Expected goals (XG) and player analysis

Раде Којчев 142/2021 kti1422021@feit.ukim.edu.mk

Факултет за електротехника и информациски технологии

Универзитет „Св. Кирил и Методиј“- Скопје, Р. Македонија

As time passes by and football as sport becomes more and more popular sports or in this case football science becomes of higher and higher value because if used correctly it can give a given team advantage over another. Expected goals (XG) is the first football metric that started the mass exploration of the match data in order to use the same data for evaluating the players, their performances and finding an objective way of defining who has performed to their expected level, but also who has under or overperformed that same level. In this paper we use three of the most used machine learning algorithms for sports science (Gradient boosting, Logistic regression and Random forest) in order to create the best model for predicting XG. The results show that all the models work almost identically when comparing their hit/miss rate but because of the different feature importances of the Gradient boosting model I declare it as the best one.

I. INTRODUCTION

In the modern day where everything is getting digitalized it was only a matter of time when the digitalization of sports will come through even though it does sound really inhumane. Expected goals is a metric that given an attempt on goal with its features can give a numeric value from 0 to 1 of what is the chance of that attempt resulting in a goal. Over the years it showed that even though people love football they can not objectively evaluate what team was better in a given match and also which players performed well or not because the people, the fans have their bias towards given teams and also given players, so XG is the perfect tool in this regard to help with that evaluation.

II. RELATED WORK

Even though expected goals and sports or to be more precise football data analysis is a relatively new branch in data science it has seen big improvements from its beginning up until now. The first XG models are quite similar to the one that we are doing in this paper but because of the mass investment and research in this field of data science there are newer and significantly better models on this topic.

Today the most famous and most used XG models are the following:

1. Opta: One of the pioneers in football analytics, Opta’s xG model is highly regarded for its extensive data coverage and reliability. Opta’s model incorporates various factors such as shot location, shot type, assist type, and defensive pressure.

2. StatsBomb: Known for its detailed data collection, StatsBomb's xG model includes unique features like pressure on the ball and player positioning. This model is used by many professional clubs and analysts for its comprehensive and granular approach.

3. FiveThirtyEight: FiveThirtyEight’s xG model is popular for its public accessibility and detailed match predictions. The model considers a range of factors including shot location, shot angle, and body part used for the shot.

4. Understat: Understat provides a transparent and user-friendly xG model, offering visualizations and detailed breakdowns of each shot's xG value. It is widely used by fans and analysts for its simplicity and depth.

5. Fbref (FootballRefference): Fbref, powered by Stats Perform data, offers an accessible xG model that is integrated into its extensive

database of football statistics. The model is valued for its accuracy and ease of use for both casual fans and serious analysts.

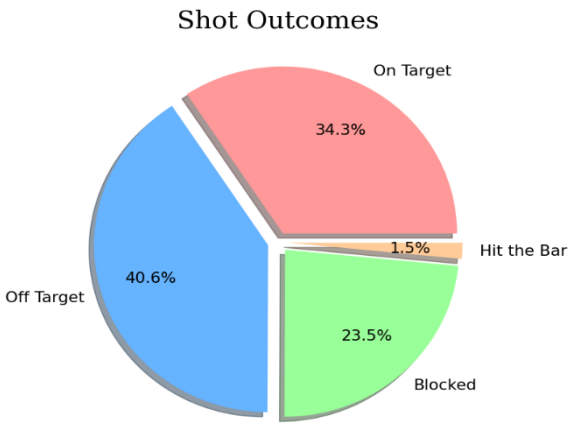
6. GoalImpact: This model focuses not only on the xG of individual shots but also on the overall impact of players and their contributions to creating and preventing chances. It's particularly useful for player scouting and performance analysis.

7. The Analyst (formerly known as Opta Analyst): This model is known for its detailed breakdowns and advanced metrics beyond just xG, such as Expected Assists (xA) and Expected Points (xP). It's widely respected in the football analytics community.

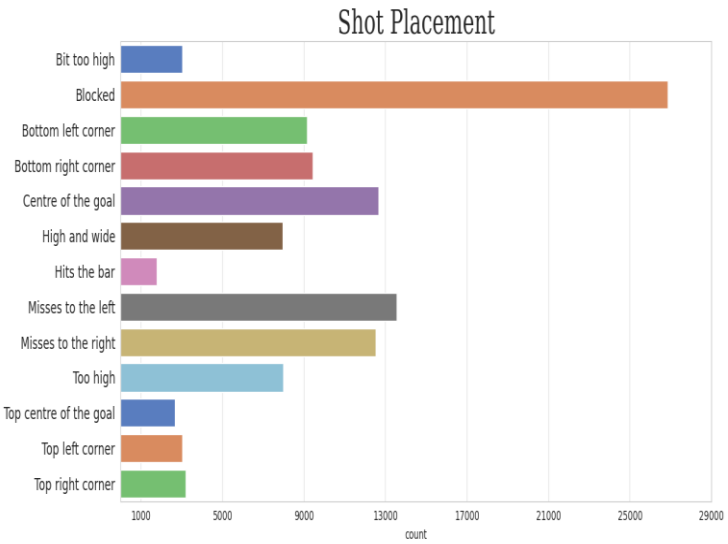
Each of these models has its strengths and may be preferred depending on the specific requirements of the analysis, such as the level of detail needed, the focus on team vs. player performance, and accessibility of the data. For the most thorough analysis, using a combination of these models can provide the best insights.

III. DATA

I used publicly available dataset for this research, which is the largest one that I could find that involves everything needed for creating a good XG model. In the dataset that we use we have information about over 900 000 actions from matches of Europe's top five leagues in the time span of 2011-2016. These actions are divided in 15 types of events from which we just need the shots and they are additionally divided in respect to the placement of the shot (13), outcome of the shot (4), location of the shot (19), body part with which the shot was taken (3), the assist method (5), and the situation in which the shot was taken (4). The mentioned dataset that we use in this paper Is genuinely a magnificent one because it does not contain any null values or anything that we need to deal with in that regard. There are only 2 things that we need to deal with and that is extracting the attempts (shots) from the other actions and then sampling the same in regard to their is_goal feature, more precisely keeping the same distribution of goals and not goals. After all of this we ended up with 114 567 shots from which 12 220 resulted in a goal or approximately 10.6% this will be of big importance for us in our further work. We can view picture 3.1 and picture 3.2 for better visualization and understanding of the nature of the data that we work with



Picture 3. 1



Picture 3. 2

As we can see from the pictures a lot of shots do not end up in goals which makes it very easy to predict non-goal attempts, more precisely not knowing the context of a shot if we say that it has a negative outcome (not a goal) we would right almost 9 out of 10 times which on the other hand makes it very hard to predict when an attempt results in a positive outcome (goal). This is something we need to take into

consideration to not get too carried away with the accuracy of the model but more so its capability of correctly predicting a positive outcome, a goal.

IV METHODS

The shots that we sampled from all of the actions are used as an input to the Gradient boosting, Random forest and Logistic regression algorithms in order to create a model that will correctly estimate the outcome of a shot. For the development of the models I used hyperparameter optimization for finding the parameters on which the algorithms would perform best, and also I got the feature importance of every algorithm to get a grip of how the models estimate the value of the target variable. This was done because having played it and also having solid knowledge on how the game is played and watching a ton of it I wanted to see if any given algorithm would find a cop out mechanism in a form of a feature with whom it will get better results on paper but those particular methods, that particular way of work would not be suitable for real life implementation. I used some of the most popular sports machine learning algorithms and the reason for using every single one of them that I used are the following:

I used Gradient boosting due to its high predictive accuracy, ability to handle various data types, feature importance insights, flexibility, capacity to model non-linear relationships, robustness to overfitting, scalability, and strong community support. These qualities make it effective for capturing the complex patterns and nuances in football data.

I used Logistic regression due to its simplicity, interpretability, ease of implementation, and efficiency. It provides clear probabilistic outputs, works well with smaller datasets, and offers straightforward insights into the importance and influence of features.

I used Random forest because of its high accuracy, robustness to overfitting, ability to handle various data types, and feature importance insights. It is also relatively easy to tune and implement, making it effective for capturing complex patterns in football data.

V EXPERIMENTAL SETUP

In order to evaluate and compare the different algorithms I used ROC-AUC, Baseline PR-AUC, PR-AUC and Cohen Kappa.

The ROC-AUC measures the ability of a model to distinguish between classes. It's calculated as the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

$$Baseline\ PR-AUC = \frac{P}{P+N}$$

Where P is the number of positive samples and N is the number of negative samples

The PR-AUC measures the area under the Precision-Recall curve, which plots precision (positive predictive value) against recall (sensitivity) at various threshold settings. While there isn't a simple closed-form formula for PR-AUC, it is often computed using numerical integration methods such as the trapezoidal rule.

$$Cohen\ Kappa: k = \frac{P_o - P_e}{1 - P_e}$$

P_0 is the observed agreement between the raters (the proportion of times both raters agree).

P_e is the expected agreement by chance, calculated as: $P_c = \frac{1}{N^2}(\sum_{\{k=1\}}^{\{K\}} n_{k1} * n_{k2})$

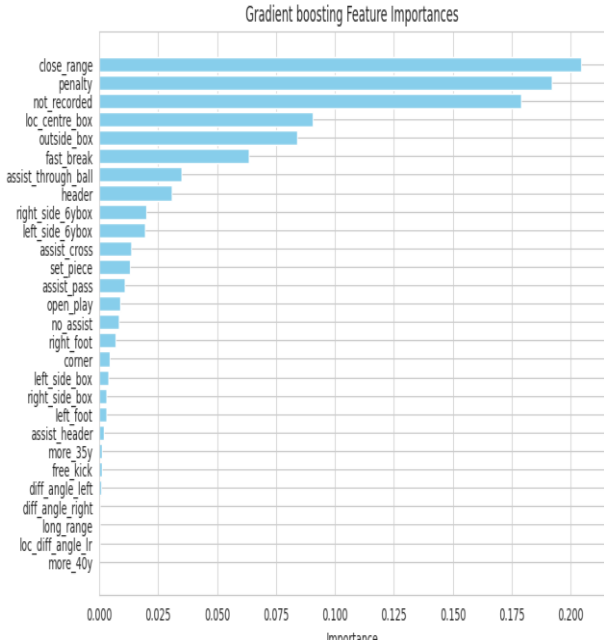
VI. EXPERIMENTAL RESULTS

The results obtained from the testing of the algorithms are shown in Table 6.1

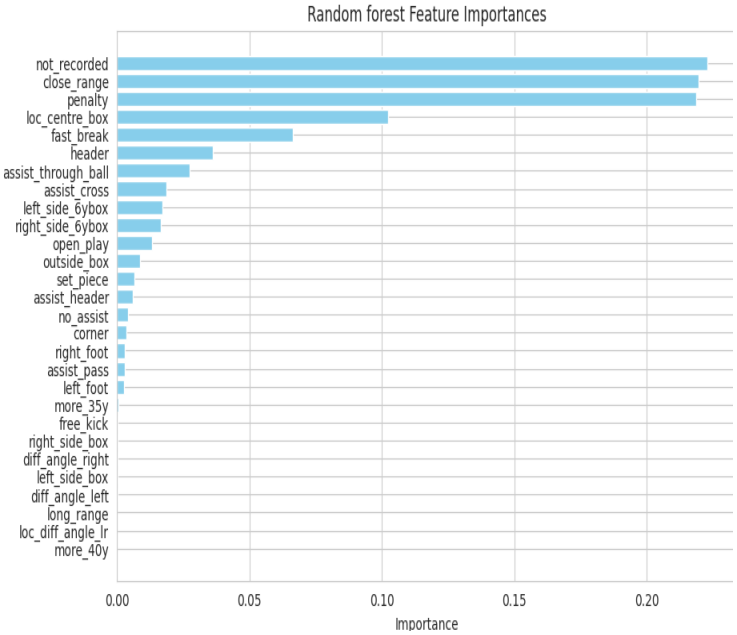
Metric	Gradient Boosting	Random forest	Logistic regression
Accuracy	91%	91%	91%
ROC-AUC	81%	81%	81%
Baseline PR-AUC	11%	11%	11%
PR-AUC	47%	46%	46%
Cohen Kappa	0.35	0.34	0.34

Table 6. 1

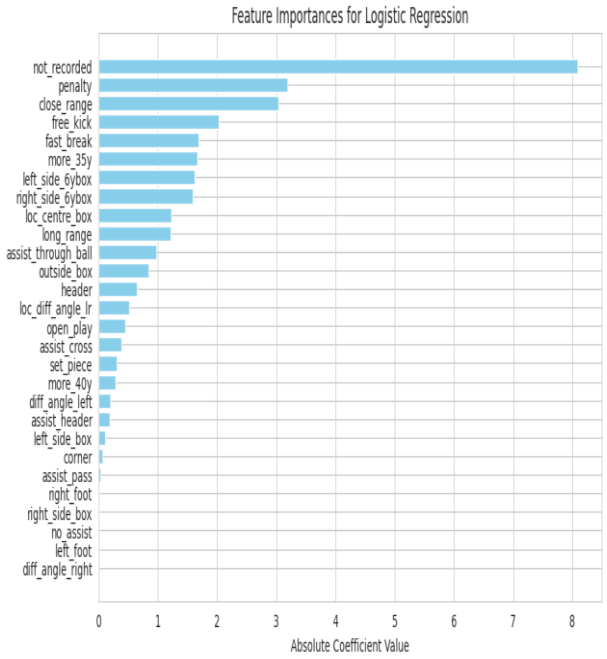
When only looking at these numbers we see that with every single technique we evaluate our models we get very similar numbers, but as I mentioned before in this paper, I wanted to dig a bit deeper and with the extraction of the feature importances see a bit into how the models operate and right there one model stood out in comparison to the other two.



Picture 6. 1



Picture 6. 3



Picture 6. 2

As we can see from pictures 6.1, 6.2 and 6.3 the Gradient boosting model greatly varies from the other two in that the feature not_recorded does not have as high importance as in the other two models. This feature is the only anomaly of the dataset which sadly is not repairable because if we remove this feature we lose valuable information from the dataset on lots of players. We can say that it is unlucky that this is the case but none the less we have created a decent model and that is also shown when we get the correlation between true and expected goals and the numbers show that these two corelate with 96% correlation which is a great percentage.

VII. CONCLUSION

In this paper we created three models for predicting Expected goals, so that with those results we can perform player analysis and see who were the best performers in terms of outperforming their XG numbers in the period of 2011-2016 in Europe's top five leagues. All three models worked well and showed solid, good numbers but the Gradient Boosting model faced off the best against the anomaly that appeared in a way of the not_recorded feature, which was a feature in which the location from which the attempt was taken was not recorded. Even the other two models that leaned more on this feature were not wrong in doing that because a lot of the not recorded attempts ended up as goals, so they were right in their predictions that they were goals. All the models show that it is pretty hard to predict whether an attempt would have a positive outcome not knowing the history of the shots taken by that individual player and his surroundings, the players around him, the position of the goalkeeper and other features that are used in newer XG models. Finally to conclude I will say that we did a good job, creating pretty solid models which with further features, as the ones mentioned before plus adding coordinates for player location instead of verbal explanation will give even better results knowing that machines work better with numbers compared to plain text.