Random Forest for Credit Card Fraud Detection

Shivabasayya Kulkarni [1], Doddamani Basavaraj[2], and Bhagyashri M[3]
[1]*Senior Scale Lecturer, Dept of CSE, Government Polytechnic for Women, Hubli*
[2] *Senior Scale Lecturer, Dept of CSE, Government Polytechnic for Women, Hubli*
[3] *Lecturer, Dept of CSE, Government Polytechnic for Women, Hubli*

***Abstract*— The proposed framework is introduced using real-world credit card transactions for fraud detection. As credit card transaction increases, fraudulent transactions also increase exponentially. Banks also seek to improve the identification of fraudulent transactions to reduce losses. The fraudulent transaction primarily occurs because the card or cardholder needs not to be present during the transaction process and the seller has no way to check whether the payment is made by the cardholder or not. The proposed method uses the algorithm, Random Forest, to locate fraud transactions and precision for such transactions. Decision trees are used for dataset classification. Random forest output is a confusion matrix used to evaluate results.**

***Index Terms*— Random Forest, Credit Card, Fraud Detection, Fraudulent Transaction.**

I. INTRODUCTION

In these days, the usage of credit card platforms, such as Amazon , Flipkart and Snap Deal, is highly common. Two forms of transactions occur, including online transactions and offline transactions. With the advent of the UPI (Unified Payment Interface), the use of online purchases is increasing dramatically.

You don't have to attend the transaction for the online scenario card. In order to render the purchase, users just need to insert the card information in the web portal. The fraudsters will rob these confidential data using software while entering information that can monitor the button click. Therefore the information should always be entered using a virtual keyboard. The bank is increasingly concerned with these fraud transactions and is working hard to minimize this issue. The card must be available in the offline scenario to do purchases. In the case of fraud purchases, fraudsters intercept the data of the issuer in the actual issuer and use another type of intercept data from RF interfaces. The computer extracts the card details as the system is in near contact to the wallet.

Credit card conveniences have rendered it much cheaper to utilize online transfers and avoid money losing. The analysis indicates an increasing growth in the amount of fraud transactions before 2021. The principal issue with both these fraud activities is that the issuer of the payment may not have to be present for these purchases and the actual device may therefore not be available. The fraudsters take advantage of both ways to conduct new fraud. The bankers are still seeking to increase the levels of fraud identification, but no progress has occurred. The purchase is called a crime if the cardholder does not execute the purchase; the behavior of the purchases would be examined. There are two forms of suspicious transactions: abuse identification and anomaly detection.

The credit cards are made of plastic, consisting of an internal chip containing cardholder banking information and a magnet strip used to access the payment system for purchases. Many credit cards have a contactless payment through which purchases can only be done by having direct contact with the payment system and inside the card is a chip NFC (Near Field Communication). Credit cards provide a easy option for cardholders to acquire items in retail shop or through an online retailer. Since online payment systems are increasingly prevalent, credit card consumers are growing-day by day, which contributed to a huge amount of fraud purchases worldwide.

In the financial processing portal, fraudsters discover the shortcut or workaround and use it to their benefit to conduct fraud transactions. Cardholders, retailers and banks incur financial damages worldwide each year as a consequence of rising fraud purchases. The banking sector still attempts to use some form of cutting-edge technology to reduce the issue in the payment portal, but fraudsters find different methods of transacting fraud. In the banking sector this has been a big concern and this initiative would undoubtedly help to reduce this issue.

II. PROBLEM STATEMENT

Since online shopping is increasing in popularity, the incidence of fraud transactions is growing dramatically, depending mainly on the bank payment passport to allow online transactions. It is becoming more difficult to detect such fraud transactions as the purchase does not require a physical card. Anybody who knows the specifics of the card will make an online purchase, and the cardholder can only get to know this after these transactions are carried out. To minimise the vast number of fraud transactions, suitable algorithms have to be implemented. This project develops and establishes a methodology that categorises fraud and non-fraud transactions.

III. LITERATURE SURVEY

Thomas G. Thomas et al., [3] The author in this article suggests an agent method based on an assembly of forests and neural networks using ensemble methods. Credit card fraud identification, by way of the Ensemble machine learning technique the activity is categorized into a legal transaction or a fraud transaction. The Learning Ensemble trains the data set by splitting the data set into study, test and validation data. For both legally-regarded activity and fraud transactions, the ratio of this data set is as follows. 20% legit and scam test data and 20% eventually validation data kit containing legitimate and illegal purchases. 20% validation data bundle. Once the data-set is broken, the algorithm is eventually applied to identify the fraud transaction.

In order for the random forest trees to operate, this proposed method [3] works well: the bootstrapped values are added first for training data to each of the forest trees. Secondly, only a limited portion of the data is chosen arbitrarily for the creation of different trees. The random method of forests is a combined "bagging" and random method of subspace. The packaging method is built using individual models on training dataset values by sampling an ensemble method. In the random subspace process , a set is created for individual trees using the random data values of an attribute. Remember that the training data set consists of N instances, there are B attributes in such cases and each tree is constructed according to ensemble method. The first step is to gather N samples of bootstraps for the sample. The second stage would be to pick the random samples of attributes b and then decide the division for the collection of attributes b and create a tree that must be planted entirely without overgrowing of any trees. In this way trees are grown independently of the conditions of each other, because the estimation performance is greater than the other approaches.

They suggested a system using a vector support machine (SVM) to classify credit card fraud [7]. Siddhartha Bhattacharyya et al. The supported vector machine is used to divide the values of the dataset into distinct classes using a line or hyperplace from which point the reference line is equivalent.. The difference between the point and the reference line is vector.

R Devaki et al . [4] presented a model utilising methodology based on distances and methodology based on marks. The current transaction is calculated by distance based comparison with previous transaction histories and when the amount of current transactions approaches the previous transaction history, the present payment is suspicious, since fraudsters have created a fraud transaction and a number of secret questions are asked to answer in order to proceed with the transaction, if the user gives the correct ans. The distance method identifies the cardholder 's expenditure behaviour and then performs the distance detection process.

This method[4] is used to evaluate the history of previous transactions using the technique for estimation of marks. Every transaction is delegated by the method to low, medium and high; these assignments represent the possibility of each transaction. The K-means approach is used to identify the distance from the transaction cluster when the gap is high, the transaction shall be treated as a fraud transaction, and if the cardholder addresses all the issues correctly, it shall cause the transaction to take effect.

They proposed utilising the Secret Markov model to identify fraud, Mohd Avesh Khan et al.[5]. The Secret Markov model is used to distinguish between low , medium and high cardholders' purchase patterns. These thresholds are allocated to each account for each cardholder. Measured to prior levels of likelihood, the actual transaction frequency would be measured whenever the machine can detect a difference below a predefined standard, so the machine would interrupt the transaction and label the transaction as a fraud activity. The efficiency of fraud detection systems would not improve in these current systems. The connexion port is not sufficiently secure for online transactions and routine maintenance of the gateway.

HMM generates its own expenditure profile in the technologies provided [5] on the basis of the features of the cardholder's purchases. The actual transactions are contrasted with the profile of the Secret Markov model when the profile has been generated and, should there be a difference, the framework flags the transaction as fraud.

IV. IMPLEMENTATION

Information on the functions used to execute the proposed approach are given in the implementation section. The accuracy of fraud is detected by machine learning techniques. The models are equipped with the random forest classifier. For the implementation of the Random Forestry Classifier, the python programming language is used.

Algorithm for Random forest classifier:

The following algorithm summarizes the predictive efficiency of the random forest classification for python libraries' fraud transactions. Machine simulation approaches are used to train data collection for the random classification of forests.

```
Random-forest-classifier:
Input: The features derived from the data collection are the algorithm input
Output: model accuracy is returned as output
Step 1: The sum 'm' is randomly chosen from the 'k' dataset such that the amount 'k' is less than 'm'.
Step 2: By using the matplotlib of the python visualize the dataset in two separate group such as nonfraud and fraud transaction by considering the transaction amount and time.
Step 3: To achieve the degree of randomness shuffle the dataset.
Step 4: By using the split ratio we can divide the decision node 'd' with respect to 'k' features.
Step 5: Using the splitting method, break into knots.
Step 6: Use the test and testing samples to practice the random forest classifier algorithm. If the model has been conditioned, offer new knowledge to test the system's consistency. By creating the uncertainty matrix, the consistency of the method can be shown.
```

Algorithm for cleaning null values:

The following method illustrates how the Kaggle repository cleans the credit card data collection.

Cleaning-Algorithm:

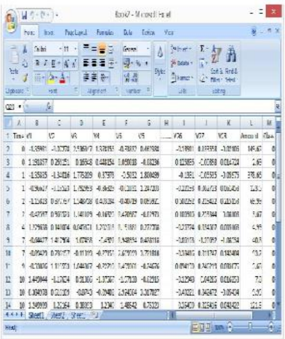
```
Input: threshold for the row and columns of the dataset Output: dataset are cleaned.
Step 1: Identify the null values present in the dataset
Step 2: count how many rows and columns of the dataset which has the null values with respect to column.
Step 3: When the count values are over the row threshold then increase the count row, print and remove row values with null values.
Step 4: Count how many columns having the null values with respect to row in the dataset.
Step 5: increase the column threshold if and only if count value is greater than column threshold and print the column number which contain the null values and drop those null values from their respective columns.
```

Algorithm for Multilayer Perceptron:

```
Input: features which are collected from the dataset
Output: model accuracy is pushed as the output.
Step 1: Using the Kaggle repository choose the accurate dataset for the proposed system.
Step 2: Design the neural network using the input, hidden and output layers. Before the hidden layer the activation function need to be chosen accurately.
Step 3: Choose the proper weights for testing the model using the testing and training dataset. The learning rate, training time are the parameters can be employed during designing he neural network.
Step 4: Measure the type of multilayer perceptrons using train and test samples and measure their precision.
```

V. EXPERIMENTAL RESULTS AND ANALYSIS

Time, quantity and classification features of the studies are regarded. The 'Time' function includes the seconds between each data set operation. The 'Sum' function is the number of transactions, the 'Level' function is the answer attribute, and the meaning is 1 for illegal and 0 for legal. This section contains the screen pictures and the description for each screenshot of the credit card scanning algorithm. There are 2,84,808 purchases in the payment card dataset. Figures 5.1 and 5.2 represent both the first and last pages of the payment card dataset.



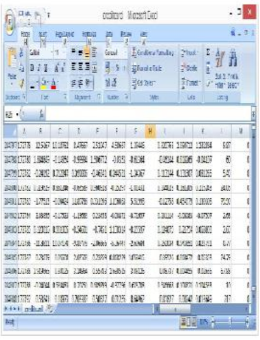
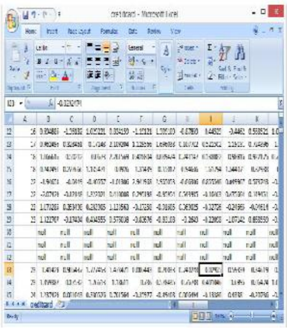


Figure 5.1: First page of credit card dataset
Figure 5.2: Current payment card collection link

The data set snapshot is seen in figure 5.4 before cleaning. The cleaning efficiency is seen in Figure 5.5, except the zero values in the dataset which occurs when the dataset is cleaned.



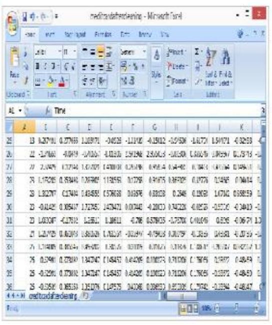
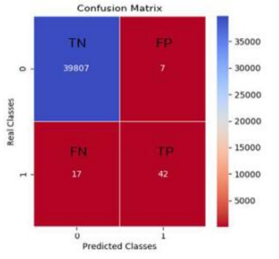


Figure 5.4: Dataset screenshot before cleaning.
Figure 5.5: Snapshot of the test after data collection cleanup.

The random forestry classifier describing Figure 5.6 and Figure 5.7 accounts for the uncertainty matrix. A matrix of uncertainty consists of the classifier's effects estimates that are both accurate and wrong. In order to determine the precision and efficiency of the classifier, the uncertainty matrix is used. The uncertainty matrix allows to calculate the complete description of the classification algorithm 's predictions.



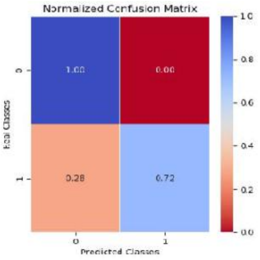
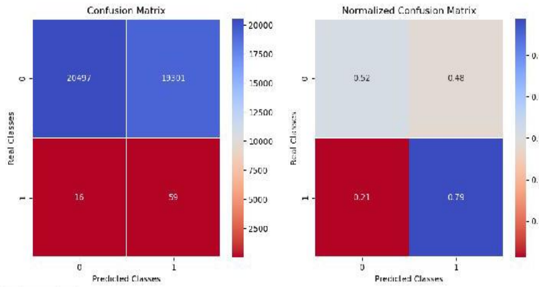


Figure 5.6: Confusion matrix.
Figure 5.7: Normalized confusion matrix.

Trained and monitored data is fitted with a 4 block matrix of ambiguity true positive, true negative, false positive, false negative. In a totally decent case, the classifier predicts that the contract is fraud and also that the actual transaction is fraud. This demonstrates that the selling was correctly identified by the classifier as a scam. The classifier forecasts the transaction as legible, in the very negative scenario, as well as the actual transaction. The transaction was correctly considered valid by the classifier here. The classifier predicts the transaction to be incorrect correctly, but the real transaction was true. In this situation, the classifier mispronounced the transaction as being fake. The classifier forecasts the transaction as valid in false negative terms and the transaction is simply fraudulent. In this scenario, the classifier mispredicted the transaction as true.

Figure 5.8 represents the multi-layer experience classifier uncertainty matrix.



Calculation of Accuracy:

Depending on the situation, classification rate or preciseness can be great, decent, bad or awful. Precision measurement for RFA and MLP,

❖ Random forest classification accuracy is determined as follows,

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (42+ 39807) / (44 + 39807 + 07 + 17)
= 99.9%

❖ Accuracy for Multilayer perceptron is calculated as,

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (59+20497)/(59 +20497+19301+16)
= 51.5%

DISCUSSION

Table 1 discusses the findings of the random forest model along with the findings of a multilayer perceptron solution applied utilising the same dataset (exactness). Table 1: Random Forest Grouping and Multilayer Perceptron Comparative Assessment.

Model	Accuracy (%)
Random Forest Model	99.9
Multilayer perceptron	51.5

VI. CONCLUSION AND FUTURE SCOPE

The proposed method investigates how the Multilayer Perceptron and Random Forest model work. This approach uses a real-life payment data set of credit cards from B2C. The random forest algorithm performs better with more training data but not time. The precision of the identification of credit card fraud by random selection is 99,93 per cent. In order to diagnose credit card theft, the Multilayer perceptron model achieved 51.58 per cent. In contrast with the multilayer perceptron model, a rogue forest model has greater precision in fraud detection.

Using the techniques of deep learning and artificial intelligence the detection of the fraud cabe improved in future.

REFERENCES

[1] J Richard Bolton and J David Hand,"Unsupervised Profiling Methods for Fraud Detection", Proceedings of the VII Conference on Credit Scoring and Credit Control, pp. 235-255, 2001

[2] T.S Jon Quah and M Sriganesh, "Real-time credit card fraud detection using computational intelligence", International Journal Expert Systems with Applications, vol. 35, no. 4, pp.1721–1732, 2008

[3] G Thomas Dietterich, "Ensemble Methods in Machine Learning", International Conference on Information Security, pp. 1-15, 2000

[4] R Devaki, V Kathiresan, and S Gunasekaran," Credit Card Fraud Detection using Time Series Analysis", International Conference on Simulations in Computing Nexus, pp. 8-10, 2014

[5] MohdAvesh Zubair Khan , Jabir Daud Pathan and Ali Haider Ekbal Ahmed", Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering", International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 2, pp.1-4, 2014

[6] EkremDuman and M Hamdi Ozcelik ,"Detecting credit card fraud by genetic algorithm and scatter search", International Journal of Expert Systems with Applications, vol. 38, no.10,pp. 13057–13063, 2011

[7] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, J ChristopherWestland,"Data mining for credit card fraud: A comparative study", International Journal of Decision Support Systems, vol.50, no.3, pp.602–613, 2011

[8] Veronique Van Vlasselaera , Cristian Bravob, Olivier Caelenc , Tina Eliassi-Radd , Leman Akoglue , Monique Snoecka and Bart Baesens, "APATE: A Novel Approach for Automated Credit Card Transaction Fraud Detection using Network-Based Extensions", International Journal of Decision Support Systems, vol.75, no.8, pp.38-48, 2015

[9] Lutao zheng, Guanjun Liu, Wenjing Luan, Zhengchuan Li, Yuwei Zhang, Chungang Yan and Changjun Jiang, “A New Credit Card Fraud Detecting Method Based on Behaviour Certificate”, IEEE 15th International Conference on Networking, Sensing and Control, pp.150-158, 2018

[10] Rong-Chang Chen, "A new binary support vector system for increasing detection rate of credit card fraud", International journal of pattern recognition and artificial intelligence, vol. 20, no. 2 ,pp.227–239,2006