JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Supervised Machine Learning Techniques for Predicting Credit Card Fraud

[1] Ngamrang Arangham, [1]Mokchen Suayang, [2]Biswajit Das, [2] Sonali Mondal

Department of Computer Science, Arunachal University of Studies, Namsai, Arunachal Pradesh, India

Abstract: Everyone is shifting to online purchases in this modern era, and credit cards have become crucial to this since they let people make payments without carrying cash. Although this method of payment is extremely helpful, there are some disadvantages as well. The total amount of credit card fraud has risen simultaneously with the increasing number of consumers with credit cards. Credit card details are being illegally collected by the fraudster, which can be used for unauthorized transactions. Some algorithms based on machine learning can be used to predict and stop fraudulent activities to tackle this problem. This paper compares a few well-known supervised learning algorithms to identify genuine and fraudulent transactions. The machine learning algorithms we use in this paper are logistic regression, linear regression, linear regression, and random forest.

Keywords: Credit card, Random Forest, Logistic Regression, Machine Learning, and Fraud Prediction

I. Introduction

In today's market, credit cards have become a necessary component of financial transactions because they allow users to make payments cashless and don't need to carry cash. Cards facilitate large-scale transactions and payments; credit cards have revolutionized the payment process by making it easier for customers. The user can pay all around the world with the card without exchanging the currency physically. But unfortunately, there are some drawbacks to this easier transaction. Credit card fraud has been rising in parallel with increasing credit card usage. The fraudster illegally collects the information of an individual and transactions the authorized payment with the help of the information, and sometimes they physically steal the card and use it for transactions without knowing the user, resulting in loss of privacy and financial damages.

To combat this problem, some machine learning algorithms are used in this project. Machine learning falls under the umbrella of artificial intelligence (AI), focusing on creating algorithms and models for computers to learn from data and enhance their performance without explicit programming. Its essence lies in enabling machines to detect patterns, make forecasts, and adapt through experience. Logistic regression, a statistical approach, deals with binary classification, estimating the likelihood of an event (like pass/fail, win/lose, or alive/dead) based on independent variables. It utilizes the logistic function to produce probabilities ranging from 0 to 1. Random Forest, an ensemble method, builds multiple decision trees in training and outputs the mode of classes (for classification) or mean prediction (for regression). Renowned for handling large, high-dimensional datasets and estimating missing data, it maintains accuracy even with significant missing data proportions. SVC, commonly denoting support vector classification in machine learning, belongs to the support vector machine (SVM) family and is employed for classification tasks. SVMs are supervised learning models used for data analysis in classification and regression scenarios.

II. Literature Review

To improve credit card fraud detection, Khalid, Abdul Rehman, et al. (2024) introduced an ensemble machine learning approach that addresses problems with drifting concepts and information imbalances in existing systems. Several classification algorithms, including KNN, Random Forest, SVM, Bagging, and Boosting, are included in their inventive model. To address dataset imbalances, under-sampling and SMOTE are also used. Using a dataset from Europe, they assessed the model and discovered that it performed better than conventional techniques in several areas.

To detect credit card fraud, Khatri, Samidha, Aishwarya Arora, and Arun Prakash Agrawal (2020) compared supervised machine learning methods. They evaluated how well algorithms such as random forest, decision tree, and logistic regression distinguished between legitimate and fraudulent transactions. The study found that the sheer number of identical transactions taking place at the same time makes it difficult to detect fraud.

To detect credit card fraud, Madhurya et al. (2020) carried out exploratory research utilizing machine learning techniques. By identifying critical trends that differentiate fraudulent transactions from legitimate ones, they sought to improve the efficacy and usability of fraud detection systems.

Aftab, A. U., et al. (2023) looked into data imbalance concerns and supervised machine learning techniques for credit card fraud detection. By utilizing SMOTE to compare Random Forest, SVM, Decision Tree, and logistic regression, they discovered that Random Forest was the most successful, with outstanding recall scores.

Machine learning was used by Sailusha, Ruttala, et al. (2020) to identify credit card fraud by contrasting several supervised algorithms and resampling techniques. To choose the ideal model, they used hyperparameter tuning and grid search.

In their discussion of machine learning techniques for detecting credit card fraud, Varmedja, Dejan, et al. (2019) emphasized the importance of effective detection systems. They looked at the results of techniques like neural networks, logistic regression, and random forest using actual datasets.

An algorithm was created by Thennakoon, Anuruddha, et al. (2019) to address information imbalances in the identification of credit card fraud. Their method used analytics to predict fraud and evaluate the authenticity of the transaction.

In 2019, Dornadula, Vaishnavi Nath, and Sa Geetha developed a novel methodology for streamed transaction data fraud detection. To effectively predict fraudulent activity, they created classification algorithms, defined groups, and categorized cardholders based on transaction quantities.

The effectiveness of algorithms like Naive Bayes, K-nearest neighbor, and logistic regression based on different variables was assessed in a comparative study of machine learning techniques for spotting credit card fraud by Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare (2017).

III. Methodology

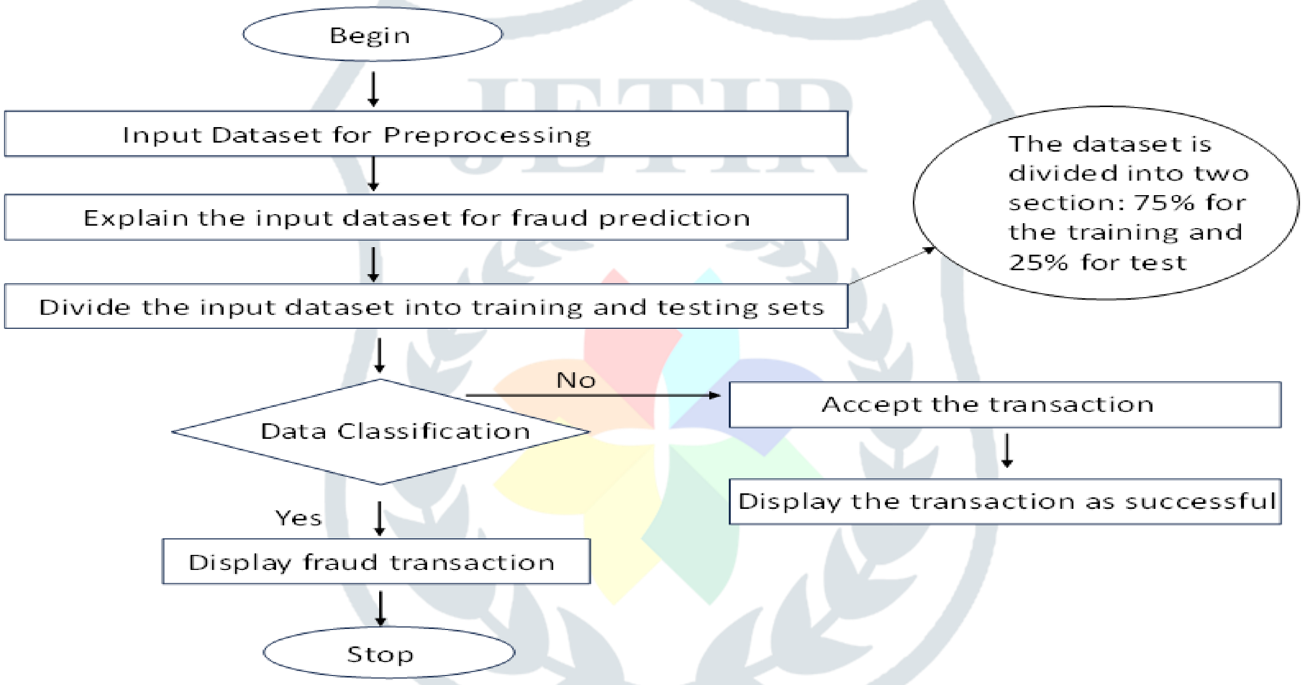

Begin
Input Dataset for Preprocessing
Explain the input dataset for fraud prediction
Divide the input dataset into training and testing sets
The dataset is divided into two section: 75% for the training and 25% for test
Data Classification
No
Accept the transaction
Display the transaction as successful
Yes
Display fraud transaction
Stop

Fig 1: Flowchart

I. Begin the fraud detection procedure.
II. Input Dataset for Preprocessing: Load the dataset so that it is prepped and ready for analysis.
III. Explain the input dataset for fraud prediction. Define the dataset's main attributes for fraud detection.
IV. Divide the input dataset into training and testing sets. Divide the dataset into two sections: 75% for training the model and 25% for determining accuracy.
V. Data Classification: Determine whether the data reflects fraudulent behaviors.
•If not, the data is not considered phony.
•Accept the transaction. Confirm the transaction.
•Display the transaction as successful. Show a confirmation that the transaction was completed successfully.
VI. Display fraud transaction: stop transactions.

IV. Experimental Results

A scatter plot graph (fig.2)

I. The graph uses a Cartesian coordinate system, with x ranging from 0 to 8000 and y ranging from 0 to 1.
II. Individual data points are represented by many blue dots distributed around the plot.
A cluster of dots is concentrated around the origin (0,0) and extends horizontally along the x-axis.
III. Several single dots at higher y-values, particularly around y = 1 and y = 0.8, are scattered over the x-axis.
IV. The plot's backdrop is white, and both axes are labelled with numerical values to indicate scale.

Systematic risk is the only independent variable for the CAPM and inflation, interest rate, oil prices and exchange rate are the independent variables for APT model.

Bar Graph (fig.3)

The bar graph compares the validation accuracies of three machine-learning techniques, including

- Logistic Regression (0.88).
- Random Forest Classifier (0.94).
- A Support Vector Classifier (SVC) with a linear kernel and a random state time of 2 (0.92).

Each classifier has an accuracy close to 0.8, showing that they are extremely efficient in predicting credit card fraud. The graph clearly shows the performance of the different machine learning algorithms: Logistic Regression (0.98), Random Forest Classifier (0.85), and Support Vector Classifier (SVC) with a linear kernel and a random state of 2 (0.92).

Algorithms	Output Accuracy
Logistic Regression	0.88
Random Forest	0.94
SVC (kernel=linear, random state=2)	0.92

Table 1: The table show differences algorithms output

The table shows the differences between algorithms, accuracy output, and time taken for the result.

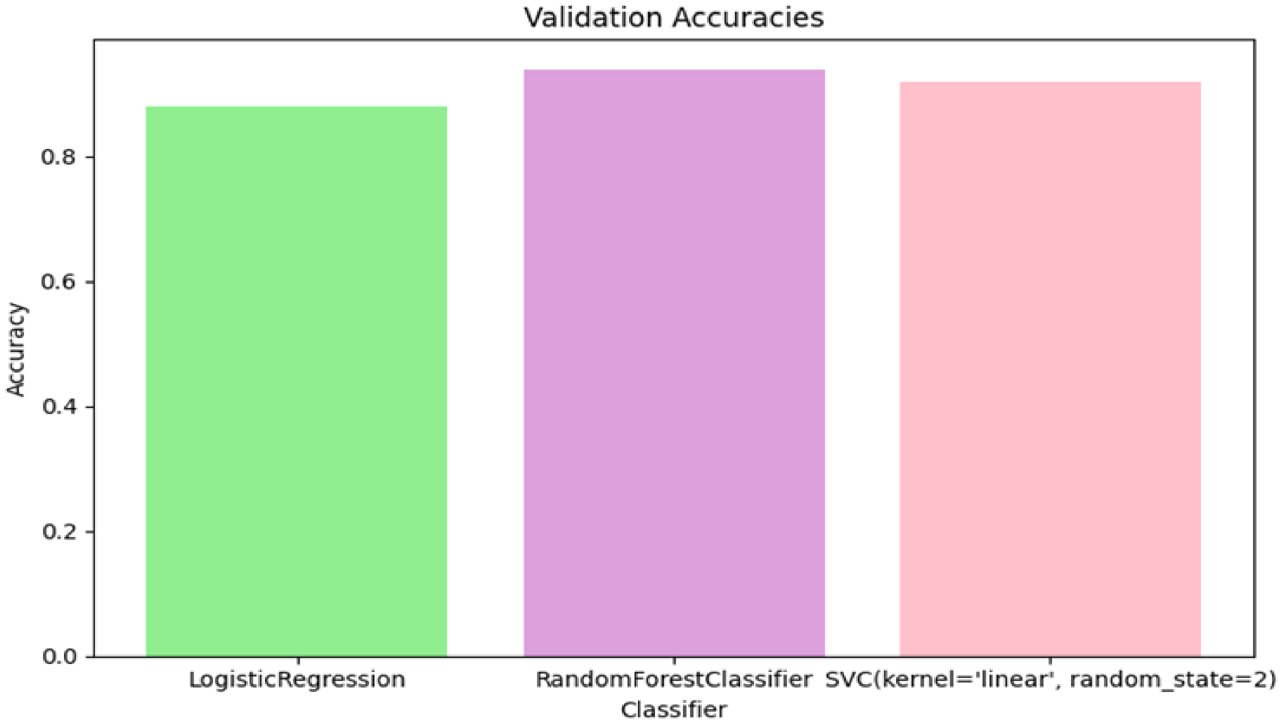


Fig 3: Bar graph It shows the validation accuracy

of various machine learning techniques.

line graph (fig.4)

A line graph has been used to examine the validation accuracy of various machine learning algorithms for credit card fraud prediction.

Accuracy using Various Classifiers

Y-Axis Represents the accuracy, which ranges from 0.88 to 0.94.

X-Axis Three different classifiers are being compared: logistic regression, random forest classifier, and SVC (kernel=linear, random state=2).

Data Points

Logistic regression has an accuracy of around 0.88.

The Random Forest Classifier has an accuracy close to 0.94.

SVC (kernel=linear, random state=2) has an accuracy close to 0.92.

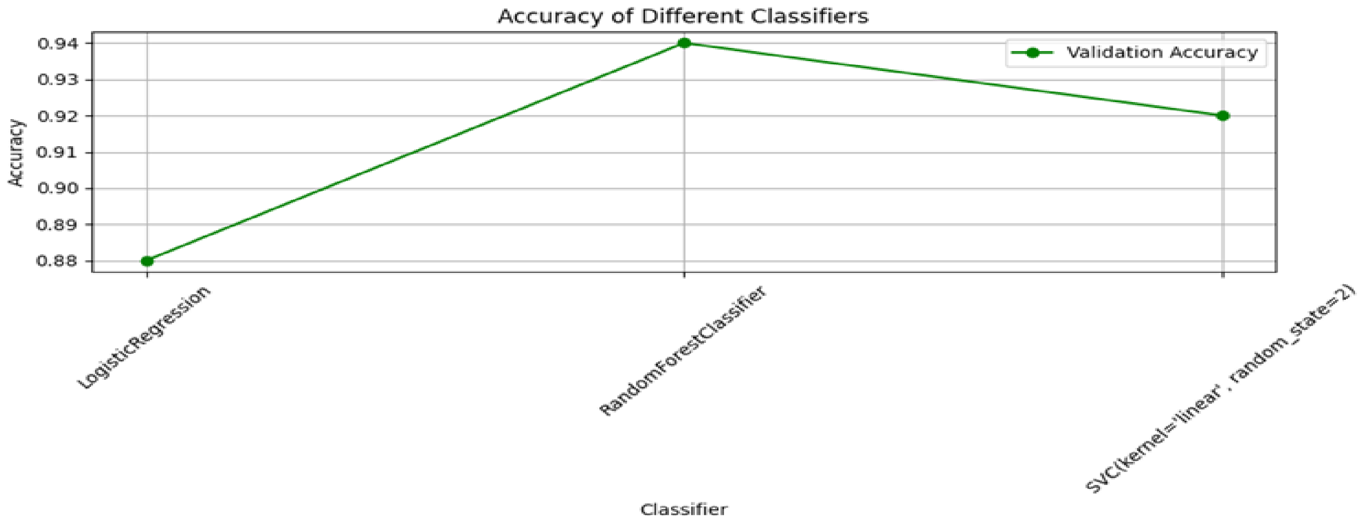
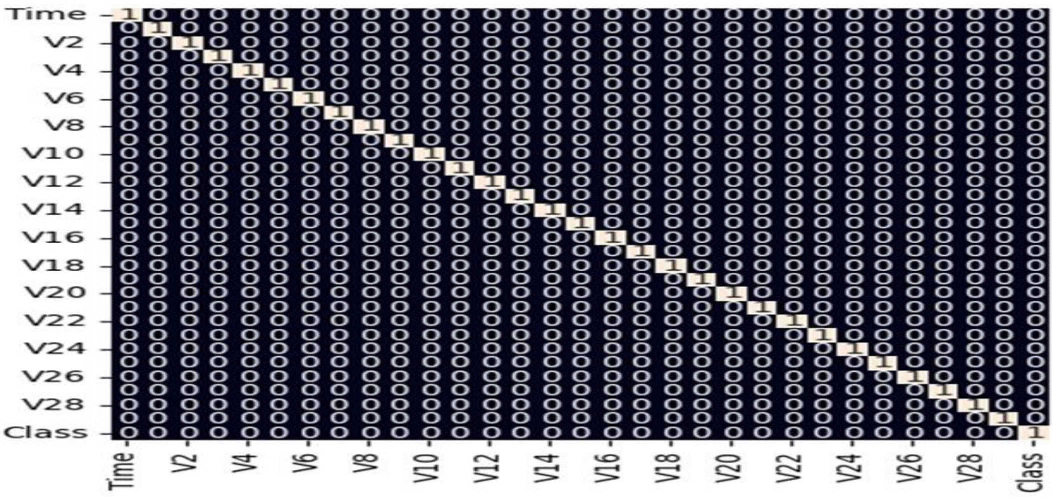


Fig 4: displays a line graph that compares the validation accuracy of various machine learning methods.

Heat map (fig. 5)

A heat map is a graphical representation of data in which individual values in a matrix are represented by colors. This heatmap depicts a correlation matrix.
I. Time from V1 to V28: These are most likely features or variables in your dataset, with 'V1' to 'V28' being principal components obtained by a PCA (Principal Component Analysis) to avoid multicollinearity, or other constructed characteristics.
II. Class: This is most likely the target variable that determines whether a transaction is fraudulent or not.
III. Color Intensity and Circle Size: The deeper the color and the larger the circle in a cell, the greater the relationship between the two variables. A perfect correlation of 1 is depicted on the diagonal from top left to bottom right, with each variable perfectly correlating with itself.
IV. Correlation values range from -1 to 1. A score close to one shows a significant positive correlation, implying that when one measure increases, so does the other. A score around -1 shows a strong negative correlation, implying that if one variable increases, the other tends to decrease. A number near zero indicates no relationship.



V. CONCLUSIONS

In this study, we examined various methods of machine learning to figure out which one is more effective for detecting credit card fraud. We employed three algorithms: logistic regression, random forest, and SVC. Our results indicate that the Random Forest outperforms the other algorithms having an output of 0.94. We utilized a supervised dataset and several kinds of machine learning techniques to figure out which approaches are more suitable for identifying credit card fraud.

Future study might utilize a greater number of machine learning algorithms to achieve higher accuracy than what we stated in our study. If more algorithms and datasets are employed in this study, subsequent research could result in higher accuracy.

After comparing multiple machine learning algorithms, we found that Random Forest performed extremely well for predicting the theft of credit cards in our study. However, new study has the potential to enhance accuracy, assisting cardholders in recognizing credit card theft and protecting themself from the fraudsters.

REFERENCES

[1] Khalid, Abdul Rehman, et al. "Enhancing credit card fraud detection: an ensemble machine learning approach." Big Data and Cognitive Computing 8.1 (2024): 6. DOI No. 10.3390/BDCC8010006.

[2] Khatri, Samidha, Aishwarya Arora, and Arun Prakash Agrawal. "Supervised machine learning algorithms for credit card fraud detection: a comparison." 2020 10th international conference on cloud computing, data science & engineering (confluence). IEEE, 2020.DOI No. 10.1109/CONFLUENCE47617.2020.905785.

[3] Madhurya, M. J., et al. "Exploratory analysis of credit card fraud detection using machine learning techniques." Global Transitions Proceedings 3.1 (2022): 31-37. DOI No.10.1109/CONFLUENCE51453.2020.9132664.

[4] Aftab, A. U., et al. "Fraud Detection of Credit Cards Using Supervised Machine Learning." Pakistan Journal of Emerging Science and Technologies (PJEST 4.3 (2023). DOI No. 10.58619/PJESTV4i3.114.

[5] Sailusha, Ruttala, et al. "Credit card fraud detection using machine learning." 2020 4th international conference on intelligent computing and control systems (ICICCS). IEEE, 2020. DOI No. 10.1109/ICICCS48265.2020.9121114.

[6] Varmedja, Dejan, et al. "Credit card fraud detection-machine learning methods." 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). IEEE, 2019. DOI No. 10.1109/INFOTEH.2019.8717766.

[7] Thennakoon, Anuruddha, et al. "Real-time credit card fraud detection using machine learning." 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019. DOI No. 10.1109/CONFLUENCE.2019.8776942.

[8] [Dornadula, Vaishnavi Nath, and Sa Geetha. "Credit card fraud detection using machine learning algorithms." Procedia computer science 165 (2019): 631-641. DOI No. 10.1016/J.PROCS.2020.01.057.