Skeleton Document

## 1.Introduction

The sports industry has witnessed a significant transformation in recent years, driven by advancements in technology and data analytics. From player performance to fan engagement, data analytics has become an integral part of how sports are played, watched, and understood. This transformative growth is particularly evident in football, where data-driven decision-making has become a crucial aspect of the game.

The sports analytics industry is experiencing significant growth, particularly in women's football, which has seen increased visibility and professionalisation. Despite these advancements, there is a notable gap in the analytical resources available for women's football compared to men's. Most statistical models, including the influential expected goals (xG) metric, have been developed and refined primarily using men's football data. This metric assesses the likelihood of a shot resulting in a goal based on various factors like shot location and play context, revolutionising football analytics by providing precise measures of player and team performance (Rathke, 2017).

## 1.1 Project Background

Technological advancements in data collection, storage, and analysis have revolutionised football analytics (Cavus & Biecek, 2022). Early uses of sports analytics were basic and manual, focusing on simple statistics like possession and shots on target. However, advancements in machine learning and big data have enabled more complex analyses, such as pass evaluation, space quantification, and expected goals (xG) modelling (Cavus & Biecek, 2022; Rathke, 2017).

Initially, sports analytics focused predominantly on men's football due to the availability of extensive data and substantial financial stakes. Early xG models and other performance metrics were developed using data from men's games, which often led to highly effective insights and strategies in that context (Rathke, 2017). However, these models have been less effective when directly applied to women's football due to differences in play style, physicality, and other factors (Mackenzie & Cushion, 2013).

Expected Goals (xG) is a crucial metric in football analytics, providing a probabilistic measure of the quality of goal-scoring opportunities. It accounts for various factors such as shot distance, angle, and type, offering a nuanced understanding of a team's or player's performance beyond just goals scored (Rathke, 2017). xG models are vital for performance evaluation and strategic planning, allowing teams to assess their offensive and defensive efficiency more accurately.

Women's football has seen significant growth over the past decade, with increasing participation rates, higher levels of competition, and greater visibility on the global stage. Despite this progress, the analytical support for women's football has lagged behind that of men's football. Historically, the lack of dedicated data and resources meant that women's football was often analysed as an extension of the men's game rather than as an independent entity (Mackenzie & Cushion, 2013). This approach fails to capture the unique dynamics and characteristics of the women's game, highlighting the need for tailored analytical models (Sarmento et al., 2018).

The development of dedicated analytical frameworks for women's football is not just about equity in data availability but also about enhancing strategic understanding and decision-making in the sport. As more data becomes available and more tailored models are developed, teams and coaches can gain deeper, actionable insights specifically relevant to improving performance in women’s football (Cavus & Biecek, 2022; StatsBomb, 2023). This evolution in analytics is expected to drive further growth and professionalisation in women’s football, helping to close the gap with men’s football in terms of tactical sophistication and audience engagement.

## **1.2 Project Objectives**

The primary objective of developing a expected goals (xG) model for women's football is to refine how performance analytics are applied within the sport. This initiative enhances tactical analysis and strategic planning while supporting the broader goal of elevating the professional stature and understanding of women's football. Tailoring analytical tools to the women's game ensures that the insights generated are genuinely reflective of the on-field scenarios, leading to better decision-making by coaches and teams and potentially improving player development and scouting within the sport.

The current data landscape in women's football has been notably limited, often requiring reliance on models developed primarily with men's football data. Recognizing this gap, efforts have been made to actively expand datasets to incorporate detailed event data from various women's football leagues across multiple seasons. This expansion has significantly enhanced the volume and quality of data available, creating a robust foundation for model training. The aim of this project is to leverage this enriched dataset to effectively develop a women-specific expected goals (xG) model, ensuring that the analytics better reflect the unique characteristics of women's football.

# 

# **2. Literature review**

**2.1 Explainable Expected Goal Models for Performance Analysis in Football Analytics**

Cavus and Biecek (2022) present a study that focuses on the development of an explainable expected goals (xG) model using data from seven seasons of the top-five European football leagues. The model integrates explainable artificial intelligence (XAI) tools to enhance interpretability without sacrificing predictive accuracy. The aim is to provide a more representative measure of team and player performance, addressing the inherent randomness in match outcomes. Initially introduced by Green, the xG metric quantifies the probability of a shot resulting in a goal, effectively addressing the low-scoring and often unpredictable nature of football.

The study utilises a comprehensive dataset of 315,430 shots from seven seasons (2014-2021) of the top-five European leagues. Machine learning models such as Random Forest, XGBoost, LightGBM, and CatBoost are employed, with training facilitated by AutoML tools. Performance is validated using metrics like precision, recall, F1 score, accuracy, and AUC. To tackle the data imbalance issue, where only about 10.66% of shots result in goals, the random over-sampling method is used.

Furthermore, Cavus and Biecek introduce aggregated ceteris-paribus (CP) profiles to provide local-level explanations, making the model more interpretable. These profiles help in evaluating player and team performance, offering insights into potential performance improvements by adjusting shot strategies. The proposed model is claimed to be the most accurate xG model to date, effectively predicting both goals and non-goals. The use of XAI tools in football analytics is highlighted, offering practical applications for performance evaluation and tactical adjustments (Cavus and Biecek, 2022).

**2.2 An Examination of Expected Goals and Shot Efficiency in Soccer**

Rathke's study investigates goal scoring in European football leagues, focusing on factors associated with predicting Expected Goals (xG). The analysis uses data from the Premier League and Bundesliga games of the 2012-2013 season to determine the impact of shot distance and angle on xG. The xG metric helps evaluate player performance, particularly strikers, by predicting the likelihood of a goal based on shot characteristics. Rathke's model divides the football pitch into zones and calculates the xG based on the shot's distance and angle, emphasizing that these factors together provide a better prediction than each variable alone (Rathke, 2017).

The study analyzes 18,218 shots from both leagues, categorizing them into eight zones based on distance and angle. Logistic regression was initially attempted but not feasible due to data limitations. Instead, a simpler calculation method using league averages for goals per shot from each zone was employed. The results highlight that shots taken closer to the goal (zones 1 and 3) have a higher xG value. The study's findings can aid managers and coaches in identifying players who consistently score or outperform their xG, which is useful for tactical decisions and player evaluations (Rathke, 2017).

**2.3. Expected Goals in Football: Improving Model Performance and Demonstrating Value**

Mead, O’Hare, and McMenemy (2023) focus on enhancing xG models by incorporating previously untested features such as player/team ability and psychological effects. The study aims to demonstrate the value of xG models in predicting football outcomes more effectively than traditional metrics. Historically, the concept of expected goals originated from studies in ice hockey and football, aimed at providing a probability measure for shot success. xG models are essential for dealing with the randomness in football, offering a more stable performance metric than goals alone (Mead, O’Hare, and McMenemy, 2023).

The study explores integrating new features, including player/team ability and psychological factors, into xG models. Data from sources like Wyscout and Transfermarkt are used, incorporating event data (passes, shots, fouls) and positional data (player locations tracked at high frequency). The models are built using various machine learning techniques, including logistic regression, gradient boosting, neural networks, and support vector machines. Enhanced xG models are shown to be superior in predicting team success and match outcomes, outperforming traditional statistics and even some industry-leading models (Mead, O’Hare, and McMenemy, 2023).

**2.4 Evolution and Application in Women's Football**

While all three studies primarily focus on men's football, the methodologies and advancements can be applied to women's football. Similar data collection methods can be employed for women's football, focusing on shot characteristics, player positions, and game contexts. Machine learning techniques can be adapted to account for the specific attributes and playing styles prevalent in women's football. Features such as player ability, team strategies, and psychological factors should be included to enhance the accuracy of xG models in women's football. Positional adjustments can be particularly useful in evaluating the performance of players in different roles, considering the unique dynamics of women's football.

Differences in playing style and physicality between men's and women's football can influence xG values. Models should be adjusted to reflect these variations, and shot distance and angle may have different impacts, requiring tailored zone definitions and probability calculations. Variations in league quality, competition levels, and team strategies across different women's leagues should be considered, and models can be adjusted to account for these differences, providing more accurate and context-specific xG values (Bransen and Davis, 2021; StatsBomb, 2023).

2.5 **Advancements and Future Directions**

The development of xG models has significantly advanced football analytics by providing a more nuanced measure of performance. The integration of machine learning techniques and explainable AI tools, as demonstrated in the works of Cavus and Biecek (2022), Rathke (2017), and Mead, O’Hare, and McMenemy (2023), has enhanced both accuracy and interpretability. These methodologies offer a robust framework that can be adapted to women's football, thereby supporting the growing interest and investment in the women's game. By incorporating specific attributes relevant to women's football, such as playing style, physicality, and league characteristics, xG models can be tailored to provide accurate and context-specific evaluations. Future research should focus on expanding data collection, exploring new features, and refining these models to cater to the unique dynamics of women's football. This approach will ensure that the advancements in xG models continue to contribute meaningfully to performance analysis and tactical adjustments in both men's and women's football (Bransen and Davis, 2021; StatsBomb, 2023; Mackenzie and Cushion, 2013; Memmert and Rein, 2018; Sarmento et al., 2014).

# **3. The Approach**

## 3.1 Understanding the Concept of xG

Expected goals (xG) is a statistical metric used in football to measure the quality of goal-scoring opportunities. The underlying principle of xG is to assign a probability to each shot based on factors that influence the likelihood of scoring, such as shot location, type of shot, and presence of defenders (Rathke, 2017). Shots from closer to the goal and in more central positions are generally assigned higher xG values, as they are more likely to result in a goal on average. For example, a shot from the penalty spot may have an xG value of 0.75, meaning it has a 75% chance of being scored on average.

By comparing a team's actual goals scored to their xG, analysts can determine if the team is over or underperforming in terms of converting their chances (Rathke, 2017). A team with a high xG but low actual goals may be unlucky or the opponent's defence is overperforming, while a team with a low xG but high actual goals may be overperforming. This information can provide valuable insights into a team's attacking efficiency and finishing ability.

## 3.2 Data Collection for xG Modeling

To develop an xG model for the 2023 Women's World Cup, the first step is to collect high-quality event data on shots taken during the tournament. The most important data points will be detailed information on each shot, including factors like shot location, shot type, and the presence of defenders (Anzer & Bauer, 2021).

Fortunately, StatsBomb has released a free dataset containing comprehensive event data for the 2023 Women's World Cup. This dataset provides the necessary granular information on shots, including the x/y coordinates, body part used, and whether the shot was assisted. By leveraging this publicly available data, we avoid the challenges of obtaining proprietary data sources and focus on the modelling and analysis (Pappalardo et al., 2021).

In addition to the shot data, we believe it will be beneficial to collect contextual information that could influence shot probabilities, such as game state (score difference, time remaining) and player characteristics (Bransen & Davis, 2021).

## 3.3 Preparing the Data for xG Modeling

Once the raw data has been collected, the next step is to prepare the dataset for modelling. This involves cleaning the data, handling missing values, and engineering relevant features that could impact shot outcomes.

For the Women's World Cup data, studies have found that female footballers tend to shoot from closer to the goal with a smaller angle compared to their male counterparts (Pappalardo et al., 2021). Accounting for these differences in the feature engineering process will be crucial for building an accurate xG model tailored to the women's game.

## 3.4 Feature Selection

After collecting and preparing the Women's World Cup data, the next crucial step is to select the most relevant features for predicting shot outcomes. Feature selection is essential for building an accurate and interpretable xG model, as it helps identify the factors that have the greatest impact on goal-scoring probabilities (Mead et al., 2021).

Common features used in xG models include shot location (distance and angle to goal), shot type (e.g., left foot, right foot, head), assist type, and game state (e.g., score difference, time remaining) (Bransen & Davis, 2021).

One approach to feature selection is to use statistical techniques such as principal component analysis, correlation analysis, mutual information, or recursive feature elimination (Mead et al., 2021). These methods can help identify the most informative features and remove redundant or irrelevant ones, improving the model's performance and interpretability.

Additionally, we hope to leverage improving our domain knowledge and consult with experts in women's football to inform the feature selection process. Incorporating insights from coaches, analysts, and players can help identify contextual factors that may influence shot outcomes but are not readily apparent in the data (Bransen & Davis, 2021).

## 3.5 Model Selection

In the realm of xG modelling, various approaches and models have been explored to predict goal-scoring probabilities accurately. Let's delve into the different models mentioned in research papers and discuss their advantages and limitations:

1. **Logistic Regression**

Logistic regression is a statistical model used to predict the probability of a binary outcome (0 - not goal: 1 - goal), making it suitable for estimating the likelihood of a shot resulting in a goal.

Advantages: Simple to implement, interpretable results, and well-suited for binary classification tasks like xG modeling.

Limitations: Assumes a linear relationship between features and the log-odds of the outcome, which may not capture complex interactions in the data.

1. **Gradient Boosting**

Gradient boosting is an ensemble learning technique that builds a strong predictive model by combining multiple weak models. Gradient boosting keeps refining the prediction by learning from errors, step by step, until it can't improve any further. It’s like gradually polishing a rough sketch into a detailed and accurate picture.

Advantages: High predictive accuracy, handles complex relationships in the data, and robust to overfitting.

Limitations: Computationally intensive, prone to overfitting if hyperparameters are not tuned properly.

1. **Neural Networks**

Neural networks are deep learning models inspired by the human brain, capable of learning complex patterns in data. Neural networks become better at their tasks by continuously learning from the data they process, improving their accuracy over time without being explicitly programmed to perform those specific tasks.

Advantages: Can capture intricate patterns in the data, suitable for nonlinear relationships, and can handle large datasets.

Limitations: Require a large amount of data for training, prone to overfitting, and complex to interpret.

1. **Support Vector Machines (SVM)**

Support Vector Machines are supervised learning models used for classification tasks, including predicting goal probabilities in xG modeling. SVMs examine the characteristics of each shot and decide how likely it is to be a goal, helping us understand and predict football match outcomes better.

Advantages: Effective in high-dimensional spaces, memory-efficient, and versatile with different kernel functions.

Limitations: Not suitable for large datasets, sensitive to the choice of kernel parameters, and may not perform well with noisy data.

1. **Tree-Based Models (Random Forest, AdaBoost, XGBoost)**

Tree-based models are ensemble learning methods that build a strong predictive model by combining multiple decision trees. Each decision tree is a simple model that makes decisions by splitting data into branches based on certain criteria. When you combine many of these trees together, they form a more powerful model that can predict more accurately than any single tree could. This approach is especially good at handling complex data by looking at it from many different angles, which helps improve the overall prediction accuracy.

Advantages: Handle nonlinear relationships, feature interactions, and are robust to outliers and missing data.

Limitations: Prone to overfitting, complex models may be hard to interpret, and sensitive to hyperparameter tuning.

Each of these models offers unique strengths and weaknesses in predicting expected goals in women's football. The choice of model should consider the specific characteristics of the data, the complexity of relationships in the game, and the interpretability required for practical applications in the football analytics domain (Mead et al., 2021).

## 3.6 Model Development

After evaluating various models and selecting the most suitable ones for predicting expected goals in women's football, the next step is to develop the chosen models. This process involves several stages, as outlined in the research papers:

1. **Model Training**

The selected models are trained on the prepared Women's World Cup dataset, using techniques such as cross-validation or 5 - fold validation using 80%-20% training and testing dataset split to ensure robust model performance (Bransen & Davis, 2021). We will ensure that the training data is representative of the women's game, accounting for any gender-specific patterns in shot characteristics and conversion rates (Pappalardo et al., 2021).

1. **Hyperparameter Tuning**

The models' hyperparameters are optimised to improve their predictive performance on the validation set (Mead et al., 2021). We will use techniques like grid search or random search to explore the hyperparameter space efficiently, and monitor for signs of overfitting during the tuning process.

1. **Feature Engineering**

Additional features may be engineered from the existing data to enhance the models' predictive power. We will explore interactions between features and incorporate domain knowledge from women's football experts to identify relevant contextual factors that could influence shot outcomes.

By following these stages and incorporating best practices from the research papers, you can develop robust xG models tailored to the women's game. These models can provide valuable insights into goal-scoring probabilities during the 2023 Women's World Cup and beyond.

## 3.7 Model Validation and Testing

Data Splitting: The Women's World Cup dataset is split into training and testing sets, typically using a 80:20 ratio (Mead et al., 2021). The split is stratified, meaning the training and testing sets have a similar distribution of positive (goals) and negative (non-goals) outcomes. This helps prevent biassed estimates of model performance.

Model Training: The xG models are trained on the training set using techniques like cross-validation to optimise their performance (Bransen & Davis, 2021). We will monitor for signs of overfitting during training, where the model performs well on the training data but fails to generalise to new samples. Techniques like regularisation and early stopping can help mitigate overfitting.

Model Evaluation: The trained models are evaluated on the held-out testing set using appropriate metrics such as log loss, area under the ROC curve (AUC), and accuracy (Mead et al., 2021). Log loss is a commonly used metric for xG models, as it measures the quality of probability estimates. Lower log loss indicates better model performance.

Calibration Checks: The calibration of the xG models is assessed by comparing the predicted probabilities to the actual outcomes (Mead et al., 2021). Our model will have predicted probabilities that closely match the observed frequencies of goals. Calibration plots and reliability diagrams can help identify any systematic biases in the models.

By following these validation and testing procedures, we will ensure that the developed xG model not only fits the Women's World Cup data well but also generalises effectively to new, unseen data. This step is crucial to provide accurate insights into goal-scoring probabilities in women's football.

## 3.8 Analysis of xG Model

After validating the performance of the developed xG models, the next step is to analyse their outputs and interpret the results. This involves evaluating the models' performance using appropriate metrics and employing techniques to enhance interpretability, which can provide valuable insights into the factors influencing goal-scoring probabilities in women's football.

**Performance Evaluation Metrics**

The xG models can be evaluated using various performance metrics, such as the Area Under the Curve (AUC), Brier score, and Log-Loss (Mead et al., 2021).  
AUC measures the model's ability to distinguish between positive (goals) and negative (non-goals) outcomes, with values closer to 1 indicating better performance. The Brier score measures the mean squared difference between predicted probabilities and actual outcomes, with lower values indicating better calibration. Log-Loss, as mentioned earlier, measures the quality of probability estimates, with lower values indicating better performance.

**Feature Importance Analysis**

To understand the contribution of each feature towards the predicted xG, techniques like SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) will be employed (Lundberg & Lee, 2017; Ribeiro et al., 2016). SHAP values provide a measure of each feature's importance in the model's predictions, considering both the feature's individual impact and its interactions with other features. LIME generates local explanations for individual predictions, identifying the most influential features for a specific shot.

By thoroughly analysing the xG model outputs and leveraging interpretability techniques, we hope that these insights can inform decision-making, enhance player and team performance, and ultimately contribute to the growth and development of the women's game.

## 3.9 Iterative Improvement

After analysing the initial outputs and performance of the xG model developed for women's football, the next step is to refine and improve the model through an iterative process. This involves carefully examining the model's strengths, weaknesses, and areas for potential enhancement, and then making targeted adjustments to enhance its predictive accuracy and practical utility.

One key aspect of the iterative improvement process is to revisit the feature engineering stage. Based on the insights gained from the feature importance analysis, we will explore incorporating additional relevant features or modifying the existing ones to better capture the nuances of goal-scoring in the women's game (Pappalardo et al., 2021). For example, if the model is underperforming in certain game situations, such as when teams are trailing or leading by a large margin, additional features related to game state could be engineered to improve the model's ability to account for these contextual factors.

Additionally, the model's hyperparameters can be further tuned to optimise its performance on the validation and testing datasets (Mead et al., 2021). This may involve experimenting with different regularisation techniques, adjusting the learning rate, or exploring alternative model architectures to find the best-performing configuration for the women's football xG task.

By embracing an iterative improvement mindset, we aspire to continuously refine and enhance the women's football xG model, ensuring that it provides accurate and actionable insights that can be effectively leveraged by coaches, analysts, and decision-makers in the women's game.

### **4. Measures of Success**

The success of developing a women-specific expected goals (xG) model for women's football will be evaluated based on its ability to provide accurate, actionable, and relevant insights tailored to the unique dynamics of the women's game. The model aims to enhance decision-making for coaches, management staff, and recruitment teams by predicting match outcomes, identifying top-performing players, guiding team selection, and estimating player value.

The following criteria will outline the specific measures of success for the xG model:

1. **Reliability in Predicting Match Outcomes**:The xG model should consistently and accurately predict the outcomes of matches. Success will be measured by the model's predictive accuracy, which can be quantified through metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and comparison with actual match results.
2. **Identification of Top Performing Players**:The model should effectively identify the top-performing players by generating comprehensive performance reports that highlight key statistics such as goals, assists, shots on target, and xG values. Success will be measured by the model's ability to provide accurate and actionable insights into player performance.
3. **Guidance for Coaches and Management**:The xG model should serve as a valuable tool for coaches and management staff, helping them to choose the best squad for a given match setting. Success will be determined by the model's usability and the positive feedback from coaches and management staff who use it to make informed decisions about team selection and tactics.
4. **Tool for Player Recruitment and Value Estimation**:The model should be an effective tool for player recruitment and player value estimation during the football transfer season. Success will be measured by the model's accuracy in assessing player potential and value, and its acceptance and usage by recruitment teams for making transfer decisions.

**Deliverables**

The successful development of the women's xG model will result in the following deliverables:

## Capstone Document and Presentation: A comprehensive report detailing the research, methodology, development process, and findings of the project, along with a presentation to summarise the key points and results.

## Codebase: The complete code used to develop, test, and validate the xG model, documented and stored in a version-controlled repository.

## Datasets: The datasets used for training and testing the model, including any pre-processing steps applied.

## Technical Report: A detailed technical report explaining the model's architecture, statistical methods used, validation results, and potential limitations.

## User Manual: A user-friendly manual to help end-users understand and implement the model in their analyses.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **References**

Bransen, L. and Davis, J., 2021. Women’s football analysed: interpretable expected goals models for women. In AI for Sports Analytics (AISA) Workshop at IJCAI 2021.

Cavus, M. and Biecek, P., 2022. Explainable expected goal models for performance analysis in football analytics.

Hewitt, J.H. and Karakuş, O., 2023. A machine learning approach for player and position adjusted expected goals in football. Franklin Open, 4, p.100034.

Mackenzie, R. and Cushion, C., 2013. Performance analysis in football: A critical review and implications for future research. Journal of Sports Sciences, 31(6), pp.639-676. doi: 10.1080/02640414.2012.746720.

Mead, J., O’Hare, A. and McMenemy, P., 2023. Expected goals in football: Improving model performance and demonstrating value. PLOS ONE, 18(4), p.e0282295.

Memmert, D. and Rein, R., 2018. Match analysis, big data and tactics: current trends in elite soccer. German Journal of Exercise and Sport Research, 48(2), pp.213-217. doi: 10.1007/s12662-018-0507-7.

Rathke, A., 2017. An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise, 12(2proc), pp.S514-S529. doi: 10.14198/jhse.2017.12.Proc2.05.

Sarmento, H. et al., 2014. Match analysis in football: a systematic review. Journal of Sports Sciences, 32(20), pp.1831-1843. doi: 10.1080/02640414.2014.898852.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2021). PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Transactions on Intelligent Systems and Technology, 10(5), 1-27. https://doi.org/10.1145/3343172

StatsBomb, 2023. Women's Football Analytics. Available at:h[ttps://statsbomb.com/womens-football-analytics](https://statsbomb.com/womens-football-analytics).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). doi: <https://doi.org/10.1145/2939672.2939778>

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4768-4777).