### **94 Model Summary Report**

#### Datasets and Preparation

* Datasets Used:
  + cleaned\_matches\_data\_df.csv: Contains cleaned match data with various features.
  + Additional datasets were referenced but not explicitly used in the final model.
* Data Preprocessing:
  + Missing Values: Checked and handled missing values.
  + Feature Selection: Selected relevant features for the model.
  + Data Splitting: Divided the data into training and testing sets. Training data included matches before November 30, 2023, while testing data included matches from and after this date.

#### Features Used in the Model

* Primary Features: Included but not limited to 'Pre-Match PPG (Home)', 'Pre-Match PPG (Away)', 'home\_team\_goal\_count', 'away\_team\_goal\_count', 'average\_corners\_per\_match\_pre\_match', 'average\_cards\_per\_match\_pre\_match', 'odds\_ft\_home\_team\_win', 'odds\_ft\_draw', 'odds\_ft\_away\_team\_win'.
* Feature Exclusion: Certain features were excluded to prevent data leakage and ensure model robustness. For example, 'FTR' (Full Time Result) was not used as it would cause data leakage.

#### Model Selection and Rationale

* Model Chosen: Logistic Regression.
* Rationale: Given its effectiveness in classification tasks, especially in predicting binary outcomes. The simplicity of logistic regression makes it a good starting point for this kind of predictive modeling.

#### Model Evaluation and Refinement

* Initial Performance: The model showed high accuracy in initial tests.
* Cross-Validation: Implemented 5-fold cross-validation to robustly assess the model's performance.
* Precision, Recall, F1-Score: Used to evaluate the model beyond accuracy, particularly due to the imbalanced nature of the dataset.

#### Critical Information for Replication

* Feature Names: It's crucial to ensure that the features used during model training are consistent with those used during prediction.
* Data Structure: The structure of the input data for predictions should match the structure of the data used for training the model.
* Preprocessing Steps: Any preprocessing applied to the training data must also be applied to the new data used for predictions.

#### Next Steps for Testing on New Data

* Acquiring New Data: Obtain updated datasets for upcoming matches.
* Preprocessing: Apply the same preprocessing steps to the new data as were applied to the training data.
* Making Predictions: Use the trained model to predict outcomes on the new data.
* Evaluation: Assess the model's performance on the new data and refine if necessary.

#### Exporting the Model and Data

* Model: Saved as a .pkl file using joblib.
* DataFrames: Exported as .csv files.
* Feature Names: Saved in a text file for consistency in future use.

### **Actions Taken and Reasons**

* Data Leakage Prevention: Certain columns like 'FTR' were not used to prevent data leakage.
* Feature Selection: Based on domain knowledge and exploratory data analysis, key features impacting match outcomes were chosen.
* Cross-Validation: To ensure the model's robustness and prevent overfitting.

### **Future Considerations**

* Model Tuning: Depending on the results with new data, further tuning of hyperparameters or trying different algorithms might be beneficial.
* Feature Engineering: Explore additional features or transformations that could improve model performance.