Выпускная квалификационная работа

по курсу «Data Science»

Слушатель: Баркинхоева З. М.

Цели и задачи

Цель исследования заключается в попытке спрогнозировать ряд конечных свойств получаемых композиционных материалов; на входе используются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.), предоставленные Центром НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Задачи:

1) Изучить теоретические основы и методы решения поставленной задачи.
2) Провести разведочный анализ предложенных данных.
3) Провести предобработку данных.
4) Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении.
5) Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.
6) Оценить точность моделей на тренировочном и тестовом датасете.

Используемые методы

Поскольку данные в датасете относятся к непрерывным, в работу берутся регрессионные модели.*

Регрессия – это технология статистического анализа, целью которой является определение лучшей модели, устанавливающей взаимосвязь между выходной (зависимой) переменной и набором входных (независимых) переменных.

В исследовании применены следующие методы:

1. Линейная регрессия
2. Регрессия ближайших соседей
3. Дерево решений (Decision Tree). В данной работе используется дерево регрессии (Decision Tree Regressor), предназначенное для непрерывных целевых переменных, а не дерево классификации (Decision Tree Classifier), необходимое для дискретных переменных.

*И действительно впоследствии попытка обучения моделей Logistic Regression (которая требует категориальных признаков) и Decision Tree Classifier привели к ошибке «ValueError: Unknown label type: 'continuous'».

Разведочный анализ данных

Вывод для каждой колонки:

- среднего значения
- медианы

Проверка на:

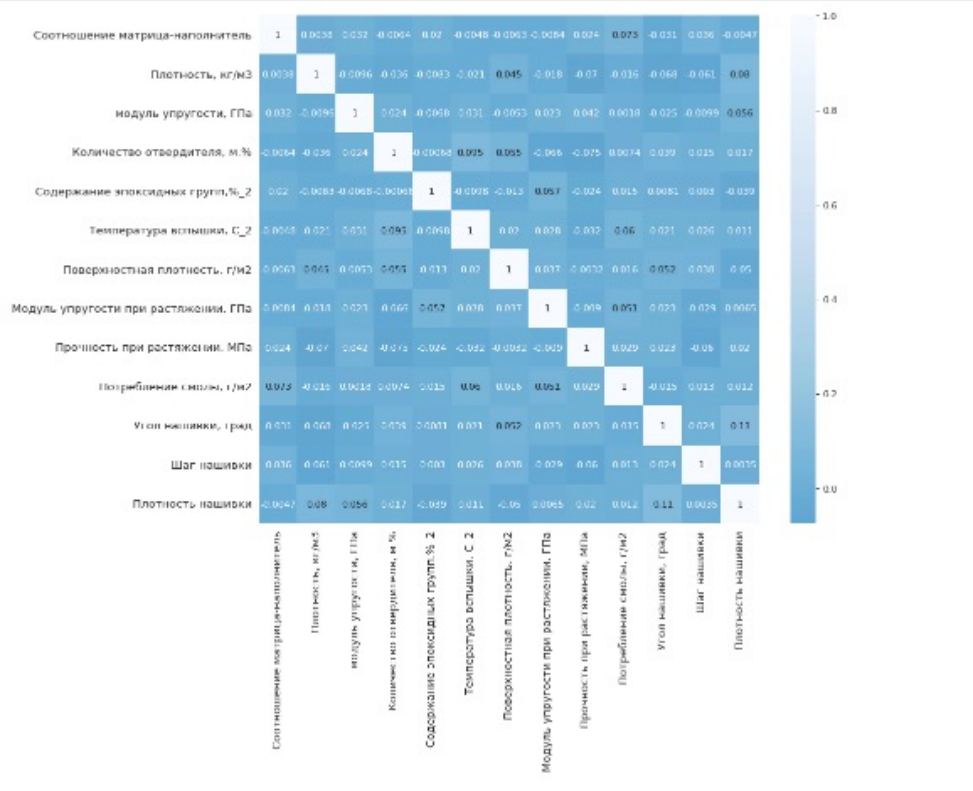
- выбросы
- пропуски

Анализ и визуализация предложенных данных:

- гистограммы распределения каждой переменной
- "ящики с усами"
- попарные графики рассеяния точек

Разведочный анализ данных

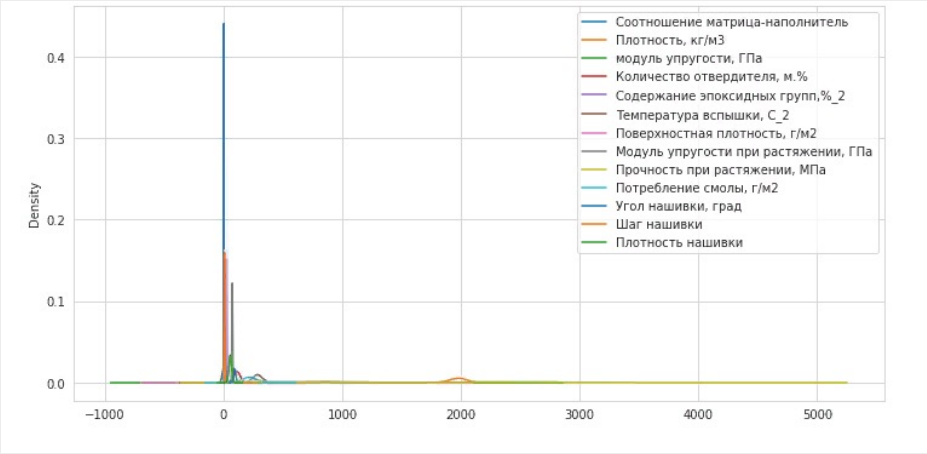
Примененные методы анализа не выявили четкой зависимости, что и подтвердила тепловая карта с матрицей корреляции: результаты близки к 0, максимальный результат – 0,11.

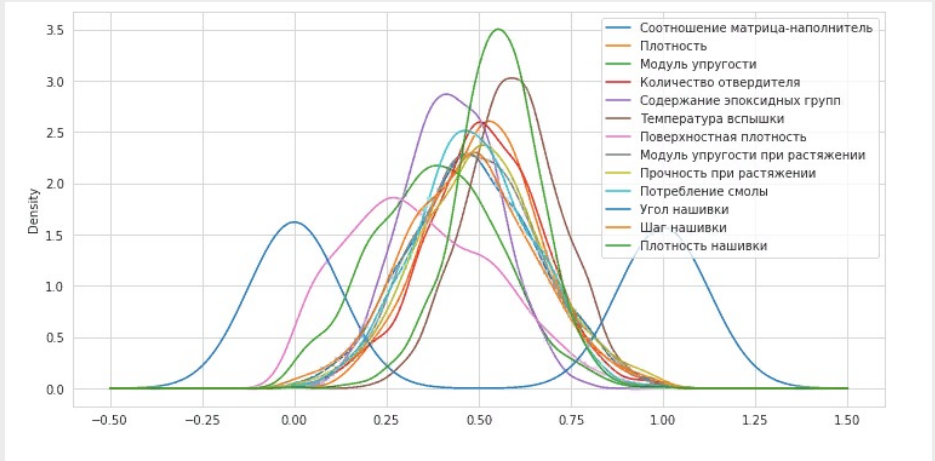


ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Предобработка данных

Условия поставленного задания требуют масштабирования данных путем нормализации. Нами выбран метод MinMaxScaler.





Алгоритмы и результаты

Разработку и обучение моделей машинного обучения было решено провести отдельно для каждого из выходных параметров (по условию «Прочность при растяжении» и «Модуль упругости при растяжении»).

Также по условию задания при построении модели необходимо 30% данных оставлялось на тестирование модели, на остальных проводилось обучение моделей. Поэтому в коде test_size равен 0.3.

После обучения моделей проведена оценка точности этих моделей на обучающей и тестовых выборках. В качестве параметра оценки модели использовалась средняя квадратическая ошибка (MSE) и коэффициент детерминации (R2 или R-квадрат), который показывает, какая доля изменчивости целевой переменной объясняется с помощью использованной модели.

Прочность при растяжении

Модель	MSE	R2
Linear Regression	0.028208863778752036	0.004230422816883017
KNeighborsRegressor	0.028208863778752036	0.004230422816883017
Decision Tree Regressor	0.028470633138137295	-0.005010004814622437

Модуля упругости при растяжении

Модель	MSE	R2
Linear Regression	3.5310902339377382e-31	1.0
KNeighborsRegressor	3.5310902339377382e-31	1.0
Decision Tree Regressor	0.002268699766192818	0.9207559947003582

Алгоритмы и результаты

При построении нейронной сети для соотношения «матрица – наполнитель» разделение на тестовую и обучающую выборки осуществлялось также по принципу 70/30 (test_size равен 0.3).

Количество оптимизируемых параметров равно 34059.

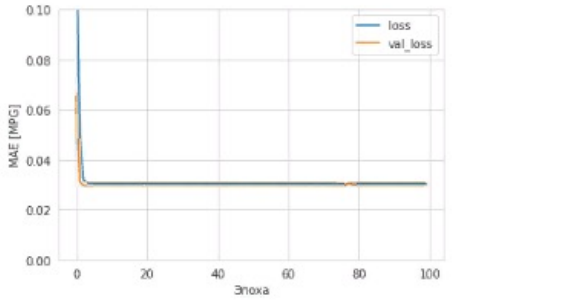
MSE: 0.031683008287056 16,

R2: -0.00039930547597832877.

```
ns = Sequential()
ns.add(layers.Dense(50, input_dim=12, activation='relu'))
ns.add(layers.Dense(128, activation='relu'))
ns.add(layers.Dense(128, activation='relu'))
ns.add(layers.Dense(64, activation='relu'))
ns.add(layers.Dense(32, activation='softmax'))
ns.add(layers.Dense(1))
ns.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_6 (Dense)             (None, 50)                650

 dense_7 (Dense)             (None, 128)               6528

 dense_8 (Dense)             (None, 128)               16512

 dense_9 (Dense)             (None, 64)                8256

 dense_10 (Dense)            (None, 32)                2080

 dense_11 (Dense)            (None, 1)                 33

=================================================================
Total params: 34,059
Trainable params: 34,059
Non-trainable params: 0
_________________________________________________________________
```

```
def plot_loss(history):
    plt.plot(history.history['loss'], label='loss')
    plt.plot(history.history['val_loss'], label='val_loss')
    plt.ylim([0, 0.1])
    plt.xlabel('Эпоха')

    plt.ylabel('MAE [MPG]')
    plt.legend()
    plt.grid(True)
plot_loss (history)
```



Выводы

- На основании проведенного анализа можно предположить несостоятельность выбранных моделей для прогнозирования, запрошенное условиями задания. Применённые модели регрессии не показали высокой эффективности в прогнозировании свойств композитов. Так, коэффициент детерминации должен принимать значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным. Но при расчете прочности при растяжении R2 в лучшем случае ближе к 0, а для дерева решений принял отрицательное значение. А при оценке моделей для прогноза значений модуля упругости при растяжении высокой оказались значения средней квадратической ошибки. По тем же соображениям считаем малоуспешной реализацию нейронной сети.

- Все же укажем, что лучшие результаты из предложенных моделей для модуля упругости при растяжении продемонстрировала линейная регрессия и метод ближайших соседей, для прочности при растяжении – дерево решений.

- Также отметим, что в последующих этапах изучения предоставленных материалов автором предполагается применение большего числа моделей, в том числе и не относящихся к регрессиям, удаление выявленных выбросов, преобразование категориальных признаков и другие манипуляции для большей эффективности исследования.

Спасибо за внимание!

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана