School of Mathematical and Physical Science
University of Sussex

Predicting Football Match Results with Machine Learning Approaches Using Expected Goals

A dissertation project report submitted for the award of:
MSc Data Science

15th August 2024

Abstract

This dissertation explores the application of machine learning models to predict football match outcomes, focusing on the integration of the Expected Goals (xG) metric to enhance the prediction outcomes. A custom Expected Goals model was developed using a dataset from the top five European football leagues, incorporating basic shot features such as shot location, shot type, and the body part used to take the shot. The Expected Goals model was then used to generate xG values to represent team performance in the match prediction models. After that, this study utilized various machine learning algorithms, including Logistic Regression, Support Vector Machines (SVMs), Random Forests, and Artificial Neural Networks (ANN) to predict match results as Home Team Win, Away Team Win or Draw. The findings indicate that while incorporating Expected Goals improved the probabilistic predictions of the models, the baseline predictions derived from bookmakers remained the most well-calibrated. Among the proposed models, Logistic Regression was performed closest to the baseline, demonstrating reasonable performance, while SVMs, despite being the most accurate in terms of classification, had the worst calibration. Whilst the result suggests a positive impact of the incorporation of Expected Goals, it also indicated the dominant influence of traditional match statistics like average number of passes and goal difference on models' predictions.

Contents

Chapter 1
Introduction

1.1 Data Science in Football

Football, also known as soccer or association football, originated in England but has rapidly transcended borders, uniting people worldwide. As the world's most popular sport, it captivates billions of fans globally.

In recent years, sports analytics, particularly football analytics, has seen an incredible amount of interest. By exploiting technological advancements and new methods of tracking players, the football science industry is experiencing a data explosion with a vast number of events and many types of data being collected in each game. Consequently, researchers are increasingly seeking advanced techniques to uncover the game's secrets and achieve informative insights. No doubt that one of the most expected insights is the predictions of match outcomes with great accuracy. The ability to predict match outcomes empowers various stakeholders in the football world, enabling them to make informed decisions that provide a competitive advantage. Information from predictions can inform strategies and tactics, allowing teams to exploit opponent weaknesses and maximise their chances of winning. For players, understanding the key factors leading to favourable results may improve their training, promoting their contribution to the team and increasing their value. Other side, accurate predictions are huge advantages in the commercial aspect of football, with the betting market, for instance. Therefore, predicting match outcomes has become a widespread practice, attracting football professionals, managers, media, passionate fans, and businesses such as investors and bookmakers.

With the well-known capability of capturing relationships between provided input and possible outcomes, data science has opened a new era of football analytics, where machine learning is an excellent tool for making match predictions.

1.2 Motivation

Despite football's immense popularity and the advancements in technology, accurately predicting match results remains challenging. The sport's stochastic nature, with numerous factors influencing the outcomes, creates a complex scenario where unpredicted events during the 90-minute window, such as injuries and red cards, can be pivotal elements affecting the match result. Additionally, the rarity of goals, which determine the results, compared to other sports adds another layer of complexity to predictions.

Traditional methods for predicting match results heavily rely on expert opinions, experience, and intuition, which can be subjective and lack specific insights. This creates an opportunity to develop better forecasting tools that provide more objective and data-driven information. One approach that has garnered significant interest is machine learning (ML), a powerful technique within artificial intelligence (AI) that has shown promising results in prediction tasks.

While conventional ML models have used features such as previous win/loss records and the number of goals scored, the inherent rarity of goals in football can lead to an

inconsistency between a team's performance and their actual goal count. A dominating team can still lose if their strikers have an unlucky day and fail to score. In ML models, such outcomes might negatively impact the predicted chances of winning future games despite good performance metrics. This is where the innovative concept of Expected Goals (xG) becomes crucial. xG is a statistical metric that estimates the probability of a goal being scored based on the quality of a shot. By measuring the quality of chances, xG provides a better representation of a team's offensive performance.

1.3 Objectives

This project aims to explore the potential of various machine learning models for predicting match outcomes, with a particular focus on incorporating the innovative concept of Expected Goals (xG) to gain deeper insights into team performance and improve prediction accuracy.

Our objectives include leveraging machine learning techniques to develop prediction models using metrics representing teams' historical performances, such as goals scored and conceded, corners, freekicks, and recent results. These models predict match outcomes as a win for the home team, a win for the away team, or a draw. Using different performance metrics, our models will be assessed against a benchmark prediction, which is bookmakers' odds. A good prediction will align with the actual outcome probability, meaning a model predicting the probability of home team wins that is closer to the real chance of a win for the home team will be considered successful. Prediction calibration will serve as the primary evaluation metric.

Additionally, we will compare the performance of these machine learning models using traditional goal scores and the innovative xG metric. The main approach we will take is to build a model for expected goal statistics in order to understand a team's performance better and thus generate more reliable predictions for the future. By incorporating expected goals with these models, we aim to gain valuable insights into the effectiveness of implementing the xG metric in predicting football match results.

Our investigation will address several key questions:

- Does incorporating Expected Goals into the analysis improve the prediction accuracy of machine learning models?
- How do our developed machine learning models predict football match outcomes compared to predictions from bookmakers?
- Which one of the proposed machine learning models performs best?

By answering these questions, this study aims to contribute to the field of football analytics and provide valuable insights, particularly about applying the Expected Goals in the industry.

Chapter 2
Background Research

2.1 Predicting Match Results with Machine Learning

In terms of Artificial Intelligence, machine learning algorithms provide computer systems with the ability to learn from data and enhance their performance on specific tasks through experience. Supervised learning, a specific type of Machine Learning, focuses on training the system using data that includes input information (features) and the desired outcome (labelled data). By learning the relationships between features and labelled data, these systems can predict the corresponding outcome for unseen inputs.

Given football's popularity as one of the most widely followed sports globally (Fan et al. 2023; Müller, Simons, and Weinmann 2017), a significant amount of research has been conducted on applying machine learning to football prediction. Previous research in this area can be broadly classified into two main categories: result-based studies and goal-based studies. Goal-based approaches focus on predicting the number of goals each team will score in a specific match, while result-based approaches are treated as a classification problem (Bunker and Thabtah 2019) and aim to predict the overall outcome of the match (win, lose, or draw). This study aligns with the latter category, focusing on predicting the match results with prediction tasks using machine learning.

In past studies, various techniques have been explored for the prediction of football match results. For instance, Shtovba, Dounias, and Tsakonas (2003) employed Support Vector Machines (SVMs) to forecast football match outcomes. They utilized data from the Ukrainian football championship over ten years. The SVM model was employed to perform a regression task, where the predicted value was compared to zero to classify the match outcome. By ignoring the draws and achieving 61.4% correct classification, their approach demonstrated the potential of SVMs in handling complex classification problems in sports analytics. Joseph, N. E. Fenton, and Neil (2006) developed a framework using Bayesian networks to predict the results of Tottenham Hotspur matches. Their study compared the Bayesian network framework with other machine learning techniques, illustrating the effectiveness of probabilistic models in sports prediction. Bayesian networks provided a structured way to incorporate prior knowledge and update predictions based on new evidence, making them suitable for dynamic and uncertain environments like football matches. Even though the study found that Bayesian networks performed well in predicting match outcomes with an accuracy of 59.21%, their model incorporated domain-specific knowledge of only one team in a specific period of time. Berrar, Lopes, and Dubitzky (2019) employed k-Nearest Neighbors (k-NN) and Extreme Gradient Boosting (XGBoost) for football match outcome prediction. Similarly, Hubáček, Šourek, and Železný (2019), winners of the 2017 Soccer Prediction Challenge, utilized Gradient Boosted Trees and Relational Dependency Networks (RDN-Boost) to develop their predictive models. Choi, Foo, and Chua (2023) investigated the use of different machine-learning approaches to predict football match outcomes. Their study compared the performance of models like Logistic Regression, Support Vector Machines, Random Forest, and Neural Networks. They found that the Linear Regression model has the best performance for multiclass

match result prediction while the Random Forest model performs better among binary models.

The studies in this area also emphasise the importance of feature engineering, with the common use of match statistics data. Choi, Foo, and Chua (2023), Baboota and Kaur (2019), and Rodrigues and Pinto (2022) computed the aggregate value of match statistic indicators for both the home and the away teams based on previous games using either average or sum. These studies highlight that engineered features significantly contribute to the performance of predictive models. It is noticeable that in the Random Forest model of Baboota and Kaur (2019)], only two out of the twelve best-performing features were not engineered, highlighting the critical role of feature engineering in predicting football match outcomes. Beyond basic match statistics, researchers have explored innovative approaches. Hvattum and Arntzen (2010) adapted the Elo rating system, originally proposed by Elo (1978) for ranking international chess players, to football match prediction. Subsequently, Constantinou and Norman Elliott Fenton (2013) developed an alternative measure called pi-ratings, which performed better than the proposed Elo ratings. Their model accounted for various factors such as team strength, form, and match location. Another interesting approach involves leveraging collective knowledge from social media platforms. Schumaker, Jarmoszko, and Labedz (2016) utilized language data from platforms like Twitter. They specifically used the sentiment analysis of tweets to predict match results, demonstrating how social media sentiment can serve as a valuable feature in the predictive modelling of football outcomes.

Among the studies, one common challenge in football match prediction is the accurate prediction of draws. Pappalardo and Cintia (2018) proposed that wins and loses are often strongly correlated with team performance metrics, but since their model was not able to capture draws, they are unlikely predictable, demonstrating that draw matches are challenging to predict using machine learning due to their ambiguous nature in performance metrics. Danisik, Lacko, and Farkas (2018) reported a similar finding that their LSTM (long short-term memory) model failed to predict draws effectively. Baboota and Kaur (2019) highlighted that all their models, including Gaussian Naive Bayes, support vector machine, random forest, and XGBoost, under-predicted draws, attributing this issue to imbalanced datasets, as draws are relatively rare compared to wins and losses.

2.2 Expected Goals (xG)

Many models mentioned above have been developed using historical information on the teams competing to estimate teams' strengths and, therefore, calculate the probabilities of the result (win, draw, loss) or predict the scoreline. These models are based on key features in football, such as goals scored and conceded, possessions, shots and passes taken, and recent results. While these machine learning models offer promising improvements for football match outcome prediction, traditional features like goals scored have limitations. For example, a team might dominate during the match, have better possession, and create high-chance scoring opportunities but still lose the game because of unexpected mistakes or the opponent's lucky goals. Recognizing this limitation, researchers have developed and begun exploring the use of Expected Goals in football analytics.

Expected Goals (xG) is a statistical metric used to measure the quality of goal-scoring opportunities and evaluate a team's performance beyond the actual goals scored. This concept aims to provide a more nuanced understanding of a team's offensive performance

by indicating how many goals a team or a player should have scored, considering the various statistical factors that have been observed during the match. The concept of Expected Goals was first introduced in hockey by Macdonald (2012), and it has since been adapted and refined for football. The core idea is to assign a probability to each shot, representing its likelihood of resulting in a goal. The Expected Goals models are predictive models that assign a probability between 0 and 1 to each shot taken, reflecting the possibility of it resulting in a goal. Expected Goals models are trained on a variety of features extracted from events and tracking data by using different approaches, including logistic regression, gradient boosting, neural networks, support vector machines, and tree-based algorithms.

Expected Goals models are primarily defined by the features they incorporate in order to capture the contextual information of a shot. Rathke's research (Rathke 2017) emphasised the importance of shot location and the angle of the shot as primary determinants of goal probability. The paper highlights that shots taken closer to the goal and at more favourable angles have higher xG values, thus indicating a greater likelihood of scoring. Other features are the type of play that led to the shot, the body part used (head, foot), the specific shot type (volley, free kick), and the shooting technique. The author also discusses how xG can provide insights into a team's offensive efficiency by comparing the actual goals scored to the expected goals, thereby identifying over- or under-performing teams and players. An online article by Robberechts and Davis (2020b) highlights the importance of feature representation in xG models. It compares eight different ways to encode shot locations, illustrating how these representations affect model performance and underscoring the critical role that feature engineering plays in enhancing the predictive power of xG models. Newer studies incorporate additional features such as the positioning of the goalkeeper and defenders, the type of pass received by the shooter (e.g., through ball, cross), and whether the player taking the shot is under pressure from defenders, providing a more holistic view of the scoring opportunity. Mead, O'Hare, and McMenemy (2023) focus on improving the predictive accuracy of xG models by incorporating various machine-learning techniques with additional features. Their study demonstrates that including these features significantly enhances the model's performance with a Brier score of 0.07908, which is superior to the score of 0.0799 from a previous study. This research underscores the value of xG in providing a more comprehensive analysis of match dynamics and player performance. Another approach is a study that develops models that consider specific players or positions. Hewitt and Karakuş (2023) introduce a machine-learning approach to adjust xG models based on player and position-specific factors. Their model incorporates features such as player quality, defensive pressure, and contextual information about the match. By adjusting for these factors, their xG model provides a more accurate reflection of a player's contribution to goal-scoring opportunities, revealing insights into player efficiency and can be used to evaluate individual performance more effectively.

Expected Goals models offer valuable insights for player and team evaluation. Compared to traditional goal-based metrics, xG offers a significant advantage – it considers the frequency of scoring chances, which are much more common than goals. With the low-scoring nature of football, analysing expected goals provides a more objective view of in-game performance, indicating how many goals a team likely scored based on the opportunities created. Brechot and Flepp (2020) proposed using xG models instead of match outcomes for performance assessment to overcome the randomness in short-term results. Eggels, Elk, and Pechenizkiy (2016) propose a method to determine the expected winner of a match based on the aggregation of xG values for all goal-scoring opportunities.

Additionally, Tureen and Olthof (2022) developed the "Estimated Player Impact" (EPI) measure, which quantifies the contribution of individual players to each xG value. Their findings reveal that Heung-Min Son, a forward for Tottenham Hotspur, had the highest positive impact on EPI per xG, suggesting his exceptional ability to create high-quality scoring chances. In an analysis by Raudonius and Seidl (2023), the authors used their xG model to discuss the differences in shooting tendencies and success rates between players in four different levels of German football. These studies present the expanding role of xG models in analysing football and show their promising impact in predicting match results.

Chapter 3
Data Processing

3.1 Dataset

The dataset used in this research encompasses matches from the 2017/2018 season of the top five European leagues: the English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, and French Ligue 1. The data was collected from a combination of publicly available sources covering two primary types of information: historical match statistics and odds data.

3.1.1 Wyscout Data

The first data source used is a public data set of spatio-temporal match events provided by Pappalardo, Cintia, et al. (2019). The data has been collected and supplied by Wyscout, a leading company in football analytics. The dataset refers to the 2017/2018 season and contains information about competitions, matches, teams, and players. Most importantly, the dataset includes detailed events that occur during each match. Event data describes all of the actions on the pitch. These are typically on-the-ball actions, which, for example, could be passes, shots, dribbles, and tackles. It is important to note that while Wyscout developed their Expected Goals metrics, this public dataset does not include this specific measure. To address this, a custom Expected Goals model will be built for this research.

While the dataset also covers other competitions, such as the FIFA World Cup 2018 and European Cup 2016, which are competitions for national teams, only data from the five mentioned national competitions in Europe will be used to ensure uniformity. This results in a usable dataset comprising 1,826 matches and 3,071,395 events (Table 3.1).

Table 3.1: An overview of Wyscout dataset.

Competition	**# matches**	**# events**
Spanish first division	380	628,659
English first division	380	643,150
Italian first division	380	647,372
German first division	306	519,407
French first division	380	632,807
Total	1,826	3,071,395

3.1.2 Bookmaker Data

The second dataset consists of odds data collected from football-data.co.uk, which will be used to compare with our predictive models. Bookmakers rely heavily on accurate match outcome predictions to set profitable odds and maintain a successful betting market.

Therefore, the accuracy of bookmakers' predictions serves as a valuable baseline for evaluating our models.

In addition to odds from various bookmakers, the football-data.co.uk dataset provides common historical statistics of matches, such as the number of shots, corner kicks, yellow cards, and red cards, which are also usable for this research.

3.1.3 Understat xG

To evaluate the impact of Expected Goals on predictive models, an additional dataset with xG measures computed by another provider will be used for comparison. For this purpose, xG values from understat.com (abbreviated as Understat) will be utilized. Understat employs neural network algorithms to train their xG models using a large amount of shot data (over 100,000 shots with more than 10 parameters for each). The xG values from Understat are well-regarded and have been adopted as benchmark xG metrics in several articles (Cefis 2024; Partida et al. 2021).

3.2 Expected Goals Model

3.2.1 Data Descriptions

The shots dataset utilized in this study was extracted from the Wyscout event dataset. Across all five competitions, there was a total of 43,040 recorded shots, which resulted in 4,790 goals. This gives a goal conversion rate of approximately 11.13%, highlighting the rare goals in professional football (Table 3.2).

Table 3.2: Shots and Goals from Wyscout dataset.

Competition	**# shots**	**# goals**	**conversion rate**
Spanish first division	8,545	993	11.62%
English first division	8,881	988	11.12%
Italian first division	9,347	978	10.46%
German first division	7,290	833	11.42%
French first division	8,977	998	11.12%

In terms of data adequacy, Robberechts and Davis (2020a) have demonstrated that five seasons of data are generally sufficient for training a complex Expected Goals model. Their findings indicate that such data does not become outdated quickly and that models trained on league-specific data do not significantly outperform those trained on more generalized datasets. This supports the use of our dataset as a robust foundation for building a reliable xG model.

To better understand the distribution and effectiveness of shots on the pitch, a series of analyses were conducted, including heatmaps of shot and goal locations. Figure 3.1 revealed that the highest concentrations of shots and goals occur in the central area of the penalty box, particularly just outside the 6-yard box. This is consistent with the common understanding that shots taken from central positions close to the goal are more likely to result in goals. Additionally, we observed significant clusters of shots taken from just

outside the penalty box. These likely represent a mix of long-range attempts, which are also common in football matches.

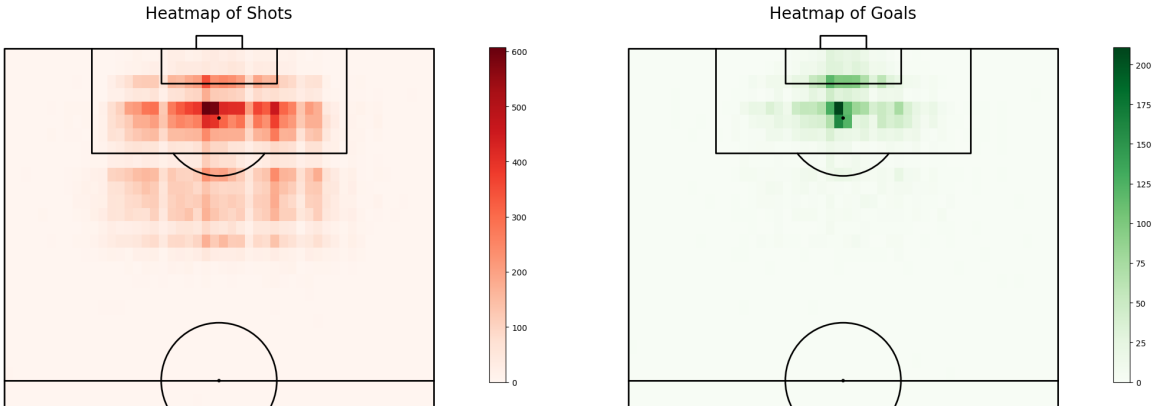

Figure 3.1: Heat maps of numbers of Shots and Goals from different areas.

Furthermore, we analysed the probability of scoring from different areas on the pitch to identify high-probability scoring zones. By applying a shot threshold, where at least five shots have been taken, we filtered out less frequent shooting locations that might otherwise appear as high-probability areas due to their small sample size. This approach provided a clearer and more accurate representation of the most effective scoring zones on the pitch. The resulting heatmap (Figure 3.2) indicates that the central area directly in front of the goal within the penalty box is the most critical zone for scoring. Conversely, while a substantial number of shots are taken from areas just outside the penalty box, these are generally less likely to result in goals.

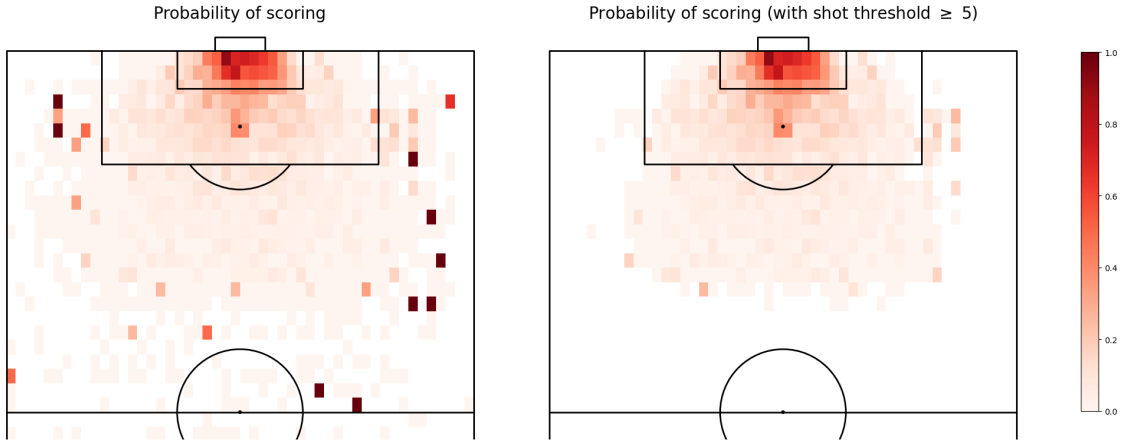

Figure 3.2: Heat maps of Probabilities of scoring from different areas.

3.2.2 Feature Engineering

Developing an accurate Expected Goals model requires constructing a robust set of features that effectively describe the context and characteristics of each shot attempt. Traditional Expected Goals models often rely solely on basic shot-specific information, such as the shot's location on the pitch. However, modern Expected Goals models incorporate additional, advanced information that can significantly enhance the model's predictive accuracy.

Since one of the goals of this study was to investigate the contribution of Expected Goals, the first step was to develop a usable Expected Goals model that generates reasonable xG values for individual shots, which can then be incorporated in match result predictive models. Due to the limitations of our dataset, we focused on extracting key features related to the shot's location, the type of shot, and the body part used to take the shot.

The first step in feature engineering involved standardising the coordinates of each shot's location to align with the standard UEFA pitch dimensions. This standardization process was crucial because it ensured that all spatial features derived from the shot location were consistent and accurate, regardless of the original source of the data. From the standardised shot location coordinates, we computed two critical spatial features, as demonstrated in Figure 3.3: the distance to the goal and the visible angle. The distance feature represents the straight-line distance from the shot's location to the centre of the goal. This is a fundamental feature in Expected Goals models, as shots taken from closer distances are generally more likely to result in goals. The visible angle measures the angle between the shot location and the two goalposts, capturing the width of the goal that is visible from the shot location. This feature was computed using the following formula:

$$\theta = \tan^{-1}\left(\frac{\text{goal\_width} \times d_1}{d_1^2 + d_2^2 - (\frac{\text{goal\_width}}{2})^2}\right)$$

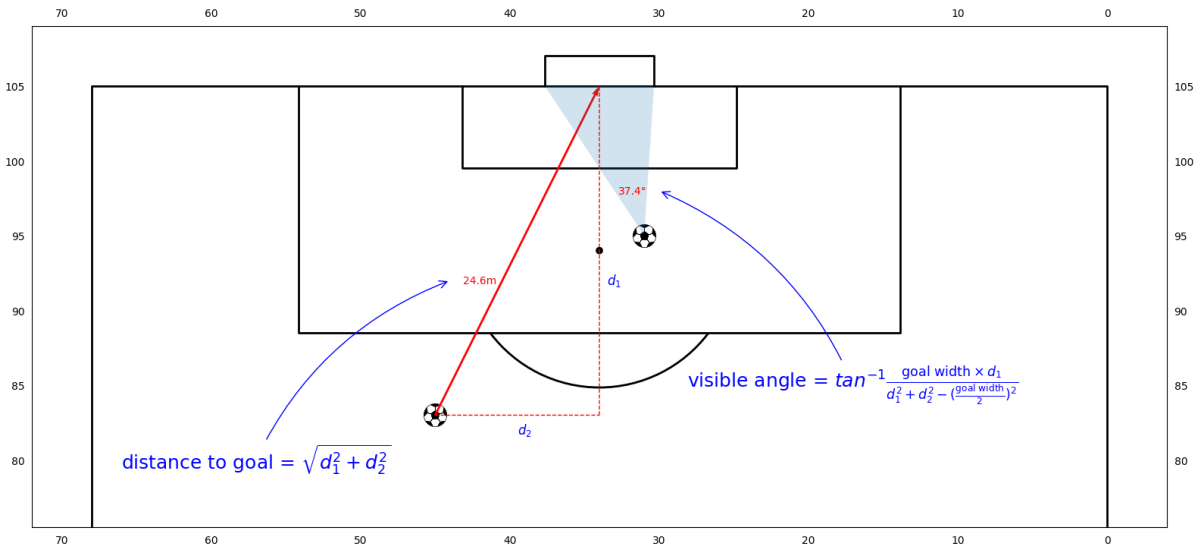

Figure 3.3: Calculation of shot's distance and shot's visible angle.

In addition to the spatial features, the feature set also included information on the type of shot and the body part used to take the shot. Utilising the player dataset, we were able to construct an additional feature that indicates whether the shot was taken with the

player's strong foot (e.g., right foot for a right-footed player). This feature adds valuable context to the shot, as shots taken with a player's strong foot are generally more accurate and powerful and, therefore, could increase the likelihood of scoring. In Table 3.3, all the variables used in the expected-goals model are listed, along with a brief description of each.

Table 3.3: Expected Goals model's features description.

Feature	**Type**	**Description**
shot_type	Categorical	Type of shot (Open play, Freekick, Penalty)
body_part	Categorical	Body part use to take the shot (Right foot, Left foot, Header/Other)
is_strong_foot	Boolean	Whether the shot is taken by the player's strong foot
x	Numerical	x coordinate of the location where the shot is taken
y	Numerical	y coordinate of the location where the shot is taken
distance	Numerical	Shot's distance to goal in metres
visible_angle	Numerical	Shot's visible angle in degrees

3.3 Match Result Predictive Models

3.3.1 Feature Engineering

Previous research, including studies by Joseph, N. E. Fenton, and Neil (2006) and Odachowski and Grekow (2013), has demonstrated that the quality of the results in football match prediction tasks is directly linked to the quality of the feature set used for modelling. Recognizing this, we consider the engineering of features as a crucial aspect of this research due to its potential enhancement in forecasting football match results. This section details the rationale behind the engineered features and their mathematical formulations.

Since both prior research and intuitive understanding suggest the well-established impact of the home/away factor on match outcomes, this feature is treated as a global characteristic in our models. Despite its apparent simplicity, this factor proved to be highly influential. Consequently, all engineered features are computed separately for both the home and away teams to capture this dynamic accurately.

Match statistics

In terms of capturing teams' performances, match statistics were extracted from the football-data.co.uk dataset. The dataset includes team names and various match statistics such as Goals, Shots, Shots on Target, Corners, Fouls, Red Cards, and Yellow Cards. Although these features can also be derived from event data provided by Wyscout, data collected by football-data.co.uk showed figures closer to those published on official league

websites, which are considered more reliable. Additional statistics derived from Wyscout's event data include the number of Passes, Free Kicks, Offsides, Clearances, and Saves (*Wyscout Glossary* 2024). To encapsulate a team's overall performance, match results were encoded into points following the common league rule: Win = 3 points, Draw = 1 point, and Lose = 0 points.

Expected Goals values

As proposed, the Expected Goals (xG) metric was incorporated as an additional performance indicator. xG values were derived using a pre-built Expected Goals model (as described later in the section below) applied to shot data recorded by Wyscout. This model generates an xG value for each shot, representing the probability of that shot resulting in a goal based on various factors such as shot location and body part used. To integrate xG into our predictive models, we explored two approaches:

In this traditional approach, the xG values for each team in a match were summed, representing the total number of goals the team was expected to score based on the quality of their chances (Eggels, Elk, and Pechenizkiy 2016). This method provides a straightforward measure of a team's attacking potential for a match. Additionally, the Understat's pre-aggregated match xG values were taken into account for comparison.

However, this aggregation of xG values over a game may not tell a full story of the team's ability to create chances. For example, a team that created only a big chance with a high xG value throughout a match can have an equal sum-up xG value with the opposite team that created multiple small chances. To investigate this, in the second approach, we utilized the xG values of all individual shots and their timing within the match. By incorporating the sequence and timing of shots, this approach aims to provide a more detailed representation of a team's offensive capabilities, potentially improving the predictive power of our models.

Historical Match Statistics

For football match prediction tasks, it is crucial to use features that are known before the start of each match to ensure the model can make pre-match predictions. As the extracted features relate to the end of each match and reflect past performance, they cannot be used directly to train the prediction models. Instead, the average values of these metrics over the past k games were considered to reflect the teams' recent performance. Following the suggestion in previous studies, an optimised value of $k = 6$ was used, resulting in the exclusion of matches from the first six rounds of each competition.

Goal Difference

The last considered feature is the Goal Difference. Empirical evidence, such as the study by Constantinou and Fenton (2013), demonstrates that Goal Difference is an important feature of predictive modelling. Goal Difference is calculated as the number of goals scored minus the number of goals conceded in all matches. For a team's n^{th} match, the Goal Difference (GD) is defined as:

$$GD_k = \sum_{i=1}^{n-1} GS_i - \sum_{i=1}^{n-1} GC_i$$

Where GS = Goals Scored and GC = Goals conceded
Table 3.4 summarises the 38 features included in our final dataset.

Table 3.4: Match result predictive model's features description.

Feature	**Type**	**Description**
home_team	Categorical	Name of the home team
away_team	Categorical	Name of the away team
avg_home_shot	Numerical	The average number of shots by the home team over the past 6 matches
avg_away_shot	Numerical	The average number of shots by the away team over the past 6 matches
avg_home_sot	Numerical	The average number of shots on target by the home team over the past 6 matches
avg_away_sot	Numerical	The average number of shots on target by the away team over the past 6 matches
avg_home_pass	Numerical	The average number of passes by the home team over the past 6 matches
avg_away_pass	Numerical	The average number of passes by the away team over the past 6 matches
avg_home_corner	Numerical	The average number of corners by the home team over the past 6 matches
avg_away_corner	Numerical	The average number of corners by the away team over the past 6 matches
avg_home_freekick	Numerical	The average number of free kicks by the home team over the past 6 matches
avg_away_freekick	Numerical	The average number of free kicks by the away team over the past 6 matches
avg_home_foul	Numerical	The average number of fouls by the home team over the past 6 matches
avg_away_foul	Numerical	The average number of fouls by the away team over the past 6 matches
avg_home_yellow	Numerical	The average number of yellow cards for the home team over the past 6 matches
avg_away_yellow	Numerical	The average number of yellow cards for the away team over the past 6 matches
avg_home_red	Numerical	The average number of res cards for the home team over the past 6 matches
avg_away_red	Numerical	The average number of res cards for the away team over the past 6 matches
avg_home_offside	Numerical	The average number of offsides by the home team over the past 6 matches
avg_away_offside	Numerical	The average number of offsides by the away team over the past 6 matches
avg_home_clearance	Numerical	The average number of clearances by the home team over the past 6 matches

avg_away_clearance	Numerical	The average number of clearances by the away team over the past 6 matches
avg_home_save	Numerical	The average number of saves by the home team over the past 6 matches
avg_away_save	Numerical	The average number of saves by the away team over the past 6 matches
avg_home_point	Numerical	The average number of points for the home team over the past 6 matches
avg_away_point	Numerical	The average number of points for the away team over the past 6 matches
home_goal_diff	Numerical	Current home team's goal difference (before the match)
away_goal_diff	Numerical	Current away team's goal difference (before the match)
avg_home_goal	Numerical	The average number of goals by the home team over the past 6 matches
avg_away_goal	Numerical	The average number of goals by the away team over the past 6 matches
avg_home_xg	Numerical	The average number of xG values for the home team over the past 6 matches
avg_away_xg	Numerical	The average number of xG values for the away team over the past 6 matches
avg_home_ud_xg	Numerical	The average number of Understat's xG values for the home team over the past 6 matches
avg_away_ud_xg	Numerical	The average number of Understat's xG values for the away team over the past 6 matches
home_xg_vector	Array	List of xG values for shots by the home team over the past 6 matches
away_xg_vector	Array	List of xG values for shots by the away team over the past 6 matches
home_xg_time_vector	Array	List of corresponding time for shots by the home team over the past 6 matches
away_xg_time_vector	Array	List of corresponding time for shots by the away team over the past 6 matches

3.3.2 Feature Selection

Having a large number of features increases the dimensionality of the feature space. Therefore, a step of feature selection was performed to identify the most impactful and relevant ones.

First, the relationship between features and match results was investigated. While most features exhibited distinct patterns in their distributions by match outcomes (Home Team Win, Away Team Win, and Draw), some features like Freekicks, Fouls, Yellow Cards, Red Cards, and Offsides did not offer visually separate distributions (Figure 3.4).

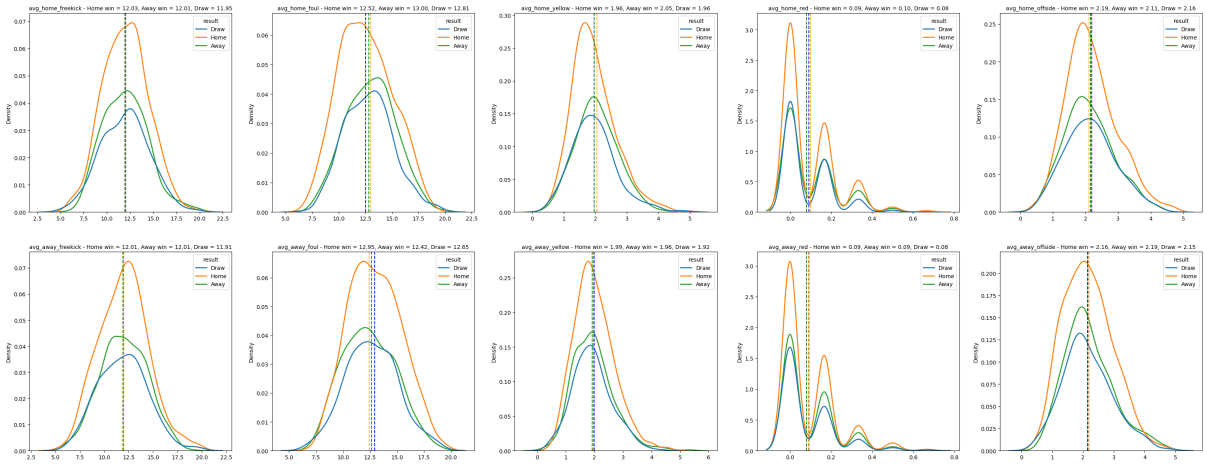

Figure 3.4: KDE plot by match results for 5 features:Freekicks, Fouls, Yellow Cards, Red Cards, and Offsides.

Given that Home Team Win and Away Team Win are opposite outcomes, differences in performance metrics for each team were expected. A Kolmogorov-Smirnov test was conducted to assess the distribution differences in performance metrics for home and away teams. The test indicated that the distributions of the 5 pairs mentioned metrics were not significantly different at the 0.1% significance level, consequently leading to the exclusion of these features from the predictive models.

A correlation matrix (Figure 3.5) was computed to identify highly correlated features. Interestingly, the analysis revealed that the home team's variables had low correlations with their counterparts for the away team. This is generally a good indication as it suggests that each set of features contains distinct and non-redundant information about the teams' performances. On the other hand, some performance metrics of a team, such as average shots, average shots on target, and average corners, showed high correlations with each other. For example, the correlation between average shots and average shots on target was as high as 0.81. These high correlations are logically reasonable, as teams that take more shots are also likely to have more shots on target and corners. However, the presence of such strong correlations suggests that these features might be redundant, which can negatively affect the model's performance.

To address this, a feature selection technique, namely Recursive Feature Elimination (RFE) was adopted. RFE works by building multiple models, sequentially eliminating features from the entire pool, and eventually returning the most relevant features based on the model's performance. After conducting RFE with the Logistic Regression algorithm, an interesting observation was made: despite the high correlations among some performance metrics, RFE retained all of them. This outcome suggests that, even though these features are highly correlated, they each contribute unique and essential information that the model relies on for accurate predictions. On the contrary, RFE removed features related to the teams' names, which aligns with expectations since the team's name itself is not a performance metric and likely does not provide predictive value independent of the teams' actual performance statistics. This process left us with a refined set of 22 features (Figure 3.6).

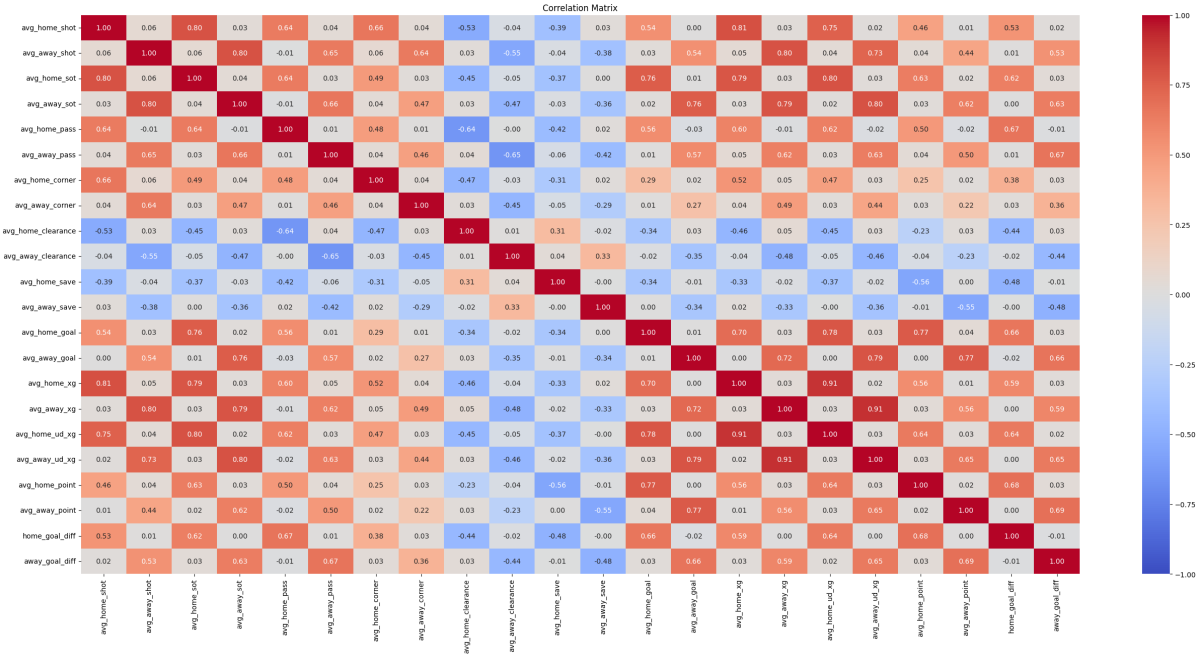

Figure 3.5: Correlation Matrix between all numerical features.

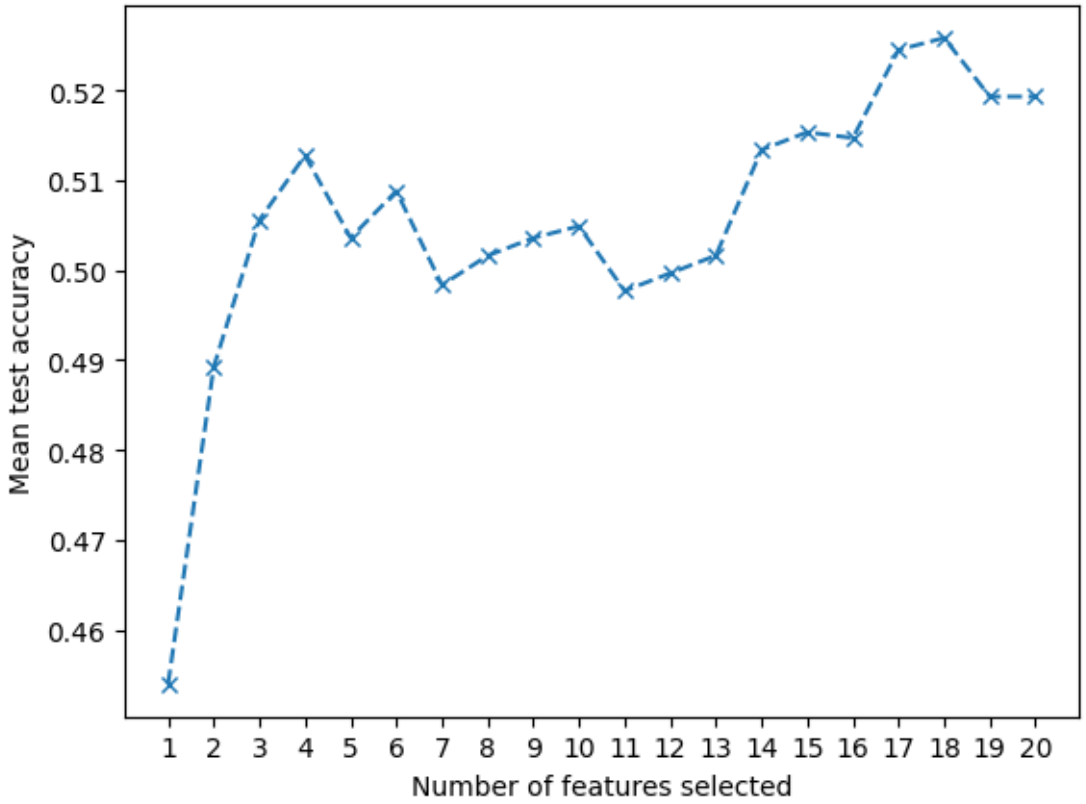

Figure 3.6: Feature selection with RFE.

Chapter 4
Models and Experiments

4.1 Expected Goals Model

In this section, we trained an Expected Goals model to address the binary classification problem of predicting whether a specific shot will result in a goal. This model is essential for generating xG values for individual shots, which can then be aggregated to calculate the cumulative xG values for each team in matches, which is a critical performance metric that we want to investigate its impact on match results predictive models.

4.1.1 Data Sampling

In machine learning, it is a common practice to split the data into a training set and a test set. The goal is to train a model that performs well on out-of-sample data, ensuring that it generalizes effectively to new, unseen instances. A standard train and test data split of 70% to 30% was chosen for all experiments assessed. In the context of our Expected Goals model, where only 11.13% of shots result in goals, we face a significant class imbalance issue within the target variable (Figure 4.1). This imbalance poses a challenge, as models trained on imbalanced data tend to be biased towards the majority class (in this case, shots that do not result in goals). To address this challenge, we investigated two data sampling strategies:

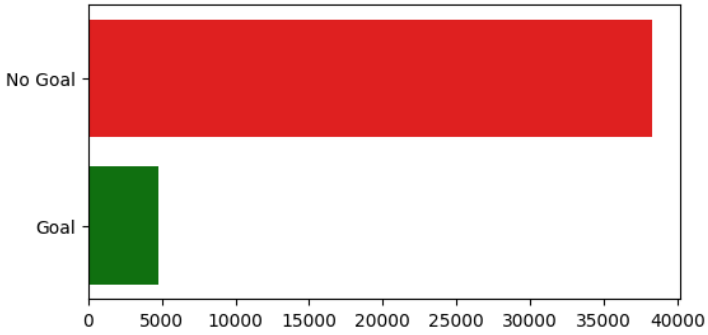

Figure 4.1: Outcomes of shots.

Stratified Splitting

The first approach was to use a stratified splitting strategy. Stratification ensures that the proportion of goals and no goals is maintained in both the training and test sets. This approach helps preserve the original distribution of the classes, which is crucial for evaluating the model's performance on the test data. Stratified splitting ensures that both sets have a representative sample of the minority class (goals), thereby providing a more balanced evaluation of the model's predictive capabilities.

Oversampling with SMOTE

The second technique we adopted was oversampling the training data using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an advanced oversampling technique that generates synthetic samples of the minority class (goals) by interpolating between existing minority class instances (Chawla et al. 2002). For example, if two shots that resulted in goals are close to each other in the feature space, SMOTE will create a new synthetic shot by randomly interpolating between these two shots. This process increases the number of goal instances in the training set, thereby balancing the class distribution and allowing the model to learn more effectively from both classes.

4.1.2 Standardization

After splitting the data, a standardization step was applied to ensure that all features were on the same scale. Standardization is particularly important for features like distance and angle, which may have different units and ranges. Without standardization, features with larger numerical ranges could disproportionately influence the model, leading to poor performance.

In this study, we utilized the `StandardScaler` from the `sklearn` package to perform standardization. The `StandardScaler` scales features so that they have the properties of a Gaussian distribution with a mean of 0 and a standard deviation of 1. It follows the formula:

$$z = \frac{x - \mu}{\sigma}$$

Where:

- z: The standardized value, represents the number of standard deviations x is away from the mean μ.
- x: The original value from the dataset.
- μ: The mean of the dataset.
- σ: The standard deviation of the dataset.

4.1.3 Model Learning

For the model learning process, we trained the Expected Goals model using the Logistic Regression algorithm. Article by Robberechts and Davis (2020a) has shown that a Logistic Regression model, even with a simple feature set like ours, only has a slightly lower performance than more complex models such as Gradient-Boosted Trees. This makes logistic regression a suitable choice for our Expected Goals model, balancing simplicity and performance.

We conducted two main experiments corresponding to the two sampling techniques:

Experiment 1 - Stratified Split without resampling

In the first experiment, we used the original dataset split according to the stratified strategy. This allowed us to maintain the class distribution while training the model, ensuring that the model could generalize effectively across both majority and minority classes.

Experiment 2 - Stratified Split and SMOTE

In the second experiment, we employed SMOTE to oversample the training data, thereby balancing the classes. This approach aimed to improve the model's ability to detect and predict goals, which are the minority class in this scenario.

In both experiments, we applied a grid search technique to fine-tune the C hyperparameter of the logistic regression model. The C parameter represents the cost of misclassification on the training data. A low C value results in a smoother decision boundary, which may lead to underfitting, while a high C value allows the model more flexibility to classify all training examples correctly, which could lead to overfitting. Therefore, it was crucial to identify the most appropriate value for C to ensure the model achieves the best balance between bias and variance.

4.2 Match Result Predictive Models

In this part, we focused on the multiclass prediction of football match outcomes, where the target variable could take one of three possible values: Home Team Win, Away Team Win, or Draw (abbreviated as Home, Away, Draw) (Figure 4.2). The objective was to accurately predict these outcomes by leveraging advanced machine learning techniques. We further explored the integration of Expected Goals (xG) values into our predictive models by incorporating the xG values generated by our pre-built Expected Goals model and comparing them with xG values obtained from Understat.

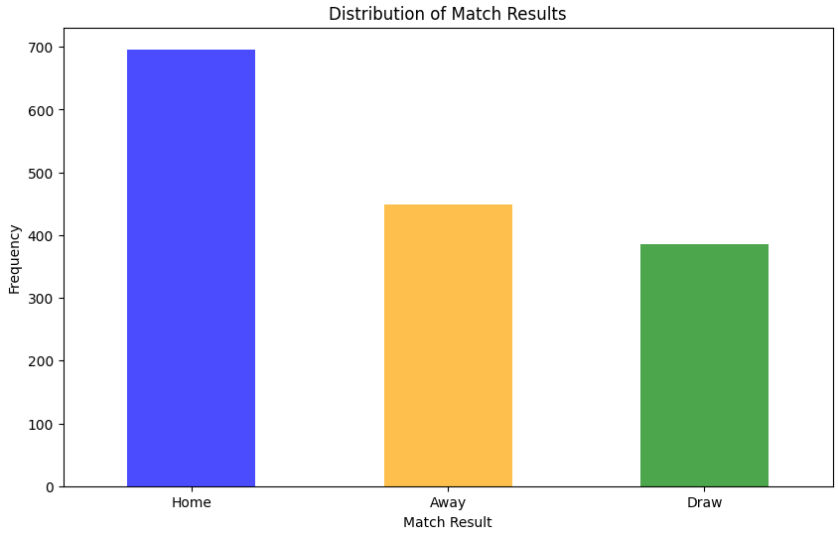

Figure 4.2: Number of matches by match result.

4.2.1 Model Selection

In this study, we compared the performance of models trained on datasets obtained. We will develop the predictive models with four common approaches, which are Logistic Regression, Support Vector Machines, Random Forest, and Neural Networks, and investigate the implementation of Expected Goals in these models to have a comprehensive view of the impact of this measure.

Logistic Regression

Logistic Regression is a supervised learning algorithm designed to solve classification problems and is commonly used for binary classification tasks. It models the relationship between the input features and the probability of a data instance belonging to a particular class. Similar to Linear Regression, Logistic Regression employs a linear combination of features but applies a sigmoid function to convert the output into a probability between 0 and 1. This probability represents the likelihood of an event belonging to a specific categorical outcome, such as Home, Away, or Draw, as a result of a football match. While primarily used for binary classification, logistic regression can be extended to handle multi-class problems through the "one-vs-rest" approach. This involves training multiple binary classifiers, each distinguishing one class from the rest. The class with the highest predicted probability is ultimately assigned to the data instance. As discussed in section 4.1.3, our models have been through a hyper-parameter tuning step to identify the optimal C, which balances the trade-off between minimizing training error and achieving good generalization on unseen data.

Support Vector Machines

Support Vector Machines (SVMs) are a powerful and flexible class of supervised learning algorithms primarily used for classification tasks but also adaptable for regression problems. SVMs work by identifying the hyperplane that best separates classes in the feature space, maximizing the margin between the closest data points of each class, known as support vectors. This method is particularly effective in high-dimensional spaces, where it can create complex decision boundaries.

In multi-class classification scenarios like predicting match outcomes (Home, Away, Draw), SVMs typically employ the "one-vs-one" or "one-vs-rest" approach, breaking down the multi-class problem into several binary classification tasks. For non-linearly separable data, SVMs can apply the kernel trick, mapping the input features into a higher-dimensional space where a linear separator can be found. Common kernels include the radial basis function (RBF) and polynomial kernels. In this study, our SVM classifier (SVC) utilized the default RBF kernel, with key hyper-parameters C (regularization) and *Gamma* (influence of a single training example) tuned to optimize model performance.

Random Forest

Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees to improve overall model accuracy and generalizability. By constructing numerous trees on different subsets of the training data, Random Forest mitigates the overfitting risk inherent in individual decision trees. Each tree in the forest produces a classification or regression output, and the final prediction is typically made by the majority vote for classification tasks or by averaging the outputs for regression tasks.

Random Forest's ability to handle complex decision boundaries and its robustness against noise makes it a popular choice for various machine learning applications, including football match outcome prediction. In this study, our Random Forest models were optimized by tuning two critical hyper-parameters: *n_estimators* (the number of trees in the forest) and *max_depth* (the maximum depth of each tree). These adjustments helped balance the model's bias-variance trade-off, ensuring accurate predictions while avoiding overfitting.

Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the human brain's network of neurons. They consist of layers of interconnected nodes, or neurons, where each connection has an associated weight. These weights are continuously adjusted during the training phase to minimize the error in predictions. Activation functions are then applied to transform the outputs of each layer, allowing ANNs to handle complex tasks due to their ability to handle non-linear relationships between input data and the desired output.

In this study, we employed a feedforward neural network architecture with fully connected layers. The input layer receives the feature vector for each match, and subsequent hidden layers transform this input through a series of weighted connections and activation functions. The output layer, which uses a SoftMax activation function, generates a probability distribution over the possible match outcomes (Home, Away, Draw). To enhance the network's performance and generalization capability, we implemented early stopping and other regularization techniques. Early stopping monitors the validation loss during training and halts the process when no further improvement is observed, thus preventing overfitting.

4.2.2 Experiment 1 - Using Aggregated Expected Goals

In order to integrate Expected Goals into match result predictive models an investigate its impact, in the first approach, we calculated the match xG values by aggregating predicted xG values for shots in the match and treated them as other match statistics. The aggregated xG represents the total number of goals a team was expected to score based on the quality of all the chances they created during a match.

For the data preparation, the dataset was split into training and testing sets with a standard 70/30 proportion. We also applied standardization using the `StandardScaler`, as detailed earlier in Section 4.1. This process converted all features to a standard normal distribution, ensuring that no single feature disproportionately influenced the model due to differing scales.

Since all the features are match statistics and quite straightforward, the following learning algorithms were chosen to predict the match results due to their ease of interpretation: Logistic Regression, Support Vector Machines (SVMs), and Random Forest. These algorithms are not only well-suited to handle the structure and relationship of the data effectively but also offer clear insights into how they made decisions by providing probabilities. The ability to analyse how each feature contributes to the predictions makes them ideal for understanding the effect of match statistics on the outcome.

To investigate the impact of xG values, the experiment was exploited with three scenarios: models without Expected Goals, models replacing Expected Goals for Goals, and models using both Goals and Expected Goals.

As shown in Figure 4.2, we observed a class imbalance in the target variable, with certain outcomes (e.g., draws) being less frequent. Therefore, we also examined the use of the SMOTE resampling technique in the models.

4.2.3 Experiment 2 - Using Individual Expected Goals

The second approach aimed to leverage the individual Expected Goals (xG) values along with their corresponding timestamps within a match. Unlike the aggregated xG values,

this approach retains the temporal sequence of shots, providing a richer representation of a team's offensive activity throughout a match.

This Expected Goals representation was constructed from the individual xG values of shots taken in the last six games, along with the time each shot was taken during those matches. This resulted in a two-dimensional vector for each match where each shot in a vector was represented by its xG value and the time in the match. One challenge with this approach was the variability in the number of shots across different teams and matches. To ensure consistent input lengths for the model, we employed zero-padding, a technique commonly used in neural networks. Zero-padding fills the shorter shot sequences with zeros, ensuring that all input vectors have the same length. The max length of the vectors was set at 20 by observing the distribution of the number of shots over the past 6 games, which means we only investigated the teams' last 20 shots.

To capture the sequential nature of the shot data, we implemented Long Short-Term Memory (LSTM) layers, a type of recurrent neural network (RNN) well-suited for processing and learning from sequences. We used separate LSTM layers for the home and away teams to model the distinct dynamics of each team. The LSTM outputs were then combined with other match statistics and passed through dense layers in a feedforward neural network to predict the match outcome. Figure 4.3 illustrates the overall architecture of the ANN model used in Experiment 2 in both scenarios of incorporating xG values and not.

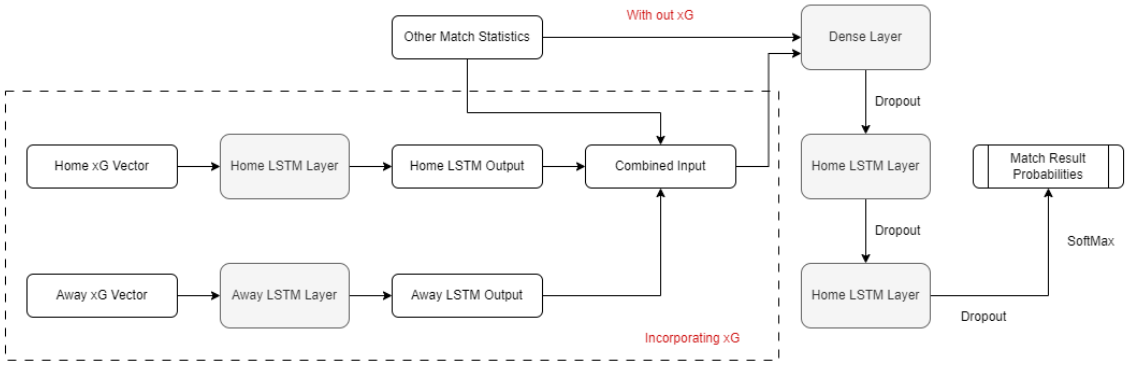

Figure 4.3: Overall architecture of ANN model with LSTM layers.

To compare with other models in Experiment 1, the dataset was also split into training and testing sets with the same proportion of 70/30, and other match statistics were applied standardization.

With the ability to handle the non-linearity, neural networks can capture complex relationships between the input and targeted output. However, this also presents overfitting as a common issue, where the model becomes too closely fitted to the training data and performs poorly on unseen data. To combat this, we employed Dropout, a regularization technique where fractions of the network's neurons are randomly dropped during training. This prevents the model from overly relying on specific paths through the network, encouraging generalization. Another technique that was applied was Early Stopping, which monitors the model's learning phase using the loss of the validation dataset. The validation set was extracted as 20% of the training set in this experiment.

Chapter 5
Results

In this section, we present the outcomes of our comprehensive analysis, focusing on the performance of the Expected Goals model and the subsequent estimation of match outcome probabilities. Various metrics were used to evaluate the models. Additionally, a feature importance analysis was conducted to gain insights into the variables that significantly influence the model's decision-making process.

5.1 Evaluation Metrics

To evaluate the proposed models, we employed multiple performance metrics, each offering valuable insights into model performance. When dealing with a classification problem, common metrics include accuracy, precision, recall, and F1 Score.

- **Accuracy**: This metric measures the proportion of correctly classified instances out of the total number of instances in a dataset. It provides a straightforward assessment of a classifier's overall performance but can be misleading when dealing with imbalanced datasets.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: Precision is a metric that quantifies the accuracy of positive predictions made by a model. It is the ratio of true positive predictions to the total number of positive predictions made by the model.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**: Recall measures the ability of a model to correctly identify all relevant instances within a dataset. It is the ratio of true positive predictions to the total number of actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score**: The F1 score combines precision and recall into a single metric, balancing the trade-off between these two metrics.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC AUC**: This is a common metric used to evaluate the performance of a binary classification task. The ROC AUC measures the trade-off between the true positive rate and the false positive rate at different classification thresholds by calculating the area under the ROC curve. A high ROC AUC score indicates strong model performance in distinguishing between classes. In multiclass classification, the "one-vs-rest" strategy is applied.

Given that the proposed models are probabilistic, it is crucial to assess them not only by their classification accuracy but also by the quality of the predicted probabilities. This is particularly important in scenarios with imbalanced classes or when the targets are probabilities, as is the case with the Expected Goals model. The appropriate methods in this case are Logarithmic Loss (Log Loss), Brier Score, and Calibration Curve.

- **Logarithmic Loss (Log Loss)**: This metric evaluates the accuracy of probabilistic predictions by penalizing confident but incorrect predictions more heavily. Log Loss is the cross-entropy between the predicted probability distribution and the true labels, making it useful for handling uncertainty in predictions.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log p_{ij}$$

  Where:

  - N: The total number of instances (or samples) in the dataset.
  - M: The total number of possible classes.
  - y_{ij}: The indicator (or binary) variable that is 1 if the actual outcome of instance i belongs to class j (e.g., the actual outcome is a Home Win) and 0 otherwise.
  - p_{ij}: The predicted probability that instance i belongs to class j.

- **Brier Score**: The Brier score measures the accuracy of probabilistic predictions by calculating the mean squared difference between predicted probabilities and actual outcomes. A lower Brier score indicates better calibration and sharpness of predictions. For multiclass problems, each possible category is treated as a binary outcome.

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(p_{ij} - y_{if})^2$$

  Where:

  - N: The total number of instances (or samples) in the dataset.
  - M: The total number of possible classes.
  - y_{ij}: The indicator (or binary) variable that is 1 if the actual outcome of instance i belongs to class j (e.g., the actual outcome is a Home Win) and 0 otherwise.
  - p_{ij}: The predicted probability that instance i belongs to class j.

- **Calibration Curve**: This visualization assesses how well-calibrated a model's predicted probabilities are. It plots the predicted probability for a bin of instances against the observed frequency, providing insights into whether the model tends to overestimate or underestimate probabilities.

5.2 Expected Goals Model

Since the primary objective of an Expected Goals model is to produce well-calibrated probability estimates rather than focusing on predicting the outcome of individual shots,

the model should be evaluated based on how accurate its probabilistic predictions are to the real-world probabilities. Therefore, it is important to look at the metrics that reflect how well-calibrated the models are, such as Log Loss and Brier Score.

Table 5.1 compares the performance of our Expected goal models with models from previous research. In addition to Log Loss and Brier Score, due to the lack of these measurements in some past works, we also included standard metrics like F1 Score, ROC AUC, Precision, and Recall for a comprehensive evaluation.

Table 5.1: Performances of our Expected Goals models and those from previous studies.

Study	Model	F1 Score	ROC AUC	Log-loss	Brier score
Robberechts and Davis (2020a)	Logistic Regression (basic features)	-	0.76	-	0.0812
	XGBoost (basic features)	-	0.76	-	0.0806
	Logistic Regression (advanced features)	-	0.77	-	**0.0783**
	XGBoost (advanced features)	-	0.77	-	**0.0783**
Van Haaren (2021)	Boosting Machine	-	0.7932	-	0.0823
Mead, O'Hare, and McMenemy (2023)	XGBoost	-	-	**0.2818**	0.0799
Eggels, Elk, and Pechenizkiy (2016)	Random Forest	0.800	0.814	-	-
	Decision Tree	0.676	0.677	-	-
	Logistic Regression	0.673	0.697	-	-
	Ada-boost	0.688	0.670	-	-
Anzer and Bauer (2021)	Gradient Boosting	-	**0.822**	-	-
	Logistic Regression	-	0.807	-	-
	Ada-boost	-	0.816	-	-
	Random Forest	-	0.794	-	-
Our study	Logistic Regression	**0.8742**	0.7905	0.2852	0.0812
	Logistic Regression (with SMOTE)	0.7811	0.7903	0.5485	0.1719

In terms of evaluation, we prioritized the use of Brier scores over AUROC due to the nature of Expected Goals purposes, which rely on accurate probability predictions rather than just classification accuracy. Compared to models from previous work, our Logistic Regression model performed similarly, with a Brier score of 0.0812 and a Log Loss of 0.2852. While this result is not the best in comparison to those from previous research, it is important to note that the mentioned studies often utilized more advanced features,

such as positional data and shot-related events, which were not available in our dataset. The calibration curve from Figure 5.1 also demonstrated that our Expected Goals model is quite well-calibrated. Additionally, in terms of average Precision, Recall, and F1 Score, our model outperformed those in previous studies.

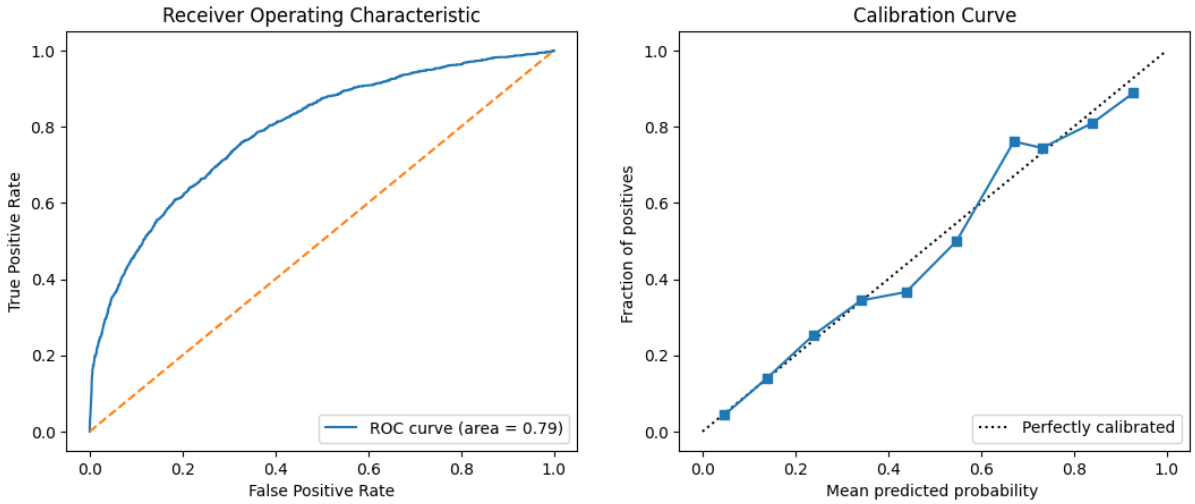

Figure 5.1: ROC Curve and Calibration Curve of our Logistic Regression model for xG.

More detailed of the model's estimation, the comparison of heatmaps between the predicted xG and the dataset's true probability of scoring (Figure 5.2) revealed that the predicted xG heatmap effectively represented the chances, with high-probability areas aligning well with observed scoring patterns. Notably, when calculating the match xG values by aggregating the xG predictions of this chosen model as suggested in previous work (Eggels, Elk, and Pechenizkiy 2016), the cumulative xG values presented a strong correlation with Understat's xG values (Figure 5.3). This high correlation underscores the model's effectiveness in producing reliable xG values.

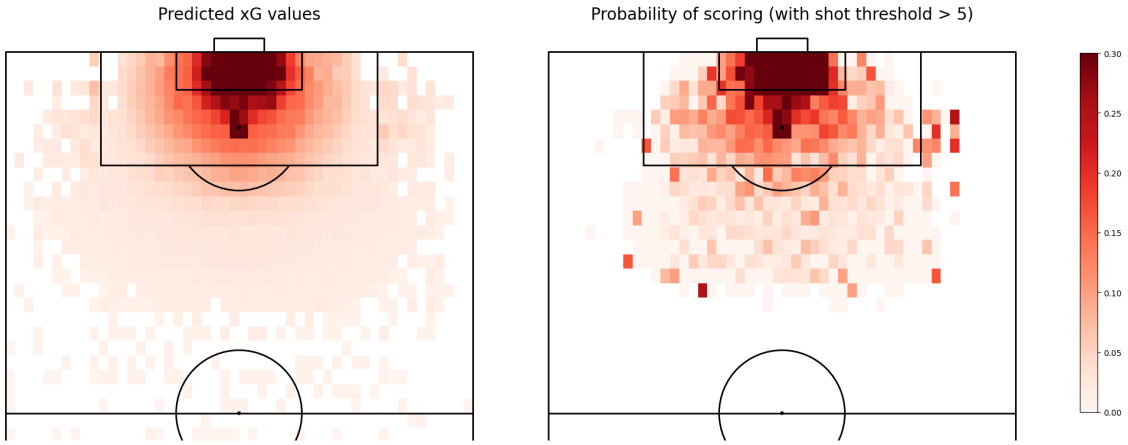

Figure 5.2: Heat maps of Predicted xG values and Probabilities of scoring from different areas.

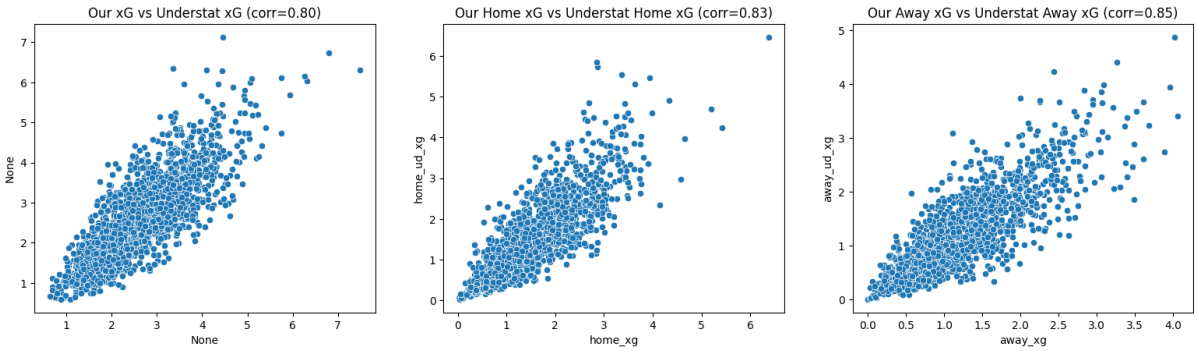

Figure 5.3: Correlation between aggregated predicted xG values and xG values from Understat.

In addressing the class imbalance, the SMOTE resampling technique showed potential applications. Although this approach led to a decrease in accuracy and precision, it improved recall for the goal class. However, this resulted in a lower F1 Score and worse probabilistic predictions, as indicated by higher Log Loss and Brier scores. Based on these evaluation metrics, the Logistic Regression model without SMOTE was chosen as the Expected Goals model for generating xG values to be used in the match results predictive model.

Using the SHAP (SHapley Additive exPlanations) library, we visualized feature importance in our model. As expected, the most important feature was the shot's distance, indicating that shots taken from closer distances are more likely to result in goals (Figure 5.4). The next most important features were the body part used to take the shot and the shot's visible angle.

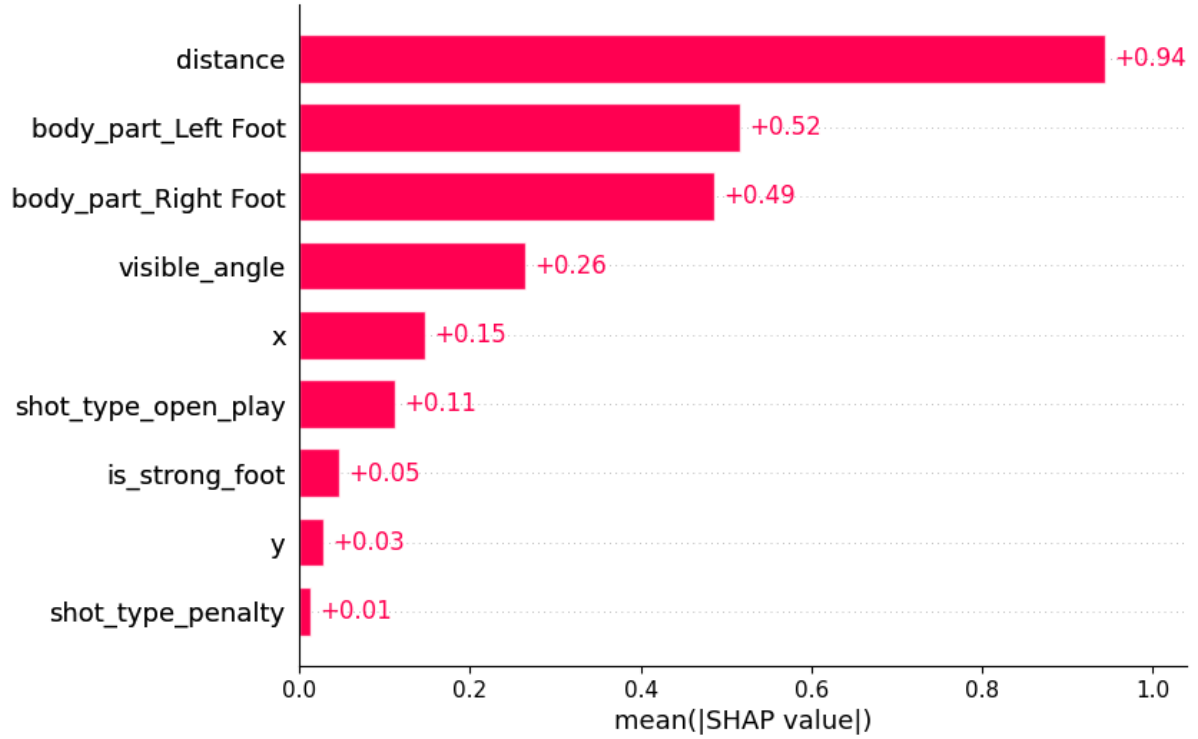

Figure 5.4: Feature importance of our Expected Goals model.

5.3 Match Results Predictive Models

5.3.1 Baseline Prediction

To effectively assess the performance of our models, we first established a baseline for comparison. This baseline serves as a point of reference, allowing us to assess the performance of our predictive models against a standard. A particularly insightful baseline involves examining the odds provided by bookmakers. Bookmakers offer the opportunity for consumers to bet on football matches, and their odds are based on internal predictions of match outcomes. In this study, we utilized the fixed odds from **Bet365** as a representation of the bookmaker's predictions.

To interpret these odds, we need to calculate the implied probability for each possible match result. A commonly used formula is:

$$p_i = \frac{1}{o_i}$$

Where p_i is the probability of result i (Home Win, Away Win, or Draw), and o_i is the odds for that result. However, as noted by Štrumbelj (2014), bookmakers typically include a margin to ensure profitability, meaning the sum of the inverse odds (often referred to as the "booksum") is always greater than 1. The bookmaker margin can be calculated using the following equation:

$$\text{Bookmaker Margin} = \frac{1}{o_h^{-1} + o_d^{-1} + o_a^{-1}}$$

Where $o_h^{-1}, o_d^{-1}, o_a^{-1}$ are odds for the results of Home Win, Draw, and Away Win respectively. Therefore, accurate probabilistic predictions from bookmakers are then derived by adjusting for this margin:

$$p_i = \frac{o_i^{-1}}{o_h^{-1} + o_d^{-1} + o_a^{-1}}$$

5.3.2 Model Performance

As detailed in section 4.2, several classifiers were employed to predict the outcomes of football matches. Our primary goal was to investigate the impact of Expected Goals on predictive models and to build the best model compared with the baseline in terms of predicting match results. Since the baseline estimated the probability of each class, we mainly assess our models by their probabilistic predictions. Other evaluation metrics based on the classification are still reported for the investigation of Expected Goals.

Classification Ability

First, in terms of accuracy, our models demonstrated promising results and accurately predicted the outcomes of approximately 54% of matches. The accuracy scores of all models are presented in Figure 5.5. Although direct comparisons with other studies are challenging due to variations in datasets, our results are notable when compared to previous work, such as the study by Joseph, N. E. Fenton, and Neil (2006), where our models outperformed many of the proposed models in the literature.

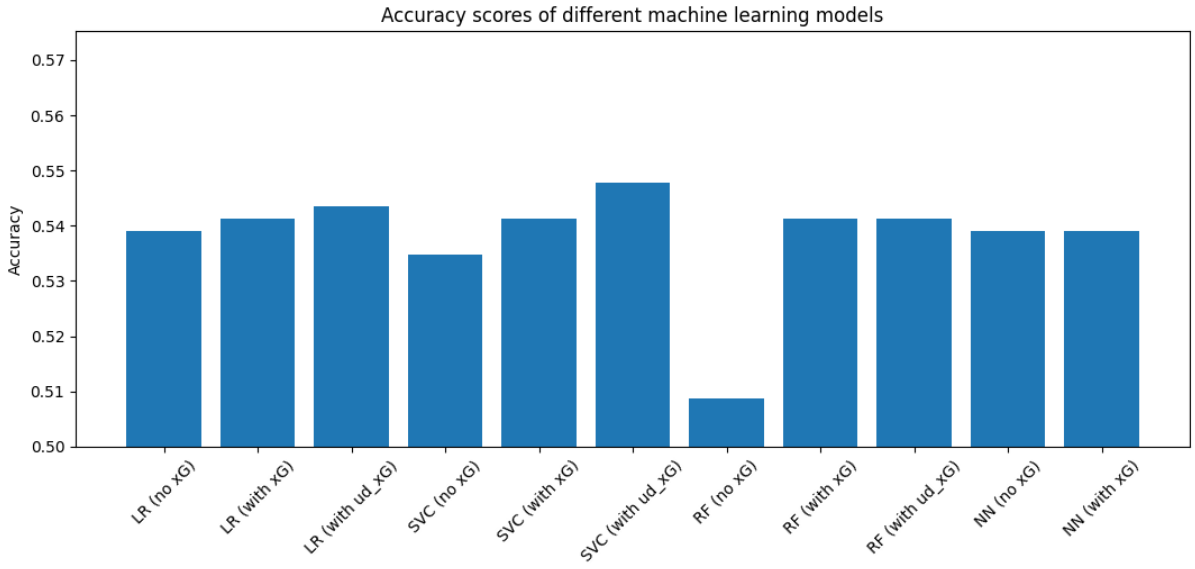

Figure 5.5: Accuracy scores of the different match result predictive models.

Among our experiments, the most accurate model was the Support Vector Machines (SVMs), particularly when incorporating xG values from Understat (ud_xG) with the highest accuracy of 54.78%. When using xG values predicted by our Expected Goals model (as described in Section 5.1), both the Logistic Regression and SVC algorithms demonstrated slight improvements in accuracy. However, the most notable increase in accuracy was observed in the Random Forest (RF) model, where accuracy improved significantly from 50.87% to 54.13% when utilizing either xG values generated by the pre-built Expected Goals model or those from Understat. In contrast, the Artificial Neural Network (ANN) model did not show any improvement when incorporating individual xG values for shots, contrary to the expectations that this approach would capture the relationship between individual xG values and lead to substantial enhancements.

To further investigate model performance, we visualized the classifiers' effectiveness using Receiver Operating Characteristic (ROC) curves. Interestingly, integrating xG values did not have a marked effect on the ROC-AUC scores for most models, contrary to what might have been anticipated. The ROC curves (Figures A1, A2, A3, A4) for the Logistic Regression and SVC models showed minimal changes with the inclusion of xG values, suggesting that these metrics did not significantly influence the models' ability to distinguish between classes. The Random Forest model exhibited a slightly improved ROC curve for the Draw class, particularly at higher probability thresholds. However, this improvement is unlikely to have a significant impact due to the infrequent occurrence of draws, making it less critical in the overall performance of the model. Detailed results, including confusion matrices, precision, and recall metrics from the best model of each algorithm, are provided in Tables 5.2, 5.3, 5.4, and 5.5.

As anticipated, the models demonstrated strong performance in predicting home wins, with the Logistic Regression model, for instance, achieving precision and recall scores of 0.55 and 0.85, respectively. However, predicting draws presented a significant challenge, a difficulty that aligns with findings from previous studies. As discussed in section 2.1, predicting draws is inherently challenging for machine learning models, primarily because draws are the least likely outcome in football match results. This difficulty was evident in our experiments, where the SVMs models failed to predict any draws and even the model

Table 5.2: Logistic Regression model (with Understat xG)

(a) Confusion matrix			
	Predicted Home	Predicted Draw	Predicted Away
Actual Home	178	2	29
Actual Draw	83	1	32
Actual Away	63	1	71
(b) Precision–recall table			
	Precision	Recall	F1 Score
class Home	0.55	0.85	0.67
class Draw	0.25	0.01	0.02
class Away	0.54	0.53	0.53

Table 5.3: SVMs model (with Understat xG)

(a) Confusion matrix			
	Predicted Home	Predicted Draw	Predicted Away
Actual Home	198	0	11
Actual Draw	92	0	24
Actual Away	81	0	54
(b) Precision–recall table			
	Precision	Recall	F1 Score
class Home	0.53	0.95	0.68
class Draw	0.00	0.00	0.00
class Away	0.61	0.40	0.48

with the highest precision for this class—the Random Forest model—only managed a recall of 0.03. These results mirror the challenges highlighted in earlier research, underscoring the complexity of accurately forecasting football match outcomes, particularly for the Draw class. The consistently poor performance across different models in predicting draws reaffirms the inherent unpredictability and difficulty in modelling such outcomes.

Probabilistic Estimation

In football match predictions, accurately predicting the match outcome is crucial, but it is also important to provide the associated probabilities for each possible result. This becomes especially significant when comparing our model's predictions with those of bookmakers, as these probabilities offer deeper insights into the confidence and reliability of the predictions. To evaluate these probabilistic predictions, we reported metrics such as Log Loss and Brier Score in Table 5.6, alongside the ROC AUC score, which was calculated using the `sklearn` library.

Our findings reveal that the incorporation of Expected Goals (xG) values enhanced

Table 5.4: Random Forest model (with Understat xG)

(a) Confusion matrix			
	Predicted Home	Predicted Draw	Predicted Away
Actual Home	183	5	21
Actual Draw	85	3	28
Actual Away	66	6	63
(b) Precision–recall table			
	Precision	Recall	F1 Score
class Home	0.55	0.88	0.67
class Draw	0.21	0.03	0.05
class Away	0.56	0.47	0.51

the probabilistic predictions of most of our models (except the case of the SVMs model). Despite these improvements, the baseline—derived from bookmakers' odds—still delivered the most well-calibrated predictions, outperforming all of our models. This suggests that bookmakers have a superior estimation process, one that is more closely aligned with the actual probabilities of match outcomes. Notably, our Logistic Regression model performed quite closely to the baseline, indicating that it is a strong candidate in terms of calibration. Visual assessment of the Calibration Curves for each match result prediction confirms that the baseline model had the most well-calibrated curves. Among our proposed models, the Logistic Regression model was the most reasonably calibrated, while the most accurate model, SVMs, had the worst calibration curves (Figure A5, A6, A7). Figure 5.6 shows how well-calibrated the Logistic Regression is in comparison to the baseline.

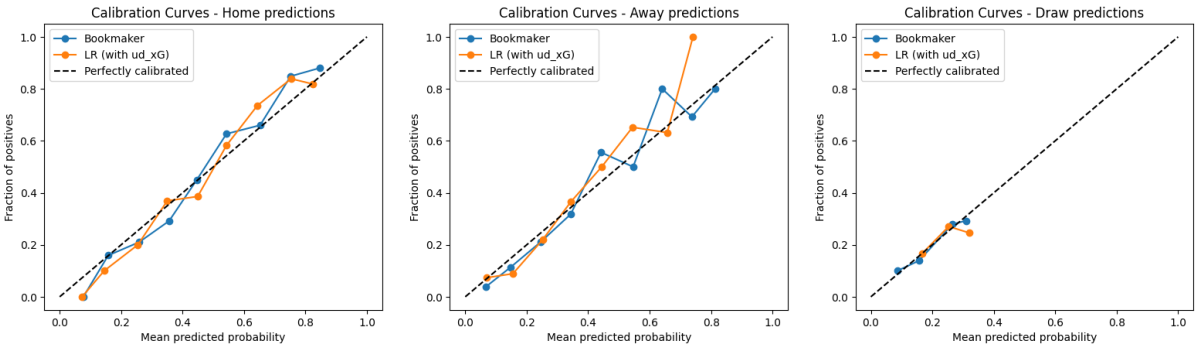

Figure 5.6: Calibration curves of Logistic Regression (with *ud_xG*) model and baseline in comparison.

SMOTE Application

When exploring the application of the SMOTE resampling technique to address the class imbalance, we observed that the Logistic Regression (with ud_xG) model using SMOTE achieved a slight improvement in predicting draws, raising the precision to 0.28 and recall to 0.06 for the Draw class. This indicates that the model became more sensitive

Table 5.5: ANN model (with xG)*

(a) Confusion matrix			
	Predicted Home	Predicted Draw	Predicted Away
Actual Home	176	1	32
Actual Draw	86	0	30
Actual Away	63	0	72
(b) Precision–recall table			
	Precision	Recall	F1 Score
class Home	0.54	0.53	0.54
class Draw	0.00	0.00	0.00
class Away	0.54	0.84	0.66

Note: This is the ANN model incorporating individual xG values of the last 20 shots for each team, as described in section 4.2.3.

to predicting draws, resulting in fewer false negatives. However, these gains were marginal and came at the expense of significantly poorer calibration, with a Log Loss of 1.0926 and a Brier Score of 0.6626.

Interestingly, the SVMs model, when combined with SMOTE, demonstrated a notable improvement in capturing draw matches, achieving a recall of 0.17. Nevertheless, this improvement was accompanied by a substantial decline in both classification performance and probabilistic prediction accuracy. The trade-off between enhancing the prediction for the Draw class and maintaining overall model performance underscores the challenges of applying resampling techniques like SMOTE in football match outcome prediction, particularly when dealing with inherently imbalanced classes like draws.

Considering all the evaluation and analysis, the Logistic Regression model incorporating Understat's xG values emerged as the best-performing model among the models developed in this study. This model not only demonstrated strong calibration but also offered competitive accuracy and reliability in its probabilistic predictions, making it the most effective choice.

5.3.3 Feature Analysis

To gain deeper insights into the influence of the Expected Goals (xG) feature, we conducted a feature analysis using the same methodology as outlined in Section 5.2.

Figures A8, A9, and A10 illustrate the importance of the feature when the Logistic Regression model (with *ud_xG*) predicted home wins, away wins, and draws, respectively. Interestingly, despite the slight performance enhancements observed with the inclusion of xG values, the *avg_home_ud_xg* and *avg_away_ud_xg* features ranked among the least important in this model. This finding clarifies why the improvements were not as substantial as initially anticipated; the xG features did not contribute significantly to the model's decision-making process.

In contrast, the Random Forest model, which experienced the most notable performance boost with the integration of xG values, highlighted the Understat's xG values features as

Table 5.6: Evaluation of probabilistic predictions from different models.

Model	ROC AUC	Log Loss	Brier score
Baseline	**0.7089**	**0.9237**	**0.5459**
Logistic Regression (without xG)	0.6829	0.9575	0.5677
Logistic Regression (with xG)	0.6852	0.9559	0.5665
Logistic Regression (with ud_xG)	0.6857	0.9560	0.5666
SVMs (without xG)	0.6432	0.9905	0.5907
SVMs (with xG)	0.6454	0.9986	0.5959
SVMs (with ud_xG)	0.6356	1.000	0.5969
Random Forest (without xG)	0.6614	0.9699	0.5757
Random Forest (with xG)	0.6681	0.9617	0.5695
Random Forest (with ud_xG)	0.6681	0.9617	0.5695
ANN (without xG)	0.6716	0.9745	0.5791
ANN (with xG)	0.6703	0.9697	0.5773

two of the top five most influential factors (Figure A11, A12, A13). This suggests that the Random Forest model is more efficient at leveraging the information provided by xG metrics, effectively utilizing these features to enhance its prediction.

For both models investigated, the average number of passes over the past six matches and the goal difference (for both home and away teams) emerged as the most impactful factors influencing the model's decisions. The SHAP values associated with these features significantly outweighed those of other variables, underscoring their dominant role in the models' predictive processes. This finding reinforces the importance of considering not only advanced metrics like xG but also other performance indicators such as passing frequency and goal difference when predicting football match outcomes.

Chapter 6
Conclusion

6.1 Summary of Findings

This study aimed to explore the use of machine learning approaches, particularly the incorporation of Expected Goals (xG) metrics, in predicting football match results. Through a detailed analysis, we developed and evaluated several predictive models, comparing their performance with a baseline derived from bookmakers' odds.

Key findings

Our investigation revealed several key findings:

1. **Incorporation of Expected Goals**: Integrating Expected Goals into our predictive models provided a better understanding of team performance and improved the probabilistic accuracy of our predictions. While traditional metrics like goals scored were informative, Expected Goals offered a more detailed view of a team's performance efficiency. Feature analysis revealed that while xG values were useful, traditional metrics like the average number of passes and goal difference had more impact on the models' decision-making processes. This suggests that while Expected Goals is a valuable addition, it should be considered alongside other match statistics for optimal predictive accuracy.

2. **Comparison with Bookmakers' Predictions**: The baseline model, reflecting bookmakers' odds, consistently outperformed our machine learning models in terms of calibration. This underscores the effectiveness of bookmakers' estimation processes, which likely benefit from a combination of sophisticated statistical methods and expert domain knowledge

3. **Model Performance**: Among the models developed, the Logistic Regression model, particularly when utilizing xG values from understat.com, demonstrated the closest performance to the baseline, which was derived from bookmakers' odds. This suggests that, despite the complexity and uncertainty in football match outcomes, Logistic Regression remains a robust method for prediction, especially when enhanced with xG metric.

4. **Challenges with Draw Predictions**: Predicting draw outcomes proved challenging, aligning with the findings of previous research. Even with the application of advanced techniques like SMOTE for addressing class imbalance, the improvements in draw prediction were minimal, highlighting the difficulty of this task.

Implications for Football Analytics

The findings of this study contribute to the growing field of football analytics by demonstrating the practical utility of Expected Goals in match result predictive models. The results suggest that while Expected Goals potentially improves the detail of

performance analysis, traditional metrics continue to play a critical role. For football professionals, this means that a balanced approach - utilizing both xG and match statistics - may result in the most reliable predictions.

6.2 Limitations and Future Research

This study faced several limitations that offer opportunities for future research. First, our Expected Goals model was developed using relatively basic features. While it demonstrated reasonably good predictive performance, it is potentially improved by incorporating more advanced features, such as positional data, shot-related events, and player-specific metrics.

Secondly, the dataset was limited to the 2017/2018 season, which may not capture long-term trends and variations in team performance. Expanding the dataset to include multiple seasons could help the model produce more reliable predictions. Future research could also explore the integration of additional innovative metrics, such as Expected Assists or defensive metrics. Additionally, most of the features used in our predictive models were derived by averaging match statistics from the previous six games. While providing an informative representation of a team's recent form, this approach raised the limitation of the inability to predict the first six matches of any given season. This restricts the application of the models in the early stages of a season since it can only make predictions from the 7^{th} match onwards.

Bibliography

Anzer, Gabriel and Pascal Bauer (Mar. 29, 2021). "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. Publisher: Frontiers. ISSN: 2624-9367. DOI: 10.3389/fspor.2021.624475.

Baboota, Rahul and Harleen Kaur (Apr. 1, 2019). "Predictive analysis and modelling football results using machine learning approach for English Premier League". In: *International Journal of Forecasting* 35.2, pp. 741–755. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2018.01.003.

Berrar, Daniel, Philippe Lopes, and Werner Dubitzky (Jan. 1, 2019). "Incorporating domain knowledge in machine learning for soccer outcome prediction". In: *Machine Learning* 108.1, pp. 97–126. ISSN: 1573-0565. DOI: 10.1007/s10994-018-5747-8.

Brechot, Marc and Raphael Flepp (May 1, 2020). "Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals". In: *Journal of Sports Economics* 21.4. Publisher: SAGE Publications, pp. 335–362. ISSN: 1527-0025. DOI: 10.1177/1527002519897962.

Bunker, Rory P. and Fadi Thabtah (Jan. 1, 2019). "A machine learning framework for sport result prediction". In: *Applied Computing and Informatics* 15.1, pp. 27–33. ISSN: 2210-8327. DOI: 10.1016/j.aci.2017.09.005.

Cefis, Mattia (2024). "A PLS-SEM Approach for Composite Indicators: An Original Application on the Expected Goal Model". In: *Intelligent Technologies for Interactive Entertainment*. Ed. by Martin Clayton, Mauro Passacantando, and Marcello Sanguineti. Cham: Springer Nature Switzerland, pp. 127–135. ISBN: 978-3-031-55722-4. DOI: 10.1007/978-3-031-55722-4_10.

Chawla, N. V. et al. (June 1, 2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. ISSN: 1076-9757. DOI: 10.1613/jair.953.

Choi, Bing, Lee-Kien Foo, and Sook-Ling Chua (Dec. 20, 2023). "Predicting Football Match Outcomes with Machine Learning Approaches". In: *MENDEL* 29, pp. 229–236. DOI: 10.13164/mendel.2023.2.229.

Constantinou, Anthony Costa and Norman Elliott Fenton (Mar. 30, 2013). "Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries". In: *Journal of Quantitative Analysis in Sports* 9.1. Publisher: De Gruyter, pp. 37–50. ISSN: 1559-0410. DOI: 10.1515/jqas-2012-0036.

Danisik, Norbert, Peter Lacko, and Michal Farkas (Aug. 2018). "Football Match Prediction Using Players Attributes". In: *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), pp. 201–206. DOI: 10.1109/DISA.2018.8490613.

Eggels, H., R. van Elk, and M. Pechenizkiy (2016). "Explaining soccer match outcomes with goal scoring opportunities predictive analytics: 3rd Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA 2016)". In: *Proceedings of the Workshop on Machine Learning and Data Mining for Sports Analytics 2016 co-located with the 2016 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2016)*. CEUR Workshop Proceedings. Publisher: CEUR-WS.org.

Fan, Mu et al. (Oct. 1, 2023). "Determinants of international football performance: Empirical evidence from the 1994–2022 FIFA World Cup". In: *Heliyon* 9.10, e20252. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2023.e20252.

Hewitt, James H. and Oktay Karakuş (Sept. 1, 2023). "A machine learning approach for player and position adjusted expected goals in football (soccer)". In: *Franklin Open* 4, p. 100034. ISSN: 2773-1863. DOI: 10.1016/j.fraope.2023.100034.

Hubáček, Ondřej, Gustav Šourek, and Filip Železný (Jan. 1, 2019). "Learning to predict soccer results from relational data with gradient boosted trees". In: *Machine Learning* 108.1, pp. 29–47. ISSN: 1573-0565. DOI: 10.1007/s10994-018-5704-6.

Hvattum, Lars Magnus and Halvard Arntzen (July 1, 2010). "Using ELO ratings for match result prediction in association football". In: *International Journal of Forecasting*. Sports Forecasting 26.3, pp. 460–470. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2009.10.002.

Joseph, A., N. E. Fenton, and M. Neil (Nov. 1, 2006). "Predicting football results using Bayesian nets and other machine learning techniques". In: *Knowledge-Based Systems*. Creative Systems 19.7, pp. 544–553. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2006.04.011.

Macdonald, Brian (2012). "An Expected Goals Model for Evaluating NHL Teams and Players". In.

Mead, James, Anthony O'Hare, and Paul McMenemy (Apr. 5, 2023). "Expected goals in football: Improving model performance and demonstrating value". In: *PLOS ONE* 18.4. Publisher: Public Library of Science, e0282295. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0282295.

Müller, Oliver, Alexander Simons, and Markus Weinmann (Dec. 1, 2017). "Beyond crowd judgments: Data-driven estimation of market value in association football". In: *European Journal of Operational Research* 263.2, pp. 611–624. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2017.05.005.

Odachowski, Karol and Jacek Grekow (2013). "Using Bookmaker Odds to Predict the Final Result of Football Matches". In: *Knowledge Engineering, Machine Learning and Lattice Computing with Applications*. Ed. by Manuel Graña et al. Berlin, Heidelberg: Springer, pp. 196–205. ISBN: 978-3-642-37343-5. DOI: 10.1007/978-3-642-37343-5_20.

Pappalardo, Luca and Paolo Cintia (May 2018). "Quantifying the relation between performance and success in soccer". In: *Advances in Complex Systems* 21.3. Publisher: World Scientific Publishing Co., p. 1750014. ISSN: 0219-5259. DOI: 10.1142/S021952591750014X.

Pappalardo, Luca, Paolo Cintia, et al. (Oct. 28, 2019). "A public data set of spatio-temporal match events in soccer competitions". In: *Scientific Data* 6.1. Publisher: Nature Publishing Group, p. 236. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0247-7.

Partida, Adan et al. (Mar. 31, 2021). "Modeling of Football Match Outcomes with Expected Goals Statistic". In: *Journal of Student Research* 10.1. ISSN: 2167-1907. DOI: 10.47611/jsr.v10i1.1116.

Rathke, Alex (2017). "An examination of expected goals and shot efficiency in soccer". In: Accepted: 2017-08-29T07:06:55Z Publisher: Universidad de Alicante. Área de Educación Física y Deporte. ISSN: 1988-5202. DOI: 10.14198/jhse.2017.12.Proc2.05.

Raudonius, Laurynas and Thomas Seidl (2023). "Shot Analysis in Different Levels of German Football Using Expected Goals". In: *Machine Learning and Data Mining for Sports Analytics*. Ed. by Ulf Brefeld et al. Cham: Springer Nature Switzerland, pp. 14–26. ISBN: 978-3-031-27527-2. DOI: 10.1007/978-3-031-27527-2_2.

Robberechts, Pieter and Jesse Davis (2020a). "How Data Availability Affects the Ability to Learn Good xG Models". In: *Machine Learning and Data Mining for Sports Analytics*. Ed. by Ulf Brefeld et al. Cham: Springer International Publishing, pp. 17–27. ISBN: 978-3-030-64912-8. DOI: 10.1007/978-3-030-64912-8_2.

— (May 14, 2020b). *Illustrating the interplay between features and models in xG*. DTAI Sports. URL: https://dtai.cs.kuleuven.be/sports/blog/illustrating-the-interplay-between-features-and-models-in-xg (visited on 08/14/2024).

Rodrigues, Fátima and Ângelo Pinto (Jan. 1, 2022). "Prediction of football match results with Machine Learning". In: *Procedia Computer Science*. International Conference on Industry Sciences and Computer Science Innovation 204, pp. 463–470. ISSN: 1877-0509. DOI: 10.1016/j.procs.2022.08.057.

Schumaker, Robert P., A. Tomasz Jarmoszko, and Chester S. Labedz (Aug. 1, 2016). "Predicting wins and spread in the Premier League using a sentiment analysis of twitter". In: *Decision Support Systems* 88, pp. 76–84. ISSN: 0167-9236. DOI: 10.1016/j.dss.2016.05.010.

Shtovba, Serhiy, Georgios Dounias, and Athanasios Tsakonas (Jan. 1, 2003). "FORECASTING FOOTBALL MATCH OUTCOMES WITH SUPPORT VECTOR MACHINES". In: *Herald of Zhytomyr EngeneeringTechnological Institute*.

Štrumbelj, Erik (Oct. 1, 2014). "On determining probability forecasts from betting odds". In: *International Journal of Forecasting* 30.4, pp. 934–943. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2014.02.008.

Tureen, Tahmeed and S. B. H. Olthof (Sept. 20, 2022). ""Estimated Player Impact" (EPI): Quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models". In: *StatsBomb Conference Proceedings*. StatsBomb Conference 2022. Wembley, London: StatsBomb.

Van Haaren, Jan (May 27, 2021). *"Why Would I Trust Your Numbers?" On the Explainability of Expected Values in Soccer*. DOI: 10.48550/arXiv.2105.13778. arXiv: 2105.13778[cs,stat].

Wyscout Glossary (2024). URL: https://dataglossary.wyscout.com/ (visited on 08/14/2024).

Appendix

Predictive Models' ROC Curves

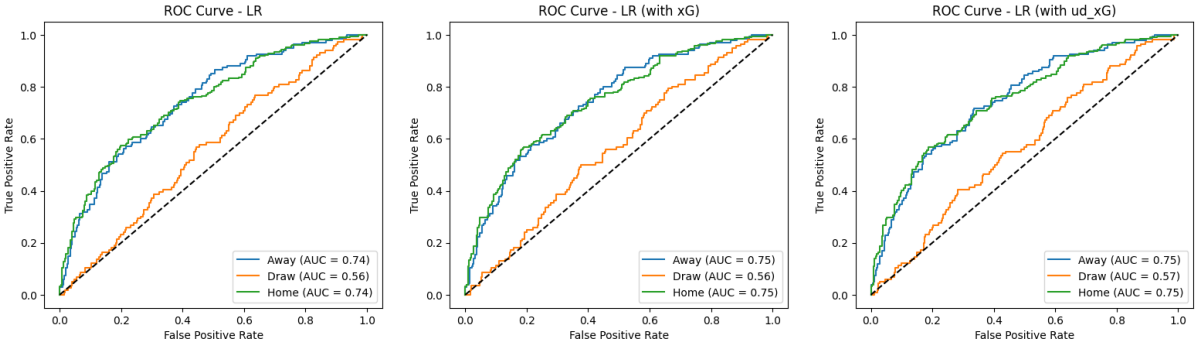

Figure A1: ROC curves of Logistic Regression model in three experimental scenarios.

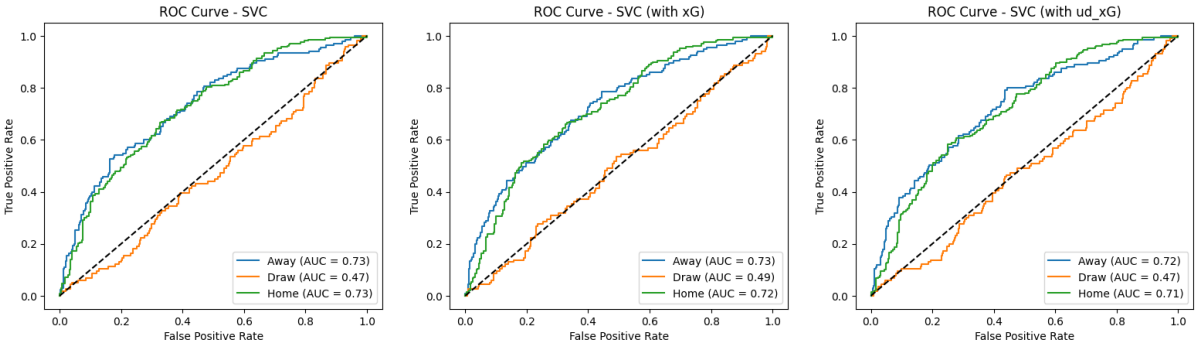

Figure A2: ROC curves of SVMs model in three experimental scenarios.

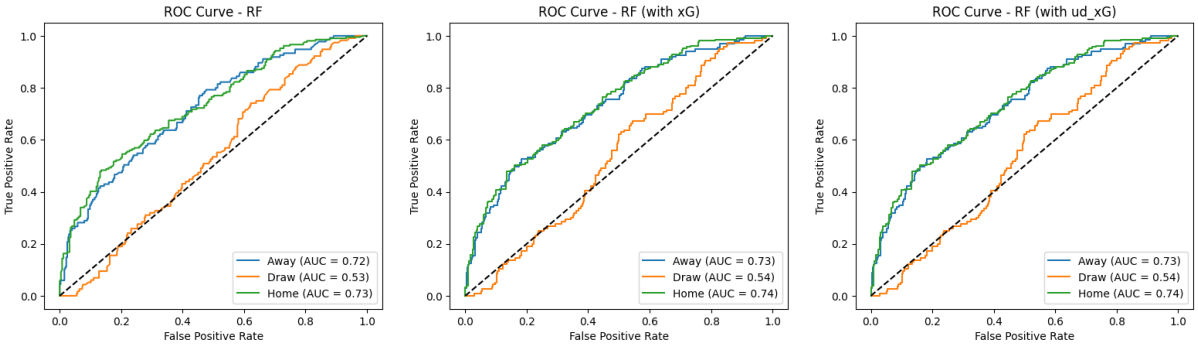

Figure A3: ROC curves of Random Forest model in three experimental scenarios.

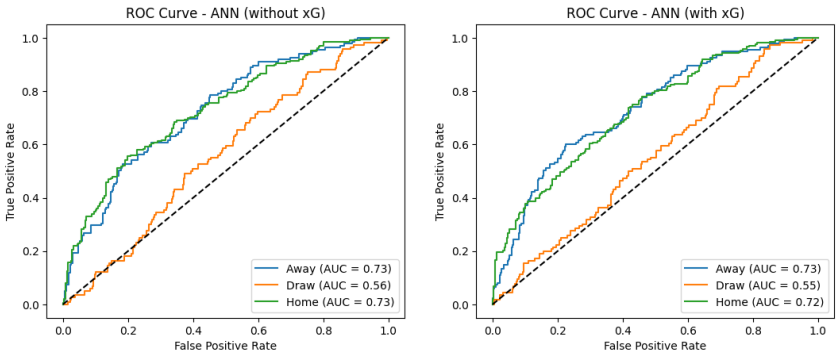

Figure A4: ROC curves of ANN model in two experimental scenarios.

Predictive Models' Calibration Curves

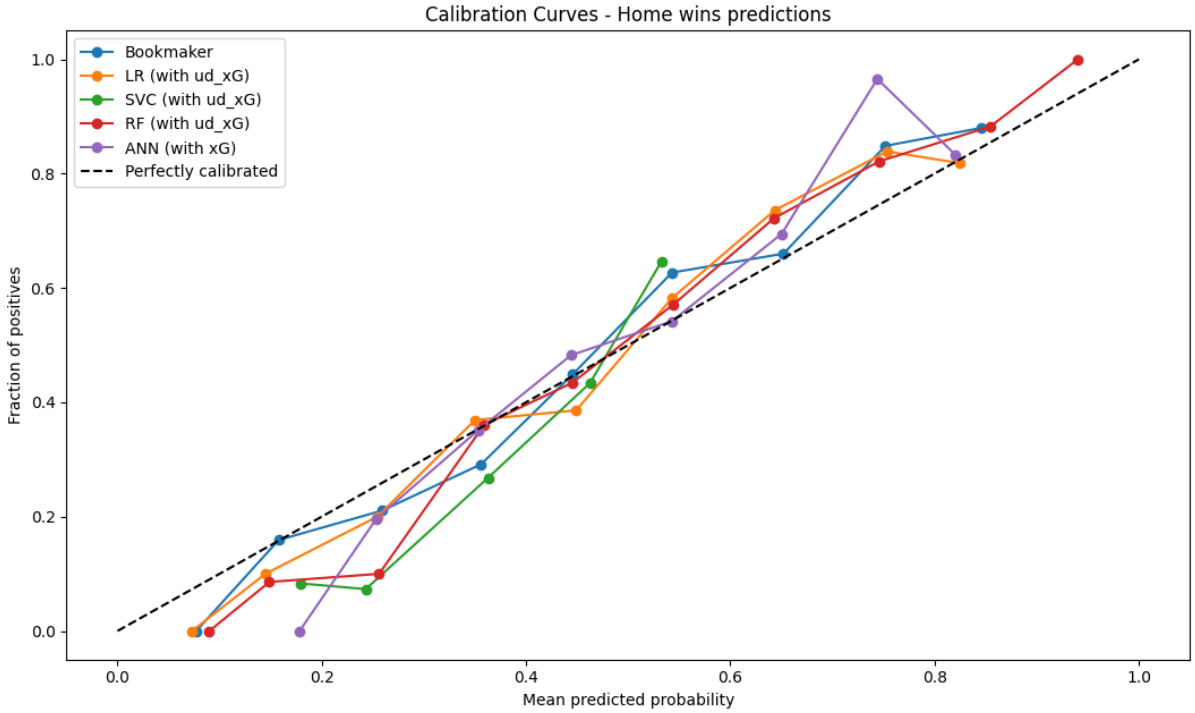

Figure A5: Calibration curves of different models when predicting Home wins.

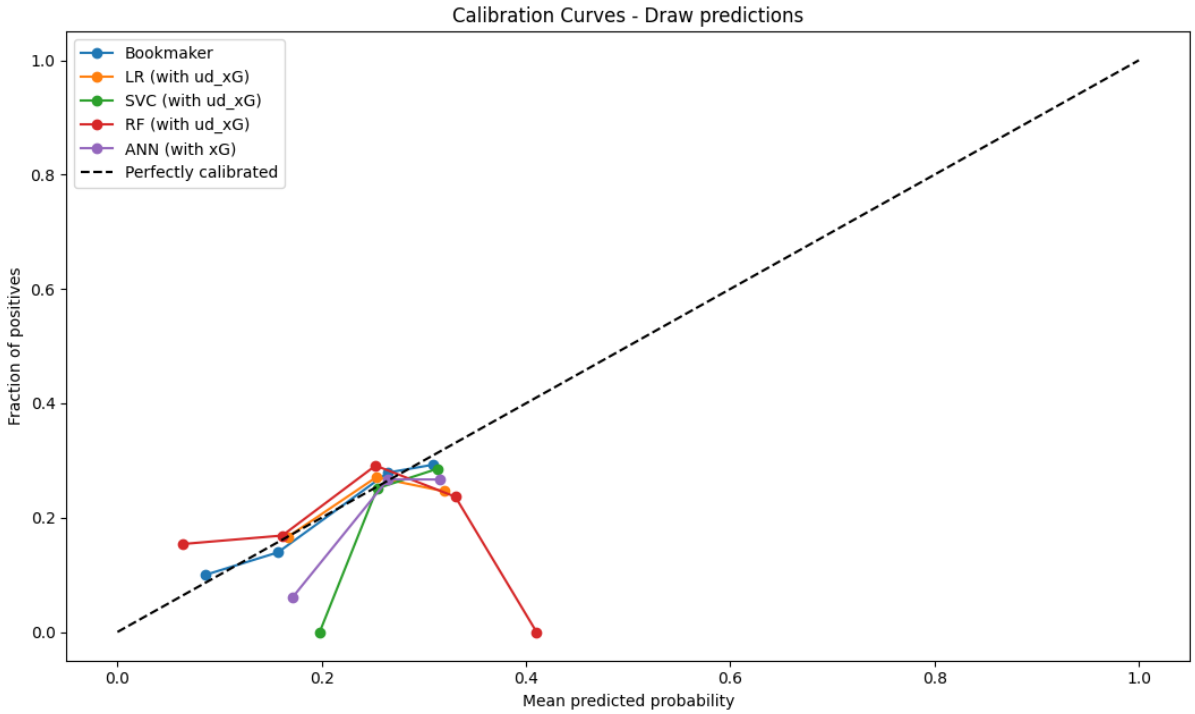

Figure A6: Calibration curves of different models when predicting Draws.

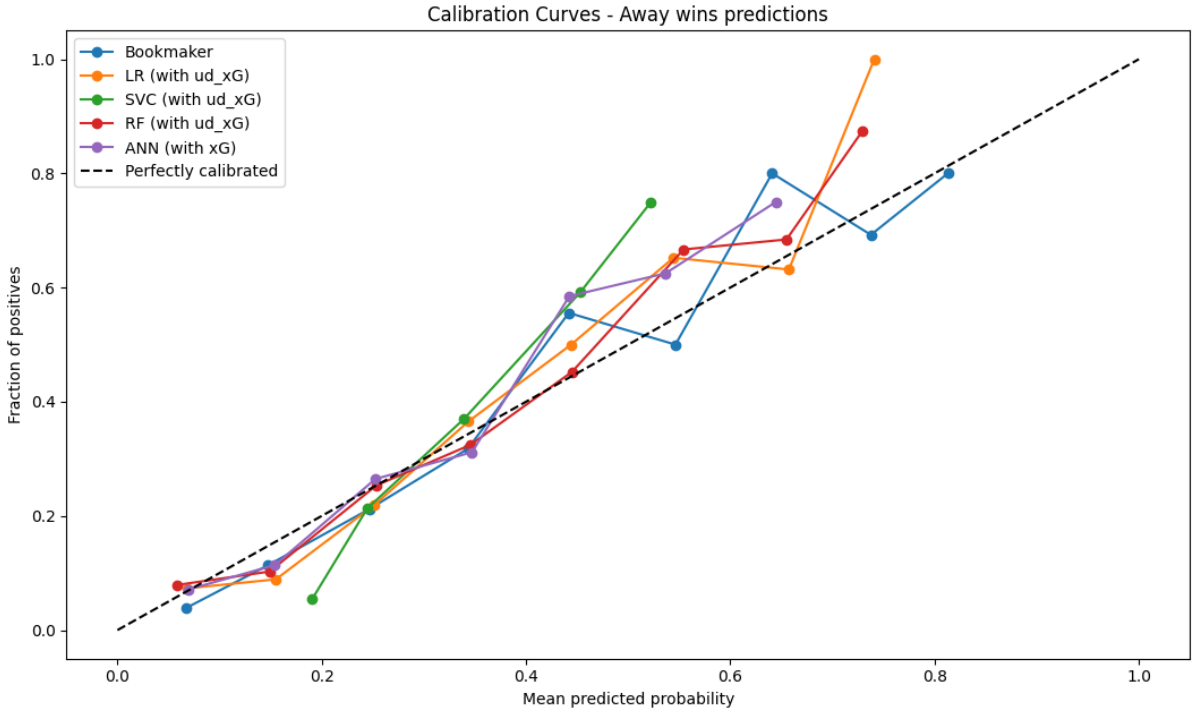

Figure A7: Calibration curves of different models when predicting Away wins.

Predictive Models' Feature Importance

Feature importance of Linear Regression (with ud_xG) model

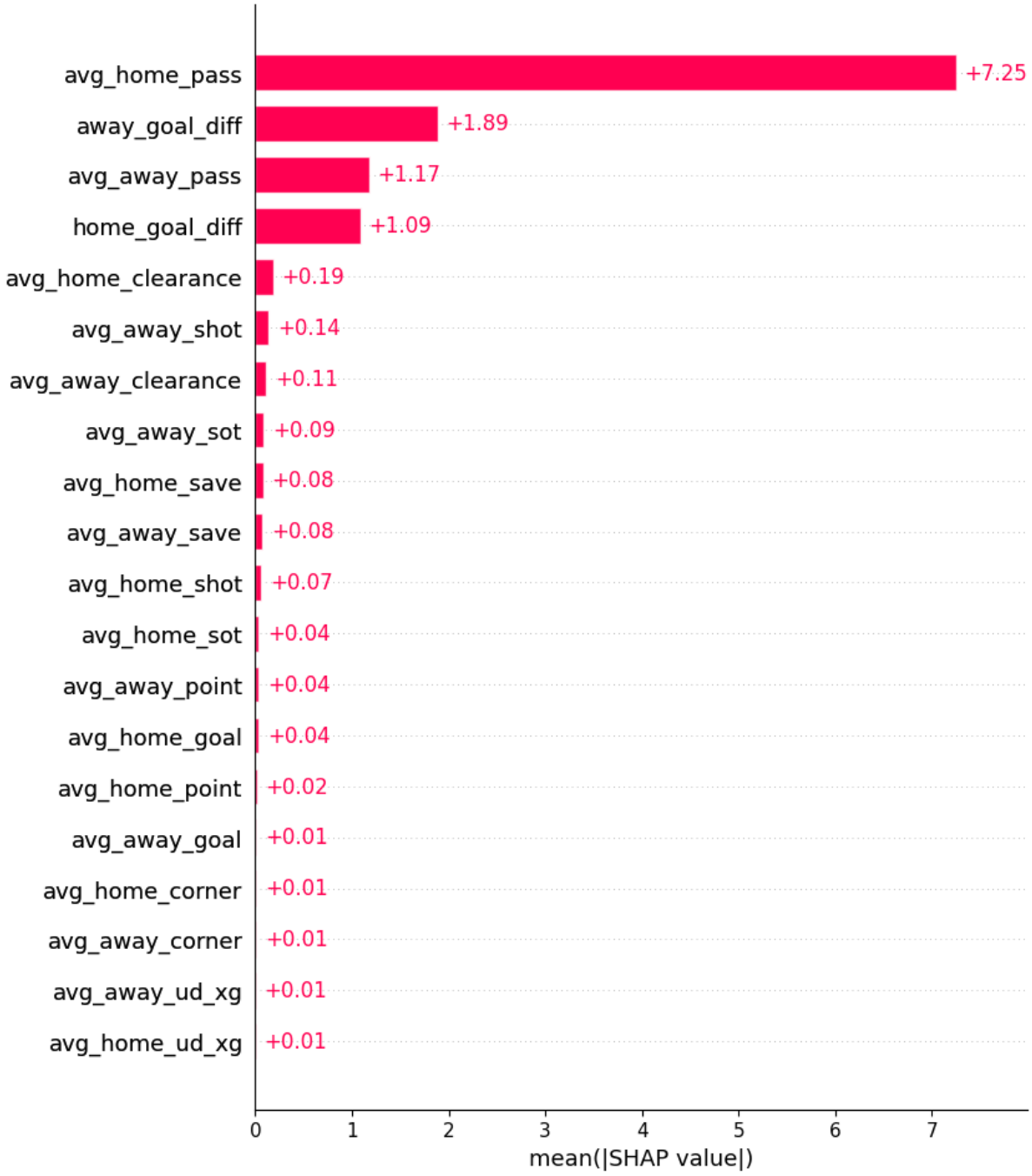

Figure A8: Feature importance when Linear Regression (with ud_xG) model predicts Home wins.

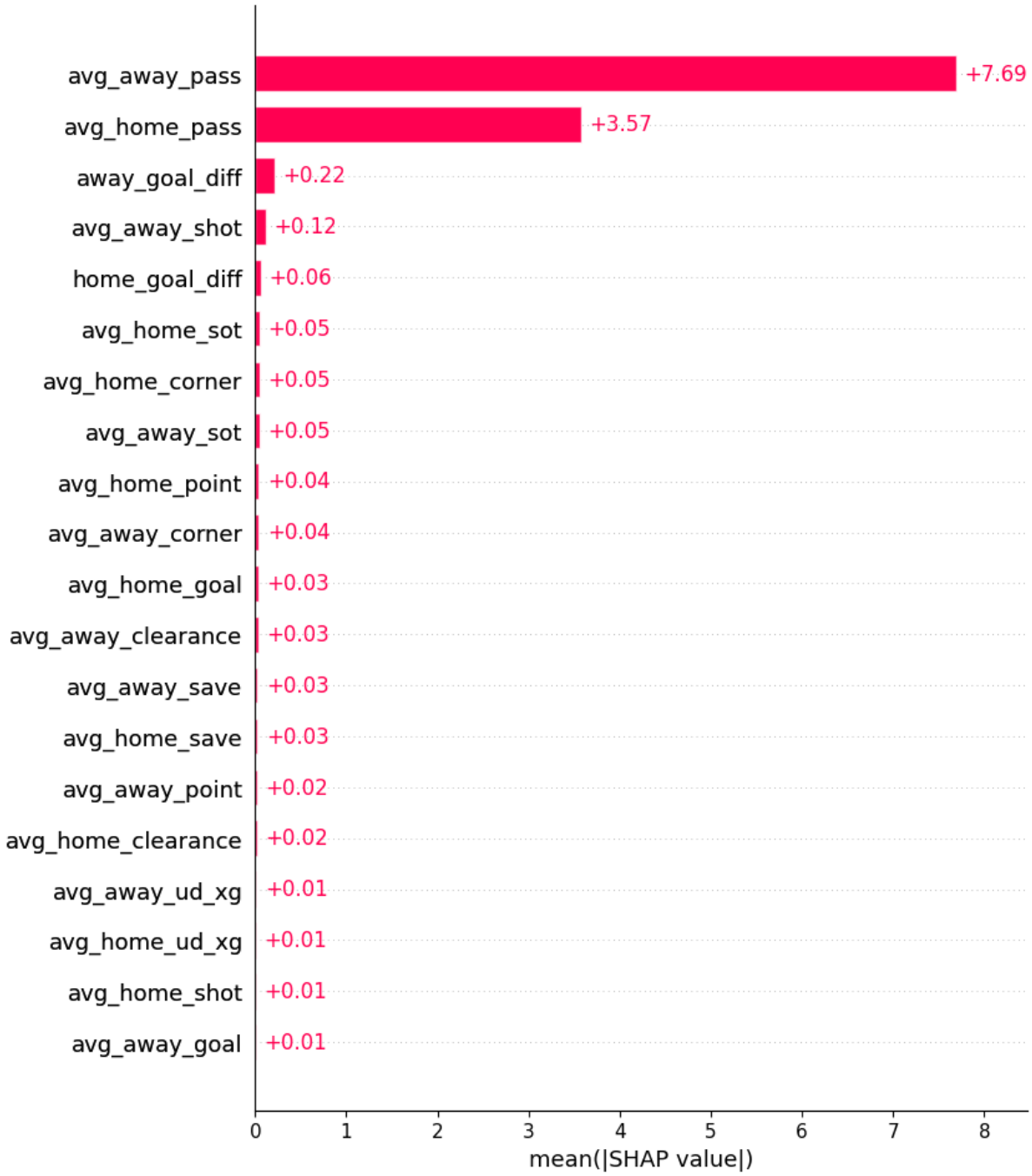

Figure A9: Feature importance when Linear Regression (with ud_xG) model predicts Draws.

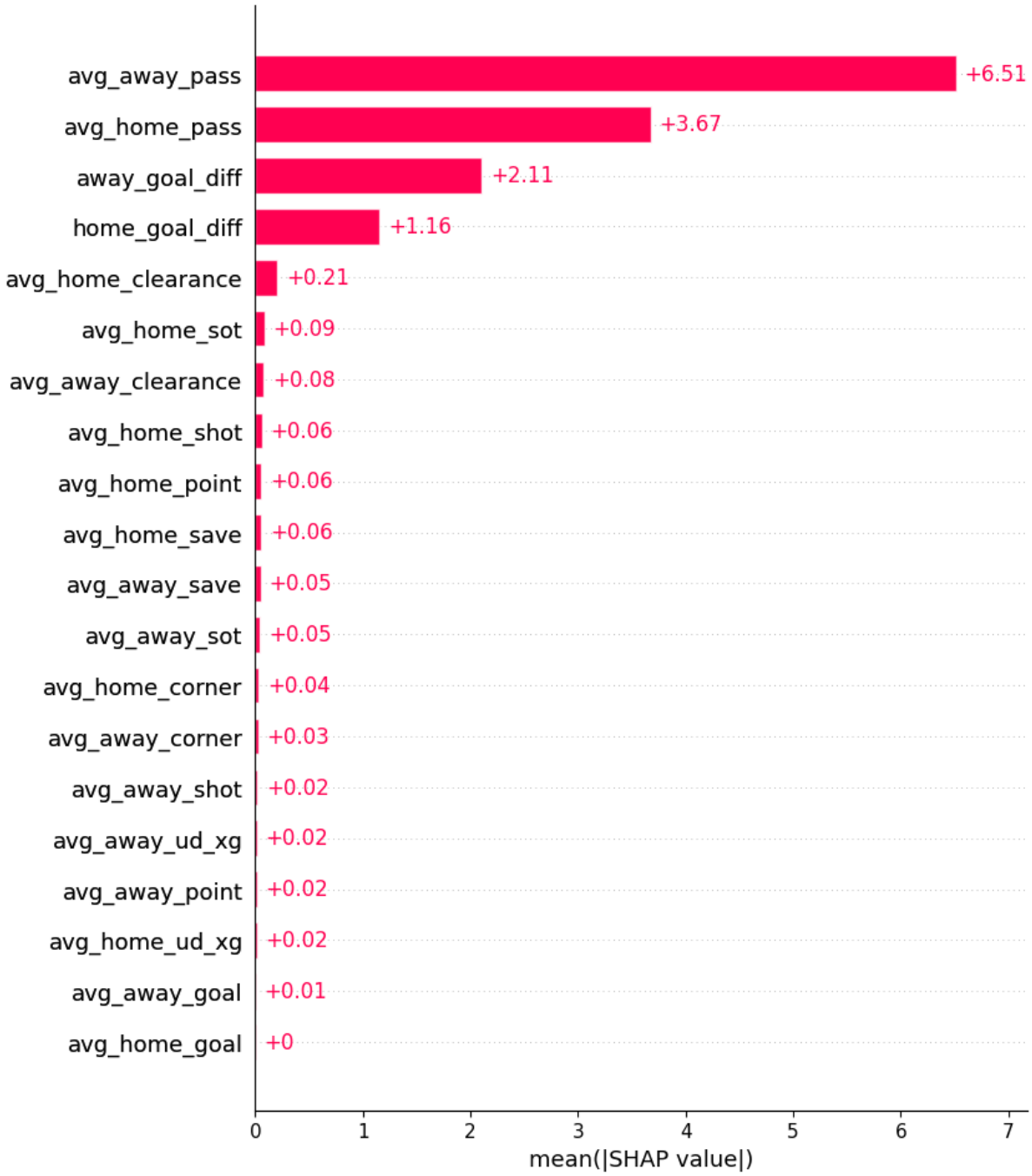

Figure A10: Feature importance when Linear Regression (with ud_xG) model predicts Away wins.

Feature importance of Random Forest (with ud_xG) model

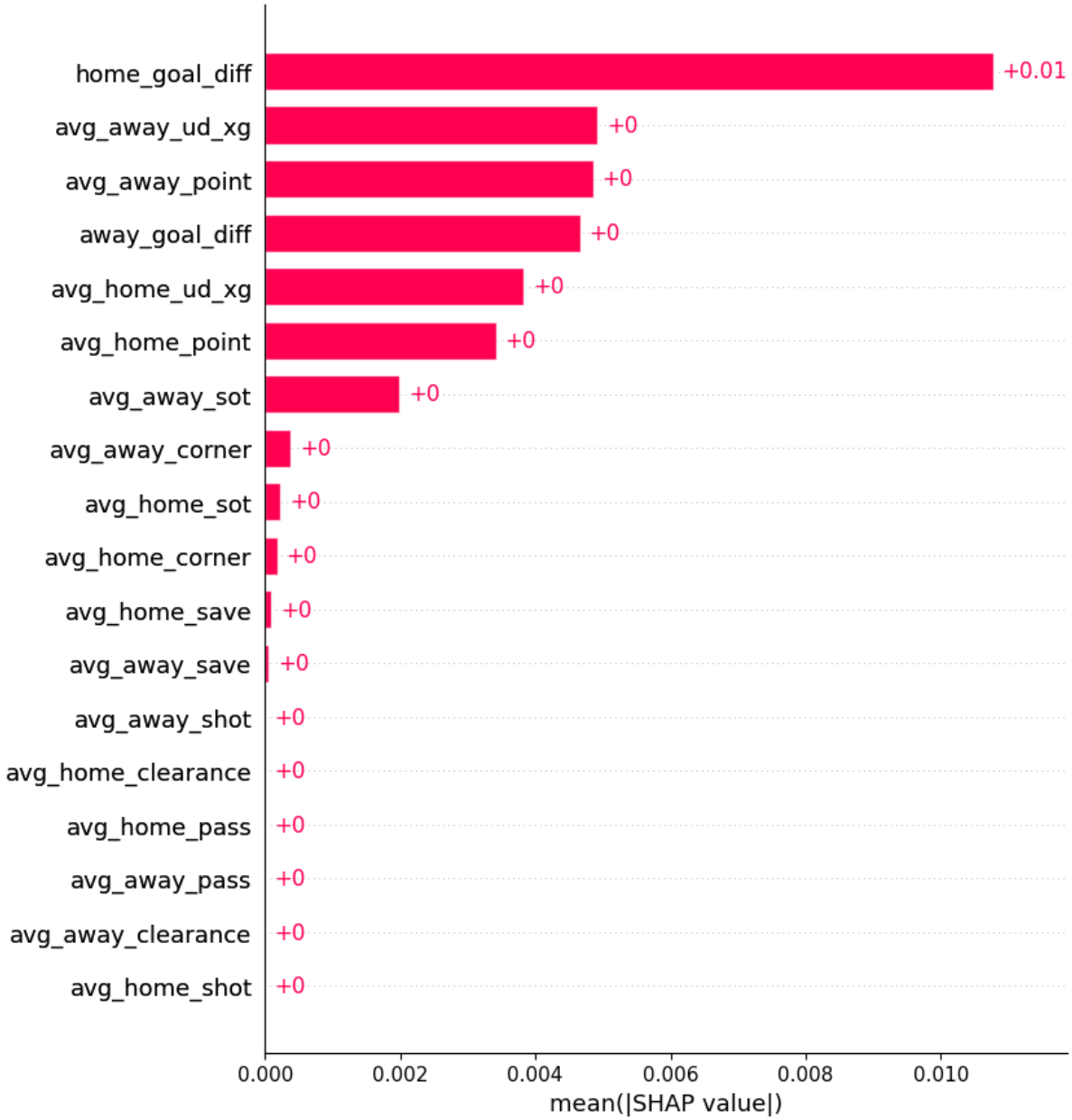

Figure A11: Feature importance when Random Forest (with ud_xG) model predicts Home wins.

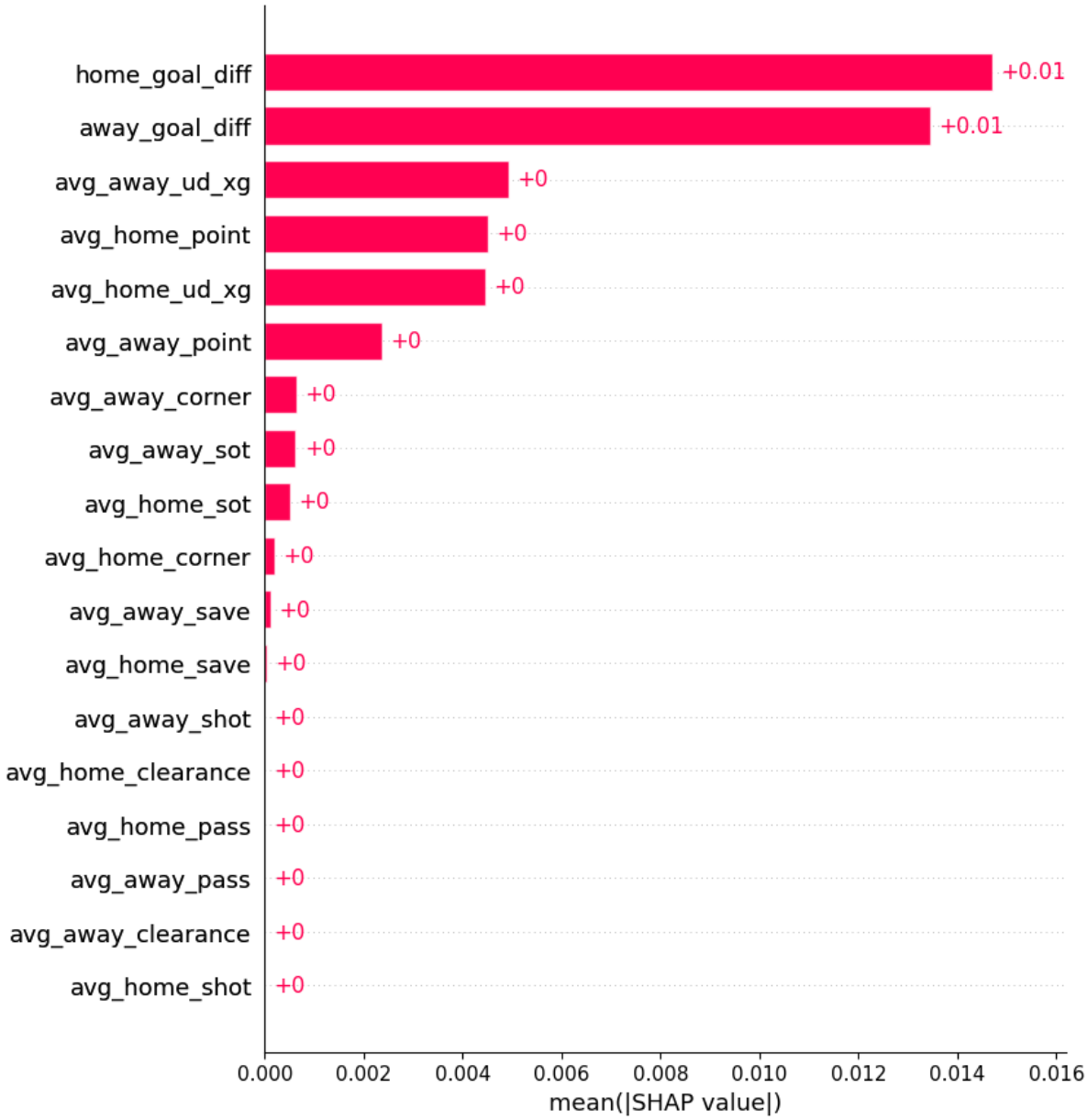

Figure A12: Feature importance when Random Forest (with *ud_xG*) model predicts Draws.

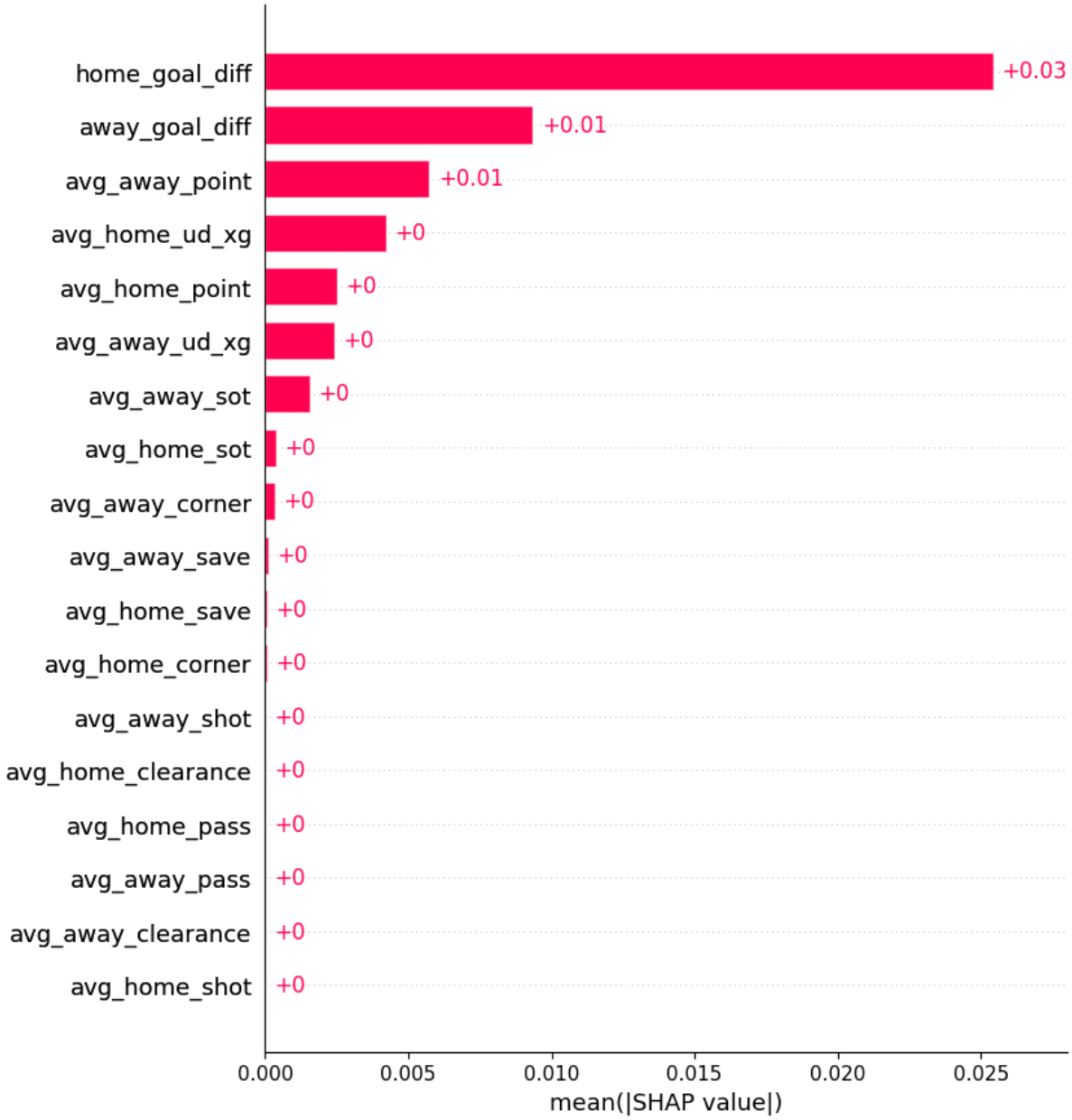

Figure A13: Feature importance when Random Forest (with ud_xG) model predicts Away wins.

List of Figures

List of Tables