iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47127

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Credit Card Fraud Detection Using Machine Learning Algorithms

Muskan Banu[1], Prof. Kavitha G[2]
[1]*M.Tech (CSE), Department of Studies in Computer Science andEngineering, University B.D.T College of Engineering, Davanagere -577004, Karnataka, India*
[2]*M.tech, Department of Studies in Computer Science andEngineering, University B.D.T College of Engineering, Davanagere - 577004, Karnataka, India*
(A Constituent College of Visvesvaraya TechnologicalUniversity, Belagavi)

Abstract: ***Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card-issuing authorities are unawareof the fact that the card is being used. Credit card frauds are easy and friendly targets. E-commerce and many other online sites have increased the online payment modes, increasing the risk for online frauds.***
In the era of digitalization, the need to identify credit card frauds is necessary. Fraud detection involves monitoring and analyzing the behaviour of various users to estimate, detect or avoid undesirable behaviour. To identify credit card fraud detection effectively, we need to understand the various technologies, algorithms and types involved in detecting credit card frauds.
The algorithm can differentiate transactions which are fraudulent or not. To find fraud, we need to pass dataset and knowledge of the fraudulenttransaction. Algorithms analyze the dataset and classify all transactions. Fraud detection involves monitoring the activitiesof populations of users to estimate, perceive or avoidobjectionable behaviour, which consist of fraud, intrusion, and defaulting. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. We have taken an imbalanced dataset of transactions to detect the frauds

I. INTRODUCTION

Credit card generally refers to a card that is assigned to the customer (cardholder), usually allowing them to purchase goods and services within credit limit or withdraw cash in advance. Credit card provides the cardholder an advantage of the time, i.e., it provides time for the customers to repay later in a prescribed time, by carrying it to the next billing cycle. Credit card frauds are easy targets. Without any risks, a significant amount can be withdrawn without the owner's knowledge, in a short period. Fraudsters always try to make every fraudulent transaction legitimate, which makes fraud detection very challenging and difficult task to detect.

Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary preventive measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, Credit Card Fraud can be defined as a case where a person uses someones card and issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting.

This isa very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance.

The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time. These are not the only challenges in the implementation of a real- world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyse allthe authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

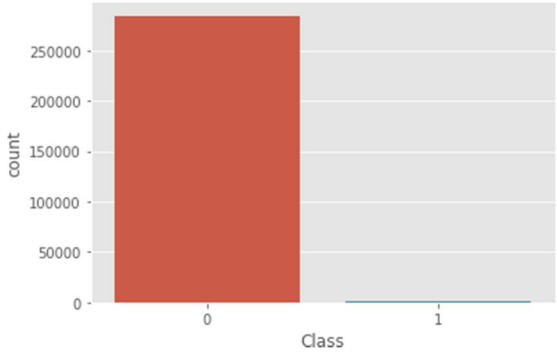
II. METHODOLOGY

Architectural design for detecting Credit Card Frauds

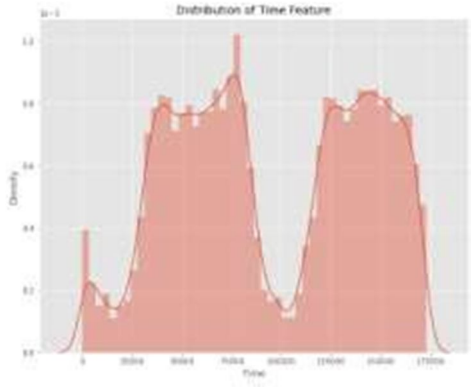
General Steps we are following in the Project are as follows;

- *Step 1:* Read the dataset.
- *Step 2:* Random Sampling is done on the data set to makeit balanced.
- *Step 3:* Divide the dataset into two parts i.e., Train datasetand Test dataset.
- *Step 4:* Feature selection are applied for the proposedmodels.
- *Step 5:* Accuracy and performance metrics has beencalculated to know the efficiency for different algorithms.
- *Step 6:* Then retrieve the best algorithm based onefficiency for the given dataset.

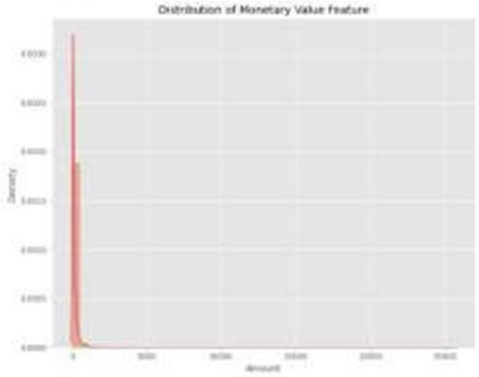
1) *DataSet:* We obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data. The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one. We plot different graphs to check for inconsistencies in the dataset and to visually comprehend it:



This graph shows that the number of fraudulenttransactions is much lower than the legitimate ones.

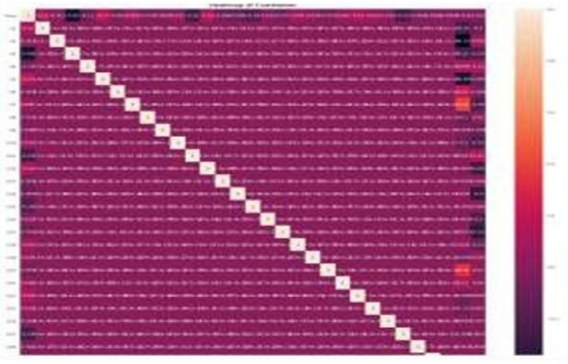


This graph shows the times at which transactions were done within two days. It can be seen that the least number of transactions were made during night time and highest during the days.



This graph represents the amount that was transacted. A majority of transactions are relatively small and only a handful of them come close to the maximum transacted amount.

After checking this dataset, we plot a histogram for every column. This is done to get a graphical representation of the dataset which can be used to verify that there are no missing value imputation and the machine learning algorithms can process the dataset smoothlyany values in the dataset. This is done to ensure that we don't.
After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable. This heatmap is shown below:



The dataset is now formatted and processed. The time and amount column are standardized and the Class column is removed to ensure fairness of evaluation. The data is processed by a set of algorithms from modules. The following module diagram explains how these algorithms work together:This data is fit into a model and the following outlier detection modules are applied on it:

Some of the currently used approaches to detection of such fraud in this paper are:

- Logistic Regression.
- K-Nearest Neighbor Classifier.
- SVM
- Naive Bayes Algorithm.
- Decision Tree Classifier.
- Random Forest Algorithm.

These algorithms are a part of sklearn. The ensemble module in the sklearn package includes ensemble-based methods and functions for the classification, regression and outlier detection. This free and open-source Python library is built using NumPy, SciPy and matplotlib modules which provides a lot of simple and efficient tools which can be used for data analysis and machine learning. It features various classification, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries. Python to demonstrate the approach that this paper suggests. This program can also be executed on the cloud using Google Collab platform which supports all python notebook files. Detailed explanations about the modules with pseudocodes for their algorithms and output graphs are given as follows

a) *Logistic Regression:* It is one of the classification algorithm, used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds are used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function.

b) *K-Nearest Neighbor Classifier:* This is a supervised learning technique that achieves consistently high performance in comparison to other fraud detection techniques of supervised statistical pattern recognition . Three factors majorly affect its performance: distance to identify the least distant neighbors, some rule to deduce a categorization from k-nearest neighbor & the count of neighbors to label the new sample. This algorithm classifies any transactions that occurred by computing the least distant point to this particular transaction and if this least distant neighbor is classified as fraudulent thenthe new transaction is also labeled as a fraudulent one. Euclidean distance is a good choice to calculate the distances in this scenario. This technique is fast and results in fault alerts. Its performance can be improved by distance metric optimization

c) *SVM:* Support vector machines or SVMs are linear classifiers as stated in that work in high dimensionality because in high-dimensions, a non-linear task in input becomes linear andhence this makes SVMs highly useful for detecting frauds.Due to its two most important features that is a kernel functionto represent classification function in the dot product of input data point projection, and the fact that it tries finding a hyperplane to maximize separation between classes while minimizing overfitting of training data, it provides a very high generalization capability .

d) *Naïve Bayes Algorithm:* Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. It isa classification technique based on Bayes' theorem with anassumption of independence between predictors. In simple terms, aNaive Bayes classifier assumes that the presence of a particularfeature in a class is unrelated to the presence of any other feature.

e) *Decision Tree Algorithm:* Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that ismostly used in classification problems. It works for bothcategorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significantsplitter / differentiator in input variables.

f) *Random Forest:* Random forest is a tree based algorithm which involves building several trees and combining with the output to improve generalization ability of the model. This method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

III. IMPLEMENTATION

This idea is difficult to implement in real life because it requires the cooperation from banks, which aren't willing to share information due to their market competition, and also due to legal reasons and protection of data of their users. Therefore, we looked up some reference papers which followed similar approaches and gathered results.

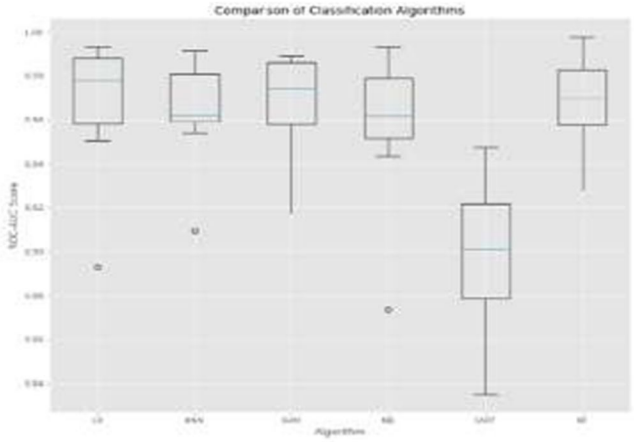
As stated in one of these reference papers: supplied by a German bank in 2006. For banking confidentiality reasons, only a summary of the results obtained is presented below. After applying this technique, the level 1 list encompasses a few cases but with a high probability of being fraudsters. All individuals mentioned in this list had their cards closed to avoid any risk due to their high-risk profile. The condition is more complex for the other list. The level 2 list is still restricted adequately to be checked on a case by case basis. Credit and collection officers considered that half of the cases in this list could be considered as suspicious fraudulent behaviour. For the last list and the largest, the work is equitably heavy. Less than a third of them are suspicious. In order to maximize the time efficiency and the overhead charges, a possibility is to include a new element in the query; this element can be the five first digits of the phone numbers, the email address, and the password, for instance, those new queries can be applied to the level 2 list and level 3 list.".

IV. RESULTS

The code prints out the number of false positives it detected andcompares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction ofdata we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end and both the results are printed. These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and1 means it was determined as a fraud transaction. This resultmatched against the class values to check for false positives Results when 10% of the dataset is used:

LR
[[82 2]
[7 33]]
Precision Score: 0.9428571428571428
Recall Score: 0.825

KNN
[[83 1]
[8 32]]
Precision Score: 0.9696969696969697
Recall Score: 0.8



SVM
[[84 0]
[7 33]]
Precision Score: 0.9232066520120
Recall Score: 0.825

NB
[[81 3]
[9 31]]
Precision Score: 0.9117647058823529
Recall Score: 0.775

CART
[[79 5]
[6 34]]
Precision Score: 0.8717948717948718
Recall Score: 0.85

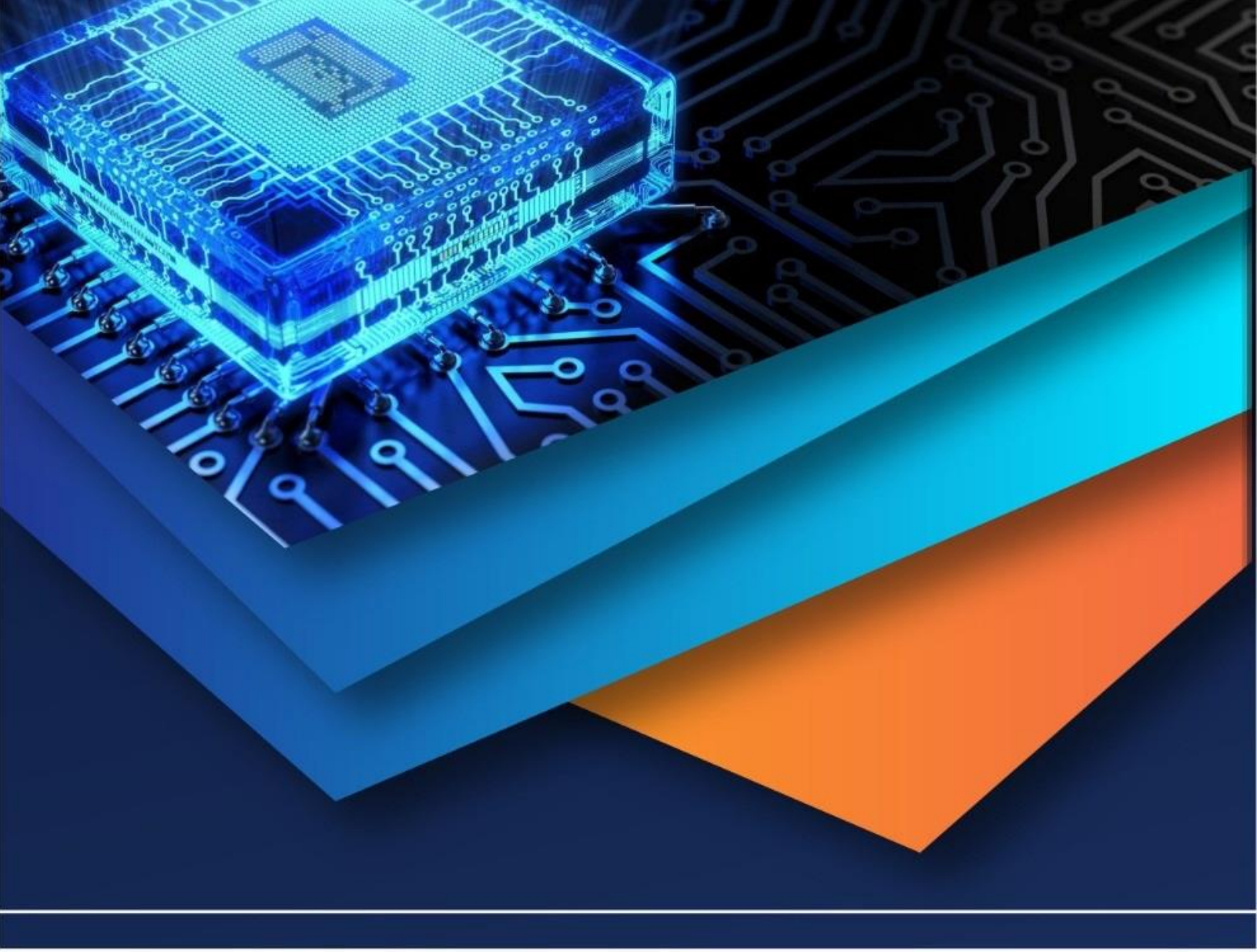
RF
[[84 0]
[7 33]]
Precision Score: 0.98203354221520
Recall Score: 0.825

V. CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results. While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of days' transaction records, its only a fraction of data that can be made available if this project were to be used on a commercial scale. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put intoit.

REFERENCES

[1] Neda Soltani Halvaiee, Msohammad Kazem Akbari "A novel model for credit card fraud detection using Artificial Immune Systems", Elsevier, pp 40-39, 2014.
[2] Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun k. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model", IEEE Transaction on dependable and secure computing, VOL.5, NO.1, January-March 2008.
[3] Adrian Banarescu "Detecting and Preventing Fraud with Data Analytics", Elsevier, pp 1827-1836, 2015.
[4] Dr. Saleh Al-Furiah, Lamia AL-Braheem "Comprehensive studyon methods of fraud prevention in credit card e-payment system", ACM 978-1-60558- 660-1/09/0012, December2009.
[5] Ms.Pratiksha L.Meshram, Prof. Tarun Yenganti "Credit and ATM Card Fraud Prevention Using Multiple Cryptographic Algorithms", ISSN: 2277 128X, Volume 3, August 2013.
[6] Aman Srivastava, Mugdha Yadav, Sandipani Basu, Shubham Salunkhe, Muzaffar Shabad "Credit Card Fraud Detection at Merchant Side using Neural Networks", 978-9-3805-4421-2, IEEE 2016.
[7] Raghavendra Patidar, Lokesh Sharma "Credit Card Fraud Detection using Neural Network", ISSN: 2231-2307, Vol. 1, June 2011.
[8] Gabriel Preti Santiago, Adriano C.M. Pereira, Roberto Hirata "Amodelling approach for credit card fraud detection in electronicpayment services", ACM 978-1-4503-3196-8/15/04, April 2015.
[9] Ekrem Duman, M.Hamdi Ozcelik "Detecting credit card fraud detection by Genetic Algorithm and scatter search", Elsevier, pp- 13057-13063, 2011.
[10] Ishu Trivedi, Monika, Mrigya Mridushi "Credit card frauddetection", ISSN: 2278-1021, vol. 5, Januray 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY