JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SMART FOOTBALL GOAL PREDICTOR

Pranav Singhal, Arihaant Arora, Om Dwivedi, Akshay Singh,
Student, Student, Student, Assistant Professor
Department of Information Technology
Maharaja Surajmal Institute of Technology, New Delhi, India

***ABSTRACT*-Prediction model for Football results using ML Approach: In numerous methods intended to predict the outcomes of professional football matches, the number of goals scored by each club has traditionally been used as a basic factor for evaluating a team's performance and projecting future results. The count of goals scored in a game is a significant random variable, though, and it frequently causes significant discrepancies between a team's play and the amount of goals they score or allow.**

The main goal of this paper is to investigate various Machine Learning methods for predicting the result of football games using in-game match events rather than the total number of goals scored by each team. We will investigate several model design hypotheses and compare the effectiveness of our models to industry standards. Instead of utilising the actual number of goals scored, we established a "anticipated goals" measure in this project to assess a team's performance We paired this metric with a computations of a team's offensive and defensive ratings that are updated after each game in order to develop a classification model that forecasts the outcomes of future matches and a regression model that forecasts the ultimate score of those matches. In relation to currently employed traditional methodologies, our models perform brilliantly and achieve a precision on par with betting exchange models.

Keywords: Football Prediction , Machine Learning, Machine learning models, Expected goals, Conventional techniques, Football matches, goals scored, Results prediction

I.Introduction

Since football is perhaps one of the most watched sports in the world, many people have always watched it with anticipation. Recently, comprehensive data that illustrates the intricacies of each shot or pass made throughout a game has been compiled for a variety of competitions across many different countries. Information science is now at the forefront of the football industry because to the collection of this data, which has a huge range of possible uses and applications:

- Match technique, methods, and examination
- Recognizing players' playing styles
- Group spending, player acquisition, and player evaluation
- Routines, timetables, and obsession with exercise
- Injury prevention and injury forecasting utilising employment and test results
- The executives and execution anticipation
- Anticipation of the game's outcome and league standings
- The occasion's planning and preparation

• The betting industry has grown quickly in recent years, in part because of the increased incorporation of live football coordinates and the increased accessibility of betting sites due to the development of portable and tablet devices. The value of the football betting market is estimated to range between 300 million and 450 million pounds today.

II.Literature Review

In this research paper we have divided the prior investigation into two groups for football predictions. First, we would look to the landscape of football prediction systems and those methods that are used previously. The studies that give models for our final talk will focus on estimating the estimated goals that a team is estimated to have reached.

Football Prediction Landscape:

Since the 20th century, creating football outcomes gauges has been a notable topic of study. In light of past group findings, Moroney (1956) and Reep (1971) displayed the number of goals scored in a football game using the Poisson conveyance and the negative binomial circulation. However, it wasn't until Slope's discovery in 1974 that it was possible to present and predict match outcomes using historical knowledge rather than depending just on chance.

When Maher used Poisson circulations to estimate the hostile and cautious characteristics of the home and away teams and predict the mean number of goals for each side, he made the first invention in 1982. In the years that followed, dixon and Coles (1997) quickly created a model that could predict match outcomes and scores using a Poisson conveyance. We will distinguish the models we create from the Dixon and Coles model, which is really thought of as a generally valid model. The Poisson relapse condition, which converts the anticipated all-out objectives per side into objective probability using the Poisson circulation, is the foundation of Dixon and Coles' technique.

P(k goals in match)= e k k!

where is the predicted number of goals in the game.

It allows for the estimation of each team's probability of scoring a particular number of goals, which may subsequently be converted into scored possibilities and, eventually, match outcome probabilities.

Expected goals xG models :

Another topic that is essential to us for our research is the normal targets model. A rather original idea seeks to analyse match data in order to determine the number of goals that a certain team should score in light of the game's remaining aspects. By using an anticipated model of objective, we might reduce some of the haphazardness associated with the actually scored objectives and provide a more accurate depiction of a group's viability and, consequently, strength. In 2012, MacDonald developed an anticipated objectives model to assess the feasibility of NHL (Hockey) games. He used two different metrics-

- Corsi and Fenwick ratings for successful and unsuccessful shots (shots, blocked shots and missed shots)

It made it possible to assess a team's performance and determine, for instance, if they wasted goal opportunities or failed **to** generate enough chances to score goals. With more accurate estimations for future outcomes, our anticipated objectives model produced very good results. Making use of the opponent's statistics to determine the anticipated goals would be a potential expansion for this study. In football, which is precisely what we will be going to do: taking data from two sides to determine number of goals both the teams was predicted to score in a game and then using this number to enhance future forecasts. In 2015, Lucey used spatio- temporal information properties to determine the chance that each shot would result in a goal. It includes elements like short-term planning, safety proximity, game stage, and so on. This gives us the ability to determine how many objectives a group would have been anticipated to score and to assess the group's performance during the game. The situation is particularly interesting to us since we have geographic information on goals and shots that take place throughout a game. We would utilise the creation of an anticipated model of advanced aims to

analyse how a group really behaved in a game in more detail. Similar to Eggels (2016), who used geological data for the shot as well as the player's body portion to categorise each scoring a lucrative open door into a chance of really accomplishing the goal. Techniques for different categorization, such as decision trees , logistic regression and random forests, were evaluated. Additionally, we will need the calculation of several approaches to calculating the likeliness of goal scoring for every opportunity. in this research paper. In order to predict the outcomes of upcoming football games, the website FiveThirtyEight uses a group assessing framework and an anticipated objectives model, which is one of the most intriguing models in our analysis. They maintain a cautious and hostile rating in light of the average goals scored and the expected amount of goals in each game. They can make use of Monte Carlo simulations to estimate future outcomes in order to attempt and anticipate competition winners. We might test the intriguing idea that, on the off occasion that a team is winning at the end of a game, the importance given to the predicted goals is less in order to obtain the most accurate estimations, we will

test a few AI computations and add attributes to the expected goals model in addition to using these methods as models for the exam paper.

Researchers have tried to implement machine learning to try and develop a system that can forecast football game results. Various researchers have used different types of parameters to build their models. Due to the unpredictable nature of the sport, achieving high accuracy of correct predictions is a tedious task.

In this section, we review the methodologies used in different research papers published and what makes this research paper different than others.

1) Ahmed Awadallah and Raghav Khandelwal of Stanford University talk about predicting match outcome based on previous performance of the target team on home and away turf. They use the head-to-head record of the target team against the opponent team. The issue with this parameter is the lack of reliability. Underdogs causing upsets are hard to predict based on head-to-head record alone. Due to the ever-changing nature of the game, past performances alone do not capture the whole performance of a team. For example- A team dominant in a particular fixture might be so due to the influence of their talismanic player. Transfer of this player might weaken the team and its chance to dominate the fixture.

2) Dwijen Rudrapal of National institute of technology, Agartala made further improvements to this approach. They further included form as a contributing factor for the prediction parameter. They gave more weightage to the team's current form, factoring in wins, losses and draws in recent fixtures. Although the algorithm performed better than its predecessor, it still didn't factor real time team performance.

3) Athiwat Thongyoo and Suphavit Thaneerat of Chulalongkorn University used parameters that were further divided into sub-parameters. To name a few, Player performance, Tactic performance, League scale and match scale are some of the umbrella features under which the sub-parameters combine. Each individual parameter contributes to the prediction. Some of these parameters are unpredictable in nature such as Tactic performance. Tactics change according to the opponents and may not always work. Due to the volatile nature of the parameters, consistency was not found in this Algorithm.

4) Tuomas Tiippana of Aalto University introduced a new method to predict the outcome. He benchmarked the expected goals predicted for a team in contradiction of the actual goals recorded by the team. The following formula was used to create the benchmark scores:

$$\textit{Goals scored} = \textit{Shots on target} * \frac{\textit{Goals scored during the season}}{\textit{Shots on target throughout the season}}$$

This score was then compared to the actual goals scored. A value greater than equal to the benchmark was regarded as a goal. Any value lesser than the benchmark was disregarded. Although the algorithm performed really well, it did have its flaws. This method is heavily dependent on the expected goals predicted by external sources. Any error occurred in calculating expected goals may result in false predictions. Also, this algorithm relies more on statistics as compared to machine learning.

5) Students of Darul Uhm University used shot distance, angle and the part of the body used to shoot the ball. This model is more about a shot resulting in a goal or not rather than capturing all the aspects of the match. Football is a lot more than just scoring goals. This model therefore fails to accurately predict the outcome of a fixture.

6) Corentin Herbinet of Imperial college, London has devised an algorithm that covers almost every essence of the game. Three major components used in this algorithm are expected shots, match xG and ELO ratings.

Expected shots show if the shot taken by the player results in a goal or not based on area of attempt and ELO rating of the player.
Match xG throws light on the performance of players working in a team. Individual player performances of a player are grouped together to predict the performance of the team.
ELO ratings are used to measure the relative skill of the target team in comparison to the opponent team. This is done by comparing individual player skills. For example- Skills of a striker will be compared to the opponent's striker, goalkeeper to goalkeeper and so on. This is done by the ratings obtained from a video game called 'FIFA'.
Despite this algorithm being near perfect, one flaw is that the ratings of players are obtained from a video game and not real life data. Although FIFA ratings give an idea of a player's calibre, human skill is hard to quantify in numerical terms. Also these ratings change once a year so player form and growth curve are also not factored in.

This work learns from the mistakes made by other researchers and therefore the chosen features try to capture the whole essence of a fixture. It tries to factor in the performance of a team based on statistics that are obtained from in-game events.

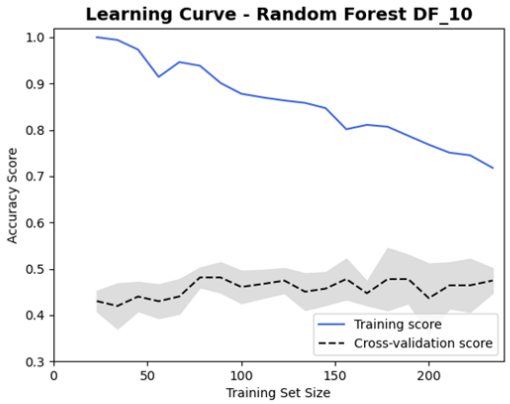
It makes use of historical as well as newly obtained data.Goal difference and shooting statistics help to analyse the attacking prowess of a team while possession and passing stats define the playmaking ability of the team.

Shots difference also give us an insight about the teams differences as a negative difference of the opponents suggest more blocks made by the target team and therefore better chance at defending a goal. This is also inferred by the goal difference statistic as a positive difference suggests more goals scored than conceded.

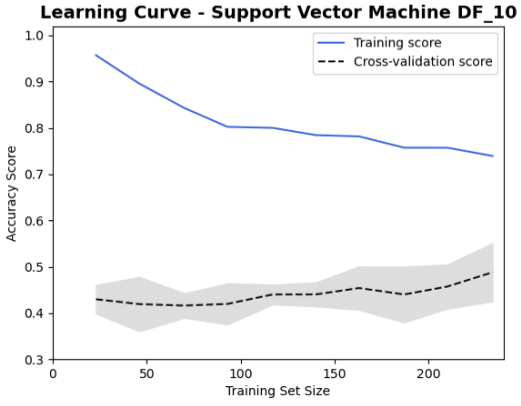
III.Scope Of Work

The seven features under consideration were for our target team. Likewise, these seven features are used for predicting the chances of the opposition team. In total, a graph plots fourteen features- seven for each team under consideration. As seen in the graphics, each of the selected features had some impact on the final result of the match. Each of these features had an impact- either big or small- with an exception being number of fouls conceded which had little relevance to the final outcome. In the about visualisation, the green dots are a representative of a win for our target team. The blue dots are an indication of a win for the opponents. The points on the graph lying in the bottom left quadrants are a suggestion that the teams contesting are of low quality in terms of their gameplay. In contrast to this, the quadrant on the top left side of the graph indicate the quality of the target team as low and that of the opponent is high. Points on the top right suggest both the target team and the opponent are of high quality. Bottom right suggests the target team being of high quality going against an opponent of low quality. The model performances were judged by creating a confusion matrix.

We find that Random forest algorithm has an accuracy of 50.8%. The SVM model had an accuracy of 46.6%.
K nearest neighbour had the accuracy of 51.5%.

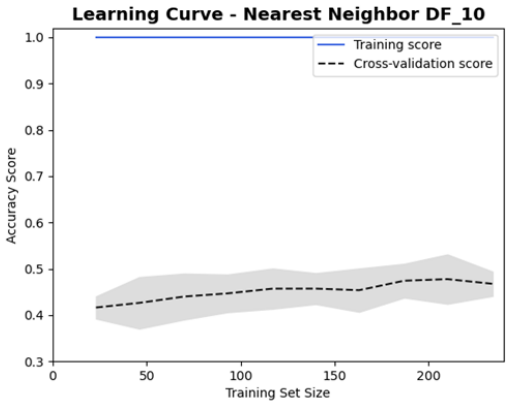


These grids quickly uncover that each of the three models are neglecting to conjecture a draw precisely. While gauging a move, nearest neighbour and SVM are both more mistaken than exact, and the irregular woodland classifier just accurately predicts 5 coaxes out of almost 300 games. The highlights displayed in figure 1 were re-plotted to show just attractive request to decide why this was occurring. Since it was anticipated that groups with comparable expertise levels would draw all the more every now and again, information bunches were expected to be situated inside the ran red lines assigning these groups. Yet, clearly the conveyance of drawn matches shows up somewhat erratically in this assortment of results from 2019 and 2020.Given the overall uncommonness of draws (around 1 of every 5 matches) and the capricious idea of the outcome in this data set, it isn't is really to be expected that the models perform ineffectively in their capacity to figure them. The irregular woodland and closest neighbour models both prevail with regards to accomplishing the review's underlying objective, which is to accomplish a test exactness of half. The SVM model was not generally thought about for additional investigation since it didn't meet this objective.



The probabilities created by the model should be exact to fulfil the subsequent point. Figure below was utilized to concentrate on this, and it shows a histogram of precise and erroneous expectations given a forecast likelihood. A great model ought to create probabilities that are dependable and steady with the outcome, for instance, a likelihood of 60% ought to be exact 60% of the time. The irregular woods model appears to help this, but the closest neighbour model yields some odd findings. Predictions with a likelihood of 80% to 82.5%, for example, are bound to be off-base than right. The arbitrary woods model was picked over the closest neighbour model hence.

Given a prediction probability, this histogram shows both the correct (green) and wrong (red) forecasts. After 50 iterations of model construction with random train-test splits, the outcome was steady.

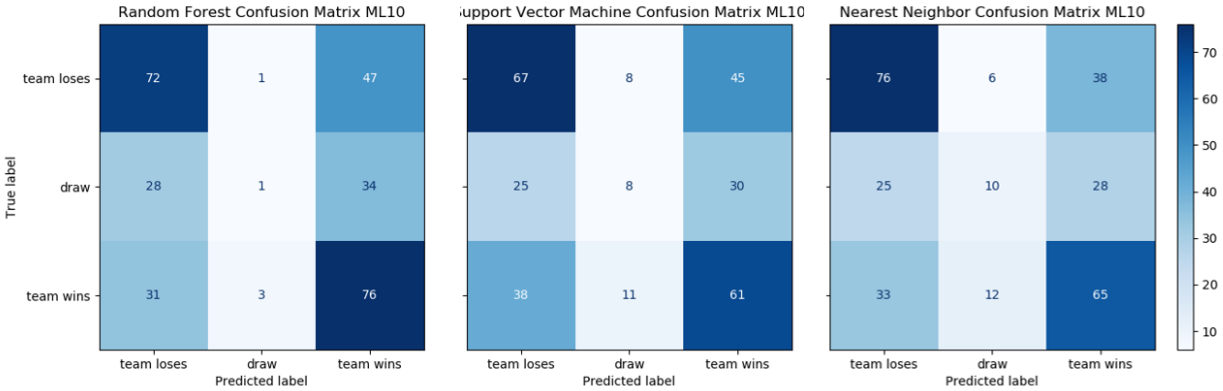


IV Result and Conclusion

As a result from the confusion matrices of the algorithms used in this research paper the model performances were judged and we found out that the least accuracy came when the Support vector machine algorithm was implemented that is 46.6%. Then came random forest algorithm with an accuracy of 50.8.2%. The algorithm that gave the best accuracy was K nearest neighbour with a percentage of 51.5

The primary goal of developing an anticipated objectives model by investigating various approaches to machine learning has been achieved. We gained knowledge of cutting-edge machine learning techniques including K closest neighbour, Support Vector Machine, and Random Forest, which enabled us to provide match results and score predictions while also reaching the highest level of accuracy.

The key problem of the project, which needed extensive information gathering, information gathering, and research labour, was to acquire appropriate data to develop an expected objectives model. After all the fabrication and experimental work and running all the machine learning algorithms K Nearest Neighbor helped us to achieve the highest accuracy.



V.Future work

There are further improvements to this project that can be implemented to make this project have more practical use cases and better prediction accuracy. Some of the suggested changes are:

· Inclusion of additional features: For this project we had selected seven features. There are many more in-game events that influence the outcome of the match such as individual brilliance, form, tactical changes etc. Inclusion of such parameters can help minimize the randomness of the game resulting in better prediction.

· Trying on new methods of features engineering: This project uses basic mathematical tools to minimize the dimensionality of data. Using complex statistical models such as standard deviation, mean deviation, Poisson distribution etc may improve the dataset and provide us with new insights.

· The current prediction model is built on in game events. Inclusion of factors like the club size, historical performances, playstyle, average points per game, current and previous league positions, type of match (friendlies, group stage, knockout stage, finals etc) may help include the mentality with which the players approach the game and therefore affects match performance.

Practical use case of project: Currently the project is just an algorithm and not of much use unless user knows how to implement the model and extract and interpret information. For this project to be user friendly and usable, an User interface must be created that is graphical in nature such as an application or a website. This can mean the predictions can be accessed by general public.

VI.References

1.https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf.
2. Rudrapal, Dwijen & Boro, Sasank & Srivastava, Jatin & Singh, Shyamu. (2020).
A Deep Learning Approach to Predict Football Match Result. 10.1007/978-981-13-8676-3_9.
3.http://cs230.stanford.edu/projects_spring_2020/reports/38854780.pdf
4. Wheatcroft, Edward. 'Forecasting Football Matches by Predicting Match Statistics'.
1 Jan. 2021 : 77 – 97.
5. https://doi.org/10.48550/arXiv.2206.09258
6.M. A. Al-Asadi and S. Tasdemır, "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques," in *IEEE Access*, vol. 10, pp. 22631- 22645, 2022, doi: 10.1109/ACCESS.2022.3154767.
7. Thongyoo, Athiwat & Thaneerat, Suphavit. (2022). Football Team Performance Evaluation with Expected Goals (xG).
8.https://aaltodoc.aalto.fi/bitstream/handle/123456789/99805/bachelor_Tiippana_Tuomas_2020.pdf?sequence=1&isAllowed=y
9.Fátima Rodrigues, Ângelo Pinto,Prediction of football match results with Machine Learning,Procedia Computer Science,Volume 204,2022.
10.Partida, A., Martinez, A., Durrer, C., Gutierrez, O., & Posta, F. (2021). Modeling of Football Match Outcomes with Expected Goals Statistic. Journal of Student.
11.Umami, Izzatul, Deden Hardan Gautama, & Heliza Rahmania Hatta. "implementing the Expected Goal (xG) model to predict scores in soccer matches." *International Journal of Informatics and Information Systems* [Online], 4.1 (2021): 38-54. Web. 25 Nov. 2022.