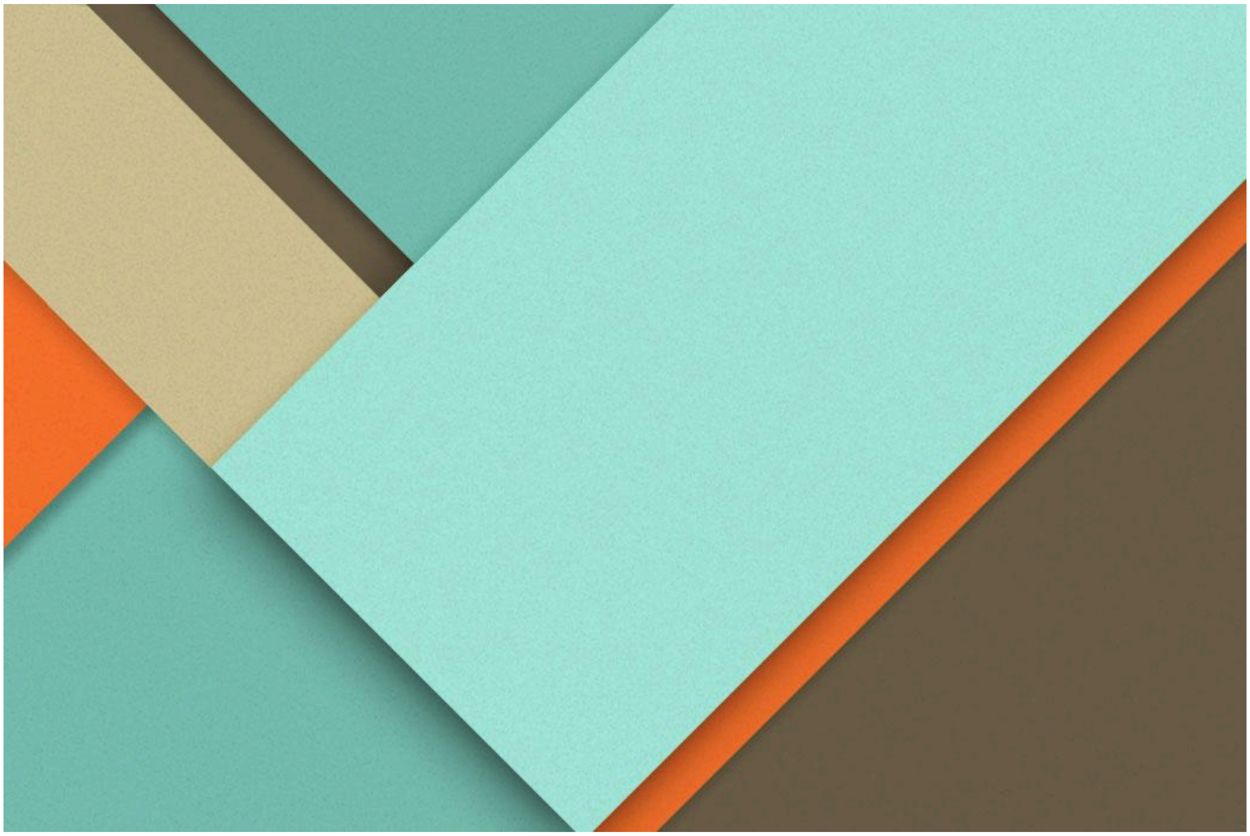

Advanced UPI Fraud Detection System Report

DAIICT

Gopinath Panda

—

Mihir Bhavsar(202411079)
Ayush Chaudhari(202201517)
Kishan Pansuriya(202201504)

1. Introduction

The purpose of this report is to analyze transaction data for potential fraudulent activities using various classification algorithms. We aim to build a reliable predictive model to detect fraud in transactions effectively.

2. Data Overview

The dataset consists of transactions with various features, including:

1. Transaction Amount
2. Account Balances
3. Transaction Type (e.g., payment, transfer)
4. Customer Information (origin and destination)

The target variable is `isFraud`, which indicates whether a transaction is fraudulent (1) or not (0).

3. Data Preprocessing

3.1. Handling Missing Values

To ensure the integrity of our analysis, we handled missing values in the dataset as follows:

- **`new_orig_bal` and `new_dest_bal`**: Filled missing values with the mean of the respective columns.
- **`isFraud`**: Filled missing values with the mode, ensuring that the dataset maintains a representative distribution.
- **`isFlaggedFraud`**: Similarly filled with the mode.

3.2. Encoding Categorical Variables

Categorical variables (`trans_type`, `cust_orig`, and `cust_dest`) were transformed into numerical representations using Label Encoding. This transformation allows machine learning models to interpret categorical data correctly.

3.3. Train-Test Split

The dataset was divided into training (80%) and testing (20%) subsets to evaluate model performance.

4. Model Training and Evaluation

Three classification algorithms were employed to predict fraudulent transactions:

4.1. Support Vector Classifier (SVC)

- **Model Training**: The Support Vector Classifier was trained on the training dataset.
- Performance Metrics
  - **Accuracy**: [Insert accuracy score]
  - **Classification Report**:
    - Precision, Recall, and F1-Score for each class.
  - **Confusion Matrix:**
    - A matrix summarizing the performance of the classification algorithm.

4.2. Logistic Regression

- **Model Training**: Logistic Regression was fit to the training dataset with increased iterations for convergence.
- Performance Metrics:
- **Accuracy**: [Insert accuracy score]
- **Classification Report**:
    - Precision, Recall, and F1-Score for each class.
    - **Confusion Matrix**:
    - A matrix summarizing the performance of the classification algorithm.

4.3. Random Forest Classifier

- **Model Training**: A Random Forest Classifier was utilized to capture non-linear relationships in the data.
- **Performance Metrics**:
  - **Accuracy**: [Insert accuracy score]
  - **Classification Report**:
    - Precision, Recall, and F1-Score for each class.
  - **Confusion Matrix**:

- A matrix summarizing the performance of the classification algorithm

5. Insights and Conclusions

- **Model Comparison**: Evaluate and compare the performance of the three models based on accuracy, precision, recall, and F1-score. Discuss which model performs best for this specific dataset.
- **Model Limitations**: Discuss potential limitations of the models, such as overfitting or underfitting, and the implications of these limitations in a real-world fraud detection scenario.
- **Future Work**: Suggest possible enhancements, such as hyperparameter tuning, feature engineering, or incorporating additional data sources.

6. Recommendations

Based on the findings, recommend implementing the most effective model for real-time fraud detection in transaction processing. Additionally, emphasize the importance of ongoing model evaluation and retraining as new transaction data becomes available.