Franklin Open 4 (2023) 100034

Contents lists available at ScienceDirect

Franklin Open

journal homepage: www.elsevier.com/locate/fraope

A machine learning approach for player and position adjusted expected goals in football (soccer)

James H. Hewitt, Oktay Karakuş *

Cardiff University, School of Computer Science and Informatics, Abacws, Senghennydd Road, Cardiff, CF24 4AG, UK

ARTICLE INFO

Keywords:
Expected goals
xG
Football
Soccer
Machine learning
Player adjusting
Position adjusting

ABSTRACT

Football is a very result-driven industry, with goals being rarer than in most sports, so having further parameters to judge the performance of teams and individuals is key. Expected Goals (xG) allow further insight than just a scoreline. To tackle the need for further analysis in football, this paper uses machine learning methods that are developed and applied to Football Event data. The proposed solution utilises StatsBomb as the data provider and an industry benchmark to tune the models in the right direction. The proposed ML solution for xG is further used to tackle the age-old cliche of: 'the ball has fallen to the wrong guy there'. To investigate this, we tackle Positional Adjusted xG, splitting the training data into Forward, Midfield, and Defence to provide insight into player qualities based on their positional sub-group. Positional Adjusted xG successfully predicts and proves that more attacking players are better at accumulating xG. Finally, this study has further developments into Player Adjusted xG to prove that Lionel Messi is statistically at a higher efficiency level than the average footballer. Thanks to this analysis, we conclude that the Messi model performs 347 xG higher than the general model outcome.

1. Introduction

Football is one of the lowest-scoring games, due to its single-scoring system where every goal is worth one point compared to for example Rugby where the minimum isolated event is worth three points (a penalty kick). This means that every chance is highly valuable. This emphasises the importance of being able to convert goals from scoring opportunities. Having predictions and probabilities of each chance has been proven to have a significant competitive advantage. If players' impact is only judged on converted goals when they are so rare, then, it is very possible to miss their actual impact.

The main focus of this paper is to create and apply an Expected Goals (xG) model 'from scratch' and predict xG values with new and highly informative features. From this base level, this paper will document the development in the analysis of expected goals (xG). Finalising with novel *Position* and *Player adjusted xG*, to provide industry competitive advantages and improvement in academic knowledge through this publication.

xG can be seen as a probability, ranging between 0 and 1, stating the chance that each shooting opportunity is converted into a goal. With the value always being within the 0–1 range but never 0 or 1, this measure suggests that a certain Goal (xG = 1.0) or certain No Goal (xG = 0.0) is not feasible. The model details include such events as the method in which the player receives the ball, the technique they use to shoot and then painting freeze frame pictures of all possible factors influencing the shot output quality. Fig. 1 shows a freeze frame example of a shot, and details factors to be considered later on: shot location (distance and angle), opposition players surrounding the shooter, and opposition between the shot and goal. In the sequel, this section deep dives into the xG revolution and investigates the development of this valuable metric in terms of both the football industry and the academic perspective.

1.1. XG in football industry

The analytics revolution and widespread usage of xG are something relatively new. Statistics were somewhat shunned by classical old-fashioned pundits in the original stages. Possession stats are being heavily used to indicate 'running of a game' despite some teams being set up to counterattack and therefore choosing to have less possession of the ball. This possession-orientated school of thought was quickly dismissed. However, in 2017 the then Arsenal Manager Arsene Wenger, following a defeat to Manchester City stated: "If you look at the expected goals, it was 0.7 for them and 0.6 for us, it was a very tight game, they created very little, had a small number of shots on target, 0.1 more than us, that's all" [1]. This was not greeted well as Arsenal

* Corresponding author.
E-mail address: karakuso@cardiff.ac.uk (O. Karakuş).

https://doi.org/10.1016/j.fraope.2023.100034
Received 3 May 2023; Received in revised form 23 July 2023; Accepted 26 August 2023
Available online 30 August 2023

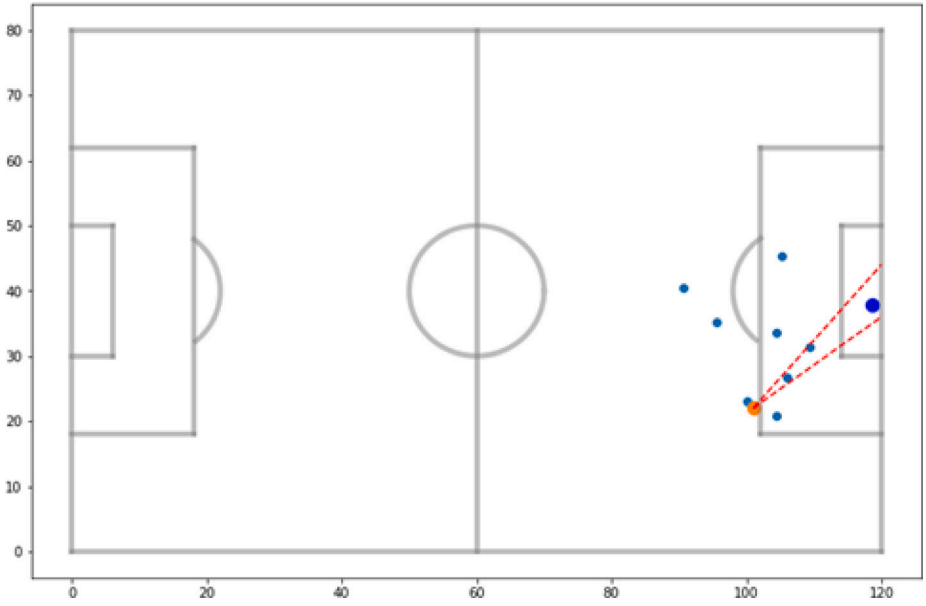

Fig. 1. A freeze frame example: shot location (distance and angle).

had lost that game 3–1. But what was going to happen from there was xG would only build traction.

At that time, it was correct for viewers and fans to be sceptical about the validity of xG in football. Following this, Sky Sports ran a report detailing early stage xG and specifically how "Expected goals also struggles to account for where the defenders are on the pitch — something that can dramatically impact the relative chances of two shots taken from precisely the same spot" [2]. Over the next few years, xG became more prominent with companies such as StatsBomb and Opta heavily utilising the metric in their data packages.

Underlying advanced metrics and analytics such as xG provide utility for *Performance Analysis* teams within the professional game, *Recruitment Departments* within those teams and also *Bettors/ Betting Companies*. Performance analysts refer to objective information used by athletes, coaches and analyst departments. Performance analysis includes video analysis of the game tape, including coaching analysis of tactics. Then, data-driven decisions are based on in-game performance such as pass completion rates, and sports science-related criteria such as heart rate monitor statistics. Analysts gain from xG by applying it to their team and opposition analysis in preparation or games using event and track data to build a number-driven approach to the game. xG allows the performance department objective evaluation on player performance, moving away from "he should've scored that" to "statistically that's a huge chance". Another application is gaining a competitive advantage by building visualisations that graphically show the areas in which xG is given up by opposing defences. This allows the potential implementation and exploitation of weaknesses in the opposition's defence. Goals may seem quite random to a casual view, however, xG can graphically provide a basis to utilise angles, distances and densities of a shooting zone and how that influences a potential goal. The idea of implementation of xG allows a different perspective for performance analysts, before the data revolution of the 2010s, was limited to game tape analysis and biased views on the difficulty of chances.

Effective recruitment analysis provides possibly the largest competitive advantage in all sports. A player's ability to be in the right zones and shooting positions is more indicative of potential success. This is where xG comes in. Finding high cumulative xG footballers at a young age allows cheaper signing and room for development. Brentford's rise is heavily linked to xG and this approach is detailed in [3]. Being able to identify undervalued players creates room for profits and improved team quality on a low budget. For attackers and attacking midfielders, xG as an underlying metric of quality can provide insight into which players may be worth signing before a goal-scoring run.

Betting companies also need the usage of xG, if a team is not conceding a lot but receiving a lot of xG against then its models will pick up on the trend. It would be fair to presume that a large element of luck behind their lack of goals was conceded considering a lot of xG against them. Likewise, if a team is not conceding goals and not conceding xG then it is safe to presume that the team is effectively defending. These underlying statistics allow a Betting Company to hedge its profit margins against values that in the past were not available to individual consumers.

As discussed in [3], *SmartOdds* and *Matthew Benham* used xG to make a stamp on the football industry. They first used expected goals, before its mainstream implementation, to 'win millions' each year by placing bets on football fixtures. With this same money, Benham purchased shares in Brentford and then used xG as a recruitment parameter and signed, as discussed before, Neil Maupay who funded a positive chain of events for the West London side. SmartOdds and Benham are living proof of the success of the application of xG.

1.2. Academic look to the xG metric

In contrast to the development in the football industry, betting companies and performance analysts, the expected goals metric is a relatively new venue with a large gap in published academic papers. This is due to the football industry keeping technological developments in-house to gain a competitive advantage against opponents. Up-to-date data is also privatised and requires large funding to acquire and test which is increasing the barriers to producing academic papers within the xG subject area.

One of the seminal pieces that started a certain data revolution is "The Expected Goals Philosophy" by James Tippett [3]. It discusses the development of the xG over time and highlights the evidence supporting xG's application to football within performance and recruitment (scouting). Tippett [3] begins discussing the "dialect" and analysis of football as "deeply flawed" and states that other sports such as Baseball, Basketball, Cricket and Hockey have already embraced strong mathematical approaches to performance analysis.

Players' xG values are covered and analysed to show the xG totals for individual players against their total goals. A player with a high xG

score suggests that they are often in goal-scoring positions, with this compared to total goals to judge how efficient they are with chances. This allows the reader to see if they are clinical, which would show a player over-performing their xG, if they are wasteful, or if they were underperforming their xG. Examples show Sadio Mane (Former Liverpool, now Al Nassr) overperforming by an "incredible" 5.24 xG in the 2018/19 Premier League season.

The application of the xG values helped clubs such as Brentford and Brighton recruit "undervalued players" and sell on for a profit. For example, Neil Maupay was transferred from St Etienne to Brentford before the 2017/18 season for a fee of €2 Million. Brentford then signed off on a further transfer for Maupay only two years later just before the 2019/20 season which saw Maupay move to Brighton for a fee of €15.56 million (TransferMarkt). This transfer fee of 678% inflated from the original value Brentford paid, along with Maupay being pivotal and scoring 41 goals in 95 appearances (TransferMarkt) for Brentford during their stay in the English Championship. The Maupay funds were invested into Ivan Toney in September 2020, who incidentally led the side to back-to-back playoff finals. From a transfer fee of €5.6 million, he helped promote the side and now is sitting on a €45 million market value (TransferMarkt) after two successful seasons in the Premier League.

At the time of writing, Tippett [3] suggested improvements worth considering in xG models, saying that the OptaSport model did not incorporate defender positioning into the xG output. Also, xG models do not include threatening attacks that do not result in a shot.

Lucey et al. [4] state that shots and shots on target values do not provide the true value of shot attempts using Spatio-temporal patterns to analyse the 10-second build-up to shot attempts. This paper assigns defender proximity-to-shooter to develop validity and to evaluate the shot quality and therefore determine whether a team was "dominant" or "lucky" with the outcome of the game. Further questioning whether the context of the scenario is important and proposing how probabilities change depending on the player attacking with regards to the opponent they are competing against within a contextual period of xG creation. In their research, Ruiz et al. [5] employ a dataset comprising 10,318 shots taken during the 2013/14 English Premier League season, categorised into open play footed shots, headers, free kicks, and penalty shots. To assess the scoring performance of Premier League footballers, the authors utilise a Multilayer Perceptron (MLP) with 6 hidden layers, constructing a goal expectancy model to estimate the conversion probability of these shots. By analysing the expected number of goals, the study provides valuable insights into striker scoring efficiency. The findings shed light on the goal-scoring capabilities of the players, enabling a comprehensive evaluation of their performance. This research contributes to the field of soccer analytics, offering a data-driven approach to objectively assess and compare the goal-scoring abilities of players in the Premier League. Eggels and Pechenizkiy [6] investigate the quantification of soccer player contributions within a match through the development of a method aimed at determining the expected winner of a match based on estimating the probability of scoring for individual goal-scoring opportunities. The research utilises data from three distinct sources: (1) main events during matches tracked by ORTEC employees, (2) data on player quality extracted from the soccer game FIFA via web sources, and (3) spatiotemporal data on players tracked by Inmotio using cameras during matches. This comprehensive dataset comprises information from seven different leagues over three seasons, encompassing a total of 5017 matches, during which 128,667 goal attempts were made, resulting in 14,109 goals. The authors employ various machine learning techniques, including Logistic Regression, Decision tree, Random Forrest, and Ada-boost, to predict the likelihood of scoring for each goal-scoring opportunity and integrate these probabilities to determine the match outcome. The study demonstrates the effectiveness of their proposed method in terms of classification performance and the calibration of probabilistic estimates. Additionally, the expected goals metric finds further application in evaluating seasons, matches, and individual players, providing valuable insights for soccer analytics and performance assessment.

In Fairchild et al. [7], based on a sample of 1115 non-penalty shots from 99 games in the Major League Soccer (MLS — USA Football League), a logistic regression model is utilised with extracted features of shot location coordinates, distance from the goal line, the angle from vertical posts, shot type (categorical variable — *cat.*), assist type (*cat.*) and play type (*cat.*). This paper concludes that spatio-temporal movements of players are essential along with fractal dimensions of ball trajectory. Their work has an important final statement of "*We envision our study to trigger more research in the connection between complex systems theory and sports.*". Following this, Herold et al. published a summary paper that provides interesting insight into machine learning applications of football [8]. The authors discuss the lack of context to xG models built at that time by sharing the statement *"They either did not acknowledge or capture opponent positioning and thus, failed to provide context that coaches and analysts can apply to the match"*.

Brechot and Flepp [9] deal with randomness in match outcomes with a large sample size of 7304 matches across four seasons in the top 5 leagues. Variables utilised include distance, angle, rule setting and body part. The empirical model used is a logistic regression for a binary response variable, finding that xG is the optimal source of a team's actual performance based on xG scorelines in comparison to the number of points collected in those games. Allowing objective assessment of performance and underperformance, Brechot and Flepp [9] state that short-run events are often based upon randomness and therefore unsustainable. Allowing clubs to be able to judge their players properly based on statistical performance rather than result outcome, therefore avoiding adjustments which may inhibit long-run performance and therefore results. The paper suggests accepting and working with randomness in football and using expected goals to navigate around its impacts, to allow sturdy decision-making processes.

In a university thesis from Universitat Politècnica de Catalunya (UPC) in accordance with FC Barcelona by Madrero [10], the authors utilise location-based, contextual data such as previous pass method and detailing the area of the pass whether inside or outside the box, and player related information based on the shot body part along with some further player-related information scraped from FIFA games. They propose using three models: Logistic Regression, XGBoost and Neural Networks. Distance to goal proves to be the most influential feature, regardless of player quality, but better 'rated' players have a larger shot rating that they can accumulate higher xG values and goals. From the La Liga sample, Messi is shown as the highest scorer and the highest in terms of cumulative xG. In the same sample, there is also team-based xG, where the number of goals and accumulated xG correlate positively with the teams' position in the table.

Umami et al. [11] use factors such as distance and angle of shot on goal. Utilising Distance and Angle as one combined variable, this paper demonstrates that it had a greater impact on calculating the xG. They also state that the process could be applied to further research to allow understanding of distance and angles influence on gaining and giving up xG and in turn goals. Another interesting development they suggest is an implementation of more difficult quantified variables such as surface type and condition, role play, and confidence.

In Cavus and Biecek [12], based on a large sample of 315,430 shots over seven seasons for European Top 5 leagues, the authors use an 'explainable' artificial intelligence approach to create 'explainable expected goals' to produce accurate expected goals for the team and player performance analysis. The features used are game minute, home or away fixture, play type (situation e.g. Open Play, Set Play), shot type (limb), last action (e.g. cross), distance from goal and angle from goal. They use tree-based classification: XGBoost, Random Forest, Light GBM and CatBoost, and also propose utilising Aggregated Profiles (AP) to demonstrate the difference in model predictions depending on a change in the value of a feature.

Tureen and Olthof [13], with a sample of 580 Premier League and 326 Women's Super League fixtures, use Statsbomb data provided for the 2022 conference and aim to quantify individual players' nested data in a hierarchal structure to reduce biased interferences. By creating the Estimated Player Impact (EPI) measure using the Generalised Linear Mixed Models (GLMM), they estimate individual players' impact on each xG valuation and are able to quantify the shot impact on shot conversion. The EPI metric quantifies a player's impact on xG estimations. Similarly, to this paper, the EPI is measured on positional subgroups: Forward, Midfielder and Defenders (split into central and wing backs). Their findings claim that Heung Min Son of Tottenham Hotspurs who plays as a forward has the largest positive impact on EPI per xG. The study conducted by Maed et al. [14] aims to address existing issues in soccer analytics by applying various machine learning techniques to model expected goal values using previously untested features. The authors collect data on all shots from the European Top 5 leagues during the 2017–18 seasons. They utilise logistic regression, XGBoost, Random Forest, AdaBoost, and MLP algorithms in their analysis. The main purpose of the research is to compare the predictive ability of traditional statistics with the newly developed metric of expected goals. The results demonstrate that the error values of the expected goals models developed in this study were competitive with optimal values reported in other papers. Additionally, some of the newly introduced features are found to have a significant impact on the outputs of the expected goals models. Notably, the study reveals that expected goals outperformed traditional statistics as a predictor of a football team's future success. Moreover, the research outcomes even surpass those obtained from an industry leader in the same field, highlighting the effectiveness of the proposed machine learning techniques in soccer analytics.

In parallel with the developments of xG, academic and sport science researchers have tried to find some different ways to quantify player actions in the game, such as the expected threads (xT) by Karun Singh [15] and Valuing Actions by Estimating Probabilities (VAEP) by Decroos et al. [16]. Even though these works are not directly xG-related, considering their influence on studies to better quantify the actions on the football pitch necessitates acknowledgement whilst finalising this section. In one of the pioneering studies, Particularly, the Expected Threat (xT) is a metric used in football analysis to assess the likelihood of a goal being scored from different zones on the pitch. It involves dividing the pitch into zones and assigning a value to each zone based on goal-scoring probabilities. Developed by Karun Singh in 2018 [15], xT is a widely recognised possession value model in the industry. However, it has some limitations. Firstly, it only considers actions that involve moving the ball between zones, like passes and carries, while disregarding defensive actions and shots. Additionally, xT is often implemented using limited event data, neglecting important factors such as the player's pressure situation when making the action. These limitations should be considered when using xT for football analysis in academic papers.

Furthermore, Decroos et al. [17] introduce a novel approach to assess the impact of individual actions performed by soccer players during games. They noted that the existing metrics often focus on rare actions like shots and goals, or overlook the context in which actions occur. In order to address this, this paper proposes (1) a new language for describing player actions and (2) a framework to value any type of action based on its impact on the game outcome while considering the context. By aggregating action values, the total offensive and defensive contributions of soccer players to their teams can be quantified. The approach highlights the importance of considering contextual information, which traditional player evaluation metrics tend to ignore. The paper demonstrates the effectiveness of the proposed method through various use cases related to scouting and playing style characterisation in the 2016/2017 and 2017/2018 seasons in Europe's top competitions.

Decroos et al. [16] present a framework for valuing actions – VAEP – for evaluating player actions in any game by considering their impact on the game's outcome and the context in which they occur. This approach allows for a comprehensive assessment of a player's performance, including the quantification of their offensive and defensive contributions to their team. The paper includes practical examples to showcase the effectiveness and advantages of the proposed framework. Lastly, the paper by Van Roy et al. [18] critically compares xT and VAEP. Whilst many existing metrics focus on shots and assists, which represent a small portion of on-the-ball actions, xT and VAEP aim to address this limitation by assessing all types of actions that affect the likelihood of yielding a goal. The paper highlights the conceptual and practical differences between these two models, including their design choices, resulting in distinct top-player rankings and significant variations in how they value specific actions (see Table 1).

2. The proposed xG model details

Based on the academic and football industry background of xG discussed above, this paper aims to

1. to create a reliable and robust xG model on open play goals.
2. to evaluate models based on their abilities to predict goals close to actual goals. This means that the direction of the project is not to build the best model according to classifier performance. But compare classifier outputs with real-life events.
3. to develop a model the output of which is useful in showing the quality of players' finishing to a high degree.
4. to achieve a significant positive correlation with industry xG providers.
5. to successfully apply adjusted xG metrics to specific matches.

To reach the aforementioned objectives, in this section, we share the proposed xG model along with the developed data set. We then share our results under four experimental scenarios of (1) a general model comparison of the baseline and proposed model, (2) position-adjusted xG analysis, (3) player-corrected xG analysis and (4) an industry benchmark testing based on the 2018 Champions League Final game of Liverpool versus Real Madrid.

2.1. Data

The whole data set in this paper has been web-scraped by using *statsbombpy* Python module that Statsbomb provides. Please visit the Statsbomb GitHub page via https://github.com/statsbomb for details. The data frame is filtered to include just open-play shots, with penalties removed as they skew the model. Leaving a sample of 15,575 data points of event data. The shot breakdown is 2873 European Championship and World Cup samples, 12,688 La Liga samples with 2301 being Messi shots and then a Game Sample of 28 shots from the Champions League. A separate xG model can be run to calculate penalty xG, however, this is not a point of interest in this project. The industry standard of 0.766 xG per penalty is an acceptable value.

From the event data sample discussed, there were originally 95 features most of which are informative features such as ID, League, Year, Day, Time, and some action-related information such as pass type, foul won, block, etc. The features were first reduced by removing all variables that had over 99% NaN or Null values. Furthermore, from the remaining features, 8 (4 binary, 3 categorical and 1 2D coordinates) were selected to be used in the model along with all the open-play shots considering their relation to goal scoring. For the supervised learning of a binary classification problem, some steps are taken to prepare the data for learning. Shot information with missing values is dropped from the data set, numeric features are scaled, categorical features are label encoded, Boolean features are hot encoded and coordinates are separated into X and Y. In total, after pre-processing, there are 26 variables in this initial model with 1 target variable of Goals (1 = Goal, 0 = No Goal). All the utilised StatsBomb variables and their details are presented in Table 2 whilst an example of 10 rows of the data set is also shown in Table 3 with StatsBomb features.

Table 1
Summary of literature.

Paper	Data	Data source	Method name	Further details
Lucey et. al. (2015) [4]	9732 shots & 10 s of gameplay video before each shot from a professional league.	Prozone (now Stats Perfofm)	Expected Goal Value (EGV)	Conditional Random Models based probabilistic model.
Ruiz et. al. (2015) [5]	10,318 shots taken during the 2013–2014 English Premier League	Prozone (now Stats Perfofm)	Goal Expectancy Model	Multilayer perceptron (MLP) (6 hidden layers)
Eggels & Pechenizkiy (2016) [6]	From seven different leagues over three seasons — 128667 (14109 resulted in a goal)	ORTEC, FIFA, Inmotio	Expected Goals	Logistic Regression, Decision tree, Random Forrest, & Ada-boost.
Fairchild et. al. (2018) [7]	1115 non-penalty shots from 99 games in the Major League Soccer (MLS — USA Football League)	Manual via Burn-Murdoch	Expected Goal Model	Logistic Regression
K Singh (2018) [15]	a Blog post explaining the method. No specific dataset was used.	N/A	Expected Threat (xT)	Iteratively solved mathematical expression
Decroos et. al. (2019) [17]	the English, Spanish, German, Italian, French, Dutch, & Belgian top divisions — 11565 games between 2012/2013 & 2017/2018.	Wyscout	VAEP (Valuing Actions by Estimating Probabilities)	CatBoost in a binary classification setup.
Brechot & Flepp (2020) [9]	170,688 shots (7,304 matches) from the English Premier League, the French Ligue 1, the German Bundesliga, the Italian Serie A, & the Spanish La Liga from 2013–2014 to 2016–2017.	Gracenote	Expected Goals	Logistic Regression
PM Pardo (2020) [10]	87161 shots, of which 8083 are goals from major European leagues.	OPTA & FIFA	Expected Goals	Logistic Regression, XGBoost & Neural Networks
Umami et. al. (2021)	More than 600K shot from 5 major European Leagues	Wyscout	Expected Goals	Logistic Regression
Cavus & Biecek (2022) [12]	315,430 shots over seven seasons for European Top 5 leagues	Understat	Explainable Expected Goals	XGBoost, Random Forest, LightGBM, & CatBoost
Tureen & Olthoff (2022) [13]	580 Premier League & 326 Women's Super League fixtures	StatsBomb	Estimated Player Impact (EPI)	Generalised Linear Mixed Models (GLMM)
Mead et. al. (2023) [14]	All shots from European Top 5 leagues in 2017-18 seasons.	Wyscout	Expected Goals	Logistic regression, XGBoost, Random Forest, AdaBoost & MLP.

Table 2
StatsBomb data features and their details.

Variables	Type	Values
Aerial shot	*Binary*	True & False
1st time	*Binary*	True & False
Open Goal	*Binary*	True & False
Pressure	*Binary*	True & False
1v1	*Binary*	True & False
Shot Technique	*Categorical*	Backheel, Diving Header, Half Volley, Lob, Normal, Overhead Kick, Volley
Pass From	*Categorical*	Corner, Counterattack, Free Kick, Goal Kick, Keeper, Kick off, Throw In, Regular Play
Shot Body Part	*Categorical*	Header, Left Foot, Right Foot, Other
Location	*Coordinates*	(x, y) coordinates — 120 yard ×80 yards
Goal	*Binary*	True & False (Target)

Table 3
Example sample of the data set for Statsbomb features.

Player	Aerial shot	1st time	1v1	Open goal	Press.	Shot tech.	Pass from	Body part	Goal
Diego Garcia	1	0	0	0	1	Normal	Free Kick	Head	1
Antoine Griezmann	0	0	0	0	0	Normal	Regular Play	L. Foot	0
Alvaro Negredo	0	0	0	0	1	Normal	Corner	L. Foot	0
Lionel Messi	0	0	0	0	0	Normal	Free Kick	L. Foot	1
Xavier Hernandez	0	1	0	0	0	Normal	Regular Play	R. Foot	0
Gerard Pique	0	0	0	0	0	Normal	Corner	Head	0
Lionel Messi	0	1	0	0	0	Normal	Free Kick	L. Foot	1
Jordi Alba	0	1	0	0	0	Half Volley	Free Kick	L. Foot	0
Martin Odegaard	0	0	0	0	1	Normal	Regular Play	L. Foot	0
Lionel Messi	0	0	0	0	0	Normal	Free Kick	L. Foot	0

2.2. xG model

Based on these variables selected, an xG model is built to provide a baseline for development within the paper with the idea of surpassing the Baseline model by including additional variables and further model tuning.

A logistic regression approach was taken to establish a baseline xG model. The aim is to create a model that shows the minimum acceptable level and then improve the results from there. Logistic regression is a traditional approach as it is easy to run, interpret and train. With the sample of 15,575 open-play shots, the baseline model is tested versus the industry benchmark of StatsBomb xG. Then, Accumulated Goals are

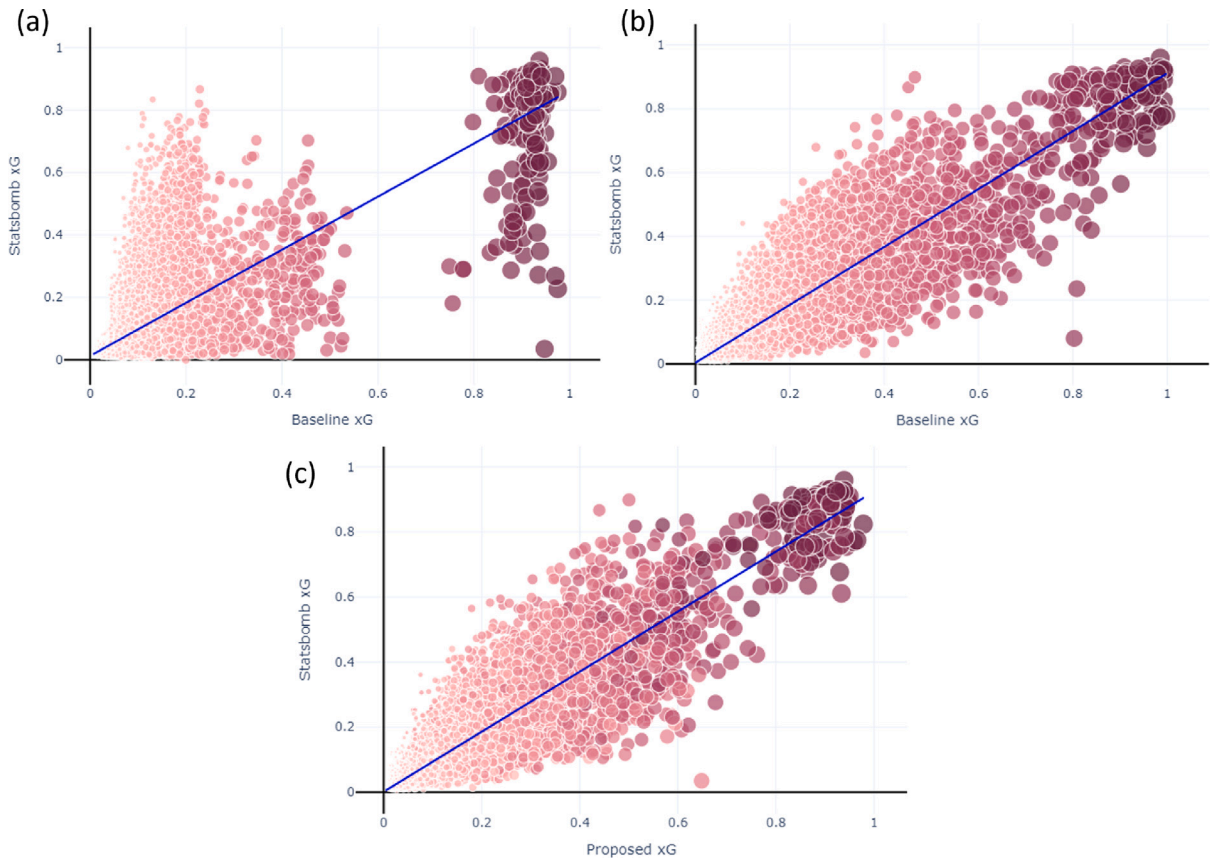

Fig. 2. (a) Baseline model results with existing 26 features. (b) Baseline model performance after adding extra features. (c) Proposed xG model performance.

used as a metric to compare to Accumulated xG with the aim that the correlation would be high, to judge whether the model created is close to the industry level of expected goals, and that accumulated Goals and xG are close.

Fig. 2(a) represents the output of the baseline logistic regression xG model compared to the xG of StatsBomb. The fitted line is $y = 0.01 + 0.85x$, and the correlation is 0.659. Graphically, there is evidence that the model is not working at maximum efficiency as there is a large area without any considerable xG values, with 15,575 shots it would be expected to see a more even distribution of shot xG values and not so skewed to the two tails. In conclusion, the distribution of xG values suggests that it is limited in predicting a certain level of expected goals. The graph suggests that the xG Model is under-predicting xG values, hence the large cluster on the left-hand side between 0.0 xG and 0.2 xG.

The lower performance of the existing baseline is not caused by the model capability but by the limited amount of information provided to the model with the existing 26 features. Some of those are *Location, Shot-technique, Pass-type* and *Direction*. It is clear from the analysis in Fig. 2(a) that existing features are not enough to provide a robust xG model.

In order to understand this result, we follow a feature analysis based on SHapley Additive exPlanations (SHAP) [19] which is a widespread technique in the field of explainable AI (xAI) that provides insights into the contributions of each feature or variable in a machine learning model's decision-making process. Based on cooperative game theory, SHAP assigns a value to each feature by quantifying its impact on the model's predictions. It calculates the average marginal contribution of each feature across all possible feature combinations, attributing credit to individual features fairly and consistently. By utilising SHAP, xAI practitioners can gain a deeper understanding of how specific features influence the model's output, enabling them to interpret and explain the model's decisions with greater transparency and accuracy.

To allow feature analysis, SHAP values of the baseline model are calculated and a beeswarm plot is plotted in Fig. 3. Interestingly, Open Goal remains a highly influential factor along with Right Footed effort and Left Footed efforts, which is expected as they are ever present in the vast majority of strikes, with the only alternative being other. Normal shot technique also proved to influence, suggesting that alternatives such as Volleys are more difficult to score. On the other hand, the SHAP values are quite small even though they contribute to the prediction stages. This concludes the need for more informative features for xG calculation.

Parallel with the suggestions from the experts such as [3], we developed various new features to better inform the proposed xG model. Some examples of additional features utilised are Goalkeeper positioning, Player Pressure Radiuses and Opposition Between the shot and the goal. Fig. 2(b) presents the same baseline modelling performance but this time with the new features along with the existing ones (40 features in total). Please see Table 4 for the proposed engineered variables for xG calculation. Following the creation and implementation of further variables, the logistic regression results are relatively better compared to the previous experiment case. With the benchmark being StatsBomb xG, the line outcome is $y = 0.00 + 0.90x$ with a correlation of 0.887. Visually, the distribution is also pleasing and the output is acceptable. The predicted total xG is at 1866 compared to the actual goals of 1887.

As is mentioned in the above sections, the purpose of this paper is not only to create novel and informative features, but also to extend and improve the baseline approach with a decision trees-based machine learning method. Decision trees are a non-parametric supervised learning method used for classification and regression. Decision tree algorithms are prominent in current Kaggle competitions and various regression-based prediction studies. To create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features, the supervised learning gradient boosting family models are promoted.

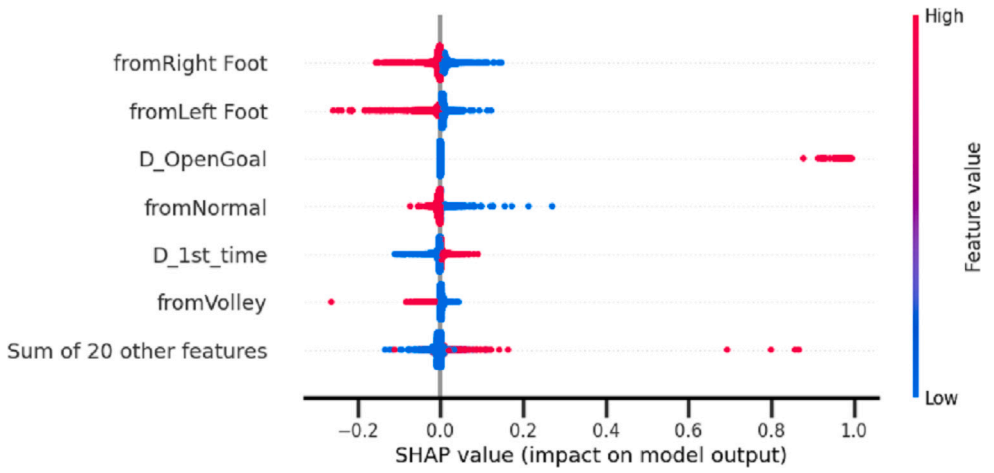

Fig. 3. SHAP Value analysis for feature importance from baseline logistic regression model with StatsBomb predictors only.

Table 4
Engineered Features and their details..

Variables	Description	Value
Strong Footed	*defines the player strong body part for shot*	(Categorical) Left, Right, Head
Within Penalty Area	*indicated whether the shot is in the penalty area*	(Binary) True & False
Distance from Goal	*distance from the shot location to the centre of the goal*	(Scaled Distance) Range (0, 1)
Angle to the centre of goal	*angle between shot location and two posts of the goal (see Fig. 1)*	(Scaled Angle) Range (0, 1)
GK distance from centre of goal	*distance between the goalkeeper and the centre of the goal*	(Scaled Distance) Range (0, 1)
Opp. players between shot and goal	*an integer number corresponds to the number of players within the shot triangle*	Continuous
Num. of opponents within 5-yard radius	*considering the player is the centre of a 5-yard radius circle, this feature refers to an integer number of players within this circle*	Continuous
Player positional sub-group	*some specific positional information of the player*	(Categorical) Attacker, GK, Central Defender, Winger Defender, Defensive Midfielder, Central Midfielder, Wide Midfielders

Table 5
Model comparison based on the number of goals predictions.

	Baseline	Proposed	Statsbomb	Exact Goals
Goals	1866	1870	1751	1887

Gradient boosting aims to combine weaker models (performing only slightly better than random choice), building upon Leslie Valiant's probably approximately correct (PAC) learning [20], which investigated the complexity of machine learning problems. Gradient tree boosting is the statistical framework, where the aim is to minimise the total loss of the models. This is achieved by combining weaker learning models using gradient descent. It has three main areas of focus (1) Loss function, (2) Weak learner and (3) Additive model.

Fig. 2(c) depicts the regression analysis plot for *the decision-trees based proposed xG model*. Examining Fig. 2(c), we can conclude that the proposed xG model is slightly better than the logistic regression. The visualisation shows exceptionally promising results with the fitted line of $y = 0.00 + 0.92x$ and a correlation of 0.902 with the StatsBomb xG. Predicting an accumulated xG of 1870 out of the 1887 actual goals as shown in Table 5.

Investigating the effect of engineered features and the proposed decision tree-based machine learning technique, we present a SHAP analysis this time for the proposed parameter set in the proposed xG model. Fig. 4 presents a beeswarm plot for SHAP values of important 9 predictors in the model. Examining Fig. 4, we can conclude that there is significant impact on model outcomes coming from proposed engineered predictors of *Angle to Goal, Distance from centre of Goal* and *Goalkeeper distance from centre of Goal*. The improvement in the xG calculations and statistical distance to StatsBomb xG values can be explained by the SHAP values presented here.

3. Results

The proposed xG model was tested on three different perspectives :

1. We first presented position-adjusted xG analysis.
2. Subsequently, a similar analysis to (1) but this time player-adjusted xG analysis was implemented using Lionel Messi as the target player.
3. Finally, we performed an industry benchmark on a specific fixture of R. Madrid vs. Liverpool in the 2018 UEFA Champions League Final.

3.1. Position adjusted xG

For this experimental step, we start by analysing the data set by separating the players in terms of their stronger positions, such as Goalkeeper, Defender, Midfielder, and Forward. The positional analysis results are presented in Table 6.

The results show some interesting outcomes. Forwards are known as the most clinical finishers in football, hence why they are entrusted with the position closest to the opposition goal with the most chances to score. Their conversion rate supports this, at 6.77 shots per goal. As expected goals represent the average probability of the shot turning to a goal, you should predict that the forwards are finishing at a rate higher than the average. Forwards scored 1276 Goals in this data, with all reference xG predictions being lower than the actual goals, suggesting an over-performance.

Midfielders show something quite different compared to Forwards. In all the models, they are underperforming by 1 and up to 13 below their xGs. This suggests that midfielders are wasteful with their chances. From the optimal selected model, midfielders are performing

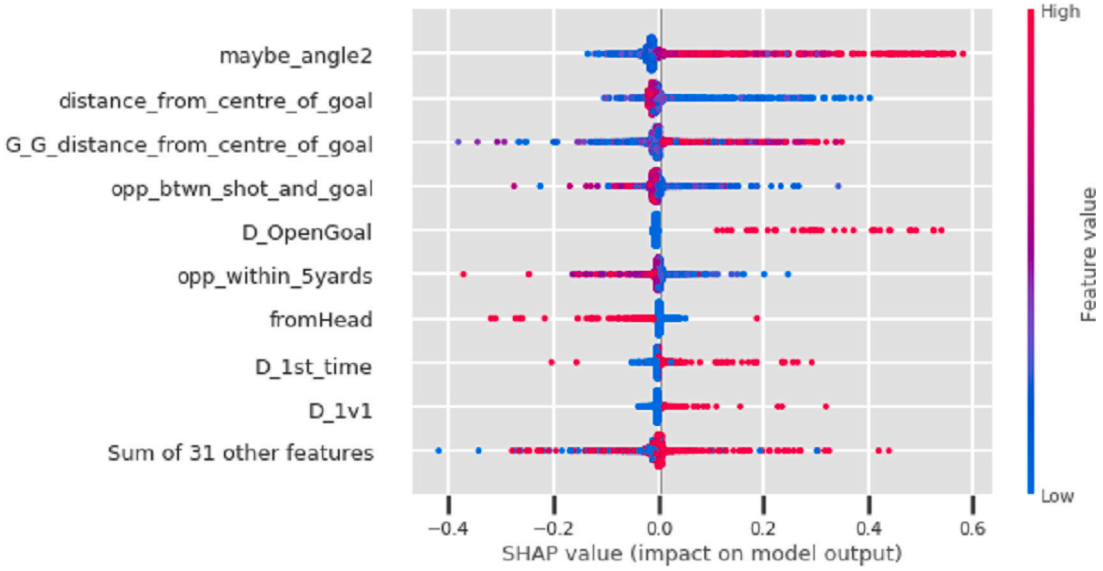

Fig. 4. SHAP Value analysis for feature importance from the proposed decision tree model with StatsBomb and Engineered predictors.

Table 6
Positional analysis of the data set.

Position	Total shots	Total goals	Shot/Goal	Baseline xG	Statsbomb xG	Proposed xG
Forward	8646	1276	6.77	1265.767	1154.624	1252.802
Midfield	4590	398	11.53	399.3236	391.2747	411.0013
Defender	2336	213	10.97	200.9173	205.6996	206.8463
Goalkeeper	2	0	0	0.096735	0.107875	0.105983

at 411 xG and scoring 398 goals. Defenders' results are surprising even based on the average xG including some high-value attempts by talented forwards. All xG predictions for defenders are below the exact number of goals, which suggests a high level of clinical nature and an overperformance of xG.

Fig. 5 depicts the density of shots for each position except GKs. From Fig. 5(a), as expected, the density of shots by forwards is predominantly within the box which suggests higher xG opportunities and therefore potentially more goals. Fig. 5(b) shows a very different story compared to the forwards with a lot of low density being in and around the six-yard box, which is where forwards showed the highest density. The shots from outside the box are at a high density and suggest that shooting from the range is more prominent with midfielders compared to any other subgroup. Long shots do not have high xG chances and may suggest why their conversion rates from xG to actual goals are lower than Forwards. From Fig. 5(c), the density of the shots is as expected that the defenders' main chances come from set pieces and corners with the highest level of density falling between the penalty spot and the six-yard box. Considering this, the higher conversion rate and goals surpassing xG are very interesting to see in comparison to Midfielders, who football fans would consider more skilled in converting goals.

As discussed in this section up to now, the results gathered suggest a deviation in xG performance between positional subgroups. To more accurately quantify the performance of each, the same classification problem is carried out. However, in this instance, the training dataset is changed on three different occasions. The model is now trained on each positional sub-group, creating Forward Adjusted xG, Midfield Adjusted xG and Defender Adjusted xG. The aim is to provide empirical evidence that each position has different levels of efficiency over a whole scale of data.

Table 7 shows, as can be easily expected by an average football fan, Forward Adjusted xG predicts the highest value, followed by Midfield and then Defenders. Forwards xG valuation increases by an absolute adjustment value of 86, this large increase suggests that if all the chances are changed to being taken by a forward-skilled player

Table 7
Positional Adjusted xG Values.

Model	xG	Forward xG	Midfield xG	Defender xG	Goals
Goals/xG	1870	1956	1728	1397	1887
Adjustment Value	0	+86	−142	−473	0

then xG increases by 86. For Midfielders, when their sample skillset is applied across the whole dataset they find a reduction from the average xG of 142. This result is an expected outcome, as most clinical players (presumed to be forwards) are removed from their training set. Defenders, follow a similar yet more drastic reduction in xG even in comparison to Midfielders. Their xG reduces by 473 and therefore shows that the model predicts a lower clinical ability and conversion rates to other positional groups. This is what was expected by the testing, and proves that the model is working effectively.

3.2. Player adjusted xG values

Along with the aforementioned positional adjusted xG, this paper also aimed to test player-specific xG. The motivation of player-adjusted xG is to assess how much better elite-level forwards are at finishing in comparison to normal players and average xG.

The utilised data set in this paper shows that most elite finishers produce Goals greater than their xG value. Lionel Messi's sample of 415 Goals from 342 xG (%21.34 higher) is a large overperformance whilst Luis Suarez has 125 goals with 107 xG (%16.82 higher). Messi's overperformance is to a degree larger than any other player in the sample at 0.82 xG per Goal (Please see Fig. 6 for more details).

Considering the data set in this paper includes a large sample of Lionel Messi fixtures in La Liga and therefore a lot of Messi shots and therefore goals, we decided to use Lionel Messi as the player adjustment

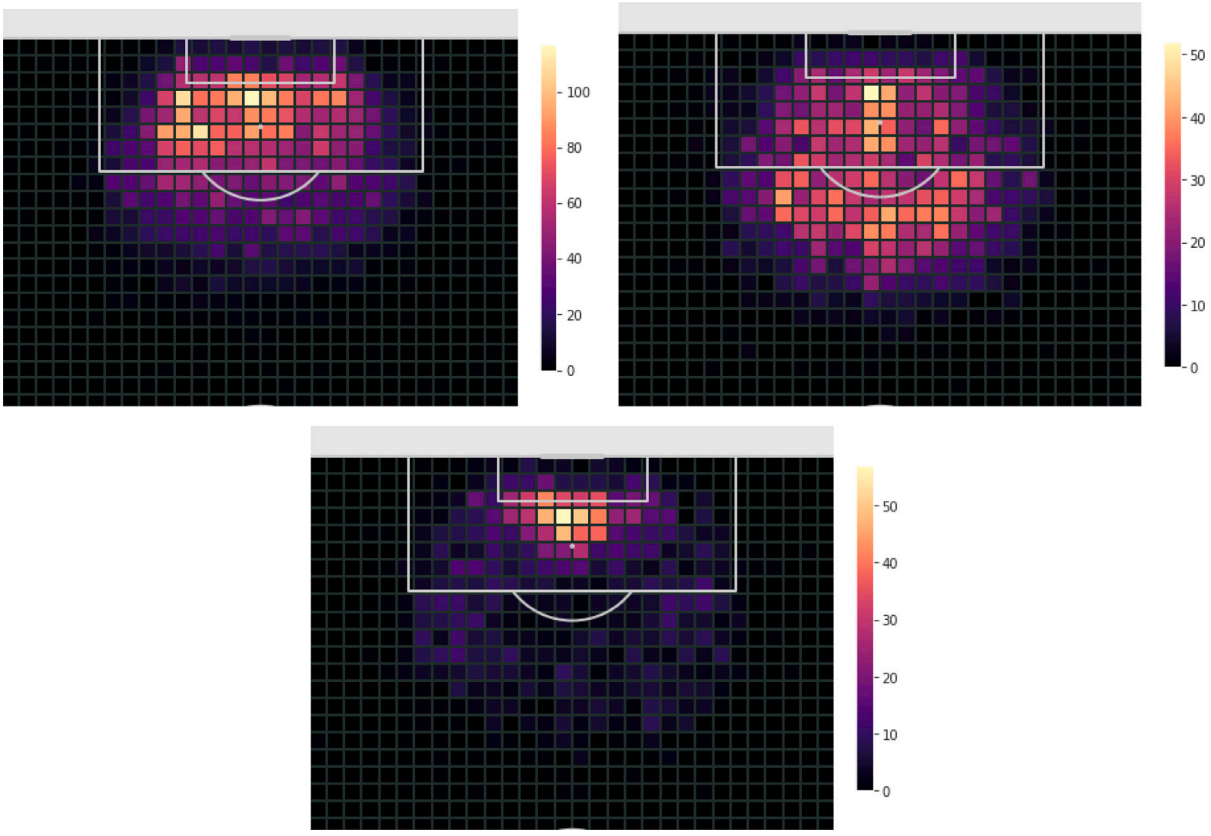

Fig. 5. Shot Density of (a) Forwards, (b) Midfielders, (c) Defenders.

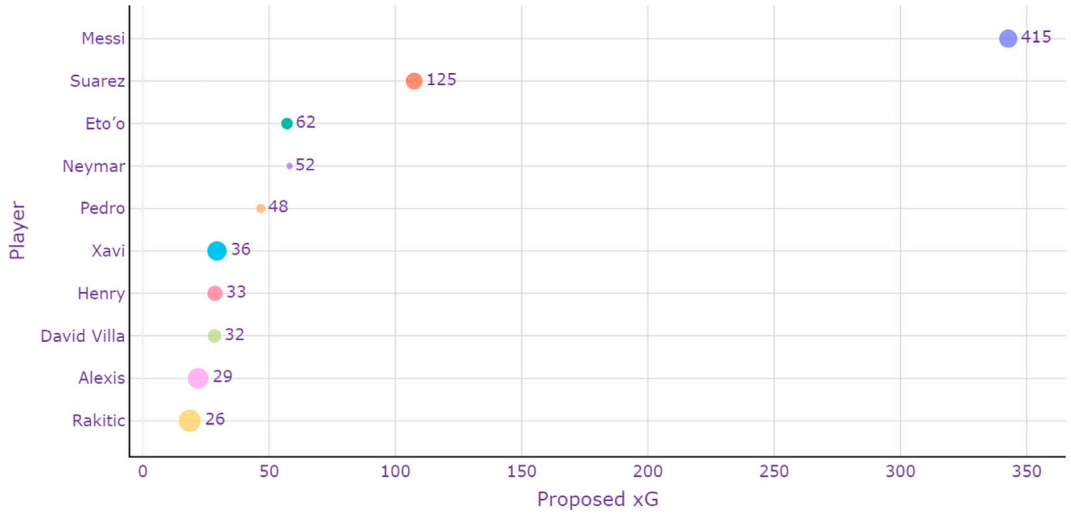

Fig. 6. Player specific xG and Goals analysis. The size of the markers shows Goals per xG metric whilst the text next to each marker is the exact number of goals scored by the player.

test case. Thus, we use Messi's 2300 shots and 415 goals to train all xG models. Another reason why Messi is selected as he is widely known as one of the all-time greats and has significantly overperformed xG in all xG models. He is a great sample to compare the "average" player against and allows us to show how significant his skill set is.

The process was similar to the positionally adjusted xG in that a subset of the data is taken with it being made up of only Messi shots, this also leaves the other subset as a sample including No-Messi shots. From these two subsets, Messi data is used to train an individual xG model, and No-Messi data is tested to evaluate the player-adjusted xG values.

The initial results of this analysis show significantly inflated levels of xG based on the Messi Adjusted xG. The results presented in Table 8 show that the Messi-trained model produces significantly more accumulated xG, at 1874 xG predicted which is an adjustment of 347 xG over the whole sample. This is as expected and confirms the hypothesis that Messi is an elite shooter.

Table 9 offers further insight into Messi's influence on the models. Specific player-adjusted values show a dramatic improvement in all players' xG. This also means e.g. if Messi shot the shoots Luis Suarez had, he could accumulate 126.21 xG with a nearly 20 xG increase compared to the proposed model's initial result of a total of 107.45

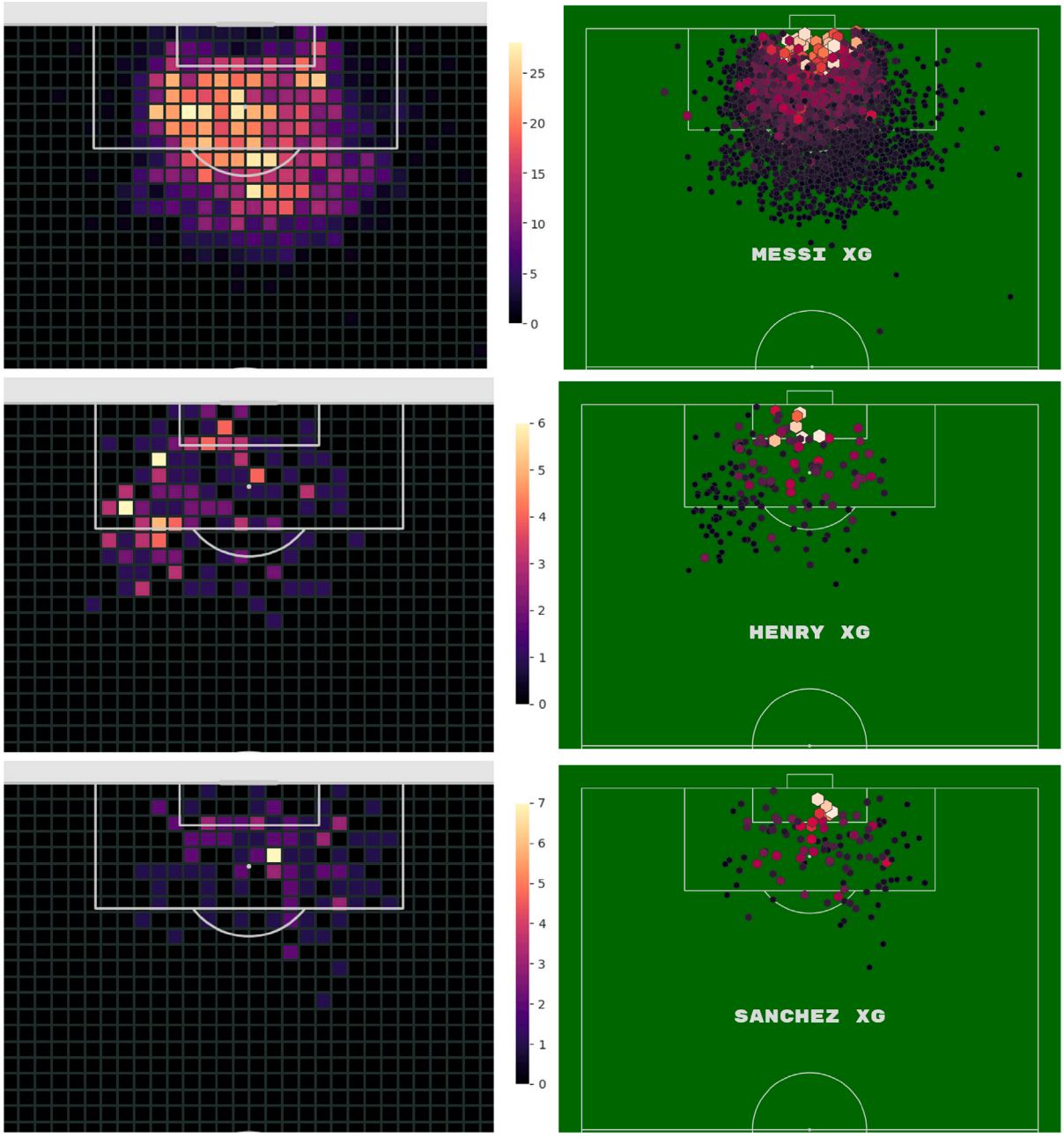

Fig. 7. (LEFT) Shot location heatmap and (RIGHT) xG location map visualisations for Lionel Messi, Thierry Henry and Alexis Sanchez from top to bottom, respectively.

Table 8
Messi Adjusted xG Values.

Model	Goals	Proposed xG	Messi adjusted xG
Score	1472	1527	1874
Adjustment Value	0	0	+347

xGs. Significantly, based on the original model, every player in the sample, except Neymar, over-performed their xG suggesting that they are clinical and finish chances greater than the average. With Messi adjusted xG, now every player falls short of their xG values, suggesting that they are underperforming.

Fig. 7 depicts shot location heatmaps and xG location maps for three important finishers namely Lionel Messi, Thierry Henry and Alexis Sanchez. In addition, Henry and Sanchez are the only two strikers who over-performed in Messi-adjusted xG values. All subfigures show how each player has some different shooting patterns whilst their xG maps are logical and aligned with the expected outcomes: The closer to the goal, The higher the xG. The left-hand side shot heatmap subfigures highlight how clinical Lionel Messi is with his capability to score over and outside of the box whilst Henry's trademark finesse shot to the far post can be seen with a high probability of shots from the left. It looks like Alexis Sanchez mostly tries shots from inside the box which gives him an over-performance on Messi-adjusted xG values.

This trial emphasises the value one of the all-time greatest – Lionel Messi – adds to the accumulated probabilities and confirms the expectation that he is an Elite Level finisher and far above the average player in the sample, despite other high-quality finishers being in the sample such as Thierry Henry.

Table 9
Player-specific results for Messi Adjusted xG.

Player	Goals	Proposed xG	Statsbomb xG	Messi adjusted xG	Over-performance?
Suarez	125	107.4513	108.2466	126.2198	No
Eto'o	62	57.1248	52.71371	63.02088	No
Neymar	52	58.12527	56.02049	66.74318	No
Pedro	48	46.75322	43.47725	53.60634	No
Xavi	36	29.35869	28.75191	41.78057	No
Henry	33	28.5776	28.18925	32.03867	**Yes**
David Villa	32	28.3766	29.0367	32.61256	No
Alexis	29	21.9309	21.33671	25.03286	**Yes**
Rakitic	26	18.57919	18.45674	26.59705	No

Table 10
Game statistics for the 2018 Real Madrid vs Liverpool, Champions League final.

Team	Goals	Shots	Statsbomb	FBRef	infogol	Proposed
Liverpool	1	14	1.31442	**1.9**	**1.88**	1.61353
R. Madrid	**3**	14	**1.367858**	1.5	1.71	**1.816377**

Table 11
2018 Real Madrid vs Liverpool, Champions League final game goals.

Team	Player	Shot technique	Minute	Proposed xG	Statsbomb xG
R. Madrid	Karim Benzema	Volley	50	0.351569	0.517137
Liverpool	Sadio Mane	Volley	54	0.599426	0.548516
R. Madrid	Gareth Bale	Overhead Kick	63	0.131235	0.022605
R. Madrid	Gareth Bale	Normal	82	0.020848	0.013965

3.3. Industry benchmark testing and application

This experimental analysis details the proposed xG model being applied to the Real Madrid vs Liverpool, Champions League final from the 2017/18 season to prove the ability to apply the xG model to a specific fixture.

This game was finalised within 90 min with shots drawn at 14 apiece, R. Madrid beat Liverpool 3 goals to 1. In Table 10, there are xG figures from StatsBomb (use their data), FB Ref (use Opta data) and infogol.com along with the proposed xG model. Comparing and contrasting these figures allows the validity of the proposed model to be shown. Accumulated xG from StatsBomb is at 2.7, FBRef falls at 3.4, infogol is at 3.59 and the proposed model is at 3.4. This also shows the different factors within the models and feature tuning between companies. The results from the proposed model are still satisfactory as they are within the same region as all the companies, despite resource constraint differences.

Fig. 8(a) shows an industry-standard recreation of the play-by-play xG events, known as an xG timeline, something commonly used to show a better understanding of which periods of the game were more productive for each side. Fig. 8(b) visualises each shot within this fixture using industry-standard recreation a shot map with the larger icon indicates a higher xG.

Table 11 details each goal within the game showing the xG valuation differences between the proposed and the StatsBomb xG models. The 2nd goal of the game by Sadio Mane and the 4th Goal of the game by Gareth Bale show similar valuations. However, Karim Benzema's goal from the Karius mistake is slightly lower from the proposed model which is surprising as the accumulated xG showed that the proposed xG model quantified higher than StatsBomb. That trend however is shown in the Gareth Bale Overhead Kick, the proposed xG shows a 0.131 value whereas StatsBomb is at 0.022. In this scenario, the result is concerning as it likely suggests an over-prediction of the xG value of overhead kicks, which are known to be a very tough skill.

4. Conclusions

In conclusion, the project achieves various important insights in terms of position and player adjusting of xGs with a machine learning model created from scratch along with various important new features. The results from Position Adjusted xG and Messi Adjusted xG align completely with the expectations and therefore are inferred to have predicted xG accurately.

The main contributions to current literature are development in xG modelling and more features than are used in conventional models. Some of the additional features which are utilised in this paper are Goalkeeper positioning, Player Pressure Radiuses and Opposition Between the shot and the goal. These features directly develop what many writers including [3] have specifically called for, to develop xG. The successful implementation and application of these allow further study such as creating a variable that can indicate the surface area that a player has available to shoot at which may require products such as StatsBomb 360 data.

The key findings are that Forwards are the most clinical and most effective shooters from the positional subgroups, regardless of adjustment or not. Further developments find that the shooting quality order is Forwards, Midfielders, and Defenders in terms of their ability to successfully gain xG. The Player Adjusted xG finds that Messi performs better than even the Forward subset. This level of analysis on player and positional adjusted xG cannot be found in current literature, and we believe it successfully fills a large gap.

Further research could include adjusting xG for Leagues and European competitions using the same subset process that this project follows. Questions can be asked such as: 'How much better are Premier League players in comparison to the top 5 leagues?' and 'How much more efficient are players performing in the Champions League subgroups to their Domestic Leagues?'.

As further the opening paragraph, football is situation based and has room to adjust metrics to increase the validity of statistics within the professional game and improve performance analysis.

CRediT authorship contribution statement

James H. Hewitt: Wrote the main manuscript text, Created data visualization outputs, Analysed the results, Reviewed the manuscript, Collected the data, Created engineered variables and conducted experiment(s). **Oktay Karakuş:** Wrote the main manuscript text, Created data visualization outputs, Analysed the results, Reviewed the manuscript, Supervised the research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

All authors approved the final version of manuscript to be published.

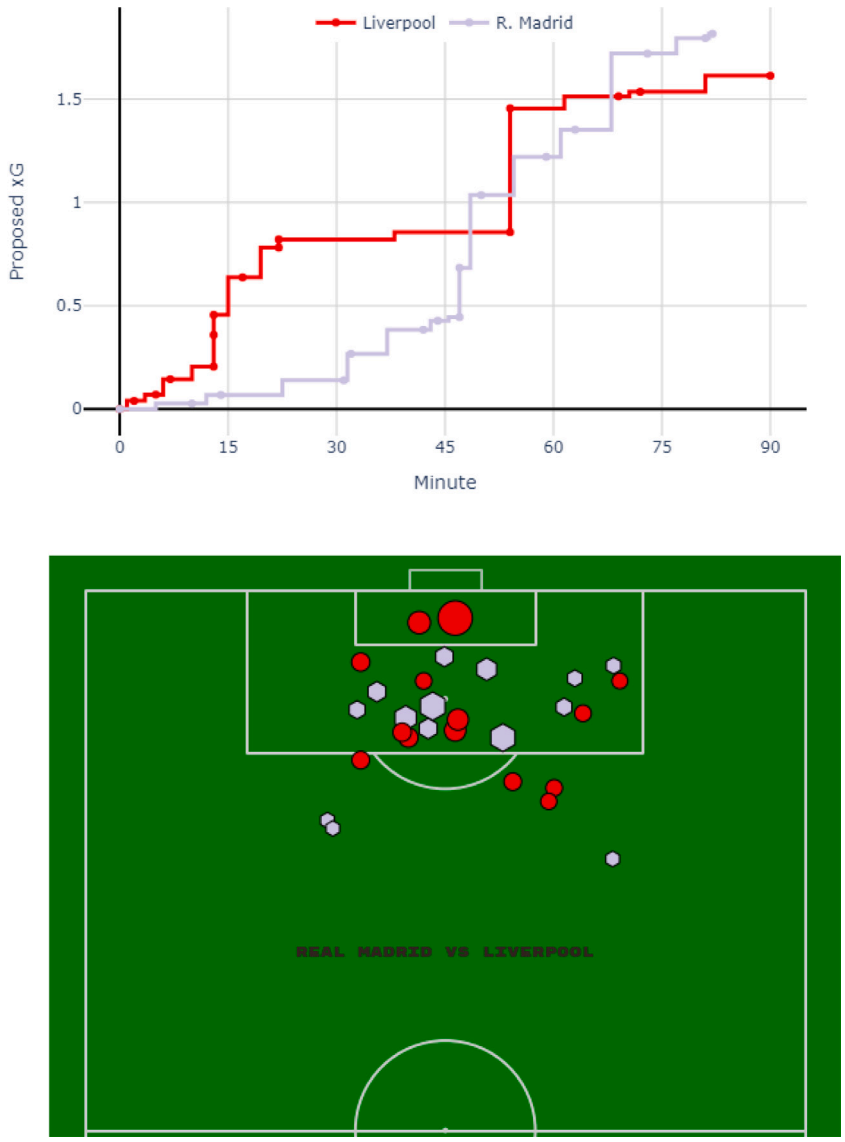

Fig. 8. 2018 Real Madrid vs Liverpool, Champions League final (TOP) xG timeline, (BOTTOM) xG location map.

References

[1] S. Winterburn, Wenger: Man City Only Beat Us by 0.1 Expected Goals, 2017, Football365, URL https://shorturl.at/lopY7.

[2] A. Bate, R. Campbell, Expected goals explained: The analysis that is changing the game, 2017, Sky Sports, URL https://shorturl.at/mvF16.

[3] J. Tippett, The Expected Goals Philosophy: A Game-Changing Way of Analysing Football, 2019.

[4] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, I. Matthews, Quality Vs Quantity: Improved Shot Prediction in Soccer Using Strategic Features from Spatiotemporal Data, MIT, 2015.

[5] H. Ruiz, P. Lisboa, P. Neilson, W. Gregson, Measuring scoring efficiency through goal expectancy estimation, in: ESANN, 2015.

[6] H. Eggels, R. Van Elk, M. Pechenizkiy, Explaining soccer match outcomes with goal scoring opportunities predictive analytics, in: Mlsa@ Pkdd/Ecml, 2016.

[7] A. Fairchild, K. Pelechrinis, M. Kokkodis, Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality, J. Sports Anal. 4 (3) (2018) 165–174.

[8] M. Herold, F. Goes, S. Nopp, P. Bauer, C. Thompson, T. Meyer, Machine learning in men's professional football: Current applications and future directions for improving attacking play, Int. J. Sports Sci. Coach. 14 (6) (2019) 798–817.

[9] M. Brechot, R. Flepp, Dealing with randomness in match outcomes: how to rethink performance evaluation in European club football using expected goals, J. Sports Econ. 21 (4) (2020) 335–362.

[10] P. Madrero Pardo, Creating a Model for Expected Goals in Football Using Qualitative Player Information (Master's thesis), Universitat Politècnica de Catalunya, 2020.

[11] I. Umami, D.H. Gautama, H.R. Hatta, Implementing the expected goal (xG) model to predict scores in soccer matches, Int. J. Inform. Inf. Syst. 4 (1) (2021) 38–54.

[12] M. Cavus, P. Biecek, Explainable expected goal models for performance analysis in football analytics, in: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics, DSAA, IEEE, 2022, pp. 1–9.

[13] T. Tureen, S. Olthof, "Estimated Player Impact"(EPI): Quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models, in: StatsBomb Conference Proceedings, StatsBomb, 2022.

[14] J. Mead, A. O'Hare, P. McMenemy, Expected goals in football: Improving model performance and demonstrating value, PLoS One 18 (4) (2023) 1–29, http://dx.doi.org/10.1371/journal.pone.0282295.

[15] K. Singh, Introducing expected threat (xt), 2018, Personal Blog, URL https://karun.in/blog/expected-threat.html. (Accessed July 2023).

[16] T. Decroos, L. Bransen, J. Van Haaren, J. Davis, VAEP: an objective approach to valuing on-the-ball actions in soccer, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 4696–4700.

[17] T. Decroos, L. Bransen, J. Van Haaren, J. Davis, Actions speak louder than goals: Valuing player actions in soccer, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1851–1861.

[18] M. Van Roy, P. Robberechts, T. Decroos, J. Davis, Valuing on-the-ball actions in soccer: a critical comparison of XT and VAEP, in: Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports. AI in Team Sports Organising Committee, 2020.

[19] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[20] L.G. Valiant, A theory of the learnable, Commun. ACM 27 (11) (1984) 1134–1142.