Chapter 1 INTRODUCTION

The rapid improvement in the technologies revolving around the collection, storage, and analysis of data has had a revolutionary effect on football analytics as well as many fields over the last couple of years. The easy accessibility of data provides a great potential to propose several key performance metrics. Some of these metrics include measuring several aspects of play such as pass evaluation, quantifying controlled space, evaluating shots, and goal-scoring opportunities through possession values. One of these prominent metrics is expected goal (xG) which is the most notable one in football talk shows on TV and end-of-match statistics nowadays. It is proposed to quantify the probability of a shot being the goal.

1.1. IMPORTANCE OF XG IN THE FIELD OF FOOTBALL

This metric is essential as it represents the low-scoring nature of football rather than the other sports. It is an ordinary story in football that a team dominated the game -they had many scoring opportunities- but could not score a goal, and the opponent won the match by converting one of the few goal opportunities to a goal they created. In this case, the xG is used as a useful indicator of the score. It can be defined as the mean of a large number of independent observations of a random variable which is the shots from the statistical point of view. It is also a good indicator which is used to predict the future team performance.

1.1.1. Challenges

One of the major challenges is that there are no publicly available datasets for training these football metric models. The dataset can only be accessed upon special request.
The data that is used for training the model is imbalanced. The vast variation between the majority and minority classes had to be dealt with. Balancing methods have a positive effect on the model's behavior.These methods for imbalanced datasets in

binary classification tasks are a very commonly used solution to improve the prediction performance of ML models.

1.1.2 Focus

The focus of this project is mainly on analyzing the xG (Expected Goal) metric in football more accurately and in an efficient manner. The factors like distance to the goal, angle to the goal, etc, are taken into consideration to build an effective machine-learning model. This is further made more impactful by deploying it on an interactive webpage.

1.1.3 Social Impact

It's actively used in predictive modeling which is utilized by bookmakers and professional gamblers. It also plays a vital role in scouting candidates for the position of Head Coach of a team, based on the effective football that they produce. xG is also an important tool in team analysis as well as player analysis, which gives in-depth details about their performance to fans and professionals alike.

1.2. SCOPE

One of the practical applications of the xG models, which is the main focus of this paper, is performance evaluation. The xG models could be for performance evaluation instead of match outcomes. Useful metrics calculated based on xG such as offensive and defensive ratios have proved to be extremely helpful in match analysis. This model could also be further tweaked to predict the probability of future goals which could be used in fantasy football and major betting.

Chapter 2 PROBLEM DEFINITION

2.1. PROBLEM

Football is the world's most popular sport played by over 250 million people in 200 different countries across the globe. It became a requirement to have various metrics to analyze the performance of players and the team as a whole on the pitch. Since football is a low-scoring game, it was proposed that the probability of a shot being the goal to be taken into consideration as one of the important analysis metrics. This came to be known as xG or Expected Goals metric which could help represent the low-scoring nature of football. An efficient machine learning model used for calculating the xG was the need of the hour. The data used to train the xG model is highly unbalanced. It causes poor prediction performance of the models on minority classes which is not ideal.

2.2. SOLUTION

In this report, we aim to propose an accurate xG model in terms of both majority and minority classes. Forester Auto ML tool, which uses tree-based classification models is put to use to implement this model with ease. With the additional help of XAI tools, we can explain a black-box machine learning model's behavior at the local and global levels. In this way, we can gather more information from the model, not only its prediction and also its behavior.

Chapter 3 LITERATURE REVIEW

The literature, or the major research report we referred to for this report is "Explainable Expected goal Models for Performance Analysis in Football Analytics**"** by Mustafa Cavus and Przemyslaw Biecek[1]. The report was published in 2022 and its basis is the Explainable AI tools (XAI) to explain the performance of xG in football analytics. The tools used to measure the xG has been closely followed in our model to get a better result. We incorporated the catboost and lightgbm model used in this paper into our model. The paper gives an overview on how to use xG in performance analysis.

Expected goals(xG) was proposed by Green[2] in, "Assessing the performance of premier league goalscorers. OptaPro Blog, 2012" to quantify the probability of a shot being a goal. Green[2] proposed the idea of a shot having a probalistic value of being a goal or not.

"How Data Availability Affects the Ability to Learn Good xG Models" by P. Robberechts and J. Davis[3] answers 3 questions on how data affects a xG Model, which are:"How much data is needed to train an accurate xG model?", "Are xG models league-specific?", and "Does data go out of date?" These questions helped in identifying the type of data and attributes needed for our model. The features used in our dataset is inspired from this paper.

"A goal scoring probability model for shots based on synchronized positional and event data in football" by G. Anzer and P. Bauer.[4] use three main strategies to train model:balancing the dataset, by using over or under-sampling methods using cost-sensitive learners, and using ensemble learning models.

"Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality" by A. Fairchild, K. Pelechrinis, and M. Kokkodis[5] focused on ways for evaluating the xG model goes beyond the accuracy offensive and defensive efficiency by comparing the xG metric with the actual goals.

"A Mathematics based new penalty area in football: tackling diving" by Morales and Caesar A[6] described how the distance to goal and the angle to the goal is calculated which we have used in our model.

Chapter 4 PROJECT DESCRIPTION

With the use of Forester AutoML, a tool to train tree-based classification models, we build a machine-learning model for predicting the probability of a goal from a shot taken on goal. The xG is an important metric as it helps to analyze the performance of both individual players and teams. Attack and defense ratios can also be calculated with this metric. The dataset worldFootballR is used in this project to train the model. It is in the form of a .csv file with many features like distance to goal, angle to goal, etc. As it uses AutoML, pre-processing steps like missing data imputation, encoding, or transformation are not provided. The data balancing is done by balancing methods that solve the skewed class proportions in the data.

Since a black box model is used to train data, XAI (explainable AI) is used further to explain the local-level and global-level behavior. The xG is then predicted accurately. For better understanding of the xG model, we have implemented this using an interactive webpage. The player positions and shot locations can be dragged and adjusted on the web page and it'll display the resulting xG.

Chapter 5 REQUIREMENTS

1) Understat shots dataset consisting of 315,430 shots across 7 years.

2) R 4.2.2 and R Studio

3) Prerequisites to scrape data

- R Studio
- worldfootballR package

4) Prerequisites to train the model

- Forester package
- DALEX package
- ROSE package
- ggplot2

5) Prerequisites to build GUI

- Tensorflowjs for backend
- Shiny.js for frontend
- Interactive SVG

6) Convert the model into tfjs format to fit the web's app directory

Chapter 6 METHODOLOGY

6.1. DATA DESCRIPTION

The dataset used should answer 3 significant questions regarding the model such as, "How much data is needed to train an accurate xG model?", "Are xG models league-specific?", and "Does data go out of date?" that may affect the performance of an xG model.

We focus in our paper on 315,430 shots-related event data (containing 33,656 goals ~ 10.66% of total shots) from the 12,655 matches in 7 seasons between 2014-15 and 2020-21 from the top-five European football leagues which are Serie A, Bundesliga, La Liga, English Premier League, Ligue 1. The dataset is collected from Understat using the R-package worldfootballR and excluded the 1,012 shots resulting in own goals due to their unrelated pattern from the model concept.

Table 6.1.: List of features used to train our model

Features	Type	Description
Status	Categorical	situation that the shot is being a goal (0: no goal, 1: goal)
Minute	Continuous	Minute of shot between 1 and 90+possible extra time
Home and away	Categorical	Status of the shooting team
situation	Categorical	Event type(Direct Free Kick, From corner-kick, open play, penalty, set play)
Last action	Categorical	Last Action before the shot(pass, rebound,cross, head pass)
Distance to goal	Continuous	Distance from where shot was taken to goal line

Features	Type	Description
Status	Categorical	situation that the shot is being a goal (0: no goal, 1: goal)
Angle to goal	Continuous	Angle of the throw to the goal line
Player Position	Continuous	Distance of the player corresponding to the shot location for defenders and goalkeepers

6.2. PRE-PROCESSING THE DATA

Standardizing the football pitch of L=105m and W=68m in size.

Distance to goal (X^{DTG}) = $\sqrt{[}105-(L \times 105)]^2 + [34-(W \times 68)^2]$

Angle to goal (X^{ATG}) = $\left|\frac{a}{b} \times \frac{180}{\pi}\right|$,

a = arctan[7.32 $\times$ [$105 - (L \times 105)$]],

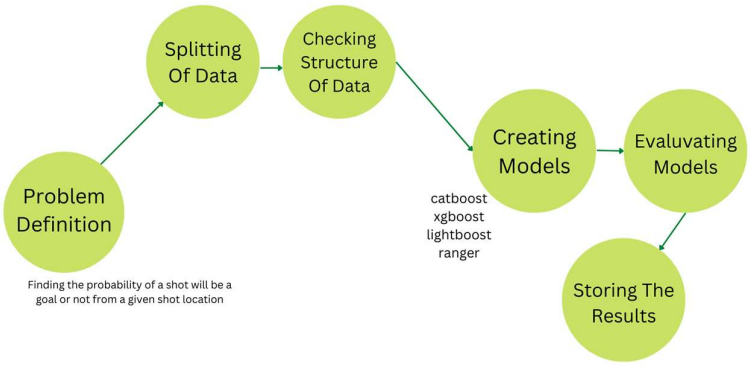
b=[105-(L$\times$ 105)]^2 + [34-(W$\times$ 66)]^2 -(7.32/2)^2

Filtering out the own goals from the dataset and selecting the features from the dataset to train the model.

6.3. MODEL TRAINING

We use the Forester AutoML tool to train various tree-based classification models like xgboost, randomforest, LightGBM, and CatBoost libraries. There are no pre-processing steps like missing data imputation, encoding, or transformation, and show quite good performance in the presence of outliers in the dataset to train the models.

Fig 6.1.: Flowchart of Forester model



We split the data to test and train (80-20) and validate the models to check the structure of the data. Due to a problem of Imbalancedness in the data, we use a type of imbalance learning where there are three strategies to train the model:

1)Using Balanced Dataset

2)Using Over Sampling Dataset

3)Using Under Sampling Dataset

Chapter 7 EXPERIMENTATION

The Data from Understat is split into 80-20 ratio with 252,344 shots used in the training model. Another 63,086 shots have been used to test the built model.

A Summary of the seven season's shots data is given below which explains the total number of shots, mean of shots taken, number of goals, and the mean of the conversion rate of these shots.

Table 7.1.: Summary statistics of shots and goals

League	No of Matches	No of Shots	μShot	No of Goals	μGoal	Conversion Rate %
Bundesliga	2,141	55,129	25.7	6,161	2.88	11.2
La Liga	2,648	62,028	23.4	6,854	2.59	11.0
Ligue 1	2,557	61,053	23.9	6,438	2.52	10.5
Serie A	2,659	70,615	26.6	7,252	2.73	10.3
EPL	2,650	66,605	25.1	6,951	2.62	10.4
Mean	2,531	63,086	24.9	6,37	2.67	10.7
Total	12,655	315,430	-	33,656	-	-

Algorithm To Calculate Xg Value For A Player/Team

1: Input: X_i , y, a player / team.

2: Train an xG model: $y \sim f(X_i)$.

3: for $i \leftarrow 1$ to n_i do

4: Predict $f(X_i)$ for i = 1, 2, ..., n_i

5: end for

6: xGplayer/team = $\sum_{i=1}^{n_i} f(X_i)$

We use the forester AutoML tool to train various tree-based classification models from XGBoost, randomForest, LightGBM, and CatBoost libraries. These models do not provide any pre-processing steps like missing data imputation, encoding, or transformation and show quite good performance in the presence of outliers in the dataset which is used to train models.

Forester function used in our model:

Fig 7.1.: Forester function definition

```
forester <- function(data, target, type, metric = NULL, data_test = NULL, train_ratio = 0.8,
fill_na = TRUE, num_features = NULL, tune = FALSE, tune_iter = 20, refclass = NULL)
```

Parameters used in the forester function:

1) data: the training data set to create the model, should contain target column
2) target: name of the target column
3) type: Defined in the task, either regression or classification. Our model uses classification.
4) metric: metric used in the model(NULL).
5) data_test: Test data used to evaluate model performance
6) train_ratio: Proportion of Splitting data train over original dataset
7) fill_na: logical parameter(TRUE=removes missing values in target column)
8) num_featues: indicates most important feature(NULL)
9) tune: logical parameter(FALSE)
10) tune_iter: total number of times optimization step is repeated. Used when tune=TRUE.
11) refclass: NULL

Chapter 8 TESTING AND RESULTS

We trained our model using Forester AutoML tool for the following tree based models:

1) Random Forest
2) XGBoost
3) LightGBM
4) CatBoost

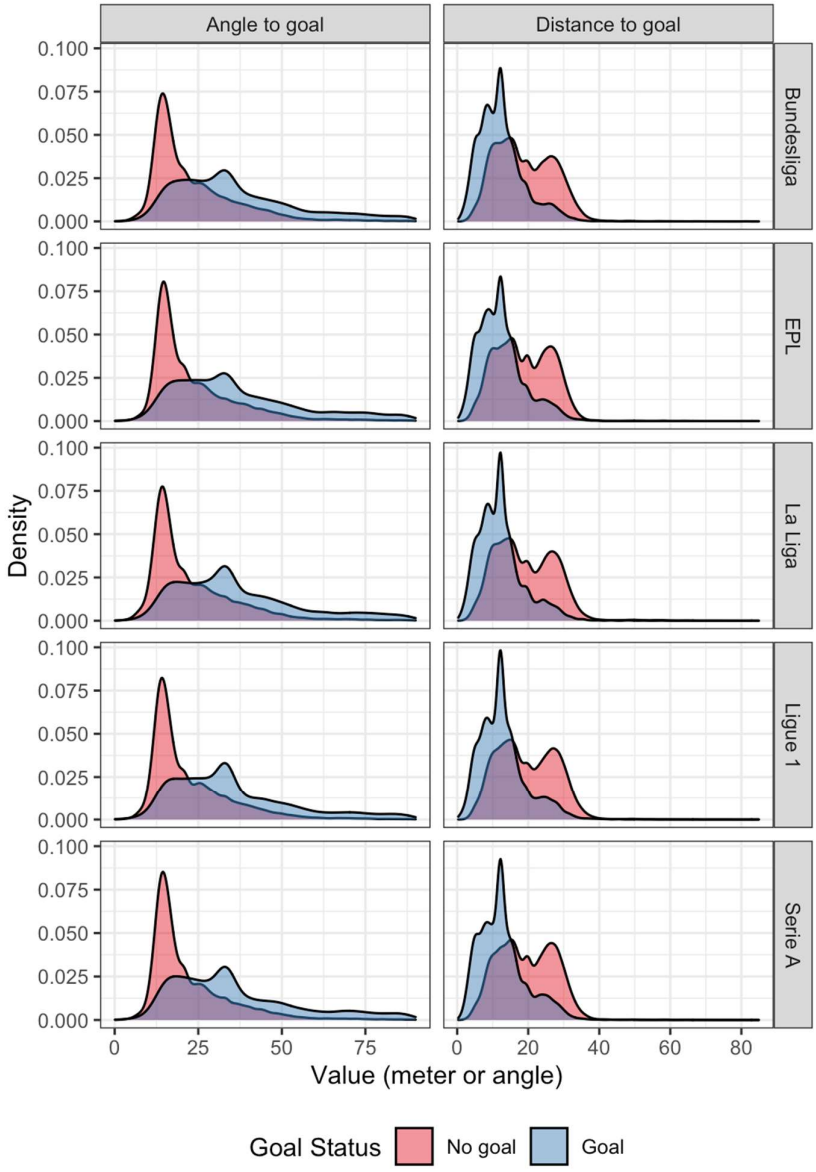
For the tree-based classifications, our model gave the following results in terms of Recall, Precision, Log Loss and Balanced Accuracy.

Table 8.1.: Performance of trained xG models

Model	Sampling	Recall	Precision	Log Loss	Accuracy
RandomForest	over	0.949	0.902	0.288	0.929
	under	0.842	0.879	0.364	0.836
	original	0.304	0.888	0.173	0.649
CatBoost	over	0.740	0.762	0.495	0.755
	under	0.728	0.756	0.507	0.746
	original	0.198	0.722	0.261	0.594
XGBoost	over	0.710	0.751	0.523	0.737
	under	0.717	0.751	0.521	0.739
	original	0.175	0.709	0.266	0.583
LightGBM	over	0.713	0.748	0.526	0.736
	under	0.714	0.748	0.525	0.736`
	original	0.179	0.689	0.267	0.585

From our training model, we found out that the distance and angle to goal for the five different leagues remain the same. We plot the density graph for the same below.

Fig 8.1.: The distribution of angle to goal and distance to goal of shots regarding goal status in the last seven seasons of top-five European football leagues



We built a GUI using Interactive SVG and TensforFlowJS which we used to integrate our model into the web app.

Following are screenshots from the web app where we can find the xG for the current shot location and how it changes if we move the defenders and goalkeeper.

Fig 8.2.: Screenshots of the web app

Fig 8.2.(a).

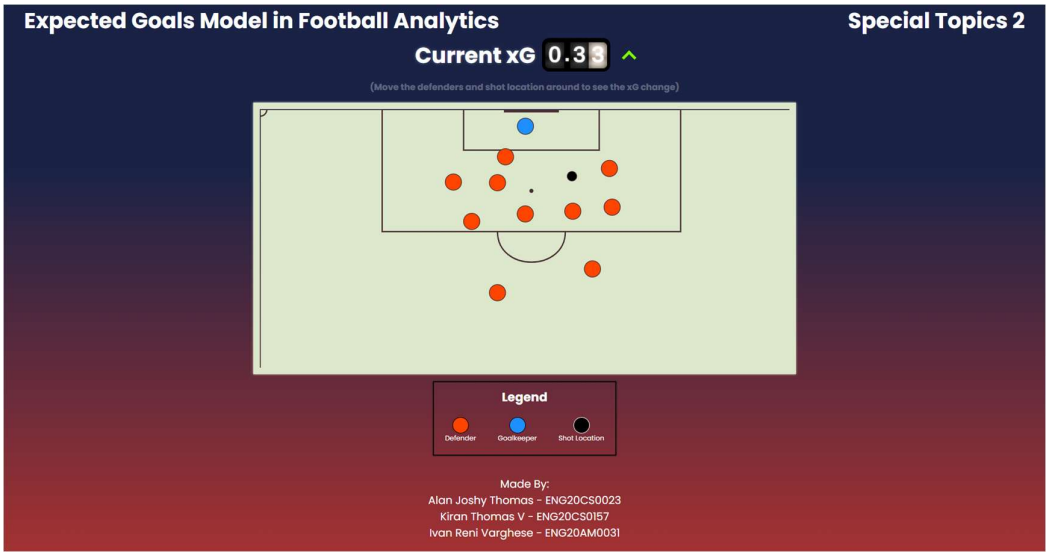
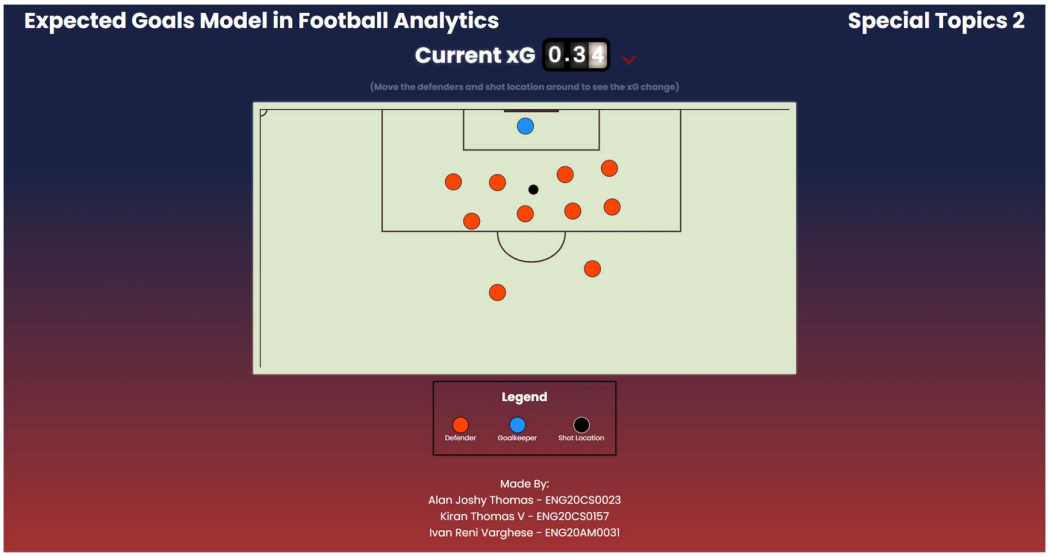


Fig 8.2.(b).



Chapter 9 CONCLUSION

The Forester AutoML tool is of great help in building a model. Most of the strenuous work is automated. The use of AutoML aids in easier machine-learning model building. For the four different machine-learning algorithms we used, Random Forest was found to have the highest accuracy of 92.9%. From the model, the already established conclusion of xG does not change with time and remains the same in spite of the country or league it is used in. Hence, the Random Forest model is preferred over CatBoost, XGBoost, and Light GBM.

By comparing the actual goals and expected goals, statistics is provided based on the difference for evaluating offensive and defensive performance of a team or a player. In this way, we can suggest how a team's performance can be improved by changing the strategies based on the features that affect the xG value such as distance to goal, angle to goal etc. We evaluated the performance both at the team level and player level by using the accurate xG model we proposed. Detailed information about the performance evaluation can be obtained with practical use of XAI tools on the xG model.

REFERENCES

[1] Mustafa Cavus and Przemyslaw Biecek . “Explainable Expected Goal Models for Performance Analysis in Football Analytics,” Front. Sports Act. Living, vol. 3, pp. 1–15, 2021.

[2] S. Green. “Assessing the performance of premier league goalscorer.” OptaPro Blog, 2012.

[3] P. Robberechts and J. Davids. “How Data Availability affects the ability to learn good xG models.” Machine Learning and Data Mining for Sports Analytics, MLSA 2020, Communications in Computer and Information Science, vol. 1324, Springer, Cham, 2020

[4] G. Anzer and P. Bauer. “A Goal scoring probability model for shots based on synchronized positional and event data in football.”Front. Sports Act. Living, vol. 3, pp. 1–15, 2021.

[5] A. Fairchild, K. Pelechrinis and M. Kokkodis. “Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality.”Journal of Sports Analytics, vol. 4, pp. 165–174, 2018.

[6] Morales and Caesar A. “A Mathematics based new penalty area in football: tackling diving”. Journal of Sports Sciences 34, no.24(2016):2233-2237.