STM JOURNALS

Journal of
Open Source Developments

ISSN: 2395-6704
Volume 11, Issue 1, 2024
January–April
DOI (Journal): 10.37591/JoOSD

http://computers.stmjournals.com/index.php?journal=JoOSD&page=index

Review JoOSD

Credit Card Fraud Detection Using Machine Learning Techniques

Avanish Kumar Singh[1,*], Himanshu Singh[1]

Abstract

Credit card fraud has become a significant concern in today's digital economy, with billions of dollars being lost annually to fraudulent transactions. Conventional rule-based approaches frequently prove inadequate in addressing the constantly changing strategies employed by fraudsters. Machine learning methods have emerged as robust solutions for detecting credit card fraud, presenting the capability to accurately identify fraudulent transactions promptly. In this study, we investigate the efficiency of three widely used machine learning algorithms—logistic regression, random forest, and decision tree—for the detection of credit card fraud. Through an extensive comparative study, we analyze the performance, accuracy, and efficiency of each algorithm on a real-world credit card transaction dataset. Our findings provide valuable insights into the strengths and limitations of these techniques in addressing the challenges posed by credit card fraud.

Keywords: Credit card frauds, training system, decision tree classifiers. random forest algorithms, logistic regression

INTRODUCTION

An increasing number of people, companies, financial institutions, and law enforcement agencies are concerned about credit card fraud, which is a type of identity theft and financial deception. It describes the unlawful use of another person's credit card details to conduct fraudulent purchases, which can cause victims to suffer short- and long-term financial losses. Credit card fraudsters have adjusted their methods as technology develops and transactions become more digital, taking advantage of flaws in payment systems, e-commerce platforms, and even individual conduct. The impact of credit card fraud on financial matters is significant. Industry studies state that fraudulent actions cost billions of dollars annually, therefore understanding the complex principles underlying these illegal operations is crucial. Furthermore, credit card fraud affects other industries as well, undermining consumer trust in financial systems, driving up costs for businesses, and necessitating the expenditure of large sums of money on mitigation and prevention strategies. A National Crime Records Bureau (NCRB) report from 2021 states that 3432 cases of credit and debit card fraud were registered in India. This was a noteworthy increase of around 20% from the previous year. Furthermore, according to the survey, the number of credit card theft cases skyrocketed in just 2 years in 2020, rising by over 70% [1]. These data demonstrate the alarming surge in credit and debit card fraud in India, emphasizing the need for workable solutions to halt this growing problem.

***Author for Correspondence**
Avanish Kumar Singh
E-mail: Avanish121299@gmail.com

[1,2]Research Scholar, MCA Thakur Institute of Management Studies, Career Development & Research (TIMSCDR) Mumbai, Maharashtra, India

Received Date: February 29, 2024
Accepted Date: March 26, 2024
Published Date: April 05, 2024

Citation: Avanish Kumar Singh, Himanshu Singh. Credit Card Fraud Detection Using Machine Learning Techniques. Journal of Open Source Developments. 2024; 11(1): 1–7p.

RELATED WORK

Studies have utilized various machine learning methods for detecting credit card fraud. The researchers aim to identify strategies that best integrate the adaptive credit card fraud detection model. The researchers have used techniques like decision tree, random forest [2] employed different techniques of machine learning and convolutional

neural network (CNN) and compared and contrasted their performance and saw that convolutional neural network to have the highest accuracy and performance measure and hence is used for credit card fraud detection [3].

They employed two distinct machine learning methods. They analyzed both decision tree and a random forest confusion matrix for the actual and expected labels. We also found evidence of a disparity in the types of transactions that fall into the fraudulent and legitimate categories [4]. Credit card fraud detection accuracy of 0.99915(99.92%) was thus obtained using a random forest algorithm [5] takes a dataset taken from Kaggle, which has 2 days of transactions of European credit card holders. A CSV (comma-separated value) file was used to store the dataset. The dataset consists of 2,84,807 transactions with only 492 fraud transactions. Due to the confidentiality problem, the attributes are PCA (principal component analysis) transformed, and the input-viable dataset is converted into numerical values. So, after transformation, there are a total of 31 features. Features like time, amount, and class are not converted. V1, V2, V3, V27, and V28 are the PCA transformation features obtained using PCA transformation. Class is binary classified, meaning only two classes are represented, 1 and 0, where 1 represents fraud, and 0 represents a legitimate transaction. Merely 0.172% of transactions are fraudulent, leading to a severely imbalanced classification challenge. Five different techniques are applied and each method's accuracy is calculated and the performance metrics like recall (sensitivity), F1-score, and precision were found, which helped select the best method among the different methods used. Random forest gave the best output among all the machine learning techniques. Khare and Sait [4] analyzed different machine learning models like logistic regression XGBoost Classifier K-nearest neighbor gradient boosting classifier.

PROPOSED METHODS

Following are the step to apply machine learning algorithms and test the accuracy and other features of the algorithms [6].

Importing Dataset along with Python Libraries

As usual, the first step is to import the libraries, which are Pandas, NumPy, and Matplotlib. Importing the dataset is done using the Pandas module in Python. Additional Python modules are imported as required [7].

Data Analysis and Preprocessing

Data analysis involves different exploratory data analysis (EDA) processes. Data preprocessing includes both cleaning and transforming the data to prepare it for training the model. A method called data preprocessing is applied to transform the unprocessed data into a clean data collection.

i. *Dataset Cleaning:* Preparation for classifier training: Dataset cleaning involves rectifying or eliminating inaccurate, corrupted, improperly formatted, duplicate, or incomplete data entries within the dataset.
ii. *Splitting of Dataset*: Two datasets, namely training data and testing data, are created from the given data. While the testing data will be used to evaluate the dataset's effectiveness and accuracy, the training data will be used to train the model.
iii. *Machine Learning Models:* In this section, we discuss the machine learning algorithms that are applied in our research on credit card fraud detection.

Logistic Regression

This traditional binary classification approach calculates the likelihood that an instance falls into a specific class. Using the training data, we trained a logistic regression model and adjusted the hyperparameters using methods such as grid search. The testing dataset was utilized to evaluate the model's performance, employing various metrics including precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (Figure 1).

Random Forest

This ensemble learning technique builds several decision trees during training and aggregates the predictions to increase accuracy. We put in place a random forest classifier, modifying variables such as the maximum depth and tree count. The model's efficacy was evaluated using the same metrics as in logistic regression, and parameter adjustment was done using cross-validation (Figure 2).

Decision Tree

Decision trees are simple yet effective classifiers that recursively split data based on the most discriminative attributes. We constructed a Decision Tree classifier and pruned the tree to prevent overfitting [8]. The model's performance was evaluated using the aforementioned metrics (Figure 3).

A table known as a confusion matrix is frequently used to explain how well a classification model performs when applied to a collection of data for which the true values are known.

The confusion matrix typically has four entries:

1. *True positives (TPs):* The number of events that were accurately predicted as positive.
2. *True negatives (TNs):* The number of instances that were accurately predicted as negative.
3. *False positives (FPs):* The number of instances that were predicted as positive but are actually negative.
4. *False negatives (FNs):* The number of instances that were predicted as negative but are actually positive.

Performance Analysis

The performance of different algorithms is measured using several performance metrics such as precision, accuracy, recall, and F1 score, and the following results were obtained as shown in Figures 4 to 6.

RESULTS AND DISCUSSION

For this a dataset of 76.3 MB containing 8 columns was taken from Kaggle and was extracted in Google Collaboratory for further study [9]. The dataset consists of 1,000,000 transactions with only 87,403 fraud transactions which is only 8.7% of total transactions which is a highly imbalanced classification problem. Libraries such as Numpy and Pandas were used to extract and analyze the behavior of the dataset whereas Matplotlib and Seaborn were used to get statistical graphics. To determine the accuracy and precision score of the system against each fraudulent transaction, additional algorithms like decision tree, logistic regression, and random forest were applied. A total of 80% is used for training, and the remaining 20% is used for testing [10]. The experiments show that random forest produces the best results compared to logistic regression and decision tree as shown in Tables 1 and 2.

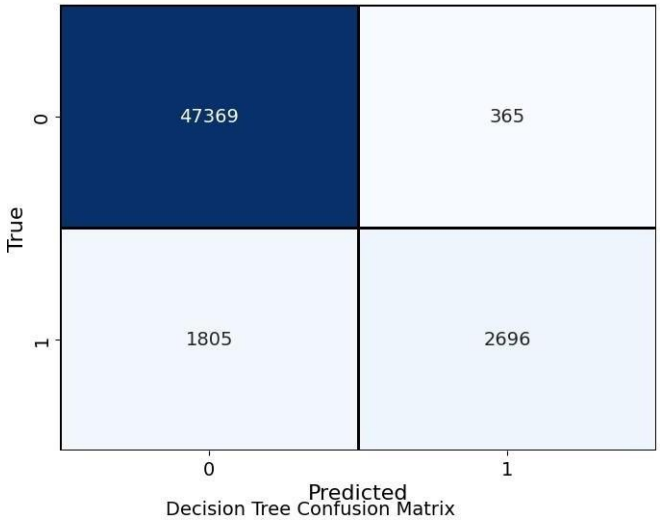


Figure 1. Confusion matrix of logistic regression.

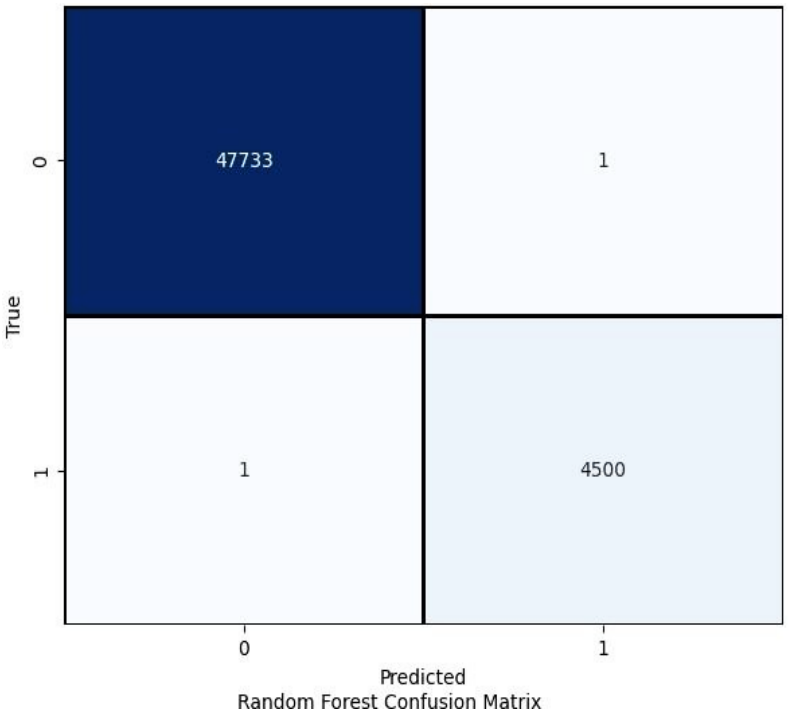


Figure 2. Confusion matrix of random forest regression.

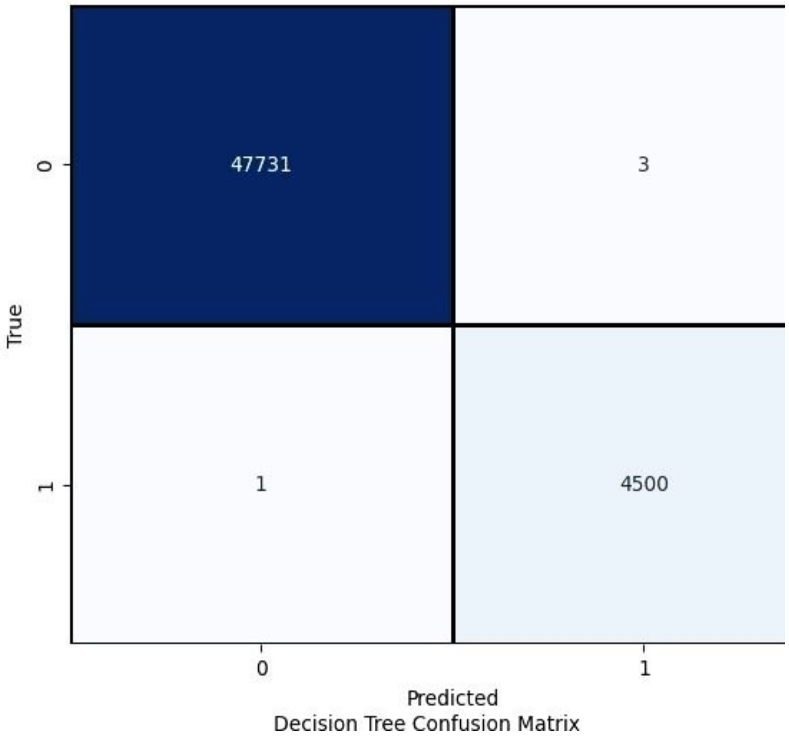
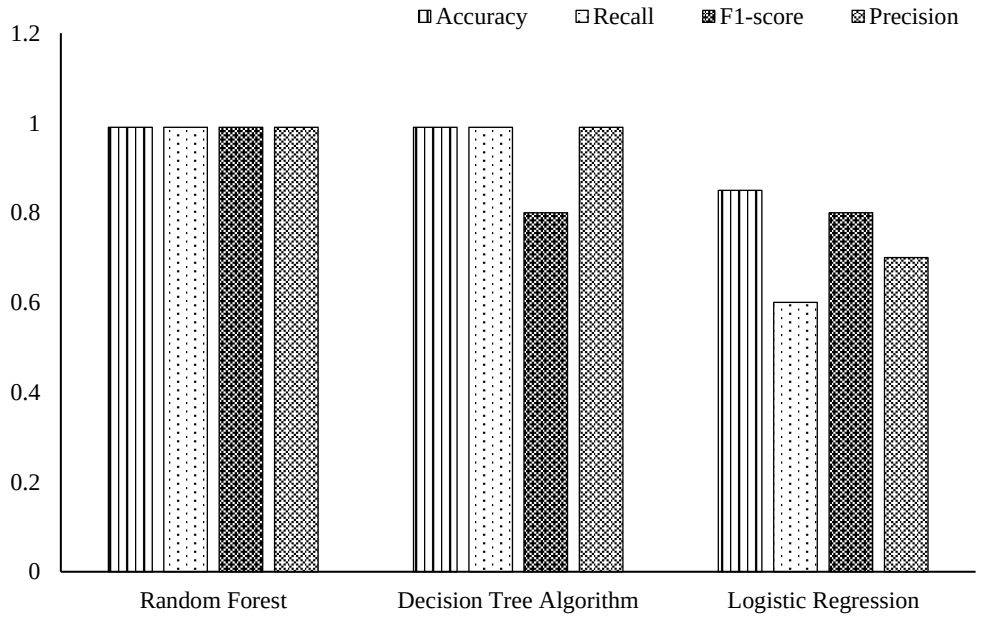


Figure 3. Confusion matrix of decision matrix.

Table 1. Performance of machine learning models with test size 20%.

Approach	Accuracy	Recall	F1-Score	Precision
Logistic regression	95.85	0.60	0.71	0.88
Random forest	99.99	0.99	0.99	1.0
Decision tree	99.99	0.99	0.99	0.99



Performance Comparison: Random Forest vs Decision Tree vs Logistic Regression

Figure 4. Performance of machine learning models with test size 20%.

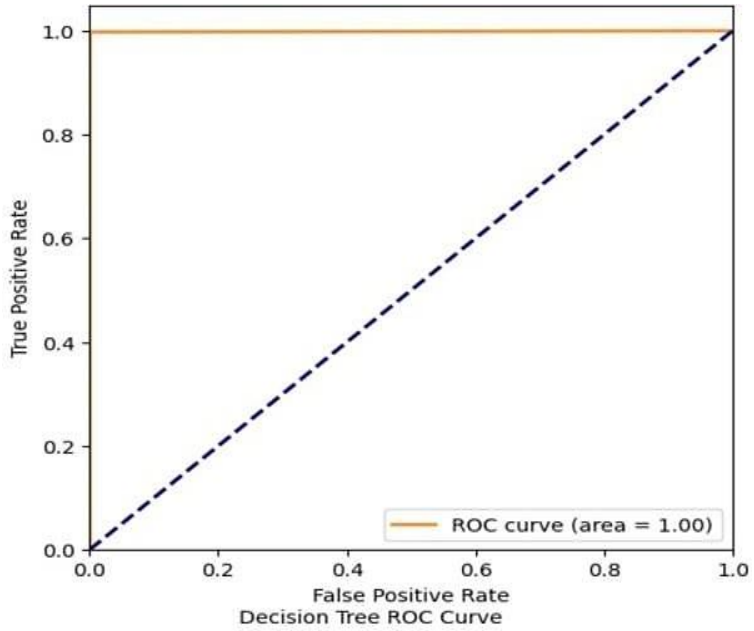


Figure 5. Decision tree receiver operating characteristic (ROC) curve.

Table 2. Best results compared to logistic regression and decision tree.

	Distance_from_home	Distance_from_last_transaction	Ratio_to_Median_purchase_price
Count	96225.0	96225.0	96225.0
Mean	26.70	5.02	1.81
Std	65.42	24.52	2.92
Min	0.021	0.00048	0.01
25%	3.86	0.29	0.47
50%	9.96	0.99	0.99
75%	25.71	3.33	2.08

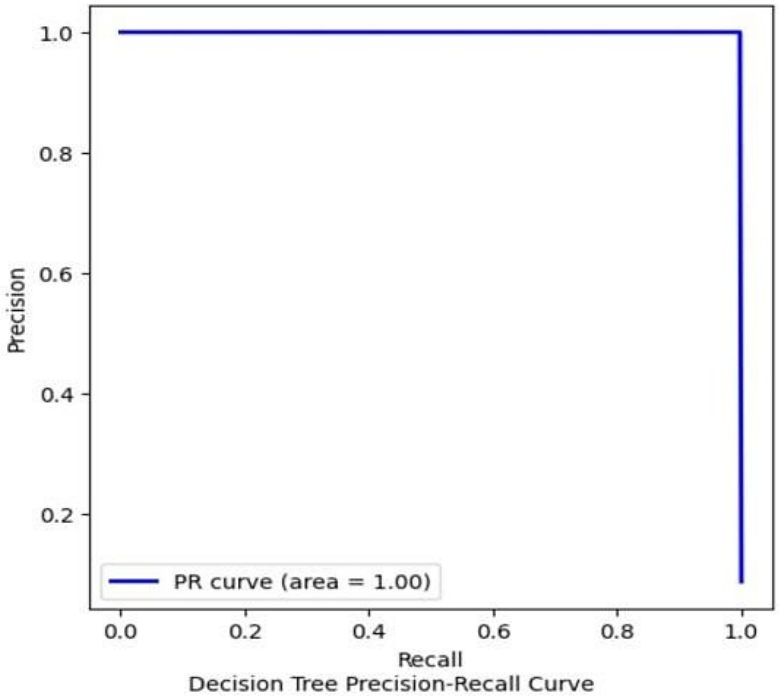


Figure 6. Decision tree precision-recall curve.

CONCLUSION

This research was accomplished by using the random forest, decision tree, and logistic regression methods. After analyzing the dataset, we can say that it is unbalanced, meaning that 87,403 of the 1,000,000 transactions were fraudulent, while 91,2597 were legitimate. As the investigation went on, under-sampling produced a fresh dataset that was meant to be used for training the computer. Graphs and statistics were employed to examine the prevalent fraudulent trends found in the dataset. The testing process employed machine learning techniques to determine the accuracy and precision scores. While decision tree and random forest have nearly the same accuracy, random forest has the best precision at 1.0 compared to decision tree's 0.99. We believe that random forest is the best algorithm for detecting credit card fraud based on precision. Higher training data yield better results for the random forest method, but smaller training data allow for quicker testing and implementation. It might be beneficial to use extra pre-processing methods. Our future work will use cutting edge advancements like artificial intelligence, deep learning, machine learning, and machine intelligence to combat credit card fraud.

REFERENCES

1. Thennakoon A, Bhagyani C, Premadasa S, Mihiranga S, Kuruwitaarachchi N. Real-time credit card fraud detection using machine learning. In: 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, January 10–11, 2019. pp. 488–493.
2. Roy A, Sun J, Mahoney R, Alonzi L, Adams S, Beling P. Deep learning detecting fraud in credit card transactions. In: 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, April 27, 2018. pp. 129–134.
3. Randhawa K, Loo CK, Seera M, Lim CP, Nandi AK. Credit card fraud detection using AdaBoost and majority voting. IEEE Access. 2018; 6: 14277–14284.
4. Xuan S, Liu G, Li Z, Zheng L, Wang S, Jiang C. Random forest for credit card fraud detection. In: 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, China, March 27–29, 2018. pp. 1–6.
5. Khare N, Sait SY. Credit card fraud detection using machine learning models and collating machine learning models. Int J Pure Appl Math. 2018; 118 (20): 825–838.
6. Melo-Acosta GE, Duitama-Munoz F, Arias-Londoño JD. Fraud detection in big data using supervised and semi-supervised learning techniques. In: 2017 IEEE Colombian Conference on

Communications and Computing (COLCOM), Cartagena, Colombia, August 16–18, 2017. pp. 1–6.
7. Jiang C, Song J, Liu G, Zheng L, Luan W. Credit card fraud detection: a novel approach using aggregation strategy and feedback mechanism. IEEE Internet Things J. 2018; 5 (5): 3637–3647.
8. Amro A, Al-Akhras M, Hindi KE, Habib M, Shawar BA. Instance reduction for avoiding overfitting in decision trees. J Intell Syst. 2021; 30 (1): 438–459.
9. Yang MY, Kumaar S, Lyu Y, Nex F. Real-time semantic segmentation with context aggregation network. ISPRS J Photogramm Remote Sensing. 2021; 178: 124–134.
10. Sun Y, Cheng H, Zhang S, Mohan MK, Ye G, De Schutter G. Prediction and optimization of alkali-activated concrete based on the random forest machine learning algorithm. Construct Build Mater. 2023; 385: 131519.

