NBA 2018 Playoffs and Championship Prediction

Amit Kumar
Jayesh Mehta
Xiaohai Su

ABSTRACT **- The NBA (National Basketball Association) is a men's professional Basketball league in North America consisting of 30 teams. Each year around 1250 matches are played among these teams to decide a winner. The goal of our project is to use machine Learning Algorithms to predict the playoffs matches for the year 2018 and finally predict a winner team for NBA 2018.**

I. INTRODUCTION

The dataset used for this project has been scrapped from www.basketball-reference.com as shown in Figure[1]. The dataset is for the past 3 years anc contains data for almost 1100 players from the 30 teams of NBA Each player has about 34 features. Since we cannot include all features in our model, we use correlation techniques and RFE to obtain the 10 most relevant features. The score from these features are used to score a team as a whole. This data is fed into Logistic Regression algorithm to determine the outcome of a match.

II. DATA PREPARATION

The data obtained from online sources is not always clean and error free. We need to make sure that the data is cleaned and ready for analysis. Different techniques are used to clean the data and remove and null values from the dataset.

There are many data entries which not complete or are out of range, all such entries need to be removed. We generated heat maps to ensure that all values were cleaned and all corrections were made.

There were instances during our project where players got injured and hence needed to be replaced from the dataset.

The dataset was then ready for Analysis & Modelling.

III. DATA ANALYSIS

The final dataset contains around 1000 entries. Since we have multiple data points for each player, we take average for each feature in the dataset.

Feature Selection - Having a good understanding of feature selection/ranking can be a great asset. Using Feature selection leads to better performing models, better understanding of the underlying structure and characteristics of the data and leads to better intuition about the algorithms that underlie many machine learning models. Hence we perform Pearson Correlation & Recursive Feature Elimination.

Pearson Correlation - One of the simplest method for understanding a feature relation to the response variable is Pearson correlation coefficient, which measures linear correlation between two variables. The resulting value lies in

[-1;1], with -1 meaning perfect negative correlation (as one variable increases, the other decreases), +1 meaning perfect positive correlation and 0 meaning no linear correlation between the two variables.

Recursive Feature Elimination - Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coefficient attribute or through a feature importance attribute. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

By using the above two techniques we are able to derive a set of 10 most relevant and important features would help us judge a player's performance. Few of the important features include 'Salary', '3PT Score' etc. These features are represented in Figure[5].

After Elimination and Averaging features of the players, we calculate the player score based on the available feature set. Each player is scored individually based on the formula we have created.

Player Scoring - According to our formula every player feature is assigned a value with the range [1,10], '1' being lowest feature score and '10' being the highest feature score. Since there are 10 features which are highly relevant, the range of every player score is from 10 to 100.

For example, one of the most important features is 'Exp'. If the highest value of 'Exp' of all the players in the league is 30, the lowest value is 10 and the mean value is 20. Then we divide the score of 'Exp' feature into two parts and give each part 5 divisions. For the high level part, which is the range that is from the mean value to the highest value.

we calculate that the length for the high level has 5 divisions is given by -

$$\frac{Highest\ Value - Mean\ Value}{5} \Rightarrow \frac{30-20}{5} \rightarrow 2$$

This means the 5 divisions of the high level are: 20~22, 22~24, 24~26, 26~28 and 28~30. We will give players whose 'Exp' features are at those ranges 6, 7, 8, 9 and 10 points respectively. Based on this, the range of the low level part is from the lowest value to the mean value, which is from 10 to 20. And the length for low level part has 5 divisions is as follows -

$$\frac{Lower\ Value - Mean\ Value}{5} \Rightarrow \frac{30-20}{5} \rightarrow 2$$

. The 5 divisions of the low level part are: 10~12, 12~14, 14~16, 16~18 and 18~20. We will give players 1, 2, 3, 4 and 5 points respectively.

We add all 10 features score for each player and consider the final score as the player score. This is how we score each player based on the features we select and formula we create.

Team Scoring - For the team score, considering each team has different number of players and players whose scores are very low don't make any effort to their teams, we decide to take into account the top 9 players for each team and make sure that every team has this number of players by checking our cleaned data.

The final team score equals to the mean value of all the top 9 players score in this team.

IV. MODELLING

Since, our aim was to predict the binary outcome. We used Logistic Regression for our project.

Logistic Regression - is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or

more nominal, ordinal, interval or ratio-level independent variables.

$$\sigma(t) = \frac{1}{1 + e^{t}}$$

$$F(x) = \frac{1}{1+e^{(\beta_0+\beta_1 x)}}$$

Assumption for LR Modelling - The dependent variable should be dichotomous in nature (e.g., presence vs. absence). There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores, There should be no high correlations (multicollinearity) among the predictors. This can be assessed by a correlation matrix among the predictors. Tabachnick and Fidell (2013) suggest that as long correlation coefficients among independent variables are less than 0.90 the assumption is met.

Overfitting - When selecting the model for the logistic regression analysis, another important consideration is the model fit. Adding independent variables to a logistic regression model will always increase the amount of variance explained in the log odds (typically expressed as R^2). However, adding more and more variables to the model can result in overfitting, which reduces the generalizability of the model beyond the data on which the model is fit.

Since, we need to predict the outcome of the matches which are yet to played. So, we first took equal number of players from each team as shown in Figure[2]. The data was fed to the Logistic Regression model and we got the predicted outcomes of each match as shown in figure[6]. Furthermore, Our model predicted the Playoff teams for both East zone and West Zone as shown in Figure[7] and Figure[7.1]. Now, we wanted to test our model against the actual match results. At the time of writing this report, around 390 matches has been played. So, we tested our model against the actual match results as shown in Figure[8] and Figure[9]. We can see our predicted wins in Figure[8] and the actual match results in Figure[9]. The Accuracy of our model against the actual match result is 73.5%.

V. CONCLUSION

After performing Analysis and Modelling we are able to conclude the Top 8 Teams from the both the East & West Divisions. We have also compared accuracy from different models and find that we get the best accuracy of 73.5% from Logistic Regression. This could be attributed to that fact that our result is purely binary and hence Logistic Regression performs better. We also can confidently predict that Houston Rockets and Cleveland Cavaliers will play in the Final game and Houston Rockets should win the Championship title.

VI. FUTURE SCOPE

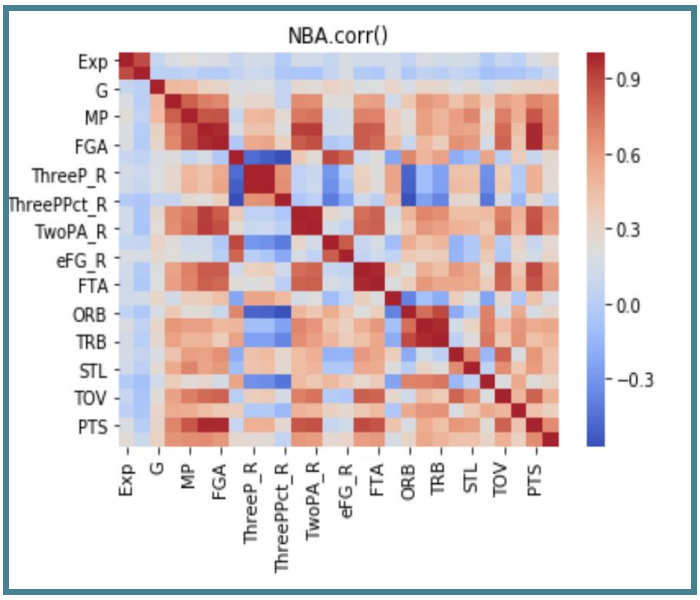
We need to take into account ever changing nature of the game. We plan to connect our model to some API which can feed the real time data to our model. This way, our model can address the constraints like player injuries, ever changing game statistics etc. Our model can utilize the real time data for better training of the model.

V. APPENDIX

Figure 1: Variables Description

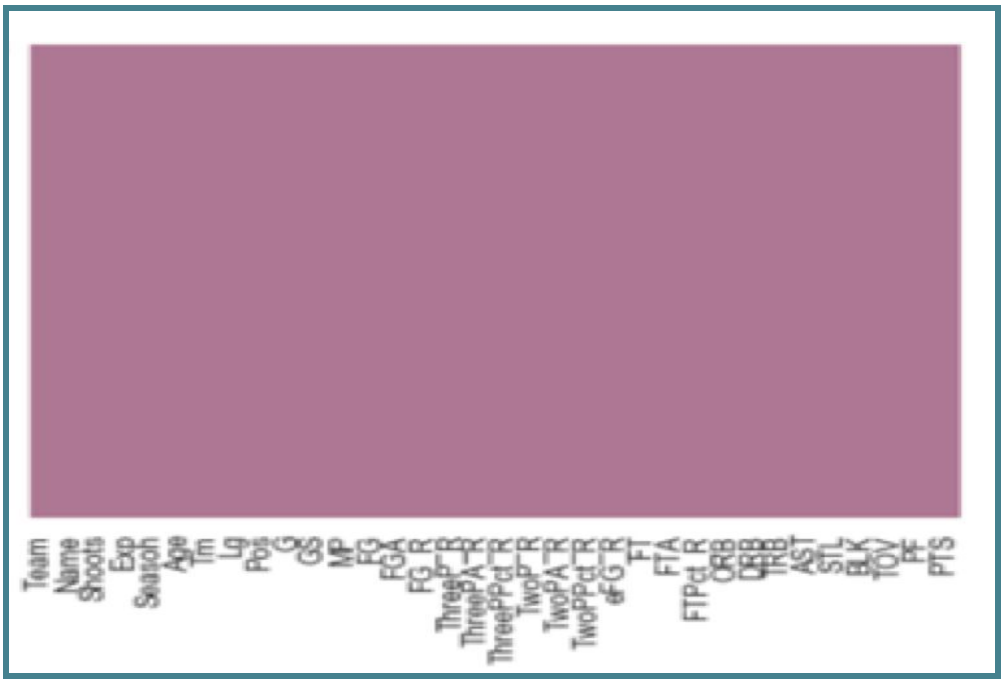
Variables	Description	Feature Importance
Team	The team names of all the teams in NBA	High
Team_Score	The mean value of the top 9 players score of every team	High
Salary	The salary of each player in 2018 season	High
Exp	Experience - How long the players play for NBA (unit is year)	High
G	Game - The total number of games that players participate in every season	High
GS	Game started - The number of games that players participate in when games started every season	High
MP	Minute per game - The average minute of how long the players play for each game (unit is minute) every season	High
FGA	Field goal attempt - The average frequency of how many times the players attempt to score per game every season	High
ThreePoint	The average frequency of how many times the players get the 3-Point successfully per game every season	High
PTS	The average points that players get per game every season	High
Player_Score	The player score calculated based on our own formula	High

Cleveland Cavaliers	9
Portland Trail Blazers	9
Golden State Warrior	9
Denver Nuggets	9
Philadelphia 76ers	9
Dallas Mavericks	9
Chicago Bulls	9
Detroit Pistons	9
Boston Celtics	9
Oklahoma City Thunder	9
Indiana Pacers	9
Washington Wizards	9
Houston Rockets	9
Toronto Raptors	9
Atlanta Hawks	9
Brooklyn Nets	9
New Orleans Pelicans	9
Milwaukee Bucks	9
Orlando Magic	9

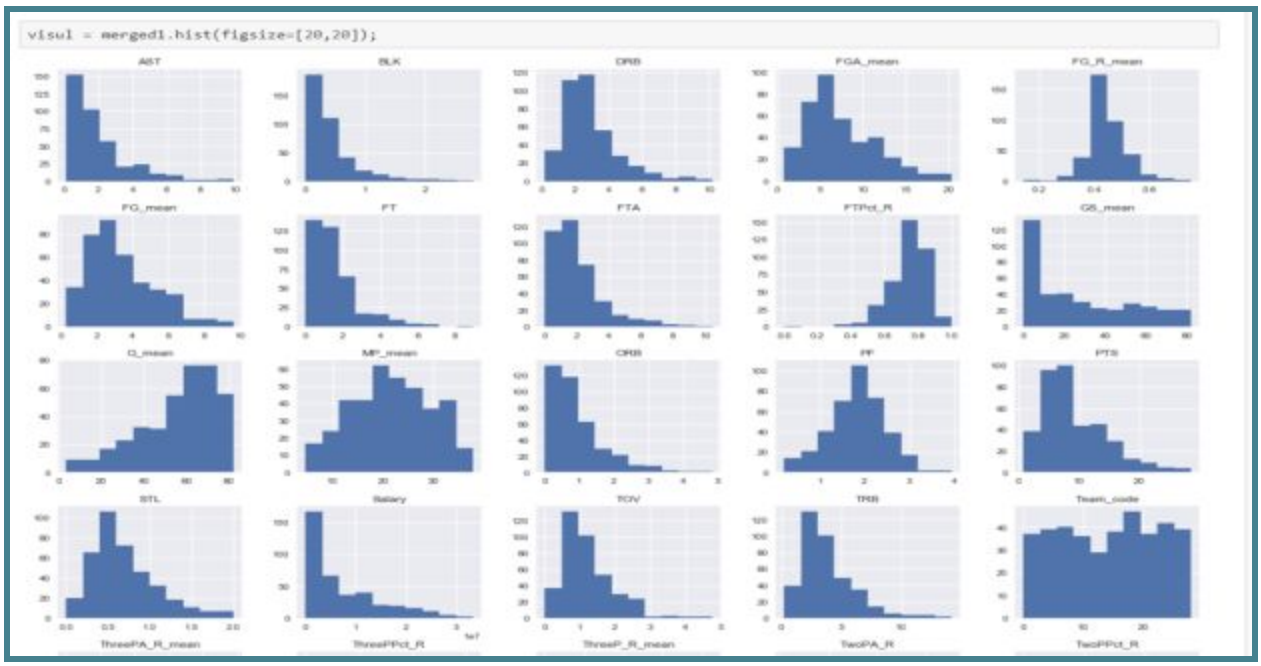


Figure[2]: Player Count

Figure[3]: Correlation Matrix



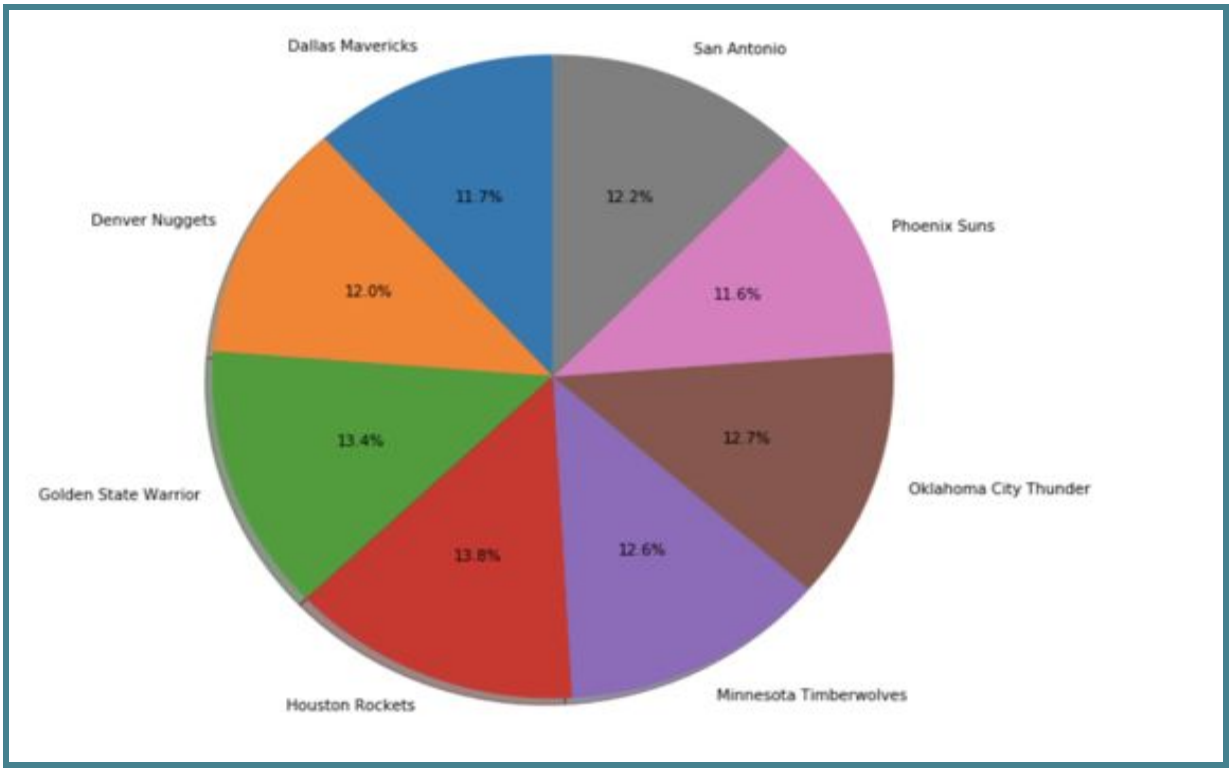
Figure[4]: HeatMap



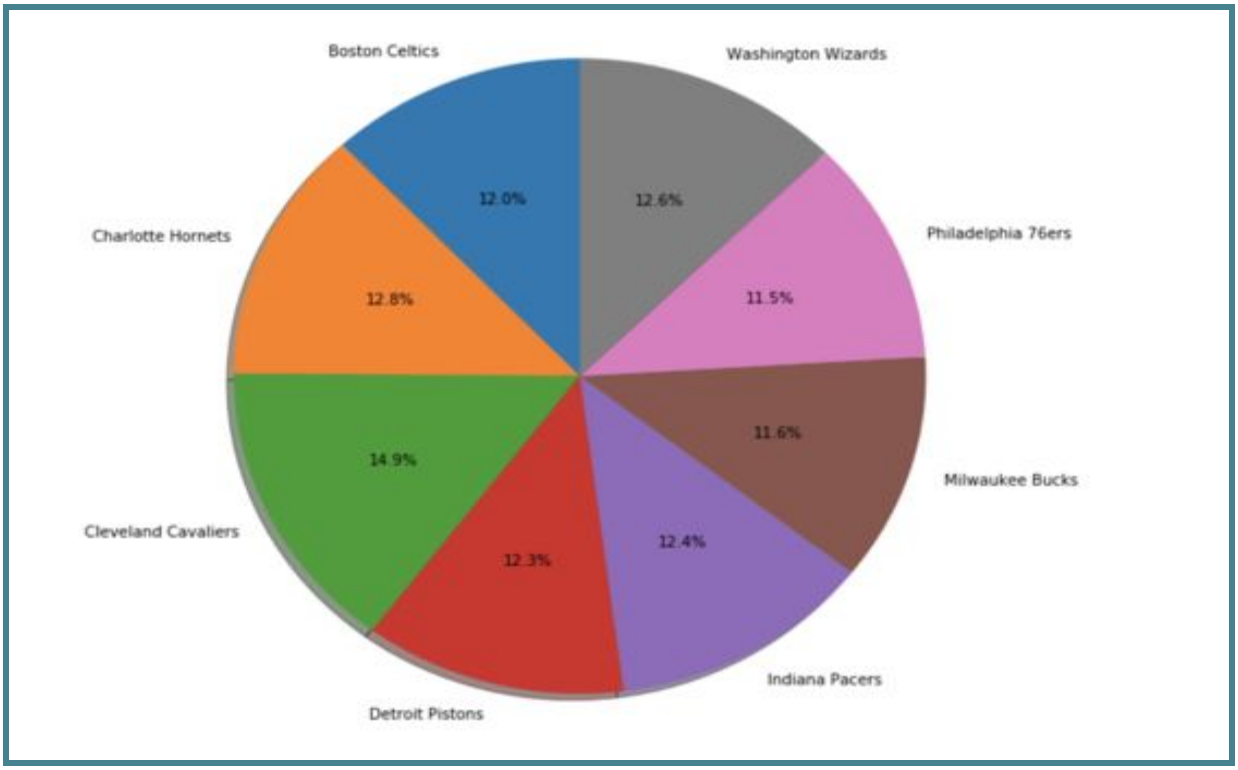
Figure[5]: Feature Representation

```
In the Match between  Boston Celtics  and   Cleveland Cavaliers  :>>  Boston Celtics will loose !
In the Match between  Houston Rockets  and  Golden State Warrior  :>>  Houston Rockets will win !
In the Match between  Milwaukee Bucks  and   Boston Celtics  :>>  Milwaukee Bucks will loose !
In the Match between  Atlanta Hawks  and   Dallas Mavericks  :>>  Atlanta Hawks will loose !
In the Match between  Charlotte Hornets  and   Detroit Pistons  :>>  Charlotte Hornets will loose !
In the Match between  Brooklyn Nets  and   Indiana Pacers  :>>  Brooklyn Nets will loose !
In the Match between  New Orleans Pelicans  and  Memphis Grizzlies  :>>  New Orleans Pelicans will win !
In the Match between  Miami Heat  and  Orlando Magic  :>>  Miami Heat will win !
In the Match between  Portland Trail Blazers  and   Phoenix Suns  :>>  Portland Trail Blazers will loose !
In the Match between  Houston Rockets  and  SAN ANTONIO  :>>  Houston Rockets will win !
In the Match between  Minnesota Timberwolves  and  Sacramento Kings  :>>  Minnesota Timberwolves will win !
In the Match between  Denver Nuggets  and   Utah Jazz  :>>  Denver Nuggets will loose !
In the Match between  Philadelphia 76ers  and   Washington Wizards  :>>  Philadelphia 76ers will loose !
In the Match between  Los Angeles Clippers  and   Los Angeles Lakers  :>>  Los Angeles Clippers will loose !
In the Match between  New York Knicks  and   Oklahoma City Thunder  :>>  New York Knicks will loose !
```

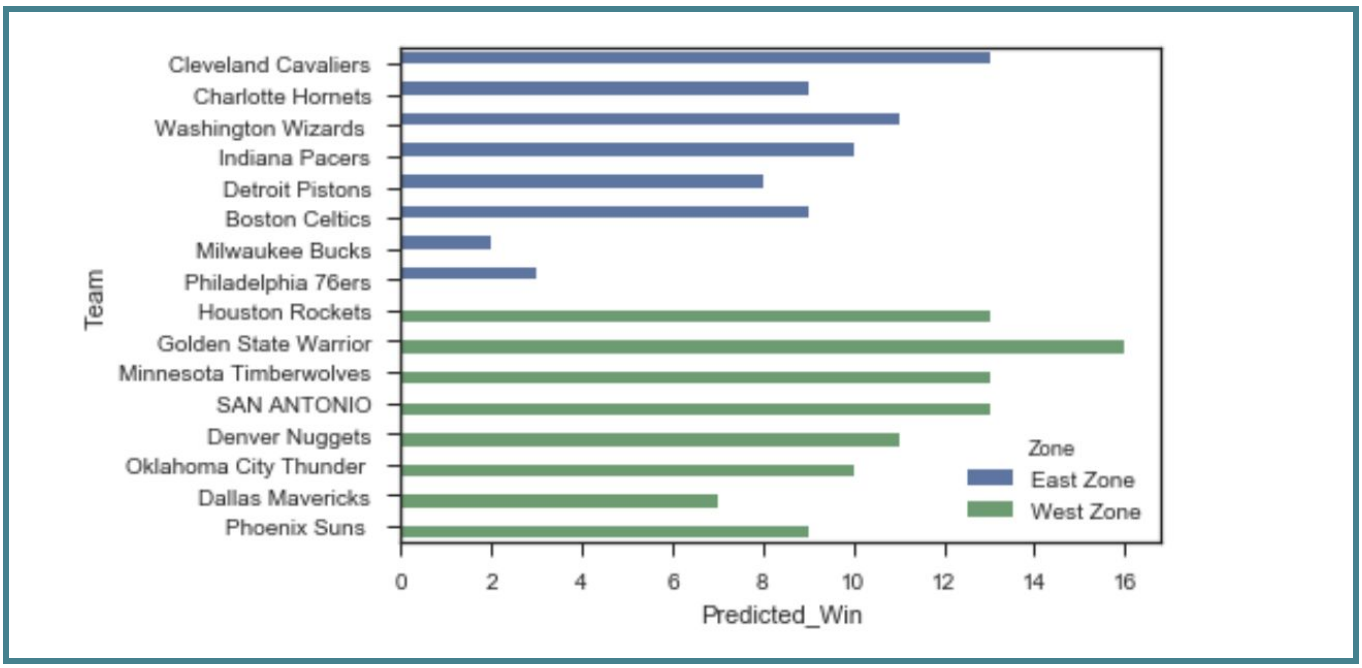
Figure[6]: Predictions of Match Results



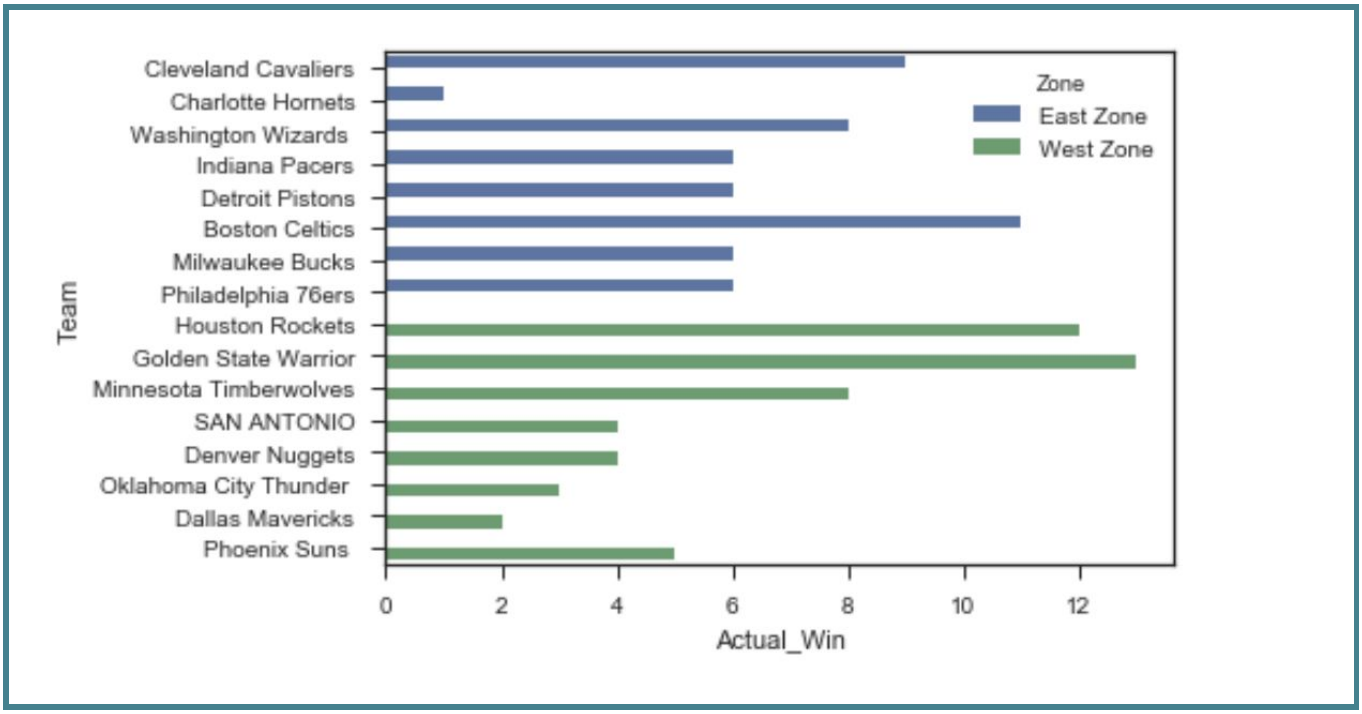
Figure[7]: Playoff Predictions for West Zone



Figure[7.1]: Playoff Predictions for East Zone



Figure[8]: Predicted wins for the 390 Matches which has been played so far.



Figure[9]: Actual wins for the 390 Matches which has been played so far.