Barclays Premier League winning team prediction Using Machine Learning

Monirul Islam Pavel, Afrin Akther, Israth Chowdhury, Shahraiar Hasan Parash , Akash Biswas
Department of Computer Science & Engineering
BRAC University
66 Mohakhali, Dhaka, Bangladesh.
E-mail :{mipavel07,afrin.jhumi,chowdhury.israth,shahriarhasan0,bbiswasakashh}@gmail.com

***Abstract*—Predicting the results of Barclays Premier League matches have become an interesting challenge due to the fact that the sport is so popular and widespread. However, predicting the winning are also a difficult problem because of the number of factors which must be taken into account that cannot be quantitatively valued or modeled. With a view to develop of the system, a number of tests have been carried out in order to determine the optimal combination of features and classifiers using logistic regression,svm, xgboost with python libraries like pandas and scikit-learn.The results of the presented system show a satisfactory capability to predict the winning team.**

Keywords—Score Predicting; Xgboost; Regression; Machine Learning

I. Introduction

Sporting events have always been very interesting to a wide range of population. One of the most popular sports is football and Barclays Premier League is the toughest and the most prestigious club football league competition in the world.Three types of outcome can come which are home team winning, draw or away team winning . Due to its popularity and the small number of possible outcomes of games, predicting results is a very interesting and seemingly simple challenge. However, it is very difficult to predict the final outcome because the way the team plays on a particular day depends on many factors, such as the current form,the last team meetings, rivalries, offensive and defensive skills, individual abilities of key players and even the psychological impact of fans in the stands. Football is a game where the average sum of scored goals is pretty little (two to three per game), which means that a moment of brilliance or stupidity of an individual can decide the final outcome. For this reason, it is a big challenge to choose the features and the way of classification which would facilitate the prediction.

II. Literature Review

Snigh T. and et. al [1] presented two methods that have been executed by using Linear Regression Classifier and Naïve Bayes Classifier for first innings and second innings respectively. Two methods, 5 over intervals have been formed from 50 overs of the match and at each interval above mentioned attributes have been noted for all non-curtailed matches played between 2002 and 2014 of every team separately. Error has been found in the results that error in Linear Regression classifier which is lesser than the Current Run Rate method in evaluating the final score and also accuracy of Naive Bayes in predicting match outcome has been 68% starting from 0-5 overs to 91% till the end of 45th over.

McCabe A. and et. all [2] displayed an extension by using of artificial intelligence to predict sports result. An expanded model is explained, as well as a broadening of the area of application of the original work. The model usedis a form of multi-layer perceptron and presented with a number of features which helps to capture the quality of various sporting teams. The system performs well and compares which is suitable for human tipsters in several environments. A study of less rigid “World Cup” formats appears, along with extensive live testing results in a major international tipping competition.

Joseph A. and et. al showed impressive results for BNs (Bayesian networks) [3] that, in a number of key respects, the study assumptions place them at a disadvantage. the system has been assumed that the BN prediction is ‘incorrect’ if a BN predicts more than one outcome as equally most likely (whereas, in fact, such a prediction would prove valuable to somebody who could place an ‘each way’ bet on the outcome). Although the expert BN has now long been irrelevant (since it contains variables relating to key players who have retired or left the club) the results here tend to conWrm the excellent potential of BNs when they are built by a reliable domain expert. The ability to provide accurate predictions without requiring much learning data are an obvious bonus in any domain where data are scarce. Moreover, the BN was relatively simple for the expert to build and its structure could be used again in this and similar types of problems.

III. Proposed Methodology

Aiming to predict the winning team data has been pre-processing and data modeled after collecting the past

datas. As the system is designed to predict either the winning team is home or away or it is a draw, so it is a multi-class classification problem. After being splitted into training and testing dataset, the processed data are classified using support vector machine, logistic regression, extreme gradient boosting (xgboost) and short out the best algorithm to predict the winning team with best accuracy.

A. Data Collection

For the proposed method data of past years Barclays Premier League is gained from http://www.football-data.co.uk/ which has datasets of several years. The dataset contains League Division,Match Date,Home Team,Away Team,Full Time Home Team Goals,Full Time Away Team Goals,Full Time Result (H=Home Win, D=Draw, A=Away Win),Half Time Home Team Goals,Half Time Away Team Goals,Half Time Result (H=Home Win, D=Draw, A=Away Win) and so on.

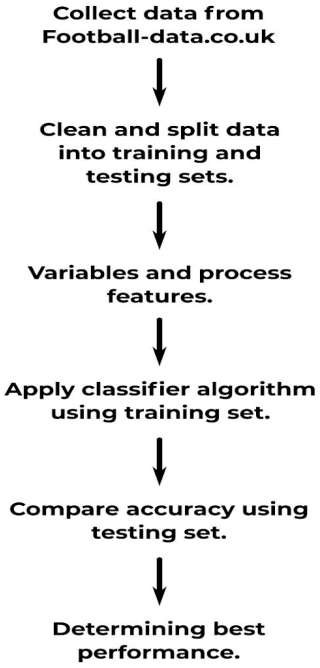

Figure 1 : Work Diagram of Proposed Model

B. Data Pre-processing

The accuracy of the outcome are highly depends on the pre-processed data model . In Barclays premier league there always two teams- home team and away team. There are bunch of different statistics in a game. By using those features we can predict our goal. For this first we need to clean our dataset which means make sure we are using those features that we need. here we will take into consideration only one feature that is FTR(full time result). As this is not a binary classification problem which only give results for win or lose rather it is a multi-class classification problem which have three possibles outcomes for which team will win- home(h), away(a) or there will be draw(d).

C. Classification

Logistic regression can analyze a dataset containing one or more independent variable to measure an end result. The result is determined with a dichotomous variable and we can get our desirable outcome which is win or lose situation of the football team. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the prediction. XGBoost stands for eXtreme Gradient Boosting, it was developed by Tianqi Chen and now is part of a wider collection of open-source libraries developed by the Distributed Machine Learning Community (DMLC). XGBoost is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources for tree boosting algorithms.

IV. Result and Analysis

The aim to predict the winning team either they are home or away team or it's draw. After loading the data the first thing is to sort out the total number of games, home winning, away winning and draw in a team home ground (fig. 2)

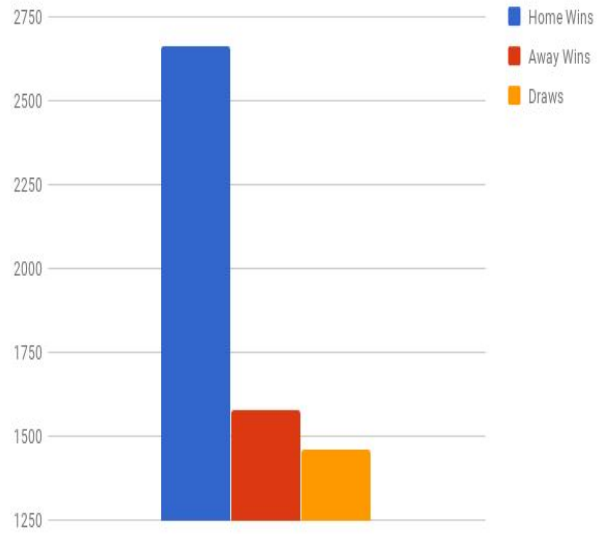

Figure 2 : Total winning strategy

Scatter diagram is plotted to visualize the negative and positive correlation of HTGD,ATGD,HTP,DiffFormPts, DiffLP.

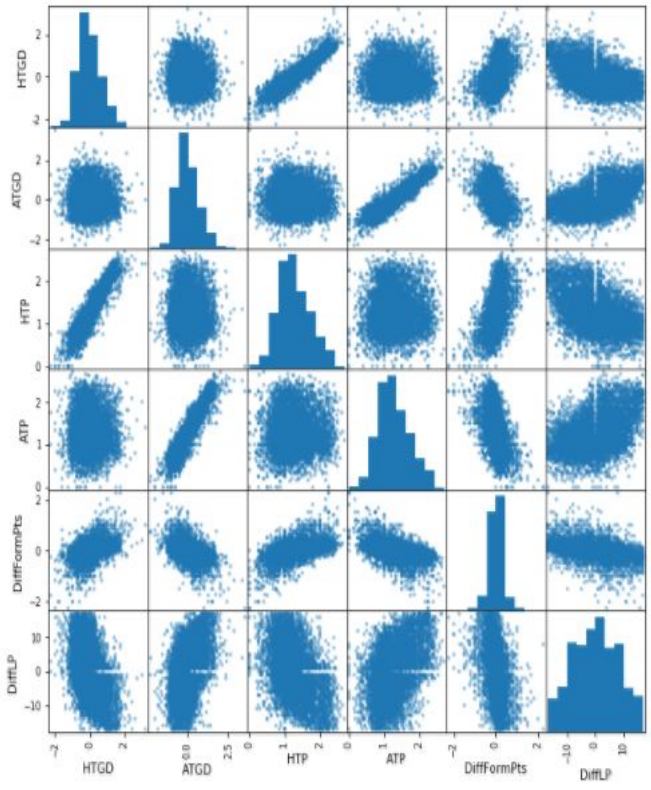

Figure 3 : Scatter Plotting(Negative and Positive

Before doing multi-class classification, the system only takes full time match results and drops all features. based on this several classification algorithms after spliting into training and testing datasets , are applied and compared. Finally, The system detect that xgboost works best with accuracy for f1 score 72.53% and in accuracy score 74%

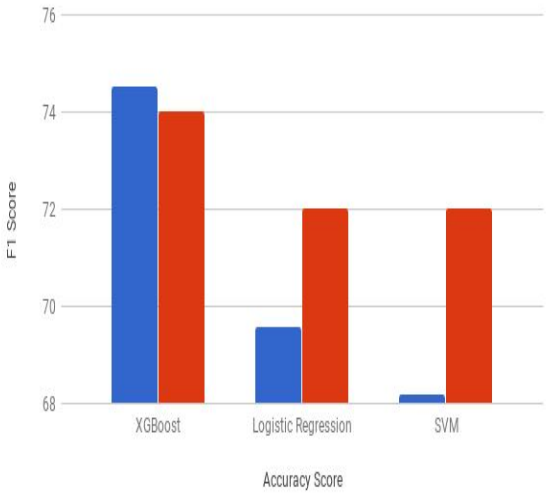

Figure 4 : F1 vs Accuracy Score

V. Conclusion

Predicting the results of football matches poses an interesting challenge due to the fact that the sport is so popular and widespread. However, predicting the outcomes is also a difficult problem because of the number of factors which must be taken into account that cannot be quantitatively valued or modeled. As part of this work, a software solution has been developed in order to try and solve this problem. During the development of the system, a number of tests have been carried out in order to determine the optimal combination of features and classificators. The results of the presented system show a satisfactory capability of prediction which is superior to the one of the reference method (most likely a priori outcome). The goal set at the start of this project (achieving accuracy around 60%) was greatly surpassed so from that point of view we can consider this project successful. Of course, there is room for further improvement, primarily in the area of feature selection. If we were to model the form for each and every player in the match we could probably achieve better results. This way we could monitor each players form during the season and determine its influence on the final score. Larger data set for learning would also help to predict future outcomes

References

[1] T. Singh, V. Singla and P. Bhatia, "Score and winning prediction in cricket through data mining," 2015 International Conference on Soft

Computing Techniques and Implementations (ICSCTI), Faridabad, 2015, pp. 60-66.doi: 10.1109/ICSCTI.2015.7489605

[2] A. McCabe and J. Trevathan, "Artificial Intelligence in Sports Prediction," Fifth International Conference on Information Technology: New Generations (itng 2008), Las Vegas, NV, 2008, pp. 1194-1197.doi: 10.1109/ITNG.2008.203

[3] A. Joseph and N.E. Fenton and M. Neil, “Predicting football results using Bayesian nets and other machine learning techniques ” Jelsevier B.V., 2006.

[4] S. J. Russell and P. Norvig, ‘’ Artificial Intelligence A ModernApproach’’, 1. Edition, 1995.

[5] J. Friedman and T. Hastie, R. Tibshirani, “Additive logistic regression: a statistical view of boosting”, Annals of Statistics, 2000.

[6] L. Breiman, “MACHINE LEARNING -Random Forests”, Kluver

[7] Academic Publishers, volume 45, 5-32, 2001

[8] K. Elissa, “Title of paper if known,” unpublished.

[9] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[10] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[11] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.