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 Серов Павел Владимирович

Москва, 2023

**Оглавление**

**Введение3**

**Аналитическая часть5**

Постановка задачи5

Описание используемых методов6

Разведочный анализ данных8

**Практическая часть4**

Предобработка данных5

Разработка и обучение модели 2

Создание и обучение нейронной сети2

Создание удаленного репозитория и загрузка 2

**Заключение 4**

**Литература4**

**Введение**

В настоящее время композитные материалы являются одними из наиболее перспективных материалов для создания новых изделий в различных отраслях промышленности. Композиты представляют собой материалы, состоящие из двух или более различных компонентов, которые объединяются для образования нового материала с уникальными свойствами.

Одним из ключевых преимуществ композитных материалов является возможность сочетания различных свойств компонентов в одном материале. Например, связующий компонент может обеспечивать прочность и жесткость, а армирующий компонент - легкость и прочность при растяжении. В зависимости от соотношения компонентов, их формы и структуры, можно получить материалы с различными свойствами и характеристиками.

Существует несколько типов композитных материалов, которые различаются по характеристикам и структуре компонентов. Например, углепластик - это композитный материал, в котором армирующим компонентом является углеродное волокно, а связующим - полимерная матрица. Волокнистые композиты, такие как стекловолокноармированный пластик (СВАП), включают стекловолокно в качестве армирующего компонента и полимерную матрицу в качестве связующего компонента. Существуют также другие типы композитов, такие как металломатрические композиты и керамические композиты, которые имеют свои уникальные свойства и применения.

В данной работе мы сосредоточимся на прогнозировании конечных свойств новых композиционных материалов, основанных на данных о характеристиках входящих компонентов. Для этого мы будем использовать методы машинного обучения и симуляционные модели для создания представительного элемента объема композита. Это позволит нам предсказать свойства новых материалов на основе свойств и характеристик входящих компонентов и определить наиболее эффективные способы сочетания компонентов для достижения желаемых свойств композитного материала.

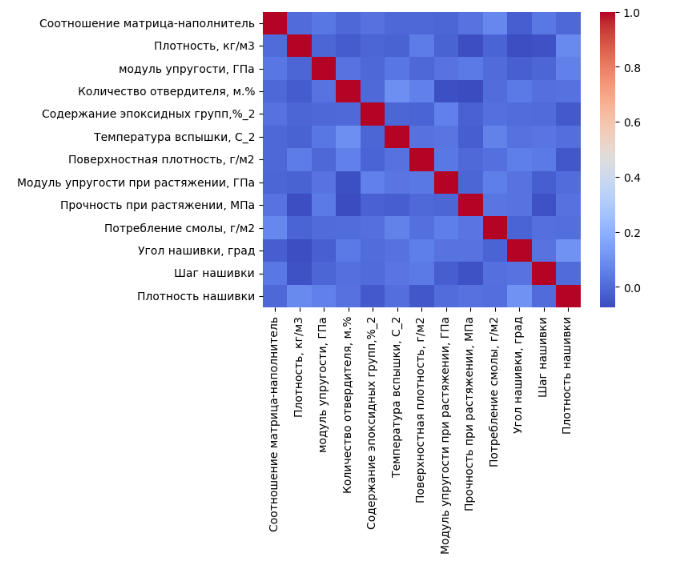
**Аналитическая часть**

Постановка задачи

В данном исследовании нам необходимо разработать модели для прогнозирования трех свойств новых материалов: модуля упругости при растяжении, прочности при растяжении и соотношения матрица-наполнитель.

Для этого у нас есть два файла, которые нужно объединить в один, используя INNER JOIN по индексу. Далее мы произведем анализ таблицы, чтобы выявить присутствующие признаки и убрать шумы. Мы сделаем описательную статистику, подсчет уникальных значений, проверку на дубликаты, проверку данных на выбросы и пропуски. Затем проведем разведочный анализ данных, используя гистограммы распределения, диаграммы ящика с усами, попарные графики рассеяния и матрицу до удаления выбросов. Мы удалим выбросы методом межквартильных расстояний, после чего проведем нормализацию и стандартизацию данных, чтобы алгоритмы машинного обучения могли работать с признаками в одном диапазоне и с наилучшей возможностью решения задачи.

Далее приступаем к написанию и обучению моделей для прогнозирования модуля упругости при растяжении и прочности при растяжении, а также нейронной сети для прогнозирования соотношения матрица-наполнитель. После проведения оценки нашей работы мы напишем файл readme.txt и создадим репозиторий на GitHub, где опубликуем код нашей работы.





**Описание используемых методов**

Задача которую нам надо решить в рамках классификации, классификация в машинном обучении относится к алгоритмам контролируемого обучения, обучения с учителем это также это задача регрессии.

В результате работы были применены следующие методы: Линейная регрессия и Случайный лес.

**Линейная регрессия** в машинном обучении с учителем. Алгоритму для обучения необходимо указать как входные данные, так и заранее подготовленные выходные данные. Все вместе это называется обучающей выборкой. Преимущества линейной регрессии в том, что она весьма проста в реализации, но надо учитывать, что выбросы могут оказывать огромное влияние на регрессию.

**Случайный лес -** это алгоритм классификации, состоящий из многих деревьев решений (ансамбль решающих деревьев), и это алгоритм машинного обучения с учителем. Он использует бэггинг и случайность признаков при построении каждого отдельного дерева, чтобы попытаться создать некоррелированный лес, прогноз которого точнее, чем у любого отдельного дерева. Плюсы метода в том, что имеет высокую точность предсказания, на большинстве задач будет лучше линейных алгоритмов; точность сравнима с точностью бустинга, практически не чувствителен к выбросам, не чувствителен к масштабированию (и вообще к любым монотонным преобразованиям) значений признаков, не требует тщательной настройки параметров, хорошо работает «из коробки». Минусы, в отличие от одного дерева, результаты случайного леса сложнее интерпретировать, алгоритм работает хуже многих линейных методов, когда в выборке очень много разреженных признаков, алгоритм склонен к переобучению на некоторых задачах, особенно на зашумленных данных, для данных, включающих категориальные переменные с различным количеством уровней, случайные леса предвзяты в пользу признаков с большим количеством уровней: когда у признака много уровней, дерево будет сильнее подстраиваться именно под эти признаки, так как на них можно получить более высокое значение оптимизируемого функционала (типа прироста информации).

**Разведочный анализ данных**

Разведочный анализ является обязательной процедурой в результате которой - получим представления о характере распределения данных, оценим качество исходных данных на наличие пропусков, выбросов. При разведочном анализе учитывается и сравнивается большое число признаков, закономерностей.

Трудно работать с данными, не понимая, что они из себя представляют, стоит ли что то удалять или нет, пропущенные данные или искаженные, безосновательно удалять что-либо не в коем случае не стоит. Именно поэтому сначала набор данных надо изучить, загрузим и выведем некоторые статистики.

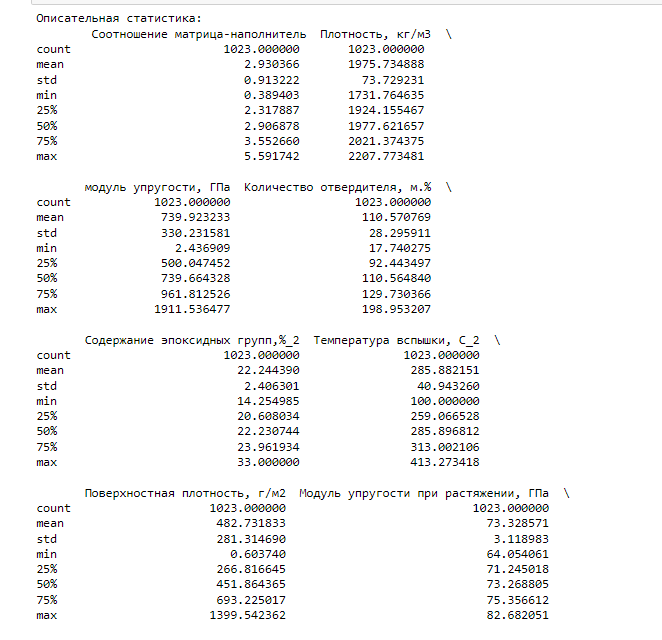


Рисунок 2.

В разведочном анализе применяются оценка характеристик датасета, гистограммы распределения, диаграммы ящика с усами, попарные графики рассеяния точек, тепловая карта, анализ и удаление выбросов, пропусков, дубликатов.

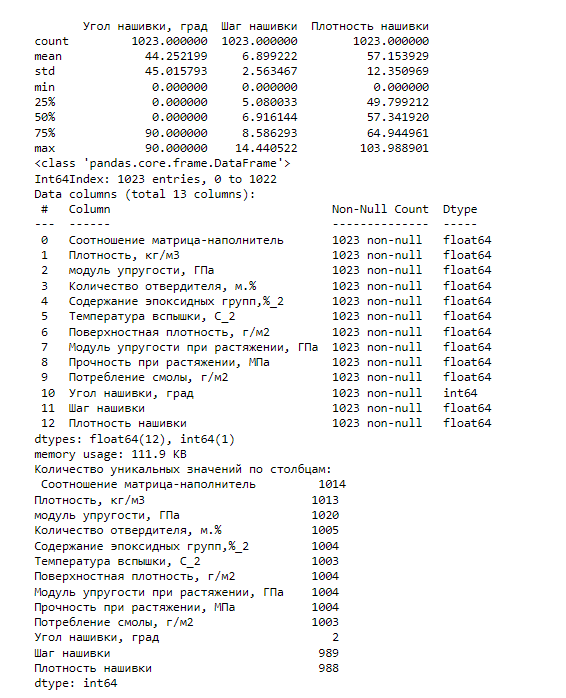


Рисунок 3. Посмотрим информацию о размерности данных.

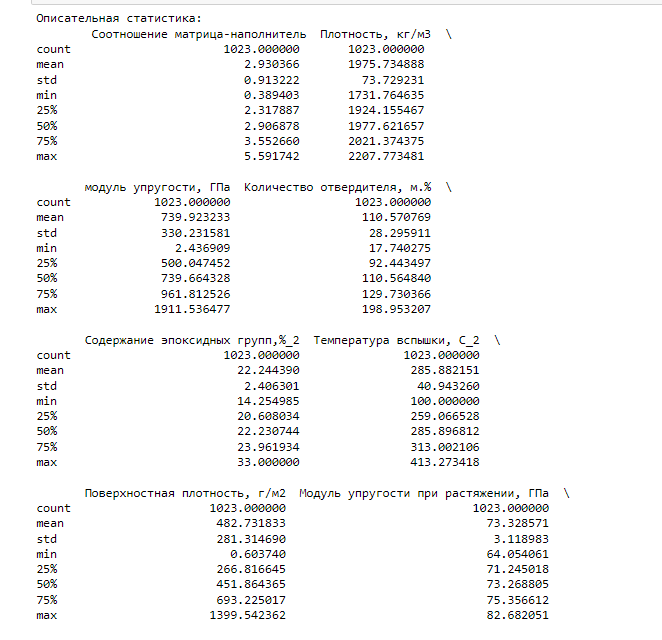


Рисунок 4. Описательная статистика.

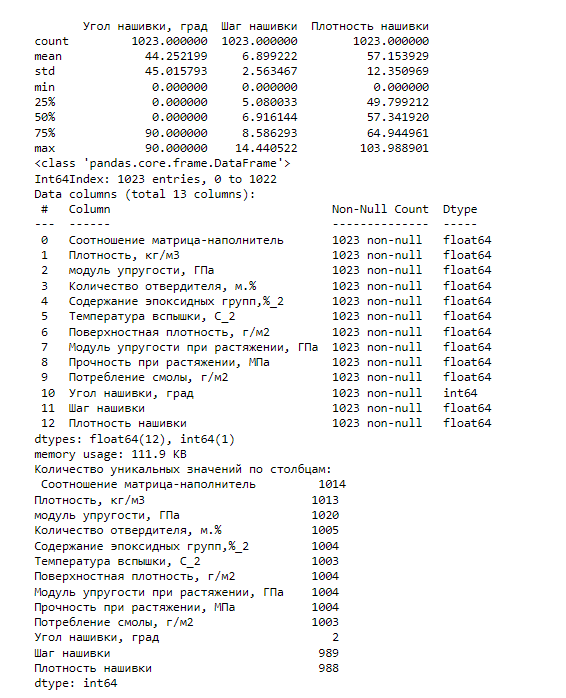


Рисунок 5. Сделаем подсчет уникальных значений по столбцам.

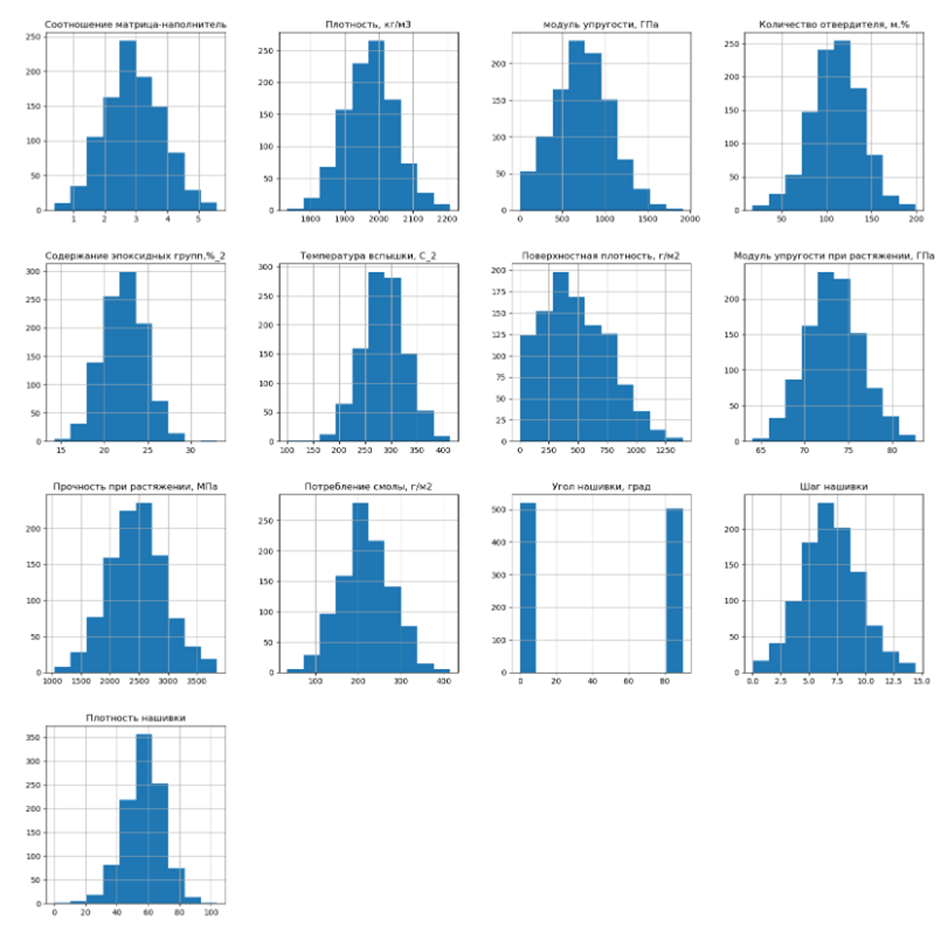


Рисунок 6. Гистограмма

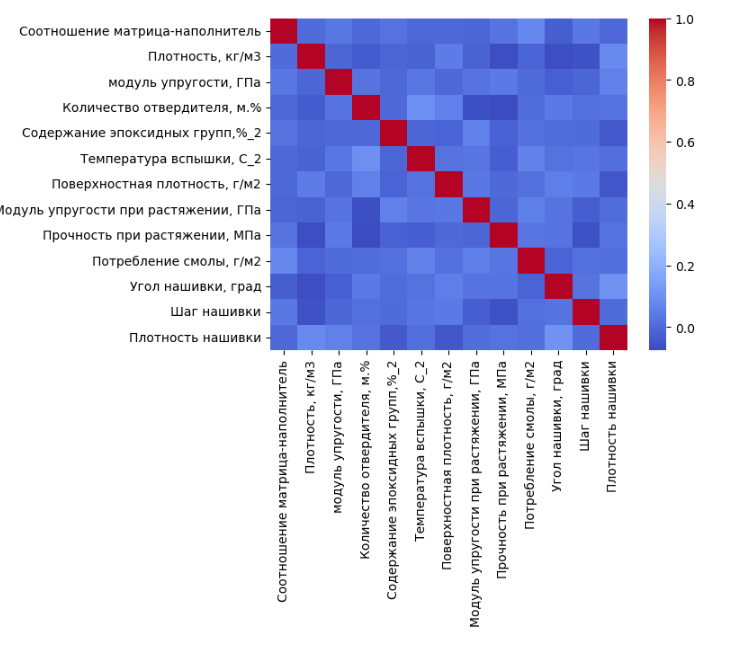


Рисунок 8. Визуализация корреляционной матрицы, тепловая карта.

Корреляция между всеми параметрами очень близка к нулю (Рисунок 8), максимальная корреляция между плотностью нашивки и углом нашивки и составляет 0.11. Все говорит об отсутствии корреляционных связей между

переменными.

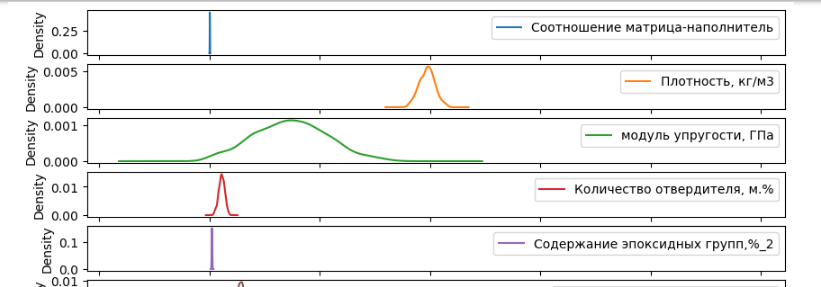
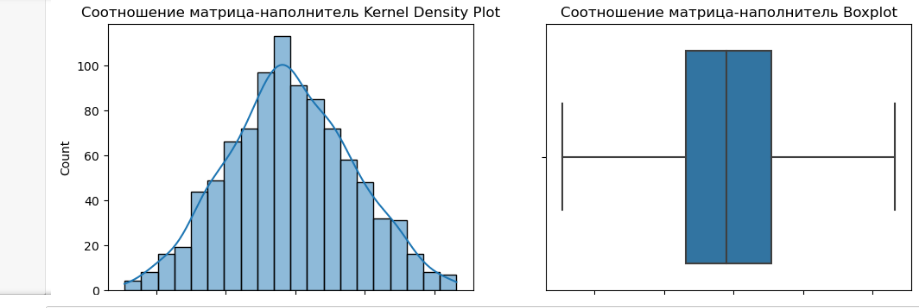


Рисунок 9. Статистика уже отобранного датасета



После анализа данных были обнаружены выбросы которые будем удалять методом межквартильных растояний IQR. IQR используется компаниями в качестве показателя уровня их дохода, также он используется для выявления выбросов и аномалий. Также может указывать на асиметрию набора данных.

**Практическая часть**

**Предобработка данных**

Если наши данные будут не в лучшем формате то мы не сможем оценить как наши признаки влияют на целевую переменную необходимо подготовить данные таким образом чтобы они давали нашим алгоритмам обучения наилучшую возможность для решения задач наша модель при обучении должна иметь возможность работать с одинаковыми признаками и иметь один диапазон. Если данные в одном порядке то необходимости к нормализации нет. Нормализация это такой процесс когда мы делаем преобразование числовых значений как правило вдиапазоне от ноля до единицы, при этом мы основываемся на максимальном и минимальном значении. Мы будем применять межквартильный интервал он определяется как разница между 75м и 25м процентилями данных.

Удаляем обнаруженные выбросы методом межквартильных интервалов.

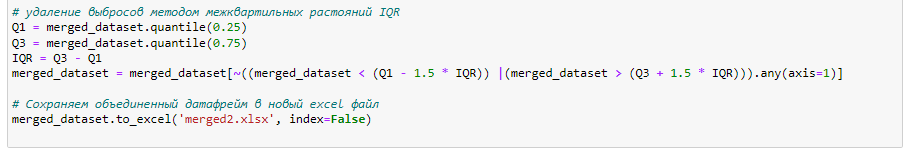


Рисунок 10.

Выбросы практически ушли дальше очищать не будем так можем утерять часть данных которые являются редкими/уникальными признаками, явные выбросы которые кординально выбиваются из общего увиденного в выборке мы убрали которые могли быть ошибкой ввода данных (аномалией).

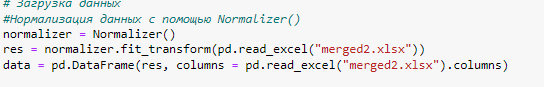
Нормализуем данные пррименяем MinMaxScaler. так как многие модели машинного обучения показывают лучшие результаты, когда значения признаков лежат в одинаковых диапазонах, а в исходных данных это не так, то используем MinMaxScaler (значения по умолчанию min=0, max=1)

, которая преобразует значения числовых признаков таким образом, что они будут лежать в диапазоне от 0 до 1.

Нормализации и приведение различных данных в самых разных единицах измерения и диапазонах значений к единому виду - сделано.

**Разработка и обучение модели**

В процессе обучения моделей проведена оценка точности этих моделей на выборках тестовой и обучающей. Разработка и обучение моделей машинного обучения согласно методам, выбранным ранее (линейная регрессия и случайный лес). Строим модели для прогноза модуля упругости при растяжении и прочности при растяжении. Загружаем данные из файла сохранения



Определяем входы и выходы для модели, разделяем датасет на тестовую и обучающую выборки рисунок 11.



Рисунок 11.

Далее применяем следующие модели: линейная регрессия\случайный лес.

Выводить оценки будем поR2 коффициент детерминации и MSE:

**Линейная регрессия**

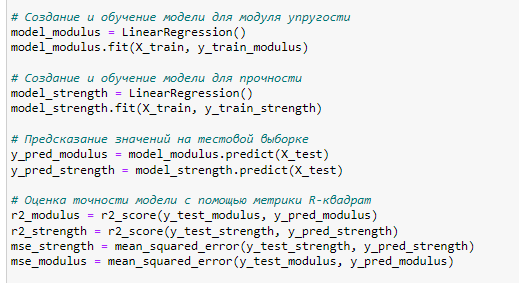
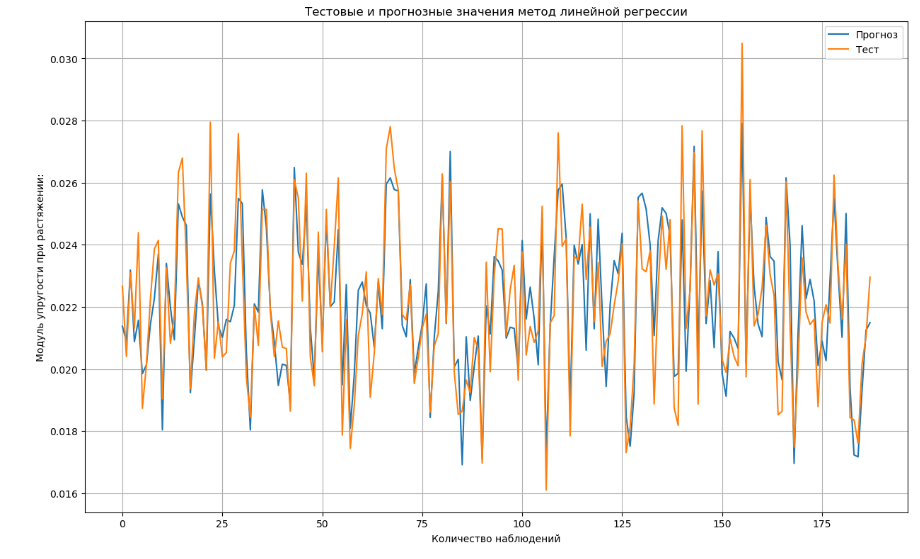


График на рисунке 12



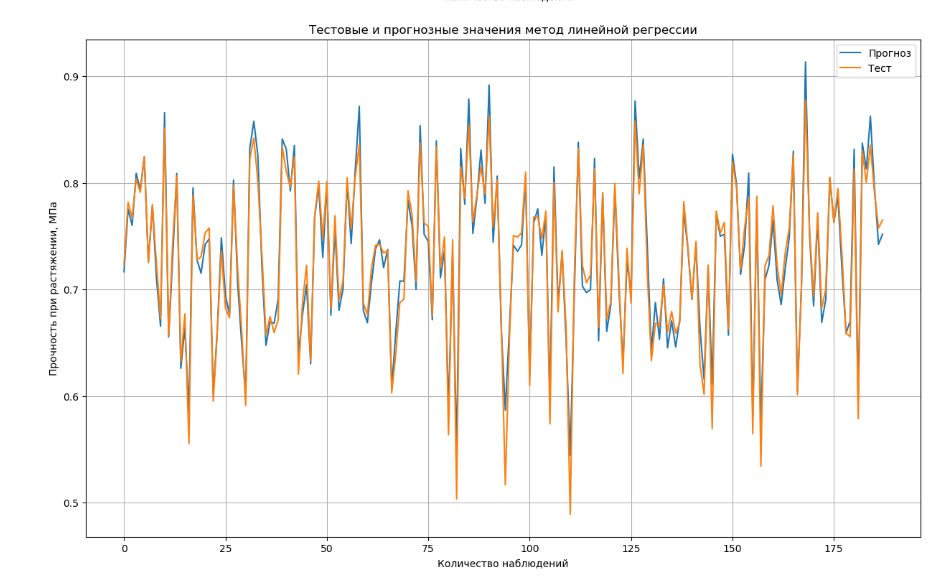


Рисунок 13.

**Случайный лес**

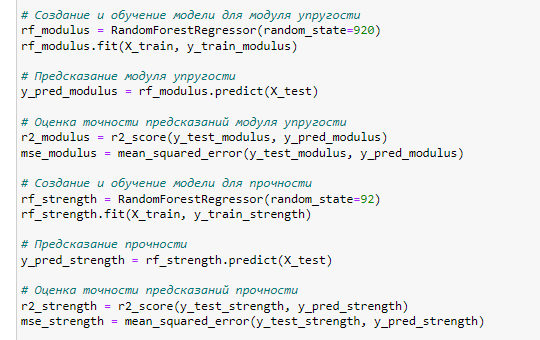


Рисунок 14.

Оценки собраны в датасет ошибок и приведена ниже:

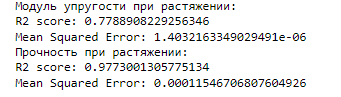


Рисунок 15.

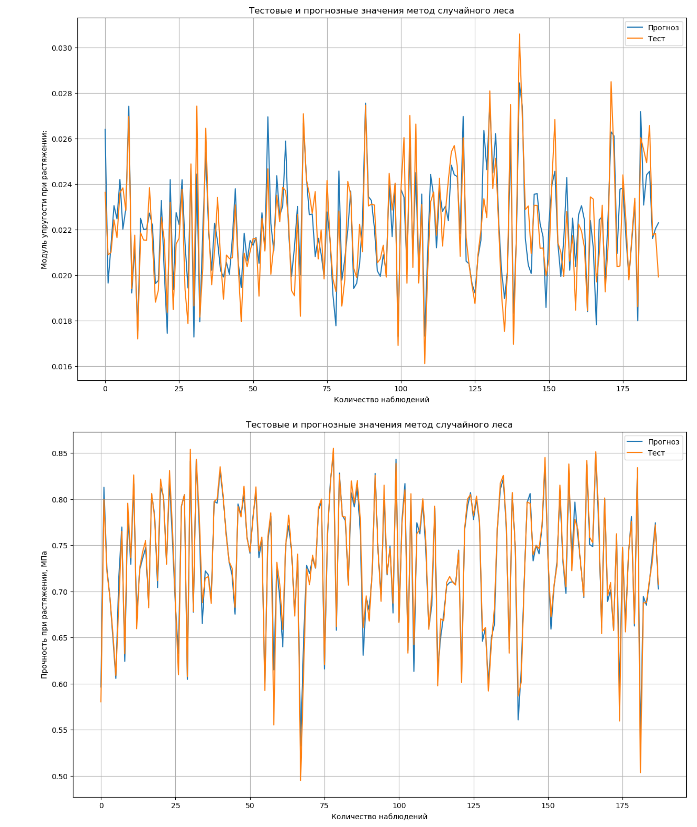


Рисунок 16.

**Создание и обучение нейронной сети**

В результате проведенной работы была создана нейро сеть при помощи библиотеки TensorFlow. Четыре скрытых слоев, конфигурации:

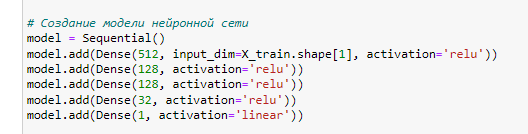


Рисунок 17.

Перед входом в нейронку данные были нормализованы. И разбиты на выборки. Подача данных в сеть. И результат на рисунке 21.

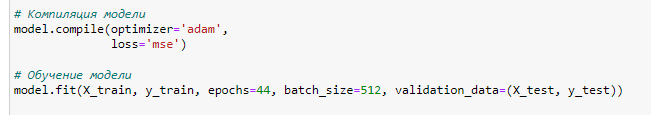


Рисунок 18.

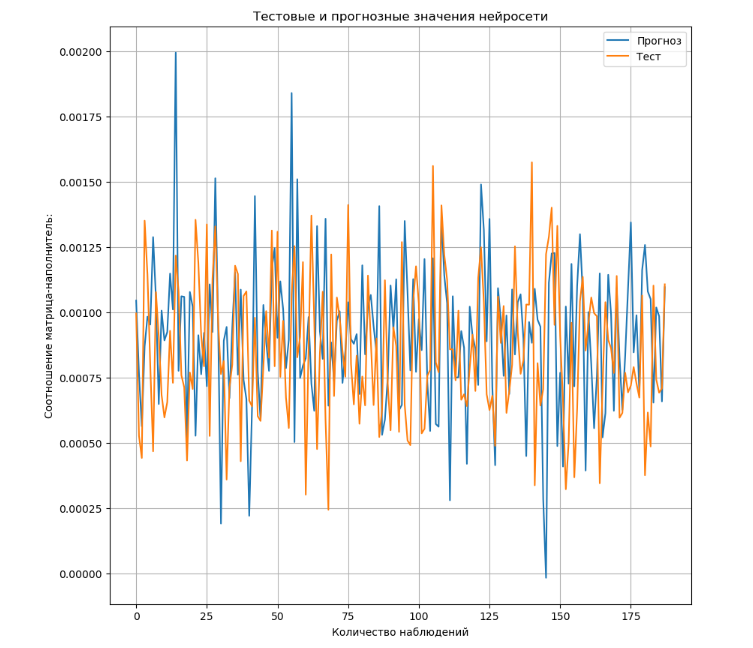


Рисунок 19.

Было использовано разное количество настроек, но все же не достаточно чтобы рекомендовать точные прогнозы. При обучении моделей было замечено что на тестовой выборке нейронная сеть ведет себя не состоятельно.

**Разработка приложения**

Приложение является частью ноутбука, туда можно вручную вписать необходимые параметры и проверить предсказание нейросети.

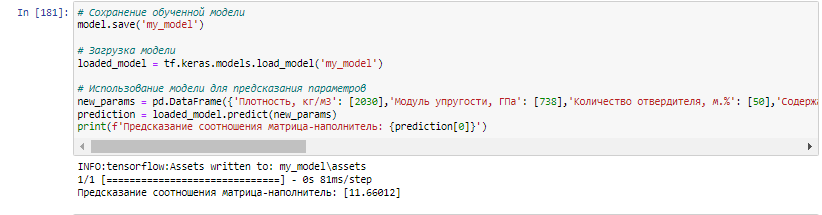
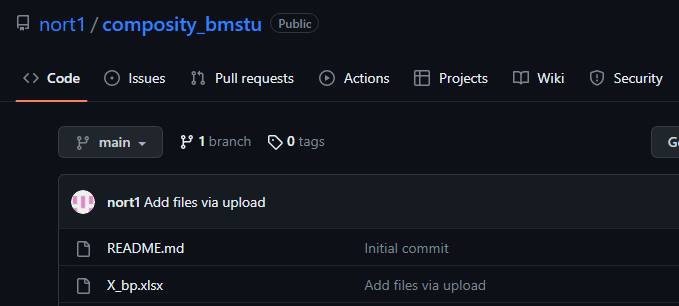


Рис. 20.

**2.6. Создание удаленного репозитория**

Репозиторий был создан на github.com по адресу: https://github.com/nort1/composity\_bmstu



**Заключение**

Машинное обучение в задачах моделей прогнозирования  – сложный процесс, требующий не только навыков программирования, но и профессиональных знаний в сфере работы с композитными материалами, в такой рааботе необходим контакт со специалистами в данной сфере для консультаций.

В ходе работы была произведена подробная опись и анализ датасета; построено множество графиков; осуществлено разбиение данных на обучающую и тестовую выборки. Для реализации моделей машинного обучения и поиска гиперпараметров были задействованы несколько алгоритмов: метод К ближайших соседей, линейная регрессия, деревья решений, опорные вектора, случайный лес. Были составлены отчеты, оценивающие качество проводимого обучения.

Было представлено сравнение результатов оценок работы моделей, графики и диаграммы, позволяющие оценить итоги проведенного обучения. Обучена и разработана нейронная сеть. В процессе выполнения данной работы получилось сделать следующие выводы. Распределение полученных данных в объединенном датасете близко к нормальному, коэффициенты корреляции между парами признаков стремятся к нулю, использованные модели не позволяют получить достаточно достоверные прогнозы, распределение полученных данных в объединенном датасете близко к нормальному коэффициенты корреляции между парами признаков стремятся к нулю использованные модели не всегда позволяют получить достоверные прогнозы лучшая метрика для прочности при растяжении, метод градиентного бустинга лучшая метрика для модуля упругости при растяжении – метод опорных векторов.

Из этого можно сделать вывод, для того чтобы определить на сколько верны полученные решения нужно дополнять датасет недостоющими данными, а для этого нужно основательно погрузится в проблематику данного направления, а именно собрать команду из специалистов по данной теме для консультаций и сбора информации.

**Библиографический список**

1. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
2. Гафаров, Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учеб. пособие /Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Издательство Казанского университета, 2018. – 121 с.
3. Джулли, Пал: Библиотека Keras - инструмент глубокого обучения / пер. с англ. А. А. Слинкин.- ДМК Пресс, 2017. – 249 с.
4. Грас, Джоэл. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.:
5. Д. Рутковская, М. Пилиньский, Л. Рутковский Нейронные сети, генетические алгоритмы и нечеткие системы. - М.: Горячая Линия - Телеком. - 2013. - 384 с. ISBN: 978-5-9912-0320-3