

Using Machine Learning Techniques to Detect Credit Card malicious

P. Varaprasada Rao
Associate Professor
Dept. of Computer Science & Engineering
Gokaraju Rangaraju Institute of Engineering & Technology
Hyderabad, India

ABSTRACT

It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications. Credit Card Fraud Detection is a typical sample of classification. In this process, we have focused on analysing and preprocessing data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm on the PCA transformed Credit Card Transaction data.

INTRODUCTION

Credit card frauds

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account.

In the Financial Year 2020, more than 50,000 cases 52,006 to be precise of fraudulent usage of debit cards, credit cards and internet banking were reported in India, the Minister of State for Electronics and IT, Sanjay Dhotre revealed in Parliament today. The amount involved in these fraudulent transactions is Rs 228.44 crore. In Financial Year 2019, more such cases 52,304 were reported, however, the amount relating to fraudulent usage was less than Financial Year 2020 at Rs 149.42 crore. The data for FY20 is till December 2019.

Also, between Financial Year 2017 and Financial Year 20, a total of 140,471 cases of fraudulent usage of debit cards, credit cards and internet banking were reported, with the amount relating to such fraudulent usage being Rs 589.14 crore. The number and amount related to such cases can potentially be higher, since fraud cases below Rs. 100,000 were not required to be reported to RBI prior to April 1, 2017.

Necessary prevention 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 someone else's credit card for personal reasons while the

owner and the card 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 is a 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.

Existing System

The existing system for Credit card Fraud Detection is very restricted and not so beneficial to Detecting or alerting the Fraud in Credit card transactions.

Till now the existing system used these techniques to detect fraud credit cards.

- ☒ Approaching Experienced Fraud Analysts
- ☒ Low level accuracy rate in detecting frauds
- ☒ The Traditional detection method mainly depends on the database system and the education of customers, which usually are delayed, inaccurate and not in-time.
- ☒ Valid only for small amount of data

Proposed System

To detect credit card fraud using Machine Learning Algorithms and find the best algorithm between Local Outlier Function Algorithm, Isolation Forest Algorithm. Although there existed several methods to detect the credit card fraud we use machine learning because it gives better accurate results quickly.

2. LITERATURE SURVEY

Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule or policy with an aim to attain unauthorized financial benefit.

Numerous literatures pertaining to anomaly or fraud detection in this domain have been published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates have revealed that techniques employed in this domain include data mining applications, automated fraud detection, adversarial detection. In another paper, Suman, Research Scholar, GJUS&T at Hisar HCE presented techniques like Supervised and Unsupervised Learning for credit card fraud detection. Even though these methods and algorithms fetched an unexpected success in some areas, they failed to provide a permanent and consistent solution to fraud detection.

A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transactions in an emulation experiment of credit card transaction data set of one certain commercial bank. Outlier mining is a field of data mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e. basically used in monetary and internet fields. It deals with

detecting objects that are detached from the main system i.e. the transactions that aren't genuine. They have taken attributes of the customer's behaviour and based on the value of those attributes they've calculated that distance between the observed value of that attribute and its predetermined value. Unconventional techniques such as hybrid data mining/complex network classification algorithm is able to perceive illegal instances in an actual card transaction data set, based on network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group have proved efficient typically on medium sized online transactions.

SYSTEM DESIGN

Architecture Diagram

An architecture diagram is a graphical representation of a set of concepts that are part of an architecture, including their principles, elements and components. architecture diagrams are the primary “artifacts” produced by semi-technical “high-level” “software architects”. Below is Detailed Architectural Diagram of Credit Card Fraud Detection using Machine Learning. This Architectural Diagram depicts all components involved in Credit card transactions and the flow.

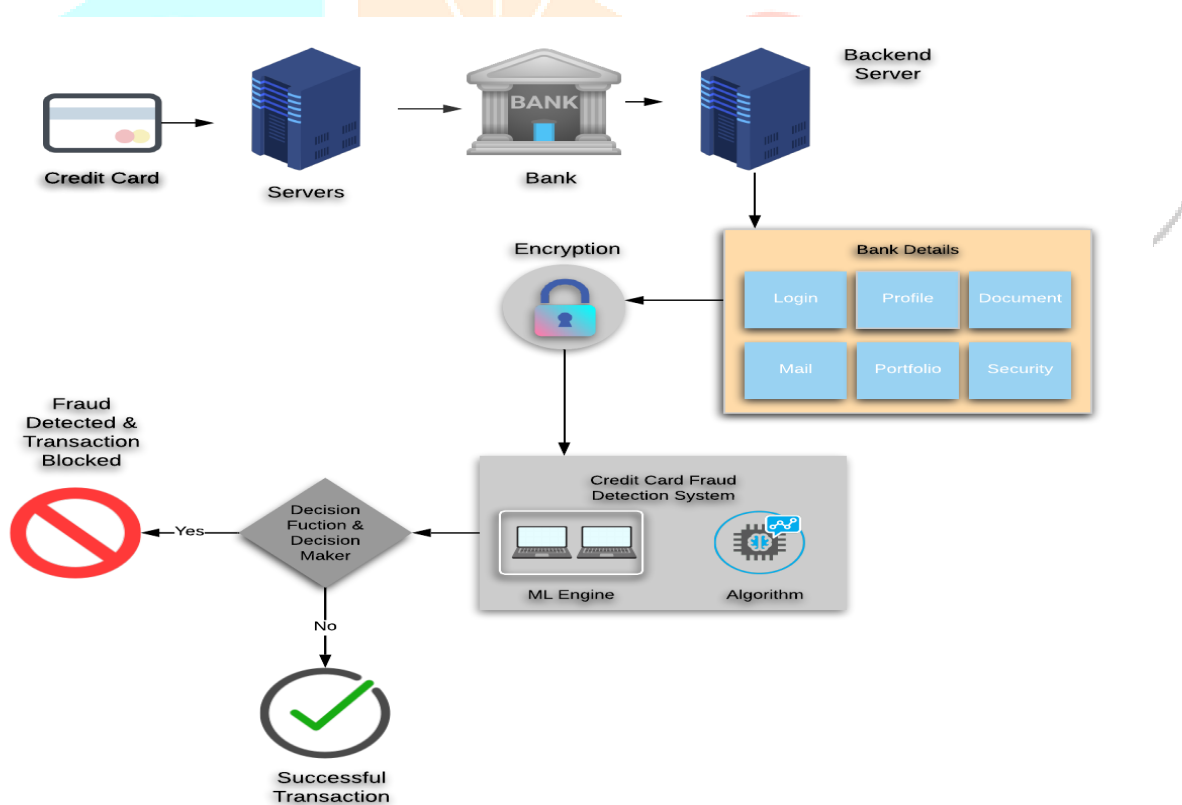


Figure 5.1.1 Architectural Diagram

6. METHODOLOGY

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers.

First of all, 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.

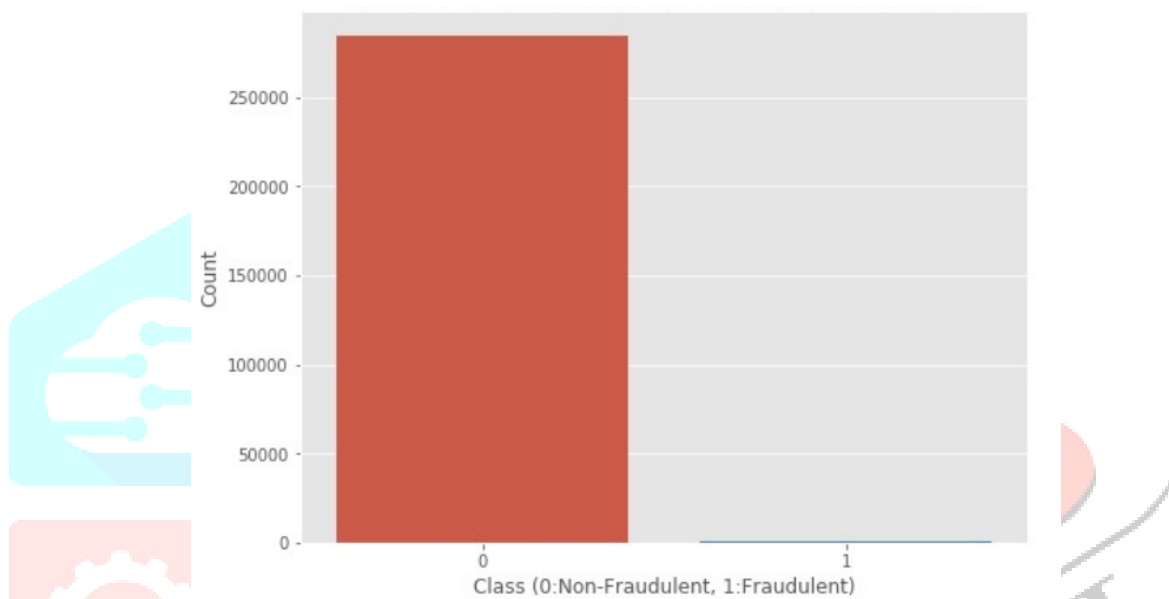


Figure 6.2 Count of Fraudulent vs Non-Fraudulent Transactions

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



Figure 6.3 Distribution of Time Feature

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.

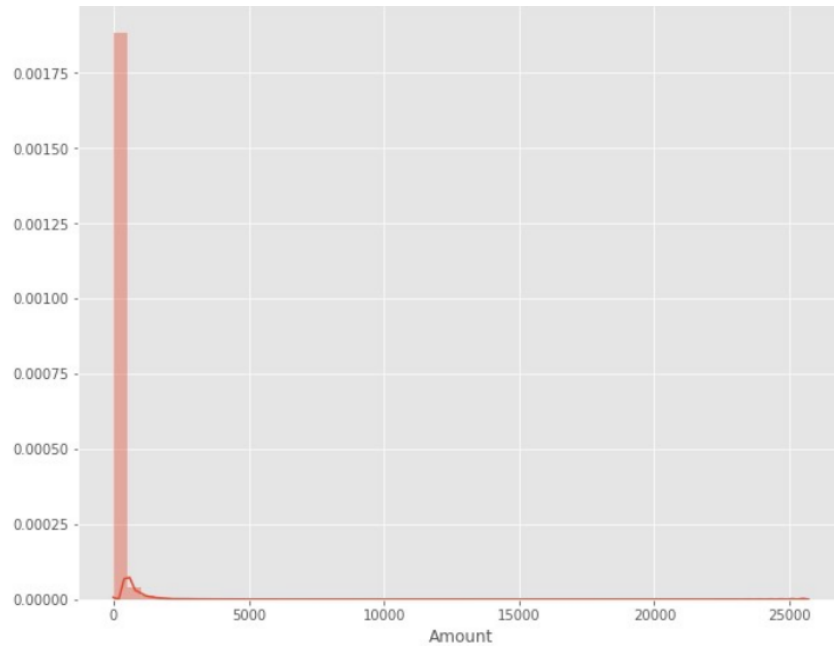


Figure 6.4 Distribution of Monetary Value Feature

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 values in the dataset. This is done to ensure that we don't require any missing value imputation and the machine learning algorithms can process the dataset smoothly.

After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between our predicting variables and the class variable. This heatmap is shown below:

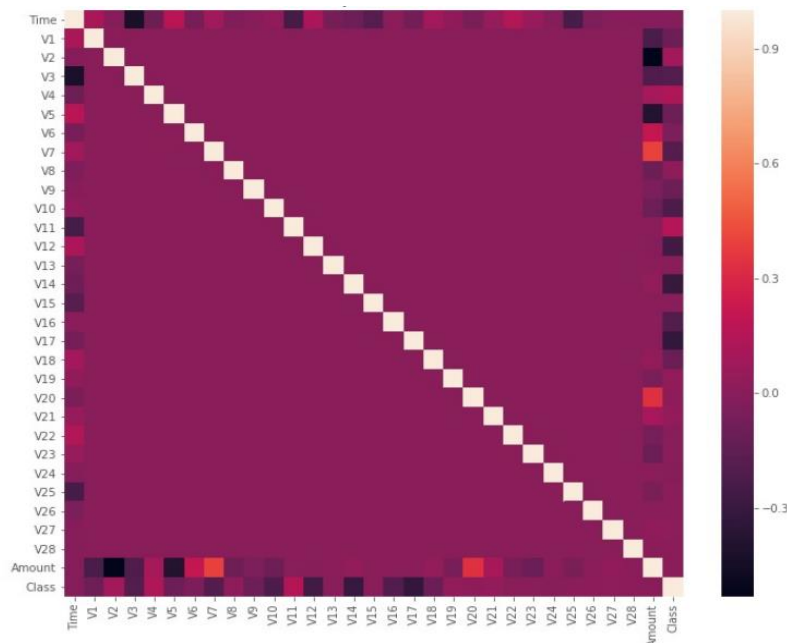


Figure 6.6 Heatmap of Correlation

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:

- ☒ Local Outlier Factor
- ☒ Isolation Forest Algorithm

These algorithms are a part of `s learn`. The ensemble module in the `s learn` 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. We've used Jupyter Notebook platform to make a program in 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:

6.1 Local Outlier Factor

It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbours. More precisely, locality is given by k -nearest neighbours, whose distance is used to estimate the local data. The pseudocode for this algorithm is written as:

On plotting the results of Local Outlier Factor algorithm, we get the following figure:

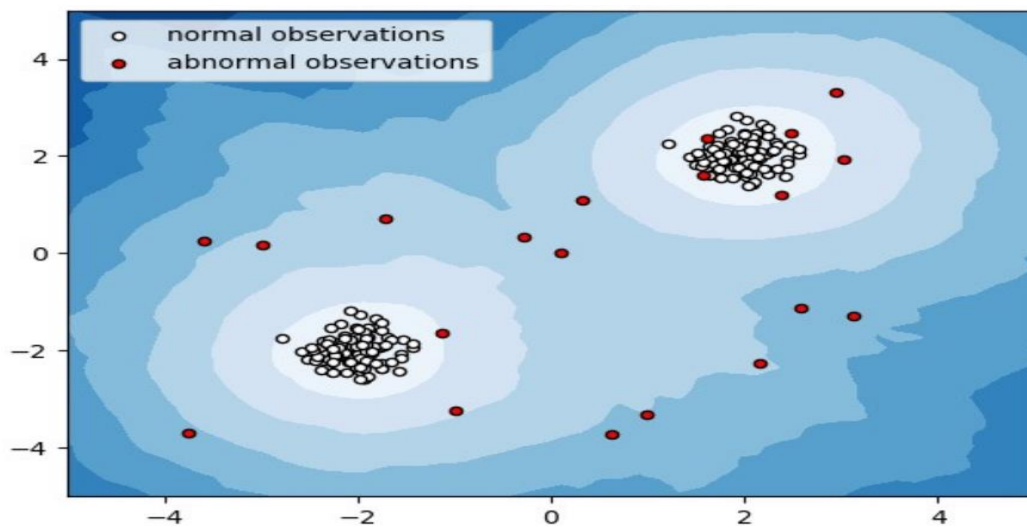


Figure 6.1.1 Location Outlier Factor

By comparing the local values of a sample to that of its neighbours, one can identify samples that are substantially lower than their neighbours. These values are quite anomalous and they are considered as outliers. As the dataset is very large, we used only a fraction of it in our tests to reduce processing times. The final result with the complete dataset processed is also determined and is given in the results section of this paper.

6.2 Isolation Forest Algorithm

The Isolation Forest 'isolates' observations by arbitrarily selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designated feature. Recursive partitioning can be represented by a tree, the number of splits required to isolate a sample is equivalent to the path length root node to terminating node. The average of this path length gives a measure of normality and the decision function which we use. The pseudocode for this algorithm can be written as:

On plotting the results of Isolation Forest algorithm, we get the following figure:

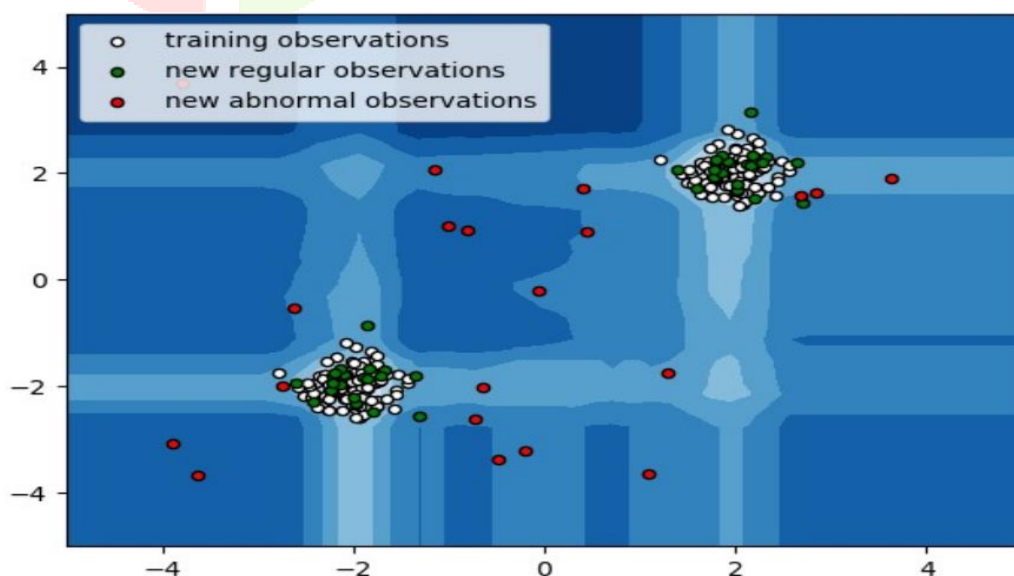


Figure 6.2.1 Isolation Forest

Partitioning them randomly produces shorter paths for anomalies. When a forest of random trees mutually produces shorter path lengths for specific samples, they are extremely likely to be anomalies. Once the anomalies are detected, the system can be used to report them to the concerned authorities. For testing purposes, we are comparing the outputs of these algorithms to determine their accuracy and precision.

RESULT

The code prints out the number of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data 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 and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives. Results when 10% of the dataset is used

Isolation Forest

Number of Errors: 71

Accuracy Score: 0.99750711000316

	precision	recall	f1-score	support
0	1.60	1.00	1.00	28432
1	1	0.28	0.29	49
accuracy			1.00	28481
macro avg	0.64	0.64	0.64	28481
weighted avg	1.60	1.00	1.00	28481

Local Outlier Factor

Number of Errors: 97

Accuracy Score: 0.9965942207085425

	precision	recall	f1-score	support
0	1.00	1.00	1.00	28432
1	0.02	0.02	0.02	49
accuracy			1.00	28481
macro avg	0.51	0.51	0.51	28481
weighted avg	1.00	1.00	1.00	28481

Results with the complete dataset is used:

Isolation Forest

Number of Errors : 659

Accuracy Score : 0.9976861523768727

	precision	recall	f1-score	support
0	1.00	1.00	1.00	284315
1	0.33	0.33	0.33	492
Accuracy			1.00	284807
macro avg	0.66	0.67	0.66	284807
weighted avg	1.00	1.00	1.00	284807

Local Outlier Factor

Number of Errors: 935

Accuracy Score: 0.9967170750718963

	precision	recall	fi-score	support
0	1.00	1.00	1.00	284315
1	0.05	0.05	0.05	492
accuracy			1.00	284807
macro avg	0.52	0.52	0.52	284807
weighted avg	1.00	1.00	1.00	284807

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 of 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 genuine transactions.

Since the entire dataset consists of only two 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 into it.

FUTURE ENHANCEMENT

While we couldn't reach our goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such projects, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision

of the algorithms increases when the size of the dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

REFERENCES

1. "Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 (IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017
2. CLIFTON PHUA, VINCENT LEE], KATE SMITH] & ROSS GAYLER² * A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia
3. "Survey Paper on Credit Card Fraud Detection by Suman", Research Scholar, GJUS&T Hisar HCE, Sonapat published by International Journal of Advanced Research in Computer Engineering & Technology (UARCET) Volume 3 Issue 3, March 2014
4. "Research on Credit Card Fraud Detection Model Based on Distance Sum - by Wen-Fang YU and Na Wang" published by 2009 international Joint Conference on Artificial Intelligence
5. "Credit Card Fraud Detection through Parenclitic Network Analysis-By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages
6. "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018
7. "Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi" published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016
8. David J.Watson,David J.Hand, M Adams,Whitrow and Piotr Juszczak "Plastic Card Fraud Detection using Peer Group Analysis" Springer, Ksue 2008.