**Submitted by: Sanskar Sharma**

**E-mail:** [**sksharma@mitaoe.ac.in**](mailto:sksharma@mitaoe.ac.in)

**Batch: ML-1/1B5**

**CONCLUSION:**

**Exploratory analysis of the data:**

The given dataset is a record of customers uniquely identified by their e-mails and their personal details and transactions they’ve made via different method providers and type which resulted in certain outcomes like order status no. of orders, payments, transactions and an attribute which says whether the customer is fraud or not.

On studying the data thoroughly and applying necessary operations certain observations were made such as,

1. The customer detail table have 8 entries with same email. E-mail is a primary key hence should be unique, therefore keeping only first entry and removing other 7 to avoid data redundancy on merging the two tables.
2. To have a better understanding of both data sets, additional columns have been added to customer table such has total transaction amount for each customer, no. of each payment method type used by customer, no. of each payment method provider used by customer, count of each order status of customers, failed transactions of each customer etc.
3. The data contained out layers such as transaction amount of one customer was 358 exceptionally high than other transactions which may hamper our model’s prediction . So such out layers have been removed.

**Summary of trends of data /visualisation and significant and insignificant factors of a fraud:**

1. Most of the not fraud customer uses “JCB 16 digit” and “Maestro” as the method provider while the fraud ones prefers “JCB 16 digit” and “VISA 16 digit”.
2. Most of the not fraud customers prefers “card” and “paypal” as payment method type whereas the fraud ones uses “card” and “bitcoin”.
3. Most of the customers have fulfilled order state.
4. In general, customers preferring bitcoin and apple pay as payment method types, VISA 13 digit and Voyager as payment method providers and have their orders fulfilled are most probable to be a **FRAUD**.
5. “**PayPal**” method type is highly secure, hence their user are generally not fraud.
6. Average transaction of a total fraud customers is higher than not fraud. That implies a large amount transaction is suspicious always,
7. Whereas, no. of orders, payments and transactions made by them are less which is quite obvious for a **FRAUD**.

**An overview of model’s implementation and success percentages.**

The problem statement demanded a predictive model on customers details and their transaction details predicting them being a fraud or not fraud.

Since, we’ve been given the labelled data so we proceeded with Supervised Learning. Problem statement demanded classification so I chose to use the most accurate and effective algorithm, **Random** **Forest**. Python provides a very comfortable and efficient platform for implementation of such glorious algorithms and also visualizations for better understanding.

**Co-relation Matrix** was used to identify the relations between pair of two attributes.

It started with importing some python defined algorithms from sklearn and under various modules like model\_selection, ensemble, datssets.

* from sklearn.model\_selection import GridSearchCV
* from sklearn.model\_selection import RandomizedSearchCV
* from sklearn.datasets import load\_digits
* from sklearn.ensemble import RandomForestClassifier

It started with splitting the dataset(merged) as test(0.2) and train(0.8).

Then, using 20 Decision trees under random forest to predict for the test data using the train data and showed the accuracy result of top 3 decision tree models out of 20.

It was the first raw approach towards the prediction model which resulted in an accuracy of **98.86%**

Now, to get the best accurate results and reduce the overfitting of data “Hyper Parameter Tuning” was must. So, by importing the GridSearchCV and RandomizedSearchCV algorithms from sklearn.model\_selection two other models were prepared with enhanced accuracy rates as **98.95%** and **99.24%** respectively.

Receiver Operating characteristic curves were plotted for all 3 Random forest models of True and False positive rates.

Lastly, by using GridSearchCV the best features or important features were identified.