Machine Learning Report

Curtis Higa

February 5, 2018

For this project, I looked at the signifcance of the three point shot in the NBA to winning. Winning can be thought of in two ways: winning a specific game or an above average winning percentage for an entire season.

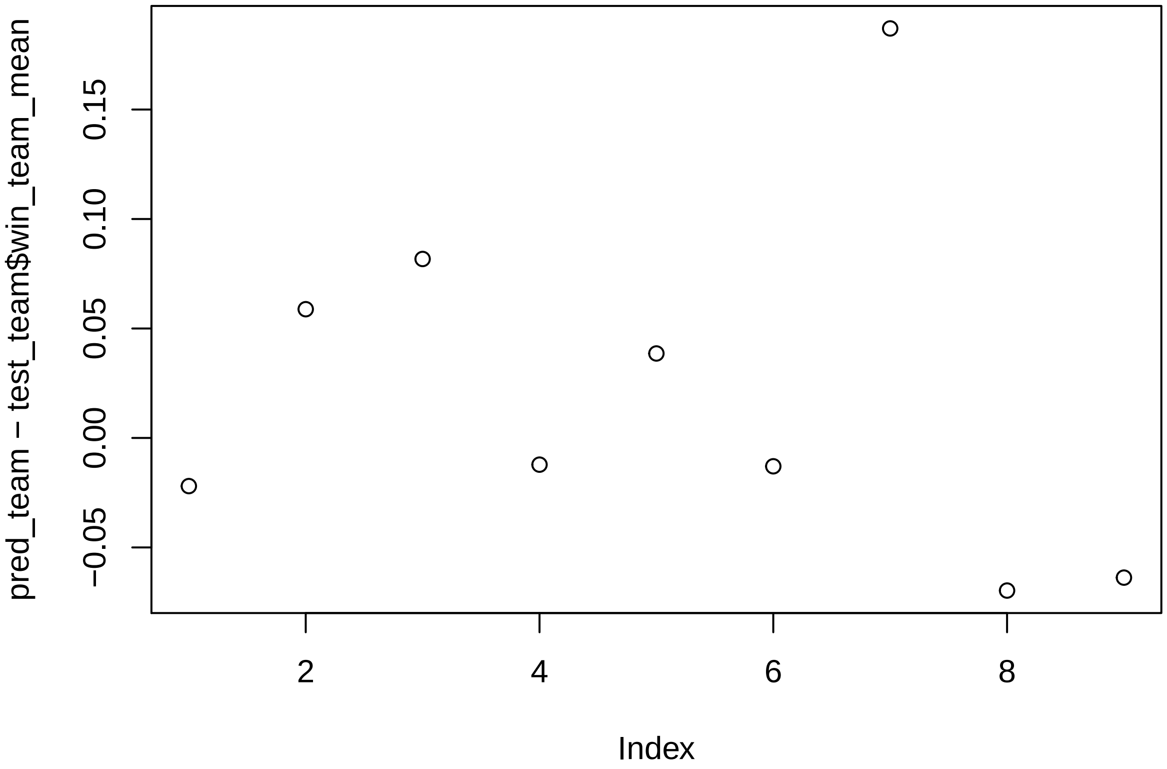
When trying to predict each teams winning percentage, a linear regression model was used using the independent variables *three_pt_Percentage_mean* and *opp_three_pt_Percentage_mean* which are the three point percentage and allowed three point percentage respectively for the team over the course of the season. The model is shown below as *reg_mode*.

```r
reg_mode <- lm(win_team_mean ~ three_pt_Percentage_mean + opp_three_pt_Percentage_mean,
               data = team_statistics)
summary(reg_mode)
```

```
##
## Call:
## lm(formula = win_team_mean ~ three_pt_Percentage_mean + opp_three_pt_Percentage_mean,
##     data = team_statistics)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17033 -0.05472  0.01265  0.06035  0.15848
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.415593   0.653277   2.167 0.039232 *
## three_pt_Percentage_mean      0.033095   0.008733   3.790 0.000770 ***
## opp_three_pt_Percentage_mean -0.058770   0.013488  -4.357 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08401 on 27 degrees of freedom
## Multiple R-squared:  0.647,  Adjusted R-squared:  0.6209
## F-statistic: 24.75 on 2 and 27 DF,  p-value: 7.848e-07
```

As shown in the summary, the R^2 term is *0.647* which indicates it's a fairly decent model and both independent variables are indicated with "***" signifying they are significant. After splitting the data set into a training and test set 70-30 respectively and creating a new model using the training data, the following prediction model was generated.

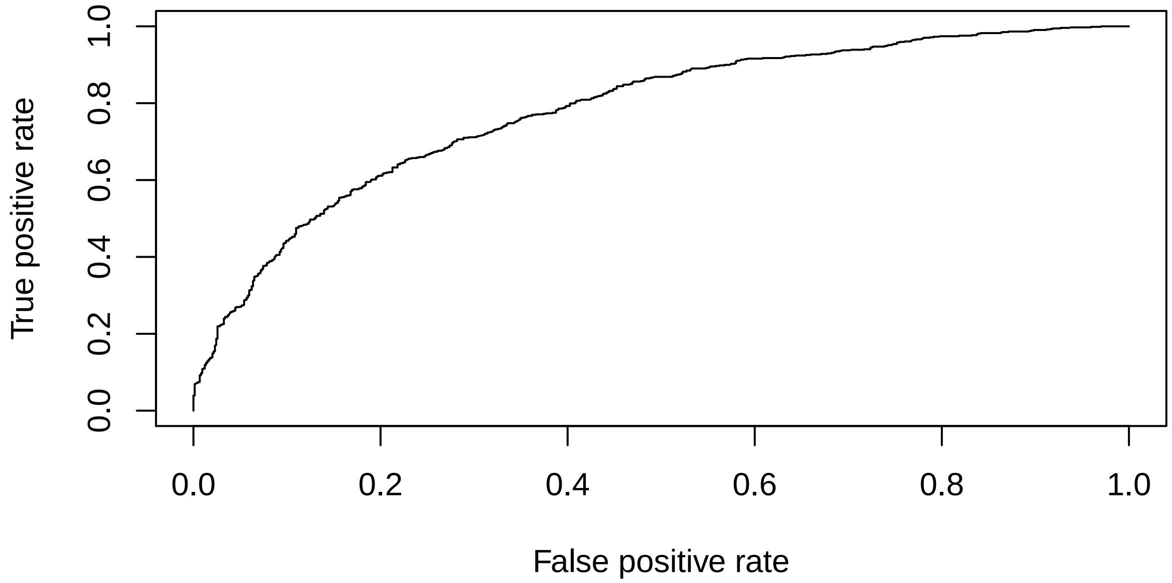
```r
mod_team <- lm(win_team_mean ~ three_pt_Percentage_mean + opp_three_pt_Percentage_mean,
               data = train_team)
pred_team <- predict(mod_team, type = "response", newdata = test_team)
plot(pred_team - test_team$win_team_mean)
```



The above plot shows how the residuals are distributed among the test set and it shows the residuals are randomly plotted. This means the model produces no bias towards one result and is therefore valid. However, the performance of the model may not be great due to a couple reasons. One of which may be because of the sample size. The team statistics are for all the teams that played in the 2016-2017 NBA regular season which is 30 teams. Another reason may be because there are other factors of basketball that may contribute more to a higher winning percentage than just focusing on three point shots. In the future, more factors and data entries could be added to improve this model.

Models predicting game-by-game outcomes were also investigated and considered as a classification problem due to there being exactly one winner and one loser per game. On a game-by-game basis, we're trying to build a model to predict a winner based on the various final three point statistics and the two three point percentage statistics were most relevant resulting in the following logistic regression model, where *train* is the training data.

```
mod1 <- glm(win_team ~ three_pt_Percentage + opp_three_pt_Percentage,
            data = train,
            family = "binomial")
ROCRperf <- performance(ROCRpred, "tpr","fpr")
plot(ROCRperf)
```



```
table(test$win_team, predictTest >= 0.5)
```

```
##
##      FALSE TRUE
##   0   375  117
##   1   138  354
```

With a threshold value of 0.5, this model correctly predicted 74% of the test data indicating it's a good model shown in the confusion matrix above. However, this model takes into account middle and end game statistics to determine a winnner which is good for which team has a high probability of winning at a specific moment of the game. It doesn't include predetermining factors similar to venue or active players to predict a winner before the game occurs. If we want to be able to predict the winner of a game before the game occurs, we'd need independent variables like active players, venue, and home team.