ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу «Data Science Pro»

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель: Поникаровских Андрей Александрович

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

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 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 и разместить на вэб-хостинге rendor.com

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных

Датасет со свойствами композитов представлен из двух таблиц формата Excel:

1) «**X_bp.xlsx**» (матрица из базальтопластика, признаков: 10 и индекс, строк: 1023),

2) «**X_nup.xlsx**» (наполнитель из углепластика, признаков: 3 и индекс, строк 1040)

Проведено объединение таблиц по индексной колонке с помощью типа объединения INNER.

- объём объединенного датасета: 1023 записи по каждому показателю
- пропуски отсутствуют (пустых значений нет), дубликатов нет, уникальные значения приведены

```
df.duplicated().sum()
```

```
np.int64(0)
```

дубликатов данных нет

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1023 entries, 0 to 1022
Data columns (total 13 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
 10  Угол нашивки, град                 1023 non-null   int64  
 11  Шаг нашивки                        1023 non-null   float64
 12  Плотность нашивки                  1023 non-null   float64
dtypes: float64(12), int64(1)
```

```
df.isnull().sum()
```

```
Соотношение матрица-наполнитель    0
Плотность, кг/м3                   0
модуль упругости, ГПа              0
Количество отвердителя, м.%        0
Содержание эпоксидных групп,%_2    0
Температура вспышки, С_2           0
Поверхностная плотность, г/м2      0
Модуль упругости при растяжении, ГПа  0
Прочность при растяжении, МПа      0
Потребление смолы, г/м2            0
Угол нашивки, град                 0
Шаг нашивки                        0
Плотность нашивки                  0
dtype: int64
```

```
df.nunique()
```

```
Соотношение матрица-наполнитель    1014
Плотность, кг/м3                   1013
модуль упругости, ГПа              1020
Количество отвердителя, м.%        1005
Содержание эпоксидных групп,%_2    1004
Температура вспышки, С_2           1003
Поверхностная плотность, г/м2      1004
Модуль упругости при растяжении, ГПа  1004
Прочность при растяжении, МПа      1004
Потребление смолы, г/м2            1003
Угол нашивки, град                    2
Шаг нашивки                         989
Плотность нашивки                   988
dtype: int64
```

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Описательная статистика

Представлены основные характеристики параметров датасета:

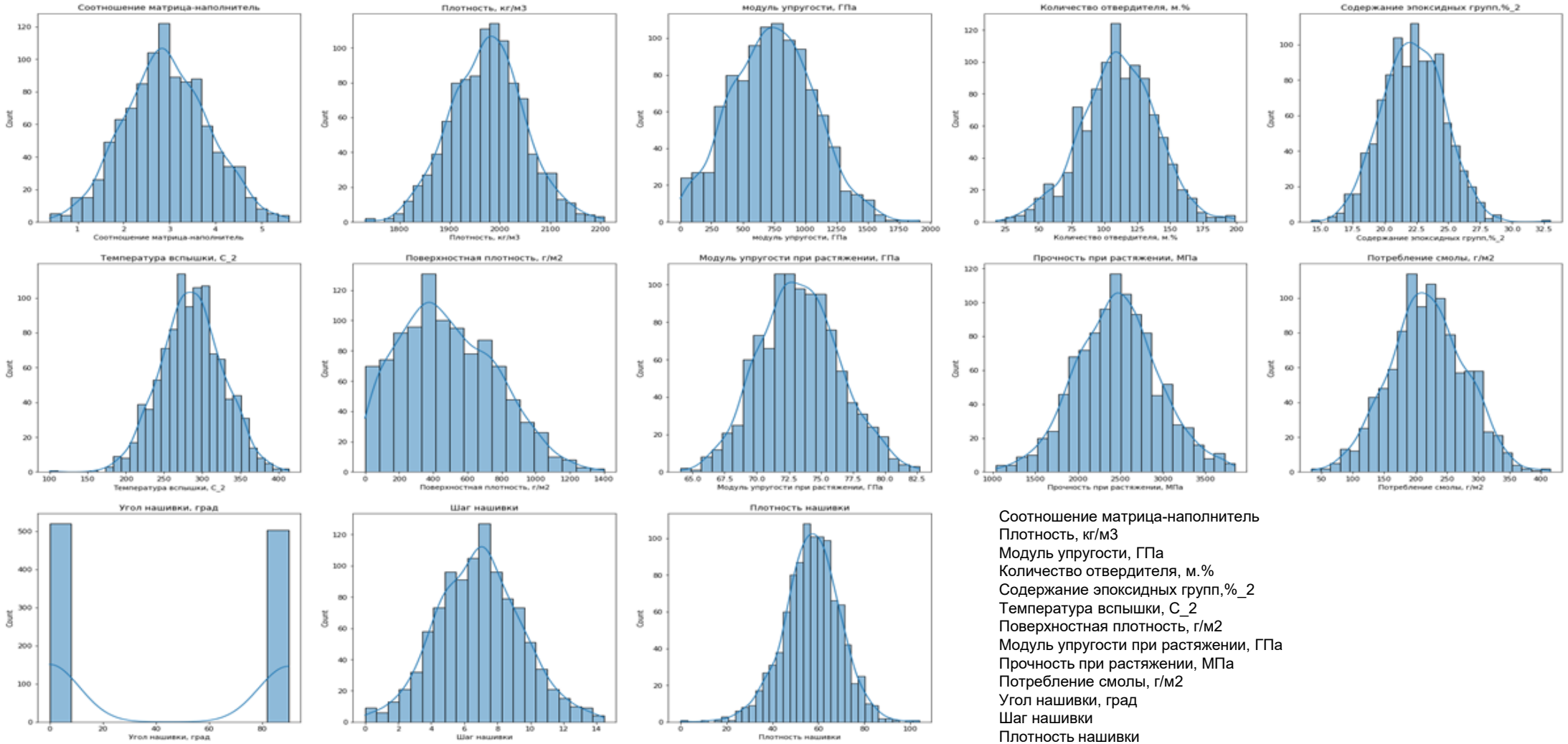
- количество элементов, средние, медианные значения параметров,
- минимальные, максимальные значения параметров, их среднеквадратичное отклонение и квартили.

```
df_descr = df.describe().T
df_descr['median'] = df.median()
df_descr.style.format(precision=4)
```

	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	57.1[illegible]9	12.3510	0.0000	49.7992	57.3419	64.9450	103.9889	57.3419

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

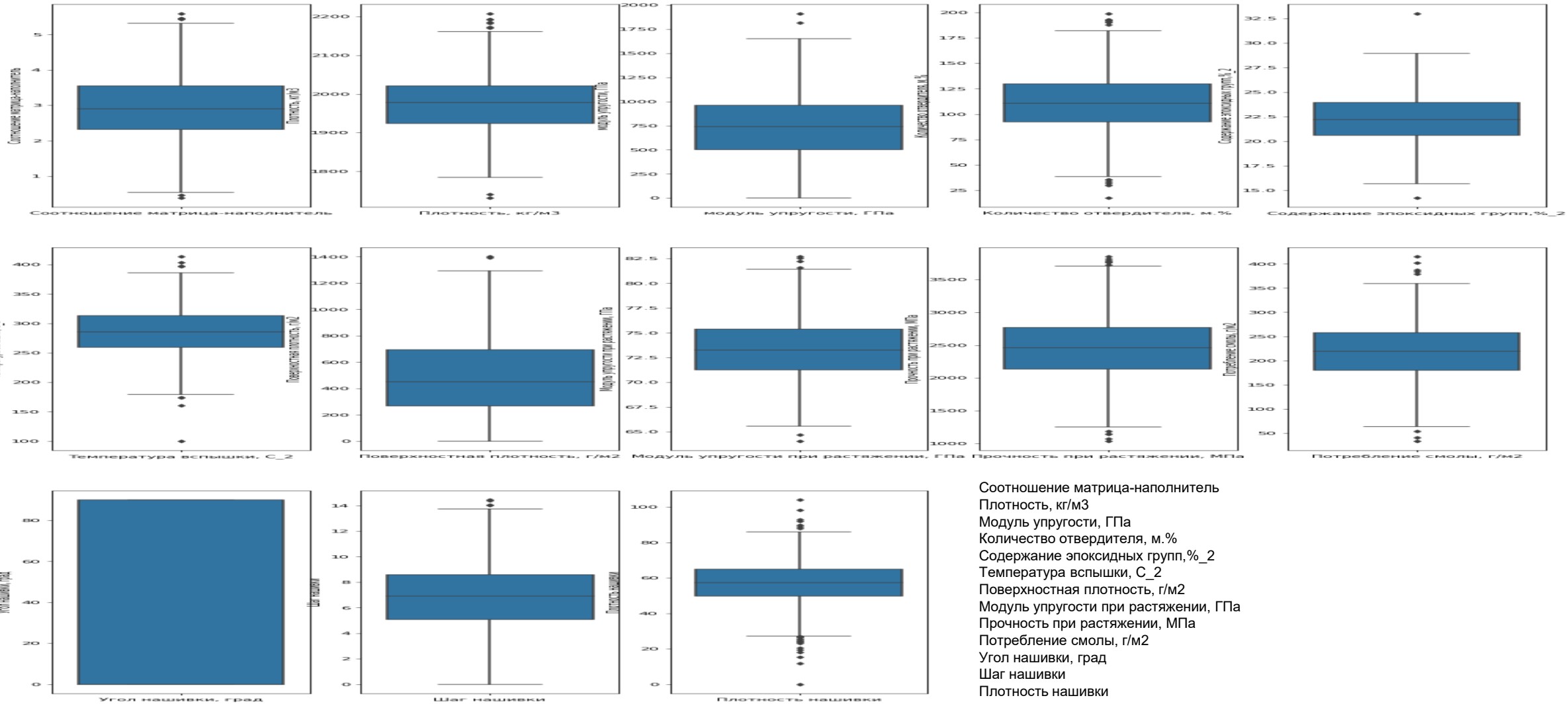
Гистограммы распределения



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

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Ящик с усами

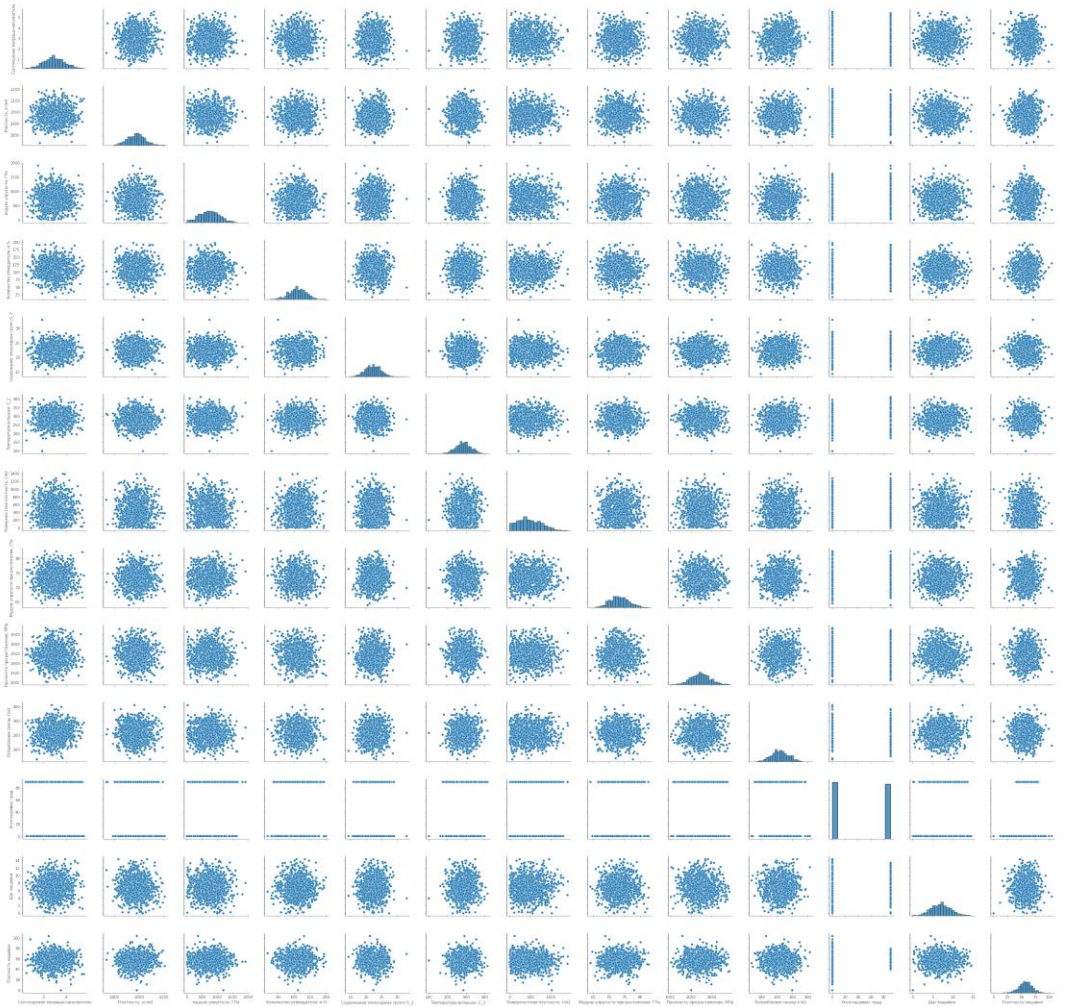


Соотношение матрица-наполнитель
Плотность, кг/м3
Модуль упругости, ГПа
Количество отвердителя, м.%
Содержание эпоксидных групп,%_2
Температура вспышки, С_2
Поверхностная плотность, г/м2
Модуль упругости при растяжении, ГПа
Прочность при растяжении, МПа
Потребление смолы, г/м2
Угол нашивки, град
Шаг нашивки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

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Попарные графики рассеивания точек

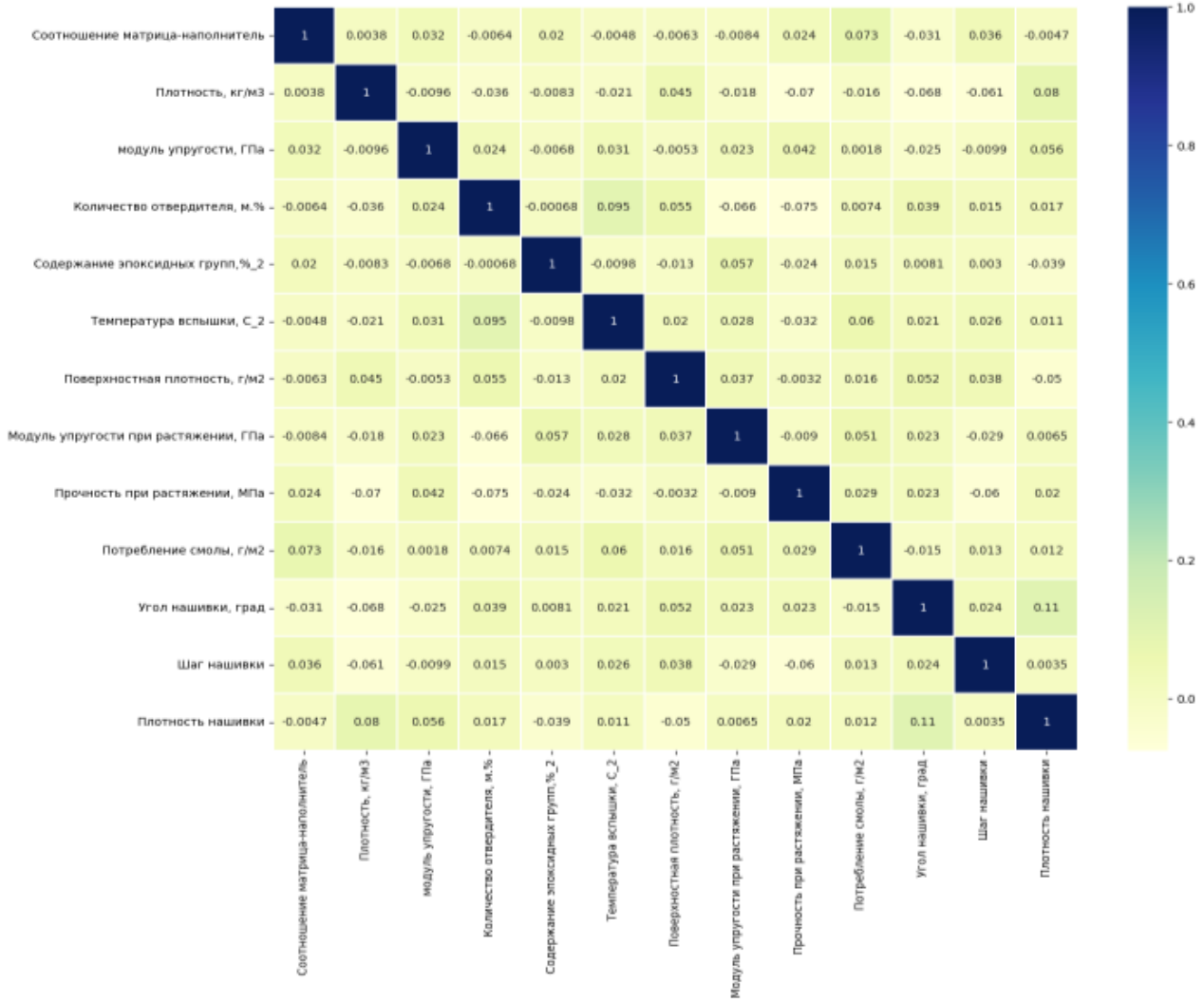
- Выбросы имеются
- Некоторые точки стоят очень далеко от облака
- Зависимости не просматриваются



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Тепловая матрица корреляции

- Коэффициенты корреляции близки к нулю;
- Самая высокая зависимость между углом нашивки и плотностью нашивки (0,11);
- Нормальное распределение у всех признаков;
- Линейная зависимость не просматривается.



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

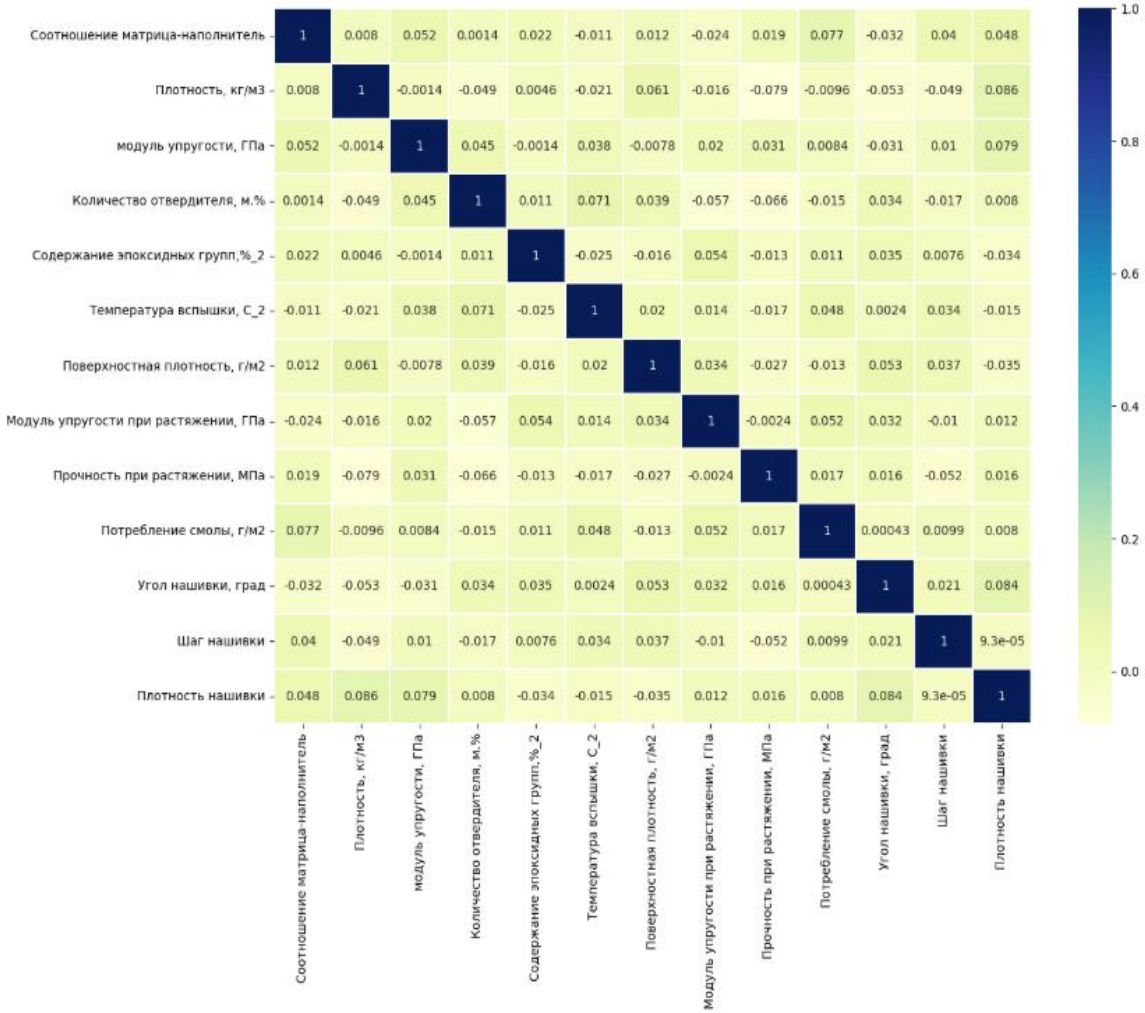
Выбросы

Найдено первоначально:

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

Для более точного построения модели использован для очистки метод межквартильных расстояний трижды:

- Осталось 922 строки очищенного датасета;
- Тепловая матрица корреляции изменилась незначительно;
- отсутствие линейной зависимости между признаками.



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Выходные переменные

	Модуль упругости при растяжении, ГПа
min	65.979990
max	81.203147
mean	73.342384
std	3.027444

	Прочность при растяжении, МПа
min	1250.392802
max	3636.892992
mean	2466.696221
std	459.451353

	Соотношение матрица-наполнитель
min	0.547391
max	5.314144
mean	2.925725
std	0.906983

Для каждого признака строим отдельная модель:

- модуль упругости при растяжении
- прочность при растяжении
- соотношение матрица-наполнитель

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Входные переменные

Значение входных признаков находятся в разных диапазонах => требуется препроцессинг

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

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

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

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Сравнение моделей машинного обучения

Метод	Интерпретируемость	Скорость обучения	Скорость предсказания	Устойчивость к шуму
DummyRegressor	Нет (тривиален)	Мгновенно	Мгновенно	—
Линейная регрессия	Очень высокая	Очень быстро	Очень быстро	Низкая
Ridge	Высокая	Быстро	Быстро	Средняя
LASSO	Высокая (разреж.)	Быстро	Быстро	Средняя
Дерево решений	Очень высокая	Быстро	Очень быстро	Низкая
Случайный лес	Низкая	Умеренно	Быстро	Высокая
Градиентный бустинг	Низкая	Медленно	Быстро	Высокая
SVM/SVR	Очень низкая	Медленно ($O(n^2–n^3)$)	Умеренно	Низкая
k-NN	Низкая	Нет обучения	Очень медленно ($O(n)$)	Низкая
Нейронные сети (MLP)	Очень низкая	Медленно (особенно без GPU)	Быстро (после обучения)	Средняя (с регуляризацией)

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Сравнение моделей машинного обучения

Метод	Требует масштабирования	Работает с нелинейностями	Отбор признаков	Априорные предпосылки
DummyRegressor	Нет	Нет	Нет	Нет
Линейная регрессия	Нет (но желательно)	Нет	Нет	Линейность, гомоскедастичность, независимость ошибок
Ridge	Да	Нет	Нет	Линейность, мультиколлинеарность допустима
LASSO	Да	Нет	Да	Линейность, разреженность истинной модели
Дерево решений	Нет	Да	Косвенно	Нет
Случайный лес	Нет	Да	Да (важность)	Нет
Градиентный бустинг	Нет	Да	Да	Нет (но лучше на табличных данных)
SVM/SVR	Да	Да (через ядра)	Нет	Нормализация, умеренный размер выборки
k-NN	Да	Да (локально)	Нет	Низкая размерность, нормализация, локальная гладкость
Нейронные сети (MLP)	Да	Да (глубоко и гибко)	Нет (но можно через attention/ важность)	Масштабирование, большой объём данных, числовые признаки

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель для модуля упругости при растяжении

Значения выхода: от 64 до 83

Лучшая модель:

- До подбора параметров Lasso R2 = -0,011
- После подбора параметров Ridge R2 = -0,007

Результаты моделей с гиперпараметрами по умолчанию

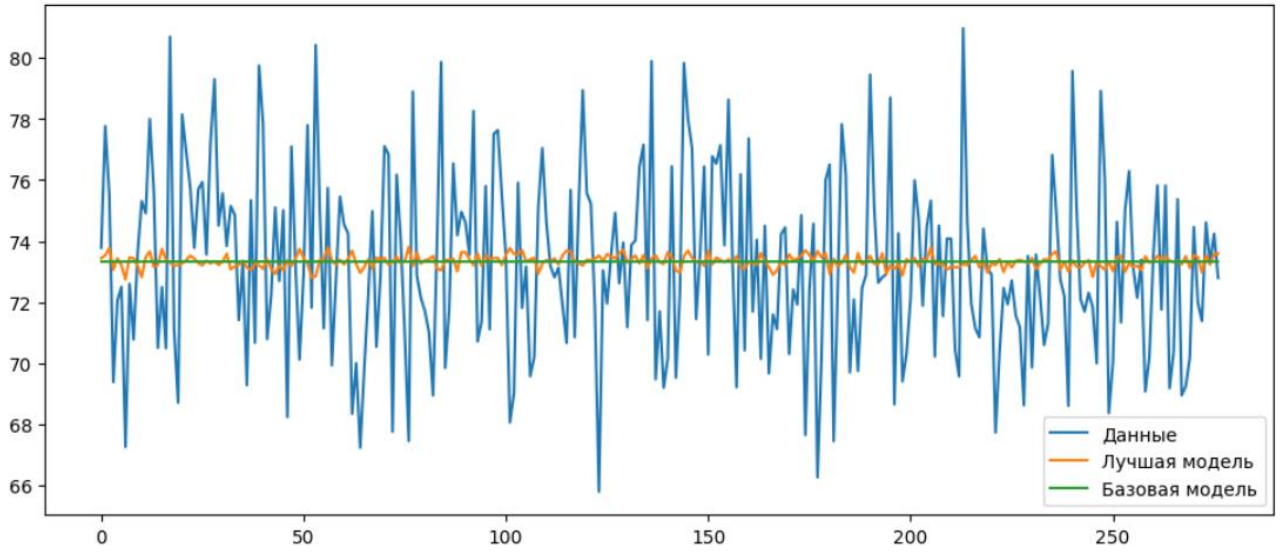
	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.011479	3.018468	2.434764	0.033263	-7.216742
LinearRegression	-0.022599	3.034619	2.453669	0.033520	-7.222509
Ridge	-0.022517	3.034496	2.453574	0.033519	-7.222363
Lasso	-0.011479	3.018468	2.434764	0.033263	-7.216742
SVR	-0.085891	3.124768	2.524366	0.034452	-7.445850
KNeighborsRegressor	-0.226448	3.311413	2.628943	0.035925	-8.330975
DecisionTreeRegressor	-1.317465	4.505732	3.660577	0.050026	-11.406496
RandomForestRegressor	-0.081441	3.117797	2.525393	0.034500	-7.342254
GradientBoostingRegressor	-0.132211	3.184520	2.554379	0.034910	-7.971823

Результаты моделей после подбора гиперпараметров

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=700, positive=True, solver='lbfgs')	-0.007231	3.011776	2.432033	0.033226	-7.126265
Lasso(alpha=0.15)	-0.009012	3.014177	2.429802	0.033195	-7.183302
SVR(C=0.02)	-0.012616	3.020104	2.437300	0.033289	-7.228111
KNeighborsRegressor(n_neighbors=29)	-0.059927	3.087279	2.493132	0.034113	-7.254115
DecisionTreeRegressor(max_depth=1, max_features=1, random_state=3128, splitter='random')	-0.011451	3.018141	2.426078	0.033145	-7.182651
RandomForestRegressor(bootstrap=False, criterion='absolute_error', max_depth=2, max_features=1, n_estimators=50, random_state=3128)	-0.011578	3.018652	2.438736	0.033310	-7.189810

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель для модуля упругости при растяжении



	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.001840	3.023023	2.480618	0.033942	7.628576
Лучшая модель (Ridge)	-0.009474	3.034520	2.475754	0.033873	7.730087

	R2	RMSE	MAE	MAPE	max_error
Модуль упругости, тренировочный	0.013531	3.004559	2.419611	0.033056	7.680848
Модуль упругости, тестовый	-0.009474	3.034520	2.475754	0.033873	7.730087

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель для прочности при растяжении

Значения выхода: 1071 до 3849

Лучшая модель:

- До подбора параметров SVR R2 = -0,013
- После подбора параметров Lasso R2 = -0,006

Результаты моделей с гиперпараметрами по умолчанию

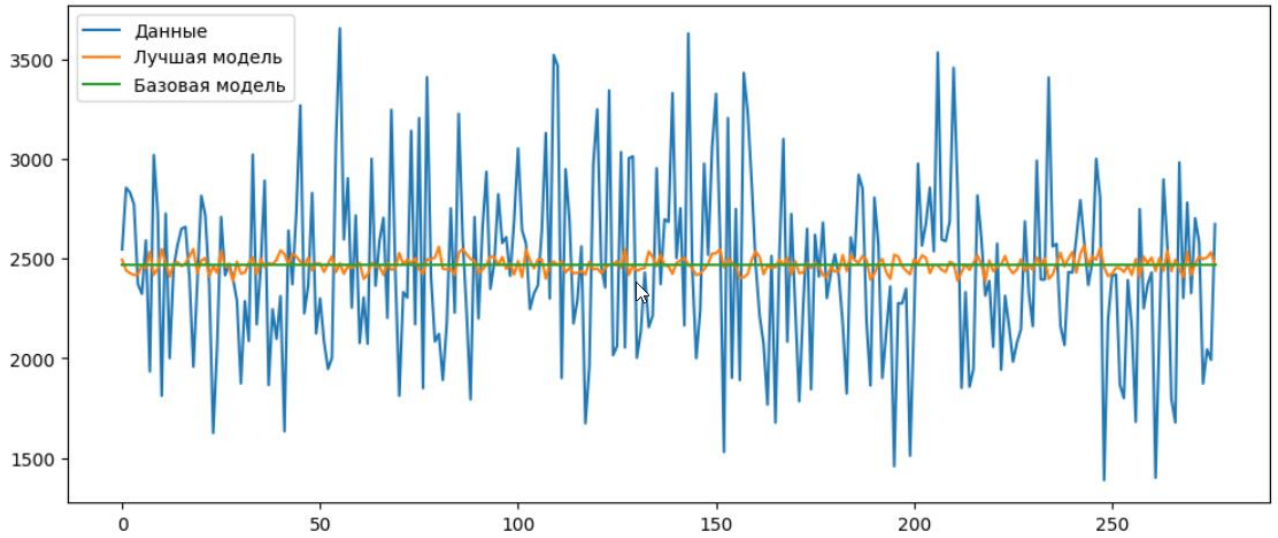
	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.014754	458.517408	366.664655	0.160630	-1095.981614
LinearRegression	-0.017429	459.567129	368.992333	0.161357	-1121.299260
Ridge	-0.017348	459.547973	368.974704	0.161350	-1121.197755
Lasso	-0.014644	458.916723	368.444713	0.161130	-1118.923935
SVR	-0.012796	458.101333	366.555234	0.160547	-1094.668346
DecisionTreeRegressor	-1.101336	648.893181	524.028712	0.223497	-1648.797617
GradientBoostingRegressor	-0.121957	482.055962	389.345452	0.169754	-1194.692003

Результаты моделей после подбора гиперпараметров

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=990, solver='lsqr')	-0.008290	457.219991	366.133785	0.160339	-1092.435894
Lasso(alpha=20)	-0.005856	456.617587	365.960810	0.160195	-1090.504240
SVR(C=0.02, kernel='linear')	-0.012688	458.075057	366.548778	0.160541	-1094.267773
DecisionTreeRegressor(max_depth=1, max_features=3, random_state=3128)	-0.020410	459.733570	365.998753	0.160581	-1107.058869
GradientBoostingRegressor(loss='absolute_error', max_depth=1, max_features=11, n_estimators=50, random_state=3128)	-0.022204	460.206446	368.817375	0.161594	-1105.850304

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель для прочности при растяжении



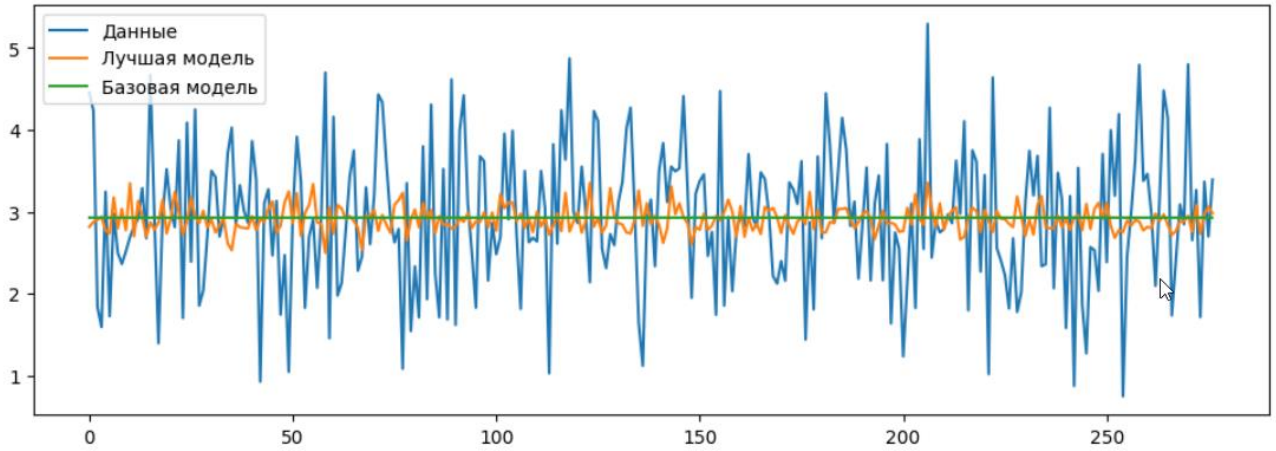
	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.001555	439.676848	350.354301	0.151168	1187.738138
Лучшая модель (Lasso)	-0.011555	441.866376	350.404421	0.151146	1178.211769

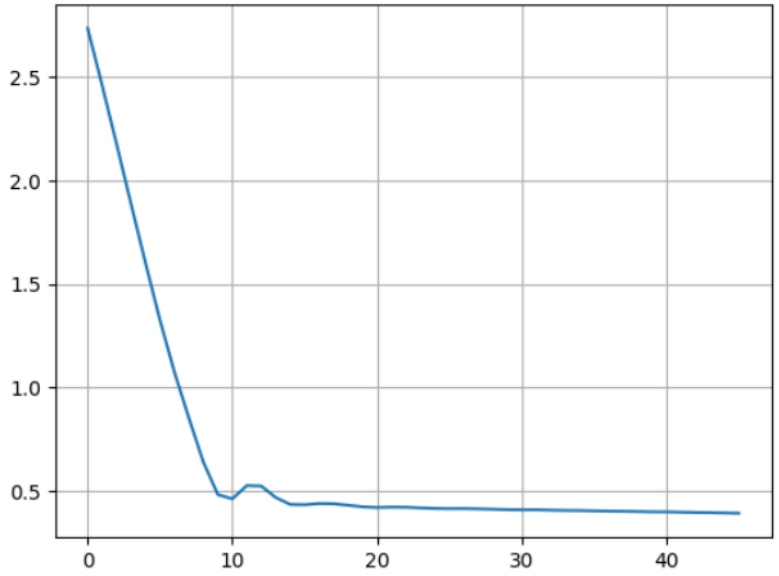
	R2	RMSE	MAE	MAPE	max_error
Прочность при растяжении, тренировочный	0.018043	454.934526	363.717442	0.159240	1295.547126
Прочность при растяжении, тестовый	-0.011555	441.866376	350.404421	0.151146	1178.211769

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель для соотношения матрица-наполнитель

MLPRegressor из библиотеки scikit-learn:
- 8 слоев;
- нейронов на каждом слое 24;
- активационная функция relu;
- оптимизатор: adam;
- пропорция разбиения данных на тестовые и валидационные 30%;
- ранняя остановка;
- количество итераций: 5000.





	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.000074	0.868130	0.691547	0.296759	2.370117
MLPRegressor	-0.064119	0.895496	0.722489	0.305808	2.201930

Значения выхода от 0,39 до 5,46

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель для соотношения матрица-наполнитель

Нейросеть из библиотеки TensorFlow:

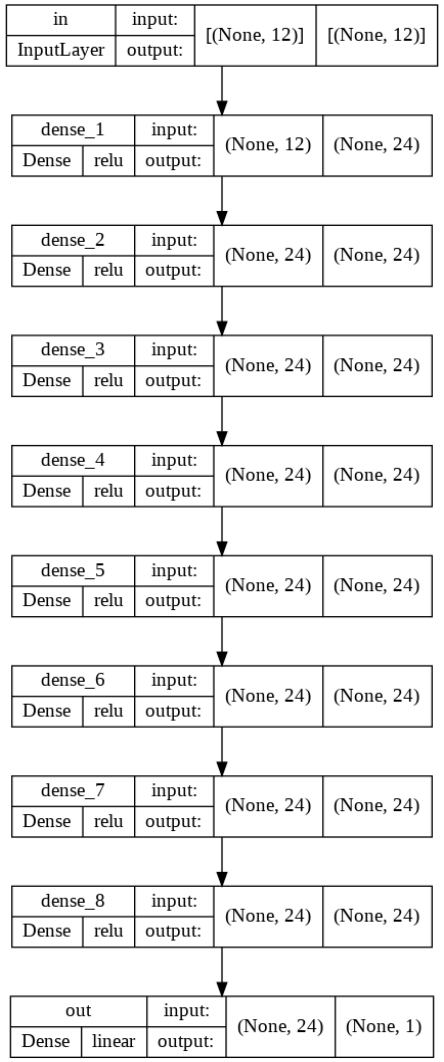
- входной слой для 12 признаков;
- выходной слой для 1 признака;
- скрытых слоев 8;
- нейронов на каждом скрытом слое 24;
- активационная функция скрытых слоев relu;
- оптимизатор Adam;
- loss-функция MeanAbsolutePercentageError

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 (17.72 KB)

Trainable params: 4,537 (17.72 KB)

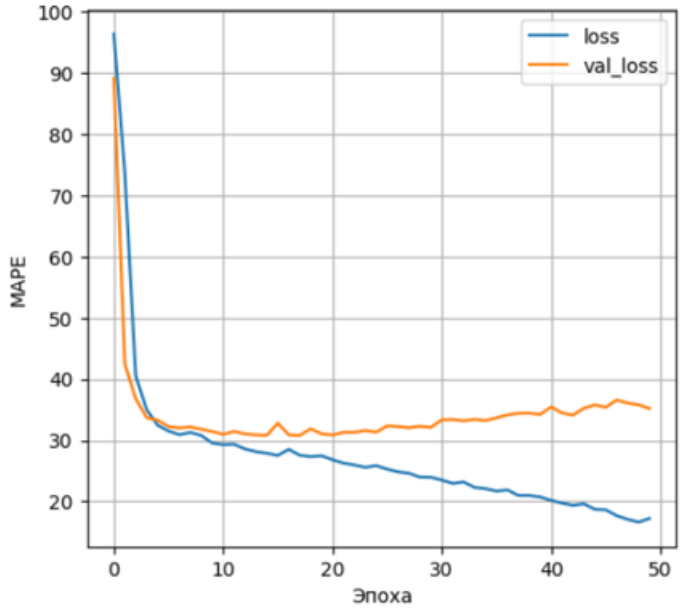
Non-trainable params: 0 (0.00 B)



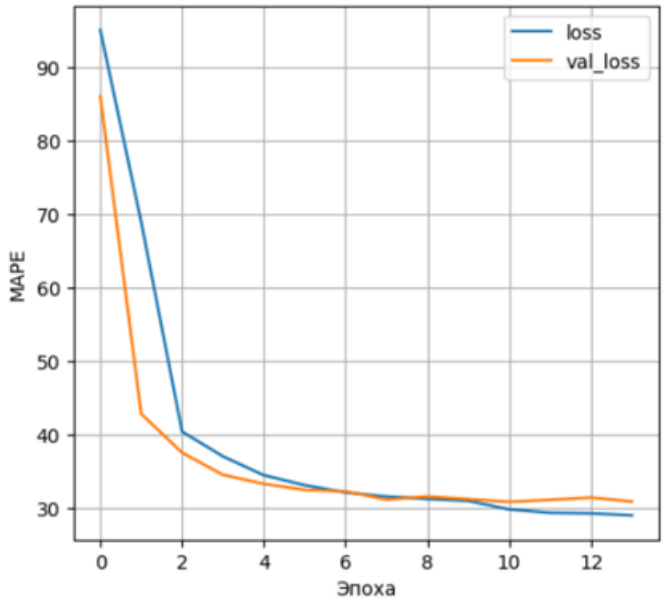
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель для соотношения матрица-наполнитель

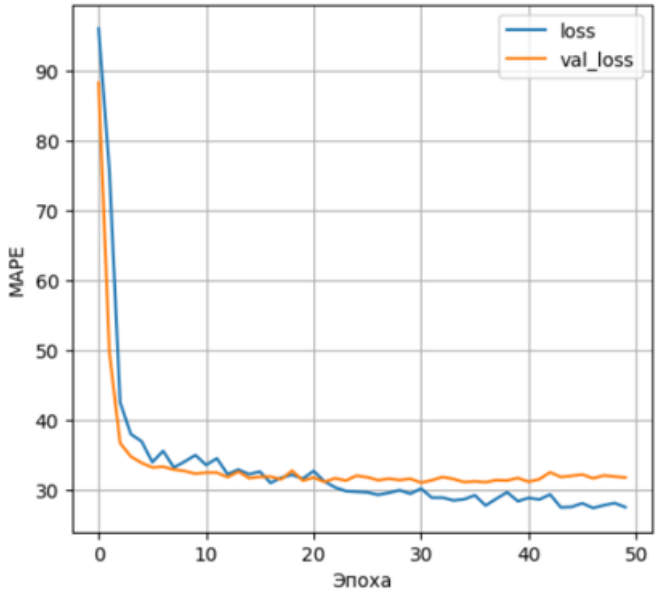
Обучение нейросети



Борьба с переобучением: ранняя остановка



Борьба с переобучением: Dropout



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

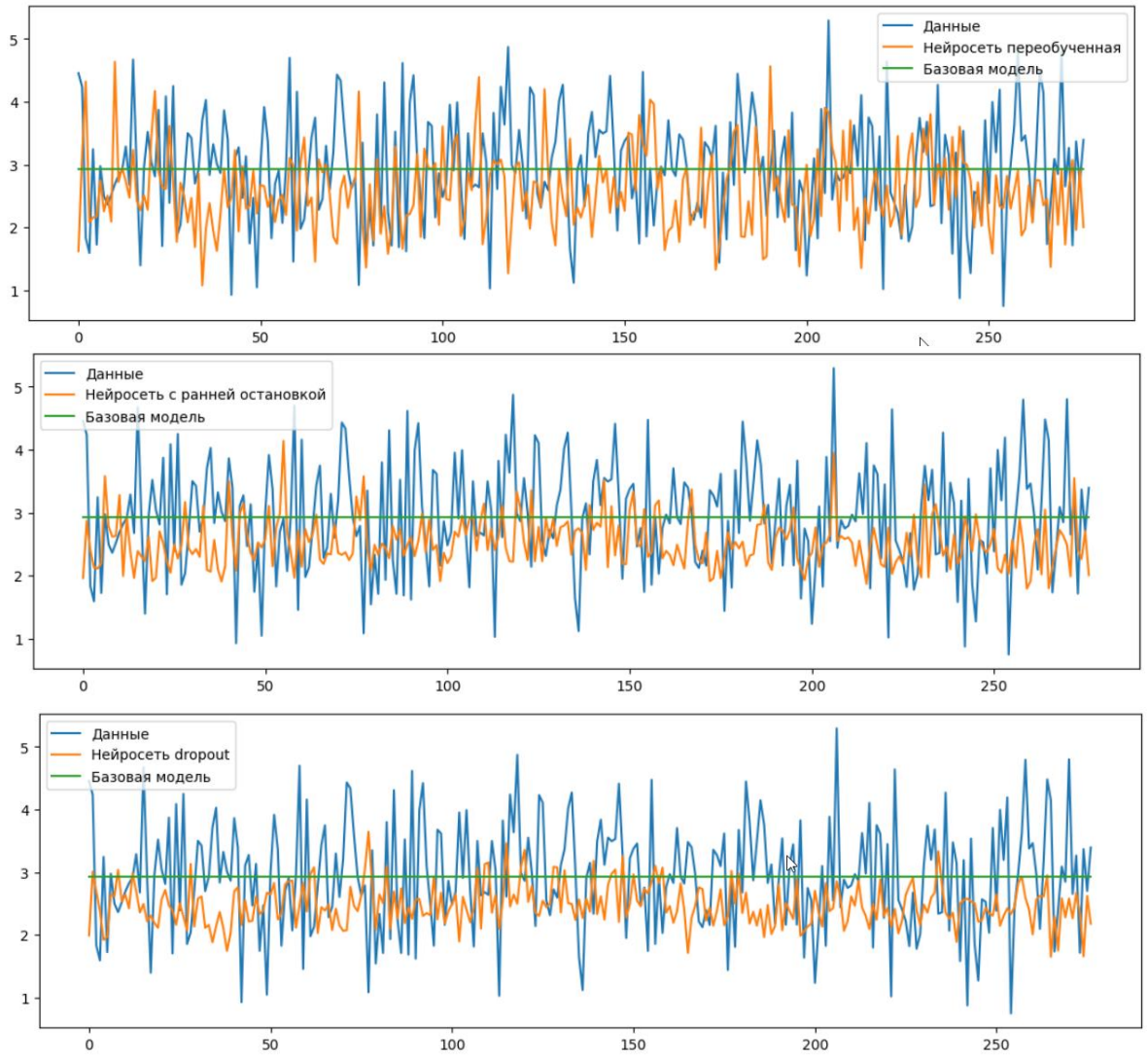
Модель для соотношения матрица-наполнитель

Значения выхода от 0,39 до 5,46

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.000074	0.868130	0.691547	0.296759	2.370117
Нейросеть переобученная	-0.830760	1.174585	0.957319	0.361797	3.604231
Нейросеть с ранней остановкой	-0.462496	1.049823	0.847659	0.311427	2.732753
Нейросеть dropout	-0.468733	1.052059	0.856727	0.310578	2.559960

Выбрана нейросеть, обученная с ранней остановкой

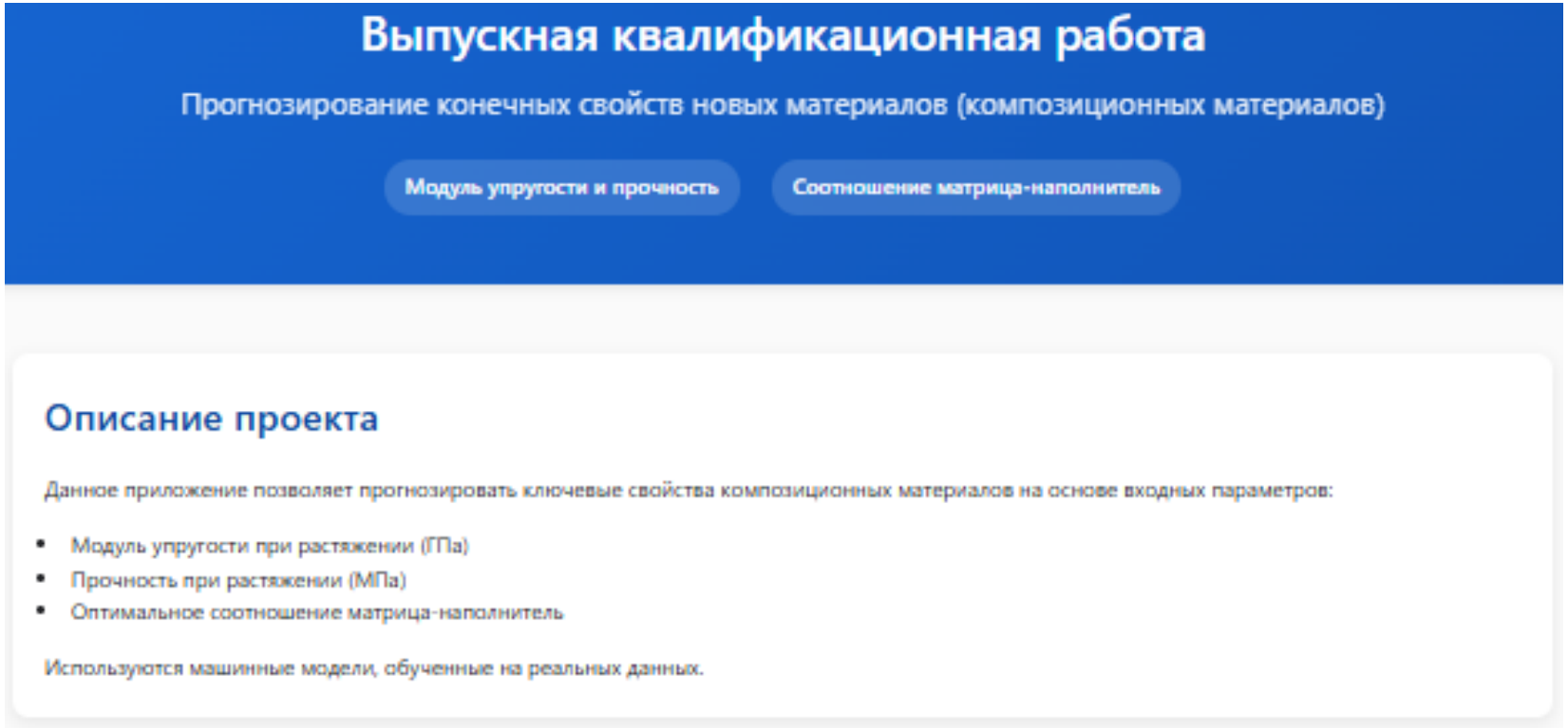
	R2	RMSE	MAE	MAPE	max_error
Соотношение матрица-наполнитель, тренировочный	-0.278531	1.024750	0.805735	0.292013	3.013253
Соотношение матрица-наполнитель, тестовый	-0.462496	1.049823	0.847659	0.311427	2.732753



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

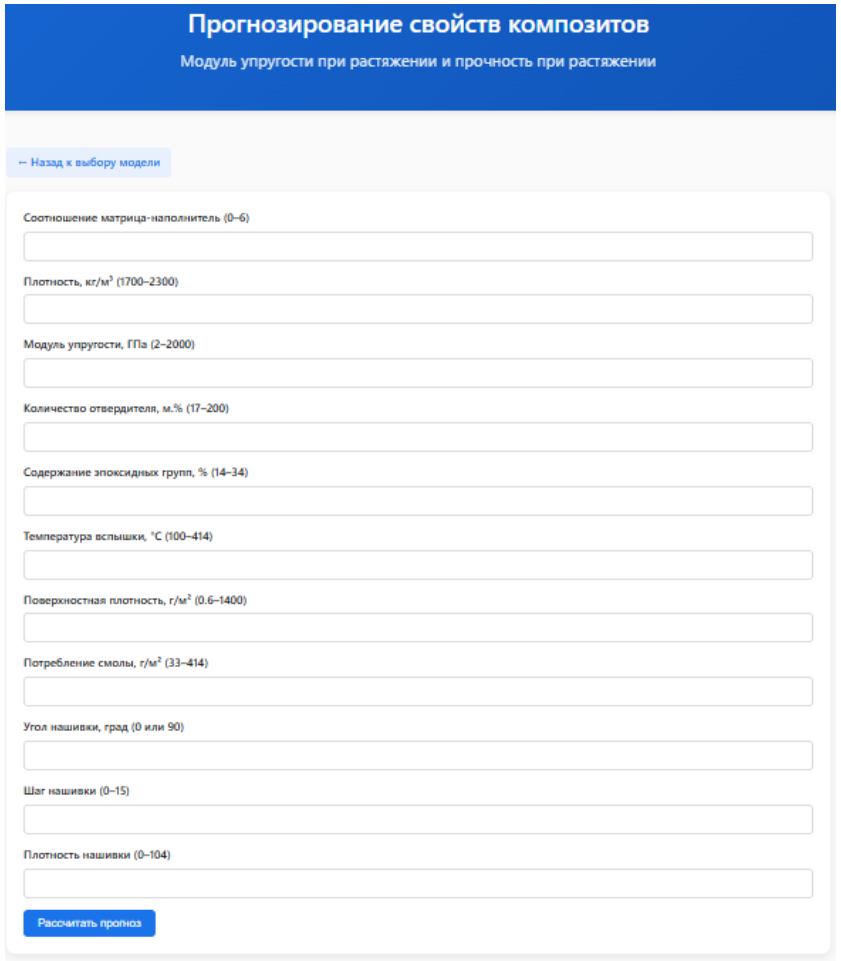
Разработка веб-приложения

Стартовая страница приложения

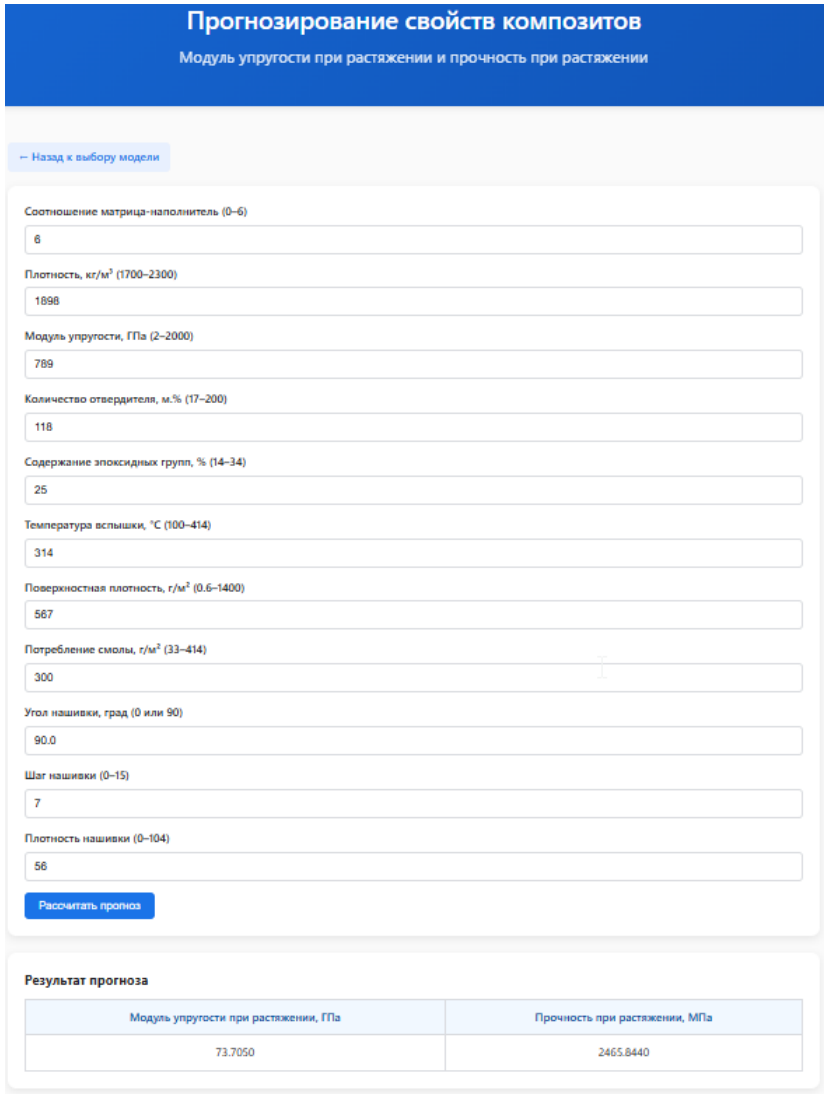


ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка веб-приложения



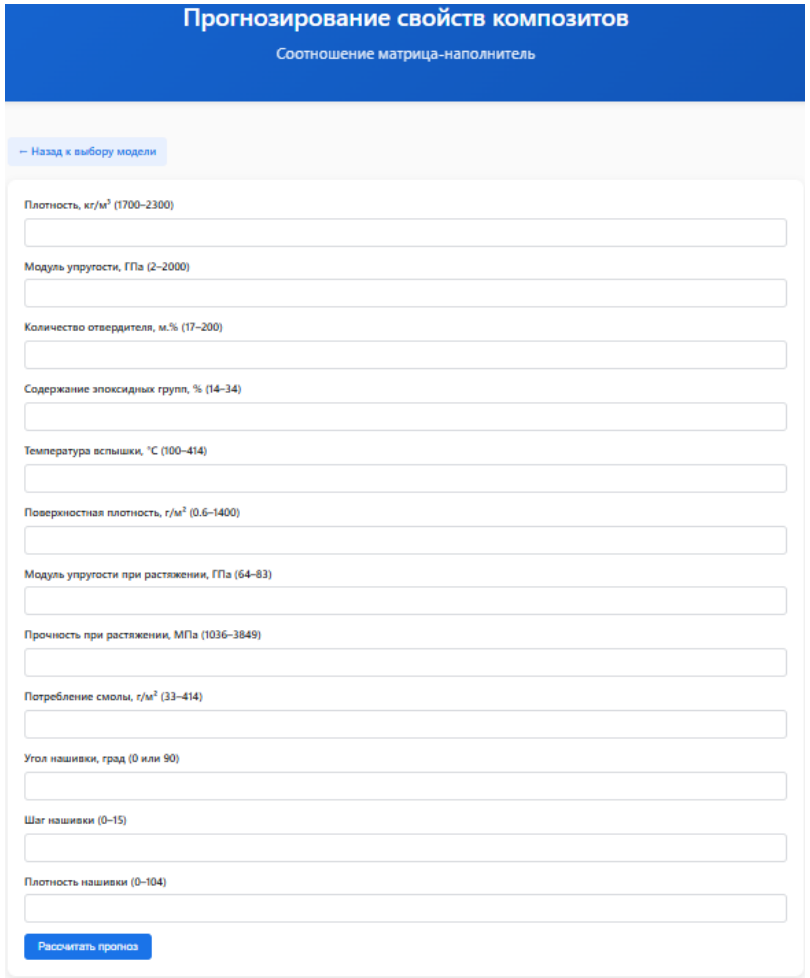
Ввод входных параметров



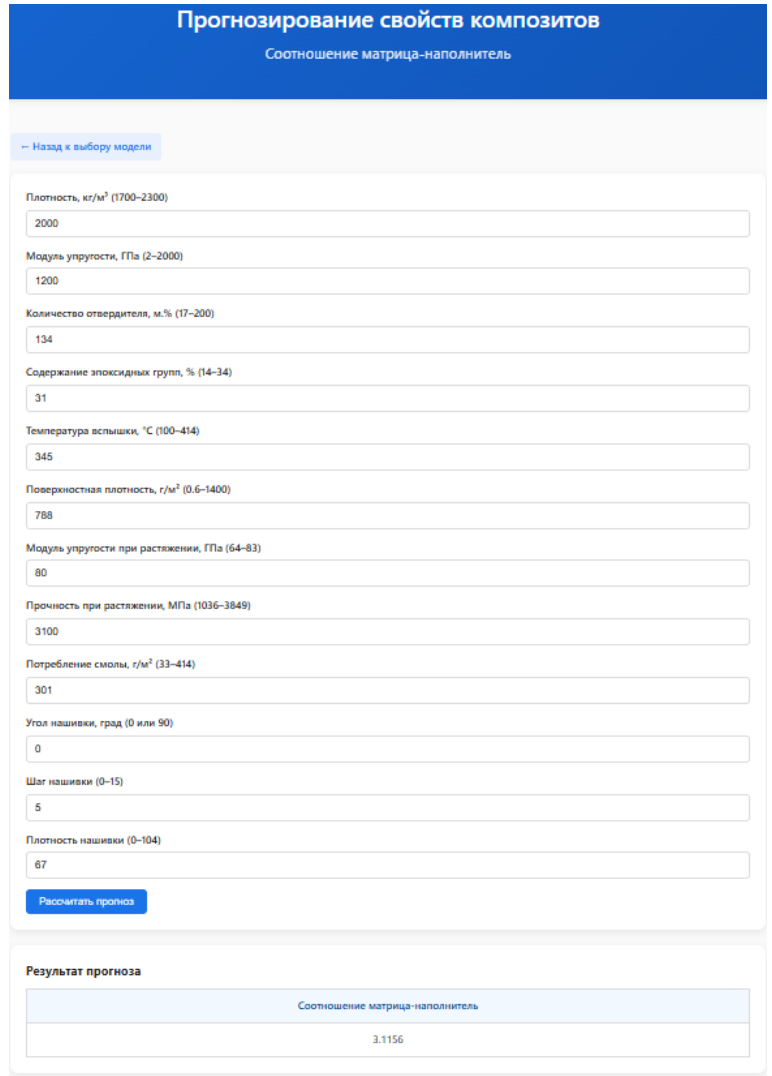
Вывод результата

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка веб-приложения



Ввод входных параметров



Вывод результата

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Итоговые результаты работы

Задача в целом не решена, модели не оптимальны

Дальнейшие поиски решения могут включать:

- консультация у экспертов в предметной области
- углубление в изучение нейросетей, использование различной архитектуры, параметров обучения
- исследовать сырые данные, использовать другие методы очистки и подготовки
- провести отбор признаков и уменьшение размерности
- провести тщательный подбор гиперпараметров градиентного бустинга

ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru