Integration of Artificial Intelligence in the Advancement of
Science and Engineering
July 2024

CREDIT CARD FRAUD DETECTION USING ML WITH ANALYSIS OF ALGORITHMS

1st ShivKumar
Computer science & Engineering
RTC Institute of Technology
Ranchi, India
Shivkumar004@gmail.com

2nd Payal Kumari
Computer science & Engineering
RTC Institute of Technology
Ranchi, India
shahp0499@gmail.com

3rd Mamta kumari
Computer science & Engineering
RTC Institute of Technology
Ranchi, India
kumarigungunbasia@gmail.com

***Abstract*—**Lately, there has been a significant surge in the use of credit cards, resulting in an increased susceptibility to credit card fraud. Financial institutions are diligently working to combat fraudulent transactions and safeguard their systems. It is evident that fraud takes on various forms and continually evolves, making it essential to implement a robust machine-learning algorithm capable of addressing these dynamic challenges. This study compares three distinct machine-learning algorithms based on their accuracy and various performance metrics. The findings indicate that the Random Forest Algorithm demonstrates the highest accuracy in detecting fraudulent activities.

Key terms - machine learning algorithm, credit card fraud, Decision Tree, Logistic Regression, Random Forest, Score, Precision, MCC, accuracy, recall

I. INTRODUCTION

Instances of credit card fraud occur when cards are lost or stolen, when mail is intercepted by criminals, or when employees of a business unlawfully obtain customer information. Methods of Credit Card Fraud:

1. Traditional Methods:

Fraud Using Paper Documents -Involves criminals using stolen or fake documents such as utility bills and bank statements to acquire valuable Personally Identifiable Information (PII) and open an account in someone else's name.

Fraudulent Applications -

ID Theft: When an individual impersonates someone else.

Financial Fraud: When someone provides false financial information to obtain credit.

2. Modern Methods:

Skimming for Fraud -A type of crime in which dishonest employees make unauthorized copies of credit or debit cards using a 'skimmer'. A skimmer is a device that captures credit card numbers and other confidential account information stored on the magnetic stripe or smart chip, and then transfers this data to a different card.

Credit card approval is a critical process within the banking sector. In the past, banks relied on manually assessing creditworthiness, which was time-consuming and susceptible to errors.

Machine Learning algorithms have the capability to analyze large amounts of data and identify patterns, making them extremely valuable in the credit card approval process.

By training ML models on historical data containing applicant information, financial behavior, and credit history, banks can make more precise and efficient predictions about creditworthiness..

II. LITERATURE REVIEW

K. Randhawa and other researchers did an empirical evaluation on a number of standard models which include NB, SVM, and DL [1].

D. Tanouz and other researchers evaluated algorithms such as decision tree, Random forest, logistic regression and naive Bayes classification [2] .

V. Jain and other researchers implemented a web based application using machine learning algorithms such as Logistic Regression, Random Forest, and AdaBoost [3].

M. R. Dileep and other researchers proposed a model where two algorithms are used viz Fraud Detection in credit card using Decision Tree and Fraud Detection using Random Forest [4] .

Donglin Li proposed combination of XGBoost model and Lasso-Logistic model. With this calculation, speed become faster and important variables can also be selected [5] .

S. K. Saddam Hussain and other researchers used Random Forest, SVM, and DL algorithms in the proposed method to identify the credit card frauds [6].

S. Khatri and other researchers used an imbalanced dataset and evaluated different supervised machine learning models to find fraudulent transaction [7].

C. Wang and other researchers proposed a system based on whale algorithm optimized BP neural Network. The aim is to solve the problems of slow convergence rate, easy to fall into local optimum, network defects and poor system [8] .

C. H. Sumanth and other researchers analysis three machine learning algorithm, viz. Navie Bayes, SVM and DNN to find fraudulent credit card transactions [9].

Kundu and other researchers proposed an algorithm named as BLAH for credit card fraud detection [10] .

D. Li researchers combined the Lasso-Logistic model and the XGBoost model. By this calculation speed is increased In [11].

III. PROBLEM STATEMENT

The use of credit cards in India has increased significantly from 2019 to 2023.Credit card defaults reached Rs 4,072 crore in FY23, representing a growth of Rs 950 crore from FY22.

The majority of credit card loans issued by banks are directed towards higher-risk borrowers. As credit card usage rises, it is imperative to have a secure system in place to prevent fraudulent transactions. Various machine learning algorithms can be utilized to detect fraudulent transactions effectively. Assessing the accuracy of algorithms based on different factors and attributes is crucial.

IV. OBJECTIVE

- Evaluating following algorithms with large dataset.
    - Decision Tree (DT),
    - Logistic Regression (LR)
    - and Random Forest(RF).
- Study and use parameters such as Accuracy, Precision, MCC and Recall score for evaluation..

V. PROPOSED SYSTEM

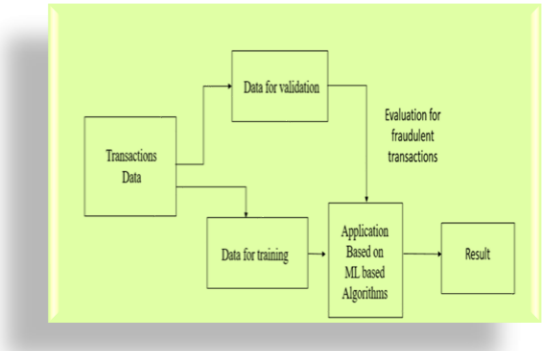


Fig 1 Workflow of Credit Card Fraud Detection System

We are using data set received from European cardholder[12],from Kaggle website. Data set is divided in 2 set, viz. Data for validation and Data for training. Training data (80% of total data) is fed to the application for training purpose. Any one ML algorithm is enabled at one time.

Testing data is passed to application, application will verify and conclude if any transaction is in fraudulent category or in normal category.

The purpose of logistic regression is to predict the outcome of a categorical dependent variable, which means the result needs to be a categorical or discrete value. This can include values like Yes or No, 0 or 1, true or False, etc. instead of providing the exact binary values, logistic regression yields probabilistic values that fall between 0 and 1. Unlike fitting a regression line, logistic regression involves fitting an "S" shaped logistic function that predicts two maximum values (0 or 1).

Decision Tree is a type of supervised learning method and is commonly used for tackling Classification problems. It is characterized by its tree-like structure. The features of a dataset are represented by the internal nodes, while the decision rules are depicted by the branches, and the outcome is represented by each leaf node. Here's how it works: Start with the root node, denoted as S, which contains the entire dataset. Next, identify the best attribute in the dataset using Attribute Selection Measure (ASM). Then, split the root node into subsets containing potential values for the best attributes. After that, create the decision tree node that comprises the best attribute. Finally, recursively generate new decision trees using the subsets of the dataset created in the previous step. Repeat this process until you reach a stage where further classification is not possible, at which point the final node is labeled as a leaf node.

Random Forest utilizes a mix of decision trees to enhance the outcomes. Each decision tree assesses different conditions. Training of decision trees is done on random datasets. The sub-trees will produce the probabilities for a transaction being 'fraud' or 'non-fraud.' The model will then classify the transaction as either 'fraud' or 'genuine' based on the combined result. Here's how it works:

Step 1: Random samples are chosen from the given data or training set.

Step 2: A decision tree is built for each training data by this algorithm.

Step 3: Voting occurs through averaging the decision tree predictions.

Step 4: Finally, the final prediction result is the one with the highest number of votes.

VI. IMPLEMENTATION

Table 1 Prerequisites tools and software

Sr.No.	Tool/Software	Version
1	Python	3.x
2	Matplotlib	3.3.4
3	Jupiter	4.4.0

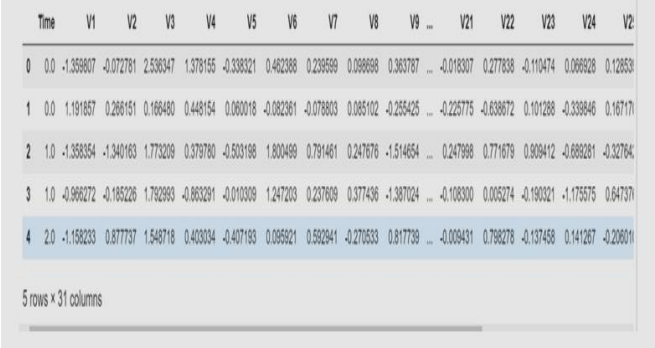
Table 2 H/W Requirement

Sr. No.	Tool/Software
1	RAM - 16GB
2	Processor - Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz
3	Windows 11 Pro

Numpy is a python library. It is used while working with arrays. It is useful for quick array based operations.

Pandas is used for working with relational data. It provide support in analyzing and manipulating data.[16]

Matplotlib is a plotting library. its numerical mathematics extension Numpy. It provides an object-riented API for embedding plots into applications.[17]



	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V2
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.12853
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.16717
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.32764
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.64737
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.20601

5 rows × 31 columns

Fig. 2. Fraud transaction data

VII. RESULT AND DISCUSSION

As shown in Fig.3, we have total 2,84807 transaction. In that, 492 are fraudulent transaction.

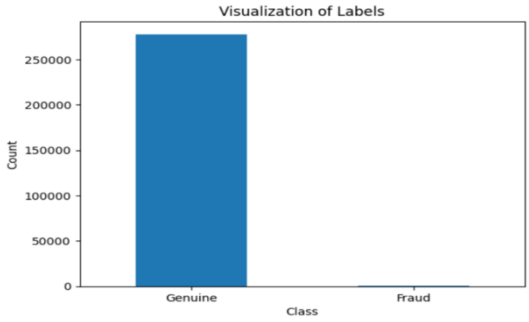


Fig.3. Fraud and Genuine Transactions

With application deployed with Random forest algorithm, we found Precision Score of 0.94167, F1 Score of 0.88281, Accuracy Score of 0.99965 and Recall Score of 0.83088.

Fig. 4 shows matrix plot of Random forest algorithm.

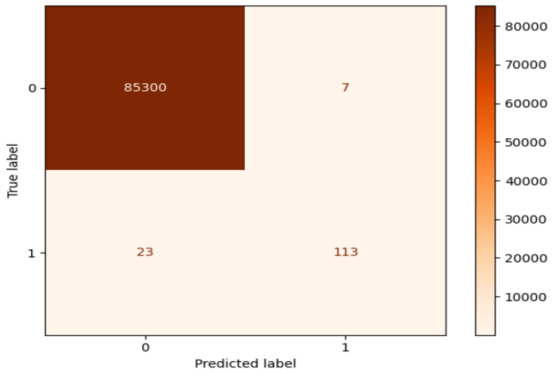


Fig.4. Random Forest algorithm Matrix

With application deployed with Logistic Regression (LR) algorithm, we found Precision Score of 0.87629, F1 Score of 0.72961, Accuracy Score of 0.99926 and Recall Score of 0.62500.

Fig. 5 shows matrix plot of Logistic Regression (LR) algorithmWith application deployed with Logistic Regression (LR) algorithm, we found Precision Score of 0.87629, F1 Score of 0.72961, Accuracy Score of 0.99926 and Recall Score of 0.62500.

Fig. 5 shows matrix plot of Logistic Regression (LR) algorithm

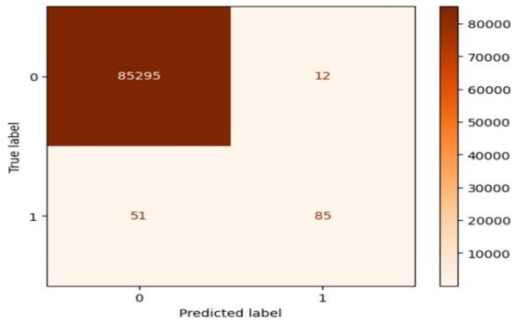


Fig. 5 Logistic Regression (LR) algorithm Matrix

With application deployed with Decision Tree algorithm, we found Precision Score of 0.76712, F1 Score of 0.79433, Accuracy Score of 0.99932 and Recall Score of 0.82353.

Fig. 6 shows matrix plot of Decision Tree algorithm.

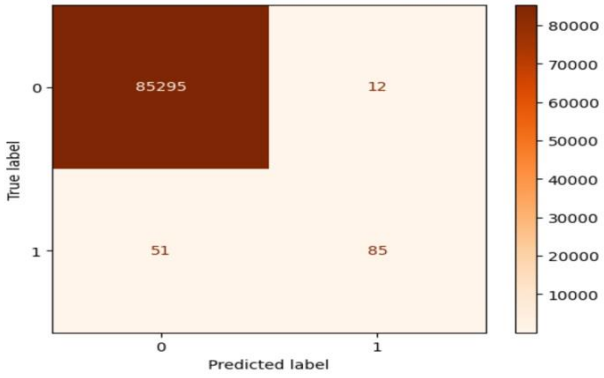


Fig. 6 Decision Tree algorithm Matrix

Below Table 2 shows the comparison of algorithms based on different parameters.

Table 3 Algorithm comparison on basis of four scores, viz. Precision, MCC, Accuracy and Recall

Table 3 Algorithm comparison on basis of four scores

		Algorithms used		
		Random Forest	Logistic Regression (LR)	Decision Tree
Score	Precision Score	0.94167	0.87629	0.76712
	MCC Score	0.88281	0.72961	0.79433
	Accuracy Score	0.99965	0.99926	0.99932
	Recall Score	0.83088	0.62500	0.82353

Precision can be considered as an indicator of excellence, while recall can be viewed as an indicator of capacity. Increased precision indicates that an algorithm produces more pertinent outcomes than irrelevant ones, and high recall indicates that an algorithm produces the majority of pertinent outcomes (regardless of whether irrelevant ones are also produced).

MCC serves as an optimal single-value classification metric that effectively summarizes the confusion matrix or an error matrix. A confusion matrix comprises four elements: True positives (TP), True negatives (TN), False positives (FP).

CONCLUSION

Based on our findings, the Random Forest algorithm outperformed all other algorithms. It achieved a Precision of 94167, MCC Score of 0.88281, Accuracy Score of 0.99965, and Recall Score of 0.83088. Among all 4 performance metrics, the Random Forest algorithm demonstrated superior performance. Its Accuracy, Precision, MCC, and Recall scores were all better.

FUTURE SCOPE

Our main goal is to calculate the average of the top two solutions and enhance the reliability of the application for future opportunities. A web-based application can play a significant role in this area. Our efforts will be directed towards developing a more engaging web application.

ACKNOWLEDGMENT

Shiv kumar received the m. Tech. Degree in computer science and engineering from mewar university chittorgargh in 2012. During 2007-2013, he stayed in canon india private limited center of excellence center and india software center noida and gurgaon of india. He knows with mewar university, chittorgargh, india.presently working in rtcit, ranchi

REFERENCES

[1] 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.

[2] D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar and C. V. N. M. Praneeth, "Credit Card Fraud Detection Using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 967-972, doi: 10.1109/ICICCS51141.2021.9432308.

[3] V. Jain, H. Kavitha and S. Mohana Kumar, "Credit Card Fraud Detection Web Application using Streamlit and Machine Learning," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), 2022, pp. 1-5, doi: 10.1109/ICDSIS55133.2022.9915901.

[4] M. R. Dileep, A. V. Navaneeth and M. Abhishek, "A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1025-1028, doi: 10.1109/ICICV50876.2021.9388431.

[5] D. Li, "Credit card fraud identification based on unbalanced data set based on fusion model," 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT), 2019, pp. 235-239, doi: 10.1109/ICCASIT48058.2019.8973167.

[6] S. K. Saddam Hussain, E. Sai Charan Reddy, K. G. Akshay and T. Akanksha, "Fraud Detection in Credit Card Transactions Using SVM and Random Forest Algorithms," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 1013-1017, doi: 10.1109/I-SMAC52330.2021.9640631

[7] S. Khatri, A. Arora and A. P. Agrawal, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison," 2020 10th International Conference on Cloud

Computing, Data Science & Engineering (Confluence), 2020, pp. 680-683, doi: 10.1109/Confluence47617.2020.9057851.

[8] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai and S. Pan, "Credit Card Fraud Detection Based on Whale Algorithm Optimized BP Neural Network," 2018 13th International Conference on Computer Science & Education (ICCSE), 2018, pp. 1-4, doi: 10.1109/ICCSE.2018.8468855.

[9] C. H. Sumanth, P. P. Kalyan, B. Ravi and S. Balasubramani., "Analysis of Credit Card Fraud Detection using Machine Learning Techniques," 2022 7th International Conference on Communication and Electronics Systems (ICCES), 2022, pp. 1140-1144, doi: 10.1109/ICCES54183.2022.9835751.

[10] A. Kundu, S. Panigrahi, S. Sural and A. K. Majumdar, "BLAST-SSAHA Hybridization for Credit Card Fraud Detection," in IEEE Transactions on Dependable and Secure Computing, vol. 6, no. 4, pp. 309-315, Oct.-Dec. 2009, doi: 10.1109/TDSC.2009.11.

[11] D. Li, "Credit card fraud identification based on unbalanced data set based on fusion model," 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT), 2019, pp. 235-239, doi: 10.1109/ICCASIT48058.2019.8973167.

[12]What is a credit card? (2003, November 19). Investopedia. https://www.investopedia.com/terms/c/creditcard.asp

[13]Credit card fraud. (2022, September 22). Wikipedia, the free encyclopedia. Retrieved May 11, 2023, from https://en.wikipedia.org/wiki/Credit_card_fraud

[14] Fraud detection algorithms | Fraud detection using machine learning. (2023, March 2). Intellipaat Blog. https://intellipaat.com/blog/fraud-detection-machine-learning-algorithms/

[15] Credit Card Fraud Detection Accessed: Jan. 20, 2022.

[Online].Available:https://www.kaggle.com/mlgulb/creditcardfrad.

[16] Wikipedia. (n.d.). pandas (software). https://en.wikipedia.org/wiki/Main_Page. Retrieved February 28, 2023, from https://en.wikipedia.org/wiki/Pandas_(software)

[17]Wikipedia.(n.d.). Matplotlib. https://en.wikipedia.org/wiki/Main_Page. Retrieved February 28, 2023, from https://en.wikipedia.org/wiki/Matplotlib