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»

Слушатель Андреев Савелий Михайлович

Москва, 2023

Содержание

Введение

После того как современная физика металлов подробно разъяснила нам причины их пластичности, прочности и ее увеличения, началась интенсивная систематическая разработка новых материалов. Это приведет, вероятно, уже в вообразимом будущем к созданию материалов с прочностью, во много раз превышающей ее значения у обычных сегодня сплавов. При этом большое внимание будет уделяться уже известным механизмам закалки стали и старения алюминиевых сплавов, комбинациям этих известных механизмов с процессами формирования и многочисленными возможностями создания комбинированных материалов. Два перспективных пути открывают комбинированные материалы, усиленные либо волокнами, либо диспергированными твердыми частицами. У первых в неорганическую металлическую или органическую полимерную матрицу введены тончайшие высокопрочные волокна из стекла, углерода, бора, бериллия, стали или нитевидные монокристаллы. В результате такого комбинирования максимальная прочность сочетается с высоким модулем упругости и небольшой плотностью. Именно такими материалами будущего являются композиционные материалы.

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. 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.

Тема выпускной квалификационной работы по курсу «Data Science»: «Прогнозирование конечных свойств новых материалов (композиционных материалов)».

Задачей данной работы является прогнозирование конечных свойств новых композиционных материалов. Для решения поставленной задачи были использованы данные о начальных свойствах компонентов композиционных материалов. Целью проведения исследования, анализа предоставленных данных было получение прогноза ряда конечных свойств получаемых композиционных материалов.

Актуальность темы заключается в том, что созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

В ходе решения данной задачи применялись методы, изученные на курсе «Data Science».

1. Аналитическая часть

1.1 Постановка задачи

Применение композитов в различных областях обосновано их высокими физико-механическими свойствами по сравнению с общепринятыми материалами. На практике заметно, что при использовании композитов уменьшается вес объекта, можно сэкономить средства на производство или строительство, при этом улучшить некоторые механические характеристики конструкции. Также благодаря своим свойствам они могут быть использованы в агрессивных средах, а их применение в целом способствует увеличению срока службы здания. Однако, недостаток композитной арматуры в том, что, обладая высокими прочностными характеристиками, она имеет низкий модуль упругости.

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).

На входе имеются данные о начальных свойствах компонентов композиционных материалов, состоящие из 13 факторов (переменных):

1) соотношение матрица-наполнитель;
2) плотность, кг/м3;
3) модуль упругости, Гпа;
4) количество отвердителя, м.%;
5) содержание эпоксидных групп,%_2;
6) температура вспышки, С_2;
7) поверхностная плотность, г/м2;
8) модуль упругости при растяжении, ГПа;
9) прочность при растяжении, Мпа;
10) потребление смолы, г/м2;
11) угол нашивки, град;
12) шаг нашивки;
13) плотность нашивки.

На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Датасет со свойствами композитов, полученный после объединения данных по типу INNER, состоит из 1 023 строк (наблюдений).

Требуется:

1) Изучить теоретические основы и методы решения поставленной задачи.

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

3) Провести предобработку данных (удаление шумов, нормализация и т.д.).

4) Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении модели необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 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, на выбор учащегося).

7) Оценить точность модели на тренировочном и тестовом датасете.

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

1.2 Описание используемых методов

При решении задачи применялись методы машинного обучения. Машинное обучение — это класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

Некоторые алгоритмы обучения делают определенные предположения о структуре данных или желаемых результатов. Если найти тот алгоритм, который соответствует потребностям, с ним можно получить более точные результаты, более точные прогнозы и сократить время обучения.

В данной исследовании поставлена задача регрессии, так как требуется получить прогноз на основе выборки объектов с различными признаками. На выходе должно получиться число.

Рассмотрим каждый из используемых в работе методов машинного обучения:

1. ***Линейная регрессия*** — это общий статистический метод, который был реализован в машинном обучении и дополнен многими новыми методами для подгонки строки и измерения ошибок. Регрессия связана с прогнозированием числовых целевых значений. Линейная регрессия — хороший выбор, когда требуется простая модель для базовой задачи прогнозирования. Линейную регрессию также обычно используют для работы с многомерными разреженными наборами данных с отсутствием сложности.

2. ***Лес принятия решений (случайный лес).*** Деревья принятия решений — это непараметрические модели, выполняющие последовательность простых тестов для каждого экземпляра, выполняя обход древовидной структуры двоичных данных до достижения конечного узла (решения).

Деревья принятия решений имеют следующие преимущества:

- они эффективны с точки зрения вычисления и использования памяти во время обучения и прогнозирования;

- они могут представлять границы нелинейного принятия решений;

- они выполняют выбор признаков и классификацию и являются устойчивыми при наличии шумовых признаков.

Эта модель регрессии состоит из совокупности деревьев принятия решений. Каждое дерево в регрессионном лесу решений выводит распределение по Гауссу в виде прогноза. По совокупностям деревьев выполняется агрегирование с целью найти распределение по Гауссу, ближайшее к объединенному распределению для всех деревьев модели.

3. ***Метод опорных векторов с линейным ядром (Support Vector Machine — SVM)*** Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как метод классификатора с максимальным зазором.

Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, создающая наибольшее расстояние до двух параллельных гиперплоскостей. Алгоритм основан на допущении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

4. ***Метод k-ближайших соседей*** В случае использования метода для регрессии, объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны. Алгоритм может быть применим к выборкам с большим количеством атрибутов (многомерным). Преимуществом метода является его хорошая математическая обоснованность, недостатком — низкая объясняющая способность.

5. Регрессия нейронной сети. Несмотря на то, что нейронные сети широко используются для углубленного обучения и моделирования сложных задач, таких как распознавание изображений, они легко адаптируются к задачам регрессии. Любой класс статистических моделей можно назвать нейронной сетью, если эти модели используют адаптивные весовые коэффициенты и могут использоваться для аппроксимации нелинейных функций входных данных. Таким образом, регрессия нейронной сети подходит для задач, которые нельзя решить с помощью более традиционных моделей.

Нейронная сеть выдаст прогнозируемое значение переменной, зависимое от множества входных параметров.

Перед тем, как производить прогноз, алгоритм обучается на тренировочном наборе данных — обучающей выборке. Каждая строка такой выборки содержит:

- в полях, обозначенных как входные — множество входных параметров;

- в поле, обозначенном как выходное — соответствующее входным параметрам значение зависимой переменной.

Технически обучение заключается в нахождении весов — коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными параметрами и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения нейронная сеть способна выдать верный результат на основании данных, которые отсутствовали в обучающей выборке, а также на неполных и/или «зашумленных», частично искажённых данных.

1.3 Разведочный анализ данных

С помощью разведочного анализа данных были изучены основные свойства данных. С помощью метода «df.isna()» получена информация об отсутствии пропусков в данных (рис. 1):

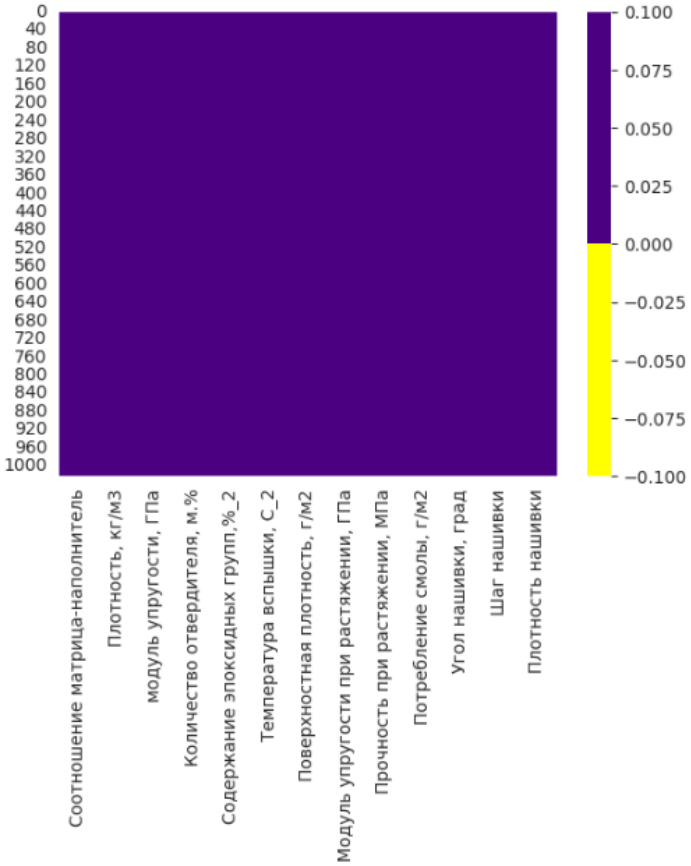


Рисунок 1 – Тепловая карта пропущенных значений

Тепловая карта пропущенных значений показала, что пропусков в полученной выборке нет.

С помощью метода Python «df.describe()» была получена описательная статистика (рис. 2).

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.930366	0.913222	0.389403	2.317887	2.906878	3.552660	5.591742
Плотность, кг/м3	1023.0	1975.734888	73.729231	1731.764635	1924.155467	1977.621657	2021.374375	2207.773481
модуль упругости, ГПа	1023.0	739.923233	330.231581	2.436909	500.047452	739.664328	961.812526	1911.536477
Количество отвердителя, м.%	1023.0	110.570769	28.295911	17.740275	92.443497	110.564840	129.730366	198.953207
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301	14.254985	20.608034	22.230744	23.961934	33.000000
Температура вспышки, С_2	1023.0	285.882151	40.943260	100.000000	259.066528	285.896812	313.002106	413.273418
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690	0.603740	266.816645	451.864365	693.225017	1399.542362
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983	64.054061	71.245018	73.268805	75.356612	82.682051
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006	1036.856605	2135.850448	2459.524526	2767.193119	3848.436732
Потребление смолы, г/м2	1023.0	218.423144	59.735931	33.803026	179.627520	219.198882	257.481724	414.590628
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1023.0	6.899222	2.563467	0.000000	5.080033	6.916144	8.586293	14.440522
Плотность нашивки	1023.0	57.153929	12.350969	0.000000	49.799212	57.341920	64.944961	103.988901

Рисунок 2 – Описательная статистика

Из данных таблицы мы можем посмотреть следующие значения для каждого признака: count - количество значений, mean - среднее значение, std - стандартное отклонение, min – минимум, 25% - верхнее значение первого квартиля, 50% - медиана, 75% - верхнее значение третьего квартиля, max – максимум.

Для оценки величины и характера разброса данных построены гистограммы распределения (рис. 3):

Рисунок 3 – Гистограммы распределения

При этом сделаны выводы о выборке:

1) Угол нашивки является дискретной величиной;

2) Значения Поверхностной плотности имеют Пуассоновское распределение;

3) Все остальные распределения близки к нормальному.

Для визуализации коэффициентов корреляции и определения того, между какими переменными установлена более тесная взаимосвязь, была построена тепловая карта коэффициентов корреляции (рис. 4):

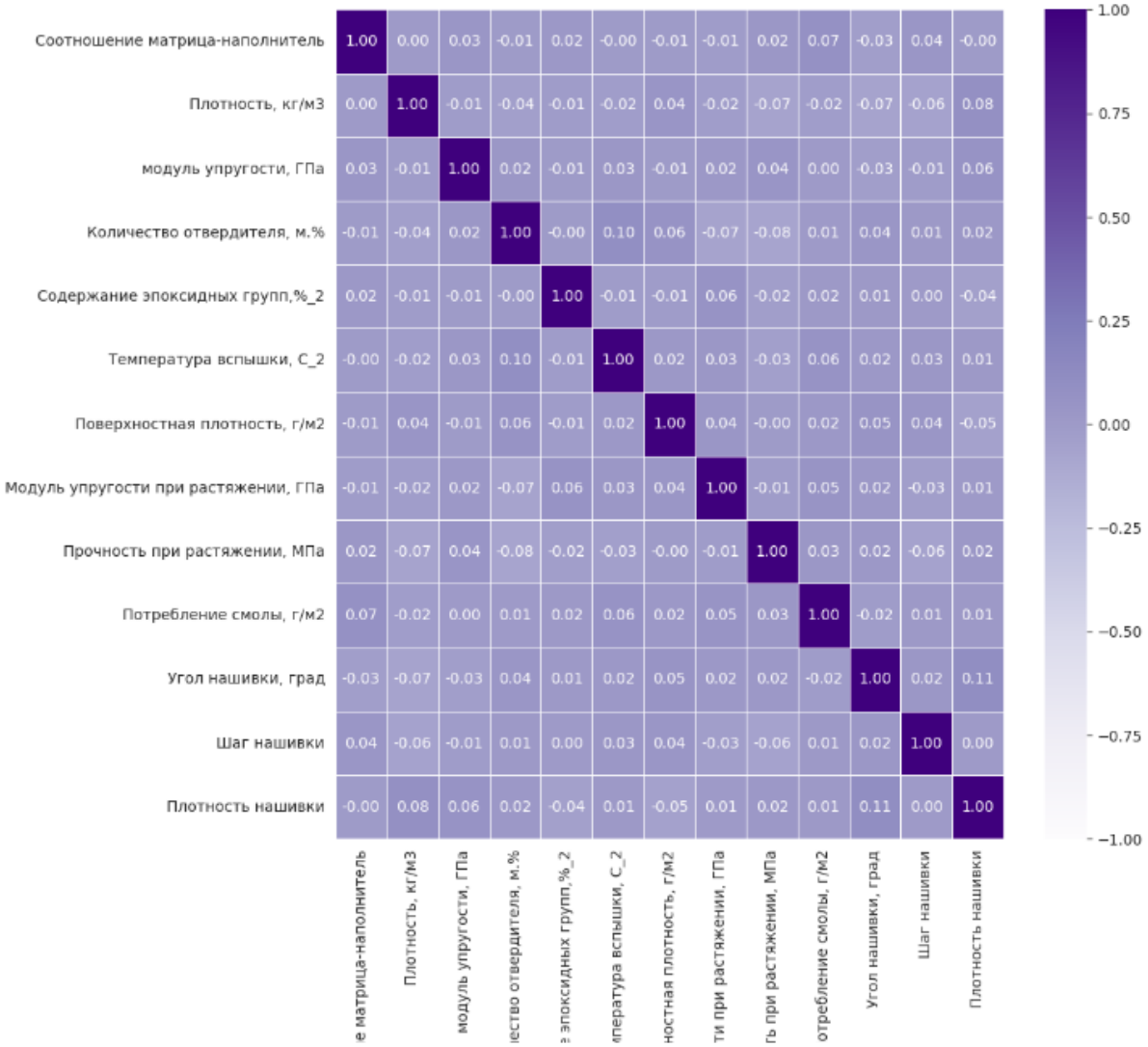


Рисунок 4 - Тепловая карта коэффициентов корреляции.

Самая большая корреляция из рассматриваемых (по убыванию):

1) Угол нашивки/Плотность нашивки 0,11

2) Температура вспышки/Количество отвердителя 0,10;

3) Плотность/Плотность нашивки 0,08;

4) Прочность при растяжении/Количество отвердителя - обратная корреляция -0,08;

5) Потребление смолы/Соотношение матрица-наполнитель 0,07;

6) Модуль упругости при растяжении/Количество отвердителя - обратная корреляция -0,07.

Однако, стоит отметить, что все параметры коррелируют между собой очень слабо. Это также доказывают построенные диаграммы рассеивания. Пример диаграмм рассеивания по показателям, которые больше всего коррелируют между собой (рис. 5):

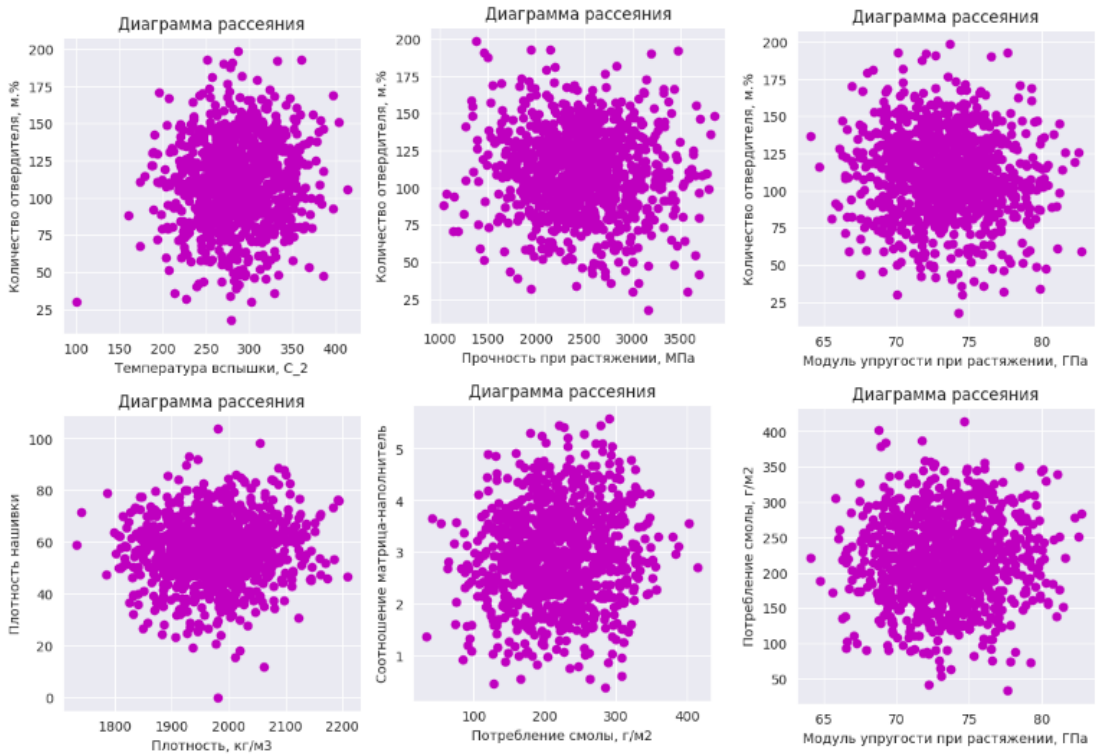


Рисунок 5 – Диаграммы рассеивания на примере параметров с наибольшей корреляцией

Для наглядности были построены гистрограммы распределения и ящики с усами вместе, с помощью функций «boxplot()» и «histplot()». Пример изображен на рисунке 6.

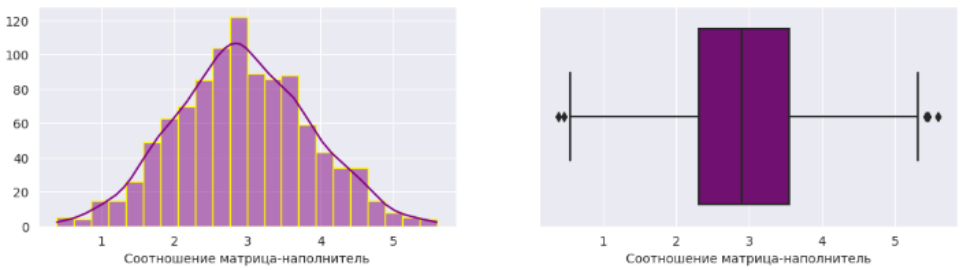


Рисунок 6 – Диаграммы размаха («ящики с усами») с гистограммами распределения

Чтобы понять взаимосвязь между всеми возможными парами числовых переменных с помощью использования библиотеки seaborn был построен попарный график, в верхнем-правом углу которого расположены графики плотности ядра, в нижнем-левом – графики рассеивания, между ними – графики распределения. (рис.7):

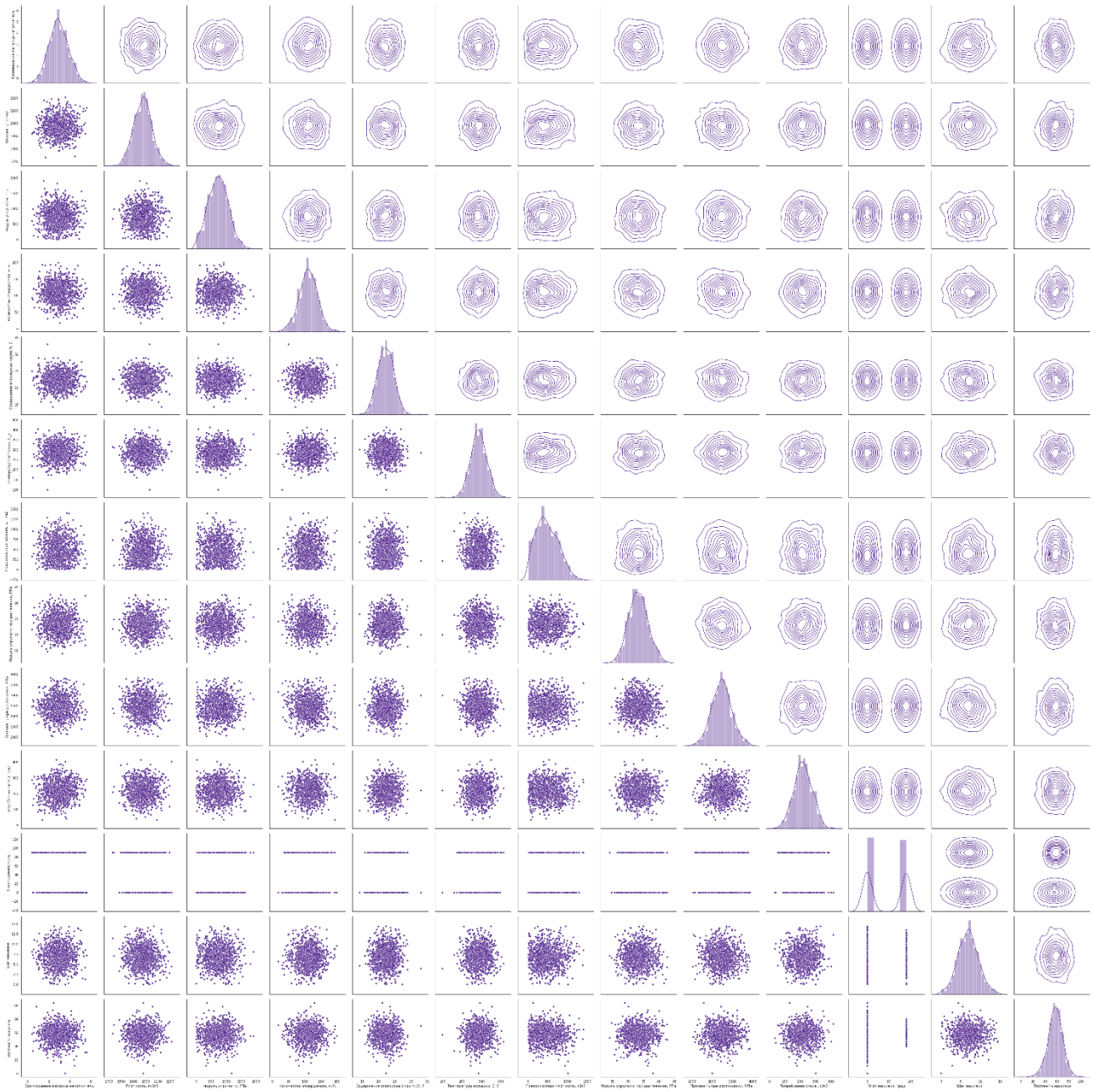


Рисунок 7 – Попарный график диаграмм рассеивания точек

2.Практическая часть

2.1 Предобработка данных

Предоставленные для решения задачи данные были предобработаны.

В целях очищения датасета от выбросов (аномалий) предпринимались попытки выявления выбросов данными методами:

1) Изолирующий лес. Этот метод «изолирует» наблюдения, случайным образом выбирая признак, а затем случайным образом выбирая значение разделения между максимальным и минимальным значениями выбранного признака. Этот алгоритм отлично работает с наборами данных очень большого размера. Однако он выявил более 200 выбросов, что очень много для общего количества данных (19,5%). Поэтому принято решение не применять этот метод исключения выбросов.

2) Метод трёх сигм. Анализ выбросов в данных методом позволяет определить аномальные значения в нестационарных рядах с распределением близким к нормальному. Основу данного метода анализа составляет расчет среднего значения ряда и среднеквадратичного отклонения.

3) Метод IQR (межквартильный размах). IQR - это понятие в статистике, которое используется для измерения статистической дисперсии и изменчивости данных путем разделения набора данных на квартили. Межквартильный размах (IQR) важен, потому что он используется для определения выбросов. Это разница между третьим квартилем и первым квартилем (IQR = Q3 -Q1). Выбросы в данном случае определяются как наблюдения ниже (Q1 - 1,5x IQR) или выше (Q3 + 1,5x IQR). По методу удаления IQR получается, что в данных 52 строки с выбросами.

Принято решение остановиться на методе трёх сигм, т.к. он удаляет меньше выбросов, а именно - 24 строки (2,3% от общего количества наблюдений). Это решение связано с тем, чтобы сохранить большее количество данных для обработки, так как их распределение не говорит о присутствии явных аномалий. После очистки данных от выбросов количество наблюдений составило 999 строк. Таким образом, можно сделать вывод, что исключение выбросов не оказало существенного влияния на размер выборки.

Будучи разными по физическому смыслу, данные сильно различаются между собой по абсолютным величинам. Работа аналитических моделей машинного обучения с такими показателями окажется некорректной: дисбаланс между значениями признаков может вызвать неустойчивость работы модели, ухудшить результаты обучения и замедлить процесс моделирования.

После нормализации все числовые значения входных признаков будут приведены к одинаковой области их изменения – некоторому узкому диапазону. Это позволит свести их вместе в одной модели и обеспечит корректную работу вычислительных алгоритмов. Для дальнейшей работы с данными, сравнения их между собой и составления модели машинного обучения была проведена нормализация данных методом «MinMaxScaler()». Этот метод нормализации включает масштабирование набора данных до диапазона [0, 1].

Масштабированием называется общий процесс изменения диапазона признака. Это необходимый шаг, потому что признаки измеряются в разных единицах, а значит покрывают разные диапазоны. Это сильно искажает результаты таких алгоритмов, как метод опорных векторов и метод k-ближайших соседей, которые учитывают расстояния между измерениями. А масштабирование позволяет этого избежать. И хотя методы вроде линейной регрессии и «случайного леса» не требует масштабирования признаков, лучше не пренебрегать этим этапом при сравнении нескольких алгоритмов.

Описание данных после проведения нормализации приведено на рисунке 8:

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	999.0	0.489727	0.174701	0.0	0.371306	0.484288	0.608487	1.0
Плотность, кг/м3	999.0	0.467798	0.178722	0.0	0.341020	0.472391	0.579760	1.0
модуль упругости, ГПа	999.0	0.446887	0.198929	0.0	0.302135	0.448458	0.581067	1.0
Количество отвердителя, м.%	999.0	0.496747	0.170875	0.0	0.384427	0.495616	0.613450	1.0
Содержание эпоксидных групп,%_2	999.0	0.493097	0.179869	0.0	0.368588	0.492051	0.624540	1.0
Температура вспышки, С_2	999.0	0.488685	0.174877	0.0	0.371822	0.488391	0.606296	1.0
Поверхностная плотность, г/м2	999.0	0.371058	0.215125	0.0	0.206249	0.348503	0.534748	1.0
Модуль упругости при растяжении, ГПа	999.0	0.501023	0.167891	0.0	0.389296	0.496176	0.610020	1.0
Прочность при растяжении, МПа	999.0	0.508273	0.172193	0.0	0.390683	0.504890	0.613078	1.0
Потребление смолы, г/м2	999.0	0.512182	0.170414	0.0	0.401086	0.512933	0.625356	1.0
Угол нашивки, град	999.0	0.496496	0.500238	0.0	0.000000	0.000000	1.000000	1.0
Шаг нашивки	999.0	0.477203	0.177675	0.0	0.351355	0.478419	0.593879	1.0
Плотность нашивки	999.0	0.507132	0.163683	0.0	0.405778	0.510118	0.612960	1.0

Рисунок 8 – Описательная статистика после нормализации данных

Ниже приведены гистограммы распределения данных после нормализации (рис.9):

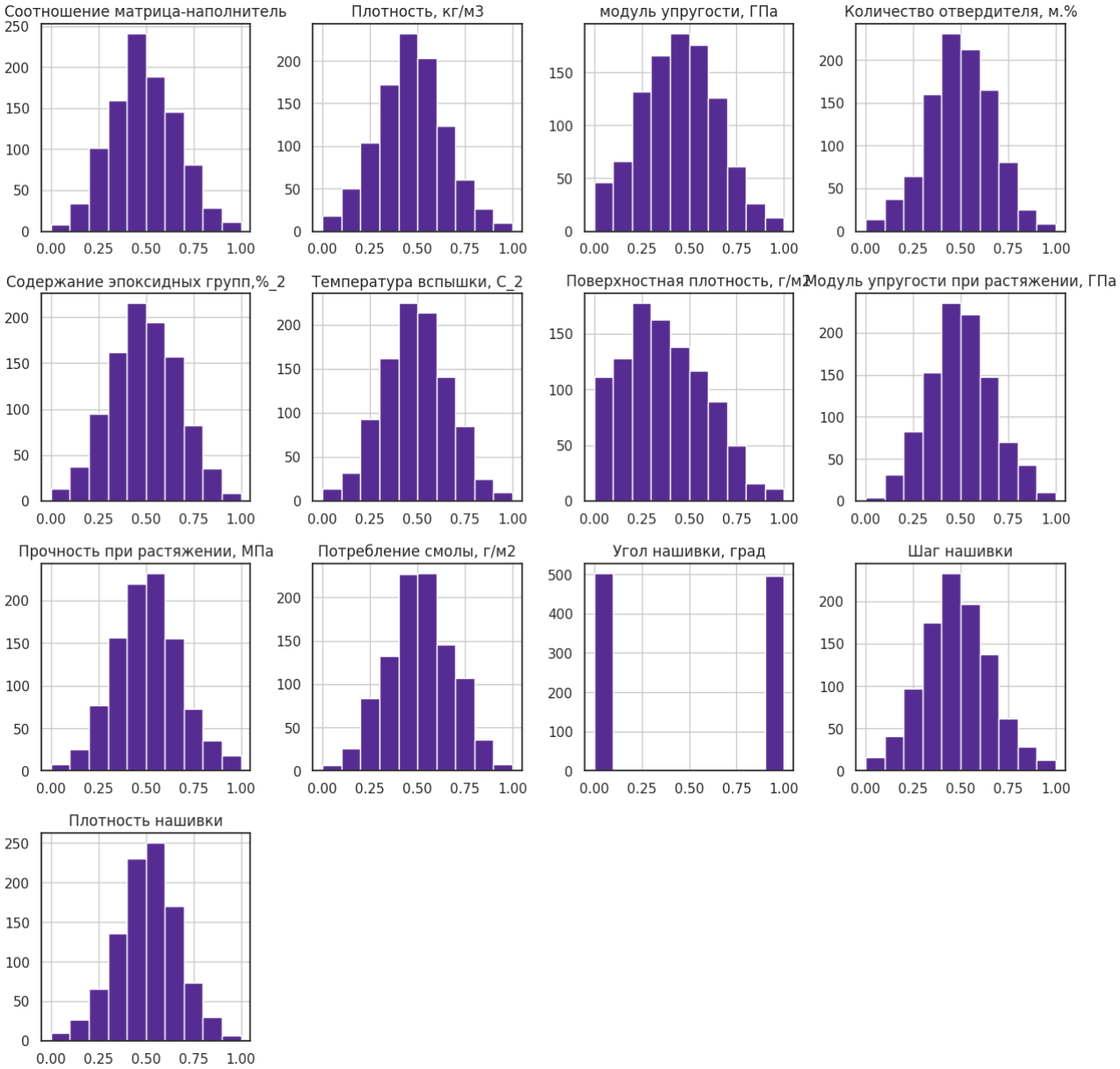


Рисунок 9 – Гистограммы распределения после нормализации

Таким образом, после предобработки данных получаем нормализованный датасет размером 999 строк, данные в котором приближены к нормальному распределению, за исключением угла нашивки (дискретная величина) и поверхностной плотности (Пуассоновское распределение).

2.2 Разработка и обучение модели

Для прогноза модуля упругости при растяжении и прочности при растяжении были использованы следующие методы решения задачи множественной регрессии с помощью Python:

1) ***Линейная регрессия*** – метод «LinearRegression»;
2) ***Метод k-ближайших соседей*** – метод «KneighborsRegressor»;
3) ***Метод опорных векторов с линейным ядром*** – метод «SVR»;
4) ***Лес принятия решений (случайный лес***) – метод «RandomForestRegressor».

Для разработки и обучения моделей прогнозные значения отделены от выборки:

```
trg = df_MinMax[['Модуль упругости при растяжении, ГПа', 'Прочность при растяжении, МПа']]

trn = df_MinMax.drop(['Модуль упругости при растяжении, ГПа', 'Прочность при растяжении, МПа'], axis=1)
```

Далее все модели были помещены в один список для удобства дальнейшего анализа:

```
models = [LinearRegression(),
          KNeighborsRegressor(n_neighbors=10),
          SVR(kernel='linear'),
          RandomForestRegressor(n_estimators=100, max_features ='sqrt')]
```

На тестирование модели оставили 30 процентов данных, а на остальных происходило обучение моделей. Далее для каждого из параметров (модуль упругости при растяжении и прочность при растяжении) были построены модели регрессии.

2.3 Тестирование модели

Оценка качества моделей проведена с помощью расчёта коэффициентов детерминации, которые показывает долю вариации результативного признака под влиянием факторного признака. При отсутствии связи эмпирический коэффициент детерминации равен нулю, а при функциональной связи — единице. Полученные результаты моделей показаны на рисунке 10:

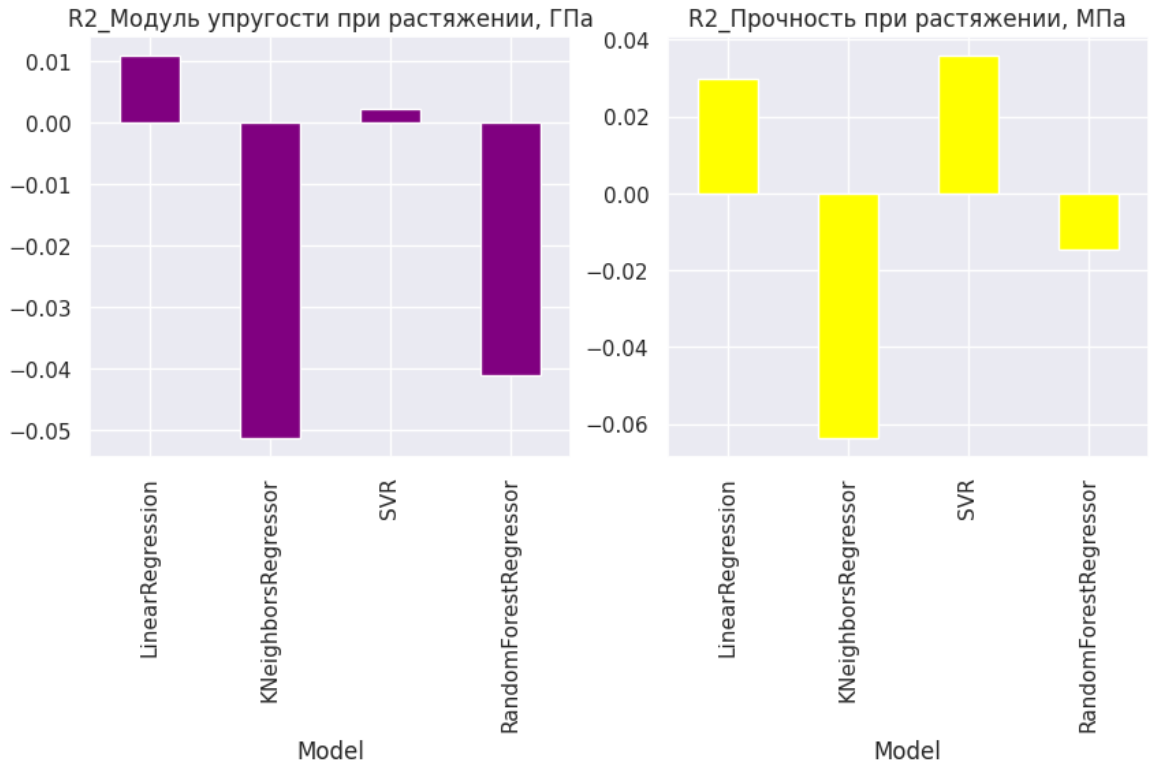


Рисунок 10 – Результаты моделей для решения задач множественной регрессии

Из графиков видно, что ни одна из моделей не справилась с задачей хорошо. При этом хуже всего показывает себя метод k-ближайших соседей. Метод линейной регрессии и метод опорных векторов дают прогнозы, приближённые к простому усреднению.

Так как лучше всего себя показала модель на основе линейной регрессии, была проведена её оценка путём вычисления среднеквадратической ошибки прогноза (MSE) и средней абсолютной ошибки (MAE). Для этого использовались методы «mean_squared_error» и «mean_absolute_error» соответственно. Результаты оценки получились следующие: «mse: 1.47, mae: 1.18».

2.4 Написание нейронной сети, рекомендующей соотношение матрица-наполнитель

Для прогнозирования соотношения матрица-наполнитель написаны модели с использованием однослойного и многослойного персептрона. В этих целях была задействована библиотека keras. Нормализация данных для нейронной сети проводилась с помощью отдельного слоя «normalizer».

При построении линейной модели зависимость соотношения матрица-наполнитель была установлена от потребления смолы, так как по тепловой карте коэффициентов корреляции видно, что с этим признаком наблюдается наибольшая взаимосвязь (коэффициент корреляции: 0,07).

Функция активации определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения. В качестве такой функции был выбран гиперболический тангенс («tanh»). Её природа нелинейна, она хорошо подходит для комбинации слоёв, а диапазон значений функции -(-1, 1).

При построении многослойного персептрона после слоя нормализации добавлен слой с 32 нейронами и активационной функцией «tanh», слой с пакетной нормализацией (BatchNormalization), слой с методом «прореживания» (Dropout), а также выходной слой с 1 нейроном.

Пакетная нормализация должна привести данные к положительному распределению со средним значением 0 и дисперсией 1, чтобы распределение данных было согласованным и исчезал градиент. Также это помогает для ускорения процесса обучения нейронной сети.

Метод «прореживания» используется для предотвращения переобучения модели. Проблема переобучения в том, что модель хорошо объясняет только примеры из обучающей выборки, адаптируясь к обучающим примерам, вместо того чтобы учиться классифицировать примеры, не участвовавшие в обучении (теряя способность к обобщению). Dropout «выключает» нейроны с

вероятностью *p* и, как следствие, оставляет их включенными с вероятностью *q=1-p*.

Все модели обучались на 100 эпохах, что является приемлемым для данного датасета. Размер тестовой выборки составил 30 процентов.

При построении линейной модели с использованием многослойного персептрона (dnn_potrsmoly_model) видно, как определяются предсказанные значения соотношения матрица-наполнитель в отличие от истинных данных (рис. 11):

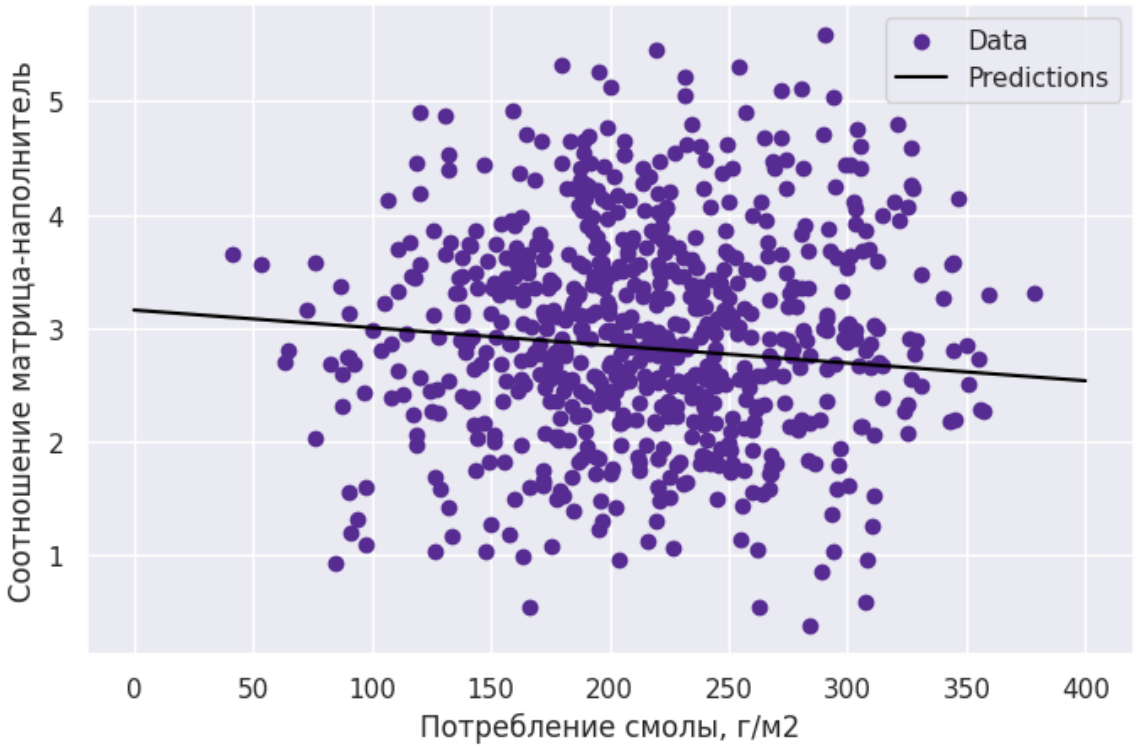
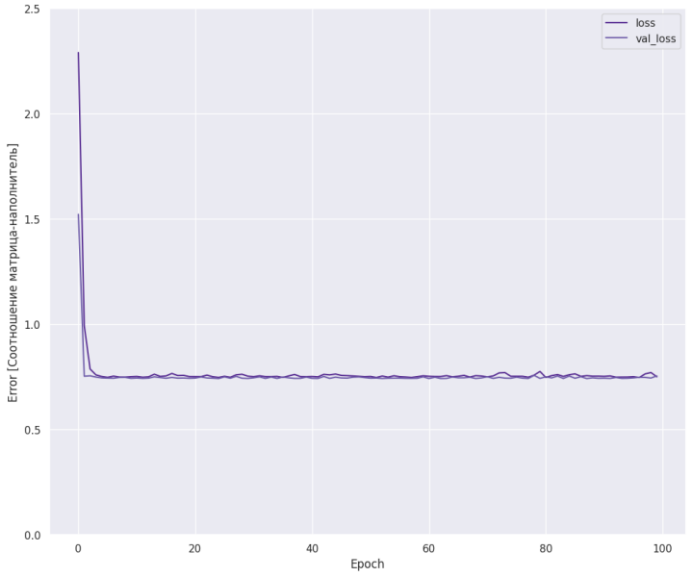
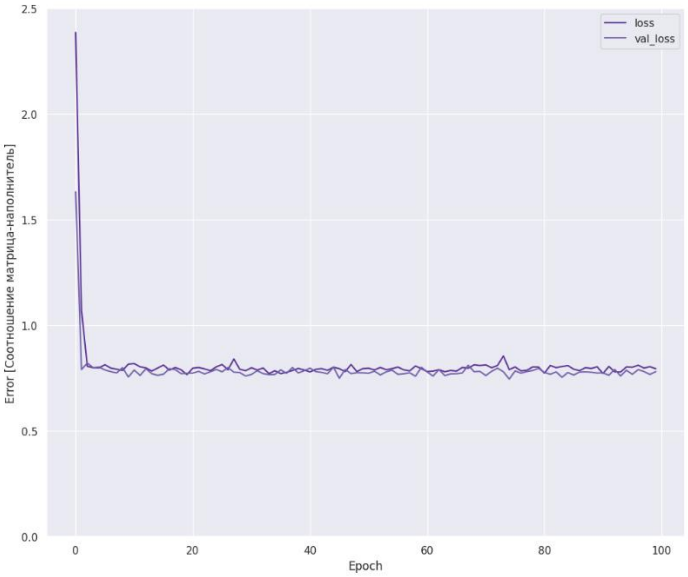


Рисунок 11 – График предсказанных значений соотношения матрица-наполнитель в модели (dnn_potrsmoly_model)

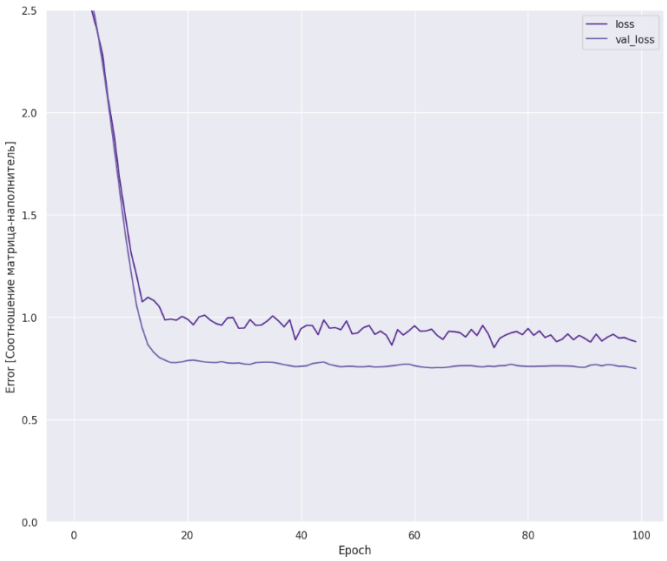
Графики, на которых отражён процесс обучения моделей, представлены на рисунке 12:



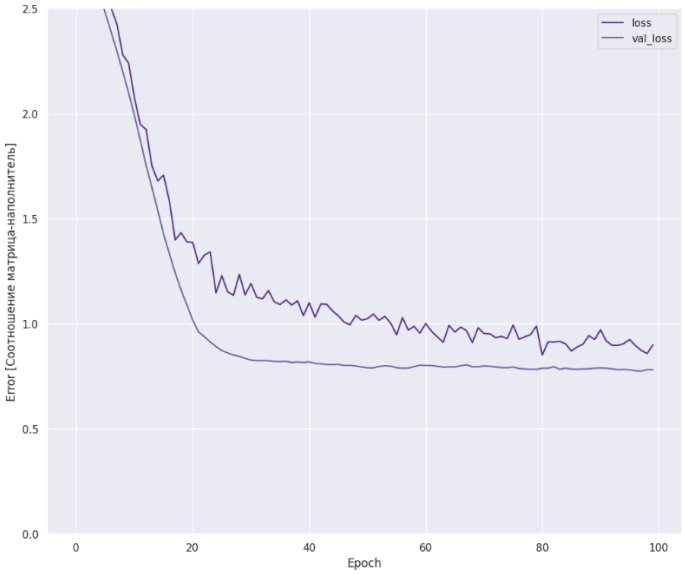
potrsmoly_model



linear_model



dnn_potrsmoly_model



dnn_model

Рисунок 12 – Графики изменения ошибки тестовой и валидационной выборки в процессе обучения моделей

Оценка качества моделей проведена с помощью метрики MAE (рис. 13).

	Mean absolute error [Соотношение матрица-наполнитель]
potrsmoly_model	0.700175
linear_model	0.691898
dnn_potrsmoly_model	0.678404
dnn_model	0.710185

Рисунок 13 – Результаты оценки качества моделей для прогноза соотношения матрица-наполнитель с помощью метрики MAE

По результатам оценки качества моделей видно, что наименьшая ошибка у многослойного персептрона по потреблению смолы. Однако, при большом количестве признаков данных, корреляция которых не является сильной, предлагается использовать модель, которая учитывает все имеющиеся признаки (dnn_model).

Для этой нейронной сети (dnn_model) был построен график, отражающий истинные и предсказанные моделью значения соотношения матрица-наполнитель (рис. 14). На нем прекрасно видно, что точность предсказаний находится на невысоком уровне:

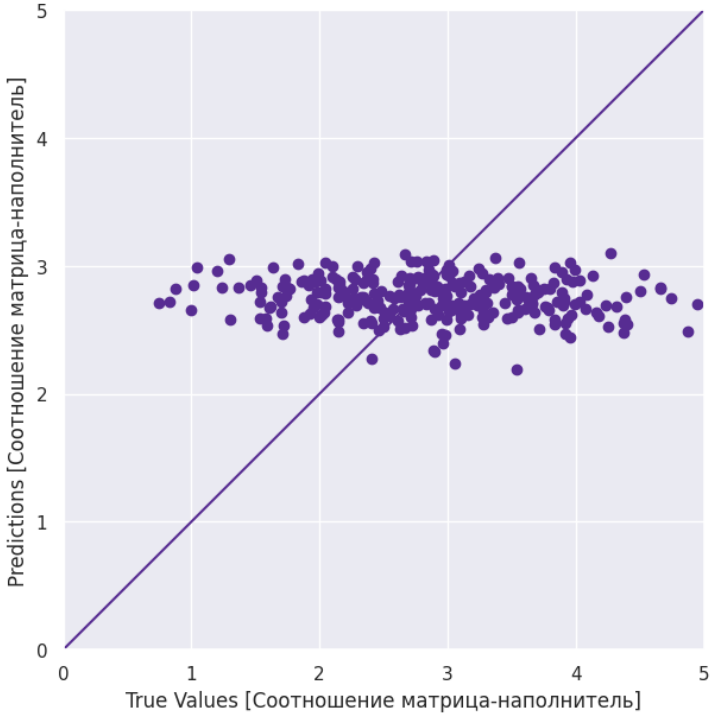


Рисунок 14 – График истинных и предсказанных значений соотношения матрица-наполнитель

Построена гистограмма распределения ошибки модели (dnn_model) (рис.15). Видно, что распределение приближено к нормальному.

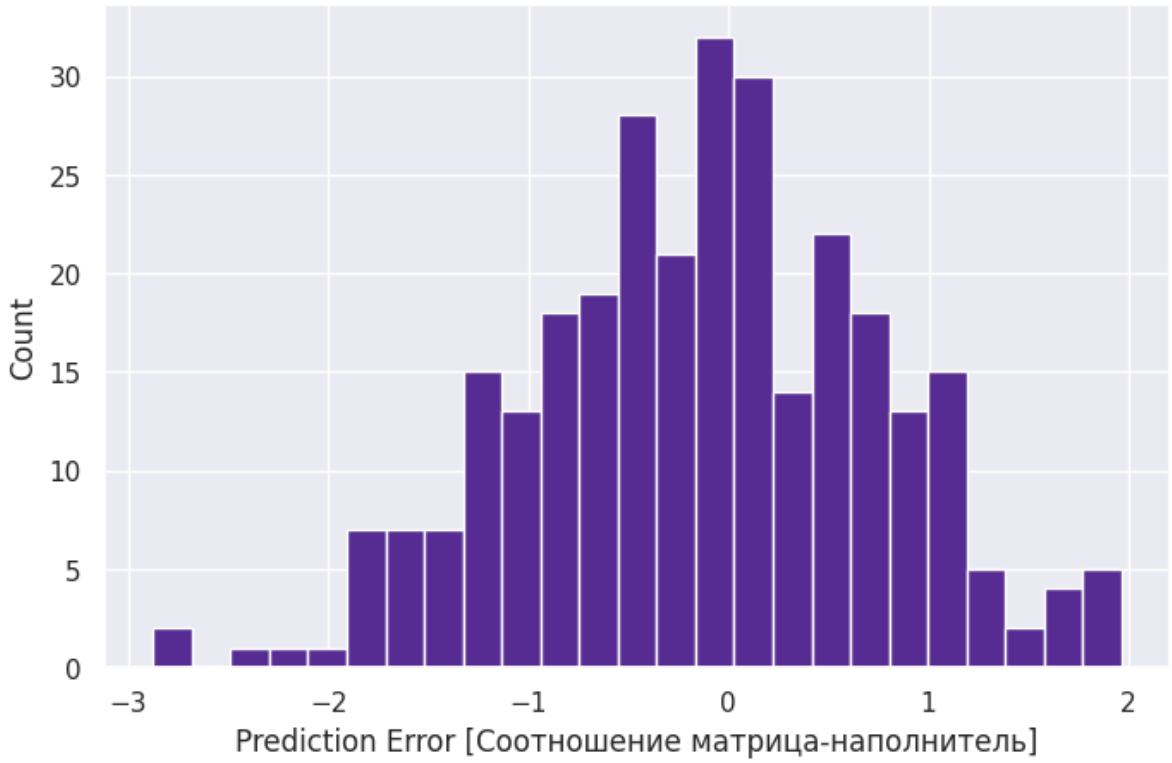


Рисунок 15 - Гистограмма распределения ошибки

2.5 Создание удаленного репозитория и загрузка результатов работы

Создан репозиторий в GitHub, где размещён код исследования. Оформлен файл README.

Страница слушателя: https://github.com/squerna

Созданный репозиторий: https://github.com/squerna/squern_model.

Заключение

В ходе решения задачи прогнозирования конечных свойств новых композиционных материалов были изучены основные теоретические основы методов машинного обучения. Проведён анализ предоставленных данных, а также получены прогнозы ряда конечных свойств получаемых композиционных материалов.

При этом в практической части задачи были применены методы машинного обучения на базе библиотек Python. Проведён анализ результатов, полученных с помощью созданных моделей, для выбора наиболее точной из них.

Полученные в ходе решения задачи модели не приносят положительного результата, так как имеют высокий уровень ошибки в предсказаниях. Однако, при проведения данного исследования получен опыт выбора наиболее подходящих методов для решения задач регрессии, опыт настройки моделей, выбора гиперпараметров, а также оценки качества моделей по разным метрикам.

Таким образом, можно сделать вывод о том, что данная работа позволила выработать навыки решения задач с помощью методов машинного обучения, определить направление для дальнейшего постижения науки о данных (Data science).

Библиографический список

1 ГОСТ 32794-2014 Композиты полимерные. - Введ. 2015-09-01. - М.: Стандартинформ, 2015

2 ГОСТ Р 57970-2017 Композиты углеродные. Углеродные композиты, армированные углеродным волокном. – Введ. 2018-06-01. - М.: Стандартинформ, 2018

3 СП 28.13330.2012. «Защита строительных конструкций от коррозии». Актуализированная редакция СНиП 2.03.11-85. Дата введения 2013.01.01.

4 Библиотека Keras - инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow / пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2018. - 294 с.

5 Аллен Б. Дауни – Основы Python. Научитесь думать как программист / Аллен Б. Дауни ; пер. с англ. С. Черникова ; [науч. ред. А. Родионов]. — Москва: Манн, Иванов и Фербер, 2021. — 304 с.

6 Avdeeva A., Shlykova I., Perez M., Antonova M., Belyaeva S.Chemical properties of reinforcing fiberglass in aggressive media. MATEC Web of Conferences. 2016. Vol. 53. 01004.

7 Астапов Р.Л., Мухамадеева Р.М. Автоматизация подбора параметров машинного обучения и обучение модели машинного обучения // Актуальные научные исследования в современном мире. 2021. № 5-2 (73). С. 34-37.

8 Barabanshchikov Y., Belyaeva S., Avdeeva A.and Perez M. Fiberglass Reinforcement for Concrete (2015) Applied Mechanchanics and Materials, Pp. 475-481

9 Вичугова A.Data Preparation: полет нормальный – что такое нормализация данных и зачем она нужна [Электронный ресурс]: – Режим доступа: https://www.bigdataschool.ru/blog/нормализация-feature-transformation-data-preparation.html (дата обращения: 13.06.2022).

10 Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 480 с.: ил. — (Серия «Бестселлеры O’Reilly»).

11 Бринк Х. Машинное обучение. / Х. Бринк, Дж. Ричардс, М. Феверолф. – СПб.: Питер,2017. 336 с.

12 Вандер Плас Дж. Python для сложных задач: наука о данных и машинное обучение. - СПб.: Питер, 2018. - 576 с.

13 Гиздатуллин А.Р., Хозин В.Г., Куклин А.Н., Хуснутдинов А.М. «Особенности испытаний и характер разрушения полимеркомпозитной арматуры».

14 Горбунов П.М., Мацкевич Ю.А., Чубарь А.В. Машинное обучение. Автоматизация подбора модели машинного обучения // Робототехника и искусственный интеллект. 2021. С. 155-160.

15 Джалилов Ш.А. Метод расчета параметров множественной линейной регрессии // Достижения науки и образования. 2020. № 3 (57). С. 24-28.

16 Джулли, Пал: Библиотека Keras - инструмент глубокого обучения / пер. с англ. А. А. Слинкин.- ДМК Пресс, 2017. – 249 с.

17 Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем.Пер. с англ. - СпБ.: ООО Альфа-книга: 2018. - 688 с.

18 Flach P. Machine learning. The art and science of building algorithms. pp. 118-142.

19 Кузнецов И.Н. Пример решения задачи множественной регрессии с помощью Python [Электронный ресурс]: – Режим доступа: https://habr.com/ru/post/206306/ (дата обращения: 13.06.2022).

20 Makusheva N.Yu., Kolosova N.B. Comparative analysis of metal reinforcement and fibre-reinforced plastic rebar Construction of Unique Buildings and Structures, 2014, No10 (25) Pp. 60-72

21 Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. / Дж. Вандер Плас. – СПб.: Питер, 2018. 576 с.

22 Сохина С.А., Немченко С.А. Машинное обучение. Методы машинного обучения // Современная наука в условиях модернизационных процессов: проблемы, реалии, перспективы. 2021. С. 165-168.

23 Субботина С.А., Шлыкова И.Д., Авдеева А.А., Одинокова Г.В., Соколова Н.В. Виды композитных материалов: стеклопластик, углепластик, базальтопластик // Синергия наук. 2017. № 18. − С. 641-645.

24 Таршхоева Ж.Т. Zзык программирования Python. Библиотеки Python // Молодой ученый. 2021. № 5 (347). С. 20-21.

25 Токарев В. Сравнение арматурных прутьев из базальтопластика и углепластика [Электронный ресурс]: – Режим доступа: http://cemgid.ru/sravnenie-armaturnyx-prutev-iz-bazaltoplastika-i-ugleplastika.html (дата обращения: 12.06.2022).

26 Щелконогов А.Н. Разработка простейших нейросетей в keras // Математическое и программное обеспечение вычислительных систем. 2019. С. 51-53.

27 Выбор алгоритмов машинного обучения Azure [Электронный ресурс]: – Режим доступа: https://docs.microsoft.com/ru-ru/azure/machine-learning/how-to-select-algorithms (дата обращения: 13.06.2022).

28 Нейросеть (регрессия) [Электронный ресурс]: – Режим доступа: https://help.loginom.ru/userguide/processors/datamining/neural-network-regression.html (дата обращения: 13.06.2022).

29 Машинное обучение [Электронный ресурс]: – Режим доступа: https://ru.wikipedia.org/wiki//Машинное_обучение (дата обращения: 13.06.2022).

30 Комплексная платформа машинного обучения с открытым исходным кодом [Электронный ресурс]: – Режим доступа: https://www.tensorflow.org/ (дата обращения: 12.06.2022).

31 Платформа scikit-learn [Электронный ресурс]: – Режим доступа: https://scikit-learn.org/stable/ (дата обращения: 12.06.2022).