Major League Baseball Player Performance and Rewards Analysis

During this time of the Covid-19 virus, professional sports leagues have essentially shutdown. Among them is Major League Baseball (MLB) whose season is yet to start, and there is uncertainty whether or not the league will even have a season this year. Fans, including my sons and myself, are eagerly waiting for games to entertain us once again. My oldest son who is still in daycare loves baseball and has shown some talent in the sport. Back when I was single, I never understood why so many families were spending so much time with children on traveling teams, but now I truly understand their passion. For my project titled, “Major League Baseball Player Performance and Rewards Analysis,” I will analyze baseball data and focus on how performance relates to rewards. This may help fill the entertainment void with the absence of a baseball, while giving me a head start in case my son pursues a baseball career.

There is substantial data available for professional baseball players and their performance. Most people are familiar with batting statistics which include homeruns, RBI’s, and batting average. Other metrics are fielding statistics, which include error and assist, and pitching statistics, which include win-loss record, innings pitched, ERA, strikeout, walk, and complete game among many others. These performance metrics may be reflected in rewards for the player which can include a high salary, awards such as MVP, Silver Slugger, or Cy Young, being named an All-Star, induction to the Hall of Fame, or winning a championship.

For this project, I will attempt to mine frequent patterns in player performance data and rewards data. There may be certain association rules that exist such as “Players who lead in homeruns, RBI’s and batting average for the season tend to win the MVP for that year” or helps to answer questions such as, “Do pitchers who have the best win-loss record for the year win the Cy Young?” I will also investigate data metrics that can be used for clustering. For example, there is likely a relationship between overall team player performance and team win-loss records.

The data for this project will be the Baseball Databank which can be found on Kaggle-<https://www.kaggle.com/open-source-sports/baseball-databank>. It is a collection of historical baseball data from 1871 to 2015. This dataset contains 20 files, and the main tables include a master table for player and biographical info, batting statistics, pitching statistics, and fielding statistics. There are also tables that contain information for player salaries, awards, All-Star appearances, and Hall of Fame votes.

In terms of the evaluation of the methods, strong association rules may help to predict whether a person receives a particular reward or not. For example, if we plug in player performance data and determine with a strong likelihood that a player should receive the MVP award, we could confirm if that player indeed won the award for that year. In a similar way for the pitching statistics, if we specify the win-loss record, it may help to identify whether the pitcher wins the Cy Young award. Clustering may yield interesting insights for player performance.

I am looking forward to applying the different concepts, techniques, and algorithms from this course to the Baseball Databank. The dataset seems to be very robust. I hope to identify interesting relationships between player performance and rewards.