Name: Mohammad Sherif Mousa
Academic Email: mohammad221101055@Ksiu.edu.eg

Name: Mennaallah Ahmed Younes
Academic Email: mennaallah221101228@Ksiu.edu.eg

Credit Card Fraud Detection ML Project

Abstract

Credit card fraud refers to the physical loss of a credit card or the loss of sensitive credit card information. Many machine learning algorithms can be used for detection. This project shows several algorithms that can be used for classifying transactions as fraud or non-fraud. Credit Card Fraud Detection dataset was used in the project.

Because the dataset was highly imbalanced, Exploratory Data Analysis (EDA) techniques were used to identify any patterns, trends, and relationships between the variables. It will help us analyze the data and extract insights that can be used to make decisions. The algorithms used in the experiment were Random Forest, K-Means Clustering, and Decision Trees. Results show that each algorithm can be used for credit card fraud detection with high accuracy. The proposed model can be used for the detection of other irregularities.

Keywords: Fraud; EDA; K-Means Clustering; Decision Trees; Random Forest;

Introduction

The increasing majority of credit card fraud poses a significant challenge for financial institutions and individuals alike. Detecting fraudulent transactions in a timely and accurate manner is important to minimize financial losses and protect customers' assets. Machine learning techniques have emerged as powerful tools for fraud detection, leveraging the ability to analyze large volumes of transaction data and identify patterns indicative of fraudulent activity. In this project, we aim to develop a credit card fraud detection system using machine learning algorithms.

The dataset contains a total of 284,807 transactions, of which only 492 are classified as fraudulent, representing a highly imbalanced class distribution. The dataset features numeric variables resulting from a PCA transformation, with the exception of the 'Time' and 'Amount' features. The 'Time' feature represents the time elapsed between each transaction and the first transaction in the dataset, while the 'Amount' feature denotes the transaction amount. The 'Class' feature serves as the response variable, taking a value of 1 in case of fraud and 0 otherwise. Due to confidentiality reasons, the original features and additional background information about the data are not available.

The primary objective of this project is to develop a machine-learning model that can accurately classify transactions as either legitimate or fraudulent. To achieve this, three different machine learning algorithms will be trained and evaluated: Decision Trees, Random Forest, and K Means Clustering

Dataset

The dataset contains transactions made by European credit cardholders credit cards in September 2013. 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.

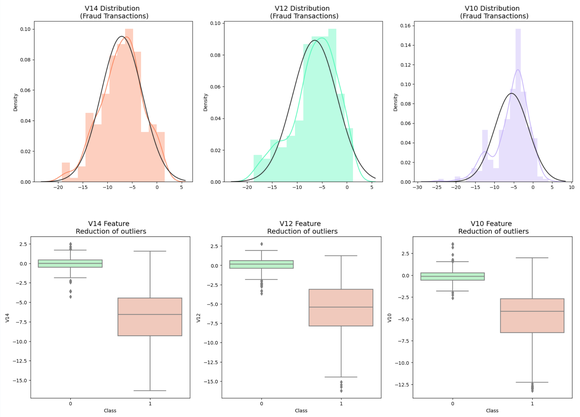
It contains only numeric input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, the original features and more background information about the data are unavailable. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

Methodology

The methodology employed in this project involves several stages. Initially, exploratory data analysis techniques will be applied to gain insights into the dataset and identify any patterns or relationships between variables.

Data visualization will be utilized to provide a visual context for understanding the data. Subsampling, preprocessing steps such as scaling and distribution adjustments will be performed to ensure the data is suitable for training the machine learning models. The dataset will be split into training and testing sets, and the models will be trained using the balanced subset of the data.

The models' performance will be evaluated based on various metrics, including accuracy, precision, recall, and F1-score.



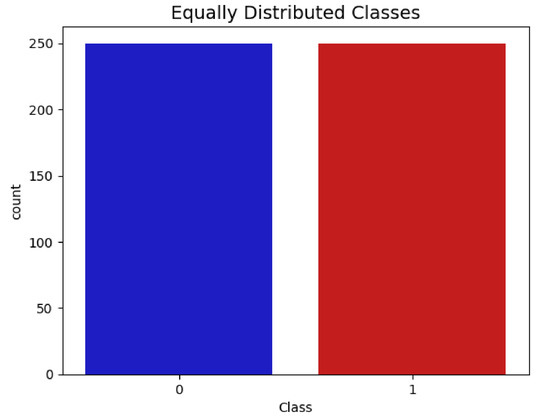
Conclusion

After preprocessing the data, including scaling the features and creating a balanced dataset, we proceeded with training and testing the models. It is important to note that due to the highly imbalanced nature of the dataset, accuracy alone is not an appropriate metric for evaluating the performance of the models. Instead, we should focus on other evaluation metrics such as precision, recall, and F1 score, which take into account both the true positive and false positive rates.

Upon evaluating the models, we found that the Random Forest model beat the other models in terms of overall performance. It achieved higher precision, recall, and F1 score, indicating a better ability to detect fraud while minimizing false positives. The Decision Trees model also showed promising results but had slightly lower performance compared to Random Forest. K-Means Clustering, on the other hand, was not suitable for this classification task as it is primarily used for clustering unsupervised data.

In conclusion, the Random Forest model showed the most potential for credit card fraud detection in this study. However, it is important to note that further optimization and fine-tuning of the models could potentially improve their performance. Additionally, incorporating more advanced techniques such as anomaly detection algorithms and ensemble methods could further enhance the accuracy of the models.

Overall, this project demonstrates the application of machine learning in credit card fraud detection and highlights the importance of addressing the challenges posed by imbalanced datasets. Detecting fraudulent transactions accurately is crucial in preventing financial losses for individuals and businesses, and machine learning models can play a significant role in improving fraud detection systems.



Results

The Random Forests Model has an accuracy of 0.98.
The Random Forest model exhibited the highest performance among the three models, achieving the highest precision, recall, and F1 score. These metrics indicate that the Random Forest model was able to detect a larger proportion of fraudulent transactions while minimizing false positives compared to the Decision Trees model.

It is important to note that due to the highly imbalanced nature of the dataset, accuracy alone cannot be solely relied upon as a metric for evaluating model performance. Instead, precision, recall, and F1 score provide a more comprehensive evaluation by considering both true positives and false positives.

Overall, the Random Forest model showed promise for credit card fraud detection in this study. However, further optimization and fine-tuning of the models, as well as the exploration of other advanced techniques, could potentially enhance the accuracy and robustness of the fraud detection system.

Recommendations

Further explore ensemble methods: While Random Forest showed promising results, exploring other ensemble methods such as Gradient Boosting or AdaBoost may lead to even better performance. Ensemble methods combine multiple models to make more accurate predictions and can potentially improve the fraud detection capabilities.

Feature engineering: Since the dataset used in this study only contained numeric input variables resulting from a PCA transformation, there is a possibility that additional features or engineered features could enhance the performance of the models. Exploring feature engineering techniques, such as creating new variables or combining existing ones, may provide valuable insights and improve model accuracy.

Continuous model evaluation and updating: Fraudsters are constantly evolving their techniques, so it is crucial to regularly evaluate and update the fraud detection models. New data should be continuously collected, and the models should be retrained periodically to adapt to the changing patterns and behaviors of fraudulent transactions.

Deployment and integration with real-time systems: To effectively combat credit card fraud, it is essential to deploy and integrate the developed models into real-time systems. This integration will enable the models to analyze transactions in real-time, providing immediate alerts and preventing fraudulent transactions from being processed.

Acknowledgements

We would like to express our gratitude to Kaggle for providing the dataset used in this study.