Salary prediction of NBA Players based on 2016-2017 Season Performance using R *

Connor Cocklin †

May 9, 2019

Abstract

This project gathered performance data from the NBA 2016-2017 NBA season and used Linear Regression modelling to predict salary and identify undervalued players. Research questions include: What statistics have the greatest impact on predicted salary? What players are among the most underpaid according to the predictive model? Which of these players add the most value to their team by measure of Win Shares. The linear model used to predict salary included a variety of metrics and resulted in an $R^2 of .5$.

Culminating factors outside of performance explain the rest of the variance that the model did not. These include salary cap restrictions, years in the league, amount of all NBA appearances, Public Opinion, and free market factors and position demand at times of signings. A further investigation should look into how much revenue a player is predicted to bring an organization. This investigation would essentially look into the marginal revenue product of the NBA Player and would be free of the culminating factors outside of the relationship between salary and performance.

*Thank you to Dr. Tyler Ransom, who has challenged and taught me more than I could imagine as it relates to coding and Data Science. Thank you to Mark Wissler, who's analytical thinking helped me organize my thoughts regarding this project and how to process the results. Thank you to Dr. Daniel Larson, who has been a mentor to me and my graduate liaison, who without, I would not be able to navigate the graduate program

†Department of Health and Exercise Science, Sports Data Analytics Lab, University of Oklahoma. E-mail address: connorcocklin@ou.edu

1 Introduction

This topic peaked my interest as a result of my curiosity into the financial workings of the NBA and other professional sports leagues. There is a real need by sport organizations to be able to quantify both a players worth in value added by revenue brought to the organization and also a fair salary that fits within the predetermined and agreed salary cap restricting all teams within the league. NBA teams are not able to pay their players what they bring in as fair as revenue because then there would be no marginal product (profit) generated to keep the organization wheel turning. When determining salary there are many spokes with the wheel of the decision making process. Understanding how to quantify performance allows NBA organizations to build a team that maximizes wins while remaining under the salary cap restriction. The salary cap restriction for the 2016-2017 NBA season was set at $94.14 million$. With 12 active players this is an average of 7.8 million per player. Performance by the players is not the only determinant when it comes to compensation as this will be showcased later when a performance based model is unable to completely explain variance within actual salary paid to the players. Other work done by researchers have created other models which include factors outside of performance and in addition to performance measures to explain salary to a more successful model. This predictive model is the first portion of this project.

In keeping with the theme of playing GM, this project will attempt to identify underpaid players among the NBA in the 2017 season and determine which of these players will maximize wins when compared to their counterparts of the same position. The final result will be a team built to be as cheap as possible while maximizing wins. This project will showcase the analytical power of R as well as showcase the thought process of front offices within sport organizations in regards to deciding on team composition. There are many obstacles in choosing the right player for the best price for the organization and this will show a small glimpse into the issues that may arise when undertaking such an action. Unfortunately performance is not the end all be all for determining what a player should be paid when the general public might understand that to be the case, performance however is a major factor in that determination and in no way will hurt a player's justification for

compensation. This data project hopes to build upon what many authors have contributed in the past and on data sets provided for public use through Kaggle by Omri GoldStein. This analysis was performed using RStudio desktop [**?**].

2 Literature Review

2.1

There have been a multitude of papers looking at performance and compensation for professional athlete's. Many other factors exist however, in [**?**] we see how fans and their racial bias's might play a role in compensation. The paper explores whether fans are a source of wage discrimination against African American players in the NBA. A multiple regression analysis found that black players are compensated fourteen to sixteen percent less than white players with comparable on-court performance. They also found that there exists statistically significant sorting by race based geographically but found that fan attendance does not move inversely of black players percentage of playing time. The authors of this study concluded that because white fans tend to want to see white players on the court there is an over representation of white player in large population areas composed of a relatively high proportion of white fans. This and the compensation of comparable black NBA players lead the authors to believe that there is a bias among the salary structure in the NBA. I believe this paper is important as it assesses factors outside of the players control. The players on court have the ability to control their performance. Ultimately in Utopian NBA performance would be the only factor deciding how to adequately compensate a player.

2.2

In the paper [**?**] the authors take a look at performance based measures to evaluate NBA Salary. The season data collected to be analyzed was from the 2013-2014 NBA season and included 243 NBA players. The authors hypothesized that scoring parameters such as points per game, field goal percentage, free throw, and three point percentage would be huge contributors towards determining salary. After they ran their analyses they found that not all their scoring measures were contributors to the extent they believed. Points per game, rebounds, and personal fouls were found to be among the stronger contributors towards determining NBA salary.

3 Data

The primary data set used was obtained through the public source Kaggle. Other sources of data were scraped from the basketball reference web page. Data was cleaned in R and merged together in order to produce a master table and add flexibility.

3.1 scraping, merging and cleaning the data

Two data sets were downloaded from Kaggle as csvs and read into R:

```
#reading in Data
seasonstats <- read.csv(file = "C:\\Users\\conno\\Desktop\\Data Science Class
    \\Final Project\\Seasons_Stats.csv", header = TRUE, sep = ",")

# Get salary dataset
#Reading in salary dataset for year end 2017.
salary.table <- read.csv(file = "C:\\Users\\conno\\Desktop\\Data Science Class
    \\Final Project\\Salary1617.csv")

#cleaning data a tad, getting stats for year ending 2017.
stats17 <-
  seasonstats %>% filter(Year >= 2017) %>%
  select(Year:G, MP, PER, FG:PTS) %>%
  distinct(Player, .keep_all = TRUE) %>%
  mutate(MPG = MP/G, PPG = PTS/G, APG = AST/G,
         RPG = TRB/G, TOPG = TOV/G, BPG = BLK/G,
```

```
          SPG = STL/G,x3PaG = X3PA/G)

#merging salary and stats into one table

Salary_Stats_2017 <- merge(stats17, salary.table,
                           by.x = "Player", by.y = "Player")
names(Salary_Stats_2017)[41] <- "Salary17"
Salary_Stats_2017<- Salary_Stats_2017[-39]
Salary_Stats_2017<- Salary_Stats_2017[-39]

#Reading in Advanced Statistics to join the party
#Get Advanced stats dataset
page <- read_html("https://www.basketball-reference.com/leagues/
    NBA_2017_advanced.html")
AdvancedStats.Table <- page %>% html_table(header = FALSE) %>% extract2(1)


#fixing headers and columns
names(AdvancedStats.Table) <- AdvancedStats.Table[1,]
AdvancedStats.Table <- AdvancedStats.Table[-1,]
AdvancedStats.Table <- AdvancedStats.Table[ , !names(AdvancedStats.Table)
                                            %in% c("NA", "Tm", "PER", "MP", "
    Age")]

AdvancedStats.Table <- AdvancedStats.Table %>% filter(Rk!="Rk")


#changing column type to Numeric instead of Character
AdvancedStats.Table <- as.data.frame(AdvancedStats.Table)
AdvancedStats.Table <- as_tibble(AdvancedStats.Table)

AS1 <- AdvancedStats.Table %>% select(Rk, Player, Pos)
As2 <- AdvancedStats.Table %>% select(-Rk, -Player, -Pos)


As2 <- As2 %>% mutate_if(is.character,as.numeric)
AdvancedStats.Table <- bind_cols(AS1,As2)

AdvancedStats.Table = AdvancedStats.Table[!duplicated(AdvancedStats.
    Table$Player),]


#Merging data into Master data set

Master_Salary_Stats_2017 <- merge(Salary_Stats_2017, AdvancedStats.Table,
                           by.x = "Player", by.y = "Player")
```

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl

hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

4 Empirical Methods

I used linear regression to predict salary and included variables such as Age, PPG, MPG, Etc. to be able to predict salary with performance measures. Many formulas that I modeled were resulting in large bouts of multicollinearity which is explained by performance enough. If I use an advanced metric that evaluates scoring as a factor in the overall metric then that will cause collinearity with lower level scoring measures such as points per game.

$$Y_{it} = \alpha_0 + \alpha_1 Z_{it} + \alpha_2 X_{it} + \varepsilon, \tag{1}$$

where Y_{it} is a continuous outcome variable for unit i in year t, and Z_{it} are characteristics about the firm at which i is working, while X_{it} are characteristics about i. The parameter of interest is α_1.

5 Research Findings

The main results are reported in Table 2. There the most significant factors are shown. One could argue for the sake of a more efficient model that some factors be dropped but in an effort of transparency and explanatory power I have left them in for the predictive model.

6 Conclusion

From these findings it is suggested that these players are the winningest and most affordable in the league at the present time. Further speculation would suggest that these individuals would be fairly compensated on their next contract in acknowledgment of their ability on the court.

References

Figures and Tables

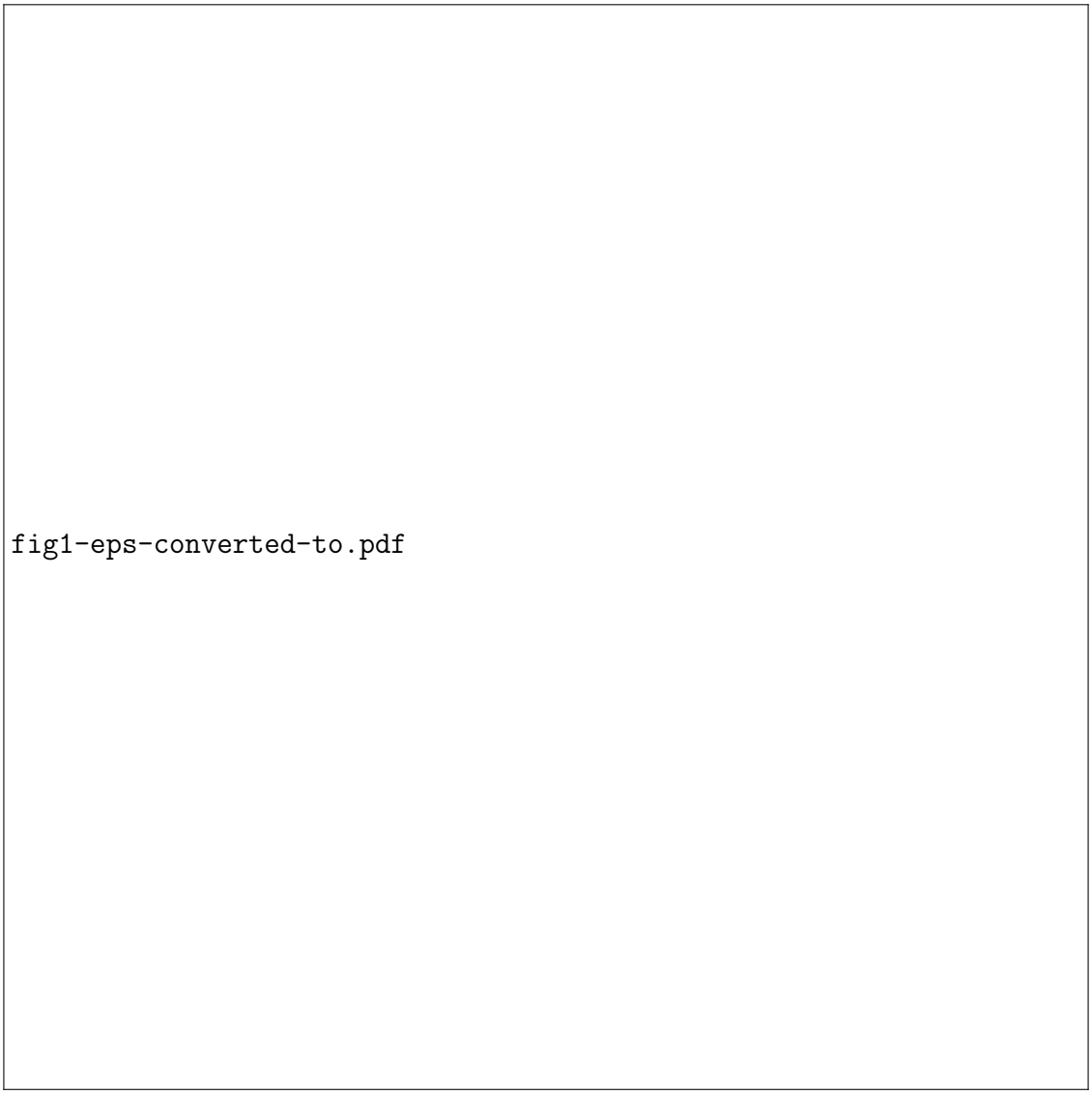

Figure 1: Figure caption goes here

References

Table 1: Summary Statistics of Variables of Interest

Panel A: Summary Statistics for Variables of Interest

	Mean	Std. Dev.	Min	Max
Outcome variable 1	4.127	1.709	0.000	8.516
Outcome variable 2	1.293	0.648	0.000	0.216
Policy variable	0.685	0.464	0.000	1.000
Control variable 1	0.451	0.497	0.000	1.000
Control variable 2	0.322	0.467	0.000	1.000

Panel B: Sample Means of Outcome Variables for Subgroups

	Group 1	Group 2	Group 3	Group 4
Outcome variable 1	1.782	2.181	3.749	4.127
Outcome variable 2	0.824	0.971	1.215	1.693
N	25,796	75,879	37,157	33,839

Notes: Put any notes about the table here. Sample size for all variables in Panel A is $N = 172,671$.

Table 2: Empirical estimates of parameter of interest

	Few Controls	Many Controls
Variable of interest	-1.977***	-0.536**
	(0.219)	(0.214)
Individual characteristics	✓	✓
Firm characteristics		✓
Location dummies		✓
N	172,671	172,671

Notes: Table notes here. Standard errors in parentheses. ***Significantly different from zero at the 1% level; **Significantly different from zero at the 5% level.