### Fraud Detection in Credit Card Transactions

**1. Introduction**

With the increasing volume of credit card transactions, traditional methods of detecting fraudulent activities have become inadequate. This project, undertaken by **XPACE TECHNOLOGIES Pvt Ltd** in partnership with a major financial institution, aims to develop a machine learning model capable of accurately predicting fraudulent credit card transactions.

**2. Objective**

The primary objectives of this project are:

* To understand the patterns that differentiate fraudulent transactions from legitimate ones.
* To build a predictive model that can effectively detect fraud.
* To provide actionable insights and recommendations for improving the fraud detection process.

**3. Dataset Description**

The dataset provided contains the following columns:

* **Transaction ID**: Unique identifier for each transaction.
* **Customer ID**: Unique identifier for each customer.
* **Transaction Date**: The date and time of the transaction.
* **Transaction Amount**: The monetary value of the transaction.
* **Merchant**: Name of the merchant where the transaction occurred.
* **Location**: The transaction location.
* **Transaction Type**: Type of transaction (e.g., online purchase, in-store purchase).
* **Card Type**: Type of credit card used (e.g., Visa, MasterCard).
* **Is Fraudulent**: Binary indicator of whether the transaction was fraudulent (Yes/No).

**4. Methodology**

**4.1 Data Understanding and Preprocessing**

* **Data Exploration**:  
  Analyzed the distribution of fraudulent vs. legitimate transactions.
* **Handling Missing Data**:  
  Rows with missing data were dropped.
* **Categorical Encoding**:  
  Categorical variables (Merchant, Location, Transaction Type, Card Type) were converted into numerical representations using LabelEncoder.
* **Balancing the Dataset**:  
  SMOTE (Synthetic Minority Over-sampling Technique) was used to handle the imbalance between fraudulent and legitimate transactions.

**4.2 Feature Engineering**

New features were created to enhance model prediction:

* **Frequency of transactions by a customer**.
* **Average transaction amount**.
* **Time of day when transactions occur**.

**4.3 Model Building**

* **Data Splitting**:  
  The dataset was split into training (80%) and testing (20%) sets.
* **Model Training**:  
  Various machine learning models were trained:
  + Logistic Regression
  + Random Forest
  + Gradient Boosting
* **Model Evaluation**:  
  The models were evaluated using:
  + Confusion Matrix
  + Classification Report (accuracy, precision, recall, F1-score)
  + ROC Curve and AUC-ROC for performance metrics.

**5. Results**

**5.1 Model Evaluation Metrics**

**Confusion Matrix for Logistic Regression**  
The confusion matrix for Logistic Regression shows the following:

* **True Positives (Fraudulent correctly predicted)**: 21
* **True Negatives (Legitimate correctly predicted)**: 20
* **False Positives (Legitimate incorrectly predicted as Fraudulent)**: 16
* **False Negatives (Fraudulent incorrectly predicted as Legitimate)**: 17

**Classification Report for Logistic Regression**

* **Precision for Legitimate transactions (class 0)**: 0.54
* **Recall for Legitimate transactions**: 0.56
* **F1-Score for Legitimate transactions**: 0.55
* **Precision for Fraudulent transactions (class 1)**: 0.57
* **Recall for Fraudulent transactions**: 0.55
* **F1-Score for Fraudulent transactions**: 0.56
* **Overall Accuracy**: 55%
* **AUC-ROC**: 0.53 (marginally better than random guessing)

**ROC Curve and AUC-ROC for Logistic Regression**  
The ROC curve shows an **AUC-ROC score of 0.53**, which reflects a weak ability to distinguish between fraudulent and legitimate transactions.

**5.2 Feature Importance**

**Logistic Regression Feature Importance**:

* **Transaction Type**: The most important feature, with the highest absolute coefficient value.
* **Card Type**: Second in importance, contributing moderately to fraud detection.
* Other features such as **Merchant**, **Transaction Amount**, and **Location** have minimal impact on the model’s prediction.

5.3 **Confusion Matrix for Random Forest:**  
The confusion matrix for Logistic Regression shows the following:

* **True Positives** : 20
* **True Negatives :** 21
* **False Positives** : 15
* **False Negatives :** 18

**Random Forest:**

**Classification Report for Random Forest**:

* **Precision for Legitimate transactions (class 0)**: 0.54
* **Recall for Legitimate transactions**: 0.58
* **F1-Score for Legitimate transactions**: 0.56
* **Precision for Fraudulent transactions (class 1)**: 0.57
* **Recall for Fraudulent transactions**: 0.53
* **F1-Score for Fraudulent transactions**: 0.55
* **Overall Accuracy**: 55%
* **AUC-ROC**: 0.55 (slightly better than logistic regression, indicating moderate improvement in distinguishing between fraudulent and legitimate transactions).

**5.2 Feature Importance**

**Logistic Regression Feature Importance**:

* **Transaction Amount**: The most important feature, with the highest absolute coefficient value.
* **Card Type**: Second in importance, contributing moderately to fraud detection.
* Other features such as **Merchant**, **Transaction Type**, and **Location** have minimal impact on the model’s prediction.

**5.3 Feature Importance**

**Logistic Regression Feature Importance**:

* **Transaction Amount**: The most important feature, with the highest absolute coefficient value.
* **Location**: Second in importance, contributing moderately to fraud detection.
* **Merchant**: The type of merchant where the transaction occurred also plays a vital role.
* **Card Type** : The type of credit card used has a moderate influence on predicting fraud.
* **Transaction Type**: The type of transaction (e.g., online purchase, in-store purchase, ATM withdrawal) shows the least importance in predicting fraud.

5.4 **Gradient Boosting:**

**Confusion Matrix for Logistic Regression**  
The confusion matrix for Logistic Regression shows the following:

* **True Positives (Fraudulent correctly predicted)**: 21
* **True Negatives (Legitimate correctly predicted)**: 17
* **False Positives (Legitimate incorrectly predicted as Fraudulent)**: 19
* **False Negatives (Fraudulent incorrectly predicted as Legitimate)**: 17

**Classification Report for Gradient Boosting**:

* **Precision for Legitimate transactions (class 0)**: 0.50
* **Recall for Legitimate transactions**: 0.47
* **F1-Score for Legitimate transactions**: 0.49
* **Precision for Fraudulent transactions (class 1)**: 0.53
* **Recall for Fraudulent transactions**: 0.55
* **F1-Score for Fraudulent transactions**: 0.54
* **Overall Accuracy**: 51%
* **AUC-ROC**: 0.53

**Logistic Regression Feature Importance**:

* **Location**: The most important feature, with the highest absolute coefficient value.
* **Transaction Amount**: Second in importance, contributing moderately to fraud detection.
* **Merchant**: The type of merchant where the transaction occurred also plays a vital role.
* **Card Type and Transaction Type**: The type of credit card used has a moderate influence on predicting fraud.

**6. Recommendations**

Based on the model evaluation, the following recommendations are provided:

1. **Real-Time Fraud Detection System**:  
   Implement a real-time fraud detection system leveraging the developed model, ensuring it can process transactions in real-time and flag suspicious activities for review.
2. **Continuous Learning**:  
   Establish a continuous learning mechanism where the model can adapt to new fraud patterns by retraining with recent transaction data.
3. **Integration with Existing Systems**:  
   Integrate the fraud detection model with existing transaction processing systems to enhance the institution's overall fraud detection capabilities.
4. **Regular Audits and Updates**:  
   Conduct regular audits of the model’s performance and update it based on the latest transaction trends and patterns observed in the data.
5. **User Education**:  
   Educate customers on identifying fraudulent activities and reporting them quickly to minimize financial loss.

**7. Conclusion**

The project successfully developed a machine learning model for detecting fraudulent credit card transactions, providing valuable insights and recommendations for improving the fraud detection process. Although the Logistic Regression model demonstrated limited performance (AUC-ROC of 0.53), additional models like Random Forest or Gradient Boosting could offer better results. Implementing the recommended strategies will enhance the institution’s ability to detect and prevent fraudulent transactions, ultimately safeguarding both its customers and its reputation.