МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования   
«Пермский государственный национальный исследовательский университет»

Физико-математический институт

Отчёт по дисциплине

«Технологии машинного обучения и нейронные сети для решения прикладных задач»

|  |  |
| --- | --- |
| Студент 4 курса | Вячин О.С. |

Пермь 2024 год

Целью работы было протестировать на своем наборе данных методы регрессии. Также важным условием было, чтобы коэффициент детерминации был минимум 0,85.

В качестве набора данных использовался датасет с платами, на которых в определенных местах могут располагаться серые зоны. Соответственно, датасет подходит для задач бинарной классификации.

Первым делом были импортированы необходимые библиотеки для работы с датасетом и последующим анализом. Далее мы загружаем датафрейм из файла dataframe.csv и печатаем первые пять записей из датасета. Следующим шагом было составить корреляционную матрицу и построить на ее основе тепловую карту, чтобы узнать корреляцию между признаками.

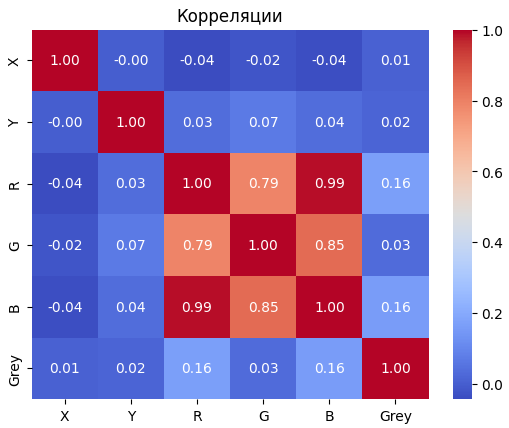


Рисунок 1. Тепловая карта матрицы корреляции

Следующим шагом выполняется нормализация и стандартизация, а также выделение целевой переменной Grey и оставшихся признаков.

Затем разделяем датасет на обучающую и тестовые выборки с отношением 8:2, где 80% - обучающая выборка, 20% - тестовая выборка.

В качестве модели выбираем линейную регрессию, метрикой будет выступать r2\_score для вычисления коэффициента детерминации. Результатом метрики составило 0.07, что очень мало и говорит о том, что линейная регрессия не способна точно предсказывать значения.

Следующей моделью был выбран метод случайного леса – RandomForestRegressor. После обучения модели была выполнена попытка предсказания и также измерили результат работы модели метрикой r2\_score. Результат составил 0.999, что очень много и говорит о переобучении модели.

Далее рассмотрим модель градиентного бустинга. После обучения модели и замера метрик выясняется, что результат составляет 0.98, что уже менее походит на переобученную модель, но при этом точность предсказания модели все также высока.

Теперь рассчитаем среднеквадратичную ошибку и среднюю абсолютную ошибку для градиентного бустинга. Показатели составили:

* MSE – 0.00017
* MAE – 0.00155
* R2 – 0.98

Чем меньше параметры MSE и MAE, тем лучше. В моем случае параметры низкие, что говорит о высокой точности модели.

Последним шагом будет визуализация модели регрессии с предсказанными и истинными значениями.

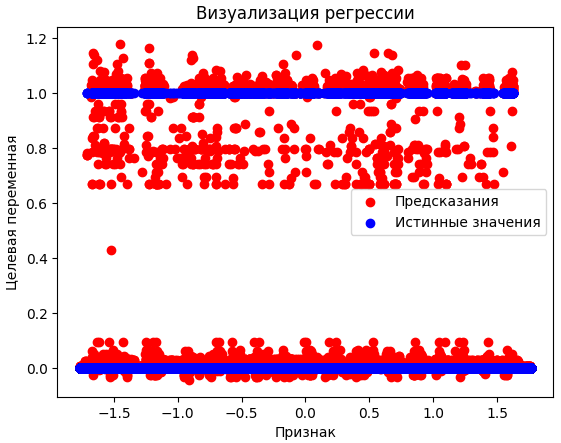


Рисунок 2. Предсказанные и истинные значения модели регрессии