IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Credit Card Fraud Detection Using Data Mining

Reshma Farooqui*, Sifatullah Siddiqi#

**M.Tech, CSE, Integral University, Lucknow, Uttar Pradesh, India*

#Assistant Professor, dept. of CSE, Integral University, Lucknow, Uttar Pradesh, India

Abstract—**Now dayscredit card fraud is a serious problem in financial services. Billions of dollars are lost due to credit card fraud every year. There is a lack of research studies on analyzing real-world credit card data owing to confidentiality issues. In this paper, machine learning algorithms are used to detect credit card fraud. Standard models are first used. Then, hybrid methods which use AdaBoost and majority voting methods are applied. To evaluate the model efficacy, a publicly available credit card data set is used. Then, a real-world credit card data set from a financial institution is analyzed. In addition, noise is added to the data samples to further assess the robustness of the algorithms. The experimental results positively indicate that the majority voting method achieves good accuracy rates in detecting fraud cases in credit cards.**

Keywords:credit card, AdaBoost, detecting fraud, accuracy rates, robustness.

1. INTRODUCTION

As per Global Payments Report 2015, Mastercard is the most noteworthy utilized installment technique around the world in 2014 contrasted with different strategies, for example, e-wallet and Bank Transfer [1]. The tremendous value-based administrations are frequently looked at by digital crooks to direct false exercises utilizing the Mastercard administrations. Visa extortion is characterized as the unapproved use of card, surprising exchange conduct, or exchanges on an idle card [2]. By and large, there are three classifications of charge card extortion specifically, ordinary cheats (for example taken, phony and fake), online cheats (for example bogus/counterfeit dealer destinations), and shipper related fakes (for example shipper arrangement and triangulation) [3]. In the past two or three the years, Mastercard breaks have been moving alarmingly. As per Nilson Report, the worldwide Mastercard extortion misfortunes came to $16.31 billion out of 2014 and it is assessed that it will surpass $35 billion of every 2020 [4]. In this manner, it is important to foster Visa extortion identification strategies as the counter measure to battle criminal operations. By and large, Mastercard extortion identification has been known as the method involved with distinguishing whether exchanges are certifiable or fake. As the information mining and AI methods are immeasurably used to counter digital lawbreaker cases, researchers frequently embraced those ways to deal with study and distinguish charge card extortion exercises. Information mining is known as the most common way of acquiring fascinating, novel and canny examples as well as finding reasonable, illustrative and prescient models from huge size of information assortments [5, 6]. The capacity of information mining methods to remove productive data from enormous size of information utilizing factual and numerical strategies would help Mastercard extortion recognition in light of separating the attributes of normal and dubious Visa exchanges. While information mining zeroed in on finding significant knowledge, AI is established in learning the knowledge and fostering its own model with the end goal of grouping, bunching or so on. The utilization of AI procedures spreads generally all through PC sciences spaces, for example, spam sifting, web looking, promotion position, recommender frameworks, credit scoring, drug plan, misrepresentation location, stock exchanging, and numerous different applications. AI classifiers work by building a model from model information sources and utilizing that to settle on forecasts or choices, as opposed to adhering to rigorously static program guidelines. There are various kinds of AI approaches accessible with the expectations to tackle heterogeneous issues. Because of the idea of this review which was centered around order, the conversation that follows depends on this subject. AI order alludes to the method involved with figuring out how to appoint cases to predefined classes. Officially, there are a few sorts of learning, for example, directed, semi-managed, solo, support, transduction and figuring out how to learn [7]. As the interest of this review was to lead managed based AI grouping, the conversations about the other strategies are disposed of from additional elaboration. In most grouping review, supervisedbased learning is inclined toward more than different strategies because of the capacity to control the classes of the occurrences with the mediations of human. In regulated learning, the classes of the occurrences would be marked preceding taking care of into classifiers. Then, at that point, by utilizing specific assessment measurements, the exhibitions of the classifiers could be estimated.

In this paper section I contains the introduction, section II contains the literature review details, section III contains the details about methodologies, section IV shows architecture details, V describe the result and section VII provide conclusion of this paper.

2. LITERATURE REVIEW

A formative information mining and AI are famous strategies to study and battle the charge card misrepresentation cases. There is countless examinations that took advantage of the strength of information mining and AI to forestall the charge card fake exercises. In light of Self-Organizing Map and

Neural Network, the investigation of [8] got Receiver Operating Curve (ROC) more than 95.00% of misrepresentation cases without phony problems rate. The Hidden Markov Model (HMM) additionally has been applied in charge card misrepresentation recognition with low level of deception rates [9]. Nonetheless, change cycle of various states and ascertaining the likelihood in HMM are exorbitant and escalated. Besides, instead of utilizing single classifiers, a portion of the Mastercard extortion identification concentrates on utilized metalearning students in view of directed learning. Stolfo et al. explored Mastercard misrepresentation identification framework utilizing four sorts of calculations to be specific Iterative Dichotomiser 3 (ID3), Classification and Regression Tree (CART), Ripper and Bayes as base students and tried with heterogeneous information circulations [10]. In view of half/half dissemination of occurrences (misrepresentation and non-extortion), the investigation discovered that metalearning involving Bayes as a base student got a higher genuine positive rate contrasted with other meta students. In any case, despite the fact that the dissemination of half/half yields great outcomes, it doesn't reflect certifiable conditions where veritable Mastercard exchanges are very higher than non-authentic exchanges. Scientists have likewise tried different sorts of meta learning classifiers, for example, Adaboost, Logitboost, Bagging and Dagging and yielded intriguing results [11].

Through our writing review, Bayesian Network is one of the classifier types that have been generally applied to recognize misrepresentation in the charge card industry. Maes et al inspected the genuine positive and misleading positive created by Bayesian Belief Network and Artificial Neural Network on ordering Mastercard misrepresentation occurrences. The investigation discovered that Bayesian organization performed around 8% higher than Artificial Neural Network and asserted that the previous' classifier handling time is more limited than the last [12]. Instead of examining utilizing customary arrangement strategies, the examination by [13] started to perform cost delicate Visa misrepresentation recognition in light of Bayes Minimum Risk strategy. The review estimated the exhibitions of Logistic Regression (LR), C4.5 and Random Forest (RF). The review showed that changing the probabilities of Bayes Minimum Risk classifier on RF order yielded reliably improved results than LR and C4.5. All through our perception and examination of past investigations, Bayesian Network classifiers have become one of the well known classifier types that are generally used to characterize Visa misrepresentation information. Hence, this review endeavored to research the characterization by a few Bayesian classifiers, for example, K2, Tree Augmented Naïve Bayes (TAN), and Naïve Bayes. In addition, this concentrate likewise estimated the exhibitions of Logistics Regression and J48 in light of the proposed philosophy. A short conversation about Bayesian Network Classifier and proposed classifiers are expressed underneath.

Essentially, the objective of misrepresentation recognition ought to be matched to an information mining technique. Information, as a rule, mining procedures can be partitioned into two kinds as far as whether the fake occasion is recognized in the past information: managed and solo [3]. Ngai et al. [4] have shown that grouping as a managed strategy is the most often involved information mining application in monetary misrepresentation identification. Regardless, a classifier ought to characterize every client into one of the two classes of typical or false clients. With a complete view, we observe that we are confronted with a specific kind of characterization issue. Taking into account a bank data set with a large number of exchanges in a day, just exactly couple of exchanges might be dubious in a month. All in all, we are confronted with a super imbalanced data set. The issue with an awkwardness informational index is the slanted dissemination of the information that makes the learning calculations ineffectual, particularly in foreseeing the minority classes. In this segment, we audit the writing in which issues with imbalanced information arrangement and charge card extortion discovery methods are. Albeit the absence of freely accessible data sets has restricted the distributions on monetary extortion identification, in this part we will survey a portion of the accessible ones.

3. METHODOLOGIES

- **Decision Tree (DT)**

The presentation of data in form of a tree structure is useful for ease of interpretation by users. The Decision Tree (DT) is a collection of nodes that creates decision on features connected to certain classes. Every node represents a splitting rule for a feature. New nodes are established until the stopping criterion is met. The class label is determined based on the majority of samples that belong to a particular leaf. The Random Tree (RT) operates as a DT operator, with the exception that in each split, only a random subset of features is available. It learns from both nominal and numerical data samples. The subset size is defined using a subset ratio parameter.

The Random Forest (RF) creates an ensemble of random trees. The user sets the number of trees. The resulting model employs voting of all created trees to determine the final classification outcome. The Gradient Boosted Tree (GBT) is an ensemble of classification or regression models. It uses forward-learning ensemble models, which obtain predictive results using gradually improved estimations. Boosting helps improve the tree accuracy.

- **Naïve Bayes (NB)**

Naïve Bayes (NB) uses the Bayes' theorem with strong or naïve independence assumptions for classification. Certain features of a class are assumed to be not correlated to others. It requires only a small training data set for estimating the means and variances is needed for classification.

- **The Random Forest (RF)**

The Random Forest (RF) creates an ensemble of random trees. The user sets the number of trees. The resulting model employs voting of all created trees to determine the final classification outcome. The Gradient Boosted Tree (GBT) is an ensemble of classification or regression models. It uses forward-learning ensemble models, which obtain predictive results using gradually improved estimations. Boosting helps improve the tree accuracy. The Decision Stump (DS) generates a decision tree with a single split only. It can be used in classifying uneven data sets.

- **AdaBoost and Majority Voting**

Adaptive Boosting or Ada Boost is used in conjunction with different types of algorithms to improve their performance. The outputs are combined by using a weighted sum, which represents the combined output of the boosted classifier. AdaBoost tweaks weak learners in favor of misclassified data samples. It is, however, sensitive to noise and outliers. As long as the classifier performance is not random, AdaBoost is able to improve the individual results from different algorithms.AdaBoost helps improve the fraud detection rates, with a noticeable difference for NB, DT, RT, which produce a perfect accuracy rate. The most significant improvement is achieved by LIR. Majority voting is frequently used in data classification, which involves a combined model with at least two algorithms. Each algorithm makes its own prediction for every test sample. The final output is for the one that receives the majority of the votes. The majority voting method achieves good accuracy rates in detecting fraud cases in credit cards.

- **Machine Learning Algorithms**

Machine learning is the science of designing and applying algorithms that are able to learn things from past cases. It uses complex algorithms that iterate over large data sets and analyze the patterns in data. The algorithm facilitates the machines to respond to different situations for which they have not been explicitly programmed. It is used in spam detection, image recognition, product recommendation, predictive analytics etc. Significant reduction of human effort is the main aim of data scientists in implementing ML. Even with modern analytics tools, it takes a lot of time for humans to read, collect, categorize and analyze the data. ML teaches machines to identify and gauge the importance of patterns in place of humans. Particularly for use cases where data must be analyzed and acted upon in a short amount of time, having the support of machines allows humans to be more efficient and act with confidence.

4. SYSTEM ARCHITECTURE

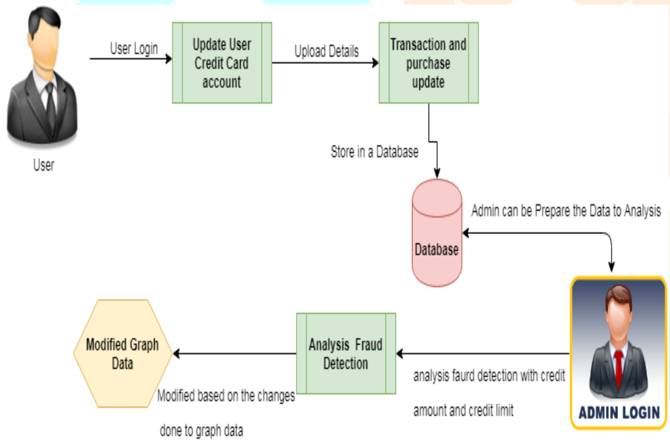
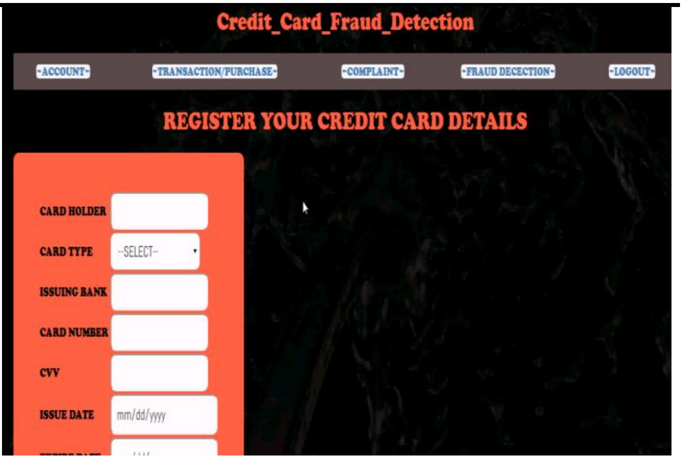

User Login
Update User Credit Card account
Upload Details
Transaction and purchase update
User
Store in a Database
Database
Admin can be Prepare the Data to Analysis
Modified Graph Data
Analysis Fraud Detection
analysis faurd detection with credit amount and credit limit
ADMIN LOGIN
Modified based on the changes done to graph data

Figure 1 System Architecture

5. RESULTS

In this paper outcome part, we step up and study the machine learning algorithms are used for detecting credit card fraud. The algorithms range from standard neural networks to deep learning models. They are evaluated using both benchmark and real-world credit card data sets. In addition, the AdaBoost and majority voting methods are applied for forming hybrid models. To further evaluate the robustness and reliability of the models, noise is added to the real-world data set. The key contribution of this paper is the evaluation of a variety of machine learning models with a real-world credit card data set for fraud detection. While other researchers have used various methods on publicly available data sets, the data set used in this paper are extracted from actual credit card transaction information over three months.



Credit_Card_Fraud_Detection

REGISTER YOUR CREDIT CARD DETAILS

CARD HOLDER

CARD TYPE --SELECT--

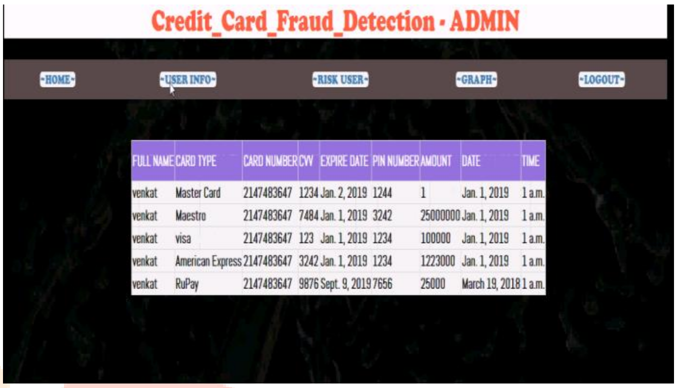
ISSUING BANK

CARD NUMBER

CVV

ISSUE DATE mm/dd/yyyy

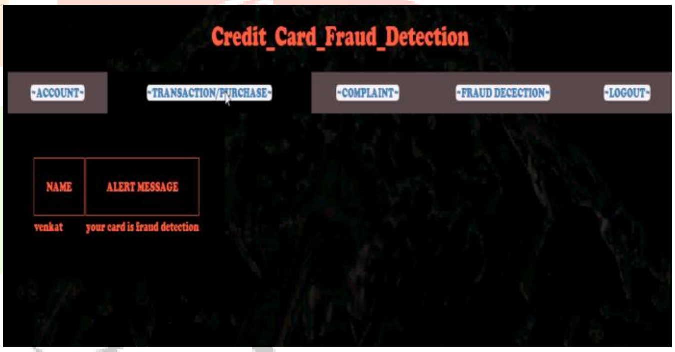
Figure 2: Register credit cards details



Credit_Card_Fraud_Detection - ADMIN

FULL NAME	CARD TYPE	CARD NUMBER	CVV	EXPIRE DATE	PIN NUMBER	AMOUNT	DATE	TIME
venkat	Master Card	2147483647	1234	Jan. 2, 2019	1244	1	Jan. 1, 2019	1 a.m.
venkat	Maestro	2147483647	7484	Jan. 1, 2019	3242	25000000	Jan. 1, 2019	1 a.m.
venkat	visa	2147483647	123	Jan. 1, 2019	1234	100000	Jan. 1, 2019	1 a.m.
venkat	American Express	2147483647	3242	Jan. 1, 2019	1234	1223000	Jan. 1, 2019	1 a.m.
venkat	RuPay	2147483647	9876	Sept. 9, 2019	7656	25000	March 19, 2018	1 a.m.

Figure 3: credit cards details



Credit_Card_Fraud_Detection

NAME	ALERT MESSAGE
venkat	your card is fraud detection

Figure 4: credit cards fraud alert message

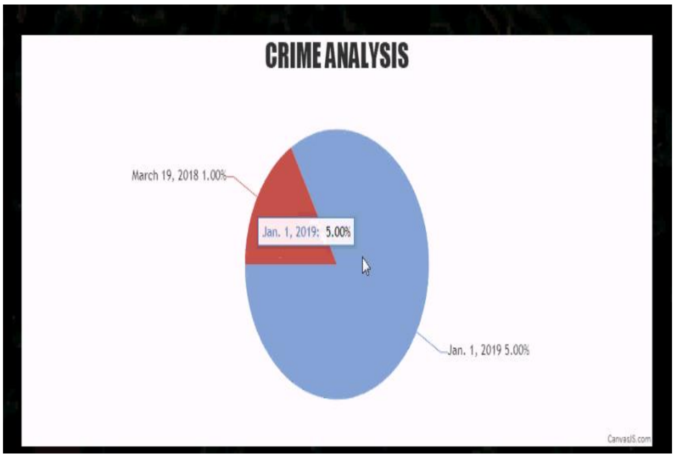


Figure 5: Crime analysis pie chart

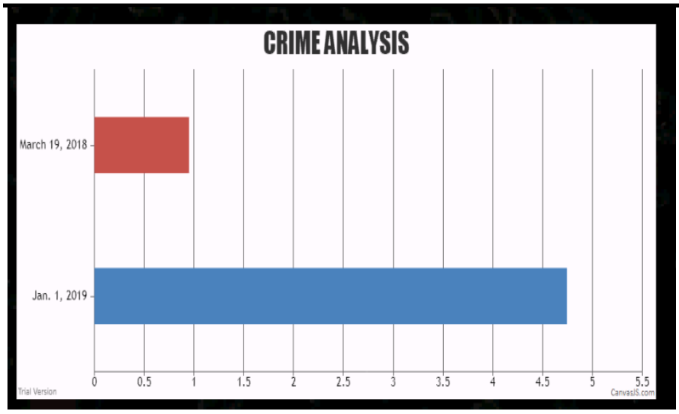


Figure 6: Crime analysis bar chart

6. CONCLUSION

In this research paper, we have considered the novel concept of study on credit card fraud detection using machine learning algorithms has been presented in this paper. A number of standard models which include NB, SVM, and DL have been used in the empirical evaluation. A publicly available credit card data set has been used for evaluation using individual (standard) models and hybrid models using AdaBoost and majority voting combination methods. The MCC metric has been adopted as a performance measure, as it takes into account the true and false positive and negative predicted outcomes. The best MCCscore is 0.823, achieved using majority voting. A real credit card data set from a financial institution has also been used for evaluation. The same individual and hybrid models have been employed. A perfect MCC score of 1 has been achieved using AdaBoost and majority voting methods. To further evaluate the hybrid models, noise from 10% to 30% has been added into the data samples. The majority voting method has yielded the best MCC score of 0.942 for 30% noise added to the data set. This shows that the majority voting method offers robust performance in the presence of noise.

REFERENCE

[1] WorldPay. (2015, Nov). Global payments report preview: your definitive guide to the world of online payments. Retrieved September 28, 2016, from http://offers.worldpayglobal.com/rs/850-JOA 856/images/GlobalPaymentsReportNov2015.pdf.

[2] Federal Trade Commision. (2008). consumer sentinel network - data book for January - December 2008. Retrieved Oct 20, 2016. From https://www.ftc.gov/.

[3] Bhatla, T.P., Prabhu, V., and Dua, A. (2003). understanding credit card frauds. Crads Business Review# 2003-1, Tata Consultancy Services.

[4] The Nilson Report. (2015). Global fraud losses reach $16.31 Billion. Edition: July 2015, Issue 1068.

[5] Y. Sahin and E. Duman, “Detecting credit card fraud by decision trees and support vector machines”, *Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2011 Vol I, IMECS 2011, March 2011.*

[6] Elkan, C. (2001). Magical thinking in data mining: lessons from COIL challenge 2000. Proc. of SIGKDD01, 426-431.

[7] Mohammed, J. Zaki., & Wagner, Meira Jr. (2014). Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press. ISBN 978-0-521-76633-3.

[8] F. N. Ogwueleka. (2011). Data mining application in credit card fraud detection system. *Journal of Engineering Science and Technology, Vol. 6, No. 3 (2011) 311 - 322.*

[9] V. Bhusari& S. Patil. (2011). Application of hidden markov model in credit card fraud detection. *International Journal of Distributed and Parallel Systems (IJDPS) Vol.2, No.6.*

[10] S.J. Stolfo, D.W. Fan, W. Lee, A.L. Prodromidis, and P.K. Chan. (1998). Credit card fraud detection using meta-learning: issues and initial results, *Proc. AAAI Workshop AI Methods in Fraud and Risk Management, pp. 83-90.*

[11] Sen, Sanjay Kumar., & Dash, Sujatha. (2013). Meta learning algorithms for credit card fraud detection. *International Journal of Engineering Research and Development Volume 6, Issue 6, pp. 16-20.*

[12] Maes, Sam, Tuyls Karl, Vanschoenwinkel Bram &Manderick, Bernard. (2002). Credit card fraud detection using bayesian and neural networks. *Proc. of 1st NAISO Congress on Neuro Fuzzy Technologies. Hawana.*

[13] A.C. Bahnsen, Aleksandar, Stojanovic., D. Aouada& Bjorn, Ottersten. (2013). Cost sensitive credit card fraud detection using bayes minimum risk. *12th International Conference on Machine Learning and Applications*.

[14] AmlanKundu, SuvasiniPanigrahi, Shamik Sural and Arun K. Majumdar. (2009). Credit card fraud detection: a fusion approach using dempster–shafer theory and bayesian learning. *Special Issue on Information Fusion in Computer Security, Vol. 10, Issue No. 4, pp.354-363.*

[15] Lam, Bacchus (1994). Learning bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence, Vol. 10, Issue No. 3, pp.269–293.*

[16] M. Mehdi, S. Zair, A. Anou and M. Bensebti (2007). A bayesian networks in intrusion detection systems. *International Journal of Computational Intelligence Research, Issue No. 1, pp.0973-1873 Vol. 3.*

[17] R.Najafi&Afsharchi, Mohsen. (2012). Network intrusion detection using tree augmented naive-bayes. *The Third International Conference on Contemporary Issues in Computer and Information Sciences (CICI) 2012.*

[18] G. Cooper, E. Herskovits (1992). A bayesian method for the induction of probabilistic networks from data. Machine Learning. 9(4):309-347.

[19] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

[20] Friedman, N. and Goldszmidt, M. (1996). Building classifiers using bayesian networks. Proc. 13th National Conference on Artificial Intelligence.Vol. 2, pp 1277-1284.

[21] Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers. machine learning,Vol. 29, pp 131-163. Kluwer Academic Publishers, Boston.