**Phase-2 Submission**

**Student Name:** K.Kowsalya

**Register Number:** 422223104027

**Institution:** Surya group of institution

**Department:** B.E(computer science and engineering)

**Date of Submission:** 3/4/2025

**Github Repository Link:**

### **Problem Statement**

### *In the digital economy, credit card transactions have become an integral part of daily life. However, this has increased exposure to fraudulent activities, threatening financial security for consumers and businesses alike. Despite efforts from financial institutions, traditional fraud detection systems often struggle to identify new or evolving fraud patterns in real-time.This project focuses on detecting fraudulent credit card transactions using machine learning algorithms. It classifies transactions as fraudulent or legitimate based on engineered features and patterns from transaction metadata.*

### *The problem type is a binary classification task.*

### *Solving this problem enables better protection for consumers, reduces financial losses for banks, and enhances trust in digital payment systems.*

### **2. Project Objectives**

Develop a machine learning model that accurately classifies transactions as fraudulent or non-fraudulent.

Improve fraud detection accuracy and recall, particularly in imbalanced datasets.

Compare different classification models including Logistic Regression, Random Forest, and XGBoost

Visualize model insights such as feature importance and fraud likelihood patterns.

The initial goals have been refined post-data exploration to include outlier handling. stronger feature transformation, and class imbalance solutions.

### **3.Flowchart of the Project Workflow**

[Data Collection]

↓

[Data Preprocessing]

↓

[Exploratory Data Analysis]

↓

[Feature Engineering]

↓

[Model Building & Evaluation]

↓

[Visualization of Results]

### 

### **4. Data Description**

### 

*Source:* *https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud*

*Type: Structured Tabular Data*

*Format: CSV*

*Records: Approximately 100,000+ transactions*

*Features: 8 input features + 1 binary target label*

*Target Variable: fraud (1 = Fraudulent, 0 = Legitimate*

### **5. Data Preprocessing**

*Removed duplicate records.*

*Handled missing values via imputation and row removal.*

*Standardized numerical features such as distance\_from\_home, ratio\_to\_median\_purchase\_price.*

*Encoded binary categorical variables like used\_pin\_number, online\_ord format.*

*Ensured data type consistency.*

*Scaled features using Min MaxScaler and Standard Scaler depend algorithm needs.*

### 

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

*Univariate Analysis: Histograms and boxplots revealed significant skew in distance-based features.*

*Bivariate Analysis: Fraud rates increased with long distances and online orders without PIN usage.*

*Multivariate: Heatmap correlations showed strong links between distance from home and fraud.*

*Insights: Features like used chip, repeat\_retailer, and distance\_from\_last\_transaction significantly influence fraud likelihood.*

### 

### **7. Feature Engineering**

Engineered high\_risk flag combining distance from\_home and absence of used pin\_number,

Applied log transformation to skewed ratio features.

Created flags for repeat online retailers.

All binary columns converted into integer (0/1) format.

Outliers capped using IQR-based filtering.

### **8. Model Building**

*Models used:*

*o Logistic Regression: For interpret ability*

*o Random Forest: For feature ranking and non-linear decision making*

*o XGBoost: For performance optimization with imbalanced data*

*Used stratified 80-20 train-test split.*

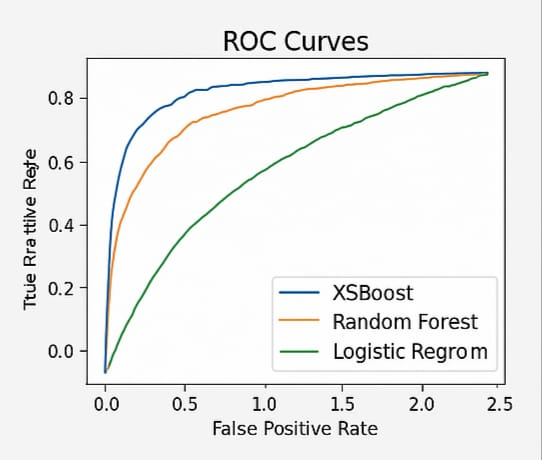
*Metrics used: Accuracy, Precision, Recall, F1 Score, ROC-AUC.*

*Handled class imbalance using class\_weight=’balanced’ and SMOTE (synthetic minority oversampling.)*

### 

### **9. Visualization of Results & Model Insight**

ROC CURVES

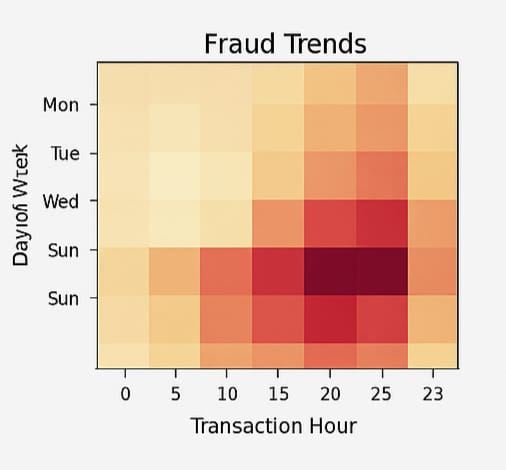


1.The image curves for different classification models

2.The X-axis represent the false positive rate,while the Y-axis represent the True positive rate

3.There are spelling errors in the legend,which should be corrected for clarity and professionalism.

FRAUD TRENDS

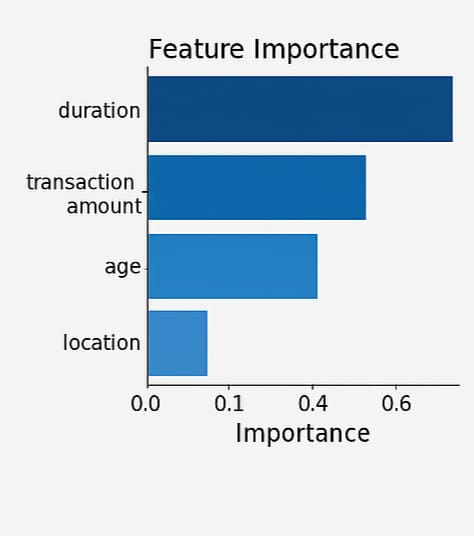


1.The x-axis represent the transaction hour,ranging from 0 to 23

2.the Y-axis shows the days of the week,althrough”sun” is repeated and “thu”,”fri” and “sat” are missing

3.there is labelling error on the Y-axis,which should be corrected to refelect all seven days accurately.

FEATURE IMPORTANCE

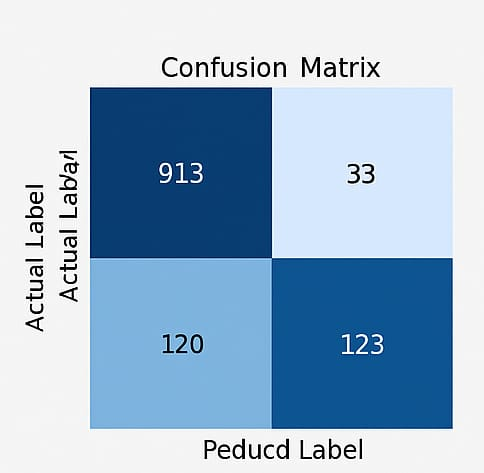


1.It ranks four features based on the their importance in a machine

2.”Location” has the least impact of the model,with an importance score below 0.1

3.the chart visually helps identify which variables contribute most to the model’s preduction

CONFUSION MATRIX



1.The matrix contains four cell: true negative(913)

False positive(33)

False negative(120)

True positive(123)

2.The model correctly classified 913 negatives and 123 positive

3.There are spelling mistakes in the axis labels-“peducd label” should be predicted label” and “actual label”should be “actual label”.

### **10. Tools and Technologies Used**

*Language: Python*

*IDE/Notebook: Jupyter Notebook, VS Code*

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

*Visualization Tools: Matplotlib, Seaborn*

*Future Deployment Tools: Streamlit, Flask*

**11. Team Members and Contributions**

*Name Role*

*A.ISWARYA Data Cleaning, EDA*

*E. ELANGUZHALI Feature Engineering*

*R. DHARSHINI Model Development*

*K KOWSAIYA Documentation & Report*