International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

[1] G Divyavani, [2] Kodaboina Ashwini, [3] Dachepalli Amukta Malyada

[1] Associate Professor, Department of Computer and Science Engineering, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India.

[1]Divyavanigu1@gmail.com

[2,3]Students, Department of Computer and Science Engineering, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India.

[2]ashwiniyadavkodaboina@gmail.com , [3] amuktamalyada23102001@gmail.com

Abstract:

Credit card is the commonly used payment mode in the recent years. As thetechnology is developing, the number of fraud cases is also increasing andfinally poses the need to develop a fraud detection algorithm to accuratelyfind and eradicate the fraudulent activities. This research work proposesdifferent machine learning based classification algorithms such as logisticregression, random forest, and Naive Bayes for handling the heavilyimbalanced dataset. Finally, this research work will calculate the accuracy,precision, recall, f1 score, confusion matrix, and Roc-auc score.

1. INTRODUCTION

The primary objective of this research work is to identify the fraudulenttransactions using credit cards. To accomplish this, it is required to classifythe fraudulent and non-fraudulent transactions. The primary goal is to makea fraud detection algorithm, which finds the fraud transactions with less timeand high accuracy by using machine learning based classification algorithms.As technology is advancing rapidly, the payment by cash is reduced andonline payment gets increased, this paves way for the fraudsters to makeanonymous transactions.

In some modes of online payments, only card number, expiration date, andcvv are required and that data may be lost without our presence, in somecases we don't even know our data is being stolen. The purchases that doneover the internet where fraudsters use phishing techniques to grab thedetails still, we do not know that our data has leaked. To do fraud he justneeds card details for some purchases and the user may not know whetherhis/her credit card information was leaked. The card details should be keptprivate. But sometimes it is not in our hands. Due to phishing sites theinformation may be leaked, Sometimes the card itself may be lost or may bestolen. The best way to find whether a transaction is fraud or not we need tofind the spending pattern of the customer by using

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

existing data and useMachine learning to find whether a is genuine or not.

2. LITERATURE SURVEY

Fraudulent activities are causing major loss, which motivated researchers tofind a solution that would detect and prevent frauds. Several methods havealready been proposed and tested. Some of them are briefly reviewed below.Classical algorithms such as Gradient Boosting (GB), Support VectorMachines (SVM), Decision Tree (DT), LR and RF proven useful. GB, LR, RD,SVM and a combination of certain classifiers was used, which led to highrecall of over 91% on a European dataset. High precision and recall wereachieved only after balancing the dataset by under sampling the data. Inpaper [6], European dataset was also used, and comparison was madebetween the models based on LR, DT and RF. Among the three models, RFproved to be the best, with accuracy of 95.5%, followed by DT with 94.3%and LR with accuracy of 90%.

k-Nearest neighbors (KNN) and outlier detection techniques can also beefficient in fraud detection. They are proven useful in minimizing false alarmrates and increasing fraud detection rate. KNN algorithm also performed wellhere the authors tested and compared it with other classical algorithms.

Unlike so far mentioned papers, a comparison was made between someclassical algorithms and deep learning techniques. All of the testedtechniques achieved accuracy of approximately 80%, set side by sidefollowing algorithms: RF, GB, LR, SVM, DT, KNN, NB, XGBoost (XGB), MLPand stacking classifier (a combination of multiple machine learningclassifiers), while using European dataset. As a result of thorough datapreprocessing, all of the algorithms accomplished high accuracy of over90%. Stacking classifier was most successful with accuracy of 95% andrecall value of 95%.

a neural network was tested on the European dataset. Experiment includedback propagation neural network that was optimized with Whale algorithm.Neural network consisted of 2 input layers, 20 hidden and 2 output layers.Due to optimization algorithm, they achieved exceptional results on 500 testsamples: 96.40% accuracy and 97.83% recall.used neural networks, in order to demonstrate improvement in results whenensemble techniques are used. In paper [15] three datasets were used forcomparison between Auto-encoder and Restricted Boltzmann Machinealgorithms, which led to the conclusion that algorithms like MLP can besuitable for credit card fraud detection.Numerous papers are focused on detecting fraudulent transactions usingdeep neural networks. However, these models are computationallyexpensive and perform better on larger datasets. This approach may lead togreat results, as we saw in some papers, but what if same results, or evenbetter, can be achieved with fewer amounts of resources? Our main goal isto show that different machine learning algorithms can give decent resultswith appropriate preprocessing. Authors of most of the mentioned paperused under sampling technique, and that was a motivation for using adifferent approach – oversampling technique. Considering given facts,authors of this paper decided to compare the suitability of LR, RF, NB andMLP for credit card fraud detection. In order to achieve that, an experimentwas conducted.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

3. EXISTING SYSTEM

Fraud in any way is a criminal activity and is an offence; credit card fraud isstealing money. There are many studies in which they tried to find whether atransaction is fraud or not. Still having many challenges and tries toovercome those problems Firstly, many used Data Mining Techniques to findfraudulent transactions by using some Traditional approach, which is notconventional and these days fraudsters are so smart that they can do.

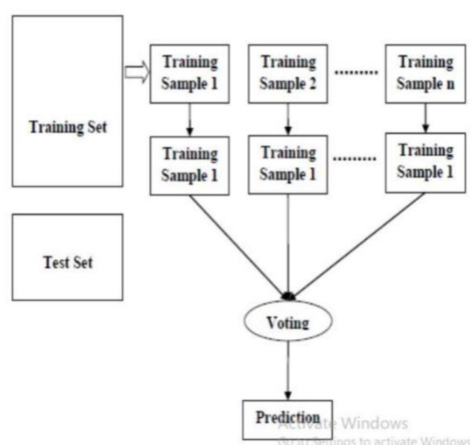
4. PROPOSED SYSTEM

Numerous papers are focused on detecting fraudulent transactions usingdeep neural networks. However, these models are computationallyexpensive and perform better on larger datasets. This approach may lead togreat results, as we saw in some papers, but what if same results, or evenbetter, can be achieved with fewer amounts of resources? Our main goal isto show that different machine learning algorithms can give decent resultswith appropriate preprocessing. authors of most of the mentioned paperused under sampling technique, and that was a motivation for using adifferent approach – oversampling technique. Considering given facts,authors of this paper decided to compare the suitability of LR, RF, NB andMLP for credit card fraud detection. In order to achieve that, an experimentwas conducted.

The detailed architecture diagram for the credit card fraud detection systemincludes many steps from gathering dataset to deploying model andperforming analysis based on results. In this model we take the Kagglecredit card fraud dataset and pre-processing is to be done for the dataset.Now to prepare the model we have to split the data into the training dataand the testing data. We use the training data to prepare the Random Forestand the Adaboost models. Then we develop both the models. Finally, theaccuracy, precision, recall, and F1-score is calculated for bot the models.Finally the comparison of the credit card fraud transactions more accurately.

5. ALGORITHMS USED

Random Forest Algorithm The Random Forest algorithm [Figure. 5]is one ofthe widely used supervised learning algorithms. This can be used for bothregression and classification purposes. But, this algorithm is mainly used forclassification problems. Generally, a forest is made up of trees and similarly,the Random Forest algorithm creates the decision trees on the sample dataand gets the prediction from each of the sample data. Then Random Forestalgorithm is an ensemble method. This algorithm is better than the singledecision trees because it reduces the over-fitting by averaging the result.



Steps for Random Forest Algorithm

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

1. Take the Kaggle credit card fraud dataset that is trained and randomlyselect some of the sample data.

2. Using the randomly created sample data now creates the Decision Treesthat are used to classify the cases into the fraud and non-fraud cases.

3. The Decision Trees are formed by splitting the nodes, the nodes whichhave the highest Information gain make it as the root node and classify thefraud and non-fraud cases.

4. Now the majority vote is performed and the decision Trees may result in 0as output which includes that these are the non-fraud cases.

5. Finally, we find the accuracy, precision, recall, and F1 -score for both thefraud and non-fraud cases.

6. IMPLEMENTATION

6.1 Dataset:

In this research the Credit Card Fraud Detection dataset was used, whichcan be downloaded from Kaggle . This dataset contains transactions,occurred in two days, made in September 2013 by European cardholders.The dataset contains 31 numerical features. Since some of the inputvariables contains financial information, the PCA transformation of theseinput variables were performed in order to keep these data anonymous.Three of the given features weren't transformed. Feature "Time" shows thetime between first transaction and the every other transaction in thedataset. Feature "Amount" is the amount of the transactions made by creditcard. Feature "Class" represents the label, and takes only 2 values: value 1in case of fraud transaction and 0 otherwise. Dataset contains 284,807transactions where 492 transactions were frauds and the rest were genuine.Considering the numbers, we can see that this dataset is highly imbalanced,where only 0.173% of transactions are labeled as frauds. Since distributionratio of classes plays an important role in model accuracy and precision,preprocessing of the data is crucial.

6.2 Preprocessing:

Feature selection is a fundamental technique, which selects the variablesthat are most relevant in the given dataset. Carefully choosing appropriatefeatures and removing the less important one can reduce overfitting,improve accuracy and reduce training time. Visualization techniques can behelpful in that process. Feature selector tool by Will Koehrsen was used inthis experiment for that purpose. By using this tool it has been determinedwhich features are the most important. Furthermore, features that do notcontribute to the cumulative importance of 95% were removed. After thefeature selection technique, 27 features were selected for additionalexperiment. Machine learning algorithms have trouble learning whenclassification categories are not approximately equally distributed.Considering given data is highly imbalanced, it is necessary to perform somekind of balancing, so that model can be efficiently trained. Frequently usedmethods for adjusting the class distribution include undersampling themajority class, oversampling the minority class, or combination of those two.

Synthetic Minority Oversampling Technique (SMOTE) is a popularoversampling method that has proven useful when used on imbalanceddataset.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

SMOTE was proposed method to improve random oversampling.Preprocessing:

6.3 Feature selection is a fundamental technique, which selects the variablesthat are most relevant in the given dataset. Carefully choosing appropriatefeatures and removing the less important one can reduce overfitting,improve accuracy and reduce training time. Visualization techniques can behelpful in that process. Feature selector tool by Will Koehrsen was used inthis experiment for that purpose. By using this tool it has been determinedwhich features are the most important. Furthermore, features that do notcontribute to the cumulative importance of 95% were removed. After thefeature selection technique, 27 features were selected for additionalexperiment. Machine learning algorithms have trouble learning whenclassification categories are not approximately equally distributed.

Considering given data is highly imbalanced, it is necessary to perform somekind of balancing, so that model can be efficiently trained. Frequently usedmethods for adjusting the class distribution include under sampling the majority class, oversampling the minority class, or combination of those two.Synthetic Minority Oversampling Technique (SMOTE) is a popular

Over sampling method that has proven useful when used on imbalanceddataset. SMOTE was proposed method to improve random oversampling.The experiment system environment is Windows 10 operating system, andthe software operating environment is Spyder, scientific python development environment, which is part of the Anaconda platform. Used libraries include numpy, pandas, matplotlib, sklearn and imblearn. Previously mentionedalgorithms used in the experiment are described in the following section.

7. CONCLUSION

Credit card frauds represent a very serious business problem. These fraud scan lead to huge losses, both business and personal. Because of that,companies invest more and more money in developing new ideas and waysthat will help to detect and prevent frauds. The main goal of this paper was to compare certain machine learning algorithms for detection of fraudulenttransactions. Hence, comparison was made and it was established thatRandom Forest algorithm gives the best results i.e. best classifies whethertransactions are fraud or not. This was established using different metrics,such as recall, accuracy and precision. For this kind of problem, it isimportant to have recall with high value. Feature selection and balancing ofthe dataset have shown to be extremely important in achieving significantresults.

REFERENCES

[1] Global Facts (2019). Topic: Startups worldwide. [online] Available at:https://www.statista.com/topics/4733/startups-worldwide/ [Accessed 10Jan. 2019].

[2] Legal Dictionary (2019). Fraud - Definition, Meaning, Types,Examples of fraudulent activity. [online] Available at:https://legaldictionary.net/fraud/ [Accessed 15 Jan. 2019].

[3] EuropeanCentral Bank (2018). Fifth report on card fraud, September 2018. [online].Available at:https://www.ecb.europa.eu/pub/cardfraud/html/ecb

.cardfraudreport201809.en.html#toc1 [Accessed 21 Jan. 2019].

[4] En.wikipedia.org. (2019).Credit card fraud. [online] Available at:https://en.wikipedia.org/wiki/Credit_card_fraud [Accessed 24 Jan. 2019].

[5] A. Mishra, C. Ghorpade, “Credit Card Fraud Detection on the SkewedData Using Various Classification and Ensemble Techniques” 2018 IEEEInternationalStudents' Conference on Electrical, Electronics and ComputerScience (SCEECS) pp. 1-5. IEEE.

[6] S. V. S. S. Lakshmi, S. D. Kavilla“Machine Learning For Credit Card Fraud Detection System”, unpublished

[7]N.Malini, Dr. M. Pushpa, “Analysis on Credit Card Fraud IdentificationTechniques based on KNN and Outlier Detection“, Advances in Electrical,Electronics, Information, Communication and BioInformatics (AEEICB), 2017Third International Conference on pp. 255- 258. IEEE.

[8] Mrs. C. Navamani,M. Phil, S. Krishnan, “Credit Card Nearest Neighbor Based Outlier DetectionTechniques”

[9] J. O. Awoyemi, A. O. Adentumbi, S. A. Oluwadare, “Creditcard fraud detection using Machine Learning Techniques: A Comparativenalysis”, Computing Networking and Informatics (ICCNI), 2017International Conference on pp. 1-9. IEEE.

[10] Z. Kazemi, H. Zarrabi,“Using deep networks for fraud detection in the credit card transactions”,Knowledge-Based Engineering and Innovation (KBEI), 2017 IEEE 4thInternational Conference on pp. 630-633. IEEE.

[11] S. Dhankhad, B. Far, E.A. Mohammed, “Supervised Machine Learning Algorithms for Credit CardFraudulent Transaction Detection: A Comparative Study”, 2018 IEEEInternational Conference on Information Reuse and Integration (IRI) pp.122-125. IEEE.

[12] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai, S. Pan, “Creditcard fraud detection based on whale algorithm optimized BP neuralnetwork”, 2018 13th International Conference on Computer Science &Education (ICCSE) pp. 1-4. IEEE.

[13] N. Kalaiselvi, S. Rajalakshmi, J.Padmavathi, “Credit card fraud detection using learning to rank approach”,2018 Internat2018 Predicted 0 1 Actual 0 56861 3 1 18 80 Predicted 0 1Actual 0 56843 21 1 18 80 International Conference on Computation ofPower, Energy, Information and Communication (ICCPEIC) ional conferenceon computation of power, energy, Information and Communication(ICCPEIC) pp. 191- 196. IEEE

[14] F. Ghobadi, M. Rohani, “Cost SensitiveModeling of Credit Card Fraud using Neural Network strategy”, 2016 SignalProcessing and Intelligent Systems (ICSPIS), International Conference of pp.1-5. IEEE.

[15] A. Pumsirirat, L. Yan, “Credit Card Fraud Detection usingDeep Learning based on Auto-Encoder and Restricted Boltzmann Machine”,2018 International journal of advanced computer science and applications,9(1), pp. 18-25

[16] Learning – Towards Data Science. [online] Available at:https://towardsdatascience.com/deep-learning-vs-classical-machinelearning-9a42c6d48aa [Accessed 19 Jan. 2019]. [17] Kaggle.com. (2019). CreditCard Fraud Detection. [online] Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud [Accessed 10 Jan. 2019].

[18] Github (2019). Featureselector. [online] Available at: https://github.com/WillKoehrsen/feature-selector [Accessed 18 Jan. 2019].

[19] Garćıa, Salvador and NiteshV.Chawla. "SMOTE for Learning from Imbalanced Data : Progress andChallenges, Marking the 15-year Anniversary." (2018), Journal of ArtificialIntelligence Research, 61, pp. 863-905.

[20] J. Wang, M. Xu, H. Wang and J.Zhang, "Classification of Imbalanced Data by Using the SMOTE Algorithmand Locally Linear Embedding", Signal Processing, 2006 8th InternationalConference on (Vol. 3). IEEE. 2006 8th international Conference on SignalProcessing, Beijing, 2006

[21] Deeplearningbook.org. (2019). DeepLearning. [online] Available at: https://www.deeplearningbook.org/[Accessed 11 Jan. 2019].