ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

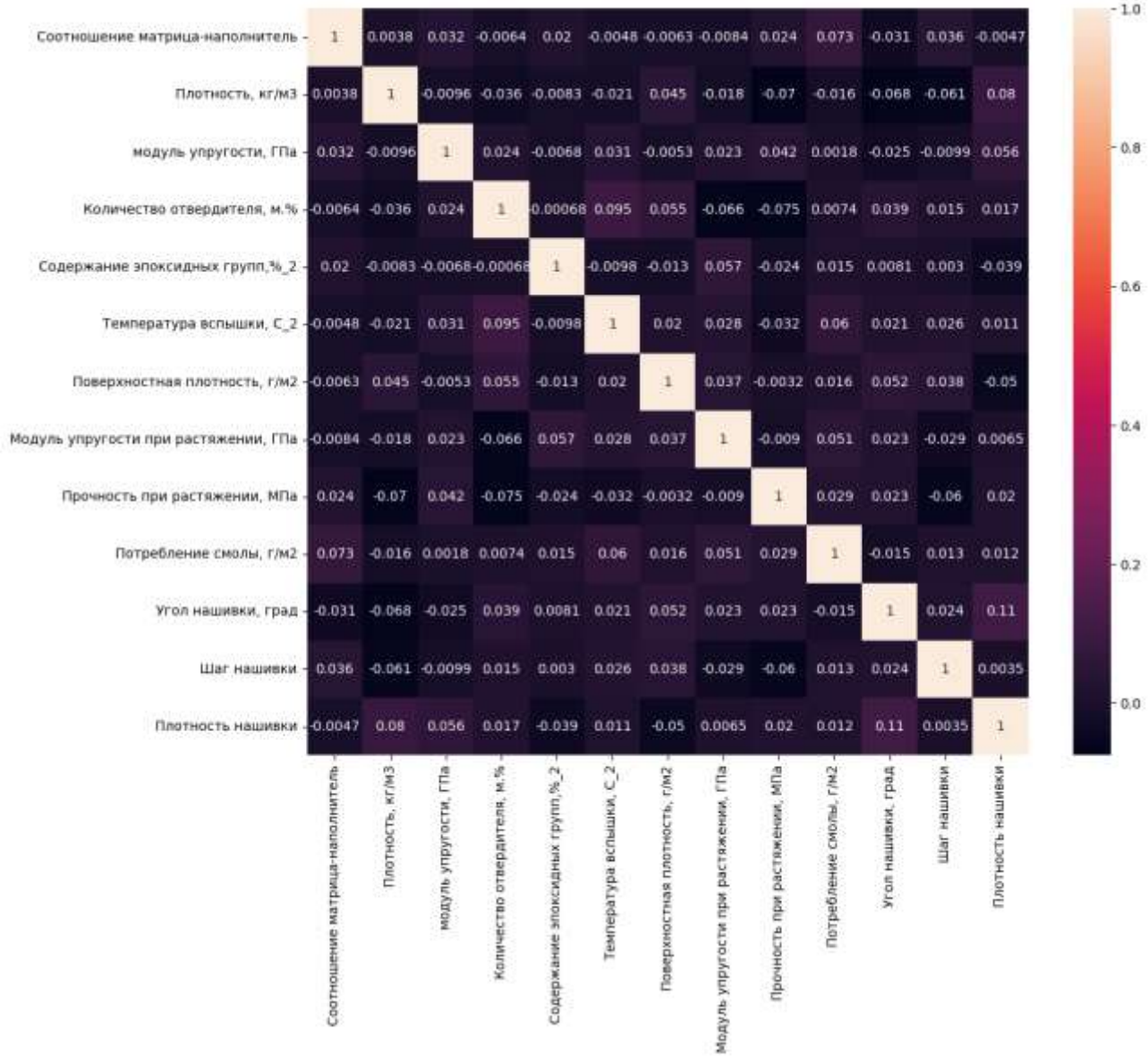
Выпускная квалификационная работа по курсу «Data Science»

Тема: «Прогнозирование конечных свойств новых материалов (композиционных материалов)»

Слушатель: Курбатов А.В.

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных



Результаты

1. Всего 1023 объекта, 13 признаков, 3 из которых целевые
2. Все характеристики являются числовыми, пропусков в данных нет
3. Каких-либо зависимостей между признаками не выявлено, попарные коэффициенты корреляции близки к 0
4. График попарного рассеяния точек не выявил зависимостей, видны выбросы

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Предобработка данных

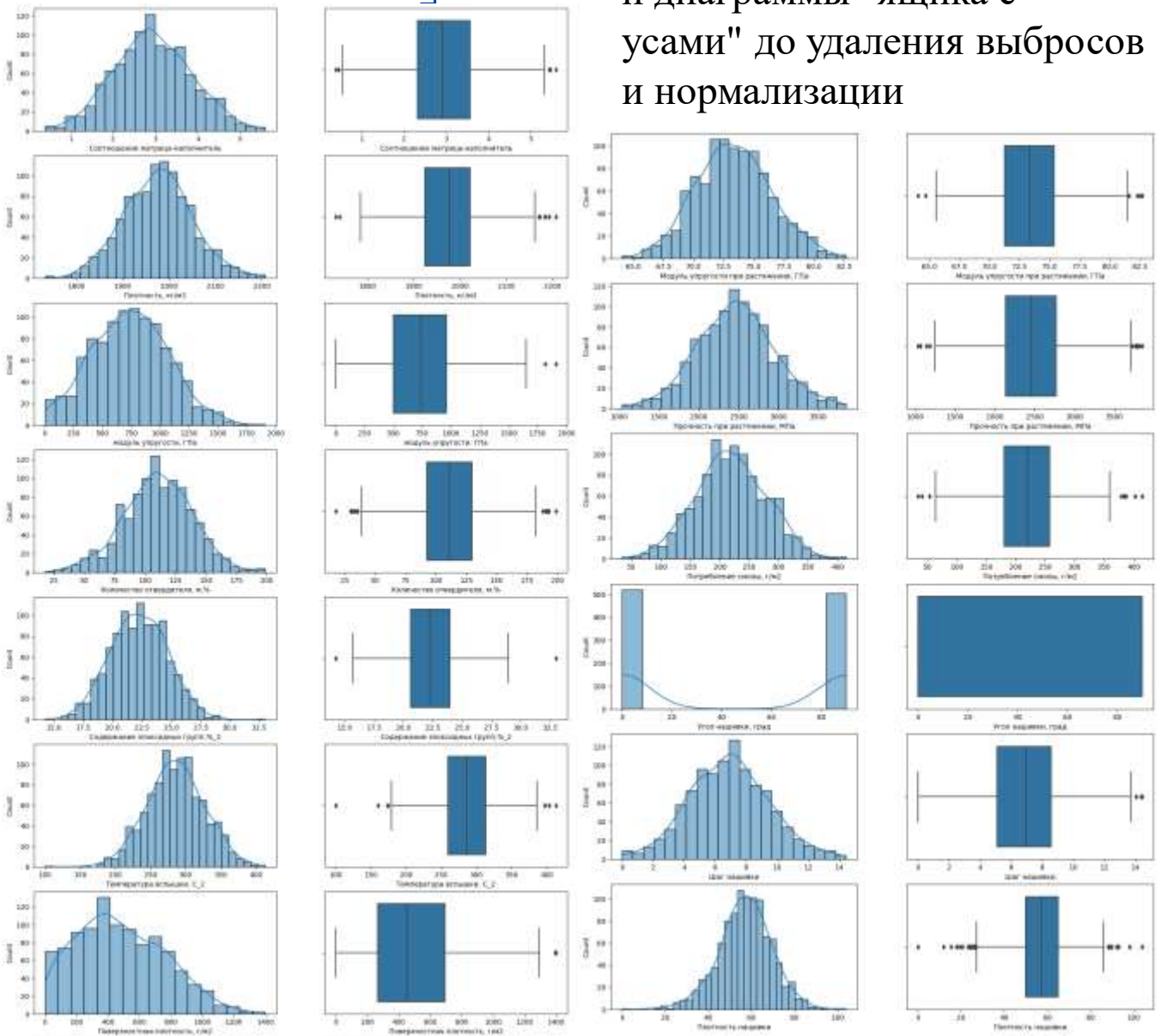
Проведены:

1. Нормализация с помощью **MinMaxScaler** библиотеки **sklearn**, привела все признаки к диапазону от **0** до **1.**
2. Удаление выбросов методом межквартильного размаха, т.к. для некоторых признаков наблюдаем ассиметричное распределение

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.930366	0.913222	0.389403	2.317887	2.906878	3.552660	5.591742
Плотность, кг/м3	1023.0	1975.734888	73.729231	1731.764635	1924.155467	1977.621657	2021.374375	2207.773481
модуль упругости, ГПа	1023.0	739.923233	330.231581	2.436909	500.047452	739.664328	961.812526	1911.536477
Количество отвердителя, м.%	1023.0	110.570769	28.295911	17.740275	92.443497	110.564840	129.730366	198.953207
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301	14.254985	20.608034	22.230744	23.961934	33.000000
Температура вспышки, С_2	1023.0	285.882151	40.943260	100.000000	259.066528	285.896812	313.002106	413.273418
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690	0.603740	266.816645	451.864365	693.225017	1399.542362
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983	64.054061	71.245018	73.268805	75.356612	82.682051
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006	1036.856605	2135.850448	2459.524526	2767.193119	3848.436732
Потребление смолы, г/м2	1023.0	218.423144	59.735931	33.803026	179.627520	219.198882	257.481724	414.590628
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1023.0	6.899222	2.563467	0.000000	5.080033	6.916144	8.586293	14.440522
Плотность нашивки	1023.0	57.153929	12.350969	0.000000	49.799212	57.341920	64.944961	103.988901

Описательная статистика до обработки

Гистограммы распределения и диаграммы "ящика с усами" до удаления выбросов и нормализации



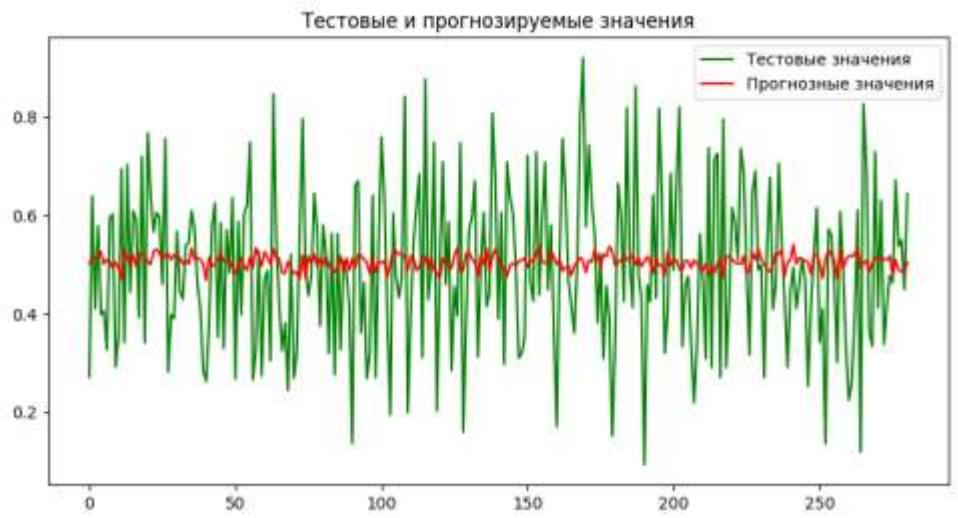
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Построение моделей прогнозирования модуля упругости при растяжении

Сравнение лучших моделей

Название модели с указанием параметров	R2	MSE	MAE	Точность, %
Метод K-ближайших соседей KNeighborsRegressor (n_neighbors=106)	0.001	0.027	0.136	72.424
Градиентный бустинг GradientBoostingRegressor (n_estimators=30, learning_rate=0.0001)	-0.001	0.027	0.135	72.559
Лассо Lasso(alpha=0.1)	-0,001	0,027	0.135	72,561
Эластичная сеть ElasticNet(alpha=0.1)	-0,001	0,027	0.135	72,561

- Датасет разделен на тренировочную (70%) и тестовую (30%) выборки.
- Рассмотрено 7 моделей в разных вариациях.
- Для 4 лучших моделей выполнен подбор гиперпараметров с помощью метода GridSearchCV библиотеки sklearn.



Метод K-ближайших соседей

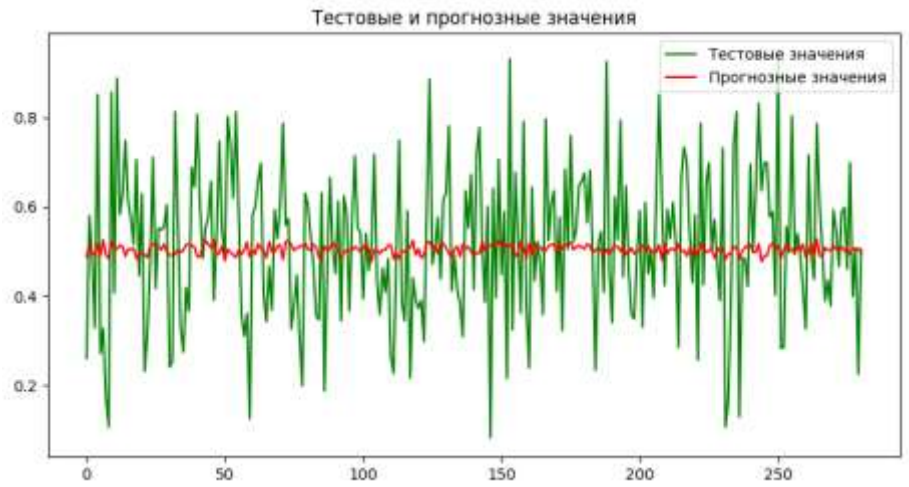
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Построение моделей прогнозирования прочности при растяжении

Сравнение лучших моделей

Название модели с указанием параметров	R2	MSE	MAE	Точность, %
Метод K-ближайших соседей KNeighborsRegressor (n_neighbors=145)	0.006	0.028	0.135	73.86
Градиентный бустинг GradientBoostingRegressor (learning_rate=0.01, n_estimators=3)	-0.006	0.029	0.136	73.73
Обобщенная линейная модель TweedieRegressor(alpha=100, max_iter=10, power=1, verbose=1)	-0.006	0.029	0.136	73.729

- Датасет разделен на тренировочную (70%) и тестовую (30%) выборки.
- Рассмотрено 4 моделей в разных вариациях.
- Для 3 лучших моделей выполнен подбор гиперпараметров с помощью метода GridSearchCV библиотеки sklearn.



Метод K-ближайших соседей

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка НС для прогнозирования Соотношения матрица-наполнитель

- **12** входных признаков, **1** - целевой.
- Нейронную сеть строю с помощью класса **Sequential** библиотеки **keras.**

Что испробовано:

- Несколько вариаций архитектуры НС.
- Общий подход к предобработке данных и вариант без удаления выбросов.
- Контроль метрик при обучении и соответствующий выбор количества эпох.
- Оптимизаторы **Adam** и **SGD**
- Различные функции активации на скрытых слоях.

Метрики:

- **MSE=0.03,**
- **R2=-0.001,**
- Точность **71%.**

Лучший вариант НС:

Параметры НС:

- Входной слой: полносвязный, **8** нейронов, ФА **tanh;**
- Один скрытый полносвязный слой с **8** нейронами, ФА **tanh;**
- Один **Dropout** слой;
- Выходной слой: полносвязный, **1** нейрон, ФА **leaner;**
- Оптимизатор: стохастический градиентный спуск **(SGD);**
- **loss**-функция: среднеквадратичная ошибка **(mean_squared_error).**

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Заключение

- Использованные при разработке моделей подходы не позволили получить сколько-нибудь достоверных прогнозов.
- Все признаки имеют очень слабую корреляцию между собой, и в датасете присутствуют выбросы, мешающие обучению моделей.
- После проведения различных стратегий предобработки данных не удалось достичь значимых результатов.
- Поставленная задача не решена. Необходимо использовать другие возможные пути решения поставленной задачи.

Спасибо за внимание