MLB Hall of Fame Predictor

Jakub Jezusek
Jack Lowrie
Tait Murphy
Jeremy Piech
Christopher Pierce

Problem

Every year, members of the Baseball Writers' Association of America (BBWAA) vote on who is to be elected into the Baseball Hall of Fame, which is the highest honor in the game. A player must receive 75% of the vote to be elected. In addition to being extremely competitive, the voting process can be highly subjective, making it the subject of much debate and speculation among baseball fans and sports writers. That being said, a trained algorithm might be able to expose trends that can correctly predict a player's chance to be inducted into the Hall of Fame after retirement. While a Hall of Fame predictor would be primarily usable by baseball fans, it can also be useful for players themselves, by giving them a means to compare their stats with those of players that have been inducted.

Data Set

We generated our data set by pulling data from baseball-reference.com, and was comprised of all players in the Hall of Fame, and all currently retired non-inductees who were active at any time between 1961-2005. We also required a minimum of 162 games played for position players and 100 innings pitched for pitchers, which we determined to be logical cutoffs. After cleaning, our data set contained 1725 batters, 128 of which were inducted, and 1327 pitchers, 60 of which were inducted. We chose to focus solely on career statistics, and therefore removed any player implicated in steroid use from the non-Hall of Fame training set, because their stats would skew the training data.

Because the stats differ for batters and pitchers, we separated the pitchers from the batters and tested on each subset independently. For pitchers, we chose the following stats, strikeouts, wins above replacement (WAR), games played, complete games (CG), shutouts, wins, losses, win-loss percentage, saves, innings pitched, earned run average (ERA), and fielding independent pitching (FIP). For batters, the features we chose were home runs, WAR, games played, runs, hits, runs batted in (RBI), walks, stolen bases, batting average, and on-base percentage plus slugging percentage (OPS).

Testing

Because the stats used to evaluate batters and pitchers are different, we trained the same models on each data set separately. Initially, it looked as though a decision tree would be the most reliable model, as it was able to classify players with 96.17% accuracy on our batter data set. Interestingly, the number of times a player was intentionally walked was a strong indicator of being inducted for these initial tests, though it was not reliable enough to consider in isolation, and we decided to remove this feature from our data set. After cleaning and expanding our data set, we compared our initial findings with bayes nets, logistic regression, multilayer perceptron, and k-nearest neighbor models. Not surprisingly, the best model for one dataset was not the the best model for the other.

Results

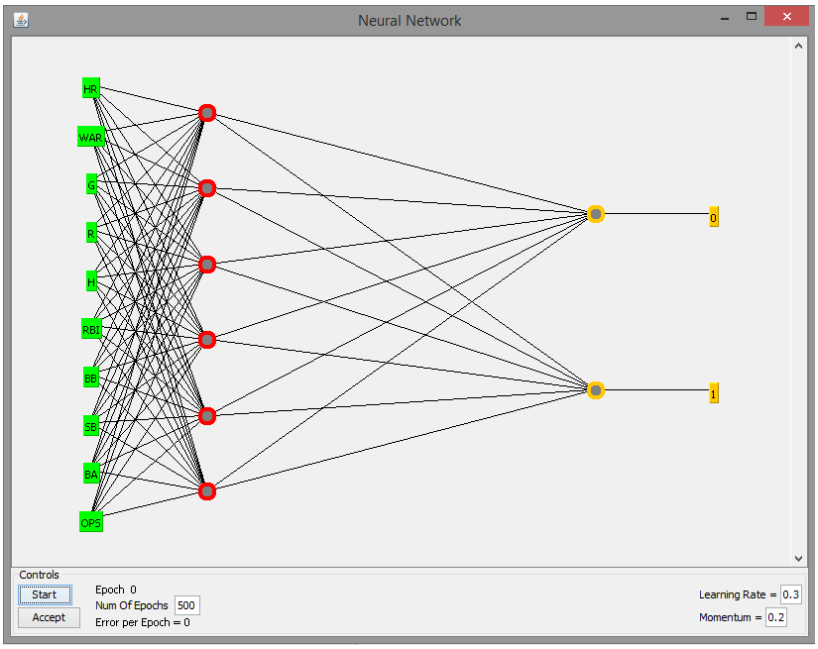
	J48(Decision Tree)	BayesNet	NaiveBayes	Logistic	Multi-Layer Perceptron	Ibk(nearest neighbor)
Accuracy-HoF	0.648	0.922	0.914	0.734	0.805	0.758
Accuracy- Non HoF	0.984	0.914	0.914	0.988	0.988	0.987
Precision - Hof	0.761	0.461	0.459	0.832	0.844	0.829
Precision - Non Hof	0.972	0.993	0.993	0.979	0.984	0.981
Recall - Hof	0.648	0.922	0.919	0.734	0.805	0.758
Recall - Non Hof	0.984	0.912	0.914	0.988	0.988	0.987
F-Measure - Hof	0.7	0.615	0.611	0.78	0.824	0.792
F-Measure - Non Hof	0.978	0.952	0.951	0.983	0.986	0.984
Accuracy	0.959	0.914	0.914	0.969	0.974	0.97
Precision	0.957	0.954	0.953	0.968	0.974	0.969
Recall	0.959	0.914	0.914	0.969	0.974	0.97
F-Measure	0.957	0.927	0.926	0.968	0.974	0.97

Model performance for batters

	J48(Decision Tree)	BayesNet	NaiveBayes	Logistic	Multi-Layer Perceptron	Ibk(nearest neighbor)
Accuracy-HoF	0.767	0.95	0.983	0.883	0.867	0.85
Accuracy- Non HoF	0.994	0.965	0.949	0.993	0.993	0.996
Precision - Hof	0.868	0.564	0.48	0.855	0.852	0.911
Precision - Non Hof	0.989	0.998	0.999	0.994	0.994	0.993
Recall - Hof	0.767	0.95	0.983	0.883	0.867	0.85
Recall - Non Hof	0.994	0.965	0.949	0.994	0.993	0.996
F-Measure - Hof	0.814	0.708	0.645	0.869	0.86	0.879
F-Measure - Non Hof	0.992	0.981	0.974	0.994	0.993	0.994
Accuracy	0.984	0.965	0.951	0.988	0.987	0.989
Precision	0.984	0.978	0.976	0.988	0.987	0.989
Recall	0.984	0.965	0.951	0.988	0.987	0.989
F-Measure	0.984	0.969	0.959	0.988	0.987	0.989

Model performance for pitchers

Based on the performance statistics, the best model for classifying batters was given by the multi-layer perceptron. It outperformed the other models in almost every respect. MLP's f-measure for hall of fame players was significantly better than the f-measure of any other classification methods, signifying that this model, while also the most accurate on non Hall of Fame players, represents the best model for correctly classifying Hall of Famers, which is the crux of the problem we are attempting to address.



Trained MLP Model for Batters

The multilayer perceptron model did not perform as well for pitchers as k-nearest neighbor, though the best model was not as immediately apparent. The f-measures of k-nearest neighbor for the different categories were all strictly better than any other classifier tested. Although its precision and recall values were not always the highest compared to its closest competitors, the multi-layer perceptron and logistic regression models, its f-measure was higher. This indicates a better average case performance for the k-nearest neighbor model on the pitcher data set.

Predictions

Running our final models for the 2016 ballot, our predictor indicates the following players will be inducted into the Hall of Fame:

Batters				Pitchers	
Name	**HoF?**	**Name**	**HoF?**	**Name**	**HoF?**
Ken Griffey	Yes	Larry Walker	Yes	Trevor Hoffman	No
Mike Piazza	Yes	Mark McGwire	Yes	Curt Schilling	No
Jeff Bagwell	Yes	Gary Sheffield	Yes	Roger Clemens	Yes
Tim Raines	No	Sammy Sosa	No	Mike Mussina	No
Barry Bonds	Yes	Jim Edmonds	Yes	Lee Smith	No
Edgar Martinez	No	Nomar Garciaparra	No	Billy Wagner	No
Alan Trammell	No	Mike Sweeney	No	Mike Hampton	No
Fred McGriff	No	David Eckstein	No		
Jeff Kent	Yes	Jason Kendall	No		
Mark Grudzielanek	No	Garret Anderson	No		
Brad Ausmus	No				

Conclusions

WAR appears to be the most significant indicator of a player's chances for being inducted: it is the most heavily weighted attribute for batter multilayer perceptron model and although the k-nearest neighbor model is more opaque, the decision tree algorithm when run on both data sets always identified WAR as the first attribute to split on. That being said, WAR is not as reliable in isolation: running our predictors with only WAR as an attribute, the f-measure for HoF batters was 70.3%, and the f-measure for pitchers is 64.3%. This means, with WAR as the baseline, our predictor is 12.1% more accurate for batters and 23.6% more accurate for pitchers.

Overall the pitcher models were usually more accurate than the models for the batters. For example, it classified Roger Clemens as a Hall of Famer, who based on the merits of his career statistics alone is widely thought to merit a spot in the Hall. Similar to the case of Roger Clemens, several batters classified as Hall of Famers would get in based on the merit of their statistics if not for their alleged steroid use. That being said, our predictor also classified several ‘clean’ batters, including Jeff Bagwell, as being inducted soon, when in reality they fell short of the 75% needed on this year's actual ballot.

Further work

Though our model is able to predict induction into the hall of fame with reliable accuracy, we realized during testing that steroid use was significantly affecting the performance of the models. We traded steroid-use and alleged use as an attribute for the ability to fine-tune our predictor on player’s career statistics, though steroids do have a drastic effect on a player’s chance of being inducted -- therefore positive tests for any PED as well as any allegations of foul play could be taken into account for greater accuracy. The need to account for PEDs is especially apparent in our predictions for the 2016 ballot: Barry Bonds, Mark McGwire, and Roger Clemens all make it in according to our predictor, though none of them will be inducted due to their steroid use. Furthermore, though our predictor’s accuracy makes it useful for sports writers and baseball fans alike, our results are useful to players in the most high-level sense: WAR as composite statistic is difficult to directly influence through training.