

Addressing Evaluation Challenges on the Expected Goals (xG) Metric in Football Analysis

**Bachelor Project submitted for the degree of Bachelor of Science HES in
Business Information Technology**

by

William SENN

Bachelor Project Mentor:

Grigorios ANAGNOSTOPOULOS, Senior Research Associate HEG

Geneva, August 5th 2024

Haute École de Gestion de Genève (HEG-GE)

Business Information Technology

Disclaimer

This Bachelor thesis is completed as part of the final examination requirements of the Haute école de gestion de Genève (HEG), for the purpose of obtaining the title Bachelor of Science HES-SO in Business Information Technology. The student has sent this document by email to the address provided by their thesis advisor for analysis using the COMPILATIO plagiarism detection software.

The student agrees, if applicable, to the confidentiality clause. The use of the conclusions and recommendations formulated in this Bachelor thesis, regardless of their value, does not engage the responsibility of the author, the thesis advisor, the jury members, or HEG.

"I certify that I have completed this work alone without using sources other than those cited in the bibliography."

Geneva, August 5th, 2024

William Senn

Acknowledgments

At the conclusion of this bachelor's thesis, I would like to extend my heartfelt thanks to all the individuals and institutions who contributed to or supported me in the completion of this work.

First and foremost, my deepest gratitude goes to my supervisor and thesis director, Dr. Anagnostopoulos Grigorios. His guidance, expertise, and unwavering support were invaluable throughout my research journey, and I am profoundly thankful for his mentorship.

I also wish to express my appreciation to the HES-SO and the program Business Information Technology. The knowledge and skills I acquired during my Bachelor of Science HES-SO in Business Information Technology have been instrumental in my academic and professional development.

Special thanks are due to the Infothèque of the Haute école de gestion and its staff. Their access to resources and valuable referencing advice significantly aided my research efforts.

Lastly, I am grateful to my family and friends for their continuous encouragement and support.

This work would not have been possible without the support and contributions of the following individuals and institutions. Thank you.

Executive Summary

The Expected Goals (xG) metric in football analysis measures the probability of a shot resulting in a goal using historical data. This study evaluates various machine learning models, including Logistic Regression, Random Forest, and XGBoost, against StatsBomb xG values to address evaluation challenges in xG models.

We analyzed data from over 3000 games and 10 million events from the StatsBomb dataset, focusing on features like shot angle and distance. The models were evaluated using precision, recall, F1 score, and ROC-AUC due to significant class imbalance, where non-goal events far outnumber goal events. Traditional accuracy metrics were less effective.

Calibration plots showed that models were well-calibrated up to a probability of 0.5 but fluctuated beyond this. Logistic Regression aligned closely with actual xG values, while all models tended to underestimate high xG values. The study highlighted the importance of high-quality contextual data over model complexity and found hyperparameter tuning less effective.

Addressing class imbalance was identified as crucial for developing high-performing models, as it significantly impacts their accuracy and reliability. The recommended strategies include resampling techniques like oversampling the minority class or undersampling the majority class, adjusting class weights to give more importance to the minority class, and using Stratified K-Fold Cross-Validation to ensure balanced class proportions during model training and validation. Implementing these methods could improve the overall performance and robustness of the models.

Continuous validation with real-world data is essential for model relevance and accuracy. The study also emphasized the need for more data on female players to improve gender-specific shooting pattern analysis and model inclusivity.

In conclusion, developing accurate xG models in football analytics requires robust performance metrics, handling class imbalance, ensuring data quality, and ongoing real-world validation. Future work should mostly focus on data refining and calibration methods but also advanced modelling techniques to improve predictive accuracy and reliability.

Table of contents

Disclaimer	i
Acknowledgments	ii
Executive Summary	iii
Table of contents	iv
List of tables	vi
List of figures	vi
Note	vii
List of Abbreviations	viii
1. Introduction	1
2. Literature Review	3
2.1 Overview of xG Metric.....	3
2.2 Other key metrics	3
2.2.1 Expected Threats (xT).....	3
2.2.2 Valuing Actions by Estimating Probabilities (VAEP)	3
2.2.3 On-Ball Value (OBV)	3
2.2.4 Goals Added (+G)	4
2.3 Types of Data in football analytics	4
2.3.1 Event data	4
2.3.2 Tracking data.....	5
2.4 Diverse xG model methodologies.....	7
2.5 Relevant evaluation metrics	8
2.6 Limitations of xG models.....	10
2.6.1 Lack of Consideration for Individual Player Abilities.....	10
2.6.2 Complexity and Variability of Football Situations	10
2.6.3 Ball Event data	10
3. Data	12
3.1 Selected data	12
3.1.1 Data description	12
3.2 Data preparation	14
3.3 Feature engineering	16
3.3.1 Angle	16
3.3.2 Distance	16
3.3.3 Final Datasets	16
4. Exploratory Data analysis	19
4.1 Distribution of the label class	19
4.2 Goal rate density and shot frequency map.....	21

4.3 Historical shot distance analysis	23
4.3.1 Shot distance comparison between male and female	24
4.3.2 Historical analysis of male shot distance trends	25
5. Methodology	28
5.1 Training	28
5.1.1 Binary Classification	28
5.1.2 Probabilistic Output	30
6. Evaluation	32
6.1 Model Benchmarking	32
6.2 Performance Metrics	32
6.3 Recommendations	34
7. Results exploration	37
7.1 Calibration Plot Analysis	37
7.2 Analysis of predicted probabilities	40
8. Conclusion	48
8.1 Summary of findings	48
8.2 Future work	50
Bibliography	52
Appendix 1: Definition of different zones of the penalty area	56

List of tables

Table 1: List of Statsbomb competitions open data	12
Table 2: Sample of training data	28
Table 3: Logistic regression parameters	29
Table 4: Random forest parameters	29
Table 5: XGBoost parameters	29
Table 6: Performance of models across various metrics	32

List of figures

Figure 1: Example of non-exhaustive event dataset (originally 91 columns).....	5
Figure 2: Representation of the local positioning systems (Serrano 2020)	6
Figure 3: Representation of the video-based optical systems	6
Figure 4 : Representation of the Global Navigation Satellite Systems	7
Figure 5: Comparative goals distribution between male and female; (A) male distribution; (B) female distribution.....	20
Figure 6: Comparative goal rate map and shot frequency map between male and female ; (A) male goal rate; (B) male shot frequency; (C) female goal rate (D) female shot frequency	21
Figure 7: Comparative average shot distance between male and female	24
Figure 8: Male average shot distance per year with number of shots	25
Figure 9: Shot distance distribution for male players (2004-2024)	27
Figure 10: Calibration Plot of Multiple Models with 20 Bins	37
Figure 11: Distribution of predicted probabilities by different models; (A) XGBoost; (B) Logistic Regression; (C) Random Forest.....	41
Figure 12: Scatter plots comparison of xG values between trained models and StatsBomb; (A) Logistic Regression; (B) XGBoost; (C) Random Forest	45

Note

This bachelor thesis utilized the assistance of ChatGPT from OpenAI to help with the reformulation of sentences. This generative AI was employed to improve the clarity and readability of the text, ensuring that the content is communicated effectively. The use of AI tools in this manner is aimed at improving the overall quality of the thesis while maintaining the integrity and originality of the research.

List of Abbreviations

AUC	Area Under the Curve
DCNN	Deep Convolutional Neural Network
EPTS	Electronic Performance & Tracking Systems
F1 Score	Harmonic mean of precision and recall
GNSS	Global Navigation Satellite Systems
HEG-GE	Haute École de Gestion de Genève
HES-SO	Haute École Spécialisée – Suisse Occidentale
IQR	Interquartile Range
LMS	Learning Management System
MCC	Matthews Correlation Coefficient
OBV	On-Ball Value
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
VAEP	Valuing Actions by Estimating Probabilities
xG	Expected Goals
xT	Expected Threats

1. Introduction

In recent years, sports analytics has experienced significant growth, with football analytics seeing a substantial increase in academic articles, highlighting its growing importance (fortunebusinessinsights 2024). The advent of data-driven technologies, such as wearable devices—electronic gadgets attached to the body—has revolutionized the discipline, transforming the primary challenge from data collection to data analysis (Rein 2016).

Data-driven technologies have impacted various aspects of football analytics. Wearable devices, for instance, quantify fatigue and physical strain, helping to prevent potential injuries. They also measure variables such as speed and distance to evaluate player performance and map out movement patterns on the field. This data-driven approach enables teams to gain valuable insights into player behavior and optimize their strategies accordingly.

The transformation and value brought about by football analytics underscore its power to reshape the sport's landscape. These advancements revolutionize how teams, coaches, and analysts approach the game, integrating data analysis as a fundamental component of modern football strategies (Cacho-Elizondo 2020). As the field continues to evolve, data-driven decision-making becomes increasingly important in enhancing performance and strategy in football.

In this constantly evolving context, Expected Goals (xG) have emerged as a key metric in football analysis. This metric quantifies the quality of scoring opportunities during a match by measuring the probability that a shot will result in a goal, using historical data on similar shots (Whitmore 2023). Various methodologies, ranging from simple models to sophisticated machine learning algorithms, have been proposed to calculate xG.

Despite the advancements in xG modeling, there is a lack of a comprehensive evaluation framework to assess how effective xG models are at quantifying goal opportunities. Given the critical role of xG metrics in tactical decisions and performance assessments, addressing the associated evaluation challenges is essential for advancing football analytics. This study aims to evaluate different xG models, identify key features influencing xG predictions, analyze evaluation challenges, and propose methods to improve model evaluation. By exploring these aspects, this research seeks to enhance the understanding and application of xG metrics in football, ultimately contributing to better data-driven decision-making in this sport.

This work has been realized primarily with Python, sklearn, and the Mplsoccer library (*mplsoccer 1.4.0 documentation*) from StatsBomb, among others. These tools have been essential, and their application was enriched by the machine learning courses and methodological tools from the Haute école de Gestion, as well as additional courses from Coursera, Fastai, and other platforms.

2. Literature Review

2.1 Overview of xG Metric

The Expected Goals (xG) metric is a pivotal tool in football analytics, designed to quantify the quality of scoring opportunities by measuring the probability that a shot will result in a goal. This probability is calculated using historical data from similar shots, considering multiple factors such as shot location, angle, the body part used to take the shot (foot, head), and contextual factors like defensive pressure and game state. The xG metric provides a nuanced understanding of a team's offensive performance, rather than simply counting goals to evaluate the quality of their scoring chances.

2.2 Other Key Metrics

Beyond xG, several other metrics are widely used in football analytics to assess team and player performance. These metrics provide complementary insights that enhance the overall understanding of the game:

2.2.1 Possession Value (PV)

Possession Value (PV) quantifies the value of maintaining possession and its contribution to creating scoring opportunities. It evaluates how possession influences the likelihood of creating chances and preventing the opposition from gaining control (Whitmore 2020).

2.2.2 Expected Threats (xT)

Expected Threats (xT) measures the potential threat created by moving the ball into advantageous positions. It assigns value to actions like passes and dribbles based on their likelihood to lead to a goal, highlighting players who create significant goal-scoring opportunities (Sumpter 2021).

2.2.3 Valuing Actions by Estimating Probabilities (VAEP)

Valuing Actions by Estimating Probabilities (VAEP) assesses the impact of every on-ball action by estimating its effect on the probability of scoring and conceding goals. It provides a comprehensive view of a player's contribution to both offensive and defensive aspects of the game (Decroos 2020).

2.2.4 On-Ball Value (OBV)

On-Ball Value (OBV) measures the effectiveness of a player's actions while in possession of the ball, such as passing, dribbling, and shooting. It helps identify players who excel in creating and capitalizing on scoring opportunities (StatsBomb 2021).

2.2.5 Goals Added (+G)

Goals Added (+G) evaluates a player's overall contribution to the team's goal-scoring efforts by valuing every on-ball action in terms of goals. It uses contextual data to estimate the impact of each action on the probability of scoring and conceding goals. By aggregating these impacts, Goals Added (+G) provides a holistic assessment of a player's influence on team performance (Muller 2020).

2.3 Types of Data in Football Analytics

In football analytics, two primary data standard types are utilized to gain insights into match dynamics and player performance: ball event data and tracking data. Each data type provides unique and complementary perspectives on the game, enabling a comprehensive analysis of various aspects of football matches.

2.3.1 Event Data

Event data, also called ball event data, are crucial for understanding match dynamics and player performance, as they specifically track actions directly involving the ball. These data provide a detailed chronicle of significant ball-related actions during a match, such as passes, dribbles, crosses, interceptions, tackles, and shots. Additionally, they include critical incidents like fouls and the issuance of cards, which can significantly influence the game's outcome.

Each recorded event is characterized by several key attributes directly linked to the ball. The timestamp marks the exact moment the event occurred, while the location on the pitch, represented by (x,y) coordinates, provides precise spatial context. The type of event (e.g., pass, cross, or foul) categorizes the specific action, and the players involved are identified to attribute actions correctly.

Depending on the nature of the event, further specific details are documented. For instance, in the case of a pass, the data include the start and end locations of the ball, giving insight into its direction and distance. For tackles, the outcome (successful or otherwise) indicates whether the defending player managed to dispossess the opponent of the ball.

Figure 1: Example of non-exhaustive event dataset (originally 91 columns)
(statsbomb/open-data 2024)

period	timestamp	possession	player_name	position_name	pass_recipient	pass_length	pass_angle	pass_height	end_x	end_y	body_part
1	00:00:00	Hoffenheim	Maximilian Be	Left Center Forw	Tim Drexler	20.41568	3.102397	Ground Pass	40.6	40.9	Right Foot
1	00:00:02	Hoffenheim	Tim Drexler	Left Center Back					41.2	41.1	
1	00:00:02	Hoffenheim	Tim Drexler	Left Center Back							
1	00:00:02	Hoffenheim	Tim Drexler	Left Center Back	Florian Grillitscl	2.280351	-1.3045443	Ground Pass	41.8	38.9	Left Foot
1	00:00:03	Hoffenheim	Florian Grillit:	Center Back							
1	00:00:03	Hoffenheim	Florian Grillit:	Center Back	Pavel Kadeřábel	59.5373	0.4426139	High Pass	95.6	64.4	Right Foot
1	00:00:05	Hoffenheim	Pavel Kadeřát	Right Wing Back							
1	00:00:05	Hoffenheim	Piero Martín I	Left Center Back							Head
1	00:00:08	Hoffenheim	Pavel Kadeřát	Right Wing Back							
1	00:00:08	Hoffenheim	Alejandro Grir	Left Wing Back							Head
1	00:00:21	Hoffenheim	Pavel Kadeřát	Right Wing Back	Wout Weghorst	23.517227	-1.6090755	High Pass	110.6	56.5	
1	00:00:23	Hoffenheim	Wout Weghor:	Right Center For							
1	00:00:23	Hoffenheim	Wout Weghor:	Right Center For					110.4	54.8	
1	00:00:23	Hoffenheim	Wout Weghor:	Right Center For							
1	00:00:24	Hoffenheim	Alejandro Grir	Left Wing Back							
1	00:00:24	Hoffenheim	Alejandro Grir	Left Wing Back					7.5	22.9	
1	00:00:25	Hoffenheim	Wout Weghor:	Right Center For							
1	00:00:26	Hoffenheim	Alejandro Grir	Left Wing Back	Piero Martín Hin	3.5608988	2.2367656	Ground Pass	5.3	25.7	Left Foot
1	00:00:26	Hoffenheim	Piero Martín I	Left Center Back							
1	00:00:26	Hoffenheim	Piero Martín I	Left Center Back		26.685202	-0.38816106	High Pass	30	15.6	Left Foot
1	00:00:28	Hoffenheim	Ozan Muhamr	Right Center Bac	Pavel Kadeřábel	16.734396	0.43157306	High Pass	105.3	71.5	Head
1	00:00:29	Hoffenheim	Pavel Kadeřát	Right Wing Back							
1	00:00:29	Hoffenheim	Pavel Kadeřát	Right Wing Back		18.970766	-1.8375107	High Pass	100.3	53.2	Head
1	00:00:31	Hoffenheim	Jonathan Tah	Center Back							Head
1	00:00:32	Hoffenheim	Tim Drexler	Left Center Back	David Jurásek	24.485914	-1.920963	Low Pass	69.6	27.1	Right Foot
1	00:00:34	Hoffenheim	David Jurásek	Left Wing Back							

Event data enables the evaluation of individual and team performances by examining how players interact with the ball and identifying strengths and weaknesses. By observing patterns and trends in ball movement and control, they can uncover strategic insights, such as preferred passing lanes or common defensive breakdowns.

Event data is collected semi-automatically. There's a meticulous process that involves a human and a computer working in tandem to collect the raw data that powers everything. Computer vision is used to help the data collector tag events such as shots, passes and tackles, and insert their locations on the pitch. According to Walid (Walid 2023), analyzing a single live match takes five people, but together they can reach 99 percent accuracy.

Moreover, event data support the development of advanced metrics and models that focus on ball movement and player interactions, enhancing predictive accuracy and strategic planning. By systematically capturing every critical moment involving the ball on the pitch, these datasets provide a foundational resource for advancing football analysis, improving performance evaluation, and driving strategic innovation in the sport (Jan Van Haaren 2019).

2.3.2 Tracking Data

There has been a significant shift towards the use of tracking data in football analytics, driven by advancements in Electronic Performance & Tracking Systems (EPTS). Three

main types of EPTS exist in the commercial market: these are Global Navigation Satellite Systems (GNSS), local positioning systems (LPS) and video-based (optical) systems.

Figure 2: Representation of the local positioning systems (Serrano 2020)

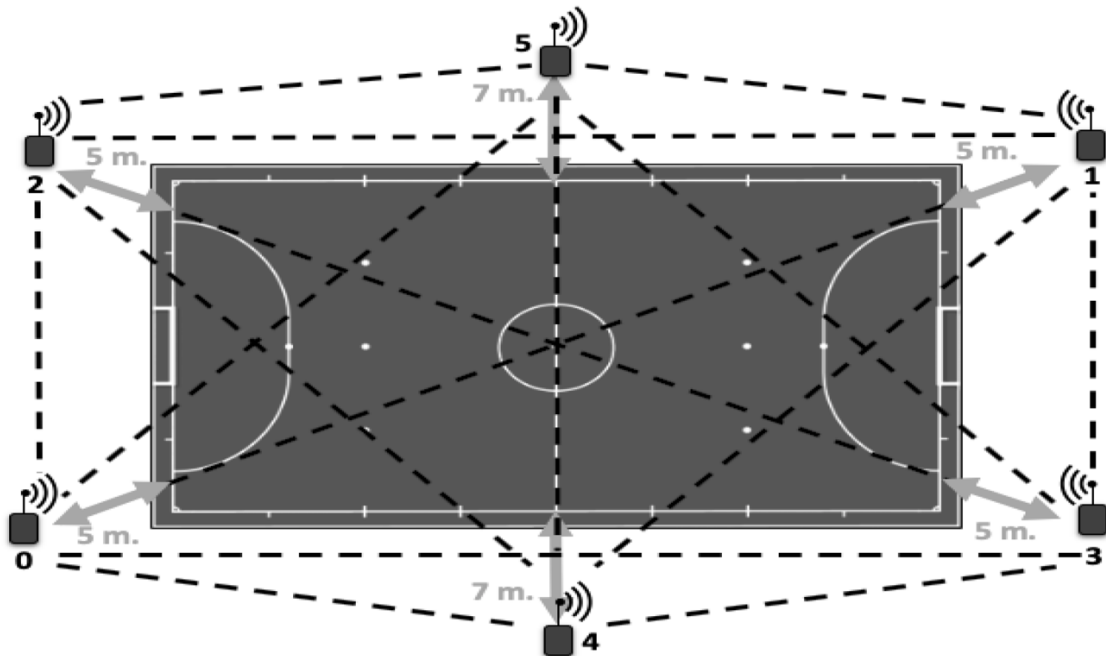
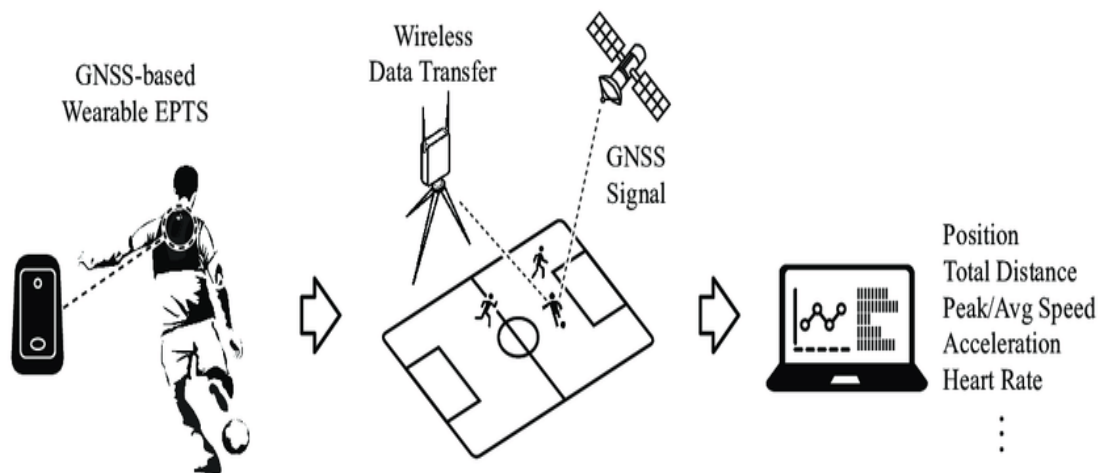


Figure 3: Representation of the video-based optical systems (Akyildiz 2022)



CC BY 4.0 Z. Akyildiz, H. Nobari, F. González-Fernández, et al., 2022. *Variations in the physical demands and technical performance of professional soccer teams over three consecutive seasons.* <https://www.nature.com/articles/s41598-022-06365-7>

Figure 4 : Representation of the Global Navigation Satellite Systems



As Vidal-Codina et al. (Vidal-Codina 2022) highlight, these systems provide a wealth of information by capturing coordinates for all players and the ball multiple times per second. Unlike event data, tracking data offers a more comprehensive view of the game, including off-ball movements and team dynamics. This rich dataset allows for the quantification of previously qualitative observations, opening new avenues for analysis. Tracking data typically contains around 3 million data points per match (at 25 Hz), compared to approximately 3,000 events in traditional event data. This vast amount of information has led to the development of advanced metrics such as team tactics evaluation, pitch control analysis, and expected possession value (Vidal-Codina 2022).

2.4 Diverse xG Model Methodologies

In the field of football analysis, numerous Expected Goals (xG) models have been developed, each incorporating a variety of variables such as location, distance, shot type, speed, and angle. Among the most notable contributors to this diversity is Michael Caley, a prominent statistician and author of "Cartilage Free Captain." Caley has published several xG models, each employing a distinct methodology (Rathke 2017).

Caley's models include various focuses:

- **Shots on Target Model:** This model concentrates exclusively on the analysis of shots on target, providing insights into the quality of these specific attempts (Caley 2013a).
- **Pitch Areas Model:** This model targets different areas of the pitch to understand how location influences scoring opportunities (Caley 2013b).

- **Comprehensive Team Quality Model:** This model evaluates multiple factors such as location, speed of attack, passes, set pieces, and player payroll to assess overall team quality (Caley 2013b).

A significant finding in the evaluation of xG models is the importance of distance as a determining factor. According to Bertin (2015) using distance alone as an indicator for calculating xG can capture 85% of the variability in scoring chances. This underscores the predictive power of shot distance in assessing the likelihood of a goal.

Additionally, defensive pressure has been identified as a critical factor in xG studies, influencing the reduction of xG value. Defensive actions can significantly impact the probability of a shot resulting in a goal, making it an essential component of accurate xG models (Caley 2015).

Despite the advancements in xG modeling, there is no universally accepted "golden rule" regarding which factors should be included in calculating xG. The literature reflects the complexity and multifactorial nature of football, indicating that various approaches can be valid depending on the specific context and objective of the analysis.

By examining these diverse methodologies and factors, researchers and analysts can develop more accurate and insightful xG models. This, in turn, enhances the strategic capabilities of football teams, allowing them to better understand and improve their offensive and defensive performances.

2.5 Relevant Evaluation Metrics

Deciding which performance metrics to use for evaluating an xG model is a complex challenge. The following metrics are commonly used in various studies to assess the accuracy and effectiveness of xG models (Cavus 2022):

Recall: Measures the ability of the model to identify all relevant instances of goals. High recall indicates that the model captures most actual goals, but it might also include many false positives.

Precision: Focuses on the accuracy of the goals predicted by the model. High precision means that most of the predicted goals are true goals, but it might miss some actual goals.

F1 Score: The harmonic mean of precision and recall, providing a balance between the two. It is useful when the need to balance false positives and false negatives is crucial.

Accuracy: The ratio of correctly predicted outcomes (both goals and no-goals) to the total number of predictions. While straightforward, it can be misleading in imbalanced datasets where one outcome (e.g., no goal) is much more common than the other.

Area Under the Curve (AUC): Measures the ability of the model to distinguish between different classes (goals vs. no goals). AUC is useful for understanding the trade-offs between true positive rates and false positive rates at various threshold settings.

Matthews Correlation Coefficient (MCC): Provides a balanced measure that considers true and false positives and negatives. It is particularly effective for binary classification tasks with imbalanced classes.

Brier Score: Measures the mean squared difference between predicted probabilities and the actual outcomes. It provides insights into the calibration and sharpness of the probabilistic predictions.

Log-loss: Evaluates the accuracy of probabilistic predictions, penalizing false classifications based on their confidence. Lower logloss values indicate better model performance.

Balanced Accuracy: Adjusts the traditional accuracy metric to account for imbalanced datasets by averaging the recall obtained on each class.

Coefficient of Determination (R^2): One of the most common metrics for evaluating xG models. It explains the percentage of total variation in the observed data that the model accounts for. Despite its widespread use, R^2 has been criticized for its inefficiency in providing a complete picture of model performance (Mackay 2016).

The evaluation challenge is particularly difficult because many researchers use different metrics without a common rule or consensus on what constitutes good performance and how to know if the model accurately represents reality. This variability complicates comparisons across studies and makes it challenging to establish a standard benchmark for model evaluation.

Moreover, Professor David Sumpter, an author and expert in football analytics, emphasizes the necessity of validating xG models through practical application and real-world comparisons. This involves testing the model against actual match data, iteratively refining it based on real-world performance, and seeking expert validation to ensure the model's predictions are realistic and useful in a football context (Friends of Tracking 2020b).

2.6 Limitations of xG Models

Even with significant advancements in xG modeling techniques, several notable limitations continue to affect their accuracy and applicability.

2.6.1 Lack of Consideration for Individual Player Abilities

One of the significant gaps in current xG models is their failure to account for individual player abilities (Pinnacle 2019). In any given situation, assigning the same xG value to two players with different abilities, such as a highly-skilled Player 1 and a less skilled Player 2, does not accurately reflect their true probabilities of scoring. This generalization overlooks crucial nuances related to individual talents and goal-scoring efficiency, which is a significant shortcoming in the current calculation and interpretation of xG (Cefis 2024).

2.6.2 Complexity and Variability of Football Situations

The limitation of xG models is further exacerbated by the lack of comprehensive data that can capture the innumerable possibilities in the field. The variety and complexity of game situations are such that it becomes challenging to account for all specific nuances related to each player. This data constraint limits the models' ability to integrate these subtleties, thereby failing to distinguish between the unique skills of players in seemingly identical situations (Haidair 2019).

By addressing these limitations, future research can aim to develop more sophisticated xG models that better incorporate individual player abilities and more detailed game data. This would significantly enhance the accuracy and applicability of xG models in football analytics, providing a more precise tool for evaluating scoring opportunities.

2.6.3 Ball Event Data

The vast majority of xG models are constructed primarily using ball-related event data, which inherently limits their scope and accuracy (Mead, O'Hare, McMenemy 2023). These models predominantly focus on the immediate context surrounding the quality of the shot, such as the location, type of assist, and defensive pressure at the time of shooting. However, this approach overlooks a crucial aspect of football: off-ball movement.

Off-ball movement, which includes player positioning, runs, and spatial manipulation without possession, plays a significant role in creating and shaping scoring opportunities. These movements can dramatically influence the quality of a chance, yet they are not typically captured in standard event data. As a result, xG models may fail to account for

the build-up play, tactical positioning, and dynamic spatial relationships that precede a shot.

This limitation means that xG models might undervalue the contributions of players who excel in off-ball movement or overestimate the quality of chances that appear favorable based solely on ball-related factors. To develop more comprehensive and accurate xG models, future research should explore ways to incorporate off-ball data, potentially through advanced tracking technologies or more sophisticated event coding systems. This would allow for a more holistic evaluation of scoring opportunities, better reflecting the complex nature of football and the myriad factors that contribute to goal-scoring potential.

3. Data

To explore, compare, develop, and evaluate xG models across both male and female football, we need high-quality data. We will outline the preprocessing steps, including filtering, event selection, and aggregation, to ensure the data's consistency and relevance, which is crucial for developing robust xG models. Our methodology emphasizes the inclusion of both male and female competitions, allowing for a comparative analysis that broadens the scope and applicability of the findings. Additionally, we will calculate specific features to enhance our analysis. It is essential to maintain separate datasets for male and female competitions to accurately reflect the differences and nuances in each.

3.1 Selected Data

For this study, the data used was collected and curated by Statsbomb through a combination of computer vision, human input, and artificial intelligence; it offers accurate and comprehensive detailed event data essential for this study. Statsbomb is a sports data business company located in England and specialized in football analytics (Viloria 2024).

For our analysis, we focus on event data due to its widespread availability and established use in xG modeling. However, future studies could potentially incorporate tracking data to enhance the depth and accuracy of xG models (StatsBomb 2018).

3.1.1 Data description

The data used in this study is obtained from Statsbomb open data (*statsbomb/open-data* 2024) and includes a diverse range of international competitions and national leagues. The Table 1 below is a summary of the competitions:

Table 1: List of Statsbomb competitions open data

International Competitions	National Leagues
African Cup of Nations	Bundesliga
Champions League	Indian Super League
Copa America	La Liga
FIFA U20 World Cup	Ligue 1
FIFA World Cup	Major League Soccer
UEFA Euro	Premier League
UEFA Europa League	Serie A

In the subsections below, we will provide an overview of how the data is organized and detail the different data types included.

3.1.1.1 Competition Data

Competition data outlines the various competitions included in the data. Each record in this file consists of the following attributes:

- Competition ID and Season ID: Unique identifiers for the competition and its respective seasons.
- Competition Name: Name of the competition (e.g., Premier League, La Liga, etc.).
- Competition Gender: Gender of the participants (male or female).
- Country Name: The country or region the competition is based in.
- Season Name: Label of the season (e.g., 2017/2018).
- Match Updated and Match Available: Timestamps indicating the last update of a match and the availability of match data.

3.1.1.2 Match Data

Match data is stored in individual JSON files within a dedicated folder, with each file named using a unique match ID (e.g., 11.json). Match data is organized by competition and season IDs. Each match record provides:

- Match ID: A unique identifier for each match.
- Home and Away Teams: Information about the teams playing, including team ID, team name, and manager details.
- Match Date and Kick-Off Time: The scheduled date and time of the match.
- Stadium: Details of the stadium including its ID, name, and location.
- Referee: Information about the referee overseeing the match.
- Match Status: Status of the match data (e.g., available, scheduled).
- Scores: Final scores of the match.
- Competition Stage: The stage of the competition (e.g., regular season, finals).

3.1.1.3 Event Data

Events data files are linked to specific matches through match IDs with each file (e.g., 1564728.json) corresponding to a specific match and containing all events for that match. Each event is recorded with:

- Event Type: Describes the nature of the event (e.g., pass, shot, goal).
- Player and Team ID: Identifiers for the player involved and their team.
- Location: Pitch coordinates (x, y) marking where the event occurred.
- Timestamp: Match time at which the event took place.
- Related Events: Links to related events within the same match.
- Tactical Setup: Information on player formations and tactical changes during the match.

3.1.1.4 Lineup Data

Lineup data is stored in separate JSON files, each linked to a specific match through a unique ID. These files (e.g., [match_id].json) contain detailed information about the teams' lineups for each match and include the starting players, their positions, jersey numbers, and substitution details.

3.2 Data Preparation

In this section, we outline the steps taken to prepare our data. To ensure our data is both current and comprehensive, we utilized a cutoff date of March 1, 2024. This guarantees consistency and accuracy throughout our study, covering over 2800 games for males with more than 10 million events and 509 games for females with more than 1.7 millions events.

We extracted competition-level data, integrated match-level data, and compiled event data, focusing on relevant shot events. Separate datasets for male and female competitions were maintained to allow independent analysis and comparison of xG patterns between genders. This thorough process ensured our final datasets encompassed all available matches, providing a solid foundation for our research.

Competition-Level Data Extraction:

We initiated the process by parsing the competition data JSON file, which serves as a master index for all leagues and tournaments. For each competition entry, we systematically extracted key identifiers: Competition ID, Season ID, and Competition Gender. This step established the foundational structure for our dataset, allowing for precise categorization and future comparative analysis.

Match-Level Data Integration:

Utilizing the extracted Competition and Season IDs, we accessed and processed the corresponding match data JSON files. From each match file, we extracted the Match ID along with other pertinent match details, such as date, teams involved, and competition stage. This integration step created a robust linkage between high-level competition data and specific match instances, crucial for contextualizing subsequent event data.

Event Data Compilation:

For each identified Match ID, we located and process the associated event data JSON file. We extracted all events from these files, applying filters to isolate relevant shot events while preserving their contextual information. This compilation resulted in a comprehensive pool of event data, spanning across all matches and competitions, providing a rich dataset for our xG model.

Gender-Based Data Segregation:

Throughout the aggregation process, we implemented a systematic approach to maintain separate datasets for male and female competitions, using the Competition Gender attribute as the discriminating factor. This segregation was crucial for our research objectives, allowing for independent analysis and comparison of xG patterns between genders.

The meticulous nature of this aggregation process ensured that our final datasets encompassed event data from every available match across all competitions in the StatsBomb dataset. This data aggregation methodology provides a solid empirical foundation for addressing our research questions regarding xG in male and female football.

3.3 Feature Engineering

Two essential features of xG models are based on the shot position and goal post locations: the angle and distance features. These features play a crucial role in analyzing and predicting shot outcomes in expected goals model.

3.3.1 Angle

The angle feature is calculated using the cosine rule. First, the shot position is defined as $p = (x, y)$ and the locations of the top and bottom goal posts are defined as $g_0 = (120, 44)$ and $g_1 = (120, 36)$, respectively. These coordinates represent the positions of the goalposts on the field, where the x-coordinate of 120 corresponds to the goal line at the far end of the pitch. The y-coordinates of 44 and 36 represent the top and bottom goalposts, reflecting the standard goal width of 8 yards (or 7.32 meters). The full size of the football field, corresponding to these coordinates, is 120 meters in length and 80 meters in width. Two vectors, $v_0 = g_0 - p$ and $v_1 = g_1 - p$, are constructed to represent the vectors from the shot position to each goal post. The magnitudes of these vectors are calculated using the Euclidean norm: $|v_0| = \sqrt{(g_{0x} - x)^2 + (g_{0y} - y)^2}$ and $|v_1| = \sqrt{(g_{1x} - x)^2 + (g_{1y} - y)^2}$. The cosine of the angle θ between v_0 and v_1 is then computed using the dot product: $\cos(\theta) = \frac{v_0 \cdot v_1}{|v_0||v_1|}$. Finally, the angle θ is obtained by taking the inverse cosine (arccos) of the cosine value and converting it from radians to degrees.

3.3.2 Distance

To calculate the distance feature, the Euclidean distance formula is used to determine the shortest distance from the shot position to a point between the goalposts. The x-distance is computed as the difference between the x-coordinate of the shot position and the x-coordinate of the goal line: $x_{dist} = 120 - x$. The y-distance depends on the relative position of the shot with respect to the goal posts. If the shot is below the bottom goal post ($y < 36$), the y-distance is calculated as $y_{dist} = 36 - y$. If the shot is above the top goal post ($y > 44$), the y-distance is calculated as $y_{dist} = y - 44$. If the shot is between the goal posts, the y-distance is considered to be zero. The Euclidean distance is then calculated using the formula $d = \sqrt{x_{dist}^2 + y_{dist}^2}$.

3.3.3 Final Datasets

The preprocessing results in two parallel datasets, one for male and one for female competitions, each structured identically. They exclusively comprise shot events from

open play during regular time, ensuring homogeneity and consistency in gameplay variables. Each dataset includes:

- All relevant shot events from their respective gender competitions
- Consistent gameplay context (open play, regular time)
- Comprehensive coverage across all available leagues and tournaments
- Additional calculated features: angle, distance
- Binary labels (goal, no-goal), along with
- xG values calculated from StatsBomb for each shot

This approach ensures that:

Gender-specific characteristics: Each dataset accurately reflects the specific dynamics of male or female football respectively.

Consistency: Having the same data structure facilitates direct comparisons between male and female football.

Data Richness: By including all available competitions and matches, we maximize the diversity and volume of our dataset, enhancing the robustness of our xG models.

Representativeness: The inclusion of various leagues, tournaments, and seasons ensures that our models are trained on a wide spectrum of football contexts, improving their generalizability.

Comparative Potential: The parallel processing of male and female data, while maintaining their separation, sets the foundation for nuanced comparative analyses, a key objective of this study.

Enhanced Features: With the addition of angle, distance, and binary labels, along with xG values from StatsBomb, we can perform more detailed benchmarking and model evaluation.

By maintaining parallel but separate datasets, we enable the development of a robust xG model using the more extensive male dataset, while still allowing for comparative analysis between genders. This structure is crucial for enhancing the accuracy of the xG model trained on men's football data and for addressing our research objectives. It allows us to explore potential differences in shot quality and goal-scoring probabilities between

men's and women's football, and to conduct comprehensive data exploration and comparative analysis.

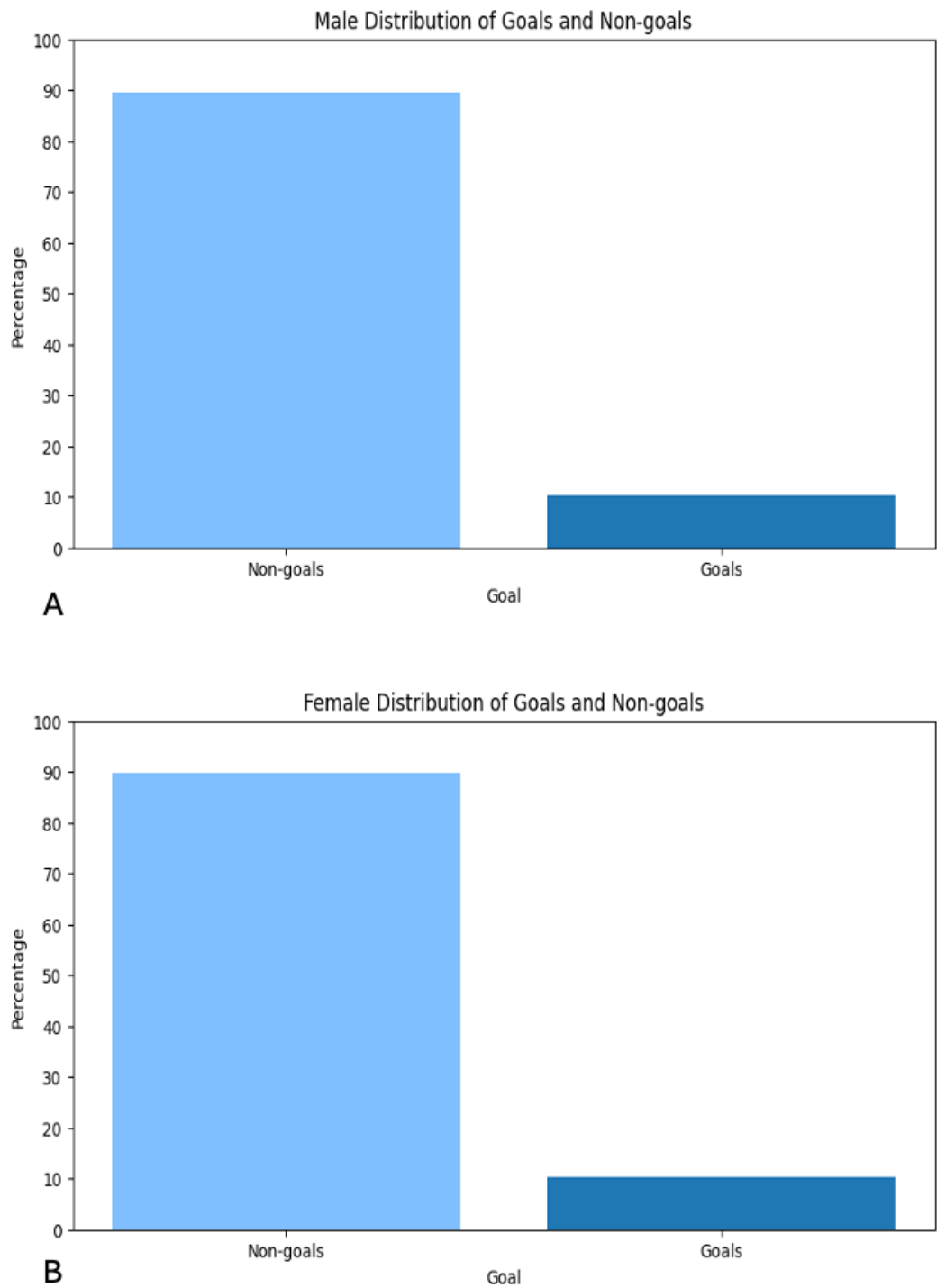
4. Exploratory Data Analysis

In this section, we delve into the exploratory data analysis of our dataset, focusing on key aspects such as the distribution of the label class, goal rate density, shot frequency maps, and historical shot distance analysis. This analysis aims to uncover patterns and insights that will inform the development and evaluation of our xG models.

4.1 Distribution of the Label Class

Figure 5 below illustrates the distribution of goals and non-goals in the dataset for both male and female players. As depicted, there is a significant class imbalance in both distributions, which are almost identical. Specifically, the number of non-goal events far exceeds the number of goal events. This class imbalance poses a challenge for predictive modeling, as models might become biased toward predicting the majority class (non-goals), thereby neglecting the minority class (goals). This distribution highlights the necessity of using evaluation metrics robust to class imbalance. Identifying the right metrics helps ensure that the model not only correctly identifies the more frequent non-goal events but also accurately predicts the less frequent goal events.

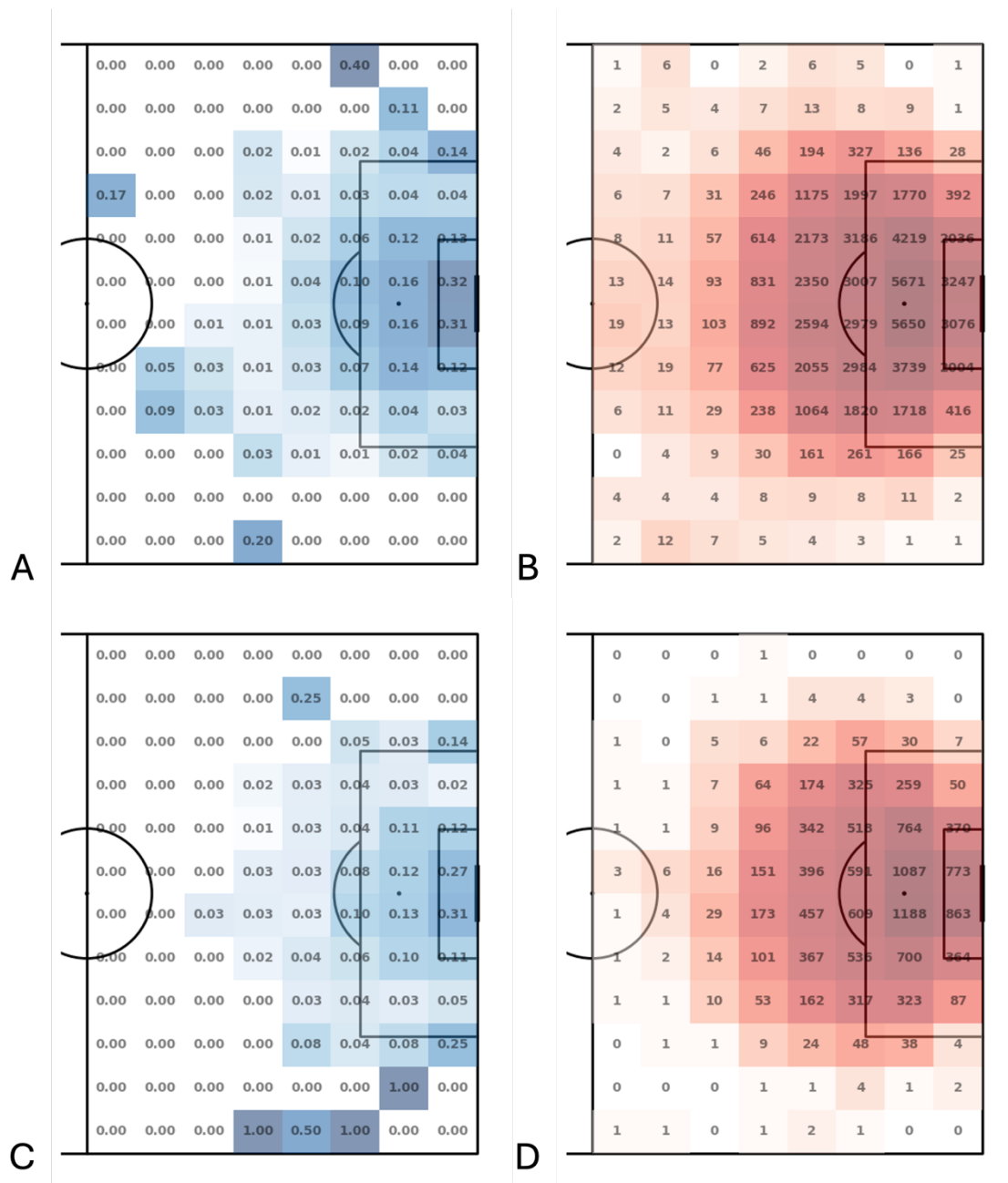
Figure 5: Comparative goals distribution between male and female; (A) male distribution; (B) female distribution



4.2 Goal Rate Density and Shot Frequency Map

The Figure 6 below compares the goal rate and shot density between men and women. They visually represent where shots are most likely to result in goals and where shots are more frequently taken, with darker shades indicating higher goal rates and shot frequencies. Note that the data for women is much smaller, so the scales have been adjusted accordingly.

Figure 6: Comparative goal rate map and shot frequency map between male and female ; (A) male goal rate; (B) male shot frequency; (C) female goal rate (D) female shot frequency



Subplot A: Male goal rate map

The heatmap illustrates the goal rate for male players, with higher values concentrated in the central areas close to the goal. As the distance from the goal increases and the position becomes less central, the values diminish.

The first noticeable feature is the presence of very high percentage values far from the goal on either side. Comparing these with the number of shots taken from those positions reveals that they resulted from a single goal scored in a few attempts, clearly indicating they are outliers.

The highest goal rates outside of obvious outliers are observed directly in front of the goal inside the six-yard box or goal area (see Appendix 1), with probabilities around 30%. This indicates that even shots perceived as the easiest have only approximately one-in-three chance of being scored, highlighting that there are no truly easy goals.

Subplot B: Male shot frequency

The heatmap displays the shot frequency for male players. The higher values appear in the central area, inside the penalty area and also at the edge of it. A lot of shots are still taken centrally outside the penalty area around the penalty arc ranging between 3% to 10% with the lowest value for the furthest shots. When comparing with the goal rate heatmap we notice that generally players definitely favor shooting in positions that have higher chances to score but will still take a considerable amount of shots.

The highest frequencies appear in the central area, close to the goal. Unlike the goal rate, the highest values are not inside the six-yard box but just outside of it, within the penalty area. This suggests that while these positions have the highest chance of scoring, they are harder to reach with the ball, resulting in fewer shots from these areas.

Subplot C: Female Goal Rate

The heatmap shows the goal rate for female players, with significant rates observed in the central and slightly off-center zones close to the goal. Similar to male players, there are outliers on each side of the goal resulting from a few attempts. The highest goal rates, excluding these outliers, are directly in front of the goal, indicating common effective shooting zones for both men and women.

Subplot D: Female Goal Frequency

The heatmap illustrates the shot frequency for female players with the highest frequencies found in the central area. The spread around the goal is more concentrated

compared to men which indicates fewer attempts from very far distances. This forms a similar shape but with a slightly narrower radius than for men.

Combined Insights

A radial heatmap system could better represent goal-scoring opportunities by accounting for the natural circular dispersion around the goal, highlighting how distance and angle affect scoring likelihood. Conversely, a grid probably fails to account for angular variations in scoring opportunities.

Both male and female players exhibit higher goal rates and shot frequencies the closer to the goal and the more central with the exception of the penalty mark having a higher shot frequency than the six-yard box. Male players display a slightly broader distribution of shooting zones.

These patterns suggest similar shooting strategies between male and female players which can be broken down into key observations:

1. **High-Probability Zones:** The areas with the highest goal rates are concentrated close to the goal, particularly within the six-yard box. The darkest shades in these zones indicate a significantly higher probability of scoring, which aligns with the intuitive understanding that shots taken from closer distances are more likely to result in goals.
2. **Central Areas:** Shots taken from central positions inside the penalty area also show higher goal rates. The probabilities in these regions, although lower than those directly in front of the goal, still reflect a substantial likelihood of scoring compared to other areas of the pitch.
3. **Wide Areas:** The goal rates diminish considerably as the shot locations move towards the wide areas of the pitch. This decline is evident in the lighter shades seen along the sides of the penalty area and further out towards the touchlines.
4. **Distant Shots:** The goal rate is very low for shots taken from outside the penalty area, as indicated by the lightest shades on the map. This reflects the general difficulty of scoring from longer distances, where the angle and power required make successful shots less likely.

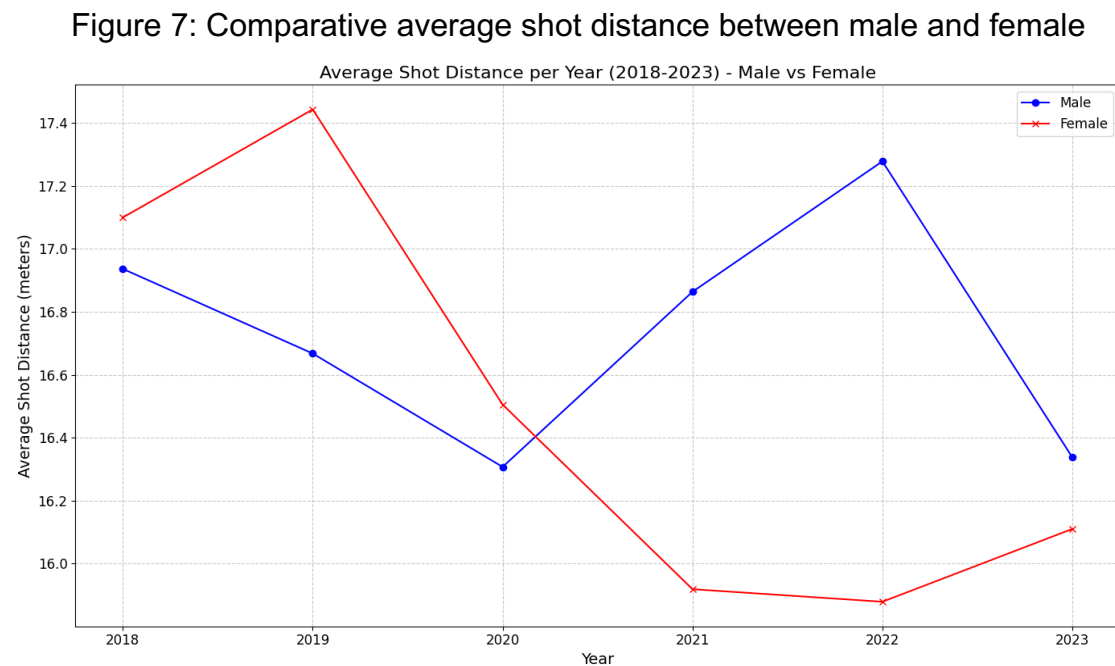
4.3 Historical Shot Distance Analysis

In this section, we analyze the historical trends in shot distances for both male and female players. By examining these trends, we aim to uncover patterns and differences in

shooting behavior over time, providing deeper insights into how shot selection and strategies have evolved.

4.3.1 Shot Distance Comparison Between Male and Female

The Figure 7 displays the average shot distance for males and females per year. The time frame was chosen as data for female games prior to 2018 was unavailable.



Trends

For male players: the average shot distance is decreasing from 2018 to 2020, followed by an increase until 2022, and a sharp decline in 2023.

For female players: the average shot distance peaks in 2019 with a slight increase from 2018 and then sharply decreases to reach a minimum in 2022 with a slight recovery towards 2023.

Key Points

- Peak distance: female shot distance peaked in 2019 (~17.4 meters), males in 2022 (~17.2 meters).
- Lowest Averages: female saw their lowest averages in 2021, males in 2020
- Divergence: significant differences from 2020 onwards with males increasing their shot distance while females stabilized at lower levels.
- Convergence: both lines converge towards ~16.2 meters in 2023.

The spectrum of average shot distances for males varies from approximately 16.3 meters to 17.3 meters. Alternatively, the range for females is slightly broader, spanning from around 15.8 meters to 17.4 meters. Contrary to initial expectations of larger disparities between male and female shot distances, the observed difference is only about 1-2 meters.

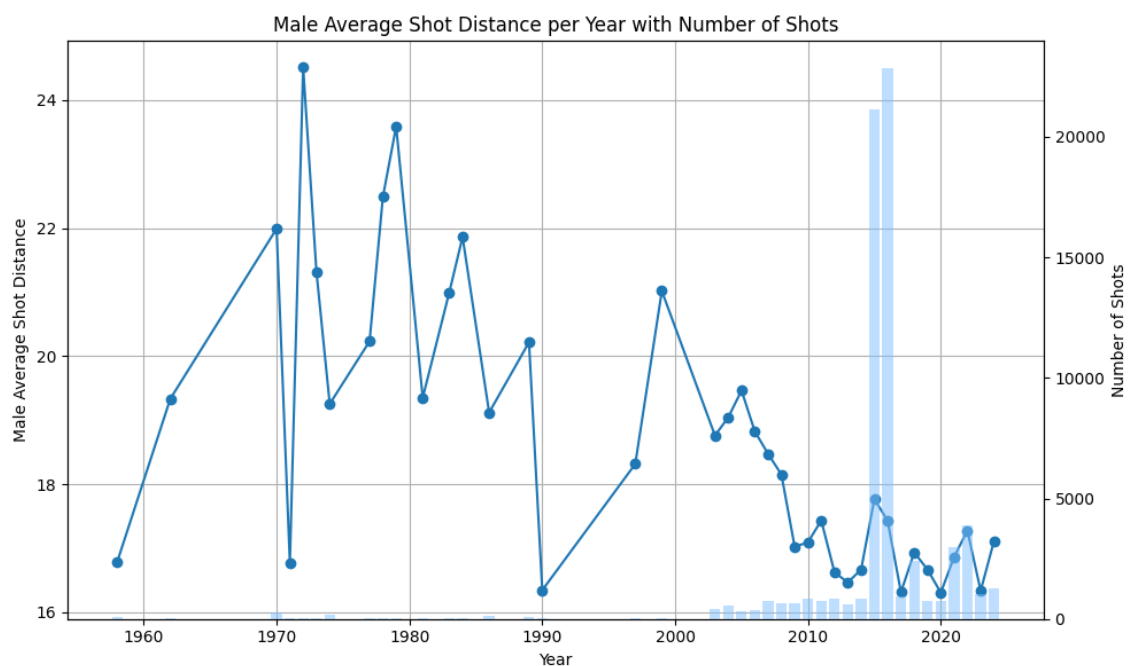
Limitations

While there appears to be a decreasing trend in the average distance for females over the years, the trend for males is less clear. The time frame of the data collection is likely too short to make definitive conclusions about these trends.

4.3.2 Historical Analysis of Male Shot Distance Trends

Since our comparison was limited by the amount of female data available, Figure 8 focuses solely on the male data to extend our time frame.

Figure 8: Male average shot distance per year with number of shots



Long-term trends

- 1960s-1970s: High volatility with peaks around 24 meters and significant dips.
- 1980s-1990s: Fluctuating but generally declining average shot distances.
- 2000s-Present: Continued decline with lower average shot distances, especially post-2010.

Number of shots

- Sharp increase in the number of shots post-2010, peaking significantly in the last decade (Bate 2017).
- The increase in the number of shots contrasts with the declining average shot distances, indicating a possible change in play style or strategy.

Conclusion of historical shots comparison between male and female

The analysis reveals fluctuating average shot distances over time, with notable changes in recent years. Male players exhibit a long-term declining trend. While data from before the 2000s may be insufficient for definitive conclusions, data from the 2000s onward indicates a steady decrease, converging around 16-17 meters since 2017. This period corresponds with the rising popularity of the term expected goal (Bate 2017), which might explain why teams began to favor higher-quality chances that are closer to the goal.

4.3.3 Recent Analysis of Shot Distance Distribution for Male Players

This analysis in Figure 9 will focus on the boxplot illustrating the distribution of shot distances for male players over multiple 2-year periods from 2004 to 2024. To improve the readability of the plot and focus on the central distribution of the data, we have applied a strategy to limit the y-axis range to the values between the 1st and 99th percentiles. This approach effectively excludes extreme outliers, allowing us to better visualize and interpret the key trends and patterns in shot distances across different periods.

The Figure 9 demonstrates the distribution of shot distances for male players, with each box representing a 2-year period. The red line within each box indicates the median shot distance, while the blue triangles represent the mean shot distance.

From 2004/2005 to 2024/2025, both the median and mean shot distances appear relatively stable, generally hovering around 17-20 meters. This consistency suggests that the typical shot distance for male players has not significantly changed over the past two decades. The interquartile range (IQR), depicted by the boxes, shows slight variations but remains fairly consistent, indicating that the central tendency of shot distances has not undergone dramatic changes.

The whiskers are purposely limited to the 1st and 99th percentiles, excluding the extreme values to improve readability. By cutting off data beyond these percentiles, we can zoom in on the central range of shot distances, making the graph more focused and easier to

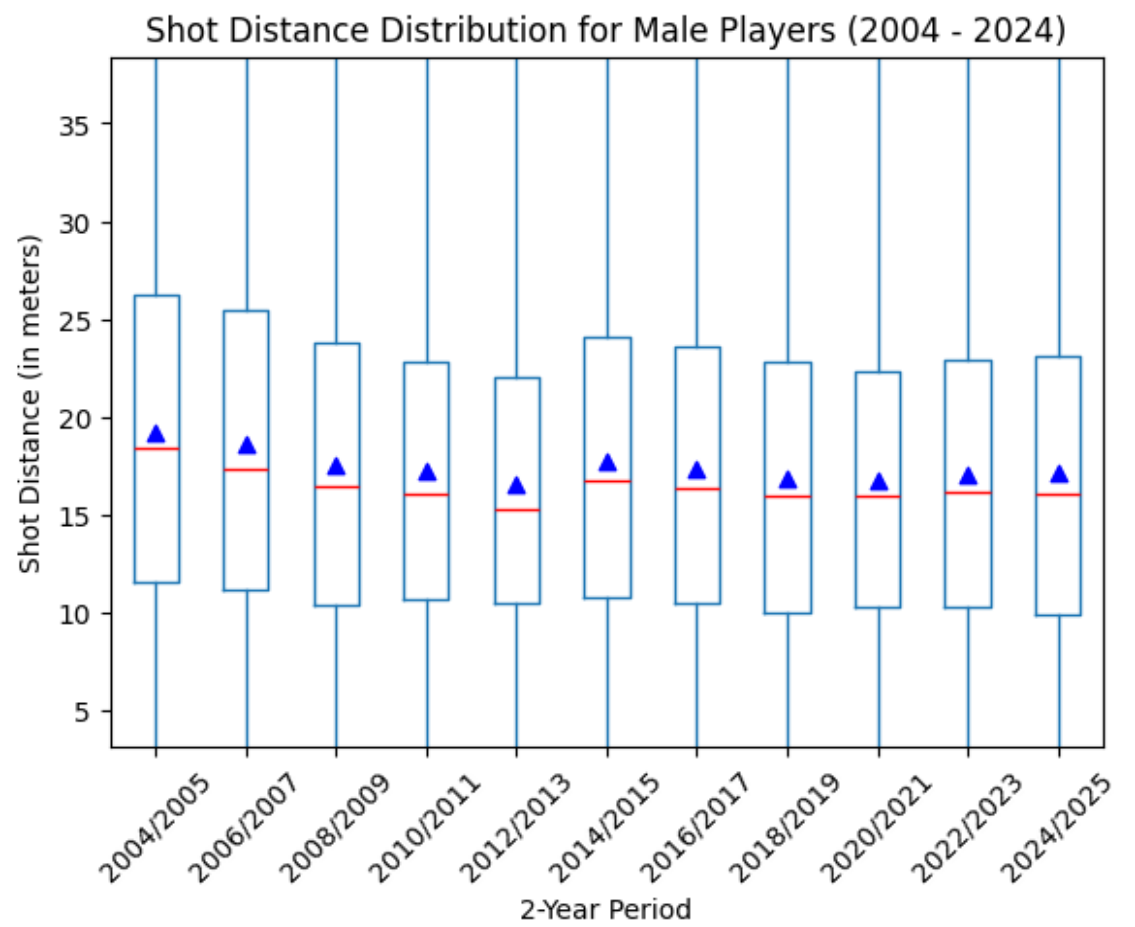
interpret. This range suggests that while there are some variations in shot distances, most shots fall within a consistent range over the years.

When compared to the historical analysis in Figure 8, it can be observed that the earlier decades (1960s and 1970s) exhibited high volatility in average shot distances, with peaks around 24 meters and significant dips. The fluctuations decreased in the 1980s and 1990s.

In contrast, the data from 2004 to 2024 shows more stability, with less fluctuation in average shot distances. This stability aligns with the strategic shift towards higher-quality, closer shots, as indicated by the rising popularity of metrics like xG in modern football analysis.

In conclusion, the analysis of the boxplot from 2004 to 2024 reveals a consistent pattern in shot distances for male players. This stability contrasts with the historical fluctuations observed in earlier decades, aligning with modern strategic shifts towards higher-quality shot opportunities.

Figure 9: Shot distance distribution for male players (2004-2024)



5. Methodology

This thesis addresses the evaluation challenges associated with xG metric. We achieve this by comparing different machine learning models to predict the likelihood of a shot resulting in a goal (xG) based on shot characteristics. The methodology involves several key components: binary classification, probabilistic output generation, and performance benchmarking.

5.1 Training

To train our models, we will focus on key calculated features: shot angle, shot distance, and the binary outcome of goal or no goal (represented as 1 or 0), which will serve as the label. These features are critical for accurately predicting expected goals (xG). Additionally, we will benchmark our model's performance against the xG values provided by StatsBomb. This benchmarking will allow us to compare our results with an established standard, ensuring the reliability and validity of our models. By concentrating on these specific features and incorporating a robust comparison framework, we aim to develop highly accurate and reliable xG models.

Table 2: Sample of training data

Angle	Distance	Goal
17.14	26.51	0
22.57	19.60	0
15.54	29.30	0
17.39	22.73	0
10.85	18.04	0

5.1.1 Binary Classification

We apply binary classification techniques to predict shot outcomes based on the derived features. The models used include logistic regression, random Forest, and XGBoost (Extreme gradient boosting), each chosen for their unique strengths in handling different types of data and relationships. All models are implemented using the Scikit-learn library, a widely-used machine learning toolkit in Python (*scikit-learn: machine learning in Python — scikit-learn 1.5.1 documentation*).

Logistic regression: A simple model that is effective for binary classification problem. It estimates the probability that a given input point belongs to a certain class, making it suitable for predicting the likelihood of a goal based on shot characteristics (*Logistic Regression — scikit-learn 1.5.1 documentation*) (Cox 1958).

Random forest: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. It is robust to overfitting and can handle a large number of input features, making it ideal for capturing complex interactions between shot angle and distance (*Random Forest Classifier — scikit-learn 1.5.1 documentation*) (Breiman 2001).

XGBoost: An optimized gradient boosting algorithm that is highly efficient and scalable. It builds an ensemble of trees sequentially, where each tree corrects the errors of the previous ones. XGBoost is known for its high performance and accuracy, making it suitable for capturing intricate patterns in the data (*Gradient Boosting Classifier — scikit-learn 1.5.1 documentation*) (Chen 2016).

5.1.1.1 Hyperparameters Tuning

The objective is to identify the best set of hyperparameters to improve the forecasting accuracy of each model in predicting goals and non-goals. To determine the best parameters for each model, cross-validation combined with a grid search method was employed to explore various parameter combinations. 20% of the dataset was reserved as a test set, and 5-fold cross-validation was performed on the remaining 80%. The F1 score metric, which is more robust to class imbalance, was used as the parameter for GridSearchCV. This methodology facilitates the determination of the best hyperparameters for each model and reduces the risk of overfitting.

Table 3: Logistic regression parameters

Parameters	Values
C	0.001, 0.01, 0.1, 1, 10, 100
max_iter	500
tol	0.0001

Table 4: Random forest parameters

Parameters	Values
n_estimators	50, 100
max_depth	None, 5, 10
min_samples_split	2, 5, 10

Table 5: XGBoost parameters

Parameters	Values
n_estimators	100, 200, 300
max_depth	3, 5, 7, 9
learning_rate	0.01, 0.05, 0.1
subsample	0.5, 0.7, 1.0
colsample_bytree	0.5, 0.7, 1.0

For logistic regression, various regularization strengths were tested to prevent overfitting by penalizing large coefficients. The model was configured to run for a maximum of 500 iterations, which limits how long the optimization process can continue. A tolerance level was set to determine when to stop the optimization process, based on minimal changes in the loss function, ensuring the model doesn't continue running when improvements are negligible.

In the case of random forest, the number of trees in the forest (estimators), the maximum depth of each tree, and the minimum number of samples required to split a node were varied. More trees generally improve performance but require more computational power. The depth of the trees impacts their complexity; deeper trees can capture more intricate patterns but may overfit. The minimum samples split parameter affects how granular the splits in the decision trees are, influencing the model's complexity and performance.

For XGBoost, parameters such as the number of boosting rounds (estimators), the depth of the trees, the learning rate, the subsample ratio, and the fraction of features used to build each tree (column sample by tree) were explored. The learning rate adjusts how much the model learns from each boosting round, with lower rates often leading to better generalization but requiring more rounds. The subsample ratio introduces randomness by using only a portion of the training data in each round, which helps prevent overfitting. The column sample by tree parameter further helps to control overfitting by ensuring each tree doesn't use all available features.

Additionally, a dummy classifier using the most frequent class was included as a baseline for comparison. This simple model always predicts the majority class, providing a reference point to measure the effectiveness of the more sophisticated models.

5.1.2 Probabilistic Output

The models were trained to provide probabilistic outputs, representing the likelihood of a shot being a goal. Logistic regression is particularly suitable for this task as it naturally

outputs probabilities. Random forest and XGBoost, while typically used for classification, can also be configured to provide probabilistic predictions. These probabilities are directly interpretable as xG values.

In scikit-learn, the `predict_proba` method is utilized to obtain probabilistic outputs. This method returns the probability estimates for each class, with the second column representing the probability of the positive class (e.g., the probability of a goal).

Random Forest achieves probabilistic predictions by aggregating the predictions from multiple decision trees. Each tree in the forest outputs a class probability distribution based on the proportion of training samples of each class in the leaf nodes. For a given instance, the probability of the positive class (goal) is calculated for each tree, and these probabilities are then averaged to produce the final probabilistic prediction.

XGBoost can also provide probabilistic predictions. This model builds an ensemble of decision trees sequentially, where each tree attempts to correct the errors of the previous trees. Initially, a base score is computed based on the log-odds of the positive class using the overall proportion of positive instances in the training data. Each subsequent tree contributes an adjustment to the log-odds score. The `predict_proba` method in XGBoost outputs the probability of each class by transforming these cumulative log-odds into probabilities using the logistic function. This transformation converts the aggregated log-odds into a probability value, enabling the interpretation of the model outputs as probabilities.

6. Evaluation

In football analytics, evaluating the performance of xG models is a critical step. The evaluation process determines how well these models predict the likelihood of a goal being scored from a particular shot. However, the choice of evaluation metrics can significantly influence the perceived efficacy of a model. In contexts with imbalanced classes, such as those often found in football data where goals are far less frequent than non-goals, traditional metrics like accuracy may not be the best indicators of performance. This section delves into the various metrics that can provide a more comprehensive understanding of model performance, highlighting the importance of selecting appropriate metrics in football analytics.

6.1 Model Benchmarking

To evaluate the performance of the models, a dummy classifier that always predicts the majority class (no goal) was used as a benchmark. This simple baseline provided a reference point to measure how the more sophisticated models were performing. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were utilized to assess the models. These metrics ensured a comprehensive evaluation of the models' predictive capabilities.

6.2 Performance Metrics

Selecting the right metrics is pivotal in assessing the effectiveness of predictive models, particularly in football analytics. Metrics provide quantitative measures that help in comparing different models and understanding their strengths and weaknesses. Given the class imbalance inherent in football shot data, where non-goals vastly outnumber goals, relying solely on accuracy can be misleading. Therefore, a variety of metrics should be considered to capture different aspects of model performance, including how well the model can identify true positives (goals) and true negatives (non-goals).

Table 6: Performance of models across various metrics

Model	Accuracy	Precision	Recall	F1 score	AUC ROC	R2 score
<i>Logistic Regression</i>	0.894	0.728	0.051	0.096	0.751	0.104
<i>Random Forest</i>	0.859	0.248	0.139	0.178	0.648	-0.158
<i>XGBoost</i>	0.891	0.536	0.071	0.125	0.718	0.065
<i>Dummy Classifier</i>	0.890	0.0	0.0	0.0	0.5	-0.123

Accuracy

The use of accuracy as a performance metric is not suitable in this context due to the significant class imbalance in the label distribution. When GridSearchCV was applied, it revealed minimal differences between the models and the dummy classifier. This observation suggests that the models are not outperforming the dummy classifier, even when optimized parameters are used. Consequently, accuracy, as expected does not appear to be the most effective metric for evaluating an expected goal model. It may be beneficial to consider instead other metrics that are more robust to class imbalance for evaluation, such as Precision, Recall, F1 score, or Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics could provide a more nuanced understanding of the model's performance.

Precision

Precision is critical in this context because it measures the accuracy of positive predictions made by the model. Given the significant class imbalance in the label distribution, precision ensures that the model does not predict too many false positives. The table shows that logistic regression achieved the highest precision (72.8%), followed by XGBoost (53.6%) and random forest (24.8%), indicating that logistic regression was the most effective at making accurate positive predictions. The models demonstrate a much higher precision compared to random guessing, which would yield a precision equal to the distribution of goals, around 10%.

Recall

The importance of recall lies in its ability to capture all actual positive cases, which is key in an expected goal model. The recall values are notably low for all models, with random forest achieving the highest recall (13.9%), followed by XGBoost (7.1%) and logistic regression (5.1%). These values are much lower than what a random classifier recall value would obtain, which is 50% in a binary classification problem. While these low recall values suggest that the models are not identifying all actual goals, this does not necessarily indicate poor performance in an expected goals (xG) model. Since many goals occur with predicted xG values below 0.5, low recall at this threshold reflects the difficulty of capturing goals that were low-probability events. The issue may stem more from the choice of threshold and the inherent class imbalance, rather than the model's ability to estimate probabilities effectively. As a result, it is important to focus on evaluation metrics that better reflect the probabilistic nature of xG models, such as log loss, Brier score, or calibration, rather than relying solely on recall.

F1 Score

The F1 score is relevant as it provides a balance between precision and recall, offering a single metric that considers both false positives and false negatives. The F1 scores in the table reflect the poor balance between precision and recall, with random forest achieving the highest F1 score (17.8%), followed by XGBoost (12.5%) and logistic regression (9.6%). These results indicate that while random forest balances precision and recall slightly better, all models have significant room for improvement.

AUC-ROC

The relevance of AUC-ROC lies in its ability to evaluate the model's performance across various threshold settings, providing an overall measure of its discriminative power. The AUC-ROC values are relatively close, with logistic regression achieving the highest (0.751), followed by XGBoost (0.718) and random forest (0.648). These values indicate that logistic regression performs slightly better in distinguishing between goals and non-goals, although all models show some level of competence.

R² Score

The R² score measures the proportion of variance in actual goals that can be explained by the predicted xG values, indicating the model's explanatory power. In this context, a higher R² score indicates that the model's predictions align more closely with the observed data. The R² scores are relatively low, with logistic regression achieving the highest (0.104), followed by XGBoost (0.065) and random forest (-0.158). Random forest has a negative R² score indicating that it performs worse than a mean-based model. These results suggest that while the models provide some explanatory power, there is significant room for improvement.

Conclusion of performance metrics analysis

The performance metrics indicate that while the models demonstrate some ability to predict goals, they are significantly impacted by class imbalance. Precision and AUC-ROC values are relatively high, suggesting that the models perform well in making accurate predictions and distinguishing between goals and non-goals. However, the low recall and F1 scores reflect the difficulty in capturing true positive cases, meaning that many actual goals are not being identified based on a binary threshold. This issue is further emphasized by the low R² scores, indicating limited explanatory power in the models' predictions.

The high AUC-ROC values show that the models can distinguish between classes across various thresholds, but this does not necessarily translate into practical performance, given that many goals are associated with lower xG values. Additionally, the precision scores reveal that when the models do predict goals, these predictions are often correct. However, the low recall may be more reflective of the choice of threshold and the probabilistic nature of xG rather than a true weakness of the models themselves.

In summary, although the models show strengths in certain metrics, focusing too heavily on recall and F1 scores may not fully capture their performance. Metrics that better align with probabilistic outputs, such as log loss, Brier score, and calibration, could provide a more comprehensive evaluation. Addressing class imbalance and refining model evaluation approaches may further enhance the reliability and robustness of the models, leading to more accurate and meaningful predictions of expected goals.

Recommendations

To enhance the predictive performance and reliability of xG models, several recommendations can be made based on the evaluation results and the challenges posed by class imbalance. The complexity of the model does not appear to be the limiting factor, as the various models perform similarly. Instead, refining the data could significantly improve the model's performance. Here are some suggestions:

Addressing Class Imbalance

Class imbalance poses a significant challenge in developing high-performing models. This issue can significantly affect the accuracy and reliability of the model's predictions. To effectively tackle class imbalance, there are a range of strategies and techniques designed to mitigate its impact. By addressing this challenge, we can improve the overall performance and robustness of the model. The following methods could be impactful:

Resampling techniques: implement resampling methods such as oversampling the minority class or undersampling the majority class to create a more balanced training dataset. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) can also be employed to generate synthetic samples of the minority class (Chawla 2002).

Class weight adjustment: adjust the class weights in the models to give more importance to the minority class. This can be done by setting higher weights for the goal class, which helps the model focus more on correctly predicting goals (Thakur 2020).

Stratified K-Fold Cross-Validation: Ensures that each fold in cross-validation has the same proportion of classes as the original dataset. This helps in providing a more balanced evaluation during model training and validation (Moreno-Torres 2012).

Implementation: In scikit-learn, StratifiedKFold can be used to maintain class proportions.

Enhance Feature Engineering:

Add more context and information, incorporate additional features that may provide more predictive power. These could include player-specific metrics, contextual information about the match, or more advanced statistics.

Validate with Real-World Data:

The models should be tested against actual match data by comparing the cumulative xG with the real-world performance. This process involves using recent and relevant datasets to ensure the models remain current and applicable. Based on the results, the models should be iteratively refined to improve their accuracy and reliability.

Utilize Robust Evaluation Metrics:

Precision, recall, F1 score, and AUC-ROC provide a more comprehensive evaluation of the models' ability to predict goals, especially in the context of imbalanced data. Unlike accuracy, which can be misleading, these metrics offer a more thorough assessment of model performance.

Calibration Metrics:

Include calibration metrics such as Brier score to assess how well the predicted probabilities of goals correspond to the actual outcomes, ensuring the models provide accurate probability estimates.

By implementing these recommendations, the models can be better equipped to handle the challenges posed by class imbalance, leading to more accurate and meaningful predictions of expected goals. This will ultimately enhance the utility of xG models, providing deeper insights into the factors that influence shot success.

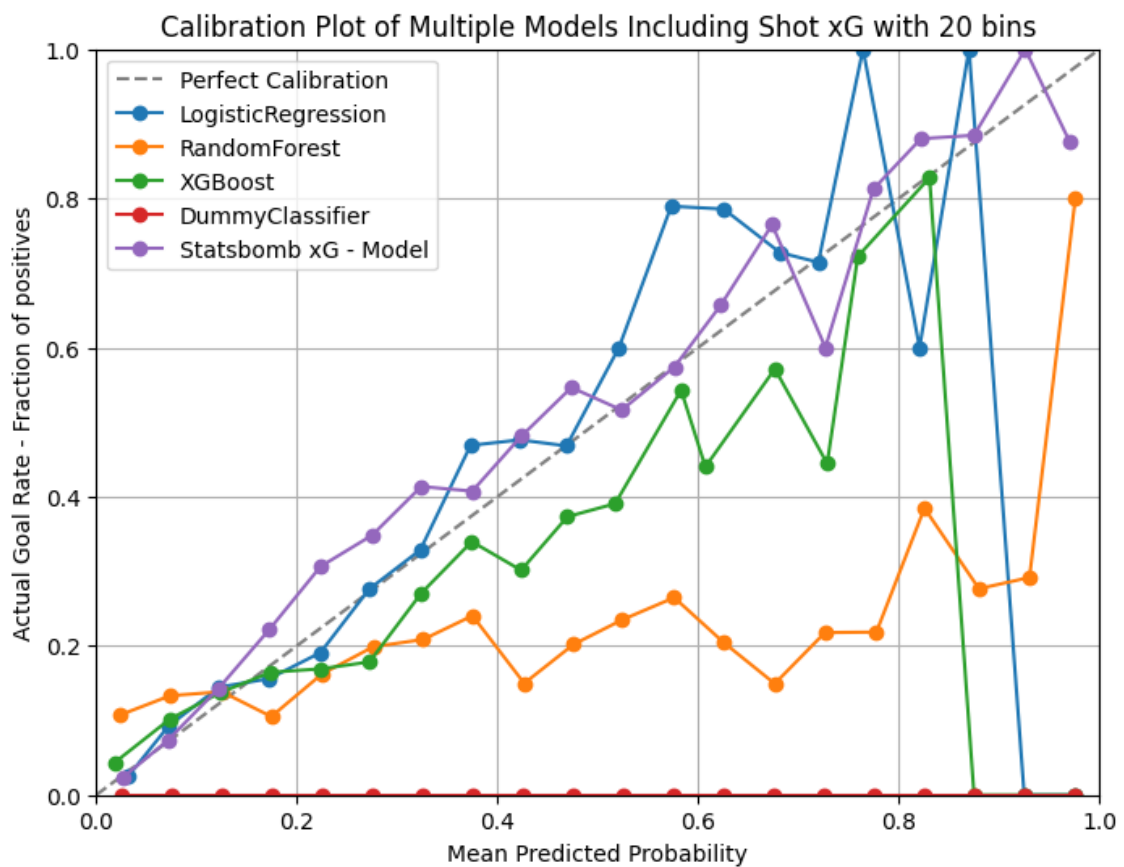
7. Results exploration

In this section, we delve into the performance and predictions of various xG models. By examining calibration plots, predicted probability distributions, and scatter plots, we aim to uncover insights into the accuracy, reliability, and tendencies of these models. This exploration will help identify strengths and weaknesses, guiding future improvements in xG model development.

7.1 Calibration Plot Analysis

The calibration plot is a graphical representation that compares the predicted probabilities of a binary outcome (in this case, the probability of scoring a goal) with the actual observed outcomes. This comparison helps assess the accuracy and reliability of probabilistic predictions made by different models. Calibration plots are crucial for diagnosing and improving model performance by visually displaying how well predicted probabilities match actual outcomes. As discussed by (ElHabr 2023), calibration plots are useful for evaluating the calibration of probabilistic models and can highlight areas where the model consistently over-predicts or under-predicts events.

Figure 10: Calibration Plot of Multiple Models with 20 Bins



Grouping Predictions into Bins

The plot uses 20 bins to group the predicted probabilities. Each bin represents a range of predicted probabilities. The range of predicted probabilities is from 0 to 1 (0-100%) with intervals of 0.05 (5%). Each bin contains predictions that fall within that range.

Calculating Mean Predicted Probability

For each bin, the mean predicted probability is calculated. This is the average of all predicted probabilities that fall into that bin. For instance, if a bin contains predictions of 0.05, 0.06, and 0.07, the mean predicted probability for that bin would be $(0.05 + 0.06 + 0.07) / 3 = 0.06$.

Calculating Actual Goal Rate

Actual Goal Rate: For each bin, the actual goal rate is calculated. This is the proportion of true positive instances (actual goals) within that bin.

Continuing the example, if there are 4 predictions in a bin and 2 of them are actual goals, the actual goal rate for that bin would be $2/4 = 0.5$.

Plotting the Calibration Curve

The calibration curve is plotted by placing the mean predicted probability on the x-axis and the actual goal rate on the y-axis for each bin.

If a model is perfectly calibrated, the points will lie on the diagonal line, indicating that the predicted probability matches the actual goal rate.

Comparing Multiple Models

Multiple Models: The plot includes curves for different models to compare their calibration.

Each line shows how well the predicted probabilities of each model correspond to the actual goal rates across different probability ranges.

Interpretation and Discussion

The calibration plot provides insights into the reliability of the models' probabilistic predictions:

Logistic regression: the blue line shows that logistic regression predictions are relatively well-calibrated, as they stay close to the diagonal line, especially at lower probabilities. This indicates that the predicted probabilities match the actual goal rates

reasonably well. Nevertheless, from 0.5 onwards, the predicted probabilities begin to fluctuate. In the 0.5 to approximately 0.7 range, the model tends to overestimate the likelihood of a goal, as the predicted probabilities are generally below the actual goal rate. Notably, a point close to 0.7 lies on the line of the actual goal rate, indicating accurate calibration at this specific probability. From 0.7 to 1.0, the predictions continue to fluctuate a lot and tend to underestimate and overestimate the likelihood of a goal. The bin with the highest predicted probabilities, which is close to 1, has a goal rate of 0, suggesting that this bin likely contains very few shots, if not a single point.

Random Forest: The orange line shows that the random forest model performs decently, like the other models, until about 0.3. However, from there until approximately 0.9, it significantly underestimates the likelihood of goals. Beyond 0.9, the model shows a sharp increase and gets closer to the actual goal rate. These fluctuations and deviations from perfect calibration indicate that the random forest model has very conservative predictions, leading it to consistently underestimate the likelihood of goals.

XGBoost: The green line shows that this model, which is the most complex among the three, is decently calibrated like the logistic regression up to 0.5. However, beyond this point, it deviates more from the actual goal rates. Unlike logistic regression, XGBoost consistently underestimates predictions. Notably, it has two bins very close to the actual goal rate at around 0.8 and then also drops close to 1.0. Comparable to logistic regression, this is likely due to a bin with few shots that are incorrectly predicted as goals.

Dummy Classifier: the red line is flat at zero, indicating that the dummy classifier always predicts the majority class (no goal) and has no discriminative power. It fails to provide useful probabilistic predictions.

StatsBomb xG Model: The purple line shows that the StatsBomb xG model has calibration akin logistic regression before 0.5 but also exhibits deviations, particularly at higher predicted probabilities. This model experiences fluctuations starting from 0.6 onwards, leading to both underestimations and overestimations. However, it remains notably closer to the actual goal rate than the models we trained, especially at higher probabilities, indicating better overall calibration.

Conclusion of the calibration plot analysis

The evaluation of various models for predicting goals reveals consistent patterns in calibration performance. Logistic regression, random forest, and XGBoost all demonstrate good calibration up to a predicted probability of 0.5, indicating that these

models are effective at representing shots with lower probabilities, which are more frequent. However, beyond the 0.5 threshold, the models begin to fluctuate significantly.

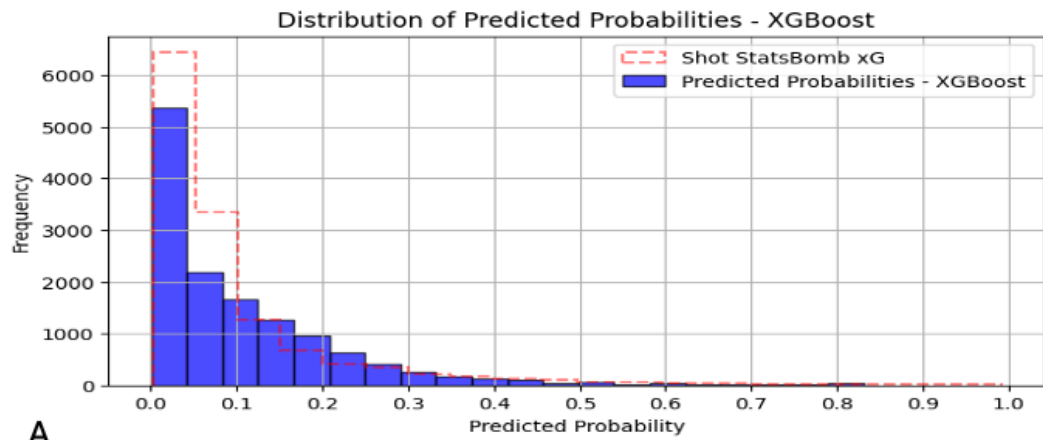
The Statsbomb xG model, although experiencing similar calibration issues, aligns more closely with actual goal rates compared to the trained models, indicating better overall calibration.

These findings suggest that the models are more reliable in predicting lower probability shots, which are more common in the dataset. Conversely, the models struggle to accurately represent higher probability shots due to their rarity, leading to greater fluctuations and deviations from actual goal rates. This underscores the need for enhanced data collection and model refinement to improve the representation and prediction of rarer, higher probability shots, ensuring more accurate and reliable expected goal models.

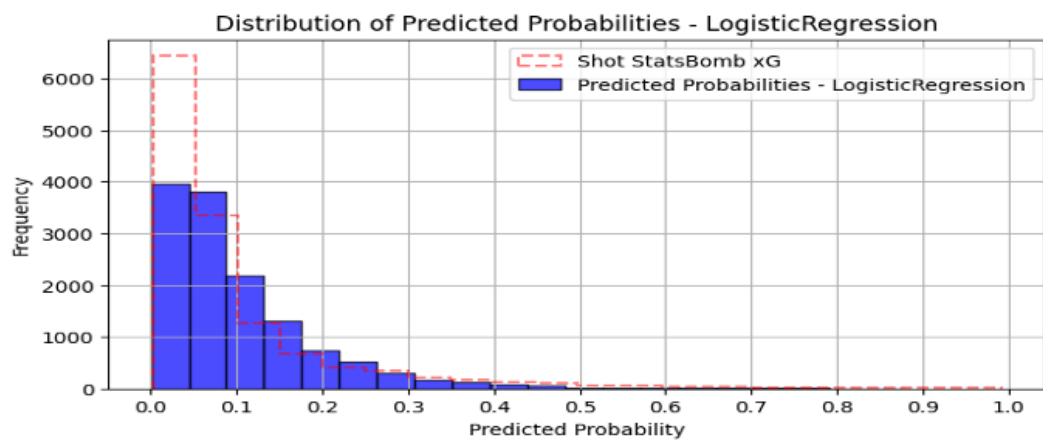
7.2 Analysis of Predicted Probabilities

The calibration plot revealed certain flaws in the xG models, particularly deviations from the actual goal rate at higher predicted probabilities. Consequently, examining the distribution of predicted probabilities within each bin of the plot is compelling to understand their composition and gain deeper insights.

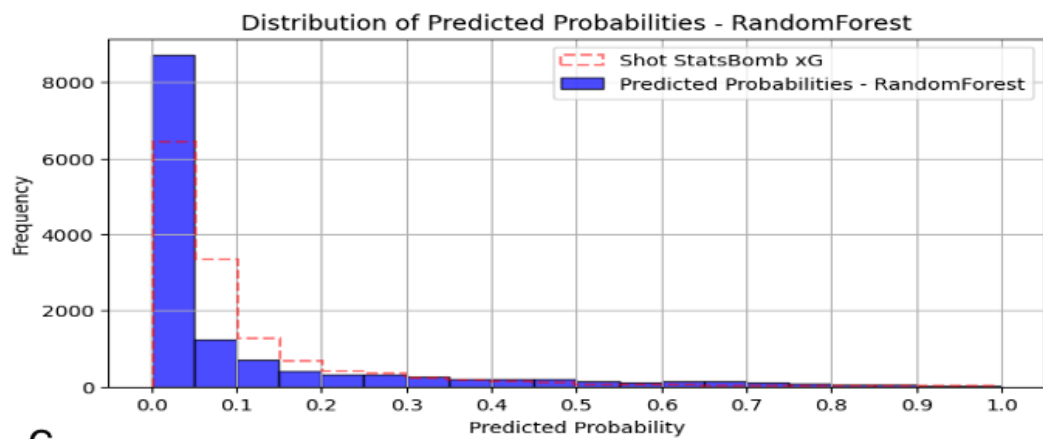
Figure 11: Distribution of predicted probabilities by different models; (A) XGBoost; (B) Logistic Regression; (C) Random Forest



A



B



C

Analysis and Discussion

Figure 11 presents the distribution of predicted probabilities by different models: (A) XGBoost, (B) Logistic Regression and (C) Random Forest. Each subplot displays the frequency of predicted probabilities on the y-axis and the predicted probabilities on the x-axis. The histograms represent the predicted probabilities by the respective models, while the red dashed lines indicate the Shot StatsBomb xG predicted probabilities for comparison.

The distributions of predicted probabilities for logistic regression, XGBoost, and random forest models exhibit striking similarities. In all three plots, the majority of predicted probabilities are concentrated at the lower end of the spectrum, predominantly between 0.0 and 0.2. This indicates that the models are conservative in their predictions, assigning low probabilities to most shots.

Alignment with StatsBomb xG:

- **XGBoost (Subplot A):**
 - The XGBoost model has a relatively higher frequency of predictions between 0.1 and 0.2 compared to StatsBomb xG, which predominantly predicts probabilities between 0 and 0.1.
 - This suggests that while XGBoost is generally accurate in predicting low-probability shots, it tends to slightly overestimate these probabilities compared to StatsBomb xG.
- **Logistic Regression (Subplot B):**
 - The Logistic Regression model also shows a higher frequency of predictions between 0.1 and 0.2, similar to XGBoost, but slightly lower than XGBoost.
 - Logistic Regression aligns closely with StatsBomb xG at very low probabilities (0 to 0.1) but then starts to overestimate slightly in the 0.1 to 0.2 range.
- **Random Forest (Subplot C):**
 - The Random Forest model has a higher concentration of predictions at the very lowest probabilities (0 to 0.1) compared to both XGBoost and Logistic Regression.

- Unlike the other models, Random Forest has fewer predictions in the 0.1 to 0.2 range compared to StatsBomb xG, indicating a more conservative approach in assigning higher probabilities to shots.

Deviation at Higher Probabilities

Beyond 0.2, the frequency of predicted probabilities drops sharply for all models. The deviations from the StatsBomb xG model become more pronounced:

- **XGBoost and Logistic Regression:**
 - Both models show a more gradual decline in predicted probabilities beyond 0.2 compared to StatsBomb xG.
 - They tend to underpredict probabilities in the higher ranges (0.3 and above), resulting in fewer high-probability predictions compared to StatsBomb xG.
- **Random Forest:**
 - The Random Forest model shows a sharper decline in predicted probabilities beyond 0.2.
 - There are significantly fewer high-probability predictions (0.3 and above) compared to StatsBomb xG, indicating a conservative bias in higher probability ranges.

Discussion

- **Model Conservatism:**
 - The conservative nature of the models, especially Random Forest, reflected in the concentration of low predicted probabilities, helps in reducing false positives but also limits their ability to predict high-probability goals accurately. This conservatism contributes to the fluctuations and deviations observed in the calibration plots, particularly at higher probabilities.
- **Data Distribution:**
 - The similarity in the distribution patterns across the three models suggests that the underlying data distribution is influencing the models' predictions. The rarity of high-probability shots leads to insufficient data

points in these ranges, causing the models to struggle with accurate predictions.

- **Model Comparison:**

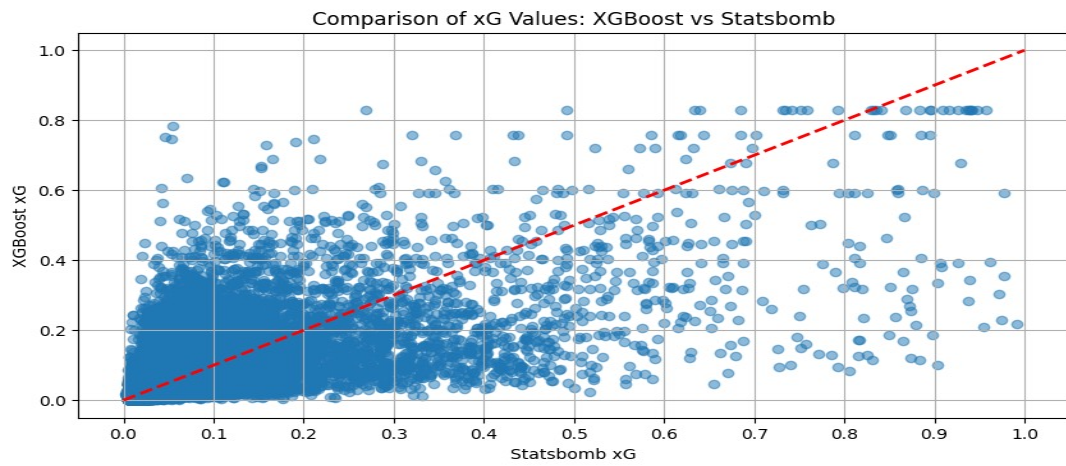
- While all three models demonstrate a general alignment with StatsBomb xG at lower probabilities, their behavior diverges at higher probabilities. XGBoost and Logistic Regression tend to overestimate low probabilities and underpredict higher probabilities. In contrast, Random Forest is more conservative overall, particularly in higher probability ranges.

Overall, the analysis underscores the importance of examining the distribution of predicted probabilities to identify potential areas for improvement and enhance the models' reliability in predicting goal probabilities.

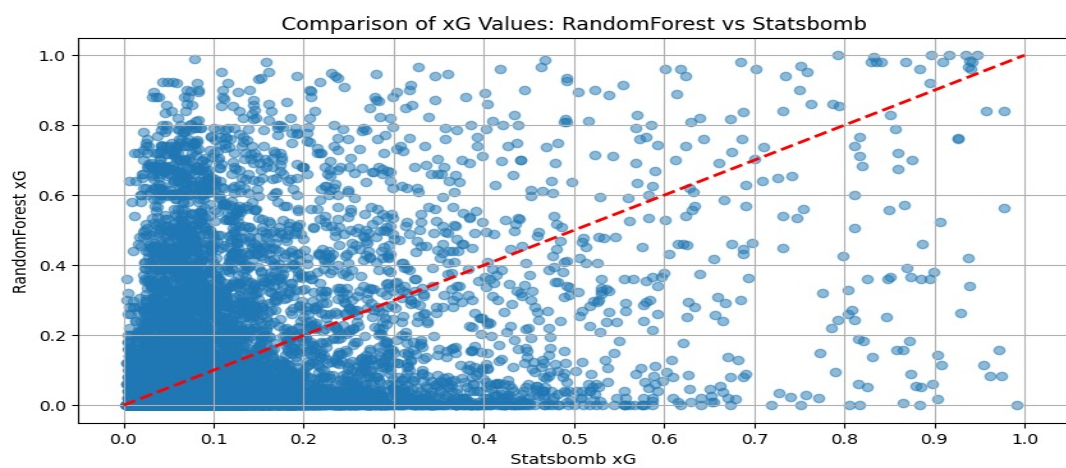
Figure 12: Scatter plots comparison of xG values between trained models and StatsBomb; (A) Logistic Regression; (B) XGBoost; (C) Random Forest



A



B



C

Figure 12 comprises three scatter plots comparing the predicted expected goals (xG) values generated by different machine learning models—logistic regression, XGBoost, and random forest—against the actual xG values provided by StatsBomb.

Trend Analysis:

- **Underestimation of High xG Values:** All three models exhibit a clear tendency to underestimate higher xG values compared to StatsBomb, as indicated by the points lying below the red dashed line (representing perfect agreement, $y = x$). This trend is particularly noticeable for xG values above 0.3.
- **Conservative Predictions:** Each model shows a high concentration of points at lower xG values (0 to 0.3), suggesting a conservative approach to predicting xG.

Concentration of Points:

- **Logistic Regression (Subplot A):**
 - The scatter plot shows that the logistic regression model tends to predict lower xG values, particularly for higher xG values.
 - There is a significant concentration of points where both the x-axis and y-axis values range between 0 and 0.2, indicating conservative predictions.
- **XGBoost (Subplot B):**
 - Similarly to logistic regression, XGBoost also underestimates higher xG values.
 - The points are more dispersed, indicating a wider range of predictions, but still with a majority concentrated between 0 and 0.3 on the x and y-axis.
- **Random Forest (Subplot C):**
 - The distribution of points is much more spread out than logistic regression and XGBoost.
 - There is an upward bias, random forest tends to predict higher xG values compared to the benchmark model. This is evident from the y-values being higher (0-0.8) when the benchmark x-values are low (0-0.2).
 - It overestimates consistently the xG values for scenarios where the benchmark model predicts low xG values

- The model underestimates and overestimates xG values across different ranges, indicating that while there is a tendency to underestimate higher xG values, it also overestimates in certain cases, particularly at lower xG values.

Underestimation of High xG Values: Consistently, all three models show an underestimation of high xG values compared to the statsbomb model. This is a critical area where model calibration could improve alignment with actual xG values provided by StatsBomb.

Conservative Predictions: The high concentration of points in the lower xG range (0 to 0.3) across all models suggests that these models are generally conservative in their predictions. This conservatism might be beneficial in reducing false positives but lead to underestimations for higher xG values.

Model Comparison:

- **Logistic Regression:** Shows the most conservative and tightly clustered predictions, indicating a higher degree of caution in predicting higher xG values.
- **XGBoost:** Displays a wider range of predictions with more dispersion than logistic regression, suggesting more variability and less conservative predictions.
- **Random Forest:** Offers the most spread predictions with a lot of shots underestimating and overestimating compared to the statsbomb model.

Conclusion of the scatter plot analysis

These observations highlight the differences in prediction tendencies between the models and suggest that while all models are generally conservative, Random Forest provides a more varied prediction range. The consistent underestimation of high xG values across all models points to a need for better calibration methods to align predictions more closely with actual outcomes. Additionally, the conclusions drawn from the distribution of predicted probabilities and the calibration plot are confirmed by these scatter plots. The lack of high probability predictions (>0.5) further underscores the necessity for improved model calibration. Improving the predictive accuracy and reliability of these models requires addressing these calibration issues, particularly for higher xG values.

8. Conclusion

This section introduces the key findings from the research and outlines potential future work. These sections provide an overview of the insights gained and recommendations for further improving xG models.

8.1 Summary of Findings

This research sought to address the evaluation challenges inherent in xG models within football analysis by comparing the performance of various machine learning models, including logistic regression, XGBoost, and random forest. The investigation concentrated on critical aspects such as data preprocessing, feature engineering, model training, and the evaluation of these models using a range of performance metrics. Through this comprehensive approach, the study aimed to uncover key insights into the effectiveness of different xG modeling techniques and their applicability in real-world football scenarios. The key findings of this study are summarized as follows:

The data exploration between male and female players revealed that both genders exhibit similar patterns in shooting positions and the xG values associated with those shots. Both male and female players demonstrated a preference for high-quality shots taken closer to the goal over lower-value xG shots. This trend aligns with the strategic emphasis on quality shooting opportunities, especially since the popularization of the xG metric.

The F1 score, which combines precision and recall, alongside the Brier score and ROC-AUC, was identified as much more appropriate than accuracy for evaluating model performance. This is due to the significant class imbalance present in the dataset, where the number of non-goal events far exceeds the number of goal events. As it has been shown accuracy can be misleading in such contexts as it may reflect the model's ability to predict the majority class rather than its overall predictive power and does not perform better than the majority class predictor. Therefore, precision, recall, and the F1 score provide a more nuanced and informative assessment of a model's performance, especially in scenarios with imbalanced data.

Handling class imbalance proved to be a considerable challenge in ensuring the models provided a realistic representation of reality. This imbalance affects the models' ability to accurately predict goals, as they tend to be biased towards the more frequent non-goal events. Addressing this issue through techniques such as resampling, class weight adjustments, and stratified sampling is recommended to enhance the models' effectiveness and ensure a more balanced evaluation of their performance.

The study emphasizes that having high-quality, contextual data is more valuable than developing more complex models. Quality data ensures that the models can learn and generalize better, leading to more accurate predictions. This insight highlights the importance of investing in data collection and preprocessing to enhance the overall performance of xG models.

Hyperparameter tuning using grid search proved to be quite inefficient, failing to yield significant improvements in various performance metrics. This finding underscores the necessity of prioritizing data quality and preprocessing over the complexity of the models themselves.

Calibration plots were found to be highly effective in evaluating how well the predicted probabilities matched actual outcomes. The analysis revealed that all models were well-calibrated up to a predicted probability of 0.5. Beyond this threshold, predictions fluctuated significantly, indicating less reliable probabilistic predictions at higher values. Calibration plots thus provide a valuable tool for assessing the reliability of probabilistic predictions and identifying areas where models may need improvement.

Scatter plots comparing model predictions against StatsBomb xG values were used to benchmark the models. These plots showed a consistent underestimation of high xG values across all models, particularly for xG values above 0.3. While scatter plots are useful for benchmarking, they are limited in their ability to fully evaluate model performance and should not be relied upon as the sole method of assessment. This benchmarking indicated that the existing model from StatsBomb might have limitations and should be interpreted with caution.

Establishing a baseline model was deemed crucial for determining the utility of the developed models. The baseline model served as a reference point, allowing for a comparative assessment of the more sophisticated models. This practice ensures an evaluation of whether the developed models provide significant improvements over simple or naive approaches. Moreover, the baseline model highlighted that accuracy, while often used as a primary metric, was not an adequate measure in this case. Since the baseline model already achieved approximately 90% accuracy, it became clear that accuracy alone failed to provide meaningful insights into model performance.

Finally, it was highlighted that models need to be evaluated in real conditions and kept up to date. Continuous validation against real-world data is essential to ensure that the models remain relevant and accurate. This ongoing evaluation helps the models adapt

to changes in playing styles and strategies, maintaining their applicability and effectiveness over time.

This study underscores the complexity and challenges of developing and evaluating xG models in football analytics. By focusing on robust performance metrics, handling class imbalance, and ensuring data quality, more accurate and reliable xG models can be developed. Continuous validation and real-world evaluation are essential for maintaining the relevance and applicability of these models in football analysis. Future work should explore advanced modeling techniques, enhance feature engineering, and refine calibration methods to improve the predictive accuracy and reliability of xG models.

8.2 Future Work

To enhance this study, several paths for future work should be considered. The first priority is to address class imbalance. Models should focus on improving recall, especially in predicting the minority class (goals). This can be achieved even with a limited number of variables, ensuring that the models are better at identifying successful outcomes.

Following the improvement of recall, the next step is data refinement. This involves incorporating more contextual information, which is crucial for enhancing model performance and reliability. Detailed positional data of defenders during the shot, the part of the body used for shooting, and other relevant contextual factors should be included. Additionally, environmental data such as weather conditions and pitch quality could be considered. Providing the models with more comprehensive information will enable them to more accurately represent reality and perform better. Capturing details about the preceding pass, including its origin and speed, can also offer valuable context.

Once a robust xG model is established as a foundation, building upon it with advanced metrics like Expected Assists (xA) and Expected Threats (xT) can provide a more comprehensive evaluation of offensive plays. These metrics complement xG by offering deeper insights into chance creation and progression, capturing actions that lead to high-quality shot opportunities. As more features are added to these models, it will be important to analyze which contribute the most to information gain and overall model improvement, ensuring a balanced and holistic understanding of team and player performance.

Gathering more data on female players is essential for gaining deeper insights into the nuances of xG between genders. Current datasets often lack sufficient female player data, which limits the ability to make comprehensive comparisons. Expanding the

dataset to include more female players will enable a more detailed analysis of gender-specific shooting patterns and tendencies, ultimately leading to more accurate and inclusive models.

Once the data quality is improved, model refinement can be pursued. This involves fine-tuning hyperparameters, exploring different algorithmic approaches, and leveraging advanced techniques such as ensemble learning. The primary focus should be on data enhancement before model optimization to ensure that any improvements in model performance are based on a solid foundation of high-quality data.

Developing a robust evaluation framework is crucial for iterative model improvement. This framework should utilize appropriate metrics such as the F1 score and ROC-AUC, which are better suited for imbalanced datasets than simple accuracy metrics. Additionally, calibration plots should be employed to assess the reliability of probabilistic predictions. A well-structured evaluation framework will facilitate frequent iteration and refinement of the models, ensuring they remain effective and accurate over time.

Evaluating models against real-world scenarios and maintaining their relevance through constant training is vital. Models should be updated regularly to reflect the latest strategies and tactics in football. One approach is to use cumulative xG, which aggregates xG values over multiple games, and compare this against actual game outcomes to gauge model accuracy. Regularly testing the models in new games will help ensure their adaptability and precision, allowing for timely adjustments based on the latest data.

Exploring models specific to different leagues could uncover varying shooting strategies and improve the model's applicability across different contexts. Each league may exhibit unique playing styles and tactical approaches, which can influence shooting behaviors and xG values. By tailoring models to specific leagues, we can better capture these differences, leading to more precise predictions and insights. This approach will enhance the overall robustness and utility of the xG models in diverse footballing environments.

In summary, future work should prioritize addressing class imbalance, followed by data refinement, and finally model refinement. This systematic approach will significantly enhance the accuracy, reliability, and applicability of xG models in football analytics, leading to deeper insights and more effective predictions.

Bibliography

- BATE, Adam, 2017. Expected goals explained: The analysis that is changing the game. *Sky Sports* [online]. June 9, 2017. Available at: <https://www.skysports.com/football/news/11661/10907419/expected-goals-explained-the-analysis-that-is-changing-the-game> [accessed July 21, 2024].
- BERTIN, Michael, 2015. The Third-to-Last Thing I'll Ever Write About Expected Goals. [online]. August 2015. Available at : <http://michaelbertin.com/2015/08/28/the-third-to-last-thing-ill-ever-write-about-expected-goals/> [accessed April 28, 2024].
- BREIMAN, L, 2001. Random Forests. *Machine Learning*. Vol. 45, pp. 5-32. DOI 10.1023/A:1010950718922.
- CACHO-ELIZONDO, Dr Silvia, 2020. Big Data in the Decision-Making Processes of Football Teams Integrating a Theoretical Framework, Applications and Reach. *Journal of Strategic Innovation and Sustainability*. . DOI <https://doi.org/10.33423/jsis.v15i2.2887>.
- CALEY, Michael, 2013a. Shot Matrix I: Shot Location and Expected Goals. *Cartilage Free Captain* [online]. November 13 2013. Available at : <https://cartilagefreecaptain.sbnation.com/2013/11/13/5098186/shot-matrix-i-shot-location-and-expected-goals> [accessed April 28 2024].
- CALEY, Michael, 2013b. Shot Matrix II: Headers and Crosses. *Cartilage Free Captain* [online]. November 15, 2013. Available at : <https://cartilagefreecaptain.sbnation.com/2013/11/15/5107438/shot-matrix-ii-pass-type-and-shot-type-or-heading-is-super-hard> [accessed April 28, 2024].
- CALEY, Michael, 2015. This is why I like expected goals. (In response to the Deadspin article.). *Cartilage Free Captain* [online]. April 10, 2015. Available at : <https://cartilagefreecaptain.sbnation.com/2015/4/10/8381071/football-statistics-expected-goals-michael-caley-deadspin> [accessed April 28, 2024].
- CAVUS, Mustafa et BIECEK, Przemysław, 2022. Explainable expected goal models for performance analysis in football analytics. In : *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1-9. October 13, 2022. DOI 10.1109/DSAA54385.2022.10032440. arXiv:2206.07212 [cs]
- CEFIS, Mattia et CARPITA, Maurizio, 2024. A new xG model for football analytics. *Journal of the Operational Research Society*. pp. 1-13. DOI 10.1080/01605682.2024.2323669.
- CHAWLA, N. V. et al., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. Vol. 16, pp. 321-357. DOI 10.1613/jair.953.
- CHEN, Tianqi et GUESTRIN, Carlos, 2016. XGBoost: A Scalable Tree Boosting System. In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. San Francisco California USA : ACM. August 13 2016. ISBN 978-1-4503-4232-2. DOI 10.1145/2939672.2939785.
- COX, D. R., 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*. Vol. 20, no 2, pp. 215-242.

DECROOS, Tom et al., 2020. VAEP: An Objective Approach to Valuing On-the-Ball Actions in Soccer (Extended Abstract). In : *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 4696-4700. Yokohama, Japan : International Joint Conferences on Artificial Intelligence Organization. July 2020. ISBN 978-0-9992411-6-5. DOI 10.24963/ijcai.2020/648.

ELHABR, Tony, 2023. Tony's Blog - Calibrating Binary Probabilities. [online]. September 12, 2023. Available at : <https://tonyelhabr.rbind.io/posts/probability-calibration/> [accessed August 4, 2024].

FORTUNEBUSINESSINSIGHTS, 2024. Sports Analytics Market Size, Share | Growth Analysis [2032]. [online]. July 15, 2024. Available at : <https://www.fortunebusinessinsights.com/sports-analytics-market-102217> [accessed August 4, 2024].

Gradient Boosting Classifier — scikit-learn 1.5.1 documentation, *scikit-learn* [online]. Available at : <https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html> [accessed July 21, 2024].

HAIDAIR, Altaie, 2019. The revolutionary new football metric: Expected goals. *Hidden Insights* [online]. April 3, 2019. Available at : <https://blogs.sas.com/content/hiddeninsights/2019/04/03/the-revolutionary-new-football-metric-expected-goals/> [accessed April 29, 2024].

JAN VAN HAAREN et al., 2019. Analyzing Performance and Playing Style Using Ball Event Data. *Proceedings of the 25th ACM SIGKDD Conference*.

Logistic Regression — scikit-learn 1.5.1 documentation, *scikit-learn* [online]. Available at : https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html [accessed July 21, 2024].

MACKAY, Nils, 2016. Mackay Analytics. [online]. March 14, 2016. Available at : <https://mackayanalytics.nl/2016/03/14/how-not-to-evaluate-your-xg-model/> [accessed June 24, 2024].

MEAD, James, O'HARE, Anthony et MCMENEMY, Paul, 2023. Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*. Vol. 18, no 4, p. e0282295. DOI 10.1371/journal.pone.0282295.

MORENO-TORRES, J. G., SAEZ, J. A. et HERRERA, F., 2012. Study on the Impact of Partition-Induced Dataset Shift on k -Fold Cross-Validation. *IEEE Transactions on Neural Networks and Learning Systems*. Vol. 23, no 8, pp. 1304-1312. DOI 10.1109/TNNLS.2012.2199516.

MULLER, John, 2020. Goals Added: Introducing A New Way To Measure Soccer. *American Soccer Analysis* [online]. May 4, 2020. Available at : <https://www.americansocceranalysis.com/home/2020/4/22/37ucr0d5urxxtryn2cfhzormdziphq> [accessed June 24, 2024].

PINNACLE, 2019. Understanding the limitations of expected goals. *Betting Resources* [online]. January 2, 2019. Available at : <https://www.pinnacle.com/betting-resources/en/soccer/understanding-the-limitations-of-expected-goals/k4gjc1wx3vs6mwvk> [accessed March 25, 2024].

Quick start — mplsoccer 1.4.0 documentation, [online]. Available at : <https://mplsoccer.readthedocs.io/en/latest/> [accessed September 22, 2024].

Random Forest Classifier — scikit-learn 1.5.1 documentation, *scikit-learn* [online]. Available at : <https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> [accessed July 21, 2024].

RATHKE, Alex, 2017. An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*. Vol. 12, no Proc2. DOI 10.14198/jhse.2017.12.Proc2.05.

REIN, Robert et MEMMERT, Daniel, 2016. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*. Vol. 5, no 1, p. 1410. DOI 10.1186/s40064-016-3108-2.

scikit-learn: machine learning in Python — scikit-learn 1.5.1 documentation, [online]. Available at : <https://scikit-learn.org/stable/> [accessed July 21, 2024].

STATSBOMB, 2018. StatsBomb Open Data. [online]. June 5, 2018. Available at : <https://github.com/statsbomb/open-data> [accessed August 2, 2024]. Platform: Github

STATSBOMB, 2021. Introducing On-Ball Value (OBV). *StatsBomb | Data Champions* [online]. September 16, 2021. Available at : <https://statsbomb.com/articles/soccer/introducing-on-ball-value-obv/> [accessed June 24, 2024].

statsbomb/open-data [logiciel] [online]. March 13, 2024. StatsBomb. [accessed March 15, 2024]. Available at : <https://github.com/statsbomb/open-data> [accessed March 15, 2024].

SUMPTER, David, 2021. Explaining Expected Threat. *Medium* [online]. August 20, 2021. Available at : <https://soccermatics.medium.com/explaining-expected-threat-cbc775d97935> [accessed June 24, 2024].

THAKUR, Ayush et ROY GOSTHIPATY, Aritra, 2020. Weights & Biases. *W&B* [online]. September 1, 2020. Available at : <https://wandb.ai/authors/class-imbalance/reports/Simple-Ways-to-Tackle-Class-Imbalance--VmlldzoxODA3NTk> [accessed August 4, 2024].

VIDAL-CODINA, Ferran et al., 2022. *Automatic event detection in football using tracking data* [online]. arXiv:2202.00804. arXiv. arXiv:2202.00804. Available at : <http://arxiv.org/abs/2202.00804> [accessed July 1, 2024]. arXiv:2202.00804 [cs]

VILORIA, Iñaki Rabanillo, 2024. Creating Better Data: How To Map Homography. *StatsBomb | Data Champions* [online]. March 26, 2024. Available at : <https://statsbomb.com/articles/football/creating-better-data-how-to-map-homography/> [accessed August 2, 2024].

WALID, Ahmed, 2023. Inspired by you: xG and beyond - visiting StatsBomb's data collection centre in Cairo. *The New York Times* [online]. March 24, 2023. Available at : <https://www.nytimes.com/athletic/4263797/2023/03/24/xg-statsbomb-inspired-by-you/> [accessed July 16, 2024].

WHITMORE, Jonny, 2020. Evolving Our Possession Value Framework. *Stats Perform* [online]. 2020. Available at : <https://www.statsperform.com/resource/evolving-our-possession-value-framework/> [accessed June 24, 2024].

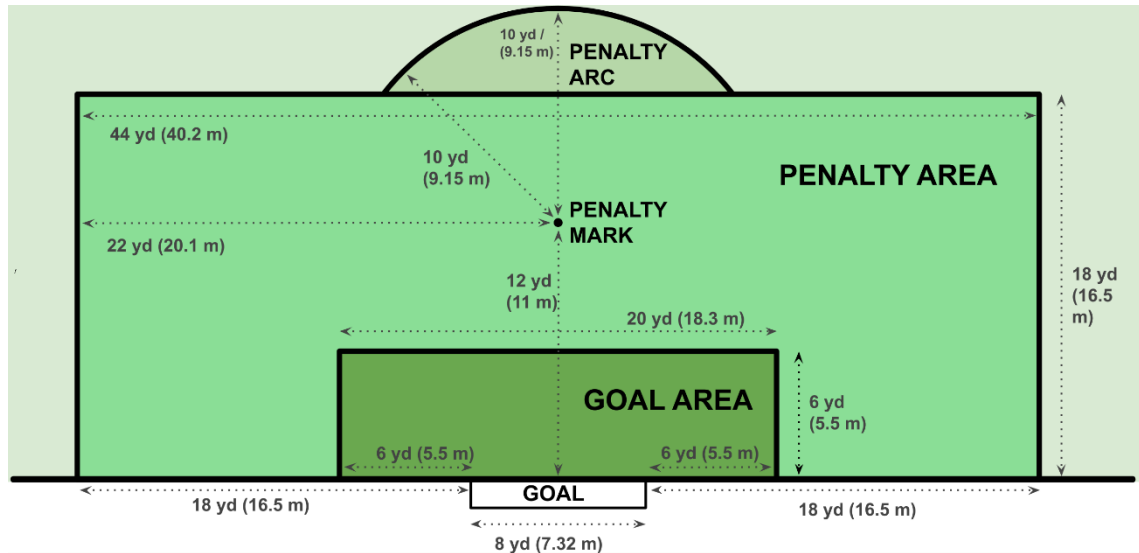
WHITMORE, Jonny, 2023. What Is Expected Goals (xG)? | Opta Analyst. *The Analyst* [online]. August 8, 2023. Available at : <https://theanalyst.com/2023/08/what-is-expected-goals-xg/> [accessed March 28, 2024].

CC BY 4.0 Serrano C, Felipe JL, Garcia-Unanue J, Ibañez E, Hernando E, Gallardo L, Sanchez-Sanchez J, 2020. *Local Positioning System Analysis of Physical Demands during Official Matches in the Spanish Futsal League*. <https://doi.org/10.3390/s20174860>

CC BY 4.0 Z. Akyildiz, H. Nobari, F. González-Fernández, et al., 2022. *Variations in the physical demands and technical performance of professional soccer teams over three consecutive seasons*. <https://www.nature.com/articles/s41598-022-06365-7>

CC BY 4.0 Hyunsung Kim, Jaehee Kim, Young-Seok Kim, Mijung Kim, Youngjoo Lee, 2020. *Energy-Efficient Wearable EPTS Device Using On-Device DCNN Processing for Football Activity Classification*. <https://doi.org/10.3390/s20216004>

Appendix 1: Definition of different zones of the penalty area



CC BY-SA 4.0 Grover Cleveland, 2019. Diagram of the goal area and penalty area in association football with detailed measurements.