Образовательный центр МГТУ им. Н.Э. Баумана

Выпускная квалификационная работа по курсу "Data Science"

Слушатель: Хайрутдинов Рамиль

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов)

Постановка задачи и план работы

- изучить предметную область
- провести разведочный анализ данных
- разделить данные на тренировочную и тестовую выборки
- выполнить препроцессинг (предобаботку)
- 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

Разведочный анализ данных

X_bp (матрица из базальтопластика):

- признаков: 10 и индекс
- строк: 1023

X_nup (наполнитель из углепластика):

- признаков: 3 и индекс
- строк: 1040

Объединение с типом INNER по индексу, получилось:

- признаков: 13
- строк: 1023

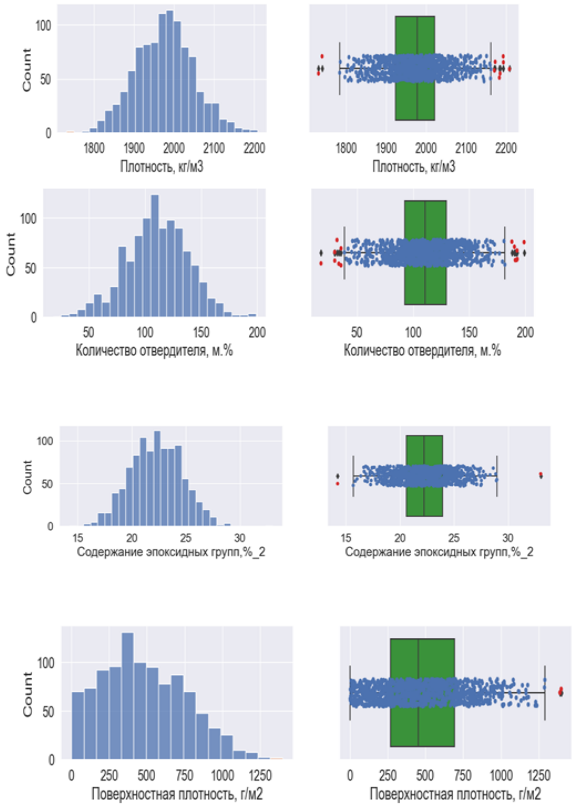
Разведочный анализ данных

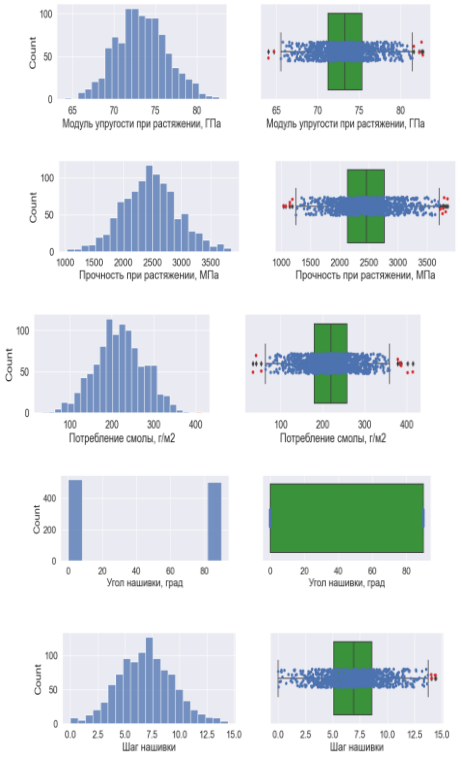
Название	Файл	Тип данных	Непустых значений	Уникальных значений
Соотношение матрица-наполнител	X_bp	float64	1023	1014
Плотность, кг/м3	X_bp	float64	1023	1013
модуль упругости, ГПа	X_bp	float64	1023	1020
Количество отвердителя, м.%	X_bp	float64	1023	1005
Содержание эпоксидных групп,%_2	X_bp	float64	1023	1004
Температура вспышки, С_2	X_bp	float64	1023	1003
Поверхностная плотность, г/м2	X_bp	float64	1023	1004
Модуль упругости при растяжении, ГПа	X_bp	float64	1023	1004
Прочность при растяжении, МПа	X_bp	float64	1023	1004
Потребление смолы, г/м2	X_bp	float64	1023	1003
Угол нашивки, град	X_nup	float64	1023	2
Шаг нашивки	X_nup	float64	1023	989
Плотность нашивки	X_nup	float64	1023	988

	Среднее	Стандартное отклонение	Минимум	Максимум	Медиана
Соотношение матрица-наполнитель	2.9304	0.9132	0.3894	5.5917	2.9069
Плотность, кг/м3	1975.7349	73.7292	1731.7646	2207.7735	1977.6217
модуль упругости, ГПа	739.9232	330.2316	2.4369	1911.5365	739.6643
Количество отвердителя, м.%	110.5708	28.2959	17.7403	198.9532	110.5648
Содержание эпоксидных групп, %_2	22.2444	2.4063	14.2550	33.0000	22.2307
Температура вспышки, С_2	285.8822	40.9433	100.0000	413.2734	285.8968
Поверхностная плотность, г/м2	482.7318	281.3147	0.6037	1399.5424	451.8644
Модуль упругости при растяжении, ГПа	73.3286	3.1190	64.0541	82.6821	73.2688
Прочность при растяжении, МПа	2466.9228	485.6280	1036.8566	3848.4367	2459.5245
Потребление смолы, г/м2	218.4231	59.7359	33.8030	414.5906	219.1989
Угол нашивки, град	44.2522	45.0158	0.0000	90.0000	0.0000
Шаг нашивки	6.8992	2.5635	0.0000	14.4405	6.9161
Плотность нашивки	57.1539	12.3510	0.0000	103.9889	57.3419

Пропусков нет

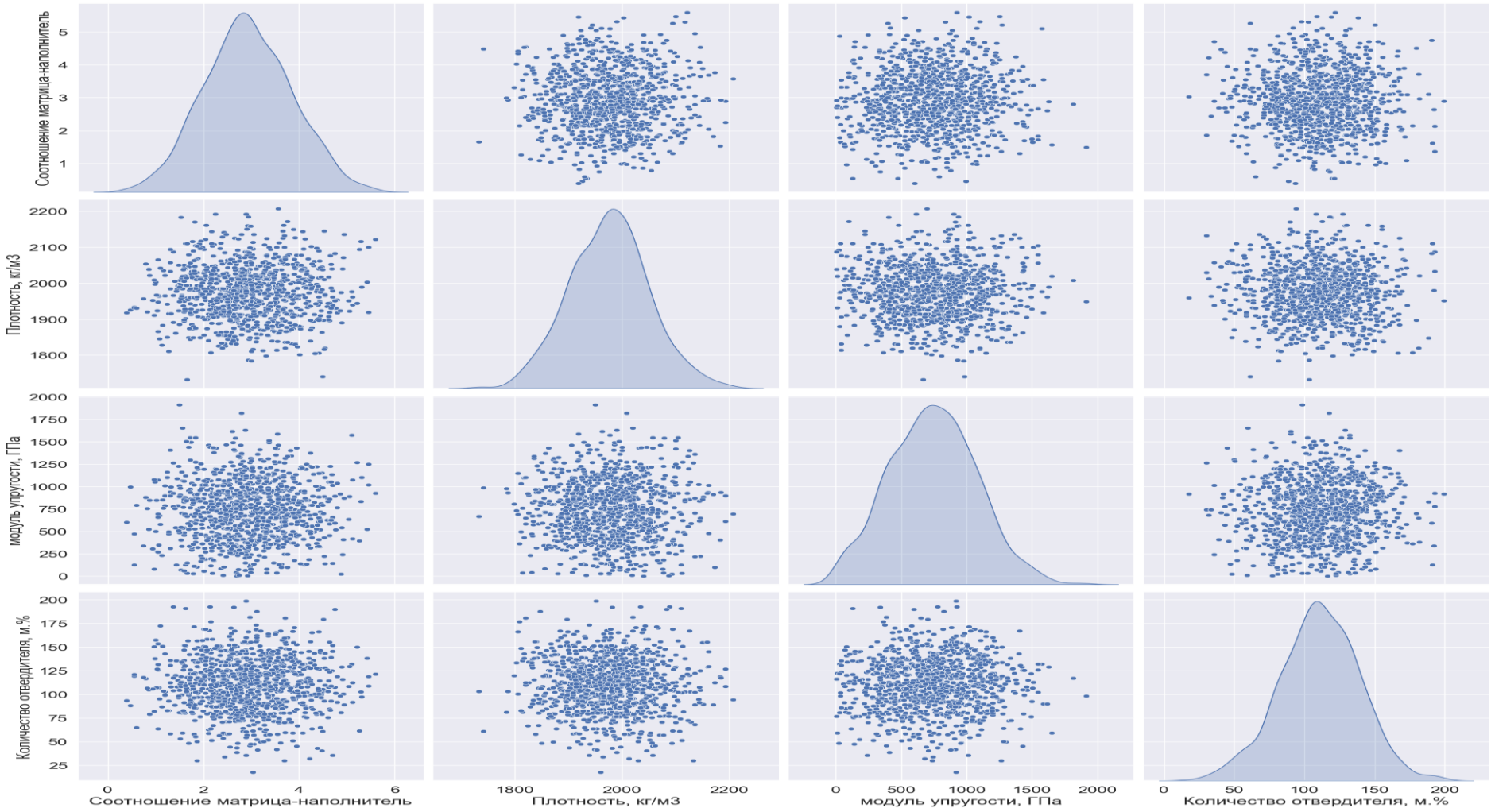
Гистограммы распределения и диаграммы “ящик с усами”



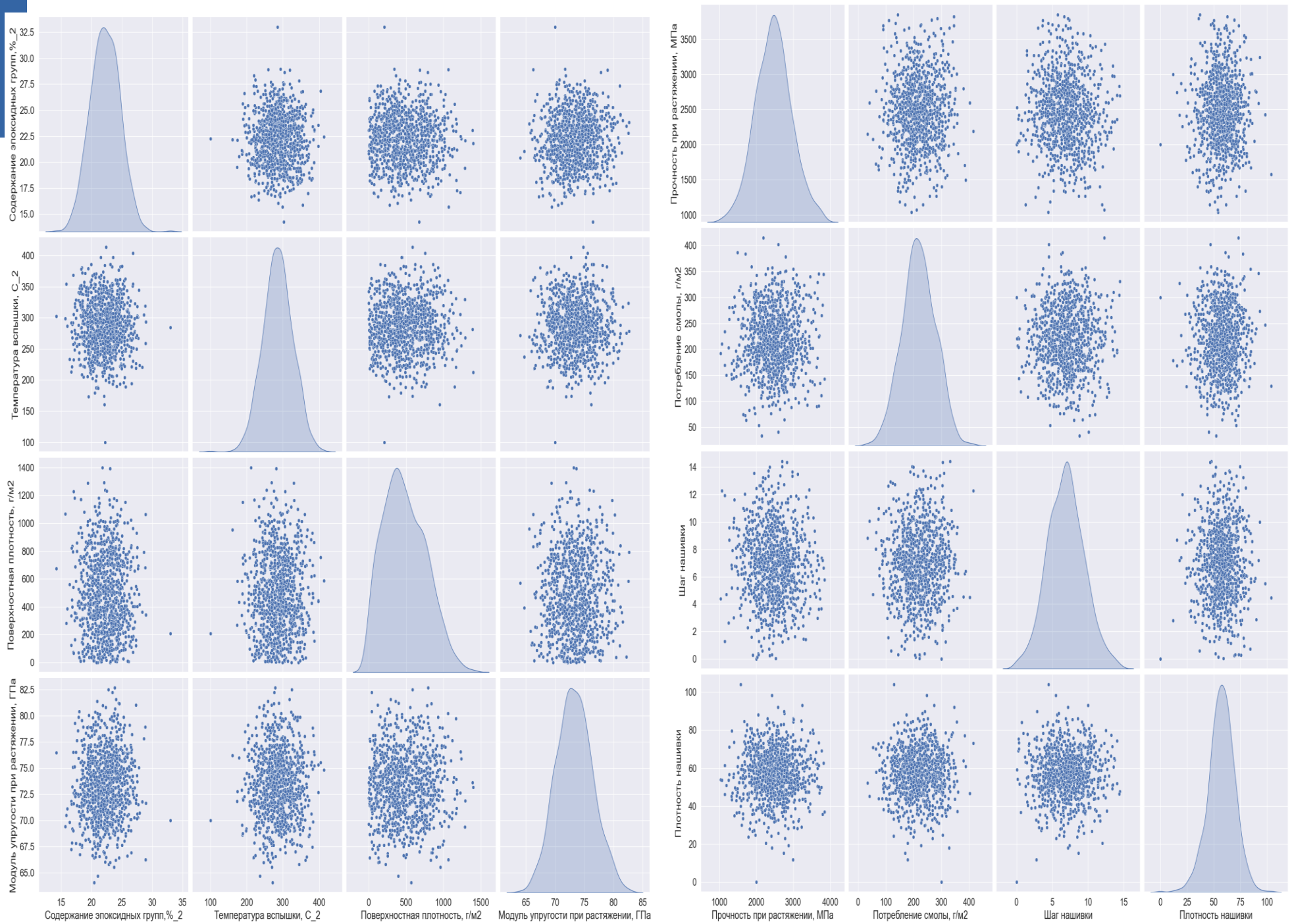


- Большинство — количественные, вещественные, положительные, нормально распределенные
- Угол нашивки — категориальный, бинарный

Попарные графики рассеяния точек



- Выбросы есть
- Зависимостей нет

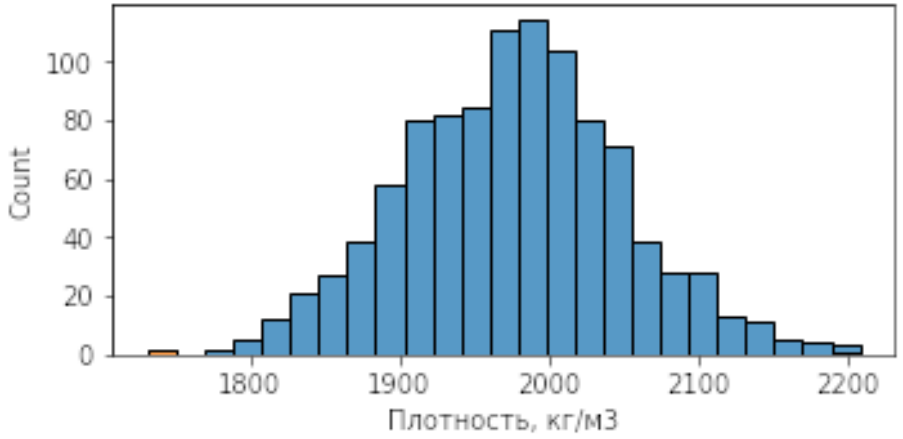

Содержание эпоксидных групп,%_2
Температура вспышки, С_2
Поверхностная плотность, г/м2
Модуль упругости при растяжении, ГПа
Прочность при растяжении, МПа
Потребление смолы, г/м2
Шаг нашивки
Плотность нашивки

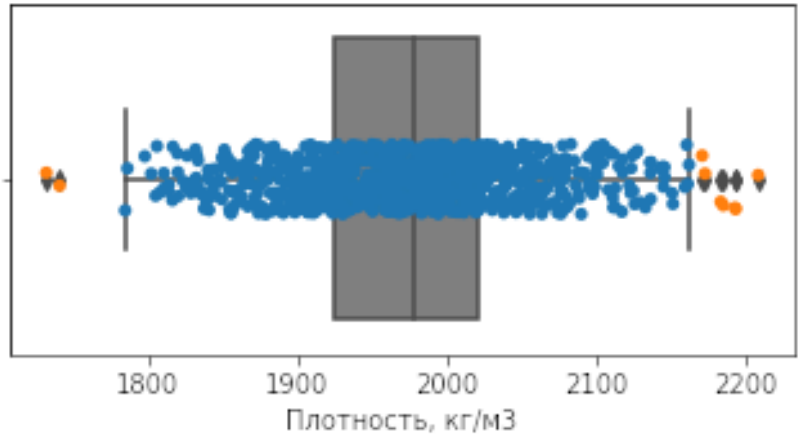
Выбросы

Найдено:

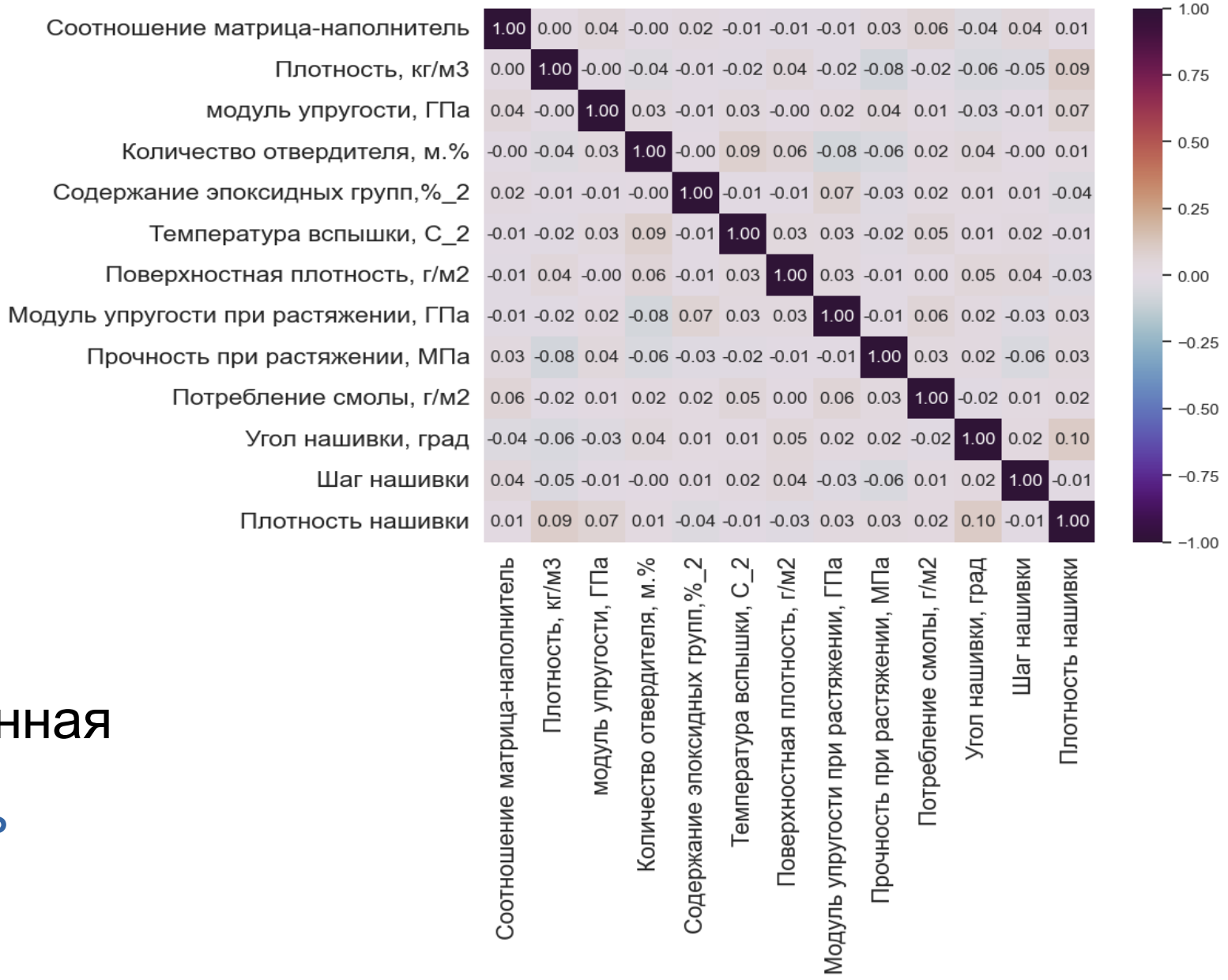
- методом 3-х сигм — 24 выброса
- методом межквартильных расстояний — 93 выброса

После удаления осталось 1000 строк





Матрица корреляции

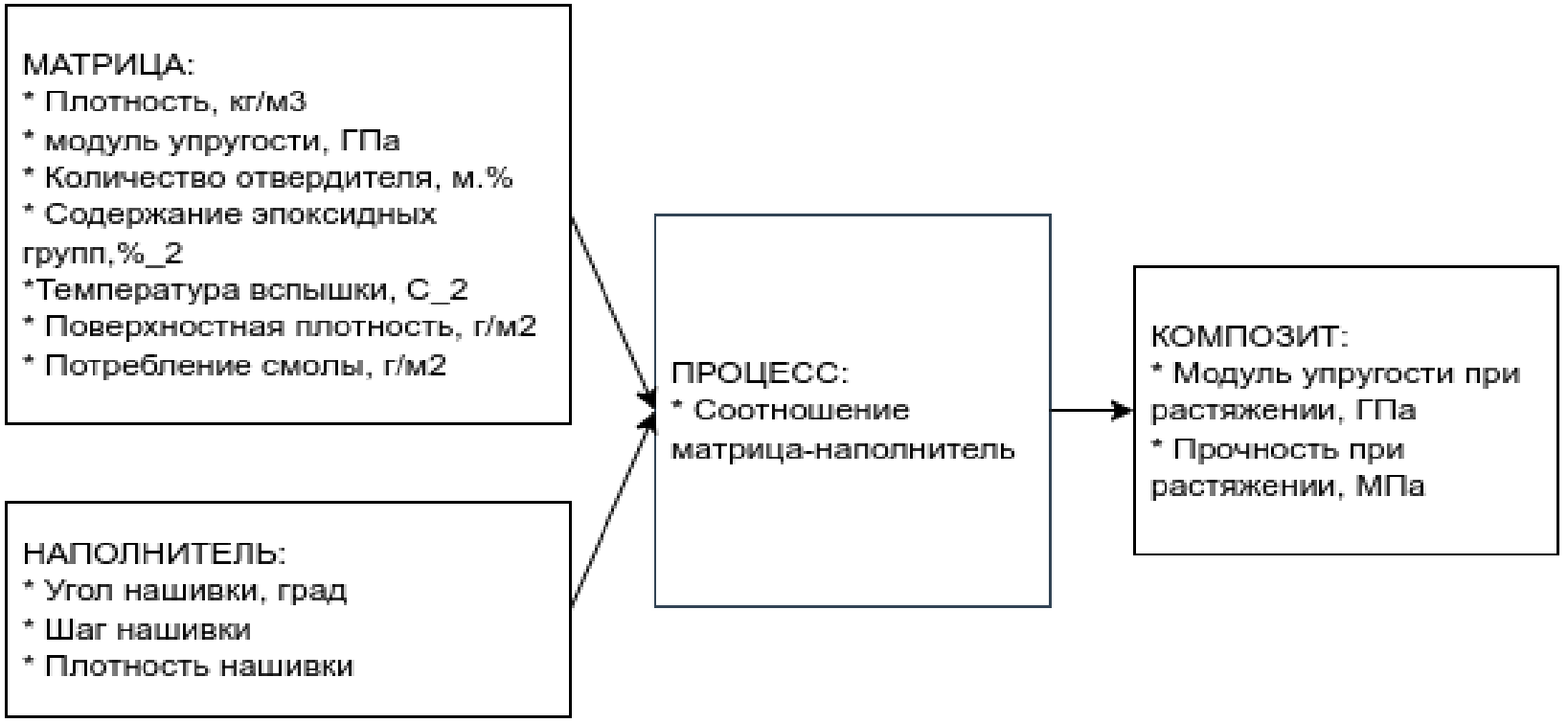


Корреляционная

зависимость

отсутствует

Предметная область: композитные материалы



Выходные переменные

```
# Описательная статистика выходной переменной
```

	Модуль упругости при растяжении, ГПа
min	64.054061
max	82.682051
mean	73.354026
std	3.066086

```
# Описательная статистика выходной переменной
```

	Прочность при растяжении, МПа
min	1071.123751
max	3848.436732
mean	2468.178562
std	487.297434

Соотношение матрица-наполнитель

	0
min	64.054061
max	82.682051
mean	73.372372
std	3.202078

Для каждого признака — отдельная модель

- модуль упругости при растяжении, Гпа (композит)=f(матрица, наполнитель, процесс);
- прочность при растяжении, Мпа(композит)=f(матрица, наполнитель, процесс);
- соотношение матрица-наполнитель (процесс) = f(матрица, наполнитель, композит).

Входные переменные

Значения признаков в разных диапазонах =>

необходим препроцессинг

- разделим на количественные и категориальные
- Категориальный один («Угол нашивки») - OrdinalEncoder
  - список значений стал [0, 1]
- количественные все остальные — StandardScaler
  - матожидание стало 0
  - стандартное отклонение стало 1
- создать объект-препроцесор, сохранить вместе с моделью
  - для train — fit_transform
  - для test — transform
  - для введенных данных — transform

Метрики качества

- R2 или коэффициент детерминации
- RMSE (Root Mean Squared Error) или корень из средней квадратичной ошибки
- MAE (Mean Absolute Error) или средняя абсолютная ошибка
- MAPE (Mean Absolute Percentage Error) или средняя абсолютная процентная ошибка
- max error или максимальная ошибка данной модели

Модели

- Линейная регрессия
- Лассо (LASSO) и гребневая (Ridge) регрессия
- Метод опорных векторов для регрессии
- Метод k-ближайших соседей
- Деревья решений
- Случайный лес
- Градиентный бустинг
- Нейронная сеть

Модель для модуля упругости при растяжении

Значения выхода от 64 до 83

По умолчанию →

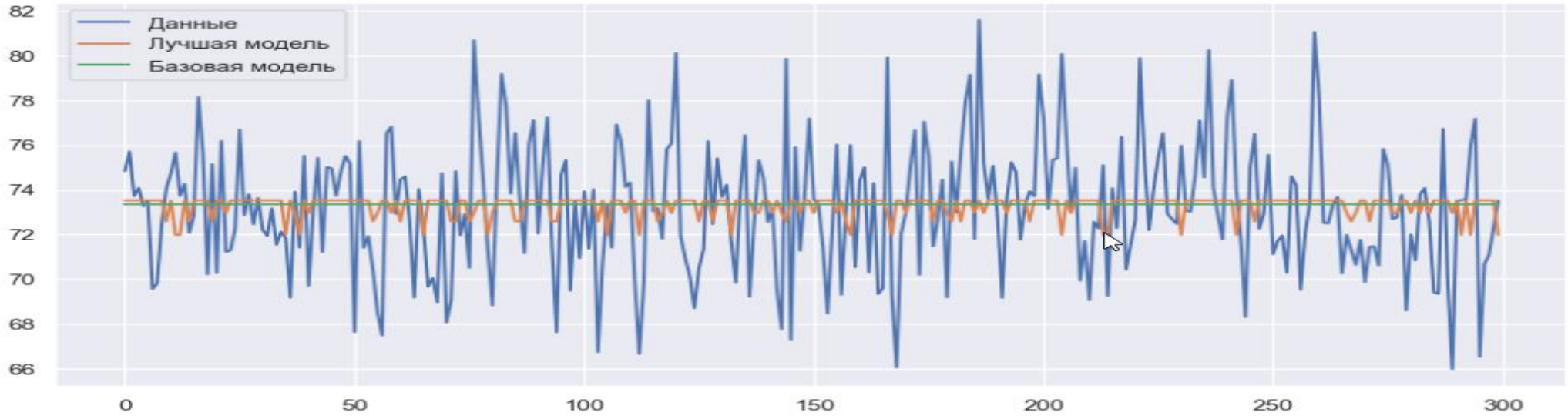
После подбора гиперпараметров ↓

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.018012	-3.196837	-2.589076	-0.035338	-7.858063
LinearRegression	-0.029339	-3.212486	-2.591121	-0.035370	-8.050245
Ridge	-0.029266	-3.212377	-2.591059	-0.035369	-8.049493
Lasso	-0.018012	-3.196837	-2.589076	-0.035338	-7.858063
SVR	-0.068756	-3.272522	-2.636092	-0.035968	-8.096169
KNeighborsRegressor	-0.228960	-3.508454	-2.826728	-0.038587	-8.809429
DecisionTreeRegressor	-1.245350	-4.727658	-3.755910	-0.051265	-12.535198
RandomForestRegressor	-0.089733	-3.304112	-2.642280	-0.036080	-8.323619

jt 63

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=70, positive=True, solver='lbfgs')	-0.022029	-3.201530	-2.577152	-0.035176	-7.959058
Lasso(alpha=0.1)	-0.019932	-3.199343	-2.581442	-0.035236	-7.930160
SVR(C=0.5, kernel='sigmoid')	-0.024054	-3.203086	-2.583316	-0.035243	-8.116314
KNeighborsRegressor(n_neighbors=29)	-0.059458	-3.259411	-2.638205	-0.036045	-7.956574
DecisionTreeRegressor(max_depth=1, max_features=1, random_state=4344, splitter='random')	-0.015694	-3.193718	-2.582909	-0.035254	-7.861263
RandomForestRegressor(bootstrap=False, criterion='absolute_error', max_depth=3, max_features=2, n_estimators=50, random_state=4344)	-0.030235	-3.215300	-2.594014	-0.035363	-8.035027

Модель для модуля упругости при растяжении



	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.003897	2.899450	2.293164	0.031448	8.222378
Лучшая модель (дерево решений)	-0.013159	2.912794	2.302184	0.031535	8.092000

	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

Значения выхода от 1036 до 3791

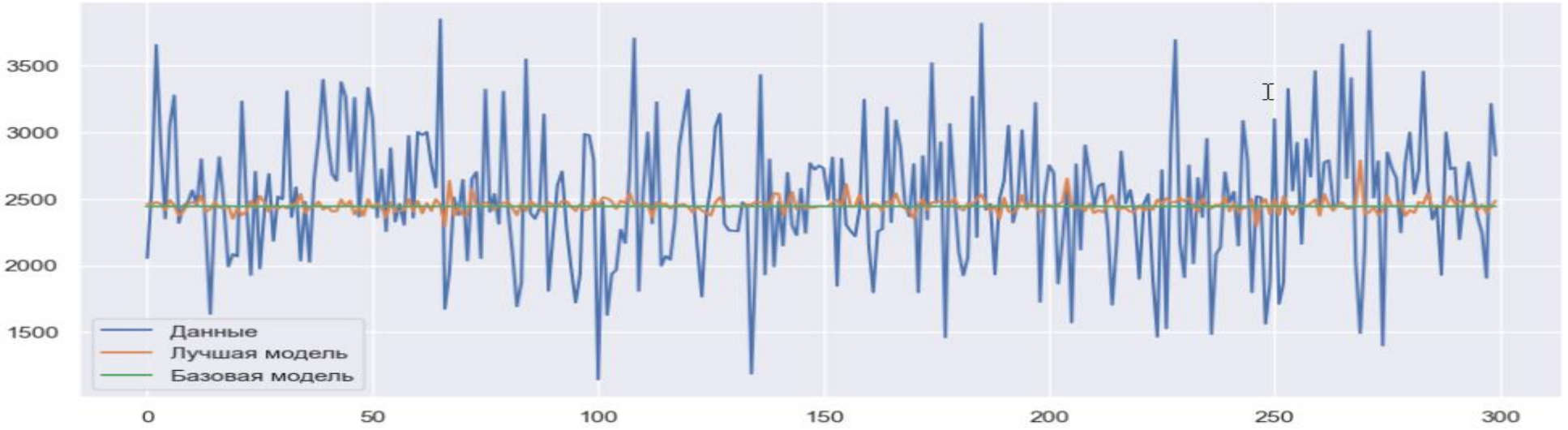
По умолчанию →

После подбора гиперпараметров ↓

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.023272	-479.709993	-381.342407	-0.169363	-1246.603286
LinearRegression	-0.034866	-482.179573	-384.522648	-0.170244	-1242.973850
Ridge	-0.034779	-482.159805	-384.504258	-0.170237	-1242.902268
Lasso	-0.034327	-482.063468	-384.364454	-0.170198	-1243.224304
SVR	-0.020411	-479.056248	-380.998666	-0.168651	-1242.263783
DecisionTreeRegressor	-1.103701	-684.934071	-554.567310	-0.239678	-1754.314464
GradientBoostingRegressor	-0.124182	-502.353407	-403.549852	-0.177907	-1304.367710

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=990, random_state=4344, solver='sag')	-0.020567	-479.004881	-381.234480	-0.169168	-1241.965662
Lasso(alpha=40)	-0.022584	-479.545607	-381.220708	-0.169277	-1247.335738
SVR(C=0.3)	-0.020458	-479.068243	-380.981776	-0.168647	-1242.642174
DecisionTreeRegressor(criterion='absolute_error', max_depth=2, max_features=3, random_state=4344, splitter='random')	-0.005679	-475.639357	-377.565204	-0.167369	-1242.778485
GradientBoostingRegressor(max_depth=1, max_features=1, random_state=4344)	-0.021589	-479.254818	-382.510490	-0.169329	-1229.007586

Модель для прочности при растяжении

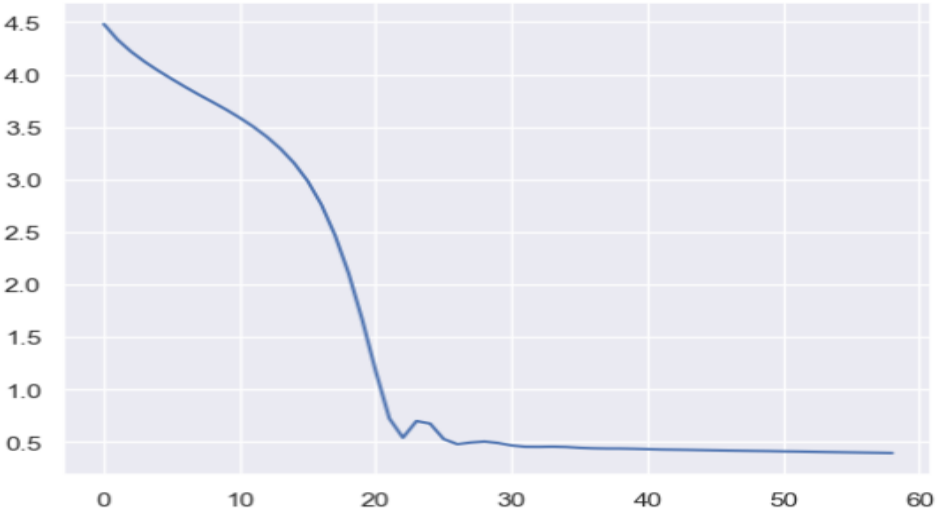


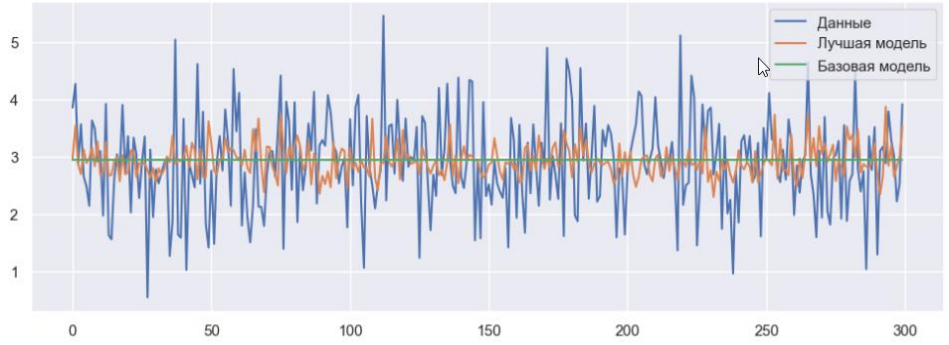
	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.016795	497.332204	385.915353	0.161962	1402.747538
Лучшая модель (градиентный бустинг)	-0.023595	498.992530	387.420047	0.163368	1392.474237

	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

MLPRegressor
из библиотеки sklearn





	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.011269	-0.911261	-0.737067	-0.299795	-2.684301
MLPRegressor	-0.052842	-0.929803	-0.751262	-0.306957	-2.790557

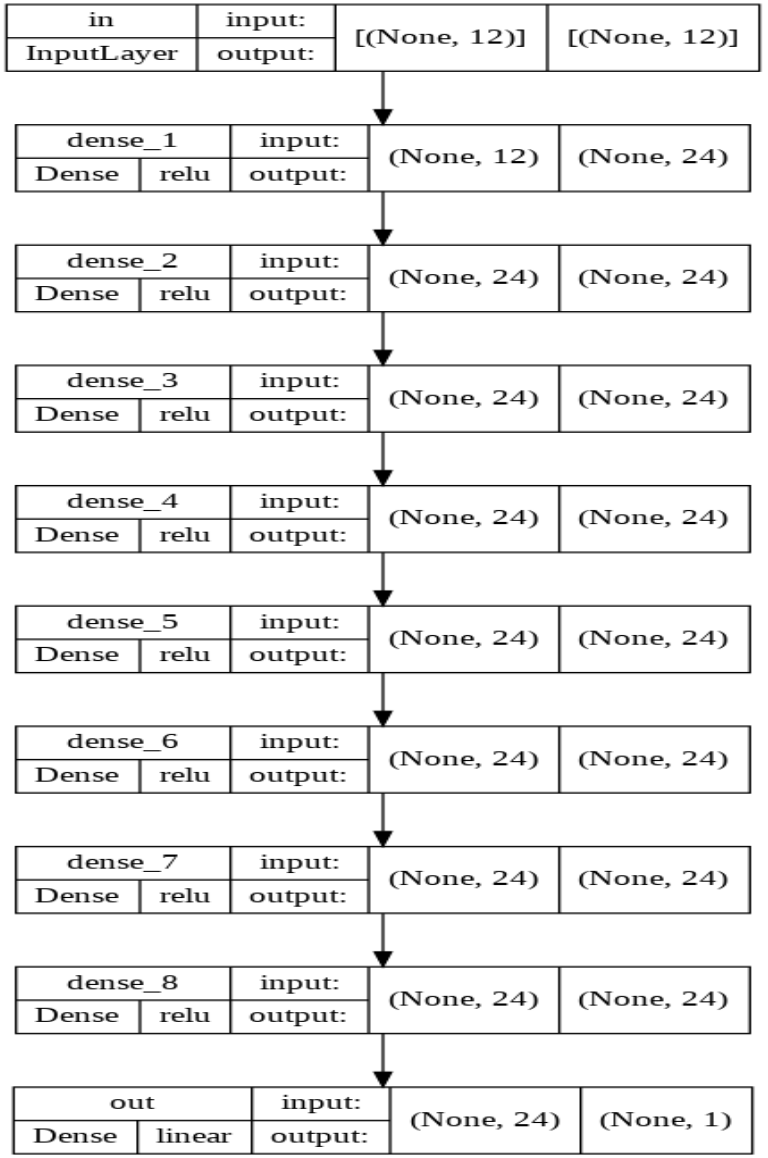
Значения выхода от 0.39 до 5.46

Модель для соотношения матрица-наполнитель

Нейросеть
из библиотеки
tensorflow

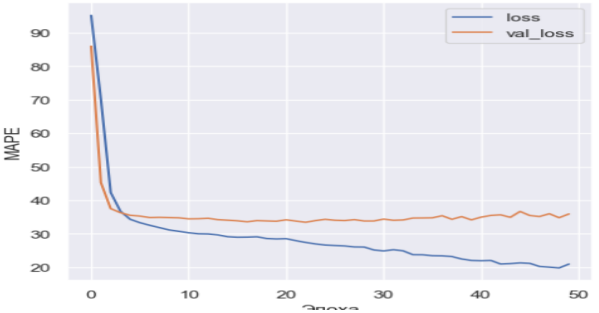
Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 24)	312
dense_2 (Dense)	(None, 24)	600
dense_3 (Dense)	(None, 24)	600
dense_4 (Dense)	(None, 24)	600
dense_5 (Dense)	(None, 24)	600
dense_6 (Dense)	(None, 24)	600
dense_7 (Dense)	(None, 24)	600
dense_8 (Dense)	(None, 24)	600
out (Dense)	(None, 1)	25

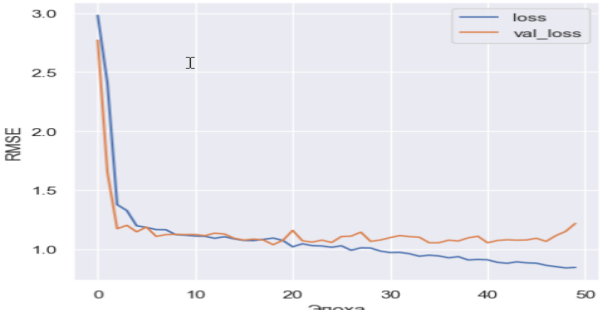
Total params: 4,537
Trainable params: 4,537
Non-trainable params: 0



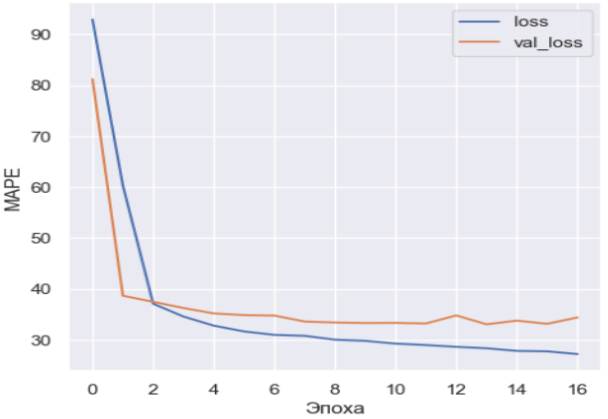
Модель для соотношения матрица-наполнитель

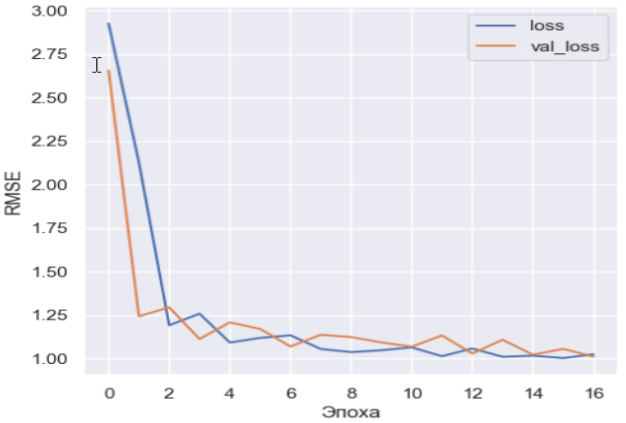
Обучение нейросети



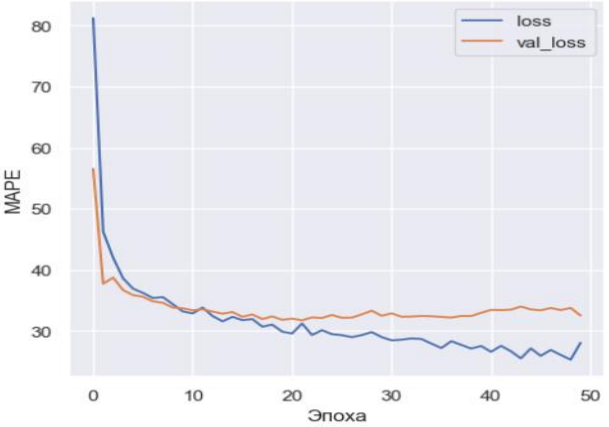


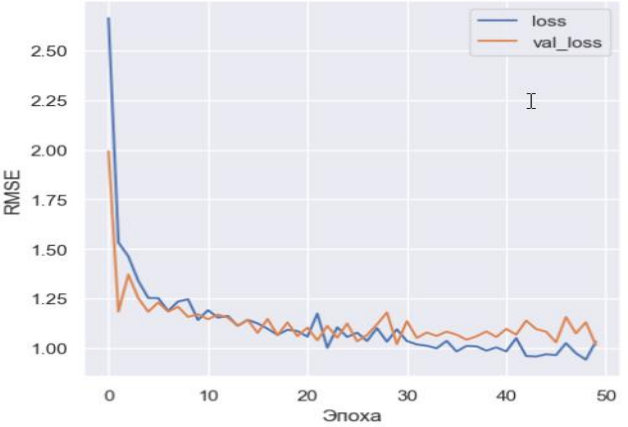
Борьба с переобучением: ранняя остановка





Борьба с переобучением: Dropout

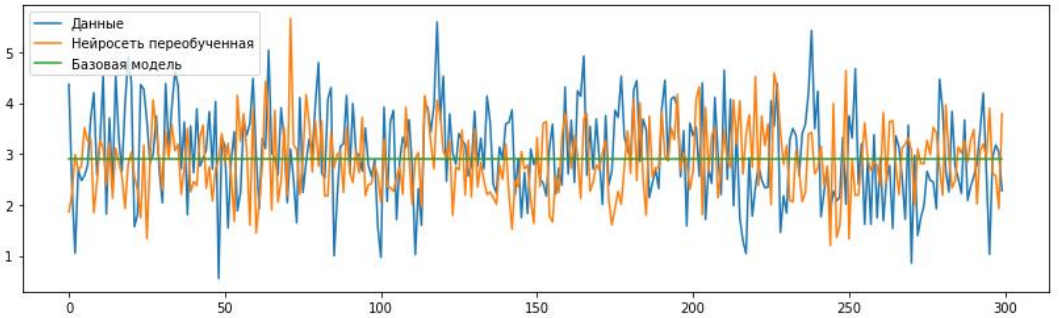


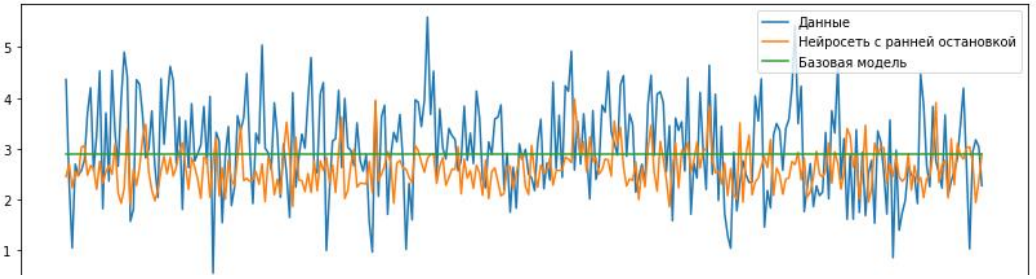


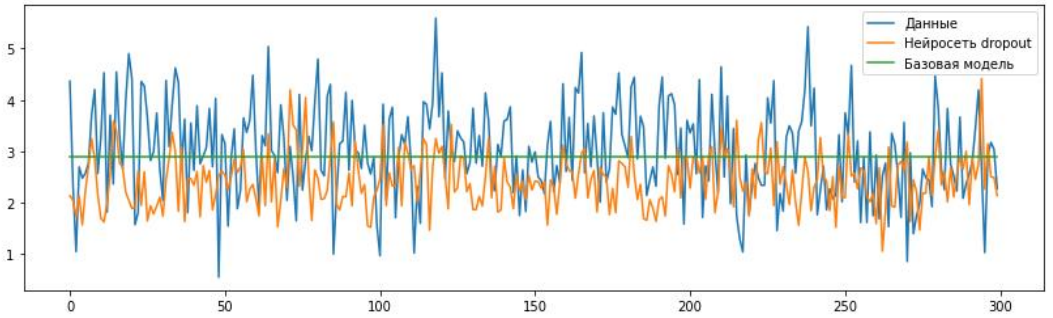
Модель для соотношения матрица-наполнитель

Значения выхода от 0.39 до 5.46

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.004784	0.837505	0.659057	0.287431	2.501931
Нейросеть переобученная	-1.055588	1.197897	0.958529	0.334777	3.424804
Нейросеть с ранней остановкой	-0.253462	0.935421	0.762494	0.303079	2.425718
Нейросеть dropout	-0.272050	0.942331	0.756197	0.293931	3.040035



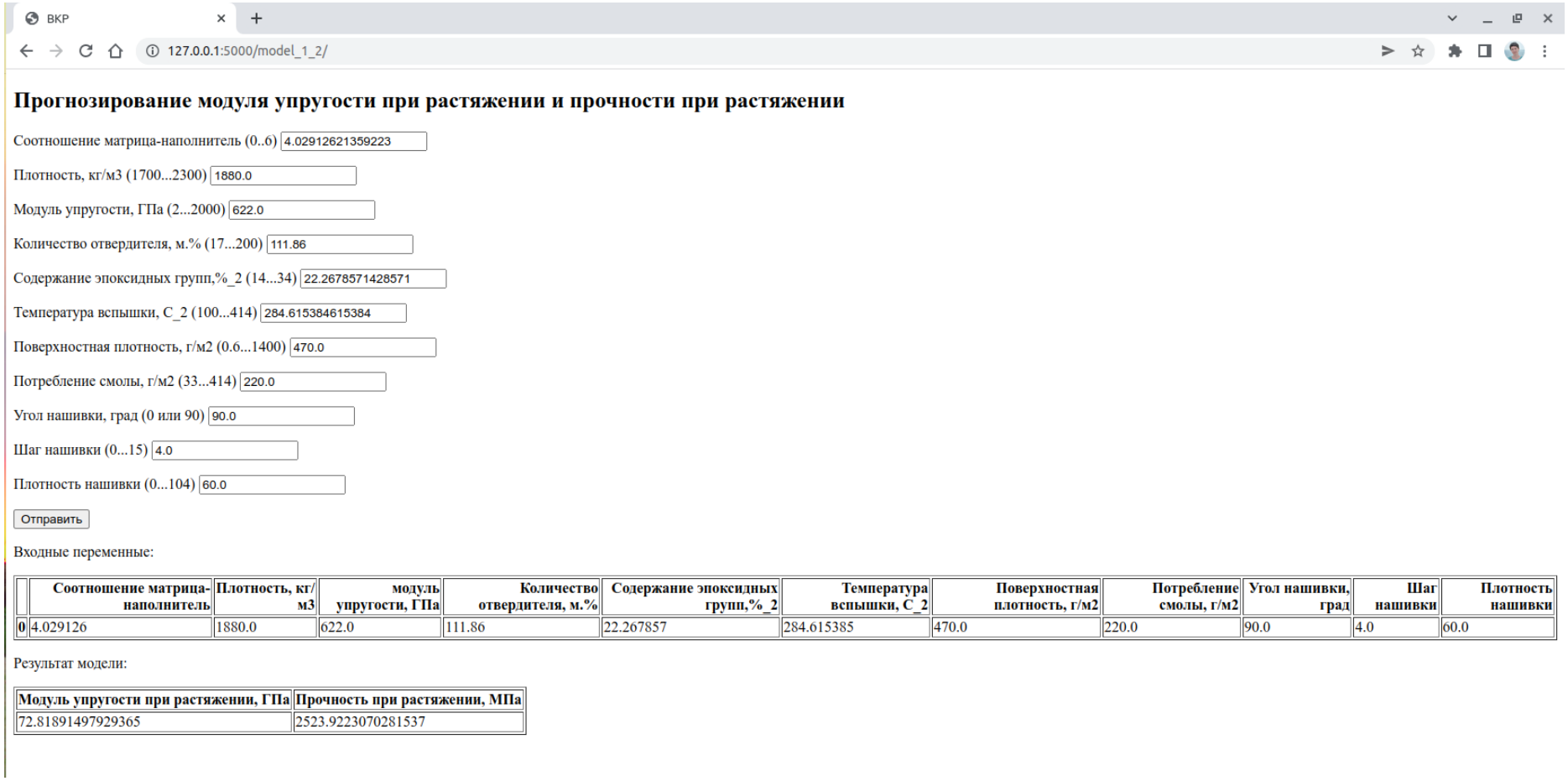




Выбираю нейросеть, обученную с ранней остановкой

	R2	RMSE	MAE	MAPE	max_error
Соотношение матрица-наполнитель, трениро...	-0.026908	0.949794	0.753971	0.296951	2.861537
Соотношение матрица-наполнитель, тестовый	-0.253462	0.935421	0.762494	0.303079	2.425718

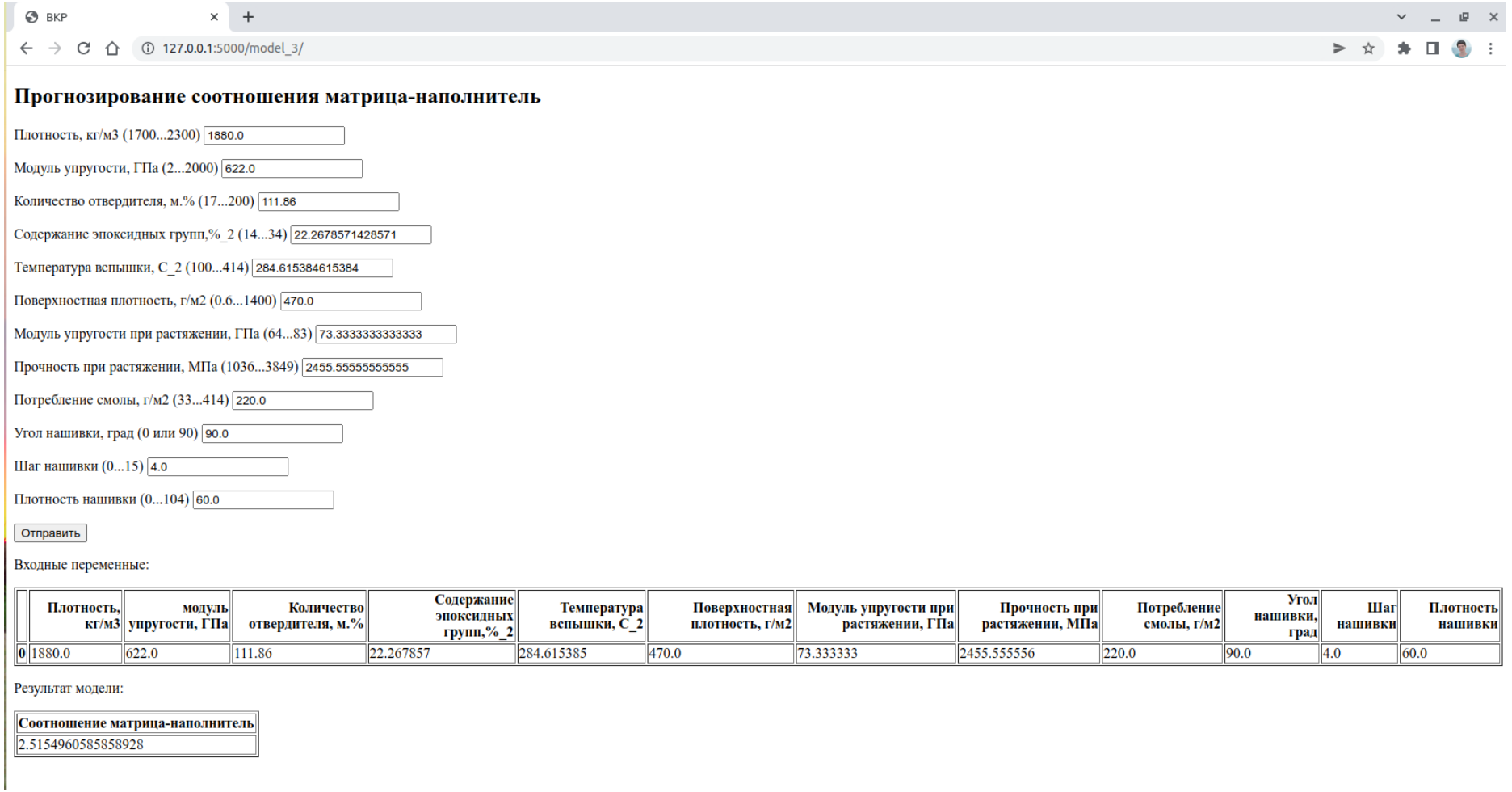
Разработка веб-приложения



	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	4.029126	1880.0	622.0	111.86	22.267857	284.615385	470.0	220.0	90.0	4.0	60.0

Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа
72.81891497929365	2523.9223070281537

Разработка веб-приложения



ВКР

127.0.0.1:5000/model_3/

Прогнозирование соотношения матрица-наполнитель

Плотность, кг/м3 (1700...2300) 1880.0

Модуль упругости, ГПа (2...2000) 622.0

Количество отвердителя, м.% (17...200) 111.86

Содержание эпоксидных групп,%_2 (14...34) 22.2678571428571

Температура вспышки, С_2 (100...414) 284.615384615384

Поверхностная плотность, г/м2 (0.6...1400) 470.0

Модуль упругости при растяжении, ГПа (64...83) 73.3333333333333

Прочность при растяжении, МПа (1036...3849) 2455.55555555555

Потребление смолы, г/м2 (33...414) 220.0

Угол нашивки, град (0 или 90) 90.0

Шаг нашивки (0...15) 4.0

Плотность нашивки (0...104) 60.0

Отправить

Входные переменные:

	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	1880.0	622.0	111.86	22.267857	284.615385	470.0	73.333333	2455.555556	220.0	90.0	4.0	60.0

Результат модели:

Соотношение матрица-наполнитель
2.515496058585928

Результаты

Цель задания решить не удалось

Дальнейшие поиски решения могли бы включать:

- консультации экспертов
- уточненную постановку задачи
- глубокое исследование первичных данных
- отбор признаков и уменьшение размерности
- эксперименты с градиентным бустингом
- углубление в нейросети

Спасибо за внимание!