International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

Credit Card Fraud Detectionusing MachineLearning

Poonam Tandon[1], Shivani Bawankule[2], Vaishnavi Bhaik[3], Kishor Sisodiya[4], Prof. Vanita P. Lonkar[5].
[1,2,3,4]*BE Student,* [5]*Assistant Professar*
Depertment Of Computer Science & Engineering (poonamtandon151999@gmail.com)
Govindrao Wanjari College Of Engineering And Technology Nagpur, RTMNU, Nagpur, Maharashtra, India

Submitted: 10-06-2021 Revised: 21-06-2021 Accepted: 24-06-2021

ABSTRACT-It's critical for credit card firms to be able to spot fraudulent credit card transactions so that customers aren't charged for things they didn't buy. Such issues can be solved with Data Science, which, together with Machine Learning, cannot be underestimated. With Credit Card Fraud Detection, this project demonstrates the modelling of a data collection using machine learning. Modelling prior credit card transactions with data from those that turned out to be fraudulent is part of the Credit Card Fraud Detection Problem. The model is then used to determine whether or not a new transaction is fraudulent. Our goal is to detect 100% of fraudulent transactions while reducing the number of inaccurate fraud classifications. Credit Card Fraud Detection is an example of a common classification sample. On the PCA converted Credit Card Transaction data, we concentrated on evaluating and pre-processing data sets, as well as deploying different anomaly detection techniques such as the Local Outlier Factor and Isolation Forest method.

Keywords—
Creditcardfraud,applicationsofmachinelearning,logistic regression, KNN, random forestalgorithm,localoutlierfactor,automatedfrauddetection.

I. INTRODUCTION

In credit card transactions, 'fraud' refers to the unlawful and unwelcome use of an account by someone who is not the account's owner. To stop this misuse, necessary preventative steps should be adopted, and the behaviour of such fraudulent acts can be analysed to decrease it and guard against future occurrences. In other words, credit card fraud occurs when a person uses another person's credit card for personal gain while the owner and card issuing authorities are unaware of the transaction.

This is a very important subject that requires the attention of fields like machine learning and data science, where the answer can be automated.

This issue is particularly difficult to solve from the standpoint of education because it is characterised by many elements such as class imbalance. The number of legitimate transactions considerably outnumbers the number of fraudulent transactions. Furthermore, transaction patterns frequently modify their statistical features over time.

However, these aren't the only difficulties that come with implementing a real-world fraud detection system. In real-world scenarios, automatic tools scan a vast stream of payment requests to identify which transactions to authorise.

Fraud detection is tracking the behaviours of large groups of people in order to predict, detect, or avert unacceptable behaviour such as fraud, intrusion, or defaulting.

To examine all permitted transactions and report suspect ones, machine learning techniques are used. Professionals evaluate these reports and call cardholders to establish whether the transaction was legitimate or fraudulent. The investigators submit feedback to the automated system, which is utilised to train and update the algorithm over time in order to improve fraud detection effectiveness.
Methods for detecting fraud are always being improved in order to protect criminals from altering their fraudulent schemes. These deceptions are categorised as follows:

- CreditCardFrauds:OnlineandOffline
- CardTheft
- AccountBankruptcy
- DeviceIntrusion
- ApplicationFraud
- CounterfeitCard
- TelecommunicationFraud

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

Someofthecurrentlyusedapproachestodetectionofsuchfraudare:

- ArtificialNeuralNetwork
- FuzzyLogic
- GeneticAlgorithm
- LogisticRegression
- Decisiontree
- SupportVectorMachines
- BayesianNetworks
- HiddenMarkovModel
- K-NearestNeighbour

II. LITERATURE REVIEW

Fraud is defined as an illegal or criminal deception intended to gain financial or personal gain. It is a purposeful act committed in violation of a law, rule, or policy with the intent of obtaining unlawful financial advantage.

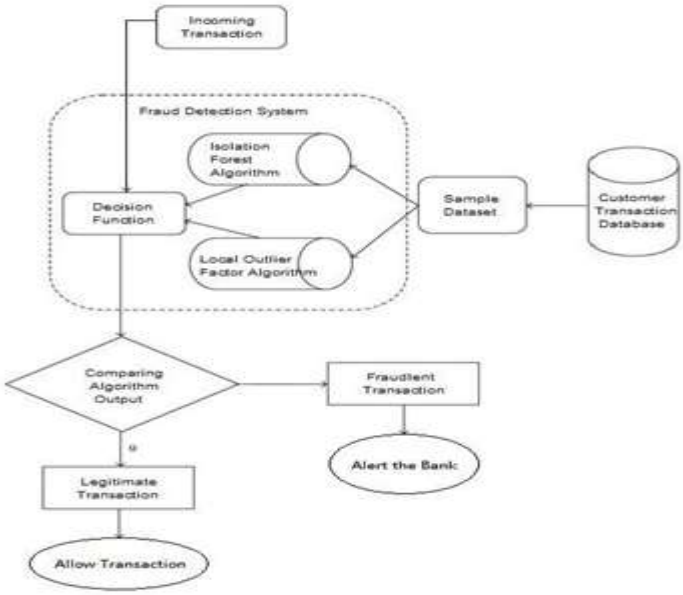
A large number of literatures on anomaly or fraud detection in this domain have previously been published and are open to the public. Data mining applications, automated fraud detection, and adversarial detection are among the strategies used in this domain, according to a comprehensive survey undertaken by Clifton Phua and his colleagues. Suman, Research Scholar, GJUS&T at Hisar HCE, proposed strategies for credit card fraud detection such as Supervised and Unsupervised Learning in another study. Despite their unexpected success in some areas, these methods and algorithms failed to provide a long-term and consistent answer to fraud detection.

Wen-Fang YU and Na Wang presented a similar study domain in which they employed Outlier mining, Outlier detection mining, and Distance sum algorithms to accurately forecast fraudulent transactions in an emulation experiment using credit card transaction data from a single commercial bank. Outlier mining is a type of data mining that is commonly utilised in the financial and internet industries. It deals with detecting items that are disconnected from the main system, such as fraudulent transactions.They took attributes of customer behaviour and estimated the distance between the observed value of that attribute and its predetermined value based on the value of those attributes.Unconventional techniques, such as hybrid data mining/complex network classification algorithm, have shown effective on medium-sized online transactions, based on network reconstruction method that allows building representations of the divergence of one instance from a reference group.

There have also been attempts to move forward from an entirely other perspective. In the event of a fraudulent transaction, efforts have been made to improve the alert-feedback interaction.

In the event of a fraudulent transaction, the authorised system will be notified, and a response will be delivered to refuse the current transaction. One of the ways that provided new light on this topic was the Artificial Genetic Algorithm, which tackled fraud from a different angle.

It was successful in detecting fraudulent transactions and reducing the amount of false alarms. Despite the fact that it was accompanied with a categorization issue with varying misclassification costs.



International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

III. METHODOLOGY

Machine Learning is the scientific study of algorithms and static models that computer system use in order to perform a specific task effectively without using the explicit instruction, relaying on patterns and interface instead.

Machine Learning algorithms build a mathematical model based on sample data, known as training data in order to make predictions and decisions.

It is the subset of artificial intelligence.

Machine Learning algorithms are used in email filtering, face recognition etc.

There are 3 types of ML algorithms – Supervised , unsupervised and reinforcement algorithm.

Supervised algorithm is of 2 types – Regression and classification.

Regression algorithms used to predict continuous values.(weather forecasting).

Classification algorithm used to predict class.(Identify fraud / binary values(yes/no))

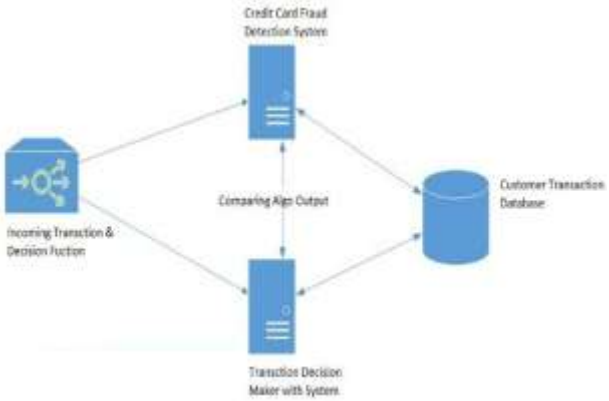
Supervised learning as the name indicates the presence of supervisor as a teacher.(historical data).

Supervised algorithm build a mathematical model of a set of data that contain both input and output.

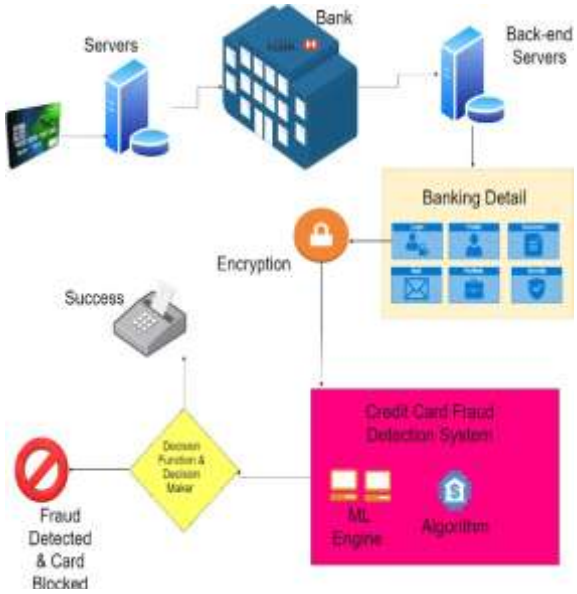
Basically supervised learning is a learning in which we teach or train the machine using data which is labelled that means some data is already tagged with the correct answer.

After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithms analyses the training data(set of training examples) and produces a correct outcome from labelled data.

The basic rough architecture diagram can be represented withthefollowingfigure:



When looked at in detail on a larger scale along with real lifeelements, the full architecture diagram can be represented asfollows:

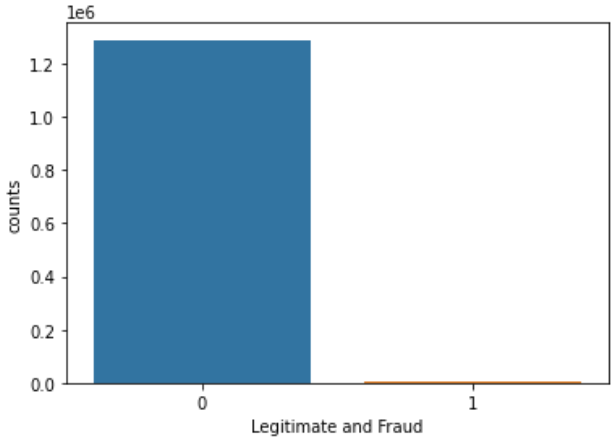


First and foremost, we got our data from Kaggle, a data analysis service that offers datasets. Inside this

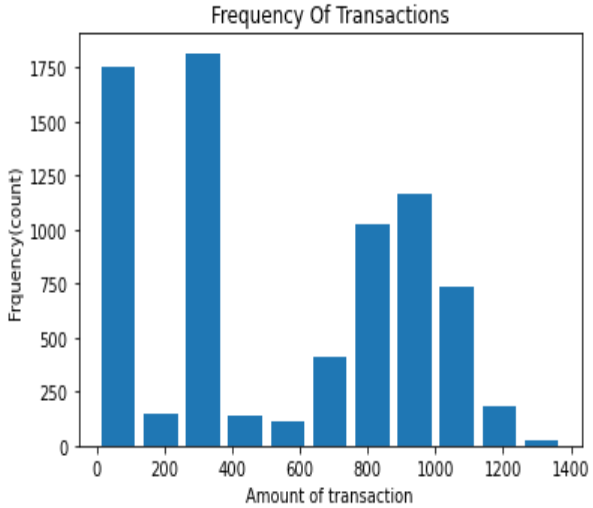
International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

dataset, there are 23 columns. We needed to perform data preparation so that system can accept the training data for its development.

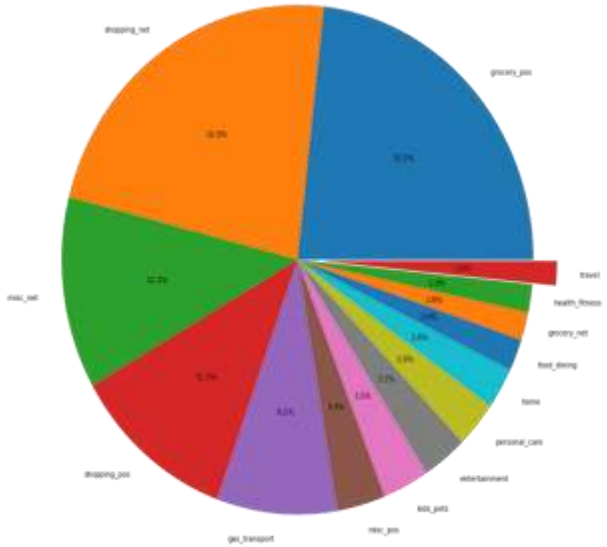
We plot different graphs to check for inconsistencies in thedatasetandtovisuallycomprehendit:



Thisgraphshowsthatthenumberoffraudulenttransactionsis much lowerthanthelegitimateones.

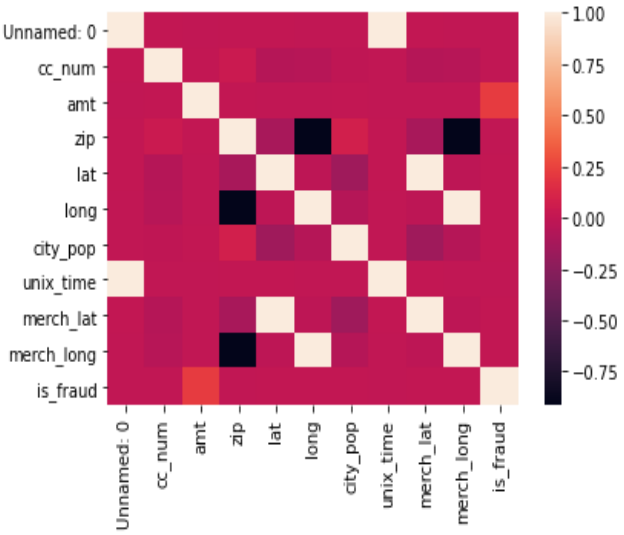


This graph shows the times frequency of transactions that of what amount of transaction has been done.



International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

For analysing categories of transaction has been done maximum and minimum time, above pie hart has been generated.
It is showing some of the categories have maximum percentage of transaction done.
For checking for outliers in amount column above graph is generated. Outliers are the distinct points in the data.
We plot a histogram for each column after checking the dataset. This is done to create a graphical representation of the dataset, which may be used to ensure that no values are missing. This is done to ensure that no missing value imputation is required and that the machine learning algorithms can analyse the dataset efficiently.
Afterthisanalysis,weplotaheatmaptogetacolouredrepresentationofthedataandtostudythecorrelationbetweenoutpredictingvariablesandtheclassvariable.Thisheatmapisshownbelow:



The dataset is now formatted and processed.A set of algorithms from modules process the data. The module diagram below depicts how these algorithms interact:

- Logistic Algorithm
- Random Forest Algorithm
- KNearest-Neighbor(KNN)

Sklearn contains these algorithms. Ensemble-based algorithms and tools for classification, regression, and outlier identification are included in the sklearn package's ensemble module.

This free and open-source Python library is made up of NumPy, SciPy, and Matplotlib modules, and it includes a number of easy and efficient data analysis and machine learning tools. It includes a number of classification, grouping, and regression methods, as well as the ability to work with numerical and scientific libraries.

We have used above given three of algorithm to build this automatic detection of fraud transaction. In this method, first have checked the accuracy of all the three algorithms, the algorithm which gave highest accuracy is used further to build the system. Followings are given findingaccuracy of each algorithm:

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

- Logistic Algorithm

```
from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import precision_score, recall_score
from sklearn.metrics import f1_score, matthews_corrcoef
from sklearn.metrics import confusion_matrix


print("The model used is Logistic Regression classifier")

acc = accuracy_score(ytest, y_pred)
print("The accuracy is {}".format(acc))

# Precision quantifies the number of positive class predictions that act
prec = precision_score(ytest, y_pred)           # Precision = TruePosit
print("The precision is {}".format(prec))

rec = recall_score(ytest, y_pred)
print("The recall is {}".format(rec))

f1 = f1_score(ytest, y_pred)
print("The F1-Score is {}".format(f1))
```

```
The model used is Logistic Regression classifier
The accuracy is 0.9954797298834742
The precision is 0.0
The recall is 0.0
The F1-Score is 0.0
```

- KNearest-Neighbor(KNN)

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=3)

# Train the model using the training sets
knn.fit(xtrain,ytrain)

#Predict Output
knn_predicted= knn.predict(xtest)
```

```
print("The model used is  KNN classifier")

acc = accuracy_score(ytest, knn_predicted)
print("The accuracy is {}".format(acc))

# Precision quantifies the number of positive class predictions
prec = precision_score(ytest, knn_predicted)  # Precision = Tru
print("The precision is {}".format(prec))

rec = recall_score(ytest, knn_predicted)
print("The recall is {}".format(rec))

f1 = f1_score(ytest, knn_predicted)
print("The F1-Score is {}".format(f1))
```

```
The model used is  KNN classifier
The accuracy is 0.9959943784538589
The precision is 0.46938775510204084
The recall is 0.2895104895104895
The F1-Score is 0.3581314878892734
```

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

- Random Forest Algorithm

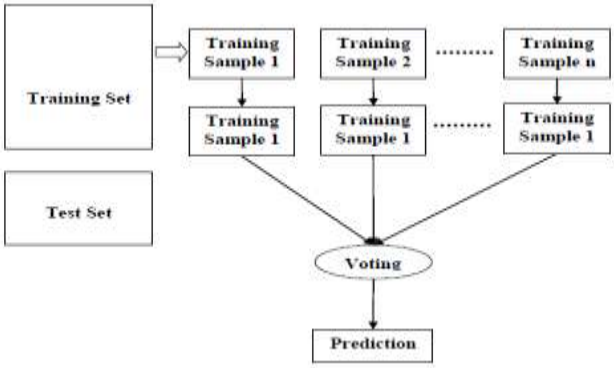
```
# Building the Random Forest Classifier (RANDOM FOREST)
from sklearn.ensemble import RandomForestClassifier
# random forest model creation
rfc = RandomForestClassifier()
rfc.fit(xtrain, ytrain)
# predictions
random_forest = rfc.predict(xtest)
```

```
print("The model used is Random forest classifier")

acc = accuracy_score(ytest, random_forest)
print("The accuracy is {}".format(acc))

# Precision quantifies the number of positive class predictions
prec = precision_score(ytest, random_forest)    # Precision = Tr
print("The precision is {}".format(prec))

rec = recall_score(ytest, random_forest)
print("The recall is {}".format(rec))

f1 = f1_score(ytest, random_forest)
print("The F1-Score is {}".format(f1))
```

```
The model used is  KNN classifier
The accuracy is 0.9974429522834383
The precision is 0.7165071770334929
The recall is 0.5585081585081585
The F1-Score is 0.6277181103222426
```

As we have seen among all three algorithm, random forest giving the highest accuracy. So random forest algorithm is the one used to build this fraud detecting system.

In this project, Supervised algorithms have been used and among them random forest algorithm gave the highest accuracy in our case. Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.



IV. IMPLEMENTATION

This concept is difficult to put into practise in practise because it necessitates the cooperation of banks, which are unwilling to share information due to market competition, as well as for legal reasons and the protection of their users' data. As a result, we searched up some reference publications that used comparable methods and gathered data.

This whole development involved processes such as data analysing, data preparation, data visualization, model selection. After model

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2592-2601 www.ijaem.net ISSN: 2395-5252

selection, training data is verified by the selected model which gave highest accuracy.
With the test data, fraud transaction detector is tested where it gave 0 or 1 as an output.

V. RESULTS

For showing the results of the system, frontend is created where test data can be applied to know the particular transaction is fraud or valid.

Below is the result generating page, where outcome of the inputs applied is shown.

Here, the outcome showing 1 i.e. the transaction is fraud.
In case if it would be 0, transaction would be valid and not the fraud.

VI. CONCLUSION

In this paper we developed a novel method for fraud detection, where fraud transaction can be identified. It helps to save the amount that every year lost by such fraud transaction. It checks the behaviour of the transaction and the behaviour in different aspects such as transaction category, transaction amount, state, etc. We have used the supervised algorithms such as logistic regression, KNN algorithm, random forest algorithm. During the model selection, we found random forest giving highest accuracy among all algorithms, So the inputs are verified by random forest model.

Credit card fraud is unquestionably a form of criminal deception. This article evaluated recent results in this field and outlined the most common types of fraud, as well as how to detect them. This automatic fraud transaction detection will give more accurate results as it will receive more accurate inputs.

VII. FUTURE ENHANCEMENT

While we didn't achieve our target of 100 % accuracy in fraud detection, we did create a system that can get extremely close to it given enough time and data. As with any effort of this nature, there is potential for improvement.

Because of the nature of this project, multiple algorithms can be merged as modules and their findings mixed to improve the accuracy of the final output.

More algorithms can be added to this model to improve it even further. These algorithms' output, however, must be in the same format as the others. The modules are simple to add once that criterion is met, as seen in the code. The project gains a lot of modularity and versatility as a result of this.

The dataset contains further room for development. As previously established, the precision of the algorithms improves as the dataset size grows. As a result, more data will undoubtedly improve the model's accuracy in detecting frauds while lowering the amount of false positives. This, however, need official backing from the banks themselves.

REFERENCES

[1]. "CreditCardFraudDetection-byIshuTrivedi,Monika,Mrigya,Mridushi" published by International Journal of Advanced Research inComputer and Communication Engineering Vol. 5, Issue 1, January2016

[2]. "Credit CardFraudDetection Based on TransactionBehaviour -byJohn Richard D. Kho, Larry A. Vea" published by Proc. of the 2017IEEE Region 10 Conference (TENCON), Malaysia, November 5-8,2017

[3]. "CreditCardFraudDetection:ARealisticModelingandaNovelLearningStrategy"publishedbyIEEETRANSACTIONSONNEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO.8, AUGUST2018

[4]. CLIFTONPHUA1,VINCENTLEE1,KATESMITH1&ROSSGAYLER2 " A Comprehensive Survey of Data Mining-based FraudDetection Research" published by School of Business Systems, FacultyofInformationTechnology,MonashUniversity,WellingtonRoad,

[5]. Clayton,Victoria3800,Australia

[6]. "Survey Paper on Credit Card Fraud Detection by Suman" , ResearchScholar,GJUS&THisarHCE,SonepatpublishedbyInternationalJournal of Advanced Research in Computer Engineering & Technology(IJARCET)Volume3Issue3,March2014

[7]. "Research on Credit Card Fraud Detection Model Based on DistanceSum–byWen-FangYUandNaWang"publishedby2009InternationalJointConferenceon Artificial Intelligence

[8]. "Credit Card Fraud Detection throughParenclitic Network Analysis-ByMassimilianoZanin,MiguelRomance,ReginoCriado,andSantiagoMoral"publishedbyHindawiComplexityVolume2018,ArticleID5764370,9pages

[9]. David J.Wetson,DavidJ.Hand,MAdams,Whitrow and PiotrJusczak"Plastic Card Fraud Detection using Peer Group Analysis" Springer,Issue 2008.

[10]. Jiang, Changjun et al. "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism." IEEE Internet of Things Journal 5 (2018): 3637-3647.

[11]. Pumsirirat, A. and Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).

[12]. Mohammed, Emad, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE

Annals of the History of Computing, IEEE,1July 2018,doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.http://www.rbi.org.in/Circular/CreditCard

[13]. https://www.ftc.gov/news-events/press-releases/2019/02/imposter-scams-top-complaints-made-ftc-2018

[14]. https://www.kaggle.com/mlg-ulb/creditcardfraud

[15]. https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset