ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА ПО КУРСУ «DATA SCIENCE»

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель

Решетников Александр Анатольевич

Постановка задачи

Цель исследования состоит в прогнозировании ряда конечных свойств получаемых композиционных материалов на основе пула входящих параметров.

Задачи исследования:

- Изучить теоретические основы и методы решения поставленной задачи;
- Провести разведочный анализ данных;
- Провести предобработку данных;
- Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении;
- Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель;
- Разработать приложение;
- Оценить точность модели на тренировочном и тестовом датасете;
- Создать удаленный репозиторий и разместить там код исследования. Оформить файл readme.

Загруженные данные

Датасет X_bp.xlsx:

Тип данных:	float64
Количество параметров:	10
Количество записей:	1023

Датасет X_nup.xlsx:

Тип данных:	float64
Количество параметров:	3
Количество записей:	1040

В датасете X_bp.xlsx присутствуют целочисленные значения, но загруженные типы данных определены как числа с плавающей запятой двойной точности. Строковых данных нет.

Датасет X_nup.xlsx имеет на 17 записей больше, чем X_bp.xlsx.

Разведочный анализ: Анализ датафреймов

Датафрейм df_bp:

Количество пропущенных значений:	0
Количество NULL значений:	0
Количество не уникальных значений:	
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

Датафрейм df_nup:

Количество пропущенных значений:	0
Количество NULL значений:	0
Количество не уникальных значений:	
Угол нашивки, град:	2
Шаг нашивки:	1006
Плотность нашивки:	1005

Датафреймы имеет не уникальные значения. Угол нашивки представлен как бинарный параметр, но т.к. предоставлен усеченный набор данных, то считаем, что параметр может и иметь другие углы нашивки, не попавшие в исходный датасет.

Разведочный анализ: Анализ датафреймов

Датафрейм df_bp:

Количество нулевых строк:	0
Количество дублирующихся строк :	0
Количество дублирующихся значений:	
Соотношение матрица-наполнитель:	9
Плотность, кг/м3:	10
модуль упругости, ГПа:	3
Количество отвердителя, м.%:	18
Содержание эпоксидных групп,%_2:	19
Температура вспышки, С_2:	20
Поверхностная плотность, г/м2:	19
Модуль упругости при растяжении, Гпа:	19
Прочность при растяжении, МПа:	19
Потребление смолы, г/м2:	20

Датафрейм df_nup:

Количество нулевых строк:	1
Количество дублирующихся строк :	19
Количество дублирующихся значений:	
Шаг нашивки:	34
Плотность нашивки:	35

Датафрейм df_nup имеет нулевые и дублирующиеся строки. Угол нашивки исключен из анализа, т.к. представлен как имеет только значения 0 и 90. Оба датафрейма имеют дублирующиеся значения, что не характерно для чисел с плавающей запятой. Удалим дубли.

Разведочный анализ: Объединение датафреймов

Датафрейм df_bp:

Тип данных:	float64
Количество параметров:	10
Количество записей:	1002

Датафрейм df_nup:

Тип данных:	float64
Количество параметров:	3
Количество записей:	1002

Датафрейм df:

Тип данных:	float64
Количество параметров:	13
Количество записей:	1002

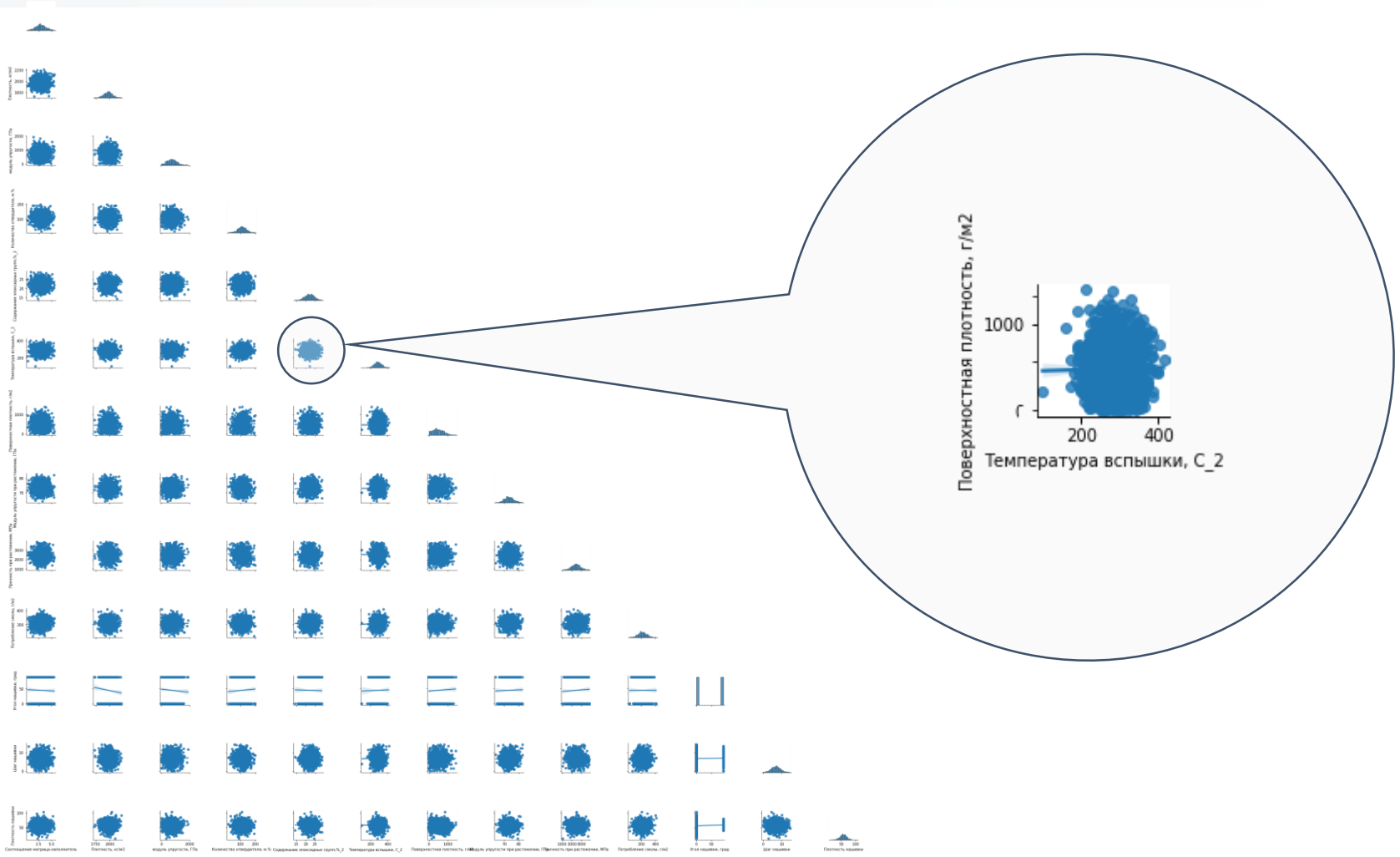
После удаления нулевых и дублирующихся строк, а так же строк с дублирующимися значениями размеры датафреймов стали равны по количеству записей, что позволило объединить их в один датафрейм.

Разведочный анализ: Описательная статистика

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1002.0	2.930165	0.913871	0.389403	2.317247	2.907832	3.552781	5.591742
Плотность, кг/м3	1002.0	1975.675440	73.757180	1731.764635	1924.370115	1977.574305	2021.186675	2207.773481
модуль упругости, ГПа	1002.0	740.098060	330.030446	2.436909	498.538615	741.148111	962.650230	1911.536477
Количество отвердителя, м.%	1002.0	110.479158	28.396466	17.740275	92.054117	110.162666	130.240418	198.953207
Содержание эпоксидных групп,%_2	1002.0	22.242882	2.404798	14.254985	20.563359	22.230761	23.981598	28.955094
Температура вспышки, С_2	1002.0	285.739807	41.343587	100.000000	258.469516	285.853960	313.472775	413.273418
Поверхностная плотность, г/м2	1002.0	482.649366	280.682398	0.603740	267.736782	451.944708	693.654483	1399.542362
Модуль упругости при растяжении, ГПа	1002.0	73.326808	3.118688	64.054061	71.297280	73.247594	75.365124	82.682051
Прочность при растяжении, МПа	1002.0	2467.050190	485.889244	1036.856605	2141.720311	2461.249253	2760.983489	3848.436732
Потребление смолы, г/м2	1002.0	218.290295	59.840786	33.803026	179.147494	217.277006	257.488673	414.590628
Угол нашивки, град	1002.0	44.910180	45.022382	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1002.0	6.907026	2.557644	0.037639	5.132313	6.909686	8.564373	14.440522
Плотность нашивки	1002.0	57.234866	12.330789	11.740126	49.922625	57.362576	65.094083	103.988901

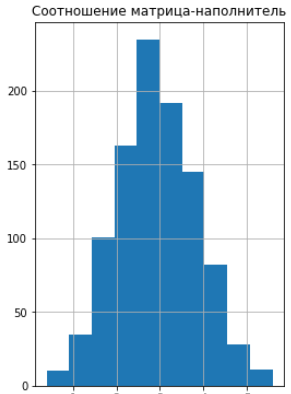
При незначительном разбросе значений соотношения матрица-наполнитель достаточно большой разброс по параметрам "Модуль упругости", "Поверхностная плотность", "Прочность при растяжении".

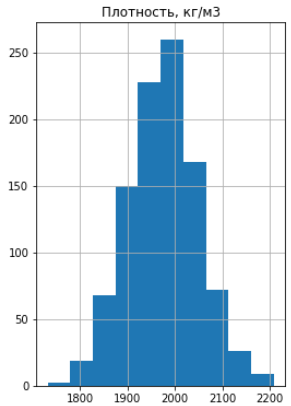
Разведочный анализ: Парные графики рассеяния точек

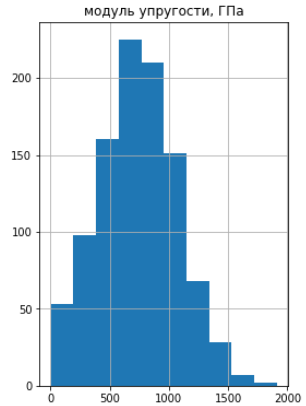


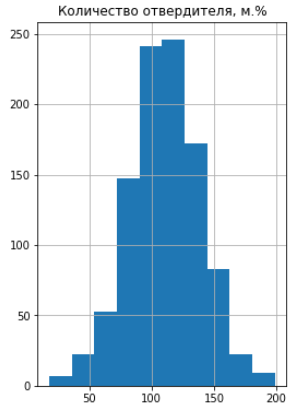
По всем параметрам наблюдается наличие выбросов и отсутствие корреляции.

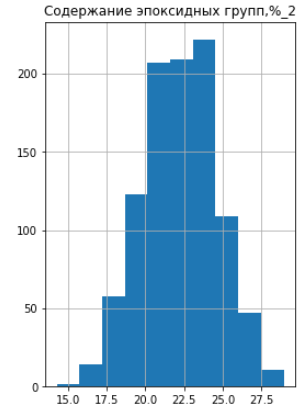
Разведочный анализ: Гистограммы распределения

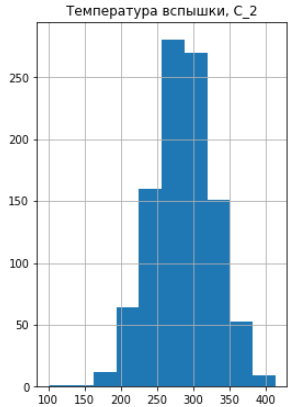


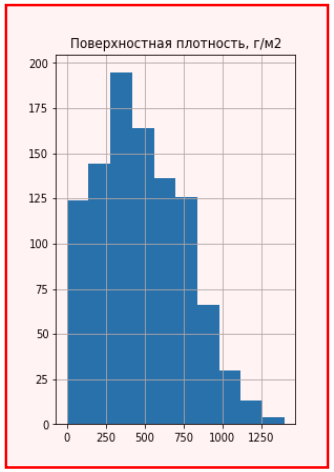


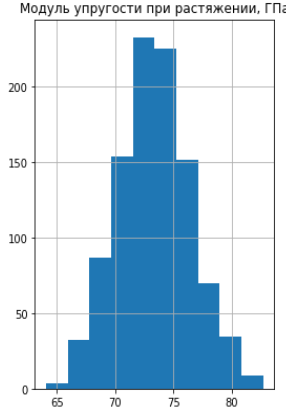


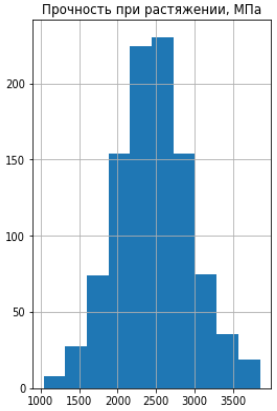


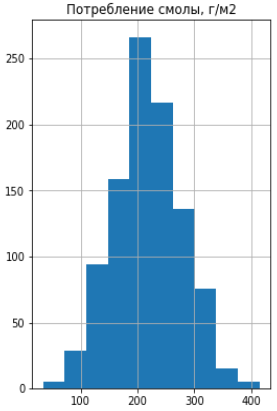


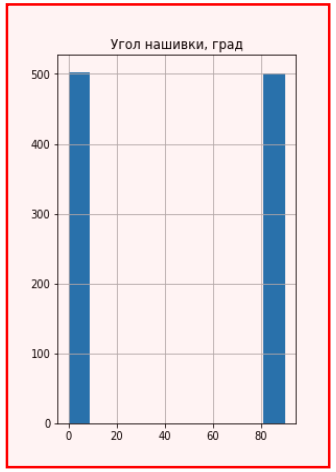


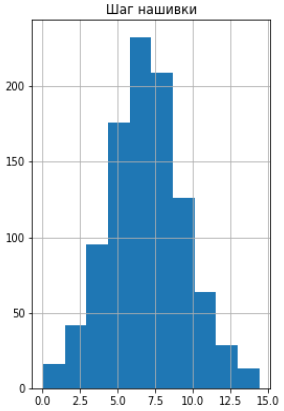


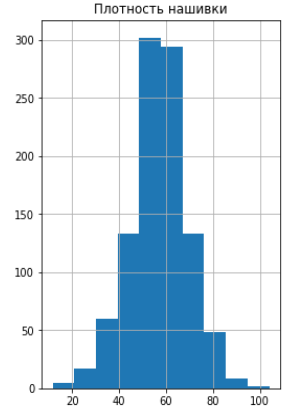












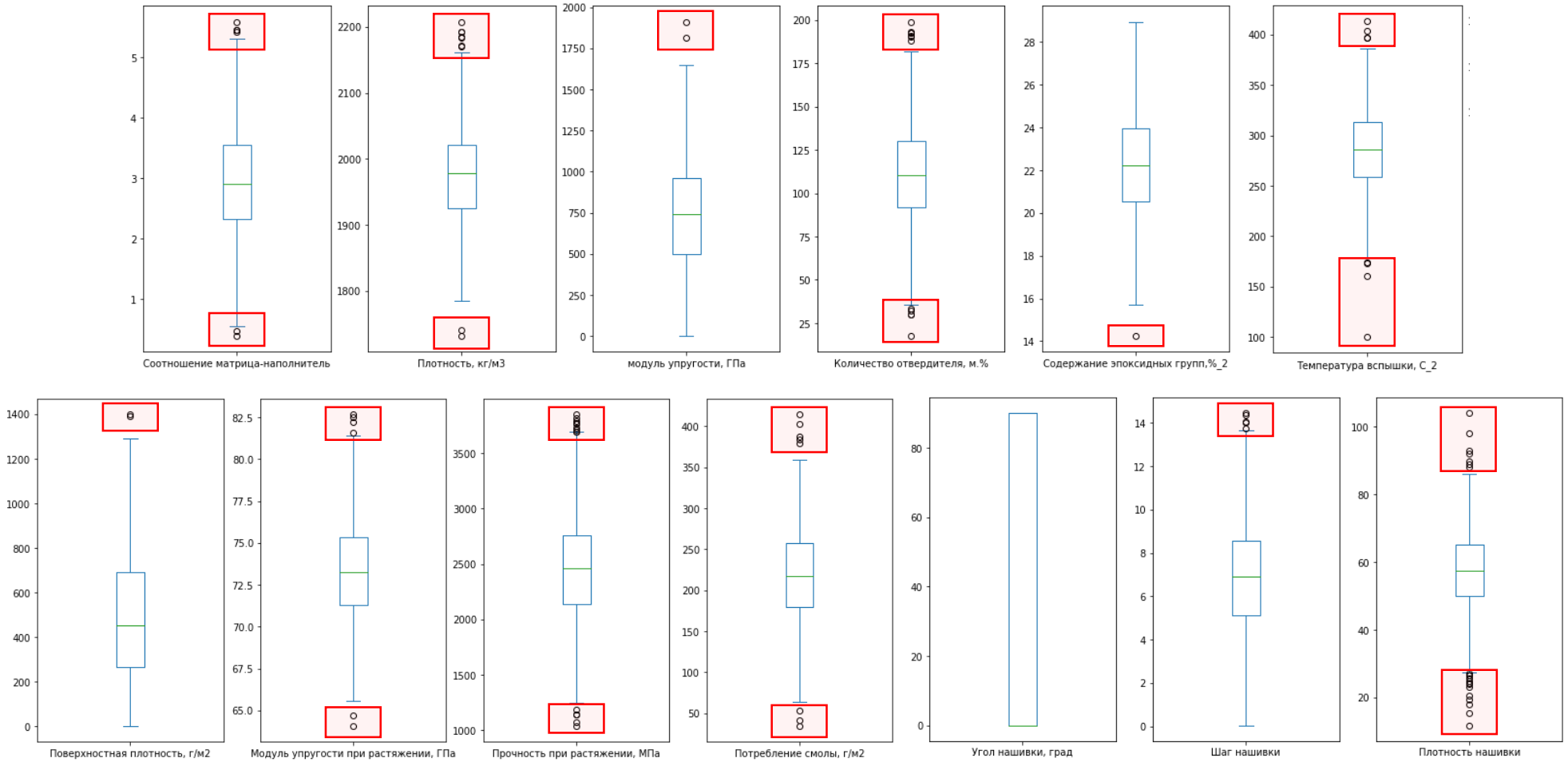
Гистограммы имеют нормальное распределение, за исключением параметров "Поверхностная плотность" и "Угол нашивки".

Разведочный анализ: Тепловая карта корреляции

	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	0.009730	0.032527	-0.013712	0.027562	-0.006074	-0.010999	-0.015854	0.033971	0.077503	-0.017127	0.004688	0.046602
Плотность, кг/м3	0.009730	1.000000	-0.017816	-0.034475	-0.011411	-0.021585	0.037709	-0.021791	-0.072360	-0.026858	-0.063110	-0.060183	0.035423
модуль упругости, ГПа	0.032527	-0.017816	1.000000	0.023692	-0.006371	0.031032	-0.017991	0.014716	0.047515	-0.008100	-0.037088	0.022978	-0.014466
Количество отвердителя, м.%	-0.013712	-0.034475	0.023692	1.000000	0.013128	0.093342	0.050919	-0.073184	-0.070399	0.006991	0.029513	-0.037946	-0.001998
Содержание эпоксидных групп,%_2	0.027562	-0.011411	-0.006371	0.013128	1.000000	-0.008176	-0.007539	0.064473	-0.030984	0.015714	-0.010078	-0.005978	-0.040914
Температура вспышки, С_2	-0.006074	-0.021585	0.031032	0.093342	-0.008176	1.000000	0.019709	0.027867	-0.031211	0.060217	0.010441	0.018672	-0.017855
Поверхностная плотность, г/м2	-0.010999	0.037709	-0.017991	0.050919	-0.007539	0.019709	1.000000	0.015804	0.012743	0.001643	0.030177	-0.009088	-0.005019
Модуль упругости при растяжении, ГПа	-0.015854	-0.021791	0.014716	-0.073184	0.064473	0.027867	0.015804	1.000000	0.009559	0.044586	0.013813	0.039314	0.015160
Прочность при растяжении, МПа	0.033971	-0.072360	0.047515	-0.070399	-0.030984	-0.031211	0.012743	0.009559	1.000000	0.028685	0.027367	-0.063270	-0.012882
Потребление смолы, г/м2	0.077503	-0.026858	-0.008100	0.006991	0.015714	0.060217	0.001643	0.044586	0.028685	1.000000	0.001263	0.017253	0.015883
Угол нашивки, град	-0.017127	-0.063110	-0.037088	0.029513	-0.010078	0.010441	0.030177	0.013813	0.027367	0.001263	1.000000	0.029419	0.115374
Шаг нашивки	0.004688	-0.060183	0.022978	-0.037946	-0.005978	0.018672	-0.009088	0.039314	-0.063270	0.017253	0.029419	1.000000	-0.001475
Плотность нашивки	0.046602	0.035423	-0.014466	-0.001998	-0.040914	-0.017855	-0.005019	0.015160	-0.012882	0.015883	0.115374	-0.001475	1.000000

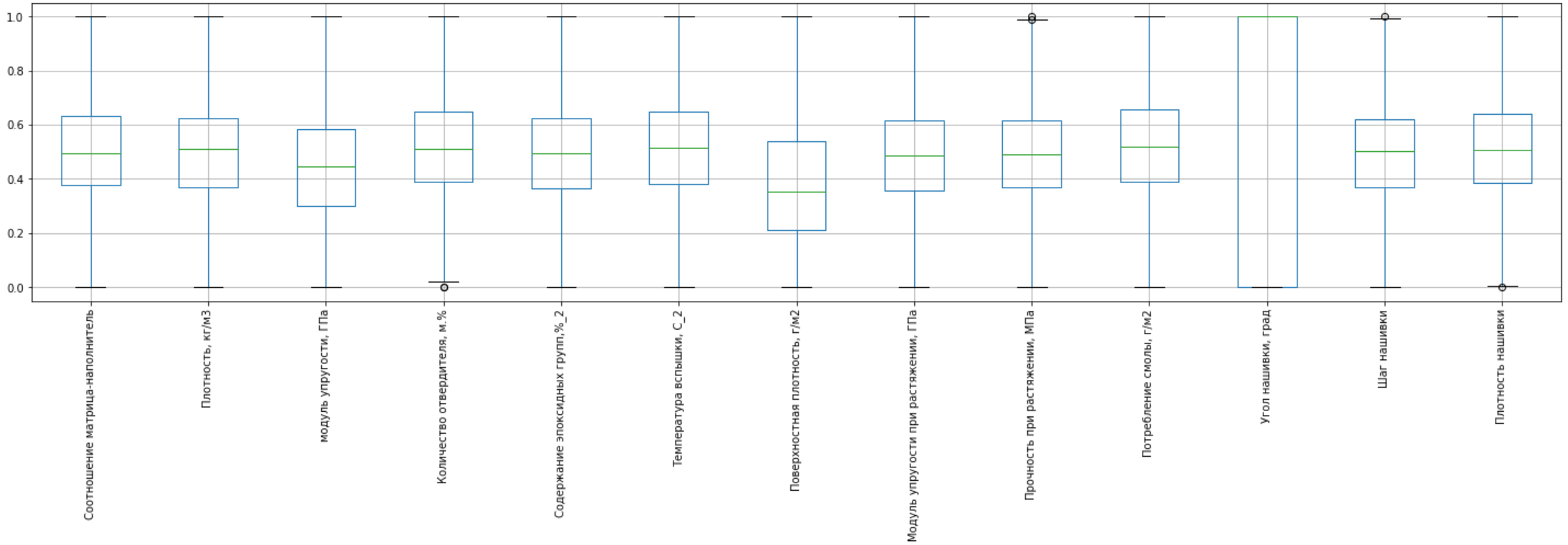
Полученная тепловая карта так же свидетельствует об отсутствии явной корреляции между признаками.

Разведочный анализ: Выбросы



Полученные ящики с усами так же свидетельствуют о наличии выбросов.

Предобработка данных



На стадии предобработки данных произведено удаление выбросов и нормализация данных.

Модели прогноза модуля упругости при растяжении

	MAE	MSE	R2
Линейная регрессия	0.159633	0.038589	-0.0259701
Дерево решений	0.15958	0.0377647	-0.00405595
Случайный лес	0.162879	0.0392461	-0.0434404

Показатели прогнозирования:

MAE - измеряет среднюю абсолютную ошибку прогнозов. Для каждой точки вычисляется разница между прогнозами и целью, а затем усредняются эти значения.

MSE - измеряет средний квадрат ошибок прогнозов. Для каждой точки вычисляется квадратная разница между прогнозами и целью, а затем усредняются эти значения.

R^2 - Коэффициент детерминации, или R-квадрат, является еще одним показателем, который мы можем использовать для оценки модели, и он тесно связан с MSE, но имеет преимущество в том, что не имеет значения, являются ли выходные значения очень большими или очень маленькими.

Отрицательное значение коэффициента детерминации близкого к нулю свидетельствует о низком качестве модели и отсутствии линейных связей. MAE и MSE показывает высокие показатели, что так же подтверждает отсутствие линейных связей.

Модели прогноза прочности при растяжении

	MAE	MSE	R2
Линейная регрессия	0.15369	0.0385339	-0.0653631
Дерево решений	0.151809	0.0377405	-0.0434284
Случайный лес	0.151761	0.0377587	-0.0439306

Показатели прогнозирования:

MAE - измеряет среднюю абсолютную ошибку прогнозов. Для каждой точки вычисляется разница между прогнозами и целью, а затем усредняются эти значения.

MSE - измеряет средний квадрат ошибок прогнозов. Для каждой точки вычисляется квадратная разница между прогнозами и целью, а затем усредняются эти значения.

R^2 - Коэффициент детерминации, или R-квадрат, является еще одним показателем, который мы можем использовать для оценки модели, и он тесно связан с MSE, но имеет преимущество в том, что не имеет значения, являются ли выходные значения очень большими или очень маленькими.

Отрицательное значение коэффициента детерминации близкого к нулю свидетельствует о низком качестве модели и отсутствии линейных связей. MAE и MSE показывает высокие показатели, что так же подтверждает отсутствие линейных связей.

Нейронная сеть: Модель

```
def __init__(self, loss='mean_squared_error', optimizer='adam'):
    norm = np.array(X_train)
    self.norm_layer = layers.Normalization(axis=None,
                                           input_dim=X_train.shape[1])
    self.norm_layer.adapt(norm)

    self.model = keras.Sequential()
    self.model.add(self.norm_layer)
    self.model.add(Dense(X_train.shape[1], activation='relu'))
    self.model.add(Dense(X_train.shape[1], activation='relu'))
    self.model.add(layers.Dense(1))
    self.model.summary()

def fit(self):
    self.history = self.model.fit(X_train, y_train, validation_split=0.2,
                        verbose=1, epochs=20, validation_data=(X_test, y_test))
```

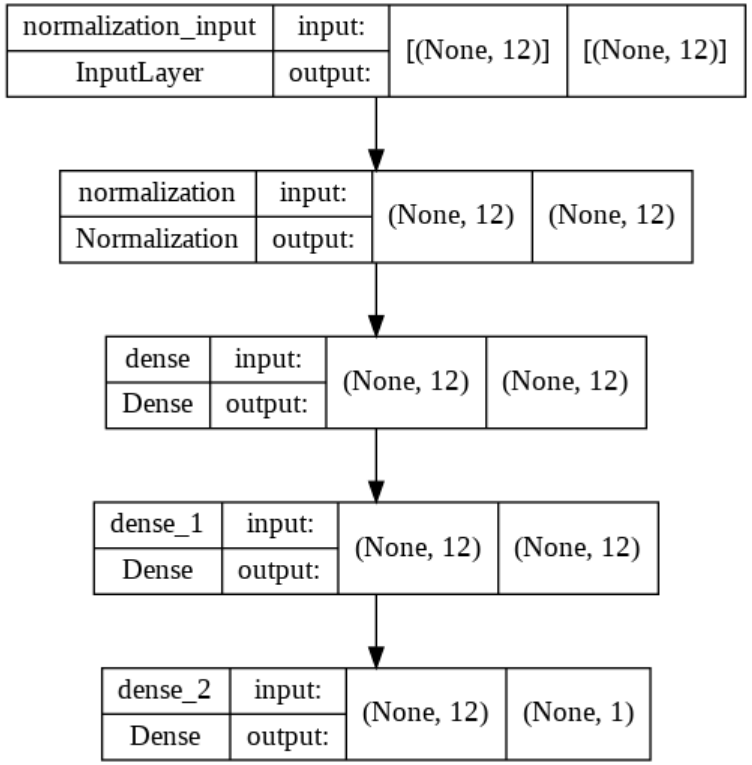
Модель состоит из входного нормализованного слоя и двух скрытых слоев размерностями соответствующими входной модели и активационной функцией ReLu, а также выходной слой размерностью в 1 нейрон. Для компиляции выбираем оптимизатор Adam, а в качестве функции потерь используем среднеквадратичную ошибку.

Нейронная сеть: Сводка

Строковая сводка сети

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization (Normalizatio  (None, 12)               3
 n)

 dense (Dense)               (None, 12)                156

 dense_1 (Dense)             (None, 12)                156

 dense_2 (Dense)             (None, 1)                 13

=================================================================
Total params: 328
Trainable params: 325
Non-trainable params: 3
```

Графа зависимостей слоев

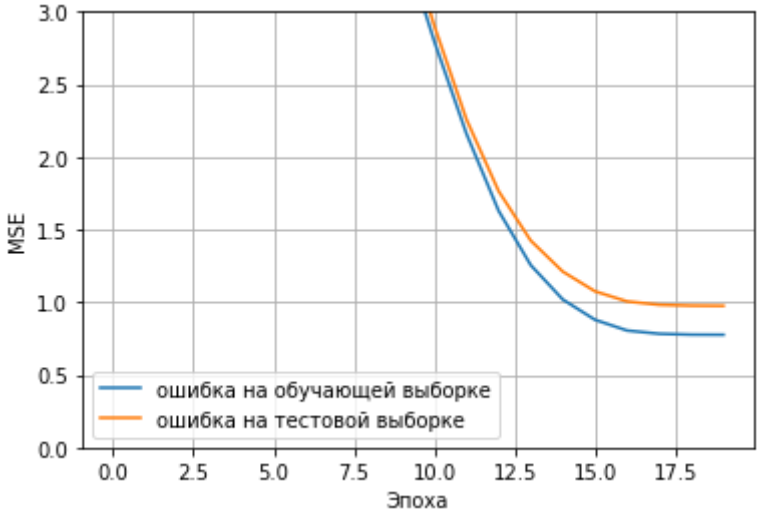


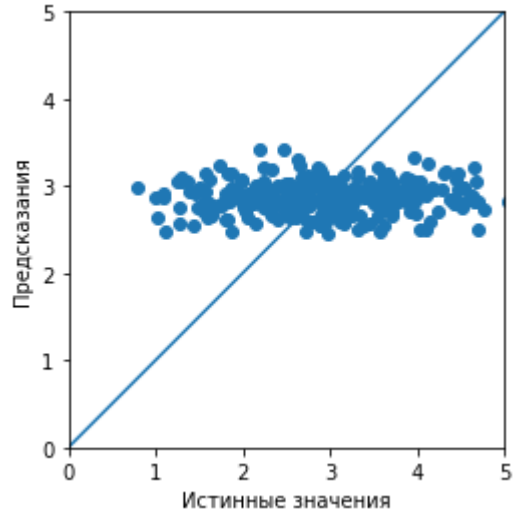
Тренировочных параметров в модели 325.

Нейронная сеть: Результаты

```
Epoch 1/20
17/17 [==============================] - 1s 12ms/step - loss: 9.2618 - val_loss: 9.5636
…
Epoch 20/20
17/17 [==============================] - 0s 4ms/step - loss: 0.7759 - val_loss: 0.9752
```

MAE: 0.7529630048280076
MSE: 0.8525619071439539
R2: -0.025004005511227057





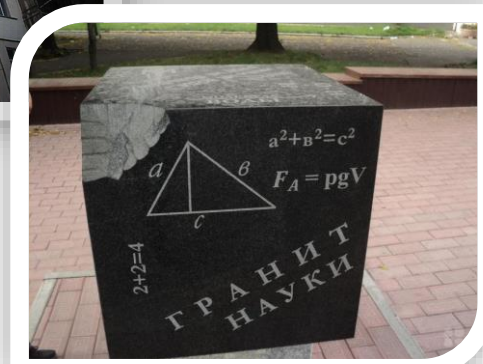
Отрицательное значение коэффициента детерминации близкого к нулю свидетельствует о низком качестве рекомендательной модели соотношения матрица-наполнитель и отсутствии линейных связей. MAE и MSE показывает высокие показатели, что так же подтверждает отсутствие линейных связей.

Результаты

Результаты построения и обучения моделей не дали положительного результата. Возможные причины неудовлетворительной работы моделей:

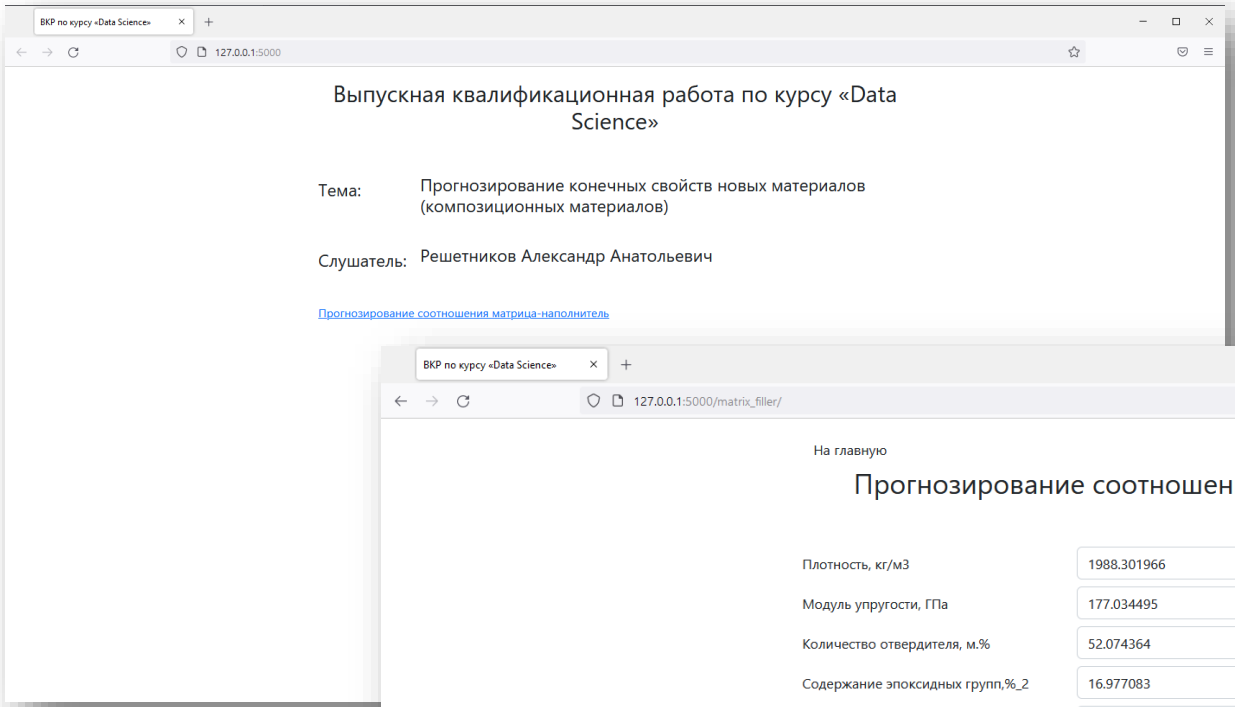
Нечеткая постановка задачи, отсутствие дополнительной информации о зависимости между входными датасетами, что привело к неверному объединению датасетов;

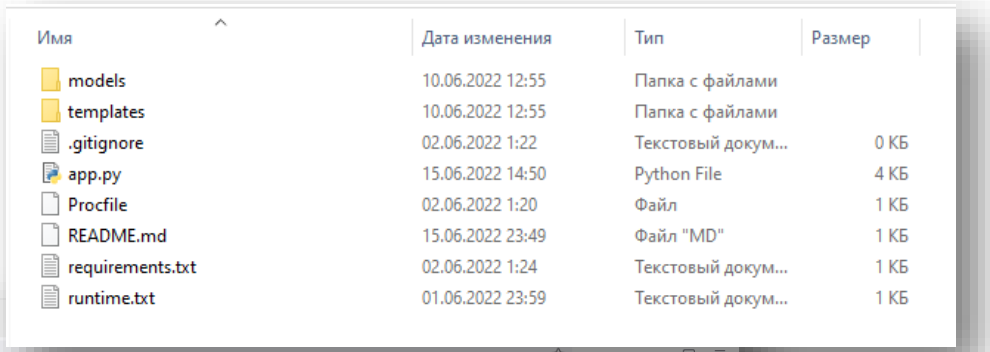
Исследование проводилось на предварительно обработанных (дополненных) датасетах. Возможно, на исходных датасетах можно было бы получить более качественные регрессионные модели;

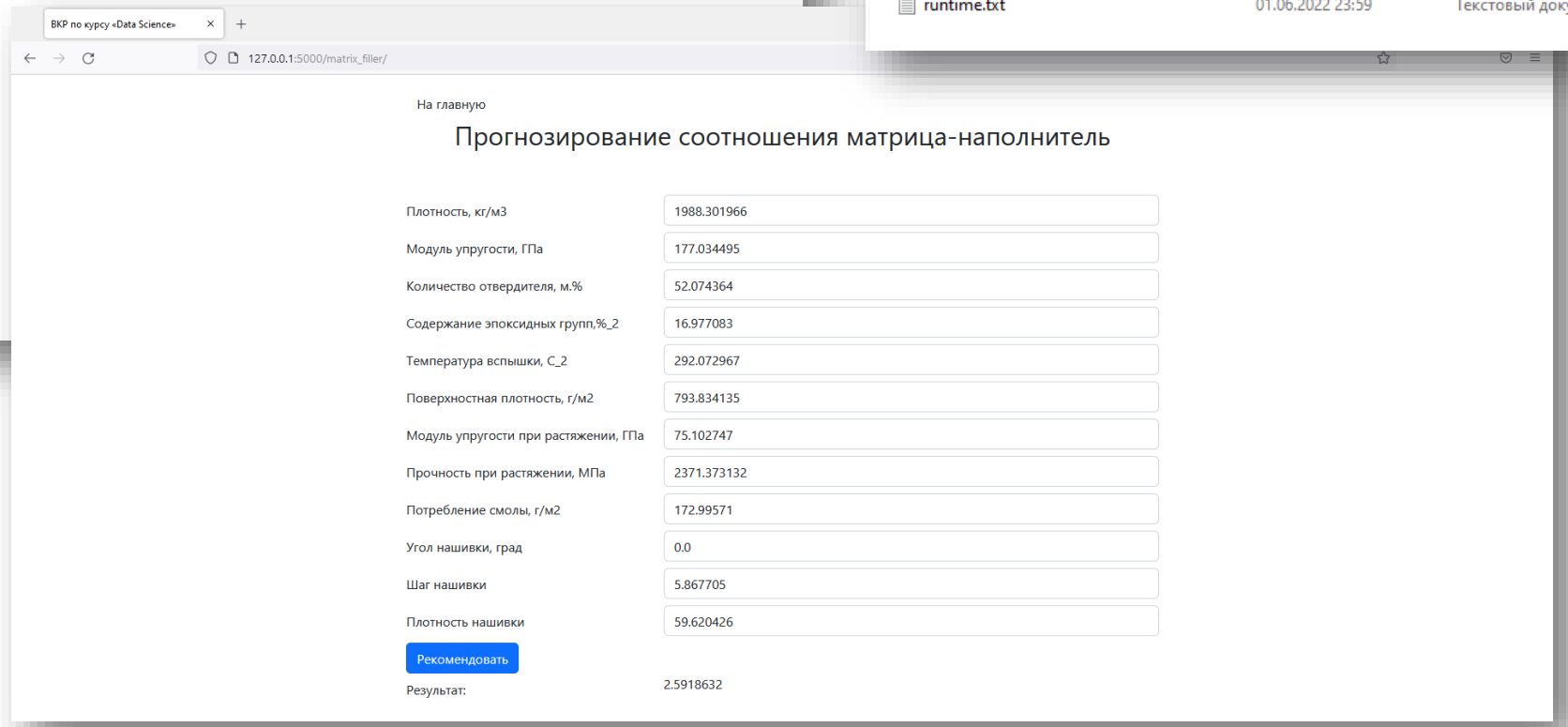


Недостаток знаний и опыта, были испробованы не все возможные модели прогноза.

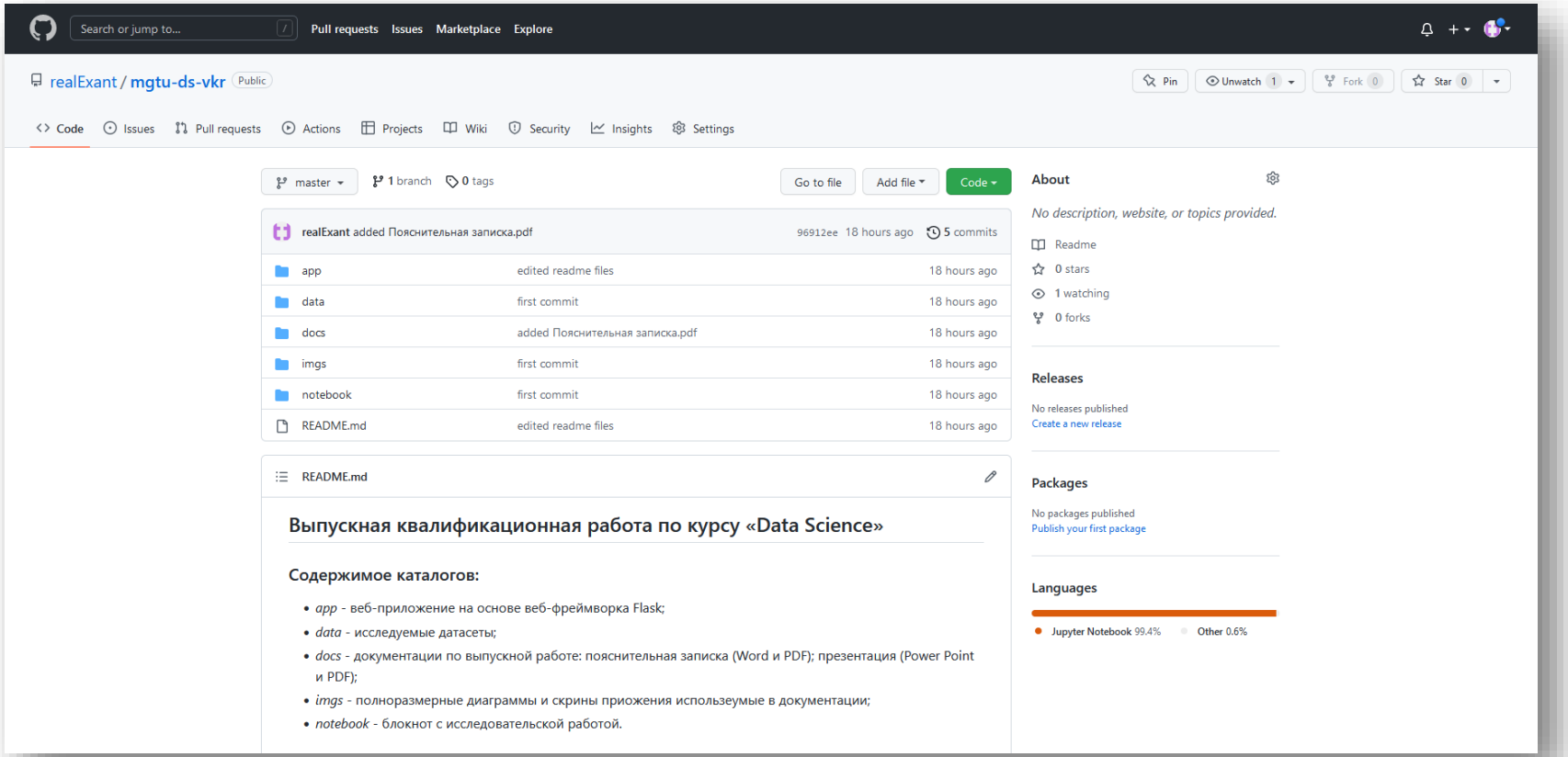
Разработанное приложение







GitHub


Search or jump to...
Pull requests
Issues
Marketplace
Explore
realExant / mgtu-ds-vkr
Public
Pin
Unwatch 1
Fork 0
Star 0
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
master
1 branch
0 tags
Go to file
Add file
Code
realExant added Пояснительная записка.pdf
96912ee 18 hours ago
5 commits
app edited readme files 18 hours ago
data first commit 18 hours ago
docs added Пояснительная записка.pdf 18 hours ago
imgs first commit 18 hours ago
notebook first commit 18 hours ago
README.md edited readme files 18 hours ago
README.md
Выпускная квалификационная работа по курсу «Data Science»
Содержимое каталогов:
app - веб-приложение на основе веб-фреймворка Flask;
data - исследуемые датасеты;
docs - документации по выпускной работе: пояснительная записка (Word и PDF); презентация (Power Point и PDF);
imgs - полноразмерные диаграммы и скрины приожения используемые в документации;
notebook - блокнот с исследовательской работой.
About
No description, website, or topics provided.
Readme
0 stars
1 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Languages
Jupyter Notebook 99.4%
Other 0.6%

Спасибо за внимание!