Daniel Angel

June 5, 2021

DSC 530: EDA

Final Term Paper

Dr. Shankar Parajulee

Summary of Exploratory Data Analysis performed on NBA Game Log Data Set

Originally, I wanted a question which came to me- the classic: “Defense wins championships,” or phrased as a question, “Does defense win championships?” Due to interest I chose NBA, specifically seasons between 1985 and 2019. I was curious what effect defense has on outcomes of professional basketball games. Since I was unable to easily access/use data for championships, I reformulated my hypothesis as **“Does defense win games?”** or put statistically, **“Do higher amounts of defensive statistics increase likelihood of winning?”**

I looked at or created variables, through transformation, which would have the biggest impact. These variables included existing ones : rebounds[‘DREB’], steals[‘STL’], blocks[‘BLK’]. Other variables required transformation, or moving it from one place to another, such as: opponents’ field goal percentage[‘OPP\_FG\_PCT’], opponents’ 3-point field goal percentage[‘OPP\_3FG\_PCT’], and opponents’ points[‘OPP\_PT’]. And one variable, rebounding rate[‘REB\_RATE’], didn’t exist but I thought it could be insightful, so I had to create it using a clever bit of calculation.

Prior to performing my analysis, the data set had to be created by merging several other data sets, 34 to be exact, which I found in a GitHub repository by the Los Angeles Times Data Desk and cleaned. The cleaning step was when I got rid of bogus, missing, empty, and null values. This was also when I did my variable creation and deletion. I removed nine variables and added four new ones. I also normalized the stats for entries(rows) where the games went into overtime past the standard 240 player minutes per game by multiplying by a weighting factor of 240/”Game Minutes”, which was one of the more difficult steps of the entire project.

After performing the various tasks of an exploratory data analysis such as looking at visualizations - histograms, probability mass function plots, cumulative density function graphs, scatter plots, and analytical distributions – analyzing and comparing various quantifiers – summary statistics, covariance, correlation, and spread - performing hypothesis testing and regression – linear fit, logistic regression, multiple regression, and permutation testing – and weighing these results against various measures of goodness of fit – accuracy, R^2, and P-Value – several interesting results revealed themselves. **At the end I had to conclude that Defense does win games.**

Rebounds had a greater effect than blocks or steals. The opponents’ points and field goal percentage had the greatest effect of all. Surprisingly, 3 point field goal percentage had a small effect and was mostly random noise, definitely correlation rather than causation.

I feel like I missed confidence intervals and plotting the percentile residuals. Also, the analysis could have been enriched by more GroupBy and slicing or figuring out how to incorporate data from playoffs or championships. Furthermore, it would have been fascinating to do a time series analysis or compare data by year, over time, or by team. It is possible that other variables could have helped like personal fouls or free throws but it was difficult for me to comprehend how. I assumed that linear models were an adequate fit for modeling my data without trying any other non-linear models like lognormal, exponential, etc.

Covariance and correlation calculations and linear fit were not working for variables which had float values. A possible solution could have been using the parameters from the logistic model to create the linear fit. I didn’t understand why my hypothesis tests looked the way they did, whether they were supposed to or if I made a mistake. Lastly, when performing cleaning and transformation functions across the entire data frame I used “for” loops when I am sure that was incorrect and there is a better, faster, more efficient way of accomplishing that.