Образовательный центр МГТУ им. Н.Э. Баумана

Выпускная квалификационная работа
по курсу «Data Science»

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель: Гутенев Евгений

Постановка задачи

- изучить и описать предметную область
- провести разведочный анализ данных
- провести предобработку данных
- разделить данные на тренировочную и тестовую выборки
- выбрать базовую модель и несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении
- обучить выбранные модели с гиперпараметрами по умолчанию
- провести поиск гиперпараметров сравниваемых моделей с помощью поиска по сетке с перекрестной проверкой
- сравнить модели после подбора гиперпараметров и выбрать лучшую
- написать нейронную сеть, которая будет рекомендовать соотношение «матрица-наполнитель»
- разработать приложение, которое будет выдавать соответствующий прогноз
- оценить точность моделей на тренировочном и тестовом датасетах
- создать репозиторий в GitHub и разместить там код исследования

Изучение и описание предметной области

Датасет со свойствами композитов:

- X_bp (матрица из базальтопластика, 10 признаков, 1023 строк)
- X_nup (наполнитель из углепластика, 3 признака, 1040 строк)

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
- модуль упругости при растяжении, ГПа
- прочность при растяжении, МПа
- потребление смолы, г/м2
- угол нашивки, град
- шаг нашивки
- плотность нашивки

Выходные переменные (исключаются из списка выходных в процессе решения задачи):

Задача 1 (регрессия):

- модуль упругости при растяжении, ГПа

Задача 2 (регрессия):

- прочность при растяжении, МПа

Задача 3 (рекомендательная система на основе нейросети):

- соотношение матрица-наполнитель

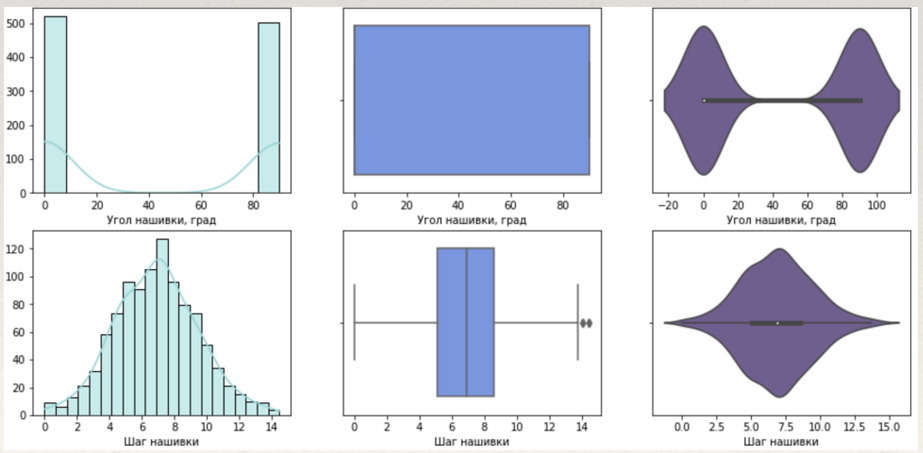
После объединения типом INNER по индексу получаем датасет, включающий 13 признаков и 1023 строк

Разведочный анализ данных

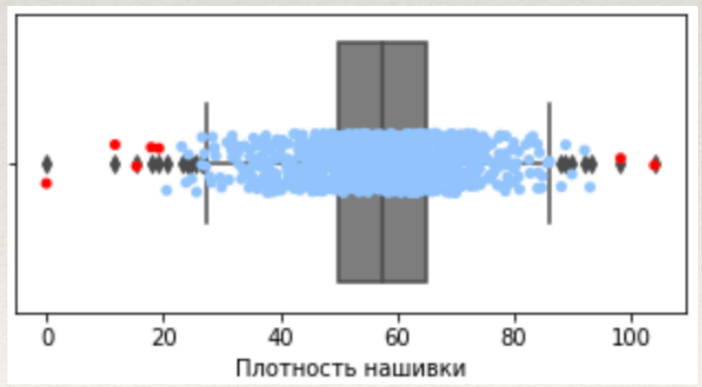
	count	mean	std	min	25%	50%	75%	max	median
Соотношение матрица-наполнитель	1023.000	2.930	0.913	0.389	2.318	2.907	3.553	5.592	2.907
Плотность, кг/м3	1023.000	1975.735	73.729	1731.765	1924.155	1977.622	2021.374	2207.773	1977.622
модуль упругости, ГПа	1023.000	739.923	330.232	2.437	500.047	739.664	961.813	1911.536	739.664
Количество отвердителя, м.%	1023.000	110.571	28.296	17.740	92.443	110.565	129.730	198.953	110.565
Содержание эпоксидных групп,%_2	1023.000	22.244	2.406	14.255	20.608	22.231	23.962	33.000	22.231
Температура вспышки, С_2	1023.000	285.882	40.943	100.000	259.067	285.897	313.002	413.273	285.897
Поверхностная плотность, г/м2	1023.000	482.732	281.315	0.604	266.817	451.864	693.225	1399.542	451.864
Модуль упругости при растяжении, ГПа	1023.000	73.329	3.119	64.054	71.245	73.269	75.357	82.682	73.269
Прочность при растяжении, МПа	1023.000	2466.923	485.628	1036.857	2135.850	2459.525	2767.193	3848.437	2459.525
Потребление смолы, г/м2	1023.000	218.423	59.736	33.803	179.628	219.199	257.482	414.591	219.199
Угол нашивки, град	1023.000	44.252	45.016	0.000	0.000	0.000	90.000	90.000	0.000
Шаг нашивки	1023.000	6.899	2.563	0.000	5.080	6.916	8.586	14.441	6.916
Плотность нашивки	1023.000	57.154	12.351	0.000	49.799	57.342	64.945	103.989	57.342

Описательная статистика

- пропуски отсутствуют
- все значения признаков (кроме признака «Угол нашивки, град») имеют нормальное распределение
- выбросы присутствуют



Распределение значений признаков «Угол нашивки» и «Шаг нашивки»



Поиск выбросов критерием трех сигм

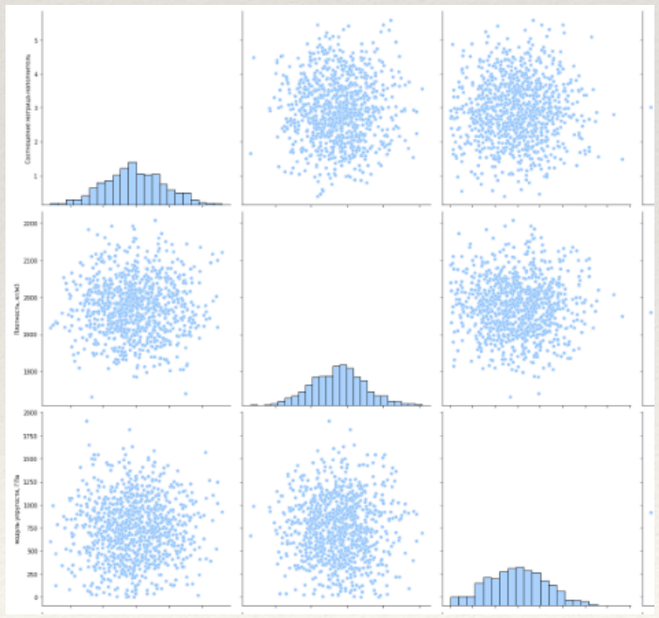
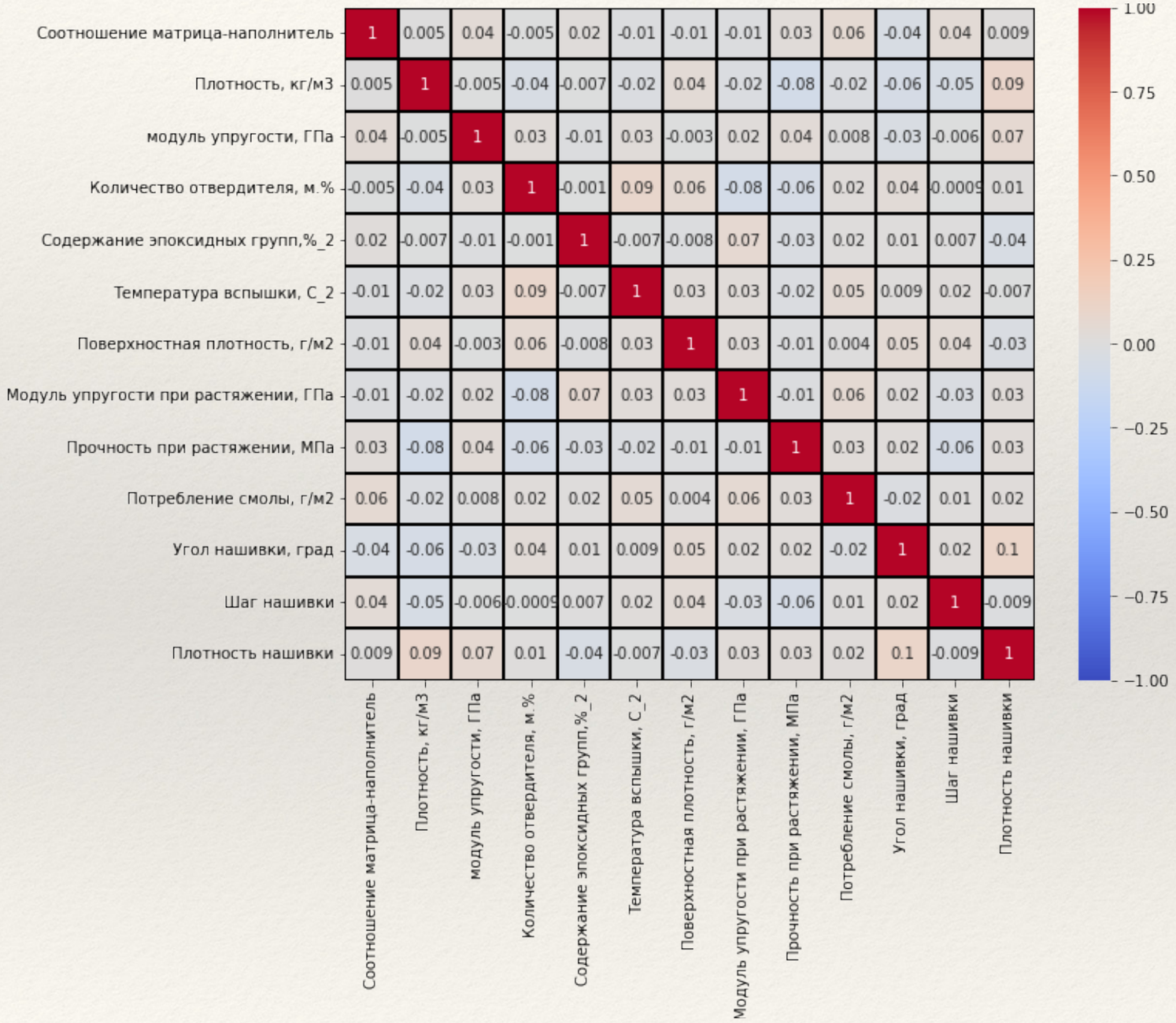


Диаграмма рассеяния первых трех признаков

Разведочный анализ данных



Тепловая карта коэффициентов корреляции

Линейная зависимость отсутствует

Предобработка данных. Выбор моделей

> Рассчитываем количество выбросов (найдено 24) и удаляем их из датасета
> Группируем признаки для прогнозирования по каждой из 3 задач
> Нормализуем и стандартизируем данные:
- используем OrdinalEncoder для категориальных признаков
- используем MinMaxScaler для количественных признаков
> Делим выборку на тренировочную и тестовую

Используемые метрики качества модели:

- R2 (коэффициент детерминации)
- RMSE (Root Mean Squared Error, корень из средней квадратичной ошибки)
- MAE (Mean Absolute Error, средняя абсолютная ошибка)
- max error (максимальная ошибка)

Базовая модель:

- DummyRegressor

Модели для прогноза:

- Линейная регрессия (LinearRegression)
- Лассо (Lasso)
- Метод опорных векторов (SVR)
- Метод k-ближайших соседей (KNeighborsRegressor)
- Деревья решений (DecisionTreeRegressor)
- Градиентный бустинг (GradientBoostingRegressor)
- Случайный лес (RandomForestRegressor)
- Нейронная сеть (MLPRegressor / Sequential)

1) Выбор моделей для прогнозирования модуля упругости при растяжении

Сравнение результативности моделей с гиперпараметрами по умолчанию

	R2	RMSE	MAE	max_error
DummyRegressor	-0.019376	-3.126837	-2.510495	-7.798105
LinearRegression	-0.018532	-3.123936	-2.502366	-8.098392
Lasso	-0.019376	-3.126837	-2.510495	-7.798105
SVR	-0.039011	-3.153628	-2.514390	-8.311025
KNeighborsRegressor	-0.227280	-3.424991	-2.724152	-8.703162
DecisionTreeRegressor	-1.034156	-4.403633	-3.589790	-11.822403
GradientBoostingRegressor	-0.098848	-3.242846	-2.593079	-8.632489
RandomForestRegressor	-0.075323	-3.208305	-2.555245	-8.380008

Описательная статистика выходных переменных

min	64.05406
max	82.23760
mean	73.39876

Сравнение результативности моделей с подобранными гиперпараметрами

	R2	RMSE	MAE	max_error
Lasso(alpha=0.01)	-0.013670	-3.116889	-2.501700	-7.985118
SVR(C=0.5)	-0.021235	-3.127617	-2.501344	-8.154217
KNeighborsRegressor(n_neighbors=91)	-0.009123	-3.110492	-2.497003	-7.900511
DecisionTreeRegressor(max_depth=2, max_features=5, min_samples_split=3, random_state=42)	-0.024121	-3.135466	-2.496037	-8.292868
GradientBoostingRegressor(learning_rate=0.01, loss='absolute_error', min_samples_split=10, random_state=42)	-0.014217	-3.118909	-2.496307	-7.994517
RandomForestRegressor(bootstrap=False, criterion='absolute_error', max_features=1, n_estimators=50, random_state=42)	-0.023475	-3.134199	-2.484176	-8.293257

2) Выбор моделей для прогнозирования прочности при растяжении

Сравнение результативности моделей с гиперпараметрами по умолчанию

	R2	RMSE	MAE	max_error
DummyRegressor	-0.022944	-493.539876	-391.010975	-1281.791709
LinearRegression	-0.014804	-491.329446	-391.262712	-1305.947015
Lasso	-0.011470	-490.565450	-390.216167	-1301.382442
SVR	-0.020978	-493.091014	-390.520034	-1278.945635
KNeighborsRegressor	-0.197623	-533.145743	-424.397300	-1408.337527
DecisionTreeRegressor	-1.097452	-700.282012	-561.980002	-1784.349498
GradientBoostingRegressor	-0.033926	-496.051787	-398.082126	-1274.138037
RandomForestRegressor	-0.036206	-496.315183	-393.991316	-1325.856363

Описательная статистика выходных переменных

min	1071.12375
max	3848.43673
mean	2469.10920

Сравнение результативности моделей с подобранными гиперпараметрами

	R2	RMSE	MAE	max_error
Lasso(alpha=1)	-0.011470	-490.565450	-390.216167	-1301.382442
SVR(C=0.01, degree=5, kernel='poly')	0.001648	-487.500183	-386.090173	-1287.643454
KNeighborsRegressor(n_neighbors=67)	-0.015582	-491.711317	-389.708036	-1286.796521
DecisionTreeRegressor(criterion='poisson', max_depth=3, max_features=10, min_samples_split=3, random_state=42, splitter='random')	-0.014634	-491.520067	-388.357164	-1282.857828
GradientBoostingRegressor(loss='absolute_error', n_estimators=75, random_state=42)	0.002556	-487.325938	-386.488031	-1250.025205
RandomForestRegressor(bootstrap=False, criterion='poisson', max_depth=4, max_features=1, n_estimators=50, random_state=42)	-0.015398	-491.694360	-388.959919	-1278.428575

3) Выбор модели для прогнозирования соотношения «матрица-наполнитель»

Сравнение результативности модели MLPRegressor

	R2	RMSE	MAE	max_error
DummyRegressor	-0.000744	-0.924041	-0.739327	-2.554458
MLPRegressor	-0.043173	-0.943426	-0.753366	-2.633232

Сравнение результативности моделей Sequential (keras)

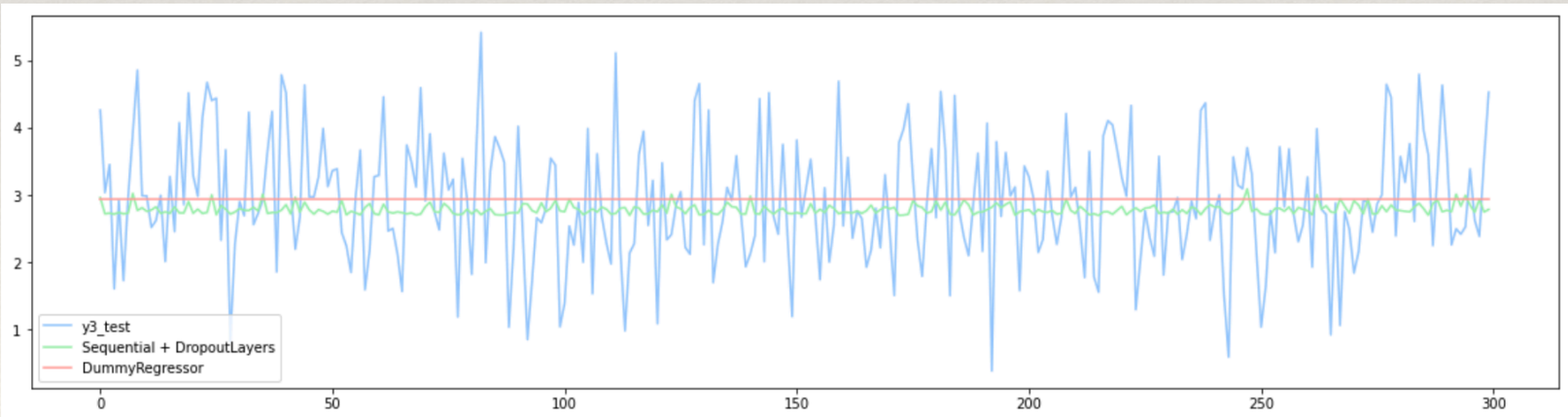
	R2	RMSE	MAE	max_error
DummyRegressor	-0.000744	-0.924041	-0.739327	-2.554458
Sequential	-0.254183	-1.034452	-0.837282	-3.580135
Sequential with EarlyStopping	-0.069018	-0.955042	-0.761271	-2.711501
Sequential with DropoutLayers	0.000587	-0.923427	-0.735780	-2.554991

Сводная информация по архитектуре сети Sequential

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, 36)	900
dense_5 (Dense)	(None, 36)	1332
dense_6 (Dense)	(None, 36)	1332
dense_7 (Dense)	(None, 36)	1332
dense_8 (Dense)	(None, 24)	888
dense_9 (Dense)	(None, 48)	1200
dense_10 (Dense)	(None, 24)	1176
out (Dense)	(None, 1)	25

Total params: 9,697
Trainable params: 9,697
Non-trainable params: 0

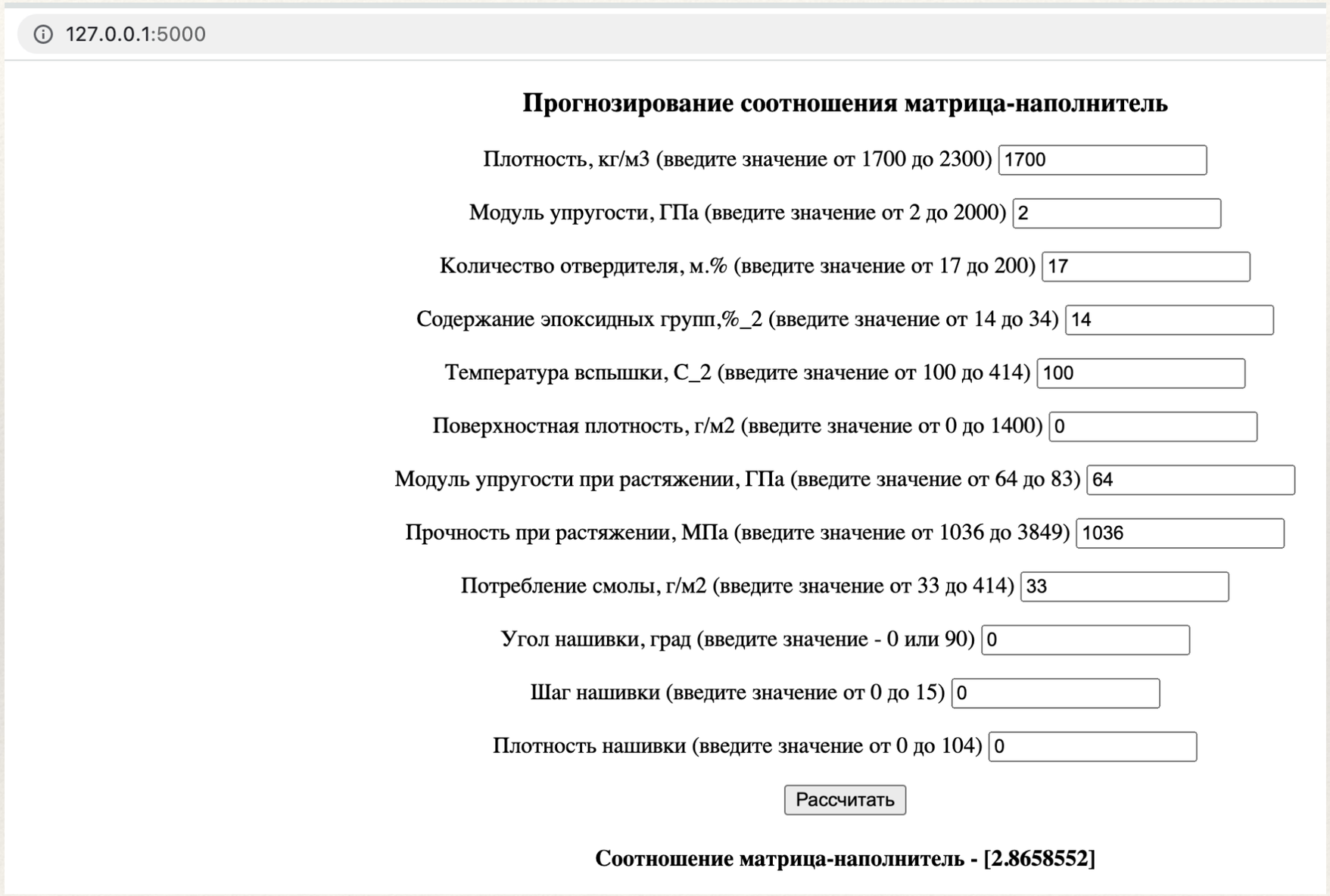
Визуализация сравнения прогноза модели Sequential с Dropout-слоями с базовой



Описательная статистика выходных переменных

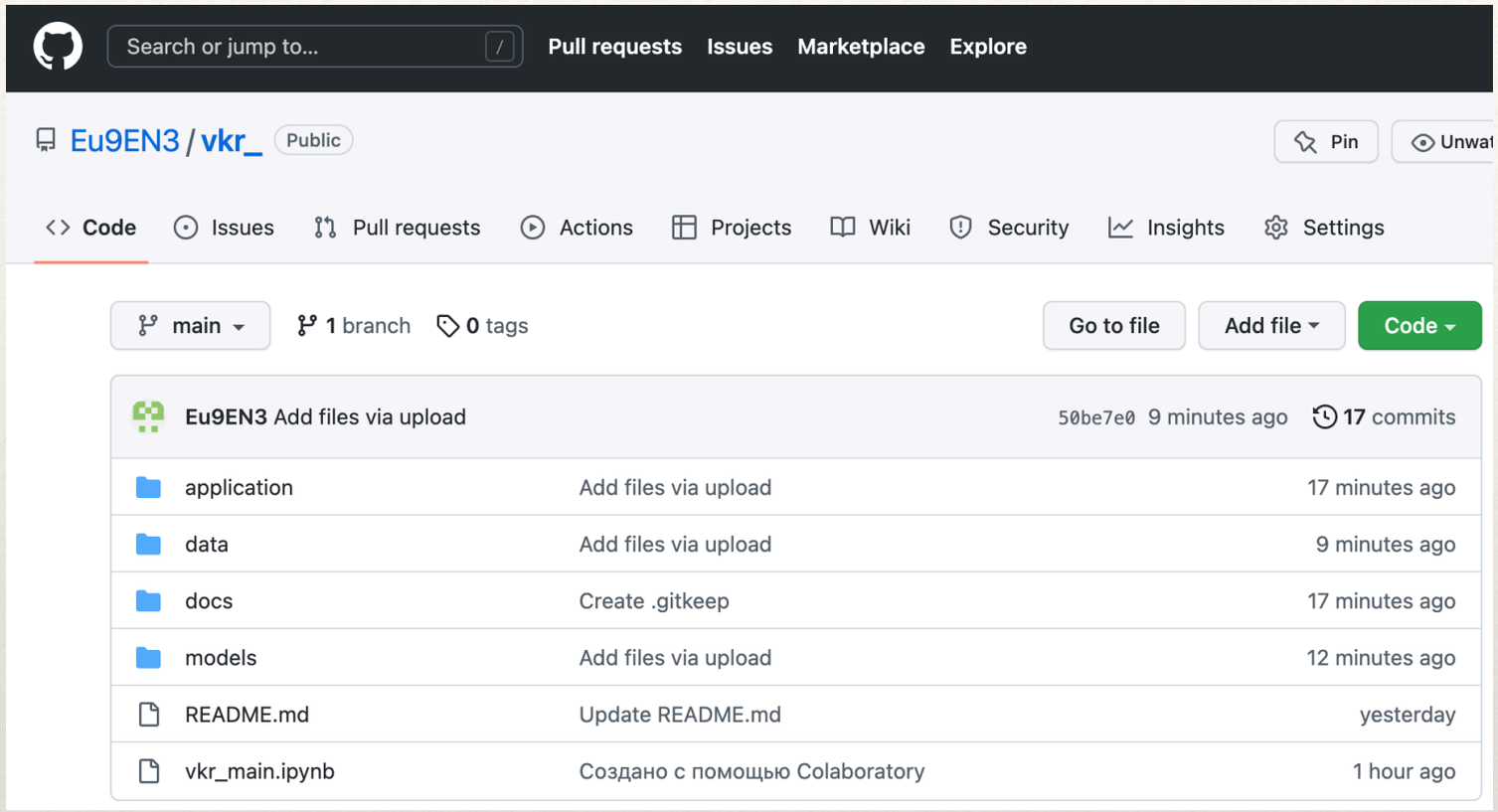
min	0.54739
max	5.59174
mean	2.94386

Разработка приложения для рекомендательной системы



Редактор исходного кода: VS Code
Язык: Python
Интерпретатор: Flask

Создание репозитория в GitHub и размещение кода исследования



Благодарю за внимание!