\*\*Log Loss\*\* is a performance metric for evaluating the accuracy of a classification model where the prediction is a probability value between 0 and 1. In the context of expected goals (xG), log loss quantifies how close the predicted probability of a goal (xG value) is to the actual outcome (whether a goal was scored or not).

### What Log Loss Indicates

- \*\*Lower Log Loss\*\*: Indicates that the predicted probabilities are close to the actual outcomes. For example, if a shot has a high xG value and results in a goal, or a low xG value and does not result in a goal, the log loss will be low.

- \*\*Higher Log Loss\*\*: Indicates that the predicted probabilities are far from the actual outcomes. For example, if a shot has a high xG value but does not result in a goal, or a low xG value but does result in a goal, the log loss will be high.

### How Log Loss is Calculated

Log loss is calculated using the formula:

\[ \text{Log Loss} = -\frac{1}{N} \sum\_{i=1}^{N} \left[ y\_i \log(p\_i) + (1 - y\_i) \log(1 - p\_i) \right] \]

where:

- \( N \) is the number of observations (shots).

- \( y\_i \) is the actual outcome (1 for a goal, 0 for a non-goal).

- \( p\_i \) is the predicted probability (xG value) for the shot.

### Example from the Dataset

Let's consider a few examples from the dataset:

#### Example 1: High xG and Goal

- \*\*Actual Outcome\*\*: Goal (1)

- \*\*Predicted xG (Probability of Goal)\*\*: 0.75

- \*\*Log Loss Contribution\*\*: \(- (1 \times \log(0.75) + 0 \times \log(0.25)) = -\log(0.75)\)

#### Example 2: Low xG and No Goal

- \*\*Actual Outcome\*\*: No Goal (0)

- \*\*Predicted xG (Probability of Goal)\*\*: 0.10

- \*\*Log Loss Contribution\*\*: \(- (0 \times \log(0.10) + 1 \times \log(0.90)) = -\log(0.90)\)

#### Example 3: High xG and No Goal

- \*\*Actual Outcome\*\*: No Goal (0)

- \*\*Predicted xG (Probability of Goal)\*\*: 0.80

- \*\*Log Loss Contribution\*\*: \(- (0 \times \log(0.80) + 1 \times \log(0.20)) = -\log(0.20)\)

#### Example 4: Low xG and Goal

- \*\*Actual Outcome\*\*: Goal (1)

- \*\*Predicted xG (Probability of Goal)\*\*: 0.05

- \*\*Log Loss Contribution\*\*: \(- (1 \times \log(0.05) + 0 \times \log(0.95)) = -\log(0.05)\)

### Explanation in Terms of xG

- \*\*Better Probabilistic Predictions\*\*: If the model's predicted xG values (probabilities) closely match the actual outcomes, the log loss will be low. This means the model is accurate in assigning higher probabilities to shots that result in goals and lower probabilities to shots that do not.

- \*\*Example\*\*: In the dataset, if a shot is predicted to have an xG of 0.75 (75% chance of scoring) and it actually results in a goal, the prediction is accurate. Conversely, if a shot is predicted to have an xG of 0.10 (10% chance of scoring) and it does not result in a goal, the prediction is also accurate. Both scenarios contribute to a lower log loss.

### Predicted Value Comparison

The predicted xG value (probability of scoring) is being compared to the actual outcome (whether the shot resulted in a goal or not):

- \*\*Predicted xG\*\*: The model's estimation of the likelihood that the shot will result in a goal (a value between 0 and 1).

- \*\*Actual Outcome\*\*: The true outcome of the shot (1 if it resulted in a goal, 0 if it did not).

In summary, log loss measures the accuracy of the model's probabilistic predictions by comparing the predicted xG values to the actual outcomes of the shots. Lower log loss indicates that the model's predicted probabilities are well-aligned with the actual results, making it a reliable metric for evaluating the performance of an xG model.

Sure, let's explain the other metrics in terms of the predicted xG values.

### 1. \*\*ROC AUC Score\*\*

The ROC AUC Score (Receiver Operating Characteristic - Area Under the Curve) measures the ability of the model to distinguish between classes (goals and non-goals).

#### ROC AUC Score Interpretation

- \*\*Value Range\*\*: 0 to 1

- \*\*0.5\*\*: The model is no better than random guessing.

- \*\*1\*\*: Perfect ability to distinguish between goals and non-goals.

- \*\*Higher ROC AUC\*\*: Indicates better performance in distinguishing between goals and non-goals.

#### Example from Dataset

- \*\*Predicted xG (Probability of Goal)\*\*: 0.80

- If this results in a goal, it's a true positive.

- If this does not result in a goal, it's a false positive.

- \*\*Predicted xG (Probability of Goal)\*\*: 0.10

- If this results in a goal, it's a false negative.

- If this does not result in a goal, it's a true negative.

The ROC AUC score evaluates the model’s performance across all possible classification thresholds, balancing the true positive rate (recall) and false positive rate.

### 2. \*\*Precision\*\*

Precision measures the ratio of true positive predictions to the total predicted positives. It indicates the accuracy of the positive predictions (goals).

#### Precision Interpretation

- \*\*Formula\*\*: \(\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}\)

- \*\*High Precision\*\*: Indicates a low false positive rate (few non-goals predicted as goals).

#### Example from Dataset

- \*\*Predicted xG\*\*: 0.80

- If this results in a goal (true positive) and there are few cases where high xG values do not result in goals (false positives), precision is high.

### 3. \*\*Recall\*\*

Recall measures the ratio of true positive predictions to the total actual positives. It indicates the model’s ability to capture all actual positive cases (goals).

#### Recall Interpretation

- \*\*Formula\*\*: \(\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}\)

- \*\*High Recall\*\*: Indicates a low false negative rate (few goals predicted as non-goals).

#### Example from Dataset

- \*\*Predicted xG\*\*: 0.10

- If this results in a goal (false negative) and there are many cases where low xG values do result in goals, recall is low.

### 4. \*\*F1-Score\*\*

F1-Score is the harmonic mean of precision and recall. It balances both metrics, providing a single measure of the model’s accuracy.

#### F1-Score Interpretation

- \*\*Formula\*\*: \(\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\)

- \*\*High F1-Score\*\*: Indicates a good balance between precision and recall.

#### Example from Dataset

- A model with high precision but low recall, or vice versa, will have a lower F1-score compared to a model that balances both.

### 5. \*\*Support\*\*

Support is the number of actual occurrences of the class in the dataset.

#### Support Interpretation

- \*\*Goals (Class 1)\*\*: 72

- \*\*Non-goals (Class 0)\*\*: 600

#### Example from Dataset

- The support indicates how many instances of goals and non-goals are present in the dataset, providing context for evaluating precision, recall, and F1-score.

### 6. \*\*Accuracy\*\*

Accuracy measures the overall ratio of correct predictions to the total number of predictions.

#### Accuracy Interpretation

- \*\*Formula\*\*: \(\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}\)

- \*\*High Accuracy\*\*: Indicates a high overall rate of correct predictions.

#### Example from Dataset

- If the model correctly predicts 600 non-goals and 50 goals out of 672 shots, the accuracy is \(\frac{650}{672} \approx 0.97\).

### 7. \*\*Macro Avg and Weighted Avg\*\*

- \*\*Macro Avg\*\*: The average of precision, recall, and F1-score for each class, treating all classes equally.

- \*\*Weighted Avg\*\*: The average of precision, recall, and F1-score for each class, weighted by the number of true instances for each class.

#### Example from Dataset

- \*\*Macro Avg\*\*:

- Average metrics for goals and non-goals.

- \*\*Weighted Avg\*\*:

- Metrics weighted by the number of non-goals and goals, giving more importance to the majority class (non-goals).

### Summary

- \*\*Log Loss\*\*: Measures the accuracy of probabilistic predictions. Lower values indicate better performance.

- \*\*ROC AUC\*\*: Measures the model’s ability to distinguish between classes. Higher values indicate better performance.

- \*\*Precision\*\*: Accuracy of positive predictions.

- \*\*Recall\*\*: Ability to capture all actual positives.

- \*\*F1-Score\*\*: Balance between precision and recall.

- \*\*Support\*\*: Number of actual occurrences of each class.

- \*\*Accuracy\*\*: Overall ratio of correct predictions.

- \*\*Macro Avg and Weighted Avg\*\*: Summarize model performance across classes, considering class balance.

In the context of xG, these metrics help evaluate how well the model predicts the probability of scoring for each shot, with lower log loss and higher ROC AUC scores indicating better model performance.

Here are the results tabulated for better comparison:

| Model | Log Loss (without tuning) | ROC AUC (without tuning) | F1 Score (Goal, without tuning) | Log Loss (with tuning) | ROC AUC (with tuning) | F1 Score (Goal, with tuning) |

|--------------------|---------------------------|---------------------------|-------------------------------|------------------------|------------------------|-----------------------------|

| Logistic Regression| 0.3791 | 0.9001 | 0.52 | 0.3906 | 0.8984 | 0.54 |

| XGBoost | 0.2718 | 0.8996 | 0.53 | 0.2081 | 0.9056 | 0.57 |

| CatBoost | 0.2064 | 0.9153 | 0.54 | 0.2078 | 0.9124 | 0.49 |

| Random Forest | 0.2545 | 0.9075 | 0.50 | 0.2151 | 0.9130 | 0.56 |

| LightGBM | 0.2698 | 0.9138 | 0.57 | 0.2050 | 0.9042 | 0.56 |

### Interpretation of Results

#### Logistic Regression

- \*\*Without tuning\*\*: The model has a Log Loss of 0.3791 and a ROC AUC of 0.9001. The F1 score for the 'Goal' class is 0.52, indicating moderate performance.

- \*\*With tuning\*\*: The Log Loss slightly increased to 0.3906, and the ROC AUC slightly decreased to 0.8984. The F1 score for 'Goal' improved to 0.54.

#### XGBoost

- \*\*Without tuning\*\*: The model has a Log Loss of 0.2718 and a ROC AUC of 0.8996. The F1 score for 'Goal' is 0.53.

- \*\*With tuning\*\*: The Log Loss significantly decreased to 0.2081, and the ROC AUC increased to 0.9056. The F1 score for 'Goal' improved to 0.57.

#### CatBoost

- \*\*Without tuning\*\*: The model has a Log Loss of 0.2064 and a ROC AUC of 0.9153. The F1 score for 'Goal' is 0.54.

- \*\*With tuning\*\*: The Log Loss slightly increased to 0.2078, and the ROC AUC slightly decreased to 0.9124. The F1 score for 'Goal' decreased to 0.49.

#### Random Forest

- \*\*Without tuning\*\*: The model has a Log Loss of 0.2545 and a ROC AUC of 0.9075. The F1 score for 'Goal' is 0.50.

- \*\*With tuning\*\*: The Log Loss decreased to 0.2151, and the ROC AUC increased to 0.9130. The F1 score for 'Goal' improved to 0.56.

#### LightGBM

- \*\*Without tuning\*\*: The model has a Log Loss of 0.2698 and a ROC AUC of 0.9138. The F1 score for 'Goal' is 0.57.

- \*\*With tuning\*\*: The Log Loss decreased to 0.2050, and the ROC AUC slightly decreased to 0.9042. The F1 score for 'Goal' decreased slightly to 0.56.

### Choosing the Best Model

- \*\*Log Loss\*\*: Lower is better, indicating more accurate probabilistic predictions.

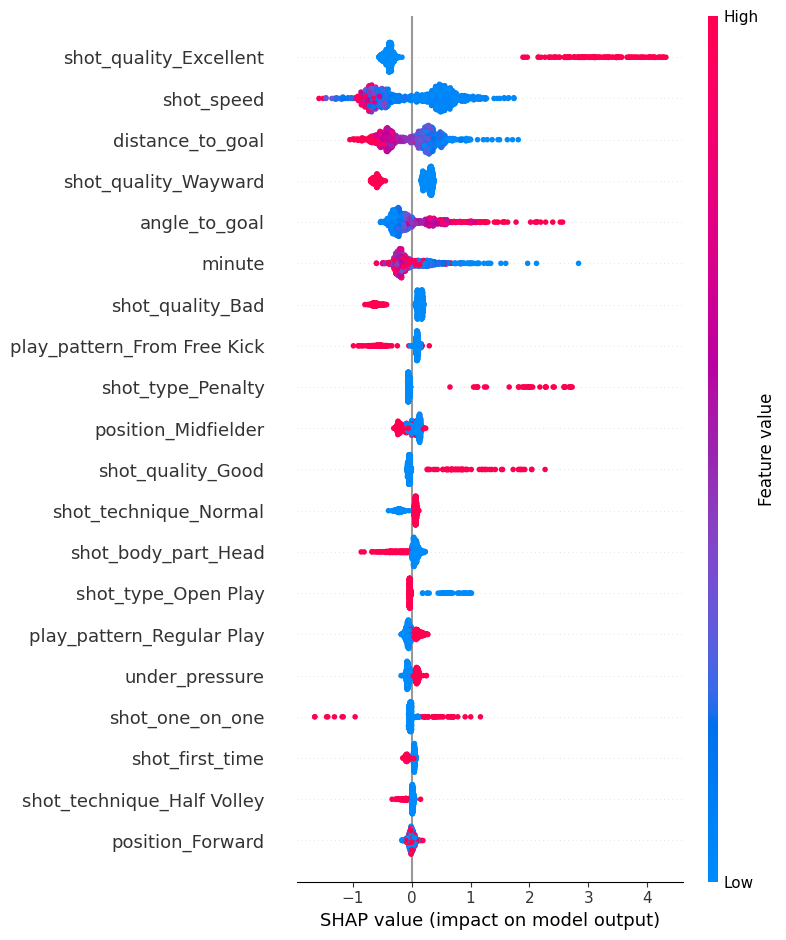
- \*\*ROC AUC\*\*: Higher is better, indicating better discrimination between classes.

- \*\*F1 Score (Goal)\*\*: Higher is better, indicating better performance in predicting goals.

The CatBoost model, without hyperparameter tuning, shows the lowest Log Loss (0.2064) and a high ROC AUC (0.9153), indicating strong probabilistic predictions and good discrimination between goals and non-goals. The F1 score for 'Goal' is also decent (0.54).

### Conclusion

Considering the balance between Log Loss, ROC AUC, and F1 Score, \*\*CatBoost\*\* without hyperparameter tuning seems to be the best model for predicting the expected goals (xG). This model provides a good balance of accurate probabilistic predictions and classification performance.

SHAP:  


Certainly! The SHAP summary plot provides a comprehensive visualization of feature importance and their impact on the model's output. Here’s a detailed explanation of the plot:

1. \*\*SHAP Value (x-axis)\*\*:

- The SHAP value represents the impact of a feature on the model's output. Positive SHAP values push the model output higher, indicating a higher likelihood of the shot being a goal. Negative SHAP values push the output lower, indicating a lower likelihood of the shot being a goal.

2. \*\*Features (y-axis)\*\*:

- The features are listed in descending order of importance from top to bottom. The most important features are at the top.

3. \*\*Feature Value (Color)\*\*:

- The color indicates the value of the feature. Red represents high feature values, while blue represents low feature values.

### Key Insights from the SHAP Summary Plot:

- \*\*shot\_quality\_Excellent\*\*:

- This is the most influential feature. High values (red dots) have a strong positive impact on the model's output, significantly increasing the likelihood of a goal. This makes sense as excellent shot quality should correlate with higher chances of scoring.

- \*\*shot\_speed\*\*:

- High shot speed (red) has a positive impact on the likelihood of a goal, whereas lower shot speeds (blue) tend to have a negative impact.

- \*\*distance\_to\_goal\*\*:

- Shots taken from a shorter distance to the goal (red) positively impact the likelihood of scoring, while shots from further distances (blue) decrease the likelihood of scoring.

- \*\*shot\_quality\_Wayward\*\*:

- This feature also has a significant impact. High values (red) have a negative impact on the likelihood of a goal, indicating that wayward shots are less likely to result in a goal.

- \*\*angle\_to\_goal\*\*:

- A lower angle to the goal (red) positively impacts the likelihood of scoring, whereas a higher angle (blue) negatively impacts the likelihood.

- \*\*minute\*\*:

- The minute in which the shot is taken has varying impacts, with no clear trend observed in the summary plot.

- \*\*shot\_quality\_Bad\*\*:

- High values (red) have a negative impact on the likelihood of scoring, indicating that bad shot quality decreases the chances of scoring.

- \*\*play\_pattern\_From Free Kick\*\*:

- Shots taken from free kicks (red) seem to have a slightly negative impact on the likelihood of scoring compared to other play patterns.

- \*\*shot\_type\_Penalty\*\*:

- Penalty shots (red) tend to have a positive impact on the likelihood of scoring, which is expected as penalties have a higher chance of resulting in goals.

- \*\*position\_Midfielder\*\*:

- Being a midfielder (red) has a slightly negative impact on the likelihood of scoring compared to other positions.

Other features like `shot\_quality\_Good`, `shot\_technique\_Normal`, `shot\_body\_part\_Head`, `shot\_type\_Open Play`, and so on, show varying degrees of influence, but the top features like `shot\_quality\_Excellent`, `shot\_speed`, `distance\_to\_goal`, and `shot\_quality\_Wayward` are the most impactful.

### Conclusion:

The SHAP summary plot reveals that shot quality, shot speed, and distance to goal are among the most critical features influencing the model's predictions. This aligns with our intuition that better quality shots, faster shots, and shots taken from closer distances are more likely to result in goals.