# DATA MINING TECHNIQUES TO ANALYSES RISK GIVING LOAN(BANK)

[1]Mrunal Surve, [2]Pooja Thitme, [3]Priya Shinde, [4]Swati Sonawane, [5]Sandip Pandit

[1]Mrunal Surve, BE, Dept. of Computer Engineering, Amrutvahini COE, Maharashtra, India
[2]Pooja Thitme, BE, Dept. of Computer Engineering, Amrutvahini COE, Maharashtra, India
[3]Priya Shinde, BE, Dept. of Computer Engineering, Amrutvahini COE, Maharashtra, India
[4]Swati Sonawane, BE, Dept. of Computer Engineering, Amrutvahini COE, Maharashtra, India
[5]Sandip Pandit, Professor,Dept. of Computer Engineering, Amrutvahini COE, Maharashtra, India

## ABSTRACT

*In this project the main focus is to identify and analyse the risk in giving loan of commercial banks.the domain is data mining so we are using data mining techniques to analyze risk giving loan.it includes analysing and processing data from various resources and summarise into a valuable information. Traditional process of lending is very tedious task because it includes a large manual work.whereas our project is design to automate all the manual work this will increase the speed and accuracy.This will improve the quality of banking system thus improving the customer retention. We are using C4.5 classification algorithm of Data mining.Applying it on a dataset of customer and predicting the risk percentage for individual to give loan.Today, customers have so many opinions with regard to where they can choose to do their business.Executives in the banking industry,therefore,must be aware if they are not giving each customer their full attention,the customer can simply find another bank. Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. Although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analyst. The banks and many investment companies are pioneers in taking advantage of Data Mining.*

**Keyword:** *- Decision tree, ID3, C 4.5, Banking System, Risk Percentage*

## 1. INTRODUCTION

The traditional lending system includes lots of manual work, which is a tedious task and may leads to some mistakes that increases risks. The main idea behind the project is to analyse the risk percentage while assessing the loan. Analysing the loan becomes automated process and it reduces the manual work and increases customer retention. IT has helped the banking industry to deal with the challenges the new economy poses. Nowadays, Banks have realized that customer relationships are a very important factor for their success. Customer relationship management (CRM) is a strategy that can help them to build long-lasting relationships with their customers and increase their revenues and profits. CRM in the banking sector is of greater importance. The CRM focus is shifting from customer acquisition to customer retention and ensuring the appropriate amounts of time, money and managerial resources are directed at both of these key tasks use data warehousing to combine various data from databases into an acceptable format so that the data can be mined. The data is then analyzed and the information that is captured is used throughout the organization to support decision-making.

In this paper we study, how the customer relationship manages in banking system and retail industries by using data mining techniques. We know that now a days several industries including like banking, finance, retail, insurance, publicity, database marketing, sales predict, etc are data Mining tools for Customer Relationship Management.

Currently, banks are using Data Mining tools for customer segmentation and benefit, credit scoring and approval, predicting payment lapse, marketing, detecting illegal transactions, etc. Data mining is nothing but identifying the patterns and relationships in the data. In data mining, Some widely used techniques are artificial neural networks, genetic algorithms, K-nearest neighbour method, decision trees, and data reduction.

Systemic risk is defined as the danger of one specific bank being in stress amplifying the panic in the whole banking system, leading to the failure of other banks, and consequently to the financial crisis. Therefore, measuring and identifying level of this kind of risk for each bank is essential for bank supervisors and policy makers as well. To explain the systemic risk, some data mining approaches such as Support Vector Machines to help predict forward CoVaR for regulatory purpose are used. The data mining techniques such as clustering, classification, association, decision tree, SVM, CoVaR, CRM, KDD,Credit Scoring, etc are used to analyze the risk in financial sector. In this paper C4.5 algorithm is used to analyse the risks in banking systems.

## 2. METHODOLOGIES

Following is the data mining techniques used for predicting and analysing the risk in banking system. More concentrated on the financial risk prediction.

### 2.1 C4.5 Algorithm

C4.5 is a program for inducing classification rules in the form of decision trees from a set of given examples.
C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan.
The C4.5 systemconsists of four principal programs:
**1) decision tree generator ('c4.5')** - construct the decision tree
**2) production rule generator ('c4.5rules')** - form production rules from unpruned tree
**3) decision tree interpreter ('consult')** - classify items using a decision tree
**4) production rule interpreter ('consultr')** - classify items using a rule set
Classification algorithms have attracted considerable interest both in the machine learning and in the data mining research areas. Among classification algorithms, the C4.5 system of Quinlan deserves a special mention for several reasons. On the one hand, it represents the result of research in machine learning that traces back to the ID3 system . As such, it has always been taken as the point of reference for the development and analysis of novel proposals. On the other hand, the results show that the C4.5 tree-induction algorithm provides good classification accuracy and is the fastest among the com pared main-memory algorithms for machine learning and data mining. It should be mentioned that several external-memory algorithms and parallel implementations have been proposed with the aim of speeding up the execution time and reasoning on very large training sets.

The algorithm constructs a decision tree starting from a training set *T S*, which is a set of cases, or tuples in the database terminology. Each case specifies values for a collection of attributes and for a class. Each attribute may have either discrete or continuous values. Moreover, the special value unknown is allowed, to denote unspeci_ed values. The class may have only discrete values. We denote with $C_1$…. $C_{NClass}$ the values of the class.

### 2.2 Decision trees

A decision tree is a tree data structure consisting of decision nodes and leaves. A leaf specifies a class value. A decision node specifies a test over one of the attributes, which is called the attribute selected at the node. For each possible outcome of the test, a child node is present. In particular, the test on a discrete attribute A has h possible outcomes A = d1, . . . , A = dh, where d1…….dh are the known values for attribute A. The test on a continuous attribute has two possible outcomes, A _ t and A > t, where t is a value determined at the node, and called the threshold. A decision tree is used to classify a case, i.e. to assign a class value to a case depending on the values of the attributes of the case. In fact, a path from the root to a leaf of the decision tree can be followed based on the attribute values of the case. The class specified at the leaf is the class predicted by the decision tree. A performance measure of a decision tree over a set of cases is called classification error. It is defined as the percentage of mis-classified cases, i.e. of cases whose predicted classes differ from the actual classes.

**2.3 The tree construction algorithm**

The C4.5 algorithm constructs the decision tree with a divide and conquers strategy. In C4.5, each node in a tree is as- asociated with a set of cases. Also, cases are assigned weights to take into account unknown attribute values. At the beginning, only the root is present, with associated the whole training set T S and with all case weights equal to 1:0. At each node the following divide and conquer algorithm (see Program 1) is executed, trying to exploit the locally best choice, with no backtracking allowed. Let T be the set of cases associated at the node. The weighted frequency freq(Ci; T) is computed (step (1)) of cases in T whose class is Ci, for i2 [1;NClass]. If all cases (step (2)) in T belong to a same class Cj (or the number of cases in T is less than a certain
value) then the node is a leaf, with associated class Cj (resp., the most frequent class). The classi_cation error of the leaf is the weighted sum of the cases in T whose class is not Cj (resp., the most frequent class). If T contains cases belonging to two or more classes (step (3)), then the information gain of each attribute is calculated. For discrete attributes, the information gain is relative to the splitting of cases in T into sets with distinct attribute values. For continuous at tributes, the information gain is relative to the splitting of T into two subsets, namely cases with attribute value not greater than and cases with attribute value greater than a certain local threshold, which is determined during information gain calculation. The attribute with the highest information gain (step (4)) is selected for the test at the node. More-over, in case a continuous attribute is selected, the threshold is computed (step (5)) as the greatest value of the whole training set that is below the local threshold. A decision node has s children if T1…… Ts are the sets of the splitting produced by the test on the selected attribute (step (6)). Obviously, s = 2 when the selected attribute is continuous, and s = h for discrete attributes with h known values. For i = [1; s], if Ti is empty, (step (7)) the child node is directly set to be a leaf, with associated class the most frequent class at the parent node and classification error 0. If Ti is not empty, the divide and conquer approach consists of recursively applying the same operations (step (8)) on the set consisting of Ti plus those cases in T with unknown value of the selected attribute.
Note that cases with unknown value of the selected attribute are replicated in each child with their weights proportional to the proportion of cases in Ti over cases in T with known value of the selected attribute. Finally, the classification error (step (9)) of the node is calculated as the sum of the errors of the child nodes. If the result is greater than the error of classifying all cases in T as belonging to the most frequent class in
T, then the node is set to be a leaf, and all sub-trees
are removed.

## 3. PROPOSED SYSTEM

Following Figure depicts the data mining techniques and algorithm that are applicable to the banking sector. Customer retention pays vital role in the banking sector. The supervised learning method Decision Tree implemented using CART algorithm is used for customer retention. Preventing fraud is better than detecting the fraudulent transaction after its occurrence. Hence for credit card approval process the data mining techniques Decision Tree, Support Vector Machine (SVM) and Logistic Regression are used. Clustering model implemented using EM algorithm can be used to detect fraud in banking sector.

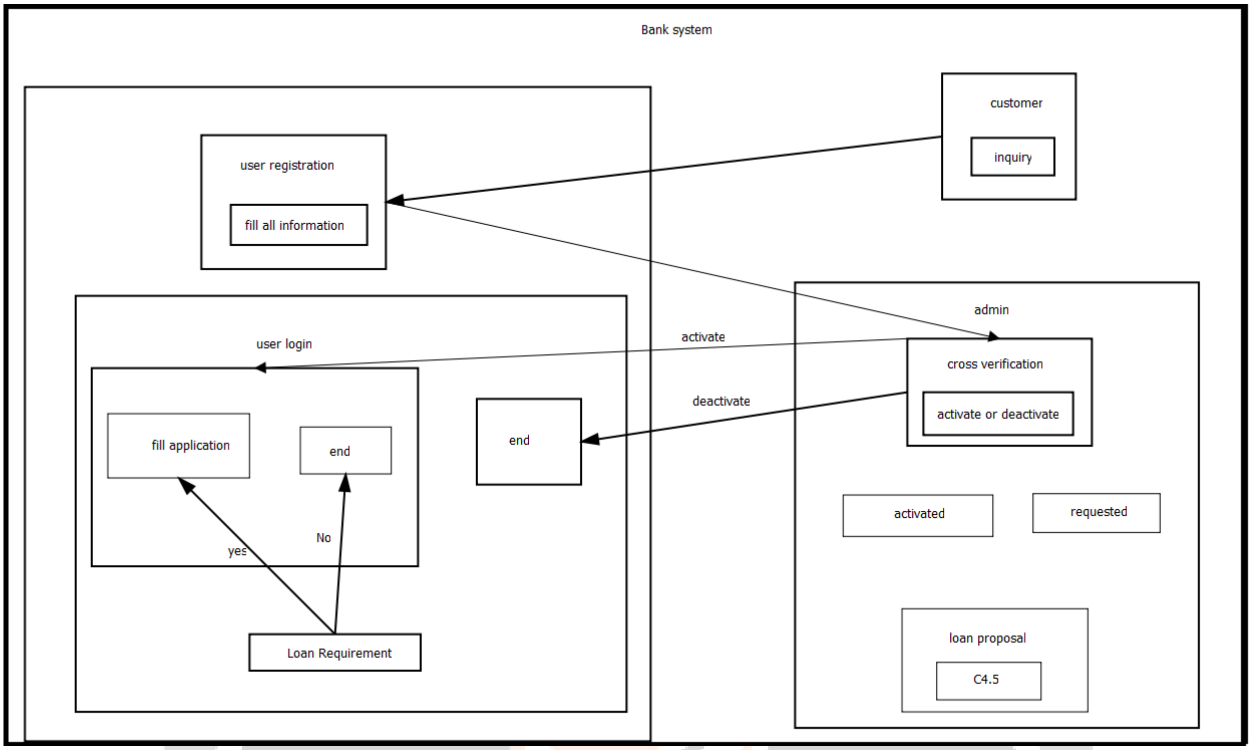

**Fig-1:**Proposed System Architecture

### 3.1 Decision Tree:

Decision trees are the most popular predictive models (Burez and Van den Poel, 2007). A decision tree is a tree like graph representing the relationships between a set of variables. Decision tree models are used to solve classification and prediction problems where instances are classified into one of two classes, typically positive and negative, or churner and non-churner in the churn classification case. These models are represented and evaluated in a top-down manner. Developing decision trees involves two phases:

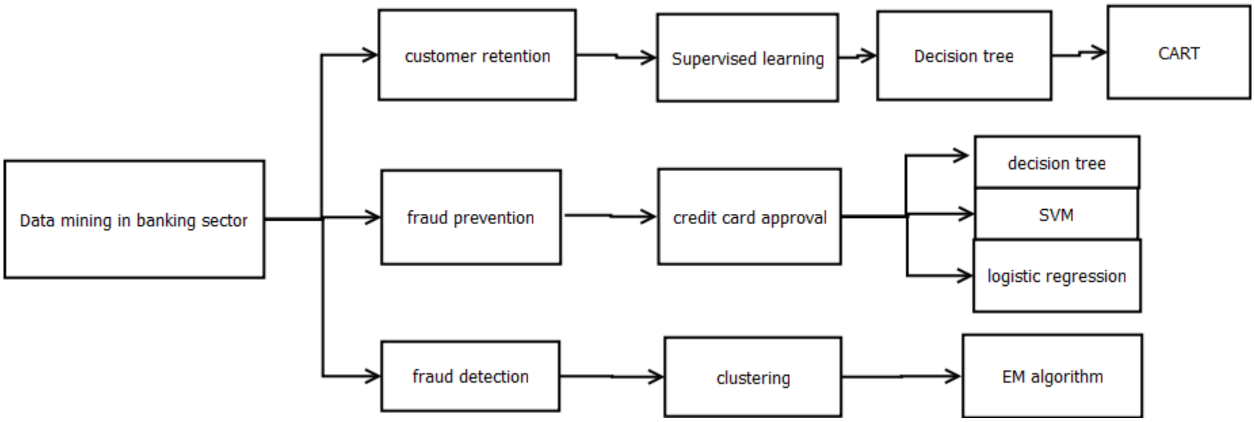

**Fig -2**: Decision Tree

**1)Tree building and tree pruning:**

Tree building starts from the root node that rep- resents a feature of the cases that need to be classified. Feature selection is based on evaluation of the information gain ratio of every feature. Following the same process of information gain evaluation, the lower level nodes are constructed by mimicking the divide and conquer strategy. Building a decision tree incorporates three key elements:

1. Identifying roles at the node for splitting data according to its value on one variable or feature.
2. Identifying a stopping rule for deciding when a sub-tree is created.
3. Identifying a class outcome for each terminal leaf node, for example,Churn or Non-churn.

Once the model is built, the decision about a given case regarding to which of the two classes it belongs is established by moving from the root node down to all the leaves and interior nodes. The movement path is determined by the similarity calculation until a leaf node is reached, at which point a classification decision is made.

**2) Value Prediction Methods:**

In this method, for example, instead of classifying new loan applications, it attempts to predict expected default amounts for new loan applications. The predicted values are numeric and thus it requires modeling techniques that can take numerical data as target (or predicted) variables. Neural Network and regression are used for this purpos e.

The most common data mining methods used for customer profiling are:

_ Clustering (descriptive)

_ Classification (predictive) and regression (predictive)

_ Association rule discovery (descriptive) and sequential pattern discovery (predictive)

## 4. RESULTS

Overall an innovative technique to analyse the risk while giving loan(bank) is presented here. The proposed methodology C4.5 is used to make decision about the risk in granting loan to the customer. This reduce human efforts and makes the banking system more efficient. Thus it results in improving the customer retention.

## 5. CONCLUSIONS

Data mining is a tool used to extract the knowledge from existing data and enable better decision-making throughout the banking industries. Data mining tool use data warehousing to collect the various data from databases into an acceptable format so that the data can be mined. This mined data is then analyzed and the information that is captured from the data which is used throughout the organization to support decision-making. Data mining techniques are used many industries effectively. data mining tools used in many practical applications in such areas as analyzing medical outcomes, detecting credit card fraud, predicting customer purchase behavior, predicting the personal interests of Web users, optimizing manufacturing processes etc. have been very successful. It has also be responsible for to a set of fascinating scientific questions about how computers might automatically learn from past experience. The marginal industry is also realizing that data mining could give them a competitive benefit. A majority of the banks in developing countries are not usually known to exploit their information "asset" for deriving business value through data mining and gain competitive advantage. But with progressive liberalization of rules on entry for private and foreign multinational banks, under the GATS framework of WTO, competitive pressure on domestic banks is increasing. Thus, customer retention and acquisition will be an important determinant of the banks‟ bottom lines. Those banks and retailers that have realized the utility of data mining and are in the process of building a data mining environment for their decision-making process will reap immense benefit and dessrive considerable competitive advantage to stand competition in future.

## 6. REFERENCES

[1]. Yibing Chen, Yong Shi,Cheng-Few Lee, Minqiang Li, Yuewen Liu "Measuring and Predicting Systemic Risk in the Chinese Banking System", 2014 IEEE International Conference on Data Mining Workshop.

[2]. Dr. K. Chitra, B. Subashini ,"Data Mining Techniques and its Applications in Banking Sector, International Journal of Emerging Technology and Advanced Engineering", Website:www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013) .

[3]. Abhijit A. Sawant and P. M. Chawan ,"Study of Data Mining Techniques used for Financial Data Analysis, International Journal of Engineering Science and Innovative Technology "(IJESIT) Volume 2, Issue 3, May 2013 .
[4]. P Salman Raju, Dr V Rama Bai, G Krishna Chaitanya,"Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries", International Journal of Innovative Research in Computer and Communication Engineering,Vol. 2, Issue 1, January 2014 .
[5]. Hamid Eslami Nosratabadi, Sanaz Pourdarab and Ahmad Nadali, "A New Approach for Labeling the Class of Bank Credit Customers via Classifcation Method in Data Mining", International Journal of Information and Education Technology, Vol.1, No. 2, June 2011.
[6]. Ketaki Chopde, Pratik Gosar, Paras Kapadia, Niharika Maheshwari, Pramila M. Chawan ,"A Study of Classi_cation Based Credit Risk Analysis Algorithm, International Journal of Engineering and Advanced Technology" (IJEAT) ISSN: 2249 8958, Volume-1, Issue-4, April 2012 .
[7]. M. P. Thapliyal , "Data Mining: A Tool for Banking Industry, International Journal of Emerging Research in Management Technology "ISSN: 2278-9359 (Volume-4, Issue-4), April 2015.
[8]. Cheng-Lung Huang , Mu-Chen Chen , Chieh-Jen Wang "Credit scoring witha data mining approach based on support vector machines" Expert Systems with Applications 33 (2007) 847856 ,Sciencedirect 2007.
[9]. Dongsong Zhang and Lina Zhou ,"Discovering Golden Nuggets: Data Mining in Financial Application" IEEE transactions on systems, man, and cyberneticspart c: applications and reviews, vol. 34, no. 4, november 2004.
[10]. Wei-Sen Chen , Yin-Kuan Du ,"Using neural networks and data mining techniques for the _nancial distress prediction model", Expert Systems with Applications 36(2009) 40754086, Elsevier 2009.