International Journal of Modern Agriculture
ISSN: 2305-7246
Volume 10 Issue 3, 2021

CREDIT CARD FRAUD PREDICTION WITH IMPROVED ACCURACY

Dr. D. Rajalakshmi[1], Deepthi K[2], Kiruthika V[3], Jothika M[4], Indhiraelanthenral Salich J.T[5]

[1] Assistant Professor, Department of CSE, RMD Engineering College, Chennai, India
[2,3,4,5] Department of ECE, RMD Engineering College, Chennai, India
Email: [1]draji2008@gmail.com, [2]uec18128@rmd.ac.in, [3]uec18225@rmd.ac.in, [4]uec18210@rmd.ac.in, [5]uec18208@rmd.ac.in

Abstract

The emergence of modern technology has made us think in an ingenious way to save our time. In such a way innovation of Credit cards in e-commerce, technology plays a major role in our day-to-day life. Transactions are made easy in a single swipe by using credit cards. Hence, the usage of credit cards has increased worldwide, so the fraudulent actions also increased rapidly. Identification of fraudulent activities is important to protect the money of every cardholder. Referring to maximum papers, we have come up with a solution for fraudulent detection using classifiers of machine learning. Instead of using a single classifier to identify the fraudulent combining all classifiers, an average of those provides highly accurate results.

Key words: Time saves, Credit-card, Transactions, Single Swipe, Fraudulent activities, Machine learning, Average, Accuracy

Introduction

Different methodologies are adopted by technicians, researchers, data scientists, emerging engineers to fulfill the human's needs and expectations to make life more comfortable and safe. Myriad developments due to technology in the E-commerce field provide a large number of real-time applications. Nowadays money plays a huge role in everybody's life. Hence safety of the money is important. If we go outside with money there will be a fear of theft activities, so we can't concentrate on our work.

Credit card:

It is a rectangular-shaped card made up of plastic or metal. It has a unique number attached to a bank account that allows the cardholder to get goods or services without using money. The cardholder receives a bill once a month for which they have purchased.

The Credit card provides digital shopping and online bill payments. There are 2.8 Billion credit cards are in worldwide. The rise in credit card transactions leads to the rise in fraudulent activities frequently.

Types of credit cards:

I. Bank-issued credit cards

II. Store/Priority cards

III. Travel/Entertainment cards

Features of credit card:

- It is compact
- It is the alternative for money
- It stores the of detail the transaction
- It has the fund backup plans which helps in emergency
- It has an EMI facility
- It provides incentives and offers
- It is a flexible way for transactions

Credit card fraud:

Any unauthorized use of one's details to plunderage the money is known as credit card fraud. Scamming happens regularly these days. This may happen in many ways such as stealing our cards physically, hacking our computer, skimming our details, calling and asking about PINs, etc. This kind of activity takes place in health, finance, public, and many sectors. Scammers are machines that record the all details of cardholders without the knowledge of the user. Due to this, an enormous amount of money is lost every year. This affects the developing country's financial status vigorously. To overcome the increasing list of fraudulent transactions, companies are implementing modern fraud detection techniques.

Fig(2.1) Credit card Fraudulence flow chart

Credit card fraud detection:

Prevention measures have to be taken to avoid fraudulent practices. It can be done by examining the fraudulent activities and finding which technology is used against the existing technology. This ensures that in the future it won't happen again. Implementation of credit card detection is done by various methodologies. This system should be simple and cost-effective.

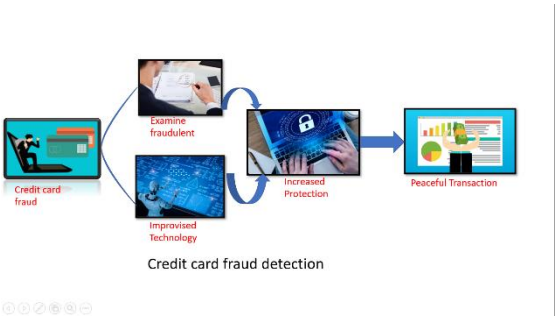


Fig 2.2 Basic Flow chart of Fraudulence

Fraud detection systems have more difficulties:

1. The credit card transaction data has contrast in nature. This is because of a very small percentage of illegal transactions happening with legal transactions.

2. In the fraud detection task, different misclassification occurs.

3. No standard method is implemented to compare the results of fraudulent transactions.

This work has come up with a comparison of some Machine learning classifiers, ensemble methods that increases the accuracy of the result. The various classifiers employed here are Logistic regression, Naive Bayes, K-nearest neighbors, Random forest, Support Vector Machine, Isolation forest.

Logistic Regression:

It is commonly used to estimate the probabilities than on instances belonging to a particular class.

Naive Bayes:

This method improves accuracy based on calculating the probabilities of required classes.

K-Nearest neighbors:

This algorithm deals with the instance's query to provide the result by using k nearest neighbors. Based on this the result is obtained.

Random forest:

This Random forest algorithm incorporates several algorithms of the same type.

Support Vector Machine:

It predicts patterns into two categories; fraudulent or non-fraudulent.

Isolation forest:

This method works based on separating irregular data. This is a very fast algorithm. It comes under the unsupervised algorithm.

All these methods mentioned are used in our paper to compare their accuracy. Comparing this most used and most important algorithm we can know which method provides the best accuracy. we do not want only the highest accuracy giving model but also we need a stable model which would predict with the best accuracy and also stable.

To make the model stable There is a method Called Ensembling,

ENSEMBLING

Ensemble methods is a technique that creates different models and later combines them to evaluate improved results. Ensemble methods generally provide more accurate results than a single model. This has been the scenario in numerous machine learning products, where the winning method is of ensemble methods. Their many methods to do ensembling but we are using only 2 methods

Taking Mean:

Averaging different predictions and producing the results

Taking mode:

Taking the most frequent prediction

Literature Survey

The major difficulties involved in credit card detection of fraud are not resolved by any method. The detection of fraud quickly and accurately are the main tasks to be deal with.

Fraud act as false deception which intends to result in personal or economic benefit. The use of Neural network technology in the detection of fraud in the banking sector has been applied to detect fraud transactions.[8,9] This was carried out by “Raghavendra Patidar” and “Lokesh Sharma”[5].An approach that was based on scattering search and the genetic algorithm was published by “EkremDuman.M Hamdi “.Their solution mainly focuses to minimize the consideration of genuine as a fraud. They merged scatter search and genetic algorithm. Among all the peer group analysis made by Whitrow and David is considered a better solution for credit card fraud detection[1].

In 2016 Shimpi surveyed different machine learning, data mining, and artificial intelligence methods used in fraud detection[4]. Also performed a comparative study in credit-card fraud detection techniques used. Tree-level security in credit card fraud detection using the hmm (Hidden Markov Model). [7]Sidharta and Yadav also used hmm for credit-card fraud detection employing dynamic random forest and K-nearest neighbor algorithms[6].

Shiyang Xuan proposed random forest and provide good result on dataset[12].Kuldeep and Randhawa used machine learning technique to detect fraud and their methods achieve good accuracy rates in fraud detection

methods[10,11].Awoyemi and john .O compare the naïve bayes and logistic regression where kNN is better.They used several algorithms to on credit card fraud data which is highly skewed[13,14,15].

Combination of supervised and unsupervised learning techniques provides higher accuracy. This was proposed by Carcillo, Le Borgne and Bontempi[17].

phau et al proposed innovative fraud detection method built upon Minority report,which deals with data mining problem.the future scope is to make appropriate than other[18,19].

Shiyang Xuan and colleagues provided a way of estimating fraudulent activities by resembling the two randomforest algorithms. They used different base classifiers[20],[21]. Always the comparison techniques provides precise results than other methodologies.

Crucial financial loses may happen if the time taken to find the illegitimate transactions is more. Hence Illegal actions towards credit cards should be found as soon as possible[22].

An approach to detect the fraudsters using classifiers of machine learning provided by V.N. Dornadula and S.Geetha[23].

Legal transactions should not be avoided while calculating the accuracy, different proposal was given by Sunil S Mhamane. He explained how the markov model used to find illegal transactions in online banking[24].

Nilson's report gives the survey on theworld wide credit card losses. More than $30 billion credit card frauds happened in 2020[25].

The fraudulent activities are happening frequently, hence determination of that activities must be fast. Profiling technique was described by Mathmound Reza Hashemi, to find the fraudsters[28].

Aditya come up with the solution for finding the fraud transactions using different classifiers of machine learning(Logistic regression and Support machine vector algorithm).This method involves to reduce the number of illegitimate transactions[29].

To overcome the fraud-related dataset with maximum accuracy machine learning techniques like Naive Bayes, Logistic Regression, K-Nearest Neighbours, Random forests. These techniques are used to find fraud transactions in real-time datasets[2,3].

The random forest method obtains good result, still, there is some problem due to imbalance data. So, the future model focus on solving the problem.

Execution of K-Nearest Neighbor, Logistic relapse, Naïve Bayes is broke down on profoundly slanted Credit Card extortion information where the examination is done on approaches that handle exceptionally irregularity charge card misrepresentation information.[26,27,28,30]

All technologies have their pros and cons as well. Our main motive is to use all technologies that can detect fraud as fast as possible and avoid loss as much as possible.

With Referring all the above concepts we proceeded with the methodology of ensembling using two simple methods i.e Taking the mean of all predictions of 6 classifiers, also Taking mode i.e taking the most frequent prediction as the result.

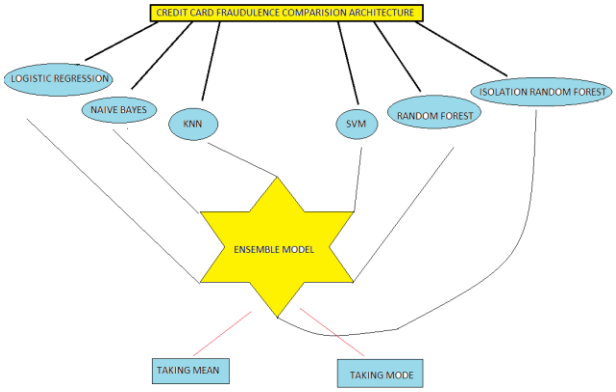


Fig 2.1 Work Flow

Methodology

Our work looks at the most recent AI classifier calculations which predict Mastercard cheats.

We have utilized a dataset given by Kaggle[16] (*It is a small data set hence we will get high accuracy than large data set, hence this work is based on this data set only*). Inside the dataset, there are 31 sections out of which 30 are utilized as highlights and the excess 1 segment is utilized as a class. Our highlights incorporate Time, Amount, and Number of exchanges

We have obtained a scatter plot (Fig 3.1) understand the data set visually

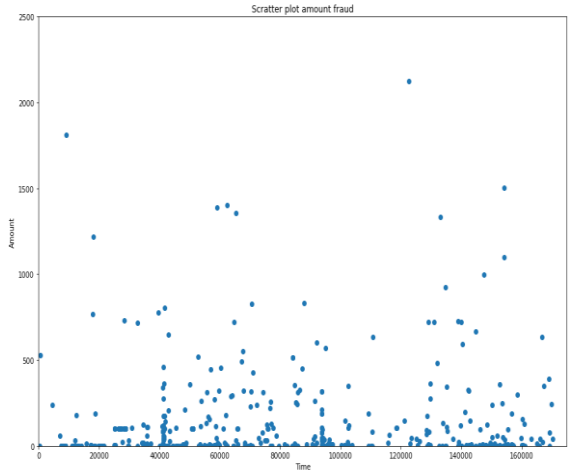


Fig 3.1 (Scatter plot)

In our work, we are trying to improve the accuracy in predicting fraud by assembling 6 machine learning classifiers as a comparative study. The classifiers we are using are

a) Logistic Regression

b) Naive Bayes

c) K-nearest neighbors

d) Random forest

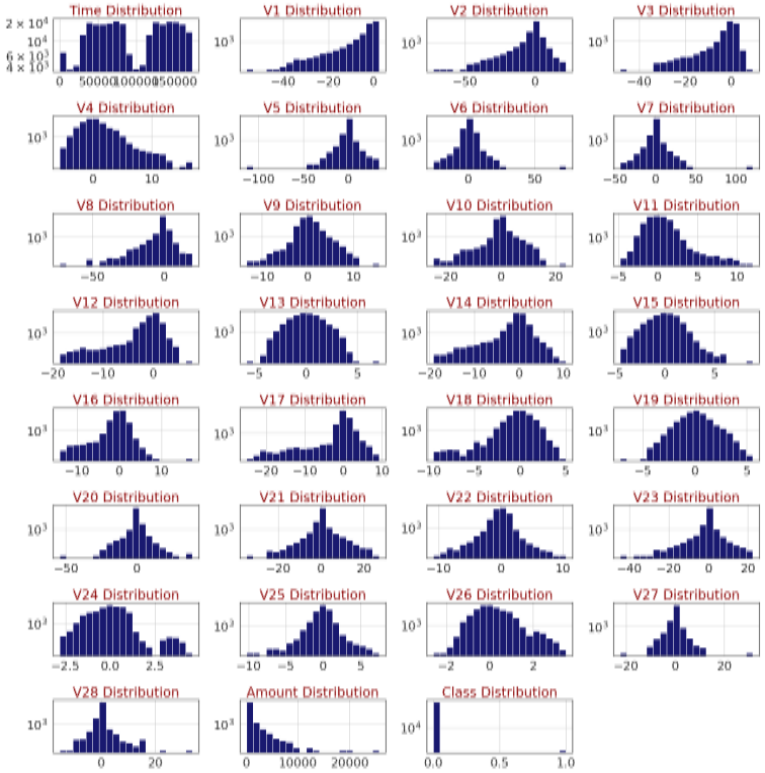
e) Isolation random forest

f) Support Vector Machine

A) Logistics Regression:

Before starting this classifier we produced a histogram visualization as found below in Fig(3.2)

International Journal of Modern Agriculture
ISSN: 2305-7246
Volume 10 Issue 3, 2021



Fig(3.2) Histogram visualization

In this visualization, we obtained a separate graph for each element for a more clear view of the dataset. Which further helps in improving the accuracy of the model.

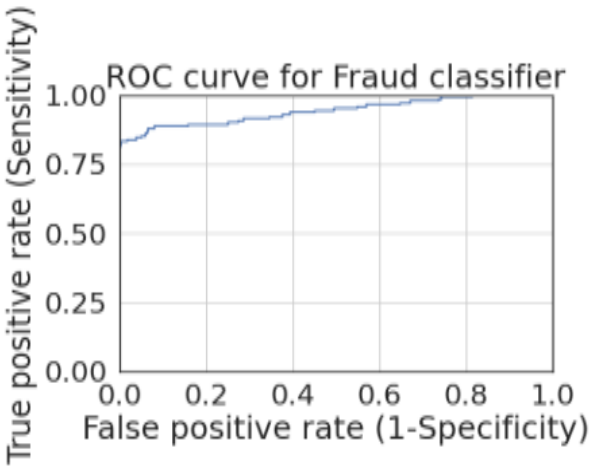
Using logistic regression we obtain an accuracy of 94.2%.

```
roc_auc_score(y_test,y_pred_prob_yes[:,1])
```

```
0.9420150279382945
```

Fig(3.2.1)Output of Logistic Regression

also with this, a graphical visualization has been obtained (Fig 3.2.2)



Fig(3.2.2) Graphical output for Logistic Regression.

B) Naive Bayes

Using Naive Bayes classification we get a prediction of about 90.8% also the code snippet is below in Fig(3.3)

```
recall score:  0.9081632653061225
precision score:  0.11323155216284987
```

(Fig 3.3) Output of Naïve bayes

C) K-nearest neighbors(KNN)

With this classifier, we obtained an accuracy up to 99.91% which was the highest accuracy among all the obtained accuracy.

```
classifier created
model evaluated
0.9991854161399961
```

(Fig 3.4)output of Knn

D) RANDOM FOREST

Random forest classifier has used both the basic random forest model and isolation random forest. By this, we could obtain 2 different accuracies where the random forest classifier obtained slightly higher accuracy than Isolation random forest.

```
the Model used is Isolation Forest
The accuracy is  0.9978933323970366
The precision is 0.375
The recall is 0.336734693877551
The F1-Score is 0.3548387096774193
The Matthews correlation coefficient is0.3543008067850027
```

Fig(3.5)Output of Isolation Random forest

```
The model used is Random Forest classifier
The accuracy is  0.9995611109160493
The precision is 0.974025974025974
The recall is 0.7653061224489796
The F1-Score is 0.8571428571428571
The Matthews correlation coefficient is 0.8631826952924256
```

Fig(3.3.3.2) Output of Random forest

E) Support Vector machine (SVM)

Using SVM, we produced a decent accuracy of prediction which was 98.99%

As shown in the output snippet

```
We have detected 175 frauds / 199 total frauds.

So, the probability to detect a fraud is 0.8793969849246231
the accuracy is : 0.9899708472111982
```

Fig (3.6)output of svm

For the same purpose, we observed different accuracy with different classifiers. Same with using different ensemble models we could obtain different accuracy. Comparing accuracy with different classifiers and different ensemble models we try to find out the best possible combination to predict the fraudulence with the highest accuracy. We are using the ensemble method to increase stability, accuracy, reduce error.

In our paper, we tend to use different Ensemble methods such as:

a) Taking MODE

b) Taking Average

a) TAKING MODE:

This method produces the most frequently occurring number found in a group of numbers. In our case, it's the prediction accuracy.

By using this method we theoretically obtain approximately > 99% accuracy.

b) TAKING AVERAGE:

This method produces the mean of all the predictions. This is also one of the simplest ensembling methods which produce accuracy > 97.2%

We can also use several other Ensembling methods, but in our paper we used only simple methods as for their simplicity in increasing stability and accuracy. As we are ensembling more than 2 models, which is already complex, a simple ensembling model will help us get a better model with a good understanding.

Results & Discussion

RANDOM FOREST	99.95%
ISOLATION RANDOM FOREST	99.78%
SVM	98.99%
KNN	99.91%
LOGISTIC REGRESSION	94.2%
NAIVE BAYES	90.8%
TAKING MODE	>=99%
TAKING MEAN	>=97.2%

Table (4.10) Accuracy Comparison table

On training each algorithm separately i.e The 6 classifiers which are 1) LOGISTIC REGRESSION; 2) KNN; 3) SVM; 4)RANDOM FOREST; 5) ISOLATION RANDOM FOREST; 6)NAÏVE BAYES.

The accuracy obtained is

1) LOGISTIC REGRESSION: 94.2%

2) KNN: 99.91%

3) SVM: 98.99

4) RANDOM FOREST: 99.95%

5) ISOLATION RANDOM FOREST 99.78%

6) NAÏVE BAYES: 90.8%

All the above accuracy of predictions are more or less similar but while observing deeply we can see that the highest produced accuracy is 99.95% which is for Is random forest. Also, we got the least prediction from the Naïve Bayes classifier.

After observing all these classifiers now we know which can produce the highest accuracy but we do not guarantee with the highest accuracy of prediction, it's also stable. So to make the model stable we shall now use the methods of ensembling. In this paper, we have used 2 methods which are

1) TAKING MEAN
2) TAKING MODE

As being the simplest method of ensembling it's easy to code with 6 classifiers. The comparison is shown in the table above. The accuracy obtained is

1) TAKING MEAN: >=97.2%
2) TAKING MODE: >=99%

5) Conclusion

A fraud detection system for credit cards by applying six different algorithms and training these algorithms with the dataset, which explains the skewness of data produces the desired outcome. Therefore, we can infer the requirement of applying this technique. Here for the dataset, we have used the Principal Component Analysis algorithm to select patterns from dataset where variance and correlation as parameters are used. After applying the six machine learning algorithms as described in the methodology, it shows high accuracy. The scores of each model were 94.2%, 90.8%, 99.91%, 99.78%,99.95%,98.99% for logistic regression, Naive Bayes, KNN, Isolation random forest, random forest respectively. After ensembling, the accuracy values theoretically are > 97.2%. And value using mode method we obtain accuracy >99%.

6) Future Scope

Reaching 100% accuracy is the target of our model.We can further implement new algorithms and classifiers. The dataset can be further improved by replacing the skewed values with normalized values for bringing a pattern that helps in building a more accurate model. These improvements will surely increase the versatility of the project and make it more accurate. Using different ensembling models which are also complex can be used to compare and get a broader view to not only get 100% accuracy but also to obtain a 100% stable model.

References

1. David J.Wetson, David J.Hand, M Adams, Whitrow and Piotr Jusczak "Credit Card Fraud Detection utilizing peer bunch investigation" Springer Issue 2008.
2. Francisca NonyelumOgwuleka, "Information Mining application in charge card misrepresentation detection"Journal of designing science and technology,vol 6 no 3, issue 2011.
3. John T.S Quah,MSriganesh "Ongoing charge card misrepresentation identification utilizing computational Intelligence" ELSEVIER Science immediate, 35(2008) 1721-1732.
4. P.R.Shimphi, 'Overview on charge card misrepresentation identification methods', Int .J.Eng.Comput.Sci.,2016.
5. Patidar and L.Sharma," Credіy card misrepresentation recognition utilizing neural systems administration " NCAI2011,13 may 2011,Jaipur,India,International Journal of delicate processing ndenngineering (IJSCE) ISSN:june 2011.
6. S.Yadav and S.Siddhartha, 'Misrepresentation recognition of Credit card by utilizing HMM model', Int.J.Res.Eng.Technol.,vol.6.no1 ,pp.41-46,2018.
7. 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.

8. Pumsirirat, An and Yan,L.(2018).Credit Card Fraud Detection utilizing Deep Learning dependent on Auto-Encoder and Restricted Boltzmann Machine. Worldwide Journal of Advanced Computer Science and Applications. Charge card Fraud Detection Based on Transaction Behavior by John Richard D.Kho,LarryA.Vea" distributed by Proc . of the 2017 IEEE Region 10 Conference (TENCON),Malaysia, November 5-8, 20107
9. Mohammed, Emad, and Behrouz Far. "Administered Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1 July 2018, doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.
10. Randhawa, Kuldeep, et al. "Charge card Fraud Detection Using AdaBoost and Majority Voting." IEEE Access, vol. 6, 2018, pp. 14277–14284., doi:10.1109/access.2018.2806420.
11. Roy, Abhimanyu, et al. "Profound Learning Detecting Fraud in Credit Card Transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS), 2018, doi:10.1109/sieds.2018.8374722.
12. Xuan, Shiyang, et al. "Arbitrary Forest for Credit Card Fraud Detection." 2018 IEEE fifteenth International Conference on Networking, Sensing and Control (ICNSC), 2018, doi:10.1109/icnsc.2018.8361343.
13. Awoyemi, John O., et al. "Visa Fraud Detection Using Machine Learning Techniques: A Comparative Analysis." 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, doi:10.1109/iccni.2017.8123782.
14. Melo-Acosta, German E., et al. "Misrepresentation Detection in Big Data Using Supervised and Semi-Supervised Learning Techniques." 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 2017, doi:10.1109/colcomcon.2017.8088206.
15. http://www.rbi.org.in/roundabout/credircard.
16. https://www.kaggle.com/mlg-ulb/credircardfraud
17. F. Carcillo, Y.- A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Consolidating unaided and regulated learning in Mastercard extortion location," Information Sciences,May 2019, doi: 10.1016/j.ins.2019.05.042
18. J.R. Dorronsoro, F. Ginel, C. Sgnchez and C.S. Cruz, ―Neural misrepresentation identification in Visa operations‖, IEEEtransaction neural organization vol. 8, no. 4, pp. 827-834, Jul.1997.
19. Sai Kiran, Jypti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya,‖Credit card misrepresentation location usingNaïve Bayes model based and KNN classifier‖, Int. Diary of Adv. Exploration , Ideas and Innovations in Technology,vol.4,2018.
20. Phua, d.Alahakoon and V.Lee "Minority report in misrepresentation discovery characterization of slanted data".ACM SIGKDD investigations pamphlet vol 6.no.pp 50-59.
21. Credit card misrepresentation discovery utilizing covered up markov model – AbinavSrivastava,AmlanKundu,ShamikSural,ArunK.majumdar
22. ShiyangXuan, GuanjunLiu, ZhenchuanLi, LutaoZheng, ShuoWang, Jiang, Random Forest for Mastercard extortion detection‖, Int.conf.on Networking, Sensing and control,2018.
23. A. G. C. de Sá, A. C. M. Pereira, and G. L. Pappa, "A modified grouping calculation for Mastercard misrepresentation location," Engineering Applications of Artificial Intelligence, vol. 72, pp. 21–29, Jun. 2018, doi: 10.1016/j.engappai.2018.03.011.
24. V. N. Dornadula and S. Geetha, "Visa Fraud Detection utilizing Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 631–641, 2019, doi: 10.1016/j.procs.2020.01.057.
25. Sunil S Mhamane and L.M.R.J Lobo "Utilization of Hidden Markov Model as Internet Banking Fraud Detection" International Journal of Computer Applications (0975 – 8887) Volume 45–No.21, May 2012
26. The Nilson Report. (2015). Worldwide extortion misfortunes reach $16.31 Billion. Version: July 2015, Issue 1068.
27. N. F. Ryman-Tubb, P. Krause, and W. Garn, "What Artificial Intelligence and AI research means for installment card misrepresentation discovery: An overview and industry benchmark," Engineering Applications of Artificial Intelligence, vol. 76, pp. 130–15 Nov. 2018.
28. Wen-Fang YU, Na Wang, Research on Credit Card Fraud Detection Model Based on Distance Sum, IEEE International Joint Conference on Artificial
29. Mathmound Reza Hashemi, Research on "Mining data from Visa time arrangement for more ideal extortion location
30. Aditya Oza, Fraud Detection utilizing MachineLearning,aditya19@stanford.edu