Credit Card Fraud Detection Report

Introduction

This report presents the analysis and implementation of a credit card fraud detection model using machine learning techniques. The dataset was pre-processed, imbalanced data was handled, and models such as XGBoost and Random Forest were trained and evaluated for fraud detection.

Problem Statement:

A credit card is one of the most used financial products to make online purchases and payments. Though the Credit cards can be a convenient way to manage your finances, they can also be risky. Credit card fraud is the unauthorized use of someone else's credit card or credit card information to make purchases or withdraw cash.
It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.
The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
We have to build a classification model to predict whether a transaction is fraudulent or not.

Data Overview

The dataset contains 284,807 rows and 31 columns. To protect user privacy, the original numerical features were transformed using Principal Component Analysis (PCA) into 28 principal components, in addition to the columns Time, Amount, and Class. The Class column is the target variable, with values indicating fraudulent (1) or non-fraudulent (0) transactions.

Data Pre-processing

- **Dataset Loading:** The dataset was loaded from a CSV file.
- **Missing Values:** Checked and handled missing values in the dataset.
- **Feature Engineering:**
    - The `Time` column was removed to avoid potential bias.
    - The `Amount` column was normalized.

- **Data Splitting:**
  - The dataset was split into training and testing sets using an 80-20 ratio.

Handling Imbalanced Data

- Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the dataset, ensuring that the minority class (fraudulent transactions) was adequately represented.

Model Training and Selection

- Implemented two machine learning models:
  - **Random Forest Classifier**
  - **XGBoost Classifier**
- Used **RandomizedSearchCV** for hyperparameter tuning to optimize model performance.

Model Evaluation

- The models were evaluated using:
  - **Accuracy**
  - **Precision, Recall, and F1-score**
  - **Confusion Matrix**
  - **ROC-AUC Curve**

Results

- The **XGBoost model** and the **Random Forest model** in detecting fraudulent transactions has achieved success .

Conclusion

The model demonstrated improved performance after addressing the class imbalance using SMOTE and training with the Random Forest and xgboost algorithm. With an accuracy exceeding 99%, the model shows strong potential for identifying fraudulent transactions in real-world applications. Moving forward, we aim to deploy the model and continuously monitor its performance to ensure its effectiveness.

References

- Libraries Used: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, XGBoost