Travis De Prima

Machine Learning I

Professor Lam

13 December 2020

Loading the Data:

At the beginning of my load_json.py file, my program lists all the variables the computer will be reading from the dataset given to us in class. Then my program opens that dataset and reads the variables listed at the beginning. I only really accounted for a few things in my model that I will get to later, but first there are some notable things to mention. In order for my program to make an assessment of the accuracy scores, I had to denote a dummy variable to a select amount of states I decided to extract from the dataset. These states are WI, NY, SC, and PA. These dummy variables were assigned through an if-else statement with if cell_state == "WI" the program assigns a 1, while else is assigned a 0. I then did the same process for "By appointment only", "Restaurant price range" and "Restaurants", which ended up being a little difficult at first, but ultimately figured it out. For the most part I kept the rest of the load_yelp_json.py file the way it was and was a great template to modify and find inspiration from. My load_json.py file then exports the append data to a file called "business_data.csv".

The Model

To combat my computer from sounding like a jet engine and a bunch of warning messages I was getting I decided to cut my dataset into a sub-dataset of 1000. I then set my target value to 2 (stars) and my data to 5,6,7,8,12,16, which corresponds to WI, NY, SC, PA, restaurant price, and restaurant. I ended up excluding sunday_closing_time as I was running into issues with getting the time variables to convert to a dummy variable and restaurant_price, which left me with a few cells still reading the word “None” and not converting it to a dummy variable. I then used a variation of the RandomForest model we used in class and used a for loop to get my accuracy scores. These accuracy scores then were converted to a csv file called forest.csv. Then, to get my comparison model, I simply used the logistic model file we were taught in class. I originally wanted to use a linear regression, but I ran into issues with my dataset size and the program simply did not want to read the original dataset correctly. After thinking through this speed bump, I realized that it would make much more sense to use a logistics model as it’s much more receptive to the type of data I was working with. The program finally creates a separate csv file that presents you with the accuracy scores.

The Results:

After running my program, I got the final results: RandomForest: [0.408, 0.356, 0.372, 0.356] and Logistic: [0.396, 0.34, 0.348, 0.352]. After analyzing my data, I found that both models were fairly close to each other without much separation. Although they were both fairly similar, I found that my RandomForest model was slightly more accurate than my logistics model. This could be because RandomForest models don’t assume that the variables in the model have a linear relationship, contrary to a logistic model.