



ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу
«Data Science»

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Олейник Владимир Александрович



Постановка задачи

- 1 Произвести анализ данных
- 2 Выполнить предварительную обработку данных
- 3 Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении
- 4 Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель
- 5 Разработать приложение

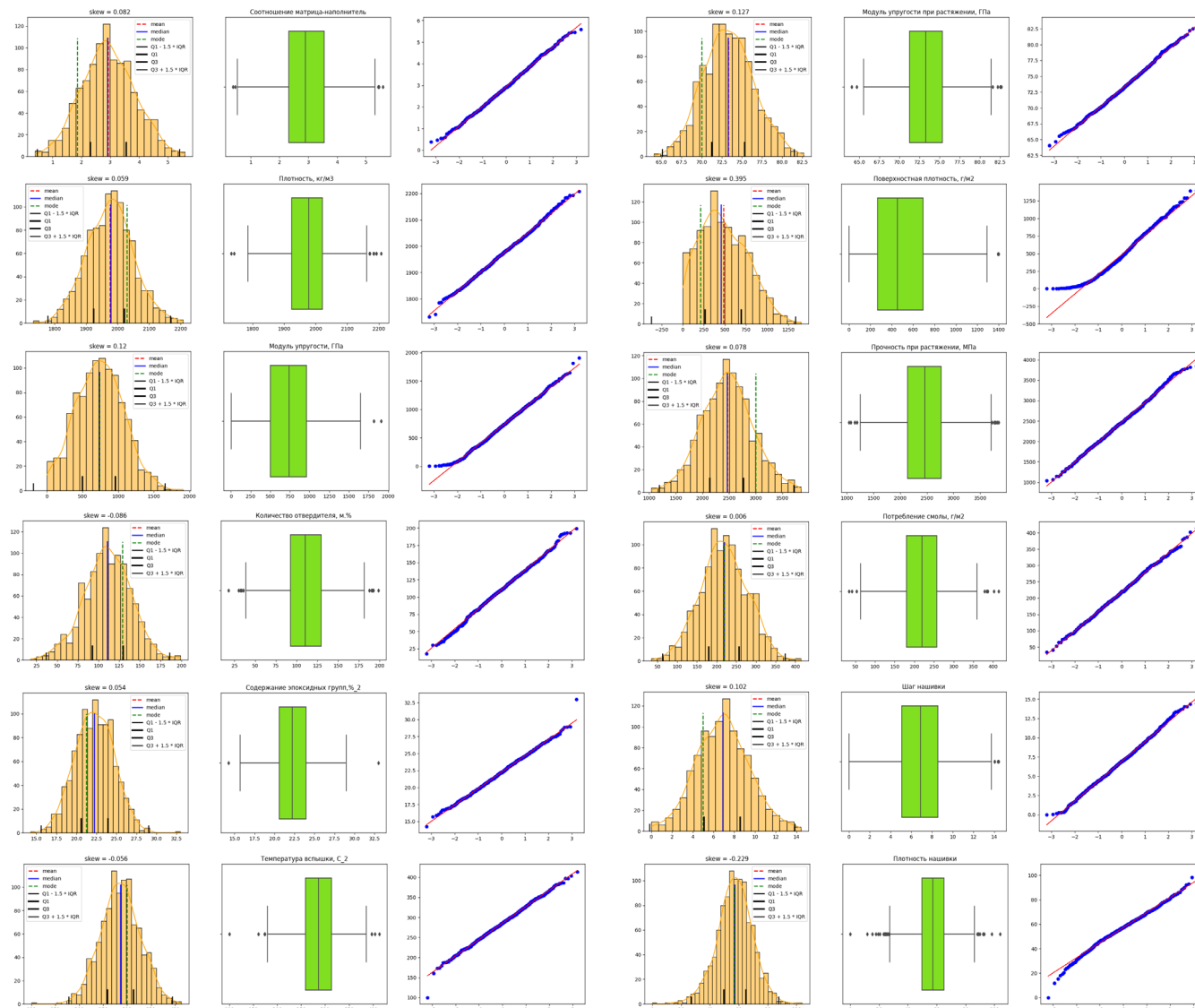
На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана)



Получена сводная и описательная статистическая информация, матрица корреляции, построены графики распределения, диаграммы «ящик с усами», графики «квантиль-квантиль», попарные графики рассеяния точек.

Результаты анализа:

- пропуски в данных отсутствуют;
- дубликаты отсутствуют;
- все признаки имеют числовой тип данных;
- признак «Угол нашивки, град» имеет только два уникальных значения, поэтому в дальнейшем изменим тип данных на категориальный;
- наличие выбросов;
- явная корреляция между признаками отсутствует;
- распределение близкое к нормальному по всем признакам, кроме «Поверхностная плотность, г/м²», признак имеет положительную асимметрию.





Предварительная обработка данных

Опробованы различные подходы устранения выбросов:

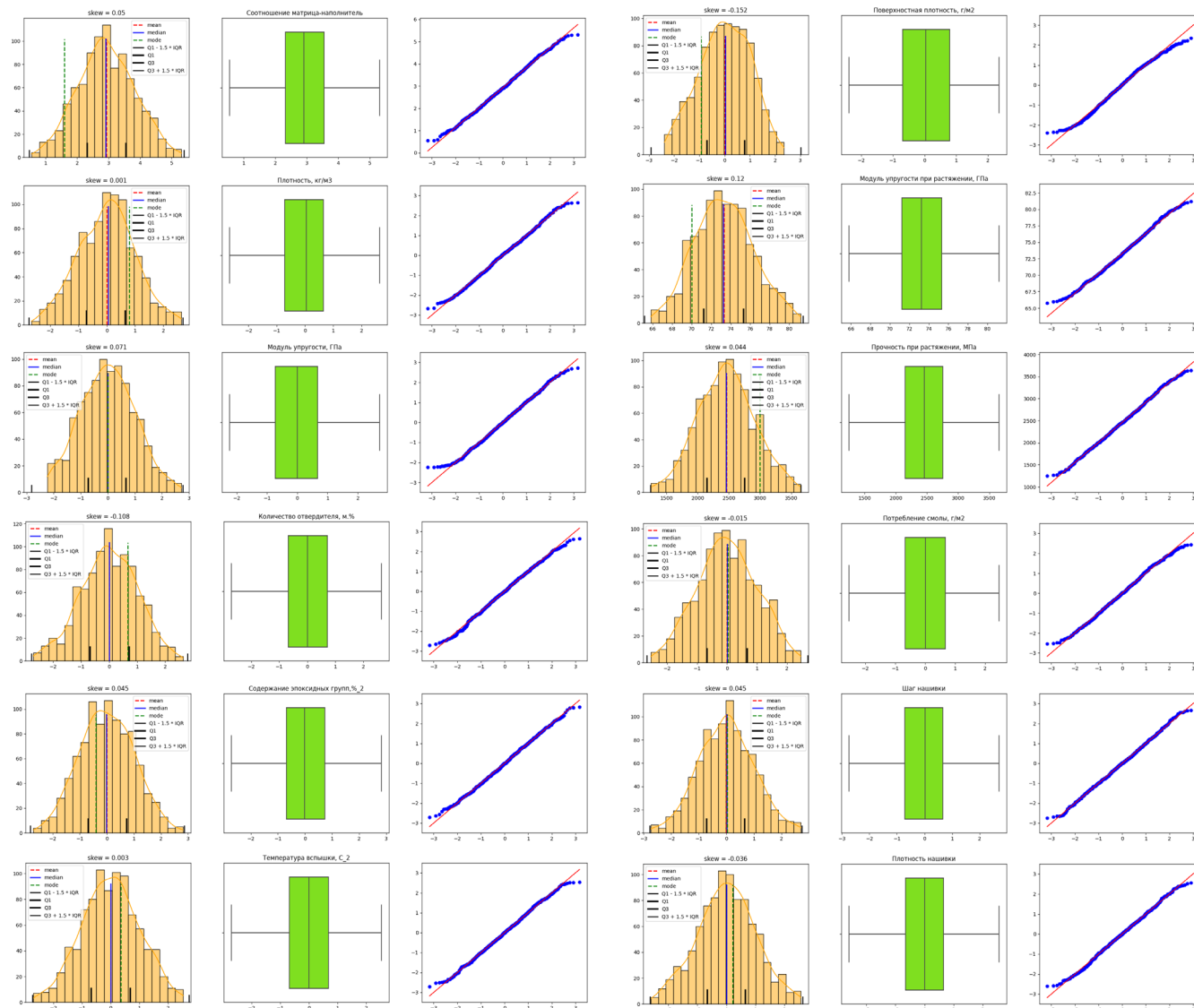
- правила трех сигм - удалено 23 строки, при повторных итерациях еще 4;
- межквартильный размах - удалено 87 строк, при повторных итерациях еще 14;
- 5- 95 квантилей - удалено 727 строк.

Выполнены различные методы преобразования данных:

- StandardScale совместно с PowerTransformer;
- MinMaxScaler;
- RobustScaler;
- QuantileTransformer.

Результаты предобработки:

- выбросы устранены методом межквартильного размах;
- предобработка данных выполнена с помощью StandardScaler и PowerTransformer.





Разработка и оценка моделей

Создан препроцессинг данных, который состоит из следующих преобразований:

- StandardScaler для числовых признаков без асимметрии;
- PowerTransformer для числовых признаков с явной асимметрией;
- OneHotEncoder для категориальных признаков.

С помощью GridSearchCV (поиск по сетке) выполнен подбор наилучших параметров моделей машинного обучения.

Создан Pipeline из препроцессинга данных и инициализированных моделей.

Обучены и сохранены модели для прогнозирования модуля упругости при растяжении и прочности при растяжении:

- LinearRegression;
- DecisionTreeRegressor;
- LinearSVR;
- KNeighborsRegressor;
- RandomForestRegressor.

	Model	MAE	MSE	R2
0	LinReg	2.520664	9.494092	-0.028532
1	DecTreeReg	3.434590	18.378696	-0.991035
2	LinSVR	2.522822	9.529168	-0.032332
3	KNNReg	2.488656	9.199464	0.003386
4	RandForReg	2.526412	9.775038	-0.058968

Метрики моделей для прогноза модуля упругости при растяжении

	Model	MAE	MSE	R2
0	LinReg	368.612465	210248.957038	-0.026610
1	DecTreeReg	546.510000	455256.167719	-1.222940
2	LinSVR	368.192588	206855.073445	-0.010039
3	KNNReg	366.052595	204508.051117	0.001421
4	RandForReg	383.447685	217757.993006	-0.063276

Метрики моделей для прогноза прочности при растяжении



Разработка нейронной сети

Модуль упругости при растяжении и прочность при растяжении будем считать дополнительными входными переменными для прогнозирования соотношения матрица-наполнитель.

Созданы два препроцессинга данных: один аналогичен препроцессингу созданному ранее, но с добавлением двух новых входных переменных, второй использует MinMaxScaler для преобразования числовых признаков и OneHotEncoder для категориальных признаков.

Инициализирована модель нейронной сети и выполнен ручной подбор параметров сети для датасета с разной предобработкой данных.

При обучении нейронной сети были использованы ModelCheckpoint для сохранения лучшей модели и EarlyStopping для ранней остановки обучения, при отсутствии улучшения точности модели после 20 эпох обучения.

Наилучший показатель детерминации показала модель со следующими параметрами сети:

- входной слой (количество нейронов - 8, инициализация матрицы весов – 'normal', функция активации – 'relu');
- выходной слой (количество нейронов - 1, инициализация матрицы весов – 'normal', функция активации – 'linear');
- функция ошибки 'mean_squared_error';
- оптимизатор - Adam (скорость обучения - 0.0001);
- метрика - 'mean_squared_error'.

	Model	MAE	MSE	R2		Model	MAE	MSE	R2
30	Best_model31	1.963414	4.923813	-5.126551	10	Best_model11	0.746346	0.859275	-0.069170
29	Best_model30	1.740889	3.973470	-3.944068	11	Best_model12	0.737074	0.838771	-0.043657
28	Best_model29	1.621751	3.734108	-3.646237	5	Best_model6	0.732213	0.826989	-0.028997
31	Best_model32	1.088815	1.701520	-1.117150	30	Best_model31	0.730297	0.816529	-0.015983
19	Best_model20	0.725915	0.807424	-0.004653	29	Best_model30	0.727404	0.803535	0.000186
21	Best_model22	0.724715	0.804841	-0.001439	15	Best_model16	0.724841	0.803500	0.000229
13	Best_model14	0.734034	0.804738	-0.001312	24	Best_model25	0.724744	0.803040	0.000801
23	Best_model24	0.725160	0.802765	0.001143	20	Best_model21	0.724952	0.802893	0.000985
27	Best_model28	0.724351	0.792748	0.013608	32	Best_model33	0.724462	0.798938	0.005906
17	Best_model18	0.724967	0.792472	0.013951	28	Best_model29	0.724881	0.798927	0.005920
33	Best_model34	0.722363	0.790095	0.016909	33	Best_model34	0.722947	0.796763	0.008611
7	Best_model8	0.715045	0.781157	0.028030	31	Best_model32	0.725978	0.796449	0.009003

Метрики моделей нейронной сети
(слева – датасет после StandardScaler,
справа – датасет после MinMaxScaler)

Дополнительно была выполнена попытка параметров сети перебором через цикл. Параметры цикла:

- количество слоев – [1, 2, 3];
- количество нейронов - [8, 16, 32, 64];
- функции активации - ['tanh', 'linear', 'relu'];
- оптимизаторы - ['sgd', 'adam'].

Полученные модели не дали коэффициент детерминации выше, чем модель, указанная выше.



Разработка приложения

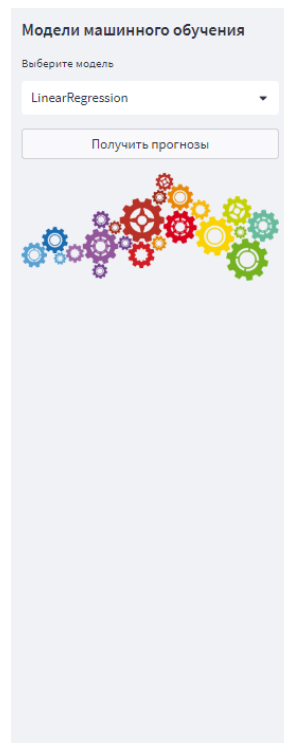
Для разработки приложения использован фреймворк Streamlit.

Разработанное приложение позволяет пользователю:

1. внести характеристики матрицы и наполнителя (по умолчанию, заполнены значениями из тестовой выборки);
2. выбрать модель машинного обучения (по умолчанию выбрана модель линейной регрессии);
3. получить прогнозы модуля упругости при растяжении, прочности при растяжении, соотношения матрица-наполнитель после нажатия кнопки «Получить прогнозы».

Принцип работы приложения после нажатия кнопки «Получить прогнозы» следующий:

1. формируется датасет из внесенных значений;
2. в зависимости от выбора пользователя загружаются модели для прогнозирования модуля упругости при растяжении и прочности при растяжении, и выполняется прогнозирование этих значений;
3. формируется датасет из внесенных и полученных значений из предыдущего пункта;
4. загружается модель препроцессинга данных и выполняется преобразования датасета из предыдущего пункта;
5. загружается модель нейронной сети и выполняется прогнозирование соотношения матрица-наполнитель;
6. вывод результатов работы.



Выпускная квалификационная работа по курсу «Data Science»

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Олейник Владимир Александрович

Плотность, кг/м3

1881,62

-

+

Модуль упругости, ГПа

663,33

-

+

Количество отвердителя, м.%

123,51

-

+

Содержание эпоксидных групп,%₂

28,32

-

+

Температура вспышки, C₂

220,23

-

+

Поверхностная плотность, г/м2

363,13

-

+

Потребление смолы, г/м2

314,70

-

+

Угол нашивки, град

0

▼

Шаг нашивки

9,59

-

+

Плотность нашивки

35,88

-

+

Соотношение матрица-наполнитель


3.0123

Модуль упругости при растяжении, ГПа

73.52

Прочность при растяжении, МПа

2476.24



Метрики 'Модуль упругости при растяжении, ГПа'

Метрики 'Прочность при растяжении, МПа'



<https://olvovchik-data-science-learn-applicationapp-tfgy2e.streamlit.app/>
ссылка на приложение в сети Интернет.



Заключение

Полученный коэффициент детерминации обученных моделей и нейронной сети практически нулевой – это означает, что связь между переменными регрессионной модели отсутствует и получаемые прогнозы ничем не отличаются от прогноза средним значением.

Итоговое решение поставленной задачи не достигнуто, требуется более детальный анализ данных (желательно с привлечением специалистов предметной области). Также можно попробовать использовать другие методы и модели прогнозирования, которые не были рассмотрены в текущей работе.



Спасибо за внимание!



ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана



do.bmstu.ru