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 Зюликов Олег Александрович

# 

Москва, 2022

**Содержание:**

1. **Аналитическая часть 4**
   1. **Постановка задачи 4**
   2. **Описание используемых методов 7**
   3. **Разведочный анализ данных 15**
2. **Практическая часть 16**
   1. **Предобработка данных 16**
   2. **Разработка и обучение модели 23**
   3. **Тестирование модели 26**
   4. **Нейронная сеть 27**
   5. **Разработка приложения 31**
   6. **Создание удаленного репозитория 32**
   7. **Заключение 33**
   8. **Список использованной литературы 34**

# **Введение**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом.

Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы 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). На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе 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.

В процессе исследовательской работы были разработаны несколько моделей для предсказания модуля упругости при растяжении и прочности при растяжении, а также была создана нейронная сеть, которая рекомендует соотношение матрица–наполнитель при определенных параметрах. На основе этой нейронной сети было создано веб-приложение с использованием фреймворка Flask, которое запрашивает у пользователя значения входных параметров и выдает рекомендацию по соотношению матрица-наполнитель.

1. **Аналитическая часть**
   1. **Постановка задачи**

Для исследовательской работы были даны два файла: X\_bp.xlsx и X\_nup.xlsx. Необходимо было загрузить эти файлы и сделать объединение в один датасет по индексу. Для этого использовать тип объединения INNER.

Целью работы является создать модели машинного обучения для прогноза модуля упругости при растяжении и прочности при растяжении. А также разработать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель при заданных параметрах.

Перед созданием моделей необходимо провести разведочный анализ данных. Для каждого признака получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; пред обработать данные: удалить шумы и выбросы, сделать нормализацию и стандартизацию.

После разработки моделей нужно создать приложение на Flask для рекомендации значения соотношения матрица-наполнитель, а также загрузить результаты работы на удаленный репозиторий GitHab.

Перед началом работы был произведен импорт необходимых для работы библиотек. Также необходимые библиотеки добавлялись в этот список по мере надобности в процессе работы над заданием.

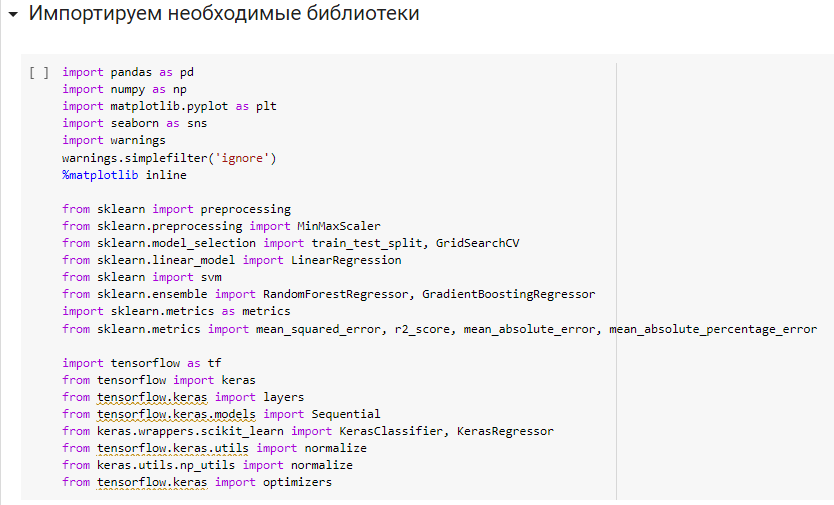


Рисунок 1 - начало работы – импорт необходимых библиотек

После этого были загружены наши файлы с данными, которые были объединены в один датасет.



Рисунок 2 – часть датасета, загруженного из файла

Для объединения файлов мы использовали тип объединения INNER, а объединение проводили по индексу .

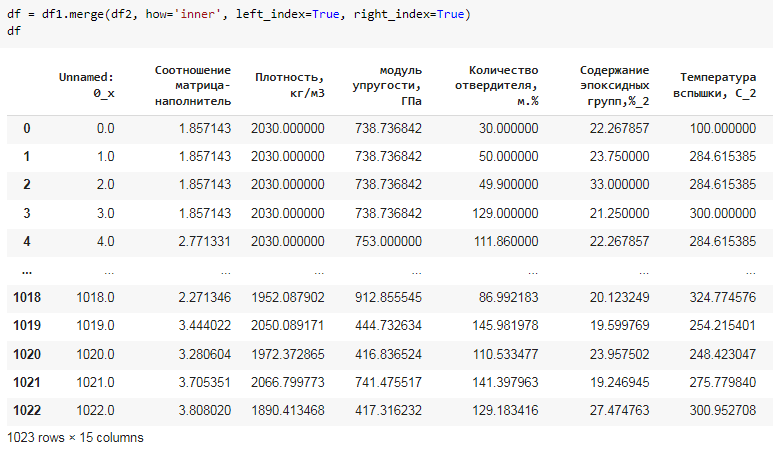


Рисунок 3 – объединение файлов в один датасет

Изначально данные из файла X\_bp.xlsx имели размер 1023 строки и 11 колонок, а данные из файла X\_nup.xlsx – 1040 строк и 4 колонки. После объединения наш датасет приобрел размер 1023 строки и 15 колонок, причем две колонки были одинаковы – это индексы из обоих файлов. Было принято решение удалить эти колонки. После удаления размер датасета стал 1023 строки и 13 колонок. То есть у нас получилось 13 признаков и 1023 строки их значений.

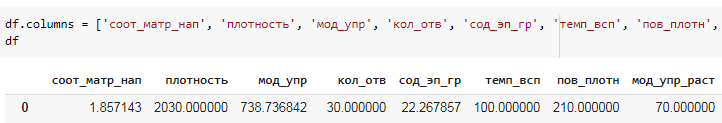


Рисунок 4 – корректировка заголовков признаков

Для удобства работы мы произвели корректировку заголовков признаков, переименовав их и сделав более короткими.

После этого наш датасет стал готов к разведочному анализу.

* 1. **Описание используемых методов**

Для решения этой задачи использовались методы машинного обучения с учителем, именно методы регрессии. В процессе работы были применены следующие методы:

* линейная регрессия;
* метод опорных векторов;
* градиентный бустинг;
* случайный лес;

Линейная регрессия — это алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Это один из самых простых и эффективных инструментов статистического моделирования. Она определяет зависимость одной (объясняемой, зависимой) переменной от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости.

Модель линейной регрессии является часто используемой и наиболее изученной. А именно изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели

Достоинства метода: быстр и прост в реализации; легко интерпретируем; имеет меньшую сложность по сравнению с другими алгоритмами;

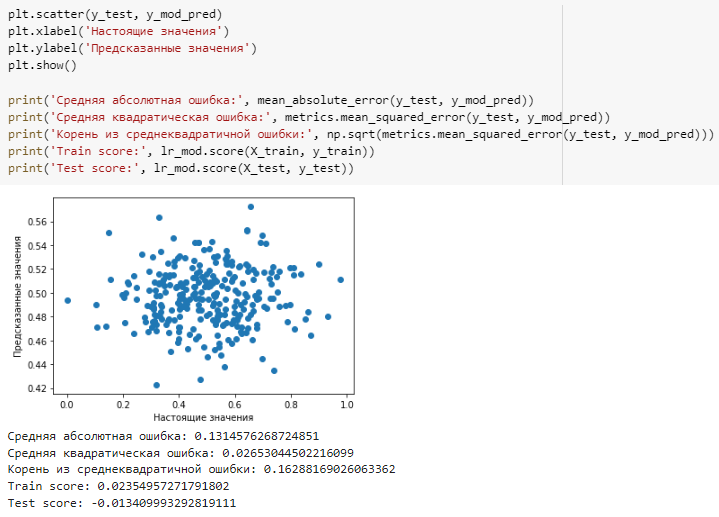


Рисунок 5 – результаты и график рассеяния настоящих и предсказанных значений методом линейной регрессии

Недостатки метода: моделирует только прямые линейные зависимости; требует прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние.

Метод опорных векторов - это метод машинного обучения, целью которого является попытка классифицировать входные наборы данных в один из двух классов. Для эффективной работы метода сначала необходимо использовать обучающую выборку, состоящую из входных и выходных данных, которая необходима для построения модели метода опорных векторов, и которую в дальнейшем можно будет использовать для классификации новых данных.

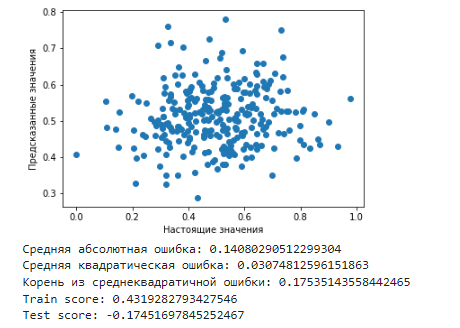


Рисунок 6 – результаты и график рассеяния настоящих и предсказанных значений методом опорных векторов

Для построения модели метода опорных векторов нужно взять обучающие входные данные, отобразить их в многомерное пространство, а затем использовать регрессию, чтобы найти гиперплоскость (гиперплоскость - это поверхность в n-мерном пространстве, которая разделяет его на два подпространства), которая лучше всего разделяла бы два класса входных данных. После обучения модели она способна классифицировать новые входные данные в один из классов при помощи разделяющей гиперплоскости.

По существу, метод опорных векторов является методом входов/выходов. Пользователь вводит входные данные, и на основе разработанной (при помощи обучения) модели получает выходные результаты. Теоретически, число входов для метода опорных векторов лежит в диапазоне от одного до бесконечности. Однако, в практическом применении, есть определенные ограничения на размер входной выборки, которые зависят от вычислительной мощности. Метод опорных векторов – это бинарный линейный классификатор, который хорошо работает на небольших датасетах. Данный алгоритм – это алгоритм обучения с учителем, использующихся для задач классификации и регрессионного анализа.

Достоинства метода: для классификации достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных

Недостатки метода: неустойчивость к шуму

Градиентный бустинг — метод машинного обучения, который создает решающую модель прогнозирования в виде ансамбля слабых моделей прогнозирования, обычно деревьев решений. Он строит модель поэтапно, позволяя оптимизировать произвольную дифференцируемую функцию потерь.

Основная идея градиентного бустинга: строятся последовательно несколько базовых классификаторов, каждый из которых как можно лучше компенсирует недостатки предыдущих. Финальный классификатор является линейной композицией этих базовых классификаторов.

Достоинства метода: новые алгоритмы учатся на ошибках предыдущих; требуется меньше итераций, чтобы приблизиться к фактическим прогнозам; наблюдения выбираются на основе ошибки; прост в настройке темпа обучения и применения; легко интерпретируем.

Недостатки метода: необходимо тщательно выбирать критерии остановки, иначе это может привести к переобучению; наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок чем нейронные сети.

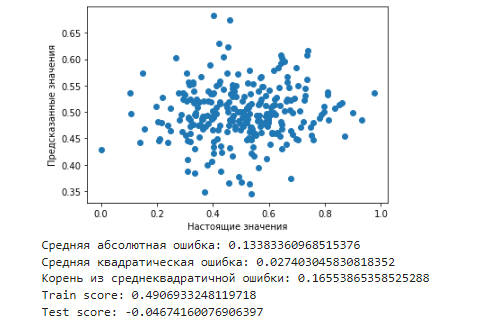


Рисунок 7 – результаты и график рассеяния настоящих и предсказанных значений методом градиентного бустинга

Случайный лес – ансамблиевый метод, способный выполнять как задачи регрессии, так и классификации с использованием нескольких деревьев решений и техники, называемой начальной загрузкой и агрегацией, обычно известной как bagging. Основная идея, лежащая в основе этого, заключается в объединении нескольких деревьев решений для определения конечного результата, а не в том, чтобы полагаться на отдельные деревья решений.

Случайный лес имеет несколько деревьев решений в качестве базовых моделей обучения. Мы случайным образом выполняем выборку объектов из набора данных, формируя образцы наборов данных для каждой модели. Эта часть называется Bootstrap. Достоинства метода: не переобучается; не требует предобработки входных данных; эффективно обрабатывает пропущенные данные, данные с большим числом классов и признаков; имеет высокую точность предсказания и внутреннюю оценку обобщающей способности модели, а также высокую масштабируемость.

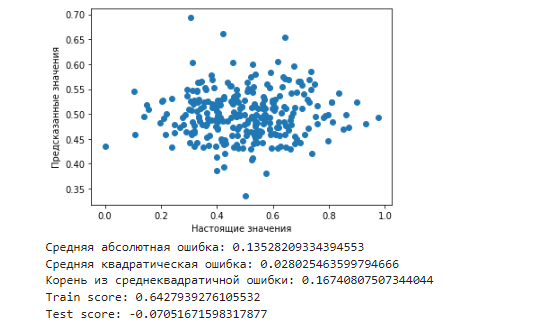


Рисунок 8 – результаты и график рассеяния настоящих и предсказанных значений методом случайного леса

Недостатки метода: построение занимает много времени; сложно интерпретируемый; не обладает возможностью экстраполяции; может недо обучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы.

При построении модели случайного леса также был произведен поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой.

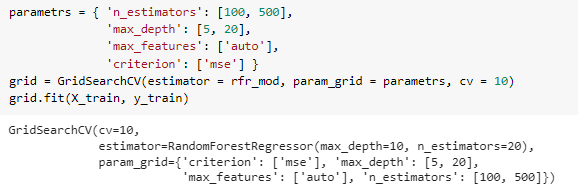


Рисунок 9 – поиск гиперпараметров для модели случайного леса

После подстановки найденных гиперпараметров в модель, алгоритм выдал более точный результат предсказания на наших данных.

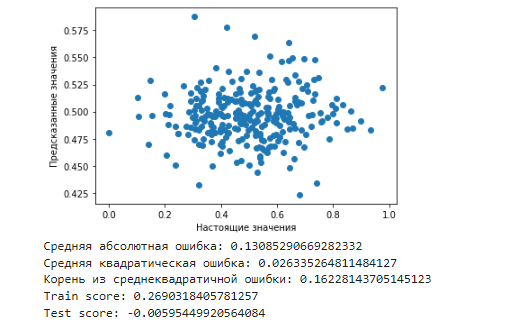


Рисунок 10 – результаты и график рассеяния настоящих и предсказанных значений методом случайного леса с подбором гиперпараметров

При сравнении эффективности алгоритмов нами были применены следующие метрики: средняя абсолютная ошибка (MAE), средняя квадратическая ошибка (MSE) и корень из среднеквадратической ошибки (RMSE), а также мы рассчитывали score на тренировочной и тестовой выборках.

Практически во всех моделях score на тренировочной выборке был значительно выше, чем на тестовой. Что говорит о недостаточном обучении нашей модели для эффективного предсказания результата.

Наиболее точные прогнозы были сделаны с помощью метода линейной регрессии, как в предсказаниях модуля упругости при растяжении, так и для предсказания прочности при растяжении.

Ниже представлены сравнительные показатели ошибок используемых моделей с помощью метрики MAE.

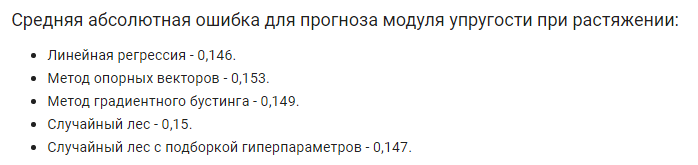


Рисунок 11 – средняя абсолютная ошибка для прогноза модуля упругости при растяжении

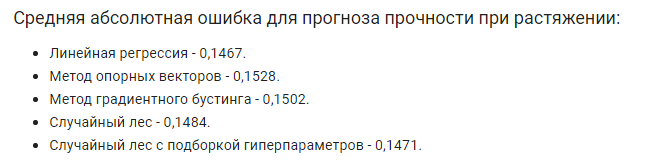


Рисунок 12 – средняя абсолютная ошибка для прогноза прочности при растяжении

* 1. **Разведочный анализ данных**

Для того, чтобы работа с данными была эффективна, необходимо их соответствующим образом обработать: выявить пропущенные значения, дубликаты, выбросы, очистить данные. Эти задачи решает разведочный анализ данных.

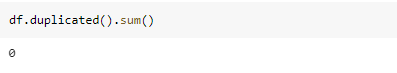


Рисунок 13 – поиск дубликатов в нашем датасете

С помощью него мы можем посмотреть на наши данные с различных сторон, изучить их для дальнейшей эффективной обработки.

В нашем наборе данных ни пропусков, ни дубликатов не оказалось.

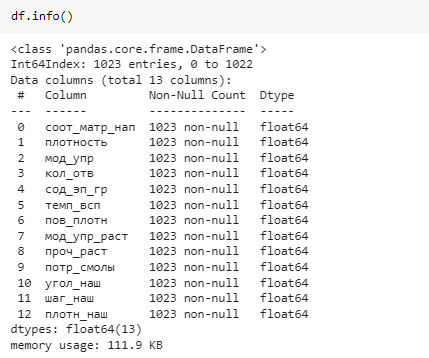


Рисунок 14 – изучение данных, поиск пропущенных значений

При разведочном анализе данных мы также в обязательном порядке должны посмотреть описательные статистики нашего датасета. С помощью описательных статистик мы провели оценку статистических характеристик датасета: среднее и медианное значение для каждой колонки, минимальное и максимальное значение, а также квартили.

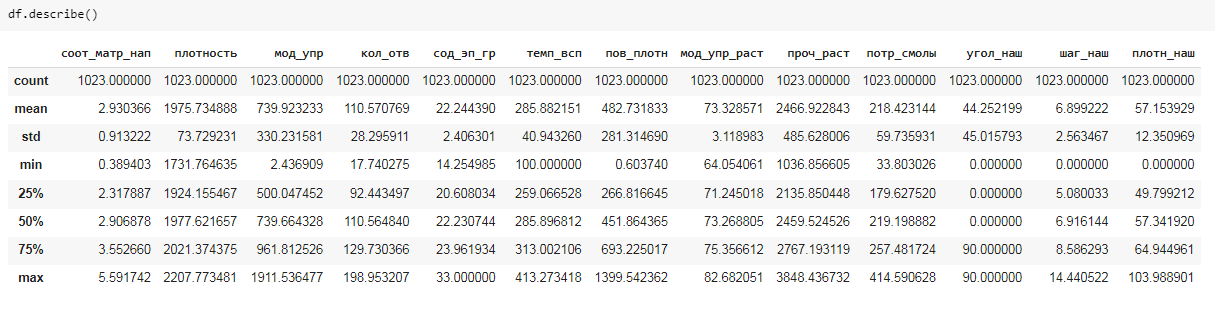


Рисунок 15 – изучение описательных статистик с помощью метода describe()

В качестве инструментов разведочного анализа данных также используются: гистограммы распределения каждой из переменной; диаграммы ящика с усами; попарные графики рассеяния точек.

1. **Практическая часть**
   1. **Предобработка данных**

Построим гистограммы для изучения распределения наших переменных.



Рисунок 16 – гистограммы распределения переменных

Также для изучения наших признаков построим попарные графики рассеяния.

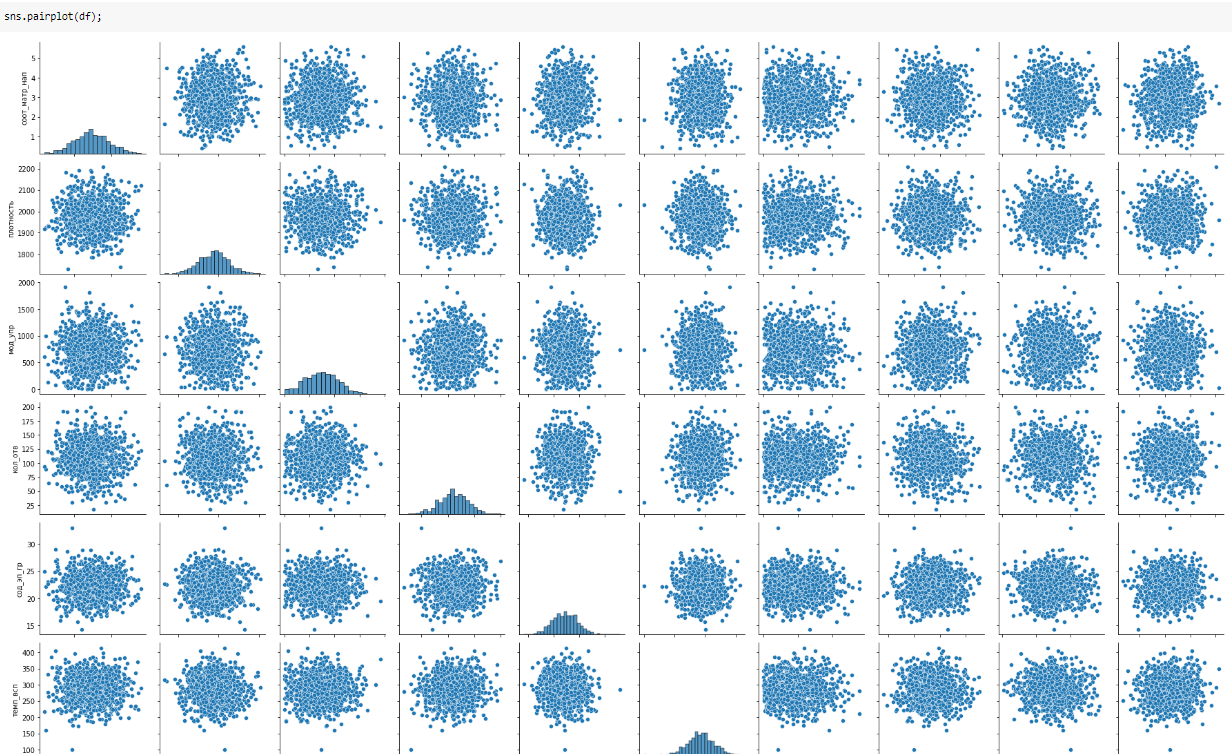


Рисунок 17 – попарные графики рассеяния

Здесь мы можем наблюдать, что какая-либо значимая корреляция между признаками визуально отсутствует.

Важным этапом обработки данных является удаление выбросов из нашего набора данных. Это необходимо для более точного предсказания наших целевых переменных, поскольку выбросы будут очень сильно влиять на итог.

Для этой цели мы будем использовать диаграммы ящика с усами. Эти диаграммы позволяют визуально определить выбросы в наших данных. Мы построим диаграммы, с их помощью определим числовые значения наших выбросов и удалим их, чтобы они не влияли на точность наших предсказаний.

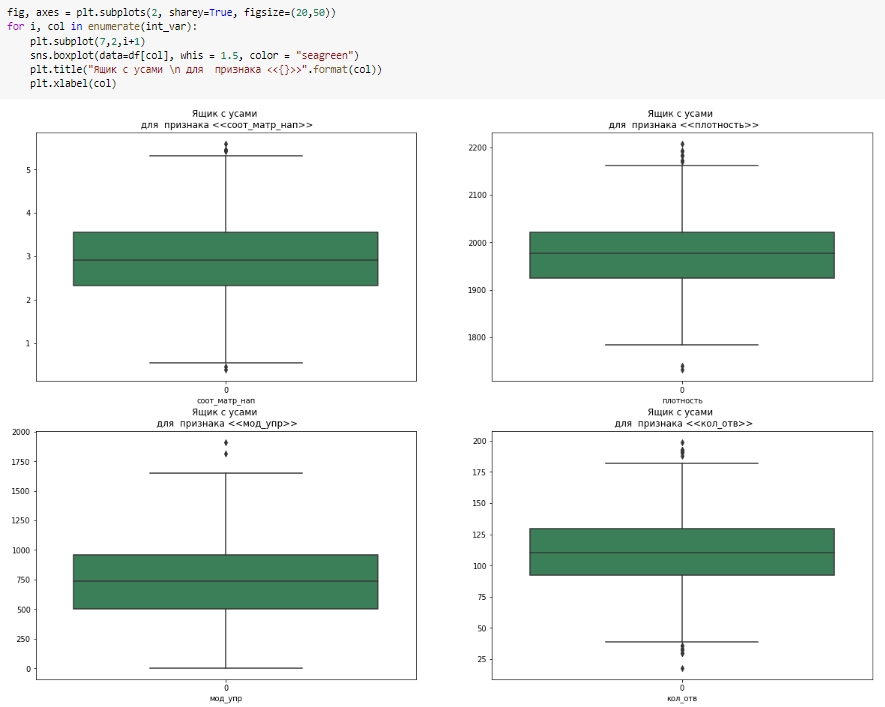


Рисунок 18 – диаграммы ящика с усами

На диаграммах видны точки за пределами – это выбросы, мы их удалим, чтобы их влияние не сказывалось негативно на предсказаниях наших моделей. Причем наши целевые признаки – соотношение матрица-наполнитель, модуль упругости при растяжении и прочность при растяжении мы не будем трогать, поскольку есть вероятность, что это не выбросы, а просто нам не хватает данных для более точной статистической оценки.

После удаления выбросов в нашем датасете осталось 924 строки.

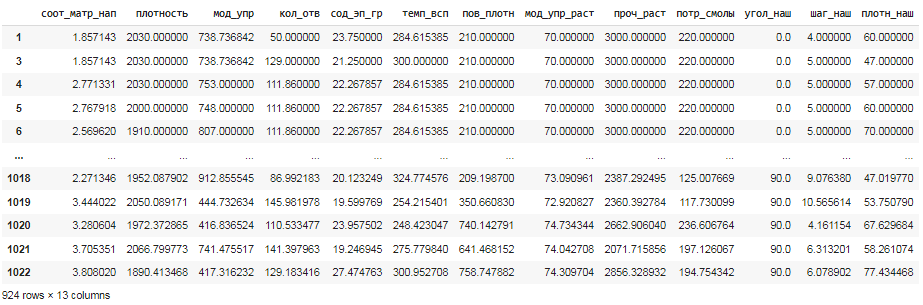


Рисунок 19 – датасет после удаления выбросов

Следующим нашим шагом будет нормализация данных. Для того, чтобы привести данные в диапазон между 0 и 1. Это позволит моделям более точно сравнивать данные друг с другом.

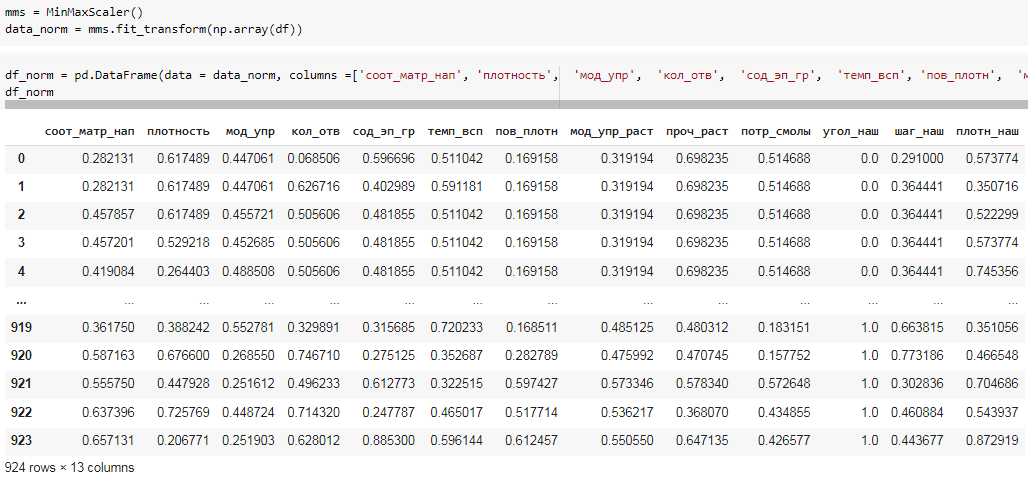


Рисунок 20 – датасет после нормализации

Проведем визуализацию наших данных после обработки.

Построим гистограммы распределения.



Рисунок 21 – гистограммы распределения признаков после обработки

Визуально мы наблюдаем, что после обработки у наших признаков распределение стало ближе к нормальному.

Построим попарные графики рассеяния и диаграммы ящика с усами.



Рисунок 22 – попарные графики рассеяния после обработки

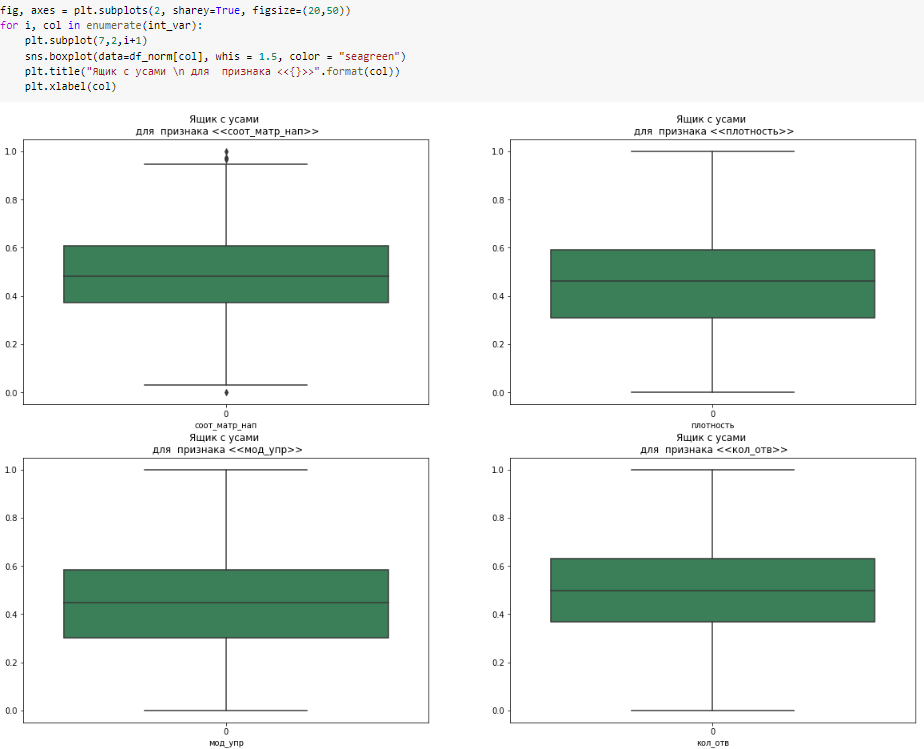


Рисунок 23 – диаграммы ящика с усами после обработки

На диаграммах ящика с усами мы видим, что у тех признаков, которые мы корректировали, выбросы отсутствуют.

Посмотрим на корреляцию наших признаков.

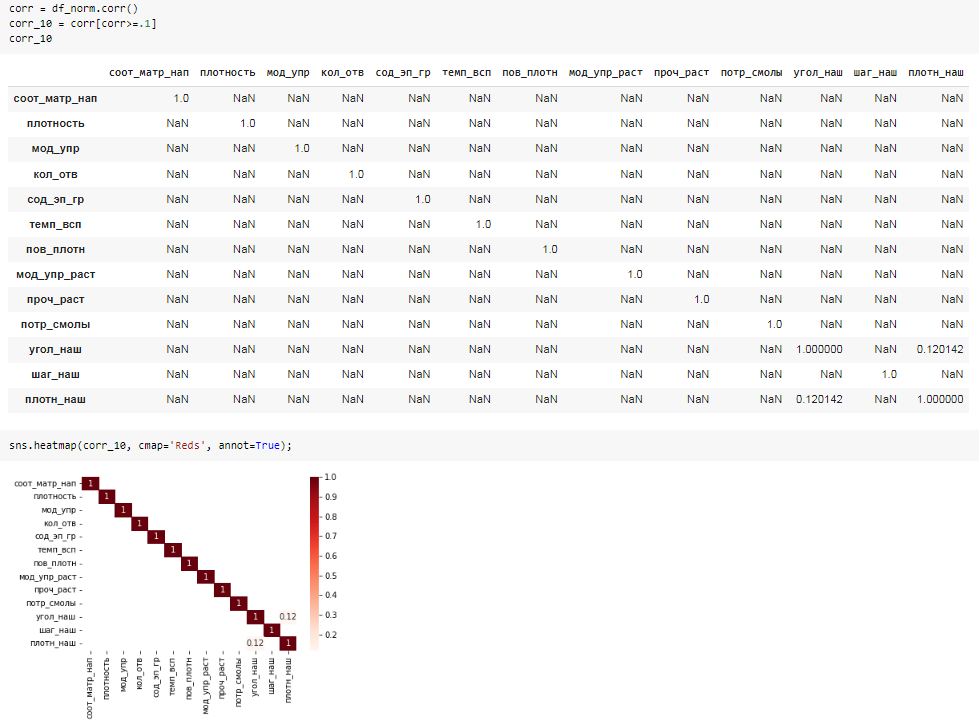


Рисунок 24 – корреляция признаков

Мы видим, что практически все признаки между собой имеют корреляцию менее 0,1.

* 1. **Разработка и обучение модели**

Разработка и обучение моделей машинного обучения осуществлялась для двух целевых переменных - прочность при растяжении и модуль упругости при растяжении.

Для решения были применим методы, описанные выше. Модели строились отдельно для каждой из целевых переменных.

Для предсказаний мы использовали датасет без наших целевых признаков. То есть признаки соотношение матрица-наполнитель, модуль упругости при растяжении и прочность при растяжении мы убрали из датасета.

Линейная регрессия:

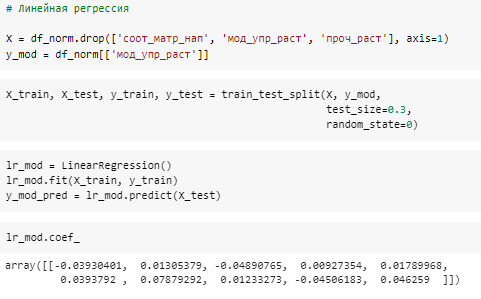


Рисунок 25 – построение модели линейной регрессии

Метод опорных векторов:

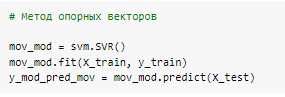


Рисунок 26 – построение модели по методу опорных векторов

Метод градиентного бустинга:

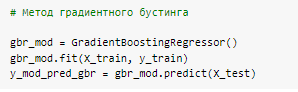


Рисунок 27 – построение модели градиентного бустинга

Случайный лес:

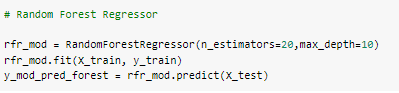


Рисунок 28 – построение модели случайного леса

Для модели случайного леса мы применили поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.

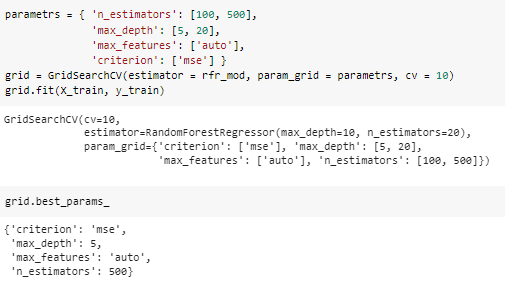
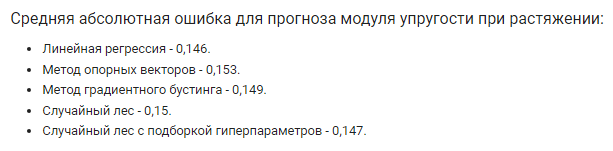


Рисунок 29 – поиск гиперпараметров для модели случайного леса

Нужно заметить, что после применения лучших гиперпараметров, модель выдала более хороший результат, чем прежде.

* 1. **Тестирование модели**

После обучения моделей была проведена оценка точности этих моделей, в качестве параметра оценки модели использовалась средняя абсолютная ошибка (MAE).



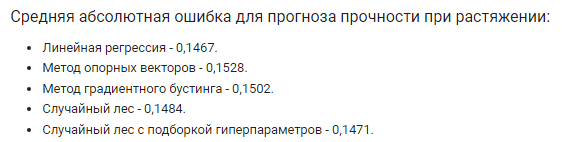


Рисунок 30 – сравнение эффективности моделей машинного обучения

Лучший результат показала модель линейной регрессии, незначительно уступила ей модель случайного леса с подобранными гиперпараметрами. На третьем месте модель случайного леса до подбора гиперпараметров. Модели, разработанные с применением метода опорных векторов и метода градиентного бустинга показали худшие результаты.

**2.4 Нейронная сеть**

Обучение нейронной сети - это процесс, в котором параметры нейронной сети настраиваются посредством моделирования среды, в которую эта сеть встроена. Тип обучения определяется способом подстройки параметров. В процессе обучения происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки.

В нашей работе для предсказания соотношения матрица-наполнитель мы будем использовать многослойный персептрон.

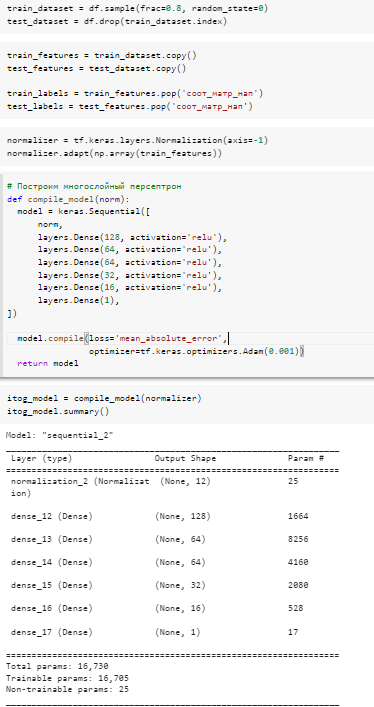


Рисунок 31 – построение нейронной сети

После построения нейронной сети, обучим ее.

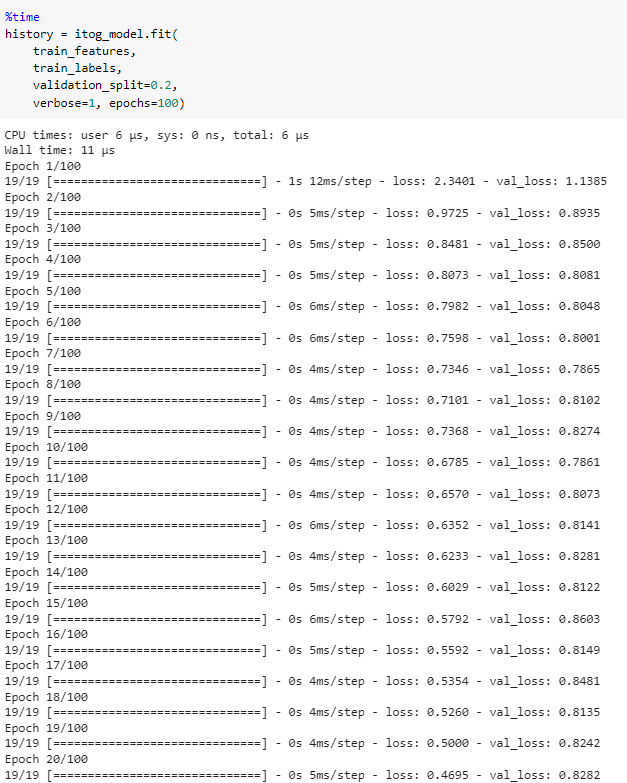


Рисунок 32 – обучение нейронной сети

Посмотрим на историю обучения нашей нейронной сети и на график потерь модели на тренировочной и тестовой выборках.

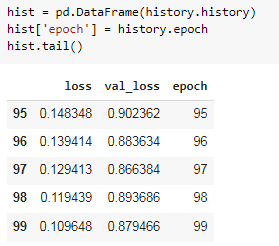


Рисунок 33 – история обучения нейронной сети

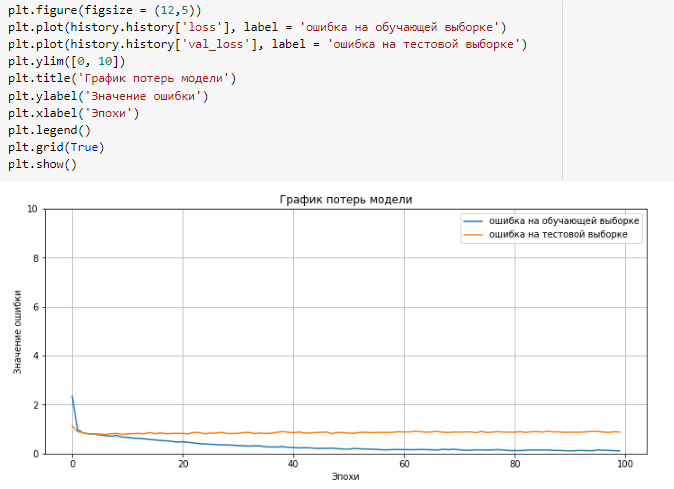


Рисунок 34 – график потерь модели

Из графика мы видим, что на обучающей выборке модель успешно обучилась и показала хороший результат, но на тестовой выборке результат предсказания оставляет желать лучшего. Это также подтверждается и графиком рассеяния.

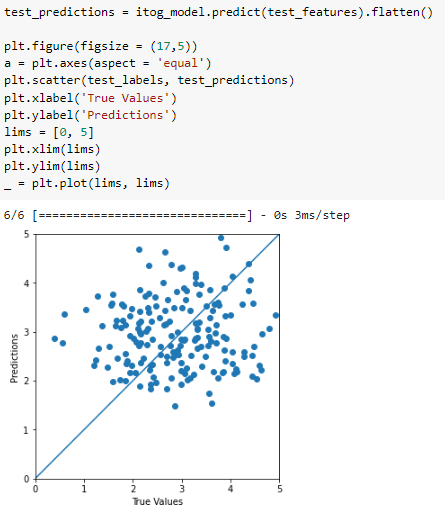


Рисунок 35 – график рассеяния предсказания и реального значения

**2.5 Разработка приложения**

Следующим этапом нашей работы стала разработка приложения на фреймворке Flask для рекомендации соотношения матрица-наполнитель по заданным пользователем параметрам.

Пользователь вводит в окно ввода значение показателей и нажимает на кнопку «Отправить».

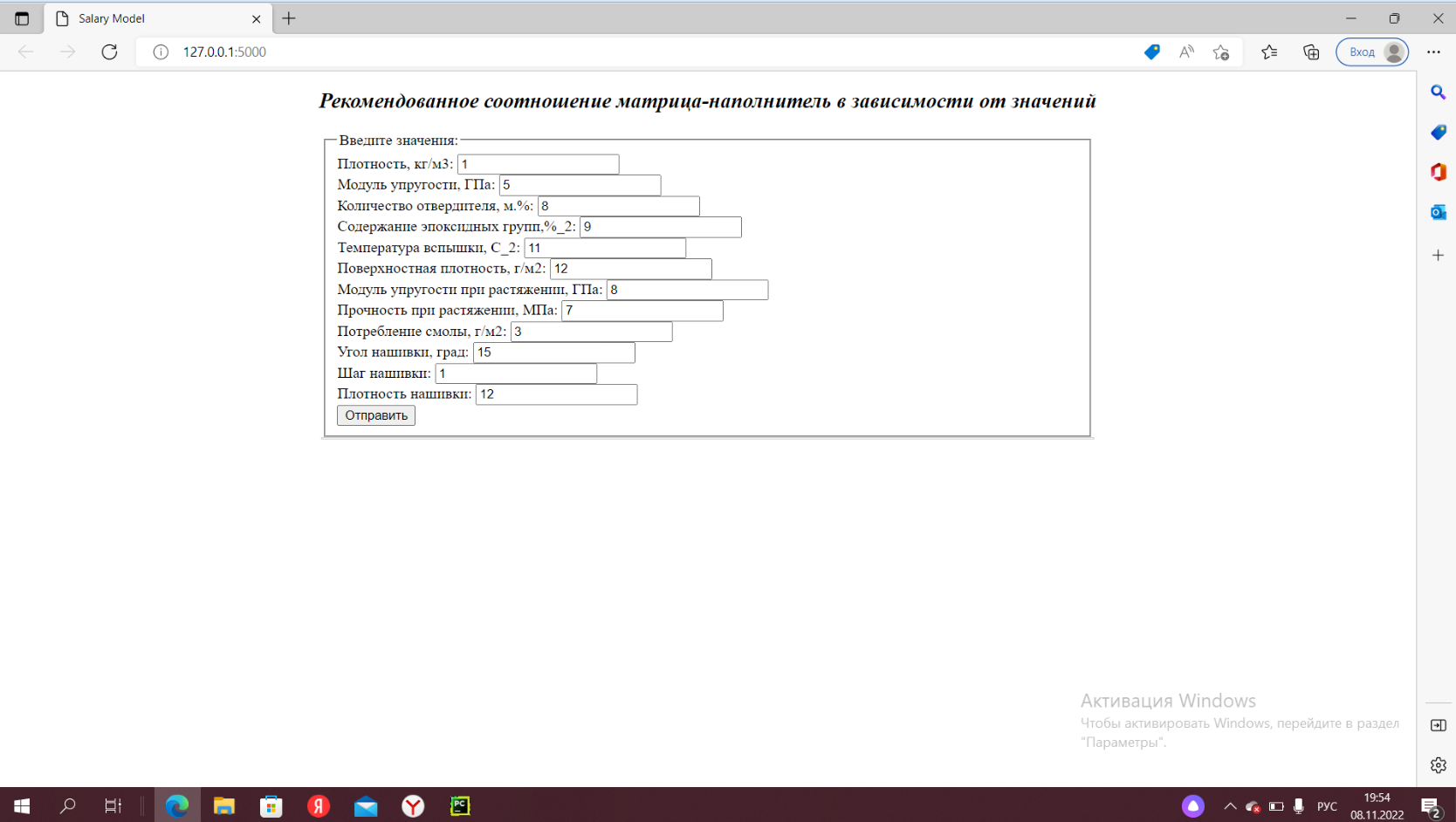


Рисунок 36 – окно ввода пользовательских значений

После этого программа выдает результат с рекомендацией соотношения матрица-наполнитель, рассчитанный на основе сохраненной модели нашей нейронной сети.

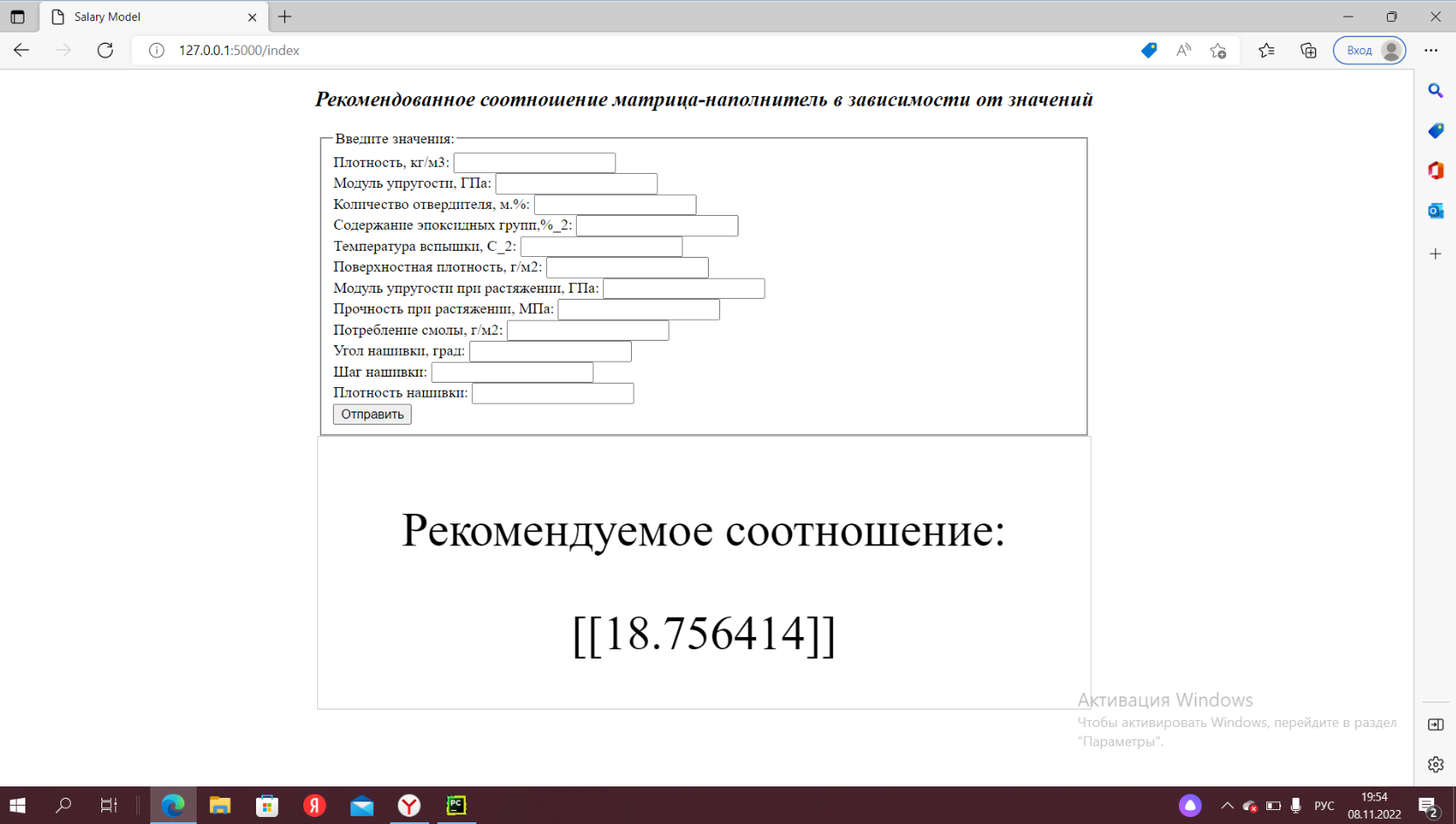


Рисунок 37 – результат работы приложения

* 1. **Создание удалённого репозитория**

Для загрузки результатов работы был создан репозиторий github.com по адресу: <https://github.com/OlegZiulikov/composite-forecast>.

Туда были загружены файлы с итоговой работой, файл созданного приложения, html-файл для приложения Flask, а также файлы с набором данных.

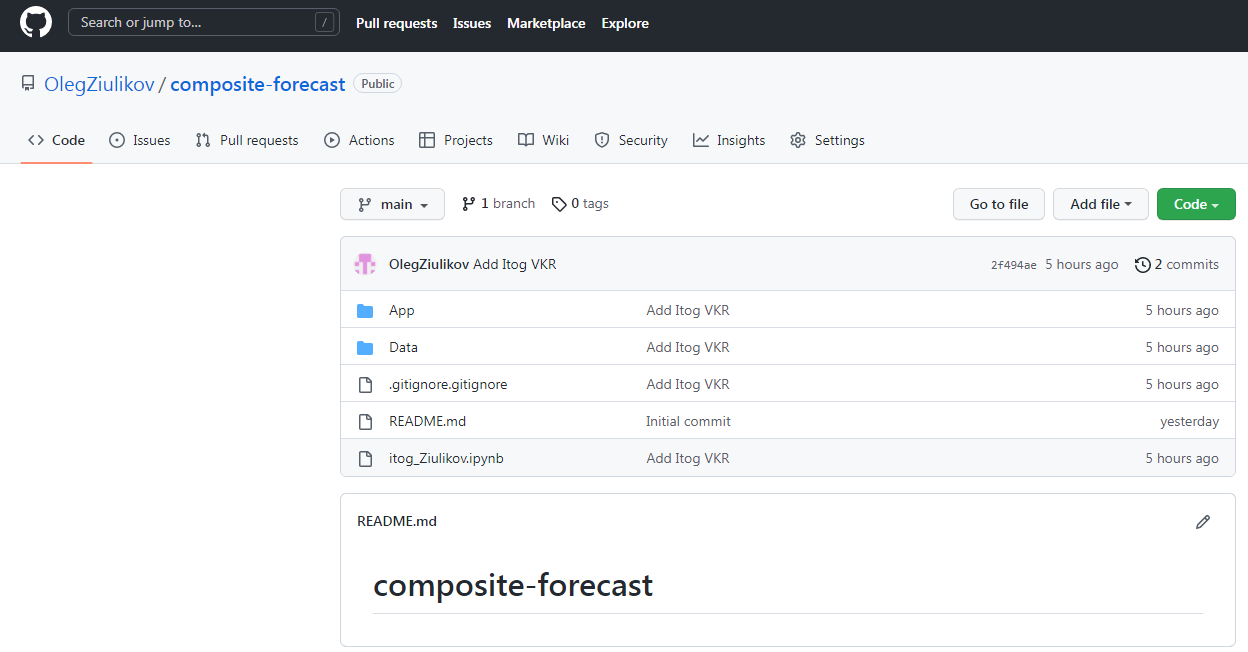


Рисунок 38 – скриншот созданного репозитория

* 1. **Заключение**

В работа был проведен анализ данных на основе объединения данных из двух файлов. Распределение полученных данных в объединённом датасете близко к нормальному, но коэффициенты корреляции между парами признаков стремятся к нулю. Поэтому применение моделей регрессии было не очень эффективно. Использованные при разработке моделей подходы не позволили получить сколько-нибудь достоверных прогнозов. Лучшей моделью для предсказания модуля упругости при растяжении и прочности при растяжении стала линейная регрессия.

Соотношение матрица – наполнитель тоже не удалось предсказать с высокой точностью. На тренировочной выборке модель показывала хорошие результаты, но на тестовой выборке результат заметно падал.

На основе полученных данных, можно сделать вывод, что для более точного предсказания нужны дополнительные данные, возможно синтезированные из уже имеющихся путем каких-либо математических преобразований. А также изучение экспертами взаимосвязи различных компонентов и получение из них дополнительного набора данных

* 1. **Список используемой литературы**

1. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.
2. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение.
3. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.