Brett Foster

DSC 530 – T302

Data Exploration and Analysis

Final Project - Baseball

November 17, 2020

Important Factors in Baseball

What are the most important factors for a baseball team to reach the pinnacle of the season, the playoffs? What factors leads a team to the most wins in the league, or to the top of their division for a chance for the ultimate trophy in baseball, *The Commissioner’s Trophy*. Statistics in baseball has been an important aspect of America’s Pastime and help explain all-stars, Most Valuable Players, and Cy Young Award winners, as well as many other attributes of the sport. Data analysis in baseball is nothing of a new subject. It is most recognizably utilized by the 2002 Oakland Athletics to reach the World Series that year, which was the inspiration for the book, and then movie, *Moneyball: The Art of Winning an Unfair Game*. The book reflects on the statistics of batting average, stolen bases, and runs batted in as ancient indicators of offensive success in the game and hope to prove that on-base percentage and slugging percentage are more important to the ultimate success of a team.

Based on my analysis, the most important statistics for wins in a season, from a modeling standpoint, are Runs Batted In, Earned Run Average, and Total Bases, in that order. These variables, in simple terms, explain a team’s overall defense and offense abilities, the team’s hitting abilities in important moments, a team’s ability to limit their opponent’s runs, and the last as a measure of the number of bases achieved through hits, walks, HBPs, and hopefully the more optimized these results are, the more success. Although there may be variables that correlate to more wins, these selective variables account for the best model for predicting outcomes.

Surprisingly, the changes to the baseball rules in 2020 had a considerably greater effect on the American League compared to the National League. The American League endured a major decrease in Batting Average that can only be attributed to the COVID-19 implementations of AL teams playing more games against NL teams than in a non-pandemic affected season.

I found this surprising until I did a more complete analysis on the American League vs National League statistics in which the National League scored stronger in 6 out of the 8 compared statistics during the 2020 season, which strongly suggests the National League is superior than the American League with all things considered. Additionally, a difference between the American League and National League did exist prior to 2020 where Designated Hitters were not utilized in the National League, causing the AL to utilized 9 professionally trained hitters where the NL only uses 8 with the pitcher being the 9th hitter, typically focused on advancing runners with bunts or maintaining a minuscule batting average. As a result, the AL does have ~3.5% greater ERA as well as slightly better Batting Average and Slugging Percentage than the NL.

Things I felt I may have missed in the analysis was a focus around player performance beyond that of just team performance. Unfortunately, I felt the focus of the analysis would have greatly shifted beyond that have simply winning percentage and league success if that were the case.