Project 2: Fraud Detection

Prepared by: Thiru Sudar S L

Role: Machine Learning Intern

Abstract

This project aims to build a machine learning model to classify whether a transaction is fraudulent or legitimate based on a simulated dataset. The dataset contains core transaction details and incorporates specific fraud scenarios to validate the effectiveness of the implemented fraud detection technique. The Random Forest model achieved an accuracy of 100%, demonstrating its capability to detect fraudulent transactions accurately. This report outlines the methodologies, results, and key insights from the project.

Introduction

Fraudulent transactions pose significant risks to financial institutions and consumers alike. Early detection and classification of such transactions can mitigate losses and enhance security measures. This project leverages machine learning techniques to identify fraudulent transactions using a simulated dataset that reflects common fraud patterns.

Objective

To develop a system that classifies transactions as fraudulent or legitimate using machine learning algorithms.

Dataset Description

The dataset is a simulated collection of original and fraudulent transactions, featuring the following columns:

- **TRANSACTION_ID**: A unique identifier for each transaction.
- **TX_DATETIME**: The date and time at which the transaction occurs.
- **CUSTOMER_ID**: A unique identifier for each customer.
- **TERMINAL_ID**: A unique identifier for each terminal (merchant).
- **TX_AMOUNT**: The amount of the transaction.

- **TX_FRAUD**: A binary variable, with 0 indicating a legitimate transaction and 1 indicating a fraudulent transaction.

Fraud Scenarios

The dataset simulates fraud using the following scenarios:

1. Any transaction amount exceeding 220 is marked as fraudulent.
2. Each day, two terminals are randomly selected, and all transactions on these terminals in the next 28 days are marked as fraudulent.
3. Each day, three customers are randomly selected, and one-third of their transactions in the next 14 days have their amounts multiplied by 5 and marked as fraudulent.

Methodology

1. **Data Exploration**: Inspected and understood the dataset structure and content.
2. **Preprocessing**: Handled missing values, converted date columns to datetime format, and created additional features relevant to fraud detection.
3. **Model Building**: Trained multiple models including Logistic Regression and Random Forest.
4. **Hyperparameter Tuning**: Optimized the Random Forest model using GridSearchCV to improve performance.
5. **Evaluation**: Measured accuracy, precision, recall, and F1-score to select the best model.
6. **Feature Importance Analysis**: Analyzed which features contributed most significantly to model predictions.
7. **Prediction Testing**: Tested the model with new transaction data to classify them as fraudulent or legitimate.

Results

1. **Model Accuracy**: The Random Forest model achieved an accuracy of 100%.
2. **Confusion Matrix Analysis**:
    - True Positives (Fraud correctly identified): 2861
    - True Negatives (Legitimate correctly identified): 347970
    - False Positives (Legitimate incorrectly identified as Fraud): 0

- False Negatives (Fraud incorrectly identified as Legitimate): 0

3. **Feature Importance**: Key predictors included TX_AMOUNT, TERMINAL_ID, and customer spending habits.

Discussion

The results indicate that the Random Forest model performs exceptionally well in classifying transactions with perfect accuracy on this simulated dataset. However, while high accuracy is achieved, it is essential to recognize that this performance may not generalize to more complex real-world scenarios where fraud patterns can be less obvious.Future work could involve testing the model on real-world datasets or integrating additional features that capture more nuanced behaviors in transaction data.

Conclusion

This project successfully developed a machine learning model capable of detecting fraudulent transactions with high accuracy. Insights gained from feature importance analysis highlight critical factors influencing fraud detection. Future enhancements could focus on deploying the model in a real-time environment or as part of an API for broader application.

Output: