Task 2: Behavioral and Transactional Risk Modeling for Online Marketplaces (Competition)

1. Project Background

"GlobalMart" is a massive online marketplace connecting millions of buyers and sellers worldwide. Due to its scale, the platform is a frequent target for Risky activities, ranging from automated bot purchases to the use of stolen user accounts. These activities not only cause financial loss but also damage the platform's reputation. As a senior data scientist on the Trust and Safety team, you are tasked with building a high-precision machine learning model to detect these Risky orders in real-time. Unlike the foundational task, this project requires you to handle a large, complex, and messy real-world dataset.

2. Task Goal

This is a **competitive task**. Your goal is to build the best possible binary classifier to identify Risky orders (IsRisky = 1). Your success will be measured by your model's F1-score on a hidden test set. This task will challenge your skills in data wrangling, feature engineering, and model interpretation.

3. Dataset Information

Your data is split into four files, simulating a real database structure:

- **globalmart_train_transactions.csv:** The main training table for transactions. Contains the IsRisky label.
- **globalmart_train_identity.csv:** Supplemental identity information for the training set. Can be joined with the main table.
- **globalmart_test_transactions.csv:** The main testing table for transactions. You need to predict the IsRisky attribute for each item in this table.
- **globalmart_test_identity.csv:** Supplemental identity information for the testing set.

Other files:

- **sample_submission_2.csv** The sample file to show the output format. The wrong format will lead to an unknown result.
- **evaluate_mac_2** The command line tool to evaluate your result on macOS. Usage: Press "command + space" to open spotlight search and type in "terminal", then type in the following command: ./evaluate_mac_2 ./submission_2.csv. Please note that "./" denotes the current position of the command line and "submission_2.csv" denotes your submission file name.
- **evaluate_windows_2.exe** The command line tool to evaluate your result on Windows. Usage: Press "command + r" and then type "cmd" in the dialog box to launch a terminal, then type in the following

command: .\evaluate_windows_2.exe .\submission_2.csv. Please note that "." denotes the current position of the command line and "submission_2.csv" denotes your submission file name.

4. Attribute Information

OrderID: Unique ID of an order.

IsRisky: Whether an order is Risky (Target Variable).

OrderTimestamp: Time elapsed since a reference date.

OrderAmount: The value of the order.

PaymentType: The method of payment.

CardNetwork: The credit card network (e.g., Visa, MasterCard).

PayerEmailProvider: The email provider of the payer.

DeviceOS: The operating system of the user's device.

CustomerBehavior1 ... CustomerBehavior14: Counting features related to the customer's historical activities.

TimeDelta1 ... TimeDelta15: Time-related delta features from previous events.

MatchStatus1 ... MatchStatus9: Features indicating whether certain pieces of information matched (e.g., billing and shipping address).

IdentityFeature1 ... IdentityFeature38: Anonymized features related to user identity and device.

5. Core Challenges & Requirements

1. **Missing Value Imputation**: This dataset contains a significant amount of missing data. In your report, you must dedicate a section to explaining your strategy for handling these missing values and justify your choices.
2. **Feature Engineering**: To achieve a high score, you **must** create **at least two new, meaningful features** from the existing data. Document the logic behind your new features in your report. Examples could include combining existing features, extracting patterns, or creating interaction terms.
3. **Advanced Modeling**: You are free to use any machine learning algorithm, including advanced techniques like Gradient Boosting (XGBoost, LightGBM) or Neural Networks.

4. **Model Interpretability Analysis (Optional)**: A prediction is not enough; we need to understand why. You can explore the importance of different values or other available analysis methods.

6. Output & Submission

You must submit a submission.csv file containing your predictions for the test set. The file must contain only two columns: OrderID and IsRisky, and follow the format shown in sample_submission.csv.