EECS 731 – Project 4 Report

10/5/2020

Brian Clark

2928899

For the fourth project of the semester, I initially began working with the MLB dataset that was provided. After removing 10-12 columns, I realized I still had a lot of data to manipulate and run our regression models on. Due to the time constraints of the assignment, I decided it would be best to work with the NFL dataset as it involves significantly less data as compared to the MLB dataset. The datasets were provided by Dr. Kuehnhausen and can be found on GitHub at <https://github.com/fivethirtyeight/data/tree/master/nfl-elo-game>

I began the project by reading the datasets directly from GitHub into my program using Pandas and specifically read\_csv. I removed columns that would not be used for training, testing, and performing the regression models on the provided data. In order to make the data more interpretable, I had to convert the date column into separate columns, and convert the initials indicating the teams into a numerical value. After doing so I was able to run Random Forest regression and Linear regression on the given dataset.

I believe the value in this data would be to help a sport’s enthusiast more accurately predict the outcome of given NFL football games. While the accuracy scores were at or under 40%, it still provides some guidance and I feel would be a helpful tool for forecasting the winner of each game. If more data were used it might be possible to attain a higher accuracy score and be able to predict the winner in a more consistent manner. I was very surprised to see that Random Forest regression actually had a higher accuracy score than Linear regression. Based on my prior experience with the two regression models, I had nearly always found that Linear regression was more accurate.