Code

Can you predict NBA three point percentage with college stats?

Introduction

In this project, I would like to answer the question “can you predict NBA three point percentage with college stats?” More explicitly, I will tackle this using college 3P% and college FT%. Analytically-minded talking heads for the NBA have been preaching the importance of not only college 3P% in predicting future NBA 3P$, but also college FT%.

The goal is to create a multiple regression model with these two explanatory variables to answer if these metrics are a good way to measure future success shooting from long range in the NBA.

Web Scraping

To start, I will be using basketball-reference.com (https://www.basketball-reference.com/) and their fantastic database of basketball data.

The methodology will be to look at all the players who have played between 2010-2024 and return a link to their college basketball page on sports-reference (https://www.sports-reference.com/cbb) and then their total college stats. We will also return each players total stats within the NBA between these years.

College Stats Web Scrape

Hide

```
start_year <- 2010
end_year <- 2024
vector_years <- start_year:end_year
years <- end_year-start_year+1

datalist = list()
datalist = vector("list", length = years)

#for loop to gather all college basketball links for each player
for (year in vector_years) {

  total_year_paste <- paste("https://www.basketball-reference.com/leagues/NBA_",year,"_totals.html", sep =
"")

  total_year_html <- read_html(curl::curl(total_year_paste))

  total_year_table <- total_year_html %>%
    html_table()

  total_year_tibble <- total_year_html %>%
    html_elements('a') %>%
    html_attr('href') %>%
    data.frame() %>%
    as.tibble()

  colnames(total_year_tibble) <- c("link")

  total_year_tibble <- total_year_tibble %>%
    filter(grepl('players/', link), grepl('.html', link))

  datalist[[year]] <- total_year_tibble

}

data <- bind_rows(datalist)

#take the unique strings in data to pass to the next for loop
data_unique <- data %>%
  unique()

player_vector <- as.vector(data_unique)
player_vector <- unlist(player_vector)
player_vector <- player_vector
player_count <- length(player_vector)
datalist_players = list()
datalist_players = vector("list", length = player_count)

for (player in player_vector) {

  player_paste <- paste("https://www.basketball-reference.com",player, sep = "")

  player_html <- read_html(curl::curl(player_paste))

  player_tibble <- player_html %>%
    html_elements('a') %>%
    html_attr('href') %>%
```

```
    data.frame() %>%
    as.tibble()

  colnames(player_tibble) <- c("link")

  player_tibble <- player_tibble %>%
    filter(grepl('cbb', link), grepl('.html', link), grepl('utm', link))

  player_list <- as.character(player_tibble)

  #some players never went to college (player internationally, etc) so they will not have a link
  if(player_list == "character(0)") {
    next
  }

  college_html <- read_html(curl::curl(player_list))

  college_name <- college_html %>%
    html_elements('h1')

  college_name <- sub(".*<span>","" ,sub("</span>.*","" ,college_name))

  college_table <- college_html %>%
    html_table()

  #depending on the page being scraped, add'l players who did not play between 2010-2024 will be added to t
he list of college players and will have played in a year where there were no three pointers so we much che
ck that there were three pointers to skip those that did not have them
  if(grepl("3PA", paste(as.character(colnames(college_table[[2]])), collapse = "")) == TRUE) {

    #starting in 2004-05, intraconference play stats were recorded within college stats which added an ad
d'l table between the per game stats and totals
    if(colnames(college_table[[2]][2]) != "Team"){
      college_table <- college_table[[2]] %>%
        filter(Season != "Career", Season != "") %>%
        select(Season, MP, `3P`, `3PA`, FT, FTA, PTS, FGA) %>%
        mutate(Player = college_name)
    } else {
      college_table <- college_table[[3]] %>%
        filter(Season != "Career", Season != "") %>%
        select(Season, MP, `3P`, `3PA`, FT, FTA, PTS, FGA) %>%
        mutate(Player = college_name)
    }

  } else {
    next
  }

  datalist_players[[player]] <- college_table

}

data_players <- bind_rows(datalist_players)


#since each web scraping can take hours, for ease of use the data is stored in a .xlsx file and called late
```

```
r
wb <- createWorkbook()
addWorksheet(wb, "Data")
writeData(wb, "Data", data_players)

saveWorkbook(wb, file = file_path, overwrite = TRUE)
```

NBA Stats Web Scrape

Data Manipulation

Since basketball-reference formatting is very uniform, there is only a small amount of manipulation to do on each table before joining them together.

We intentionally did not take the 3P% or FT% per year for either NBA or college because we wanted to calculate these ourselves after grouping each player together. We will also filter both NBA and college tibbles to take only the top 75th percentile in minutes played (MP), three pointers attempted (3PA), and free throws attempted (FTA) to make sure each player has a large enough sample for analysis. Our last filters will involve making sure the NBA tibble only includes stats for players in their first three years and after Stephen Curry's first MVP in 2015 when the NBA three point revolution officially began.

Hide

```
library("tidyverse")
library("readxl")
library("openxlsx")
library("lubridate")
library("rvest")
library("curl")

nba <- read_excel("/Users/ijspeelman/OneDrive - Integrity Express Logistics/Desktop/Storage/Analytics/Projects/Personal/College v NBA Threes/nba stats.xlsx")
college <- read_excel("/Users/ijspeelman/OneDrive - Integrity Express Logistics/Desktop/Storage/Analytics/Projects/Personal/College v NBA Threes/college.xlsx")

nba_mutate <- nba %>%
  group_by(Player) %>%
  mutate(year_rank = rank(year)) %>%
  ungroup() %>%
  filter(
    MP >= quantile(MP, .25, na.rm = TRUE),
    FTA >= quantile(FTA, .25, na.rm = TRUE),
    `3PA` >= quantile(`3PA`, .25, na.rm = TRUE),
    year_rank != year - 2009,
    year_rank <= 3,
    year > 2015
  ) %>%
  select(Player, MP, `3PA`, `3P`, `FTA`, `FT`) %>%
  group_by(Player) %>%
  summarize(MP = sum(MP), `3PA` = sum(`3PA`), `3P` = sum(`3P`), FT = sum(FT), FTA = sum(FTA)) %>%
  ungroup()

college_mutate <- college %>%
  filter(
    MP >= quantile(MP, .25, na.rm = TRUE),
    FTA >= quantile(FTA, .25, na.rm = TRUE),
    `3PA` >= quantile(`3PA`, .25, na.rm = TRUE)
  ) %>%
  group_by(Player) %>%
  summarize(MP = sum(MP), `3PA` = sum(`3PA`), `3P` = sum(`3P`), FTA = sum(FTA), FT = sum(FT)) %>%
  ungroup()

head(nba_mutate)
```

Player <chr>	**MP** <dbl>	**3PA** <dbl>	**3P** <dbl>	**FT** <dbl>	**FTA** <dbl>
AJ Griffin	1401	259	101	42	47
Aaron Gordon	4161	409	119	285	410
Aaron Holiday	3439	530	197	183	220
Aaron Nesmith	3059	537	186	162	196
Aaron Wiggins	3734	386	152	177	227
Abdel Nader	2083	293	103	101	141

6 rows

Hide

```
head(college_mutate)
```

Player <chr>	MP <dbl>	3PA <dbl>	3P <dbl>	FTA <dbl>	FT <dbl>
A.J. Lawson	1771	302	105	243	169
A.J. Price	2914	433	158	371	268
AJ Green	3226	677	255	355	321
Aaron Brooks	3218	472	177	334	280
Aaron Gordon	1187	45	16	180	76
Aaron Harrison	2307	361	121	277	218

6 rows

Hide

```
NA
```

Then we join nba_mutate and college_mutate together and create our explanatory variables 3P%_college and FT%_college and our response variable 3P%_nba. We make sure to remove any player in the NBA that was not in college for our analysis by removing those with NAs in the college columns.

Hide

```
nba_college <- nba_mutate %>%
  left_join(college_mutate, by = c("Player"), suffix = c("_nba","_college")) %>%
  filter(is.na(MP_college) == 0 | is.na(FTA_college) == 0 | is.na(`3PA_college`) == 0) %>%
  mutate(`3P%_nba` = `3P_nba`/`3PA_nba`, `FT%_nba` = FT_nba / FTA_nba, `3P%_college` = `3P_college`/`3PA_co
llege`, `FT%_college` = FT_college / FTA_college)
```

Exploratory Analysis

To see how college 3P% and college FT% will affect our model, first we check their correlation directly with NBA 3P% and then graph them against NBA 3P%.

The correlation coefficients come out to .264 for college 3P% and .391 for college FT% indicating FT% as a better indicator of NBA 3P%.

Point for the talking heads.

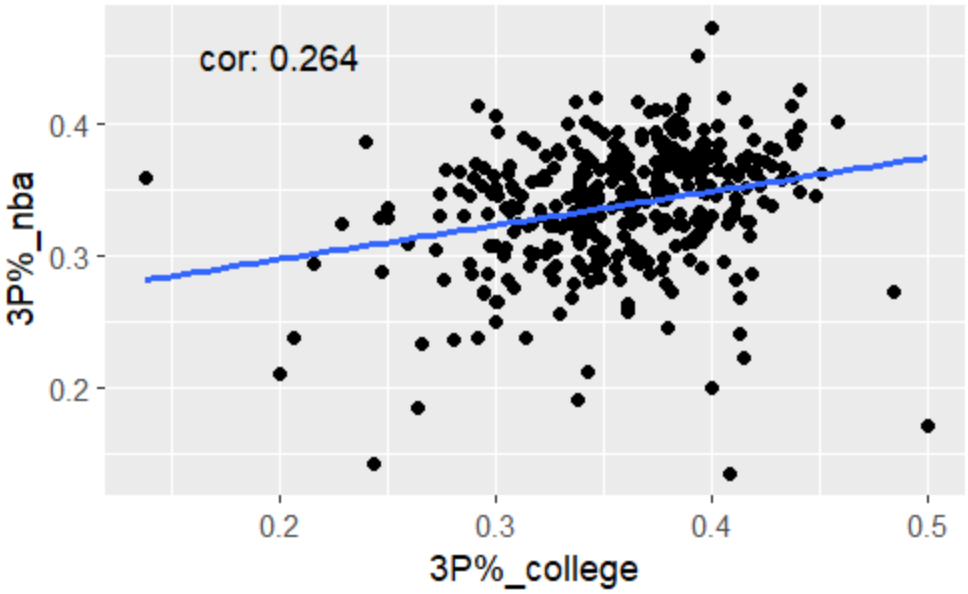
However, these are not high correlation coefficients and only indicate moderate correlation with NBA 3P%.

Our last graph shows how both college 3P% and FT% model against NBA 3P% by letting the colors of the points be NBA 3P%.

Hide

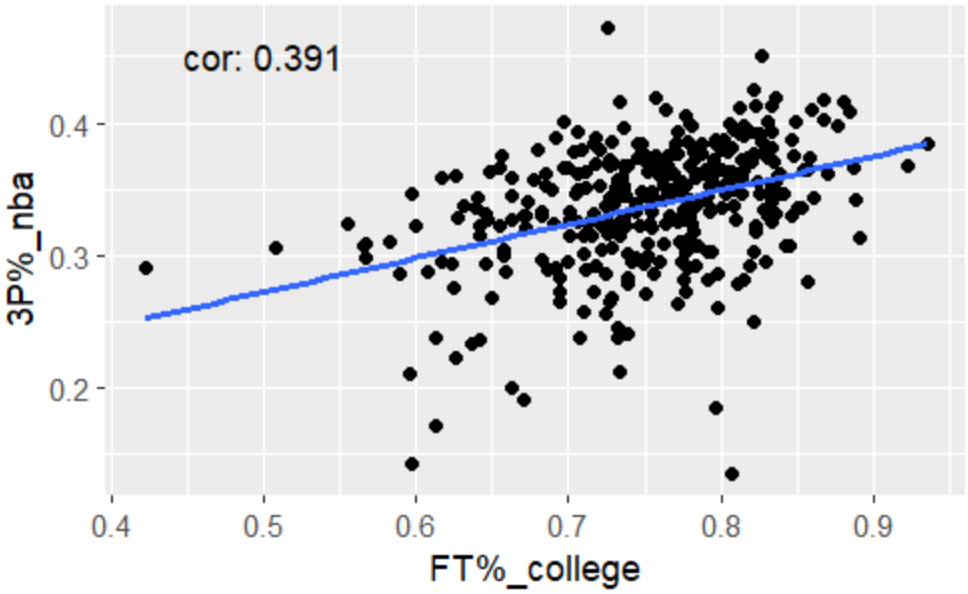
```
cor_3P <- round(cor(nba_college$`3P%_nba`, nba_college$`3P%_college`), digits = 3)
cor_FT <- round(cor(nba_college$`3P%_nba`, nba_college$`FT%_college`), digits = 3)

nba_college %>%
  ggplot(aes(`3P%_college`, `3P%_nba`)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  annotate("text", x = .2, y = .45, label = paste("cor: ",cor_3P, sep = ""))
```



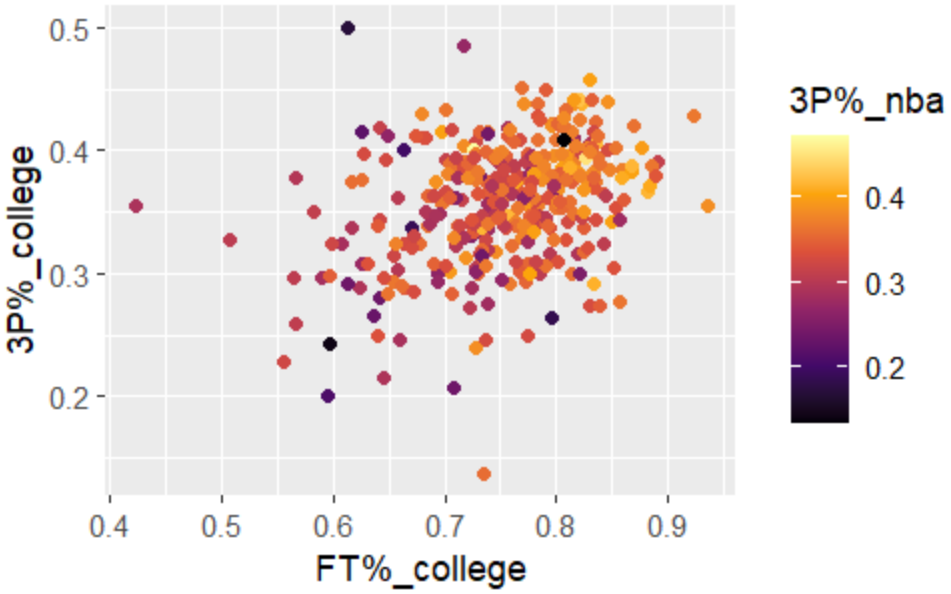
Hide

```
nba_college %>%
  ggplot(aes(`FT%_college`, `3P%_nba`)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  annotate("text", x = .5, y = .45, label = paste("cor: ",cor_FT, sep = ""))
```



Hide

```
nba_college %>%
  ggplot(aes(`FT%_college`, `3P%_college`, color = `3P%_nba`)) +
  geom_jitter() +
  scale_color_viridis_c(option = "inferno")
```



Data Modelling

Create our multiple linear regression has us take our response variable, 3P%_nba, and have it predicted by our explanatory variables, 3P%_college and FT%_college with an interaction between them. This results in a R-squared value of 0.1791 and a RSE of 0.043. This indicates an error of around 4% in either direction on a predicted 3P%. This could result in a player's prediction being either one of the greatest shooters in the league or a well-below average shooter.

Hide

```
nba_model <- lm(`3P%_nba` ~ `3P%_college` * `FT%_college`, data = nba_college)

summary(nba_model)
```

```
Call:
lm(formula = `3P%_nba` ~ `3P%_college` * `FT%_college`, data = nba_college)

Residuals:
      Min        1Q    Median        3Q       Max
-0.224334 -0.019068  0.005395  0.028894  0.136525

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.34967    0.15832   2.209   0.0279 *
`3P%_college`              -0.52132    0.44945  -1.160   0.2469
`FT%_college`              -0.09713    0.21681  -0.448   0.6544
`3P%_college`:`FT%_college`  0.91361    0.60949   1.499   0.1348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04305 on 346 degrees of freedom
Multiple R-squared:  0.1791,    Adjusted R-squared:  0.172
F-statistic: 25.17 on 3 and 346 DF,  p-value: 9.437e-15
```

To show the predictions visually, we will re-graph the college 3P% vs college FT% with the color scale as NBA 3P%, but with an overlay of predicted data as a scatter to show how the predictions compare to actual data.

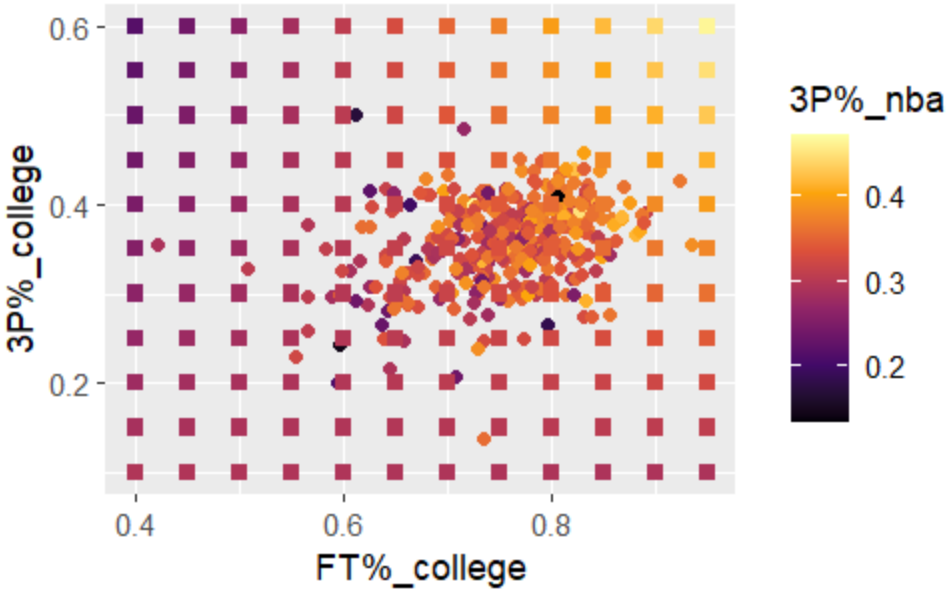
To create predicted data, we create two sequences for 3P%_college and FT%_college to create multiple theoretical data points for the model to predict their 3P%_nba.

Hide

```
explanatory_data <- expand_grid(
  `3P%_college` = seq(.1,.6,.05),
  `FT%_college` = seq(.4,.95,.05)
)

prediction_data <- explanatory_data %>%
  mutate(`3P%_nba` = predict(nba_model, explanatory_data))

ggplot(nba_college, aes(`FT%_college`, `3P%_college`, color = `3P%_nba`)) +
  geom_jitter() +
  scale_color_viridis_c(option = "inferno") +
  geom_point(data = prediction_data, shape = 15, size = 2)
```



Conclusion

While the model shows that using college 3P% and FT% to predict future NBA 3P% success is not nearly perfect, the analytical basketball talking heads were correct to see FT% as an additional indicator of NBA 3P%.

Using a model that only uses college 3P% or college FT% results in a lower R-squared and higher RSE. This shows that their interaction is a useful indicator despite not being a full proof one.

Hide

```
three_model <- lm(`3P%_nba` ~ `3P%_college`, data = nba_college)

summary(three_model)
```

```

Call:
lm(formula = `3P%_nba` ~ `3P%_college`, data = nba_college)

Residuals:
      Min        1Q    Median        3Q       Max
-0.215146 -0.023709  0.007006  0.028978  0.124033

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.24566    0.01811  13.568  < 2e-16 ***
`3P%_college`  0.25631    0.05021   5.105 5.46e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0457 on 348 degrees of freedom
Multiple R-squared:  0.06967,   Adjusted R-squared:  0.067
F-statistic: 26.06 on 1 and 348 DF,  p-value: 5.457e-07
```

Hide

```
FT_model <- lm(`3P%_nba` ~ `FT%_college`, data = nba_college)

summary(FT_model)
```

```
Call:
lm(formula = `3P%_nba` ~ `FT%_college`, data = nba_college)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.215847 -0.024293  0.005925  0.030715  0.142282 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.14392    0.02450   5.875 9.90e-09 ***
`FT%_college`  0.25654    0.03236   7.928 3.06e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0436 on 348 degrees of freedom
Multiple R-squared:  0.153, Adjusted R-squared:  0.1505 
F-statistic: 62.85 on 1 and 348 DF,  p-value: 3.06e-14
```