Data Mining Final Project: What Makes a Winner

Colin McNally

5/7/2022

Abstract

Individual success is hard to measure in most professions, but because sports have well documented statistics it is easy to measure if a player has a good or bad season. Especially in a sport like baseball where almost every aspect of the game can be broken down into a statistic. As well it is quite easy to measure individual success on a baseball field. This is because baseball is an incredibly lonely team game. It is always one batter against one pitcher, and this allows for individuals in baseball to easily rise above or fall short of others. When beginning this research I wanted to find out what makes a legend on the mound or at the plate? With methods learned throughout the semester, like *Random Forests*, *Scaling*, and *Train Test Splits* I set out to find what makes a legendary baseball player. What I found was that *Wins* and *Strikeouts* make great pitchers, and that *Batting Average* and *RBIs* creates a standout hitter. As well I was able to determine which individual season performances were the most remarkable. The standout seasons for pitchers in my model were Sandy Koufax 1963, Dwight Gooden 1985, Pedro Martinez 1999, and Bob Gibson 1968. Hitters on the other hand were dominated by a single player especially: Barry Bonds. Our model predicted other players had great seasons like Babe Ruth in 1923 and Larry Walker in 1997, but Barry Bonds had 4 out of the 6 best hitting seasons ever according to our model.

Introduction

Sports are endlessly debatable, there's always a discussion of who was the best player of all time or which season was the greatest of all time. In a sport with such a rich history like baseball there have been thousands of player to play over the past 100-plus years. In those years there have been a lot of great players, but which one was the best? Which one had the best season ever? And what stats exactly qualify someone to be named the best player? First we must separate players into to two types of players, Pitchers and Hitters. Pitchers and Hitters have to wildly different jobs in the game of baseball, a pitcher's job is to get outs for their team and prevent hitters from scoring runs. A hitter's job is quite the opposite, their job is to get hits and put their team in position to score runs to win. For the sake of comparison, it is almost impossible to compare the greatness of a pitcher to a hitter. So in this study we will treat these two types of players separately. Now the questions change to what was the greatest pitching season of all time, or what stats make a great hitter? As an avid baseball fan I grew up watching the likes of Sammy Sosa hitting 64 *Homeruns* and thinking that for certain he was the greatest baseball player ever. I was only about 3 years old when this happened, and did not know that baseball had such a rich history of hitters. Many hitters in fact had accomplished much more than Sammy Sosa ever did, and they did it without the *help* Sosa used. As I got older, I learned more and more about the statistical side of baseball and how every aspect of baseball was tracked with statistics. Now with statistical training and new data science methods, I would be able to truly know if Sammy Sosa had the greatest season of all-time when he had 160 *RBIs*, had a *Batting Average* of .328, and hit 64 *Homeruns*.

Methods

I started this project by initially looking for free public use data that was readily available. Through *GitHub* I was able to find Lahman's Baseball Database, a free to use database that tracks basic baseball statistics such as *Hits*, *Homeruns*, *ERA*, and *Strikeouts* per player per season. In the tables I had to create some of my own stats that were not tracked but countable from the data given like *Battting Average*, *Slugging Percentage*, *OPS*, and *WHIP*. The data sets in Lahman's Baseball Database did have some limitations on statistics because of the simplicity of its measurements. As it only measured the basics of hitting and pitching it was missing many advanced statistics that pervade the modern game of baseball. Statistics such as *WAR*, *FIP*, *wOBA*, *BABIP*, *Spin Rates*, and *ERA+* were all unavailable in this data set. In modern era baseball sabermetrics such as *WAR* and *Spin Rates* are often cited in MVP and Cy Young debates along with basic stats like *Batting Average* and *ERA*. As well, I excluded data on fielding from my analysis. The reason for the exclusion was that fielding, though it is important to the game of baseball, does not factor into the MVP or Cy Young Award debates very much. Basic statistics for fielding as well have various problems. An error for one fielder is a base hit for another fielder. As some players are faster or have stronger arms, the ability to measure errors and true fielding ability is imprecise. Finding precise data on fielding takes sabermetric level data which I do not have access to, thus I must omit fielding from my models. On top of hitting and pitching data, Lahman's Baseball Database also had data on every major award that had been given out since 1911, including the Cy Young Award and MVP. The Cy Young Award is handed out to the best pitcher in each league every year. The MVP is handed out to the best player in each league each year. Though pitchers can win the MVP, I used the MVP specifically for hitting data. As pitchers have their own award and no pitcher has won the MVP without also winning the Cy Young Award. In my research I purposely excluded any pitchers from the MVP Award. In the data set on awards, there was data on the individual votes for each player to win the award and the maximum amount of votes they could earn. As there have been changes in the number of maximum votes each year, I decided instead to use the *Vote Share* for awards as a measurement for success. If a player continually has high *Vote Shares* then the public is recognizing their achievements in baseball that year. A player's individual stats are commonly cited as reasons for them achieving an MVP or Cy Young Award. This stat of *Vote Shares* would be my metric for success of an individual player, the higher the *Vote Share* the more dominant the player was that season. Measurements like *Vote Share* are imperfect as there is no criteria for what makes an MVP or Cy Young, but that is what I will be testing in this paper. Sometimes position makes a difference in voting, such as a catcher can win MVP with lower statistical output because their position is not expected to be great hitters. As well closers can have incredibly low *WHIP* and *ERA* but are largely excluded from Cy Young discussions as they do not pitch as many innings as a starter. *Vote Shares* is the only way with the data available that I could measure individual success of hitters and pitchers. With *Vote Shares* I constructed a model based on hitting and pitching data that would predict the expected *Vote Share* of a player based on their stats. The model I used was a *Random Forest* that would optimize which statistics mattered most for determining a player's success. To make sure that I was minimizing my error, I split my data into *Training* and *Testing* sets so that I could measure the out of sample accuracy of my model. Then after the *Random Forest* discovered the best model for the data, I fitted the model onto players to quantify the greatness of their individual seasons.Now the fitted value of *Vote Share* would be based on the statistical achievements of the players rather than the voting of baseball journalists.

Results

Pitching

When fitting the model for pitching I included 8 variables for determining *Vote Share*. The variables included for the *Random Forest* to use were: *Wins*, *Strikeouts*, *WHIP*, *Losses*, *ERA*, *Opponent's Batting Average*, *Saves*, and *Homeruns* given up. I first put all the variables through a single decision tree first to then compare later to the *Random Forest*.

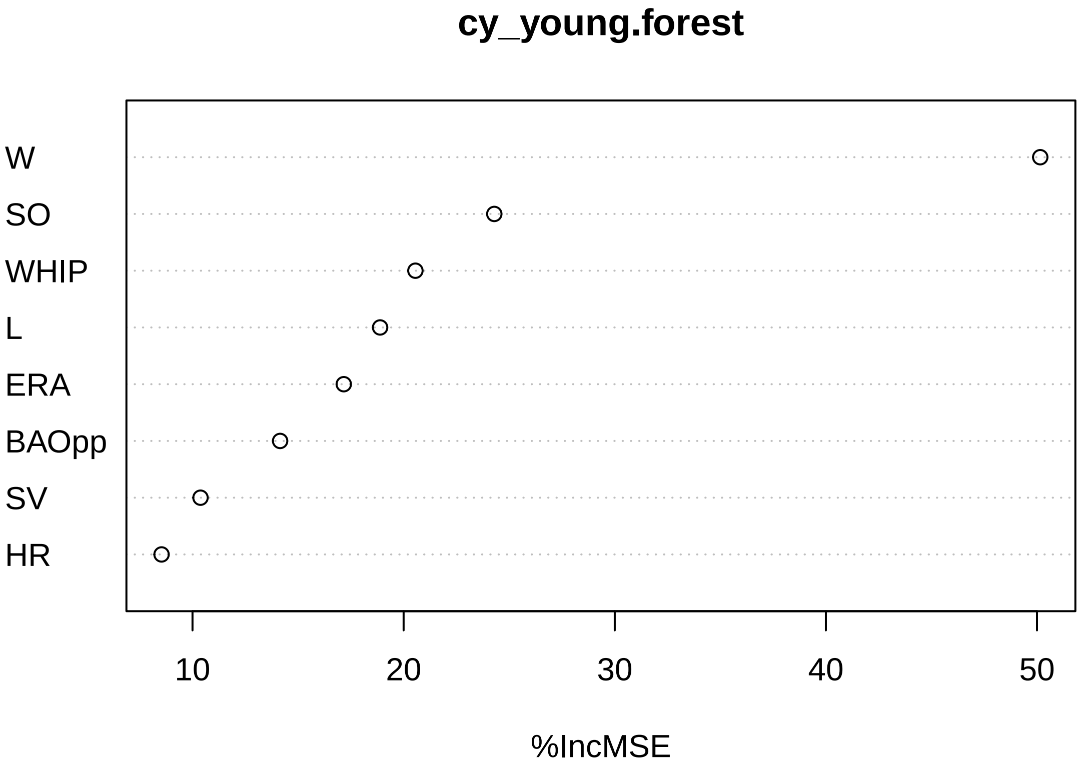
Single Decision Tree RMSE

```
## [1] 0.2625239
```

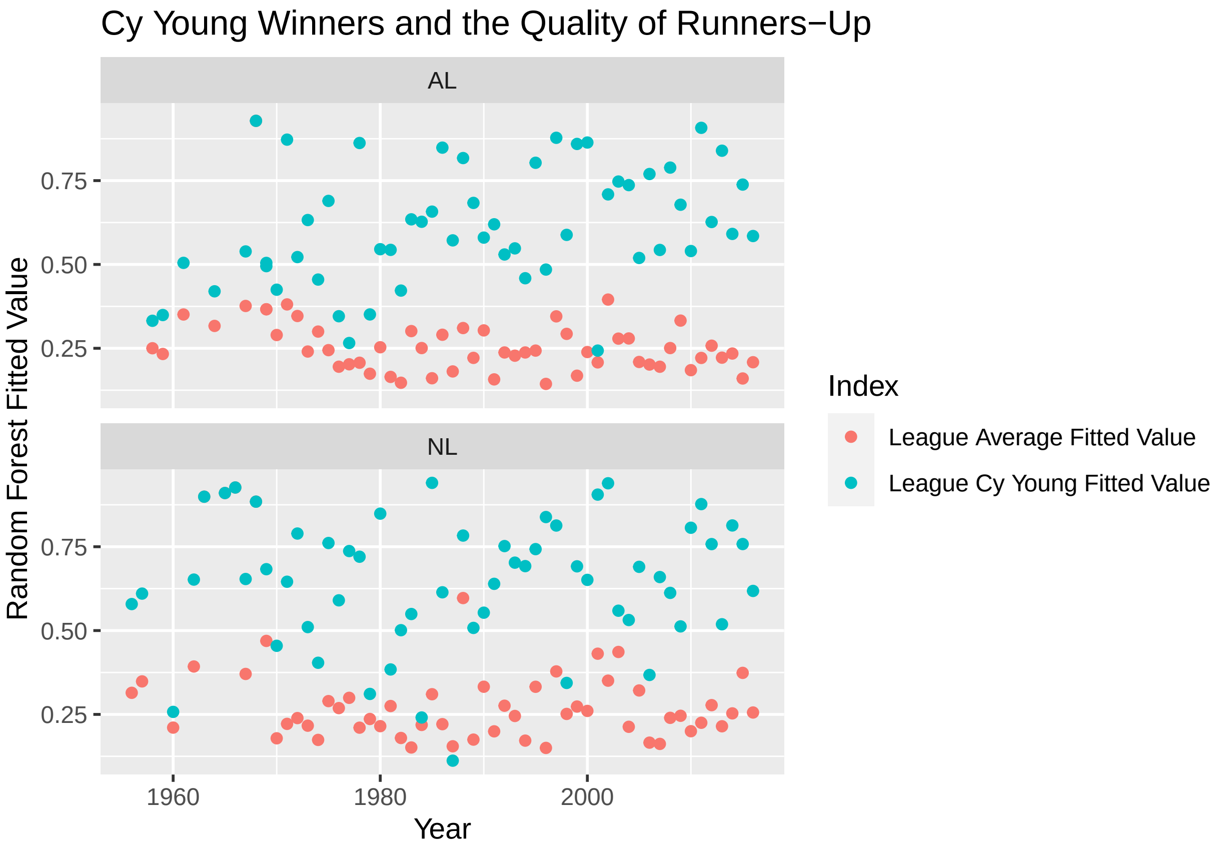
Random Forest RMSE

```
## [1] 0.2274575
```

After passing data through both of the models we can see that the *Random Forest* has a much lower out of sample Root Mean Squared Error. The change in RMSE was a decrease of about 13.36%. This decrease in Root Mean Squared Error show that our *Random Forest* was a more accurate estimate of *Vote Share* than just our single decision tree.



After creating the *Random Forest*, I wanted to know which variables had the largest impact on *Vote Share*. So, I put the *Random Forest* into a *Variable Importance Plot* which shows the percentage increase in Root Mean Squared Error when omitting a variable. In the Cy Young Award, the omission of *Wins* saw an increase of about 50% in Root Mean Squared Error. This was the highest of any variable meaning it was the most important variable in predicting *Vote Share* for the Cy Young Award. The lowest of any of the variables was *Homeruns* surrendered by the pitcher at around an increase of only about 8% when omitted.



Next I wanted to see how Cy Young Award winners performed comparatively to the league they played in to see how much better the winner was to league average.The plot above shows a gulf between most Cy Young Award winners and their competition. This is to be expected as these pitchers were the best performers in the league that year in the opinion of journalists. The gulf between the winners and league average also tells us that our model is correctly predicting that these players were standouts that year as a pitcher. Some anomalies did occur such as in 1987 in the NL the Cy Young had a fitted value lower than the league average. The award winner was Steve Bedrosian a closer with good stats but was a controversial winner at the time. We also see on this graph years where there were no league average such as in 1963 or 1968 in the NL. This is because the pitchers in those years were so dominant that there was not data on the league average for *Vote Share* because these pitchers were the league average. There was no other data on any other pitchers receiving votes because these were unanimous selections. Of course these were completely dominant seasons as this only happened 4 times. As well every single one of those seasons under our model were a top 10 single season performance.

Hitting

The model for hitting had more variables as there are more basic statistics that are tracked in the hitting process. There were 11 variables included in the model. The varaibles included were *Batting Average*, *On Base Percentage*, *Slugging Percentage*, *OPS*, *RBIs*, *Doubles*, *Triples*,*Homeruns*,*Runs*, *Walks*, and *Stolen Bases*. I then put all of these variables into a single decision tree model to use as comparison against the *Random Forest*.

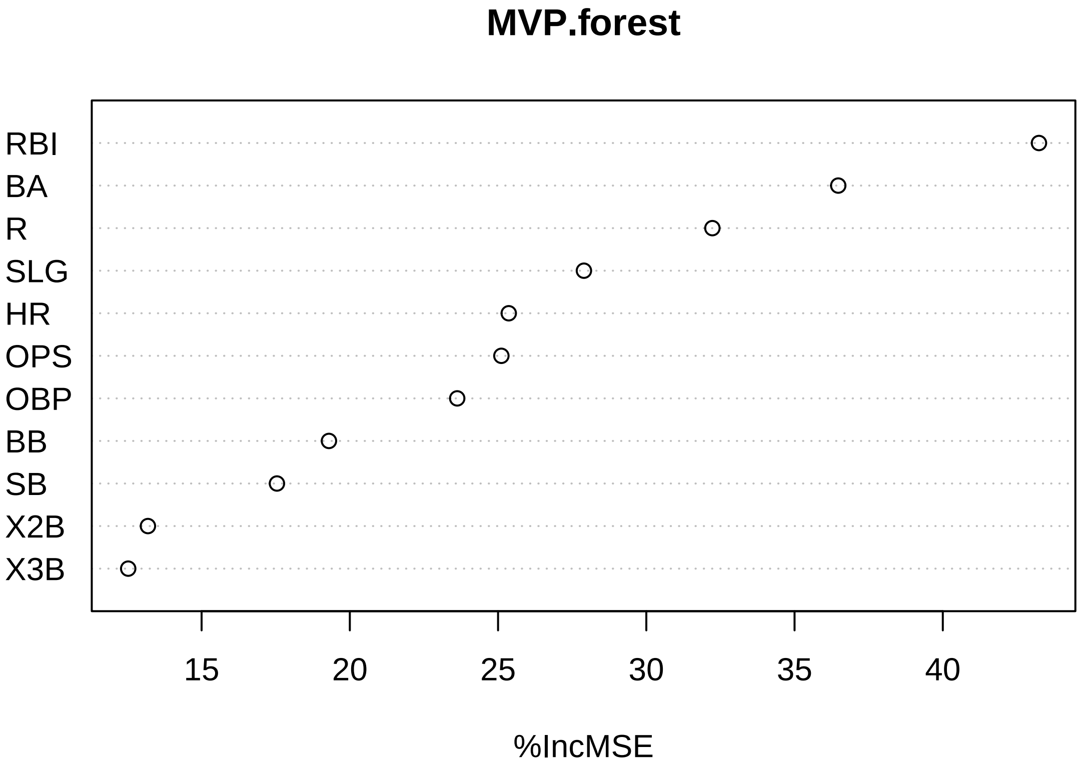
Single Decision Tree RMSE

```
## [1] 0.2173242
```

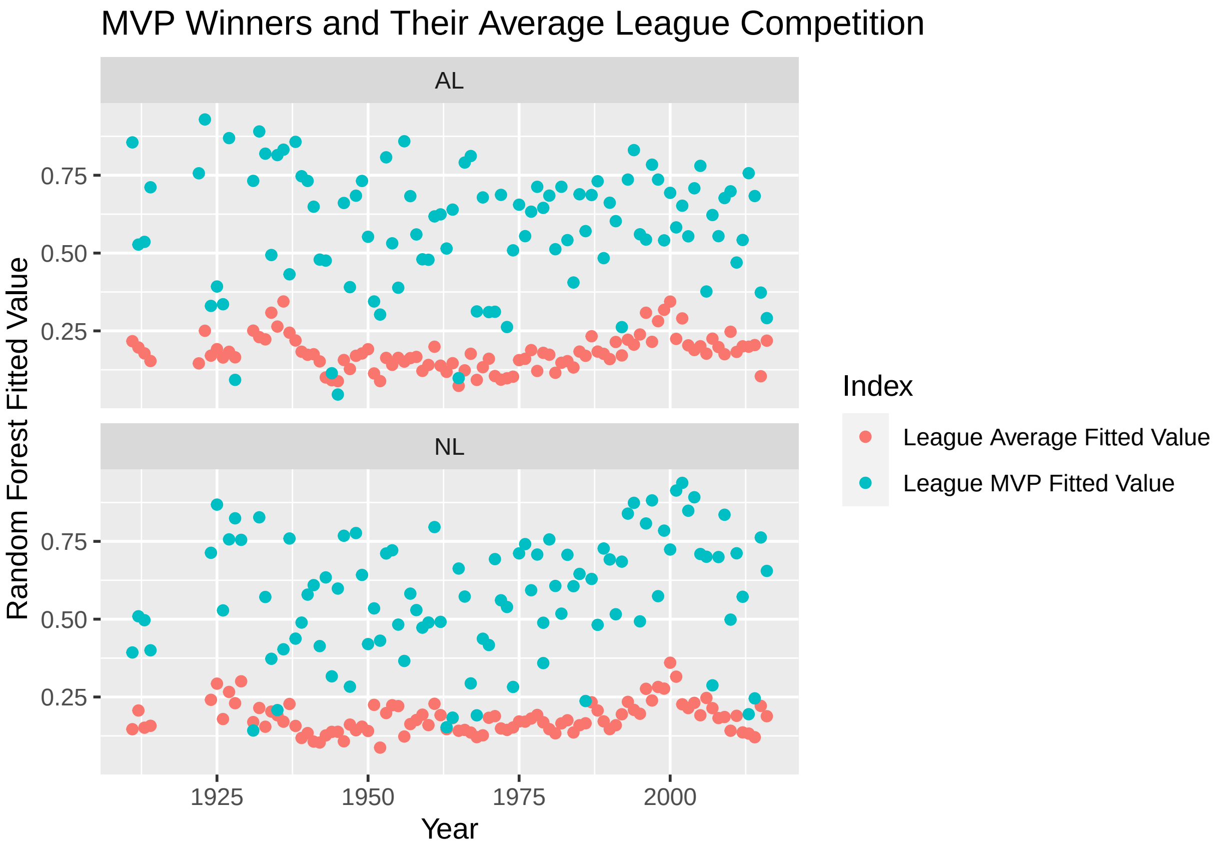
Random Forest RMSE

```
## [1] 0.1932394
```

The change in RMSE here from the single decision tree to the *Random Forest* was about an 11.08% decrease. The *Random Forest* is performing better than the single decision tree. As well this *Random Forest* is performing better than the Cy Young *Random Forest* but this is likely due to the large amount of observations in the set of MVP data compared to the Cy Young data.



Next to understand the importance of our variables I created another *Variable Importance Plot*. This time we found that the most important variable in our model for the MVP was *RBIs* which came in at just about 43% increase in Root Mean Squared Error when omitted from our model. While the least important of our variables was *Triples* which saw an increase in Root Mean Squared Error of about 12% when omitted.



The plot is again the same for MVP's as it was for Cy Young winners a large gulf between the winners and league average performance. There are more anomalies in this data set though compared to the Cy Young graphs. There are multiple MVP winners that perform worse than the league average performance. One instance of this is the 1931 when Frankie Frisch won the MVP, when he had less *Homeruns*, *RBIs*, *Runs* and a worse *Batting Average* and *Slugging Percentage* than the 2nd place finisher and 3rd place finisher. Frankie Frisch only lead the league in *Stolen Bases* that year. The big gulfs between the MVP's fitted value and the average league fitted value inform us of years when MVPs truly stood head and shoulders above the rest. While years like 1931 tells us that voters really did get things wrong that year and should have voted for someone else.

Conclusion

After running the model it was interesting to see which each individual seasons were the most impressive. Barry Bonds was always a great performer with our model and had some of the greatest hitting seasons of all time. Our model showed that his stretch from 2001-2004 were 4 of the 6 best hitting seasons of all time. Most likely making that stretch of 4 years the most dominant stretch of consecutive years hitting ever. The problem with this is that Bonds began using steroids around this time and that highly impacts his reputation. Had he not abused steroids there would be no doubt that he was the greatest hitter of all time. For pitchers there was a similar dominant performer: Sandy Koufax. Koufax was the premier pitcher of the 1960's and had 2 of the best pitching seasons ever according to our model. No pitcher other than Koufax had 2 seasons in the Top 10 of greatest pitching seasons according to our model. As well we get to see what voter's seem to value most. When it comes to winning the Cy Young Awards a pitcher almost is required to be on a winning team in order to win the award. *Wins* were far and away the most important statistic for predicting *Vote Share* for the Cy Young Award. To become an MVP on the other hand, hitters needed to get hits and especially when runners are on. *RBIs* and *Batting Average* were the best predictors for *Vote Share* in our model. Our model also predicted widely accepted award snubs, such as in 1987 with

Steve Bedroisian winning the Cy Young Award. Many analysts now look back and prefer the season Nolan Ryan had to Bedroisian. It is wildly regarded that Bedroisian won because every starting pitcher had a blemish on their resumé. Nolan Ryan's blemish was that he had too many *Losses*. Under the new model though *Losses* would hold less importance as only the 4th most important variable. As well we look back and see with our MVP data that Frankie Frisch was an unworthy candidate for MVP in 1931. Our model had Frisch below the league average for *Vote Share* and his statistics more than back that up. Frisch only had 4 *Homeruns* and 82 *RBIs* that year while batting 0.311. All of these numbers were quality but considering his competition for MVP his candidacy is diminished. The 2nd place in MVP that year, Chuck Klein, had 31 *Homeruns*, 121 *RBIs*, and batted 0.337. All of these stats better than Frankie Frisch and even his own teammate, Chick Hafey, had better *Batting Average*, *RBI*, and *Homerun* numbers and he finished in 5th. To finally circle back to Sammy Sosa and if he was the greatest player ever. In a very short answer, no he was not. Sosa did not even have a top 10 greatest hitting season according to our model. His MVP winning season of 1998 where he hit 66 *Homeruns* and had 158 *RBIs* was considered only the 23rd greatest season in history for batters. That being said 11 of the top 25 MVP hitting seasons ever were during the steroid era thus juicing the statistics. When it was all said and done our model was a very accurate predictor for success in baseball. Barry Bonds without steroids would be considered the greatest hitter of all time. While Sandy Koufax is routinely in discussions for the best pitcher ever.