# Machine Learning Portfolio Project: Credit Fraud Detection

## Overview

\*\*Project Description\*\*: The objective of this project is to develop an effective credit fraud detection system that can accurately identify and prevent fraudulent transactions within a financial institution's ecosystem. With the increasing sophistication of fraudsters and the evolving nature of fraudulent activities, there is a critical need for a robust and adaptive solution that can detect anomalies and suspicious patterns in real-time.

The current scenario highlights a growing concern for financial institutions, as they face significant losses due to fraudulent activities. Traditional rule-based systems and basic anomaly detection methods have proven insufficient in keeping pace with the evolving tactics of fraudsters. This necessitates the development of an advanced, machine learning-based solution that can adapt to new patterns of fraudulent behaviour.

\*\*Technologies Used\*\*: Python (Pandas, Numpy, Scikit-learn, Matplotlib, Seaborn)

## Table of Contents

1. [Introduction](#introduction)

2. [Data Preprocessing](#data-preprocessing)

3. [Exploratory Data Analysis (EDA)](#exploratory-data-analysis)

4. [Feature Engineering](#feature-engineering)

5. [Model Development](#model-development)

6. [Model Evaluation](#model-evaluation)

7. [Deployment (if applicable)](#deployment)

8. [Conclusion and Future Improvements](#conclusion-and-future-improvements)

## 1. Introduction

\*\*Objective\*\*: To develop an effective credit fraud detection system that can accurately identify and prevent fraudulent transactions within a financial institution's ecosystem.

\*\*Dataset\*\*:

Downloaded the dataset from Kaggle. The Dataset has a size of 76.3MB in total. It has 1000000 records and 8 features columns (including target). Loaded the data as a Dataframe using Pandas.

**Feature Description:**

distance\_from\_home - the distance from home where the transaction happened.

distance\_from\_last\_transaction - the distance from last transaction happened.

ratio\_to\_median\_purchase\_price - Ratio of purchased price transaction to median purchase price.

repeat\_retailer - Is the transaction happened from same retailer.

used\_chip - Is the transaction through chip (credit card).

used\_pin\_number - Is the transaction happened by using PIN number.

online\_order - Is the transaction an online order.

fraud - Is the transaction fraudulent.

## 2. Data Preprocessing

\*\*Data Cleaning\*\*: [Describe the steps taken to clean the dataset, handling missing values, outliers, etc.]

Steps:

* Converted the numerical columns and categorical columns to its lower byte representation for memory efficiency. At the end of this process the file size has been reduced from 76.3 MB to 16.2 MB.
* Verified the Number of missing values in the dataset.
* Analysed the outliers in the numerical columns using Box Plot and Histogram Plot Visuals from Seaborn Library.
* Calculated the IQR (Inter Quartile Range) of each numerical column and capped the outliers to the lower and upper limit.

\*\*Data Transformation\*\*: [Explain any transformations performed on the data (e.g., normalization, encoding categorical variables, etc.)]

Steps:

* Performed normalization of Numerical features using Standard Scaler from Scikit-learn Library.

## 3. Exploratory Data Analysis (EDA)

Plotted multiple visualization charts including:

* Histograms
* Count Plots
* Heatmap
* Scatter Plot
* Pie Chart
* Box Plot

To analyse the distribution of each feature and interaction of different features to each other.

\*\*Summary Statistics\*\*:

* The average distance from home for a transaction is 26.52 Miles. 50% of the Transactions happened within the radius of 9 miles from the home.
* The average distance of last transaction varies around 5 Miles from the previous transaction. 75% of the transactions happened within the radius of 3 Miles from the previous transactions.
* The average ratio of current transaction to median of the purchase transaction is around 1.82.

\*\*Visualization\*\*: [Include relevant visualizations to help understand the data distribution, relationships, and patterns.]

Insights from Exploratory Data Analysis:

## 4. Feature Engineering

\*\*Feature Selection\*\*: [Explain how features were selected or engineered for model training.]

## 5. Model Development

\*\*Model Selection\*\*: [Explain the choice of machine learning algorithm(s) and any ensemble methods used.]

\*\*Training and Validation\*\*: [Detail the process of training and validating the model, including parameter tuning.]

## 6. Model Evaluation

\*\*Evaluation Metrics\*\*: [Specify the metrics used to assess the model's performance (e.g., accuracy, precision, recall, F1-score, etc.)]

\*\*Results\*\*: [Present the model's performance on the validation/test set.]

\*\*Visualizations (if applicable)\*\*: [Include relevant visualizations to support the evaluation.]

## 7. Deployment (if applicable)

\*\*Deployment Platform\*\*: [Explain how the model is deployed (e.g., web app, API, etc.)]

\*\*Instructions for Usage\*\*: [Provide instructions for users on how to interact with the deployed model.]

## 8. Conclusion and Future Improvements

\*\*Summary\*\*: [Summarize the key findings and outcomes of the project.]

\*\*Future Enhancements\*\*: [Discuss potential improvements or extensions to the project.]