Final Porject Proposal - NBA Players’ Salary Prediction

Group 4 Yuka Chen, Jack Lo, Adela Yang

2022-03-23

```
library(tidyverse)
library(broom)
```

1. Problem Statement:

Many basketball players want to play in the National Basketball Association (NBA) because of the high salary. NBA players are the highest-paid athletes on average in the world. According to Statista.com, The average salary in the NBA for 2021-2022 season is 7.3 million and the top ten highest salaries for NBA players in the season are all over 39 million U.S. dollars, including superstars Stephen Curry from the Golden State Warriors,Kevin Durant from the Brooklyn Nets, and LeBron James from the Los Angeles Lakers.

Since of one our group members is a super fan of NBA, he believes that those basketball players are paid by their season performance and they were not overpaid. However, other members in our group think otherwise.

Through this project, we want to find out whether these NBA players were overpaid, whether majority of players have salary increased each contract they signed or it was only for a few specific players who have well performance, as well as to see whether their salary were higher each year due to the inflation in the US dollar currency.

We want to create a prediction model for NBA player’s salary and help the players and NBA team to predict their annual salary by their seasonal performance, such as scoring, rebound, field goal percentage etc.

The outcome we expect from this proect is that not only high efficient scorers and all-star players get high paid, but also some specific roles of the teams get highly paid. The possible reason is that there are lots of different and tough tasks need to be finished on court. For instance, the mission for a great point guard is high assist/turnover ratio and free throw percentage. Secondly, the mission for a great center is as much as rebounds he can grab and how many blocks he can reach. Thirdly, the mission for a great shooter is nearly 40 percentage of three-point. Last but not least, the mission for a great forward is how comprehensive his stats is.

Purpose of this project:

1. Discover which statistics are the best predictors of an NBA player’s salary
2. Use a multiple regression model to predict NBA salaries
3. Determine which players have been overvalued and undervalued according to their given vs. predicted salary

Potential problems: The salary glowing may because inflation and increased business value due to globalization.

2. Identify a data set that is relevant to your question:

The two data set were collected from Basketball Reference (https://www.basketball-reference.com/) by Chris Davis (https://data.world/datadavis).

- Data set assess link: https://data.world/datadavis/nba-salaries

The first data set includes the information of NBA players backgroun information such as ID, Name, DOB, weights, heights, position, shoots etc, as well as their career performance.

```
players <- read_csv("players.csv") ##Rmarkdown and dataset are saved in same folder
## 24 variables and 4584 observations; 20 characters and 4 doubles variables
#spec(players) #to check variable types
names(players) #to check all variable names
```

```
##  [1] "_id"         "birthDate"   "birthPlace"  "career_AST"  "career_FG%"
##  [6] "career_FG3%" "career_FT%"  "career_G"    "career_PER"  "career_PTS"
## [11] "career_TRB"  "career_WS"   "career_eFG%" "college"     "draft_pick"
## [16] "draft_round" "draft_team"  "draft_year"  "height"      "highSchool"
## [21] "name"        "position"    "shoots"      "weight"
```

The second data set includes the playerID, team name, season periods (year) and their salaries the 1984-1985 season to the 2017-2018 season.

```
salaries <- read_csv("salaries_1985to2018.csv") #same folder
## 7 variables and 14163 observations ; 4 characters and 3 double variables
#spec(salaries) #to check variable types
names(salaries) #to check all variable names
```

```
## [1] "league"       "player_id"    "salary"       "season"       "season_end"
## [6] "season_start" "team"
```

3. Proposed Method of Analysis:

1. we would combine two data set into one.
2. data exploratory analysis for variables visualization and see if there are any potential linear relationship or outlinears for the models that we are interested
3. to verify the linear relationship by t.test and other statistical information
4. using simple linear relationship to create a model for predicting salaries
5. using multiple linear relationship to see if there are other variables hae stronger relationship with salaries than the ones we expected
6. stepwise regression to find the best fited prediction variables for our models
7. visualization for our models