The Proposal of Mid-Semester Project

Student: Zheng Fang and Pengfei Yang

DA 210-02 / CS 181-02: Data Systems

Spring 2023

Instructor: Dr. Tanya Amert

**The Central Question:**

Which player not inducted into the NBA Hall of Fame between 1999 and 2016 is potential to become a member of Hall of Fame in the future?

By using three NBA career statistical databases, establish a model and discuss the following questions:

(1) Identify common traits among players who have been inducted into the NBA Hall of Fame.

(2) Predict whether players who have not yet been inducted into the Hall of Fame will be inducted in the future using this model.(Rank players who have not yet been inducted into the NBA Hall of Fame.)

(3) Evaluate the accuracy of this model.

**Data Sets:**

We will utilize three NBA player career statistical datasets:

1. NBA Rookies Performance Statistics and Minutes[[1]](#footnote-1)
   1. In this dataset, we have two CSV files and we will use one of them named " NBA Rookies by Year\_Hall of Fame Class.csv".
   2. In this CSV file, there are 22 columns which are:

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Introduction | Variable Type |
| Name | The name of the rookie | String |
| Year Drafted | The year the rookie was drafted | Integer |
| GP | The number of games played by the rookie | Integer |
| MIN | The number of minutes played by the rookie | Integer |
| PTS | The number of points scored by the rookie | Integer |
| FGM | The number of field goals made by the rookie | Integer |
| FGA | The number of field goals attempted by the rookie | Integer |
| FG% | The field goal percentage of the rookie | Float |
| 3P Made | The number of three pointers made by the rookie | Integer |
| 3PA | The number of three pointers attempted by the rookie | Integer |
| 3P% | The three point percentage of the rookie | Float |
| FTM | The number of free throws made by the rookie | Integer |
| FTA | The number of free throws attempted by the rookie | Integer |
| FT% | The free throw percentage of the rookie | Float |
| OREB | The number of offensive rebounds by the rookie | Integer |
| DREB | The number of defensive rebounds by the rookie | Integer |
| OREB | The number of offensive rebounds by the rookie | Integer |
| AST | The number of assists by the rookie | Integer |
| STL | The number of steals by the rookie | Integer |
| BLK | The number of blocks by the rookie | Integer |
| TOV | The number of turnovers by the rookie | Integer |
| EFF | The efficiency rating of the rookie | Float |

1. NBA Players stats since 1950[[2]](#footnote-2)
   1. In this dataset, there are three CSV files and we will use one of them named " Seasons\_Stats.csv".
   2. In this CSV file, there are 53 columns which are:

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Introduction | Variable Type |
| Year | Season | Integer |
| Player | Name | String |
| Pos | Position | String |
| Age | Age | Float |
| Tm | Team Name | String |
| G | The number of games played | Integer |
| GS | The number of games Started | Integer |
| MP | Minutes Played | Float |
| PER | Player Efficiency Rating | Float |
| TS% | True Shooting % | Float |
| 3PAr | 3-Point Attempt Rate | Float |
| FTr | Free Throw Rate | Float |
| ORB% | Offensive Rebound Percentage | Float |
| DRB% | Defensive Rebound Percentage | Float |
| TRB% | Total Rebound Percentage | Float |
| AST% | Assist Percentage | Float |
| STL% | Steal Percentage | Float |
| BLK% | Block Percentage | Float |
| TOV% | Turnover Percentage | Float |
| USG% | Usage Percentage | Float |
| blanl | empty | Float |
| OWS | Offensive Win Shares | Float |
| DWS | Defensive Win Shares | Float |
| WS | Win Shares | Float |
| WS/48 | Win Shares Per 48 Minutes | Float |
| blank2 | empty | Float |
| OBPM | Offensive Box Plus/Minus | Float |
| DBPM | Defensive Box Plus/Minus | Float |
| BPM | Box Plus/Minus | Float |
| VORP | Value Over Replacement | Float |
| FG | Field Goals | Float |
| FGA | Field Goal Attempts | Float |
| FG% | Field Goal Percentage | Float |
| 3P | 3-Point Field Goals | Float |
| 3PA | 3-Point Field Goal Attempts | Float |
| 3P% | 3-Point Field Goal Percentage | Float |
| 2P | 2-Point Field Goals | Float |
| 2PA | 2-Point Field Goal Attempts | Float |
| 2P% | 2-Point Field Goal Percentage | Float |
| eFG% | Effective Field Goal Percentage | Float |
| FT | Free Throws | Float |
| FTA | Free Throw Attempts | Float |
| FT% | Free Throw Percentage | Float |
| ORB | Offensive Rebounds | Float |
| DRB | Defensive Rebounds | Float |
| TRB | Total Rebounds | Float |
| AST | Assists | Float |
| STL | Steals | Float |
| BLK | Blocks | Float |
| TOV | Turnovers | Float |
| PF | Personal Fouls | Float |
| PTS | Points | Float |

1. NBA Playoffs Player Statistics 1950-2022[[3]](#footnote-3)
   1. In this dataset, there is one CSV file and we will use it named "playoffStats 2.csv"
   2. In this CSV file, there are 51 columns which are:

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Introduction | Variable Type |
| season | NBA Season. 2022 would represent the 2021-2022 season. | Integer |
| player | Player name | String |
| pos | Player position | String |
| age | Player Age | Integer |
| team\_id | Player team | String |
| g | Number of playoff games in season played | Integer |
| gs | Number of playoff games started in season | Integer |
| mp\_per\_g | Average minutes played | Float |
| fg\_per\_g | Average field goals made | Float |
| fga\_per\_g | Average field goals attempted | Float |
| fg\_pct | Average field goal percentage | Float |
| fg3\_per\_g | Average three point shots made | Float |
| fg3a\_per\_g | Average three point shots attempted | Float |
| fg3\_pct | Average three point percentage | Float |
| fg2\_per\_g | Average two point shots made | Float |
| fg2a\_per\_g | Average two point shots attempted | Float |
| fg2\_pct | Average two point showing percentage | Float |
| efg\_pct | Effective shooting percentage | Float |
| ft\_per\_g | Free throws made per game | Float |
| fta\_per\_g | Free throws attempted per game | Float |
| ft\_pct | Free throw percentage per game | Float |
| orb\_per\_g | Offensive rebounds per game | Float |
| drb\_per\_g | Defensive rebounds per game | Float |
| trb\_per\_g | Total rebounds per game | Float |
| ast\_per\_g | Assists per game | Float |
| stl\_per\_g | Steals per game | Float |
| blk\_per\_g | Blocks per game | Float |
| tov\_per\_g | Turnovers per game | Float |
| pf\_per\_g | Personal fouls per game | Float |
| pts\_per\_g | Points per game | Float |
| ast\_pct | Assist percentage per game | Float |
| blk\_pct | Block percentage per game | Float |
| bpm | Box plus minus | Float |
| dbpm | Defensive box plus minus | Float |
| drb\_pct | Defensive rebounding percentage | Float |
| dws | Defensive win share | Float |
| fg3a\_per\_fga\_pct | Three point shot attempts per field goal attempted | Float |
| fta\_per\_fga\_pct | Free throw attempted per field goal attempted percentage | Float |
| mp | Total minutes played | Float |
| obpm | Offensive box plus minus | Float |
| orb\_pct | Offensive rebounding percentage | Float |
| ows | Offensive win share | Float |
| per | Player Efficiency Rating | Float |
| stl\_pct | Steal percentage | Float |
| tov\_pct | Turnover percentage | Float |
| trb\_pct | Total rebound percentage | Float |
| ts\_pct | True shooting percentage | Float |
| usg\_pct | Usage percentage | Float |
| vorp | Value Over Replacement Player | Float |
| ws | Win Share | Float |

**Outline and Functions of the Project:**

Step 1: Establish a function called "Hall of Fame Member Differentiation" according to “NBA Rooks Year\_ Hall of Fame Class.csv.” Use the function to distinguish the players who have entered the Hall of Fame between 1999 and 2016 and those who have not been included in the Hall of Fame between 1999 and 2016 and establish two lists for these two types of players.

Step 2: Establish a function called "Search for Hall of Fame data and establish career files." According to the list of players who have been selected in the Hall of Fame in Step 1, find the regular season and playoff game average data of each Hall of Fame player in the regular season database (Seasons\_Stats. csv) and the playoff season game average database (playoffStats. csv). This function should finally return a detailed player career profile, including which seasons each player participated in and the regular season and playoff performance of each participating season (if entering the playoffs).

Step 3: Establish a function called "Common characteristics of Hall of Fame players." Through the career profile of each player in Step 2, compare the detailed data in the game performance to find the median, lower quartile, and upper quartile of these technical statistics (for example, the median, lower quartile, and upper quartile of scoring of Hall of Fame players from 1999 to 2017 are: ...). From this, we can get the technical statistic range of players who we think can enter the Hall of Fame (for example, the score range is between 21 and 35, and the field goal percentage is between 51% and 65%) and establish a data model of potential Hall of Fame members to match the actual players.

Step 4: Establish a function called "Search for potential Hall of Fame players," and search for qualified players in the list of players who failed to be included in the Hall of Fame established in step 1 with two databases (Seasons\_Stats.csv and playoffStats.csv) through the data model in step 3.

Step 5: Establish a function called "Rank the candidates" and judge the probability of entering the Hall of Fame by the coincidence of the data of these players and the model of potential Hall of Fame members. For example, if all the data of Player 1 fall into all numerical ranges of potential Hall of Fame players, his selection priority will be the highest. Rank all the players in the list who failed to be included in the Hall of Fame to find out which players are most likely to enter the Hall of Fame after 2017.

Step 6: Based on the actual situation of the basketball hall of fame from 2018 to 2022, check whether our prediction is reasonable, and give feasible optimization plans to improve our prediction.

1. <https://www.kaggle.com/datasets/thedevastator/nba-rookies-performance-statistics-and-minutes-p> [↑](#footnote-ref-1)
2. <https://www.kaggle.com/datasets/drgilermo/nba-players-stats?select=Seasons_Stats.csv> [↑](#footnote-ref-2)
3. <https://www.kaggle.com/datasets/robertsunderhaft/nba-playoffs> [↑](#footnote-ref-3)