Anticipating Goal Scoring Outcomes using Shot Information in Soccer

Chirag Sreenivasamurthy Panchakshari, Karthik Garimella, Sandeep Alfred, William Roland

Abstract

Soccer, like all sports, is played to win. Teams do their best to score as many points as possible and achieve victory. In order to increase the chances of victory, this paper explores the possibility of using machine learning to increase a team's chances of achieving those aims by giving them indicators for where shots should be taken. With this in mind, data on the top 5 European Soccer Leagues was used to analyze several features, such as Expected Goals (xG), Shot Type, Last Action, etc., to compare XGBoost, Logistic regression, ADABoost and Random Forests via grid search and gauge their accuracy in determining the desired target: Goal or No Goal. The resulting best model, ADABoost, was then used to indicate the probability of scoring a goal from a given position on the pitch with a heat map. ADABoost gave an accuracy of 91%, precision of 89% and recall of 93% for No Goal, and precision of 93% and recall of 89% for Goal indicating that there was room for improvement, but it is rather good at prediction for this type of data. The principal objective of the paper is to explore machine learning's potential role in giving coaches and players insights to improve their game and guide future work in which models may be worth exploring further.

Introduction

In soccer, more widely known as football globally, strikers are defined as the goal-scoring threat for any opposing team. Strikers are the primary source of goals; however, most strikes do not lead to a goal, even for the best players. While taking shots on goal is an important duty, strikers are often tasked with delaying tactics when they receive the ball in the opponent's zone to let their teammates converge and set up better plays. They also contribute to the team's attacking link-up plays, put pressure on the opponents to increase the chance of retaining control of the ball during rebounds, help with clearing aerial balls in defense during corners, and usually have the highest chance of taking penalty shots. It is known that goals are the most important aspect in football because all matches are decided on which team scored the most.

The top 5 European soccer leagues (English Premier League, La Liga, Bundesliga, Serie A and Ligue 1) (UEFA, 2024) consist of the highest quality strikers, yet the conversion rate (percentage of shots taken leading to a lead) is barely above 30% (Scholars Den, 2022). Soccer is the lowest scoring game when compared to basketball, ice hockey, tennis and table tennis. A soccer match is played in 2 halves with each half consisting of at least 45 minutes of play time excluding added time for things such as penalties, time-outs, overtime, injuries, etc.. The field size of a soccer pitch is larger when compared to almost any other sport which leads to the game being played mostly away from the goal posts on either end of the field. This larger field makes it challenging for the players to attempt shots because almost every shot requires a large chain of events to reach a conclusion.

One of the primary focuses of any team is to not concede any goals because a goalless game can still mean points for both teams in the league table. For this reason, soccer formations usually have more players in defense than in attack, i.e., a 4–3-3 formation of 4 defenders, 2 midfielders that focus equally on attack and defense and 1 midfielder that focuses more on defensive support, and 3 attackers. Goalkeepers are usually excluded from the formation since every team fields one in a typical game; although a 5-3-2 formation with 5 defenders may be seen in rare cases. Any in-game actions taken by players occur within the confines of the game's rules. Even with these rules, the unpredictability of real-world conditions and direct pressure from opponents on the pitch mean that strikers have limited time to take shots before they risk being dispossessed by the opposition. Every split second decision made without an instinctual strategy can lead to a wasted shot. A good coach utilizes informed decisions to hone those instincts during training to increase the chances of scoring goals. This analysis could help coaches train their strikers to improve their chances of scoring goals on the pitch in an official game which can lead to titles.

Every shot taken on the pitch has an X,Y coordinate and additional information such as which body part was used to take the shot, and from which circumstances the play was made, i.e. a corner, a freekick, in open play, a counter-attack, etc. The metric Expected Goals (xG) gives further insight on the quality of the shot taken. Expected goals measure the quality of a chance by calculating the likelihood that it will be scored by using information on

similar shots in the past. This metric ranges from 0-1, where zero indicates an impossible chance of scoring and one indicates a player scoring every time. It takes into account every aspect of the action before a shot attempt.

Every shot has a situation leading up to it with what kind of scenario, how the shot was taken, the xG of the shot and how the player received the ball before attempting the shot. Different machine learning models were employed through a Grid Search with multiple tuning hyperparameters to determine which model achieves the best training accuracy. The best model is then selected for testing to achieve a high testing accuracy with a focus on the recall value specifically.

Related Work

Scarcity and quality of data makes it difficult to develop appropriate models which can be used for predicting or assessing soccer teams or players. The shots information was obtained from the German Bundesliga season of 2019/2020 and 105, 627 shots were considered. The major challenge was synchronizing the event tracking data with the shots taken. Own-goals and penalties were discarded from the data in pre-preprocessing as the situations are entirely different from the in-game play. Each teams' shooting behavior was examined by taking shots and xG values of those shots into consideration.

The models tested in (Anzer G and Bauer P, 2021) were XGBoost, Logistic regression, ADABoost and Random Forests. The features taken were changed for different iterations to test accuracy of their models. Different models were considered for different situations as well including leg-shot, header and direct freekick. To investigate their efficiency, they tested on all shots taken instead of filtering the data on specific shot types.

The density of a shot taken from a specific position for an elite team; the distance of the closest defender was correlated with improved shot placements, with shots from a close range are affected by the site of the goal and the shots from longer distances did not have any positional variables correlated with the shot outcome (Schulze, E., Mendes, B., Maurício, N., Furtado, B., Cesário, N., Carriço, S., & Meyer, T. , 2018).

A soccer pitch can be divided into multiple zones where each zone can describe the danger of a goal being scored with the angle, opposition players and goalkeeper positioning. The density of a shot is determined using geometry (Link D, Lang S, Seidenschwarz P, 2016).

Data

The data was scraped in two stages from the soccer database website Understat (Understat, 2024) to collect player shot information. First the team collected all striker information for players currently in the top 5 European soccer leagues (English Premier League, La Liga, Bundesliga, Serie A and Ligue 1). Using the unique ID of each player, each striker's shot history was scraped using the Selenium WebDriver (Selenium, 2024). The initial unfiltered data consisted of 111,031 records of 587 players over 8 features, listed in Table 1. The data was then filtered to each player's shot history within the last 5 years in the leagues, which is from 2018 to 2023; hence, the number of shots considered for this project is 98,723 for each feature.

Features	**Description**
X	X coordinate of the shot on the pitch {numeric}
Y	Y coordinate of the shot on the pitch {numeric}
xG	Probability of the player scoring a goal {numeric}
Home/Away	Whether the game was played home or away {categorical}
Situation	Which situation led to the shot being taken {categorical}
Shot Type	Which body part was used to take the shot {categorical}
Last Action	How the player taking a shot received the ball {categorical}
Result	Outcome, whether the shot resulted in a goal or not {categorical}

Table 1: Description of each considered feature.

The collected shot information, visualized in Figure 1, for a single player, in this case, Haaland, shows the probability of a shot leading to a goal as determined by the xG feature. The blue circles represent goals and white circles with stripes represent not-a-goal with the size of each displaying the probability of a goal from that location. From this

figure, it can be seen that not all high xG shots lead to a goal and some low xG shots do lead to a goal.

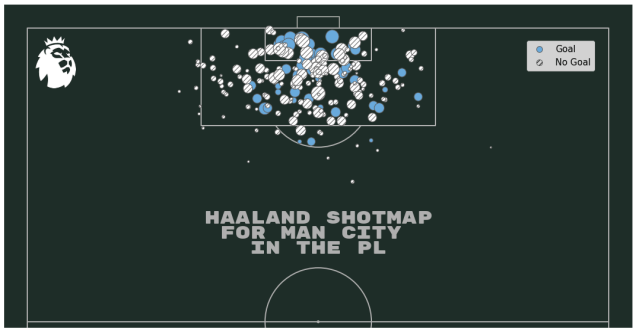

Figure 1: A striker's (Haaland) shotmap in the Premier League.

The target class, Result, in the dataset has 12,795 data points for shots resulting in goals with the other 85,928 data points belonging to the non-goal shots. Seen in Figure 2, there is a substantial class imbalance in the ratio of 3:25 for this class and provides insight on how shots taken lead to more non-goals than goals.

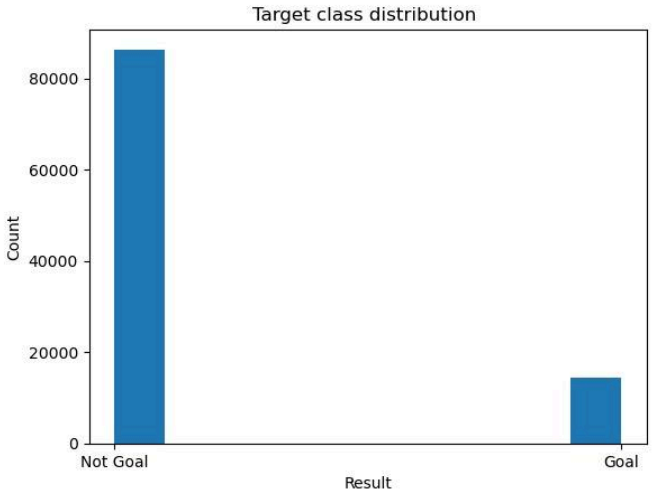

Figure 2: Bar plot representing the class imbalance for the target variable 'Result'

Methods

The first step of pre-processing of the data involved the removal of records with NaN values. For example, the Last Action feature had 10,246 records with NaN values with no information on how the player received the ball before taking a shot. Penalties taken by players were then excluded since penalties are a special case where only the player and goalkeeper are involved. Own-goals are also excluded as they cannot be considered as a genuine shot attempt by a player as it is unintentional.

The features Home/Away, Situation, Shot Type and Last Action are categorical variables. The Home/Away feature only has 2 classes: 'H' for home matches and 'A' for away matches. Last Action has 31 subsets which have been categorized into 4 top-level classes: 'Cross', 'Pass', 'Dispossessed' and 'Other'. An example of this consolidation are the classes 'Aerial', 'Cross' and 'Chipped' all being defined as 'Cross'. The Results feature has 'Missed Shot', 'Saved Shot', 'Blocked Shot', 'Shot on Post' and 'Goal' where 'Goal' is encoded as 1 and 0 for all other variables as they are shots that did not lead to a goal. All other categorical variables were one-hot encoded to make them suitable for training the models which increases the number of features to 14.

For the initial training of the models, 50% of the entire data is considered for inspection if the models achieve a high training accuracy. This was decided upon to make iterating faster during the beginning of the project. After the data transformation, the data is divided into 'features' for the models to train on and 'target' for predictions. An 80:20 train-test split was used to train the models and test the predictions. The data is split by giving equal weightage to each class.

A pipeline was employed with Logistic Regression, Support Vector Classifiers (SVC), XGBoost and ADABoost classifiers with a random state seed of 42 for reproducibility of the results. The Logistic Regression model hyperparameters considered were *solver* with 'saga' and 'lbfgs' parameters and *C* (inverse regularization strength) with 0.1 and 1 parameters. The SVC model hyperparameters were *C*, *gamma* and *degree*. The parameters considered for C are 0.1 and 1, for degree are 1 and 2, and for gamma are 0.1 and 1. For XGBoost classifier, *n_estimators* with 10 and 100 as parameters and *max_depth* with 10 and 50 as parameters. Regarding ADABoost classifier, the *base_estimator* with DecisionTreeClassifier, *learning_rate* with 0.01, 0.1, 1, *base_estimator__criterion* with 'gini' and 'entropy' and *base_estimator__max_depth* with 1, 5 and 9. All the 4 models with their various hyperparameters mentioned above were run through a grid search with a 5 fold cross validation to find the best model. After finding the best model with the parameters, the test data was used to predict the target class labels. The accuracy, precision and recall scores were checked and assessed for acceptability.

For improving the model accuracy, a higher cutoff taken from the data for the number of shots per player. This was considered as not all players take a high number of shots. Hence, the top 75 percentile of players with the most shots are accounted for. The data for the shots of these players

was trained on using the exact same models and hyperparameters. Another improvement that was tested was the ADASYN technique to address the class imbalance between 'Goal' and 'Not Goal'. ADASYN is a synthetic sampling technique to address the problem of class imbalance in datasets (Imbalanced Learn, 2024). This technique is designed to generate synthetic data points for the minority class. During testing, it was found that SVC was not the best class and was the slowest to train; therefore, it was replaced at this stage by a KNN supervised learning model for classification. The KNN hyperparameters are n_neigbors with 3, 4, 5 as the parameters and p with 1, 2 as parameters.

The resampled data is used in the new pipeline with the hyperparameters for training, returning the best model and acquiring the test scores. After checking the performance of the model on just half the dataset, the entire dataset was taken into account for training and testing. The test scores for accuracy, precision and recall were compared. The ROC curve, Figure 4, with the AUC score was also inspected.

The predicted labels were compared with the true labels using heatmaps. The heatmap, Figure 7, for the predicted labels '1' and '0' were added into the bins of the heatmap with a bin size of 35 which provided enough granularity to distinguish areas of significance. The heatmap was imposed on the top half of a soccer pitch in a grid with the goal lines, penalty areas, touch lines and corner arc. This plot gives a visual representation of how many shots taken in a given zone lead to a goal.

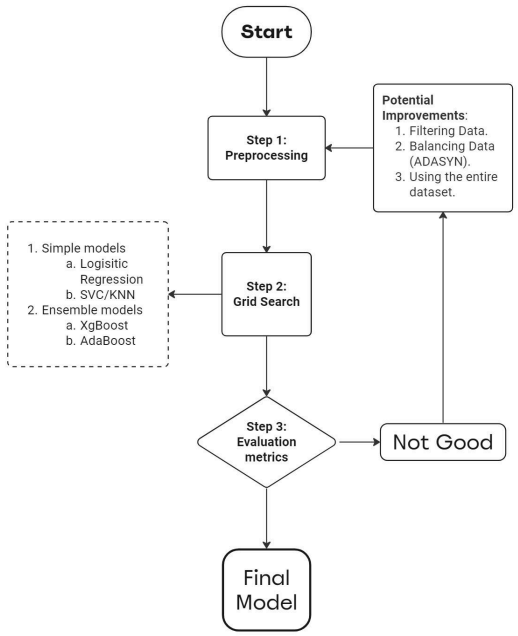

Figure 3: Project Flowchart

Experiments and Results

The Grid Search with Cross Validation of 5 on the initial run of only the preprocessed data, produced the ADABoost classifier as the best model. This is the baseline model for the project. The model achieved a decent accuracy score of 87%, a reasonable precision of 60% for goals, but performed poorly on recall with only 20% for goals. Accuracy, despite measuring overall correctness of the model, can be misleading in imbalanced datasets (Akosa, J.S., 2017). Given the heavy imbalance in the target variable 'Result', precision and recall metrics were used as the main evaluation criteria. In the confusion matrix, Figure 4, we get better insight into the exact classifications of each class such as the gross misclassification of 'Not Goals' as 'Goals' which has the direct consequence of a low recall for the minority class.

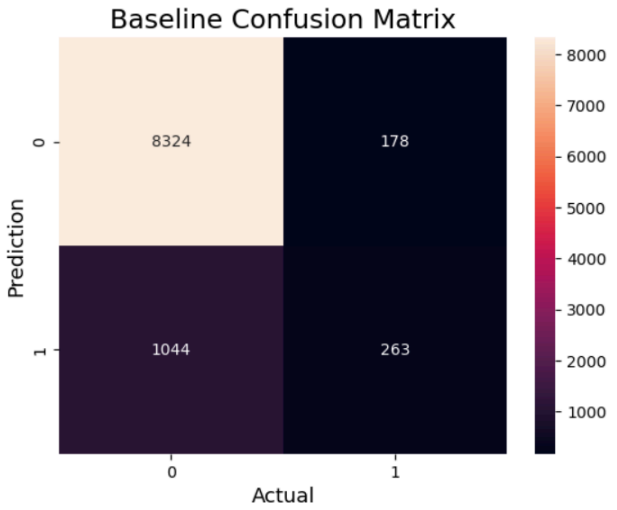

Figure 4: Baseline Model Confusion Matrix

To improve the recall of shots that lead to a goal, the first potential improvement, mentioned in Figure 3; i.e., filtering the data was considered. Filtering the data to players who have taken more shots was expected to account for some imbalance in the target variable. Running the Grid Search on the second iteration with the filtered data improved the recall by only 2% for goals from to 22%, and no changes were seen in precision for the goal class.

Simply filtering the data was not sufficient in order to improve both precision and recall of the minority class 'goal'; therefore, oversampling using ADASYN applied. This method helped balance the target variable 'Result' by increasing the number of shots that lead to a goal. Before oversampling, out of every 100 shots taken, about 15 goals were scored. After applying ADASYN, this number was increased to about 99 goals per 100 shots. The other improvement made was to replace SVC with KNN, since

SVC was never the best model in the grid search and was taking an excessive amount of execution time within the constraints of the project timeline which became several times worse when attempting to train on the entire data set.

With the aforementioned improvements the Grid Search was run for the third iteration. The precision of the goal class increased from 58% to 92%. A similar improvement was observed in recall of the goal class with a 60% increase to 87%. Balancing the dataset proved to be effective.

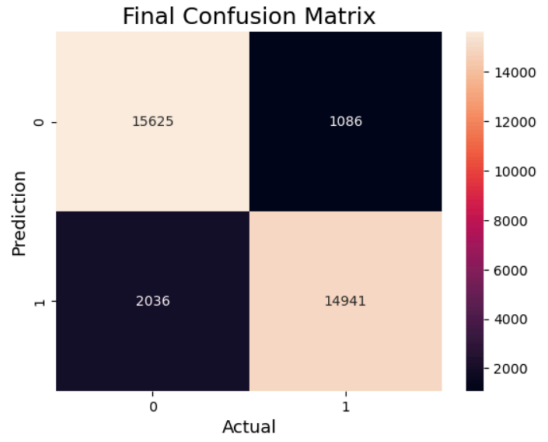

Figure 5: Final Model Confusion Matrix

For the fourth iteration of the Grid Search, the entire dataset of 98,723 observations was considered. The model's precision and recall of the goal class increased by 2% from the previous iteration, indicating the model is able to identify the goal class better, but not so significantly that it should be the first step in future work due to its increased time-to-train. With this final iteration, the confusion matrix in Figure 5 also shows vastly significant improvements over the baseline confusion matrix, Figure 4, with the majority of the classes being labeled correctly except for a few misclassifications.

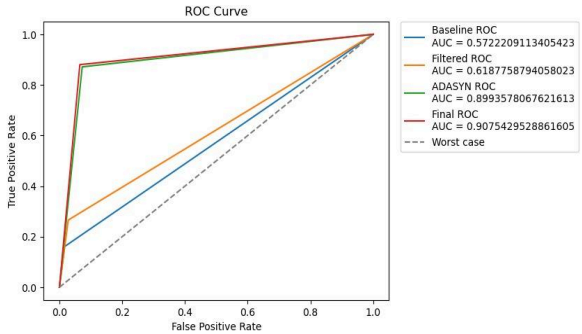

Figure 6: ROC-AUC curve for the Baseline, Filtered, ADASYN and Final models

The ROC-AUC curve in Figure 6 explains the trade-off between the true positive rate versus the false positive rate. Figure 4 shows that the Baseline ROC and Filtered ROC do not perform well for classifying predictions of the positive and negative across various thresholds. When compared with the worst case scenario, random, they are nearly random in their prediction capabilities. The ADASYN ROC and Final ROC perform exceptionally well and are very similar in identifying the positive classes as positive and negative classes as negative. This provides more confidence in our model selection for the binary classification problem.

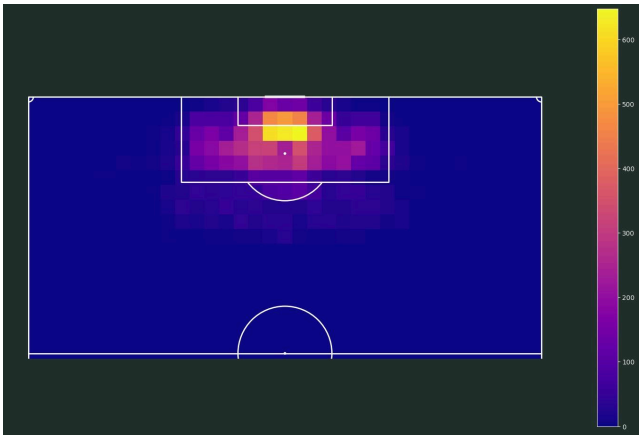

Figure 7: A heatmap of the number of shots predicted leading to a goal

The heatmap in Figure 7, is of shots taken on the soccer field that led to a goal with our predicted values. What can be inferred from the figure is that shots taken closer to the goal lead to goals more often than not for the brighter bins which refer to the zones on the soccer pitch. There is a lot of activity above the penalty spot, even after penalty shots are removed, which leads to the conclusion that more shots are taken from inside the penalty area closer to the goal due to its advantageous nature. There seems to be less activity in the zones extremely to the goal indicating that fewer shots are taken inches away from the goal. The figure signifies that shots taken from far away do not yield a better outcome when compared to shots taken further up the pitch. Another observation suggests that shots taken from tight angles, i.e., from zones where there are acute angles, tend to result in fewer goals being scored. This analysis agrees with the visual test as well.

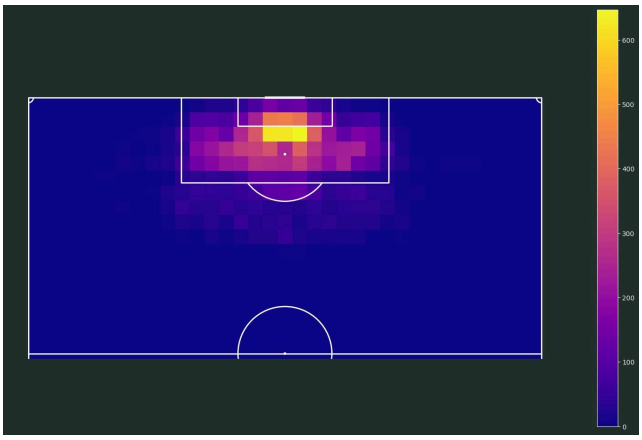

Figure 8: A heatmap of shots actually leading to a goal

Figure 8 is a heatmap of the shots that led to a goal using the true class labels. With the high accuracy rate, it meets expectations of being extremely similar to the predictions, Figure 7, due to its 90% accuracy. This comparison provides enough confidence in the predictions for how likely it is for a player in the top 5 men's European leagues to score a goal from a given position on the pitch with other factors taken into consideration that future work in this area seems fruitful.

Conclusions

The findings demonstrate that shots taken closer to the goals lead to a higher chance of scoring a goal, especially around the penalty spot. Although, in theory, the closer one player is closer to the goal, one would expect a higher number of goals scored from near the goal line. But the heatmap unveils that more shots are taken from nearby the penalty spot which seems convincing of the model's ability to make good predictions. It is more likely for a shot with high probability of scoring to occur around the penalty area as players have the most options to strike at the net from this location. Players would have fewer shot attempts right beside the goal line as the opposition would more likely contain or restrict the players to score from such a close distance.

The ADASYN technique was used to increase the recall value of the minor class as class imbalance affects the recall of the minority class by a substantial margin. The best model for the binary classification of 'Goal' or 'Not a Goal' is the ADABoost classifier with an overall accuracy of 91%, precision of 93% and 89% and recall of 89% and 93% for goals and non-goals respectively.

References

1. Anzer G and Bauer P (2021) A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer). *Front. Sports Act. Living* 3:624475. doi: 10.3389/fspor.2021.624475

2. UEFA *(*2024, April 23). *UEFA rankings Association club coefficients.* https://www.uefa.com/nationalassociations/uefarankings/country/?year=2024

3. Scholars Den (2022, December 14). *The Best Strikers in the World Based on Conversion Rate.* https://scholarsden.net/conversion-rate/

4. Understat (2024). https://understat.com/

5. Selenium (2024). *WebDriver.* https://www.selenium.dev/documentation/webdriver/

6. Schulze, E., Mendes, B., Maurício, N., Furtado, B., Cesário, N., Carriço, S., & Meyer, T. (2018). Effects of positional variables on shooting outcome in elite football. Science and Medicine in Football, 2(2), 93–100. https://doi.org/10.1080/24733938.2017.1383628

7. Link D, Lang S, Seidenschwarz P. Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data. PLoS One. 2016 Dec 30;11(12):e0168768. doi: 10.1371/journal.pone.0168768. PMID: 28036407; PMCID: PMC5201291.

8. Imbalanced Learn (2024), *Over-sampling* https://imbalanced-learn.org/stable/over_sampling.html

9. Akosa, J.S. (2017). Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data.