Project 1

Andrew Curtis James Seelos

January 23, 2023

1 Introduction

The purpose of this analysis is to elucidate the distribution of baseball salaries for recent years and their relation to certain batting statistics. This analysis could be used by a baseball club manager to determine what an expected salary for a player should be. When negotiating player contracts or trades, the manager could determine if they are getting an above, fair or below average expected value for the player.

2 Dataset

The Lahman's baseball dataset contains a number of baseball statistics for players and teams. Many options for analysis are available but we will focus on player's salaries and batting statistics. The years available for the dataset run from 1871 to 2021 for batting statistics and up to 2016 for salaries. In order to keep the analysis relevant to the current time period, we have chosen to limit the data set to the four most recent years available, 2012-2016.

3 Analysis Technique

Three types of machine learning regressions were used to determine which batting statistics were features important to determining player salary. These were decision tree, figure 1, linear regression, and random forest, figure 2. From there a reduced model and individual batting statistics are investigated further.

4 Results

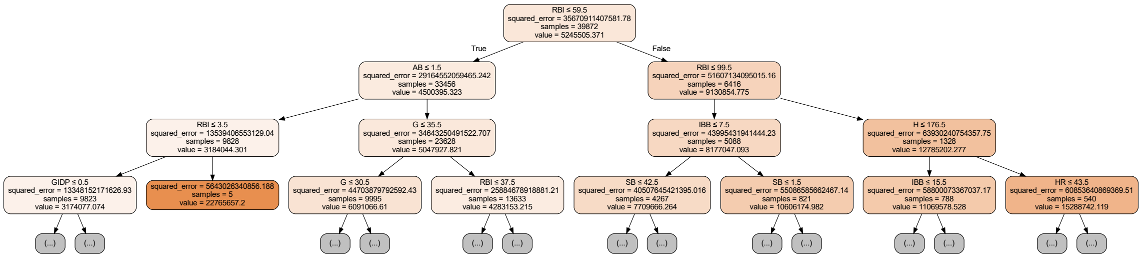

Figure 1: Decision tree regression of salary on batting statistics. Value is the contract salary in $ dollars.

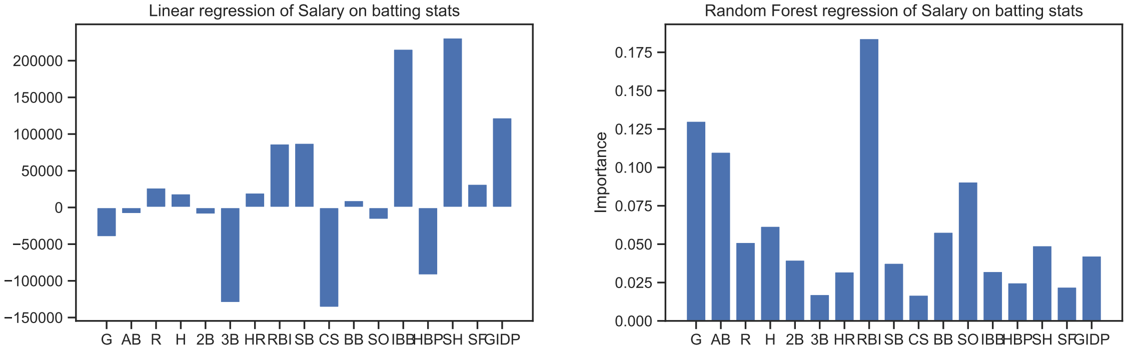

Figure 2: Regressions of salary on batting statistics. The importance of each feature is shown for two methods. Linear Regression (left) and Random Forest (right)

Based on the machine learning models a reduced amount of features were selected for further investigation, including: hits(H), runs batted in(RBI), walks(IBB), home runs (HR), and games played(G). These statistics correlations with salary and each other are presented below in the scatterplot matrix. From this visualization we can determine there is some possible multiple collinearity. This could be researched in further work. It appears that higher amounts of hits, games played, and walks are all important features for predicting salary. However, players that are starters would tend to accrue higher numbers for these statistics and would also tend to receive a higher salary since they are assumed to be a better player. So instead we select RBI's and HR's specifically as statistics to regress salary on. These could be more reliable batting statistics to use as predictors. Linear regressions for those two are shown below in figure 3. Following these regression lines you could say look at the RBIs a player had in the previous season and determine what salary would be fair compensation.

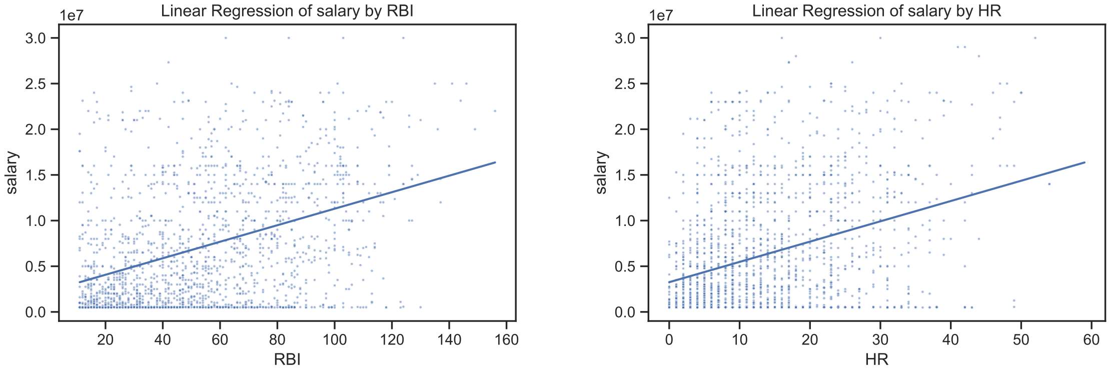

Figure 3: Regressions of salary on RBIs(left) and home runs(right).

5 Technical

Several steps were taken for data preparation. Datasets were merged on playerIDs. The data was limited to only the four most recent years in the dataset, 2012-2016. In order to ensure that players were active (played and were not injured all season), the dataset was limited to players that played more than 10 games. There were many batting statistics available as features to use as predictors for salary. We initially wanted to see which might be important so the machine learning regression techniques were used on all features and importance was viewed. Decision tree and random forest selected similar statistics and those were chosen to be investigated further. Statistics that would be accrued from playing more games were generally more important. This could be because starting players played more of these games. RBIs and home runs were selected for reduced linear regression models for predicting salary. Multicollinearity is possible for some of the features and that could be investigated in further work. It would also be interesting to feature engineer some other batting statistics like batting average, slugging and ops for use in the models. However, we do believe these final two models presented could be useful as is.