ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ МГТУ им. Н.Э. Баумана

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э.
Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу

«Data Science Pro»

Слушатель Петров Сергей Вячеславович

Москва, 2024

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

1. Введение

Тема данной работы - прогнозирование конечных свойств новых материалов (композиционных материалов). Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними.

Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т.е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита — железобетон.

Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

2. Аналитическая часть

2.1 Постановка задачи.

На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.).

На выходе необходимо:

1. Обучить алгоритм машинного обучения, который будет определять значения:
   - модуля упругости при растяжении, ГПа;
   - прочности при растяжении, МПа.
2. Написать нейронную сеть, которая будет рекомендовать: соотношение матрица-наполнитель.

Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов и цифровыми двойниками новых композитов.

Датасет со свойствами композитов представлен в виде двух файлов формата Excel.

X_bp.xlsx, состоит из 11 столбцов, первый из которых является порядковым номером (индексом) для строк, общее количество которых составляет 1023.

X_nup.xlsx, состоит из 4 столбцов, первый из которых является порядковым номером (индексом) для строк, общее количество которых составляет 1040.

В соответствии описанием поставленной задачи, данные массивы данных необходимо объединить по индексу. Объединенный датасет представляет из себя все строки из первого файла и присоединенные к ним справа

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

строки из второго файла, связующим ключем являются индексы каждого файла.

Объединенный датасет имеет следующую структуру и типы данных:

#	Название столбца	Количество ненулевых	Общее количество	Тип данных
0	Соотношение матрица-наполнитель	1023	1023	float64
1	Плотность, кг/м3	1023	1023	float64
2	модуль упругости, ГПа	1023	1023	float64
3	Количество отвердителя, м.%	1023	1023	float64
4	Содержание эпоксидных групп,%_2	1023	1023	float64
5	Температура вспышки, С_2	1023	1023	float64
6	Поверхностная плотность, г/м2	1023	1023	float64
7	Модуль упругости при растяжении, ГПа	1023	1023	float64
8	Прочность при растяжении, МПа	1023	1023	float64
9	Потребление смолы, г/м2	1023	1023	float64
10	Угол нашивки, град	1023	1023	int64

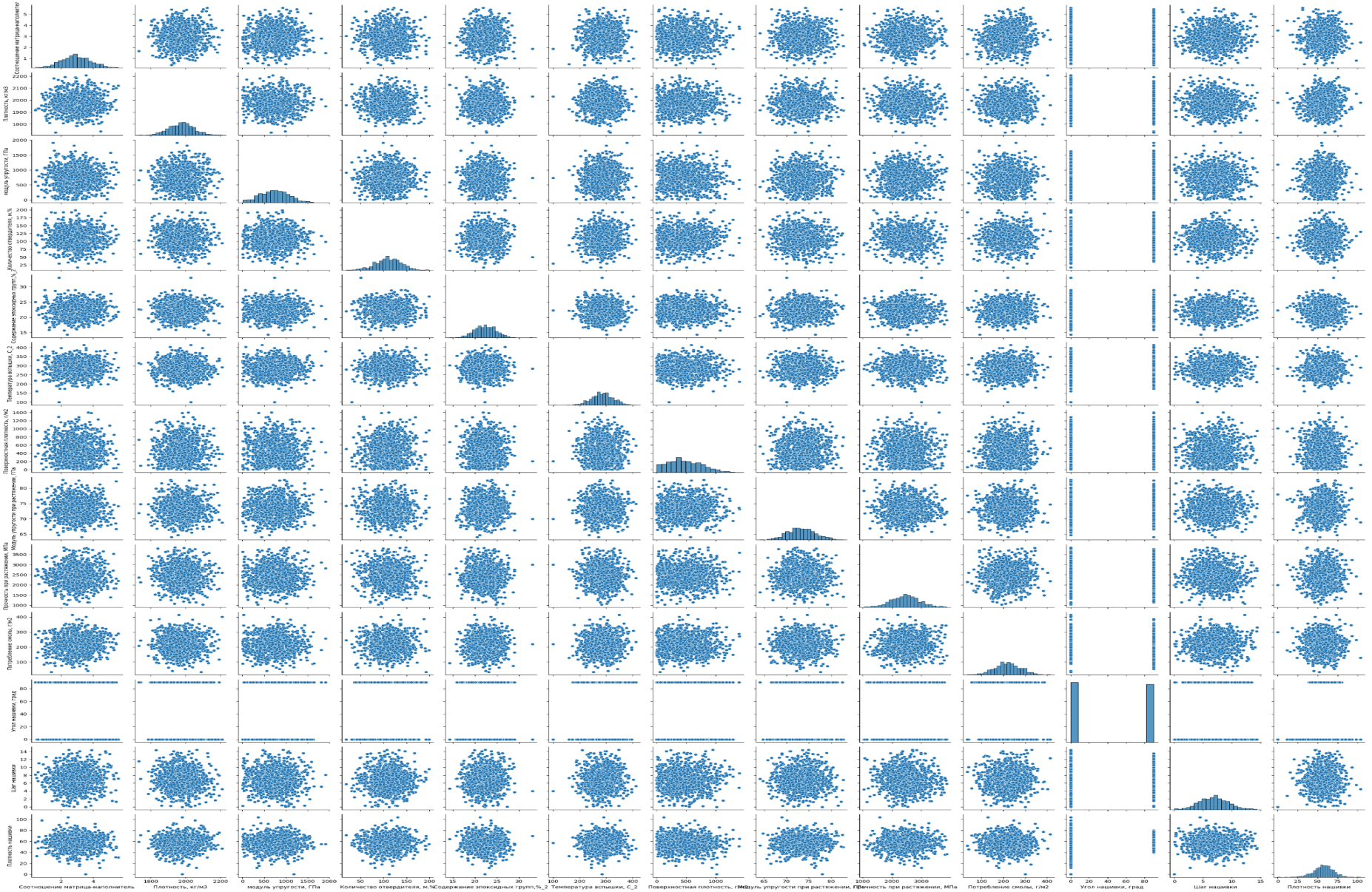
Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

11	Шаг нашивки	1023	1023	float64
12	Плотность нашивки	1023	1023	float64

Все представленные данные имеют числовой формат, пропуски отсутствуют.

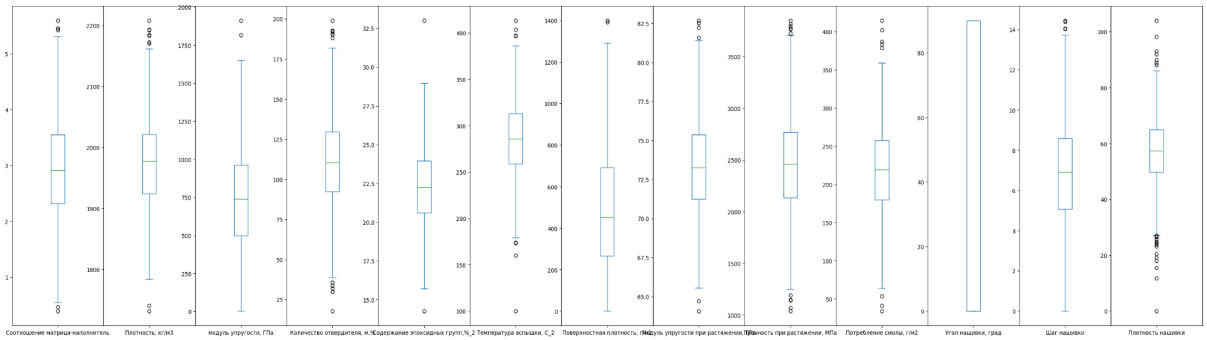
Для сравнения распределения данных использована гистограмма попарного распределения:

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru



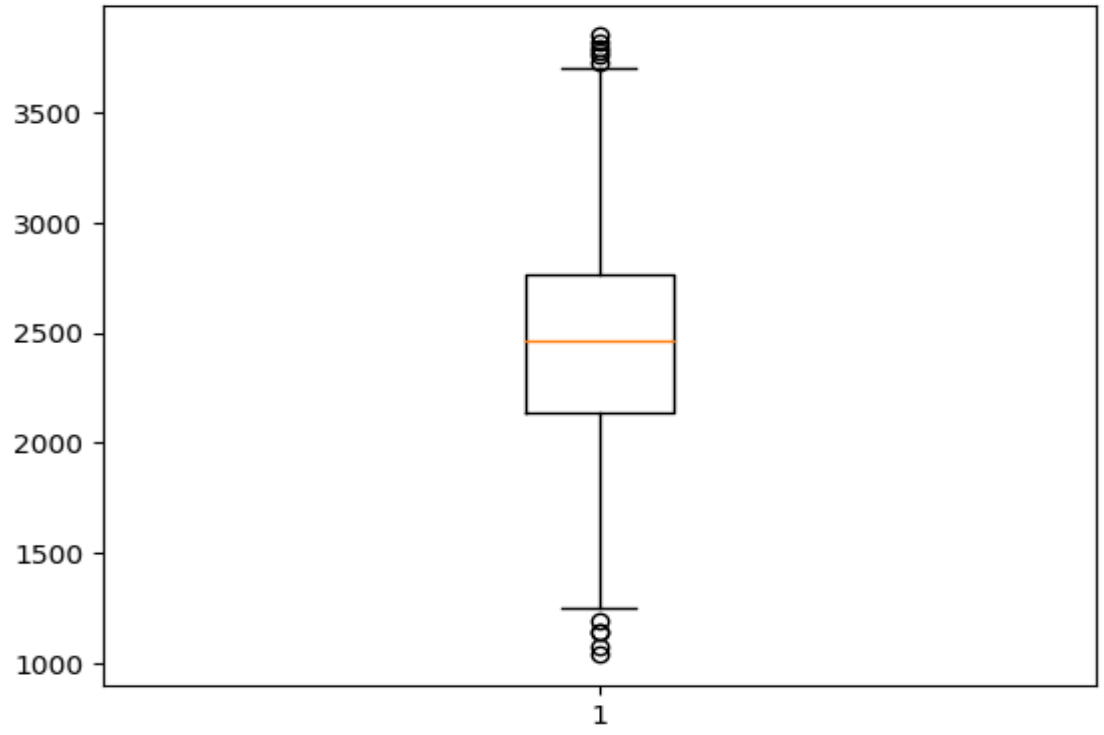
Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

На основании графиков следует вывод об отсутствии линейных зависимостей между данными, как следствие предполагается низкая эффективность определения целевых значений. Однако, стоит отметить, что каждый набор данных имеет близкое к нормальному распределению, что в свою очередь может свидетельствовать о предварительной подготовке датасета или его синтетической природе. В данном датасете столбец “Угол нашивки, град” представлен двумя величинами: 0 и 90. Выбросы в данных практически незначительные, что демонстрирует следующая диаграмма:



Выбросы целевых показателей выглядят следующим образом:

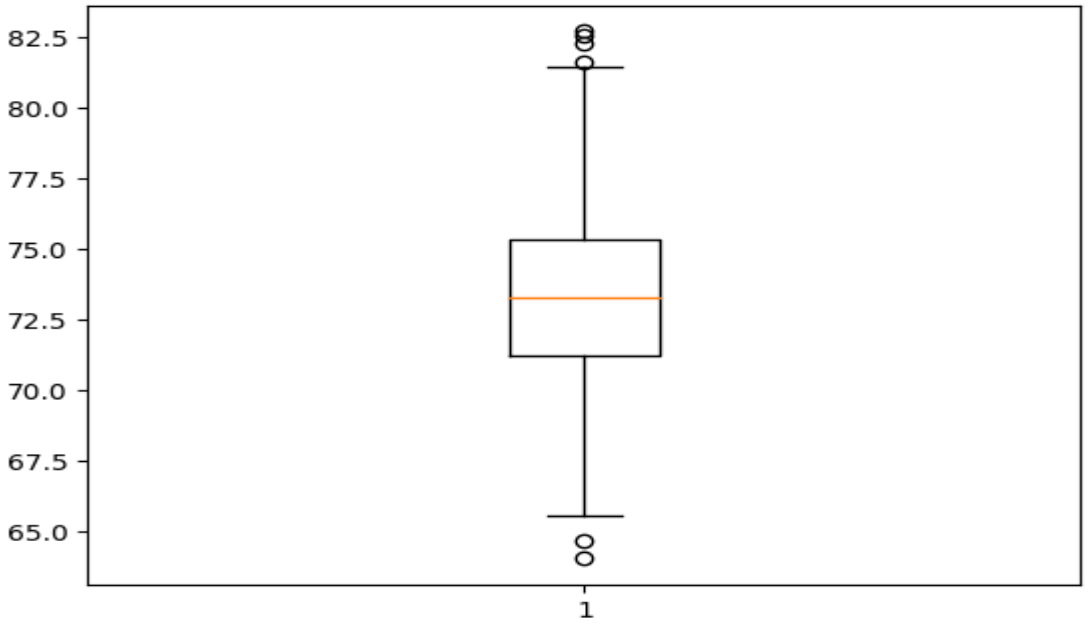
для “Прочность при растяжении, МПа”



Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

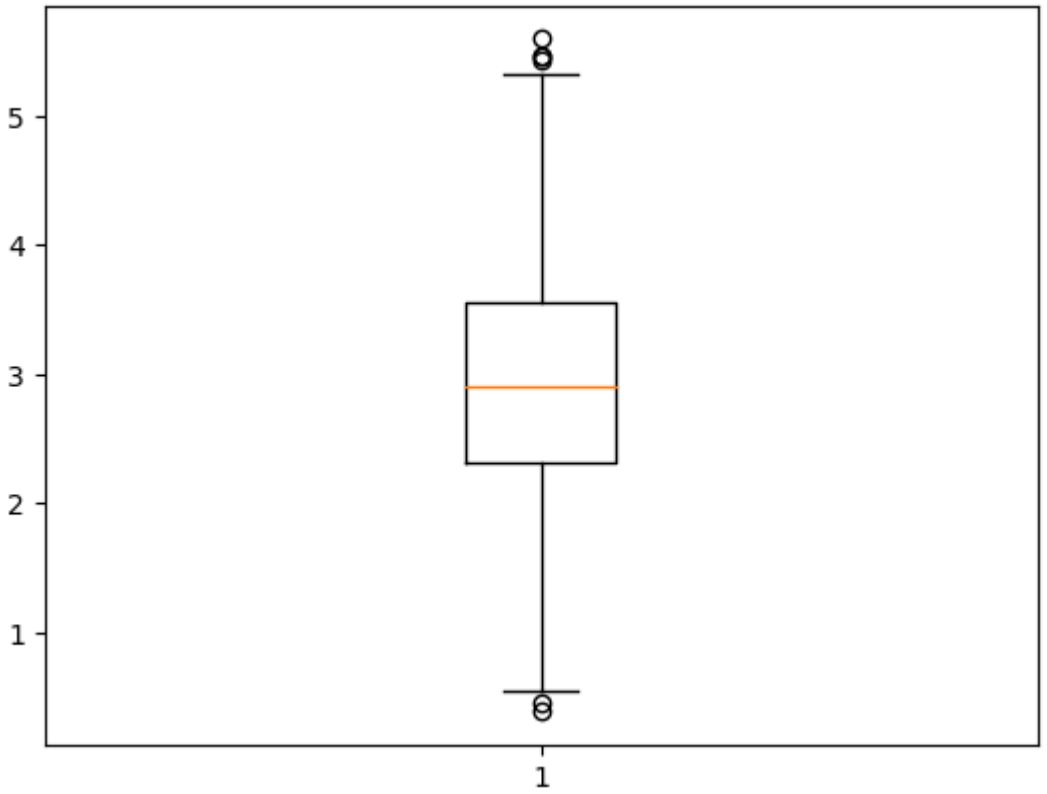
общее количество - 11;

для “Модуль упругости при растяжении, ГПа”



общее количество - 6;

для “Соотношение матрица-наполнитель”



общее количество - 6;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

	Среднее	Стандартное отклонение	min	50%	max
Соотношение матрица-наполнитель	2.93033	0.9132	0.3894	2.9068	5.5917
Плотность, кг/м3	1975.7348	73.7292	1731.7646	1977.6216	2207.7734
модуль упругости, ГПа	739.92321	330.2315	2.4369	739.6643	1911.5364
Количество отвердителя, м.%	110.57074	28.2959	17.7402	110.5648	198.9532
Содержание эпоксидных групп,%_2	22.24438	2.40630	14.2549	22.2307	33.0
Температура вспышки, С_2	285.8821	40.9432	100.0	285.8968	413.2734
Поверхностная плотность, г/м2	482.7318	281.3146	0.6037	451.8643	1399.5423
Модуль упругости при растяжении, ГПа	73.3286	3.1189	64.0540	73.2688	82.6820
Прочность при растяжении, МПа	2466.9228	485.6280	1036.8566	2459.5245	3848.4367
Потребление смолы, г/м2	218.4231	59.7359	33.8030	219.1988	414.5906
Угол нашивки, град	44.2521	45.0157	0.0	0.0	90.0

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Шаг нашивки	6.8992	2.5634	0.0	6.9161	14.4405
Плотность нашивки	57.1539	12.3509	0.0	57.3419	103.9889

Для минимизации потерь данных мы формируем отдельные датасеты под целевые переменные, так как выбросы не пересекаются.

2.2 Описание используемых методов.

Предсказание значений вещественной, непрерывной переменной — это задача регрессии. Эта зависимая переменная должна иметь связь с одной или несколькими независимыми переменными, называемых также предикторами или регрессорами. Регрессионный анализ помогает понять, как «типичное» значение зависимой переменной изменяется при изменении независимых переменных.

В настоящее время разработано много методов регрессионного анализа. Например, простая и множественная линейная регрессия. Эти модели являются параметрическими в том смысле, что функция регрессии определяется конечным числом неизвестных параметров, которые оцениваются на основе данных.

Линейная регрессия

Простая линейная регрессия имеет место, если рассматривается зависимость между одной входной и одной выходной переменными. Для этого определяется уравнение регрессии (1) и строится соответствующая прямая, известная как линия регрессии:

$y = ax + b$ (1)

Коэффициенты a и b, называемые также параметрами модели, определяются таким образом, чтобы сумма квадратов отклонений точек,

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

соответствующих реальным наблюдениям данных, от линии регрессии была бы минимальной. Коэффициенты обычно оцениваются методом наименьших квадратов.

Если ищется зависимость между несколькими входными и одной выходной переменными, то имеет место множественная линейная регрессия. Соответствующее уравнение имеет вид:

$Y = b0 + b1 * x1 + b2 * x2 + \cdots + bn * xn,$

где n - число входных переменных.

Очевидно, что в данном случае модель будет описываться не прямой, а гиперплоскостью. Коэффициенты уравнения множественной линейной регрессии

подбираются так, чтобы минимизировать сумму квадратов отклонения реальных

точек данных от этой гиперплоскости. Линейная регрессия — первый тщательно изученный метод регрессионного анализа. Его главное достоинство — простота. Такую модель можно построить и рассчитать даже без мощных вычислительных средств. С линейной регрессии целесообразно начать подбор подходящей модели.

На языке линейная регрессия реализована в бибилиотеке sklearn.linear_model.LinearRegression.

Гребневая (Ridge) регрессия

Гребневая регрессия или ридж-регрессия — так же вариация линейной регрессии, очень похожая на регрессию LASSO. Она так же применяет сжатие и хорошо работает для данных, которые демонстрируют сильную мультиколлинеарность.

Самое большое различие между ними в том, что гребневая регрессия использует регуляризацию L2, которая взвешивает ошибки по их квадрату, чтобы сильнее наказывать за более значительные ошибки.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Регуляризация позволяет интерпретировать модели. Если коэффициент близким к 0, значит данный входной признак не является значимым.

Этот метод реализован в sklearn.linear_model.Ridge.

Метод опорных векторов для регрессии

Метод опорных векторов (Support Vector Machine, SVM) — один из наиболее популярных методов машинного обучения. Он создает гиперплоскость или набор гиперплоскостей в многомерном пространстве, которые могут быть использованы для решения задач классификации и регрессии. Чаще всего он применяется в постановке бинарной классификации. Основная идея заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Интуитивно, хорошее разделение достигается за счет гиперплоскости, которая имеет самое большое расстояние до ближайшей точки обучающей выборке любого класса. Максимально близкие объекты разных классов определяют опорные вектора.

Если в исходном пространстве объекты линейно неразделимы, то выполняется переход в пространство большей размерности.

Решается задача оптимизации. Для вычислений используется ядерная функция, получающая на вход два вектора и возвращающая меру сходства между ними:

− линейная;

− полиномиальная;

− гауссовская (rbf).

Эффективность метода опорных векторов зависит от выбора ядра, параметров ядра и параметра C для регуляризации. Преимущество метода— его хорошая изученность.

Недостатки:

− чувствительность к выбросам;

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

− отсутствие интерпретируемости.

Вариация метода для регрессии называется SVR (Support Vector Regression).

В python реализацию SVR можно найти в sklearn.svm.SVR.

Градиентный бустинг

Градиентный бустинг (GradientBoosting) — ансамблевых методов. Бустинг реализует последовательное построения линейной комбинации алгоритмов. Каждый следующий алгоритм старается уменьшить ошибку предыдущего.

Чтобы построить алгоритм градиентного бустинга, нам необходимо выбрать базовый алгоритм и функцию потерь или ошибки (loss). Loss-функция – это мера, которая показывает насколько хорошо предсказание модели соответствуют данным. Используя градиентный спуск и обновляя предсказания, основанные на скорости обучения (learning rate), ищем значения, на которых loss минимальна.

Бустинг, использующий деревья решений в качестве базовых алгоритмов, называется градиентным бустингом над решающими деревьями. Он отлично работает на выборках с «табличными», неоднородными данными и способен эффективно находить нелинейные зависимости в данных различной природы. Один из самых эффективных алгоритмов машинного обучения. Широко применяется во многих задачах. Из недостатков алгоритма можно отметить только затраты времени на вычисления и необходимость грамотного подбора гиперпараметров.

Catboost

CatBoost — это алгоритм градиентного бустинга на деревьях решений. Он разработан исследователями и инженерами Яндекса и используется для поиска, рекомендательных систем, персонального

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

помощника, беспилотных автомобилей, прогнозирования погоды и многих других задач в Яндексе и в других компаниях, включая CERN, Cloudflare, Careem taxi. Он находится в открытом исходном коде и может быть использован кем угодно.

Его особенность в том, что он хорошо подходит для работы с разнородными данными. Кроме того, градиентный бустинг даёт точные результаты даже там, где данных относительно мало. Этим он отличается от нейросетей, которые обучаются на огромном массиве однородных данных.

ADABoost

AdaBoost означает «Adaptive Boosting» или адаптивный бустинг. Он превращает слабые обучающие алгоритмы в сильные. AdaBoost можно использовать для повышения производительности алгоритмов машинного обучения. Он лучше всего работает со слабыми обучающими алгоритмами, поэтому такие модели могут достигнуть точности гораздо выше случайной. Наиболее распространенными алгоритмами, используемыми с AdaBoost, являются одноуровневые деревья решений.

Нейронная сеть

Нейронная сеть (также искусственная нейронная сеть, ИНС, или просто нейросеть) — математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации биологических нейронных сетей — сетей нервных клеток живого организма. Это понятие возникло при изучении процессов, протекающих в мозге, и при попытке смоделировать эти процессы. Нейронная сеть представляет собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Такие процессоры обычно довольно просты (особенно в сравнении с процессорами,

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

используемыми в персональных компьютерах). Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам. И, тем не менее, будучи соединёнными в достаточно большую сеть с управляемым взаимодействием, такие по отдельности простые процессоры вместе способны выполнять довольно сложные задачи.

У полносвязной нейросети выход каждого нейрона подается на вход всем нейронам следующего слоя. У нейросети имеется:

− входной слой — его размер соответствует входным параметрам;

− скрытые слои — их количество и размерность определяем специалист;

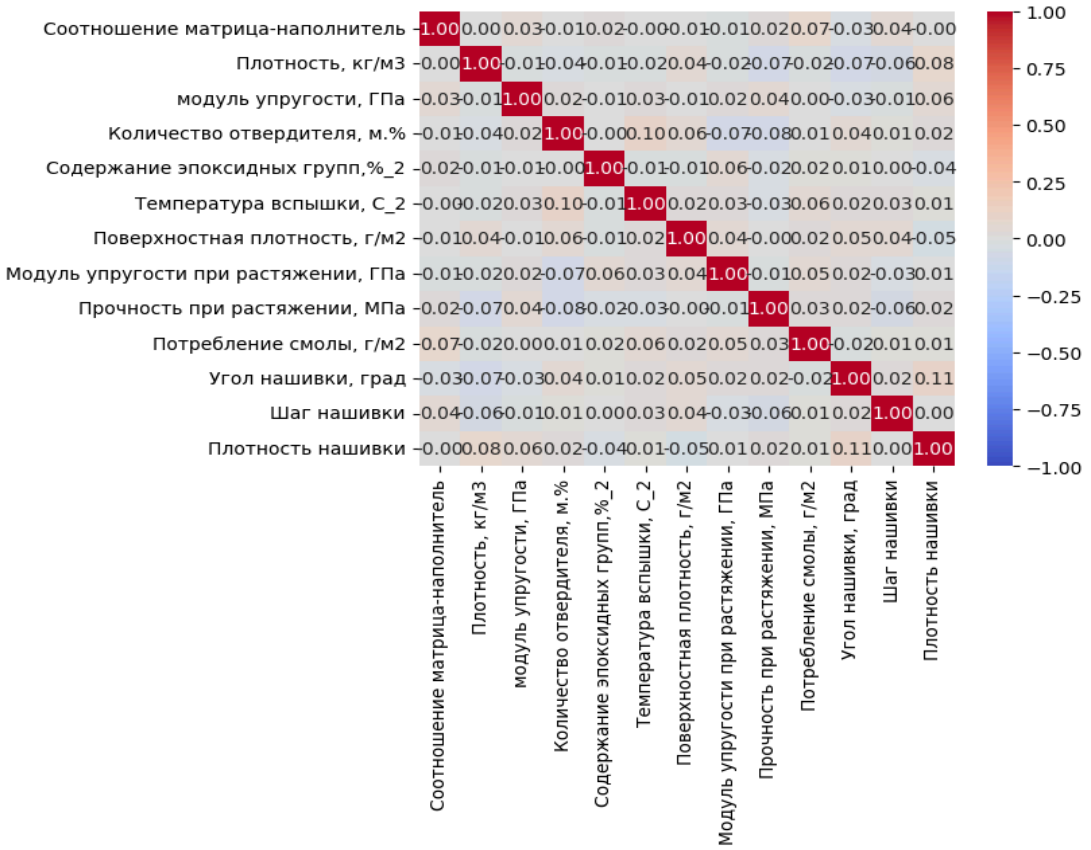
− выходной слой — его размер соответствует выходным параметрам.

Прямое распространение – это процесс передачи входных значений в нейронную сеть и получения выходных данных, которые называются прогнозируемым значением. Прогнозируемое значение сравниваем с фактическим с помощью функции потери. В методе обратного распространения ошибки градиенты (производные значений ошибок) вычисляются по значениям весов в направлении, обратном прямому распространению сигналов. Значение градиента вычитают из значения веса, чтобы уменьшить значение ошибки. Таким образом происходит процесс обучения. Обновляются веса каждого соединения, чтобы функция потерь минимизировалась. Для обновления весов в модели используются различные оптимизаторы. Количество эпох показывает, сколько раз выполнялся проход для всех примеров обучения. Нейронные сети применяются для решения задач регрессии, классификации, распознавания образов и речи, компьютерного зрения и других. На настоящий момент это самый мощный, гибкий и широко применяемый инструмент в машинном обучении.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

2.3. Разведочный анализ данных.

Выявление зависимостей на первом этапе - одна из задач стоящих перед специалистом, проведение данного этапа может существенно сократить время на разработку подходов или принятие решения о целесообразности использования той или иной модели машинного обучения. Тепловая карта корреляции рассматриваемого датасета представлена ниже:



Матрица корреляции с около нулевыми значениями наглядно демонстрирует отсутствие линейных зависимостей и сильных признаков датасета.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

3. Практическая часть

3.1. Предобработка данных.

Предварительная обработка данных необходима для корректной работы моделей машинного обучения. В нашем случае все признаки являются вещественными числами с близким к нормальному распределению, за исключением признака “Угол нашивки, град”, который принимает значения 0 и 90. Исходя из этого его можно принять за категориальный признак и воспользоваться одним из энкодеров (LabelEncoder, OrdinalEncoder), однако, указанный признак несет физическое значение и в перспективе может принять значения отличные от данных, в связи с чем принято решение использовать для всех наборов признаков MinMaxScaler(), который математическими операциями приводит значения в интервал от 0 до 1. Данную трансформацию возможно обратить соответствующим методом, при необходимости.

При использовании моделей машинного обучения для задач регрессии большую роль играет верификация моделей. В практической части это решено разделением каждого из датасета на признаки и целевую переменную, которые в свою очередь разделены на обучающую (70%) и тестовую(30%) выборки.

3.2. Разработка и обучение модели.

Модели машинного обучения, проходили обучение и тестировались на одинаковых датафреймах для соответствующих признаков. Это позволяет достоверно сравнивать метрики моделей для оценки их эффективности и принятия решения какая модель для какой целевой переменной пойдет в дальнейшем на PROD и будет использоваться при необходимости.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

В соответствии с заданием для для целевых признаков “Прочность при растяжении, МПа" и “Модуль упругости при растяжении, ГПа” применялись модели классического машинного обучения, а именно:

- Catboost
- LinearRegression
- Ridge
- GradientBoostingRegressor
- SVR(Suport vector machine)
- DecisionTreeRegressor
- ADABoostRegressor

Описание указанных моделей приводилось ранее.

Для целевого признака “Соотношение матрица-наполнитель” применялась нейросеть Sequential.

3.3. Тестирование модели.

Существует множество различных метрик качества, применимых для регрессии. В этой работе использованы:

− R2 или коэффициент детерминации измеряет долю дисперсии, объяснен-
ную моделью, в общей дисперсии целевой переменной. Если он близок к еди-
нице, то модель хорошо объясняет данные, если же он близок к нулю, то про-
гнозы сопоставимы по качеству с константным предсказанием;

− RMSE (Root Mean Squared Error) или корень из средней квадратичной
ошибки принимает значениях в тех же единицах, что и целевая переменная.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Метрика использует возведение в квадрат, поэтому хорошо обнаруживает гру-
бые ошибки, но сильно чувствительна к выбросам;

− MAE (Mean Absolute Error) - средняя абсолютная ошибка так же прини-
мает значениях в тех же единицах, что и целевая переменная;

− MSE (Mean Squared Error) это оценка среднего значения квадрата ошибок, различие между предсказанием и фактическим значением. Эту метрику удобно использовать для выявления аномалий.;

В процессе обучения и тестирования моделей были достигнуты следующие значения метрик:

Метрики моделей ML для целевого показателя "Прочность при растяжении, МПа"

	Модель ML	R^2	MAE	MSE	RMSE
0	Catboost	-0.016551	374.605949	235875.695449	485.670357
1	Base model	-0.006117	370.405644	233454.752973	483.171556
2	Ridge	-0.005958	370.369408	233417.834390	483.133351
3	GBR	-0.130478	407.950646	262310.863534	512.162927
4	SVR	-0.001587	369.486575	232403.458449	482.082419
5	Tree	-0.080094	391.281075	250619.852791	500.619469

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

6	ADABoost	-0.027919	385.143599	238513.427215	488.378365

Метрики моделей ML для целевого показателя "Модуль упругости при растяжении, ГПа"

	Модель ML	R^2	MAE	MSE	RMSE
0	Catboost	-0.019474	2.154454	7.539076	2.745738
1	Base model	-0.020051	2.162465	7.543342	2.746514
2	Ridge	-0.009862	2.142459	7.481941	2.726259
3	GBR	-0.128717	2.208512	8.346933	2.889106
4	SVR	-0.013820	2.150660	7.497260	2.738112
5	Tree	-0.374640	2.425285	10.165544	3.188345
6	ADABoost	-0.073479	2.203727	7.938440	2.817524

В процессе обучения моделей применялись различные гиперпараметры, однако существенного изменения метрик они не вносили. Итогом проверок метрик стал выбор двух моделей для реализации на PROD.

Для целевого показателя "Прочность при растяжении, МПа" выбрана модель SVR (Support vector machine), которая показала лучшие метрики, а для целевого признака "Модуль упругости при растяжении, ГПа" выбрана “Ridge”.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

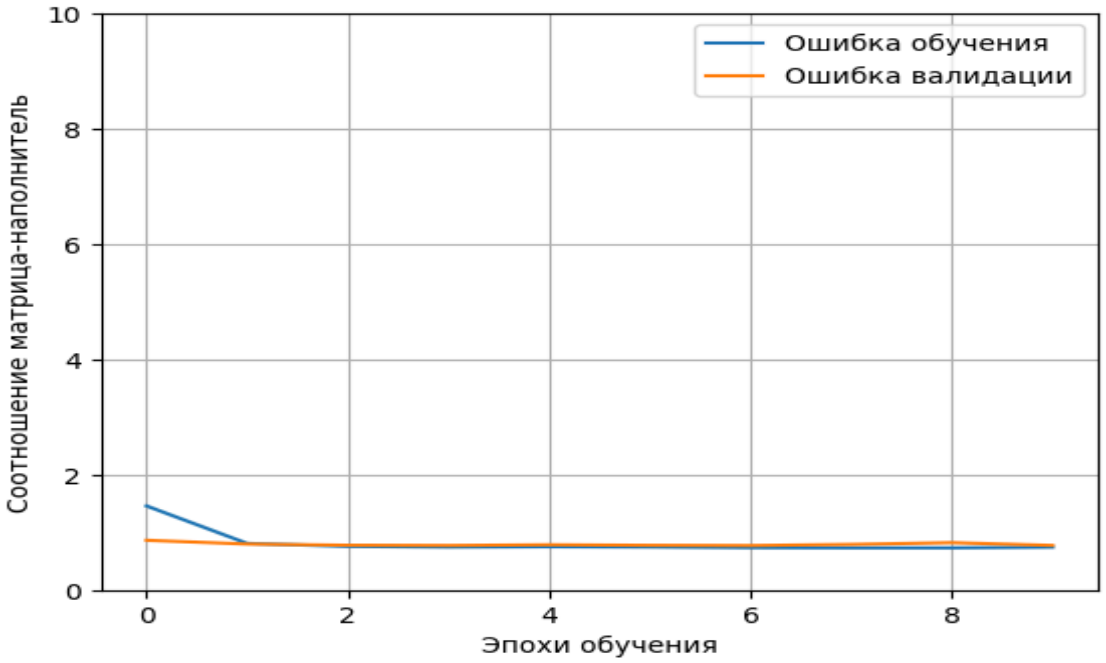
Хотя показатели данных моделей лучшие среди представленных, они все равно далеки от хороших. Данные модели не смогут оказать значимой помощи специалисту при работе с ними.

3.4. Нейронная сеть.

В соответствии с заданием для целевого признака “Соотношение матрица-наполнитель” необходимо построить нейросеть.

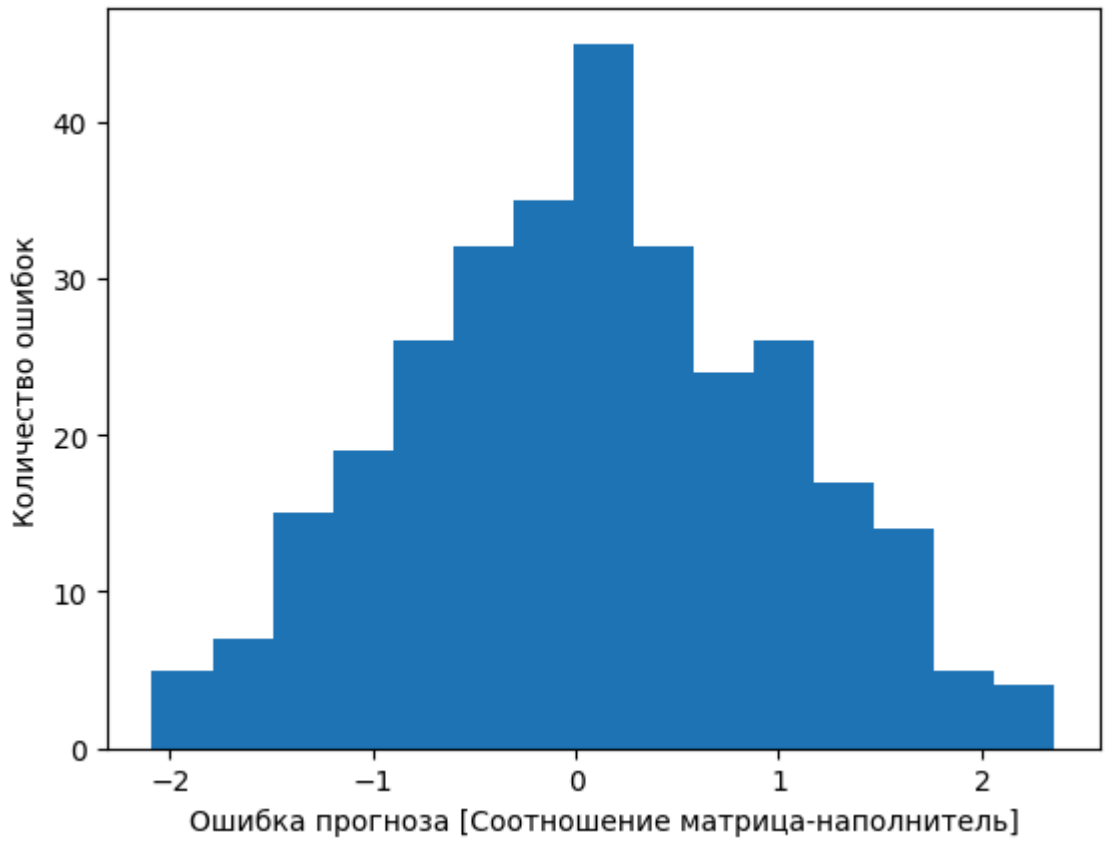
Была выбрана нейросеть Sequential из библиотеки tensorflow.keras. Данная сеть обладает широкими настройками количества последовательных слоев, чисел нейронов в слоях, активационной функции. Как ранее отмечалось, для данного целевого признака был выделен собственный датафрейм, выбросы удалены, разделен на датафрейм тестовый и валидационный наборы.

Первой итерацией была обучена базовая нейросеть, состоящая из одного слоя. Обучение происходило в соответствии с графиком:



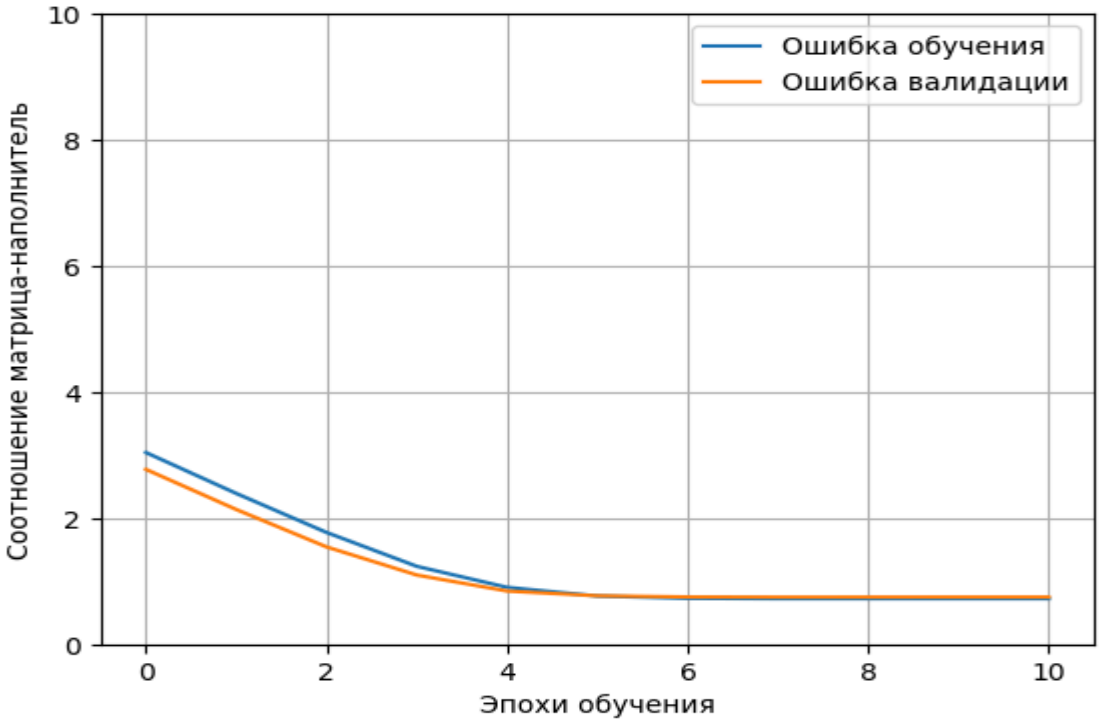
Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Распределение ошибок прогноза при проверке нейросети на тестовых данных выглядят следующим образом:

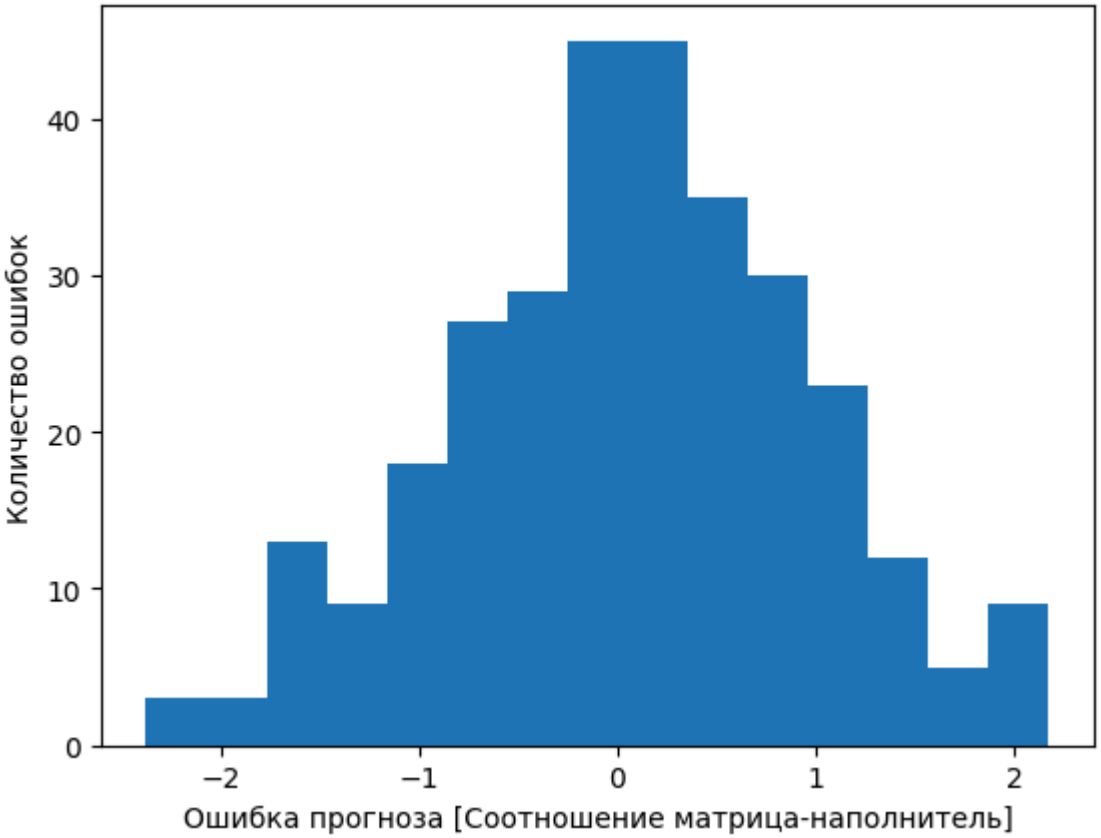


Следующим этапом нейросеть была усложнена, теперь она состоит из 4 скрытых слоев по 24 нейрона в каждом, функцией активации выбрана “sigmoid” и выходного слоя с одним нейроном и активационной функцией “linear”.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru



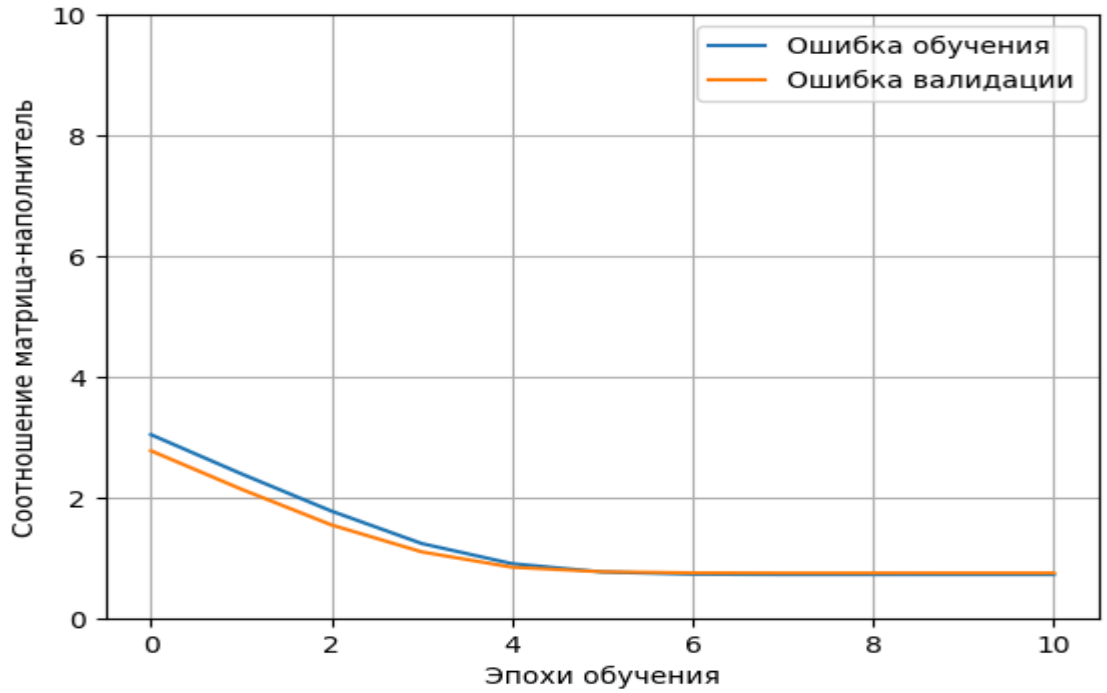
Распределение ошибок прогноза при проверке нейросети на тестовых данных выглядят следующим образом:



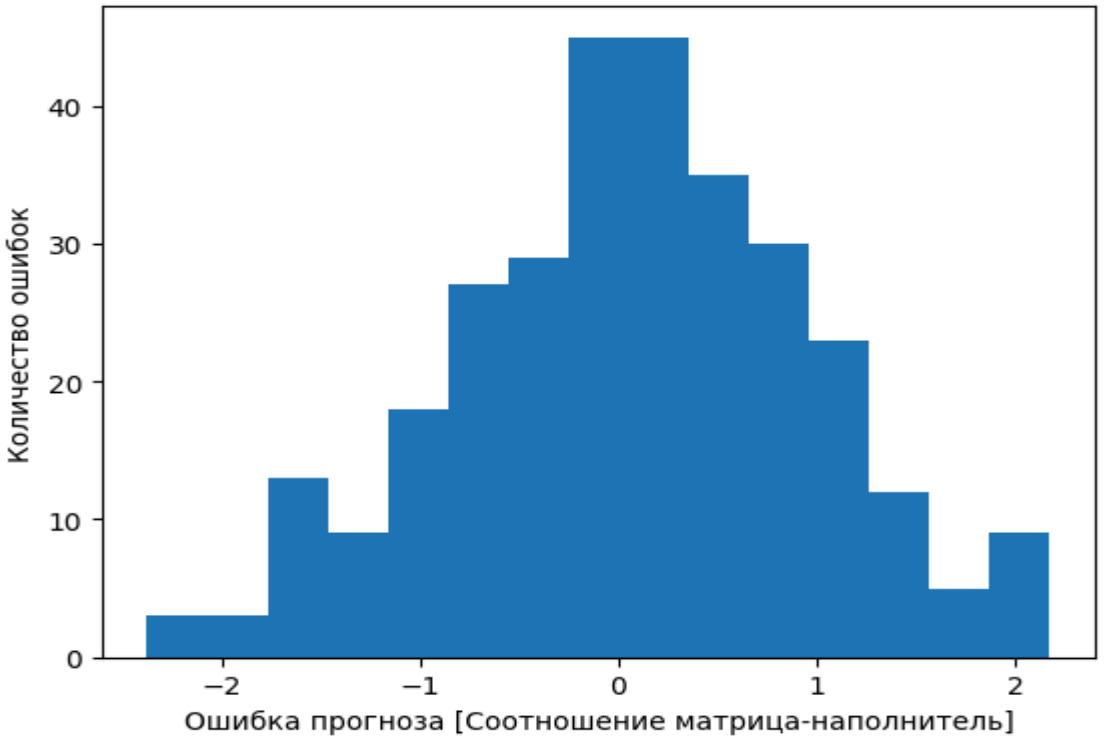
Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Новые настройки позволили добиться роста точности нейросети.

Третьим вариантом нейросети была сеть, аналогичная предыдущей, с настроенным ранним остановом обучения.



Распределение ошибок прогноза при проверке нейросети на тестовых данных выглядят следующим образом:



Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Анализ метрик нейросети позволяет говорить о значительном повышении качества последней нейросети, но все равно неудовлетворительными. Однако, по заданию, необходимо выбрать лучшую, которая пойдет на PROD.

	Нейросеть	R^2	MAE	MSE	RMSE
0	Lmodel	-0.068298	0.737604	0.837703	0.915261
1	Sequential_many_dense	-0.011836	0.703780	0.793429	0.890746
2	Earl_Sequential_many_dense	-0.005385	0.705452	0.788370	0.887902

Для добавления в приложение выбрана обученная нейросеть с ранним остановом.

3.5. Разработка приложения.

Разработка веб-приложение осуществлялась с помощью языка Python, фреймворка Flask.

Приложение доступно в сети интернет по адресу: https://vkr-baum-petrovsv.onrender.com

В приложении необходимо реализовать следующие функции:

− ввод входных параметров;

− проверка введенных параметров;

− загрузка сохраненной модели, получение и отображение прогноза выход-

ных параметров.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Задача решена, приложение доступно.

3.6. Создание удаленного репозитория.

Для данного исследования был создан удаленный репозиторий на GitHub,
который находится по адресу:
Petrovsv84/VKR_Baum (github.com)
На него были загружены результаты работы: исследовательский notebook, код приложения.

4. Заключение

Данная работа включает в себя большую часть задач, которые необходимо решать в процессе анализа данных и построения моделей. Мы провели исследование предоставленных исходных данных, выполнили разведочный анализ, осуществили подбор, настройку, обучение различных моделей машинного обучения, включая построение нейросети. Выбрали наилучшие модели и обосновали их выбор, подтвердив различными метриками качества. Разработали приложение, в котором использовали полученные модели для предсказания целевых значений, разместили его в сети интернет для доступа из любого уголка планеты. Немаловажную роль играет созданный нами репозиторий, который позволяет модернизировать наши модели в условиях сохранения версионности, делиться ими с комьюнити и просто интересующимися.

Качество полученных моделей низкое, что обусловлено, помимо прочего, характером предоставленных данных и их зависимостями. Вместе с тем исследование проведено добросовестно, осознано, и может быть использовано в

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

дальнейшем как для усовершенствования моделей, так и для определения нового набора признаков.

5. Список литературы

1. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и BigData. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
2. Документация по языку программирования python: – Режим доступа: https://docs.python.org/3.8/index.html.
3. Документация по библиотеке numpy: – Режим доступа: https://numpy.org/doc/1.22/user/index.html#user.
4. Документация по библиотеке pandas: – Режим доступа: https://pandas.pydata.org/docs/user_guide/index.html#user-guide.
5. Документация по библиотеке matplotlib: – Режим доступа: https://matplotlib.org/stable/users/index.html.
6. Документация по библиотеке seaborn: – Режим доступа: https://seaborn.pydata.org/tutorial.html.
7. Документация по библиотеке sklearn: – Режим доступа: https://scikit-learn.org/stable/user_guide.html.
8. Документация по библиотеке keras: – Режим доступа: https://keras.io/api/.
9. Руководство по быстрому старту в flask: – Режим доступа: https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html.
10. Loginom Вики. Алгоритмы: – Режим доступа: https://wiki.loginom.ru/algorithms.html.
11. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: – Режим доступа:https://habr.com/ru/company/vk/blog/513842/.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

12. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19.

13. Yury Kashnitsky. Открытый курс машинного обучения. Тема 3. Классификация, деревья решений и метод ближайших соседей: – Режим доступа: https://habr.com/ru/company/ods/blog/322534/.

14. Yury Kashnitsky. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес: – Режим доступа: https://habr.com/ru/company/ods/blog/324402/.

15. Alex Maszański. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest): – Режим доступа: https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12