NBA’s Winning Factor: How to Make Playoff

Code

Sung In Cho, Yoon Seo Jang

Dec 16, 2016

- File : `2014~2016.11.18 NBA Season Game Log` , `2006~2016 NBA Season Team` csv Data
- Used package: `dplyr` , `tidyr` , `xts` , `lubridate` , `qtlcharts` , `forecast` , `tseries` , `leaflet` , `ggplot2` , `plotly` , `dygraphs` , `viridis` , `graphics`

---

1 Introduction

This is an exploratory analysis for data collected from **NBA**.

These days, many sport clubs are using statistical analysis to run the club more efficiently. Due to a development of technology, many types of data now can be collected from many sport games. Nowadays, sport of data does not only refer to baseball. Especially, basketball, for instance, also provides variety of large amount of data. We thought, it would be interesting to analyze sports data which is not about baseball. That’s why we made an analysis on NBA data.

First, we wanted to figure out which factor is most influential on outcomes of games and, by extension, making playoff. Making the playoff is one of the most important goal of the season. It gives invaluable experience to the team. We thought key factors of winning one game and going to the playoffs may be different. There, we made two individual analyzations to see if there is really a difference.

The source of the data is “http://stats.nba.com/ (http://stats.nba.com/)”. This page provides information of games and teams of the NBA. The match data are collected from 2014 to Nov 18th, 2016 and the season average data of teams are collected from 2006 to 2016.

1.1 Loading packages

Code

Now that our packages are loaded, let’s read in and check the attributes of data.

---

2 General Winning Factors of Basketball (NBA)

Every team has its own winning strategy. For instance, "Golden State Warriors" is the famous team for high percentage of three-pointers and "San Antonio Spurs" prefers to pass the ball to one another until they get a perfect chance of scoring. However, we assumed that there would be common winning factors and we wanted to figure it out. Using the data, we made a model that can predict the winning rate. Additionally, we tested the model with test data from the latest season.

2.1 Attributes of data and Handling

We spent very long time on data handling. But I thought the presentation time may not be enough to explain all of it. Therefore, we are going to skip our explanation on the data handling. If you want to know about what we have done, please check for the given handout.

Code

wl	THPM	THPA	THPP	FTM	FTA	FTP	OREB	DREB	REB	AST	STL	BLK	TOV	PF	TPA	TPM	TPP
W	11	29	37.9	19	24	79.2	11	38	49	39	8	6	17	23	80	46	57.50000
L	15	43	34.9	22	25	88.0	27	34	61	29	10	1	15	26	82	38	46.34146
W	18	35	51.4	20	28	71.4	11	28	39	37	4	2	9	16	47	33	70.21277

Code

```
## 'data.frame':    4920 obs. of  18 variables:
##  $ wl  : Factor w/ 2 levels "L","W": 2 1 2 2 2 2 2 2 2 1 ...
##  $ THPM: int  11 15 18 15 12 11 14 4 8 13 ...
##  $ THPA: int  29 43 35 20 28 35 28 15 16 27 ...
##  $ THPP: num  37.9 34.9 51.4 75 42.9 31.4 50 26.7 50 48.1 ...
##  $ FTM : int  19 22 20 12 22 44 30 19 29 29 ...
##  $ FTA : int  24 25 28 16 31 49 34 24 33 35 ...
##  $ FTP : num  79.2 88 71.4 75 71 89.8 88.2 79.2 87.9 82.9 ...
##  $ OREB: int  11 27 11 3 18 19 11 4 10 7 ...
##  $ DREB: int  38 34 28 31 28 31 40 30 29 32 ...
##  $ REB : int  49 61 39 34 46 50 51 34 39 39 ...
##  $ AST : int  39 29 37 31 22 22 32 35 23 17 ...
##  $ STL : int  8 10 4 13 10 6 4 8 9 7 ...
##  $ BLK : int  6 1 2 3 6 4 5 1 7 3 ...
##  $ TOV : int  17 15 9 13 13 14 16 6 17 14 ...
##  $ PF  : int  23 26 16 30 25 27 26 23 27 35 ...
##  $ TPA : int  80 82 47 64 67 64 58 76 73 63 ...
##  $ TPM : int  46 38 33 41 40 30 32 52 41 33 ...
##  $ TPP : num  57.5 46.3 70.2 64.1 59.7 ...
```

These are the names, class type of variables. We can also check first few observations. In total, there are 4920 observations and 18 variables. Simple description of the variables is as follows. :

Variable Name	Description	Variable Name	Description
wl	Game Result	OREB	Offensive Rebound
TPM	2 Points Made	DREB	Defensive Rebound
TPA	2 Points Attempted	REB	Total Rebound
TPP	2 Points Percentage	AST	Assist
THPM	Three Points Made	STL	Steal
THPA	Three Points Attempted	BLK	Block
THPP	Three Points Percentage	TOV	Turn Over
FTM	Free Throw Made	PF	Personal Foul
FTA	Free Throw Attempted		

Variable Name	Description	Variable Name	Description
FTP	Free Throw Percentage		

2.2 Variable Selection

We used random forest method and correlation matrix to select variables.

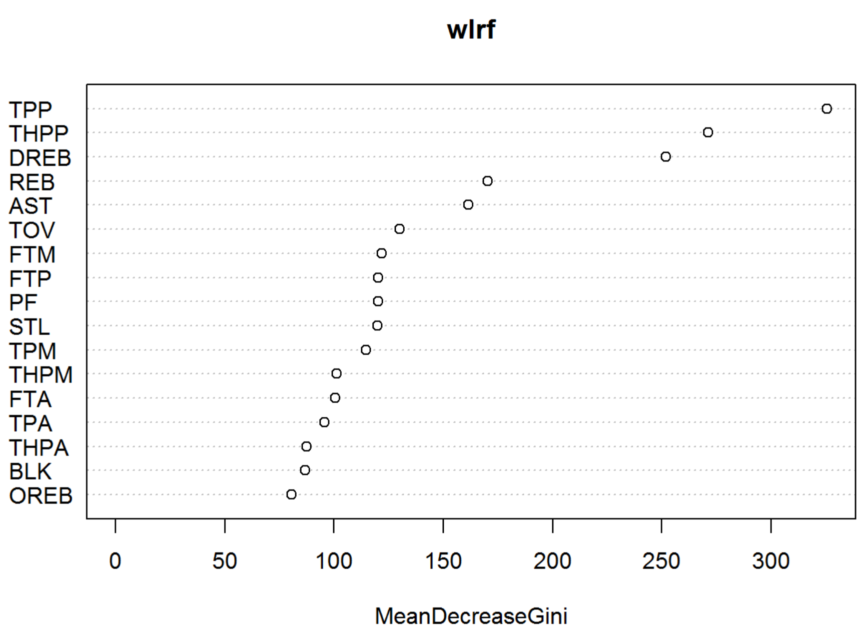
2.2.1 RandomForest

2.2.1.1 Importance of Variables

Code

```
##
## Call:
##  randomForest(formula = wl ~ ., data = wl, ntree = 200, proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 20.3%
## Confusion matrix:
##      L    W class.error
## L 1977  483   0.1963415
## W  516 1944   0.2097561
```

Code



- This is the result of the random forest. The error rate is 20.3%. Since this is the prediction of winning rate, 80% of accuracy seems about right.
- `TPP` and `THPP` showed the largest importance. It may be reasonable to say that scoring is the first prerequisite for winning.
- `DREB` and `REB` were also very important variables. Rebound refers to consistency of the team and consistency is about making less mistakes.

2.2.1.2 Accuracy of Random Forest, using Test Data

Code

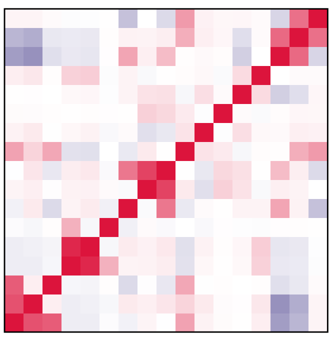
```
## $`confusion matrix`
##
## wlPred   L   W
##      L 166  36
##      W  45 175
##
## $accuracy
## [1] 0.8080569
##
## $error
## [1] 0.1919431
```

This is the result of the random forest applied to the test data. The error rate is 19.1%, which is about the same with the above, 20.3%.

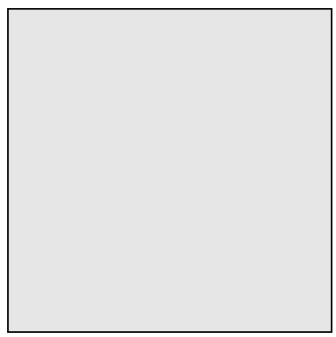
2.2.2 Correlation of Variables

Code

NBA Season



Scatterplot



`(TPA, TPM)`, `(TPP, TPM)`, `(THPA, THPM)`, `(THPP, THPM)`, `(FTA, FTM)`, and `(REB, DREB)` showed the absolute value of correlation coefficient higher than 0.6. When you see the above result of variable importance from random forest, `TPP` and `THPP` showed the largest importance. Therefore, we removed `TPM` and `THPM`. `DREB` showed larger importance than `REB` and `FTM` showed larger importance than `FTA`. Hence, we removed `REB` and `FTA`.

2.3 Logistic Regression modeling

2.3.1 Modeling Procedure

- We first ran glm against dependent variable, `wl`; game result. If there had been invalid variables, we removed them and ran glm again.
- If there had only been valid variables in glm, we ran anova to see whether we can reduce the model. Since size of the logistic regression model does not increase the explanatory power, smaller size of model that has same explanatory power is always better. After removing certain variables we ran glm again.

Code

```
##
## Call:
## glm(formula = wl ~ TPP + THPP + DREB + TOV + FTP + FTM + STL +
##     PF + BLK + OREB, family = binomial, data = wl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4759  -0.5558  -0.0018   0.5385   3.3925
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -28.511079   0.933406 -30.545  < 2e-16 ***
## TPP           0.237567   0.008768  27.094  < 2e-16 ***
## THPP          0.132376   0.005219  25.365  < 2e-16 ***
## DREB          0.286553   0.010682  26.826  < 2e-16 ***
## TOV          -0.203480   0.011548 -17.620  < 2e-16 ***
## FTP           0.027811   0.004276   6.504 7.80e-11 ***
## FTM           0.088723   0.007600  11.674  < 2e-16 ***
## STL           0.273486   0.015607  17.524  < 2e-16 ***
## PF           -0.115415   0.010251 -11.259  < 2e-16 ***
## BLK           0.111024   0.016527   6.718 1.84e-11 ***
## OREB          0.166996   0.012138  13.758  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6820.6  on 4919  degrees of freedom
## Residual deviance: 3705.2  on 4909  degrees of freedom
## AIC: 3727.2
##
## Number of Fisher Scoring iterations: 6
```

The final model includes `TPP`, `THPP`, `DREB`, `TOV`, `FTP`, `FTM`, `STL`, `PF`, `BLK` and `OREB` for explanatory variables.

$$p = \frac{e^{\beta}}{1+e^{\beta}}$$

$$\beta = -28.51 + 0.24 \cdot TPP + 0.13 \cdot THPP + 0.29 \cdot DREB - 0.20 \cdot TOV + 0.03 \cdot FTP$$

$$+0.09 \cdot FTM + 0.27 \cdot STL - 0.12 \cdot PF + 0.11 \cdot BLK + 0.17 \cdot OREB$$

- These are the results of logistic regression. `DREB` showed the largest coefficient followed by `STL` and `TPP`.

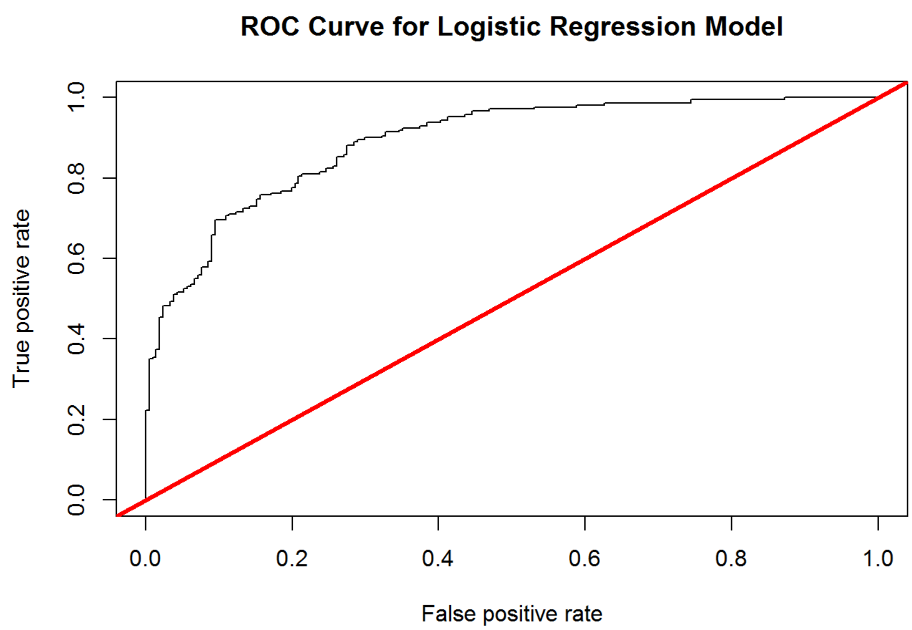
2.3.2 Confusion Matrix, Accuracy, Error, using Test Data

Code

```
## $`confustion matrix`
##       Predicted
## Actual   L   W
##      L 159  52
##      W  39 172
##
## $accuracy
## [1] 0.7843602
##
## $error
## [1] 0.2156398
```

The model’s accuracy was 78.4%. We assumed that the model is worth of using.

Code



Code

```
## [1] 0.890591
```

We drew Receiver Operating Characteristic(ROC) Curve to evaluate the model. Area Under the Curve(AUC) was 0.89. It seems like logistic regression model classifies the data very well.

2.4 Performance of Logistic Regression

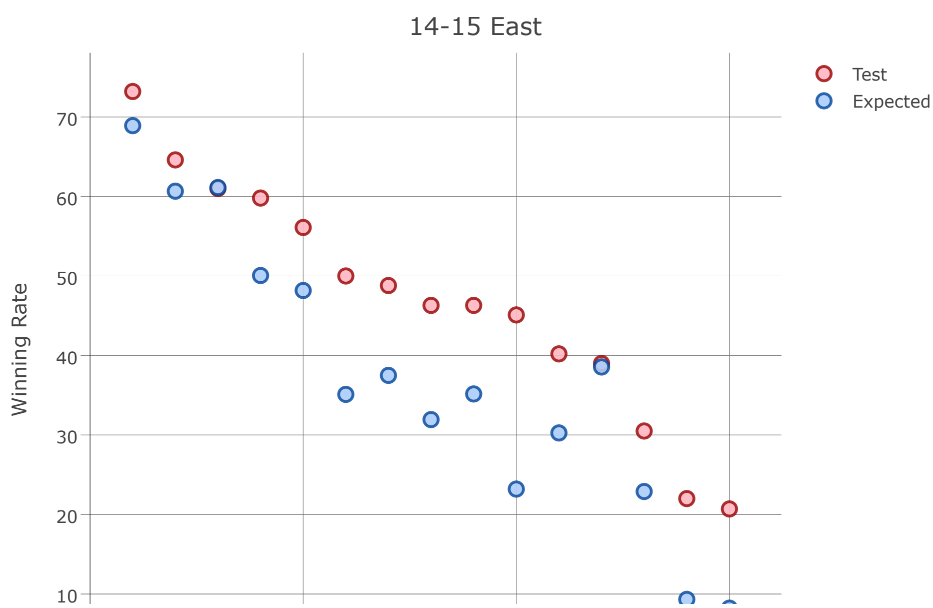
We put each teams’ values of variables into the model to find out how well the model can predict the winning rate.

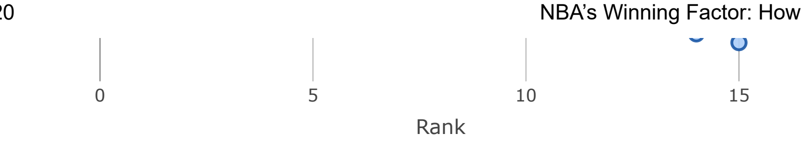
2.4.1 Comparing Expected Winning rate with Real Winning rate of the teams

Code

2.4.1.1 East Conference (14-15)

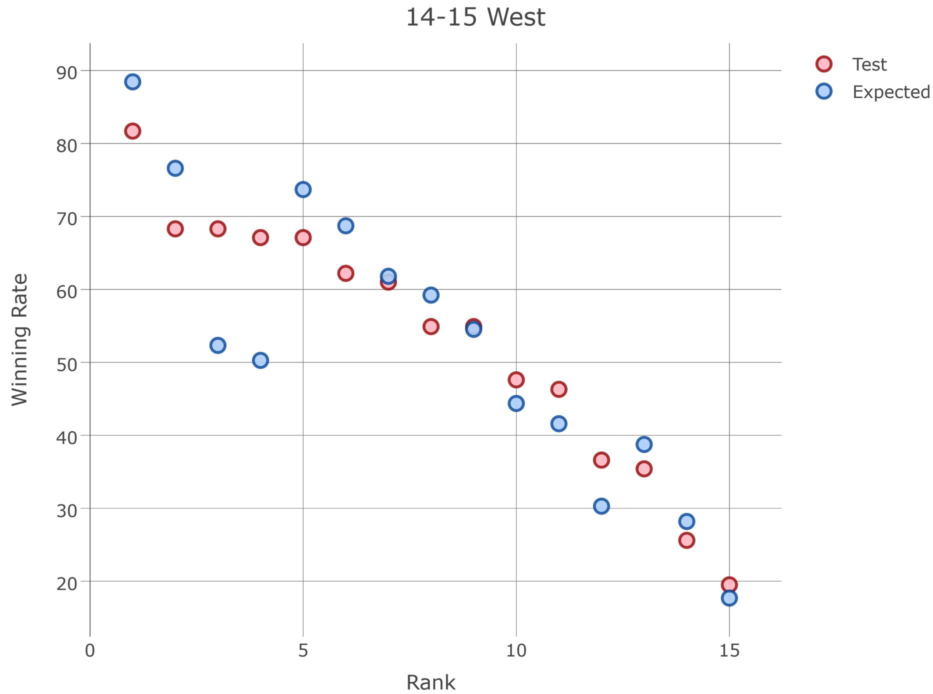
Code





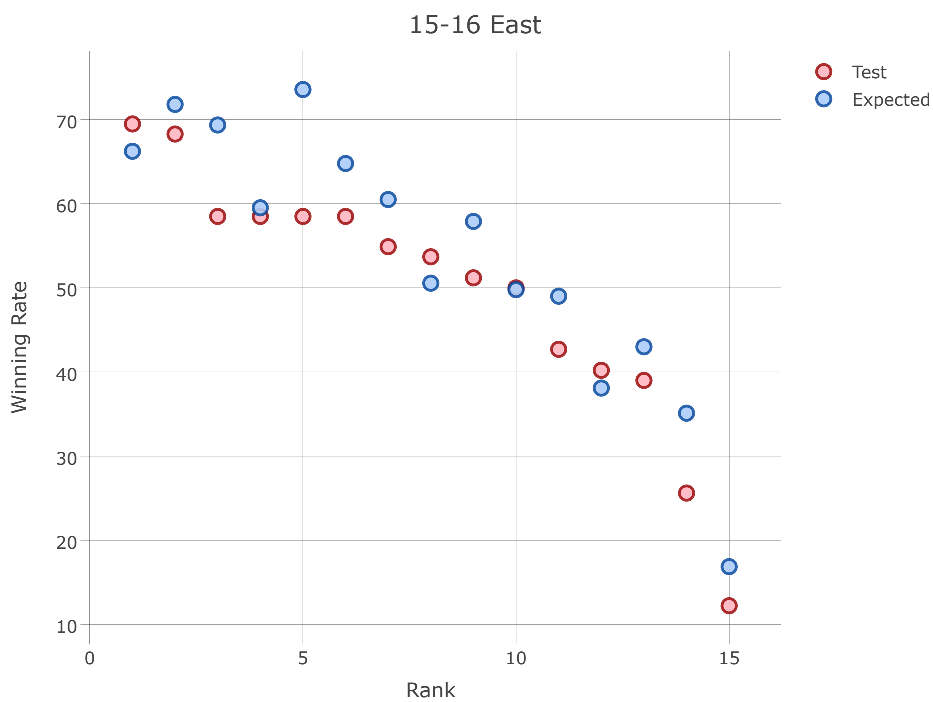
2.4.1.2 West Conference (14-15)

Code



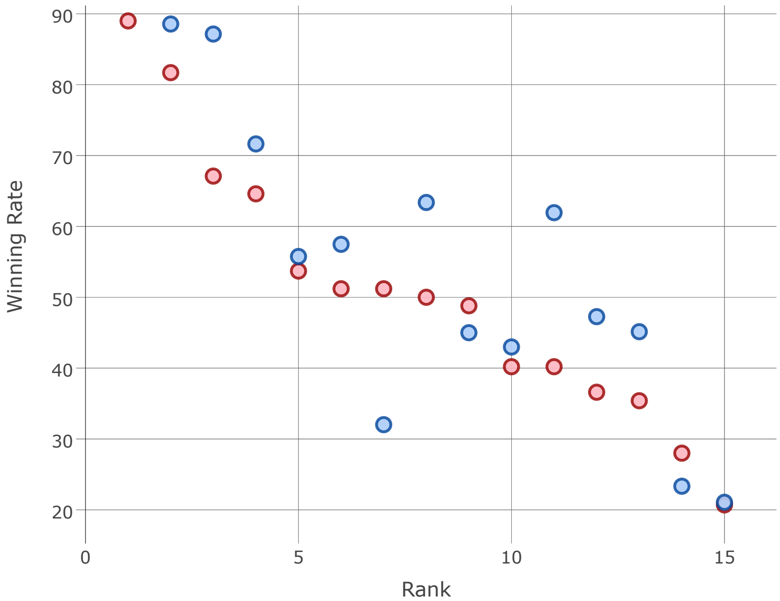
2.4.1.3 East Conference(15-16)

Code



2.4.1.4 West Conference(15-16)

Code



Generally,the model predicted winning rate of each teams very well. However, some of the predictions were not accurate.

2.4.2 Why aren't they expected well?

We drew radar plot with few selected variables to see why these teams' winning rates are not expected well.

2.4.2.1 Choosing Variables

Coefficients of the model refers to influence of explanatory variables. However, if the value of variables differ largely, it will be hard to say, variables having the largest coefficients have the largest influence on the winning rate. Therefore, we multiplied coefficients of the model with mean of each variables to measure the influence. We used variables that had five largest values on the radar plot.

$$(coefficient) * (mean.of.variables)$$

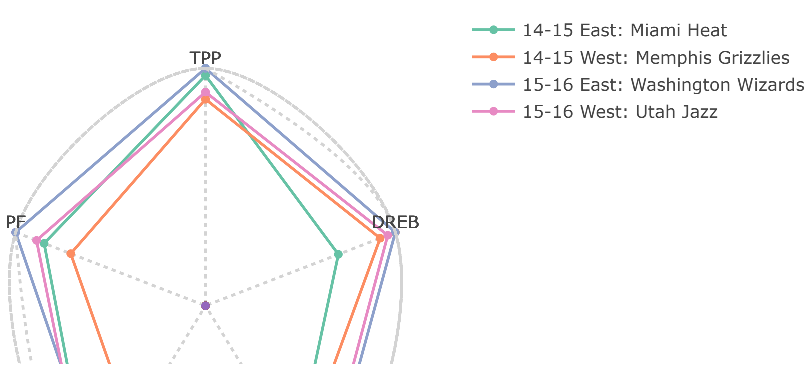
Code

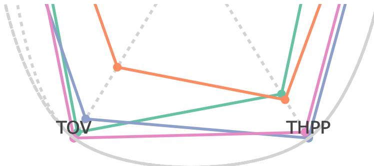
```
##       TPP      DREB      THPP       TOV        PF
## 11.645877  9.420357  4.631249 -2.923783 -2.336037
```

`TPP` , `DREB` , `THPP` , `TOV` , and `PF` are chosen in descending order of influence.

2.4.2.2 Radar Plot

Code





Code

team	TPP	DREB	THPP	TOV	PF
14-15 East: Miami Heat	0.97	0.70	0.77	0.97	0.85
14-15 West: Memphis Grizzlies	0.87	0.92	0.80	0.63	0.71
15-16 East: Washington Wizards	1.00	1.00	1.00	0.90	1.00
15-16 West: Utah Jazz	0.90	0.96	0.97	1.00	0.89

Although “Miami Heat” and “Memphis Grizzlies” had less personal fouls and turn overs than other two, the model expected them to have lower winning rates as their percentage of two-pointers, number of defensive rebounds, and percentage of three-pointers are smaller. We figured out that, if the team’s values of these variables are very different from other teams, then the performance of the model may be very poor. However, except for these kinds of cases, the model would predict team’s winning rate very well.

3 Making Playoffs

Every team wants to make it to the playoffs. Although only one team can earn championship title, making playoffs would give them invaluable experience. At the middle of the season, people often starts to assume that certain teams will be able to make playoffs. However, there is rarely a solid evidence in their reasoning. By using the similar procedure that we used in previous analysis, we are going to find out important factors of making playoffs. We made a model that can classify playoff teams.

3.1 Attributes of data and Handling

Code

THPM	THPA	THPP	FTM	FTA	FTP	OREB	DREB	REB	AST	TOV	STL	BLK	BLKA	PF	PFD	TPA	TPM	TPI
10.7	29.6	36.2	16.3	21.7	74.8	10.6	33.9	44.5	22.7	13.6	6.7	3.9	4.4	20.3	20.6	54.4	28.0	51.4705
8.6	23.3	37.0	20.8	26.7	77.7	10.2	33.2	43.4	18.7	13.1	7.8	5.5	5.4	19.6	22.0	58.0	28.1	48.4482
9.9	28.4	35.0	15.6	20.0	78.3	8.3	33.8	42.1	25.6	15.0	9.1	5.9	5.0	19.1	18.3	56.0	28.7	51.2500

Code

```
## 'data.frame':    300 obs. of  20 variables:
##  $ THPM: num  10.7 8.6 9.9 8.7 10.6 6.1 8.1 9 7.9 8.6 ...
##  $ THPA: num  29.6 23.3 28.4 26.1 29.4 18 23 26.2 21.4 24.2 ...
##  $ THPP: num  36.2 37 35 33.5 36.2 33.6 35.1 34.5 37.1 35.8 ...
##  $ FTM : num  16.3 20.8 15.6 18.5 18.7 17.1 17.4 17.1 16.5 16.5 ...
##  $ FTA : num  21.7 26.7 20 23.5 23.7 23 22.8 25.5 21 22.5 ...
##  $ FTP : num  74.8 77.7 78.3 78.8 79 74.5 76.4 66.8 78.7 73 ...
##  $ OREB: num  10.6 10.2 8.3 11.6 9 9.8 10.3 12.5 11.1 9.1 ...
##  $ DREB: num  33.9 33.2 33.8 33.3 35 34.3 33.9 33.9 35.2 32.8 ...
##  $ REB : num  44.5 43.4 42.1 44.9 43.9 44.1 44.2 46.3 46.3 41.8 ...
##  $ AST : num  22.7 18.7 25.6 24.2 21.7 20.8 21.2 19.4 22.8 24.5 ...
##  $ TOV : num  13.6 13.1 15 13.7 12.5 14.1 14.9 13.5 13.9 14.5 ...
##  $ STL : num  6.7 7.8 9.1 9.2 7.3 6.7 9 7 6 8.6 ...
##  $ BLK : num  3.9 5.5 5.9 4.2 5.3 6.5 4.8 3.7 5.7 3.9 ...
##  $ BLKA: num  4.4 5.4 5 5.5 5.5 4.1 4.5 4.5 5.7 4.3 ...
##  $ PF  : num  20.3 19.6 19.1 21.9 18.1 18.3 20 19 18.8 20.8 ...
##  $ PFD : num  20.6 22 18.3 21 20.4 19.6 20.4 21.6 18.7 20.1 ...
##  $ TPA : num  54.4 58 56 63.1 55 63.7 62.2 60.2 66 61.6 ...
##  $ TPM : num  28 28.1 28.7 30.5 26.4 32.3 30.2 28.9 30.7 30.9 ...
##  $ TPP : num  51.5 48.4 51.2 48.3 48 ...
##  $ PO  : Factor w/ 2 levels "F","P": 2 2 2 2 2 2 2 2 1 1 ...
```

These are the names, class type of variables after handling data. We can also check first few observations. In total, there are 300 observations and 20 variables. Simple description of the variables is as follows. :

Variable Name	Description	Variable Name	Description
PO	Playoff or Failure	OREB	Offensive Rebound
TPM	2 Points Made	DREB	Defensive Rebound
TPA	2 Points Attempted	REB	Total Rebound
TPP	2 Points Percentage	AST	Assist
THPM	Three Points Made	TOV	Turn Over
THPA	Three Points Attempted	STL	Steal
THPP	Three Points Percentage	BLK	Block
FTM	Free Throw Made	BLKA	Blocked Shots
FTA	Free Throw Attempted	PF	Personal Foul
FTP	Free Throw Percentage	PFD	Personal Foul Drawn

Code

We produced training data set and test data set by generating random numbers.

3.2 Variable Selection

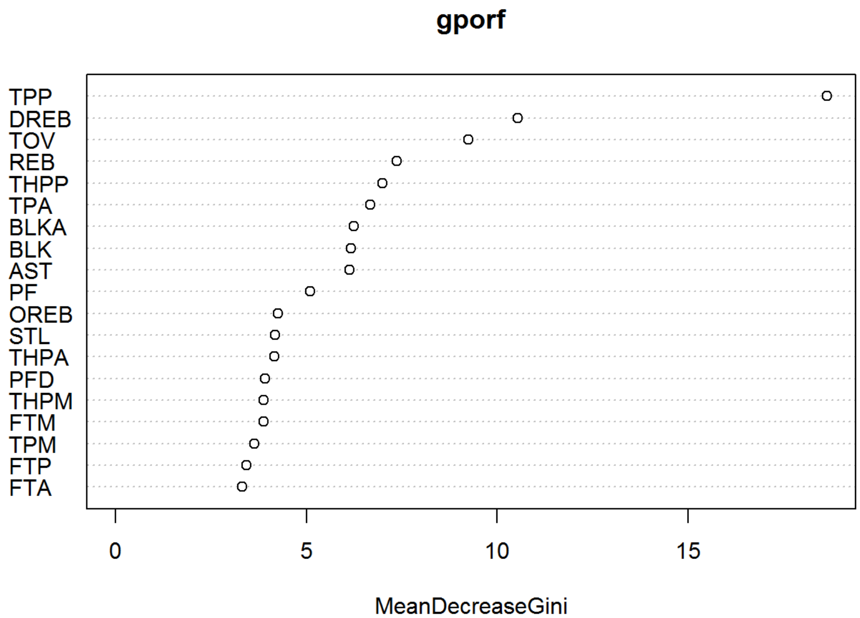
3.2.1 RandomForest

3.2.1.1 Importance of Variables

Code

```
##
## Call:
##  randomForest(formula = PO ~ ., data = traingpo, ntree = 300,      proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 300
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 21.25%
## Confusion matrix:
##    F   P class.error
## F 76  29   0.2761905
## P 22 113   0.1629630
```

Code



- This is the result of the random forest. The error rate is 21.25%.
- `TPP` showed the largest importance. Since `TPP` was the most important factor in winning the game, it seems reasonable.
- `TPP` was followed `DREB` and `TOV`. As we mentioned in the first analysis, rebound refers to making less mistakes. Turn overs are typical variable representing frequency of made mistakes. Variables, representing mistakes, showed large influence in making playoffs.

3.2.1.2 Accuracy of Random Forest, using Test Data

Code

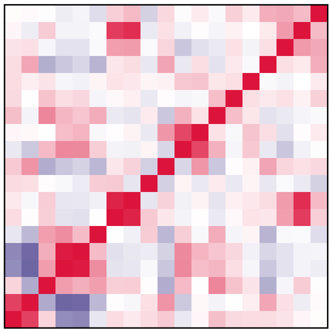
```
## $`confusion matrix`
##
## gpoPred  F  P
##        F 19  2
##        P 14 25
##
## $accuracy
## [1] 0.7333333
##
## $error
## [1] 0.2666667
```

This is the result of the random forest applied to the test data. The error rate is 26.7%, which is little more higher than above, 21.25%.

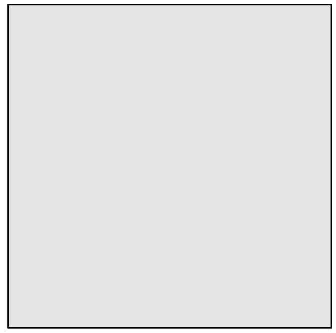
3.2.2 Correlation of Variables

Code

Playoff



Scatterplot



`(TPM, TPA)`, `(THPM, TPA)`, `(THPA, TPA)`, `(THPA, THPM)`, `(FTA, FTM)`, `(FTM, PFD)`, `(FTA, PFD)`, and `(REB, DREB)` showed the absolute value of correlation coefficient higher than 0.6. When you see the above result of variable importance from random forest, `TPA` showed the larger importance than `TPM`, `THPM`, `THPA`. Therefore, we only selected `TPA` among them. `DREB` showed larger importance than `REB` and `PFD` showed larger importance than `FTM` and `FTA`. Hence, we removed `REB`, `FTM` and `FTA`.

3.3 Linear Discriminant Analysis (LDA)

- We used LDA instead of logistic regression, as logistic regression becomes unstable when the classes are well separated and when there are only few data. Although some of the explanatory variables did not satisfy the normality assumption, we had to use LDA for better classification.
- `PO` is used as target variable because the analysis is to classify teams that made playoffs. `TPP`, `DREB`, `TOV`, `THPP`, `TPA`, `BLKA`, `BLK`, `AST`, `PF`, `OREB`, `STL`, `PFD`, and `FTP` were chosen as explanatory variables.

Code

```
## Call:
## lda(PO ~ TPP + DREB + TOV + THPP + TPA + BLKA + BLK + AST + PF +
##     OREB + STL + PFD + FTP, data = traingpo)
##
## Prior probabilities of groups:
##      F      P
## 0.4375 0.5625
##
## Group means:
##        TPP     DREB      TOV     THPP      TPA     BLKA      BLK      AST
## F 47.63082 30.34762 14.79619 34.80190 63.67238 5.108571 4.660952 20.99524
## P 49.43702 31.75037 14.13407 36.16741 61.25481 4.647407 5.013333 22.11333
##         PF     OREB      STL      PFD   FTP
## F 21.09524 11.22190 7.394286 20.42667 75.54
## P 20.33259 10.79333 7.625926 20.79556 75.98
##
## Coefficients of linear discriminants:
##              LD1
## TPP   0.18470909
## DREB  0.27101461
## TOV  -0.61307000
## THPP  0.11809325
## TPA  -0.03471521
## BLKA -0.20815296
## BLK   0.40362758
## AST   0.04709369
## PF   -0.20485638
## OREB  0.20007467
## STL   0.32509012
## PFD   0.32247847
## FTP   0.01473884
```

$$D = 20.77 + 0.18 \cdot TPP + 0.27 \cdot DREB - 0.61 \cdot TOV + 0.11 \cdot THPP - 0.03 \cdot TPA - 0.21 \cdot BLKA +$$
$$0.40 \cdot BLK + 0.05 \cdot AS.T - 0.20 \cdot PF + 0.20 \cdot OREB + 0.33 \cdot STL + 0.32 \cdot PFD + 0.01 \cdot FTP$$

These are the result of LDA. `TOV` showed the largest absolute coefficient. Followed by `BLK`, `STL`, and `PFD`. Variables representing mistakes showed relatively larger influence.

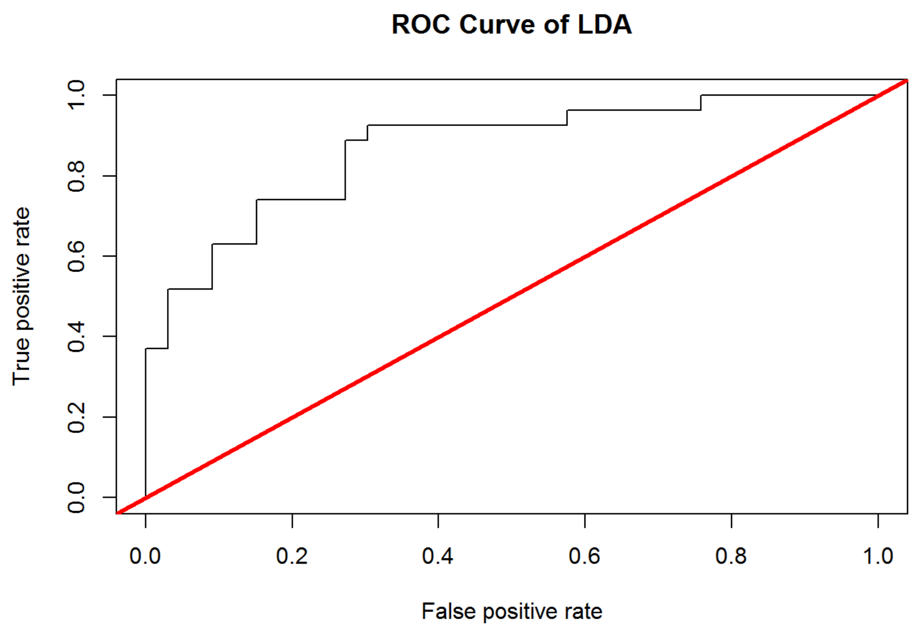
3.3.1 Confusion Matrix, Accuracy, Error, using Test Data

Code

```
## $`confustion matrix`
##       Predicted
## Actual  F  P
##      F 22 11
##      P  2 25
##
## $accuracy
## [1] 0.7833333
##
## $error
## [1] 0.2166667
```

The model’s accuracy was 78.3%. To be specific, most of the teams that have made playoffs were predicted correctly (93% accuracy). However, some teams that failed to enter playoffs were predicted incorrectly (67% accuracy).

Code



Code

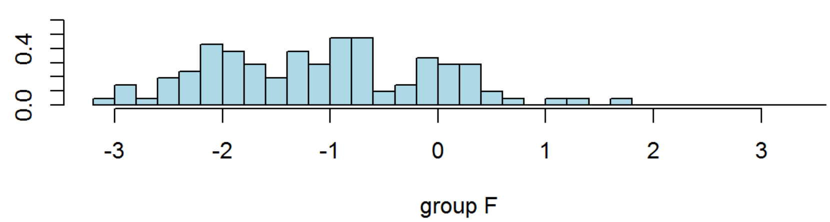
```
## [1] 0.8675645
```

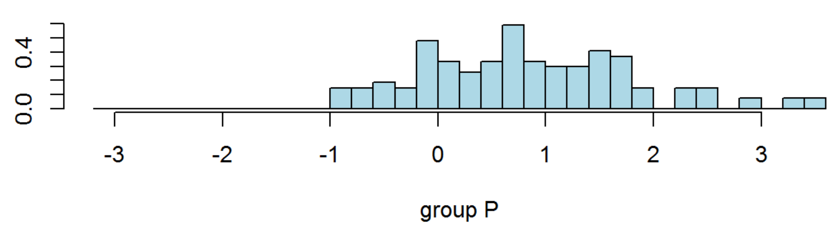
We drew Receiver Operating Characteristic(ROC) Curve to evaluate the model. Area Under the Curve(AUC) was 0.87. LDA classified the data very well.

3.4 Performance of LDA

3.4.1 Histogram of LDA

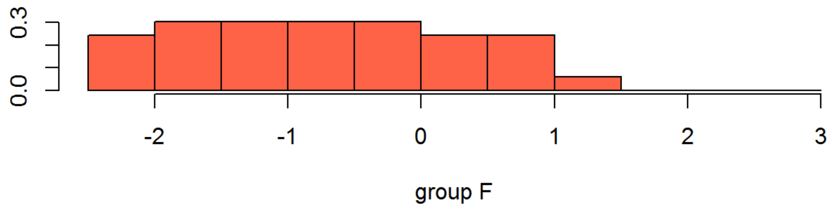
Code

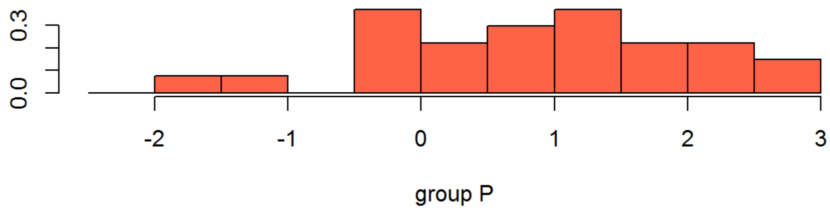




This is a histogram of value of discriminant function, using training data set. They seem to be discriminated quite well.

Code

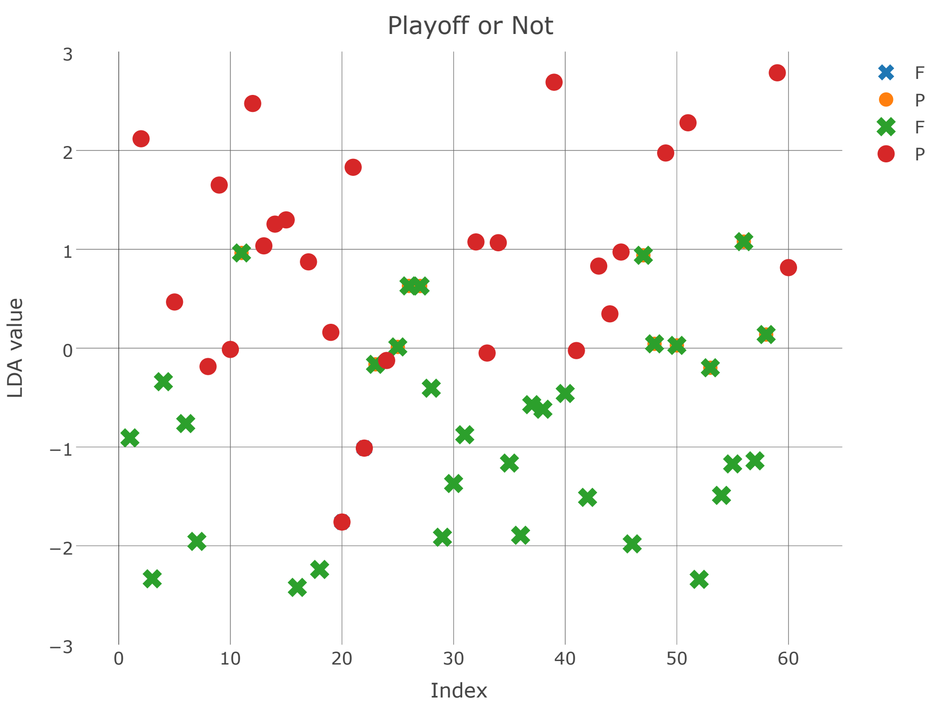




This is a histogram of value of discriminant function, using test data set. They also seem to be discriminated well.

3.4.2 Comparing predicted results with test data

Code



"X" colored in blue and circle colored in orange are the points of predicted results. "X" colored in green and circle colored in red are the points from real test data set. As you can see, even though some of them are not discriminated correctly, it seems fair to say that LDA classified very well in general.

3.4.3 What makes playoffs teams and non-playoffs teams different?

- Choosing Variables

To draw radar plot, we multiplied coefficients of the discriminant function with mean of each variables again, and put variables that had five highest values into the radar plot.

Code

```
##       TPP       TOV      DREB       PFD        PF
##  8.981049 -8.845987  8.444454  6.656708 -4.229465
```

`TPP` , `TOV` , `DREB` , `PFD` , and `PF` were chosen in descending order of influence.

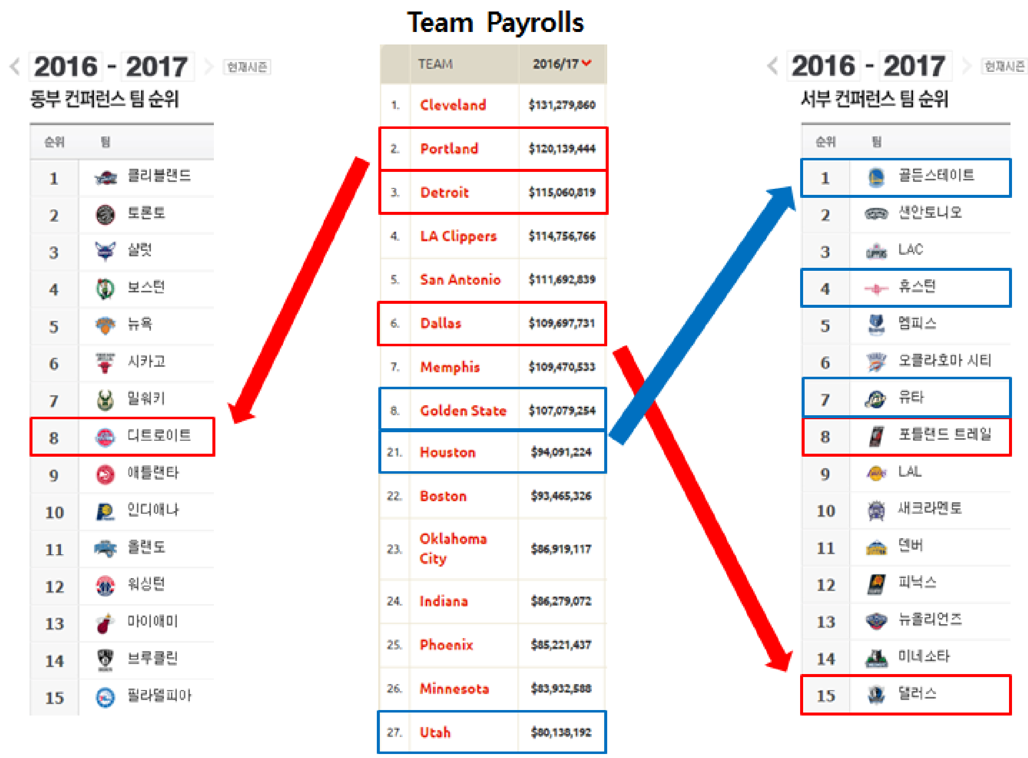
Code

Playoffs teams made less `PF` and `TOV` than non-playoffs teams. In `TPP` , `DREB` , and `PFD` , they were greater than non-playoffs teams. It can be interpreted that playoffs teams make less mistakes, score more points, and draw more mistakes from their opponents.

---

4 Conclusion

- In winning the games, scoring was the most important factor among all. Making less mistakes was also important. However, scoring was much more important.
- In making playoffs, making less mistakes was the most important factor.
- Teams with great players are more likely to score more. However, to possess those players, it requires unmeasurable amount of money. Also, teams with bad teamwork usually makes alot of mistakes. Teams with great teamwork require much less amount of money and are more likely to achieve a long term goal of the season. Nowadays, too much money are spent in the field of sports. The owners of the teams often think money would bring them championship very easily. As you can see from this analysis, it's not. Teams must put more efforts on building team's chemistry to accomplish their long term goals rather than just buying some valuable players.




```
* Writer: Sung-In Cho, Yoon-Seo Jang
* Creation date: Dec 16, 2016
```