Football Prediction using XGBoost Algorithm: A Literature Review

Wasim Gourh[1], Keshav Poojary[1], Mallika Vengarai[1], Nida Parkar[2]

[1]Department of Computer Engineering, Student,University of Mumbai, Atharva College of Engineering, Malad, Mumbai, India
[2]Department of Computer Engineering, Assistant Professor,University of Mumbai, Atharva College of Engineering, Malad, Mumbai, India

Publication Info

Article history:
Received : 10 February 2020
Accepted : 25 May 2020

Keywords:
EPL,PCA RPS

Corresponding author:
Wasim Gourh
e-mail: *wesgourha@gmail.com*

Abstract

Sports is one of the most prevalent activities of all time. About half of the world has been interested in different activities, soccer or soccer, as it is commonly known. Football is now an activity with massive venture resources and annual sales of billions, not just in terms of sports. In the last year, the Premier League itself has raised more than $1 billion. Because betting is now permitted in most European nations, citizens continue to participate week after week. Betting companies have their predictors that are the basis for betting scores. If people are willing to resolve these challenges, they will indeed help take a huge deal. Although there may be three outcomes to win, lose, or draw results in a soccer match, it may be difficult to predict these outcomes. The objective of this paper is therefore to create a model that aims to overcome these odds utilizing limited data.

1. INTRODUCTION

1.1. Need

Sport is one of the most followed events in this world and football is right on top of the list. The world cup, which takes place every four years, comes with lots of surprises. So much so, that in 2010 it was not the nations but a mere octopus that was able to predict the winner of each match successfully. Prediction of the football match has always been something that fans tend to do before matches be it for fun, enjoyment or even betting to make easy bucks. So, as a result, various models were developed to predict the match results. Current models estimate the results roughly with an accuracy of about 50%. We in this paper are trying to improve the current models developed by individuals online.

1.2. Basic Concept

Football is a sport where the duration of the matches is fixed and the team either loses, wins or draws. In past models, it has been observed that people have developed algorithms to find out how well a team can score goals. These algorithms were used to conclude results. In recent times, the goals being scored by teams are fewer as they are turning towards the defensive approach. Also, the result of the match can be decided by the factor whether the team's star player is in the lineup or not i.e. the impact of an individual player on the match. So, in this paper along with team stats we have tried to incorporate individual player's stats and how that player can impact the team's overall performance and change the result.

2. RELATED WORK

2.1. Bayesian hierarchical model for the prediction of football results

Authors: G. Baio and M. Blangiardo.
In recent years, the issue of football modeling has become more and more common and several various models have been introduced to predict the features that lead a team to lose or win a game or to forecast a result for a specific game. To meet these two objectives and check the Bayesian hierarchical model focused on data from the 1991-1992 Italian Series A championship. They recommend a more complex blend model that matches the observed data to solve the over-reduction problem generated by the Bayesian hierarchical model. The Italian Series A championship 2007-2008 is an illustration for checking its results.

2.2. Predicting football scores using machine learning techniques

Authors: J. Hucaljuk and A. Rakipovid

The key aim of the paper is to test numerous methods for machine learning to forecast the outcome and result of football matches by utilizing in-game match activities rather than the number of goals scored by each side. They have tested different model architecture theories and analyze the efficiency of their models against benchmark techniques. In this paper, they have established an' anticipated objective' measure to help us assess the success of a team rather than use the specific achieved goals. This measure is paired with the measurement of an offensive and defensive team ranking update during-game to construct a classification model that

predicts the outcomes of future matches and a regression model that predicts future matches. The efficiency of their models correlates well with the current mainstream strategies and is close to that of bookmakers.

2.3. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach

Authors: N. Tax and Y. Jousts

A framework for the Dutch Eredivisie focused on public data is defined in this article. A systematic literature review described the variables of predictive utility for the match outcomes. Candidate characteristics have been created. Self-made public data collection, consisting of 13 Dutch Eredivisie match data seasons, was accompanied by modeling preparation. A variety of variations have been evaluated on public data testing developed in the dimensional reduction techniques and classification algorithms. A mixture of PCA (with a difference of 15 percent) and a Naive Bayes or Multilayer Perceptron classifier obtained the best detection precision for the public data feature collection. Models for betting odds and a hybrid feature set (common data union and wagering odds features) have been created. Check McNemar has found no substantial gap in the accuracy of the model with the lowest accuracy hybrid function setting and the low precision betting odds, but the findings do lift the supposition that a mixed combination of betting odds and public data will defeat the bookmaker. The results obtained can be seen as a positive sign that competitive structures for supporting betting decisions based on open data can be created.

2.4. Predictive analysis and modelling football results using machine learning approach for English Premier League

Authors: Baboota, Rahul & Kaur, Harleen

This paper demonstrates their research in developing a common statistical model for the English Premier League games. They have built a feature collection using software engineering and an exploratory data review, which evaluates the main factors for predicting football match outcomes and thus develop a highly detailed predictive framework by machine learning. They have demonstrated that their model's success is highly based on important characteristics. In the EPL aggregated during two seasons (2014-and 2015–2016) their best model with the gradient boosts achieved a performance of 0.2156 in the probability (RPS) metric for the game week 6 to 38 whereas, the betting organizations they consider (Bet365 and Pinnacle Sports) received RPS value of 0.2012 for the same period. Because the low RPS value reflects a higher predictive accuracy, given encouraging performance, their model did not surpass the forecasts of the bookmaker.

2.5. Effects of expertise on football betting

Authors: Khazaal, Y., Chatton, A., Billieux, J

The goal of this analysis was to determine whether football experts could forecast football match scores than non-experts. The precision of football match prognoses does not seem to be affected by experience, age, and sex. The assumption that soccer expertise enhances betting skills is, therefore, nothing more than a cognitive delusion known as the "illusion of control;" gamblers may profit from psychological therapies that work on the illusion of control ties that their perceptions have between betting skills and soccer expertise. The practice that needs to be taken into account is the public safety agenda to discourage football gambling.

2.6. Using Twitter to predict football outcomes

Authors: Kampakis, Stylianos and Andreas Adamides

The objective of this investigation was to determine how Twitter data can be used. They have built a model based on tweets and have examined whether these models can overcome predictive models using solely historical data and simple soccer statistics, for the outcome of football matches in the English Premier League. Moreover, both Twitter and historical data are used to construct combined models. The final Twitter-based model performs significantly better than chance when calculated by the Cohen Kappa and is equivalent to a model that uses simple statistics and historical data. The convergence of the two models increases performance better than the other model. This work also shows that technologies obtained from Twitter can theoretically provide valuable information to predict soccer outcomes.

2.7. An Introduction to Logistic Regression Analysis and Reporting

Authors: Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll

This article shows the preferred pattern for using logistic methods with an example of the logistic regression used for a data set to test a research hypothesis. Reasonable reporting formats and minimum observationto-predictor ratio for logistic regression tests are also suggested. In eight papers released in The Journal of Educational Research between 1990 and 2000, the scientists discussed the application of and understanding of functional regression. All 8 reports also shown that the prescribed requirements were fulfilled or surpassed.

2.8. Predicting football match results with logistic regression

Authors: D. Prasetio and D. Harlili

The model is developed from 2010-2011 up to 2015-

2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)

2016 using variations of training data. Logistic regression is a classification tool used to forecast sports outcomes that may offer empirical insight by coefficients of regression. The criteria used are "Home Offensive," "Home Defense," "Forgotten," and "Forgotten." By modifying the training data used, they performed experiments. The statistical precision of the model being developed is 69.5%. They arrived at the conclusion that "Home Defense" and "Away Security" were the relevant variables.

2.9. Goal detection in football by using support vector machines for classification

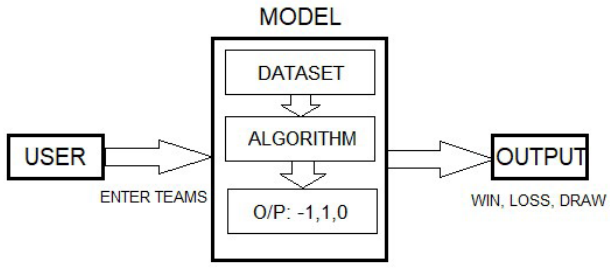
Authors: N. Ancona, G. Cicirelli, A. Branca and A. Distante

In this article, they have tackled the question of achieving a target during a football match utilizing strategies and devices that neither alters the ball nor the goalmouth. Automatic goal identification in football is a popular problem and the referee associations, the press and fans pay special attention to it. Sometimes, the official and the workers (linemen) are not able to discern the goal or, what is more significant, they do not assign the award accurately to either side. There are several unusual cases in which the goal happens. The occurrence happens by detecting the ball and measuring its place on the location of the postage in the photograph. The strategy for ball detection is based on a controlled classification method named vector aids. Examples are used for training the correctly filtered and previously recorded picture pattern variant of the views of the entity to be detected. On actual photographs in which the ball is both fully transparent and partly occluded, they have checked thoroughly the methodology. Detection rate, false-positive and precise ball location in the image measure performance of the proposed detection scheme.

2.10. An Implementation of Naive Bayes Classifier

Authors: Yang and Feng-Jen

In order to decide the most appropriate classification for the specified data in a problem field, the Naive Bayes classification consists of a collection of probabilistic calculations. This paper explains the application of the grouping of Naive Bayes. This framework will be viewed as a generic resource kit for specific classified areas. A reference data collection is chosen to check this definition for the precision of all probabilistic estimates concerned.



Block Diagram

3. PROPOSED METHODOLOGY

The proposed system or model for the prediction will use the XGBoost algorithm as the primary algorithm for the work. The algorithm will be tested against many other algorithms to see how it works in comparison with them. The data set used will be the same among all algorithms. The dataset will be filtered out to seek out only the important data that affects the match predictions greatly. After selecting the desired data set after few testing and different combinations to find the perfect data set, we'll perform optimization or normalization on the so to see if it affects the outcome of the prediction in any way. The predicted output will be in form of certain numerical values which will represent the three natural outcomes of the game respectively i.e. win, draw or loss. These values will be used to produce the proper output in a human language like which team has won or lost.

4. WORKING

First, the dataset will be filtered and only a few of the features will be selected among them. This dataset will be used to train the model over the period. The model will be trained over different combinations of the data set to find out the best combination. Then after selecting the suitable data set format for the model and trying out the various combo, the result will be converted into 3 values using a small code, the values being -1,0,1 each representing loss draw and win respectively as seen in Figure 1. The user will be provided with an option to enter the teams for the matches were in the first team entered will be home and later away. After entering the user will be provided output in a simple English language i.e. win, lose or draw.

5. CONCLUSION

The proposed model could perform better than many other models using only the publicly available data due to the use of XGBoost algorithm. This will benefit people who are interested in betting to see how the results turn out to be in comparison to the odds being set by the bookies.

Also, it can be used in journalism or for the pre-match or post-match talks based on the results it generated.

6. REFERENCES

[1] G. Baio and M. Blangiardo. "Bayesian hierarchical model for the prediction of football results." University College London Department of Statistical Sciences, Gower Street, London WC1 6BT

[2] J. Hucaljuk and A. Rakipovid. "Predicting football scores using machine learning techniques." University of Zagreb, Faculty of Electrical Engineering and Computing Unska 3, 10000 Zagreb, Croatia

[3] N. Tax and Y. Jousts. "Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approwh."

[4] Baboota, Rahul & Kaur, Harleen. (2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. 35. 10.1016/j.ijforecast.2018.01.003.

[5] Khazaal, Y., Chatton, A., Billieux, J. Effects of expertise on football betting. *Subst Abuse Treat Prev Policy* **7,** 18 (2012). https://doi.org/10.1186/1747-597X-7-18

[6] Kampakis, Stylianos and Andreas Adamides. "Using Twitter to predict football outcomes." *ArXiv* abs/1411.1243 (2014): n. pag.

[7] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[8] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), George Town, 2016, pp. 1-5.

[9] N. Ancona, G. Cicirelli, A. Branca and A. Distante, "Goal detection in football by using support vector machines for classification," IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), Washington, DC, USA, 2001, pp. 611-616 vol.1.

[10] Yang, Feng-Jen. (2018). An Implementation of Naive Bayes Classifier. 301-306. 10.1109/CSCI46756.2018.00065.

AUTHORS

Wasim Gourh
Student Atharva College of Engineering

Keshav Poojary
Student Atharva College of Engineering

Mallika Vengarai
Student Atharva College of Engineering

Nida Parkar
Assistant Professor
Atharva College of Engineering

2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)