Soccer Analytics using Machine Learning

Abha Tewari, Asst. Prof, VESIT
Tushar Parwani, VESIT, Mumbai
Ajinkya Phanse, VESIT, Mumbai
Akshay Sharma, VESIT, Mumbai
Anush Shetty, VESIT, Mumbai

ABSTRACT

Sports Analysis is rapidly growing area of sports science with the ever increasing easy internet accessibility and recognition of Machine Learning. This can be a motivating space of analysis for soccer, as soccer is considered way more complicated and dynamic when put next to a couple of different sports. Additionally its the world's most liked sport, played in over two hundred countries. Many methodologies, approaches and measures are being taken to develop prediction systems.The paper is developed to predict the outcome of the matches in English Premier League(EPL), by studying the trends from the previous matches and identifying the foremost vital attributes that are required to accurately predict the result. XGBOOST, SUPPORT VECTOR MACHINE and LOGISTIC REGRESSION models were taken into consideration and chosen the most effective among them to build the prediction model. This model is applied on real team information and fixture results gathered from http://www.football-data.co.uk/ for the past few seasons.

Keywords

Football, Prediction, Machine Learning, F-SCORE

1. INTRODUCTION

Prediction systems have proven to be of importance in a range of fields like stock markets, sports, on-line searching, and so on. In sports, these systems can be especially useful for coaches to investigate the performance of the squad, enhance their game setup, etc. Sports' card-playing conjointly has been growing in integer rates over the past few years. As a result, Machine Learning is presently a extremely trending approach. For the prediction of the likely outcome of the most-watched football event, numerous simulations were performed and three modelling approaches adopted : Poisson regression models, random forests, and ranking methods .To model the previous scores of the competing teams as (conditionally) independent variables, Poisson regression approach was used.With love for the game and inspiration from these researchers, decision was took to predict the results of football matches in the Barclays Premier League, that is hailed to be the foremost exhilarating league of soccer within the world. The League operates on a promotion and relegation basis with twenty groups competing with one another to accomplish their ambitions as a club. Before continuing to the most important section of the paper, need to review a couple of previous works during this field.

2. LITERATURE SURVEY

By observing the results from paper [1], understood the techniques to improve the efficiency of the prediction. Traditionally many models have been built to predict the results using goals scored by each team as a metric. Using the paper, got to know the inconsistencies introduced by it. But this paper has used a "expected goals" metric which takes into consideration the teams performance rather than just the goals scored. By analysing the devised of techniques to clean the dataset and introduce new attributes that would provide in depth metrics for accurate prediction of the winning team.

In paper [2], authors have proposed a logistic regression model to estimate 2015/2016 Barclays' Premier League match results with an accuracy of around 69.5%. They develop this model with the help of data from Barclays Premier League and sofifa.com using four significant variables: Home Attack, Home Defense, Away Attack, and Away Defense. They implement this method in software called Football Predictor. Their work predicts who is going to win a match (home/away), and list out details regarding the odds and probability, and the coefficients of regression. This model comprises of just four variables but gives strong prediction accuracy. Various techniques have been utilized to develop result prediction systems. In explicit, football match result prediction systems have been developed with techniques such as artificial neural networks, naive Bayesian system, k-nearest neighbor algorithms (k-nn), and others.The choice of any technique depends on the application as well as the feature sets. The priority of a system developer or designer in most cases is to get a high prediction accuracy. The objective of [3] study is to investigate the performance of a Support Vector Machine (SVM) with respect to the prediction of football matches.The findings showed 53.3% prediction accuracy.From [4] the conclusin comes that the XGBoost is quite an efficient algorithm for predicting the results. In [4] the results obtained were quite up to the mark as the only parameters considered were ranking and points data. But, if the algorithm runs efficiently for the set of attributes needs to be tested.

In the paper [5], the authors study multiple techniques in data mining and their prediction results are correlated to devise a good model for predicting matches of the Dutch football team. They use three major models namely Generalized Boosted Models (GBM), K-nearest neighbor and Naive Bayes classification. Using GBM, they attained 60.22% accuracy on average, while the other models were not as accurate. The results of the paper were based on a data-set that only included information about the Dutch team but no data regarding the opponent except for their FIFA ranking. To further improve this research, more data and statistics could be taken into account such as the opponent team's overall form in that season, and other factors such as head-to head results or information about each team's previous games.

Different evaluation processes gauge different characteristics of machine learning algorithms. The factual evaluation of algorithms and classifiers is a matter of on-going debate amongst researchers. Most measures in use today focus on a classifier's ability to identify classes correctly, [12] Note other useful properties, such as failure avoidance or class discrimination, and it suggest measures to evaluate such properties. The measures named Youden's index, likelihood, Discriminant power are used in medicinal diagnosis. It also lists other learning problems which may benefit from the application of these measures.

3. PROPOSED METHOD

First of all, the Dataset is passed to a system. Now this Dataset will be going through Cleaning Process and hence use the data to find the correlation between parameters by plotting the scatter plot define key attributes which will be used using Jupyter Notebook.This cleaned Dataset will be given to the system along with the upcoming matches of all the teams.Using the algorithm like XGBoost, Scikit Learn and Pandas library, the data was used to calculate the past records of the specific teams.Future match schedule will be provided to the system.Now this Past record analysis will be used to calculate the WIN percentage or WIN Prediction for the future matches of the desired teams.The final output or target label is the Full time Match Result (FTR). This label will indicate a Home team win (H), an Away team win (A), or a Draw (D).

4. ARCHITECTURE/DESIGN

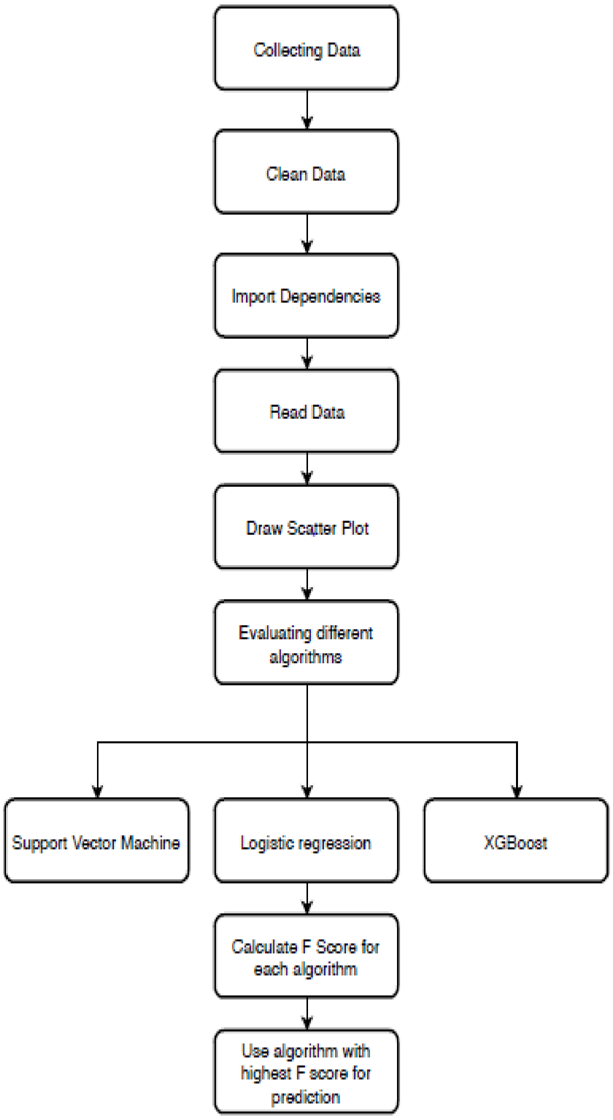

Fig. 1 Architecture Diagram

In the above fig. 1, it shows the complete flow of the system or can say, the complete procedure in a system. As discussed earlier, the initial phase will be data collection and some algorithms will be used to check the F score of each of the algorithm and the algorithm giving the highest F score will be used for the Prediction purpose. In this way, the predicted output will be most accurate as compared to all other algorithms.

A. Dataset Description

The dataset was extracted from http://www.football.co.uk/ which contained of 35 attributes which were refined and cleaned and as a result obtained 8 attributes which helped in predicting the outcomes of the matches with efficiency. Also in the dataset, the attribute for the team form was incorporated by considering the results of the team for the past five matches. This introduces a factor of the updation of the team form to predict results on the current status of team and not on past statistics which are of no importance now.

B. Pre-Processing

The data-set was obtained from football.co.uk consists of many attributes from every season. Most of them are just unnecessary and just complicate the process of prediction. Hence, the primary task is to scrub the data-set and to solely retain the options or attributes that are needed the foremost. To check the correlation of the attributes with each other, scatter matrix was plotted. This helps in visualising the relation of each attribute with every other. For those attributes in which no variation is seen when the other attribute is varied can be ignored.

C. Data Splitting:

Taking all available labeled data, and splitting it into training and testing subsets, with a ratio of 70-80 percent for training and 20-30 percent for testing. The Machine Learning system uses the training data to train models to see patterns, and uses the testing data to evaluate the predictive quality of the trained model.

D. Model Building

For building the model for prediction we consider various classification techniques and compute the F-SCORE for each of them which provides an excellent base for comparing their prediction accuracy.

F-SCORE:

F-SCORE is a measure of the accuracy of the test. It considers two parameters namely precision(p) and recall(r). p is the no. of correct positive results divided by the no. of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that are ought to be known as positive). F-SCORE gives the harmonic average of the precision and recall, having best value as 1 and worst value as 0.

F-SCORE = [(p^(-1)+r^(-1))/2]^(-1) = 2(p*r)/(p+r)

Logistic Regression: Logistic regression seeks to design the possibility of an event occurring based on values of independent variables, that could be categorical or numerical. It is a statistical method that operates on data-sets having one or more independent variables which decide an outcome. The objective of this method is to compute the perfect fitting model that describes the interrelationship amidst the dependent variable and a list of independent (predictory) variables. [2] problem is a multi-class classification problem as there exists more than two possible outcomes i.e, Home Win, Draw and Away Win. Hence we are going to be using Multinomial Logistic Regression.

Support Vector Machines: Support Vector Machines are models in Machine Learning that is useful for regression analysis and classification tasks. Mapping each data item as a point in a space of n-dimensions (n being the number of features) in which each feature-value corresponds to a particular coordinate. The target is to obtain a hyperplane that classifies all training vectors into two classes. The finest choice is the hyper-plane that leaves the maximum margin from both the classes. [3].

XGBoost: XGBoost algorithm is used to develop a predictive model that is based on an ensemble of decision trees. XGBoost is an algorithm that is on the rise recently dominating applied machine learning for tabular or structured data. This algorithm is an application of gradient-boosted decision trees designed for performance and speed. [4]

F-SCORE for three classification techniques i.e Support Vector Machine, Logistic Regression and XGBoost was computed. After comparing the accuracy of all the three techniques, it was observed that XGBoost to have the highest F-SCORE, thereby providing the highest accuracy for prediction. Hence, XGBoost is used to build the model.

5. CONCLUSION

It is possible to make predictions in the sports and specially in Football, also called as Soccer as it is one of the highest paid or also can say highest fan-following sport as it has fan-following all around the world.Modelling prediction in sports, soccer in particular is tasking and very intricate but this can be made even simpler and unchallenging by the implementation of machine learning.Thus, can conclude that the process of prediction can be simplified and made efficient by the slashing down of unwanted attributes that contribute to increased complexity. Also they have incorporated the team form giving our predictions the edge to make predictions only on the current statistics of the team.

It can be concluded that the Analysis and Prediction for the future event can be successfully done using the past records and data of that event with the highest accuracy rate.

6. REFERENCES

[1] CORENTIN HERBINET.Predicting Football Results Using Machine Learning Techniques.IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE.

[2] Darwin, P & Dra, H (2016). Predicting Football Match Results with Logistic Regression. 2016 International Conference On Advanced Inform

[3] Support Vector Machine–Based Prediction System for a Football Match Result Chinwe Peace Igiri (Computer Engineering, Rivers State College of Arts and Science, Nigeria)

[4] https://medium.com/@viv.mishra04/aspiring-octopus-predicting-world-cup-2018-results-using-xgboost-d18479462a46

[5] Abel Hijmans. Dutch football prediction using machine learning classifiers (unpublished)

[6] Hucaljuk, J & Rakipović, A. (2011). Predicting football scores using machine learning techniques. 2011 Proceedings of the 34th International Convention MIPRO,, 1623-1627.

[7] BARCLAYS PREMIER LEAGUE, 2014.League Table Barclays Premier League current & previous standings [online] Barclays Premier League. Available from:http://www.premierleague.com/engb/matchday/league-table.html[accessed 1/17/15].

[8] MALARIĆ R, KATIĆ T, SABOLIĆ D. 2008“The market efficiency of the soccer fixed odds internet betting market”. Applied Economics Letters [serial online]. 15 (3), pp. 171-174.

[9] POMORSKI D., PERCHE, P.B. 2001“Inductive learning of decision trees: application to fault isolation of an induction motor”.Engineering Applications of Artificial Intelligence, 14 (2), Pages 155-166

[10] Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches.

[11] Predicting the winner of NFL-games using Machine and Deep Learning

[12] Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation.