ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science»
Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель: Иващенко Анастасия Сергеевна

Основные задачи

- изучить предметную область;
- провести разведочный анализ данных;
- разделить данные на тренировочную и тестовую выборки;
- выполнить препроцессинг (предобработку);
- выбрать базовую модель и модели для подбора;
- сравнить модели с гиперпараметрами по умолчанию;
- подобрать гиперпараметры с помощью с помощью поиска по сетке с перекрестной проверкой;
- сравнить модели после подбора гиперпараметров и выбрать лучшую сравнить качество лучшей и базовой моделей на тестовой выборке
- сравнить качество лучшей модели на тренировочной и тестовой выборке разработать приложение.

Используемые библиотеки и модули

```
#Импорт нужных библиотек
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score, cross_validate
from sklearn.model_selection import GridSearchCV
from sklearn.base import BaseEstimator
from sklearn.dummy import DummyRegressor
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn import metrics
import pickle

import tensorflow as tf
from tensorflow import keras
print(tf.__version__)
```

```
2.9.2
```

Характеристика анализируемого датасета

Датасет состоит из двух файлов: X_bp и X_nup .
Файл X_bp содержит:
признаков: 10 и индекс;
строк: 1023.
Файл X_nup содержит:
признаков: 3 и индекс;
строк: 1040.
По заданию файлы требуют объединения с типом INNER по индексу.

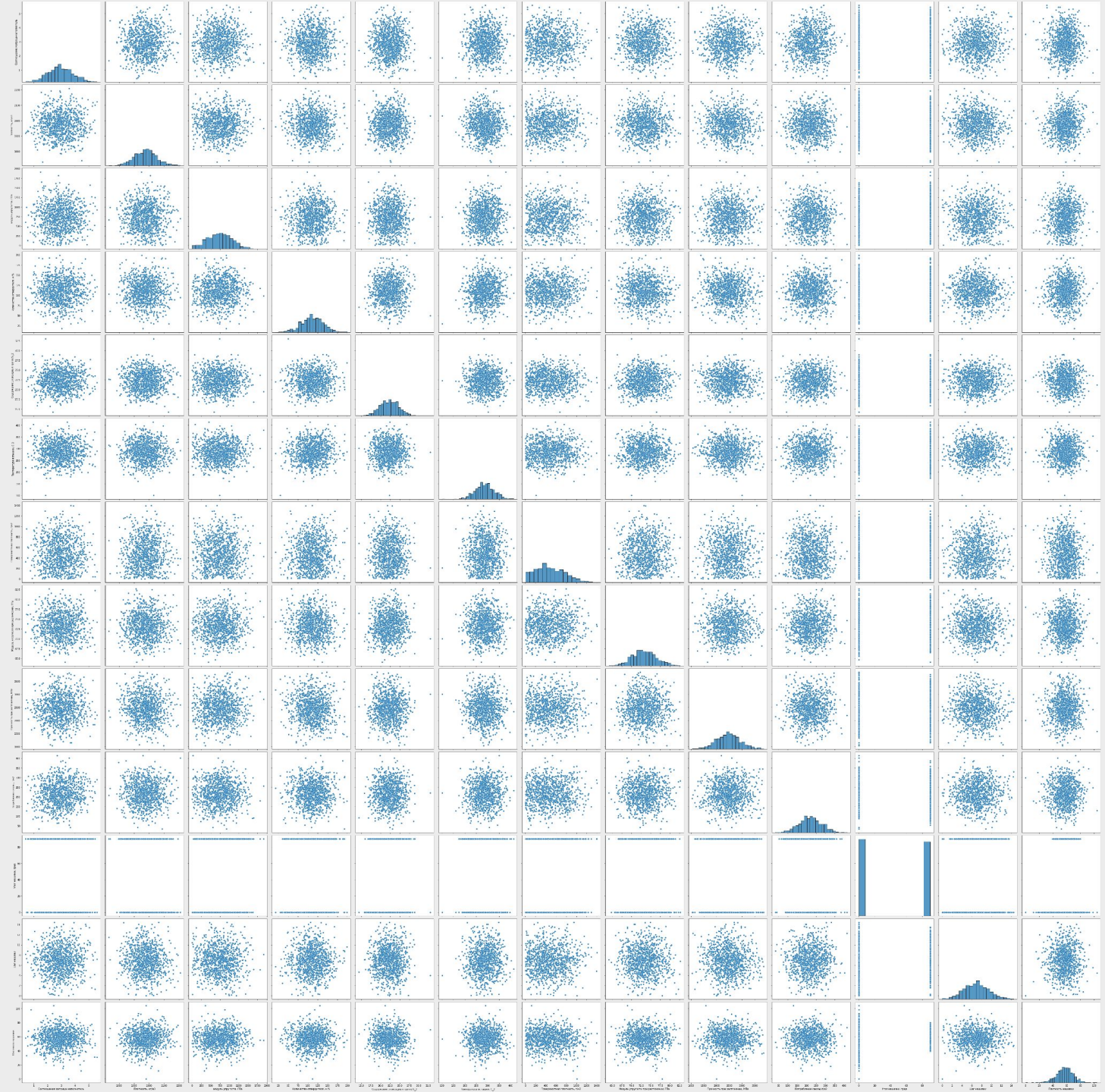
После объединения часть строк из файла X_nup была отброшена. И дальнейшие исследования проводим с объединенным датасетом, содержащим 13 признаков и 1023 строк.

Типы данных свойств в датасете

```
 #    Column                                 Non-Null Count  Dtype
---   ------                                 --------------  -----
 0    Соотношение матрица-наполнитель        1023 non-null   float64
 1    Плотность, кг/м3                       1023 non-null   float64
 2    модуль упругости, ГПа                  1023 non-null   float64
 3    Количество отвердителя, м.%            1023 non-null   float64
 4    Содержание эпоксидных групп,%_2        1023 non-null   float64
 5    Температура вспышки, С_2               1023 non-null   float64
 6    Поверхностная плотность, г/м2          1023 non-null   float64
 7    Модуль упругости при растяжении, ГПа   1023 non-null   float64
 8    Прочность при растяжении, МПа          1023 non-null   float64
 9    Потребление смолы, г/м2                1023 non-null   float64
 10   Угол нашивки, град                     1023 non-null   float64
 11   Шаг нашивки                            1023 non-null   float64
 12   Плотность нашивки                      1023 non-null   float64
dtypes: float64(13)
memory usage: 111.9 KB
```

Описательная статистика датасета

	count	mean	std	min	25%	50%	75%	max	median
Соотношение матрица-наполнитель	1023.0000	2.9304	0.9132	0.3894	2.3179	2.9069	3.5527	5.5917	2.9069
Плотность, кг/м3	1023.0000	1975.7349	73.7292	1731.7646	1924.1555	1977.6217	2021.3744	2207.7735	1977.6217
модуль упругости, ГПа	1023.0000	739.9232	330.2316	2.4369	500.0475	739.6643	961.8125	1911.5365	739.6643
Количество отвердителя, м.%	1023.0000	110.5708	28.2959	17.7403	92.4435	110.5648	129.7304	198.9532	110.5648
Содержание эпоксидных групп,%_2	1023.0000	22.2444	2.4063	14.2550	20.6080	22.2307	23.9619	33.0000	22.2307
Температура вспышки, С_2	1023.0000	285.8822	40.9433	100.0000	259.0665	285.8968	313.0021	413.2734	285.8968
Поверхностная плотность, г/м2	1023.0000	482.7318	281.3147	0.6037	266.8166	451.8644	693.2250	1399.5424	451.8644
Модуль упругости при растяжении, ГПа	1023.0000	73.3286	3.1190	64.0541	71.2450	73.2688	75.3566	82.6821	73.2688
Прочность при растяжении, МПа	1023.0000	2466.9228	485.6280	1036.8566	2135.8504	2459.5245	2767.1931	3848.4367	2459.5245
Потребление смолы, г/м2	1023.0000	218.4231	59.7359	33.8030	179.6275	219.1989	257.4817	414.5906	219.1989
Угол нашивки, град	1023.0000	44.2522	45.0158	0.0000	0.0000	0.0000	90.0000	90.0000	0.0000
Шаг нашивки	1023.0000	6.8992	2.5635	0.0000	5.0800	6.9161	8.5863	14.4405	6.9161
Плотность нашивки	1023.0000	57.1539	12.3510	0.0000	49.7992	57.3419	64.9450	103.9889	57.3419



Попарные графики рассеивания

Выбросы

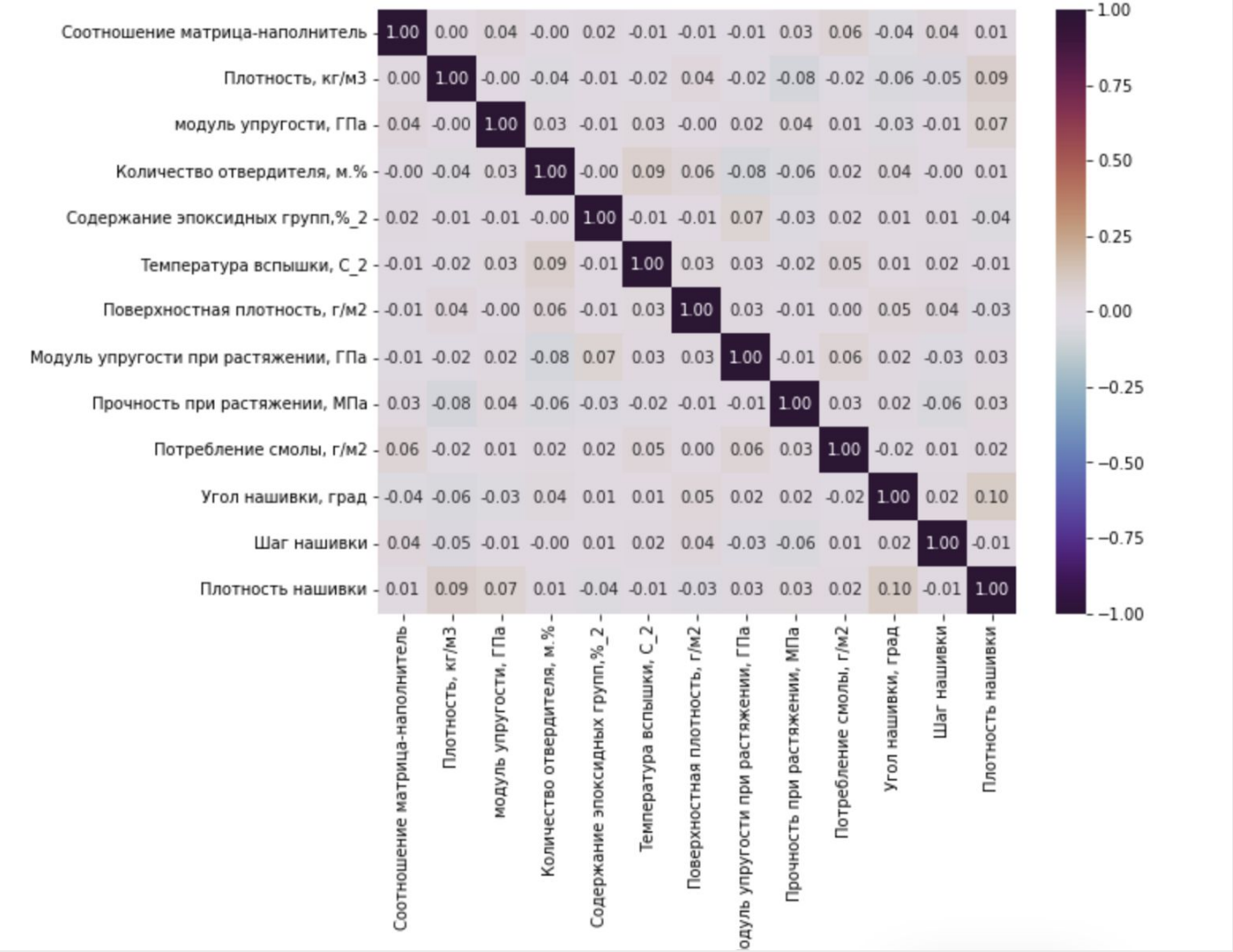
Найдено:

методом 3-х сигм — 24 выброса

методом межквартильных расстояний — 93 выброса

Применим метод 3-х, чтобы потерять меньше значимой информации, так как датасет уже очищен от лишней информации.

Матрица корреляции



Модель для модуля упругости при растяжении, ГПа

Сравнение моделей с подобранными параметрами, поиск лучшей

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=480, solver='lsqr')	-0.013299	-3.046623	-2.455526	-0.033517	-8.071899
Lasso(alpha=0.15)	-0.019048	-3.055423	-2.459921	-0.033574	-8.102101
SVR(C=0.015, kernel='linear')	-0.016521	-3.052020	-2.456808	-0.033549	-8.140634
KNeighborsRegressor(n_neighbors=25)	-0.030786	-3.074728	-2.461113	-0.033581	-8.031419
DecisionTreeRegressor(criterion='absolute_error', max_depth=2, max_features=10, random_state=3128, splitter='random')	-0.009281	-3.041407	-2.435050	-0.033185	-8.004156
RandomForestRegressor(bootstrap=False, criterion='absolute_error', max_depth=4, max_features=2, random_state=3128)	-0.015396	-3.049810	-2.446070	-0.033369	-8.275716

Сравнение предсказаний базовой модели и лучшей модели на тестовом множестве

	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.001377	-3.222954	-2.577796	-0.035319	-7.800690
Лучшая модель (дерево решений)	-0.035776	-3.277844	-2.610243	-0.035707	-8.152045

Модель для прочности при растяжении, МПа

Сравнение моделей с подобранными параметрами, поиск лучшей

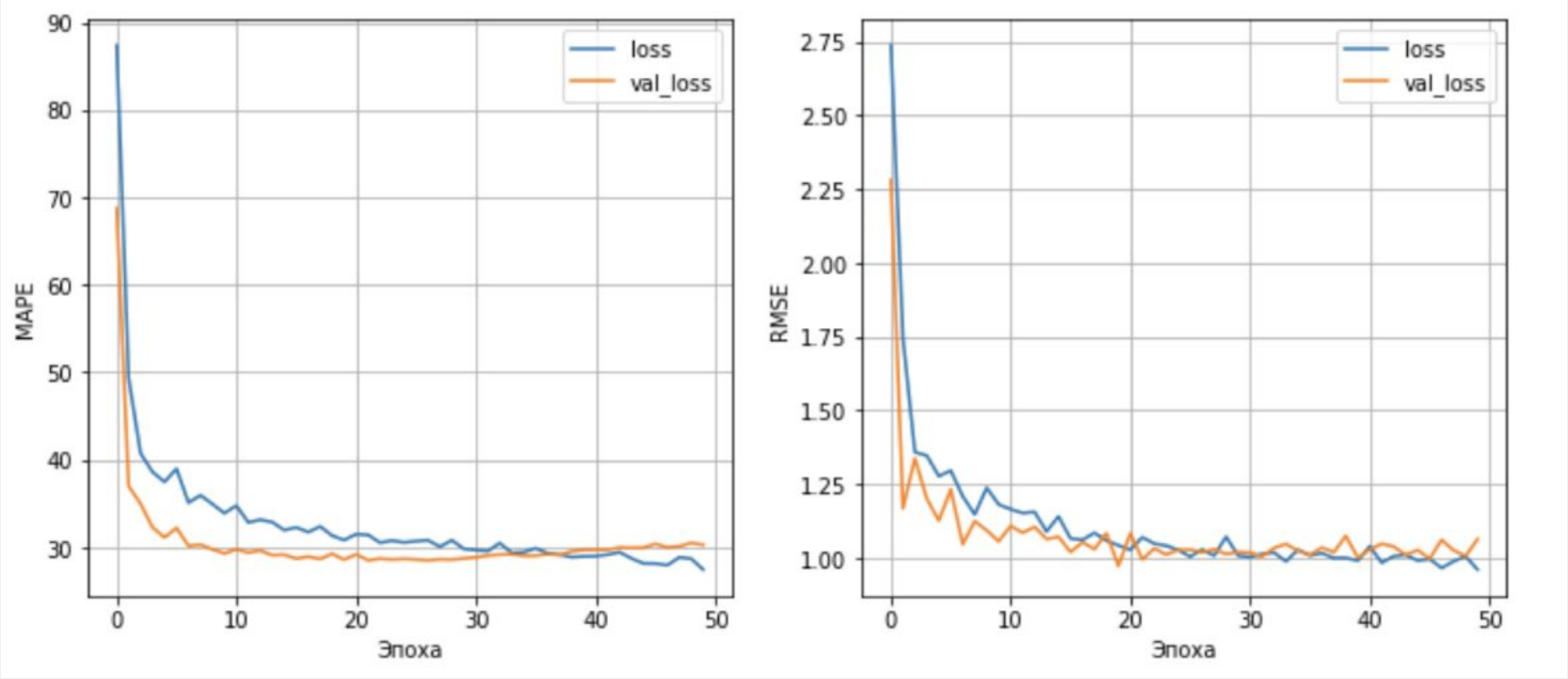
	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=990, solver='sparse_cg')	-0.010764	-484.199853	-385.891069	-0.169828	-1233.196571
Lasso(alpha=50)	-0.012988	-484.654884	-385.827028	-0.169931	-1228.780064
SVR(C=0.2)	-0.012246	-484.489867	-385.724279	-0.169413	-1232.341495
DecisionTreeRegressor(criterion='poisson', max_depth=3, max_features=6, random_state=3128, splitter='random')	-0.009440	-483.713960	-384.045197	-0.169031	-1244.359901
GradientBoostingRegressor(max_depth=1, max_features=1, n_estimators=50, random_state=3128)	-0.005486	-483.026609	-385.268908	-0.169409	-1231.878292

Сравнение предсказаний базовой модели и лучшей модели на тестовом множестве

	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.000531	-479.694153	-375.066608	-0.165566	-1431.321957
Лучшая модель (градиентный бустинг)	0.004028	-478.600202	-376.647056	-0.166046	-1384.841404

Модель для соотношения матрица-наполнитель

График обучения нейросети с Dropout-слоем



Сравнение предсказаний базовой модели и лучшей модели на тестовом множестве

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.011269	-0.911261	-0.737067	-0.299795	-2.684301
Нейросеть переобученная	-0.654114	-1.165445	-0.954058	-0.373033	-2.762262
Нейросеть dropout	-0.727036	-1.190858	-0.970366	-0.334842	-3.112869

Оценка точности модели на тренировочном и тестовом датасете

Модель для модуля упругости при растяжении

	R2	RMSE	MAE	MAPE	max_error
Модуль упругости, тренировочный	0.017295	-3.037284	-2.410294	-0.032850	-9.008468
Модуль упругости, тестовый	-0.035776	-3.277844	-2.610243	-0.035707	-8.152045

Модель для прочности при растяжении

	R2	RMSE	MAE	MAPE	max_error
Прочность при растяжении, тренировочный	0.057141	-472.832206	-374.670333	-0.164825	-1383.885510
Прочность при растяжении, тестовый	0.004028	-478.600202	-376.647056	-0.166046	-1384.841404

Модель для соотношения матрица-наполнитель

	R2	RMSE	MAE	MAPE	max_error
Соотношение матрица-наполнитель, тренировочный	-0.323193	-1.044150	-0.795323	-0.267259	-3.750573
Соотношение матрица-наполнитель, тестовый	-0.727036	-1.190858	-0.970366	-0.334842	-3.112869

Создание удаленного репозитория

Адрес страницы удаленного репозитория:
https://github.com/ivashenkoas/vkr_datascience22

edu.bmstu.ru

+7 495 182-83-85

edu@bmstu.ru

Москва, Госпитальный переулок , д. 4-6, с.3

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана