Using Machine Learning to Predict MLB Hall of Famers

Chauncey Tanner Murphey

Spring 2022

Abstract

The sport of baseball is likely the most statistically analyzed (or over-analyzed, depending on who you ask[5] sports in the world. The sheer volume of available data tracking all the way back to the first days of the sport make it ripe for analysis. In this project, a collection of player-season level data is analyzed in order to predict which players are inducted into the coveted Baseball Hall of Fame. The player-season numbers are first aggregated into a variety of player-career level numbers. These numbers are then run through a variety of different supervised machine learning models and evaluated based on their accuracy. Similar stats are also compared against each other to see which ones are better at correctly predicting which players get into the Hall. Other than the initial dataset, which was found online, every Python file and Jupyter Notebook file was typed from scratch by me. The csv files with the processed data were generated by me as well.

1 Introduction

Since its inception, baseball has been at the forefront of sports analytics. Game numbers can be found all the way back to 1871. Stats like batting average (AVG) and earned-run average (ERA) have been used to evaluate hitters and pitchers for decades. The discrete nature of the sport in both its flow and outcomes (each pitch is a single event with a finite amount of outcomes) allows numbers to easily be tabulated and paint a near-complete picture of what transpired in the game. While nowadays many more details are recorded (pitch speed/selection, launch angle, etc.), major game stats like hits, runs, and strikeouts have been recorded for over 150 years.

In the early 2000s though, the analytics revolution began when Oakland A's General Manager Billy Beane started using on-base percentage (OBP) instead of just AVG to gain a competitive edge in scouting player talent[4]. Since then, analytics has exploded as teams and individuals continue to try to quantify talent in baseball. Nowadays, websites like Baseball-Reference and FanGraphs have their own advanced models for evaluating how much a player contributes to their team. While MLB teams use their own data and models to try to gain an edge over their opponents, there exists plenty of publicly available data for amateur statisticians to do a variety of analyses (like predict Hall of Famers).

2 Tools

2.1 Data

The data used for this project was a collection of player-season level numbers for every MLB player dating back to 1871[3]. It was collected by journalist Sean Lahman over the course of several years. Within the database are the hitting, fielding, and pitching tallies along with a few other miscellaneous pieces of data (like Hall of Fame voting). The hitting and pitching data of Hall of Fame-eligible players is passed through a variety of functions I wrote myself to calculate a collection of popular stats like ERA, AVG, and others. These stats were tested on individual players and verified via Baseball-Reference and Fangraphs The resulting collection of data what is then run against the models to see how telling they are for the different types of machine learning algorithms.

2.2 Libraries

The only libraries used are the ones typical of scientific computing and data analysis. They include pandas, scikit-learn, and matplotlib. The handling and preliminary processing of the data is handles via pandas dataframes while scikit-learn handles the ML models. The classifier models used are a decision tree, K-nearest neighbors, support vectors, and logistic regression.

3 Techniques

3.1 Initial Data Processing

The first step was slimming down Lahman's collection of players down just to those that would have been eligible for Hall of Fame induction and made it through their window of eligibility. This cleans out a large amount of unhelpful data, since ineligible players have no chance to make it into the Hall regardless of their stats and those not yet through their window of eligibility have yet to have their fate determined. The relevant rules, as set by the Baseball Writers Association of America (BBWA) are as follows [1]:

In order for a player to be eligible, they first must have played in 10 different seasons in the MLB. After they have retired as a player for 5 years, they enter a 10 year period of eligibility where they can be voted on by members of the BBWA. If a player receives 75% of the possible votes, they are inducted into the Hall of Fame that year. BBWA members are allowed to vote for a maximum of 10 players in any given year.

Trimming this down is trivial: only include players who've played for at least 10 years and have been retired for 15 years. This cuts down the number of players down by around 90%, showing how rare it is for a player to even play long enough to be eligible. Some basic tallies of the remaining players a calculated (awards won, World Series titles, Hall of Fame induction (true outcome), etc.). The players are then split into pitchers and hitters based on if more than half of their appearances were as a pitcher. While there have been two-way players like Babe Ruth who were famous for both their pitching and non-pitching abilities, they so rare that the work it would take to account for them would be impractical.

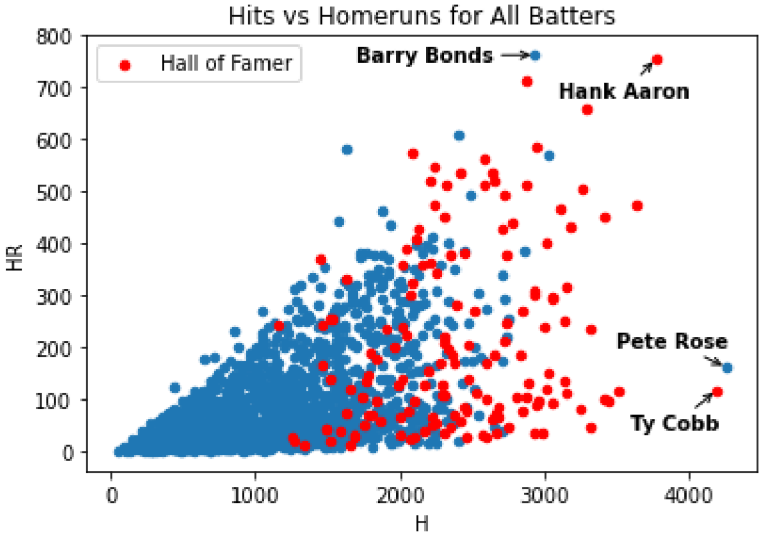

Figure 1: Career hits vs home runs for all eligible batters. Red dots are Hall of Famers

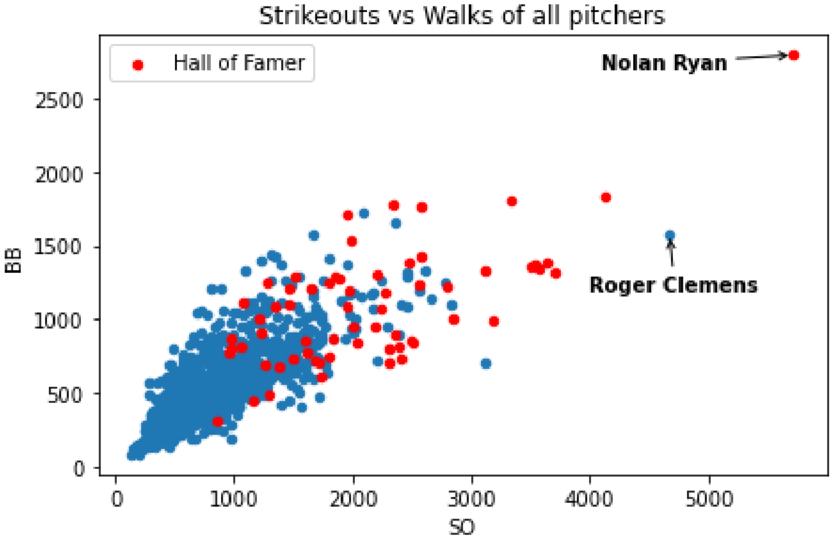

Figure 2: Career strikeouts vs walks for eligible pitchers. Red dots are Hall of Famers

3.2 Pitchers and Hitters

The pitchers and hitters were completely separated due to the fact that the qualifications for each to get into the Hall of Fame are vastly different. Not only do pitchers have their pitching stats that hitters don't, the hitting stats of pitchers are irrelevant for all intents and purposes. Pitchers in the American League haven't even needed to hit since 1973. They were replaced by a position called the Designated Hitter, whose sole job is to hit instead of pitchers. The National League just followed suit this year[2] as well. Therefore, the most logical procedure is to ignore hitting for pitchers and make the two groups completely separate. The a variety of stats were calculated for pitchers including strikeouts, hits allowed, ERA, WHIP, FIP, and ERA+, among others (see `baseballstats.py` for more info). For batters, stats like hits, home runs, AVG, SLG, wOBA, and OPS+ were calculated. After dataframes were constructed containing all of the stats for the players, they were each exported to their own CSV files for later use.

4 Stats

For hitters, the final dataset included their career totals in hits, runs, home runs, runs batted in, stolen bases, and walks. Additionally, their batting average (AVG), on-base percentage (OBP), slugging percentage (SLG), on-base plus slugging+ (OPS+), and weighted on-base average (wOBA) were calculated. Batting average is just the number of hits a batter has divided by their total number of qualified at-bats. On-base percentage is their hits, plus walks and hit-by-pitches, all divided by the number of plate appearances a batter has (different than at-bats). Slugging percentage weights the number of hits the player has by the number of bases that hit was worth (1 for singles, 2 or doubles, etc.). OPS+ is the on-base percentage and slugging percentages normalized to the league averages in each and summed together. Lastly, wOBA is Fangraphs improved version of OPS+, instead using a custom weighting for each outcome in on-base and slugging percentages to better evaluate how good a hitter was. These weights change slightly every year.

For pitchers, their dataset included their career totals in hits, earned runs, and home runs allowed as well as their wins, losses, strikeouts, and walks. Additionally, their career earned-run average (ERA), walks and hits per inning pitched (WHIP), earned-run average+ (ERA+), and fielding-independent pitching (FIP). ERA is simply the amount of earned runs allowed by the pitcher per 9 innings that they pitch. WHIP is self-explanatory, the sum of their hits and walks divided by the total innings pitched. ERA+ is simply league average ERA divided by the pitcher ERA (so that higher number is better). Fielding independent pitching assigns weights to each outcome the pitcher has sole control over (strikeouts, walks, and home runs) divided by the total innings pitched.

4.1 Machine Learning

There were 5 different machine learning classifiers used and compared against each other. Two of these were decision trees and random forests. The decision tree breaks apart the data along various columns. If enough breaks are made, the data can be broken down entirely and therefore overfit. Random forests, on the other hand, make several decision trees and average them together to make a model that is much less likely to overfit at the cost of being much more difficult to visualize.

K-nearest neighbor, another classifier used, constructs its model by classifying data points by whatever the result of the nearest k points are. If $k = 5$, and 3 of the 5 closest points to the test point are Hall of Famers, then the test point is classified as one as well. This method works well if the data is clustered together, but not if the data is mixed together.

Logistic regression can be used as a classifier, as this project has, but it is not inherently one. Logistic regression attempts to assign weights to different stats to convert a linear combination of this data to a probability. The classifier just then evaluates this probability to assign a prediction to the test data.

The last method used was a support vector classifier. This method looks for margins in the data to divide between the predictions. It then uses a softer margin on either side of the middle one to account for an imperfect fit. This one, like the decision tree, is prone to overfitting and also, like the nearest neighbor, is not suited for data that isn't well separated. The consequences of this can be seen in the results section.

5 Results

Batting	Basic Stats		Advanced Stats		**Pitching**	Basic Stats		Advanced Stats	
Classifier	0	1	0	1	Classifier	0	1	0	1
Decision Tree	0.98	0.71	0.96	0.61	Decision Tree	0.98	0.73	0.98	0.69
K-Neighbors	0.96	0.48	0.97	0.63	K-Neighbors	0.98	0.48	0.97	0.36
Support Vectors	0.96	0.42	0.96	0.50	Support Vectors	0.98	0.38	0.97	0.33
Random Forest	0.98	0.71	0.96	0.63	Random Forest	0.99	0.79	0.98	0.71
Log. Regression	0.97	0.63	0.96	0.62	Log. Regression	0.99	0.83	0.98	0.69

Table 1: F-1 scores for the all of the classifiers for both pitchers and hitters. Of the 1743 hitters eligible for the Hall, only 151 were inducted as players. For pitchers, only 63 of 991 eligible were inducted for their playing careers.

Table 1 shows the f-1 scores for each classifier for both the basic and advanced stat runs. For hitters, AVG, OPS, and SLG were swapped out for OPS+ and wOBA. For pitchers, ERA and WHIP were swapped out for ERA+ and FIP. The decision tree, logistic regression, and random forest classifiers seem to have the best accuracy, with the random forest having the slightest edge. K-nearest neighbor and support vector classifiers both perform significantly worse at predicting who got inducted (value 1). KNN being that bad is particularly interesting, since it means that players with very similar stats across all that were compiled may have very different results. SVC performing less is more understandable, since there is no significant transition along any stats for inducted vs non-inducted. These differences are more pronounced in the pitching table than the batting, likely due to the relative lack of eligible pitcher compared to batters.

Another noticeable feature is that the classifiers were much more accurate at predicting who was not inducted (0) than they were at predicting at who was (1). Looking at Tables 1 and 2, can see that there are large portions of eligible player not inducted that aren't close to the Hall of Famers, but many of those inducted are tangled in with players not inducted. While these two stats are far from the only ones used, they show how the rarity of Hall of Famers allows the non-inductees to create a lot of noise in the disputed areas while having their own portion of the chart with few or no inductees.

When comparing the basic stats vs the advanced ones, There are some particularly interesting results. The first of which is that all three of the ones that did well with the basic stats did worse with the advanced ones for both pitching and hitting. The ones that did poorly before, however, improved for the batting data when using the advanced numbers but did strangely did much worse in the pitching table.

6 Conclusion

The sport of baseball arguably has more stats baked into it than any other major sport in the country. Every day the Hall of Fame worthiness of great players, both current and former, is discussed. This makes analyzing some of this data and using it to predict one of the most discussed topics within the sport a fun and interesting project. Different types of classifiers were compared in how well they predict both pitchers and hitters getting into the Hall. The random forest model posted the highest accuracy across the board while the support vectors did the worst. Additionally, there were also mixed results when swapping out several basic stats for their more advanced counterparts.

References

[1] Baseball Writers Association. *Hall of Fame Election Requirements*. URL: `https://bbwaa.com/hof-elec-req/`.

[2] Major League Baseball. *Designated Hitter Rule*. URL: `https://www.mlb.com/glossary/rules/designated-hitter-rule`.

[3] Sean Lahman. *Baseball Archive*. URL: `https://www.seanlahman.com/baseball-archive/`.

[4] Michael Lewis. *Moneyball: the Art of Winning an Unfair Game*. W. W. Norton & Company, 2004. ISBN: 0393324818.

[5] Abraham Wyner. *Changing the Game: How Analytics is Upending Baseball*. URL: `https://knowledge.wharton.upenn.edu/article/analytics-in-baseball/`.