Final Project Submission

Credit Card Fraud Detection Using Regression Analysis

Course Name: Introduction to Data Science with Python

Course Code:- WM-ASDS04

Section: A, BATCH 9TH

Submitted to:

Farhana Afrin Duty

Asst. Professor, Dept. of Statistics,

Jahangirnagar University

Submitted By

Md. Tanvir Siraj , ID: 20229007

Mukesh Saha Dipto, ID:- 20229008

Irtefa Waseek, ID:- 20229012

Ajmir Al-Raji Chowdhury, ID:- 20229040

Ehsanuzzaman Surid, ID:- 20228013

Table of Contents

Problem Statement:

Credit card fraud is a significant problem that has been on the rise in recent years. According to a report by Nilson, a trade publication covering the payments industry, losses due to credit card fraud worldwide reached over $27 billion in 2018 and are projected to increase in the coming years. The complexity and sophistication of fraudulent activities have made detecting and preventing fraudulent transactions challenging. The traditional rule-based systems that banks and financial institutions have relied on are no longer effective in detecting sophisticated fraud.

Machine learning solutions have emerged as a powerful tool to detect fraudulent transactions. By analyzing large amounts of data, machine learning algorithms can learn patterns of normal and fraudulent behavior, thus enabling more accurate and effective detection of fraudulent transactions. Machine learning-based fraud detection systems have the potential to save billions of dollars in losses and prevent financial crimes.

1. Introduction

Introduction:

Credit card fraud has become a significant problem for banks, financial institutions, and their customers. Fraudsters use a variety of techniques to steal sensitive information, such as card numbers, passwords, and personal information, and use it to make unauthorized transactions. These unauthorized transactions can result in significant financial losses for both individuals and organizations.

The traditional rule-based systems used by banks and financial institutions to detect fraudulent transactions have proven to be insufficient in detecting and preventing sophisticated frauds. As a result, machine learning solutions have emerged as a powerful tool to detect and prevent credit card fraud.

Machine learning-based fraud detection systems use sophisticated algorithms to analyze large amounts of data, including transaction data, customer data, and historical data, to identify patterns of normal and fraudulent behavior. These algorithms can detect even the most subtle signs of fraudulent activity and can adapt to changing patterns of fraudulent behavior over time.

This report aims to provide an overview of machine learning-based credit card fraud detection systems. We will discuss the different types of fraud detection techniques and algorithms, the challenges in building and deploying these systems, and the benefits of using machine learning-based fraud detection systems. We will also explore some of the key considerations banks and financial institutions need to consider when implementing these systems.

1.2 Research Objectives

The main objectives of this project are:

1. To develop a credit card fraud detection system using machine learning algorithms.
2. To analyze and preprocess the credit card transaction data to extract meaningful features for the fraud detection model.
3. To train and evaluate different machine learning models to identify the best-performing algorithm for credit card fraud detection.
4. To achieve high accuracy, precision, recall, and F1-score in fraud detection while minimizing the false positive and false negative rates.
5. To improve the fraud detection system by continuously monitoring its performance and updating the model with new data to adapt to new fraud patterns.
6. To provide insights and recommendations based on the performance analysis of the fraud detection system to help prevent future fraudulent transactions.

1.3 Research Questions

Some potential research questions that could be answered through the credit card fraud detection project are:

1. What are the common types of credit card fraud and what are their characteristics?
2. What are the current methods used for credit card fraud detection and how effective are they?
3. What machine learning algorithms are best suited for credit card fraud detection and why?

4. How can feature engineering techniques improve the performance of machine learning models for credit card fraud detection?

1.4 Limitations of the Study

Credit card fraud is a significant and persistent problem in the financial industry, costing billions of dollars annually. Traditional rule-based fraud detection systems have limitations in detecting evolving and sophisticated fraud patterns. Machine learning-based approaches have shown promise in improving fraud detection accuracy. This project aims to develop and evaluate machine learning models for credit card fraud detection to improve the efficiency and effectiveness of fraud detection in the financial industry. The research questions will focus on identifying the best-performing machine learning algorithms, feature selection techniques, and performance evaluation metrics. The results of this project can be used to enhance current fraud detection systems and ultimately reduce financial losses due to fraud. However, limitations may include the availability and quality of data, as well as the potential for the models to be evaded by sophisticated fraudsters.

2. Research Methodology

2.1 Method

1. Data collection: Gather a large dataset of credit card transactions, including both fraudulent and non-fraudulent ones. This dataset will serve as the basis for training and testing the machine learning algorithms.

2. Data pre-processing: Clean and pre-process the data to ensure it is in a usable format. This includes removing duplicates, handling missing values, and scaling the features.

3. Feature selection: Identify the most important features that are most indicative of fraudulent transactions. This may involve using techniques such as correlation analysis and feature ranking.

4. Model training: Train a logistic regression model and a classification model (such as a random forest or support vector machine) using the pre-processed data and selected features. These models will learn to identify patterns in the data that distinguish between fraudulent and non-fraudulent transactions.

5. Model evaluation: Evaluate the performance of the trained models using various metrics such as accuracy, precision, recall, and F1 score. This will help determine which model is best suited for the task of credit card fraud detection.

6. Model optimization: Fine-tune the selected model by adjusting the model parameters or using techniques such as cross-validation. This will improve the model's accuracy and reduce the risk of false positives and false negatives.

It's important to note that this process is iterative, and may require going back and forth between different steps to refine the model and improve its performance.

Handling imbalanced datasets is a common challenge in machine learning, especially in fraud detection. In our project, we first analyzed the distribution of the data and found that the number of fraud cases was much smaller than non-fraud cases. To address this issue, we used the NearMiss algorithm to create a sub-data frame with an equal ratio of fraud and non-fraud transactions.

After balancing the dataset, we trained and tested several classifiers, including logistic regression, and random forest to find the one with the highest accuracy. We also paid attention to other metrics such as precision, recall, and F1 score, as accuracy can be misleading in imbalanced datasets.

We also made sure to understand common mistakes made with imbalanced datasets, such as using accuracy as the only evaluation metric or oversampling the minority class without considering the potential overfitting issue. Overall, by balancing the dataset and carefully selecting and evaluating the classifiers, we were able to develop a robust fraud detection model.

PCA transformation is a technique used for dimensionality reduction of high-dimensional datasets. It is used to transform the dataset into a lower-dimensional space while retaining as much of the original information as possible. In this credit card fraud detection dataset, all the features except for time and amount have been transformed using PCA.

Scaling is an important preprocessing step for PCA, as it is sensitive to the scale of the features. Therefore, it is essential to ensure that the features are scaled before applying PCA. In this dataset, the V features have been scaled or at least assumed to be scaled by the dataset developers. Scaling is also important for other machine learning algorithms as it helps to improve the performance and accuracy of the model.

I. Understanding our data

Before we can start working with the data, it is important to have a good understanding of it. This includes knowing what the data represents, what types of values are present in the data, and any patterns that may exist within the data. To accomplish this, we can use various tools and techniques such as data visualization and descriptive statistics.

II. Preprocessing

Once we have a good understanding of our data, we can begin the preprocessing stage. This involves cleaning and transforming the data to make it more suitable for analysis. One important step in this stage is scaling and distributing the data. Scaling is important to ensure that all features are given equal weight in the analysis while distributing the data is necessary to prevent any particular feature from dominating the analysis.

Another important step in the preprocessing stage is splitting the data into training and testing sets. This is necessary to evaluate the accuracy of our model.

III. Random Under-Sampling and Oversampling

One of the main challenges with credit card fraud detection is that the dataset is usually highly imbalanced, with very few cases of fraud compared to legitimate transactions. To address this issue, we can use techniques such as random under-sampling and oversampling.

Random under-sampling involves reducing the size of the majority class to match the size of the minority class. This can help to balance the dataset and improve the accuracy of the model. Oversampling, on the other hand, involves generating new samples of the minority class to increase its representation in the dataset.

Other techniques that can be used in this stage include anomaly detection, dimensionality reduction, clustering, and the use of different classifiers. For example, t-SNE can be used for dimensionality reduction and clustering to help identify any patterns in the data.

IV. Testing

Once we have trained our model, we can test its accuracy using various metrics such as precision, recall, and F1-score. This involves testing the model on the testing set and comparing the results with the actual values.

We can also compare the accuracy of different models using techniques such as logistic regression and neural networks. Additionally, we can test the effectiveness of under-sampling versus oversampling to determine which technique is more effective for our specific dataset.

The credit card fraud detection process involves several stages, including understanding the data, preprocessing, random under-sampling and oversampling, and testing. By using various techniques and tools such as scaling, distributing, and dimensionality reduction, we can improve the accuracy of our model and better detect instances of credit card fraud.

2.2 The Dataset Justification

The Google Big Query public data set for credit card fraud detection is a great dataset for our analysis because it contains a large amount of real-world credit card transactions, which is essential for training our machine learning models. The dataset includes over 284,000 transactions with 30 features, of which only 0.17% are fraudulent transactions, making it a challenging dataset for fraud detection. Furthermore, the dataset is updated and maintained regularly by Google, ensuring that we have access to the most up-to-date and relevant data for our analysis. Overall, this dataset provides a comprehensive and reliable source of data for our credit card fraud detection project.

3. Results

3.1 The Process

For EDA, we started by exploring the dataset's statistical properties and distribution of each feature. We also checked for any anomalies or outliers and analyzed their impact on the dataset. To handle missing values, we identified the features with missing data and then used different techniques such as imputation and deletion, depending on the missing data's characteristics.

To decide on the algorithm, we experimented with different machine learning algorithms like Logistic Regression, Random Forest, and XGBoost. We evaluated each model's performance using metrics like precision, recall, and F1-score. Based on the evaluation results, we selected the algorithm that provided the best performance on the dataset.

To build our model, we first preprocessed the dataset by performing feature scaling and splitting the dataset into training and testing sets. We then trained the selected algorithm on the training set and used the testing set to evaluate the model's performance. We fine-tuned the model by adjusting its hyperparameters to improve its performance. Finally, we validated the model's performance using techniques such as cross-validation and AUC-ROC curve analysis.

3.2 The Code Results

5. Discussion and Conclusions

Discussion:

For our credit card fraud detection project, we explored the use of machine learning algorithms, specifically logistic regression, and classification, to identify fraudulent transactions.

During the preprocessing phase, we performed scaling and distribution on the data and split it into training and testing sets. We also addressed the imbalanced nature of the dataset through random under-sampling and oversampling techniques. Additionally, we performed anomaly detection, dimensionality reduction, and clustering using t-SNE to better understand the distribution of our data.

We then experimented with various classifiers, ultimately selecting logistic regression and comparing its performance with oversampling using SMOTE. We also explored neural network testing with both under-sampling and oversampling techniques.

Through our analysis, we found that logistic regression, with oversampling using SMOTE, performed well in identifying fraudulent transactions, achieving a high recall rate while still maintaining a reasonable precision rate. Our results showed that oversampling can be an effective method for handling imbalanced datasets in fraud detection tasks.

Conclusion:

In conclusion, the use of machine learning algorithms, particularly logistic regression, and classification, can be an effective approach to credit card fraud detection. Preprocessing techniques, such as scaling and distribution, and addressing imbalanced datasets through random under-sampling and oversampling, can significantly improve model performance.

Our analysis showed that logistic regression, with oversampling using SMOTE, achieved a high recall rate and reasonable precision rate in identifying fraudulent transactions. This indicates that oversampling can be a useful technique in handling imbalanced datasets for fraud detection tasks.

Overall, our study highlights the importance of utilizing machine learning algorithms and preprocessing techniques in credit card fraud detection and provides insights into the efficacy of oversampling in addressing imbalanced datasets.