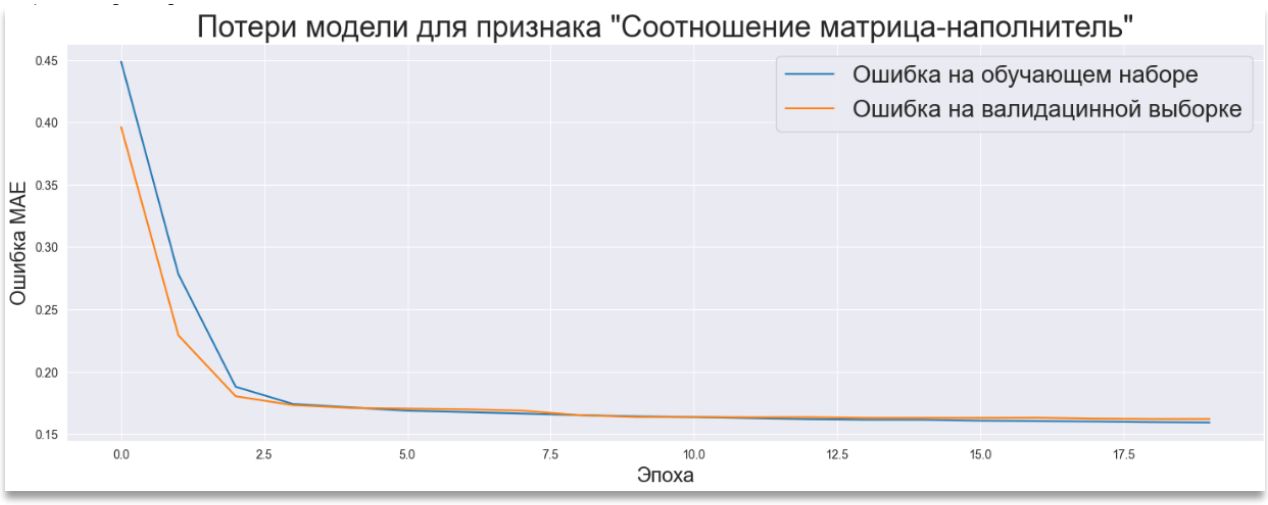
```
model.summary()
model.compile(optimizer='Adam', loss='mae', metrics=['mae'])
history = model.fit(X_train, y_train, validation_split=0.1, verbose=1, epochs=20)
```

Рисунок—27 Компиляция модели нейросети

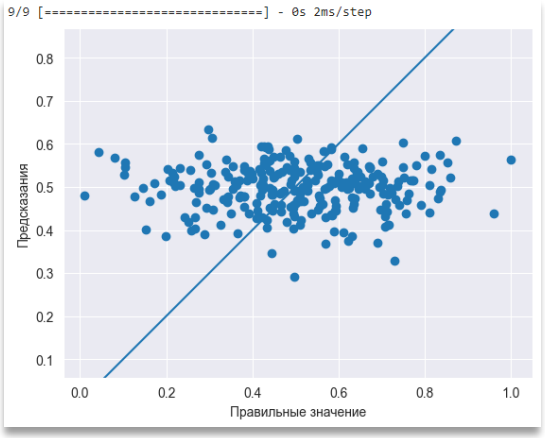
Визуализируем тестовую и прогнозную работу модели нейросети, рисунок 28.



Рисунок—28 Визуализация тестовой и прогнозной работы модели нейросети



Рисунок—29 Графики ошибки модели нейросети



Рисунок—30 Разброс предсказаний модели нейросети

Предсказанные значения довольно плохо соотносятся с правильными ответами (точки предсказаний не проецируется вдоль линии), рисунок 30.

Результаты работы нейросети оценим по основным метрикам в таблице 20. Как видим по MAE нейронная сеть отработала хуже, чем отдельные регрессоры при построении моделей.

Таблица 20—Основные метрики нейросети

	Модель	MAE	MSE	R2
0	NS_train	0.152	0.035	-0.095
1	NS_test	0.159	0.040	-0.080

Выполним сохранение модели нейросети, рисунок 31.

```
[ ]  from joblib import dump, load
     with open('model.pkl', 'wb') as f:
         dump(model, f)
```

Рисунок—31 Сохранение модели нейросети

Проверим загрузку модели нейросети, рисунок 32.

```
Model: "sequential_20"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_66 (Dense)            (None, 10)                130

 dense_67 (Dense)            (None, 5)                 55

 dense_68 (Dense)            (None, 1)                 6

=================================================================
Total params: 191
Trainable params: 191
Non-trainable params: 0
_________________________________________________________________
```

Рисунок—32 Загрузка модели нейросети

2.5 Разработка приложения

На основе имеющегося набора данных, используя фреймворк Flask, разработано приложение, прогнозирующее «Соотношение матрица-наполнитель».

В приложении вводятся значения всех признаков, соответствующие набору исходных данных, а полученный ответ соответствует прогнозу параметра «Соотношение матрица-наполнитель» при заданных признаках.

Инструкция по использованию приложения:

1) Запустить приложение, перейти по ссылке (в блокноте);
2) ввести значения признаков;
3) нажать на кнопку «Вычислить».

2.6 Создание удаленного репозитория и загрузка результатов работы

На GitHub создан репозиторий.

Ссылка на репозиторий: https://github.com/StoyanovP/VKR

Заполнен README.

Файлы по ВКР загружены в репозиторий.

2.7 Выводы к разделу:

1) выбросы нельзя однозначно интерпретировать как выбросы, соответственно нужно рассматривать варианты дальнейшей работы с удалением и без удаления выбросов;
2) удаление выбросов, практически не оказало влияние на работу моделей прогноза, подтверждает предположение, что это не классические выбросы;
3) RapidMiner показал сопоставимые одинаково неудовлетворительные результаты как на исходных данных (без удаления выбросов и стандартизации), так и на предобработанных (без выбросов, стандартизированных данных). Существенная разница получилась только в одном варианте модели прогноза методом SVN (при настройке обучения удалили половину признаков и т.п.);
4) нельзя полагаться на стандартные методы работы с датасетом и стандартные настройки Avto ML, т.к корреляция не показала значимые зависимости, а наоборот практически полное отсутствие. Нужна помощь специалиста, владеющего всем объемом методов машинного обучения и интерпритации результатов;
5) датасет в 1000 строк является небольшим, сформировать достаточную выборку наивысших значений целевых признаков для отдельного изучения не представляется возможным;
6) вариант с изучением выборок из датасета топ 25 наивысших значений целевых признаков показал достаточно интересные результаты, возможно нужно было остановиться на нем подробнее;
7) также определенные закономерности выявило деление на группы (по три группы, по десять групп) значений целевых признаков, возможно надо

было двигаться дальше в этом направлении и поработать с моделями классификации;

8) свод результатов работы всех моделей прогноза представлен в обобщающей таблице;
9) с учетом полученных неудовлетворительных результатов в качестве прогноза для модуля упругости при растяжении и прочности при растяжении по предоставленному датасету предлагается использовать среднее значение признаков (датасет без удаления выбросов) либо строить прогнозы на выборках топ 25 с генерацией дополнительных строк;
10) без консультации с экспертом по свойствам композитов, изучение данных и интерпретация результатов работы моделей, на мой взгляд, слабо отличается от метода работы на угад, или метода «черного ящика», выражаясь языком тестировщиков.

Заключение

За композитами—будущее, но на данном этапе развития отрасли, есть ряд существенных недостатков:

1) анизотропия – одни и те же свойства композитного материала могут в десятки раз различаться в зависимости от направления внешнего воздействия (вдоль волокон или поперек);
2) большой удельный объем;
3) гигроскопичность;
4) токсичность (при изготовлении и в процессе эксплуатации эти материалы могут выделять вредные для здоровья человека пары);
5) высокая цена (при производстве часто используется дорогостоящее оборудование) и т.п.

Создание прогнозных моделей поможет сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

В данной работе выполнены все этапы Pipeline построения моделей машинного обучения на структурированных данных, а также рассмотрены теоретические основы работы с Big Data. Изучены характеристики данных, проведен их анализ, выполнена предобработка, построены предсказывающие значения целевых переменных модели, сделаны выводы, разработано приложение для автоматизации прогноза.

Результаты работы в целом неудовлетворительные. Без консультации с экспертом по свойствам композитов, без постановки более точной бизнес-задачи нет возможности правильно идентифицировать выбросы, оценить корреляцию, также нельзя выбрать и настроить подходящую модель (т.к. не можем отобрать нужные признаки, определить их вес и т.п.).

Для достижения лучшего результата важно, с учетом полученных разъяснений от эксперта и бизнеса, определить новую оптимальную стратегию работы с данными и заново выполнить весь Pipeline.

Список литературы

1 Википедия, Композитные материалы [Электронный ресурс]: – Режим доступа: https://www.ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D0%BE%D0%B7%D0%B8%D1%82%D0%BD%D1%8B%D0%B9_%D0%BC%D0%B0%D1%82%D0%B5%D1%80%D0%B8%D0%B0%D0%BB. (дата обращения: 26.03.2023).

2 Библиотека Tensorflow [Электронный ресурс]: – Режим доступа: https://www.tensorflow.org/api_docs/python/tf/keras/metrics. (дата обращения: 28.03.2023).

3 Библиотека keras [Электронный ресурс]: – Режим доступа: https://keras.io/api/metrics/. (дата обращения: 28.03.2023).

4 Библиотека matplotlib [Электронный ресурс]: – Режим доступа: https://matplotlib.org/stable/plot_types/index.html. (дата обращения: 28.03.2023).

5 Библиотека numpy [Электронный ресурс]: – Режим доступа: https://numpy.org/doc/1.22/user/c-info.html. (дата обращения: 28.03.2023).

6 Библиотека pandas [Электронный ресурс]: – Режим доступа: https://pandas.pydata.org/docs/user_guide/io.html. (дата обращения: 28.03.2023).

7 Библиотека scikit-learn [Электронный ресурс]: – Режим доступа: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics. (дата обращения: 28.03.2023).

8 Библиотека seaborn [Электронный ресурс]: – Режим доступа: https://seaborn.pydata.org/tutorial/relational.html. (дата обращения: 28.03.2023).

9 Документация по языку программирования python [Электронный ресурс]: – Режим доступа: https://docs.python.org/3.10/. (дата обращения: 28.03.2023)

10 Инжиниpиум МГТУ им. Н.Э. Баумана, Интересные факты о композитах [Электронный ресурс]: – Режим доступа: https://dzen.ru/a/XsPAVKI-FWARPzyn. (дата обращения: 26.03.2023).

11 Кодкамп, Полное руководство: когда удалять выбросы в данных [Электронный ресурс]: – Режим доступа: https://www.codecamp.ru/blog/remove-outliers/. (дата обращения: 26.03.2023).

12 Независимая газета, Будущее сделано из композитов [Электронный ресурс]: – Режим доступа: https://www.ng.ru/nauka/2022-06-07/13_8455_future.html. (дата обращения: 26.03.2023).

13 Реутов Ю.А.: Прогнозирование свойств полимерных композиционнных материалов и оценка надёжности изделий из них, Диссертация, Томск 2016: – Режим доступа [Электронный ресурс]: http://ams.tsu.ru/TSU/QualificationDep/co-searchers.nsf/ECF749E40C9E580247258044003491 89/$file/%D0%A0%D0%B5%D1%83%D1%82%D0%BE%D0%B2_%D0%AE.%D0%90._%D0%94%D0%B8%D1%81%D1%81%D0%B5%D1%80%D1%82%D0%B0%D1%86%D0%B8%D1%8F.pdf. (дата обращения: 28.03.2023)

14 Среда разработки Jupyter Notebook [Электронный ресурс]: – Режим доступа: https://docs.jupyter.org/en/latest/projects/content-projects.html. (дата обращения: 28.03.2023)

15 Хабр, История композиционных материалов [Электронный ресурс]: – Режим доступа: https://habr.com/ru/post/362189/. (дата обращения: 26.03.2023).

16 Чун-Те Чен и Грейс Х. Гу. Машинное обучение для композитных материалов (март 2019г.) [Электронный ресурс]: – Режим доступа: https://www.cambridge.org/core/journals/mrs-communications/article/machine-%20learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15. (дата обращения: 28.03.2023)

17 LibTime, Как создают композиты [Электронный ресурс]: – Режим доступа: https://libtime.ru/science/kak-sozdayut-kompozity.html. (дата обращения: 26.03.2023).

18 Longinom [Электронный ресурс]: – Режим доступа: https://wiki.loginom.ru/visualization.html. (дата обращения: 28.03.2023).