ESP-JETA
ESP Journal of Engineering & Technology Advancements
Volume 3 / Issue 1 Paper Id: JETA-V3I1P104 / Page Count: 22-27

Research Article

Expected Goals Prediction in Football using XGBoost

Harish S[1], Abishek Kevin A[2], Harsha Vardhan U[3], Sharon Femi P[4]
[1,2,3]*UG Scholar, of Information Technology Sri Venkateswara College of Engineering, Sriperumbudur, India.*
[4]*Associate Professor, of Information Technology Sri Venkateswara College of Engineering, Sriperumbudur, India.*

Received Date: 10 January 2023 Revised Date: 27 January 2023 Accepted Date: 02 February 2023

Abstract: *Expected goals of a football match determine whether a team have won or lost. When considering the expected goal results, a team may appear to have lost the game but actually win it, and vice versa. The expected goal is the amount of goals a team should have scored based on the information available for that particular game. Numerous machine learning algorithms are employed to measure the effectiveness of shots in football. In this paper, we develop a gradient Boosting model to evaluate the scoring opportunities using event data collected from live football matches. This method can be used to show the players who are most likely to score at any given time throughout the game as well as where on the field they are most likely to do so. Experimental results demonstrate that they recognise and assess significant match opportunities and analyze the players in a football match depending upon their performance using Expected goals(xG) as an evaluation metric*.

Keywords: *Expected Goals, Extreme Gradient Boosting, Logistic Regression*

I. INTRODUCTION

In the modern football industry, data analytics has taken on a significant role. The game of football has developed so much that the statistics involving the game is not just the mere number of goals scored by each side and the ball possession each team has. With a clear distinctive difference between football in the olden days and the modern game, it requires more from the footballing teams to increase their performance and recruit the right profile of players. Making a decision about an event that is happening on the go has been a challenge to many. In a football match, something that is all about events like that needs a complex model to simplify the process of understanding those events. This rapid growth of the game resulted in defining "Expected goals".

This is defined as the probability of a shot ending up as a goal. Big data is enabling clubs to extract insights to improve player performance, reduce injuries, and increase their commercial efficiency as they seek to acquire a competitive edge both on and off the field. It is not always simple to contextualize the data and draw out insightful conclusions. In this paper, a novel method is presented to predict the expected goals of each team in a football match using expert feature engineering techniques. This helps the stakeholders related to football in making better decisions. The major stakeholders being the football clubs, football analysts and sports betting enthusiasts, a method to maximize the positive outcomes will cater to their needs. The chance of a goal to be scored will be calculated simultaneously as the match happens. The objectives are to assist football-related stakeholders in making wiser decisions, to forecast a player's capacity to capitalise on opportunities when they arise and to forecast the strength and weakness of the team.

The first section of this paper introduces the need for Expected Goal prediction. It further goes on to emphasize on the importance of data in the game of football. The next section discovers the previous works done on Expected Goals and Football Analytics altogether. The third section describes the proposed system of this paper. The next section details about the description on all the modules present in the system. The next section describes the results obtained when we deploy the model. The Extreme Gradient Boosting model gives better results compared to Logistic Regression. The performance evaluation of this model is described in the next section.

II. RELATED WORKS

Spearman et al [1] introduced an approach for identifying and analyzing significant opportunities during the match by highlighting the areas of the pitch where specific players create off-ball scoring opportunities to aid opponent analysis. The model used is Potential Pitch Control Field to detect which team has ball more time with them without any interference. The scoring probability solely depends on the distance between the goal and shooting position, invalidating other features.

To forecast the outcome of the English Premier League, Baboota and Kaur [2] built a generalised predictive model. The most crucial elements for forecasting the outcome of a football game are identified via feature engineering and exploratory data analysis, which culminates in the development of a highly accurate prediction system.

Berrar et al [3] proposed probabilistic, bootstrapped models that are built to predict the actual scorelines of the matches beforehand using domain knowledge. Feature extraction methods were employed for fine-tuning significant features: recency feature extraction and rating feature learning. Models used were k-nearest neighbors and xGBoost. Ranked probability score (RPS) metric used to evaluate two supervised ml models.

To forecast the results of football, Rodrigues and Pinto [4] used machine learning techniques that use a variety of statistics from prior matches and player traits from both sides as inputs. A number of prediction models were evaluated, and the results of the experiment indicated that they performed well in terms of the profit margin of football bets.

Kollar [5] introduced a method called Pitch Partitioning- a process of splitting the pitch into 8 partitions and assigning weighted average values to these parts based on the dawson model. In addition to the pre-existing features in the evolved model, possession sequence- a record of the ball activity throughout the two halves of a game proves to be a significant addition to the feature set.

Bauer et al [6] detected defensive traits to reduce goal threat. Model that was deployed is the gradient boosting algorithm. Ball possession and ball status are the features that tell which team is in control of the attack are the key performance indicators. The disadvantage is that it excludes the set-piece possessions and ball throw-ins. Expert based feature extraction for fetching event data Speed of opposition in defense excluded.

Link et al [7] suggested a method for representing probability of goals scored during every second of the game Pitch split into different zones and are given weight values accordingly Ball action and pressure zones are metrics used to define the defensive actions Model used here is logistic regression with k3 constant Calibration process done to optimize the model constants. The approach described here has the drawback of not accounting for elements such as player and ball movement dynamics, player gaze patterns, player positioning in respect to the ball, and the availability of teammates.

Recurrent back propagation is a method that Staudemeyer et al. [8] devised to address the issues of learning to store information over long time intervals. Additionally, LSTM resolves difficult, artificial long-time lag tasks that no previous recurrent network algorithms have ever been able to. It still remains an upgrade to the existing recurrent neural networks model with capacity to retain memory.

By analyzing the threat of the play sections that came before them, Merhej et al. [9] proposed using deep learning techniques to construct a novel metric that rewards such defensive responses. It values defensive actions according to what it prevented from occurring throughout the game. Using the Levene test to combat overfitting. It uses the RNN-LSTM model.

According to Ferraresi et al. [10], the essential features include the amount of assists, shots allowed, saves made by the goal keeper, precise passes made relative to total passes, and shoots on goal. Additionally, this study comes to the conclusion that offensive behaviors outnumber defensive ones. Bayesian model averaging and Monte Carlo-Markov Chain hybrid Models are used. Disadvantages are that it discredits the defensive position that contributes to a goal scoring opportunity unique from other approaches as they've included the ball retained from the opposition half as an attacking feature.

In order to forecast the results of matches played by the Tottenham Hotspur Football Club, Joseph et al [11] contrasted the performance of expertly created Bayesian Networks(BN) with other machine learning algorithms.

III. PROPOSED SYSTEM

The current system that this paper talks about is loosely based on the gradient boosting models explained in the literature survey. The system is built as an extreme gradient boosting model that predicts the probability of goals that can be scored. The earliest papers that covered this problem evaluated a model based on two parameters. The parameters are the distance from which the shooter shoots the ball, and the angle between the shooting position and the goal posts. As a matter of fact, the addition of the position of the ball tends to give more accurate results.

Another important aspect that we found to be crucial is the position of the ball before striking it towards the goal. Various positions of the ball result in various types of outcomes. This feature is the unique point of improvisation in this model. Leveraging these features, XGBoost is used to train the Kaggle Dataset. The proposed system includes collecting the data for training the model to bootstrapping them to the best performing model. The developed model is deployed and evaluated using various performance metrics. The proposed architecture diagram for predicting the expected goals in football is shown in Fig 1.

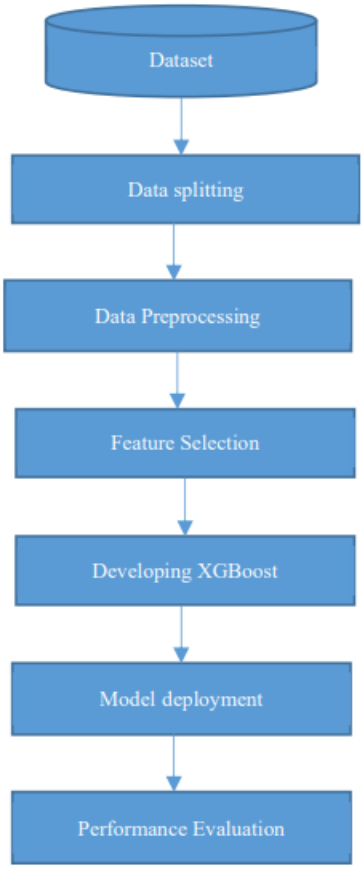

Figure 1: Proposed Goal Prediction Architecture Diagram

IV. MODULES DESCRIPTION

A. Data Collection

The majority of publicly accessible football statistics are only comprised of compiled information like Goals, Shots, Fouls, and Cards. This straightforward aggregation, which lacks context, might be misleading when evaluating performance or developing prediction models. For instance, a team with 10 long-range shots on target has a lesser chance of scoring than one with the same number of attempts from inside the box. However, measures derived from this straightforward count of shots will evaluate the two teams in a similar way. The dataset that's used is collected by scraping websites that provide users with intricate event data. The dataset offers an in-depth look at 9,074 matches, comprising 941,009 occurrences, from the top five European football leagues: England, Spain, Germany, Italy, and France, from the 2011–2012 season to the 2016–2017 season as of 25.01.2017..

B. Data Splitting

A ratio of 80:20 was used to divide the dataset into training and testing datasets. Overfitting is a problem that arises frequently during model training. When a model excels on the data we used to train it, but underperforms when applied to fresh, untested data, overfitting occurs. Usually, a model's likelihood of being overfitted increases with model complexity. When the model performs poorly even on the training set of data, underfitting has taken place. By creating different samples to train and test can be used to resolve these issues. The testing data is used to evaluate the model, while the training data is utilised to train.. Two-thirds of the data set can be used for training and one-third for testing and the model is evaluated with the performance metrics.

C. Data Preprocessing

Data Pre-processing is the process of transforming the features to understandable data. The majority of real-world datasets will have some missing values because the system generating the data may have made an error leading to missing observations, or a value may be missing because it is not relevant for a particular sample. The K-Nearest Neighbors algorithm is a commonly used technique for replacing missing values. This model uses a distance metric, such as the Euclidean distance, to determine a specified set of nearest neighbors and imputes the mean value for those neighbors. Another pre-processing step transforms the data to take the shape of a normal distribution. This data normalization converts the data values to a common range of values so that it is easy to draw conclusions.

D. Feature Selection

Feature selection identifies the features that are of importance to build the model. The features of the dataset that are correlated are determined to be the features. Correlation metric, namely Pearson correlation coefficient, is used in identifying the features that are most important to be selected as features. Pearson correlation coefficient is the measure of the strength of a linear connection between two variables. It displays the degree to which the data points fit this new model or line of best fit. The value of correlation coefficient ranges is between -1 to +1, where -1 indicates that the features have no correlation whereas +1 indicates that the features are strongly correlated.

E. Developing Model

An extreme gradient boosting model is built and trained for the dataset. Extreme Gradient Boosting is one kind of a gradient boosting algorithm. It is made up of a group of decision trees. We can reduce the variance of our predictions within the well-known bias versus variance tradeoff by creating hundreds of alternative trees using various predictors and samples because these trees have a tendency to overfit to the training data. It is predicated on the hunch that when prior models are coupled with the best possible upcoming model, the overall prediction error is minimized. The desired result for each instance in the data depends on how modifying that case's forecast affects the overall prediction error; if a case's prediction changes slightly but results in a significant decrease in error, the case's desired result is a high value. The error will be reduced by predictions from the new model that are close to their targets, and the case's next target outcome will be zero if a slight change in the forecast remains unchanged. This forecast cannot be altered to reduce the error.

Due to its ability in managing overfitting and skewed data, the XGBoost is utilised to construct the ultimate output for optimal feature space and increase the prediction gradually. Recent studies show that XGBoost is more effective than other ensemble approaches at handling skewed datasets [12]. In order to regularise the parameters, reduce overfitting, and improve speed and performance, it makes use of the gradient boosting concept. During sequential learning, gradient boosted trees replace regression trees. These regression trees base their ultimate prediction on the continuous score given to each leaf and added up.

In this procedure, the weight w is determined for each iteration to forecast the outcome as the tree grows. The best score for each leaf and the tree's total score are then determined using gradient descent. A regularisation or penalty term,, that lowers the complexity of the regression tree functions and lowers the variance is a component of the gradient descent loss function. This parameter is adjustable and accepts values that are equal to or greater than 0.

F. Model Deployment

Model deployment is the action of implementing machine learning models. This enables the predictions of the model accessible to users, developers, or systems, allowing them to interact with their application or take business decisions based on data.

G. Performance Evaluation

The model is evaluated using precision, recall and F1-score. The quantity of accurate positive forecasts serves as a measure of precision. Recall measures the number of positive cases correctly predicted by the model in comparison to the total positive cases of the data. F1-score combines the precision and recall and is the average of both precision and recall. The model is evaluated using ROC-AUC, PR AUC and kappa statistic

a. ROC-AUC: ROC AUC score helps to calculate the rank correlation in between prediction and target. This metric is more helpful because it informs us that this metric demonstrates how well your model ranks predictions. It reveals the likelihood that a positively picked case will be ranked higher than a negatively chosen one.

b. PR AUC: Precision (PPV) and Recall (TPR) are combined into a single visualisation by this curve. The PR AUC is determined by averaging the precision ratings determined for each recall threshold.

c. Kappa statistic: Interrater dependability is typically evaluated using the kappa statistic. The significance of rater dependability depends on how well the data collected for the study reflect the factors assessed.

V. EXPERIMENTAL RESULTS

A. Data

For this analysis, the event data, shot placement and shot outcome data have been taken from Kaggle football analytics competition. The event data contains all the information about the football match, such as the distance from which the shot was taken, whether the shot resulted in a goal, what type of shot that is and so on. The visualization of the shot outcome and shot placement are shown in Fig. 2 and Fig. 3.

B. Data Preprocessing

The event data for a particular football match is retrieved in csv format. The event data primarily contains categorical features. Data pre-processing involves eliminating the categorical features by converting them into numerical data type using one-hot encoding.

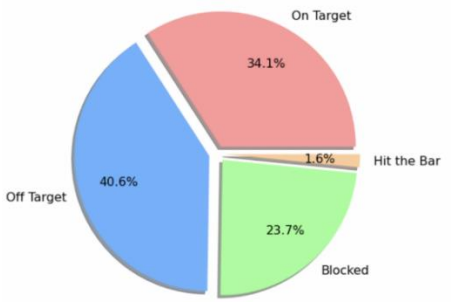

Figure 2: Shot Outcomes

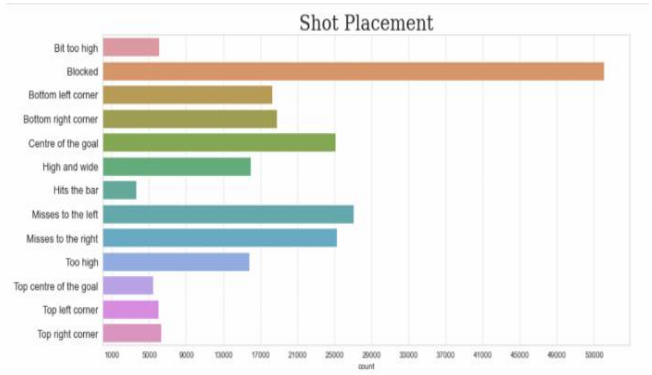

Figure 3: Shot Placement

C. Correlation of Data

There are over 12 features present in our dataset. Handling them all can be an extensive task. Using correlation as a metric, the features that are most important to predict the probability are determined. Out of the 12 features, “location”, “is_goal”, “shot_outcome” and “assist_ method” came out to be the features with high correlation. Similarity scores are derived by attribute-by-attribute comparisons of two data objects, with the correlation score being the result of adding the squares of the magnitude differences between the two attributes. Pearson's correlation, one of the most often used correlation techniques, generates a score that can range from -1 to +1. Two uncorrelated objects would have a Pearson score close to zero. Two objects with a high score are very similar to one another. Two items with inverse correlation would have a Pearson score close to -1.

D. Model Building

We built an Expert Gradient Boosting Algorithm to predict the probability of goals scored by the teams. This model takes in the features shortlisted by feature selection using Pearson’s correlation coefficient as the input.

E. Performance Evaluation

The confusion matrix, which summarizes all predictions, reveals that, our model correctly classified 70,781 of them as no-goals and 913 goals. The model erred in 6,238 instances when it incorrectly predicted that the shot would not result in a goal yet it did and 2266 successful shots were not projected to result in goals. The model performs exceptionally well at no-goal (class 0) prediction but poor at goal prediction (class 1). The model resulted in 71% of precision, 27% of recall and F1 score of 0.39.

XGBoost model predicts whether a shot is a goal or not 91% of the time. Also, ROC-AUC score of 0.82 is obtained. But, these metrics do not take the dataset's extreme imbalance into account. There are a lot more attempts than attempts that result in goals. Therefore, the accuracy attained will be already 89% if a shot is randomly predicted not to be a goal each and every time. So, in order to properly know how effective our model is, we need additional measures. For this reason, Cohen's Kappa and PR-AUC (Precision-Recall Under the Curve) are used to assess the models. Given that they both account for the imbalance in our data, both of these are more pertinent in this situation.

The PR-AUC and Cohen-Kappa scores for our XGBoost model are as follows: The baseline performance for PR-AUC is 0.11. This is the PR-AUC that would result from a purely arbitrary guess. The proposed model obtains an PR-AUC of 47.35 and a Cohen Kappa of 0.35.

VI. CONCLUSION

In this study, we use XGBoost to predict the results of football matches in European football leagues based on the performance metrics of the players. The striker's shot, the distance of the high-speed striker and midfielder, the successful pass of the player and fair play can all improve the chances of winning a football match. With the growing focus on sports statistics, Expected Goals aims to benefit both footballers and spectators. The proposed model achieves 80% accuracy in distinguishing winning and losing teams in European soccer matches.

VII. REFERENCES

[1] Spearman, W., 2018, February. Beyond expected goals. In Proceedings of the 12th MIT sloan sports analytics conference (pp. 1-17).

[2] Baboota, R. and Kaur, H., 2019. Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), pp.741-755.

[3] Partida, A., Martinez, A., Durrer, C., Gutierrez, O. and Posta, F., 2021. Modeling of Football Match Outcomes with Expected Goals Statistic. Journal of Student Research, 10(1).

[4] Rodrigues, F. and Pinto, Â., 2022. Prediction of football match results with Machine Learning. Procedia Computer Science, 204, pp.463-470.

[5] Berrar, D., Lopes, P. and Dubitzky, W., 2019. Incorporating domain knowledge in machine learning for football outcome prediction. Machine learning, 108(1), pp.97-126.

[6] Bauer, P. and Anzer, G., 2021. Data-driven detection of counterpressing in professional football. Data Mining and Knowledge Discovery, 35(5), pp.2009-2049.

[7] Link, D., Lang, S. and Seidenschwarz, P., 2016. Real time quantification of dangerousity in football using spatiotemporal tracking data. PloS one, 11(12), p.e0168768.

[8] Staudemeyer, R.C. and Morris, E.R., 2019. Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586.

[9] Merhej, C., Beal, R.J., Matthews, T. and Ramchurn, S., 2021, August. What Happened Next? Using Deep Learning to Value Defensive Actions in Football Event-Data. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3394-3403).

[10] Zambom-Ferraresi, F., Rios, V. and Lera-López, F., 2018. Determinants of sport performance in European football: What can we learn from the data?. Decision Support Systems, 114, pp.18-28.

[11] Joseph, A., Fenton, N.E. and Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), pp.544-553.

[12] Femi, P.S. and Vaidyanathan, S.G., 2022. An efficient ensemble framework for outlier detection using bio-inspired algorithm. International Journal of Bio-Inspired Computation, 19(2), pp.67-76.