Event-Based Expected Goals Modelling with Interpretability and Calibration: Evidence from the 2018 FIFA World Cup

Zexuan Xu & Mohd Rahimi Che Jusoh

Abstract **–We develop an interpretable and well-calibrated expected goals (xG) model using only event data from all 64 matches of the 2018 FIFA World Cup. Pitch coordinates are first normalized so that all teams attack from left to right by period, and timestamps are mapped to absolute match time. Shots are labeled by matching to goals from the same team within a 0–10 second window, with conservative fallbacks using event tags to capture immediate rebounds and deflections. On this basis, we construct a compact but rich feature set including (i) shot geometry (distance, longitudinal and lateral position, goal-mouth angle and selected nonlinear or interaction terms), (ii) temporal and game-state context (match minute, score difference, time since previous event, displacement and implied speed), (iii) chance-creation and execution indicators (through ball, cross, set piece, rebound, header, weak foot), and (iv) out-of-fold hierarchical encodings for shooter, attacking team and opponent, capturing latent finishing and defensive tendencies without leakage. We compare ridge-regularized logistic regression with quadratic terms to random forest and gradient boosting under match-grouped cross-validation, and apply isotonic regression for probability calibration. The selected logistic model attains competitive discrimination on the held-out test set (ROC-AUC ≈ 0.78, PR-AUC ≈ 0.32), while calibration substantially improves probability accuracy (Brier score 0.176→0.077). Spatial maps, distance and angle response curves, and team-level aggregates confirm that predicted xG is consistent with known football regularities and remains reliable across pitch zones and teams. SHAP analyses on a comparable tree model highlight the dominant role of geometric variables and the incremental value of contextual and hierarchical features. The framework provides a practical xG solution for settings where only event data are available, balancing predictive performance, calibration quality and tactical interpretability.**

Keywords **– Expected goals (xG); Event data; Probability calibration; Model interpretability; SHAP**

I. INTRODUCTION

Football analytics has increasingly embraced data-driven metrics to quantify performance, with expected goals (xG) emerging as a cornerstone in recent years. The xG metric represents the probability that a given shot will result in a goal, providing a more nuanced measure of chance quality than raw shot counts or on-target statistics. It has become widely adopted by clubs and analysts as a tool for evaluating players and team attacking performance (Mead et al., 2023).

By assigning a goal probability to every shot, xG offers insight into "what could have happened" in a match rather than just what did happen, helping to identify whether teams are over- or under-performing relative to the chances they created (Brechot & Flepp, 2020; Smith, 2022). In practice, xG-based evaluations inform decision-making ranging from in-game tactics to player transfers, solidifying its status as a fundamental metric in modern football analysis.

However, the reliability and acceptance of xG models depend on more than just predictive accuracy. Many early xG models were simplistic, relying only on basic geometric features like shot distance and angle. Such models often ignored important contextual factors. For example, whether the shot was preceded by a through-ball or a rebound, the match phase and scoreline pressure, or the intrinsic finishing skill of the player. This omission of key features has been noted as a shortcoming of traditional xG approaches and partly why some practitioners remain skeptical of xG's outputs. Moreover, the probabilistic calibration of xG estimates is critical: if a model assigns 0.8 xG to a chance, that shot should actually result in a goal about 80% of the time in reality. Poorly calibrated models can mislead, even if a model discriminates well between goals and misses, it might systematically over- or under-estimate the true scoring probabilities, eroding trust when xG is used as an interpretable probability. Yet, many studies and commercial implementations focus on maximizing discrimination rather than ensuring the outputs align with observed frequencies (Fairchild et al., 2018; Hewitt & Karakuş, 2024). In addition, recent xG modeling efforts have trended toward complex machine learning techniques trained on large datasets to squeeze out performance gains (Cavus & Biecek, 2022). While effective, these black-box models often sacrifice interpretability, making it difficult for coaches and analysts to understand why a given shot's xG is high or low (Iapteff et al., 2025). The lack of transparency can further hinder adoption, as football practitioners value clear insights into the factors behind a metric's predictions.

In this paper, we address these challenges by developing an event-based xG model that prioritizes interpretability and calibration without requiring high-dimensional tracking data. Using the event data from all 64 matches of the 2018 FIFA World Cup, we construct a modeling pipeline that integrates rich spatial and contextual features while enforcing that the predicted probabilities closely reflect actual scoring rates. We deliberately limit our inputs to event data (e.g. shots, passes, fouls recorded with locations and timestamps) since such data is widely available and standardized, unlike player tracking data. Within this scope, our approach incorporates: (a) fine-grained spatial features describing shot location (distance, angle, and relative positioning on the pitch); (b) situational context such as the assist type (through-ball, set-piece, etc.), phase of play,

Zexuan Xu, City University Malaysia, Malaysia (Email address: 779509853@qq.com).
Mohd Rahimi Bin Che Jusoh, City University Malaysia, Malaysia (Email address: rahimi.jusoh@city.edu.my).

and game state which means time and score line when the shot occurred; and (c) latent player and team effects to account for finishing skill and team playing style differences. We train a logistic regression-based model with carefully engineered features and apply isotonic regression post-processing to ensure that the output xG probabilities are well-calibrated. Throughout, we use cross-validation grouped by match to evaluate performance, guarding against overfitting to the small dataset. The result is an xG model that achieves competitive predictive accuracy comparable to more complex models in the literature, while its probability estimates can be interpreted with confidence and its key drivers can be explained in football terms.

To demonstrate the usefulness of our approach, we present analyses at multiple levels of granularity. At the shot level, we examine how factors like distance, angle, assist type, and pressure situations influence the chance of scoring, drawing comparisons between the model's learned probabilities and the empirical success rates. At the spatial level, we visualize xG across different field locations and check that regions of the pitch with higher or lower actual conversion rates are accurately reflected by the model. At the team level, we aggregate xG over the tournament for each team to see if the model's estimates align with actual goals scored, which speaks to the model's validity in evaluating team performance. Finally, we employ interpretability techniques (such as SHAP values) to break down the model's predictions, illustrating how the model can be used not just as a black box predictor but as a tool for tactical insight. Our contributions thus lie in marrying data-driven modeling with practical interpretability and reliability, an approach we believe is valuable for sports analytics practitioners and extends the current state of event-based xG modeling.

II. PROBLEM STATEMENT

This study aims to answer how an expected goals model can be built and interpreted using only event data from a major tournament, addressing both predictive performance and practical usefulness. In particular, we focus on three research questions:

RQ1: How can we construct a fine-grained expected goals model using event data from the 2018 World Cup that produces well-calibrated goal probability estimates? Here we seek to develop an xG model that leverages the available information to predict the scoring probability of each shot as accurately as possible, while ensuring that the predicted probabilities align with observed frequencies. This involves selecting appropriate features, modelling techniques, and calibration methods suitable for the World Cup data, which poses challenges such as limited sample size (only 64 matches) and varied team styles.

RQ2: What is the influence of different spatial and contextual factors on the probability of a shot resulting in a goal? By analysing the model and the data, we aim to quantify how various features affect scoring likelihood. For example, how strongly do shot distance and angle determine the chance of scoring? Beyond basic geometry, we examine the impact of the preceding play, the game state means does a team's lead or deficit influence shooting quality or success rate? and the time since the last action. Answering this question provides football insights into which factors most significantly elevate or diminish scoring probability, in the specific context of World Cup matches.

RQ3: Does the proposed xG model maintain good calibration and explain ability across different conditions, such as various field locations and team aggregates? We investigate whether the model's predicted probabilities are consistent with actual conversion rates when broken down by location on the pitch and by team. Good calibration in these dimensions is crucial for the model's credibility as a true probability model. Additionally, we address whether the model's decisions can be interpreted: for instance, can we explain why the model gives a particular shot a high or low xG value? We use techniques like SHAP and visualizations to interpret the model's behaviour. This interpretability check ensures the model is not only statistically sound but also offers transparency, which is vital for adoption in coaching and analysis settings.

By answering these questions, our work intends to develop an xG modeling framework that is accurate, reliable, and insightful. The focus on the World Cup data provides a rigorous test case for this framework, as it involves a diversity of teams and play styles under high-pressure conditions, all within a relatively small dataset that demands careful model design to avoid overfitting. Ultimately, addressing the above questions will result in an xG model that not only predicts goal outcomes but also deepens our understanding of how and why goals are scored in elite football. Such a model can be a powerful tool for analysts seeking to evaluate performance and inform strategy when only event data are available.

III. LITERATURE REVIEW

Evolution of Event-Based xG Models

The expected goals concept was introduced to football analytics over a decade ago and has since seen rapid development, especially with the increasing availability of detailed event data (Brechot & Flepp, 2020,). Early xG models typically used logistic regression with a few core features, most notably the distance of the shot from goal and the angle to the goalmouth, as these geometric factors have an obvious influence on scoring probability. For instance, a simple model using only distance and angle (plus an indicator for header vs foot) can already provide a baseline level of predictive performance (Iapteff et al., 2025). These basic event-based models demonstrated that shots taken closer to the goal and from more central angles are far more likely to score, which matches intuition. However, such parsimonious models leave out many nuances of real-game situations.

In recent years, research has expanded xG models by adding a variety of spatial, temporal, and contextual features made available in event datasets. One important extension has been to include the type of preceding play or assist. Studies have shown that how a chance is created – for example, whether via a through-ball, a cross, a set-

piece, or a rebound – affects the probability of conversion (Bandara et al., 2024). Bandara et al. (2024) in particular proposed a framework that leverages event sequences before the shot: by incorporating information about one or two actions leading up to the shot, their random forest model significantly improved xG prediction accuracy compared to single-event model. This finding underline that a shot is not an isolated event; the context of play (such as a fast break vs. a slow build-up) and the quality of the final pass are important determinants of scoring success.

Another line of enhancement has focused on game context and pressure. The state of the match can psychologically and tactically influence shooting outcomes – for instance, a team that is trailing might take more desperate shots from poor positions, or conversely, a player in a tied knockout game might shoot more cautiously. While these effects are harder to quantify, some xG models have started to include features like the current score difference and time of the match at the shot (Mead et al., 2023). Mead et al. (2023) noted that such "psychological" factors and match context variables had been largely neglected in earlier models, and their work demonstrated that including them can yield measurable improvements in model performance. Similarly, features capturing whether the shooting team was on a strong momentum or coming straight from a turnover can also influence xG, though these are complex to derive from event data alone and remain an area of ongoing exploration.

Player and Team Effects

A significant assumption of classic xG models is that any two identical shots have the same probability of scoring, regardless of who takes them. This ignores potential differences in player finishing skill or tactical nuances between teams. Recently, researchers have questioned this assumption and attempted to adjust xG for player- or team-specific effects. For example, Madrero Pardo (2020) incorporated player skill ratings (extracted from a popular video game) into an xG model, finding that accounting for a player's finishing ability improved the prediction of season-long goal totals. A more principled approach is to use statistical hierarchical models or mixed-effects models: Hewitt and Karakuş (2024) constructed a Bayesian hierarchical logistic regression to examine position and player-level adjustments on xG. They found that, in a basic model with only distance and angle predictors, certain positions (notably strikers and attacking midfielders) showed higher baseline scoring probabilities than others, reflecting their typically better shooting skills (Hewitt & Karakuş, 2024). When additional features were included, positional differences diminished, suggesting that much of the positional advantage is explained by factors like shot selection and situation; however, individual player effects persisted (Hewitt & Karakuş, 2023). In particular, elite finishers, the oft-cited example being Lionel Messi were found to consistently exceed the expected conversion rates of average players for equivalent opportunities (Hewitt & Karakuş, 2023). These findings support the intuitive notion that xG models can benefit from modest adjustments for player skill, especially when evaluating performance over many shots: a goal by a highly skilled forward might be less "surprising" than the same goal scored by a less prolific player. Methodologically, such adjustments can be implemented via out-of-fold encoding or mixed-effects regression, which allow the model to learn latent talent factors from the data without leaking information (Mead et al., 2023). Our study follows this approach by incorporating player and team identifiers in a manner that infers their scoring effectiveness while avoiding overfitting in the small World Cup sample.

Advanced Modelling Techniques vs. Interpretability

Alongside feature enrichment, there has been exploration of various modeling approaches for xG. On large datasets complex machine learning models have been applied to capture nonlinear interactions among features. Cavus and Biecek (2022) built an xG model on 315,000+ shots from seven seasons using ensemble methods, and they reported that a Random Forest achieved better predictive performance than logistic regression or even gradient boosting in their comparisons (Cavus & Biecek, 2022). Some studies have also experimented with neural networks for xG, though gains over tree-based methods are not consistently large and come at the cost of a loss in interpretability (Herold et al., 2019; Anzer & Bauer, 2021). Indeed, model interpretability has become an important consideration in the literature. While black-box models might marginally improve accuracy, they make it difficult to communicate insights to coaches or players. In response, researchers have applied eXplainable AI tools to xG models. Cavus and Biecek (2022) exemplified this by using SHAP values to explain their model's predictions, confirming that traditional factors like distance and angle are dominant determinants of xG, while also uncovering how secondary features contribute to each prediction. Similarly, Iapteff et al. (2025) argue for a more interpretable approach by using a Bayesian generalized linear mixed model for xG. Their model, constrained to only seven key variables related to shot location and situation, achieved an Area Under Curve (AUC) nearly on par with a proprietary industry model (StatsBomb's xG) despite its simplicity. This demonstrates that with careful feature selection, even simpler models can remain competitive, all while providing clear coefficients and credible intervals that stakeholders can understand. Furthermore, Iapteff et al. (2025) highlight that by leveraging transfer learning, pre-training on a larger dataset and fine-tuning on a smaller one , one can handle small sample scenarios (like national team tournaments) without resorting to pure black-box solutions, maintaining interpretability.

Calibration and Evaluation

A theme receiving growing attention is the calibration of xG models. An xG model is not just a classifier but a probabilistic estimator, so its output should ideally reflect true scoring probabilities. Poor calibration

means the xG values might systematically overestimate or underestimate goal likelihoods, which is problematic when xG is used for crucial analyses such as evaluating team finishing (are they clinical or wasteful?) or assessing whether a result was due to bad luck. Some researchers have explicitly discussed calibration in the context of xG. Fairchild et al. (2018) recommended using the Brier score , a proper score measure of probability accuracy, to evaluate xG models, and they employed reliability curves to ensure the predicted probabilities matched observed goal frequencies in each probability band. Despite such recommendations, many subsequent works still primarily report discrimination metrics like AUC or logarithmic loss. There is a recognition, however, that calibration is particularly important when xG is communicated to a broad audience. To bridge this gap, recent xG modeling efforts including ours have adopted calibration techniques. For instance, some studies apply isotonic regression or Platt scaling as a post-processing step on top of the raw model outputs to fine-tune the probability estimates (Bandara et al., 2024; Kumar, 2019). Ensuring calibration does not usually improve ranking performance, but it makes the xG values more trustworthy as "expected goals" in the literal sense. Our work contributes to this aspect by emphasizing and evaluating calibration at different aggregation levels, which is less commonly reported in existing literature.

In summary, the literature on event-based xG models has progressed from simple logistic regressions on basic shot characteristics to sophisticated frameworks incorporating multiple facets of the game. There is an evident trade-off between complexity and interpretability. The most recent studies strive to find a balance, enriching models with critical context while either using inherently interpretable methods or applying post-hoc explanation techniques. However, gaps remain in fully calibrating these models and validating them in small but high-impact datasets like international tournaments. This paper builds on the state of the art by developing an xG model that integrates many of the advancements noted above and testing it on the 2018 World Cup data. In doing so, we aim to demonstrate that even with limited event data, a carefully crafted model can produce well-calibrated, interpretable, and insightful expected goals estimates, complementing the trends observed in broader league data and contributing a case study to the sports analytics literature on applying xG in practice.

IV. METHOD

Data and Event Representation

We use event data from all 64 matches of the 2018 FIFA World Cup. Each event record contains a match identifier mmm, a team identifier jjj, a player identifier iii, an event type label, a time stamp within the period in seconds, and raw pitch coordinates (x^{raw}, y^{raw}) measured in units that can be linearly mapped to metres. For each match we also have structured information on home and away teams and on the period of play, including first half, second half, and extra time.

Let L denote the length of the pitch in metres and W denote the width. We adopt the convention that the attacking team always attacks from left to right in the transformed coordinate system. This requires two steps: estimation of team attacking direction by period and spatial normalization of all events.

Pitch Normalization and Temporal Alignment

Attacking direction by team and period

For each match m, team j , and period r, we estimate an attacking direction variable $s_{mjr} \in \{-1,+1\}$. The value $s_{mjr} = +1$ corresponds to a team that attacks towards increasing x, and $s_{mjr} = -1$ corresponds to a team that attacks in the opposite direction. The direction is inferred in a robust sequence:

We compute the median x coordinate of all shot events taken by team j in period r. If there are too few shots in that period to reliably determine direction, we instead use the median x-coordinate of all events (e.g., passes, duels, etc.) by that team in the same period. If this still does not yield a clear attacking direction, we infer it from the direction identified for the opposing team in the same period or, if necessary, from the direction already established for the same team in the previous period. Finally, for a team that attacks towards increasing x, we set $s_{mjr} = +1$. Otherwise, we set $s_{mjr} = -1$.

Spatial normalization

For an event that belongs to match m , team j and period

r with raw coordinates (x_{raw}, y_{raw}), we define normalized coordinates (a_x, a_y) as

$$a_x = \begin{cases} x^{raw}, s_{mjr} = +1 \\ L - x^{raw}, s_{mjr} = -1 \end{cases},$$

$$a_y = \begin{cases} y^{raw}, s_{mjr} = +1 \\ W - y^{raw}, s_{mjr} = -1 \end{cases}$$

Thus, all attacking events for a team are mapped into a unified frame in which the goal they attack lies on the right-hand side.

We place the centre of the attacking goal at $(x_g, y_c) = (L, W/2)$. The left and right posts have coordinates

(x_g, y_L) and (x_r, y_R), where $y_L = W/2 - \omega_{goal}/2$ and $y_R = W/2 + \omega_{goal}/2$, and ω_{goal} is the width of the goal.

Absolute time within match

Event time stamps are recorded relative to the start of the period. To compare and order events on a shared time axis, we define an absolute time in seconds. Let t_{mjr}^{rel} be the recorded time in period r. Let C_r fixed offset in seconds for each period, for example zero for first half, 45×60 for

second half, 90×60 for the first extra period, and 105×60 for the second extra period. The absolute time is

$$t^{abs} = C_r + t^{rel}_{mjr}$$

This absolute time variable is used to align sequences of events and to define the matching between shots and subsequent goals.

Shot Sample and Outcome Labelling

Identification of shots and goals

We construct the analysis sample by extracting all events registered as shots. Let the set of all shot events be indexed by $i = 1, \dots, N$. For each shot we retain the match identifier m_i, team identifier j_i, player identifier p_i, normalized coordinates $(a_{x,i}, a_{y,i})$, and absolute time t_i^{abs}.

Goal events are identified in a similar way using the event type. For each goal we record the match, team, normalized coordinates (a_x^{goal}, a_y^{goal}) and absolute time t_{goal}^{abs}.

We restrict attention to shots taken in regulation time and extra time, that is first half, second half, and both extra periods. Penalty shootouts are excluded.

Matching shots to subsequent goals

For each match m, team j and period r we sort both shots and goals by absolute time. For a given shot i belonging to match m_i and team j_i, we search for a goal by the same team that occurs in a time window immediately after the shot. The window length is 10 seconds, which is sufficient to cover rebounds and very short sequences without merging independent attacks.

Formally, define the set of candidate goals for shot i as

$$\mathcal{G}_i = \{\, g : m_g = m_i,\ j_g = j_i,\ t_g^{\text{abs}} \in \left[t_i^{\text{abs}},\ t_i^{\text{abs}} + \Delta\right] \},$$

where Δ=10 seconds. If $\mathcal{G}_i$ is non-empty, then shot i is labelled as a goal. If $\mathcal{G}_i$ is empty but the event tags associated with shot i indicate that it directly resulted in a goal, we also label it as a goal. All remaining shots are labelled as misses. The binary outcome variable is

$$y_i = \begin{cases} 1, & if\ shot\ i\ is\ matched\ to\ a\ goal \\ 0, & otherwise \end{cases}$$

Feature Engineering

We construct a feature vector x_i or each shot i. The features are grouped into geometric features, temporal and match dynamics, previous event features, player action features, and hierarchical encodings for players and teams.

Geometric features

The Euclidean distance from the shot location to the centre of the goal is

$$\text{dist}_i = \sqrt{\left(x_g - a_{x,i}\right)^2 + \left(y_c - a_{y,i}\right)^2}.$$

We also compute the opening angle to the goal posts. The angles from the shot location to the left and right posts are

$$\theta_{L,i} = \arctan 2\left(y_L - a_{y,i},\ x_g - a_{x,i}\right),$$
$$\theta_{R,i} = \arctan 2\left(y_R - a_{y,i},\ x_g - a_{x,i}\right).$$

and the goal angle is

$$\alpha_i = |\theta_{R,i} - \theta_{L,i}|.$$

We use both the angle in radians and derived transformations such as $cos(\alpha_i)$. The longitudinal distance from the goal line and the lateral offset from the central line are

$$x_i^{\text{front}} = x_g - a_{x,i}, \quad y_i^{\text{lat}} = a_{y,i} - y_c, \quad |y_i^{\text{lat}}|$$

These variables capture how deep in the penalty area the shot is taken and how far it is from the centre of the goal. We also consider interaction and non-linear terms such as

$$\frac{1}{1 + \text{dist}_i}, \qquad \text{dist}_i^2, \qquad \alpha_i^2, \qquad \text{dist}_i \cdot \alpha_i.$$

These terms allow the logistic model to capture curvature without relying on very high model complexity.

Temporal and match dynamics

For each shot we compute the minute of the match as

$$\text{minute}_i = \frac{t_i^{\text{abs}}}{60}.$$

To capture match dynamics, we construct goal counts for the shooting team and the opponent up to just before shot i. Let

GF_i denote the number of goals scored by the shooting team before time t_i^{abs}, and GA_i the goals conceded. The score difference at the time of the shot is

$$\text{scoreDiff}_i = \text{GF}_i - \text{GA}_i.$$

This variable expresses whether the team is leading, drawing, or trailing at the moment of the shot. We also consider the absolute value of score difference and simple categorical groupings for interpretation in downstream analysis.

Previous event features

To describe the immediate attacking context, we examine the event that precedes the shot for the same team in the same match. For each shot i we identify the previous event index k_i in that team's event sequence. We record the previous event's coordinates $(a_{x,k}, a_{y,k})$ and the time t_k^{abs}. The time difference is

$$\delta_{t_i} = t_i^{abs} - t_k^{abs}$$

and the displacement vector from the previous event location to the shot location is

$$\Delta x_i = a_{x,i} - a_{x,k}, \qquad \Delta y_i = a_{y,i} - a_{y,k}.$$

The total displacement and an implied speed of ball or attacking movement are

$$\text{distPrev}_i = \sqrt{(\Delta x_i)^2 + (\Delta y_i)^2}, \text{speedPrev}_i = \frac{\text{distPrev}_i}{\Delta t_i}.$$

for $\Delta_i > 0$, We also compute the lateral displacement from the previous event

$$\text{latPrev}_i = |(a_{y,i} - y_c) - (a_{y,k} - y_c)|.$$

From the event tags of the previous action we derive indicator variables that capture how the chance was created. Examples include indicators for through ball, direct free kick, indirect free kick, corner, or cross. These features capture the idea that not all shots from the same location are equally dangerous, because a clean through ball or a fast cutback can create better shooting conditions.

Player action and footedness

We parse the tags associated with the shot to identify whether the attempt was taken with the left foot, the right foot, or the head. This yields binary variables such as $leftFoot_i, rightFoot_i, header$.

From the player table we extract the dominant foot of each player, denoted $f_{p_i} \in \{left, right, both\}$, We then define a weak foot indicator

$$\text{weakFoot}_i = \begin{cases} 1, & if\ shot\ uses\ the\ non-dominant \\ 0, & otherwise \end{cases}$$

This variable captures potential reductions in scoring probability when the shot is taken with the non-dominant foot.

Hierarchical encodings for players and teams

To represent latent finishing skill and team style without introducing a very large number of categorical parameters, we use out-of-fold target encoding for players, attacking teams, and opponents.

Consider an entity e which can be a player, a team, or an opponent team. Let S_e be the set of shots taken by that entity in the training data, and let $n_e = |S_e|$ be the number of such shots. The empirical goal rate for entity e is

$$\widehat{\mu_e} = \frac{1}{n_e} \sum_{i \in S_e} y_i.$$

Let μ be the overall mean goal rate in the training set. We define a smoothed estimate

$$\widetilde{\mu_e} = \frac{n_e \widehat{\mu_e} + \lambda \mu}{n_e + \lambda}.$$

where $\lambda > 0$ is a regularization strength that shrinks entities with few observations towards the global mean. To avoid optimistic bias, we compute these encodings in an out-of-fold manner. We partition the training data into K folds grouped by match, fit $\widehat{\mu_e}$ on K−1 folds, and assign the encoding to the held-out fold. This yields three encoded features: one for shooter, one for the attacking team, and one for the opponent.

Model Specification and Training Procedure

Logistic regression with engineered features

The main model is a logistic regression with the engineered features described above. For each shot i we have a feature vector $x_i \in R^p$. The model specifies the goal probability as

$$\Pr(y_i = 1 \mid x_i) = \sigma(\beta_0 + x_i^{\top} \beta).$$

where, $\sigma_z = \frac{1}{1+e^{-z}}$ is the logistic function, β_0 is an intercept, and β is the vector of the coefficients.
We include both linear terms and selected quadratic and interaction terms inside x_i, for example $dist_i^2$ and $dist_i \cdot \alpha$. This allows the model to capture non-linear relationships while retaining interpretability through explicit coefficients.
The parameters are estimated by maximizing the regularized log-likelihood

$$\mathcal{L}(\beta_0, \beta) = \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] - \lambda_{\text{reg}} ||\beta||_2^2.$$

where $\rho_i = Pr(y_i = 1|x_i)$ and λ_{reg} is a regularization parameter that controls the strength of the ridge penalty.

Tree based baseline models

To compare the performance of the logistic model with more flexible alternatives, we also train tree-based models on the same feature set. These include a random forest classifier and a gradient boosting classifier. These models use the same input features but allow complex non-linear interactions through ensembles of decision trees.
The tree-based models are not the main focus of the work, but they serve as reference points for predictive performance and for interpretability analyses based on Shapley values.

Train–Test Split and Cross Validation

To prevent information leakage between matches and to respect the temporal and tactical structure of individual

games, we split the data by match rather than by individual shot.

Train–test partition

Let $\mathcal{M}$ be the set of all matches. We randomly partition

$\mathcal{M}$ into a training subset $\mathcal{M}_{train}$ a test subset $\mathcal{M}_{test}$, with approximately 80 percent of matches assigned to training and 20 percent to testing. All shots from matches in

$\mathcal{M}_{train}$ form the training set, and all shots from matches in $\mathcal{M}_{test}$ form the test set. This ensures that shots from the same match never appear in both training and test sets.

Cross validation grouped by match

Within the training set, we apply cross-validation that preserves match groupings and balances the outcome distribution by using a stratified grouped K-fold scheme: for each fold k, we select a subset of matches for the validation fold such that the proportion of goals and misses is approximately the same across all folds, assign all shots from these matches to the validation subset, and use the remaining matches (and their shots) as the fitting subset for that fold.

For fold k, the model is fitted on the fitting subset and generates predictions ρ_i^k for each shot i in the validation subset. Collecting predictions over all folds yields out-of-fold predictions on the entire training set. These out-of-fold probabilities are used to fit the calibration model and to compute cross-validated performance metrics. The grouped scheme ensures that evaluation reflects the ability to generalize to unseen matches.

Calibration via Isotonic Regression

Logistic regression produces uncalibrated probabilities p_i^{raw}. To obtain probability estimates that align closely with observed frequencies, we apply isotonic regression as a non-parametric calibration step.

Let $(p_i^{raw}, y_i)_{I\in\tau}$ be the out-of-fold predictions and labels in the training set. We seek a non-decreasing function $g: [0,1] \to [0,1]$ that minimises the sum of squared errors

$$g^* = \arg\min_{g\in J} \sum_{i\in T} \left(y_i - g(p_i^{\text{raw}})\right)^2.$$

where I is the set of all non-decreasing functions on [0,1]. In practice we approximate g^* by a piecewise constant function computed by the pool adjacent violators algorithm.

The calibrated probability for shot i is then

$$\widehat{p_i} = g^*(r_i^{raw})$$

After choosing the final model, we refit it on the full training set, generate raw probabilities for both training and test shots, and apply the same isotonic function g^* to obtain calibrated probabilities for all shots.

Model Selection and Evaluation

We evaluate each candidate model on both cross-validated training predictions and on the held-out test set. The main performance criteria are:

i. Discrimination, measured by the area under the receiver operating characteristic curve (ROC AUC) and the area under the precision–recall curve (PR AUC).
ii. Calibration error, measured by the Brier score,
$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{p_i})^2.$$
iii. Classification quality at a threshold of 0.5, summarised by precision and recall at that threshold.

The primary selection criterion is cross-validated PR AUC, since the positive class is rare and the precision–recall curve is more sensitive to performance on rare events. If two models have similar discrimination, we favour the one with better Brier score after calibration.
Once the final model is selected, we retrain it on the full training set, calibrate it with isotonic regression, and then fix both the model and the calibration function for all subsequent analyses.

Aggregation for Spatial and Team Level Analysis

To support spatial and tactical interpretation, we aggregate calibrated probabilities and outcomes over locations and over teams.

For spatial analysis we partition the pitch into a grid of cells indexed by c. For each cell we compute the empirical conversion rate and the mean xG as

$$\text{empiricalConv}_c = \frac{\sum_{i\in c} y_i}{\sum_{i\in c} 1}, \quad \text{meanXG}_c = \frac{\sum_{i\in c} \widehat{p_i}}{\sum_{i\in c} 1},$$

The difference $meanXG_c - empiricalConv_c$ allows inspection of overestimation and underestimation in different regions of the pitch.
For team level analysis we aggregate by team j. Let S_j be the set of shots taken by team j. The total xG and total goals for that team are

$$\text{xG}_j = \sum_{i\in S_j} \widehat{p_i}, \quad \text{Goals}_j = \sum_{i\in S_j} y_i.$$

The relation between xG_j and $Goals_j$ across teams illustrates team level calibration and supports tactical interpretation of which teams created chances of higher quality than their raw goal totals suggest.

V. RESULTS

From the 2018 FIFA World Cup event dataset, we extracted $n = 1419$ shots and 128 goals, yielding a base conversion rate of 9.02%. The group-aware split resulted in 1153 training shots (104 goals, 9.0%) and 266 test shots (24 goals, 9.0%).

Model Selection and Discriminative Performance

Using the protocol described in Section, we compared several classifiers and selected the model with the highest out-of-fold (OOF) PR-AUC on the training set. Logistic regression achieved the best OOF PR-AUC (0.319) and strong discrimination (OOF ROC-AUC $\approx$ 0.754) while remaining stable under calibration. On the held-out test set, the selected model achieved ROC-AUC of 0.778 and PR-AUC of 0.325 (Table I). Comparative ROC and PR curves for all models are shown in Figure 1.

TABEL I: TEST-SET PERFORMANCE OF THE SELECTED MODEL

Metric	Value
ROC-AUC	0.7779
PR-AUC	0.3248
Brier score (raw)	0.1763
Brier score (calibrated)	0.0767
Precision@0.5(calibrated)	1.0000
Recall@0.5(calibrated)	0.08333

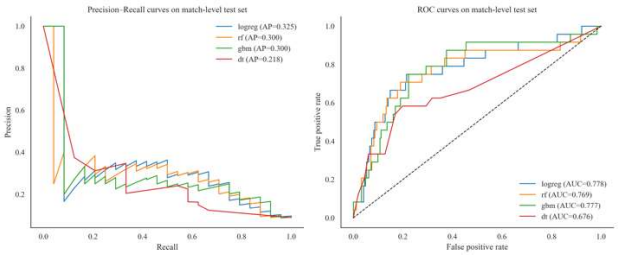

Figure1 Discriminative performance on the match-level test set.

Probability Calibration and Reliability

Isotonic regression markedly improved probability accuracy. On the test set, the Brier score decreased from 0.176 (raw) to 0.077 (calibrated). Reliability curves indicate close alignment to the diagonal across quantile bins (Figure 2).

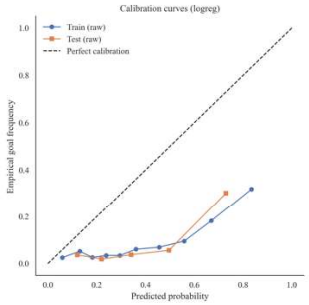

Figure 2. Reliability analysis for the selected model: empirical goal frequency vs. predicted probability on train (OOF) and test.

Spatial Behaviour and Face Validity

Predicted xG adheres to established soccer regularities. Mean xG concentrates centrally and at short range (Figure 3), and empirical conversion patterns match this structure (Figure 4). Location-wise calibration maps show small, spatially localized deviations between predicted and empirical rates .

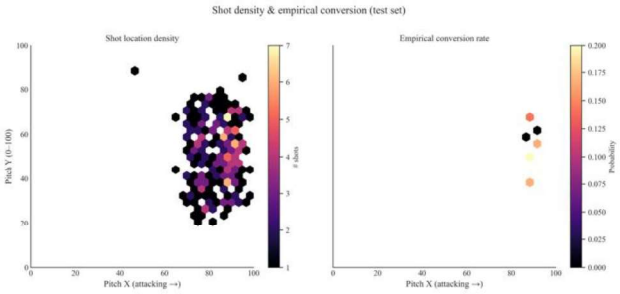

Figure 3. Shot-location density (left panel) and empirical conversion rate (right panel) on the test set.

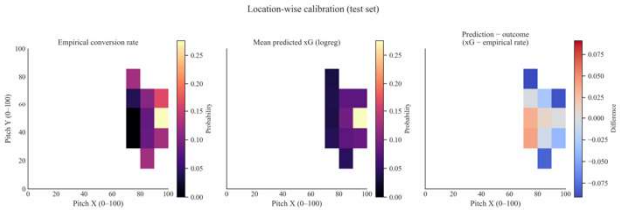

Figure 4. Location-wise calibration: empirical goal rate, mean predicted xG, and their difference (xG - empirical) on the test set.

Binned response curves confirm that both empirical conversion and predicted xG decrease with distance and with more acute angles. The model tracks these monotone trends closely (Figure 5).

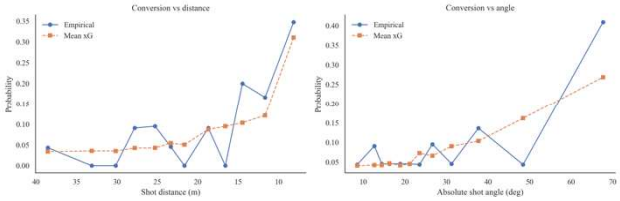

Figure 5. Empirical conversion (solid) and mean predicted xG (dashed) across binned shot distance (left) and absolute angle (right) on the test set

At team level, aggregate xG aligns with realized goals near the identity line (Figure 6), and predicted xG distributions show clear separation between goals and non-goals (Figure 7), supporting calibration and separability.

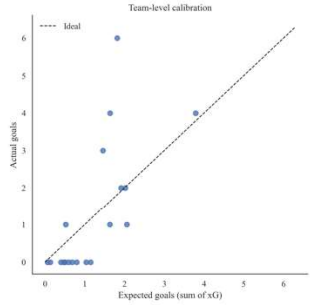

Figure 6. Team-level calibration on the test set: total xG vs. total goals with identity line.

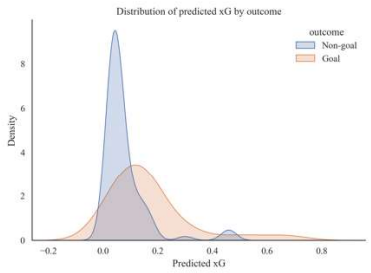

Figure 7. Predicted xG distributions for goals vs. non-goals on the test set.

5.4 Interpretability via SHAP

Although the best-performing model is logistic regression (logreg), for richer local explanations we additionally trained a comparably performing tree-based model (e.g. Gradient Boosting or XGBoost) and used SHAP values for global and local attribution. This design preserves the calibrated, parsimonious logreg model for reporting and aggregation, while leveraging tree-model SHAP to examine non-linear interactions in a post hoc, interpretation-first manner.

Consistent with soccer geometry, SHAP importance ranks shot distance and angle among the dominant contributors, with larger distances contributing negatively and more favorable shooting angles contributing positively to goal probability. Contextual features, which including pre-shot score differential and previous-event dynamics (time gap, displacement, speed), provide additional, situation-aware signal. Technical-action indicators (weak-foot, headers) exhibit intuitive effects (weak-foot often decreasing probability, all else equal), while hierarchical encodings (player/team/opponent OOF priors) capture persistent finishing/defending tendencies without leaking target information. Local SHAP explanations further reveal heterogeneous pathways to chance quality across pitch zones and play patterns, complementing the aggregate spatial diagnostics.

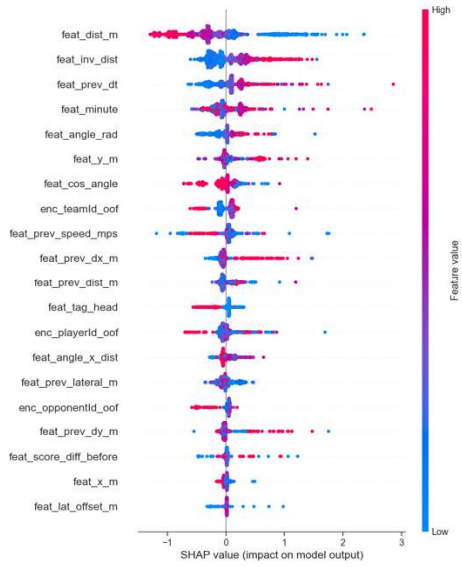

Figure 8. Global SHAP summary (tree model). Higher absolute SHAP values indicate stronger feature influence on predicted goal probability; color encodes feature value.

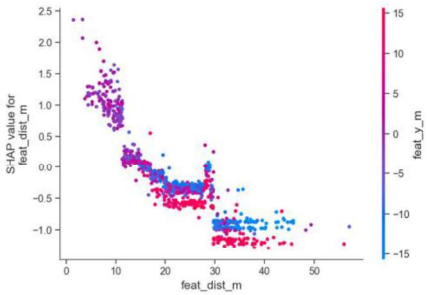

Figure 9. SHAP dependence for shot distance (meters). Points show per-shot contributions; color can reflect an interacting covariate if provided.

The proposed calibrated xG model exhibits competitive discrimination and strong probability calibration on the held-out test set. Spatial diagnostics and response curves indicate face-valid patterns consistent with the game's geometry. Aggregation to team level preserves calibration, making the estimates suitable for match- and tournament-level analyses.

VI. DISCUSSION

Interpretation of Model Behaviour

The selected logistic regression model, trained on geometry, immediate temporal context, technical-action tags, and hierarchical encodings, achieves test ROC-AUC ≈ 0.778 and PR-AUC ≈ 0.325. These values are consistent with reported ranges for event-only xG models in elite competitions, reflecting meaningful discriminative ability under class imbalance. Notably, isotonic calibration reduces the test Brier score from 0.176 to 0.077, indicating well-calibrated probabilities suitable for aggregation across time and entities (players, teams, matches).

Spatial diagnostics reveal strong face validity. Mean predicted probability concentrates in central, close-range regions, mirroring empirical conversion (Figure 3). Location-wise calibration maps (Figure 4) show small, localized deviations, which is expected given finite sample sizes in spatial bins and unmodelled contextual factors (e.g., defensive pressure, goalkeeper positioning). Distance and angle response curves (Figure 5) further support that the model captures monotone geometrical effects without pathological artefacts.

Operational Considerations

Thresholder metrics at 0.5 show very high precision but low recall ($Precision = 1.00$, $Recall = 0.083$), a common outcome under class imbalance with conservative calibration. For decision-support use cases that value coverage (e.g., identifying chance quality across a full match), alternative operating points or cost-sensitive thresholds should be considered. Because the model is well calibrated, expected value analyses and cumulative xG aggregates are robust to threshold choices.

At the team level, predicted xG aligns well with realized goals (Figure 6), suggesting that aggregation preserves calibration and supports comparative analyses (e.g., performance sustainability, finishing variance).

Outcome-stratified xG distributions (Figure 7) show clear separation, reinforcing that the model provides informative probability mass even when events are rare.

Limitations

The model is built on event data without freeze-frame features; thus, it lacks direct measures of pressure, defensive proximity, goalkeeper position, and obstruction. Possession-sequence context is approximated via the previous event; richer sequence modelling could capture attack dynamics more faithfully. The evaluation is limited to a single tournament (64 matches), which constrains external validity; transportability should be tested across seasons and leagues. Finally, hyperparameters were intentionally not exhaustively tuned to emphasize methodological clarity and probability calibration; modest gains may be achievable with targeted tuning or assembling.

Implications and Future Directions

For practitioners, calibrated xG supports: (i) match analysis and post-match review via spatial maps and aggregates, (ii) player and team scouting with hierarchical priors, and (iii) research into tactical patterns through location-wise residuals. Future work should integrate freeze-frame covariates (pressure, angles to nearest defenders, GK depth), pitch control estimates, and multi-event sequence representations. Calibration could be revisited with Bayesian or hierarchical approaches to stabilize estimates in low-sample spatial regions. Cross-competition validation will clarify generalization and inform domain adaptation strategies.

VII. CONCLUSION

We develop a transparent, calibrated xG pipeline for the 2018 FIFA World Cup that combines interpretable geometry, immediate temporal context, technical-action indicators, and out-of-fold hierarchical encodings to mitigate leakage. Across several classifiers, logistic regression delivers the best precision–recall trade-off, and isotonic calibration substantially improves probability accuracy (Brier 0.176 $\rightarrow$ 0.077) without sacrificing discrimination. Spatial diagnostics and aggregation behaviour exhibit face-valid patterns, and team-level calibration indicates suitability for applied analysis.

The framework is readily extensible: adding freeze-frame and pitch-control features, enhancing sequence context, and validating across leagues are natural next steps. Given its calibration quality and interpretability, the model provides a solid foundation for both academic study and practitioner-oriented performance analysis.

REFERENCES

Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. Plos one, 18(4), e0282295.

Brechot, M., & Flepp, R. (2020). Dealing with randomness in match outcomes: how to rethink performance evaluation in European club football using expected goals. Journal of Sports Economics, 21(4), 335-362.

Smith, R. (2022). Expected goals: The story of how data conquered football and changed the game forever.

Fairchild, A., Pelechrinis, K., & Kokkodis, M. (2018). Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality. Journal of Sports Analytics, 4(3), 165-174.

Scholtes, A., & Karakuş, O. (2024). Bayes-xG: player and position correction on expected goals (xG) using Bayesian hierarchical approach. Frontiers in Sports and Active Living, 6, 1348983.

Cavus, M., & Biecek, P. (2022, October). 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). IEEE.

Iapteff, L., Le Coz, S., Rioland, M., Houde, T., Carling, C., & Imbach, F. (2025). Toward interpretable expected goals modeling using Bayesian mixed models. Frontiers in Sports and Active Living, 7, 1504362.

Brechot, M., & Flepp, R. (2020). Dealing with randomness in match outcomes: how to rethink performance evaluation in European club football using expected goals. Journal of Sports Economics, 21(4), 335-362.

Bandara, I., Shelyag, S., Rajasegarar, S., Dwyer, D., Kim, E. J., & Angelova, M. (2024). Predicting goal probabilities with improved xG models using event sequences in association football. Plos one, 19(10), e0312278.

Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. Plos one, 18(4), e0282295.

Madrero Pardo, P. (2020). Creating a model for expected goals in football using qualitative player information (Master's thesis). Universitat Politècnica de Catalunya.

Hewitt, J. H., & Karakuş, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). Franklin open, 4, 100034.

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. International Journal of Sports Science & Coaching, 14(6), 798-817.

Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). Frontiers in sports and active living, 3, 624475.

Fairchild, A., Pelechrinis, K., & Kokkodis, M. (2018). Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality. Journal of Sports Analytics, 4(3), 165-174.

Kumar, A., Liang, P. S., & Ma, T. (2019). Verified uncertainty calibration. Advances in neural information processing systems, 32.