NBA Season Rankings and Predictions

2023-06-10

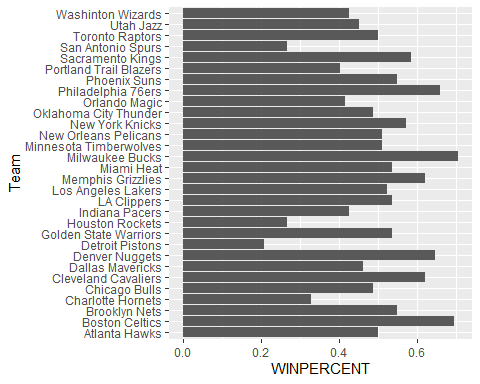
The purpose of this project is to provide a comprehensive list of the team standings based on key statistics from the NBA website. Based on these stats, we’d like to predict what the standings will be next season using machine learning.

data <- read\_excel("NBA Standings Data.xlsx")

sorted\_df <- arrange(data, desc(W))  
sorted\_df

## # A tibble: 30 × 27  
## Team GP W L WINPERCENT MIN PTS FGM FGA FGPERCENT THREEPM  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Milwa… 82 58 24 0.707 48.4 117. 42.7 90.4 47.3 14.8  
## 2 Bosto… 82 57 25 0.695 48.7 118. 42.2 88.8 47.5 16   
## 3 Phila… 82 54 28 0.659 48.5 115. 40.8 83.8 48.7 12.6  
## 4 Denve… 82 53 29 0.646 48.2 116. 43.6 86.4 50.4 11.8  
## 5 Cleve… 82 51 31 0.622 48.5 112. 41.6 85.2 48.8 11.6  
## 6 Memph… 82 51 31 0.622 48.2 117. 43.7 92.1 47.5 12   
## 7 Sacra… 82 48 34 0.585 48.4 121. 43.6 88.2 49.4 13.8  
## 8 New Y… 82 47 35 0.573 48.7 116 42 89.4 47 12.6  
## 9 Brook… 82 45 37 0.549 48.1 113. 41.5 85.1 48.7 12.8  
## 10 Phoen… 82 45 37 0.549 48.2 114. 42.1 90.1 46.7 12.2  
## # ℹ 20 more rows  
## # ℹ 16 more variables: THREEPA <dbl>, THREEPPERCENT <dbl>, FTM <dbl>,  
## # FTA <dbl>, FTPERCENT <dbl>, OREB <dbl>, DREB <dbl>, REB <dbl>, AST <dbl>,  
## # TOV <dbl>, STL <dbl>, BLK <dbl>, BLKA <dbl>, PF <dbl>, PFD <dbl>,  
## # `PLUS/MINUS` <dbl>

ggplot(data = sorted\_df) +  
 geom\_col(aes(x = WINPERCENT, y = Team))

 The Boston Celtics had the highest percentage of wins during the 2022-23 regular season, with the Pistons in the lowest spot.

data <- sorted\_df %>%   
 select(WINPERCENT, PTS, FGPERCENT, THREEPPERCENT, FTPERCENT, OREB, DREB, REB, AST, TOV, STL, BLK, PF, PFD)  
  
data

## # A tibble: 30 × 14  
## WINPERCENT PTS FGPERCENT THREEPPERCENT FTPERCENT OREB DREB REB AST  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.707 117. 47.3 36.8 74.3 11.1 37.5 48.6 25.8  
## 2 0.695 118. 47.5 37.7 81.2 9.7 35.6 45.3 26.7  
## 3 0.659 115. 48.7 38.7 83.5 8.7 32.2 40.9 25.2  
## 4 0.646 116. 50.4 37.9 75.1 10.1 32.9 43 28.9  
## 5 0.622 112. 48.8 36.7 78 9.7 31.4 41.1 24.9  
## 6 0.622 117. 47.5 35.1 73.3 12 34.6 46.6 26   
## 7 0.585 121. 49.4 36.9 79 9.5 32.9 42.5 27.3  
## 8 0.573 116 47 35.4 76.1 12.6 34 46.6 22.9  
## 9 0.549 113. 48.7 37.8 80 8.2 32.3 40.5 25.5  
## 10 0.549 114. 46.7 37.4 79.3 11.8 32.4 44.2 27.3  
## # ℹ 20 more rows  
## # ℹ 5 more variables: TOV <dbl>, STL <dbl>, BLK <dbl>, PF <dbl>, PFD <dbl>

model <- lm(WINPERCENT ~ PTS + FGPERCENT + THREEPPERCENT + FTPERCENT + OREB + DREB + REB + AST + TOV + STL + BLK + PF + PFD, data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = WINPERCENT ~ PTS + FGPERCENT + THREEPPERCENT + FTPERCENT +   
## OREB + DREB + REB + AST + TOV + STL + BLK + PF + PFD, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.085087 -0.030251 -0.004282 0.035745 0.111159   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.837082 1.206414 -2.352 0.031831 \*   
## PTS -0.001665 0.009071 -0.184 0.856668   
## FGPERCENT 0.002805 0.018942 0.148 0.884108   
## THREEPPERCENT 0.063862 0.014892 4.288 0.000564 \*\*\*  
## FTPERCENT 0.002065 0.005804 0.356 0.726635   
## OREB -0.286082 0.509699 -0.561 0.582386   
## DREB -0.280666 0.507759 -0.553 0.588072   
## REB 0.307813 0.510150 0.603 0.554715   
## AST 0.001663 0.012569 0.132 0.896371   
## TOV -0.031931 0.017098 -1.867 0.080262 .   
## STL 0.051478 0.021747 2.367 0.030872 \*   
## BLK 0.015707 0.019494 0.806 0.432209   
## PF -0.039829 0.019225 -2.072 0.054814 .   
## PFD 0.028372 0.020532 1.382 0.186020   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0644 on 16 degrees of freedom  
## Multiple R-squared: 0.8471, Adjusted R-squared: 0.7228   
## F-statistic: 6.816 on 13 and 16 DF, p-value: 0.0002677

According to the model, the best predictors of win % are 3P%, TOV, STL, and PF. This doesn’t entirely make sense according to the hypothesis that PTS, Field Goals, and Assists/Rebounds would be the most significant indicators. But given that these are all variables that vary by team, it’s logical that they wouldn’t be the best predictors given a linear model.

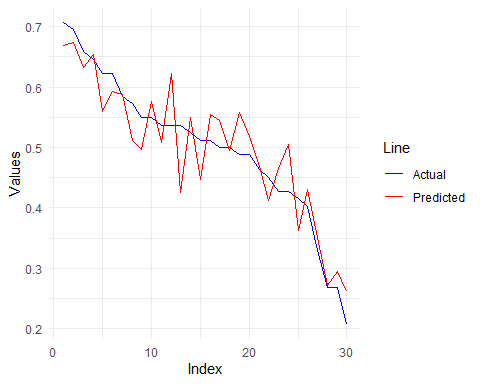
predictions <- predict(model)

actual <- data$WINPERCENT  
predicted <- predictions  
differences <- actual - predicted  
team <- sorted\_df$Team

results <- tibble(team = team, acutal = actual, predicted = predicted, difference = differences)  
results

## # A tibble: 30 × 4  
## team acutal predicted difference  
## <chr> <dbl> <dbl> <dbl>  
## 1 Milwaukee Bucks 0.707 0.669 0.0379   
## 2 Boston Celtics 0.695 0.673 0.0218   
## 3 Philadelphia 76ers 0.659 0.633 0.0259   
## 4 Denver Nuggets 0.646 0.654 -0.00788  
## 5 Cleveland Cavaliers 0.622 0.559 0.0630   
## 6 Memphis Grizzlies 0.622 0.593 0.0292   
## 7 Sacramento Kings 0.585 0.587 -0.00202  
## 8 New York Knicks 0.573 0.512 0.0610   
## 9 Brooklyn Nets 0.549 0.497 0.0517   
## 10 Phoenix Suns 0.549 0.576 -0.0273   
## # ℹ 20 more rows

ggplot(results, aes(x = 1:length(actual))) +  
 geom\_line(aes(y = actual, color = "Actual")) +  
 geom\_line(aes(y = predicted, color = "Predicted")) +  
 labs(x = "Index", y = "Values", color = "Line") +  
 scale\_color\_manual(values = c("Actual" = "blue", "Predicted" = "red")) +  
 theme\_minimal()



predictions <- arrange(results, desc(predicted)) %>%   
 mutate(teamabr = c("BOS", "MIL","DEN","PHI","LAC","MEM","SAC","PHX","CLE","CHI","NOP","LAL","ATL","OKC","NYK","GSW","WAS",  
 "BKN","TOR","DAL","IND","MIN","POR","MIA","UTA","ORL","CHA","SAS","HOU","DET"))  
predictions

## # A tibble: 30 × 5  
## team acutal predicted difference teamabr  
## <chr> <dbl> <dbl> <dbl> <chr>   
## 1 Boston Celtics 0.695 0.673 0.0218 BOS   
## 2 Milwaukee Bucks 0.707 0.669 0.0379 MIL   
## 3 Denver Nuggets 0.646 0.654 -0.00788 DEN   
## 4 Philadelphia 76ers 0.659 0.633 0.0259 PHI   
## 5 LA Clippers 0.537 0.622 -0.0851 LAC   
## 6 Memphis Grizzlies 0.622 0.593 0.0292 MEM   
## 7 Sacramento Kings 0.585 0.587 -0.00202 SAC   
## 8 Phoenix Suns 0.549 0.576 -0.0273 PHX   
## 9 Cleveland Cavaliers 0.622 0.559 0.0630 CLE   
## 10 Chicago Bulls 0.488 0.558 -0.0696 CHI   
## # ℹ 20 more rows

data <- arrange(predictions, desc(actual))  
  
# Plot the bar chart using ggplot2  
plot <- ggplot(data, aes(y = teamabr)) +  
 geom\_bar(aes(x = actual, fill = "Actual"), stat = "identity", position = "dodge") +  
 geom\_bar(aes(x = predicted, fill = "Predicted"), stat = "identity", position = "dodge") +  
 geom\_bar(aes(x = difference, fill = "Difference"), stat = "identity", position = "dodge") +  
 labs(title = "Multiple Bars in Each Column", x = "Team", y = "Values") +  
 scale\_fill\_manual(values = c("Actual" = "red", "Predicted" = "blue", "Difference" = "orange")) +  
 theme\_bw()  
  
plot

