International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-8, June 2021

Financial Fraud Detection in Plastic Payment Cards using Isolation Forest Algorithm

Ankaj Kumar, Gouri Sankar Mishra, Parma Nand, Madhav Singh Chahar, Sonu Kumar Mahto

Abstract: ***The need for technology has always found space in Financial Transaction as the number of fraud in financial transactions increases day by day. In this research we have proposed a new methodology by using the isolation forest algorithm and local outlier detection algorithm to detect the financial fraud. A standard data set is used in experimentation to classify a transaction occurred is a fraudulent or not. We have used neural networks and machine learning for classification. We have focused on the deployment of anomaly detection algorithms that is Local Outlier Factor and Isolation Forest algorithm (IFA) on financial fraud transactions data.***

Keywords: ***In This Research We Have Proposed A New Methodology By Using The Isolation Forest Algorithm And Local Outlier Detection Algorithm To Detect The Financial Fraud.***

I. INTRODUCTION

In this digital world in an online transaction, there are many fraud cases like credit card fraud, debit card cloning etc. Fraud is an act of deception used to illegally depose another money, property and legal rights. A typical organization loses 5% of their yearly revenues [ ] in fraudulent transactions. According to an RTI (Right To Information) 2480 cases [ ] in public sectors banks, it is found that fraudulent transaction occurred around Rs.31865.8 crore. In the financial year 2018-19, RBI has reported total number of 911 credits card fraudulent transactions [ ] happened amounting Rs.65.98 crore loss of customers. This research is about detecting the financial fraud detection using machine learning so that the user will be able to detect fraudulent transactions occurring in his debit or, credit cards. Neural Network and fuzzy logic is widely used in this method to detect credit card fraud detection [ ] and we can't ignore the use of artificial intelligence in this field without using fuzzy databases and machine learning techniques.

Manuscript received on April 28, 2021.
Revised Manuscript received on June 13, 2021.
Manuscript published on June 30, 2021.
* Correspondence Author
Ankaj Kumar, Department of Computer Science and Engineering, Sharda University, Greater Noida (Uttar Pradesh), India. Email: ankajraj106@gmail.com
Gouri Sankar Mishra, Department of Computer Science and Engineering, Sharda University, Greater Noida (Uttar Pradesh), India. Email: gourisankar.mishra@sharda.ac.in
Parma Nand, Department of Computer Science and Engineering, Sharda University, Greater Noida (Uttar Pradesh), India. Email: parma.nand@sharda.ac.in
Madhav Singh Chahar, Department of Computer Science and Engineering, Sharda University, Greater Noida (Uttar Pradesh), India. Email: madhavc232@gmail.com
Sonu Kumar Mahto*, Department of Computer Science and Engineering, Sharda University, Greater Noida (Uttar Pradesh), India. Email: sonumahtoraj121@gmail.com

II. LITERATURE SURVEY

It has been reported some research papers in the study of detecting the credit card frauds and used different methodologies i.e. example sliding window technique [ ], Artificial Neural Networks [ ], decision tree [ ], fuzzy logic [ ], Support Vector Machine [ ], Hidden Markov Model[ ]. To summarize the existing systems are in following categories of application of this problem:

(1) Certain sets of rules are made by observing previous behavior of frauds.

(2) The historical data set of transactions is used to train the machine learning model and predict fraud and legit behavior of transactions for the new transaction.

(3) The real-time application of the technique in the payment process of businesses.

A general approach flow of fraud detection in transaction [ ] can be shown as:

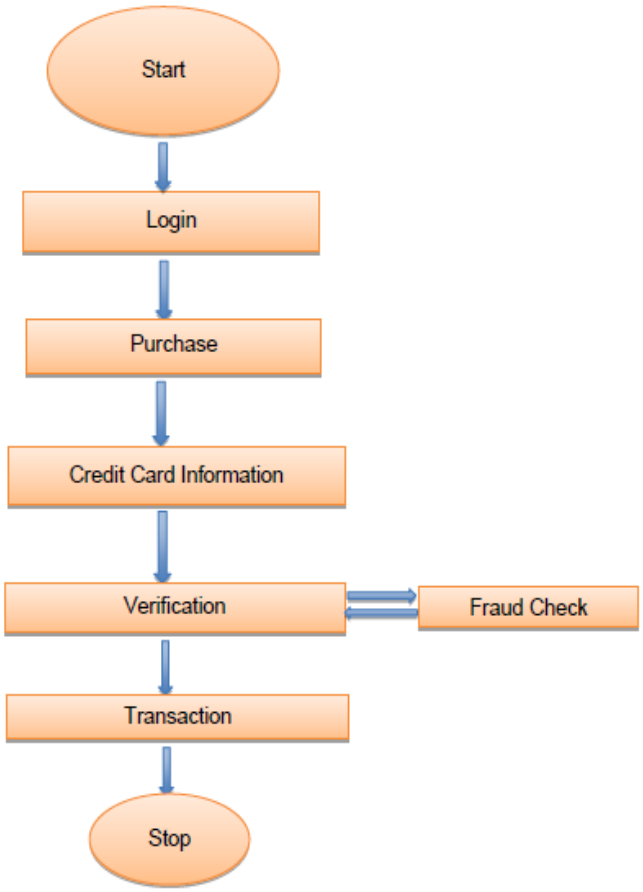

Fig.1: A general approach flow of fraud detection

Retrieval Number: 100.1/ijitee.G88730510721
DOI: 10.35940/ijitee.G8873.0610821
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

Methodologies implemented widely in existing models are as follows:

2.1. Artificial Neural Networks

Artificial neural networks (ANN) mimic the human brain, having nodes connected with each other to communicate just like the neurons in human brain. ANN use previous information to make decision and produce outputs. Similarly in validating the transactions historic data is given as input and based on that it recognizes whether the transaction was legit or fraud [ ]. The ANN can be trained using supervised and unsupervised learning. In supervised learning the result of check is known and output is matched with the actual value to check whether it is correct or not. In unsupervised learning there is no data to compare the result with, hence there is uncertainty about the result.

2.2. Decision Tree

Decision tree is type of supervised learning where the data set need to be trained on the fraudulent transaction data set [ ]. The data is split on decision nodes and leaves contain the final output. For this kind of data in particular the decision tree has the leaf nodes that contain the final class label and the paths contain the results. In decision tree data is partitioned recursively either using breadth first search or depth first search until the all the element for the current class are assigned. Partitioning is most effective when the subparts of data do not overlap.

2.3. Fuzzy Logic

Generally Fuzzy logic methodology is used when the values are not in true or false (or 0 or 1) nature whereas the logic is defined with the degree of true or false. As this is different case than others because the behavior is continuous in nature. So certain rules must be followed to deal with it. For analyzing whether the transaction is fraudulent or not [ ], following rules are be followed in given order: a) fuzzification, b) Rule based, c) defuzzification. The transactions are labelled in three categories as high, medium and low based on their given values associated with transaction. Then second step is followed to classify the transaction based on customer behavior. Finally defuzzification is done, if the transaction doesn't align with the predefined behavior then it is flagged and not allowed to move forward until crosscheck verification from the user.

2.4. Support Vector Machines

SVM is very effective in dealing with the complex data and build a robust machine learning model. Like decision tree, it is also the supervised machine learning algorithm which is generally used for classification by using the number of feature and the value to plot in 'n' dimension space where 'n' is number of features and determines the hyper plane. The points near to a hyper-plane are called as support vectors. These support vectors helps in the classification of the point based on which side of the plane it lies. In the fraud detection historic data is fed to the model [ ]. Therefore the behavior of the transactions is known. After the data is trained, new transactions can be identified as fraud or legit.

2.5. Bayesian Network

Bayes theorem for conditional probability is used in this model [ ]. This is a probabilistic approach is used where the graphical model having set of values and their conditional probabilities are interdependent. This graphical model is represented in a directed acyclic graph. Each node of this graph represents the variables and the edges represent the relationship with each other. The historic data is fed which has values as fraud and legit already known so the probabilities of transaction being fraud or legit is calculated. The transaction can be flagged fraud when its probability is less than minimum probability of begin legit and greater than maximum probability of being fraud.

2.6. K- Nearest Neighbor

K-Nearest Neighbor (KNN) method can be used for both classification and regression. So the outcomes may vary based on the purpose of use [ ]. The outcomes of KNN depends on three parameters which are:

1) Value of K is the number of comparisons with the neighboring data points.

2) Distance rule is useful in classifying the new points into one class by comparing its features with that of neighboring data point.

3) Distance metrics contains the distance of nearest neighbor for the next coming point.

The value of k is calculated from the graph based on the validation curve, now this value of k will remain constant for all the predictions.

2.7. Hidden Markov Model

Hidden Markov Model (HMM) is the foundation of sequence analysis and is used in the technologies like profile searches and gene finding. In this model the state change with respect to time is key feature. With the above property, the hidden states can't be identified directly. But the feature that relates with the current one, can be used for observation of the new state changes. The model is trained on features like the expenditure behavior of user [ ]. When a new transaction occurs, it is analyzed based on the variance of value compared with the threshold.

2.8. Logistic regression

This is a supervised learning algorithm. Unlike the linear regression the output value is either zero or one instead of some other numeric value. This can also be implemented for real time application as it works on clusters of the data and while the transaction is running, it flag whether it should be proceeded or not [ ].

III. PROPOSED METHODOLOGY

The objective of the proposed system is to train the data set in order to produce precise prediction on the behavior of the transaction whether it is legit or fraud. We have used the isolation forest algorithm and local outlier detection algorithm in the proposed methodology. The challenge is to normalize the data set as the fetched data set is imbalanced in fraudulent and legitimate cases, else it will favor the majority which is legit cases and decreases the accuracy of prediction. Therefore, in order to resolve the problem of imbalanced data we have followed the following methods:

Retrieval Number: 100.1/ijitee.G88730510721
DOI: 10.35940/ijitee.G8873.0610821
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

1) *Random over sampling:* In this resampling method, the value from minority class are copied to balance the difference between majority and minority class.

2) *Random under sampling:* In this resampling technique the values are deleted randomly from majority class for the balancing of data.

3) *SMOTE:* To solve the binary classification problem while working on it, we need Smote because it helps in generating samples from the minority class, unlike the above two techniques it does not copy the minority class values but instead it synthesizes new values similar to minority class.

The combination of isolation forest and local outlier factor is to make an improvement over each of them separately. Isolation Forest is better for global outliers but is poor in dealing with local outliers whereas the local outlier factor performs well in local outlier detection. It has high time complexity too. In order to resolve the shortcomings of both, they are coupled together to complement each other and produce better results than performing alone.

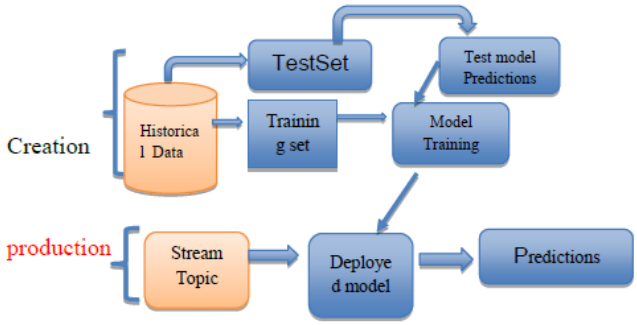

Figure-2: Proposed methodology

3.1. Local Outlier Factor

Anomaly score of every sample is known as Local Outlier Factor. It is used to find how much local deviation has happened with respect to neighborhood. Locality is given by K-nearest algorithm. Which is use to find the median. The following figure shows the data points and outlier scores.

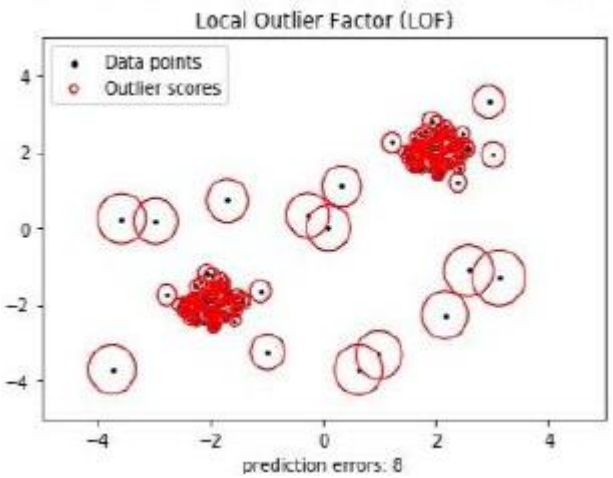

Figure-3: Loical Outlier Factor

3.2. Isolation Forest

The Isolation Forest is based on random forest technique. It isolates observations by randomly selecting a feature and then a random value is selected between the maximum and minimum value of selected 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.

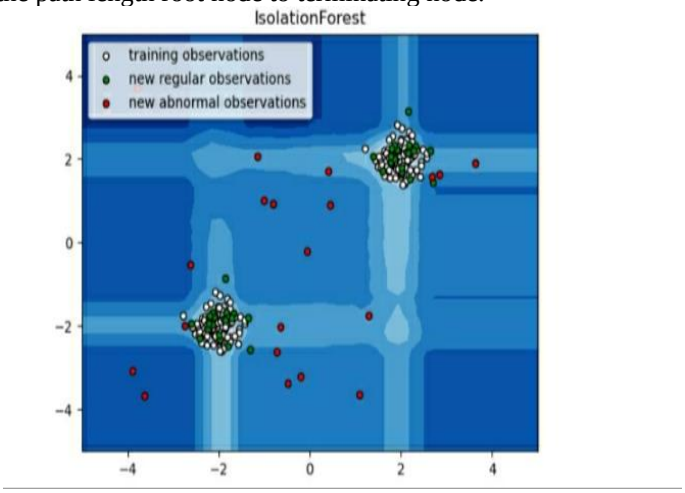

Figure-4: Isolation Forest

The first step in the proposed method is to describe and visualize data in various ways to get the features correlate with each other. A correlation matrix and the details of rows and columns is given below.

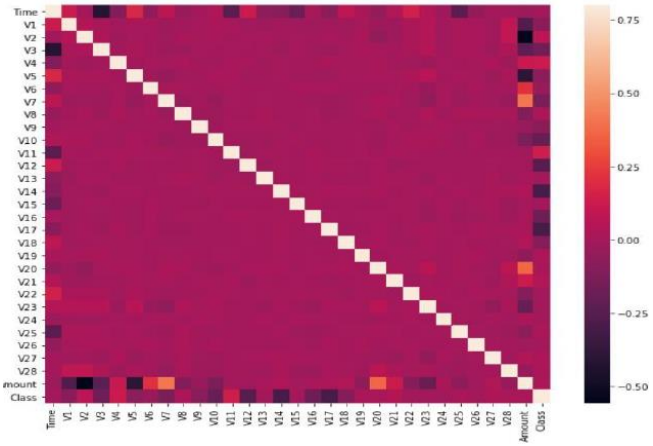

Figure-5: Correlation matrix

It can be observed the feature correlation from the heatmap given in Fig.5. Most of the features do not correlate although some features that show positive or negative relationships with each other. For example V2 and V5 have a negative relationship with feature amount.

IV. EXPERIMENTAL RESULTS

Two major observations have been found from the experimental results out of that, it gives the accuracy and recall which are basically the percentage of predicting a legit transaction as true and recall is prediction false as false. So, it is observed that the proposed methodology reaches accuracy of 99.7 (~99.8) % with precision value at 36% and that too for only 10% of total data set. Its performance will increase as the number of transactions will increase.

Retrieval Number: 100.1/ijitee.G88730510721
DOI: 10.35940/ijitee.G8873.0610821
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

```
Isolation Forest: 63
0.9977879990168884
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.36      0.37      0.36        49

    accuracy                           1.00     28481
   macro avg       0.68      0.68      0.68     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor: 97
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
```

Figure-6: Experimental results

The comparison of the results found from the experimentation of proposed methodology with the results of existing methodologies of fraudulent transaction detection is given in Table-1.

Table-1: Comparison of proposed method with existing methods

Method Used	Frauds	Genuines	MCC
Naïve Bayes	83.130	97.730	0.219
Decision Tree	81.098	99.951	0.775
Random Forest	42.683	99.988	0.604
Gradient Boosted Tree	81.098	99.936	0.746
Decision Stump	66.870	99.963	0.711
Random Tree	32.520	99.982	0.497
Deep Learning	81.504	99.676	0.812
Neural Network	82.317	99.966	0.787
Multi Layer Perceptron	80.894	99.998	0.806
Linear Regression	54.065	99.985	0.683
Logistic Regression	79.065	99.975	0.787
Support Vector Machine	79.878	99.971	0.812

V. CONCLUSION

Various machine learning algorithms are reviewed for detecting the fraudulent transactions through debit and credit cards. The proposed methodology is experimentally found to achieve 99.6% accuracy. It is also found that the precision is 36% since the data set was taken from the recorded transaction of only two days which is only 10% of the entire data set. The precision will improve with respect time as the model will learn from subsequent transaction data.

REFERENCES

1. "S P Maniraj,Aditya Saini,Shadab Ahmed And Swarna Deep Sarkar"Credit Card Fraud Detection using machine learning and Data Science," 2019 International Journal of Engineering Research & Technology (IJERT), 2019,IJERTV8IS090031
2. "Suchita Anant Padvekar,Pragati Madan Kangane,Komal Vikas Jadhav "Credit Card Fraud Detection System" International Journal of Engineering and Computer Science(IJECS) ISSN: 2319-7242, Volume-5, Issue-4 April 2016,Page No.16183-16186
3. "Vaishnavi Nath Dornadula,Geetha S "Credit Card Fraud Detection using Machine Learning Algorithms " International Confrence on Recent Trends in Advanced Computing(ICRTAC) ,Procedia Computer Science165(2019)631-641
4. "Yashvi Jain,Narmata Tiwari,Shripriya Dubey,Sarika jain "A Comparative Analysis of Various Credit Card Fraud Detection Techniques" International Journal of Engineering Research & Technology (IJERT),ISSN:2277-3878, Volume-7, Issue-5S2,January 2019
5. "E.Burnaev,P.Erofeev,A.papanov "Influence of Resampling on Accuracy of Imblanced Classification" arXiv.org arXiv:1707.03905v1[stat.ML]12 July 2017
6. "M. Carminati, L. Valentini, and S. Zanero", "A Supervised Auto-Tuning Approach for a Banking Fraud Detection System," in *Proceeding of the International Conference on Cyber Security Cryptography and Machine Learning*, Springer, Cham, Switzerland, 2017.
7. "F. Fadaei Noghani and M. Moattar", "Ensemble Classification and Extended Feature Selection for Credit Card Fraud Detection",*Journal of AI and Data Mining*, vol. 5, no. 2, pp. 235–243, 2017.
8. "M. Kamboj and G. Shankey", "Credit Card Fraud Detection and False Alarms Reduction using Support Vector Machines," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 2, no. 4, 2016.
9. "C.-C. Chiu and C.-Y. Tsai", "A web services-based collaborative scheme for credit card fraud detection," in *Proceedings of the e-Technology, e-Commerce and e-Service, 2004. EEE'04. 2004 IEEE International Conference on*, IEEE, 2004.

10. "K. K. Sherly and R. Nedunchezhian", "Boat adaptive credit card fraud detection system," in *Proceedings of the Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on*, pp. 1–7, 2010.
11. "K. RamaKalyani and D. UmaDevi", "Fraud detection of credit card payment system by genetic algorithm," *International Journal of Scientific & Engineering Research*, volume 3, no. 7, 2012.
12. "E. Aleskerov, B. Freisleben, and R. Bharat", "Cardwatch: A neural network based database mining system for credit card fraud detection," in *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, 1997.
13. "M. Bansal and Suman", "Credit card fraud detection using self organised map," *International Journal of Information & Computation Technology*, vol. 4, no. 13, pp. 1343–1348, 2014.
14. "K. K. Sherly and R. Nedunchezhian", "Boat adaptive credit card fraud detection system," in *Proceedings of the Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on*, pp. 1–7, 2010.
15. "V. V. Vlasselaer et al. ", "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions," *Decision Support Systems*, vol. 75, 2015.

AUTHORS PROFILE

Sonu Kumar Mahto, CSE, Sharda University I have completed my UG course and completed the internship in Ksolves. I am interested in Machine Learning and done basic course like Andrew Nag course on Coursera ,NPTEL and Udemy. I am undergoing rigorous training of Machine Learning . I will do further research on Fraud detection of financial transaction in recent years.

Dr. Gouri Sankar Mishra, a researcher in the field of Machine Translation, Natural Language Processing. He is working as Assistant Professor in Sharda University, Greater NOIDA and has been contributing in academics since last 20 years. Author of 11 international publications. Delivered invited talks in conferences and lectures in faculty development programs organized in different Universities.

Prof. Parmanand, A visionary, eminent academician, researcher in the field of Algorithms, AI & Machine learning, NLP, wireless networks etc. with PhD in Computer Science & Engineering from IIT Roorkee and Bachelor & Masters degree from in Technology from IIT Delhi, India. Having vast experience of 26 years, he is the author of more than 70 research papers in international publications. Being a constant source of inspiration within the students and researchers, he has delivered talks on International conferences, workshops and symposiums.

Madhav Singh Chahar, CSE, Sharda University
I have Completed my Engineering in computer science Major, Currently i work as React Developer at Daffodil Software Ltd. I am Well versed with web technologies, but i chose
Machine learning as my research subject because it has always fascinated me, I look
Forward to working on more of machine learning and Artificial intelligence projects.

Ankaj Kumar, I completed my bachelor's degree from the School of engineering and Technology Sharda University, Greater noida Uttar Pradesh in 2021. In Computer Science & Engineering specialization. I am currently thinking of pursuing M.tech from any reputed university. I want to do work on research on a machine learning project. I got a job offer letter from two or three companies. I currently done my research on financial fraud detection. that was a very interesting journey for me i learn very much thing from their.