The Finisher's Signature: A Data-Driven Analysis of Messi vs. Mbappé

An Application of an Expected Goals (xG) Model to Profile Elite Goal-Scorers

Prepared by: Chadrick Clarke

Date: July, 2025

email: clarke.chadrick@gmail.com

LinkedIn: https://www.linkedin.com/in/chadrickclarke/

Prepared by: Chadrick Clarke
Date: July, 2025

Table of Contents

Prepared by: Chadrick Clarke
Date: July, 2025

Executive Summary

In modern football, technical directors and managers understand that assessing a player's finishing ability simply on goal counts is misleading. It does not provide enough insight to say whether a striker was fortuitous in the past season or his goal tally is a sustainable output. If you are a sporting director or manager you can readily see the implications this potentially has on your transfer plans for the upcoming season.

This project presents as a solution, a custom Expected Goals (xG) model developed to quantify the quality of every shot, enabling a more objective and data-driven assessment of a player's finishing skill.

To demonstrate the value this tool offers a team's analytics department, a post season analysis of the finishing profile for Kylian Mbappe and Lionel Messi were conducted. Based on the analysis:

1. Both Messi and Mbappe are confirmed as elite, long-term overperformers. Analysis of four seasons of data shows Mbappe scored 18 more goals than expected, while Messi scored 5.4 more; proving their finishing is a durable, world-class skill.
2. The players achieve this through distinct finishing signatures. Our model diagnostics revealed Mbappe excels as a *Dynamic Finisher*, converting difficult chances in chaotic situations. In contrast, Messi acts as a *Technical Maestro*, engineering geometrically perfect scoring opportunities through superior technique and space creation.
3. The custom model proved to be a powerful diagnostic tool. While simpler than industry benchmarks, its disagreements with more complex models were instrumental in identifying these nuanced player archetypes

With the aid of this analytical framework, the project successfully demonstrates how modeling a player's finishing throughout a season provides not only a more accurate and reliable measure of performance but a deeper insight into the unique stylistic signatures of world-class players.

Prepared by: Chadrick Clarke
Date: July, 2025

Introduction

Football is arguably the world's most popular sport, with competitions like the English Premier League generating billions in viewership annually. Yet, the event that drives this global passion,the scoring of a goal, is a statistical rarity. Only 10-15% of shots taken typically result in a goal, it is a game defined by low-probability events of immense importance. In order to win, a team must score, thus the deconstruction of what turns a simple shot into a goal becomes a fundamental but critical analytical challenge.

This report aims to meet that challenge. First by building a statistical Expected Goals (xG) model to quantify the factors that define a high-quality scoring opportunity and then employing this model to create a detailed 'finishing profile' for two of the modern era's greatest players: Kylian Mbappé and Lionel Messi. This will provide the answer to an important question: 'What does the data say about their finishing skill?

To deconstruct a goal, the report focuses on the metric of Expected Goals (xG) . As leading football analyst David Sumpter (2023) argues, any useful metric must be explainable in terms of player actions. xG excels here: it is a value between 0 and 1 that represents the probability of a shot becoming a goal based on its characteristics. A shot with an xG of 0.7 is a high-quality chance expected to be scored 70% of the time, while a shot with an xG of 0.02 is a low-quality chance.

For those who watch the sport, intuitively you know the most critical factors are the shot’s distance and angle to the goal. Hence, goals scored from far distances and impossible angles have been described as “worldie” or “spectacular”. This principle is echoed by (Sumpter, 2023) who states that when evaluating a shot the viewing angle the player has of the goal is important to his chances of scoring. This was confirmed by the model's feature importance analysis (see Figure 8). In other words A 'worldie' is, by definition, a goal scored from a low-xG situation.

In deconstructing a goal and the elements that contribute to it, a statistical model was built using logistic regression. This algorithm was chosen for its interpretability and ease of understanding how different features contribute to a goal. The dataset used for this analysis was taken from StatsBomb free football tracking data found on its Github repository.

Exploratory data analysis and model development was done on data aggregated from 6 major tournaments; Africa Nations Cup 2023, Copa America 2024, Uefa Euros 2020 and 2024 and the World Cup 2018 and 2022. Combined this represented 7863 recorded shot events with 42 features.

The saved model was then used to assess the finishing profile for Kylian Mbappe and Lionel Messin over two seasons in Ligue 1 (2021-2023). However, it should be noted that the free packaged provided by StatsBomb does not provide the entire number of shot events for every team and player across these competitions or all the data points that are captured. This was one limitation identified from the dataset.

Notwithstanding there were enough data points to draw some generalised conclusions about what the features of a good shot are. This will be discussed in detail in the next section.

Exploratory Analysis - The Anatomy of a Shot

Exploratory data analysis (EDA) main purpose is to help better understand patterns found within data, detect outliers and anomalies events as well as find interesting relationships among variables (IBM, 2025). This exploration provided insights into the characteristics surrounding a shot.

Finding 1: Shots are Concentrated in a "Prime Scoring Zone

Analysis of the 7863 shot events reveal a startling clear and deliberate strategy from players; get the ball into a high danger area before shooting. The distribution of shots is heavily concentrated between 5 and 35 yards from the goal, with a distinct peak at approximately 12 yards; right around the penalty spot. Also, the most frequent shot angle was observed around 0.25 radians (approximately 14 degrees), indicating a preference for taking shots with a relatively central open view of the goal.

This deliberateness from players intuitively supports the notion that the odds of scoring are increased as the player approaches the goal and has a better viewing angle.

The concentration of shots in a prime zone (penalty area) directly impacts shot quality. Despite players' attempts to get into good scoring positions, a majority of these shots will represent low probability events, aligning with the 10-15% conversion rate observed across the sport. This assessment is supported by the distribution of Expected Goals (xG) which is heavily skewed towards lower values, with a peak at a modest 0.02 xG.

Finding 2: The Dominance of Open Play

While set-pieces and penalties provide unique opportunities, the vast majority of shots originate from open play. This reinforces the idea that most scoring chances are not from structured, dead-ball situations but are created during the dynamic flow of the game. Interestingly a high concentration of these shots also fall within the 15% probability range (see Figure 1) for scoring according to (Sumpter, 2023).

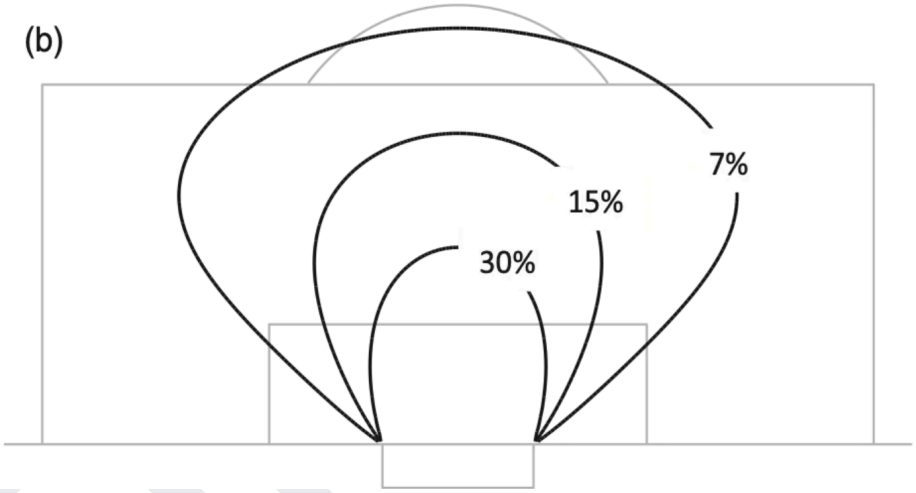

Figure 1: Plot showing probability for different scoring zones

Finding 3: The Inherent Trade-off Between Distance and Quality

A correlation analysis between key shot features confirms the intuitive principles of goal-scoring; the closer to the goal the better the odds of scoring. As shown in the scatter plot below there is a strong inverse relationship between shot_distance and shot_xg, with shots taken from further away systematically having a lower xG value. This visualizes the fundamental trade-off every player faces: shooting from distance is a low-percentage play compared to working the ball closer to the goal.

Prepared by: Chadrick Clarke
Date: July, 2025

A Note on Data Distribution and Outliers

All key numerical features (shot_distance, shot_angle and shot_xg) exhibit a right-skewed distribution for shots taken. Thus the average value for each is influenced by extreme values (i.e. Long range shots, shots from impossible angles). It is common practice in many modeling scenarios to manage these outliers; however, for this analysis, these “outliers” represent legitimate football events.

Also considering the statistical rarity of a goal, these data points were kept in order to build a model that understands the full spectrum of the game, including the probability of a 40 yard “screamer”.

Finding 4: Where are the goals scored?

Not surprisingly most goals are scored within the 18 yard box with the most dense concentration of goals around the penalty mark towards the yard box. Those goals with the highest xG as expected were found within the 6 yard box with lower xG goals fanning outwards. This further reinforces the simple principle the closer to the goal the better the odds of scoring.

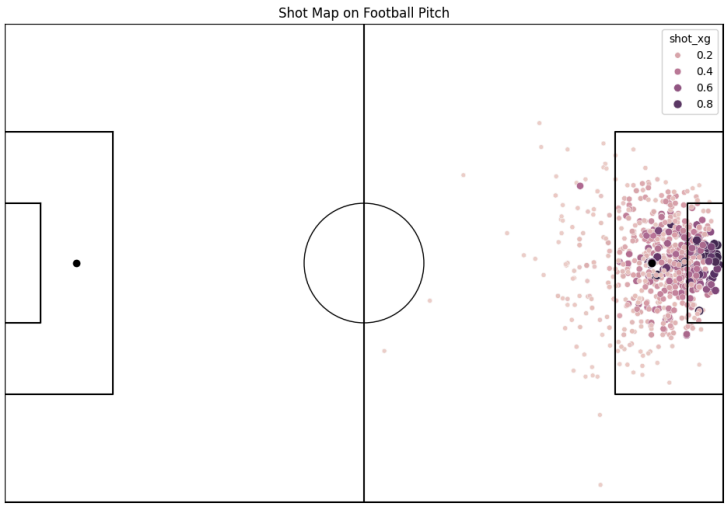

Figure 2: Shot Map on football pitch for shot’s resultign in a goal

Conversely for shots which did not result in a goal there is a high concentration of low xG shots at the 18 yard line and towards the half circle.

Prepared by: Chadrick Clarke
Date: July, 2025

Finding 5: Which play pattern resulted in higher quality shots?

As mentioned earlier a majority of shots are taken from regular or open play but on average the xG is only 0.09. But this low xG is expected due to the larger sample size. Shots taken from a counter have the second highest xG with that of .14 but these events are significantly smaller than regular play. The play event only described as other by Statsbomb has the highest xG at 0.75 but also has a low number of shot events. The documentation did not provide much detail on this category, but as the name suggests and the high xG these are unusual phases of play which may occur in the game (i.e. shot taken after a restart to a game).

However, cumulatively there is a great opportunity for players and coaching staff to create high scoring situations from dead ball situations such as throw-ins, corner kicks, goal kicks etc.

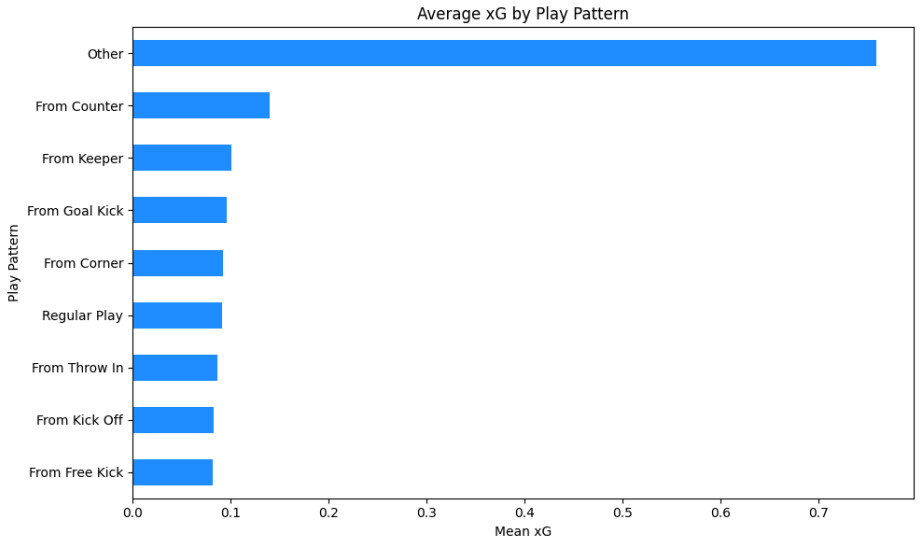

Figure 3: Horizontal bar plot of average xG by Play Pattern.

Finding 6: Which body part is mostly used at the end of a play?

Right foot and left foot are the preferred instruments by which a shot is taken across most play patterns except corners. The use of the head is preferred in these situations, however, these shooting opportunities are not always high probability situations despite the closeness to goal.

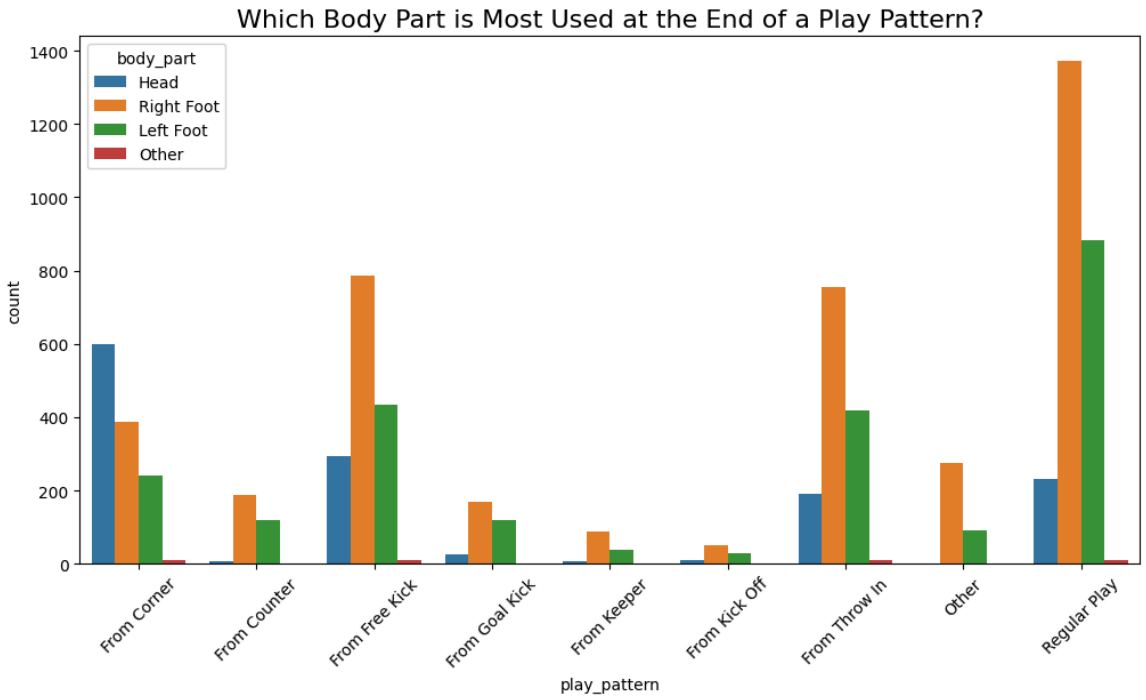

Figure 4: Barplot showing which body part is most used by play pattern.

These opportunities are usually within a high traffic zone, under pressure and require more skill in directing a shot with the head. It's not surprising the xG for shots taken with the head are just 0.10, considering the small number of headed shot events compared to left and right foot shots.

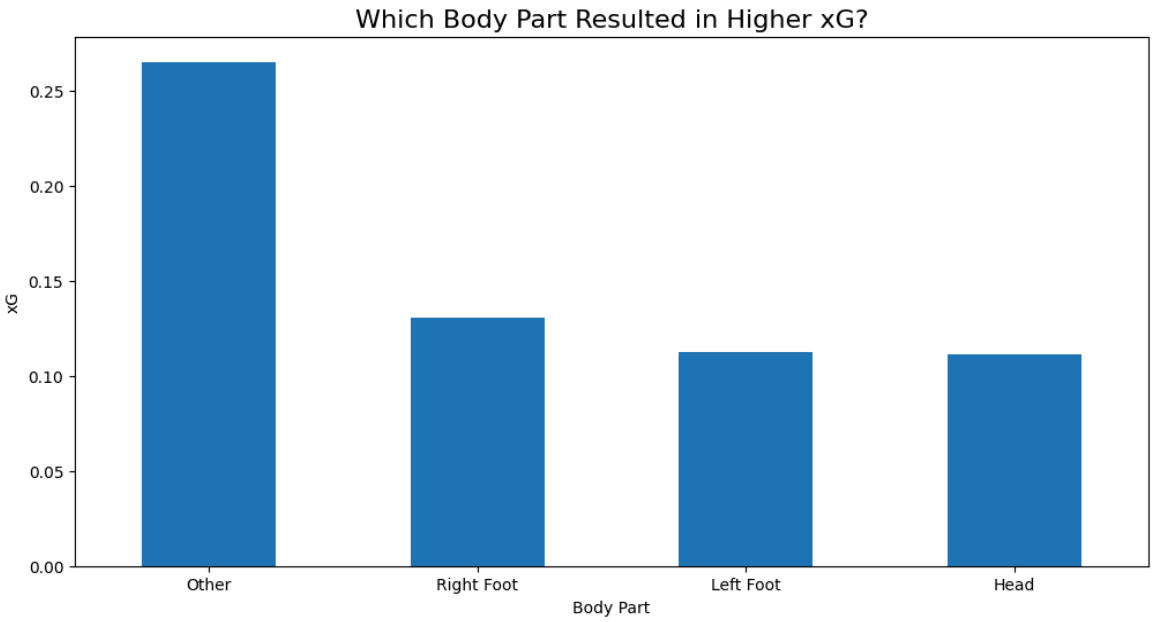

Figure 5: Barplot showing which body part results in high quality shots.

Finding 7: What is the average shot distance per body part

Consistent with the trend of the data the average distance for shots taken for right and left are approximately 20 yards from goal. This underscores the notion of football being a low scoring game, though the reasons are varied and dynamic, a high number of shots on average are taken from low probability zones.

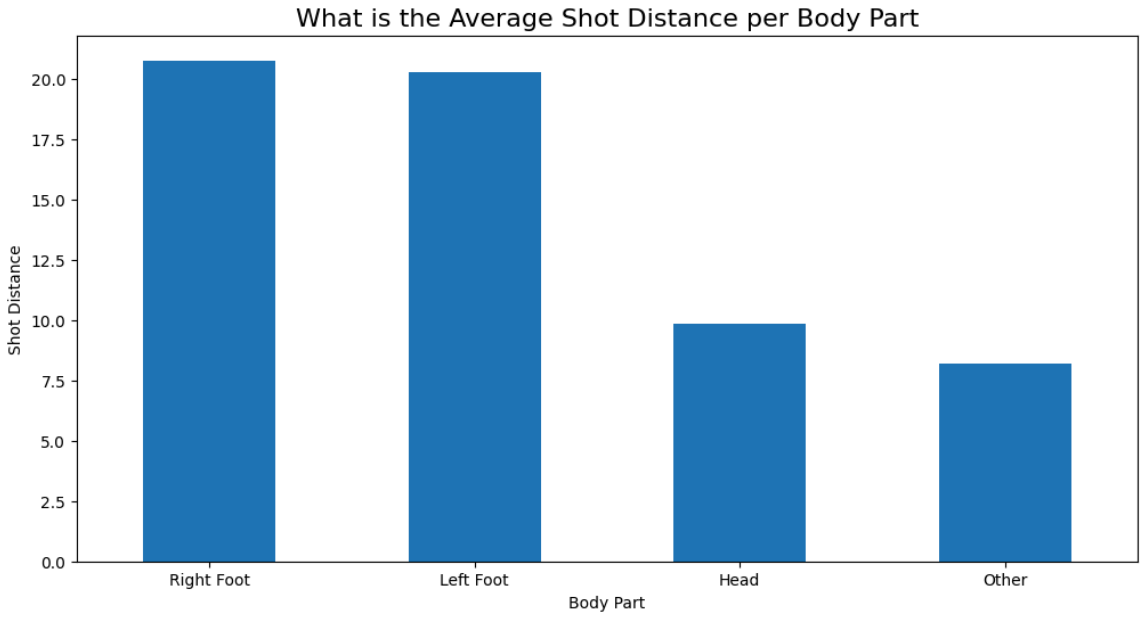

Figure 6: Barplot showing average distance for shot by body part.

Building the xG Model - Quantifying the Opportunity

Model Choice and Objective

In order to quantify shot quality, a logistic regression model was built. This algorithm was chosen for its interpretability and its natural output of probabilities between 0 and 1, which in this context would represent a shot's Expected Goals (xG) value. The main challenge in this task is the inherent class imbalance as observed in the training data, where only 11.6% of shots resulted in a goal. Thus when evaluating model performance the focus should not just be on overall accuracy, but on the model's ability to correctly identify these rare but critical goal-scoring events.

Model Performance: A Robust and Reliable Predictor

The model demonstrates strong and stable performance, generalizing well from the training set to the unseen test set. The primary metric for a classifier on an imbalanced dataset is the ROC AUC score; this indicates how good a classifier is at distinguishing between classes. The model scored an AUC of 0.83 on the test set; thus affirms that the model is significantly better than random chance and shows a strong capacity to separate high-quality chances from low-quality ones.

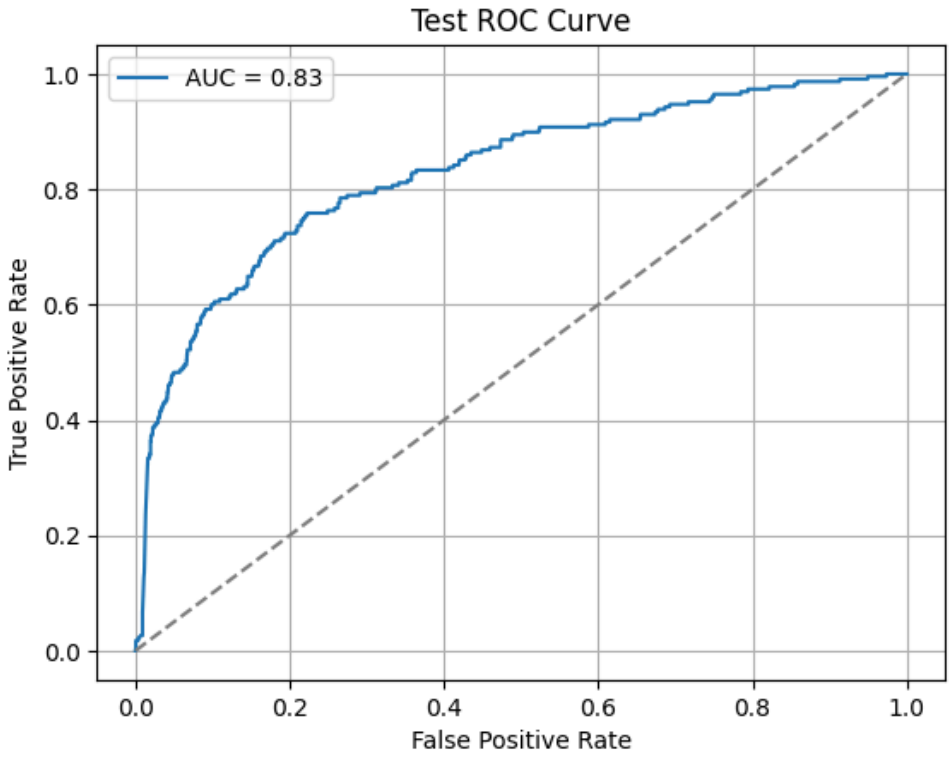

Figure 7: Plot of the ROC curve of the model's performance during the test phase.

This is further supported by the recall score for the 'Goal' class, which was 0.72. This means the model successfully identified 72% of all actual goals scored in the test set. While the precision of 0.32 is modest, it is expected in a problem where positive events are so rare. The key takeaway is that the model is highly effective at its primary job: finding the shots that are most likely to be goals.

Metric	Train	Test (Tuned)
Precision (Goal)	0.32	0.32
Recall (Goal)	0.70	0.72
AUC	0.83	0.83
Brier Score	0.159	0.076
Accuracy	0.79	0.79

Table 1: Model performance metrics

Calibration: Are the model's probabilities accurate?

Beyond just separating classes, an xG model's probabilities must be reliable. In order to determine how reliable these probabilities are, the Brier Score was usd. This technique measures the accuracy of these probabilities, with values close to 0 being ideal. Initially the model had a score of 0.159, but after applying calibration techniques, this improved dramatically to 0.076, indicating a much more trustworthy model.

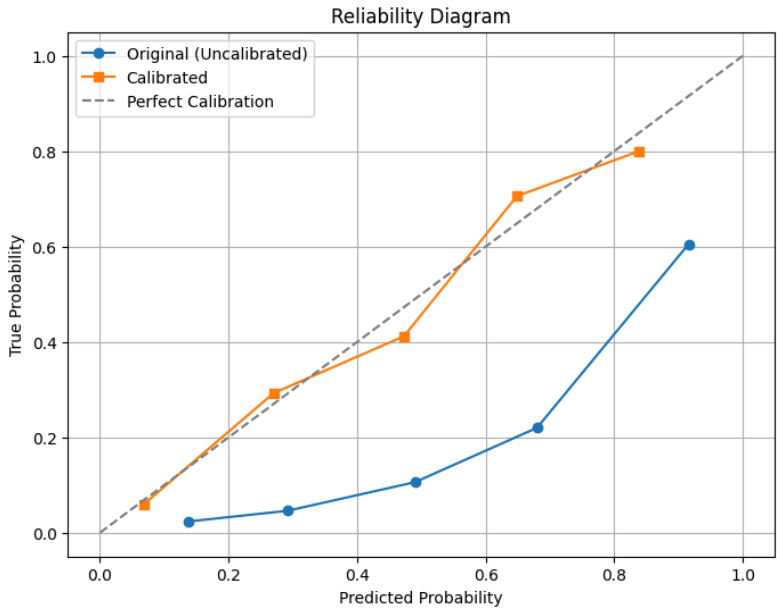

Figure 8: Plot of the model's reliability graph showing uncalibrated predicted probabilities versus calibrated.

The reliability diagram visually confirms this. The uncalibrated model (blue line) was overconfident, systematically underestimating probabilities. The calibrated model (orange line) tracks the 'Perfect Calibration' line much more closely, ensuring that when the model predicts a 40% chance of a goal, the real-world frequency is indeed around 40%.

Model Interpretation: The Anatomy of a High Quality Shot

By examining the model's internal logic using SHAP feature importance and odds ratios, we can deconstruct the key elements of a goal-scoring opportunity.

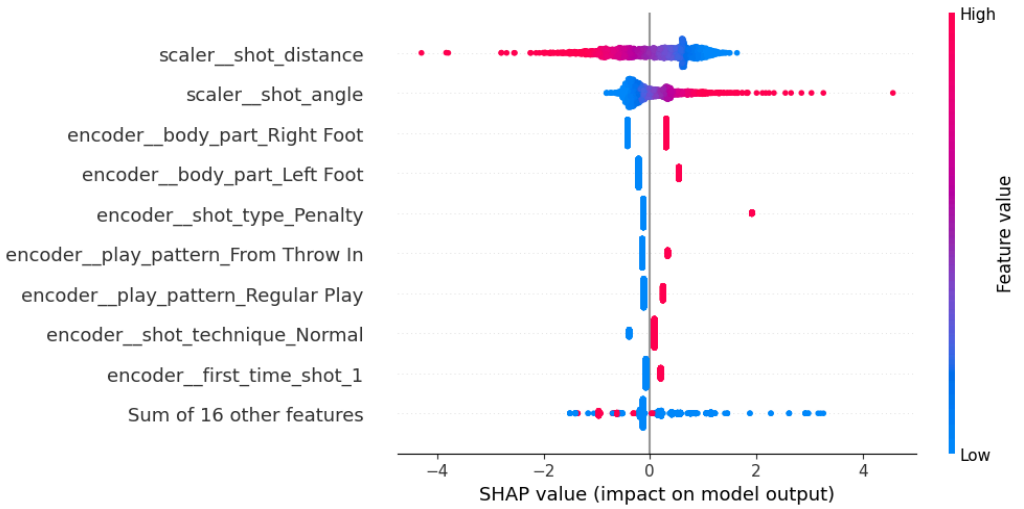

Figure 9: Plot of the model's feature importance

Location is King

As expected, shot_distance and shot_angle are by far the most influential features. The SHAP plot shows that low shot distance (blue dots) and high shot angle (red dots) have a strong positive impact on the model's output, pushing the prediction towards 'Goal'. Our odds ratio analysis quantifies this: for every one-yard increase in distance, the odds of scoring drop by approximately 55%, holding all else constant.

Technique Matters

The model confirms football intuition about shot difficulty. An Overhead Kick is incredibly difficult, reducing the odds of scoring by 73% compared to a standard shot. Conversely, a Lob shot, often taken when a goalkeeper is out of position, dramatically increases the odds of scoring by over 9 times, making it one of the most potent shot types.

An Interesting Quirk

The Left-Foot Advantage? An unexpected finding was the model's slight preference for left-footed shots over right-footed ones. Both are positive predictors, but the coefficient and SHAP values suggest left-footed shots are considered marginally more dangerous. This could be a statistical artifact of the players in this dataset (e.g., a high number of elite left-footed finishers) and warrants further investigation in future work.

Game State Context

Penalties, as expected, have a massive positive impact, increasing the odds of a goal by over 7 times. This confirms the model is correctly identifying these unique, high-leverage game events.

Feature	Coefficient	Odds ratios
Shot Technique - Lob	2.257113	9.555459
Shot Type - Penalty	1.970728	7.175901
Body Part- Left Foot	0.756524	2.130856
BodyPart - Right Foot	0.726814	2.068480
Shot Angle	0.506814	1.659994
Shot Technique Volley	-0.101841	0.903173
Shot Distance	-0.802180	0.448350
Shot Technique- Overhead Kick	-1.297239	0.273285

Table 2: Model coefficients and Odds ratios

Methodology and Limitations

Data Sources

This analysis triangulates findings from two primary sources to ensure robustness:

- Event-Level Data: Sourced from the StatsBomb open data repository. This provided granular, on-the-ball event data for model building and diagnostics. The training dataset comprised 7,863 shot events from six major international tournaments (AFCON 2023, Copa America 2024, etc.), while the player-specific analysis utilized data from the 2021-2023 Ligue 1 seasons.
- Summary-Level Data: Sourced from FBRef.com (powered by Opta). This provided high-level, longitudinal performance statistics used to validate the overall finishing trends for Messi and Mbappé over multiple seasons.

Modeling Approach

- Algorithm Choice: A Logistic Regression model was selected for this project. As a binary classification algorithm, it is well-suited to the "Goal" vs. "No Goal" problem. Its primary strength lies in its interpretability and its natural output of probabilities, which directly serve as the Expected Goals (xG) value for each shot.
- Key Features: The model was trained on a set of core features describing each shot event, including (but not limited to) shot location coordinates (shot_distance, shot_angle), body_part, play_pattern, and shot_technique.

Core Methodological Decision

- Retaining Outliers: A crucial decision was made to not remove statistical outliers from the feature data. In the context of football, an "outlier", such as a very long-range shot is not a data error but a legitimate, albeit rare, event. Retaining these data points is essential for the model to learn the full spectrum of shot probabilities and to accurately model the low-probability nature of goal-scoring.

Acknowledged Limitations

To properly contextualize the findings of this report, two primary limitations must be acknowledged:

- Sample Data: The StatsBomb open data represents a sample of all available matches, not a complete archive. While sufficient for building a robust model and drawing conclusions, the findings are based on the specific matches included in this public dataset.
- Feature Constraints: The custom model does not include advanced, proprietary features such as the real-time positions of defenders and the goalkeeper. These context-rich features are present in industry-leading models (like StatsBomb's full commercial version) and their absence in our model likely explains a significant portion of the variance observed when comparing predictions, particularly for dynamic players like Mbappe.

Player Finishing Profile

The Verdict: Consistent Overperformance(FBRef Data)

To establish a real-world baseline for finishing skill, data from FBRef (powered by Opta) was used. This high-level view, covering four full seasons, provides undeniable evidence that both Kylian Mbappe and Lionel Messi are elite finishers who consistently score more goals than expected.

Player	Period	G	xG	G-xG
Mbappe	2021-2023	57	50.1	6.9
	2023-2025	58	46.9	11.1
Messi	2021-2023	22	25.5	-3.5
	2023-2025	31	22.1	8.9

Table 3: G-xG Performance for Messi and Mbappé (2021-2025).

Over the four-year period from 2021 to 2025, Mbappe scored an astonishing 18 more goals than the quality of his chances suggested. This overperformance was not a lucky streak but a demonstration of consistent world-class finishing ability.

Similarly, Messi overperformed his xG by 5.4 goals over the same period. While his overperformance was more pronounced in the latter two seasons, it confirms that even in the late stages of his career, he continues to maintain his status as a legendary finisher who adds value beyond that of an average player. This robust, large-scale data confirms that both players are, without question, in the absolute top tier of goal-scorers.

The Signature: Uncovering How They Overperform

The FBRef data tells us that they overperform but to really understand how their styles differ the custom xG model was used as a diagnostic tool; evaluating its predictions to the StatsBomb benchmark on a shot-by-shot basis. This revealed two distinct "finishing signatures.

Mbappe: The “Dynamic Finisher

The analysis of Mbappé's shots reveals a systematic disagreement between the custom model and the benchmark. As seen in the plot (see Figure 10), the vast majority of points fall below the y=x line of perfect agreement. This means the model, based on simpler geometric features, consistently rated his chances as lower quality than the more context-rich StatsBomb model.

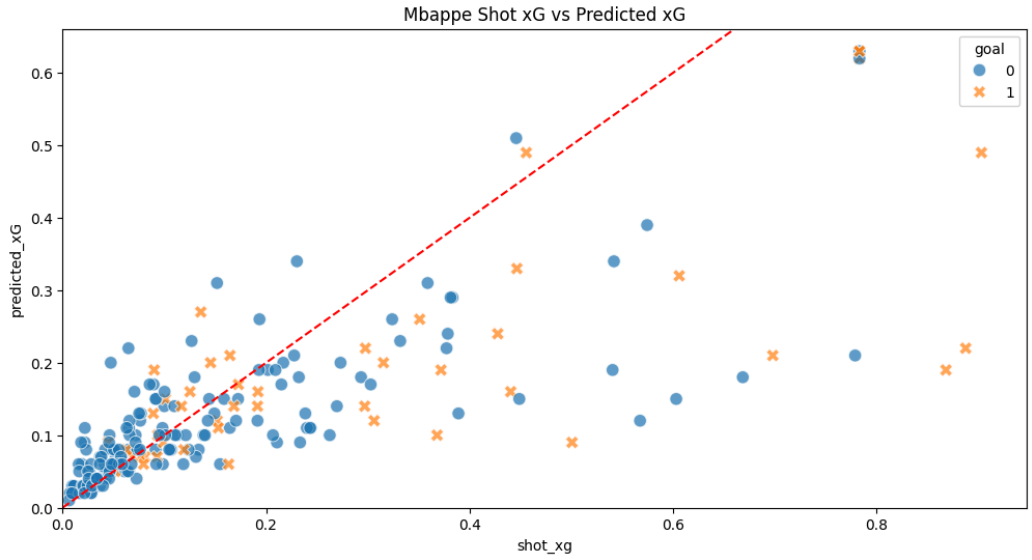

Figure 10: A scatter plot of Mbappe’s shot xG (benchmark) versus Predicted xG

This means Mbappé's genius lies in his ability to score from dynamic, often chaotic situations created by his explosive pace. He thrives in moments where defender positions and high-pressure states of play, features the model cannot see, are critical. In other words he is a master of converting chances that look difficult from a standard analytical perspective, hence the "Dynamic Finisher" archetype.

Messi: The “Technical Maestro

In stark contrast to Mbappe, the plot for Messi shows a much stronger agreement between our model and the benchmark. The points are clustered far more tightly around the y=x line.

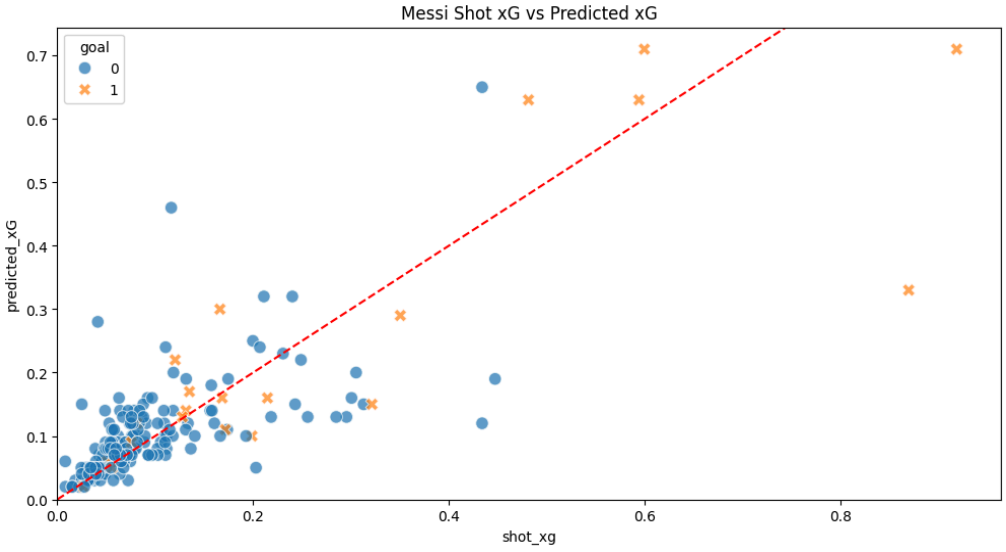

Figure 11: A scatter plot of Messi’s shot xG (benchmark) versus Predicted xG

This means Messi's style relies on his technical understanding of the game in order to create geometrically "clean" scoring opportunities. His movement and control create situations so optimal that even a simpler model can accurately recognize their high quality. He doesn't just finish chances; he engineers them to be as high-probability as possible before the shot is taken. This is the signature of the “Technical Maestro”.

Synthesizing the Evidence: A Multi-Model Conclusion

This multi-faceted analysis enables the creation of a complete picture. On one hand the FBRef data provides the robust, high-level verdict of sustained overperformance while on the other the custom model diagnostics then provide the nuanced story of how this overperformance is achieved.

The custom model was also validated using these benchmarks (StatsBomb and Fbref). For instance, the custom model predicted a massive +19.52 G-xG for Mbappe but the FBRef and StatsBomb results predicted around +7 to +11 per two seasons. This provides sound context to be able to establish that the model correctly identified the signal of his elite finishing but inflated the magnitude due to its systematic underestimation of his shots. This is a crucial finding, demonstrating how simpler models can be used as diagnostic tools even when they don't perfectly match more complex benchmarks.

Player	Period	G	xG(predicted	xG(statsBomb)	G-xG(predicted)	G-xG
Mbappe	2021-2023	49	29.38	38.52	19.52	10.48
Messi	2021-2023	22	24.19	23.13	-2.19	-1.13

Table 4: G-xG Performance for Messi and Mbappé (2021-2023) based on custom model.

Ultimately, by triangulating evidence from the custom model, StatsBomb, and FBRef, the conclusion is that both players are generational finishers, but they achieve their results through different, statistically identifiable means.

Conclusion and Future Work

In conclusion both Kylian Mbappe and Lionel Messi are elite finishers who consistently outperform their stated xG; this underscores the world-class talent they possess. The insight from the analysis supports the conclusion that Mbappe’s ability to score from less structured scoring situations is emblematic of a Dynamic Finisher. On the other hand Messi’s ability to create high quality geometrically perfect chances, which are easy for the model to ascertain is indicative of a Technical Maestro.

Despite the model’s tendency to be pessimistic with xG predictions compared to benchmarks,highlights the model’s challenge in properly model dynamic situations (like those Mbappe thrives in) and would benefit from richer situational data such as but not limited to defender/keeper position, teammate position, etc.

The strength of this analysis lies in its multi-faceted approach. By triangulating findings from our custom model, the StatsBomb benchmark, and high-level FBRef data, we were able to build a robust and highly credible picture of player performance. This highlights the power of using simpler models not just for prediction, but as diagnostic tools to probe and understand more complex systems

To further improve this model and its predictive power, experimenting with an ensemble method like xgBoost could prove useful. Ideally increasing the feature space would be beneficial and this would involve exploring the use of StatsBomb 360 data, which provides crucial information on player positions surrounding each event

References

1. Sumpter, D. (2023, August 14). *The geometry of shooting*. Medium. https://soccermatics.medium.com/the-geometry-of-shooting-7cbdd0d5da3b
2. IBM. (2025, June 4). *What is exploratory data analysis?*. IBM. https://www.ibm.com/think/topics/exploratory-data-analysis#:~:text=The%20main%20purpose%20of%20EDA%20is%20to%20help,anomalous%20events%2C%20find%20interesting%20relations%20among%20the%20variables.