Determining a Baseball Hall of Fame Candidate

Project Report

Kevin Braun
Univ. of Notre Dame
Computer Science and Engineering
Notre Dame, IN
kbraun@nd.edu

Blake Hartz
Univ. of Notre Dame
Mathematics
Notre Dame, IN
bhartz@nd.edu

John Leyhane
Univ. of Notre Dame
Computer Science and Engineering
Notre Dame, IN
fleyhane@nd.edu

Daniel McGee
Univ. of Notre Dame
Computer Science and Engineering
Notre Dame, IN
dmcgee@nd.edu

ABSTRACT

Baseball, more than any other sport, is a game of statistics. Applications of data mining to baseball and other sports are fairly new and we propose to advance these studies. This project seeks to determine what a "Hall of Fame baseball player" truly is based on statistical data, and predict which recent players are most likely to be inducted into the Baseball Hall of Fame. We will combine widely available baseball data with our domain knowledge of the game to create a classification scheme suitable for deployment.

1. INTRODUCTION

The game of baseball has always been regarded as "America's Pastime" and is one of the most popular and profitable sports in America. Teams are continually seeking new ways to gain a competitive advantage. Because there is an extensive amount of baseball statistical data available, it makes sense for baseball clubs to attempt to analyze this data to look for some pattern that will predict the success of a player. Many people would find it helpful to know the definition of a "Hall of Famer"; baseball clubs would find it especially attractive to have a future Hall of Fame player on their roster. The goal of our project is to predict whether a player will be inducted into the Hall of Fame based on statistics from the player's career, or at the least, see if this prediction is possible. Our challenge will be to examine the use of trees, rules, and other meta- and ensemble methods as techniques to sift through the large amount of data that will need significant preprocessing. These methods will then be applied to current and recently retired players to see which ones are likely to end up in the baseball Hall of Fame at the conclusion of their career. While this project may not yield results that are particularly important outside of professional baseball, we hope that it provides a somewhat useful model for describing Hall of Fame players which could be an objective piece of evidence amidst the flurry of opinionated sports media discussions that surround the Hall of Fame candidates annually.

Presented at:
MINE@ND Workshop for
Data Mining and Pattern Recognition
CSE 40647, Fall Semester 2006
December 7, 2006

2. RESOURCES

2.1 Previous Sports Data Research

2.1.1 Other Baseball Applications

Our approach towards searching for meaningful patterns in baseball statistical data is not unique. Notably, in-depth baseball data analysis has been used before by small-market baseball teams seeking players who are undervalued from an economic perspective. This process is described in a case study of the Oakland A's in the book *Moneyball* [4]. However, one major difference between our model and the *Moneyball* techniques is that our model will not be taking salary into consideration. This is not to say that we will be examining different statistics, as both our goals and those of the book seek to identify quality players. Our definition of "quality" varies somewhat from the economic approach of *Moneyball*. It would seem that Hall of Fame-caliber statistics would be a cut above those of a typical economically-undervalued player, although it is possible for a Hall of Fame-caliber player to be undervalued himself. Furthermore, the definition of "quality" described in *Moneyball* emphasized contributions to the success of the team; i.e. whether or not acquiring a player will result in more wins for the team. It is not likely that the properties so valued by the Oakland A's are of the same importance to Hall of Fame voters.

The subjects of *Moneyball* use statistics derived from what is known as "sabermetrics" which is a movement among baseball researchers to come up of with objective baseball knowledge based on statistical analyses. One of the pioneers of this field is Bill James, who developed many of the statistics of *Moneyball*. James has also considered the Hall of Fame, though, and he describes four models in his book on the subject [3]. One of these models is 'similarity scores' which attempts to argue for a player's induction based on his similarity to someone who is already a Hall of Famer. A similar concept could be achieved by use of nearest-neighbor algorithms. Another metric he uses is a system in which a player's career is valued by a set of weights applied to the number of seasons he has led the league in certain statistical categories. James's other two methods are similar concepts, in which a linear metric is developed based on the standard statistical categories. While these methods may be useful to understanding the practicalities of voting results, it is important to understand that these methods (and the corre-

Instance Type	Number of Instances
Batting	87,308
Pitching	36,898
Fielding	126,130
All Star	4,115
Hall Of Fame	3,369
Master	16,556

Table 1: Selected Row Counts of Baseball DataBank Tables; Master represents number of unique playerID's under consideration.

sponding weights) are not algorithmically generated. This is the major difference between our project and the work that James has done previously.

2.1.2 Data Quality

Another successful implementation of data mining in sports-related data is Advanced Scout [1]. Advanced Scout works with data from NBA basketball games and provides interesting patterns to NBA coaching staffs. These coaching staffs use Advanced Scout as a valuable part of their game preparation. The implementation required extensive use of data preprocessing and cleaning techniques, just as our project requires. The data preparation stage mainly uses consistency checks, which involve domain knowledge of how the game is played. This eliminates impossible situations that arise from conflicting data or erroneous entry. A similar system of consistency checks was considered for the Hall of Fame problem but was eventually abandoned after careful examination of the data.

2.2 Data Sources

We used data sources available from the Baseball DataBank [2]. These statistics are quite extensive, although there may be some deficiencies in the completeness of the records. Early statistics-keeping was not as consistent as it is today, so care must be taken in considering records from the early years of baseball. Table 1 shows the number of records available for use in our research and discovery. Although not overwhelming in size, this data covers nearly the full history of baseball from its beginnings in the late 1800's through the 2005 Major League Baseball season.

There are several recorded instances of players being banned from baseball for life which includes removing their eligibility for the Hall of Fame. We used an official list of these ineligible players from the research library of the Baseball Hall of Fame [7]. The players on this list must be considered as "noise" to our dataset, and the way our classification handles these points will greatly affect its accuracy.

3. UNDERSTANDING THE DATA

3.1 Organization

Baseball DataBank [2] organizes its data in an SQL database in which the various tables are organized by player and if necessary by year. The 'Master' table contains an entry for each person, using playerID as the primary key, that has played or participated in the game of baseball. The 'Batting' table contains rows of statistics keyed off of unique playerID's and yearID's of the season in which the statistics were recorded. The 'Pitching' table has a similar format. Other tables that apply to our research include 'HallOfFame', 'Allstar', and 'AwardsPlayers'.

As mentioned above, we were initially skeptical of the quality of the data and intended to implement a system of consistency checks similar to the NBA Scout [1] application. Fortunately, much of this work is incorporated into the raw Baseball DataBank data. Further efforts to this effect seem to be largely unnecessary as many of the missing values we thought would be problematic in the data are for features that we can confidently say should not be important features to the classification schemes. For example, the 'IBB' statistic ('intentional bases on balls') has only been recorded recently and is fairly indicative of a prolific power hitter. However, the fact that this statistic is not available for some of the classic power hitters who are undisputed Hall of Famers suggests that this statistic is not very useful to modeling. The current practice of intentionally walking power hitters was not common as recently as twenty years ago.

However, even the diligent and thorough work of the Baseball DataBank does not provide perfect data. This is largely due to historical problems of data completeness. For instance, the common batting statistic 'strike outs' ('SO' in the database) is not recorded for any players in the seasons 1897 to 1907.

3.2 Real-life Hall of Fame Selection Process

The Baseball Writers' Association of America (BBWAA) is responsible for selection of Hall of Famers by a voting process. The Hall of Fame selection process proceeds as follows [6]. A player is eligible to be nominated for the Hall of Fame if his career was ten or more seasons long and he has been retired for five years at the time of nomination. Each year, new players become eligible and a list of these players is given to a screening committee of the BBWAA. The screening committee selects which players will be presented on the final ballot that is sent to the full BBWAA. There have been a few exceptions to these eligibility requirements, but they have only resulted in two actual inductees.

Understanding the induction process has implications in how we approach our classification problem. The inclusion of players who have never been eligible for induction does not aid in creating models, so we have restricted our dataset to players who meet the career length requirements and who are not recently active. For our purposes, since our dataset is current through the 2005 season we have excluded players who have statistics recorded for the 2000 season or later.

Additionally, the fact that a player makes it past the screening committee and onto the full ballot is some indicator of that player's career quality, so we have approached the classification problem in two ways: as a two-class problem (inducted, not inducted) and as a three-class problem (inducted, on ballot, and all others).

Furthermore, as previously mentioned, there are the several recorded instances of players being banned from baseball for life which requires removing their eligibility for the Hall of Fame. Most of these ineligibilities are due to gambling; the well-known examples of this are players involved in fixing

the 1919 World Series (the 'Black Sox' scandal) and Pete Rose, who bet on baseball while he was a manager. Using the official list of these ineligible players obtained from the research library of the Baseball Hall of Fame [7], we removed these "noisy" items from our training data. It is worth noting that ineligible player Jean Dubuc was a Notre Dame student at one point!

The number of Hall of Famers is quite small in comparison to all the players who are present in the available statistics, but by employing the restrictions discussed above about eligibility (10 year career, 5 years out of baseball, and not ineligible), our dataset size is greatly reduced. The Hall of Fame class is definitely still the minority even though its relative abundance is higher in the restricted dataset. The sizes, after imposing these eligibility restrictions, of the various datasets are given in Table 2.

Dataset	Inductees	On Ballot	Not HoF
All Hitters	138	371	1047
All Pitchers	59	191	591
Post-War Hitters	63	227	697
Post-War Pitchers	27	134	451

Table 2: Number of Eligible Players in each dataset and class.

212 individuals have been inducted into the Hall of Fame as players. The datasets above account for 197 of these inductees. Two of the missing players, Lou Gehrig and Addie Joss, had medical conditions that shortened their otherwise excellent careers. Seeing as these are rarely occurring exceptions from the standard eligibility requirements, we have not changed our criteria to accommodate these players in the restricted dataset. Other players are omitted from the dataset because they have no recorded statistics. One example of this is Joe Williams, who played in the Negro Leagues in the early 1900's when statistics are not readily available.

Despite this discussion as to how the Hall of Fame induction process actually occurs, it is important to understand that this is not how the process has always worked and some, like James [3], would argue that some of the selections made prior to the current system are essentially mistakes. It is possible that our models could help to identify players who should not have been inducted statistically.

3.3 Splitting the Dataset

Our initial work focused on trying to classify hitters (non-pitchers) based on their batting statistics. However, the inclusion of pitchers (who are notoriously poor hitters) in batting datasets is essentially noise. Therefore, we needed to split the dataset based on the positions of players.

In order to split our dataset, we had to consider a few options. We wanted to include as many players as possible in each dataset, but we wanted to eliminate the chance of a Hall of Fame pitcher being included in the hitting dataset because that would negatively affect our classifier (even the best pitchers tend to have very poor hitting statistics). There is also the possibility that a Hall of Fame hitter be included in the pitching dataset if he had to pitch an inning one day when the team was desperate for a pitcher as can sometimes happen in extremely lopsided games. We also needed to consider Babe Ruth, who was a Hall of Fame hitter and also a very good pitcher for several years, as a special case to our splitting of the dataset. We decided to eliminate pitchers from the hitting dataset by ensuring that every player in the hitting dataset had played at least twenty games in at least one season at a fielding position. The number twenty was arbitrary, but we found that it did eliminate the pitchers from the dataset without removing any hitters. To eliminate hitters from the pitching dataset, we made sure that all the members of the pitching dataset had pitched in at least three games in a single season at some point in their career. Hitters rarely pitch in multiple games. In most cases, this happens once or twice in a career whereas all pitchers who have played ten years are guaranteed to have pitched in three games in most of those years.

3.4 Feature Creation

Our first attempts at classifying players were based on career statistics. We decided to attack the problem of classifying the hitters first because they account for 150 of the 260 Hall of Fame inductees. These were not difficult to obtain because they are just the sum over years of a career of all the various statistical categories. However, the results using this dataset are not very impressive (described fully in "Experiments" section) so we needed to seek out combinations of features that are more useful to discriminate between Hall of Fame-caliber players and average ones.

After several runs with many different classifiers, we saw that none of the classifiers worked very well with our starting feature set. We deduced that this could be a result of two factors: we had not included many statistics related to overall career performance, and statistics from different eras could have different meanings. For example, Hall of Famer Joe Kelley hit .321 with 65 home runs, 1194 RBI and 2245 hits in his career from 1891 to 1908. In the past ten years, 65 home runs or more have been hit by a player in a single season four times. Because of this, we decided to make this a consideration for our dataset.

The calculation of certain common batting statistics, like slugging percentage (SLG) and on-base percentage (OBP) and their sum (OPS), were some early additions to the feature set. OBP is one of the key statistics used in Moneyball [4]. We have also tried to quantify the average hitting production based on career length. For example, a ratio given by

$$HitAverage = \frac{\sum_{career} Hits}{\sum_{career} Games}.$$

was used to try and reduce the effect of career length on the hits statistic. Similar ratios were used with other batting statistics. However, these new features did not do much better than the career total statistics.

We next approached the angle of features based on a player's era. This would allow us to address the issue of players simply outperforming their peers to gain lasting recogni-

tion. We developed a statistic that compares a given player's statistics to the average statistics of all players who played in his career span. The Hall of Famer should be able to separate himself from the field using these ratios. An example of this new kind of statistic is given by

$$HomeRunRate(player) = \frac{\frac{\sum_{career} HomeRuns}{\sum_{career} AtBats}}{\frac{\sum_{career}\sum_{allplayers} HomeRuns}{\sum_{career}\sum_{allplayers} AtBats}}.$$

Descriptions of the feature sets for both hitters and pitchers are available in Appendix B. The most distinguishing attributes, as determined by GainRatio, are in **boldface** in Tables 5 and 6.

4. EXPERIMENTS

4.1 Setup

4.1.1 Initial Experiments

Our first major experiment was completed using Weka Experimenter. We tested four different hitter datasets: 2-class simplified, 2-class expanded, 3-class simplified, and 3-class expanded. The simplified feature set has 20 features, while the expanded feature set (which includes comparisons of players to others of their era) has 37. For this initial experiment, we did not use feature selection beyond what is included in the algorithms. The experiment used twelve different learning algorithms to classify these datasets, and the results were mixed.

Our second major experiment used the above explained hitters datasets and added an additional two datasets for pitchers: 2-class and 3-class. Both pitchers datasets have 28 features. We also revised our set of algorithms to include some with and without feature selection, and eliminated a few algorithms that performed poorly in our first experiment. This round used fifteen different learning algorithms. Once again, the results were mixed.

4.1.2 Further Experiments

Even though the results of the initial experiments were not very good, the wide variety of classifiers and options we had chosen enabled us to narrow our search in finding an acceptable model. After the initial experiments, two techniques stood out as significantly improving performance. Not surprisingly, these methods were ensemble techniques and feature selection.

The final experiments tested a number of different meta-learners that used various combinations of the standard basic classifiers. The value of different feature selection methods (CfsSubset, InfoGain, and GainRatio) were evaluated both as preprocessing steps to single classifiers and within the decision structures of meta-learners.

Another variable we changed in these later Experimenter runs was the temporal range of the dataset. The motivation for this is that baseball was played very differently in its earlier history, which results in the accumulation of different sets of statistics that measure performance. There are a number of possible historical eras that the data can be restricted to but we have chosen to examine one, the post-war era (1946-present). This selection is based largely on domain knowledge about trends in the way games were played and the fact that we retain 54 years of data.

4.2 Evaluation

Evaluation of models for this problem was not easy because of the extremely unbalanced class distribution. We needed to consider what performance indicators would be most valuable. Since there are so few Hall of Fame players, we decided that we wanted our models to include as many Hall of Famers as possible. Therefore, our evaluation scheme focused on reducing the number of false negatives (Hall of Famers classified as not inducted). The most natural statistic to use for this is recall.

However, this is not to say that our only goal was reducing false negatives. Our ideal classifier would also have a reasonably low number of false positives. We say 'reasonably low' because no classification scheme will be perfect and we expect there to be some players who were not popular enough to be voted into the Hall of Fame, despite having statistical qualifications to be Hall of Famers.

We used an evaluation measure that combines both precision and recall in order to reduce both components of error. The natural choice for this is the f-measure. For our application, we want want to weight the f-measure to favor recall. The evaluation parameter can be described as F_α, which is defined as

$$F_\alpha = \frac{(1+\alpha)PR}{(\alpha P) + R}$$

in the standard literature where P is precision and R is recall. We used $\alpha = 1, 2$, and 3 to examine the effect of increasing the weight of R and help to correctly classify known Hall of Famers.

4.3 Results

Our initial results were promising but not spectacular. The best learning algorithms from early experiments were AdaBoost with JRip and RandomForest. Surprisingly, a Naive Bayes classifier using attribute selection also performs to the level of the other two classifiers when $\alpha = 3$. However, this is due to Bayesian classifiers having a tendency to produce high recall at the cost of low precision. This explains why Bayes did not perform well with a lower α value.

Our further experiments used meta-learners and feature selection to improve performance. Performance of all classifiers on the 3-class problem was quite poor (hundreds of misclassified instances) so we abandoned that line of inquiry to focus on producing an acceptable model The performance of the various models we tried is included in Table 3. We used f-measure with $\alpha = 1$ and $\alpha = 3$ to evaluate classification models. $\alpha = 3$ was especially helpful because it distinguished between classifiers that were similar at lower values of α. The table only includes results of experiments on the post-war hitting dataset because the post-war data provided

more accurate models and the hitting dataset provided more training instances than that of the pitching dataset.

Classifier	F_1	F_3
JRip	0.73	0.72
RandomForest	0.78	0.75
AdaBoost(JRip)	0.78	0.75
GainRatio(RandomForest)	0.80	0.76
GainRatio(Vote(RandomForest*5))	0.79	0.76

Table 3: Performance of various classifiers on postwar hitting dataset throughout experimental phase using $\alpha = 1$ and $\alpha = 3$.

The final model selected for deployment is a voting method among five classifiers. These five classifiers are all of type RandomForest with different values used for the random seed. We used gain ratio with a threshold value of 0.07 to select attributes before building this model. An interesting fact about this model is if the instance order is randomized before model compilation, it is slightly different each time. This is due to the RandomForest being quite random in its feature selection. However, using five RandomForests in a voting situation helps to smooth possible shortcomings of the classifiers. Although Table 3 shows good performance for a single RandomForest after using gain ratio for attribute selection, this effectiveness is not guaranteed with every run of the classifier as randomness plays a large role in the performance; i.e. a single RandomForest was not stable across builds with different seeds.

5. DEPLOYMENT

5.1 2007 Hall of Fame Ballot

The Hall of Fame ballot for 2007 was released in late November 2006 [5]. The voting algorithm we decided upon for our model yielded the following predictions about the 2007 Hall of Fame Ballot. Cal Ripken, Jr., Tony Gwynn, Mark McGwire, and Harold Baines are the predicted Hall of Famers. All of the players are hitters; the pitching prediction models does not project that any of the pitchers currently on the ballot will be inducted. It is important to remember that this prediction is only for eventual induction, not necessarily that these players will be voted in this year.

Some of the predicted inductees have non-statistical issues that could hinder or help their induction chances in the upcoming election. Mark McGwire has been linked to steroid scandals. This decreases his actual chance of induction because the actual voting is based on human opinion, even though the model outputs McGwire's probability at 0.82.

The fact that the model predicts that Harold Baines as a Hall of Fame is evidence of one limitation to our dataset, namely the feature of fielding position. Baines' position is designated hitter, a player in the American League who only bats (and doesn't play defense in the field); there are zero designated hitters in the Hall of Fame. The inclusion of position in our dataset could potentially improve our models significantly. Further discussion on positional analysis can be found in the 'Conclusions' section of this report.

A full list of the probabilistic predictions for the 2007 Hall of Fame ballot is included in Appendix A. It should be noted that our model outputs a probability of 0.52 for Harold Baines' induction, very close to the cutoff, which indicates his chances are less than those of all other probable inductees. If the probability threshold is raised to a higher value, it may help better reflect the fact that being inducted into the Hall of Fame takes far more than a majority vote of the writers (75% of the vote is required to be inducted. If the value is raised to 0.60, only three players make the cut: Gwynn, Ripken, and McGwire, who have clearly superior statistics.

5.2 Current Players

We did not evaluate current players in the final model because their career total statistics are unavailable. It is quite possible, however, to compute their likelihood of induction given that 2005 (the upper limit of the dataset) was the last year of their career, i.e. a question like "If he were to be considered right now, what are his chances of being inducted?" is feasible in the current implementation of our system. We did not develop a system to project career total statistics for players who have not played 10 years or have not likely played most of their careers.

6. CONCLUSIONS

Our work has shown us that determining a baseball Hall of Fame candidate is no easy task by the numbers alone, because the ultimate selection is done by humans who look at more than just statistics. It is this subjective human bias that allows Hall of Fame balloting to be so hotly debated for years before and after a selection. Pure statistical data has a hard time capturing this unknown feature. However, after analysing some of the false negatives, it is clear there are some features that could be added to improve the performance of the classifier because many of these left out Hall of Famers have a lot in common.

There were several interesting things that we learned from analyzing our false values. The average run produced around 25-30 false values. We analyzed the results of each run and tried to note similarities among the false values. Nearly all the players that were coming up as false negatives were always the same and many of the false positives were also the same. Since the runs were very similar, a closer look at one run will shed light on the reasons for the false values. One particular run had 46 true positives, 9 false positives, and 17 false negatives. Of the 17 false negatives, 15 began their careers before 1958 and 13 began their careers before 1948. This might suggest that maybe we should have used only players that did not play at all before the postwar era instead of players that played some in the postwar era.

Upon further analysis, the two players that were false negatives that began their careers after 1958 were Gary Carter and Tony Perez. One would expect that a good classifier would pick Perez, although Perez's stock was raised in the minds of voters because of his tremendous postseason success as a member of the 1970s Cincinnati Reds. Adding postseason stats to the feature set might improve the classifier. Carter on the other hand has decent hitting numbers, but his stock was raised because he played catcher. He won several silver slugger awards as the best hitting catcher of his time. He also won three gold gloves and was considered a tremendous defensive player. Both positional analysis and

defensive analysis would improve could improve our classifier. We did not have any fielding features or positional features. Yogi Berra was another catcher included in the false negatives. Players that reached the Hall of Fame based on their fielding statistics would not be included in our model. This was evident with other false negatives Bill Mazeroski, Phil Rizzuto, and Richie Ashburn. Mazeroski also would benefit from postseason analysis. His World Series winning homerun in 1960 is one of the most famous homeruns of all time. It is ironic that he is remembered for a homerun when he won 8 gold gloves and is considered a Hall of Fame player for his defense.

In conclusion, it was difficult to create a perfectly accurate classifier for a wide variety of reasons. Since players are voted in by writers, there is not any objective criteria for voting. However, one would hope that there is some commonality that an algorithm could find among Hall of Fame players. The classifier actually was fairly effective in classifying players. When you consider all the true negatives, it was around 97could be improved by including other features that took into account positions, fielding statistics, and postseason performance. All things considered, our classifier is effective in finding some objectivity in a very subjective process.

7. REFERENCES

[1] I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pratap, and K. Ramanujam. Advanced scout: Data mining and knowledge discovery in NBA data. *Data Mining and Knowledge Discovery*, 1(1):121–125, Nov. 1997.

[2] S. Forman. Baseball-DataBank.org. `http://www.baseball-databank.org`, Aug. 2005.

[3] B. James. *The Politics of Glory*. Macmillan Publishing Company, New York, 1994.

[4] M. Lewis. *Moneyball: The Art of Winning an Unfair Game*. W.M. Norton & Company, Inc., New York, 2003.

[5] National Baseball Hall of Fame. 2007 BBWAA hall of fame ballot released. `http://www.baseballhalloffame.org/news/2006/061127.htm`, Nov. 2006.

[6] National Baseball Hall of Fame. History of BBWAA hall of fame voting. `http://www.baseballhalloffame.org/history/hof_voting/default.htm`, Jan. 2006.

[7] National Baseball Hall of Fame. Major League Baseball list of ineligible players. Received through email correspondence, Nov. 2006.

APPENDIX

A. BBWAA 2007 HALL OF FAME BALLOT

The following players (Table 4) are listed on the 2007 ballot as eligible for induction [5] along with their probability of induction, as determined by our deployed model.

Name	Probability
Harold Baines	**0.52**
Albert Belle	0.16
Dante Bichette	0.00
Bert Blyleven	0.26
Bobby Bonilla	0.04
Scott Brosius	0.00
Jay Buhner	0.00
Ken Caminiti	0.00
Jose Canseco	0.38
Dave Concepcion	0.10
Eric Davis	0.04
Andre Dawson	0.20
Tony Fernandez	0.02
Steve Garvey	0.24
Rich Gossage	0.12
Tony Gwynn	**0.94**
Orel Hershiser	0.04
Tommy John	0.08
Wally Joyner	0.00
Don Mattingly	0.16
Mark McGwire	**0.82**
Jack Morris	0.10
Dale Murphy	0.02
Paul O'Neill	0.04
Dave Parker	0.26
Jim Rice	0.148
Cal Ripken Jr.	**0.72**
Bret Saberhagen	0.22
Lee Smith	0.04
Alan Trammell	0.00
Devon White	0.02
Bobby Witt	0.00

Table 4: Probability of Induction. Predicted inductees are in boldface.

B. ATTRIBUTES USED

Table 5 lists the attributes used in the hitting dataset. Features with suffix 'r' are a ratio of a player's statistic to the average statistic during his career. See the subsection 'Feature Creation' for more information. Similarly, Table 6 lists the attributes of our dataset for pitchers.

The possible class values for both sets are Y (inducted), B (on the ballot), and N (neither Y nor B). These three classes could also be represented in a two class system by merging classes B and N. Attributes used for classification were selected using GainRatio with a cutoff of 0.07 and are printed in **boldface**.

C. MLB LIST OF INELIGIBLE PLAYERS

The following players are ineligible for induction into the Baseball Hall of Fame and were removed from our dataset [7]: Eddie Cicotte, William Cox, Cozy Dolan, Phil Douglas,

Feature	Description
FY	First year in professional baseball
LY	Last year in professional baseball
Y	Seasons played
ASY	**All-Star years**
AWD	**Number of season awards received**
G	**Games played**
BA	**Career Batting Average**
SLG	**Slugging Percentage**
OBP	On-Base Percentage
OPS	**On-Base Plus Slugging (SLG+OBP)**
R	**Runs Scored**
H	**Hits**
EB	**Extra-Base Hits (2B+3B+HR)**
2B	**Doubles**
3B	**Triples**
HR	**Home Runs**
RBI	**Runs Batted In**
SB	Stolen Bases
BB	**Bases on Balls**
SO	**Strikeouts**
HBP	Hit By Pitch
SH	Sacrifice Hits
SF	**Sacrifice Fly**
BAr	BattingAverage Rate
SLGr	**Slugging Percentage Rate**
OBPr	**On-Base Percentage Rate**
OPSr	**OPS Rate**
Rr	**Run Rate**
Hr	Hit Rate
EBr	**Extra-Base Hit Rate**
2Br	Double Rate
3Br	Triple Rate
HRr	Home Run Rate
RBIr	**Runs Batted In Rate**
SBr	Stolen Base Rate
SOr	Strike Out Rate

Table 5: Hitters Feature Description

Jean Dubuc, Happy Felsch, Ray Fisher, Chick Gandil, Joe Gedeon, Benny Kauff, Joe Jackson, Fred McMullin, Lee Magee, Jimmy O'Connell, Gene Paulette, Swede Risberg, Pete Rose, Buck Weaver, Lefty Williams, Heinie Zimmerman.

Feature	Description
FY	First year in professional baseball
LY	Last year in professional baseball
Y	**Seasons played**
ASY	**All-Star years**
AWD	**Number of season awards received**
G	Games played
ERA	**Earned Run Average (ER*9)/IP**
BAOpp	Opponents' Batting Average
WHIP	**Walks+Hits / IP (BB+H)/IP**
GGper	**"Good Games" % = (W+SV)/G**
W	**Wins**
L	**Losses**
GS	**Games Started**
CG	**Complete Games**
SHO	**Shutouts**
SV	Saves
IPOuts	**3*Innings Pitched**
H	**Hits Allowed**
ER	**Earned Runs Allowed**
HR	**Home Runs Allowed**
BB	**Bases on Balls**
SO	**Strikeouts**
BFP	**Batters Faced by Pitchers**
R	**Runs Allowed**
ERAr	**Earned Run Average Rate**
BAOppr	Opponents' Batting Average Rate
WHIPr	**WHIP Rate**

Table 6: Pitchers Feature Description