UNIVERSITY OF NEW SOUTH WALES, SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Predicting Basketball Hall of Fame Membership

COMP9417 - Machine Learning and Data Mining: Project

Chi-Chun Tiao

June 9, 2013

1 INTRODUCTION

Enshrinement into the Naismith Memorial Basketball Hall of Fame is an honour bestowed upon exceptional basketball players, coaches, referees, executives and other major contributors to the sport across the world. As one might expect, the population of inductees is dominated by members in the players category - 158 members in the players category out of a total of 326 members as of the enshrinement class of 2013[1]. Of the 158 players in the Hall of Fame, 103 played in the National Basketball Association (NBA), easily rendering it the league with the most players in the Hall of Fame[1]. As such, we shall limit our discussion to individuals who have played in the NBA. [2]

Unlike other major sport's Hall of Fames where sportswriter voters openly debate their choices, the Naismith Hall of Fame has an opaque and closely held selection process[2]. As a consequence of the screening committee's lack of transparency, the criteria for induction is vague at best and discussion regarding the merits of players is subjective. It would be useful to form an objective basis upon which a player's merits can be analysed and used to predict whether he will be inducted into the Hall of Fame. The goal of this project is to develop such a method using machine learning techniques.

[1] Other inductees of this category include college, international and female players, etc., most of whom have their own shrines already: The College Experience (NCAA), the FIBA Hall (international) and the WBHOF (women's).

[2] Indeed, this inevitably places prejudice against players who had decorated international or college careers, but relatively short or less impressive stints in the NBA, but might nevertheless be considered worthy of Hall of Fame induction.

Currently, some popular methods used include the *Keltner List*, a less objective but systematic method adopted from baseball (as many advanced basketball metrics often are) and also a logistic regression model[4], developed by *Basketball Reference*, a site that provides comprehensive basketball statistics. However, no attempts have been made to develop a machine learning model (to the best of my knowledge). Work has been done to solve the same problem in baseball using a Radial Basis Function Network by Smith, Lloyd and Downey, James (2009)[5], and some of the methods here shall be based upon this.

In this project, I strive to harness the capabilities and advantages of machine learning techniques to build a model with better predictive accuracy than that of the aforementioned methods, by experimenting with various different datasets, features and learning algorithms. This report documents the methods employed, work done to collect and preprocess the data, experimentations involved and analysis of the various results.

2 Experimentation

2.1 Domain-Specific Features

In the initial stages of this project, simple basketball statistics were incorporated as features into the training set, alongside other metrics that might be indicative of the relative merit of particular players with respect to other players who played in the league during his time:

- Points Per Game
- Assists Per Game
- Rebounds Per Game
- Career Points
- Career Assists
- Career Rebounds
- Most Valuable Player Awards
- All-Star Game Appearances
- Year Played in the League

Taking into account the fact that a player must be fully retired for five years before being eligible for induction[6] and that the inaugural NBA All-Star Game took place in 1951, we used data from players who debuted in the league during or after the 1951-52 season and retired before the 2008-09 season. This not only avoids differences in basketball eras but ensures comprehensive statistics for each player (no missing or incomplete features for any instance) as all the above statistics were being recorded in that time period. Other key statistics such as *steals*, *blocks* and *field goal percentage* were not incorporated for this very reason, as many players either did not have these statistics recorded at all or only had it recorded for some seasons. Other metrics like player's *draft pick*, *weight* and *height* were also considered but these failed to improve performance.

In an effort to improve predictive accuracy on this model built on a relatively small number of features, several advanced basketball statistics were considered. Namely, *win shares*[9], a metric adopted from baseball that is designed to estimate a player's contribution to his team's win, which is made up of two other metrics, *defensive win shares* and *offensive win shares*, each of which considers a player's offensive and defensive contributions respectively. Since it has been said that certain players are inducted into the Hall of Fame based but on their defensive prowess, which may not reflect well on a stat sheet, I decided it would be beneficial to incorporate both defensive and offensive win shares. Another statistic considered was the *Player Efficiency Rating (PER)*, a rather arcane but popular metric which summarises a player's performance and contributions into a one number using a very detailed formula[10]. Finally, the statistics *effective field goal percentage (eFG%)* and *true shooting percentage (TS%)* were considered, both of which incorporate 3-point and 2-point field goals, the former of which adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal and the latter also considers free throw percentage.

The last metric considered was the number of *NBA Championship Titles* the player has won and the player's *longest stint with any team*. The latter is potentially interesting and valuable as this could serve as an indication that he was a franchise player, an elite status bestowed upon (usually) the best players on their respective teams, whom the franchise is built around. Unfortunately, this was not experimented with due to the difficulty in obtaining this feature.

As noted earlier, since many statistics were not being recorded prior to the 1973-74 season, we cannot simply incorporate these advanced basketball metrics into our model as most rely on these missing statistics. Luckily, *Basketball Reference* provides a complete system with estimations to account for missing statistics in particular eras, the formulae for which are quite detailed. The method for calculating win shares and player efficiency rating can be found at `http://www.basketball-reference.com/about/ws.html` and `http://www.basketball-reference.com/about/per.html`.

2.2 Data Collection and Preparation

A considerable challenge faced in this project was the absence of a readily available data repository relevant to the problem, so a substantial amount of work was devoted overcoming this by web data extraction (scraping) and parsing data into a format that could be analysed with machine learning software. The documentation of the data collection process is kept as brief as possible since while part of the overall machine learning and data mining process, it has little to do with machine learning itself.

At first, official NBA stats portal `http://stats.nba.com/` seemed like a promising avenue for data collection. However, not only were there no datasets available for download, the site did not provide any APIs. To avoid resorting to scraping the pages, I inspected the source of the site's pages to find ways to access the backend scripts that drove the site's data and was able to find Javascript sources that provided results in a friendly Javascript Object Notation (JSON) format. Since each source only provided a subset of features for the particular player queried, I was required to cross reference these data sources to create a complete set of features for each player. The data sources and respective features provided are described below:

- `http://stats.nba.com/stats/commonallplayers` - *List of all NBA players*

- `http://stats.nba.com/stats/commonplayerinfo/?PlayerID=<player_id>` - *All-Star Appearances (Not available for active players), Position*
- `http://stats.nba.com/feeds/players/profile/<player_id>_Profile.js` - *Draft Pick, Weight, Height*
- `http://stats.nba.com/feeds/players/awards/<player_id>_Award.js` - *Most Valuable Player Awards, Hall of Fame Inductee*
- `http://stats.nba.com/stats/playerprofile` - *Points Per Game, Assists Per Game, Rebounds Per Game, Career Points, Career Assists, Career Rebounds*

At this stage of the data collection process, I hadn't conceived of using the advanced statistics described in the previous section, so these features seemed to be enough to get started on preliminary experimentation. This data source obviously failed to yield the number of championship titles, the aforementioned advanced statistics and the full set of simple statistics that were required to calculate complex metrics like the PER. This posed a problem after no further improvements could be obtained with the original dataset and its limited feature space and an alternative data source was required.

Since Basketball Reference provides all of the aforementioned features, with all of the advanced metrics already calculated (and much more), I finally resorted extracting data from its web pages since there was no API or datasets available for download. The list of players and their respective identifiers were obtained from `http://www.basketball-reference.com/players/<first_letter_last_name>` and a players complete profile could be obtained from `http://www.basketball-reference.com/players/<first_letter_last_name>/<identifier>.html`.

The upside of this was that unlike the official NBA stats portal where it was necessary to piece together data from different sources, it was possible to obtain all the information required from a single page. Using the Python library *Beautiful Soup*, I was able to extract all the required features for every player. Additional statistics were also obtained in case they were required, such as statistics from Playoff games, All-Star Games, etc. This data source seemed far more comprehensive, and appeared to also include players from the American Basketball Association (ABA), which was later merged with the NBA in 1976. This also meant a larger training set to work with.

Furthermore, the Most Valuable Player Award data is expressed in terms of cumulative award shares for the player's career, which is calculated by the number of points a player received for a particular award over the total points of all first-place votes[7] - far more useful than merely the raw number of MVP awards in the case of players such as Jason Kidd, who has never won the award but has been a serious contender in many seasons due to his stellar performance.

Both of the aforementioned datasets have included in the submission files in the *JSON* format and can also be parsed into Weka's *ARFF* format or vectorised into a *numpy* array to work with the Python machine learning library *scikit-learn* using the scripts included.

2.3 TRAINING THE MODEL

In the initial experiment, the original dataset described in the previous section was used and trained on using different learning algorithms with the popular machine learning workbench *Weka*[8]: *Logistic Regression*, *Multilayer Perceptrons*, *Decision Trees* and *Radial Basis Function Networks*, using the default parameters for each respective algorithm. To remove irrelevant instances, the dataset filtered out players who played for less than one season (this might be a poor criterion of relevance, since players who play for many seasons don't play many actual games. Games played were not available in this dataset so could not be filtered accordingly).

The performance of these algorithms were estimated with 10-fold cross-validation repeated 10 times and the predictive accuracy of models were compared using a paired t-test. Surprisingly, the logisitic regression model outperformed all other models, but not by a statistically significant amount, with the exception of the tree learner. As seen in table 2.1, the percentage of correct classifications under these data schemes were all estimated to be above 97%.

Table 2.1: Predictive Accuracy (Percent Correct)

Dataset	Logistic	RBF Network	Multilayer Perceptron	Decision Tree (J48)	
Initial NBA Data	98.08±0.83	97.79±1.01	97.83±0.83	97.41±0.93	•

∘, • statistically significant improvement or degradation

The suspiciously (and somewhat deceptively) good predictive accuracy attained by these learning schemes might be explained by the unquestionable skewness of the training set toward negative examples. Indeed, the prestigious honour of enshrinement into the Hall of Fame is only bestowed upon a very select few individuals - 80 out of the 1875 players (4.2%) in the training set (in fact one could attain 95.8% predictive accuracy simply by classifying all instances as negative!).

Consequently, it was necessary to consider other cost-sensitive metrics to evaluate the performance of the learning schemes, the key one being *recall*.

Table 2.2: Recall

Dataset	Logistic	RBF Network	Multilayer Perceptron	Decision Tree (J48)
Initial NBA Data	0.67±0.16	0.70±0.16	0.65±0.16	0.60±0.18

∘, • statistically significant improvement or degradation

In table 2.2, we indeed see that the employed learning schemes did not have high recall, so 30-40% of actual Hall of Fame members were being misclassified as not being in the Hall of Fame, with the RBF network model yielding the highest recall [3].
Table 2.3 shows the confusion matrix for the RBF network model. Optimisations were per-

[3] Obviously, it was also important not to compromise the prestige of the Hall of Fame by classifying instances as positive too liberally.

Table 2.3: Confusion Matrix

		Predicted		
		HOF	not HOF	Total
Actual	HOF	57	23	80
	not HOF	16	1779	1795
	Total	73	1802	1875

formed on this model[4] in an effort to minimise the false negatives and thereby increase recall. The first attempt was to optimise the parameters of the model using the `CVParameterSelection` meta-classifier in Weka, namely, number of clusters, minimum standard deviation for the clusters, and the random seed to be used by k-means clustering. Additionally, I tried to use Weka's feature selection options to either eliminate features that yielded little information gain, or otherwise transform the possibly correlated features into linearly uncorrelated features using principle component analysis. Unfortunately, these efforts were mostly in vain, as these had little to no effect on the predictive accuracy of the model or its ability to correctly classify positive instances. Some minor improvements where obtained, however, with a cost-sensitive meta-classifiers using two different methods: reweighting training instances according to the total cost assigned to each class and predicting the class with minimum expected misclassification cost[11]. The respective results of which can be seen in table 2.4. Note the overall predictive accuracy remained practically unchanged at around 97%.

Table 2.4: Confusion Matrix

		Predicted		
		HOF	not HOF	Total
Actual	HOF	61	19	80
	not HOF	24	1771	1795
	Total	85	1790	1875

(a) reweighted training instances

		Predicted		
		HOF	not HOF	Total
Actual	HOF	59	21	80
	not HOF	21	1774	1795
	Total	80	1795	1875

(b) minimize expected misclassification cost

This was a good improvement, but I wanted to see if further advances could be made. Unfortunately, as mentioned in the data preparation section, this initial experiment only used the very basic set of features and more features were required at this point to make progress on this problem.

With the new dataset acquired from Basketball Reference, players who played less than 100 games were filtered out, along with any players who did not have the complete records of their respective statistics. Consequently, the training set consisted of all the features described in the previous section (including the features in the original dataset), with a total of 1591 play-

[4] There was no reason why the other learning schemes could not have been optimised as well. Radial Basis Function networks seemed especially appealing and promising due to the results achieved in solving the problem for the baseball Hall of Fame.

ers and 85 Hall of Fame members. Initial experimentation with this dataset used all of the features in training the new models.

Table 2.5: Predictive Accuracy (Percent Correct)

Dataset	Logistic	Decision Tree (J48)	Multilayer Perceptron	RBF Network
Basketball Reference Data	97.81±1.04	97.27±1.17	97.83±0.98	97.18±1.11

○, • statistically significant improvement or degradation

Again, as shown in table 2.5, the models were all of comparable predictive accuracy. Upon further inspection, table 2.6 shows RBF Network model yielded higher recall.

Table 2.6: Recall

Dataset	Logistic	Decision Tree (J48)	Multilayer Perceptron	RBF Network
Basketball Reference Data	0.74±0.15	0.68±0.16	0.73±0.15	0.80±0.13

○, • statistically significant improvement or degradation

Table 2.7 shows the confusion matrix for the RBF network model. As one might expect, though there were relatively few false negatives, the false positives were quite high. Consequently, I sought to strike a good balance between precision and recall using similar approaches taken in the initial experiment.

Table 2.7: RBF Network Model - Confusion Matrix

		Predicted		
		HOF	not HOF	Total
Actual	HOF	68	17	85
	not HOF	28	1478	1506
	Total	96	1495	1591

Precision	0.708
Recall	0.8
F-Measure	0.751
Percent Correct	97.17%

Firstly, by analysing the various new features used, a few of them yielded relatively low information gain, and were thus removed. The features that reduced overall predictive accuracy or had little effect were TS%, eFG%, total games played, and the offensive and defensive win shares respectively. Table 2.8 shows the improvements gained from this.

Table 2.8: RBF Network Model (irrelevant features removed) - Confusion Matrix

		Predicted		
		HOF	not HOF	Total
Actual	HOF	67	18	85
	not HOF	20	1486	1506
	Total	87	1504	1591

Precision	0.77
Recall	0.788
F-Measure	0.779
Percent Correct	97.61%

Finally, I sought to optimise the parameters of the RBF network model. By using 3 as the clustering seed with 2 clusters and the minimum standard deviation of 0.1 in addition to removing the win shares feature. Good results were obtained, as depicted in table 2.9.

Table 2.9: RBF Network Model (final model) - Confusion Matrix

		Predicted		
		HOF	not HOF	Total
Actual	HOF	68	17	85
	not HOF	17	1489	1506
	Total	85	1506	1591

Precision	0.8
Recall	0.8
F-Measure	0.8
Percent Correct	97.86%

A comparison of the model's F-measure subject to the variations of parameters of the model and features of the dataset can be seen in table 2.10. The overall predictive accuracy across these models remained unchanged.

Table 2.10: F-measure

Dataset	All features	Filtered features 1	Filtered features 2
RBF Network (default)	0.75± 0.11	0.77±0.11	0.78±0.11
RBF Network (optimised parameters)	0.74± 0.11	0.77±0.11	0.79±0.10

∘, • statistically significant improvement or degradation

Evidently, the final model achieved a good balance between precision and recall as a result of optimisations and feature selection. These approaches were experimented on the different learning schemes, logistic regression, tree learning and multilayer perceptrons but none outperformed the RBF network model, even with cost-sensitive classifiers. The learning schemes that did perform well were on par with the RBF network model.

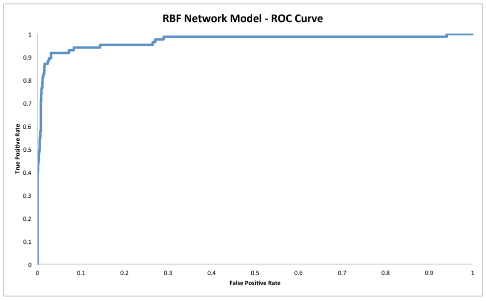

Figure 2.1: ROC Curve

The ROC curve of the RBF Network model is depicted in 2.1 with an area of 0.971 under the curve (AUC).

3 PREDICTIONS

Using the predictive model developed in the previous section, it would be interesting for practical purposes to apply domain knowledge to reason about the misclassifications made by the model and also to analyse the generalisation power of the model by evaluating it on active or recently retired players.

Table 3.1 shows some of the Hall of Fame members with all the members misclassified as nonmembers bolded. A few of these false negatives can be justified by players with relatively short but illustrious careers or short lives in extreme cases, as is the case with Drazen Petrovic died at age 28 and was posthumously honoured by the Hall.

In the book titled *The Book of Basketball* by influential sports columnist Bill Simmons, wherein almost 400 pages are devoted to analysing the Hall of Fame, Simmons breaks down the Hall of Fame into five levels, with level one being the "just-made-it" Hall of Famers: either "having a great career but not long enough", "very good for a long time but never great" and also "memorable career but never won anything". All players that fall into this category are listed by Simmons and many of the members misclassified as nonmembers are on this list: Gail Goodrich, Connie Hawkins, Arvydas Sabonis, Jack Tywman, Chris Mullin, Joe Dumars, Earl Monroe, just to name a few.

Such is the case with players not in the Hall of Fame, in table 3.2 (only shows players with HOF probability above 5%). Many of the nonmembers of the Hall of Fame misclassified as members fall into level one: Robert Horry, Sidney Moncrief, Chris Webber, Shawn Kemp, Kevin Johnson - again, just to name a few. Not also that a large number of these "false positives" are in fact bound for the Hall of Fame, but have yet to be inducted. By way of example, *Not in Hall of Fame* is a site "dedicated to promoting those who have achieved levels of excellence deserving of the Hall of Fame". A look at the top players listed as eligible for the enshrinement class of 2014 (`http://www.notinhalloffame.com/basketball/the-basketball-futures/2014-basketball-eligibles`), we indeed see players like Alonzo Mourning, Anfernee Hardaway, Chris Webber at the top of the list.

The model was evaluated on active or recently retired players, the result of which is shown in table 3.3 (only those with probability greater than 5%). Simmons also lists active and recently retired players in the respective levels and a quick cross-reference finds most of the players with high probability of enshrinement on this list. However, here is where some faults of the model become apparent. In Simmons's system (or pyramid, as he calls it), the top level of the pyramid which consists of "twelve greatest players of all-time, the best of the best" includes an active and recently retired player - the great Tim Duncan and Shaquille O'Neal, both of whom have 100% probability of Hall of Fame enshrinement. However, table 3.3 shows 14 other players who also have 100% probability, some of whom are "level one" players at best. Additionally, for active players, this model really answers the question of how likely they are to be inducted into the Hall of Fame in the future if they were to hypothetically retire *today*. As such, a player like Dwight Howard with 100% probability of induction might outrage some fans. For prediction of the outcome, this model might be fine, but for a realistic assessment of the likelihood of the prediction, this might not be. One might incorporate a quadratic loss function or information loss function to more precisely assess this.

Table 3.1: Hall of Fame Members

Name	Probability	PTS	AST	REB	PPG	APG	RPG	PER
Michael Jordan	1.0	32292	5633	6672	30.1	5.3	6.2	27.9
David Robinson	1.0	20790	2441	10497	21.1	2.5	10.6	26.2
Wilt Chamberlain	1.0	31419	4643	23924	30.1	4.4	22.9	26.1
Bob Pettit	1.0	20880	2369	12849	26.4	3.0	16.2	25.3
Kareem Abdul-Jabbar	1.0	38387	5660	17440	24.6	3.6	11.2	24.6
Charles Barkley	1.0	23757	4215	12546	22.1	3.9	11.7	24.6
Magic Johnson	1.0	17707	10141	6559	19.5	11.2	7.2	24.1
Karl Malone	1.0	36928	5248	14968	25.0	3.6	10.1	23.9
Julius Erving	1.0	30026	5176	10525	24.2	4.2	8.5	23.6
Hakeem Olajuwon	1.0	26946	3058	13748	21.8	2.5	11.1	23.6
Larry Bird	1.0	21791	5695	8974	24.3	6.3	10.0	23.5
		⋮						
Dennis Johnson	0.485	15535	5499	4249	14.1	5.0	3.9	14.6
Gail Goodrich	0.421	19181	4805	3279	18.6	4.7	3.2	16.7
Chris Mullin	0.368	17911	3450	4034	18.2	3.5	4.1	18.8
Joe Dumars	0.31	16401	4612	2203	16.1	4.5	2.2	15.3
Connie Hawkins	0.309	11528	2556	5450	18.7	4.1	8.8	19.9
Jack Twyman	0.298	15840	1861	5424	19.2	2.3	6.6	17.8
Richie Guerin	0.192	14676	4211	4278	17.3	5.0	5.0	17.3
Gus Johnson	0.071	10243	1603	7624	16.2	2.5	12.1	16.7
Earl Monroe	0.038	17454	3594	2796	18.8	3.9	3.0	17.2
Roger Brown	0.035	10498	2315	3758	17.4	3.8	6.2	17.2
Jamaal Wilkes	0.032	14644	2050	5117	17.7	2.5	6.2	16.5
Calvin Murphy	0.024	17949	4402	2103	17.9	4.4	2.1	18.0
Arvydas Sabonis	0.021	5629	964	3436	12.0	2.1	7.3	21.2
Ralph Sampson	0.021	7039	1038	4011	15.4	2.3	8.8	16.0
Tom Gola	0.021	7871	2962	5617	11.3	4.2	8.0	14.1
Bill Bradley	0.021	9217	2533	2354	12.4	3.4	3.2	12.2
Drazen Petrovic	0.009	4461	701	669	15.4	2.4	2.3	16.4
Bob Houbregs	0.001	2611	500	1552	9.3	1.8	5.5	17.0

4 CONCLUSION

Using a Radial Basis Function Network model and by optimising its parameters and training set, a good level of predictive accuracy (97.86%) was obtained. Using domain knowledge, the performance of the model was also assessed subjectively and found it yielded accurate outcome predictions. However, when it comes to a realistic assessment of the likelihood of the prediction, the model was poor at yielding probabilities that accurately reflected the class membership probability.

5 Further Work

Some further work and lines of inquiry not yet investigated includes training a separate model for each position, to account for the unquestionable positional factor involved, or perhaps just a model for likely geographic position of the court, e.g. guards predominantly occupy the backcourt while forwards and centres for the most part occupy the front court. As such, point guards aren't likely to be inducted for their rebounding prowess, so this feature might be excluded when training the model for backcourt players.

For active players, specifically young but dominant players, the model might not accurate reflect their dominance due to their lack of experience. This might be remedied by training a model specific to their experience in the league, say 3 years, so the training set could include statistics of players for up to only their first 3 years in the league.

References

[1] *"Hall of Famers"* Retrieved 6 June 2013 from Naismith Memorial Basketball Hall of Fame: `http://www.hoophall.com/hall-of-famers-index/`

[2] Aschburner, Steve. (12 August 2012) *"Hall of Fame selection process leaves much to be desired"*. Retrieved 6 June 2012 from NBA: `http://www.nba.com/2010/news/features/steve_aschburner/08/12/hall.process/index.html`

[3] Ziller, Tom. (30 March 2010) *"Fans to Vote for Basketball Hall of Fame Inductees"*. Retrieved 1 June 2013 from AOL News: `http://nba.fanhouse.com/2010/03/30/fans-to-vote-for-basketball-hall-of-fame-inductees/`

[4] *"Hall of Fame Probability"*. Retrieved 1 June 2013 from Basketball Reference: `http://www.basketball-reference.com/about/hof_prob.html`

[5] Smith, Lloyd and Downey, James (2009) *"Predicting Baseball Hall of Fame Membership using a Radial Basis Function Network,"* Journal of Quantitative Analysis in Sports: Vol. 5: Iss. 1, Article 6.

[6] *"Guidelines For Nomination and Election Into the Naismith Memorial Basketball Hall of Fame"* Retrieved 7 June 2013 from Basketball Hall of Fame: `http://www.hoophall.com/enshrinement-process/`

[7] James Bowman *"WNBA Most Valuable Player Shares"* Retrieved 7 June 2013 from Swish Appeal: `http://www.swishappeal.com/2010/2/21/1320444/wnba-most-valuable-player-shares`

[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009) *"The WEKA Data Mining Software: An Update,"* SIGKDD Explorations, Volume 11, Issue 1.

[9] *"Calculating Win Shares"* Retrieved 7 June 2013 from Basketball Reference: `http://www.basketball-reference.com/about/ws.html`

[10] *"Calculating PER"* Retrieved 7 June 2013 from Basketball Reference: `http://www.basketball-reference.com/about/per.html`

[11] *"CostSensitiveClassifier"* Retrieved 7 June 2013 from Weka Documentation: `http://weka.sourceforge.net/doc/weka/classifiers/meta/CostSensitiveClassifier.html`

Table 3.2: Hall of Fame Nonmembers

Name	Probability	PTS	AST	REB	PPG	APG	RPG	PER
Alonzo Mourning	0.777	14311	946	7137	17.1	1.1	8.5	21.2
George McGinnis	0.738	17009	3089	9233	20.2	3.7	11.0	20.0
Chris Webber	0.734	17182	3526	8124	20.7	4.2	9.8	20.9
Jack Sikma	0.676	17287	3488	10816	15.6	3.2	9.8	17.3
Tim Hardaway	0.672	15373	7095	2855	17.7	8.2	3.3	18.6
Sidney Moncrief	0.671	11931	2793	3575	15.6	3.6	4.7	18.7
Tom Sanders	0.671	8766	1026	5798	9.6	1.1	6.3	11.9
Spencer Haywood	0.67	17111	1541	8675	20.3	1.8	10.3	19.1
Anfernee Hardaway	0.669	10684	3525	3146	15.2	5.0	4.5	17.4
Robert Horry	0.668	7715	2343	5269	7.0	2.1	4.8	13.4
Kevin Johnson	0.659	13127	6711	2404	17.9	9.1	3.3	20.7
Mitch Richmond	0.652	20497	3398	3801	21.0	3.5	3.9	17.6
Walter Davis	0.647	19521	3878	3053	18.9	3.8	3.0	19.1
Mark Jackson	0.619	12489	10334	4963	9.6	8.0	3.8	16.0
Zelmo Beaty	0.601	15207	1365	9665	17.1	1.5	10.9	18.7
Shawn Kemp	0.566	15347	1704	8834	14.6	1.6	8.4	19.1
Terry Cummings	0.562	19460	2190	8630	16.4	1.9	7.3	18.3
Marques Johnson	0.555	13892	2502	4817	20.1	3.6	7.0	20.1
Sam Cassell	0.507	15635	5939	3221	15.7	6.0	3.2	19.5
Jo Jo White	0.427	14399	4095	3345	17.2	4.9	4.0	14.2
Ron Harper	0.362	13910	3916	4309	13.8	3.9	4.3	15.6
Louie Dampier	0.354	15279	4687	2543	15.9	4.9	2.6	15.2
Lou Hudson	0.352	17940	2432	3926	20.2	2.7	4.4	17.4
Jim Loscutoff	0.248	3156	353	2848	6.2	0.7	5.6	8.3
Mark Aguirre	0.235	18458	2871	4578	20.0	3.1	5.0	19.0
Reggie Theus	0.217	19015	6453	3349	18.5	6.3	3.3	16.6
Buck Williams	0.147	16784	1646	13017	12.8	1.3	10.0	15.3
Gus Williams	0.124	14093	4597	2222	17.1	5.6	2.7	18.5
Mark Price	0.119	10989	4863	1848	15.2	6.7	2.6	19.6
Rod Strickland	0.102	14463	7987	4084	13.2	7.3	3.7	18.0
Brad Daugherty	0.089	10389	2028	5227	19.0	3.7	9.5	18.9
Maurice Lucas	0.067	14857	2498	9306	14.6	2.4	9.1	16.2
Maurice Cheeks	0.061	12195	7392	3088	11.1	6.7	2.8	16.5
Horace Grant	0.056	12996	2575	9443	11.2	2.2	8.1	16.0
Bill Laimbeer	0.053	13790	2184	10400	12.9	2.0	9.7	16.1
Tom Chambers	0.052	20049	2283	6703	18.1	2.1	6.1	16.5
Charlie Scott	0.051	14837	3515	2846	20.7	4.9	4.0	15.8
Paul Westphal	0.05	12809	3591	1580	15.6	4.4	1.9	19.4

Table 3.3: Active or recently retired players

Name	Probability	PTS	AST	REB	PPG	APG	RPG	PER
LeBron James	1.0	21081	5302	5553	27.6	6.9	7.3	27.6
Shaquille O'Neal	1.0	28596	3026	13099	23.7	2.5	10.9	26.4
Chris Paul	1.0	10311	5449	2426	18.6	9.8	4.4	25.5
Dwyane Wade	1.0	16453	4049	3364	24.7	6.1	5.1	25.5
Tim Duncan	1.0	23785	3612	13219	20.2	3.1	11.2	24.7
Kevin Durant	1.0	12258	1447	3153	26.6	3.1	6.8	23.6
Dirk Nowitzki	1.0	25051	2923	9096	22.6	2.6	8.2	23.5
Kobe Bryant	1.0	31617	5887	6575	25.5	4.8	5.3	23.4
Kevin Garnett	1.0	25274	5224	13843	19.1	3.9	10.5	23.1
Dwight Howard	1.0	12731	1043	9017	18.3	1.5	12.9	22.2
Tracy McGrady	1.0	18381	4161	5276	19.6	4.4	5.6	22.1
Allen Iverson	1.0	24368	5624	3394	26.7	6.2	3.7	20.9
Paul Pierce	1.0	24021	4305	6651	21.8	3.9	6.0	20.6
Vince Carter	1.0	22223	4030	5350	20.8	3.8	5.0	20.3
Steve Nash	1.0	17285	10249	3613	14.4	8.5	3.0	20.0
Ray Allen	1.0	23804	4218	5067	19.4	3.4	4.1	18.8
Jason Kidd	0.999	17529	12091	8725	12.6	8.7	6.3	17.9
Carmelo Anthony	0.855	17846	2181	4551	25.0	3.1	6.4	20.8
Chris Bosh	0.715	13970	1485	6370	19.5	2.1	8.9	20.8
Amar'e Stoudemire	0.691	14242	944	5761	21.3	1.4	8.6	22.2
Yao Ming	0.686	9247	769	4494	19.0	1.6	9.2	23.0
Grant Hill	0.678	17137	4252	6169	16.7	4.1	6.0	19.0
Derrick Rose	0.671	5858	1911	1071	21.0	6.8	3.8	19.9
Tony Parker	0.671	14917	5247	2645	17.1	6.0	3.0	19.1
Chauncey Billups	0.671	15730	5594	2964	15.4	5.5	2.9	19.0
Jermaine O'Neal	0.671	12960	1344	7019	13.4	1.4	7.3	18.0
Dikembe Mutombo	0.665	11729	1240	12359	9.8	1.0	10.3	17.2
Pau Gasol	0.661	15535	2780	7756	18.4	3.3	9.2	21.6
Stephon Marbury	0.482	16297	6471	2516	19.3	7.6	3.0	18.7
Shawn Marion	0.387	16633	2023	9402	16.1	2.0	9.1	19.4
Blake Griffin	0.364	4653	821	2368	20.4	3.6	10.4	22.5
Elton Brand	0.242	16242	2070	8516	17.4	2.2	9.1	21.0
Deron Williams	0.148	10386	5241	1877	17.8	9.0	3.2	19.3
Andre Miller	0.13	15496	7956	4461	13.8	7.1	4.0	17.7
Kevin Love	0.1	4979	557	3490	17.3	1.9	12.2	22.1
Joe Johnson	0.1	15927	3948	3698	17.6	4.4	4.1	16.3
Manu Ginobili	0.091	10819	2892	2839	14.9	4.0	3.9	21.6
Antawn Jamison	0.066	19958	1754	8102	18.8	1.7	7.6	18.2
Peja Stojakovic	0.056	13647	1408	3782	17.0	1.8	4.7	17.1