Assignment: Detecting Financial Fraud using Machine Learning

1. Assignment Overview

The cost of financial fraud globally is estimated in the billions. In this assignment, you will assume the role of a Senior Data Scientist at a digital bank. Your task is to build a robust fraud detection system capable of flagging suspicious credit card transactions.

You will face the same challenges real-world analysts face: massive class imbalance, anonymized features, and the trade-off between catching fraud (Recall) and annoying legitimate customers (Precision).

2. The Dataset

We will be using the **Credit Card Fraud Detection Dataset**.

- **Source:** Dataset attached with the assignment.
- **Context:** The dataset contains transactions made by credit cards in September 2013 by European cardholders.
- **Structure:** It contains only numerical input variables which are the result of a PCA transformation (V1, V2, ... V28). The only features which have not been transformed are Time and Amount.
- **The Challenge:** The dataset is highly unbalanced. Out of 284,807 transactions, **only 492 are fraudulent** (0.172%).

3. Technical Requirements

You must submit a Jupyter Notebook (.ipynb) containing the following four sections.

Part 1: Exploratory Data Analysis (EDA) & Preprocessing

Before modelling, you must understand your data.

1. **Visual Inspection:** Visualize the class imbalance.
2. **Feature Analysis:** Analyse the Amount and Time distributions for Fraud vs. Normal transactions. Do fraudulent transactions tend to be larger or smaller?
3. **Correlation:** Check if specific PCA features correlate strongly with the target variable (Class).
4. **Scaling:** The PCA features (V1-V28) are already scaled, but Amount and Time are not. Apply an appropriate scaler to these columns.

Part 2: Unsupervised Learning (Anomaly Detection)

Scenario: Imagine you do not have historical labels for fraud. You must find the "odd ones out" based purely on data structure.

1. **Drop the Label:** Remove the Class column from your training set for this section.
2. **Implement an Unsupervised Model:** Choose any algorithms suitable for outlier detection.
   - *Suggestions:* Isolation Forest, Local Outlier Factor (LOF), or One-Class SVM.
3. **Prediction:** Use the model to predict anomalies on your test set and compare them against the actual labels you set aside.

Part 3: Supervised Learning (Classification)

Scenario: You now utilize the historical labels to train a classifier.

1. **Handle Imbalance:** You **must** implement a technique to address the 0.17% fraud rate.
   - *Options:* SMOTE (Synthetic Minority Over-sampling Technique), Random Undersampling, or using Class Weights in the model.
2. **Implement a Supervised Model:** Choose any classifiers.
   - *Suggestions:* Logistic Regression, Random Forest, XGBoost, or Neural Networks.

Part 4: Evaluation & Comparison (Critical)

CRITICAL: Do not rely on "Accuracy." If your model predicts 100% "Normal," it will have 99.8% accuracy but will be useless for the bank.

1. **Metrics:** Calculate **Precision**, **Recall**, **F1-Score**, and **AUPRC** (Area Under the Precision-Recall Curve).
2. **Confusion Matrix:** Plot the confusion matrix for both your Supervised and Unsupervised models.
3. **Commentary:** In markdown cells, explain which model performed better. Discuss the trade-off: Which is more important for the bank—catching every fraud (high Recall) or minimizing false alarms (high Precision)?

4. Deliverables

You are required to submit two files:

1. The Jupyter Notebook (StudentID_Name.ipynb)

- **Code:** Must be runnable and error-free.
- **Documentation:** Use Markdown cells to explain your logic. If you choose a specific hyperparameter (like contamination=0.01), explain *why*.
- **Conclusion:** A final paragraph summarizing your findings.

2. The Executive Presentation (StudentID_Name.pdf or .pptx)

- **Length:** 5-7 Slides.
- **Audience:** Non-technical stakeholders (e.g., The Bank's Vice President of Risk).
- **Content:**
  - **Slide 1:** Title & Executive Summary.
  - **Slide 2:** The Data Story (What patterns did you find?).
  - **Slide 3:** Methodology (Unsupervised vs Supervised approach).
  - **Slide 4:** Results (Show the Confusion Matrices and key metrics).
  - **Slide 5:** Recommendation (Which model should the bank deploy and why?).

Hints

- **Hint 1:** For the Unsupervised model (e.g., Isolation Forest), be careful with the contamination parameter. It represents the expected proportion of outliers in the data set.
- **Hint 2:** When splitting your data for the Supervised model, ensure you use stratify=y to maintain the same percentage of fraud in both training and testing sets.
- **Hint 3:** In your presentation, do not show raw code. Show charts, business impact, and concepts like "Money Saved vs. Customers Annoyed."