ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ МГТУ им. Н.Э. Баумана

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э.
Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу
«Data Science Pro»

«Прогнозирование конечных свойств новых материалов (композиционных материалов).»

Слушатель Батраков Данил Романович (ФИО)

Москва, 2024

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Аннотация

В данной работе представлен анализ данных производства композитных материалов, а также представлено построение регрессионных моделей на предоставленных данных с их анализом и сравнением.

Abstract

This paper presents an analysis of data from the production of composite materials, and also presents the construction of regression models on the provided data with their analysis and comparison.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Оглавление

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Введение

Композитные материалы – материалы, созданные из двух или более составляющих материалов со значительно различными физическими и химическими свойствами, которые в комбинации придают итоговому материалу характеристики отличные от характеристик отдельных компонентов и не являющиеся простой их суперпозицией.

Ключевые компоненты композитных материалов:

Матрицу материала – это основной материал, окружающий и связывающий вместе армирующие материалы. Матрица обеспечивает защиту и поддержку армирующего материала. Часто, в качестве матрицы используются полимеры.

Армирующий материал – материалы, внедряющиеся в матрицу материалов для улучшения определенных свойств, таких как, например, прочность, жесткость или ударная вязкость. В качестве наполнителей композитов как правило выступают углеродные или стеклянные волокна.

Сочетание разных компонентов позволяет улучшить характеристики материала и делает его одновременно лёгким и прочным. При этом отдельные компоненты остаются таковыми в структуре композитов, что отличает их от смесей и затвердевших растворов. Варьируя состав матрицы и наполнителя, их соотношение, ориентацию наполнителя, получают широкий спектр материалов с требуемым набором свойств. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

1. Постановка задачи

1.1. Теоретическое обоснование исследуемых параметров

Согласно заданию необходимо написать алгоритм, который исходя из известных реальных данных будет определять значения:

- Модуль упругости при растяжении, ГПа;
- Прочность при растяжении, МПа.

Так же, согласно заданию, необходимо написать алгоритм, определяющий:

- Соотношение матрица-наполнитель;

Задача стоит в построении регрессионной модели машинного обучения.

В представленном датасете содержатся следующие данные:

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

Модуль упругости при растяжении (модуль Юнга), равен отношению напряжения к относительному удлинению, вызванному этим напряжением. Этот модуль характеризует способность материала упруго сопротивляться растяжению (после снятия нагрузки образец возвращается к первоначальному размеру).

$$E = \frac{s}{e};$$

Где,

E – упругость при растяжении;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

S – нормальное растяжение;

E – относительное удлинение.

Прочность при растяжении обозначает максимальное механическое растягивающее напряжение, с которым можно нагружать образец. При превышении прочности при растяжении материал разрушается: приложение усилия снижается, пока образец, наконец, не порвется.

$$\mathrm{Rm} = \frac{Fm}{S0};$$

Где,

Rm - прочность при растяжении;

Fm - максимальное усилие растяжения;

S0 - площадь поперечного сечения.

1.2. Разведочный анализ данных

1.2.1. Анализ выбросов

	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
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	2.930612	1975.666791	739.950520	110.541116	22.243850	285.911287	483.024665	73.328462	2467.184290	218.386876	45.000000	6.909904	57.181470
std	0.913939	73.796817	330.327013	28.304470	2.406999	40.962756	280.811736	3.119584	485.624616	59.819777	45.022517	2.560031	12.304355
min	0.389403	1731.764635	2.436909	17.740275	14.254985	160.255843	0.603740	64.054061	1036.856605	33.803026	0.000000	0.037639	11.740126
25%	2.318526	1924.203433	498.438068	92.170589	20.558296	258.539199	268.057473	71.301753	2143.834592	179.190489	0.000000	5.116518	49.821889
50%	2.907832	1977.574305	741.148111	110.162666	22.230761	285.853960	452.972263	73.247594	2461.249253	217.277006	45.000000	6.913444	57.362576
75%	3.552539	2021.159498	962.851423	130.311975	23.982115	313.581449	694.210382	75.379739	2760.163022	257.495647	90.000000	8.585130	64.986942
max	5.591742	2207.773481	1911.536477	198.953207	28.955094	413.273418	1399.542362	82.682051	3848.436732	414.590628	90.000000	14.440522	103.988901

Рисунок 1 - Описание датасета

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Из описания датасета, представленного на рисунке 1, можно увидеть минимальные, средние, медианные, максимальные значения, а также квантили распределения данных 25%, 50%, 75%.

```
<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)
memory usage: 111.9 KB
```

Рисунок 2 - Общая информация о предоставленных данных

На рисунке 2 мы видим описание столбцов датасета, в частности тип данных, хранимых в столбцах, имеются ли пропуски в столбцах.

	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
279	4.434429	2061.918771	328.876626	81.071393	23.213218	244.874100	934.780246	76.238258	2995.952606	178.066150	0	2.788476	11.740126
548	2.505018	2043.225729	947.563433	107.172872	21.141384	334.538100	665.277923	79.892087	2461.043398	164.365856	90	9.379064	49.540731
287	3.183372	1953.008677	1409.793826	63.548428	24.861055	225.514802	633.151898	73.551929	2324.703278	235.951687	0	11.361013	50.031691
640	3.603262	1998.709005	760.881311	125.495521	25.906781	263.805111	378.020970	73.595083	3249.856261	171.711493	90	4.945255	51.945279
791	3.426368	2028.026074	453.458891	105.674852	22.415611	413.273418	586.715020	74.838137	2334.649515	233.293376	90	5.256326	46.966045

Рисунок 3 - Случайные 5 записей из датасета

На рисунке 3 отображены случайно взятые 5 записей.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

	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	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.000000	3000.000000	220.0	0	4.0	57.0
1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.000000	3000.000000	220.0	0	4.0	60.0
2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.000000	3000.000000	220.0	0	4.0	70.0
3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.000000	3000.000000	220.0	0	5.0	47.0
4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.000000	3000.000000	220.0	0	5.0	57.0
5	2.767918	2000.0	748.000000	111.86	22.267857	284.615385	210.0	70.000000	3000.000000	220.0	0	5.0	60.0
6	2.569620	1910.0	807.000000	111.86	22.267857	284.615385	210.0	70.000000	3000.000000	220.0	0	5.0	70.0
7	2.561475	1900.0	535.000000	111.86	22.267857	284.615385	380.0	75.000000	1800.000000	120.0	0	7.0	47.0
8	3.557018	1930.0	889.000000	129.00	21.250000	300.000000	380.0	75.000000	1800.000000	120.0	0	7.0	57.0
9	3.532338	2100.0	1421.000000	129.00	21.250000	300.000000	1010.0	78.000000	2000.000000	300.0	0	7.0	60.0
10	2.919678	2160.0	933.000000	129.00	21.250000	300.000000	1010.0	78.000000	2000.000000	300.0	0	7.0	70.0
11	2.877358	1990.0	1628.000000	129.00	21.250000	300.000000	1010.0	78.000000	2000.000000	300.0	0	9.0	47.0
12	1.598174	1950.0	827.000000	129.00	21.250000	300.000000	470.0	73.333333	2455.555556	220.0	0	9.0	57.0
13	2.919678	1980.0	568.000000	129.00	21.250000	300.000000	470.0	73.333333	2455.555556	220.0	0	9.0	60.0
14	4.029126	1910.0	800.000000	129.00	21.250000	300.000000	470.0	73.333333	2455.555556	220.0	0	9.0	70.0
15	2.934783	2030.0	302.000000	129.00	21.250000	300.000000	210.0	70.000000	3000.000000	220.0	0	10.0	47.0
16	3.557018	1880.0	313.000000	129.00	21.250000	300.000000	210.0	70.000000	3000.000000	220.0	0	10.0	57.0
17	4.193548	1950.0	506.000000	129.00	21.250000	300.000000	380.0	75.000000	1800.000000	120.0	0	10.0	60.0
18	4.897959	1890.0	540.000000	129.00	21.250000	300.000000	380.0	75.000000	1800.000000	120.0	0	10.0	70.0
19	3.532338	1980.0	1183.000000	111.86	22.267857	284.615385	1010.0	78.000000	2000.000000	300.0	0	0.0	0.0
20	2.877358	2000.0	205.000000	111.86	22.267857	284.615385	1010.0	78.000000	2000.000000	300.0	90	4.0	47.0
21	1.598174	1920.0	456.000000	111.86	22.267857	284.615385	470.0	73.333333	2455.555556	220.0	90	4.0	57.0
22	4.029126	1880.0	622.000000	111.86	22.267857	284.615385	470.0	73.333333	2455.555556	220.0	90	4.0	60.0

Рисунок 4 - Результат отбора целочисленных значений

При визуальном анализе данного датасета было вявлено что значения преимущественно носят числовой, преимущественно дробный харакетр (см. рисунок 2), однако на части данных были замечены цельночисловые значения (см. рисунок 4), так как данные числа явно выбиваются из остального датасета, по причине явного округления, они могут повлиять на ход обучения регрессионных моделей, в связи с чем было принято решение об их удалении. Явно округленные значения находятся только в первых 22 записях и наблюдаются на столбцах:

- Плотность, кг/м3;
- модуль упругости, Гпа;
- Количество отвердителя, м.%;
- Содержание эпоксидных групп,%_2;
- Температура вспышки, С_2;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

- Поверхностная плотность, г/м2;
- Модуль упругости при растяжении, Гпа;
- Прочность при растяжении, Мпа;
- Потребление смолы, г/м2;

Можно заметить что вышеописанные данные лежали в первой части датасета, а также в этих данных имели место быть и другие «подозрительные» последовательности, такие как, например, на рисунке 5, значение с плотностью нашивки 0.

	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
19	3.532338	1980.000000	1183.000000	111.860000	22.267857	284.615385	1010.000000	78.000000	2000.000000	300.000000	0	0.000000	0.000000
117	4.136691	1944.518365	589.646853	102.349135	19.580547	247.744983	371.247656	80.075363	2588.981282	346.276191	0	7.190538	92.042139
140	0.790499	2054.123810	876.794802	116.889391	23.564969	353.561943	643.647660	70.080210	2424.148613	247.193190	0	5.506840	98.202603
193	2.348185	1929.677180	422.524663	104.737519	23.645353	305.799316	554.507496	76.436218	3238.218206	273.174370	0	9.992951	92.963492
464	2.440166	1980.132394	357.938256	130.541848	21.177094	290.618547	119.756192	68.900703	1577.288189	129.038238	0	4.447498	103.988901

Рисунок 5 - Выборка данных, где плотность нашивки лежит за диапазонами [0, 90]

Был проанализирован параметр «Угол нашивки, град», где значения являются либо 0 либо 90, на предмет соотношения количества двух значений, полученный резльтат равен 1.0337972166998013 , что близко к единице, следовательно количество записей примерно совпадает друг с другом.

1.2.2. Анализ корреляционных составляющих

При вычислении корреляции между параметрами явной зависимости между параметрами не было выявлено (см. рисунок). Выявлены единичные случаи корреляции около 0.1

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

	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

Рисунок 6 - Вычисление корреляции между параметрами

Для наглядности данная таблица представлена в виде тепловой карты на рисунке 7.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

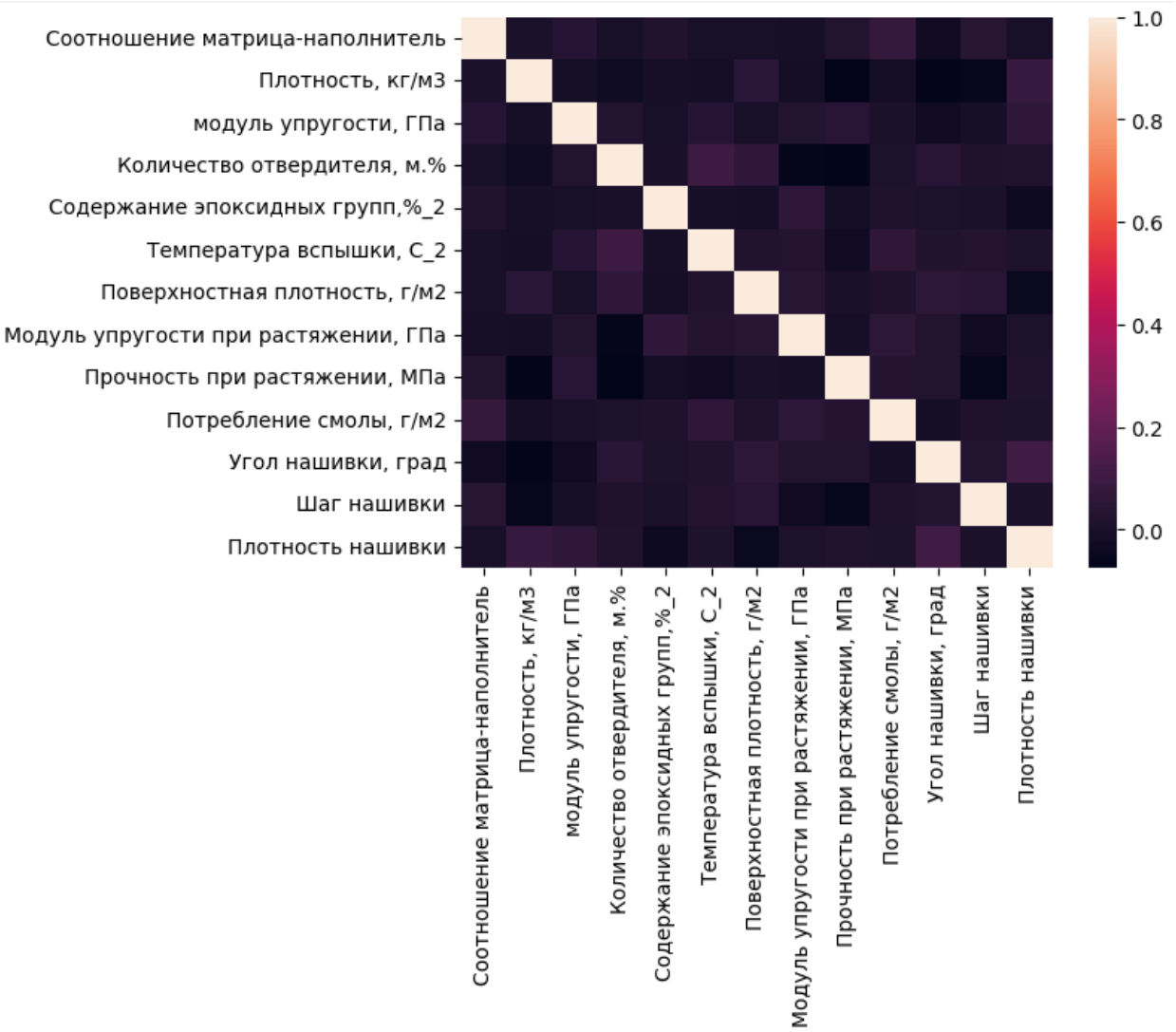

Рисунок 7 - "Тепловая карта" визуально представляющая полученные данные о корреляции

1.2.3. Анализ распределения данных

Распределение данных по каждому столбцу представлено на рисунке 8 ниже:

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

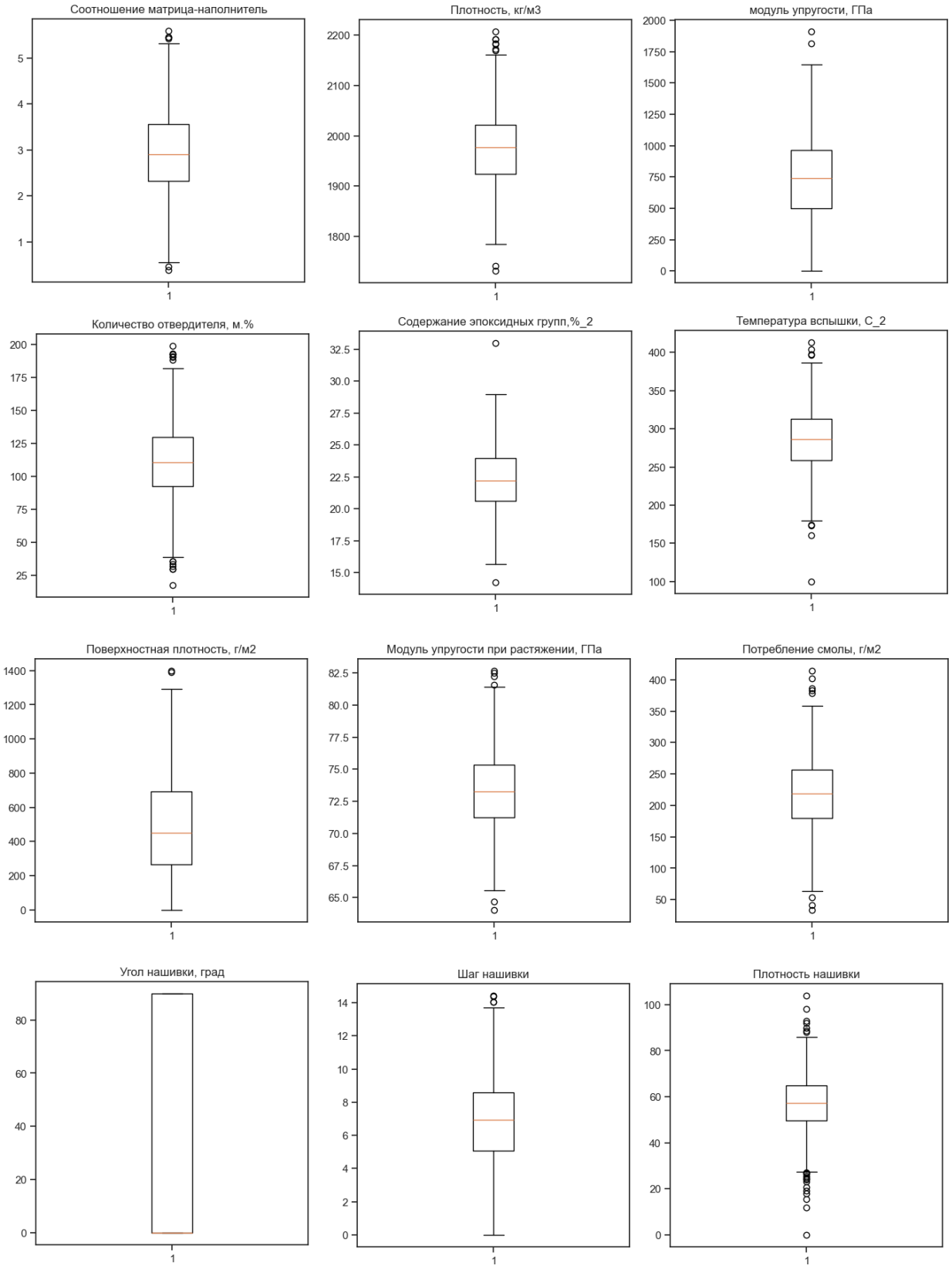

Рисунок 8 - Анализ распределения данных

Из анализа распределения данных видно, что они распредлены не равномерно, присутствуют явные многочисленные выходы за рамки

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

нормального распределения, однако, так как количество представленных экземпляров в датасете составляет около 1000 строк, удаление всех этих значений приводит к слишком маленькому количеству данных.

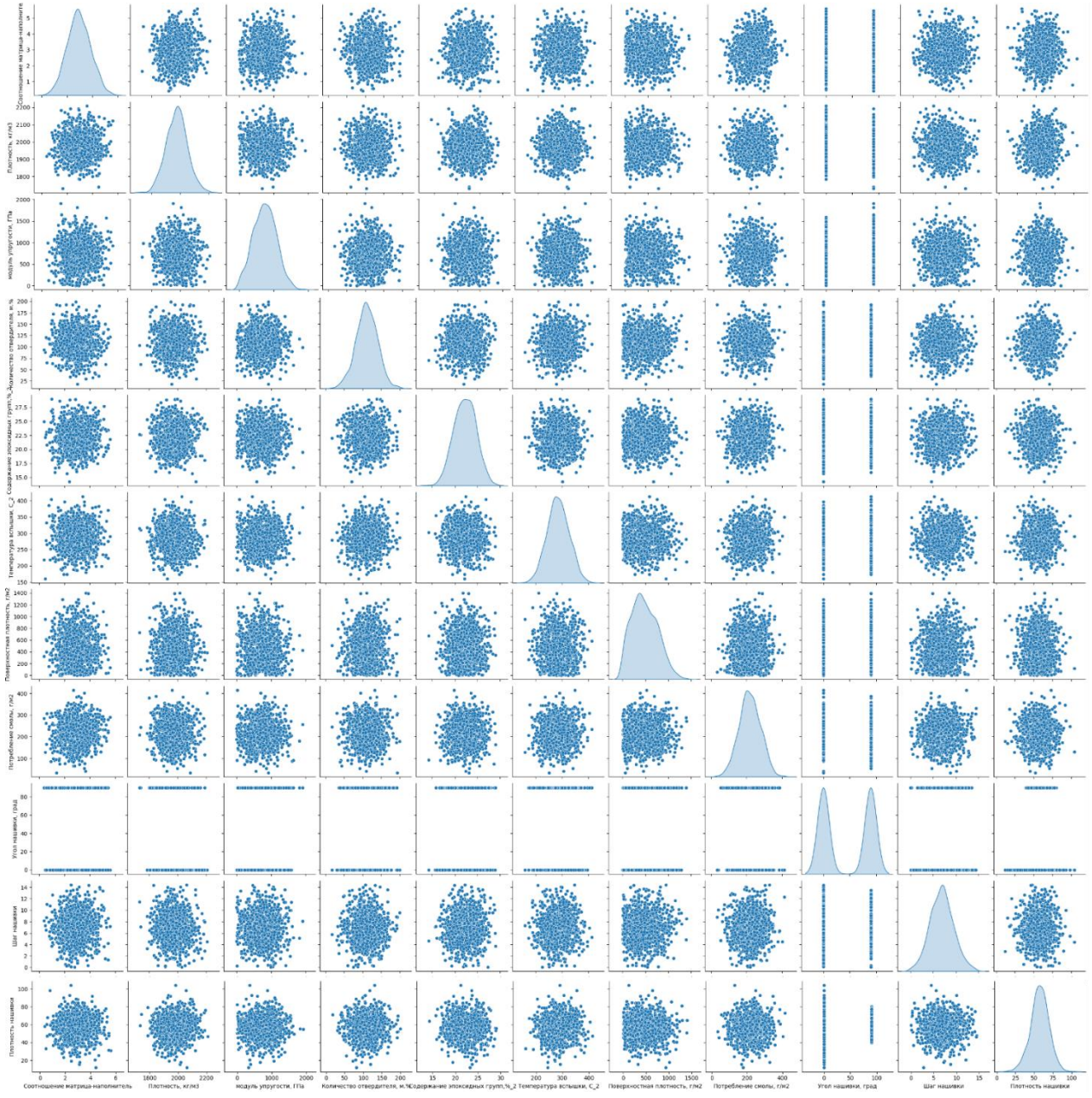

Рисунок 9 - Зависимость распределения параметров относительно друг друга

На рисунке 9 можно видеть зависимость распределения данных по параметрам относительно друг друга. Линейная зависимость не наблюдается.

Было решено оставить данные выпадающие значения в датасете и списать неравномерное распределение на «особенность» данных.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

1.2.4. Анализ данных и поиск новых значений с целью повысить корреляционную зависимость

При анализе данных были выдвинуты предположения о том, что данные были предварительно прологарифмированы ($\log_{10} x$, $\log x$, $\ln x$) или, наоборот, возведены в некоторую степень, для удобства подсчета (2, e, 1/2), однако ни одно из предположений не повысило, сколько-нибудь значимо, корреляционные зависимости.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

2. Описание используемых методов

В данной работе будут используются методы кластеризации, визуализации, машинного обучения, нейронных сетей.

Основная идея разработки регрессионной модели, после проведения анализа данных (см. пункт 1.2), строится на идее разбиения датасета на составляющие его кластеры. В качестве алгоритма кластеризации был выбран алгоритм K-means, в паре с алгоритмом приведения n-мерного пространства к 2-мерному виду T-SNE, оба эти алгоритма реализованы в библиотеке sklearn [1]. После разбиения данных на отдельные кластеры строятся регрессионные модели:

- LinearRegression - метод наименьших квадратов;
- RandomForestRegressor - случайный лес;
- KNeighborsRegressor - метод ближайших соседей;
- SVR - метод опорных векторов с линейным ядром;
- DecisionTreeRegressor - деревья решений.

Также, на основании разбитого на кластеры датасета строится алгоритм нейронной сети, методы необходимые для построения нейросетевой модели реализованы в библиотеке tensorflow [2] в модуле keras [3].

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

3. Построение моделей машинного обучения

3.1. Построение модели кластеризации данных

Для разбиения данных ан кластеры используется алгоритм K-Means (см. пункт 2). Для поиска оптимального количества кластеров для данного датасета сначала был проведен анализ межкластерного расстояния для различного колличества кластеров, именуемый «метод локтя». Межкластерное расстояние перестает значимо уменьшаться в приближении к плато графика. Код и результат данного метода представлен в приложенном к работе файлу «Composits_4» в разделе Разобьем данные на кластеры -> Подбор оптимального количества кластеров -> Метод локтя для подбора оптимального количества кластеров. Также результат работы метода представлен на рисунке ниже:

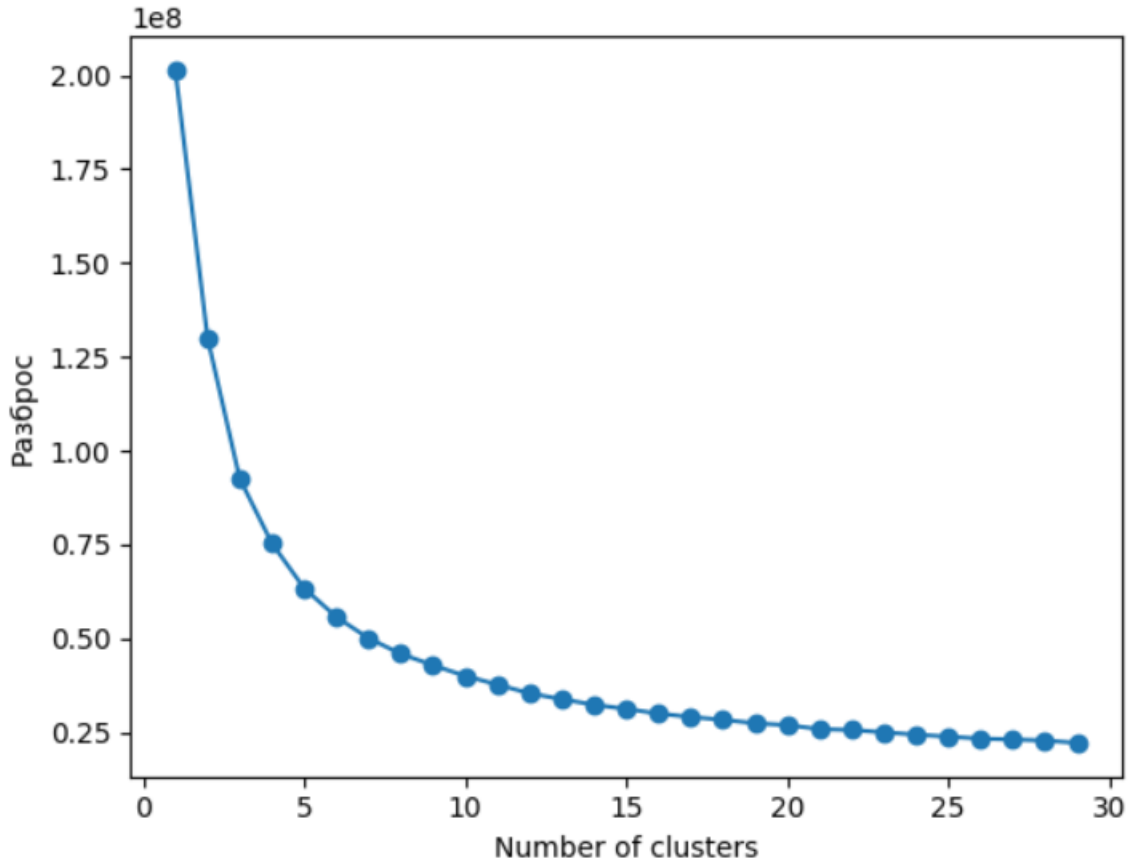

Рисунок 10 - Резльат поиска оптимального количества кластеров методом "локтя"

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

По результатам работы данного методы, можно заметить, что, разброс перестает значимо уменьшаться в районе 10-15 кластеров.

Как говорилось выше, для визуализации разбиения на кластеры использовался метод снижения размерности T-SNE (см. пункт 2).

Подробный анализ разбиения данных на кластеры представлен в приложенном к работе файлу «Composits_4» в разделе Разобьем данные на кластеры -> Подбор оптимального количества кластеров -> На основании вышеполученной информации, попроубем разбить данные на N класстеров, где N – число кластеров, лежит в диапазоне [8, 15].

По результатам, преимущественно, визуального анализа, было определено оптимальное количество в 9 кластеров. При данном числе кластеров выбросы из кластеров имеются, но их единицы.

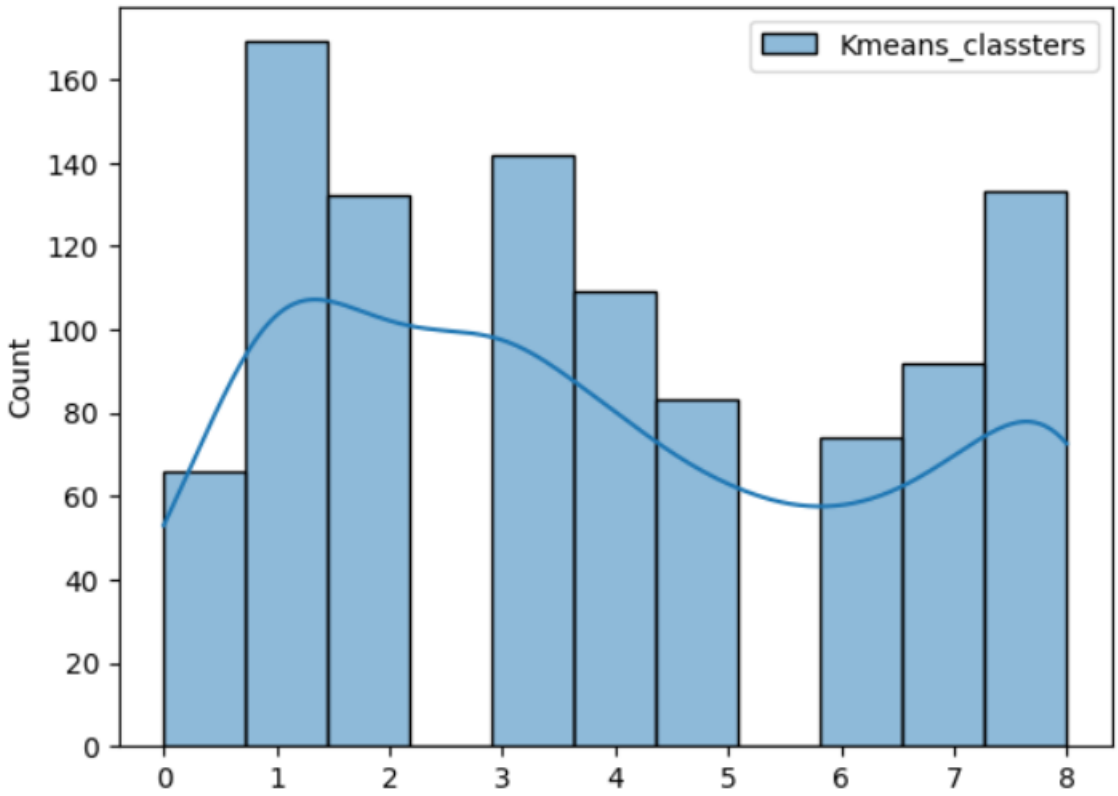

Рисунок 11 - График количества записей в каждом кластере

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

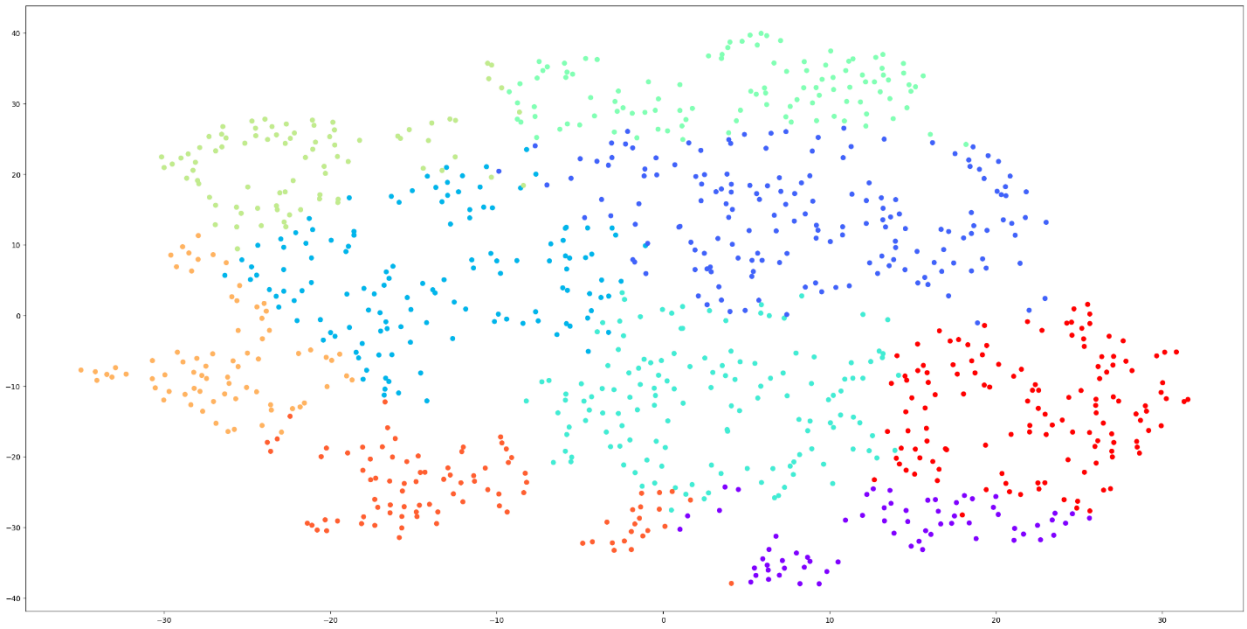

Рисунок 12 - Визуализация разбиения на кластеры представленная в двумерном пространстве

3.2. Построение модели классификации данных по заранее заданным кластерам

Для построения модели классификации, на основе полученных кластеров, была выбрана модель DecisionTreeClassifier, алгоритм которой описан в библиотеке sklearn в модуле tree [1]. Данный классификатор был выбран по причине возможности отображения по каким параметрам модель определяет класс, и можно было проанализировать и убрать неинформативные для модели поля.

Подробно модели представлены в приложенном к работе файлу «Composits_4» в разделе «Обучим классификатор "дерево решений" и посмотрим по каким параметрам определяется класс», для моделей машинного обучения, и в разделе «Обучим и сравним регрессионные модели» -> «Обучим классификатор для нейронной сети». Ввиду большого размера изображений отображающих работу классификаторов, они приложены к работе отдельно файлами «Dtree.png» и «Dtree_NN.png», для регрессионных моделей и нейронных сетей, соответственно.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Точность моделей превышает 0.9, большую точность на предложенных мне данных получить не удалось, в связи с малой выборкой.

3.3. Построение моделей машинного обучения

Для выбора модели с наибольшей предсказательной способностью были обучены все, представленные и описанные в пункте 2, модели.

Первым шагом было построение модели без учета кластеров, подробный код представлен в приложенном к работе файлу «Composits_4» в разделе «Обучим и сравним регрессионные модели» -> «Проведем анализ эффективности разных моделей на всех кластерах датасета» -> «Построение общих моделей на всем датасете (без учета кластеров)».

Все модели представили крайне слабую предсказательную способность. В качестве метрики был выбран коэффициент детерминации R2 [4]. Коэффициент детерминации - Статистический показатель, отражающий объясняющую способность регрессии f: X → Y и определяемый как доля дисперсии зависимой переменной, объяснённая регрессионной моделью с данным набором независимых переменных. Обычно определяется как единица минус доля необъяснённой дисперсии, т.е

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}, (1)$$

- SSE=∑i(yi−^yi)2 — сумма квадратов остатков (ошибок) регрессии (sum square of errors),
- SST=∑i(yi−¯¯yi)2 — полная сумма квадратов (sum square total), т.е. сумма квадратов отклонений точек данных от среднего значения,
- Xn=(xi,yi)ni=1 — набор данных из n наблюдений,
- yi∈Y, ¯¯yi=∑iyin,
- ^yi=f(xi).

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

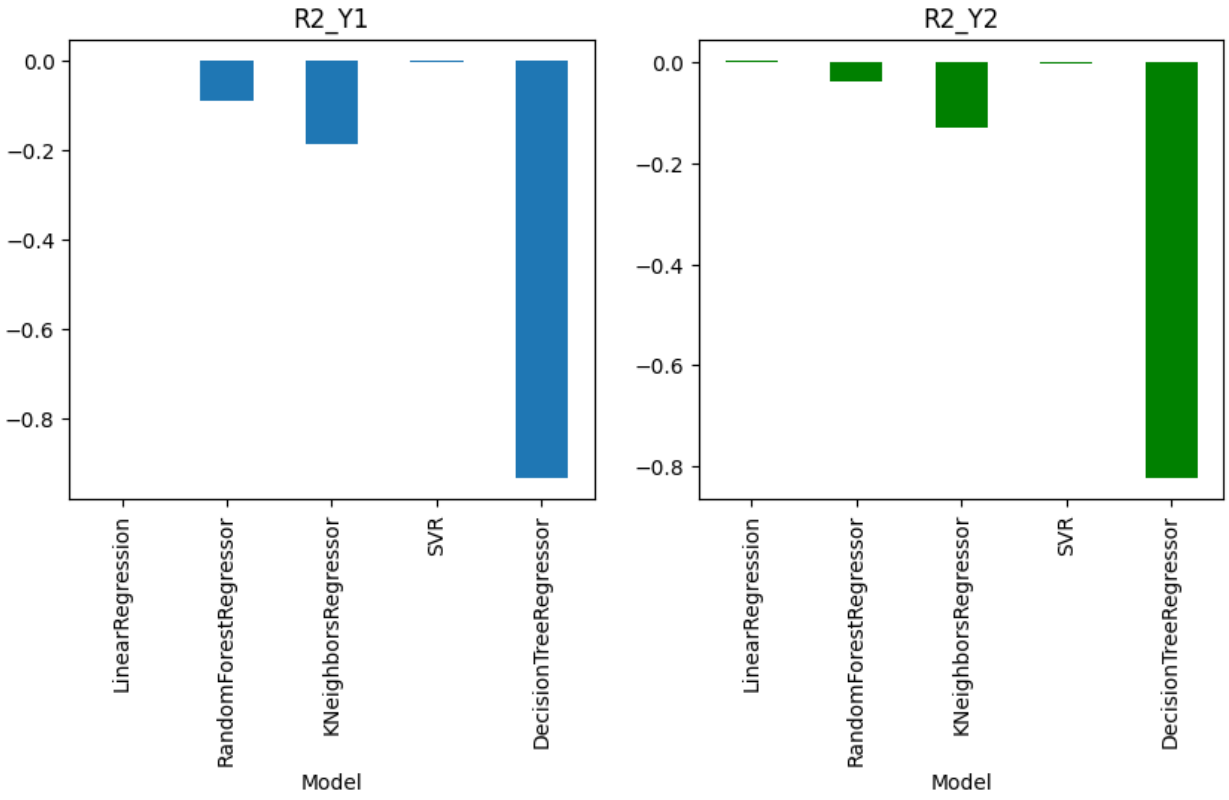

Рисунок 13 - Результаты работы регрессионных моделей

В качестве R2_Y1 было принято предсказание параметра «Модуль упругости при растяжении, Гпа».

В качестве R2_Y2 было принято предсказание параметра «Прочность при растяжении, МПа».

Построить модели с хоть сколько-нибудь значимой R2 метрикой не удалось, однако в процессе подстройки параметров был заметен небольшой, слабозначимый, прогресс. Результаты представлены ниже, на изображениях 14-23. Подробный код представлен в приложенном к работе файлу «Composits_4» в разделе «Обучим и сравним регрессионные модели» -> «Проведем анализ эффективности разных моделей на всех кластерах датасета» -> «Построение моделей с учетом кластеров».

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

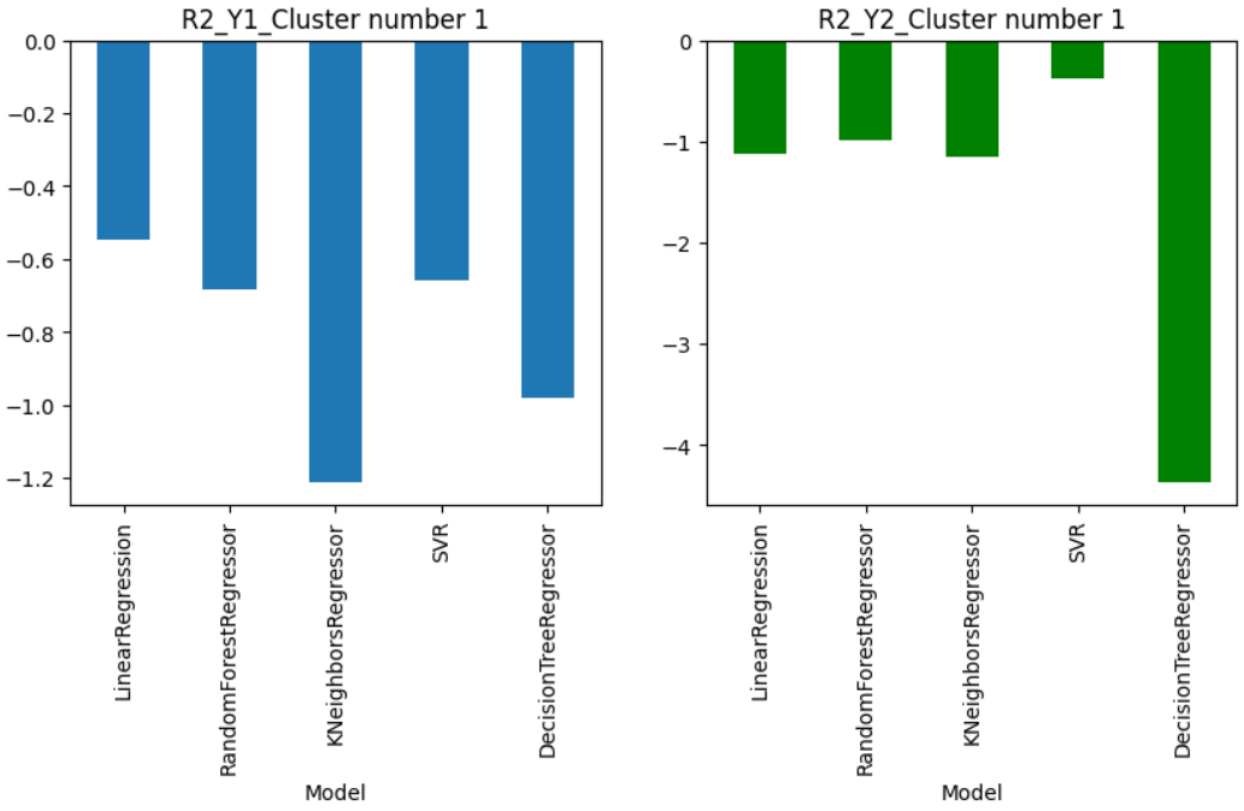

Рисунок 14

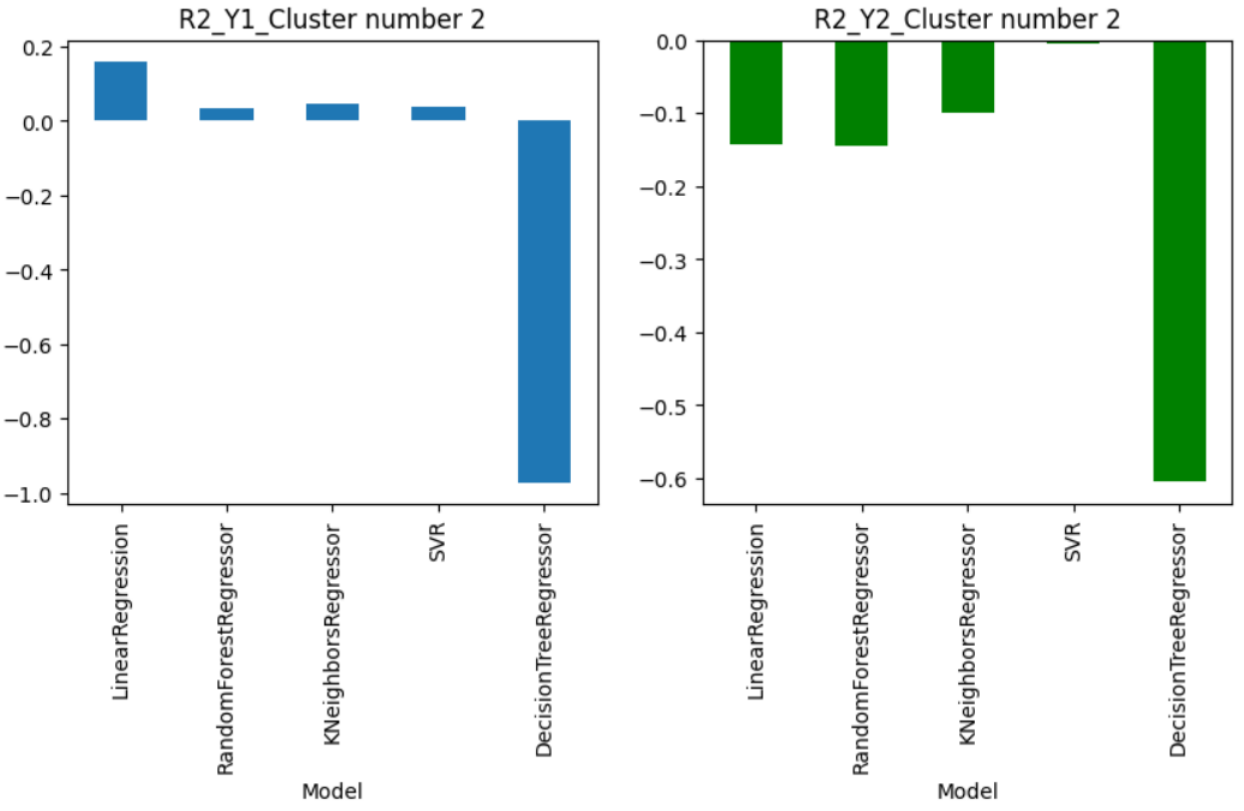

Рисунок 15

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

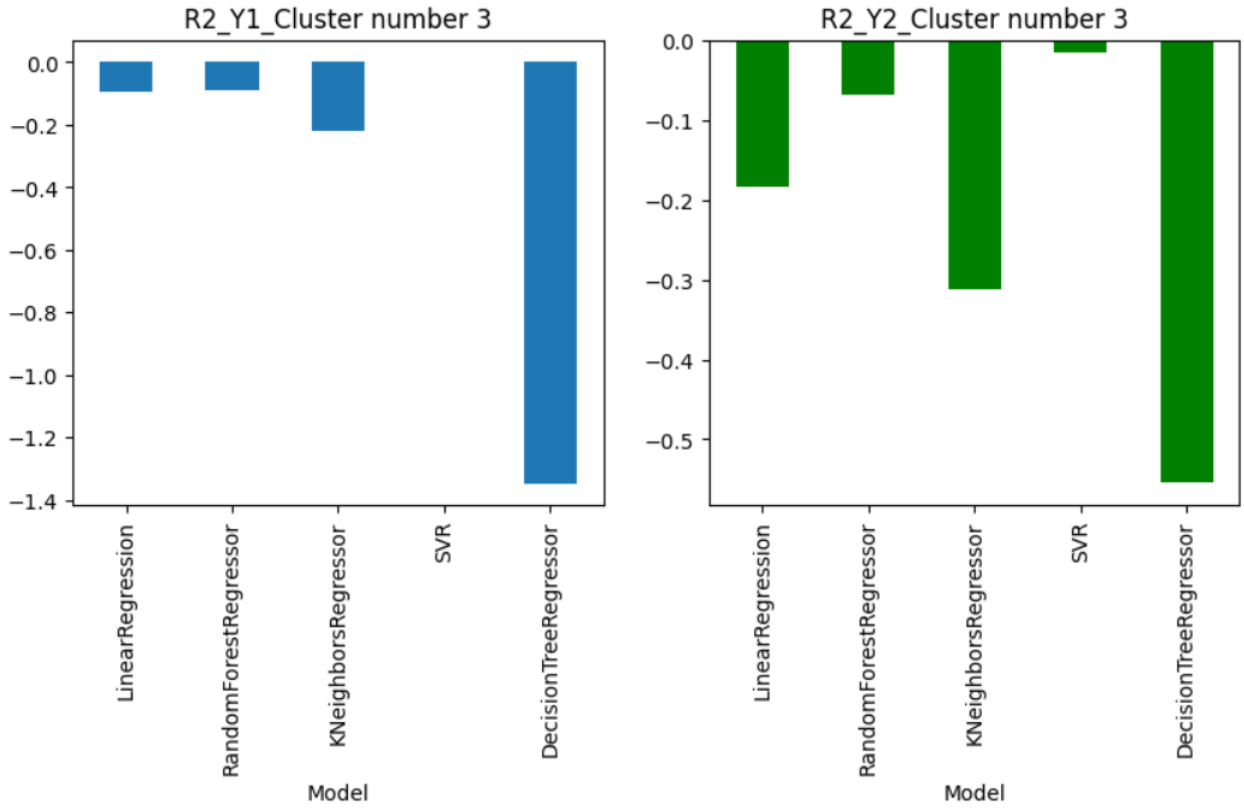

Рисунок 16

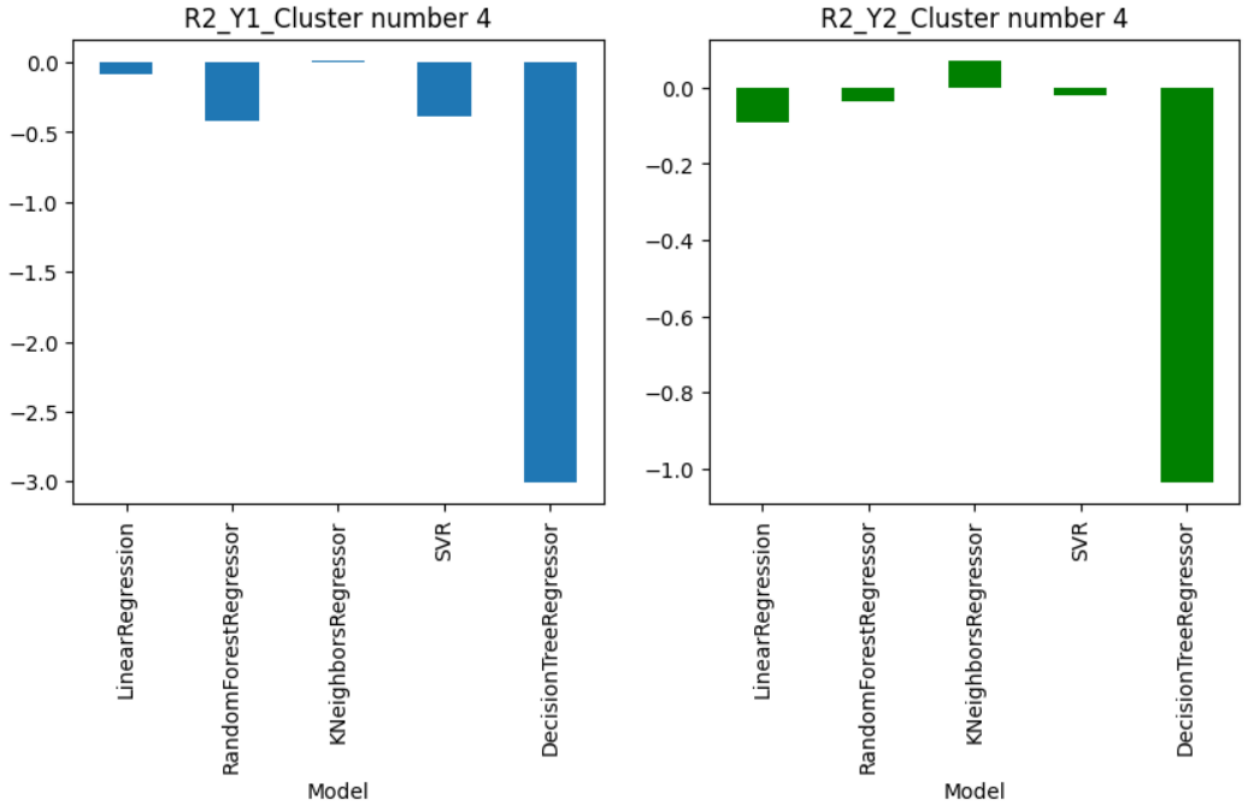

Рисунок 17

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

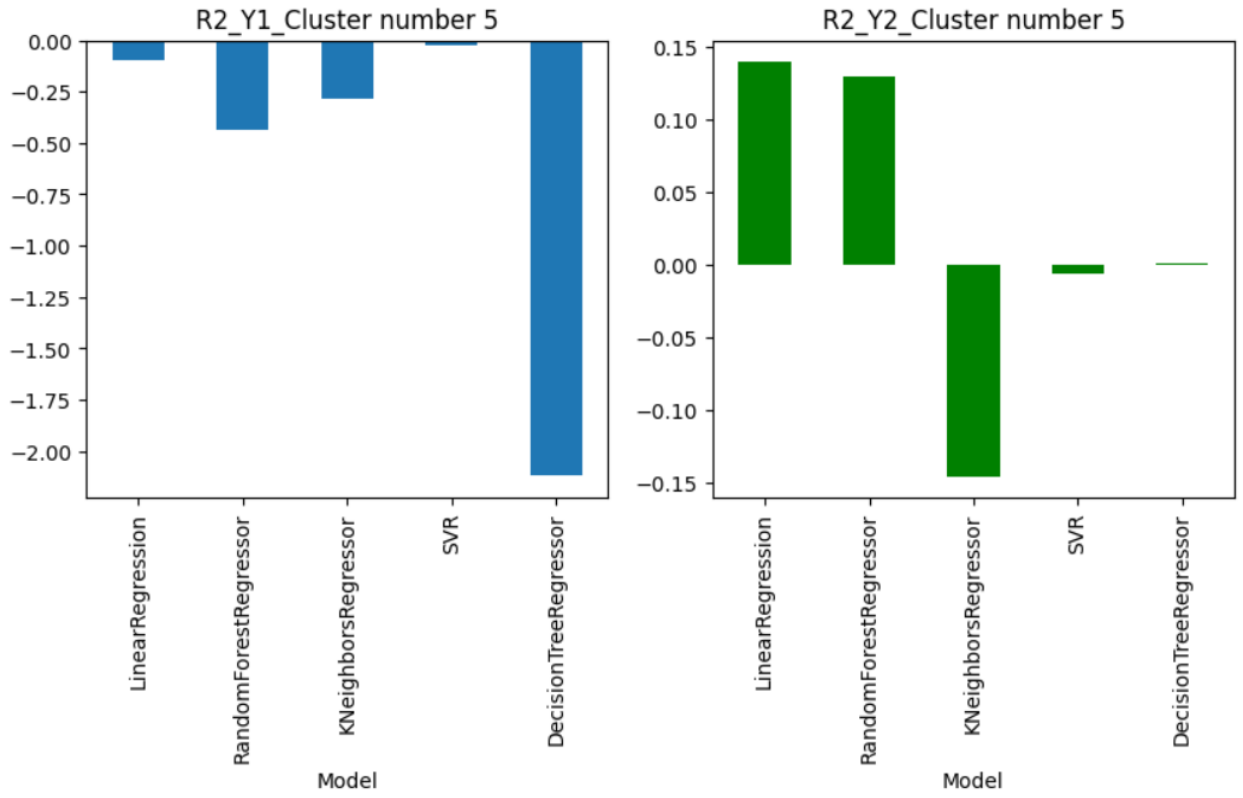

Рисунок 18

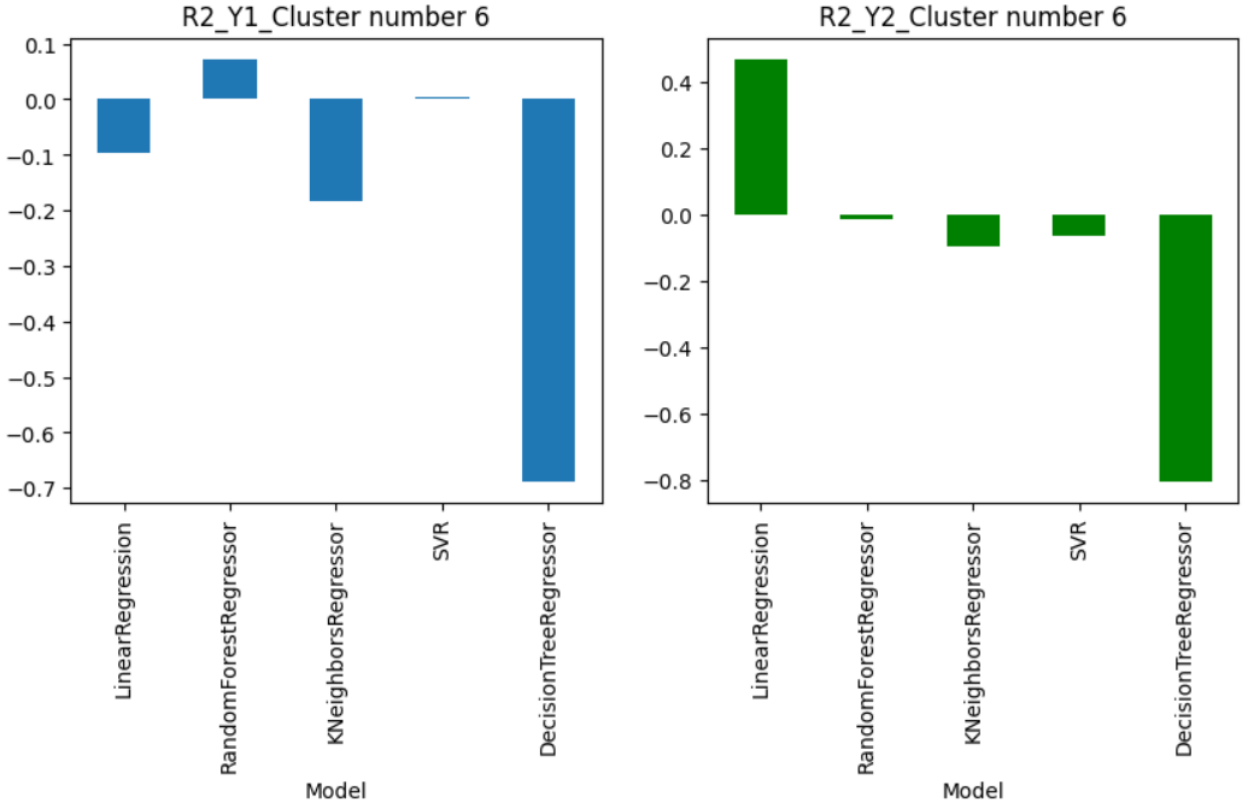

Рисунок 19

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

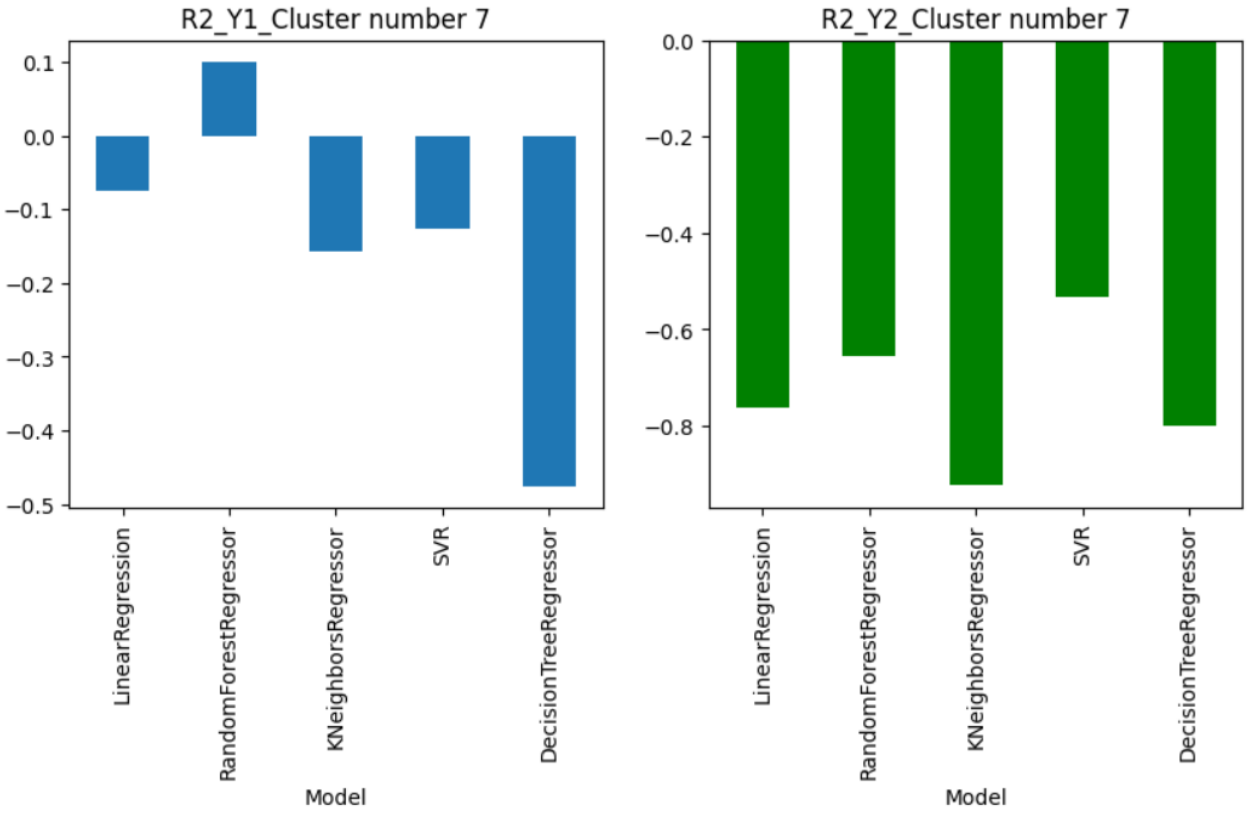

Рисунок 20

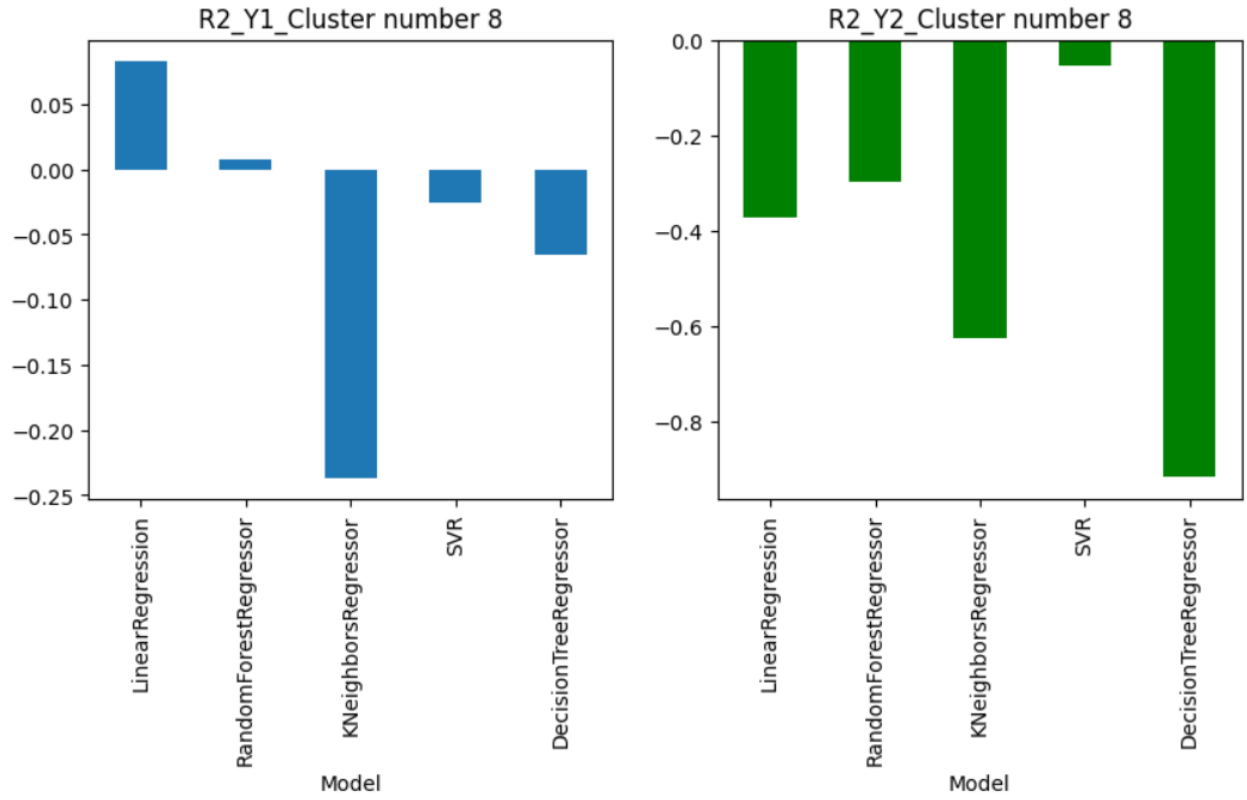

Рисунок 21

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

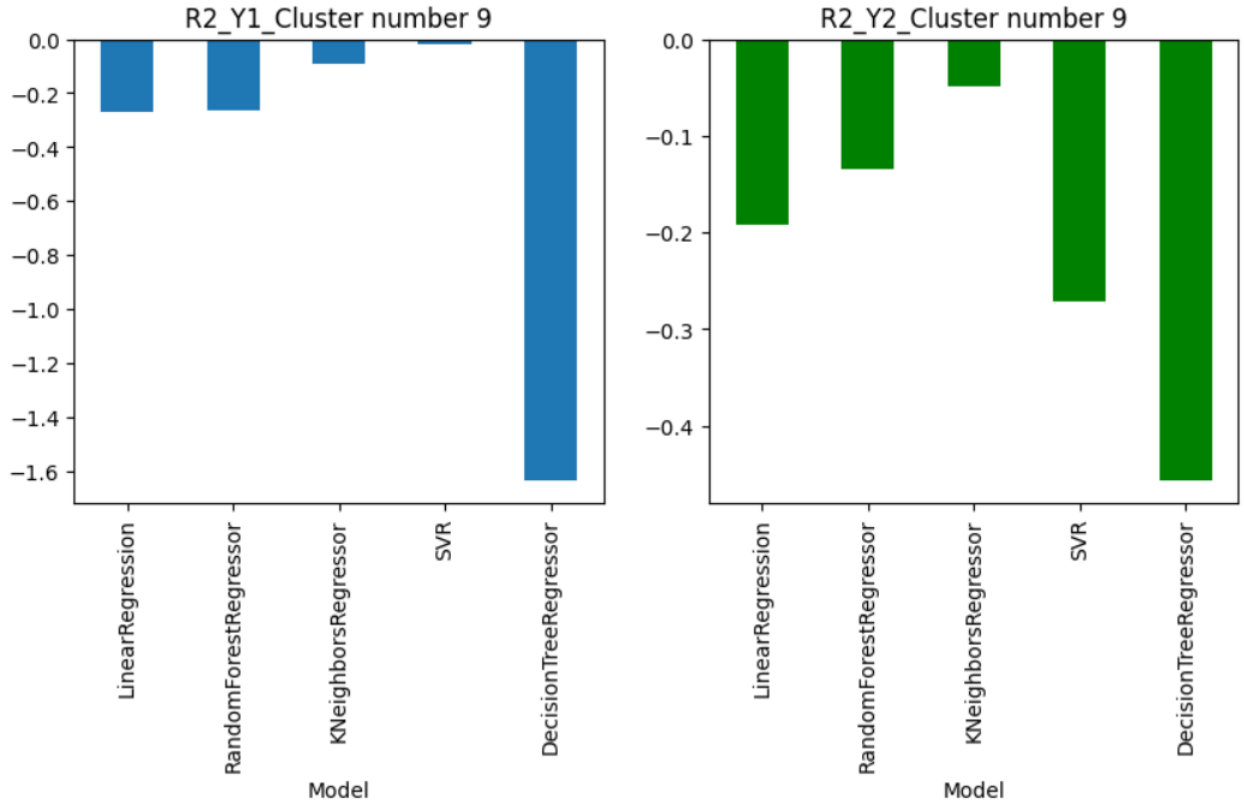

Рисунок 22

3.4. Построение моделей нейронных сетей

Первым этапом, как и в случае, описанном в пункте 3.3, было построение модели для всего объема данных, дабы сравнить эффективность.

Задача нейронной сети стоит предсказать «Соотношение матрица-наполнитель».

При построении моделей нейронных сетей были опробованы различные архитектуры, отличающиеся количеством нейронов в слоях, количеством скрытых слоев, активационными функциями в каждом слое, а также были опробованы различные оптимизаторы, и различные параметры размера «BATCH_SIZE» и различным шагом оптимизации ошибки.

При построении модели без учета кластеров, модель стремилась выдать среднее значение выходного параметра. Подробнее этот феномен отражен в приложенном к работе файле «Composits», график

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

соотношения полученных результатов относительно реальных данных представлен ниже на изображении .

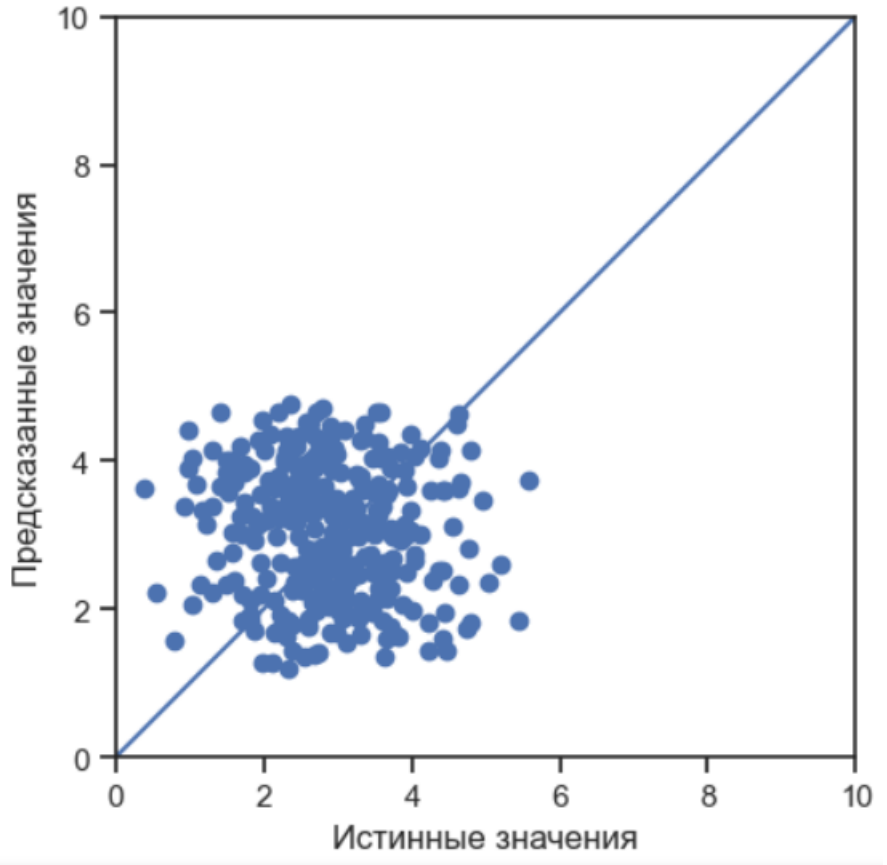

Рисунок 23 - график зависимости предсказанных значений от реальных данных

Как видно из выше представленного графика модель обладает слабой предсказательной способностью, значения метрик ошибок среднего и абсолютного значений:

MSE: 1.7967436698816572

MAE: 1.0847012372838152

Однако, при удалении выбросов в выходном значении путем вычисления квантилей 25% и 75% и отбора параметров в промежутке (-σ: σ). В датасете осталось всего 511 примеров, однако модель показала высокую предсказательную способность, представленную на изображении ниже.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

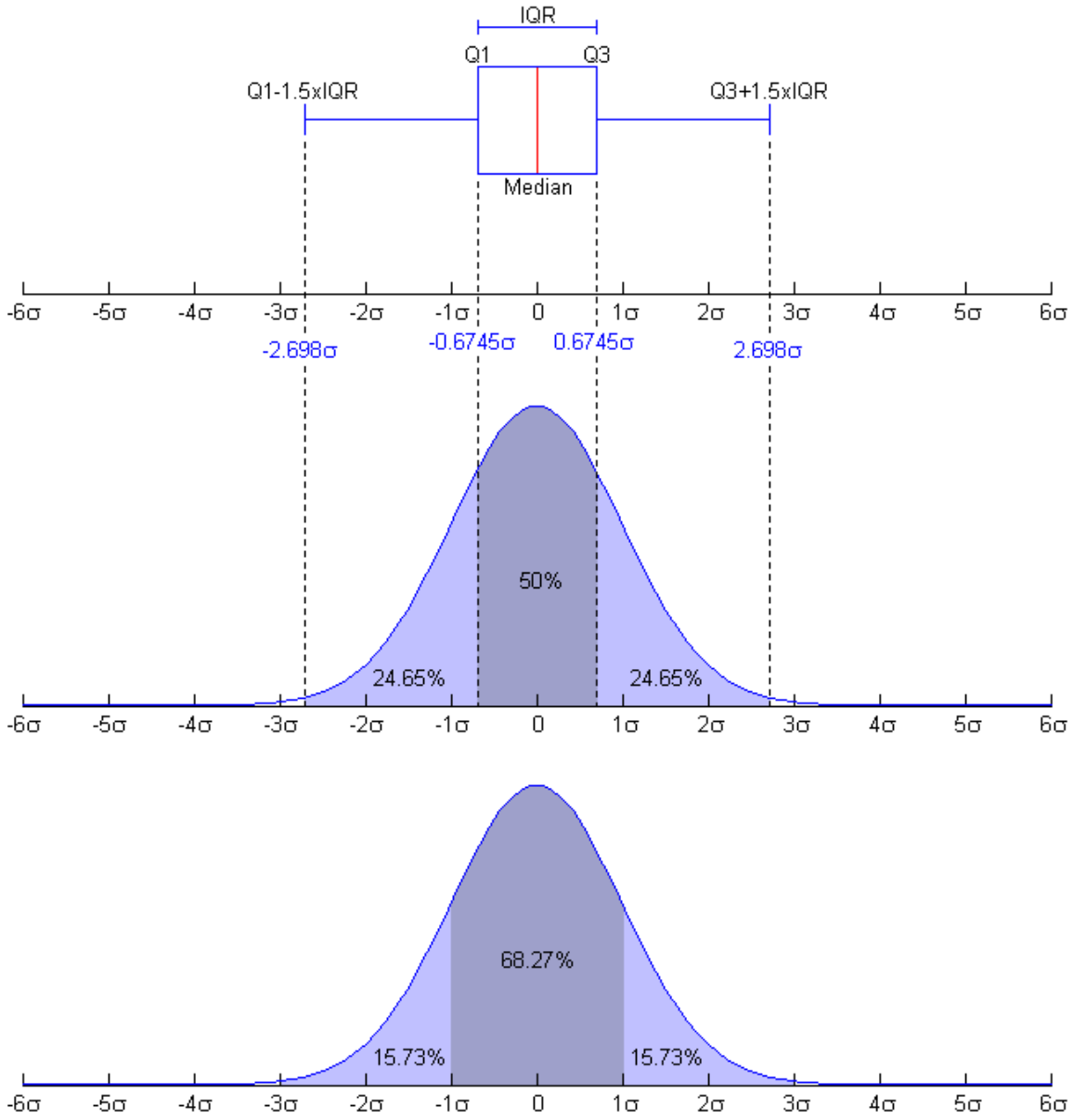

Рисунок 24 - Квантили распределения данных

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

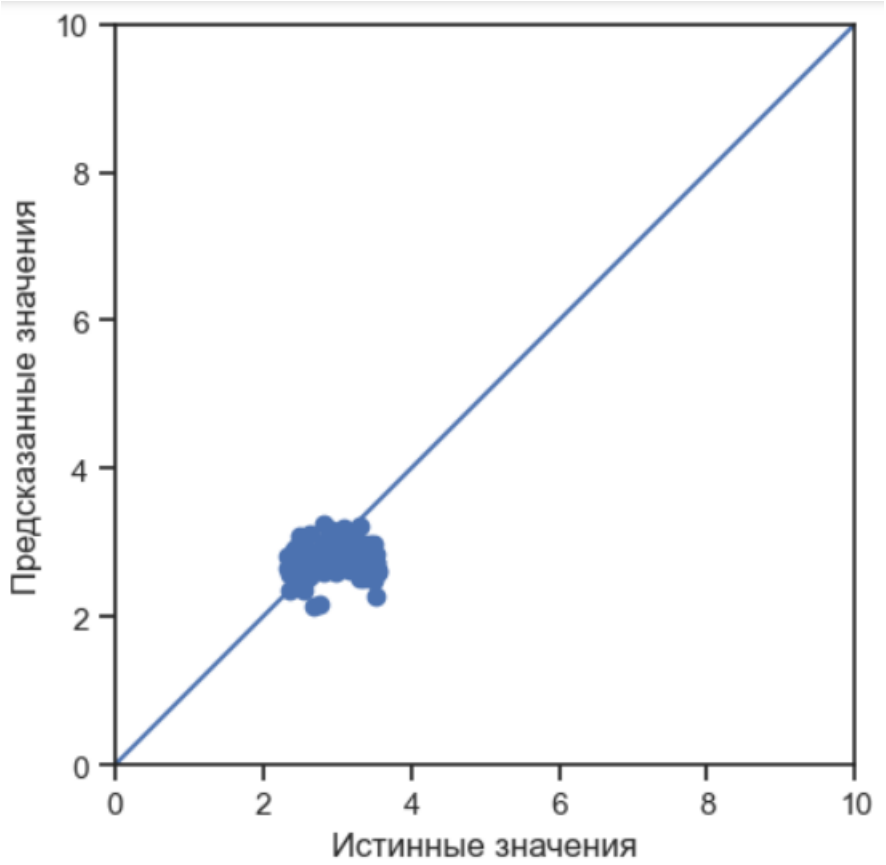

Рисунок 25 - график зависимости предсказанных значений от реальных данных

Как видно из выше представленного графика модель обладает высокой предсказательной способностью, значения метрик ошибок среднего и абсолютного значений:

MSE: 0.19210222967694704

MAE: 0.34745544799657413

Подробнее с данной моделью можно ознакомиться в приложенном к работе файле «Composits».

Однако, так как модели было представлено только около 50% исходных данных, в условиях реальной эксплуатации данная модель может вести себя некорректно.

В процессе разработки нейронных моделей, императивным путем, были выделены активационные функции и оптимизаторы выдающие лучшие результаты для данной задачи.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Оптимизаторы:

- Adam;
- Lion;
- Nadam;
- RMSprop.

Активационные функции:

- mean_absolute_percentage_error;
- huber_loss;
- log_cosh;
- mean_absolute_percentage_error;
- cosine_similarity;
- mean_absolute_error;
- mean_squared_error;
- mean_squared_logarithmic_error.

При построении модели без разбития на кластеры, но с указанием его в параметрах датасета бинарным указателем нейронная сеть предсказывает среднее значение предсказываемой выборки, что отчетливо видно на ниже представленном графике на рисунке 26.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

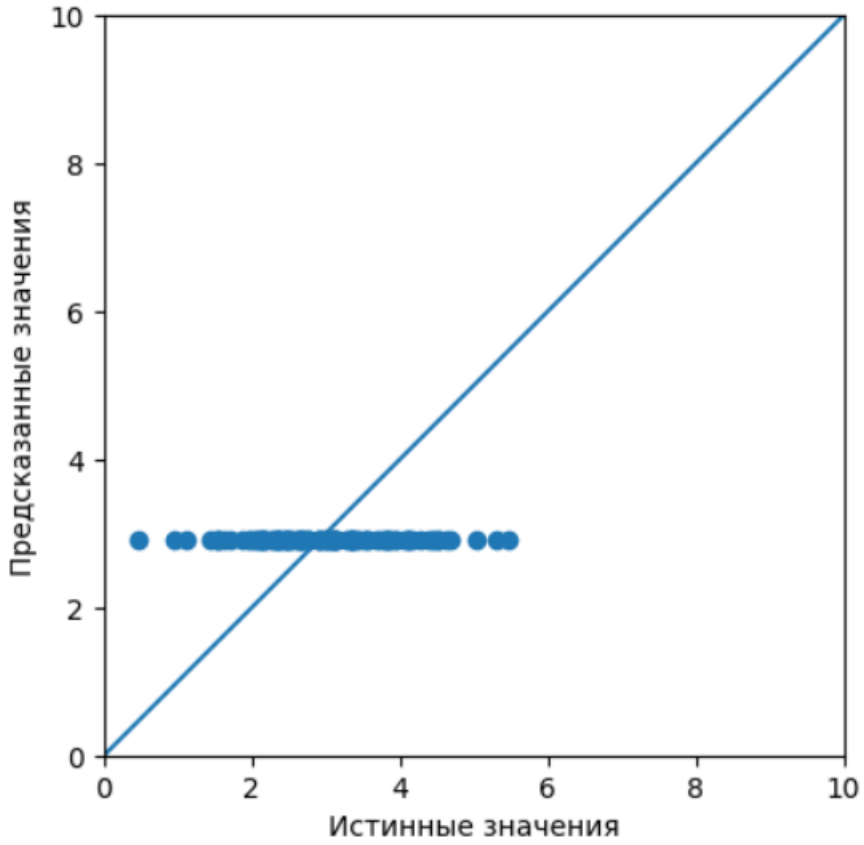

Рисунок 26 - график зависимости предсказанных значений от реальных данных

Как видно из выше представленного графика модель обладает низкой предсказательной способностью, однако имеет относительно хорошие значения метрик ошибок среднего и абсолютного значений:

MSE: 0.9181766605715228

MAE: 0.762379275068659

Далее были построены модели на основе датасетов по каждому кластеру отдельно, выделенных из исходного датасета. Так как в данном случае данных для обучения моделей было крайне мало (график распределения значений по кластерам представлен на рисунке 11 в пункте 3.1), некоторые модели показывали слабую предсказательную способность. Ниже представлены графики зависимости предсказанных значений от реальных значений, по каждой модели, по каждому кластеру, а также метрики ошибок MSE и MAE.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

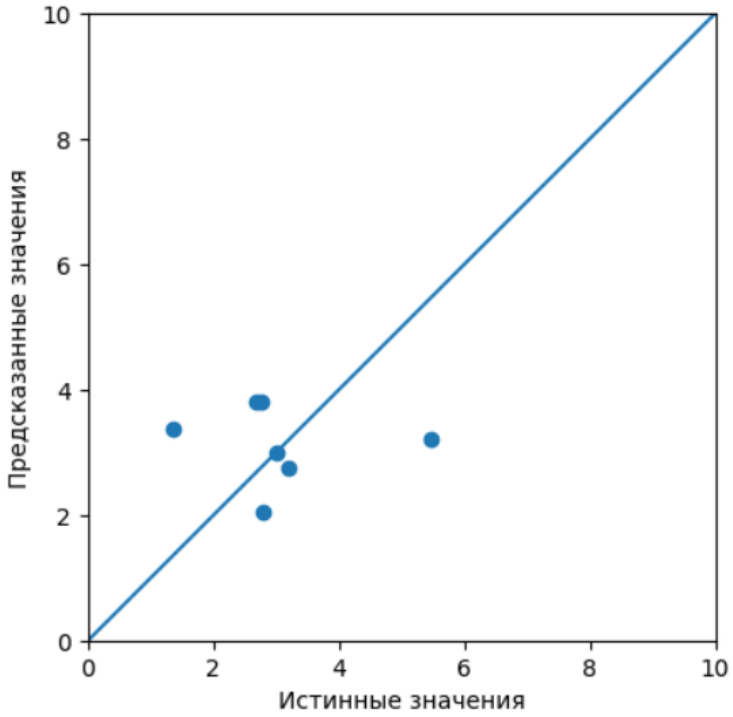

Рисунок 27 - график зависимости предсказанных значений от реальных данных. Кластер 1

MSE: 1.726057299883385

MAE: 1.0845084271180496

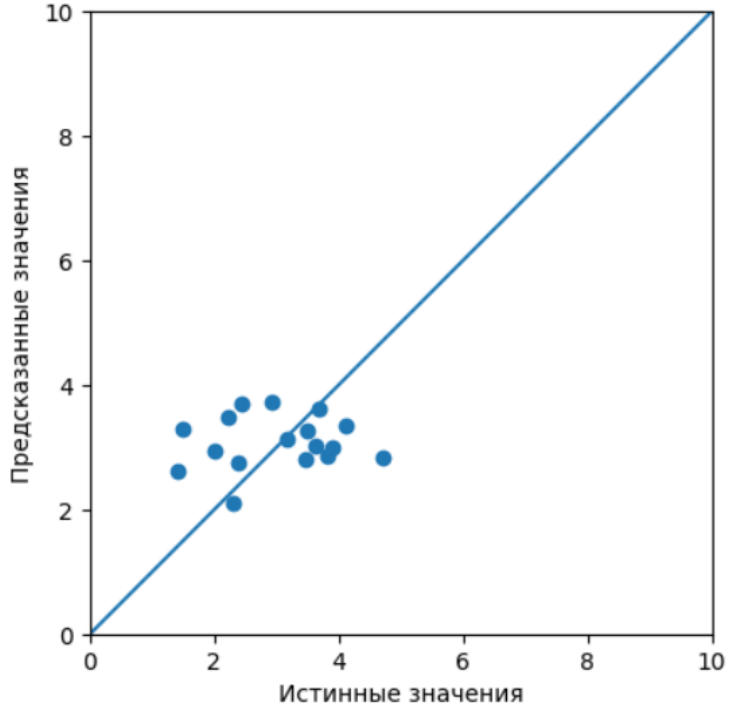

Рисунок 28 - график зависимости предсказанных значений от реальных данных. Кластер 2

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

MSE: 0.9541391192833084

MAE: 0.8204983900564922

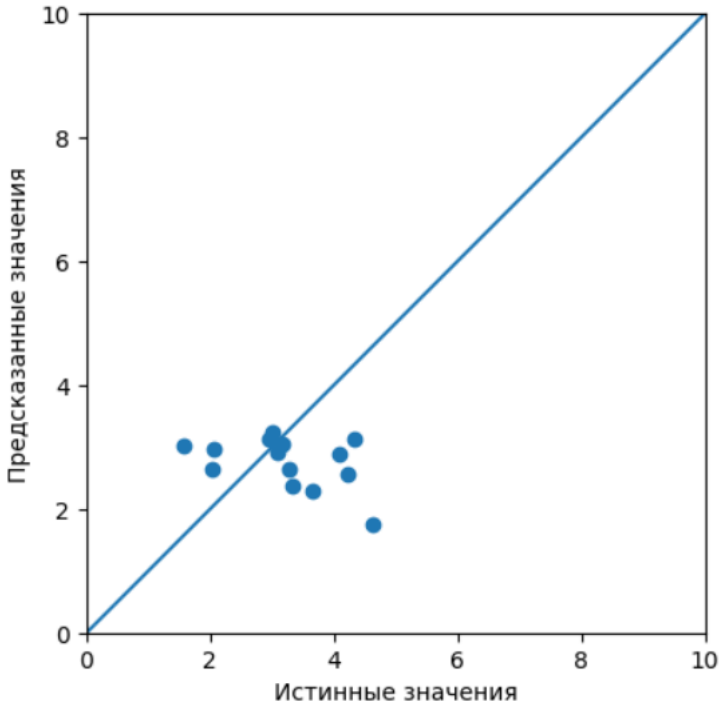

Рисунок 29 - график зависимости предсказанных значений от реальных данных. Кластер 3

MSE: 1.474501586086653

MAE: 0.9732379831518747

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

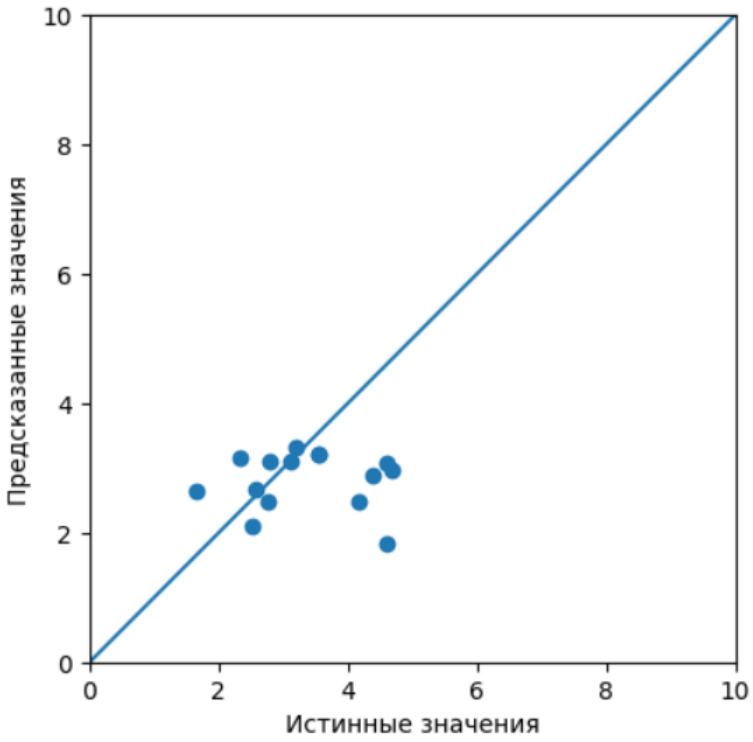

Рисунок 30 - график зависимости предсказанных значений от реальных данных. Кластер 4

MSE: 1.3572798063226463

MAE: 0.8637601013722993

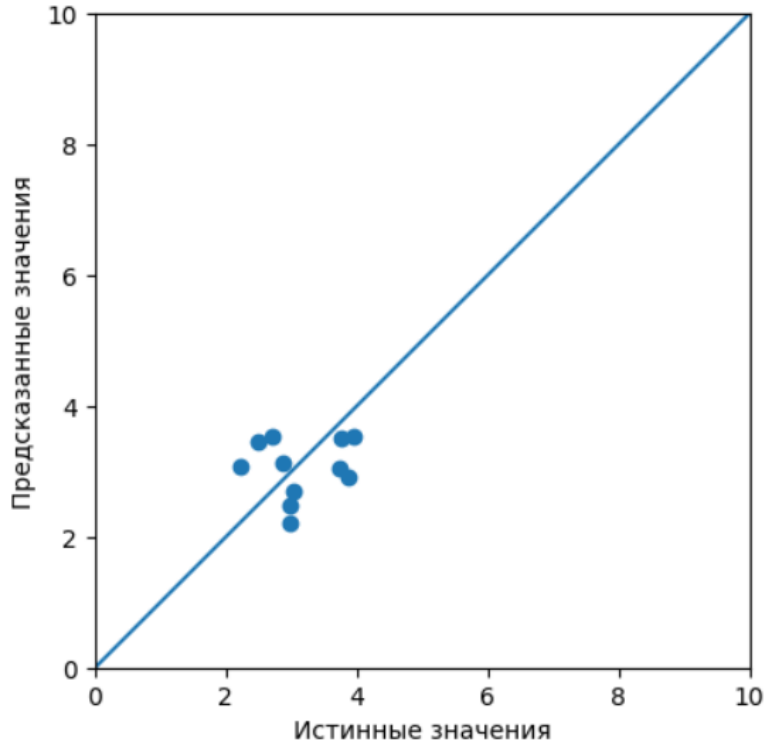

Рисунок 31 - график зависимости предсказанных значений от реальных данных. Кластер 5

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

MSE: 0.45287324222281367

MAE: 0.6213970056410778

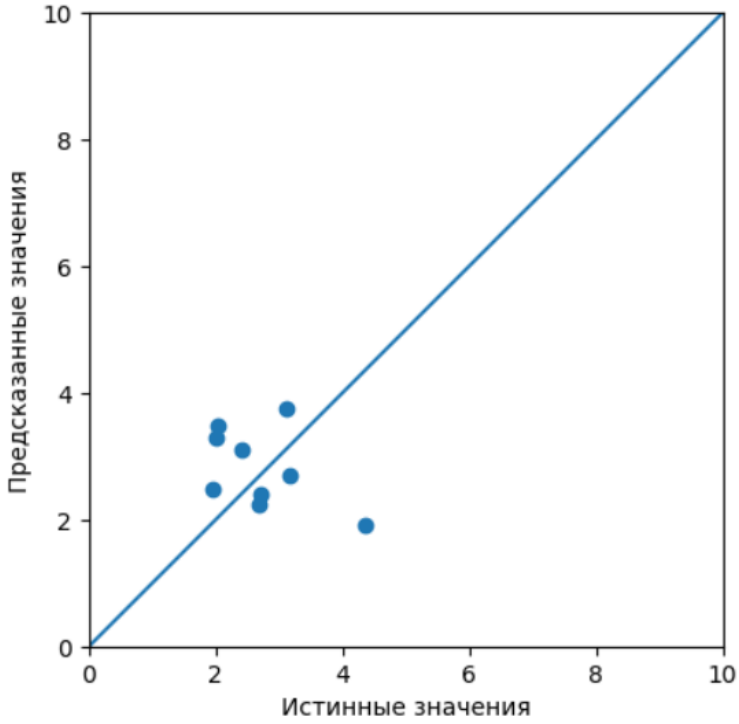

Рисунок 32 - график зависимости предсказанных значений от реальных данных. Кластер 6

MSE: 1.2621897469113819

MAE: 0.9165320609118083

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

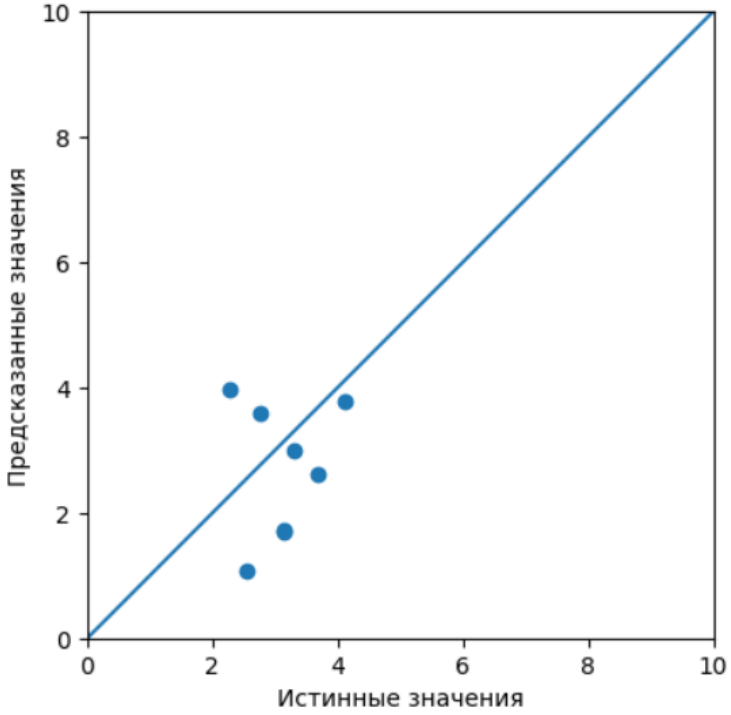

Рисунок 33 - график зависимости предсказанных значений от реальных данных. Кластер 7

MSE: 1.4187558395440685

MAE: 1.080128836572404

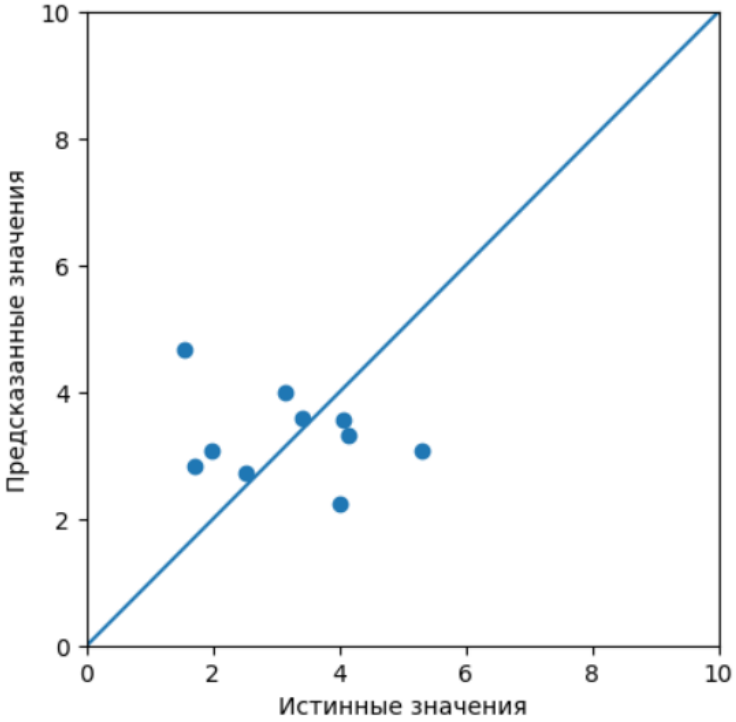

Рисунок 34 - график зависимости предсказанных значений от реальных данных. Кластер 8

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

MSE: 2.197816314711701

MAE: 1.1856097208214451

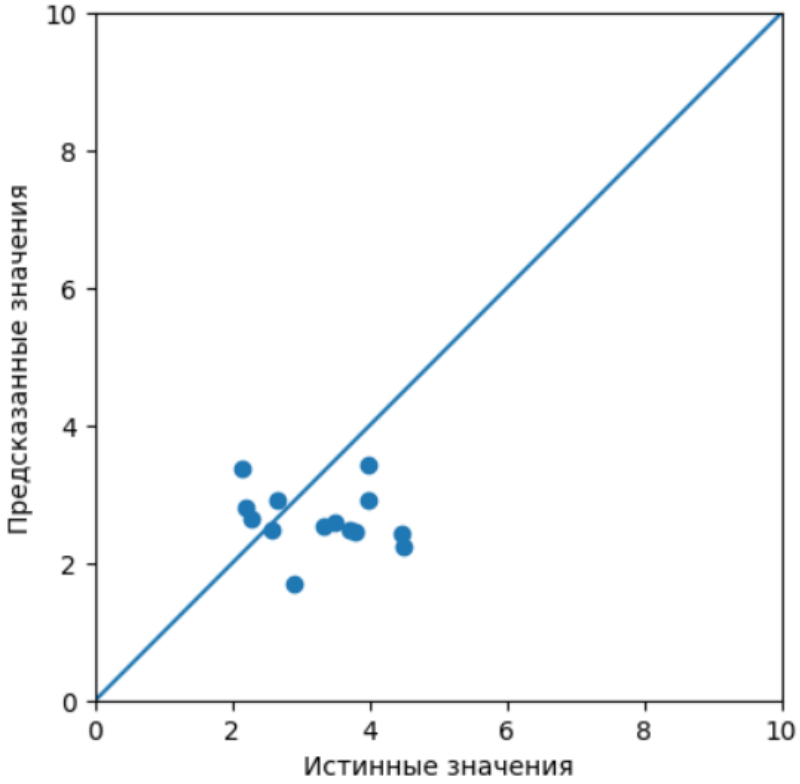

Рисунок 35 - график зависимости предсказанных значений от реальных данных. Кластер 9

MSE: 1.348130332618307

MAE: 0.9938701539967868

Данные модели обладают худшей предсказательной способностью, однако на графике видно лучшее предсказание результатов, лучшие учитывается разброс данных.

Подробнее с данными моделями и результатами работы моделей можно ознакомиться в приложенном к работе файлу «Composites_4».

Подробнее с приложением можно ознакомиться в репозитории на git-hab: https://github.com/cupiscube/Composite_materials

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

4. Построение приложения для демонстрации применимости моделей в условиях производства

Приложение было построено в виде back-end сервиса принимающего на вход индекса /get_predict серилиализуемый JSON файл в виде датасета, возвращает JSON файл с номером определенного класса и предсказание.

Приложение было построено на фреймворке flask. Приложение при запуске загружает построенные, сохраненные модели, при обращении к приложению происходит определение класса объекта, затем нормализация данных, затем предсказание моделью значения.

Подробнее с приложением можно ознакомиться в репозитории на git-hab: https://github.com/cupiscube/composites_app

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

5. Выводы

Нейронные сети показывают наибольшую гибкость при построении моделей, хотя и значительно сложнее при их проектировании. На выданных мне данных получилось построить некоторые удачные модели нейросетей, однако к применению их в промышленности я рекомендовать не могу. Задача является довольно перспективной для внедрения полученных моделей в производство и последующее его оптимизирование. Однако, так как выборка является относительно маленькой и дополнительно предобработанной для улучшения результатов требуется дальнейшее изучение данных с привлечением экспертов области, и с увеличением исходных данных.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Библиографический список

1. https://scikit-learn.org/stable/
2. https://www.tensorflow.org/api_docs
3. https://keras.io/
4. https://wiki.loginom.ru/articles/coefficient-of-determination.html