**JOMMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY**

****

**A project report on**

# “Predicting Results of Football Matches using Machine Learning”

Submitted in fulfillment for the computer systems project

**by**

**Denis Kamau Thuo (sct212-0077/2017)**

**Under the guidance of**

**Mr.** **Isaiah Onando Mulang**

**(SCIT) School of Computing and Information Technology -** **JOMMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY**

|  |  |
| --- | --- |
| **TITLE** | **PAGE NO.** |
| **Chapter 1** |  |
| 1.1Title And Abstract | **3** |
| 1.2 Background and literature review | **8** |
| **Chapter 2** |  |
| 1.3 Introduction and Project Scope | 19 |
| 1.4 Hypothesis | **26** |
| 1.5 Problem Statement | **27** |
| 1.6 Objectives | 27 |
| **Chapter 3** |  |
| 2.1 System Design And Flow | 26 |
| 2.2 SYSTEM REQUIREMENTS SPECIFICATION | 26 |
| 2.3 Functional Requirements | 28 |
| 2.4 Non Functional Requirements | 29 |
| **Chapter 4** |  |
| 2.5 APPROACH AND DESIGN | 30 |
| 2.6 UML Diagrams | 31 |
| 2.7 System development methodology | 33 |
| **Chapter 5** |  |
| 3.1Plan for implementation | 41 |
| 3.2RESULTS AND INTERPRETATIONS | 45 |
| 3.3REGRESSION | 59 |
| **CHAPTER 6** |  |
| 3.4 conclusion | 64 |
| 3.5FUTURE SCOPE AND DEVELOPMENT  3.6REFERENCES AND APPENDIX | 65 |
| 3.7RANKING ALGORITHM  3.8 Github link | 71 |

**CHAPTER 1**

**1.1 TITLE AND ABSTRACT**

**ANALYSIS AND PREDICTION FOR ENGLISH PREMIER LEAGUE**

In today’s date data analysis is need for every data analytics to examine the sets of data to extract the useful information from it and to draw conclusion according to the information. Data analytics techniques and algorithms are more used by the commercial industries, which enables them to take precise business decisions.

The objective of this study is to analyze the gathered data of the three seasons (2015/16, 2016/17 and 2017/18) of English Premier League and to predict the outcome of the league for the upcoming season(2018/19).

The entire analysis conducted with the help of datasets and R tool. Datasets are files in CSV format. The analysis is done with the premier leagues data of the seasons 2015/16, 2016/17 and 2017/18*.*

The English Premier League data consisted the list of teams that played across the three seasons mentioned above. Along with the list of teams, data had various statistics, which were helpful in determining the performance of each team.

In Football; wins, goals and clean sheets play a major role for forecasting the outcome. Hence, analyzing wins, goals and clean sheets of each team can help in predicting the ranking of teams for the next season.

The main objective of this project is to give the team players information about how each statistics makes a difference to the game. In addition, give feedback of how the teams can improve their own performance in each game. And give have a better planning of how the match should be played overall by the whole team.

Framework Design involves keeping the prerequisites at the top of the priority list the framework details are made an interpretation of into a product representation. In this stage, the fashioner underlines on calculation, information structure, programming design and so on.

In recent years, the analytics is being used in the field of sports to predict and draw various insights. Due to the involvement of money, team spirit, city loyalty and a massive fan following, the outcome of matches is very important for all stake holders.

In this paper, the past seven year’s data of football containing the player’s details, match venue details, teams, ball to ball details, is taken and analyzed to draw various conclusions which help in the improvement of a player’s performance. Various other features like how the venue or toss decision has influenced the winning of the match in last seven years are also predicted.

Various machine learning and data extraction models are considered for prediction are Linear regression, Decision tree, K-means, Logistic Regression etc. The cross validation score and the accuracy are also calculated using various machine learning algorithms. Before prediction we have to explore and visualize the data because data explora.

The algorithm creates the forest with number of decision trees. The more the trees in the forest the more number of predictions and thus high accuracy. This method’s main advantage is that it’s very simple to explain the relationship between output variable and input. We can say that the presence of a risk factor increases the probability of a given outcome by a specific percentage.

We proposed a model of football match prediction by using the data of every match in English Premier League for the last one year and trained the algorithm based on that data. The algorithm learns from training data so that it forms different rules and pattern and based on that the algorithm makes a decision on some new data. The algorithm keeps on learning by using the feedback with every input

Prediction helps the managers and the clubs in making the right decision to win the leagues and the tournaments. The present study shows that earlier research on data mining systems to predict the results and evaluate the advantages and the disadvantages of each system. Prediction has been successfully applied in all sports. Although in many aspects, this application has been of very small limits.

It is very important to look into the applications of the machine learning in these instances and see if its application can give better results in the analysis. By making use of datasets that is more precise and machine learning, this research aims to offer a solution that will help to make predictions be more accurate and precise than the earlier systems.

The management should also be able take decisions during a match and to monitor a match in real time. For all these requirements, the prediction application should be able to provide ideas based upon the current situations. We have used multiple machine learning algorithms wherein Random Forest algorithm gave us the best accuracy

Prediction is the heart of remarkable disciplines in science and that is the reason why philosophy of prediction is employed in many companies. Machine learning which an area of intelligent systems is will be used in this report to provide solutions to the problems in aspects of predicting the results of various

sports.

Several times in the history of soccer, the world was shocked by results that were considered previously as impossible. The most remarkable recent example is about Leicester City’s fairytale story in the 2015/16 season of the Premier League, which is England’s highest soccer division. Leicester City succeeded in winning the Premier League, while before the season started the odds for Leicester City to win the league were one against 5000.

Until then, Leicester City had never won a top-flight title in its 132-year existence. The most unimaginable about this fairytale is that Leicester City was at the bottom of the Premier League in April 2015, nevertheless managed to avoid relegation to the second division in their last matches, to subsequently become champions of England around a year later.

Incidentally, all this did not happen as a result of large financial investments by a wealthy club owner, but through smart transfer policy and by putting team interest in the first place. Moreover, the manager of Leicester City in the 2015/16 season was the Italian Claudio Ranieri, who was sacked as manager of the Greece national team half a year earlier after losing against the Faroe Islands. Speaking of the Greece national team, they also managed to realize something which was held as very unlikely to happen.

They won the European Championship in 2004 as outsiders after beating Portugal, which hosted the tournament, in the final in Lisbon. Even though no other soccer story in terms of impact is comparable to the Leicester City fairytale, a European Championship winner whose odds in advance were estimated at one against 150, can rightly be seen as an unbelievable performance.

These fascinating stories show that soccer is a very unpredictable sport. In contrast to a single match that went differently than predicted in advance, the unique achievements of Leicester City and the Greece national team were examples of a series of matches that, against all the odds, turned out to be a success for the underdog. Even though predicting soccer match results is not an easy task, provide ample motivation.

Contradictory to Arsène Wenger’s judgment, the goal of this study is to be able to predict the results of soccer tornaments correctly by means of data science. The application of data science is becoming more and more popular in soccer. Many soccer clubs realize that when data is used properly, this could bring them steps ahead of their competitors. Data analytics and the application of machine learning techniques is interesting for different tasks in soccer, such as monitoring the physical health of players (Rossi et al., 2017), scouting players based on their predicted potential

Keywords: Random Forest algorithm performing regression

and Database

**1.2 BACKGROUND AND LITERATURES REVIEW**

The Premier League is considered as the premium league in the game of Football. It is the most followed football league in the world. It also gains huge attention from the broadcasters of sports from around the world. The teams playing in the league are under constant pressure for performing well and securing a respectable spot in the table.

Predictive analysis is the process of analyzing data using automated statistical processes and summarizing results into useful information.

Analysis of different aspects of the game which are responsible for determining the outcome of the match are helpful for team management which can then take measures to improve in those aspects in which they lack. Improving those can help strengthen the team.

Premier League also has a strong hold in betting market. Huge number of bets are placed to determine the winning team. Analysis on various aspects of the game can help the betters to place correct bets.

In the whole world football is the most famous and popular sport. Among all the sport, football prediction is the most widely researched. The main use of our project is to create software with which we can predict the upcoming match results based on many factors. We will give the teams name as the input and it will show which team will win the match.

Not only who wins the match, the game lineup also will be also predicted. In the previous systems they used very less parameters which resulted in less accuracy of the prediction. We will be using many extra factors like player details, half-time score, full-time score and many more will be considered and the accuracy of the prediction increases

(Aditya Srinivas Timmaraju, Aditya Palnitkar, VikeshKhanna,2014) estimated that 4.7 billion people watched 2010-11 season. (FarzinOwramipur, ParinazEskandarain, and Faezeh Sadat Mozneb,2015) used Bayesian Network to predict the score of Spanish Team, Barcelona.

They found the model probability and relationship among the given domain. They collected information from valid websites that offered the statistics of the football. They also used two factors for the match prediction psychological and non- psychological factors through which they can predict the final result.

### 1.2.1 Logistic regression concepts

(Igiri, Chinwe Peace, Nwachukwu, Enoch Okechukwu, 2014) used Artificial Neural Network (ANN)and logistic regression (LR) techniques with Rapid Miner as data mining tool which result of 85% and 93% prediction accuracy respectively. They compared the existing system of prediction with their system and found that their system were twice more accurate than the current existing system. They took certain factors that affects the result of the match like Home advantage effect on team’s performance, the effect of injuries of key players on team performance, effect of external cup on league performance. They have seen these factors and related that the effect of the certain factors can also be seen in the team.

(YueWengMak,2013) used a three multilayer perceptron concept and took factors like last 5 matches of each team and their last 3 encounters between the team and added home advantage and ranking inputs to predict the result which predicted that the more the consideration of the factors the more likely the prediction will be correct.

### 1.2.2Bayesian network concepts

1. Joseph, N.E. Fenton, M. Neil, 2006) used Bayesian network approach to predict the result of Spurs only. They predicted containing certain factors of Bayesian network like MC4 Learner, Naïve Bayesian Network, Hugin Bayesian Network, Expert Bayesian network and KNN (K Nearest Neighbour) algorithm.

They used the process of machine learning with two tangible benefits, understanding and prediction. The MC4 learner identifies those attributes, which have the largest effect on the outcome of the game. It shows their relationships to each other in terms of their effect on the outcome of the game.

This is a very simplified model of the game itself. The naïve Bayesian learner doesn’t construct a model as such its model is predefined. The learning process for the naïve Bayesian learner is then simply one of discovering the relative strength, and polarity, of the effect of each attribute with respect to the result. KNN does not construct a model as such it simply uses the existing data and provides a likeness comparison with any test data.

Thus KNN doesn't significantly enhance our understanding. The expert constructed Bayesian Network represents the knowledge of the expert, that is, it is a model is the expert’s belief of the interrelationships between the attributes and their relative importance.

(Jeffrey Alan Logan Snyder, 2013) reviewed past and current research efforts in soccer prediction, categorizing their approaches and conclusion. It treats a match as a result- producing black box, ignoring the noisy, but beautifully complex processes that contribute to each shot and goal. As XY data makes its way into the hands of researches, more detailed models of in-game processes may be built.

This will in turn open up new problems and avenues for analysis, hopefully leading to a deeper understanding of the game itself. It also has begun the process of determining the relative importance of these types of data in prediction and analysis.

They also have presented an approximately optimal betting strategy for use in betting simultaneously on multiple games with mutually-exclusive outcomes, which performs substantially better than other strategies used in academic betting simulations.

### 1.2.3 Probabilistic concepts

(AdityaSrinivasTimmaraju, Aditya Palnitkar, VikeshKhanna,2014)selected the important characteristics for a feature they are Incorporate a notion of the competing nature of the problem, Be reflective of the recent form of a particular team and Manifest the home advantage factor. They took three approaches and they are:-

1. Approach 1:-The first approach they took was using Multinomial Logistic Regression. They considered the performance metrics derived from the current match, rather than taking the average over the last “k” matches. During testing, they predicted the match outcome of team A vs team B, where they arrived at the feature vector using KPP.
2. Approach 2 :-In second approach, they trained the same way and more precisely, in the training phase too, instead of using the feature vector as the performance metric vector corresponding to the current match, they used KPP. This meant that the trained parameters now inform their beliefs about the result of a match based on the performance in last “k” matches. In this approach, they also used TGKPP instead of KPP.
3. Approach 3 :-All approaches they tried will find a global set of parameters, which were independent of the competing teams. So, the given past “k” performances of the team playing home and the team playing away, our model was agnostic to the identity of the actual teams playing. They felt they were missing some team-specific trends using this method. So, in this approach, they trained different models for the different teams. However, this approach placed a limitation on the data they could use to test/train our model. They could no longer combine data from two different seasons due to the form of the team varying between seasons, and major players being traded between teams. So, due to the limited data, and the increased noise induced by increasing the granularity in the model, they ended up getting a lower accuracy (an average of 47%).

With these approaches, they proposed a metric, which is computed as the geometric mean of the predicted probabilities of actual outcomes, which is called PLS.

They obtain a PLS of 0.357 for EPL which had the best value of PLS as 0.36007. (AlbinaYezus, 2014) divided the work in the steps as choosing match set that is to be analyzed, deciding on key features, data extraction, testing various machine learning algorithms and improving the implemented algorithms. He found out that it is possible to find a classifier that predicts the outcome of soccer matches with the precision of more than 60%.

### 1.2.4 Markov chain and Monte carlo concept

(Havard Rue and Oyvind Salvesen, 1999) used MCMC (Markov Chain and Monte Carlo Methods) in which they produced as an irreducible aperiodic transition kernel. They presented that this approach seems superior to the earlier attempts to model the football games as it allows for coherent inference of the properties between the teams easily account for the joint uncertainty in the variables which in important in prediction allows for doing various interesting retrospective analysis of a season and finally provides a framework, where is it easy to change parts or parameterization in the model.

They also have improved on the data, parameter estimation, goal model and home field advantage, which make that the point of the prediction highly rises than before with these parameters. (DouweBuursma, 2013) found out the selection of the feature set led to the important insights and the performance of the classifier kept improving the bigger set of recent matches which showed the optimal number of matches to lie around 10, 20 and performance keeps improving.

The history of opponents of home teams does not seem to play as important a role as the history of opponents of the away team. He trained a number of data, which made the system fare, which makes that the analysis can use more data. The feature set like Match History, Classifiers helps to get the correct model and it also helps to make the data more useful for the training set.

The Classifiers classifies all matches as home wins, draws or away wins, depending on the features belonging to that match. He used classifiers like ClassificationViaRegression, MultiClassClassifier, Rotation Forest, LogitBoost, BayesNet, NaiveBayes and Home Wins. He concluded that the football matches would always be very hard to predict. (Francisco Louzadal, Adriano

K. Suzuki and Luis E.B. Salasar, 2014) & (Brijesh Kumar Bhardwaj, Saurabh Pal, 2011) used the data mining concept for the prediction over the improvement using classification. In data mining process, they first prepared the data over the size of 300 and they selected only the field that was necessary for data mining.

The predictor and variables were derived from the database and they used Bayesian Classification for the implementation of the mining model. (GianlucaBaio, Marta A. Blangiardo, 2010) used Bayesian network and poisson distribution for the analysis of the model where the parameters defined was home advantage, the scoring intensity.

They have showed that the home team have greater potential to win and the team with the low points are to concede more goals at home and lose. The estimation of the team was poor when bivariate poisson was used and the models parameters perform better than bivariate poisson.

. Some of the important conclusions were made through those are listed below.

* + 1. **Kalpdrum Passi and Niravkumar Pandey** discussed about the prediction accuracy in terms of runs scored by batsman and the no. of wickets taken by the bowler in each team [1].
    2. **P. Wickramasinghe** proposed a methodology to predict the performance of batsman for the previous five years using hierarchial linear model [2].
    3. **R.P.Schumaker** et. al, discussed about different statistical simulations used in predictive modeling for different sports [3].
    4. **John McCullagh** implemented neural networks and datamining techniques to identify the talent and also for the selection of players based on the talent in Australian Football League[4].
    5. **Bunker et. al**, proposed a novel sport prediction framework to solve specific challenges and predict sports results [5].
    6. **Ramon Diaz-Uriarte et. al**, investigated the use of random forest for classification of microarray data and proposed a new method of gene selection in classification problem based on random forest [6].
    7. **Rabindra Lamsal and Ayesha Choudhary**, proposed a solution to calculate the weightage of a team based on the player’s past performance of IPL using linear regression [7].
    8. **Akhil Nimmagadda et. Al**, proposed a model using Multiple Variable Linear Regression and Logistic regression to predict the final league standings and also the winner of the match using Random Forest algorithm.
    9. **Ujwal U J et. Al**, predicted the outcome of the given cricket match by analyzing previous cricket matches using Google Prediction API [9].
    10. **Rameshwari Lokhande and P.M.Chawan** came up with live cricket score predicton using linear regression and Naïve Bayes classifier [10].
    11. **Abhishek Naik et. Al**, proposed a new model used matrix factorization technique to analyze and predict the winner in ODI cricket match [11].
    12. **Esha Goel and Er. Abhilasha** discussed the improvements in Random Forest Algorithmand described the usage in various fields like agriculture, astronomy, medicine, etc. [12].
    13. **Amit Dhurandhar and Alin Dobra** proposed a new methodology for analysing the error of classifiers and model selection measures to analyse the decision tree algorithm [13].
    14. **H. Yusuff et. Al**, performed logistic regression using mammograms to find the accuracy with valid samples [14]. A. Joseph, N. F. (2006, April 6). Predicting football results using Bayesian nets and other

machine learning techniques. p. 10.

**1.2.5 REFERENCES**

.Aditya Srinivas Timmaraju, A. P. (2013). *Game ON! Predicting English Premier League Match Outcomes.*

Blundell, J. D. (2014). *Numerical Algorithms for Predicting Sports Results.* School of Computing, Faculty of Engineering.

Brijesh Kumar Bhardwaj, S. P. (April, 2011 ). Data Mining: A prediction for performance improvement using classification . *(IJCSIS) International Journal of Computer Science and Information Security,* .

Buursma, D. (2013). *Predicting sports events from past results towards effective betting on football matches.* Netherland: University of Twente.

Byungho Min, C. C. *A Compound Approach for Football Result Prediction.* Seoul, Korea: School of Computer Science and Engineering.

Farzin Owramipur, P. E. (October, 2013). Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. *International Journal of Computer Theory and Engineering, Vol. 5, No. 5* .

Francisco Louzada, A. K. (2014). Predicting Match Outcomes in the English Premier League:Which Will Be the Final Rank? *Journal of Data Science* , 235 - 254.

Gianluca Baio, M. A. (2013). *Bayesian hierarchical model for the prediction of football results.* London: University College London.

Håvard Rue, Ø. S. (1997). *Prediction and Retrospective analaysis of Soccer Matches in a league.* Trondheim, Norway: Norges Teknisk-Naturvitenskapelige Universitet.

Igiri, C. P. (December, 2014). An Improved Prediction System for Football a Match Result . *IOSR Journal of Engineering (IOSRJEN)* , 12 - 20.

Johanne Birgitte Linde, M. L. (June, 2014). *Predicting Outcomes of Association Football Matches Based on Individual Players' Performance.* Norwegian University of Science and Technology.

LANGSETH, H. (2014). *Beating the bookie: A look at statistical models for prediction of football matches.* Trondheim, Norway: Department of Computer and Information Science, Norwegian University of Science and Technology.

Mak, Y. W. (2013). *Prediction on Soccer Matches using MultiLayer Perceptron.*

Snyder, J. A. (2013). *What Actually Wins Soccer Matches: Prediction of the 2011-2012 Premier League for Fun and Profit.*

Stylianos Kampakis, A. A. (n.d.). Using Twitter to predict football outcomes.

* + 1. Yezus, A. (2014). *Predicting outcome of soccer matches using machine learning.* Saint- Petersburg State University, Mathematics and Mechanics Faculty

**CHAPTER 2**

**1.3 PROJECT INTRODUCTION AND SCOPE**

Machine Learning is a branch of Artificial Intelligence that aims at solving real-life engineering problems. This technique requires no programming, whereas it depends on only data learning where the machine learns from pre-existing data and predicts the result accordingly.

Machine Learning methods have benefit of using decision trees, heuristic learning, knowledge acquisition, and mathematical models. It thus provides controllability, observability, stability and effectiveness.

Football is played in all countries around the world. There are many domestic and international tournaments being held in many countries. The football game has various forms such as Friendly Matches, Internationals, and Internationals one day, etc. Premier League is also one of them, and has great popularity among them. The league was conducted annually. There are twenty teams which representing cities which are chosen from performance. These teams compete against each other for the trophy.

The whole tournament depends on the luck for the team, player’s performance and lot more parameters that will be taken in to the consideration. The match that is played before the day is also will make a change in the prediction. The stakeholders are much more benefited due to the huge popularity and the huge presence of people at the venue. The accuracy of a data depends on the size of the data we take for analysing and the records that are taken for predicting the outcome.

Football is a game played between two teams comprising of 11 players in each team. The result is either a win, loss or a tie.. Moreover, this game is also extremely unpredictable because at every stage of the game the momentum shifts to one of the teams between the two.

A lot of times the result gets decided on the last ball of the match where the game gets really close. Considering all these unpredictable scenarios of this unpredictable game, there is a huge interest among the spectators to do some prediction either at the start of the game or during the game. Many spectators also play betting games to win money.

**1.4 HYPOTHESIS**

**1.4.1Research Questions**

In this study, we addressed the following problem statement: To what extent can we improve upon the prediction results of Ulmer et al. (2013)? In order to formulate an answer to this problem statement, we formulated four research questions, which will be discussed in the next paragraphs. 9 To start with, being able to validate our prediction results relative to the study of Ulmer et al. (2013), we first had to replicate their prediction models. Hence, the first research question reads as follows:

RQ 1: To what extent can we replicate the study of Ulmer et al. (2013)?

For this investigation, a dataset containing soccer match statistics is used to predict the FTR of soccer matches. This dataset, which we call the original dataset, consisted of the number of shots, shots on target, fouls, corners, yellow cards, and red cards that occurred in a match for both the home and away team. The second challenge of this study was to determine which of these match statistics of the original dataset, which we call original features, are most informative regarding match result prediction. In order to identify which original features are most predictive regarding match results, the second research question reads as follows:

RQ 2: Which original features are most informative for predicting tournament results?

Based on the results of the experiment to answer the second research question, feature selection is applied. Subsequently, the selected original features are used to create new features through feature extraction, which we call engineered features. These engineered features, which are gathered into our engineered dataset, are constructed by relying on ideas from related studies (Baboota & Kaur, 2018; Ulmer et al., 2013) and based on our understanding of the problem domain. We attempted to improve upon the study of Ulmer et al. (2013) by means of the newly created engineered features which are expected to be more informative than the features they used. Therefore, the third research question reads as follows:

RQ 3: To what extent does the addition of engineered features contribute to the prediction of match results?

Apart from the fact that the dataset that is used in this study contained a larger variety of original features than the dataset of Ulmer et al. (2013), we also had access to data of more soccer matches. The use of more soccer matches could lead to an increase in the predictive ability of the model. Therefore, the effect of extending the size of the dataset is investigated. Hence, the last research question reads as follows:

RQ 4: To what extent does the extension of the dataset contribute to the prediction of match results?

Structure The remainder of this thesis is organized as follows. The first chapter provides an overview of related work which already has been performed. After the Related Work chapter, the Method chapter follows, which describes the dataset and experimental setup of this study. The performance of the model is presented in the Results chapter. The last chapter of this thesis consists of the Discussion section, the Conclusion section, 10 and the Future Work section. In the Discussion, our findings are used to refer back to the goal of the study and used to put this study in perspective regarding other literature. In the Conclusion section, the research questions will be answered. Finally, we suggest future work based on the implications of this study.

There are 6 major factors which contribute to the result of the match, they are:-

1. Number of goals scored.

2. Number of clean sheets kept.

3. Shots on target made by the team.

4. Shots taken from inside the box.

5. Goals conceded.

6. Saves made by the goalkeeper of the team.

Ranking Algorithm helps to determine where a team stands comparative to other teams in the above mentioned departments. Use of Ranking Algorithm can also provide the team with their ranking in all these fields individually, so that they can figure out the departments which needs improvement and work on the same.

The final ranking of the teams in the table is determined on number of wins. Generally, team with highest number of wins ends up winning the league. “*Goals*” and “*clean sheets”* are the two factors on which “*wins”* is dependent.

Both are directly proportional to wins. I.e. the more the number of goals scored and clean sheets kept, more are the chances of winning.

Furthermore, for determining “*goals”*, “*shots on target”* & “*shots from inside box”* are the two main factors; and for determining “*clean sheets”, “goals conceded”* & *“saves”* are the factors. Analysis of all these factors leads us to predict the overall rank of the team, which also can be considered as the finishing position in the table at the end of the season.

Football is one of the most famous games for the people worldwide. It is a game played by two teams, who needs to hit the ball on its opponent team’s goal post. The way the football is getting worldwide and the fans are also crazy about it. Audience wants to know who wins the game and who loses. Prediction of such football game is hence popular. Also many people do the betting for the favourite team. They simply see their luck keeping the money on the team and they actually don’t know who will win and who will lose. But, it’s actually difficult to predict the result of the game. Many of the football experts also finds difficult to predict the score line or result of the game. The league that is famous among all the football lovers is English Premier League where 20 English teams compete with each other to win the league. (Reilly, Thomas; Gilbourne, 2003)

The craziness is increasing day by day and each time the teams focuses on the league. The point system of the game is simple whichever team wins get 3 points and the team that loses get 0 point and if the game is draw then each team will share one point each.

The team with highest point at the last wins the league. Each time different winner is seen in every year, but the most dominant team is Manchester United with 13 Premier League’s trophies. The bottom three teams are relegated and replaced by other teams from lower leagues who perform better.

Each team plays every other team twice, once at home and once away. Premier League is one of the most-watched football (soccer) leagues today within the world which is the league between the states of England clubs. It is the league between 20 teams where bottom 3 teams gets relegated to another sub division league of England.

Prediction makes the game more really exciting and focused to the supporting team for the wait of the chances of the team to win the game. The prediction will be focused within the teams of English Premier League only. Premier League Game Result Prediction predicts the result of

the game between two teams performed in every weekend. (F. Halicioglu, 2005) predicted Euro 2000 winner and conformed that the prediction is possible and hence the prediction was started.

## 1.5 Problem Statement

Premier League fans are more excited to know the results of the game. They are more enthusiastic to know the result before the game. However, the main problem is to make the accurate result of the prediction between the teams.

## 1.6 Objectives

The main objective of the Premier League Game Result Prediction is:

1. To predict the result of each game of the team of the premier league.
2. To implement Back Propagation algorithm for training and testing the model.

## 1.6.1 Limitation

The limitation of the Premier League Game Result Prediction is:

1. It can't predict the result of cups matches like Spanish League, French League, German League, World Cup, Euro Cup, etc.
2. It can't predict the score line.
3. It doesn't contains parameters like individual players details, key players, player transfer details.

**CHAPTER 3**

**2.1 SYSTEM DESIGN AND FLOW**

**2.2 SYSTEM REQUIREMENT SPECIFICATION**

A System Requirement Specification (SRS) is basically an organization’s understanding of a customer or potential client’s system requirements and dependencies at a particular point prior to any actual design or development work. The information gathered during the analysis is translated into a document that defines a set of requirements. It gives the brief description of the services that the system should provide and also the constraints under which, the system should operate.

Generally, SRS is a document that completely describes what the proposed software should do without describing how the software will do it. It’s a two-way insurance policy that assures that both the client and the organization understand the other’s requirements from that perspective at a given point in time.

SRS document itself states in precise and explicit language those functions and capabilities a software system (i.e., a software application, an ecommerce website and so on) must provide, as well as states any required constraints by which the system must abide.

SRS also functions as a blueprint for completing a project with as little cost growth as possible. SRS is often referred to as the “parent” document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it.

Requirement is a condition or capability to which the system must conform. Requirement Management is a systematic approach towards eliciting, organizing and documenting the requirements of the system clearly along with the applicable attributes. The elusive difficulties of requirements are not always obvious and can come from any number of sources.

With all the literature review, the most common thing was choosing the correct parameter is the first way to get the prediction. The more the parameters, the more chances of getting the result of the prediction correct. The prediction that was difficult for the experts have made some easy task due to several prediction methods.

The parameters or factors such as home advantage, injuries of the players, cup game effect on league, team recent form and the head-to-head matches between the opponent need to be analyzed which adversely affect the result of the match. (Igiri, Chinwe Peace, Nwachukwu, Enoch Okechukwu, 2014) used the Hidden Markov Process Model and Ordered Probit Regression model with only three parameters like home advantage, injury of the key players and cup game effect on league. I

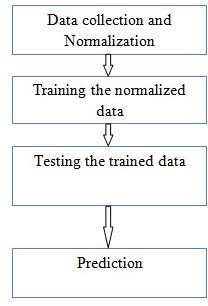
will be using the same models but with more parameters like the team recent form and the head-to-head matches between the opponents which might affect the accuracy of the prediction then (Igiri, Chinwe Peace, Nwachukwu, Enoch Okechukwu, 2014). With lots of data, the data-mining tool is used to extract the information. Artificial Neural Network (ANN) and Regression techniques are two data-mining techniques that will be used. I will first collect the previous results of the matches with every history of the team and the teams they played with.

Then, from the collected data will extract the features such as Home and Away Goal Difference, Points, Attack and Defense Skills, which is not needed for the techniques. Then a collective database will be built to collect all the necessary data and will be kept in MS Excel spreadsheet. Using an algorithm will use the parameters to adjust the training algorithms to make it more useful.

The data-mining tool, which holds the better results and improves the learning rate, also finds the optimal values and an improved model will be built. The two models Artificial Neural Network (ANN) and Logistic Regression (LR) are used to make the improved models for the system. So, the record of the match and the result of the match will be predicted by the use of ANN and LR model. The accuracy was 75.04% according to (Igiri, Chinwe Peace, Nwachukwu, Enoch Okechukwu, 2014)but after the

addition of the parameters that helps to estimates the model more might increase the accuracy by 85%.

So, with the known of the several factors, I came to use different input variables, output variables, algorithms for the process of prediction.



- Method for prediction

Figure 4 illustrates the method that needs to be followed while making the prediction. First, the data is collected and it is being normalized by using sigmoid function. Then the normalized data is trained by using neural network and the trained data are being tested for the prediction accuracy. Then after the testing the prediction can be done.

Table - Functional and Non-functional requirement

|  |  |
| --- | --- |
| **Functional requirement** | **Non-functional requirement** |
| Predicts the result of the game between two teams. | Prediction is based on the basis of team name. |
| Compares the result of the data that are trained. | A result will be shown on which team will win, lose or draw the game. |

The basic functionality of this application is that it predicts the result of the game between two teams. The results are then compared with the data that are being trained which is described in Table 1. The non-functional requirement includes displaying of the input field and a predict result button, then a result will be shown comparing the data with the trained data

**2.3 Non Functional Requirement**

Non-functional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviors. They may relate to emergent system properties such as reliability, response time and store occupancy. Non-functional requirements arise through the user needs, because of budget constraints, organizational policies, the need for interoperability with other software and hardware systems or because of external factors such as :-

* Performance Requirements
* Design Requirements
* Security Constraints
* Basic Operational Requirements

**2.4 Functional Requirements**

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality.

Following are the functional requirements on the system:

1. The whole process can be handled at minimal human interaction with android and web both.

2. The application automatically receives the captured data from the csv file.

3. The user can call emergency, map location and ECG graph on demand

4. The system gives a warning message.

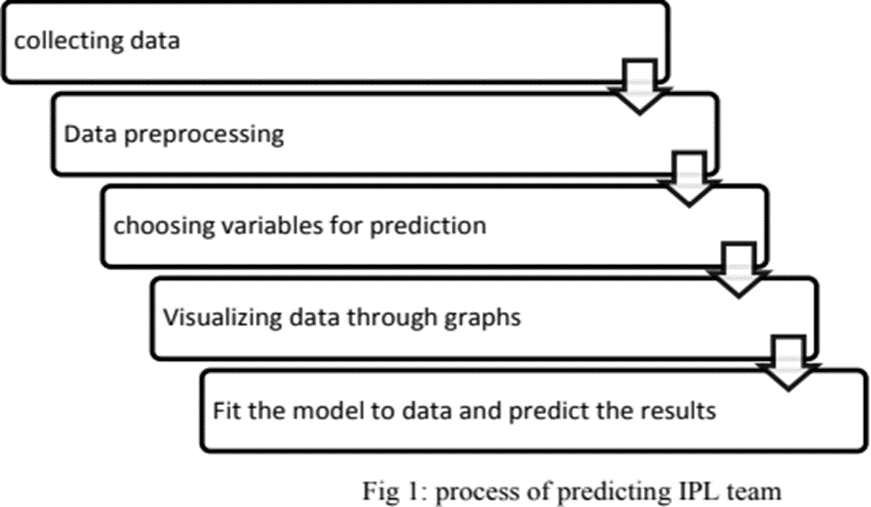
### 2.4.1 Product Requirements

* + - * **Platform independency:** A progressive web app will be developed and deployed so that users with a smartphone or a computer can access the voting site to cast their vote.
      * **Ease of use:** The progressive web app provides an interface which is easy to use and eliminates the need for the voter to go to a voting booth.
      * **Modularity:** The complete product is broken up into modules and well-defined interfaces developed to explore the benefit of flexibility of the product.
      * **Robustness:** This software being developed in such a way that the overall performance is optimized, and the user can expect the results within a limited time with utmost relevancy and correctness.

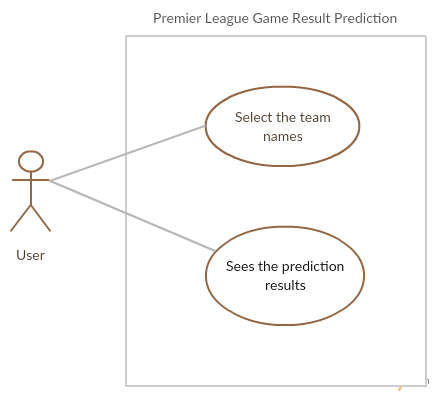
**CHAPTER 5**

**2.5 APPROACH AND DESIGN**

The below figure explains the approach we have taken into building the predictive model using machine learning algorithms.



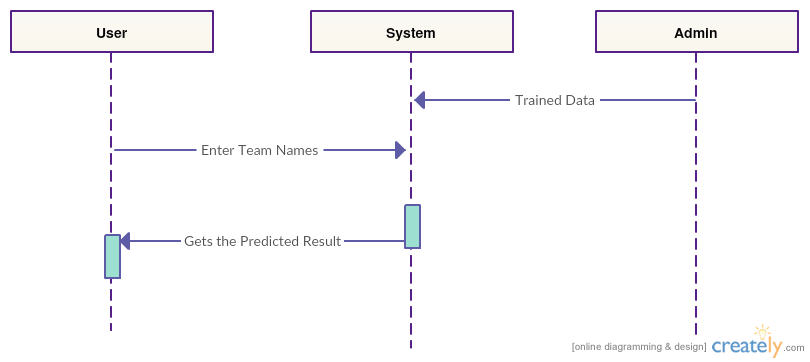
**2.6 UML DIAGRAMS FOR THE PREDICTION SYSTEM**

**2.6.1use case diagram**

Use case diagram of Premier League Game Result Prediction

The Figure describes about the use case diagram of the system. The figure illustrates that user can first select the team names and sees the prediction results.

### 2.6.2 Sequence diagram

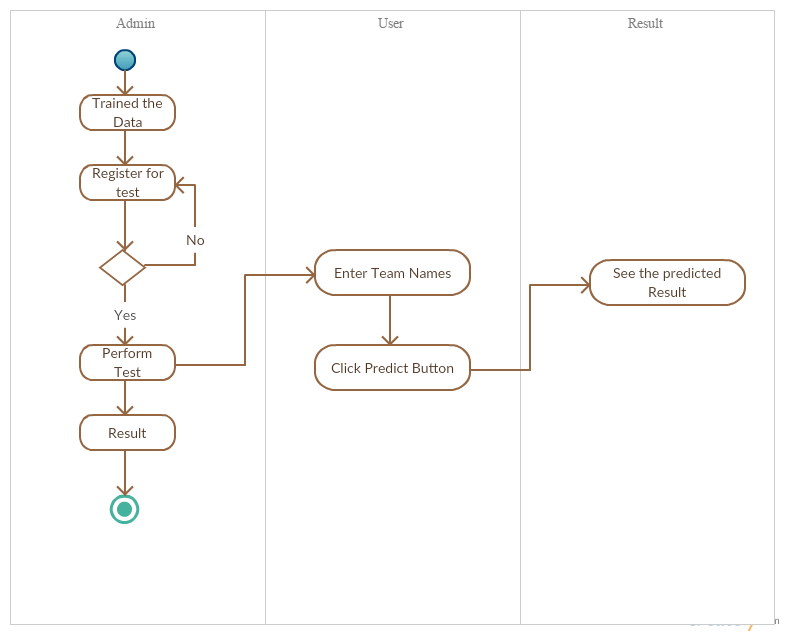


Sequence diagram of Premier League Game Result Prediction

The figure illustrates the sequence diagram of the application Premier League Game Result Prediction. It gives the sequence of the application from the user to the system and vice- versa.

When the user chooses the team names from the available teams with trained set of data, it is interpreted and the system returns the predicted result to the user. Further, displayed on the application where admin gives trained and tested result.

### 2.6.3 Event diagram

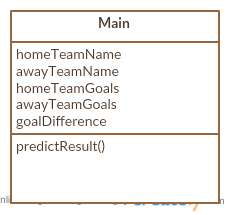


Event diagram of Premier League Game Result Prediction

The figure illustrates the event diagram of the Premier League Game Result Prediction which displays events and the process that are interrelated to each other. First, admin trains the data which is registered for the testing purpose and then test is performed.

After the test is performed, a user-friendly User Interface is built and the user enters the team names into the input fields and clicks the predict button from which the result can be seen by the user.

### 2.6.4 Class diagram



Class diagram of Premier League Game Result Prediction

The figure shows the class diagram of this application which consists of a Main Class only where a function named as predict Result() represents the prediction of the result. The parameters or variables include Home Team Name, Away Team Name, Home Team Goals, Away Team Goals and Goal Differences.

## 2.7 System development methodology

System development method is a process through which a product will get completed or a product gets rid from any problem. Software development process is described as a number of phases, procedures and steps that gives the complete software. It follows series of steps which is used for product progress. The development method followed in this project is waterfall model.

### 2.7.1 Model phases

The waterfall model is a successive programming improvement process, in which advance is seen as streaming relentlessly downwards (like a waterfall) through the periods of Requirement start, Analysis, Design, Implementation, Testing and upkeep.

**Prerequisite Analysis:** This stage is worried about gathering of necessity of the framework. This procedure includes producing record and necessity survey.

**Framework Design:** Keeping the prerequisites at the top of the priority list the framework details made an interpretation of into a product representation. In this stage the fashioner underlines on calculation, information structure, programming design and so on.

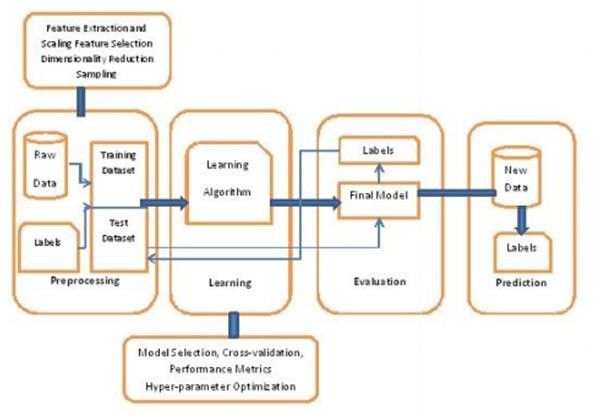
**Coding:** In this stage developer begins his coding with a specific end goal to give a full portray of item. At the end of the day framework particulars are just changed over into machine coherent register code.

**Usage:** The execution stage includes the genuine coding or programming of the product. The yield of this stage is regularly the library, executables, client manuals and extra programming documentation.

**Testing:** In this stage, all projects (models) are coordinated and tried to guarantee that the complete framework meets the product prerequisites. The testing is worried with check and approval.

**Support:** The upkeep stage is the longest stage in which the product is upgraded to satisfy the changing client need, adjust to suit change in the outside environment, right mistakes and oversights beforehand undetected in the testing stage, improve the proficiency of the product.

### 2.7.2 System Architecture



**Dataset**: The datasets used are of xls and csv formats.

**Dataset origin**: [www.kaggle.com](http://www.kaggle.com/)

**R Studio and R Language:**

In this system the Software used is RStudio:

As RStudio is a free and open-source integrated development environment (IDE)

for R, a programming language for statistical computing and graphics

In this System the Technology used is R:

R is an open-source language and environment for statistical computing and data visualization, supporting data manipulation and transformations, as well as sophisticated graphical displays.

**Packages Used:**

**ggplot2**:

ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.

**plotly**:

Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration.

**Dplyr**:

Dplyr is the next iteration of plyr, focused on tools for working with data frames

(hence the d in the name).

**Shiny**:

Shiny is a new package from RStudio that makes it incredibly easy to build interactive web applications with R.

**Shiny dashboard**:

Create dashboards with Shiny. This package provides a theme on top of Shiny, making it easy to create attractive dashboards.

**RColorBrewer**:

Provides color schemes for maps (and other graphics) designed by Cynthia Brewer as described at [http://colorbrewer2.org.](http://colorbrewer2.org/)

**Session:**

Utility functions for interacting with R processes from external programs. This package includes functions to save and restore session information (including loaded packages, and attached data objects), as well as functions to evaluate strings containing R commands and return the printed results or an execution transcript.

**3.1 Plan of Implementation**

The project can be broken down into 7 main steps which are as follows:

1. Understand the dataset.

2. Clean the data.

3. Analyze the candidate columns to be Features.

4. Process the features as required by the model/algorithm.

5. Train the model/algorithm on training data.

6. Test the model/algorithm on testing data.

Premier League Game Result Prediction is the ability to predict the game result. The main purpose of implementing Premier League Game Result Prediction is to predict result of matches before the game. The past dataset needs to be trained and tested and with the trained dataset, we need to predict the result. In the implementation process, we have around 9000 data of which 80% are used for the training purpose whereas remaining 20% are used for the testing purpose.

Training done by using Back Propagation algorithm in octave application where the data are differentiated into Home Team, Away Team, Home Team Goals, Away Team Goals and Goal Differences in one file and the result is set in another file. The supervised learning is done so the test and trained result output must be given before processing the application.

Table - Sample trained data used for implementation



Table 9 shows the sample input data that are being trained. The data are kept are the input parameter for the training set.

Trained data result used for implementation

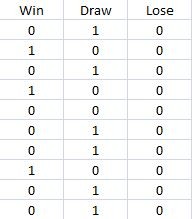


Table shows the output that is seen while data are trained. The data represents win, lose and draw. 010 represents draw, 100 represents win and 000 represents lose. Testing is done as per the result obtained from the training sets in octave application where the output pattern is already known by the application.

Test data used for implementation



Table shows the input data that are being tested. The data are kept are the input parameter for the testing purpose.

Test Data Result Used For Implementation

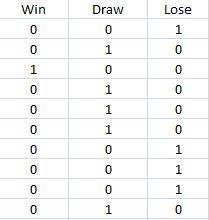


Table 12 shows the output that is seen while testing the data. The data represents win, lose and draw. 010 represents draw, 100 represents win and 000 represents lose.Testing is done as per the result obtained from training sets in octave where the output pattern is already known by the application.

e the model/algorithm for higher accuracy.

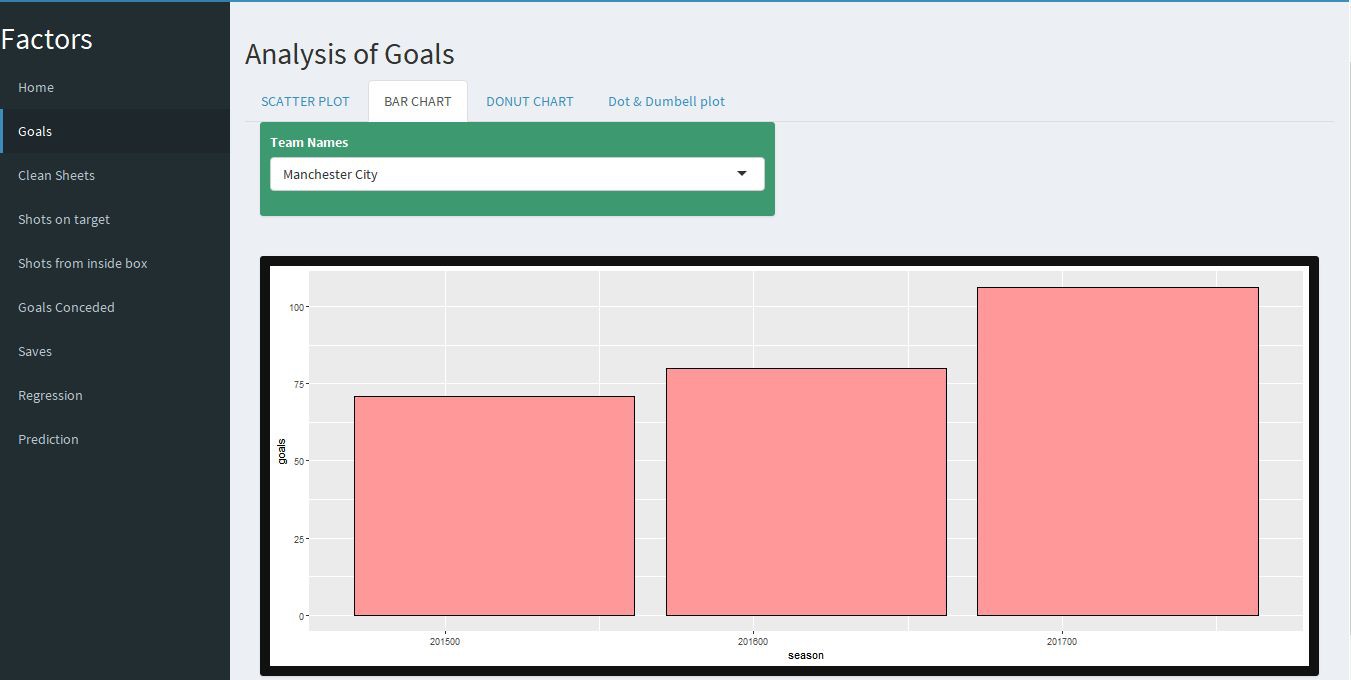
**3.2 RESULTS AND INTERPRETATIONS**

**A**. *FACTOR - GOALS*

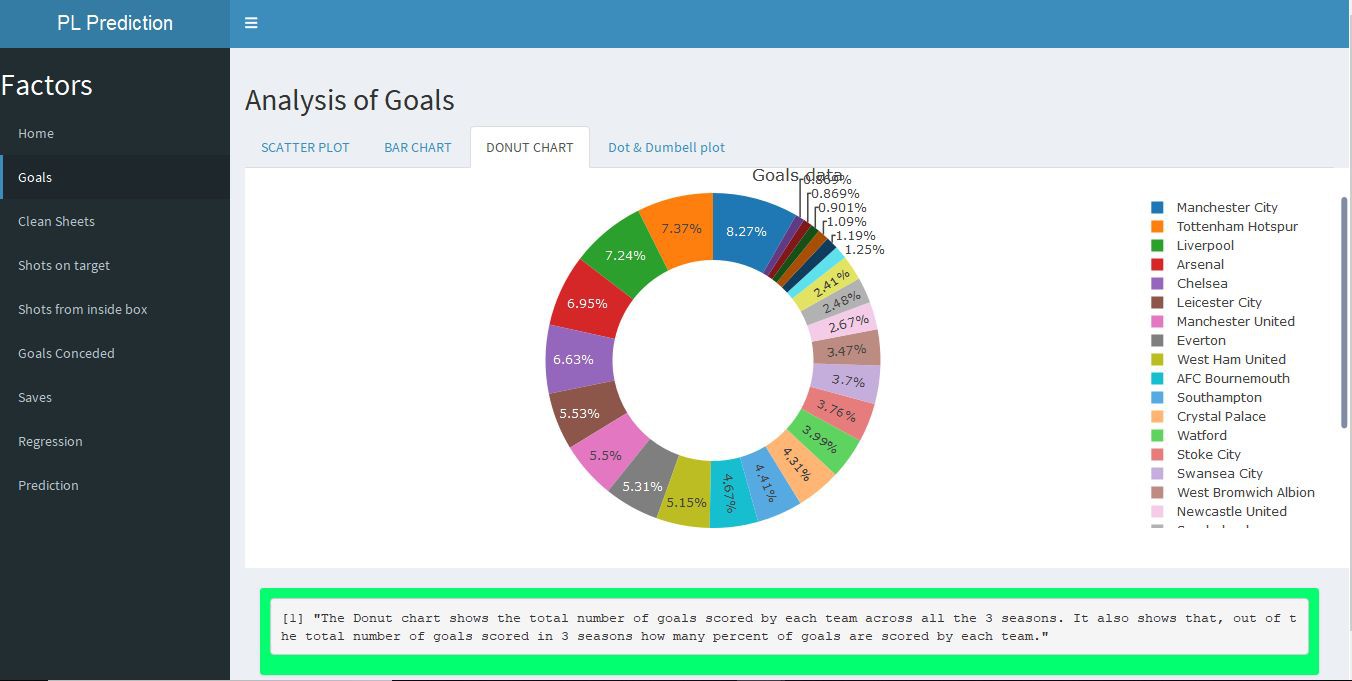
1. Graph showing direct relation between goals and wins.



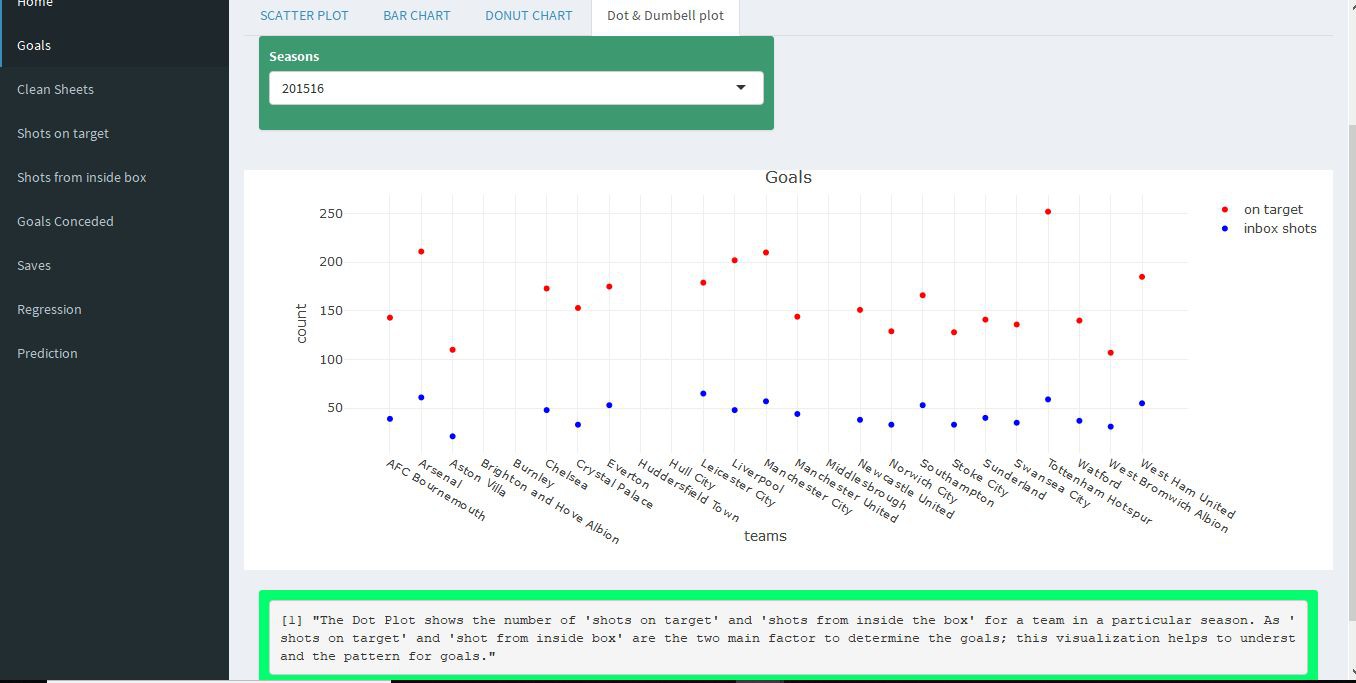
2. In which season a particular team has scored more number of goals ?



3. Which team has scored the highest number of goals in 3 seasons combined ?



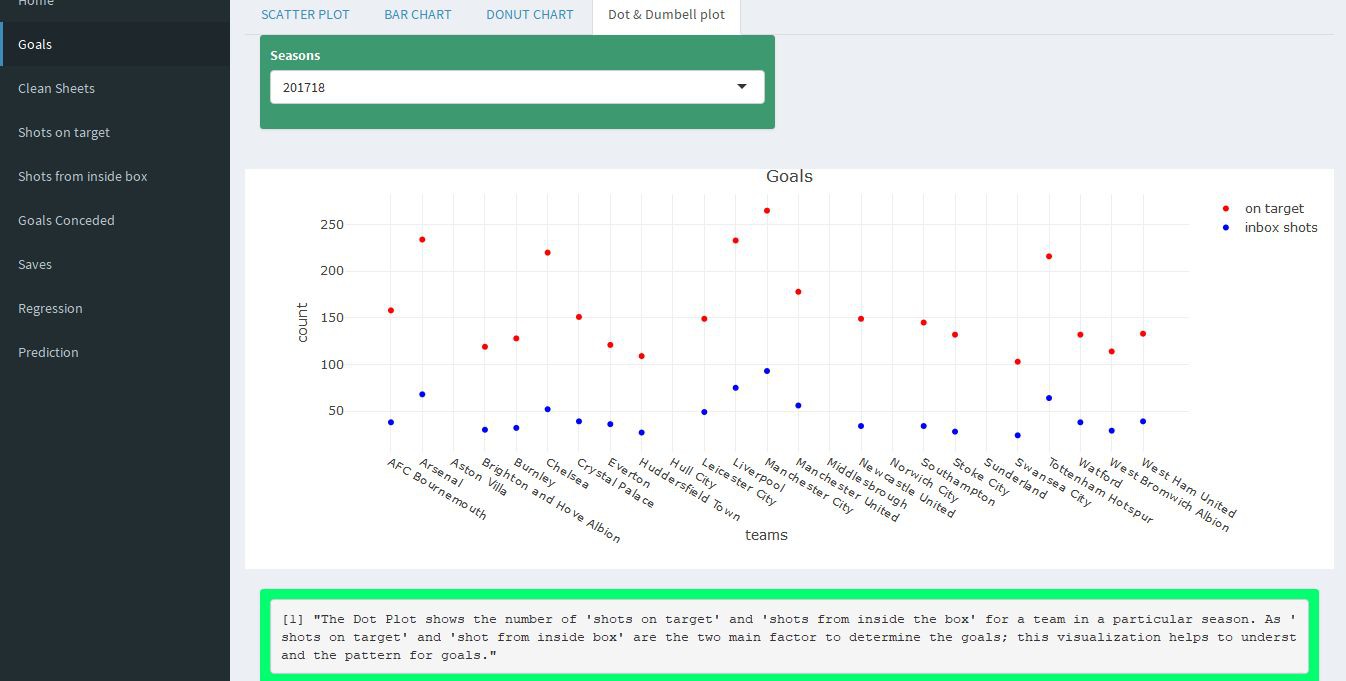
4. Which team had most number of “shots on target” and “shots from inside the box” in season 2015/16 ?



5. Which team had most number of “shots on target” and “shots from inside the box” in season 2016/17 ?



6. Which team had most number of “shots on target” and “shots from inside the box” in season 2017/18 ?

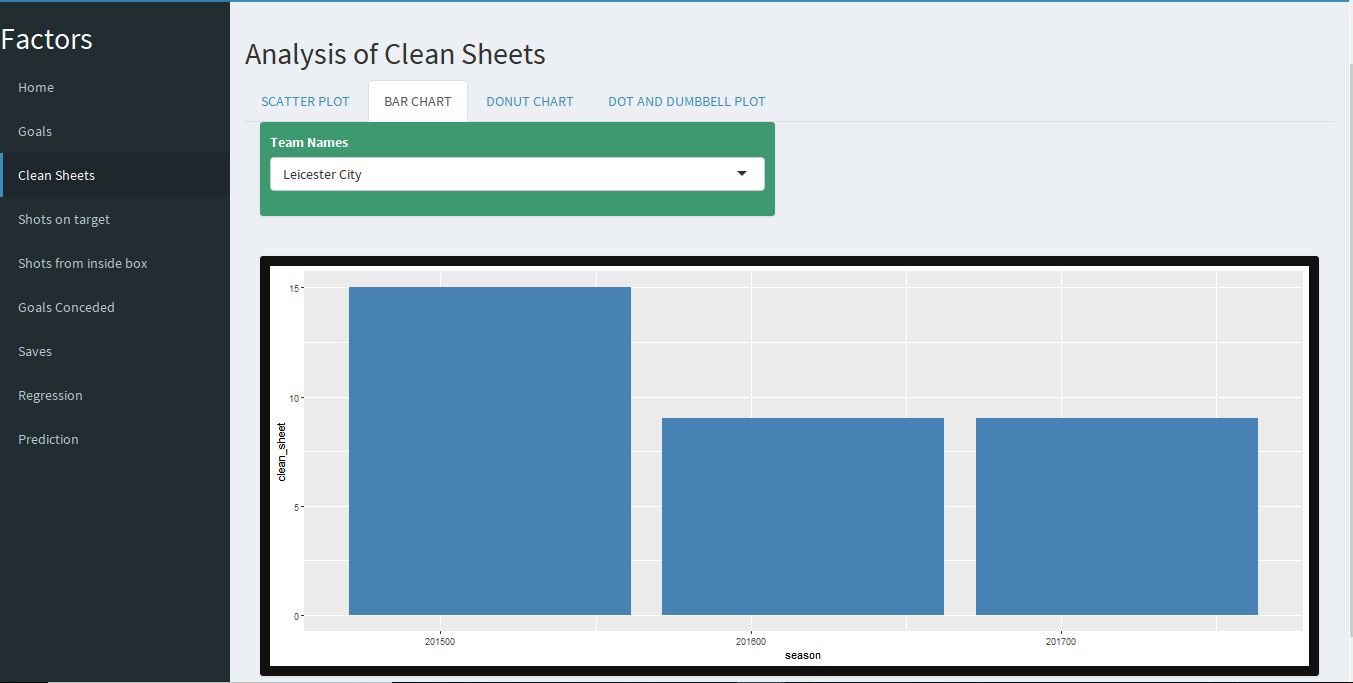


**B**. *FACTOR - CLEAN SHEETS*

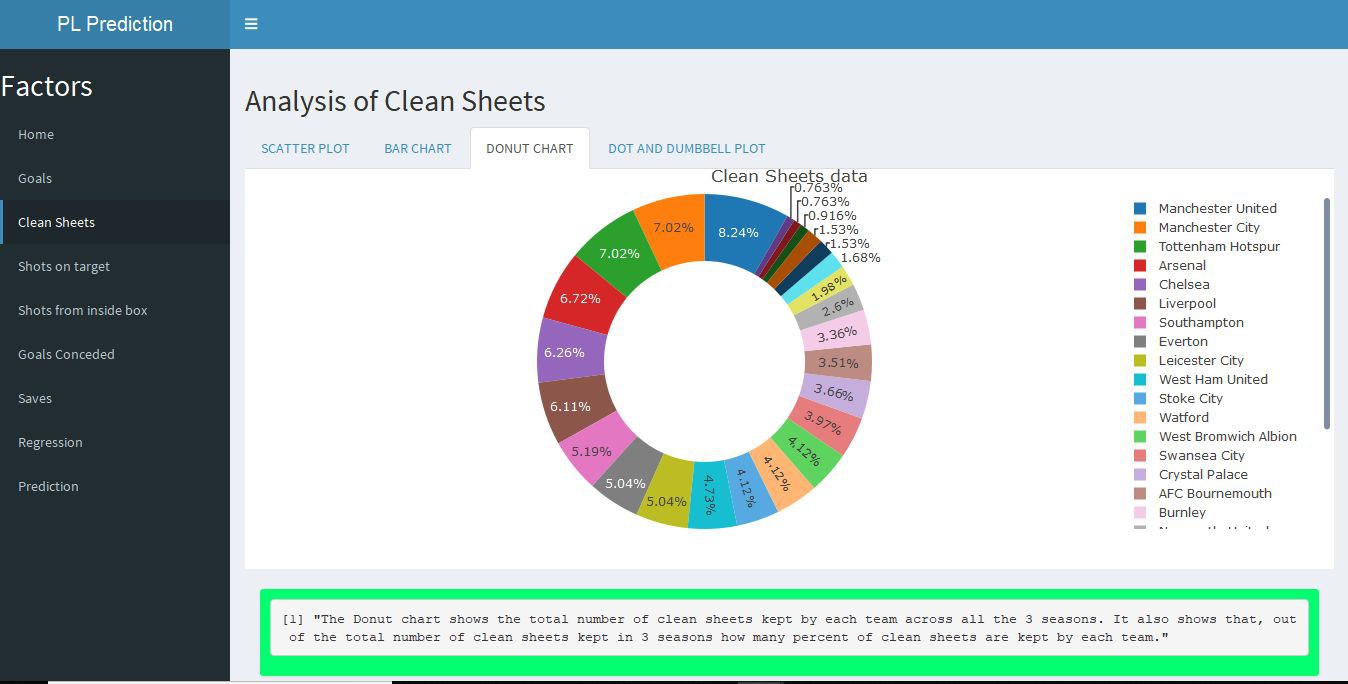
1. Graph showing direct relation between clean sheets and wins.



2. In which season a particular team has kept more clean sheets?

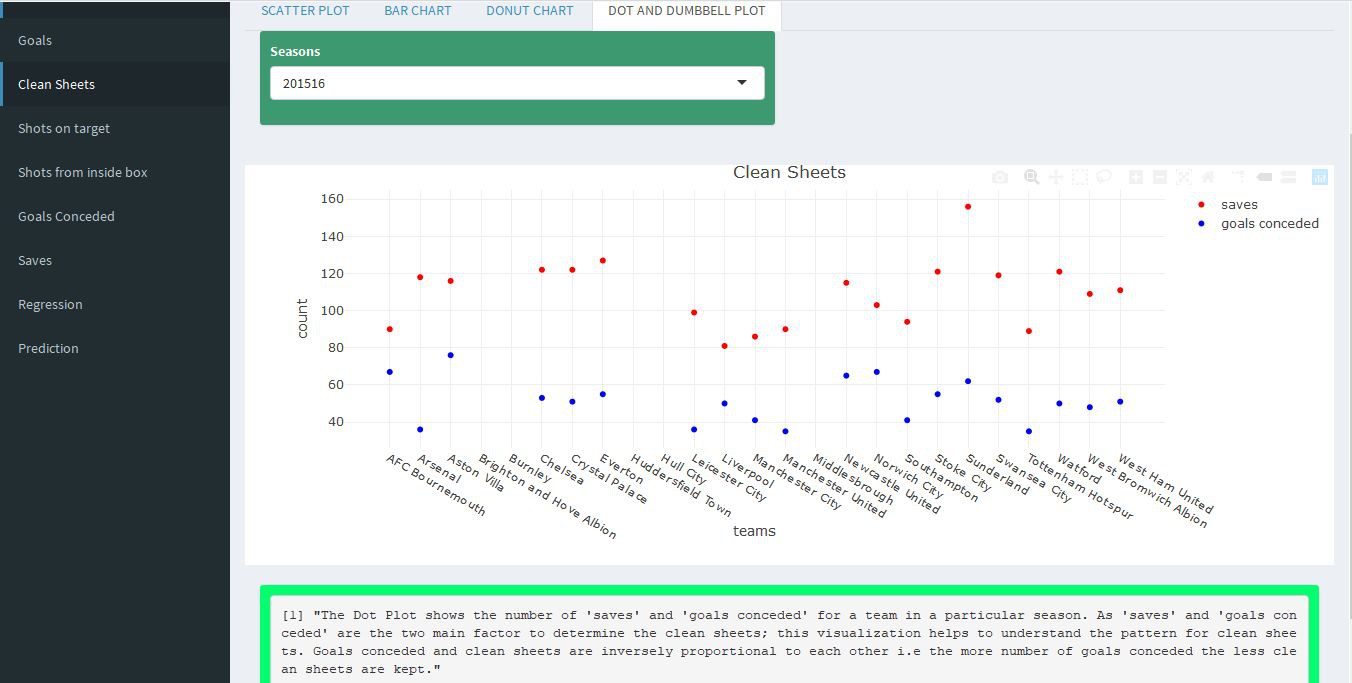


3. Which team has kept the highest number of clean sheets in 3 seasons combined ?



4. Which team had most number of “goals conceded” and “saves” in season

2015/16 ?



5. Which team had most number of “goals conceded” and “saves” in season

2016/17 ?



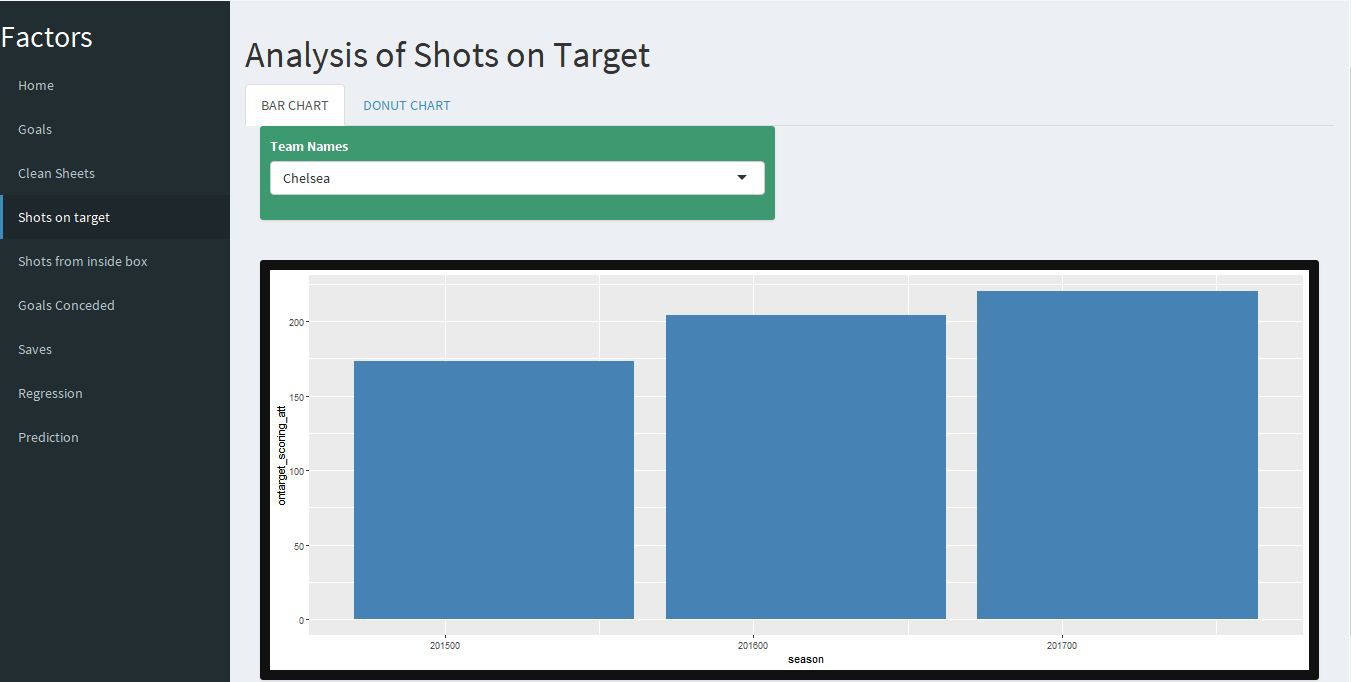
6. Which team had most number of “goals conceded” and “saves” in season

2017/18 ?

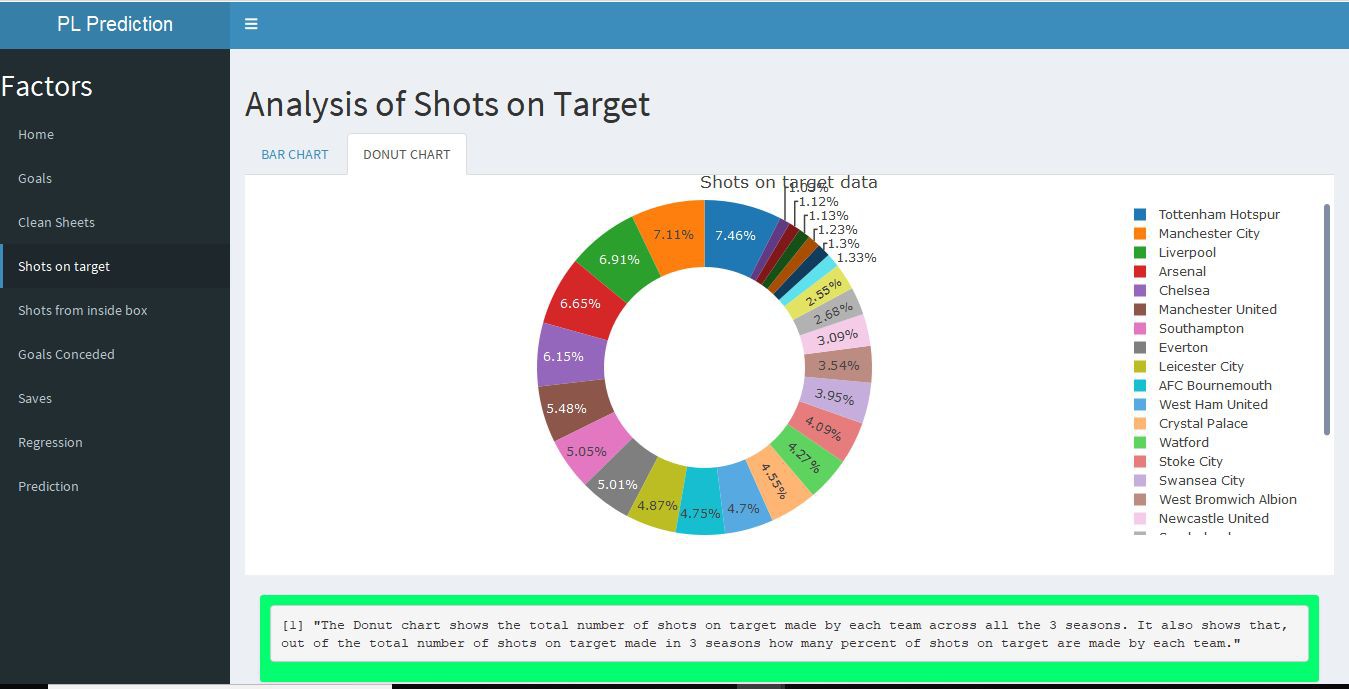


**C**. *FACTOR - SHOTS ON TARGET*

1. In which season a particular team has attempted more shots on target ?

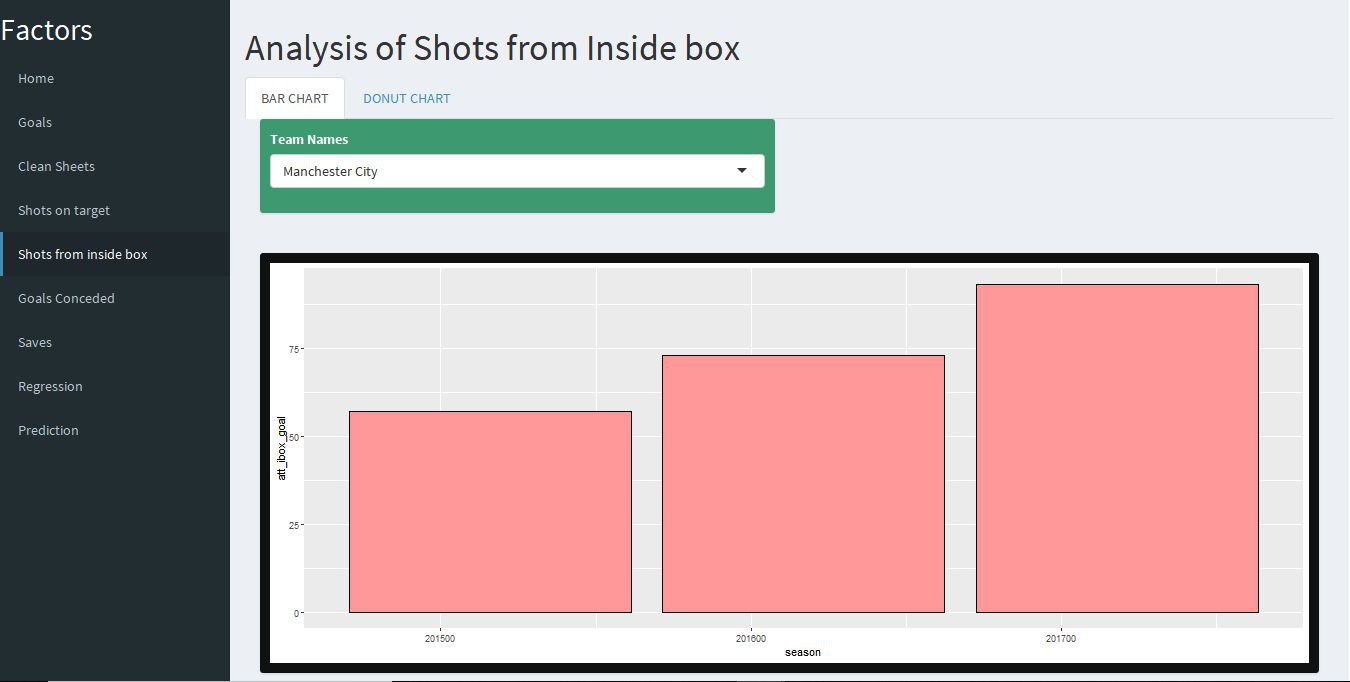


2. Which team has attempted the highest number of shots on target in 3 seasons combined ?

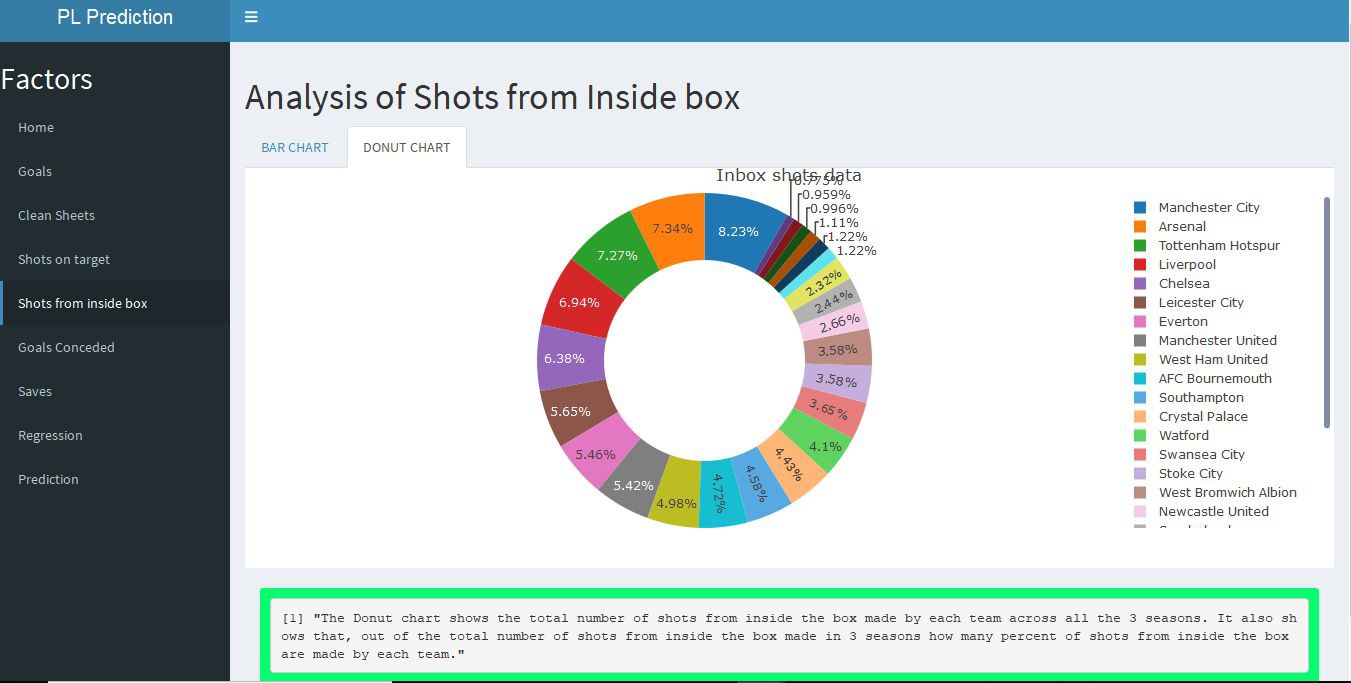


**D**. *FACTOR - SHOTS FROM INSIDE THE BOX*

1. In which season a particular team has attempted more shots from inside the box ?

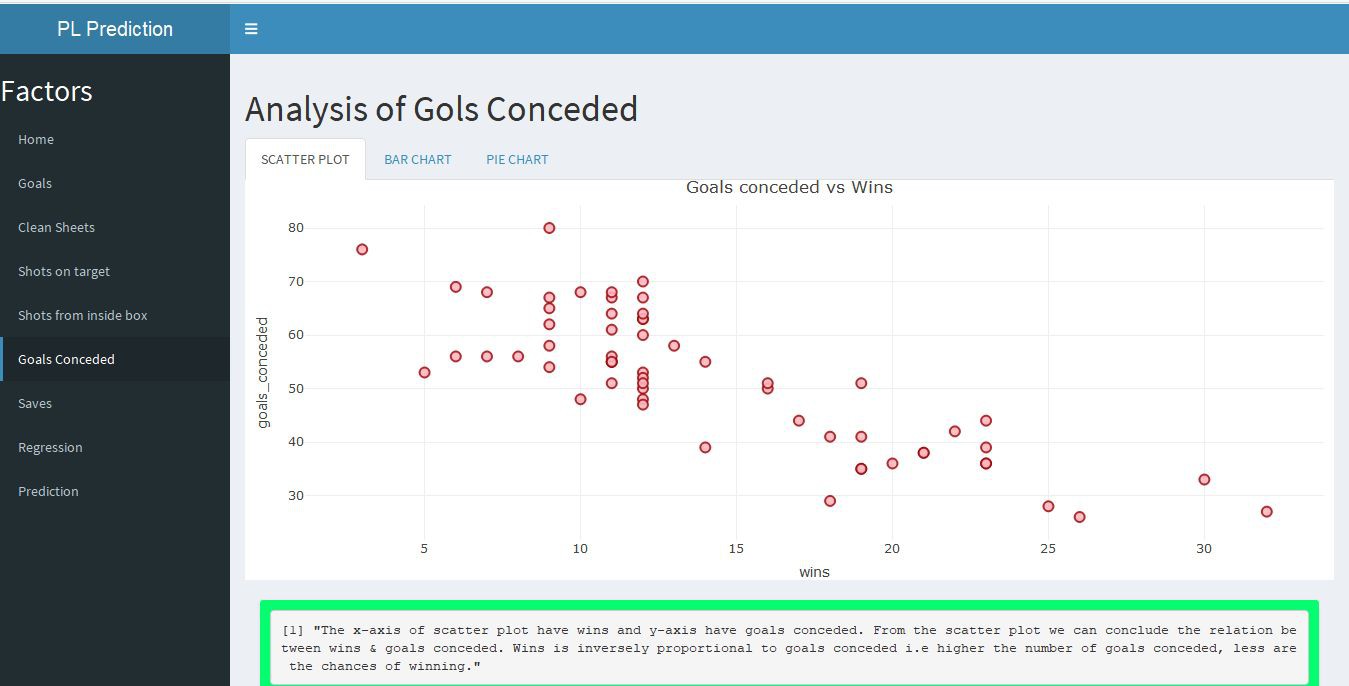


2. Which team has attempted the highest number of shots from inside the box in 3 seasons combined ?



**E**. *FACTOR - GOALS CONCEDED*

1. Graph showing direct relation between goals conceded and wins.



2. In which season a particular team has conceded more number of goals ?

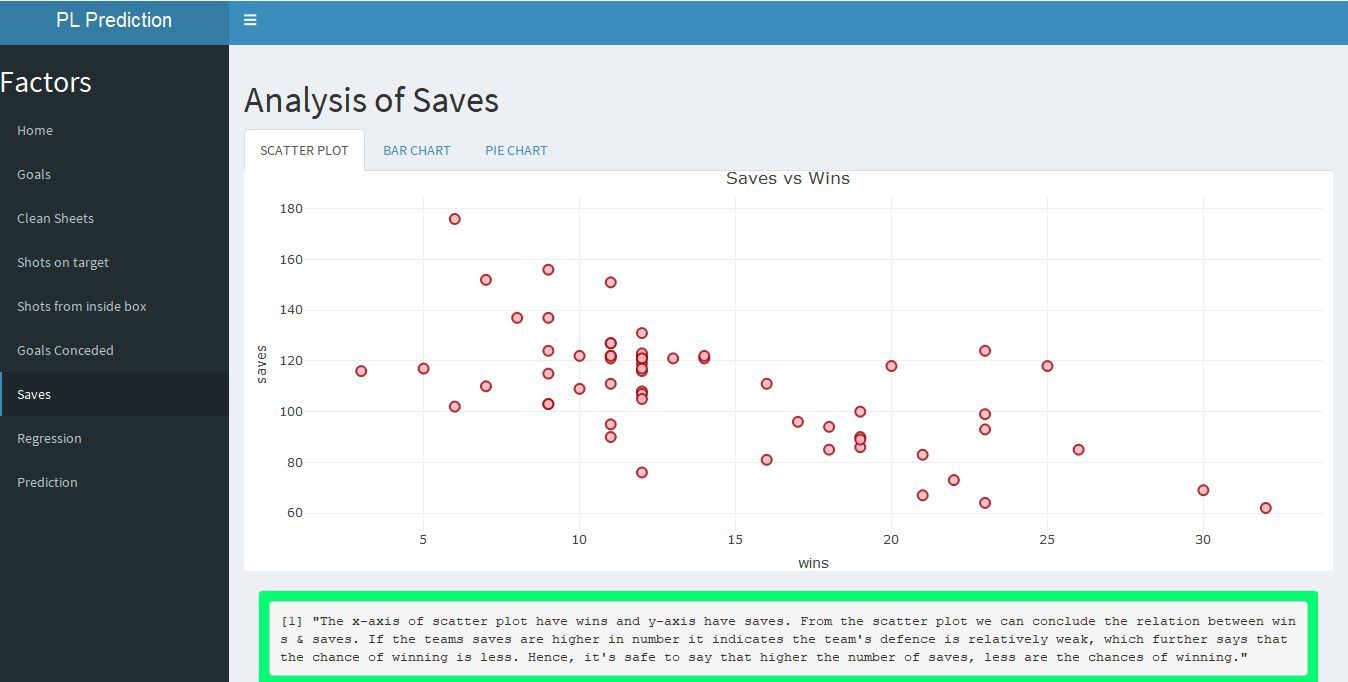


3. Which team has conceded most goals in 3 seasons combined ?



**F**. *FACTOR - SAVES*

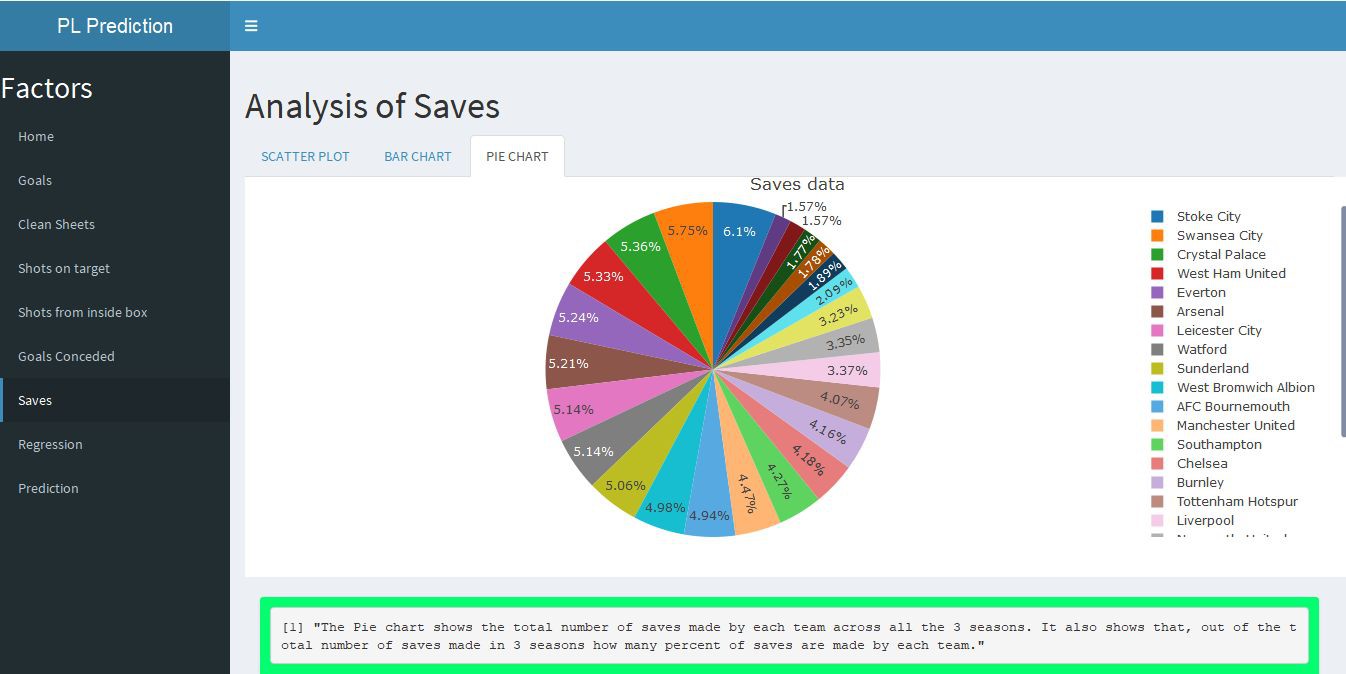
1. Graph showing direct relation between saves and wins.



2. In which season a particular team has made more number of saves ?

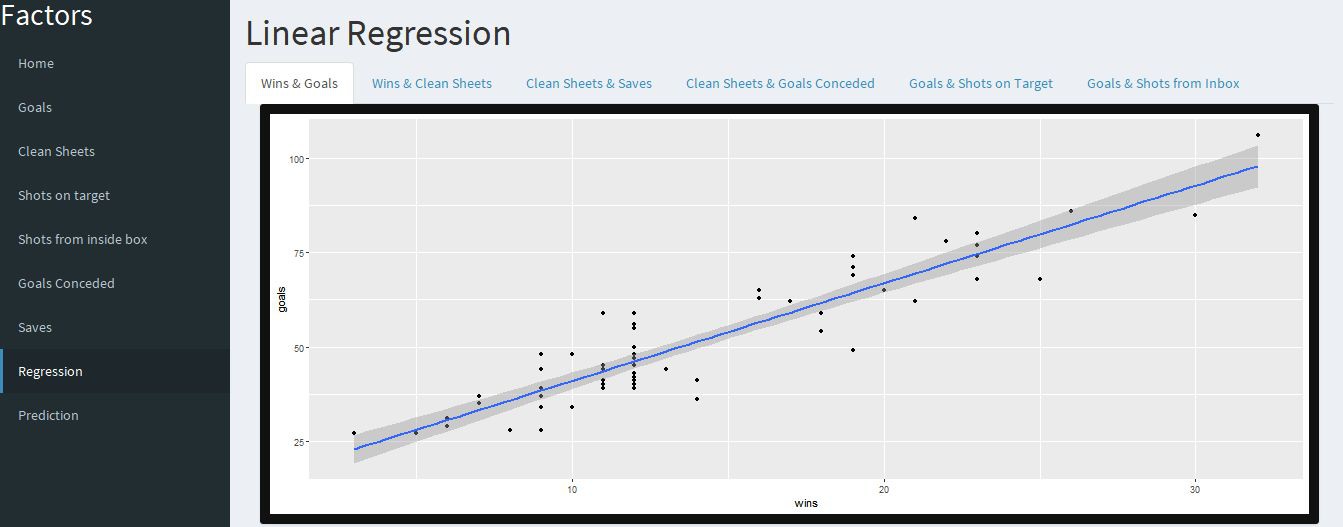


3. Which team has made most saves in 3 seasons combined ?

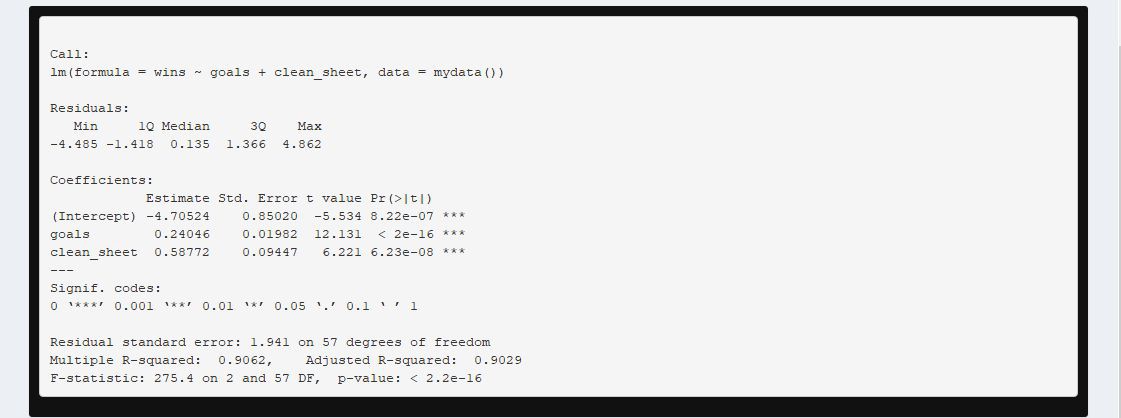


**3.3 REGRESSION**

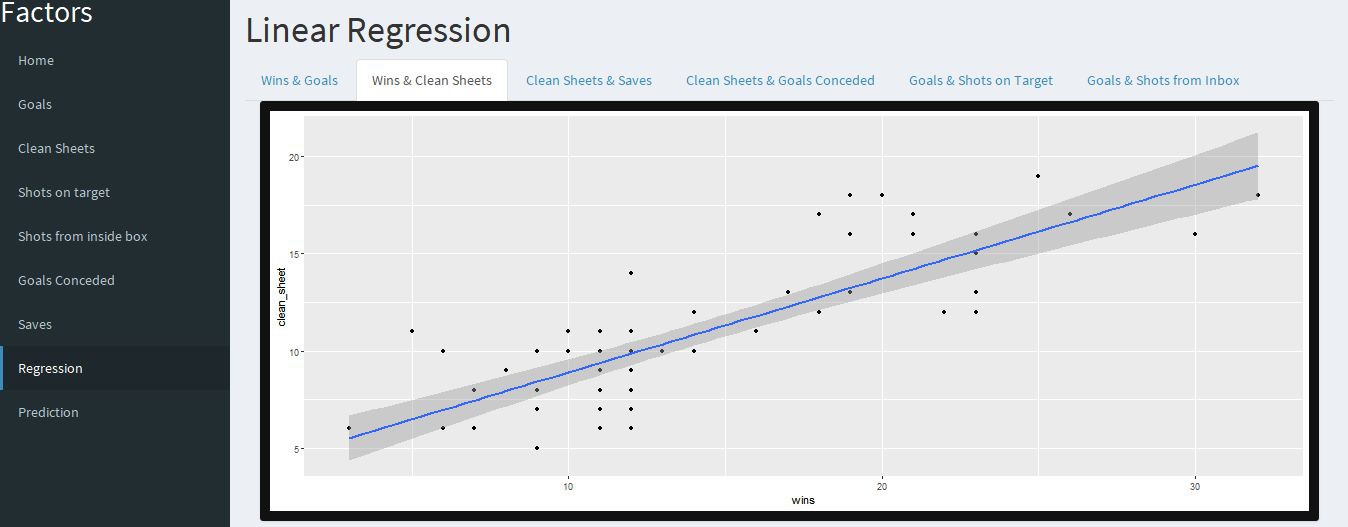
1. Graph for - Wins vs Goals



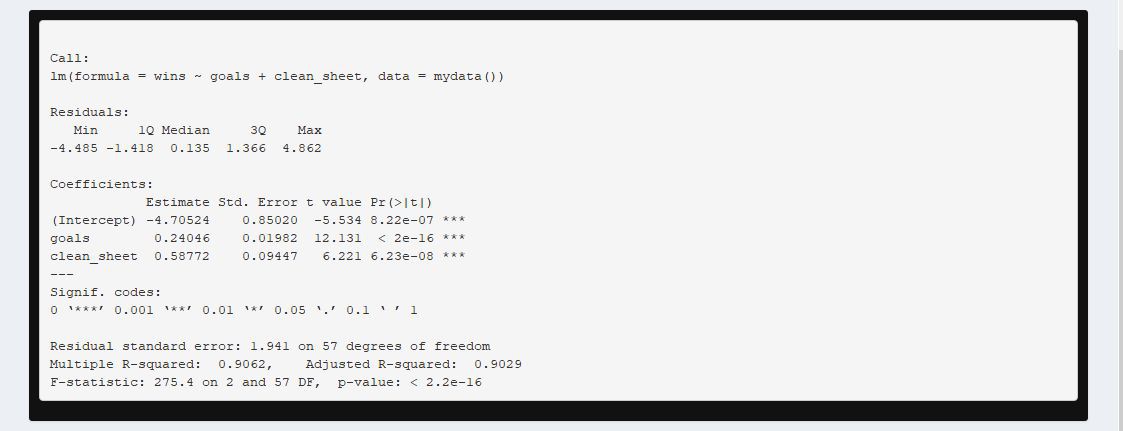
Summary for - Wins vs Goals



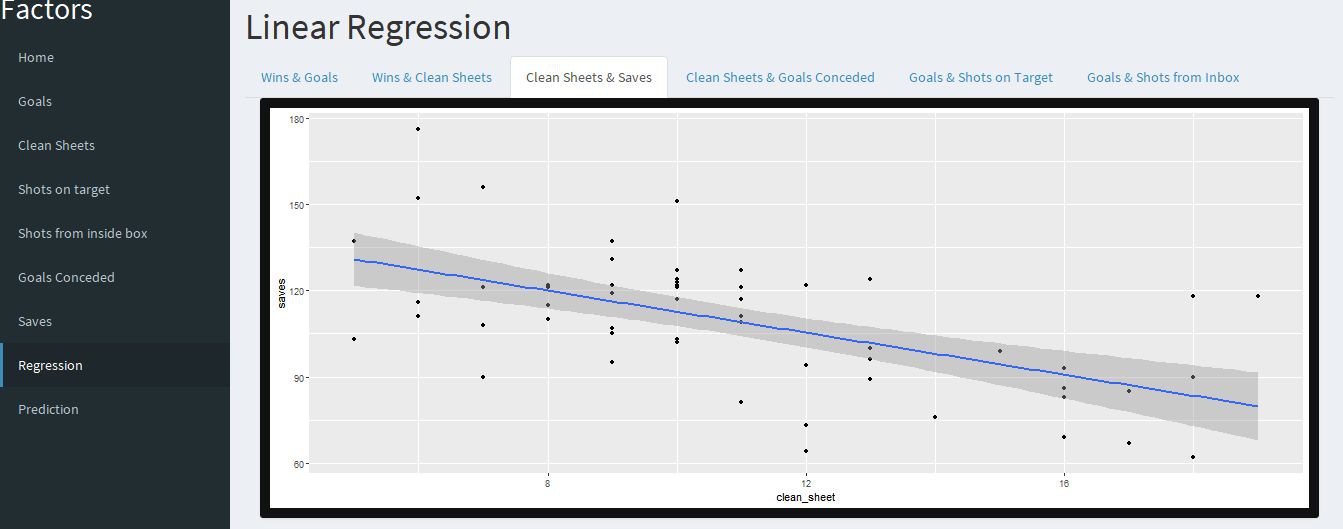
2. Graph for - Wins vs Clean Sheets



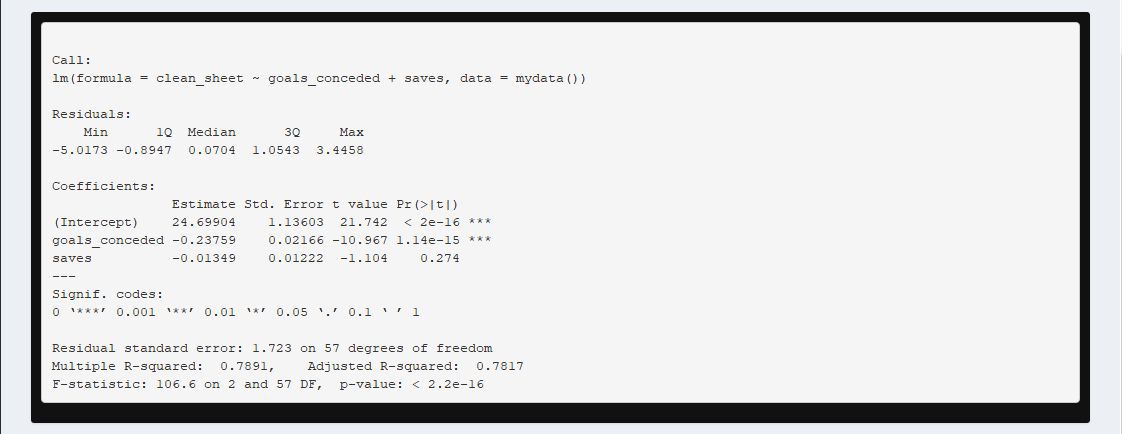
Summary for - Wins vs Clean Sheets



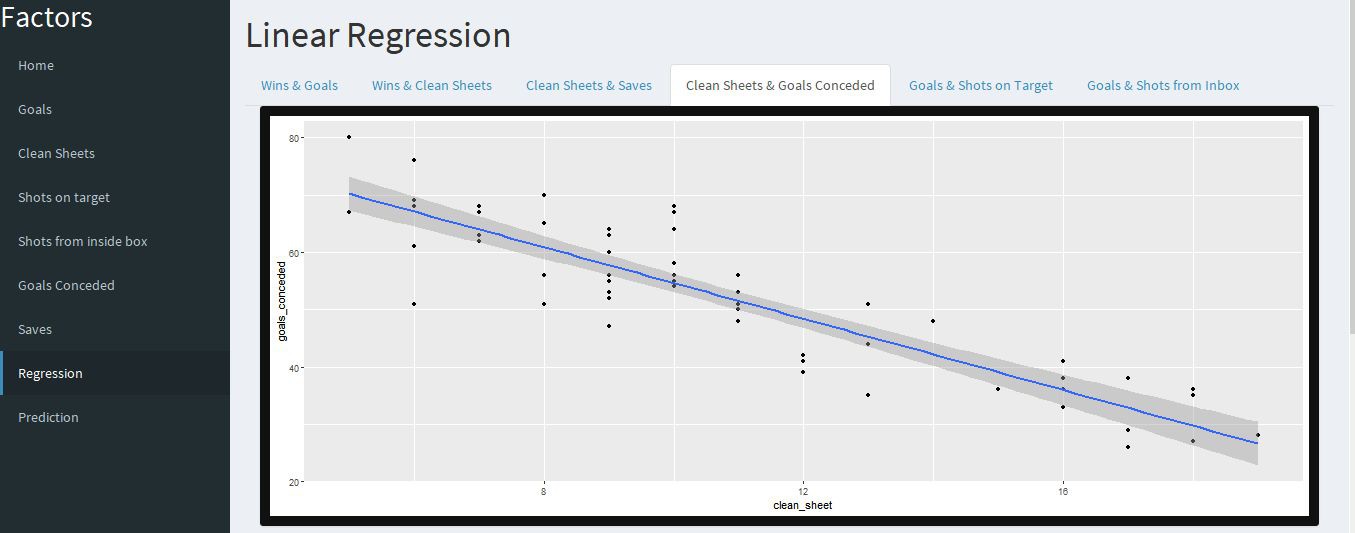
3. Graph for - Clean Sheets vs Saves



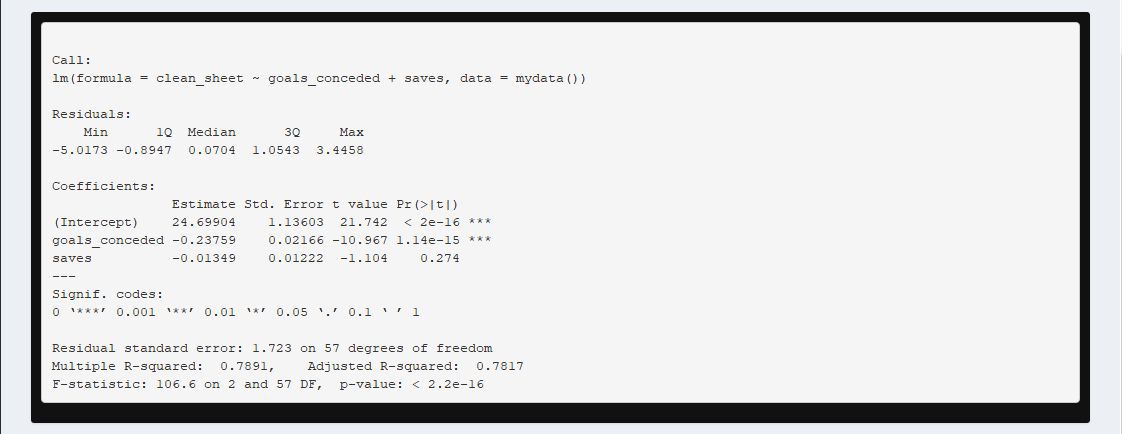
Summary for - Clean Sheets vs Saves



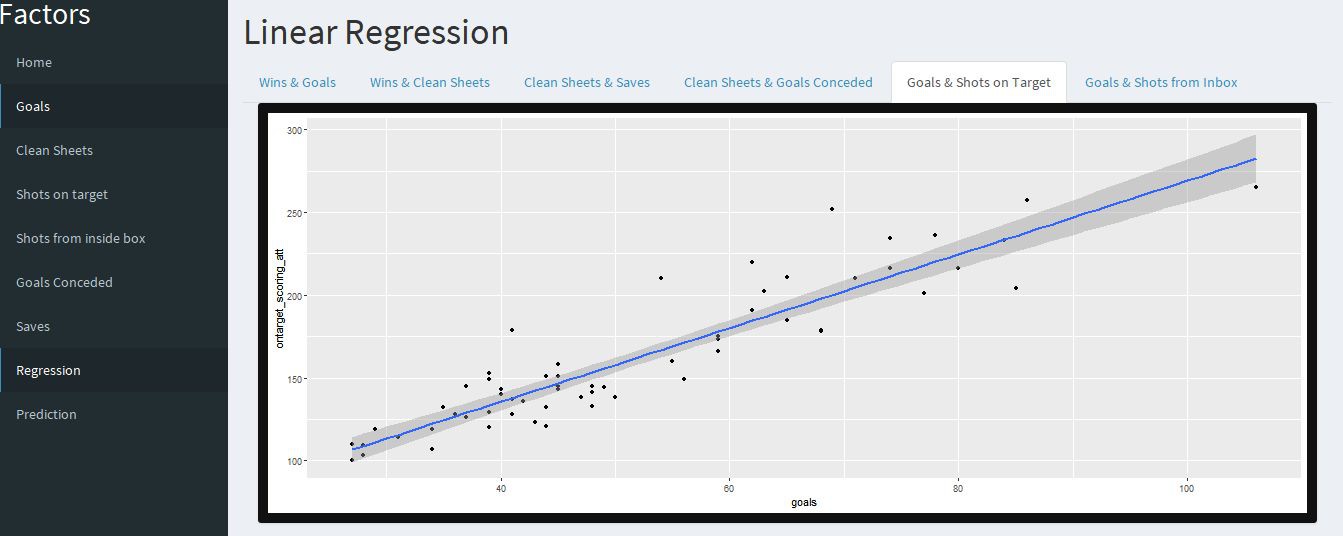
4. Graph for - Clean Sheets vs Goals Conceded



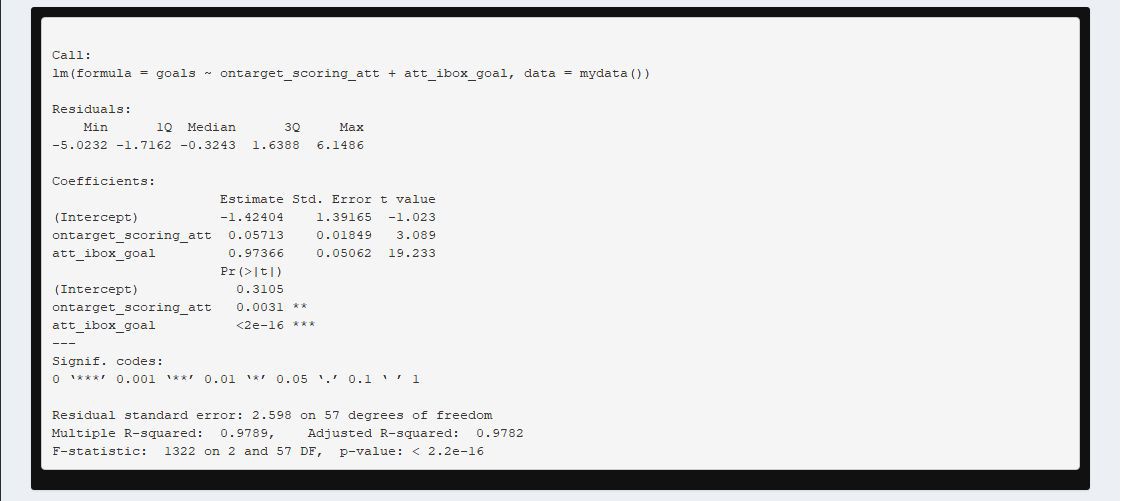
Summary for - Clean Sheets vs Goals Conceded



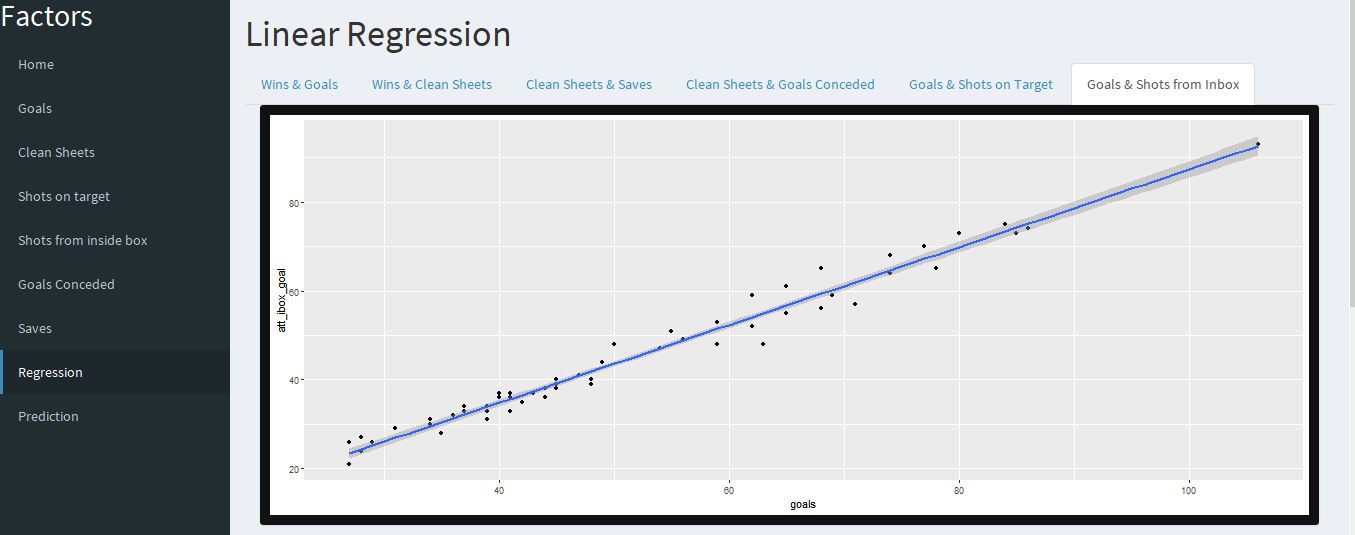
5. Graph for - Goals vs Shots on Target



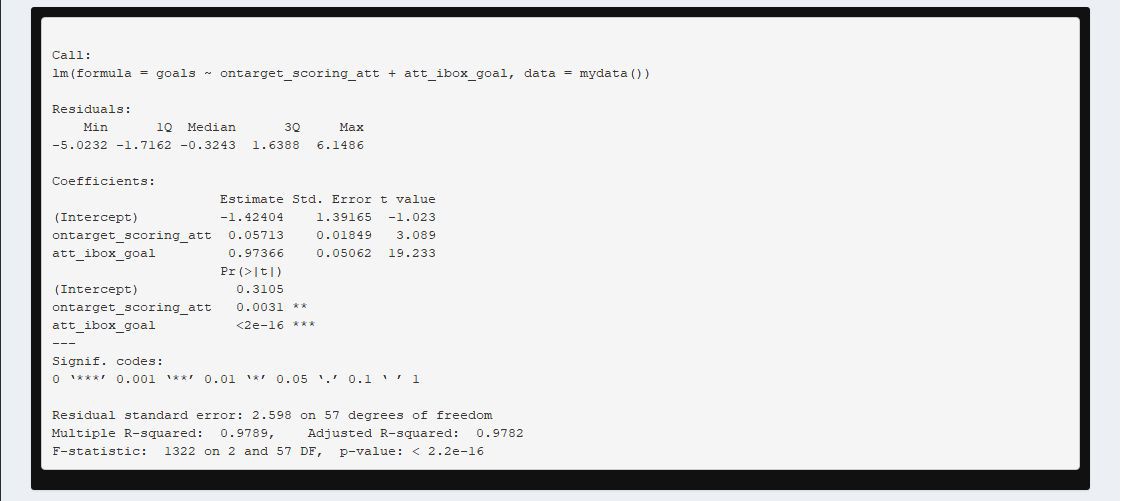
Summary for - Goals vs Shots on Target



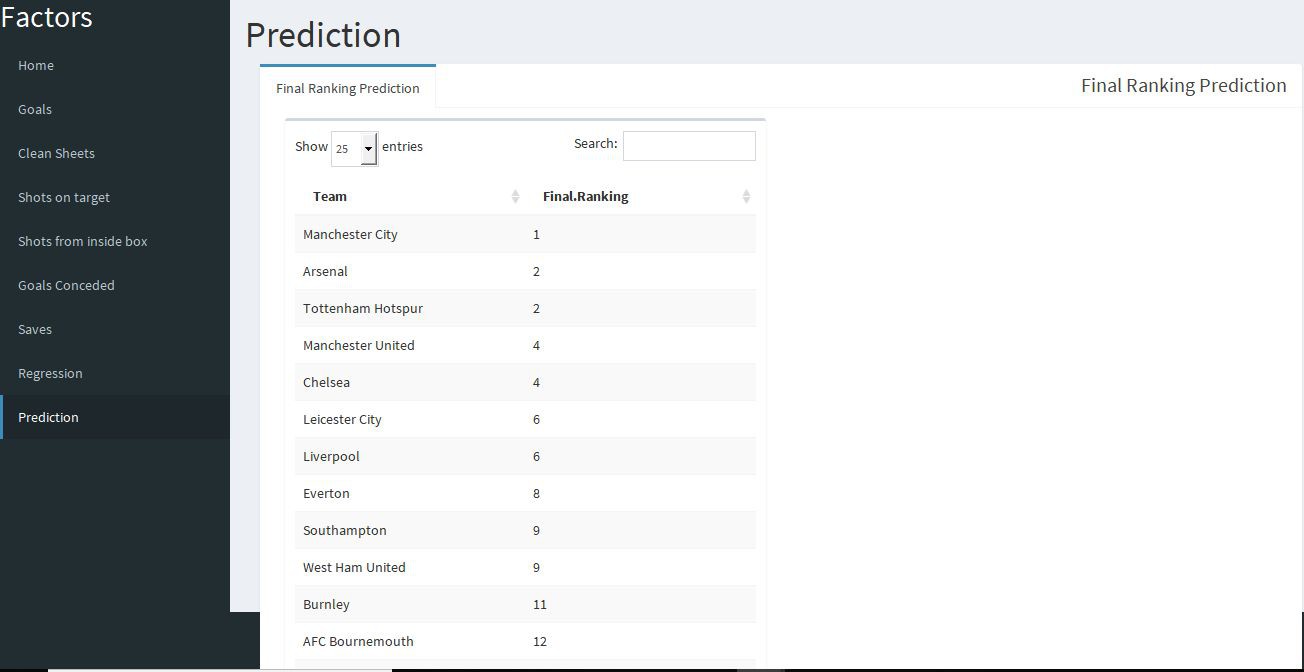
6. Graph for - Goals vs Shots from inside box



Summary for - Goals vs Shots from inside box



**FINAL PREDICTED STANDINGS :-**



**CHAPTER 6**

**3.4 CONCLUSIONS**

1) The 3 Linear Regression models has R-Squared value of 0.9062, 0.7891 and

0.9789.

2) This models fairly predicts the result of English Premier League for next season.

3) The confidence level for the model is fairly low.

4) The model works satisfactorily.

**3.5 FUTURE SCOPE AND DEVELOPMENT**

1. Take into consideration players of the team and their performance.

Sentimental Analysis can be used to predict whether the player’s attitude will affect the overall team’s performance.

2. Take into consideration the result of each and every match played between all the teams and then predict in the basis of the Home or Away wins.

3. Take into consideration the list of impact players who got transferred from one team to another which can affect the team’s result relative to previous season.

**S/W System Configuration:**

* Operating System - XP/7/8/8.1/10
* Coding Language - R –Language,Python.

**Hardware Requirements**

• Processors - Pentium IV Processor

• Speed - 3.00 GHZ

• RAM - 2 GB

• Storage - 20 GB

**Software Requirements**

• Operating system - Windows 10 Professional

• IDE used - Visual Studio Code

**3.6 REFERENCES AND APPENDIX**

**REFERENCES** :-

**1. Stern Hal. (1995) *Who's Number 1 in College Football?...And How Might We***

***Decide?*Chance, Summer, 7-14.**

2. [**^ Moroney M. J. (1956)**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Moroney_article_2-0) ***Facts from figures*. 3rd edition, Penguin, London.**

3. [**^ Reep C. Benjamin B. (1968)**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Reep_article_3-0) ***Skill and chance in association football*. Journal of the Royal Statistical Society, Series A, 131, 581-585.**

4. [**^ Hill I.D. (1974),**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Hill_article_4-0) ***Association football and statistical inference*. Applied statistics, 23, 203-208.**

5. [***a***](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Maher_article_5-0) [***b***](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Maher_article_5-1) [***c***](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Maher_article_5-2) [***d* Maher M.J. (1982),**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Maher_article_5-3) ***Modelling Association Football scores*. Statistica**

**Neerlandica, 36, 109-118**

6. [**^ Caurneya K.S. and Carron A.V. (1992)**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Caurneya_article_6-0) ***The home advantage in sports competitions: a literature review*. Journal of Sport and Exercise Physiology,**

**14, 13-27.**

7. [**^ Knorr-Held, Leonhard (1997)**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Knorr_article_7-0) ***Dynamic Rating of Sports Teams*. (REVISED**

**1999). Collaborative Research Center 386, Discussion Paper 98**

8. [***a***](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Kuonen_article_8-0) [***b* Diego Kuonen (1996)**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Kuonen_article_8-1) ***Statistical Models for Knock-out Soccer Tournaments***

9. [**^ Lee A. J. (1997)**](https://en.wikipedia.org/wiki/Statistical_association_football_predictions#cite_ref-Lee_article_9-0) ***Modeling scores in Premier League: is Manchester United really the best*. Chance, 10, 15-19**

10. **Kalpdrum Passi and Niravkumar Pandey discussed about the prediction accuracy in terms of runs scored by batsman and the no. of wickets taken by the bowler in each team [1].**

**11. P. Wickramasinghe proposed a methodology to predict the performance of batsman for the previous five years using hierarchial linear model [2].**

**12. R.P.Schumaker et. al, discussed about different statistical simulations used in predictive modeling for different sports [3].**

**13. John McCullagh implemented neural networks and datamining techniques to identify the talent and also for the selection of players based on the talent in Australian Football League[4].**

**14. Bunker et. al, proposed a novel sport prediction framework to solve specific challenges and predict sports results [5].**

**15. Ramon Diaz-Uriarte et. al, investigated the use of random forest for classification of microarray data and proposed a new method of gene selection in classification problem based on random forest [6].**

**16. Rabindra Lamsal and Ayesha Choudhary, proposed a solution to calculate the weightage of a team based on the player’s past performance of IPL using linear regression [7].**

**17. Akhil Nimmagadda et. Al, proposed a model using Multiple Variable Linear Regression and Logistic regression to predict the score in different innings and also the winner of the match using Random Forest algorithm [8].**

**18. Ujwal U J et. Al, predicted the outcome of the given cricket match by analyzing previous cricket matches using Google Prediction API [9].**

**19. Rameshwari Lokhande and P.M.Chawan came up with live cricket score predicton using linear regression and Naïve Bayes classifier [10].**

**20. Abhishek Naik et. Al, proposed a new model used matrix factorization technique to analyze and predict the winner in ODI cricket match [11].**

**21. Esha Goel and Er. Abhilasha discussed the improvements in Random Forest Algorithmand described the usage in various fields like agriculture, astronomy, medicine, etc. [12].**

22. Amit Dhurandhar and Alin Dobra proposed a new methodology for analysing the error of classifiers and model selection measures to analyse the decision tree algorithm [13].

23. H. Yusuff et. Al, performed logistic regression using mammograms to find the accuracy with valid samples [14].

**APPENDIX :-**

***Multiple Linear Regression***

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS)

regression that involves more than one explanatory variable.

The Formula for Multiple linear regression is :-

***Yi* = *β*0 + *β*1*xi*1 + *β*2*xi*2 +...+ *βpxip* +** ϵ

*where, for i=n observations:*

***Yi* = *dependent variable***

***xi* = explanatory variables**

***Β*0 = y-intercept (constant term)**

***Βp* = slope coefficients for each explanatory variable**

ϵ **= the model’s error term (also known as the residuals)**

The multiple regression model is based on the following assumptions:

● There is a [linear relationship between the dependent variables and the](https://www.investopedia.com/terms/l/linearrelationship.asp) independent variables.

● The independent variables are not too highly [correlated with each other.](https://www.investopedia.com/terms/c/correlation.asp)

● yi observations are selected independently and randomly from the population.

● Residuals should be [normally distributed with a mean of 0 and](https://www.investopedia.com/terms/n/normaldistribution.asp) [variance *σ.*](https://www.investopedia.com/terms/v/variance.asp)

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of a multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

**3.7 RANKING ALGORITHM :-**

**Learning to rank** or **machine-learned ranking** (MLR) is the application of machine learning, typically supervised, semi-supervised or reinforcement learning, in the construction of ranking models for information retrieval systems.[[2] Training data consists](https://en.wikipedia.org/wiki/Learning_to_rank#cite_note-2) of lists of items with some [partial order specified between items in each list. This order is](https://en.wikipedia.org/wiki/Partial_order) typically induced by giving a numerical or ordinal score or a binary judgment (e.g. "relevant" or "not relevant") for each item. The ranking model's purpose is to rank, i.e. produce a permutation of items in new, unseen lists in a way which is "similar" to rankings in the training data in some sense.

Learning to rank algorithms have been applied in areas other than information retrieval:

● In machine translation for ranking a set of hypothesized translations;

● In computational biology for ranking candidate 3-D structures in protein structure prediction problem.

● In recommender systems for identifying a ranked list of related news articles to recommend to a user after he or she has read a current news article.

● In software engineering, learning-to-rank methods have been used for fault localization.

**INDEX AND ACRONYMS**

Business 5 ,6, 7 Analysis on Shots 20,21

from inside box

Customer 6 Analysis on Goals 22,23

conceded

Analysis on 10,11,12, Analysis on 24,25

Goals 13 Saves

Analysis on 14,15,16, Regression 26,27,28,

Clean Sheets 17 29,30,31

Analysis on 18,19 Prediction 32

Shots on Target