Basketball Statistics Analysis

Introduction

We are attempting to use machine learning & statistical analysis techniques to understand which statistics the best variables are to predict which teams will win more than half of their games.

We have done this by using 2 different logistic regression models & also a Random Forest model, to begin to understand which statistics are the most key variables in predicting which teams are winning more often than not.

The reason we are using 2 different logistic regression models, is that there are 3 variables which are not really predictor variables of the Win result. These being Net Rating, O-Rating & D-Rating. O-Rating is the points scored per 100 possessions, D-Rating is points scored against per 100 possessions, Net Rating is the difference of the two.

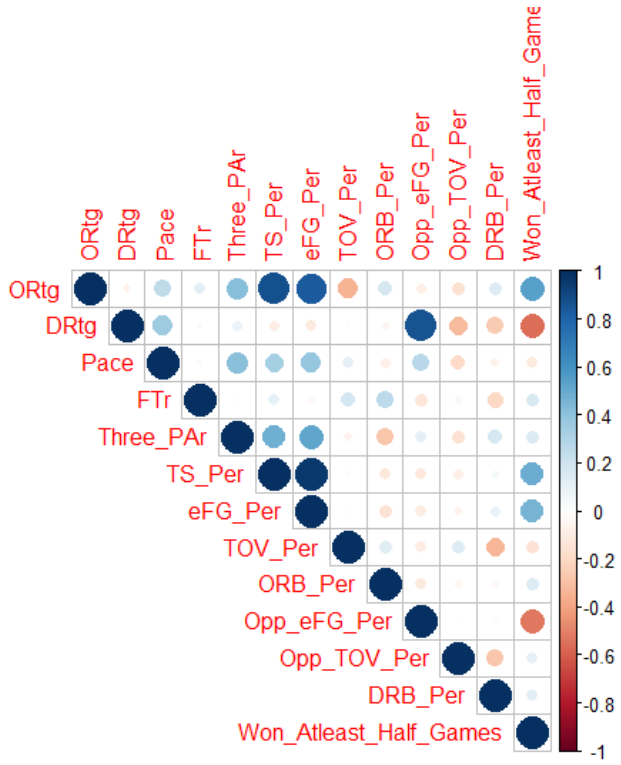
We are going to exclude the Net Rating from both models, as obviously the team that scores more points per possession than the other team is going to win more games. We are also going to exclude O-Rating & D-Rating from the 2nd Logistic Regression model, O-Rating is an obvious predictor, as scoring more points per possession is going to make you win more, D-Rating is also an obvious predictor, as stopping the other team from scoring is going to make you win more. What we want to look into is what variables other than scoring more points or stopping the other team from scoring points will have a relationship with winning games.

Explanation of Variables

- ORtg: offensive rating; number of points scored per 100 possessions
- DRtg: offensive rating; number of points allowed per 100 possessions
- NRtg: net rating, differential between ORtg and DRtg
- Pace: number of possessions per 48 minutes
- FTr: free throw rate; number of free throws for every shot attempt
- 3Par: 3-point attempt rate: percentage of shots that were 3-point attempts
- TS%: true shooting percentage; shooting efficiency that takes into account 2-point shots, 3-point shots, and free throws
- eFG%: effective field goal percentage; field goal percentage that takes into account that a 3-point shot is worth more than a 2-point shot
- TOV%: turnover percentage; percentage of possessions that end in a turnover
- ORB%: offensive rebound percentage; percentage of shots that were offensive rebounded
- Opp_eFG%: opponent effective field goal percentage
- Opp_TOV%: opponent turnover percentage
- DRB%: defensive rebound percentage; percentage of shots that were defensive rebounded
- Won_Atleast_Half_Games: did the team win at least half its games

Correlation Analysis

We plot the correlation matrix to analyse the dataset as a whole:



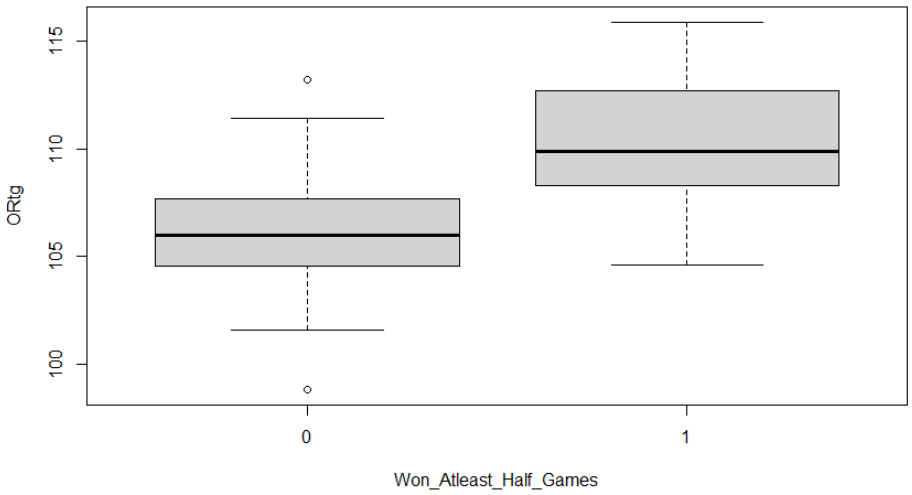
The main things we can gain from this are:

- Winning at Least Half of Games are highly correlated to higher ORtg, lower DRtg, higher TS_Per, higher eFG_Per and lower Opp_eFG_Per. So, the main factors are the Points Per Possession (ORtg & DRtg), True Shooting % Per game and Effective FG% Per game (For & Against)
- ORtg Is positively correlated to TS_Per & eFG_Per. Also weakly positively correlated to Three_Par & Pace. While TOV_Per is weakly correlated to ORtg
- DRtg is positively correlated to Opp_eFG_Per. It's also weakly positively correlated to Pace and weakly negatively correlated to Opp_TOV_Per & DRB_Per

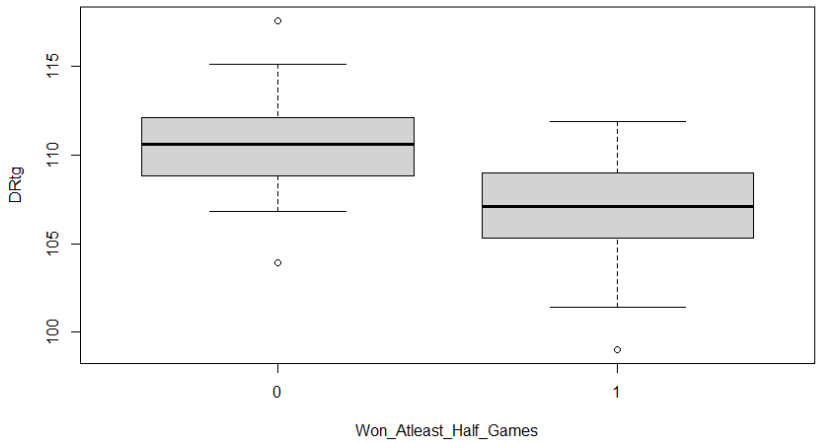
Data Visualisation

Based on this Correlation Analysis, we probably want a slightly better understanding of some of these key variables.

We'll start with ORtg, we can see that teams that are Winning more than half of their games have consistently higher ORtg.

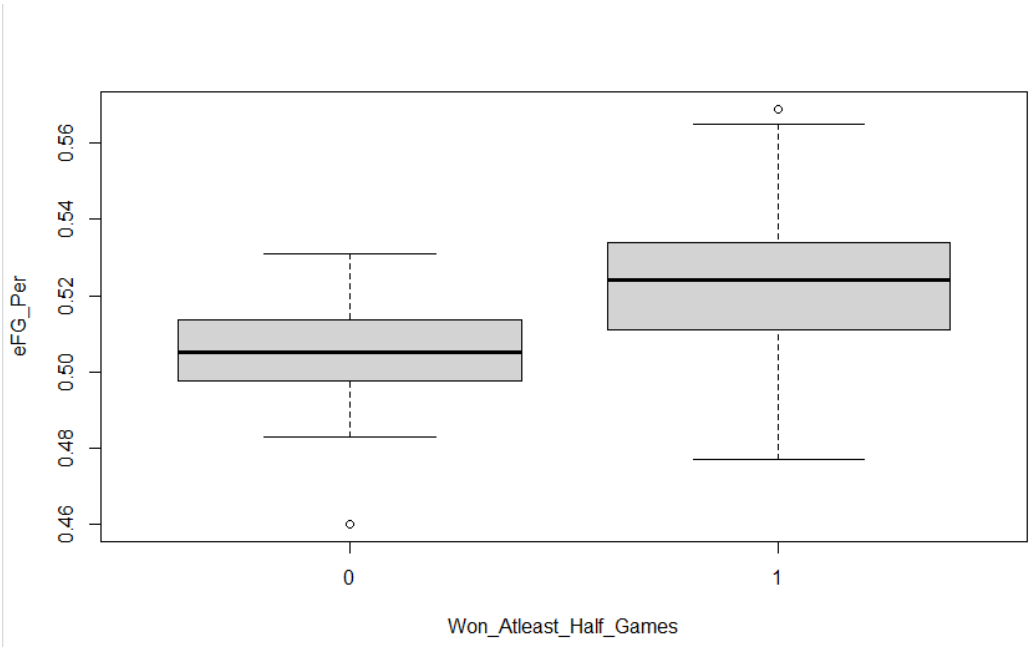


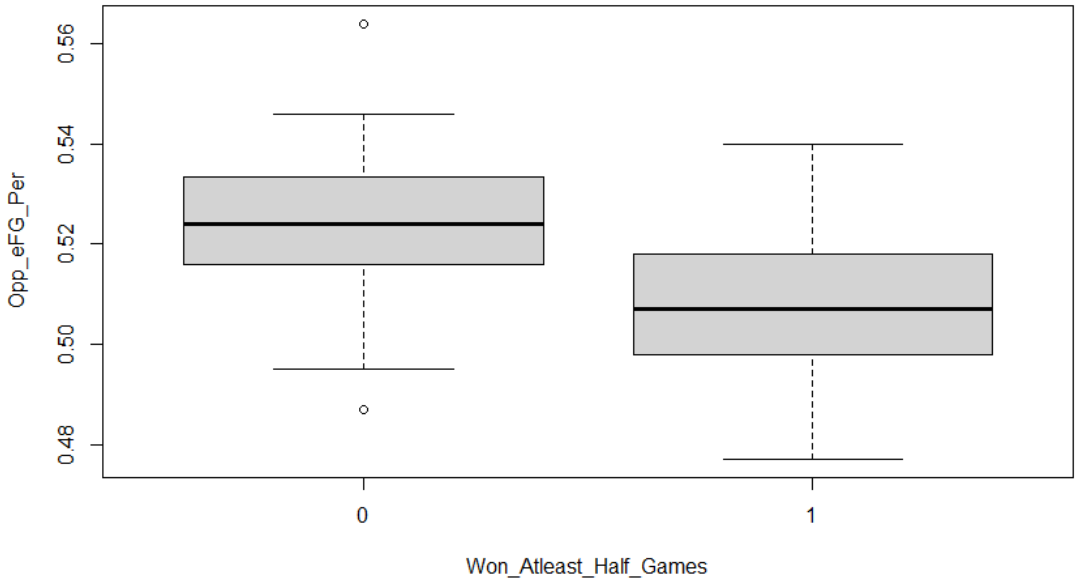
In alignment with the results we had in our Correlation Analysis, we can also see that teams Winning more also have lower DRtg's.



The next plots we're going to analyse is eFG_Per and Opp_eFG_Per.

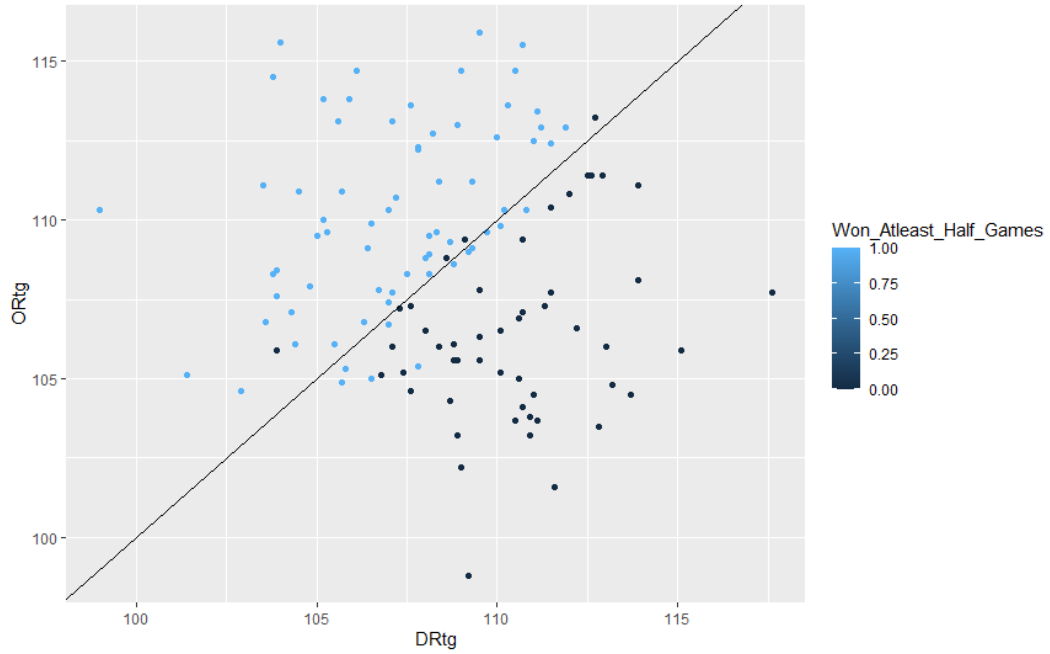
As you can see below, winning teams consistently have higher eFG than losing teams & lower Opponents eFG than losing teams.



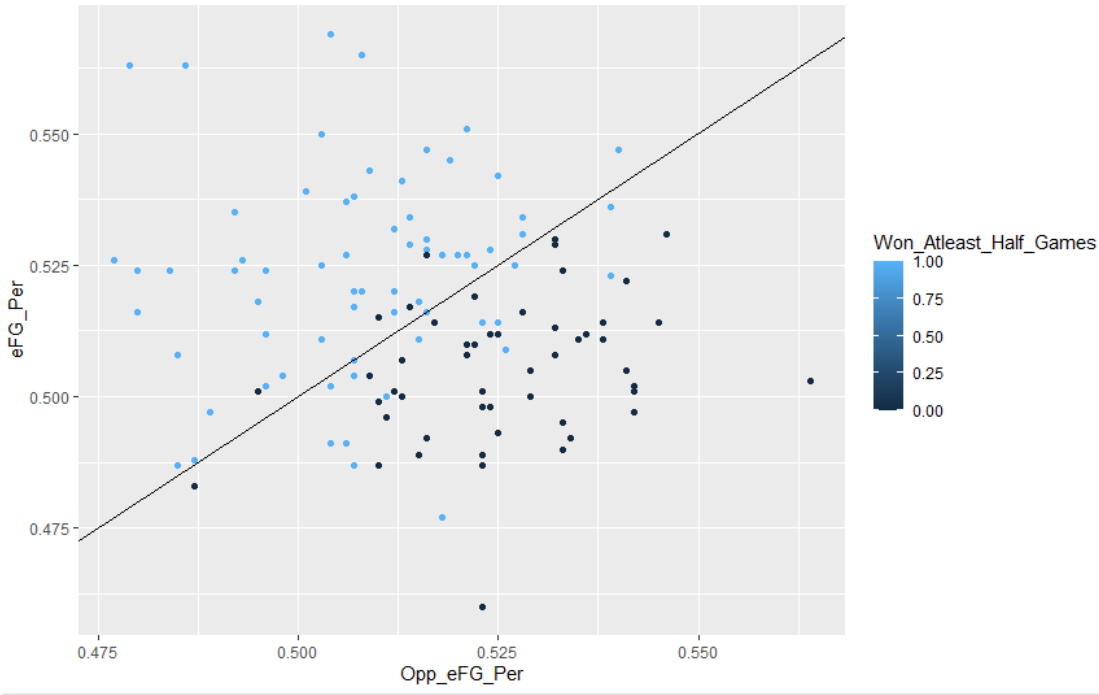


Where it does get interesting is when we analyse ORtg & DRtg against one another.

Here we show a scatter plot of ORtg & DRtg, where losing teams are in Light Blue, Winning teams are in DarkBlue. You can see that most winning teams sit to the Top left of the average line, meaning that they have above average ORtg divided by DRtg, meaning they are scoring more often than their opponent per possession



We are also going to do some similar analysis on eFG vs Opp_eFG. With extremely similar results. The only real difference is that the eFG to Opp_eFG slope line is flatter than the ORtg to DRtg slope line. Which suggests that ORtg & DRtg have a larger spread and are possibly more volatile team-to-team than eFG & Opp_eFG.



Logistic Regression

As previously mentioned, we have derived two logistic regression models, one including ORtg & DRtg, one excluding them. Both sets exclude NetRtg. The data was split into training and testing data, using a 70/30 split.

The results are as below:

Model 1

```
Call:
glm(formula = as.factor(Won_Atleast_Half_Games) ~ ORtg + DRtg,
    family = "binomial", data = training_bball)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.20040  -0.04702   0.00344   0.17604   1.88450

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  71.8416    34.3019   2.094 0.036225 *
ORtg          1.5629     0.4636   3.371 0.000748 ***
DRtg         -2.2093     0.6629  -3.333 0.000860 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102.765  on 75  degrees of freedom
Residual deviance:  23.497  on 73  degrees of freedom
AIC: 29.497

Number of Fisher Scoring iterations: 8
```

We can see that the predictor variables DRtg and ORtg are the contributing factors. With an increase in DRtg having a negative relationship with winning, and an increase in ORtg having a positive relationship with winning. Furthermore, DRtg has a larger coeeficient than ORtg, but this may be due to the intercept being so large.

Using the logistic regression equation $p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots}}$ logistic regression model equation is as follows:

$$p = \frac{e^{71.8416-2.2093(DRtg)+1.5629(ORtg)}}{1+e^{71.8416-2.2093(DRtg)+1.5629(ORtg)}}$$

This model has an AUC of 0.95. The confusion matrix below shows an accuracy of 89%.

	0	1
0	15	5
1	0	24

Model 2

In Model 2, we removed ORtg & DRtg, to try to dig deeper into the factors of the game which contribute towards winning.

```
Call:
glm(formula = as.factor(Won_Atleast_Half_Games) ~ TS_Per + TOV_Per +
    ORB_Per + Opp_eFG_Per + Opp_TOV_Per, family = "binomial",
    data = training_bball)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.61609  -0.00082   0.00000   0.00247   2.41507

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -25.0406    29.3341  -0.854   0.3933
TS_Per       694.2020   295.4942   2.349   0.0188 *
TOV_Per       -5.5305     2.4766  -2.233   0.0255 *
ORB_Per        1.7527     0.6922   2.532   0.0113 *
Opp_eFG_Per -761.9332   304.7618  -2.500   0.0124 *
Opp_TOV_Per    5.4584     2.1450   2.545   0.0109 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102.765  on 75  degrees of freedom
Residual deviance:  14.799  on 70  degrees of freedom
AIC: 26.799

Number of Fisher Scoring iterations: 10
```

We can see that first of all, we get much more statistically significant variables, which means we probably have a more suitable model where the underlying variables are not simply being masked by the more accurate, but less insightful ORtg & DRtg metrics. The variables found to be significant are true shooting percentage, turnover percentage, offensive rebound percentage, opponent effective field goal percentage, and opponent turnover percentage.

The equation produced by this 2nd Logistic Regression model is as follows:

$$p = \frac{e^{-25.0406+694.2020(TS\ \%)-5.5305(TOV\ \%)+1.7527(ORB\ \%)-761.9332(Opp\ EFG\%)-0.028(Opp\ TOV\ \%)}}{1+e^{-25.0406+694.2020(TS\ \%)-5.5305(TOV\ \%)+1.7527(ORB\ \%)-761.9332(Opp\ EFG\%)-0.028(Opp\ TOV\ \%)}}$$

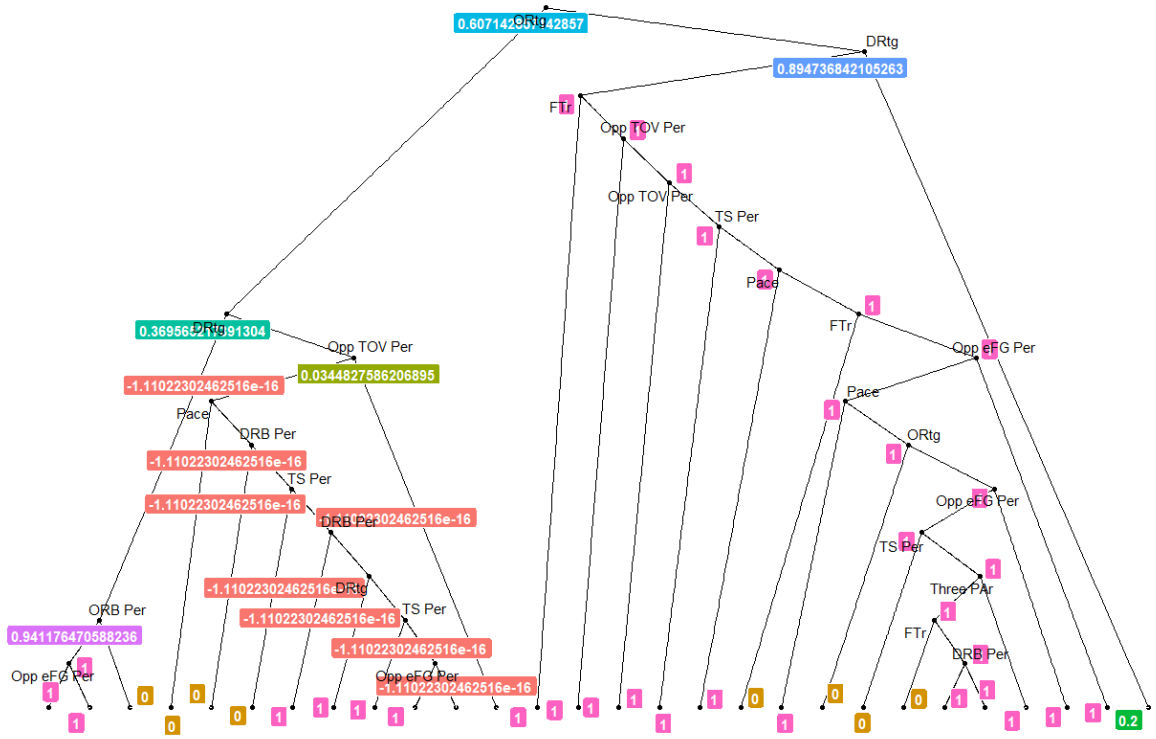
This model has an AUC of 0.97. The confusion matrix below shows an accuracy of 91%. Thus, this logistic regression model performed a little better than the first one.

	0	1
0	16	4
1	0	24

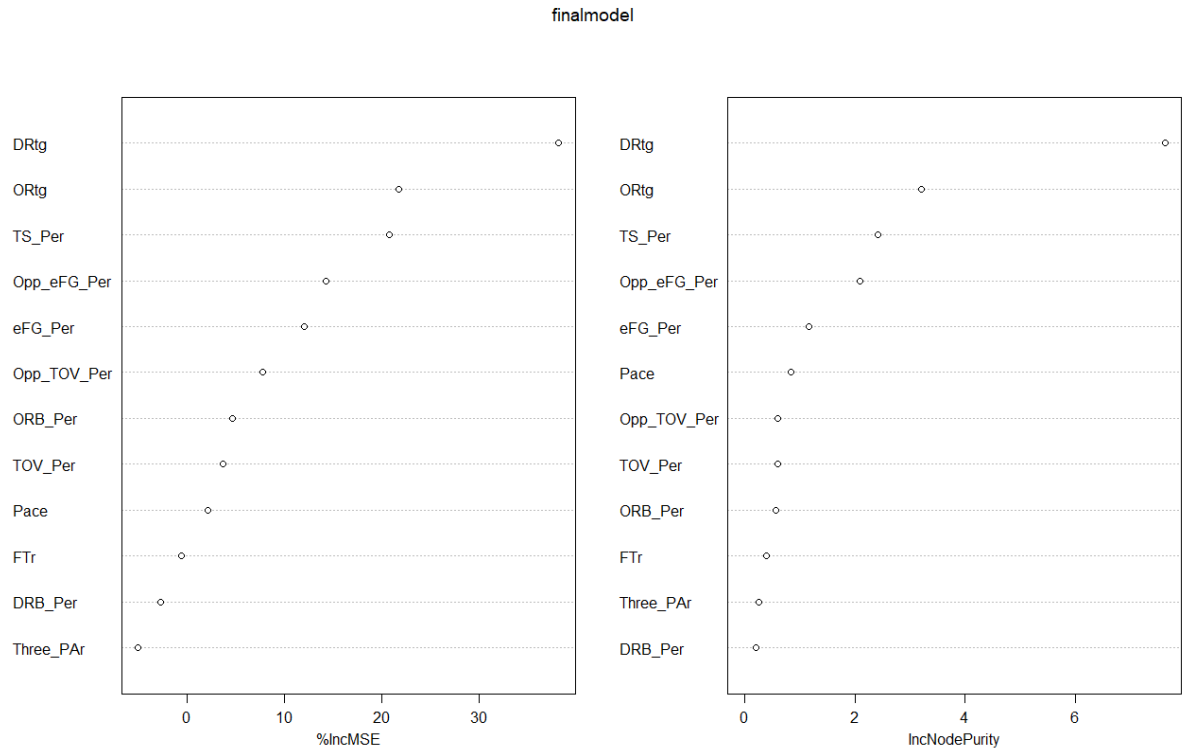
Random Forest Model

We have also developed a Random Forest model, which will help us get an understanding of which variables are most important in predicting Winning Games. After experimenting with various mtry values in the random forest process, which dictates the number of variables randomly selected at each split. After some experimentation, we found that the mtry value & max depth of the tree which optimised the % of Var explained was mtry = 5 and max depth = 18 (with a maximum of 59.84% of Variance in Winning Half of Games explained by the model).

The Tree Diagram of the Random Forest Model can be seen below:



And we can see the importance of each variable below:



What is promising from this Random Forest model, is that it confirms our previously discussed variables. We see DRtg, ORtg, both Opponent & Team eFG_Per both high on the list, TS_Per high on the list and again Opp_TOV_Per showing to be somewhat relevant. So this helps us confirms our original analysis from the Correlation Plot & Logistic Regression models.

The AUC of this model is 0.99. The confusion matrix below shows an accuracy of 94%. This shows that when you have all of these statistics available you can almost completely predict if a team will win half its games.

```
   0  1
0 13  1
1  1 21
```

Conclusion

As you would expect the 2 most important factors are DRtg & ORtg, as they are quite simply measures of how well a team can score more points & stop the other team from scoring points, per possession.

Once we look past those 2 factors, a number of interesting factors come into play.
The main ones are Effective FG% (eFG_Per & Opp_eFG_Per), True Shooting % (TS_Per)

and Opponent Turnovers per game (Opp_TOV_Per). These factors make a lot of sense as to how they are important in winning games, Effective Field Goal % and True Shooting % are both weighted measures of field goal percentage. For example, in Effective Field Goal %, shooting 30% on 3 pointers and 45% on 2 pointers, would give you an equivalent eFG%. As from 100 shots, both of those scenarios is expected to score you 90 points.

The other main predictor is Opponent Turnovers per game, if a team can force the other team to turn the ball over, then the team will get more possessions. If 2 teams both have a 40% True Shooting %, the pace is 200 possessions per game between the 2 teams, each team will have 100 possessions each & score 80 points each. However, if one team turns the ball over 10 times more than the other team, then one team gets 110 shots, and the other gets 90. Then we see the final score being 88 to 72. So, it makes sense that forcing the other team into turning the ball over will help you win games.

One of the logistic regression models also suggests that getting more Offensive & Defensive Rebounds also helps teams win games, this also helps teams get more possessions. If you get an Offensive Rebound, you get an extra shot. If you get a Defensive Rebound, you are stopping the other team from getting an Offensive Rebound and thereby them getting another shot.

So, to summarise, the blueprint for winning games is:

- Be more efficient shooting the ball than the other team (Higher eFG_Per, Lower Opp_eFG_Per and Higher TS_Per)
- Make the other team turn the ball over more (Higher Opp_TOV_Per)
- There is also some evidence that Rebounds will help you win games also (Higher ORB Per & Higher DRB Per)