Quest Journals
Journal of Software Engineering and Simulation
Volume 9 ~ Issue 5 (2023) pp: 55-60
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Credit Card Fraud Detection Using Machine Learning

Parvathi R[#1], Dhyuthi S Vijay[#2], Aparna Madhu[#3], Georgey Daniel[#4]
Final year Students Of Computer Science and Engineering Department, Mahaguru Institute Of Technology,
APJ Abdul Kalam Technological University, Alappuzha, Kerala

Abstract— *In this study, people can use credit cards for online transactions as they provide an efficient and easy-to-use facility. With the increase in usage of credit cards, the capacity for credit card misuse has also increased. Credit card fraud causes significant financial losses for both cardholders and financial companies. In this research study, the main aim is to detect such frauds, including the accessibility of public data, high-class imbalance data, changes in fraud nature, and high rates of false alarm. The relevant literature presents many machine learning-based approaches for credit card detection, such as the Extreme Learning Method, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and XG Boost. However, due to low accuracy, there is still a need to apply state-of-the-art deep learning algorithms to reduce fraud losses. The main focus has been to apply the recent development of deep learning algorithms for this purpose. A comparative analysis of both machine learning and deep learning algorithms was performed to achieve efficient outcomes. A machine learning algorithm was first applied to the dataset, which improved the accuracy of the detection of the frauds to some extent. Later, three architectures based on a convolutional neural network are applied to improve fraud detection performance. The further addition of layers further increased the accuracy of detection. A comprehensive empirical analysis has been carried out by applying variations in the number of hidden layers, epochs, and the latest models. The proposed model outperforms state-of-the-art machine learning and deep learning algorithms for credit card detection problems. In addition, we have performed experiments by balancing the data and applying deep learning algorithms to minimise the false-negative rate. The proposed approaches can be effectively implemented for the real-world detection of credit card fraud. . We use algorithms such as Logistic Regression, Support Vector Machine, XG boost, Random Forest, Decision Tree, and KNN. Over sampling method is used to balance the dataset. Here we use SMOTE[Synthetic Minority Oversampling Technique]. oversampling. In our model, the support vector machine gives more accuracy. The accuracy is given by the ROC [Receiver Operating Characteristic] curve.*

Keywords— *Fraud Detection Deep learning, Machine learning, Credit card frauds.*

***R**eceived 12 May, 2023; **R**evised 24 May, 2023; **A**ccepted 26 May, 2023 © The author(s) 2023.*
P**ublished with open access at **www.questjournals.org

I. INTRODUCTION

Nowadays Credit card usage has drastically increased across the world; now people believe in going cashless and are completely dependent on online transactions[20]. The credit card has made digital transactions easier and more accessible. A huge number of dollars are lost every year due to criminal credit card transactions. Fraud detection is the process of monitoring the transaction behaviour of a cardholder to detect whether an incoming transaction is authentic and authorised or not; otherwise, it will be detected as illicit[13]. There is a rapid growth in the number of credit card transactions, which has led to a substantial rise in fraudulent activities. Credit card fraud detection is a significant, but also popular, problem to solve. In our proposed system, we built the credit card fraud detection using machine learning[14]. Machine learning has been identified as a successful measure for fraud detection. We use algorithms such as Logistic Regression, Support Vector Machine, XG boost, Random Forest, Decision Tree, and KNN. Over sampling method is used to balance the dataset. Here we use SMOTE[Synthetic Minority Oversampling Technique]. oversampling. In our model, the support vector machine gives more accuracy. The accuracy is given by the ROC [Receiver Operating Characteristic] curve.

II. PROBLEM STATEMENT

Our aim here is to detect fraudulent transactions while minimising incorrect fraud classifications.

III. LITERATURE REVIEW

In Credit card fraud detection using classification, unsupervised, neural networks model, International journal of engineering research and technology[IJERT] ,published in 2020 states a large variety of parameters are used to choose and classify. The models they have used for detecting the credit card fraud are: Logistic Regression, K-Means, Convolutional Neural Networks. The different models gives the accuracy of Logistic regression -99.88%, K-means-54.27%, CNN-99.61%. It was clear from their model the Logistic Regression model performed well in dataset. The main disadvantages of this model is here for this dataset grouping is a difficult task because fraud and genuine transactions look very similar. So, it is very difficult to put fraud and genuine transactions into separate groups[2].

In Credit card fraud detection using machine learning, the International Journal of Creative Research Thoughts [IJCRT], published in 2022, states that the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today are to attack end-to-end technology and exploit human vulnerabilities. The algorithms used for this model are random forest, logistic regression, and Nave Bayes. This model includes higher accuracy in fraud detection, less manual work needed for additional verification, and the ability to identify new patterns and adapt to changes[1].

Credit card fraud detection, International Journal of Engineering Research and Technology (IJERT), published in 2021 It focuses on credit card fraud and its detection measures. Credit card fraud occurs when one individual uses another individual's card for their personal use without the knowledge of its owner. In this paper, machine learning algorithms are used to detect credit card fraud. To evaluate the model's efficacy, a publicly available credit card data set is used. Its main objective is the implementation and evaluation of the framework as a tool for credit card fraud detection[4].

In Credit Card Fraud Detection using Machine Learning Algorithms, published in 2019, the objective is to design and develop a novel fraud detection method for streaming transaction data with the objective of analysing the past transaction details of the customers and extracting behavioural patterns. Where cardholders are clustered into different groups based on their transaction amount. Then, using the sliding window strategy, aggregate the transactions made by the cardholders from different groups so that the behavioural patterns of the groups can be extracted respectively. Later, different classifiers are trained on the groups separately. And then the classifier with a better rating score can be chosen as one of the best methods to predict fraud. Thus, followed by a feedback mechanism to solve the problem of concept drift, In this paper, they worked with a European credit card fraud dataset. They use multiple supervised and semi-supervised machine learning techniques for fraud detection. Different supervised machine learning algorithms like decision trees, naive Bayes classification, least squares regression, logistic regression, and SVM are used to detect fraudulent transactions[3].

IV. SYSTEM MODEL AND METHODOLOGY

A. System Overview

In our system, we use some algorithms such as Logistic Regression (LG), Support Vector Machine (SVM), XG boost, Random Forest, Decision Tree, and K-nearest neighbours (KNN).

Logistic regression is used for predicting the categorical dependent variable using a given set of independent variables. Instead of giving the exact value as 0 and 1, it gives the probabilistic values, which lie between 0 and 1.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the boundary, which is called a hyperplane.

Works by finding the K nearest neighbours to a given data point based on a distance metric and then making a prediction based on the class or value of those neighbours.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction The Decision of the majority of the trees is chosen by the random forest as the final decision

In our system, we use the oversampling method for data balancing. After the testing, we can see that SVM has higher accuracy.

The accuracy is given by the ROC [Receiver Operating Characteristic]curve.

B. Dataset details

The dataset contains transactions made by credit cards in September 2013 by European cardholders. 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) accounts for 0.172% of all transactions. It contains only numerical input variables, which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features or more background information about the data. Features V1, V2,... V28 are the principal components obtained with PCA; the only features that have not been transformed with PCA are "time" and "amount." Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction amount; this feature can be used for

example-dependent, cost-sensitive learning. Feature 'Class' is the response variable, and it takes a value of 1 in the case of fraud and 0 otherwise. Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification[6].

C. PCA

Principal Component Analysis (PCA) is a dimensionality- reduction technique that is often used to transform a high- dimensional dataset into a smaller-dimensional subspace. ● PCA is mathematically defined as on orthogonal linear transformation that transform the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first some projection principal component), the second greatest variance on the second coordinate, and so on.

D. SMOTE

SMOTE stands for Synthetic Minority Oversampling Technique. This is a statistical technique to increase the number samples in the minority class in the dataset to make it balanced. It works by generating new instances of data from existing data by taking feature space of each target class and its nearest neighbours.

E. Proposed system

We are considering an open source dataset (credit card.csv file) whose reading can be done using the Pandas library. We are doing preprocessing steps on the above dataset. It involves null value checking, scaling, etc.
For model building, we are dividing the dataset into two parts: train and test. We are using the Sklearn library for this.
Model building can be done using Machine learning algorithms such as Logistic Regression (LG), Support Vector Machine (SVM), XG boost, Random Forest, Decision Tree, and K-nearest neighbours (KNN).
Model training can be done by passing the data into each of the algorithms. and plotting the roc curve using the algorithms
Model testing can be done by passing test values to the trained model. So it will predict the output, and we can compare these with the original values. Performance can be evaluated using a confusion matrix. A model can be finalised based on test results general accuracy.

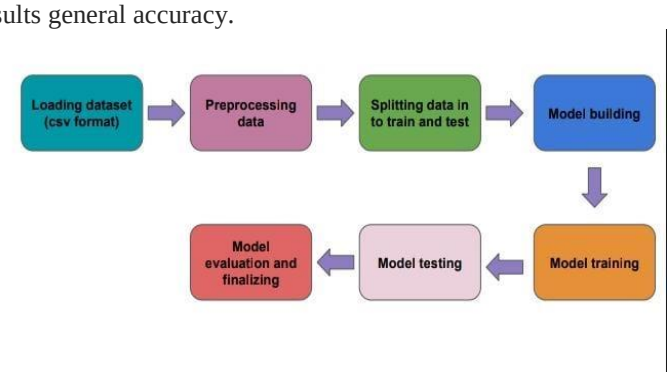


Fig : Proposed system

F. Software Requirements

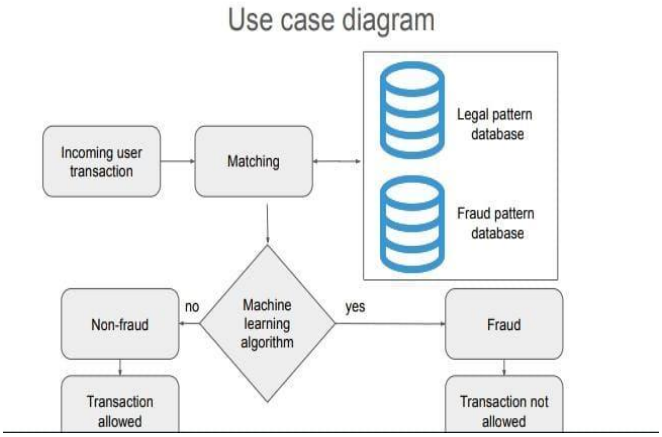
1. **Jupyter Notebook:** The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualisations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. The Jupyter Notebook is not included with Python, so if you want to try it out, you will need to install Jupyter[7].
2. **Anaconda Navigator:** Anaconda Navigator is a desktop graphical user interface (GUI) that allows to launch applications and manage conda packages, environments, and channels without using comment line interface (CLI) commands, which is available for Windows, macOS, and Linux[8].

LIBRARIES USED

1. **NumPy**: It is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, the Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project, and you can use it freely. NumPy stands for numerical Python[15].
2. **Pandas:** It is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "panel data", and "python data analysis[16]."

3. **Keras :** It is a high-level, deep learning API developed by Google for implementing neural networks. It is written in Python and is used to make the implementation of neural networks easy. It also supports multiple backend neural network computations[17].
4. **Sklearn:** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling[18].

V. USE CASE DIAGRAM



VI. EXPERIMENT AND RESULTS

We use algorithms such as Logistic Regression, Support vector Machine, XG boost, Random Forest, Decision Tree, and KNN. In our model, the support vector machine gives more accuracy.

A. *Implementation*

Data Collection: 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) accounts for 0.172% of all transactions. It contains only numerical input variables, which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features or more background information about the data. Features V1, V2,... V28 are the principal components obtained with PCA.

Data preprocessing: We are doing preprocessing steps on the above dataset. It involves null value checking, scaling, etc.

Model building: For model building, we are dividing the dataset into two parts: train and test. We are using the Sklearn library for this. Model building can be done using Machine learning algorithms such as Logistic Regression (LG), Support Vector Machine (SVM), XG boost, Random Forest, Decision Tree, and K-nearest neighbours (KNN).

Model training: Model training can be done by passing the data into each algorithm and plotting the roc curve using the algorithms.

Model testing: Model testing can be done by passing test values to the trained model. So it will predict the output, and we can compare these with the original values. Performance can be evaluated using a confusion matrix. After testing, the support vector machine gives higher accuracy.

.

B. *Results*

In our system we use some algorithms such as Logistic Regression(LG), Support Vector Machine(SVM), XG Boost, Random Forest, Decision tree, K-Nearest Neighbours(KNN).

Logistic Regression used for predicting the categorical dependent variable using a given set of independent variables .Instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in boundary is called a hyperplane.

Works by finding the K nearest neighbours to a given data point based on a distance metric, and then making a prediction based on the class or value of those neighbours.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction The Decision of the majority of the trees is chosen by the random forest as the final decision

In our system, we use the oversampling method for data balancing. After the testing, we can see that SVM has higher accuracy.

SVM gives accuracy of 97%

The accuracy is given by the ROC [Receiver Operating Characteristic]curve.

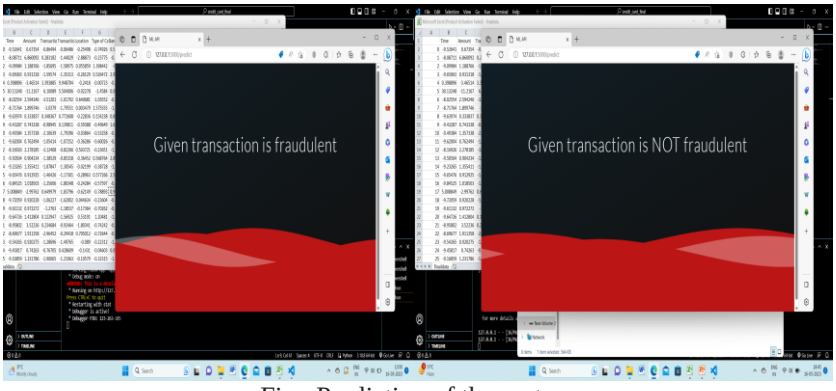
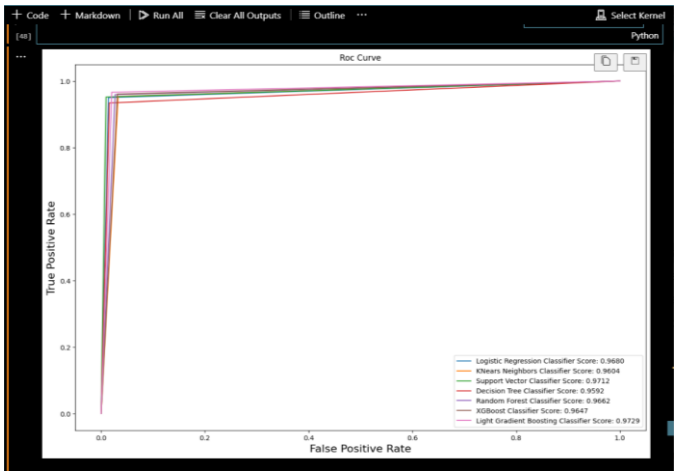


Fig : Prediction of the system

VII. ROC CURVE



VIII. CONCLUSION

Credit Card Fraud is an increasing threat to financial institutions. Fraudsters tend to constantly come up with new fraud methods. Accurately predicting fraud cases and reducing false-positive cases is the foremost priority of a fraud detection system. The performance of ML methods varies for each individual business case[19]. The type of input data is a dominant factor that drives different ML methods. For detecting CCF, the number of features, number of transactions, and correlation between the features are essential factors in determining the model's performance. We use csv formatted dataset. The datas are highly private. Imbalanced data that is most of the transactions are non fraudulent which make it really hard for detecting the fraudulent ones. It is difficult to obtain available credit card data sets since the security, privacy and cost issues. an abnormal transaction detection system is important for fast and accurate detection, and research is needed to improve the algorithm. In future some other datasets can be used for further testing of proposed mechanisms. So, we conclude that the system is 97.3% accurate from the above ROC curve.

REFERENCES

[1]. https://ijcrt.org/papers/IJCRT2108185.pdf
[2]. https://www.ijert.org/research/credit-card-fraud-detection-using-classification-unsupervised-neural-networks-models-IJERTV9IS040749.pdf
[3]. https://ieeexplore.ieee.org/document/9718341
[4]. https://www.ijert.org/credit-card-fraud-detection
[5]. https://towardsdatascience.com/the-random-forestalgorithm-d457d499ffcd
https://www.xoriant.com/blog/productengineering/decision-trees-machine-learningalgorithm.html
[6]. https://www.google.com/search?q=kaggle+dataset+of+credit+card+fraud+detection&oq=kaggle&aqs=chrome.0.69i59j46i199i433i465i512j0i433i512l2j0i512j0i131i433i512j0i512j0i131i433i512.4712j0j15&sourceid=chrome&ie=UTF-8
[7]. https://realpython.com/jupyter-notebook-introduction/
[8]. https://docs.anaconda.com/free/navigator/index.html
[9]. https://towardsdatascience.com/credit-card-fraud-detection-using-machine-learning-python-5b098d4a8edc
[10]. https://ijarcce.com/wp-content/uploads/2015/12/IJARCCE-94.pdf
[11]. https://www.researchgate.net/publication/317609932_Database_Implementation_and_Testing_of_Dynamic_Credit_Card_Fraud_Detection_System
[12]. https://towardsdatascience.com/credit-card-fraud-detection-using-machine-learning-python-5b098d4a8edc
[13]. https://www.sciencedirect.com/science/article/pii/S187705092030065X
[14]. https://ieeexplore.ieee.org/document/9755930
[15]. https://www.w3schools.com/python/numpy/numpy_intro.asp
[16]. https://www.geeksforgeeks.org/introduction-to-pandas-in-python/
[17]. https://www.tutorialspoint.com/keras/keras_introduction.htm
[18]. https://en.wikipedia.org/wiki/Scikit-learn
[19]. https://www.ripublication.com/ijaer18/ijaerv13n24_18.pdf
[20]. https://ijcsmc.com/docs/papers/April2021/V10I4202112.pdf
[21]. https://www.ijraset.com/research-paper/credit-card-fraud-detection-using-ml