Analysis And Prediction Model For Football Scores and Results

Utkarsh Devgan **Suchit Mathur** **Ayush Verma**
Department of Information And Communication Technology, Manipal Institute Of Technology

I. ABSTRACT

Predictive analytics encompasses a variety of statistical techniques from predictive modelling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events. Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the various models is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance.We have used knowledge discovery in databases (KDD) to develop a football match result predictive model and player clustering model by utilising various parameters that affect the result of football matches. Clustering, Regression and Classification techniques have been utilised in the implementation of the project.

II. PROBLEM STATEMENT

Each manager managing a team in the English Premier League is given a set of objectives by the Board of Directors of the particular club. These are the minimum goals the manager is expected to achieve during the course of the season. Managers employ varying techniques and latest technology to achieve these objectives. The manager should not only play the right squad in each match by assessing the overall and opponent specific performance of a player, but also transfer out or loan the player if his performance is not up to the mark. He should also be able to purchase new players within the allocated budget to strengthen the squad so that the team finishes at a good ranking at the end of the season. Hence, this crucial decision making prompted us to take up this score prediction project as the mini project for Data Mining and Predictive Analysis Lab.

III. OBJECTIVES

The football match result and score predictor has been developed keeping the following objectives in mind:

1. To implement score prediction for a match between any two teams in the English Premier League.
2. Classification of players according to their in-match performances.
3. Prediction of confidence levels of the score predicted.
4. To provide data visualisation for a better and more accurate understanding of the results obtained.

IV. INTRODUCTION

Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data

mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.

Football is the most popular sport played in the world. 2014 FIFA World Cup, held in Brazil was watched by a television audience of 3.2 Billion people around the globe. 280 Million people watched the match online or on a mobile handset. The 2014 FIFA World Cup broke numerous television records including the United States and Germany. On the club level, the English Premier League has an in-home audience of 3.0 Billion people, and is watched in roughly 730 Million homes.

There are numerous predicting techniques that are used to employ match result and score prediction techniques. However, our data prediction model also classifies potential transfer market targets into the best, good and average strikers and defenders. Prediction systems like ours are therefore heavily employed in industries, especially in stock markets and score predictions. A lot of factors actually help in predicting the score outcome or player performance. These are shots on target, passes, goals, assists, number of cards, number of successful interceptions, number of tackles etc. All these parameters therefore need to be considered for the correct and accurate functioning of the data prediction model. Weights are therefore associated with the parameters and used for analysing the scores and player performances. The best performers can thus be identified by the manager and hopefully purchased in the transfer market to make their respective squads stronger so as to achieve a better rank than the previous season. Thus, better understanding of the transfer market can prove beneficial in the long time run.

V. METHODOLOGY

R-Studio is the software suite that has been incorporated for implementation purposes. The prediction process is utilised for two main functions, which are score prediction and player classification followed by ranking them. For score prediction, the process begins with cleansing of data. Individual player data was present in the data set which was aggregated by team name, opposition name and venue to obtain the playing squads of each team. Now, a linear model was related the number of goals, key passes, shots on target and the venue. This gave a holistic overview of team performance. Using each team's performance data as parameters, a decision tree was created in which the leaf nodes (classes) were set as the number of goals. Hence, the score prediction model was developed which can be used to predict the match outcome between any two teams in the English Premier League.

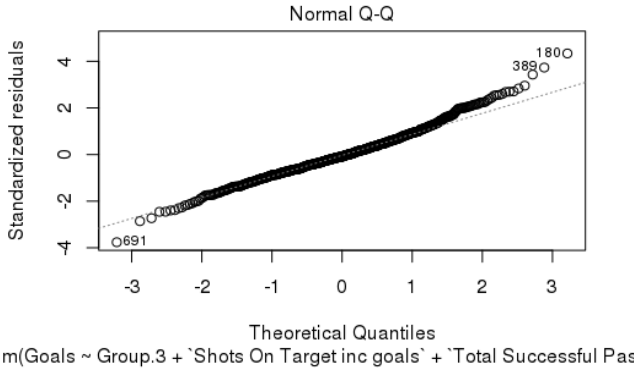


Figure 1

The shift from the dotted line represents the difference between the actual results and the predicted results.

The plot suggests that there is a decreasing linear relationship between number of goals and the linear dependency of the respective parameters which affect the number of

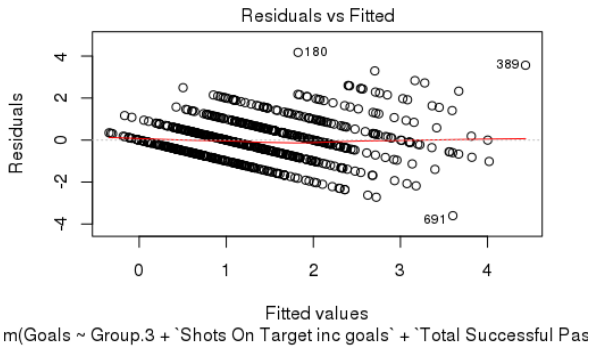


Figure 2

goals, namely venue, shots on target and total successful passes. It also suggests that there are no unusual data points i the dataset and it illustrates that the variation around the estimated regression line is constant, suggesting that the assumption of equal error variances is reasonable.

The decision tree consists of the most important performance parameters of the team that affects the final result. Iterating through the tree, we can predict the score result.

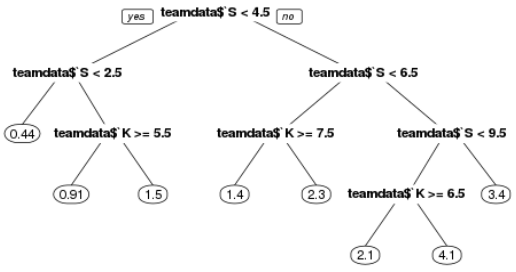
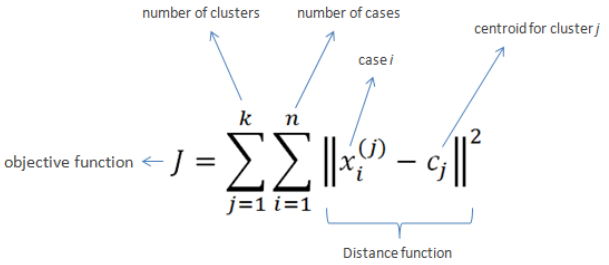


Figure 3

For the player performance and ranking, data was cleaned and the player classification was done into attackers and midfielders. The important parameters considered for classification of players into attackers comprised of shots on target and the number of goals scored. Similarly for classification into midfielders, the key parameters considered were total successful passes, key passes and number of assists. Based on these parameters, the best, average and below average attackers and midfielders are segregated into different clusters. K-means clustering has been used for cluster formation and to find the separation of cluster means from the origin. *k*-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



The players were given some cluster id and players were grouped according to cluster ids. Based on the cluster ids, cluster means were calculated and the distance of the means from the origin was calculated and were sorted into decreasing order. If the top k- results were required and k value was being satisfied by the topmost cluster, the top k players from the best cluster were picked. If the value of k was less, a certain number of players from the second best cluster were picked and adjusted in the top cluster to obtain the top k players.

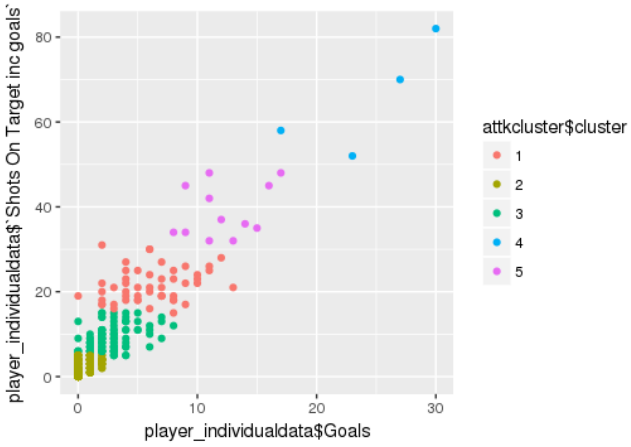


Figure 4

Figure 4 shows the plot of various attackers clusters which can be identified by the different colours. The parameters considered are goals and shots on target. The issue with this clustering is that the weights assigned to both the parameters were same. This resulted in inaccurate predictions .Therefore, both the parameters were assigned relevant weights and clustered again.

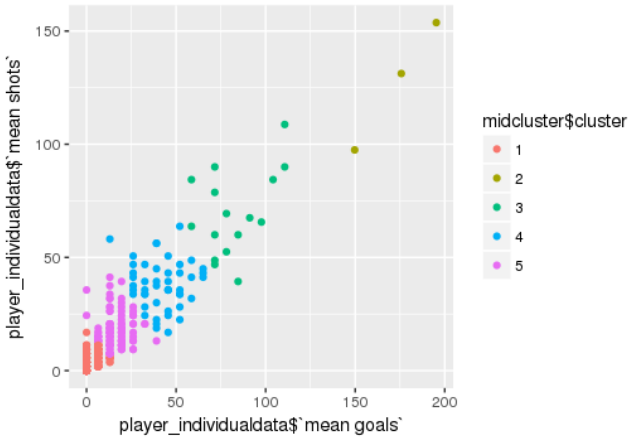


Figure 5

The new clusters obtained are shown in figure 5 and they represent a more accurate classification of the strikers. The new parameters are as follows:
Goals = Goals x Mean(Shots on Target)
Shots on target = Shots on Target x Mean (Goals)

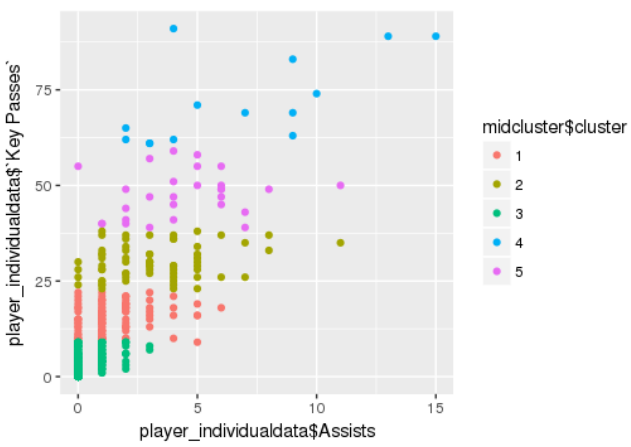


Figure 6

Similarly for midfielders, the key clustering parameters considered were assists and key passes. Figure 6 shows midfielder clusters formed by considering these two parameters. Here again, the same weight assignment to both the key parameters resulted in slightly inaccurate predictions. Therefore, relevant weights were assigned and the new plot can be seen in figure 7.

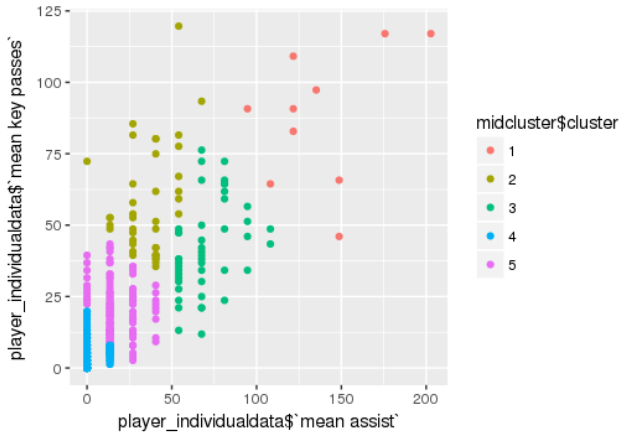


Figure 7

The new parameters are as follows:

Assists = Assists x Mean(Key Passes)
Key Passes = Key Passes x Mean (Assists)

VI. CONCLUSION

The above models for score prediction and player classification and ranking can be used to make smart decision while buying a team and training the team for a particular game event to strategize and attack over the weak areas of the opponent and to keep pre-planned threshold score target to beat the opponent easily.

VII. REFERENCES

[1] http://www.football-data.co.uk/data.php

[2] Ben Ulmer and Matthew Fernandez; Predicting Soccer Match results in the English Premier League, cs229, 2014.

[3] http://www.homepages.ucl.ac.uk/~ucakche/presentations/darksideposter.pdf

[4]. The R Project for Statistical Computing. Available from: http://www.r-project. org [cited 2005-11-28].

[5]. Moroney, K., 2014. Comeback Kings: An analysis of football scores and resilient teams. Management report 1. pp.2-6.

[6] https://ww2.coastal.edu/kingw/statistics/R-tutorials/simplelinear.html