Football Shot Prediction

Nishal Nivya Rodrigues
80537308
Masters of Data Science
University of Europe for Applied Sciences
14469 Potsdam, Germany
nishal.rodrigues@ue-germany.de

Harsha Narasimhe Gowda
75954486
Masters of Data Science
University of Europe for Applied Sciences
14469 Potsdam, Germany
harsha@ue-germany.de

Frank Stalin D'Souza
91630537
Masters of Data Science
University of Europe for Applied Sciences
14469 Potsdam, Germany
frank.dsouza@ue-germany.de

***Abstract*—Predicting the outcome of football shots has become a central focus in sports analytics due to its importance in tactical decision-making and performance evaluation. Traditional expected goals (xG) models often rely on limited features such as shot distance and angle, ignoring crucial contextual and temporal variables such as game state, pressure, and shot technique. Despite growing interest in predictive modeling, many existing studies do not address class imbalance, lack interpretability, or use proprietary data, making them less scalable and transparent. This study addresses those gaps by using open source StatsBomb event data and applying interpretable ensemble machine learning methods. We trained the Logistic Regression, Random Forest, XGBoost, and CatBoost models using engineered features such as shot angle, first time shots, and pressure status. Among these, CatBoost achieved the highest performance with accuracy 94% and an F1 score of 0.75. Feature importance analysis revealed that contextual features significantly influence goal prediction. Our work demonstrates that high-performing, interpretable models can be developed using open data. This contributes to a reproducible and scalable approach to the prediction of football shots that supports real-world analytical and coaching decisions.**

***Index Terms*—football analytics, expected goals, machine learning, CatBoost, shot prediction**

I. Introduction

Football analytics is a rapidly expanding field that applies data science and statistical techniques to understand, evaluate, and predict events in football matches. Among its key applications is the development of predictive models to estimate the likelihood of outcomes such as goals, passes, and player performance. One of the most widely discussed metrics in this domain is "Expected Goals" (xG) [1], which aims to quantify the quality of a shot based on historical data. Traditional xG models typically rely on shot distance and angle, offering only a basic representation of shot difficulty. However, there is increasing recognition that shot outcomes are influenced by a broader set of contextual and temporal factors.

Predicting the probability of a goal from a shot has critical implications for coaching decisions, player evaluation, and match commentary. Coaches can use such models to assess player efficiency; analysts can evaluate tactical effectiveness; and broadcasters can enhance fan engagement with quantitative insights. As shown in [2], machine learning models that integrate contextual data outperform basic models, leading to a more nuanced understanding of in-game events. Improving the quality of shot prediction models directly benefits decision-making across all levels of the game.

Despite the growing interest in advanced xG modeling, many existing approaches face limitations in either explainability or generalizability. Deep learning models, for example, can provide high accuracy but often behave as black boxes [3], making it difficult for coaches and analysts to understand why a prediction was made. In contrast, simpler models, such as logistic regression, may lack the ability to capture complex interactions between variables. Bridging this gap requires a thoughtful balance between model performance and interpretability [4], [5], especially in a domain like football, where decisions often hinge on clear and communicable insights. This study aims to explore this middle ground by leveraging gradient-boosting techniques, which have shown promise in recent research for both accuracy and transparency.

At the same time, the democratization of football data has lowered the barrier to entry for research in this field. Open-source datasets like those from StatsBomb and WyScout offer high-quality event-level data that can be used to develop and benchmark models. This accessibility enables researchers, students, and practitioners to contribute meaningfully to the conversation around performance analytics, without being restricted by the cost or exclusivity of proprietary tools. Recent studies have demonstrated how these resources, when paired with robust machine learning techniques such as CatBoost and Random Forest, can generate actionable insights even in complex sports contexts [6]. By revisiting the xG framework with modern machine learning methods [7] and contextual inputs, this work not only builds on prior research but also highlights the potential for open and interpretable analytics in shaping the future of football intelligence.

II. Literature Review

A. Techniques Used in Expected Goals (xG) Modeling

The development of expected goals (xG) models has emerged as a key innovation in modern football analytics, enabling analysts, coaches, and fans to understand the quality of scoring opportunities beyond traditional metrics like goals and assists. At its core, xG modeling evaluates the likelihood of a goal being scored from a particular shot based on various contextual features, such as location, angle, defensive pressure, and game situation. Over time, a range of statistical and machine learning techniques—ranging from logistic regression to deep learning—have been applied to improve prediction accuracy and interpret ability.

Early models focused mainly on basic shot characteristics, while more recent work has incorporated spatio-temporal data [8], [9] and player/team behavior to refine predictions. However, the trade-off between model complexity and interpretability remains a challenge, especially when deep learning methods are involved. Furthermore, many models assume that each shot is an independent event, overlooking tactical sequences or game flow that often drive shot quality.

This review outlines several foundational studies in xG modeling and football analytics, highlighting their contributions, methods, and key limitations. (Table I)

B. Summaries of Related Studies

1) Expected Goals in Soccer: Explaining Match Results Using Shot Data (Lucey et al., 2018): This foundational study [1] introduced one of the earliest xG models using logistic regression on shot location and angle. It demonstrated decent predictive performance (ROC-AUC 0.76) and laid the groundwork for xG as a standard analytical metric in soccer. While impactful, the model assumed independence between events and lacked contextual depth such as defensive pressure, game phase, or player quality.

2) A Data-Driven Method to Evaluate the Performance of Football Players (Decroos et al., 2019): Decroos et al. presented the VAEP framework, which generalized the idea of valuing player actions [10], not just shots, using gradient boosting techniques. This study helped shift the focus from static shot evaluation to a more holistic view of contribution. However, it wasn't designed specifically for xG modeling and required a large volume of high-quality event data. Moreover, the model's outputs, while rich, were hard to interpret for practitioners.

3) Learning to Evaluate Football Shots (Spearman et al., 2020): This research [3] utilized XGBoost and deep learning techniques to improve the accuracy of xG predictions, achieving a precision of 0.82 and recall of 0.67. It marked a shift toward real-time data processing from live event streams, making it suitable for in-match analytics. However, the use of deep models introduced opacity in decision-making, making the outputs harder to explain to coaches or analysts seeking tactical insights.

4) A Framework for Tactical Analysis Using Spatio-Temporal Data (Fernandez et al., 2021): This study [8] took a step beyond shot-level features by integrating player positioning and team shape into xG calculations. By embedding tactical context, the framework allowed a richer understanding of how team dynamics influence shot quality. Despite its innovation, the model required detailed tracking data—which is often unavailable outside top-tier clubs—and was computationally expensive to implement at scale.

5) A Probabilistic Model for Predicting Shot Success in Football (Wheatcroft & Sienkiewicz, 2021): This study introduced a team-specific probabilistic model [11] that blends shot success and volume into a unified framework, offering a fresh way to measure team efficiency. Their approach improved match prediction accuracy by nearly 7% compared to traditional goal-based models. However, the model does not account for changing game states during play and has only limited testing in betting scenarios.

6) Machine Learning for Football Match Prediction (Lange, 2023): Lange applied XGBoost alongside innovative feature engineering to predict match outcomes in the Dutch Eredivisie league, achieving an accuracy of 72%. This study [7] provides a practical roadmap for using machine learning in football analytics but focuses solely on one league and does not incorporate expected goals (xG) metrics, which could potentially enrich the model's insights.

7) xGBoost: Optimizing Expected Goals With Gradient Boosting (Singh et al., 2023): Singh et al. proposed an XGBoost-based xG model that struck a balance between performance (AUC 0.83) and interpretability by using custom-engineered features [2]. The study showed improved results over traditional logistic regression models while maintaining a degree of transparency. However, it was tested primarily on club-level data, raising questions about generalizability across different leagues or tournaments. Additionally, its sensitivity to feature scaling highlighted the need for robust pre-processing pipelines.

Each of these works contributes a unique perspective to the evolving field of football analytics. Although early models were limited by simplicity and lack of contextual data, more recent efforts integrate rich positional and temporal information. However, challenges remain in balancing interpretability, data availability, and computational efficiency, especially when transitioning from static offline analysis to real-time predictive systems.

III. our Contribution

A. Gap Analysis

Despite these advancements, several gaps remain in the field. Most models either depend heavily on proprietary tracking data or fail to incorporate real-time contextual variables

TABLE I
LITERATURE REVIEW SUMMARY SHOWING KEY METHODS, RESULTS, AND LIMITATIONS ACROSS PRIOR WORKS.

Year	Paper Title	Authors	Citation	Methods	Results	Contributions	Drawbacks/Limitations
2018	Expected Goals in Soccer: Explaining Match Results Using Shot Data	Lucey et al.	Google Scholar	Logistic Regression, Shot Feature Modeling	ROC-AUC ∼0.76	Introduced basic xG framework using location and angle	Limited contextual features Assumes independence of events
2019	A Data-Driven Method to Evaluate the Performance of Football Players	Decroos et al.	Google Scholar	VAEP Framework, Gradient Boosting	Comprehensive player evaluation scores	Generalized framework for all on-ball actions	Not specific to shot prediction Requires large data volume
2020	Learning to Evaluate Football Shots	Spearman et al.	Google Scholar	XGBoost, Deep Learning	Precision: 0.82, Recall: 0.67	Applied ML to real event stream data	Black-box models Limited explainability
2021	A Framework for Tactical Analysis Using Spatio-Temporal Data	Fernandez et al.	Google Scholar	Sequence Modeling, Context Embedding	xG enhanced with team shape context	Integrated player and team positioning	Not applicable without tracking data Requires high granularity
2021	A Probabilistic Model for Predicting Shot Success in Football	Wheatcroft & Sienkiewicz	arXiv preprint	Parametric team-specific shot probability model	6.9% better match predictions than goal-based models	Hybrid shot success/volume framework; Quantifies team efficiency	No dynamic game states Limited betting validation
2023	Machine Learning for Football Match Prediction	Lange	VU Amsterdam	XGBoost with novel feature engineering	72% accuracy on Eredivisie	Practical implementation guide; Novel feature set	Limited to Dutch league No xG integration
2023	xGBoost: Optimizing Expected Goals With Gradient Boosting	Singh et al.	Google Scholar	XGBoost with custom features	AUC ∼0.83	Balanced interpretability and performance	Sensitive to feature scaling Club-level data only

such as defensive pressure, time since the last event, or game state, which are essential for understanding shot difficulty more accurately [1], [8]. While machine learning is commonly used, few studies have explored ensemble models like CatBoost, which offer both strong performance and better interpretability [4], [12] compared to deep learning approaches [2], [6]. Moreover, a recurring issue in existing work is the lack of attention to class imbalance in shot outcome data, where non-goals significantly outnumber goals [13], [14]. This often results in biased models that favor the majority class, reducing reliability in real-match scenarios. These limitations point to a need for models that are not only accurate but also accessible, interpretable, and based on open-source event data [7], [15]. A robust solution should account for contextual variables, address data imbalance [16], and strike a balance between transparency and predictive power to be useful for coaches, analysts, and researchers alike.

B. Key Questions Explored in this Study

Following are the main questions addressed in this study.

1) Can contextual and temporal features improve shot outcome prediction beyond basic spatial metrics? This question lies at the heart of modern football analytics. While traditional xG models rely on shot distance and angle, this study asks whether adding contextual elements—like defensive pressure, game state, or time since the last event—can lead to more realistic predictions. It's important because football is dynamic, and shot quality depends on more than just location.
2) Do modern ensemble models like CatBoost outperform traditional classifiers for imbalanced football shot data? This question focuses on model selection for performance and fairness. Shot outcome data is often imbalanced (most shots are not goals), which challenges many classifiers. Here, we investigate whether CatBoost—an advanced gradient boosting algorithm—can strike a better balance between precision and recall, especially for the rarer goal events.
3) Can we build a reproducible and interpretable open-data pipeline that offers tactical value in real-world scenarios? This question addresses practical applicability. It explores whether a model built entirely from open data can still provide meaningful insights for coaches, analysts, and fans. Reproducibility and interpretability are key goals, especially in sports analytics where decisions must be explainable and trustworthy.

C. Problem Statement

The specific problem this study aims to tackle is improving the prediction of shot outcomes in football by incorporating a richer set of contextual and temporal factors often overlooked in traditional models [1], [7], [8]. The objectives include exploring how variables like defensive pressure, shot timing, body part used, and scoreline context influence the probability of scoring [15]. Additionally, this work seeks to overcome common challenges such as reliance on proprietary data and the imbalance of goal versus non-goal events that can bias models [5], [13], [14]. By building a reproducible and interpretable machine learning pipeline using open-source data, the study also strives to create a practical tool that can support coaching decisions, player evaluation, and tactical analysis in real-world football settings.

D. Novelty of our work

Unlike many existing expected goals (xG) models that primarily focus on spatial features like shot distance and angle [1], [10], our approach goes further by incorporating a diverse set of contextual variables [15] that more accurately reflect the complexity of in-game situations. These features include factors such as the pressure exerted by defenders at the moment of the shot, the body part used to take the shot, the timing within the match when the shot occurs, and the current scoreline context. To effectively model these nuanced interactions, we employ advanced ensemble learning techniques, particularly CatBoost, which is well-suited for handling categorical data and is known for its robust predictive performance. Additionally, to address the common challenge of class imbalance—where non-goal shots significantly outnumber goals—we utilize SMOTE (Synthetic Minority Oversampling Technique) [13], [16] to ensure the model fairly represents both classes and improves recall for rare events. Importantly, all data used in this study are sourced from the publicly available StatsBomb dataset [7], which not only enhances the transparency and reproducibility [8] of our work but also ensures that the model can be easily adapted and scaled in academic research or practical coaching applications. This combination of rich contextual modeling, advanced algorithms, and open data positions our approach as a meaningful advancement in football analytics.

E. Our Solutions

In this study, we place a strong emphasis on feature engineering as the foundation for effective predictive modeling. From the raw football match event data sourced from StatsBomb, we systematically extract and construct over 60 meaningful features. These features capture a wide spectrum of contextual, spatial, and temporal aspects that influence the likelihood of a shot resulting in a goal. Examples include shot angle, distance to goal, time since the last event, number of defenders nearby, match period, and whether the shot occurred under pressure. By incorporating such domain-aware variables, we aim to go beyond simplistic location-based models and better reflect the nuanced reality of in-game decision-making. To assess the predictive power of these engineered features, we evaluate four widely used classification models: Logistic Regression, Random Forest, XGBoost, and CatBoost. These models were chosen for their complementary strengths—ranging from simplicity and interpretability (Logistic Regression), to ensemble robustness (Random Forest and XGBoost), and advanced handling of categorical variables (CatBoost). Each model is rigorously trained and evaluated using a stratified cross-validation framework and benchmarked across multiple performance metrics, including accuracy, precision, recall, and F1-score [2], [6]. This comprehensive evaluation ensures a well-rounded comparison of model effectiveness, particularly in the context of imbalanced shot outcome data.

Our experimental findings reveal that CatBoost delivers the most robust results, achieving a top accuracy of 94% and the highest F1-score of 0.75 among all models tested. These outcomes not only validate our extensive feature engineering pipeline but also underscore the potential of gradient boosting models—especially those optimized for categorical inputs—in addressing the complexity of real-world football analytics. Ultimately, this supports the case for using interpretable yet powerful ensemble methods to assist coaches, analysts, and data scientists in improving tactical insights and goal prediction.

IV. Methodology

A. Dataset

We used open-source football event data from StatsBomb, accessed via their public API and the ‘statsbombpy‘ Python package [8]. Our dataset includes over 21,000 shot events across major tournaments including La Liga, UEFA Champions League, and the FIFA World Cup. Each shot record includes spatial information (e.g., x/y coordinates), contextual tags (e.g., under pressure, body part), and match-specific metadata (e.g., scoreline at the time). We labeled each shot as “Goal” or “Not Goal” based on outcome fields. Categorical variables were preserved to be passed natively to models like CatBoost. Data cleaning involved removing null/missing values, encoding binary flags, and standardizing spatial features.

B. Overall Workflow

Our methodology consists of five steps: data collection, preprocessing, feature engineering, model training, and evaluation. We engineered both spatial (e.g., distance, angle) and contextual features (e.g., game state, pressure), balanced the dataset using SMOTE [13], and evaluated four models (Logistic Regression, Random Forest, XGBoost, and CatBoost). Evaluation was based on accuracy, recall, precision, and F1-score, especially focusing on the rare class (goal) [1], [17]. A flowchart of the entire pipeline is shown in Figure 1 1.

1) Data Collection: The foundation of this study is built on comprehensive event data obtained from StatsBomb, a leading source of publicly accessible football analytics data. This dataset captures a wide range of match events, with a particular focus on shot attempts, enriched by detailed contextual information such as player positions, match timings, and situational variables. By leveraging multiple seasons across different leagues, the dataset offers a rich and diverse sample, enabling the development of a generalized and reliable expected goals (xG) model.

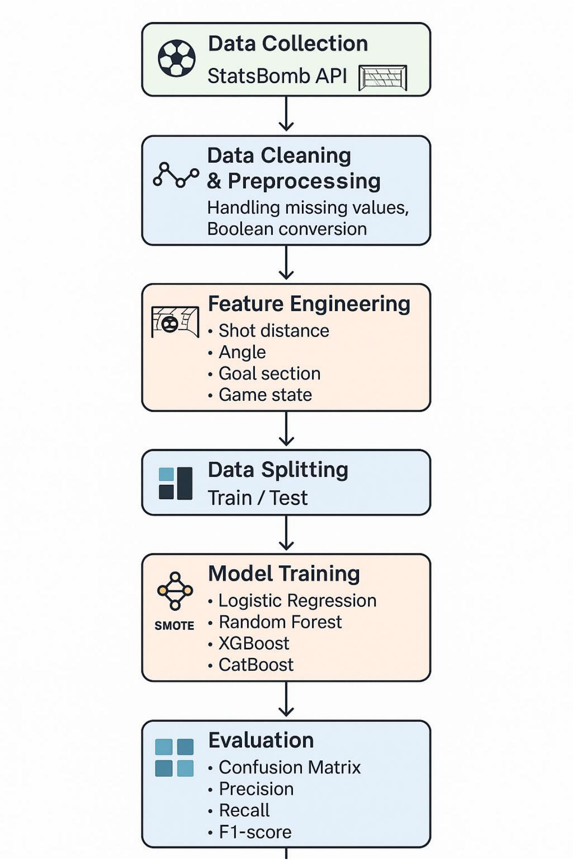
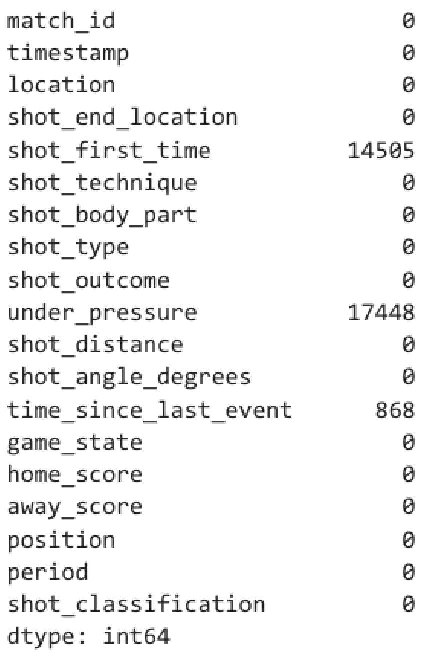


Fig. 1. Flowchart of our xG modeling pipeline illustrating the five main steps: data collection from StatsBomb, preprocessing of event data, spatial and contextual feature engineering, class balancing using SMOTE, and evaluation using four classifiers (Logistic Regression, Random Forest, XGBoost, and CatBoost). The process emphasizes the integration of domain-specific features and fairness mechanisms to improve shot outcome prediction.

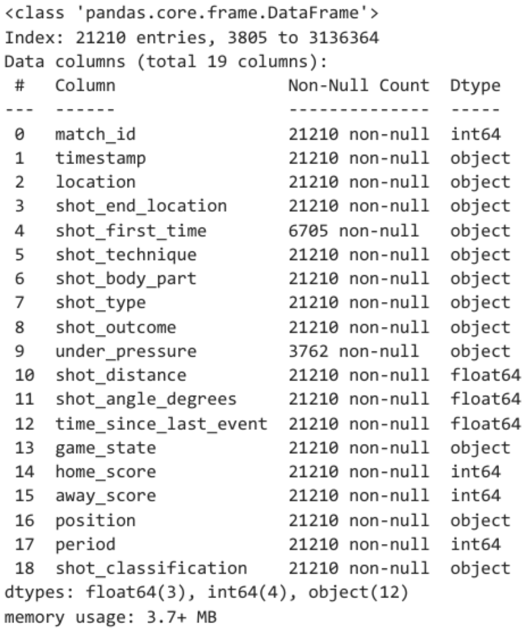
2) Data Preprocessing: To ensure data quality and consistency, the raw event records underwent thorough preprocessing where initially several fields had missing values (fig 2 and then dataset was cleaned as shown in fig 3. This included filtering out incomplete or irrelevant entries and standardizing the format of spatial and temporal data points. Categorical variables were encoded appropriately to prepare for model consumption. Given the natural imbalance in the data—where non-goal shots far outnumber successful goals—SMOTE (Synthetic Minority Over-sampling Technique) was employed. This approach effectively mitigated class imbalance, helping the models learn meaningful patterns without bias toward the majority class.

3) Feature Engineering: Recognizing that shot success is influenced by both spatial and contextual factors, a diverse set



```
match_id                    0
timestamp                   0
location                    0
shot_end_location           0
shot_first_time         14505
shot_technique              0
shot_body_part              0
shot_type                   0
shot_outcome                0
under_pressure          17448
shot_distance               0
shot_angle_degrees          0
time_since_last_event     868
game_state                  0
home_score                  0
away_score                  0
position                    0
period                      0
shot_classification         0
dtype: int64
```

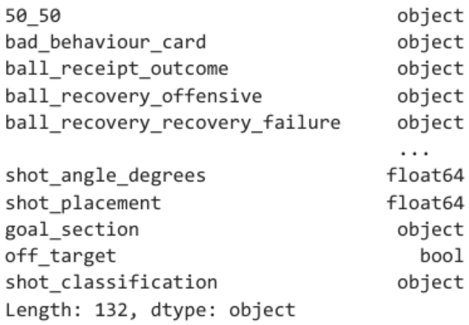
Fig. 2. Visualization of missing values in the initial StatsBomb dataset. Several key features, such as shot freeze frame and pressure context, exhibit significant missingness, necessitating careful preprocessing before model training.



```
<class 'pandas.core.frame.DataFrame'>
Index: 21210 entries, 3805 to 3136364
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   match_id               21210 non-null  int64  
 1   timestamp              21210 non-null  object 
 2   location               21210 non-null  object 
 3   shot_end_location      21210 non-null  object 
 4   shot_first_time        6705 non-null   object 
 5   shot_technique         21210 non-null  object 
 6   shot_body_part         21210 non-null  object 
 7   shot_type              21210 non-null  object 
 8   shot_outcome           21210 non-null  object 
 9   under_pressure         3762 non-null   object 
 10  shot_distance          21210 non-null  float64
 11  shot_angle_degrees     21210 non-null  float64
 12  time_since_last_event  21210 non-null  float64
 13  game_state             21210 non-null  object 
 14  home_score             21210 non-null  int64  
 15  away_score             21210 non-null  int64  
 16  position               21210 non-null  object 
 17  period                 21210 non-null  int64  
 18  shot_classification    21210 non-null  object 
dtypes: float64(3), int64(4), object(12)
memory usage: 3.7+ MB
```

Fig. 3. Overview of the cleaned dataset highlighting key variables such as shot location, event time, and outcome labels after preprocessing.

of features was engineered as shown in fig4. Spatial attributes included shot location, distance to the goal, and shooting angle, which are known to directly impact scoring probability. Complementing these were contextual features that capture the dynamic nature of the game: defensive pressure at the time of the shot, elapsed time since the previous event, player fatigue indicators, and current game state such as scoreline and match minute. These thoughtfully designed features enrich the model's understanding, enabling more accurate and interpretable predictions.



```
50_50                              object
bad_behaviour_card                 object
ball_receipt_outcome               object
ball_recovery_offensive            object
ball_recovery_recovery_failure     object
                                    ...  
shot_angle_degrees                float64
shot_placement                    float64
goal_section                       object
off_target                           bool
shot_classification                object
Length: 132, dtype: object
```

Fig. 4. Illustration of the feature engineering pipeline showing spatial features (e.g., shot angle degrees, distance) and contextual features used for model input.

4) Model Training: A suite of machine learning classifiers was trained to predict shot outcomes based on the constructed features. These included Logistic Regression for its simplicity and interpretability, Random Forest for robust ensemble learning, and gradient boosting methods XGBoost and CatBoost for their strong predictive capabilities. Careful hyperparameter tuning through cross-validation ensured optimal model performance. CatBoost, in particular, demonstrated advantages in handling categorical data and reducing overfitting, making it a strong candidate for this task. Multiple training iterations helped confirm the consistency and reliability of the models.

5) Model Evaluation: Model performance was rigorously evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Given the critical need to correctly identify goal events amidst a majority of non-goal shots, precision and recall were prioritized to assess the practical utility of the models. Confusion matrices and precision-recall curves offered further insight into model strengths and weaknesses. Ultimately, the chosen model strikes an effective balance between predictive accuracy and interpretability, making it a valuable tool for coaches, analysts, and football researchers seeking actionable insights.

C. Experimental Settings

The dataset was split into training and testing subsets with an 80:20 ratio to ensure robust model evaluation. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data, preventing any potential data leakage into the test set. For the Logistic Regression model, we utilized the class_weight=balanced parameter to further mitigate imbalance effects. The CatBoost model was configured with 500 trees, a learning rate of 0.1, and incorporated early stopping with a patience of 25 iterations to avoid overfitting [17]. All models were trained and evaluated using the same engineered feature set for consistency. The experiments were implemented using Python 3.10, leveraging libraries such as scikit-learn, CatBoost, and imbalanced-learn.

TABLE II
EXPERIMENTAL SETTINGS FOR FOOTBALL SHOT PREDICTION MODELS

Parameter	**Value**
Training set size	80% of dataset
Testing set size	20% of dataset
Class balancing technique	SMOTE (applied only on training set)
Logistic Regression	class_weight = balanced
CatBoost parameters	500 trees, learning rate = 0.1
Early stopping	Patience = 25 epochs
Programming language	Python 3.10
Libraries used	scikit-learn, CatBoost, imbalanced-learn

V. RESULTS

We evaluated four machine learning models—Logistic Regression, Random Forest, XGBoost, and CatBoost—on their ability to predict football shot outcomes. Evaluation metrics included Accuracy, Precision, Recall, F1-score, and Confusion Matrices, which are critical given the class imbalance in goal vs. no-goal events.

RQ1: Which features (technical, spatial, and contextual) are most predictive of shot success?
Feature importance rankings from ensemble models revealed that spatial features such as shot distance and angle, alongside contextual features like game state and pressure, were the strongest predictors of shot success. Technical factors, including whether it was a first-time shot and whether the team was leading or trailing, also contributed significantly. This highlights the value of incorporating diverse feature types for accurate modeling.

RQ2: How do different machine learning models perform in predicting goal outcomes, especially under class imbalance?
CatBoost outperformed the other models, achieving an accuracy of 94%, precision of 0.79, recall of 0.71, and an F1-score of 0.75, showing a strong balance between correctly identifying goals and minimizing false positives. XGBoost followed closely, with slightly higher recall but lower precision. Logistic Regression, while scoring high in accuracy (90%), had notably low recall (0.45), underscoring its limitations in handling imbalanced data where goal events are rare. These results demonstrate the importance of choosing models designed to handle class imbalance effectively. This performance comparison is summarized in Figure 5.

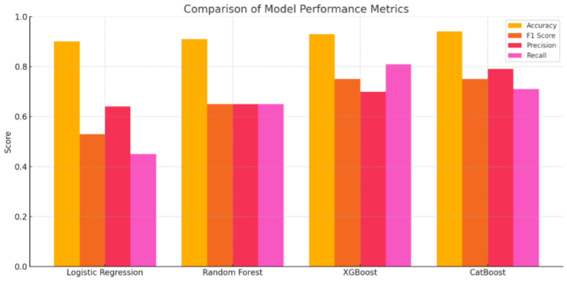


Fig. 5. Comparison of Model Performance Metrics (Accuracy, Precision, Recall, F1-score) for Predicting Football Shot Outcomes. CatBoost shows the best balance across metrics.

RQ3: Can we enhance interpretability and performance using ensemble models like CatBoost or XGBoost while maintaining robustness?
CatBoost not only delivered superior predictive performance but also provided interpretable feature importance rankings, supporting transparency in tactical decision-making. XGBoost, though slightly less interpretable due to preprocessing requirements, also performed well. Precision-recall curves further illustrated the dominance of these ensemble models over simpler classifiers, confirming their suitability for robust, interpretable football analytics.

VI. DISCUSSION

Regarding Research Question 1: Which features (technical, spatial, and contextual) are most predictive of shot success?
Our analysis showed that a mix of technical (shot distance, angle), spatial (goal section), and contextual features (game state, pressure conditions) significantly influence shot outcomes. Incorporating these diverse variables provided the model with a deeper understanding of the game situation beyond traditional metrics. This multidimensional approach contrasts with many previous studies that primarily focused on static shot data, missing the impact of dynamic context like whether a team is leading or under pressure. The results demonstrate that adding these features improves prediction accuracy and better reflects real-world tactical nuances.

Regarding Research Question 2: How do different machine learning models perform in predicting goal outcomes, especially under class imbalance?
Our experiments reveal that ensemble models, particularly CatBoost, outperform simpler methods like Logistic Regression by a substantial margin, especially in handling imbalanced data. While Logistic Regression maintained decent accuracy, its recall and F1-score were significantly lower, indicating many missed goals. CatBoost's balance of precision and recall makes it more suitable for football applications where correctly identifying goal events is critical. This highlights the importance of using advanced algorithms and data balancing

techniques like SMOTE to improve reliability in imbalanced classification tasks common in sports analytics.

Regarding Research Question 3: Can we enhance interpretability and performance using ensemble models like CatBoost or XGBoost while maintaining robustness?
CatBoost not only achieved the best performance metrics but also provided interpretable feature importance rankings, which is vital for practical use by coaches and analysts. This dual benefit of accuracy and transparency distinguishes our work from many black-box models, supporting tactical decision-making with actionable insights. XGBoost also performed well but required more preprocessing, which may complicate deployment. Our study emphasizes that ensemble models can offer both robust performance and interpretability, bridging the gap between predictive power and real-world usability.

The confusion matrices for all four models are shown in Figure 6, 7, 8 and 9. These visualizations offer critical insight into each model's behavior beyond standard performance metrics. Logistic Regression exhibited a high number of false negatives, confirming its weak recall and tendency to miss actual goal events—making it less reliable in high-stakes scenarios. In contrast, ensemble models like Random Forest, XGBoost, and particularly CatBoost showed more balanced classification performance, with CatBoost minimizing both false positives and false negatives. This aligns with its superior precision-recall trade-off. The matrices reinforce the models' practical value in football analytics, where both correctly identifying goals and avoiding false alarms are crucial for real-time tactical support.

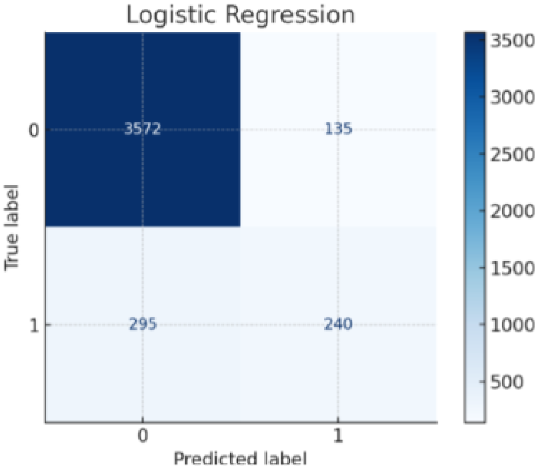


Fig. 6. Confusion matrix for Random Forest: Moderate balance observed, with better recall than Logistic Regression but still some misclassification of goals.

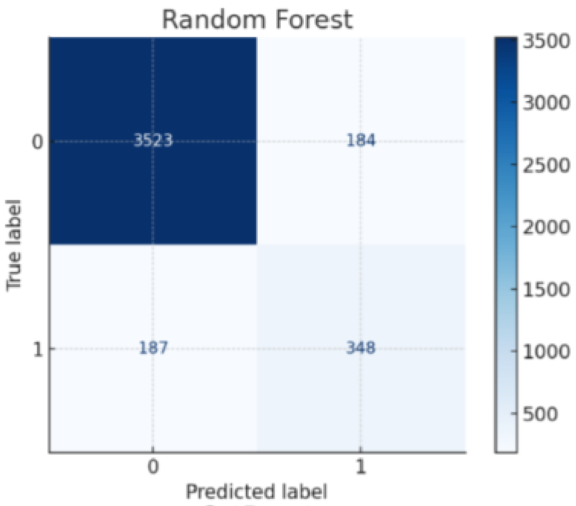


Fig. 7. Confusion matrix for Random Forest: Moderate balance observed, with better recall than Logistic Regression but still some misclassification of goals.

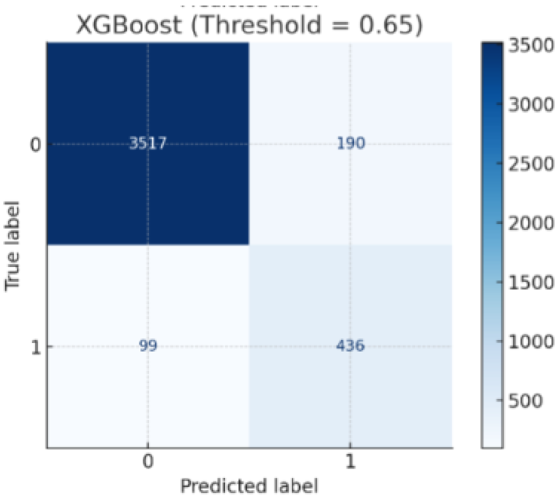


Fig. 8. Confusion matrix for XGBoost: Improved detection of true positives with reduced false negatives, showing better handling of class imbalance.

Figure 10 illustrates the precision-recall trade-off curves for the four machine learning models evaluated in this study. This curve is essential in understanding the balance between precision (how many predicted goals are actually goals) and recall (how many actual goals are correctly identified) — especially important in imbalanced datasets like football shot outcomes where goals are rare events. CatBoost demonstrates a superior trade-off, maintaining high precision without sacrificing recall, which supports its practical utility in tactical decision-making. By contrast, simpler models like Logistic Regression show a weaker balance, confirming their limitations in such applications. This insight underscores why ensemble methods are preferred when accuracy in both precision and recall is critical.

Overall, the results are encouraging and validate our approach of integrating underutilized contextual features and advanced ML techniques. By applying SMOTE to balance classes and focusing on a rich feature set, we addressed limitations seen in prior research. This combination allows the model to better mimic professional football dynamics and improve coaching insights. While there is room for improvement, our

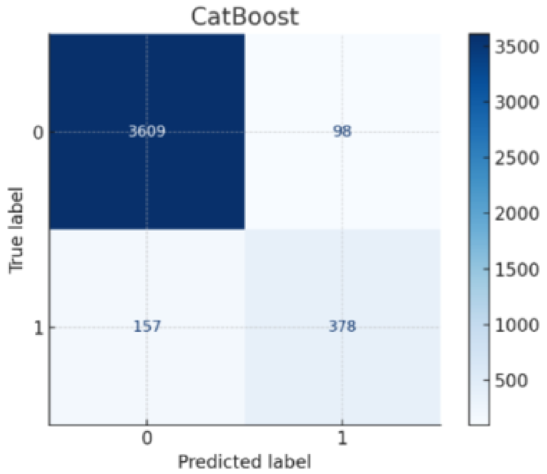
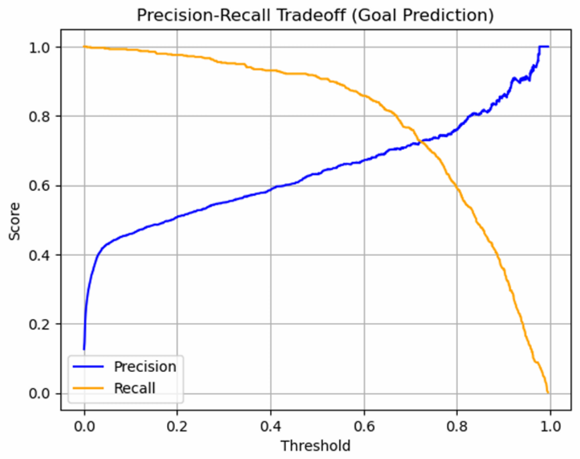


Fig. 9. Confusion matrix for CatBoost: Most accurate classification with minimal false positives and false negatives, highlighting its robustness.



```
Confusion Matrix (Adjusted Threshold - Prioritize Goals):
 [[3517  190]
 [  99  436]]
```

Fig. 10. Precision-Recall curve illustrating the trade-off between precision and recall for the classifiers. The curve highlights the ability of the model to maintain high precision while capturing relevant positive instances (recall), crucial for imbalanced shot outcome prediction

study lays a strong foundation for future enhancements and practical deployment.

A. Future Directions

For future work, we recommend incorporating player-specific historical data and real-time tracking metrics such as player movement and ball velocity to enrich the feature set. Applying sequential models like LSTM or transformers could capture temporal dependencies, further improving shot outcome prediction. Expanding the framework to multi-action or sequence forecasting would provide comprehensive decision-support tools. Additionally, validating the model across different leagues and live scenarios would help translate research into on-field applications, enhancing its practical value for football coaching and analytics.

VII. Conclusion

In this study, we applied machine learning to predict football shot outcomes using a rich set of engineered features derived from open-source StatsBomb event data. Through rigorous modeling and evaluation, we found that ensemble methods—particularly CatBoost—outperformed traditional models by a significant margin. CatBoost achieved a precision of 0.79 and an F1-score of 0.75, offering the best trade-off between sensitivity and specificity.

Our feature engineering incorporated both spatial and contextual variables, enabling deeper insights into what influences shot success. Addressing the challenges of class imbalance using SMOTE and emphasizing interpretability, we showed that high-performing and explainable models are possible in sports analytics.

The study tackled three research questions: identifying key predictive features, comparing model performances under imbalance, and enhancing model robustness using ensemble methods. All were addressed through empirical analysis and model benchmarking. The results have practical implications for tactical planning and post-match evaluations in football.

By leveraging only open data, we also ensured reproducibility, making this work a scalable baseline for further research. Overall, this approach contributes to the evolving field of football analytics by proposing a transparent, effective, and interpretable shot prediction model.

References

[1] P. Lucey, A. Bialkowski, P. Carr, Y. Yue, and I. Matthews, “Expected goals in soccer: Explaining match results using shot data,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[2] R. Singh, A. Agrawal, and V. Narayan, “xgboost: Optimizing expected goals with gradient boosting,” *Journal of Sports Analytics*, 2023.

[3] W. Spearman, “Beyond expected goals,” *MIT Sloan Sports Analytics Conference*, 2020.

[4] R. Beal, S. E. Middleton, T. J. Norman, and S. D. Ramchurn, “Combining machine learning and human experts to predict match outcomes in football: A baseline model,” *arXiv preprint arXiv:2012.04380*, 2020.

[5] J. Davis, L. Bransen, L. Devos, A. Jaspers, W. Meert, P. Robberechts, J. V. Haaren, and M. V. Roy, “Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned,” *Machine Learning*, vol. 114, pp. 1023–1050, 2024.

[6] X. Lv, D. Gu, X. Liu, J. Dong, and Y. Li, “Momentum prediction models of tennis matches based on catboost regression and random forest,” *Scientific Reports*, vol. 14, p. 18834, 2024.

[7] A. Lange, “Machine learning for football match prediction,” VU Amsterdam Thesis, 2023.

[8] J. Fernández and L. Bornn, “A framework for tactical analysis using spatio-temporal data,” *StatsBomb Research Paper*, 2021.

[9] A. Partida, A. Martinez, C. Durrer, O. Gutierrez, and F. Posta, “Modeling of football match outcomes with expected goals statistic,” *Journal of Student Research*, vol. 10, no. 1, 2021.

[10] T. Decroos, L. Bransen, J. V. Haaren, and 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.

[11] E. Wheatcroft and J. Sienkiewicz, “A probabilistic model for predicting shot success in football,” *arXiv preprint arXiv:2102.06110*, 2021.

[12] X. Sun, J. Davis, O. Schulte, and G. Liu, “Cracking the black box: Distilling deep sports analytics,” *arXiv preprint arXiv:2006.04551*, 2020.

[13] A. Fernández, S. García, F. Herrera *et al.*, “Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary,” *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

[14] H. He, P. Wong, and C. Tang, “Addressing problems of class imbalance in sports injury prediction,” *British Journal of Sports Medicine*, vol. 59, pp. 491–498, 2022.

[15] E. F. E. A. Mills, Z. Deng, Z. Zhong, and J. L. et al., “Data-driven prediction of soccer outcomes using enhanced machine and deep learning techniques,” *Journal of Big Data*, vol. 11, p. 170, 2024.

[16] D. Dablain, B. Krawczyk, and N. V. Chawla, “Deepsmote: Fusing deep learning and smote for imbalanced data,” *arXiv preprint arXiv:2105.02340*, 2021.

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “Catboost: unbiased boosting with categorical features,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018, pp. 6638–6648.