Analysis and Prediction of Football Statistics using Data Mining Techniques

Anurag Gangal
VESIT, Mumbai

Abhishek Talnikar
VESIT, Mumbai

Aneesh Dalvi
VESIT, Mumbai

Vidya Zope
VESIT, Mumbai

Aadesh Kulkarni
VESIT, Mumbai

ABSTRACT
To solve the problem of loss of interest in Fantasy Football over the season, a game-changing strategy was thought of which led to the creation of this idea. Powered by an exhaustive dataset of all football statistics from 1992 i.e. the start of the Premier League era, it seemed exciting to allow the use of Data Mining techniques to forecast future statistics. A points system based on the success of predictions (explained later in detail), which in turn allow buying/auctioning better players adds a greater interactive feeling to the existing FPL system. This would prevent the churning of players of the season, since they would be attracted to getting more points and better players through such predictions.

Keywords
data-mining; sports; football; prediction; statistics; analysis; fantasy football;

1. INTRODUCTION
Football has been a game that was always the most popular sport to be played and viewed in most European and South American countries. The popularity of the sport, however, recently started booming in the Indian subcontinent since the early 21st century, probably owing to the popularity of televisions and global broadcasting of the games.

India is steadily becoming a global figure in the Football (American: Soccer) world, with more and more official football events happening and also major international stars participating in the new Indian Super League.

Keeping this in mind, it was clear that such a market of growing enthusiasts required to be tapped in with a "Fantasy" football concept. While the existing Fantasy Premier League is rather popular, a common problem faced by it is churn of players over the season i.e. players lose interest and don't come back to check it out every week.

To solve this problem, a game-changing strategy was thought of which led to the creation of this idea. Powered by an exhaustive dataset of all football statistics from 1992 i.e. the start of the Premier League era, it seemed exciting to allow the use of Data Mining techniques to forecast future statistics. A points system based on the success of predictions (explained later in detail), which in turn allow buying/auctioning better players adds a greater interactive feeling to the existing FPL system. This would prevent the churning of players of the season, since they would be attracted to getting more points and better players through such predictions.

1.1 The Existing System
Fantasy Premier Leagues work on the following concept:

A player signs up and makes his own team of players from various real-life teams. Each player has a pre-defined (slightly variable) value based on his previous form, which has to be paid to add the player to your team. Likewise, a team of 11 players + 5 substitutes has to be formed from a predefined budget which is the same for all players.

Now, as the actual games are played and the season progresses, points are gained by each player based on his real life performances - and likewise every player from a user's team will get certain points. Based on the summation of all points the user gets from his team, his position in a League (all users, friends, etc) is determined.

1.2 The Change
Now, the problem with the existing system is that users, especially indian users (show small survey) do not play throughout the season. To prevent this, using our data mining algorithm and the existing dataset, we add a system of predictions.

Predictions will be made before every match played in real life. Users can each make one prediction, before 1 hour of the actual kick-off time. Based on our own prediction algorithm, we then allot points to the user if his prediction was right.

These points will then be used to buy packs of players, which randomly generate players (gold ,silver, bronze). These players obtained from packs can then be added to teams or sold on the transfer market to other users for the player's value.

Now since the user has better players from the more points he got from making right predictions, the user has a better fantasy team which will land him more points.

Thus, our approach adds elements of exciting predictions and pack openings to the existing system, thus ensuring a more fun, interesting and season-long experience for the users.

2. LITERATURE SURVEY
We surveyed some of the literature which included some of the prediction algorithms to predict football results for different results.

Paper [1] used the algorithm implementing Bayesian Network together with machine learning techniques including a decision tree learner (MC4) and K-nearest neighbor (KNN) to predict the results of the games played by Tottenham Hotspur football club. They show that the expert Bayesian Network is generally superior to other machine learning techniques in terms of the prediction accuracy in this domain. They also

claim that the overall average accuracy for the expert Bayesian Network is 59.21% for the 3 outcomes prediction (win, draw or lose).

Paper [2] used different learning algorithms like Naive Bayes, Bayesian network, LogitBoost, KNN, random forest and Artificial Neural Network (ANN) to predict the game results of Europe Champions League. A software solution has been developed in order to try and solve this problem. To design this system for classification, the feature selection and choice of the learning algorithm can greatly affect the performance of classification. In Feature Selection, various important features like the current form of teams shown on the basis of results achieved in the last six games, the outcome of the previous meeting of the teams that play the game, the current position in the rankings, number of injured players from the first team, the average number of scored and received goals per game, which affect a football match are been chosen. After choosing correct features, a learning algorithm among the ones mentioned above is selected. During the development of the system, a number of tests have been carried out in order to determine the optimal combination of features and classifiers. The results of the presented system show a satisfactory capability of prediction which is superior to the one of the reference method. Paper [2] claims that the best accuracy is achieved by using an ANN which is around 68.8%.

A neural network method is adopted to predict the football game's winning rate of two teams according to their previous stage's official statistical data of 2006 World Cup Football Game in [5]. The input data are transformed to the relative ratios between two teams of each game. New training samples are added to the training samples at the previous stages. The adopted prediction model is based on multi-layer perceptron (MLP) with back propagation learning rule. They input the average data of each team into the weIl-trained MLP, and then they compare the output value to determine the relationship between victory and defeat. The team with bigger output value, which means the more ability to win the game, is the winner. This paper[5] applies the Artificial Neural Network (ANN) to the official statistical data of 2006 FIFA World Cup and based on the adopted MLP prediction method, it claims that the accuracy can be achieved is 76.9% if the draw games are not considered (i.e. 2 outcomes, win or not lose). Note that this approach will give a higher prediction rate than the three outcomes approach.

Several soft computing techniques like Fuzzy Logic, ANN and GP to perform the prediction task were used in [4]. The best overall accuracy is achieved by GP (around 76%).

Direct comparison between different work is difficult as they use different data sets and features, it seems that the soft computing based techniques (such as the ANN with back propagation [2], [5], GP[4]) have obtained a better performance over other machine learning techniques in terms of the prediction accuracy. Bayesian Networks [4] is also good, compared to the poor performance of other techniques (like MC4 or random forest [4]).

After studying all the existing literature on this topic, we came to the following conclusions:

Various algorithms were considered which were the following:

If three scenarios are considered, ie. Win, Lose or Draw, the following algorithms could be implemented:

Bayesian Networks - which gave an average accuracy of ~59.21%

Artificial Neural Networks (ANN) - which gave an accuracy of ~68.8%

Genetic Programming - which gave an accuracy of ~76%

If two scenarios are considered, ie. Win or Lose, the following algorithms could be implemented, as seen in Table 1:

Artificial Neural Networks (ANN) - gave an accuracy of ~76.9%

Table 1

Algorithm	Accuracy
Bayesian Network	59.21%
Artificial Neural Network	68.8%
Genetic Programming	76%

Even though the ANN in a two - outcome scenario (win or lose) gave the highest accuracy i.e. 76.9%, it is not favorable in the case of Football prediction, since Draw is an equally important outcome. Hence, the best algorithm in the three outcome scenario i.e. Genetic Programming was finalized to be used in our project.

3. PROPOSED SYSTEM

Figure 1 shows our proposed system. It improves the existing model of Fantasy Football by adding an element of Prediction to it. The system has been described in the following block diagram:

Our System will take Premier League Data set as input for performing Prediction , generating Statistical information for User Profiles and Betting Module. Also, An Internal Database will store all the Updated Premier League Information such as which team won the previous Game week matches, which player scored the goal and so on, which will not be present in the dataset. The User Profile roles and action will be tracked from the User Profile Module. As with the legacy Systems, our System also gives Users the fundamental feature of making their own Teams under the Team Formation Module. After team formations are done and bettings are placed, The Entire System gets updated using the Updation module. The Updation module makes changes in the User points, resets the Teams for each user, Updates the Leaderboard and other such functionalities.

4. PROPOSED ALGORITHM

The algorithm is a weighted equation which equates a total score for each team based on it's performances in various time frames and many other factors. This algorithm is used on the holistic dataset from 1992, thus giving rich data-driven solutions.

The algorithm will be used to come up with a score called "AlgoScore", which will be a variable score that each team will have when all the parameters in the algorithm are considered and equated.

While looking to make predictions, the AlgoScores of the two teams playing against each other will be calculated. The team with a higher AlgoScore will be the better team as per the algorithm.

Thus, predicting a win of a better team will land the user fewer points than predicting a win of the weaker team - since the weaker team is not expected to win the game.

Rewards of making a right prediction are allotted to users as follows :

Case A: If Team A has an AlgoScore X and Team B has an AlgoScore Y such that X>Y and either team wins ; a user who predicts a win of Team A and Team A wins, gets 100 (the prediction token cost) + (Y).

a user who predicts a win of Team B, gets 100 (the prediction token cost) + (X)

Since X>Y, the user who predicts the weaker team's win gets a greater reward than the user who predicts the expected result.

Case B: If Team A has an AlgoScore X and Team B has an AlgoScore Y such that X>Y and the match is a draw ; the user who predicts a draw, gets 100 (the prediction token

cost) + Z

where Z = (X+Y)/2

The factors that affect the results of a football match are segregated as follows:

4.1 A. Variable Factors

Variable Factors are the factors that vary throughout the football league season. These are parameters such as Current Form, Injuries, Suspensions, Previous Games against tougher opponents, possible effects of other tournaments etc.

These factors will change in value for every game that is played and have to be calculated before every game.

Each factor is given a specific weight which is determined from experimental observations.

4.2 B. Fixed Factors

These factors do not change over the course of the season. Such factors are calculated before every season commences and remain fixed throughout the next season. Variable and Fixed factors are shown in the following figure.

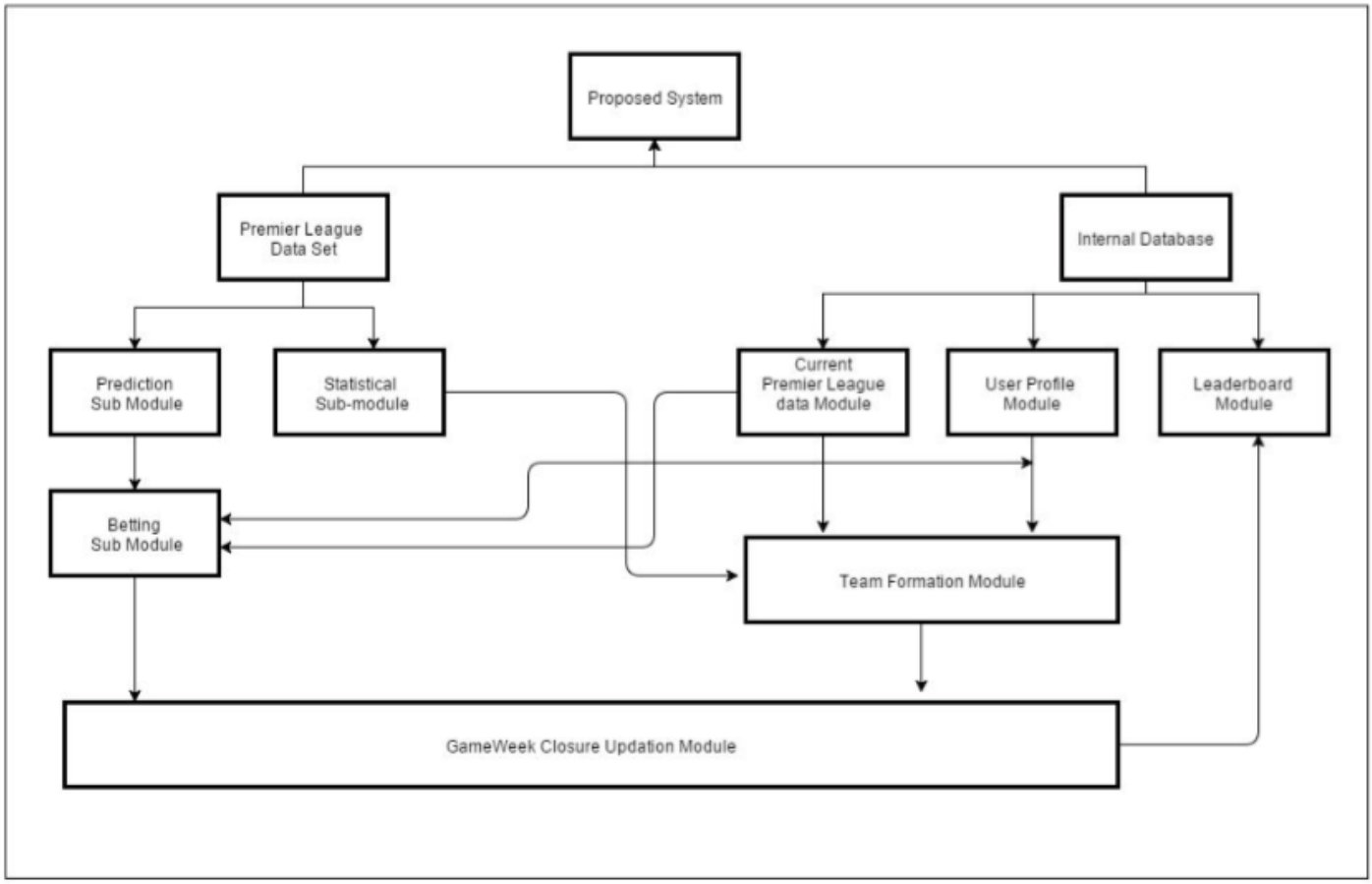

Fig 1: System Block Diagram

5. IMPLEMENTATION

The proposed system shall be added on top of the existing fantasy league system. It would developed on a web portal for which the following technologies would be used: Python, JavaScript, PHP, CSS, WEKA tool, Photoshop.

6. CONCLUSION

As seen in the above paper, and through surveys undertaken, the existing model of the Fantasy Football has the problem of churning of customers within the duration of the season. This problem is therefore addressed by using interactive models of Predictions where a user predicts the results of each game in order to be rewarded which would further help him strengthen his Fantasy squad. This is done using a data set of all statistics

of games played since the inception of the Premier League in 1992. Using GP function on this dataset, we can successfully predict the outcome of matches and thus award points to users based on their predictions. The project thus, aims not only to attract more users to this game that is Fantasy Football, but also aims at improving the general attraction to the Premier League. This happens because in a predictive model, a user makes a prediction on every game, and ends up watching that game to check if his prediction is going right.

Thus our project will not only improve the existing system of Fantasy Football, but will also augment the reach of Football in India.

7. REFERENCES

[1] Kou-Yuan Huang and Wen-Lung Chang. A neural network method for prediction of 2006 world cup football game. In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1 –8, july 2010.

[2] J. Hucaljuk and A. Rakipovic. Predicting football scores using machine learning techniques. In MIPRO, 2011 Proceedings of the 34th International Convention, pages 1623 –1627, may 2011.

[3] A. Joseph, Norman E. Fenton, and Martin Neil. Predicting football results using bayesian nets and other machine learning techniques. Knowledge-Based Systems., 19(7):544–553, 2006.

[4] A. Tsakonas, G. Dounias, S. Shtovba, and V. Vivdyuk. Soft computingbased result prediction of football games. In V. Hrytsyk, editor, The Ist International Conference on Inductive Modelling (ICIM'2002), pages 15–23, Lviv, Ukraine, 20-25 May 2002.

[5] Kou-Yuan Huang and Wen-Lung4 Chang. A neural network method for prediction of 2006 world cup football game. In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1 –8, july 2010