Rahul Patel

Prof. Vijayakanthan

CSIT 114 - Python II

3 May 2023

Predicting Professional Baseball Games - An Analysis of Team Statistics in the MLB

Abstract

As it is America's pastime, baseball is a sport with a rich history and has been recorded diligently for over a century. It involves complex statistics, and the rise of machine learning and data analytics has opened the door to newer and better approaches to predicting games. This project explores using algorithms and statistical analysis to predict the outcome of Major League Baseball (MLB) games in 2023. This project involves gathering data on specific teams and evaluating their performance in categories such as walks, strikeouts, and hits. It is then used to generate a model that shows the score of a game after inputting the two teams and the location of where the game is being played. Visualizations are also included in determining which variables have correlations and whether they should be considered. The project results show the importance and potential of statistical analysis to provide insight into a team's performance and the outcome of their games. In the predictive model, we can identify critical factors contributing to a team's success and make well-informed decisions about future matchups using historical data. Overall, this project shows the influence and power of data analytics in sports and emphasizes the potential for further research.

Keywords

MLB, Statistics, Data Analysis, Game Prediction, Team Performance, Predictive Modeling, Regression Plots

Introduction

As baseball is one of the most watched and played sports in the world, it is one of the most interesting to statisticians because of all the considered variables. Variables, such as strikeouts, errors, and walks, are extensively used to determine who should play against specific teams and what defensive position players should be in when a particular batter is up to the plate. There are endless possibilities for what can happen in a game, so statistics can help managers and players gain a competitive edge to win games and have a successful season. In the past decade, there has been a spike in interest from the sports community to provide a tool that fans and analysts can use to predict games based on historical data from previous games. As we analyze hitting, fielding, and pitching, the project will help identify patterns and trends that reveal which teams have the advantage. To accomplish this goal, we have developed a program that uses Python techniques to analyze large amounts of data. The program involves basic formulas and principles to identify positive or negative correlations, which can then be used to determine the outcome of future games. This will not only help us better understand the game and make more accurate predictions but also give the baseball community a tool that will help them make well-informed decisions in the future. Thus, we hope to provide fans and analysts with a more engaging experience with detailed and accurate statistics.

Design/Methodology

This project's design and methodology involve using historical data and definitions to analyze and predict MLB games based on team statistics. The first step was to collect the data and clean it. We extracted the data from three websites: FanGraphs and Baseball Savant, and Team Rankings. The primary reason for obtaining data from FanGraphs and Baseball Savant was to determine which variables are the most important to include in the model. We copied and pasted the data values into a spreadsheet and then organized it by removing the unnecessary rows and

columns between the values. We then exported the spreadsheet as a CSV file. In order to find any correlations between the variables, we had to create regression plots. There were five regression plots made:

1. On-Base Plus Slugging vs. Weighted Runs Created
2. Batting Average vs. Strikeout Rate
3. Isolated Power vs. Slugging
4. Stolen Bases + Caught Stealing Percentage vs Ultimate Base Running
5. Park Factor vs. Clutch

Next, we determined which variables should be included in the predictive model from the regression plots and extracted the data from Team Rankings. If the variables had a positive or negative correlation, we noted it and incorporated them into the model. Next, we copied and pasted the values again, organized the spreadsheet, and exported it as a CSV file. We then created the Python code and used formulas from Team Rankings that allowed us to determine the runs created and those allowed by the home and away teams. The Python script read the CSV file uploaded in the Jupyter Notebooks and executed it after the user imputed three items: Home Team, Away Team, and Location of the Game. The output was the score of the game (i.e., "Team A wins with a score of X to Y"). As the model was executed without problems, we tested it on past games. We took note of the results and will provide them here.

Results, Evaluation, and Test Cases

From the regression plots, we determined that the following had a strong correlation: (1) On-Base Plus Slugging vs. Weighted Runs Created and (2) Isolated Power vs. Slugging. Although the other plots did not have a strong correlation, they did show strong characteristics for the formulas, runs created, and runs allowed. Although no strong correlation, Stolen Bases + Caught Stealing Percentage vs. Ultimate Base Running is essential for runs created. If players

could steal bases and get into scoring position, they have a higher probability of scoring when their teammates are batting. If they can't steal bases and aren't in a scoring position, their chances of scoring will slim to 10-20%. A surprising factor that most analysts don't take into account is the location of where the game is being played. It makes a significant difference if the game is on the west rather than the east. For example, the Colorado Rockies Stadium (Coors Field) is the best stadium for hitters and the worst for pitchers. The stadium is one mile above sea level, so there is more scoring. Thus, managers might want to play pitchers who give up more ground balls than pop flies, as they might not give up so many home runs. If the game were located on the east coast, managers would focus more on hitters than pitchers. So, after considering everything and creating the model, we tested it on previous games to see whether or not it was accurate. We tried 90 games over the past few months, all selected randomly and three from each team. The model was 71.1% accurate. This means that 64/90 games predicted the winner correctly. Here are a couple of games that were picked and were the most accurate:

1. 4/11/23 - Red Sox @ Rays
    a. Actual Score: Red Sox 2 - Rays 7
    b. Model Score: Red Sox 1 - Rays 6
2. 4/17/23 - Diamondbacks @ Cardinals
    a. Actual Score: Diamondbacks 6 - Cardinals 3
    b. Model Score: Diamondbacks 6 - Cardinals 4
3. 4/28/23 - Reds @ Athletics
    a. Actual Score: Reds 11 - Athletics 7
    b. Model Score: Reds 10 - Athletics 7

In order for a model to remain accurate, it should have at least a 57.8% accuracy. We surpassed the margin by 13.3%. Therefore, the predictive model is statistically valid, and we can use it to make well-informed decisions about future games.

Conclusion

In conclusion, we have proven that predictive models can be applied to analyze statistical data in professional baseball and provide users with information to help them make wise decisions about predicting future games. In addition, the runs created and runs allowed formulas allow the model to predict accurately and encourage team managers to make data-driven decisions that will develop effective strategies for offense and defense. Overall, we expect more accurate insights and predictions as further advancements in this field of data analysis continue to evolve.

References:

Major League Team Stats 2023 Advanced Statistics: FanGraphs Baseball. Major League Team Stats 2023 Batters Advanced Statistics | FanGraphs Baseball. (n.d.). Retrieved May 3, 2023, from https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=0&type=1&season=2023&month=0&season1=2023&ind=0&team=0%2Cts&rost=0&age=0&filter=&players=0&startdate=2023-01-01&enddate=2023-12-31

MLB STATS. MLB STATS | TeamRankings.com. (n.d.). Retrieved May 3, 2023, from https://www.teamrankings.com/mlb/stats/

Statcast Park Factors. baseballsavant.com. (n.d.). Retrieved May 3, 2023, from https://baseballsavant.mlb.com/leaderboard/statcast-park-factors?type=year&year=2023&batSide=&stat=index_wOBA&condition=All&rolling=&sort=3&sortDir=desc