Author: Roshni Vadiraja

Introduction

Sports Analytics is one of the fastest growing areas under analytics. All professional sports teams and leagues make use of analytics to analyze their performance, identify areas of strengths and weaknesses and make improvements as and when needed. They also use them to study their opponents' strengths and weaknesses and game plan against them effectively. Analytics in sports are also used to identify the workload on players, provide them with ample rest in order to avoid recurring injuries. There is a large volume of interesting and varied data available which makes it an interesting field to work in.

The objective of this project is to use an NBA statistics dataset, identify strengths and key performance indicators.

Research Questions

1. Is home advantage across teams an important factor affecting wins and losses
2. Mapping a teams' scoring pattern using a decision tree across 4 seasons. This allows us to understand their scoring range and allow us to know the keys to scoring efficiently.
3. Predicting wins and losses using a few key performance indicators

Chosen Data set

We chose the NBA statistics for teams from 2014-2018 that was available on Kaggle. The data set contains 41 columns containing data points for each game such as result for the home and away team, points scored, shots attempted, scored, free throws attempted vs conceded etc. The statistics are presented for each game across 4 seasons and include all home and away games. 4920 games are covered, and each game has 2 records one for the home team

and the other as the away team. This makes it a total of 9840 rows in the data set. We have a lot of data on hand which allows us to run multiple different tests and make predictions.

Columns available in the dataset are listed below

S.No
Team
Game
Date
Home
Opponent
WINorLOSS
TeamPoints
OpponentPoints
FieldGoals
FieldGoalsAttempted
FieldGoals.
X3PointShots
X3PointShotsAttempted
X3PointShots.
FreeThrows
FreeThrowsAttempted
FreeThrows.
OffRebounds
TotalRebounds
Assists
Steals
Blocks
Turnovers
TotalFouls
Opp.FieldGoals
Opp.FieldGoalsAttempted

Opp.FieldGoals.
Opp.3PointShots
Opp.3PointShotsAttempted
Opp.3PointShots.
Opp.FreeThrows
Opp.FreeThrowsAttempted
Opp.FreeThrows.
Opp.OffRebounds
Opp.TotalRebounds
Opp.Assists
Opp.Steals
Opp.Blocks
Opp.Turnovers
Opp.TotalFouls

Analysis

Identifying Home Advantage

NBA teams place a lot of emphasis on home advantage. Teams that have home advantage in play offs tend to win those matches. Teams try to win more matches in regular season in order to gain home advantage for playoffs. Here we use proportion test to examine statistically if home advantage really exists. We take an example of 2 teams that make play off frequently (Boston and Toronto).

First step would be to load the data set

```
NBAStats <- read.csv(file = 'nba.games.stats 3.csv') # Reading the data file
```

Next would be split the data by teams

```
df <- split(NBAStats, NBAStats$Team) # splitting the data by team
TorontoStats <- df$TOR # Contains the data of Toronto team
BostonStats <- df$BOS # Contains the data of Boston team
```

The data frame can be split into home and away games

```
#For Toronto games
THome <- df1$Home # contains the data of Home games
TAway <- df1$Away # contains the data of Away games

#For Boston games
BHome <- df2$Home
BAway <- df2$Away
```

This can again be broken up into wins and losses, home and away

```
#For Toronto games
THomeProp <- table(THome$WINorLOSS) # number of wins and losses at home
TAwayProp <- table(TAway$WINorLOSS) # number of wins and losses away

#For Boston games
BHomeProp <- table(BHome$WINorLOSS)
BAwayProp <- table(BAway$WINorLOSS)
```

The next step is to construct a proportion table

```
TProptable <- rbind(THomeProp,TAwayProp) # Constructing proportion table for Toronto
BProptable <- rbind(BHomeProp,BAwayProp) # Constructing proportion table for Boston
```

This data set can be used to verify the Null hypothesis

Null Hypothesis (H_0): There is no difference in the number of wins home and away

Alternative Hypothesis (H_1): The number of losses at home is less than the number of losses away

```
prop.test(TProptable, alternative = "less") # proportion test for the hypothesis for Toronto; "losing percentage in home games is less than away games"
```

```
2-sample test for equality of proportions with continuity correction

data:  TProptable
X-squared = 9.1265, df = 1, p-value = 0.00126
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.07352279
sample estimates:
   prop 1    prop 2
0.2621951 0.4268293
```

prop.test(BProptable, alternative = "less") # proportion test for the hypothesis for Boston;

```
2-sample test for equality of proportions with continuity correction

data:  BProptable
X-squared = 2.8525, df = 1, p-value = 0.04562
alternative hypothesis: less
95 percent confidence interval:
 -1.000000000 -0.002829048
sample estimates:
   prop 1    prop 2
0.3536585 0.4512195
```

The test performed above performs a 2 proportion Z-test. We can compare the P-Value versus the significance level to validate the hypothesis. We can see from both results that the P-value is less than the Significance level. This implies that the null hypothesis can be rejected which in our case is no difference between proportions home and away.

Decision Tree to understand scoring pattern

A decision tree is a graph based structure used in data mining to dig through available data, map options and to predict outcomes. This is typically to find the best possible course of action when multiple options are available and to predict results based on the multiple paths taken.

With the dataset that is being used, the most obvious analysis a team would want to know is, what is the most efficient way to score. We used a decision tree to try and use assists, 3 point shots scored and free throws scored to understand how they affect scoring and what would be the best way to score.

Here data is divided into 2 sets. A train set consisting of 70% of the data and a test set containing the other 30%. We will use the train set to create the model and the test set to create decision using the model created.

First step is to import necessary libraries and load the data set

```
#import libraries

install.packages("rpart")

library(rpart)

install.packages("party")

library(party)

#to invoke the external dataset in NBAStats variable

NBAStats <- read.csv(file = 'nba.games.stats.csv')
```

Next step is to split the dataset by team and isolate Bostons dataset

```
#split the data by team

team <- split(NBAStats, NBAStats$Team)

team

#Boston team data
```

```
BostonStat <- team$BOS

BostonStat
```

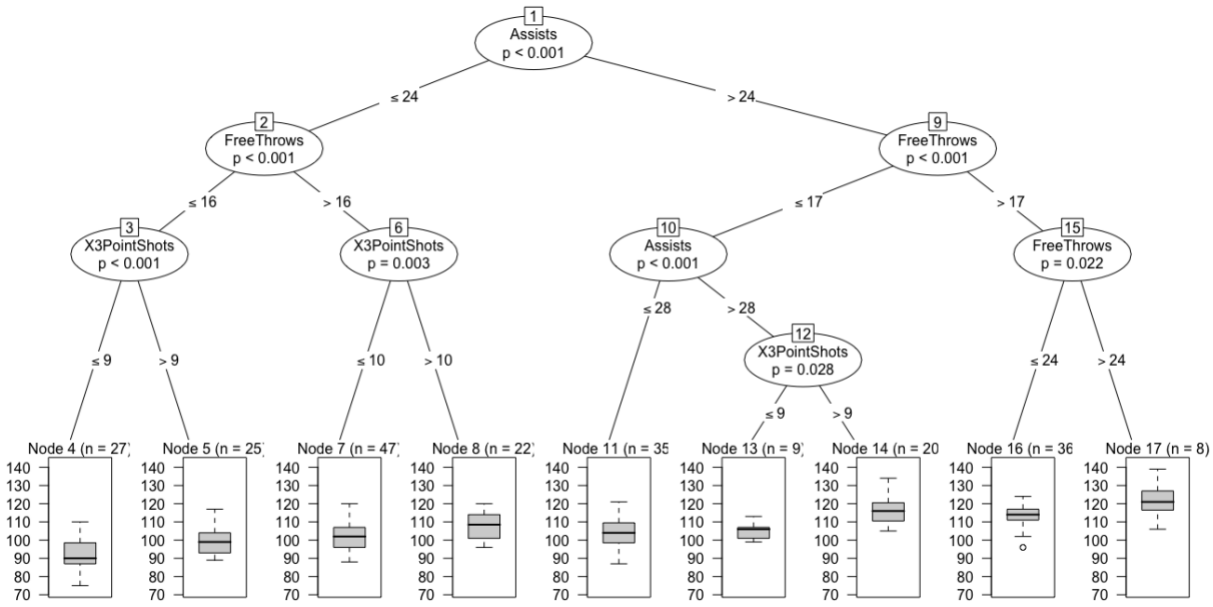
Next up, we build the train and test data sets

```
set.seed(101)

alpha     <- 0.7 # percentage of training set

inTrain   <- sample(1:nrow(BostonStat), alpha * nrow(BostonStat))

train.set <- BostonStat[inTrain,]

test.set  <- BostonStat[-inTrain,]
```

The last step is to build the decision tree and use it to make predictions

```
# Decision Tree

BosTree <- ctree(TeamPoints~ X3PointShots+FreeThrows+Assists, data = train.set)

plot(BosTree)

#Predictions

predict(BosTree, test.set, type="prob")
```

The decision tree we created looks like this



Here we can clearly see that the teams highest scoring averages are when the assists and free throws are > 24 each or Assists are greater than 24 and 3 point shots scored are > 9. So, this clearly shows that when the team is sharing the ball, scoring improves. Scoring more 3 point shots and free throws help increase the score as well

Logistic Regression to Predict Wins and Losses using 3 key performance indicators

Linear regression is a form of predictive analysis where one dependent variable depends on one or more independent variable. The outcome here is a binary variable say 1 or 0, Right or left, Success or Failure and Win or Loss.

Here , we try to predict wins or loss using logistic regression. The dependent variable is the result (win or loss). The dependent variables we are using are Field goals taken, 3-point shots taken, and free throws taken.

Here, we divided the data into 2 sets. A train set consisting of 80% the data and a test set consisting of 20% of the data. We will use the train data to create the model and use this model the predict the results on the test data set.

First step is to load the data set

```
library(caTools)

NBAStats <- read.csv(file = 'nba.games.stats 3.csv') # Reading the data file

NBAStats
```

This data is split into train and test data as discussed

```
# Split Data

split <- sample.split(NBAStats, SplitRatio=0.8)

split

train <- subset(NBAStats, split=TRUE)

test <- subset(NBAStats, split=FALSE)
```

The model will be built next

```
#Building binomial logistic regression

model<- glm(WINorLOSS~FieldGoals+X3PointShots+FreeThrows, data = train, family =
"binomial")

summary(model)
```

From the model built, the results are as follows

```
Call:

glm(formula = WINorLOSS ~ FieldGoals + X3PointShots + FreeThrows,

    family = "binomial", data = train)

Deviance Residuals:

    Min       1Q   Median       3Q      Max
```

-2.7740 -0.9592 -0.0263 0.9687 2.5847

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.891577	0.248451	-39.81	<2e-16 ***
FieldGoals	0.190936	0.005614	34.01	<2e-16 ***
X3PointShots	0.097926	0.006875	14.24	<2e-16 ***
FreeThrows	0.094389	0.004093	23.06	<2e-16 ***

Here we can see that the Z values for the data set are 34.01, 14.24 and 23.06 for field goals, 3-point shots and free throws. This means that Win or Loss records are most dependent on field goals, followed by free throws and lastly 3-point shots taken.

The residuals as we can see are quite symmetric around the 0 which means that the model, we have built is a good one. The significance value is also very low and close to zero. This means that the dependent variable is quite dependent on the independent variables we have chosen.

The predictions are created in the next step

#Predictions

```
res <- predict(model,train, type = "response")

res

res <- predict(model,test, type="response")

res
```

We use a confusion matrix to validate our results

#Build confusion matrix

```
confmatrix <- table(Actual_Value=test$WINorLOSS, Predicted_Value= res>0.5)

confmatrix
```

	Predicted_Value	
Actual_Value	FALSE	TRUE
L	3432	1488
W	1497	3423

Here we were able to get 6855 results right and 2975 results wrong which is a success rate of 69.67%. The model we built was able to predict more than 2/3rds of the results correctly.

Conclusion

From this, we can make the following 3 conclusions

Using a 2 proportion Z-test, we can compare the P-Value versus the significance level to validate the hypothesis. We can see from both results that the P-value is less than the Significance level. This implies that the null hypothesis can be rejected which in our case is no difference between proportions home and away.

The proportion of home losses is less than the proportion of away losses for both Boston and Toronto. This shows that home advantage is real. Proportion tests don't give us a reason for this. We should analyze other factors such as shooting percentages and free throws to understand this.

From the decision tree built, we understand that assists play the most important role in increasing scoring. The more a team plays unselfishly and the more they pass the ball, scoring improves. High number of assists also result in higher number of free throws taken and higher number of 3 point shots scored which increase scoring as well. So, playing unselfishly is the key to scoring well.

From the results of the logistic regression test, we can conclude that given the low significance value for the 3 independent variables, win or loss depends on number of field goals taken, number of 3-point shots taken and field goals taken significantly. Off these, field goals have the most weightage, followed by free throws and 3-point shots taken respectively. Using these 3 variables, we can predict a win or loss with an accuracy of close to 70%.