PRACTICAL DATA SCIENCE
COSC 2670
COMPUTER SCIENCE &
SCHOOL OF SCIENCE

# Statistical Classification of NBA Data

## Data Modelling and Presentation of NBA Statistics

**Supervisor: Professor Yongli Ren**

Aidan Cowie – S3429481

Oliver Eaton – S3641518

*26 May, 2018*

# 1 Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Aidan Cowie and Oliver Frederick

*26 May, 2018*

# 2 Abstract

The purpose of this research is to examine the use of different statistical classification methods for predicting the outcome of basketball matches in the National Basketball Association (NBA). Statistical data from 4920 NBA games were collected and used to train and test a variety of classification techniques such as the KNN method and Decision Tree.

Classification was chosen as the model to identify new observations as a “Win” or “Loss”. Methods used in this research are the supervised machine learning algorithms K-Nearest Neighbours (KNN) and Decision Trees. The hill climbing method, with random feature selection, was used to determine the value of neighbours and features used in the models.

The experiment showed how adding and removing features differed the overall results in the predictability of the outcome of an NBA game via Decision Trees and the K-NN Methods. The K-NN method provided a more accurate predictability in determining the outcome of an NBA basketball match with the Cross validation across five folds of data produced precision ratings between 76 - 79%.

Future research could look at analysing individual players and see how their performances influence match outcome.

# 3 Contents

# 4 List of Figure and Tables

## 4.1 List of Figures

## 4.2 List of Tables

# 5 Introduction

In this paper we examine the use of machine learning as a tool for predicting the success of basketball teams in the National Basketball Association (NBA). Further, we investigate which subset of features input to the different methods are the most salient features for prediction. When it comes the National Basketball Association (NBA) the seriousness of competition becomes real. Players strive to be the best, coaches strive to devise a unique style of play to beat competitors and franchisees pay an exuberant amount of money to put together the best team to win a championship. So, when it comes to playing 'styles', which style is superior to other? Which style is more likely to determine the success of a team, coach and franchise? This research aims to analyse the driving factors behind classifying a match in the NBA as a win or loss. Methods used in this research are machine learning algorithms K-Nearest Neighbours (KNN) and Decision Trees.

## 5.1 KNN Literature Review

K-Nearest neighbours is a machine learning classification technique that incorporates two important concepts. The first concept is the method of calculating distance between two points or observations in a dataset. Calculating 'distance' can take various forms, it can be calculated using Manhattan distance; a combination of a horizontal and vertical path, Euclidean distance; a diagonal or more "direct" path; or more arbitrarily using Minkowski distance which generalises both Euclidian and Manhattan distance (Hassan, Aickelin, & Wagner, 2014). Figure one shows a representation of the diverse types of distances; the red line is Manhattan distance; the green line is Euclidian distance and the blue line is Minkowski distance. The choice of distance metric used depends on the dataset being analysed and is ultimately decided upon by the scientist carrying out the research.

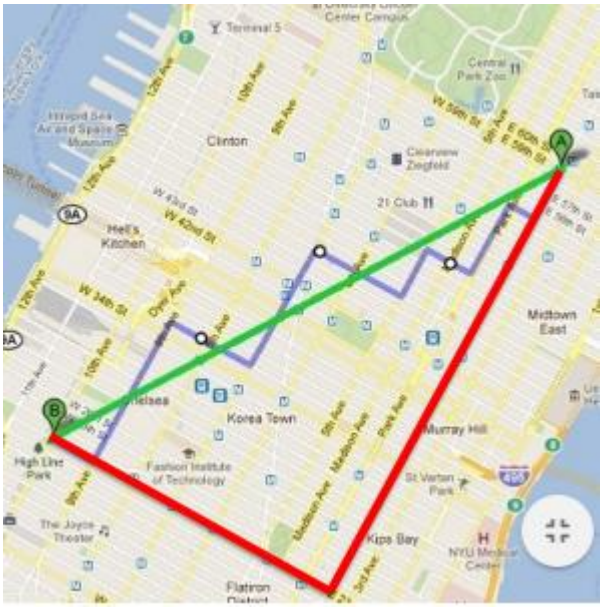



*Figure 1 A representation of Manhattan, Eculidian and Minkowski distance metrics used in the KNN algorithm. (Zhang, 2016)*

The second major concept is the number of neighbours that defines a 'neighbourhood', the *k* parameter. The appropriate choice of k has impacts diagnostic performance of the algorithm. A large k reduces variance caused by random error but runs the risk of ignoring small but important patterns. An appropriate k value strikes a balance between overfitting and underfitting (Zhang, 2016) . Figure two shows a representation of over an overfitted and well fitted model; the green path is overfitted, the black path is well fitted. Overly fitted models although represent the current dataset perfectly, does not generalise well on unseen data.

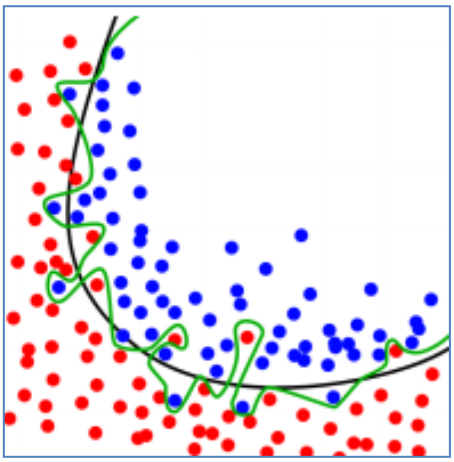


*Figure 2 A representation of overfitting and underfitting a model (Zhang, 2016)*

# 6 Decision Tree Literature Review

Decision Trees is a non-parametric supervised learning model for classification and regression (Sehra, 2018). The trees learn from the data set with a set of if-the-else statements from here they approximate a sine curve.

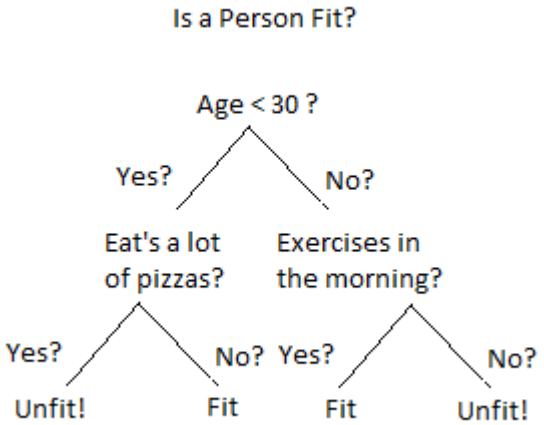



*Figure 3 Decision Tree Example (Sehra, 2018)*

The tree brakes down the data set into small subsets, and incrementally develops decision nodes and leaf nodes for the data subsets (Sehra, 2018). Thus, decision trees can tend to over fit the data as they keep making nodes and leaf nodes until it fits all the exceptions in the data set shown in Figure 2 (Zhang, 2016). A decision node shown in figure 3 has two or more branches that stems down to the next best predictor. The top most node is called the root node which corresponds to having the best

predictability properties. The leaf nodes represent the decision or classification that path or node which the tree expanded to (Sehra, 2018).

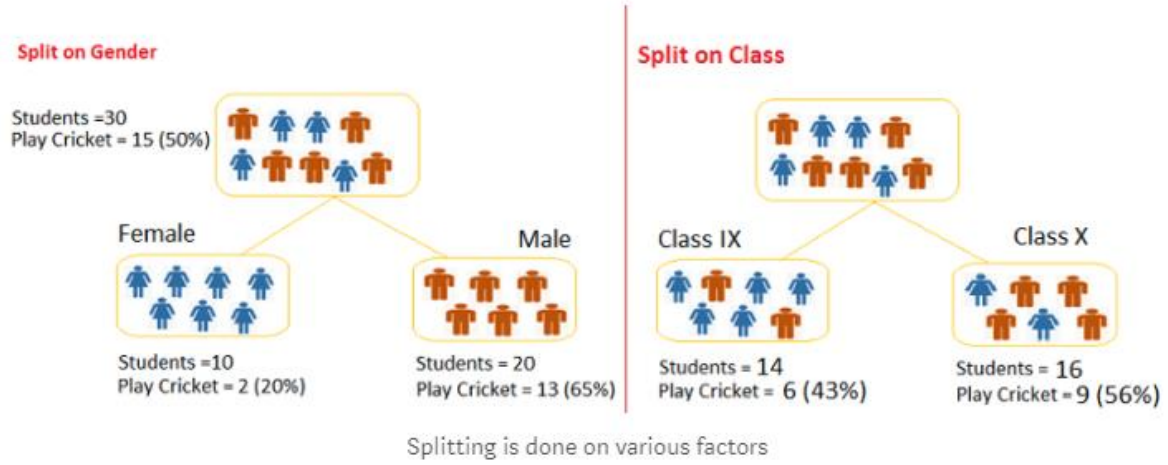
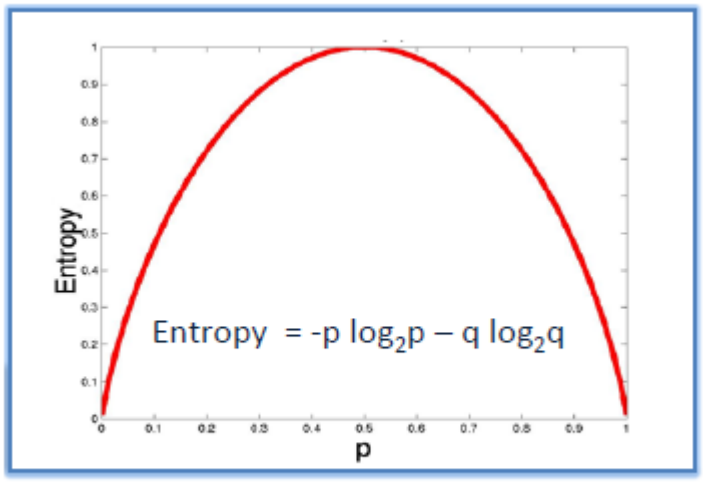



*Figure 4 Example of splitting (Sehra, 2018)*

Figure 4 demonstrates and example of the tree spitting the data into smaller subsets on the left-hand side we can see the splitting of male and female, and on the right shows the splitting of classes. The process of splits is based of particular variables called features (Sehra, 2018).

Pruning of the trees helps to eliminate over fitting its shortening the branches, the process of size reduction by transforming decision nodes into leaf's. It also removes all of the corresponding and dependent branches off that node. The shorter and simpler the tree, the less over-fitting errors when coming into classifying new data (Sehra, 2018).

Entropy is a major factor on finding the smallest tree which fits the data, this corresponds with the tree that has the lowest cross-validation error (Sehra, 2018).




$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

*Figure 5 Entropy Graph (Sehra, 2018)*

As decision trees are built top-down starting at the root node, involving partitioning the data into homogeneous subsets. Entropy for a homogeneous subset is zero and if the sample is equally divided by out comes the entropy is one thus, the tree uses entropy to determine the homogeneous subsets shown in figure 5 (Sehra, 2018).

# 6 Materials and Experimental Methodology

## 6.1 Materials

The dataset in focus consists of 9841 observations and 41 variables. The complete list of variables and its description is provided in **Table 1**.

| Variable Name | Description | Variable type |
|---|---|---|
| Assists | Team assists | Ratio |
| Blocks | Team blocks | Ratio |
| Date | Date of match played | Date |
| Field Goals | Team field goals (2-point shots) | Ratio |
| FieldGoals. | % of successful team field goals | Ratio |
| FieldGoalsAttempted | Team field goal attempts | Ratio |
| FreeThrows | Team free throws | Ratio |
| FreeThrows. | % of successful team free throws | Ratio |
| FreeThrowsAttempted | Team free throw attempts | Ratio |
| Home | Whether team is playing at home | Binary {Home, Away} |
| OffRebounds | Team offensive rebounds | Ratio |
| Opp.3PointShots | Opponents three points shots | Ratio |
| Opp.3PointShots. | % of opponents successful three point shots | Ratio |
| Opp.3PointShotsAttempted | Opponents three points shots attempted | Ratio |
| Opp.Assists | Opponents assists | Ratio |
| Opp.Blocks | Opponents blocks | Ratio |
| Opp.FieldGoals | Opponents field goals | Ratio |
| Opp.FieldGoals. | % of opponent's successful field goals | Ratio |
| Opp.FieldGoalsAttempted | Opponents field goals attempted | Ratio |
| Opp.FreeThrows | Opponents free throws | Ratio |
| Opp.FreeThrows. | % of successful opponents free throws | Ratio |
| Opp.FreeThrowsAttempted | Opponents free throws attempted | Ratio |
| Opp.OffRebounds | Opponents offensive rebounds | Ratio |
| Opp.Steals | Opponents steals | Ratio |
| Opp.TotalFouls | Opponents total fouls | Ratio |
| Opp.TotalRebounds | Opponents total rebounds | Ratio |
| Opp.Turnovers | Opponents turnovers | Ratio |
| Opponent | Team playing against | Categorical |
| OpponentPoints | Opponents match points | Ratio |
| Steals | Team steals | Ratio |
| Team | Team | Categorical |
| Game | Unique number of a match in the dataset | Categorical |
| TeamPoints | Team match points | Ratio |
| TotalFouls | Team total fouls | Ratio |
| TotalRebounds | Team total rebounds | Ratio |
| Turnovers | Team turnovers | Ratio |
| WINorLOSS | Whether team won or lost | Binary {Win, Loss} |
| X3PointShots | Team three point shots | Ratio |
| X3PointShots. | % of successful team three point shots | Ratio |
| X3PointShotsAttempted | Team three point shots attempted | Ratio |

*Table 1 Variables and its Description*

An unnamed variable was dropped because it did not contain any information. Each recorded match has a row from the perspective of the home team and another row from the away team, resulting in two rows per game. Because each row also contains statistics of the "opposition" when developing a KNN

model, the dataset was filtered to include only records from the home team perspective. After exploration of the dataset, new attributes were created. These include:

*WinPoints, LossPoints, WinFieldGoals, LossFieldGoals, WinFieldGoalsAttempted, LossFieldGoalsAttempted, WinFieldGoals., LossFieldGoals., Win3PTShots, Loss3PTShots, Win3PTAttempted, Loss3PTAttempted, Win3PT., Loss3PT., WinFreeThrows, LossFreeThrows, WinFreeThrow., LossFreeThrow., WinOffRebounds, LossOffRebounds, WinTotalRebounds, LossTotalRebounds, WinAssists, LossAssists, WinSteals, LossSteals, WinBlocks, LossBlocks, WinTurnovers, LossTurnovers, WinTotalFouls, LossTotalFouls.*

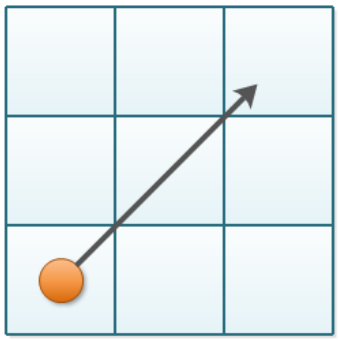
The new attributes enable exploration of each column to be tabulated by the non-mutually exclusive categories home team, away team, winning team and losing team.

## 6.2 Data Modelling

### 6.2.1 KNN Model

Classification was chosen as the model to identify new observations as a "Win" or "Loss". Because observations in the dataset were tagged as a win or lose, this is a supervised learning problem and the K-Nearest Neighbours algorithm was chosen as the classification technique. The hill climbing method, with random feature selection, was used to determine the value of neighbours and features used in the model. The 'weight' used in the model was *uniform,* which weights point in a neighbourhood equally. *P*, the power parameter for the Minkowski metric, was set to 2. This is equivalent to using "euclidian distance" so the distance between two points was calculated as a straight line.

**Euclidean Distance**



$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

In the initial model the training set contained variables:

*FieldGoals., X3PointShots., FreeThrows., OffRebounds, TotalRebounds, Assists, Steals, Blocks, Turnovers, TotalFouls, Opp.FieldGoals., Opp.3PointShots., Opp.FreeThrows., Opp.OffRebounds, Opp.TotalRebounds, Opp.Assists, Opp.Steals, Opp.Blocks, Opp.Turnovers, Opp.TotalFouls.*

Percentage values for field goals, three-point shots and free throws were chosen over successful shots and attempts because it penalises unsuccessful attempts. The dataset was split into training and test data at a split of 70% and 30%, respectively. The hill climbing method, used to select the k value and

features, looped through a training set and iteratively increased the value of neighbours by one with each loop ranging from one to ten. In each iteration, features were randomly selected and added to the model until predictive power stopped increasing. The number of neighbours, features present and predictive power of each iteration was output in a table and graph.

From the first model some trends contain few features but have similar predictability to iterations containing over 10 features. The trends with high predictability but few features contain the features shot percentage (*FieldGoals.*), opponent shot percentage (*Opp.FieldGoals.*), three-point shot percentage (*3PT*), and opponent three-point shot percentage (*Opp.3PT*). It seems intuitive that features directly *are* the match outcome and including these features does not satisfy the curiosity to determine the driving factors behind NBA match outcome. A new model was run on a data frame with these features removed. The new data frame contained features:

*OffRebounds, TotalRebounds, Assists, Steals, Blocks, Turnovers, TotalFouls, Opp.OffRebounds, Opp.TotalRebounds, Opp.Assists, Opp.Steals, Opp.Blocks, Opp.Turnovers, and Opp.TotalFouls*

that were available for model selection.

As can be seen in **Figure 38**, classification rate increases with the value of k. Therefore, a k-value of 10 was determined. By analysing the common features in each iteration, the features chosen to include in the model were the features which appeared most frequently. These features were: *Assists, TotalRebounds, Blocks, Steals, Opp.Assists, Opp.TotalRebounds, Opp.Blocks,* and *Opp.Steals. We* used cross-validation to address overfitting by training and testing data on different portions of the dataset.

### 6.2.2 Decision Tree Model

We chose the decision tree as the second method for modelling the data set as it could handle categorical data, giving us a larger range of features to test the probability that effects the outcome of determining the “Win” or “Loss” of a game from our data set. We separated the data set into two different separate data sets one with just the home games as we found the games were doubling up effecting our decision trees model’s accuracy, and one that left bot the home and away games for each team. Each of the features with categorical values had to be reassigned numerical values as place holder and have no incremental relevance so that the “.fit()” method would run in python. Once the data was curated our “target” and “data” frames were created consisting with the features of “WinorLoss” for our target and; *OffRebounds, TotalRebounds, Assists, Steals, Blocks, Turnovers, TotalFouls, Opp.OffRebounds, Opp.TotalRebounds, Opp.Assists, Opp.Steals, Opp.Blocks, Opp.Turnovers, and Opp.TotalFouls,* for our data to replicate the KNN method to compare results. Using the Hill-Climbing technique in combination with the decision tree method to determine the features that have the best predictability properties. Using the test size of 30% of the data set and a max depth on the decision tree to having 4 branches to limit over fitting. These features were determined to be; OffRebounds, TotalRebounds , Steals , Turnovers, Opp.OffRebounds, Opp.TotalRebounds, Opp.Steals. To enforce cross—validation the 5-fold method was used with a random state of 4.

All the variables were kept to the standard and enforced in each stage so that there was consistency though this experiment. The features of “Team” and “Opponent” were added into the initial features for the second round of each data set, keeping all variables the same as the first iteration. In the third

iteration we added “FieldGoalsAttempted”, “X3PointShotsAttempted” and “FreeThrowsAttempted” on top of all the other variables to see which features were selected. For the Final experiment in the Home and Away data set we still used the “Target” data set however for the “Data” set we only had “Team”, “Home” and “Opponent” as our features to predict future outcomes just based on these attributes. No Hill-Climb method was required in this iteration however the 5-Fold method was included, and all variables and parameters were constant as the predeceasing iterations.

# 7 Results And Discusion

## 7.1 Data Exploration

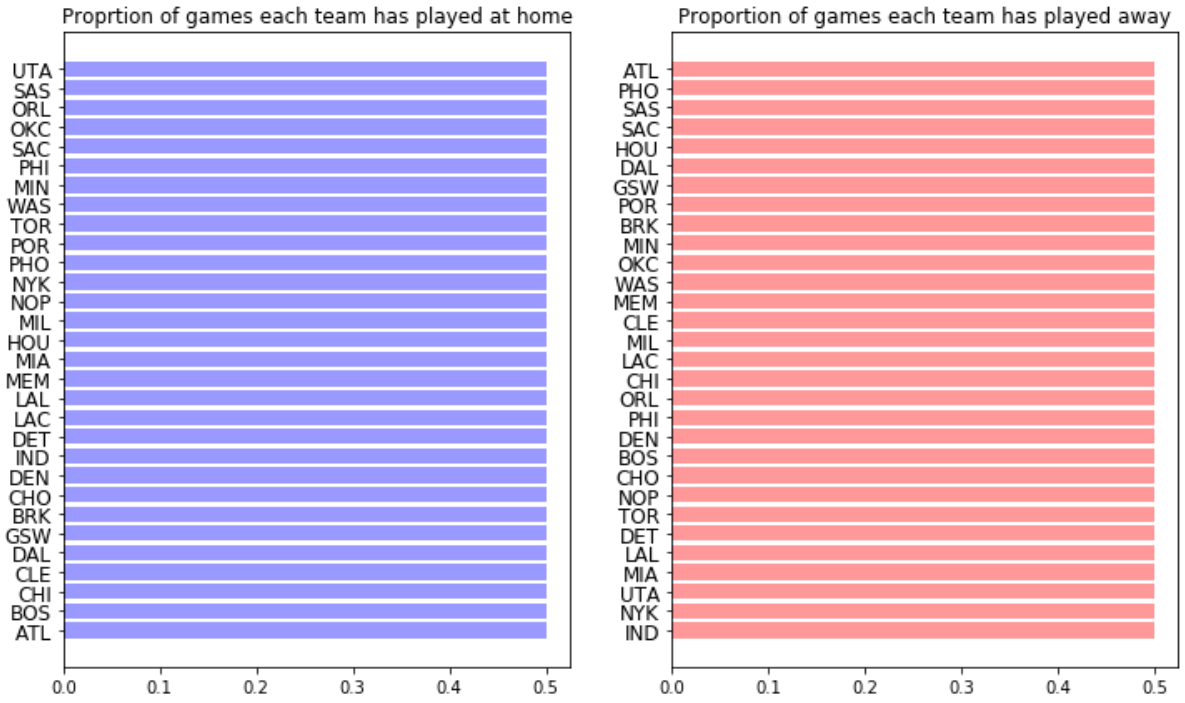



*Figure 6 Proportion of home and away matches each team has played in the NBA between 2014 to 2018*

**Figure 6** shows the proportion of home and away games that each team has played between 2014 – 2018. As expected, every team has played an equal proportion of home and away matches.

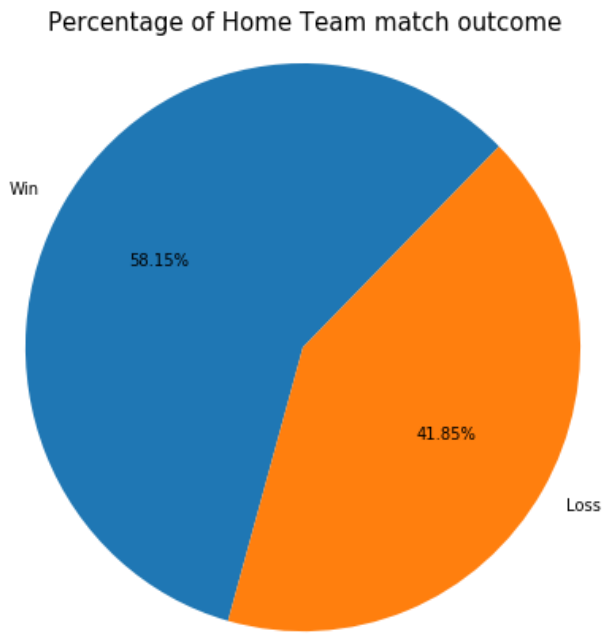



*Figure 7 Percentage of home team match outcome. Teams playing at home win 58.15% of the time.*

**Figure 7** shows the proportion of win loss rate for home teams. Home teams have a slight advantage of winning over away teams.


*Figure 8 Match frequency across each season. It is not expected that the number of games played daily affects the outcome of an individual match.*

**Figure 8** shows the frequency of matches played per day for each season in the NBA. As can be seen in the charts, the NBA takes a small break from playing in February each season. It is not expected that the date a match is played or the frequency of matches per day will affect win loss rate being analyse.

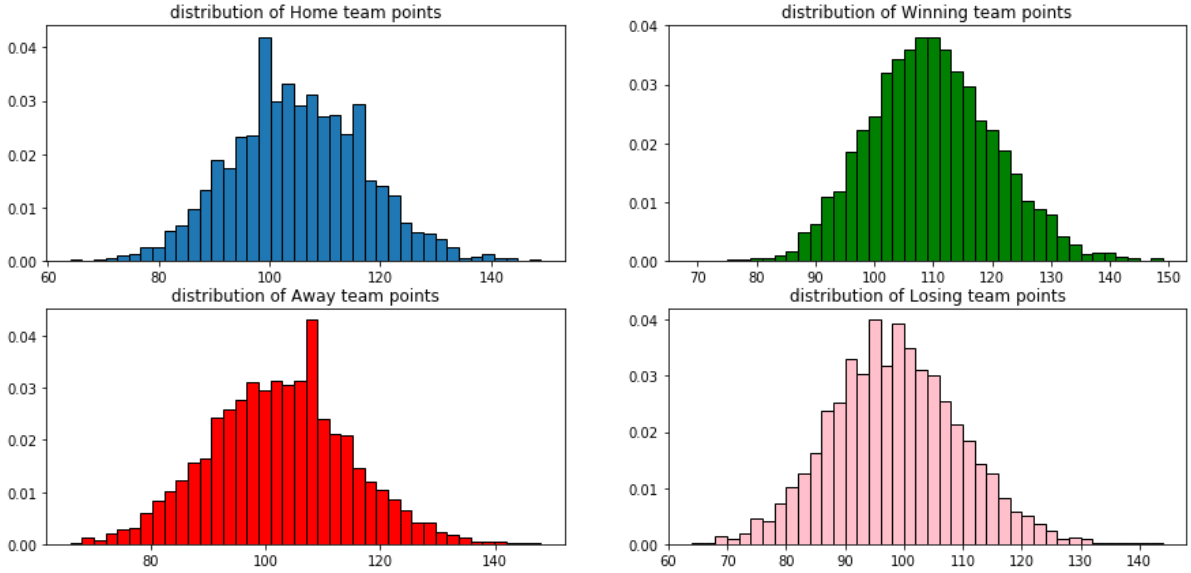



*Figure 9 Distribution of team points for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 9** shows the distribution of home, away, winning and losing team match points per match. As anticipated, the winning team has a slightly greater mean team point value than the losing teams.

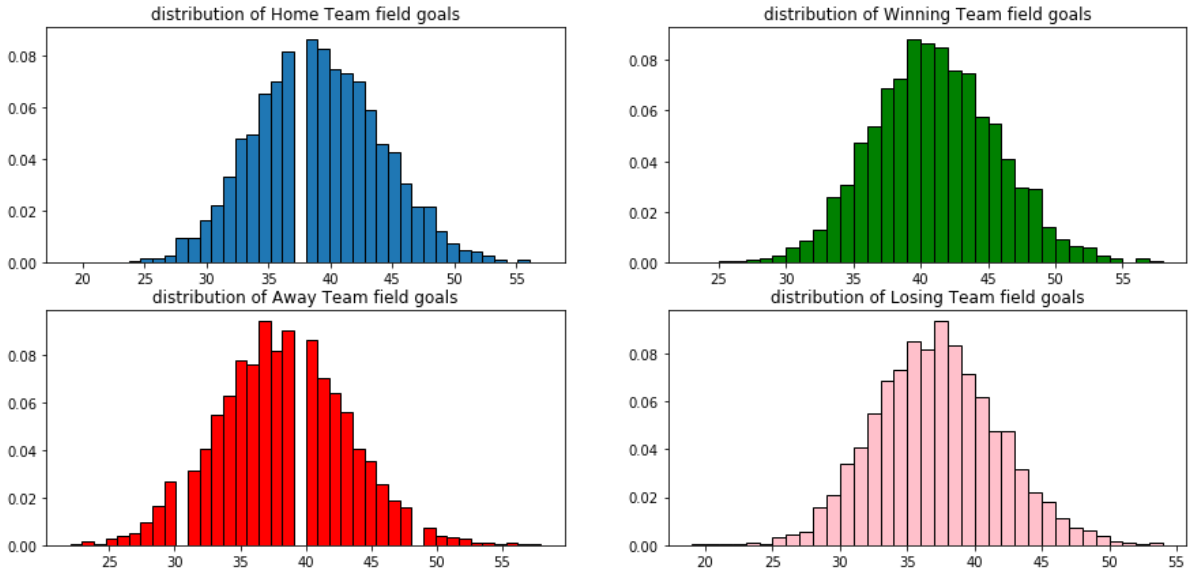



*Figure 10 Distribution of team field goals for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 10** shows the distribution of home, away, winning and losing team field goals per match. Congruent with team points, winning team's mean field goals is greater than losing teams.

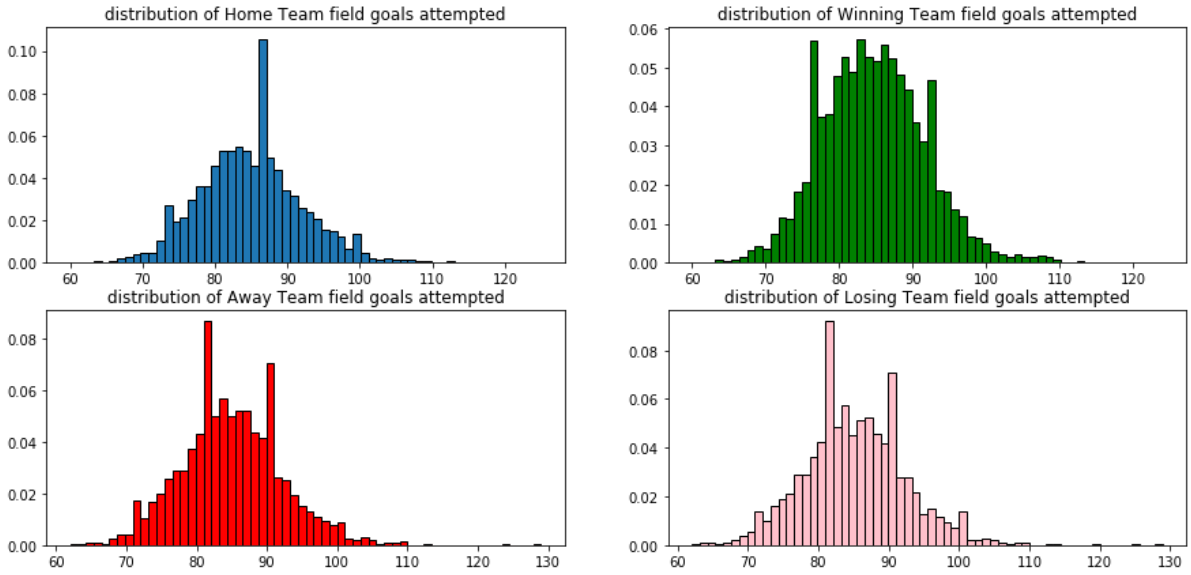



*Figure 11 Distribution of team field goals attempted for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 11** shows the distributions of team field goals attempted per match. The amplitude in the winning team distribution does not contain extremities like the other distributions.

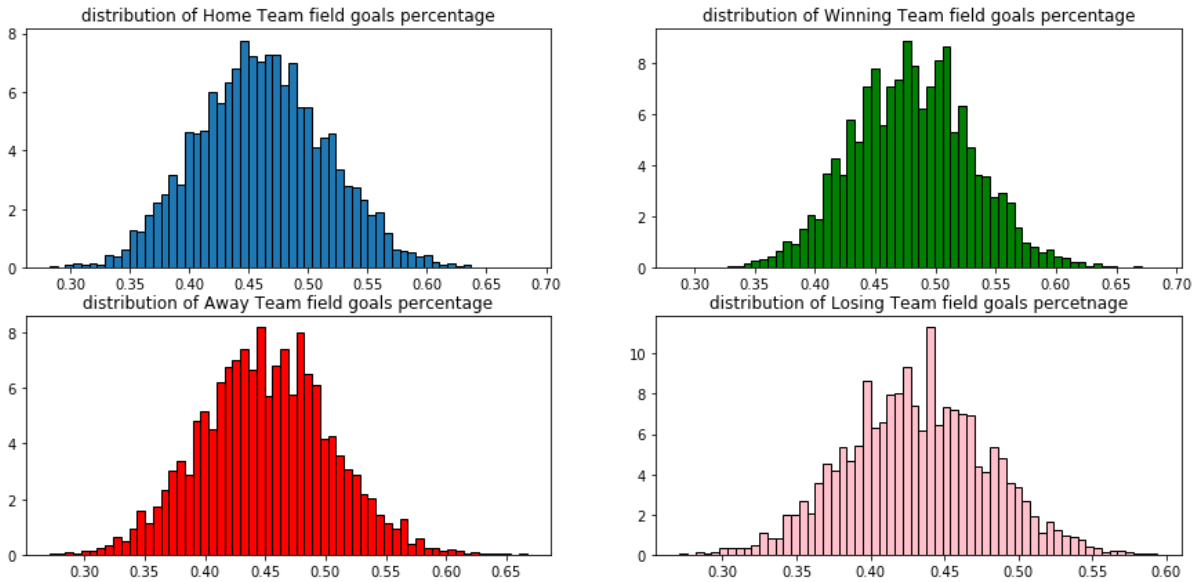



*Figure 12 Distribution of team field goals percentage for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 12** shows the distributions field goal percentage. The percentage indicates the proportion of successful field goal attempts. The losing team distribution shows greater amplitude in some cases. Maybe in these situations the winning team achieved more three-point shots.

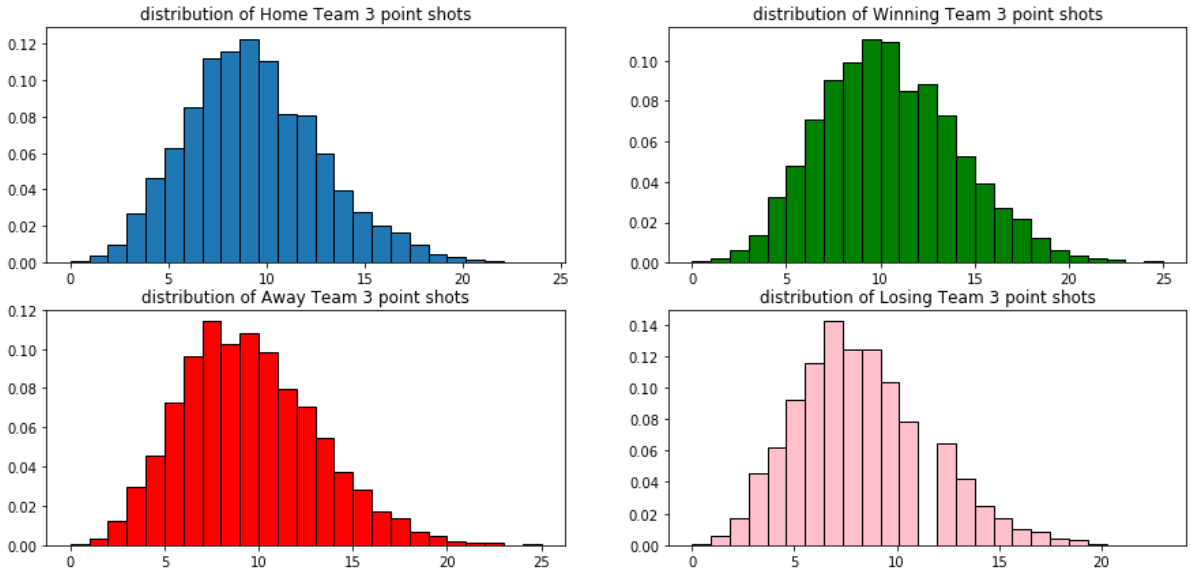



*Figure 13 Distribution of team three-point shots for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 13** shows the distributions of three-point shots made. All distributions were slightly skewed to the right; however, the winning and home team's distributions are less skewed compared to away and losing teams.

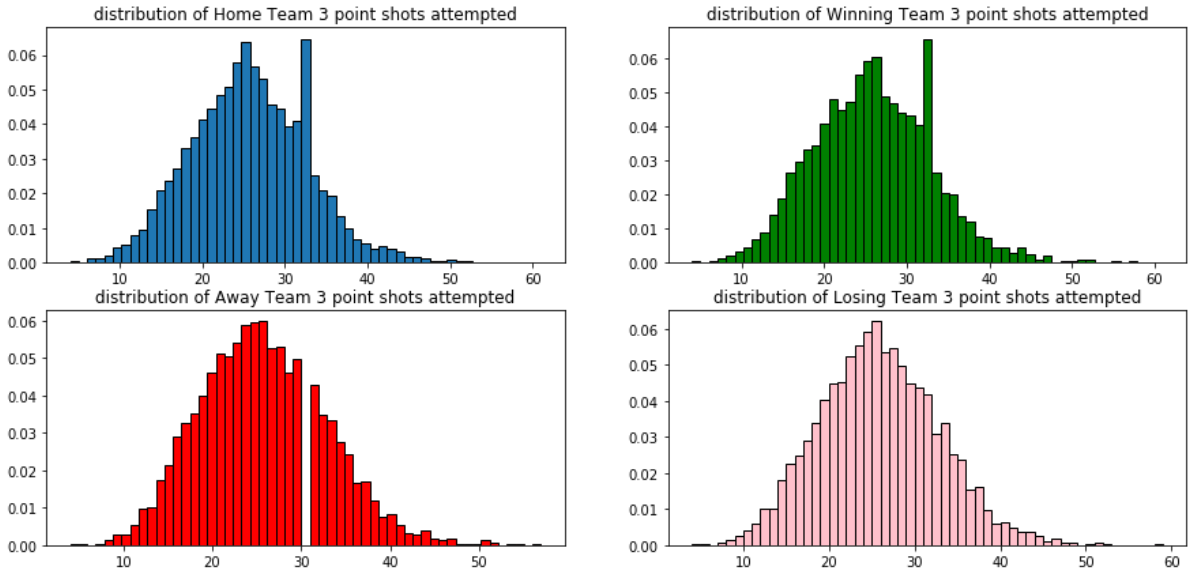



*Figure 14 Distribution of team three-point shots attempted for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 14** shows the distributions of three-point shots attempted. Home and winning team distributions were slightly skewed to the right with greater amplitude around thirty-three shots attempted.

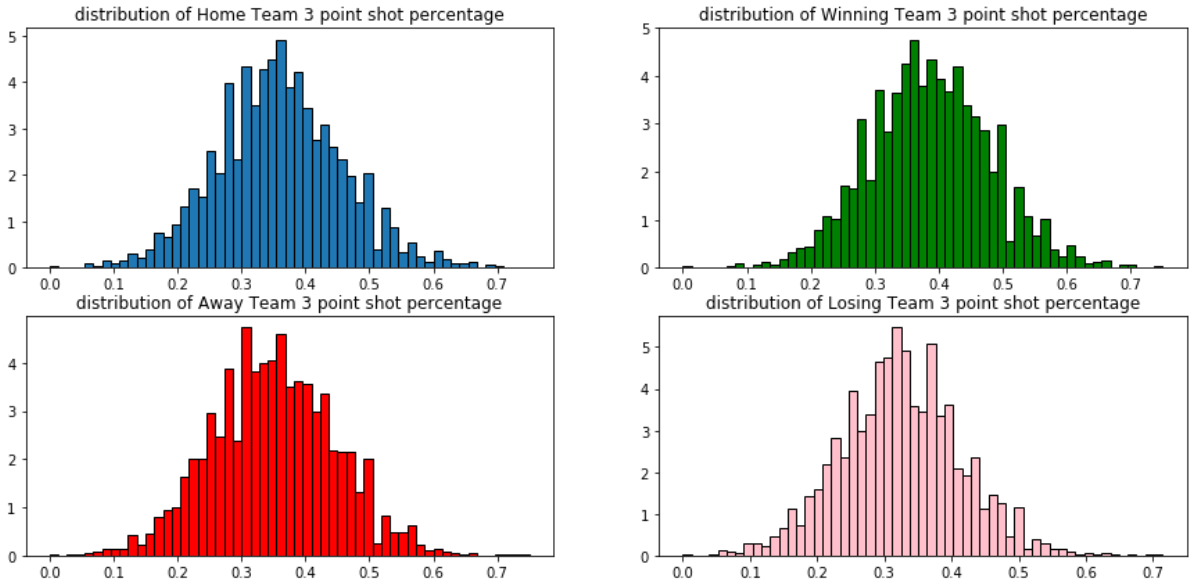



*Figure 15 Distribution of team three-point shots attempted for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 15** shows the distributions of three-point shot percentage. As anticipated, the mean value of winning team three-point shot percentage was greater than the losing teams.

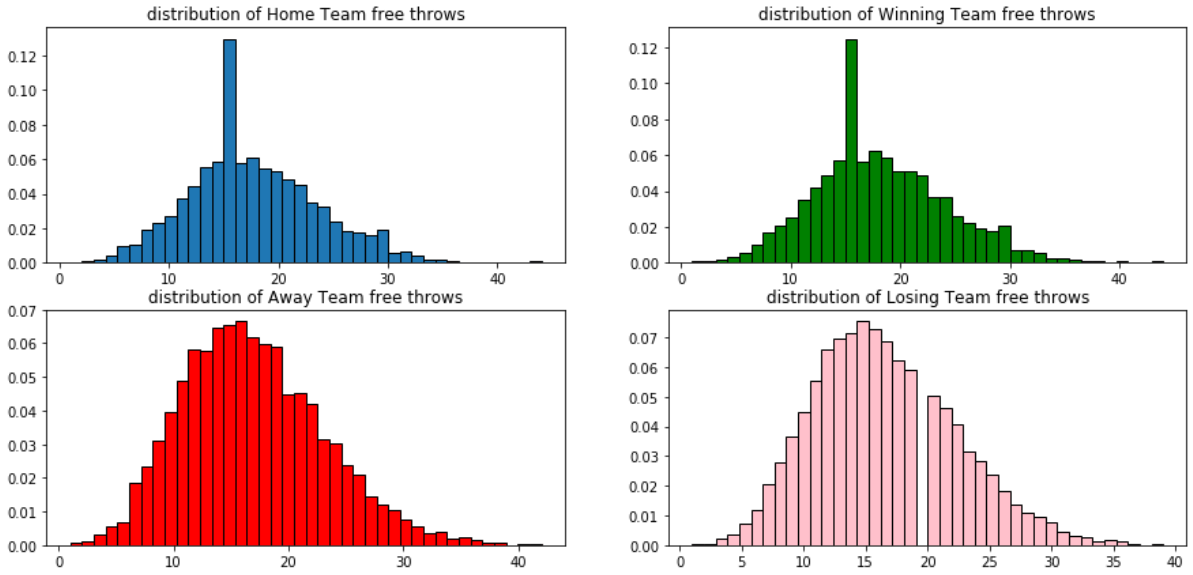



*Figure 16 Distribution of team free throws for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 16** shows the distribution of teams' free throw per match. The winning and home team distributions were both slightly skewed to the right, with an extreme amplitude being observed at around sixteen free throws. Both away team and losing team free throw distributions were slightly skewed to the right.

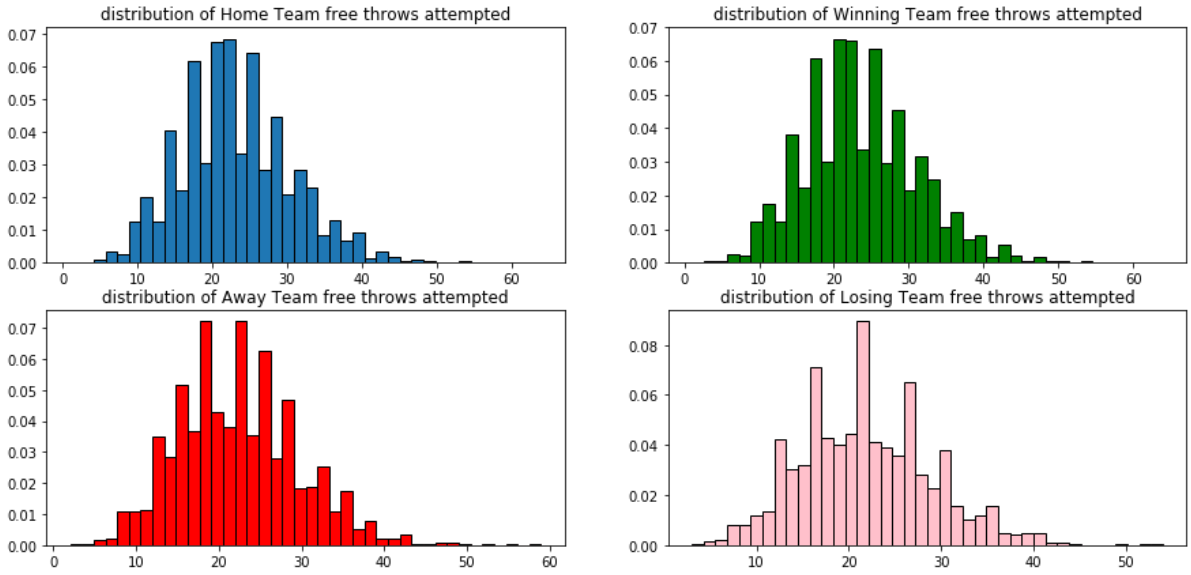



*Figure 17 Distribution of team free throws attempted for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 17** shows the distributions of team free throws attempted. All distributions were slightly skewed to the right. The losing team distribution shows less amplitude than the other distributions.

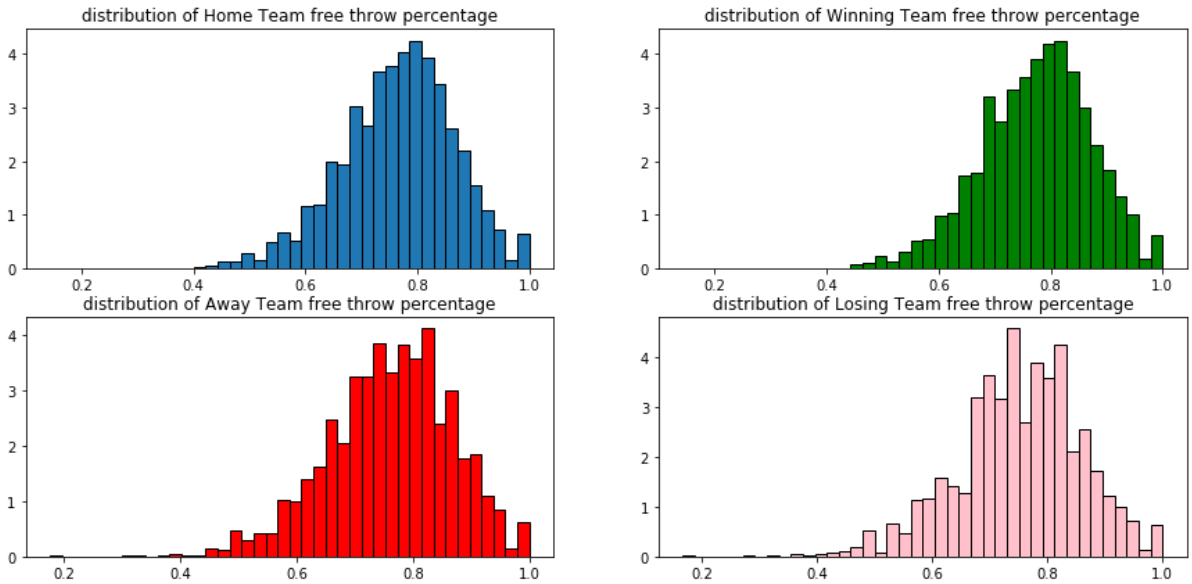



*Figure 18 Distribution of team free throws percentage for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 18** shows the distributions of free throw percentage. All distributions were skewed to the left and have a mean value of around 0.8.

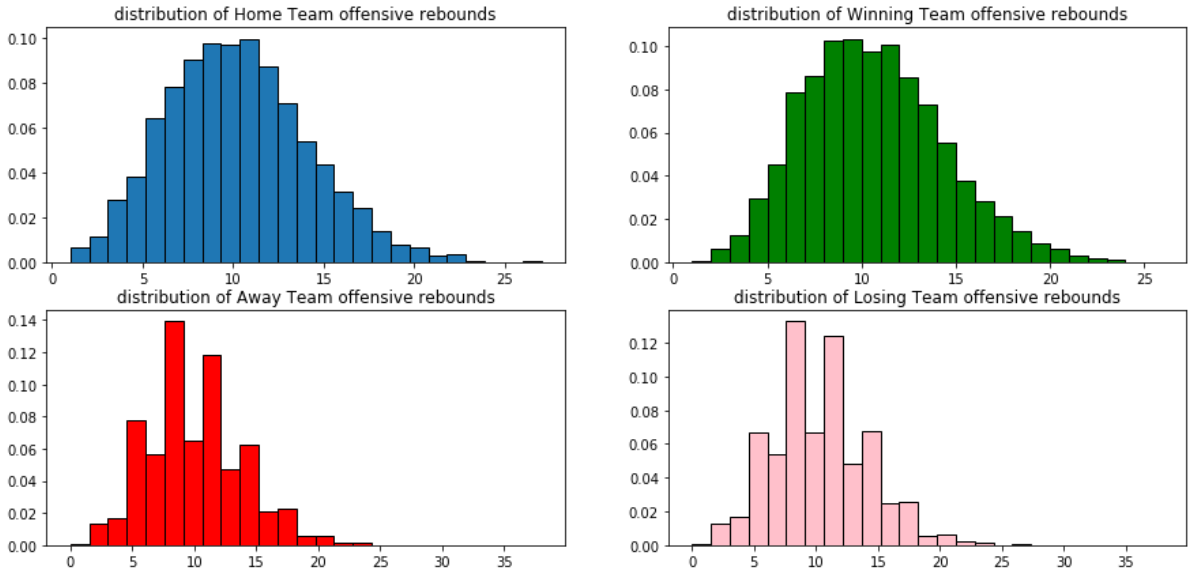



*Figure 19 Distribution of offensive rebounds for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive*

**Figure 19** shows the distributions of offensive rebounds. All distributions were slightly skewed to the right. Losing and away team distributions have less shape than home and winning team distributions.

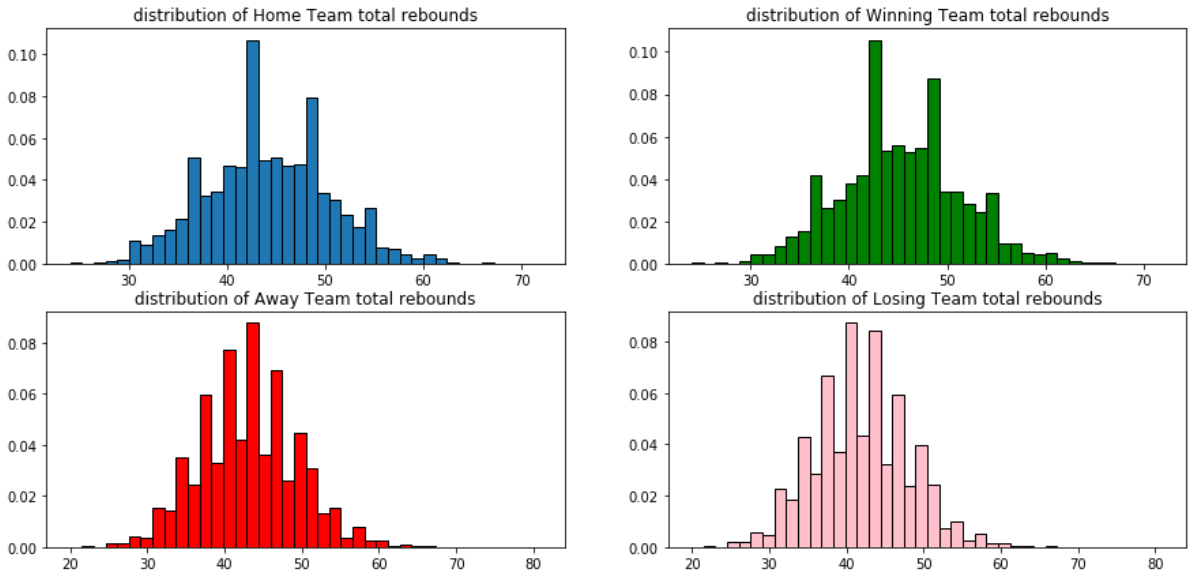



*Figure 20 Distribution of team total rebounds for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 10** shows the distributions of total rebounds. The winning team distribution has a greater mean value than the losing team distribution.

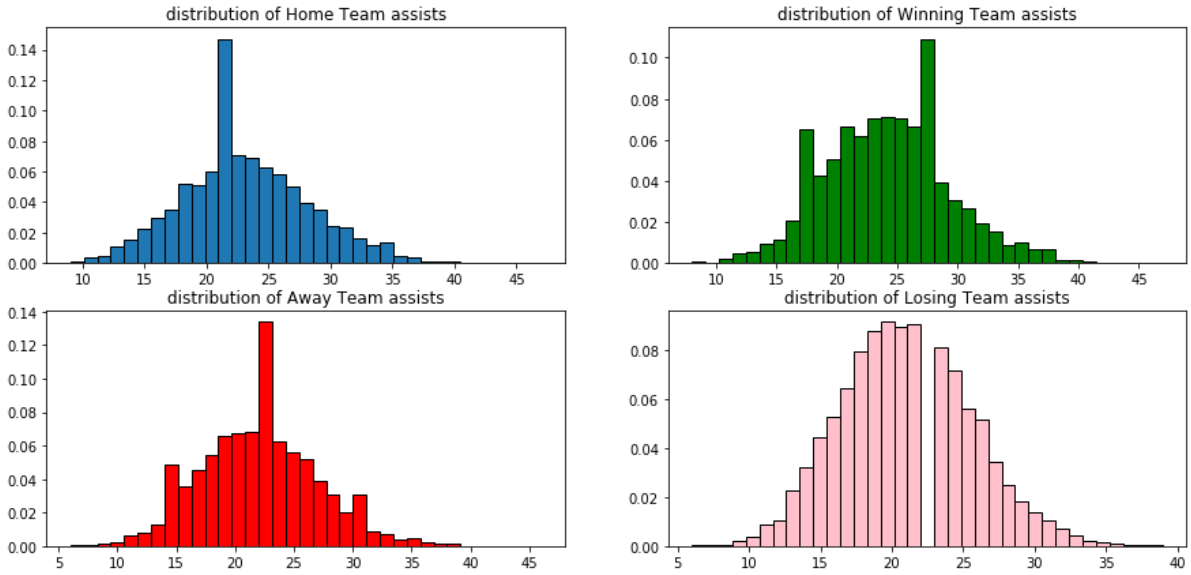



*Figure 21 Distribution of team assists for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive*

**Figure 21** shows the distributions of assists. The winning team distribution has a greater mean value than the losing team distribution.

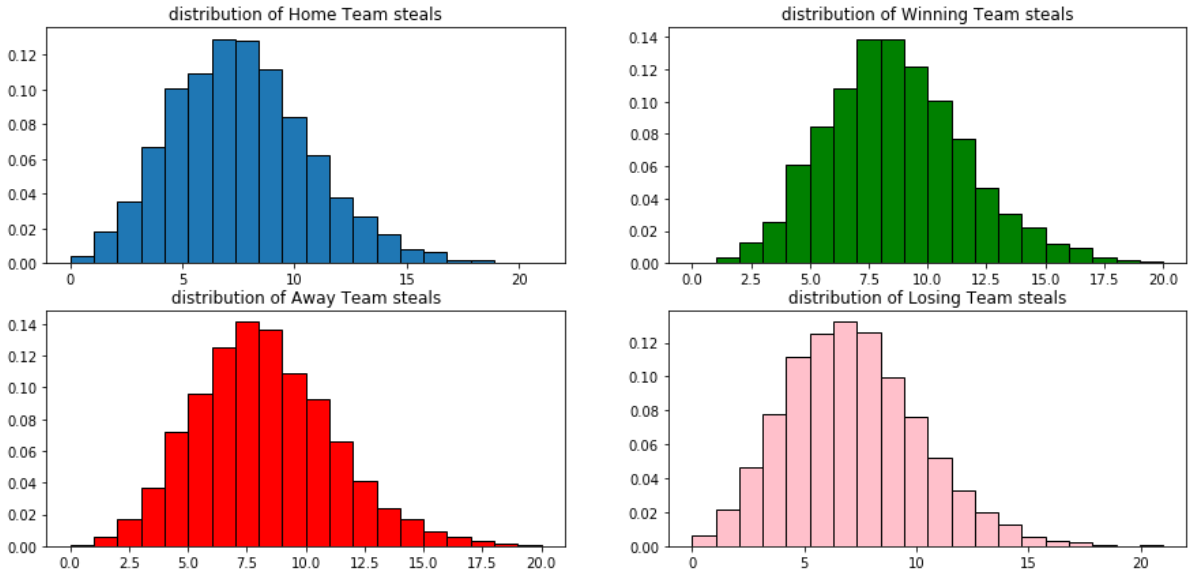



*Figure 22 Distribution of team steals for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 22** shows the distributions of steals. All distributions are slightly skewed to the left. The winning team distribution has a greater mean value than the losing team. Interestingly, the away team distribution has a greater mean value and amplitude than the home team distribution.

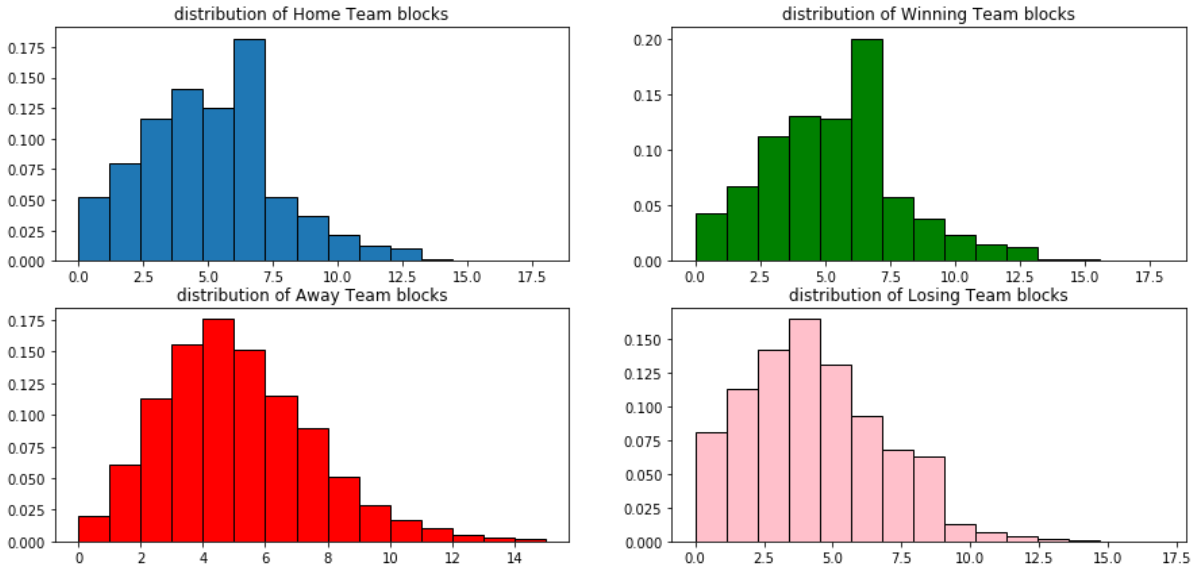



*Figure 23 Distribution of team steals for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive.*

**Figure 23** shows the distributions of blocks. All distributions were skewed to the right. Home and winning team distributions show inflated amplitude around six blocks.

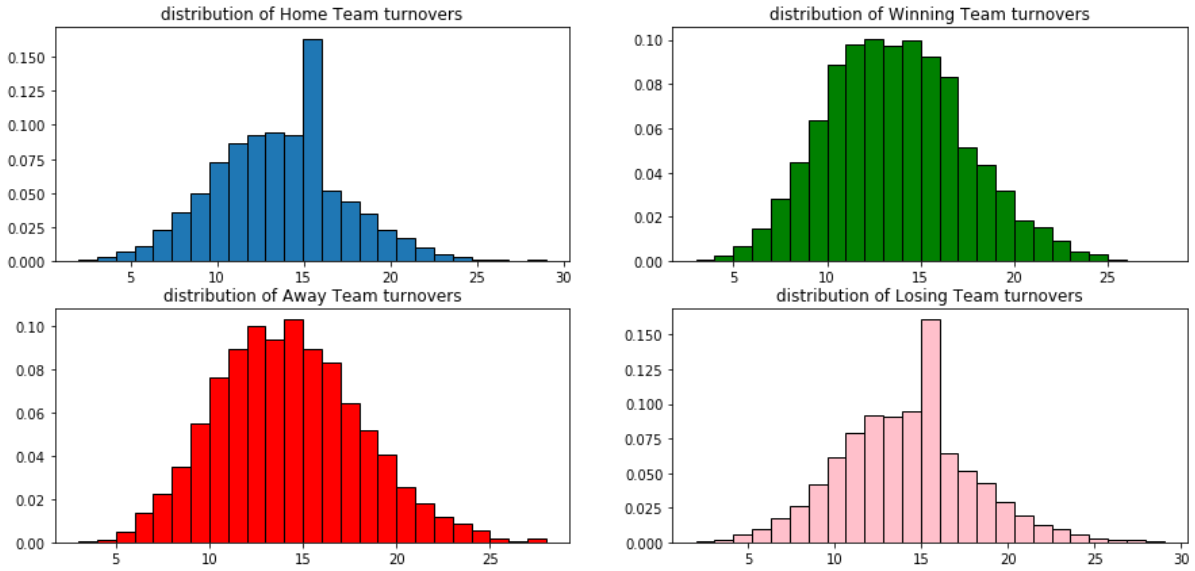



*Figure 24 Distribution of team steals for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive*

**Figure 24** shows the distributions of turnovers. The distributions of winning and away team were relatively normal. However, home and losing team distributions show inflated amplitude around sixteen turnovers.

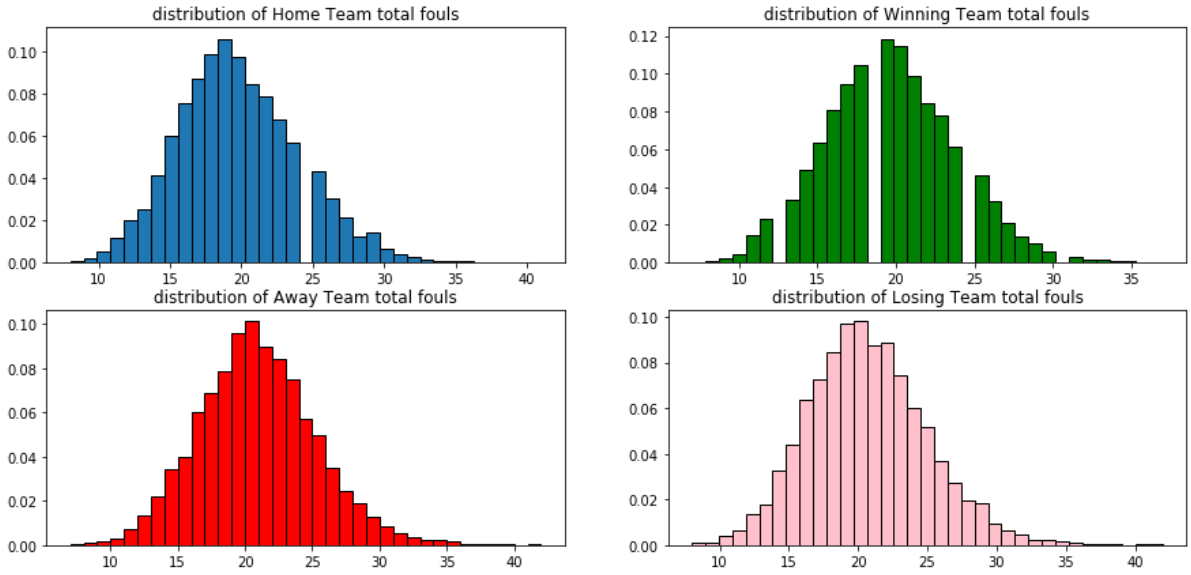



*Figure 25 Distribution of team total fouls for teams playing at home & away and winning & losing teams. These distributions are not mutually exclusive*

**Figure 25** shows the distributions of total team fouls. The home team distribution was slightly skewed to the left. Also, the home and winning team distributions have a marginally smaller mean value of total fouls, at around 18.

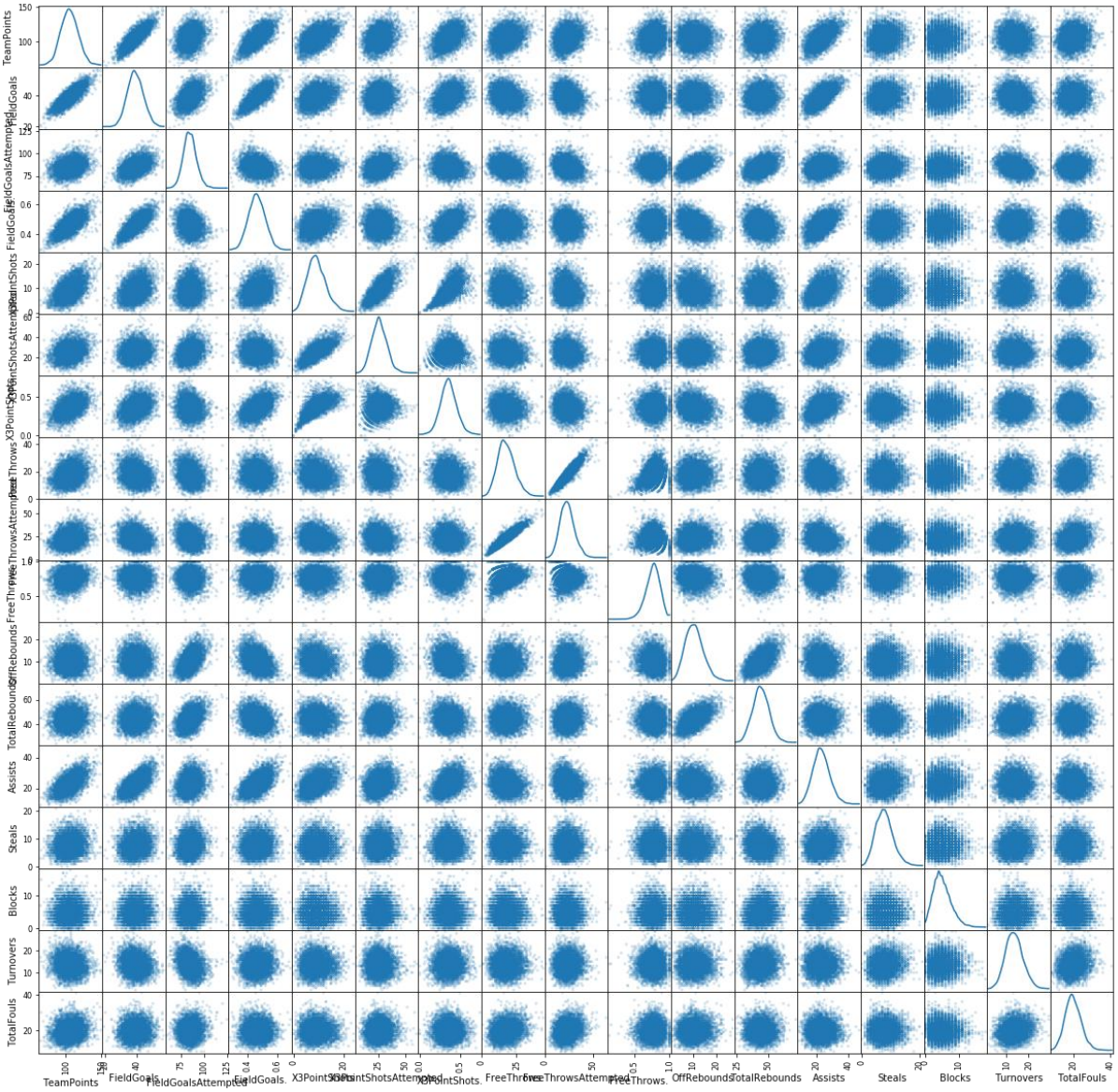



*Figure 26 A scatter matrix for all numerical columns in the dataset*

**Figure 26** shows a scatter plot matrix for all original numerical columns in the dataset. At a glance most variables did not display a correlation or direct relationship. Some relationships that show strong correlation were attempts by percentage, for example free-throws attempted by percentage of free-throws.

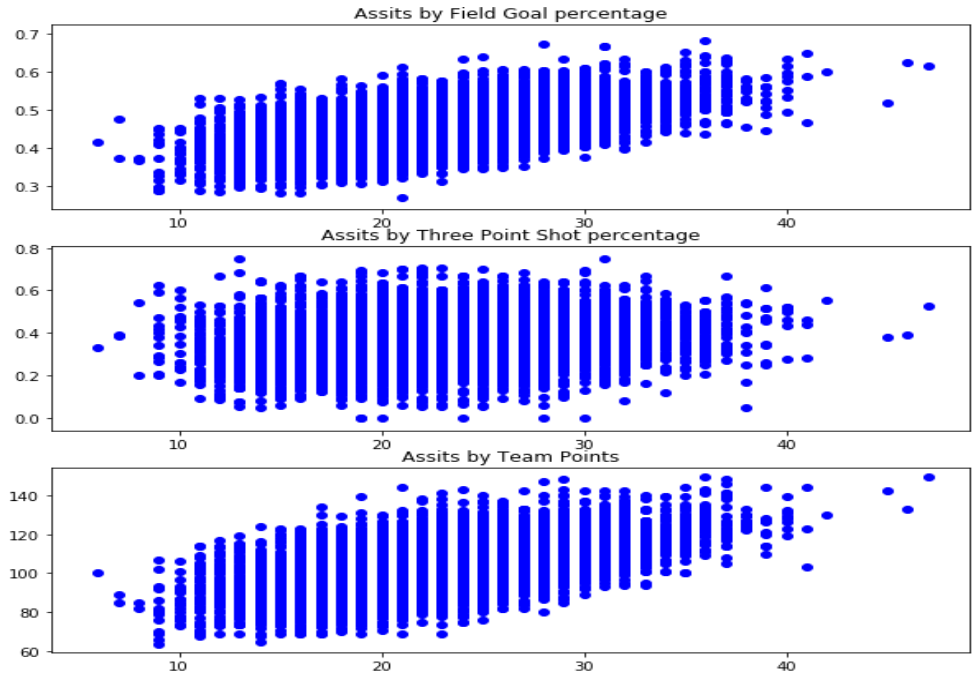



*Figure 27 Assists by Field Goals Percentage, Three Point Shot percentage and Team Points*

**Figure 27** shows the relationship between assists and field goals percentage; three point shot percentage and team points. Assists by field goal percentage and team points both showed a positive relationship, as assists increased team points and field goal percentage increased. However, the relationship between assists and three point shot percentage showed a relatively stable relationship.

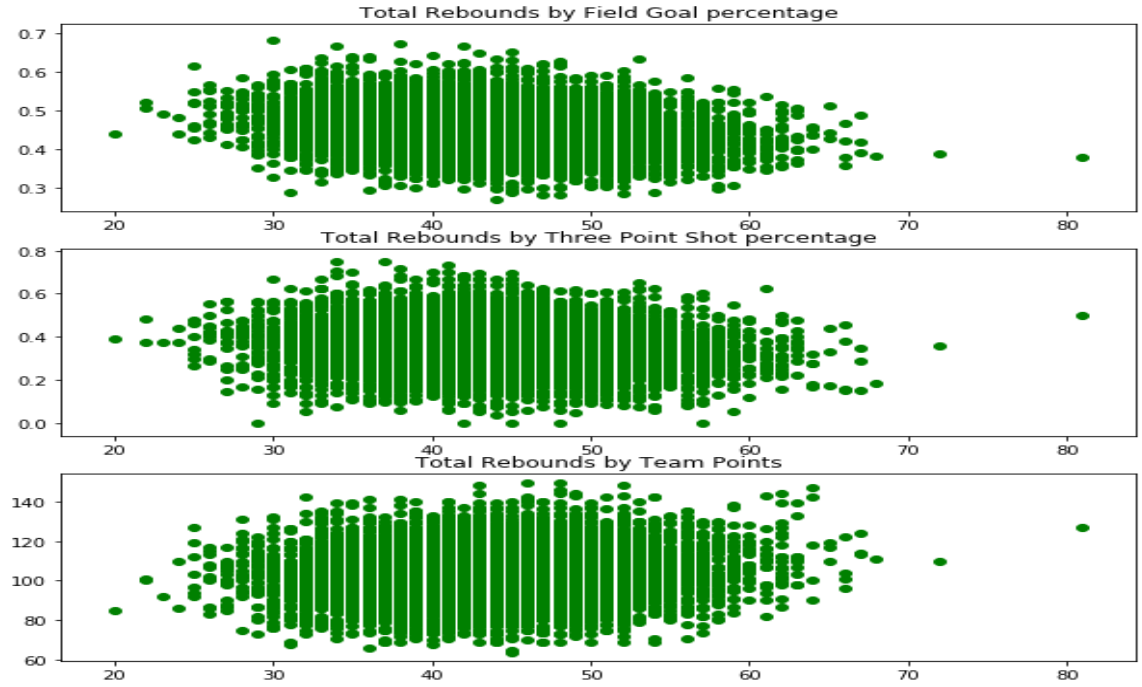



*Figure 28 Total Rebounds by Field Goals Percentage, Three Point Shot percentage and Team Points*

**Figure 28** shows the relationship between total rebounds and field goals percentage; three point shot percentage and team points. The relationship between total rebounds and field goal percentage and three point shot percentage showed a slight negative relationship. As total rebounds increased field

goal percentage and three point shot percentage decreased. Total rebounds by team points had a relatively stable relationship.

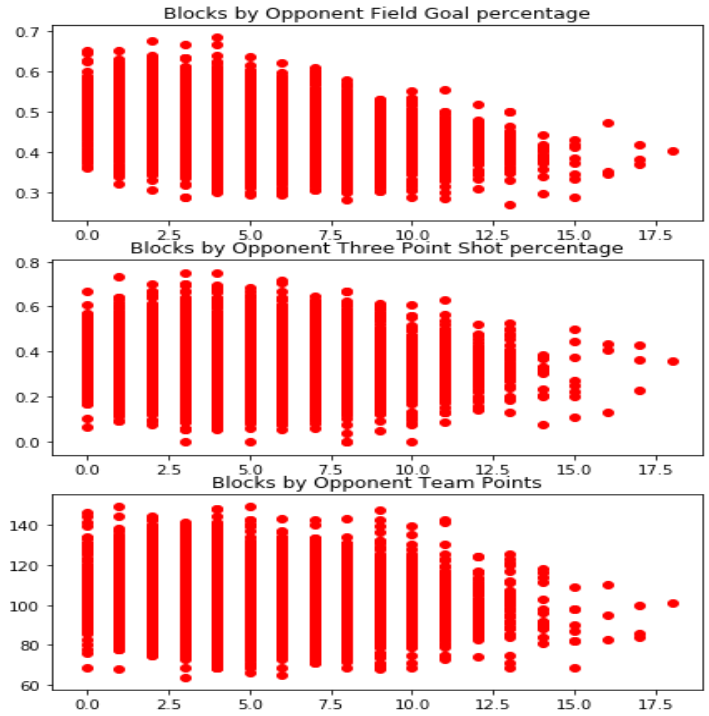



*Figure 29 Blocks by Opponent Field Goals Percentage, Three Point Shot percentage and Team Points*

**Figure 29** shows the relationship between blocks and opponent field goal percentage; three point shot percentage and team points. All trends showed a negative relationship. As blocks increase the opponents field goal percentage, three point shot percentage and team points decreased.

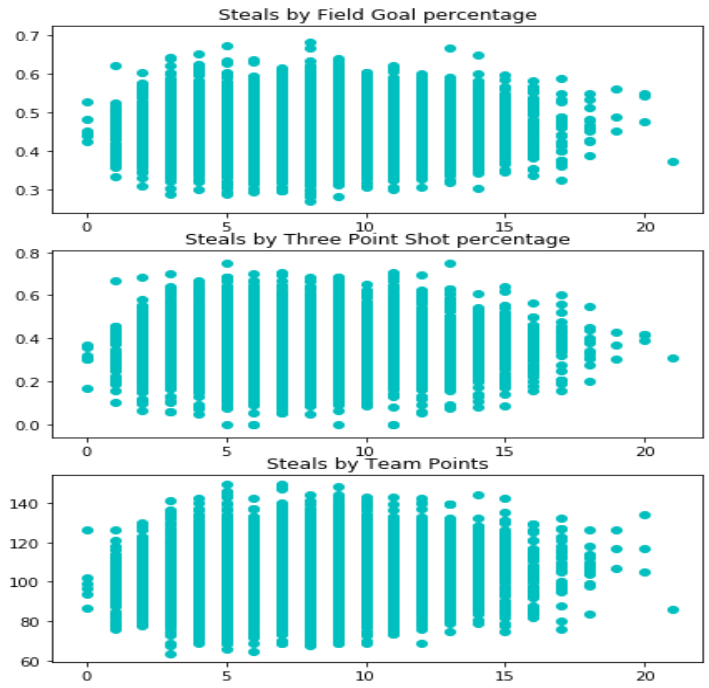



*Figure 30 Steals by Field Goals Percentage, Three Point Shot percentage and Team Points*

**Figure 30** shows the relationship between blocks and opponent field goal percentage; three point shot percentage and team points. No charts showed a positive or negative trend.

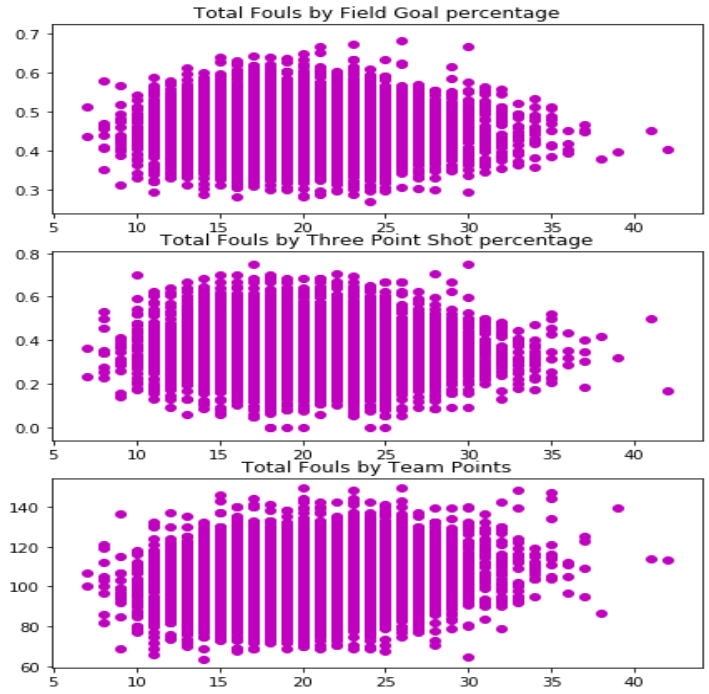



*Figure 31 Total Fouls by Field Goals Percentage, Three Point Shot percentage and Team Points*

**Figure 31** shows the relationship between total fouls and opponent field goal percentage; three point shot percentage and team points. No charts showed a positive or negative trend.

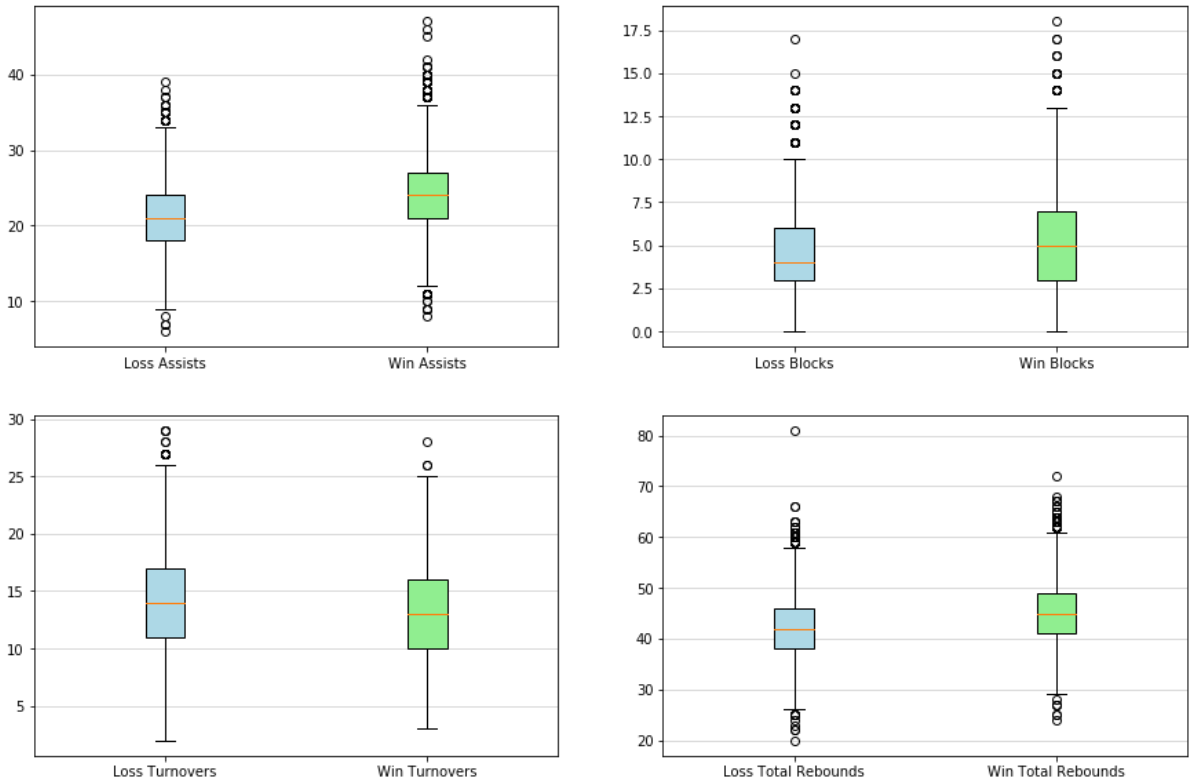


*Figure 32 Comparison assists, blocks, turnovers and total rebounds between winning and losing teams*

**Figure 32** shows the comparison of winning and losing statistics for assists, blocks, turnovers and total rebounds. In the assists quadrant the mean value for winning teams is greater than losing teams. In the blocks quadrant the interquartile range has a greater area than the losing teams interquartile range. In the turnovers quadrant losing teams mean value is slightly greater than the winning teams. In the total rebounds quadrant, the winning teams mean value is greater than the losing team

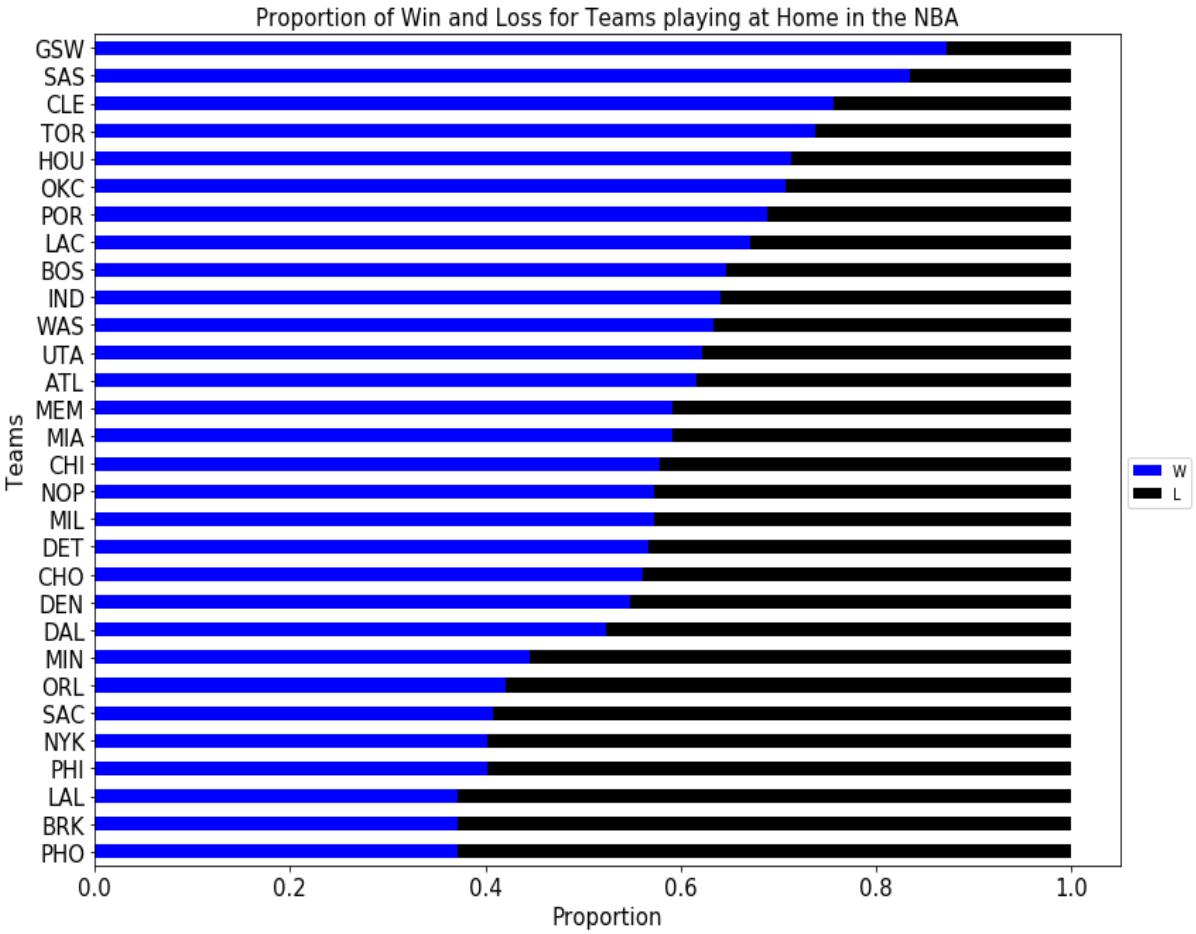



*Figure 33 The win and lose proportion of each teams track record while playing at home.*

**Figure 33** shows the proportion of each teams win and loss rate when playing at home between the 2014 to 2018 seasons. Golden State Warriors (GSW) have the highest proportion of wins when playing at home. Phoenix Suns (PHO) have the lowest proportion of wins when playing at home.

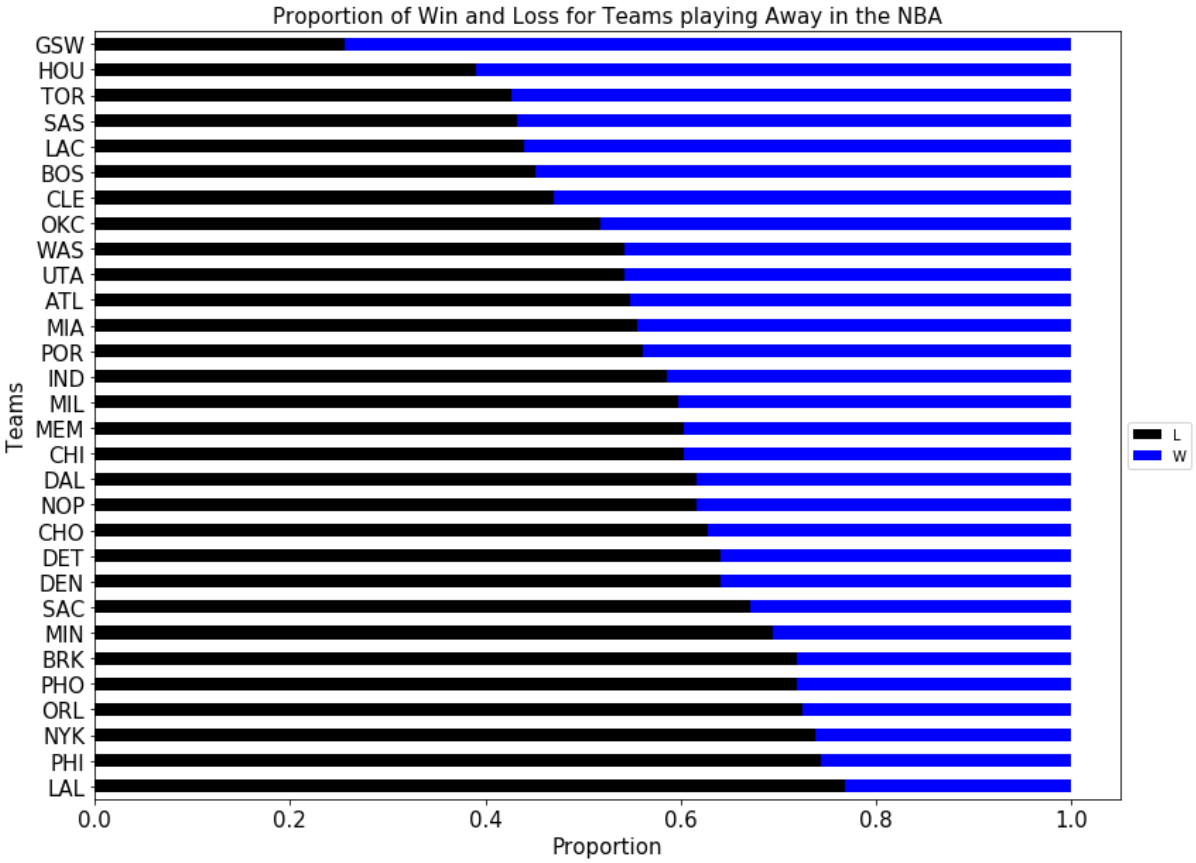



*Figure 34 The win and lose proportion of each teams track record while playing away.*

**Figure 34** shows the proportion of each teams win and loss rate when playing away between the 2014 to 2018 seasons. Congruent with playing at home, GSW have the highest proportion of wins when playing away. Los Angeles Lakers (LAL) have the lowest proportion of wins when playing away.

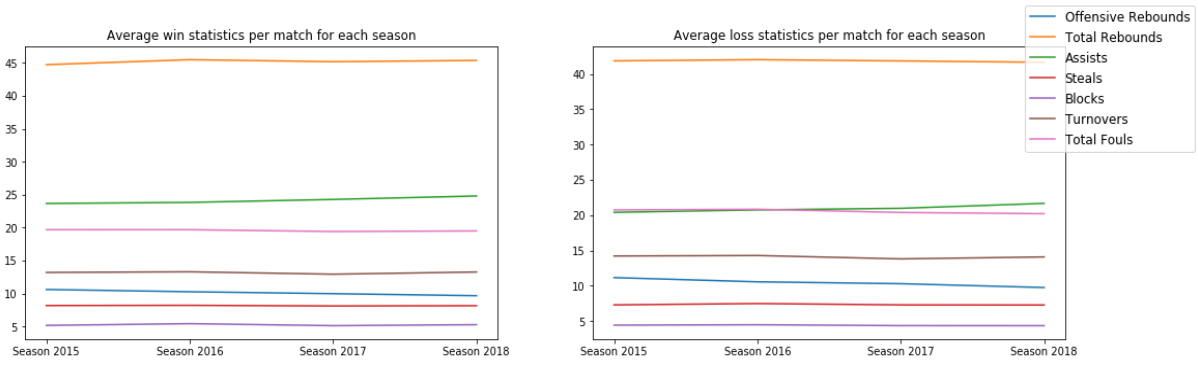



*Figure 35 The average rate of offensive rebounds, total rebounds, assists, steals, blocks, turnovers and total fouls per match across seasons 2014 to 2018 for winning and losing teams*

**Figure 35** shows the average rate of offensive rebounds, total rebounds, assists, steals, blocks, turnovers and total fouls per match across seasons 2014 to 2018 for winning and losing teams. The average rate of assists per match slightly increases across the 2017 to 2018 season for both winning

and losing teams. However, all other trends remain stable across all seasons for both winning and losing teams.

## 7.2 KNN – Model

This model attempted to address the question "what the driving factors behind are classifying a win or loss in the NBA between the 2014 to 2018 seasons?". By using the hill climbing technique, eight factors were determined to be more "explanatory" of match outcome than the rest of the variables in the dataset analysed. They were *Assists, TotalRebounds, Blocks, Steals, Opp.Assists, Opp.TotalRebounds, Opp.Blocks* and *Opp.Steals.*

```
knn - 1 with 15 features [16, 4, 5, 7, 9, 11, 15, 10, 17, 14, 18, 1, 19, 3, 0]  predictive value: 0.7134146341463414
knn - 2 with 5 features [1, 4, 12, 0, 10]  predictive value: 0.6222415795586528
knn - 3 with 4 features [14, 1, 10, 0]  predictive value: 0.6437282229965157
knn - 4 with 6 features [9, 1, 0, 12, 11, 10]  predictive value: 0.7212543554006968
knn - 5 with 4 features [10, 4, 0, 1]  predictive value: 0.7543554006968641
knn - 6 with 3 features [10, 6, 0]  predictive value: 0.6980255516840883
knn - 7 with 14 features [18, 15, 9, 3, 7, 16, 14, 5, 4, 10, 19, 0, 8, 13]  predictive value: 0.7839721254355401
knn - 8 with 5 features [4, 12, 1, 10, 0]  predictive value: 0.7409988385598142
knn - 9 with 16 features [4, 3, 1, 14, 17, 15, 9, 13, 5, 11, 10, 7, 0, 8, 18, 6]  predictive value: 0.7833914053426249
knn - 10 with 3 features [10, 4, 0]  predictive value: 0.7694541231126597
```

*Figure 36 Model iterations that increase the value of k by one with each loop, the index value of features within the model and the predictive power of each iteration*

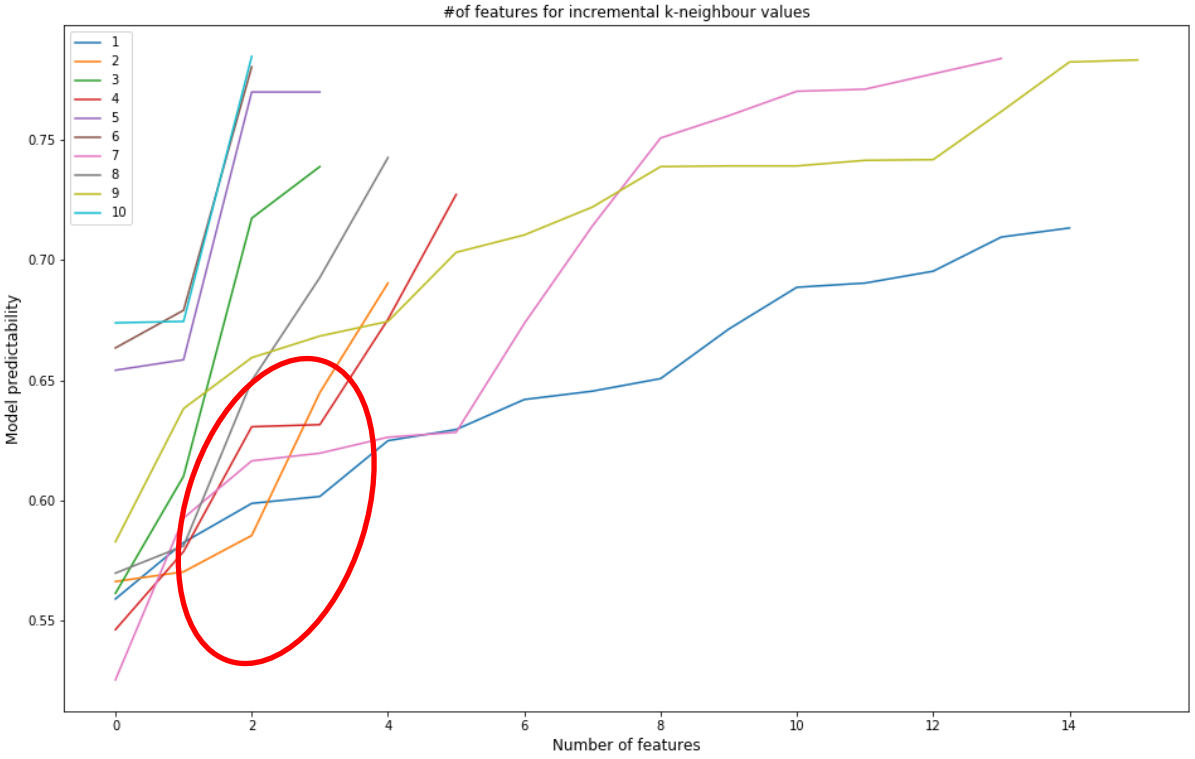



*Figure 37 Predictability output for each model produced in the initial hill climbing technique. Red circle outlines the trends with high predictability but few features.*

By using the K-nearest neighbours algorithm, with the size of a neighbourhood equal to ten and implementing Euclidian distance as the distance metric, the model was able to correctly determine the win, loss outcome of a match on a test data set of size 984 to a rate of around 76.5% correct classifications.

```
knn - 1 with 10 features [1, 11, 0, 9, 8, 10, 3, 4, 5, 2]  predictive value: 0.6991869918699187
knn - 2 with 13 features [4, 6, 3, 8, 9, 1, 10, 13, 7, 12, 5, 2, 0]  predictive value: 0.7113821138211383
knn - 3 with 9 features [13, 12, 6, 2, 9, 11, 1, 5, 8]  predictive value: 0.7299651567944251
knn - 4 with 13 features [10, 1, 2, 13, 0, 4, 6, 12, 8, 9, 3, 11, 5]  predictive value: 0.7491289198606271
knn - 5 with 10 features [11, 13, 3, 6, 2, 10, 8, 1, 4, 9]  predictive value: 0.7412891986062717
knn - 6 with 13 features [2, 8, 6, 5, 13, 3, 9, 1, 7, 12, 4, 11, 0]  predictive value: 0.7807781649245064
knn - 7 with 11 features [9, 4, 2, 13, 5, 12, 8, 6, 7, 11, 1]  predictive value: 0.7587108013937283
knn - 8 with 13 features [2, 13, 3, 12, 8, 9, 4, 6, 10, 5, 7, 0, 1]  predictive value: 0.789198606271777
knn - 9 with 12 features [12, 10, 3, 1, 11, 0, 8, 2, 4, 7, 5, 9]  predictive value: 0.7796167247386759
knn - 10 with 10 features [11, 7, 8, 6, 4, 2, 3, 9, 10, 1]  predictive value: 0.7598722415795587
```

*Figure 38* Model iterations that increase the value of k by one with each loop, the index value of features within the model and the predictive power of each iteration

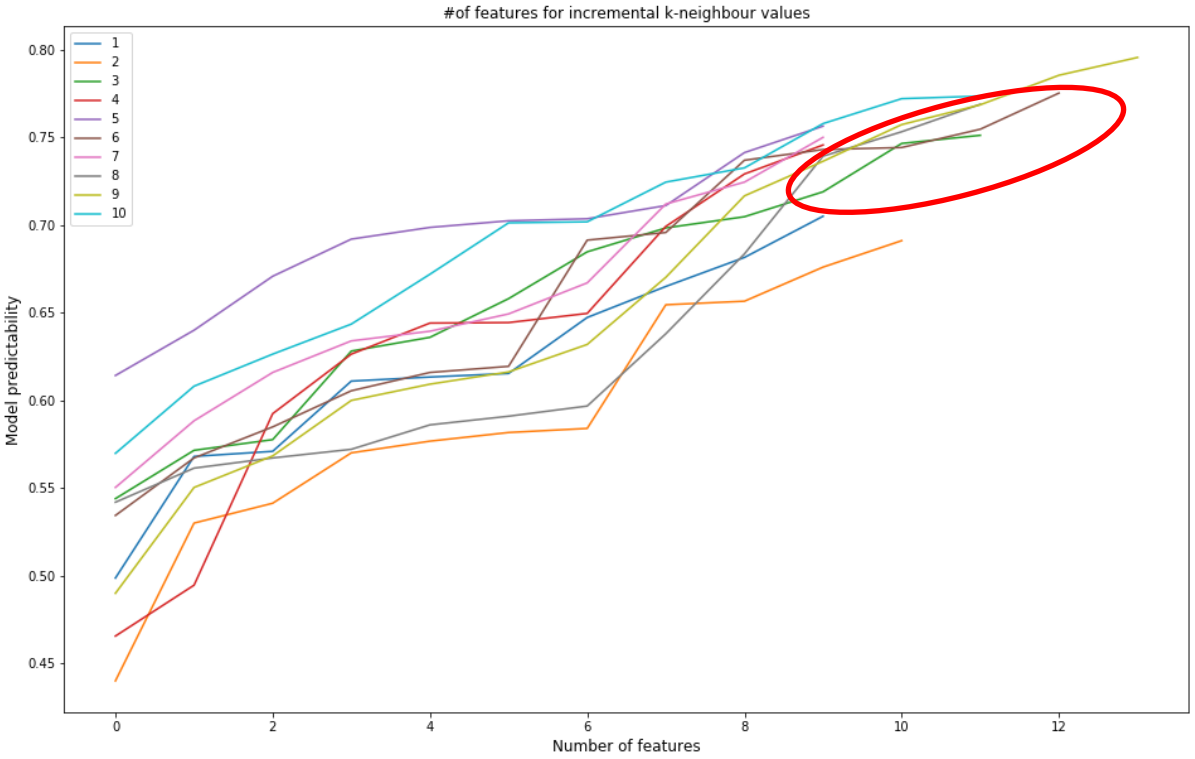



*Figure 39* Predictability output for each model produced in the initial hill climbing technique. Red circle outlines high predictability trends with large k-value.

| Index | Feature | Frequecy |
|---|---|---|
| 9 | Opp.Assists | 10 |
| 8 | Opp.TotalRebounds | 10 |
| 3 | Steals | 10 |
| 2 | Assists | 10 |
| 1 | TotalRebounds | 10 |
| 11 | Opp.Blocks | 8 |
| 10 | Opp.Steals | 8 |
| 4 | Blocks | 8 |
| 0 | OffRebounds | 8 |
| 13 | Opp.TotalFouls | 7 |
| 12 | Opp.Turnovers | 7 |
| 6 | TotalFouls | 7 |
| 5 | Turnovers | 6 |
| 7 | Opp.OffRebounds | 5 |

*Figure 40 The most frequent features that appeared in the hill climbing technique. Features highlighted green were chosen for the final model*

```
fold 1 score: 0.7723577235772358
             precision    recall  f1-score   support

          L       0.76      0.67      0.71       409
          W       0.78      0.85      0.81       575

avg / total       0.77      0.77      0.77       984

fold 2 score: 0.7632113821138211
             precision    recall  f1-score   support

          L       0.74      0.69      0.71       420
          W       0.78      0.82      0.80       564

avg / total       0.76      0.76      0.76       984

fold 3 score: 0.7916666666666666
             precision    recall  f1-score   support

          L       0.80      0.68      0.73       417
          W       0.79      0.87      0.83       567

avg / total       0.79      0.79      0.79       984

fold 4 score: 0.774390243902439
             precision    recall  f1-score   support

          L       0.80      0.65      0.71       427
          W       0.76      0.87      0.81       557

avg / total       0.78      0.77      0.77       984

fold 5 score: 0.7896341463414634
             precision    recall  f1-score   support

          L       0.76      0.68      0.72       386
          W       0.81      0.86      0.83       598

avg / total       0.79      0.79      0.79       984
```

*Figure 41 Cross-Validation for K-neighbour values for the second iteration.*

Cross validation across five folds of data produced precision ratings between 0.76 – 0.79, meaning the model correctly labelled samples 76% to 79% of the time. And, the f1-score, which is a weighted average between precision and recall, between 076 – 0.79, or 76% to 79% of the time.

```
[[372 151]
 [ 95 612]]
             precision    recall  f1-score   support

          L       0.80      0.71      0.75       523
          W       0.80      0.87      0.83       707

avg / total       0.80      0.80      0.80      1230
```

*Figure 42 Final K-neighbour values with a confusion matrix*

In **Figure 41** is a confusion matrix and table containing metrics for the knn model. This shows that for home teams, the model correctly predicts a loss with 372 observations and incorrectly predicts a loss as a win with 95 samples. It also incorrectly predicts a win as a loss with 151 observations and correctly predicts a win with 612 observations

The drawbacks from this model are that only observations from the perspective of the *home* team were included in the model. This limits model findings to be applicable only when analysing team performance at a home stadium. Secondly, the dataset only includes aggregate match statistics of a *team's* performance, statistics of *individual players* are not recorded. This limits the analysis to the assumption that the players in a team are stagnant and do not change. In reality this is not the case. Players are constantly traded from team to team between and even during seasons. Finally, the model assumes that a particular playing 'style' is superior in that in a team that can achieve more assists, blocks, rebounds and steals, while minimising their opponents assists, blocks, rebounds and steals. Where this might not be the case when playing against a team with an alternative playing 'style'.

## 7.3 Decision Trees

As with the KNN model, the Decision Tree method was administrated to contrast and compare certain features to address the question "what the driving factors behind are classifying a win or loss in the NBA between the 2014 to 2018 seasons?". By using the same original factors as the KNN model the Hill-Climbing method determined which features where more "explanatory" of match outcome than the rest of the variables in the dataset analysed.

### 7.3.1 First Iteration of Experiments

```
Decision Tree with  1 [8]  selected features: 0.6070460704607046
Decision Tree with  2 [8, 6]  selected features: 0.6253387533875339
Decision Tree with  3 [8, 6, 4]  selected features: 0.6429539295392954
Decision Tree with  4 [8, 6, 4, 11]  selected features: 0.6443089430894309
Decision Tree with  5 [8, 6, 4, 11, 2]  selected features: 0.6585365853658537
Decision Tree with  6 [8, 6, 4, 11, 2]  selected features: 0.6551490514905149
Decision Tree with  7 [8, 6, 4, 11, 2, 9]  selected features: 0.6944444444444444
Decision Tree with  8 [8, 6, 4, 11, 2, 9, 1]  selected features: 0.7086720867208672
Decision Tree with  9 [8, 6, 4, 11, 2, 9, 1, 7]  selected features: 0.7086720867208672
Decision Tree with  10 [8, 6, 4, 11, 2, 9, 1, 7, 10]  selected features: 0.7086720867208672
Decision Tree with  11 [8, 6, 4, 11, 2, 9, 1, 7, 10]  selected features: 0.7073170731707317
Decision Tree with  12 [8, 6, 4, 11, 2, 9, 1, 7, 10, 0]  selected features: 0.7086720867208672
Decision Tree with  13 [8, 6, 4, 11, 2, 9, 1, 7, 10, 0, 5]  selected features: 0.7086720867208672
Decision Tree with  14 [8, 6, 4, 11, 2, 9, 1, 7, 10, 0, 5, 12]  selected features: 0.7086720867208672
```

*Figure 43 Number of features for incremental Decision Tree values using Hill-Climbing Method first iteration*

The features that were select was the 8th iteration seen in **Figure 42.** This was attributed to knowing decision trees tend to over fit the data resulting in large cross-validation errors stated in the literature review.

'OffRebounds', 'TotalRebounds', 'Steals', 'Turnovers', 'Opp.OffRebounds', 'Opp.TotalRebounds' and 'Opp.Steals', were determined from the Hill-Climbing method to be the best features for predicting the outcome of a win or loss. Highly suggesting that the amount of ball possession has a major factor in the overall outcome of the game.

```
[[322 208]
 [223 477]]
             precision    recall  f1-score   support

          L       0.59      0.61      0.60       530
          W       0.70      0.68      0.69       700

avg / total       0.65      0.65      0.65      1230
```

*Figure 44 Final Decision Tree values with a confusion matrix first iteration.*

The confusion matrix shown in **Figure 44**, depicts that the accuracy of being able to predict a team winning is a momentous gap then the accuracy of being able to predict a team losing the game. 59% was the fraction of correctly predicted out comes for a team losing compared to the 70% of accurately predicted teams winning. Which averages out to being 65%, when we compare this number against the 5-Fold method shown in **Figure 45,** the 5-fold produces a predictability of 65-69% giving a 4% of cross-validation error.

```
fold 1 score: 0.6778455284552846
             precision    recall  f1-score   support

          L       0.64      0.59      0.61       426
          W       0.70      0.75      0.72       558

avg / total       0.68      0.68      0.68       984

fold 2 score: 0.6829268292682927
             precision    recall  f1-score   support

          L       0.61      0.60      0.61       398
          W       0.73      0.74      0.73       586

avg / total       0.68      0.68      0.68       984

fold 3 score: 0.6565040650406504
             precision    recall  f1-score   support

          L       0.62      0.49      0.55       417
          W       0.68      0.78      0.72       567

avg / total       0.65      0.66      0.65       984

fold 4 score: 0.6900406504065041
             precision    recall  f1-score   support

          L       0.62      0.63      0.63       407
          W       0.74      0.73      0.74       577

avg / total       0.69      0.69      0.69       984

fold 5 score: 0.665650406504065
             precision    recall  f1-score   support

          L       0.60      0.62      0.61       411
          W       0.72      0.70      0.71       573

avg / total       0.67      0.67      0.67       984
```

*Figure 45 5-Fold Predictability First Iteration*

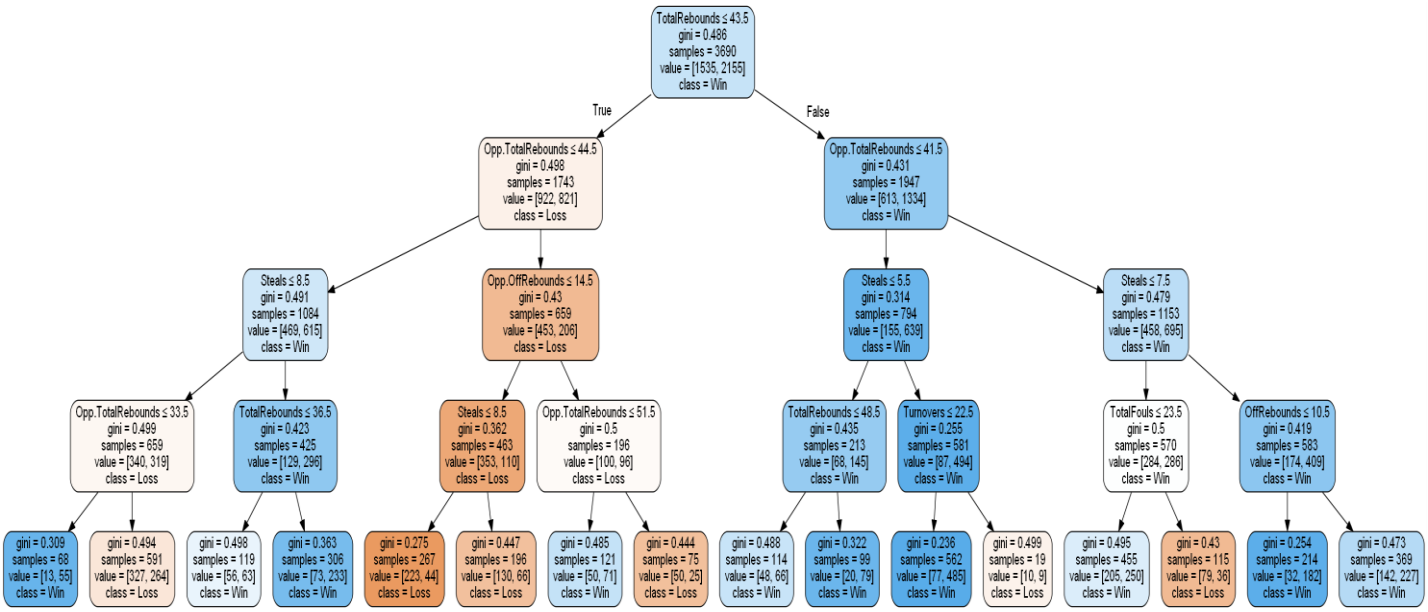



*Figure 46 Visualisation of the Decision Tree for the First Iteration*

From **Figure 46** we can see that TotalRebounds is the root node followed by Opp.TotalRebounds, furthering the indication that possession is a key indicator for determining the overall outcome of an NBA game. This is most likely because at the profession level, the players and team have high shot percentage so the more possession designates the amount of shots taken by the team and theoretically less by the opposition.

### 7.3.2 Second Iteration of Experiments

The second iteration of the experiment, the features had an additional 2 called Team and Opponent added to the original data frame used for the Decision Tree method. This was to see if the teams and opposition contributed to the outcomes of a basketball game. In theory we should to see that the teams make no or insignificant difference to the overall outcome. NBA teams have salary caps; however, some players take salary cuts to play in a winning team. Thus, the teams are unbalanced this was shown in the data set as the Golden State Warriors won the champion ship consecutive years in a row.

```
Decision Tree with  1 [1]  selected features: 0.6287262872628726
Decision Tree with  2 [1]  selected features: 0.6117886178861789
Decision Tree with  3 [1]  selected features: 0.6239837398373984
Decision Tree with  4 [1]  selected features: 0.6212737127371274
Decision Tree with  5 [1]  selected features: 0.6111111111111112
Decision Tree with  6 [1, 4]  selected features: 0.6314363143631436
Decision Tree with  7 [1, 4]  selected features: 0.6280487804878049
Decision Tree with  8 [1, 4]  selected features: 0.6294037940379403
Decision Tree with  9 [1, 4, 10]  selected features: 0.6565040650406504
Decision Tree with  10 [1, 4, 10, 7]  selected features: 0.6571815718157181
Decision Tree with  11 [1, 4, 10, 7, 15]  selected features: 0.6592140921409214
Decision Tree with  12 [1, 4, 10, 7, 15, 11]  selected features: 0.6951219512195121
Decision Tree with  13 [1, 4, 10, 7, 15, 11, 3]  selected features: 0.7066395663956639
Decision Tree with  14 [1, 4, 10, 7, 15, 11, 3, 0]  selected features: 0.7066395663956639
Decision Tree with  15 [1, 4, 10, 7, 15, 11, 3, 0]  selected features: 0.7052845528455285
Decision Tree with  16 [1, 4, 10, 7, 15, 11, 3, 0, 12]  selected features: 0.7066395663956639
```

*Figure 47 Number of features for incremental Decision Tree values using Hill-Climbing Method second iteration*

The features that were select was the 13th iteration seen in **Figure 47**, to once again limit the amount of over fitting in the model.

*'Turnovers', 'Steals', 'TotalRebounds', 'Opp.TotalRebounds', 'Opponent', 'TotalFouls', 'Team', 'Blocks', 'Opp.OffRebounds' and 'OffRebounds'* were determined from the Hill-Climbing method to be the best features for predicting the outcome of a win or loss for the second iteration .

```
[[281 243]
 [149 557]]
             precision    recall  f1-score   support

          L       0.65      0.54      0.59       524
          W       0.70      0.79      0.74       706

avg / total       0.68      0.68      0.68      1230
```

*Figure 48 Final Decision Tree values with a confusion matrix second iteration.*

The confusion matrix shown above in **Figure 48**, demonstrates the predictability of a team losing increasing from 60-65% and the predictability of a team wining remains the same at 70% Which averages out to being 68%, this makes this iteration's feature selection much more accurate than its predecessor. The 5-fold produces a predictability of 66-69% giving a 3% of cross-validation error, shown in **Figure 49**, however majority of the precision values lie between 66-67%.

```
fold 1 score: 0.6646341463414634
             precision    recall  f1-score   support

          L       0.64      0.52      0.57       426
          W       0.68      0.77      0.72       558

avg / total       0.66      0.66      0.66       984

fold 2 score: 0.6900406504065041
             precision    recall  f1-score   support

          L       0.62      0.60      0.61       398
          W       0.74      0.75      0.74       586

avg / total       0.69      0.69      0.69       984

fold 3 score: 0.6666666666666666
             precision    recall  f1-score   support

          L       0.63      0.52      0.57       417
          W       0.69      0.78      0.73       567

avg / total       0.66      0.67      0.66       984

fold 4 score: 0.6717479674796748
             precision    recall  f1-score   support

          L       0.60      0.60      0.60       407
          W       0.72      0.72      0.72       577

avg / total       0.67      0.67      0.67       984

fold 5 score: 0.6626016260162602
             precision    recall  f1-score   support

          L       0.59      0.61      0.60       411
          W       0.72      0.70      0.71       573

avg / total       0.66      0.66      0.66       984
```

*Figure 49 5-Fold Predictability Second Iteration*

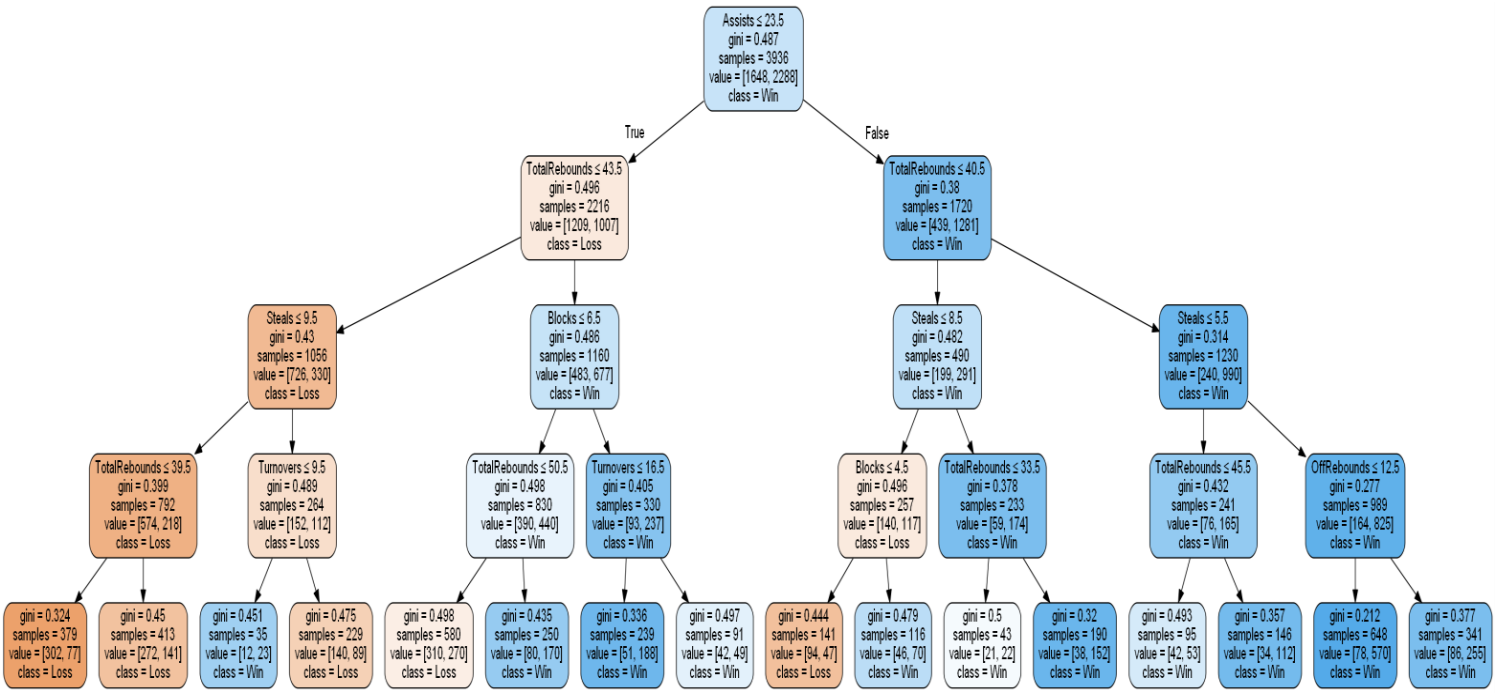



*Figure 50 Visualisation of the Decision Tree for the Second Iteration*

From **Figure 50** it can be shown that now Assists is the root node followed by TotalRebounds, furthering the indication that possession is no longer the key indicator but still a very high one for determining the overall outcome of an NBA game. Assists rates can be as high as 60% of every Field Goal in the NBA, making it a major factor to how many points a team scored and overall how well a team preformed on the day. Which its surprising that in the first iteration it was ignored as the root node and even a feature of the tree.

### 7.3.3 Third Iteration of Experiments

The third iteration or experiment consisted of *'OffRebounds', 'TotalRebounds', 'Assists', 'Steals', 'Blocks', 'Turnovers' and 'TotalFouls'.* The home team was used to compare, if there was a significant difference in how well the home team played comparison to the opposition. As in the last iteration the opposition data have very little to no significance in determining the outcome.

```
[[290 254]
 [131 555]]
             precision    recall  f1-score   support

          L       0.69      0.53      0.60       544
          W       0.69      0.81      0.74       686

avg / total       0.69      0.69      0.68      1230
```

*Figure 51 Final Decision Tree values with a confusion matrix third iteration*

The third experiments confusion matrix shown in **Figure 51**, portrays that the accuracy of being able to predict a team winning and the accuracy of being able to predict a team losing the game as the same at 69%. Stating that this method is an improvement compared to the second iteration of features, indicating that there may be an over fitting in the number of features for the first two iterations.

The 5-Fold method shown below in **Figure 52,** specifies a predictability of 65-69% giving a 4% of cross-validation error. Suggesting that still the second iteration has the best features for the determining the outcome of an NBA basketball game.

```
fold 1 score: 0.6341463414634146
             precision    recall  f1-score   support

          L       0.62      0.69      0.65       426
          W       0.74      0.67      0.71       558

avg / total       0.69      0.68      0.68       984

fold 2 score: 0.6443089430894309
             precision    recall  f1-score   support

          L       0.58      0.59      0.58       398
          W       0.72      0.71      0.71       586

avg / total       0.66      0.66      0.66       984

fold 3 score: 0.6117886178861789
             precision    recall  f1-score   support

          L       0.62      0.57      0.59       417
          W       0.70      0.74      0.72       567

avg / total       0.67      0.67      0.67       984

fold 4 score: 0.649390243902439
             precision    recall  f1-score   support

          L       0.67      0.51      0.58       407
          W       0.70      0.82      0.76       577

avg / total       0.69      0.69      0.68       984

fold 5 score: 0.6270325203252033
             precision    recall  f1-score   support

          L       0.59      0.54      0.56       411
          W       0.69      0.73      0.71       573

avg / total       0.65      0.65      0.65       984
```

*Figure 52 5-Fold Predictability Third Iteration*

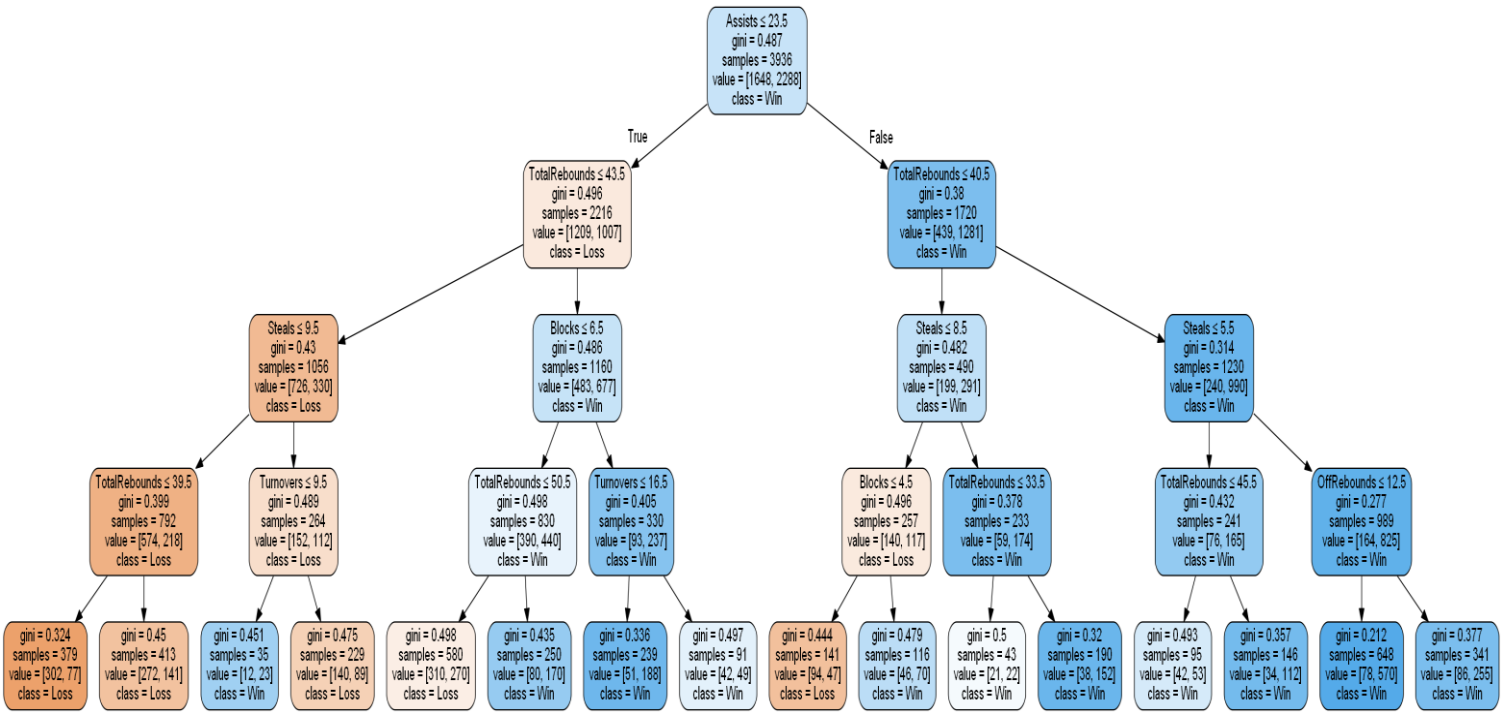



*Figure 53 Visualisation of the Decision Tree for the Third Iteration*

From **Figure 51** and **Figure 53,** strengthens our cause in saying that the opposing team has a limiting influence on the overall outcome of the game. If the teams keep up an elevated level of possession and assists, passing the ball around till they get a good shot opening and control the rebounds they have a very high chance of winning despite the opponent's skill level. This will need further investigation to determine however there is a strong basis for an additional study in this area.

### 7.3.4 Fourth Iteration of Experiments

In the fourth iteration of the experiment a further two features were added to the initial data frame. This was to see the implications of shots attempted and how they would change the predictability of the overall outcome of a game. The features that were select was the last 10th iteration seen in **Figure 54**, this was due to a significant increase in the predictability by nearly 2%. The features that were determined from the Hill-Climbing method consisted of; 'FieldGoalsAttempted', 'X3PointShotsAttempted','OffRebounds', 'TotalRebounds', 'Assists', 'Steals', 'Blocks' and 'Turnovers'.

```
Decision Tree with  1 [2]  selected features: 0.5799457994579946
Decision Tree with  2 [2, 8]  selected features: 0.5860433604336044
Decision Tree with  3 [2, 8, 4]  selected features: 0.6327913279132791
Decision Tree with  4 [2, 8, 4, 9]  selected features: 0.6348238482384824
Decision Tree with  5 [2, 8, 4, 9, 1]  selected features: 0.6348238482384824
Decision Tree with  6 [2, 8, 4, 9, 1, 6]  selected features: 0.6558265582655827
Decision Tree with  7 [2, 8, 4, 9, 1, 6, 7]  selected features: 0.6565040650406504
Decision Tree with  8 [2, 8, 4, 9, 1, 6, 7]  selected features: 0.6517615176151762
Decision Tree with  9 [2, 8, 4, 9, 1, 6, 7]  selected features: 0.6510840108401084
Decision Tree with  10 [2, 8, 4, 9, 1, 6, 7, 5]  selected features: 0.6686991869918699
```

*Figure 54 Number of features for incremental Decision Tree values using Hill-Climbing Method fourth iteration*

```
[[320 180]
 [206 524]]
             precision    recall  f1-score   support

          L       0.61      0.64      0.62       500
          W       0.74      0.72      0.73       730

avg / total       0.69      0.69      0.69      1230
```

*Figure 55 Final Decision Tree values with a confusion matrix fourth iteration.*

The confusion matrix shown in **Figure 55**, portrays that the gap between accuracy of being able to predict a team winning and losing is once again getting bigger. 61% was the fraction of correctly predicted out comes for a team losing compared to the 74% of accurately predicted teams winning. Which averages out to being 69%, when we compare this number against the 5-Fold method shown in **Figure 56,** the 5-fold produces a predictability of 64-69% giving a 5% of cross-validation error.

```
fold 1 score: 0.657520325203252
             precision    recall  f1-score   support

          L       0.65      0.45      0.53       426
          W       0.66      0.81      0.73       558

avg / total       0.66      0.66      0.64       984

fold 2 score: 0.7164634146341463
             precision    recall  f1-score   support

          L       0.69      0.54      0.61       398
          W       0.73      0.83      0.78       586

avg / total       0.71      0.72      0.71       984

fold 3 score: 0.6788617886178862
             precision    recall  f1-score   support

          L       0.62      0.65      0.63       417
          W       0.73      0.70      0.72       567

avg / total       0.68      0.68      0.68       984

fold 4 score: 0.6798780487804879
             precision    recall  f1-score   support

          L       0.61      0.65      0.63       407
          W       0.74      0.70      0.72       577

avg / total       0.68      0.68      0.68       984

fold 5 score: 0.6371951219512195
             precision    recall  f1-score   support

          L       0.56      0.59      0.58       411
          W       0.69      0.67      0.68       573

avg / total       0.64      0.64      0.64       984
```

*Figure 56 5-Fold Predictability Final Iteration*

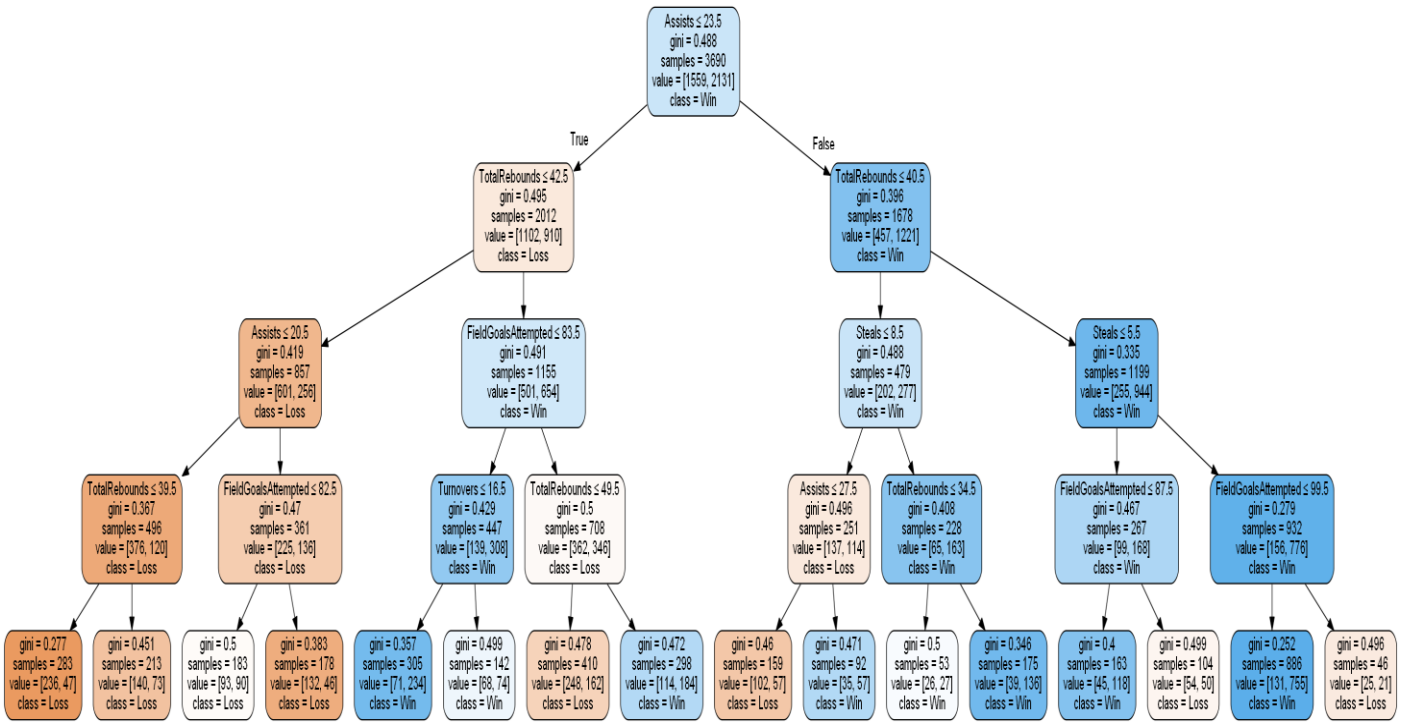



*Figure 57 Visualisation of the Decision Tree for the Final Iteration*

Frome the **Figure 57,** it can be seen that FieldGoalsAttempted has a somewhat significant role in our tree in the 4th row or branch of our tree, showing that going above 80 field goal attempts actually decreases your chances of winning. This can be attributed to the rut of missing a shot and the losing confidence in taking another, or other risky shots as well as just shooting the ball more times than your enemy and not taking good and accurate shots.

# 8 Conclusion and Recommendations

The experiment showed how adding and removing features differed the overall results in the predictability of the outcome of an NBA game via Decision Trees and the K-NN Methods.

The K-NN method provided a more accurate predictability in determining the outcome of an NBA basketball match with the Cross validation across five folds of data produced precision ratings between 76 - 79%, compared with the 66 - 69% of the Decision Tree model. This may be attributed to the Hill-Climb method selecting different features respectively and further research in this area would determine whether or not this was the case. *Assists, TotalRebounds, Blocks, Steals, Opp.Assists, Opp.TotalRebounds, Opp.Blocks* and *Opp.Steals,* were the features determined by this study to yield the greatest predictability in whether or not a team would win or lose.

Future research could look at analysing individual players and see how their performances influence match outcome. In addition, analysing how groups of players complement each other's playing styles might be beneficial.

# 9 Bibliography

Hassan, D. A. U. &. W. W., 2014. *Comparison of distance metric for hierachial data in medical databases.* s.l.:s.n.

Sehra, C., 2018. *Decision Trees Explained Easily*. [Online]
Available at: https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248
[Accessed 26 May 2018].

Zhang, Z., 2016. Introduction to machiene learning: K-nearest neighbhours.. *Annals of Translational medicine.*