Cesar Nunez Rodriguez and Raymond Ogunjimi
Dr. Jiao Xun
Applied Machine Learning
7 March 2022

Mini Project Report

Public link to our Google Colab Workspace:
https://colab.research.google.com/drive/1CSt3y3byClvw3ktZbeNoEHtLYCV-a6f2?usp=sharing

Our model is designed to predict the beer rating of a certain beer. We found through a trial and error method that for one of the most accurate possible predictions our model should use the following features as inputs: the beer's ABV, age of the user, time, and how the user rated the palate, taste, and aroma of the beer. These features are a part of the provided .JSON file and were parsed from the HTTP URL using the given interpretation algorithm. Since all of the features used were quantitative, our model is based on linear regression. To feed data into the model, the train_test_split( ) function is utilized to split the data into training (80% of the data) and testing (remaining 20% of the data) sets.

In terms of the process used to create the linear regression algorithm, the numpy.linalg.lstsq( ) function is used to determine the thetas of our equation. Afterwards, the model makes predictions on the remaining inputs and uses MAPE to compare with the actual rating given by the users. In the training phase, the theta coefficients/weights are taken along with the feature variables to produce an equation tha is used for the prediction in the testing phase. When testing the model, appropriate feature values are substituted into the equation to produce a prediction output. This output rating is then evaluated against the true output as given by the .JSON file provided.

In order to evaluate, the Mean Absolute Percent Error (MAPE) method is used. The higher the MAPE the less accurate the model is. The best error as calculated by the MAPE function was about 9% which can be considered a good accuracy and fit for the data using 6 features. The accuracies can change based on the randomized split of the data for training and testing. Although the group was able to achieve good accuracy, the model could definitely be improved by using features that are not quantitative such as the beer brand and flavor, and integrating the pattern of those features with the linear regression algorithm. However, if a few quantitative features are able to predict the ratings and only miss by about 9% of the rating on average, this model can be safely relied on for rating prediction.