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 Калиновская М.В.

Москва, 2022

Содержание:

1. Аналитическая часть

1.1. Постановка задачи.

1.2. Описание используемых методов

1.3. Разведочный анализ данных

2. Практическая часть

2.1. Предобработка данных

2.2. Разработка и обучение модели

2.3. Тестирование модели

2.4. Построение нейронной сети для рекомендации соотношение матрица-наполнитель.

Заключение

Список использованной литературы

1. Аналитическая часть

1.1. Постановка задачи.

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

Описание:

В настоящее время композиционные материалы (КМ) на полимерных и металлических матрицах находят всё более широкое применение в различных отраслях промышленности в качестве конструкционных материалов. Внедрение КМ обусловлено стремлением использовать их преимущества по сравнению с традиционно используемыми металлами и сплавами. Уникальность композиционных материалов проявляется в их высоких значениях удельной жесткости (отношения модуля упругости к плотности) и удельной прочности (отношения предела прочности к плотности), химической и коррозионной стойкости к агрессивным средам, анизотропии свойств и возможности их варьирования для наилучшего восприятия действующих нагрузок. Внедрение КМ в конструкцию различных агрегатов и узлов позволяет снизить массовые характеристики изделия, увеличить ресурс и срок службы, уменьшить издержки, связанные с обслуживанием композитных конструкций в эксплуатации.

Разработка новых энерго- и ресурсосберегающих технологий для изготовления деталей и агрегатов из КМ способствует снижению себестоимости изделий, уменьшению издержек, меньшему негативному влиянию на окружающую среду.

Расширение использования композитов в различных отраслях связано с возможностью реализации таких свойств КМ, как:

Повышенная вибрационная стойкость, что позволяет использовать КМ в зонах действия повышенных вибрационных нагрузок;

Высокий коэффициент затухания волн в КМ, что обеспечивает надежное гашение вибраций, особенно высокого значения декремента затухания колебаний возможно достичь при применении в вибропоглощающих конструкциях органопластиков – материалов на основе пара- и метаарамидных волокон и полимерных связующих;

Хорошие демпфирующие свойства стеклопластиков, базальтопластиков и органопластиков, что позволяет применять КМ в качестве материала демпферов, защитных кожухов, корпусов и гасителей ударных динамических воздействий на узлы;

Высокие значения шумопоглощения, что позволяет снизить вредное акустическое воздействие на обслуживающий персонал;

Высокие прочностные и жесткостные свойства конструкционных углепластиков, что дает возможность применять КМ в средне- и высоконагруженных узлах и агрегатах;

Химическая и коррозионная стойкость КМ, что позволяет внедрять подобные материалы в эксплуатацию в агрессивных средах;

Высокая усталостная прочность, обусловленная анизотропией свойств композиционных пластиков, в том числе слоистых, что приводит к наличию у КМ высоких коэффициентов трещиностойкости, и как следствие, к высоким параметрам усталостной прочности.

В то же время композитам присущи высокие жесткостные и прочностные свойства в направлении армирования – ориентирования волокон – что достигается применением материалов на основе углеродных, борных волокон и тонких металлических проволок. Удельные прочностные и жесткостные характеристики однонаправленных КМ в несколько раз выше по сравнению с традиционными материалами, что позволяет разрабатывать из КМ конструкции высоконагруженных деталей и узлов, несущих элементов конструкций, изделий с высокой весовой отдачей.

Применение КМ в ответственных деталях, агрегатах и изделиях накладывает на них весьма жесткие требования по прочности, выносливости, усталости, ресурсу и ремонтопригодности.

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 (количество связующего, наполнителя, температурный режим отверждения и т.д.). 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).

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

1.2. Описание используемых методов

Для решения поставленной задачи регрессии применим следующие методы:

* линейная регрессия;
* градиентный бустинг;
* К-ближайших соседей;
* дерево решений;
* метод опорных векторов;
* Лассо;
* «Случайный лес».

Градиентный бустинг, он же Gradient Boosting, Gradient Boosted-Tree, более правильно будет назвать ансамбль деревьев решений, обученный с использованием градиентного бустинга.

Градиентный бустинг представляет собой ансамбль деревьев решений. В основе данного алгоритма лежит итеративное обучение деревьев решений с целью минимизировать функцию потерь. Благодаря особенностям деревьев решений градиентный бустинг способен работать с категориальными признаками , справляться с нелинейностями.

Бустинг – это ансамблевый метод, в котором алгоритмы применяются последовательно. Этот метод использует логику, в которой последующие модели учатся на ошибках предыдущих. Следовательно, Наблюдения (Observation) имеют неодинаковую вероятность появления в последующих моделях, а наблюдения с наибольшей ошибкой появляются чаще. (Таким образом, наблюдения выбираются не на основе процесса начальной загрузки, а на основе ошибки). Из используемых методов перечислим деревья решений, регрессоры, классификаторы и т.д. Поскольку новые алгоритмы учатся на ошибках, совершенных предшественниками, требуется меньше времени / итераций, чтобы приблизиться к фактическим прогнозам. Но мы должны тщательно выбирать критерии остановки, иначе это может привести к Переобучению (Overfitting).

Метод ближайших соседей (kNN - k Nearest Neighbours) - метод решения задач классификации и задач регрессии, основанный на поиске ближайших объектов с известными значения целевой переменной.

Алгоритм находит расстояния между запросом и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии).

Преимущества данного метода – простая реализация, низкая чувствительность к выбросам, отсутствие необходимости строить модель, настраивать несколько параметров или делать дополнительные допущения, универсальность.

Метод k-ближайших соседей (k-nearest neighbors) – это простой алгоритм машинного обучения с учителем, который можно использовать для решения задач классификации и регрессии. Он прост в реализации и понимании, но имеет существенный недостаток – значительное замедление работы, когда объем данных растет.

Дерево принятия решений (также называют деревом классификации или регрессионным деревом) — средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике. Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны признаки, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — признаки, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

Метод Опорных Векторов или SVM (от англ. Support Vector Machines) — это линейный алгоритм используемый в задачах классификации и регрессии. Данный алгоритм имеет широкое применение на практике и может решать как линейные так и нелинейные задачи. Суть работы “Машин” Опорных Векторов проста: алгоритм создает линию или гиперплоскость, которая разделяет данные на классы.

Данный метод используется для задач классификации и регрессионного анализа. Основной метод опорных векторов принимает набор входных данных и прогнозирует для каждого данного входа одну из двух возможных форм выхода. Благодаря такому процессу, данный метод случайным является бинарным линейным классификатором.

Учитывая набор обучающих наблюдений (обучающую выборку), каждое из которых помечена как принадлежащая к одной из двух категорий, алгоритм обучения метода опорных векторов строит модель, которая определяет новые наблюдения в одну из категорий.

Модель метода опорных векторов – отображение данных точками в пространстве, так что между наблюдениями отдельных категорий имеется разрыв и он максимален.

Затем новые наблюдения отобразятся в том же пространстве и будут относиться к одной из категорий в зависимости от того, по какую сторону разрыва они отобразились.

Lasso (Least absolute shrinkage and selection operator) - метод оценивания коэффициентов линейной регрессионной модели.

Метод заключается во введении ограничения на норму вектора коэффициентов модели, что приводит к обращению в 0 некоторых коэффициентов модели. Метод приводит к повышению устойчивости модели в случае большого числа обусловленности матрицы признаков X, позволяет получить интерпретируемые модели - отбираются признаки, оказывающие наибольшее влияние на вектор ответов.

Случайный лес (random forest) — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

* Выбирается подвыборка обучающей выборки размера samplesize (м.б. с возвращением) – по ней строится дерево (для каждого дерева — своя подвыборка).
* Для построения каждого расщепления в дереве просматриваем max\_features случайных признаков (для каждого нового расщепления — свои случайные признаки).
* Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Такая схема построения соответствует главному принципу ансамблирования (построению алгоритма машинного обучения на базе нескольких, в данном случае решающих деревьев): базовые алгоритмы должны быть хорошими и разнообразными (поэтому каждое дерево строится на своей обучающей выборке и при выборе расщеплений есть элемент случайности).

1.3. Разведочный анализ данных

В ходе решения поставленной задачи применим следующие инструменты разведочного анализа:

* гистограммы распределения каждой из переменной;
* диаграммы ящика с усами;
* попарные графики рассеяния точек;
* описательная статистика для каждой переменной;
* анализ и исключение выбросов;
* проверка наличия пропусков.

Для знакомства с данными изучим описательную статистику:

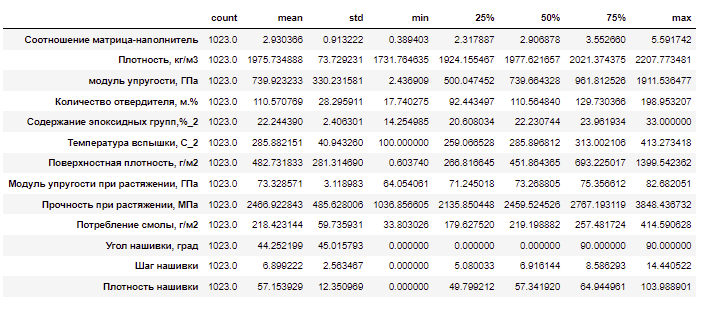


Рисунок 1. Описательная статистика

В наборе данных 13 параметров. Каждый параметр содержит 1023 значения.

Гистограммы используются для изучения распределений частот значений переменных. Такое частотное распределение показывает, какие именно конкретные значения или диапазоны значений исследуемой переменной встречаются наиболее часто, насколько различаются эти значения, расположено ли большинство наблюдений около среднего значения, является распределение симметричным или асимметричным, многомодальным (т.е. имеет две или более вершины) или одномодальным и т.д.

По форме распределения можно судить о природе исследуемой переменной (например, бимодальное распределение позволяет предположить, что выборка не является однородной и содержит наблюдения, принадлежащие двум различным множествам, которые в свою очередь нормально распределены).

При проведении анализа выявлены параметры близкие к нормальному:

* 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;
* Шаг нашивки;
* Плотность нашивки.

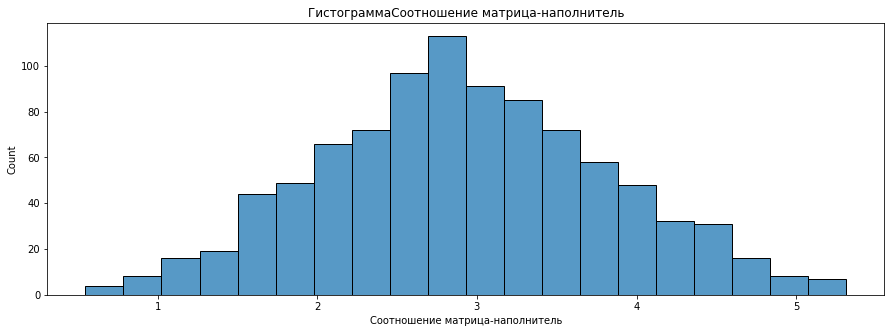
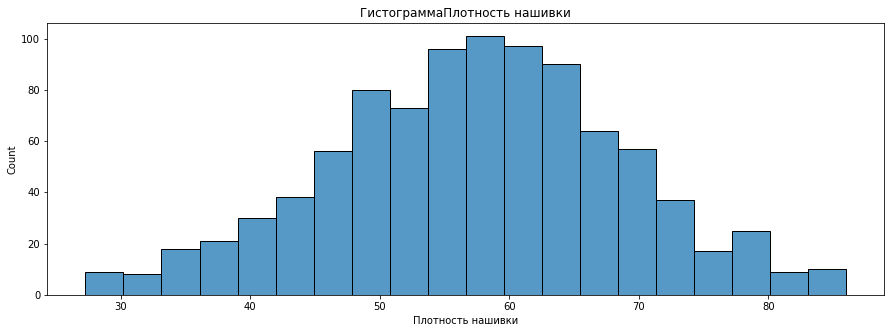
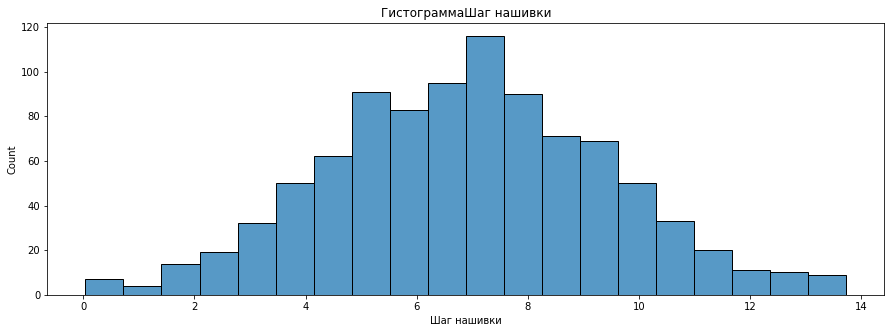
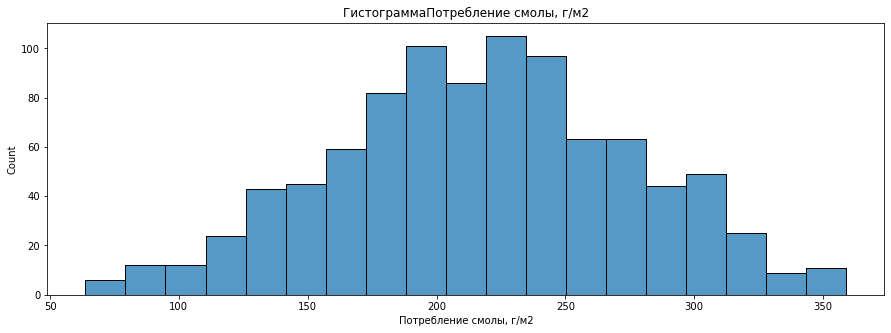
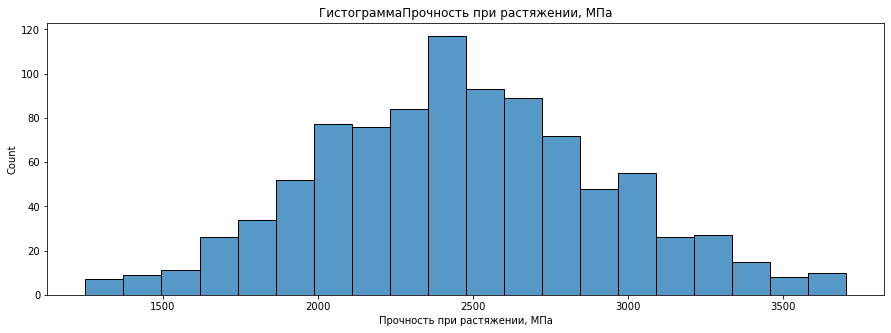
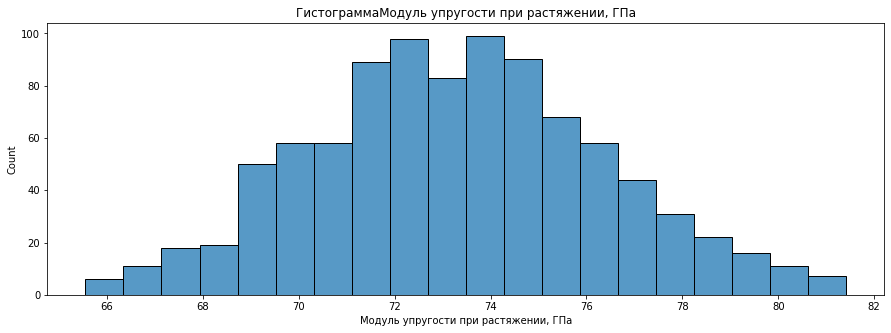
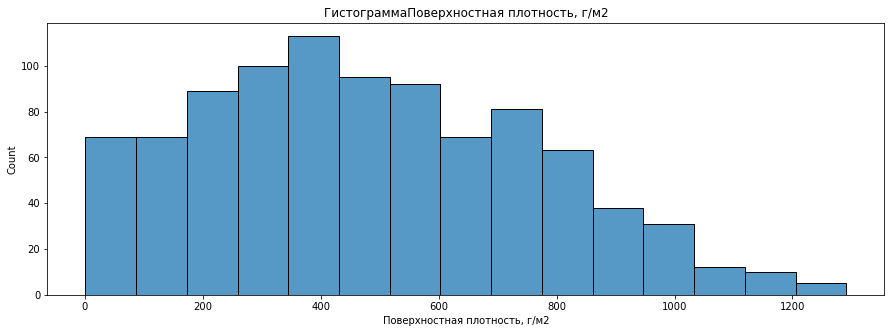
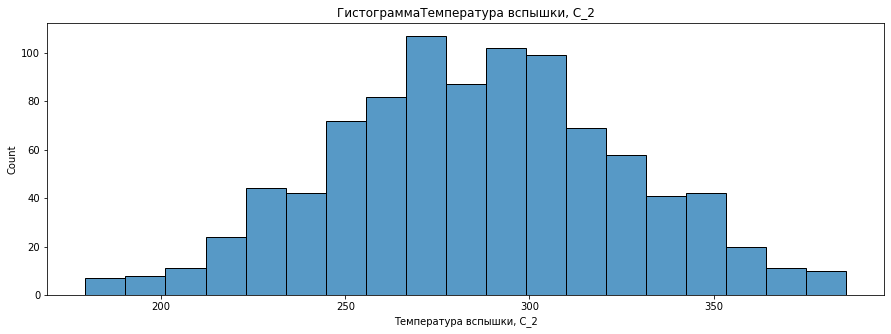
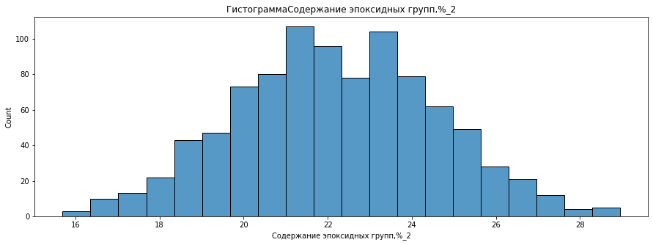
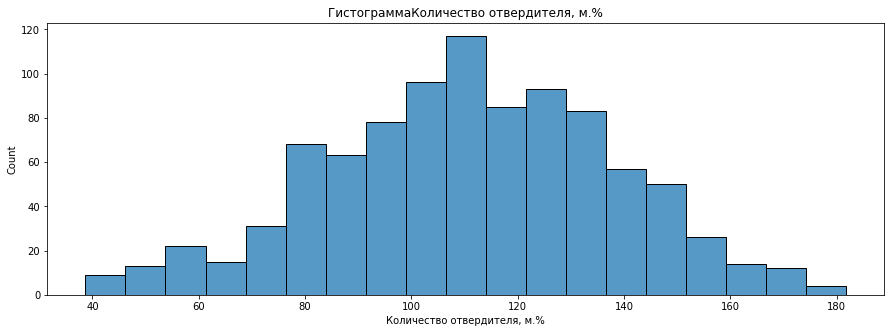
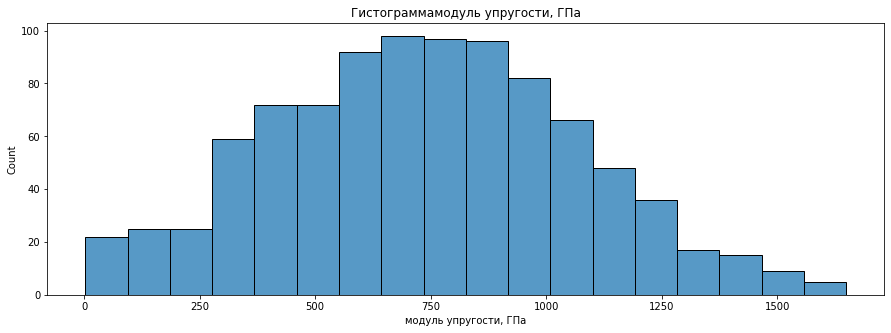
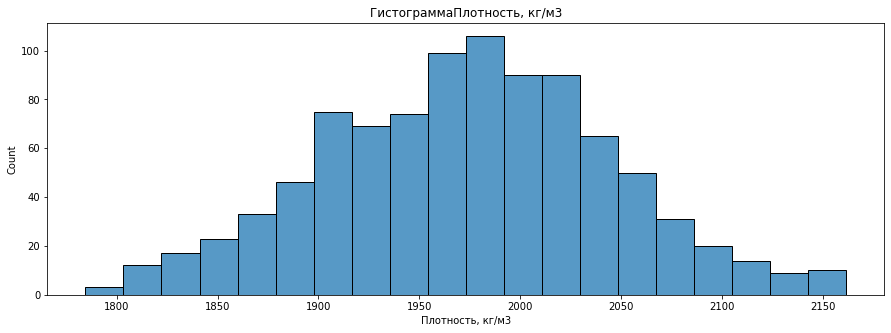


Рисунок 2. Гистограммы переменных с нормальным распределением

Параметр, имеющий дискретное распределение:

* Угол нашивки, град.



Рисунок 3. Гистограмма переменной «Угол нашивки»

Двумерные диаграммы рассеяния используются для визуализации взаимосвязей между двумя переменными X и Y (например, весом и ростом). На этих диаграммах отдельные точки данных представлены маркерами на плоскости, где оси соответствуют переменным. Две координаты (X и Y), определяющие положение точки, соответствуют значениям переменных. Если между переменными существует сильная взаимосвязь, то точки на графике образуют упорядоченную структуру (например, прямую линию или характерную кривую). Если переменные не взаимосвязаны, то точки образуют "облако", как в нашем случае.

С помощью диаграмм рассеяния можно исследовать и нелинейные взаимосвязи между переменными. При этом не существует каких-либо "автоматических" или простых способов оценки нелинейности. Стандартный коэффициент корреляции Пирсона r позволяет оценить только линейность связи, а некоторые непараметрические корреляции, например, Спирмена R, дают возможность оценить нелинейность, но только для монотонных зависимостей. На диаграммах рассеяния можно изучить структуру взаимосвязей, чтобы затем с помощью преобразования привести данные к линейному виду или выбрать подходящую нелинейную подгонку

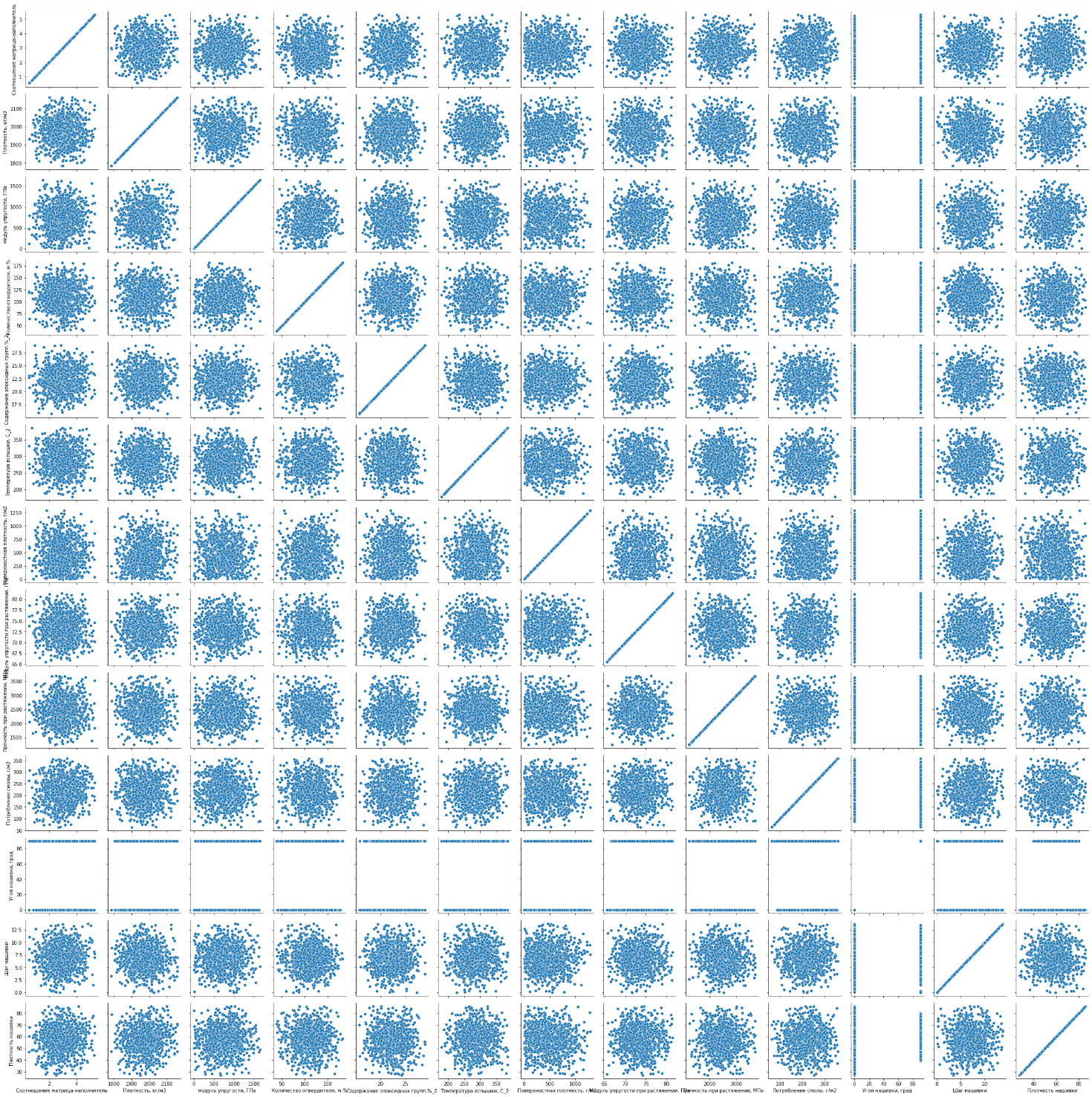


Рисунок 4. Диаграммы рассеивания

Ящик с усами, диаграмма размаха (англ. box-and-whiskers diagram or plot, box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

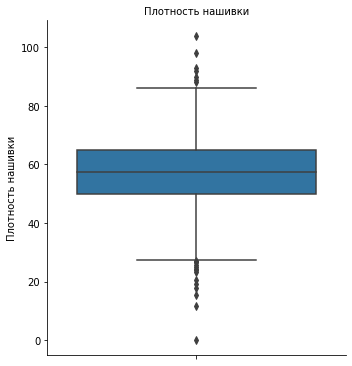
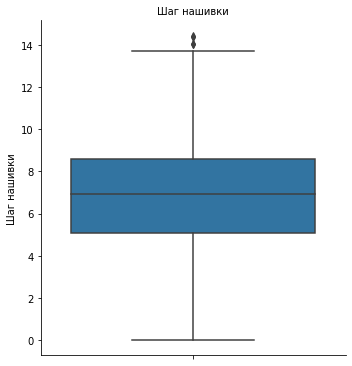
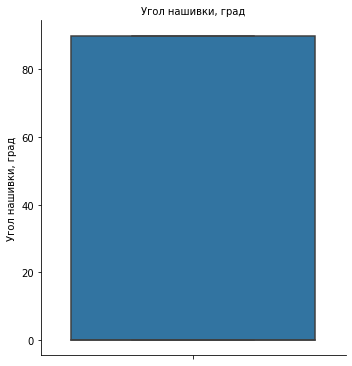
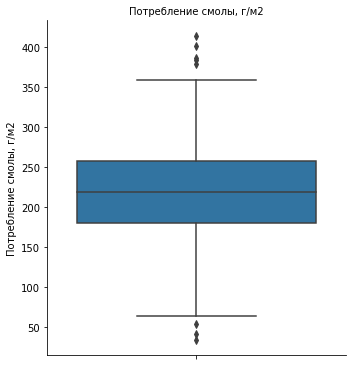
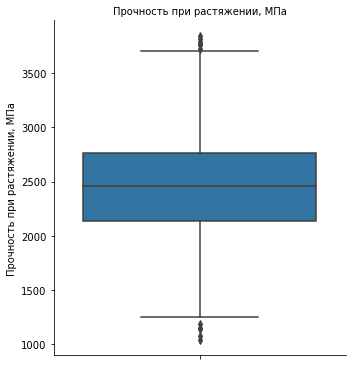
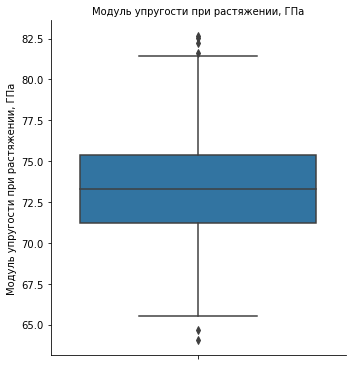
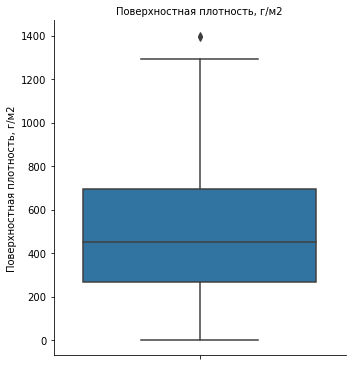
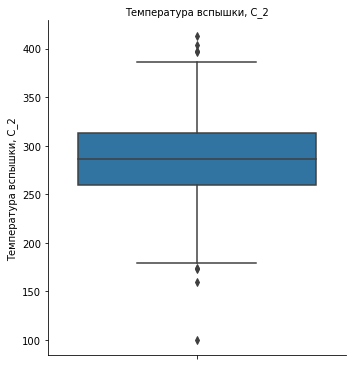
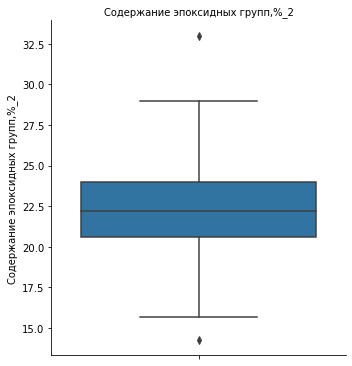
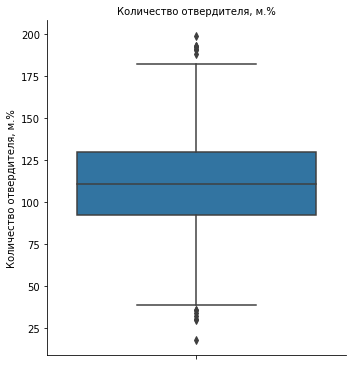
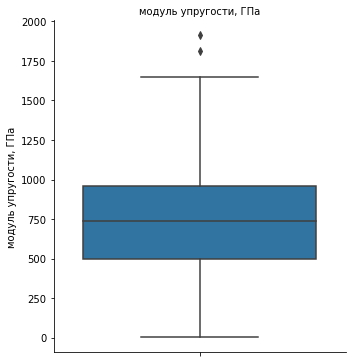
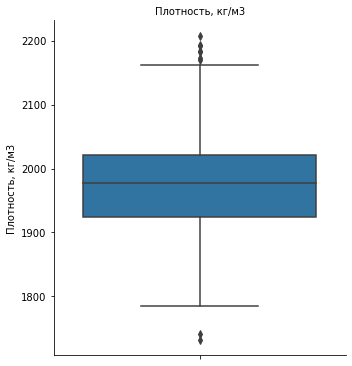
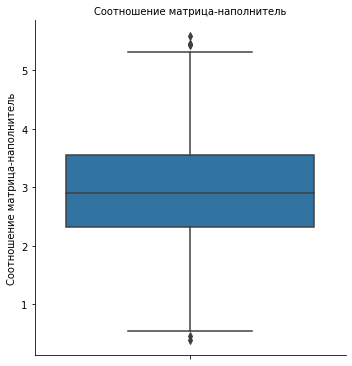


Рисунок 5. Ящик с усами

Проанализировав диаграммы ящика с усами делаем вывод, что пригодным является параметр "Угол нашивки". В остальных данных необходима работа с выбросами.

Выброс — это экстремальные значения во входных данных, которые находятся далеко за пределами других наблюдений.

Многие алгоритмы машинного обучения чувствительны к разбросу и распределению значений признаков обрабатываемых объектов. Соответственно, выбросы во входных данных могут исказить и ввести в заблуждение процесс обучения алгоритмов машинного обучения, что приводит к увеличению времени обучения, снижению точности моделей и, в конечном итоге, к снижению результатов.

Причины возникновения выбросов: сбой работы оборудования; человеческий фактор; случайность; уникальные явления и др.

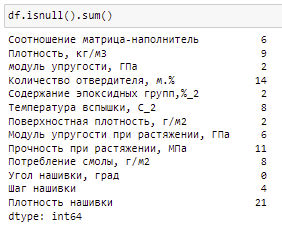


Рисунок 6. Анализ наличия выбросов

Максимальное количество выбросов содержит переменная «Плотность ошибки». Так количество выбросов небольшое принимаем решение удалить выбросы.

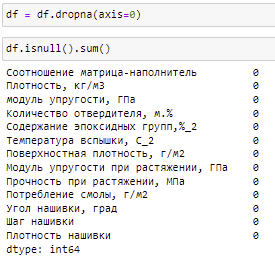


Рисунок 7. Удаление выбросов

Наряду с некорректными типами данных одна из самых частых проблем -отсутствующие значения. Они могут отсутствовать по разным причинам, и перед обучением модели эти значения нужно либо заполнить, либо удалить. Проведем анализ отсутствующих значений.

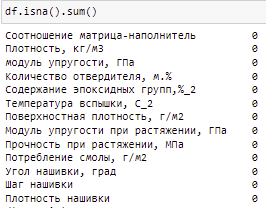


Рисунок 8. Анализ отсутствующих значений

Как видим пропуски отсутствуют. Можем переходить к следующему шагу.

Построим тепловую карту:

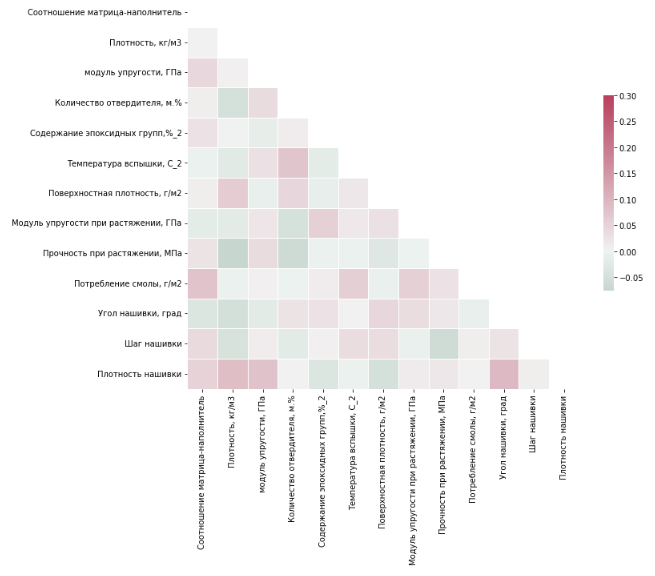


Рисунок 9. Тепловая карта

Проанализировав тепловую карту получаем вывод - корреляция очень низкая.

2. Практическая часть

2.1. Предобработка данных

В ходе проведенного анализа принимаем решение закодировать столбец "Угол нашивки" с помощью LabelEncoder.

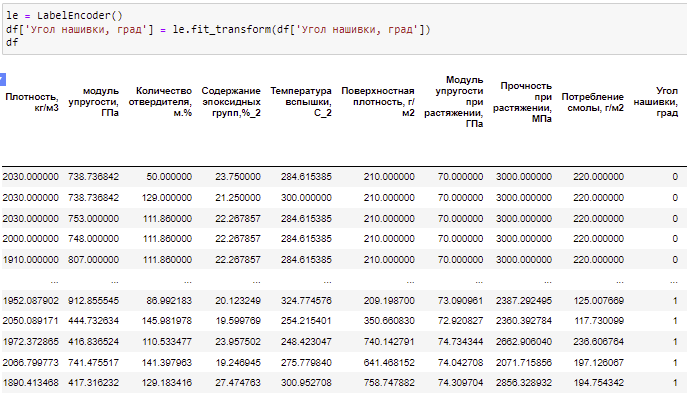


Рисунок10. Использование LabelEncoder

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.

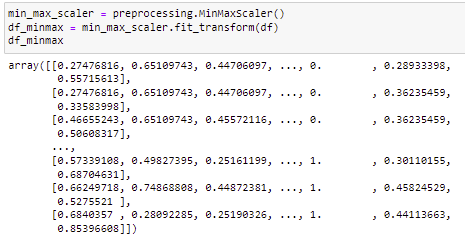


Рисунок 11. Использование MinMaxScaler

Масштабировать будем с помощью приведения каждого признака к диапазону от 0 до 1 с помощью метода MinMaxScaler.

2.2. Разработка и обучение модели

После всех подготовительных работ переходим к процессу создания, обучения моделей. Мы будем использовать в Python библиотеку Scikit-Learn.

В качестве базового уровня предскажем медианное значение цели на обучающем наборе для всех примеров в тестовом наборе. В качестве метрики возьмём среднюю абсолютную ошибку (mae) в прогнозах. Для обучения используем 70 % данных, а для тестирования — 30 %.

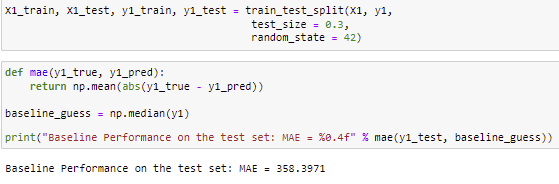


Рисунок 12. Построение базовой модели для прогноза прочности при растяжении

Для решения поставленной задачи регрессии применим следующие методы:

* линейная регрессия;
* градиентный бустинг;
* К-ближайших соседей;
* дерево решений;
* метод опорных векторов;
* Лассо;
* «Случайный лес».

Реализуем в Scikit-Learn модели машинного обучения для прогноза прочности при растяжении.

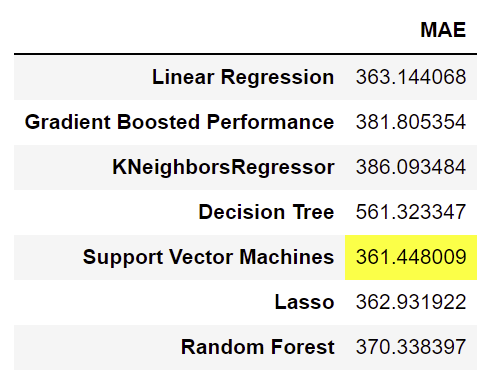


Рисунок 13. Построение моделей для прогноза прочности при растяжении

Чтобы объективно оценивать модели, мы с помощью медианного значения цели вычислили базовый уровень и получили 358,3971. А полученные результаты немного хуже.

Лучший результат показал метод "Метод опорных векторов". Реализуем метод настройки гиперпараметров с помощью поиска по сетке с перекрёстной проверкой для метода опорных векторов.

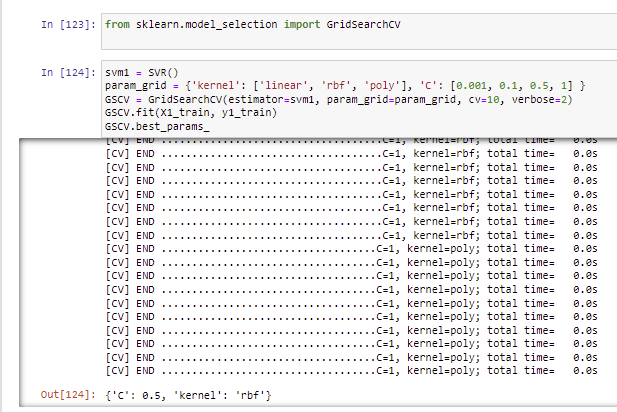


Рисунок 14. Метод настройки гиперпараметров модели для прогноза прочности при растяжении

Модель после настройки гиперпараметров показала результат немного лучше. Однако, ниже, чем базовая модель.

Прочность при растяжении не имеет линейной зависимости. Все использованные модели не справились с задачей.

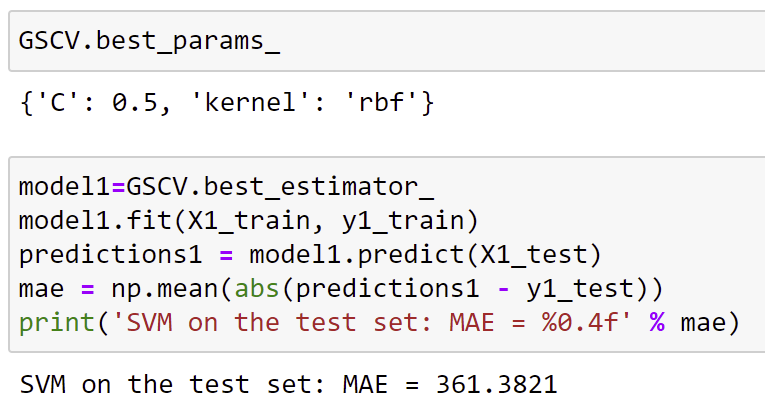


Рисунок 15. Результаты настроенной модели для прогноза прочности при растяжении

Повторим алгоритм для прогноза модуля упругости при растяжении.

Построим базовую модель.

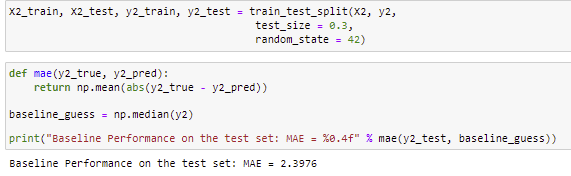


Рисунок 16. Построение базовой модели для прогноза модуля упругости при растяжении

Реализуем в Scikit-Learn модели машинного обучения.

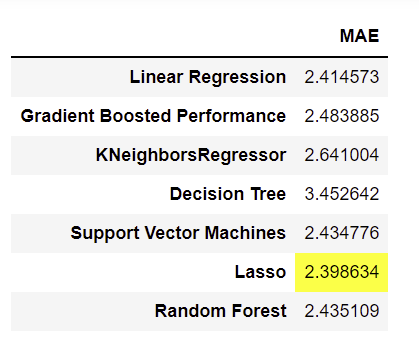


Рисунок 17. Построение моделей для прогноза модуля упругости при растяжении

Результаты использованных моделей, как и в предыдущем случае, оказались немного хуже базовой модели. Попробуем провести настройку гиперпараметров модели «Лассо».

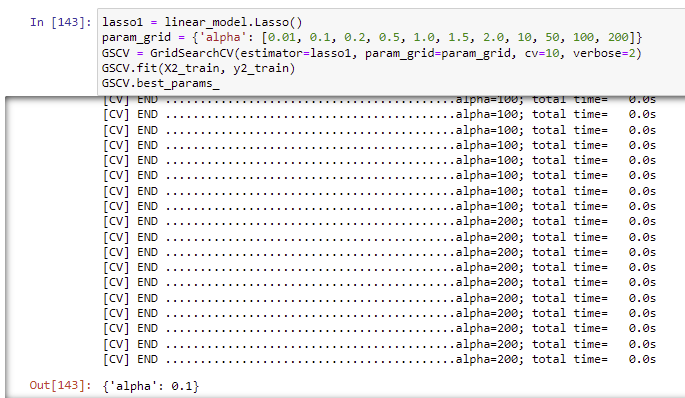


Рисунок 18. Метод настройки гиперпараметров модели для прогноза модуля упругости при растяжении.

Настроенная модель показала результат немного лучше базовой модели. Примем результат как неудовлетворительный. Модуль упругости не имеет линейной зависимости. Все использованные модель не справились с задачей. Свойства композитных материалов в первую очередь зависят от используемых материалов.

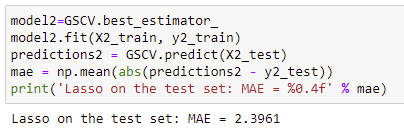


Рисунок 19. Результаты настроенной модели для прогноза модуля упругости при растяжении.

2.3. Тестирование модели

Проведем оценку точности каждой из настроенных моделей.

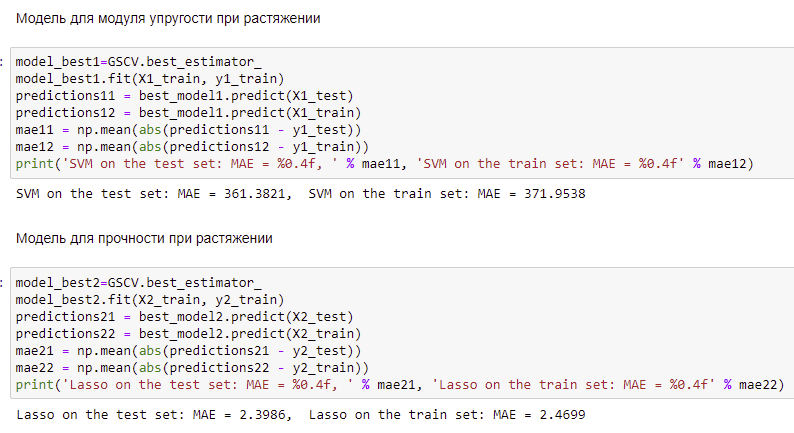


Рисунок 20. Оценка точности моделей.

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовых выборках. В качестве параметра оценки модели использовалась средняя абсолютная ошибка (MAE). Обе модели даже на тренировочном датасете не смогли обучиться и приблизиться к исходным данным. Поэтому ошибка на тестовом датасете выше.

2.4. Написать нейронную сеть, которая будет рекомендовать соотношение матрица.

Построим базовую модель.

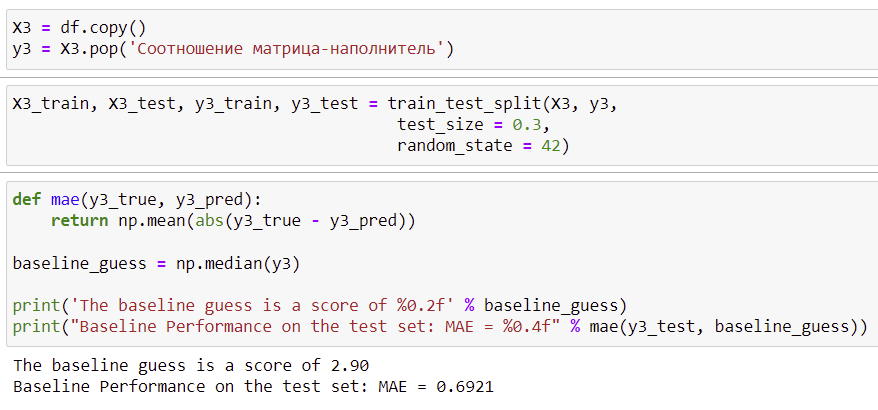


Рисунок 21. Базовая модель для прогнозирования соотношения матрица-наполнитель.

Для рекомендации соотношения «матрица-наполнитель» была разработана простая модель глубокого обучения.

Архитектура нейронной сети может быть описана следующим образом.

Модель состоит из четырех скрытых уровней. Первый содержит 64 нейрона. Последующие скрытые уровни – они содержат 16, 4 и 4 нейрона. Снижение числа нейронов на каждом уровне сжимает информацию, которую сеть обработала на предыдущих уровнях.

Для эксперимента был выбран relu (выпрямленная линейная единица).

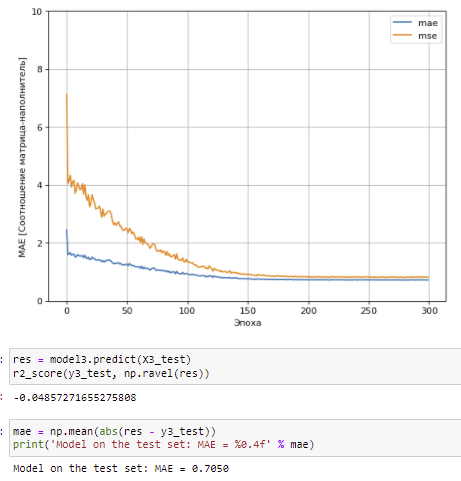


Рисунок 22. Результаты нейронной сети для соотношения матрица-наполнитель.

После обучения модели для была определена средняя абсолютная ошибка и среднеквадратическая ошибка на тестовом наборе данных в соответствии с рисунком 22. Результат оказался хуже базовой модели.

Заключение

В связи с неудовлетворительными результатами использованных моделей считаю разработку приложения нецелесообразным.

Результаты работы размещены на GitHub:

<https://github.com/00Maria/VKR.git>

Список используемой литературы

[1] А.А. Ларин, Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с.

[2] Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.

[3] Язык программирования Python: - Режим доступа: https://www.python.org/ (дата обращения 27.04.2022)

[4] Библиотека Pandas – Режим доступа: https://pandas.pydata.org/ (дата обращения 27.04.2022)

[5] Библиотека Matplotlib – Режим доступа: https:// https://matplotlib.org/ (дата обращения 01.04.2022)

[6] Библиотека Sklearn – Режим доступа: https://scikit-learn.org/stable/ (дата обращения 27.04.2022)

[7] К. Андерсон, Аналитическая культура. От сбора данных до бизнес-результатов: монография. Москва: O’Reilly, 2017, 392 с.

[8] Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) – Режим доступа: https://habr.com/ru/post/428503/ (дата обращения 27.04.2022)