Aaron Drexler
11/20/2021
DSC 550
Werner

Conclusion

For my final project, I decided to look at baseball. Baseball has been a passion of mine for a long time and one that drives much of my excitement and joy in data science. The problem that I was looking to solve is what is the salary efficiency point for both batters and pitchers. I want to see where is the line that indicates a “good” contract compared to a “bad” contract.

To do this, I wanted to create a ratio between an advanced statistic and that of the players salary. For batters, I used weighted on base average, which is a good neutral statistic that factors out bias such as slugging and walks to indicate a true reflection on a batter’s individual game impact. To achieve a similar independent statistical impact, I used fielding independent pitching. This factors out how well the fielders surrounding the pitcher perform. I took key statistics provided in the data using feature extraction, then used the unique formula to find these advanced saber metric datapoint.

There is one challenge that I did have with the data. The data I had access to only had regular statistics in the database, no advanced statistics. Advanced statistics do a better job of calculating a players value, which would have been much better for this project. If there was a way to have said data, I would have used WAR (wins above replacement) to calculate both batter and pitcher value and then I would be able to combine my two models as opposed to have two separate models for each type of player. However, WAR, along with most other types of advanced saber metrics, are not calculable from just the basic statistics, leaving me to be use wOBA and FIP instead.

With these saber metrics and key statistics, I included a number of visualizations. These visualizations were, in order, I created histograms for the salaries of the players, ERA, Hits, wOBA, a box plot for ERA versus FIP, a scatter plot for FIP ratios with line of best fit and mean line, a scatter plot for wOBA ratios with line of best fit and mean line, a side by side line graph of the 2 lines of best fit and the 2 mean lines from the previous 2 visualizations, and a comparison of the two mean lines evened out flat. Each visualization tells me something different about my data and gives me a unique insight into the data visualized.

Along with numbers other observations and visualizations, I set up a linear regression model to determine the where the league average ratio line was located. In theory, all points below the line indicated poor performance compared to league average while the scatter points above indicated the opposite. With this model, we will be able to identify which players outperformed their salaries, thus creating a market gap for a team to exploit and gain the financial advantage.

There are really three different ways to use this model. First, you can look for outliers to find a diamond in the rough that is over performing his salary. If you look at the model, you will see a larger cluster of points at a very low salary, which is due to the financial system set up. However, once further right on the model, you can see several outliers far above the line that would indicate players far exceeding their salary. The other method of using these models is by looking at a group of specific points. Whether it is just for a specific team, player, or year. Looking at the standard deviation of the line from each of those points would create a large data set value to show how that group compares to the total data set. Lastly, we can look at a players projection and past performance, and compare with our model to determine the best salary to offer to ensure a contract that the compensation is not above performance. By looking at a players data points, you can determine how much you should offer contractually. This gives us a competitive edge understanding how to best allocate resources and what is overpaying for a player.