WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predicting Premier league Match odds using Machine Learning

Anish Yekhande[1], Yadav Deepak Kumar[2], Kanojiya Sanjay[3], Ashwini Phalke[4]

[1]Student, Dept. of I.T. ,VPPCOE & VA,Mumbai University,Mumbai,India
[2] Professor, Dept. of I.T. ,VPPCOE & VA,Mumbai University,Mumbai,India

ABSTRACT-

Sports Analytics is a leading industry and one of the best real-world applications of Data Science.Many techniques to predict the outcome of professional football matches have traditionally used the number of goals scored by each team as a base measure for evaluating a team's performance and estimating future results. Although, the number of goals scored during a match possesseslarge inconsistencies in many games between a team's performance and number of goals scored or conceded. The main objective of this project is to explore different Machine Learning techniques to predict the odds offootball team, using the last year standingsand the number of goals scored by each team until half-time. We will traversevarious model design hypothesis and assess our models performance against standard techniques. model predicting the outcome of future matches, as well as a regression model predicting the score of future games. Our models' performance compare favourably to existing traditional techniques and achieve a similar accuracy to bookmakers' models.

***Key Words*:**Data Science , Machine learning

1.INTRODUCTION

Football is very famous sport in the world.The Governing body of football, the InternationalFederation of Association Football (FIFA),estimated that at the turn of the 21st century there were approximately 300million football players and over 1.5 billion people interested in popularity of this game created huge number of people's association.Hence , estimation of match result in advance is very attractive to the experts and re-searchers. But it is very difficult to guesstheresult of a football by expertsor past statistics. There are many factors who can influence the match outcomelike, skills, player combination, key players forms, teamwork, home advantageand many others. Prediction is much harder morewhen match stats are influenced by extra time or substitute players or injuries.

Football associated sports data is now publicly available and grown curosity in innovative intelligent system to speculate the outcomes of matches. In thelast twodecades, researchers proposed several advancedtechniques to predictoutcome using previous and available data. Most of the past research devise theprediction problem as a classification problem. The classifier forecast theresult with one class among win, loss or draw class.

With the exposure of powerful machine learning practises, the predictive performance of classification problems have been improved over last few years. This project explores different data preparation techniques and algorithms to predict the result of a football match in terms of odds of winning, losing or drawing. The proposed model needs to be feed the data regarding team ratings of the attacking anddefensive team and the goals scored by them at half time. This has been possible due to the huge dataset available publicly.The data for the research is of good quality and extracted from credible data sources. There are huge number of statistics available, however, for the ease of accessibility of the system, only a few would be considered for now. The main goal is to build a model capable of understanding team's historic performances and predicting the odds at the half-time of the concerned match. The important aspects of the research are building a training and testing pipeline that can compare the benefits of adding new features and using other models on the result

2. LITERATURE SURVEY

In this section, past research done on the subject is detailed along with the linkage to this research paper. The paper 'Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers of 2021 defined many parameters which can be used to assess a player and the success of a transfer. The research used algorithms like Random Forest, AdaBoost and Naïve Bayes to train and test the classifiers, which are some of the algorithms that would be used in this research also. Numerous experiments have been performed with different parameter weights and this forms a basis of experiments performed in this research. The paper 'A Machine learning Approach to Football Match Result prediction' by Luca Carloni

compared his proposed approach usingclassical machine learning algorithms based on features. Experiments were performed for selecting and rating features that would help in automatic forecasting of football match results. The inspiration of features used in this research is drawn from this paper. Another deep learning-based paper called 'Survey on predicting the winning football team using Machine Learning algorithms' by the same researcher excellently estimated the results of FIFA 2018 world cup by segregating data into training, test and validation. The paper narrowed down the research to one single league which is taken as reference for this research on English Premier League data in the similar manner. This system predicts whether the home team will win the match or not with the help of 3 different models: , Support Vector Classifer and Gradientboosting Classifiers trained and parameters are tuned for best performance. The best results were given by Gradient boosting Classifier, i.e., an accuracy of 70%. This was taken as a reference for the comparisons and research performed in this paper.

3. Dataset

A.Origin:

For this research, one of the most eminent international football leagues, Premier League or English Premier League was chosen. 20 teams take part in the league with each team playing around 40 matches throughout the season. The premier league was organisedfirst in 1990 and since then many new English clubs have come up, so, for this research, data from the year 2005 to 2020 was considered.

The data used in this project was extracted from the Football-Data.co.uk. The website inventory is quite large as it has all the results of five English leagues for every year starting from 2005. Seventeen different datasets were taken from the website so that the results of the year 2005 to 2020 and ongoing season2021-2022 act as the training data and the behaves as the test dataset. Every datasets have statistics about Endgame result and halftime results, match stats, total goals and Away/Home odds

B. Features:

In the dataset that we have used from the website there were a large number of features provided. Although , in the datacleaning step , many of them are removed and only the most easy ones to understand, features/labels are used. All is this feature segregation is important to create an application capable of predicting the results of English Premier League matches which can be used by anyone. So, for such an application, the input from the user needs to be minimum and easily available. Although for optimising the prediciton in future, more featuresshould beenumerated as more the data, better the model will predict to recognize patterns. For this research, the following features are considered

C.Data Pre-processing :

Goals scored by each team as well as point gained or lost are important parameters in football analysis as the final scoreboard is dependent on it. As Home-Team and Away-Team are categorical variables, for the model to understand this, the variables need to be one-hot encoded. As a result, a 40x40 binary matrix is created where 40 is the number of teamsthat ever participated in the Premier League from 2005 to 2020, i.e., every year only 20 clubs participate but they might not necessarily be the same. Now, every team isencoded as a unique string of 0s and 1s.

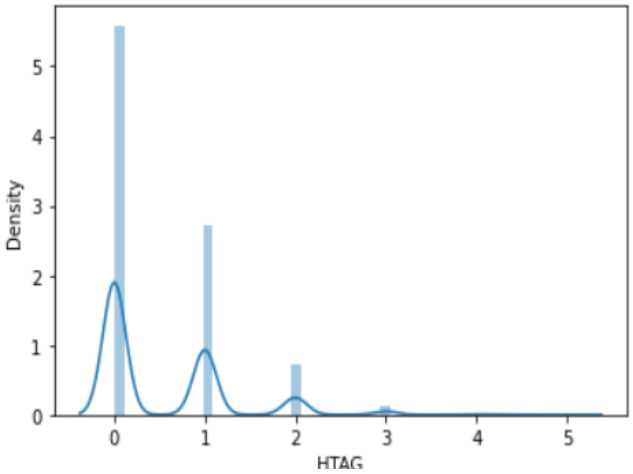

Fig 1 Density plots for HTAG

Name of the feature	Data Type	Definition
Home Team	Categorical object	Name of the home team
Away Team	Categorical object	Name of away team
HTHG	Integer	Half time Home goals
HTAG	Integer	Half time Away goals
FTR	Integer	Full Time Result
HomeTeamLP	Categorical Integer	Home Team Leader Position
AwayTeamLP	Categorical Integer	Away Team Leader Position

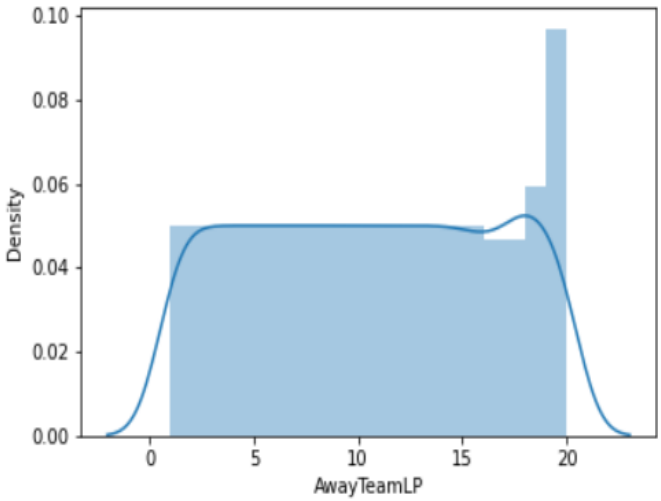

Fig 2 Density plot for AwayTeamLP

4. Analysis and Modelling

A.Exploratory Analysis:

The end task is to predict the winner of the football so first there is a need to dig out factors that might influence the aggregate win percentage.The first factor that can largely impact is the stadium or the ground. In the last 15 years, 48% of times a team has won when it
Is playing on the home ground , the figure 3 given below shows the aggregate win percentage of past 15 years of Premier league.

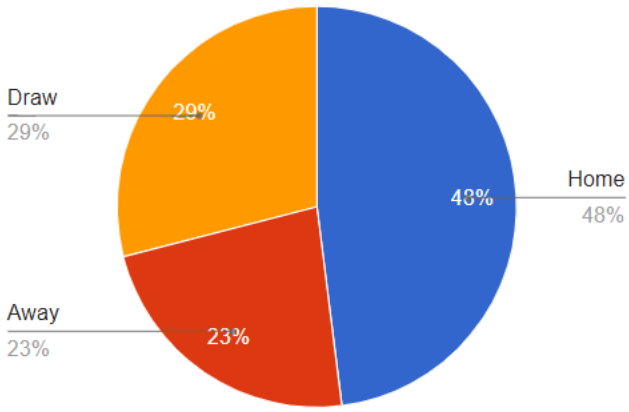

Fig 1 Classification of match wins from 2005 - 2020

As HomeTeam and AwayTeam are categorical variables,
for the model to understand this, the variables need to be one-hot encoded. As a result, a 40x40 binary matrix is created where 40 is the number of teamsthateverparticipated in the Premier League from 2005 to 2020, i.e., every year only 20 clubs participate but they might not necessarily be

B. Modelling and Tuning:

The goal is to predict the probability of both winning and loosing team and a draw.For this, it is essential to select an algorithm capable of not giving the bestclassification accuracy and then use that algorithm capable of not giving the best classification accuracy and then use that algorithm for predicting the probability using the predict_proba_function.As it is a multi-classification problem , the first approach was to use classification algorithms.For this the frequently used classification algorithms like XG Boost, Gradient Boosting Classifier , Logistic Regressor .
For validation , two different methods were considered:

- Random
The training set in this case consisted of 6060 matches from 2005 to 2020. While, validation set has any 20 matches apart from the training set.
- Sequential The training set in this case consisted of 6060 matches from 2005 to 2020 but in order. The validation set in this case consisted of the last 20 matches from the season 2020-2021

6 . Conclusion:

Sports analytics is an interesting yet sparsely explored area of machine learning because of the pre-requisite knowledge of the sport, its rules and key-performance indicators. Thus, the goal was to create a football match result predictor with least input from the user with the best
possible accuracy. In this paper, the data of one renowned league was taken into consideration however, the approachescan be extended to any football league, national or international. Here, by inputting only 6 features and implementing 3 state-of-the-art algorithms a satisfactory accuracy has been reached. Given that sports do not run by numbers but players and playing conditions, expecting a very high accuracy would not be possible. However, there are a huge number of statistical indicators and parameters that are left out from this research. In future, more experimentation will be carried out with extra features like results of previous five matches, shots taken, shots at target, fouls, etc. by the half-time to make themodelunderstand better. Not only this, the research can be improved by using neural networks and pre-trained
models. Apart from the winningprobability, other predictions can also be made such as expected goals and the goals at full time which would make it a regression problem

REFERENCE

[1] An Improved Prediction System for Football a Match Result by Igiri, Chinwe Peace1; Nwachukwu, Enoch Okechukwu2 IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org ISSN (e): 2250-3021, ISSN (p): 2278-8719 Vol. 04, Issue 12 (December 2014), ||V4|| PP 12-20

[2]A Machine Learning Approach to Football Match Result Prediction by lucacarloni,Giuseppe Sansonetti Part of the Communications in Computer and Information Science book series (CCIS,volume 1420) available: : https://www.researchgate.net/publication/352940839_A_Machine_Learning_Approach_to_Football_Match_Result_Prediction

[3]M. A. Raju, M. S. Mia, M. A. Sayed and M. Riaz Uddin, "Predicting the Outcome of English Premier League Matches using Machine Learning," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp.1-6, doi: 10.1109/STI50764.2020.9350327.

[4]Kundu, Tuhin & Choudhury, Akash & Rai, Sruti. (2021). Predicting English Premier League Matches Using Classification and Regression. 10.1007/978-981-15-5077-5_50.

[5]Ulmer, B., & Fernandez, M. (2014). Predicting Soccer Match Results in the English Premier League.

[6]Ćwiklinski, Bartosz &Giełczyk, Agata &Choraś, Michał. (2021). Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers. Entropy. 23.90. 10.3390/e23010090.

[7]Herbinet, C., 2018.Predicting Football Results Using Machine Learning Techniques. [online] Imperial College-ofLondonRahman, M.A. A deep learning framework for football match prediction. SN Appl. Sci.2,165
(2020). https://doi.org/10.1007/s42452-019-1821-5019-1821-5[9]Rana, D., 2019.PREMIER LEAGUE MATCH RESULT
PREDICTION USING MACHINE LEARNING[online] Jaypee University of Information Technology Waknaghat, Solan- 173234, Available at:http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/22987/1/Premier%20League%20Match%20Result%20Prediction%20Using%20Machine%20Learning.pdf

[10]Yadav A, Sharma A, Gautam A, Bathla G, Jindal R (2017) Predicting English Premier League Results using Machine Learning. J ComputerEng Inf Technol 6:1. doi: 10.4172/2324-9307.1000165

[11]Campanelli, N. (2019, May 22). Betting on the English Premier League.Towards Data Science.https://towardsdatascience.com/betting-on-the-english-premier-
league-making-money-with-machine-learning-fb6938760c64