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»

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

Слушатель Сидоров Дмитрий Сергеевич

Москва, 2023

Содержание

Введение

В последнее время практически во всех сферах жизнедеятельности используются разработки, созданные благодаря машинному обучению: программы, способные самостоятельно выдавать результаты анализа больших данных и их закономерностей. Технология машинного обучения на основе анализа данных берёт начало в 1950 году, когда начали разрабатывать первые программы для игры в шашки. За прошедшие десятилетий общий принцип не изменился.[1]

Для запуска процесса машинного обучения требуется набор данных, на которых алгоритм учится обрабатывать запросы. Результатом обучения является модель, выдающая готовое прогнозное значение.

В данной выпускной квалификационной рассмотрена задача регрессии на примере свойств новых композитных материалов.

Композиционный матери́ал или композитный матери́ал (КМ), сокращённо компози́т — многокомпонентный материал, изготовленный (человеком или природой) из двух или более компонентов с существенно различными физическими и/или химическими свойствами, которые, в сочетании, приводят к появлению нового материала с характеристиками, отличными от характеристик отдельных компонентов и не являющимися простой их суперпозицией. В составе композита принято выделять матрицу/матрицы и наполнитель/наполнители, последние выполняют функцию армирования (по аналогии с арматурой в таком композиционном строительном материале, как железобетон). В качестве наполнителей композитов как правило выступают углеродные или стеклянные волокна, а роль матрицы играет полимер. Сочетание разных компонентов позволяет улучшить характеристики материала и делает его одновременно

[1] Бринк Х. Машинное обучение / Х. Бринк, Дж. Ричардс, М. Феверолф. — пер. с англ. Рузмайкина И. — Санкт-Петербург: Питер, 2017. — с 2, 336 с.

лёгким и прочным. При этом отдельные компоненты остаются таковыми в структуре композитов, что отличает их от смесей и затвердевших растворов. Варьируя состав матрицы и наполнителя, их соотношение, ориентацию наполнителя, получают широкий спектр материалов с требуемым набором свойств. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик.[2]

Развитие композитных материалов уходит корнями в древний Египет, где при строительстве зданий начали использовать саманный кирпич как смесь из глинистого грунта, соломы и песка.

Прорывом в разработке новых композитов стала эпоха СССР, когда были разработаны бетон и цемент, успешно применяемый в современности. В условиях цифровой трансформации материаловедение - передовое научное направление. Революцию в материаловедении произвело распространение композиционных, или сложных неоднородных материалов, состоящих из армирующего компонента и матрицы и обладающих повышенной прочностью, легкостью и пластичностью.

В данной работе проведено исследование двух современных композитов, широко применяемых в строительстве для изготовления арматурных прутьев: базальтопластика и углепластика. Так, компонент, распределенный по всему объему материала, называют матрицей, внутри которой другие компонеты – наполнители.

[2] Композитный_материал [Электронный ресурс]: Википедия, Свободная энциклопедия – Режим доступа https://ru.wikipedia.org/wiki/Композитный_материал (дата обращения 12.03.2023)

Целью выпускной квалификационной работы является разработка Flask-приложения, которое будет прогнозировать конечные свойства новых композиционных материалов.

Задачи выпускной квалификационной работы:

1) Изучить теоретические основы и методы решения поставленной задачи.

2) Провести разведочный анализ предложенных данных: нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек; получить среднее, медианное значение; провести анализ и исключение выбросов, проверить наличие пропусков.

3) Провести предобработку данных (удалить шумы, нормализовать данные и т.д.).

4) Обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении (30% данных оставить на тестирование модели. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой с количеством блоков равным 10.)

5) Написать нейронную сеть, которая будет рекомендовать соотношение «матрица-наполнитель».

6) Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании 4 или 5 (один или два прогноза).

7) Оценить точность модели на тренировочном и тестовом датасете.

8) Создать репозиторий в GitHub / GitLab и разместить там код исследования. Оформить файл README.

Объектом исследования в выпускной квалификационной работе являются данные о начальных свойствах компонентов двух композиционных материалов – базальтопластика и углепластика.

Предметом исследования – модели, прогнозирующие свойства композиционных материалов.

Выпускная квалификационная работа состоит из двух глав: аналичиеская часть и практическая часть.

1. Аналитическая часть

1.1 Постановка задачи

В целях прогнозирования конечных свойств новых композиционных материалов для анализа данных были предоставлены два файла формата Excel: датасет X_bp.xlsx с параметрами базальтопластика, изначально состоящий из 1023 строк и 10 столбцов (рисунок 1), и датасет X_nup.xlsx с информацией об углах нашивок углепластика из 1040 строки и 3 столбцов (рисунок 2).

	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	0.0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0
1	1.0	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0
2	2.0	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0
3	3.0	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0
4	4.0	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0

Рисунок 1 – данные о X_bp.xlsx

	Unnamed: 0	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	0.0	0.0	4.0	57.0
1	1.0	0.0	4.0	60.0
2	2.0	0.0	4.0	70.0
3	3.0	0.0	5.0	47.0
4	4.0	0.0	5.0	57.0

Рисунок 2 – данные об X_nup.xlsx

В соответствии с заданием после нужно провести разведочный анализ данных, нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек.

Для каждой колонки нужно посчитать среднее и медианное значения, провести анализ и исключение выбросов, проверить наличие пропусков; обработать данные: удалить шумы и выбросы, сделать нормализацию и стандартизацию. Обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель. Разработать приложение с графическим интерфейсом, которое будет выдавать прогноз соотношения «матрица-наполнитель». Оценить точность модели на тренировочном и тестовом датасете. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

1.2 Описание используемых методов

В общем случае задача обучения по прецедентам заключается в том, чтобы по заданной выборке пар «объект-ответ» восстановить функциональную зависимость между объектами и ответами, то есть построить алгоритм, способный выдавать адекватные ответы на предъявляемые объекты. Когда множество допустимых ответов конечно, говорят о задачах классификации или распознавания образов. Когда множество допустимых ответов бесконечно, например, является множеством действительных чисел или векторов, говорят о задачах восстановления регрессии. Когда объекты соответствуют моментам времени, а ответы характеризуют будущее поведение процесса или явления, говорят о задачах прогнозирования.[3]

[3] Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / под ред. О.Б.Лупанова. 2004. Вып. 13 С. 5–36.

Все модели машинного обучения разделяются на обучение с учителем (на размеченных данных) и без учителя (на неразмеченных данных). В первую категорию входят регрессионная и классификационная модели. В целом классификация и регрессия - это методы обучения для создания моделей прогнозирования на основе собранных данных. Ключевое отличие между ними заключается в том, что в классификации размеченные данные являются категориальными и неупорядоченными, а в регрессии зависимые переменные являются непрерывными или упорядоченными целыми значениями [4]

Задача, поставленная в рамках выпускной квалификационной работы, имеет набор размеченных числовых данных и, соответственно, является задачей регрессии.

Данные представляют собой таблицу, в которой по столбцам перечислены начальные свойства (характеристики) композитных материалов, а строки содержат объекты измерений, на пересечении строк и столбцов - значение данной характеристики у данного объекта. Цель - предсказать с помощью машинного обучения конечные свойства (характеристики) композитных материалов, определить и минимизировать при этом функции потерь. Функция ошибки измеряет отклонения предсказанных значений от эмпирических (от тех, которые есть в данных)., она необходим для организации процесса обучения: чем выше значение функции ошибки, тем хуже модель соответствует имеющимся данным, хуже описывает их. Если модель полностью соответствует данным, то значение функции ошибки будет нулевым.

С целью прогноза модулей упругости при растяжении и прочности при растяжении было выбрано три метода анализа, подходящих для решения задачи регрессии:

[4] Jackson L Разница между классификацией и регрессией: – Режим доступа:. https://ru.strephonsays.com/classification-and-vs-regression-13803 (дата обращения: 12.03.2023).

– метод К-ближайших соседей (K-Neighbors Regressor);
– случайный лес (Random Forest Regressor);
– дерево решений (Decision Tree Regressor).
– стохастический градиентный спуск (SGDRegressor)

1.2.1 Метод К-ближайших соседей (K-Neighbors Regressor)

При использовании данного метода для решения поставленной задачи регрессии, объекту присваивается среднее значение по (k) ближайшим к нему объектам, значения которых уже известны. Перед применением алгоритма нужно определить функцию расстояния; классический вариант такой функции — евклидова метрика как расстояние между двумя точками евклидова пространства.[5] Метод находит расстояния между искомым запросом (свойством) и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, далее усредняет метки.

Достоинства метода ближайших соседей: позволяет настроить несколько параметров одновременно; не чувствителен к выбросам в данных; является универсальным для решения задач как с» учителем», так и без него.

Недостатки метода: замедляется с ростом объёма данных; не создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных исторических данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоёмкость.

[5] 34. Евклидова метрика [Электронный ресурс]: Википедия, Свободная энциклопедия – Режим доступа https://ru.wikipedia.org/wiki/ Евклидова_метрика (дата обращения 12.03.2023)

1.2.2 Дерево решений (Decision Tree Regressor)

Дерево решений – средство поддержки принятия решений для прогнозных моделей. Суть его работы заключается в последовательном разбиении множества данных на непересекающиеся классы, которые в свою очередь также подвергаются разбиению по каким-либо критериям с оценкой эффективности разбиения.[6]

Дерево решений состоит из «узлов», «листьев» и «веток». «Ветки» содержат записи атрибутов, от которых зависит целевая функция, «листья» – значения целевой функции, а «узлы» – остальные атрибуты, по которым происходит классификация. Чаще всего выделяют два типа деревьев: для классификации (в этом случае предсказываемый результат – класс, которому принадлежат данные) и для регрессии (результат – прогнозируемое значение целевой функции). Второй тип использовался для решения поставленной в данной работе цели.

Если говорить о достоинствах деревьев решений, то можно выделить следующие. Во-первых, простота понимания и интерпретации. Во-вторых, минимальные требования к подготовке данных, а также способность работы с большими объемами данных. В-третьих, метод одинаково хорошо работает с разными видами признаков. В-четвертых, является надежным методом и позволяет оценить модель статистическими тестами.

Недостатки данного метода: подверженность переобучению; не для всех задач может быть получено решение удовлетворительного качества.[7]

[6] Паклин, Н. Б. Глава 9, // Бизнес-аналитика: от данных к знаниям : учебное пособие / Н. Б. Паклин, В. И. Орешков. – 2-е изд.. – СПб. : Питер, 2013. – С. 444.

[7] Воронина В. В., Михеев А. В., Ярушкина Н. Г., Святов К. В. // Теория и практика машинного обучения : учебное пособие /В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. –Ульяновск : УлГТУ, 2017. –. – С. 56.

1.2.3 Случайный лес (Random Forest Regressor)

Метод является универсальным для большинства моделей машинного обучения как «с учителем», так и без него. Данный представитель ансамблевых методов успешно применяется для решения задач кластеризации, классификации, регрессии. Random Forest Regressor представляет собой множество (лес) независимых деревьев решений.

Дерево решений строится на основе всего набора данных с использованием всех интересующих объектов (переменных), в то время как случайный лес случайным образом выбирает наблюдения (строки) и конкретные объекты (переменные) для построения нескольких деревьев решений, а затем усредняет результаты.

Достоинства данного метода: делает достаточно точные предсказания; обрабатывает пропуски в наборе данных; не переобучается; не требует предварительной обработки входных данных; хорошо масштабирует.

Таким образом, сделан вывод о том, что при выборе модели для решения задачи регрессии следует учитывать, что любая может выдавать совершенно разные значения для одних и тех же входных данных, если в функции будут разные параметры. Основная цель алгоритма обучения - подобрать значения параметров таким образом, чтобы для объектов обучающей выборки, для которых мы уже знаем правильные ответы, предсказанные значения были как можно ближе к тем, которые есть в датасете, истинным значениям.

1.2.4 Стохатический градиентный спуск (SGDRegressor)

Стохастический градиентный спуск (SGD) — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под

выпуклые функции потерь, такие как (линейные). Метод опорных векторов и логистическая регрессия

Несмотря на то, что SGD существует в сообществе машинного обучения уже давно, совсем недавно он привлек значительное внимание в контексте крупномасштабного обучения.

SGD успешно применяется для решения крупномасштабных и разреженных задач машинного обучения, часто встречающихся при классификации текста и обработке естественного языка. Учитывая, что данные немногочисленны, классификаторы в этом модуле легко масштабируются для решения задач с более чем 105 обучающими примерами и более чем 105 функциями.

Строго говоря, SGD — это просто метод оптимизации и не соответствует конкретному семейству моделей машинного обучения. Это всего лишь *способ* обучить модель.

1.3 Разведочный анализ данных

Разведочный анализ данных (англ. exploratory data analysis, EDA) — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации.[8]

Цель разведочного анализа – представить наблюдаемые данные компактной и простой форме, позволяющей выявить имеющиеся в них закономерности и связи.

Разведочный анализ включает преобразование данных и способы наглядного их представления, выявление аномальных значений, грубую оценку типа распределения, сглаживание. Термин разведочный анализ применяется

[8] П. Брюс, Э. Брюс. 1. Разведочный анализ данных // Практическая статистика для специалистов Data Science. — СПб.: БХВ-Петербург, 2018. — С. 19—58. — 304 с.

также в более широком смысле, чем предварительная обработка данных. Так, в многомерных процедурах, таких как факторный анализ, многомерное шкалирование данных, цель разведочного анализа, кроме анализа первичных данных, заключается в определении минимального числа факторов, которые удовлетворительно воспроизводят ковариационную (корреляционную) матрицу или матрицу близостей наблюдаемых переменных. В процессе разведочного анализа в данных определяется структура, основные переменные, зависимость между отдельным характеристиками, отклонения и аномалии в массиве данных, количество пропущенных и уникальных значений. На этапе разведки данные группируются, разъединяются, отображаются в таблицах, графиках, матрицах и рисунках.

Использование при «разведке» функции описательной статистики позволяет определить основные статистические характеристики: среднее, медианное, минимальное и максимальное значения.

С помощью графиков плотности распределения данных и «ящиков с усами» определяются тип распределения и наличие аномальных выбросов.

Следует отметить высокую важность разведочного анализа в Data Science, поскольку именно этот этап машинного обучения требует внушительных временных затрат, а от его результата зависит выбор дальнейшего направления работы.

Задачи машинного обучения с учителем как правило состоят в восстановлении зависимости между парами в наборе исходных данных. Алгоритмы машинного обучения служат для построения модели, аппроксимирующей эту зависимость. Оценивается качество работы алгоритмов с помощью метрик. С целью оценки выбранных для решения задачи регрессионного анализа алгоритмов используются три основных метрики:

- MAE;

- MSE;
- R2.

MAE - метрика, которая вычисляет среднюю абсолютную разницу между прогнозируемыми значениями и фактическими значениями в наборе данных. Чем ниже MAE, тем лучше модель соответствует набору данных.

MSE - метрика, которая вычисляет среднеквадратичную разницу между прогнозируемыми значениями и фактическими значениями в наборе данных. Чем ниже MSE, тем лучше модель соответствует набору данных.

Коэффициент детерминации (R2) - доля дисперсии (вариации) целевой переменной, объясненная данной моделью. Если модель всегда предсказывает идеально, данная метрика будет равна 1. Когда предсказания модели сводятся к среднему значению, а метрика будет равна 0, такую модель назовем тривиальной. Если модель хуже идеальной, но лучше тривиальной, то метрика будет в диапазоне от 0 до 1, причем чем ближе к 1 - тем лучше. Если же модель предсказывает такие значения, что отклонения их от теоретических получаются больше, чем от среднего значения, то метрика будет принимать отрицательные значения, это означает, что модель хуже, чем тривиальная

1. Практическая часть

1.1. Предобработка данных

В соответствии с заданием к выпускной квалификационной работе объединение двух датасетов сделано по индексу с типом объединения INNER.

	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
0	1.857143	2030.000000	738.736842	30.000000	22.267857	100.000000	210.000000	70.000000	3000.000000	220.000000	0.0
1	1.857143	2030.000000	738.736842	50.000000	23.750000	284.615385	210.000000	70.000000	3000.000000	220.000000	0.0
2	1.857143	2030.000000	738.736842	49.900000	33.000000	284.615385	210.000000	70.000000	3000.000000	220.000000	0.0
3	1.857143	2030.000000	738.736842	129.000000	21.250000	300.000000	210.000000	70.000000	3000.000000	220.000000	0.0
4	2.771331	2030.000000	753.000000	111.860000	22.267857	284.615385	210.000000	70.000000	3000.000000	220.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...
1018	2.271346	1952.087902	912.855545	86.992183	20.123249	324.774576	209.198700	73.090961	2387.292495	125.007669	90.0
1019	3.444022	2050.089171	444.732634	145.981978	19.599769	254.215401	350.660830	72.920827	2360.392784	117.730099	90.0
1020	3.280604	1972.372865	416.836524	110.533477	23.957502	248.423047	740.142791	74.734344	2662.906040	236.606764	90.0
1021	3.705351	2066.799773	741.475517	141.397963	19.246945	275.779840	641.468152	74.042708	2071.715856	197.126067	90.0
1022	3.808020	1890.413468	417.316232	129.183416	27.474763	300.952708	758.747882	74.309704	2856.328932	194.754342	90.0

1023 rows × 13 columns

Рисунок 1 – объединенный датасет: ds

При использовании данного метода объединяются только те значения, которые можно найти в обеих таблицах. Так, семнадцать строк об углепластике из X_nup.xlsx были удалены, поскольку не имели соотношений в таблице с данными о бальзатопластике X_bp.xls. Объединенный датасет из 13 столбцов и 1023 строк представлен на рисунке 3 и назван dat.

С помощью функции info была проведена оценка данных: количество строк и столбцов, название столбцов, тип данных и количество ненулевых значений данных в каждом из столбцов (рисунок 4):

```
ds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1023 entries, 0 to 1022
Data columns (total 13 columns):
 #   Column                                 Non-Null Count  Dtype  
---  ------                                 --------------  -----  
 0   Соотношение матрица-наполнитель        1023 non-null   float64
 1   Плотность, кг/м3                       1023 non-null   float64
 2   модуль упругости, ГПа                  1023 non-null   float64
 3   Количество отвердителя, м.%            1023 non-null   float64
 4   Содержание эпоксидных групп,%_2        1023 non-null   float64
 5   Температура вспышки, С_2               1023 non-null   float64
 6   Поверхностная плотность, г/м2          1023 non-null   float64
 7   Модуль упругости при растяжении, ГПа   1023 non-null   float64
 8   Прочность при растяжении, МПа          1023 non-null   float64
 9   Потребление смолы, г/м2                1023 non-null   float64
 10  Угол нашивки, град                     1023 non-null   float64
 11  Шаг нашивки                            1023 non-null   float64
 12  Плотность нашивки                      1023 non-null   float64
dtypes: float64(13)
memory usage: 111.9 KB
```

Рисунок 4 – Типы данных

Данные объединенного датасета представлены числами, категориальных признаков нет, данные со значением «NaN» также отсутствуют. Преобразования

типов не требуется, так как для машинного обучения и требуются числовые признаки.

Практически все значения являются уникальными (рисунок 5) за исключением столбца с углом нашивки, которые представлены всего двумя значениями:0 и 90 градусов.

```
Соотношение матрица-наполнитель         1014
Плотность, кг/м3                        1013
модуль упругости, ГПа                   1020
Количество отвердителя, м.%             1005
Содержание эпоксидных групп,%_2         1004
Температура вспышки, С_2                1003
Поверхностная плотность, г/м2           1004
Модуль упругости при растяжении, ГПа    1004
Прочность при растяжении, МПа           1004
Потребление смолы, г/м2                 1003
Угол нашивки, град                         2
Шаг нашивки                              989
Плотность нашивки                        988
dtype: int64
```

Рисунок 5 – Уникальные значения

Рисунок 6 – датасет после предварительной подготовки

Анализ показал близость по значению среднего и медианного значений (рисунок 7).

	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
mean	2.930366	1975.734888	739.923233	110.570769	22.244390	285.882151	482.731833	73.328571	2466.922843	218.423144
50%	2.906878	1977.621657	739.664328	110.564840	22.230744	285.896812	451.864365	73.268805	2459.524526	219.198882

Рисунок 7 – Уникальные значения

Данное наблюдение позволяет сделать вывод о том, что наше распределение не просто нормальное, а практически симметричное., поскольку для симметричных распределений эти две стороны совпадают.

Отсутствие пропусков в данных было установлено с помощью функции ds.isnull().sum(),

Гистограммы распределения признаков показали, что они близки к нормальному,

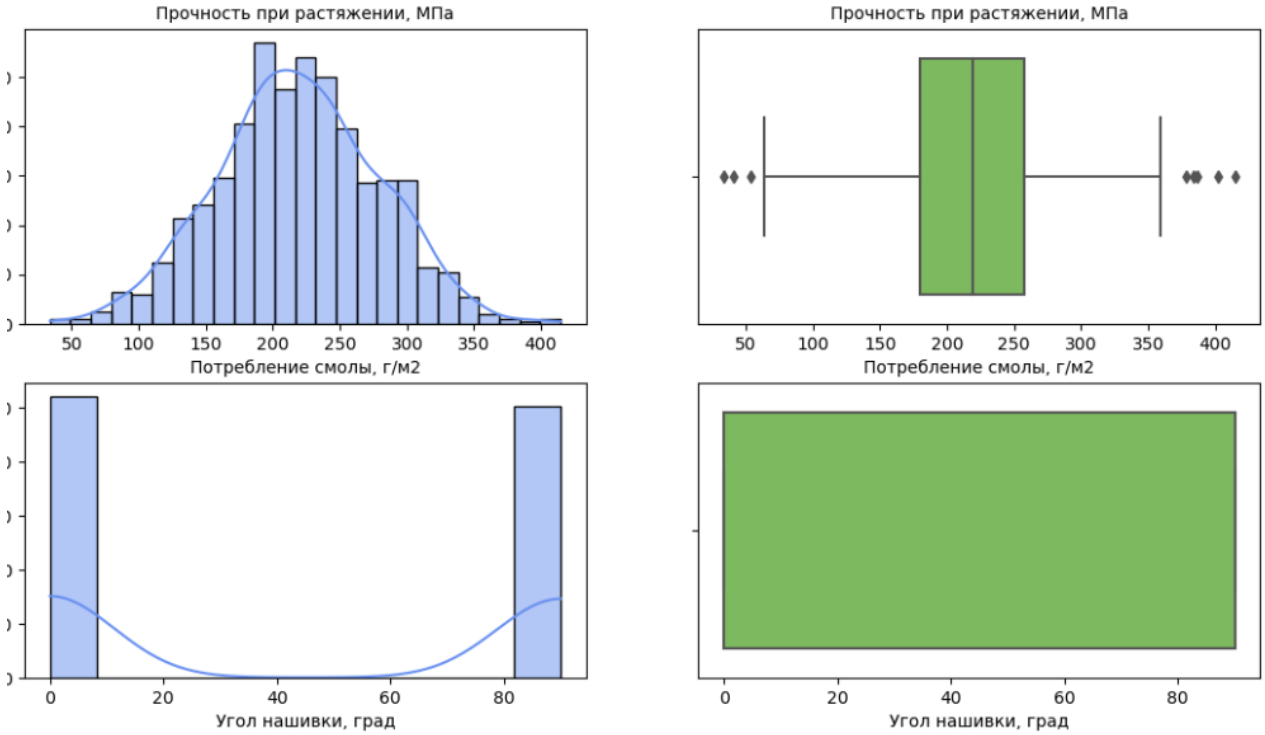

Рисунок 9 – Гистограммы распределения признаков

за исключением преобразованного столбца с углом нашивки (рисунок 9).

Данное наблюдение ставит под сомнение реальность исходных данных.

Работа с выбросами проводилось с использованием диаграмм «Ящик с усами» (рисунок 10):

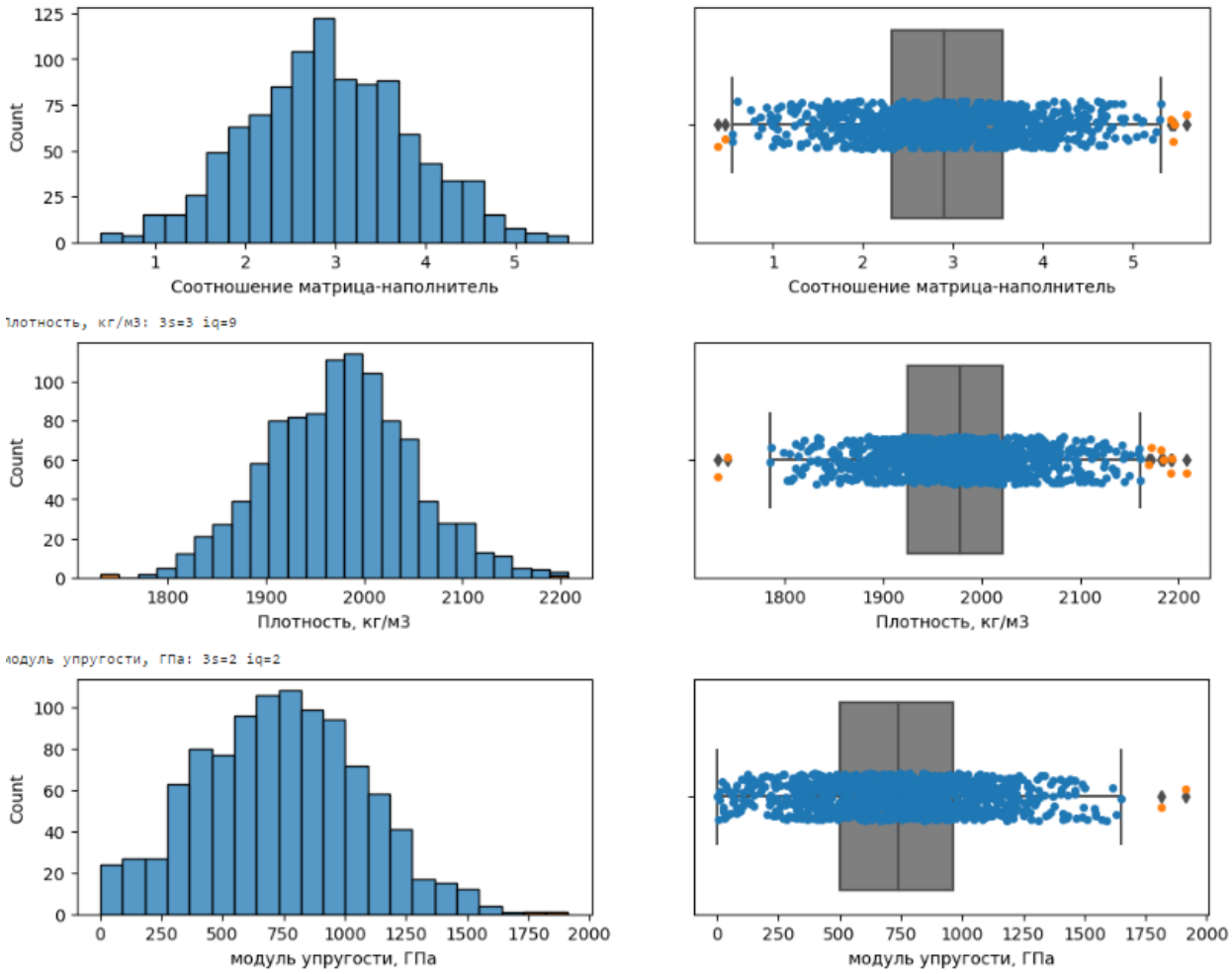

Рисунок 10 – «Ящики с усами»

Подсчет и удаление выбросов осуществлены один раз методом 3сигм для сохранения большего обьема информации

Корреляция данных посредством функции dat.corr() была практически не обнаружена. Максимальное значение – 0,11.

В целях анализа данных была произведена оценка плотности ядра и анализ распределения каждого признака (рисунок 11):

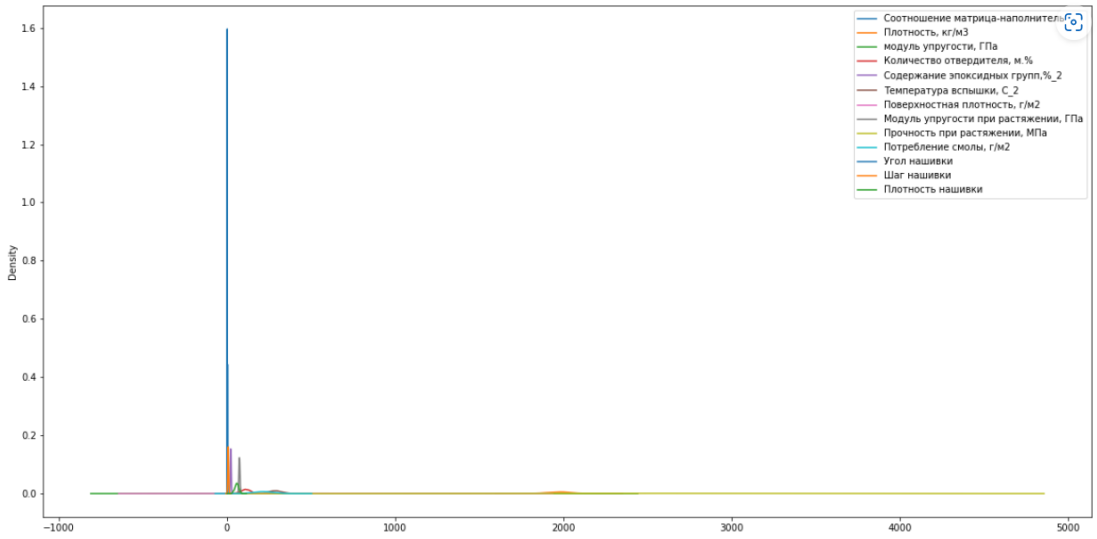

Рисунок 11 – Оценка плотности ядра

По причине нахождении данных в разных диапазонах оценку плотности ядра провести достаточно сложно.

Нормализация данных с целью приведения всех признаков к одному порядку была проведена с помощью методов fit_transform и RobustScaler), полученные в результате преобразования данные представлены в диапазоне от -2.5 до 2.5 (рисунок 12):

5]:

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1000.0	0.022264	0.736063	-2.040588	-0.477487	-1.798454e-16	0.522513	2.173033
Плотность, кг/м3	1000.0	-0.019671	0.748007	-1.977208	-0.550517	0.000000e+00	0.449483	2.208715
модуль упругости, ГПа	1000.0	-0.005365	0.710699	-1.602835	-0.521559	-1.233280e-16	0.478441	1.970734
Количество отвердителя, м.%	1000.0	0.004535	0.746589	-2.161747	-0.485646	0.000000e+00	0.514354	2.202000
Содержание эпоксидных групп,%_2	1000.0	0.004153	0.702967	-1.923987	-0.483025	0.000000e+00	0.516975	1.985333
Температура вспышки, С_2	1000.0	0.001916	0.746060	-2.083791	-0.496059	5.270307e-16	0.503941	2.184484
Поверхностная плотность, г/м2	1000.0	0.068275	0.654125	-1.060573	-0.433143	6.700369e-17	0.566857	1.979675
Модуль упругости при растяжении, ГПа	1000.0	0.021532	0.763612	-2.250338	-0.485942	-1.742530e-15	0.514058	2.317866
Прочность при растяжении, МПа	1000.0	0.013538	0.775247	-2.268983	-0.514457	-3.634333e-16	0.485543	2.225035
Потребление смолы, г/м2	1000.0	-0.005713	0.758988	-2.287452	-0.500697	1.830133e-16	0.499303	2.165854
Угол нашивки, град	1000.0	0.496000	0.500234	0.000000	0.000000	0.000000e+00	1.000000	1.000000
Шаг нашивки	1000.0	-0.003329	0.734346	-1.976602	-0.521495	0.000000e+00	0.478505	2.158562
Плотность нашивки	1000.0	-0.013012	0.787728	-2.453821	-0.503992	0.000000e+00	0.496008	2.360136

Рисунок 12 – Нормализованный датасет без выбросов

С целью наглядного представления распределения данных по характеристикам были использованы «ящики с усами», что, в сравнении с подобным распределением ненормализованных данных, выглядит гораздо понятнее показано (рисунок 13).

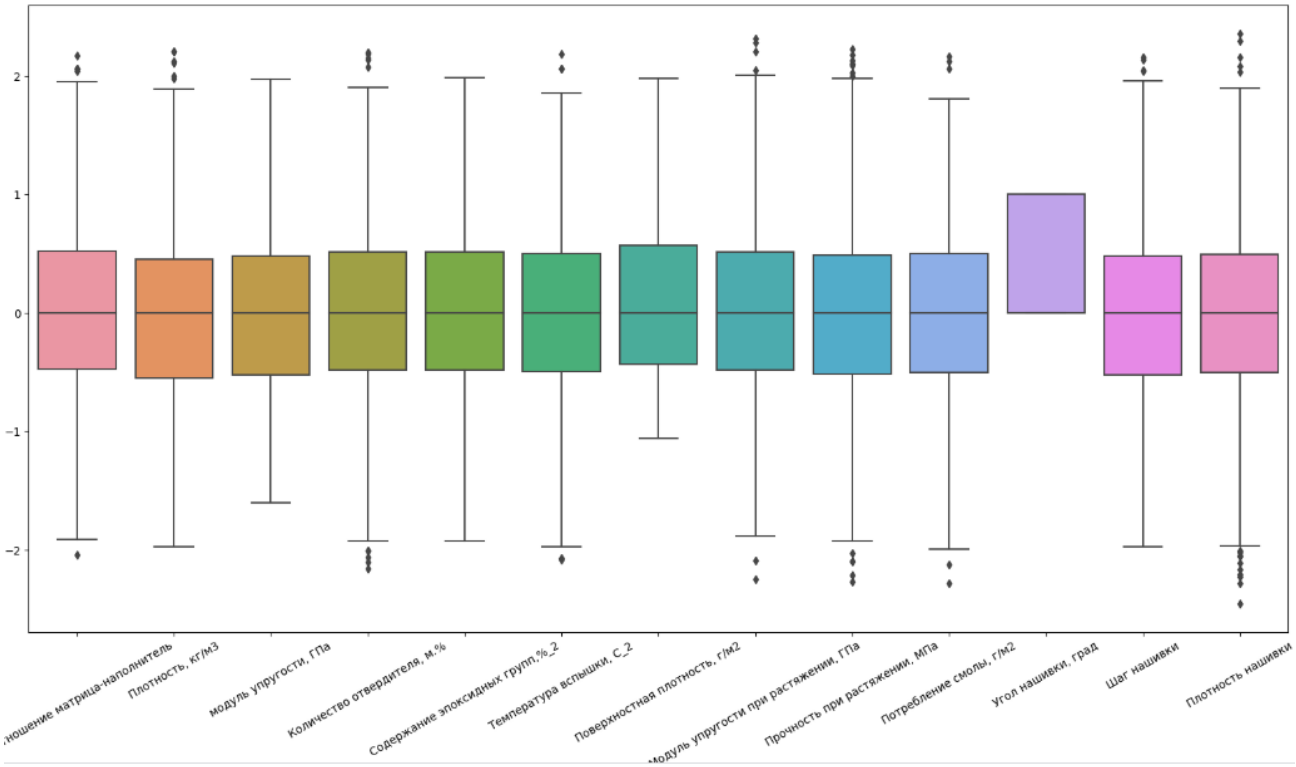

Рисунок 13 Визуализация распределения данных по характеристикам

Нормализация данных позволила также адекватно оценить плотность ядра (рисунок 14):

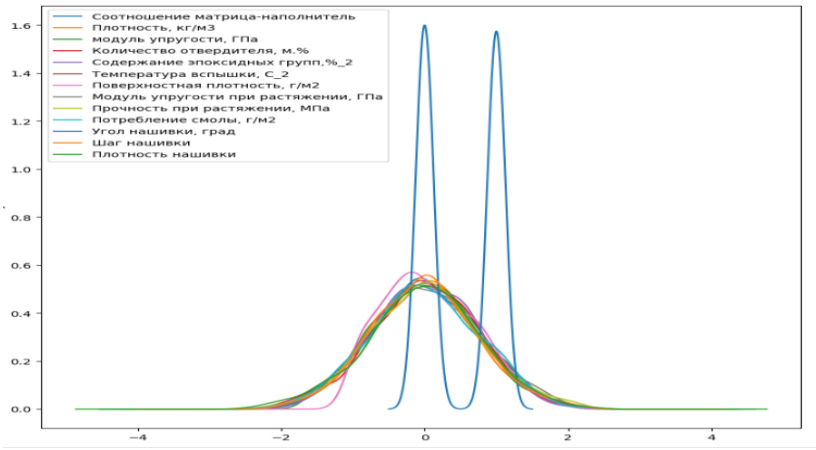

Рисунок 14 – Распределение плотности ядра

Подтвердилось отсутствие корреляции между признаками и после нормализации, это подтверждают графики рассеяния точек (рисунок 15):

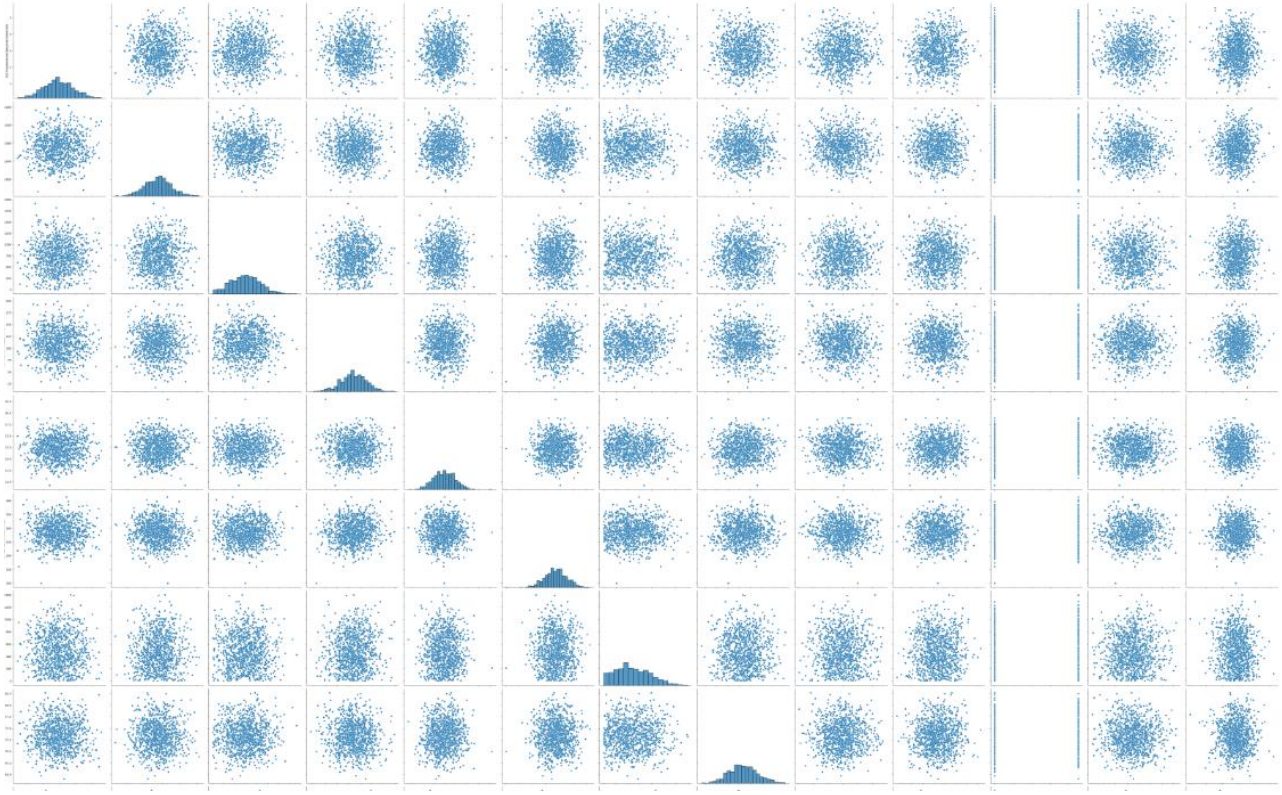

Рисунок 15 - Попарные графики рассеяния точек

Подготовленный и нормализованный датасет без выбросов (рисунок 16) был сохранен для разработки алгоритмов обучения, прогнозирующих значения модулей прочности и упругости при растяжении композитного материала.

	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
count	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000
mean	0.499412	0.502904	0.451341	0.506200	0.490578	0.516739	0.373295	0.487343	0.503776	0.507876	0.510846
std	0.187858	0.188395	0.201534	0.186876	0.180548	0.190721	0.217269	0.196366	0.188668	0.199418	0.500154
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.371909	0.368184	0.305188	0.378514	0.366571	0.386228	0.204335	0.353512	0.373447	0.374647	0.000000
50%	0.495189	0.511396	0.451377	0.506382	0.488852	0.516931	0.354161	0.483718	0.501481	0.510143	1.000000
75%	0.629774	0.624719	0.587193	0.638735	0.623046	0.646553	0.538397	0.617568	0.624299	0.642511	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Рисунок 16 – Нормализованный датасет для разработки моделей

1.2. Разработка и обучение модели

В соответствии с заданием, при построении модели необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. После проверки работы алгоритма на стандартных параметрах, необходимо провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.

Для предсказания модуля упругости и прочности при растяжении было обучено шесть моделей машинного обучения трех видов: K-Neighbors Regressor; Random Forest Regressor; Decision Tree Regressor, SGDRegressor результаты работы моделей представлены на рисунке 17.

Оба выходных параметра прогнозировались отдельно друг от друга, однако результаты обучения моделей, прогнозирующих оба параметра близки по значению, поэтому для примера покажем визуализацию работы алгоритмов, прогнозирующих модуль упругости при растяжении:

Рисунок 17 – Прогноз модуля упругости при растяжении со стандартными параметрами

В процессе разработки модели для прогноза модуля упругости и прочности при растяжении нормализованные данные были разделены на две части: согласно заданию. Данные были разделены. Первая выборка была обучающей, она составляла примерно 70% от общего числа данных. На ее примере мы и строили алгоритмы обучения. Вторая выборка – тестирующая, или проверочная, – составляла оставшиеся 30% данных. По ней проводилась оценка качества модели путем сравнения этих данных с прогнозом, сделанным по построенной модели.

По каждой из моделей был проведен поиск сетки гиперпараметров для оптимизации, сравнение работы всех моделей.

1.3. Тестирование модели

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовых выборках. Результаты работы регрессионных моделей, предсказывающих значения модуля упругости и модуля прочности при растяжении (по отдельности) представлены в Таблице 1.

Для анализа были использованы следующие метрики:

- MAE;
- MSE;
- R2.

Из таблицы видно, что самые низкие среднеквадратичные ошибки, как показатель, считающий среднее расстояние между прогнозируемыми значениями из модели и фактическими значениями в наборе данных, показали все регрессионные модели с гиперпараметрами.

По данным таблицы видно, что мы имеем сравнительно небольшую среднеквадратичную ошибку (MSE) и среднюю абсолютную ошибку (MAE). Однако коэффициент детерминации показывают, что полученные модели объясняют большую часть данных, но прогноз каждой точным не будет.

Таблица 1 – Результаты обучения регрессионных моделей

	Регрессор	MAE	MSE	Качество обучения	Точность предсказания
0	KNeighbors	0.63	0.62	0.0	0.006
1	RandomForest	0.63	0.62	-0.008	0.029
2	DecisionTree	0.66	0.65	-0.059	0.057
3	SGDRegressor	0.63	0.62	-0.001	0.018

Модуль упругости при растяжении

	Регрессор	MAE	MSE	Качество обучения	Точность предсказания
0	KNeighbors	0.63	0.62	0.0	0.006
1	RandomForest	0.63	0.62	-0.008	0.029
2	DecisionTree	0.66	0.65	-0.059	0.057
3	SGDRegressor	0.63	0.62	-0.001	0.018

Модуль прочности при растяжении

Анализ работы регрессионных моделей показал отрицательное значение коэффициента детерминации, что в очередной раз подтверждает крайне слабую степень линейной зависимости между переменными, за исключением метода ближайших соседей.

1.4. Написание нейронной сети, рекомендующей соотношение «матрица – наполнитель»

Для описания алгоритмов и устройств в нейроинформатике выработана специальная «схемотехника», в которой элементарные устройства (сумматоры, синапсы, нейроны и т.п.) объединяются в сети, предназначенные для решения задач.[9]

Слоистые сети: нейроны расположены в несколько слоев. Нейроны первого слоя получают входные сигналы, преобразуют их и через точки ветвления передают нейронам второго слоя. Далее срабатывает второй слой и т.д. до k-го слоя, который выдает выходные сигналы для интерпретатора и пользователя. Если не оговорено противное, то каждый выходной сигнал i-го слоя подается на вход всех нейронов i+1-го. Число нейронов в каждом слое может быть любым и никак заранее не связано с количеством нейронов в других слоях.

[9] Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей / А.Н. Горбань // Сиб. журн. вычисл. математики. – 1998. – Т. 1, № 1. – 21 с.

Закрытый процесс самообучения происходит через повторную активацию некоторых нейронных соединений. Так увеличивается вероятность вывода нужного результата при соответствующей входной информации. Такой вид обучения использует обратную связь - при правильном результате нейронные связи, которые выводят его, становятся более плотными.

При написании модели нейронной сети был использован стандартный способ подачи входных сигналов: все нейроны первого слоя получают каждый входной сигнал. Это трехслойная сеть, в которой каждый слой имеет свое наименование: первый входной, второй скрытый, третий выходной

В процессе обучения была предпринята попытка минимизировать потери функции, параметры обновлялись для повышения точности.

Построение нейронной сети было проведено с помощью класса keras.Sequential. Дополнительная нормализация данных не проводилась, поскольку датасет был нормализован ранее в процессе обработки, и на вход нейросеть получила нормализованные от 0 до 1 данные.

С целью решения поставленной в выпускной квалификационной работе задачи было решено использовать полносвязную нейронную сеть (feed forward neural network) (рисунок 18):

```
# Создание модели
model = Sequential()

# Добавление скрытого слоя с 10 нейронами
model.add(Dense(10, input_dim=x_train.shape[1], activation='selu'))
model.add(Dense(8, activation='selu'))
model.add(Dense(1))
```

Рисунок 18 – Архитектура нейронной сети

Так, нейронная сеть действительно представляет комбинацию нескольких слоев: входной слой, куда передаются все данные датасета; с функциями

активации selu ; выходной слой с функцией активации selu одним линейным нейроном.

Визуально качество работы нейросети можно оценить на рисунке 19:

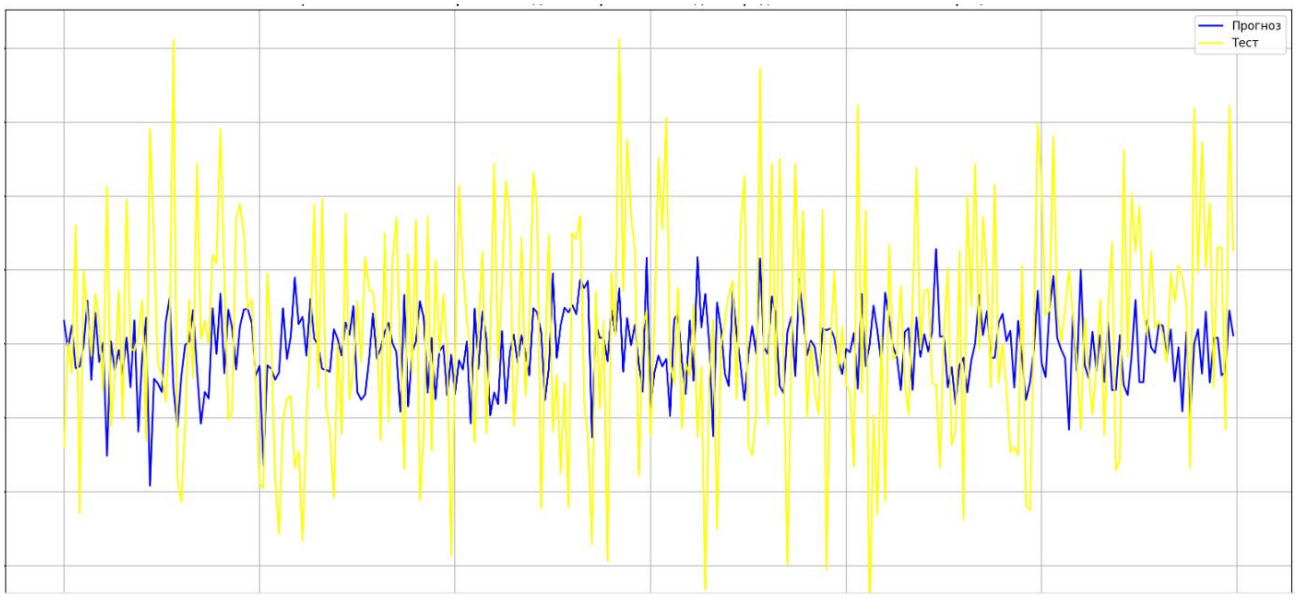

Рисунок 19 – Работа нейронной сети

Для достижения лучших результатов при компиляции сети был выбран оптимизатор RMSProp, который ограничивает колебания в вертикальном направлении и, соответственно, позволяет увеличить скорость обучения алгоритма, поскольку он идет быстрее в горизонтальном направлении.

В качестве функции ошибки (функции потерь нейронной сети) как математической дифференцируемой функции, характеризующей разницу между «истинным» значением целевой переменной и предсказанным нейронной сетью значением, была выбрана MSE (mean square error, средне-квадратичная ошибка).

Метрикой избран MAE (средний модуль ошибки mean absolute error).

Результат работы нейросети представлен в таблице 2:

Таблица 2 – Результаты обучения и тестирования работы нейронной сети

	Модель	MAE	MSE	R2
0	NS_train	0.533961	0.450960	0.180901
1	NS_test	0.625225	0.604375	-0.170581

Таким образом, можно сделать вывод, что по функциям ошибки нейросеть показала результат лучший, чем любая из моделей регрессии. Коэффициент детерминации выдал результаты, близкие по значению к рассчитанным в предыдущих моделях.

2.5 Разработка приложения

2.6. Создание удаленного репозитория и загрузка результатов работы

В соответствии с заданием на github.com был создан репозиторий: https://github.com/DimonSidorov/vkr

Рисунок 20 – Репозиторий на github.com

Заключение

В ходе исследования были разработаны и обучены несколько моделей нейронных сетей для предсказания значений целевых признаков. Однако, несмотря на это, удалось получить только относительно низкие значения коэффициента детерминации, что говорит о том, что модели не смогли точно предсказать значения целевых признаков. Это может быть связано с недостаточным количеством данных, не оптимальным выбором архитектуры нейронной сети, а также не оптимальным подбором параметров модели. Для улучшения качества модели, возможно, стоит рассмотреть следующие шаги: – сбор большего количества данных для обучения моделей; – использование более сложных архитектур нейронных сетей, с большим количеством слоев и нейронов; – использование более продвинутых методов оптимизации весов и настройки параметров модели; – анализ выборки данных для выявления выбросов, отсутствия данных и других ошибок. Таким образом, дальнейшее исследование и оптимизация моделей нейронных сетей могут привести к более точным предсказаниям значений целевых признаков и улучшению качества работы моделей.

Библиографический список

1. Воронина В. В., Михеев А. В., Ярушкина Н. Г., Святов К. В. // Теория и практика машинного обучения : учебное пособие /В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. –Ульяновск : УлГТУ, 2017. – 291 с.
2. Паклин, Н. Б. Глава 9, // Бизнес-аналитика: от данных к знаниям : учебное пособие / Н. Б. Паклин, В. И. Орешков. – 2-е изд.. – СПб. : Питер, 2013. – 706 с.
3. П. Брюс, Э. Брюс. 1. Разведочный анализ данных // Практическая статистика для специалистов Data Science. — СПб.: БХВ-Петербург, 2018. — С. 19—58. — 304 с.
4. Бринк Х. Машинное обучение / Х. Бринк, Дж. Ричардс, М. Феверолф. — пер. с англ. Рузмайкина И. — Санкт-Петербург: Питер, 2017. — 336 с.
5. Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей / А.Н. Горбань // Сиб. журн. вычисл. математики. – 1998. – Т. 1, № 1. – 21 с.
6. Паклин, Н. Б. Глава 9, // Бизнес-аналитика: от данных к знаниям : учебное пособие / Н. Б. Паклин, В. И. Орешков. – 2-е изд.. – СПб. : Питер, 2013. – 706 с.
7. Прикладной системный анализ : учебное пособие / Ф.П. Тарасенко. — М. : КНОРУС, 2010. — 224 с.
8. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / под ред. О.Б.Лупанова. 2004. Вып. 13 С. 5 – 36.
9. Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. СПб. : Питер, 2018. – 576 с.
10. Jackson L Разница между классификацией и регрессией: [Электронный ресурс] – Режим доступа: https://ru.strephonsays.com/classification-and-vs-regression-13803 (дата обращения: 12.03.2023).
11. Евклидова метрика [Электронный ресурс]: Википедия, Свободная энциклопедия – Режим доступа https://ru.wikipedia.org/wiki/ Евклидова_метрика (дата обращения 12.03.2023).

12. Машинное обучение. Конспект лекций. Разинков Е.В. Казань, 2015 – 28 с. Язык программирования Python [Электронный ресурс]- Режим доступа: https://web-creator.ru/articles/python (дата обращения: 13.03.2023)

13. Deep Learning with Keras via Artificial Neural Network [Электронный ресурс]- Режим доступа: https://rpubs.com/A_Rodionoff/Regression-Keras(Дата обращения: 20.03.2023).

14. Композитный_материал [Электронный ресурс]: Википедия, Свободная энциклопедия – Режим доступа https://ru.wikipedia.org/wiki/Композитный_материал (дата обращения 12.03.2023).

15. Машинное обучение на практике с Python и Keras [Электронный ресурс]. – Режим доступа: https://pythonru.com/primery/mashinnoe-obucheniena-praktike-s-python-i-keras (дата обращения 24.03.2023).

16. Продукция из композитных материалов. – Текст : электронный //– [Электронный ресурс]. – Режим доступа: https://bazaltural.ru/wpcontent/uploads/2018/03/Produktsiya.pdf (Дата обращения 08.03.2023).

17.