Applied Information Systems and Management (AISM)
Volume 4, (2) 2021, hal 71-76
P-ISSN: 2621-2536 ;E-ISSN: 2621-2544; DOI: https://doi.org/10.15408/aism.v4i2.20570

Credit Card Fraud Detection Using Machine Learning Approach

Kanal Bhadresh Soni[1], Madhuri Chopade[2], Rahul Vaghela[3]

Abstract **—Using new spam technologies to carry out internet banking fraud refers to shifting and withdrawing money from the user's balance account without it's authorization. Credit card fraud pops into the mind so far in the current scenario when the concept of fraud bursts into some conversation. Credit card fraud has escalated tremendously in recent times due to the incredible growth in credit card purchases. In order to assess, identify or prevent undesirable conduct, fraud detection requires tracking the purchase behavior of users/customers. The purpose of this project is to predict the genuine and fraud transactions with respect to the amount of the transaction utilizing various machine learning approaches like Logistic Regression, Decision Trees, Support Vector Machine, Naïve Bayes, Random Forest and K-Nearest Neighbor. The model built who has greater accuracy and precision is considered to be best fit for this system.**

Index Terms-Application of Machine Learning, Decision Trees, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine.

I. INTRODUCTION

Fraud in simple words can be termed as an unfair or fraudulent activity expected to result in personal and financial gain, or to injure another individual without actually contributing to clear legal impacts. The two key measures to eliminate frauds and damages due to the unethical activities are fraud avoidance and fraud detection systems. The constructive mechanism with the aim of blocking the phenomenon of fraud is fraud prevention. The constructive method with the objective of preventing the incidence of fraud is fraud prevention. When scammers overtake the fraud prevention networks and initiate a fraudulent transaction, fraud detection systems come into consideration. No one can really recognize whether the prevention procedures have been activated by a fraudulent transaction. The intention of detection techniques is often to evaluate each transaction for the likelihood of fraud, irrespective of the prevention techniques, and to detect fraud ones as rapidly as possible after a fraudulent transaction has started to be executed by the fraudster. The most popular forms of cheating are fraud activities in credit card and e-commerce networks, laundering in financial systems, computer network cyberattacks, fraudulent conversations or utilization of some services in the field of healthcare and telecommunication structures.

Credit card typically refers to a card granted to the consumer (credit card issuer), generally enabling them to buy products or services or borrow cash in advance under the credit limits. The credit card gives the cardholder the benefit of the moment to pay the bills later in the next cycle. By bringing it through the next payment period, the credit card provides the cardholder with a benefit of time or that moment. As a very significant unique card number, each card's safety relies primarily on the physical safety of the card and the secrecy of the card number.

Credit card fraud is a common criminal activity undertaken as a fraudulent source of transaction payments using only a credit card or other related method of payment. The intention may be to receive goods from an account without spending, or to gain unauthorized funding. Often, credit card fraud is an addendum to fraudulent activity. The fraud starts with either the stealing of the physical card or the misuse of account-related information, like the account number of the card or other records that during a legitimate transaction will commonly and necessarily be accessible to a customer. The compromise can arise on several common routes and can often be conducted without the card holder, the customer or the issuer being warned off, at least before the account is finally used for fraud. Before obtaining an expense report, which may be sent frequently, the cardholder may not identify fraudulent usage [11].

The rest of the paper is assembled in Sections. Section-2 elaborates the literature review of the respective problem, Section-3 represents the information of utilized dataset along with used environment for implementation part, Section-4 describes the approaches for implementation, and finally Section-5 consists of the generated conclusions of the shown study.

A. Types of Credit Card Frauds

The credit card scams are carried out in many ways. In this paper, many of it's forms are covered completely.

1) Application fraud: It commonly occurs in combination with identity fraud. It takes place when offenders request on your behalf for a new credit card.

2) Magnetic/Electronic Card Fraud: Electronic Card Imprints: This indicates that the data put on the card's magnetic strip is browsed by another.

3) CNP Fraud: CNP refers to Card Not Present. In this form of fraud, offenders will perform CNP fraud against you if anybody discovers the end date of your card and account

Received:2 May 2021 ; Revised: 23 September 2021 ; Accepted: 26 September 2021

[1]K. B. Soni, *Gandhinagar Institute of Technology Gandhinagar, India* (email: kanalsoni015@gmail.com)
[2]M. Chopade, *Gandhinagar Institute of Technology, Gandhinagar, India* (email: madhuri.chopade@git.org.in)
[3]R. Vaghela, *Gandhinagar Institute of Technology, Gandhinagar, India* (email: ravaghea83@gmail.com)

number of your card. This fraud can be completed by mobile, email or the web.

4) Intercept Fraud: This form of fraud is also termed as Mail non-receipt credit card scam. You were awaiting a new card or alternative in this scenario and an attacker is able to intercept them.

5) Suspected Identity Fraud: With suspected identity theft, to acquire a credit card, a suspect may use a provisional address and a false name.

6) Doctored Credit Card Fraud: A doctored credit card is a card where the metallic band has been removed by a powerful magnet. This is done by offenders and they then attempt to alter the information on the card itself to fit those of legal cards.

7) Fake Card Making Fraud: The production of fake card is hard. But the thieves can construct them by placing the chip, magnetic band and also holograms in the fake card. They may use false numbers and names to counterfeit this type of credit card and can carry out transactions with that card.

8) Acquisition of Account: One of the really popular types of credit card fraud is simply the acquisition of accounts. Apparently, somehow, a thief can manage to get control of all of your data and related documents. Normally, this is achieved online.

9) Stolen or Lost Credit Card Fraud: Here, either by robbery or because you dropped it, your card will be removed from your control. In order to make transactions, the offenders who get their hands on it would then use it.

10) Card Id Fraud: Card ID Fraud occurs when an offender becomes notified of your card information, and this data can then be used to control over a card account or open the new one. For all this, your name is being used.

The rough design of how the credit card fraud is taken in control is represented in the figure 1.

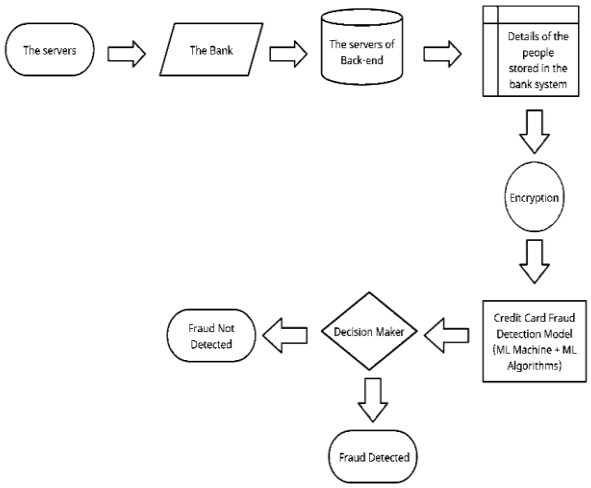

Fig. 1. Architecture of Credit Card Fraud Detection

In this paper, various Machine Learning approaches are implemented to verify that which algorithm of Machine Learning delivers the most efficient outcome and detects the fraud fast underneath whatever the situations and they are as follows:

1) K-Nearest Neighbor (KNN)

2) Logistic Regression (LR)

3) Support Vector Machine (SVM)

4) Naïve Bayes (NB)

5) Decision Trees (DT or CART)

6) Random Forest (RF)

II. Literature Review

Fraud is an illegitimate or unethical activity designed to give birth to financial or personal gain. It is a malicious attack that, in order to acquire illegitimate monetary advantage, is against the guideline, law or policy. Several researches on phenomenon or fraud detection in this context already have been released and are open for general usage.

A detailed survey done by Clifton Phua and his collaborators disclosed that the approaches used in this area include aspects of data mining, automated detection of fraud, and adversarial tracking. In further study, approaches such as Unsupervised and Supervised Learning for credit card fraud detection were proposed by Suman, Research Scholar at Hisar HCE.

Abhimanyu Roy [1] have presented deep learning structures for online currency transfer fraud detection. This method is derived from the artificial neural network of long-term or short-term memory as well as many other factors with in-built time and memory resources.

Wen-Fang YU and Na Wang [2] proposed a related field of investigation in which they used Outlier mining, Outlier detection mining and Distance total algorithms to accurately determine fraudulent transactions in a credit card transaction data set simulation experiment of the certain banking system.

Zahra Kazemi [3] suggested a Deep auto - encoder that is used to retrieve the key qualities of the credit card transaction records. Softmax tools will also be applied to deal with the problems with class labels. To map the information into a high dimensional feature space, an overcomplete auto - encoder is being used and a simple model is being used in a detailed way that provides advantages for the detection of a form of fraud.

Dheeraj Singh and Rinky Patel talk about imbalanced datasets in their proposed research and also how to deal with it and also talk about how to function on huge datasets. These issues have been resolved by the work enacted [4].

Krishna Modi [5] explored various approaches used to track fraudulent activity and made a comprehensive analysis between them. By using any any of these or using both of these approaches, fraudulent transactions can be tracked.

Andreas Prodromidis and Salvatore Stolfo utilized an ensemble model that was based on risk to acquire best outcomes and to erase the disturbance in whole process of transaction [6].

Applied Information Systems and Management (AISM)
Volume 4, (2) 2021, hal 71-76
P-ISSN: 2621-2536 ;E-ISSN: 2621-2544; DOI: https://doi.org/10.15408/aism.v4i2.20570

The researcher named Shiyang Xuan [7] relates the results of the analysis of credit card fraud with those two random forests, which are identified on the basis of their classifier.

Randhawa Kuldeep [8] applied some noise between 10-30% to the sample recorded data for further analysis of the hybrid models. A strong ranking of 0.942 for 30 percent additive noise has been received by plenty of voting methods. It was found that the voting system offered very reliable efficiency in the presence of noise.

S. Ghosh along with D.L. Reilly [9] concluded after their research that K-Nearest Neighbor algorithm provides the best sensitivity and specificity of given parameters but this algorithm does not provide better accuracy.

III. Dataset Information and Used Environment

In this project, the utilized dataset has all the transaction history of Europe. It was revealed by the Europe credit-card holders in the month of September, 2013. The data set contains 284807 rows and 31 columns out of which 492 cases of credit card fraud are detected which covers 0.2% of the complete data set. Genuine transactions found in the data set covers 99.8% of the data.

The data set taken here is imbalanced and it contains only numeric values that are the outcomes of the PCA transformation. In the data set, only the columns named 'Time' and 'Amount' aren't transformed. The column named 'Class' is our response variable which displays 0 in the case of genuine or not fraud transactions and 1 in the case of not genuine or fraud transactions. Further, there are no null values in the whole data set which shows that the whole data is clean. The details of the type of the columns is represented in table 1.

Table 1.
Table Name

Column	*Names Non-Null*	*Count Dtype*
Time	284807 non-null	float64
V1 to V28	284807 non-null	float64
Amount	284807 non-null	float64
Class (0=not fraud,1=fraud)	284807 non-null	int64

The distribution of data is done by portioning it into train set and test set where train set contains 70% of the data and test set contains 30% of the data. The model is built using six machine learning approaches and the approach giving maximum accuracy, recall score and precision score will be considered a best fit for this problem.

Anaconda navigator [10] is utilized as it provides many platforms to play with Python language. Jupyter Notebook is utilized here in this project as working with it is comparatively simpler than other IDEs.

IV. Approaches for Implementation

Many machine learning approaches can be applied to this data set to get the outcome. But, in this project, the approaches that provides the best outcomes are displayed. They are.

1) SVM
2) CART
3) RFC
4) Logistic Regression
5) NB
6) KNN

The preprocessing on data set is done by removing the unwanted columns that are present in the data set. The column named 'Time' is dropped or removed here as it's not needed. Every value in this whole data set lied in some range of values. Every column had it's range and the data in it lied in that particular range except the column named 'Amount'. So, I have changed the values of the column 'Amount' to a range of numbers that is smaller than the actual one. The new column of Amount is named as 'New_amount'. Now, the data is ready to work with.

The figure 2 indicates the dataset's correlation matrix. This matrix clarifies that the class of attributes is independent of the amount of the transaction. It is also evident from the matrix that the transaction type is dependent on the attributes added by the PCA.

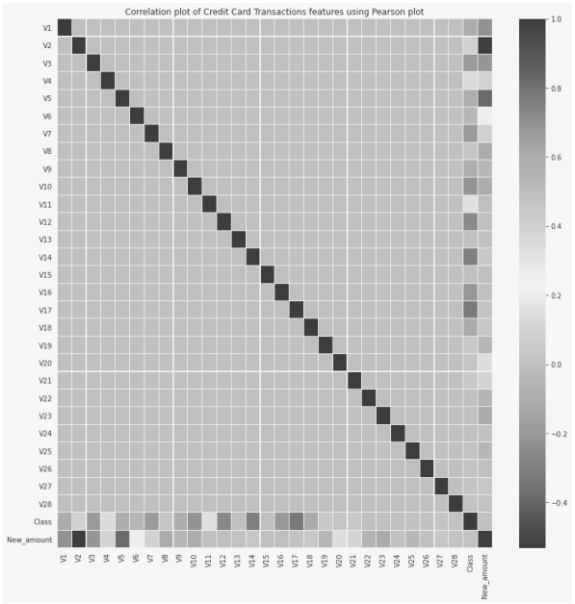

Fig. 2. Correlation Map for the data

A. Logistic Regression (LR):

The Logistic Regression is a model of classification used especially for datasets of binary classifications which utilizes the sigmoid function. As our dataset is a dataset for classification, we have been utilizing this Logistic Regression. In order to classify the fraud in the credit card fraud activity, it primarily classifies the dataset into two binary values that are ultimately 0's and 1's. The data is loaded initially with the assistance of the pandas collection. The dataset is split in the next step into values of x and y and sizes of both the values are written. To carry out the procedure of training and testing, the method named train_test_split() is used. After distributing the data into train and test sets, LogisticRegression() algorithm is implemented. Firstly, train the dataset in this model and then evaluate the remaining dataset with the help of the prediction method for the remaining data. This approach gives the accuracy of 99.908%. And lastly, the confusion matrix is created for this approach utilizing confusion_matrix() methodology. It is shown as below:

```
[[85280    13]
 [   65    85]]
```

B. Decision Tree (CART):

For the classification and regression issues that function for both, the decision tree may be used, but certain formulas can differ [13]. The classification problem utilizes entropy and information gain for the construction of the model of the decision tree. Entropy says how random the data is and how much details we can get from this function is information gained. The data is imported utilizing help of the pandas library. The method named train_test_split() for the procedure of training and testing. Then after, the approach of DecisionTreeClassifier() is utilized. This approach gives the accuracy of 99.932%. The confusion matrix is displayed using confusion_matrix() method. It is shown as below:

```
[[85286     7]
 [   51    99]]
```

C. Naïve Bayes (NB):

Naïve Bayes is the classification dilemma machine learning algorithm, which operates on the Bayes theorem concept . It can be implemented using some independent features as an input and dependent feature as an output in the data set, this very same thing that is behind the Naïve Bayes theorem is used here to measure the dependent feature's likelihood with regard to independent features. The data is loaded with the same methodology of utilizing the pandas library. The dataset is distributed utilizing train_test_split() method and then after the Naïve Bayes approach is implemented. Here, Gaussian Naïve Bayes classifier is utilized. It is implemented by using GaussianNB() method. We train the dataset first and then we implement our approach on the remaining data by utilizing the prediction method for that remaining data. This approach gives the accuracy of 97.803%. Then creation of confusion matrix is carried out at last using confusion_matrix() methodology. It is shown as below:

```
[[83444  1849]
 [   28   122]]
```

D. Support Vector Machine (SVM):

A supervised learning algorithm that filters data into two categories is a support vector machine. It is trained in two forms with a set of data originally classified, constructing the desired model as it is trained originally. The role of a SVM classifier is to decide in which category, the new point of data can be included. This enables SVM a kind of linear classifier that is not binary. This approach is also termed as “Support Vector Network” which is shortly abbreviated as SVN.

In this study, the importing method and method of distributing the data into train and test set is same as above. After doing this, SVC() methodology is implemented. Here, the ‘Radial Basis Function’ kernel is utilized which is shortly abbreviated as ‘rbf’. This approach give us the accuracy of 99.936%. And the confusion matrix can be displayed and calculated using confusion_matrix() methodology. It is shown as below:

```
[[85286     7]
 [   47   103]]
```

E. K-Nearest Neighbors (KNN):

In Machine Learning, KNN is one of the most simple but important algorithm for classification. It belongs to the category of supervised learning and finds intensive application in the identification of patterns, data mining and detection of intrusion.

It is fast supervised machine learning algorithm which is implemented easily and it is utilized to resolve the problems of regression and classification in an efficient way.

After importing the data set and applying train_test_split() method, KNeighborsClassifier() approach is implemented. Here, n_neighbors=10 is taken as it’s an essential hyperparameter, which is utilized at the time period of building our model. This approach gives the accuracy of 99.937%. After that, confusion matrix is calculated and displayed utilizing confusion_matrix() methodology. It is shown as below:

```
[[85284     9]
 [   44   106]]
```

F. Random Forest Classifier (RF):

The random forest classifier picks the characteristics that are independent variables and also chooses the rows by row sampling randomly, and the quantity of decision tree can be evaluated by optimising the hyper parameter. The output for the classification research problem is the maximum occurring outcomes within the random forest from each decision tree classifier. That's one of the frequently used algorithms for machine learning in real - world contexts and models deployed.

After loading the data and doing training and test splits, RandomForestClassifier() methodology is implemented. Here, we n_estimators=10 is taken that specifies the total number of forests utilized in our Random Forest model. This approach gives the accuracy of 99.947%. The confusion matrix is calculated and displayed utilizing confusion_matrix() methodology. It is shown as below:

```
[[85284     9]
 [   36   114]]
```

G. Comparison of Ml Approaches:

Precision Score, Recall and Accuracy are the factors that are taken in consideration to choose the best model among the six implemented algorithms. Precision score is the model accuracy score reflects the ability of the model to predict exactly the positive out of all of the positive predictions it generated. The recall evaluates the effectiveness of the proposed model to identify positive samples. The larger the recall, the more positive specimens that have been detected. Recall is also referred as "Sensitivity". The accuracy score of the model reflects the capacity of the model to predict exactly both positive and negative out of all outcomes.
Delineating mathematically;

Precision = True Positives / (True Positives + False Positives)
Recall = True Positives / (True Positives + False Negatives)
Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

Here, six approaches of Machine Learning are utilized for building the best model for detecting the credit card frauds. The comparison of different approaches along with their accuracies, precision scores and recall scores is shown in figure 3.

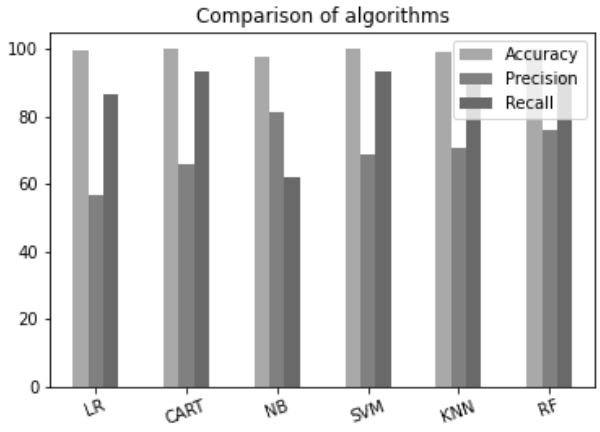

Fig. 3. Comparison of algorithms along with acuracy, precision_score and recall_score

Representing the comparison of the algorithms in the form of table, along with the precision score, recall score and accuracy with their respective algorithms. It is mentioned in table 2.

Table 2.
Table Name

Name of Algorithms	Accuracy	Precision	Recall
Logistic Regression (LR)	99.908%	56.666%	86.734%
Decision Trees (CART)	99.932%	66%	93.396%
Naïve Bayes (NB)	97.803%	81.333%	61.897%
Support Vector Machine (SVM)	99.936%	68.666%	93.636%
K-Nearest Neighbor (KNN)	99.379%	70.666%	92.173%
Random Forest (RF)	99.947%	76%	92.682%

V. CONCLUSION

As utilization of credit cards have become more frequent in every domain of the everyday life, credit card frauds are occurring more frequently. To prevent attacks of the financial transaction systems in an autonomous and reasonable manner, one of primary duties for financial firms is to develop a precise and accurate credit card fraud detection mechanism.

A variety of techniques, processes and models are built and implemented to fight the credit card frauds and the researchers have a lot of interest in building accurate credit card fraud detection system.

By observing the comparison of different machine learning approaches, it is clear from the models we built that Random Forest gave us the best accuracy of 99.947%, and it worked well with our collection. It is efficient in every term that is in accuracy score, precision score and recall score.

Other data mining techniques, including various versions of Artificial Neural Networks (ANN), can be used in future research to construct new classification techniques on the very same dataset collected, and the performance of the new models should be matched with the performance of the models presented in this paper.

REFERENCES

[1] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2018, pp. 129-134, doi: 10.1109/SIEDS.2018.8374722.

[2] W. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum," 2009 International Joint Conference on Artificial Intelligence, Hainan Island, 2009, pp. 353-356, doi: 10.1109/JCAI.2009.146.

[3] Z. Kazemi and H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, 2017, pp. 0630-0633, doi: 10.1109/KBEI.2017.8324876.

[4] Rinky D. Patel and Dheeraj Kumar Singh, "Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm", published by

International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[5] K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp. 1-5, doi: 10.1109/I2C2.2017.8321781.

[6] .[Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to Fraud Detection"; Department of Computer Science- Columbia University; 2000, https://doi.org/10.7916/D86Q28GG

[7] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random forest for credit card fraud detection," *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, Zhuhai, 2018, pp. 1-6, doi: 10.1109/ICNSC.2018.8361343.

[8] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," in *IEEE Access*, vol. 6, pp. 14277-14284, 2018, doi: 10.1109/ACCESS.2018.2806420.

[9] Ghosh and Reilly, "Credit card fraud detection with a neural-network," *1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences*, Wailea, HI, USA, 1994, pp. 621-630, doi: 10.1109/HICSS.1994.323314.

[10] Varun Kumar K S , Vijaya Kumar V G , Vijayshankar A , Pratibha K, 2020, Credit Card Fraud Detection using Machine Learning Algorithms, *International Journal of Engineering Research & Technology (IJERT)*, Volume 09, Issue 07 (July 2020)

[11] Prof. Vijayalaxmi Kadroli, P. R. S. (2015). Survey on Credit Card Fraud Detection Techniques. *International Journal of Engineering and Computer Science*, *4*(11). Retrieved from http://103.53.42.157/index.php/ijecs/article/view/2895

[12] Dahl, J.: Card Fraud. In: Credit Union Magazine (2006).

[13] C. Wirawan, "Teknik Data Mining Menggunakan Algoritma Decision Tree C4.5 untuk Memprediksi Tingkat Kelulusan Tepat Waktu," Appl. Inf. Syst. Manag., vol. 3, no. 1, pp. 47–52, 2020, doi: 10.15408/aism.v3i1.13033