ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу «Data Science»

Слушатель: Четвериков Артем Васильевич

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Начальные условия

Изначально даны два датасета X_bp.xlsx и X_nup.xlsx со свойствами композитного материала.

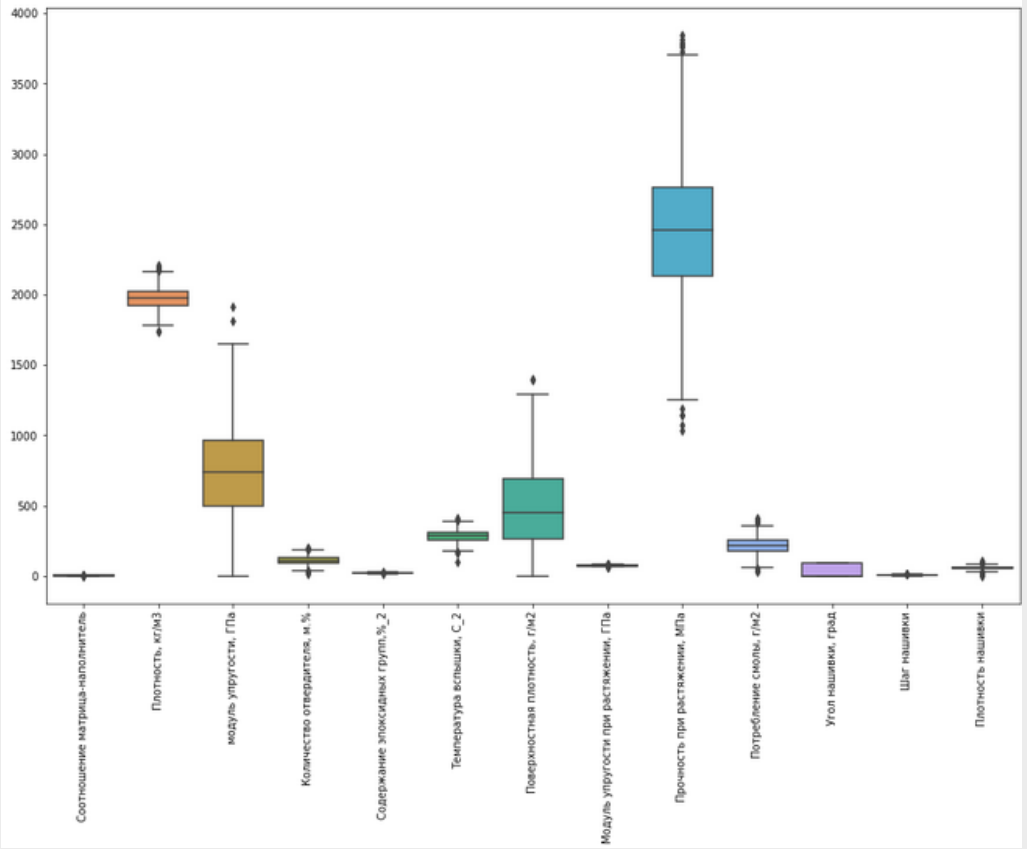
Датасет X_bp.xlsx содержит 11 столбцов и 1023 строки.

Датасет X_nup.xlsx содержит 4 столбца и 1040 строк.

Для начала необходимо объединить датасеты по типу объединения INNER и удалить дублирующие столбцы с индексами.

После объединения мы имеем датасет с 13 столбцами и 1023 строками, все строки не имеют пустых значений, не имеют дубликатов, тип данных в столбцах — float64. В основном в каждом столбце содержатся только уникальные значения, но в столбце "Угол нашивки" всего 2 значения.

Разведочный анализ

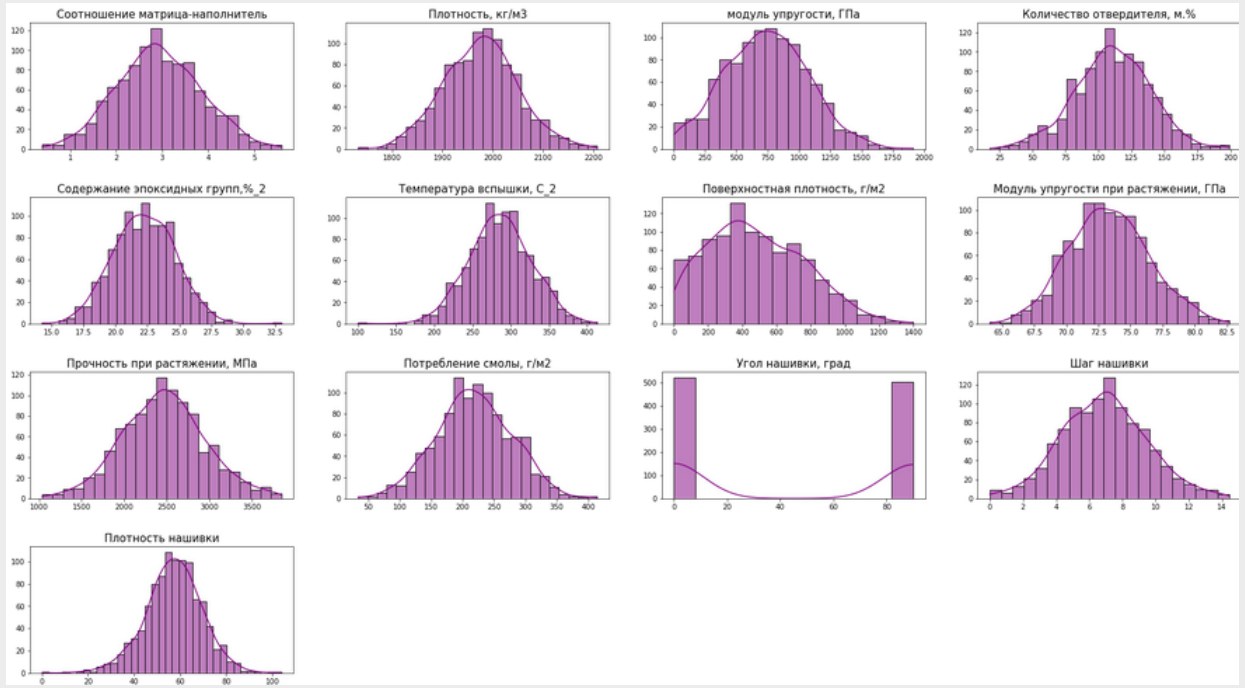


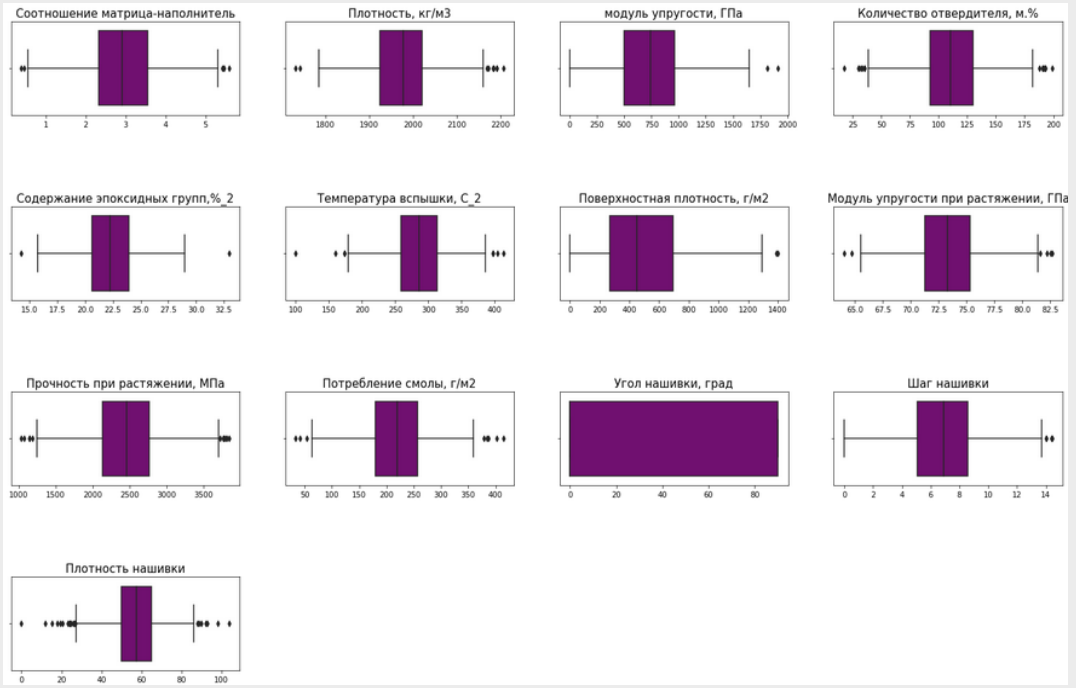
	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
Соотношение матрица-наполнитель	1.000000	0.003841	0.031700	-0.006445	0.019766	-0.004776	-0.006272	-0.008411	0.024148	0.072531	-0.031073	0.036437	-0.004652
Плотность, кг/м3	0.003841	1.000000	-0.009647	-0.035911	-0.008278	-0.020695	0.044930	-0.017602	-0.069981	-0.015937	-0.068474	-0.061015	0.080304
модуль упругости, ГПа	0.031700	-0.009647	1.000000	0.024049	-0.006804	0.031174	-0.005306	0.023267	0.041868	0.001840	-0.025417	-0.009875	0.056346
Количество отвердителя, м.%	-0.006445	-0.035911	0.024049	1.000000	-0.000684	0.095193	0.055198	-0.065929	-0.075375	0.007446	0.038570	0.014887	0.017248
Содержание эпоксидных групп,%_2	0.019766	-0.008278	-0.006804	-0.000684	1.000000	-0.009769	-0.012940	0.056828	-0.023899	0.015165	0.008052	0.003022	-0.039073
Температура вспышки, С_2	-0.004776	-0.020695	0.031174	0.095193	-0.009769	1.000000	0.020121	0.028414	-0.031763	0.059954	0.020695	0.025795	0.011391
Поверхностная плотность, г/м2	-0.006272	0.044930	-0.005306	0.055198	-0.012940	0.020121	1.000000	0.036702	-0.003210	0.015692	0.052299	0.038332	-0.049923
Модуль упругости при растяжении, ГПа	-0.008411	-0.017602	0.023267	-0.065929	0.056828	0.028414	0.036702	1.000000	-0.009009	0.050938	0.023003	-0.029468	0.006476
Прочность при растяжении, МПа	0.024148	-0.069981	0.041868	-0.075375	-0.023899	-0.031763	-0.003210	-0.009009	1.000000	0.028602	0.023398	-0.059547	0.019604
Потребление смолы, г/м2	0.072531	-0.015937	0.001840	0.007446	0.015165	0.059954	0.015692	0.050938	0.028602	1.000000	-0.015334	0.013394	0.012239
Угол нашивки, град	-0.031073	-0.068474	-0.025417	0.038570	0.008052	0.020695	0.052299	0.023003	0.023398	-0.015334	1.000000	0.023616	0.107947
Шаг нашивки	0.036437	-0.061015	-0.009875	0.014887	0.003022	0.025795	0.038332	-0.029468	-0.059547	0.013394	0.023616	1.000000	0.003487
Плотность нашивки	-0.004652	0.080304	0.056346	0.017248	-0.039073	0.011391	-0.049923	0.006476	0.019604	0.012239	0.107947	0.003487	1.000000

Построив график типа боксплот («ящик с усами») для исходного датасета, наглядно видно, что порядок значений переменных сильно различается − практически в 1000 раз.

В целях выявления зависимостей между переменными построим тепловую карту коэффициентов корреляции. Коэффициенты корелляции предварительно показывают, что явная зависимость между переменными датасета отсутствует.

Разведочный анализ





По гистограммам распределения видно, что распределение величин близко к нормальному для большей части параметров, за исключением поверхностной плотности – большое смещением влево и угла нашивки – дискретная величина.

Построив диаграммы «ящик с усами» для переменных датасета, видно что практически у всех переменных имеются выбросы, кроме угла нашивки, так как данный параметр принимает дискретные значения и диаграмма «ящик с усами» для него не показательна.

Предобработка данных

До удаления выбросов

```
0 выбросов в признаке Соотношение матрица-наполнитель
3 выбросов в признаке Плотность, кг/м3
2 выбросов в признаке модуль упругости, ГПа
2 выбросов в признаке Количество отвердителя, м.%
2 выбросов в признаке Содержание эпоксидных групп,%_2
3 выбросов в признаке Температура вспышки, С_2
2 выбросов в признаке Поверхностная плотность, г/м2
0 выбросов в признаке Модуль упругости при растяжении, ГПа
0 выбросов в признаке Прочность при растяжении, МПа
3 выбросов в признаке Потребление смолы, г/м2
0 выбросов в признаке Угол нашивки, град
0 выбросов в признаке Шаг нашивки
7 выбросов в признаке Плотность нашивки
 Всего 24 выброса
```

После удаления выбросов методом 3-х сигм

```
0 выбросов в признаке Соотношение матрица-наполнитель
0 выбросов в признаке Плотность, кг/м3
0 выбросов в признаке модуль упругости, ГПа
0 выбросов в признаке Количество отвердителя, м.%
0 выбросов в признаке Содержание эпоксидных групп,%_2
0 выбросов в признаке Температура вспышки, С_2
0 выбросов в признаке Поверхностная плотность, г/м2
0 выбросов в признаке Модуль упругости при растяжении, ГПа
0 выбросов в признаке Прочность при растяжении, МПа
0 выбросов в признаке Потребление смолы, г/м2
0 выбросов в признаке Угол нашивки, град
0 выбросов в признаке Шаг нашивки
0 выбросов в признаке Плотность нашивки
 Всего 0 выброса
```

Для начала удалим наиболее сильные выбросы. Воспользуемся методом 3-х сигм, так как датасет у нас не большой, а метод 3-х сигм удаляет только экстремальные выбросы и больше данных останется для дальнейшей работы.

Дальнейшая подготовка сводится к нормализации (масштабированию) и стандартизации данных. Наиболее распространенные методы для нормализации (масштабирования) данных − это MinMaxScaler(), Normalizer(), для стандартизации данных − StandardScaler().

Предобработка данных

	Соотношение матрица-наполнитель	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
Соотношение матрица-наполнитель	1.000000	0.272299	-0.069123	0.090689	0.219430	0.159199	-0.110231	0.261391	-0.096178	0.126405	-0.043725	0.130376	0.124241
Плотность, кг/м3	0.272299	1.000000	-0.341492	0.309659	0.669756	0.530541	-0.249897	0.874108	-0.349855	0.239540	0.005302	0.271541	0.403467
модуль упругости, ГПа	-0.069123	-0.341492	1.000000	-0.106706	-0.242820	-0.201703	-0.162351	-0.315704	-0.429844	-0.151229	-0.092625	-0.098019	-0.112122
Количество отвердителя, м.%	0.090689	0.309659	-0.106706	1.000000	0.243083	0.262897	-0.063120	0.290212	-0.223266	0.115824	0.025950	0.105673	0.142027
Содержание эпоксидных групп,%_2	0.219430	0.669756	-0.242820	0.243083	1.000000	0.398158	-0.188621	0.675816	-0.232214	0.196907	0.026647	0.221185	0.249596
Температура вспышки, С_2	0.159199	0.530541	-0.201703	0.262897	0.398158	1.000000	-0.161187	0.532456	-0.273932	0.188526	0.023446	0.184560	0.217420
Поверхностная плотность, г/м2	-0.110231	-0.249897	-0.162351	-0.063120	-0.188621	-0.161187	1.000000	-0.253748	-0.388675	-0.118802	0.006821	-0.026174	-0.172718
Модуль упругости при растяжении, ГПа	0.261391	0.874108	-0.315704	0.290212	0.675816	0.532456	-0.253748	1.000000	-0.296862	0.259627	0.035005	0.271795	0.380408
Прочность при растяжении, МПа	-0.096178	-0.349855	-0.429844	-0.223266	-0.232214	-0.273932	-0.388675	-0.296862	1.000000	-0.157079	-0.066763	-0.149099	-0.158493
Потребление смолы, г/м2	0.126405	0.239540	-0.151229	0.115824	0.196907	0.188526	-0.118802	0.259627	-0.157079	1.000000	-0.027645	0.093918	0.095083
Угол нашивки, град	-0.043725	0.005302	-0.092625	0.025950	0.026647	0.023446	0.006821	0.035005	-0.066763	-0.027645	1.000000	0.041268	0.119339
Шаг нашивки	0.130376	0.271541	-0.098019	0.105673	0.221185	0.184560	-0.026174	0.271795	-0.149099	0.093918	0.041268	1.000000	0.124782
Плотность нашивки	0.124241	0.403467	-0.112122	0.142027	0.249596	0.217420	-0.172718	0.380408	-0.158493	0.095083	0.119339	0.124782	1.000000

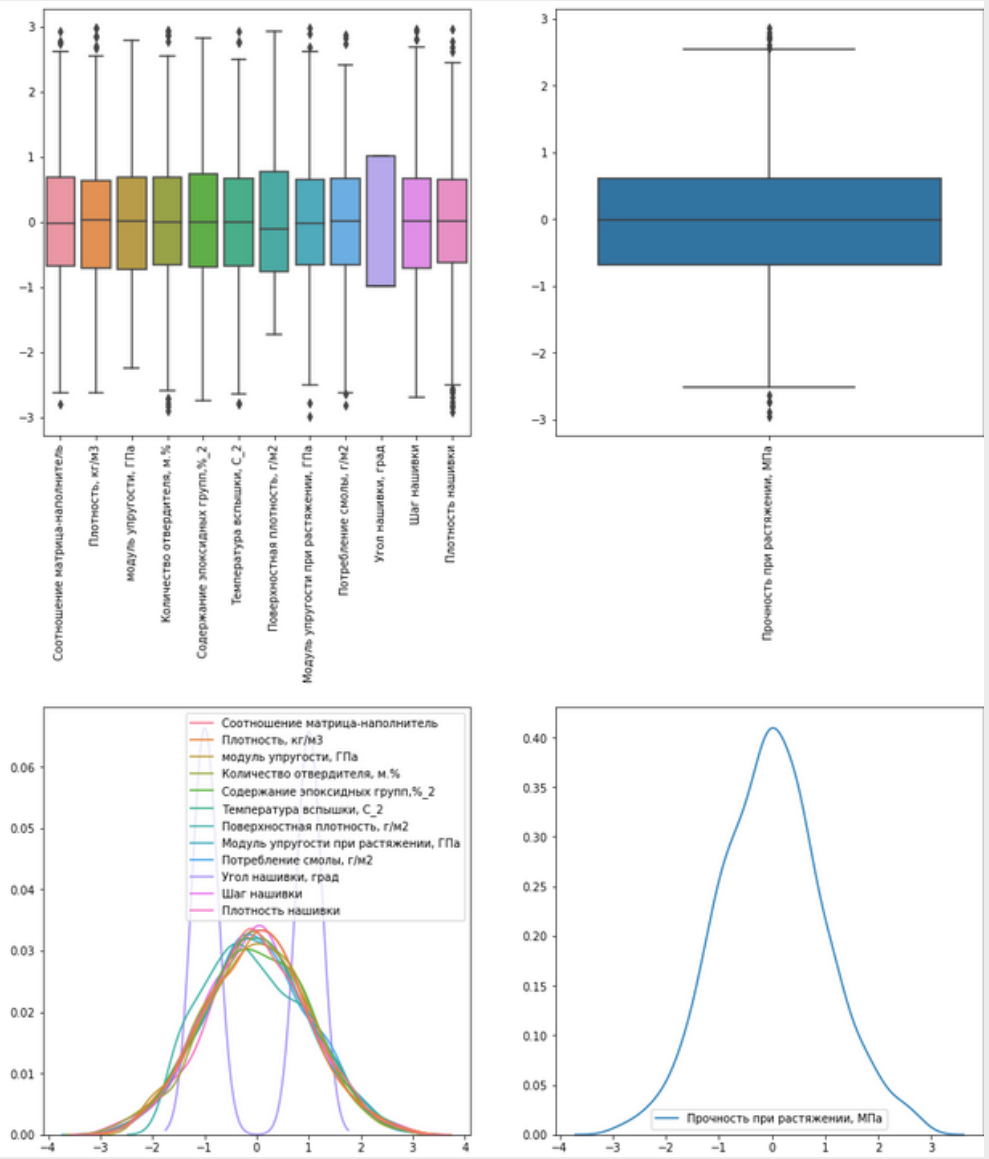
Интересный эффект дает применение Normalizer() к датасету без транспонирования, Normalizer() нормализует в датасете каждую отдельную строку — единичное наблюдение и свяжет в ложной зависимости признаки и целевую переменную.

На таком датасете практически все модели отлично обучаются, и показывают хороший результат на тестовой выборке.

А вот дальнейшее применение обученной модели на рабочих данных вызывает проблему, так как не возможно применить нормализатор, идентичный нормализатору на обучающей выборке: в рабочих данных отсутствует целевая переменная.

Для дальнейшей работы разделим наш датасет на два — целевая переменная и признаки и применим к ним методы нормализации и стандартизации по отдельности.

Предобработка данных



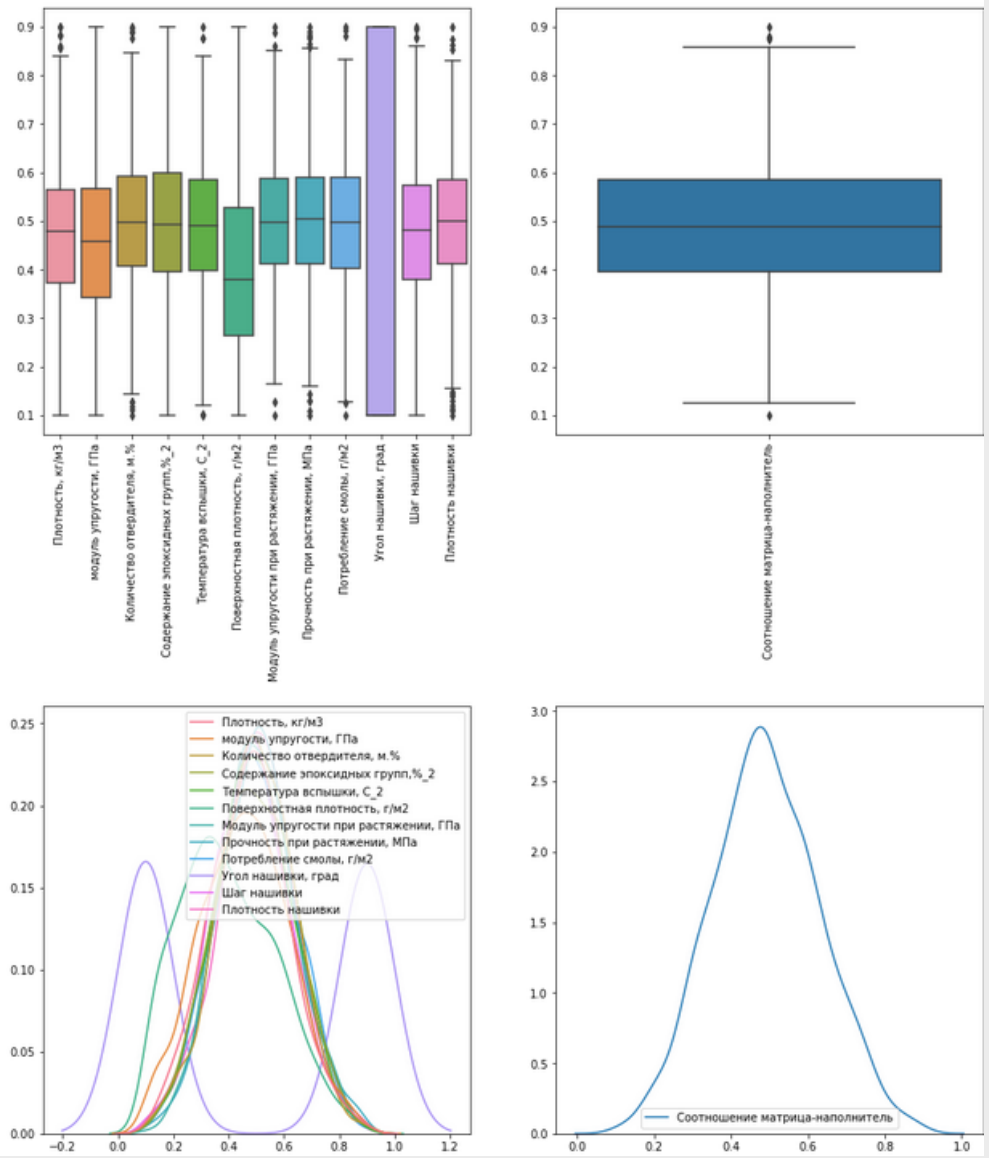
	count	mean	std	min	25%	50%	75%	max
Прочность при растяжении, МПа	996.0	8.159470e-17	1.000502	-2.951505	-0.683292	-0.018213	0.609486	2.858263

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	996.0	-7.735891e-17	1.000502	-2.803667	-0.678604	-0.029453	0.679380	2.924821
Плотность, кг/м3	996.0	4.345029e-16	1.000502	-2.617348	-0.710194	0.027578	0.625685	2.974847
модуль упругости, ГПа	996.0	1.276311e-16	1.000502	-2.249733	-0.726070	0.006244	0.678775	2.778542
Количество отвердителя, м.%	996.0	-1.245100e-16	1.000502	-2.907555	-0.657806	-0.001975	0.685531	2.942771
Содержание эпоксидных групп,%_2	996.0	-1.702342e-15	1.000502	-2.738354	-0.691286	-0.009945	0.733182	2.817417
Температура вспышки, С_2	996.0	1.447972e-16	1.000502	-2.791963	-0.670053	-0.002121	0.673185	2.922531
Поверхностная плотность, г/м2	996.0	7.446074e-17	1.000502	-1.722929	-0.766767	-0.103865	0.761024	2.923701
Модуль упругости при растяжении, ГПа	996.0	-2.309676e-15	1.000502	-2.982683	-0.667639	-0.030928	0.648376	2.973738
Потребление смолы, г/м2	996.0	-3.518493e-16	1.000502	-2.808015	-0.655024	0.005891	0.662979	2.871884
Угол нашивки, град	996.0	7.989593e-15	1.000502	-0.995992	-0.995992	-0.995992	1.004024	1.004024
Шаг нашивки	996.0	-9.809199e-18	1.000502	-2.688098	-0.711760	0.005528	0.660806	2.954293
Плотность нашивки	996.0	-5.996988e-17	1.000502	-2.908316	-0.619488	0.018682	0.650048	2.958674

Для случая прогнозирования прочности при растяжении применением StandardScaler().

Аналогично применением StandardScaler() для случая прогнозирования модуля упругости при растяжении.

Предобработка данных



	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	996.0	0.49154	0.139723	0.1	0.396771	0.487427	0.586418	0.9
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Плотность, кг/м3	996.0	0.474429	0.143128	0.1	0.372831	0.478374	0.563937	0.9
модуль упругости, ГПа	996.0	0.457933	0.159180	0.1	0.342415	0.458927	0.565927	0.9
Количество отвердителя, м.%	996.0	0.497592	0.136813	0.1	0.407641	0.497322	0.591335	0.9
Содержание эпоксидных групп,%_2	996.0	0.494308	0.144067	0.1	0.394766	0.492876	0.599882	0.9
Температура вспышки, С_2	996.0	0.490861	0.140065	0.1	0.397057	0.490564	0.585103	0.9
Поверхностная плотность, г/м2	996.0	0.396633	0.172254	0.1	0.264620	0.378751	0.527657	0.9
Модуль упругости при растяжении, ГПа	996.0	0.500601	0.134376	0.1	0.410931	0.496447	0.587683	0.9
Прочность при растяжении, МПа	996.0	0.506420	0.137768	0.1	0.412331	0.503912	0.590345	0.9
Потребление смолы, г/м2	996.0	0.495502	0.140918	0.1	0.403244	0.496332	0.588881	0.9
Угол нашивки, град	996.0	0.498394	0.400198	0.1	0.100000	0.100000	0.900000	0.9
Шаг нашивки	996.0	0.481129	0.141855	0.1	0.380213	0.481913	0.574821	0.9
Плотность нашивки	996.0	0.496567	0.136425	0.1	0.412096	0.499114	0.585205	0.9

Для случая предсказания нейросетью соотношения матрица-наполнитель применим MinMaxScaler()

Для дальнейшего использования нормализаторы были сохранены в отдельные файлы при помощи библиотеки Joblib.

Разработка, обучение и тестирование моделей

В ранной работе применялись наиболее часто используемые методы и модели машинного обучения из широкоизвестной библиотеки scikit-learn. При разработке и обучении моделей был проведен поиск оптимальных гиперпараметров моделей с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10, для чего был применен метод GridSearchCV() с параметрами: количество перекрестных проверок cv = 10, сравнение качества моделей по коэффициенту детерминации scoring = 'r2'.

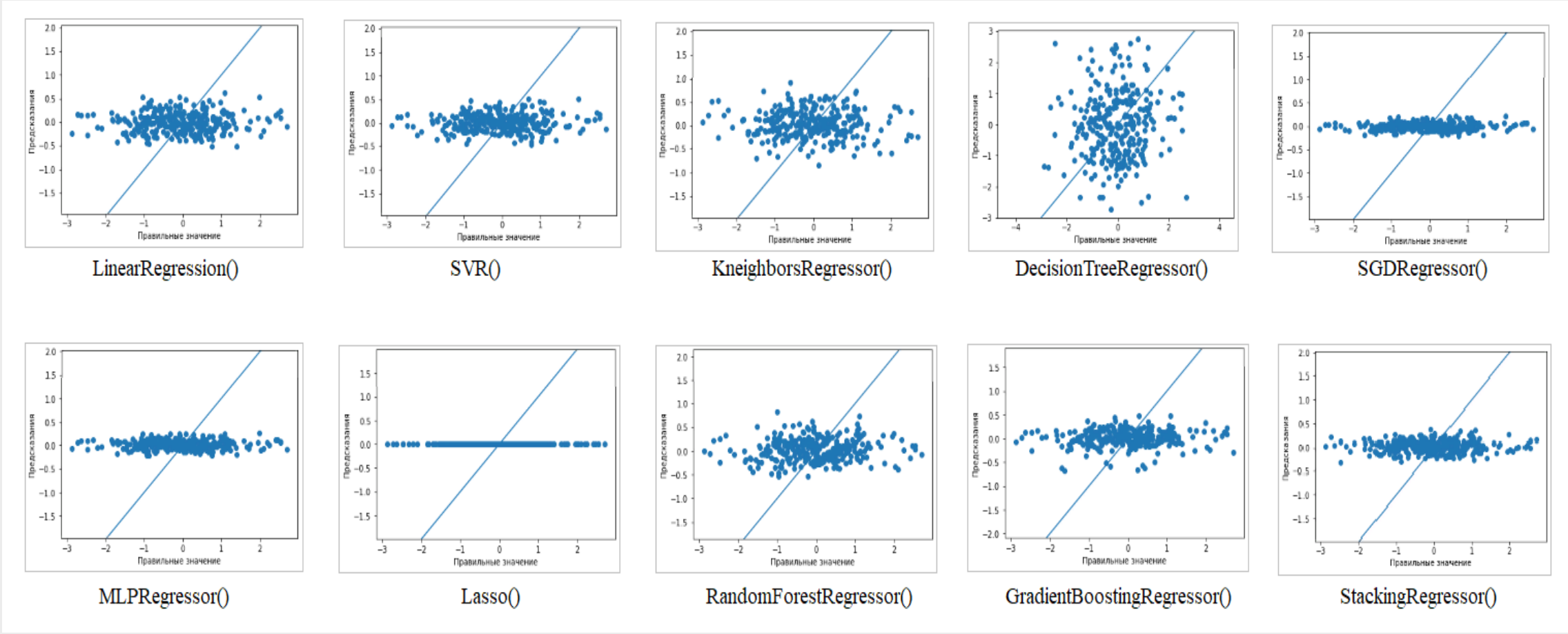
Перед обучением моделей датасеты были разделены на обучающую и тестовую выборки, в соответствии с условием задачи 70% на обучение и 30% на тестирование.

№ п/п	Модель	Коэффициент детерминации для тестовой выборки R^2
1.	LinearRegression	-0.03
2.	SVR	-0.02
3.	KneighborsRegressor	-0.14
4.	DecisionTreeRegressor	-1.04
5.	SGDRegressor	-0.00
6.	MLPRegressor	-0.01
7.	Lasso	-0.00
8.	RandomForestRegressor	-0.02
9.	GradientBoostingRegressor	-0.02
10.	StackingRegressor	-0.03

Модели будем сравнивать по коэффициенту детерминации, как наиболее показательной характеристике.

Из сводной таблицы видно, что при предсказании прочности при растяжении не удалось приблизится к идеальному результату более, чем до предсказания среднего значения.

Разработка, обучение и тестирование моделей



Графики сравнения прогнозируемых значений с наблюдаемыми значениями

Разработка, обучение и тестирование моделей

№ п/п	Модель	Коэффициент детерминации для тестовой выборки R^2
1.	LinearRegression	0.00
2.	SVR	-0.04
3.	KneighborsRegressor	-0.09
4.	DecisionTreeRegressor	-0.81
5.	SGDRegressor	-0.01
6.	MLPRegressor	-0.01
7.	Lasso	-0.01
8.	RandomForestRegressor	-0.03
9.	GradientBoostingRegressor	-0.04
10.	StackingRegressor	-0.04

Так же ощутимых результатов не удалось добиться при предсказании модуля упругости при растяжении

Из сводной таблицы видно, что при предсказании модуля упругости при растяжении не удалось приблизится к идеальному результату более, чем до предсказания среднего значения.

Разработка нейронной сети для рекомендации соотношения матрица-наполнитель

Model: "model_5"

Layer (type)	Output Shape	Param #	Connected to
input_10 (InputLayer)	[(None, 12)]	0	[]
dense_82 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_83 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_84 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_85 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_86 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_87 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_88 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_89 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_90 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_91 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_92 (Dense)	(None, 2)	26	['input_10[0][0]']
dense_93 (Dense)	(None, 2)	26	['input_10[0][0]']
concatenate_5 (Concatenate)	(None, 24)	0	['dense_82[0][0]', 'dense_83[0][0]', 'dense_84[0][0]', 'dense_85[0][0]', 'dense_86[0][0]', 'dense_87[0][0]', 'dense_88[0][0]', 'dense_89[0][0]', 'dense_90[0][0]', 'dense_91[0][0]', 'dense_92[0][0]', 'dense_93[0][0]']
dense_94 (Dense)	(None, 24)	600	['concatenate_5[0][0]']
dense_95 (Dense)	(None, 1)	25	['dense_94[0][0]']

Total params: 937
Trainable params: 937
Non-trainable params: 0

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 16)	208
dense_1 (Dense)	(None, 16)	272
dense_2 (Dense)	(None, 1)	17

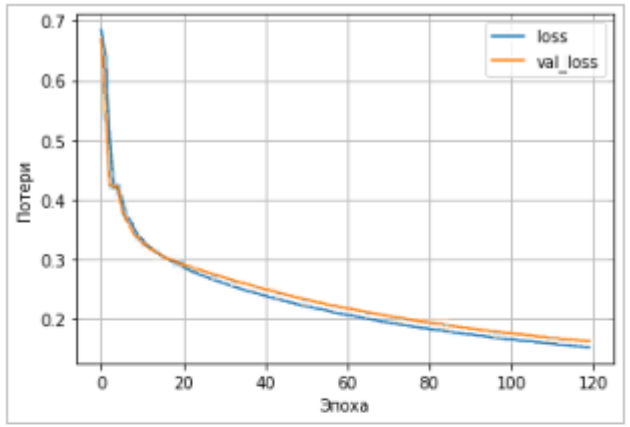
Total params: 497
Trainable params: 497
Non-trainable params: 0

Для рекомендации оотношения матрица-наполнитель было выбрано две модели нейронных сетей с разными архитектурами:

- первая нейросеть имеет два скрытых слоя по 16 нейронов, активационная функция на внутренних слоях "tanh", на выходном слое "relu";

- вторая нейросеть имеет разветвленную структуру: после входного слоя разделяется по числу входных признаков на 12 потоков по 1 слою из 2 нейронов, далее потоки объединяются в один слой из 24 нейронов, активационная функция на внутренних слоях "tanh", на выходном слое "relu"

Разработка нейронной сети для рекомендации соотношения матрица-наполнитель



Обучение первой нейросети

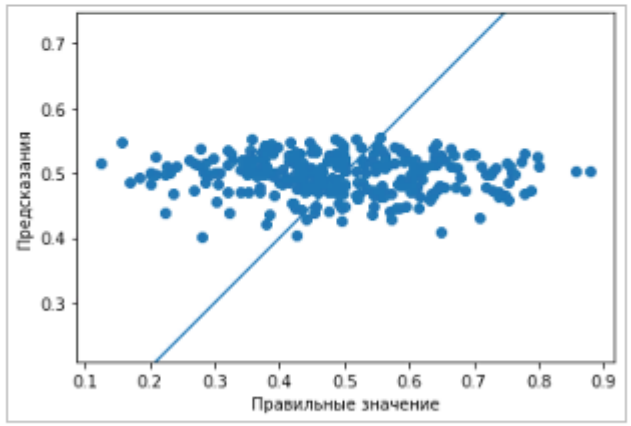
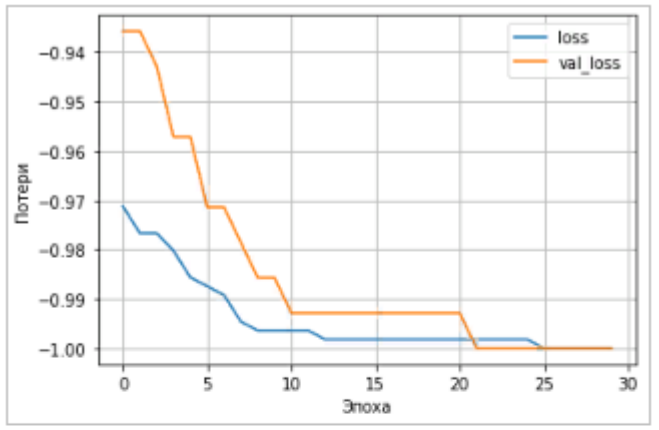


График сравнения прогнозируемых значений с наблюдаемыми значениями для первой нейросети



Обучение второй нейросети

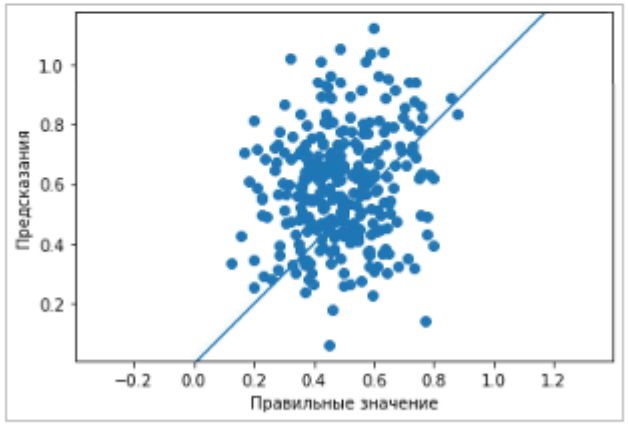
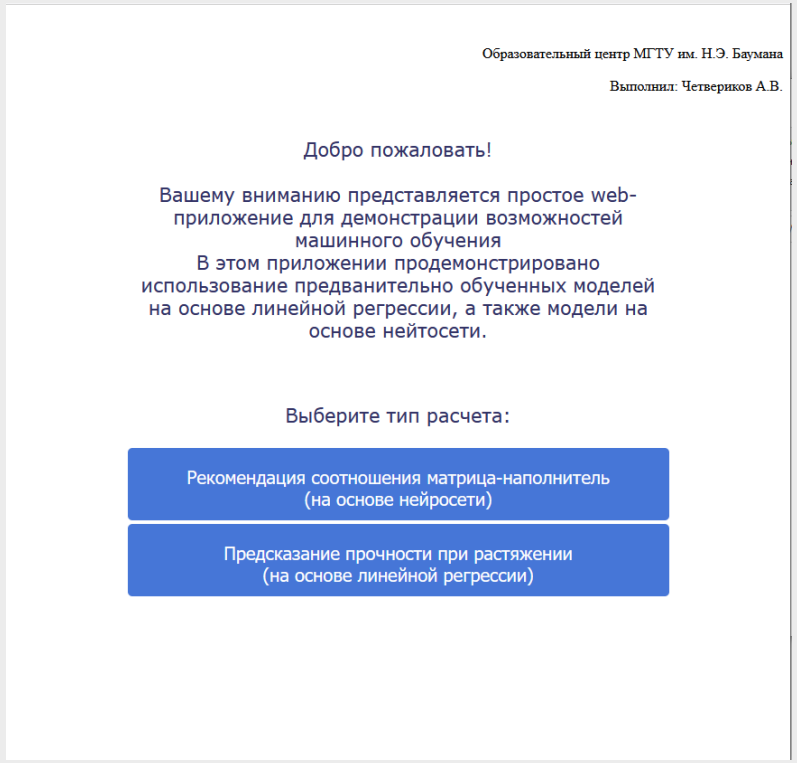


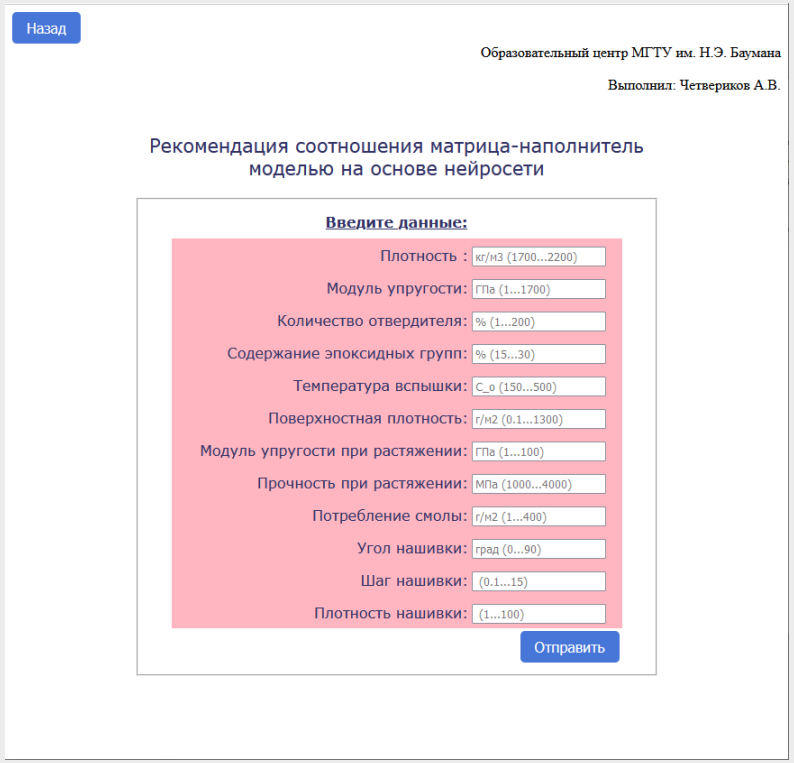
График сравнения прогнозируемых значений с наблюдаемыми значениями для второй нейросети

№ п/п	Модель	Коэффициент детерминации для тестовой выборки R^2
1.	Модель №1	-23.16
2.	Модель №2	-0.63

Из сводной таблицы видно, что при вычислении рекомендуемого соотношения матрица-наполнитель ни одной модели не удалось приблизится к идеальному результату, даже предсказание среднего значения удается с трудом.

Разработка приложения





Разработанные и обученные модели теперь необходимо как-то применять на практике. Для этих целей было разработано удобное web-приложение с использованием библиотеки Flask. В приложении представлены расчеты на основе двух моделей машинного обучения: расчет для рекомендации соотношения матрица-наполнитель выполняется нейросетью, а расчет для предсказания прочности при растяжении выполняется моделью на основе линейной регрессии.

Приложение можно запустить локально из среды Python 3.9. Также приложение доступно по адресу: http://avchetverikov.pythonanywhere.com/

Заключение

Если обобщить полученные результаты, то можно сделать вывод, что взаимосвязь, если она есть, между параметрами датасета очень слабая, и выявить ее наиболее распространенными моделями машинного обучения не удалось. Возможно, требуется более точная настройка гиперпараметров моделей, занимающая довольно много времени, и в рамках данной работы это не представляется возможным сделать.

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