Credit Card Fraud Detection Project Report

Introduction

Credit card fraud is a significant issue in the financial industry, leading to substantial financial losses for both consumers and businesses. Detecting fraudulent transactions in real-time is crucial to mitigate these losses. This project focuses on building a machine learning model to detect fraudulent credit card transactions using the **Logistic Regression** algorithm. The dataset used for this project is highly imbalanced, with a majority of legitimate transactions and a small fraction of fraudulent ones.

Project Overview

The goal of this project is to classify credit card transactions as either **legitimate (0)** or **fraudulent (1)**. The dataset contains anonymized features (V1-V28) derived from PCA (Principal Component Analysis), along with the transaction amount and time. The dataset is highly imbalanced, with only 492 fraudulent transactions out of 284,807 total transactions.

Key Steps:

1. **Data Collection and Preprocessing**: Load and explore the dataset.
2. **Handling Imbalanced Data**: Use undersampling to balance the dataset.
3. **Model Building**: Train a Logistic Regression model.
4. **Model Evaluation**: Evaluate the model's performance using accuracy.
5. **Predictive System**: Build a system to predict whether a transaction is fraudulent.

Data Collection and Preprocessing

Dataset Overview

The dataset contains 284,807 transactions, with 31 features:

- **Time**: Time elapsed between transactions.
- **V1-V28**: Anonymized features derived from PCA.
- **Amount**: Transaction amount.
- **Class**: Target variable (0 = Legitimate, 1 = Fraudulent).

Data Exploration

- The dataset has no missing values.
- The distribution of the target variable is highly imbalanced:
    - Legitimate transactions: 284,315
    - Fraudulent transactions: 492

Handling Imbalanced Data

To address the imbalance, we used **undersampling**:

- Randomly sampled 492 legitimate transactions to match the number of fraudulent transactions.
- Combined these samples to create a balanced dataset of 984 transactions.

Model Building

Logistic Regression

Logistic Regression was chosen because it is well-suited for binary classification problems. The model was trained on the balanced dataset.

Steps:

1. **Feature Scaling**: Standardized the features using StandardScaler to ensure all features contribute equally to the model.
2. **Train-Test Split**: Split the data into training (80%) and testing (20%) sets.
3. **Model Training**: Trained the Logistic Regression model on the training data.

Model Evaluation

Accuracy Scores

- **Training Data Accuracy**: 95.17%
- **Testing Data Accuracy**: 92.89%

The model performs well on both the training and testing datasets, indicating that it generalizes well to unseen data.

Predictive System

A predictive system was built to classify new transactions as either legitimate or fraudulent. The system takes input data, standardizes it, and uses the trained Logistic Regression model to make predictions.

Key Insights

1. **Imbalanced Data Handling**: Undersampling was effective in balancing the dataset, allowing the model to learn from both classes equally.
2. **Model Performance**: The Logistic Regression model achieved high accuracy on both training and testing datasets, demonstrating its effectiveness in detecting fraudulent transactions.
3. **Feature Importance**: The anonymized features (V1-V28) derived from PCA played a significant role in distinguishing between legitimate and fraudulent transactions.

Conclusion

This project successfully built a Logistic Regression model to detect credit card fraud. By addressing the imbalanced dataset and standardizing the features, the model achieved high accuracy and demonstrated its ability to classify transactions effectively. The predictive system can be integrated into real-time transaction processing systems to flag potentially fraudulent activities, helping financial institutions reduce losses and protect consumers.

References

- Dataset: Credit Card Fraud Detection Dataset
- Scikit-learn Documentation: Logistic Regression
- Pandas Documentation: DataFrame Operations

This report provides a comprehensive overview of the project, highlighting the key steps, challenges, and outcomes. The model's high accuracy and the predictive system's effectiveness make this project a valuable contribution to the field of fraud detection