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 Козырев Виталий Владимирович

Москва, 2023

Содержание

[1. Аналитическая часть 3](#_Toc129567798)

[1.1. Постановка задачи 3](#_Toc129567799)

[1.2. Описание входных данных 4](#_Toc129567800)

[1.3. Разведочный анализ данных 7](#_Toc129567801)

[2. Практическая часть 16](#_Toc129567802)

[2.1. Предобработка данных 17](#_Toc129567803)

[2.2. Выбор наилучшего метода 20](#_Toc129567804)

[3. Нейронная сеть по оценке показателя «Соотношение матрица-наполнитель» 28](#_Toc129567805)

[4. Создание приложения 30](#_Toc129567806)

[5. Создание репозитория на GitHub 32](#_Toc129567807)

[Заключение 33](#_Toc129567808)

[Список литературы и веб-ресурсов 34](#_Toc129567809)

# Аналитическая часть

## Постановка задачи

Композиционные материалы - это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т.е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом.

Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства.

Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик.

Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента). 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.

Данная ВКР основана на реальных данных одной из производственных задач Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

В ходе ВКР необходимо выполнить нижеперечисленные задачи:

1. Да входе имеются два датасета с данными, которые нужно объединить по индексу (тип объединения INNER).

1. Обучить алгоритм машинного обучения, который будет определять значения двух целевых переменных:

* Модуль упругости при растяжении, ГПа
* Прочность при растяжении, Мпа

1. Написать нейронную сеть, которая будет рекомендовать ещё одну целевую переменную:

* Соотношение матрица-наполнитель

1. Написать приложение, которое будет выдавать прогноз, полученный в задании 2 или 3 (один или два прогноза, на выбор учащегося).

## Описание входных данных

В первом датасете (файл “X\_bp.csv”) содержится 10 признаков (характеристик), описывающих свойства компонентов композиционных материалов:

* Соотношение матрица-наполнитель
* Плотность
* Модуль упругости
* Количество отвердителя
* Содержание эпоксидных групп
* Температура вспышки
* Поверхностная плотность
* Модуль упругости при растяжении
* Прочность при растяжении
* Потребление смолы

При этом «Соотношение матрица-наполнитель», «Модуль упругости при растяжении» и «Прочность при растяжении» - это объясняемые переменные, т.е. целевые переменные (targets), все остальные признаки – это объясняющие переменные (features).

Второй датасет (файл “X\_nup.csv”) содержит ещё три объясняющих переменных (характеристики/признаки исходных компонентов):

* Угол нашивки
* Шаг нашивки
* Плотность нашивки

По своему типу все исходные данные – это непрерывные числовые данные с плавающей запятой, кроме признака «Угол нашивки», который имеет всего два значения - 0 и 90 градусов, поэтому далее приведём этот признак к категориальному типу.

Graphical user interface, text

Description automatically generatedText

Description automatically generated

Объединение двух датасетов будем производить по уникальному индексу (по условиям задачи), при этом они (датасеты) имеют различие (несовпадение) по 17 позициям.

A screenshot of a computer

Description automatically generated with medium confidence

Осуществляем объединение двух датасетов через метод JOIN, и в результате получаем таблицу данных размерностью 1023 строки (наблюдений) на 13 колонок (, т.е. признаков, среди которых находятся наши целевые переменные).

A picture containing graphical user interface

Description automatically generated

Для удобства дальнейшей работы переименуем данные (колонки датасета) – приведём к более лаконичной форме наименований.

Text

Description automatically generated

И ещё один штрих - приводим значения признака «Угол нашивки» к категориальному типу.

Graphical user interface, application

Description automatically generated

## Разведочный анализ данных

Первое, что обращает на себя внимание – это наличие в исходном датасете ряда повторяющихся целочисленных значений (выделены на изображении красным цветом), что выглядит подозрительно на фоне остальных нецелых данных с плавающей запятой.

Есть предположение, что такие значения заносились в базу данных вручную с грубым округлением или иными упрощениями, а, значит, такие значения могут являться ненужным мешающим шумом, искажающим истинные взаимосвязи и зависимости между данными.

С учётом достаточного количества наблюдений (порядка 1023) можно безболезненно избавиться от таких подозрительных позиций – в итоге, их всего 23.

Chart

Description automatically generated

Ещё из общей картины выбивается одно нулевое значение признака «Плотность нашивки» (cover\_strength), но оно благополучно попадает в ряды выше обозначенных «бракованных» записей, от которых осознанно решаем избавиться.

Graphical user interface, table

Description automatically generated with medium confidence

Далее смотрим на основные описательные статистики данных и видим, что…

* значения признаков имеют разный масштаб, т.е. неизбежен шаг по нормализации данных;
* судя по схожести значений средних значений (mean) со значениями второго квартиля (50%), данные, скорее всего, имеют распределение, близкое к нормальному, что хорошо, с т.з. дальнейшей работы алгоритмов с такими данными.

Graphical user interface, text, application

Description automatically generated

Через построение гистограмм визуально убеждаемся в предположении о почти нормальном распределении наших трёх целевых переменных

Chart, histogram

Description automatically generated

Что касается остальных признаков – объясняющих переменных, то также оцениваем тип распределения через тест Shapiro-Wilk и построение QQ-plots.

Text

Description automatically generated

Видим, что три признака имеют распределение, отличное от нормального, и с этим можно поработать – попытаться исправить через Box-Cox преобразование.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Пытаемся привести распределение значений трёх признаков - модуль упругости (elasticity), поверхностная плотность (surface density) и плотность нашивки (cover strength) - максимально близко к нормальному через Box-Cox преобразование.

Chart, line chart

Description automatically generated

Shape

Description automatically generated with low confidence

A picture containing shape

Description automatically generated

В целом, можно почистить датасет от выбросов, которые можно наблюдать на диаграмме «ящик с усам».

Chart, box and whisker chart

Description automatically generated

Избавляемся от выбросов через подход с межквартальным интервалом.

A screenshot of a computer

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generated

Смотрим на взаимосвязи данных через график парного рассеивания точек (pairplot) и приходим к следующим наблюдениям:

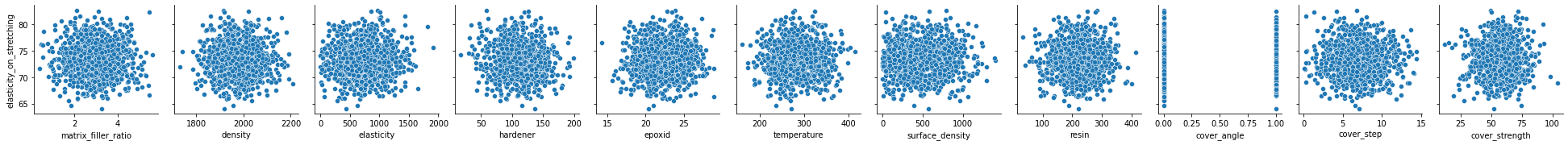
* ещё раз убеждаемся в предположении от том, что распределение данных всех признаков очень близко к нормальному, что хорошо;
* не очень радует наблюдение относительно отсутствия каких-либо явных линейных или нелинейные взаимосвязей между данными, что скорее плохо для задачи регрессии

A picture containing shoji, building

Description automatically generated

Отдельно смотрим на поведение данных во взаимосвязи с нашими целевыми переменными – модуль упругости при растяжении (elasticity on stretching) и прочность при растяжении (solidity on stretching) - фиксируем факт отсутствия прямой линейной зависимости, что, повторюсь, скорее всего, означает сложности при решении задачи регрессии и построения прогноза характеристик новых материалов.

График парного рассеивания точек по целевой переменной «Модуль упругости при растяжении» (elasticity on stretching).



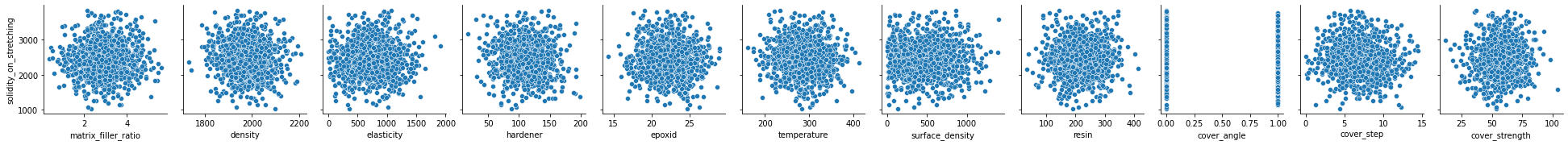
Корреляция между целевой переменной и объясняющими переменными отсутствует.

Text

Description automatically generated

Аналогичная картина и для второй целевой переменной.

График парного рассеивания точек по целевой переменной «Прочность при растяжении» (solidity on stretching).



Text

Description automatically generated

Хорошее возможное (или, возможно, хорошее) решение по улучшению результатов регрессионной модели при работе с такими данными – это создание новых (синтетических) признаков, т.е. речь идёт про feature engineering, в частности, можно добавить дополнительные признаки через …

* попарное перемножение признаков
* возведение признаков в степень (но, похоже, с такими данными это не сильно поможет)
* через метод главных компонент (PCA)

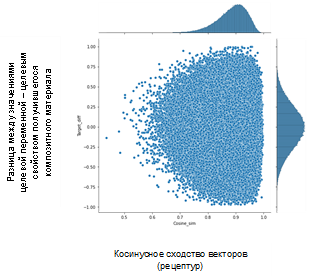
… Что и будет проделано в практической части.

**Однако посмотрим на взаимоотношение данных иначе.**

Если нет линейной зависимости (отсутствует корреляция) между целевой переменной и отдельными признаками (т.е. по вертикали = feature-wise), то, возможно, целевая переменная реагирует на набор признаков, как единое целое – собственно, это и есть "рецептура" (состав композитного материала). Другими словами, важна взаимосвязь признаков между собой по горизонтали - по строкам (sample-wise).

Тогда можно предположить, что относительно схожим значениям целевой переменной (т.е. разница д.б. близка к 0 или около того) должны соответствовать оносительно схожие по своему составу ("рецептуре") испытываемые образцы композитных материалов.

Отсюда сделаем следующее упражнение...Попарно сравним все наблюдения (эксперименты) между собой на предмет схожести / разности через косинусное расстояние (строка признаков – это вектор, описывающий "рецептуру" разных образцов композитного материала) и при этом будем фиксировать разницу между значениями целевой переменной сравниваемых векторов.



В итоге получаем следующую картину – см. изображение.

Это пример по «Модуль упругости при растяжении» (elasticity on stretching). Аналогичный результат получается и для «Прочность при растяжении» (solidity on stretching).

Получается странная картина: чем больше разница между векторами "рецептур" (т.е. косинусное расстояние дальше от «1»), тем меньше разница в значениях целевых переменных, т.е. конечном свойстве композитного материала.

И наоборот - чем выше сходство векторов признаков по горизонтали (по экспериментам = по набюджениям), тем больше разница (дисперсия) между получаемыми конечными свойствами композитного материла.

И такое наблюдение ещё больше вносит сомнений в успешность получения качественной регрессионной модели.

**Выводы, основанные на предварительных попытках решить данную задачу через классическую регрессию:**

1. С таким набором данных со слабыми связями с целевой переменной (т.е. отсутствием корреляций) классическая регрессия, скорее всего, не даст хорошего результата.
2. В условиях, когда, практически, все эксперименты (все наблюдения) дают разные результаты (т.е. нет систематических повторений значений целевой переменной от схожих исходных параметров) при выборе наилучшего метода лучше использовать подход с кросс-валидацией по всему датасету, вместо его разбиения на train и test подвыборки.
   * Возможно, самое идеальное решение – это сравнение моделей через LeaveOneOut кросс-валидацию, но с датасетом такого размера это очень ресурсоёмко, в т.ч. и по времени.
3. Задачу с прогнозом свойств композитного материала можно решить альтернативным путём, перейдя от регрессии к задаче классификации, разбив целевую переменную на интервалы значений (bins), т.е. классы, и каждому классу (интервалу) будет соответствовать некое среднее значение целевой переменной.
   * Такой подход рассмотрим в практической части.

# Практическая часть

Важно обозначить, что по данному проекту, ввиду отсутствия должного опыта и знаний в области композиционных материалов, у меня были уточняющие вопросы относительно целевых переменных (их три – см. выше часть п.1 Постановка задачи) и их взаимосвязи, что важно для корректной сборки датасета для моделей с соблюдением причинно-следственных связей характеристик компонентов.

Diagram

Description automatically generatedВ ходе серии консультаций в рамках подготовки данного ВКР была зафиксирована ниже следующая схема связей трех параметров из исходного датасета.

Такая схема кросс-зависимостей немного смущает, но она была подтверждена – с этим пониманием и была продолжена дальнейшая практическая работа.

Text

Description automatically generated

## Предобработка данных

Помимо ранее озвученных преобразований значений некоторых признаков (а именно, преобразовали «Угол нашивки» в категориальный тип данных и применили Box-Cox трансформацию для «Модуля упругости», «Поверхностной плотности» и «Плотности нашивки»), была проведена следующая предварительная работа с данными…

*(Далее представлена последовательность преобразований данных на примере датасета по одной из трёх целевых переменных – Модуль упругости при растяжении. По остальным целевым переменным и их моделям преобразования аналогичные).*

* + 1. Для начала исходный датасет разбиваем на две подвыборки – train и test, и убеждаемся в их гомогенности.

Graphical user interface

Description automatically generated

И временно убираем категориальный признак («Угол нашивки»), который не подлежит какой-либо дополнительной трансформации. Мы его потом вернём в обучающую выборку.

Text

Description automatically generated

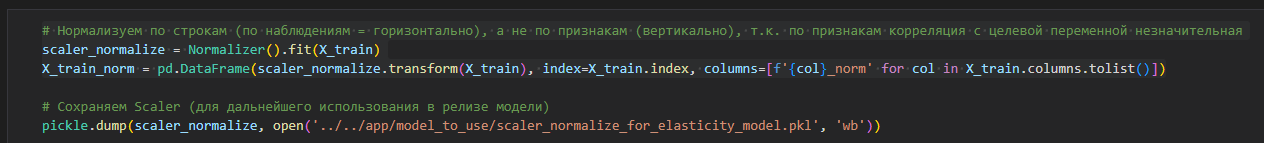
Далее проводим работу с обучающей выборкой …

* + 1. Приводим данные к единому масштабу через MinMaxScaler (с сохранением параметров преобразования для тестовой выборки).

A screenshot of a computer

Description automatically generated with medium confidence

* + 1. Создаём новые признаки: нормализуем данные по строкам (по наблюдениям = экспериментам = sample-wise), исходя из той логики, что строки – это разная рецептура (состав и технология), т.е. важна взаимосвязь признаков между собой по горизонтали.

Плюс помним об отсутствии связей (корреляции) между целевой переменной и отдельными признаками (т.е. отсутствие зависимостей по вертикали = feature-wise).

* + 1. Создаём новые признаки через попарное перемножение исходных признаков – таким образом пытаемся учесть возможные совокупные (суммарные) эффекты признаков при взаимодействии друг с другом.

Text

Description automatically generated

* + 1. Создаём новые признаки через метод главных компонент, предварительно стандартизировав исходные данные через StandardScaler.

Text

Description automatically generated

Собираем финальный датасет обучающей выборки с новыми добавленными признаками, не забывая вернуть обратно категориальный признак.

A picture containing text

Description automatically generated

И проделываем аналогичные трансформации с тестовой выборкой (X\_test), используя настройки методов трансформации (MinMaxScaler, Normalizer, StandardScaler и PСA), обученных на X\_train.

## Выбор наилучшего метода

*(Далее представлены комментарии и рассуждения на примере обучения модели по целевой переменной – «Модуль упругости при растяжении». По второй целевой переменной – «Прочность при растяжении», логика, комментарии и заключения аналогичные).*

Изначально к данному проекту я подходил, как к решению задачи регрессии, что по умолчанию подразумевала само задание ВКР.

Однако в ходе предварительной разведки с перебором ряда базовых популярных регрессионных методов (LinearRidge, BayesianRidge, RandomForestRegressor, XGBoostRegressor, LGBMRegressor, CatBoostRergressos) столкнулся с тем, что систематически получаю отрицательный R2 (коэффициент детерминации) по тестовой выборке.

Text

Description automatically generated

По всей видимости, причина в качестве и специфике самих данных, о чём было сказано выше в части про разведочный анализ данных.

A picture containing rectangle

Description automatically generatedНиже график со значениями показателя корреляции признаков с целевой переменной (модуль упругости при растяжении) из обучающей выборки.

Rectangle

Description automatically generated with medium confidenceНиже график со значениями показателя корреляции признаков с целевой переменной (модуль упругости при растяжении) из тестовой выборки – значения и расстановка признаков по коэффициенту корреляции сильно отличается от аналогичных значений обучающей выборки, что, как мне кажется, уже изначально предполагает несходимость результатов модели между обучающей и тестовой выборок.

Поэтому в поисках альтернативного решения по прогнозу целевой переменной – одного из конечных свойств композиционного материала (будь то показатель «модуль упругости при растяжение» или «прочность при растяжении»), я принял решение перейти от задачи регрессии к задаче классификации, для чего был проделан ряд преобразований целевой переменной, обозначенных далее.

Шаг 1 – Группируем значения целевой переменной (далее показано на примере «Модуль эластичности при растяжении») на интервалы.

Chart, histogram

Description automatically generated

Экспериментальным путём и экспертным решением определил оптимальное количество интервалов в размере 5-ти.

Шаг 2 – Переводим интервалы в лейблы (наименования категории) нашей новой целевой переменной и сохраняем средние значения целевой переменной по каждому интервалу для дальнейшего предсказания абсолютного значения признака по лейблу.

Text

Description automatically generated

Text

Description automatically generated

Шаг 3 – Исправляем дисбаланс классов через oversampling.

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Далее делим датасет с новой целевой переменной, которая на этот раз является категориальной из 5-ти классов, на обучающую и тестовую выборки и производим все выше обозначенные преобразования объясняющих переменных.

В ходе решения задачи классификации:

* В качестве метрики ориентировался на accuracy
  + Альтернативный вариант F1-score (взвешенный)
* Был задействован ряд методов для сравнения их результатов и выбора наилучшего из них
  + При этом использовались базовые настройки методов для осознанного упрощения объёма работ по ВКР ввиду ограниченности времени.
  + В реальной жизни нужно и важно подбирать гиперпараметры для каждого метода и я бы это делал через GridSearch и StratifiedKFold, например так, как в представленном ниже коде

Text

Description automatically generated

Ниже представлены результаты ряда использованных методов.

* + 1. XGBoostClassifier

Calendar

Description automatically generatedText

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* + 1. GaussNB

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

* + 1. AdaBoostCalassifier

Calendar

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* + 1. KNeighborsClassifier

Calendar

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* + 1. SVC

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

* + 1. MLPClassifier

Text

Description automatically generated

Calendar

Description automatically generated with low confidence

* + 1. GradientBoostingClassifier

A picture containing text

Description automatically generatedCalendar

Description automatically generatedText

Description automatically generated

* + 1. LogisticRegression

A screenshot of a computer

Description automatically generated with medium confidenceCalendar

Description automatically generated

Методом, показавшим наилучший результат accuracy по тестовой выборке, является SVC, который объявляем «победителем» и выбираем для дальнейшей работы.

В качестве дополнительного шага производим поиск наилучших гиперпараметров через метод GridSearch и StartifiedKFold кросс-валидацию.

Text

Description automatically generated

Аналогичным образом решаем задачу по выбору наилучшего классификатора по второй целевой переменной – «Прочность при растяжении».

Все скрипты выложены в GitHub.

# Нейронная сеть по оценке показателя «Соотношение матрица-наполнитель»

Тип задачи, для которой сбираем нейронную сеть – регрессия.

Исходим из следующих базовых параметров сети:

* полносвязанная
* на входном слое количество нейронов = количеству признаков (объясняющих переменных, у нас их получилось 87)
* не менее 2-х скрытых слоёв
* активационная функция внутренних слоёв ReLu, выходного слоя - Linear
* добавляем минимум один слой Dropout для борьбы с переобучением с учётом специфики данных (о чём говорили выше)
* оптимизатор – SGD
* функция потерь – mean\_squared\_error

В ходе ряда экспериментов, как с количеством слоёв, функциями потерь (ReLu vs.Tanh) и числом эпох останавливаемся на следующей конфигурации сети.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated

y\_train против y\_train\_predict

Chart, scatter chart

Description automatically generated

y\_test против y\_test\_predict

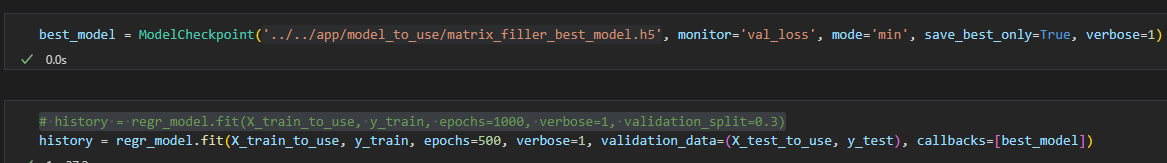
Chart, scatter chart

Description automatically generated

В целом, качество модели низкое – плохо описывает дисперсию целевой переменной и в основном выдаёт её среднее значение с небольшим стандартным отклонением.

Помимо того, что нужно ещё поэкспериментировать с архитектурой и настройками нейронной сети, основная причина, видимо, в специфике самих данных.

В ходе обучения нейронной сети задаём Callback на сохранение лучшего варианта модели для дальнейшего использования в приложении.



# Создание приложения

Создано приложение по прогнозу всех трёх параметров:

* Модуль эластичности при растяжении
* Прочность при растяжении
* Соотношение матрица-наполнитель

Часть кода приложения

Text

Description automatically generated

Работа с приложением…

Шаг 1 - Ввод исходных параметров

Graphical user interface, text, application

Description automatically generated

Шаг 2 – Результат прогноза целевых характеристик композиционного материала

Graphical user interface, text, application

Description automatically generated

# Создание репозитория на GitHub

https://github.com/VitalyKoz/baumanka\_data\_science\_vkr\_composite

Graphical user interface, text, application, email

Description automatically generated

# Заключение

1. Все основные этапы ВКР, обозначенные в п.1 (Постановка задачи), выполнены.
2. Качество представленных моделей можно и нужно улучшать, но на это нужно больше времени. Тоже касается и приложения.
3. Конечная цель данной ВКР не столько в достижении безупречной точности моделей, сколько в широте и полноте охвата всех заявленных задач и этапов.
4. Представленный датасет с исходными данными представляется крайне специфичным и даже сложным, с т.з. получения красивых результатов, поэтому осмелюсь предложить в будущем заменить данную тему по композиционным материалам, предлагаемую всем слушателям по умолчанию для ВКР, на какую-нибудь иную с более «отзывчивыми» («дружелюбными») данными.

# Список литературы и веб-ресурсов

1. Stella Hrehova and Lucia Knapcikova. Все основные этапы ВКР, обозначенные в п.1 (Постановка задачи), выполнены. The Study of Machine Learning Assisted the Design of Selected Composites Properties
2. Vade Shah, Steven Zadourian, Charles Yang, Zilan Zhang, Grace X. Gu. Data-driven approach for the prediction of mechanical properties of carbon fiber reinforced composites
3. Документация по библиотеке scikit-learn: – Режим доступа: <https://scikit-learn.org/stable/user_guide.html>
4. <https://machinelearningmastery.com/>
5. <https://github.com/>
6. <https://keras.io/>