Phase-2 Submission

**Student Name : Thamiyazh . M**

**Register Number:** 622623205053

**Institution:** Shreenivasa Engineering college **Department:** B.Tech - Information technology **Date of Submission:** [17-05-2025]

## Github Repository Link:https://github.com/ramyamssm/Card

1. **Problem Statement :**

The increasing volume of online financial transactions has led to a corresponding surge in credit card fraud cases. This project aims to develop an AI- powered system to detect and prevent fraudulent credit card transactions in real time. The problem is a binary classification task — predicting whether a transaction is fraudulent or legitimate. Solving this issue has direct applications in banking and e-commerce sectors, where fraud prevention is critical to financial stability and customer trust.

## Project Objectives :

Develop machine learning models to accurately classify transactions as fraudulent or not.Improve model interpretability for better insights into fraud patterns.

Ensure high precision and recall, minimizing false positives and negatives. Integrate data-driven decision-making for real-time fraud alerts.

## Flowchart of the Project Workflow :

(Use a diagram in your final report. Steps should include:)

Data Collection → Data Cleaning → Exploratory Data Analysis → Feature Engineering → Model Selection → Training & Testing → Evaluation → Insights & Visualization

## Data Description :

**Source:** Kaggle - Credit Card Fraud Detection Dataset

**Type:** Structured data (CSV)

**Records:** ~284,807 transactions

**Features:** 30 (28 anonymized numerical inputs, Time, Amount, Class) Static Dataset

**Target Variable:** Class (0 = Legitimate, 1 = Fraudulent)

## Data Preprocessing :

**Missing Values:** No missing values in dataset.

**Duplicates:** Removed ~100 duplicate rows.

**Outliers:** Handled using IQR-based filtering for Amount and Time.

**Encoding:** Not required, all features are numeric.

**Normalization:** StandardScaler applied to Amount and Time to match scaled anonymized features.

## Exploratory Data Analysis (EDA) :

**Univariate:** Histograms showed high skew for Amount; Class is highly imbalanced (fraud < 0.2%).

**Bivariate:** Strong correlation of some PCA features with fraud class.

**Multivariate:** PCA plots and t-SNE visualizations indicated separability in certain dimensions.

**Insights:**Fraudulent transactions typically have lower amounts. Time of transaction affects likelihood of fraud.

# Feature Engineering :

Created HourOfDay from Time to reflect transaction timing. Used Amount\_log = log(Amount + 1) to reduce skew.

Dropped Time after transformation.

Dimensionality reduction (optional): PCA not applied since features are already anonymized via PCA.

# Model Building :

**Models implemented:**

**Logistic Regression:** Baseline classifier.

**Random Forest Classifier:** Handles imbalance and nonlinear patterns.

**XGBoost:** For high performance and feature importance analysis.

**Data Split:** 80% train, 20% test with stratified sampling.

# Performance Metrics:

**Logistic Regression:** AUC = 0.96, Recall = 0.70

**Random Forest:** AUC = 0.99, Recall = 0.87

**XGBoost:** AUC = 0.998, Recall = 0.919. Visualization of Results & Model Insights

**Confusion Matrix:** Displayed for all models to evaluate FP/FN rates. **ROC Curve:** XGBoost and Random Forest showed highest AUC. **Feature Importance:** Identified top 5 contributing PCA features.

**Precision-Recall Curve:** Helped tune decision threshold for fraud detection.

# Tools and Technologies Used :

**Programming Language:** Python

**IDE:** Jupyter Notebook

**Libraries:** pandas, numpy, scikit-learn, seaborn, matplotlib, xgboost

**Visualization:** seaborn, matplotlib, plotly

# Team Members and Contributions

**M.s.s. Ramya:** Data preprocessing, Feature Engineering

**Priyadarshini,Sona:** EDA, Model Development

**Thamiyazh . M :** Model Evaluation, Report Writing & GitHub Integration