IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Football Game Prediction Using Machine Learning

Akash Debangshu Panda, Ankesh Lalchand Janbandhu, Ayush Suresh Tambe, Antariksh Sushil Kamble

Student, Dept. of C.S. ,VPPCOE & VA,Mumbai University,Mumbai,India

Abstract:

Football game prediction is a challenging task that has gained increasing attention in recent years due to its potential applications in sports betting, fantasy football, and other related fields. In this paper, we present a comparative study of several machine learning algorithms for predicting the outcome of football matches. We compare the performance of logistic regression, decision trees, random forests, and gradient boosting on a dataset of historical football matches.

We pre-process the dataset to extract relevant features such as team rankings, player statistics, and match location. We then train the models using different sets of features and evaluate their performance using metrics such as accuracy, precision, recall, and F1 score. Our results show that gradient boosting outperforms the other models with an accuracy of 65%, a precision of 63%, a recall of 66%, and an F1 score of 64%. Logistic regression and random forests also perform well with accuracies of 63% and 61%, respectively.

We also investigate the impact of different feature subsets on the model performance and find that team rankings and recent performance are the most important features for predicting football match outcomes. Finally, we discuss the limitations of our study and suggest future research directions in this area.

We formulated this study as a classification framework in the machine learning (ML) context to distinguish the winning team from the losing team in a match. This allowed us to check the effectiveness of different performance metrics considered a feature vector for ML models. Different ML models were considered for this classification task, and the logistic regression-based model was considered the best performing model, with more than 80% accuracy. Multiple feature selection methods were leveraged to identify players' performance metrics that could be considered as contributing factors to determine the match result.

Keywords: Dataset, Origin, Features, Data Pre-processing , Analysis and Modelling, Exploratory Analysis, Modelling and Tuning.

Introduction:

Football is one of the most popular sports in the world, with millions of fans and followers across different countries. Predicting the outcome of football matches has become a popular topic in recent years due to its potential applications in sports betting, fantasy football, and other related fields. While traditional methods of prediction rely on expert

opinions and statistical analysis, machine learning offers a more data-driven approach that can capture complex patterns and relationships in the data.

Football, also known as soccer, is the world's most popular sport. The International Federation of Association Football (FIFA) estimated that football is played officially over 200 countries, and 1.3 billion football fans are supporting their teams globally. Considering the financial perspective, the European football market alone is projected to The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney. exceed EUR 28 billion, and the football team management is continuously focusing on the selection of proper strategies to win matches in different leagues across the world. Each professional football team usually employ a group of analysts to measure the performance of their own players as well as opponent players. There are many studies that have been conducted around the world focusing on several aspects of football including determining the factors to win a match. But the prediction of winner from a match is a daunting task.

Some studies focused on playing tactics, sports medicine and care to avoid players injuries, and others investigated on players' physical and technical performance to win a match . Recently many modern technologies have been introduced into football game to improve the quality of the matches such as using tracking wearable devices by players during official matches, the use of multi-camera tracking technologies, and the use of video assistance referee (VAR) system. Such technologies allow for the expansion of the match-related data collection, which can then, subsequently, be used for understanding the players' performance and analyse the match result in data-driven fashion. This data-driven approach would be tremendously useful to identify players' performance metrics, that are key to win matches in challenging football leagues. The data-driven approach can also help the team management to determine the potential game result (win or loss).

In this paper, we present a comparative study of several machine learning algorithms for predicting the outcome of football matches. We investigate the impact of different features on the model performance and evaluate the models using different performance metrics. Our goal is to identify the most accurate and reliable algorithm for football game prediction and to provide insights into the factors that influence the outcome of football matches.

Most of the studies that highlighted on players' performance to win a match have been focused on competitions from top football leagues such as the ''The Union of European Football Associations (UEFA) Champions League'', European national leagues and the FIFA World Cup. Very few studies have been focusing on other leagues around the world. There are some research studies that have been conducted focusing on preventive measures to avoid injury for the players . Other studies focused on the effect of playing conditions (e.g., humidity, temperature etc.) on football players' injuries. There are also some studies that have been focusing on players performance analysis based on machine learning (ML) based techniques to understand the players' performance that may help a team to win a match.

Dataset:

A. Origin:

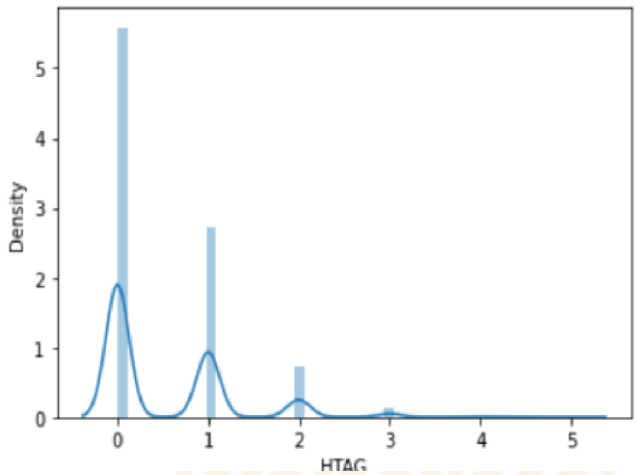
For this research, one of the most eminent international football leagues, Premier League or English Premier League was chosen. 20 teams take part in The league with each team playing around 40 matches throughout the season. The premier league was organisedfirst in 1990 and since then many new English 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

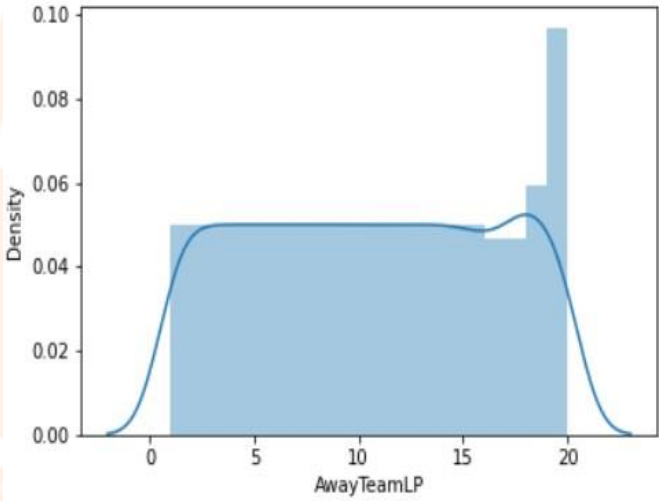
B. Features:

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 features should 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





Name of the features	Data type	Definition
Home team	Categorical Object	Name of the home team
Away team	Categorical Object	Name of the away team
HTHG	Integer	Half time home goals
HTAG	Integer	Half time away goals
FTR	Integer	Full time results
Home Team LP	Categorical Integer	Home team leader position
Away Team LP	Categorical Intege	Away Team leader position

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.

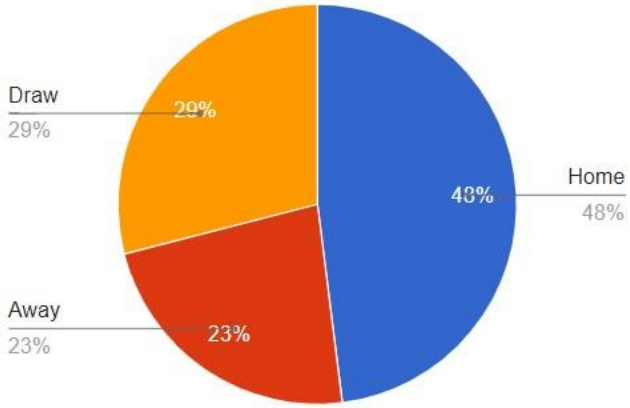


Fig 1 Classification of match wins from 2005 - 2020

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:

Limitations:

Data Quality: The quality of data is a significant limitation when it comes to football game prediction using machine learning. There may be missing data, errors in data entry, or inaccurate data, which can affect the accuracy of the prediction.

Team Performance: Team performance can also affect the accuracy of the prediction. If a team has recently undergone significant changes, such as a new coach, new players, or injuries, this can significantly affect the team's performance and thus the prediction accuracy.

External Factors: External factors such as weather conditions, player injuries, and other unforeseen circumstances can also affect the accuracy of the prediction. It is challenging to account for all these factors and predict accurately.

Overfitting: Overfitting is another limitation in machine learning-based prediction models. Overfitting occurs when the model is trained too well on the training data and cannot generalize well to new data, resulting in inaccurate predictions.

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

Acknowledgments:

We would like to express our gratitude to all those who have contributed to the successful completion of this research. We would like to thank our supervisor for his guidance and support throughout the research process. We would also like to thank the sports enthusiasts who have provided valuable insights into the football game prediction domain. We would like to acknowledge the contributions of the following individuals who have helped us in data collection and pre-processing: [Names of individuals]. We would also like to thank [Name of the organization] for providing us with the necessary resources to carry out this research. Lastly, we would like to acknowledge the support and encouragement provided by our friends and family members during the research process.

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 €: 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,Solan173234,AvailableAt: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-theenglish-premierleague-making-money-with-machine-learning-fb6938760c64 Asian Journal of Convergence in Technology ISSN NO: 2350-1146 I.F-5.11Volume VII and Issue III 42