Name : Nishan Maharjan

Id : 302763006

course title : CSC 180-01 Intelligent Systems

assignment id : Project_1

due date: 10:30 am, Wednesday, September 25, 2024

Problem Statement

Predicting Business Ratings Based on Customer Reviews and Other Business Attributes.

Given a dataset of businesses with attributes such as customer reviews, location (latitude and longitude), and review count, can we develop a predictive model that accurately estimates the star rating of a business? The model should learn from the provided features and be able to predict star rating of that business.

Methodology

1. Data Gathering and Preprocessing

- Converted Json data to pandas data frame extracting 1000000 rows. Build two data frames called **review** and **business** Removed businesses with fewer than 20 reviews from the review table.
- Grouped reviews by `business_id`.
- Merged this Data Frame with the business table using `business_id`, including relevant features like latitude, longitude, review count, stars, and name.

2. Feature Extraction and Transformation

- Applied TF-IDF vectorization to extract features from review texts.
- Converted the Data Frame into a NumPy array representation.
- Converted additional features (latitude, longitude, review count) to a NumPy array and concatenated it with the TF-IDF features.

3. Data Splitting

- Defined features (X) and target variable (y, stars) and split the dataset into training and testing sets.

4. Model Building

- Developed various models using different activation functions (Relu, Sigmoid, Tanh) and optimizers (Adam, SGD).

5. Evaluation and Visualization

- Plotted graphs to visualize model performance.
- Created a table displaying business names, actual ratings, and predicted ratings.

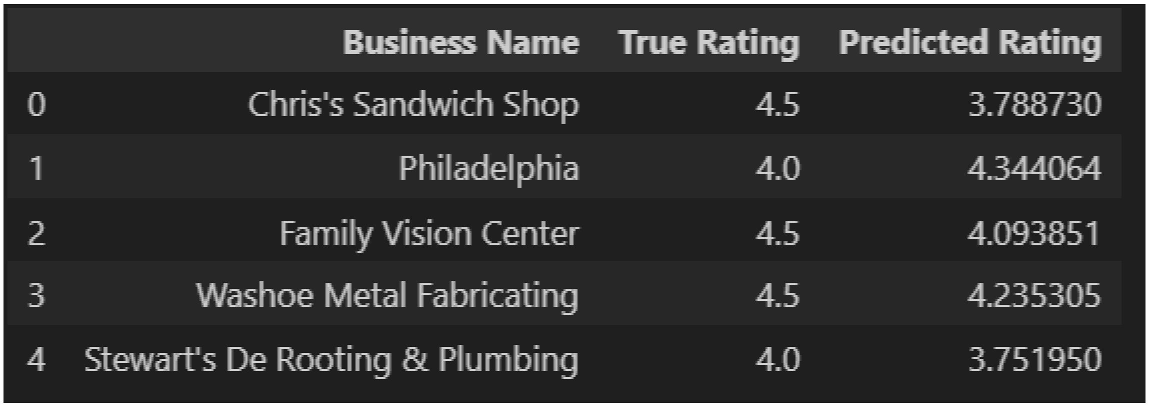
Experimental Results and Analysis

Model 1

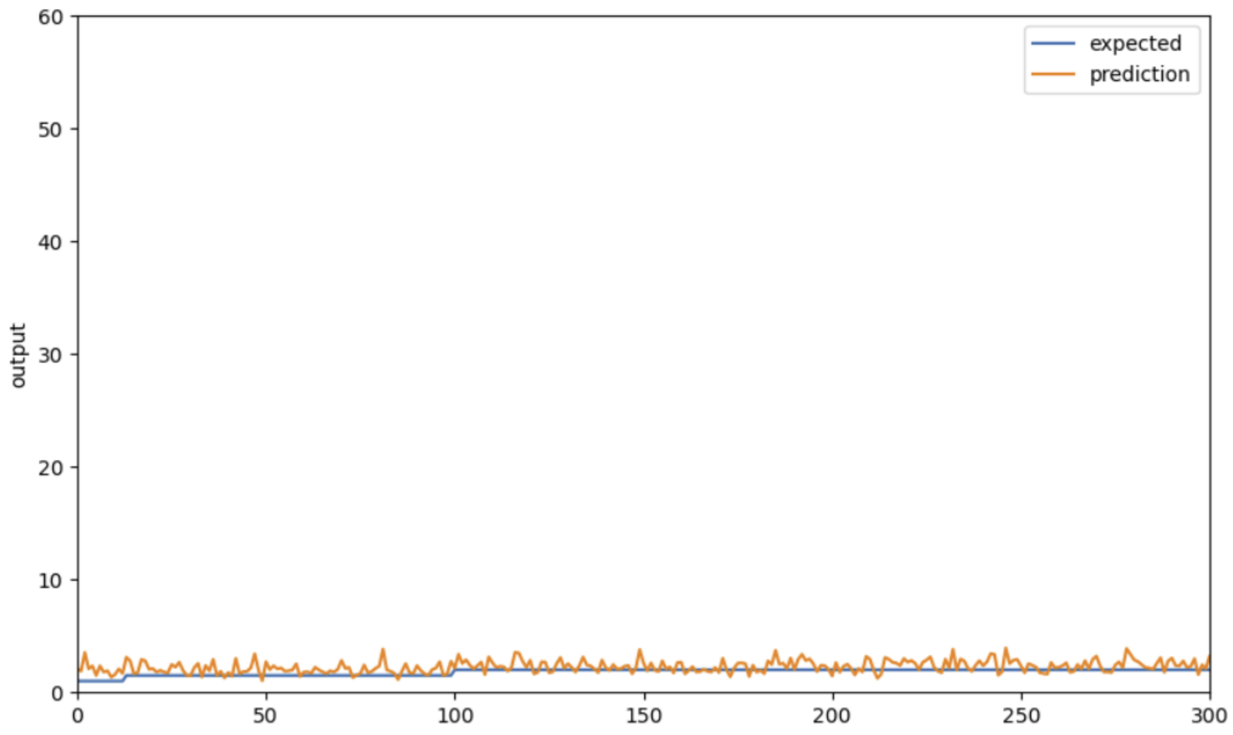
Model Configuration: The model was built using a sequential architecture with four hidden layers, utilizing the Relu activation function and the Adam optimizer.

Performance: The model achieved an RMSE of approximately 0.4543, indicating that, on average, predicted ratings deviate by about 0.4543 stars from actual ratings.

Evaluation: This RMSE value suggests reasonable predictive accuracy, but there is potential for improvement.



	Business Name	True Rating	Predicted Rating
0	Chris's Sandwich Shop	4.5	3.788730
1	Philadelphia	4.0	4.344064
2	Family Vision Center	4.5	4.093851
3	Washoe Metal Fabricating	4.5	4.235305
4	Stewart's De Rooting & Plumbing	4.0	3.751950

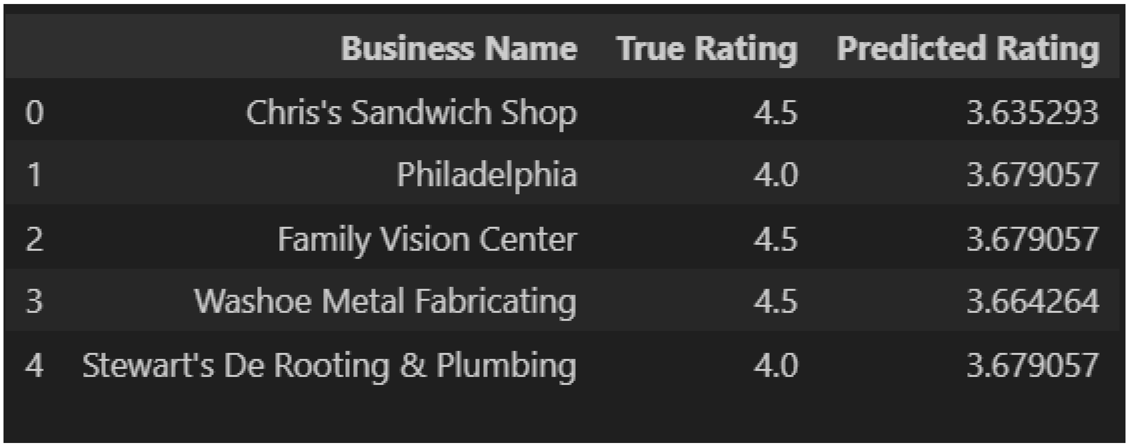


Model 2

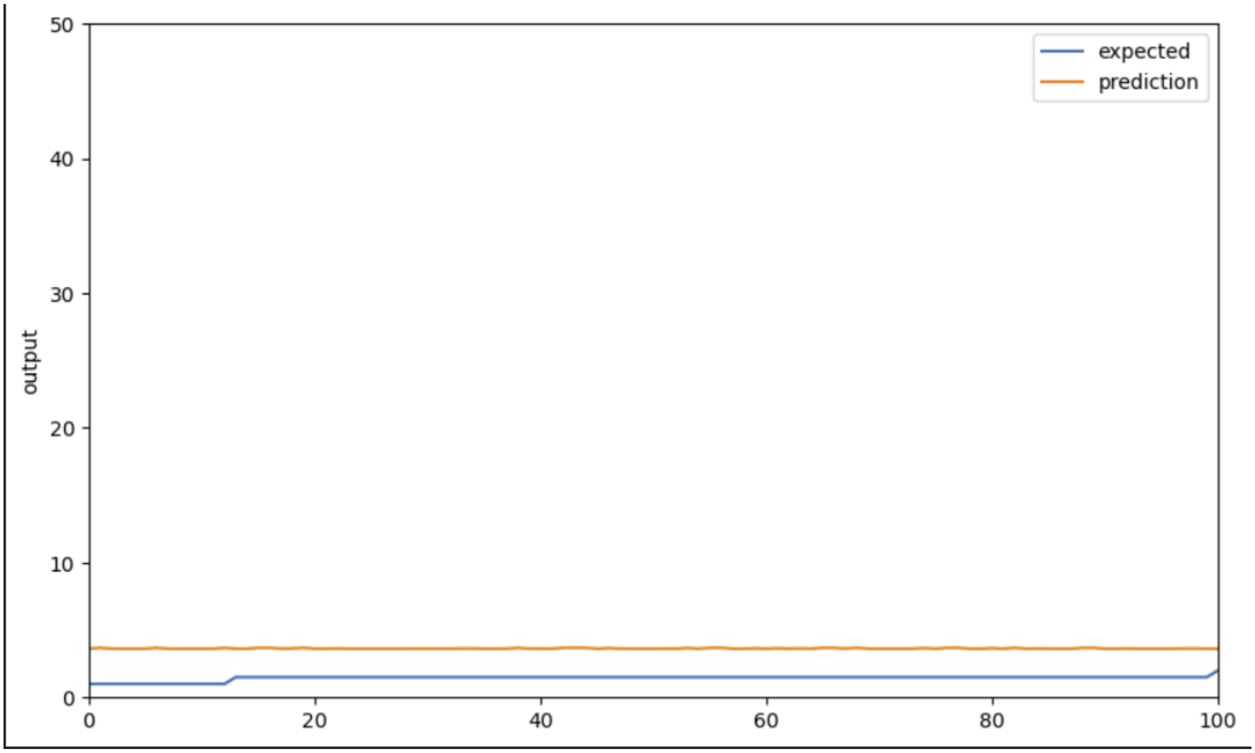
Model Configuration: The model was constructed with four hidden layers using the sigmoid activation function and the SGD optimizer.

Performance: The model achieved an RMSE of approximately 0.8218, indicating a larger average deviation from actual ratings compared to the previous model.

Evaluation: This RMSE suggests lower predictive accuracy, highlighting that the sigmoid activation may not be optimal for this problem.



	Business Name	True Rating	Predicted Rating
0	Chris's Sandwich Shop	4.5	3.635293
1	Philadelphia	4.0	3.679057
2	Family Vision Center	4.5	3.679057
3	Washoe Metal Fabricating	4.5	3.664264
4	Stewart's De Rooting & Plumbing	4.0	3.679057

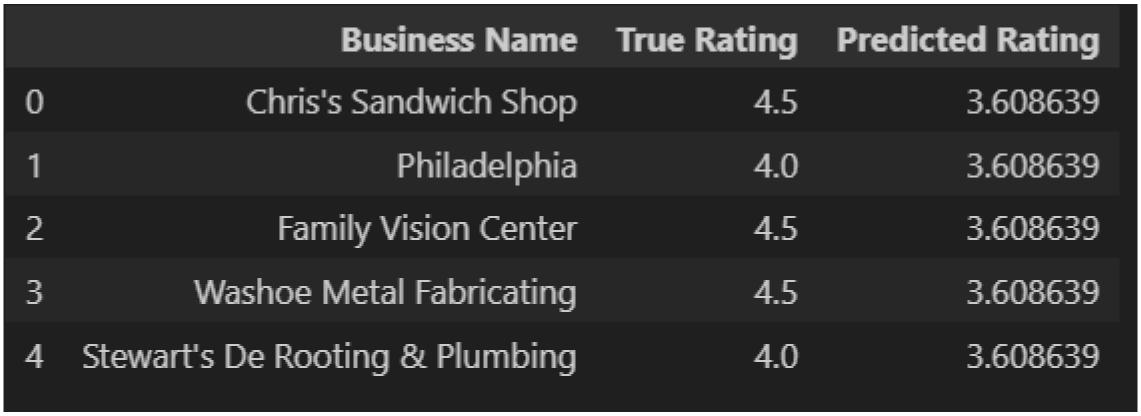


Model 3

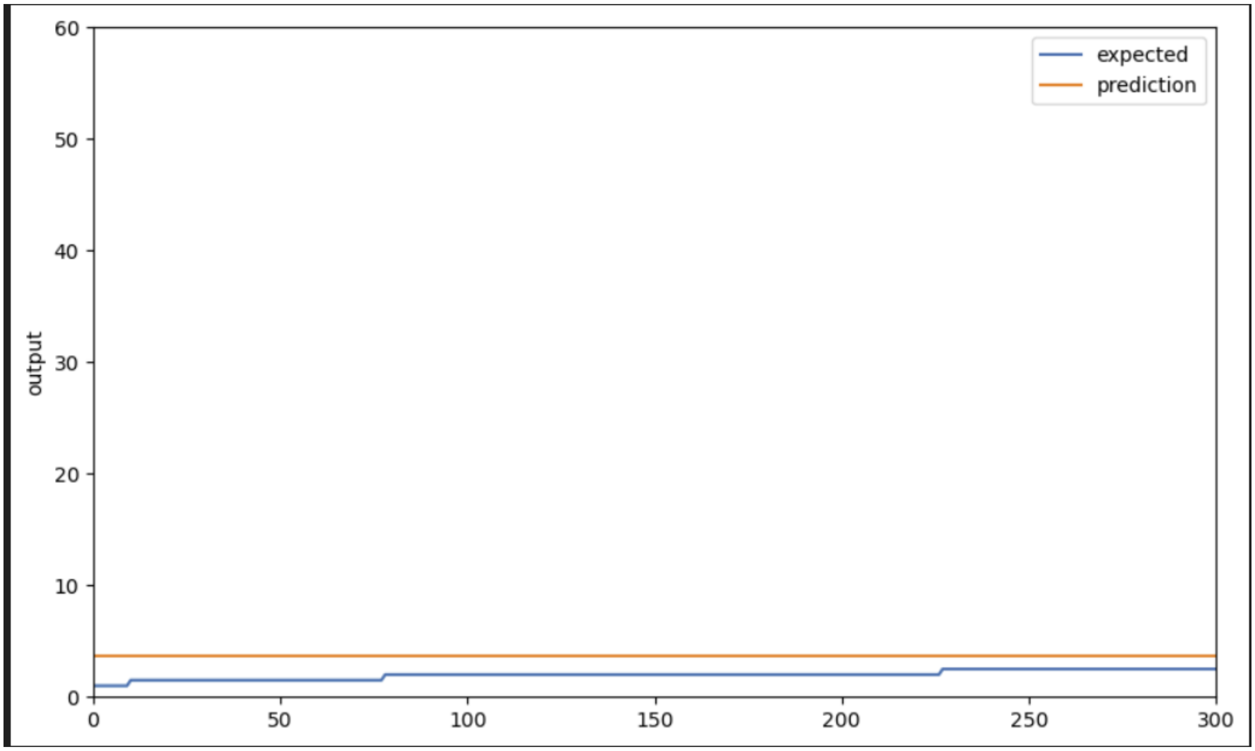
Model Configuration: The model consisted of three hidden layers, utilizing the tanh activation function and the Adam optimizer.

Performance: The model achieved an RMSE of approximately 0.8323, indicating a further increase in average deviation from actual ratings compared to previous models.

Evaluation: This RMSE suggests that the tanh activation may not be the best choice for this dataset, leading to less accurate predictions.



	Business Name	True Rating	Predicted Rating
0	Chris's Sandwich Shop	4.5	3.608639
1	Philadelphia	4.0	3.608639
2	Family Vision Center	4.5	3.608639
3	Washoe Metal Fabricating	4.5	3.608639
4	Stewart's De Rooting & Plumbing	4.0	3.608639



Task Division and Project Reflection

Task Division:

- Since the project was completed individually, I was responsible for all aspects, including data collection and preprocessing, feature engineering, model development, and evaluation.

Challenges Encountered:

- **Chart Creation**: One significant challenge was creating a proper chart for data visualization at the end.

Learning Outcomes:

- **Concatenating Arrays and Data Frames**: Gained practical experience in combining data structures for model input.
- **Experimenting with Activations and Optimizers**: Learned the effects of different activation functions and optimizers on model performance.
- **Creating Charts**: Developed skills in visualizing data effectively to communicate findings.
- **Understanding Neuron Impact**: Gained insights into how the number of neurons in a model architecture influences predictive capability.