# Predicting Whether or Not a Major League Baseball Player Will be Inducted into the Hall of Fame

Elliott Evans
Northwestern University
elliottevans2015
@u.northwestern.edu

Jon Ford
Northwestern University
jonford
@u.northwestern.edu

Corey McMahon
Northwestern University
coreymcmahon2015
@u.northwestern.edu

## Keywords

PED: Performance Enhancing Drugs
WAR: Wins Above Replacement

## 1. WHAT LEARNING PROBLEM DO YOU WANT TO SOLVE?

We would like to build a system that, given the lifetime statistics of a Major League Baseball player, predicts whether or not that player will be inducted into the Hall of Fame. We will construct two separate models: a model for position players and a model for pitchers.

## 2. WHAT IS THE INTELLECTUAL INTEREST AND PRACTICAL UTILITY OF DOING THIS PROJECT?

Come election time for the Major League Baseball Hall of Fame, various players, commentators, and writers attempt to predict whether or not a given player will gain entry. An accurate regression model using the proper attributes to associate with each player would give both the casual fan and the knowledgeable player more confidence in Hall of Fame predictions.

## 3. DESCRIBE THE DATASET YOU WILL USE TO TEST AND TRAIN YOUR SYSTEM.

We will represent every player as a collection of attributes. Each attribute will be a career statistic of the respective player. Another attribute will be included to gauge whether or not a player has been involved in PED controversy of any kind. This attribute will default to 0 unless a player has been involved in PED controversy, in which case it will be set to 1. The classification for each player will be a boolean variable, which will be 1 if the player made the Hall of Fame and 0 otherwise. As a possible extension of the project, we may wish to include a player's variation in performance over his career as an attribute in the model (if a player is bad for a few years, maybe this affects their chances of being inducted).

### 3.1 Where will you get it?

We will utilize fangraphs.com as the source of our data. The site contains records of every player's relevant hitting, pitching, and defensive statistics.

### 3.2 Is the data labeled?

The data is already labeled, with the first row of the data being headers of each statistic. We will have to add in our own label for the PED information we plan on using.

### 3.3 How big is it?

We will have at our disposal the records of every player. This includes thousands of players. However, some of these records are unreliable or not complete. For instance, stolen bases is not something that was recorded in the early 1900's. However, we will ignore data that is unreliable by using records from more dependable years.

### 3.4 How is it encoded?

The data from fangraphs.com may be exported as a csv file. From there, we may alter it using various methods, such as adding a new PED attribute via Excel.

## 4. WHAT IS THE EXACT TASK THAT THE SYSTEM WILL LEARN TO DO?

We will predict the probability that a player is inducted into the Hall of Fame, given his lifetime statistics. We will utilize two linear regression models, one for position players and one for pitchers. The attributes for each player will include sabremetric-minded statistics (e.g. WAR), basic statistics (e.g. hits) defensive statistics, awards, and accolades. The model will output a real number between 0 and 1, thus giving us the probability of a player making the hall of fame. One issue we see here is that the linear discriminate will not necessarily output a number between 0 and 1 for all inputs. A possible solution to this is to develop some sort of mapping function from the linear regression output to the real numbers 0 to 1 (but that would be a difficult task in itself). An alternative would be to use a linear discriminant model, in which case we may need to map to a higher dimension so that training data is linearly separable.

# 5. WHAT ASPECT OF THE TASK IS YOUR SYSTEM GOING TO OPTIMIZE/ IMPROVE/ LEARN?

This is a prediction task. We will use a regression model, which will learn the probability of a player gaining entrance into the hall of fame. The model will learn from samples of players that are no longer playing today

# 6. WHAT MEASURE WILL YOU USE TO EVALUATE PERFORMANCE OF THE SYSTEM?

We will use n-fold cross validation (the value of n currently undetermined) to determine the best model to use. In the case of linear regression, we will measure error by the absolute differece between the binary classifier (1 if a player is in the Hall of Fame, 0 otherwise) with the model's output (a real number between 0 and 1). In the case of linear discriminant, we will measure error based on the number of misclassifications (possible with some different weighting for false positives versus false negatives). In addition, we will compare the predictions for training data players with the baseline approach using a weighted number of misclassifications approach.

# 7. WHAT IS THE BASELINE APPROACH YOU WILL COMPARE TO?

The baseline approach is the simplistic assumption that the only way for a player to get in is to have more career WAR that the average career WAR for players of his same position already in the Hall of Fame right now.

For example, the average career WAR among catchers in the Hall of Fame right now is 52.4, so our baseline approach assumes that a player is inducted to the Hall of Fame only if he has more than 52.4 career WAR.

# 8. WHAT SOFTWARE WILL YOU NEED TO WRITE?

We wil need to write MATLAB software to perform the linear regression/linear discriminant. We will also need to write software to digest the data from the .csv files.

## 8.1 Who will write it?

The writing involved in the project will be a shared effort between the three of us.

## 8.2 How long will it take?

Given the goals of the project, it will likely take approximately 20 hours of work to write the code.

# 9. DESCRIBE ANY EXISTING SOFTWARE PACKAGES YOU WILL USE.

MATLAB.

# 10. RELATED PAPERS TO READ.

Barry, D. and Hartigan, J.A. 1993. Choice models for predicting divisional winners in major league baseball. In *Journal of the American Statistical Association*. American Statistical Association. Alexandrian, VA, USA.
http://go.galegroup.com.turing.library.northwestern.edu/ps/retrieve.do?sgHitCountType=Nonesort=RELEVANCEinPS=trueprodId=AONEuserGroupName=northwesterntabID=T002searchId=R1resultListType=RESULT$_{L}IST content Segment = searchType = AdvancedSearchFormcurrent Position = 1contentSet = GALE\%7CA176779738docId = GALE|A176779738docType = GALErole = $.

Grossman, M., Kimsey, T., Moreen, J. and Owings, M. Steroids and Major League Baseball. University of California Berkeley. http://faculty.haas.berkeley.edu/rjmorgan/mba211/steroids%20and%20major%20league%20baseball.pdf

Mills, B., and Salaga, S. 2011. Using Tree Ensembles to Analyze National Baseball Hall of Fame Voting Patterns: An Application to Discrimination in BBWAA Voting. In *Journal of Quantitative Analysis in Sports.* American Statistical Association. Alexandria, VA, USA.
http://www.researchgate.net/publication/227378890_Using_Tree_Ensembles_to_Analyze_National_Baseball_Hall_of_Fame_Voting_Patterns_An_Application_to_Discrimination_in_BBWAA_Voting/file/d912f50b3c2fce3dd2.pdf

# 11. WHAT IS THE JOB OF EACH TEAM MEMBER?

Each member of the team will participate in every portion of the project.

# 12. MILESTONES

Software written to convert .csv formatted data into a format usable by MATLAB: Elliott - Nov 16
Any additional data (PED usage) added to model, any software needed to compute this data written: Corey - Nov 20
Initial linear regression or discriminant model completed: Jon - Nov 28
Experiments completed: Elliott - Dec 2
Initial results report: Corey - Dec 2
Initial website completed: Jon - Dec 2
Extended abstract completed: Corey - Dec 13
Project website completed: Jon - Dec 13
Poster completed: all 3 - Dec 13

## 12.1 Who is responsible for each milestone?

We will assign responsibility for tasks based on a rotation. Although one member will be responsible for ensuring that a task is complete, the other members will also be responsible for working together on all tasks.