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 Волотова Юлия Викторовна

Москва, 2023

Содержание

Введение

Данная работа выполнена в рамках курса Data Science.

В качестве анализируемой задачи принята тема «Прогнозирование конечных свойств новых материалов (композиционных материалов).

1 Аналитическая часть

1.1 Постановка задачи

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов).

Достаточно известно определение, согласно которому: композиты – это материалы, состоящие из двух или более компонентов (армирующих элементов и скрепляющей их матрицы) и обладающие свойствами, отличными от суммарных свойств компонентов.

При этом предполагается, что компоненты, входящие в состав композита, должны быть хорошо совместимыми и не растворяться или иным способом поглощать друг друга.

В широком смысле композиционный материал – это любой материал с гетерогенной структурой, т. е. со структурой, состоящей минимум из двух фаз.

Такое определение позволяет отнести к композиционным материалам абсолютное большинство металлических материалов, поскольку они либо намеренно создаются многофазными, либо считаются однофазными, но в них есть неметаллические включения. Полимерные материалы также можно отнести к композитам, поскольку кроме основного компонента (полимера) в них присутствуют различные наполнители, красители и др. Материалы природного происхождения (кости человека и животных, древесина) также можно отнести к композиционным.

Например, древесина представляет собой композицию из пучков целлюлозных волокон трубчатого строения, скрепленных матрицей из органического вещества – лигнина.

Для того чтобы выделить композиционные материалы искусственного происхождения, подчеркнуть их характерные особенности наиболее полным считается определение, согласно которому к композитам относятся материалы, обладающие рядом признаков:

1. состав, форма и распределение компонентов материала «запроектированы заранее»;

2. материал не встречается в природе, а создан человеком;

3. материал состоит из двух или более компонентов, различающихся по химическому составу и разделенных выраженной границей;

4. свойства материала определяются каждым из его компонентов,

которые должны присутствовать в материале в достаточно больших количествах (больше некоторого критического содержания);

5. материал обладает такими свойствами, которых не имеют его

компоненты, взятые в отдельности;

6. материал неоднороден в микромасштабе и однороден в макромасштабе

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 известны 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).

В качестве входных данных приняты данные о начальных свойствах компонентов композиционных материалов:

- Соотношение матрица-наполнитель;
- Плотность;
- Модуль упругости;
- Количество отвердителя;
- Содержание эпоксидных групп;
- Температура вспышки;
- Поверхностная плотность;
- Модуль упругости при растяжении;
- Прочность при растяжении;
- Потребление смолы;
- Угол нашивки;
- Шаг нашивки;
- Плотность нашивки.

Общее количество параметров для анализа – 13.

Датасеты были объединены по индексу тип объединения INNER.

Итоговый датасет нормализованный, пропуски отсутствуют. Элементы массива соответствуют типу float64.

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.

1.2 Описание используемых методов

Для решения поставленной задачи выбраны методы:

– решение задачи регрессии для прогнозирования параметров: модуля упругости и прочности при растяжении. Для решения задачи регрессии использовались: линейная регрессия и случайный лес.

Оценка качества моделей указана на рисунке 1, где «МУ» – модель для параметра «Модуль упругости»; «ПР» - модель для параметра «Прочность при растяжении».

```
[452] scaler_LN_1 = preprocessing.MinMaxScaler()
      df_LN_1_1 = scaler_LN_1.fit_transform(np.array(df_new['Модуль упругости при растяжении, ГПа']).reshape(-1,1))

      scaler_LN_2 = preprocessing.MinMaxScaler()
      df_LN_2_1 = scaler_LN_2.fit_transform(np.array(df_new['Прочность при растяжении, МПа']).reshape(-1,1))
```

```
[453] # выведем метрики MAE, MSE и R2 в таблицу
      df_errors_LrLN_1 = errors(LN_1, X1_train, X1_test, y1_train, y1_test, name = LN_1, trg = LN_1, scaler = scaler_LN_1)
      df_errors_LrLN_2 = errors(LN_2, X2_train, X2_test, y2_train, y2_test, name = LN_2, trg = LN_2, scaler = scaler_LN_2)
      df_errors_Lr = pd.concat([df_errors_LrLN_1, df_errors_LrLN_2], axis=0).reset_index(drop = True)
      df_errors_Lr
```

	Модель	Целевой параметр	MAE(test)	MAE(train)	MSE(test)	MSE(train)	R2 score
0	LinearRegression()	LinearRegression()	0.840720	0.784548	1.035301	0.968276	-0.019872
1	LinearRegression()	LinearRegression()	0.802764	0.789657	0.987849	0.985416	-0.025589

Рисунок 1 - Оценка качества моделей

Средняя абсолютная ошибка (mean_absolute_error) для всех моделей находится около 0,84 (это средняя разница между фактическим значением данных и значением, предсказанным моделью).

Коэффициент детерминации (r2_score) принимает отрицательные значения для всех моделей.

1.3 Разведочный анализ данных

Для разведочного анализа данных использованы методы описательной статистики.

Датасет был проверен на наличие пропусков в значениях (команда df.isna().sum()). Пропусков обнаружено не было.

Команда Describe () позволила выявить наличие дискретной величины, принимающей значения 0 и 90 (параметр «Угол нашивки»), а также основные значения для всех параметров.

С помощью построения гистограмм было выявлено распределение величин, близкое к нормальному, для большей части параметров.

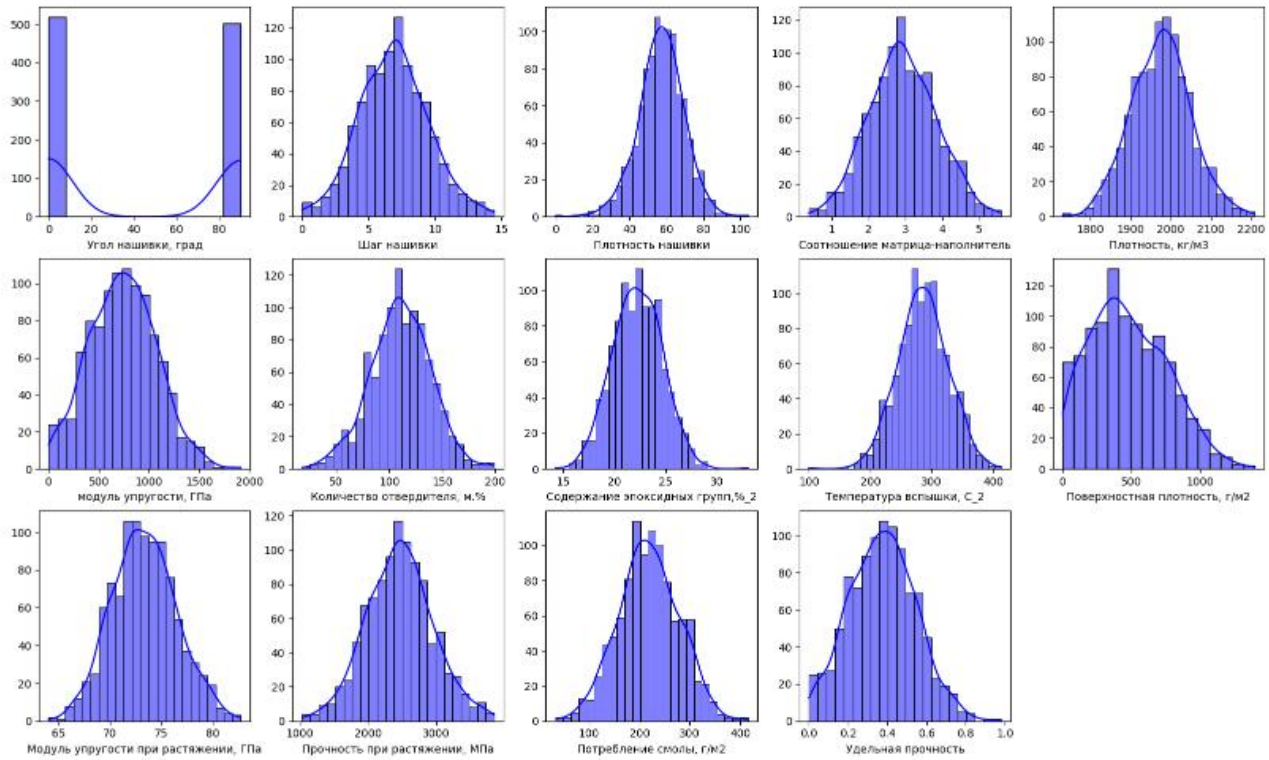

Рисунок 1. Гистограмма распределения параметров.

С помощью диаграммы «ящик с усами» для всех параметров были выявлены выбросы.

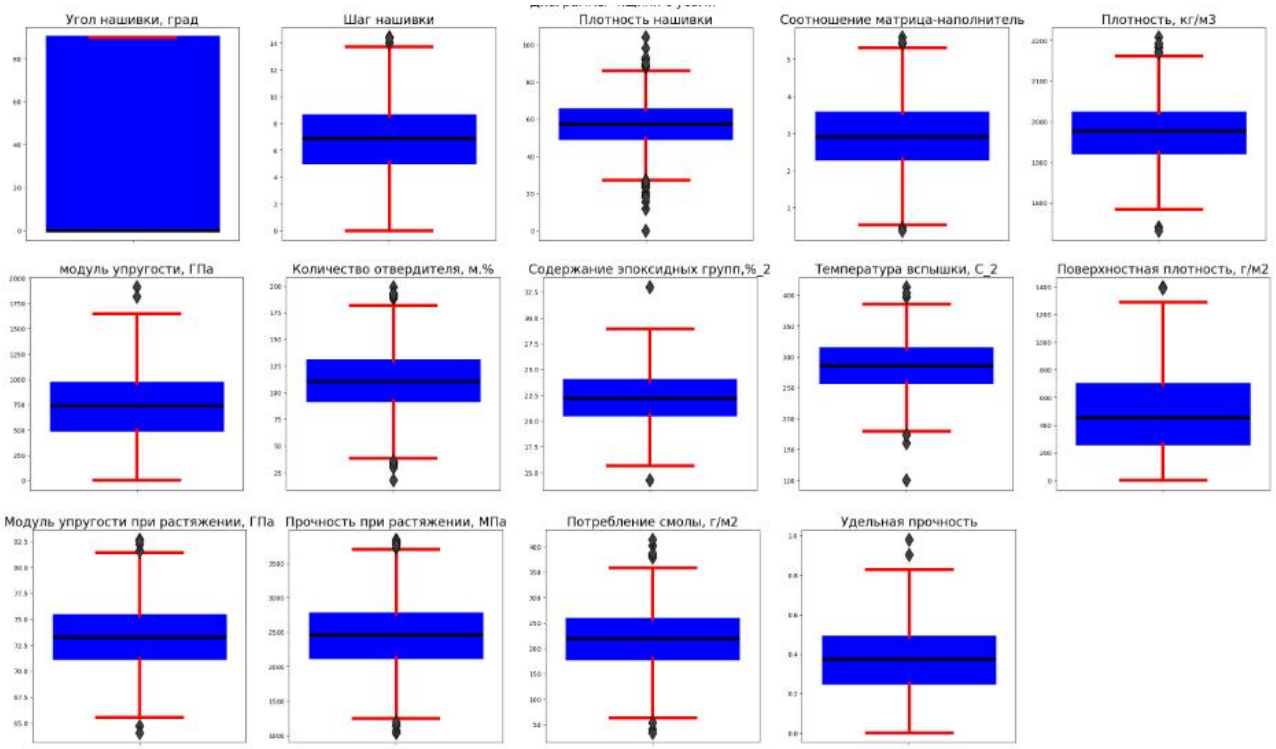

Рисунок 2 – Диаграмма распределения параметров

Выбросы наблюдаются по всем параметрам, кроме угла нашивки, т.к. данный параметр принимает дискретные значения и диаграмма «ящик с усами» для него не показательна.

Также следует отметить наличие выбросов с двух сторон (например, прочность при растяжении), наличие выбросов со стороны наименьших значения (например, содержание эпоксидных групп), наличие выбросов со стороны наибольших значений (например, шаг нашивки).

Для разведочного анализа было использовано также построение попарных графиков рассеяния точек. Графики для всех параметров показали отсутствие зависимости между переменными датасета.

На рисунке 3 приведена тепловая карта коэффициентов корреляции, значения которой показывают, что все полученные коэффициенты корреляции находятся в промежутке значений от - 0,11 до 0,11 (не считая удельную прочность, расчитанный показатель удельная прочность (модуль)=модуль упругости/плотность) .

	Угол нашивки, град	Шаг нашивки	Плотность нашивки	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Удельная прочность
Угол нашивки, град	1.000000	0.023816	0.107947	-0.031073	-0.068474	-0.025417	0.038570	0.008052	0.020695	0.052299	0.023003	0.023398	-0.015334	-0.021023
Шаг нашивки	0.023816	1.000000	0.003487	0.036437	-0.061015	-0.009875	0.014887	0.003022	0.025795	0.038332	-0.029468	-0.059547	0.013394	-0.004946
Плотность нашивки	0.107947	0.003487	1.000000	-0.004652	0.080304	0.056346	0.017248	-0.039073	0.011391	-0.049923	0.006476	0.019604	0.012239	0.047581
Соотношение матрица-наполнитель	-0.031073	0.036437	-0.004652	1.000000	0.003841	0.031700	-0.006445	0.019766	-0.004776	-0.006272	-0.008411	0.024148	0.072531	0.031866
Плотность, кг/м3	-0.068474	-0.061015	0.080304	0.003841	1.000000	-0.009847	-0.035911	-0.008278	-0.020695	0.044930	-0.017602	-0.069981	-0.015937	-0.094616
модуль упругости, ГПа	-0.025417	-0.009875	0.056346	0.031700	-0.009847	1.000000	0.024049	-0.006804	0.031174	-0.005306	0.023267	0.041868	0.001840	0.995693
Количество отвердителя, м.%	0.038570	0.014887	0.017248	-0.006445	-0.035911	0.024049	1.000000	-0.000684	0.095193	0.055198	-0.065929	-0.075375	0.007446	0.024870
Содержание эпоксидных групп,%_2	0.008052	0.003022	-0.039073	0.019766	-0.008278	-0.006804	-0.000684	1.000000	-0.009769	-0.012940	0.056828	-0.023899	0.015165	-0.005123
Температура вспышки, С_2	0.020695	0.025795	0.011391	-0.004776	-0.020695	0.031174	0.095193	-0.009769	1.000000	0.020121	0.028414	-0.031763	0.059954	0.033870
Поверхностная плотность, г/м2	0.052299	0.038332	-0.049923	-0.006272	0.044930	-0.005306	0.055198	-0.012940	0.020121	1.000000	0.036702	-0.003210	0.015692	-0.010099
Модуль упругости при растяжении, ГПа	0.023003	-0.029468	0.006476	-0.008411	-0.017602	0.023267	-0.065929	0.056828	0.028414	0.036702	1.000000	-0.009009	0.050938	0.023776
Прочность при растяжении, МПа	0.023398	-0.059547	0.019604	0.024148	-0.069981	0.041868	-0.075375	-0.023899	-0.031763	-0.003210	-0.009009	1.000000	0.028602	0.046957
Потребление смолы, г/м2	-0.015334	0.013394	0.012239	0.072531	-0.015937	0.001840	0.007446	0.015165	0.059954	0.015692	0.050938	0.028602	1.000000	0.003232
Удельная прочность	-0.021023	-0.004946	0.047581	0.031866	-0.094616	0.995693	0.024870	-0.005123	0.033870	-0.010099	0.023776	0.046957	0.003232	1.000000

Рисунок 3 – Тепловая карта коэффициентов корреляции

2 Практическая часть

2.1 Предобработка данных

Предобработка данных осуществлялась на основании разведочного анализа данных, который показал наличие выбросов. Для удаления выбросов был произведен расчет количества выбросов для каждого параметра, данные приведены на рисунке 4.

```
df.isnull().sum()

Угол нашивки, град                    0
Шаг нашивки                           4
Плотность нашивки                    21
Соотношение матрица-наполнитель       6
Плотность, кг/м3                      9
модуль упругости, ГПа                 2
Количество отвердителя, м.%           14
Содержание эпоксидных групп,%_2       2
Температура вспышки, С_2              8
Поверхностная плотность, г/м2         2
Модуль упругости при растяжении, ГПа  6
Прочность при растяжении, МПа         11
Потребление смолы, г/м2               8
Удельная прочность                    2
dtype: int64
```

Рисунок 4 – Расчет количества выбросов

Количество выбросов говорит о их незначительности, соответственно, данные значения можно удалить из датасета.

Размер датасета после удаления выбросов и проверка наличия пропусков:

– Количество строк в очищенном датасете: 936;
– Количество столбцов (переменных) в очищенном датасете: 14;
– Количество пропусков в данных очищенного датасета: 0.

Вызов команды df_new.describe() для датасета показывает, что количество строк датасета уменьшилось. Датасет очищен от выбросов.

Для выделения наиболее весомых признаков датасета использован факторный анализ.

С помощью метода главных компонент были получены значения влияния факторов (данные приведены на рисунке 5). Указанные значения позволяют сделать вывод о том, что влияние очень слабое.

```
[419] fa = FactorAnalysis(n_components=5)
      fa.fit(df_new)
      pd.DataFrame(fa.components_, columns=df.columns)
```

	Угол нашивки, град	Шаг нашивки	Плотность нашивки	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Удельная прочность
0	0.728230	-0.153352	0.228183	0.023111	-5.457260	23.689078	-1.688165	-0.017609	-0.161279	-13.083214	-0.003339	463.247398	1.630090	0.013024
1	-1.066228	0.038700	0.869226	0.037531	0.604984	327.847921	1.096562	-0.030134	1.160152	-5.252541	0.063230	-16.547451	0.146242	0.165445
2	2.075172	0.076528	-0.519907	0.008528	4.236416	4.305162	1.191170	-0.032704	0.758210	279.683415	0.091092	6.609212	-0.403295	0.001266
3	-2.373496	-0.131148	0.993023	0.002123	70.366982	0.190320	-1.570766	-0.002029	-1.011243	-0.147278	-0.071350	0.094343	-0.366186	-0.013554
4	0.465496	-0.022991	-0.010394	-0.067108	-0.265024	0.003833	0.018072	-0.028121	-2.276826	-0.000892	-0.167464	0.019933	-57.684294	-0.000010

Рисунок 5 – Вклад влияния факторов

С целью определения весов характеристик датасета для первых пяти факторов влияния был проведен анализ, последовательно добавляя по одному фактору. Полученные результаты

– Для двух факторов: выделить наименование и смысл нового фактора сложно. но можно увидеть высокие доли характеристик внутри факторов: модуль упругости, ГПа, Температура вспышки, С_2 Плотность нашивки;

– Для трех факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки;

– Для четырех факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки Угол нашивки, град;

– Для пяти факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки Угол нашивки, град Потребление смолы, г/м2.

Анализ полученных результатов показывает, что возможно характеристики: Количество отвердителя, Плотность нашивки, Плотность, Угол нашивки, Потребление смолы, являются наиболее существенными для построения будущих моделей.

Для дальнейшей разработки и обучения модели была выполнена нормализация данных с помощью масштабирования.

```
from sklearn import preprocessing
df_scaled = pd.DataFrame(preprocessing.scale(df_new),columns = df_new.columns)
df_scaled
```

2.2 Разработка и обучение модели

В качестве модели выбрана линейная регрессия и случайный лес. Обучение и тестирование производилось для двух параметров: модуль упругости и прочность при растяжении (в соответствии с условиями задачи).

При построении моделей «Случайный лес» был осуществлен поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой На рисунке 1 приведен анализ качества разработанных моделей.

2.3 Тестирование модели

Из датасета была выделена прогнозируемая (зависимая) переменная. Далее, выборка была разделена на обучающую и тестовую выборки, в соответствии с условием задачи 70% (на обучение) /30% (на тестирование).

Размер обучающей выборки: 655.

Размер тестовой выборки: 281.

2.4 Нейронная сеть

В качестве нейронной сети был принят многослойный персептрон.

Гиперпараметры модели:

- количество скрытых слоев = 3;
- количество нейронов на слое = 256 и 64;
- активационная функция «relu»;
- количество нейронов на выходном слое = 1;
- оптимизатор «Adam»;

Обучение модели происходило за 60 эпох (количество задано после поиска опытным путем наиболее приемлемого распределения MSE и общих результатов обучения модели). Результат изменения MSE модели указан на рисунке 31. MSE уменьшается со временем по мере выполнения алгоритма. Это означает, что модель приближаемся к оптимальному решению.

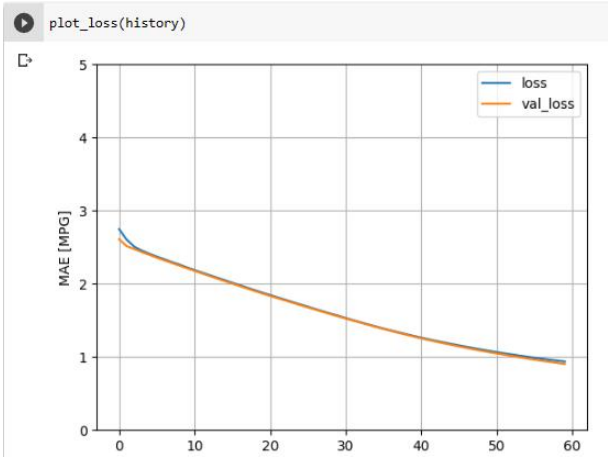

Рисунок 6 – Изменение MSE за время обучения модели

Ошибка в основном распределяется между 1 и 2. Это показывает, что модель обучения не совсем подходящая и требует дальнейшего обучения с целью привидения значение ошибки к 0. Гистограмма распределения ошибки приведена на рисунке 7.

```
error = test_predictions - y5_test
plt.hist(error, bins=25)
plt.xlabel('Prediction Error [MPG]')
_ = plt.ylabel('Count')
```

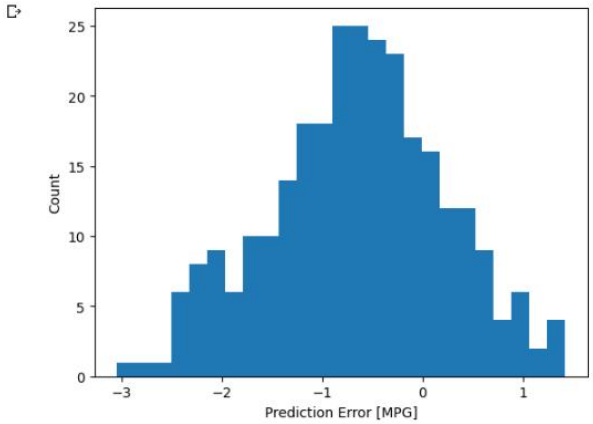

Рисунок 7 – Распределение ошибки

Так как полученная модель не оправдала ожидания, то была построена вторая модель на данных, нормализованных с помощью MinMaxScaler, смотрите Рисунок 8.

```
plot_loss(history2)
```

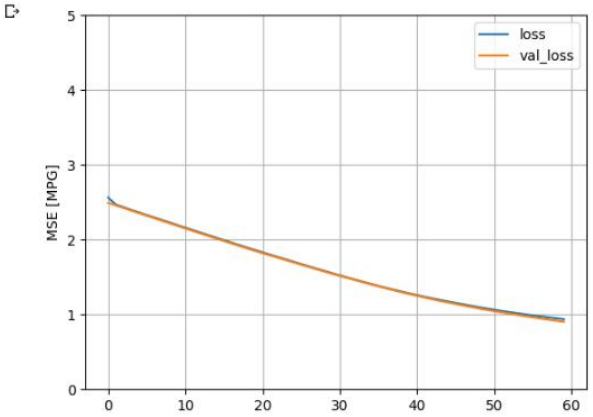

Рисунок 8 – Изменение MSE за время обучения модели

Ошибка в основном распределяется между 1 и 2. Это показывает, что модель требует дальнейшего обучения с целью привидения значение ошибки к 0. Гистограмма распределения ошибки приведена на рисунке 10. А диаграмма на Рисунке 9 показывает распределение прогнозируемый и актуальных значений. Как мы видим, то прогнозируемые значения распределены в горизонтальной плоскости и имеют значения в пределах 2-3,5.

```
test_predictions = dnn_goal_model.predict(X6_test).flatten()

a = plt.axes(aspect='equal')
plt.scatter(y6_test, test_predictions)
plt.xlabel('True Values')
plt.ylabel('Predictions')
lims = [0, 5]
plt.xlim(lims)
plt.ylim(lims)
_ = plt.plot(lims, lims)
```

```
9/9 [==============================] - 0s 5ms/step
```

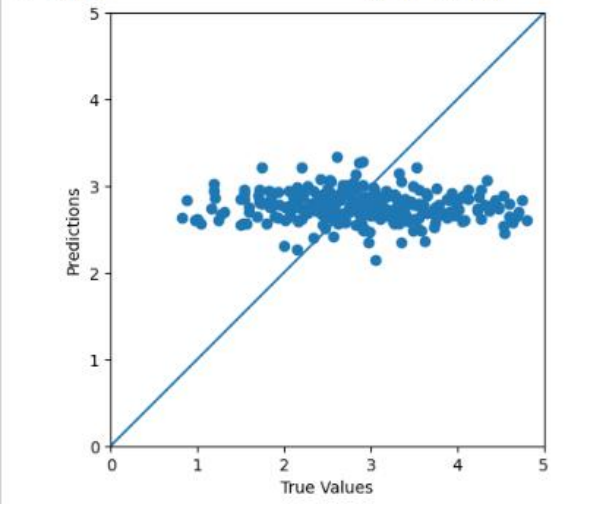

Рисунок 9. Диаграмма распределение предсказанных и актуальных значений.

Рисунок 10. Рисунок 7 – Распределение ошибки

3 Создание удаленного репозитория

Страница слушателя на GitLab
Созданный репозиторий:
https://github.com/JUPKSUP/VKR
Страница слушателя, созданный репозиторий, коммиты в репозитории.
https://github.com/JUPKSUP/VKR

Заключение

В ходе выполнения ВКР были изучены способы анализа и предобработки данных. Построенные модели показали, что исходный датасет является предобработанным и не содержит реальных значений для отработки обучения и тренировки моделей.

Полученная модель нейронной сети не идеальна, но позволяет предсказывать значения, близкие к средним значениям параметров.

Библиографический список

1. Л.И. Бондалетова, В.Г. Бондалетов Полимерные композиционные материалы: - Режим доступа - https://portal.tpu.ru/SHARED/b/BONDLI/stud_work/p_k_m_m/Tab1/Posobie_PCM.pdf. (дата обращения - 10.06.2022).
2. Электронная Книга / Основы Data Science и Big Data. Python и наука о данных - Д. Силен, А. Мейсман, М. Али
3. Теоретический минимум по Big Data. Всё что нужно знать о больших данных. Москва, изд Питер, ISBN 978-5-4461-1040-7
4. Интернет