To what extent are probabilistic and regression-based models such as expected goals (xG) useful in evaluating team performances in professional football?

Owen Liu

Introduction

Expected goals (xG) has become a central metric in modern football analytics. It aims to quantify the quality of a goal-scoring opportunity by assigning a probability that a given shot will result in a goal. Expected goals allow us to measure the quality of chances during a game, which provides a way to account for the unpredictability of a low-scoring sport like football. From a mathematical standpoint, expected goals may be viewed as a probabilistic model, where the outcome of a shot is treated as a binary variable (goal, or no goal).

Sports analytics companies construct expected goals models using large datasets and machine learning techniques. Opta's model, for instance, is trained using the data from nearly one million historical shots. Not only is a large collection of data used, but also over twenty variables are incorporated into the calculations, including distance, angle and defensive pressure.[1] While the exact equations are not publicly available, this approach is consistent with statistical methods such as logistic regression, which models probabilities based on explanatory variables.

1 Mathematical Foundations of Expected Goals

Expected goals (xG) can be mathematically understood as a probabilistic model designed to estimate the likelihood that a specific shot results in a goal.[2] Each shot is treated as a binary outcome, simply meaning two possibilities - that being, either a goal (1) or not a goal (0) - and a probability is assigned based on the circumstances and characteristics of the given shot.[3] In football, xG values range between 0 and 1 where a value which is closer to one suggests a greater chance of scoring. An example of this would be an xG value of 0.79 suggesting that a shot of similar characteristics usually results in a goal 79% of the time according to historical data.

1.1 Logistic Regression

A pivotal mathematical method for estimating expected goals is logistic regression. This is a probabilistic technique used to model binary outcomes,[4] in this case whether a shot results in a goal or not. Logistic regression predicts the probability ***p*** that a particular shot will be scored by taking a combination of explanatory variables into account. These variables may include goalkeeper position, shot type or pattern of play. The probability can be expressed with the following function:

$\boldsymbol{p} = 1/(1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_k x_k)})$

where $x_1, x_2, \ldots, x_k$ represent the characteristics of a shot such as the distance from goal or shot type. The coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are estimated from historical data.

Example: Suppose $\beta_0 = -2$ $\beta_1 = -0.1$ $\beta_2 = 0.03$
We can consider a realistic shot where the distance (x_1) = 12m and the angle (x_2) = 45°

---

[1]1 J. Whitmore, What is expected goals (xG)?, 2023. Available at: https://theanalyst.com/articles/what-is-expected-goals-xg (Accessed: 25 March 2026).
[2] Mead J., O'Hare A. and McMenemy P., 'Expected goals in football: Improving model performance and demonstrating value', PLoS One, vol. 18, no. 4, 2023, e0282295. Available at: https://doi.org/10.1371/journal.pone.0282295 (Accessed: 25 March 2026).
[3] Anzer G. and Bauer P., 'A goal scoring probability model for shots based on synchronized positional and event data in football (soccer)', Frontiers in Sports and Active Living, vol. 3, 2021, 624475. Available at: https://doi.org/10.3389/fspor.2021.624475 (Accessed: 25 March 2026).
[4] DeMaris A., 'A tutorial in logistic regression', Journal of Marriage and Family, vol. 57, no. 4, 1995, pp. 956–968. Available at: https://doi.org/10.2307/353415 (Accessed: 26 March 2026).

$\boldsymbol{p} = 1/(1+e^{-(-2+(-0.1 \times 12)+(0.03 \times 45))})$
$\boldsymbol{p} = 0.1358\ldots \approx 0.14$
Using the logistic regression formula with a distance of 12m and angle of 45 degrees, the calculated xG is 0.14. This can also be interpreted using expected value. If we define a random variable X, where X = 1 for a goal and X = 0 for a miss, then
$E[X] = (1 \times p) + (0 \times (1-p)) = \boldsymbol{p}$
This interpretation is consistent with standard probability theory, where the expected value of a Bernoulli random variable is equal to its probability.[5] Therefore the xG value of 0.14 represents the expected number of goals from this shot.

The coefficient β_0, called the intercept, represents the baseline log-odds of a shot resulting in a goal when all other variables are zero. Each variable, such as distance, adjusts the baseline to give a final probability of scoring.[6] The coefficients (β_i) quantify how strongly each variable influences the outcome (goal or no goal).[7] Some variables have larger coefficients than others, indicating a stronger effect on scoring probability. For instance, distance from goal typically has a negative coefficient, meaning that shots taken closer to the goal are much more likely to be scored.[8] Conversely, a 90-degree shooting angle would imply that the goal is perfectly aligned in front of the shooter, allowing for a higher chance to score a goal.[9]

1.2 Machine Learning

Logistic regression provides a clear mathematical framework for estimating expected goals, however modern models often use more advanced machine learning techniques to improve predictive accuracy. Most organisations, such as Opta and Statsbomb, use a model called XGBoost (extreme gradient boost), which is based on a method which is known as gradient boosting.[10] This approach builds upon simpler models called decision trees, which make predictions by splitting data based on conditions such as distance from goal or shooting angle.[11]

Decision trees work by repeatedly dividing the data according to different variables. For example, the model may first split shots based on whether the distance is less than a certain value, then further divide those shots based on them having a shooting angle. Each final group is assigned a probability of scoring, producing an xG value. Unlike logistic regression, which assumes a linear relationship between variables and the log-odds of scoring, decision trees are able to model more complex and non-linear relationships between variables.[12]

Gradient boosting improves upon this by combining numerous decision trees. Each new tree is constructed to correct the errors made by the previous trees. After a tree makes its predictions, the model calculates the residuals (the difference between the predicted probabilities and the actual outcomes). The next tree is trained to predict these residuals, correcting the mistakes of the previous tree. This process is repeated many times, gradually improving the accuracy of the model. By combining all the trees, gradient boosting captures complex patterns and interactions, allowing xG models to account for non-linear relationships between shot features.[13] For

---

[5] J.K. Blitzstein and J. Hwang, Introduction to Probability, Boca Raton: CRC Press, 2015, sec. 3.3 (Bernoulli and Binomial), pp. 112–115. Available at: https://uni.dcdev.ro/y2s2/ps/Introduction%20to%20Probability%20by%20Joseph%20K.%20Blitzstein,%20Jessica%20Hwang%20(z-lib.org).pdf
[6] Pennsylvania State University, STAT 501: Regression methods – Logistic regression, 2026. Available at: https://online.stat.psu.edu/stat501/book/export/html/1011 (Accessed: 26 March 2026).
[7] Pennsylvania State University, STAT 462: Applied regression analysis – Logistic regression, n.d. Available at: https://online.stat.psu.edu/stat462/node/207/ (Accessed: 26 March 2026).
[8] Anzer G. and Bauer P., 'A goal scoring probability…'Available at: https://doi.org/10.3389/fspor.2021.624475
[9] Anzer G. and Bauer P., 'A goal scoring probability…'Available at: https://doi.org/10.3389/fspor.2021.624475
[10] Chen T. and Guestrin C., 'XGBoost: A scalable tree boosting system', arXiv preprint, 2016. Available at: https://arxiv.org/abs/1603.02754 (Accessed: 27 March 2026).
[11] Sagi O. and Rokach L., 'Approximating XGBoost with an interpretable decision tree', Information Sciences, vol. 572, 2021, pp. 522–542. Available at: https://www.sciencedirect.com/science/article/pii/S0020025521005272 (Accessed: 28 March 2026)
[12] James G., Witten D., Hastie T. and Tibshirani R., An introduction to statistical learning with applications in R, 2nd edn. New York: Springer, 2023, ch. 8, pp. 327–339.
[13] Chen T. and Guestrin C., 'XGBoost: A scalable tree boosting system'

example, shots taken from central positions inside the penalty box may have a high probability of scoring, but this probability could decrease if multiple defenders are positioned between the shooter and the goal. This demonstrates how the effect of one variable can depend on another variable. Therefore, models such as XGBoost can account for the interactions between variables.

However, despite their increased accuracy, machine learning models such as XGBoost have limitations. One key issue is that these models are often considered “black box” models. This means that it can be difficult to interpret how individual variables contribute to the final xG value.[14] Whereas, logistic regression provides coefficients which clearly show how each variable affects the likelihood of scoring.[15] This lack of interpretability can be problematic when evaluating team or player performance, as it may be unclear why a particular xG value has been assigned to a certain shot. Furthermore, machine learning models require large datasets (such as the historical collection of over one million shots used by Opta) to perform effectively, which may not always be available in all contexts.[16] Overall, while models such as XGBoost may improve predictive accuracy, logistic regression remains valuable due to its simplicity and transparency.

1.3 Shot Features and Data Inputs

Expected goals models rely on the selection of shot features and data inputs used to estimate the probability of scoring. Both logistic regression and advanced machine learning models rely on a range of variables, which describe the shot context. Common features include distance from goal and type of assist. These features are chosen because they have a clear impact on the likelihood of the goal being scored and can be quantified using a historical database of shots.[17]

In logistic regression models, the importance of each feature is determined by the magnitude of its coefficient. Variables with larger coefficients have a greater influence on the xG value. For example, distance from goal typically has a large negative coefficient, indicating that shots which are taken closer to the goal are more likely to result in a goal. This allows analysts to interpret how each factor contributes to the overall xG in a clear and measurable way.

However, feature selection in xG models has limitations. Contextual factors such as a player’s form, finishing skill or atmosphere can influence the outcome of a shot, but are difficult to include in an xG model. As a result, even highly sophisticated models may not fully capture the genuine quality of a chance. Suggesting that while xG can be useful for evaluating performance, it in isolation should not be relied on as it can overlook qualitative aspects of the game.[18]

2 Practical Usefulness of xG

Expected goals provides a nuanced evaluation of team and player performance than traditional statistics, such as goals scored. One of the main uses of xG is assessing team performances during matches by capturing the quality of shooting chances rather than the quantity. A team’s actual number of goals scored in a match can be compared to its cumulative xG. Teams that consistently score less goals than predicted by their season-total xG values may be considered wasteful with their chances. Whereas teams who outperform their xG could be considered clinical. By quantifying the quality of the shots taken by a team during a match rather than the amount of shots, xG provides a more reliable measure of offensive performance.

Shot and xG comparison (2022/23 Premier League season)

Team	Number of shots	Total xG	Goals Scored

[14] Hassija V., Chamola V., Mahapatra A., Singal A., Goel D., Huang K., Scardapane S., Spinelli I., Mahmud M.S. and Hussain A., ‘Interpreting black-box models: A review on explainable artificial intelligence’, Cognitive Computation, vol. 16, no. 1, 2024, pp. 45–74. Available at: https://doi.org/10.1007/s12559-023-10179-8 (Accessed: 28 March 2026).
[15] James G., Witten D., Hastie T. and Tibshirani R., An introduction to statistical learning
[16] Chen T. and Guestrin C., ‘XGBoost: A scalable tree boosting system’
[17] Hudl, What are Expected Goals (xG)?. Available at: https://www.hudl.com/blog/expected-goals-xg-explained (Accessed: 29 March 2026)
[18] J. Pipping-Gamón, T. Feng and P. Sabin, Beyond expected goals: A probabilistic framework for shot occurrences in soccer, arXiv 2512.00203 [stat.AP], January 2026. Available at: https://arxiv.org/pdf/2512.00203 (Accessed: 29 March 2026).

Manchester United	592	69.70	58
Manchester City	600	81.97	94

[19]

In the 2022/23 Premier League season, Manchester United and Manchester City took a similar number of shots, but Manchester City generated a significantly higher xG. This indicates that Manchester City created higher-quality chances. Also, Manchester City overperformed by approximately 12 goals, whereas Manchester United underperformed by a similar margin. This suggests that Manchester City were more clinical in their finishing, whereas United struggled to convert chances efficiently.

Expected goals can also be useful for evaluating individual players. Summing a player's xG over a match or season can highlight who capitalises on chances. A player consistently exceeding xG may be an exceptional finisher, while one who underperforms may have low composure in front of goal. Tracking a player's xG over multiple matches or an entire season also allows analysts to identify trends in performance, such as consistent chance creation or recurring finishing issues. This information can be used to compare players, allowing coaches and scouts to identify which players are reliably effective in front of goal.

3 Limitations and Contextual Considerations of xG

Despite its many uses, xG is not without its limitations that must be considered. One major issue is that quick rebounds can produce multiple shots within a very short time span. This could create the illusion that a team created a lot of high-quality chances, even when they all stem from the same play. In extreme cases, the cumulative xG values can even exceed one implying that they were expected to score more than one goal. As well as this, xG only considers when a shot is taken, ignoring near-chances when tackles, blocks and interceptions possibly prevent dangerous shots from occurring.

Contextual factors also pose a challenge. Psychological pressure, fatigue and extreme match circumstances are not included in standard xG models, despite their strong influence. An example of this is when Randal Kolo Muani took a last-minute shot in the 2022 World Cup Final. A lack of defenders between him and the goal, distance from the goal and fairly straight angle towards goal led to the xG calculated being 0.30. However in reality the extreme situational pressure, the goalkeeper Emiliano Martinez's skill and form, and the mental demands of the moment made the chance much more difficult than the numbers alone suggest.

Additionally, xG models are subject to statistical variance, even if a team generates an expected total of 2.0 xG, the actual number of goals is subject to probabilistic variation, meaning significant deviations from expectations can occur purely by chance. This can reflected mathematically, as the variance in total goals is given by $Var(G) = \Sigma p_i(1-p_i)$, where p_i is the probability of each shot resulting in a goal.

[19] StatMuse, Most shots in Premier League by team, 2022/23 season . Available at: https://www.statmuse.com/fc/ask?q=most+shots+in+premier+league+by+team+in+22%2F23+season (Accessed: 1 April 2026).