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. Although the data was altered to fit a 162-game season, the above changes played a massive role in the statistics of that era. Even past that, the dead ball and live ball era's, additions of designated hitters and steroids all played factors as time went on.

I altered the data in a few distinct ways in order to remove outliers or NA inputs. When the original dataset had NA's, I either used the median of the rest of the column or a ratio. I determined which to use by seeing if there was another factor like the one, I was solving for. For example, there would be a ratio for how many times a team successfully stole a base and when they were caught stealing. These ratios multiplied by the factor present in the original data set would create a good estimate for the other factor. Other columns did not have such similar data points. The number of times a team was hit by a pich 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.

At the end of the day, I would not recommend using the entire baseball dataset when building a model. Since the game has change dramatically over the years, baseball statistics inherit many misconceptions. I would use team data in the past five years and remove outliers from there when building any sort of predictive model.