ACENET
Microcredential in Advanced Computing
ISP Report

Project title: Predicting Future MLB Hall of Fame Players and Past Snubs, and Outliers

Participant name: Daniel Daye

Date: 31 July 2024

Abstract

I will analyze the statistics of all MLB players who have and hadn't been inducted into the Baseball Hall of Fame (HoF). Using advanced stats, I will set a baseline for the statistics of all players inducted into the HOF and see if any have made it in with lower than baseline statistics. I will then use this baseline to determine if any non-inductees have been snubbed based purely on their statistics. Finally, I will determine which players are on track to meet the baseline numbers and predict who may eventually be inducted into the HoF when eligible. The data will not take into account those inducted or not inducted due to off-field, non-statistical data.

Introduction

The ACENET Microcredential in Advanced Computing is meant to introduce participants to advanced computing concepts and other topics, such as Python programming, data analytics and visualization, shell scripting, high performance computing, and machine learning. The culmination of the 5 month program is the Independent Study Project, or ISP, in which participants will apply the skills learnt throughout the program into one final project. All results presented in this report were obtained from utilizing the skilled acquired during this program.

Each participant was free to choose a topic of their choice. The topic I selected was analyzing baseball Hall of Famers. Specifically, I set out to analyze the stats of hall of fame players and use that to set a threshold of eligibility based on statistics. Using this threshold, I would determine whether any players inducted into the HoF did not meet the threshold for induction, or whether past players exceeded the threshold but were not inducted. Similarly, this threshold will applied to current, active players, or those who played within the past 5 years, to predict who may be inducted in the future.

Background

The National Baseball Hall of Fame (which may be referred to as HoF in this document), established in 1936 and located in Cooperstown, NY, was created with the goal of preserving and recognizing the greatest players, umpires, managers, executives, and pioneers in the sport of baseball. Only those who have demonstrated exceptional skill and contributions to the sport may be considered for election into the Hall of Fame.

There are two ways to ways to be elected into the HoF, either through the Baseball Writers' Associate of America (BBWAA) or the Era Committees. According to the National Baseball Hall of Fame's website, "The annual BBWAA election considers recently retired Major League players. The Era Committees consider retired Major League players no longer eligible for election by the BBWAA, along with managers, umpires and executives." As this project will examine

players, and not umpires, managers, or executives, the BBWAA election criteria will be referenced. As per the BBWAA's website, to be considered, candidates must meet the following eligibility criteria:

1. A baseball player must have been active as a player in the Major Leagues at some time during a period beginning twenty (20) years before and ending five (5) years prior to election.

2. Player must have played in each of ten (10) Major League championship seasons, some part of which must have been within the period described above.

3. Player shall have ceased to be an active player in the Major Leagues at least five (5) calendar years preceding the election but may be otherwise connected with baseball.

4. In case of the death of an active player or a player who has been retired for less than five (5) full years, a candidate who is otherwise eligible shall be eligible in the next regular election held at least six (6) months after the date of death or after the end of the five (5) year period, whichever occurs first.

5. Any player on Baseball's ineligible list shall not be an eligible candidate.

This analysis will primarily be concerned with the listed above points 1 and 3.

Analysis and Results

Initial Steps – Creating the Datasets

The first step was deciding which data to use, and where to obtain it. I used Stathead, which is a suite of advanced statistical tools provided by Baseball-Reference, a baseball statistics database. Using specific queries, I created four datasets:

- All Hall of Fame Hitters;
- All Hall of Fame Pitchers;
- All Non-Hall of Fame Hitters with at least 162 games played;
- All Non-Hall of Fame Pitchers with at least 32 games played.

The reason for filtering the non-hall of famers by games played was to (1) generally rule out opposite position players from the opposite dataset (such as having hitters who pitched a few innings in a pinch not be included in the pitchers dataset, etc.) and (2) remove players with a really small sample size of statistics, such as a hitter who only played a few games and has inflated numbers.

Overview

Once the datasets were created, a course of action was determined for the analysis. First, selected stats for the HoFers would be compared and visualized against the stats of the non-HoFers. This stats would then be plotted on histograms to compare the numbers between the two groups of players.

Secondly, the datasets would be split into a train and test dataset, and a confusion matrix would be used to predict outcomes for players based on the HoF threshold. Players who were predicted to make the HoF but actually did not, and players who were not predicted to make the HoF but actually did, resulted from the 0,1 and 1,0 quadrants of the confusion matrix.

Finally, the HoF threshold was applied to a dataset of players who have played in the past 5 years (and therefore were guaranteed to not be HoF players based on the eligibility requirements), to predict whether or not they may be elected into the HoF based on their current statistics.

Step 1: Comparison

The first step was to compare the stats of the HoF players to that of the non-HoF players. First was deciding which stats to compare. As hitters and pitchers have different stats to track, their stats will be compared separately. The stats selected to be compared are as follows:

For Hitters:

- BA
  - Batting Average; the percentage of time that a player gets on base by getting a hit.
- OBP
  - On-base Percentage; the percentage of time that a player gets on base by any means necessary, such as a hit, walk, or hit-by-pitch.
- SLG
  - Slugging percentage; measures a player's power hitting ability, it is the total number of bases a player has earned on a hit divided by there total number of at-bats.
- OPS
  - On-base plus Slugging; a players combined OBP and SLG, which gives a single measure of a players ability to get on base and hit for power.
- OPS+
  - Adjusted On-base plus Slugging; OPS normalized across the league to account for external factors such as different ballpark dimensions.

For Pitchers:

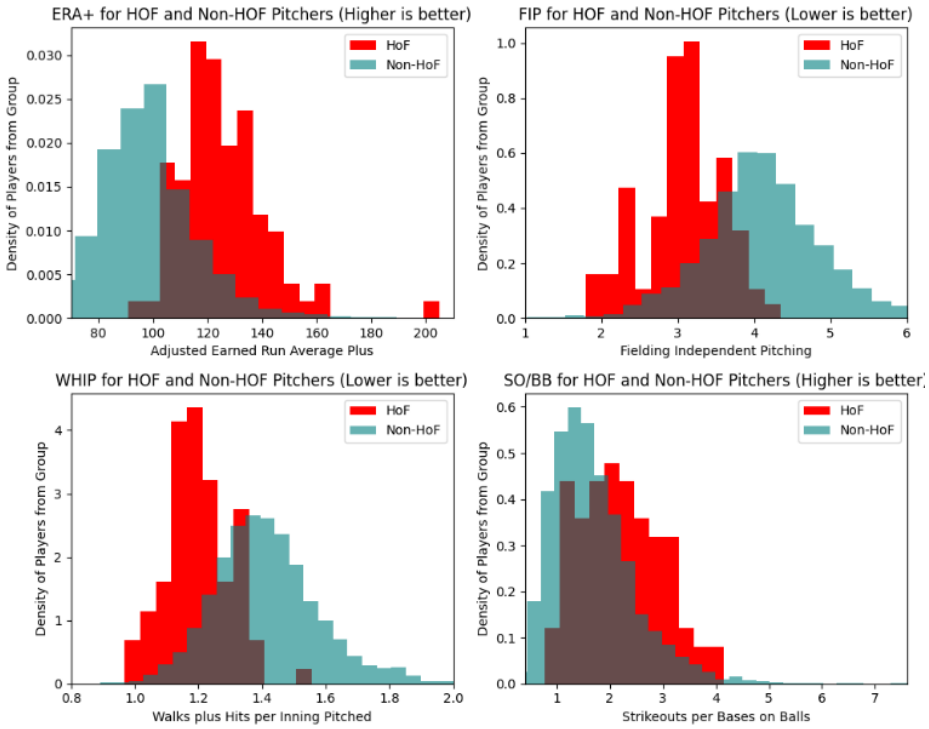
- ERA+
  - Adjusted Earned Run Average; a players Earned Run Average (ERA), which is the amount of runs that pitcher allows over nine innings, normalized across the league to account for external factors like ballparks and opponents.
- FIP
  - Fielding Independent Pitching; similar to ERA, but focuses solely on the events the pitcher has the most control over – strikeouts, walks, hit-by-pitches, and home runs. It removes results on balls hit into the field of play.
- WHIP
  - Walks and Hits per Inning Pitched; the sum of a pitchers walks and hits, divided by the number of innings pitched.
- SO/BB
  - Strikeout to Walk Ratio; sometimes abbreviated as K/BB, it's the number of strikeouts a pitcher records divided by the number of walks they allow.

The datasets for each of HoF Hitters and non-HoF Hitters, and HoF Pitchers and non-HoF Pitchers, were then imported and plotted against each other on two separated histograms on the same axis.

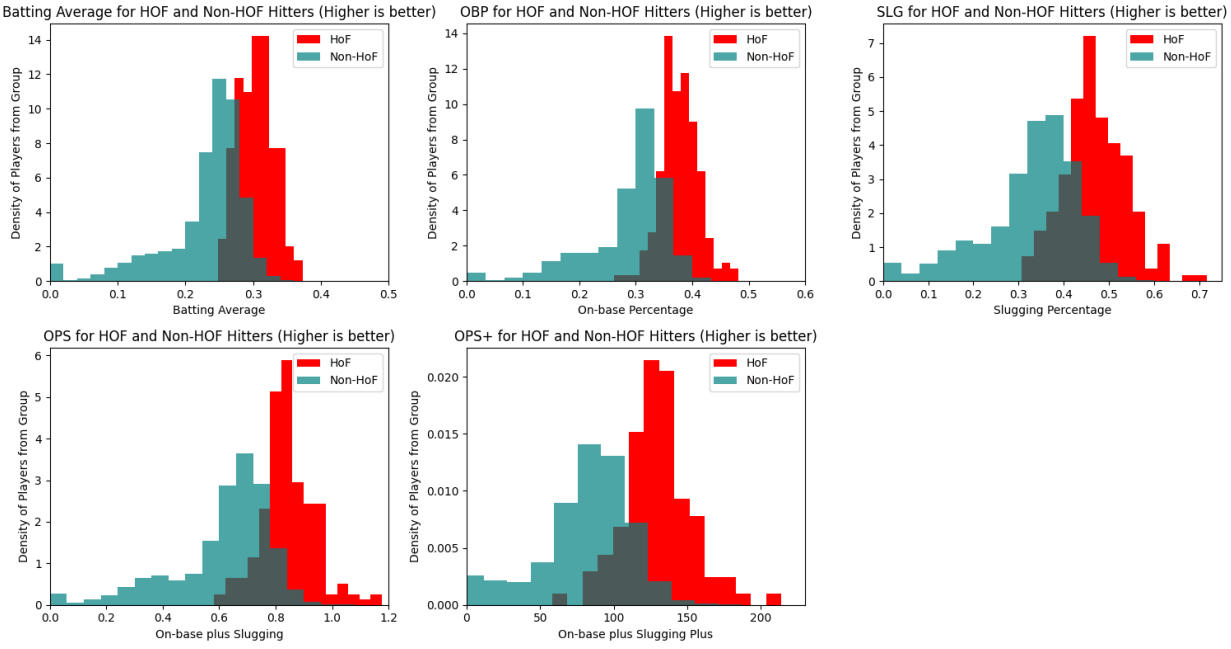
```
# BA for both overlayed
axs[0, 0].hist(p_hof_hitters_ba, bins=10, color='red', density=True, label='HoF')
axs[0, 0].hist(p_non_hof_hitters_ba, bins=50, color='teal', alpha=0.7, density=True, label='Non-HoF')
axs[0, 0].set_xlabel('Batting Average')
axs[0, 0].set_ylabel('Density of Players from Group')
axs[0, 0].set_title('Batting Average for HOF and Non-HOF Hitters (Higher is better)')
axs[0, 0].set_xlim(0,0.5)
axs[0, 0].legend()
```

The code used to create one of the comparison histograms.

The results, with the HoF players stats plotted in red and the non-HoF players plotted in teal, are as follows:



HoF and non-HoF Pitching stats compared.



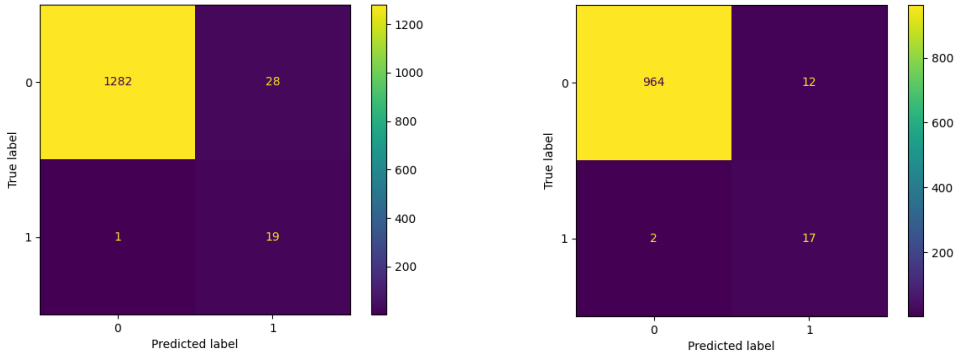
HoF and non-HoF Hitting stats compared.

Part 2: Prediction and Analysis

For the prediction and analysis, the datasets for HoF and non-HoF hitters, and for HoF and non-HoF pitchers, were combined. There was a column added to the data to denote whether a player had obtained HoF status or not. Next, the

data was separated into those who played prior to 2020, and those who played since 2020. This separation is to ensure that there is a dataset that has no HoF players, (the since-2020 dataset), which the model can be applied to to make blind predictions.

First, the pre-2020 dataset was split into a train and test dataset. The HoF players stats from the train dataset were analyzed and used to create a threshold for HoF entry using a support vector machine (SVM) model. Then it was applied to the test dataset to make predictions on which players should or shouldn't be in the HoF based on the threshold. These players were plotted on a confusion matrix, with top left quadrant being non-HoF players predicted to be non-HoF, the bottom right quadrant being HoF players predicted to be HoF, the top right quadrant being HoF players predicted to be non-HoF, and the bottom left quadrant being non-HoF players predicted to be HoF.



The resulting confusion matrices for hitters (left) and pitchers (right).

HoF hitters predicted to be non-HoF:

- Johnny Bench
- Roger Bresnahan
- Roy Campanella
- Jimmy Collins
- Earle Combs
- Buck Ewing
- George Kell
- Joe Mauer
- Kirby Puckett
- Ryne Sandberg
- Ray Schalk
- Ozzie Smith
- Hack Wilson

HoF pitchers predicted to be non-HoF:

- Ray Brown
- Jack Chesbro
- Waite Hoyt
- Bob Lemon
- Joe McGinnity
- Jack Morris
- Hilton Smith

Non-HoF hitters predicted to be HoF:

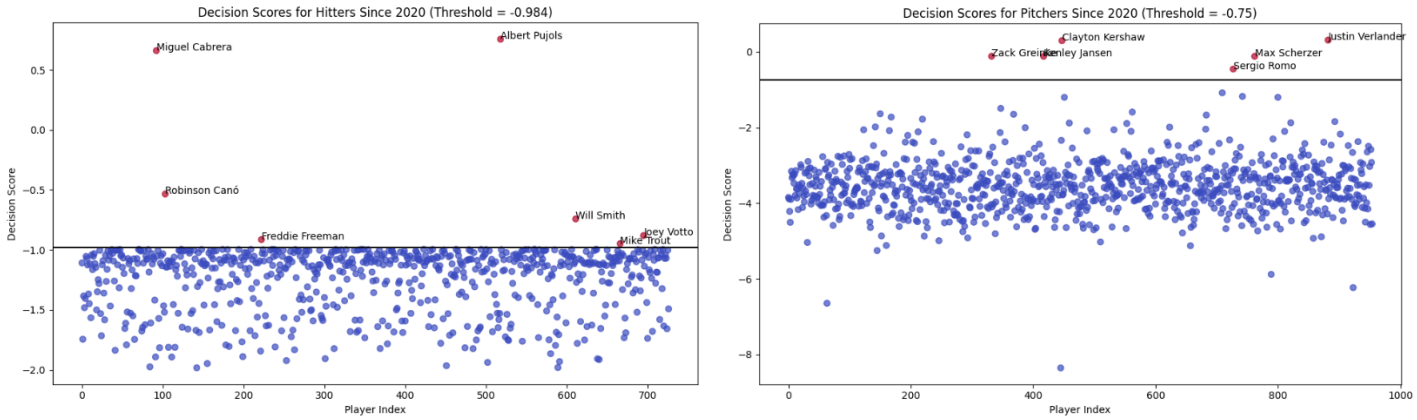
- George Van Haltren

Non-HoF pitchers predicted to be non-HoF:

- Tommy John
- Smokey Joe Wood

Note: Some HoF pitches were initially included in the HoF hitters group due to possessing hitting stats, and some HoF hitters were initially included in the HoF pitchers group due to possessing some pitching stats. These players were removed in the results.

Next, the SVM model was applied to the dataset for the since-2020 players. The model computed a decision score or each individual player. The decision score is the SVM model's confidence in a players HoF status based on their stats. Then a threshold was computed on the decision scores, and any players who exceeded the threshold were predicted future HoFers. The results were plotted for visualization purposes:



The resulting confusion matrices for hitters (left) and pitchers (right).

This resulted in the following players being predicted as future Hall of Famers:

Hitters:

- Miguel Cabrera
- Robinson Canó
- Freddie Freeman
- Albert Pujols
- Will Smith
- Mike Trout
- Joey Votto

Pitchers:

- Zack Greinke
- Kenley Jansen
- Clayton Kershaw
- Sergio Romo
- Max Scherzer
- Justin Verlander

Discussion and Conclusion

This analysis may be continued by applying similar methods to minor league players. HoF players that previously played in the minor leagues can have they minor stats examined and compared to their minor league peers who did not make the HoF. A baseline can then be established and then applied to current minor league players to predict if any of them are showing the stuff to be a future Hall of Famer.

The results of this project were interesting but there are some results that were unexpected, notably, some players that are generally agreed upon to be future HoFers that the model did not predict. This would be another good area for future examination to see why these players we or were not selected over others.

References

https://baseballhall.org/hall-of-fame/election-rules

https://bbwaa.com/hof-elec-req/

https://www.mlb.com/glossary/advanced-stats

https://stathead.com/baseball/player-batting-season-finder.cgi?request=1&match=player_season_combined&order_by_asc=1&order_by=name_display_csk&ccomp%5B1%5D=gt&cval%5B1%5D=162&cstat%5B1%5D=b_games&is_hof=N

https://stathead.com/baseball/player-pitching-season-finder.cgi?request=1&match=player_season_combined&order_by_asc=1&order_by=name_display_csk&ccomp%5B1%5D=gt&cval%5B1%5D=32&cstat%5B1%5D=p_g&is_hof=N

https://stathead.com/baseball/player-batting-season-finder.cgi?request=1&match=player_season_combined&order_by_asc=1&order_by=name_display_csk&is_hof=Y

https://stathead.com/baseball/player-pitching-season-finder.cgi?request=1&match=player_season_combined&order_by_asc=1&order_by=name_display_csk&is_hof=Y

Supplementary Materials

https://github.com/DannyDaye/isp/tree/main