Tom Bash and Conor Keating

Methods of Data Science

April 6, 2021

Project Proposal

Section 1 – Introduction

Research Question: What are the most influential variables when trying to predict a Major League Baseball position players’ first year arbitration salary.

**Important details to understand:**

Players receive Major League service time for each day spent on the 26-man Major League roster. Service time is used to determine when players are eligible for salary arbitration. Each Major League regular season consists of 187 days and each day spent on the active roster or injured list earns a player one day of service time. A player is deemed to have reached one year of Major League service upon accruing 172 days in a given year.   
 All players with between three and six years of Major League service time become eligible for salary arbitration. They can earn substantial raises relative to the Major League minimum salary. Additionally, Major League Baseball each year identifies the group of players that ended the prior season with between two and three years of Major League service and at least 86 days of Major League service in that season and designates the top 22 percent -- in terms of service time -- as arbitration eligible. Those in the top 22 percent -- "Super Two" players -- are also eligible for salary arbitration despite having less than three years of Major League service. If the club and player have not agreed on a salary by a deadline in mid-January, the club and player must exchange salary figures for the upcoming season. After the figures are exchanged, a hearing is scheduled in February. If no one-year or multi-year settlement can be reached by the hearing date, the case is brought before a panel of arbitrators. After hearing arguments from both sides, the panel selects either the salary figure of either the player or the club, not one in between, as the player's salary for the upcoming season.

In order to create an accurate model, we will just be trying to predict salary for first year arbitration eligible players.

Our data is collected from a combination of different websites including BaseballReference, Spotrac, MLB.com, and MLBTradeRumors.

The goal is to be able to build a model in order to calculate what a player’s salary in the future should be.

Variables Identified:

Batting Average, On Base Percentage, Home Runs, Runs Batted In, Slugging Percentage, On Base + Slugging Percentage, Stolen Bases, Innings Played, Plate Appearances, Strikeout Percentage, Walk Percentage, Position Played, and Defensive Runs Saved

Section 2 – Data

All data has been extracted from [baseballreference.com](https://www.baseball-reference.com/). All the batting and fielding statistics are displayed neatly, and the player salaries are available to see on another page.

Section 3 – Data analysis plan

The target variable is to predict salary. The independent variables are Batting Average, On Base Percentage, Home Runs, Runs Batted In, Slugging Percentage, On Base + Slugging Percentage, Stolen Bases, Innings Played, Plate Appearances, Strikeout Percentage, Walk Percentage, Position Played, and Defensive Runs Saved

We made many graphs to compare one individual at a time to salary. After examining each of the graphs, it looks as if Batting Average and Home Runs are most commonly paired with a high salary.

We will be looking at linear regressions and viewing the R-squared value for each graph. This will show us the strength of the correlation between the variables. We will also be using summary statistics in R to build a model. We could do a linear model, but we will test to see what kind of model will be the best for our data. We would expect home runs and batting average to be the best predictors of salary, but we also want to build a model for future use. We would like to plug in a player's stats before they file for arbitration to see if we can accurately predict their salary. We could also see if players are underpaid or overpaid.

Sources:

<https://www.mlbtraderumors.com/2017/01/2017-mlb-arbitration-tracker.html>

[Baseballreference.com](https://www.baseball-reference.com/)

<https://www.spotrac.com/mlb/arbitration/2020/infield/settlement/>

<https://www.mlb.com/glossary/transactions/service-time>

A couple of notes:

We had a hard time importing the excel file into GitHub, so we copied and pasted all the cells into the text box.

Also, we will include the graphs we used for EDA later.