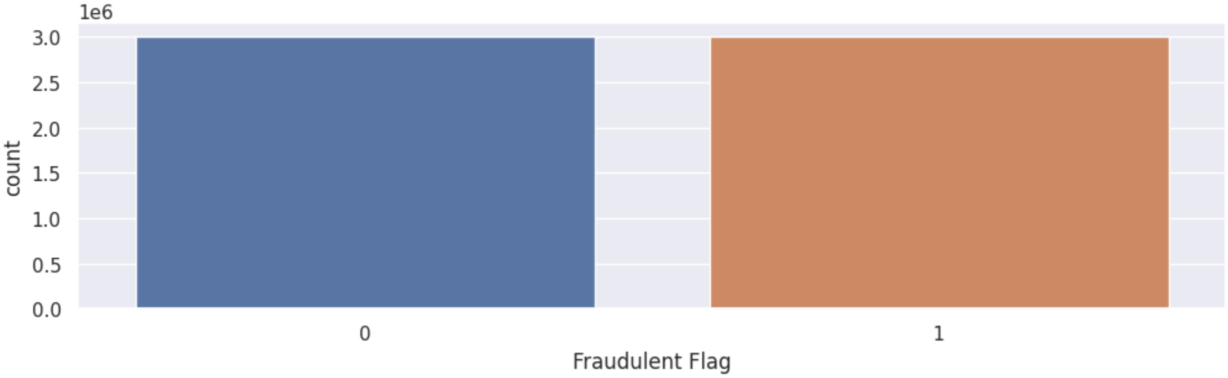
Oluwafemi’s Executive Report On DataFest23’ Datathon

Assessing Data:

Started my analysis by assessing the dataset provided which contain 6 million rows and 32 columns out of which was 18 categorical column, 13 numerical column and 1 datetime column. Went ahead to investigate the data by checking the data size, duplicates, missing values, data types, summary statistics, and other observation as seen in the code.

I plotted a histogram plot of my target variable which was quite balanced

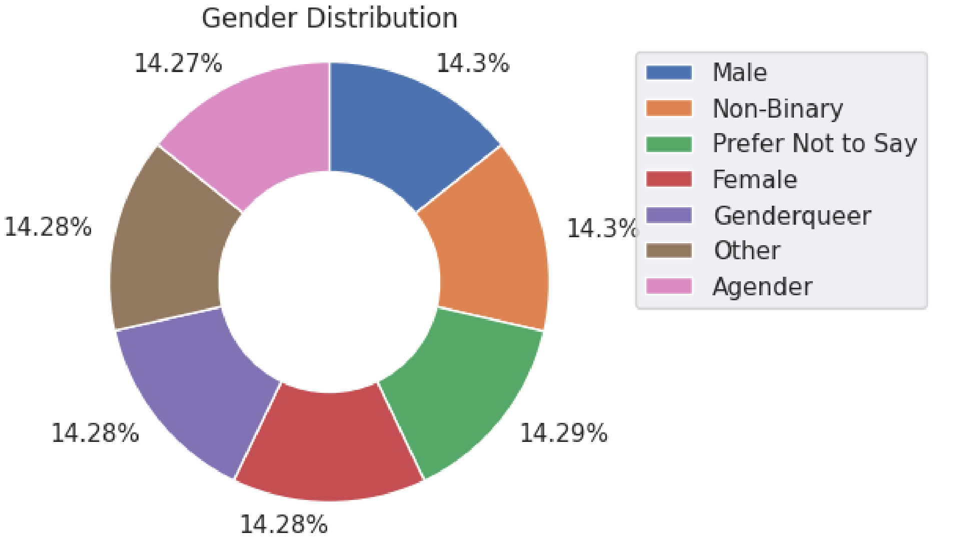


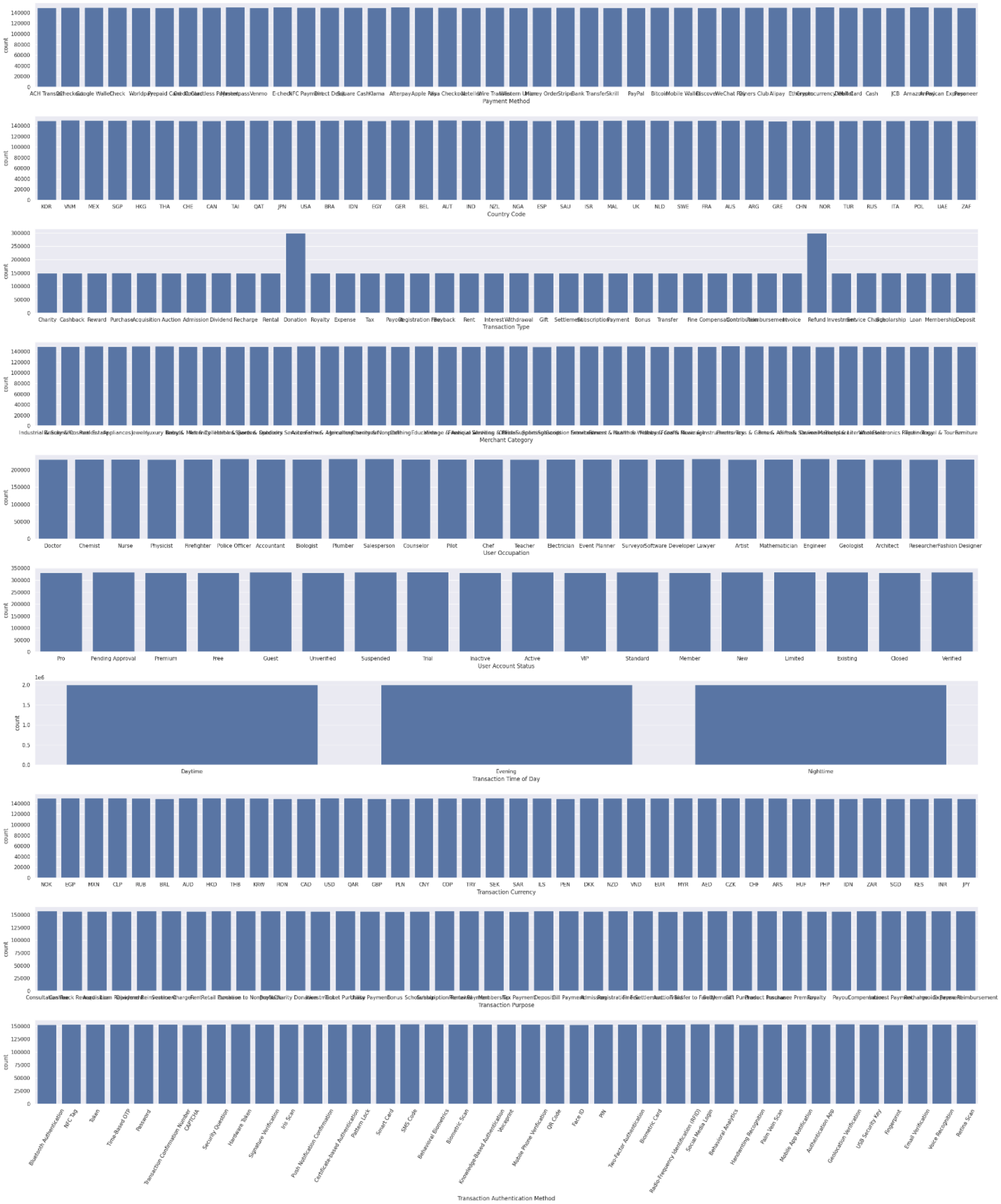
Data Exploration:

Univariate Analysis

Carried out Univariate Analysis by Checking Frequency for categorical variables and distribution for numerical variables.

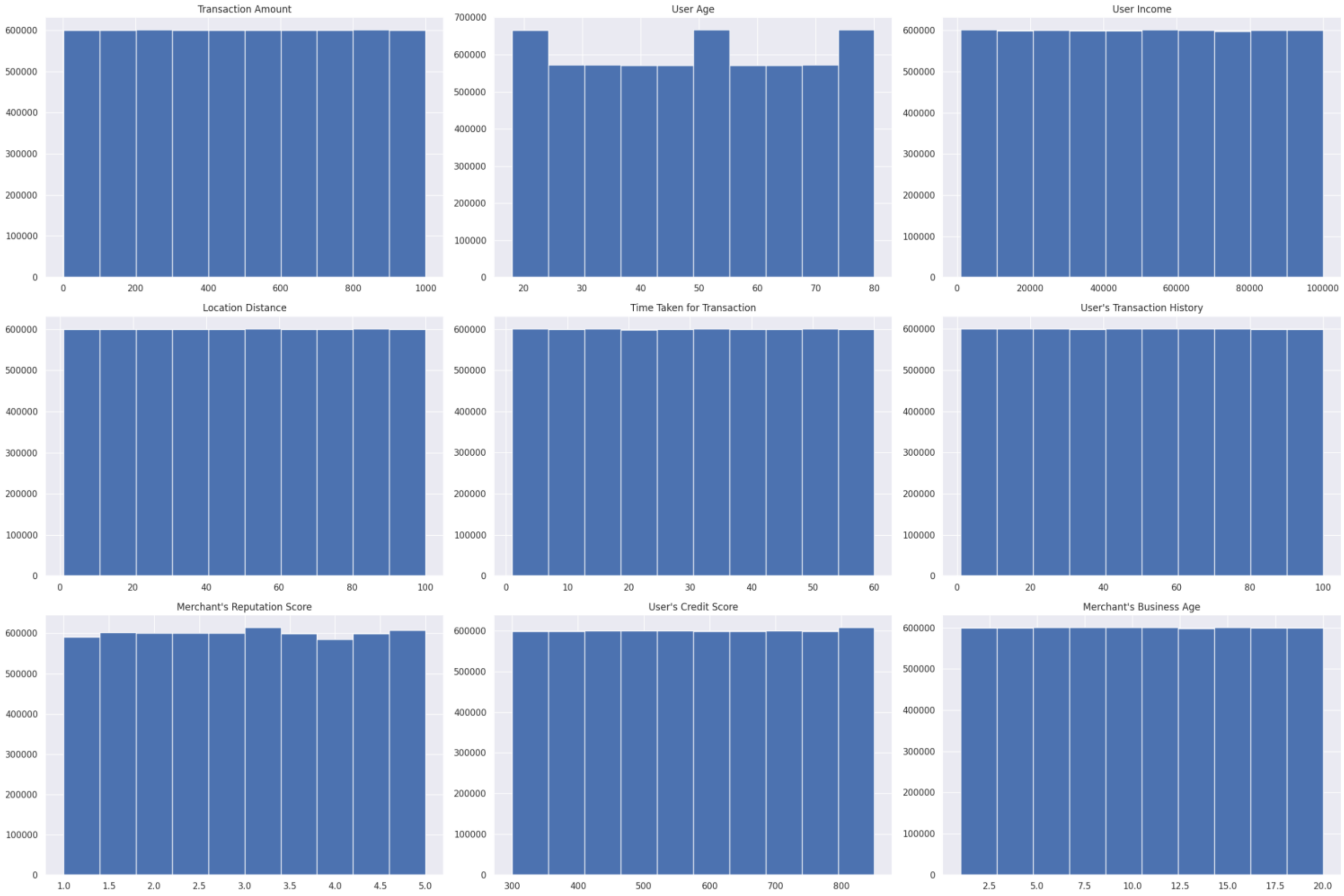
Categorical Variable





One quick observation is that the distribution is uniform and there’s nothing unusal in the pattern.

Numerical Variable:



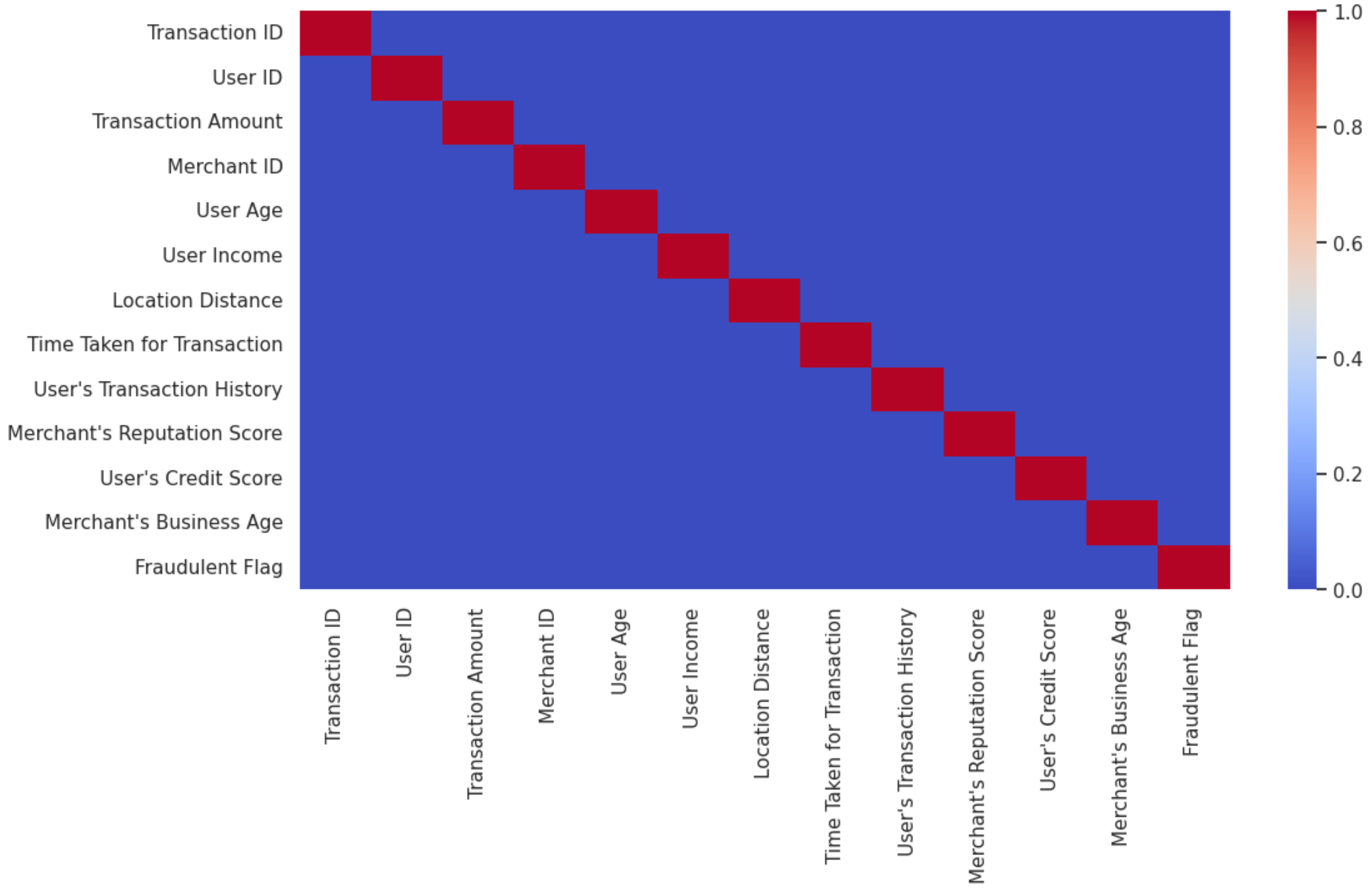
There seem not to be any pattern in the numerical distribution as well.

Bivariate Analysis

I proceeded with Bivariate analysis, where I compared each variable with the target variable (Fraudalent Flag), which didn't provide much insight as there seems to be a uniform distribution for the target

Multivariate Analysis

I proceeded to check the relationship with all categorical variable with fewer cardinality (lass than 41) to check for correlation and multicollinearity as seen below.



Feature Engineering:

I proceeded to extract new varibales from Transaction date and time including, Day of the week, Month of the year, and Day of the month and conducted bivariate analysis to observe any pattern.

Model Training and Evaluation:

Used Random Forest Classifier model and stratified Kfold (folds=4) for data splitting and set some parameters. Performed the prediction with an accuracy of slightly above the baseline accuracy of 0.5

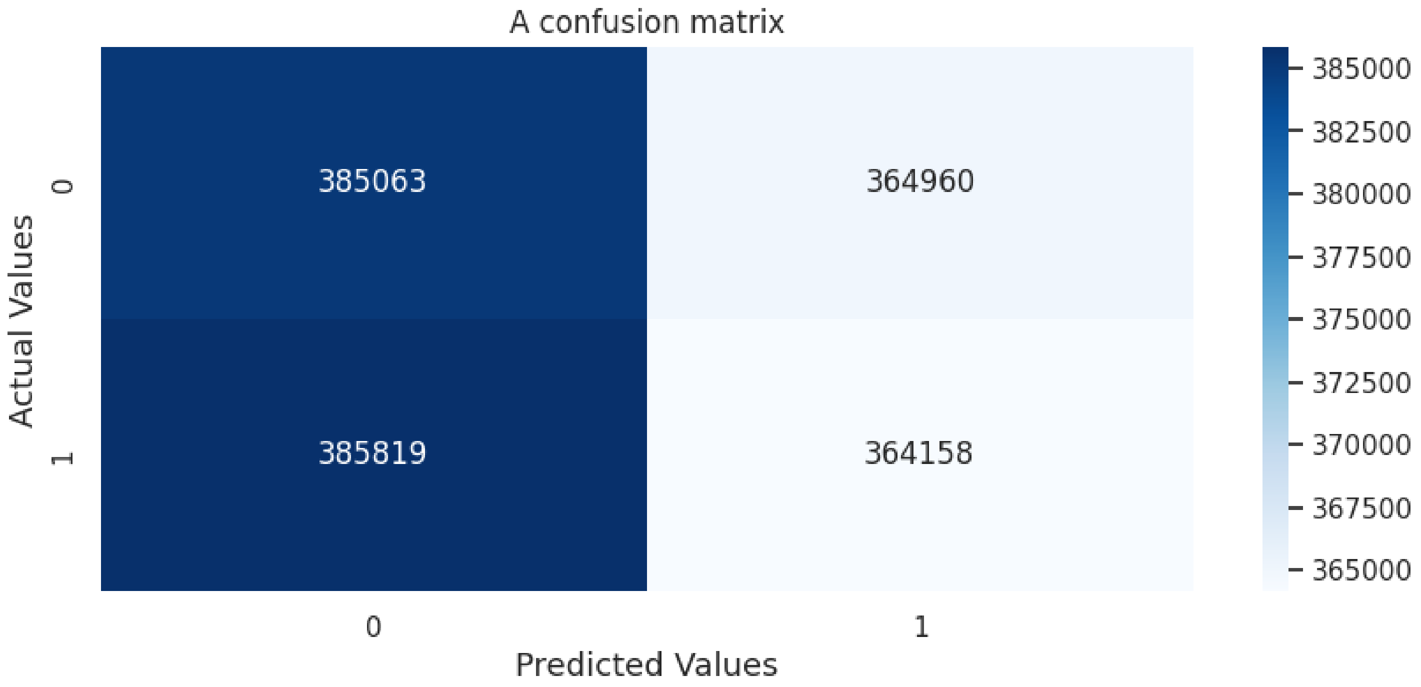
Communicating Result:

Confusion Matrix Interpretation

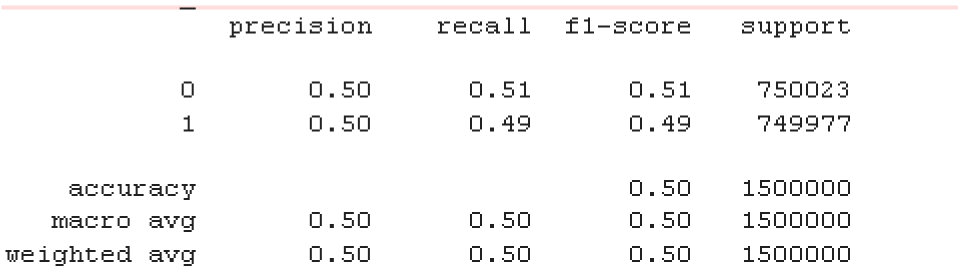
The confusion matrix you sent shows the results of a classification model for predicting fraudulent transactions. The model has two possible predictions: fraud or not fraud. The actual values are shown on the left side of the matrix, and the predicted values are shown on the right side.

This means that the model correctly predicted fraud in 385,063 cases and correctly predicted not fraud in 364,960 cases. However, the model also made one false positive prediction (predicting fraud when the actual value was not fraud) and one false negative prediction (predicting not fraud when the actual value was fraud).

Overall, the model has a high accuracy, with 50% of its predictions being correct.



Classification Report here shows and f1 score od 51% for the Not Fraudalent Transaction Transaction and 49% for the Fraudulent Transaction

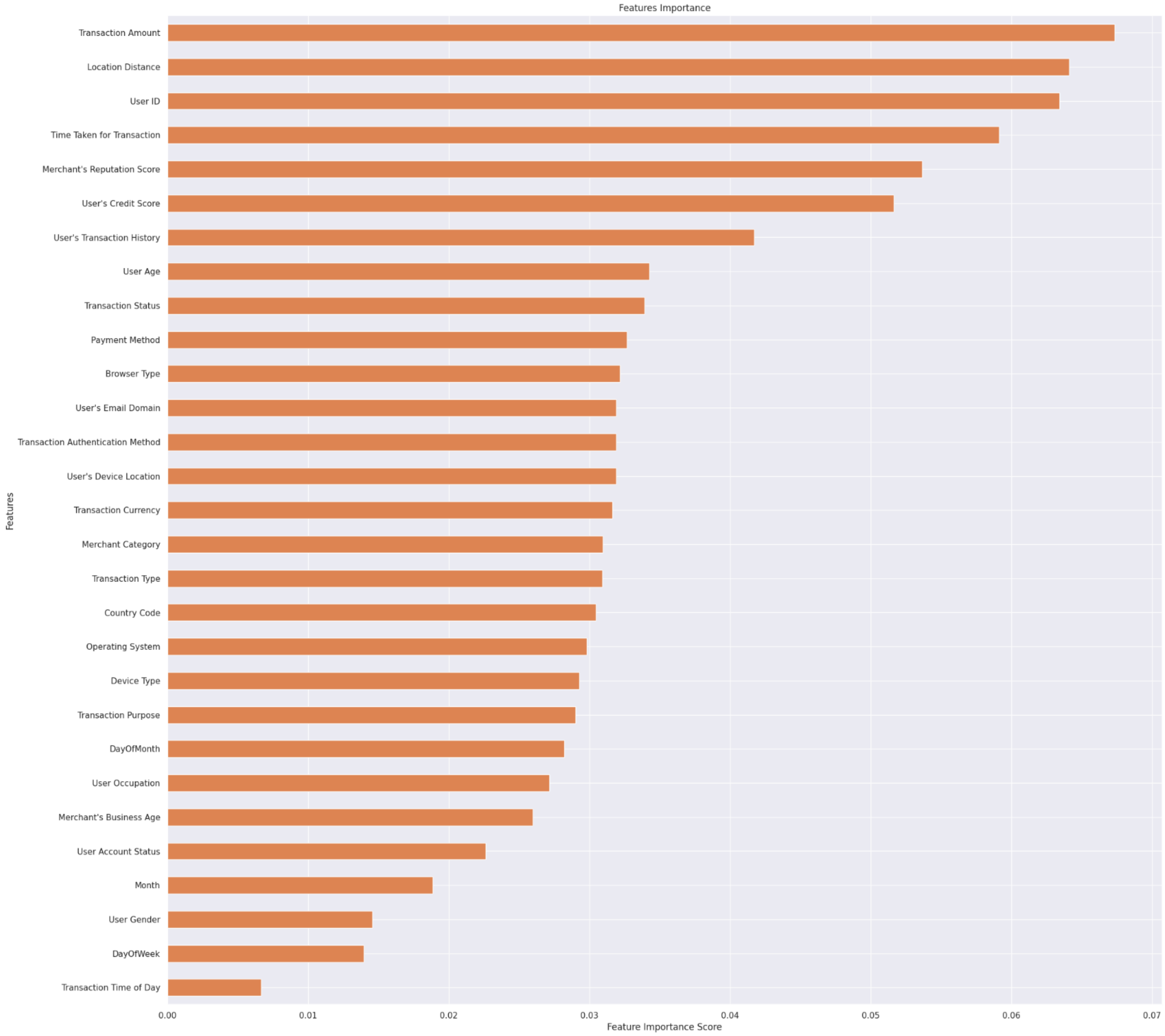


```
              precision    recall  f1-score   support

           0       0.50      0.51      0.51    750023
           1       0.50      0.49      0.49    749977

    accuracy                           0.50   1500000
   macro avg       0.50      0.50      0.50   1500000
weighted avg       0.50      0.50      0.50   1500000
```

Feature Importance


Features Importance
Transaction Amount
Location Distance
User ID
Time Taken for Transaction
Merchant's Reputation Score
User's Credit Score
User's Transaction History
User Age
Transaction Status
Payment Method
Browser Type
User's Email Domain
Transaction Authentication Method
User's Device Location
Transaction Currency
Merchant Category
Transaction Type
Country Code
Operating System
Device Type
Transaction Purpose
DayOfMonth
User Occupation
Merchant's Business Age
User Account Status
Month
User Gender
DayOfWeek
Transaction Time of Day
Features
Feature Importance Score