OBJECTIVE

The primary purpose of this model is to **classify corner kicks by threat level** using **CatBoost**, a gradient boosting algorithm optimized for categorical data.

In this study, a "threatening" corner is defined based on its expected goals (xG) value within 20 seconds after the corner kick. The model classifies each play as either "High-Threat" or "Low-Threat", providing insights into the tactical and contextual factors that increase the likelihood of a goal-scoring opportunity.

Why this model

Traditional xG models estimate the continuous probability of scoring a goal, but the aim here is different. Instead of predicting a numeric xG value, we adopt a binary thresholding approach that identifies which corners fall within the top quartile of threat generation. This formulation offers a more interpretable distinction between dangerous and less dangerous situations, facilitating more precise tactical analysis and decision-making.

DATA AND PREPROCESSING

The preprocessing phase ensures that only valid, contextually rich plays are analyzed and that the model's inputs are clean, consistent, and free of information leakage.

1. **Event filtering**
   Only events containing Freeze Frame data were included, both at P0 (the corner event) and P1 (the subsequent play).
   Any sequence missing a freeze-frame at either stage was excluded from the analysis, as the absence of positional data prevents reliable computation of spatial and contextual features.
2. **Input Data**
   The dataset included multiple contextual and spatial variables describing each corner event. These features captured game context, temporal dynamics, passing characteristics, goalkeeper positioning, and player density during both the delivery phase (P0) and the subsequent play (P1). Together with the expected goals value recorded within 20 seconds after the corner (xg_20s), these variables provide a comprehensive representation of team behavior, player positioning, and event outcomes. You can refer to the **data dictionary** for detailed documentation of all variables used in this study.

3. **Excluded Data**
   To avoid redundancy, reduce leakage, and focus on interpretable predictors, several columns were excluded before model training.

   These fall into three main categories:

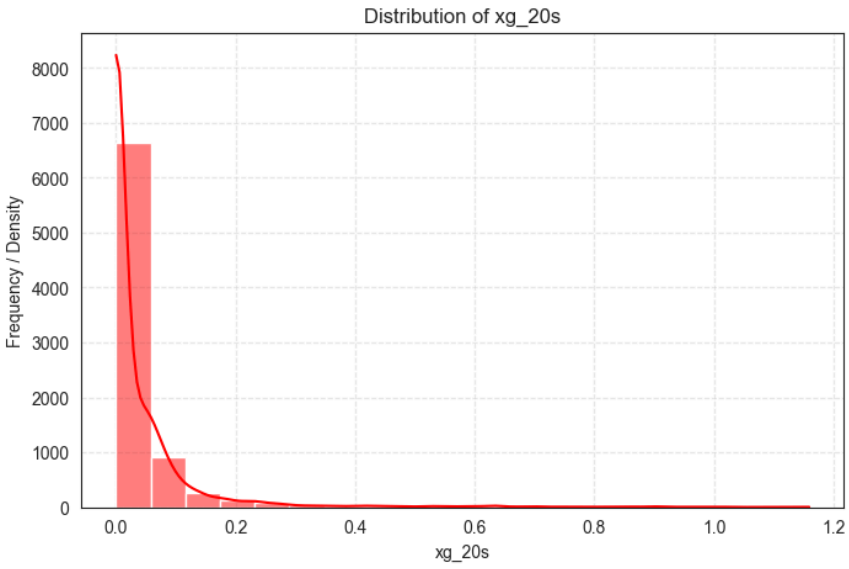
- Identifiers and metadata: *match_id, event_id, minute, second, period, match_date, home_team, away_team, player, team, recipient.*
- Variables directly related to the target or to subsequent outcomes: *goal_20s, goal_20s_def, xg_20s_def, P1_event_id, P1_timestamp, P1_index, P1_type, P1_team.*
- Positional or redundant features not used for inference: *P0_index_x, P0_index_y, P0_index, P1_coordinates_normalized, all zone_name columns.*

The final feature space, therefore, includes only the variables that provide relevant tactical, spatial, or contextual information to help distinguish between High-Threat and Low-Threat corners.

TARGET SELECTION

For every corner event (denoted as P0), the algorithm identifies all subsequent events (P1, P2, …) occurring within 20 seconds of the corner's timestamp. It aggregates the expected goals (xG) values generated by the attacking team during that period. If no shots are recorded within the window, the xG values default to zero. This consistent temporal framework enables an objective, comparable assessment of each corner's effectiveness in creating goal-scoring opportunities.

Percentile	Value
25th	0.00
50th	0.00
75th	0.0435
90th	0.0933
95th	0.1596
99th	0.521
100th	1.1591



The xg_20s distribution is highly skewed, with most corners generating no threat (median = 0). Only a few plays produce high xG_20s values, resulting in a long right tail. Therefore, the 75th percentile (0.0435) was chosen as the threshold to capture the top quartile of genuinely dangerous corners while preserving enough positive samples for model training.

$$\text{target} = \begin{cases} 1, & \text{if } xg_{20s} > 0.0435 \text{ (High-Threat)} \\ 0, & \text{if } xg_{20s} \leq 0.0435 \text{ (Low-Threat)} \end{cases}$$

MODEL CONFIGURATION AND TRAINING

1. Data division (train/test)

Data from seasons 2021/2022 to 2024/2025 were used to train and evaluate the model.

This broader range allows the model to learn from multiple seasons, capturing both stable patterns and new tactical trends in corner kick execution.

Including recent data helps the model generalize better to evolving styles of play while still preserving temporal consistency and avoiding data leakage between training and evaluation phases.

2. Configuration

The model was implemented using CatBoostClassifier, a gradient-boosting algorithm optimized for tabular data and categorical features.

The following parameters remained consistent across all experiments:

Parameter	Description	Value
loss_function	Objective for binary classification	Logloss
eval_metric	Evaluation metric	AUC
iterations	Maximum boosting rounds	2000
od_type	Early stopping method	Iter
od_wait	Early stopping patience	200
auto_class_weights	Automatic class balancing	SqrtBalanced
random_seed	Randomization Control	42

Additional considerations:

- Class imbalance correction was applied automatically using the square-root weighting strategy, giving higher weight to the minority “High-Threat” class.
- Categorical features were automatically detected and handled internally by CatBoost, avoiding the need for manual encoding.

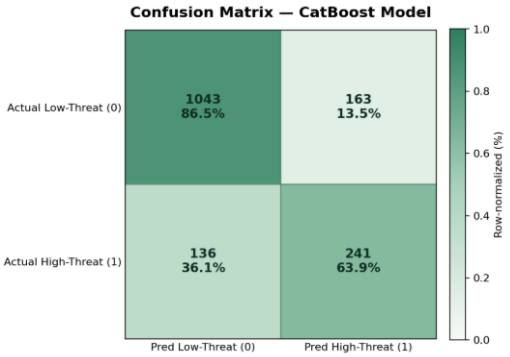
- The validation metric (AUC) guided early stopping and model selection

3. **Baseline model**

   The baseline model used the initial CatBoost configuration without any hyperparameter optimization. It served as a reference to evaluate the effect of tuning on model performance. The chosen parameters were:

Parameter	Description	Value
depth	Maximum depth of each decision tree, controlling model complexity.	6
learning_rate	Step size used to update weights during boosting.	0.05
l2_leaf_reg	L2 regularization coefficient.	3.0
border_count	Number of discrete splits to bin continuous features.	254
random_strength	Degree of randomness added to leaf value calculation.	1
bagging_temperature	Controls sampling diversity of data	1

The baseline CatBoost model achieved an AUC of 0.82 and an average precision of 0.69, indicating strong discriminative ability between high-threat and low-threat corners.
The F1-score of 0.61 reflects a reasonable trade-off between precision and recall, with precision = 0.59 and recall = 0.63 for the high-threat class. This suggests that the model can correctly identify a meaningful portion of dangerous corners, though it remains conservative in classifying them as high-risk.



Metric	Low-Threat	High-Threat
precision	0.88	0.59
recall	0.86	0.63
f1-score	0.87	0.61
support	1206	377

The confusion matrix shows that most low-threat corners were correctly classified, while the model captured a fair proportion of high-threat plays, despite the natural class imbalance (only 25% positives).

Overall, the baseline configuration demonstrated stable convergence and balanced predictive behavior, making it a solid starting point for further optimization or temporal evaluation.

FEATURE IMPORTANCE AND SHAP ANALYSIS

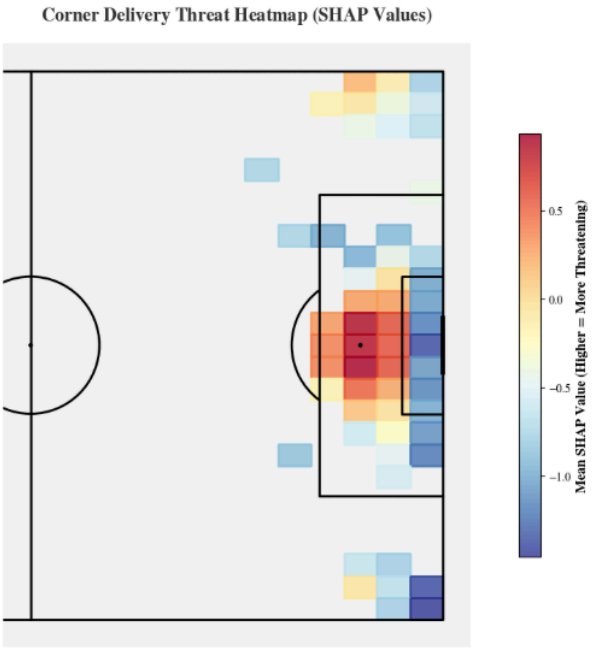
Absolute SHAP values measure the overall impact of each feature on the model's predictions, regardless of direction. In this context, they indicate which variables most strongly influence whether a corner results in a High-Threat or Low-Threat play, providing a global view of feature importance across all samples. These are the Top 10 features ranked by absolute SHAP values:

Top 10 features ranked by absolute SHAP	
end_location_x	0.52
P1_GK_x	0.33
pass_length	0.22
P1_n_att_zone_9	0.19
P1_GK_y	0.16
end_location_y	0.16
P0_GK_y	0.13
P0_GK_x	0.12
P1_n_def_zone_6	0.10
P1_n_att_zone_6	0.08

To facilitate interpretation, the ten most influential features were grouped according to their functional relationships.

Several variables describe related spatial or tactical dimensions (such as ball placement, goalkeeper positioning, or player distribution) so analyzing them together provides a clearer understanding of how different aspects of play contribute to threat generation after corners.

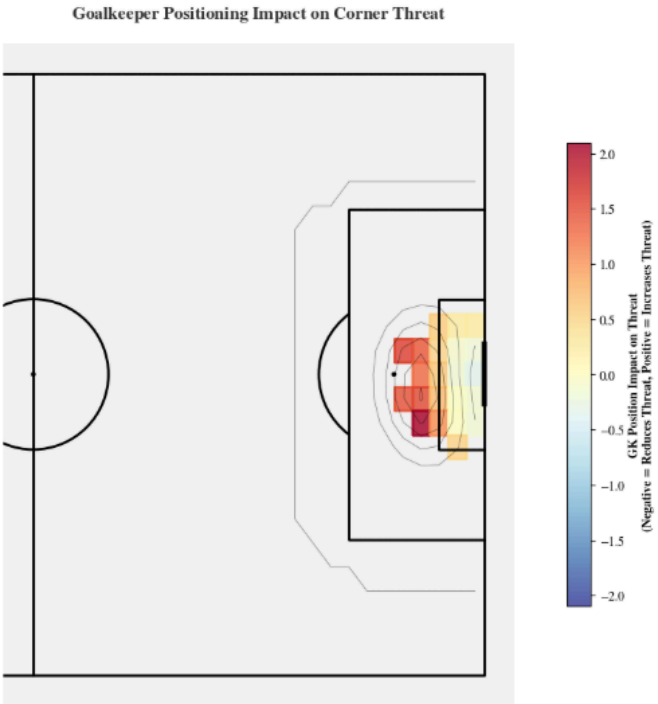
1. Ball trajectory and placement



***Variables**: end_location_x, end_location_y, and pass_length*

These features describe where and how far the ball travels after the corner delivery. Their high SHAP values indicate that the spatial outcome of the pass, particularly its endpoint and distance, is one of the strongest determinants of whether a corner becomes a High-Threat play.

2. Goalkeeper positioning

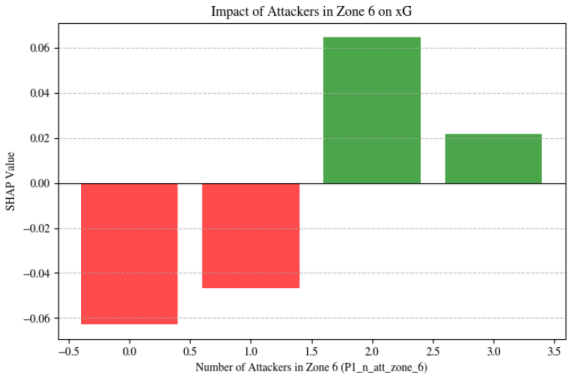


***Variables**: P1_GK_x, P1_GK_y*

The goalkeeper's position during the subsequent play (P1) plays a decisive role in determining the threat.

Higher x-coordinates (further from the goal line) and vertical displacement along the y-axis are often associated with higher threat values, suggesting that when the goalkeeper moves out of optimal positioning, the likelihood of conceding a dangerous opportunity increases.

3. Overload Architecture in P1, zone 6 attackers



***Variables**: P1_n_att_zone_6*

Threat increases when teams overload Zone 6, forcing defensive adjustments. Two attackers create the optimal pressure; a third maintains the effect, but adding more players provides no extra benefit.