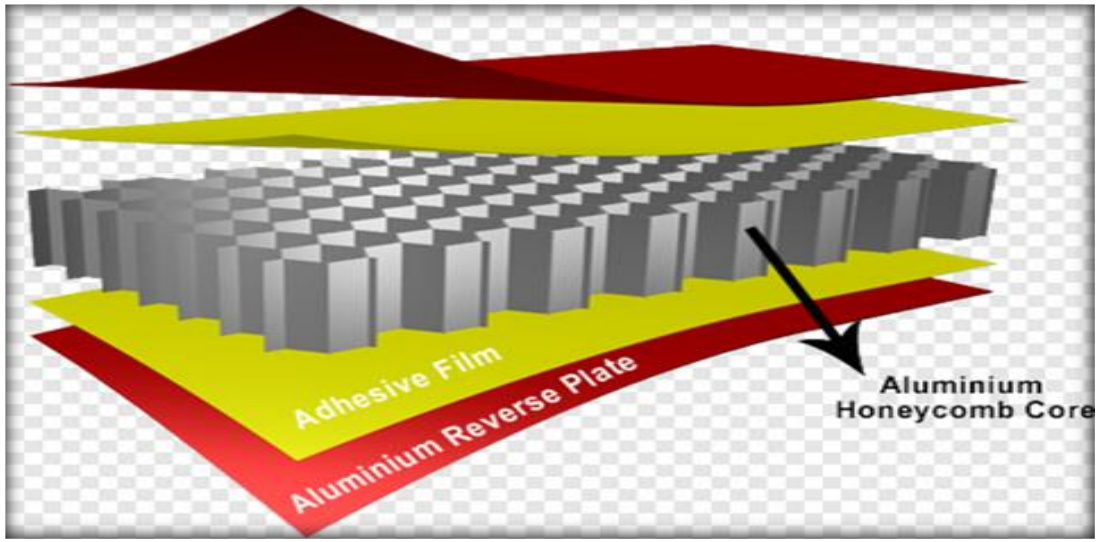
ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу «Data Science»

ТЕМА:
«Прогнозирование конечных свойств новых материалов (композиционных материалов)»



Слушатель Стоянов Павел Александрович

Поставленные задачи в рамках ВКР:

- Дать характеристику датасету
- Описать теоретические основы используемых методов изучения данных
- Провести разведочный анализ данных
- Выполнить предобработку данных
- Разработать и обучить несколько моделей для прогноза целевых признаков «Модуль упругости при растяжении» и «Прочность при растяжении»
- Написать нейронную сеть для прогноза «Соотношение матрица-наполнитель»
- Разработать приложение

Аналитическая часть

Описать теоретические основы используемых методов изучения данных

Модели для анализа:

--Linear Regression

--Lasso

--RandomForestRegressor

--KNeighborsRegressor

--SVR

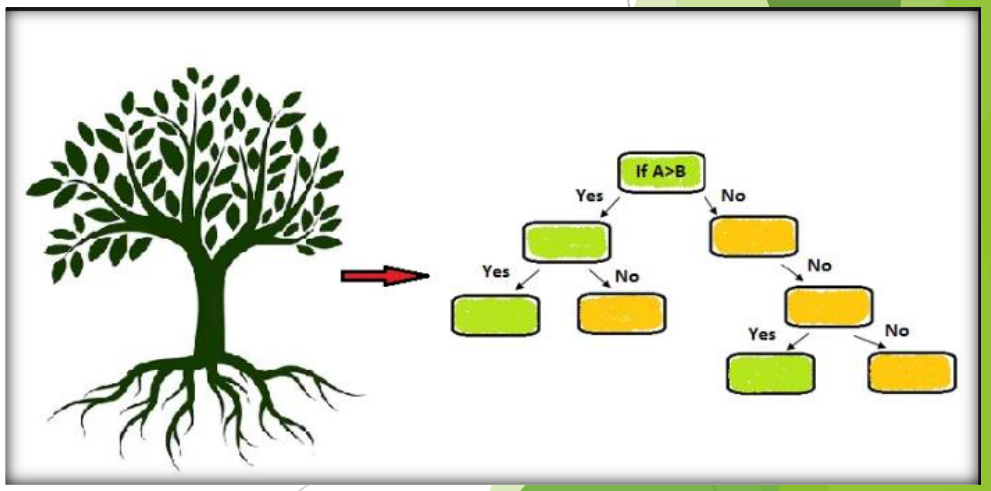
-DecisionTreeRegressor

--AdaBoostRegressor

--Нейронная сеть (Sequential)

Метрики для сравнения моделей и оценки точности их работы:

- MAE—абсолютная ошибка
- MSE--среднеквадратичная ошибка
- MAPE--средняя абсолютная ошибка в процентах
- *RMSE*—корень из среднеквадратичной ошибки
- *R*2—коэффициент детерминации



Характеристика исходных данных

- На входе имеется набор данных (файлы X_bp.xlsx и X_nup.xlsx) с начальными свойствами компонентов композиционных материалов.
- Объединение делается по индексу, тип объединения INNER

X_bp.head(5)

	Unnamed: 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
0	0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0
1	1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0
2	2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0
3	3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0
4	4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0

Файл X_bp.xlsx содержит 1023 записи и одиннадцать столбцов с признаками

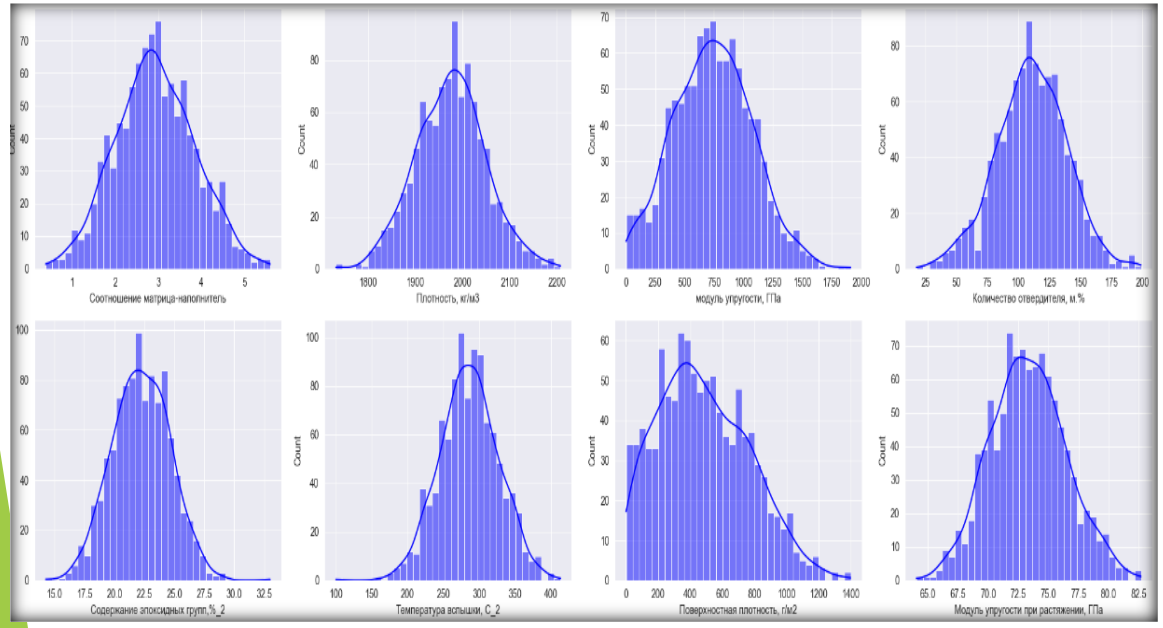
X_nup.head(5)

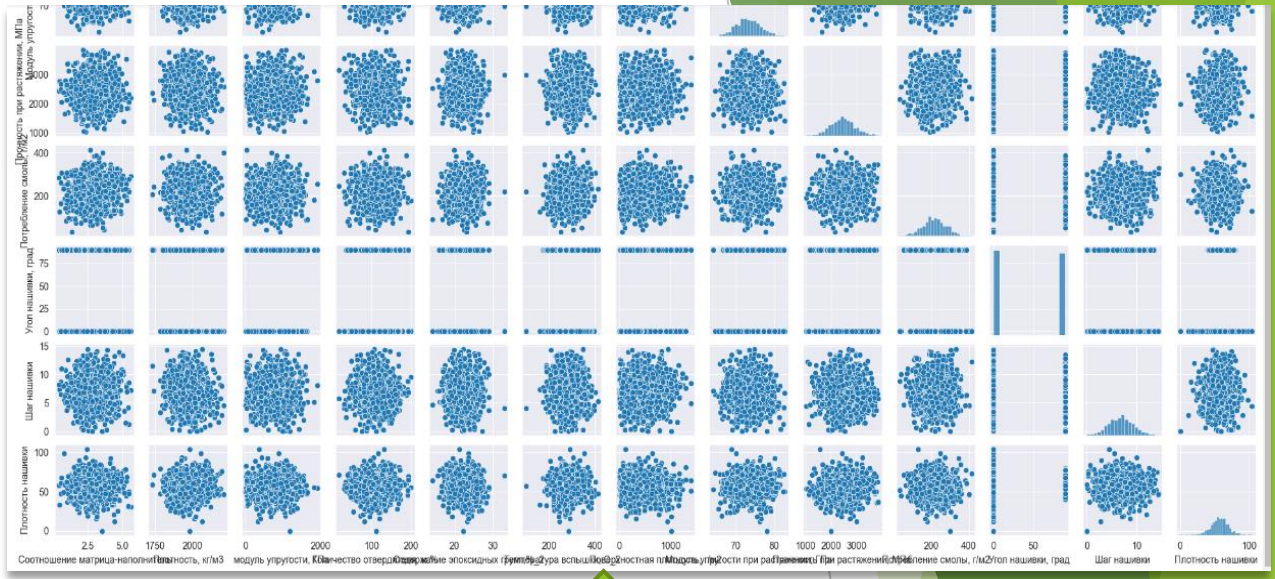
	Unnamed: 0	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	0	0	4.0	57.0
1	1	0	4.0	60.0
2	2	0	4.0	70.0
3	3	0	5.0	47.0
4	4	0	5.0	57.0

Файл X_nup.xlsx содержит 1040 записей и четыре столбца с признаками

EDA: Визуализация данных

Гистограммы распределения переменных





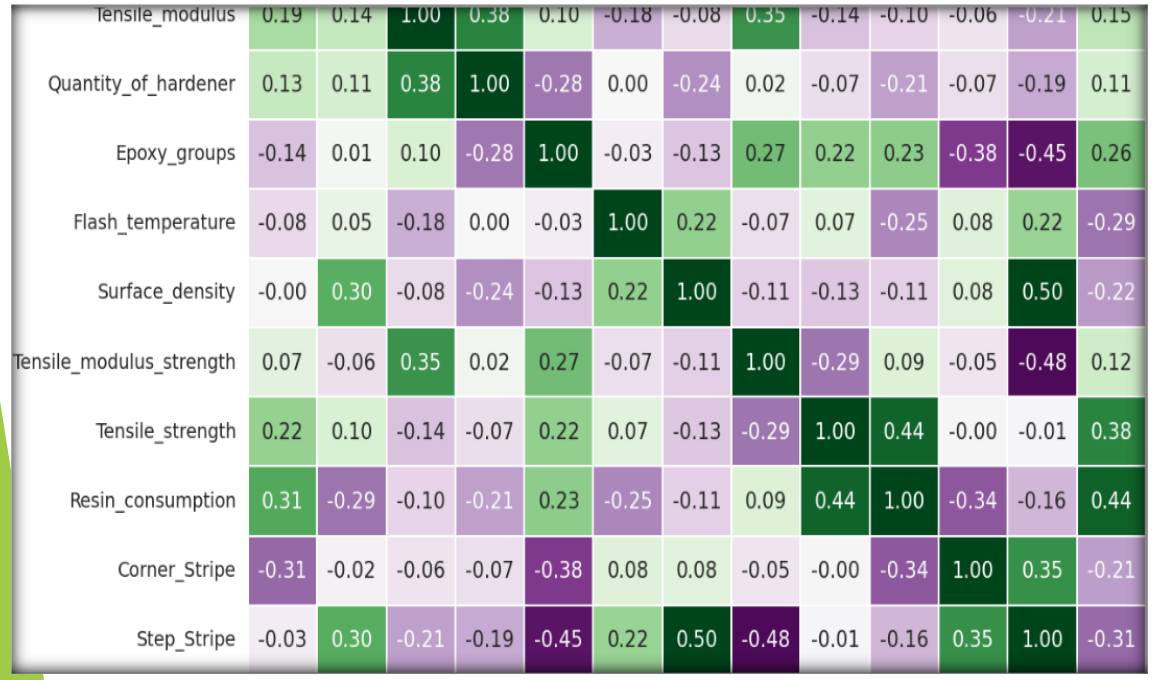
Попарные графики рассеяния точек (скаттерплоты)

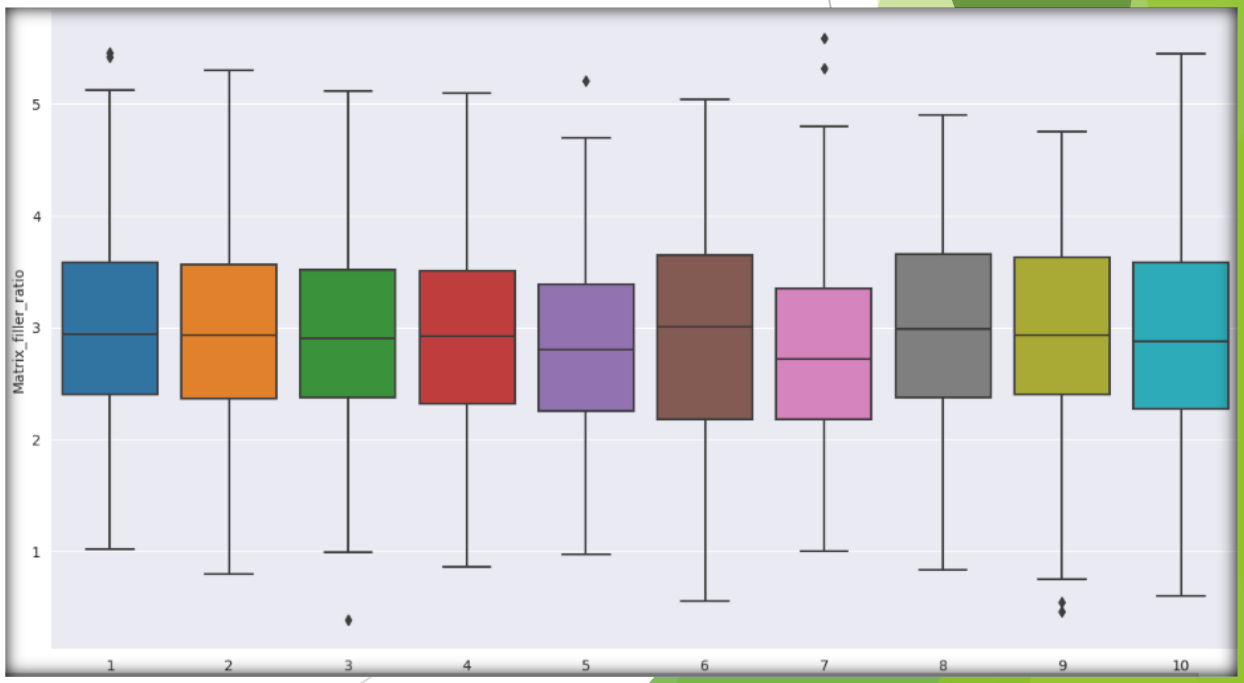
EDA: Оценка зависимостей между признаками

--корреляция признаков в датасете очень низкая

--выборки из датасета «Топ 25 наивысших значений целевых признаков»

--деление на группы (по три группы, по десять групп) значений целевых признаков





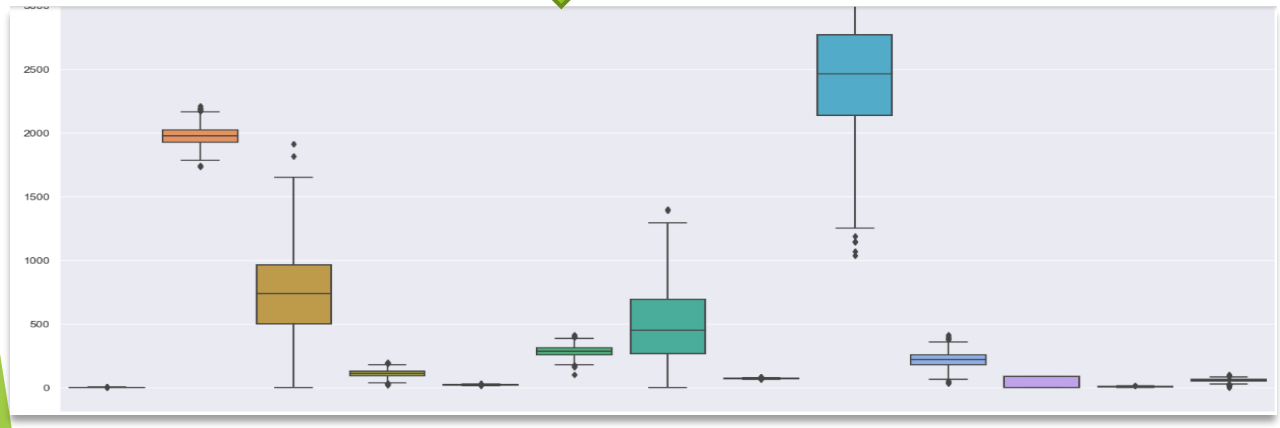
EDA: Идентификация наличия выбросов

(выбросы неоднозначны, соответственно нужно рассматривать варианты дальнейшей работы с удалением и без удаления выбросов)

	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

Boxplot «Ящик с усами»

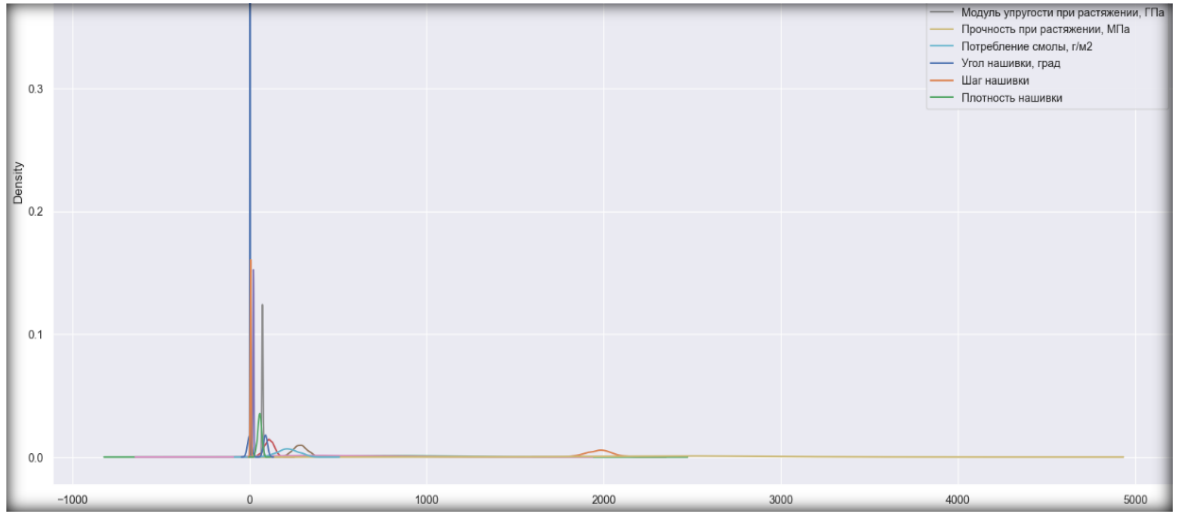
(выбросов мало, они компактны, хвосты очень короткие)



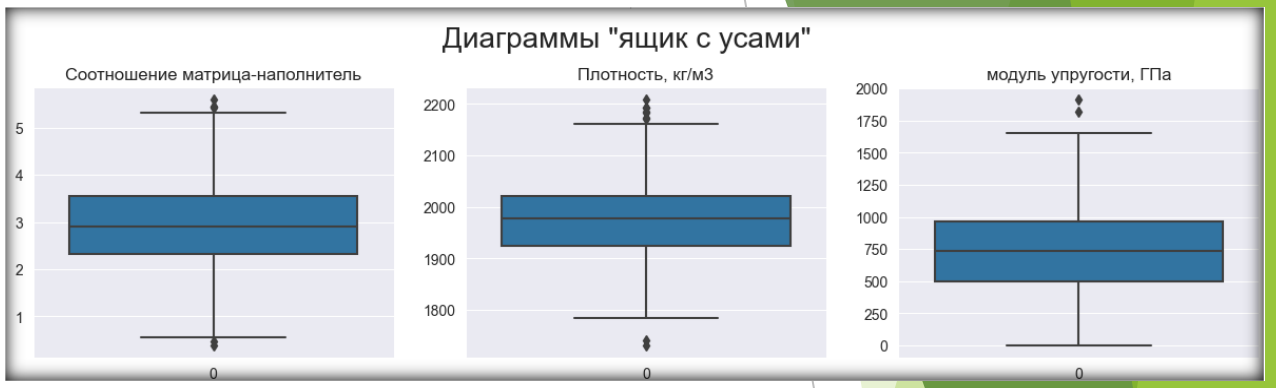
Медиана, среднее значение переменных

(значения очень близки, отклонение только по двум признакам)

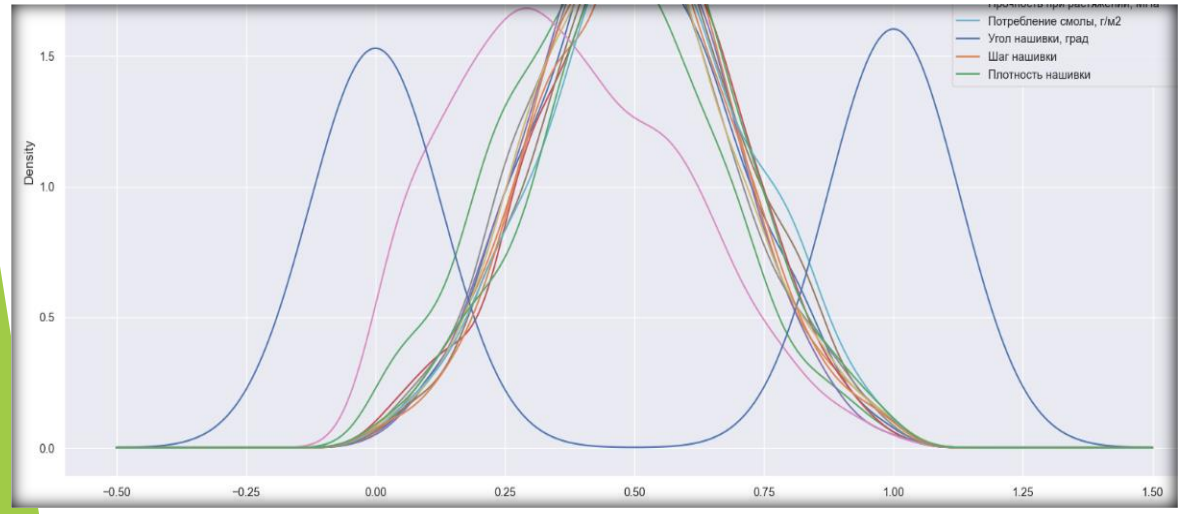
Предобработка данных



Оценка плотности ядра до нормализации



Boxplot до нормализации

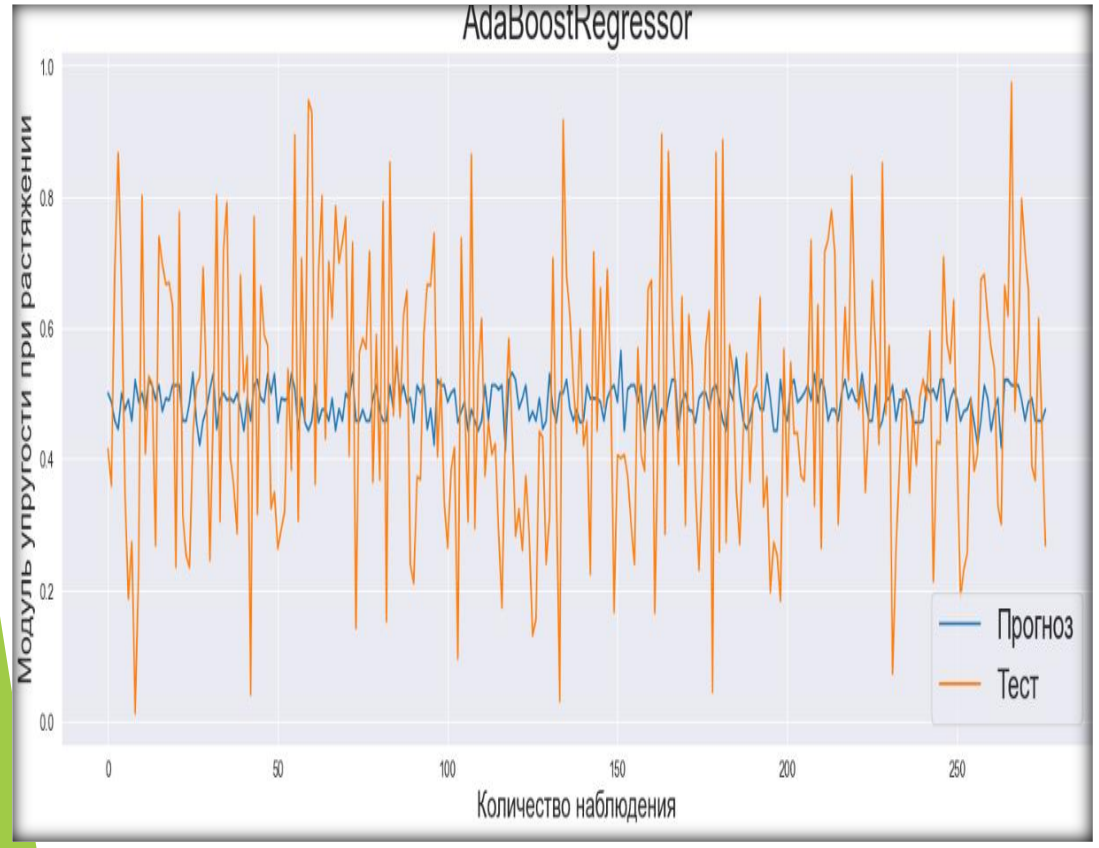


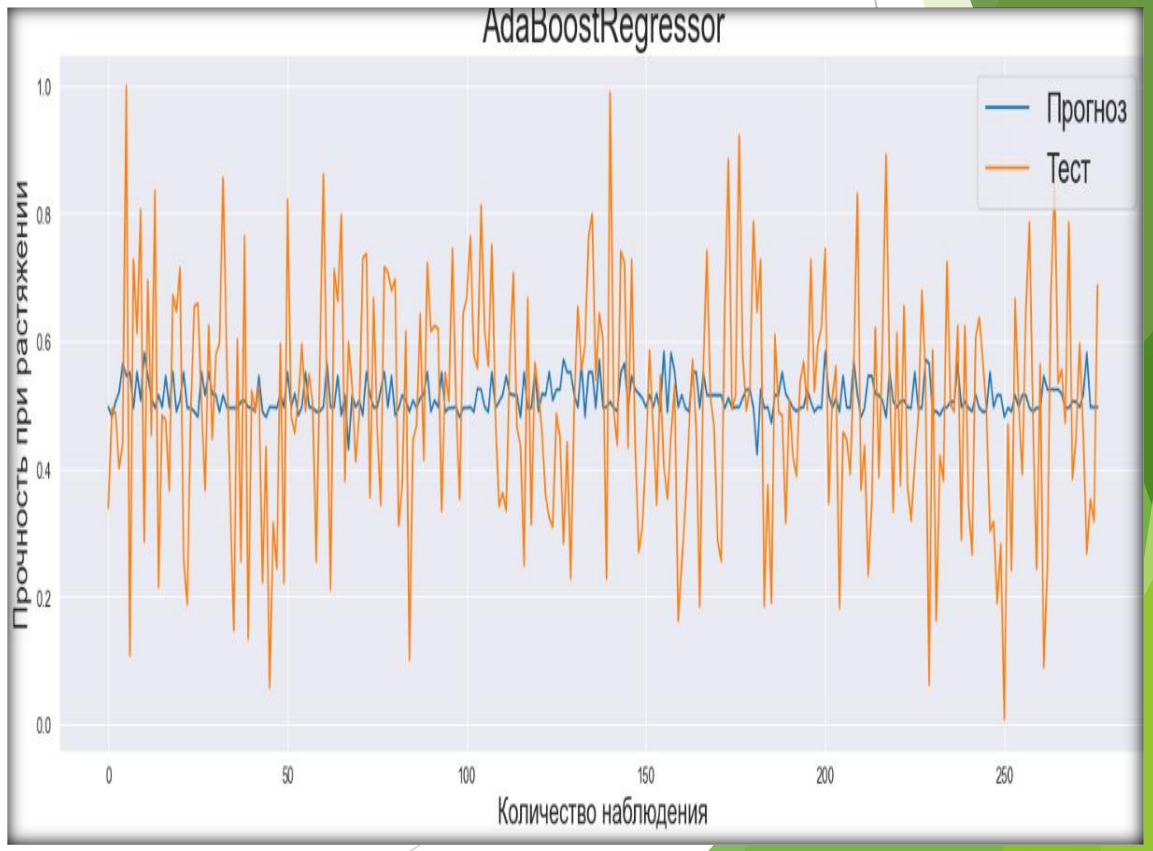
Оценка плотности ядра после нормализации

Boxplot после нормализации

Разработка и обучение моделей

- Разделение на обучающую тестовую выборки (по условию 30% данных оставить на тестирование моделей, на остальных провести обучение моделей)
- Поиск гиперпараметров моделей с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10 (метод GridSearchCV).
- Подстановка гиперпараметров в модель, обучение на тренировочных данных





Оценка точности моделей по метрикам

Оценка точности моделей по метрикам

«Модуль упругости при растяжении»

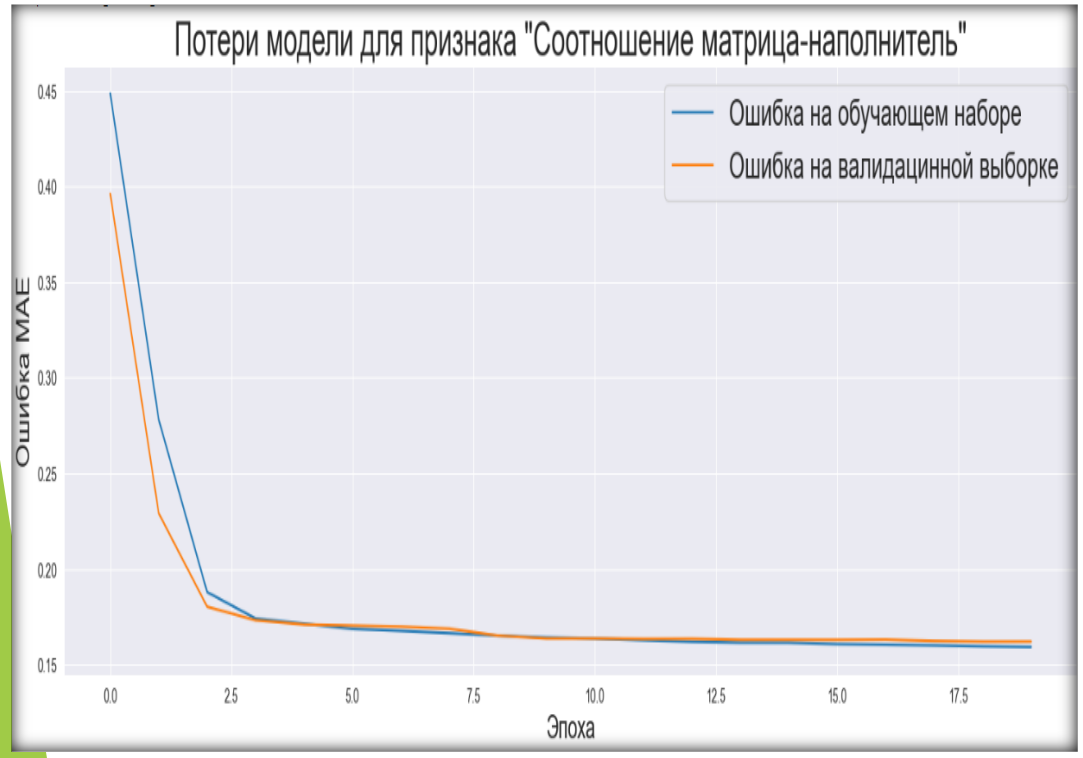
Модели	MAE (абсолютная ошибка)		MSE (среднеквадратичная ошибка)		MAPE (средняя абсолютная ошибка в процентах)		RMSE (корень из среднеквадратичной ошибки)		R-квадрат (коэффициент детерминации)		SCORE	
	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)	стандартные параметры	гиперпараметры GridSearchCV (10)
LinearRegression	0.165	0.165	0.039	0.039	0.697	0.697	0.198	0.198	-0.014	-0.014	0.019	0.019
Lasso	0.163	0.163	0.039	0.039	0.699	0.699	0.197	0.197	-0.000	-0.000	0.000	0.000
RandomForestRegres	0.163	0.163	0.039	0.039	0.692	0.695	0.198	0.197	-0.013	-0.001	0.847	0.018
KNeighborsRegresso	0.179	0.165	0.047	0.039	0.741	0.700	0.217	0.198	-0.215	-0.008	0.171	0.011
SVR	0.177	0.163	0.047	0.039	0.735	0.693	0.216	0.197	-0.203	-0.001	0.451	0.002
DecisionTreeRegress	0.227	0.164	0.081	0.039	0.800	0.705	0.285	0.198	-1.096	-0.010	1.000	0.010
AdaBoostRegressor	0.164	0.163	0.039	0.039	0.695	0.706	0.197	0.197	-0.002	0.001	0.094	0.043
Лучшее значение	0.163	0.163	0.039	0.039	0.692	0.693	0.197	0.197	-0.000	0.001	0.000	0.000

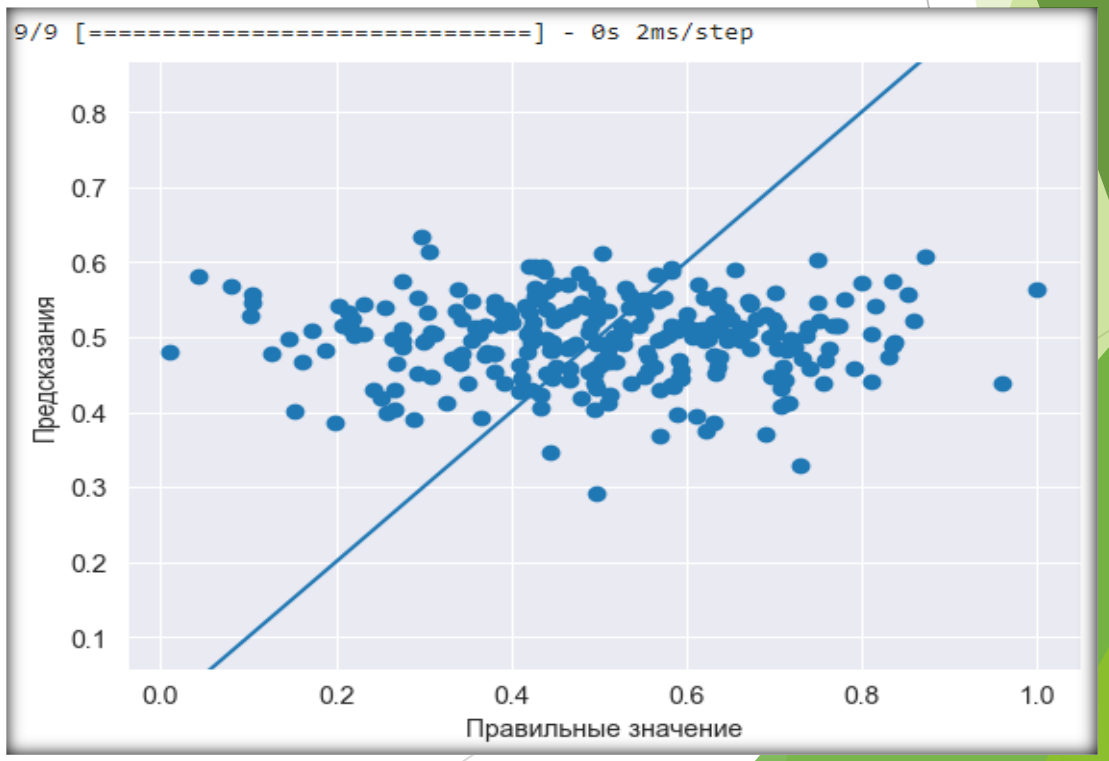
«Прочность при растяжении»

LinearRegression	0.147	0.147	0.034	0.034	0.661	0.661	0.184	0.184	0.009	0.009	-0.014	-0.014
Lasso	0.148	0.148	0.034	0.034	0.676	0.676	0.185	0.185	-0.003	-0.003	-0.009	-0.009
RandomForestRegres	0.150	0.147	0.035	0.034	0.651	0.666	0.187	0.184	-0.025	0.005	-0.417	-0.011
KNeighborsRegresso	0.164	0.147	0.041	0.034	0.725	0.678	0.202	0.185	-0.199	-0.002	-0.231	-0.006
SVR	0.159	0.147	0.040	0.034	0.726	0.669	0.199	0.184	-0.164	-0.000	-0.271	-0.005
DecisionTreeRegress	0.217	0.148	0.073	0.034	0.838	0.640	0.270	0.184	-1.147	0.001	-0.964	-0.014
AdaBoostRegressor	0.148	0.147	0.034	0.034	0.673	0.641	0.183	0.184	0.010	0.008	-0.054	-0.013
Лучшее значение	0.147	0.147	0.034	0.034	0.651	0.640	0.183	0.184	0.010	0.009	-0.009	-0.005

Нейронная сеть, рекомендующая «соотношение матрица-наполнитель»

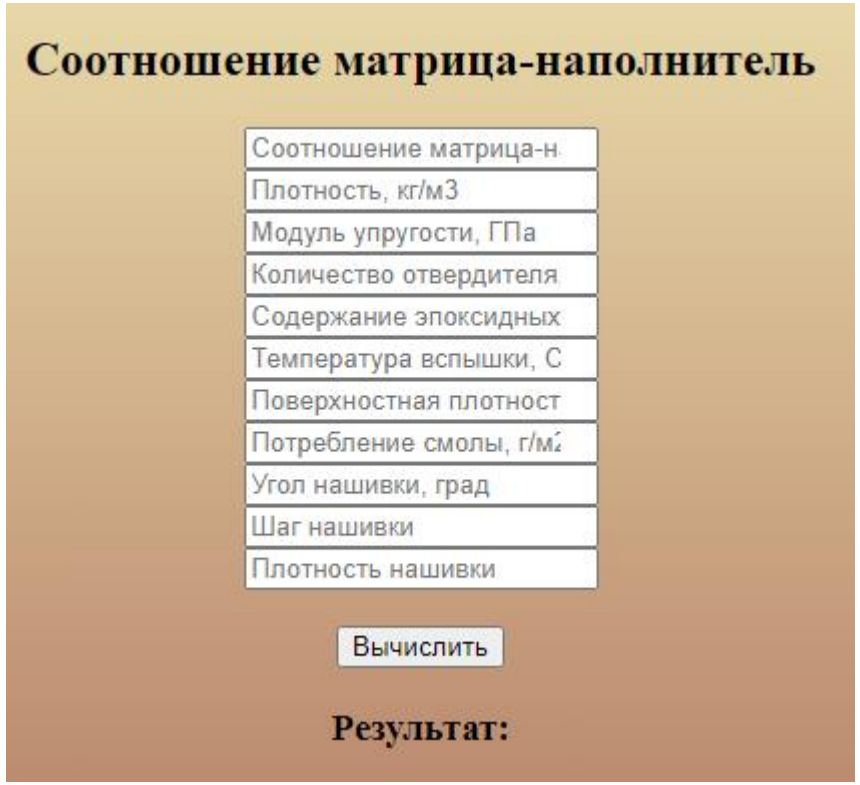
	Модель	MAE	MSE	R2
0	NS_train	0.152	0.035	-0.095
1	NS_test	0.159	0.040	-0.080

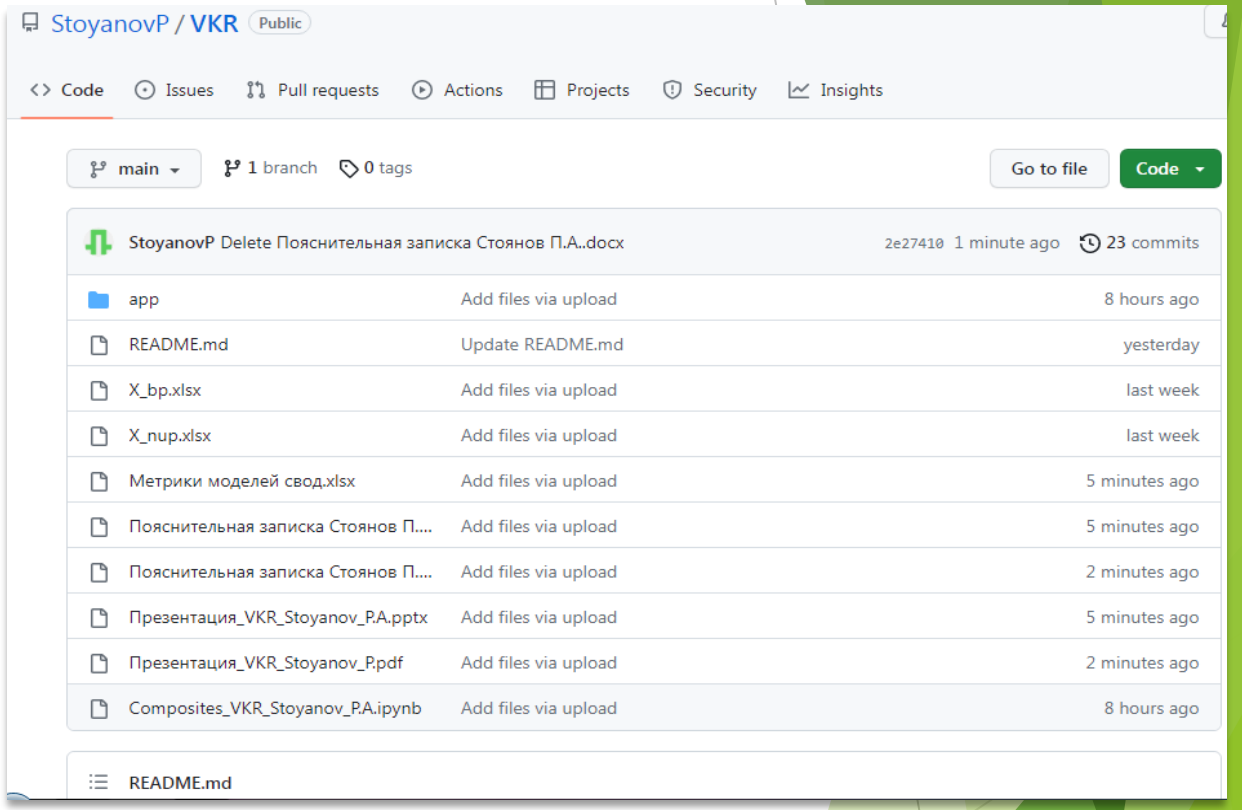




Репозиторий на GitHub

Приложение прогнозирующее «Модуль упругости при растяжении»



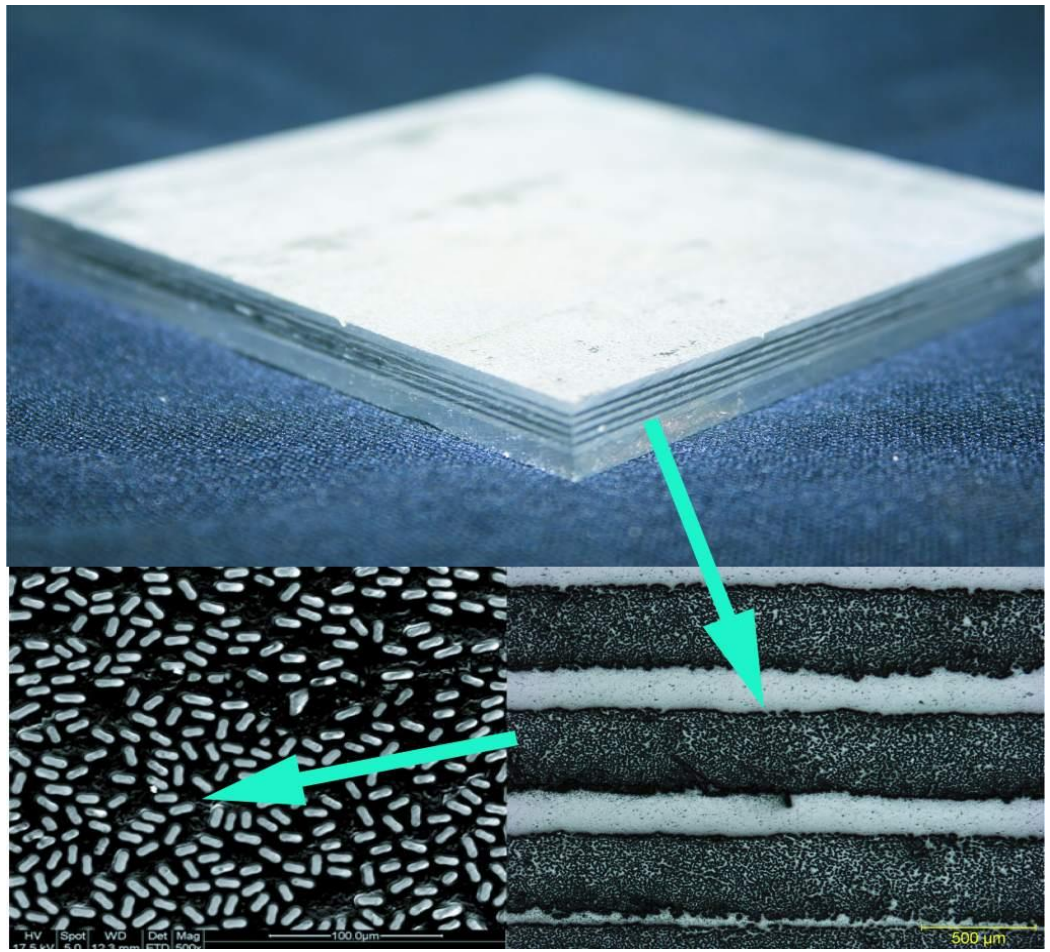


Выводы

- Изучен материал о композитах, проанализированы данные, отработаны гипотезы, отражены теоретические основы, построены модели прогноза, создана нейронная сеть, разработано приложение, создан **репозиторий** на GitHub.

- Качество работы моделей и нейронной сети неудовлетворительные. Без консультации с экспертом по свойствам композитов, без постановки более точной бизнес-задачи нет возможности правильно идентифицировать выбросы, оценить корреляцию, также нельзя выбрать и настроить подходящую модель (т.к. не можем отобрать нужные признаки, определить их вес и т.п.).

- Для достижения лучшего результата важно, с учетом полученных разъяснений от эксперта и бизнеса, определить новую оптимальную стратегию работы с данными и заново выполнить весь Pipeline.



СПАСИБО ЗА ВНИМАНИЕ