IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detect Credit Card Fraud Transactions Using Random Forest and Cart Algorithm

CHANDAKA INDRA RAO [*1], S KESAVARAO[*2]

[*1]M. Tech Scholar, Department of Computer Science & Engineering,

[*2]Associate Professor, Department of Computer Science & Engineering,

Avanthi Institute of Engineering and Technology, (Affiliated to Jawaharlal Nehru Technological University, Kakinada) ,Cherukupally, Vizianagaram.

ABSTRACT

In current days we can find out lot of loss occurred mainly from credit card transactions, almost billions of dollars are getting lost through credit card fraud transactions. In general to identify such fraud transactions there is no accurate approach and hence the design of efficient fraud detection algorithms is the only solution for reducing these losses and avoids fraud transactions. This problem is very difficult because the credit card fraud data is almost unbalanced and unstructured so there is lot of confusion for the one to classify the fields and then find out the fraud transactions manually. In this thesis we try to design a model using Random forest and CART and take sample dataset collected from KAGGLE website and then check the performance of our model using the algorithms. Here we can see random forest can able to predict and find out the fraud activity very easily and accurately compared with several primitive models. The proposed random forest algorithm achieved more than 99.7 percent accuracy for finding the credit card fraud transactions from a large dataset.

KEY WORDS:

Credit Card, Fraud Transactions, Random Forest Algorithm, CART Algorithm, KAGGLE.

1. INTRODUCTION

Falsification of the credit card can be defined as the unapproved use of a customer's card data to create purchases or to dismiss funds from the cardholder's record. The misconduct extortion starts from the credit card when somebody incorrectly acquires the number printed on card or the essential records for the card to be operated [9,10]. The owner of the card, the agent by whom card is issued and even guarantor of a card might not be informed of the fraud until the record is used to create purchases. As shopping through internet-based applications and paying bills online has been come into practice, there is no longer requirement of a physical card to create purchases. Figure 1 shows the taxonomy of frauds. Frauds can be categorized in three ways: financial frauds, communication frauds and online marketing frauds. Credit card frauds come under financial frauds. These frauds must be prevented and detected in time. In this direction, many researches are carried out by various researchers to devise the effective and efficient techniques [14, 15]. Hackers and intruders are trying different new approaches to breach the security [13]. Therefore, there should always be a safety alert against such frauds. Several machine learning based algorithms have been proposed in this direction. A learning based technique is proposed for detecting the credit card frauds.

The online shopping growing day to day. Credit cards are used for purchasing goods and services with the help of virtual card and physical card where as virtual card for online transaction and physical card for offline transaction. In a physical-card based purchase, the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In online payment mode, attackers need only little information for doing fraudulent transaction (secure code, card number, expiration date etc.).

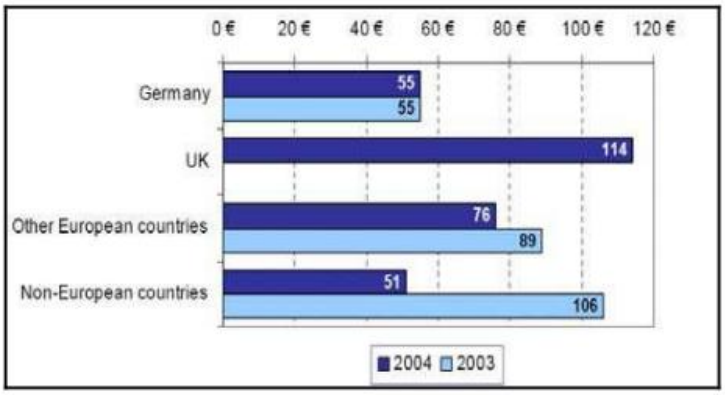
In this purchase method, mainly transactions will be done through Internet or telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyze the spending patterns on every card and to figure out any inconsistency with respect to the "usual" spending patterns. Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behaviorist profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system[1]-[9].

2. LITERATURE SURVEY

In this section we will mainly discuss about the background work that is carried out in order to prove the performance of our proposed Method. Literature survey is the most important step in software development process. For any software or application development, this step plays a very crucial role by determining the several factors like time, money, effort, lines of code and company strength. Once all these several factors are satisfied, then we need to determine which operating system and language used for developing the application. Once the programmers start building the application, they will first observe what are the pre-defined inventions that are done on same concept and then they will try to design the task in some innovated manner.

MOTIVATION

Credit card fraud stands as major problem for word wide financial institutions. Annual lost due to it scales to billions of dollars. We can observe this from many financial reports. Such as (Bhattacharyya et al., 2011) 10th annual online fraud report by Cyber Source shows that estimated loss due to online fraud is $4 billion for 2008 which is 11% increase than $3.6 billion loss in 2007and in 2006, fraud in United Kingdom alone was estimated to be £535 million in 2007 and now[16] costing around 13.9 billion a year (Mahdi et al., 2010). From 2006 to 2008, UK alone has lost £427.0 million to £609.90 million due to credit and debit card fraud (Woolsey &Schulz, 2011). Although, there is some decrease in such losses after implementation of detection and prevention systems by government and bank, card-not-present fraud losses are increasing at higher rate due to online transactions. Worst thing is it is still increasing un-protective and un-detective way[17].



For the credit card listed, the customers are contacted and if they do not react, the card is blocked. Other reports are vintage reports which identify delinquent customers, i.e. transaction reports which identify suspicious transactions. A fraudulent transaction is difficult to detect and to define. Nevertheless, ATM transactions of large amounts are suspicious and demand contact with the customer. Purchases of goods for a larger amount than normal will also be notified to the

customer as well as abnormal overseas spending patterns. Fraudulent transactions are usually impossible to prevent as they occur in a really short period of time. However, once a card is identified, the card is blocked.

3. EXISTING METHODOLOGY

In the existing system there are lot of machine learning approaches implemented for detecting credit card fraud detection which are implemented based on AE, IF, LOF and K- Means which are giving 70%,73% and 71%accuracisrespectively. In this project we are trying to improve the accuracy by refining the data in a better way for efficient detection of credit card fraud. . By using that k-means they try to cluster the dataset into 2 clusters : 0 being non-fraud and 1 as Fraud parameters, but they can't able to classify clearly all the fields. If there are less dimensions the k-means can easily able to cluster the data and find the fraud activities but if the same dataset contains more dimensions this may not generate the accurate results.

LIMITATION OF EXISTING SYSTEM

1. The Clustering produce the less accuracy when compared to Regression methods in scenarios like credit card fraud detection.
2. Comparatively with other algorithms k-means produce less accurate scores in prediction in this kind of scenarios.
3. This is accurate if we use for less dimensions
4. This is not accurate for large dimensional dataset.

4. PROPOSED MODEL

Our goal is to implement machine learning model in order to classify, to the highest possible degree of accuracy, credit card fraud from a dataset gathered from Kaggle. After initial data exploration, we knew we would implement a logistic regression model for best accuracy reports. For that we try to use CART algorithm and classify each and every record which is present in that dataset. This CART is mainly used to classify individual records into two binary classes.I.e Either true or False. Python sklearn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 for no fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train_test_split for data extraction. Now we apply Random Forest Algorithm on training data and calculate the accuracy of the model. By using RF model we can get nearly 99.76 % accuracy for detecting credit card fraud transactions and this is very high compared with primitive algorithms.

Advantages of Proposed System

1. The results obtained by the Random Forest Algorithm is best compared to any other Algorithms.

2. The Accuracy obtained was almost equal to 99.7 percent which proves using of Random Forest gives best results.

3. The plots that were plotted according to the proper data that is processed during the implementation.

5. PROPOSED MODULES

Implementation is a stage where the theoretical design is converted into a programmatic manner. The Application is mainly divided into 5 modules. They are as follows:

1. Load Dataset Module
2. Generate Test and Train Data
3. Run Random Forest Algorithm
4. Detect Fraud from Test Dataset
5. Fraud Transaction Detection Graph

1. LOAD DATASET MODULE

In this module we try to load the dataset which is collected from Kaggle website and then try to give that excel file information as input to the next module.

Dataset URL: https://www.kaggle.com/mlg-ulb/creditcardfrau

To provide privacy to users transaction data kaggles peoples have converted transaction data to numerical format using PCA Algorithm. Below is some example from dataset

2. GENERATE TEST AND TRAIN DATA MODULE

Here we try to divide the data into test and train datasets and we used 70: 30 percent ratio for dividing the whole dataset into multiple parts. Here 70 percent data records are used for training the system and 30 percent data is used for testing the model.

3. RUN RANDOM FOREST ALGORITHM MODULE

Here we try to run the RF algorithm on the train dataset and try to check the probability of each and every attribute which is present in that record. Once all the records are processed now we try to find out which records are having fraud activity and which are having normal activities. Once

we use RF on training dataset ,we can get accuracy of nearly 99.7 percent which is very high compared with several primitive algorithms.

4. DETECT FRAUD FROM TEST DATASET MODULE

Here we try to apply RF model and check the model on test data.Once the test data is given as input we can see the data can be categorized into 2 categories where how many records are found fraud activity and how many are having normal activities.

5.FRAUD TRANSACTION DETECTION GRAPH MODULE

In this module we finally can see that how many records are present in fraud and how many are normal and how many total records are present in the dataset. This module is designed using Matplot library used in Python.

6. RANDOM FOREST IMPLEMENTATION

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

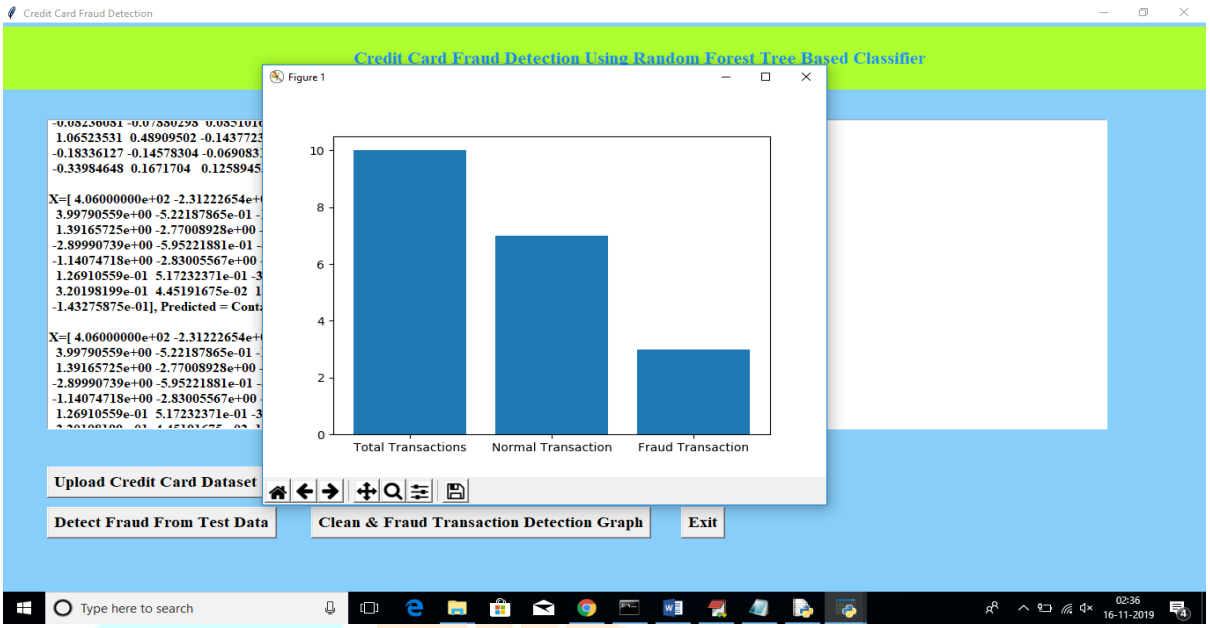
WORKING OF RANDOM FOREST ALGORITHM

The following are the basic steps involved in performing the random forest algorithm

1. Pick N random records from the dataset.
2. Build a decision tree based on these N records.
3. Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
4. For classification problem, each tree in the forest predicts the category to which the new record belongs.

7. RESULT AND DISCUSSION

PERFORMANCE ANALYSIS



From the above window we can clearly identify the total transactions present in the dataset and from that set of normal transaction as well as set of fraud transaction is also identified very accurately in very less time by using RANDOM FOREST Algorithm.Hence from this we can clearly tell that in future credit card fraud transaction detectuon,the random forest plays an important role in order to generate the transaction reports very accurately.

8. CONCLUSION

In this proposed article we for the first time designed an application to detect credit card fraud transactions based on some random forest and CART Algorithms. In this project we have used CART algorithm to classify the genuine transactions as well as fraud transactions from a certain period of time. We have achieved more than 90 percent of accuracy by using this proposed Random Forest Model compared with primitive machine learning algorithms for identifying the fraud transactions present in the credit data gathered from bank database.

9. REFERENCES

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of social media," Business horizons, vol. 53, no. 1, pp. 59–68, 2010.

[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and metaanalysis of cyberbullying research among youth." 2014.

[3] M. Ybarra, "Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression," National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda, 2010.

[4] B. K. Biggs, J. M. Nelson, and M. L. Sampilo, "Peer relations in the anxiety–depression link: Test of a mediation model," Anxiety, Stress, & Coping, vol. 23, no. 4, pp. 431–447, 2010.

[5] S. R. Jimerson, S. M. Swearer, and D. L. Espelage, Handbook of bullying in schools: An international perspective. Routledge/Taylor & Francis Group, 2010.

[6] G. Gini and T. Pozzoli, "Association between bullying and psychosomatic problems: A meta-analysis," Pediatrics, vol. 123, no. 3,pp. 1059–1065, 2009.

[7] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime," Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[8] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, 2012, pp. 656–666.

[9] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. ACM, 2014, pp. 3–6.

[10] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying." in The Social Mobile Web, 2011.

[12] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," Communications in Information Science and Management Engineering, 2012.

[13] M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, "Improved cyberbullying detection using gender information," in Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop(DIR2012). Ghent, Belgium: ACM, 2012.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in Advances in Information Retrieval. Springer, 2013, pp. 693–696.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.

[16] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7, p. 43, 2012.

[17] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," arXiv preprint arXiv: 1206.4683, 2012.