Classifying Basketball Players by Hall of Fame Merit

ABSTRACT

Throughout basketball history there have been some truly exceptional players in the National Basketball Association (NBA) and American Basketball Association (ABA). There players are rewarded with the honor of being inducted into the Hall of Fame. However, there is not an exact science as to how voters for the Hall of Fame vote on which players are most deserving. This often leads to many players being unfairly excluded despite having truly exceptional statistics. In this paper, we analyze every professional basketball player in NBA&ABA history and try to decide whether they deserve to be in the Hall of Fame based on analyzing the players that are currently in the Hall of Fame. Our intent is to apply data mining and machine learning techniques to provide more solid ground and advanced analytics for the voting committee when they vote for deserving players.

CCS CONCEPTS

• Applied computing • Information systems

KEYWORDS

Hall of Fame, Player Index, Machine Learning, Prediction Models, Confusion Matrix

1 Introduction

Basketball has been played at the professional level since 1946, and most of the relevant data is available and can be queried on basketball reference's Player Index [1]. Using this tool, we were able to collect the data of every professional basketball player in history and divide them into categories of players that are currently in the Hall of Fame vs players that are not currently in the Hall of Fame. Since a player needs to be retired for four full seasons before they are eligible for the Hall of Fame, we decided to exclude any player that has played in the NBA in the past four years.

As the Hall of Fame committee votes on certain players to enter the Hall of Fame each year, sports experts debate about which players deserve to be in the Hall of Fame and which ones do not. However, they often do not have a strong statistical background nor a strong analytical basis for their claims. Our goal is to use data mining and machine learning techniques to provide a starting point for these experts to have some analytics to cite reasons as to why a certain player deserves to be in the Hall of Fame.

In Section 2, we will show exactly how we gathered and cleaned our data. In Section 3, we will show our results and the level of confidence. In Section 4, we will open it up to a discussion for our findings and explain what we found. Section 5 maps out possible directions for future work.

2 Data Collection and Preparation

To collect data, we built a web scrapper using the BeautifulSoup Python package [2] to scrape data from basketball-reference.com. We scraped the data of players in and not in the Hall of Fame separately. Unfortunately, the Player Index had some bugs in that it would skip over random players. For example, Ray Allen was a professional basketball player that retired in 2014 and was inducted into the Hall of Fame in April 2018. However, he was not found in either list of players in or not in the Hall of Fame. For these players, we simply excluded them from our dataset and ended up with 3,704 players. Unfortunately, the exact number of excluded players remains unknown due to the bug on the website. Table 1 shows the features of the dataset.

Table 1: Features of the basketball player dataset.

Feature	Explanation
Rk	Rank
Player	Player
Season	If listed as single number, the year the season ended. ★ - Indicates All-Star for league. Only on regular season tables.
Age	Age of Player at the start of February 1st of that season.
Tm	Team
Lg	League
WS	Win Shares: An estimate of the number of wins contributed by a player.
G	Games
GS	Games Started
MP	Minutes Played
FG	Field Goals
FGA	Field Goal Attempts
2*P*	2-Point Field Goals
2*PA*	2-point Field Goal Attempts
3*P*	3-Point Field Goals
3*PA*	3-Point Field Goal Attempts
FT	Free Throws
FTA	Free Throw Attempts
ORB	Offensive Rebounds
DRB	Defensive Rebounds
TRB	Total Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls
PTS	Points
FG%	Field Goal Percentage
2*P%*	2-Point Field Goal Percentage
3*P%*	3-Point Field Goal Percentage
eFG%	Effective Field Goal Percentage: This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.
FT%	Free Throw Percentage
TS%	True Shooting Percentage: A measure of shooting efficiency that considers 2-point field goals, 3-point field goals, and free throws.

Once we compiled our datasets, we had to clean the data by filling in missing values. To fill in missing values, we first decided which features we could replace with zeroes and which we need to predict a value for. We decided that the missing values in 3*P*, 3*PA*, and 3*P%* (see Table 1) should be replaced with zeroes. The reason is that the 3-Point line was not implemented into the ABA and NBA until 1967 and 1979, respectively. It is safe to say that players who only played before the induction of the 3-Point line in their respective league never attempted a 3-Pointer, therefore zero values for these three features.

For the missing *FG%, FT%, and* 2*P*% values (see Table 1), we saw that the corresponding data points also have zeroes for *FG*, *FT*, and 2*P*, respectively. According to the definitions in Equations 1-3, we filled those blank *FG%, FT%, and* 2*P*% fields with zeroes. For *WS*, also known as win shares, there were only two missing data points, and since this small number of missing values have very little statistical impact on the data or the other features, we simply pad them with zeros.

$$FG\% = \frac{FG}{FGA} \quad (1)$$

$$FT\% = \frac{FT}{FTA} \quad (2)$$

$$2P\% = \frac{2P}{2PA} \quad (3)$$

For the other missing values, we need to find a way to predict them. To do this, we measured each feature's correlation to every other feature. This is illustrated in Figure 1. We then find a best-fit-line (regression line) between two features and predict the missing value of one feature based on the value of the other feature if the latter is not missing. For this, we made sure that the feature used to predict the missing value of another feature is not missing any value. This means we would not necessarily predict the missing values of one feature with the feature that is the most correlated to it. We would do prediction using the feature most correlated to the predicted feature and not missing any value. Once a feature no longer has any missing values, it can be used to predict the missing values in another feature. The missing values were calculated using the order illustrated below:

Step 1: PF → TRB → DRB → BLK
Step 2: TRB → ORB
Step 3: FGA → MP → TOV → STL

To calculate the best fit line (regression line), we first tried to use the correlation as the slope of the regression line and calculate the *y*-intercept against samples of our dataset. This sometimes resulted in predictions with unrealistic statistics (such as having 2,164 *TRB* and 0 *PF*), as illustrated in Figure 2(a) where the predicting feature is on X-axis and the predicted feature is on Y-axis. A first attempt to remedy this issue was to assume that the data more closely followed an exponential relationship, so we used the logarithmic value (e.g., using log(*PF*) to predict *TRB*) for prediction, as illustrated in Figure 2(b). However, these predictions were overly sensitive to outliers. This method also opened the possibility to runtime errors if the *y*-intercept was negative.

Finally, we settled on using a non-linear least square regression (LSR) line. This ensures that we will not have any negative values. The formulas for least square regression line are in Equations 4-6, where *n* is the total number of data points and (*x*, *y*) are the values of a pair of features in a correlation study. The resulting regression lines using this methodology are shown in Figure 3(a)-(c), obviously better than that of the previous attempts.

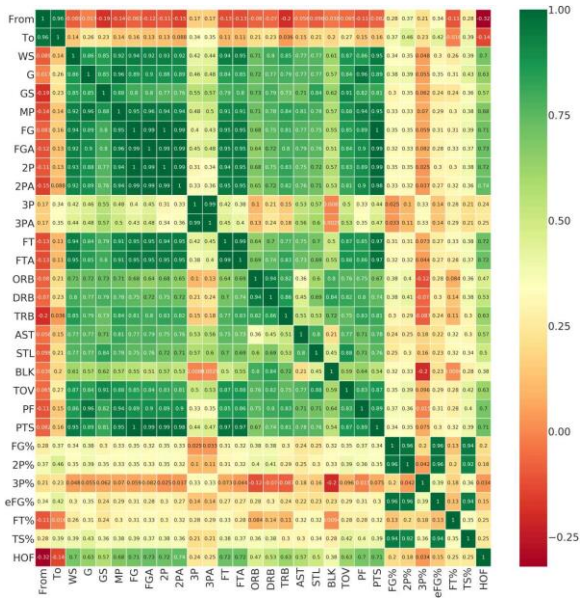
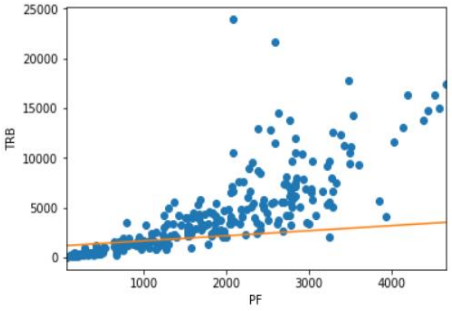
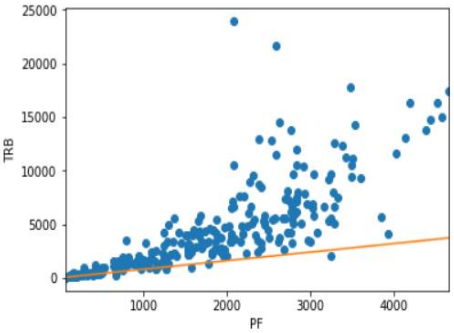


Figure 1: Plot of the Correlation Table



(a)



(b)

Figure 2: (a) TRB by PF using best fit correlation (b) TRB by PF using logarithmic correlation

$$y = ax + b \quad (4)$$

$$b = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x^2)} \quad (5)$$

$$a = \frac{\Sigma y - b\Sigma x}{n} \quad (6)$$

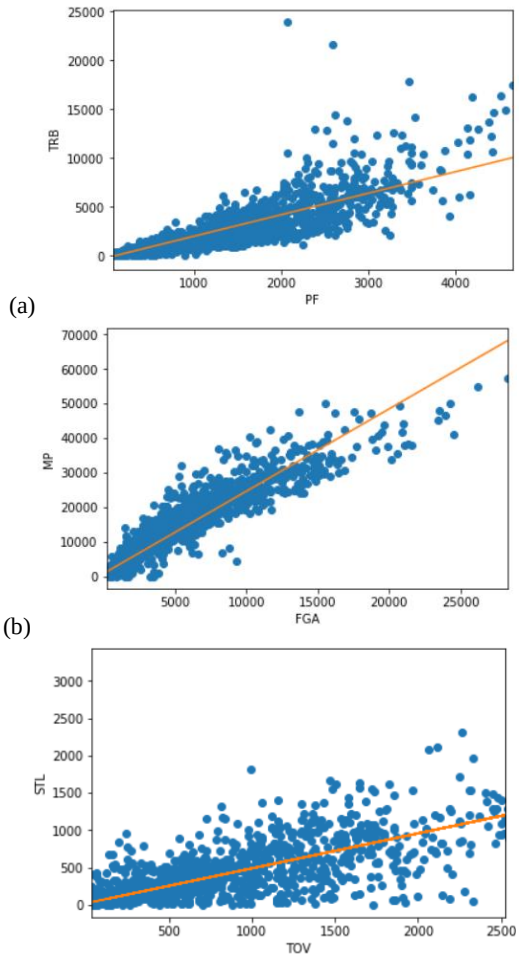


Figure 3: (a) MP by FGA using LSR correlation, (b) MP by FGA using LSR correlation, and (c) STL by TOV using LSR correlation.

Once we had the dataset ready, we checked to see how many players were in the Hall of Fame vs players that were not in the Hall of Fame. Since the disparity was so great and might cause bias in training the prediction/classification model, we added a few pseudo-duplicate Hall of Fame players until we had roughly reduced the percentage of players not in the Hall of Fame down to two thirds. The way we added duplicates was varying the value of each feature by a small amount (e.g., 0.01) for generating pseudo-duplicates. This is to trick the model to think they are two different players. You can see the before and after distributions of our dataset in Figure 4(a)-(b).

3 Prediction Models and Results

To train the prediction model, we compared and contrasted several different methods. These methods include Radical Basis Function Support Vector Machine (rbf SVM) [3-4], Logistic Regression [4-5], Decision Tree, K-Nearest Neighbors (KNN), Gaussian Naïve Bayes, Random Forest, and Ensemble [4].

First, we tested the original dataset without any added duplicates. These tests failed to yield strong results, because the dataset is extremely imbalanced with only 3.6% positive samples (players in the Hall of Fame.) The accuracy (precision) of the original dataset was extremely high, but the F-measure [4] (a commonly adopted measure that considers not only precision but also recall) was extremely low because although it correctly predicted most of the negatives, it missed most of the positives, i.e., a high true negative rate but a high false negative rate.

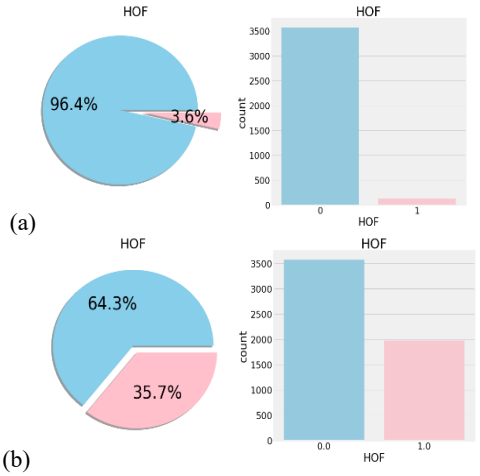


Figure 4: (a) The original count and percentage of players in the Hall of Fame. (b) The count and percentage of players in the Hall of Fame after we added pseudo-duplicates.

Table 2: The accuracy and F-Measure of the original dataset

	ACCURACY	F-MEASURE
RADIAL SVM	**0.964365**	**0.000000**
LOGISTIC REGRESSION	**0.951811**	**0.280374**
KNN	**0.937780**	**0.210325**
DECISION TREE	**0.810004**	**0.481159**
NAÏVE BAYES	**0.880599**	**0.130703**
RANDOM FOREST	**0.887788**	**0.314815**
AVERAGE	**0.905391**	**0.236229**

We ran the data through each prediction model using 10-fold cross-validation. As shown in Table 2, for the original dataset, the cross-valued mean of accuracy was relatively high for each model, but when we used the more balanced dataset, the cross-valued mean of accuracy was even higher (see Table 3). To further illustrate the point that the more balanced dataset was a better choice, we calculated the F-Measure of each model for both datasets. You can clearly see the significantly higher F-measures when the more balanced dataset is used (Table 3), regardless of the prediction model used. You can also see the differences reflected in the confusion matrices between the two datasets in Figure 5. The improvement is again because when using the original, extremely unbalanced dataset for training, although an overwhelming number of negatives were accurately predicted (high true negative rate), many of the positives were not correctly predicted (high false negative rate.)

Since using the more balanced dataset seems more promising, we further ran an ensemble of predictions models using the following schemes: Voting, Bagging KNN, bagging decision tree, AdaBoost, and Gradient Boosting [4]. The accuracies, F-Measures, and mean of the cross-value scores at 10-fold validation are presented in Table 4. As can be seen from Table 4,

the average F-measure of ensemble methods is significantly better than the average performance of using an individual prediction model alone. Out of the 5 ensemble methods we tried, 4 of them have a higher F-measure than the average F-Measure of using individual prediction methods. This gives us a high confidence in our later justification of which other players also deserve to be in the Hall of Fame.

Table 3: The accuracy and F-Measure of the edited dataset

	ACCURACY	F-MEASURE
RADIAL SVM	**1.000000**	**1.000000**
LOGISTIC REGRESSION	**0.952214**	**0.562963**
KNN	**0.967063**	**0.683938**
DECISION TREE	**0.990821**	**0.876812**
NAÏVE BAYES	**0.896058**	**0.357262**
RANDOM FOREST	**0.994600**	**0.926471**
AVERAGE	**0.966793**	**0.734574**

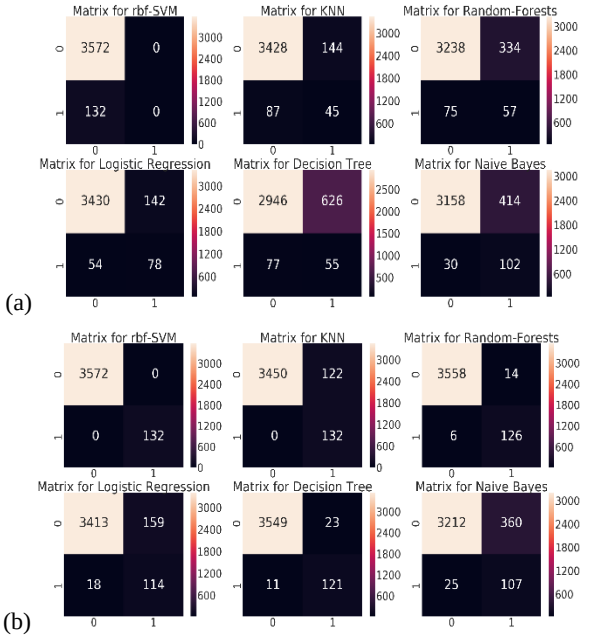


Figure 5: The 4 cells in each of the confusion matrix above, from top to bottom and from left to right, represent true negatives, false negatives, false positives, and true positives, respectively. (a) The confusion matrices of the original dataset and (b) the confusion matrices of the dataset with duplicate positives.

4 Discussions and Conclusions

Based on the promising results generated using our more balanced dataset, we are confident about the legitimacy of our model. We looked for the basketball players that were false positives for every model except rbf SVM (because it did not have any false positives), and concluded that there are 144 players who are not in the Hall of Fame but highly deserve to be in it.

Right after we completed this study, the Hall of Fame inducted 5 more players. Of these players, we correctly predicted 3 of them (Jason Kidd, Horace Grant, and Maurice Cheeks), and the other 2 (Steve Nash and Ray Allen) were not in our original dataset due to the Play Index experiencing a bug. We believe that we can use this model to strongly debate which players should be inducted into the Hall of Fame in future years, and hopefully the basketball experts will consider our model in future votes.

Table 4: 10-fold cross validation results of an Ensemble analysis using the above prediction methods

	ACCURACY	CV MEAN	F-MEASURE
VOTING	**0.984712**	**0.908763**	**0.901408**
BAGGING KNN	**0.979212**	**0.934830**	**0.774194**
BAGGING DECISION TREE	**0.993251**	**0.881061**	**0.907749**
ADABOOST	**0.967626**	**0.882951**	**0.647059**
GRADIENT BOOSTING	**0.988309**	**0.873001**	**0.928302**
AVERAGE	**0.982622**	**0.896121**	**0.831742**

5 Future Work

We will try to clean our data further by categorizing our data. For example, it would be beneficial to see which Hall of Fame players played for at least 10 years, as that would help account for some outliers that might result in false negatives.

We can also use time series data by mining the players' stats by season instead of over their entire career. Using this time series data, we can potentially make better predictions using Recurrent Neural Networks (RNN) [6] and Long Short-Term Memory (LSTM) [7].

ACKNOWLEDGMENTS

We drew inspiration from a project on Kaggle called EDA to Titanic (DieTanic) by Ashwini Swain [8]. Swain did a survival analysis of the Titanic's passengers and we believed we could relate this work to our basketball predictions. A UAB PhD student named Pravinkumar G. Kandhare provided alternative viewpoints and ideas.

REFERENCES

[1] "Play Index | Basketball-Reference.com." [Online]. Available: https://www.basketball-reference.com/play-index/. [Accessed: 08-Sep-2018].

[2] "Beautiful Soup: We called him Tortoise because he taught us." [Online]. Available: https://www.crummy.com/software/BeautifulSoup/. [Accessed: 08-Sep-2018].

[3] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[5] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, 1967.

[6] H. Goel, I. Melnyk, and A. Banerjee, "R2N2: Residual Recurrent Neural Networks for Multivariate Time Series Forecasting," *arXiv Prepr.*, 2017.

[7] Y. Zhao, R. Yang, G. Chevalier, R. C. Shah, and R. Romijnders, "Applying deep bidirectional LSTM and mixture density network for basketball trajectory prediction," *Optik (Stuttg).*, 2018.

[8] A. Swain, "EDA To Prediction(DieTanic) | Kaggle." [Online]. Available: https://www.kaggle.com/ash316/eda-to-prediction-dietanic. [Accessed: 08-Sep-2018].