Образовательный центр МГТУ им. Н.Э. Баумана

Выпускная квалификационная работа
по курсу
«Data Science»

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель: Костромина О.С.

Основные задачи

- Подготовка данных
- Разведочный анализ данных
- Предобработка данных
- Построение, обучение и тестирование моделей для прогноза модуля упругости при растяжении и прочности при растяжении
- Построение, обучение и тестирование нейронных сетей для прогноза соотношения матрица-наполнитель
- Разработка приложения для прогнозирования конечных свойств композиционных материалов

Подготовка данных

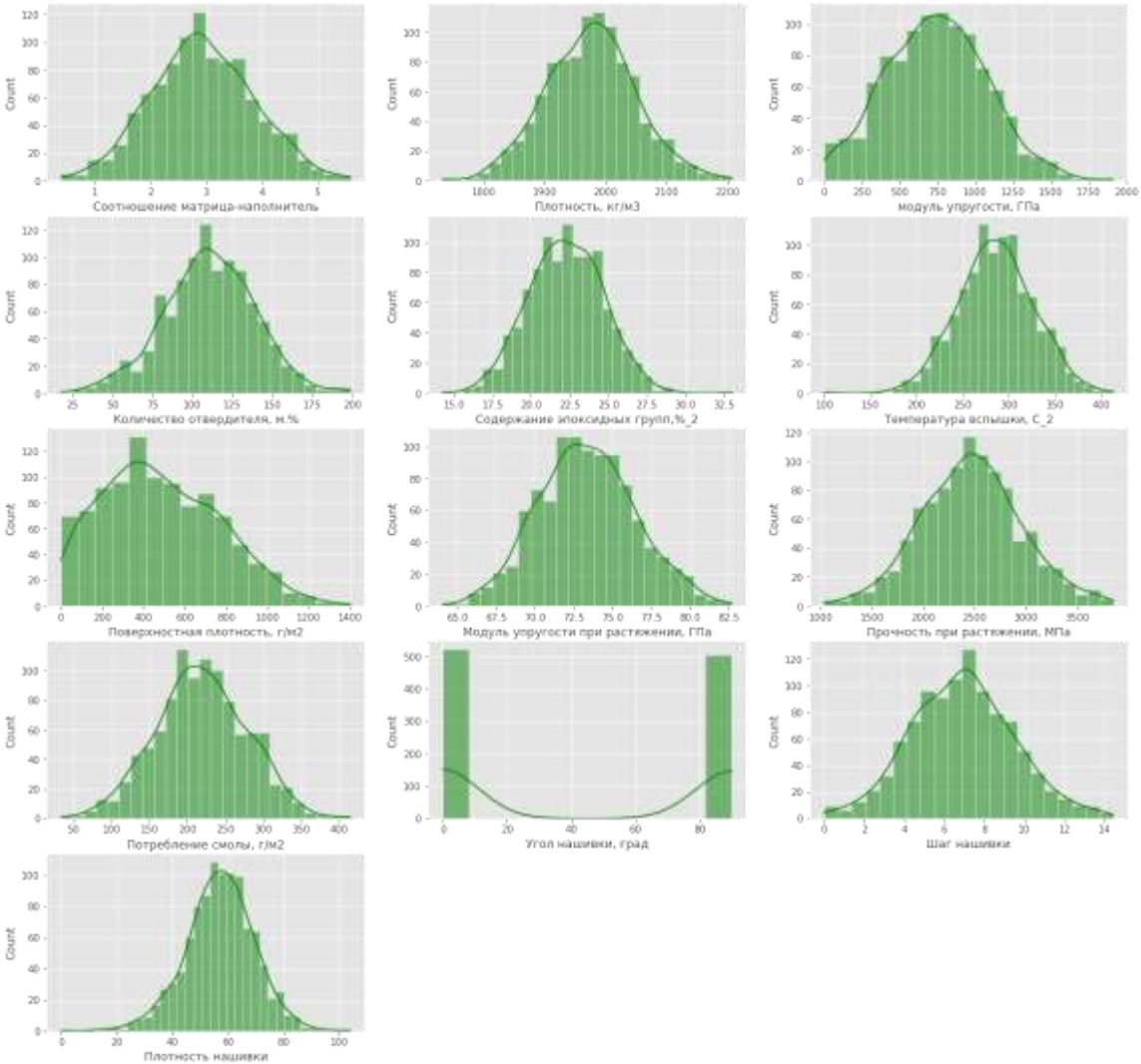
	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	1.857143	2030.000000	738.736842	30.000000	22.267857	100.000000	210.000000	70.000000	3000.000000	220.000000	0.0	4.000000	57.000000
1	1.857143	2030.000000	738.736842	50.000000	23.750000	284.615385	210.000000	70.000000	3000.000000	220.000000	0.0	4.000000	60.000000
2	1.857143	2030.000000	738.736842	49.900000	33.000000	284.615385	210.000000	70.000000	3000.000000	220.000000	0.0	4.000000	70.000000
3	1.857143	2030.000000	738.736842	129.000000	21.250000	300.000000	210.000000	70.000000	3000.000000	220.000000	0.0	5.000000	47.000000
4	2.771331	2030.000000	753.000000	111.860000	22.267857	284.615385	210.000000	70.000000	3000.000000	220.000000	0.0	5.000000	57.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1018	2.271346	1952.087902	912.855545	86.992183	20.123249	324.774576	209.198700	73.090961	2387.292495	125.007669	90.0	9.076380	47.019770
1019	3.444022	2050.089171	444.732634	145.981978	19.599769	254.215401	350.660830	72.920827	2360.392784	117.730099	90.0	10.565614	53.750790
1020	3.280604	1972.372865	416.836524	110.533477	23.957502	248.423047	740.142791	74.734344	2662.906040	236.606764	90.0	4.161154	67.629684
1021	3.705351	2066.799773	741.475517	141.397963	19.246945	275.779840	641.468152	74.042708	2071.715856	197.126067	90.0	6.313201	58.261074
1022	3.808020	1890.413468	417.316232	129.183416	27.474763	300.952708	758.747882	74.309704	2856.328932	194.754342	90.0	6.078902	77.434468

1023 rows × 13 columns

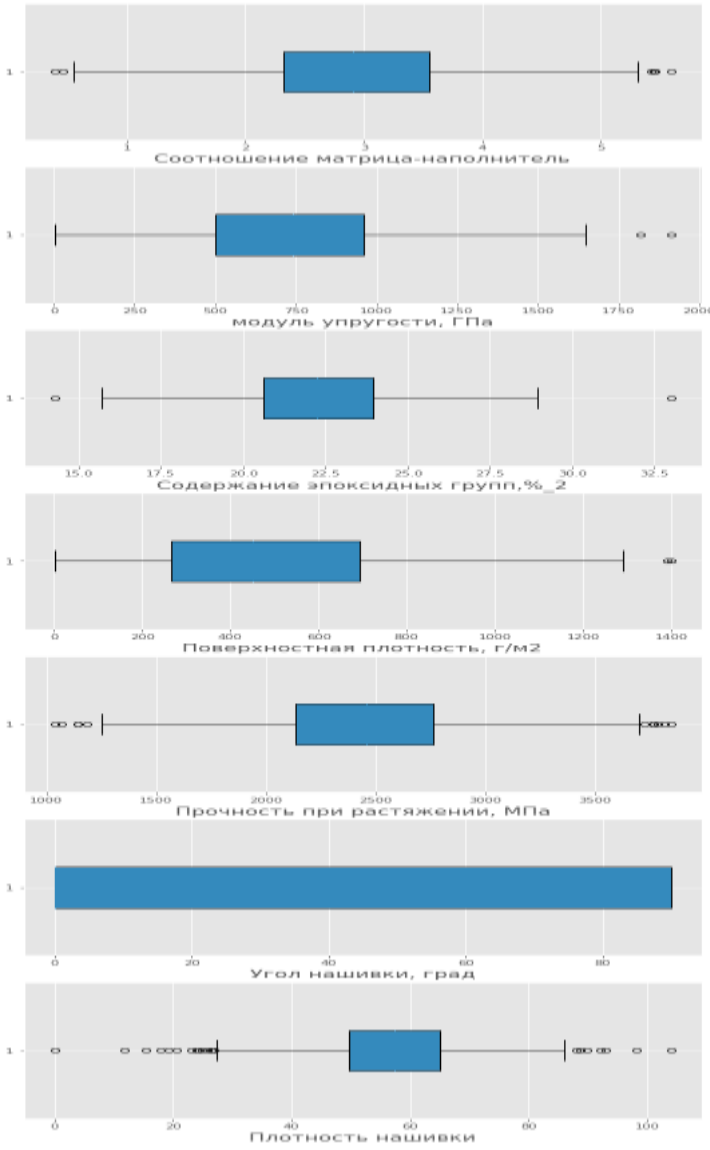
	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.930366	0.913222	0.389403	2.317887	2.906878	3.552660	5.591742
Плотность, кг/м3	1023.0	1975.734888	73.729231	1731.764635	1924.155467	1977.621657	2021.374375	2207.773481
модуль упругости, ГПа	1023.0	739.923233	330.231581	2.436909	500.047452	739.664328	961.812526	1911.536477
Количество отвердителя, м.%	1023.0	110.570769	28.295911	17.740275	92.443497	110.564840	129.730366	198.953207
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301	14.254985	20.608034	22.230744	23.961934	33.000000
Температура вспышки, С_2	1023.0	285.882151	40.943260	100.000000	259.066528	285.896812	313.002106	413.273418
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690	0.603740	266.816645	451.864365	693.225017	1399.542362
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983	64.054061	71.245018	73.268805	75.356612	82.682051
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006	1036.856605	2135.850448	2459.524526	2767.193119	3848.436732
Потребление смолы, г/м2	1023.0	218.423144	59.735931	33.803026	179.627520	219.198882	257.481724	414.590628
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1023.0	6.899222	2.563467	0.000000	5.080033	6.916144	8.586293	14.440522
Плотность нашивки	1023.0	57.153929	12.350969	0.000000	49.799212	57.341920	64.944961	103.988901

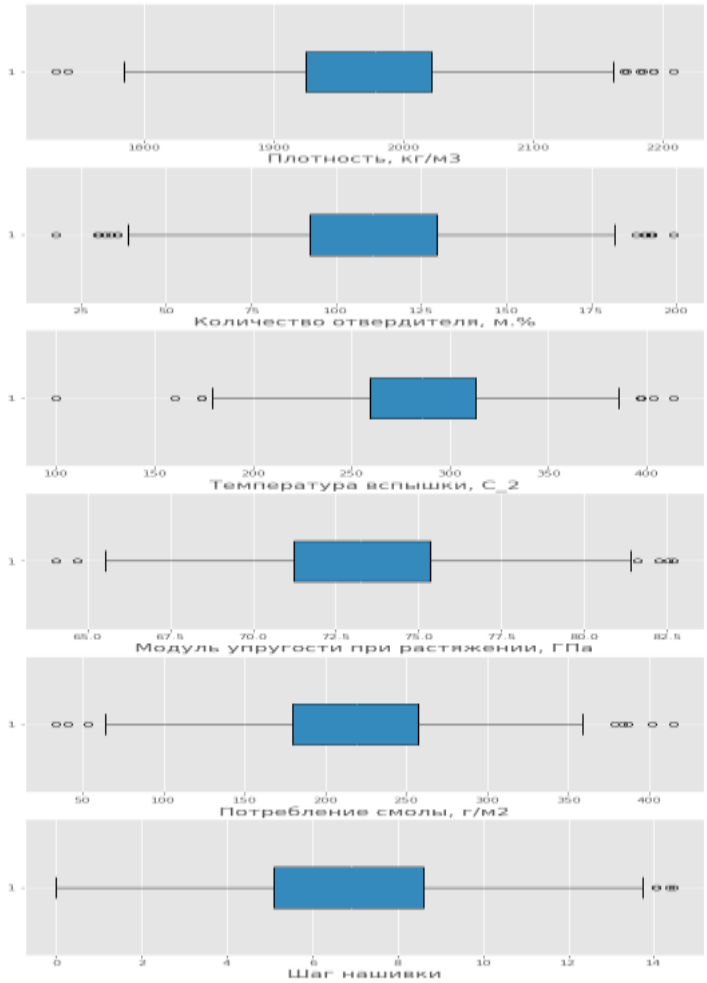
Разведочный анализ данных

Гистограммы распределения



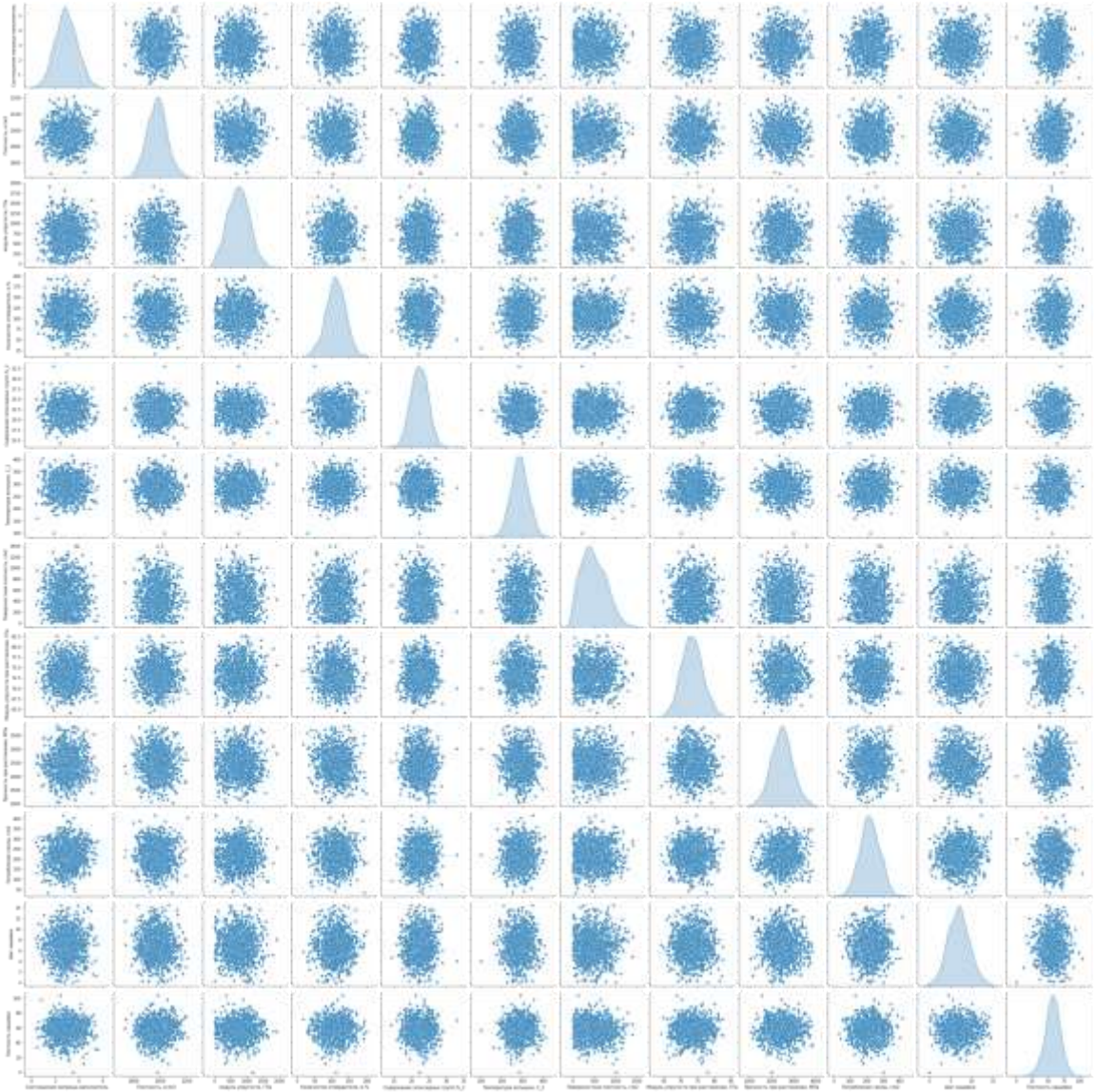
Разведочный анализ данных
«Ящики с усами»



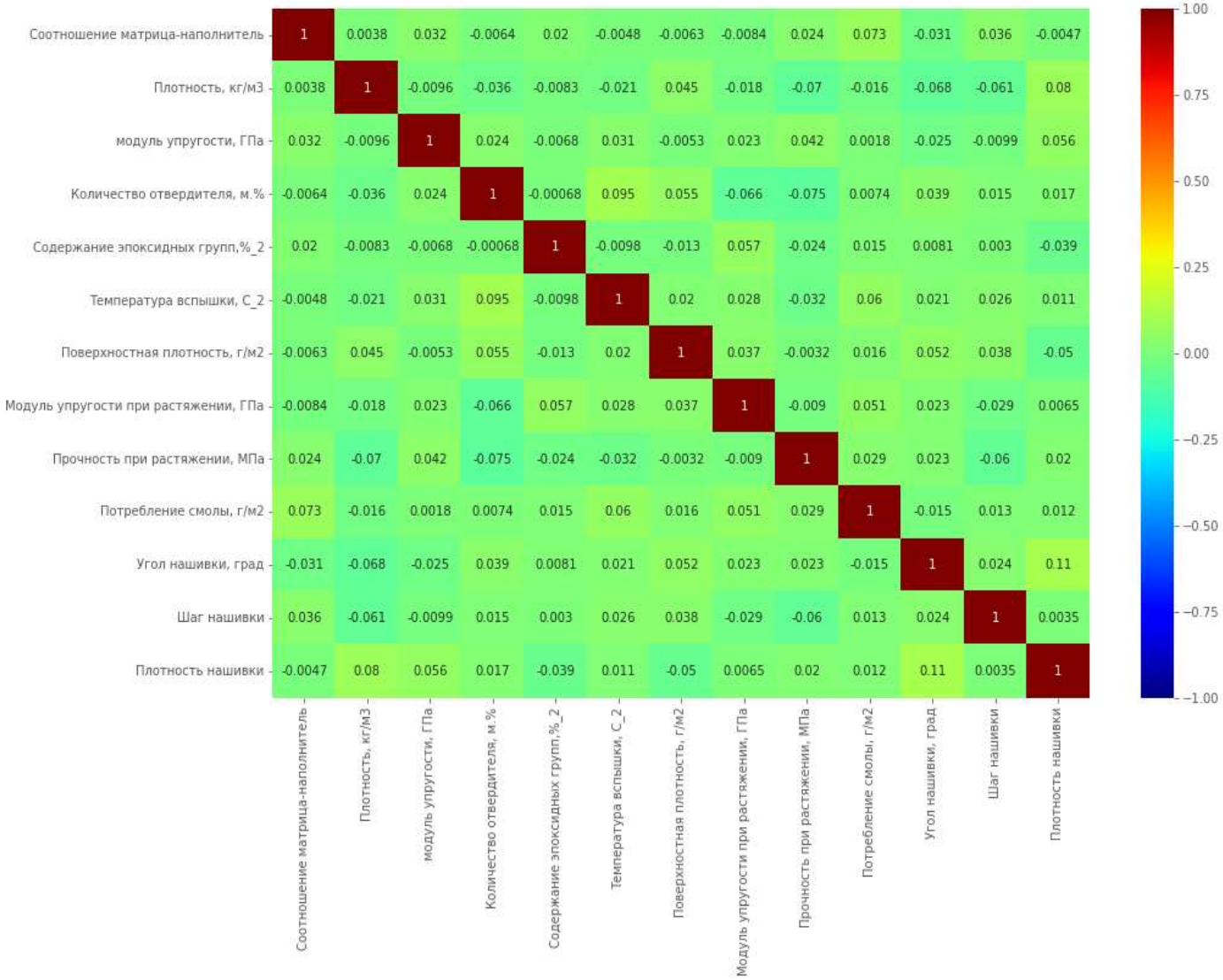


Разведочный анализ данных

Матрица рассеяний



Разведочный анализ данных
Матрица корреляций



Предобработка данных
Удаление выбросов

- Предобработка данных состоит из трех этапов: анализ данных на пропуски и дубликаты, удаление выбросов и нормализация данных.
- Пропусков, как и дубликатов в нашем датасете не обнаружено.
- Удаление выбросов происходит на основе межквартильного расстояния.

```
Соотношение матрица-наполнитель          6
Плотность, кг/м3                         9
модуль упругости, ГПа                    2
Количество отвердителя, м.%             14
Содержание эпоксидных групп,%_2          2
Температура вспышки, С_2                 8
Поверхностная плотность, г/м2            2
Модуль упругости при растяжении, ГПа     6
Прочность при растяжении, МПа           11
Потребление смолы, г/м2                  8
Угол нашивки, град                       0
Шаг нашивки                              4
Плотность нашивки                       21
dtype: int64
```

Предобработка данных

Нормализация с помощью MinMaxScaler

	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
0	0.274768	0.651097	0.452951	0.079153	0.607435	0.509164	0.162230	0.272962	0.727777	0.514688	0.0	0.289334	0.546433
1	0.274768	0.651097	0.452951	0.630983	0.418887	0.583596	0.162230	0.272962	0.727777	0.514688	0.0	0.362355	0.319758
2	0.466552	0.651097	0.461725	0.511257	0.495653	0.509164	0.162230	0.272962	0.727777	0.514688	0.0	0.362355	0.494123
3	0.465836	0.571539	0.458649	0.511257	0.495653	0.509164	0.162230	0.272962	0.727777	0.514688	0.0	0.362355	0.546433
4	0.424236	0.332865	0.494944	0.511257	0.495653	0.509164	0.162230	0.272962	0.727777	0.514688	0.0	0.362355	0.720799
...	...	...	...	...	...	...	...	...	...	...	...	...	...
917	0.361662	0.444480	0.560064	0.337550	0.333908	0.703458	0.161609	0.473953	0.472912	0.183151	1.0	0.660014	0.320103
918	0.607674	0.704373	0.272088	0.749605	0.294428	0.362087	0.271207	0.462512	0.461722	0.157752	1.0	0.768759	0.437468
919	0.573391	0.498274	0.254927	0.501991	0.623085	0.334063	0.572959	0.580201	0.587558	0.572648	1.0	0.301102	0.679468
920	0.662497	0.748688	0.454635	0.717585	0.267818	0.466417	0.496511	0.535317	0.341643	0.434855	1.0	0.458245	0.516112
921	0.684036	0.280923	0.255222	0.632264	0.888354	0.588206	0.587373	0.552644	0.668015	0.426577	1.0	0.441137	0.850430

922 rows × 13 columns

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	922.0	0.499412	0.187858	0.0	0.371909	0.495189	0.629774	1.0
Плотность, кг/м3	922.0	0.502904	0.186395	0.0	0.368184	0.511396	0.624719	1.0
модуль упругости, ГПа	922.0	0.451341	0.201534	0.0	0.305188	0.451377	0.587193	1.0
Количество отвердителя, м.%	922.0	0.506200	0.186876	0.0	0.378514	0.506382	0.638735	1.0
Содержание эпоксидных групп,%_2	922.0	0.490578	0.180548	0.0	0.366571	0.488852	0.623046	1.0
Температура вспышки, С_2	922.0	0.516739	0.190721	0.0	0.386228	0.516931	0.646553	1.0
Поверхностная плотность, г/м2	922.0	0.373295	0.217269	0.0	0.204335	0.354161	0.538397	1.0
Модуль упругости при растяжении, ГПа	922.0	0.487343	0.196366	0.0	0.353512	0.483718	0.617568	1.0
Прочность при растяжении, МПа	922.0	0.503776	0.188668	0.0	0.373447	0.501481	0.624299	1.0
Потребление смолы, г/м2	922.0	0.507876	0.199418	0.0	0.374647	0.510143	0.642511	1.0
Угол нашивки, град	922.0	0.510846	0.500154	0.0	0.000000	1.000000	1.000000	1.0
Шаг нашивки	922.0	0.503426	0.183587	0.0	0.372844	0.506414	0.626112	1.0
Плотность нашивки	922.0	0.503938	0.193933	0.0	0.376869	0.504310	0.630842	1.0

Построение и обучение моделей

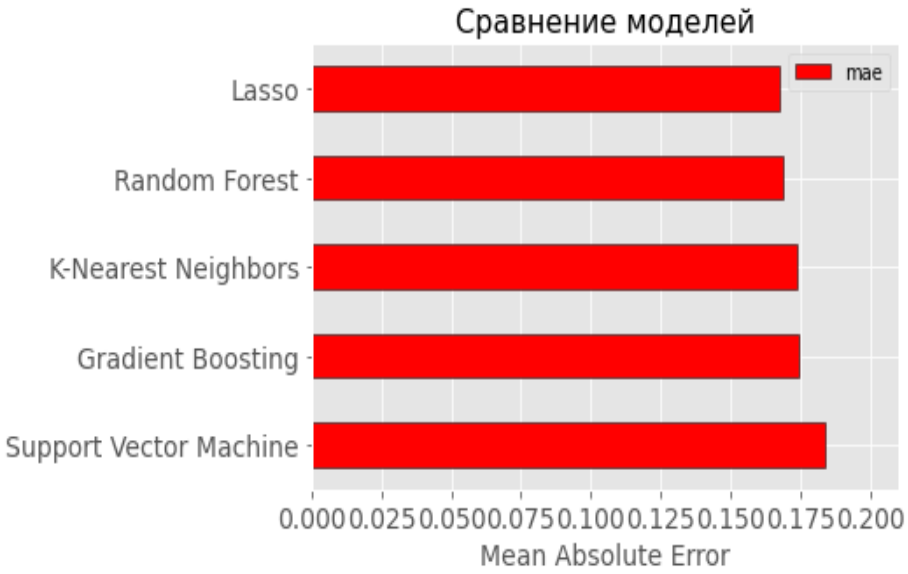
- Для свойства модуля упругости при растяжении разработаны и обучены следующие три модели:
- - модель на основе градиентного бустинга (GradientBoostingRegressor());
- - случайный лес (RandomForestRegressor());

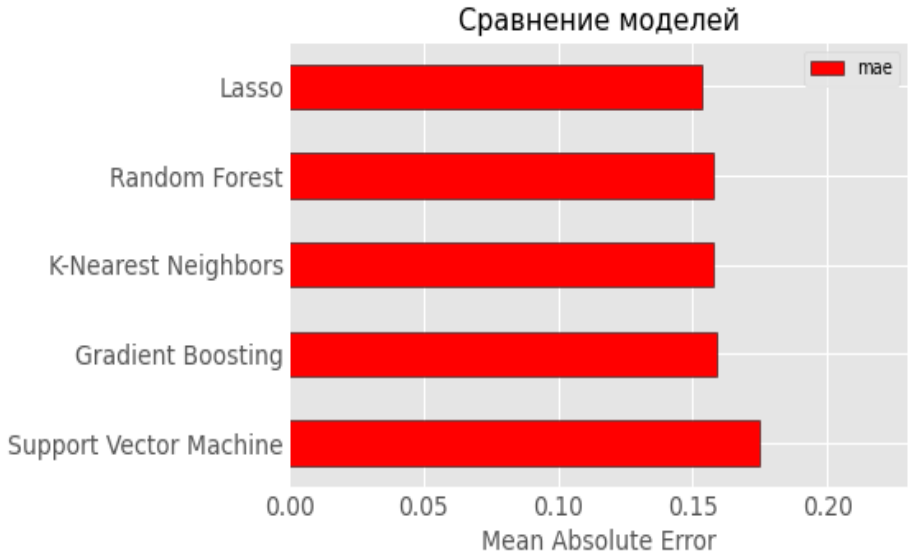
- Для свойства прочности при растяжении разработаны и обучены следующие четыре модели:
- - линейная модель Лассо (Lasso());
- - модель k-ближайших соседей (KNeighborsRegressor());
- - модель на основе метода опорных векторов (SVR());

- Настройки гиперпараметров производились случайным поиском с 10-блочной перекрёстной проверкой.
- Задаём сетку гиперпараметров.
- Случайно выбираем комбинацию гиперпараметров (RandomizedSearchCV из Scikit-Learn)
- Создаём модель с использованием этой комбинации.
- Оцениваем результат работы модели с помощью k-блочной перекрёстной проверки (GridSearchCV из Scikit-Learn).
- Решаем, какие гиперпараметры дают лучший результат.

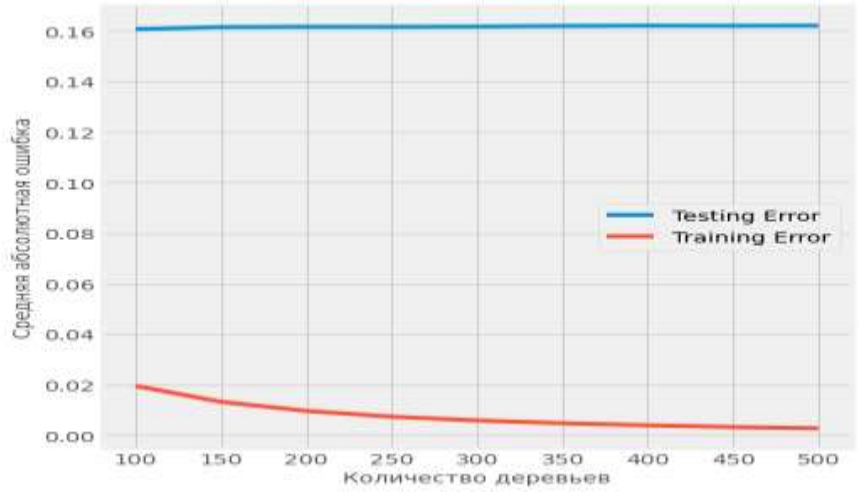
Тестирование моделей

Сравнение оценок моделей со стандартными параметрами для прогнозирования свойств модуля упругости при растяжении (слева) и прочности при растяжении (справа)



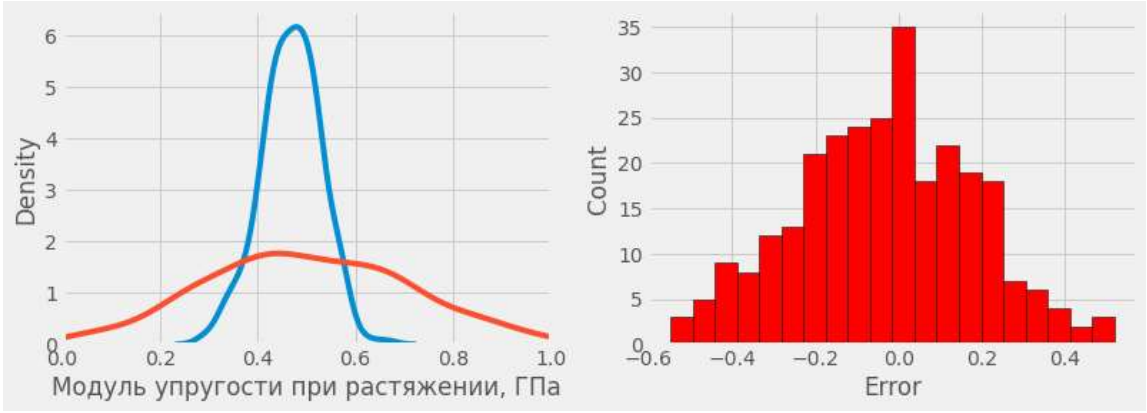


Градиентный бустинг для прогнозирования свойства модуля упругости при растяжении



	Тестовые данные	Спрогнозированные данные
319	0.387108	0.395278
377	0.742719	0.487172
538	0.364855	0.448914
296	0.435807	0.456200
531	0.468100	0.426032
...	...	...
420	0.359362	0.475689
133	0.446885	0.369484
490	0.609638	0.440140
558	0.372351	0.439003
363	0.392277	0.291119

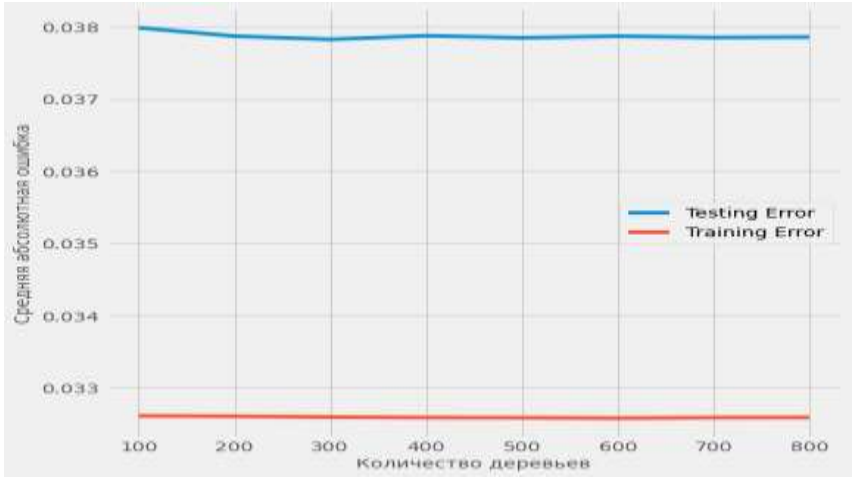
277 rows × 2 columns

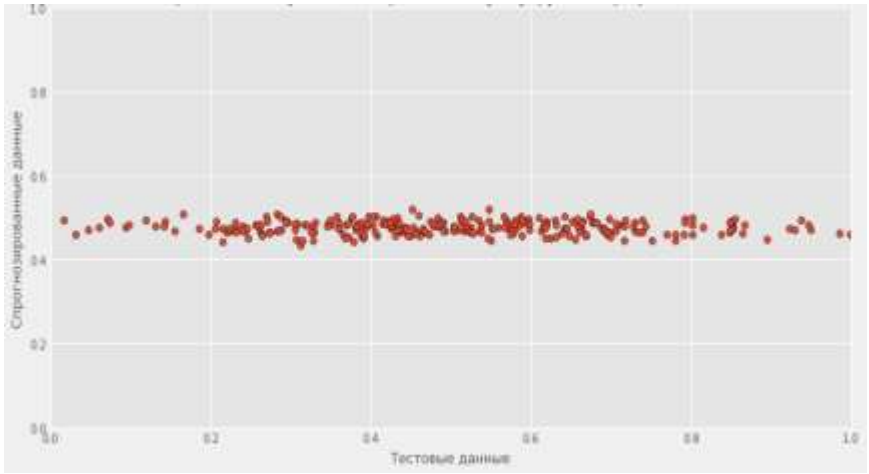


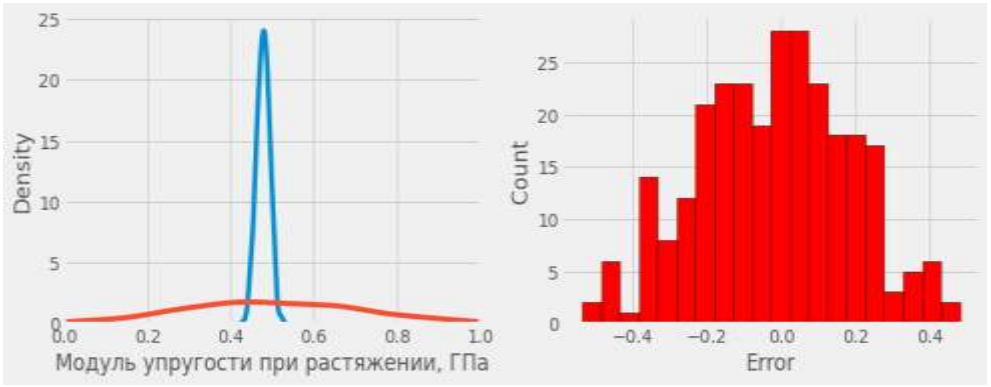
Средняя абсолютная ошибка модели со стандартными параметрами на тестовом наборе: MAE = 0.1747.

Средняя абсолютная ошибка настроенной модели на тестовом наборе: MAE = 0.1728.

Случайный лес для прогнозирования свойства модуля упругости при растяжении







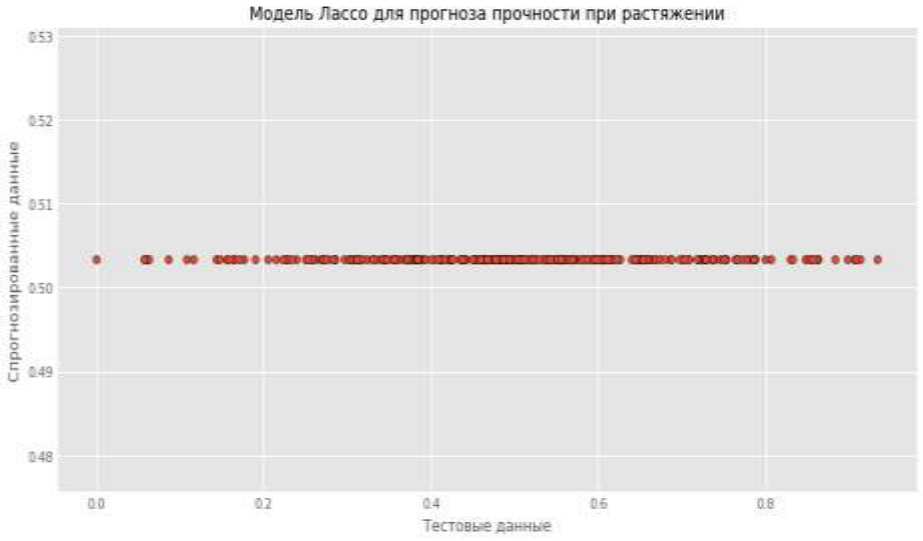
	Тестовые данные	Спрогнозированные данные
319	0.387108	0.459092
377	0.742719	0.480743
538	0.364855	0.461990
296	0.435807	0.465367
531	0.468100	0.467930
...	...	...
420	0.359362	0.473950
133	0.446885	0.459269
490	0.609638	0.501269
558	0.372351	0.488084
363	0.392277	0.459069

277 rows × 2 columns

Средняя абсолютная ошибка модели со стандартными параметрами на тестовом наборе: MAE = 0.1677.

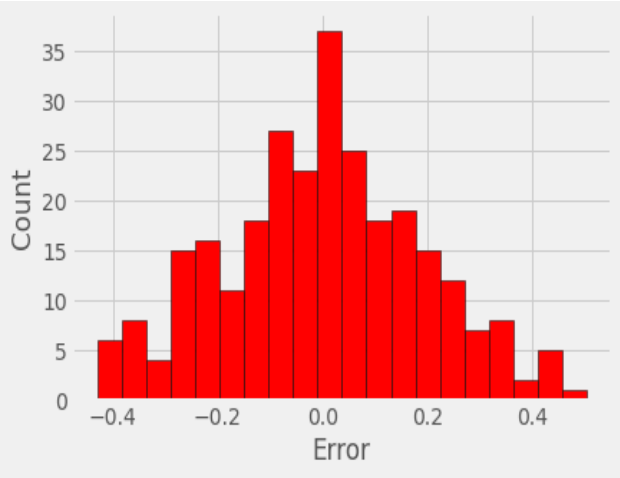
Средняя абсолютная ошибка настроенной модели на тестовом наборе: MAE = 0.1677.

Линейная модель Лассо для прогнозирования свойства прочности при растяжении



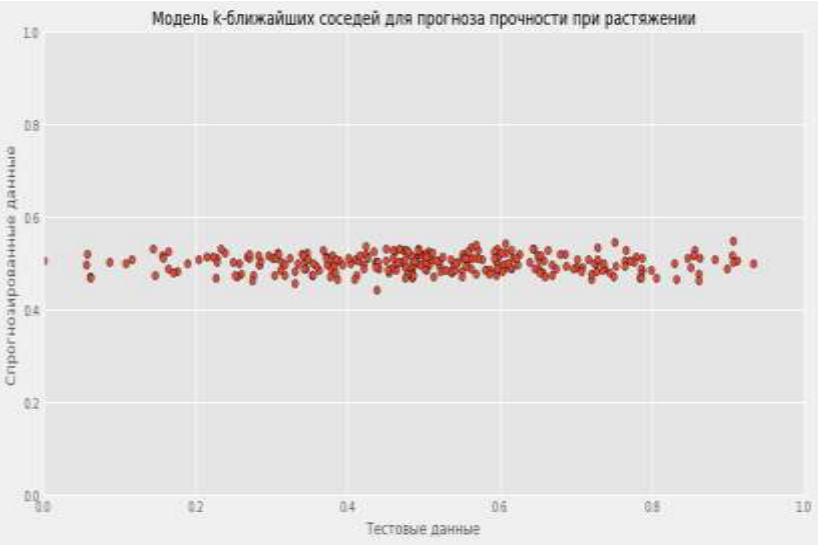
	Тестовые данные	Спрогнозированные данные
319	0.381499	0.503339
377	0.605408	0.503339
538	0.708160	0.503339
296	0.438781	0.503339
531	0.061865	0.503339
...	...	...
420	0.854766	0.503339
133	0.347529	0.503339
490	0.503825	0.503339
558	0.568094	0.503339
363	0.510845	0.503339

277 rows × 2 columns



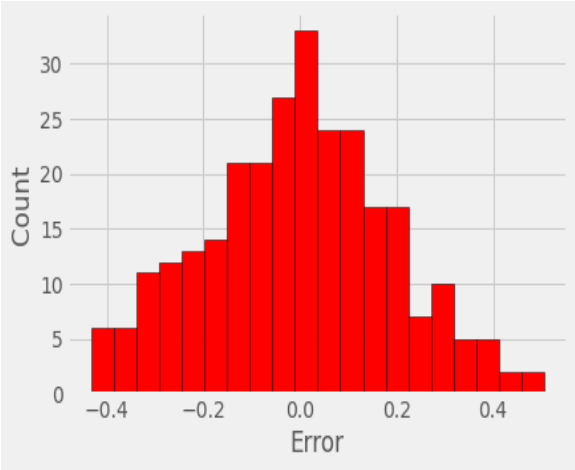
Средняя абсолютная ошибка модели со стандартными параметрами на тестовом наборе: MAE = 0.1534.
Средняя абсолютная ошибка настроенной модели на тестовом наборе: MAE = 0.1534.

Метод k-ближайших соседей для прогнозирования свойства прочности при растяжении



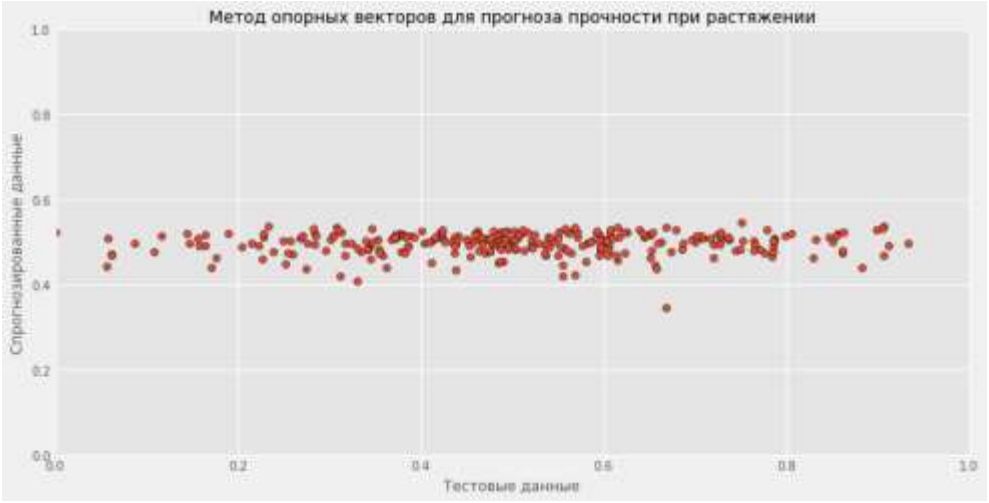
	Тестовые данные	Спрогнозированные данные
319	0.381499	0.497051
377	0.605408	0.495601
538	0.708160	0.492720
296	0.438781	0.443859
531	0.061865	0.472416
...	...	...
420	0.854766	0.516820
133	0.347529	0.509669
490	0.503825	0.515138
558	0.568094	0.540368
363	0.510845	0.524114

277 rows × 2 columns



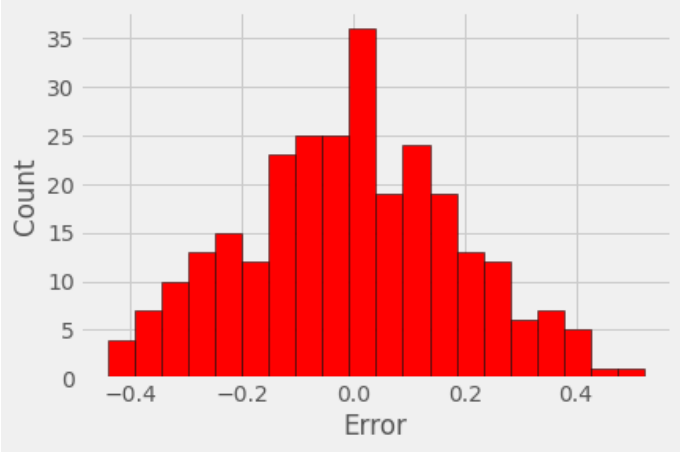
Средняя абсолютная ошибка модели со стандартными параметрами на тестовом наборе: MAE = 0.1666.
Средняя абсолютная ошибка настроенной модели на тестовом наборе: MAE = 0.1541.

Метод опорных векторов для прогнозирования свойства прочности при растяжении



	Тестовые данные	Спрогнозированные данные
319	0.381499	0.518916
377	0.605408	0.493429
538	0.708160	0.509172
296	0.438781	0.436584
531	0.061865	0.471816
...	...	...
420	0.854766	0.515011
133	0.347529	0.499941
490	0.503825	0.486532
558	0.568094	0.534347
363	0.510845	0.528162

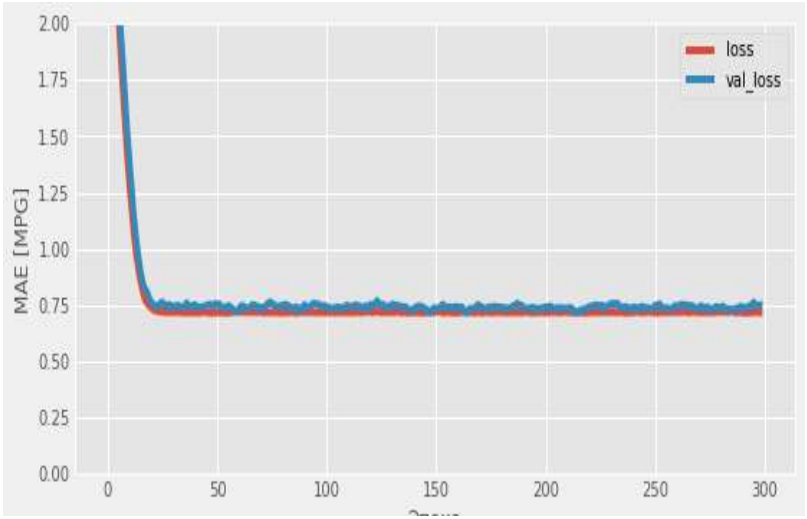
277 rows × 2 columns

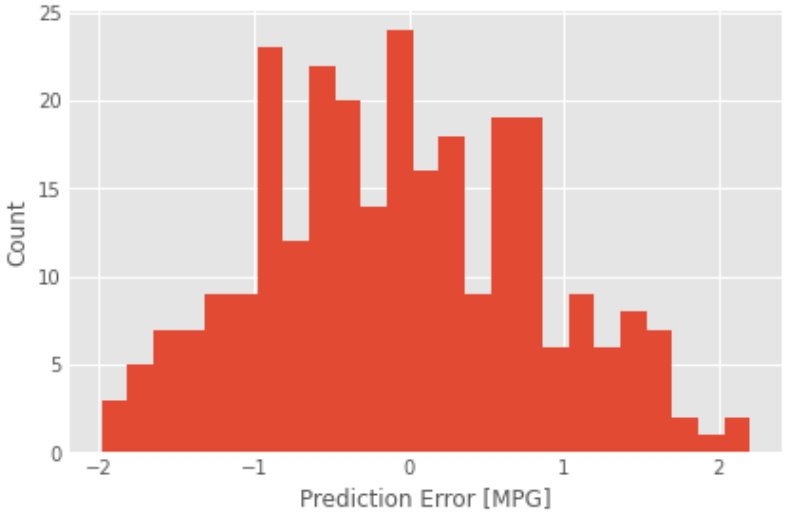


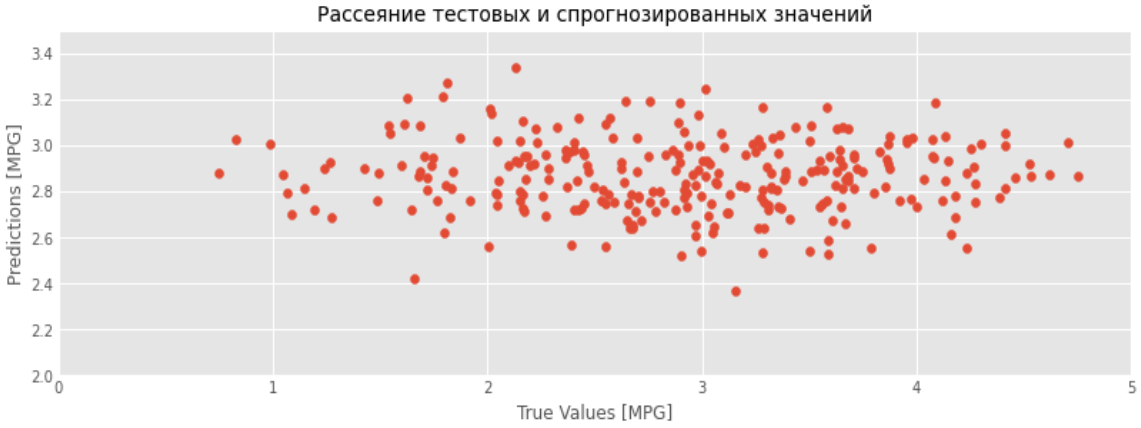
Средняя абсолютная ошибка модели со стандартными параметрами на тестовом наборе: MAE = 0.1663.
Средняя абсолютная ошибка настроенной модели на тестовом наборе: MAE = 0.1541

Нейронные сети

Линейная модель







	Тест	Прогноз
360	2.097010	2.911542
427	4.073344	2.952605
611	1.647649	2.720194
336	3.391009	2.889096
604	3.578784	3.164991
...	...	...
479	2.424798	2.719007
149	1.916843	2.760573
561	3.011978	3.248160
632	2.622594	2.901837
411	1.273273	2.689256

277 rows × 2 columns

Средняя абсолютная ошибка: 0.7294415831565857

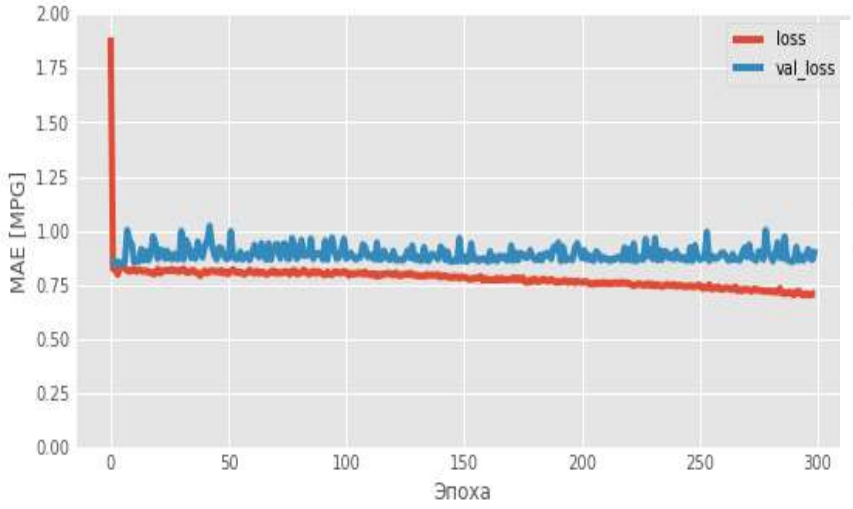
Нейронные сети

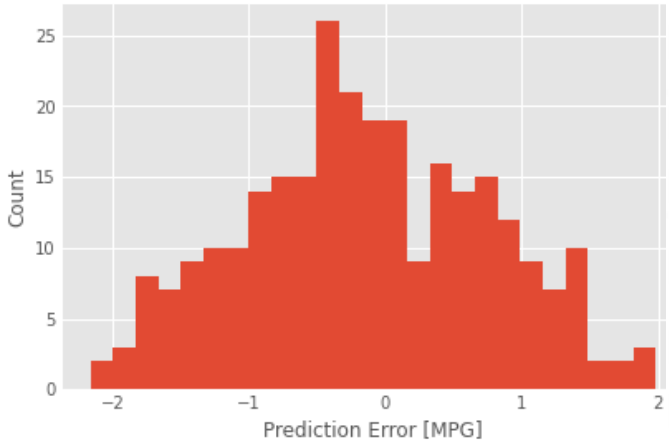
Многослойный персептрон

```
def build_and_compile_model0(norm):
    model0 = keras.Sequential([
      norm,
      keras.layers.Dense(128, activation='sigmoid'),
      keras.layers.Dense(64, activation='sigmoid'),
      keras.layers.Dense(1)
    ])

    model0.compile(loss='mean_squared_error',
                optimizer=tf.keras.optimizers.RMSprop(0.001))
    return model0

dnn_model0 = build_and_compile_model0(normalizer)
dnn_model0.summary()
```





	Тест	Прогноз
360	2.097010	2.698929
427	4.073344	2.787695
611	1.647649	2.989741
336	3.391009	2.654714
604	3.578784	2.083353
...	...	...
479	2.424798	2.508657
149	1.916843	2.745739
561	3.011978	2.896571
632	2.622594	2.435316
411	1.273273	2.652678

277 rows × 2 columns

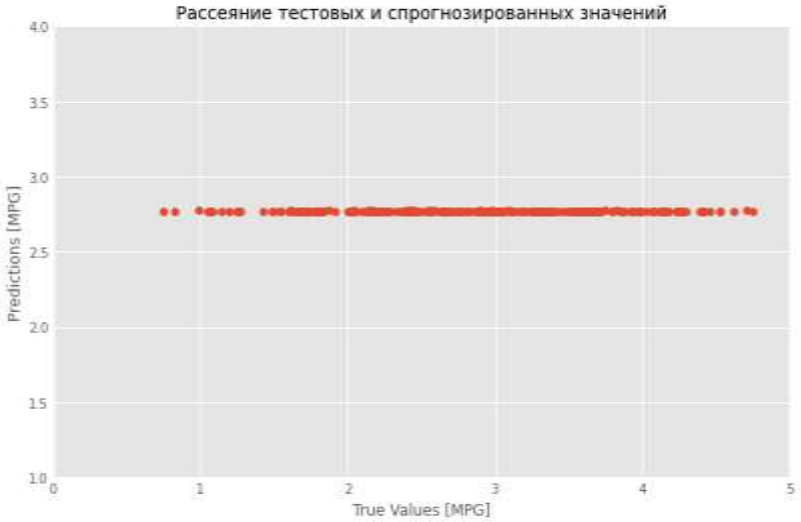
Средняя абсолютная ошибка: 0.8168711066246033

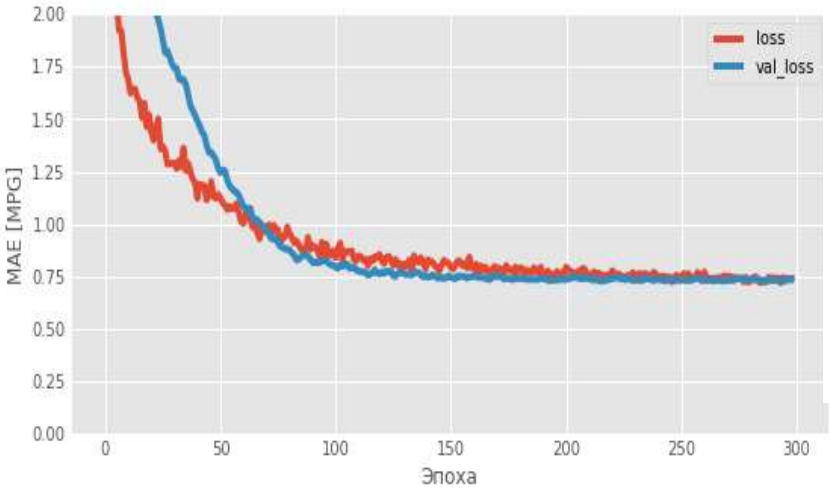
Нейронные сети

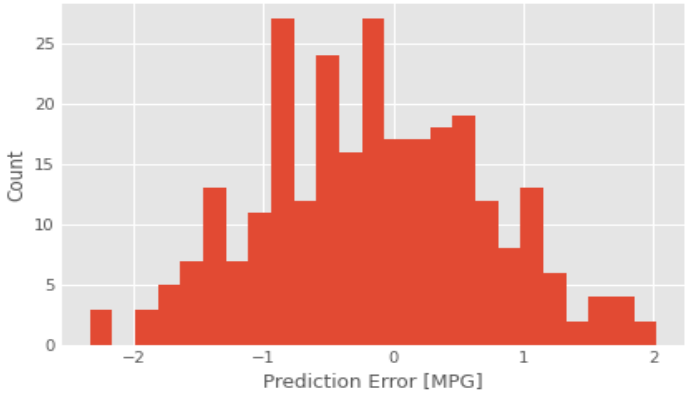
Многослойный персептрон

```
def build_and_compile_model1(norm):
    model1 = keras.Sequential([
      norm,
      keras.layers.Dense(256, activation='relu'),
      keras.layers.Dropout(0.8),
      keras.layers.Dense(192, activation='relu'),
      keras.layers.Dropout(0.8),
      keras.layers.Dense(128, activation='relu'),
      keras.layers.Dropout(0.8),
      keras.layers.Dense(64, activation='relu'),
      keras.layers.Dropout(0.8),
      keras.layers.Dense(1)
    ])

    model1.compile(loss='mean_absolute_error',
                   optimizer=tf.keras.optimizers.Adam(learning_rate=0.001))
    return model1
```







	Тест	Прогноз
360	2.097010	2.771945
427	4.073344	2.771527
611	1.647649	2.772239
336	3.391009	2.771825
604	3.578784	2.772597
...	...	...
479	2.424798	2.771859
149	1.916843	2.772062
561	3.011978	2.771788
632	2.622594	2.772402
411	1.273273	2.771854

277 rows × 2 columns

Средняя абсолютная ошибка: 0.7274526953697205

Результаты

- Несмотря на большую проделанную работу, результаты которой лишь частично представлены в отчете, построенные и обученные модели не решают поставленных задач прогнозирования модуля упругости при растяжении и прочности при растяжении композиционных материалов. Все модели не удовлетворительно описывают исходные данные. Построенные и обученные нейронные сети также не справились с задачей рекомендации соотношения матрица-наполнитель.

- Приложение для прогнозирования конечных свойств композиционных материалов не разработано.

Спасибо за внимание!