ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

Pattern Analysis and Fraud Detection Using Hadoop Framework

S.V.Phulari, Umesh Shantling Lamture, Sumit Vilas Madage, Kunal Tirupati Bhandari

Abstract— An improved approach for identifying patterns and detecting a bank fraud in any organization and to investigate the process for bank frauds and to the proper implementation of preventive security controls in banking industry transactions. According to RBI records , 22 million of the 589 million bank account holders use mobile banking apps.The volume of mobile banking transactions has risen from around 18,190 million INR in 2011–12 to approximately 1,018,510 million INR in 2014–15.Regular check on transaction needs to be done to avoid frauds.

Index Terms- **Hadoop, Sqoop, Hidden Markov Model (HMM), Map Reduce, HDFS.**

I. INTRODUCTION

In today's electronic world e-commerce has become an essential mode for global business. Electronic commerce, commonly known as e-commerce or e-commerce, is a type of industry where the buying and selling of products or services is conducted over electronic systems such as the Internet and other computer networks. According to Nielsen study conducted in 2008, 1/10th world's total population has been using internet for shopping and transaction. The most common method of payment for online purchase is credit card. As number of credit card user's increases daily there is rhythm in the people's life and the same time the credit card fraud ratio is also increases. For which crimes involving in credit card are increasing that disturbs the organization financial order seriously and hence there is a great loss to bank and card holder that affects the development of banks.

Credit card can be used to purchases goods and services using online and offline transaction mode. It can be divided into two types:
A. Physical Card
B. Virtual Card

In the physical card based purchase, card holder has to produce the card at the merchant counter and merchant will sweep the card in the EMV (Euro pay, MasterCard and Visa) machine. Fraud transaction can be happened in this mode, only after the card has been stolen. It will be difficult to detect fraud in this type of transaction. If the card holder does not realize loss of the card and does not report to police or card issuing company, it can give financial loses to issuing authorities. In the second method of purchasing i.e. online, these transactions generally happen on telephone or internet and to make this kind of transaction, the user will need some important information about a credit card (such as credit card number, validity, CVV number, name of card holder). To make fraud transaction to purchase goods and services, fraudster will need to know all these details of card only then he/she will make transactions. Most of the time, the cardholder may or may not know that when or where any person will be seen or stolen card information. To detect this kind of fraud transaction, we have proposed a Hidden Markov Model which is studying spending profile of the card holder. An HMM is to analyze the spending profile of each card holder and to find out any discrepancy in the spending patterns. Fraud detection can be detected on analyzing of previous transactions data which helps to form spending profile of the card holder. Every card holder having unique pattern contains information about amount of transactions, details of purchased items, merchant information, date of transaction etc. It will be the most effective method to counter fraud transaction through internet. If any deviation will be noticed from available patterns of the card holder, then it will generate an alarm to the system to stop the transaction.

A. Overview

In one of the real scenario, An Insurance based product offers policies to customers like Flexi funeral plan. Every policy is taken for different types of needs; therefore the conditions for your policy vary according to the Plan and Term. It also provides the commencement date on which the policy is issued, date of birth of policy holder, date of maturity of policy, premium due date and months in which the renewal premiums are to be paid etc. This

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

information validates while capturing the policy insurance product undergoes various customers screening to ensure the civil score do not indicate any implications. Current system works in the manner that it takes only basic information document of customer like CIF key, Name, organization, nationality as Identity on which basic screening is done to make person eligible for investment plan.

B. Brief Description

Insurance company is facing financial issue due to collection of Debit order payments timely basis resulting into rejections in payments because of insufficient amount in customers account. Debit order collections attempt (standing instructions) are being made on clients bank account due to insufficient balance i.e. the balance in client's bank account is not greater than actual premium amount for collection resulting rejection of debit order . Company is currently undergoing financial and reputational losses number of payment rejections coming across the policies and hence CRC rating is declining. CRC is Credit rating and collection generally engaged in credit rating, risk management and debt recovery business. It handles millions of customer data holding various policies and payment transactions that are being processed every day. At times processing huge data becomes challenging part. Traditionally, businesses have been able to store their data in local data centers comprised of a few machines. However, as data outgrows the capacity of local datacenters, Company is looking forward to rethink traditional data storage and retrieval models which indirectly proportional to the revenue growth and customer engagement.

C. Purpose

The purpose of this project is to will provide effective, efficient, quick and precise method to avoid the fraud. Fraud is a billion-dollar business and it is increasing every year. The PwC global economic crime survey of 2014 suggests that close to 30 percent of companies worldwide have reported being victims of fraud in the past year.

This analysis can help an organization interested in fraud detection build a knowledge base of fraud. The ultimate objective would be the creation of supervised learning model that is focused on uncovering fraudulent transactions. The project will helps us to guess "Fraud Detection" before it happens. This will save million dollars of organization in a better way.

D. Scope

To the Bank/Insurance product: Debit order collecting rating is at decline when a customer doesn't sufficient fund in his/her bank account. Having a policy holding by that customer, standing premium instruction will attempt to debit policy premium amount from customer's bank account. With insufficient balance in the account, Debit order rejections will happen, resulting payment not successful.

To the Customer: Missed premium also affects customer to repay the penalty, Overhead the burden of paying such premiums if the improper planning on the expenditure continues.
Faster Retrieval of customer Data: Customer experiences slowness in accessing data due to the huge no of records consumes time for processing and retrieval in faster manner.
Access Customer data from anywhere: Customer experience the space issue, they require storing their data file on the local hard drive and could not accessible from anywhere.

E. Applying software engineering approach

Incremental Model

The product is decomposed into a number of components, each of which are designed and built separately (termed as builds). Each component is delivered to the client when it is complete. This allows partial utilization of product and avoids a long development time. It also creates a large initial capital outlay with the subsequent long wait avoided. This model of development also helps ease the traumatic effect of introducing completely new system all at once. There are, overall, few problems with this model.

The incremental build model is a method of software development where the model is designed, implemented and tested incrementally (a little more is added each time) until the product is finished.

It involves both development and maintenance. The product is defined as finished when it satisfies all of its requirements. This model combines the elements of the waterfall model with the iterative philosophy of prototyping.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

The product is decomposed into a number of components, each of which are designed and built separately (termed as builds). Each component is delivered to the client when it is complete. This Customizable Indoor Location And Navigation System Based On Bluetooth allows partial utilization of product and avoids a long development time. It also creates a large initial capital outlay with the subsequent long wait avoided. This model of development also helps ease the traumatic effect of introducing completely new system all at once. There are, overall, few problems with this model.

Benefits-

1. Any faulty piece of software can be identified easily as very few changes are done after every iteration.
2. It is easier to test and debug as testing and debugging can be performed after each iteration.
3. This model does not affect anyone's business values because they provide core of the software which customer needs, which will indeed help that person to keep run his business.
4. After establishing an overall architecture, system is developed and delivered in increments.

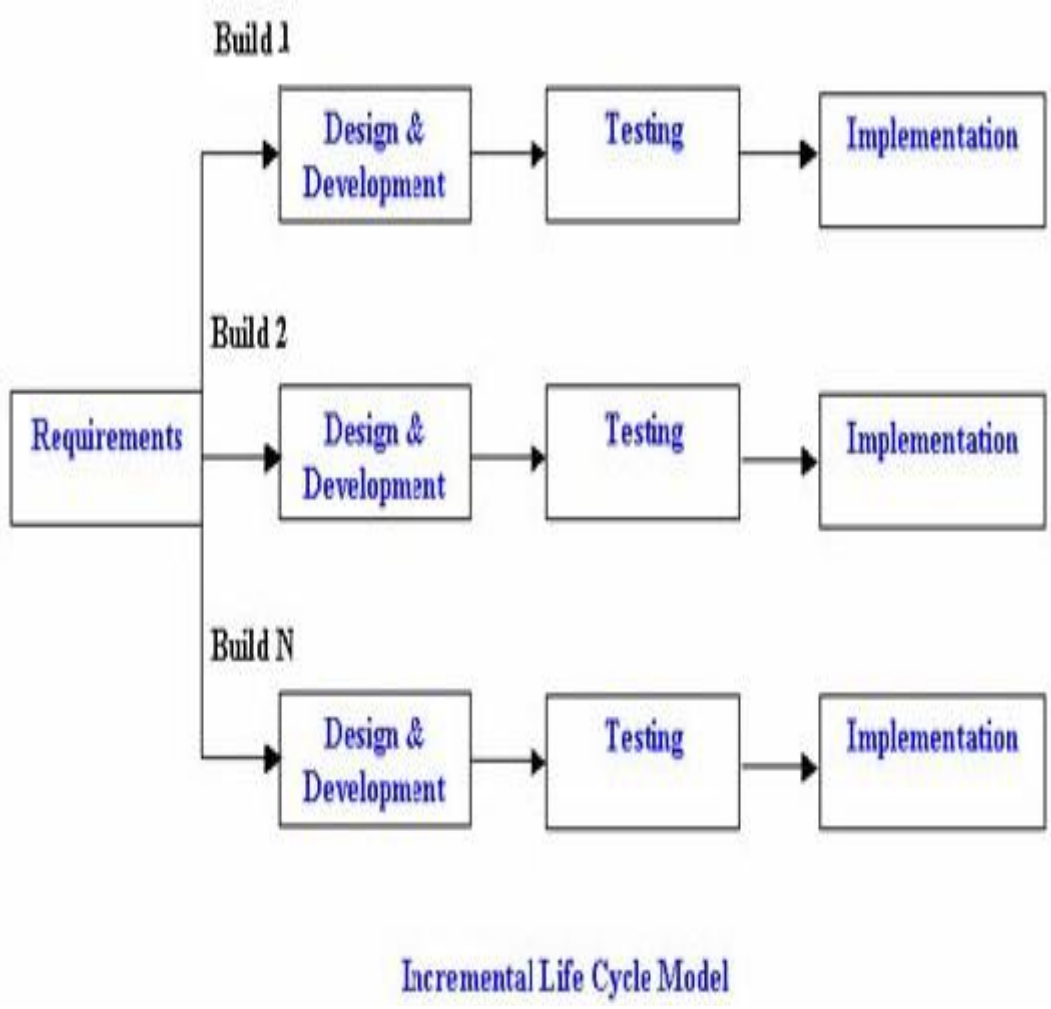


Fig 1: Incremental Model

F. Steps

The steps in the hadoop and data mining process are:

- Data collection and enhancement :

  This step involves joining the multiple data sources into a flat file structure. This step sometimes requires that decisions be made on the level of measurement (e.g. do some data get summarized.) As data from disparate sources are joined, it may become evident that the information contained in the records is

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

insufficient. For instance, it may be found that the data on vendors is not specific enough. It may be necessary to enrich data with external (or other) data.

- Loading data into Hadoop cluster :
  After collecting the data, we will load this data into Hadoop cluster. This data will store in distributed manner in data nodes. Once the data is loaded into cluster it is ready for analytics.
- Modeling strategies :
  Data mining strategies fall into two broad categories: supervised learning and unsupervised learning. Supervised learning methods are deployed when there exist a target variable with known values and about which predictions will be made by using the values of other variables as input. Unsupervised learning methods tend to be deployed on data for which there does not exist a target variable with known values, but for which input variables do exist.
- MR technique :
  Hadoop enables resilient, distributed processing of massive unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage. MapReduce serves two essential functions: It parcels out work to various nodes within the cluster or map, and it organizes and reduces the results from each node into a cohesive answer to a query.
- Training, validation, and testing of models :
  Model development begins by partitioning data sets into one set of data used to train a model, another data set used to validate the model, and a third used to test the trained and validated model. This splitting of data ensures that the model does not memorize a particular subset of data.
- Analyzing results :
  Diagnostics used in model evaluation vary in supervised and unsupervised learning. For classification problems, analysts typically review gain, lift and profit charts, threshold charts, confusion matrices, and statistics of fit for the training and validation sets, or for the test set. Business domain knowledge is also of significant value to interpreting model results.
- Implementing results :
  Clustering models can be evaluated for overall model performance or for the quality of certain groupings of data. Overall model diagnostics usually focus on determining how capable the model was in dividing the input data into discrete sets of similar cases.

G. Product Features

- ❖ Analyze and identify the no of rejections happened against client and policy level, this will help to understand the customer score card since the policy was issued. Based on score card customer in an insurance company will be rated good or bad customer in terms of paying premiums. Say for E.g.: Policy commencement date of a policy is 30 June 2014 and customer doesn't have any premium rejections so far, thus the customer score cord would be rated GREEN. But other side, If the Customer had multiple times premium rejections then he/she would be rated RED. And if customer had very few premium rejections then he/she would be rated AMBER
- ❖ Analyze the customer history transactions (Cr and Dr) and provide better Debit dates. Customer is failing to a pay premium on timely basis due to which so many rejections that the company is facing then the solution is to revise the payment day for the Debit order collection would minimize or prevent from no of premium rejections happening in the past. Say for e.g. A customer is having payment day on 1st of every month, but the salary in that customer's end on 10th , practically this might left with insufficient balance by the time of standing instruction trigger on 1st and that's why the number premium rejections implies with different other customers.
- ❖ Analyze the customer profile and perform screening / eligibility check. Recollecting to one of an example given above 1.1) that the Customer is has normally been screened based on the total CTC pa amount and not target CTC pa, Customer profile screening will validate any illegitimate case where a customer is over burdened or doesn't fit into eligibility criteria for policy premium payment that causes financial impact to both Insurance company as well as customer.
- ❖ Analyze and provide suitable future investments plan for the customer and increase the sales. Having Customer profile eligibility check performed on every customer, we would come to know the list of customers who were actually not eligible for investment contract but taken up the policy and due to which multiple premium rejections have been experienced. On other hand analyzing the customer's historic

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

transactions and its expenditure will come to know if the customer is remaining with sufficient balance in his/her bank account then insurers can offer suitable investment plan that will increase company's sale and customer engagement assurance.

II. LITERATURE SURVEY

Beirstaker, Brody, Pacini (2005) proposed numerous fraud protection and detection techniques. These various techniques include fraud policies, telephone hotlines, employee reference checks, fraud vulnerability reviews, vendor contract reviews and sanctions, analytical reviews (financial ratio analysis), password protection, firewalls, digital analysis and other forms of software technology, and discovery sampling.

Calderon and Green (1994) made an analysis of 114 actual cases of corporate fraud published in the Internal Auditor between 1986 and November 1990. They found that limited separation of duties, false documentation, and inadequate or nonexistent control account for 60 percent of the fraud cases. Moreover, the study found that professional and managerial employees were involved in 45 percent of the cases. Based on the findings, they recommended the following:

- To deter fraud, internal auditors should ensure that strong prevention systems based on the
- Fundamental principles of good internal control be established and used.
- To detect and investigate fraud, organizations must ensure the existence of strong internal audit
- Departments with sufficient resources to pursue the increased responsibilities faced by internal auditors.

To Commercial Angels‟ Newsletter (2001), the best way of preventing fraud was to understand why it happened. Fraudster generally identifies loopholes in control procedures and then assess whether their potential rewards will outweigh the penalties should they be caught. A regular control is most effective for prevention of frauds and normally requires little management time or effort. Prevention of frauds starts with identification of weakness in current systems of the organization. Next the organization must improve those systems with new or better controls. The introduction and enforcement of controls will reduce the opportunities for frauds. The control warns potential fraudster that the management is actively monitoring the business and in turns deters frauds. Education, training and awareness programm are informal intervention measures that should be implemented to prevent frauds.

Ganesh and Raghurama (2008), believe that training improves the capabilities of employees by enhancing their skills, knowledge and commitment towards their work In the survey conducted by them, about 80 executive from Corporation Bank and Karnataka Bank Ltd of India, were requested to rate their subordinates in terms of development of their skills before and after they underwent certain commonly delivered training programs. Responses revealed that for the seventeen skills identified there was improvement in the skills statistically. The paired t-test was applied individually for the seventeen skills, and all these skills have shown statistical significance. The seventeen skills include analytical skill, human resource skill, marketing skill, communication skill, accounting skill, credit appraisal skill, cash management skill, time management skill, inter-branch reconciliation skill, conceptual skill, information technology related skill, technical skill, role identification skill, problem solving skill, behavioral skill, risk management skill and customer service skill.

Haugen and Selin (1999) discussed the value of internal controls. Internal control system has four broad objectives: to safeguard assets of the firm, to ensure the accuracy and reliability of accounting records and information, to promote efficiency in firm‟s operations and to measure compliance with management prescribed policies and procedures. The effectiveness of internal controls depends largely on management's integrity. There are many other reasons for employee fraud, the more common being revenge, overwhelming personal debt, and substance abuse. Business today is very competitive, and employees often stressed. As a result, they have a feeling of being overworked, underpaid, and unappreciated. If employees are also struggling with serious personal problems, their motivation to commit fraud is very high. Adding to the situation of poor internal controls, the readily available computer technology also assists in the crime, and the opportunity to commit fraud becomes a reality.

Harris and William (2004) examined the reasons for loan frauds in banks and emphasized on due diligence program. This is a proactive approach, with each business line within the institution establishing policies and procedures for conducting due diligence investigations for both new and existing customers .They indicated that lack of an effective internal audit staff at the company, frequent turnover of management or directors, appointment

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

of unqualified persons in key audit or finance posts, customer's reluctance to provide requested information or financial statements and fictitious or conflicting data provided by the customers are the main reasons for loan frauds. Fraud thrives when conditions are right. A "fraud-friendly environment" is characterized by lax corporate culture on the enforcement of internal controls; deficient and/or absence of requisite risk controls, staff apathy and overconfidence.

Jeffords (1992) examined 910 cases submitted to the Internal Auditor during the nine-year period from 1981-1989 to assess the specific risk factors cited in the Tread way Commission Report. Approximately 63 percent of the 910 cases are classified under the internal control risks that include: lack of regular independent checks in performance, inadequate organizational control methods, inadequate methods of communicating or enforcing the assignment of authority and responsibility; and unauthorized access and physical control of assets, records, computer programs, or data.

Bhasin (2007) examined the reasons for cheques frauds, the magnitude of frauds in Indian banks, and the manner, in which the expertise of internal auditors can be integrated, in order to detect and prevent frauds in banks. He emphasized that though the head of the branch holds the responsibility for ensuring adherence to prescribed systems and procedures, the bank's internal auditors also occupy a special position in the detection and prevention of frauds. In addition to considering the common types of fraud „signals', auditors can take several „proactive' steps to combat frauds. Checking frauds requires training, account screening, signature verification and information sharing with regulators and local authorities. One important challenge for banks, therefore, is the examination of new technology applications for control and security issues.

Sharma and Brahma (2000) have emphasized on Banker's responsibility on frauds. They indicated that bank frauds could crop up in all spheres of bank‟s dealing, like cheque frauds, deposit account frauds, purchase bill fraud, hypothecation fraud, loan fraud, frauds in foreign exchange transactions and inter-branch account. Major cause for perpetration of fraud is laxity in observance in laid down system and procedures by supervising staff. Unscrupulous constituents commit frauds by taking advantage of the laxity on the part of the officials in observance of time-tested safeguards established by Reserve Bank of India (RBI). The RBI has set up an investigation cell in its central office. Ace investigator of high and vast experience mans it. The bank team goes deep into the root cause of bank frauds and suggests exhaustive preventive measures. The RBI carries out detailed studies and researches in the commission of bank frauds and recommends innovations to prevent frauds. The authors have further suggested that the need of the hour is not another piece of complex high-powered body of RBI, but analysis and concerted application of controls by bank management and their operational staff.

Smith (1995) offered a typology of individuals who embezzle. He indicated that embezzlers are "opportunist's type", who quickly detects the lack of weakness in internal control and seizes the opportunity to use the deficiency to his benefit. To deter embezzlement he recommended the following measures:-

- Institute strong internal control policies, which reduce the opportunity of crime.
- Conduct an aggressive and thorough background check prior to employment.

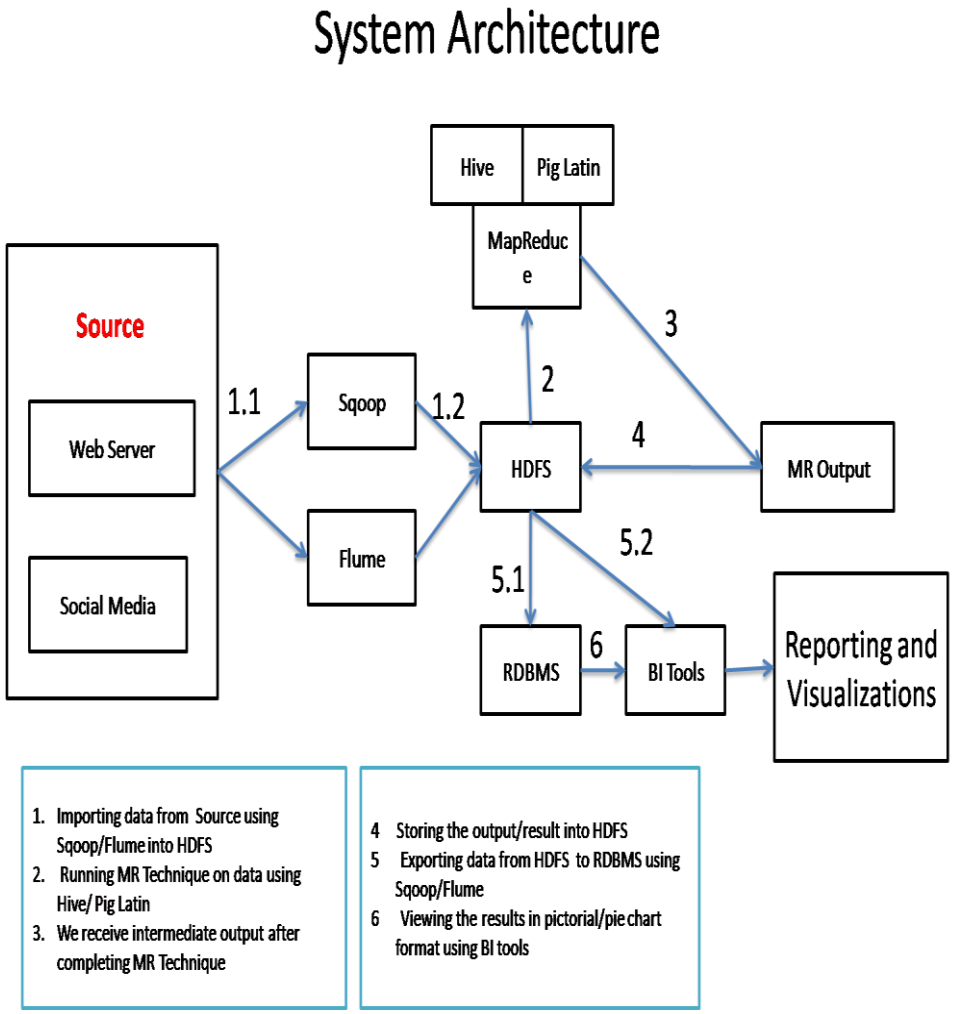
Willson (2006) examined the causes that led to the breakdown of Barring bank, in his case study: "the collapse of Barring Banks". The collapse resulted due to the failures in management, financial and operational controls of Baring Banks. The failures that were evident include the following areas:-

- Failure in management supervision
- Lack of segregation between front and back offices of Baring Futures, Singapore.
- Insufficient actions taken by Barring‟s management in response to warning signals.
- No risk management or compliance function in Singapore
- Weak financial and operational control over the activities and funding of Baring Futures Singapore at group level.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

III. SYSTEM DESIGN

A. System Architecture



IV. CONCLUSION

The bank employees do not give due importance to the problem of frauds. The awareness level of bank employees regarding bank frauds is not very satisfactory, and majority of them do not dispose favorable attitude towards RBI procedures as they find difficulty in following them due to workload and pressure of competition. Moreover employees are not well trained to prevent bank frauds. Training positively affects the compliance level of employees and improves the attitude towards RBI‟s procedure.

V. FUTURE SCOPE

This survey has explored almost all published fraud detection studies. It defines the adversary, the types and subtypes of fraud, the technical nature of data, performance metrics, and the methods and techniques. After identifying the limitations in methods and techniques of fraud detection, this paper shows that this field can benefit from other related fields. Specifically, unsupervised approaches from counterterrorism work, actual monitoring systems and text mining from law enforcement, and semi supervised and game-theoretic approaches from intrusion and spam detection communities can contribute to future fraud detection research. However, Fawcett and Provost show that there are no guarantees when they successfully applied their fraud detection method to news story

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

monitoring but unsuccessfully to intrusion detection. Future work will be in the form of credit application fraud detection.

VI. APPLICATION

- Banking Industry
- Private sector
- Public Sector
- Govt. Sector

REFERENCES

[1] Card Fraud Detection. Proc. of the IEEE/IAFE on Computational Intelligence for Financial Engineering, 220-226.

[2] Artis, M., Ayuso M. & Guillen M. (1999). Modeling Different Types of Automobile Insurance Fraud Behavior in the Spanish Market. Insurance Mathematics and Economics 24: 67-81.

[3] Barse, E., Kvarnstrom, H. & Jonsson, E. (2003). Synthesizing Test Data for Fraud Detection Systems. Proc. of the 19th Annual Computer Security Applications Conference, 384-395.

[4] Belhadji, E., Dionne, G. & Tarkhani, F. (2000). A Model for the Detection of Insurance Fraud. The Geneva Papers on Risk and Insurance 25(4): 517-538.

[5] Bell, T. & Carcello, J. (2000). A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. Auditing: A Journal of Practice and Theory 10(1): 271-309.

[6] Beneish, M. (1997). Detecting GAAP Violation: Implications for Assessing Earnings Management Among Firms with Extreme Financial Performance. Journal of Accounting and Public Policy 16: 271-309.

[7] Bentley, P. (2000). Evolutionary, my dear Watson: Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. Proc. of GECCO2000.

[8] Bentley, P., Kim, J., Jung., G. & Choi, J. (2000). Fuzzy Darwinian Detection of Credit Card Fraud. Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society.

[9] Bhargava, B., Zhong, Y., & Lu, Y. (2003). Fraud Formalization and Detection. Proc. of DaWaK2003, 330-339.

[10] Bolton, R. & Hand, D. (2002). Statistical Fraud Detection: A Review (With Discussion). Statistical Science 17(3): 235-255.

[11] Bolton, R. & Hand, D. (2001). Unsupervised Profiling Methods for Fraud Detection. Credit Scoring and Credit Control VII.

[12] Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D. (1999). A Classification-based Methodology for Planning Auditing Strategies in Fraud Detection. Proc. of SIGKDD99, 175-184.

[13] Brause, R., Langsdorf, T. & Hepp, M. (1999). Neural Data Mining for Credit Card Fraud Detection. Proc. of 11th IEEE International Conference on Tools with Artificial Intelligence.

[14] Brockett, P., Derrig, R., Golden, L., Levine, A. & Alpert, M. (2002). Fraud Classification using Principal Component Analysis of RIDITs. Journal of Risk and Insurance 69(3): 341-371.

[15] Brockett, P., Xia, X. & Derrig, R. (1998). Using Kohonen's Self Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud. Journal of Risk and Insurance 65(2): 245-274.

[16] Burge, P. & Shawe-Taylor, J. (2001). An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection. Journal of Parallel and Distributed Computing 61: 915-925.

[17] Cahill, M., Chen, F., Lambert, D., Pinheiro, J. & Sun, D. (2002). Detecting Fraud in the Real World. Handbook of Massive Datasets 911-930.

[18] Caruana, R. & Niculescu-Mizil, A. (2004). Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria. Proc. of SIGKDD04, 69-78.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

AUTHOR BIOGRAPHY

First Author :

Prof.S.V.Phulari
Lecturer at College of engg Manjari BK. Pune, Maharashtra, India.

Second Author:

Umesh Lamture
He is pursuing BE(COPMUTER) degree from College of Engg Manjari BK. Pune, Maharashtra, India.

Third Author :

Sumit Madage
He is pursuing BE(COPMUTER) degree from College of Engg Manjari BK. Pune, Maharashtra,India.

Fourth Author:

Kunal Bhandari
He is pursuing BE(COPMUTER) degree from College of Engg ManjariBK. Pune, Maharashtra,India.