CREDIT CARD FRAUD DETECTION USING ANOMALY CLASSIFERS

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

Computer Science and Engineering

(2020-2024)

SUBMITTED BY

Prerna Singh (University Roll no –2000290100104)

Khyati Singla (University Roll no –2000290100082)

Prince Piyush (University Roll no –2000290100106)

KIET Group of Institutions, Delhi-NCR,

Ghaziabad (UP)

Department of Computer Science and Engineering

Table of content

Introduction

Credit card fraud has been a persistent issue, and with the advancement of technology, it continues to evolve. Scammers constantly devise new schemes to target unsuspecting victims, taking advantage of the growth in online banking, digital payments, and card-based transactions. While these innovations offer convenience, they also foster financial fraud, including money laundering, card theft, and identity theft, causing substantial losses to businesses and users.

Efforts to combat credit card fraud have led to the development of fraud detection systems using data mining and machine learning. However, challenges persist, partly due to the sensitive nature of credit card information, which necessitates anonymized datasets for model training. Furthermore, credit card fraud detection is complex as fraudulent activities continuously evolve.

Global credit card fraud losses have risen significantly in recent years, reaching $31 billion in 2020, affecting a substantial portion of card users. Detecting these frauds is crucial, as they impact financial transactions and institutions. Despite efforts to combat this issue, it remains a significant concern with the potential for further growth, making robust fraud detection systems an ongoing necessity.

Rationale

Credit card fraud remains an ongoing threat that adapts with technological advancements. As online banking and digital payments grow, financial crimes such as money laundering, card theft, and identity theft have surged. While these methods offer convenience, they also drive more sophisticated fraud techniques, resulting in substantial losses for businesses and individuals. To combat credit card fraud, data mining and machine learning are employed, yet challenges persist. Fraud can be categorized as physical card theft and the illicit acquisition of card-related data, leading to unauthorized transactions and significant financial losses. In recent years, global credit card fraud losses have amounted to $31 billion, affecting 65% of debit and credit card users, while identity theft through credit cards remains a major concern, with 390,000 allegations reported to the Federal Trade Commission in 2021. Credit card theft is projected to reach $38.5 billion by 2027. Detecting these frauds is vital, but machine learning faces challenges due to the sensitive nature of credit card data and the ever-evolving tactics employed by fraudsters. Efforts continue to refine methods to safeguard against credit card fraud in our increasingly digital financial landscape.

Objectives

- One of the primary objectives of the project is to identify and evaluate the most effective machine learning algorithms for credit card fraud detection.
- The project will delve into the strengths and weaknesses of each algorithm, considering factors like false positive rates, sensitivity, and computational efficiency.
- It will also address data imbalances between legitimate and fraudulent transactions, employing techniques like oversampling, under-sampling, or synthetic data generation to mitigate this imbalance. The goal is to improve the model's ability to differentiate between genuine and fraudulent transactions.
- Credit card fraud detection projects need to be mindful of ethical considerations and privacy protection.
- Exploring real-time detection techniques, like anomaly detection will be a vital aspect of the to detect the fraudulent transactions over normal transactions.

Methodology/ Planning of work

•Begin by obtaining the European and German credit card fraudulent transactions datasets, which serve as the foundation for your analysis. Ensure that the datasets include information on both legitimate and fraudulent transactions.

•Perform an initial exploratory data analysis to understand the dataset's characteristics. This step involves assessing data distribution, identifying missing values, and visualizing key features that may be relevant for anomaly detection.

•Implement different anomaly detection models, on both the European and German datasets. Train these models to identify fraudulent transactions.

•Evaluate the performance of these models using standard metrics such as accuracy, precision, recall, and F1-score. Create confusion matrices to visualize how the models classify transactions into true positives, false positives, true negatives, and false negatives.

•Address the class imbalance issue in the dataset. Implement various under-sampling and oversampling techniques, to balance the proportion of fraudulent and legitimate transactions in both datasets.

•Compare the new evaluation results with the baseline results obtained before balancing the data.

•Summarize the findings, draw conclusions about the effectiveness of the anomaly detection models, and discuss potential future directions for the project, such as exploring additional anomaly detection algorithms, optimizing the chosen models, or implementing real-time fraud detection systems.

Facilities required for proposed work.

The knowledge of the following technology and tools is required to complete the proposed project:

- **Machine learning algorithms:**
  1. Isolation Forest
  2. Local Outlier Factor
- **Resampling Techniques:**
  1. Random Under-sampling
  2. AllKNN Under-sampling
  3. SMOTE (Synthetic Minority Oversampling Technique)
  4. SMOTE-ENN (Synthetic Minority Oversampling Technique - Edited Nearest Neighbor)
- Performance analysing and determining factors such as accuracy, precision, recall, f1-score, confusion matrix.

Tools and a platform are required in addition to the tech stack mentioned.:

- **Google Colab:** Google Colab (short for Colaboratory) is a free cloud-based platform provided by Google that offers a Jupyter Notebook environment for coding in Python. It allows users to run code, conduct data analysis, and perform machine learning tasks using Google's cloud infrastructure.

Expected Outcome

The research presents a comprehensive analysis of the performance results of various machine learning models, coupled with the application of multiple resampling techniques. The objective is to effectively identify fraudulent transactions in two distinct datasets related to credit card fraud detection: one from the European context and the other from the German context. The study evaluates the effectiveness of these models and techniques in detecting fraudulent activities, aiming to provide a detailed insight into their respective capabilities in ensuring financial security. This analysis encompasses a wide range of scenarios and methods to enhance the understanding of fraud detection in credit card transactions.

Literature Review

S.No.	Title	Technology	Outcome
1	Credit card fraud detection using local outlier factor and isolation forest	IF and LOF algorithm	This paper presents a study on the detection of credit card fraud using Machine Learning algorithms, specifically Local Outlier Factor and Isolation Forest, applied to a publicly available dataset. The system described in the paper has been implemented in the Python programming language. After analyzing the dataset, it was found that the Local Outlier Factor algorithm achieved the highest accuracy rate of 97%, followed by the Isolation Forest algorithm with an accuracy rate of 76%.
2	Anomaly detection in credit card transactions using machine learning	IF Model	In this research paper, an enhanced version of the Isolation Forest model was investigated for the purpose of identifying fraudulent transactions more efficiently. To provide a more robust evaluation of the methodology, the study employed the Area Under the Precision-Recall curve (AUC), which yielded superior results compared to the Area Under the ROC curve. Ultimately, the research demonstrates the effectiveness of the proposed approach in a fraud detection model, achieving an observed efficiency rate of 98.72%. This performance signifies a notably improved approach compared to other fraud detection techniques.
3	Detection of credit card fraud using isolation forest algorithm	IF and LOF model	This research paper focuses on the utilization of the Isolation Forest (IF) algorithm for credit card fraud detection and compares it to the

			Local Outlier Factor (LOF) algorithm. The paper demonstrates that IF surpasses LOF in terms of error reduction, accuracy score, and recall for both algorithms. The rate of successful fraud detection achieved by the Isolation Forest is approximately 27%, a significant improvement compared to LOF's detection rate of only 2%. Additionally, the Isolation Forest model exhibits an impressive precision rate of 99.774%, outperforming LOF, which achieved a precision rate of 99.65%.
4	Credit card fraud detection using machine learning	RF and AdaBoost algorithm	The algorithms used are random forest algorithm and the Adaboost algorithm. The algorithms used are random forest algorithm and the Adaboost algorithm.
5	Comparison and analysis of logistic regression, Na¨ıve Bayes and KNN machine learning algorithms for credit card fraud detection	LR, NB, KNN algorithms with random under-sampling	Random undersampling was used to assess the performance of Logistic Regression (LR), Naive Bayes (NB), and K-Nearest Neighbors (KNN) on imbalanced datasets. LR consistently demonstrated the best results, achieving a peak accuracy of 95%. NB followed with a 91% accuracy, while KNN lagged behind at 75%. These findings suggest that LR is the most effective choice among these models for addressing the data imbalance issue and improving classification accuracy.

References

- https://colab.google/Virtual online consultations: advantages and limitations (VOCAL) study: https://bmjopen.bmj.com/content/6/1/e009388

- https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.htmlStripe: https://stripe.com/docs/api

- https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html

- https://imbalancedlearn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

- https://imbalancedlearn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

- https://imbalancedlearn.org/stable/references/generated/imblearn.combine.SMOTEENN.html