**Advancing Expected Goals (xG) Models in Football: A Hybrid Feature Engineering Approach Leveraging Spatiotemporal StatsBomb Data**

**Guechtouli Anis, Benhamdi Anis, Limani Mohamed Amine and Bouhouita Hamza**

University of Algiers 1

Algiers, Algeria

{a.guechtouli, a.benhamdi, am.limani, h.bouhouita}@univ-alger.dz

**Abstract**

The use of data in sports, particularly football, is not a new concept. Since the early 1960s, clubs have leveraged analytics for recruitment and injury prevention, though these insights rarely reached public audiences. In recent years, however, data integration has surged, transforming fan engagement and tactical discourse. Among these advancements, Expected Goals (xG)—a metric quantifying shot quality based on contextual variables—has emerged as a cornerstone of modern analysis since its popularization by OPTA in 2012. In this study, we develop an xG model using machine learning and spatiotemporal feature engineering, trained on a decade of StatsBomb’s open men’s competition data. By synthesizing traditional variables (distance, angle) with advanced metrics (defender positioning, pressure), our model achieves a robust ROC-AUC of 78%–82%, outperforming baseline benchmarks given the dataset’s scale and complexity. These results highlight the practical utility of incorporating detailed defensive metrics, such as defender positioning and pressure, into xG frameworks, while offering fans and analysts an accessible, interpretable tool to evaluate shot outcomes in professional football.

**1 Introduction**

Football’s relationship with data has evolved significantly over the decades, transitioning from rudimentary manual observations to sophisticated, algorithm-driven insights powered by technological advancements. The journey began in the 1950s with Charles Reep, a British analyst who manually collected match data to develop his controversial “long ball” strategy. While his conclusions were flawed, Reep’s work underscored the potential of data in football. Decades later, the publication of *Moneyball: The Art of Winning an Unfair Game* by Michael Lewis in 2003 inspired football clubs to embrace data-driven approaches, challenging traditional scouting and tactical methods. Clubs like Liverpool F.C., Brentford F.C., FC Midtjylland, and AZ Alkmaar have since demonstrated the competitive edge that data analytics can provide when applied effectively.

This study focuses on constructing an Expected Goals (xG) model using the Hudl StatsBomb open-access dataset, a rich resource containing detailed event-level match data. Unlike traditional xG models, which rely on basic variables like shot distance and angle, our approach incorporates advanced and engineered features such as the shooter’s preferred foot, defensive pressure, and dynamic player positioning. By leveraging these nuanced factors, we aim to bridge the gap between professional analytics and public research, demonstrating how accessible tools and open data can yield insights comparable to those driving decisions at elite clubs.

Expected Goals (xG) quantifies the probability of a shot resulting in a goal based on historical data of similar shots, considering factors such as shot location, angle, and contextual variables. Ranging from 0 to 1, an xG value of 0 indicates no chance of scoring, while 1 represents a certainty. Popularized by OPTA, xG has become a cornerstone of modern football analysis, used in post-match evaluations and even fan discussions. While xG cannot predict match outcomes or player performance, it is a powerful tool for assessing chance quality, creative play, and player efficiency (e.g., identifying players who consistently outperform their xG).

In Section 2, we explore the StatsBomb dataset, review related research, and outline the methodology for selecting our study range. In Section 3 We then identify and justify critical features—such as shot distance, angle, and defensive pressure—through exploratory visualizations, followed by preprocessing steps to engineer variables like “favorite foot,” dynamic player density, and geometric shot angles.

Section 4 details the training of machine learning models on the preprocessed dataset, incorporating cross-validation and hyperparameter tuning to enhance robustness. In Section 5, we evaluate the models using a dual approach: as a regression model with RMSE and as a classification model with accuracy and other metrics.

Finally, in section 6 we discuss the results, including model performance and computational efficiency, before concluding with the broader implications of our work. By democratizing advanced analytics through open-data models, this study aims to inform tactical, scouting, and developmental decisions in football, making high-level insights accessible to a wider audience.

**Resources and Tools**

This project utilized a variety of tools and resources for data analysis and processing. The computational environment was set up using **WSL 2 Ubuntu** for seamless integration with Linux-based tools. Key software and libraries included:

* **Python 3.12.3**: The primary programming language for analysis and processing.
* **Hadoop 3.3.6**, **Spark 3.5.3**, and **PySpark 3.5.4**: For handling and processing large-scale data efficiently.
* **statsbombpy 1.14.0** and **mplsoccer 1.4.0**: Libraries specifically used for football analytics and visualizations.
* **Kaggle**: A resource for accessing datasets and hosting analysis.

The computations were performed on

* HP Victus 16 Intel Core i5-11400H 2.70 GHz, 16 GB RAM, NVIDIA GeForce RTX 3060 6 GB

**2 Dataset**

**Source**

The data used for this study was sourced from the publicly available **StatsBomb Open Data** repository, accessible via the following link: [Hudl StatsBomb Open Data](https://github.com/statsbomb/open-data).

The repository includes JSON files containing the dataset, which is regularly updated whenever StatsBomb announces new free data. Additionally, a documentation folder is provided, offering detailed explanations of the dataset and its features.

The dataset covers the last 10 years (2014–2024) and includes data exclusively from men's competitions. This decision aligns with a prior study demonstrating that mixing data from men’s and women’s competitions for expected goals (xG) modeling is not fair to either group due to inherent differences in playstyle and context.

To extract and utilize this data, Hudl StatsBomb offers an API that supports Python and R, enabling efficient access to the free data. The data scraping process, which took approximately **2 hours and 51 minutes**, is detailed in the Jupyter Notebook, *statsbomb\_data.ipynb*, demonstrating the use of the StatsBomb API for extracting and preparing the dataset.

**Role of the Dataset in Research**

Since 2019, Hudl StatsBomb has been hosting an annual conference at the beginning of the football season (September to October). The conference invites speakers and participants who have contributed to research over the past year. As part of the event, a research competition is held, ensuring at least nine studies are conducted using the StatsBomb dataset each year. Additionally, the dataset is actively utilized by StatsBomb’s partners, including professional football clubs such as Clube Atlético Mineiro, Club Brugge KV, and AZ Alkmaar.

Research papers from the StatsBomb conferences can be accessed through the following links:

* [StatsBomb Conference 2019](https://statsbomb.com/news/statsbomb-conference-research-papers-1/)
* [StatsBomb Conference 2021](https://statsbomb.com/news/statsbomb-conference-2021-research-papers/)
* [StatsBomb Conference 2022](https://statsbomb.com/articles/soccer/statsbomb-conference-2022-research-papers/)
* [StatsBomb Conference 2023](https://statsbomb.com/news/statsbomb-conference-2023-research-papers/)
* [StatsBomb Conference 2024](https://statsbomb.com/news/statsbomb-conference-2024-research-papers/)

**Dataset Structure**

**Competitions** table provides information about the competition, including details such as the competition's ID, season, country or continent, and whether the competition is international. It contains 6 features and 26 unique values.

**Matches** table contains data on individual matches, including match dates, the teams involved, scores, match week, competition stage, stadium, referee, and the managers of both teams. It includes 14 features and 2,450 records.

|  |  |
| --- | --- |
| Competition | Matches |
| La Liga | 590 |
| Ligue 1 | 435 |
| Serie A | 380 |
| Premier League | 380 |
| Bundesliga | 340 |
| FIFA World Cup | 128 |
| UEFA EURO | 102 |
| Africa CAN | 52 |
| Copa America | 32 |
| MLS | 6 |
| Champions League | 5 |

Table 01: Number of matches per competition

**Lineups** table provides information about the players in each match, including details such as the player’s identity, team, jersey number, the cards received, and the positions they played during the match, along with the transitions in tactics and the minutes in which these occurred. It contains 7 features and 95,630 records.

**Events** table is the most comprehensive and critical table in the dataset, containing detailed event-by-event data for each match. It includes 122 features and a total of 8,710,714 records. The table captures various match events such as passes, shots, dribbles, tackles, fouls, ball recoveries, goalkeeper actions, and substitutions. Most of the features are stored in a dummy-like structure, where values like *True* or *Null* indicate whether a specific event detail is relevant (e.g., the 50\_50 feature is True if the event involves a 50-50 duel). A ";" separator was used due to the presence of dictionary-like fields that PySpark couldn't read directly.

|  |  |  |  |
| --- | --- | --- | --- |
| index | match\_id | play\_pattern | type |
| 37 | 3943077 | From Kick Off | Shot |
| 182 | 3943077 | From Keeper | Shot |
| 241 | 3943077 | From Throw In | Shot |
| 474 | 3943077 | From Corner | Shot |

Table 02: Example of events data

**Frames** table provides a view of the stadium field during each event, capturing the positions of all players involved in that specific moment. Each row corresponds to one player and their position during an event, enabling detailed spatial analysis. The table includes 7 features and a total of 10,259,434 records. Key features include player roles such as teammate, actor (the player performing the action), and keeper, as well as their exact locations on the field.

|  |  |  |  |
| --- | --- | --- | --- |
| match\_id | teammate | actor | location |
| 3942819 | true | false | [22.96, 41.02] |
| 3942819 | false | false | [46.54, 46.65] |
| 3942819 | true | false | [70.01, 41.18] |
| 3942819 | true | true | [60.00, 40.00] |

Table 03: Example of frames data

**Dataset Limitations**

The dataset has several limitations that may affect its comprehensive analysis. One of the primary issues is the lack of continuity, as the data is only available from sporadic years, such as 2014, 2015, and 2017, with no data for certain years, creating gaps in the timeline. Additionally, only the 2015 season is available in full for the top 5 leagues, which limits the consistency and comparability of the data. The dataset also lacks contextual information, such as weather conditions. Player details are limited, with no information on height, weight, footedness, which are critical for understanding player performance in different contexts. Lastly, while the dataset provides detailed event information, the granularity of these events can complicate the analysis, especially due to the use of dummy-like structures and dictionaries in some fields, which may be challenging to parse for less experienced users.

**3 Feature Selection and Engineering**

There are many variations of an xG model, often determined by the selected features. Some models are more sophisticated, incorporating extensive details. In this section, we will identify relevant features from our dataset, explore their meaning and values, and assess their importance within the context of xG.

The simplest xG models typically rely on two key features: distance to goal and shot angle. In our dataset, this information is embedded within the shot location data, so our first step will be to select the relevant features already present in the dataset.

**1. Location** is the most critical feature in an xG model, as it directly represents the position from which a shot was taken. It provides valuable insights, such as the player's proximity to the goal and the angle of the shot.

In the dataset, location is captured as a pair of values: the first value corresponds to the horizontal coordinate (x), while the second represents the vertical coordinate (y). The StatsBomb field dimensions range from 0 to 120 for the x-axis (horizontal distance from the goal line) and 0 to 80 for the y-axis.

|  |  |  |
| --- | --- | --- |
| type | location | xg |
| Shot | [105.6, 44.0] | 0.07213958 |
| Shot | [98.6, 25.2] | 0.02977089 |
| Shot | [105.5, 47.3] | 0.07589752 |
| Shot | [113.4, 38.7] | 0.15686217 |
| Shot | [106.2, 36.8] | 0.15133068 |

Table 03: Example of location values in the dataset

A football field with orange and green dots

Description automatically generated

Figure 01: Sampled Distribution of 1,000 Shots on the Pitch Using Clustering

**2. Categorical Variables**

**a. Play pattern** is another important feature in the dataset as it indicates the type of play during the event leading to the shot. This information helps to distinguish the context in which the shot occurred, providing insight into whether the shot was made during open play, a set piece, or from a counterattack, among other scenarios. The values in the play pattern column can help us understand how different game situations influence shot outcomes.

In the dataset, the "play pattern" is represented by categorical values, with 9 distinct types:

1. **Regular Play:** Shots taken during normal game flow.
2. **From Free Kick:** Shots originating from free kick situations.
3. **From Throw In:** Shots taken after a throw-in.
4. **From Counter:** Shots made following a counterattack.
5. **From Goal Kick:** Shots taken after a goal kick.
6. **From Keeper:** Shots made after the goalkeeper’s play.
7. **From Kick Off:** Shots right after a kick-off.
8. **Other:** Other types of plays.

This feature represents the tactical context of the shot and can reveal how the situation impacts shot quality and conversion likelihood. For example, shots from "Regular Play" may generally have more dynamic and varied angles and distances, whereas shots from a "Free Kick" might be more controlled or involve a set strategy.

|  |  |  |
| --- | --- | --- |
| type | play\_pattern | xg |
| Shot | From Kick Off | 0.07213958 |
| Shot | From Keeper | 0.02977089 |
| Shot | From Throw In | 0.07589752 |
| Shot | From Corner | 0.15686217 |
| Shot | From Throw In | 0.15133068 |

Table 04: Example of play pattern values in the dataset

A graph of blue rectangular bars with white text

Description automatically generated

Figure 02: Distribution of play patterns for goals

A graph of a number of blue rectangular bars with white text

Description automatically generated

Figure 03: Distribution of play patterns for no-goal shots

**b. Shot Type** refers to the category of play in which the shot was taken. This feature helps to distinguish between shots based on the nature of the play or situation leading to the shot. Understanding the shot type is important as different shot types typically have different expected goal (xG) values associated with them. For example, penalties often have a higher chance of scoring than shots taken during open play.

In our dataset, the shot type is represented by a categorical variable with four possible values:

1. **Open Play** – A shot taken during the flow of the game, without any set-piece situation.
2. **Corner** – A shot that is taken from a corner kick.
3. **Free Kick** – A shot that results from a free kick awarded during the game.
4. **Penalty** – A shot taken from the penalty spot, usually awarded after a foul in the box.

|  |  |  |
| --- | --- | --- |
| type | shot\_type | xg |
| Shot | Corner | 0.000180 |
| Shot | Free Kick | 0.039566 |
| Shot | Open Play | 0.072140 |
| Shot | Penalty | 0.783500 |

Table 05: Example of shot type values in the dataset

A blue circle with orange and green segments

Description automatically generated

Figure 04: Proportion of shot types resulting a goal

**c. Shot Body Part** refers to the part of the body used by the player to take the shot. The dataset contains the following values for this feature:

* **Right Foot**: The shot was taken using the player's right foot.
* **Left Foot**: The shot was taken using the player's left foot.
* **Head**: The shot was taken using the player's head.
* **Other**: The shot was taken using another body part or a unique situation.

|  |  |  |
| --- | --- | --- |
| type | shot\_body\_part | xg |
| Shot | Right Foot | 0.07213958 |
| Shot | Right Foot | 0.02977089 |
| Shot | Right Foot | 0.07589752 |
| Shot | Head | 0.15686217 |
| Shot | Left Foot | 0.15133068 |

Table 06: Example of shot body part values in the dataset

A pie chart with different colored circles

Description automatically generated

Figure 06: Proportion of Shot Body Parts resulting a goal

**d. Shot Technique** refers to the method or technique used by the player to execute the shot. The dataset includes the following values for this feature:

* **Backheel**: The shot was taken using the backheel technique.
* **Diving Header**: The shot was taken using a diving header.
* **Half Volley**: The shot was taken using a half-volley technique.
* **Overhead Kick**: The shot was taken using an overhead kick.
* **Volley**: The shot was taken using a volley technique.
* **Lob**: The shot was taken using a lob technique.
* **Normal**: The shot was taken using a standard or regular shot technique.

|  |  |  |
| --- | --- | --- |
| type | shot\_technique | xg |
| Shot | Half Volley | 0.07213958 |
| Shot | Normal | 0.02977089 |
| Shot | Half Volley | 0.07589752 |
| Shot | Normal | 0.15686217 |
| Shot | Normal | 0.15133068 |

Table 07: Example of shot technique values in the dataset

A graph with blue squares

Description automatically generated

Figure 08: Distribution of shot techniques for goals

**3. Boolean Variables**

**a. Under Pressure** indicates whether the player was under pressure (i.e., being defended or challenged) when taking the shot.

# References

1. *An xG of Their Own: Using Expected Goals to Explore the Analytical Shortcomings of Misapplied Gender Schemas in Football.* **Narayanan, Sachin & Pifer, Nathan.** 2023, Journal of Sport Management, pp. 1-18.

2. *Footedness: How important is it? Which is better, a brilliant right footer or an average left footer?* **Carter, Mark.** Manchester : StatsBomb, 2024. StatsBomb Conference. p. 22.

3. **Hakim, Alfian.** How to Build Your Own Expected Goals (xG) Model. *Medium.* [Online] April 12, 2023. https://medium.com/@alf.19x/how-to-build-your-own-expected-goals-xg-model-2bd186dccdf7.

4. **Whitmore, Jonny.** What Is Expected Goals (xG)? *The Analyst.* [Online] August 8, 2023. https://theanalyst.com/2023/08/what-is-expected-goals-xg.

5. **Roccetti, Marco & Berveglieri, Filippo & Cappiello, Giuseppe.** *Football Data Analysis: The Predictive Power of Expected Goals (xG).* 2024.

6. *Expected goals in football: Improving model performance and demonstrating value.* **Mead, James & O'Hare, Anthony & McMenemy, Paul.** 2023, PLOS ONE, p. 18.

7. *Improving the Expected Goal value in football using Multilayer Perceptron Networks.* **Manuel Mendez, Carlos Montero, and Manuel Nuñez.** Phuket, Thailand : s.n., 2023. 15th Asian Conference on Intelligent Information and Database Systems. p. 12.

8. **THE FOOTBALL ANALYTICS HANDBOOK. *soccerment.* [Online] 2020. https://soccerment.com/wp-content/uploads/2020/07/handbook\_fa.pdf.**