IDS 575

Statistical Models and Methods for Business Analytics

PREMIER LEAGUE PREDICTOR

2016-17

GROUP MEMBERS

Sudhir SB

660388617 | ssuhas2@uic.edu

Mihir Singh

668554460 | msingh44@uic.edu

Sandeep G

653635677 | sgopin5@uic.edu

Debdeep Dhar

673827520 | ddhar3@uic.edu

Adarsh Tomar

675593372 | atomar3@uic.edu

CONTENTS

1. OVERVIEW

Premier League

The Premier League is an English professional league for men's association soccer clubs. At the top of the English football league system, it is the country's primary football competition. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League.

A season runs from August to May having total 380 matches. During the season, each club plays the others twice (a double round-robin system), once at their home stadium and once at that of their opponents', for a total of 38 games. Teams receive three points for a win and one point for a draw. No points are awarded for a loss. Teams are ranked by total points, then goal difference, and then goals scored. If still equal, teams are deemed to occupy the same position. If there is a tie for the championship, for relegation, or for qualification to other competitions, a play-off match at a neutral venue decides rank. Team finishing first at the end of season would be crowned champions. The three lowest placed teams are relegated into the lower division, EFL Championship, and the top two teams from the Championship, together with the winner of play-offs involving the third to sixth placed Championship clubs, are promoted in their place.

2. OBJECTIVE

The Premier League is a corporation in which the 20 member clubs act as shareholders. The wins/losses of a club affect its brand and ultimately revenue through merchandise, sponsorship deals etc. There are numerous factors affecting the wins/losses in a game of a club. So, the main aim of our research is to identify these factors for a club, specifically Manchester United, and predict its final league standing against its top five competitors at the end of the current season 2016-17.

We hope to achieve this is 3 steps:

1. Create 3 logit models:

   a) Logit Model to predict outcome based on Pre-game statistics- We have certain variables that define whether the match is a Home/Away game, the opponent team category which could help us predict the odds of wins better

   b) Logit Model to predict outcome based on In-game statistics- Using variables which include shots on goal, total free kicks

   c) Mixed- Model which uses a combination of the above variables to predict an outcome using logit

2. Choose the best model from the above using necessary parameters.

3. Use Time series (ARIMA) to predict variable values for the remaining games of the top 6 teams. Use the chosen logit model on the forecasted match statistics to predict win/loss. This will give us a ranking of Manchester United in respect to its top 5 competitors.

3. DATASET

Original Dataset-

We have identified our dataset from the site: http://www.football-data.co.uk/englandm.php for our research objective of forecasting the outcome of a soccer game in the English Premier League. The source aggregates the data for every given season. We have taken the data for 5 seasons from 2012-13 season to 2016-17 season. The data is separately taken for each of the five seasons i.e. five datasets files comprising the data for five seasons separately. The data is year-wise which help for time series-forecasting. Also, the data is available on a game-week level, data is available for each match-day fixture between two given teams.

The observations and variables count of each of the files is shown below:

Season	Observation count	Variables Count
Season 2012-13	380	74
Season 2013-14	380	68
Season 2014-15	381	68
Season 2015-16	380	65
Season 2016-17	283	65

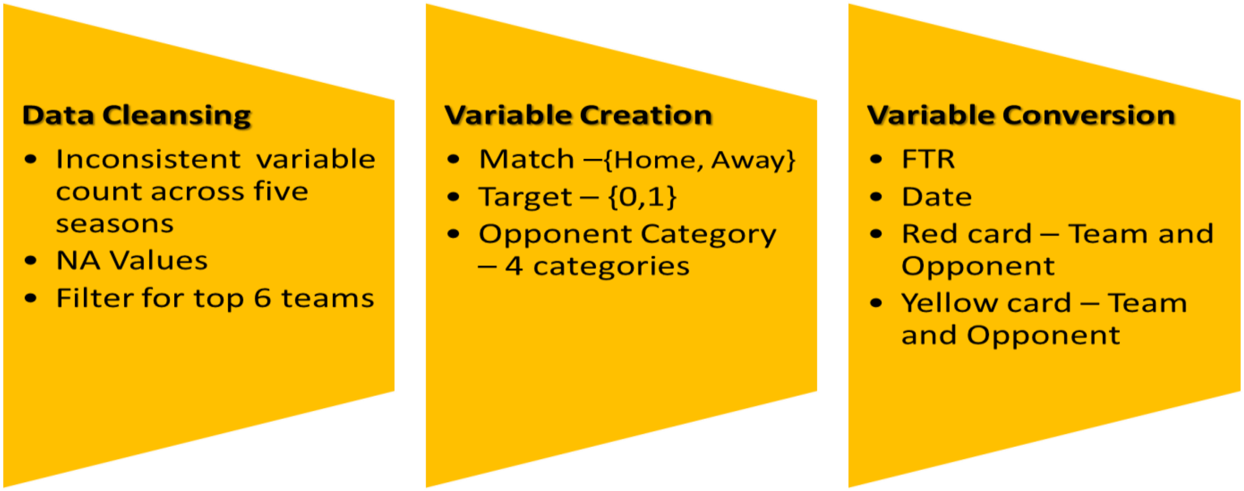
Few of the variables of our dataset are Match Date, Opponent team, Corners, Free kicks, Home Team, Away Team, Fouls, Referee, Yellow and red cards, Half-time goals, Full-time goals, Half-time Result, Full-time Result etc.

Final Dataset-

Our final dataset was obtained after executing all the data preprocessing steps mentioned in the next page.

1. For Logit modelling – The data was at a Match level, with variables like Date, Match (Home/Away), Opponent team category, Team shots on target, corners, Total free kicks and other variables that seemed important. We did however remove the betting odds from our data. We were left with 8 numerical variables and a few categorical variables in the end
2. For Time series – After predicting the values of the above variables for the future games, we split the data up into 6 individual datasets per team. The outcome of each match for each team was done separately.

4. DATA PREPROCESSING



Data Cleansing

- Inconsistent variable count across five seasons
- NA Values
- Filter for top 6 teams

Variable Creation

- Match –{Home, Away}
- Target – {0,1}
- Opponent Category – 4 categories

Variable Conversion

- FTR
- Date
- Red card – Team and Opponent
- Yellow card – Team and Opponent

We did a fair amount of data preparation and preprocessing to arrive at the final dataset that would be fit for modelling –

1) <u>Data Cleansing:</u>

- There was inconsistency in the variables count across five seasons mainly due to the set of betting variables, which, anyways, we were not factoring in our analysis so we removed these set of betting variables ultimately maintaining the same variable count overall.
- There were few NA values, which were basically the unavailable values of the variables of the future games.
- We filtered the dataset for only the consistent top 6 teams across the five seasons to remain focused on our goal of predicting the league standing of Manchester United against its five competitors at the end of the current season 2016-17.

2) <u>Variable Creation:</u>

- Instead of using the original "HomeTeam" and "AwayTeam" variables separately, we created the variable “Match” for noting whether it was a Home match or an away match for a team to help us in our analysis.

- We created a new variable “Opponent Category” having 4 categories, TT- Top Teams, FR- Fighting Relegation, R- Relegated and MT- Mid Tier.
- We created the variable “Target” – {0,1} as not a win or a win as opposed to the original FTR (Full Time Result) showing a win, lose or a draw, to better cater to our objective.

3) Variable Conversion:

- Converted FTR variable from Character to Factor
- Converted Date variable from Character format to Date format
- Red Card – Team & Opponent variables also converted to factor variable with values Yes & No.
- Yellow Card – Team & Opponent variables binned and converted to factor variable with values Yes & No.

5. DATA ANALYSIS AND PATTERNS

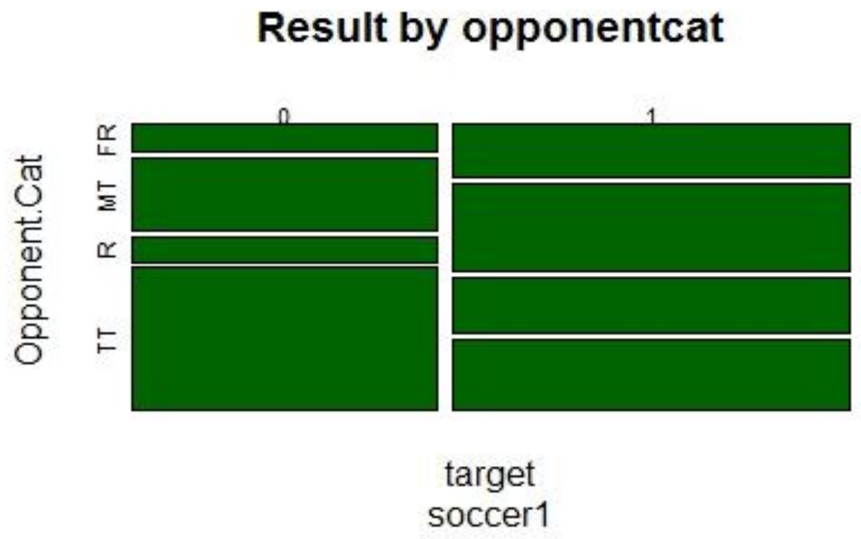
Exploratory data analysis was done to identify the relationships between various variables in the dataset. The following relationships were done with various tests, based on the type of variables (Categorical , numerical etc.):

1. Between Categorical Variables:

Target and Opponent Category:

Before this analysis, the premier league teams were broadly categorized into 4 – Top Team(TT) , Mid Table(MT) , Fighting for relegation(FR) , Relegation(R). The analysis between the target and the opponent category is done to identify if the opponent team’s performance history has an influence on the result.

The following stacked column chart depicts that the probability of losing a match against top team is higher than against a team which is fighting for relegation



Analysis 1 : Target and Opponent category

Target and Red cards:

The analysis between the target and the red cards was done to check if there was any significant impact on the results when the red card is given. The Chi-square test revealed the relation between Red cards for team and their wins/losses.

Redcards Target	0	1	2
0	432	42	4
1	597	23	0

X-squared = 15.04, df = 1, p-value = 0.0001053.

From the p- value , it is evident that this relationship is significant and has an impact on the match result.

Target and Yellow cards:

The original dataset had the count of yellow cards in each match. The data was primarily concentrated within 0 1 2 yellow cards, hardly any values for the high numbers. Hence it was transformed into categorical variable.

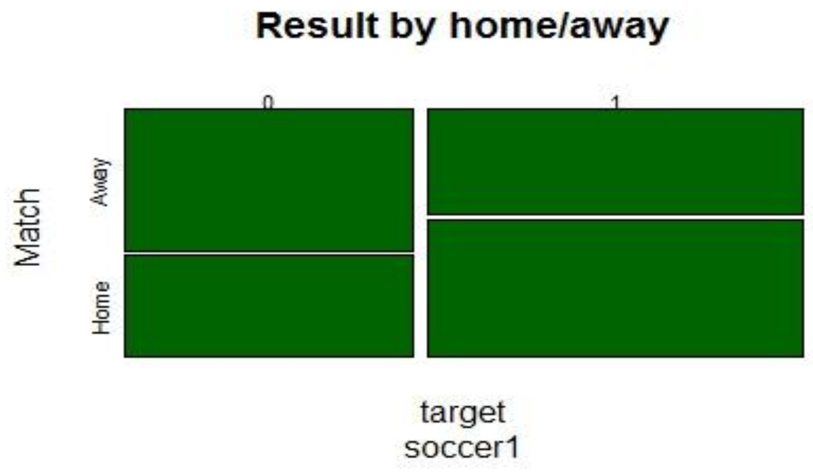
<u>Chi squared test for Transformed variable:</u>

X-squared = 10.962, df = 1, p-value = 0.0009298.

As it can be observed from the above test result, this relationship is significant.

Target and Match:

This analysis was done to identify if a team’s performance is influenced by the home/away match factor. The following stacked column chart shows that the most match results were in favor of the home team.

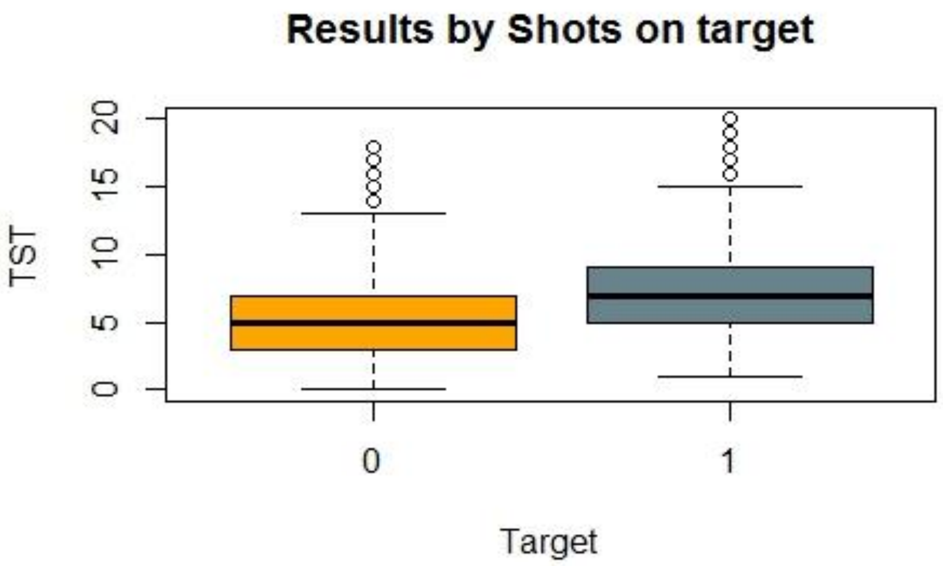


Analysis 2: Target and Match

2. **Between Categorical and Numeric variable:**

Target and TST(Team shots on Target):

The analysis between the total shots on target and the target(win/loss) yielded the following box plot :

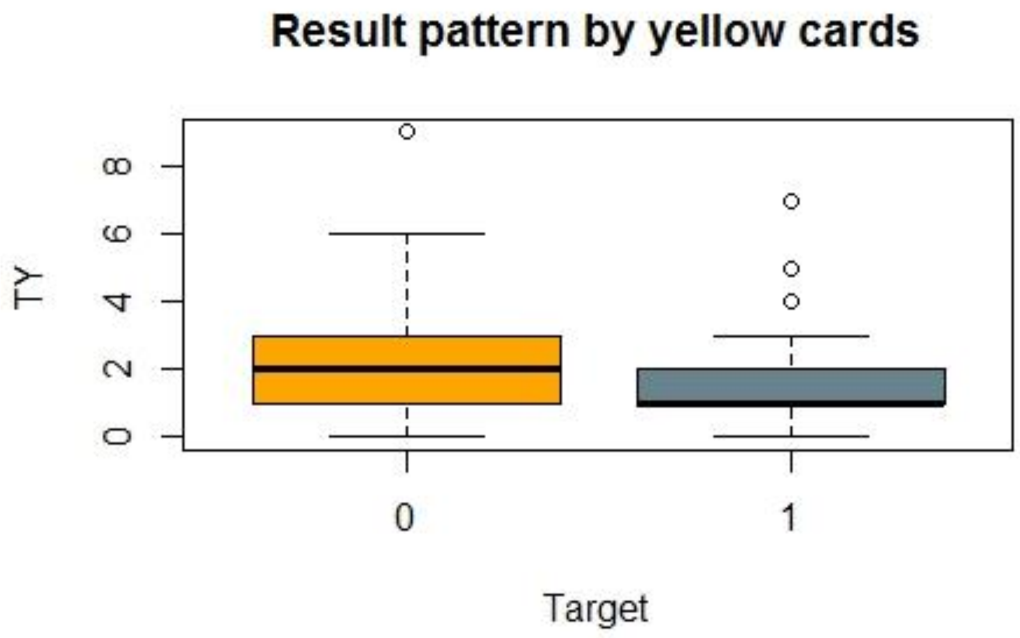


Analysis 3: TST and Target

In the above boxplot, it can be observed that the team with more shots on target have better chances of winning the match. More than 50% of the win results has more than 7 shots on target by the winning team.

Target and Team Yellow(TY)

The analysis between the Target and Team yellow variable was to identify the yellow card influence on the match result and revealed the following plot :

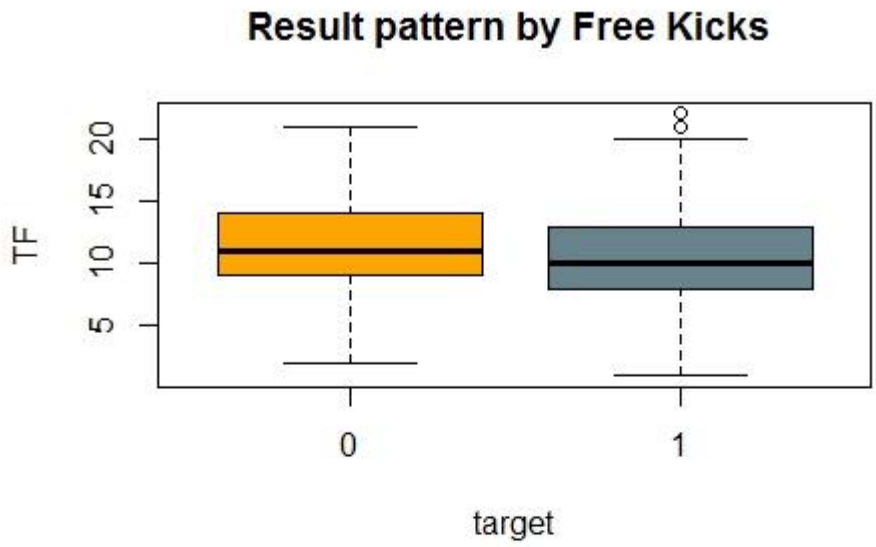


Analysis 4: Target and Team Yellow

The boxplot shows that the teams which had more than 3 yellow cards had more chances of losing the match.

Target and Free Kicks:

The analysis between the Target and Free Kicks variable was to identify the number of free kicks influence on the match result and revealed the following plot :



Analysis 5: Target and Free Kicks

The above boxplot shows that the number of free kicks does not have much influence on the final match result. This is an expected trend because the result would depend on the conversion rates of these free kicks as well.

3. Between Numeric variables

The correlation test was performed on various numeric variables and a set of results are presented in the following table :

Variable combination	Correlation Value	Observation
Team Shots, Team Corners	0.5138017	correlated. This is because the probability of a shot on target increases with number of corners
Team Shots on Target, Team Corners	0.316158	
Team Goals , Team shots on target	0.4100111	Correlated
Team Goals , Team Free kicks	-0.08672609	Negatively correlated. This is interesting . The reason could be that the conversion rate of the free kicks/corners are poor.
Team Goals, Team Corners	-0.08363878	

6. APPROACH

Our main goal of predicting the final league standings of the top 6 teams of the English Premier League for the current season (2016-2017) could be achieved in many ways through our dataset.

Dividing our predictors into pre-game statistics such as Referee, Opponent Team, Home or Away match etc. and in-game statistics such as Free kicks, corners, shots, red and yellow cards etc., we come up with three approaches or models:

1. Mixed Model:

For the Mixed Model, both the pre-game and in-game statistics are considered for making the Logistic Regression model. The aim of the Logistic Regression is to consider the significance of all the variables to predict the dependent variable which is 'Target'. Initially the model is built with all the given variables. Based on the significance of all the variables, an iteration of Logistic Regression is carried with only the significant variables. Third Iteration begins with running the Principal Component Analysis, Principal components are then combined with the categorical variables for making a Logistic Regression model for predicting the 'Target' Variable. For the fourth iteration, factor analysis was done to determine the important factors, which are then combined with the remaining categorical variables to make another Logistic Regression Model. Model with the best AIC is considered for the model that will be used to predict the final results.

2. In-game Statistics Model:

The same approach is followed as the Mixed Model, except this time only the in-game statistics like: Total Freekicks, Corners and the cards are considered as the independent variables for the Logistic Regression Model. Like the Mixed-model, the analysis begins with making a Logistic Regression model for all the independent variables. Based on the result of the first iteration, only the significant variables are considered for the second iteration. Third Iteration, Principal Component Analysis was done for all the numeric variables and the main Principal Components were combined with the categorical variables for the model creation. Fourth Iteration, Factor Analysis was done to identify the key factors, which was used to generate the Logistic Model. The model with the best AIC was selected.

3. Pre-Game Model:

For the Pre-game model only two important variables are considered: The Type of match i.e. either Home or Away, Type of opponent: top tier, mid-table, relegation threatened, teams relegated. The model is not expected to have a very high level of accuracy as more variables need to be considered to make a full-fledged model. Variables like Number of Suspensions, Injuries etc could be considered while making the pre-game model in addition to the two variables considered above.

From the three different models considered, the best model amongst them is selected. Using time series forecasting, the values for the upcoming games are forecasted. Then the values are fed back into the Logistic Regression Model to generate the results that will help in obtaining the Final standings.

7. Logistic Regression Models

Mixed Game Model-

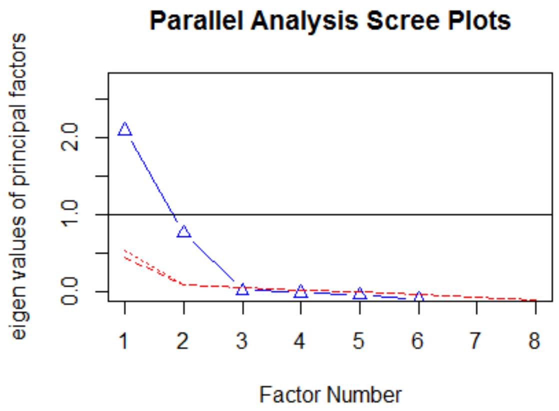
Modelling for the response “Target”, considering predictors such as opponent team, corners, free kicks, home or away match, yellow and red cards etc.

The Logistic Regression model is built on the data of the six teams under consideration from the season 2012/13 to the current season. For the analysis, the data was divided into training and test for the measurement of performance.

The training data sample contained 75% of the data and test contained the other 25%.

We generated various models for the following scenario and analyzed the results based on the training and testing performance:

- Considering All the variables: All the independent variables were considered for making the Logistic Regression Model. After Making the model, the AUC and the performance for the model was calculated.
- Considering the significant Variables: From the logistic regression run for all the variables, only the significant variables were considered for the model generation. Like the previous model, the AUC and the performance metrics were measured.
- Principal Component Analysis: PCA test was carried out for all the numeric data and the PC's that covered up to 85% of variance were considered for the model creation. Five Principal Components and the categorical variables were used as the independent variables in the Logistic Regression. The AUC and the performance metrics were also considered.
- Developing on the above model, another model with only the significant Variables and PC's was used for making another Logistic Regression model.
- Factor Analysis: Factor Analysis was done to determine the important Factors amongst all the variables considered. The analysis yielded three important Factors: Factor 1 : Opponent Shots and Opponent Shots on Target, Factor 2: Total Shots and total corners Factor 3:Total shots on Target. The factors along with the categorical variables were used for the Logistic Regression.



- **From the scree plot we see that 3 factors is ideal for the analysis.**

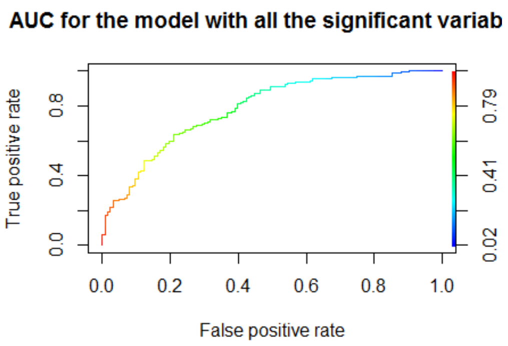
VIF test was performed in the first scenario itself, observing that there is no multicollinearity.

The following table summarizes the different performance measures for the scenarios-

Model	Training Perf	Testing Perf	AIC	AUC
All variables	74.6	73.03	916.91	0.81
Significant Variables	74.2	72.9	873.06	0.808
With PCA	71.5	71.4	921.44	0.808
Significant Variables with PCA	72.3	71.9	948.68	0.801
Factor Analysis	71.9	72.9	929.44	0.764

The model selection was done based on the AIC value. Here amongst all the iterations, the model with all the significant variables is chosen.

ROC Plot – For the model with all the significant variables.



- Other tests were also considered:

  Durbin-Watson: no correlation between residuals for all the models.

  Over-dispersion: close to 1 for all the models

- Different seeds were considered, and we observed that our Final model is the most robust to different sample data.

- Interactions were considered, and we observed that there were not any significant ones.

Significant predictors – OS**,** TST**,**OST,OF,TC**,**OC,O_R**,**Match**,**Opponent.Cat

<u>In-Game Model-</u>

Modelling for the response "Target", considering of only predictors that were logged during the course of the football game. Independent Variables that were considered during the analysis were Team Corners, Opponent Corners, Team Freekicks, Opponent Freekicks etc.

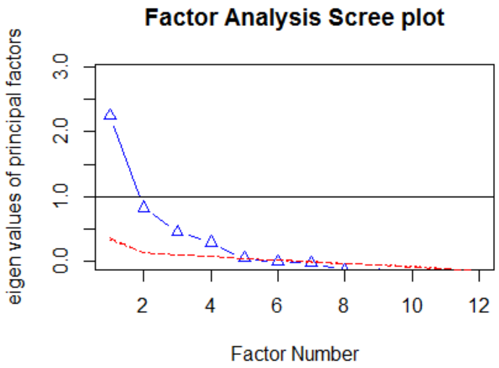
The Logistic Regression model is built on the data of the six teams under consideration from the season 2012/13 to the current season.

For the analysis, the data was divided into training and test for the measurement of performance.

The training data sample contained 75% of the data and test contained the other 25%.

We generated various models for the following scenario and analyzed the results based on the training and testing performance:

- Considering All the variables: All the independent variables were considered for making the Logistic Regression Model. After Making the model, the AUC and the performance for the model was calculated.
- Considering the significant Variables: From the logistic regression run for all the variables, only the significant variables were considered for the model generation. Like the previous model, the AUC and the performance metrics were measured.
- Principal Component Analysis: PCA test was carried out for all the numeric data and the PC's that covered up to 85% of variance were considered for the model creation. Five Principal Components and the categorical variables were used as the independent variables in the Logistic Regression. The AUC and the performance metrics were also considered.
- Developing on the above model, another model with only the significant Variables and PC's was used for making another Logistic Regression model.
- Factor Analysis: Factor Analysis was done to determine the important Factors amongst all the variables considered. The analysis yielded three important Factors: Factor 1 : Opponent Shots and Opponent Shots on Target, Factor 2: Total Shots ,Factor 3:Total shots on Target. The factors along with the categorical variables were used for the Logistic Regression.



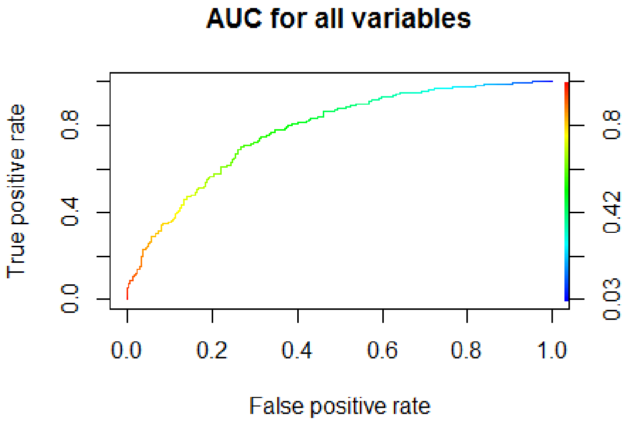
From the Scree plot, 3 is the ideal number of factors that must be considered for the Analysis.

VIF test was performed in the first scenario itself, observing that there is no multicollinearity.

The following table summarizes the different performance measures for the scenarios-

Model	Training Perf	Testing Perf	AIC	AUC
All variables	75.6	73.3	874.43	0.809
Significant Variables	71.6	72.9	931.04	0.78
With PCA	70.6	70.4	948.83	0.753
Significant Variables with PCA	70.5	69.7	951.66	0.747
Factor Analysis	43	47	1120.7	0.47

For the In-game model too we select the model with all the all Variables.



- Other tests were also considered:
  Durbin-Watson: no correlation between residuals for all the models.
  Over-dispersion: close to 1 for all the models

- Different seeds were considered, and we observed that our Final model is the most robust to different sample data.

- Interactions were considered, and we observed that there were not any significant ones.

Significant predictors – TST,OST,OF,TC,OC,Y,O_R.

Pre-Game Model :

Modelling for the response “Target”, considering of only predictors that influence the game before the kickoff. The variables considered here are: Type of game: Home/Away and Type of Opponent: Top tier, Mid-table or Relegation threatened.

The Logistic Regression model is built on the data of the six teams under consideration from the season 2012/13 to the current season.

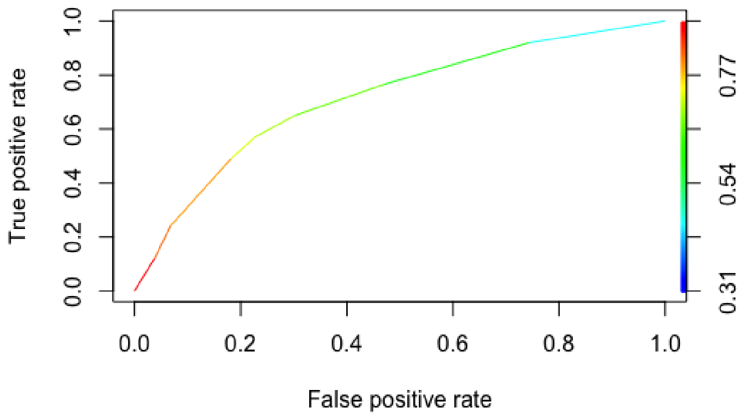
For the analysis, the data was divided into training and test for the measurement of performance.

The training data sample contained 75% of the data and test contained the other 25%.

Since the model is built only on two variables, the accuracy for the particular model is not very high.

For the given Logistic Regression model for the Pregame variables:

- AIC: 1023.2
- Training Performance:66.7
- Test Performance:67.4.
- AUC: 0.714. The curve for the same is below:



The pregame model was aimed to give us a little bit more insight on the betting odds that we had in that data, perhaps improve it. On searching about how the betting odds are calculated, we saw that a lot of factors play a role, for example recent performances of individual players, team performance in the last Home/Away match and even on how much bets are expected on that match day.

Considering the lack of such information in our dataset, the low accuracy is justified, and we shall not be using this model to predict win/loss or analyze the betting odds.

This can perhaps be a future endeavor, beyond this project.

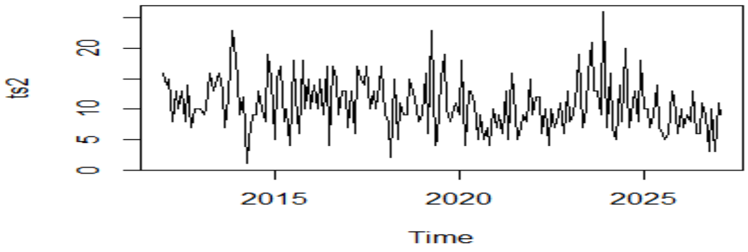
Considering the In-game, Pre-game and the Mixed Game models, we see that the AIC is the least for the Mixed Game model where only the significant variables are considered This is the model that will be used for predicting the final standings.

For all the six teams we forecast the values of all the numeric variables for the upcoming six games in order to determine the standing at the end of the season.

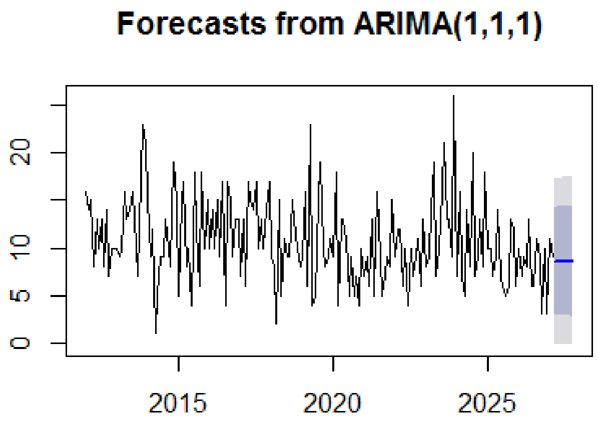
Time Series Model:

In order to generate the time series model, we divide the data for each of the teams individually thus resulting in six datasets for the six teams under consideration. In order to spread the games across equal intervals, we destroyed the conventional time-frame as it did not matter. Each game was considered as one point on the time series. We exclusively forecasted for only those variables that were significant in our logistic regression model. Below is a Time series and a forecast Plot:

Time series Plot:



Forecasting Plot :



Time Series and Forecasting Methodology:

- Each game is considered a one data point (1 month)
- Used ARIMA to predict values for next 6 points(games) for variable TS, total shots.

The forecasted data was plugged into the Mixed model significant variables Logistic Regression Model. Probabilities of win/no-win were calculated. Probability thresholds were calculated for all the teams to determine if it was a win or a no-win. Three points were awarded for a win and none to a no -win.

8. Results

For each team, probability of win was calculated for individual matches. Now the probability threshold that justified a win for a team was calculated separately based on the threshold that gave the highest accuracies between predicted results Vs actual results for the past matches.

Please refer the Rcode where the individual probability were calculated. Most teams had a probability cutt-off of about 0.5.

Now in case of a "Top 6" team playing against another, for example Liverpool Vs Arsenal, the probability of win was calculated once for Liverpool and also for Arsenal. If both teams satisfied their individual probability threshold 3 points were awarded to each of them. If only one of them satisfied their threshold, only one was awarded 3 points.

This method is not ideal but in our case we want to calculate positions relative to Manchester United and not their absolute points.

The standings of the six teams at the end at the end of the season based on the predictions from the Mixed model (Significant Variables) is:

Team	Current Points	Wins	Predicted	Total	Current Standing	Final Standing
Chelsea	75	4	12	87	1	**1**
Tottenham	68	5	15	83	2	**2**
City	61	6	18	79	4	**3**
Liverpool	63	5	15	78	3	**4**
United	57	5	15	72	5	**5**
Arsenal	54	3	9	63	6	**6**

The mixed model gave the best results based on which the standings were determined. The predictions that were made with the model were compared against actual results, the model was able to predict the results correctly for the Liverpool vs Watford and the Tottenham Hotspur vs Arsenal game

Additionally, we computed the standings based on the In- game and Pre-game models too. The results that were generated by the In-game model were similar to the Mixed model.

Standings Based on In-game model:

Team	Current Points	Wins	Predicted	Total	Current Standing	Final Standing
Chelsea	75	4	12	87	1	**1**
Tottenham	68	5	15	83	2	**2**
City	61	6	18	79	4	**3**
Liverpool	63	5	15	78	3	**4**
United	57	4	12	69	5	**5**
Arsenal	54	3	9	63	6	**6**

Standings based on Pre-game model:

Team	Current Points	Wins	Predicted	Total	Current Standing	Final Standing
Chelsea	75	4	12	87	1	**1**
Tottenham	68	4	12	80	2	**2**
City	61	6	18	79	4	**3**
Liverpool	63	5	15	78	3	**4**
United	57	4	12	69	5	**5**
Arsenal	54	5	15	69	6	**6**

9.Conclusion

From the model, Manchester United most likely seems to be finishing in the fifth position. This means that they would miss out on the Uefa Champions League spot for yet another season. Missing out on this prestigious tournament will cause a fair amount of Financial Damage.

Amongst all the models considered, mixed model which is a combination of Pre-game and In-game model gives us the best results although Im-game gave a pretty similar result.