ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА по курсу «Data Science»

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель: Леонтьева Ксения Александровна
DS11838/2

Разведочный анализ данных

```
X_bp

<class 'pandas.core.frame.DataFrame'>
Float64Index: 1023 entries, 0.0 to 1022.0
Data columns (total 10 columns):
 #   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
dtypes: float64(10)
memory usage: 87.9 KB

----------------------------------------------------------------------
X_nup

<class 'pandas.core.frame.DataFrame'>
Float64Index: 1040 entries, 0.0 to 1039.0
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Угол нашивки, град  1040 non-null   float64
 1   Шаг нашивки         1040 non-null   float64
 2   Плотность нашивки   1040 non-null   float64
dtypes: float64(3)
memory usage: 32.5 KB

----------------------------------------------------------------------
```

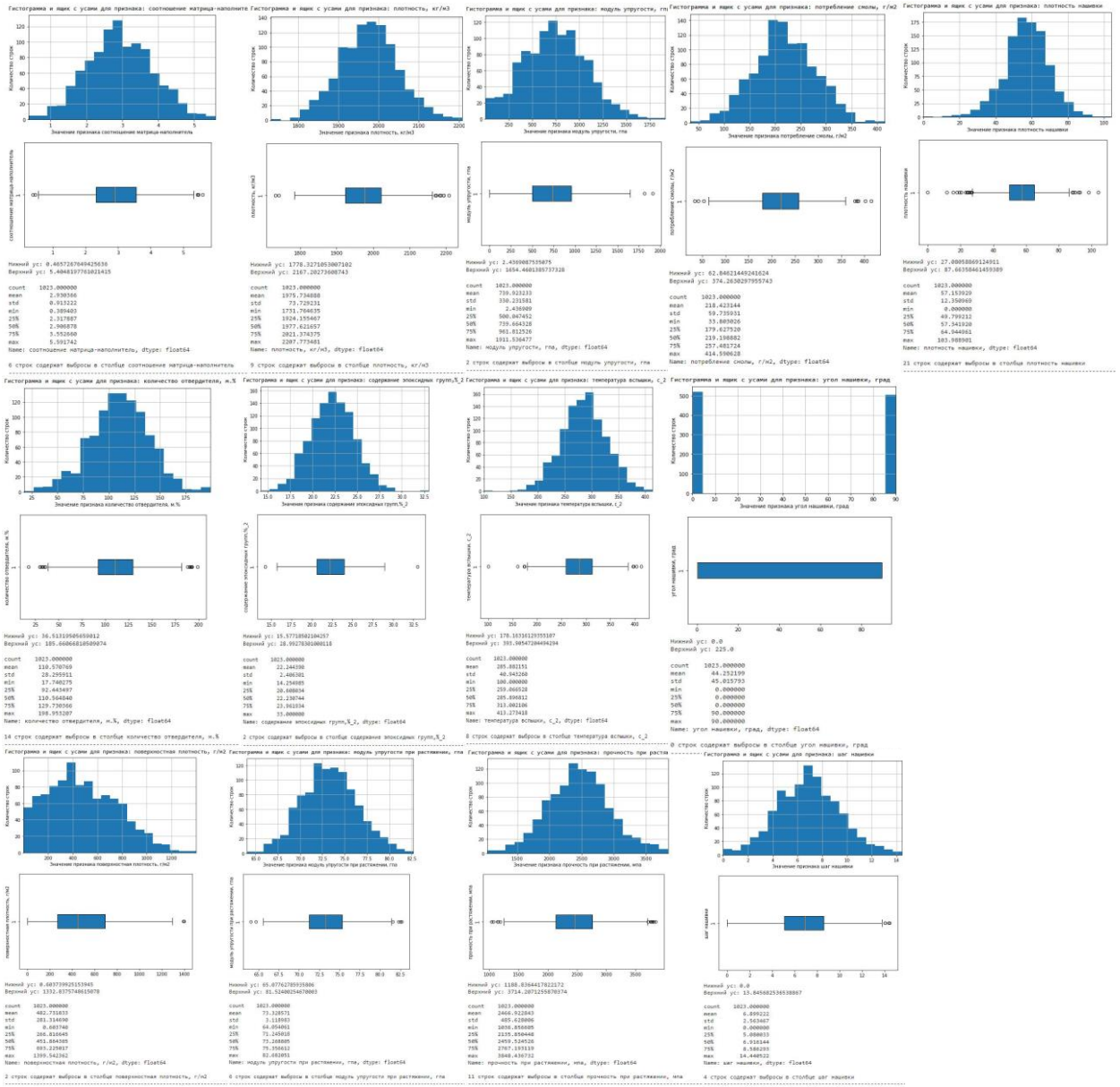
	0.0	1.0	2.0	3.0	4.0
Соотношение матрица-наполнитель	1.857143	1.857143	1.857143	1.857143	2.771331
Плотность, кг/м3	2030.000000	2030.000000	2030.000000	2030.000000	2030.000000
модуль упругости, ГПа	738.736842	738.736842	738.736842	738.736842	753.000000
Количество отвердителя, м.%	30.000000	50.000000	49.900000	129.000000	111.860000
Содержание эпоксидных групп,%_2	22.267857	23.750000	33.000000	21.250000	22.267857
Температура вспышки, С_2	100.000000	284.615385	284.615385	300.000000	284.615385
Поверхностная плотность, г/м2	210.000000	210.000000	210.000000	210.000000	210.000000
Модуль упругости при растяжении, ГПа	70.000000	70.000000	70.000000	70.000000	70.000000
Прочность при растяжении, МПа	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
Потребление смолы, г/м2	220.000000	220.000000	220.000000	220.000000	220.000000
Угол нашивки, град	0.000000	0.000000	0.000000	0.000000	0.000000
Шаг нашивки	4.000000	4.000000	4.000000	5.000000	5.000000
Плотность нашивки	57.000000	60.000000	70.000000	47.000000	57.000000

Итоговая таблица "df" состоит из 13 признаков и 1023 наблюдений.

Пропусков в столбцах нет. Дубликатов в данных нет.
Все признаки имеют тип данных float64.

В итоговой таблице 1023 объекта и 13 признаков

Разведочный анализ данных

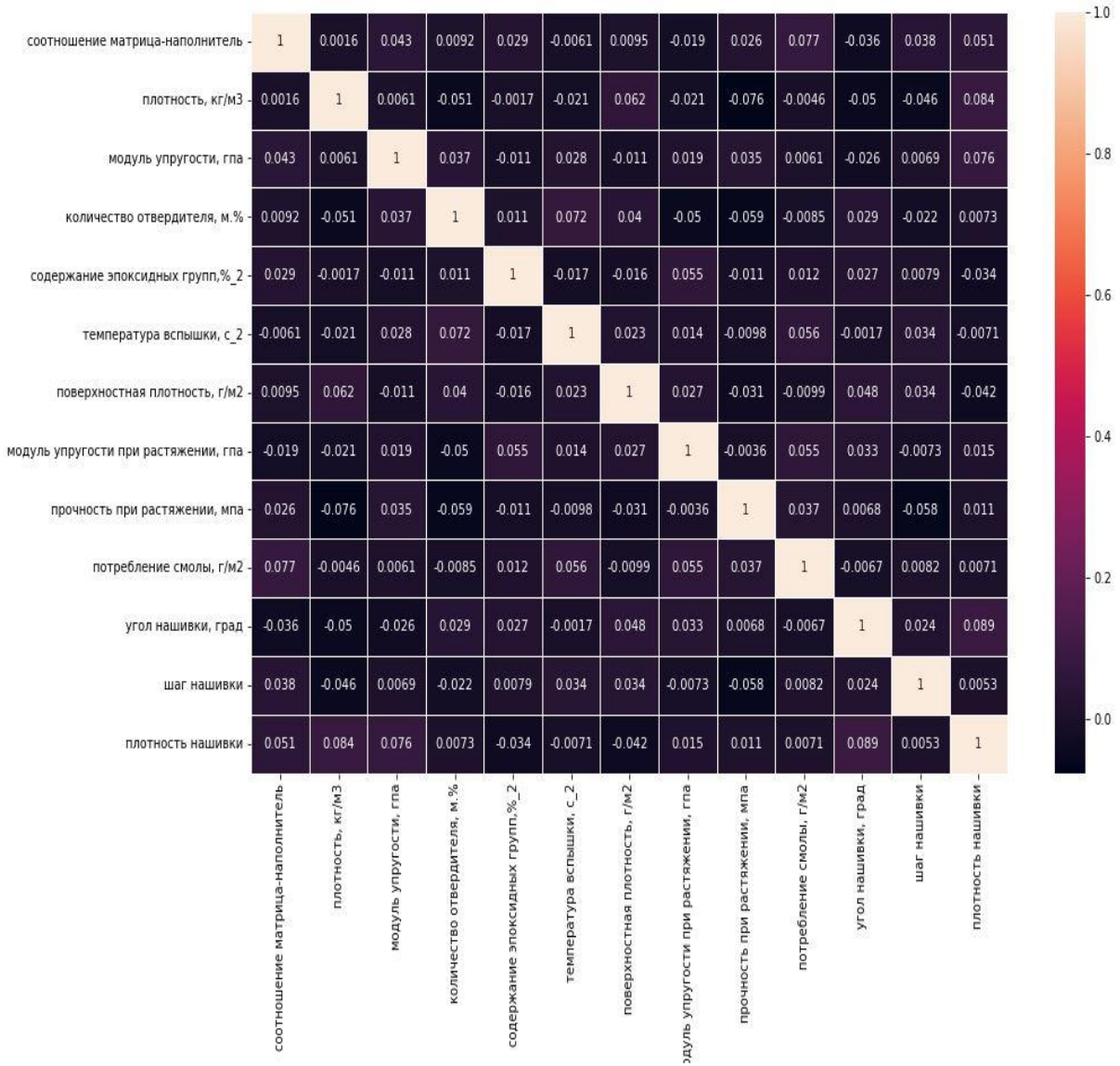


В данных присутствуют выбросы.

Всего строк с выбросами 93 (9,09% от всех данных.

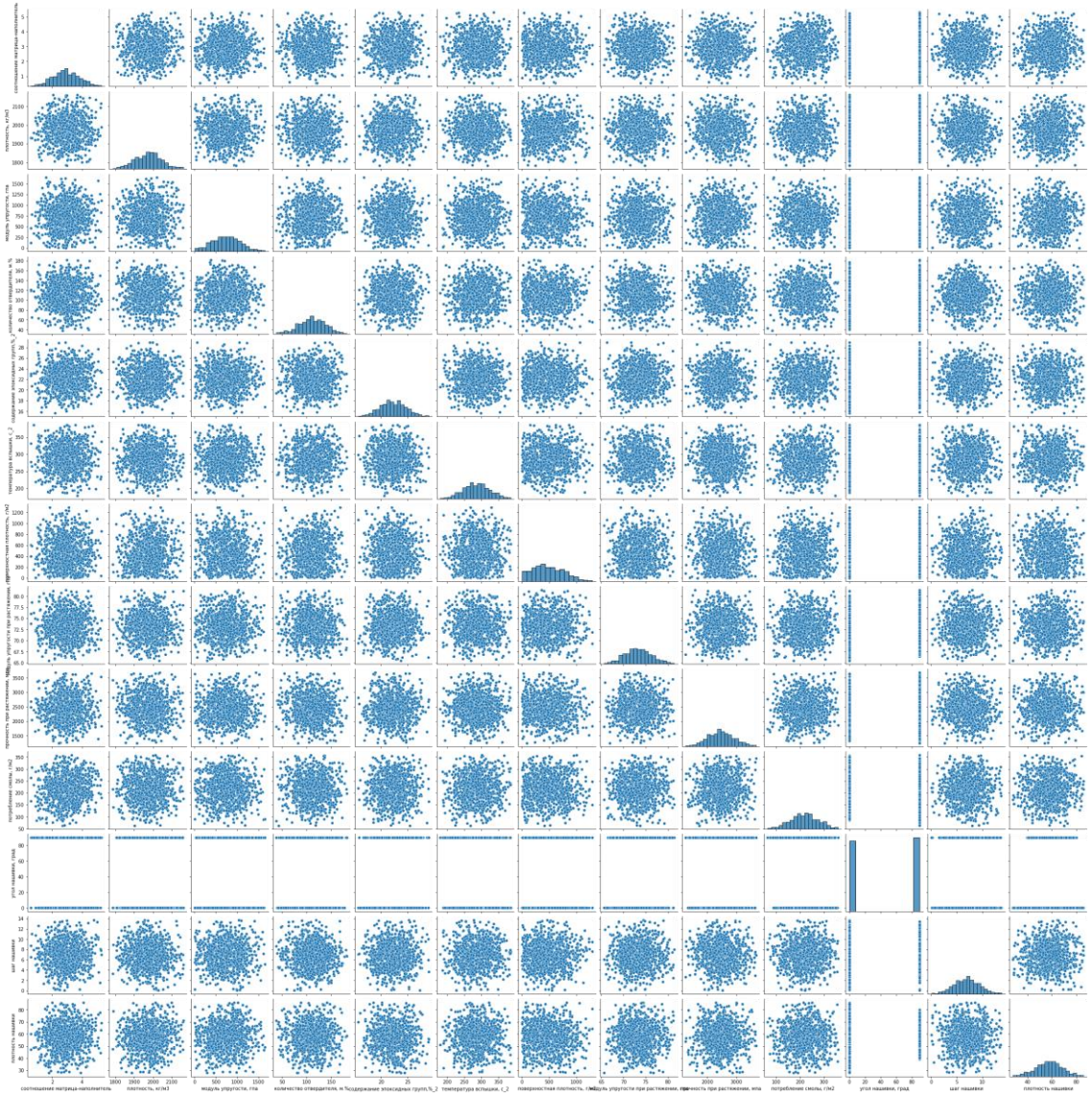
У всех переменных распределение близко к нормальному распределению.

Разведочный анализ данных



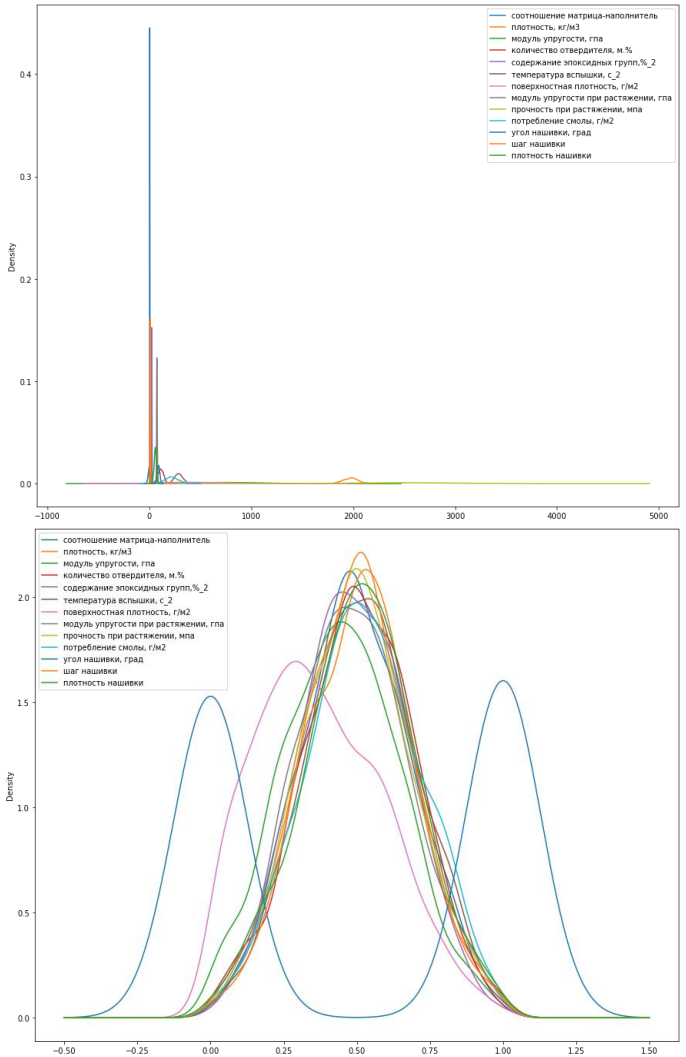
Корреляция переменных очень низкая.

Разведочный анализ данных



Линейная зависимость отсутствует.

Предобработка данных



	1.0	3.0
соотношение матрица-наполнитель	1.857143	1.857143
плотность, кг/м3	2030.000000	2030.000000
модуль упругости, гпа	738.736842	738.736842
количество отвердителя, м.%	50.000000	129.000000
содержание эпоксидных групп,%_2	23.750000	21.250000
температура вспышки, с_2	284.615385	300.000000
поверхностная плотность, г/м2	210.000000	210.000000
модуль упругости при растяжении, гпа	70.000000	70.000000
прочность при растяжении, мпа	3000.000000	3000.000000
потребление смолы, г/м2	220.000000	220.000000
угол нашивки, град	0.000000	0.000000
шаг нашивки	4.000000	5.000000
плотность нашивки	60.000000	47.000000

	0	1
соотношение матрица-наполнитель	0.274768	0.274768
плотность, кг/м3	0.651097	0.651097
модуль упругости, гпа	0.446422	0.446422
количество отвердителя, м.%	0.079153	0.630983
содержание эпоксидных групп,%_2	0.607435	0.418887
температура вспышки, с_2	0.509164	0.583596
поверхностная плотность, г/м2	0.161539	0.161539
модуль упругости при растяжении, гпа	0.280303	0.280303
прочность при растяжении, мпа	0.717396	0.717396
потребление смолы, г/м2	0.529221	0.529221
угол нашивки, град	0.000000	0.000000
шаг нашивки	0.289334	0.362355
плотность нашивки	0.552440	0.328767

Данные прошли нормализацию методом MinMaxScaler().

Обучение моделей

Делим нормализованный датасет на обучающую и тестовую выборки в соотношении 70:30.

Для прогнозирования модуля упругости при растяжении и прочности при растяжении были использованы следующие модели:

1. LinearRegression;
2. Ridge;
3. Lasso;
4. DecisionTreeRegressor;
5. RandomForestRegressor;
6. Support Vector Regression;
7. XGBoost;
8. CatBoostRegressor;
9. Neural network;
10. DummyRegressor.

Обучение моделей

```
results_1.sort_values('MSE модели').style.apply(highlight_min, axis=None, subset=['MSE модели', 'r2 модели'])
```

	Модель	MSE модели	r2 модели	Время обучения, с	Время предсказания, с	Общее время, с	Подбор гиперпараметров	Значения гиперпараметров	Данные
21	Neural network	0.0261	-0.0135	-	-	-	-	-	df_s
12	Forest_GridSearch	0.0290	-0.0302	0.02	0.00	0.02	да	{'max_depth': 1, 'max_features': 'log2', 'max_leaf_nodes': None, 'min_samples_leaf': 9, 'n_estimators': 11}	df_s_1
20	CatBoostRegressor_GridSearch	0.0290	-0.0306	0.13	0.00	0.14	да	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 20}	df_s_1
19	XGBoost_GridSearch	0.0291	-0.0312	0.01	0.00	0.01	да	{'learning_rate': 0.1, 'max_depth': 1, 'n_estimators': 10}	df_s_1
16	svr_lin_GridSearch	0.0291	-0.0320	0.00	0.00	0.00	да	{'C': 11, 'degree': 1, 'epsilon': 0.8}	df_s_1
15	svr_rbf_GridSearch	0.0291	-0.0320	0.00	0.00	0.00	да	{'C': 1, 'cache_size': 200, 'degree': 1, 'epsilon': 1.1}	df_s_1
8	Lasso_GridSearch	0.0291	-0.0323	0.00	0.00	0.00	да	{'alpha': 1, 'copy_X': True, 'fit_intercept': True, 'max_iter': 1, 'normalize': True, 'tol': 0.001}	df_s_1
0	DummyRegressor	0.0291	-0.0323	0.00	0.00	0.00	нет	-	df_s_1
6	Ridge_GridSearch	0.0291	-0.0325	0.00	0.00	0.01	да	{'alpha': 10, 'copy_X': True, 'fit_intercept': True, 'max_iter': None, 'normalize': True, 'solver': 'sag', 'tol': 0.09}	df_s_1
3	Lasso	0.0291	-0.0323	0.00	0.00	0.00	нет	-	df_s_1
7	SGDRegressor_GridSearch	0.0291	-0.0311	0.00	0.00	0.00	да	{'alpha': 2, 'eta0': 0.01, 'shuffle': False, 'warm_start': True}	df_s_1
11	Tree_GridSearch	0.0293	-0.0403	0.00	0.00	0.00	да	{'max_depth': None, 'max_leaf_nodes': 2, 'min_samples_leaf': 1}	df_s_1

Обучение моделей

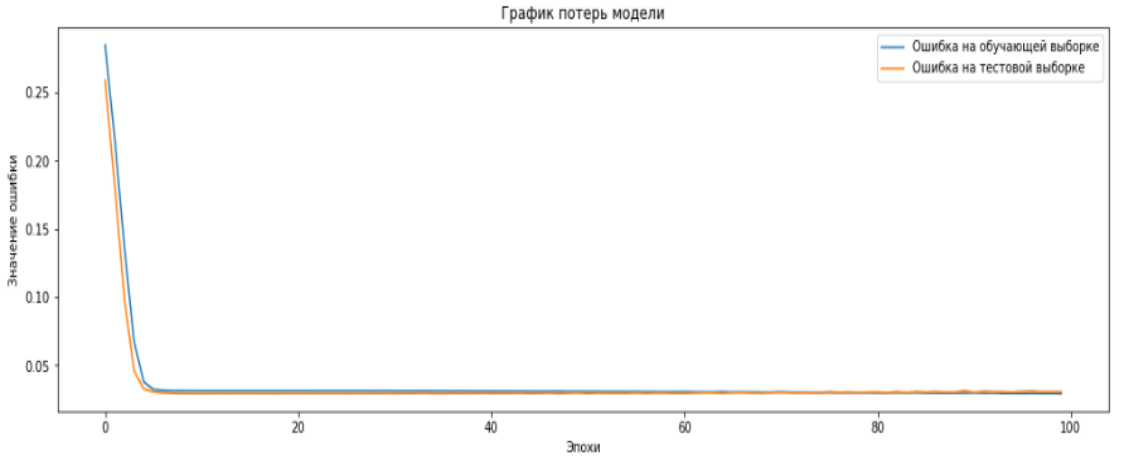
```
results_2.sort_values('MSE модели').style.apply(highlight_min, axis=None, subset=['MSE модели', 'r2 модели'])
```

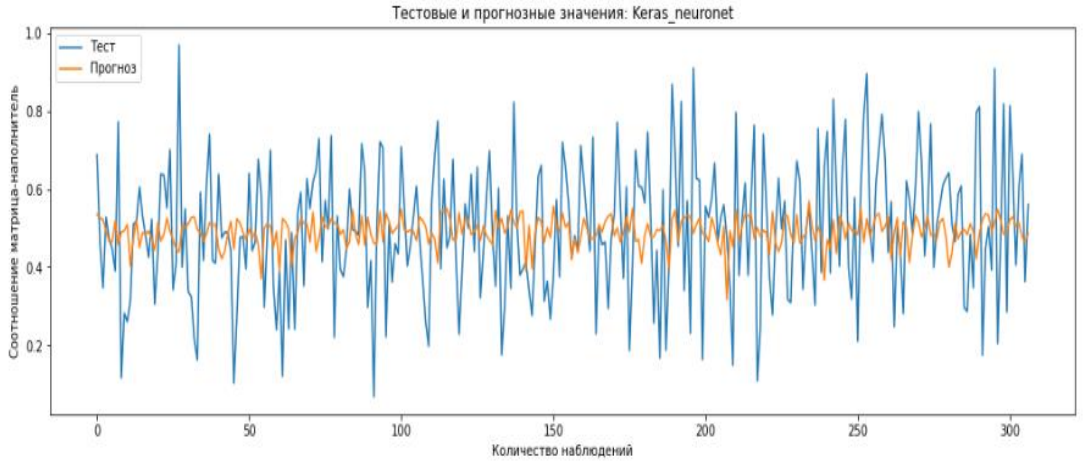
	Модель	MSE модели	r2 модели	Время обучения, с	Время предсказания, с	Общее время, с	Подбор гиперпараметров	Значения гиперпараметров	Данные
19	XGBoost_GridSearch_1	0.0289	0.0049	0.07	0.00	0.07	да	{'learning_rate': 0.2, 'max_depth': 1, 'n_estimators': 80}	df_s_1
20	CatBoostRegressor_GridSearch_1	0.0291	-0.0026	0.09	0.00	0.09	да	{'learning_rate': 0.3, 'max_depth': 1, 'n_estimators': 60}	df_s_1
21	Neural network_1	0.0296	-0.0015	-	-	-	-	-	df_s
6	Ridge_GridSearch_1	0.0296	-0.0179	0.01	0.00	0.01	да	{'alpha': 10, 'copy_X': True, 'fit_intercept': True, 'max_iter': None, 'normalize': False, 'solver': 'sag', 'tol': 0.09}	df_s_1
1	lr_1	0.0297	-0.0247	0.00	0.00	0.01	нет	-	df_s_1
2	Ridge_1	0.0297	-0.0229	0.00	0.00	0.01	нет	-	df_s_1
5	lr_GridSearch_1	0.0297	-0.0247	0.01	0.00	0.01	да	{'copy_X': True, 'fit_intercept': True, 'n_jobs': None, 'normalize': False, 'positive': False}	df_s_1
14	svr_lin_1	0.0297	-0.0231	0.10	0.02	0.11	нет	-	df_s_1
7	SGDRegressor_GridSearch_1	0.0299	-0.0267	0.01	0.00	0.01	да	{'alpha': 3, 'eta0': 0.01, 'shuffle': True, 'warm_start': True}	df_s_1
16	svr_lin_GridSearch_1	0.0300	-0.0291	0.00	0.00	0.01	да	{'C': 11, 'degree': 1, 'epsilon': 0.8}	df_s_1
15	svr_rbf_GridSearch_1	0.0300	-0.0291	0.01	0.00	0.01	да	{'C': 1, 'cache_size': 200, 'degree': 1, 'epsilon': 1.1}	df_s_1
12	Forest_GridSearch_1	0.0300	-0.0288	0.68	0.04	0.72	да	{'max_depth': None, 'max_leaf_nodes': 2, 'min_samples_leaf': 7, 'n_estimators': 122}	df_s_1

Обучение моделей

```
model_80.summary()
```

```
Model: "sequential_79"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_470 (Dense)           (None, 64)                832

 dense_471 (Dense)           (None, 64)                4160

 dense_472 (Dense)           (None, 32)                2080

 dense_473 (Dense)           (None, 1)                 33

=================================================================
Total params: 7,105
Trainable params: 7,105
Non-trainable params: 0
_________________________________________________________________
```





Архитектура модели состоит из пяти слоёв Dense:

входного, трех скрытых полносвязных и выходного.

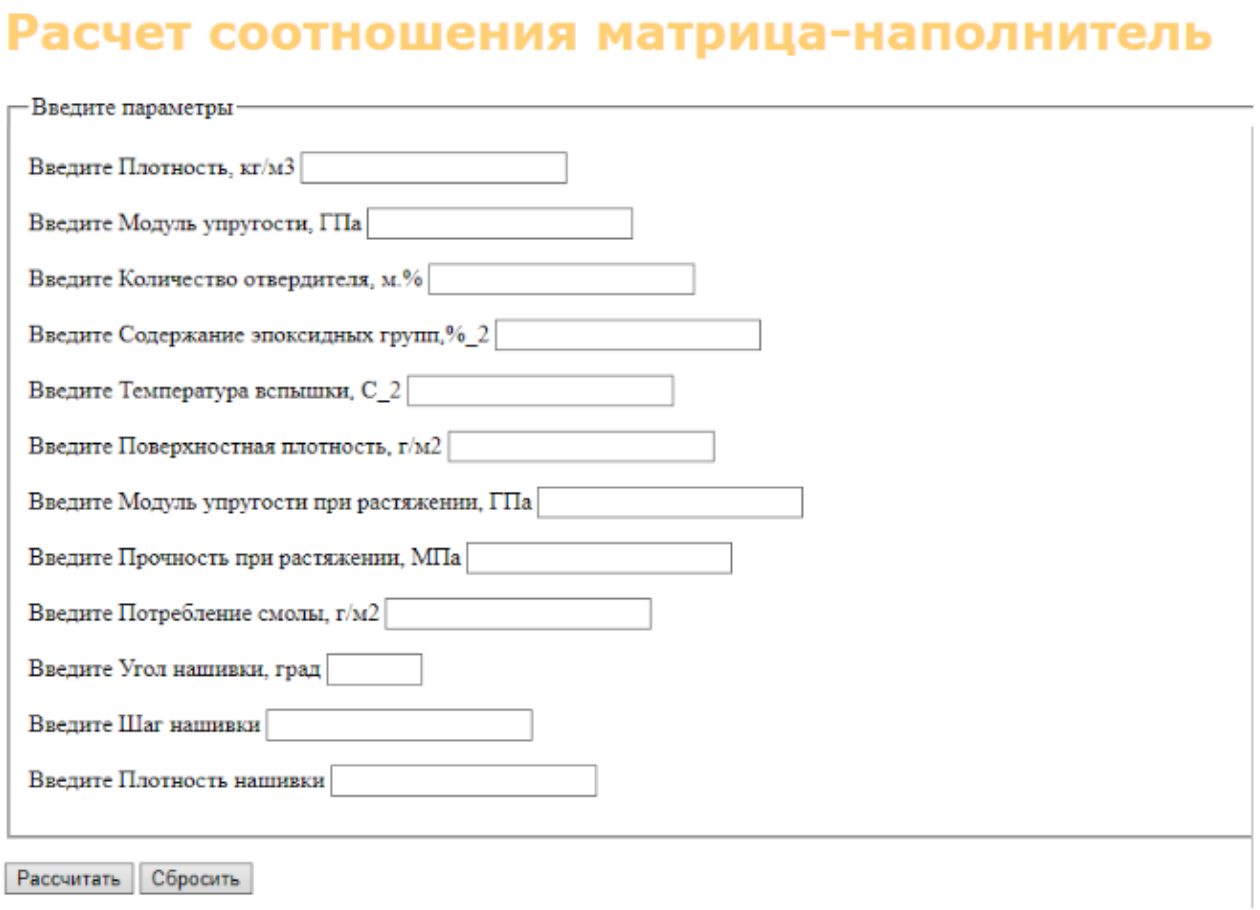
Функция активации 'relu' для скрытых слоев.

оптимизатор: Adam, loss-функция: 'mean_squared_error'.

Входной слой с 12 нейронами по количеству признаков.

Выходной слой с 1 нейроном для 1 признака. Нейронов в скрытых слоях: 64, 64 и 32.

Разработка приложения



Было разработано приложение для прогноза «Соотношение матрица-наполнитель» на основе разработанной нейронной сети. Приложение приложение написано при помощи фреймворка Flask. Для прогноза необходимо ввести 12 основных параметров.

Заключение

Качество предсказаний всех моделей получилось низкое. Не удалось найти модели, которые хорошо бы предсказывали целевые признаки. Большинство моделей показали метрики близкие или такие же как у модели предсказывающей среднее.

Для успешного прогнозирования данных недостаточно, около 1000 строк это слишком мало, так как даже удаление выбросов повлияло на метрику в худшую сторону, возможно также, недостаточно признаков. Также необходима дополнительная информации о зависимости признаков с точки зрения физики процесса. Возможно, стоит применить более сложные модели для прогнозирования.

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