TWO STEP CREDIT RISK ASSESMENT MODEL FOR RETAIL BANK LOAN APPLICATIONS USING DECISION TREE DATA MINING TECHNIQUE

Sudhakar M[1], Dr. C. V. K Reddy[2]

[1]*Research Scholar, D*epartment of Computer Science and Technology
Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India
[2]*Professor, D*epartment of Physics, Rayalaseema University
Kurnool, Andhra Pradesh, India

Abstract

Data mining techniques are becoming very popular nowadays because of the wide availability of huge quantity of data and the need for transforming such data into knowledge. In today's globalization, core banking model and cut throat competition making banks to struggling to gain a competitive edge over each other. The face to face interaction with customer is no more exists in the modern banking world. Banking systems collect huge amounts of data on day to day basis, be it customer information, transaction details like deposits and withdrawals, loans, risk profiles, credit card details, credit limit and collateral details related information. Thousands of decisions are taken in a bank on daily basis. In recent years the ability to generate, capture and store data has increased enormously. The information contained in this data can be very important. The wide availability of huge amounts of data and the need for transforming such data into knowledge encourage IT industry to use data mining. Lending is the primary business of the banks. Credit Risk Management is one of the most important and critical factor in banking world. Without proper credit risk management banks will face huge losses and lending becomes very tough for the banks.

Data mining techniques are greatly used in the banking industry which helps them compete in the market and provide the right product to the right customer with less risk. Credit risks which account for the risk of loss and loan defaults are the major source of risk encountered by banking industry. Data mining techniques like classification and prediction can be applied to overcome this to a great extent. In this paper we introduce an effective prediction model for the bankers that help them predict the credible customers who have applied for loan. Decision Tree Induction Data Mining Algorithm is applied to predict the attributes relevant for credibility. A prototype of the model is described in this paper which can be used by the organizations in making the right decision to approve or reject the loan request of the customers.

Keywords— Banking industry; Data Mining; Risk Management; Classification; Credit Scoring; Non-Performing Assets; Default Detection; Non-Performing Loans Decision Tree; Credit Risk Assessment; Classification; Prediction

---***---

1. INTRODUCTION

In the financial services industry throughout the world, the traditional face-to-face customer contacts are being replaced by electronic points of contact to reduce the time and cost of processing an application for various products. Since 1990‟s the whole concept of banking has been shifted to centralized databases, online transactions and ATM‟s all over the world, Core banking, internet banking and mobile banking which are made banking system technically strong and more customer oriented. Banks have realized that retaining the customers and preventing fraud must be the strategy tool for healthy competition [5]. Business decisions can be optimized through data mining [3].

The computerization of financial operations, use of internet and automated software's has completely changed the basic concept of business and the way the banking operations are being carried out. The face to face bank to customer interaction is no more exists in the modern banking world. Customer interaction with bank in most of the cases is through electronic mode like using the core banking services online or using the ATM for most of the basic banking services. Customer relation with bank is mostly on virtual mode and bank officers never get a chance to meet the customer in today's banking world. It is becoming hard for the bankers to understand each customer, maintain a personal relationship and understand the customer risk profile.

All Rights Reserved © 2016 IJARCET

As lending is the primary business of the banks, banks have to always find the right customer to lend at low risk to avoid any future non-performing loans. If a bank approves a loan to a borrower and if borrower not repaying the principal amount, bank will lose the principal and interest income. Thus credit risk assessment is very critical while approving a loan to customer.

Credit risks which account for the risk of loss and loan defaults are the major source of risk encountered by banking industry [2]. Data mining techniques like classification and prediction can be applied to overcome this to a great extent. There are mainly two objectives that is to be achieved through these techniques. They are:

1) Identification of the relevant attributes that signal the capacity of borrowers to pay back the loan, and

2) Determining the best model to evaluate credit risk.

Decision Tree Induction Algorithm is one of the best techniques to achieve this objective [4]. The model thus developed will provide a better credit risk assessment, which will potentially lead to a better allocation of the bank's capital.

In this regard, a study is conducted and an efficient prediction model which helps to reduce the proportion of unsafe borrowers is introduced here with. Due to the significance of credit risk analysis, this study helps banking industry by providing additional information to the loan decision-making process, potentially decreases the cost and time of loan applications appraisal, and decreases the level of uncertainty for loan officers by providing knowledge extracted from previous loans. Decision Tree Induction Algorithm used in this model is the data mining technique for predicting credible customers.

The remaining sections of the paper are organized as follows: In Section 2 we cover brief review about credit risk and credit scoring. In Section 3 we cover Data Mining in Banking and some of the related works is presented. Existing credit approval Model and issues in existing model are covered in section 4. Proposed Model and the Architecture of Proposed Model are described in Sections 5 respectively. The experimental results and the prototype for prediction are given in Section 6. The conclusion and future directions are summed up in Section 7.

1.1 Quick facts about Indian Banking System

The economy of India is the 10th largest in the world by nominal GDP and the 3rd largest by purchasing power parity (PPP). The country is one of the G-20 major economies and a member of BRICS. On a per-capita-income basis, India ranked 141st by nominal GDP and 130th by GDP (PPP) in 2012, according to the IMF. India is the 19th-largest exporter and the 10th-largest importer in the world. Stable financial and banking industry is the back bone of the financial stability of any country in the world. India's financial system—comprising its banks, equity markets, bond markets and myriad other financial institutions—is a crucial determinant of the country's future growth trajectory.

The RBI-Reserve Bank of India is the banker's bank. It is India's central banking institution which controls the monetary policy and the Indian rupee. The Indian banking sector is broadly classified into scheduled banks and non-scheduled banks. There are 27 Public Sector Banks (Govt..), 30 Private Sector Banks and 40 Foreign Banks. There are 68 Co-operatives banks across India.

Indian Banking Industry currently employees 1,175,149 employees and has a total of 109,811 branches in India and 171 branches abroad and manages an aggregate deposit of ₹67504.54 billion (US$1.1 trillion or €840 billion) and bank credit of ₹52604.59 billion (US$880 billion or €650 billion).

There are 1, 80,000+ ATM's across India, 57, 45,76,541 debit cards and 4,76,87,652 credit cards from all the banks. Around 38% Indians have a bank account. On an average there are around 75, 00, 00, 000 ATM transactions executed per month across India in the all banks. The net profit of the banks operating in India was ₹1027.51 billion (US$17 billion or €13 billion) against a turnover of ₹9148.59 billion (US$150 billion or €110 billion) for the fiscal year 2012-13.

2. CREDIT RISK

The credit function is the heart of banking. Interest income is the main source of income for any bank. Risk is inherent part of bank's business. Granting any loan to customer always involves some risk. A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments. In the first resort, the risk is that of the lender and includes lost principal and interest, disruption to cash flows, and increased collection costs.

Credit risk is the bank's risk of loss arising from a borrower who does not make payments as promised. Another term for credit risk is default risk. Effective risk management is critical to any bank for achieving financial soundness. Credit risk arises whenever a borrower is expecting to use future cash flows to pay a current debt. Banks are compensated for assuming credit risk by way of interest payments from the borrower or issuer of a debt obligation.

The lack of general credit review system in many banks and the lack of precise methods for measuring credit risk are two important reasons why an expert support system is necessary. It is with this spirit researchers have taken up the tasks for checking the applicability of the integrated model on the data collected from the Indian Banks.

Before approving the credit, proper evaluation process has to be followed. Before a potential debtor wants to obtain credit, he must be evaluated on certain areas. There are five C's involved in credit evaluation. As discussed in [1] they are: **credit report/score, character, collateral, capacity and cash flow.**

Important in a bank relationship is the "know your client" (**KYC**) principle. KYC is the main term used to gather and understand the customer profile. It is important that banks deal with customers with sound reputation and credit-worthiness. Therefore banks need not only manage the credit risk in their credit portfolio but also that in any individual credit or transaction.

Effective credit risk management encompasses identification, measurement, monitoring and control of the credit risk exposures. The effective management of credit risk is a critical component and essential for the long term success of a banking organization.

2.1 What is Credit Score?

A credit score is a number used by lenders as an indicator of how likely an individual is to repay his debts and the probability of going into default. Credit scoring can also be formally defined as a statistical (or quantitative) method that is used to predict the probability that a loan applicant or existing borrower will default or become delinquent [Mester, 1997]. It is an independent assessment of the individual's risk as a credit applicant. This helps to determine whether credit should be granted to a borrower [Morrison, 2004]. Credit scoring can also be defined as a systematic method for evaluating credit risk that provides a consistent analysis of the factors that have been determined to cause or affect the level of risk [Fensterstock, 2005]. A credit score is a numerical expression to represent the creditworthiness of the person.

Credit scoring is now used in almost all forms of consumer lending — credit cards, personal loans, car finance, insurance policies, and utility payments. Credit scoring can be defined as a technique that helps credit providers decide whether to grant credit to customers. Its increasing importance can be seen from the growing popularity and application of credit scoring in consumer credit.

The objective of credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively. In the United States, the Circuit Court has found considerable actuarial evidence that credit scores are a good predictor of risk of loss [Johnson-Speck, 2005]. Similarly, a recent actuarial study has concluded that credit scores are one of the most powerful predictors of risk; they are also the most accurate predictor of loss seen in a long time [Miller, 2003].

There are advantages not only to construct effective credit scoring models to help improve the bottom-line of credit providers, but also to combine models to yield a better performing combined model. This paper has two objectives. First, it illustrates the use of data mining techniques to construct credit scoring models. Second, it illustrates the combination of behavioral and financial credit scoring models to give a superior final model

2.2 Credit Scoring Companies in India

In India, there are four credit information companies licensed by Reserve Bank of India. The Credit Information Bureau (India) Limited (CIBIL) has functioned as a Credit Information Company from January 2001.[8] Subsequently in 2010, Experian,[8] Equifax[9] and Highmark[10] were given licenses by Reserve Bank of India to operate as Credit Information Companies in India.

The CIBIL TransUnion Score is a 3 digit numeric summary of your credit history which indicates your financial & credit health. The higher your score, the higher are the chances of your loan application getting approved. Your CIBIL Credit information report (CIR) is provided to you along with your score, because it is the basis on which your Credit Score is generated. It's a record of your credit history. i.e past loans or credit cards availed from various loan providers who are members of CIBIL.

Although all the four credit information companies have developed their individual credit scores, the most popular is CIBIL credit score. Lenders such as banks and credit card companies, use credit scores to evaluate the potential risk posed by lending money to consumers and to mitigate losses due to bad debt. Lenders use credit scores to determine who qualifies for a loan, at what interest rate, and what credit limits. Lenders also use credit scores to determine which customers are likely to bring in the most revenue. The use of credit or identity scoring prior to authorizing access or granting credit is an implementation of a trusted system.

The CIBIL TransUnion Score is a three-digit numeric summary (ranging from 300 on the lower side to 900 at the higher side) of a consumer's credit history, compiled from information received from lenders who are members of CIBIL. The model followed by CIBIL predicts the likelihood of an individual missing more than three payments on a credit line over the next 12 months.

Lenders are readily giving loans to customers who have a credit score of 800 and above, therefore, it is in the interest of a loan-seeker to maintain a healthy credit score.

A credit score provided by credit bureaus like Credit Information Bureau (India) Ltd (CIBIL) is an assessment of an individual's probability of default over a period of the next 12 months. The 'Score' will help credit institutions estimate the likelihood of repayment of loan, based on an individual's past pattern of credit usage and loan repayment behavior. The closer the score is to 900, the more confidence the credit institution will have in the individual's ability to repay the loan and hence, the better the chances of the individual's application getting approved.

According to information available, in 2013, the percentage of new loans sanctioned went up to 63% from 37% in 2008, for customers who have a credit score of 800 and above. This means more and more lenders are choosing to give loans to customers with better credit scores.

The higher the score, the more favorably it is viewed by credit institutions. However, every credit institution has its own benchmark of what constitutes a good credit score. CIBIL does not recommend any cut-off score for loan application eligibility.

Below is the grading of CIBIL credit scoring numbers:

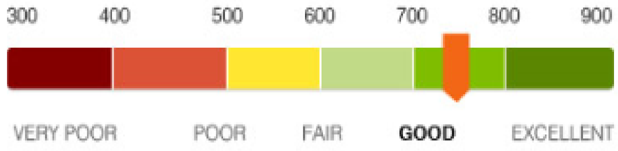


Fig 2.1: CIBIL Credit scoring categorization

2.3 Factors impacting Credit score

The exact formula for calculating credit scores is kept a secret by credit bureaus. Each major credit bureau in the US (Experian, Equifax, TransUnion) uses a proprietary tool developed by the Fair Issac Corp to calculate a FICO score. The components and their weightage in a credit score, as disclosed by FICO are:

1. Payment history -35%: Late payments on bills, such as a mortgage, credit card or automobile loan, can cause a FICO score to drop. Paying bills on time will improve your FICO score.
2. Credit utilization -30%: The ratio of current revolving debt (such as credit card balances) to the total available revolving credit, or credit limit. You can improve your FICO scores by paying off debt and lowering the credit utilization ratio.
3. **Length of credit history -15%:** As your credit history ages, it can have a positive impact on the FICO score.
4. Types of credit used -10%: You can benefit by having a history of managing different types of credit, like instalment, revolving, consumer finance and mortgage.
5. Recent search of credit -10%: Credit inquiries, which occur when you are seeking new credit, can hurt your score.

Credit scoring is not limited to banks. Other organizations, such as mobile phone companies, insurance companies, landlords, and government departments employ the same techniques. Credit scoring also has much overlap with data mining, which uses many similar techniques. These techniques combine thousands of factors but are similar or identical.

Credit score is just one factor used in the application process. Other factors apart from your credit report, such as your annual salary, length of employment, bankruptcy/litigation information, number of credit facilities may also be taken into consideration by lenders during a loan application.

2.4 Factors to keep in mind to improve the credit score

- **a)** Timely repayments:
- **b)** Age of Credit:
- **c)** How much credit limit is being used:
- **d)** Skewed towards Unsecured loans or Secured loans:
- **e)** On the lookout for more and more credit

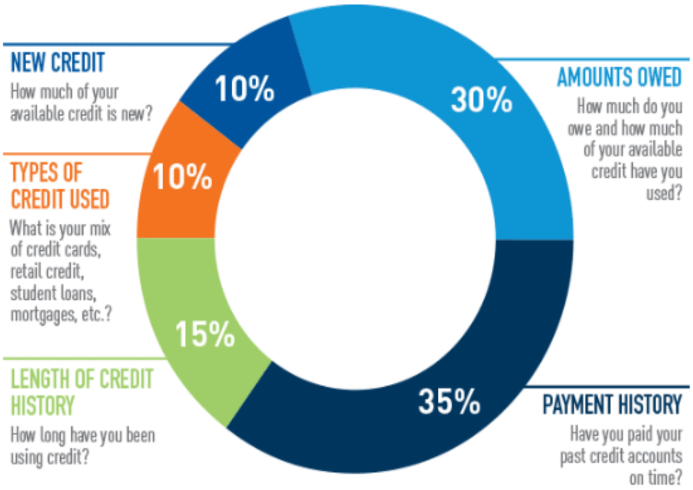


Fig 2.4 Factors impacting Credit score

By assessing your credit report, a lender makes up his mind on whether to approve your loan or to adjust their terms based on it or to decline the proposal completely. Believe it or not, but you can break or make your credit health depending on how well you manage your credit score. The more conscious you are of the way you handle your finances the better it will be for a credit friendly future.

2.5 Overview of the methods used for credit scoring and behavioral scoring

So what are the methods used in credit granting? Originally it was a purely judgmental approach. Credit analysts read the application form and said yes or no. Their decisions tended to be based on the view that what mattered are the 5Cs:

1) The character of the person — do you know the person or their family?
2) The capital - how much is being asked for?
3) The collateral - what is the applicant willing to put up from their own resources?
4) The capacity - what is their repaying ability. How much free income do they have?
5) The cashflow - what are the conditions in

Behavioral scoring systems allow lenders to make better decisions in managing existing clients by forecasting their future performance. The decisions to be made include what credit limit to assign, whether to market new products to these particular clients, and if the account turns bad how to manage the recovery of the debt. The extra information in behavioral scoring systems compared with credit scoring systems is the repayment and ordering history of this customer. Behavioral scoring models split into two approaches — those which seek to use the credit scoring methods but with these extra variable added, and those which build probability models of customer behavior. The latter also split into two classes depending on whether the information to estimate the parameters is obtained from the sample of previous customers or is obtained by Bayesian methods which update the firm’s belief in the light of the customer’s own behavior. In both cases the models are essentially Markov chains in which the customer jumps from state to state depending on his behavior.

Credit scoring and behavioral scoring are the techniques that help organizations decide whether or not to grant credit to consumers who apply to them. There are two types of decisions that firms who lend to consumers have to make. Firstly should they grant credit to a new application. The tools that aid this decision are called credit scoring methods. The second type of decision is how to deal with existing customers. If an existing customer wants to increase his credit limit should the firm agree to that? What marketing if any should the firm aim at that customer? If the customer starts to fall behind in his repayments what actions should the firm take? Techniques that help with these decisions are called behavioral scoring

In the credit scoring approaches to behavioral scoring one uses the credit scoring variables and includes others which describe the behavior. These are got from the sample histories by picking some point of time as the observation point. The time preceding this say the previous 12 months — is the performance period and variables are added which describe what happened then — average balance, number of payments missed. etc. A time some 18 months or so after the observation point is taken as the performance point and customer’s behavior by then is assessed as good at good or bad in the usual way.

Hopper and Lewis (1992) give a careful account of how behavioral scoring systems are used in practice and also how new systems can be introduced.

3. LITERATURE REVIEW

3.1 Data Mining

Data Mining refers to extracting knowledge, hidden trends and patterns from large amounts of data. Data Mining is about explaining the past and predicting the future by means of data analysis. Data mining is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology. The various steps involved in extracting knowledge from raw data as depicted in figure-1. There are different types of data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. [5]. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large [7]. Credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision Tree based Classification Algorithms. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.

Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. [5].

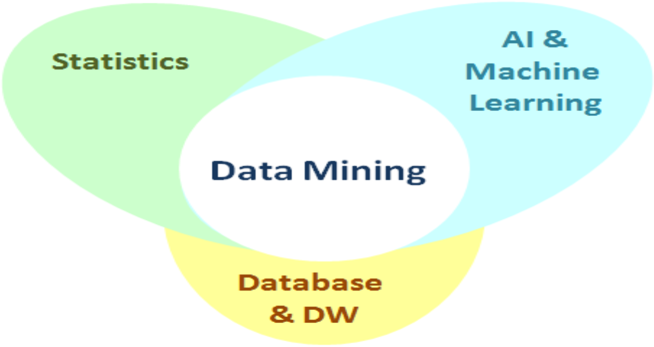


Fig 3.1: Data Mining-mix of core areas

Below flow diagram explains the steps involved in knowledge discovery using Data Mining.

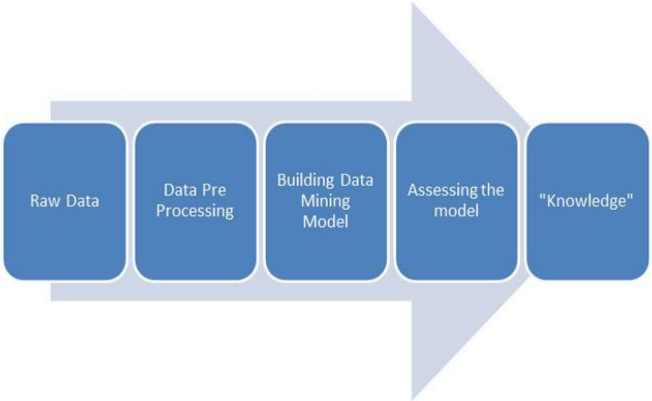


Fig. 3.1.1 Steps in Knowledge Extraction

3.2 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large [7]. Fraud detection and credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision Tree based Classification Algorithms. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

In general, in classification you have a set of predefined classes and want to know which class a new object belongs to. Clustering tries to group a set of objects and find whether there is some relationship between the objects. In the context of machine learning, classification is supervised learning and clustering is unsupervised learning.

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction.

3.3 Decision Tree

A Decision Tree is most popular classification technique. It is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. An example of Decision Tree is depicted in figure3.3.

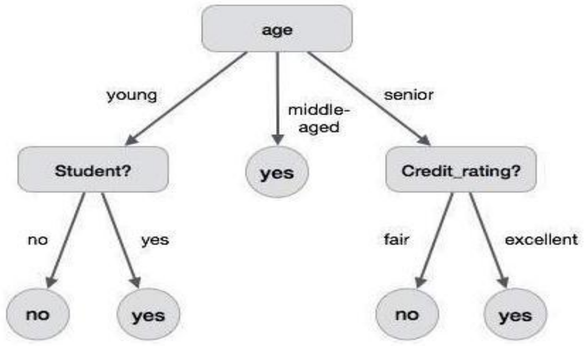


Fig. 3.3. Decision Tree Induction

Decision trees are particularly useful for classification tasks. Like Radial Basis Neural Networks, decision trees learn from data. Using search heuristics, decision trees are able to find explicit and understandable rules-like relationships among independent and dependent variables. The purpose of the logistic regression model is to obtain a regression equation that could predict in which of two or more groups an object could be placed (i.e. whether a credit should be classified as approved or rejected).

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute [3].

Advantages of using decision learning tree algorithms are:
1) They generalize in a better way for unobserved instances, once examined the attribute value pair in the training data.
2) They are efficient in computation as it is proportional to the number of training instances observed.
3) The tree interpretation gives a good understanding of how to classify instances based on attributes arranged on the basis of information they provide and makes the classification process self-evident.

The operation of decision tree is based on ID3 or C4.5 algorithms. It builds tree based on the information (information gain) obtained from the training instances and then uses the same to classify the test data. ID3 algorithm generally uses nominal attributes for classification with no missing values. ID3 can even work well on datasets with missing attribute values to certain extent.

C4.5 handles both continuous and discrete attributes. While handling the data C4.5 allows missing attribute values to be marked as (?). Missing attribute values are simply not used in gain and entropy calculations. Decision tree are self-explanatory and can be easily converted to set of rules so they are used in credit evaluation process.

Artificial neural networks are one of the most common data mining tools. Neural networks are particularly useful for the tasks of classification, prediction, and clustering in business applications. Neural network models are characterized by three properties: the computational property, the Credit risks which account for the risk of loss and loan defaults are the major source of risk encountered by banking industry [2]. Data mining techniques like classification and prediction can be applied to overcome this to a great extent. It is an area in which even a small performance improvement can mean a tremendous increase in profit to the lender because of the volume and quantity of the lending amounts.
Today, banks are realizing the various advantages of data mining. It is a valuable tool by which banks can identify potentially useful information from the large amounts of data. This can help banks to gain a clear advantage over its competitors. Data mining can help banks in better understanding of the vast volume of data collected by the CRM systems.

3.4 Data Mining in Banking

Due to tremendous growth in data the banking industry deals with, analysis and transformation of the data into useful knowledge has become a task beyond human ability [9]. Data mining techniques can be adopted in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases [7]. 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 customers are likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable [4]. Globalization and the stiff competition had led the banks focus towards customer retention and fraud prevention. To help them for the same, data mining is used. By analyzing the past data, data mining can help banks to predict credible customers. Thus they can prevent frauds; they can also plan for launching different special offers to retain those customers who are credible. Certain areas that effectively utilize data mining in banking industry are

a) Risk Management
b) Customer Relationship Management.
c) Marketing
d) Default Detection
e) Demand forecasting
f) Non-Performing Loans prediction
g) Anti-Money Laundering

3.4 Retail loans

Retail loans are those loans which are given by the banks to individuals so as to meet there personal needs, retail loans are smaller in size as compared to corporate loans. Given below are various types of retail loans which are given by the banks –

Housing Loans – Most individuals take housing loans and when it comes to retail loans, housing loans is right there at the top. Banks give housing loans to individuals so that can buy apartment or construct new house if they already have the land.

Educational Loans – This type of loans is given by the banks to students so that they can pay for the tuition fees, hostel expenses, foreign education and other such expenses.

Vehicle or Auto Loans – This type of loans are given to individuals who are looking for buying cars whether new or second hand, auto loans are also given for two wheelers to individuals.

Personal Loans – Personal loan are the loans which are given to individuals for purposes such as marriage, traveling to abroad, loans for covering hospital expenses and other such loans which individual may need depending on his or her needs and situations.

3.5. Secured Loans and Unsecured Loans

In the secured loans, the borrower has to pledge some assets (such as property) as collateral. Most common secured loan is Mortgage loan in which people mortgage their property or asset to get loans. Other example is Gold Loan, Car Loan, Housing loan etc.

In unsecured loans, the borrower's assets are not pledged as collateral. Examples of such loans are personal loans, education loans, credit cards etc. They are given out on the basis of credit worthiness of the borrowers. We note here that the interest rates on unsecured loans are higher than the secured loans. This is mainly because the options for recourse for lender in case of unsecured loans are limited

The growth in retail banking has been quite prominent retail in the recent years. Retail banking has been supported by growth in banking technology and automation of the banking process. The company A.T. Kearney, a global management consulting firm, has identified India as the second most attractive retail destination out of 30 emergent markets. The considerable recent retail banking growth in India is expected to continue in the future. Retail lending is the exhortation in India. Most banks have the retail segment on around 20% of their total lending portfolio, being this segment growing at an unnatural rate of 30 to 35% per annum. Retail lending has been the key profit driver in the banking sector in recent times.

4. Motivation for designing two step credit risk assessment model for retail bank loan applications processing

4.1 Significance and Objective of Study

Credit scoring is a very important task for lenders to evaluate the credit applications they receive from customers as well as for insurance companies, which use scoring systems today to evaluate new policy holders and the risks these prospective customers might present to the insurer.

Credit scoring systems are used to model the potential risk of credit applications, which have the advantage of being able to handle a large volume of credit applications quickly with minimal labor, thus reducing operating costs, and they may be an effective substitute for the use of judgment among inexperienced credit officers, thus helping to control bad debt losses. This study explores the performance of credit scoring models using approaches: Logistic regression, Multilayer Perceptron Model, Radial basis neural network, SVM and decision trees (C4.5).

Data mining techniques have been applied to solve classification problems in [7] for a variety of applications such as credit scoring, bankruptcy prediction, insurance underwriting, and management fraud detection. The lack of research in combining data mining techniques with domain knowledge has prompted researchers to identify the fusion of data mining and knowledge-based expert systems as an important future direction. Here by combining the advantages of data mining classification methods logistic regression, decision tree, Multilayer Perceptron Model, SVM and radial basis neural network a new integrated model will generate which will be best for a credit approval system.

Objective of the study is to find the Integrated model which can be constructed by combining the advantages of existing credit scoring model and behavior scoring model to be applied for different types of credits for credit approval process such that there will be minimum defaulters and credit risks.

4.2 Existing Credit Approval Process

When it is required to obtain credit scoring, one has to undergo a process of evaluation before the credit score is sanctioned. This process is called as credit evaluation, which may take time, but concludes in either an approval or a rejection.

Before a potential debtor wants to obtain credit, he must be evaluated on certain areas. There are five C's involved in credit evaluation. As discussed in [1] they are: credit report, character, capacity, cash flow, and collateral. Below is the existing workflow credit approval process followed by bankers.

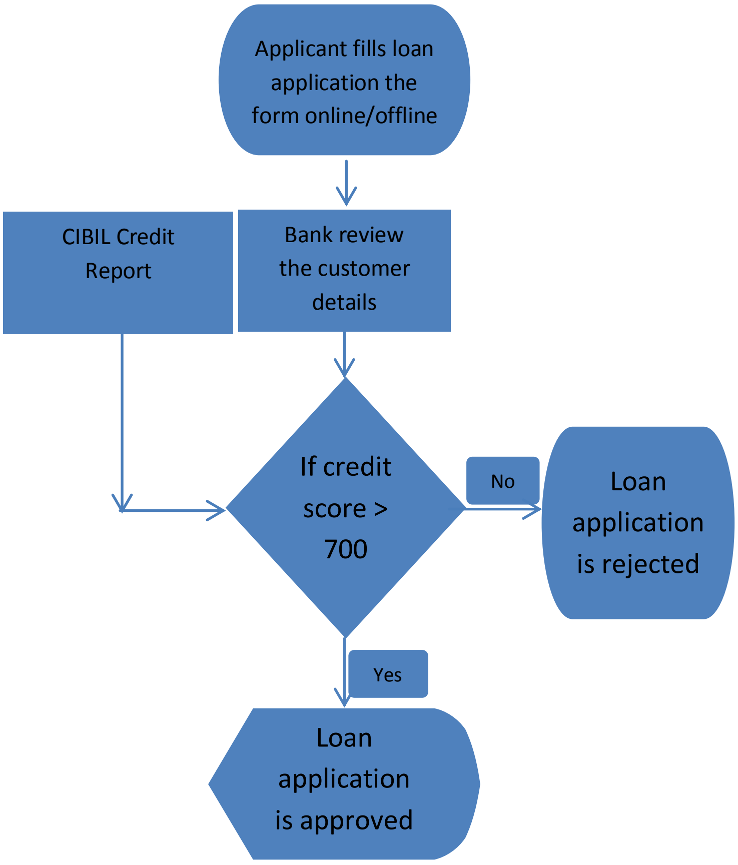

Applicant fills loan application the form online/offline
CIBIL Credit Report
Bank review the customer details
If credit score > 700
No
Loan application is rejected
Yes
Loan application is approved

Fig 4.2: Existing Credit Approval Process

The character of a person applying for a credit is a big factor to the decision for credit approval. A person with a sound financial objective is likely to be granted a credit approval quickly and are possibly than an individual who is in bad shape, not just on the financial facet, but also on other aspect.

Credit history is another important factor considered by lenders in their decision to grant and approve credit applications. The credit report is a record of an individual's past borrowing and reimbursing transactions. It also includes information about late payments and bankruptcy.

A credit report can be tarnished. A credit score can be at its low. Under these circumstances it is unlikely for you to earn the confidence of the lender for a credit approval. However, if your cash flow is good, there is a possibility of getting the credit approval. Lenders may also have to check the liquidity of an individual. This can be done by checking the bank statements of an individual borrower. In the case of businesses, lenders may have to obtain a copy of the audited financial statements. The financial statements of businesses and bank statements can be utilized to show the capacity of a borrower to settle and repay a line of credit. The capacity of the borrower to pay a credit is determined during credit evaluation and approval.

Collateral is a common term in credit. A lender seeks for security whenever the borrower defaults the credit payment. If no collateral is present as security for a credit, it is likely that the lender will give the borrower a high-interest rate credit. Credit evaluation is a process taken by the lender with the participation of the credit applicant. If you want to undergo this process, it is important to make substantial preparation so you are more likely to obtain a credit quickly and less expensively.

4.3 PROBLEM DEFINITION

The credit function is the heart of banking, under the ever changing market conditions. The lack of general credit review system in many banks and the lack of precise methods for measuring credit risk are two important reasons why an expert support system is necessary. It is with this spirit researchers have taken up the tasks for checking the applicability of the integrated model on the data collected from the Indian Banks.

CIBIL is already providing all the credit scores for existing customers. CIBIL credit score is built by not considering the behavior/character of the applicant. Another drawback of using only CIBIL model is that there is no credit score for new customers and it takes lot of time to build credit scores for new customers. Due to this limitation banks are unable to lend to new customers in right time and taking long time to process the new loan applications.

Assessing the personality of the applicant is very important before granting the loan to predict credit risks [4]. By combining the CIBIL credit score model with behavior scoring model, banks can leverage the power of technology to grant the loans to right customer at right time and reduce the risk up to great extent.

5. TWO STEP CREDIT RISK ASSESMENT MODEL FOR RETAIL BANK LOAN APPLICATIONS USING DECISION TREE DATA MINING TECHNIQUE

The main objective of the new proposed model is leveraging the power of credit scoring and behavior scoring/analysis. The new model classifies the loan applications using credit scoring and behavior scoring.

5.1 Research Methodology

The reference model for our work is cross-industry standard process for data mining (CRISP-DM) fig 3 [2] which is a well-known to develop Data Mining projects. The proposed model focuses on predicting the credibility of customers for loan repayment by analyzing their behavior. The input to the model is the customer behavior collected. Based on the output from the classifier, decision on whether to approve or reject the customer request can be made. Decision Tree Induction data mining technique is used to generate the relevant attributes and also make the decision in the model. Data mining model of the proposed system is as depicted in figure4.

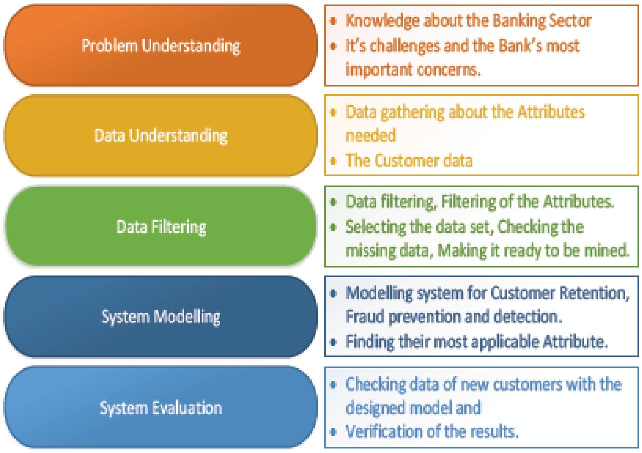

Problem Understanding
Knowledge about the Banking Sector
It's challenges and the Bank's most important concerns.
Data Understanding
Data gathering about the Attributes needed
The Customer data
Data Filtering
Data filtering, Filtering of the Attributes.
Selecting the data set, Checking the missing data, Making it ready to be mined.
System Modelling
Modelling system for Customer Retention, Fraud prevention and detection.
Finding their most applicable Attribute.
System Evaluation
Checking data of new customers with the designed model and
Verification of the results.

Fig 5.1: Data mining model

A. Problem Understanding

The data mining model is initiated with collection of details regarding the banking sector and the existing loan processing procedures. The challenges and the main risks associated with the loan approval/rejection in banking sector are thus better understood.

B. Data Understanding

The bank dataset of customer details which are required for data mining are collected and got familiarized with. Various attributes needed are also studied.

C. Data Filtering

The attributes in the bank data set are filtered and the relevant attributes needed for prediction are selected. After that the incomplete and noisy records in the dataset are removed and prepared for mining.

D. System Modelling

In this stage the system is developed in an efficient and user-friendly manner so that even those users with less technical knowledge can also use it comfortably. The system provides the most relevant attributes that help in determining whether to approve or reject the loan application. This aids in predicting the credibility of future customers.

E. System Evaluation

In the final stage, the designed system is tested with test set and the performance is assured.

5.2 Identification of Independent and Dependent Variables

The data set used in this research is divided into training and testing data sets. All training cases are set by default taking into account the banks' guidelines for personal credit approval in the banks. Data used is of 500 customer's data. The data required for the current study was collected from different banks such as SBI, IDBI, AXIS and Syndicate banks. It consists of different independent variables and one dependent variable.

Variables are the conditions or characteristics that he investigator manipulates, controls or observes. It is necessary to optimize variables by using SVM as mentioned in [8], [9], [10]. Variables are classified as dependent and independent variables. An independent variable is the condition or characteristic that affects one or more dependent variables: its size, number, length or whatever exists independently and is not affected by the other variable. A dependent variable changes as a result of changes to the independent variable.

Independent Variables:

1) Age of customer
2) Sex
3) Marital status
4) Saving account
6) Occupation
7) Home ownership
8) Time at Address
9) Education Qualification
10) Parent education
11) Family Occupation
12) No of Dependents
13) Loan Type
14) Place of living
15) Type of Employment
16) Company Name
17) Service period
18) Total years of experience
19) Monthly Take home salary
20) Relationship with Bank
21) Type of Loan

Dependent variable:

1) Credit (Approved or Not)

Using this data set, a model is built, which consists of a decision tree model (C4.5) to predict whether a future applicant's a credit is approved or rejected.

We can use the decision tree node to classify observations by segmenting the data created according to a series of simple rules. We can use the entropy gain reduction method to build the tree. The regression node fitted the logistic regression model to the data. The ensemble node combines the five models by averaging the posterior probabilities for the class target variable.

5.3 Decision Process for Credit Evaluation

Credit managers rely heavily upon external data sources to guide them in the credit decision process. To approve or reject a credit request is a delicate task. A credit manager must evaluate the risk associated with extending credit and declining an applicant based on numerous factors [2].
The need for sufficient and reliable information is the foundation of a successful credit decision. A credit manager may call on references, run background checks, pull a credit report, verify bank accounts or ask questions of the applicant to validate the information on the credit application. Credit managers are challenged with the task of obtaining readily available information to support their decision while sending a timely response to the applicant. A major obstacle in achieving this task is the turnaround time associated with checking references. The process varies from business to business and may include a background check, a verification of a bank deposit or credit references with existing suppliers.

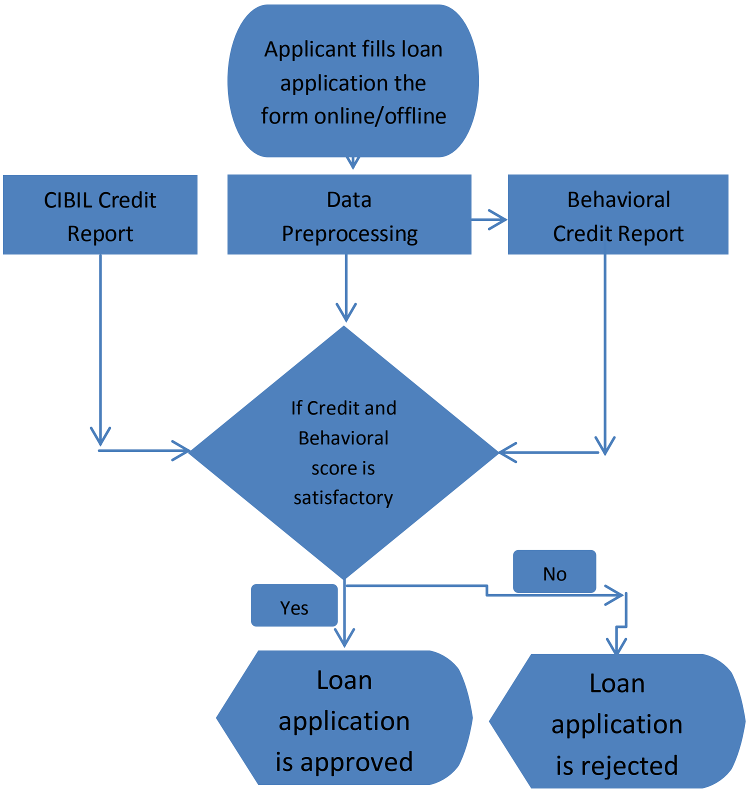

Applicant fills loan application the form online/offline
CIBIL Credit Report
Data Preprocessing
Behavioral Credit Report
If Credit and Behavioral score is satisfactory
Yes
No
Loan application is approved
Loan application is rejected

Fig 5.3: Architecture of the proposed model

5.4 Requirements and Architecture of proposed Model

Credit evaluation system requires the following four components to work with data: computer system, WEKA software, CIBIL credit score and customer data. Credit valuation system can be built by combining the advantages of logistic regression and decision tree.

5.4.1 J48 Algorithm

J48 builds decision trees from a set of labelled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created to the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case j48 creates a decision node higher up in the tree using expected value of the class. j48 can handle both continuous and discrete attributes; training data with missing attribute values and attributes with differing costs. Architecture of the two step credit risk assessment model is as shown in the figure5.

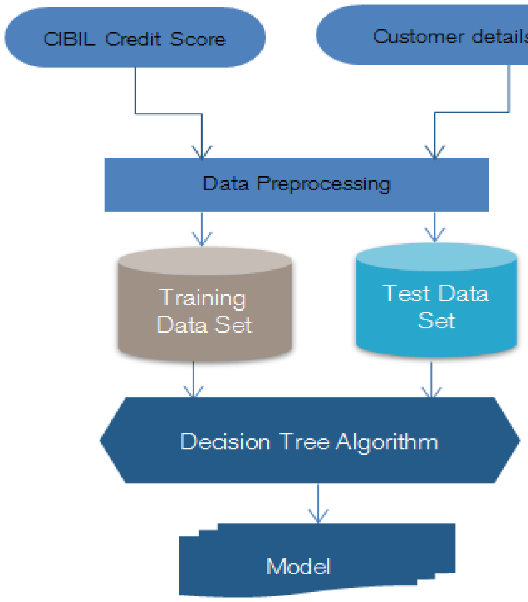


Fig 5.4: Proposed two-step credit risk assessment model

A. Input

The main highlight of this Loan Credibility Prediction System is that it uses Decision Tree Induction Data Mining Algorithm to screen/filter out the loan requests. A Decision Tree is developed by performing data mining on an existing bank dataset containing 1140 records and 24 attributes.

B. Data Pre-processing

Initially the Attributes which are critical to make a Loan Credibility Prediction is identified with Information Gain as the attribute-evaluator and Ranker as the search-method. Manual preprocessing is also performed.

C. Data Filtering

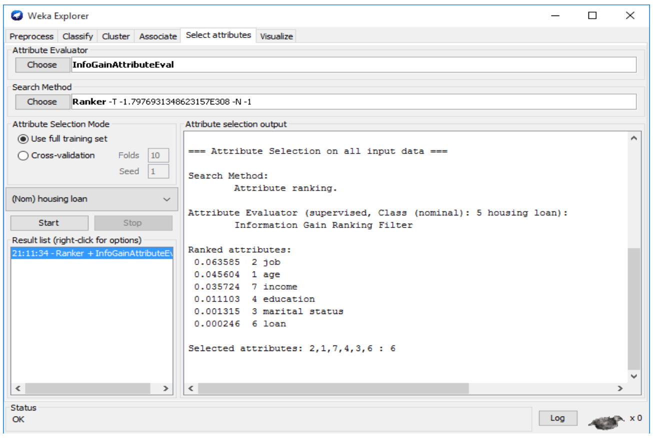
Final dataset after preprocessing is divided in such a way that there is 66 % training set and 34 % test set. Test set is used to validate the final result of the classifier.

D. Decision Tree Algorithm

An efficient Decision Tree is formulated with Decision Tree Induction Algorithm. It produces a model with the most relevant 6 attributes. Attribute with rank-1 is placed as the root node of the Decision tree, other attributes from Rank-2 to Rank-24 constitute the intermediate nodes. A decision is made at each node and the leaf node gives us the final result. That is, if the customer possesses the minimum loan repayment capacity, then the future risks can be avoided. The main benefit of applying Data Mining is that we can always rely on the result of the algorithm to accept or reject the loan application.

6. Experimental results and Prototype validation

The results of the experimental analysis in predicting the loan repayment capacity are presented in this section. We have implemented our proposed model in WEKA. An existing bank dataset has been used for the prediction. We have used a bank dataset of moderate size (1140) for the experimental analysis. After the pre-processing phase where dimensionality reduction was done manually and the dataset was reduced to a size of 854. Ranks of the attributes are found out by manually adding and applying Information Gain as attribute evaluator and Ranker as search.



InfoGainAttributeEval

Ranker -T -1.7976931348623157E308 -N -1

● Use full training set
○ Cross-validation Folds 10 Seed 1

(Nom) housing loan

21:11:34 - Ranker + InfoGainAttributeEv

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 housing loan):
        Information Gain Ranking Filter

Ranked attributes:
 0.063585  2 job
 0.045604  1 age
 0.035724  7 income
 0.011103  4 education
 0.001315  3 marital status
 0.000246  6 loan

Selected attributes: 2,1,7,4,3,6 : 6
```

Fig. 6.1 Attribute Ranking

The ranks of the Attributes (generated using Ranker) are as listed in the following table1.

Rank	Attribute
1	Job
2	Age
3	Education
4	Marital Status
5	Spouse Education
6	Gender
7	Occupation
8	Home Ownership
9	Family Occupation
10	No of Dependents
11	Purpose of the loan
12	Income Level
13	Time at Address
14	Type of Employment
15	Current Employer
16	Employment Period
17	Monthly Salary
18	Relationship with Bank
19	Place of Living
20	Parents Education Details
21	Loan Type
22	Saving Account AQB
23	Loans from Other Banks
24	Previous Employer

Fig. 6.2: Relevant attributes along with rank (Table 1)

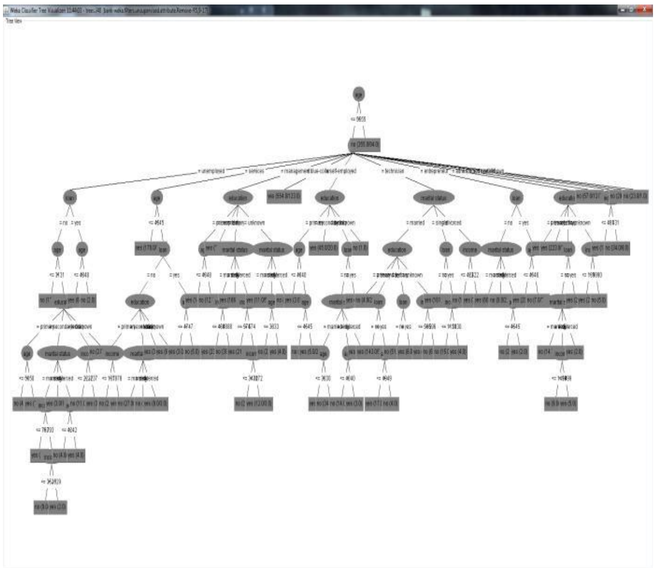
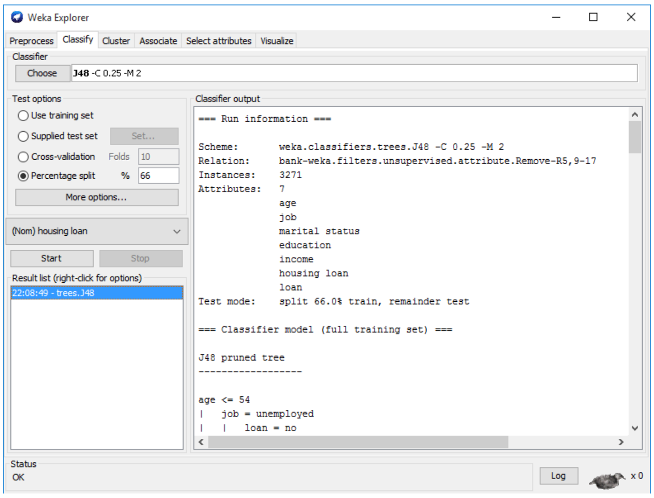


Fig.6.3: Decision Tree



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) housing loan

22:08:49 - trees.J48

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     bank-weka.filters.unsupervised.attribute.Remove-R5,9-17
Instances:    3271
Attributes:   7
              age
              job
              marital status
              education
              income
              housing loan
              loan
Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

age <= 54
|   job = unemployed
|   |   loan = no
```

Fig. 6.4: Decision Tree generation

The following figure shows the performance measure of the model for varying number of customers.

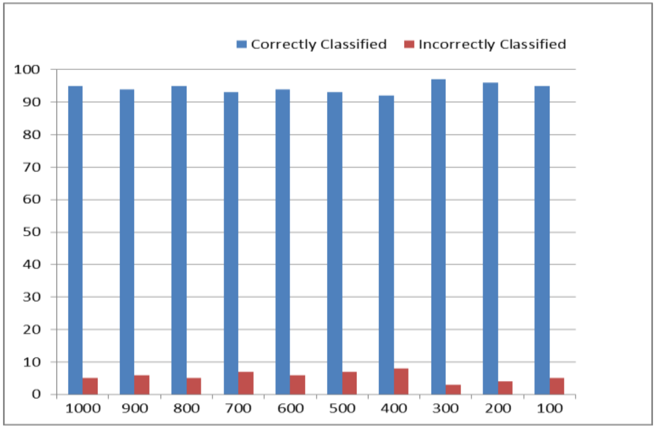


Fig. 6.5: Accuracy vs. number of customers

7. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have presented a two-step loan credibility prediction system that helps the organizations in making the right decision to approve or reject the loan request of the customers. This will definitely help the banking industry to open up efficient delivery channels. Decision Tree Induction Algorithm is used for the prediction. Incorporation of other techniques that outperform the performance of popular data mining models has to be implemented and tested for the domain. Data mining is the process to extract knowledge from existing data. It is used as a tool in banking and finance in general to discover useful information. Credit risk management is critical for successful bank lending. We attempt to model the loan approval process at one of India's midsized banks. We obtained statistically significant linear and nonlinear models to accomplish the above.

A two-step credit scoring or combined credit scoring model is very useful and accurately classifies the loan applications using traditional credit scoring and improved behavior scoring. This model is very useful in decision making for approving loan applications for existing and new customers. We propose to extend the two step credit approval model by including collateral, capacity and cash-flow parameters in the future research areas.

6. REFERENCES

[1] Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4

[2] Dileep B. Desai, Dr. R.V.Kulkarni "A Review: Application of Data Mining Tools in CRM for Selected Banks", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2)
, 2011, 199 – 201.

[3] Bhambri, V., 2011. Application of data mining in banking sector. Internat. J. Comput. Sci. Technol., 2:199-201.

[4] Chopra, B., V. Bhambri and B. Krishnan, 2011. Implementation of data mining techniques for strategic CRM issues. Int. J. Comput. Technol.Appli., 2: 879-883.

[5] Dr. K. Chitra1, B. Subashini , "Data Mining Techniques and its Applications in Banking Sector " , International Journal of Emerging Technology and Advanced Engineering(ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013)

[6] Deshpande, M.S.P. and D.V.M. Thakare, 2010. Data mining system and applications: A review. Int. J. Distrib. Parallel Syst., 1: 32-44.

[7] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Pre-processing for Supervised Leaning", International Journal of Computer Science, 2006, Vol 1 N. 2, pp 111–117.

[8] Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4

[9] Vivek Bhambri "Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology Vol. 2, Issue 2, June 2011

[10] P.Sundari, Dr.K.Thangadurai "An Empirical Study on Data Mining Applications", Global Journal of Computer Science and Technology, Vol. 10 Issue 5 Ver. 1.0 July 2010.

[11] Kazi Imran Moin, Dr. Qazi Baseer Ahmed "Use of Data Mining in Banking",International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,vol.2,Issue2,Mar-Apr2012, pp.738-742 738

[12] Rajanish Dass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad.

[13] Hamid Eslami Nosratabadi and Ahmad Nadali ,"A New Approach for Labeling the Class of Bank Credit Customers via Classification Method in Data Mining", International Journal of Information and Education Technology, Vol. 1, No. 2, June 2011

[14] Hamid Eslami Nosratabadi and Ahmad Nadali ,"A New Approach for Labeling the Class of Bank Credit Customers via Classification Method in Data Mining", International Journal of Information and Education Technology, Vol. 1, No. 2, June 2011

[15] Costa, G., F. Folino and R. Ortale, 2007. Data mining for effective risk analysis in a bank intelligence scenario. Preccedings of the 23rd International Conference on Data Engineering Workshop, Apr. 17-20, IEEE Xplore Press, Istanbul,pp:904-911. DOI: 10.1109/ICDEW.2007. 4401083