Real Time Loan Fraud Detection in Banking Sector Using Machine Learning and DM Technique

Brunda Shree V[1], Bhavana L[2], Harshavardhana Doddamani[3]

[1, 2] *Engineering Student in Dept. of Computer Science Engineering, SJCIT, Chickballapur, Karnataka, India*

[3]*Assistant Professor in Dept. of Computer Science Engineering, SJCIT, Chickballapur, Karnataka, India*

Abstract- **There is a loss of billions of dollars worldwide each year because of frauds happening in financial sectors. There is a problem with internal auditing system of financial sectors because of which there is a need for fraud detection. This paper gives the details about fraud detection in financial sectors using machine learning and data mining techniques. Bayesian and K-means algorithms are used for fraud detection here.**

Index Terms- Frauds, Financial Detection, Auditing, Worldwide.

I. INTRODUCTION

There are many types of bank fraud in this world like accounting fraud, Cheque kiting, forgery and altered cheques, loan fraud. Among all types of bank frauds loan frauds can put a bank into a huge financial loss. The paper mainly focuses on loan frauds in banking sector. There are many bank loan scams that shook the Indian Financial sector they are-Industrialist Bipin Vohra and others were booked by CBI for allegedly cheating CBI by obtaining the loan with forged documents, Vijay Mallya is wanted in India for loan fraud of around 9000 crore and many more cases like that. Manual verification for detecting loan frauds in large banks is time consuming and can be impossible sometimes. This paper deals with implementing a loan fraud detecting system using machine learning and data mining techniques. After accepting the loan application banks has to put that application to test and detect whether there is any fraud in the details given by a customer or not, if there is any fraud the application has to be rejected else accepted. Clustering is used to group similar type of data that helps in uncomplicated retrieval of data. Classification is data mining technique based on machine learning that involves learning a function that a data item into one of the previously defined classes.

II. LITERATURE REVIEW

The paper [1] proposes based on protected loans are loans that rely on an asset. In the state of loan failure to pay, the lender can possess the asset and make use of it to cover up the loan. High well being rates for secured loans may be lower than those for unsecured loans. The documents may need to be evaluated before you can have a loan of a secure type. Unsecured loans lend may be more complicated to get and have higher concern rates. Unsecured loans rely just on your credit history and your revenue to meet the criteria for the loan. If someone fails to pay back loan, the lender has to look for alternate plan to get his money back.

The approach [2] introduces by using data mining techniques to analyze patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which category of customers are likely to accept new product offers, which category of customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more efficient.

The paper [3] approaches many risks related to bank loans, for the bank and for those who get the loans. Risk is the probability of certain outcomes--or the uncertainty of them--especially an existing threat for trying to achieve a current bank operation. Risk in bank loans involve: credit risk, security risk, the risk that the loan won't be return back on time or at all; liquidity risk, the risk that many deposits will be withdrawn quickly, leaving the bank short on immediate cash; and interest rate risk, the risk that the interest rates priced on bank loans will be low to earn the bank adequate money.

The approach proposed by [4] explains various internet fraud complaints include auction fraud, credit and debit card fraud, non-delivery of goods or services. We are all vulnerable to illegal scams via internet. Online Credit Fraud includes auction fraud, credit and debit card fraud, bank details fraud, social security number fraud, valuable personal information and identity fraud through fake scams. Thus internet creates some major new challenges for consumers and organization.

The paper [5] approaches by bank took in consideration all parameters which lead in internet banking fraud. Analysts established many detection rules. Despite of initial ones, new rules are added, when analysis finds out suspect patterns and behaviors. Those rules are implemented in an offline fraud detection system. System is offline because of its database update. Data Set is imported in database in constant time frames, not in real time. As of now there is no immediate need to upgrade system in online mode. Analysis, design and implementation of the system took part in-house. Due to the bank's data sensitivity, one of the prerequisites was in-house set up and function of such system. Data mining and predictive analytics tools contributed in all phases of project and they are part of the system.

The paper [6] explains about Fraud Detection System is a system of analyzing the terminal data, IP address, and transaction details used in electronic financial transaction to detect suspicious transactions and block frauds. Fraud detection system consists of four functions: data gathering, analysis and detection, response, and monitoring and audit.

The paper approaches by [7] focuses on bankruptcy fraud. Bankruptcy fraud means misuse of credit card when a card holder is not present. Most complicated type of fraud to predict is Bankruptcy fraud. Bank uses some methods and techniques to predict users of card holder. One of the possible ways to prevent bankruptcy fraud is to pre-check the credit card with credit bureau in order to be informed about the past banking history of its customers.

The author [8] explains about different intelligent approaches to fraud detection which are both statistical and computational though the performance was differed each technique was shown to be reasonably capable at detecting various forms of financial fraud. The ability of the computational methods such as neural networks and support vector machines to learn and adapt to many new techniques is highly effective to the evolving of tactic fraudsters. Initial fraud detection studies focused heavily on statistical models such as logistic regression, as well as neural networks. Neural networks are used for financial applications such as forecasting. Neural network are well established history with fraud detection. But they require high computational power for training and operation, making it unsuitable for real time function. Potential for over fitting if training set is not a good representation of the problem domain, so requires constant retraining to adapt to new methods of fraud. In this paper the author says about the different kinds of frauds by insurance fraud, mortgage fraud, health insurance fraud, telecommunication fraud, credit card fraud. Different techniques have been defined for different kinds of frauds defining the parameters like entropy, sensitivity and comparing the efficiency of the different kinds of algorithms and representing them in a graphical representation.

The approach [9] proposes a user rating system for the internet and the authority for internet security about E-business; e-business is the vital uses of Internet. Web is the fundamental apparatus for e-business and banks have changed their plan of action with the assistance of web. Banks broadened their offices by means of on the web and along these lines e-transaction has expanded quickly in the keeping money division. The development of on-line exchange gives a colossal chance to banks and buyers. Be that as it may, credit extortion discovery and aversion framework in the managing an account part is still stayed unsecured. Keeping money speaks to the reflection of economy; extortion brings colossal misfortunes that stun all the performing exercises. Inside managing an account extortion constitutes an forceful nearness in this division. Along these lines, allurement is consistently developing and circling all through the whole keeping money framework. The measure of e-extortion was little in the exact start of e-keeping money movement.

The approach proposed by [10] explains about borrower risk and the price terms of the bank loans, there are numerous dangers identified with bank loans, for the bank and for the individuals who get the advances. The investigation of hazard in bank credits requires understanding what the significance of hazard is. Hazard is indicates to the likelihood of specific results - or the vulnerability of them- -particularly a current negative danger for attempting to accomplish a current money related task. Hazard in bank advances include: credit chance, the hazard that the credit won't be return back on time or by any stretch of the imagination; liquidity chance, the hazard that excessively numerous stores will be pulled back too rapidly, leaving the bank short on quick money; and premium rate hazard, the hazard that the financing costs valued on bank credits will be too low to acquire the bank satisfactory cash division and productivity, high hazard advance candidates, foreseeing installment default, marketing, credit examination, positioning speculations, fake exchanges, enhancing stock portfolios, money administration and determining activities, most beneficial Credit Card Customers and Cross Selling. There are various kinds of advances you need to consider when you're hoping to acquire cash and it's essential to know your choices.

III. PROPOSED METHOD

A. Objectives.

In our proposed methodology, we supposed that a fraud detecting system has the following objectives:

1. To avoid real time loan fraud at a maximum level.
2. To increase the confidence of customers in the banking system.
3. To discourage fraudsters.

In this detecting system, we are predicting loan frauds using Machine learning techniques like classification and clustering. The purposed system has two modules and they are:

1. Clustering module
2. Detection module.

The clustering module browses the data sets, cluster the data sets according to number of clusters mentioned by the bank management and visualize the clustered data. The Detection module browses the trained and test data of loan applications and predicts whether the application has fraud or no fraud.

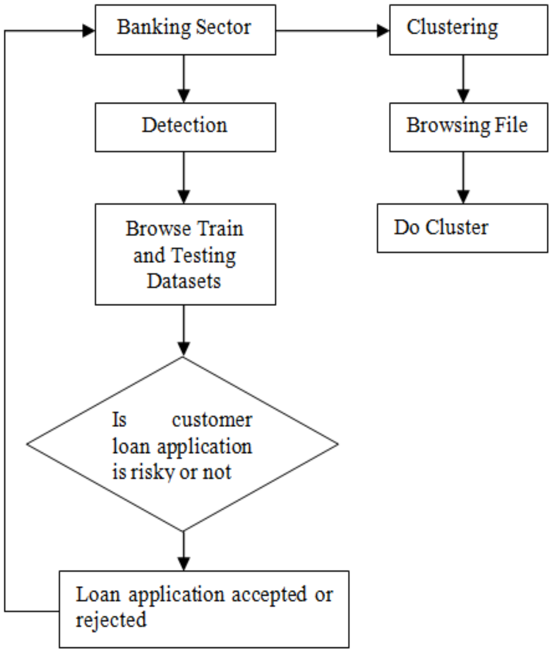

Banking Sector
Clustering
Detection
Browsing File
Browse Train and Testing Datasets
Do Cluster
Is customer loan application is risky or not
Loan application accepted or rejected

Figure -1: Data Flow Diagram of Fraud Detection System

B. Classification.

Classification is a supervised learning in machine learning, where data sets contains both dependent and independent variable. Classification techniques are Naïve Bayes Classifier, Decision Tree Classifier, Neural Networks and support vector machines. Classification techniques are used in fraud detection in credit cards, healthcare.

C. Clustering.

Clustering is an unsupervised learning, where the data sets contain only one variable and identifies the similar classes of objects. Clustering is used to partition the datasets into many dissimilar clusters so that data sets in one cluster are same and are different from the data sets in other clusters. Clustering is also called as Data segmentation. Clustering techniques are K-nearest, self-organizing maps.

D. Visualization.

Visualization refers to presentation of data sets into graphical patterns that makes the users to view the complicated data sets into clear patterns. The data sets patterns can have different color, position, size and other visual characteristics. Visualization is the best tool to represent the data sets into clear patterns.

IV. IMPLEMENTATION

Loan fraud detecting system is implemented using k-means clustering algorithm for cluster Analysis and Naïve Bayes Classifier for predicting the frauds in the data sets.

A. K-Means Clustering Algorithm.

K-Means is clustering algorithm for partitioning the data sets into different clusters. Here, k is a number of clusters specified. This algorithm calculates the minimum centroid between the cluster and given data and appends the given data to cluster and terminates when lowest distance observed.

The sample data in initially partitioned into k clusters and given data are assigned to different clusters according to following steps:

Step 1: Calculate the centroid between the cluster and the given data.

Step 2: Assign the data to a cluster if centroid is nearer or Else reject.

Step 3: Repeat Step 1 and Step 2 until each sample is assigned to a cluster and clusters are stable.

A bank management has to check the data in order to know which loan application is risky or which are safe. There is an enormous amount of data which one can actually retrieved from the banking database. The data can be used to detect the loan frauds based on their personal information like age, education, income, Debit score, and Credit score. The proposed system categorizes loan frauds based on some personal information. The system uses the concept of k-means clustering to cluster the customer based upon the similarities or the patterns they share among each other.

E. Naïve Bayes Classifier.

Naïve Bayes Classifier is a classification technique based on the probabilistic theorem .Probabilistic theorem calculates the probability of hypothesis according to the given trained data sets.

Bayes' Theorem is stated as:

$$P (p/q) = (P (q/p) * P (p)) / P (q) \qquad (1)$$

Where

P (p/q) is the probability of hypothesis p given the data q. This is called the posterior probability.

P (q/p) is the probability of data q given that the hypothesis p was true.

P (p) is the probability of hypothesis p being true (regardless of the data). This is called the prior probability of p.

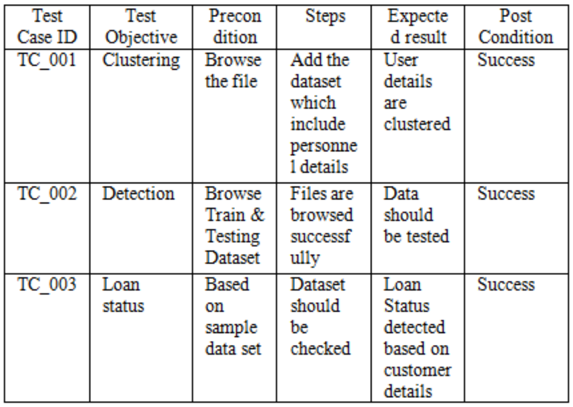
P (q) is the probability of the data (regardless of the hypothesis).

Naïve Bayes is used to classify and detect the frauds using training data sets and after classification of test data sets it predicts the frauds before proceeding the loan application.

System design

The results of experimental system in detecting loan frauds in banking system are presented in this division. We have implemented our proposed model in CORE-JAVA. We have used a sample bank dataset for experimental analysis. Here the detection of loan frauds done on using K-Means clustering and Classification Naïve Bayesian algorithm.

Table 1. Test cases for fraud detecting system



Test Case ID	Test Objective	Precondition	Steps	Expected result	Post Condition
TC_001	Clustering	Browse the file	Add the dataset which include personnel details	User details are clustered	Success
TC_002	Detection	Browse Train & Testing Dataset	Files are browsed successfully	Data should be tested	Success
TC_003	Loan status	Based on sample data set	Dataset should be checked	Loan Status detected based on customer details	Success

V. CONCLUSION

This paper begins with a concept of data mining and loan fraud detection, followed by a discussion of evolution, characteristics, and techniques. Data mining: It is a process to extract knowledge from existing data. It is used in banking and financial sector in general to discover useful information from the operational and historical data to enable better decision-making. The proposed system based on K-means Clustering and Naïve Bayesian algorithms was used to analyze and detect the loan frauds in banking sector. Here detecting the loan frauds before proceeding the loan application using customer personal details. This will definitely help the banking industry to save the huge amount.

FUTURE WORK

- Future work requires change in focus. To minimize losses and trust banks should change the strategies and priorities.
- Focus should be more self- protecting the application from data breaches and use data mining to detect pattern and fraud with Actionable Auto Intelligence.

REFERENCES

[1] "Financial Interconnectedness and Financial Sector Reforms" in the Caribbean Prepared by Sumiko Ogawa, Joonkyu Park, Diva Singh, and Nita Thacker1 in 2011

[2] Dr. Madan Lal Bhasin, "Data Mining: A Competitive Tool in the Banking and Retail Industries", the Chartered Accountant October 2006.

[3] Strahan, Philip E. "Borrower risk and the price and nonprice terms of bank loans." FRB of NewYork Staff Report 90 (1999).

[4] Palshikar, Girish Keshav. "The hidden truth-frauds and their Control: A critical application for business intelligence." Intelligent Enterprise 5.9 (2002): 46-51.

[5] C. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud detection Research", Artificial Intelligence Review, 2005.

[6] Guide to Fraud Detection System Technology, Financial Security Institute, 2014.

[7] Foster & Stine (2004) presented a model to forecast personal bankruptcy among users of credit card.

[8] Jarrod [West, Maumita Bhattacharya]"Intelligent Financial fraud detection" in 2014

[9] Gonca T. Y. and Faruk K. (2009), User Rating System for the Internet (URSI) and Central Authority for Internet Security (CAIS).

[10] Demsetz, Rebecca S., Marc R. Saidenberg and Philip E. Strahan, 1998, "Agency Problems and Risk-Taking at Banks," Federal Reserve Bank of New York, Staff Report no. 9709.