



Образовательный центр МГТУ им. Н.Э. Баумана

Выпускная квалификационная работа по курсу "Data Science"

Тема: Прогнозирование конечных свойств
новых материалов (композиционных материалов)

Слушатель: Лапшин Олег

Постановка 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 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_br (матрица):

- признаков: 10 и индекс
- строк: 1023

X_nir (наполнитель):

- признаков: 3 и индекс
- строк: 1040

Объединение с типом INNER по индексу, получилось:

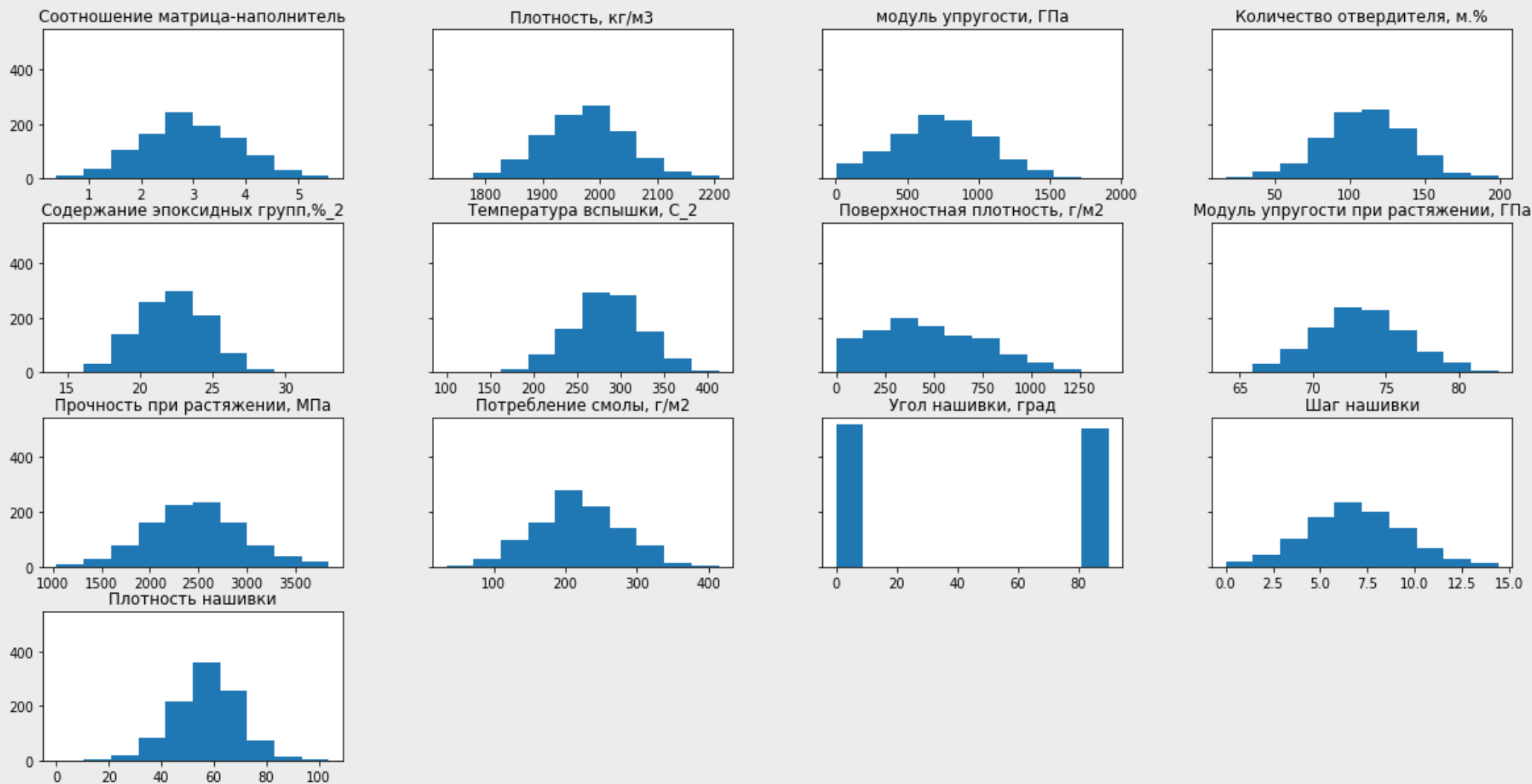
- признаков: 13
- строк: 1023

Разведочный анализ данных

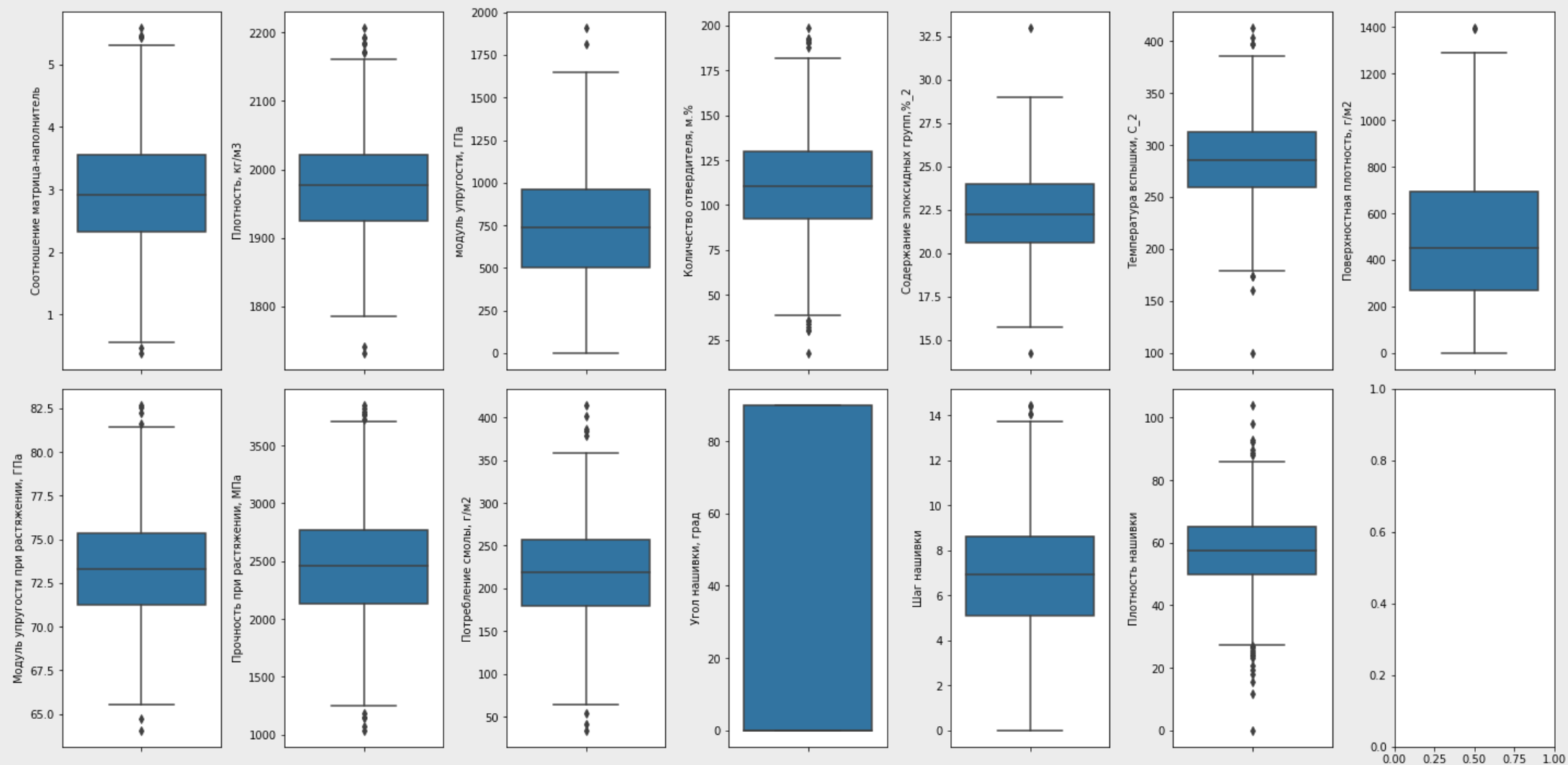
Название	Тип данных	Непустые значения	Уникальные значения
Соотношение матрица-наполнитель	float64	1023	1014
Плотность, кг/м3	float64	1023	1013
модуль упругости, ГПа	float64	1023	1020
Количество отвердителя, м.%	float64	1023	1005
Содержание эпоксидных групп,%_2	float64	1023	1004
Температура вспышки, C_2	float64	1023	1003
Поверхностная плотность, г/м2	float64	1023	1004
Модуль упругости при растяжении, ГПа	float64	1023	1004
Прочность при растяжении, МПа	float64	1023	1004
Потребление смолы, г/м2	float64	1023	1003
Угол нашивки, град	float64	1023	2
Шаг нашивки	float64	1023	989
Плотность нашивки	float64	1023	988

	<u>mean</u>	<u>std</u>	<u>min</u>	<u>max</u>	<u>median</u>
Соотношение матрица-наполнитель	2.9304	0.9132	0.3894	5.5917	2.9069
Плотность, кг/м3	1975.73	73.7292	1731.76	2207.77	1977.62
модуль упругости, ГПа	739.923	330.231	2.4369	1911.53	739.664
Количество отвердителя, м.%	110.570	28.2959	17.7403	198.953	110.564
Содержание эпоксидных групп,%_2	22.2444	2.4063	14.2550	33.0000	22.2307
Температура вспышки, C_2	285.882	40.9433	100.000	413.273	285.896
Поверхностная плотность, г/м2	482.731	281.314	0.6037	1399.54	451.864
Модуль упругости при растяжении, ГПа	73.3286	3.1190	64.0541	82.6821	73.2688
Прочность при растяжении, МПа	2466.92	485.628	1036.85	3848.43	2459.52
Потребление смолы, г/м2	218.423	59.7359	33.8030	414.590	219.198
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	103.988	57.3419

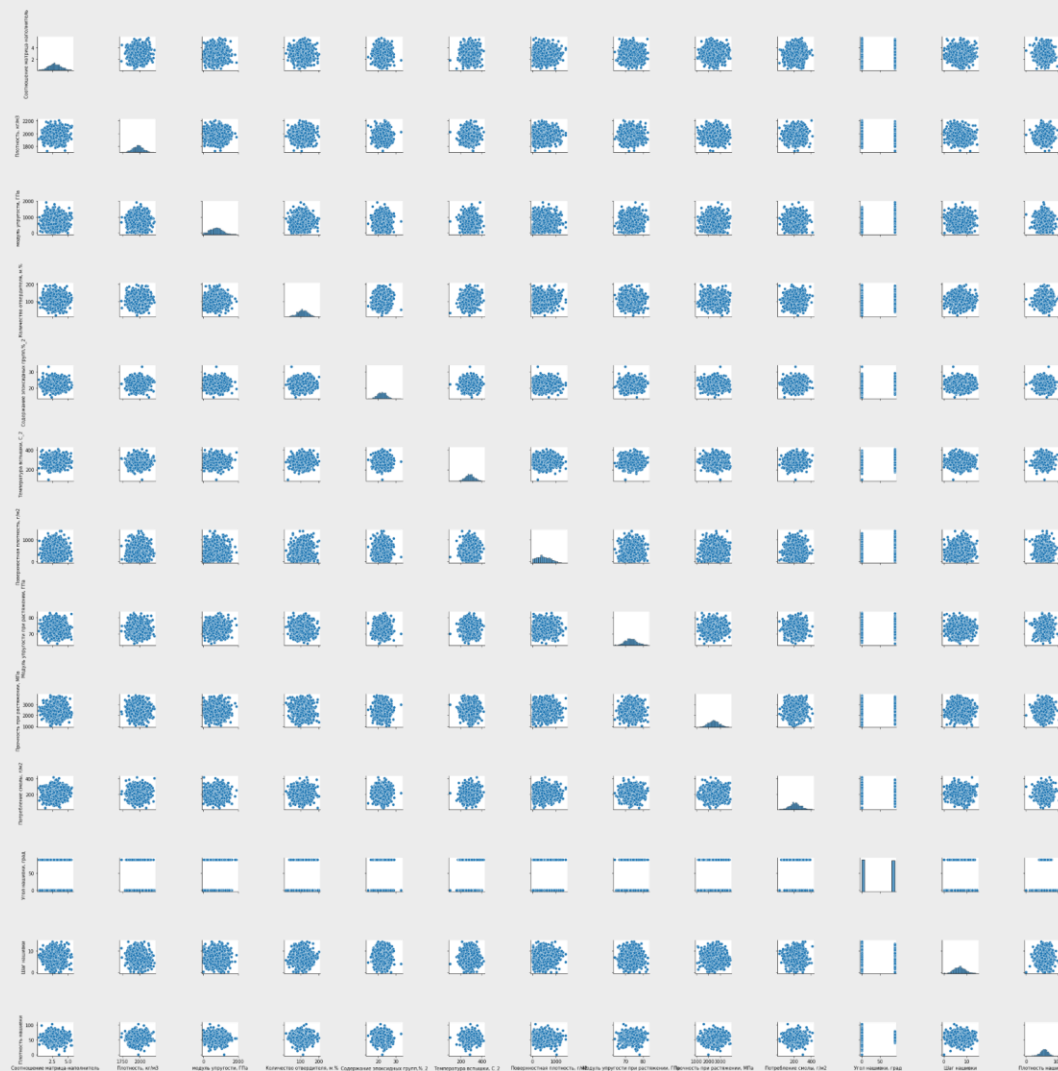
Гистограммы распределения



"Ящик с усами"



Попарные графики рассеяния точек



Выбросы

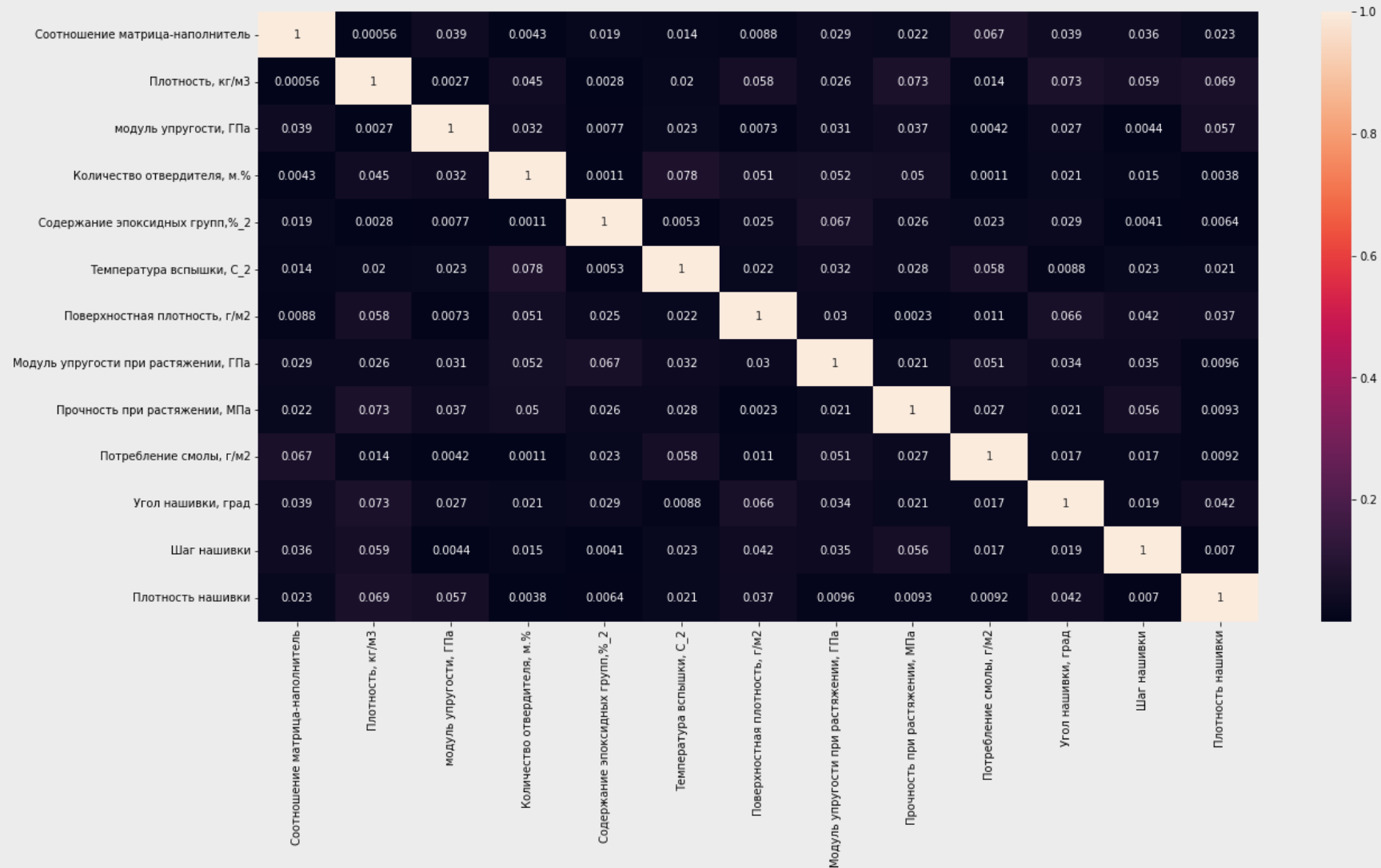
Применив метод межквартильных расстояний для удаления выбросов в параметрах с нормальным распределением, было обнаружено:

- "Соотношение матрица-наполнитель" кол-во выбросов = 0.59%
- "Плотность, кг/м3" кол-во выбросов = 0.88%
- "модуль упругости, ГПа" кол-во выбросов = 0.20%
- "Количество отвердителя, м.%" кол-во выбросов = 1.37%
- "Содержание эпоксидных групп,%_2" кол-во выбросов = 0.20%
- "Температура вспышки, С_2" кол-во выбросов = 0.78%
- "Поверхностная плотность, г/м2" кол-во выбросов = 0.20%
- "Модуль упругости при растяжении, ГПа" кол-во выбросов = 0.59%
- "Прочность при растяжении, МПа" кол-во выбросов = 1.08%
- "Потребление смолы, г/м2" кол-во выбросов = 0.78%
- "Угол нашивки, град" кол-во выбросов = 0.00%
- "Шаг нашивки" кол-во выбросов = 0.39%
- "Плотность нашивки" кол-во выбросов = 2.05%

и удалено 69 выбросов.

После этого осталось в датасете осталось 954 строк и 13 параметров

Тепловая карта



Входные переменные

По условиям задания нашими целевыми переменными являются:

- Модуль упругости при растяжении, ГПа;
- Прочность при растяжении, МПа;
- Соотношение матрица-наполнитель

Статистика входных признаков до и после предобработки

	min	max	mean	std
0	0.389403	5.455566	2.930646	0.904203
1	1731.764635	2207.773481	1975.747669	73.820992
2	2.436909	1911.536477	743.484597	330.231615
3	40.304806	179.645962	110.688614	26.781546
4	14.254985	33.000000	22.208086	2.421878
5	173.484920	413.273418	286.559313	40.450441
6	0.603740	1399.542362	481.074070	278.814346
7	1036.856605	3848.436732	2467.771681	481.224022
8	41.048278	414.590628	218.328594	59.498348
9	0.000000	90.000000	46.792453	44.987872
10	0.037639	14.440522	6.917423	2.572487
11	35.005121	84.840888	58.059589	10.465658

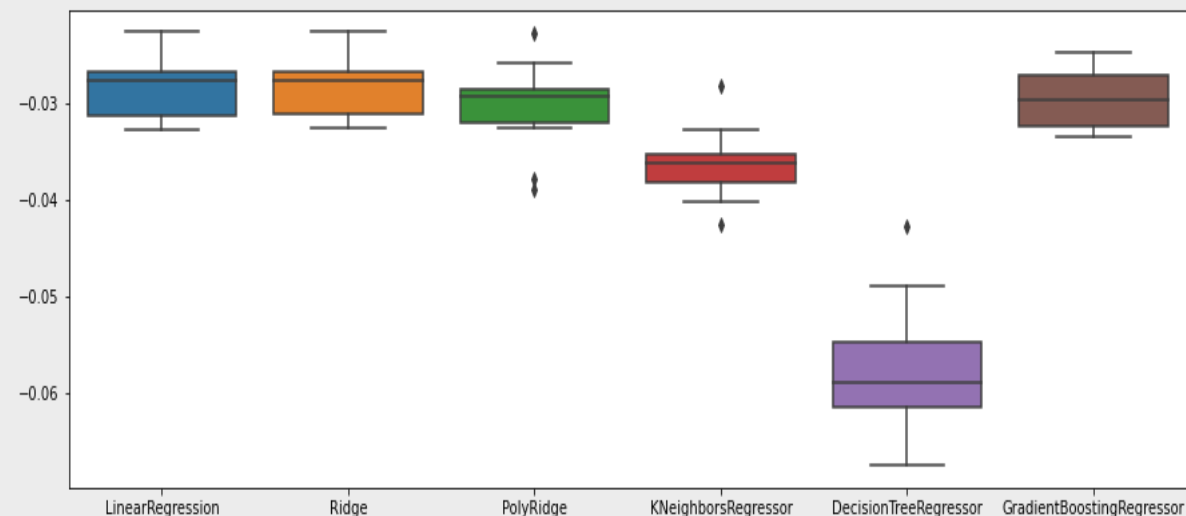
	min	max	mean	std
0	0.031185	1.000000	0.505959	0.175854
1	0.000000	0.967488	0.511649	0.157760
2	0.000996	0.851482	0.389434	0.171242
3	0.000000	0.980428	0.508197	0.193462
4	0.000000	1.000000	0.426132	0.129633
5	0.002039	1.000000	0.474961	0.168977
6	0.000761	1.000000	0.346048	0.200115
7	0.000000	1.000000	0.503434	0.174262
8	0.060602	1.000000	0.477176	0.156139
9	0.000000	1.000000	0.515742	0.500127
10	0.000000	0.995552	0.470768	0.180465
11	0.000000	1.000000	0.473552	0.215424

Метрики работы выбранных моделей

Для подбора лучшей модели для выполнения этой задачи были выбраны следующие модели:

- LinearRegression - линейная регрессия;
- Ridge - гребневая регрессия;
- KNeighborsRegressor - метод ближайших соседей;
- DecisionTreeRegressor - деревья решений;
- GradientBoosting - градиентный бустинг

	LinearRegression	Ridge	PolyRidge	KNeighborsRegressor	DecisionTreeRegressor	GradientBoostingRegressor
0	-0.027368	-0.027314	-0.028680	-0.036453	-0.048984	-0.027950
1	-0.029411	-0.029362	-0.038978	-0.042571	-0.062243	-0.031969
2	-0.032681	-0.032670	-0.037773	-0.040280	-0.062273	-0.033579
3	-0.031824	-0.031780	-0.032665	-0.036422	-0.059673	-0.032422
4	-0.027056	-0.026993	-0.028455	-0.035373	-0.067604	-0.027373
5	-0.027973	-0.027918	-0.029938	-0.035367	-0.059443	-0.031220
6	-0.022654	-0.022637	-0.022830	-0.028270	-0.057153	-0.024795
7	-0.031897	-0.031843	-0.030194	-0.038765	-0.058517	-0.033444
8	-0.024399	-0.024430	-0.028706	-0.036104	-0.054112	-0.027095
9	-0.026571	-0.026562	-0.025804	-0.032760	-0.042749	-0.026685



Метрики качества выбранных моделей

Обычная линейная регрессия методом наименьших квадратов

MAE: 0.13271867359870188
MSE: 0.02794886397265417
RMSE: 0.16717913737262247

Линейный метод наименьших квадратов с регуляризацией l2

MAE: 0.13263628212128165
MSE: 0.027938115050893025
RMSE: 0.16714698636497466

PolyRidge

MAE: 0.13543274634462302
MSE: 0.028756498486995827
RMSE: 0.16957741148807476

Gradient Boosting ...

MAE: 0.13792449178645536
MSE: 0.030205009756914706
RMSE: 0.17379588532791768

Regression based on k-nearest neighbors ...

MAE: 0.15030780504373686
MSE: 0.03414234603445326
RMSE: 0.18477647586869184

Дерево решений ...

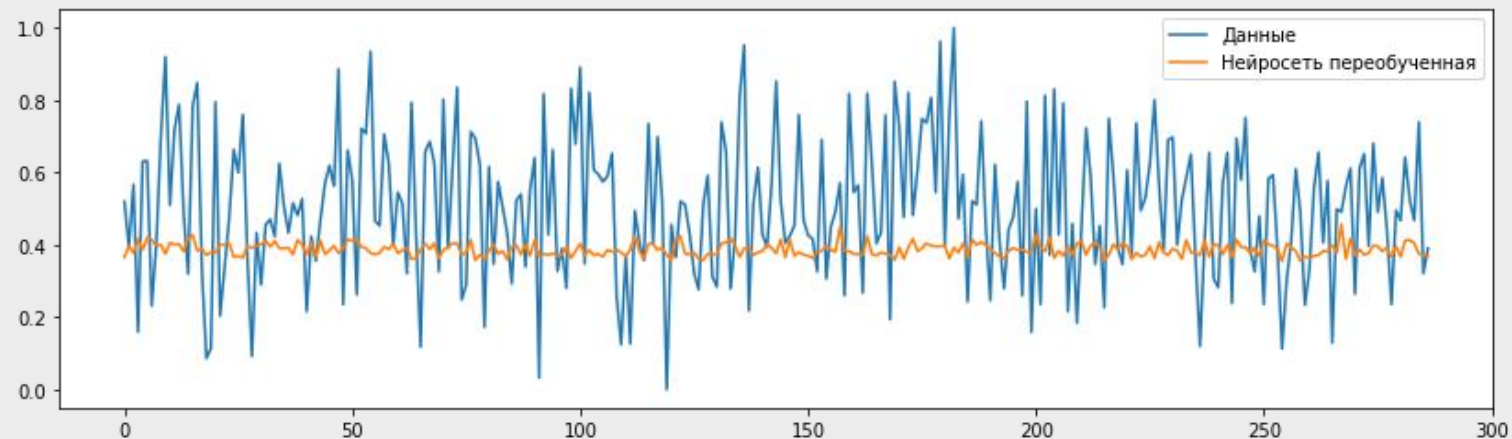
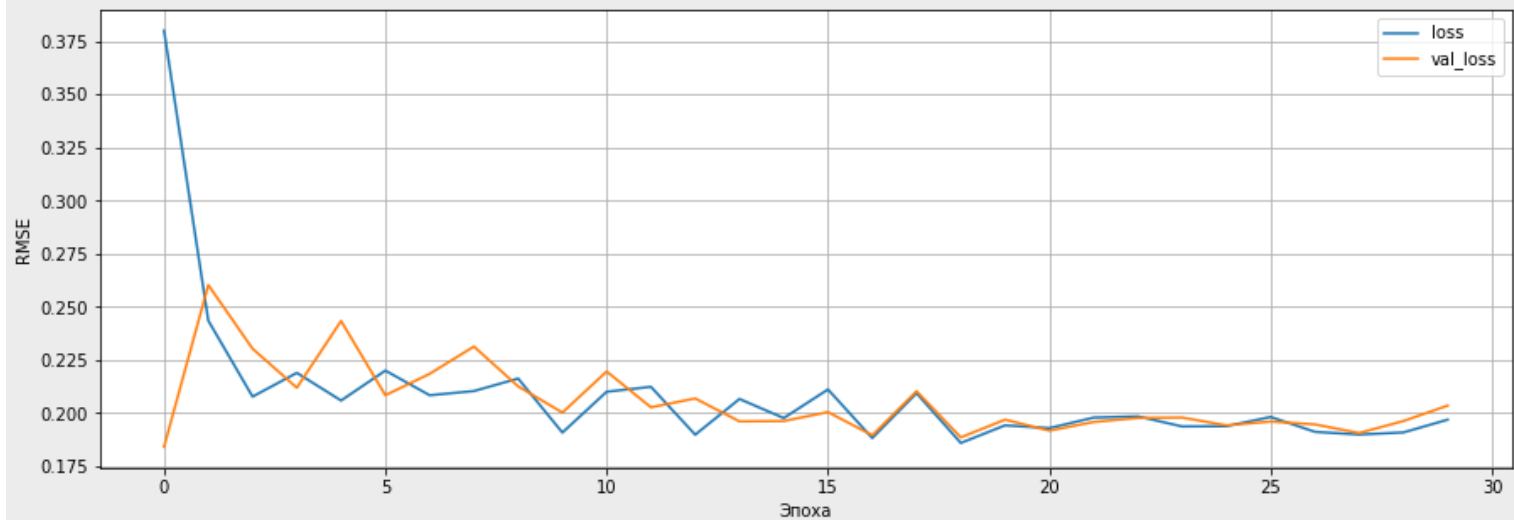
MAE: 0.17860060477088957
MSE: 0.05176681084076325
RMSE: 0.22752320945513063

Нейронная сеть

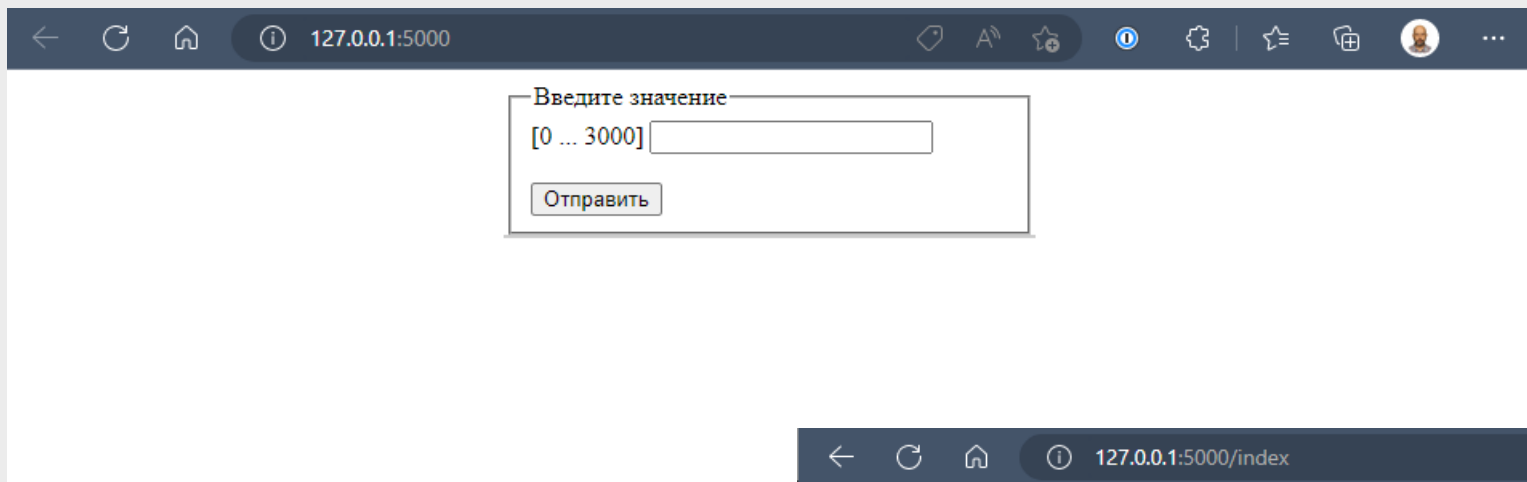
Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	832
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 64)	4160
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 64)	4160
dropout_2 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 64)	4160
dropout_3 (Dropout)	(None, 64)	0
dense_4 (Dense)	(None, 64)	4160
dropout_4 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 64)	4160
dropout_5 (Dropout)	(None, 64)	0
dense_6 (Dense)	(None, 1)	65

=====
Total params: 21,697
Trainable params: 21,697
Non-trainable params: 0
=====

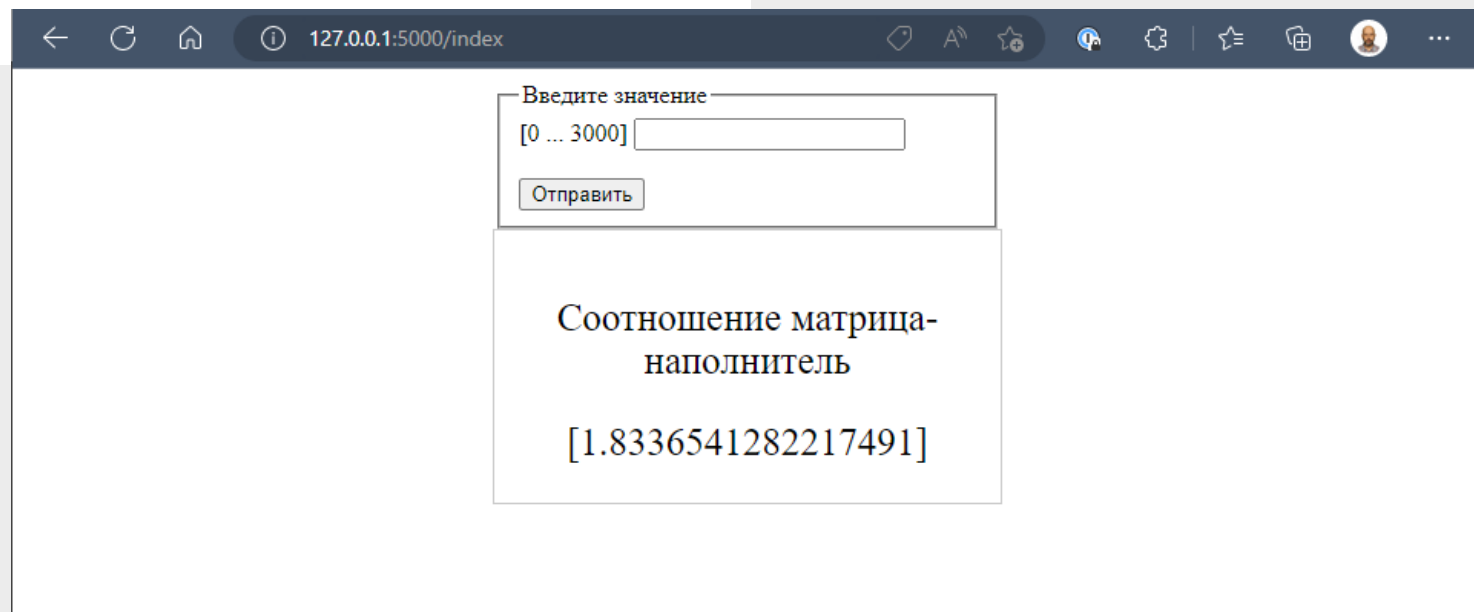


Разработка веб-приложения



Введите значение

[0 ... 3000]



Введите значение

[0 ... 3000]

Соотношение матрица-
наполнитель

[1.8336541282217491]

Что еще?

Дальнейшие поиски решения могли бы включать:

- проконсультироваться у экспертов
- исследовать сырые данные
- провести отбор признаков и уменьшение размерности
- поэкспериментировать с градиентным бустингом
- углубиться в нейросети



edu.bmstu.ru

+7 495 182-83-85

edu@bmstu.ru

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