Brady Price

The main question that I am trying to answer is if I can predict a player’s position based on how he does in the season that he plays. I am looking at MLB baseball data and I am looking to see how they do in categories such as homeruns, second-base hits, batting average, and so forth to predict what position they play on defense.

I am getting my data from MLB’s website and I am pulling data from the last 30 years and it takes about the top 150 players from every year and it puts them all into one data frame. The target of the data frame is the position. However, right now there is not a position column. That is because for the name of the player it will have their first and last name, and then it will have their position right after. For example, the first person in the data frame shows up Juan J SotoSotoLF. So, Juan Soto plays in Left field. I will need to make sure to do some cleaning so that we can split this apart and so that we don’t get repeating last names.

The features of the data frame are team, games played in that season, at-bats in that season, runs in that season, hits in that season, second-base hits in that season, third-base hits in that season, homeruns in that season, runs batted-in in that season, walks in that season, strike-outs in that season, stolen-bases in that season, hitting average in that season, on-base percentage in that season, slugging percentage in that season, OPS in that season, and year of the season.

I don’t have any plans to add any features right now. However, as I get more into the data, I may see a need to add additional features based on the features that I have now. I have nearly 4,000 observations in my dataset. The data will be able to help answer my question because it has a lot of the information and data on how each player did in the season as well as their position. It should give me the opportunity to try machine learning models out to see if it can predict the player’s position.