Project: Credit Card Fraud Detection

Abstract

Credit card fraud detection is a critical task in the financial sector, given the growing volume of transactions and the potential losses associated with fraudulent activities. This project uses machine learning techniques to identify fraudulent credit card transactions in a highly imbalanced dataset. Data preprocessing included feature scaling, addressing class imbalance using SMOTE, and feature selection through Ridge and Lasso regression. Various models such as Logistic Regression, Random Forest, and Gradient Boosting were trained and optimized using GridSearchCV, achieving a high accuracy of **99.89%** and a recall of **85.84%**. This project demonstrates the application of advanced machine learning techniques to tackle real-world financial challenges.

Dataset Overview

- **Source**: Kaggle Credit Card Fraud Detection Dataset.
- **Size**:
    - Total Transactions: 284,807
    - Fraudulent Transactions: 492 (~0.17%)
    - Non-Fraudulent Transactions: 284,315
- **Features**:
    - 28 anonymized numerical features derived from PCA (V1 to V28).
    - Time: Seconds elapsed between the first transaction and each subsequent transaction.
    - Amount: Transaction amount.
    - Class: Target variable (0 for non-fraud, 1 for fraud).

Exploratory Data Analysis (EDA)

Class Distribution

The dataset is highly imbalanced:

Class (Target Variable)	Count	Percentage
Non-Fraudulent (0)	284,315	99.83%
Fraudulent (1)	492	0.17%

Key Insight: The imbalance necessitates techniques like SMOTE or weighted algorithms to prevent bias toward the majority class.

Feature Correlations

- A **correlation heatmap** was plotted to analyze relationships among features.
- The features are decorrelated due to PCA, which makes direct feature correlations insignificant.

Visualization of Amount and Time

- The Amount feature was log-transformed to normalize its distribution for model training.
- Fraudulent transactions tend to have smaller amounts on average compared to non-fraudulent transactions.

Transaction Type	Mean Amount ($)	Median Amount ($)
Fraudulent	122.21	9.21
Non-Fraudulent	88.29	26.98

Data Preprocessing

Handling Class Imbalance with SMOTE

- The **Synthetic Minority Oversampling Technique (SMOTE)** was used to oversample the minority class (1 for fraud) to balance the dataset.

Class	Before SMOTE	After SMOTE
0	227,451	227,451
1	394	227,451

Model Building

Algorithms Used

- **Logistic Regression**: Used as a baseline model.
- **Gradient Boosting Classifier**: Achieved the best results.

Hyperparameter Tuning

GridSearchCV

- Performed hyperparameter tuning for **Gradient Boosting**:
    - n_estimators: Number of trees.
    - max_depth: Depth of trees.
    - learning_rate: Step size for updates.

Best Parameters:

Parameter	Value
n_estimators	200
max_depth	5
learning_rate	0.1

Model Evaluation

Metric	Logistic Regression	Gradient Boosting
Accuracy	98.09%	99.89%
Recall	90.2%	85.8%

Conclusion

This project successfully implemented advanced machine learning techniques to detect fraudulent credit card transactions. By addressing class imbalance with SMOTE, applying feature selection with Lasso, and optimizing Gradient Boosting Classifier, the model achieved exceptional

performance with a recall of 85.84%. The project highlights the importance of handling imbalanced datasets and using ensemble methods for real-world challenges.