Predicting Baseball Hall of Famers

MA 544 Final Project by Joel Simon

1. Abstract

In this project, I attempt to use 150 years of data to predict which current baseball players will end up in the Baseball Hall of Fame. Pitching and batting data ranging from 1871 to 2015 is used with kmeans to find clusters of good players. For each year, every player is given a score based on how close they were to award winning players. From there, logistic regression is used to do a final prediction as to whether or not a player should be in the hall of fame. After cross-validation, the best regression resulted in a recall of 65%, a precision of 69%, and an F1-score of 67%.

2. Introduction

In baseball, the most notable achievement is to be inducted into the Baseball Hall of Fame in Cooperstown, NY. Only about 1% of players are talented enough to reach that level. To get into the Hall of Fame, players must be retired for 5 years, and then be voted on by the members of the Baseball Writers' Association of America. In this project, I will be using multiple techniques to try to predict who should be in the Baseball Hall of Fame.

3. Completed Work

For this project, I was able to make a python program which takes in many years of data and eventually outputs whether a person should be in the Hall of Fame. This was done in two parts. The first is creating a score for each player. The second is using the scores to predict the outcome.

3.1 Data Processing

The data was downloaded off of Lahman's Baseball Database on data.world. From there, the important data I used was the pitching and batting data from the regular and post season. I also kept track of the award winners (Most Valuable Player, Gold Glove, etc.), All-Stars, and Hall of Famers. The pitching data had 25 features and the batting data had 17 features. Finally, I used the standard scaler on all the features to condense them to a gaussian distribution.

3.2 Scoring

To be able to effectively measure who should be in the hall of fame, I needed a way to score players based on the year. The reason it was done on a year-by-year basis is that tactics change drastically over time. For example, in the 1930s, it was very common for pitchers to throw a full game every 5 days. Now, that is extremely unlikely because of the increase in analytics use. The way I predicted if a player has done well is through Euclidean distance kmeans. By using the unsupervised model, I am able to cluster good players together. After clustering, I look at all of the award winners that year. All players in the same cluster as an award winner get a point.

Additionally, the award winner themselves get an extra point. Therefore, in the end, players near the best get rewarded. This process is then repeated for every year which results in a large, sparse matrix.

3.3 Prediction

After running through all of the data, the output is a 18847x145 integer matrix filled with mostly zeros. I employed the use of a CSR matrix to reduce the size and computation time. Then, I did an extra filtering of the data where I removed ineligible players who have not retired in the last 5 years because it is unknown whether they will be inducted or not. Additionally, I removed all the zero columns because I want the prediction to emphasize learning on players who may be good, and not those who were not close. From the remaining data, I split so that I had 30% testing data. I used logistic regression with an L2 penalty to predict whether a player was going to make it into the Hall of Fame.

4. Results

I tested the data many times, changing the number of clusters each time. I tried all the values between 4 clusters and 17 clusters to find the most optimal group. In every iteration, the accuracy was above 98%. However, that is because out of the up to 4000 testing points, only about 2% were positive. Therefore, I will look primarily at the precision, recall, and F1-score. From the best test, 14 clusters, I got a precision of 69%, a recall of 65%, and an F1-score of 67%. However, most of the time, the recall would hover around 50%. Additionally, there was a

tradeoff between precision and recall. Either way, in this case, I was able to correctly predict a Hall of Famer should be there about 65% of the time.

I am very content with this because there are many factors that go into whether a player should be in the hall. One example of this is Barry Bonds. He was an amazing baseball player in the late 1990's and early 2000's who broke many records and should be in the Hall of Fame. However, the reason he did so well was because he was taking performance enhancing drugs, so none of the writers wanted him to taint the reputation of the Hall of Fame. There are many examples from the steroids era and other times when players look like they should be inducted but are not.

5. Improvements

Even though I am content with the results, there are many ways that could be looked into to improve performance. The first is using principal component analysis on the features to reduce the dimensions and pick out the important features. Next, there are other methods that could be substituted for kmeans which could allow for better grouping. Because there were so many dimensions, it is not easy to see if the clustering worked or was effective. Furthermore, more thought could be put into how each element was scaled. This could potentially allow for closer relationships between better players. Finally, the last improvement that could be made is a better scoring system all together. The scoring system felt adequate, but was in no way scientific or mathematically backed. Therefore, a better scoring system could lead to better results.

6. Conclusion

Overall, I was able to predict a player is in the Hall of Fame 65% of the time. I used many techniques from the course including kmeans, scaling, logistic regression, confusion matrices, and sparse matrices to help me get to the final result. With a little bit of tweaking, this program can also be used to predict a current player's chances of making it into the Hall of Fame, and what they would need in the upcoming years to make it.

7. Resources

Gadoci, Brandon. "Lahmen's Baseball Database." 2016.

Simon, Joel. "Joelsimon2/MA544-Final." *GitHub*, 2023, github.com/joelsimon2/MA544-Final.

"Voting Rules History." Voting Rules History | Baseball Hall of Fame, baseballhall.org/hall-of-famers/rules/voting-rules history. Accessed 9 May 2023.