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……………………………………………………………………….3

1 Аналитическая часть………………………………………………………..6

* 1. Постановка задачи………………………………………………….6
  2. Характеристика датасета…………………………………………..6

1.3 Описание используемых методов………………………………...12

1.4 Разведочный анализ данных……………………………………...19

1.5 Выводы к разделу………………………………………………….27

2.Практическая часть………………………………………………………...28

2.1 Предобработка данных……………………………………………28

2.2 Разработка и обучение модели……………………………………33

2.3 Тестирование модели……………………………………………..35

2.4 Написание нейронной сети, рекомендующей соотношение «матрица-наполнитель»…………………………………………………………..40

2.5 Разработка приложения…………………………………………...43

2.6 Создание репозитория, загрузка результатов работы…………...44

2.7 Выводы к разделу………………………………………………….44

Заключение…………………………………………………………………...45

Список использованной литературы……………………………………….46

**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.

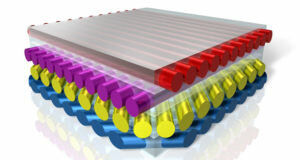


Рисунок 1—Структура композиционного материала

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](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D0%B8%D0%BC%D0%B5%D1%80%D1%8B) 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](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%B0%D0%BB%D0%BB) в качестве матрицы применяют [алюминий](https://ru.wikipedia.org/wiki/%D0%90%D0%BB%D1%8E%D0%BC%D0%B8%D0%BD%D0%B8%D0%B9), [магний](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D0%B3%D0%BD%D0%B8%D0%B9), [никель](https://ru.wikipedia.org/wiki/%D0%9D%D0%B8%D0%BA%D0%B5%D0%BB%D1%8C), [медь](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B4%D1%8C) и так далее. [Наполнителем](https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D0%BF%D0%BE%D0%BB%D0%BD%D0%B8%D1%82%D0%B5%D0%BB%D1%8C) служат или высокопрочные [волокна](https://ru.wikipedia.org/wiki/%D0%92%D0%BE%D0%BB%D0%BE%D0%BA%D0%BD%D0%BE), или тугоплавкие, не растворяющиеся в основном металле частицы различной [дисперсности](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%BD%D0%BE%D1%81%D1%82%D1%8C).

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**

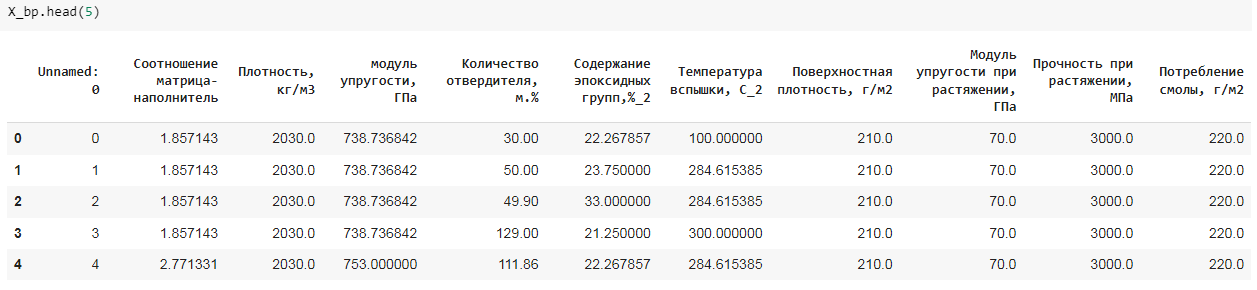
Начнем с описательного анализа данных, в большинстве случаев он используется для первичного определения типов информации.

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) 



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

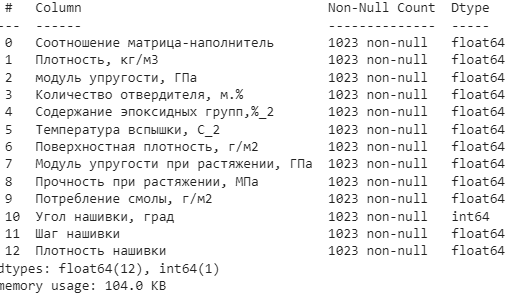
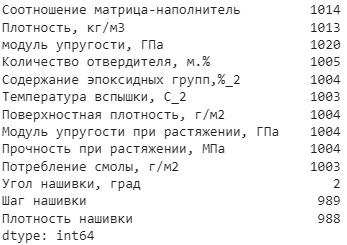
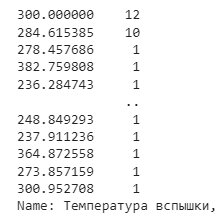
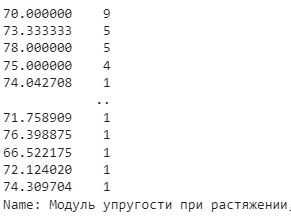


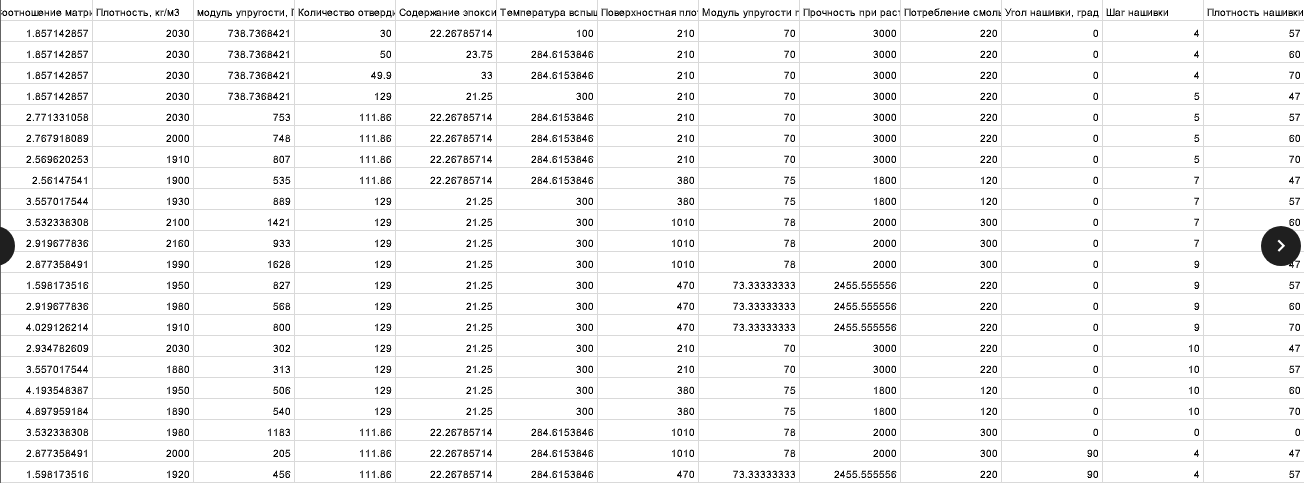
Таблица 3—Наличие уникальных значений



Таблица—4 Наиболее часто встречающиеся значения

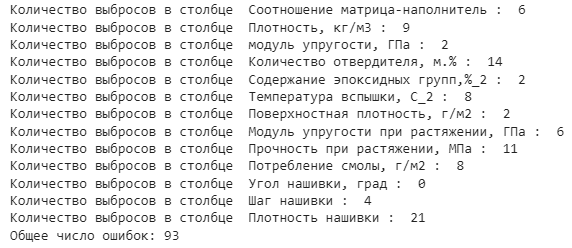
 

Таблица—5 Сгенерированные строки в датасете с 0 по 22



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

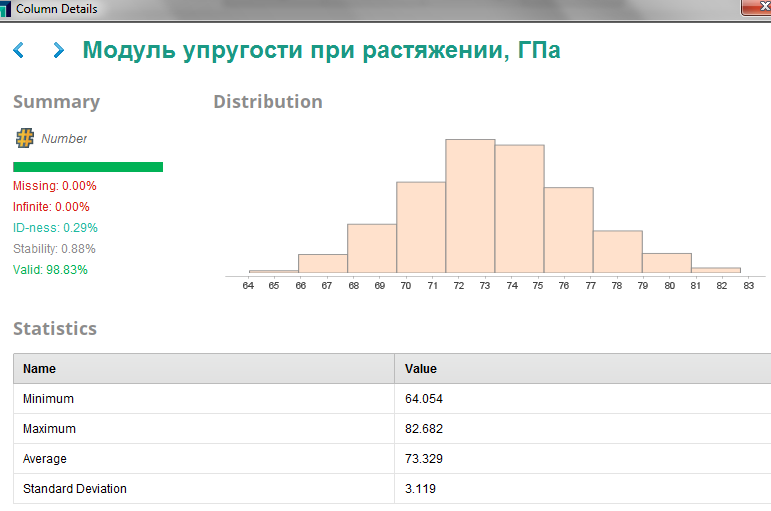
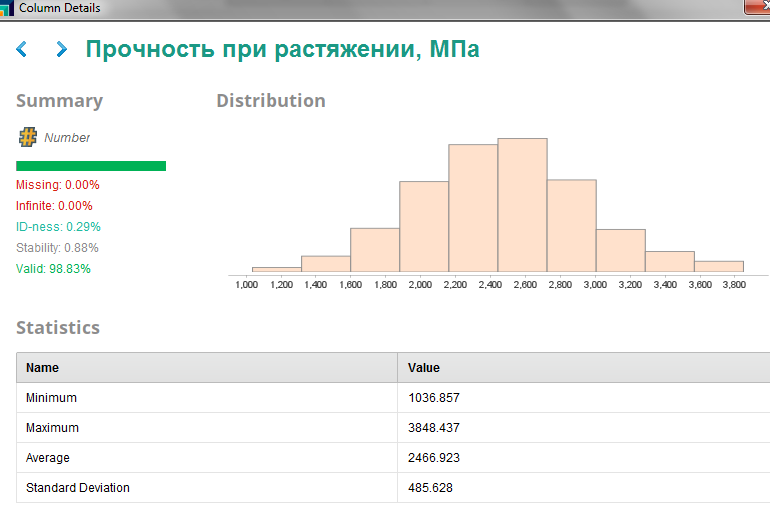
Таблица—6 Количество предполагаемых выбросов в датасете



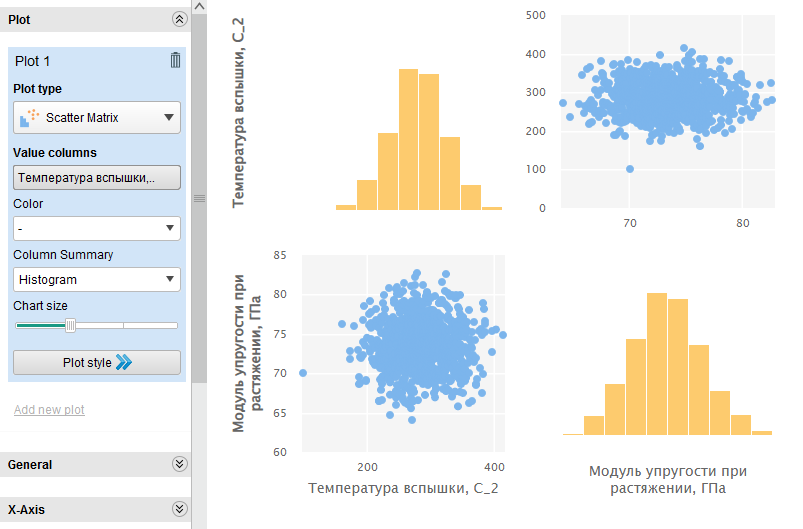
Чтобы идентифицировать их таковыми и принять решение удалить, нужно выбросы оценить, иначе информация может быть упущена (выполним это в подразделе 1.4). В сочетании с фактическими данными выбросы можно разделить на "истинные аномалии" и "ложные аномалии". Экспертная оценка выбросов нам не доступна (не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

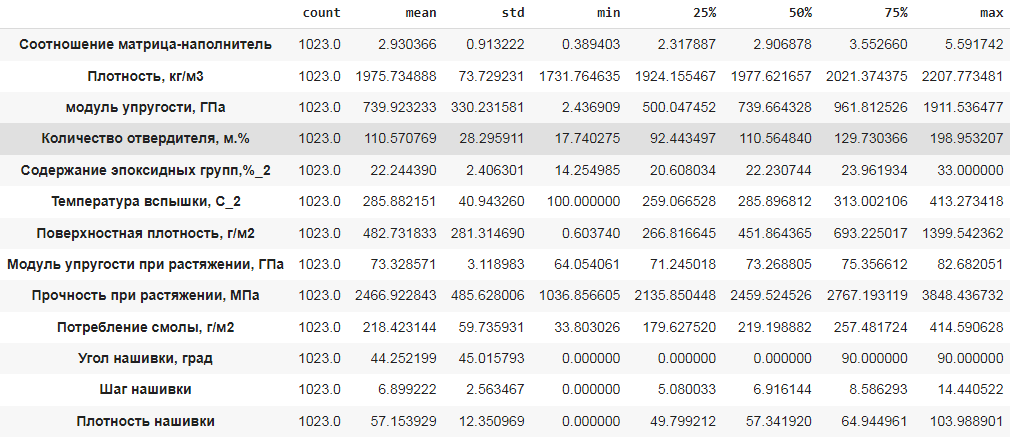


Рисунок—3 Визуализация в RapidMiner

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



**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 (МSE) 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. Помимо 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.

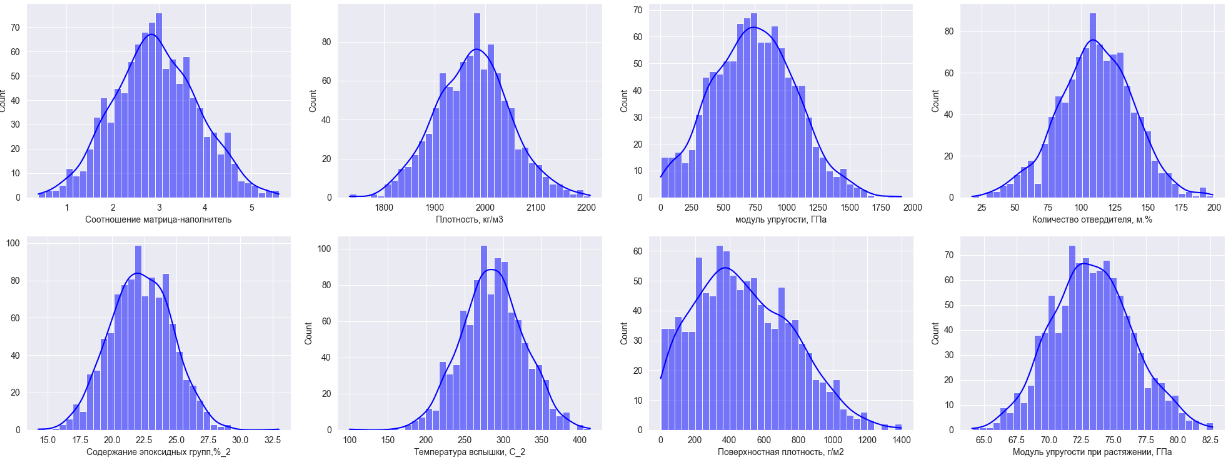
К основным методам разведочного анализа относится:

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.

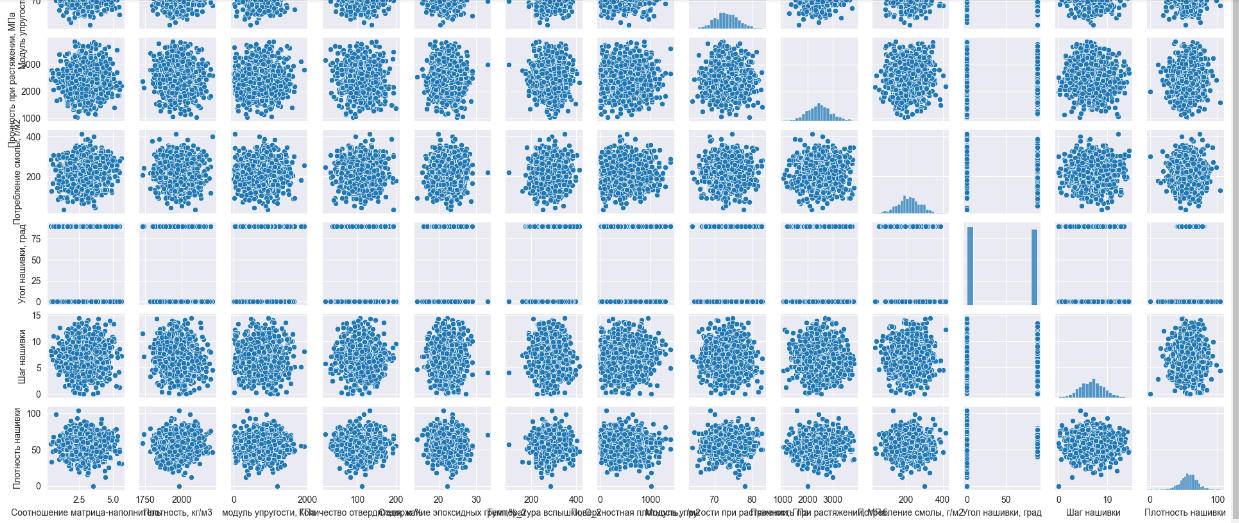
Визуализации выделяют взаимосвязи в данных и раскрывают информацию, видимую человеческому глазу, которую нельзя передать только числами и цифрами. Используем для этого библиотеки Pandas, Seaborn, Matplotlib.

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.

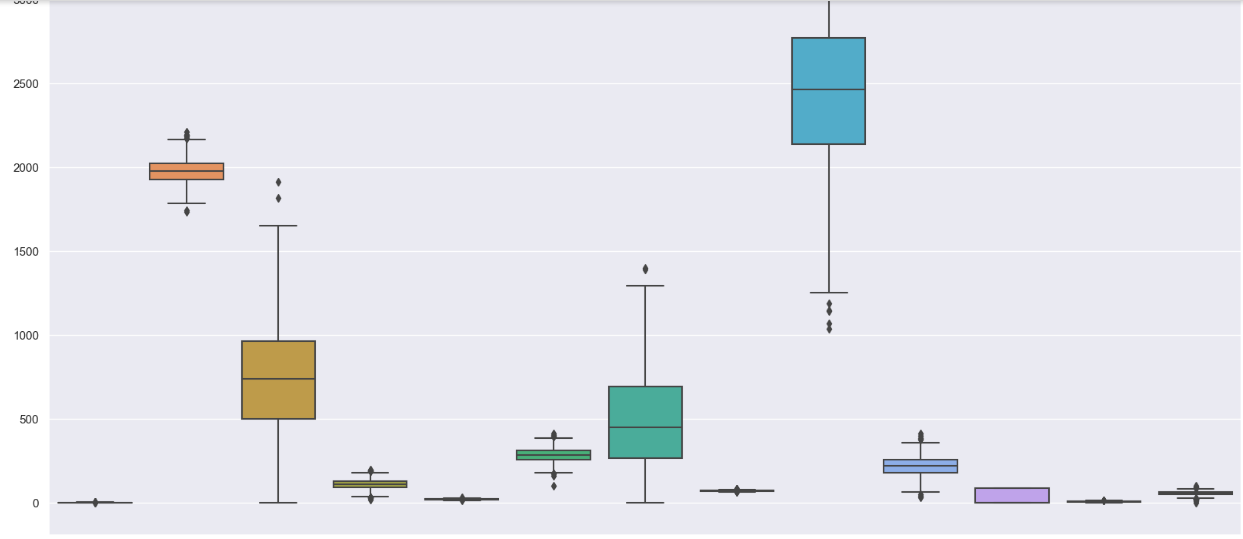
Проведем визуальный поиск наличия аномалий и выбросов с помощью метода Boxplot (ящик с усами), рисунок 6. Точки за границами "усов" (полутора межквартильных расстояний от первого и третьего квантиля), не что иное, как выбросы.



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.

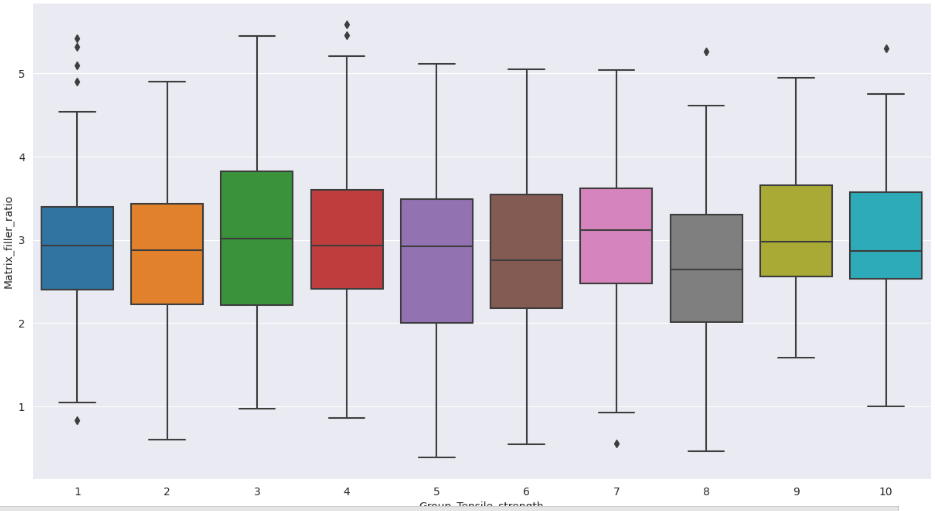
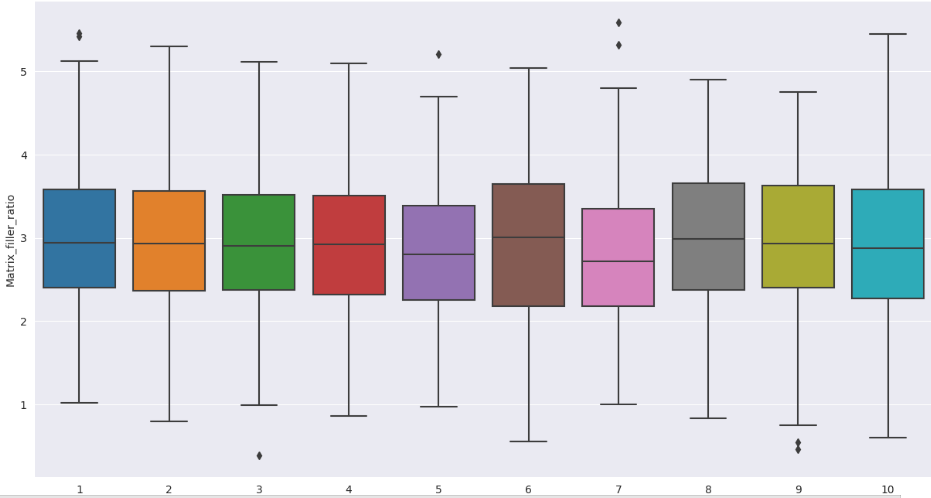
 

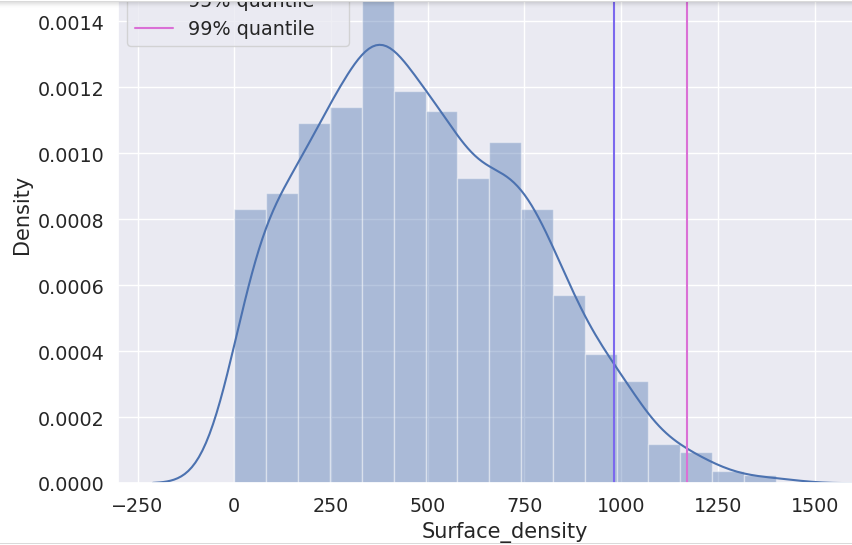
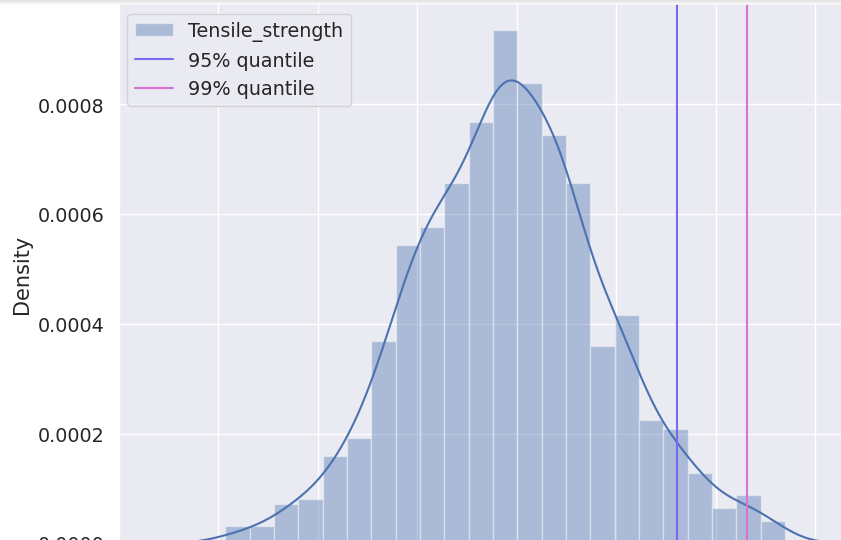
Рисунок 7 Group Tensile strength Рисунок 8 Group Tensile modulus strengt

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.

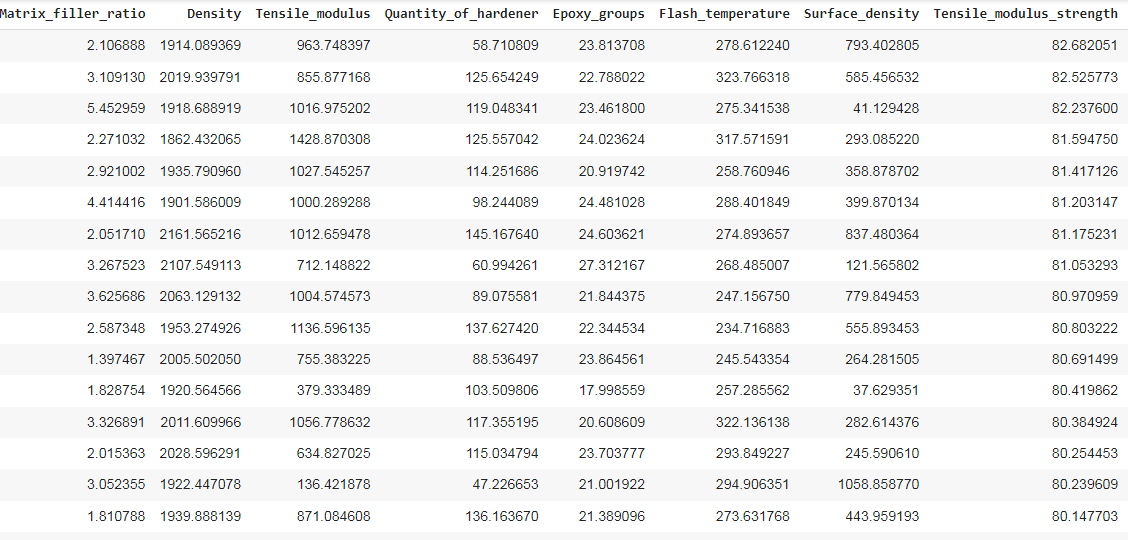
Посмотрим на значения находящиеся в соответствии с рисунком 9 за 95 и 99 процентными квантилями.

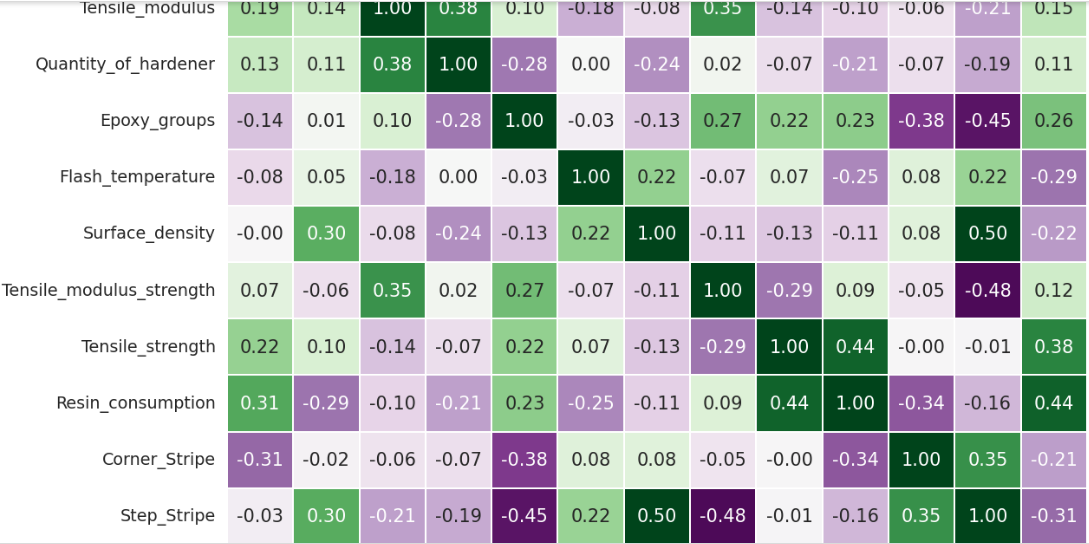


Рисунок—9 Значения, находящиеся за 95 и 99 квантили.

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»

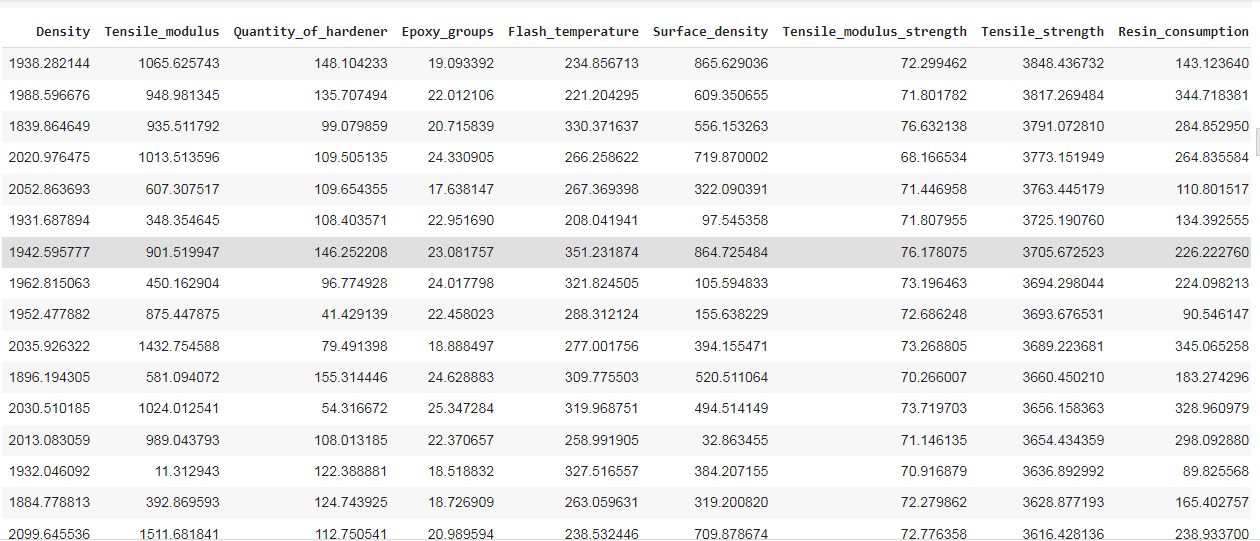


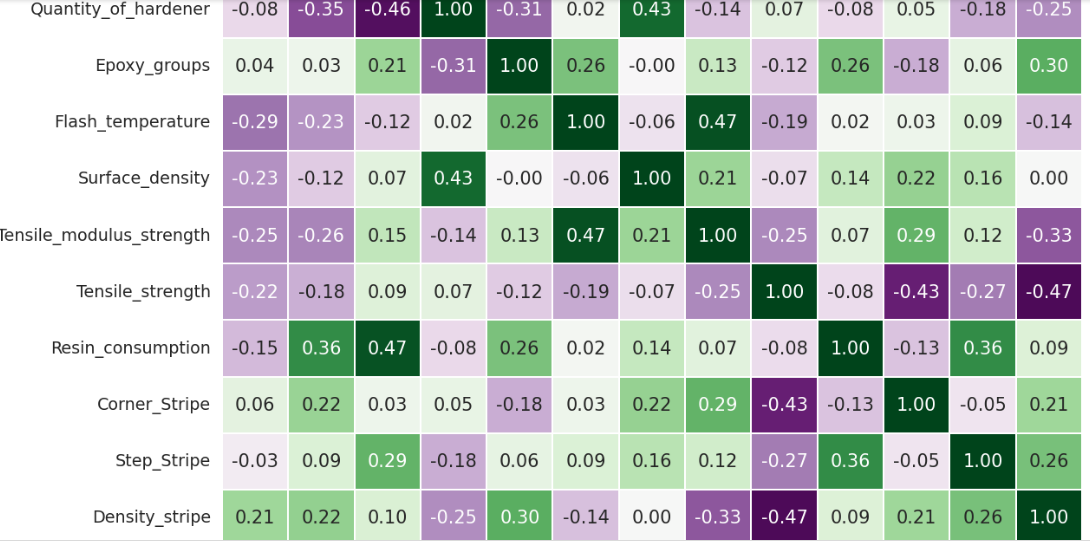


Рисунок—10 Корреляция топ 25 значений «Модуль упругости при растяжении»

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»





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.

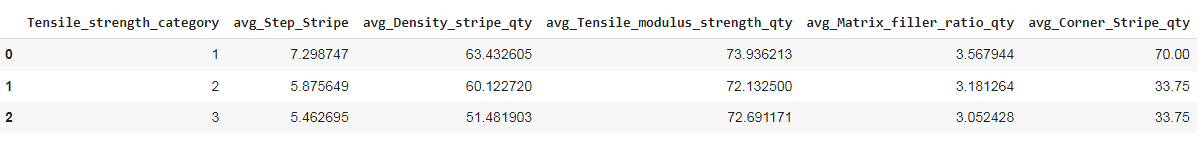


Рисунок 12—Средние значения по группам в топ 25 «Прочность при растяжен, категории»

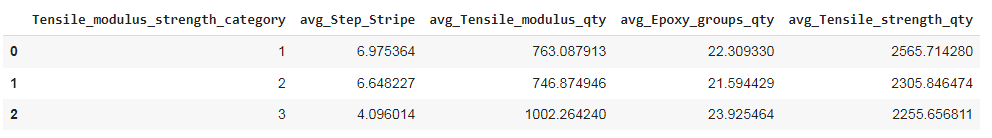


Рисунок 13—Средние значения по группам в топ 25 «Модуль упругости при растяжен, категории»

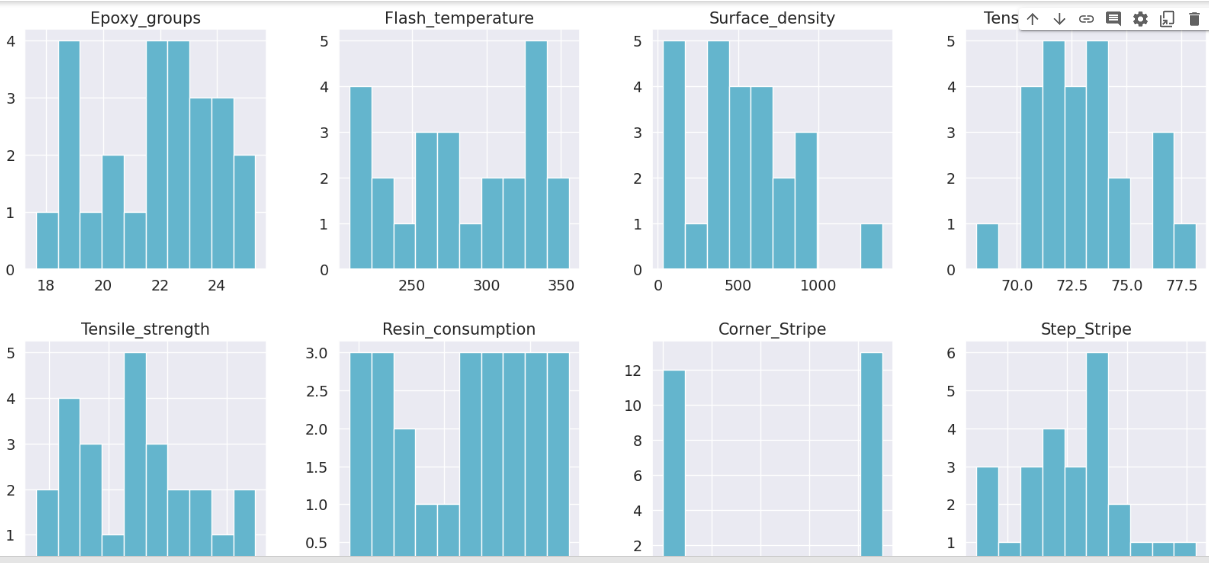


Рисунок 14—Гистограммы топ 25 «Прочность при растяжении»

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.

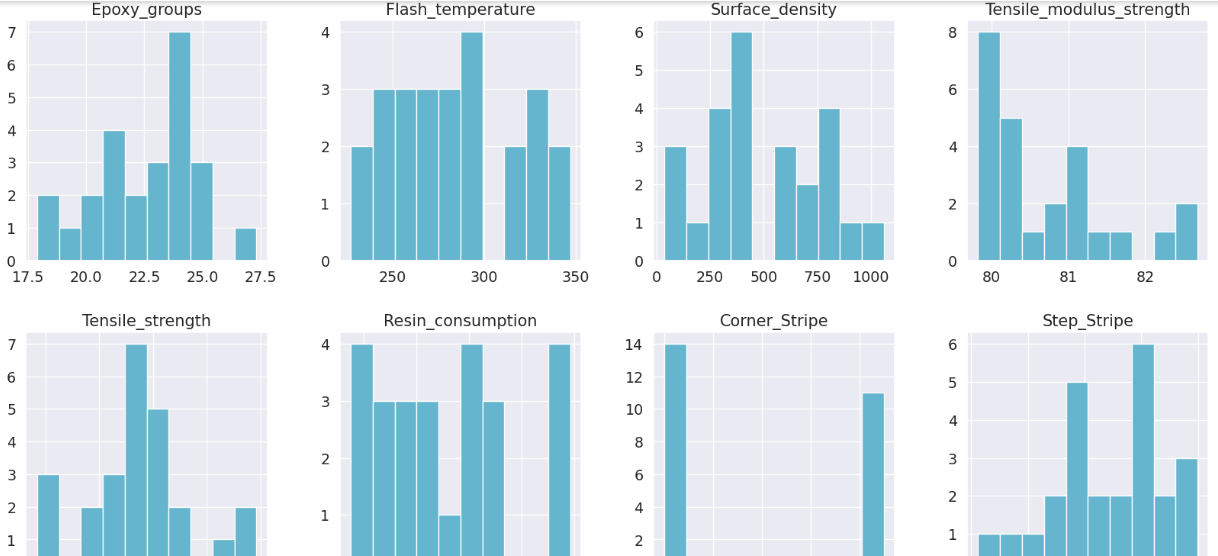


Рисунок15—Гистограммы топ 25 «Модуль упругости при растяжении»

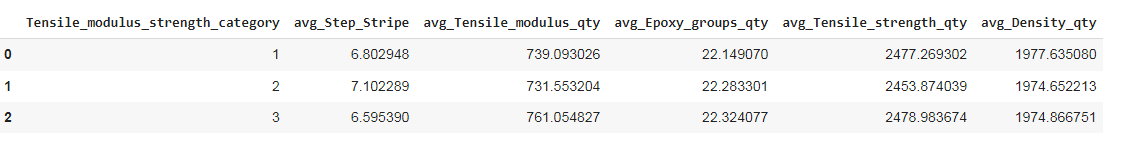
По такому же принципу продолжим работу с полным датасетом. Создадим дополнительный столбец «Модуль упругости при растяжении, категории»: 3категория включает 150 самых больших значений данного признака, 2 и 1 категории создадим как равные половины оставшихся значений.

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

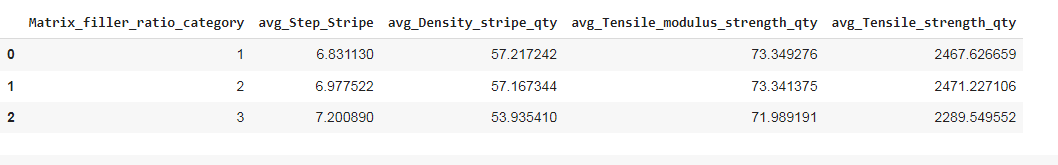
1. категория 3—значения от 4;
2. категория 2—значения от 2.5 до 4;
3. категория 1—значения от 0 до 2.5.

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

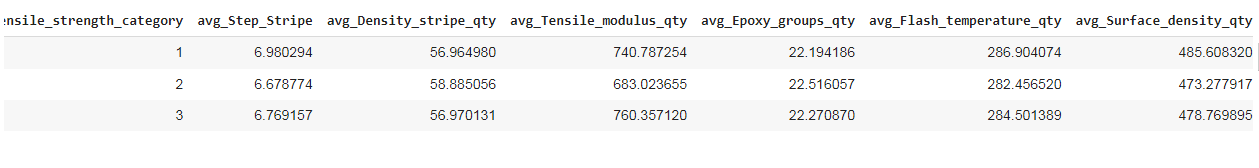
Таблица—9 Средние значен для «Модуль упругости при растяжен, категории»



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



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



**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.

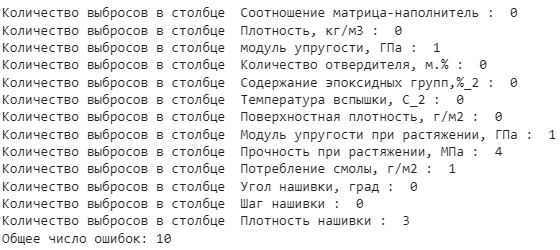
**2.Практическая часть**

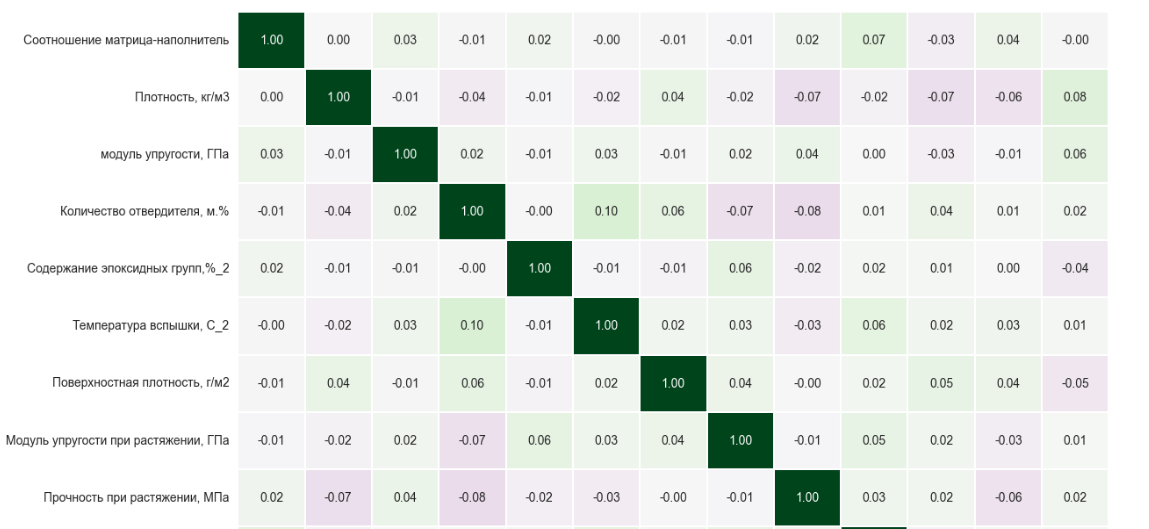
**2.1 Предобработка данных**

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





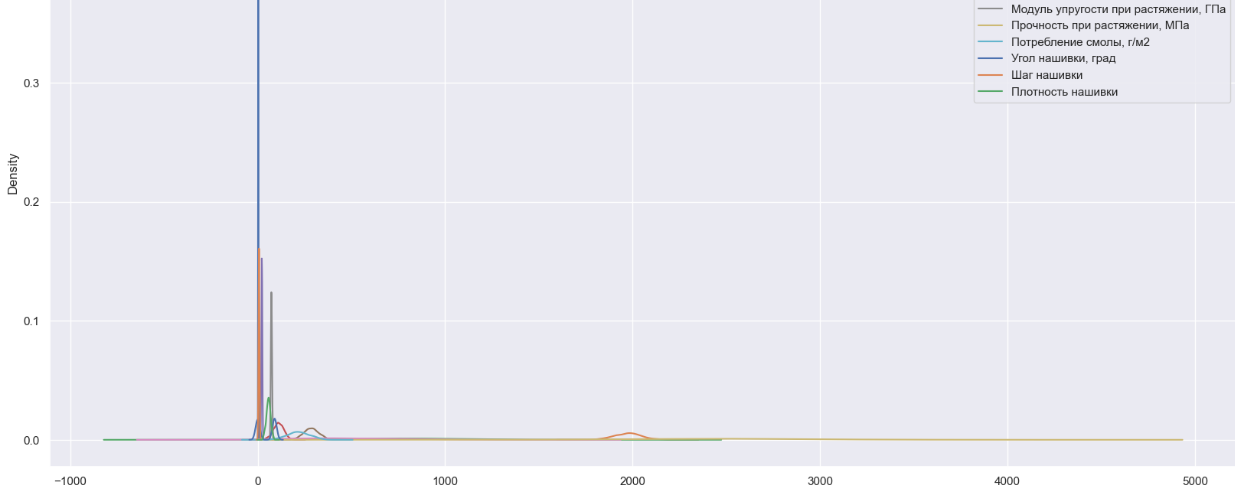
Рисунок—14 Корреляционная тепловая карта Пирсона

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.

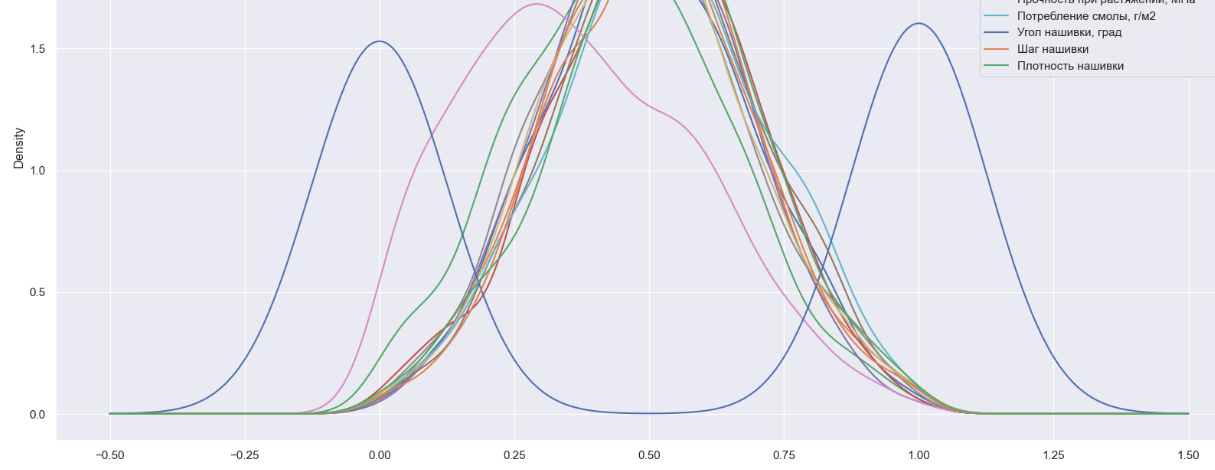


Рисунок—15 Топ 4 по корреляции для «Соотношение матрица-наполнитель»

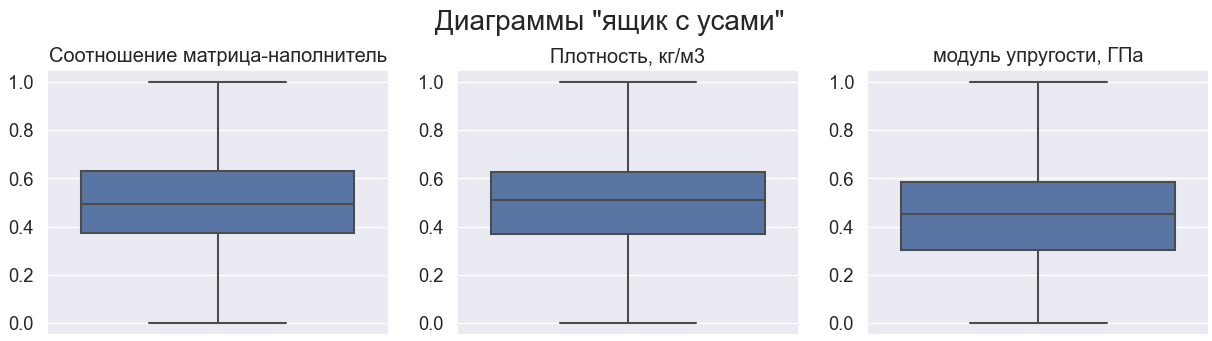


Рисунок—16 Оценка плотности ядра до нормализации

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



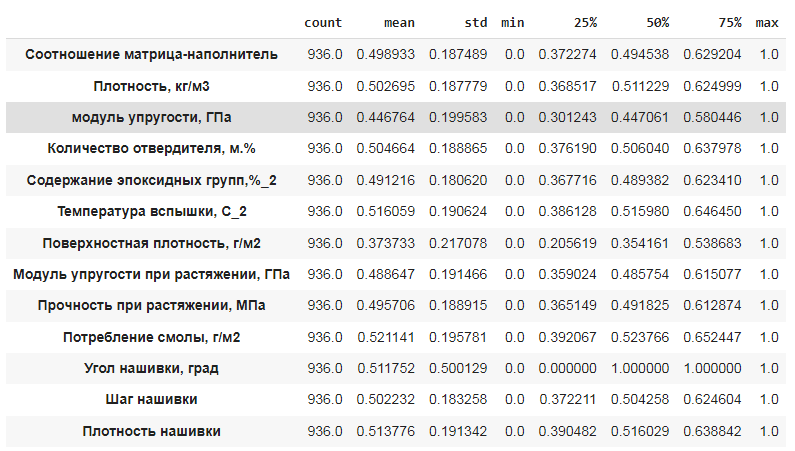
Рисунок—1 «Ящик с усами» после нормализации

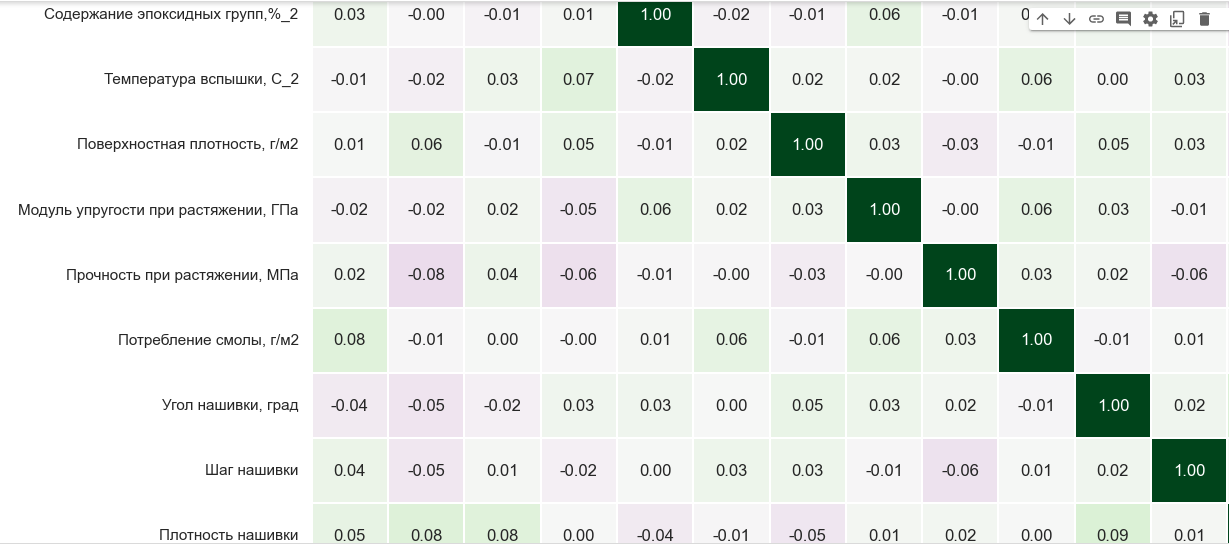
Медианное и среднее значение, в соответствии с таблицей 13, в пределах предыдущих значений после удаления выбросов.

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





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 иR2.

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



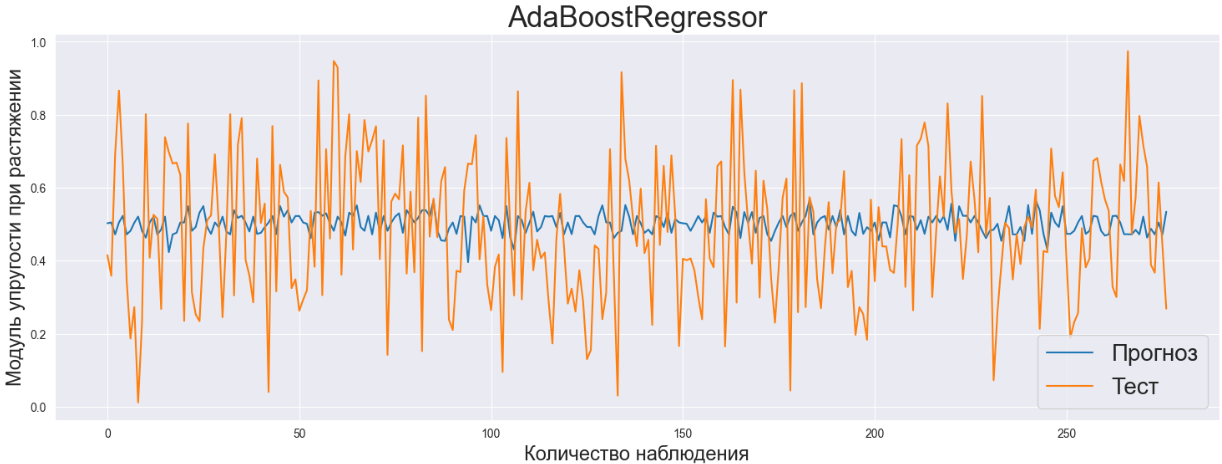


Рисунок 19—Визуализация результатов работы одной из моделей, прогнозирующих «Модуль упругости при растяжении» при стандартных параметрах.

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



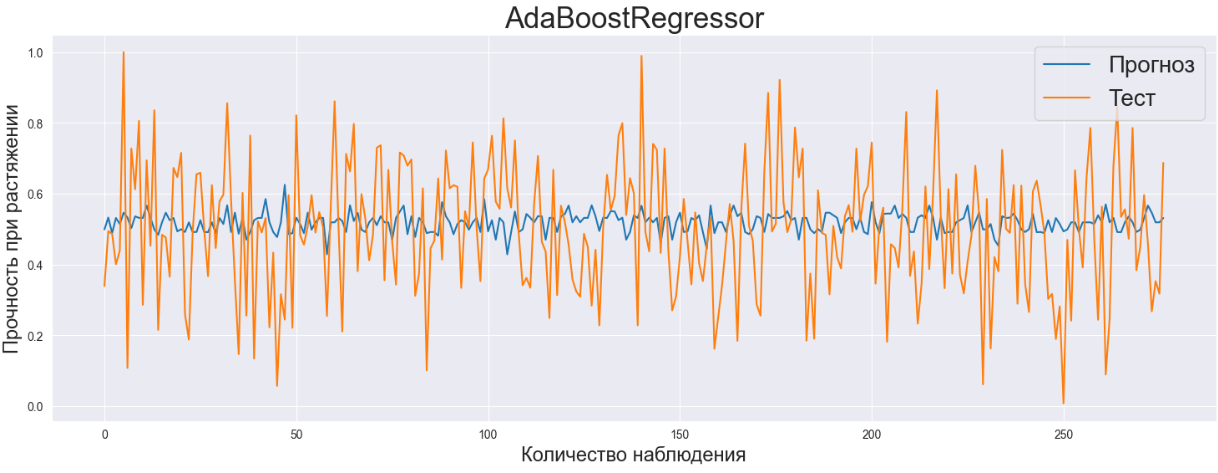


Рисунок 20—Визуализация результатов работы одной из моделей, прогнозирующей «Прочность при растяжении» при стандартных параметрах.

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.

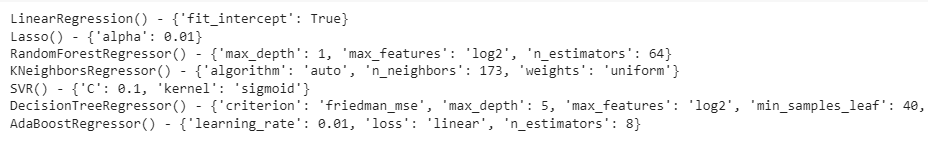
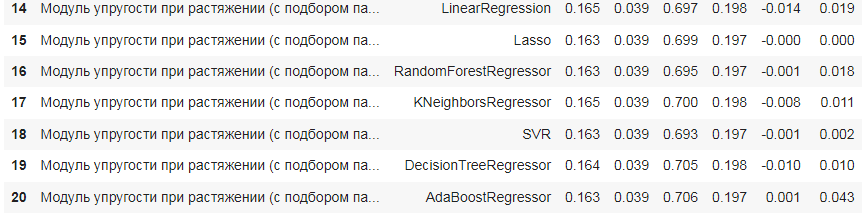


Рисунок 21—Лучшие гиперпараметры работы моделей

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



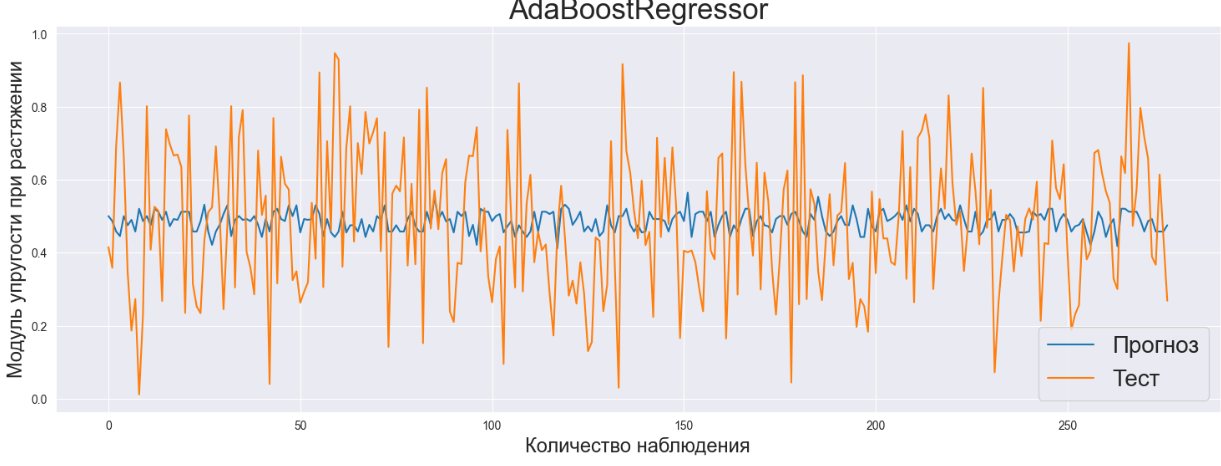
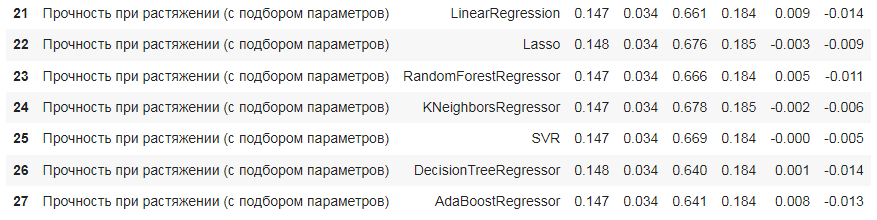


Рисунок 22—Визуализация результатов работы одной из моделей, прогнозирующих «Модуль упругости при растяжении» с лучшими гиперпараметрами

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



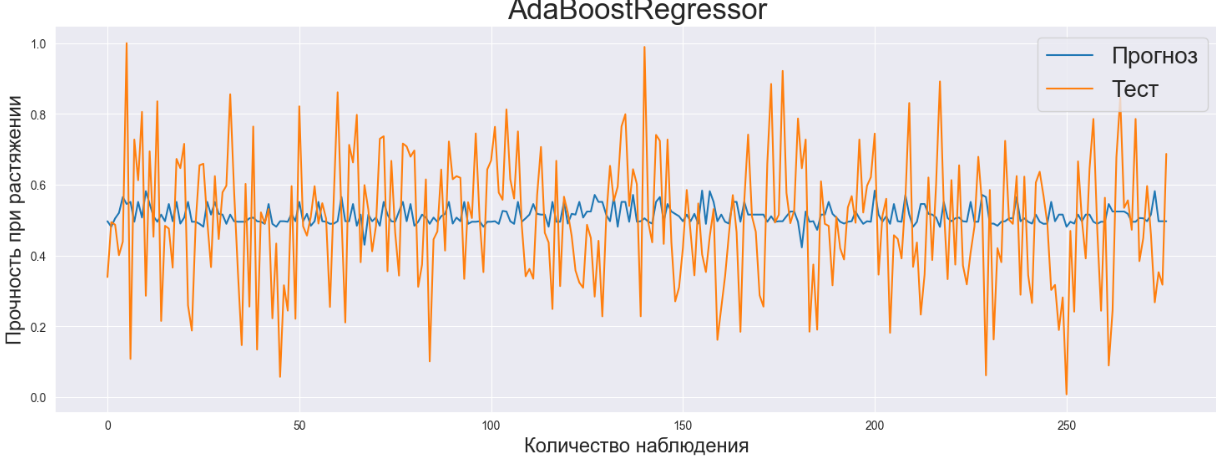
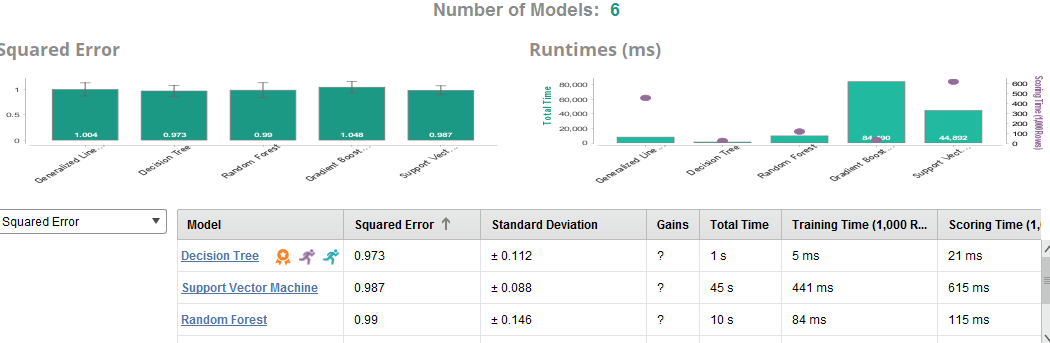
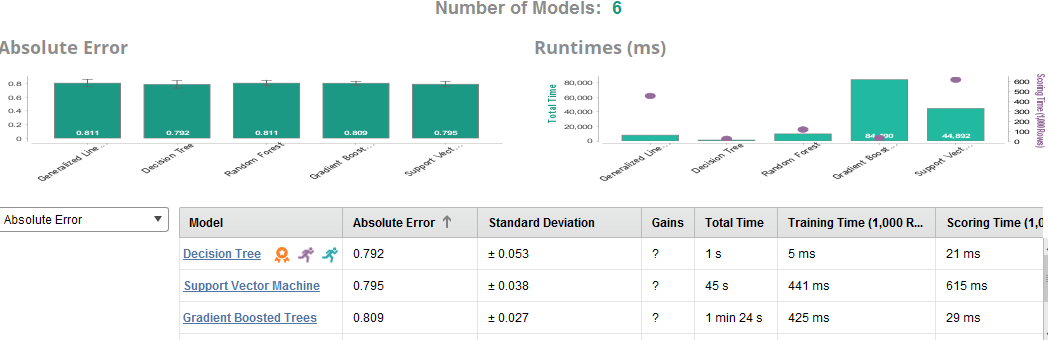


Рисунок 23—Визуализация результатов работы одной из моделей, прогнозирующих «Прочность при растяжении» с лучшими гиперпараметрами

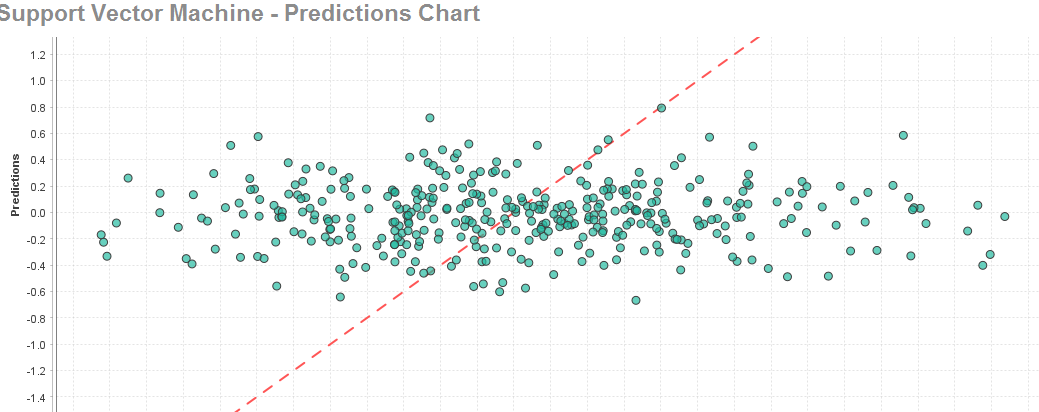
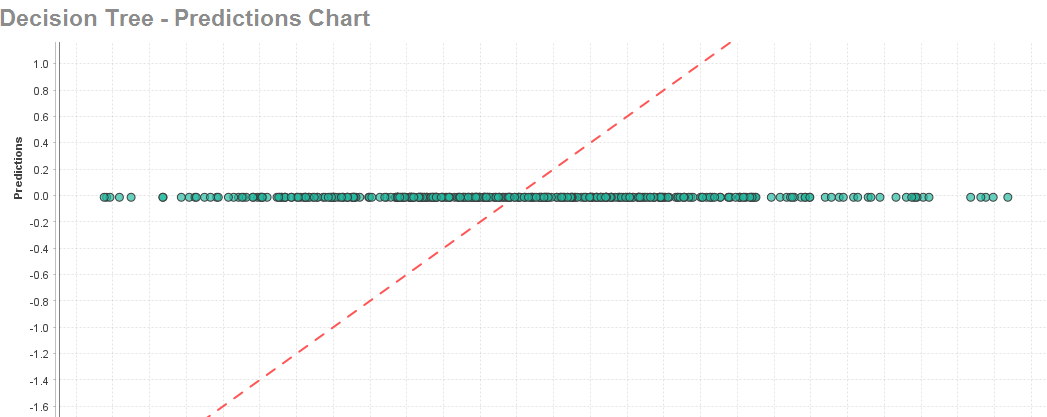
Для сравнения используем построение моделей в программной многопользовательской платформе RapidMiner. Для целевых переменных (для каждой по отдельности) построим модели (Generalised Linear Model, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine) в RapidMiner. Изначально, на 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

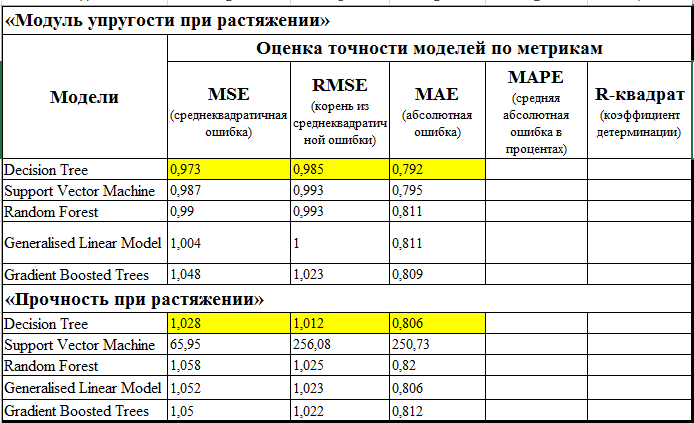


Рисунок—25 Предсказание модели в RapidMiner

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



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.

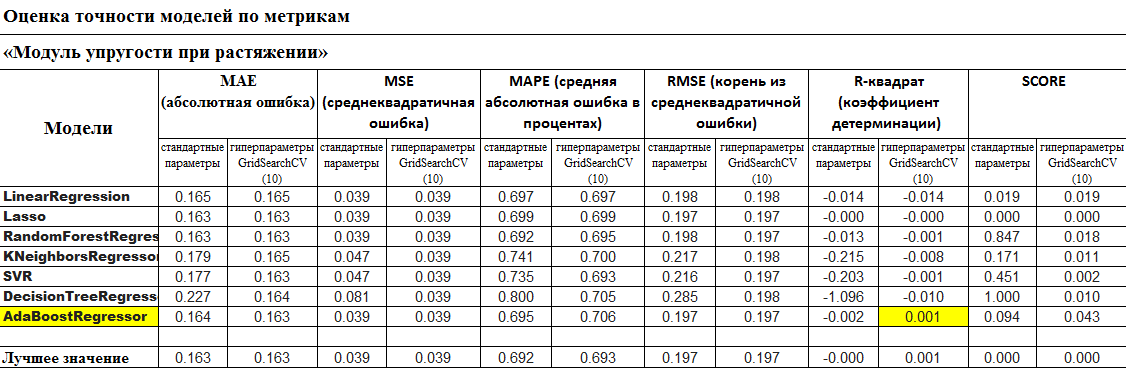
Выполнили также вариант с подачей на вход датасета стандартизированного, но с выбросами. Метрики результата отличались от варианта с полностью обработанным датасетом на 10%.

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

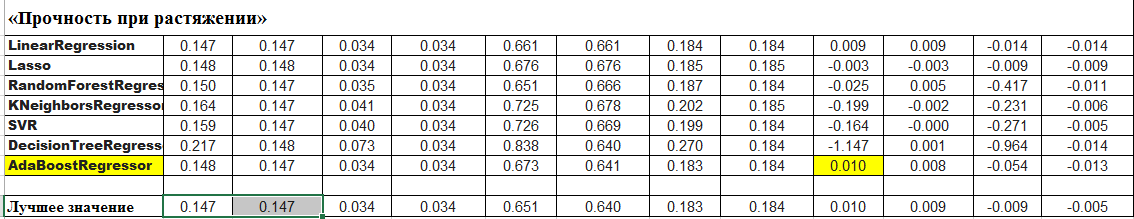
Лучшая модель для прогноза «Модуль упругости при растяжении» по метрике коэффициента детерминации R2 AdaBoostRegressor с лучшими гиперпараметрами.

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



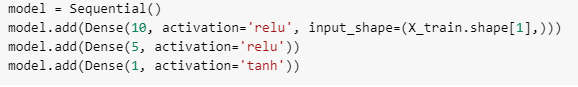




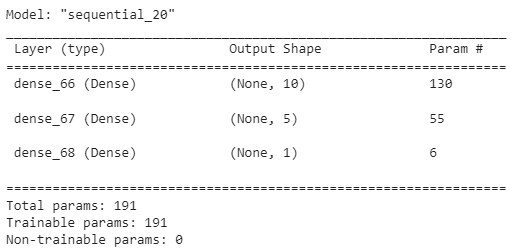
**2.4 Написание нейронной сети, рекомендующей соотношение «матрица-наполнитель»**

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



Рисунок—26 Архитектура модели нейросети



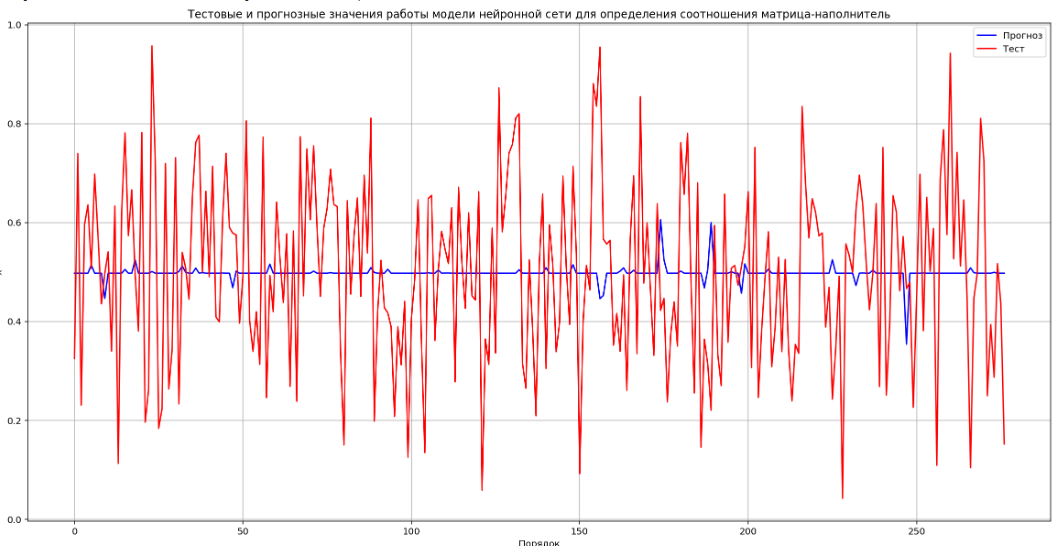
Рисунок—26 Архитектура модели нейросети

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

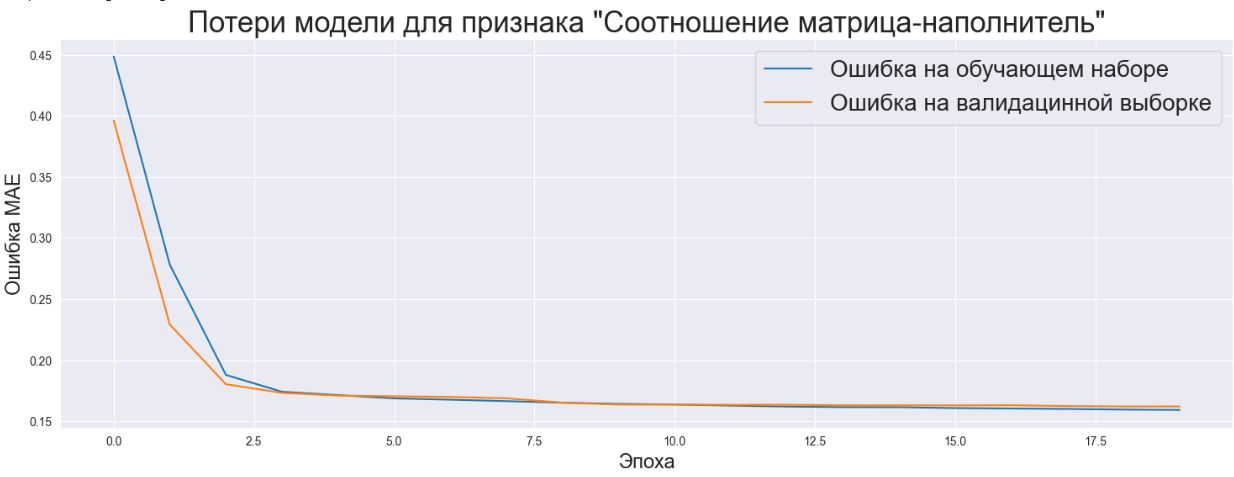


Рисунок—27 Компиляция модели нейросети

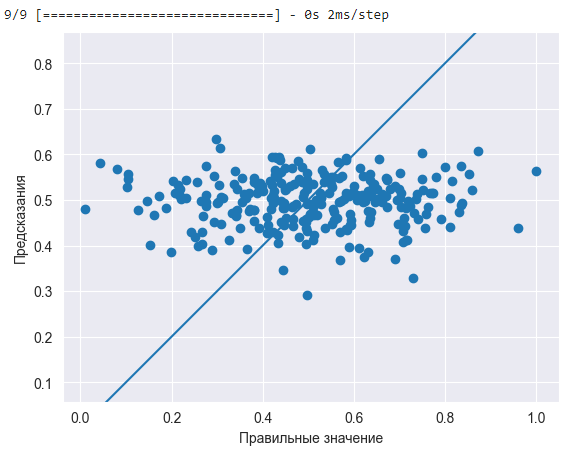
Визуализируем тестовую и прогнозную работу модели нейросети, рисунок 28.



Рисунок—28 Визуализация тестовой и прогнозной работы модели нейросети



Рисунок—29 Графики ошибки модели нейросети

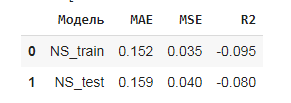


Рисунок—30 Разброс предсказаний модели нейросети

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

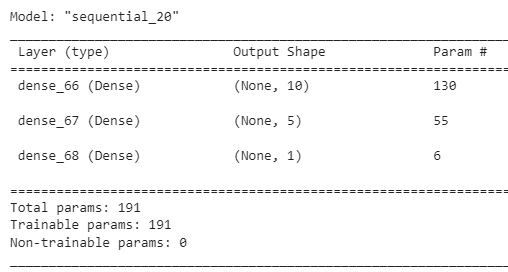


Выполним сохранение модели нейросети, рисунок 31.



Рисунок—31 Сохранение модели нейросети

Проверим загрузку модели нейросети, рисунок 32.



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 Инжинириум 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/ECF749E40C9E58024725804400349189/$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](https://libtime.ru/), 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).