G H Patel College of Engineering & Technology

(The Charutar Vidya Mandal (CVM) University)

Vallabh Vidyanagar

COMPUTER ENGINEERING DEPARTMENT

AI-ML report on

Credit Card Fraud Detection System

Submitted By

Name of Student : Ruchi Tandel

Enrollment Number : 12202040501055

Name of Student : Dev Manojkumar Ray

Enrollment Number : 12202040501069

Guided By

Dr. Priyang Bhatt

Artificial Intelligence and Machine Learning

(202040601)

A.Y. 2024-25 EVEN TERM

Objectives:

The primary objective of this project is to design and implement a robust, scalable, and intelligent machine learning system capable of detecting fraudulent credit card transactions in real time. The system is built with the following goals:

- **Real-Time Fraud Detection**: Enable immediate analysis of each transaction as it occurs, helping financial institutions take quick action to prevent losses and protect users.
- **Model Integration**: Combine the strengths of **Random Forest** and **XGBoost** classifiers to boost detection performance, reduce false positives, and improve model reliability.
- **Advanced Feature Engineering**: Apply geospatial analysis—like calculating the distance between cardholder and merchant—to identify abnormal behavior and enhance model accuracy.
- **Effective Data Preprocessing**: Use encoding techniques (label encoding, one-hot encoding) and handle data challenges such as missing values and class imbalance, ensuring consistency across features.
- **Scalability and Adaptability**: Build a system that can manage high volumes of transaction data and adapt to evolving fraud trends through regular model evaluation and updates.
- **User-Friendly Streamlit Interface**: Develop an interactive web app that allows users to input transaction details and receive instant predictions, with clear output and intuitive design.
- **Decision Support Tool**: Support financial institutions by providing actionable insights and helping prioritize which transactions to review, making fraud investigations more efficient.

Dataset Used:

Dataset Name: Credit Card Transactions Dataset

Source: https://drive.google.com/file/d/1118Jwzj51KpXd0T5jiebn9ykCygwbkhn/view

Description:The dataset contains historical credit card transactions with labeled outcomes (fraudulent or legitimate). It includes various features such as transaction amount, time of transaction, merchant details, and geolocation data.

Preprocessing Steps:

- Handling missing values and outliers
- Encoding categorical variables using label encoders and one-hot encoding
- Feature scaling and alignment to the expected model input format

Model Chosen:

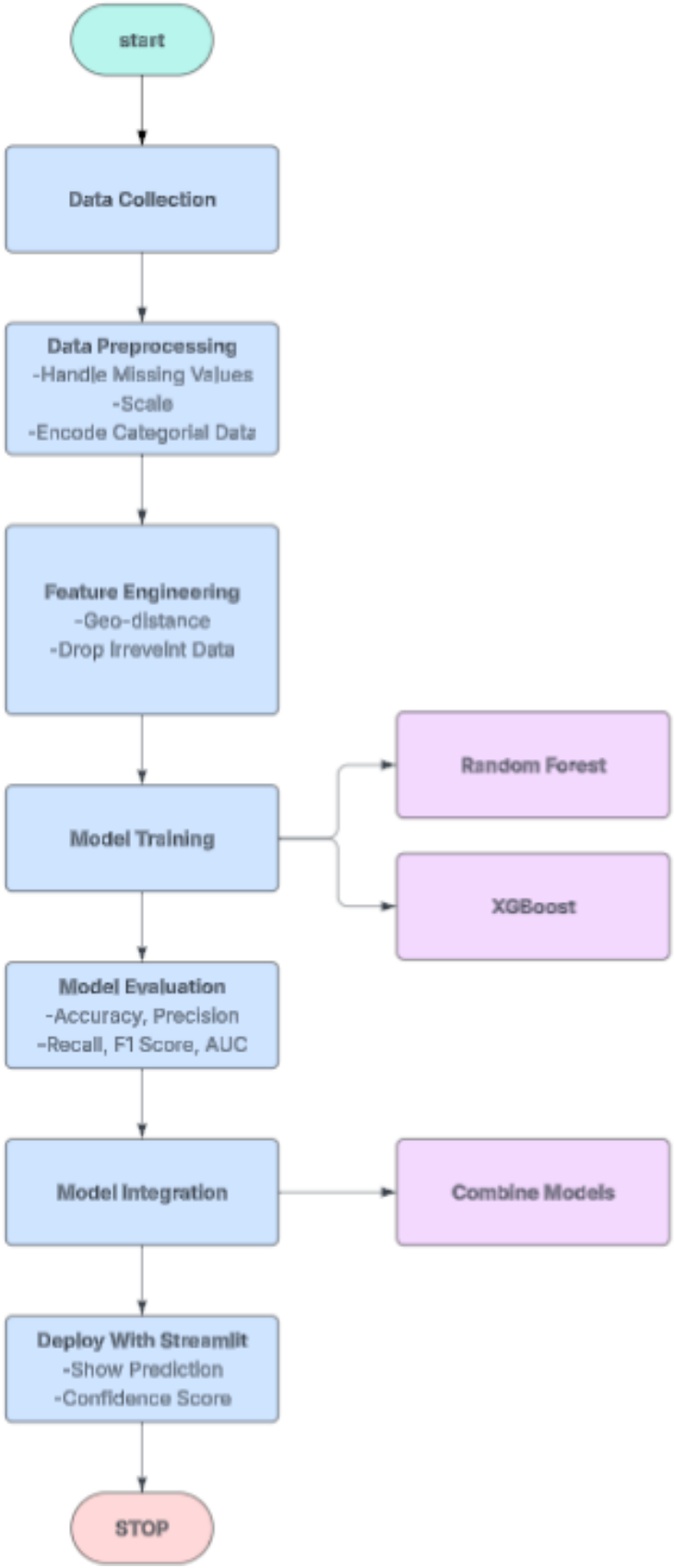
Random Forest Classifier

- Ensemble Method: Combines multiple decision trees to improve accuracy and reduce overfitting.
- Robustness & Interpretability: Handles numerous features effectively and provides insights through feature importance scores.
- Hyperparameter Tuning: Parameters such as tree depth and number can be adjusted to optimize performance.

XGBoost Classifier

- Gradient Boosting: Sequentially builds trees to correct previous errors, resulting in high accuracy.
- Efficiency: Optimized for speed and scalability, making it suitable for real-time fraud detection.
- Built-in Handling: Manages missing data and reduces overfitting through regularization techniques.

Flowchart:


start
Data Collection
Data Preprocessing
-Handle Missing Values
-Scale
-Encode Categorial Data
Feature Engineering
-Geo-distance
-Drop Irrevelnt Data
Model Training
Random Forest
XGBoost
Model Evaluation
-Accuracy, Precision
-Recall, F1 Score, AUC
Model Integration
Combine Models
Deploy With Streamlit
-Show Prediction
-Confidence Score
STOP

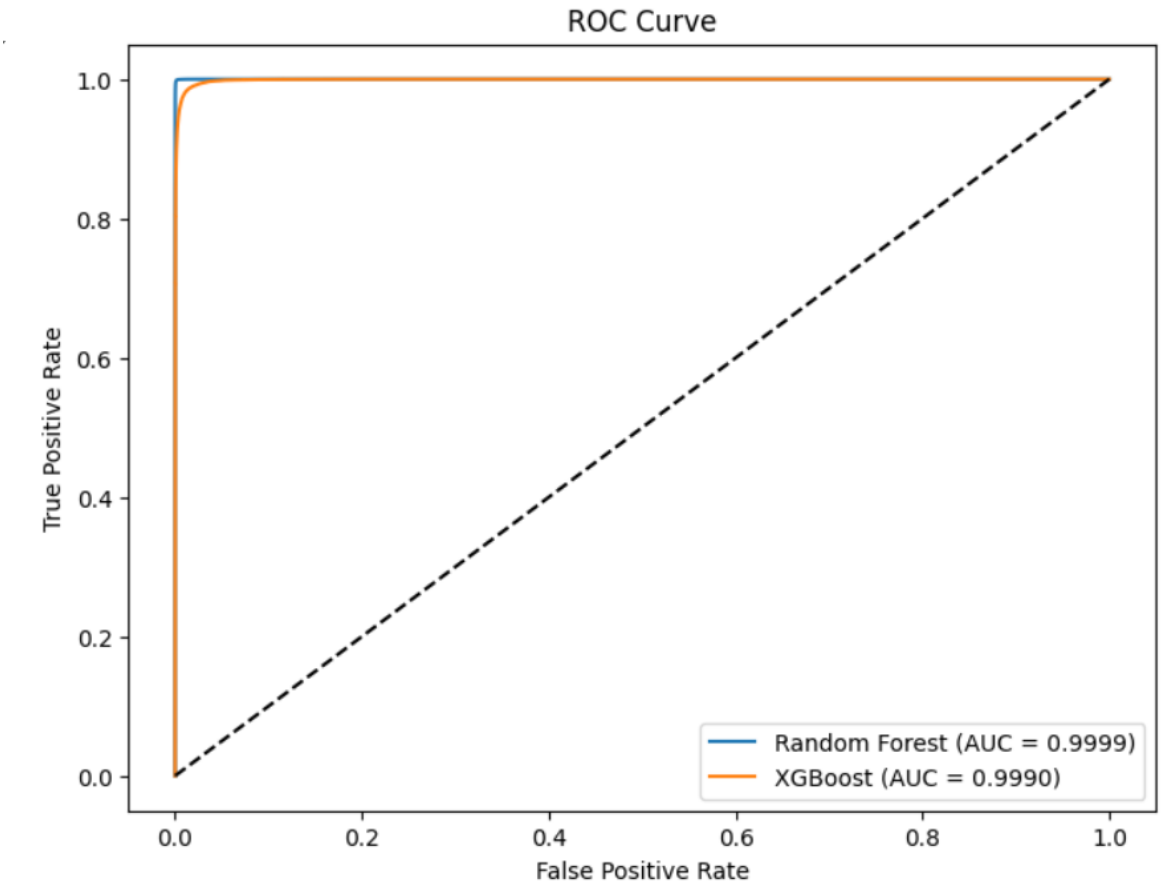
Performance Metrics:

To evaluate models effectively, we used several key performance metrics:

- **Accuracy**:
  Measures the overall proportion of correct predictions. It provides a general sense of model performance but may be misleading in cases of class imbalance.
- **Precision**:
  Indicates the fraction of correctly identified fraudulent transactions out of all transactions flagged as fraud. High precision means fewer false positives, which is critical in reducing unnecessary alerts and investigations.
- **Recall (Sensitivity)**:
  Reflects the proportion of actual fraudulent transactions that the model successfully detects. High recall minimizes false negatives, ensuring that most fraudulent cases are caught.
- **F1-Score**:
  The harmonic mean of precision and recall. This metric provides a balanced measure, especially useful when dealing with imbalanced datasets, by combining both false positives and false negatives into one number.
- **AUC (Area Under the ROC Curve)**:
  Evaluates the model's ability to differentiate between fraudulent and legitimate transactions across various threshold settings. A higher AUC indicates better overall model discrimination.

```
Random Forest Performance:
Accuracy: 0.9981
Precision: 0.9970
Recall: 0.9993
F1 Score: 0.9981
Confusion Matrix:
 [[256397    789]
 [   180 258302]]

XGBoost Performance:
Accuracy: 0.9855
Precision: 0.9875
Recall: 0.9835
F1 Score: 0.9855
Confusion Matrix:
 [[253978   3208]
 [  4276 254206]]
```



Challenges & Learnings :

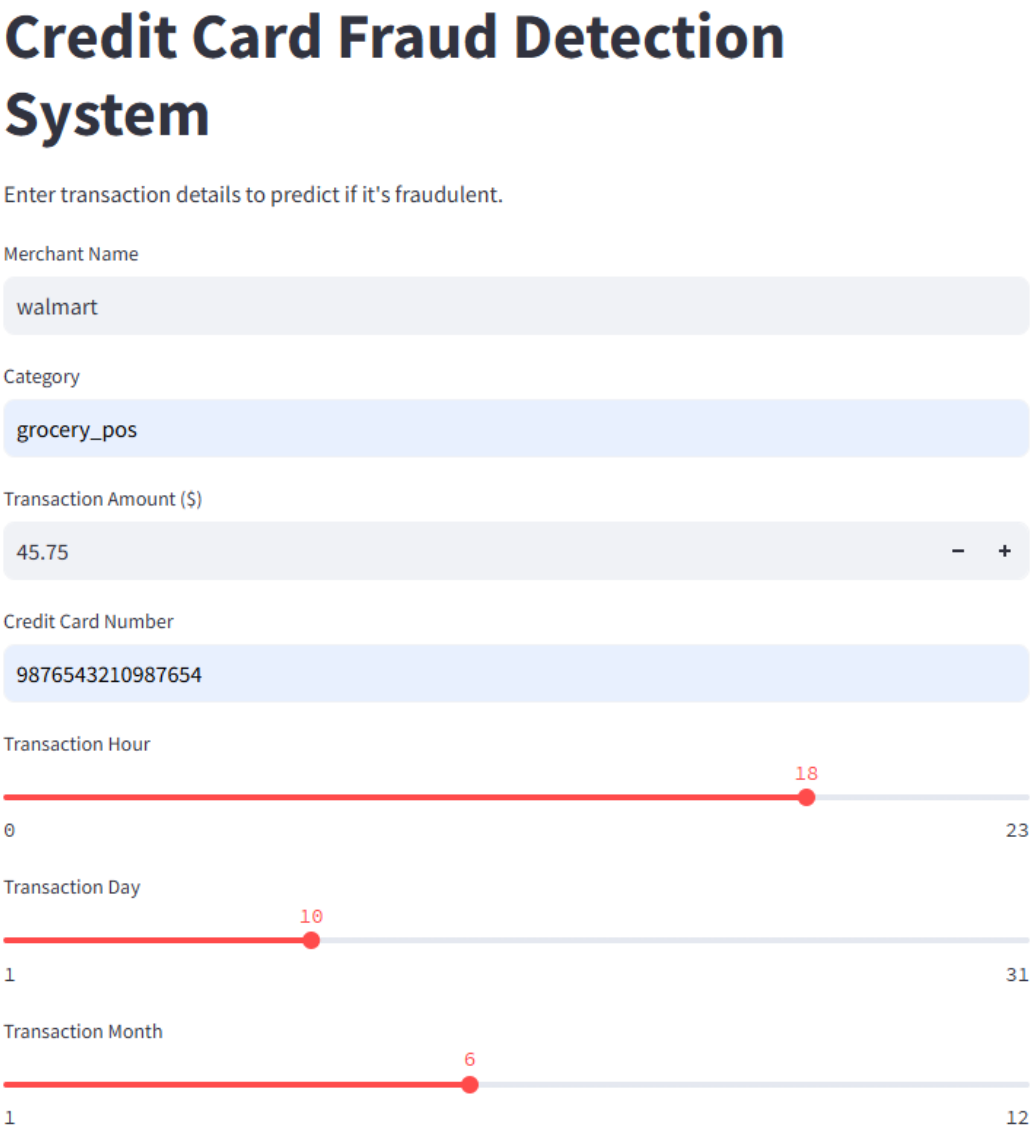
Challenges

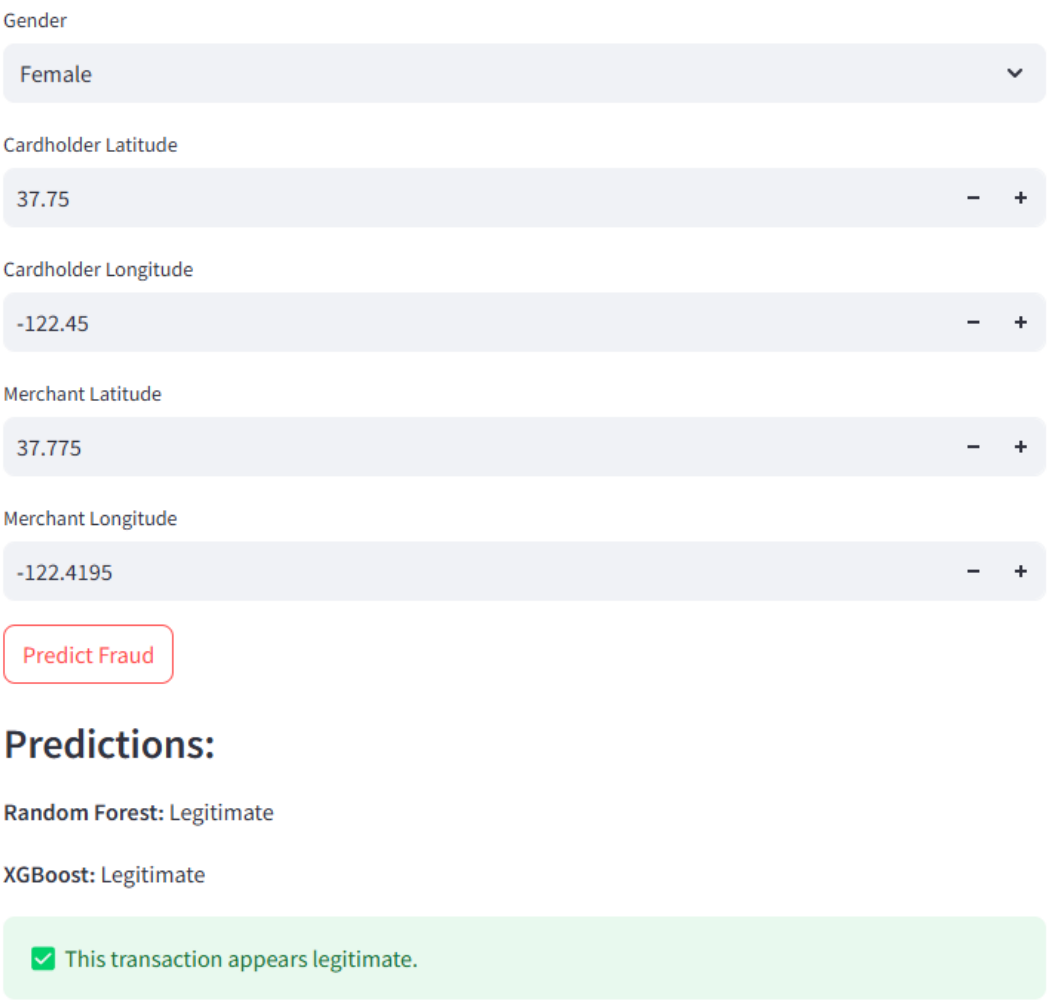
- Data Imbalance: The dataset is highly imbalanced, with fraudulent transactions being a minority. This required special techniques like oversampling or using evaluation metrics that better reflect the performance on the minority class.
- Feature Engineering: Creating meaningful features from the raw data, such as calculating the geographical distance between cardholder and merchant, was challenging but proved crucial for the detection process.
- Model Integration: Combining predictions from multiple models (Random Forest and XGBoost) and ensuring the input feature order consistency added complexity to the deployment pipeline.
- Real-Time Prediction: Ensuring that the system can provide real-time fraud predictions while maintaining high accuracy and low latency.

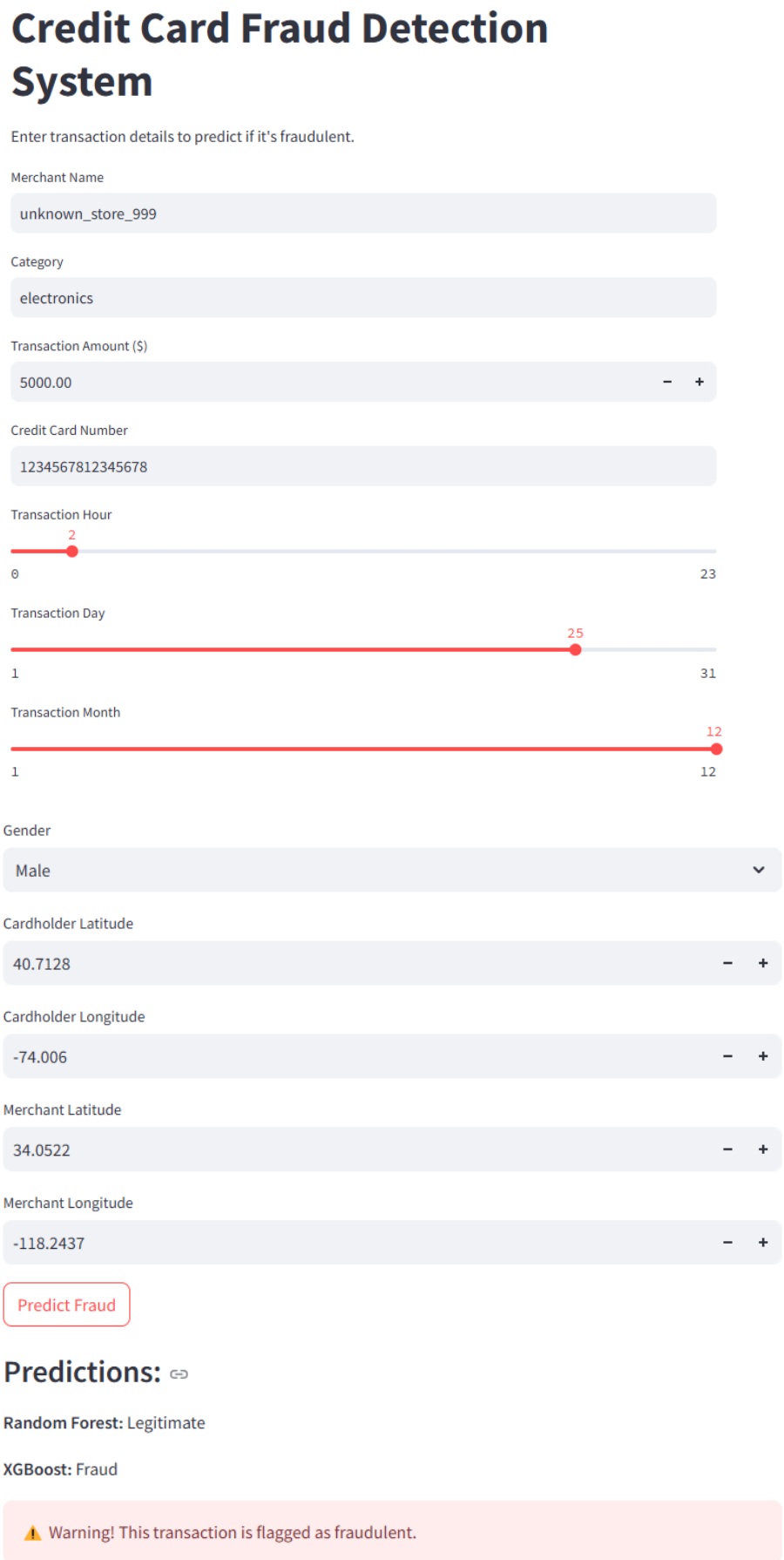
Learnings

- Preprocessing Importance: Proper data cleaning, encoding, and feature alignment are fundamental for model performance.

- Model Comparison: Testing multiple models provided insights into the strengths and weaknesses of different approaches.
- Deployment: Using Streamlit for a user-friendly interface demonstrated how powerful visualization and interactivity can be integrated with machine learning models.
- Error Handling: Building robust error-handling mechanisms and validating input data are essential steps in creating a production-ready system.




Gender
Female
Cardholder Latitude
37.75
Cardholder Longitude
-122.45
Merchant Latitude
37.775
Merchant Longitude
-122.4195
Predict Fraud
Predictions:
Random Forest: Legitimate
XGBoost: Legitimate
This transaction appears legitimate.


Credit Card Fraud Detection System
Enter transaction details to predict if it's fraudulent.
Merchant Name
unknown_store_999
Category
electronics
Transaction Amount ($)
5000.00
Credit Card Number
1234567812345678
Transaction Hour
2
0
23
Transaction Day
25
1
31
Transaction Month
12
1
12
Gender
Male
Cardholder Latitude
40.7128
Cardholder Longitude
-74.006
Merchant Latitude
34.0522
Merchant Longitude
-118.2437
Predict Fraud
Predictions:
Random Forest: Legitimate
XGBoost: Fraud
Warning! This transaction is flagged as fraudulent.