

Образовательный центр МГТУ им. Н.Э. Баумана

# **Выпускная квалификационная работа по курсу "Data Science"**

Слушатель: Комаров Николай

**Тема: Прогнозирование конечных свойств  
новых материалов (композиционных материалов)**

# Постановка задачи

- изучить предметную область
- провести разведочный анализ данных
- разделить данные на тренировочную и тестовую выборки
- выполнить препроцессинг (предобработку)
- выбрать базовую модель и модели для подбора
- сравнить модели с гиперпараметрами по умолчанию
- подобрать гиперпараметры с помощью поиска по сетке с перекрестной проверкой
- сравнить модели после подбора гиперпараметров и выбрать лучшую
- сравнить качество лучшей и базовой моделей на тестовой выборке
- сравнить качество лучшей модели на тренировочной и тестовой выборке
- разработать приложение

# Разведочный анализ данных

X\_br (матрица из базальтопластика):

- признаков: 10 и индекс
- строк: 1023

X\_pur (наполнитель из углепластика):

- признаков: 3 и индекс
- строк: 1040

Объединение с типом INNER по индексу, получилось:

- признаков: 13
- строк: 1023

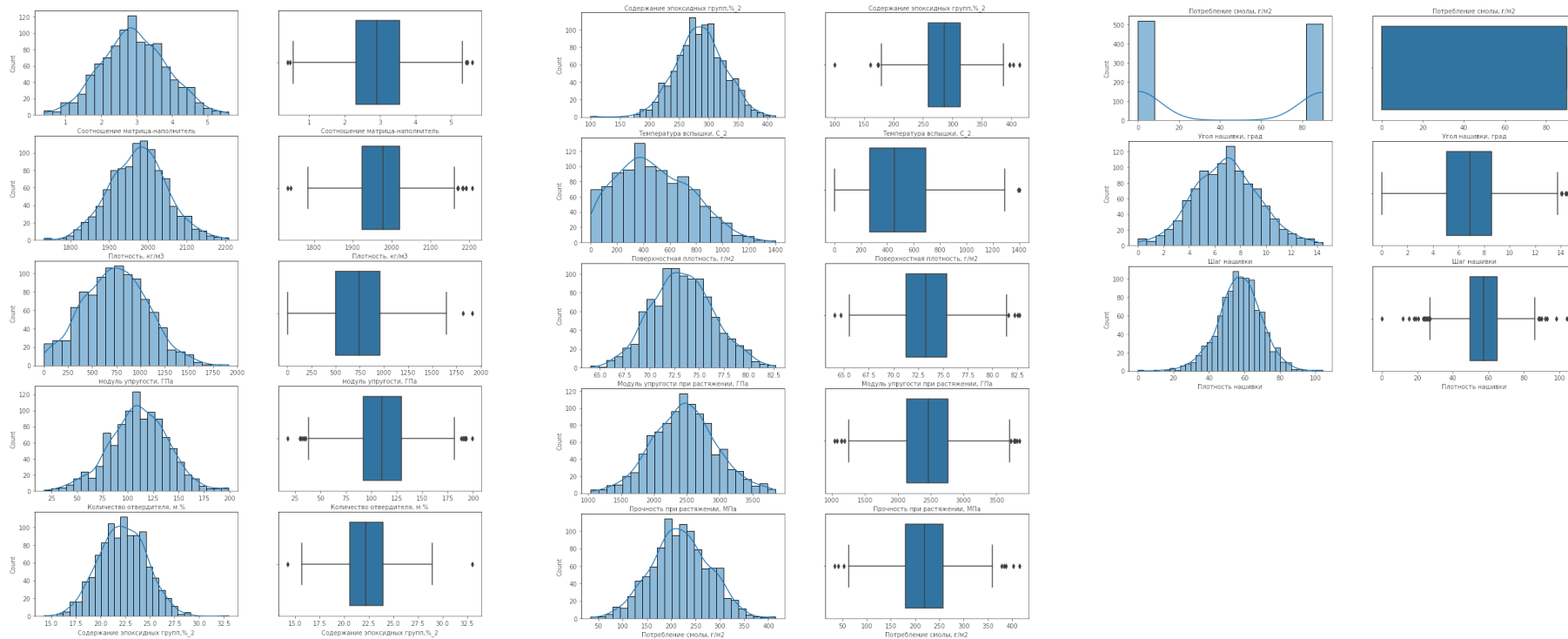
# Разведочный анализ данных

Название	Файл	Тип данных	Непустых значений	Уникальных значений
Соотношение матрица-наполнитель	X_bp	float64	1023	1014
Плотность, кг/м3	X_bp	float64	1023	1013
модуль упругости, ГПа	X_bp	float64	1023	1020
Количество отвердителя, м.%	X_bp	float64	1023	1005
Содержание эпоксидных групп, %_2	X_bp	float64	1023	1004
Температура вспышки, C_2	X_bp	float64	1023	1003
Поверхностная плотность, г/м2	X_bp	float64	1023	1004
Модуль упругости при растяжении, ГПа	X_bp	float64	1023	1004
Прочность при растяжении, МПа	X_bp	float64	1023	1004
Потребление смолы, г/м2	X_bp	float64	1023	1003
Угол нашивки, град	X_nup	float64	1023	2
Шаг нашивки	X_nup	float64	1023	989
Плотность нашивки	X_nup	float64	1023	988

	Среднее	Стандартное отклонение	Минимум	Максимум	Медиана
Соотношение матрица-наполнитель	2.9304	0.9132	0.3894	5.5917	2.9069
Плотность, кг/м3	1975.7349	73.7292	1731.7646	2207.7735	1977.6217
модуль упругости, ГПа	739.9232	330.2316	2.4369	1911.5365	739.6643
Количество отвердителя, м.%	110.5708	28.2959	17.7403	198.9532	110.5648
Содержание эпоксидных групп, %_2	22.2444	2.4063	14.2550	33.0000	22.2307
Температура вспышки, C_2	285.8822	40.9433	100.0000	413.2734	285.8968
Поверхностная плотность, г/м2	482.7318	281.3147	0.6037	1399.5424	451.8644
Модуль упругости при растяжении, ГПа	73.3286	3.1190	64.0541	82.6821	73.2688
Прочность при растяжении, МПа	2466.9228	485.6280	1036.8566	3848.4367	2459.5245
Потребление смолы, г/м2	218.4231	59.7359	33.8030	414.5906	219.1989
Угол нашивки, град	44.2522	45.0158	0.0000	90.0000	0.0000
Шаг нашивки	6.8992	2.5635	0.0000	14.4405	6.9161
Плотность нашивки	57.1539	12.3510	0.0000	103.9889	57.3419

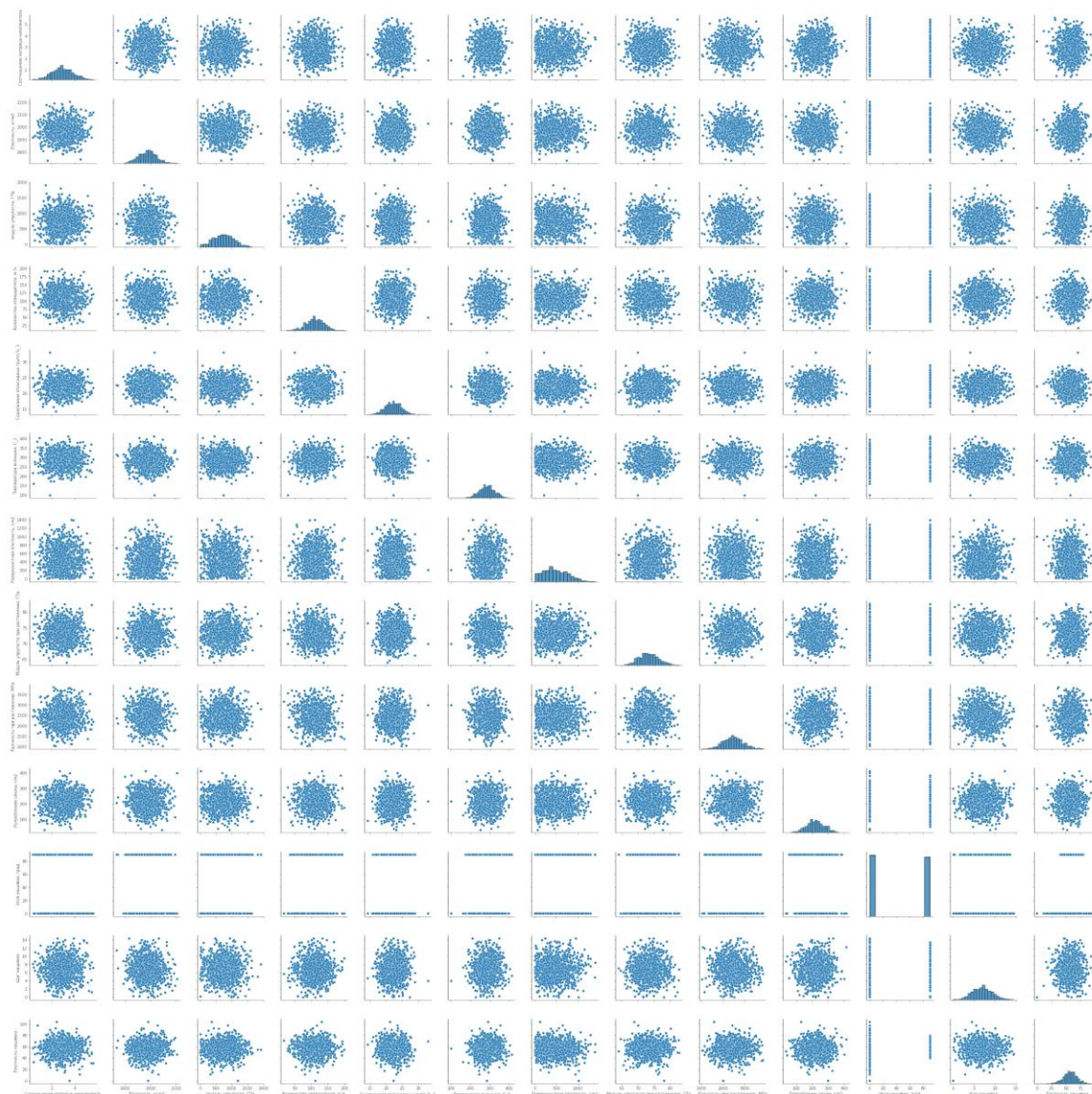
Пропусков  
нет

# Гистограммы распределения и диаграммы “ящик с усами”



- Большинство — количественные, вещественные, положительные, нормально распределенные
- Угол нашивки — категориальный, бинарный

# Попарные графики рассеяния точек



- Выбросы есть
- Зависимостей нет

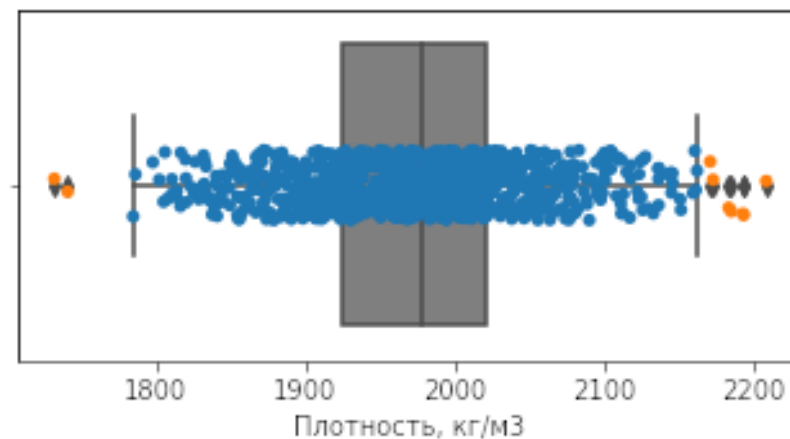
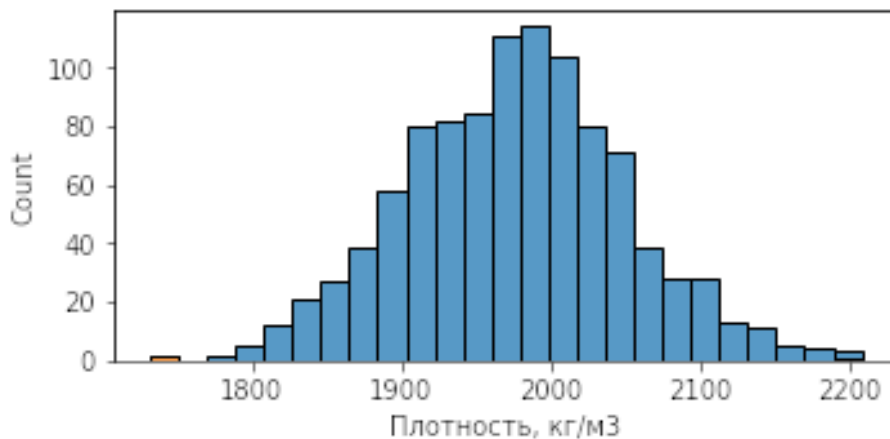
# Выбросы

Найдено:

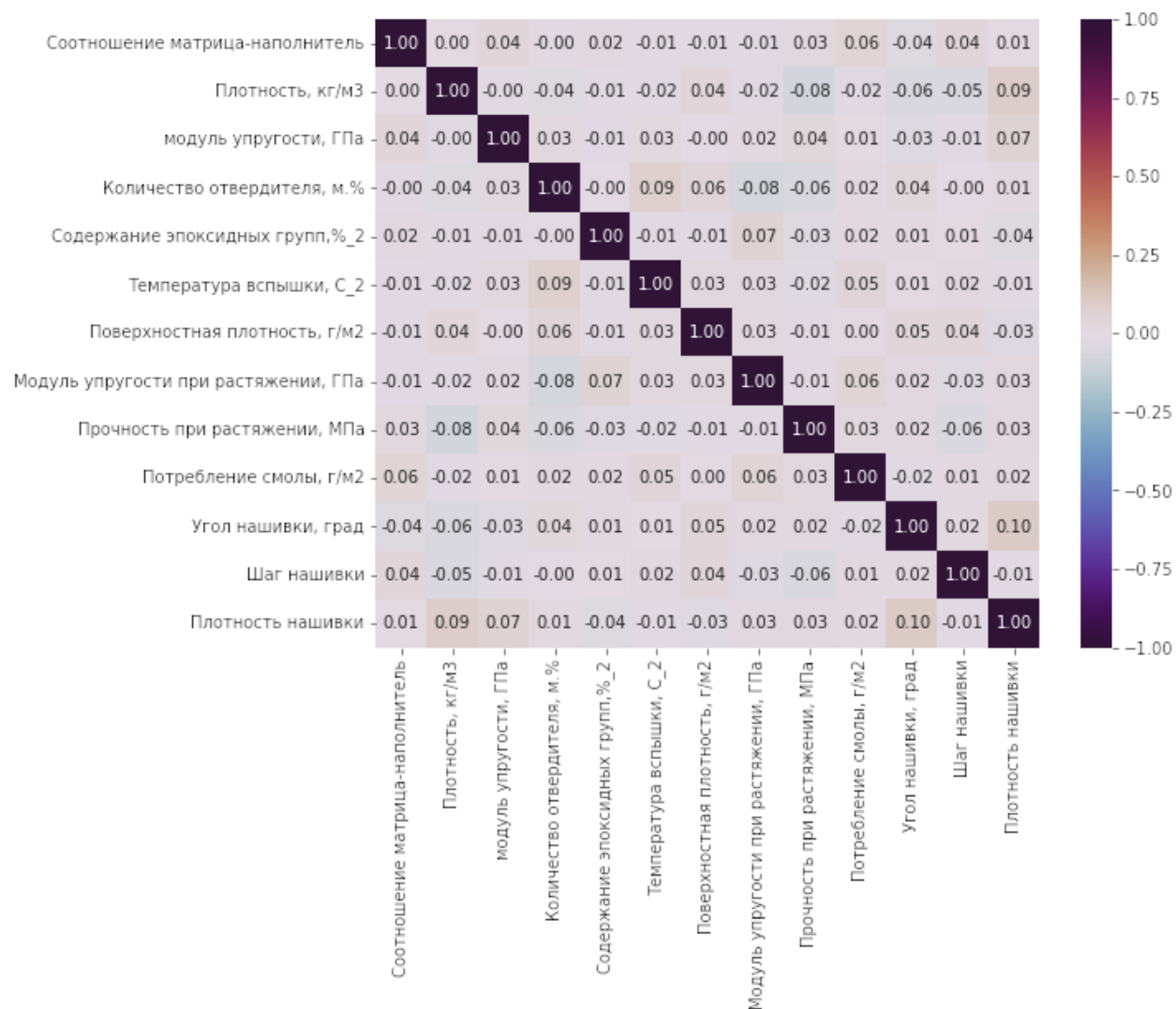
- методом 3-х сигм — 24 выброса
- методом межквартильных расстояний — 93 выброса

Удалить

осталось 1000 строк



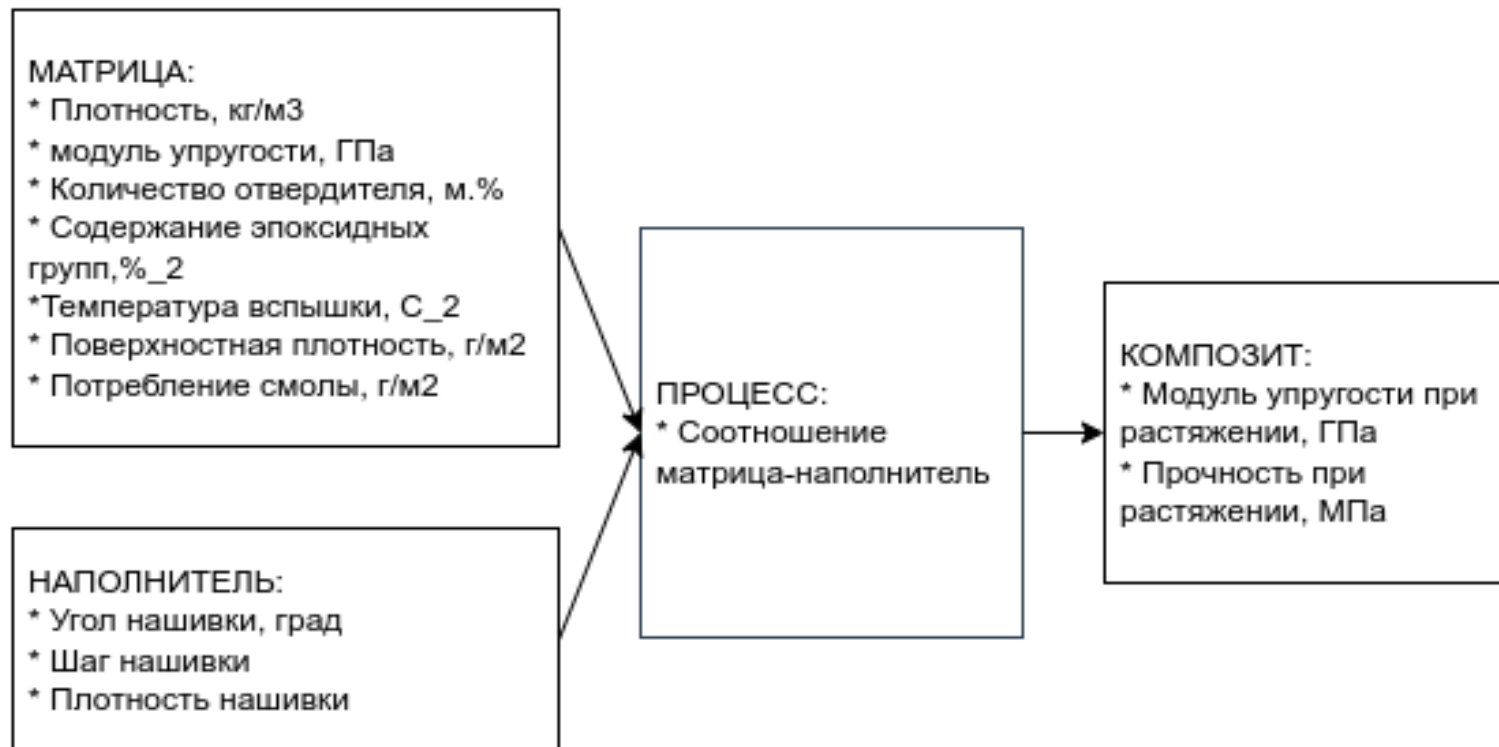
# Матрица корреляции



Линейной  
зависимости  
нет



# Предметная область: КОМПОЗИТНЫЕ материалы



# Выходные переменные

# Описательная статистика выходной переменной

Модуль упругости при растяжении, ГПа

min	64.054061
max	82.682051
mean	73.354026
std	3.066086

# Описательная статистика выходной переменной

Прочность при растяжении, МПа

min	1071.123751
max	3848.436732
mean	2468.178562
std	487.297434

# Описательная статистика выходной переменной

Соотношение матрица-наполнитель

min	0.389403
max	5.455566
mean	2.907441
std	0.908368

Для каждого признака — отдельная модель

- модуль упругости при растяжении
- прочность при растяжении
- соотношение матрица-наполнитель

# Входные переменные

Значения признаков в разных диапазонах =>  
необходим препроцессинг

- разделить на количественные и категориальные
- категориальные («Угол нашивки») - OrdinalEncoder
  - список значений стал [0, 1]
- количественные (остальные) — StandardScaler
  - матожидание стало 0
  - стандартное отклонение стало 1
- создать объект-препроцесор, сохранить вместе с моделью
  - для train — fit\_transform
  - для test — transform
  - для введенных данных — transform

# Метрики качества

- $R^2$  или коэффициент детерминации
- RMSE (Root Mean Squared Error) или корень из средней квадратичной ошибки
- MAE (Mean Absolute Error) или средняя абсолютная ошибка
- MAPE (Mean Absolute Percentage Error) или средняя абсолютная процентная ошибка
- max error или максимальная ошибка данной модели

# Модели

- Линейная регрессия
- Лассо (LASSO) и гребневая (Ridge) регрессия
- Метод опорных векторов для регрессии
- Метод k-ближайших соседей
- Деревья решений
- Случайный лес
- Градиентный бустинг
- Нейронная сеть

# Модель для модуля упругости при растяжении

Значения выхода  
от 64 до 83

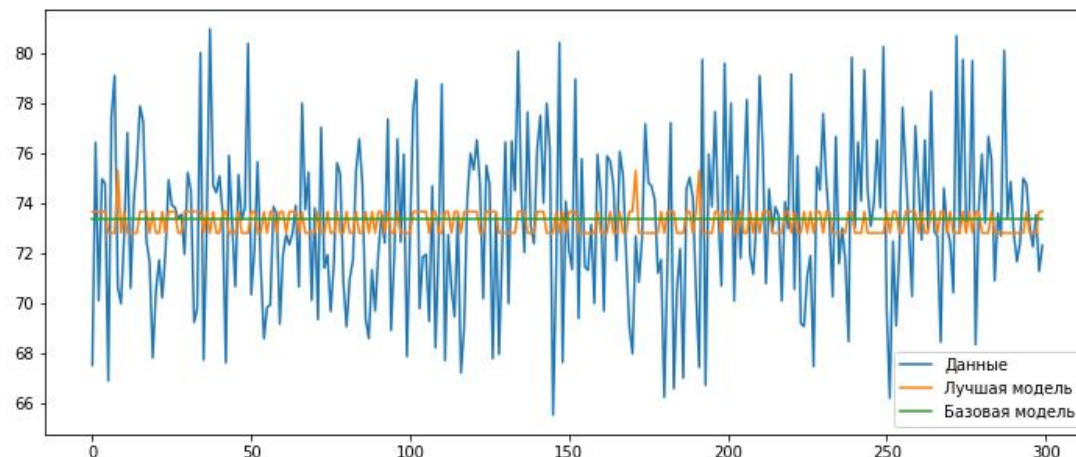
По умолчанию →

После подбора  
гиперпараметров ↓

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.021502	-3.059339	-2.465060	-0.033641	-8.053111
LinearRegression	-0.022620	-3.059379	-2.464305	-0.033641	-8.139731
Ridge	-0.022538	-3.059264	-2.464226	-0.033640	-8.139352
Lasso	-0.021502	-3.059339	-2.465060	-0.033641	-8.053111
SVR	-0.037763	-3.082058	-2.472179	-0.033767	-8.146369
KNeighborsRegressor	-0.197298	-3.312241	-2.624624	-0.035795	-8.876770
DecisionTreeRegressor	-1.229594	-4.485293	-3.545377	-0.048431	-12.178495
RandomForestRegressor	-0.061516	-3.117096	-2.485271	-0.033934	-8.457280

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=480, solver='lsqr')	-0.013299	-3.046623	-2.455526	-0.033517	-8.071899
Lasso(alpha=0.15)	-0.019048	-3.055423	-2.459921	-0.033574	-8.102101
SVR(C=0.015, kernel='linear')	-0.016521	-3.052020	-2.456808	-0.033549	-8.140634
KNeighborsRegressor(n_neighbors=25)	-0.030786	-3.074728	-2.461113	-0.033581	-8.031419
DecisionTreeRegressor(criterion='absolute_error', max_depth=2, max_features=10, random_state=3128, splitter='random')	-0.009281	-3.041407	-2.435050	-0.033185	-8.004156
RandomForestRegressor(bootstrap=False, criterion='absolute_error', max_depth=4, max_features=2, random_state=3128)	-0.015396	-3.049810	-2.446070	-0.033369	-8.275716

# Модель для модуля упругости при растяжении



	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.001377	-3.222954	-2.577796	-0.035319	-7.800690
Лучшая модель (дерево решений)	-0.035776	-3.277844	-2.610243	-0.035707	-8.152045

	R2	RMSE	MAE	MAPE	max_error
Модуль упругости, тренировочный	0.017295	-3.037284	-2.410294	-0.032850	-9.008468
Модуль упругости, тестовый	-0.035776	-3.277844	-2.610243	-0.035707	-8.152045

# Модель для прочности при растяжении

Значения выхода  
от 1071 до 3849

По умолчанию →

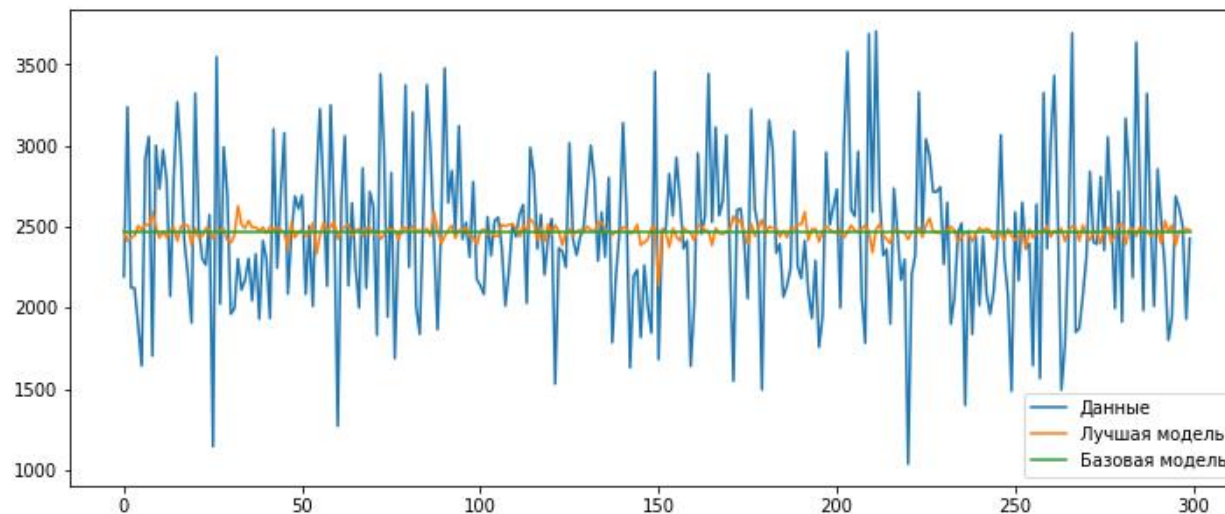
После подбора  
гиперпараметров ↓

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.012988	-484.654884	-385.827028	-0.169931	-1228.780064
LinearRegression	-0.022969	-487.063246	-388.303827	-0.170559	-1249.517419
Ridge	-0.022896	-487.046319	-388.290667	-0.170555	-1249.460177
Lasso	-0.021388	-486.695829	-387.988314	-0.170448	-1248.210674
SVR	-0.011952	-484.429045	-385.715018	-0.169382	-1232.355369
DecisionTreeRegressor	-1.187233	-702.791415	-555.350332	-0.238620	-1927.849316
GradientBoostingRegressor	-0.084580	-500.230316	-398.052645	-0.174164	-1312.873325

	R2	RMSE	MAE	MAPE	max_error
Ridge(alpha=990, solver='sparse_cg')	-0.010764	-484.199853	-385.891069	-0.169828	-1233.196571
Lasso(alpha=50)	-0.012988	-484.654884	-385.827028	-0.169931	-1228.780064
SVR(C=0.2)	-0.012246	-484.489867	-385.724279	-0.169413	-1232.341495
DecisionTreeRegressor(criterion='poisson', max_depth=3, max_features=6, random_state=3128, splitter='random')	-0.009440	-483.713960	-384.045197	-0.169031	-1244.359901
GradientBoostingRegressor(max_depth=1, max_features=1, n_estimators=50, random_state=3128)	-0.005486	-483.026609	-385.268908	-0.169409	-1231.878292



# Модель для прочности при растяжении



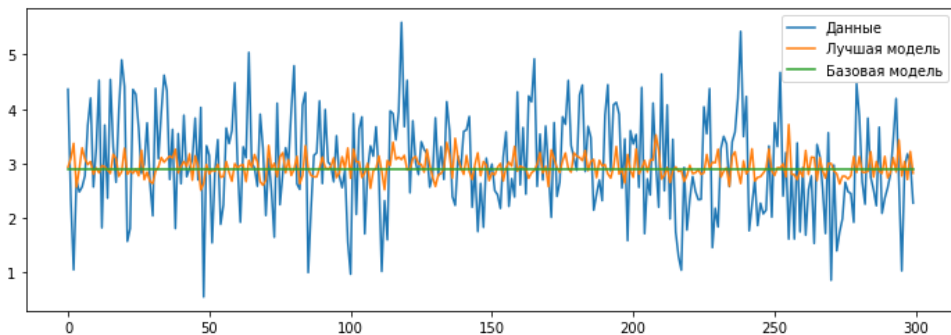
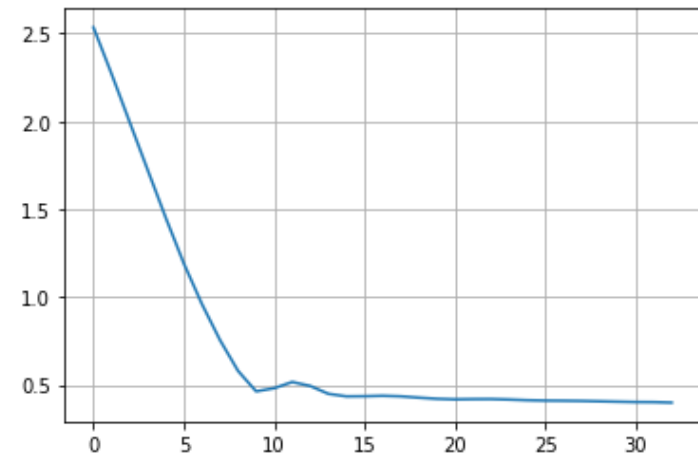
	R2	RMSE	MAE	MAPE	max_error
Базовая модель	-0.000531	-479.694153	-375.066608	-0.165566	-1431.321957
Лучшая модель (градиентный бустинг)	0.004028	-478.600202	-376.647056	-0.166046	-1384.841404

	R2	RMSE	MAE	MAPE	max_error
Прочность при растяжении, тренировочный	0.057141	-472.832206	-374.670333	-0.164825	-1383.885510
Прочность при растяжении, тестовый	0.004028	-478.600202	-376.647056	-0.166046	-1384.841404

# Модель для соотношения матрица-наполнитель

MLPRegressor  
из библиотеки sklearn



	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.011269	-0.911261	-0.737067	-0.299795	-2.684301
MLPRegressor	-0.052842	-0.929803	-0.751262	-0.306957	-2.790557

Значения выхода от 0.39 до 5.46

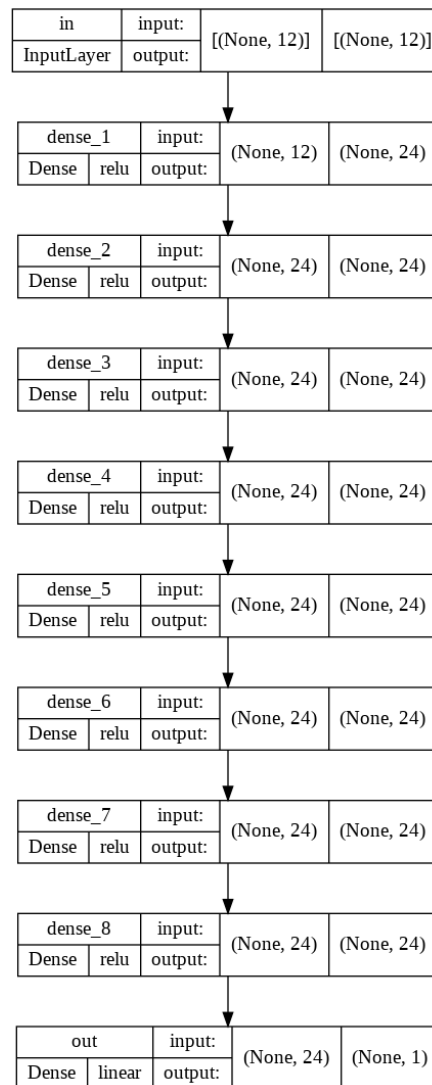
# Модель для соотношения матрица-наполнитель

Нейросеть  
из библиотеки  
tensorflow

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 24)	312
dense_2 (Dense)	(None, 24)	600
dense_3 (Dense)	(None, 24)	600
dense_4 (Dense)	(None, 24)	600
dense_5 (Dense)	(None, 24)	600
dense_6 (Dense)	(None, 24)	600
dense_7 (Dense)	(None, 24)	600
dense_8 (Dense)	(None, 24)	600
out (Dense)	(None, 1)	25

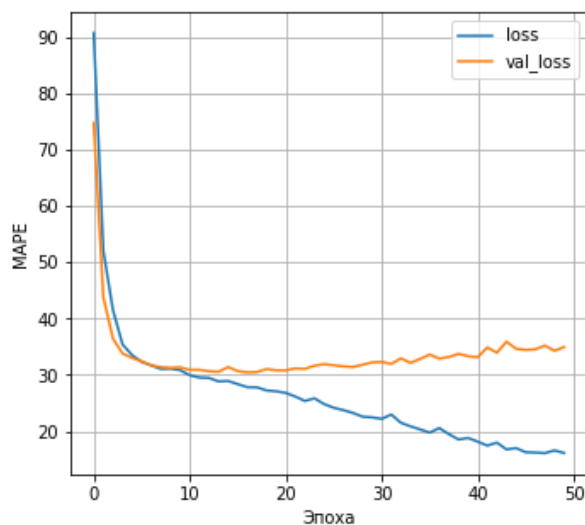
---

Total params: 4,537  
 Trainable params: 4,537  
 Non-trainable params: 0

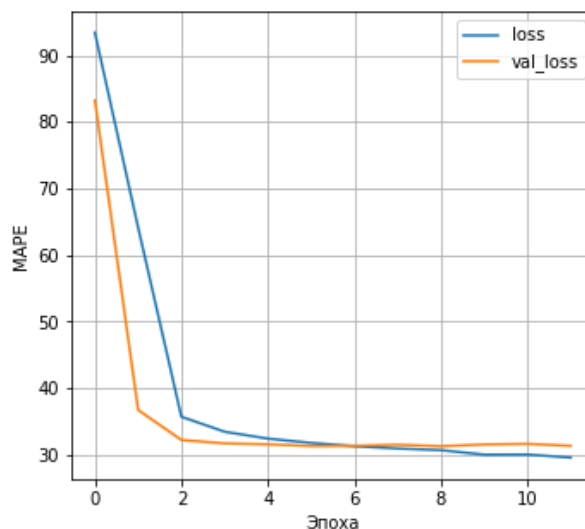


# Модель для соотношения матрица-наполнитель

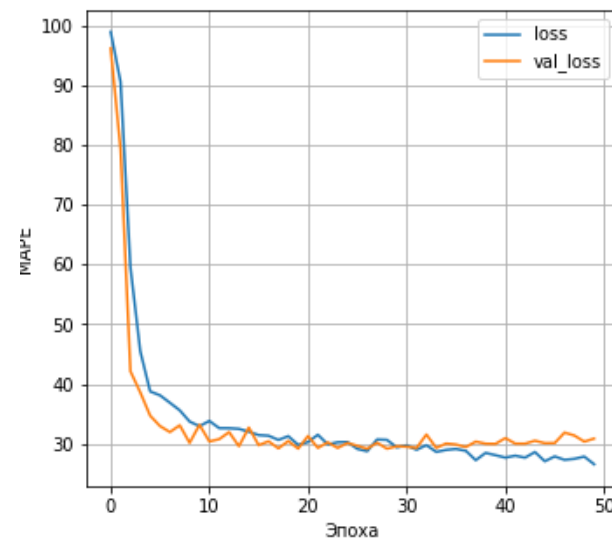
Обучение  
нейросети



Борьба с  
переобучением:  
ранняя остановка



Борьба с  
переобучением:  
Dropout

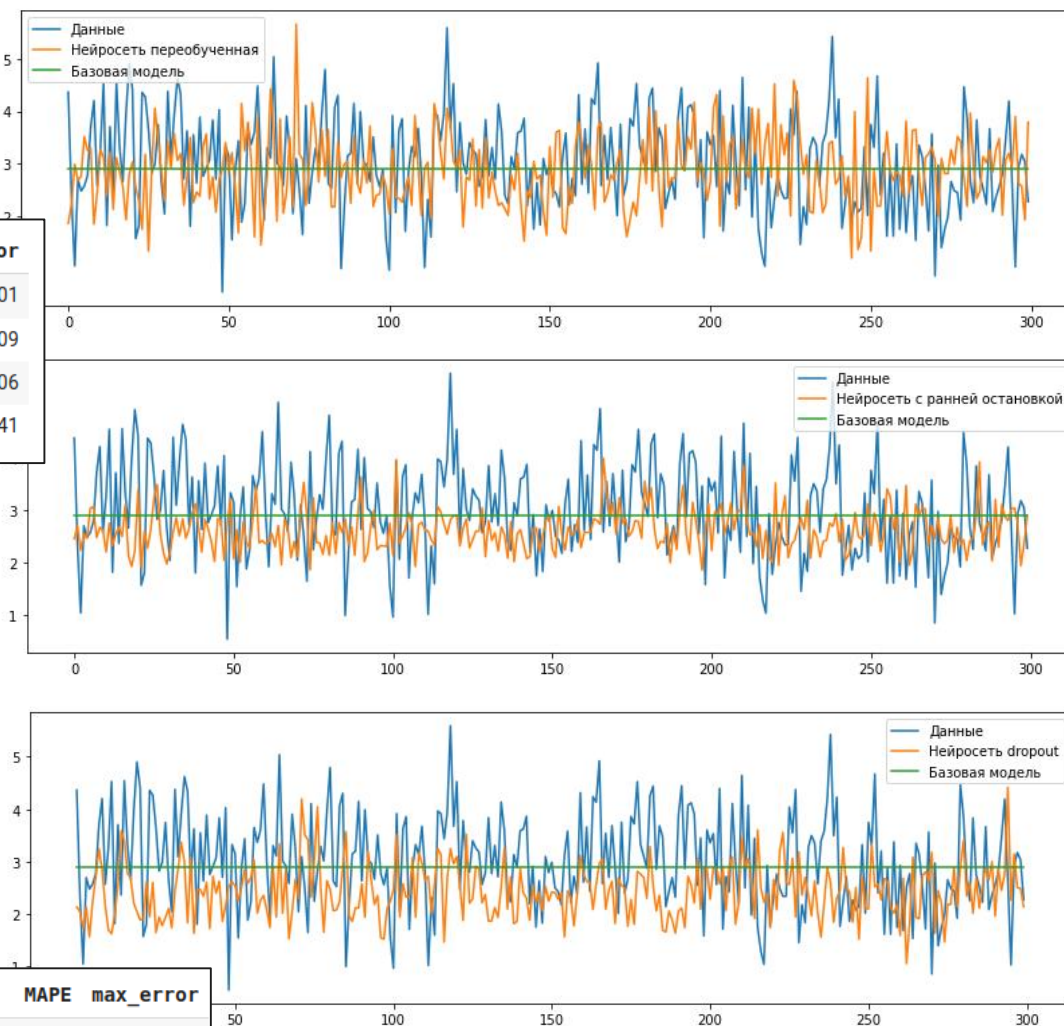


# Модель для соотношения матрица-наполнитель

Значения выхода от 0.39 до 5.46

	R2	RMSE	MAE	MAPE	max_error
DummyRegressor	-0.011269	-0.911261	-0.737067	-0.299795	-2.684301
Нейросеть переобученная	-0.624376	-1.154922	-0.938195	-0.373712	-2.868809
Нейросеть с ранней остановкой	-0.322407	-1.042058	-0.852214	-0.312846	-2.781806
Нейросеть dropout	-0.628132	-1.156256	-0.960385	-0.343979	-2.903841

Выбираю нейросеть,  
обученную  
с ранней остановкой



	R2	RMSE	MAE	MAPE	max_error
Соотношение матрица-наполнитель, тренировочный	-0.212722	-0.999613	-0.787676	-0.298627	-3.084322
Соотношение матрица-наполнитель, тестовый	-0.322407	-1.042058	-0.852214	-0.312846	-2.781806

# Разработка веб-приложения

ВКР x +

127.0.0.1:5000/model\_1\_2/

## Прогнозирование модуля упругости при растяжении и прочности при растяжении

Соотношение матрица-наполнитель (0..6)

Плотность, кг/м3 (1700...2300)

Модуль упругости, ГПа (2...2000)

Количество отвердителя, м.% (17...200)

Содержание эпоксидных групп, %\_2 (14...34)

Температура вспышки, С\_2 (100...414)

Поверхностная плотность, г/м2 (0.6...1400)

Потребление смолы, г/м2 (33...414)

Угол нашивки, град (0 или 90)

Шаг нашивки (0...15)

Плотность нашивки (0...104)

Входные переменные:

	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп, %_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	4.029126	1880.0	622.0	111.86	22.267857	284.615385	470.0	220.0	90.0	4.0	60.0

Результат модели:

Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа
72.81891497929365	2523.9223070281537

# Разработка веб-приложения

ВКР

127.0.0.1:5000/model\_3/

### Прогнозирование соотношения матрица-наполнитель

Плотность, кг/м3 (1700...2300)

Модуль упругости, ГПа (2...2000)

Количество отвердителя, м.% (17...200)

Содержание эпоксидных групп, %\_2 (14...34)

Температура вспышки, C\_2 (100...414)

Поверхностная плотность, г/м2 (0.6...1400)

Модуль упругости при растяжении, ГПа (64...83)

Прочность при растяжении, МПа (1036...3849)

Потребление смолы, г/м2 (33...414)

Угол нашивки, град (0 или 90)

Шаг нашивки (0...15)

Плотность нашивки (0...104)

Входные переменные:

	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп, %_2	Температура вспышки, C_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	1880.0	622.0	111.86	22.267857	284.615385	470.0	73.333333	2455.555556	220.0	90.0	4.0	60.0

Результат модели:

Соотношение матрица-наполнитель
2.515496058558928

23

# Результаты

## Задача не решена

Дальнейшие поиски решения могли бы включать:

- проконсультироваться у экспертов
- уточнить постановку задачи
- исследовать сырые данные
- провести отбор признаков и уменьшение размерности
- поэкспериментировать с градиентным бустингом
- углубиться в нейросети





**Спасибо за внимание!**