

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу

«Data Science»

Тема: Прогнозирование конечных свойств новых
материалов
(композиционных материалов)

Слушатель: Григорьев Сергей Алексеевич

БЛОК № 1

Датасет X_bp.xlsx:

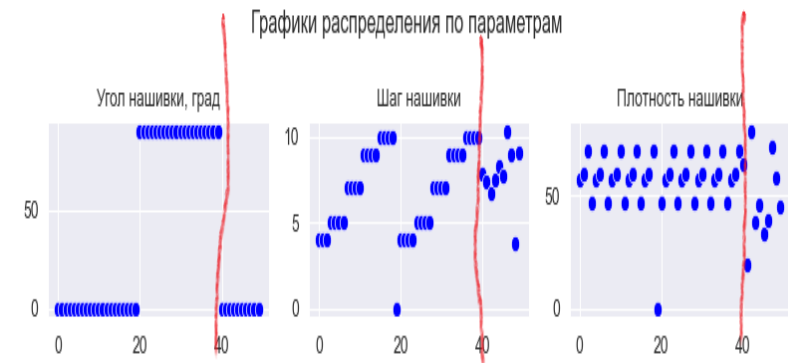
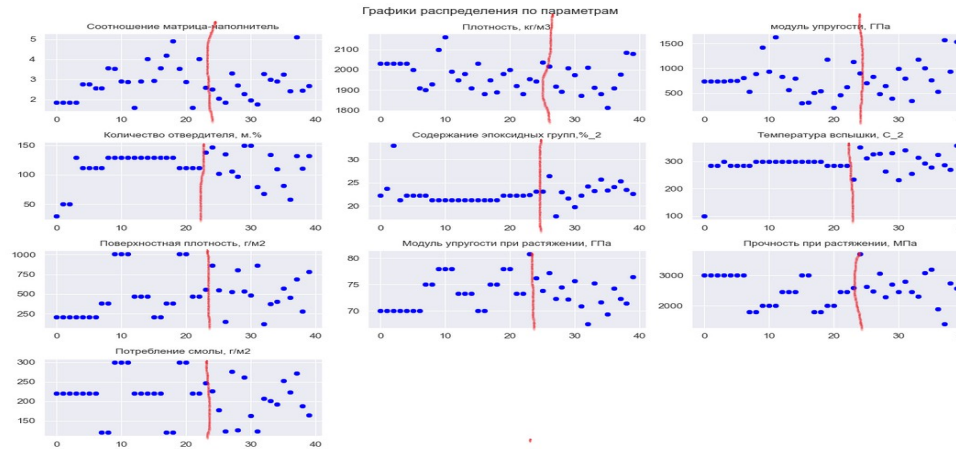
Тип данных: float64
Количество параметров: 10
Количество записей: 1023

Датасет X_nup.xlsx:

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

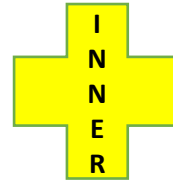
✓ Объединение по индексу INNER:

- Импортируем необходимые библиотеки и модули-функции;
- Загружаем файлы;
- Посмотрим размерность;
- Убираем детерминированные синтетические данные, обнуляем индексы
- Объединим оба файла по индексу по типу объединения INNER
- Заменяем названия столбцов на английский язык



Датасет X_bp.xlsx:

Тип данных: float64
Количество параметров: 10
Количество записей: 1000



Датасет X_nup.xlsx:

Тип данных: float64
Количество параметров: 3
Количество записей: 1000



Датафрейм 'data_main':

Тип данных: float64
Количество параметров: 13
Количество записей: 1000

```
data_main.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
#   Column                                     Non-Null Count  Dtype  
---  -
0   Угол нашивки, град                       1000 non-null   int64  
1   Шаг нашивки                             1000 non-null   float64
2   Плотность нашивки                        1000 non-null   float64
3   Соотношение матрица-наполнитель         1000 non-null   float64
4   Плотность, кг/м3                         1000 non-null   float64
5   модуль упругости, ГПа                    1000 non-null   float64
6   Количество отвердителя, м.%              1000 non-null   float64
7   Содержание эпоксидных групп,%_2         1000 non-null   float64
8   Температура вспышки, С_2                1000 non-null   float64
9   Поверхностная плотность, г/м2           1000 non-null   float64
10  Модуль упругости при растяжении, ГПа     1000 non-null   float64
11  Прочность при растяжении, МПа            1000 non-null   float64
12  Потребление смолы, г/м2                  1000 non-null   float64
dtypes: float64(12), int64(1)
memory usage: 101.7 KB
```



```
data_main.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
#   Column                                     Non-Null Count  Dtype  
---  -
0   pattern_angle                             1000 non-null   int64  
1   step_strip                               1000 non-null   float64
2   density_strip                             1000 non-null   float64
3   ratio_filler_matrix                       1000 non-null   float64
4   density                                   1000 non-null   float64
5   elasticity_module                         1000 non-null   float64
6   number_hardeners                         1000 non-null   float64
7   content_epoxy_groups                     1000 non-null   float64
8   flash_temperature                         1000 non-null   float64
9   surface_density                           1000 non-null   float64
10  elasticity_module_stretching              1000 non-null   float64
11  strapery_strength                        1000 non-null   float64
12  resin_consumption                         1000 non-null   float64
dtypes: float64(12), int64(1)
memory usage: 101.7 KB
```



**ОБРАЗОВАТЕЛЬНЫЙ
ЦЕНТР** МГТУ им. Н. Э. Баумана

БЛОК № 1

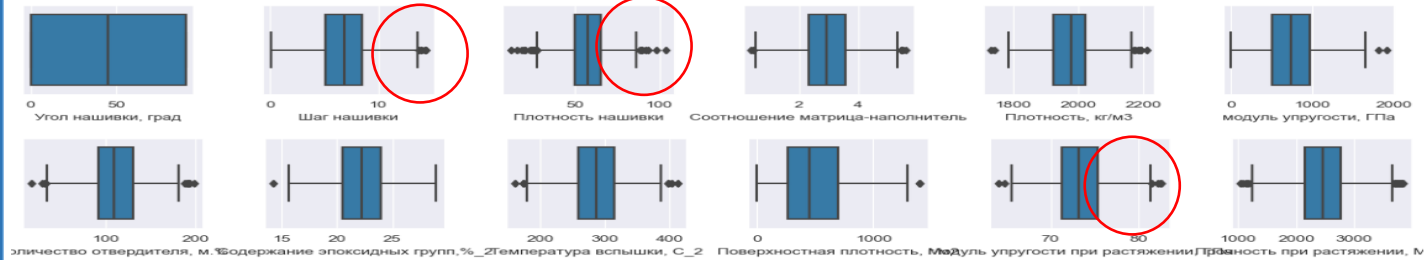
Графики boxplot

Этапы очистки

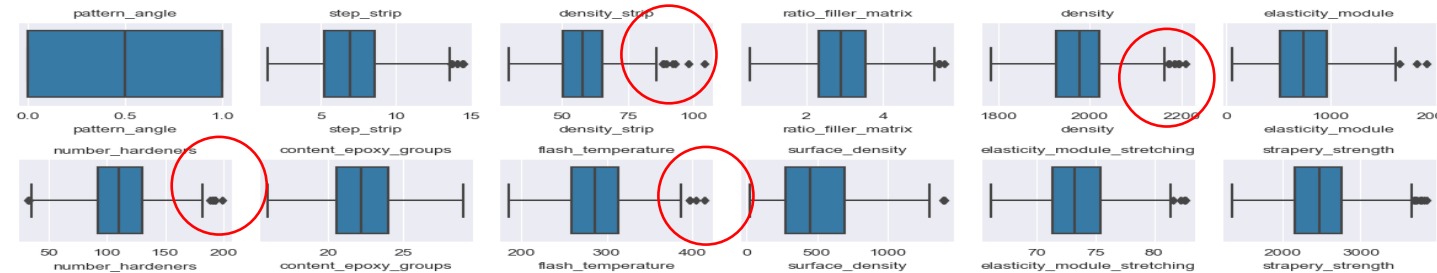
Этап очистки датасета от выбросов:

- Создана функция ЗАМЕНЫ 'подозрительно малых значений' СЛЕВА по оценкам Экспертным МИН и МАХ на основе ближайших 5 значений вместо замены на медиану;
- Проведена очистка по методу IQR (3 сигм) и Z_score;
- Создана функция ЗАМЕНЫ 'подозрительно малых значений' СПРАВА;
- Провели итерацию 4 раза с уточнением МИН и МАХ.
- Получили 'чистый' датасет.

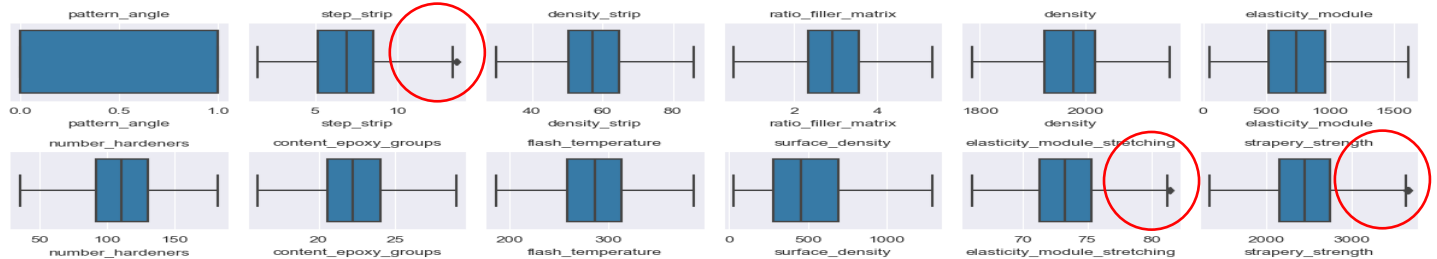
drawing_graphs(data_main, 6, 12, 2, 'boxplot')



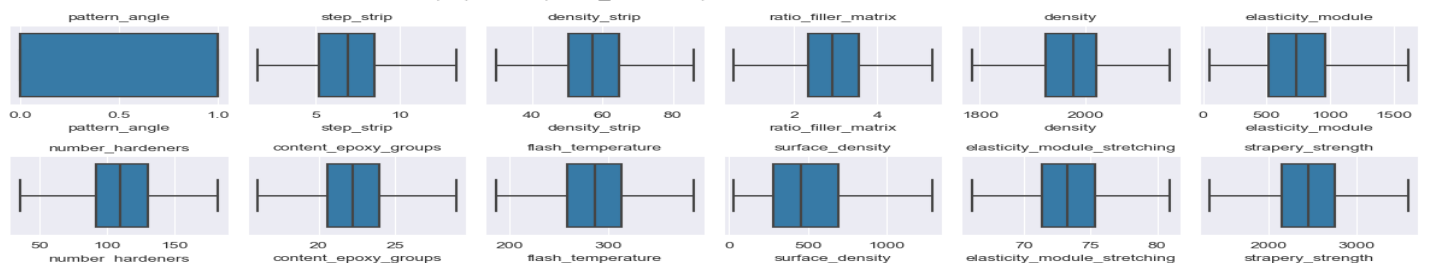
Графики boxplot df_main выбросов ДО очистки



Графики boxplot df_clean выбросов после очистки IQR



Графики boxplot df_clean выбросов ПОСЛЕ ПОЛНОЙ очистки



До начала мероприятий по очистке
1000 строк

Этап замены подозрительных значений на среднее 5 соседей
110 строк

Этап очистки методом IQR
(55 строк удалено)
Осталось **945 строк**

Этап очистки СЛЕВА и СПРАВА
(10 строк удалено)
Осталось **935 строк**
Чистый датасет!



**ОБРАЗОВАТЕЛЬНЫЙ
ЦЕНТР** МГТУ им. Н. Э. Баумана

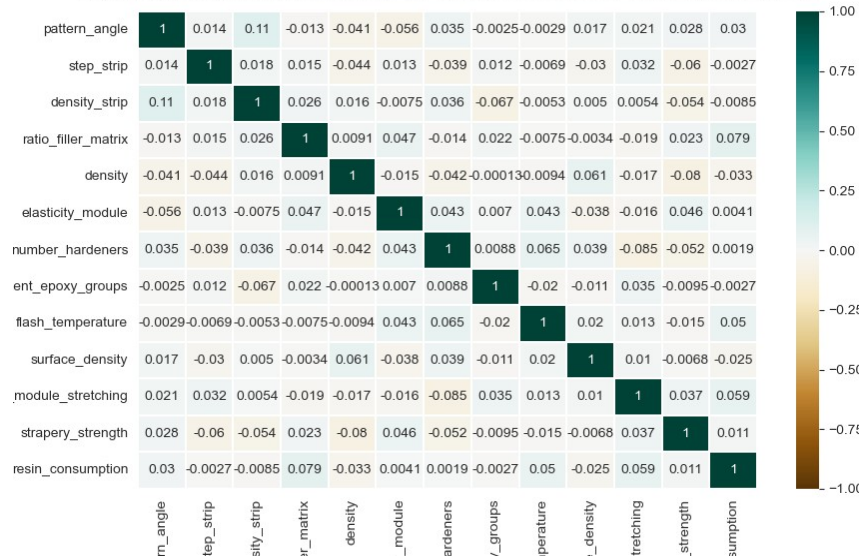
БЛОК № 2

Корреляционный анализ:

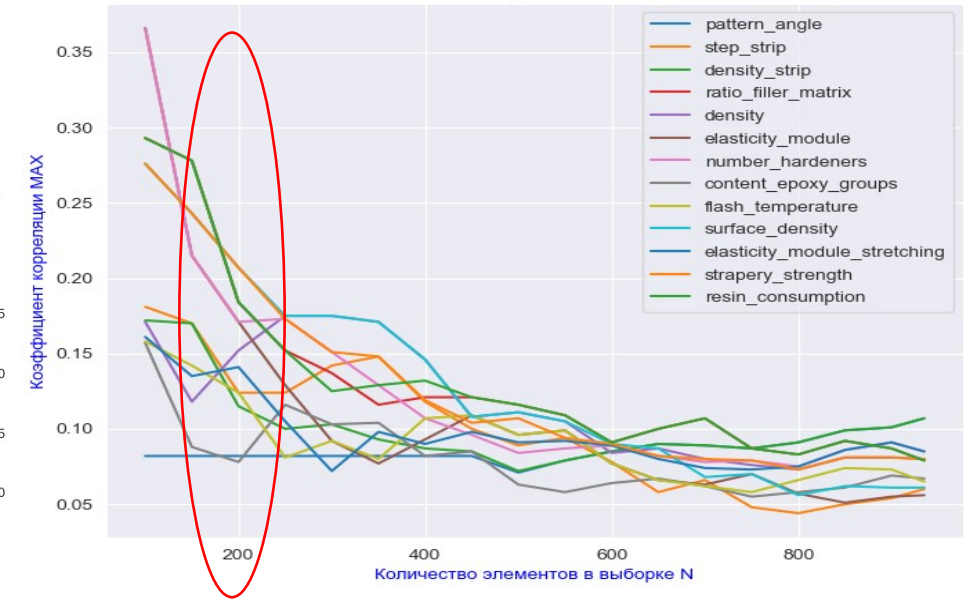
✓ Исследование падения корреляции о числа выборки:

- Построили корреляционную матрицу датасета - коэффициенты близки к 0;
- Создали массив изменения MAX коэф. корреляции в зависимости от N числа выборки;
- Построим графики, как общий, так и для каждого параметра отдельно – Коэффициенты стремительно падают с увеличением N;
- При N =200 еще сохраняется приемлемый уровень коэф.
- Выбрали фитчи для целевых переменных

Корреляционная матрица очищенного DataSet data_main_clean



Графики изменения MAX размера корреляции от количества N строк выборки

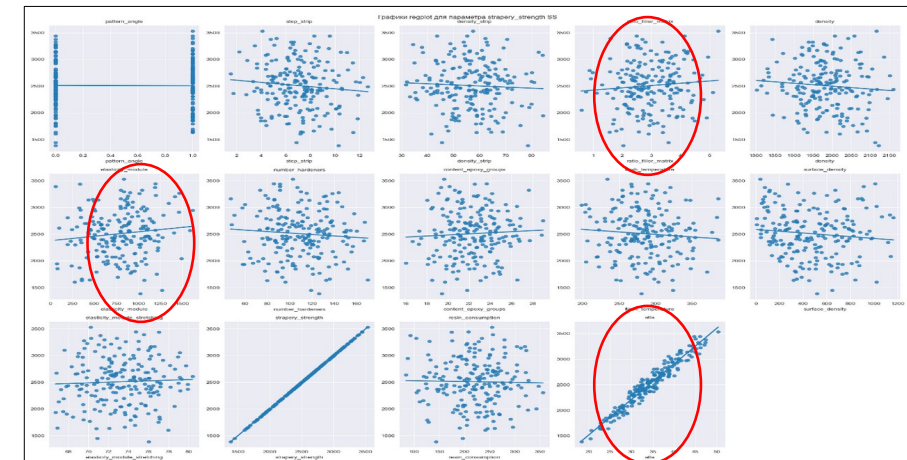


Выбор параметров (фитчи) для целевых переменных

'elasticity_module_stretching_EMS'



'strapery_strength_SS'



'elasticity_module', 'content_epoxy_groups', 'alfa'

'ratio_filler_matrix', 'elasticity_module', 'alfa'



ОБРАЗОВАТЕЛЬНЫЙ
ЦЕНТР МГТУ им. Н. Э. Баумана

$$\sigma_p = f(\alpha * E),$$

Введем новый признак 'ALFA' - 'удельную энергию' $\alpha = \sigma_p / E$

Новый параметр:

В соответствии с теорией хрупкого разрушения (теория Гриффита) [29] прочность (σ_p) определяется удельной энергией (α) вновь образованной поверхности разрушения: $(\sigma_p) = f(\alpha * E)$, где (σ_p) это наш параметр 'Прочность при растяжении' - 'strapery_strength', а E это 'Модуль упругости при растяжении' - 'elasticity_module_stretching'.

Учитывая такую физическую связь между параметрами, введем новый признак ALFA- 'удельную энергию' (α) = σ_p / E в наш новый датасет 'df_add_col'.

Создадим для N=200 DF_TRAIN и для N=100 DF_TEST

Корреляция для N =200 еще есть!



Датафрейм df_add_col:

Тип данных: float64

Количество параметров: 14

Количество записей: 1000



```
df_Train_lineReg = df_add_col.sample  
(n=200, random_state=42)
```

```
df_Test_lineReg = df_add_col.sample  
(n=100, random_state=50)
```

```
# Максимальные корреляции в нашем df для полного  
print(data_main_clean.corr().abs().apply(lambda
```

pattern_angle	0.107487
step_strip	0.060461
density_strip	0.107487
ratio_filler_matrix	0.078895
density	0.080297
elasticity_module	0.056466
number_hardeners	0.084787
content_epoxy_groups	0.066516
flash_temperature	0.065471
surface_density	0.060695
elasticity_module_stretching	0.084787
strapery_strength	0.080297
resin_consumption	0.078895
dtype: float64	

```
# Максимальные корреляции в нашем df для СЛУЧАЙНЫХ 200 строк  
print(df_add_col.sample(n=200, random_state=42).corr().abs().
```

pattern_angle	0.148352
step_strip	0.098215
density_strip	0.148352
ratio_filler_matrix	0.223873
density	0.092884
elasticity_module	0.153614
number_hardeners	0.080614
content_epoxy_groups	0.132030
flash_temperature	0.097019
surface_density	0.153614
elasticity_module_stretching	0.190468
strapery_strength	0.971485
resin_consumption	0.223873
alfa	0.971485
dtype: float64	

БЛОК № 3

Регрессионный анализ Результаты тестирования моделей

✓ Числовые показатели:

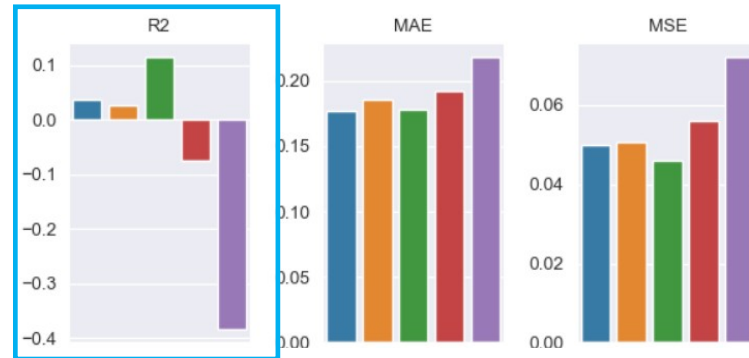
- Проведены исследования для двух целевых переменных `elasticity_module_stretching_EMS` и `'strapery_strength_SS'` по 6 различным моделям;
- Для целевой переменной `'elasticity_module_stretching_EMS'` полученные значения говорят о полном отсутствии возможности прогнозирования этой переменной на основе использованных моделей;
- Для целевой переменной `'strapery_strength_SS'` полученные значения говорят о ВОЗМОЖНОМ прогнозировании этой переменной на основе использованных моделей, $R^2_{rfr} = 0.944$!!!!!!!;



**ОБРАЗОВАТЕЛЬНЫЙ
ЦЕНТР** МГУ им. Н. Э. Баумана

'elasticity_module_stretching_EMS'

Графики изменения размера оценочного параметра EMS от типа мс



```
# OLS - ordinary least squares - метод наименьших квадратов
model_sm = sm.OLS(y_Train_EMS, X_train_EMS_add).fit()
```

`model_sm.params`

```
const          0.548250
elasticity_module -0.068939
content_epoxy_groups 0.159847
alfa           -0.219414
dtype: float64
```

`model_sm.summary()`

OLS Regression Results

Dep. Variable:	elasticity_module_stretching	R-squared:	0.058
Model:	OLS	Adj. R-squared:	0.044

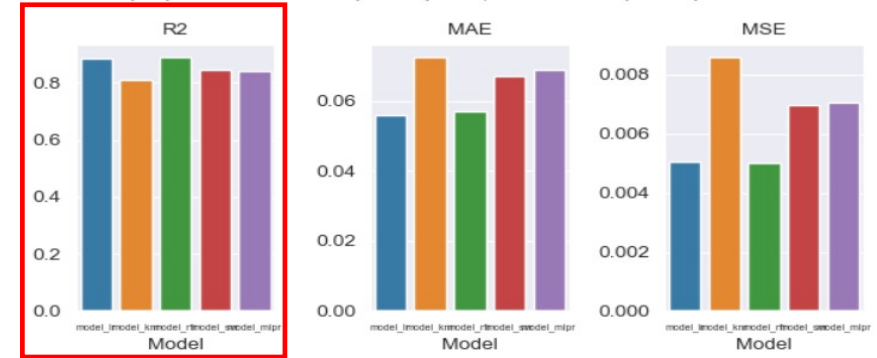
Худшая mlpr $R^2 = -0.383$

Лучшая rfr $R^2 = 0.114$

sm.OLS_EMS $R^2 = 0.058$

'strapery_strength_SS'

Графики изменения размера оценочного параметра SS от типа м



```
# OLS - ordinary least squares - метод наименьших кв
model_sm = sm.OLS(y_Train_SS, X_train_SS_add).fit()
```

`model_sm.params`

```
const          0.011492
ratio_filler_matrix 0.001450
elasticity_module -0.021967
alfa           1.043117
dtype: float64
```

`model_sm.summary()`

OLS Regression Results

Dep. Variable:	strapery_strength	R-squared:	0.944
Model:	OLS	Adj. R-squared:	0.943

Худшая knr $R^2 = 0.808$

Лучшая rfr $R^2 = 0.888$

sm.OLS_SS $R^2 = 0.944!!!$

Создание нейронной сети

✓ Обучение модели:

- Загрузили очищенный датасет из Блока 1 из 13 параметров
- Обосновали выбор количества слоев и провели расчет количества нейронов в каждом слое;
- Выбрали активационную функцию SELU;
- Выбрали оптимизатор ADAM;
- Выбрали оценочный параметр MAE;
- Создали индикатор переобучения модели;
- Модель прервала работу по условию индикатора переобучения на 35 эпохе;
- Лучшие показатели достигнуты на 9 эпохе;
- MAE модели составило 0.1634.

Создаем нейросеть из 4 слоев:

1-й - Входной слой размерность $n = 12$ - по количеству параметров в нашем df
 # 2-й - Скрытый слой размерностью $k1 = 6$ - получен из расчета по правилу геомет
 # 3-й - Скрытый слой размерностью $k2 = 3$ - получен из расчета по правилу геомет
 # 4-й - Выходной слой размерностью $m = 1$ - по количеству целевых параметров

```
model_RFM.add(Dense(12)) # входной полносвязный слой
model_RFM.add(Dense(6, kernel_initializer='lecun_normal', activation='selu'))
model_RFM.add(AlphaDropout(0.25))
model_RFM.add(Dense(3, kernel_initializer='lecun_normal', activation='selu'))
model_RFM.add(AlphaDropout(0.25))
model_RFM.add(Dense(1, activation='linear')) # выходной слой
```

```
[33]: # Оценка модели по лучшему расчетному показателю MAE
evaluation_model = model_RFM.evaluate(x= X_test_RFM,
                                      y=y_test_RFM)
print("Оценка модели = ", evaluation_model)
```

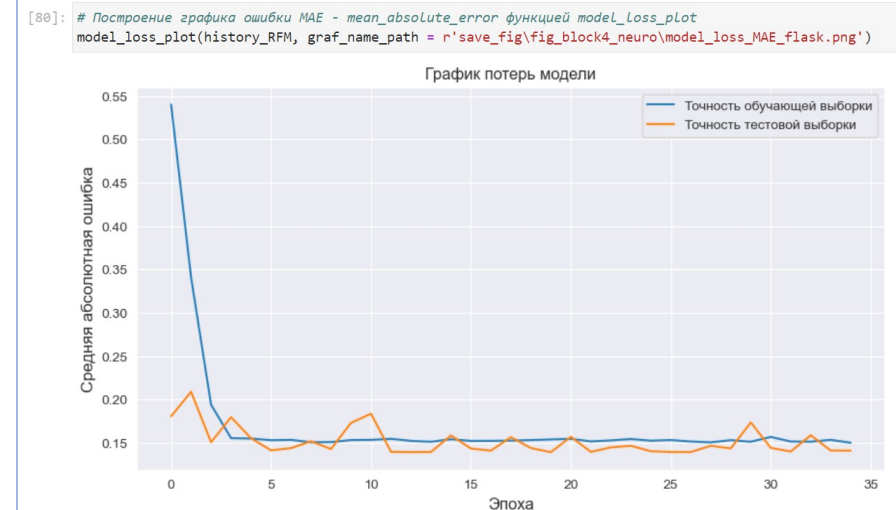
```
9/9 [=====] - 0s 1ms/step - loss: 0.1634
```

Оценка модели = 0.1633787751197815

```
[ ]: # Лучший показатель MAE был достигнут во время 9 эпохи и составил MAE(9) = 0.1633
```

```
[32]: %%time
history_RFM = model_RFM.fit(
    X_train_RFM,      # входящая выборка
    y_train_RFM,      # целевая выборка
    batch_size = 32,  # вычисляем градиенты каждые 32 наблюдений
    epochs=100,        # max 100 эпох, stop_RFM у нас создан
    verbose=1,         # индикатор выполнения
    validation_split = 0.2, # доля обучающих данных, для проверки
    callbacks = [stop_RFM] # условие остановки при прекращении обучения в течении 10 эпох
)
```

```
17/17 [=====] - 0s 3ms/step - loss: 0.1506 - val_loss: 0.1467
Epoch 29/100
17/17 [=====] - 0s 3ms/step - loss: 0.1532 - val_loss: 0.1436
Epoch 30/100
17/17 [=====] - 0s 3ms/step - loss: 0.1514 - val_loss: 0.1736
Epoch 31/100
17/17 [=====] - 0s 3ms/step - loss: 0.1569 - val_loss: 0.1443
Epoch 32/100
17/17 [=====] - 0s 3ms/step - loss: 0.1516 - val_loss: 0.1401
Epoch 33/100
17/17 [=====] - 0s 3ms/step - loss: 0.1514 - val_loss: 0.1588
Epoch 34/100
17/17 [=====] - 0s 3ms/step - loss: 0.1535 - val_loss: 0.1412
Epoch 35/100
17/17 [=====] - 0s 3ms/step - loss: 0.1502 - val_loss: 0.1411
Epoch 35: early stopping
Wall time: 2.81 s
```



БЛОК № 5

✓ Пользовательское приложение

- Сохранил модель нейронной сети для разработки веб-приложения для прогнозирования соотношения "матрица-наполнитель" в фреймворке Flask;
- При запуске приложения, пользователь переходит на: <http://127.0.0.1:5000/>;
- В открывшемся окне пользователю необходимо ввести в соответствующие ячейки значения в соответствии с рекомендованными и нажать на кнопку «Рассчитать значение».
- В случае ошибки — нажать кнопку «Очистить форму»;
- На выходе пользователь получает результат прогноза для значения параметра «Соотношение «матрица – наполнитель»».
- Приложение успешно работает

✓ Репозиторий на github.com

- https://github.com/Grain1963/MHTS_Diploma



**ОБРАЗОВАТЕЛЬНЫЙ
ЦЕНТР** МГТУ им. Н. Э. Баумана

МHTS_Diploma/translate_list... Дзен Diploma_bloc... - JupyterLab 127.0.0.1:5000

Программа расчета
прогнозного значения параметра
«Соотношение матрица-наполнитель»

Заполните ячейки и нажмите "РАССЧИТАТЬ ЗНАЧЕНИЕ" или "ОЧИСТИТЬ ФОРМУ"

Название параметра	Рекомендованное значение ОТ_ДО	Значение параметра
Угол нашивки, град	0.0 или 1.0	Заполните значение
Шаг нашивки	1.6 - 13.0	Заполните значение
Плотность нашивки	30.0 - 80.0	Заполните значение
Плотность, кг/м3	1800.0 - 2150.0	Заполните значение
Модуль упругости, ГПа	55.0 - 1550.0	Заполните значение
Количество отверстий, м.5%	40.0 - 180.0	Заполните значение
Содержание эпоксидных групп, %_2	17.0 - 28.0	Заполните значение
Температура вспеники, С_2	190.0 - 380.0	Заполните значение
Поверхностная плотность, г/м2	30.0 - 1200.0	Заполните значение
Модуль упругости при растяжении, ГПа	66.0 - 81.0	Заполните значение
Прочность при растяжении, МПа	1330.0 - 3600.0	Заполните значение
Потребление смолы, г/м2	80.0 - 350.0	Заполните значение

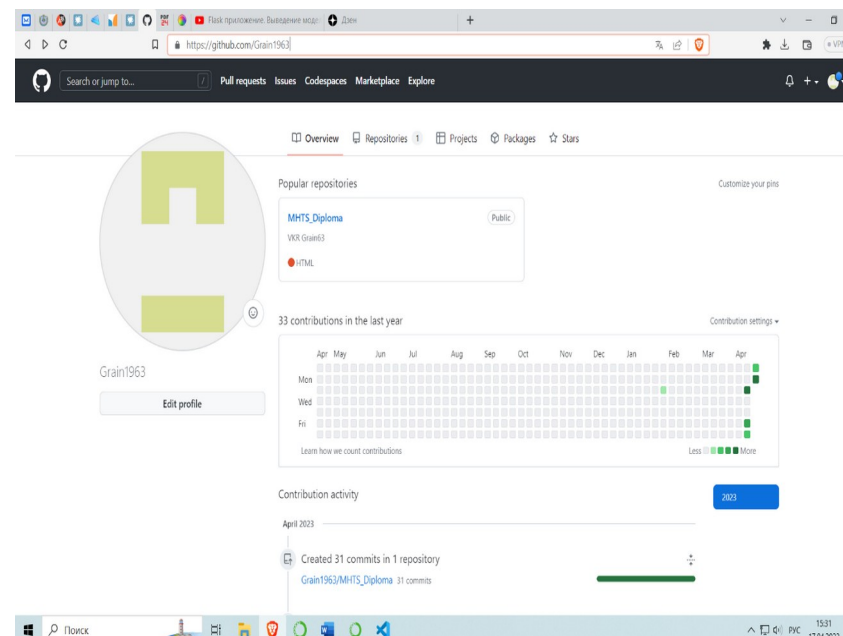
РАССЧИТАТЬ ЗНАЧЕНИЕ

ОЧИСТИТЬ ФОРМУ

Результат прогноза:

[[2.741078]]

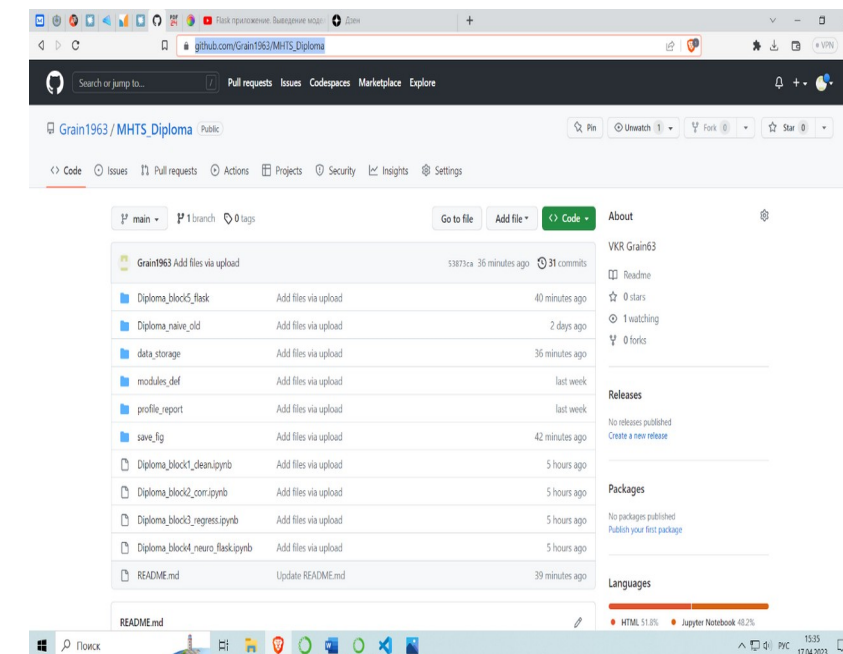
ельного центра МГТУ им. Н.Э. Баумана©, апрель 2023 г.



Файл Правка Выделение Вид Переход Выполнить ... info.py - Diploma_block5_flask - Visual S

main.html app.py info.py 9+ X

```
info.py > ...
1
2 PS C:\Users\grain\Work_folder\MGTU_Diplom\Diploma_block5_flask> &
3 C:/Users/grain/anaconda3/envs/MyENV_JupitePS
4 C:\Users\grain\Work_folder\MGTU_Diplom\Diploma_block5_flask> &
5 C:/Users/grain/anaconda3/envs/MyENV_JupiterLab/python.exe
6 c:/Users/grain/Work_folder/MGTU_Diplom/Diploma_block5_flask/app.py
7 * Serving Flask app 'app'
8 * Debug mode: off
9 WARNING: This is a development server. Do not use it in a production deployment
10 Use a production WSGI server instead.
11 * Running on http://127.0.0.1:5000
12 Press CTRL+C to quit
13 127.0.0.1 - - [17/Apr/2023 14:41:32] "GET / HTTP/1.1" 200 -
```





ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана



do.bmstu.ru