Written By: Acyuta Raman

Restaurant Ratings Analysis

Introduction

The goal of this project was to develop a linear regression model to predict the overall rating (out of 5 stars) for a business. The datasets were originally provided by Yelp, and they were directly received from Codecademy. There were various datasets for businesses, users, reviews, check-ins, tips, and photos. All analysis and visualizations were conducted via python. The scikit-learn library was used to conduct linear regression on the data. This analysis is purely for academic purposes.

Body

A correlation matrix was used to check both the collinearity of our features and their individual correlations with our target variable. The only independent variable that strongly correlated with business ratings was the average review sentiment. This number was a score that represented the average positivity or negativity of each review. Average review sentiment had a correlation coefficient of approximately 0.782 with business rating. There is a strong, positive, linear correlation between average review sentiment and business ratings for this dataset.

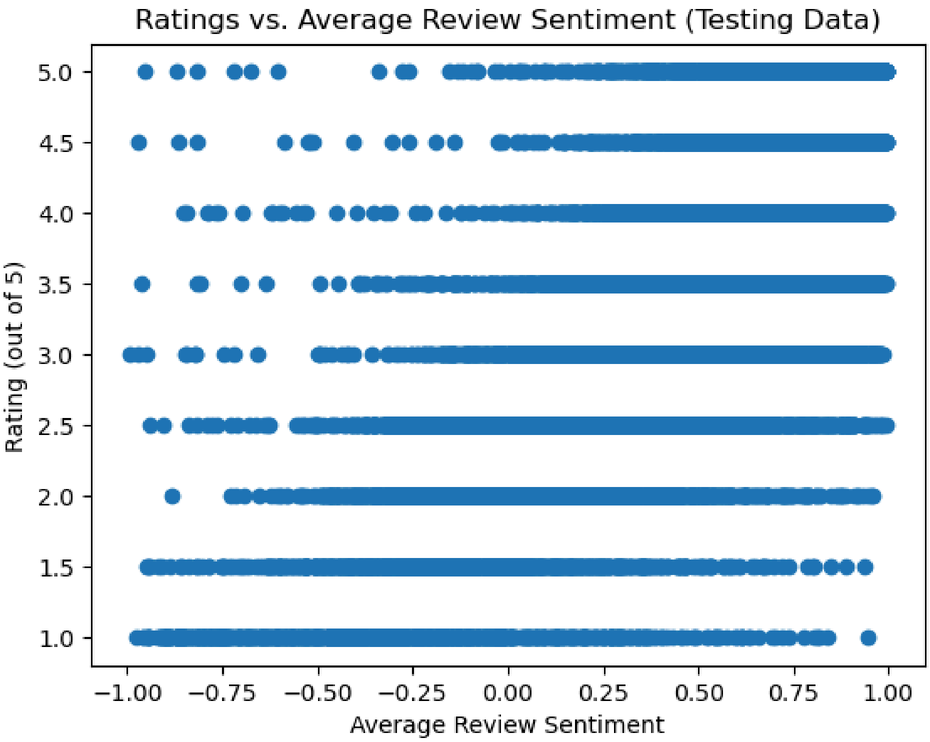

Figure 1: Rating vs. Average Review Sentiment

As the Average Review Sentiment increases, the business ratings become increasingly clustered towards the right. A line-like group of points indicates that many points are clustered in that region.

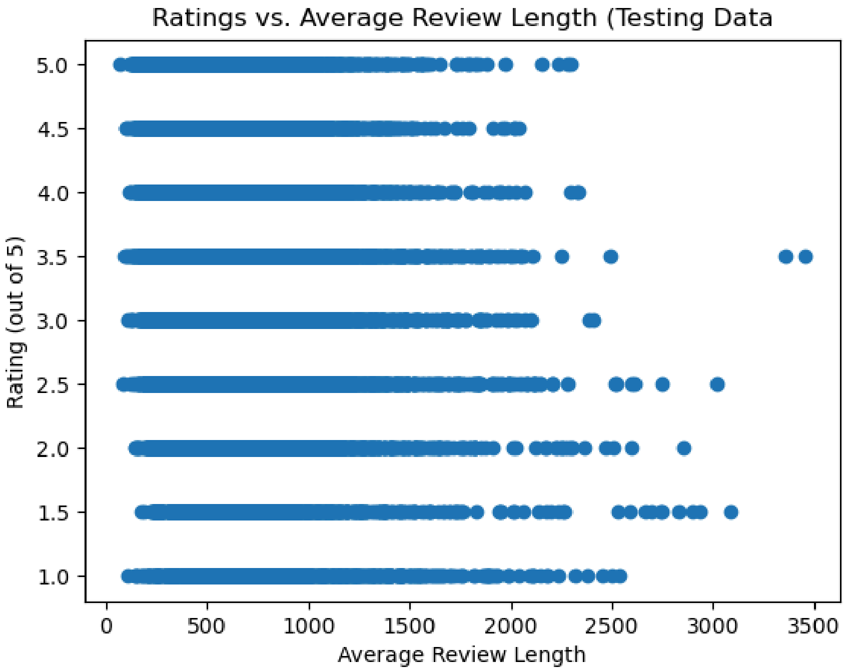

Figure 2: Rating vs. Average Review Length

The feature with the second highest correlation coefficient, with respect to business ratings, is the average length of a review. As rating increases, the line-like clusters of points shift leftward. This indicates that the majority points with higher ratings have lower average review lengths. Additionally, the correlation coefficient of average review length is approximately -0.277. Overall, there is a weak, negative, linear correlation between average review length and business ratings for this dataset.

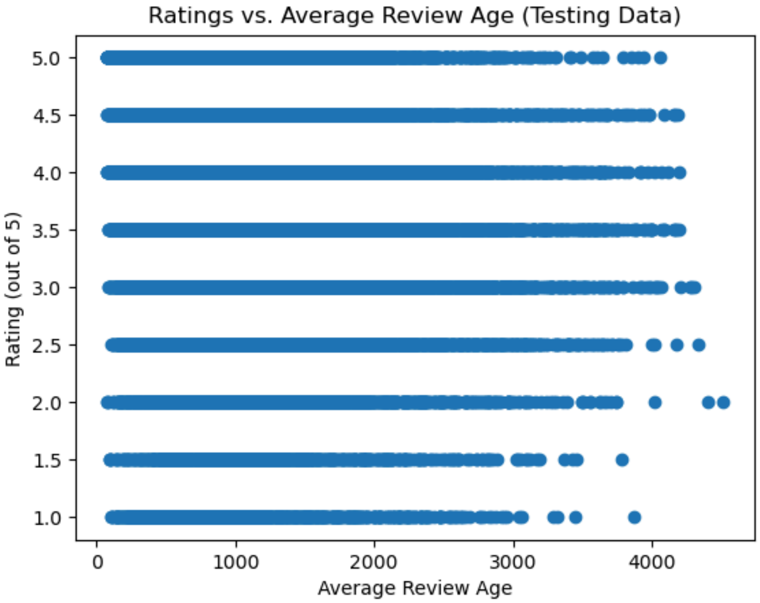

Figure 3: Rating vs Average Review Age

The average review age of a business has a correlation coefficient of approximately -0.126 with rating. Although there is a somewhat similar trend with respect to the graph for Rating vs. Average Review Length, the signs of decrease in rating are less apparent. This is likely due to the weak correlation coefficient. Overall, there is a weak, negative, linear correlation between average review age and business ratings for this dataset.

The data was split into portions for training and testing. The linear regression equations is as follows:

Ratings = 2.24294742*Average_Review_Sentiment -5.94532770*10^{-4}*Average_Review_Length -1.51216295*10^{-4}*Average_Review_Age

The model's correlation coefficients for the training data and testing data are approximately 0.6513 and 0.6526 respectively. When all variables are used as inputs, the model's correlation coefficients for training and testing data are approximately 0.6800 and 0.6809 respectively. This shows a moderate, positive, linear correlation with business ratings.This is due to the weak correlation coefficients for all the other inputs. Due to the two absolute values for the coefficients of average review length and average review age in the regression equation, average review sentiment has the most significance when predicting business ratings. This, in combination with its correlation coefficient of approximately 0.789, signifies that average review sentiment as the strongest predictor of business ratings in this dataset.

Conclusion

The linear regression model for ratings prediction has a moderately strong, positive, correlation with respect to actual business ratings. Although it included a few independent variables, its correlation strength was very close to that of a linear regression model using all possible variables. The strongest predictor of business ratings was average review sentiment, both in correlation coefficient and the model's formula. Average review length and average review age were both relatively weak in terms of correlation, with the absolute values of their correlation coefficients being less than 0.3.

Appendix

Linear Regression: The process of approximating a relationship between multiple variables with a line. This "line of best fit" is calculated by minimizing the total squared difference of the actual values and predicted values of the dependent variable.

Correlation Coefficient: a measure of the strength of the linear relationship between two variables. It ranges from -1 to 1.