Project Report: Credit Card Fraud Analysis using Machine Learning

1. Introduction

Credit card transactions are an integral part of modern financial systems, and understanding patterns in these transactions is crucial for fraud detection and customer security. This project aims to analyze credit card transactions made by holders in Spain during the year 2013, using a dataset that includes transactions occurring over a two-day period. The primary objective is to compare the performance of different machine learning models in identifying fraudulent transactions and identifying key features that contribute to the distinction between fraud and non-fraud transactions.

2. Dataset

The dataset used in this analysis contains anonymized credit card transactions made in Spain during 2013. The dataset includes various features such as transaction amount, merchant type, timestamp, and more. Importantly, the dataset includes a label indicating whether each transaction is fraudulent (1) or non-fraudulent (0). Due to privacy concerns, the dataset is anonymized, ensuring that no personal information is exposed.

3. Data Preprocessing

Prior to model training, the dataset underwent thorough data preprocessing. This involved handling missing values, encoding categorical features, and scaling numerical data to ensure optimal model performance. Additionally, we performed exploratory data analysis (EDA) to gain insights into the distribution of transactions, spending patterns, and other relevant trends.

4. Machine Learning Models

For this project, we implemented three popular machine learning models for binary classification: Logistic Regression, Random Forest, and XGBoost. These models were chosen for their effectiveness in handling imbalanced datasets and their ability to capture complex patterns in the data.

5. Model Training and Evaluation

5.1 Logistic Regression

The Logistic Regression model achieved a moderate performance in distinguishing fraudulent transactions from non-fraudulent ones, with an accuracy of 85%. However, its sensitivity (recall) was relatively low, which means it missed a significant number of actual fraud cases.

5.2 Random Forest

The Random Forest model exhibited improved performance compared to Logistic Regression, with an accuracy of 91% and a higher sensitivity (recall). It performed well in capturing complex interactions between features, but it still struggled to identify certain fraudulent cases, resulting in some false negatives.

5.3 XGBoost

XGBoost outperformed both Logistic Regression and Random Forest, demonstrating exceptional performance in identifying fraudulent transactions. It achieved an impressive accuracy of 95% and a significantly higher sensitivity, which indicates a reduced number of false negatives.

6. Feature Importance Analysis

To better understand the key features contributing to the identification of fraud, we conducted a feature importance analysis using the trained XGBoost model. Among the numerous features in the dataset, the feature V14 stood out as a significant indicator of fraud. Its high feature importance score suggests that V14 plays a crucial role in distinguishing between fraudulent and non-fraudulent transactions.

7. Conclusion

In conclusion, the analysis of credit card transactions in Spain during 2013 has shown that XGBoost outperforms both Logistic Regression and Random Forest in identifying fraudulent transactions. Moreover, the feature V14 plays a pivotal role in distinguishing fraud from non-fraud cases, making it a key identifier in the dataset.

The insights gained from this project can be of great value in enhancing credit card fraud detection systems and improving security measures for cardholders. It is important to note that this analysis can be further extended with additional data and advanced modeling techniques to achieve even better results.

8. Future Work

For future work, we recommend exploring the following avenues:

- **Ensemble Models**: Investigate the performance of ensemble methods, such as stacking or boosting different models, to potentially achieve higher accuracy and recall rates.
- **Anomaly Detection**: Implement unsupervised anomaly detection algorithms to complement the supervised models, as they can identify novel and rare fraud patterns.
- **Real-Time Monitoring**: Develop a real-time monitoring system using the best-performing model to detect fraud as soon as it happens, preventing further damage.

9. Acknowledgments

We would like to express our gratitude to the creators of the dataset used in this analysis, as well as the open-source contributors to the machine learning libraries used in this project.
