Cranfield University

School of Aerospace, Transport and Manufacturing

Thesis Project

Analysing Expected Goals (xG) in Football

MSc. Computational and Software Techniques in Engineering

Authors:
Ayush Maria - s349256

Module Supervisor:
Dr. Jun Li

August 25, 2023

Contents

List of Figures

List of Tables

Nomenclature

xG	Expected Goals

Abstract

The influence of AI has been consistently growing in sports through distinct channels. Through the help of AI, teams and players have been able to gauge their performances and improvise their game-play. AI has taken concrete steps in football with the help of Data Science and Statistics to analyse and profile players and teams not only to improve performance but also optimise transfer market investment and player scouting. The purpose of this study is to develop a metric called Expected Goals (xG) to accomplish the aforementioned. This study has obtained data from StatsBomb, a company specialising in event data for football and built four different Machine Learning based models to analyse Expected Goals (xG) in the Spanish Football League a.k.a La Liga. The purpose of this study is to highlight existing features that impact xG values and identify newer features that can be used to improvise existing xG models.

Introduction

Football is one of the most well-known athletic activities in the world [9]. People affiliated with football clubs and teams, as well as fans , are occasionally faced with scenarios in which it is impossible to predict how the teams perform in critical junctures. This acts as the source of cognitive bias, outcome bias, and the all-important "availability heuristic" within these people's perceptions. [5]

Anticipating football match results may be a valuable technique for judging a club's or team's preparedness before a game. Though forecasting sounds like a deal maker, it could be in the betting industry. This is due to the fact that football is governed by chance. [5] Forecasts may be impacted not only by significant elements such as possession, amount of shoots, precision in the final third, and so on, but also by weather, turf conditions, player mindset, the number of fans present in the game, and a plethora of others. Such criteria can quickly reveal the underlying cause of any specific team's performance. To better react to the scenario, teams must use and adopt different plans during the game. The goal of this research project is to create a framework that analyses an essential statistic in football that centres around the chance of a football shooting chance from a specific location, hence improving football strategy.

Expected Goals (xG) is the name of this measure. It may be influenced by a plethora of variables, such as the range from the goal, whether it was shot with a stronger foot, the number of people surrounding the shooter and the goal, the shot angle, the build-up play to the goal, and more. Expected Goals is essential for determining player profiles, transfer market trends, and club tactics. This statistic has been used by several clubs and teams, including Liverpool F.C., and has assisted them in winning championships like the Premier League and the UEFA Champions League. Luis Diaz, who was bought by Liverpool F.C. in January and has already helped the team reach its second Champion's League final in three years, is the most notable example of a player Liverpool F.C. has profiled in the transfer window. [16]

1.1 Aims and Objectives

- Measure the quality of chances created by footballers and footballing teams
- Highlight the use of Expected Goals (xG) in player profiling and match strategy
- Introduce how AI based on Probability is used in football
- Establish a bedrock for future metrics

Literature Review

Numerous sports, like basketball [23] and baseball [2], have made use of artificial intelligence to glean insights from data and adjust their tactics. On the other hand, because it is a closed sector, football has only lately witnessed the integration of such technologies within its systems. Data was previously ignored in this sport since the hierarchies are driven by professionals like former players and scouts. [15]

A plethora of studies attempted to forecast football match results by taking into account characteristics of the tournaments or leagues these matches were organized in. Although initially promising, as shown in some of the publications mentioned below, this strategy does not aid football teams in formulating plans during live matches. Here are some examples of publications that explore this topic:

- Authors Huang and Chang [12] stated that their study used an Artificial Neural Network to the dataset from the 2006 FIFA World Cup and claimed that if the drawn matches are not taken into account, an accuracy of 76.9% may be attained. This investigation established a standard for football statistics. It is essential to remember that every football tournament is unique, and training a model for every competition is not creative.
- Authors Joseph, Fenton, and Neil [13] study aims to predict game outcomes for the Tottenham Hotspur football team. In addition to Decision Tree algorithm and the K-nearest neighbours method, a Bayesian network was used in the implementation. The Bayesian Network fared better than other algorithms in terms of prediction accuracy in this domain, according to the authors' research. The study's overall accuracy came at 59.21%. (win, draw or lose).
- Genetic programming was used by Cui et al. [7] to predict the outcomes of English Premier League games, which may end in a win, loss, or a tie. As an input to the Genetic Program system, they selected 25 traits from each game, and the system then created a function to predict the result. A single GP-generated function's prediction accuracy underwent an experimental test, and the results looked good. For each League game played for each season, this study manually fed 25 different characteristics. The shortcomings of forecasting football games are shown by this as well as the fact that the resultant accuracy was 70%.
- To find non-linear correlations, authors Rotshtein, Posner, and Rakityanskaya [22] created a model that uses fuzzy logic as the foundation. To fine-tune fuzzy rules and get acceptable simulation results, they used competition data. The tuning method involved determining the parameters of the fuzzy membership functions and rule weights using a combination of genetic and neural optimization approaches. The model is promising, but it does not take into account

variables like injured players, players who have been booked and benched, referee prejudice, and weather conditions. Additionally, owing to its nature, it only forecasts results based on previous matches.

These publications made it simple to understand how football match prediction models take into account variables that have nothing to do with the tactics used, but rather with the patterns and results of earlier games. This went against the objective of the case study.

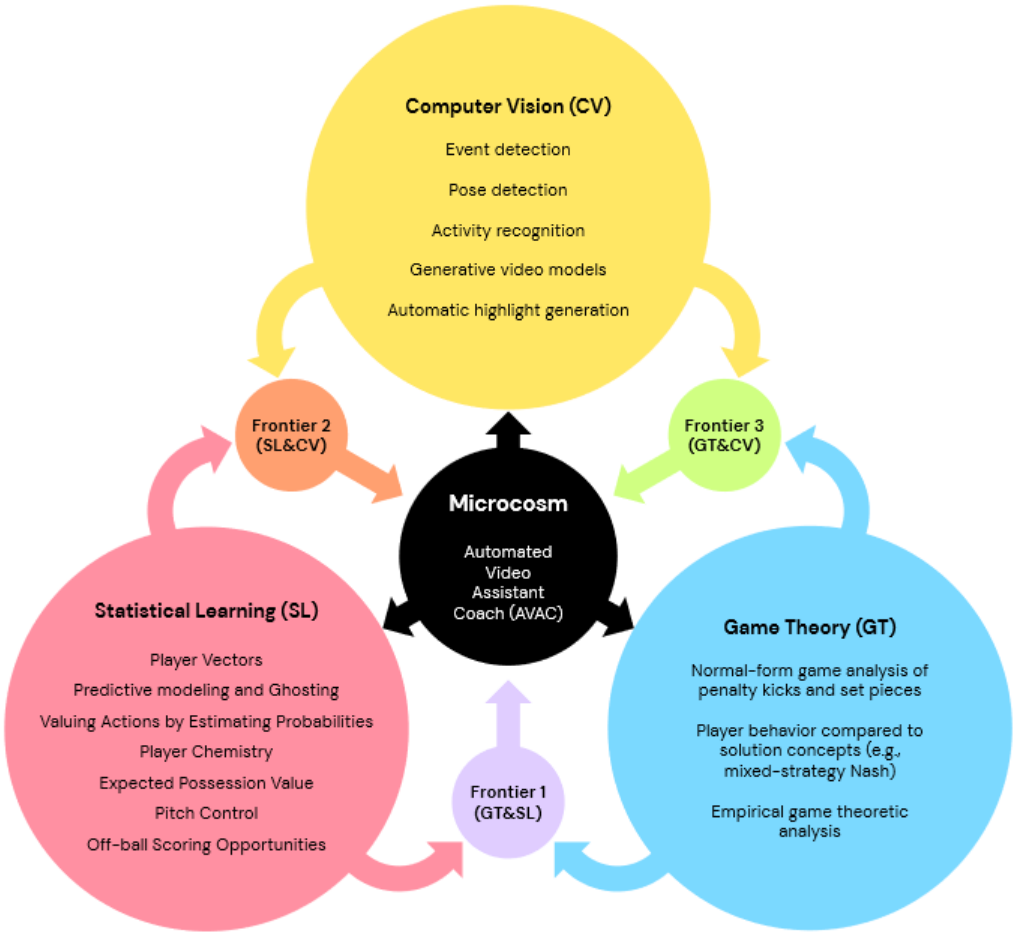


Figure 2.1: Proposed model by deepmind [26]

The study carried out by DeepMind Technologies, a division of Alphabet Inc. [26], illustrates the present impact of several branches of artificial intelligence on football analytics. As shown in the graphic 2.1, statistical learning, computer vision, and game theory work together to produce a possible system for football analytics that has the potential to spark a data revolution.

This case study sought to demonstrate statistical learning using the Expected Goal (xG) measure, which Vic Barnett and Sarah Hilditch initially introduced in their 1993 publication. [4]

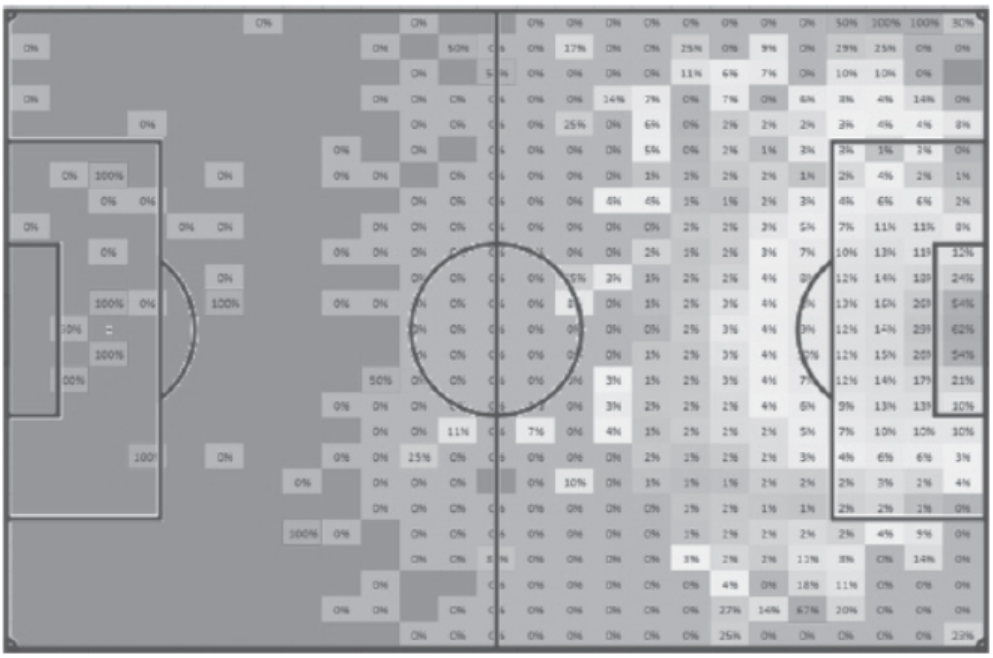


Figure 2.2: Shot Analysis
[5]

In his work, the author Biermann has offered a thorough explanation of how football strategies are developed using the xG measure. Football is a game of probability, as was stated in the previous section 1. The likelihood that a wonder goal will be scored from various locations on the football field is shown in the figure 2.2. The application of xG metric is demonstrated in this investigation. The shot's viewpoint was also taken into account in order to give this statistic additional context. For instance, shots from a distance of 12 yards have a higher likelihood of being successful than headers do, as do shots from counterattacks since the opposing defence is more unfocused in these circumstances. There are specific probability for attempts made in dead-ball situations as well.

Additionally, as seen in figure 2.3, the work [5] has also emphasised time series graphs that indicate how goals scored correspond to goal scoring possibilities.

These methods have aided managers and coaches in putting tactics into practise that have improved player placement and game play.

Rathke looked at goals scored in European football leagues to determine whether factors are related to predicting expected goals (xG). This idea helped them evaluate players, especially strikers, according to the number of goals they score year after year. As a consequence, this research examined shots from 380 and 306 Premier League and Bundesliga games played in the 2012–2013 seasons, respectively. On the playing field, all of the shots were separated into regions, and each region was given a hypothetical goal value. The shot's distance from the goal and angle with respect

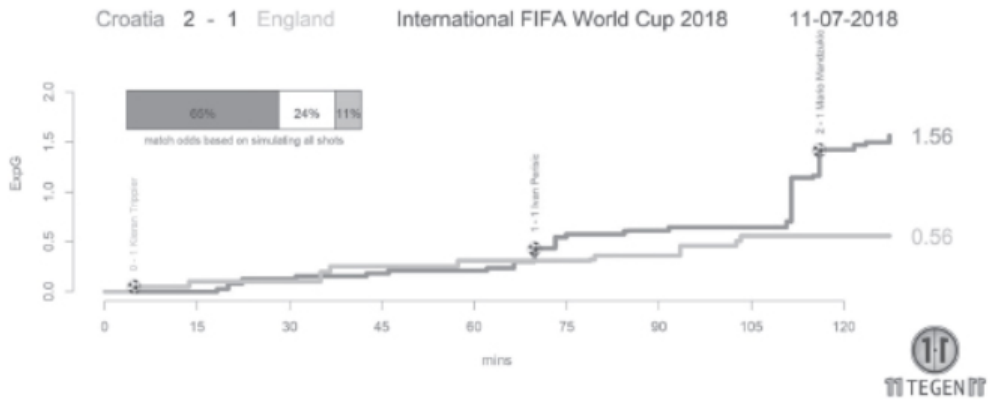


Figure 2.3: Time Series Graph
[5]

to the goal were two factors that were looked at. It was postulated that the shots' range and aiming angle had an impact on the calculation of xG. [21] The middle-of-the-table teams' goals and goals against were precisely identified by the algorithm. Teams at the top and bottom performed poorly, correspondingly.

The model created in [21] only took into account two aspects, but it may still be improved. Additionally, it hasn't been tested with newer data.

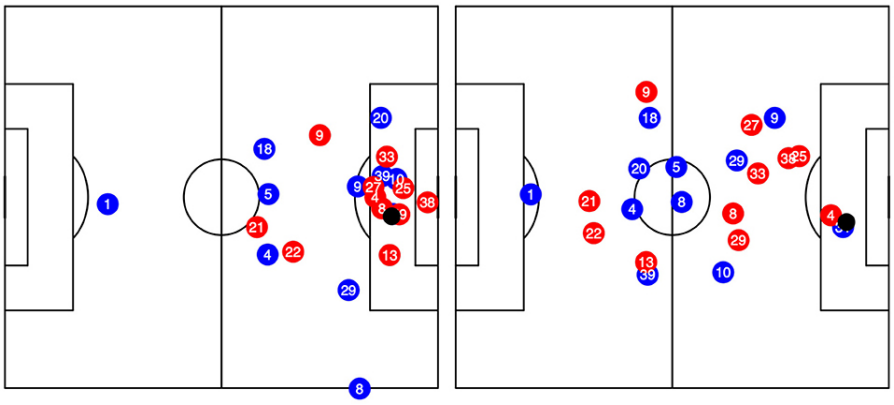


Figure 2.4: Gabriel Anzer and Pascal Bauer
[3]

The authors Anzer and Bauer assessed the calibre of each individual shot in their own model that investigates the xG metric. [3] Their strategy has been examined by seasoned football experts and scientifically proven. The best model uses the XGBoost methodology and was constructed using precisely calibrated parameters obtained from synchronizing positional and event data of 105,627 shots in the Football League played in Germany. More precision than any other model that has been formally published may be shown in this model. This model has a 0.197 RPS (ranked probability score).

Another study [17] considered the ability of the shooter from different data-sets. The features that this considered were, Long Shots, Finishing, Heading, Shooting and Weak Foot. Along with these features there were other features that includes, x and position of the player, player ability, post rebound, back pass etc. An ANN (Artificial Neural Network) was implemented along with XGBoost algorithm and Logistic Regression with XGBoost performing the best. Due to lack of data the models were rather poor in performance. The best performing model for this study was XGBoost Algorithm.

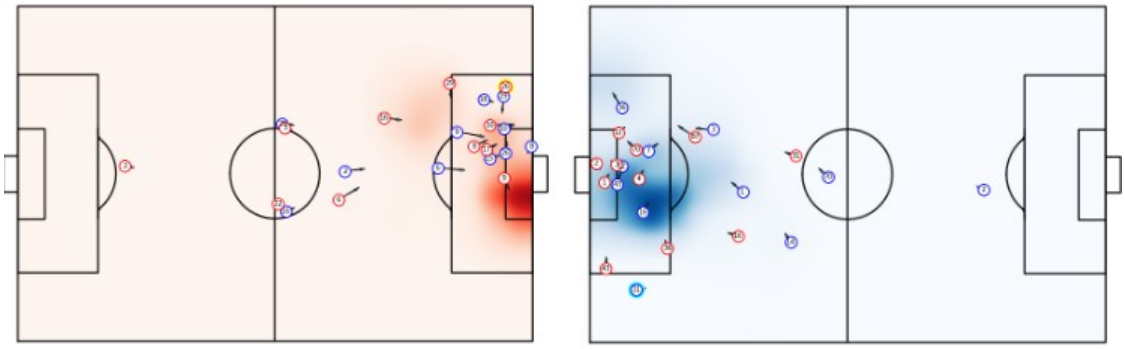


Figure 2.5: Time Series Graph
[24]

Many methods have been developed to assess the quality of soccer shots. In this work, we assess the efficiency of off-ball positioning prior to shoots that potentially result in goals. Consider a tall, unmarked centre forward who is positioned near the far post during a corner kick. When the cross comes in, the centre forward simply heads it in; other times, the cross sails over his head. Another example is a winger playing onside while breaking through the defensive line. Sometimes the through-ball arrives; other times, the winger may cut their run short when a teammate fails to provide a timely pass. Even though they never got the ball, the offensive player has generated a chance in both cases. The author Spearman [24] built a probability and physics-based model that leverages spatiotemporal player monitoring data to assess such off-ball scoring possibilities in this research (OBSO). This model may be used to predict which players, if any, are most expected to score at any moment throughout the game, as well as where on the field they are likely to score. The author demonstrates three significant applications of their model:

1. Identifying and analysing key opportunities throughout a game
2. to aid opposition analysis by indicating areas of the pitch where individual players or clubs are more likely to produce off-ball scoring chances.
3. to automate talent discovery by identifying players throughout a whole league who excel at providing off-ball scoring opportunities.

2.1 Summary

The studies analysed in this section consolidated the nature of this case study and confirmed that the direction undertaken is a fruitful path with a lot of potential answers to how data science can be used in football. Furthermore, the works analysed in this study also highlight the future work of the metric that was developed in this study.

Methodology

3.1 Data Specification

The data was obtained in .JSON format. Thus it was required to be cleaned off the values irrelevant to the purpose of this model. Furthermore it was needed to be restructured in a format which is accessible and easy to use.

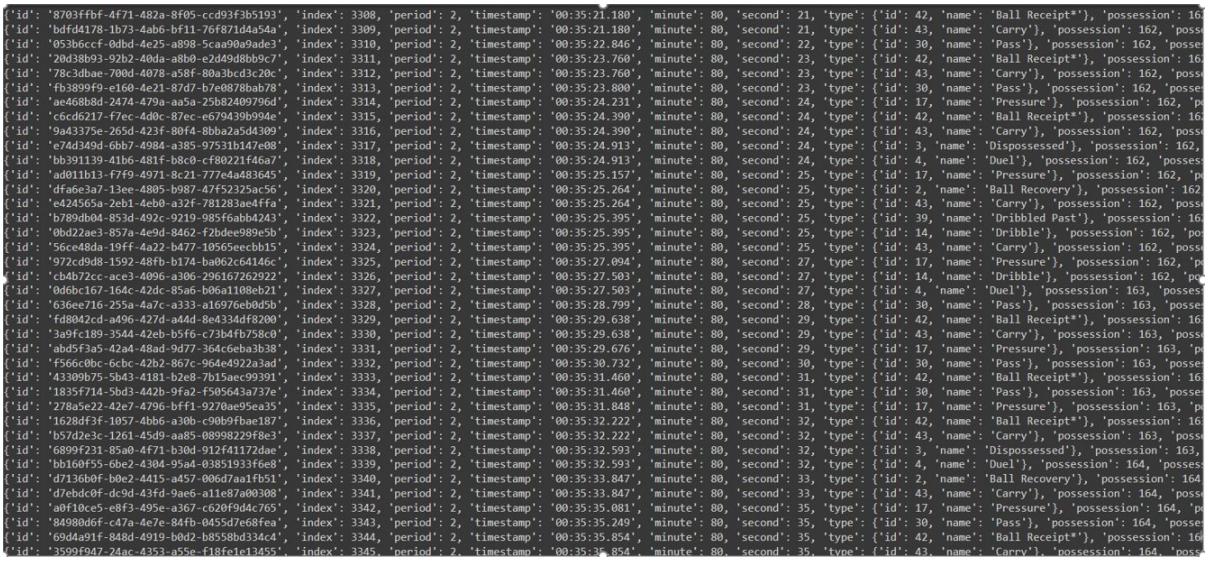


Figure 3.1: Event Data Format

A total of three .JSON files were obtained that included the following:

- Competition Data: Information of the competition
- Match Data: Folders containing the match data for each match within in season for a particular competition
- Event Data: Folders containing the events and lineups of each match

A visual of the event data obtained is highlighted in figure 3.1 where each row highlights the a specific event during a 90 minute football match. Some examples of match events are passes, dribbles, tackles etc.

The workflow and the methods utilised to develop the Expected Goals (xG) model include the following milestones as highlighted in the list below:

- Data Cleaning/Wrangling
- Exploratory Data Analysis

- Feature Engineering
- Model Testing and Hyper-Parameter Tuning
- Machine Learning

Furthermore, the data specification included important information regarding data visualisation. To visualise a football pitch, a specific scale of co-ordinates was necessary to follow. This is covered by figure 3.2 obtained from the Data Source i.e., StatsBomb [25].

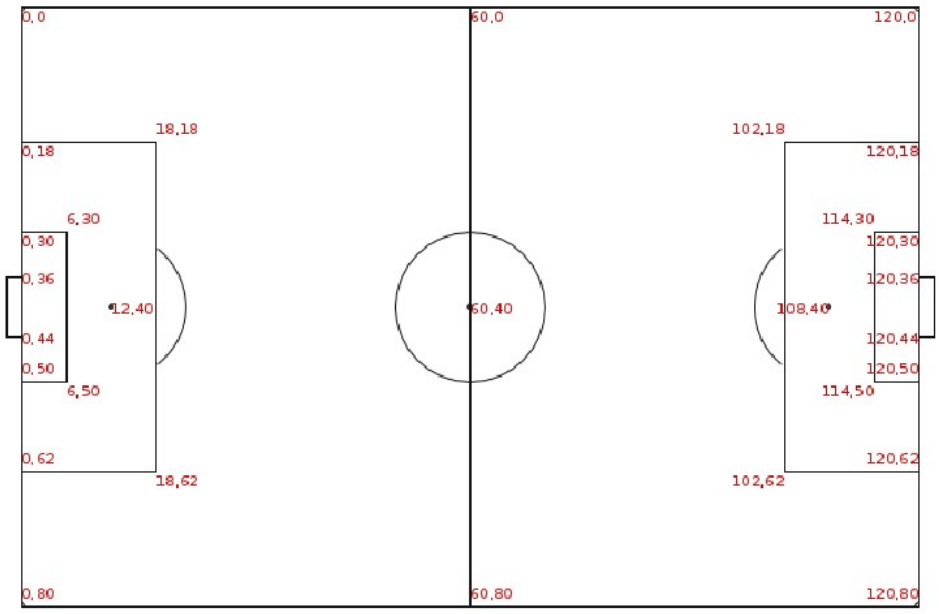


Figure 3.2: Pitch Visualisation Specification

3.2 Data Wrangling/Cleaning

Data Wrangling can be defined as a process that is used to handle raw data to convert it into meaningful information in order to clearly understand it before analysis.

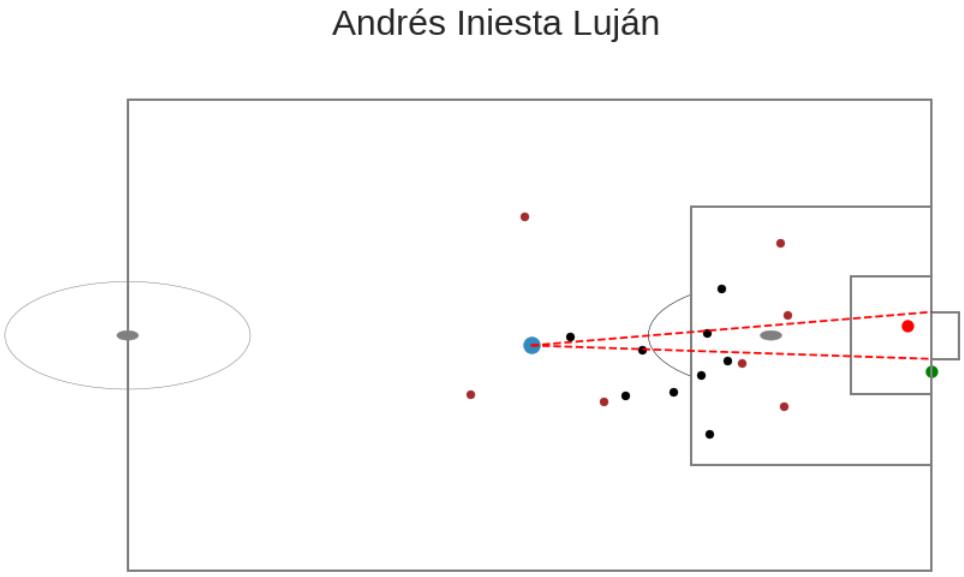


Figure 3.3: Football Pitch Visualisation

An example of the raw data obtained is highlighted in figure 3.1. The Expected Goals (xG) model being built required information related to football shots. The following data was extracted for this purpose:

- Shot Id: ID of the Shot Event
- Play Pattern: Play Pattern preceding the Shot
- X Location Shot: X Coordinate of the Shot
- Y Location Shot: Y Coordinate of the Shot
- Duration: The timestamp of the Shot
- Outcome: Outcome of the Shot
- Technique Used: The type of Technique that was used to shoot the ball
- First Time: Whether the ball was shot on First Touch or not
- X GK Location: X Coordinate of the Goalkeeper
- Y GK Location: Y Coordinate of the Goalkeeper
- Body Part: The body part used

- Type of Shot: If the Shot was a Free Kick, Open Play or Penalty
- Player Name: Name of the Player
- Team Name: Name of the Team
- Official xG: xG provided by StatsBomb
- Pass id: Event ID of the Pass prior to the Shot
- Pass Type: If the Pass is Aerial or not

Further features were engineered from the data above. They will be covered later in 3.28.

3.3 Exploratory Data Analysis

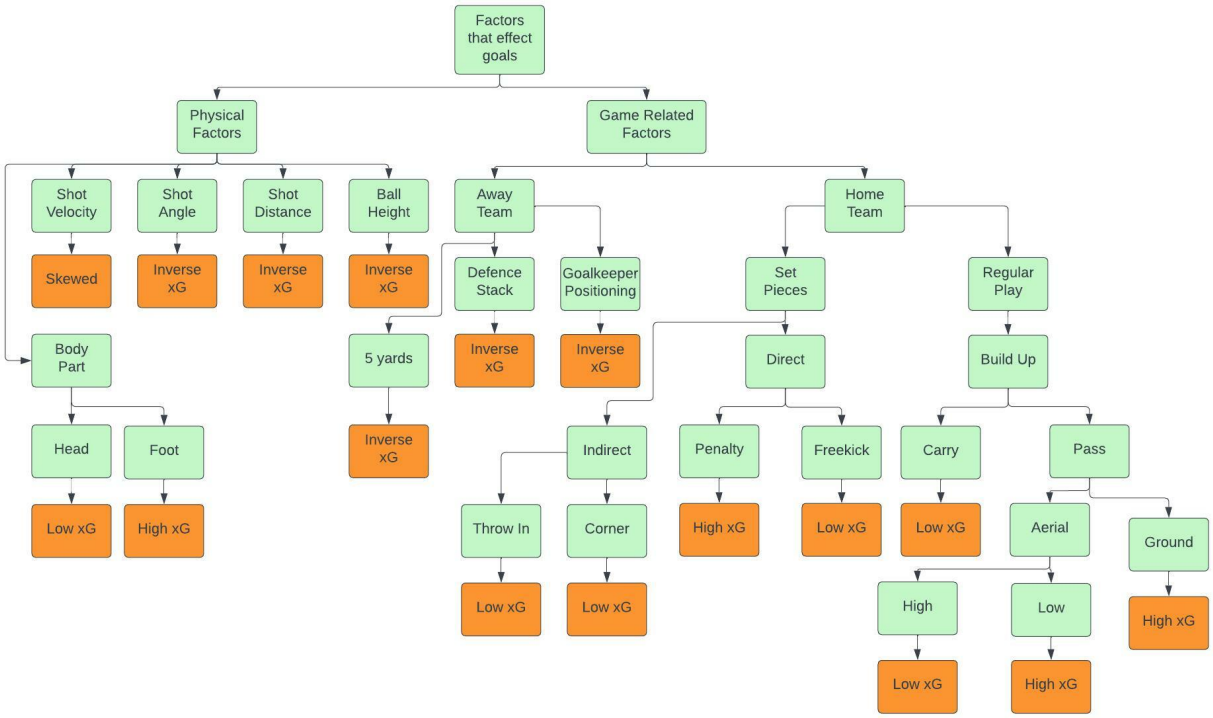


Figure 3.4: Hypothesis

After the data was wrangled and arranged in an acceptable format it was passed on to the next stage in the pipeline where important trends and insights were highlighted in order to determine which features were important for Machine Learning. This stage is called Exploratory Data Analysis.

In order to conduct this analysis, an initial hypothesis was created. This hypothesis was used to direct the analysis and provide a road map for reference. The road-map for feature analysis followed is envisioned in figure 3.4.

Each feature is analysed critically, starting with Shot Distance.

3.3.1 Shot Distance

Shot Distance is an important aspect in Football. It is intuitively understood that shot distance has a negative impact on goal scoring probability. This was stated in the null hypothesis in figure 3.4. This hypothesis was proven to hold as determined in figures 3.5 and 3.6.

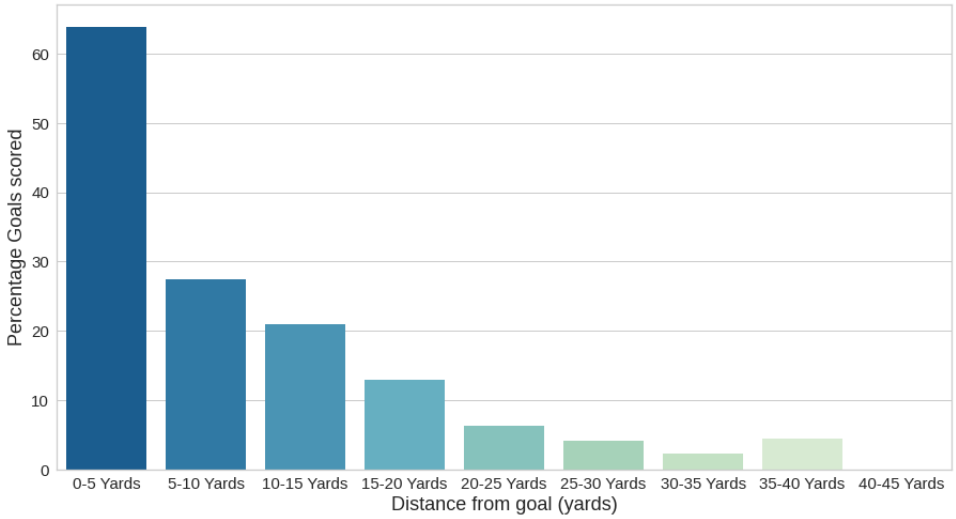


Figure 3.5: Goal Percentage from 5 yard Intervals

The decreasing trend in the number of goals per 5 yards can be witnessed in figure 3.5 which reaffirmed the null hypothesis further. Shot Distance was determined to be a critical feature to be considered.

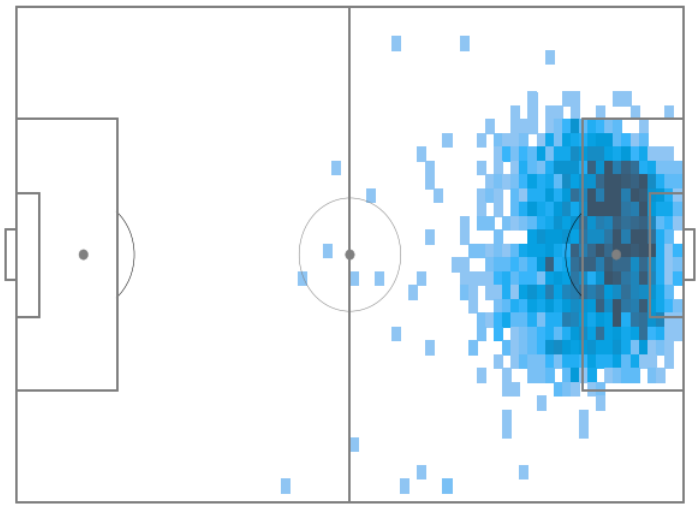


Figure 3.6: Shot Density

3.3.2 Shot Angle

Consider the triangle with its vertices being, the shot location and the two goal posts. The Shot Angle is the angle contained between the goal posts and the vertex at the shot location. Thus the angle considered is located by the blue dot in figure 3.7.

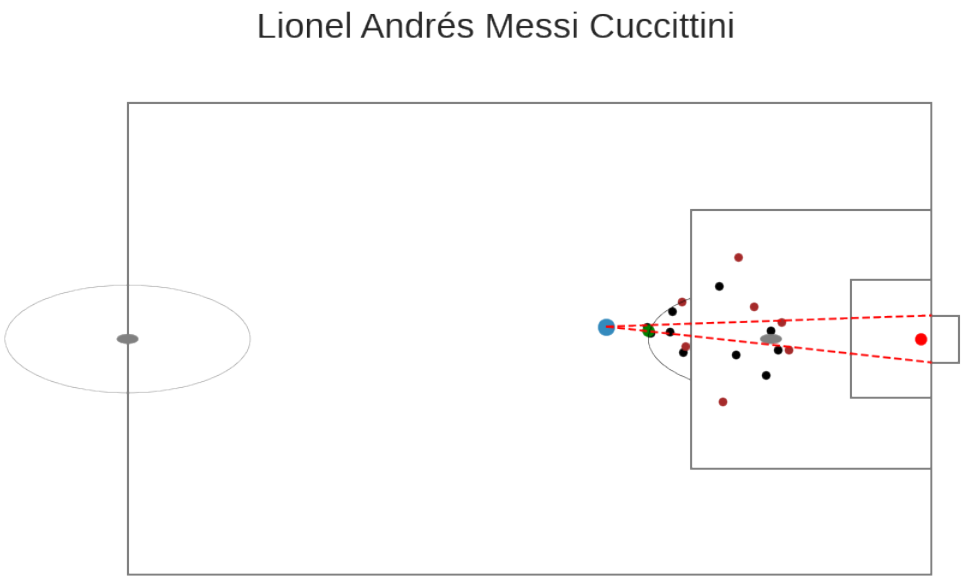


Figure 3.7: Shot Angle

Upon analysis of the Shot Angle's impact, it was determined that it is directly proportional to the xG value of the shot, hence, an important feature. This means that the xG value increases with the increase in Shot Angle as per the box-plot in figure 3.8.

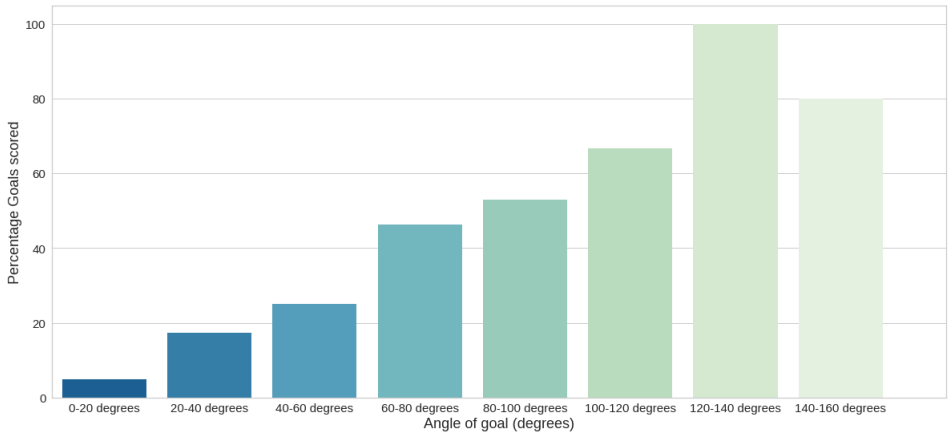


Figure 3.8: Shot Angle Box Plot

3.3.3 Body Part

The Body Part that is used to shoot the ball is the third feature that was analysed as per the road map in figure 3.4. In football, there are two body parts that are usually preferred to shoot the ball, head and feet.

It is quite rare for footballers to use any other body part as the data in figure 3.9. The first impression indicated that headed goals contributed to lesser amount of goals. This was explored further through the data available.

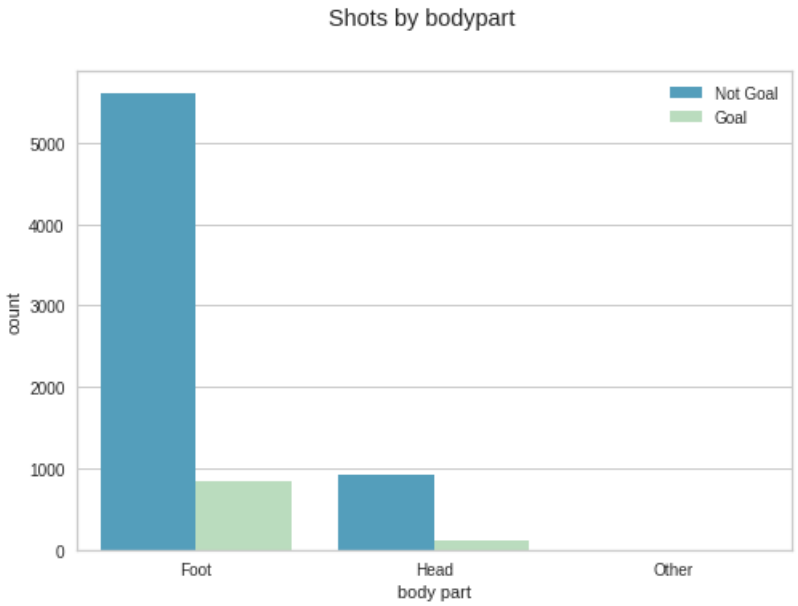


Figure 3.9: Shots by Body Part

Thus the percentage contribution was highlighted as visible in 3.10. The ratio of footed shots to goals was noted at 13.4% while it was at 10.6% for headers. It was concluded that feet contribute to more goals.

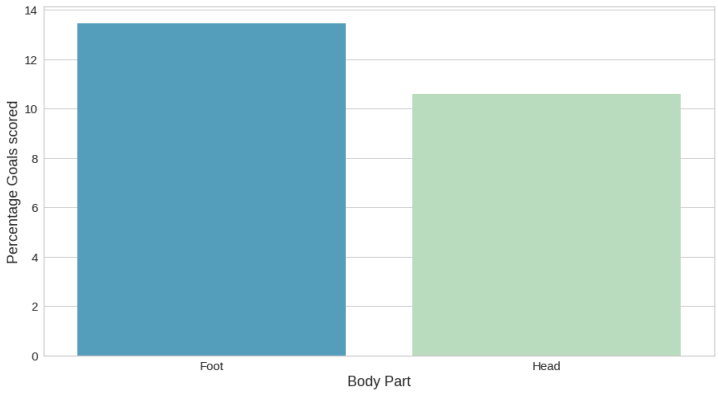


Figure 3.10: Goals by Body Part before Skewness

There was a skewness in this data to be considered which figure 3.10 did not account for. Since all headed shots are attempted from a maximum of 18 yards, it would be wrong to consider footed shots beyond that distance. In order to compare the contribution of shots this skewness, figure 3.11, was catered to.

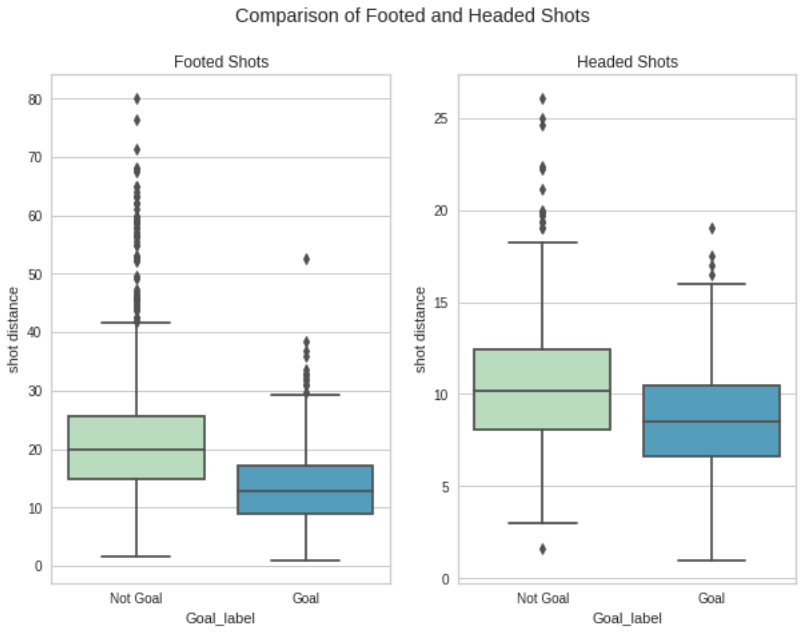


Figure 3.11: Body Part - Shot Distance Box Plot

Sure enough, when all the shots were considered from a maximum of 18 yards from the goal, the results yielded trends similar to that of figure 3.10, yet more clear.

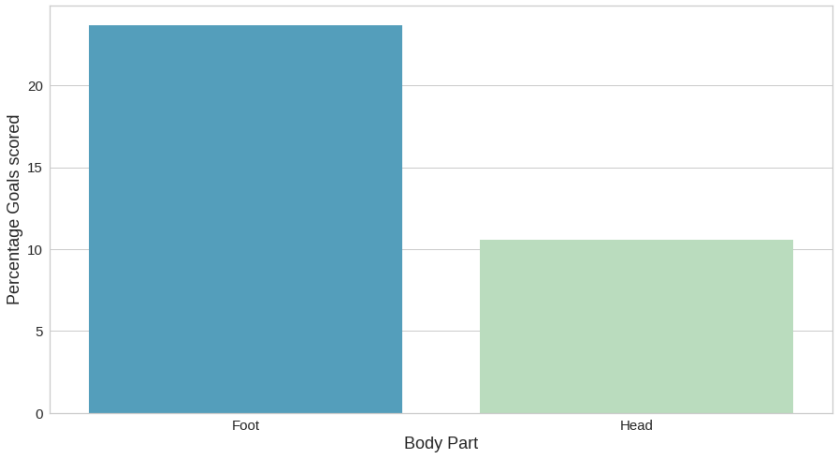


Figure 3.12: Goals by Body Part after Skewness

A whopping 23.7% goal to shot ratio was noted for footed goals while it remained the same, i.e. 10.6% for headers as previously seen in figure 3.10. These findings followed the null hypothesis mentioned in figure 3.4.

3.3.4 Pressure

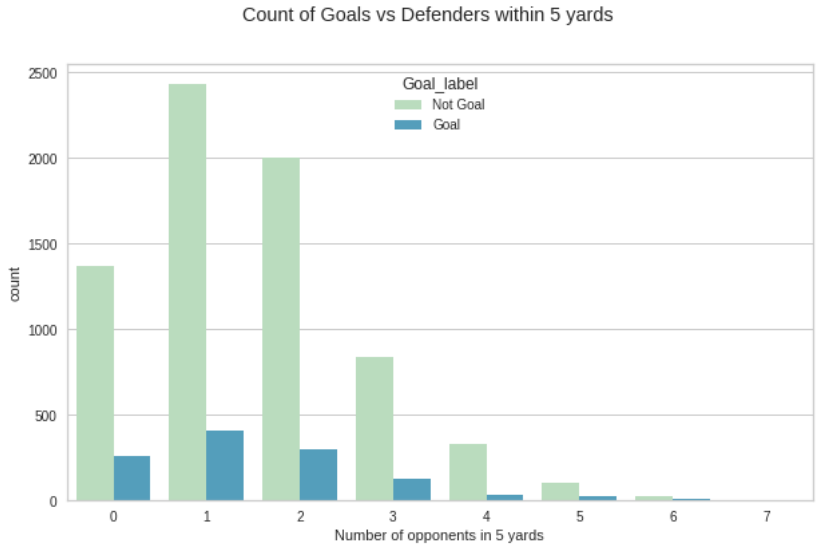


Figure 3.13: Goal Percentage vs Pressure

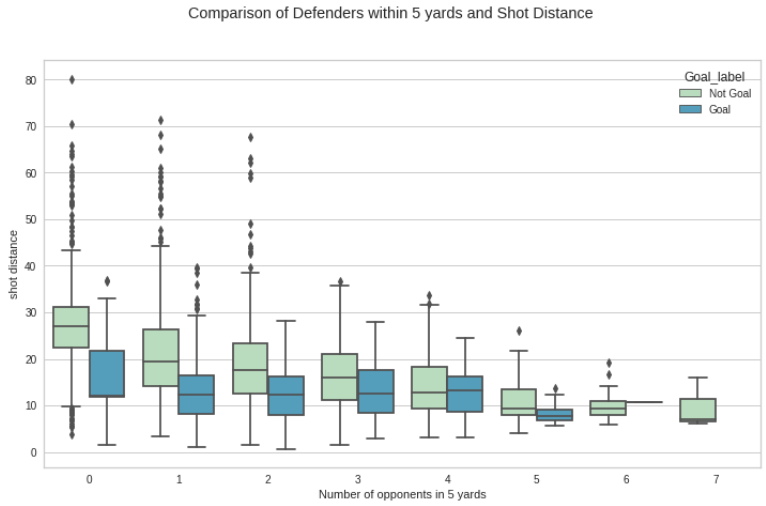


Figure 3.14: Shot Distance vs Pressure

When a player creates a chance in a football match, it is important for him/her to understand if the pressure from the opponent at the time is subpar or not. This was inculcated in the model by calculating the number of opponents within a 5 yard distance from the ball when the chance is being created.

It was clear figure 3.13 that the increase in pressure leads to a decrease in xG approving the null hypothesis in figure 3.4. A further study into the relation of the shot distance and pressure indicated that pressure does increase as the shot distance decreases and is displayed in figure 3.14.

3.3.5 Number of Players in front of Goal

Consider the triangle created when a shooting chance is created by a football player in figure 3.7. The opponents within this triangle have the best opportunity to deflect the shot and consciously try to do so (scenario highlighted in figure 3.17). Hence, it is important to consider the number of players within this triangle at the time of chance creation. A study into this aspect displayed through figure 3.15, highlighted an inverse relation to xG, further approving the null hypothesis and confirming the importance of this feature.

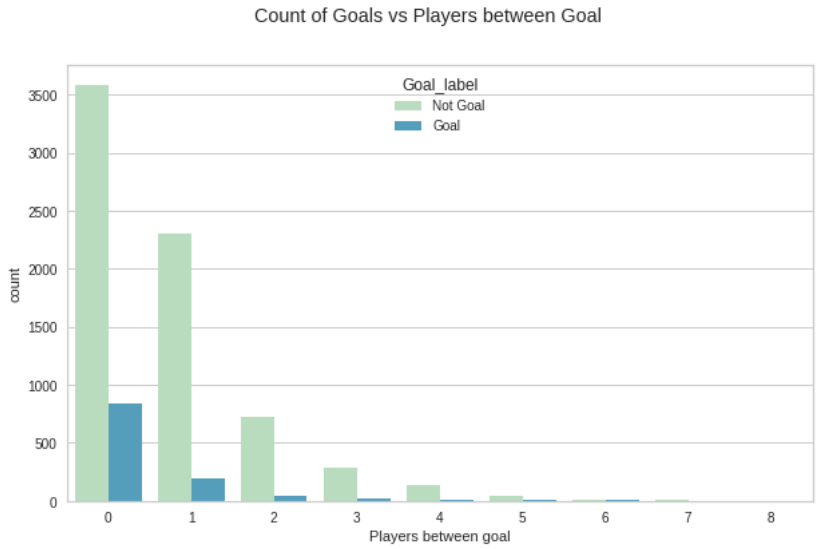


Figure 3.15: Goal Percentage vs Players between Goal

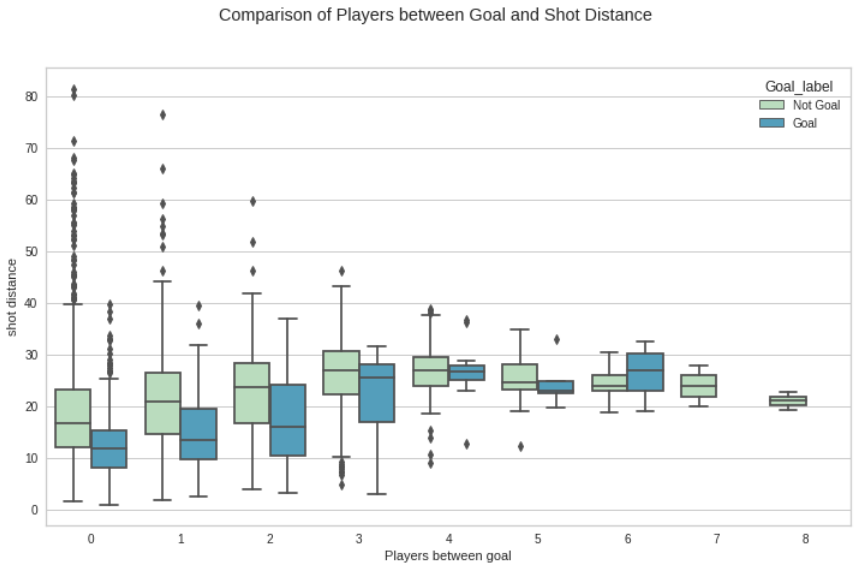


Figure 3.16: Shot Distance vs Players between Goal

A study comparing this feature with shot distance, akin to the one conducted in section 3.3.4, was performed. The resultant visualisation (figure 3.16) showed that the number of players increases with the increase in distance further impacting the xG negatively. It has been highlighted earlier in section 3.3.2, how it was determined that shot distance is inversely proportional to the shot angle. This provides a perspective on how the xG rapidly tends to 0% as multiple factors add up.

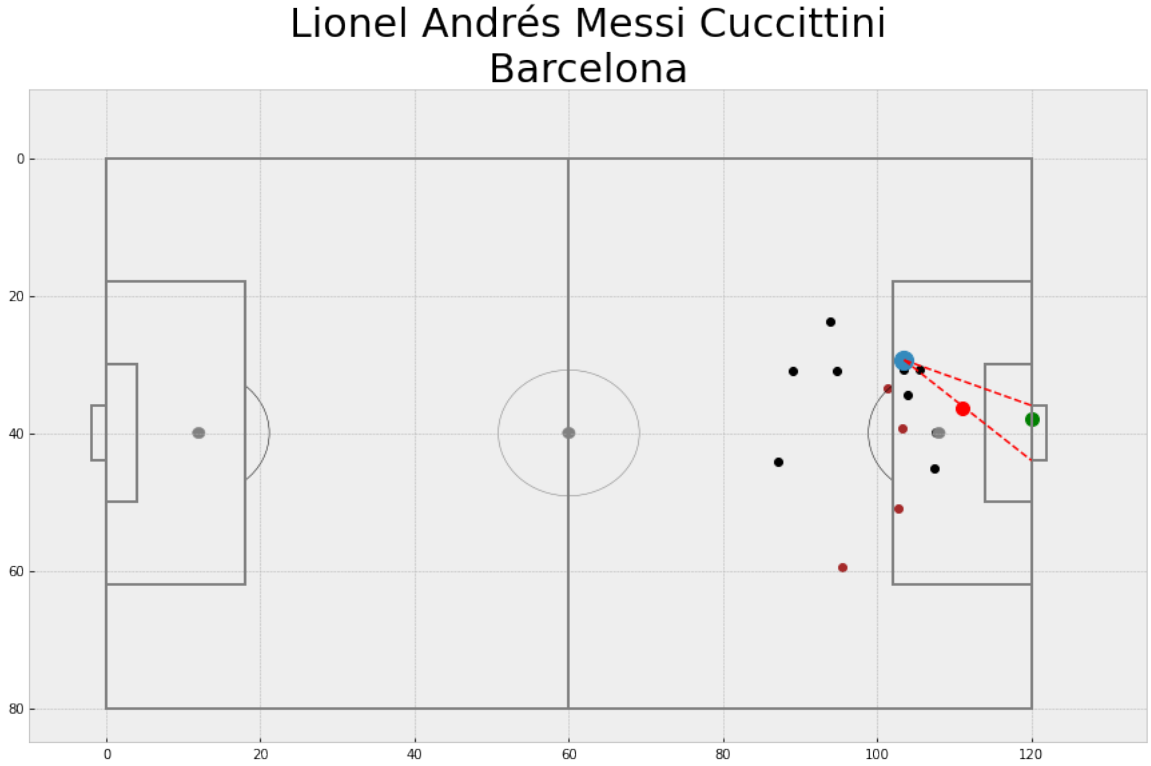


Figure 3.17: Players between Goal

3.3.6 Goal Keeper Distance and Location

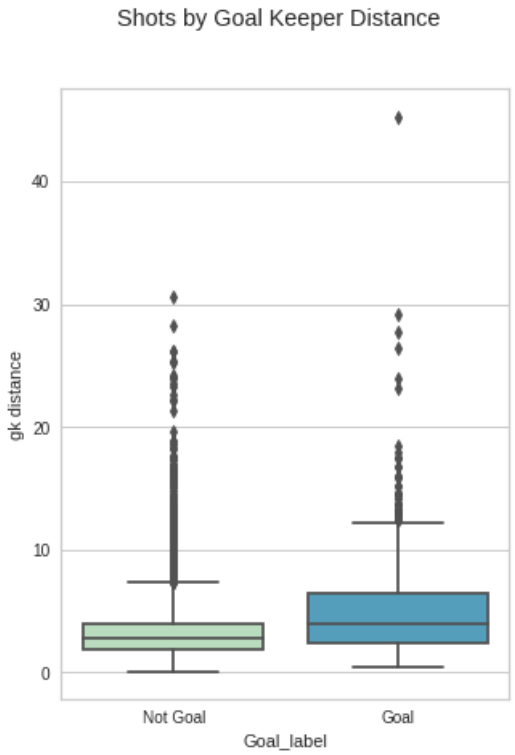


Figure 3.18: Goal Keeper Distance vs Goals Scored

Though a Goal Keeper stays near the goal throughout a football match, there are situations when he/she has to move away from the goal. For example, in a one - on - one situation, goal keepers tend to move closer to the striker to reduce shot angle. Thus, the distance of the goal keeper was analysed to understand the impact it created on chance creation.

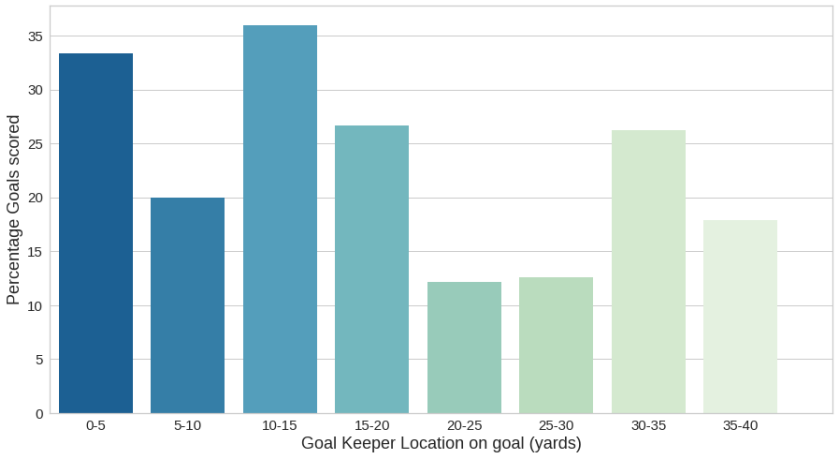


Figure 3.19: Goals Percentage vs Goal Keeper Y coordinate

As proposed by the null hypothesis in figure 3.4, it was noted that the number of goals increases with the increase in distance of the Goal Keeper through the box plot analysis in figure 3.18.

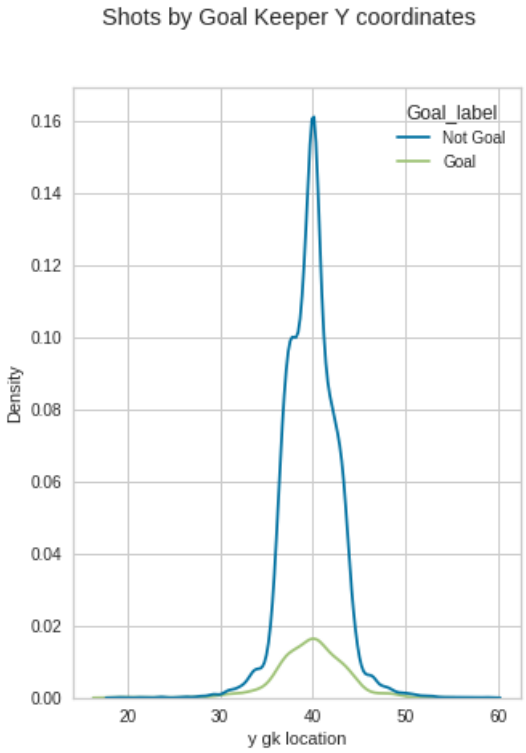


Figure 3.20: Goal Keeper Location vs Football Shots

Another aspect of to be considered was the Y coordinate of the goalkeeper. The Y coordinate of the goalkeeper indicates where the goalkeeper is on the goal line. An analysis into this feature determined that most football shots are attempted when the goalkeeper was at y = 40. This is highlighted in figures 3.20 and 3.19. Thus both of these features were inculcated into the model.

3.3.7 Preceding Play Pattern

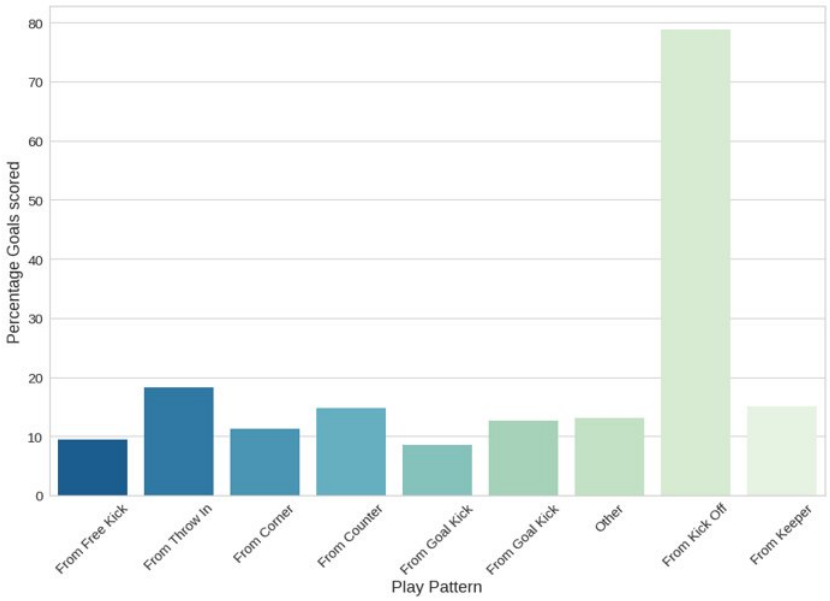


Figure 3.21: Play Pattern vs Goal Percentage

The preceding play pattern in a football game is a crucial factor for pre-shot analysis. The play patterns defined in the data that was made available by [25] are:

- 'From Counter': The match event was involved in a counter attack:
  - The possession began with an open play turnaround outside the counter-attacking team's final third.
  - The possession was at least 75% direct towards goal.
  - The attack travelled a minimum of 18 yards towards goal.
- 'From Throw In': The match event was involved in the passage of play after a throw-in.
- 'Regular Play': The match event was not involved in any of the following play patterns.
- 'From Corner': The match event was involved in the passage of play after a corner.
- 'From Free Kick': The match event was involved in the passage of play after a free-kick.
- 'From Goal Kick': The match event was involved in the passage of play after a goal kick.
- 'From Kick Off': The match event was involved in the passage of play after the kick off.

- 'From Keeper': The match event was involved in the passage of play after a keeper distribution.

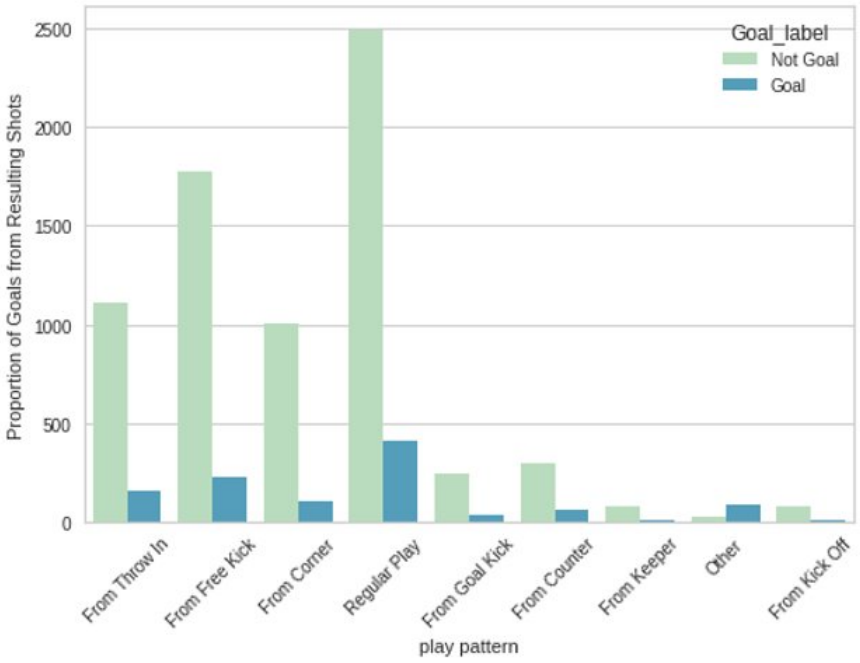


Figure 3.22: Play Pattern vs Football Shots

The play patterns mentioned above serve as a build up to the shot attempt and thus must not be confused as a direct goal from events like 'From Free Kick'. The exploratory data analysis performed to gain a perspective on play patterns provided insight into the importance counter attacks. It is important to note that counter attacks lead to more goals and thus this was added to the model. An initial look at figure 3.21 indicated that a plethora goals result from kick off which is true. This is clarified in figure 3.22.

Most teams try to score goals early on during a football match and build an early lead and this is quite apparent from the number of chances created from kick-off events.

3.3.8 Pass Type and Ball Height

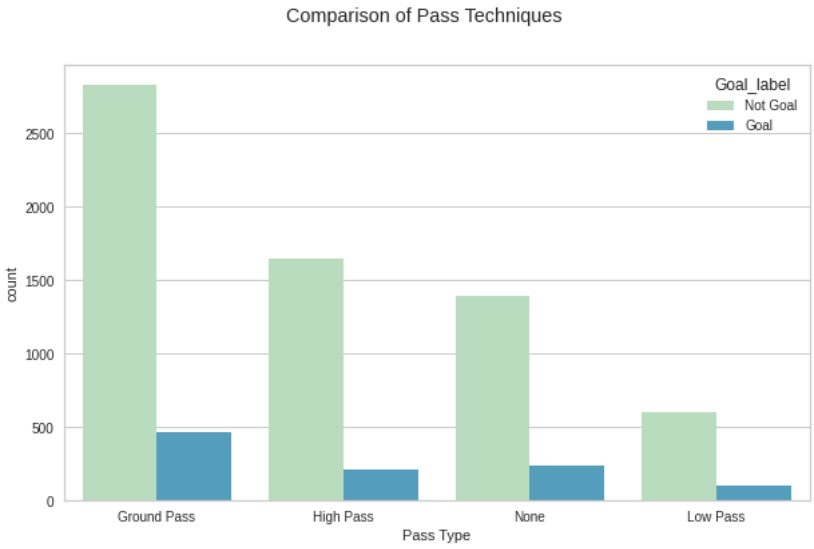


Figure 3.23: Goals vs Pass Types

There are two types of passes in football, namely aerial and ground passes. The data indicated that aerial passes are categorised into high and low passes. To analyse the efficiency of goals that result from the distinct types of key passes to the striker, the null hypothesis specified that low and ground passes resulted in high xG where as high passes did not. Thus, the next step of the exploratory data analysis was performed through figure 3.23.

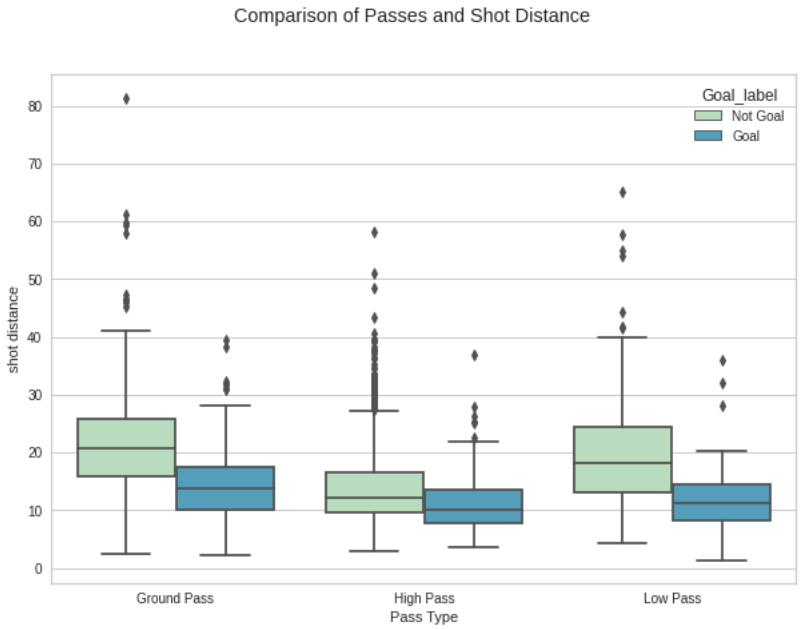


Figure 3.24: Shot Distance vs Pass Types

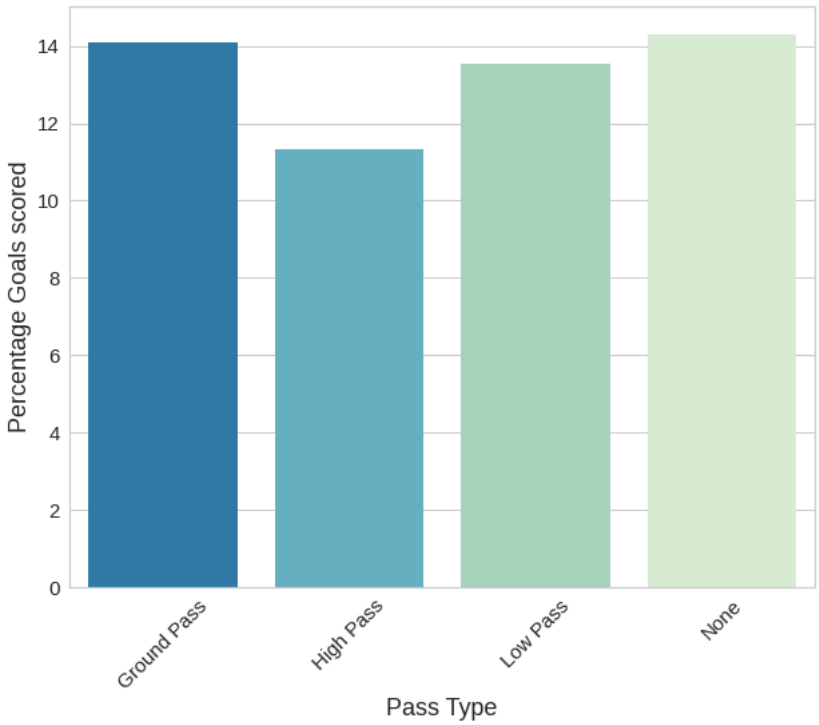


Figure 3.25: Goals vs Pass Types

As per initial observation, it was noted that a plethora of chances are created from ground passes and high passes. The shot distance was analyzed next using box plots. It was noted that high passes led to chances that were created closer to goals where as ground passes and low passes led to chances that were created from a higher shot distance. This could be due to the fact that most high passes lead to headed goals and, as observed in section 3.3.3, headed goals have a lower average shot distance.

The goal contribution from each type was analysed next. The result held true to the null hypothesis that high passes indeed contributed to low xG where as low passes and ground passes contributed to goals with high xG. This could also have to do with the fact that high passes serve as key passes to headed goals.

3.3.9 Direct Goals From Set Pieces

A set piece happens whenever play is restarted due to a foul or the ball being out of play. Typically, these take the shape of corner kicks, indirect free kicks or direct shots from free kicks and penalties. [20]

Two important free kicks to be considered for direct goals are free kicks and penalties. It should be noted that penalties have a universal xG value of 0.79 xG and free kicks are modeled separately. Hence this model focused on open play goals.

3.3.10 First Time Shot

A first touch shot in football is one that is shot the first time. This feature was the last feature to be analysed and an important one. When a ball is shot the first time by a striker, the defence and the goal keeper do not have a lot of time to adapt to it. Keeping this in mind it was hypothesised that first time shots have higher xG while the others had lower xG. The study carried out indeed held true to the null hypothesis.

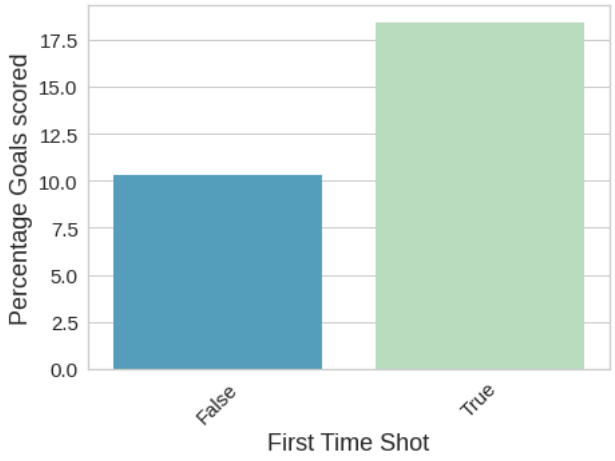


Figure 3.26: Goals vs First Time Shots

With this, the Exploratory Data Analysis, which was the second part of the pipeline was concluded.

3.3.11 Technique Used and Shot Velocity

The technique used by a footballer when they shoot the ball seemed like an important aspect to be considered along with the shot velocity. Since they can determine a lot about the shot quality. Upon research, these two features were dropped from the final list of features. The logic behind this reasoning lies in the very definition of the model. Expected Goals is a metric that analysis chance creation and is a tool for **Pre-Shot Analysis**. Thus, the technique used and the shot velocity are measures that do not serve the purpose of this model. The use of these features will be highlighted in section 6.

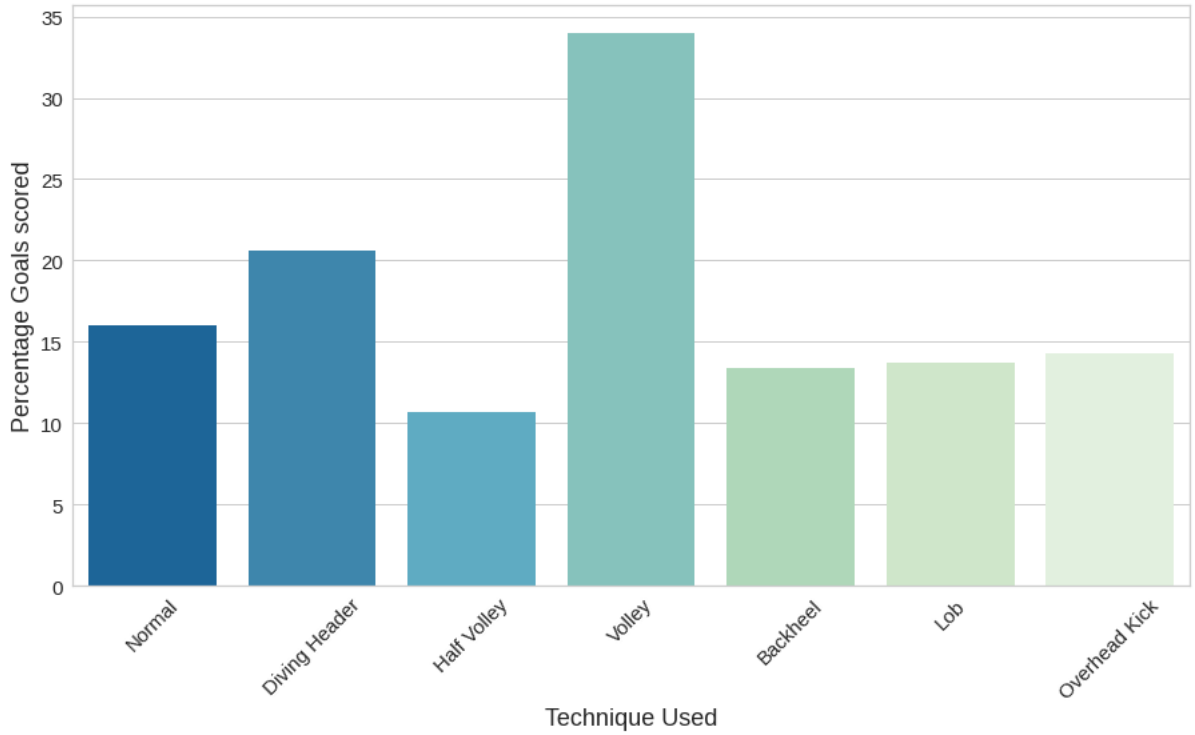


Figure 3.27: Shot Technique

Figure 3.27 shows that the most contributing technique to goals was the volley technique. Though that might be true, it was also noted that volleys are extremely hard to execute and the sheer frequency of volleys is less in football. Hence, due to a lack of data and the aforementioned reasons these features were not considered.

3.4 Feature Engineering

Feature Engineering is the process of building new features from the existing data in order to further enhance predictive modeling. The features mentioned below were created after an exhaustive research into Expected Goals and finalized after carrying out multiple tests on distinct machine learning models.

The Exploratory Data Analysis carried out in section 3.3 covers how each of these features cause an impact to xG.

These features were engineered from the existing features:

- Number of opponents in 5 yards: Number of players around the shot taker within a 5 yard distance
- Players between goal: Number of Players between the triangle formed by the Ball and the two goal posts
- Shot Distance: Distance of the ball at the time of Shot
- Goalkeeper Distance: Distance of the Goalkeeper from the ball
- Shot Angle: Angle formed between the ball and the goal posts
- Goal Label: Goal or No Goal
- Binary Outcome: Based on Goal Label, Goal being 1 and No Goal being 0

Each of these features can be observed in figure 3.3. In order to engineer the above mentioned features, a few mathematical concepts were incorporated in the Data Wrangling procedure.

3.4.1 Mathematical Concepts Used

As highlighted in [18], the distance formula is one of the most basic yet powerful theorems which yields the distance between any two given points.

$$d = \sqrt{x^2 + y^2} \tag{3.1}$$

Equation 3.1 can also be written as,

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.2}$$

Where x_2 and x_1 are the X Coordinates of the two points and y_2 and y_1 are the Y Coordinates of the two points respectively.

Furthermore, in order to calculate the Shot Angle, the Law of Cosines [11] was further incorporated after the distances were calculated.

$$a^2 + b^2 + 2ab\cos(c) = c^2 \tag{3.3}$$

where a and b are the distances of the ball from the different goal posts, while c is the distance between both the goal posts which is a constant (8 yards). An example can be seen in 3.3, where the red lines each highlight a and b respectively.

The following variant of the equation 3.3 was utilized in order to calculate the angle between the goal posts.

$$c = \arccos\left((c^2 - a^2 - b^2)/2ab\right) \tag{3.4}$$

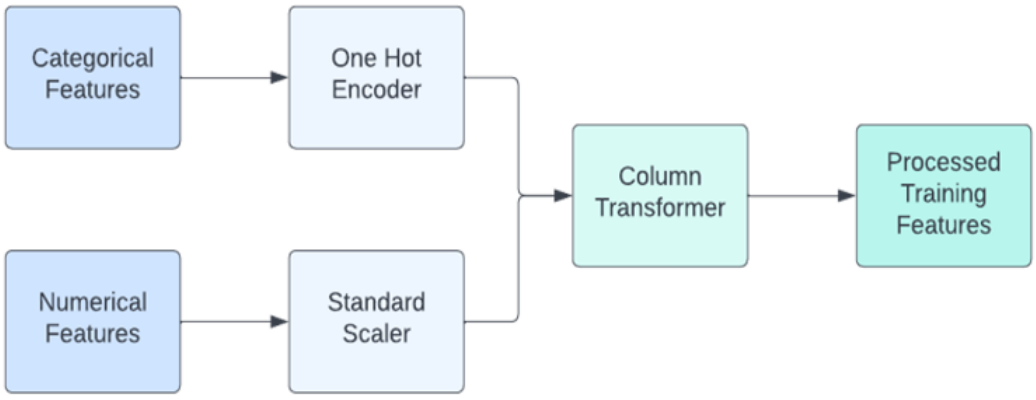


Figure 3.28: Feature Processing

A total of 18 features were finalized. These 18 features were a mix of numerical and categorical features. Since there were two distinct types of features, it was crucial to

transform them in a suitable format for the purpose of Machine Learning. This task was performed using One Hot Encoder, Standard Scaler and Column Transformer obtained from the Sci-Kit Learn library [19].

One Hot Encoding is implemented on the categorical features, for example, Body Part, while Standard Scaling is implemented on the numerical features for example, Shot Distance. These encoded features were then passed onto the Column Transformer which created the final processed features.

The pipeline in figure 3.28 highlights the process of feature transforming for model training.

3.5 Model Testing and Hyper-Parameter Tuning

The process of discovering the optimal blend of hyper - parameters that enhances model performance is known as hyper - parameter tuning. It works by combining several trials into a single training session. Each trial is a full execution of your training application with values for the selected hyper - parameters set within the defined bounds. Once completed, this method will output the collection of hyper - parameter settings that are most appropriate for the model to produce optimal outcomes.[1] The algorithms chosen for Machine Learning to predict the xG values are listed as follows:

- Logistic Regression
- XGBoost
- CatBoost
- LightGBM

3.5.1 Logistic Regression

Logistic Regression is frequently used for categorization and predictive analytics. Based on a collection of independent variables, logistic regression calculates the probability of an event, such as goal or not goal. As the result is probabilistic in nature, the dependent variable lies in the range of 0 to 1. A logit transformation is performed to the odds in logistic regression, which is the likelihood of success divided by the likelihood of failure. [10]

3.5.2 LightGBM

LightGBM is a toolkit for gradient boosting that employs tree-based learning techniques and is developed by Microsoft.[14] It is intended to be widely dispersed and effective, with the following benefits:

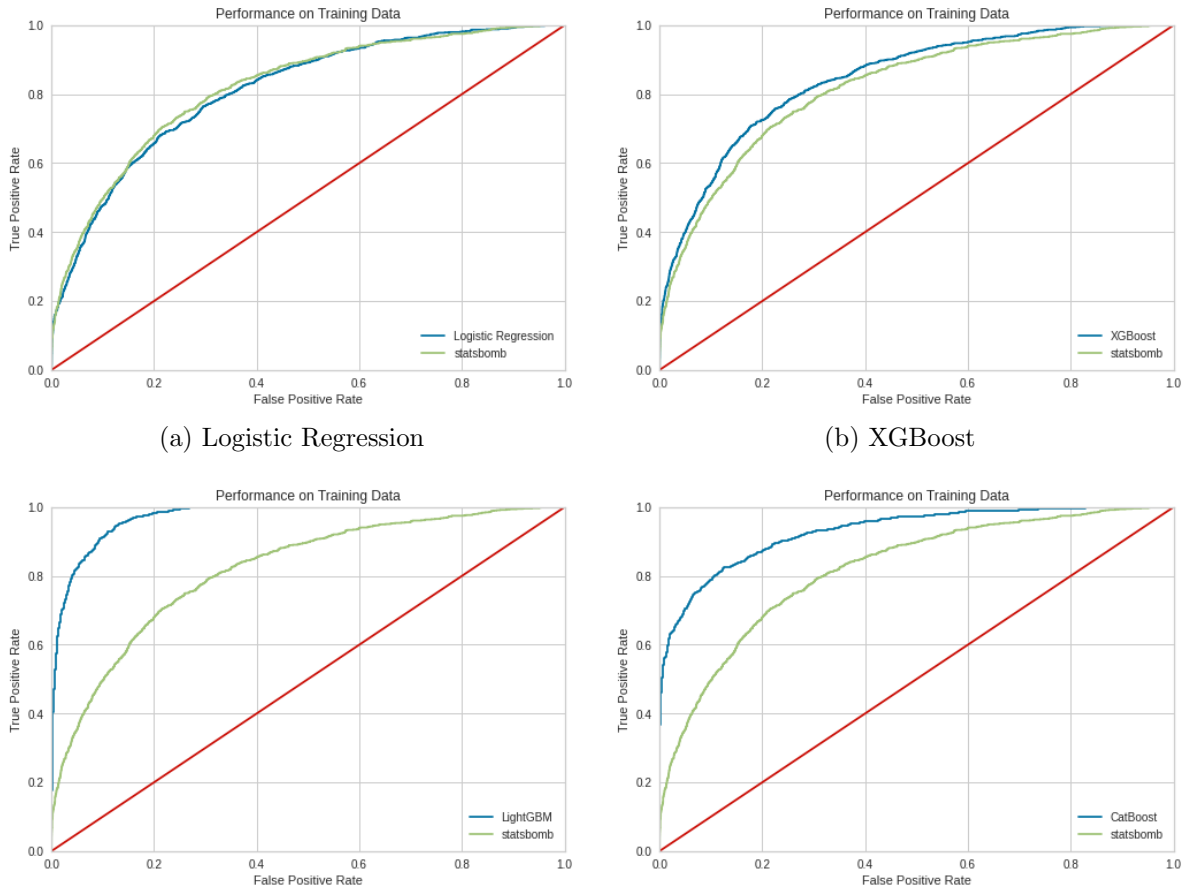
- Increased training pace and efficiency.
- Parallel, distributed, and GPU learning are all supported.
- Capable of managing enormous amounts of data.

3.5.3 XGBoost

XGBoost is a distributed gradient boosting toolkit that has been developed to be very effective, adaptable, and accessible. It uses the Gradient Boosting architecture to construct machine learning algorithms. XGBoost offers parallel tree boosting to address a wide range of data science challenges promptly and reliably. The same code may tackle problems with billions of instances in major distributed environments (Hadoop, SGE, MPI).[6]

3.5.4 CatBoost

Yandex's CatBoost is a newly developed algorithm for machine learning. It integrates easily with deep learning systems such as Google's TensorFlow and Apple's Core ML. It can operate with a variety of data formats to assist organisations tackle a wide range of challenges. It also has the highest accuracy in its class.[8]



(a) Logistic Regression

(b) XGBoost

(c) LightGBM

(d) CatBoost

Figure 3.29: Model Performance before Tuning

The performance of these models were evaluated on the basis of AUC score and Log Loss score.

Each of these metrics is elucidated briefly below:

- The AUC - ROC curve is an effectiveness metric for classification issues at various threshold levels. AUC is the magnitude or degree of separability, whereas ROC is a probability curve. It indicates how well the model can differentiate between classes. The greater the AUC score, the more effectively the model predicts distinguishes between 0 and 1 (No Goal and Goal respectively).
- The log-loss number indicates how near the predicted probability is to the orig-

inal value (0 or 1 for binary classification). The greater the difference between the expected and actual probability, the greater the log-loss value. The values 0 and 1 correspond to No Goal and Goal respectively. Thus, in conclusion a good model has low log loss scores. A log loss score below 0.5 indicates good performance.

In order to understand the raw model behaviour, each model was fit on the finalized features.

Logistic Regression performed similar to that of the Statsbomb model when their AUC curves were compared, while the XGBoost, LightGBM and CatBoost models were overfitting. Thus, hyper - parameters were chosen to cater to this overfitting. These hyper-parameters are listed below:

- Max Depth: This is used to control the depth of the tree thus helps to avoid the algorithm from learning relations specific to each data point
- N Estimators: This is used to control the number of trees within the algorithm
- Learning Rate/ETA: Learning Rate and ETA are analogous terms used for different models. They make the learning process better for the model by avoiding overfitting through a step size reduction. The feature weights shift after every step, thus the smaller the step the more robust the model.

Once the hyper-parameters were identified, RandomSearchCV (available in the Sci-Kit Learn library) developed by Pedregosa et al.[19] was used to find the optimum values for each hyper-parameter.

Results

4.1 Model Validation

Model Validation is the process of testing the model on a small amount of data which the model has not been trained on. This validates the hyper-parameters along with the performance of the model.

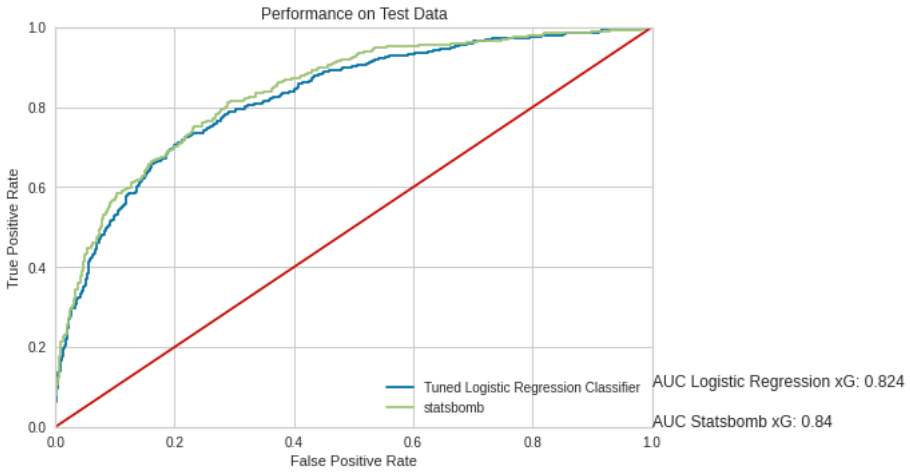


Figure 4.1: Logistic Regression Model after Tuning

The AUC-ROC curve was plotted after hyper-parameter tuning to understand how each model was impacted. As visible in figures 4.1, 4.3, 4.4 and 4.5 the AUC scores of each model were compared against the benchmarked score of 0.84 of the Statsbomb model.

- Logistic Regression: 0.824
- XGBoost: 0.801
- LightGBM: 0.816
- CatBoost: 0.823

In comparison to the Statsbomb model, the AUC scores (see section 3.5) look good considering the size of the dataset they were trained on. The Log Loss score obtained for each model was in the vicinity of 0.3 which was a good sign. The validated models were passed on to Machine Learning for predictive analysis.

y=1 top features

Weight[2]	Feature
+0.273	Pass Type_Ground Pass
+0.072	play pattern_Non Regular Play
-0.034	play pattern_Regular Play
-0.037	first time_True
-0.052	Number of opponents in 5 yards
-0.082	shot distance
-0.100	Pass Type_High Pass
-0.148	y location shot
-0.391	first time_False
-0.393	Pass Type_Low Pass
-0.394	Players between goal
-0.405	x location shot
-0.440	x gk location
-0.443	body part_Foot
-0.556	y gk location
-0.559	gk distance
-0.617	Pass Type_None
-0.916	shot_angle
-0.999	<BIAS>
-1.530	body part_Head

(a) Logistic Regression

Weight	Feature
0.1921	Pass Type_Low Pass
0.1254	shot distance
0.1060	Pass Type_Ground Pass
0.0677	body part_Head
0.0569	first time_False
0.0543	Players between goal
0.0462	play pattern_Regular Play
0.0455	y gk location
0.0447	y location shot
0.0414	Number of opponents in 5 yards
0.0399	Pass Type_None
0.0382	play pattern_Non Regular Play
0.0377	first time_True
0.0371	gk distance
0.0356	Pass Type_High Pass
0.0313	x location shot
0	x gk location
0	shot_angle
0	body part_Foot

(b) XGBoost

Weight	Feature
0.4746	Pass Type_Ground Pass
0.1962	play pattern_Regular Play
0.1238	Pass Type_Low Pass
0.0729	body part_Head
0.0487	shot distance
0.0248	first time_False
0.0157	first time_True
0.0111	Pass Type_None
0.0095	y gk location
0.0058	Pass Type_High Pass
0.0053	Players between goal
0.0043	play pattern_Non Regular Play
0.0032	Number of opponents in 5 yards
0.0029	gk distance
0.0010	x location shot
0	x gk location
0	y location shot
0	shot_angle
0	body part_Foot

(c) LightGBM

Weight	Feature
0.2619	Pass Type_Ground Pass
0.1220	body part_Head
0.1138	Pass Type_Low Pass
0.1100	first time_False
0.0859	play pattern_Regular Play
0.0720	Pass Type_None
0.0432	play pattern_Non Regular Play
0.0312	shot_angle
0.0308	Pass Type_High Pass
0.0215	shot distance
0.0198	Number of opponents in 5 yards
0.0159	y location shot
0.0156	y gk location
0.0139	first time_True
0.0136	Players between goal
0.0135	body part_Foot
0.0116	gk distance
0.0030	x gk location
0.0007	x location shot

(d) CatBoost

Figure 4.2: Model Weights

As observed in figure 4.2, it is clear that each model assigned weights to features in a trend similar to the other albeit the CatBoost and LightGBM models failed to give much importance to numerical features.

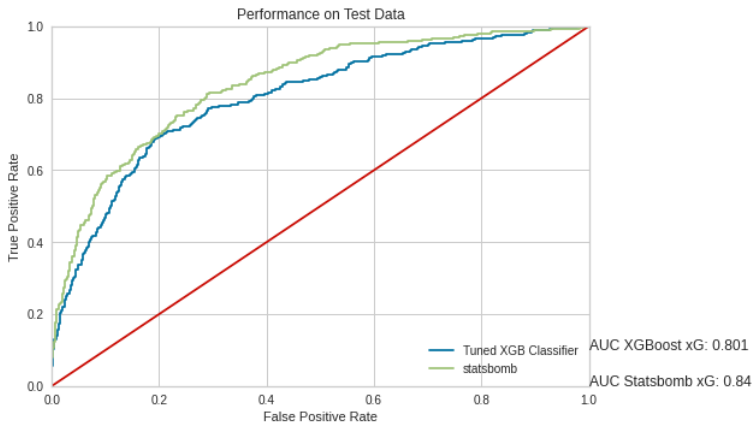


Figure 4.3: XGBoost Model after Tuning

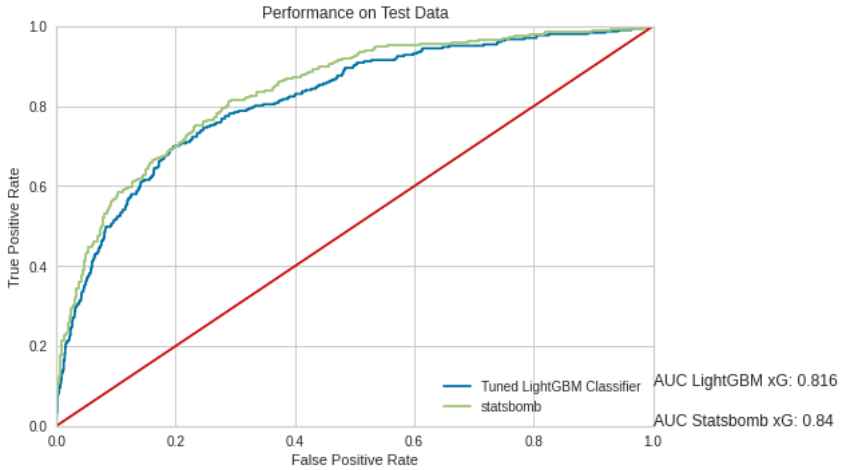


Figure 4.4: LightGBM Model after Tuning

Upon further analysis and visualisation it was observed that the XGBoost and Logistic Regression models recognised numerical features far more efficiently. A better visualisation (see figure 4.6) was created to gauge and compare the feature importances of each model.

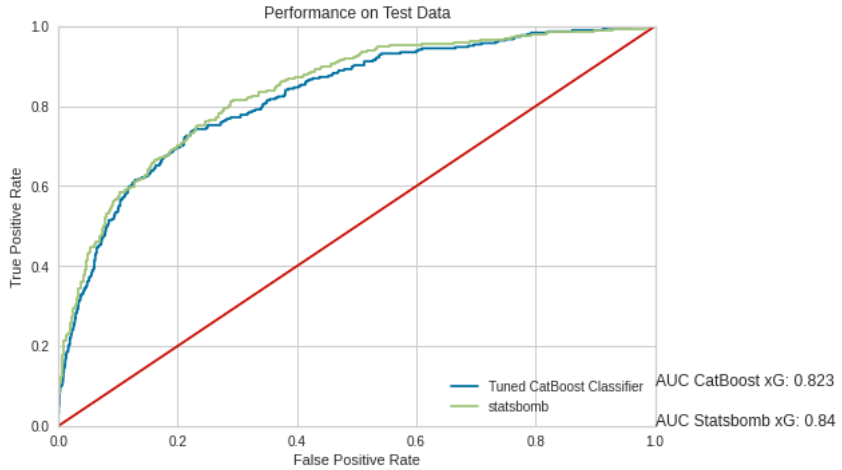


Figure 4.5: CatBoost Model after Tuning

4.1.1 Logistic Regression Feature Importance

The feature importance chart in figure 4.6 provided critical insight into how each algorithm made decisions. The key numerical features for Logistic Regression were

- Shot Angle
- Y GK Location
- GK Distance
- Players between goal

The categorical features that most impacted this algorithm's decision making were

- Ground Pass
- Low Pass
- No Pass

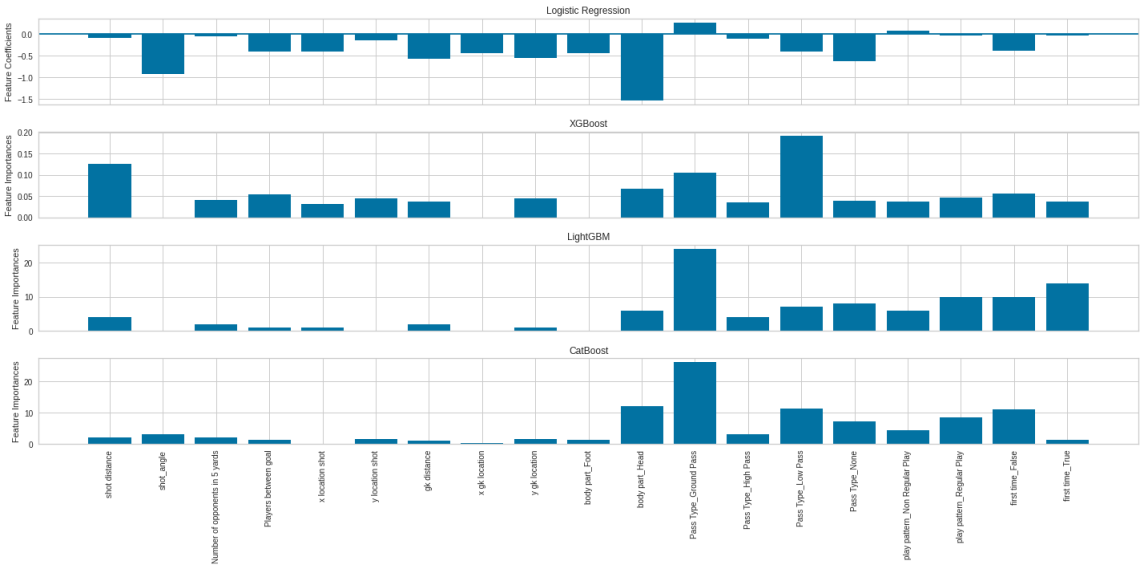


Figure 4.6: Final Model Weights

Surprisingly, numerical features like shot distance, density and number of opponents within 5 yards were not prioritised by this algorithm. This might be due to the hindrance caused by other features. Thus, leaving space for more feature engineering.

4.1.2 XGBoost Feature Importance

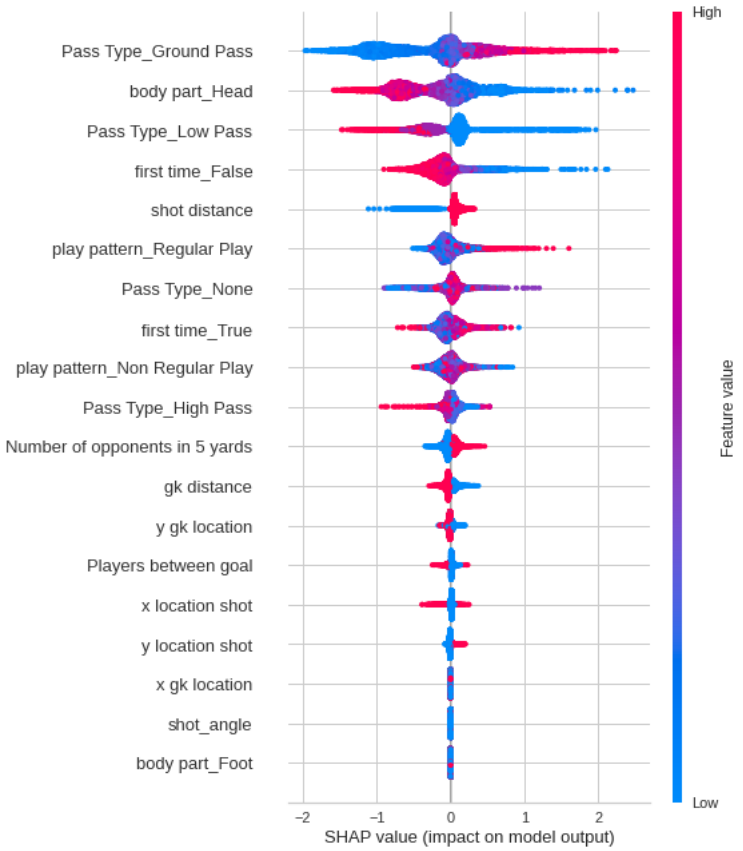


Figure 4.7: Final XGBoost Weights

The feature importance of the XGBoost model varied to that of Logistic Regression highlighted in figure 4.7. The numerical features that this model prioritised were

- Shot Distance
- Players Between Goal
- Y GK location
- Number of Opponents within 5 Yards

The categorical feature trend for this model displayed the following key features

- Low Pass
- Ground Pass
- First Time - False

4.1.3 LightGBM Feature Importance

LightGBM being a categorical boosting model did not assign high importance to numerical features highlighted in figure 4.8. That said, the key numerical features highlighted by this model were

- Shot Distance
- Number of Opponents within 5 Yards
- GK Distance

The categorical features this model prioritised were

- Ground Pass
- First Time - True
- First Time - False
- Play Pattern - Regular

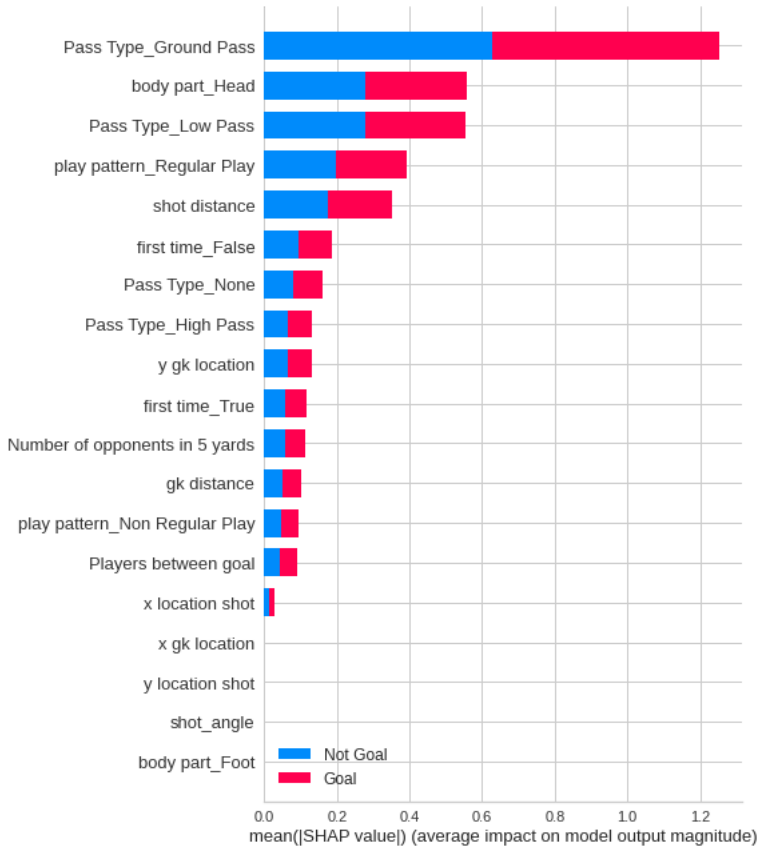


Figure 4.8: Final LightGBM Weights

4.1.4 CatBoost Feature Importance

The CatBoost model performed similar to the LightGBM model for numerical features. Though it's weights differed from LightGBM, it prioritised categorical features just like the latter highlighted in figure 4.9. The numerical features that were prioritised were

- Shot Angle
- Number of Opponents within 5 Yards
- Shot Distance

The categorical features deemed important by this model after training were

- Ground Pass
- First Time - False
- Low Pass
- Play Pattern - Regular

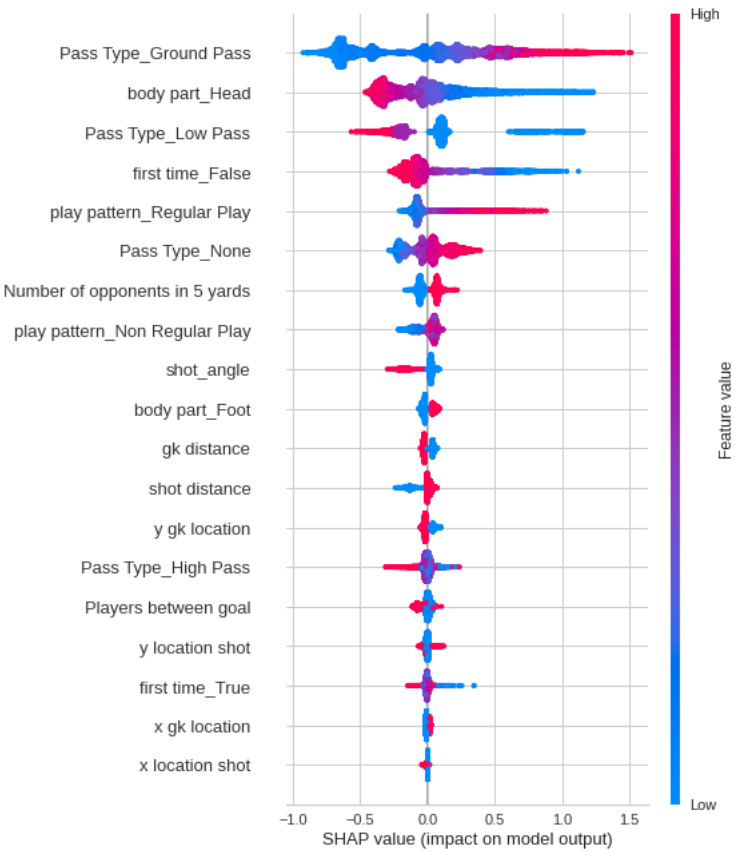


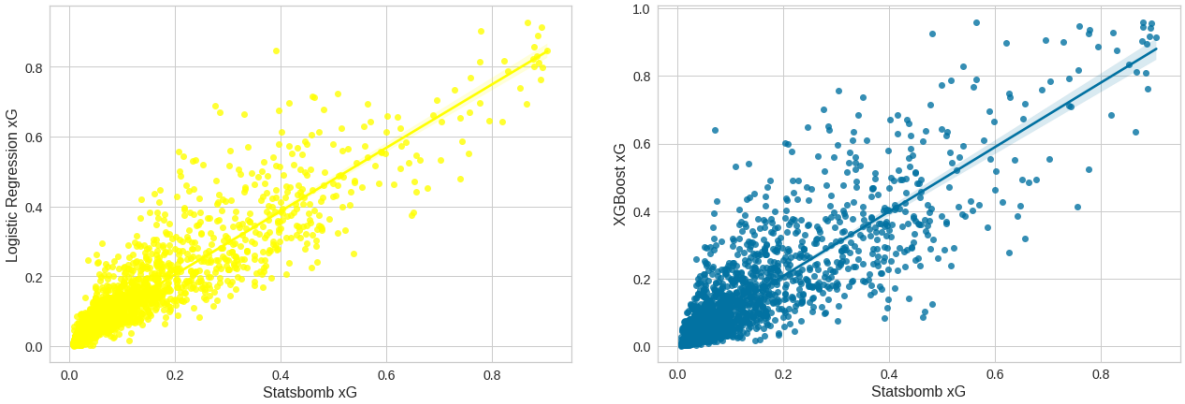
Figure 4.9: Final CatBoost Weights

Quite notably, the models either prioritised the shot distance or shot angle. None of the models use both of these features together for the decision making.

4.2 Model Outcomes

The predictions made by each model were further visualised alongside the Statsbomb through scatter plots to understand the deviation between the two. This can be seen in figures 4.10 and 4.11.

The predictions made by XGBoost were well spread as figures 4.10 and 4.11, report while the other models had a propensity to be less well spread albeit all the models shared high correlation with each other and the StatsBomb predictions.

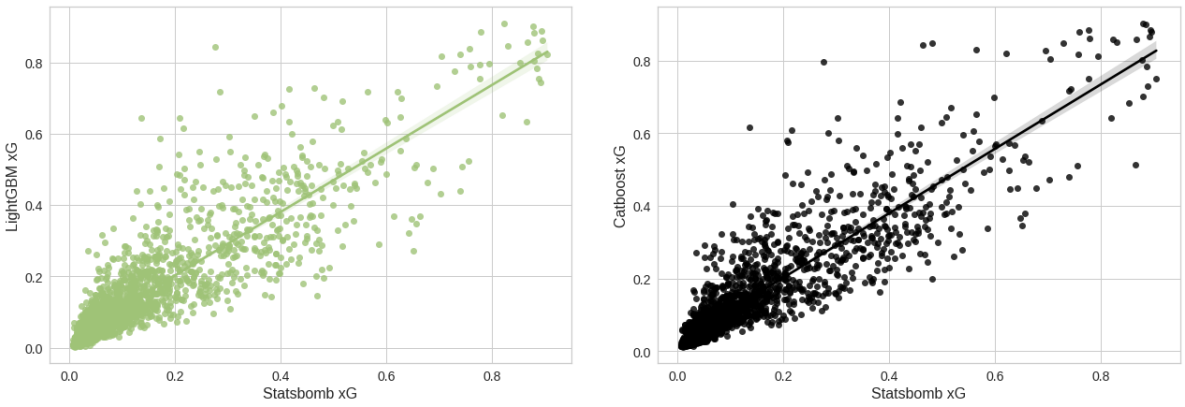


(a) Logistic Regression (b) XGBoost

Figure 4.10: Scatter Plots for Logistic Regression and XGBoost

The scatter plots of LightGBM and CatBoost shows how the models had similar prediction trends due to the similar nature of the categorical boosting models.

With this in mind, a Heat Map of xG values generated by the four models was visualized to identify the correlation with the StatsBomb model. This map is made available by figure 4.12.



(a) LightGBM (b) CatBoost

Figure 4.11: Scatter Plots for CatBoost and LightGBM

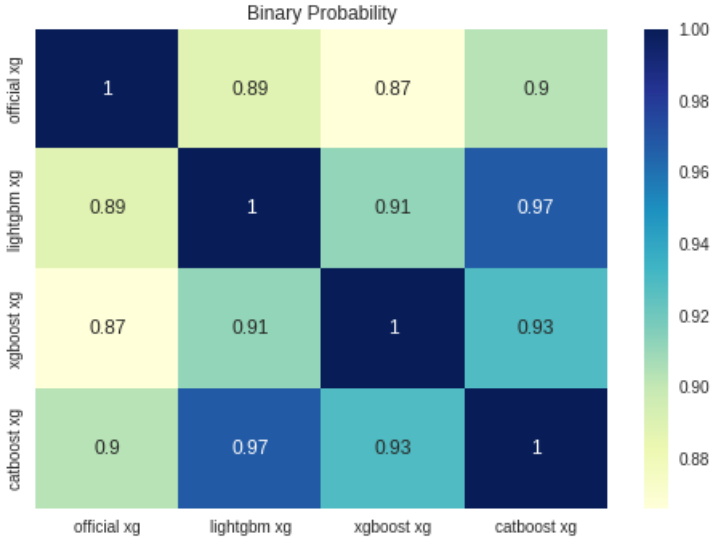


Figure 4.12: xG Values Heat Map

It was noted that the XGBoost algorithm had the least correlation with the StatsBomb xG even though it recognized the features better than other models. Conversely, CatBoost shaered the most correlation with that of StatsBomb xG values.

4.2.1 Real-Time Outcomes

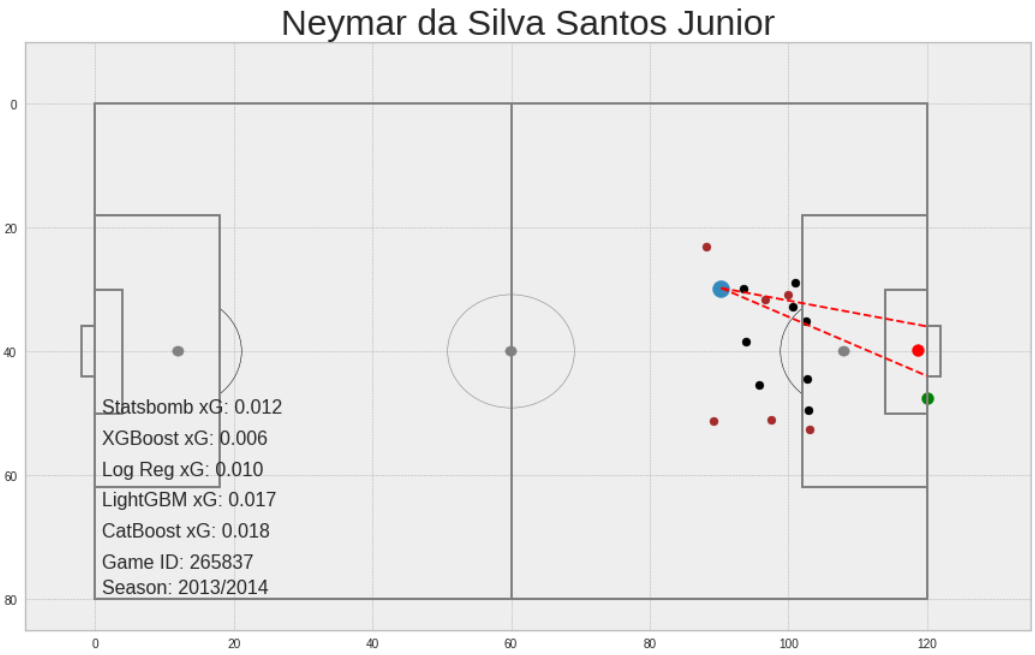


Figure 4.13: xG Value Neymar da Silva Santos Junior

Each model was deployed for pre-shot analysis for distinct football matches played by F.C. Barcelona and a few other teams available in the dataset. Different scenarios were highlighted and xG values were produced for them respectively by each model.

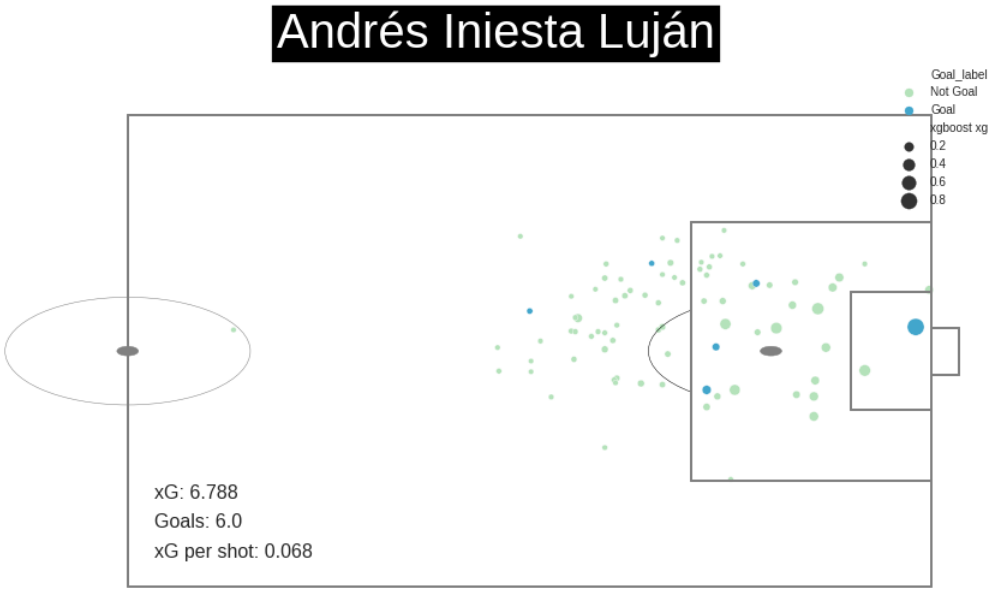


Figure 4.14: 100 shots xG - Andres Iniesta

Figure 4.13 shows the distinct values generated by the models for a chance created by Neymar da Silva Santos Junior, aka Neymar for F.C. Barcelona during a particular match in the 2013/14 season of the Spanish League. Furthermore, figure 4.14 displays the cumulative xG, valued at 6.788 for chances created by Andres Iniesta, indicating that for every 100 chances he creates he will score approximately 6-7 goals. This is a sample utilisation of a player's seasonal performance. This figure was created to signify the kind of chances each player creates on a normal basis.

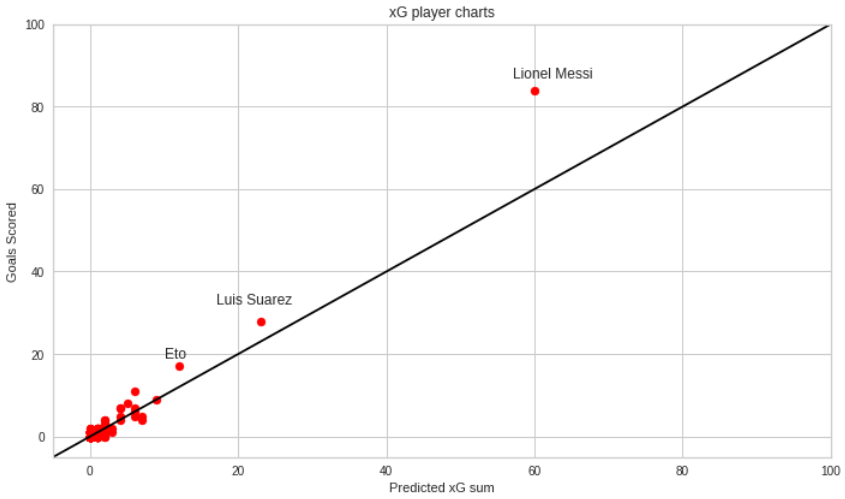


Figure 4.15: Goal Scorers for F.C. Barcelona

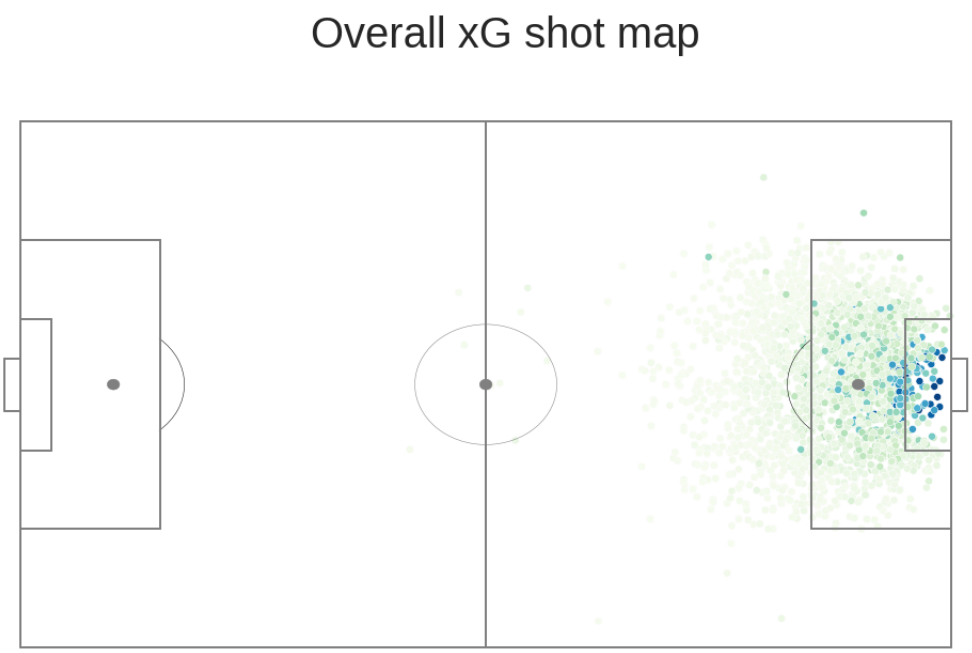
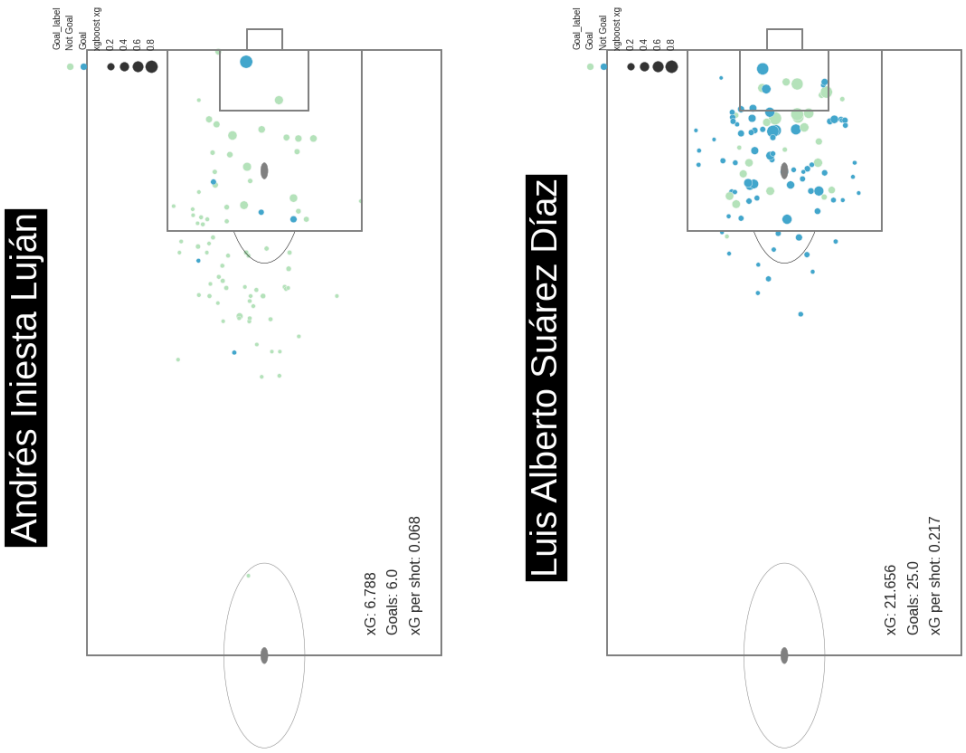


Figure 4.16: xG for F.C. Barcelona

This was taken a step further. All the goal scorers were compared and the quality

of their goals were visualised in the graph in figure 4.15, giving clarity about players who create chances higher in quality than others. The top three goal scorers for the team clearly outperformed their xG by miles. There are a few players who score the same number of goals as others but create chances that are poor in quality.

The chances created by F.C. Barcelona were visualised to highlight the goal scoring xG as a result of the chances created by them. The resultant figure 4.16 displays all the chances created by the team in test dataset and the corresponding xG values in terms of color. The shots that correspond to a higher shade of blue have a higher xG when compared to the others.



(a) Iniesta (b) Suarez

Figure 4.17: Player Comparison

Player comparison was performed by plotting Andres Iniesta and Luis Suarez's goal profiles available in figure 4.17. After analysing one could say that Iniesta is not a good goal scorer since the number of goals he scored after a 100 shots was a mere 6. Yet, it would be wrong to say that Iniesta is a poor player altogether. This was because the average xG of the shots he scored was noted to be around 6.8 xG. According to the average, for every 100 shots Iniesta scores between 6-7 goals. It could be safely stated after this result that Iniesta is performing consistently if not over-performing. When compared to Luis Suarez, who as a matter of fact is a prolific goal scorer, Iniesta seemed to be under-performing but clearly was not. For a player that dominated the midfield it clarified several misconceptions. Furthermore, if Suarez's profile was to be

considered, it justified why he scored a lot of goals every season. This was because of the amount of chances he created with a high xG.

Similarly, teams were also compared. A visualisation of the chances created by each team in the dataset were made available through figure 4.18. Due to lack of availability it was difficult to analyse and compare teams.

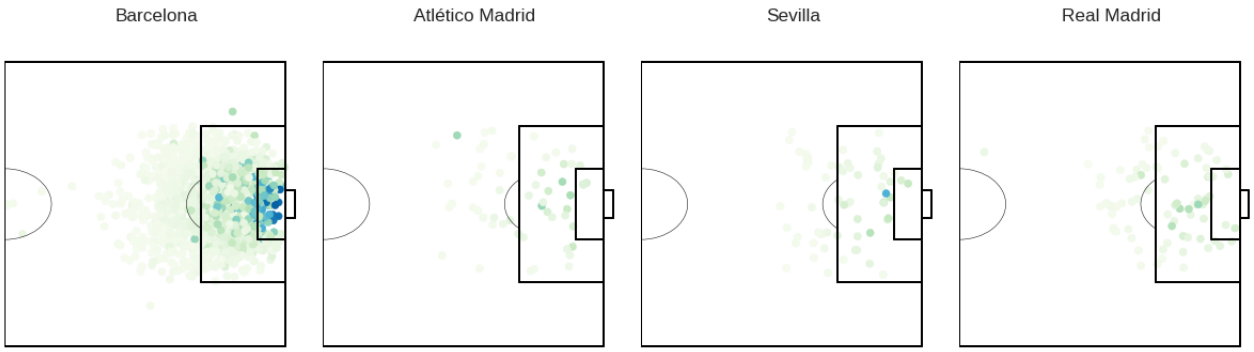


Figure 4.18: Team Comparison

Conclusion

Through this study four distinct Expected Goals (xG) models were generated to understand the following:

- Introduce probability based AI in football
- Chance Creation Quality
- Understanding Player Profiling and Match Strategy
- Establish a bedrock for future metrics

This study caters to certain features that a few other studies conducted before (highlighted in section 2) do not. The features that this study inculcates include, goalkeeper positioning and the players blocking the goal which according to section 2 the model in [21] previously did not.

Furthermore, after the feature weights were analysed carefully, the following factors were concluded from this study as far as the most optimal algorithm for an Expected Goals model is concerned.

- Logistic Regression: This algorithm understood numerical weights better than two out of the three boosting algorithms utilized in this study. This model did fail to allot the appropriate weights to categorical variables in accordance to the feature engineering and exploratory data analysis. It could be considered as a prospect for an xG model, though it needs to highlighted that XGBoost algorithm can perform better.
- XGBoost: This algorithm performed well and followed the trends highlighted in the Exploratory Data Analysis well for both, categorical and numerical features. The only feature that it did not consider well was Shot Angle. This can be catered to with thorough feature engineering. As far as this study is concerned, XGBoost algorithm was the best performing model.
- LightGBM: Though this model predicted appropriate xG values, its the feature weights were wrongly allotted. Thus utilizing this algorithm for an xG model in real-time would not be fruitful.
- CatBoost: Much like the LightGBM model, this model prioritised categorical features with similarly assigned weights. Thus it would not be right to utilize this model in real-time.

The results of the study emphasized in section 4 show how to use an xG model to profile players on the basis of the chances they create during a football match. Figure 4.14 displayed that though Andres Iniesta scored only 6 goals after a 100 shots on goal, this is not because he is a poor finisher but rather because of the positions he got

into and the chances he created. Similarly, figure 4.16 puts into perspective how often a top tier team like F.C. Barcelona creates chances that have high xG values.

This study also indicates which preceding play patterns and key passes help towards better chance creation. These features were not considered in prior studies conducted in [21], [3] and [17]. While the features considered in this study are distinct from that of many others, the fact that XGBoost algorithm followed closely by Logistic Regression should be considered for building Expected Goals models is something that was found to be a common link with others.

Future Work

The work presented in through this study successfully forms a bedrock for future metrics. An important question that arises within the minds of those who actively follow xG models is

Why does an xG model not inculcate the shot technique used and the shot velocity?

The answer to this question serves as the provenance to a new metric. The reason why those two features are not utilized is that an xG model is a *pre-shot* analysis. It simply measures a player/team's intelligence through understanding the kind of positions they get into during a football match and make the art of scoring easier. Therefore, to utilize a player's capability as a striker and a finisher, a metric to understand *post-shot* analysis needs to be created. This metric is called, Expected Goals on Target (xGOT). The following points need to be considered to develop the xGOT metric

- Placement of the shot
- Technique Used
- Velocity of the Shot
- The xG value of the Shot

A poor finisher could be hypothesised as one, who places a high xG valued shot in an easy to save position. This could be an example of how xGOT can be utilised to quantify player ability. Furthermore, xGOT can also be used to quantify the ability of the Goal Keeper saving the shot. A good Goal Keeper can be hypothesised as one who consistently saves football shots with high xGOT values.

Figure 6.1 highlights different coordinates of a goal for reference to shot placement. A good finisher can be hypothesised as one who places the shot aimed close to the four corners. Similarly a good goal keeper is one who consistently saves shots aimed at these four corners.

Another metric that is derived from xG is Expected Goals Against (xGA). This metric provides perspective in how any given team allows others to create chances against themselves. A defensively good footballing team would allow others to create chances with less xG values.

A third metric, that can be developed is one that analyses off-ball positioning first mentioned by author Spearman in [24]. This study aims at analysing the positioning

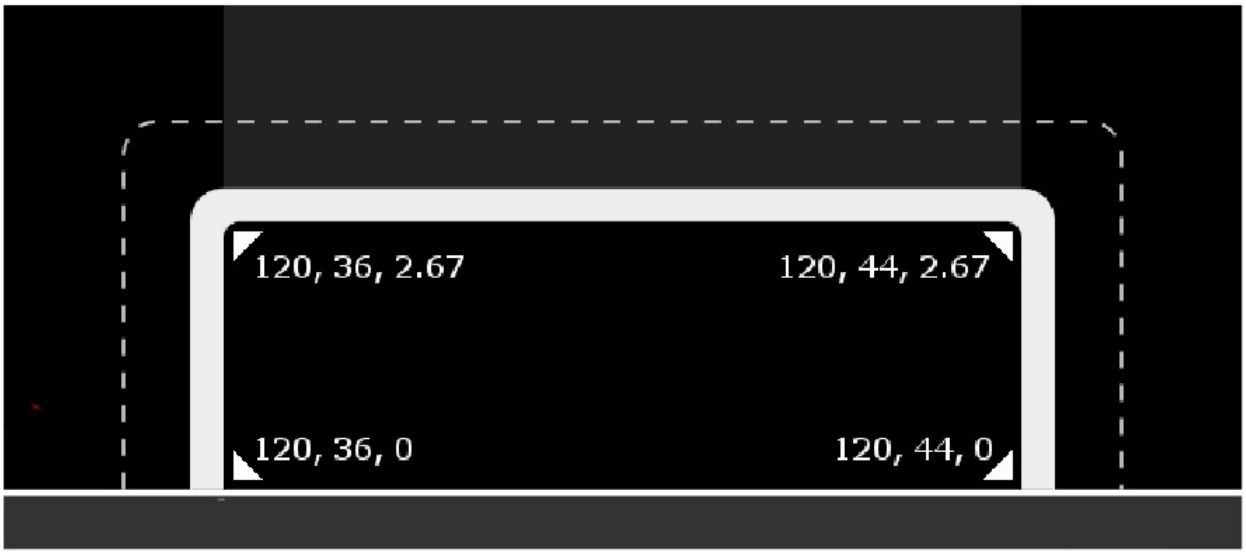


Figure 6.1: Shot Placement

of players and determining how likely they are to score while positioned without the ball highlighted in figure 6.2.

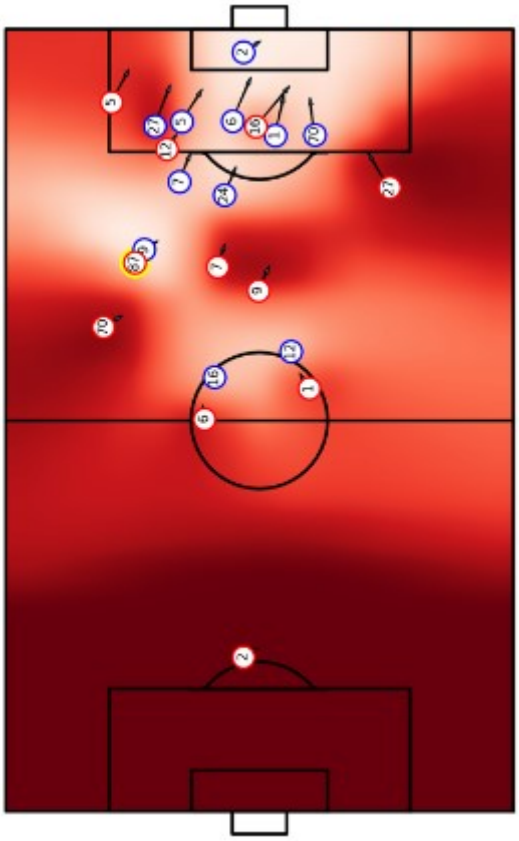


Figure 6.2: Off Ball Scoring

Appendix A: Project Plan

7.1 Introduction

This model was to be designed to understand Expected Goals. This chapter presents the project plan that was developed initially:

- Project Deliverables
- Document Conventions & Instructions to run the Project
- Milestones

7.2 Project Deliverable

The following are the deliverable of the project once completed:

- Project plan
- Project report
- Power Point Presentation
- Poster
- Source Code
- Meeting minutes

7.3 Conventions & Instructions to run the Project

The whole pipeline was implemented in Python. The libraries used are mentioned below:

- NumPy
- Pandas
- Beautiful Soup
- Matplotlib
- Plotly
- Seaborn
- Sci-Kit Learn
- LightGBM
- XGBoost
- CatBoost
- SHAP
- Google Colab
- Google Drive

7.3.1 Document Writing

This report was created in latex, through the Overleaf platform.

Naming Convention for Python

Python as a programming language employs indentation to split code chunks. Nevertheless, there are some code quality standards that must be fulfilled in order to create code that is both appealing and understandable. The following steps are followed to maintain quality:

1. **Imports** The import must be completed in a specific order. First, import the standard libraries, then third-party libraries, and ultimately local libraries.
2. **Indentation** Tab spaces are used for indentation.
3. **Blank Spaces** Unnecessary blank spaces must be prevented; just one blank space must be used around each side of any operator inculcated, one following a comma, and none within the beginning or closure of parenthesis.

The source code was submitted via drop-off, and is not included in the report.

7.3.2 Instructions for the Pipeline

There are five Jupyter Notebooks that this study has utilized:

- Data Extraction.ipynb
- Data Cleaning.ipynb
- EDA+Model.ipynb
- Pitch_Plot.ipynb
- Half_Pitch.ipynb

Data Extraction

- Contains three functions
- Needs to be called by Data Cleaning.ipynb
- Fetches Competition, Match and Event Data

Data Cleaning

- Contains functions that plot freeze frames and create a DataFrame from the data obtained from Data Extraction
- To run it needs to be connected to the user's google drive and the necessary files must be present in the file path provided

- Creates a final DataFrame and pickles it for further use

EDA+Model

- Calls the pickled DataFrame and implements Exploratory Data Analysis on it
- Includes functions to perform feature engineering on the features
- Includes the final Machine Learning and Hyper-Parameter Tuning

Half_Pitch and Pitch_Plot

- Half_Pitch.ipynb contains a function that allows visualisation of half a football pitch
- Pitch_Plot.ipynb contains a function that allows visualisation of a football pitch
- Both these files should be present in the main directory and can be called from any file

7.4 Milestones

This section was created after the study was completed, and the table below depicts the key milestones. This table solely displays the study's milestones.

Milestone	**Description**	**Date**
Draft Creation	Draft Literature Review Creation	04/05/2022
Project Targets	Project Targets were created	10/05/2022
Data Extraction	Data was extracted successfully	25/05/2022
Data Cleaning	Data Cleaning and Wrangling was carried out.	10/06/2022
Data Visualisation and Exploratory Analysis	Hypothesis Testing on data was carried out and data was visualised	18/06/2022
Feature Engineering	Initial features were created and Column Transformer was implemented	24/06/2022
Hyperparameter Tuning and Machine Learning	Machine Learning Models were shortlisted and Hyper-parameters were tuned	30/06/2022
Initial Results	Initial Results were obtained and Visualisation was initiated	07/06/2022
PowerPoint Presentation	Initial results were inculcated in the presentation and the presentation was finalized	10/06/2022
Further Model Tuning	Machine Learning Models were tuned further to finalize outputs	15/06/2022
Report Writing and Thesis Poster	Report Writing was initiated and Points for Poster were shortlisted	20/06/2022
Report Writing and Thesis Poster	Report Writing was finalized and Thesis Poster was finalized	13/08/2022

Table 7.1: Milestones

Bibliography

[1] Neptune AI. *Set Pice*. 2022. URL: https://neptune.ai/blog/hyperparameter-tuning-in-python-complete-guide (visited on 07/21/2022).

[2] Jim Albert. "Baseball data at season, play-by-play, and pitch-by-pitch levels". In: *Journal of Statistics Education* 18.3 (2010).

[3] Gabriel Anzer and Pascal Bauer. "A goal scoring probability model for shots based on synchronized positional and event data in football (soccer)". In: *Frontiers in Sports and Active Living* 3 (2021), p. 53.

[4] V Barnett and S Hilditch. "The effect of an artificial pitch surface on home team performance in football (soccer)". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 156.1 (1993), pp. 39–50.

[5] Christoph Biermann. *Football hackers: The science and art of a data revolution*. Kings Road Publishing, 2019.

[6] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[7] Tianxiang Cui et al. "An ensemble based Genetic Programming system to predict English football premier league games". In: *2013 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. 2013, pp. 138–143. DOI: 10.1109/EAIS.2013.6604116.

[8] Anna Veronika Dorogush et al. "Fighting biases with dynamic boosting". In: *CoRR* abs/1706.09516 (2017). arXiv: 1706.09516. URL: http://arxiv.org/abs/1706.09516.

[9] Richard Giulianotti. "Football". In: *The Wiley-Blackwell encyclopedia of globalization* (2012).

[10] T. Haifley. "Linear logistic regression: an introduction". In: *IEEE International Integrated Reliability Workshop Final Report, 2002.* 2002, pp. 184–187. DOI: 10.1109/IRWS.2002.1194264.

[11] Thomas Little Heath et al. *The thirteen books of Euclid's Elements*. Courier Corporation, 1956.

[12] Kou-Yuan Huang and Wen-Lung Chang. "A neural network method for prediction of 2006 World Cup Football Game". In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. 2010, pp. 1–8. DOI: 10.1109/IJCNN.2010.5596458.

[13] Anito Joseph, Norman E Fenton, and Martin Neil. "Predicting football results using Bayesian nets and other machine learning techniques". In: *Knowledge-Based Systems* 19.7 (2006), pp. 544–553.

[14] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[15] Simon Kuper and Stefan Szymanski. *Soccernomics: Why England loses, why Germany and Brazil win, and why the US, Japan, Australia, Turkey–and even Iraq–are destined to become the kings of the world's most popular sport*. Hachette UK, 2018.

[16] Ulrich Lichtenthaler. "Mixing data analytics with intuition: Liverpool Football Club scores with integrated intelligence". In: *Journal of Business Strategy* (2020).

[17] Pau Madrero Pardo. "Creating a model for expected goals in football using qualitative player information". MA thesis. Universitat Politècnica de Catalunya, 2020.

[18] Eli Maor. *The Pythagorean theorem: a 4,000-year history*. Princeton University Press, 2019.

[19] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[20] Washington Post. *Set Pice*. 2018. URL: `https://www.washingtonpost.com/news/fancy-stats/wp/2018/06/20/why-set-pieces-are-dominating-scoring-so-far-at-the-world-cup/` (visited on 06/20/2018).

[21] Alex Rathke. "An examination of expected goals and shot efficiency in soccer". In: *Journal of Human Sport and Exercise* 12.2 (2017), pp. 514–529.

[22] Alexander P Rotshtein, Morton Posner, and AB Rakityanskaya. "Football predictions based on a fuzzy model with genetic and neural tuning". In: *Cybernetics and Systems Analysis* 41.4 (2005), pp. 619–630.

[23] Brian Skinner. "The price of anarchy in basketball". In: *Journal of Quantitative Analysis in Sports* 6.1 (2010).

[24] William Spearman. "Beyond expected goals". In: *Proceedings of the 12th MIT sloan sports analytics conference*. 2018, pp. 1–17.

[25] Statsbomb. *Statsbomb open data*. 2018. URL: `https://github.com/statsbomb/open-data` (visited on 01/01/2017).

[26] Karl Tuyls et al. "Game Plan: What AI can do for Football, and What Football can do for AI". In: *CoRR* abs/2011.09192 (2020). arXiv: `2011.09192`. URL: `https://arxiv.org/abs/2011.09192`.