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: Черкасова Ольга Олеговна

Москва, 2023

**Оглавление**

[Введение 3](#_Toc132583205)

[1 Аналитическая часть 5](#_Toc132583206)

[1.1 Постановка задачи 5](#_Toc132583207)

[1.2 Описание используемых методов 5](#_Toc132583208)

[1.2.1 Методы регрессии 6](#_Toc132583209)

[1.2.2 Нейронные сети 6](#_Toc132583210)

[1.3 Разведочный анализ данных 8](#_Toc132583211)

[2 Практическая часть 10](#_Toc132583212)

[2.1 Предобработка данных 10](#_Toc132583213)

[2.2 Разработка и обучение модели 13](#_Toc132583214)

[2.3 Тестирование модели 17](#_Toc132583215)

[2.4 Написать нейронную сеть 18](#_Toc132583216)

[2.4.1 Нейронная сеть для прогноза упругости 18](#_Toc132583217)

[2.4.2 Нейронная сеть для прогноза прочности 20](#_Toc132583218)

[2.4.3 Нейронная сеть для прогноза прочности и упругости 21](#_Toc132583219)

[2.5 Разработка приложения 23](#_Toc132583220)

[2.6 Создание удаленного репозитория и загрузка работы на него 24](#_Toc132583221)

[Заключение 25](#_Toc132583222)

[Библиографический список 26](#_Toc132583223)

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

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

Датасет со свойствами композитов. Объединение делать по индексу тип объединения INNER

<https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw_SOi-vyNf2/view?usp=sharing>

1 Аналитическая часть

1.1 Постановка задачи

Целью работы является разработка прогнозных моделей для определения характеристик композитных материалов на основе данных о свойствах входящих компонентов. Для достижения этой цели необходимо решить следующие задачи:

- собрать и проанализировать данные о начальных свойствах компонентов композитных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.);

- создать представительный элемент объема композита и провести симуляцию для прогнозирования конечных свойств материала;

- сравнить результаты прогнозирования с физическими испытаниями образцов материалов для оценки точности модели;

- использовать полученные прогнозные модели для оптимизации производства и улучшения характеристик композитов.

Развитие прогнозных моделей для композитных материалов позволит сократить затраты на разработку новых материалов, улучшить их характеристики и оптимизировать производственный процесс.

На представленном датасете мы имеем 11 входных и 2 выходные величины. Размер выборка датасета небольшая, всего 1023 значения. Аномальных выбросов не наблюдается, но есть много повторений данных, из чего можно сделать вывод о его происхождении. Пропуски в датасете отсутствуют, что отлично, ведь на маленьком датасете они могут сильно повлиять на качество.

1.2 Описание используемых методов

По техническому заданию нам необходимо обучить модель и написать нейронную сеть. Так как нам необходимо решить задачу регрессии, то и модели будем использовать соответствующие.

### 1.2.1 Методы регрессии

Методы регрессии — это класс методов машинного обучения, который используется для построения моделей, которые могут предсказывать значение некоторой целевой переменной на основе значений одного или нескольких предикторов (признаков). Цель регрессионного анализа заключается в том, чтобы определить отношения между предикторами и целевой переменной и использовать эти отношения для прогнозирования будущих значений целевой переменной.

Наиболее распространенными методами регрессии являются:

1. Линейная регрессия — это метод регрессии, при котором предполагается, что между предикторами и целевой переменной существует линейная зависимость. Данная модель позволяет найти такие значения коэффициентов, при которых сумма квадратов разностей между значениями целевой переменной и значениями, предсказанными моделью, минимальна.

2. Логистическая регрессия — это метод регрессии, который используется для прогнозирования бинарных результатов. Она рассматривает логарифм шансов вместо прямого прогнозирования бинарного значения. Таким образом, данная модель может использоваться для категоризации наблюдений.

3. Регрессия решающих деревьев — это метод, который основан на построении дерева, в котором каждый узел ветвления представляет один из предикторов, а каждый лист дерева - прогноз для целевой переменной. Дерево строится путем разделения данных на группы на основе значения каждого предиктора.

Кроме того, существуют и другие методы регрессии, такие как полиномиальная регрессия, регрессия с обратным распределением Гаусса и т.д., которые могут применяться в зависимости от задачи и условий ее решения.

### 1.2.2 Нейронные сети

Нейронная сеть — это компьютерная модель, которая может имитировать работу мозга человека. Она состоит из большого количества связанных друг с другом элементов - нейронов. Нейроны представляют собой функциональные элементы, которые принимают входные данные, обрабатывают их и выдают результат на выход.

Нейронные сети широко используются для решения задач различного типа, в том числе для решения задач регрессии. Решение регрессионной задачи с помощью нейронных сетей основывается на обучении сети на основе имеющихся данных, после чего сеть может использоваться для прогнозирования значений целевой переменной на основе значений предикторов.

Процесс обучения нейронной сети заключается в том, что на вход сети подаются обучающие данные, которые включают значения предикторов и соответствующие значения целевой переменной. На основе этих данных нейронная сеть рассчитывает веса связей между ее элементами, чтобы минимизировать ошибку прогнозирования.

После обучения нейросеть может использоваться для прогнозирования значений целевой переменной на основе значений предикторов. Для этого на вход сети подаются новые данные, которые сеть обрабатывает и выдает результат на выход.

Нейронные сети могут использоваться для решения задач различного уровня сложности, но их применение в задачах регрессии особенно популярно благодаря их способности моделировать сложные нелинейные зависимости между переменными. Это позволяет создавать более точные прогнозные модели, чем традиционные методы регрессии.

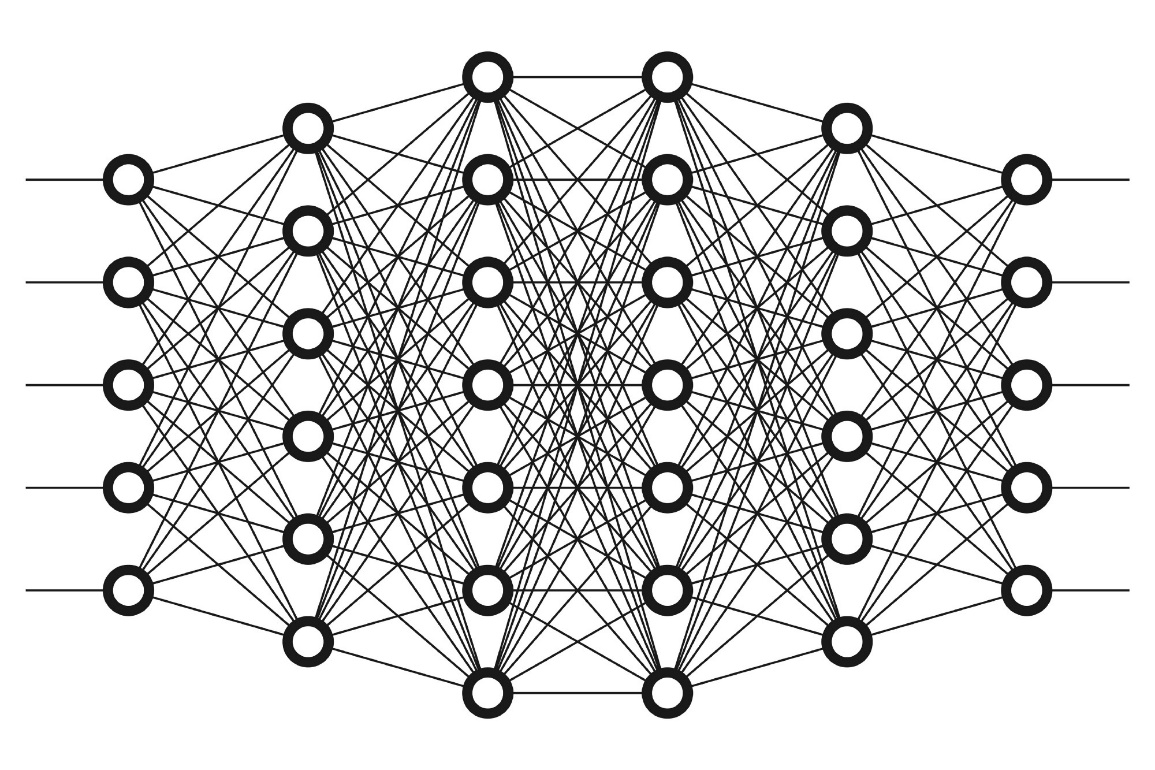


Рисунок 1.1 – Иллюстрация нейронной сети

1.3 Разведочный анализ данных

1. Рассмотрение статистических характеристик. Это метод, который используется для описания и анализа данных. Он включает в себя подсчет таких статистических характеристик, как среднее значение, медиана, мода, стандартное отклонение, дисперсия и корреляция. Этот метод помогает исследователям понять, какие свойства присутствуют в данных, и как они распределены.

2. Построение гистограмм. Этот метод используется для визуализации данных и построения графического представления распределения вероятностей. Гистограмма представляет собой столбчатую диаграмму, где по оси X отображаются значения переменной, а по оси Y - частота появления этих значений в данных. Гистограммы позволяют быстро определить, какие значения в данных наиболее часто встречаются, как они распределены и какие экстремальные значения могут быть в данных.

3. Корреляционный анализ. Этот метод используется для изучения связи между двумя или более переменными. Он позволяет оценить силу и направление взаимосвязи между переменными. Обычно для этого используются показатели корреляции, такие как корреляционный коэффициент Пирсона или коэффициент ранговой корреляции Спирмена. Корреляционный анализ не дает причинно-следственных связей между переменными, но помогает определить, насколько две переменные движутся вместе и могут ли они быть связаны между собой.

2 Практическая часть

2.1 Предобработка данных

Рассмотрев данные, аномалий не выявилось, так что перешли к визуализации данных. Например, рассмотрим распределение величин выборки для упругости:



Рисунок 2.1 – Диаграммы разброса для упругости

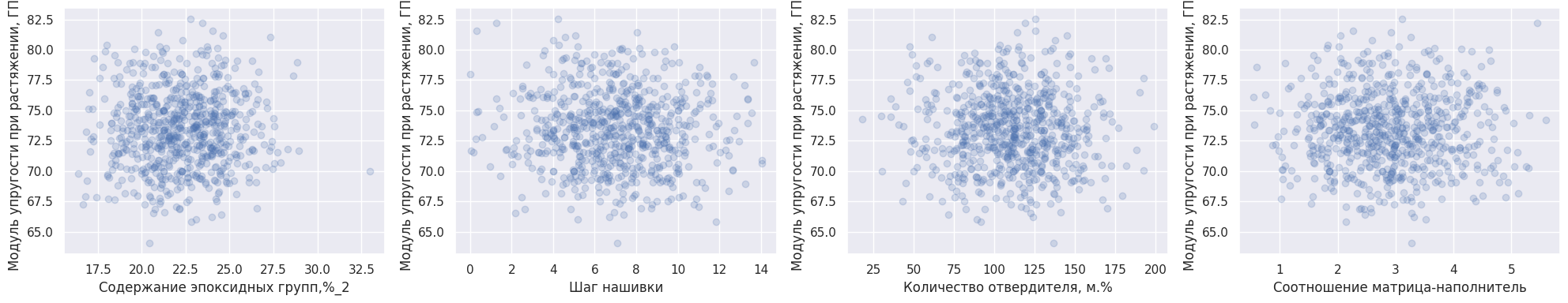
Рисунок 2.2 – Диаграммы разброса для упругости



Рисунок 2.3 – Диаграммы разброса для упругости

Диаграммы разброса для прочности:

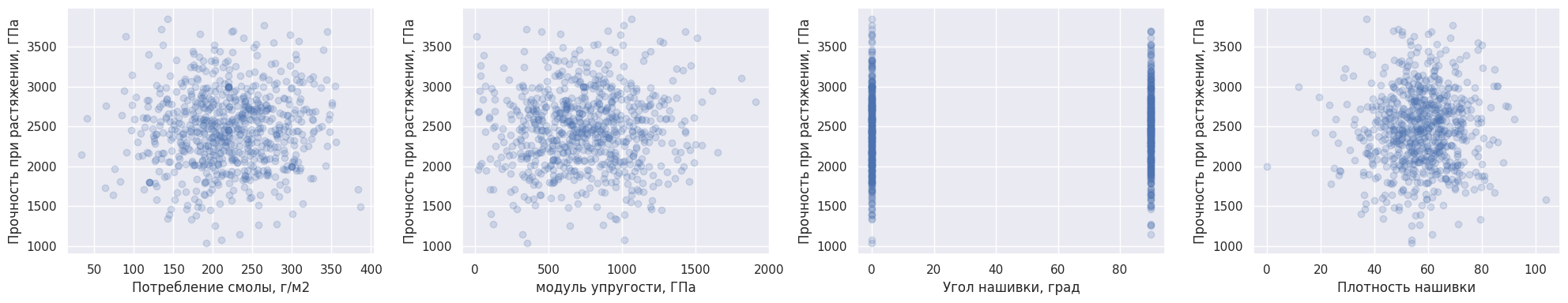


Рисунок 2.4 – Диаграммы разброса для прочности

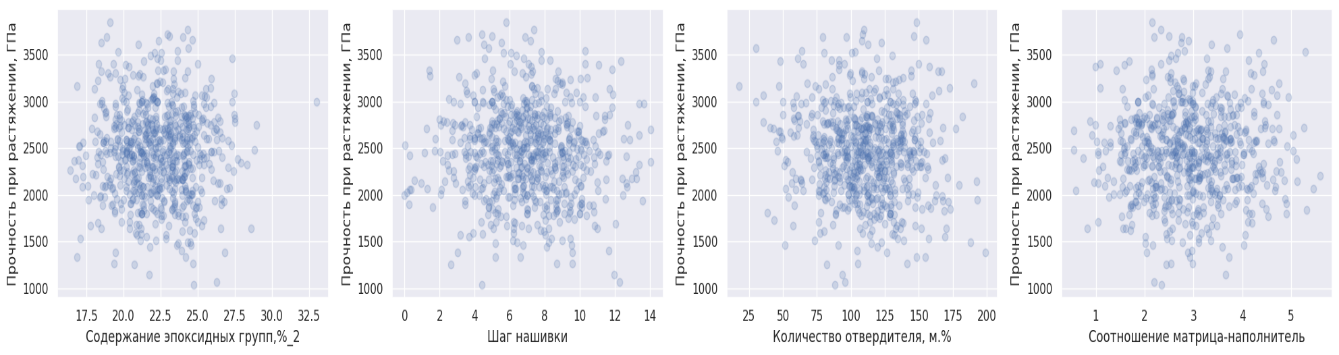


Рисунок 2.5 – Диаграммы разброса для прочности

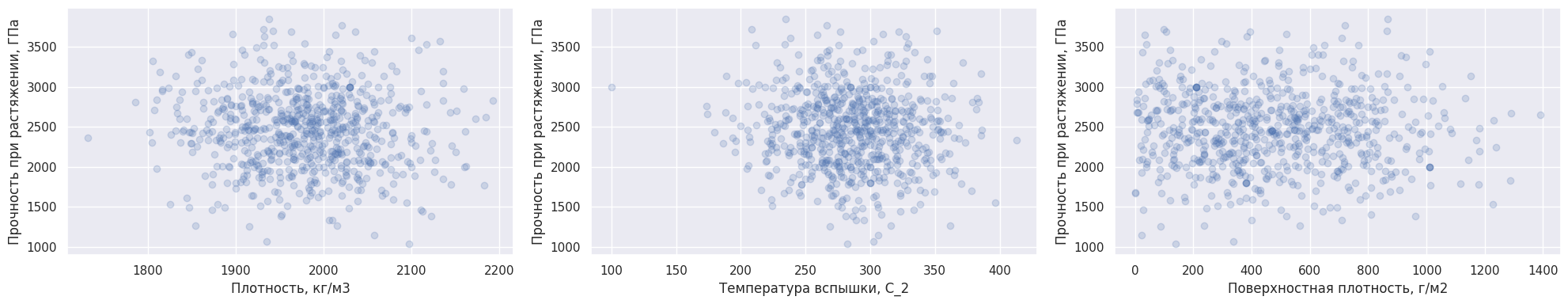


Рисунок 2.6 – Диаграммы разброса для прочности

Как видим разброс величин большой, по центру у некоторых параметров существуют скопления. Высокий разброс в анализе данных говорит о том, что значения переменной в данных сильно расходятся между собой. Он может быть вызван тем, что данные являются неоднородными или несбалансированными, а также возможно наличие определенных выбросов в данных.

Высокий разброс может говорить о том, что данными нужно более глубоко исследовать, чтобы выявить, какие переменные могут влиять на этот разброс. Это может помочь исключить возможные ошибки в данных, выделить важные переменные или атрибуты и провести дополнительные анализы.

Также, при высоком разбросе в данных, может быть сложно выделить общую тенденцию или основные паттерны, особенно если данные имеют выбросы или большое количество значений, которые существенно отличаются от среднего значения.

Проанализируем общую корреляцию со всеми параметрами: Корреляция для упругости:

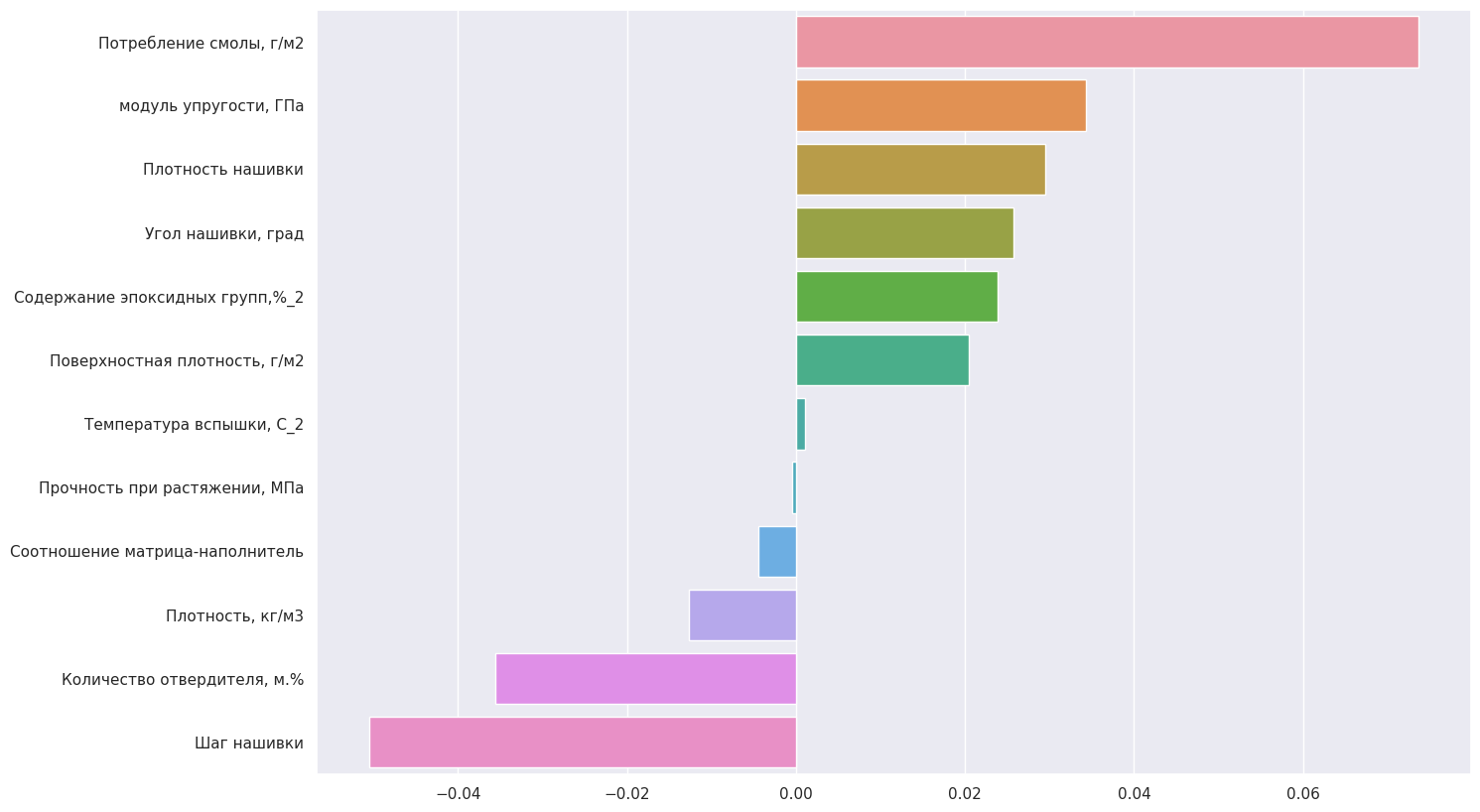


Рисунок 2.7 – Корреляция с модулем упругости при растяжении

Корреляция для прочности:

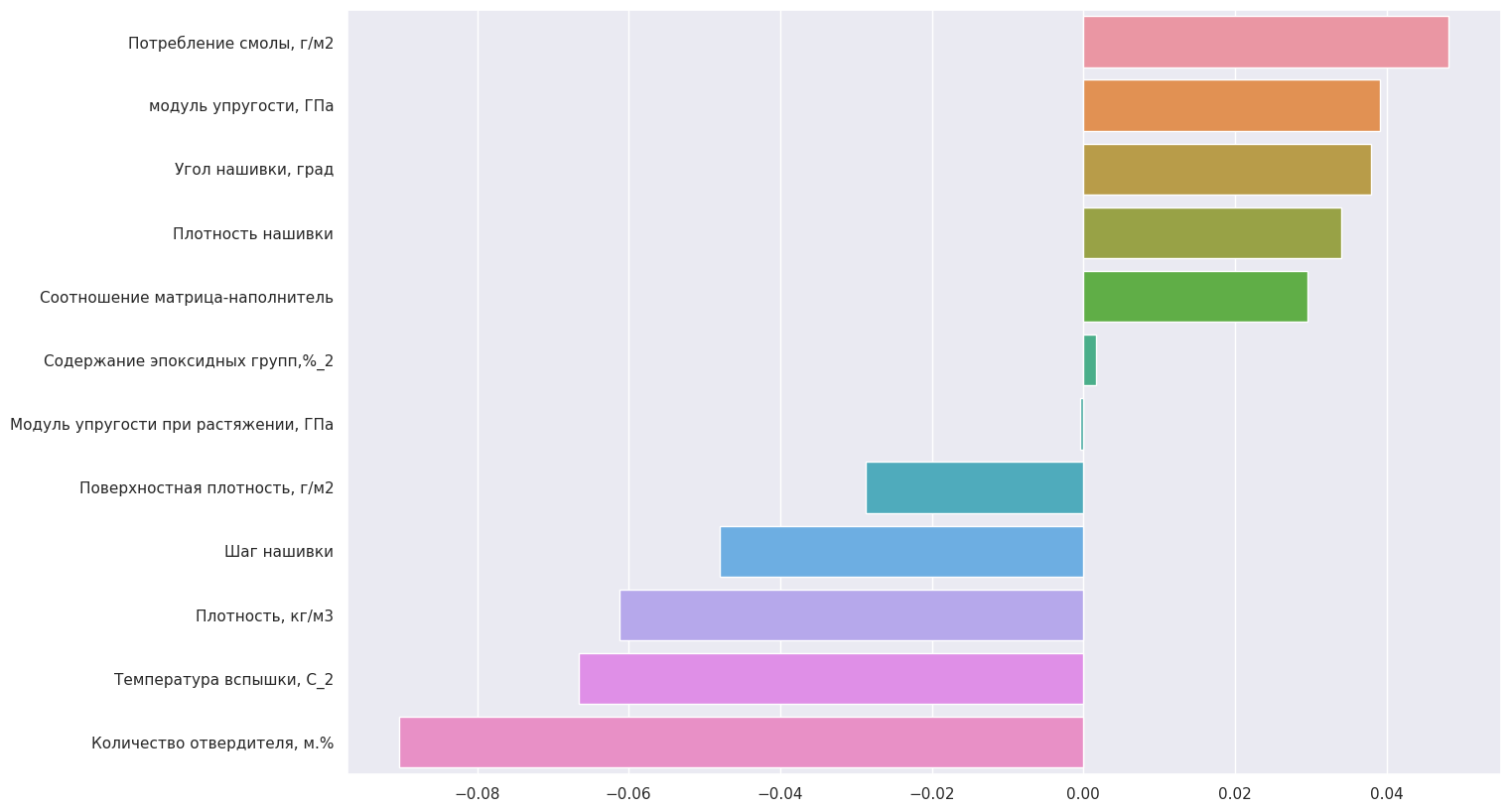


Рисунок 2.8 – Корреляция с прочностью при растяжении

Как видим, корреляция очень слабая, это означает что связь между параметрами мала или, возможно, она имеет нелинейный характер.

2.2 Разработка и обучение модели

Регрессионная задача заключается в построении модели на основе данных, для предсказания значения непрерывной переменной по заданным признакам. Для решения такой задачи используются различные алгоритмы, такие как Lasso, LinearRegression, Ridge и ElasticNet.

Для начала, рассмотрим модель линейной регрессии(Рисунок 1.1), которая представляет собой простую модель, на основе которой вычисляются коэффициенты уравнения линейной зависимости между переменными. Данный метод позволяет с достаточной точностью проводить прогнозирование по значению одной или нескольких переменных. Стоит отметить, что линейная регрессия является одним из самых простых и распространенных методов анализа данных, что делает ее доступной для использования в различных областях, таких как бизнес, финансы, экономика, медицина и др. Кроме того, понимание линейной регрессии может использоваться для построения более сложных моделей, таких как множественная регрессия и логистическая регрессия.

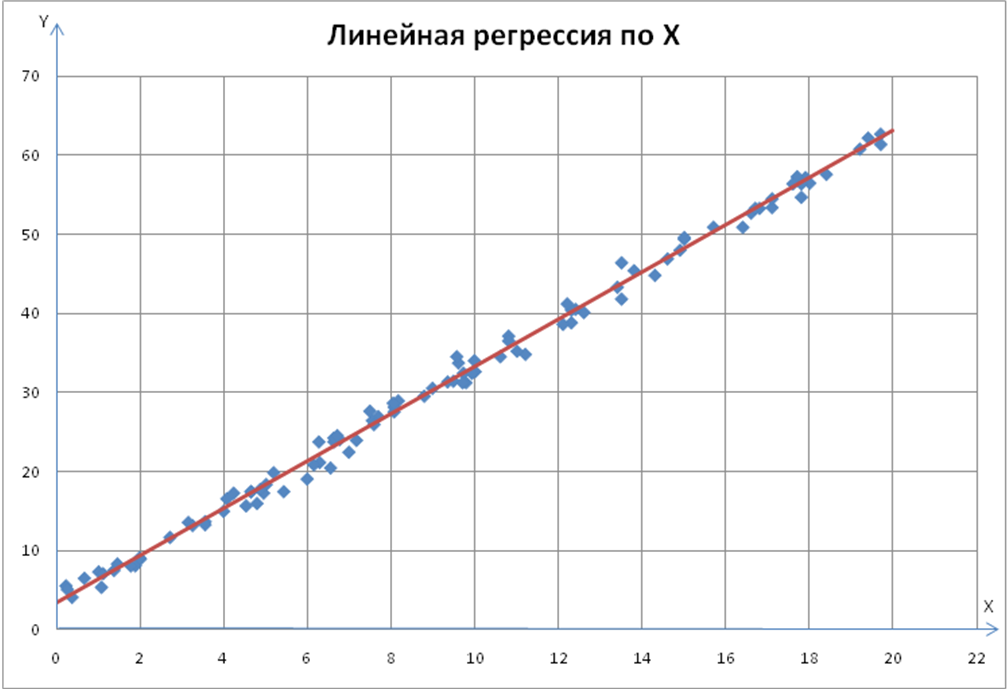


Рисунок 2.9 – Пример линейной регрессии на прямой.

Lasso является методом, решающим проблему переобучения путем регуляризации коэффициентов модели. Таким образом, этот метод выбирает наиболее информативные признаки и убирает незначимые, что позволяет уменьшить размерность пространства признаков и избежать переобучения. Алгоритм Lasso работает следующим образом:

1. Определяется коэффициент регрессии для каждой независимой переменной.

2. Оценивается качество модели, используя кросс-валидацию.

3. Применяется штраф на коэффициенты регрессии, который имеет формулу: штраф = λ \* коэффициент регрессии.

4. Штрафы суммируются вместе с оценками качества модели для определения оптимального значения λ.

5. Независимые переменные, у которых абсолютное значение коэффициента регрессии меньше, чем пороговое значение, выбрасываются из модели.

Применение Lasso позволяет уменьшить размерность данных, подчистив и избавившись от "лишних" независимых переменных. Это упрощает модель и помогает улучшить ее точность. Lasso может использоваться в различных областях, таких как биоинформатика, обработка изображений, климатические модели, экономика и многих других.

Ridge - это также метод регрессии, который использует регуляризацию для снижения весов коэффициентов модели, чтобы избежать переобученияАлгоритм Ridge (гребневая регрессия) - это метод регрессии, который также используется для сокращения количества независимых переменных в модели, но в отличие от Lasso, уменьшает весовые коэффициенты, а не исключает переменные из модели.

Алгоритм Ridge работает следующим образом:

1. Определяется коэффициент регрессии для каждой независимой переменной.

2. Оценивается качество модели, используя кросс-валидацию.

3. Применяется штраф на коэффициенты регрессии, который имеет формулу: штраф = λ \* сумма квадратов коэффициентов регрессии.

4. Штрафы суммируются вместе с оценками качества модели для определения оптимального значения λ.

5. Меньшие значения параметра λ приводят к более точным моделям (без штрафа), а большие значения допускают использование менее важных переменных (с учетом штрафа).

Применение алгоритма Ridge снижает величину коэффициентов регрессии, что снижает влияние шумовых переменных на модель и позволяет сделать более устойчивую регрессионную модель. Обычно алгоритм Ridge применяется в тех случаях, когда необходимо снизить влияние шумовых переменных, уменьшить размерность данных и получить более точные прогнозы. В отличие от Lasso, алгоритм Ridge сохраняет все переменные в модели и уменьшает значение их коэффициентов.

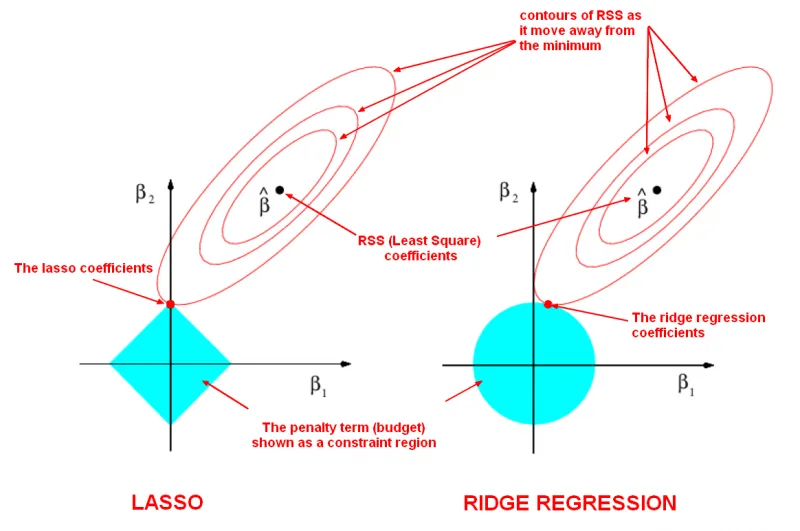


Рисунок 2.10 – Пример различия алгоритмов Лассо и Ridge

Алгоритм ElasticNet - это комбинация алгоритмов Ridge и Lasso, который используется для устранения нескольких проблем Lasso и Ridge. В отличие от Ridge и Lasso, ElasticNet способен справляться с сильно скоррелированными признаками и бороться с небольшим количеством признаков в обучающем наборе.

Алгоритм ElasticNet работает следующим образом:

1. Определяется коэффициент регрессии для каждой независимой переменной.

2. Оценивается качество модели, используя кросс-валидацию.

3. Применяются штрафы на коэффициенты регрессии, как в Lasso и Ridge.

4. Происходит настройка двух параметров регуляризации - α (коэффициент уравнения регрессии) и λ (сила регуляризации).

5. Значения коэффициентов регрессии обновляются на основе новых штрафов.

6. Повторяется до сходимости.

Применение алгоритма ElasticNet позволяет получить оптимальную регрессионную модель, которая сочетает преимущества алгоритмов Lasso и Ridge. ElasticNet может использоваться в различных областях, таких как биоинформатика, финансы, климатические модели, медицина и многих других.

Каждый из описанных методов имеет свои преимущества и недостатки, и выбор метода зависит от конкретной задачи и данных, на которых он будет применен. LinearRegression - хороший метод для данных с небольшим количеством признаков, Lasso - для выбора наиболее значимых признаков, Ridge - для уменьшения влияния шумовых признаков и ElasticNet - для работы с мультиколлинеарностью в данных.

2.3 Тестирование модели

Обучив выборук на всех данных алгоритмах мы получили следующие значения средних ошибок на тестирующей выборке для модуля упргости при растяжении(Рисунок 1.3) и прочности при растяжении(Рисунок 1.4).

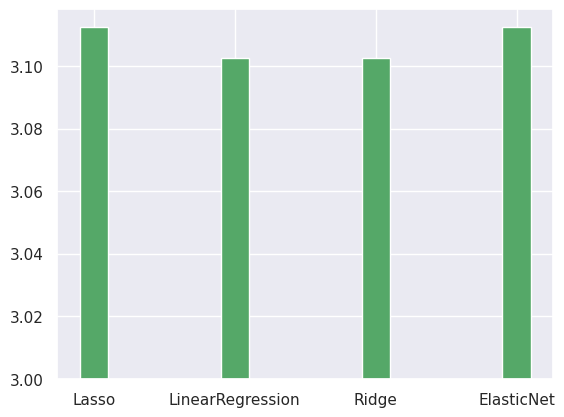


Рисунок 2.11 - Гистограмма средней ошибки для упругости

Разница между алгоритмами не сильно большо,но как видим лучше всего себя показали алгоритмы линейной регрессии и Ridge.

Ridge показало среднюю ошибку: 3.10252300597454. Это примерно 4% от среднего значения. Неплохой результат.

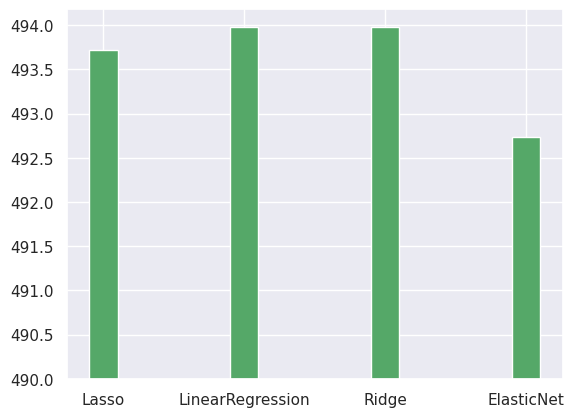


Рисунок 2.12 – Гистограмма средней ошибки прочности

Тут уже ElasticNet, показал себя лучше всех. Средняя ошибка: 492.7320465803211. Это примерно 20% от средней прочности при растяжении.

2.4 Написать нейронную сеть, рекомендующую соотношение матрица

Для сравнения напишем 3 нейронные сети:

1. Будет прогнозировать значение упругости по входным данным
2. Будет прогнозировать значение прочности по входным данным
3. Будет прогнозировать значение прочности и упругости по входным данным

### 2.4.1 Нейронная сеть для прогноза упругости

Первая нейронная сеть будет иметь по сути два внутренних полносвязных слоя по 8 нейронов + 1 выходной:

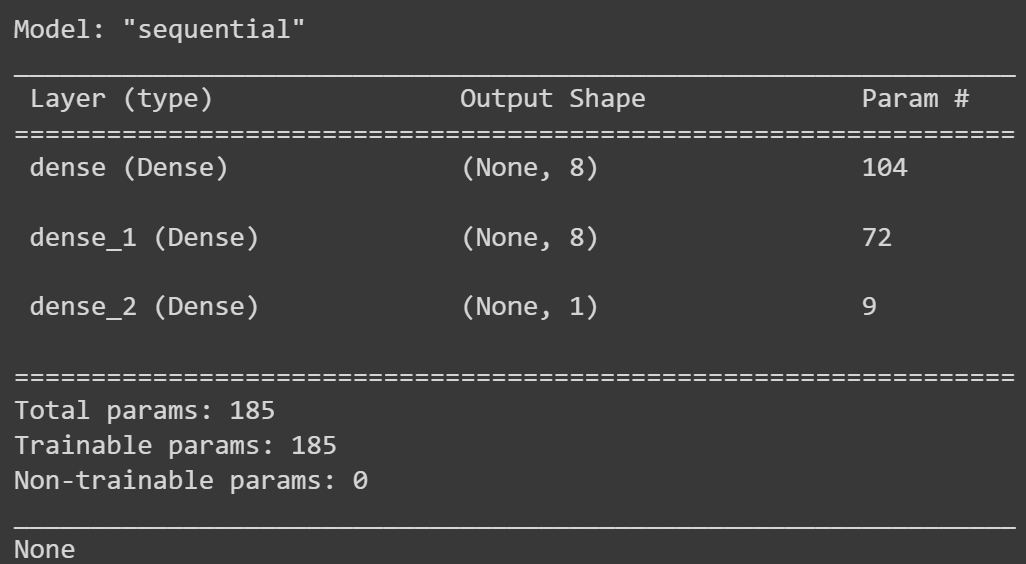


Рисунок 2.13 – Схема 1 нейронной сети по слоям

Обучим данную нейросеть на 10000 эпох, критерием будет ошибка. Такое количество эпох выбрано для анализа прогресса нейронной сети.

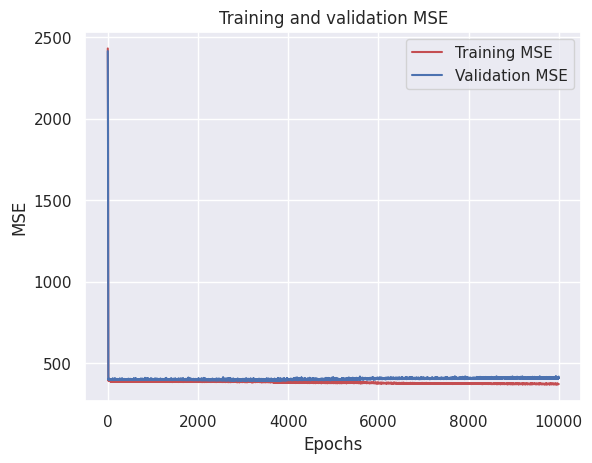


Рисунок 2.14 – График ошибки обучения нейронной сети по эпохам

Как видим нейронная сеть начинает деградировать после 2000-3000 эпох, наблюдаем эффект переобучения. Получили значение ошибки: 525, что больше, чем у модели машинного обучения. После переобучения на 2000 эпохах получили ошибку 519, что меньше, чем было, но все еще больше, чем у модели.

### 2.4.2 Нейронная сеть для прогноза прочности

Данная нейронная сеть будет иметь такую же структуру, как и предыдущая, изменяться лишь данные для обучения.

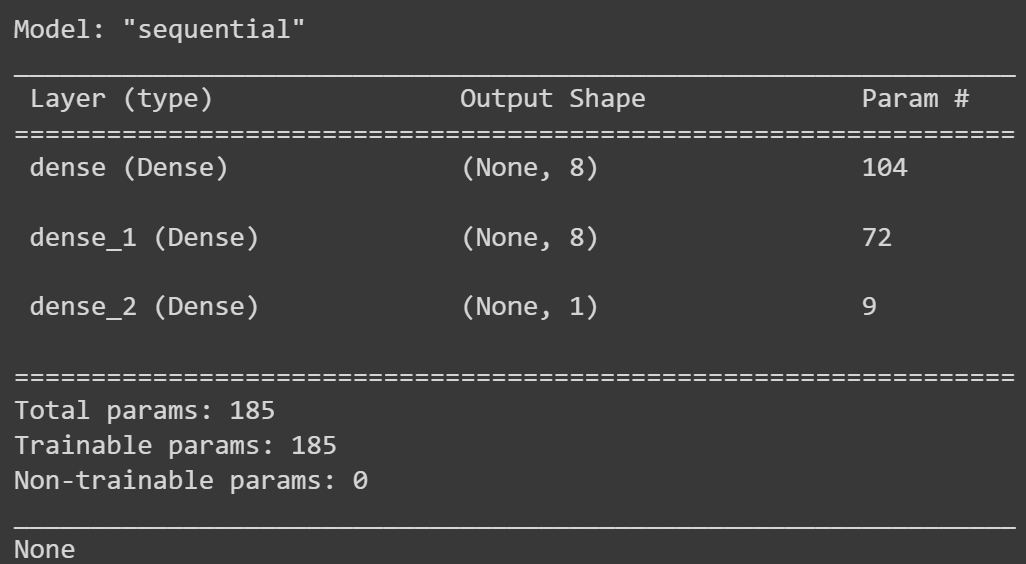


Рисунок 2.15 – Схема 2 нейронной сети по слоям

Обучим данную нейросеть на 10000 эпох, критерием будет ошибка.

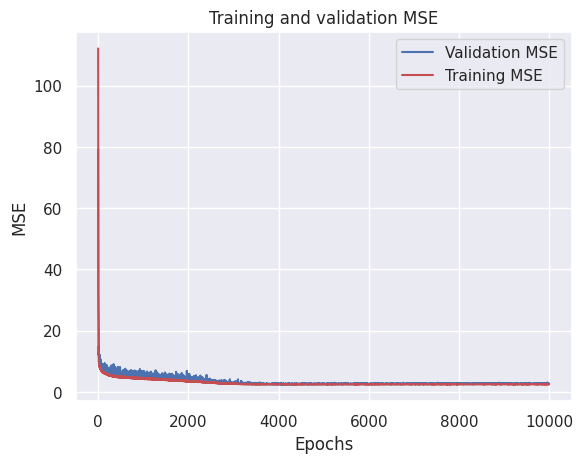


Рисунок 2.16 – График ошибки обучения нейронной сети по эпохам

Как видим данная нейронная сеть успешно обучается, не деградирует. Но смысла обучения после 4000 эпох не имеет. Средняя ошибка: 3.2, что очень близко к алгоритму Ridge машинного обучения.

### 2.4.3 Нейронная сеть для прогноза прочности и упругости

Данная нейронная сеть тоже похожа на предыдущие, за исключением последнего выходного слоя, он имеет 2 нейрона так как предсказываем два значения.

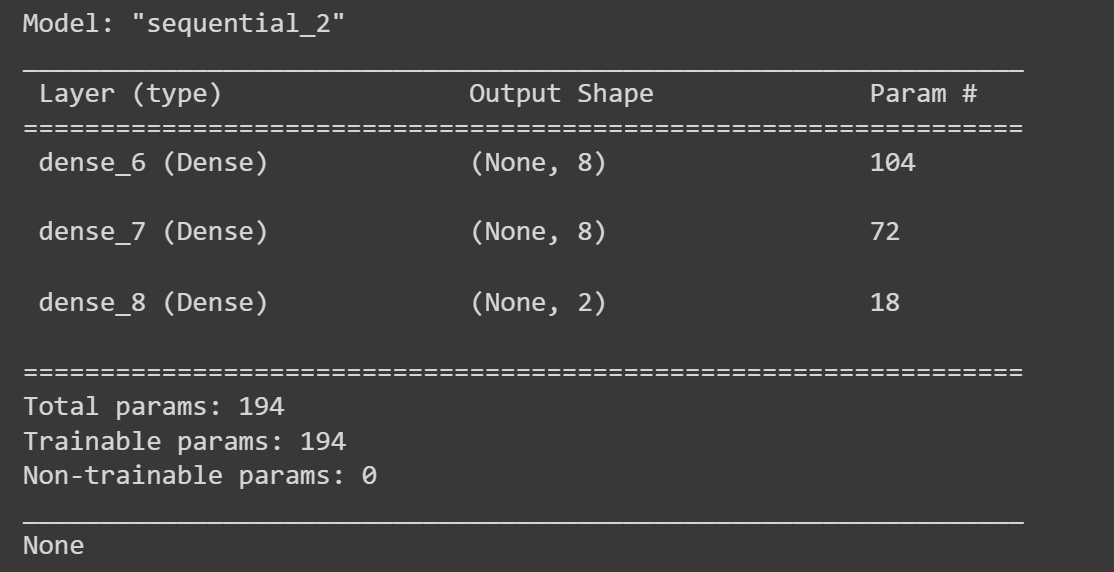


Рисунок 2.17 – Схема 3 нейронной сети по слоям

Обучим данную нейросеть на 10000 эпох, так же критерием будет ошибка.

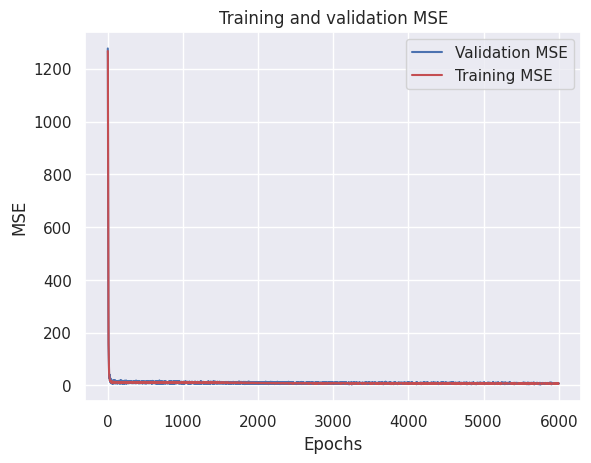


Рисунок 2.18 – График ошибки обучения нейронной сети по эпохам

Как видим данная нейронная сеть успешно обучается, не деградирует. Но смысла обучения после 1500 эпох не имеет. Средняя ошибка по эластичности: 7, что хуже, чем у отдельной нейросети. Средняя ошибка по прочности около 10. Что сильно лучше, чем у отдельной нейросети и модели машинного обучения.

2.5 Разработка приложения

Так как код написан на python, то приложение будет создано с помощью утилиты pyinstaller. Данная утилита поможет создать консольное приложение с помощью сохранённых моделей. Модели уже загружены внутрь приложения. Исходя из предыдущей главы, были выбраны две нейросети - нейронная сеть для прогноза прочности и упругости, и нейронная сеть для прогноза упругости. По первой будем прогнозировать прочность, а по второй упругость.

Консольное приложение имеет меню:

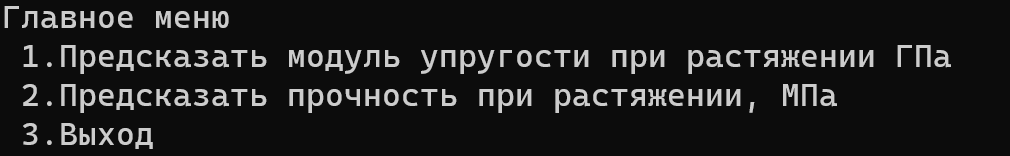


Рисунок 2.19 – Главное меню приложения

С помощью выбора команды можно спрогнозировать оба параметра:

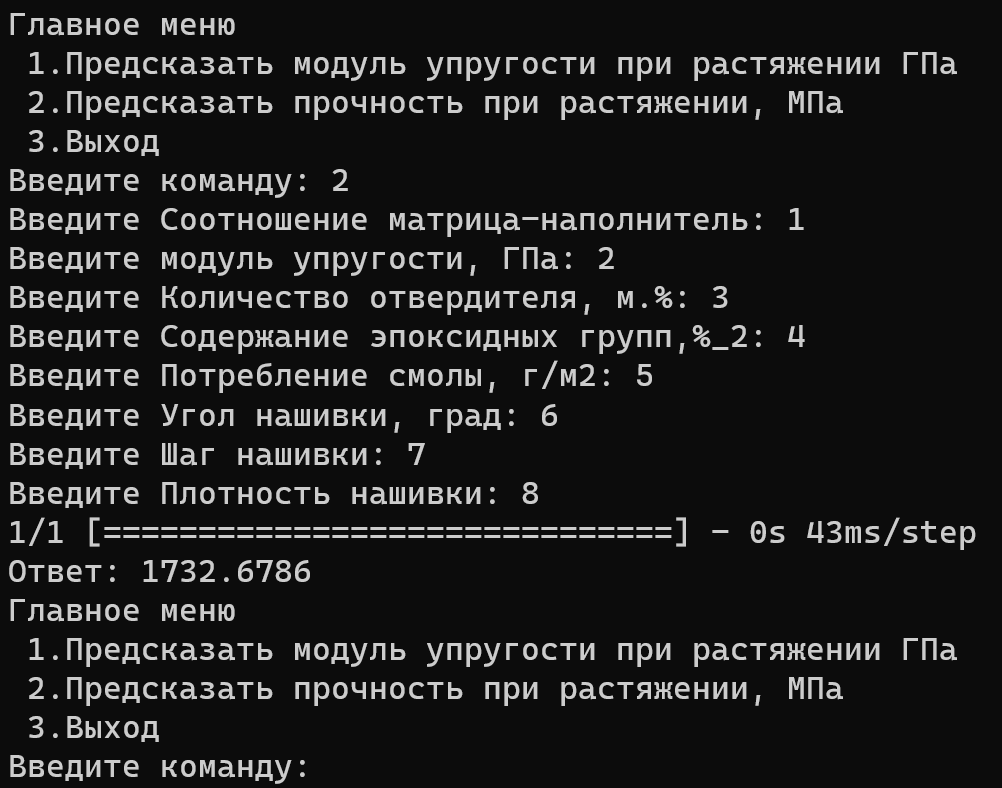


Рисунок 2.20 – Пример прогноза прочности в приложении

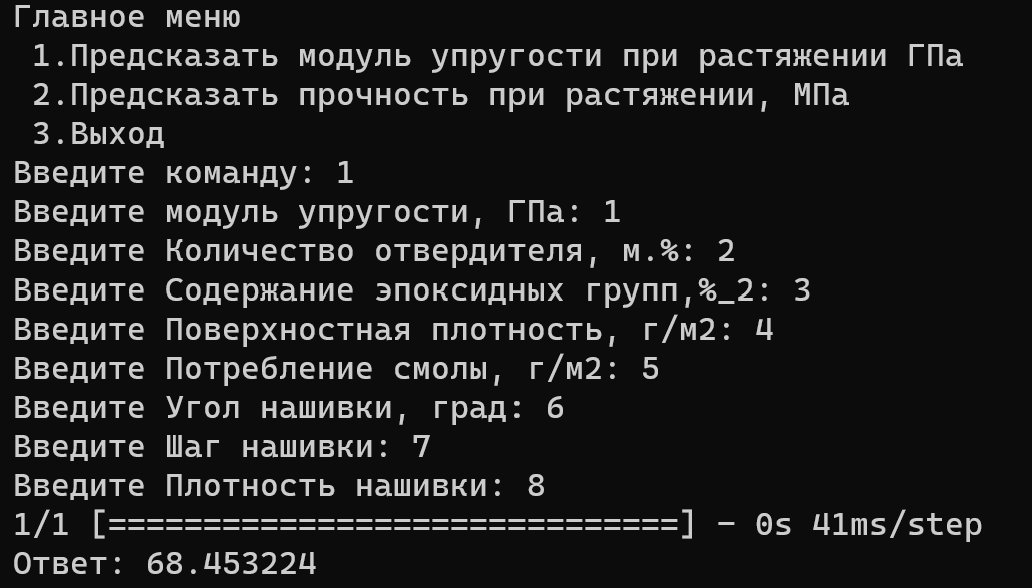


Рисунок 2.21 – Пример прогноза упругости в приложении

2.6 Создание удаленного репозитория и загрузка работы на него

Для доступа к исходному коду, будет использован публичный репозиторий github. Была создана страничка слушателя: <https://github.com/Olechka8513>. На нем был создан репозиторий: <https://github.com/Olechka8513/composits>. И два коммита: Инициализирующий коммит и коммит с добавлением всех файлов. Также ссылка на google collab: https://colab.research.google.com/drive/1qyXl8W7y9Z8qKPKrcjtxcVPy5l0BbgiX?usp=sharing#scrollTo=g-f5XP0aq8An

Заключение

В результате данной работы была проведена аналитическая работа по анализу данных образцов композитов. Были проанализированы свойства данных, изучены статистические показатели и распределения значений, построены графики корреляции.

Для прогнозирования прочности и упругости были применены различные методы машинного обучения: LinearRegression, Lasso, Ridge и ElasticNet. Каждый метод имел свои преимущества и недостатки, и был выбран на основе поставленной задачи.

Также были разработаны нейронные сети для предсказания прочности и упругости, а также для предсказания прочности и упругости одновременно. Нейронные сети были обучены и показали свою эффективность в сравнении с методами машинного обучения.

В результате было разработано консольное приложения для прогнозирования прочности и упругости композитов на основе моделей машинного обучения и нейронных сетей. Все материалы загружены на удаленный публичный репозиторий.

В дальнейшем возможно улучшение точности прогнозов путем использования других методов машинного обучения или изменения структуры нейронной сети.

Библиографический список

1. Habr.com. Библиотеки для глубокого обучения: Keras. [Электронный ресурс] : – Режим доступа: https://habr.com/ru/companies/ods/articles/325432/. (дата обращения: 10.04.2023).
2. Scikit Learn. Linear Models [Электронный ресурс]: – Режим доступа: https://scikit-learn.org/stable/modules/linear\_model.html. (дата обращения: 10.04.2023).
3. python-scripts.com. Машинное обучение для начинающих: создание нейронных сетей [Электронный ресурс]: – Режим доступа: https://python-scripts.com/intro-to-neural-networks. (дата обращения: 11.04.2023).