# **Capstone Project**

Topic:Credit Card Fraud Detection

Prepared by:**Maimana Kowatly**

batch:February 2024 / **ML C62**

Credit Card Fraud Detection

[**Capstone Project**](#_ptsomuranrfq) **1**

[**Introduction**](#_kpedn0kfcnl2) **1**

[Objective](#_yctpd6eaguzs) 2

[Business Context](#_jxlgvolc32vb) 2

[Understanding the Data](#_mi8z6nw6q295) 2

[Approach and Methodology](#_9rvnfrf1jnc) 2

[1. Exploratory Data Analysis (EDA)](#_d6woc0a463kv) 2

[2. Data Preprocessing](#_8cg680ss394b) 3

[3. Handling Class Imbalance](#_rfgilcy50hx) 3

[4. Model Development](#_c0tokl1bq9d1) 3

[Model Evaluation and Selection](#_m5crf332z7hl) 3

[Why XGBoost with ADASYN?](#_j6ym1e4ukv3d) 4

[Conclusion](#_rtgki3qeobti) 4

# **Introduction**

In an increasingly digital world, the risk of financial fraud has grown substantially. Among the most pressing challenges is credit card fraud — a threat that not only causes monetary losses but also erodes the trust between banks and their customers. This project focuses on developing an effective machine learning model to detect fraudulent credit card transactions with a strong emphasis on recall, as missing a fraudulent case can have serious financial and reputational consequences.

## **Objective**

The goal of this project is to design and evaluate models capable of identifying fraudulent transactions in highly imbalanced datasets. The dataset used contains over 280,000 transactions, of which less than 0.2% are labeled as fraud. This significant imbalance presents a unique challenge, requiring specialized handling during both preprocessing and model training.

## **Business Context**

Credit card fraud has become a global concern, with projected losses reaching up to $30 billion in recent years. The implications go far beyond financial loss — frequent false negatives can impact customer loyalty, while false positives may inconvenience legitimate users. Therefore, an intelligent fraud detection system must strike a delicate balance between accuracy and sensitivity, favoring recall to ensure fraudulent cases are not overlooked.

## **Understanding the Data**

The dataset is anonymized using Principal Component Analysis (PCA) to protect customer information. It includes 28 transformed features (V1 to V28), along with 'Time', 'Amount', and a binary 'Class' label indicating whether a transaction is fraudulent.

* **Time**: Seconds elapsed from the first transaction
* **Amount**: Transaction value
* **Class**: Target label — 1 for fraud, 0 for legitimate

The strong class imbalance in this dataset was the first key challenge addressed in the pipeline.

## **Approach and Methodology**

The project followed a structured workflow:

### **1. Exploratory Data Analysis (EDA)**

EDA helped in understanding the distribution of features and highlighted the extreme skew between fraudulent and legitimate transactions. Visualizations revealed patterns and guided decisions on feature transformation and sampling strategies.

### **2. Data Preprocessing**

The dataset was split into training and testing sets using stratified sampling to maintain the class ratio. Feature scaling was applied to ensure uniformity across model inputs.

### **3. Handling Class Imbalance**

Four different strategies were explored to address the imbalance:

1. **Imbalanced**: Training models on the original skewed data
2. **Balanced (Random Oversampling)**: Duplicating minority class instances
3. **SMOTE (Synthetic Minority Over-sampling Technique)**: Creating synthetic samples from minority class neighbors
4. **ADASYN (Adaptive Synthetic Sampling)**: Similar to SMOTE, but focuses more on harder-to-classify samples

These sampling methods were applied to evaluate how model performance, particularly recall, was impacted under different balancing techniques.

### **4. Model Development**

Three classification algorithms were implemented and evaluated:

* **Logistic Regression**: A reliable, interpretable baseline model
* **Decision Tree**: A non-linear model capturing feature interactions
* **XGBoost**: An advanced boosting algorithm known for handling complex patterns and imbalanced data effectively

Each model was trained under the four sampling methods mentioned above. Performance metrics included ROC AUC, precision, and — most importantly — recall.

## **Model Evaluation and Selection**

Among all combinations tested, the **XGBoost model with ADASYN sampling** consistently outperformed the others.

* **ROC AUC**: Achieved a near-perfect score of **98%**, demonstrating excellent discriminative ability
* **Recall**: Reached **86%**, which was the highest among all models — a critical result given the project's emphasis on catching fraudulent transactions

## **Why XGBoost with ADASYN?**

XGBoost combines the strengths of ensemble learning and gradient boosting, making it robust against noise and well-suited for imbalanced datasets. ADASYN further enhanced this by focusing oversampling efforts on more difficult, borderline samples. This synergy led to better generalization and improved identification of fraud cases.

Prioritizing recall was a deliberate choice. In real-world banking systems, the cost of missing a fraudulent transaction is significantly higher than flagging a legitimate one. Therefore, the ability of the XGBoost + ADASYN combination to maximize recall made it the most compelling solution for this task.

## **Conclusion**

This project demonstrated that thoughtful preprocessing, careful handling of class imbalance, and the selection of a suitable model architecture are all crucial for building an effective fraud detection system. XGBoost paired with ADASYN delivered the best results, balancing high recall with overall model accuracy.

By applying these techniques, financial institutions can greatly enhance their fraud detection capabilities, reduce losses, and protect customer trust in an era where digital transactions are both common and vulnerable