IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105114

Credit Card Fraud Detection using Machine Learning Techniques

[1]Mrs. H. A. Shinde[1], [2]Riddhi Vikas Bhosale, Siddhi Vikas Bhosale, Chaitanya Pandey, Bhushan Tamhane

[1]Lecturer, Department of Computer Engineering, AISSMS Polytechnic, Pune, India

[2]Students, Department of Computer Engineering, AISSMS Polytechnic, Pune, India

Abstract– The immense growth of e-commerce and increased online based payment possibilities has result in credit card fraud, which has become deeply relevant global issue. Financial fraud is an ever growing issue in the financial industry. Data mining plays important role in the detection of credit card fraud while doing online transactions.
Visa misrepresentation identification, which is an information mining issue, gets testing because of two significant reasons – first, the profiles of ordinary and deceitful practices change continually and also, Credit card extortion informational indexes are exceptionally slanted. The presentation of misrepresentation location in Credit card exchanges is significantly influenced by the testing approach on data-set, determination of factors and discovery technique(s) utilized.
Data-set of Credit card exchanges is sourced from European cardholders containing 284,807 exchanges. Notwithstanding, number of difficulties shows up, for example, absence of freely accessible datasets, exceptionally imbalanced class sizes, variation deceitful conduct and so forth A half and half method of under-examining and oversampling is completed on the slanted information. Our proposed framework will utilize Disengagement Woods Calculation which is a solo learning calculation. for peculiarity location that chips away at the standard of separating inconsistencies, rather than the most well-known methods of profiling ordinary focuses. The work is executed in Python. One of the upsides of utilizing the confinement backwoods calculation is that it recognizes oddities quicker as well as requires less memory contrasted with other oddity location calculation.

Keywords— Credit Card, Fraud, Forest Isolation Technique, comparative analysis, Machine Learning, Local Outlier Factor.

I. INTRODUCTION

As Credit Card turns into the most broad method of installment (both on the web and standard buy), misrepresentation rate will in general speed up. Distinguishing false exchanges utilizing customary strategies for manual location are tedious and incorrect; hence the approach of enormous information had made these manual techniques more unreasonable. Nonetheless, monetary establishments have gone to insightful strategies. These savvy misrepresentation methods involve calculation al insight (CI)- based procedures.
Factual misrepresentation location strategies have been isolated into two general classifications: managed and unaided. In directed extortion recognition strategies, models are assessed dependent on the examples of fake and genuine exchanges to characterize new exchanges as deceitful or authentic while in unaided misrepresentation identification, exception exchanges are distinguished as possible occasions of false exchanges.
Fraud detection methods are continuously developed for adapting to new fraudulent strategies. The frauds are classified as:
Credit Card Frauds are: Online and Offline

- Card Theft.
- Account Bankruptcy.
- Device Intrusion.
- Application Fraud.
- Counterfeit Card.
- Telecommunication Fraud

Some of currently used algorithms for detection of such frauds are:

- Artificial Neural Network.
- Fuzzy Logic.
- Genetic Algorithm.
- Logistic Regression.
- Decision tree.

This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105114

- Support Vector Machines.
- Bayesian Networks.
- Hidden Markov Model.
- K-Nearest Neighbour.

II. LITERATURE SURVEY

[1]Financial fraud is an ever growing menace with far reaching consequences in the finance industry, corporate organizations, and government. Fraud can be defined as criminal deception with intent of acquiring financial gain. High dependence on internet technology has enjoyed increased credit card transactions.[2]Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behavior, 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.[3][3]Data mining technique is one notable methods used in solving credit fraud detection problem. Credit card fraud detection is the process of identifying those transactions that are fraudulent into two classes of legitimate (genuine) and fraudulent transactions.[4] Credit card fraud detection is based on analysis of a card's spending behavior. Many techniques have been applied to credit card fraud detection, artificial neural network.[5][6] genetic algorithm.[7] is evaluated on credit card fraud data. Decision tree, neural networks and logistic regression are tested for their applicability in fraud detection[8] This paper seeks to carry out comparative analysis of credit card fraud detection using naive Bayes, k-nearest neighbor and logistic regression techniques on highly skewed data based on accuracy, sensitivity, specificity and Matthews's correlation coefficient (MCC) metrics. This paper extends the handling of highly imbalanced credit card fraud data in [9] It experiment on 50:50, 10:90 and 1:99 distributions of fraud to legitimate cases reports that 10:90 distribution has the best performance (regarding the performance comparisons on the 1:99 set) as it is closest to the real distribution of frauds and legitimates. Stratified sampling is also applied in[10] The variables that Form the card usage profile and techniques used affect the Performance of credit card fraud detection systems. These Variables are derived from a combination of transaction and Past transaction history of a credit card. These variables fall Under five main variable types, namely all transactions Statistics, regional statistics, merchant type statistics, time based amount statistics and time-based number of transactions Statistics.[11]The study shows that innovative use of naive Bayesian (NB), C4.5, and back-propagation (BP) classifiers to process the same partitioned numerical data has the potential of getting better cost savings. An adaptive and robust model learning method that is highly adaptive to concept changes and is robust to noise is presented.[12]The results show that given a skewed distribution in the original data, artificially more balanced training data leads to better classifiers. It demonstrate how meta-learning can be used to combine different classifiers and maintain, and in some cases, improve the performance of the best classifier. Multiple algorithms for fraud detection are investigated in.[13]A meta-classification strategy is applied in improving credit card fraud detection.[14] The results of the analysis shows that even though the discriminative logistic regression algorithm has a lower asymptotic error, the generative naive Bayes classifier may also converge more quickly to its (higher) asymptotic error. There are a few cases reported in which logistic regression's performance underperformed that of naive Bayes, but this is observed primarily in particularly small datasets. Another comparative study on credit card fraud detection using Bayesian and neural networks is done.

III. PROPOSED SYSTEM:

Today current society is the usage of credit score playing cards for range on reasons. Similarly fraud in credit score card transactions has been developing in latest years. Each year, a big quantity of monetary losses are due to the unlawful credit score card transactions. Fraud may also arise in range of various forms and can be limited. Therefore there may be want to remedy the problems of fraud detection in credit score card. Additionally, with the improvement of latest technology criminals unearths new ways to devote fraud. To triumph over this hassle the proposed gadget for fraud detection in credit score card transactions will be designed the usage of ML approach as a way to offer investigator a small dependable fraud alert.

The proposed device will obtain following important objectives:

1. To teach the version the usage of feedbacks and delayed samples and sum up their probability to become aware of alert.
2. To enforce device mastering method to deal with idea go with the flow and sophistication imbalance issue.
3. To increase a mastering to rank method to increase alert precision.
4. To introduce overall performance degree the ones are taken into consideration in real-global FDS.

We advise a Fraud Detection System (FDS), which mainly specializes in statistics pushed version and mastering to rank method. It additionally specializes in alert remarks interplay that tests the manner latest supervised samples are provided.

This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105114

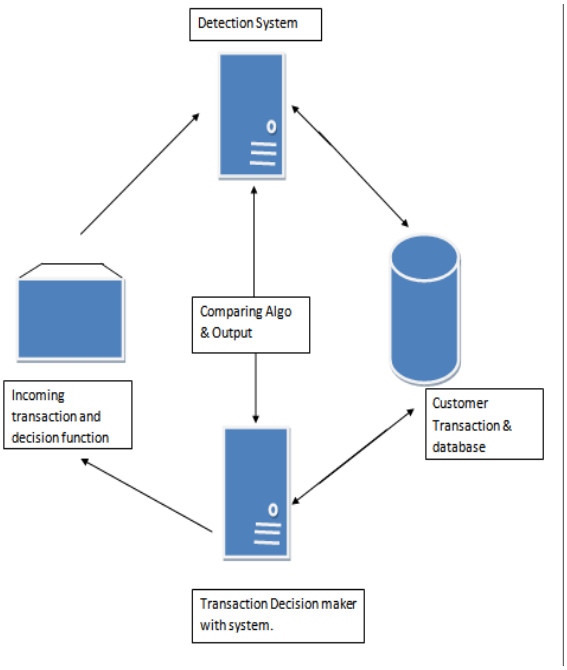

Fig.(a) – System Architecture.

We have obtained our dataset from Kaggle, a data analysis website which provides datasets. In 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. In this time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted.As shown in below fig.

```
Index(['Time', 'V1' , 'V2' , 'V3' , 'V4' , 'V5' , 'V6' ,
       'V7' , 'V8' , 'V9' , 'V10' , 'V11' , 'V12' , 'V13' ,
       'V14' , 'V15' , 'V16' , 'V17' , 'V18' , 'V19' , 'V20' ,
       'V21' , 'V22' , 'V23' , 'V24' , 'V25' , 'V26' , 'V27' ,
       'V28', 'Amount', 'Class'],
      dtype = 'object')
```

Fig.(b) – Explore dataset.

Here we've plot histogram of $1/10^{th}$ dataset. In this,Class 0 represents a valid transaction. Class 1 represents a fraudulent one.

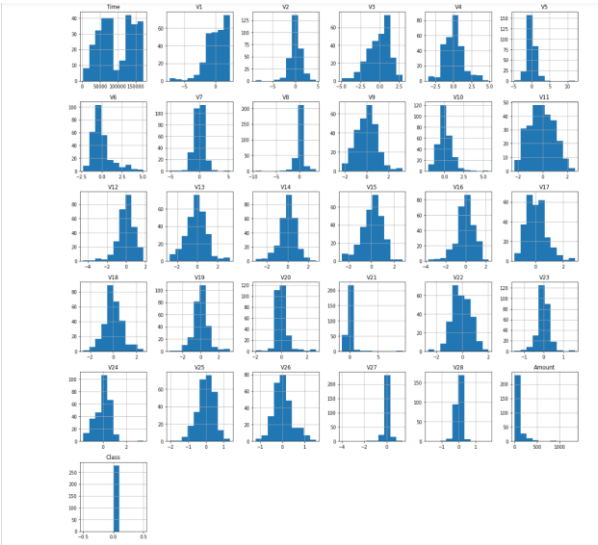

Fig.(c) – Histogram of each parameter in dataset.

This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105114

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

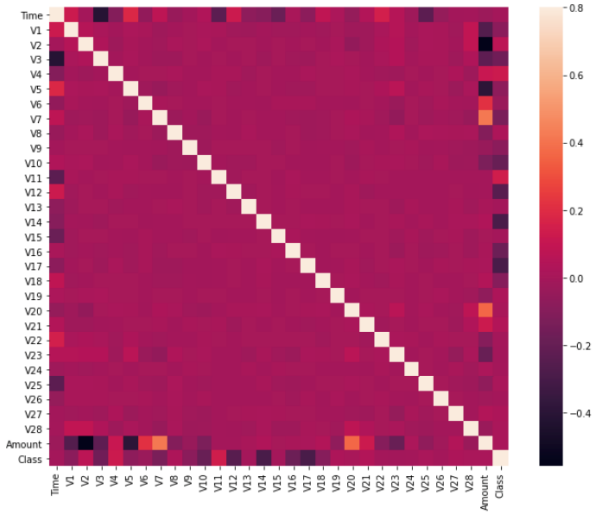

Fig.(d) – Heatmap

Now the dataset is formatted and processed. The time and amount column are standardized, and to ensure fairness of evaluation we have removed the class column. Here the data is processed by a set of algorithms from modules. The module diagram explains how these algorithms works together:
Next, this data is fitted 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 sklearn. We have ensemble module in the sklearn package which includes ensemble-based methods and functions for the classification, regression and outlier detection.
The platform that we've used is Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests.

1. Local Outlier Factor

The anomaly score of every sample is named the local Outlier factor. LOF calculates the local deviation of the density of a given sample in relation to its neighbors. It is known as local because the anomaly score depends on the object isolated the thing is with relation to the encircling neighborhood.

2. Isolation Forest Algorithm

The Isolation Forest isolates the observations by indiscriminately choosing a feature by randomly selecting a value and splitting it in between the at most maximum and minimum values of the chosen feature. The recursive partitioning method is denoted by a tree-like structure, the quantity of splitting required to isolate a sample is comparable to the trail length from the root node to the terminating node.
We have completed this fraudulent transactions detection activity by following three phases,

1. Data Exploration Steps:
a. Load dataset
b. Preprocess dataset
c. Perform graphing
d. Display dataset
2. Data Preprocessing Steps:
a. Load dataset
b. Remove Null values
c. Split dataset
d. Move to training phase
3. Data Classification Steps:
a. Train the dataset
b. Develop classifier
c. Isolation Forest
d. Perform Classification

This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105114

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 purpose, we have compared the outputs of these algorithms to determine their accuracy and precision.

IV. RESULTS

The code prints out the amount of false positives it detected and compares it with the particular values. This is wont to calculate the accuracy score and precision of the algorithms. The fraction of knowledge we used for faster testing is 10% of the whole dataset. The complete dataset is additionally used at the top and both the results are printed. These results along side the classification report for every algorithm is given within the output as follows, where class 0 means the transaction decided to be valid and 1 means it had been determined as a fraud transaction. This result matched against the category values to see for false positives.

```
Isolation Forest:5
0.99822695035461
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2815
           1       0.50      0.60      0.55         5

    accuracy                           1.00      2820
   macro avg       0.75      0.80      0.77      2820
weighted avg       1.00      1.00      1.00      2820

Local Outlier Factor:11
0.9960992907801418
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2815
           1       0.00      0.00      0.00         5

    accuracy                           1.00      2820
   macro avg       0.50      0.50      0.50      2820
weighted avg       1.00      1.00      1.00      2820
```

Results when 10% of the dataset is used:

Results with the complete dataset is used:

```
Isolation Forest:61
0.9978365725634842
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28148
           1       0.37      0.38      0.37        48

    accuracy                           1.00     28196
   macro avg       0.68      0.69      0.69     28196
weighted avg       1.00      1.00      1.00     28196

Local Outlier Factor:95
0.9966307277628033
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28148
           1       0.02      0.02      0.02        48

    accuracy                           1.00     28196
   macro avg       0.51      0.51      0.51     28196
weighted avg       1.00      1.00      1.00     28196
```

ACKNOWLEDGEMENT

No project is ever complete without the guidance of that expert who have already traded this past and hence become master of it and as a result, our leader. So we wish to take this chance to require all those individuals who have helped us in visualizing this project.

We express our deep gratitude to our guide Mrs. H. A. Shinde for providing timely assistant to our query and guidance that she gave due to her experience during this field for past many years. She had indeed been a lighthouse for us in this difficult journey.

This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105114

We also wish to take this opportunity to thank our project co-ordinate Prof. A. N. Gedam for his guidance in selecting this project and also for providing us all this details on proper presentation of this project.

We extend our sincerity appreciation to all our teachers from A.I.S.S.M.S Polytechnic Pune for their valuable inside and tip during the designing of the project. Their contributions are valuable in numerous ways in which we discover it difficult to acknowledge all of them individually.

We also are grateful to our principal Prof. S. K.Giram and HOD Mr.V. N. Kukre for extending his help directly and indirectly through various channels in our project work.

In the last we would express gratitude to all our fellow classmates for their encouragement. Without them the timely completion of this project would not have been possible.

V. CONCLUSION

Credit card fraud cases are increasing day by day and it is one of the major concerns in financial service sectors. This happens when are no legitimate safety efforts are thought about. In this paper an endeavor is made to distinguish the quantity of fake exchanges in a specific data-set by utilizing AI calculations, for example, nearby anomaly factor and segregation timberland strategy. Just a piece of data-set was utilized to accelerate the computational interaction.
This venture has likewise clarified exhaustively, how AI can be applied to improve brings about extortion discovery alongside the calculation, clarification its execution and results. While the calculation comes to more than 99.6% exactness, its accuracy stays just at 28% when a 10th of the informational index is mulled over.
Notwithstanding, when the whole data-set is taken care of into the calculation, the accuracy ascends to 33%. This high level of precision is to be required because of the colossal awkwardness between the quantity of legitimate and number of authentic transactions.

VI. REFERENCES

[1] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare . Credit card fraud detection using Machine Learning Techniques:A Comparitive Analysis.
[2] S P Maniraj, Aditya Sani, Swarna Deep Sarkar, Shadab Ahmed.Credit Card Fraud Detection using Machine Learning and Data Science. Vol.8, Issue 09, September-2019.
[3] Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B., (2002). Credit card fraud detection using Bayesian and Neural networks. Proceeding International NAISO Congress on Neuro Fuzzy Technologies.
[4] Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science And Technology, Vol. 6, No. 3, pp. 311 – 322
[5] RamaKalyani, K. and UmaDevi, D., (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm, International Journal of Scientific & Engineering Research, Vol. 3, Issue 7, pp. 1 – 6, ISSN 2229-5518
[6] Meshram, P. L., and Bhanarkar, P., (2012). Credit and ATM Card Fraud Detection Using Genetic Approach, International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 10, pp. 1 – 5, ISSN: 2278-0181
[7]Sahin, Y. and Duman, E., (2011). Detecting credit card fraud by ANN and logistic regression. In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on (pp. 315- 319). IEEE,40(15), 5916-5923.
[8] Fahmi, M., Hamdy, A. and Nagati, K., (2016). Data Mining Techniques for Credit Card Fraud Detection: Empirical Study, In Sustainable Vital Technologies in Engineering and Informatics BUE ACE1, pp. 1 – 9, Elsevier LLtd.
[9] Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications.
[10]Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). Cost sensitive credit card fraud detection using Bayes minimum risk. In Machine Learning and Applications (ICMLA), 2013 12th International Conference on (Vol. 1, pp. 333-338). IEEE. [11] Chu, F., Wang, Y., & Zaniolo, C. (2004). An adaptive learning approach for noisy data streams. In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on (pp. 351-354). IEEE.
[12] Wheeler, R., and Aitken, S. (2000). Multiple algorithms for Fraud detection. Knowledge-Based Systems, 13(2), 93-99. Elsevier.
[13] Pun, J., and Lawryshyn, Y. (2012). Improving credit card fraud detection using a meta-classification strategy. International Journal of Computer Applications, 56(10).
[14]Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

This work is licensed under a Creative Commons Attribution 4.0 International License