Charles Cantwell

Data Analysis of NBA 2016-2017 Season

Friday, December 9, 2022

Resources:

https://statisticsglobe.com/split-data-frame-in-r

https://www.quora.com/What-is-a-non-technical-way-to-interpret-varImpPlot-chart-from-Random-Forest-in-R

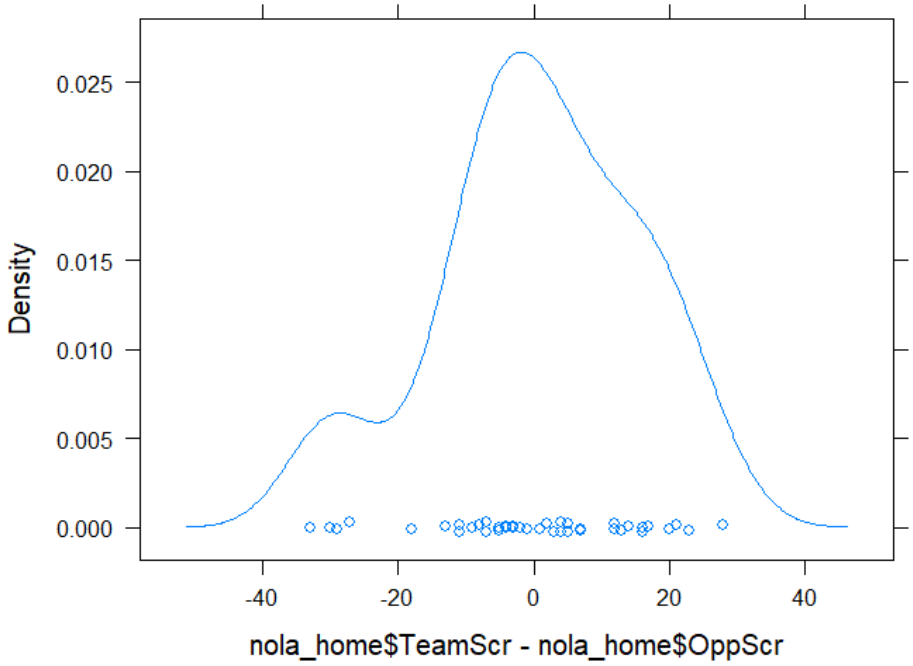
https://en.wikipedia.org/wiki/Efficiency_(basketball)

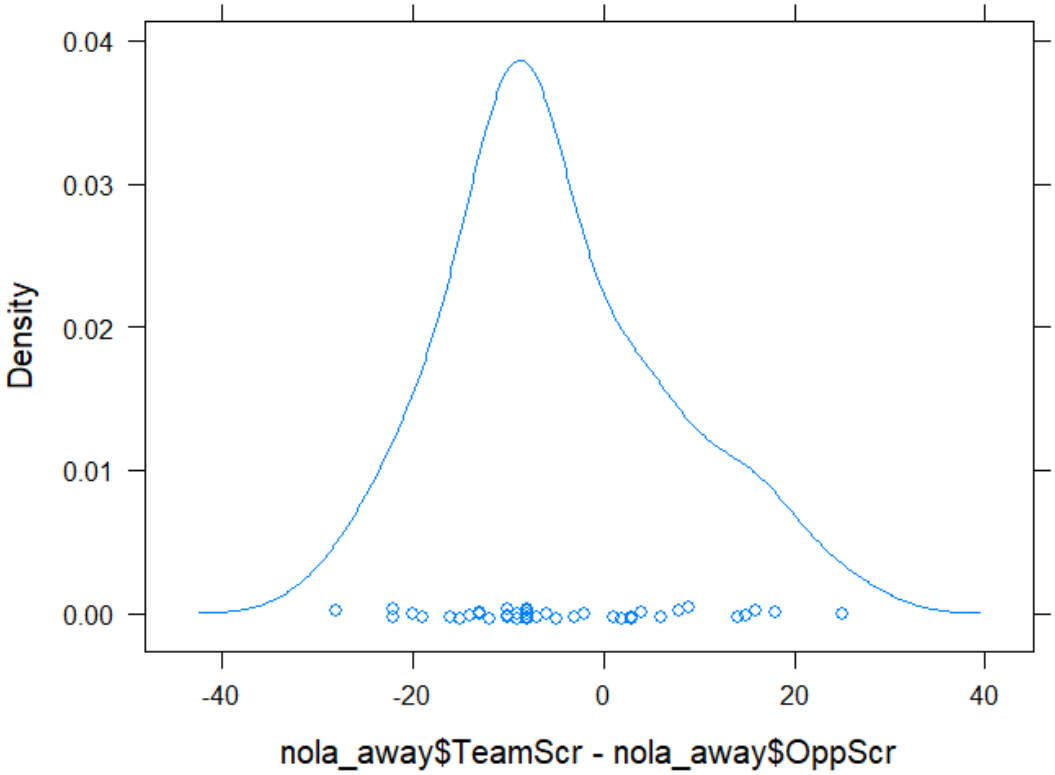
https://www.digitalocean.com/community/tutorials/r-squared-in-r-programming

Abstract:

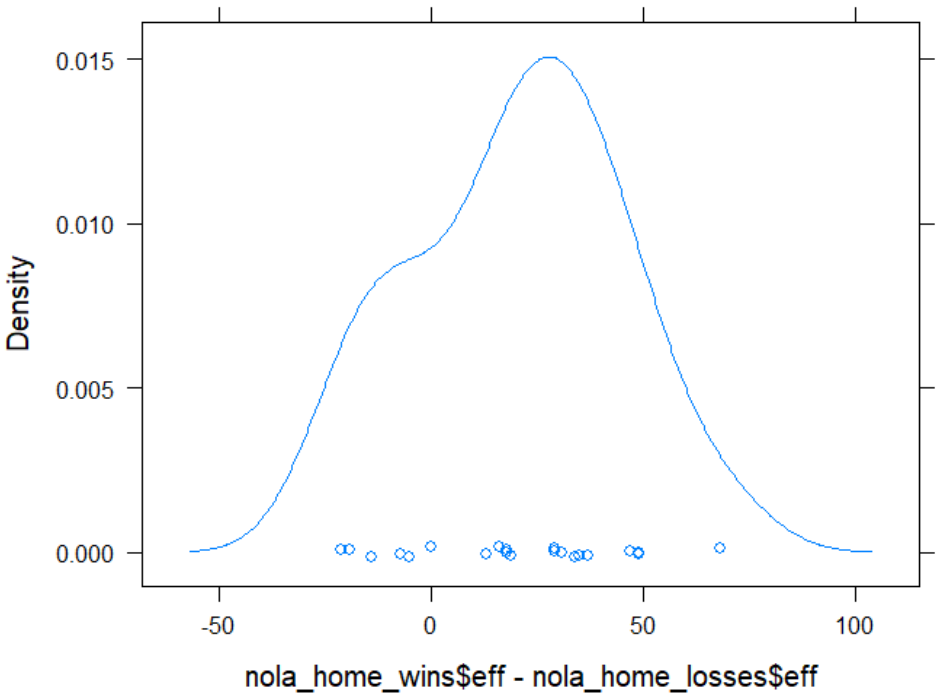
An analysis of the New Orleans Pelicans for the 2016-2017 season was conducted in order to help determine various factors that would help predict a win for the Pelicans in their 82 games that season. It was found that for away games, the Pelicans seemed to lose more at home than when the Pelicans were playing away. Additionally, at the away games, the Pelicans would lose very consistently by anywhere between 20 to 3 points. However, the difference between the Pelican's score and the opponents score is more evenly distributed about the center of 0 point difference. Additionally, it was found that Opponent Score (OppScr), Efficiency (eff), Pts (Points), Team Score (TeamScr), Defensive Rebounds (Dreb), and Field Goals Made (Fgm) are effective predictors for predicting a win for the Pelicans as visualized in the VarImpPlot. However, the direct use of OppScr minus TeamScr was not used to determine a win for the Pelicans using the Random Forest model. The second part of this analysis is to predict the scores of the home and away team for NBA playoff games using a random forest model with data from the regular season. In addition, the championship series scores were predicted.

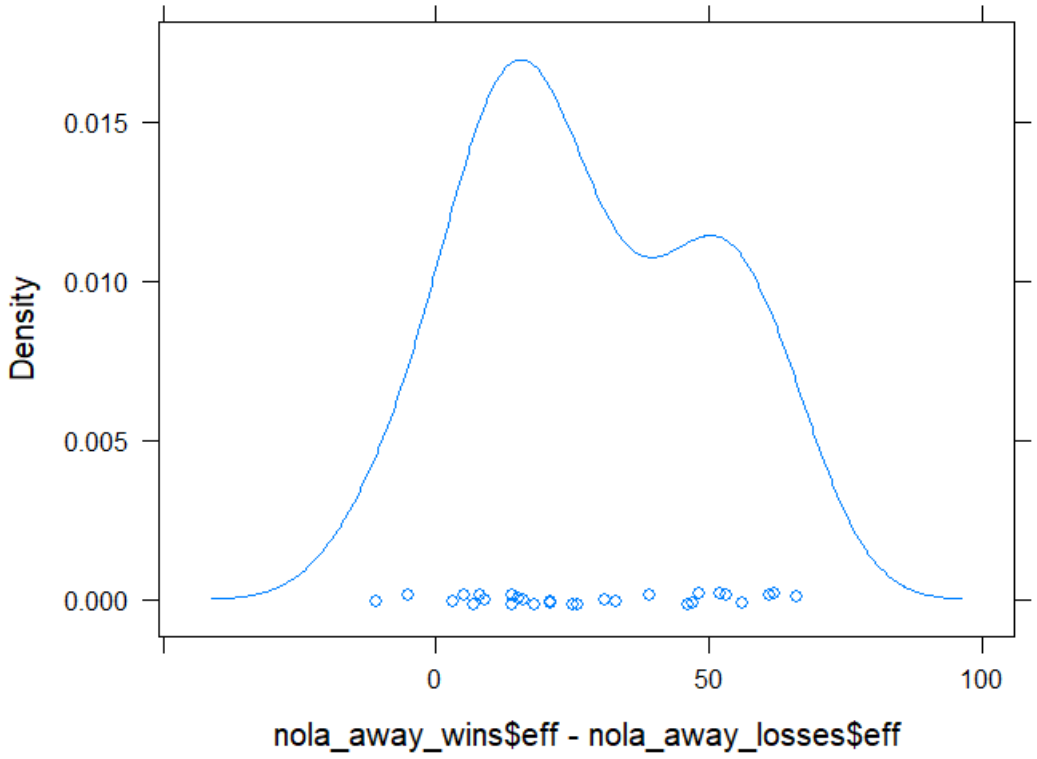
For the New Orleans Pelicans, the 2016-2017 season was a losing season with a record of 48 losses and 34 wins with a total of 82 games played. The density plots shown below show the difference between the Pelicans' scores and the opponent for home and away games. Given that the Pelicans lost 8 more games when playing away, it makes sense that the distribution of score differences is centered more to the left for away games. There is also a more centralized peak in the away graph with a more normal and right skewed distribution.


Density
0.025
0.020
0.015
0.010
0.005
0.000
-40
-20
0
20
40
nola_home$TeamScr - nola_home$OppScr


Density
0.04
0.03
0.02
0.01
0.00
-40
-20
0
20
40
nola_away$TeamScr - nola_away$OppScr

A random forest model was fitted to the data to help determine which variables would be effective in predicting a win for the Pelicans. The predictors OppScr (Opponent Score), eff (Efficiency), TeamScr (Team Score), Dreb (Defensive Rebounds), Fgm (Field Goals Made) seem to be the most effective in predicting Team Win for the Pelicans. Using the eff predictor to compare Pelican wins and losses and subtracting the win minus the lost distributions of eff, we see a large difference between the eff of wins and losses eff distribution. Only a few exceptions were found in both home and away games. This shows that eff is a good predictor of a win.


0.015
0.010
0.005
0.000
Density
-50
0
50
100
nola_home_wins$eff - nola_home_losses$eff


0.015
0.010
0.005
0.000
Density
0
50
100
nola_away_wins$eff - nola_away_losses$eff

Again a random forest model was used to predict the scores of both teams because random forest tends to be very effective when given a mix of categorical and regressive (numerical) data. However the random forest model is not easily interpretable like a linear regression model with easily understood coefficients. So, a VarImpPlot (Variable Importance Plot) is used to see which predictors is most effective in predicting TeamScr (Team Score) and, in the second model, OppScore (Opponent Score). These models are trained on games played in the regular season and will be used to predict TeamScr and OppScr for the playoff games and the championship series between the Golden State Warriors and Cleveland Cavaliers.

As for prediction accuracy of the TeamScr model for the playoff games, the random forest model performed very well with a R squared of 0.998 and a percent error of 0.378%. However the prediction accuracy of OppScr is not very high as expected (since we are using one team's data of rebounds, free throws, field goals made or attempted) to predict the score of another team. The model for OppScr prediction has a R squared of 0.621 and a percent error of 5.80%. Thus, I chose to use the model that predicts TeamScr to predict the scores of the Cleveland Cavaliers and Golden State Warriors championship series. The results of these predictions was the model predicted TeamScr of Cleveland with an R squared of 0.996 and a percent error of 0.794%, and the model predicted TeamScr of Golden State with an R squared of 0.9998 and a percent error of 0.0978%.

By using the statistic of eff for each team in the regular season and the final series, we can see that the Warriors got worse in the finals and the Cavaliers got worse versus the regular season. The Warriors averaged 143.8415 eff in the regular season and 143 eff in the finals, and the Cavaliers averaged 123.4146 eff in the regular season and 122.8 eff in the finals