**Group 24 : Ng Wei Han, Dylan Wo, Kenny Low, Khoo Wu Chuan, Jeyakumar Jegan**

**Motivation**

* Social Implication:
  + Improved Security: Fraud detection helps increase the security of credit card transactions, preventing financial hardship and stress for customers.
  + Trust in Financial Systems: When customers feel secure in their transactions, they are more likely to continue using credit cards that provide convenience to both customers and merchants.
* Business Implication:
  + Loss Prevention: Fraud detection is able prevent any fraudulent transactions from going through, safeguarding customers’ bank accounts.
  + Reduce Costs: Being able to detect fraud automatically can reduce the cost as opposed to if detection was being done manually.
  + Improved User Experience: Effective fraud detection systems can identify and block fraudulent transactions in real-time, enabling legitimate transactions proceed smoothly, enhancing overall user experience.
  + Brand Reputation: A business with strong security measures and low fraud risk will have a positive brand reputation, attracting more customers and partners.

**Dataset**

Source:<https://www.kaggle.com/datasets/kartik2112/fraud-detection/data>

Real credit card transaction datasets often contain sensitive and personal information, which makes it extremely challenging to find a public credit card transaction dataset. Therefore, we decided to make use of a simulated credit card transaction dataset generated using [Sparkov Data Generation](https://github.com/namebrandon/Sparkov_Data_Generation).

The dataset contains legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers the credit cards of 1000 customers doing transactions with a pool of 800 merchants. It has a total of 22 columns and 1296675 rows.

**Data Preprocessing/ML Models**

**EDA**

* **Data cleaning** → Checking for null values, anomalous values etc.
* **Checking target imbalance** → Target imbalance is apparent in fraud datasets, consider if oversampling / undersampling methods could be used.
* **Plotting distributions of features for Fraudulent against Non-Fraudulent transactions** → Find the notable features that greatly influence whether the transaction is a fraud.

**Feature Engineering**

* **Encoding** → Observe the amount of categorical features in the dataset, decide whether to encode them, disclude them etc.
* **Normalization** → Check if features need to be normalized / have their distributions transformed.
* **ML modeling for feature importance** → PCA / Random Forest etc.

**Implementing ML models**

* **Implementation** → Random Forest / XGBoost / LightGBM
* **Evaluation methods** → ROC / Confusion Matrix (F1 Score, Recall etc.)