Credit Card Fraud Detection in Banking/Finance

# Problem

Credit card fraud is a critical issue in the banking and finance sector, resulting in financial losses for banks, financial institutions, and individuals. According to a report by Nilson, in 2019 global losses due to credit card fraud amounted to $28.65 billion, and by 2023, this figure exceeded $32 billion. The impact of credit card fraud extends far beyond financial damage, affecting victims psychologically, legally, and through identity theft. Identity theft can have long-term consequences on victims' credit scores, ability to secure loans, and even employment opportunities.

This project aims to develop a machine learning model capable of detecting fraudulent transactions within a large, imbalanced dataset of credit card transactions.

# Anticipated Hurdles

1. Imbalanced Data: The dataset is imbalanced, with fraudulent transactions being a small minority compared to legitimate ones. Handling this imbalance is critical to develop a steady model.
2. Feature Engineering: Determining the best features and algorithms for this is challenging.
3. Evaluation Metrics: Traditional accuracy metrics are insufficient for imbalanced datasets, so advanced metrics like Precision, Recall, and F1-Score.

# Pertinent Previous Studies and Their Limitations

Many studies have addressed fraud detection using various machine learning techniques. However, they often fall short in handling imbalanced data effectively. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) and undersampling are not always optimized, leading to either overfitting or underfitting. Furthermore, the use of standard metrics such as accuracy instead of metrics designed for imbalanced datasets limits the reliability of the results.

# Data Sources

The dataset for this project is sourced from a well-known credit card fraud detection dataset available on Kaggle [here](https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). It contains thousands of anonymized transactions, where a small percentage are labeled as fraudulent.

# Evaluation Methodology

The model's performance will be evaluated using advanced metrics suited for imbalanced datasets, including:  
- Precision: To assess the accuracy of fraud predictions.  
- Recall: To evaluate how well the model captures actual fraud cases.  
- F1-Score: To balance precision and recall.  
- ROC-AUC: To provide insight into the model's ability to distinguish between fraud and non-fraud transactions.  
  
Models like Logistic Regression, Random Forest, and XGBoost will be tested, and techniques such as SMOTE and undersampling will be applied to handle the data imbalance.

# Group Members

1. [Group Member Name 1]  
2. [Group Member Name 2]  
3. [Group Member Name 3]