Contents

I.Problem Description

Predict if the Merchant is Fraudster or not for an e-commerce client

'XYZ' is a large e-commerce company with its operations in several countries. As the online giant grows, so has the number of fraudster merchants are. They deliver counterfeits or, in some cases, nothing at all. Such schemes leave customers duped, and place both legitimate merchants and the company itself in a constant battle to rid the marketplace of scammers. Determining this is also important in budgeting for fraud investigation. It's a well-known problem both to the company and to merchants, which they say hasn't effectively addressed the issue. They are serious about it and want to protect themselves from these fraudulent merchants using technology.

You are expected to create an analytical and modelling framework to predict the Merchant Fraudulency(yes/no) based on the quantitative and qualitative features provided in the dataset while answering other questions too cited below.

II.The datasets are provided as cited below:

Target attribute: "fraudster" (yes – 1, no – 0)

Train:

- train_merchant_data.csv : Merchant Information
- train_order_data.csv : Order Information
- train.csv : Target Label Information

Test:

- test_merchant_data.csv : Merchant Information
- test_order_data.csv : Order Information
- test.csv : Target is not available as it is to be predicted

➔ ip_boundaries_countries.csv : IP addresses boundaries for each country (common for both train and test)

All Attributes names are self-explanatory.

III.Tasks:

Model Building:

You are expected to create an analytical and modelling framework to predict the Merchant Fraudulency based on the quantitative and qualitative features provided in the datasets. You may derive new features from the existing features and also from the domain knowledge, which may help in improving the model efficiency.

Visualization Tasks:

Exploratory Data Analysis using visualizations in R Notebook or Jupiter notebook format. (all train data to be used for this task)

- List down the insights/patterns observed from the visualizations
- Explain the impact of most important attributes on target attribute observed from the visualizations.

Observations:

Is there any overfitting or underfitting problem? If yes, how do you address it?

IV. **Evaluation Metric:**

- Consider 'F1-score' of the fraudulent class as the error metric for classification task to tune the model and for submissions in the tool.

V. Hints:

Both Python and R provides functions to convert IP string to numeric format which makes the number comparison easier.