NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 3, ISSUE3, MAR.-2016

FRAUD ANALYTICS: A SURVEY ON BANK FRAUD AND FRAUD PREDICTION USING UNSUPERVISED LEARNING BASED APPROACH

Shashank Sharma
Data Analytics Associate, Lera Technologies, Hyderabad

Arjun Roy Choudhury
Data Analytics Associate, Lera Technologies, Hyderabad

ABSTRACT
Fraud in banks has been steadily growing over the past years and is a challenge to banks worldwide. The complexity involved in detection of such fraudulent activities further adds to the problem. A thorough examination of fraud and its possibilities is necessary to pinpoint and distinguish the few fraudulent cases within the vast volumes of banking data. In this paper we have discussed various scenarios in which fraud could take place and applied unsupervised learning approaches to detect fraudulent acts in areas such as credit cards, money laundering and financial statements. We have keenly analyzed various attributes which would be necessary in detection of culprits who may cause a loss to the banks/organizations. Our analysis assists in discovering anomalous behavior among peer groups to more consistently uncover frauds with lesser amount of false positives.

INTRODUCTION
With the evolution of internet in the banking sectors, people have changed the way they used to bank. But this digital evolution is also creating new opportunities for fraudsters to hack into personal accounts. Banking sector frauds have been in existence for centuries, with the earliest known frauds pertaining to insider trading, stock manipulation, accounting irregularity/ inflated assists etc. Fraud is a dominant form of white collar crime that continues to extract a significant toll not only on the organizations, but also on investors, financial institutions, and the economy in general. There are many issues that make effective fraud management a challenging task. These include: enormous and ever-expanding volumes of data, the growing complexity of systems, changes in business processes and activities and continuous evolution of newer fraud schemes to bypass existing detection techniques. Detecting fraudulent financial statements is a difficult task when using normal audit procedures due to limitation in understanding the characteristics of financial statements, lack of experience and dynamically changing strategies of fraudsters.
According to the Basel II definition, Fraud is a part of operational risk and has been classified as Internal and External fraud. Internal Fraud is the risk of unexpected financial, material or reputational loss as the result of fraudulent action of persons internal to the firm. Losses are due to acts of a type intended to defraud, misappropriate property or circumvent regulations, the law or company policy, excluding diversity/discrimination events, which involves at least one internal party. It includes misappropriation of assets, tax evasion, intentional mismarking of positions, bribery. The Basel Committee is the primary global standard-setter for the prudential regulation of banks and provides a forum for cooperation on banking supervisory matters.
Reserve Bank of India has defined fraud as "All instances wherein Banks have been put to loss through misrepresentation of books of accounts, fraudulent encashment of instruments like cheques, drafts and bills of exchange, unauthorized handling of securities charged to

banks, misfeasance, embezzlement, theft, misappropriation of funds, conversion of property, cheating, shortages, irregularities etc."

According to a survey done by EY which included more than 2,700 executives across 59 countries, the risks businesses are facing are not receding. More than 1 in 10 executives surveyed reported their company as having experienced a significant fraud in the past two years. This implies that there is a 5-7% chance of occurrence of fraud cases within a year.

Also, according to a Delloite survey, 93% respondents indicated that there has been an increase in fraud incidents in the banking industry in the last two years.

There are dozens of ways in which an individual can commit bank fraud. Some of these schemes are more complex, and affect more people or institutions, garnering harsher penalties than others do.

Typically, fraud in banks can be categorized into 3 main categories: Corruption, Asset Misappropriation and Financial Statement Fraud (Fig 1). Corruption includes cases of conflict of interest, bribery, illegal gratuities and extortions. Asset misappropriation may be embezzlement or inventory related fraud. Financial statement fraud involves overstating or understating the assets or revenue generated. Below listed are some typical fraud scenarios in the financial domain:

Fraud cases involving theft of identity are a serious and growing problem in the era of internet banking. With so many transactions done online, hackers have the ability to frequently access bank account and credit card information from unwitting consumers. Fraudsters can also use obtained names and addresses to apply for fraudulent accounts, credit cards and loans.

- Embezzlement could occur when a bank employee misappropriates funds from customers or from the bank itself. Banks usually guard rigorously against embezzlement in a variety of ways, since this type of bank fraud can be extremely harmful to the institution's reputation. Bank fraud cases involving internal theft usually are managed by people with considerable power within a bank branch, since they have the most access and opportunity and are generally perceived as trustworthy.

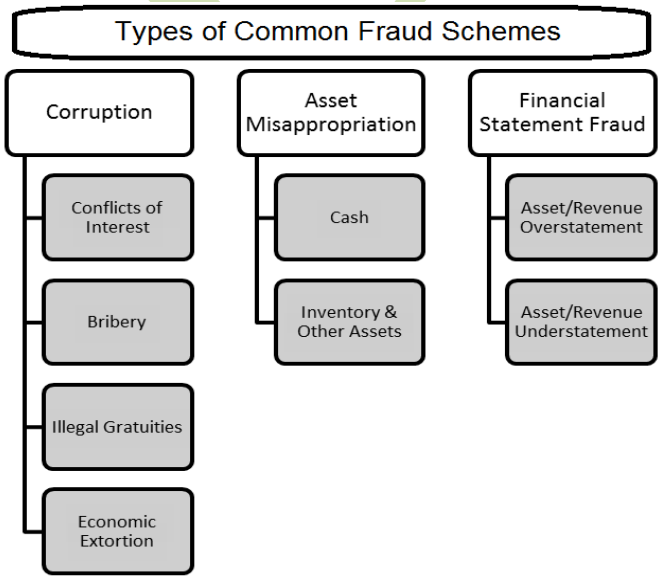

Fig. 1 Types of Common Fraud Schemes

- Bank Impersonation is where one or more individuals act as a financial institution, often by setting up fake companies, or creating websites, in order to lure people into depositing funds.

- In the context of bank fraud, internet fraud occurs when someone creates a website for the purpose of presenting themselves as a bank or other financial institution, to fraudulently obtain money deposited by other people or get the login credentials of the customers of the bank.
- A fraudulent loan is one in which the borrower is an individual or a business entity controlled by a dishonest bank officer or an accomplice; the "borrower" then declares bankruptcy or vanishes and the money is gone. The borrower may even be a non-existent entity and the loan simply a deception to conceal a theft of a large amount of money from the bank. An individual who takes out a loan, knowing that he will immediately file bankruptcy, has committed bank fraud.
- When money is obtained from criminal acts such as illegal gambling or drug trafficking, the money is considered "dirty" in that it may seem dubious if deposited directly into a bank or other financial institution. Since the criminal needs to create financial records ostensibly showing where the money came from, the money must be "cleaned," by running it through a number of legitimate businesses before depositing it, hence the term "money laundering." Because the act is specifically used to hide illegally obtained money, it too is unlawful.

A lot of work has already been done in the field of fraud analytics where most researchers have used supervised classification methods like logistic regression, decision tree, neural networks and SVM. In these kinds of methods there is the need for the model training process where the class labels of the cases in the sample to be used to train should be known first. But in practical scenario, where the probability of occurrence of a fraud is less, it is very difficult to find the train dataset from an organization's database. Also, being dependent on the train set to predict the future activities as fraudulent is a naïve approach and the probability of detecting a fraudulent attack is low.
As the probability of occurrence of the next consequent attack to follow the same pattern as of the previous attackers is also very less, the supervised trained model won't be as effective as unsupervised technique.
In this paper we have discussed various scenarios in which an individual could commit a fraud. We have adopted an unsupervised clustering based approach (Self organizing map) in combination with association rule mining to predict fraudulent activities as this approach does not need the class labels of the cases in the sample.

RELATED WORKS

Detecting management fraud is a difficult task when using normal audit procedures [10]. First, there is a shortage of knowledge concerning the characteristics of management fraud. Secondly, given its infrequency, most auditors lack the experience necessary to detect it. Finally, managers deliberately try to deceive auditors [6]. These limitations suggest that there is a need for additional analytical procedures for the effective detection of management fraud. Fraud classification model using neural networks [7] has been developed. Neural, operation based, real-time fraud detection systems are not only technically feasible, but highly interesting from a purely economic point of view [5].
Neural network with statistical methods have been used to detect fraud [6]. Statistical regression analysis [1] and statistical method of logistic regression have also been tried to detect fraud in banks [16].
Calibrated probabilities followed by Bayes minimum risk outperformed raw probabilities with a fixed threshold to identify credit card fraud [2]. A stepwise logistic regression model was developed and the attributes were analysed, which were required to detect a fraudulent

financial act [9]. Financial statement for fraud prediction were analysed and concluded that the earnings-operating cash flow relation provides important information in identifying financial statement fraud, especially when considering in comparison with other factors associated with fraud risk [18].

A sample of 77 fraud engagements and 305 non-fraud engagements was considered as sample dataset for detecting frauds and a logistic regression model that estimates the likelihood of fraudulent financial reporting for an audit client, conditioned on the presence or absence of several fraud-risk factors [19]. To detect credit card fraud better, support vector machines and random forests together with the LR were used [8]. Random forests demonstrated better overall performance across performance measures.

To detect money laundering patterns, rules and activities, core decision tree with clustering based algorithm such as BIRCH and K means algorithms have been proposed. To identify the abnormal information classification rules with core decision tree has been proposed [14].

Credit card fraud detection system was installed at Mellon Bank which was trained using neural network on fraud due to

lost cards, stolen cards, application fraud, counterfeit fraud etc. and presented that a trained neural network model was performing better than rule based fraud detection procedures [15].

Any target account showing spending behaviour anomalous to that of its peer group is flagged for further inspection; unsupervised profiling based method was adopted to identify fraudulent transactions [12].

Questionnaire Respondent Transaction (QRT) data of new users was collected by using an online questionnaire system and using this data as train set to Support vector machines (SVM) and Back propagation network (BPN) was used to detect credit card fraud for new users. SVM performs better than BPN for smaller dataset [13]. One major obstacle for using neural network training techniques is the high necessary diagnostic quality. Since only one financial transaction of a thousand is invalid no prediction success less than 99.9% is acceptable [4].

Agglomerative hierarchical clustering algorithms were used to rank the fraud activity with no significant additional computational costs [20]. The losses due to fraud have been reduced by using calibrated probabilities with Bayes minimum Risk. To predict a fraudulent transaction, Bayes minimum risk (BMR) was used. This approach required good calibrated probabilities in order to correctly estimate the individual transactions expected costs [3].

OUR APPROACH

Once we have obtained the clusters by using SOM technique we revalidate our clusters by using association rules on each cluster. To apply association rules we need to provide categorical data as input, so we convert numeric data into categorical data based on few criteria. These criteria are defined after a lot of analysis of the attributes.

Association rules are if-then statements that assist in identifying relationships among attributes in unrelated data in a database. Relationships between objects which frequently occur together are identified by association rules. Support and confidence are two primary criteria used by association rules. They help in identifying the relationships in the data by analysing on frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support and a user specified minimum confidence at the same time.

Support gives an idea of the frequency with which the items appear in the data, and confidence describes the proportion in which the if/then statements have been found to be true.

Apriori algorithm [22] is used in mining frequent item set and association rule learning. A level-wise search is used in the algorithm, where k-itemsets are used to explore (k+1)-itemsets, to mine frequent itemsets from transactional database for Boolean association rules.

An itemset which contains k items is known as k-itemset. In this algorithm, frequent subsets are added one item at a time. This step is called as candidate generation process. Testing of groups of candidates is done against the data. To count candidate item sets efficiently, this algorithm uses breadth-first search method and a hash tree structure. The frequent individual items are identified in the database and added to larger and larger item sets as long as those item sets appear frequently often in the database. Apriori assists in determining frequent item sets that may be used to identify association rules which depict general trends in the data.
We have selected the rules with low support and high confidence values for predicting fraud cases. These set of rules are applied to bank transactions to search the customers user ids (UID) which follow the corresponding rules / pattern. This list of users is further added to the watch list for prudently looking into their transactions.

$$wj(t+1) = wj(t) + \alpha\,(t)hj,i(x)(t)\,(x(t) - wj(t)).$$

The amount of model vector movement is guided by a learning rate α, decreasing in time and neighbourhood function hj,i(x). The neighbourhood function is a unimodal function which is symmetric around the location of the winner and monotonically decreasing with increasing distance from the winner.

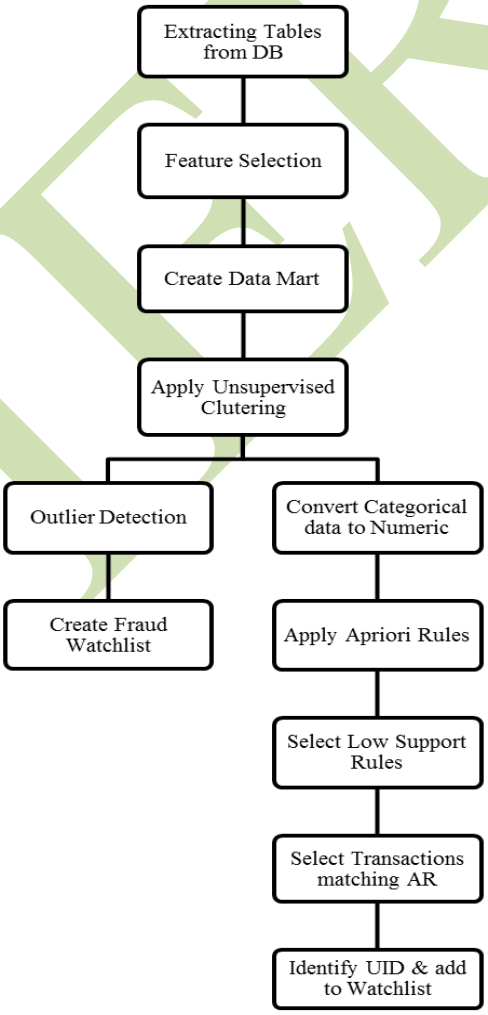

Fig. 2. Flowchart of fraud detection model

Neighbourhood function is given as:

$$hj,i(x) = \exp(\|rc\text{-}ri\|2/\ 2.\ \Delta(t)2)$$

Where $\|rc - ri\|^2$ denotes the distance between units c and i within the output space, with ri representing the two-dimensional location vector of unit i within the grid. The spatial width of the kernel is reduced gradually during training and winner is adapted at the end. [11]

Phase I: Outlier detection is done once we obtain clusters by using SOM technique. We find outliers by identifying the attributes which vary the most from the mean. Taking these set of attributes into consideration we search for the corresponding user ids which fall under this category. This set of user ids are added to the watch list which are presented to the bank to keenly observe their banking transactions.

Phase II: Once we have obtained the clusters by using SOM technique we revalidate our clusters by using association rules on each cluster. To apply association rules we need to provide categorical data as input, so we convert numeric data into categorical data based on few criteria. These criteria are defined after a lot of analysis of the attributes.

Association rules are if-then statements that assist in identifying relationships among attributes in unrelated data in a database. Relationships between objects which frequently occur together are identified by association rules. Support and confidence are two primary criteria used by association rules. They help in identifying the relationships in the data by analyzing on frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support and a user specified minimum confidence at the same time.

Support gives an idea of the frequency with which the items appear in the data, and confidence describes the proportion in which the if/then statements have been found to be true.

Apriori algorithm [22] is used in mining frequent item set and association rule learning. A level-wise search is used in the algorithm, where k-itemsets are used to explore (k+1)-itemsets, to mine frequent itemsets from transactional database for Boolean association rules. An itemset which contains k items is known as k-itemset. In this algorithm, frequent subsets are added one item at a time. This step is called as candidate generation process. Testing of groups of candidates is done against the data. To count candidate item sets efficiently, this algorithm uses breadth-first search method and a hash tree structure. The frequent individual items are identified in the database and added to larger and larger item sets as long as those item sets appear frequently often in the database. Apriori assists in determining frequent item sets that may be used to identify association rules which depict general trends in the data.

We have selected the rules with low support and high confidence values for predicting fraud cases. These set of rules are applied to bank transactions to search the customers user ids (UID) which follow the corresponding rules / pattern. This list of users is further added to the watch list for prudently looking into their transactions.

CONCLUSION

With the growth in Internet Banking, fraud in banks has become much more common. We have proposed a novel profit-based unsupervised learning based fraud prediction model to improve security of the financial transaction systems in an automatic and effective way which is going to make a stronger impact on fraud detection and predictions. Here, we have prepared a watch-list of the fraudulent customers which could be helpful for the respective banks to look into.

In the end, no automated system can exactly detect the fraudulent person. However as a preventive measure, we can prepare a record of the customers who may be harmful to the bank financially and protect the banks reputation. Also, frauds done
physically (corruption, forgery of cheques) for which database records are not available are difficult to identify using data analytics.
As for the future research, we are working on models which shall be able to predict the approximate loss to the bank due to fraudulent activities by the customers. We are also working on other probabilistic and statistical based approach to predict the fraud.

ACKNOWLEDGMENT

We are deeply indebted to LERA Technologies for the support and opportunity to pursue this research on Fraud Analytics in Banking. A hearty thanks to Mr. Shanthi Narayan (MD), Mr. Murali Valiveti (CEO) and Mr. Muralidhar Jupudi (Head – BSG) for their guidance and motivating us to undertake the project. Finally we would like to thank our families for their enduring support in all our endeavours'

REFERENCES

[1] Abbot, J. L., Park, Y., & Parker, S. (2000). *The effects of audit committee activity and independence on corporate fraud. Managerial Finance, 26(11), 55–67*

[2] Alejandro Correa Bahnsen, Aleksandar Stojanovic, Djamila Aouada, and Bjorn Ottersten. (2014). *Improving Credit Card Fraud Detection with Calibrated Probabilities. Proceedings of the 2014 SIAM International Conference on Data Mining, 677-685.*

[3] Alejandro Correa Bahnsen, Aleksandar Stojanovic, Djamila Aouada, and Bjorn Ottersten. (2013). *Cost Sensitive Credit Card Fraud Detection using Bayes Minimum Risk. International Conference on Machine Learning and Applications.*

[4] Brause, R., Langsdorf T. and Hepp M. (1999). *Neural Data Mining for Credit Card Fraud Detection. Proceedings: 11th IEEE International Conference on Tools with Artificial Intelligence.*

[5] Dorronsoro, J. R., Ginel F., Sanchez C. and Cruz C. S. (1997). *Neural Fraud Detection in Credit Card Operations. IEEE Transactions on Neural Networks 8(4), 827-834.*

[6] Deloitte: India Banking Fraud Survey Edition II, *https:// www2 .deloitte. com/ content / dam/Deloitte/in/Documents/finance/in-fa-banking-fraud-survey-noexp.pdf*

[7] Ernst & Young: Overcoming compliance fatigue, *http:// www .ey. com/ Publication / vwLUAssets/EY-13th-Global-Fraud-Survey/$FILE/EY-13th-Global-Fraud-Survey.pdf*

[8] Fanning, K., & Cogger, K. (1998). *Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management, 7(1), 21–24*

[9] Green, B. P., & Choi, J. H. (1997). *Assessing the risk of management fraud through neural-network technology. Auditing: A Journal of Practice and Theory, 16(1), 14–28.*

[10] Homma, N.; Gupta, M.M. (2004). *Fuzzy self-organizing map in cerebral cortical structure for pattern recognition. In: Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting, vol.2, no., pp.539-544 Vol.2, 27-30*

[11] Intellinx: Basel II, *http://www.intellinx-sw.com/solutions/basel-ii/*

[12] India Infrastructure Finanace Company Ltd: *Fraud Prevention & Detection Policy, http://www.iifcl.co.in/Content/FraudPolicy.aspx*

[13] J. Sanjeev, M. Guillen and J.C. Westland. (2012*). Employing transaction aggregation strategy to detect credit card fraud. Expert Systems with Applications, vol. 39, pp. 12650–12657.*

[14] O.S. Persons. (1995). *Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research, vol. 11, pp. 38-46.*

[15] Porter, B., & Cameron, A. (1987). *Company fraud—what price the auditor? Accountant's Journal (December), 44 47*

[16] Rauber, A.; Merkl, D.; Dittenbach, M. (2002). *The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. Neural Networks, IEEE Transactions, vol.13, no.6, pp.1331-1341.*

[17] R.J. Bolton and D. J. Hand. (2001). *Unsupervised profiling methods for fraud detection. Credit Scoring and Credit Control VII*

[18] Rong-Chang Chen, Shu-Ting Luo, Xun Liang, Lee, Vincent C S. (2005). *Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud. Neural Networks and Brain. ICNN&B '05. International Conference on vol.2, no., pp.810-815, 13-15.*

[19] Rui Liu; Xiao-long Qian; Shu Mao; Shuai-zheng Zhu. (2011). *Research on anti-money laundering based on core decision tree algorithm. Control and Decision Conference (CCDC), vol., no., pp.4322-4325, 23-25*

[20] S. Ghosh and D. Reilly. (1994). *Credit card fraud detection with a neural-network. Proceedings of the 27th Annual Hawaii International Conference on System Science, volume 3, Los Alamitos, CA.*

[21] Spathis, C. (2002). *Detecting false financial statements using published data: some evidence from Greece. Managerial Auditing Journal, 17(4), 179–191.*

[22] Tank and Darshan M. (2014). *Improved Apriori Algorithm for Mining Association Rules. International Journal of Information Technology and Computer Science (IJITCS) 6, no. 7.*

[23] T.Kohonen. (1990). *The self-organizing map. Proceedings of the IEEE, vol.78, no. 9, pp.1464-1480.*

[24] T.A. Lee, R.W. Ingram and T.P. Howard. (1999). *The Difference between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud. Contemporary Accounting Research, vol. 16, pp. 749-786.*

[25] T.B. Bell and J.V. Carcello. (2000). *Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. Auditing: A Journal of Practice & Theory, vol. 19, pp. 169-184.*

[26] Torgo, Luis. (2007). *Resource-bounded fraud detection. Artificial Intelligence. Springer Berlin Heidelberg, 449-460.*