Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Data Mining Techniques to Detect Insider Fraud in Banks

Mr. Vimal Kumar M
Asst. Manager – Systems
State Bank of Mysore
Bangalore, India

Mr. Sriganga B K
Business Consulting Consultant
NTT DATA
Bangalore, India

Abstract: - Banks are an integral part of a country's economy, contributing to both people and governments. In recent past a lot of fraudulent activities have been reported in banks because of people with vested interests. This paper throws light on common insider frauds occurring in banks and also tries to categorize them into different types. Here we give a perspective of definition, factors related to such categories of fraud and also the challenges one faces in detecting frauds. It is very important to detect such fraudulent activities automatically before it is too late and bring the people or group of people into terms. Data mining technique comes handy as it helps to identify unusual patters in given data set. This paper focusses on different generic data mining techniques and in specific, the techniques used for detecting insider frauds. We also list out best available techniques in realizing insider frauds with relevant illustrations. As it is very evident in today's world, Big Data challenges are popping up day-in and day-out, we have gone ahead and presented as a future enhancement a few big data challenges while handling banking data.

Keywords: Insider Fraud, Account Payable, Money Laundering, Miss-Selling, Fraud Detection, Data Mining, Benford's Law, Fuzzy Logic, Clustering

I. INTRODUCTION

"New Delhi, August 28: The Banks opened a record 15 million accounts on Thursday, against a target of 10 million, this was Prime Minister Narendra Modi launching his government's massive financial inclusion programme, the Pradhan Mantri Jan Dhan Yojana. The scheme is mainly aimed to take banking facilities to 75 million households by January 26, 2015, against the earlier deadline of August 15, 2015." – THE BUSINESS STANDARD, 29th August, 2014 [1]

India is the largest democracy of the world & the second largest economy. Banking is one of the primary tools to run the nations' economy. In the last decade, the Indian Banking sector grew at an average of 18% compared to over 7% GDP growth. Hence Our Prime Minister Narendra Modi picked the Financial Inclusion program as the first step towards development of the country, through which he wants to achieve 100% banking penetration in the country.

Banking industry accommodates 1,175,149 employees and had a total of 109,811 branches in India and 171 branches abroad and manages an aggregate deposit of INR 67,504.54 billion (US$1.1 trillion or €840 billion) and bank credit of INR 52,604.59 billion (US$870 billion or €650 billion). The net profit of the banks operating in India was INR 1,027.51 billion (US$17 billion or €13 billion) against a turnover of INR 9,148.59 billion (US$150 billion or €110 billion) for the financial year 2012-13.[2]

It is inevitable that such a huge industry is also prone to many frauds. As per the 'India fraud survey' conducted by KPMG in 2012,[3] states that, '*"Despite having a strong regulator, the financial services sector has emerged as the most susceptible sector to fraud. The misuse of technology in the banking sector includes use of banking access for overpayments to vendors or self-bank account, sharing of potential confidential information and misuse of the company's technology resources for unauthorized activities which includes conflicting business relationship. Also, providing services on mobile and social media platforms with limited knowledge of the security requirements, poses lot of threats to customers as well as the financial institutions."* As the single-biggest perpetrators to frauds, employees were on the top of the list accounting for 36% of the total, according to the survey respondents.

A study shows that the Public sector banks have cumulatively lost a huge sum of Rs.22,743 crore due to the activities of cheating and forgery during 2010-2013[4]. According to the study, more than 6000 employees of the different banks are under the scanner for involvement in these cases. These are not found in just lower level employees or the mid-level employees, but also in some cases, CMDs and directors of different banks. It is also important to remember here the incident of the recent arrest of the CMD of Syndicate Bank in a bribe-for-loan scandal with Bhushan Steel [5].

These statistics shows that Insider Fraud problem is one of the biggest concerns for banking industry. Because, an Insider fraud not only causes financial loss to organization, it also affects the good will of the organization. Financial institutes are the custodians of the public's money. They only run on trust of the public. If you lose the trust, you will lose everything. We can see below the share price of Syndicate Bank b/w 29.07.2014 to 29.08.2014. On 1st of August, it was trading at 149.80Rs/Share. On 2nd August, we got the news that CMD of Syndicate bank was arrested for Cash-for-loan scam. The share price of the company immediately started following. It went down to 118.45Rs/share within a week[figure 1].

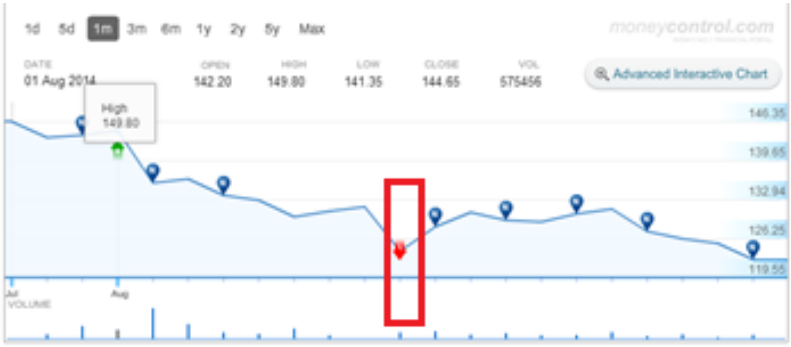

Figure 1: Share price movement of Syndicate Bank - Jul-Aug 2014(Courtesy: Money Control)

According to the annual report of Chief Vigilance Office, around 50% of the frauds proceedings carried out in PSU are from banking sector. The statistics from CVO Annual Report 2013[6] are as per the table 1.

In this paper we will try to understand the challenges faced by the banking industry in detecting & preventing the Insider Frauds and then we will review some of the Data Analysis techniques which can play a big role in detecting the insider frauds.

II. WHAT IS A FRAUD

How to define a fraud? Oxford dictionary defines fraud as:

"An act or instance of deception, an artifice by which the right or interest of another is injured, a dishonest trick or stratagem."

The Institute of Internal Auditors'[7] International Standards for the Professional Practice defines frauds as:

"… any of the illegal acts which are characterized by deceit, concealment, or violation of trust. These acts are not dependent upon the threat of violence or the physical force. Frauds are perpetrated by parties and organizations to obtain money, property or services; to avoid payment or loss of services; or to secure the personal and the business advantage."

When we define a fraud from professional perspective, any act which are done in a dishonest manner, violates the trust and laws of the organization can be considered as fraud.

III. FACTORS CAUSING THE FRAUD:

Why do people commit fraud? One of the famous criminologists Donald Cressey proposed a fraud triangle model[8] which was used for explaining the factors that cause someone to commit occupational fraud. It consists of mainly three components which lead to fraudulent behavior, as listed and explained below:

1. **Perceived unsharable financial need/Motivation:** Motivation which can be also referred to as incentive, is another aspect of the fraud triangle, it is the pressure or the "need" felt by the person who commits the fraud. It could be a real financial or other type of need, such as high medical bills or debts. Or it could also be a perceived financial need, such as a person who has a desire for the material goods but not the means to get those material goods.
2. **Perceived opportunity:** Opportunity can define as the ability to commit a fraud. Because fraudsters do not wish to be getting caught, they must also believe that that their activities will not be detected.
3. **Rationalization:** Employees may *rationalize* such behavior by determining that committing fraud is OK for a variety of reasons. Rationalization is the crucial component in most types of frauds.

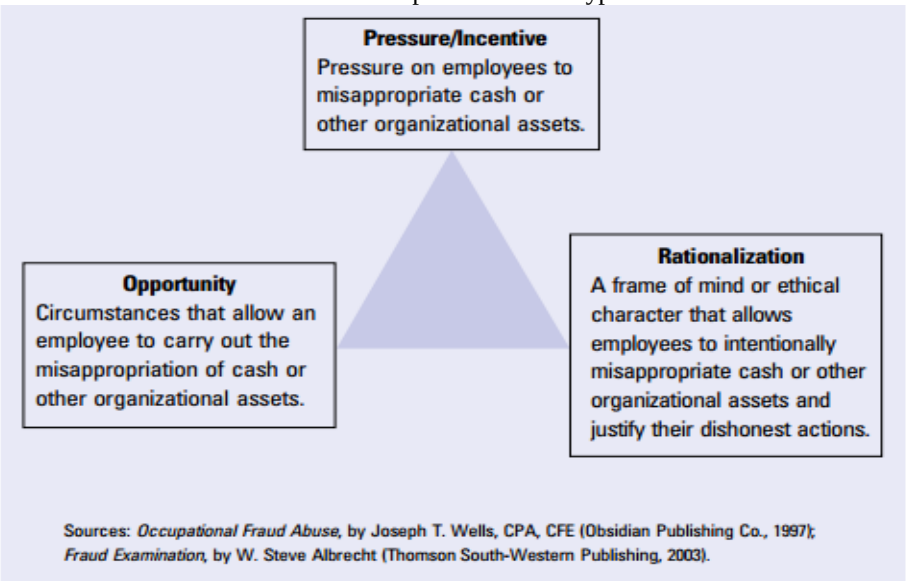

Figure 2 : The fraud triangle which was originated from the Donald Cressey's hypothesis

IV. COMMON TYPES OF INSIDER FRAUDS

Based on the person who committed the fraud, we can broadly classify them into two categories. When the fraud is committed by an outside person of an organization – say the customer – then it is called a Outsider Fraud. And when the fraud is committed by a person associated with organization – say the employee – then it is called an insider fraud.

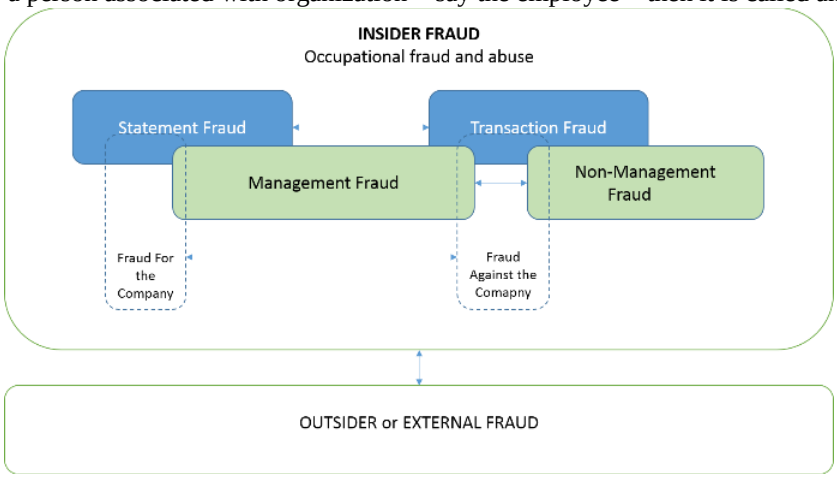

Figure 3 Relations between the fraud taxonomies

There may be ways in which an insider commits the fraud. Following are the some common areas in which most of the inside frauds are reported:

1. Billing- Frauds related to procurement and disposal of assets / Account Payable Fraud:

These are some kind of frauds which are common in every industry and having high commit rate. Banking is not an exception to these. Some of the types are Fictious Vendors, Altered invoices, Fixed Bidding, Goods not Received, Duplicated Invoices, Inflated Prices, Excess Quantities Purchased and Duplicate Payments and Duplicate Serial Numbers, Payroll fraud, Account Payables[8].

2. Corruption- Frauds related to Lending & financial transactions

These are frauds which can happen only in a financial institution like financial institutes. Banks basically being the custodian of people's money, and their main source of income is through leading the money to needy. To do this they have been authorized to exercise some financial powers. But some official miss-use these power. Recent incident of CMD of Syndicate Bank, bribe-for-loan case is a good example of for the depth of corruption in the banks. There are many other ways of coming frauds, by exercising financial powers. For instance, BGL frauds, TDS frauds, misuse of unclaimed deposits, misuse of dormant accounts, ATM Related frauds etc. An interesting point to be noted here is, we might think that, in the days CBS, these kind of frauds are committed by the individuals having good system knowledge. But, the fact is these frauds are committed by ordinary officers. But due to lack of efficient employees or mechanism to identify such frauds, has led these frauds escape un-noticed.

3. Financial Statement Frauds

Financial Statement frauds or Window dressing is one of the major problems in banks. Because each branch acts as a separate entity and each branch has to achieve their targets. Some branches adopt wrong ways to meet the target, which intern causes wrong Financial Statement at the head office level. Some ways of doing that are, transferring funds from unutilized CC accounts to SB/CA's to show positive growths in deposits and advances, transferring some amount to an overdrawn CC account, to make it below the limit, intern avoiding the account becoming NPA, Raising funds for short term, such that at the end of the Financial period the figures show positivity, Postponing the remittance to government account etc.

4. Expense Reimbursement - Frauds in availing facilities such as Entertainment, travels expense, out of pocket expense, Leave travel concession:

Banks provide many facilities for their employees in view of smooth functioning of banks. For example: Entertainment allowance, which can be utilized to make expense on offering of Soft drinks, coffee, tea, snacks to valuable customers, LTC(Leave travel concession), which is provided to go for a tour with the family. In the case of LTC (Annual Report – CVC – 2013) the modus operandi adopted includes use of forged/fake Air India tickets & boarding passes, claiming irregular reimbursements and in many cases officials have not travelled at all. The officials indulge in irregular claims like travelling by flexi/easy fare tickets by Air India and receiving cash discounts for the difference between LTC fare and flexi/easy fare from travel agents. In PSEs and banks, the LTC facility is allegedly used by officials for visiting abroad in collusion with certain airlines and travel agents. Commission also noticed instances where officers of Public Sector Banks visited foreign destinations and thereafter visit the designated place in India using a circuitous route on flexi/easy fare and claims were settled on the basis of full fare of entitled class to the designated placed in India. The guidelines and interpretation of 'circuitous route' were being abused in many cases. Instances of receipt of cash discounts

from travel agents have also been observed. In case entertainment allowance or out of pocket expenses, we can see examples of Over Limits, unusual or inappropriate expenses, miscellaneous/sundry expenses, split or duplicate expenses.

5. Insurance related frauds - Cross-Selling or Miss-selling?

Cross-Selling means encouraging a customer who buy product to buy a related or complementary product, with a view to expand banking business, reduce the per customer cost of operation and provide more satisfaction and value to the customer.

What is mis-selling? As per the IRDA, mis-selling can be defined as:

'By definition, mis-selling means selling a product by giving a wrong picture of a product, it may include, giving wrong information, giving unrealistic information, not giving full information about the product. You must have heard an insured, saying – but this was not I asked for. And, your agent accusing, but then I did mentioned all the details upfront, didn't I? Insurance is a business of selling commitments and here is a case where this was broken. Unfortunately the product was mis-sold. Mis-selling is not unique to insurance and happens in various lines of businesses (loans, credit cards, investment products, pharmacy, hospitality etc.), but Insurance being an intangible service – the principle of Caveat emptor prevails in insurance'. The **table 2** shows the complaints received by IRDA on 'UNFAIR BUSINESS PRACTICES'. We can observe from the statistics the complaints are increasing and which implies the mis-selling is also increasing.

As rightly said by Mr.P.Chidambaram, in an interview : "The reason why insurance is stumbling in India is because of the mis-selling of products and complex products. Also, If you want to sell insurance to India, you must sell simple products and must make it absolutely clear to the agents and to other officers that they should not mis-sell." LIVE MINT, FEB 12, 2013

6. Money Laundering:

Out of 140 countries India is ranked 93rd, 70th & 88th in 2012, 2013 & 2014 respectively compared to Norway, which has ranked 1st in Anti Money Laundering (AML) Basel Index[9]. This clearly show that India, in the present scenario is very vulnerable to money laundering activities and is a high risk zone. Money laundering refers to conversion of illegal or black money in such a way that it appears that it is obtained from a legitimate source. In India it's done through a system called "Hawala" which means transfer of money. It's done in 3 stages called, Placement, Layering and Integration. Banks are used intensively in the 2nd stage, i.e., layering where fund collected is channelized through different instruments. Usually they use many fictitious account created. Even though strict implantation of KYC (Know your customer) is mandatory as per the guidelines of RBI, its difficult implement the same in rural areas. Fraudsters, with the help of corrupt bankers exploit these loop holes to run their show.

7. Identity theft :

Identity theft is committed by accessing personally identifiable information of individuals/ entities without their permission with the objective to misuse this information and gain undeserving benefits. One of the primary reasons for growth in identity theft has been the proliferation of the Internet. In India, the number of internet users has grown from 7 million in 2001 to over 98 million by 2011 and is expected to reach the 300 million mark by 2015. However, controls and regulation aimed at protecting privacy have not kept pace with this growth. Fraudsters are making use of these gaps in controls to target individuals and organizations and misuse confidential data. According to the Norton Cybercrime Report 2011, four out of five online adults in India were victims of identity theft in 2011. Considering that many employees in the corporate work force use their office laptop/computer for personal transactions (such as online banking, shopping, payment of bills etc), identity theft can compromise not just their private information, but also the companies they work for. Banks can also be a source of identify, as banks will have all the personal/confidential information of an individual.

It's not only customers' data, we have seen the incidents of employees' identity theft. It's more dangerous than identity theft of customer data. As the thief can use that for fraudulent transactions and can easily escape. Hence it is very much essential for banks to identify & prevent the chances of thefts.

V. EXISTING MECHANISHMS FOR FRAUD DETECTIONS:

Before understanding the challenges we need know, at present, how a fraud is detected in Banks? Following are the some of the ways in which a fraud is detected:

1. Whistle Blowers ("Public Interest Disclosure and Protection of Informer" (PIDPI))/ anonymous complaints.
2. Internal Audits or central statutory audit
3. By Accident/ Random Verification of Vigilance Department

According to the 'India banking fraud survey – 2012' conducted by Deloitte, 53% of the responded that frauds were discovered through internal audit reviews. Significant majority of the incidents were detected though a formal or informal complaint mechanism. Around 43% and 37% respondents responded that frauds were detected trough anonymous complaints and whistle blower mechanism respectively. But shocking outcome is more than 20% frauds were discovered through accidents are random verification of vigilance department. [10]

VI. CHALLENGES FACED IN FRAUD DETECTION:

The existing mechanisms explained above have inbuilt deficiency against fraud detections:

1. Most of the frauds were discovered at least after six months and out of 23% discovered after a year [11]. As we know the longer the duration the larger the impacts. Usually the insider fraudster uses the 'low and slow' approach to fraud accomplishment. In other words, the insiders stole "low" amounts of money and conducted their activities "slowly" over a long period of time, possibly to avoid detection.

   'The lower 50% of cases (under 32 months in length) had an average actual monetary impact of approximately $382,750, while the upper 50% (at or over 32 months in length) had an average actual monetary impact of approximately $479,000.' [12]

   -Insider Fraud in Financial Services, 2012, Carnegie Mellon University
2. Internal audits and Vigilance verifications are done randomly. Because each of the branches will have 1000's of records and most of the banks have undergone computerization and CBS, It's impossible to verify each and every record manually. But fraud doesn't happen randomly. The fraudster plans in such a way that it can't be detected in random check.
3. Lack of expertise is also a very important reason why frauds go undetected. Because most of our bank official/internal auditor doesn't know how a CBS really work. How to verify the audit trail. How to identify a identify theat. This give's lot of room for a fraudster to escape.
4. In some of the banks the Vigilance departments are trying to leverage technology to identify the frauds. But most of the focus is on the traditional rule based, descriptive queries and analytics. Most of the organizations use tools such as spreadsheet tools such as Microsoft Excel and database tools such as MS Access and MS SQL Server. While these tools are important in every data analytics program, they often focus on the matching, grouping, ordering, joining or filtering of data that is primarily descriptive in nature. But generally these tools are slow and less efficient in fraud detection. Because these tools are not meant for that. There more sophisticated tools based on data mining techniques' which can provide far better results.

VII. DATA MINING TECHNIQUES: A GENERAL PERSPECTIVE

The process of deriving/discovering useful patterns from a large database is known as data mining.

Data mining is the set of methods and techniques for exploring and analyzing data sets (which are often large), in an automatic or semi-automatic way, in order to find among these data certain unknown or hidden rules, associations or tendencies; special systems output the essentials of the useful information while reducing the quantity of data. Briefly, data mining is the art of extracting information – that is, knowledge – from data.

Data mining is therefore both descriptive and predictive: the descriptive (or exploratory) techniques are designed to bring out information that is present but buried in a mass of data (as in the case of automatic clustering of individuals and searches for associations between products or medicines), while the predictive (or explanatory) techniques are designed to extrapolate new information based on the present information, this new information being qualitative (in the form of classification or scoring4) or quantitative (regression).[13]

Data mining process occurs in different stages[14]. It starts with Selection, followed by pre-processing, transforming, data mining and finally interpreting and evaluating.

The **selection** process involves, the task of acquiring only that part of an enormously large database required for the purpose of our business. Following the selection process some **pre-processing** of raw selected data is required, which is in turn followed by **transforming** the data to a suitable from to mine it. After all the pre requisite steps, actual **data mining** is done to give useful results. The results are further on **interpreted and evaluated** to get required knowledge about the data mined. The knowledge which is acquired at the end of all stages can be used for decision making.

Data Mining Techniques: There have been lot of data mining techniques employed generically for various field of studies. To limit the scope of this study we will consider, widely cited, top 10 data mining algorithms [15]. Those are:

1. **C4.5 and beyond:** Systems that are used to construct classifiers are one of the most commonly used tools in the field of data mining. Such systems take the input as a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can be used to accurately predict the class to which a new case will belong to.
2. **The k-means algorithm:** k-means clustering is the method of vector quantization which originates from signal processing that is very popular for cluster analysis in data mining. K-means clustering aims in partitioning n observations into k clusters in which, each observation belongs to a cluster with the nearest mean, serving as a prototype of cluster.
3. **Support vector machines:** Distinct versions of SVM will use different kernel functions to handle different types of the data sets. Linear and Gaussian kernels are well supported. SVM classification will attempt to separate the target classes with widest possible margin. SVM regression also tries to find a continuous function, such that the maximum number of data points will lie within an epsilon-wide tube around it.
4. **The A-priori algorithm:** A-priori is the algorithm for frequent item set mining and association rule learning over the transactional databases. A-priori proceeds by identifying frequent individual items in the database and extending them to much larger item sets as long as those item sets appear sufficiently often in database. The frequent item sets determined by this algorithm can be used to determine the association rules which highlight general trends in database. A-priori algorithm has applications in domains such as market basket analysis.
5. **The EM algorithm**: Finite mixture distributions will provides a flexible and mathematical-based approach to modeling and the clustering of data observed on few random phenomena. We focus here also on use of normal

mixture models, which will be used to cluster continuous data and to also estimate the underlying density functions. These mixture models can be fitted by the maximum likelihood via the Expectation–Maximization (EM) algorithm.

6. **PageRank**: PageRank algorithm was presented and published by Sergey Brin and Larry Page at the 17th International World Wide Web Conference (WWW7) in Apri-1998. Page Rank is a search ranking algorithm using hyperlinks on Web. Based on PageRank algorithm, they built the search engine Google, which has been very huge success.
7. **AdaBoost:** The AdaBoost was proposed by Yoav Freund and Robert Schapire is one of the most important ensemble method. It has a very solid theoretical foundation, very accurate prediction, great simplicity (Schapire said about AdaBoost, that it needs only "just 10 lines of code"), and wide and a successful applications.
8. **k-nearest neighbor classification:** k-nearest neighbor classification(Knn) , finds a group of k objects in training set that are closest to a given test object, and bases the assignment of a label on predominance of a particular class in this neighborhood. There are 3 key elements of this approach: a set of labeled objects, Ex: a set of stored records, a distance or similarity metric to compute the distance between objects, and value of k, the number of the nearest neighbors. To classify an unlabeled object, distance of this object to the labeled objects is first computed, its k-nearest neighbors are then identified, and the class labels of the nearest neighbors are finally used to determine the class label of object.
9. **Naive Bayes:** Naive Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence, as observed in the data.
10. **CART:** 1984 monograph, "CART: Classification and Regression Trees," which was co-authored by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone, represents the major milestone in evolution of Machine Learning ,Artificial Intelligence, non-parametric statistics, and the data mining. The work is very much important for the comprehensiveness of its study of the decision trees, the technical innovations it had introduced, its sophisticated discussion of the tree structured data analysis, and finally its authoritative treatment of large sample theory for trees. The CART citations can be found in almost any domain, far more appear in the fields such as electrical engineering, medical research , biology, and financial topics than, for example, in marketing research or in the field of sociology where other tree methods are more popular.

We can broadly classify above mentioned algorithms into six main categories based on the type of their objectives and also based on their learning modes we can classify them as supervised and unsupervised learning:[16]

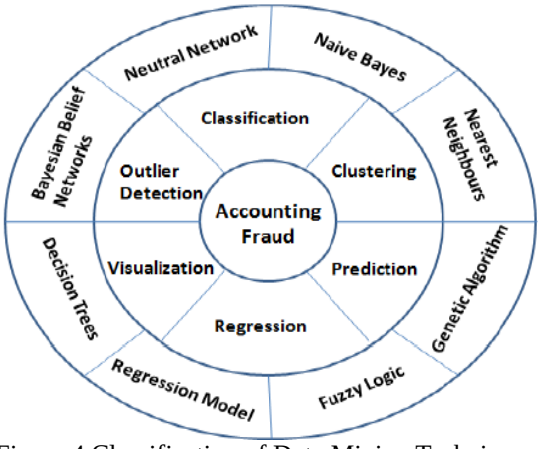

Figure 4 Classification of Data Mining Techniques

VIII. DATA MINING TECHNIQUES FOR FOR INSIDER FRAUD DETECTION IN BANKS

Internal Fraud detection framework:

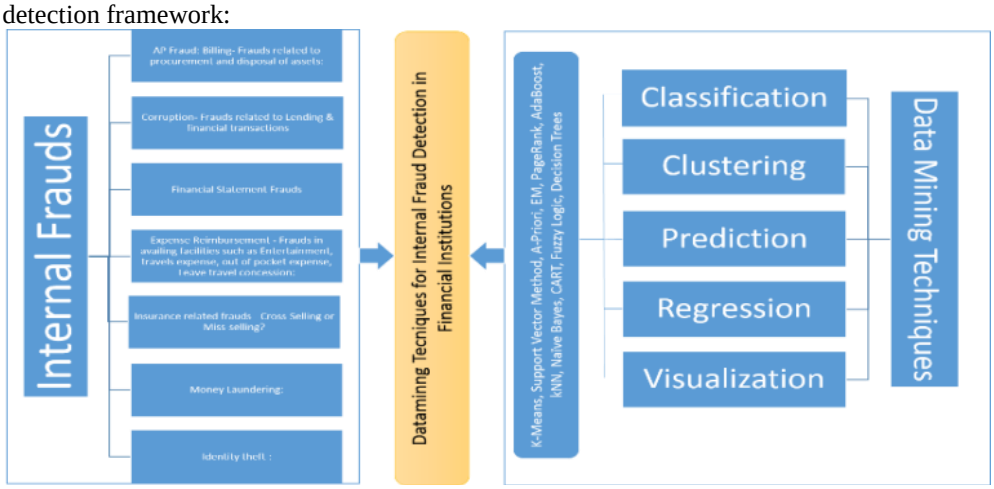

Figure 5 Internal Fraud Detection Framework

Steps in Data Mining for fraud detection:

We can use the following steps while applying the data mining techniques for fraud detection

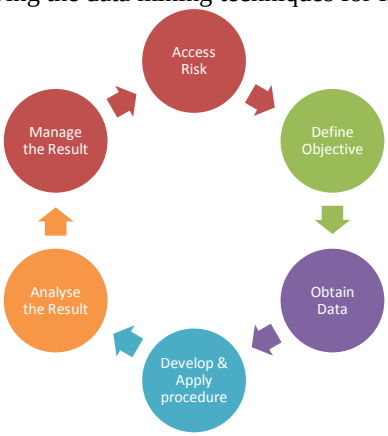

Figure 6: Steps involved in Data Mining for Fraud Detection

VIII. TECHNIQUES TO BE USED TO DETECT THE ABOVE FRAUDS

1. Benford's Law to detect financial transaction frauds

Benford's Law [17] (which was mentioned in 1881 by the astronomer Simon Newcomb) states : "if we randomly select a number from a table of physical constants or statistical data, then the probability that the first digit will be a "1" is about 0.301, rather than 0.1 as we might expect if all the digits were equally likely." In general, this law says that probability of the first digit being a "d" is

$$P\{d\} = \frac{\ln\left(1+\frac{1}{d}\right)}{\ln(10)}$$

Where 'ln' refers to natural log (base e). This numerical phenomenon was first published by Newcomb in a paper with title "Note on the Frequency of Use of the Different Digits in Natural Numbers", which appeared in The American Journal of Mathematics ,1881, 4, 39-40. It was re-discovered by Benford in 1938, and Benford published an article called "The Law of Anomalous Numbers" in Proc. Amer. Phil. Soc 78, pp 551-72.

We can use the same concept to detect the money laundering. If we analyze our payments, we can compare the frequencies of the amount digits. Then we can identify the digits which are violating the normal behavior. For example, Benford's conclusion is that, the probability of digit "1", repeating is 30%. And similarly the number "8", is 5%. If, we see more transactions with the higher numbers say "8" or "9", then we can conclude that, the transactions may not be legitimate, and we can examine those transactions in depth to identify the possibility of AML. Fraudsters will often create an amount that starts with a higher number, like 8 or 9, not knowing that auditors are now equipped to identify these abnormal payments. We can also use these techniques to identify the fictitious payments, while examining account payable frauds also.

2. Using fuzzy logic to identify AP Fraud

Paper [17] suggests implementing "similar fuzzy-matching" instead of exact matching yields an approach more accurate and powerful than many.

For instance, we can consider invoice numbers are similar if they appear same after removing the leading and trailing zeros. Or any alphabets as well as punctuation characters.

Invoice dates are considered same if the difference between the dates are less than the threshold limits such as one week. For example if you set threshold has one week then all the invoice falling within the difference of one week is considered as similar. They have suggested that, the generally threshold would be 3 weeks to detect the duplicate payments. This would often eliminates the usual monthly payments, such as rent, electricity bills etc.

They suggest to use one of the following criteria to detect the similarity among the amounts:

- The amounts compared are having difference of ±5%
- The amounts are multiples of each other. I.e. say one amount is Rs.240 and other is 480.
- The matching first 4 digits of the amounts, i.e. Rs. 123.45 and Rs. 1,234.55

We can use similar matching techniques to detect the other account payable frauds and duplicate payment frauds while doing the internal audit.

3. Cluster Based internal fraud detection

The survey paper [18], reviews 27 articles on Cluster based fraud detection techniques. The paper concludes that Compared to other domains where clustering is being applied to identify outliers, intrusion detection, etc.., clustering

based fraud detection techniques tend to use established clustering techniques. The paper states that based on the surveyed papers, we can observe that almost three quarters of the employed clustering techniques were partitional. And some papers, [19], [20] were using partitional clustering techniques with hierarchical clustering techniques as combined approach. Among partitional clustering methods, k-means clustering and its variants, along with Euclidian distance as dissimilarity metric are the one of the most commonly used ones. Hierarchical clustering techniques come in the second place being used in one quarter of the surveyed papers. Other methods like, visualization, interactive clustering techniques were also used, but in only in very small number of cases. Further, the paper speaks about the way the clustering techniques were used or combined in conjunction with other mining methods. As per that, the some were using single algorithms, some were using combined techniques by employing two or more clustering and some were using hybrid techniques, bust mostly were classifier based on decision trees, neural network and SVMs. These all were efficient in detecting the frauds and we can employee similar methods to detect all the insider frauds discussed in the section 3.

4. Data mining techniques to detect financial statement fraud

This is one such an area where an extensive work has done. Because, this kind of fraud can directly affect the economy of the country. The Research paper [21] explores the effectiveness of Data Mining (DM) classification techniques in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS. The paper suggest that to accomplished the taks of detecting the management fraud, the auditors could be facilitated with the data mining techniques in their work. The paper investigated the usefulness of Neural Networks, Decision Trees and Bayesian Belief Networks in the identification of fraudulent financial statements. The inputs for these techniques were the ratios derived from the financial statements. The paper found that these techniques are very efficient in detecting the FFS.

In the work[22] the review literature describes the use of data mining algorithms including statistical test, regression analysis, decision tree, Neural Network ,Bayesian network etc for the financial accounting fraud detection. The regression analysis is used widely for the fraud detection since it has a great ability of providing explanation. Some of the different regression methods employed by researchers are Logit, UTADIS, Step-wise Logistic and EGB2 etc. Neural Networks are very important tools for mining the data. Advantages of the Neural Network are that there are no strict requests for data, it has a strong generalization and adjustment. After the correct allocation and the proper training, Neural Network may perform great classification comparing with regression model.

The paper[23] reviews outlier in the financial statement to detect the financial statement frauds. The paper recommends a framework to detect the financial statement frauds and states that,

"Considering the dire need of analytical tool for prevention and detection of financial statement fraud, we presented a data mining framework for financial statement fraud risk reduction. This framework has following major contributions. The core of this framework is data mining as it follows the traditional information flow of data mining. Our framework recommends the use of association rules, a descriptive data mining technique, for preventing fraudulent financial reporting along with the use of predictive data mining techniques such as classification for successful identification and detection of financial statement fraud. Rule engine module of our framework will generate interesting association rules, which will be used by rule monitor in order to raise an alarm regarding financial statement fraud. Hence, our framework is capable of preventing the fraudulent financial reporting at the first place and detection of fraud once prevention mechanism is failed."

By looking at the research work, we can conclude that, the data mining techniques can employed efficiently by the auditors to detect the financial statement frauds.

5. Data mining application for detecting money laundering:

Currently RBI is monitoring all the high amount transaction (>10lac) to detect the possibility of money laundering. Which is kind of prediction methodology for detection of money laundering. There are similar methods used across the word to detect the money laundering. Such methods are discussed in the paper [24]. The Financial Crimes Enforcement Network AI System (FAIS) operates with the expert system using Bayesian inference engine to detect the suspicion scores and with the link analysis is used to visually examine the selected transactions or accounts. Supervised techniques such as case based reasoning, nearest neighbor retrieval and decision trees are used due to propositional approaches, lack of clearly labelled positive examples, and scalability issues. The unsupervised techniques were avoided due to the difficulties in deriving appropriate attributes. It has added effectiveness in the manual investigations and has gained insights of the policy decisions towards money laundering.

The paper [25] discusses a DM based framework to help AML. The frame suggests following methods to detect the ML.

- Using Bayesian inference and decision trees to detect the suspicious transactions based on probability computations so as to help anti money laundering focus on the most likely suspected transactions.
- Using link analysis, consolidation, social network etc, to detect central members, subgroups and intergroup transactions patterns in Money laundering network.
- Employing regression and case study based reasoning to discover hidden leads and patterns that may prove important, timely and predicting prospective trends and
- Using support vector machine (SVM) to handle high dimensionality heterogeneous data sets.

6. Detecting Identity theft & Mis-Selling

These are some areas which are still in primitive stages. Not much noticeable work has done in these are areas. We can use some predictive algorithms as we discussed in AP frauds to detect the mis-selling. But more work/research need to be done in that area. One probable solution to avoid these kind of frauds is to educate the customer. Because these are the frauds which directly affect the customer and in most of the cases negligence/lack of awareness of the customer encourages the fraudster to commit these frauds.

IX. FUTURE WORK: A BIGGER PERSPECTIVE

Indian banking sector consisting of 157 scheduled commercial banks, having 109811 branches & 162543 ATM and serving 58.6% of the households. Everyday around Rs.2,00,000 crores of business transactions are done through banks. This numbers will increase exponentially in coming future, as our PM's Narendra Modi's ambitious PMJDY scheme, has added nearly 70 million accounts in just 4 months. With This much transactions and number of accounts, will also leads to more number of frauds and traditional data mining techniques discussed above may not work as efficiently as expected. Hence our future goal should to focus on improving the mining techniques to adopt to data intense scenarios or we can simply say it should BIG DATA ready.

From past few years there is buzz around the word big data. And also there is a confusion about the definition of the big data. The survey paper [26] defines big data with four dimensions as " Four V's – Volume, Velocity, Variety & Veracity". The same four V's can be applied to Indian banking sector also.

1. Volume: With increasing customer base and having a 1.1 billion population, certainly the amount data produced every day will be huge. Hence the banking sector need to gear up to process massive amount of data.

2. Velocity: Having, having 109811 branches & 162543 ATMs, the number of transaction per day is around Rs. 2,00,000 crores. Implies the velocity of the inflow of data is immense.

3. Variety: Latest technologies have brought different MIS systems to banking industry, CBS, ATMs, CRM, ERPs, LMS, HRMS and so on. Each application produces data in different formats. Some are structured and some are unstructured. Hence handling variety is also new challenge.

4. Veracity: Some data is inherently uncertain, for example: sentiment and truthfulness in humans; Uncertainty manifests itself in banking in many ways. It is in the skepticism that surrounds data created in human environments like social networks; in the unknowingness of how the future will unfold and of how people, nature or unseen market forces will react to the variability of the world around them. The future technology need to acknowledge and embrace this uncertainty.

X. CONCLUSION

This paper helps us to understand what a fraud in banking sector is. It goes on to categorize different types of frauds, their definitions, factors affecting them and the challenges faced in detecting them. The paper lists out different data mining techniques with their generic use, also with respect to the insider fraud detection, we have listed out and explained the best available data mining techniques, proposed by many researchers and currently employed in different industries. As a part of future work we have thrown light on big data perspective of banking and fraud detection. Finally we can conclude that fraud detection and prevention is the prime requirement of the banking industry and the data mining techniques can really come in handy in reducing the fraud cases. We can use any or all the methods explained above to achieve the target.

TABLE 1: CVO 2013 - FRAUD PROCEEDINGS STATISTICS OF PSUS

Cases where punishment Awarded	Cut in Pension	Dismissal/ Removal/ Compulsory Retirement	Reduction to lower time scale/ rank	Other Major penalties	Minor penalties other than censure/ warning	Censure/ Warning	No action	Total
Total	212	1097	2267	1530	782	441	599	6928
Of Which Banks	37	603	990	1160	316	186	139	3431

TABLE 2: 'UNFAIR BUSINESS PRACTISES' COMPALINTS 2012-13

Segment	2011-12	2012-13
Conventional	55896	89384
Health	888	691
Pension	1592	1635
ULIP	36702	42596
Others	5722	34174
Totals	100800	168480

Table 1: list of tests to can be used to detect ap fraud [27]

Sl.No	Type of Fraud	Tests Used to Discover This Fraud
1	Fictitious vendors	Compare the postel address of Vendor and employees to identify the matching entries
		Check for the vendor name, which sounds similar or compare the address and phone numbers
2	Altered invoices	Search for duplicates
		compare invoice amounts and contract/purchase orders
3	Fixed bidding	Compare vendors for over the years and check if the same vendor is winning the most of the contracts
		Compare the bidding close dates and contract submission to check if the late bidder is winning consistently
4	Goods not received	look for purchase quantities which does not match with order places
		Monitor the levels of inventory and goods delivered for appropriate propositions
5	Duplicate invoices	Check for duplicate date, invoice numbers and amounts
6	Inflated prices	Compare the prices of the vendors with the market price, for unreasonable difference
7	Excess quantities purchased	Check for unexxplained increase in the inventory of good
		Determine if the quantities of raw material purchased is proposinate to the quantity produced
8	Duplicate payments	Check for similar invoice number and payment amounts
		seach for frequest request for refunds for invoices paid already
9	Carbon copies	Check all the checks cashed and look for gaps in the check numbers
10	Duplicate serial numbers	Check for the repurchase of already pruchased items and for same vendors
11	Payroll fraud	Check for all the terminated employess, to the employees on the payroll for the matches and extract all the payments made to them
12	Accounts payable	link AP files to contract, inventory files to examin the trasactions not matching to contract amounts, contract date, prices, quantity etc

REFERENCES

[1] THE BUSINESS STANDARD, 29th August, 2014

[2] http://en.wikipedia.org/wiki/Banking_in_India#cite_note-rbiPublication20131121-8

[3] 'India fraud survey' conducted by KPMG in 2012,

[4] 'Frauds ripped public sector banks of Rs. 23,000 crore' - Sandeep Pai and Mahua Venkatesh, Hindustan Times, New Delhi, January 30, 2014

[5] 'A Game of Shadows- The recent bribe-for-loan scandal….' The Business Today, Published on August 28, 2014

[6] Annual CVO Report – 2012 – Chief Vigilance Office, India

[7] 'Fraud Detection in the Banking industry', Discussion Paper, ACL *http://www.acl.com/pdfs/DP_Fraud_detection_BANKING.pdf*

[8] Donald R. Cressey, *Other People's Money* (Montclair: Patterson Smith, 1973) p. 30.

[9] Public Basel AML Index scores – 2012, 2013, 2014 – index.baselgovernance.org Coderre, David G. Computer Aided Fraud Prevention and Detection: A Step-by-Step Guide. John Wiley & Sons, 2009.

[10] 'India banking fraud survey – 2012' by Deloitte

[11] Navigating the Challenging environment - India Banking Fraud Survey – 2012

[12] Insider Fraud in Financial Services, 2012, Carnegie Mellon University

[13] Data Mining and Statistics for Decision Making - Paolo Giudici et al – Wiely Series

[14] Top 10 algorithms in data mining – IEEE ICDM '06

[15] A Review of Financial Accounting Fraud Detection based on Data Mining Techniques, Anuj Sharma, Prabin Kumar Panigrahi, International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012

[16] 'Accounts Payable Fraud: Ten Ways to Identify it' By Christine L. Warner
[17] Survey of Clustering based Financial Fraud Detection Research, Andrei Sorin SABAU, Informatica Economică vol. 16, no. 1/2012
[18] Rui Liu, Xiao-long Qian, Shu Mao, and Shuai-zheng Zhu, "Research on antimony laundering based on core decision tree algorithm,", Control and Decision Conference (CCDC), 2011 Chinese, 2011, pp. 4322-4325.
[19] C. Holton, "Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multibillion dollar problem," Decision Support Systems, vol. 46, no. 4, pp. 853– 864, 2009.
[20] Data Mining techniques for the detection of fraudulent financial statements, Efstathios Kirkos a,1, Charalambos Spathis b,*, Yannis Manolopoulos, www.elsevier.com/locate/eswa
[21] A Data Mining Framework for Prevention and Detection of Financial Statement Fraud, Rajat Gupta & Nasib Singh Gill, International Journal of Computer Applications (0975 – 8887) Volume 50 – No.8, July 2012
[22] A Comprehensive Survey of Data Mining-based Fraud Detection Research, CLIFTON PHUA1*, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2, http://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf
[23] Senator, T., Goldberg, H., Wooton, J., Cottini, M., Khan, U., Klinger, C., Llamas, W., Marrone, M. & Wong, R. (1995). The Financial Crimes Enforcement Network AI System (FAIS). AAAI16(4): Winter, 21-39
[24] A framework for data mining-based anti-money laundering research - Zengan Gao and Mao Ye - www.emeraldinsight.com/1368-5201.htm
[25] Analytics: The real-world use of big data - How innovative enterprises extract value from uncertain data – IBM – 2012
[26] Fraud Detection, Using Data Analysis Techniques to Detect Fraud, Coderre, David G. Fraud Detection, Using Data Analysis to Detect Fraud, (Vancouver, BC: Global Audit Publications, 1999): 50-202

