Credit Card Fraud Detection Model Report

Author name :- Priyanka Richhariya

1. Introduction

Objective : To develop a robust classification model capable of identifying fraudulent credit card transactions.

Problem : The dataset is highly imbalanced, with fraudulent transactions accounting for only 0.172% of all transactions.

Approach : Use data balancing techniques (SMOTE), train various models (Logistic Regression, Random Forest), and evaluate model performance using relevant metrics (Accuracy, ROC AUC, Confusion Matrix).

2. Dataset Overview

Source : Credit card transactions dataset (downloaded).

Number of records : 284,807 transactions.

Number of features : 31 (including 'Time', 'Amount', and 28 PCA-transformed features).

3. Data Preprocessing and Exploration

3.1 Handling Missing Values

No missing values found in the dataset.

3.2 Data Transformation

Converted 'Time' to datetime format.

Added new features like 'Transaction_hour' and 'Normalized_amount'.

Removed constant features.

3.3 Handling Imbalanced Dataset

Technique: SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset.

Original Distribution : Non-fraudulent: 99.83%, Fraudulent: 0.172%.

Post-SMOTE Distribution : Balanced dataset with equal class distribution.

4. <u>Exploratory Data Analysis (EDA)</u>

4.1 Class Distribution

Visualized the distribution of fraudulent and non-fraudulent transactions.

4.2 Time vs Amount Analysis

Scatter plots to analyze transaction amounts over time for both fraudulent and non-fraudulent transactions.

4.3 Correlation Analysis

Generated a correlation matrix to identify relationships between features.

5. Feature Engineering

PCA : Applied PCA for dimensionality reduction.

SelectKBest: Selected top 5 features using ANOVA F-test.

Final Features Used : ['V1', 'V3', 'V12', 'V14', 'Transaction_hour'].

6. Modeling

6.1 Model 1: Logistic Regression

Training: Model trained on SMOTE-resampled data.

Evaluation Metrics : Accuracy (92.17%), Balanced Precision, Recall, F1-score.

Confusion Matrix : Displayed in the heatmap.

6.2 Model 2: Random Forest Classifier

Training: 100 estimators, random_state=42.

Cross-validation score: Mean cross-validation score: 99.76%.

Evaluation Metrics : Accuracy (99.76%), Confusion matrix, ROC AUC Score (calculated).

Hyperparameter Tuning: Best parameters found using GridSearchCV.

7. Model Comparison

Compared the performance of Logistic Regression and Random Forest Classifier based on accuracy, precision, recall, and ROC AUC.

8. Visualizations

8.1 Class Distribution

- [Download Class Distribution](class_distribution.jpg)

8.2 Correlation Matrix

- [Download Correlation Matrix](correlation_matrix.jpg)

8.3 Actual vs Predicted Classes

- [Download Actual vs Predicted](actual_vs_predicted.jpg)

8.4 Transaction Volume Over Time

- [Download Transaction Volume Over Time](transaction_volume_over_time.jpg)

9. Conclusion

- The Random Forest Classifier outperformed Logistic Regression in terms of accuracy and AUC.

- Logistic Regression may still be preferred when model interpretability and computational efficiency are priorities.

- Further steps include model interpretability analysis and feature importance exploration.

10. Model Deployment

Best model (Random Forest Classifier) saved as 'credit_card_fraud_detection_model.pkl' for future use.