# What NBA Players Need to do to Make the Naismith Basketball Hall of Fame

**Harry Wang**
Statistics 440: Case Studies in the Practice of Statistics
{ harry.s.wang}@duke.edu

## Abstract

This case study develops predictive models using logistic regression and other learning classifiers to forecast the probability of a basketball player making it into the Naismith Hall of Fame. Our dataset utilizes player statistics compiled over the years by Basketball Reference, a popular sports statistics site. The two primary models we utilized in our study were the Logistic Regression classifier and the Random Forest classifier. In addition, we wanted to explore Occam's Razor and interpretability so we implemented Recursive Feature Elimination in lieu with our classifiers. The preliminary results from our studies are promising, as all of our models achieve an accuracy score of greater than 0.98. Furthermore, our models highlight several important predictors in determining if a player will make the HOF. By leveraging the results in our study, we hope to gain greater insight into making HOF selections more objective and fair.

## 1 Introduction

### 1.1 Motivation

The National Basketball Association (NBA) has been home to some of the most talented athletes in the world since its inception in 1946. Among these, a select few have achieved the prestigious honor of being inducted into the Naismith Memorial Basketball Hall of Fame (HOF), a testament to their exceptional skills, contributions, and lasting impact on the sport. The decision to induct a player into the Hall of Fame is multifaceted, influenced by both quantitative metrics and qualitative assessments of a player's career.

In recent years, the advent of advanced analytics in sports has provided new tools to quantify an athlete's career and forecast their likelihood of receiving this honor. This study aims to leverage a comprehensive dataset of NBA player statistics to develop a predictive model that can effectively estimate the probability of a player being inducted into the Hall of Fame based on their career achievements and performance metrics. By applying statistical techniques, this research not only seeks to predict future HOF inductees but also to uncover the key factors that most significantly influence HOF induction.

Furthermore, understanding the qualifications into the HOF can provide important baselines for the future admittance of players. It is important that the level of accomplishment needed to enter the HOF remains unchanged throughout the years in order to ensure fairness in the entire process. Being able to track the statistical levels of all players throughout the many generations of NBA players will allow future delegates to fairly elect players into the Hall.

Therefore, in this study, we will seek to answer two major questions:

1. Can we build a model to predict a future player's chances at making the HOF?
2. What statistics are the most important indicator of a player's probability of making the HOF?

# 2 Methodology

## 2.1 Data Sources

The dataset employed in this study was compiled through an extensive web scraping effort targeting Basketball Reference [1], a comprehensive online database that provides detailed statistics on all aspects of professional basketball, particularly the National Basketball Association (NBA). The data includes a wide array of player statistics, ranging from basic metrics such as points per game and rebounds to more advanced metrics like win shares and value over replacement player scores.

The data was collected to include a variety of player statistics from the entirety of their careers, encompassing all players who have played in the NBA/ABA throughout their careers. Active players and players current ineligible to make the Hall of Fame are also included in the scrape, will be omitted for modelling purposes.

This rigorous approach to data collection ensures a robust foundation for subsequent analyses, facilitating a comprehensive exploration of the factors that influence Hall of Fame induction in professional basketball. The dataset's breadth and depth allow for a nuanced understanding of performance metrics that potentially predict Hall of Fame status, providing a significant contribution to sports analytics and prediction models in basketball.

## 2.2 Features

We will give a brief description of several key features in our dataset:

1. Awards: Awards are awarded throughout and after the NBA season to individuals and players. Most awards included in the dataset are known to be pinnacles of an athlete's career. Most Valuable Player (MVP), Finals Most Valuable Player (Finals MVP), All NBA team selection (first through third teams), Defensive Player of the Year, etc.
2. Individual Per Game stats: These stats are accumulated throughout all the games a player has played in their career. It consists of points per game, field goal percentage, 3-point percentage, triple-doubles, etc.
3. Advanced Statistics: These statistics usually measure how much a player impacts a certain area of the game. For example, a Plus Minus measures how much the game score changes when a player is on the court. Offensive/Defensive Win Shares, Plus Minus, Player Efficiency Rating, etc.

   It is also important to discuss that most of the variables in our dataset are positively corrected with making the Hall of Fame. This is depicted in Figure 2. This is because in basketball, there are not that many stats that negatively affect HOF chances, but instead, it is the lack of certain statistics that prevent most players from not making it.

## 2.3 Covariate Selection

Since the dataset we are using contains over 80 different features, we needed to perform covariate filtering. Selecting the right covariates can dramatically increase the accuracy and predictive power of a model. By including only those variables that have a significant impact on the target variable, the model is less likely to learn noise and more likely to capture the true underlying patterns in the data. This also deals with overfitting, which is when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. By reducing the number of features, covariate selection helps in simplifying the model, thereby making it more generalizable to new, unseen data. Furthermore, models with fewer variable are more easy to understand and interpret. This will be very useful as we are interested in which particular statistics will affect HOF chances the most.

### 2.3.1 Recursive Feature Elimination

For this study, Recursive Feature Elimination (RFE) was employed as the method of covariate selection. RFE is a backward selection approach that recursively removes the least important features based on the model weights. It is particularly useful when the goal is to identify a subset of features that contribute the most to predicting the target variable.

- Effectiveness: RFE has proven to be effective in various scenarios where feature selection plays a crucial role in predictive accuracy.

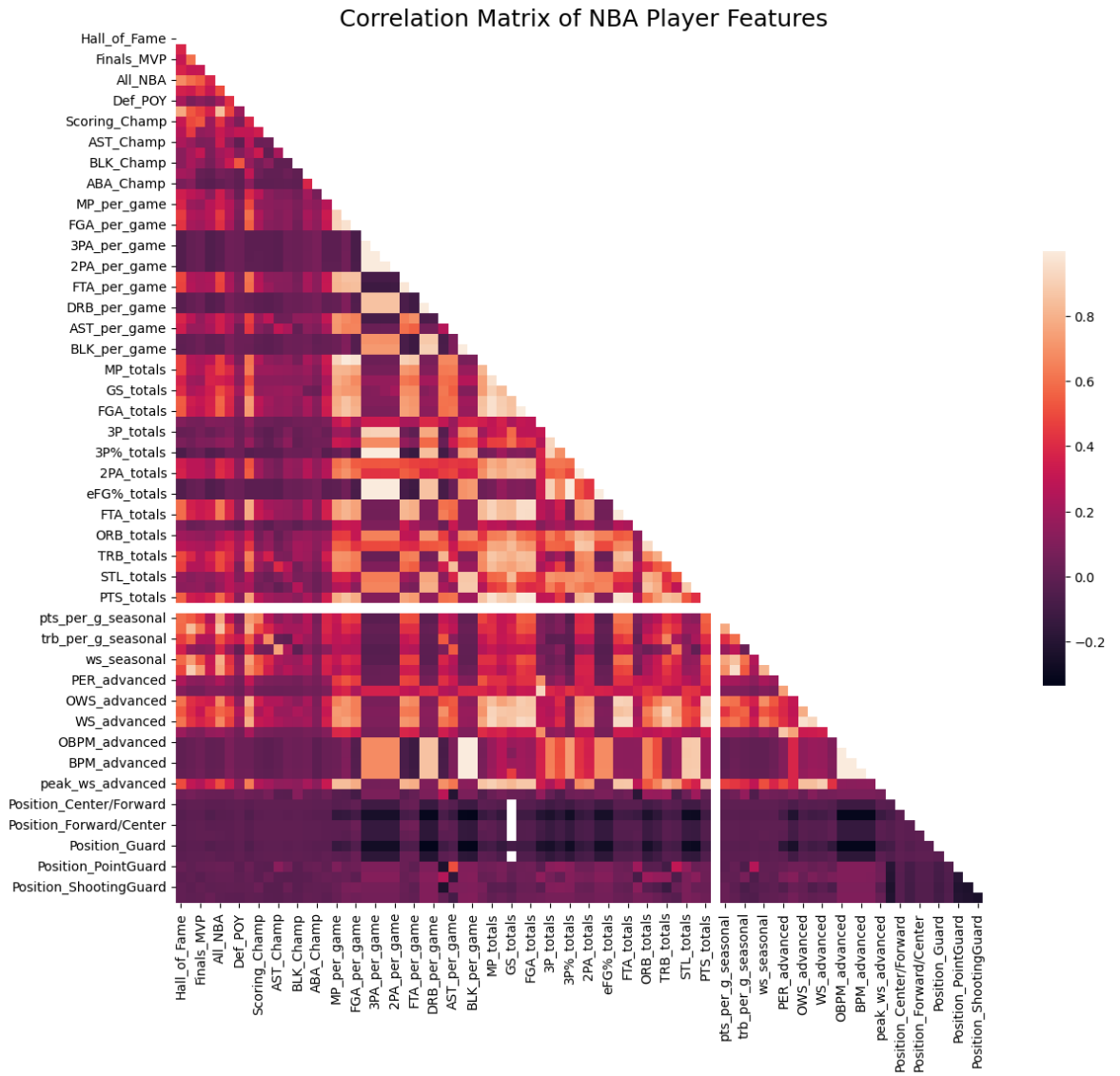


Figure 1: Correlation Heatmap between all features in dataset

- Compatibility: RFE can be used with any estimator that assigns weights to features (either through coefficients or through feature importance), making it versatile across different modeling techniques.
- Optimality: It provides an efficient way to select features by recursively considering smaller and smaller sets of features.

Thus, RFE was chosen to ensure that the model developed not only performs well on historical data but also generalizes effectively to new, unseen data by focusing on the most predictive features.

## 2.4 Models

### 2.4.1 Logistic Regression

Logistic Regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature (e.g., true/false or success/failure).

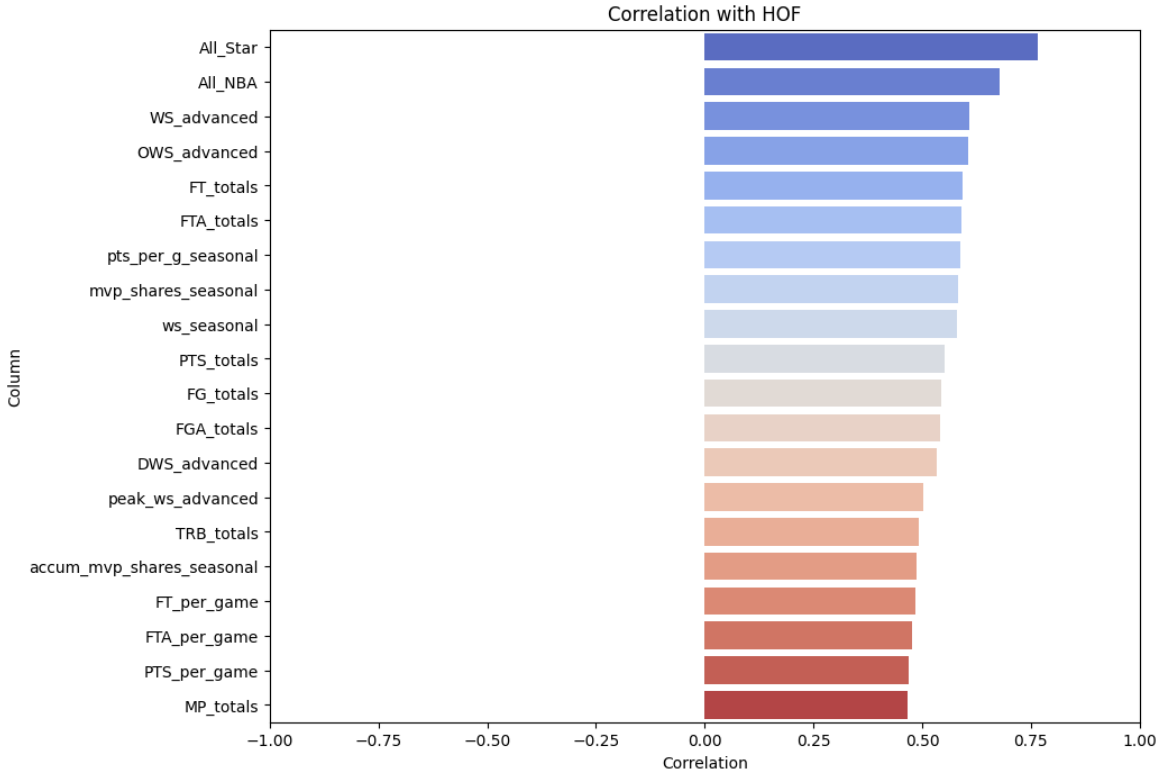


Figure 2: Top 20 features in terms of absolute correlation with making Hall of Fame

#### 2.4.1.1 Assumptions

In order to perform a logistic regression model, a few assumptions must be satisfied first. All these assumptions tests can be found in the Appendix Section 5.2.

- The dependent variable should be binary. This is already satisfied, as a player either makes the HOF or doesn't.
- The independent variables should be independent of each other, i.e., little or no multicollinearity.
- There should be a linear relationship between the logit of the outcome and each predictor variable.
- The sample size should be sufficiently large.

#### 2.4.1.2 Model Equation

The logistic function, used to model the probability of a certain class or event, is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where $p$ is the probability of the presence of the characteristic of interest, $\beta_k$ is the coefficient for the k-th covariate, and $\beta_0$ is the intercept. The equation can be rearranged into the logistic function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

### 2.4.2 Random Forest

Random Forest is an ensemble learning method for classification and regression that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes or mean prediction of the individual trees. We decided to select a random forest model in addition to Logistic Regression because we wanted to provide a layer of complexity in our analysis. Random forests are generally known for their high accuracy and we wanted to see if a more complex model could further improve our results. Furthermore, they can be used for feature selection because if you fit the algorithm with features that are not useful, the algorithm simply won't use them to split on the data.

#### 2.4.2.1 Assumptions

- Random Forest assumes that the predictors are not highly correlated to each other.
- The model works well with a large number of training samples and handles higher dimensionality very well.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

**2.4.2.2 Model Equation** The prediction of a Random Forest model is given by the average of the predictions of the ensemble of trees:

$$Y = \frac{1}{K} \sum_{k=1}^{K} f_k(X)$$

where $K$ is the number of trees, $f_k$ represents the prediction of the $k$-th tree, and $X$ are the input variables.

# 3 Results

## 3.1 Logistic Regression

The results for our basic Logisitic Regression classifier can be seen in Table 1. High precision (0.99 for Class 0 and 0.87 for Class 1) with slightly lower recall for Class 1 (0.57) indicates the model is quite reliable when it predicts an instance as positive (e.g., Hall of Fame) but misses around 30% of actual positive cases. F1-scores of 0.99 for Class 0 and 0.68 for Class 1 suggest good performance for Class 0 but room for improvement for Class 1, especially in improving recall without sacrificing precision. In terms of accuracy, at 0.9857, the model is decent, predicting correctly for most cases. A ROC AUC Score score of 0.9898 indicates excellent discrimination between the positive and negative classes across various threshold settings. Overall, it seems that our baseline model performs decently well.

Table 1: Performance Metrics and Parameters of the Logistic Regression Model

| Metric | Reject HOF | Made HOF | Macro Avg | Weighted Avg |
|---|---|---|---|---|
| Precision | 0.99 | 0.87 | 0.93 | 0.98 |
| Recall | 1.00 | 0.57 | 0.78 | 0.99 |
| F1-Score | 0.99 | 0.68 | 0.84 | 0.98 |
| Support | 817 | 23 | 840 | 840 |
| Accuracy | | 0.9857 | | |
| ROC AUC Score | | 0.9898 | | |
| Best Parameters: $\{C : 10$, class_weight: None, solver: 'lbfgs'$\}$ | | | | |

## 3.2 Random Forest Model

The results of our Random Forest model can be seen in Table 2. Similar to the Logistic regression model, this model has high precision and recall for data points that did not make the Hall of Fame. In addition, the model is slightly better at predicting a player making the HOF, with precision and recalls of 0.93 and 0.61, both of which are step ups from the logistic regression classifier. There are also minute improvements in accuracy and ROC AUC score. However, all of these improvements make sense, as random forests sacrifice interpretability in exchange for ability to more accurately predict on high dimensionality data.

## 3.3 Logistic Regression with Feature Selection

After both model tests, we performed a Logistic Regression with feature selection (via recursive feature elimination). As shown in Table 4, the feature selection narrowed down the number of covariates down to 10. Based on the performance metrics depicted in Table 3, there were relatively similar precision numbers as the previous two models; however, there was noticeable improvement in the recall for making HOF (0.70). As the recall measures how often our model correctly identifies true positives, this is a good sign, as we are more interested in predicting making the HOF rather than not having enough statistics to make the HOF. In

Table 2: Performance Metrics and Parameters of the Random Forest Model

| Metric | Reject HOF | Made HOF | Macro Avg | Weighted Avg |
|---|---|---|---|---|
| Precision | 0.99 | 0.93 | 0.96 | 0.99 |
| Recall | 1.00 | 0.61 | 0.80 | 0.99 |
| F1-Score | 0.99 | 0.74 | 0.87 | 0.99 |
| Support | 817 | 23 | 840 | 840 |
| Accuracy | | | 0.9881 | |
| ROC AUC Score | | | 0.9929 | |
| Best Parameters: {max_depth: None, min_samples_leaf: 2, min_samples_split: 5, n_estimators: 200} | | | | |

addition, it's encouraging that this logistic regression model has fewer covariates than the other two models but maintains relatively equal if not better accuracy and ROC AUC scores.

Furthermore, when examining the 10 features selected via RFE, we are able to gather conclusions regarding significant covariates. NBA Championships, All Star selections, and rebounding champion all seem to have p-values below the significance level of 0.05, indicating that they have statistically significant effects on making it into the Hall of Fame. The All star selection covariate is also particularly interesting, as this was the variable that EDA found to be the most correlated with HOF selection. According to our results, for every additional All-Star selection that a player receives, the odds of making the Hall of Fame increases by $e^{1.8146} = 6.14$. This is interesting because the more All Star selections that a player accumulates, the lesser impact it will have on their chances, which makes sense in the real world. After an NBA player has passed a certain baseline for ALl Star selections, further selections will not impact their career accolades that much. However, given our NBA knowledge, it is true that accumulating All Stars will definitely help a player's chances at making the Hall of Fame.

Table 3: Performance Metrics of Logistic Regression with Feature Selection

| Metric | Rejected HOF | Made HOF | Macro Avg | Weighted Avg |
|---|---|---|---|---|
| Precision | 0.99 | 0.89 | 0.94 | 0.99 |
| Recall | 1.00 | 0.70 | 0.85 | 0.99 |
| F1-Score | 0.99 | 0.78 | 0.89 | 0.99 |
| Support | 817 | 23 | 840 | 840 |
| Accuracy | | | 0.9983 | |
| ROC AUC Score | | | 0.9875 | |

Table 4: Logistic Regression Coefficients

| Variable | Coefficient | Std. Err | $z$-value | $P > \|z\|$ | 95% CI | |
|---|---|---|---|---|---|---|
| | | | | | [0.025 | 0.975] |
| Finals_MVP | 0.5479 | 1.207 | 0.454 | 0.650 | -1.818 | 2.914 |
| NBA_Champ | 0.6292 | 0.112 | 5.634 | 0.000 | 0.410 | 0.848 |
| Def_POY | 1.1655 | 2.887 | 0.404 | 0.686 | -4.493 | 6.824 |
| All_Star | 1.8146 | 0.093 | 19.512 | 0.000 | 1.632 | 1.997 |
| TRB_Champ | 2.8283 | 1.049 | 2.696 | 0.007 | 0.772 | 4.884 |
| BLK_Champ | 0.8202 | 0.691 | 1.186 | 0.236 | -0.535 | 2.175 |
| 3PA_per_game | 0.0022 | 0.000 | 9.138 | 0.000 | 0.002 | 0.003 |
| AST_per_game | -2.3094 | 0.112 | -20.631 | 0.000 | -2.529 | -2.090 |
| Position_Center | -3.9868 | 0.428 | -9.320 | 0.000 | -4.825 | -3.148 |
| Position_SmallForward | -2.9029 | 0.343 | -8.470 | 0.000 | -3.575 | -2.231 |

### 3.4 Limitations and Future Work

There are a few limitations in our analysis. First and foremost, we would like to mention that our data scraping was incomplete. Since the Naismith Basketball Hall of Fame encompasses all types of basketball (including overseas European, Olympic, women's, college basketball), there is not one standard for determining whether someone makes it or not. There could be drastically different standards between NBA and Europeann EuroLeague players. In our data scraping attempts, we were unable to find complete data across all basketball leagues. In our data, there are some European players who only played a few years in the NBA, yet got crowned in the HOF for their spectacular Olympics performances. If there were more time, we would like to analyze the different basketball leagues across the world for a more general study.

Another improvement could be scaling NBA statistics by time period. Basketball has been around for a long time, and the game has definitely changed. The statistics used back in the 1960s don't mean the same in the 2020s. For example, the rise in popularity of the three point shot has led league scoring averages and three-point percentages to go up, so a high scoring game in the 1960s may no longer be considered high scoring in the 2020s. I think it is important to scale certain statistics by the time period in which they were collected, because players make the HOF based on their performance during their time period. Doing this would require more extensive research into each time period and which stats to scale. It is definitely crucial for future works.

## 4 Conclusion

In this paper, we used logistic regression and random forest classifiers to help us model the probability a player will make the Hall of Fame as well as what specific career statistics are important. Both our data analysis and model results showed that there are definitely key predictors, such as All Star Selections and NBA championships.

# References

[1] URL https://www.basketball-reference.com/.

# 5 Appendix