Credit card fraud detection using Isolation Forest

Snehanka Bhosale, Prof. A.D.Gotmare

Department of Computer Engineering BDCE Wardha

Abstract

Now a days due to de-monetization everyone had started using credit cards for different types of transactions. So there will be a more chances for occurring fraud. Banks have many and enormous databases. Important business information can be extracted from these data stores. Fraud is an issue with far reaching consequences in the backing industry, government, corporate sectors and for ordinary consumers. Increasing dependence on new technologies such as cloud and mobile computing in recent years has encountered the problem. Physical detections are not only time consuming they are costly and they don't give accurate results. Not surprisingly economic institutions have turned to automated process using numerical and computational methods. Traditional approaches relied on manual techniques such as auditing, which are inefficient and unreliable due to the difficulty of the problem. Data mining based approaches have been shown to be useful because of their ability to identify small anomalies in large data sets. So we have used some of the supervised algorithms to detect the fraud which gives accurate results. There are many different types of fraud, as well as a variety of data mining methods, and research is continually being undertaken to find the best approach for each case. Financial fraud is a term with various potential meanings, but for our purposes it can be defined as the on purpose use of illegal methods or practices for the purpose of obtaining financial gain . Fraud has a large negative impact on business and society credit card fraud alone accounts for billions of dollars of lost revenue each year. We have used Isolation forest for classification of fraud and valid record from dataset. Dataset used is downloaded from Kaggle based on 20 different attributes.

Keywords: Fraud detection, Decision tree, Isolation Forest

I. INTRODUCTION

Fraud refers to the abuse of a profit organization's system without necessarily leading to direct legal concerns. Fraud is a universal act in order to deceive another person or organization for financial benefits. Credit card fraud detection is the process of identify those transactions that are false into two classes of lawful and fake transactions. These kind of frauds can be broadly classified into three categories that is traditional card related frauds and internet frauds .The fraud which is committed by individuals exterior to the organization is called as customer fraud or external fraud where when a fraud is committed by top-level management is known as management fraud or internal fraud. Fraud detection being part of all the overall fraud control, automates and helps reduce the manual parts of a screening process.

A credit card is a thin handy plastic card that contains identification information such as a signature or picture, and authorizes the person named on it to charge purchases or services to his account - charges for which he will be billed periodically. Today, the information on the card is read by automated teller machines (ATMs), store readers, bank and is also used in online internet banking system. They have a unique card number which is of utmost importance. Its security relies on the physical security of the plastic card as well as the privacy of the credit card number. There is a rapid growth in the number of credit card transactions which has led to a substantial rise in fraudulent activities. Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card as a fraudulent source of funds in a given transaction. Generally, the statistical methods and many data mining algorithms are used to solve this fraud detection problem. Most of the credit card fraud detection systems are based on artificial intelligence, Meta learning and pattern matching. The Genetic algorithms are evolutionary algorithms which aim to obtain the better solutions in eliminating the fraud. A high importance is given to develop efficient and secure electronic payment system to detect whether a transaction is fraudulent or not. In this paper, we will focus on credit card fraud and its detection measures.

II. LITERATURE SURVEY

[Jarrod West, Maumita Bhattacharya]"Intelligent Financial fraud detection" This author explains about different intelligent approaches to fraud detection which are both statistical and computational though the performance was differed each technique was shown to be reasonably capable at detecting various forms of financial fraud. The ability of the computational methods such as neural networks and support vector machines to learn and adapt to many new techniques is highly effective to the evolving of tactic fraudsters. Initial fraud detection studies focused heavily on statistical models such as logistic regression, as well as neural networks. Neural networks are used for financial applications such as forecasting. Neural network are well established history with fraud detection. But they require high computational power for training and operation, making it unsuitable for real-time function. Potential for over fitting if training set is not a good representation of the problem domain, so requires constant retraining to adapt to new methods of fraud. In this paper the author says about the different kinds of frauds i.e., insurance fraud , mortgage fraud , health insurance fraud , telecommunication fraud , credit card fraud. Different techniques have been defined for different kinds of frauds defining the parameters like entropy, sensitivity and comparing the efficiency of the different kinds of algorithms and representing them in a graphical representation.

[Rasa kanapickiene, Zivile Grundiene] ”The model of fraud detection by means of financial ratios” This author explains about how financial ratios are analysed in order to determine the most fraud-sensitive ratios of financial statements with regard to company managers’ and employees’ motivation to commit fraud. It was found out that in most cases fraud is committed to show that the company keeps growing and to fulfill obligation conditions. Literary sources offer a wide range of such ratios. Theoretical analysis showed that profitability, liquidity, activity and structure ratios are analyzed most often. Theoretical survey revealed that, in scientific literature, financial ratios are analyzed in order to designate which ratios of the financial statements are the most sensitive in relation with the motifs of executive managers and employees of companies to commit frauds. The logistic regression model of fraud detection in financial statements has been developed.

[Fletcher H. Glancy, Surya B. Yadav] “A computational model for financial reporting fraud detection” This author explains that the computational fraud detection model is possible to detect financial exposure fraud from the text of annual filings with the Security and Exchange Commission. The model is generalizable because it specifies automatable steps that can be adapted to other domains and genres. A potential application for CFDM is to screen companies for investigation of potential fraud by the SEC (Security and exchange commission). Additional potential applications include financier analysis, e-mail spam detection, and business intelligence validation. A computational fraud detection model (CFDM) was proposed for detecting fraud in financial reporting. CFDM uses a quantitative approach on textual data. It incorporates techniques that use essentially all of information contained in the textual data for fraud detection.

III. PROPOSED METHODOLOGY

There are lots of issues that make this procedure tough to implement and one of the biggest problems associated with fraud detection is the lack of both the literature providing experimental results and of real world data for academic researchers to perform experiments on. The reason behind this is the sensitive financial data associated with the fraud that has to be kept confidential for the purpose of customer’s privacy. Now, here we enumerate different properties a fraud detection system should have in order to generate proper results:

- The system should be able to handle skewed distributions, since only a very small percentage of all credit card transactions is fraudulent. There should be a proper means to handle the noise. Noise is the errors that is present in the data, for example, incorrect dates. This noise in actual data limits the accuracy of generalization that can be achieved, irrespective of how extensive the training set is.

- Another problem related to this field is overlapping data. Many transactions may resemble fraudulent transactions when actually they are genuine transactions. The opposite also happens, when a fraudulent transactions appears to be genuine.

- The systems should be able to adapt themselves to new kinds of fraud. Since after a while, successful fraud techniques decreases in efficiency due to the fact that they become well known because an efficient fraudster always find a new and inventive ways of performing his job.

- There is a need for good metrics to evaluate the classifier system. For example, the overall accuracy is not suited for evaluation on a skewed distribution, since even with a very high accuracy; almost all fraudulent transactions can be misclassified.

- The system should take care of the amount of money that is being lost due to fraud and the amount of money that will be required to detect that fraud. For example, no profit is made by stopping a fraudulent transaction that is way lesser than the amount of money that will be required to detect it.

Isolation Forest

The IsolationForest ‘isolates’ observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node.

This path length, averaged over a forest of such random trees, is a measure of normality and our decision function. Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.

The idea of identifying a normal vs. abnormal observation can be observed in Figure 1 from [1]. A normal point (on the left) requires more partitions to be identified than an abnormal point (right).

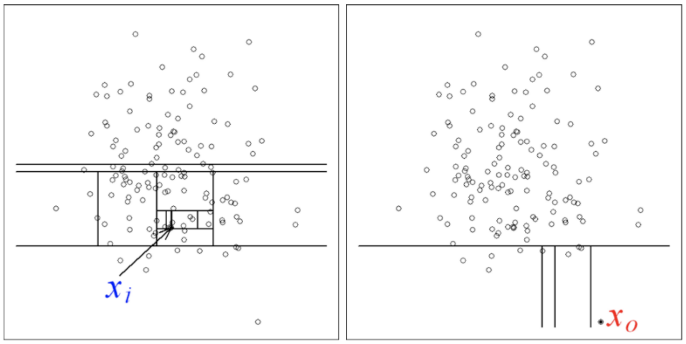

Figure 1 Identifying normal vs. abnormal observations

As with other outlier detection methods, an anomaly score is required for decision making. In case of Isolation Forest it is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where *h(x)* is the path length of observation *x*, *c(n)* is the average path length of unsuccessful search in a Binary Search Tree and *n* is the number of external nodes. More on the anomaly score and its components can be read in [1]. Each observation is given an anomaly score and the following decision can be made on its basis:

- Score close to 1 indicates anomalies
- Score much smaller than 0.5 indicates normal observations

If all scores are close to 0.5 than the entire sample does not seem to have clearly distinct anomalies

IV. CONCLUSION

Credit card fraud has become more and more rampant in recent years. To improve merchants' risk management level in an automatic and effective way, building an accurate and easy handling credit card risk monitoring system is one of the key tasks for the merchant banks. One aim of this study is to identify the user model that best identifies fraud cases. There are many ways of detection of credit card fraud. If one of these or combination of algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks. And a series of anti-fraud strategies can be adopted to prevent banks from great losses before and reduce risks. This project gives contribution towards the credit card fraud detection using the supervised learning algorithms. We have used Isolation Forest for outlier detection and obtained accuracy of 99.67% in terms of total classified records.

V. REFERENCES

[1] Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), "Credit card fraud and detection techniques: a review", Banks and Bank Systems, Volume 4, Issue 2, 2009.

[2] Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, "A review of Fraud Detection Techniques: Credit Card", International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012 .

[3] Vladimir Zaslavsky and Anna Strizhak," credit card fraud detection using self organizing maps", information & security. An International Journal, Vol.18,2006.

[4] L. Mukhanov, "Using bayesian belief networks for credit card fraud detection," in Proc. of the IASTED International conference on Artificial Intelligence and Applications, Insbruck, Austria, Feb. 2008, pp. 221– 225.

[5] John T.S Quah,M Sriganesh "Real time Credit Card Fraud Detection using Computational Intelligence" ELSEVIER Science Direct,35 (2008) 1721-1732.

[6] Joseph King –Fung Pun, "Improving Credit Card Fraud Detection using a Meta Heuristic Learning Strategy" Chemical Engineering and Applied Chemistry University of Tornto 2011.

[7] Kenneth Revett,Magalhaes and Hanrique Santos "Data Mining a Keystroke dynamic Based Biometric Dtatabase Using Rough Set" IEEE

[8] Linda Delamaire ,Hussein Abdou and John Pointon, "Credit Card Fraud and Detection technique", Bank and Bank System,Volume 4, 2009.