NBA Hall of Fame Prediction

Justin Mai

2025-05-20

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.

3.3 Data Transformation

```
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

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

```r
train<- train %>%
  select(!c(orb, drb, orb_per_game, drb_per_game, pf, fg_per_game, fga_per_game, fg_percent, x1

test <- test %>%
  select(!c(orb, drb, orb_per_game, drb_per_game, pf, fg_per_game, fga_per_game, fg_percent, x1

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

```
## Warning in cor(train[sapply(train, is.numeric)], use = "complete.obs"): the
## standard deviation is zero
```

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

```r
cols_to_exclude <- c("player_id", "player", "mvp_flag", "active_2025")

cols_to_sum <- setdiff(names(test), cols_to_exclude)

test <- test %>%
  group_by(player) %>%
  summarise(across(all_of(cols_to_sum), sum, na.rm = TRUE), .groups = "drop") %>%
  left_join(distinct(test[, cols_to_exclude]), by = "player")
```

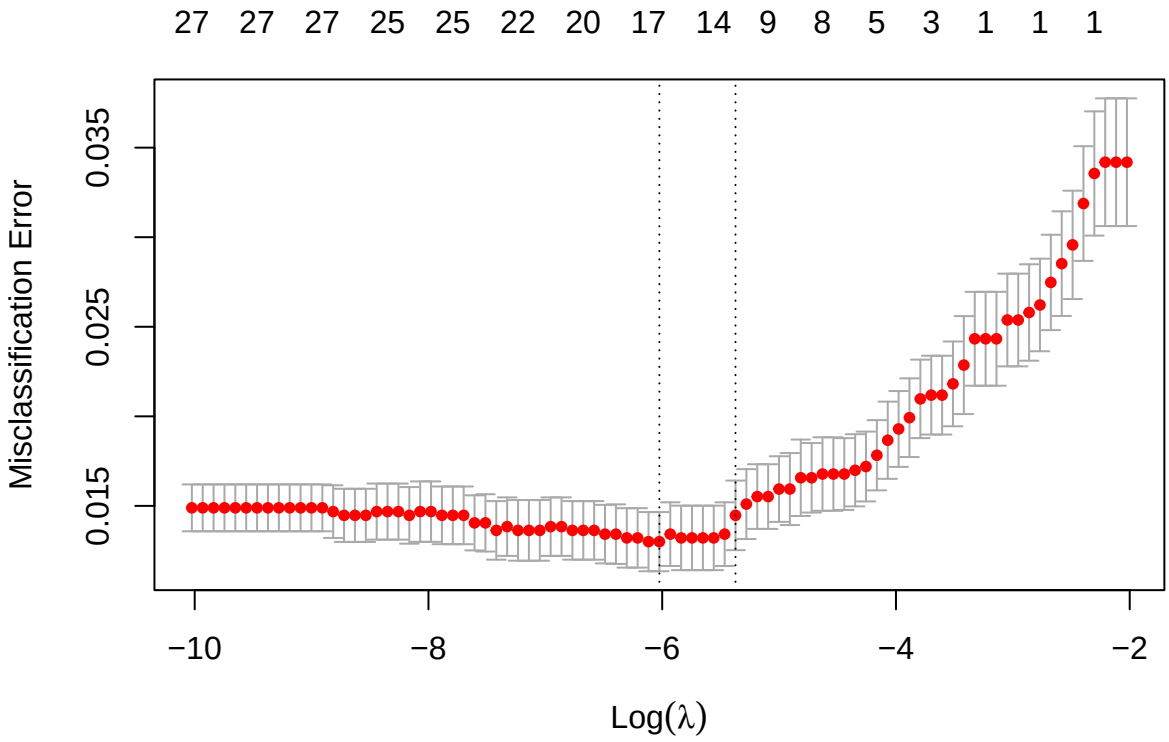
```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(all_of(cols_to_sum), sum, na.rm = TRUE)`.
## i In group 1: `player = "A.J. Green"`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```r
test <- test
```

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

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

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




```
## [1] 0.002419841
```

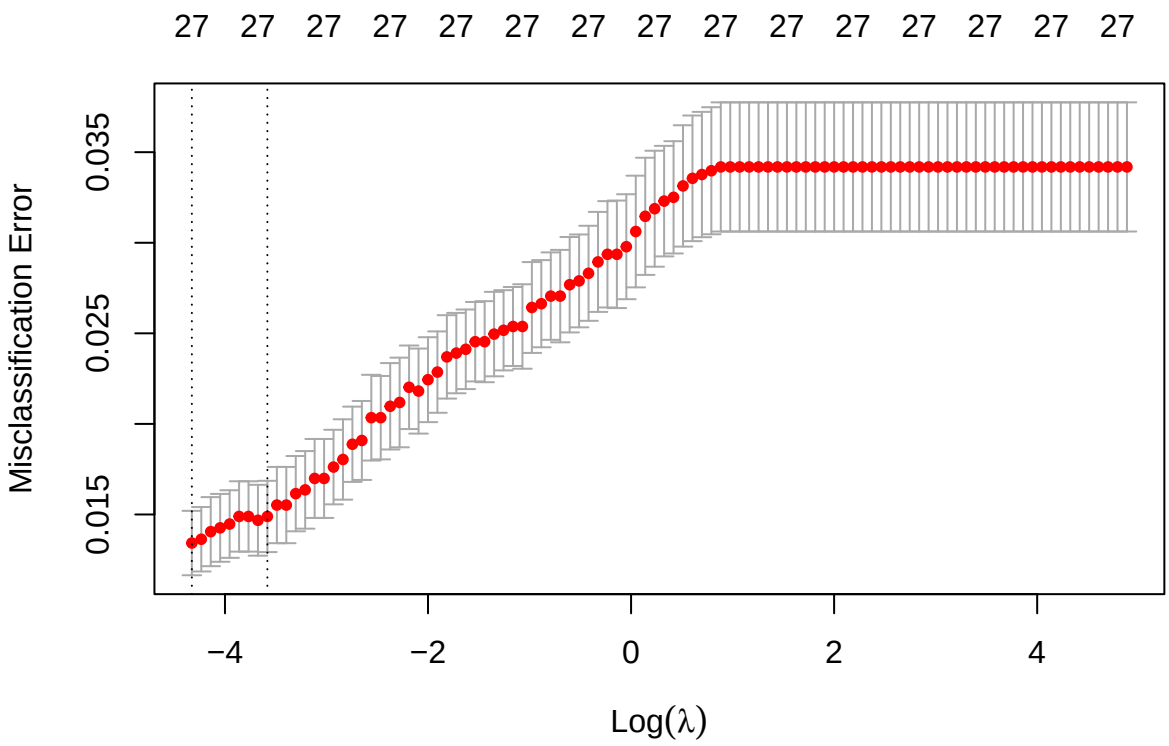
Table 1: LASSO Model Coefficients

Variable	Coefficient
(Intercept)	-7.516
num_all_star	0.554
num_season	0.000
dpoy	0.108
mip	-0.080
nba.roy	0.000
smoy	0.590
All.Defense.1st	0.362
All.Defense.2nd	-0.109
All.NBA.1st	0.123
All.NBA.2nd	0.434
All.NBA.3rd	0.000
All.Rookie.1st	0.010
All.Rookie.2nd	0.000
trb	0.000
ast	0.000

Variable	Coefficient
stl	0.000
g	0.016
gs	-0.007
ft_percent	0.000
trb_per_game	0.033
ast_per_game	0.240
stl_per_game	0.000
blk_per_game	0.000
tov_per_game	-0.628
pf_per_game	0.626
pts_per_game	0.081
mvp_flag1	1.281

Table 2: Ridge Model Coefficients

Variable	Coefficient
(Intercept)	-7.187
num_all_star	0.348
num_season	0.030
dpoy	0.205
mip	-0.858
nba.roy	-0.059
smoy	0.598
All.Defense.1st	0.386
All.Defense.2nd	-0.186
All.NBA.1st	0.294
All.NBA.2nd	0.571
All.NBA.3rd	0.231
All.Rookie.1st	0.360
All.Rookie.2nd	0.293
trb	0.000
ast	0.000
stl	0.000
g	0.012
gs	-0.013
ft_percent	0.412
trb_per_game	0.087
ast_per_game	0.211
stl_per_game	-0.310
blk_per_game	-0.067
tov_per_game	-0.387
pf_per_game	0.350
pts_per_game	0.074
mvp_flag1	1.325




```
## [1] 0.01321784
```

```
cv_lasso$cvm[cv_lasso$lambda == cv_lasso$lambda.min]
```

```
## [1] 0.01300336
```

```
cv_ridge$cvm[cv_ridge$lambda == cv_ridge$lambda.min]
```

```
## [1] 0.01342282
```

4 Results

```
x_test <- model.matrix(~ ., data = test_model)[, -1]
```

```
pred <- predict(cv_lasso, newx = x_test, s = "lambda.min", type = "response")

test_with_preds <- test %>%
  mutate(
```

```
    hof_prob = as.vector(pred)
  )

test_with_preds <- test_with_preds %>%
  select(player, hof_prob) %>%
  distinct(player, .keep_all = TRUE) %>%
  arrange(desc(hof_prob))
```

```
test_with_preds
```

```
## # A tibble: 563 x 2
##    player                hof_prob
##    <chr>                    <dbl>
##  1 LeBron James             1.00
##  2 Chris Paul               1.00
##  3 Kevin Durant             1.00
##  4 Giannis Antetokounmpo    0.995
##  5 Stephen Curry            0.995
##  6 Russell Westbrook        0.992
##  7 James Harden             0.989
##  8 Anthony Davis            0.949
##  9 Nikola Jokić             0.948
## 10 Damian Lillard           0.935
## # i 553 more rows
```

5 Discussion

No Rudy Gobert

5.1 Limitations

5.2 Next Steps

6 Appendix

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