Thomas Courtney

Machine Learning, I

Professor Wilck

1/16/2021

Through the Moneyball data exploration, I found myself researching baseball. Baseball today differs greatly from 1871 to 2006. In 1871, there were different rules regarding balls and strikes, there was no batter's box, and no stealing. This is much different then the game we see today. Even further, the game we see today has better athletes, equipment and different stadiums. Additionally, I researched statistics that historically had the biggest impact on winning.

I altered the data in a few distinct ways in order to remove outliers or NA inputs. I determined which to use by seeing if there was another factor like the one, I was solving for. Other columns did not have such similar data points. The number of times a team was hit by a pitch does not have another column to factor against. I used the median here. There were many missing variables in this column described, so I would not recommend using this column in a predictive model. For outliers or zeros, I used either IQR's or percentiles. I did some online research and based off actual baseball records determined the stopping point. For example, the fewest home runs hit by a team in a season was three. Since there was nothing available for the fewest number of HR's given up, I used the fewest number of homes runs a team has hit as a lower bound. If any team had 3 or less homeruns on a season, I increased them to the tenth percentile. I continued using this strategy based off baseball records and intermixed between using IQR and percentiles depending on the number of outliers. The greater number of outliers, I would use IQR.

When building my model, I looked up a lot of important baseball statistics that were not related to our model. These include slugging percentage, on base percentage and batting average. These greatly influenced my model. One of the largest factors was estimated runs scored, runs given up and final score. This is calculated as for every 1.87 hits, a run is scored in baseball. This can determine the number of runs scored and given up by the team. These factors obviously have the greatest affect on teams actually winning.

I used a system of trial and error to multiply different factors together. This helped determine which factors influence each other in the model.