NBA Hall of Fame Prediction

Justin Mai

2025-05-19

Contents

1 Abstract

2 Introduction

The NBA Hall of Fame inducts the most influential players, coaches, teams, and referees yearly. There have only been just over 150 NBA players inducted to the Hall of Fame which started with the inaugural class of 1959. So what classifies an NBA player as a Hall of Famer compared to other NBA players? This research paper will identify the probability of current NBA players one day making the Hall of Fame based on historical trends.

2.1 Accolades and Awards

A common debate within the basketball community is the infamous "LeBron vs. Michael Jordan" debate to crown on player as the Greatest of All Time (G.O.A.T.). Analysts will often start with the quantitative in game statistics by looking at all time averages for both players. Looking at the primary statistics, throughout Jordan's career, he averaged 30.1 points per game, 6.2 rebounds per game, and 5.3 assists per game. On the other end, LeBron averages 27.0 points per game, 7.5 rebounds per game, and 7.4 assists per game. While Jordan has the edge on scoring, LeBron has the edge in the other primary statistics so its difficult to make a clear conclusion based on this. However, this isn't the primary argument for both players, if you've ever been part of this debate you'll often hear the notion that "Jordan went 6 for 6 in championship games". The accolades and awards that each player compiles is often the primary argument.

While there is no clear calculator for identifying if a player will make the Hall of Fame, in all cases, accolades and awards will be a significant predictor to identifying HOFers. These accolades will consist of **Regular Season MVPs, Championship Wins, Finals MVPs, All-NBA Selections, All-Star Selections, End-of-Season Awards** and possibly much more. These awards all signify the impact that a player has had on their respective teams, demonstrating how their contribution leads to the team's success.

2.2 Player Impact / Other Considerations

Within the G.O.A.T debate, a primary argument for LeBron would be his longevity and consistent impact on the game and the teams he goes to, the qualitative factors that goes beyond the box score and award counts. While accolades and awards provide a quantitative summary of a player's career, qualitative aspects such as leadership, clutch performances, career longevity, influence on team culture, and global popularity often shape the broader legacy of a player.

For example, LeBron's ability to lead multiple franchises to the NBA Finals—winning championships with three different teams is a testament to his versatility and value as a player. Similarly, players like Allen Iverson and Vince Carter are celebrated not only for their statistics and accolades, but also for their cultural impact, influence on future generations, and overall contribution to the evolution of the game.

When making predictions for Hall of Fame inductees, there are also many traits outside of the box score that contributes to the selection process. However, this type of impact often correlates with the accolades and overall quantitative statistics so it will be contributing factor within the analysis. Something like longevity will also be taken into account through the number of years they played all together and the number of years they played for one team.

3 Methods

3.1 Data Collection

The data collected starts with all **5311** players who have played at least one game in the NBA since 1947 (when it was known as the BAA) to 2025. The data was collected by a Kaggle user named *Sumitro Datta* in the page **NBA Stats (1947-present)**. The data was gathered using IMPORTHTML from Google Sheets on Basketball-Reference's Play Index now known as **Stathead**.

3.2 Data Manipulation

The majority of our data manipulation process in creating the primary dataset we would use was done using Python and JupyterNotebook.

With the several datasets provided, there were different statistics that we required merging before curating our logistic regression model. We first looked at **player accolades** to get player award counts from the All-star, Awards, and End-Season-Teams tables. We turned each award into its own column so that every observation would be the number of that specific award a player won. We then combined these awards with the number of all-star games and number of seasons a player played in the NBA. This was all joined using a player's *player_id*

We then wanted to combine in-game player statistics with the accolades dataset. These statistics will consist of both total career statistics and player averages to give us flexibility in choosing our model. For many of the players, there were NA values for some statistics that weren't yet accounted for (such as 3 pointers and some awards). These NA values were replaced with 0.

```
df <- read.csv("data/final.csv")
```

```
train <- read.csv("data/train.csv")

train$mvp_flag = factor(ifelse(train$nba.mvp > 0, 1, 0))

test <- read.csv("data/test.csv")

test$mvp_flag = factor(ifelse(test$nba.mvp > 0, 1, 0))

# Removing irrelevant columns

train <- train %>%
  select(!c(aba.mvp, aba.roy, clutch_poy, All.ABA.1st, All.ABA.2nd, All.BAA.1st, All.BAA.2nd, n
```

```
train<- train %>%
  select(!c(orb, drb, orb_per_game, drb_per_game, pf, fg_per_game, fga_per_game, fg_percent, x2

cor_mat <- cor(train[sapply(train, is.numeric)], use = "complete.obs")

cor_high <- which(cor_mat > 0.85 & cor_mat < 1, arr.ind = TRUE)
```

```r
data.frame(
  var1 = rownames(cor_mat)[cor_high[, 1]],
  var2 = colnames(cor_mat)[cor_high[, 2]],
  correlation = cor_mat[cor_high]
)
```

```
## [1] var1        var2        correlation
## <0 rows> (or 0-length row.names)
```

```r
set.seed(123)
ctrl <- trainControl(method = "cv", number = 10, savePredictions = TRUE)

train_model <- train[, !names(train) %in% c("player_id", "player", "active_2025", "nba_mvp")]

M_full <- train(
  hof ~ .,
  data = train_model,
  method = "glm",
  family = "binomial",
  trControl = ctrl
)

M_full$resample
```

```
##          RMSE  Rsquared        MAE Resample
## 1  0.14224995 0.5787346 0.03045210   Fold01
## 2  0.12069719 0.4977208 0.02332446   Fold02
## 3  0.12398488 0.5263507 0.02585886   Fold03
## 4  0.11122066 0.5876602 0.02406121   Fold04
## 5  0.10731370 0.6077575 0.02192303   Fold05
## 6  0.10368447 0.7071084 0.02408266   Fold06
## 7  0.10148168 0.7521662 0.02020384   Fold07
## 8  0.10251668 0.6048304 0.02143543   Fold08
## 9  0.08825464 0.7611934 0.01936756   Fold09
## 10 0.11762392 0.5535615 0.02411036   Fold10
```

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

```r
summary(M_full)
```

```
##
## Call:
## NULL
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.309e+01  1.952e+00  -6.706 2.00e-11 ***
## num_all_star     4.814e-01  9.435e-02   5.102 3.36e-07 ***
## num_season       1.812e-01  8.048e-02   2.251  0.02438 *
## dpoy            -5.195e-04  4.791e-01  -0.001  0.99913
## mip             -1.086e+00  1.370e+00  -0.793  0.42795
## nba.mvp          9.740e+00  8.431e+02   0.012  0.99078
## nba.roy          5.419e-03  6.824e-01   0.008  0.99366
## smoy             1.092e+00  6.104e-01   1.790  0.07353 .
## All.Defense.1st  6.203e-01  2.051e-01   3.025  0.00249 **
## All.Defense.2nd -4.735e-01  2.874e-01  -1.648  0.09942 .
## All.NBA.1st      5.963e-01  2.973e-01   2.005  0.04492 *
## All.NBA.2nd      6.674e-01  2.199e-01   3.036  0.00240 **
## All.NBA.3rd      5.505e-01  3.246e-01   1.696  0.08992 .
## All.Rookie.1st   6.361e-01  4.095e-01   1.553  0.12038
## All.Rookie.2nd   1.604e+00  6.985e-01   2.296  0.02168 *
## trb             -1.945e-04  1.369e-04  -1.421  0.15538
## ast             -3.341e-04  3.070e-04  -1.088  0.27649
## stl             -3.236e-04  1.131e-03  -0.286  0.77482
## g                4.099e-02  1.636e-02   2.506  0.01219 *
## gs              -1.810e-02  1.329e-02  -1.362  0.17317
## x3p_percent_y   -3.981e+00  2.119e+00  -1.879  0.06030 .
## ft_percent       3.869e+00  2.329e+00   1.661  0.09664 .
## trb_per_game     1.832e-01  1.263e-01   1.450  0.14704
## ast_per_game     5.935e-01  2.690e-01   2.206  0.02737 *
## stl_per_game    -1.200e-01  9.524e-01  -0.126  0.89970
## blk_per_game    -8.240e-03  5.161e-01  -0.016  0.98726
## tov_per_game    -5.013e-01  2.805e-01  -1.788  0.07386 .
## pf_per_game      8.403e-01  3.020e-01   2.783  0.00539 **
## pts_per_game    -8.669e-03  5.770e-02  -0.150  0.88057
## mvp_flag1       -7.472e+00  8.431e+02  -0.009  0.99293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1420.92  on 4767  degrees of freedom
## Residual deviance:  410.58  on 4738  degrees of freedom
## AIC: 470.58
##
```

```
## Number of Fisher Scoring iterations: 17
```

4 Results

5 Discussion

5.1 Limitations

5.2 Next Steps

6 Bibliography

https://www.sportingnews.com/us/nba/news/michael-jordan-vs-lebron-james-goat-debate/sl8xdozy5u1m1s4t5m3npeqo1

https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats?select=Player+Career+Info.csv