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»

Слушатель Репин Максим Павлович

Москва, 2022

Содержание

[Введение 3](#_Toc106308308)

[1. Аналитическая часть 4](#_Toc106308309)

[1.1. Постановка задачи 4](#_Toc106308310)

[1.2. Описание используемых методов 7](#_Toc106308311)

[1.3. Разведочный анализ данных 12](#_Toc106308312)

[2. Практическая часть 16](#_Toc106308313)

[2.1. Предобработка данных 16](#_Toc106308314)

[2.2. Разработка и обучение модели 21](#_Toc106308315)

[2.3. Тестирование модели 24](#_Toc106308316)

[2.4. Нейронная сеть для рекомендации соотношения матрица-наполнитель 24](#_Toc106308317)

[2.5. Разработка приложения 26](#_Toc106308318)

[Заключение 28](#_Toc106308319)

[Библиографический список 29](#_Toc106308320)

## **Введение**

Пояснительная записка подготовлена в рамках выпускной квалификационной работы по курсу «Data Science» на тему «Прогнозирование конечных свойств новых материалов (композиционных материалов)».

Для решения поставленной задачи изучены теоретические основы предметной области, рассмотрены представленные материалы, проведены исследования и анализ данных с использованием теоретических материалов и практических навыков, полученных на курсе «Data Science».

## **Аналитическая часть**

### **Постановка задачи**

**Описание:**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые отсутствуют у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Например, композитная арматура имеет эксплуатационные характеристики по многим показателям выше металлического аналога. Композит имеет более высокую прочность, чем у металлических стержней.

Композиционный материал состоит из наполнителя, который обеспечивает прочностные и жесткостные характеристики композиционного материала, локализует появившиеся трещины и т.д., и матрицы, которая передает нагрузки между элементами арматуры, предохраняет ее от внешних повреждающих воздействий, обеспечивает монолитность, фиксирует форму и размеры изделий из композиционного материала.

Базальтопластики сочетают в себе высокую прочность, термо- и хемостойкость, экологическую чистоту, долговечность и пожаробезопасность. Применяют их в качестве строительных материалов, арматуры и конструкций, ответственных изделий в машино- и авиастроении и др. В качестве армирующей основы служат базальтовые нити, ровинги, ткани, холсты, маты, а связующей матрицей – органические и неорганические полимеры. Углепластики (карбопластики, углеродопласты) — это композиты, содержащие в качестве наполнителя углеродные волокна. При сочетании углеродных и базальтовых составляющих возможно получить композит с новыми свойствами и характеристиками.

Характеристики получаемого композиционного материала определяют либо по результатам проведения физических испытаний образцов материалов, что является ресурсоемким процессом, либо прогнозированием характеристик посредством симуляции представительного элемента объема композита на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

Поставленная задача: На основе данных о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.) необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов.

Актуальность:Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

На входе имеются данные о начальных свойствах и характеристиках исходных композиционных материалов: датасет с информацией о десяти параметрах одного композиционного материала (базальтопластик) и датасет с информацией о трех параметрах другого композиционного материала (нашивки углепластика).

Для решения задач два датасета были объединены в один датасет по индексу с типом объединения INNER и приведены к одной длине в 1023 наблюдения с 13 переменными, из которых десять переменных являются входными:

1. угол нашивки, град;
2. шаг нашивки;
3. плотность нашивки;
4. плотность, кг/м3;
5. модуль упругости, Гпа;
6. количество отвердителя, м;
7. содержание эпоксидных групп,%\_2;
8. температура вспышки, С\_2;
9. поверхностная плотность, г/м 2;
10. потребление смолы, г/м2.

Три переменных объединенного датасета являются выходными:

1. соотношение матрица-наполнитель;
2. модуль упругости при растяжении,Гпа;
3. прочность при растяжении, Мпа.

При проведении анализа объединенного датафрейма было установлено, что наблюдения не имеют пропусков.

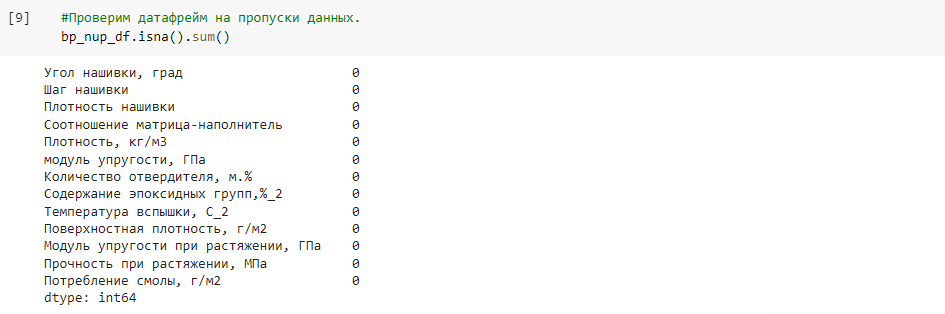


Рисунок 1 - Результаты проверки датафрейма на пропуски данных

Все переменные относятся к количественным, поэтому для решения поставленной задачи предполагается применение регрессионных моделей, поэтому все входные переменные рассматриваются в качестве независимых, регрессоров, а выходные в качестве зависимых переменных.

Для прогнозирования значений выходных переменных «модуль упругости при растяжении» и «прочность при растяжении» требуется провести предобработку данных, разведочный анализ, построить и сравнить модели прогноза, а для выходной переменной «соотношение матрица-наполнитель» создать рекомендательную систему (на основе данных об остальных входных параметрах датасета рассчитать с помощью нейронной сети абсолютную величину переменной).

### **Описание используемых методов**

1.2.1 Линейная регрессия (Linear regression) — один из фундаментальных методов статистического и машинного обучения, применяется как модель зависимости переменной x от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости. Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных. Модель линейной регрессии является часто используемой и наиболее изученной в эконометрике.

Линейная регрессия – хорошо контролируемый алгоритм обучения, который применяется, когда известно, что связь между ковариатами и переменной отклика линейна.

Основное преимущество метода заключается в том, что наиболее подходящей является линия с минимальной ошибкой во всех отношениях, он имеет высокую эффективность, но иногда эта высокая эффективность создает недостаток, который склонен к переобучению данных (т.е. некоторые зашумленные данные, которые также считаются полезными данными), а также его нельзя использовать, когда взаимосвязь между зависимой и независимой переменной нелинейная.

Недостатки линейной регрессии: упрощает многие проблемы реального мира, чаще всего ковариаты и переменные ответа не демонстрируют линейной зависимости, очень чувствительна к аномалиям в данных (или выбросам), если есть несколько параметров, а не количество доступных выборок, то модель начинает моделировать шум, а не взаимосвязь между переменными.

Метод наименьших квадратов - наиболее простой метод определения коэффициентов регрессии при выполнении регрессионного анализа.

Наиболее типичными мерами качества в задачах регрессии являются:

1. Средняя квадратичная ошибка (англ. Mean Squared Error, MSE).

MSE применяется в ситуациях, когда надо подчеркнуть большие ошибки и выбрать модель, которая дает меньше больших ошибок прогноза. Грубые ошибки становятся заметнее за счет того, что ошибку прогноза мы возводим в квадрат. Модель, которая дает нам меньшее значение среднеквадратической ошибки, имеет меньше грубых ошибок.

MSE подходит для сравнения двух моделей или для контроля качества во время обучения, но не позволяет сделать выводов о том, на сколько хорошо данная модель решает задачу. Например, MSE = 10 является очень плохим показателем, если целевая переменная принимает значения от 0 до 1, и очень хорошим, если целевая переменная лежит в интервале от 10000 до 100000. В таких ситуациях вместо среднеквадратичной ошибки полезно использовать коэффициент детерминации — R2.

1. Коэффициент детерминации (R2) измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Фактически, данная мера качества — это нормированная среднеквадратичная ошибка. Значение R2 принимает значение от 0 до 1 и показывает долю объяснённой дисперсии объясняемого ряда. Чем ближе R2 к 1, тем лучше модель, тем меньше доля необъяснённого.

Возможные проблемы: значение R2 не уменьшается при добавлении в уравнение факторов, сколь плохи бы они ни были. Коэффициент гарантированно будет равен 1, если мы добавим в модель столько факторов, сколько у нас наблюдений. Поэтому сравнивать модели с разным количеством факторов, используя R2, не имеет смысла.

1. Средняя абсолютная ошибка (англ. Mean Absolute Error, MAE)

Среднеквадратичный функционал сильнее штрафует за большие отклонения по сравнению со среднеабсолютным, и поэтому более чувствителен к выбросам. При использовании любого из этих двух функционалов может быть полезно проанализировать, какие объекты вносят наибольший вклад в общую ошибку — не исключено, что на этих объектах была допущена ошибка при вычислении признаков или целевой величины.

1. Средняя абсолютная процентная ошибка (англ. Mean Absolute Percentage Error, MAPE). Это коэффициент, не имеющий размерности, с очень простой интерпретацией. Его можно измерять в долях или процентах. Если у вас получилось, например, что MAPE=11.4%, то это говорит о том, что ошибка составила 11,4% от фактических значений. Основная проблема данной ошибки — нестабильность.

1.2.2 Случайный лес (Random forest) - алгоритм машинного обучения, предложенный Лео Брейманом и Адель Катлер, заключающийся в использовании комитета (ансамбля) решающих деревьев. Алгоритм применяется для задач классификации, регрессии и кластеризации. Основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим.

Алгоритм состоит из четырех этапов:

1. создать случайные выборки из заданного набора данных;
2. для каждой выборки построить дерево решений и получить результат предсказания, используя данное дерево;
3. провести голосование за каждый полученный прогноз;
4. выбрать предсказание с наибольшим количеством голосов в качестве окончательного результата.

Преимущества метода:

1. способность эффективно обрабатывать данные с большим числом признаков и классов;
2. нечувствительность к масштабированию (и вообще к любым монотонным преобразованиям) значений признаков;
3. одинаково хорошо обрабатываются как непрерывные, так и дискретные признаки. Существуют методы построения деревьев по данным с пропущенными значениями признаков;
4. существуют методы оценивания значимости отдельных признаков в модели;
5. внутренняя оценка способности модели к обобщению (тест по неотобранным образцам);

Недостатки: из-за большого размера получающихся моделей требуется большой объем памяти для хранения модели.

1.2.3 Нейро́нная сеть (также искусственная нейронная сеть) — математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма.

Искусственная нейронная сеть представляет собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Такие процессоры обычно довольно просты (особенно в сравнении с процессорами, используемыми в персональных компьютерах). Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам.

Нейронные сети не программируются в привычном смысле этого слова, они обучаются. Возможность обучения — одно из главных преимуществ нейронных сетей перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке, а также неполных и/или «зашумленных», частично искажённых данных.

Преимущества искусственной нейронной сети:

1. высокая точность классификации;
2. сильная способность параллельной распределенной обработки, сильная распределенная память и способность к обучению;
3. обладает высокой устойчивостью к шумовым нервам и может полностью аппроксимировать сложные нелинейные отношения;

Недостатки искусственных нейронных сетей:

1. нейронные сети требуют большого количества параметров, таких как топология сети, начальные значения весов и пороговых значений;
2. невозможно наблюдать процесс обучения между ними, а полученные результаты трудно объяснить, что повлияет на достоверность и приемлемость результатов;
3. время обучения слишком велико и может даже не достичь цели обучения.

Для создания рекомендательной системы «Соотношение матрица – наполнитель» была использована нейронная сеть архитектуры многослойный персептрон. Особенностью архитектуры является наличие более чем одного обучаемого слоя. Существует предположение, что, используя большее число слоёв, можно уменьшить число элементов в них, то есть суммарное число элементов в слоях будет меньше, чем если использовать один скрытый слой. Это предположение успешно используется в технологиях глубокого обучения и имеет обоснование.

Для решения задачи использовалась открытая программная библиотека для машинного обучения TensorFlow, разработанная компанией Google для решения задач построения и тренировки нейронной сети с целью автоматического нахождения и классификации образов.

### **Разведочный анализ данных**

Для разведочного анализа данных использовались методы описательной статистики и графические методы – гистограммы плотности распределения, «ящиковые» диаграммы, графики рассеяния, а также аналитические: методы определения выбросов наблюдений (метод трех сигм и метод межквартальных интервалов), корреляционный анализ (для установления статистических зависимостей между переменными и другие).

1.3.1 Описательная статистика – это краткие описательные коэффициенты, которые суммируют данный набор данных, который может быть либо представлением всей совокупности, либо выборкой совокупности. Описательная статистика подразделяется на меры центральной тенденции и меры изменчивости (распространения). Меры центральной тенденции включают среднее, медиану и моду, в то время как меры изменчивости включают стандартное отклонение, дисперсию, минимальные и максимальные переменные, эксцесс и асимметрию.

Методы описательной статистики:

* bp\_nup\_df.info() – выведена общая информация о датафрейме: тип данных, количество строк и столбцов, наименование переменных;
* sorted(bp\_nup\_df.[column].unique() - определена уникальность значений в датафрейме;
* bp\_nup\_df.isna().sum() – проверка данных на наличие пропусков. Пропусков не обнаружено;
* bp\_nup\_df.describe() – получены основные статистические характеристики датафрейма, в том числе средние и медианные значения характеристик, стандартное отклонение, минимальное и максимальное значения, 1-й перцентиль, 2-й перцентиль, 3-й перцентиль, количество переменных.

1.3.2 Гистограмма плотности распределения — это столбиковая диаграмма, которая показывает, как данные распределяются по группам значений. Собранные данные представляют в виде ряда прямоугольников, одинаковых по ширине и различающихся по высоте. Анализ характера изменения высот позволяет оценить динамику процесса.

Гистограмму плотности распределения используют, чтобы наглядно показать, в каком интервале располагаются наиболее часто встречающиеся значения и как вообще распределяются данные. Кроме того, эта гистограмма позволяет определить наилучшие результаты процесса. Такое графическое изображение динамики процесса дает возможность наметить приоритетные задачи по его улучшению.

* + 1. «Ящик с усами» или диаграмма размаха — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.
    2. Попарные диаграммы рассеяния и тепловая карта корреляций

Корреляционный анализ (correlation analysis) – статистический метод изучения взаимосвязи между двумя и более случайными величинами. В качестве случайных величин в эмпирических исследованиях выступают значения переменных, измеряемые свойства исследуемых объектов наблюдения. Суть корреляционного анализа заключается в расчете коэффициентов корреляции. Коэффициент корреляции может принимать, как правило, положительные и отрицательные значения и колеблется от -1 до +1. Знак коэффициента корреляции позволяет интерпретировать направление связи, а абсолютное значение – силу связи. Нулевая корреляция означает, что две переменные не зависят друг от друга. Положительная корреляция указывает на то, что переменные движутся в одном направлении, а отрицательная корреляция указывает на противоположное

Способ расчета коэффициентов корреляции зависит от шкал измерения переменных, между которыми исследуется взаимосвязь. Для переменных, измеряемых в количественной шкале (интервальной шкале или шкале отношений), рассчитывают ковариацию или корреляционный момент, а на его основе линейный коэффициент корреляции (коэффициент корреляции Пирсона).

Диаграмма рассеяния позволяет оценить зависимости между парами соответствующих элементов. Этим методом можно без математической обработки данных установить зависимости по графическому представлению. Тепловая карта используется для графического представления значений матрицы с разными цветовыми оттенками для разных значений. Он очень четко визуализирует общую матрицу.

* + 1. Правильно 3-х сигм (поиск выбросов)

Стандартное или среднеквадратичное отклонение — это наиболее частый показатель рассеивания значений величины относительно математического ожидания. Обозначается символом σ, который произносится как «сигма».

Правило трех сигм заключается в том, что при нормальном распределении практически все значения величины с вероятностью 0,9973 лежат не далее трех сигм в любую сторону от математического ожидания, то есть находятся в диапазоне [μ−3σ;μ+3σ]. Математическое ожидание — это среднее значение случайной величины. Обозначается как μ.

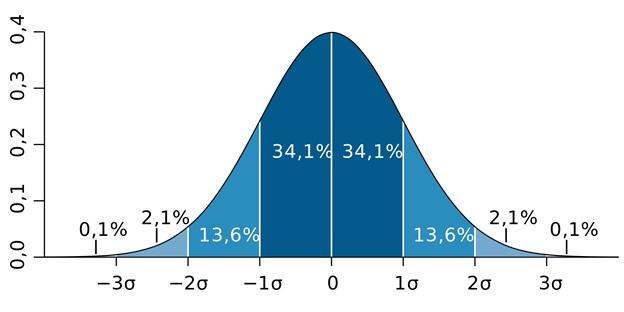


Рисунок 2 – График нормального распределения

Приблизительно 99,7% всех значений лежат в пределе трех сигм от математического ожидания, около 95% — в пределах двух сигм, а примерно 68% значений лежат в пределах всего одной сигмы.

Те значения, которые выходят за рамки 3 сигм, принято считать грубыми ошибками. Большое количество таких ошибок может свидетельствовать о том, что распределение на самом деле не является нормальным. В этом заключается практическая польза правила 3 сигм.

## **Практическая часть**

### **Предобработка данных**

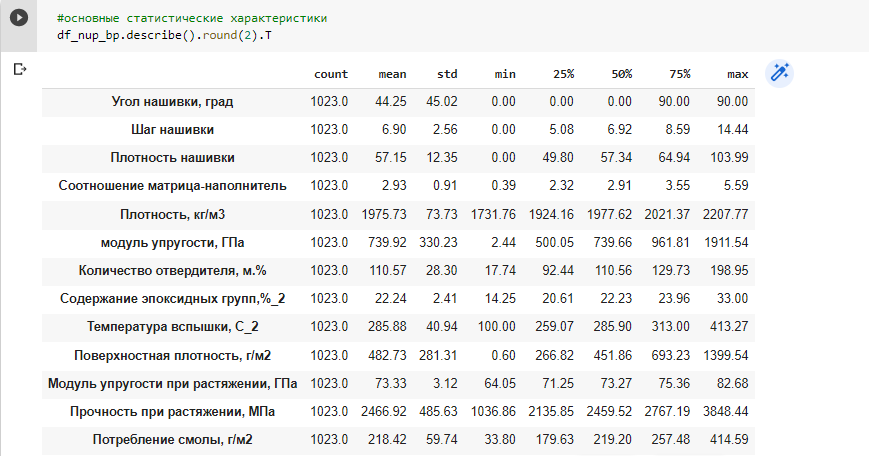
Описательные статистики переменных объединенного датасета см. Рисунок 4

Рисунок 3 – Основные статические характеристики

Для разведочного анализа построены гистограммы плотности распределения значений каждой переменной данных и установлено, что для большинства характеристик распределение близко к нормальному. Исключение составляют:

- поверхностная плотность, г/м2 - нормальное распределение со смещением вправо. Что свидетельствует о большем количестве измерений с меньшим показателем поверхностной плотности.

- угол нашивки, град - дискретное распределение, так как колонка содержит всего два уникальных значения.

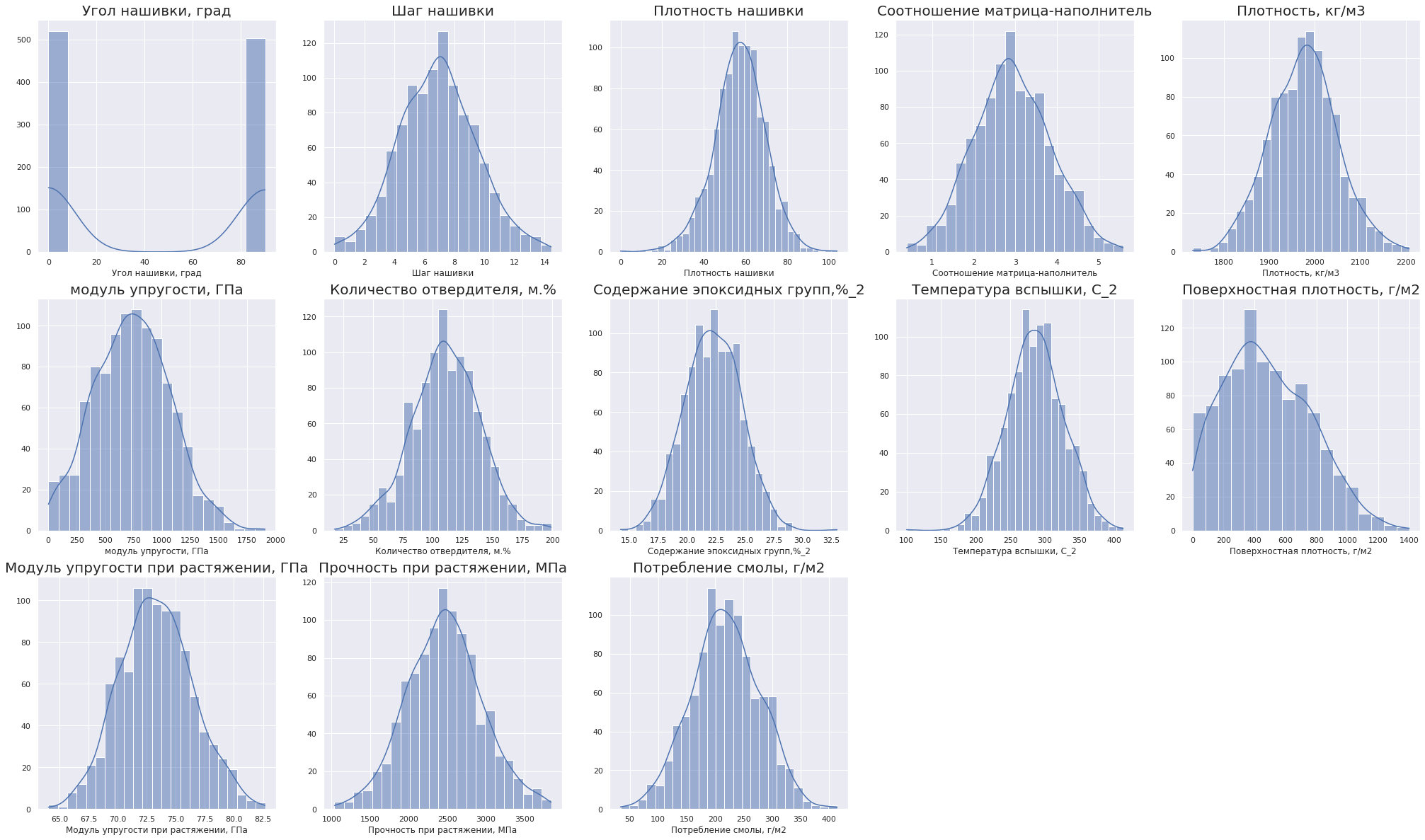


Рисунок 4 – Гистограммы плотности распределения переменных

Для поиска выбросов построены диаграммы Ящика с усами, которые позволяют разбить данные на квантили 25, 50,75, 100 и воспользоваться статистическим методом трех сигм.

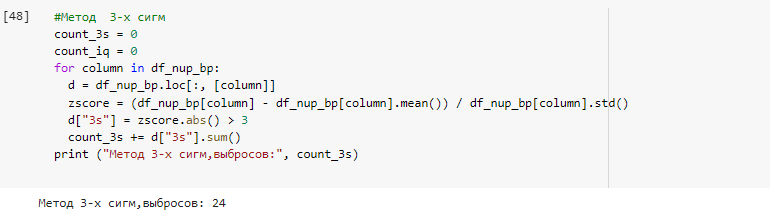


Рисунок 5 – Поиск выбросов методов 3-х сигм

Отклонение значения нормально распределённой случайной величины X от её математического ожидания М(х) не превосходит утроенного среднеквадратического отклонения σ с вероятностью около 0,9973. Иначе говоря, с вероятностью 0,9973 значение нормально распределённой случайной величины Х находится в интервале [М(х) - 3σ ... М(х) + 3σ].

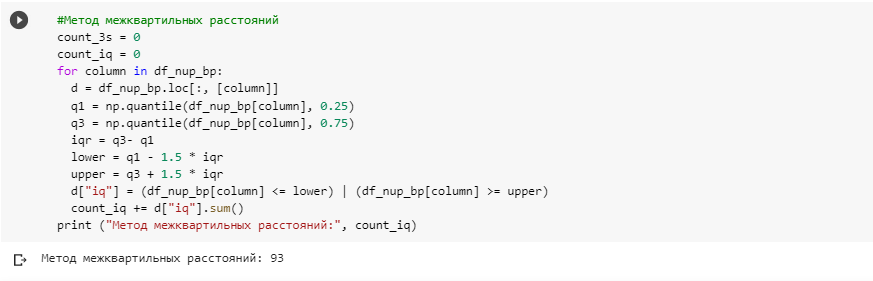


Рисунок 6 – Поиск выбросов методом межквартальных расстояний

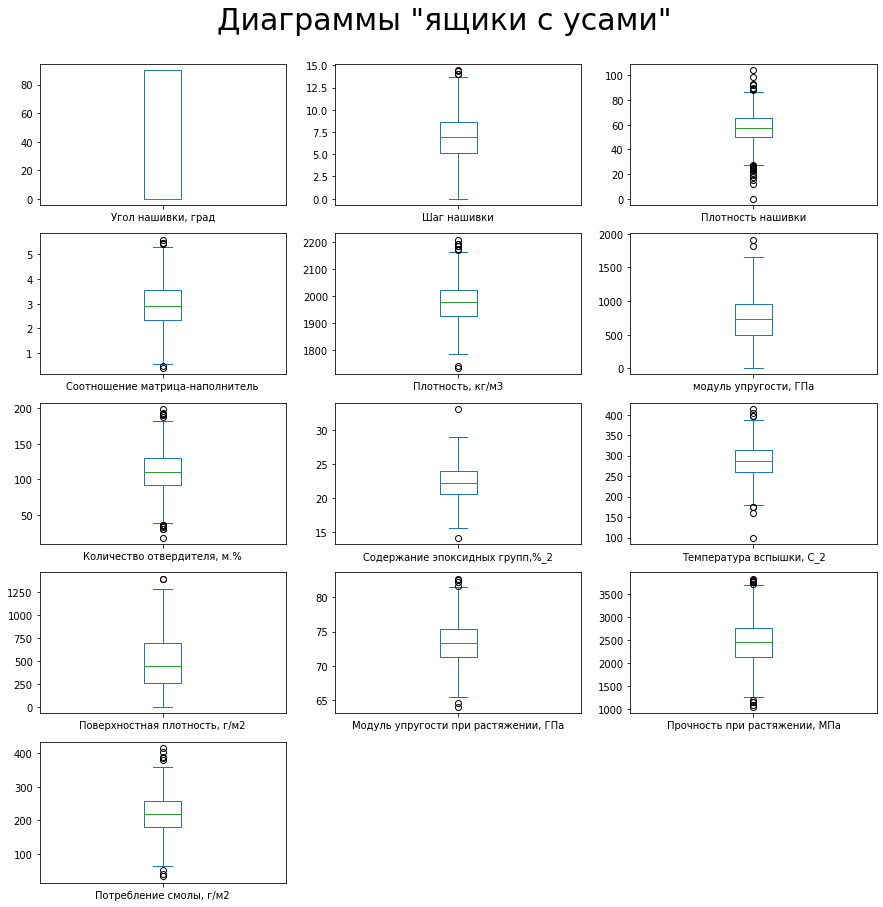


Рисунок 7 – Диаграммы «Ящик с усами»

По диаграммам «Ящик с усами» фиксируем выбросы по всем переменным датафрейма, кроме переменной «Угол нашивки», для которой эта диаграмма неинформативна.

Для дельнейшего исследования данных был рассчитан коэффициент корреляции для каждой пары входных параметров. Для этого были построены попарные графики точек рассеяния, а также тепловая карта коэффициентов корреляции.

Попарные графики рассеяния показываю очень слабовыраженную регрессионную зависимость между переменными. Возможно переменные связывает нелинейная зависимость, что тоже является регрессией.

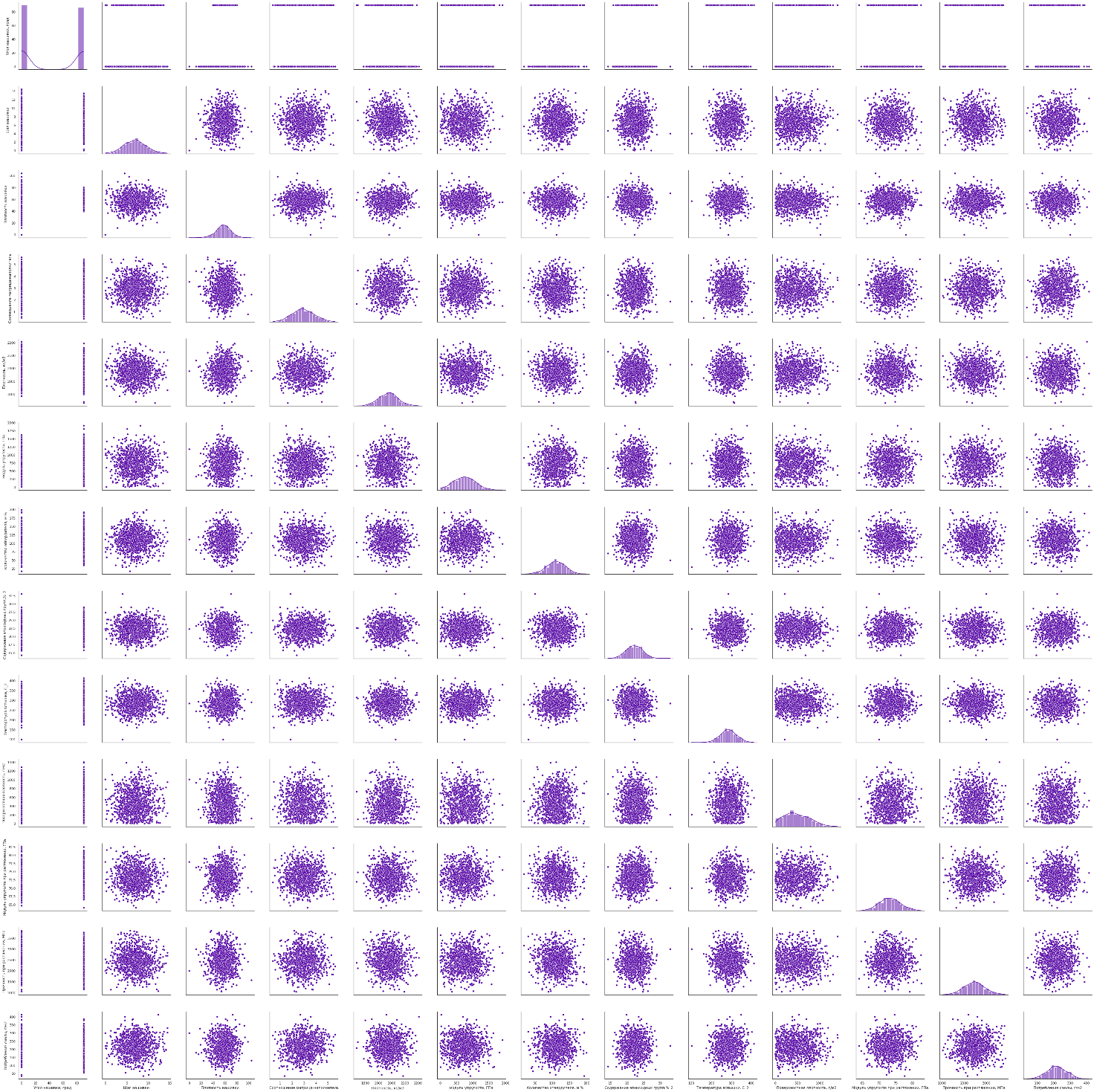


Рисунок 8 – Попарные графики рассеяния

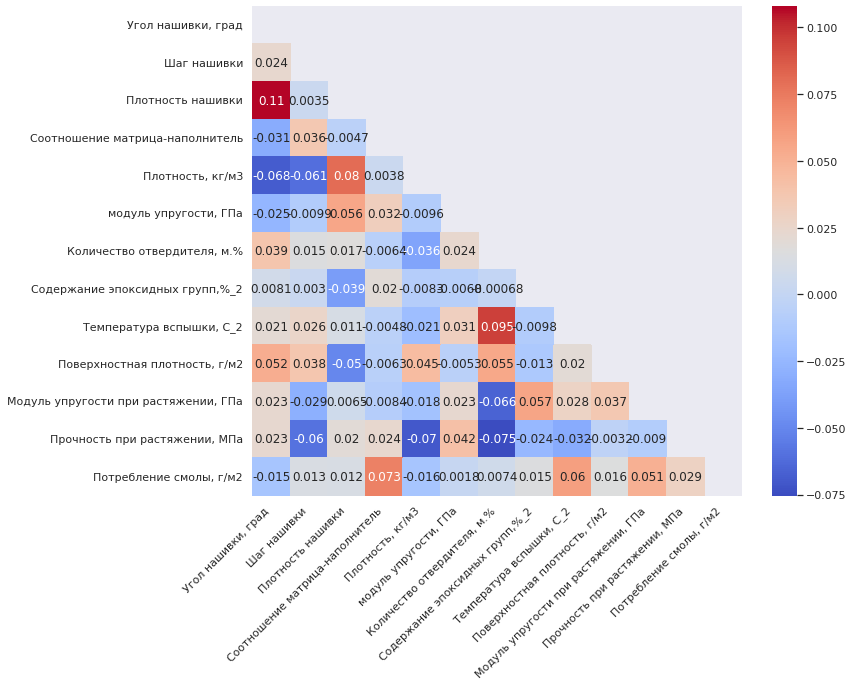


Рисунок 9 – Тепловая карта коэффициентов корреляции

Построенная тепловая карта отображает наличие слабых связей между переменными в датафрейме. Максимальная корреляция между «Плотностью нашивки» и углом нашивки и составляет 0.11, что говорит об отсутствии зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, что говорит об отсутствии корреляционных связей между переменными.

В связи с незначительным количеством выбросов их возможно удалить из датафрейма.

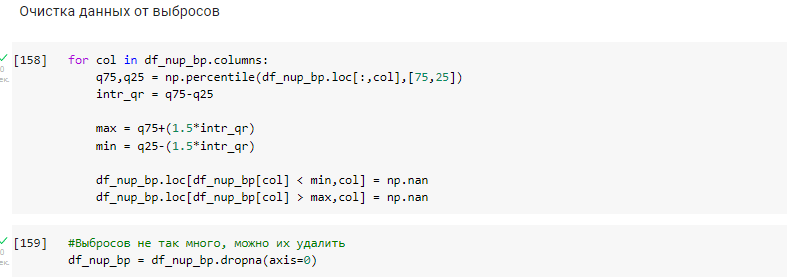


Рисунок 10 – Очистка данных от выбросов

Выполним нормализацию данных (приведение к диапазону от 0 до 1) с помощью метода MinMaxScaler.

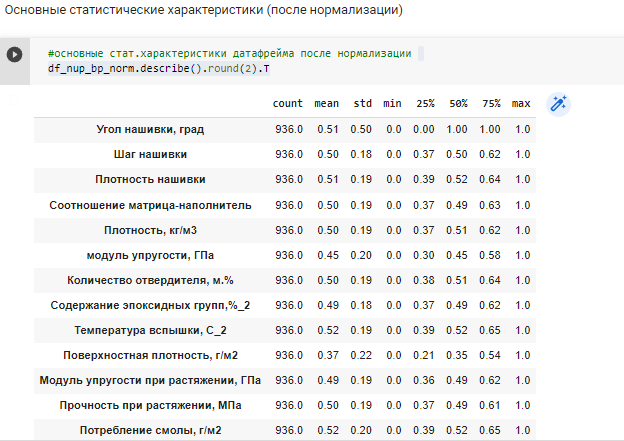


Рисунок 11 – Основные статистические показатели нормализованного

датафрейма

### **Разработка и обучение модели**

Для прогноза прочности при растяжении и модуля упругости при растяжении использовались следующие методы: случайный лес, линейная регрессия. Предварительно датафрейм был разбит тестовую (30%) и обучающую выборки с выделением предикторов и целевой переменной.

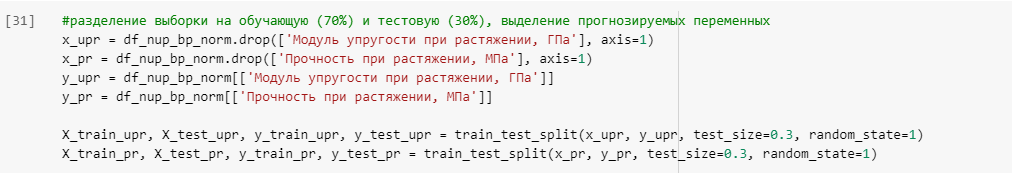


Рисунок 12 – Разделение выборки на обучающую и тестовую, выделение прогнозируемых переменных

Для прогноза прочности при растяжении и модуля упругости при растяжении использовались следующие модели: случайный лес (RandomForestRegressor\_pr) и линейная регрессия (LinearRegression\_pr ).

В качестве метрик оценки качества были выбраны MAE и R2.



Рисунок 13 – Код модели линейной регрессии для модуля упругости при растяжении



Рисунок 14 – Код модели линейной регрессии для плотности при растяжении.

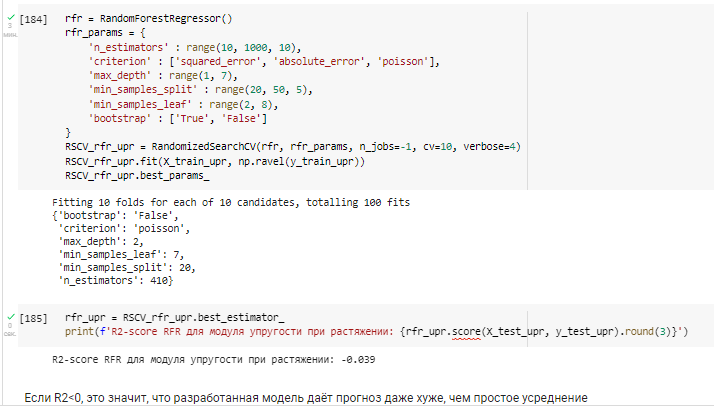


Рисунок 15 – Код модели «Случайный лес» для модуля упругости при растяжении.

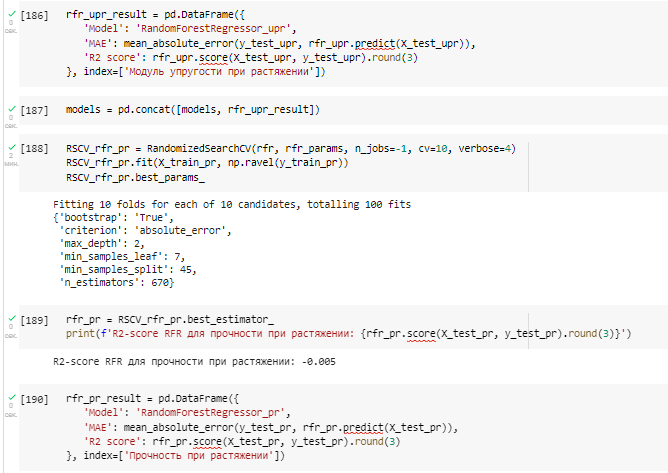


Рисунок 16 – Код модели «Случайный лес» для прочности при растяжении.

### **Тестирование модели**

Для каждой выходной переменной и по каждой модели (случайный лес и линейная регрессия) были определены ошибки:

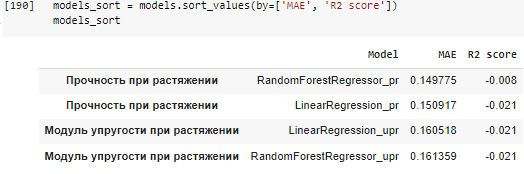


Рисунок 17 – Рассчитанные ошибки

Таким образом, ни одна из регрессионных моделей (дают очень близкие результаты) не позволяет по имеющимся данным прогнозировать модуль упругости при растяжении и прочность при растяжении.

### **Нейронная сеть для рекомендации соотношения матрица-наполнитель**

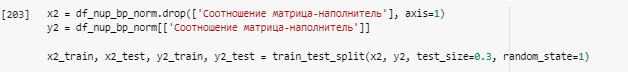
Предварительно датафрейм был разбит тестовую (30%) и обучающую выборки с выделением предикторов и целевой переменной. 

Рисунок 18 – Разделение выборки на тестовую и обучающую

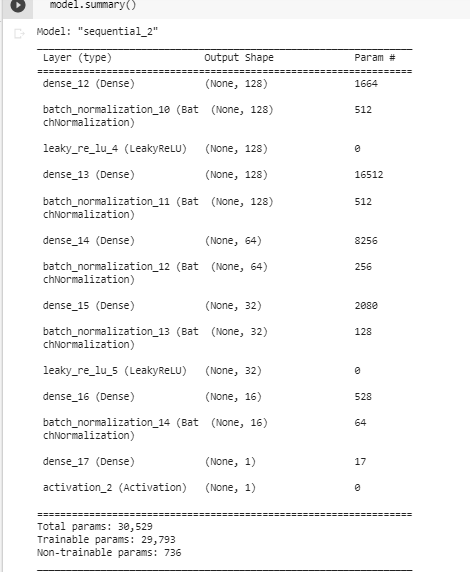
****

Рисунок 19 - Параметры нейронной сети

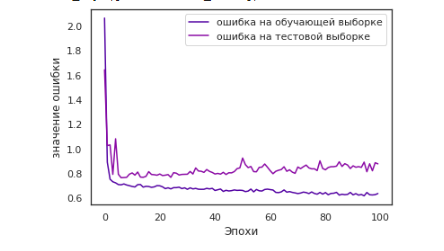
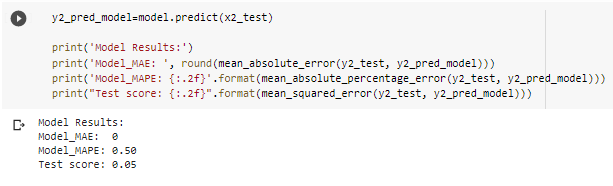
****

Рисунок 20 - График ошибки нейронной сети



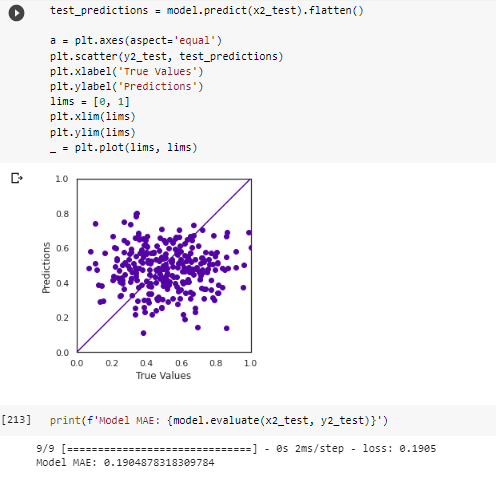


Рисунок 21 – Результаты работы нейронной сети

Таким образом, модель на основе искусственной нейронной сети не позволяет по имеющимся данным рекомендовать соотношение матрица-наполнитель.

### **Разработка приложения**

Для рекомендательной системы «Соотношение матрица-наполнитель» разработано вэб-приложение app.py. Интерфейс приложения позволяет вводить входные переменные (предикторы, независимые переменные) и применять к ним модель искусственной нейронной сети. Приложение разработано в среде разработки PyCharm. Для разработки приложения был использован интерпретатор Python для запуска веб-приложения Flask.

Приложение исполняется в среде программирования Pycharm или VSCode, создает ссылку на веб-страницу, на которой осуществляется ввод значений входных переменных и в дальнейшем вызывается исполнение модели/

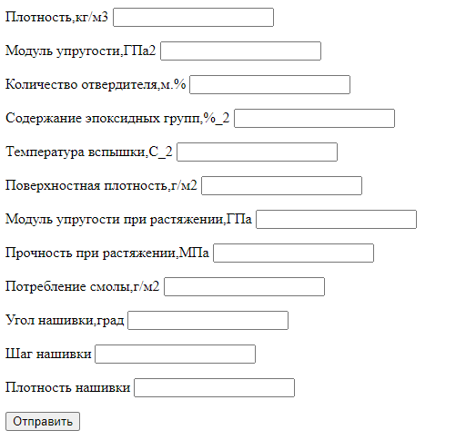


Рисунок 22 - Внешний вид приложения

Пользователь заходит на вэб-страницу (сайт) приложения, вводит необходимые 12 параметров в соответствующие окна и при нажатии кнопки «Отправить» приложение выводит в верхнем левом углу экрана расчетное значение показателя «Соотношение матрица-наполнитель».

Ссылка на страницу удаленного репозитория - <https://github.com/Max7772022/MPR>

Коммиты: пояснительная записка (файл «MyVKR.ipynb»), датасеты (папка «Data»), модели (папка «[models/my\_model](https://github.com/Max7772022/MPR/tree/main/models/my_model)»), файл README.

## **Заключение**

С целью выполнения поставленной задачи по прогнозированию ряда конечных свойств получаемых композиционных материалов в выпускной квалификационной работе были изучены теоретические основы методов машинного обучения, изучены основные библиотеки и методы Python, как одного из основных инструментов для анализа данных, было проведено теоретико-методологическое обоснование и практическое применение в исследовании некоторых основных регрессионных методов (случайного леса, линейной регрессии, разработана искусственная нейронная сеть).

При проведении анализа установлено, что распределение значений переменных близко к нормальному, имеется очень слабовыраженная зависимость между переменными, возможно переменные связывает нелинейная зависимость, что тоже является регрессией, корреляционная зависимость между переменными отсутствует.

Для прогноза прочности при растяжении и модуля упругости при растяжении были разработаны по 2 модели, показавшие близкие результаты. Подбором гиперпараметров определены наилучшие модели.

Для рекомендации соотношения матрица-наполнитель разработана модель на основе искусственной нейронной трехслойной сети с применением пакетной нормализации (batch-normalization).

В процессе построения и обучения моделей машинного обучения получен практический опыт по подготовке и анализу данных, выбору модели для решения задач регрессии, настройке таких моделей.

Разработано Flask приложение для применения модели искусственной нейронной сети при рекомендации соотношения матрица-наполнитель.

Результаты работы размещены в удаленном репозитории GitHub.

## **Библиографический список**

1. Андерсон К. Аналитическая культура. От сбора данных до бизнес-результатов. Пер. с англ. – М: Манн, Иванов и Фербер, 2017 – 336 с.
2. Артеменко С.Е., Кадыкова Ю.А., Васильева О.Г. Новые материалы и технологии и УДК 678.046.3 Базальтопластики – полимерные композиционные материалы ХХI века [Текст] / Вестник СГТУ. - 2005. №2(7). –32 с.
3. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science. Пер. с англ. – Спб.: БХВ-Петербург, 2018. – 304 с. Буланов И.М., Воробей В.В.Технология ракетных и аэрокосмических кон-струкций из композиционных материалов. –М.: МГТУ им. Н.Э. Баумана, 1998. –516 с.
4. Волокнистые и дисперсноупрочненные композиционные материалы /Под ред. Н.В. Агеева и др. –М.: Наука, 1976. –215 с Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение = Deep Learning. — М.: ДМК-Пресс, 2017. — 652 с. — ISBN 978-5-97060-554-7.
5. Дж. Себер. Линейный регрессионный анализ. — М.: Мир, 1980. — 456 с. — 13 700 экз.
6. Композиционные материалы: Справочник /Под. ред. В.В. Васильева, Ю.М.Тарнопольского. –М.: Машиностроение, 1990. –512 с.
7. Композиционные материалы: В 8 т. /Под ред. Л. Браутмана, Р. Крока. –М.: Машиностроение, 1978.-Т.3: Применение композиционных материалов в техни-ке/Под ред. Б. Нотона. –М.: Мир, 1978. –511 с.
8. Композиционные материалы: Справочник /Под ред. Д.М. Карпиноса. –Киев: Наукова думка, 1985. –592 с.
9. Миронов А.А. Машинное обучение часть I ст.9 – Режим доступа: http://is.ifmo.ru/verification/machine-learning-mironov.pdf. (дата обращения 15.04.2022)
10. Шитиков В.К., Мастицкий С.Е. Классификация, регрессия и другие алгоритмы Data Mining с использованием R: - Режим доступа:<https://ranalytics.github.io/data-mining/index.html>
11. Шоу Зед. Легкий способ выучить Python. Пер. с англ. –Москва: Эксмо, 2019 – 352
12. Документация по библиотеке numpy: – Режим доступа:<https://numpy.org/doc/1.22/user/index.html#user>.
13. Документация по библиотеке pandas: – Режим доступа:<https://pandas.pydata.org/docs/user_guide/index.html#user-guide>.
14. Документация по библиотеке matplotlib: – Режим доступа:<https://matplotlib.org/stable/users/index.html>.
15. Документация по библиотеке seaborn: – Режим доступа:<https://seaborn.pydata.org/tutorial.html>.
16. Документация по библиотеке sklearn: – Режим доступа:<https://scikit-learn.org/stable/user_guide.html>.
17. Документация по библиотеке keras: – Режим доступа:<https://keras.io/api/>.
18. Руководство по быстрому старту в flask: – Режим доступа:<https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.htm>
19. Сравнение арматурных прутьев из базальтопластика и углепластика: - Режим доступа: <http://cemgid.ru/sravnenie-armaturnyx-prutev-iz-bazaltoplastika-i-ugleplastika.html> (дата обращения: 07.06.2022)
20. Критерии оценки качества регрессионной модели, или какая модель хорошая, а какая лучше – Режим доступа:<https://www.fsight.ru/blog/kriterii-ocenki-kachestva-regressionnoj-modeli-ili-kakaja-model-horoshaja-a-kakaja-luchshe/>