National Conference on Recent Advances in Computer Science and IT (NCRACIT)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 1 | ISSN : 2456-3307

Predictive Risk Categorization of Retail Bank Loans Using Data Mining Techniques

M Mubasher Hassan[1], Tabasum M[2]

[1]Dept. of ITE BGSB University Rajouri, Jammu & Kashmir, India
[2]Education Department, Govt. of J&K, Jammu & Kashmir, India
mubasher2003@gmail.com[1], tabasum.mirza@gmail.com[2]

ABSTRACT

In the present highly competitive environment of the banking industry, reducing default and preventing NPA loans in retail banking is a major challenge. Data mining techniques are already popular in different banking sectors for mining important information to discover knowledge that can be used for marketing, analysis and predictive purpose for tremendous available data of already existing customers. We are using classification algorithms SVM, CART, j48 algorithm for predicting risk and then categorization of loan customer into any of three risk categories i.e 'low risk', 'medium risk' and 'high risk'. The risk category of customer will be used as a suggestive indicator for customization of the repayment schedule and follow up procedure required.

Keywords: Data Mining, Loan Prediction, Classification Algorithm, Credit Risk Assessment and J48 Algorithm

I. INTRODUCTION

Lending advances remain as the most important business of banking. One of the fundamental tasks of the bank is to sanction credit facility to customers as per requirement and most important cause of poor credit quality is ineffective credit risk assessment system. Data mining techniques can be used by analyzing repayment patterns and credit history of loan customers in previous loans availed or other credit products availed currently. The knowledge derived by implementing data mining can be used for predicting and evaluating various parameters associated with the repayment nature of the customer. Data mining techniques can be used for predicting credit risk associated with loans and then the classification of loans based on the amount of risk associated with them [2]. Risk assessment can be done assessing different parameters like credit history of prospective customers, his repayment pattern in previous loans or other credit products availed etc [14]. Based on the risk perceived we can classify prospective loan customer into different categories i.e; 'low risk', 'average risk' and 'high risk'. This classification will be used to customize repayment schedules for customers falling into different categories. Customized repayment schedules will vary in EMI amount, repayment period, etc. Risk category will be indicative for monitoring for the loan, whether weekly, monthly or quarterly follow-up is required for the loan. High risk loans need regular strict monitoring procedure, average risk loans needs average supervision while as low risk loans need lenient follow-up after some intervals i.e. follow-up after weeks or months [3].

II. DATA MINING TECHNIQUES

Data mining is the computing process that explores large sets of data and extracts different data patterns by analyzing relationships between the parameters of data to discover knowledge. This predictive knowledge can be used in different applications.

III. STAGES OF DATA MINING

A. Data Cleaning

Real world data is in random form which cannot be useful on its own. So, this data needs to be inspected, refined, cleaned from errors, inconsistent and redundant values and arranged into useful attributes. The output of this stage is cleansed data that can be further processed.

B. Data Integration

In this phase data from different sources stored in different formats is integrated so that redundancy is minimized.

C. Data selection

Selecting required data sets from enormous data available that can be analyzed properly.

D. Data transformation

Involves transformation of data by normalization, aggregation and generalization process to be used for data mining purpose.

E. Pattern evaluation

This stage identifies hidden patterns in data that can be used for deriving knowledge categorizing data from data sets

F. Knowledge Representation

This process extracts and represents knowledge in an understandable form for application use.

IV.DATA MINING ALGORITHMS FOR CLASSIFICATION

A. J48 algorithm

J 48 is an open source JAVA implementation of C4.5 algorithm mainly used for classification problems. This algorithm generates rules from a particular identity of data in a data set of prediction of the target variable with the help of decision trees [11]. The objective is a generalization of decision tree to achieve equilibrium. Extension of ID3. Additional features incorporated in this algorithm are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. Determination of the range of the communication on the device is based on sequential value, reduces search time when the data elements are sorted the value is regenerated for each iteration with minimal tree approach but algorithm requires prior sorting of data elements [1]

B. Support Vector Machine (SVM)

SVM are supervised machine learning algorithms that are used for classification and regression of data and generates accurate results when sample size is small and the data set is well segregated and are resistant to over fitting. It's mainly used in classification problems works effectively with high dimensional data. SVM uses kernel functions for transformations and then classifies data by finding the optimal boundary between the possible outputs [12].SVM is able to model complex nonlinear decision boundaries and effectiveness of SVM depends on the selection of kernel parameters

C. Classification and Regression Tress (Cart)

This data-mining algorithm uses decision trees, learning technique for classification and regression of input data by predicting the value of the target or dependent variable based on input (independent variables) to identify class within which target variable would fall into and using regression trees where the target variable is continuous and tree is used to predict its value. This algorithm used in combination with other prediction methods for classification or regression predictive modeling machine learning. CART incorporates both testing and validation of the test data set for accurate assessment [13]. CART is based on binary splitting of attributes, inbuilt features that deal with missing attributes uses both numeric and categorical attributes for building the decision tree [5].

V. WEKA

Is an open source machine learning software collection of algorithms used for solving real life data mining problems? The algorithms can be applied directly to data set or implemented in JAVA code

VI. CONCEPTUAL FRAMEWORK

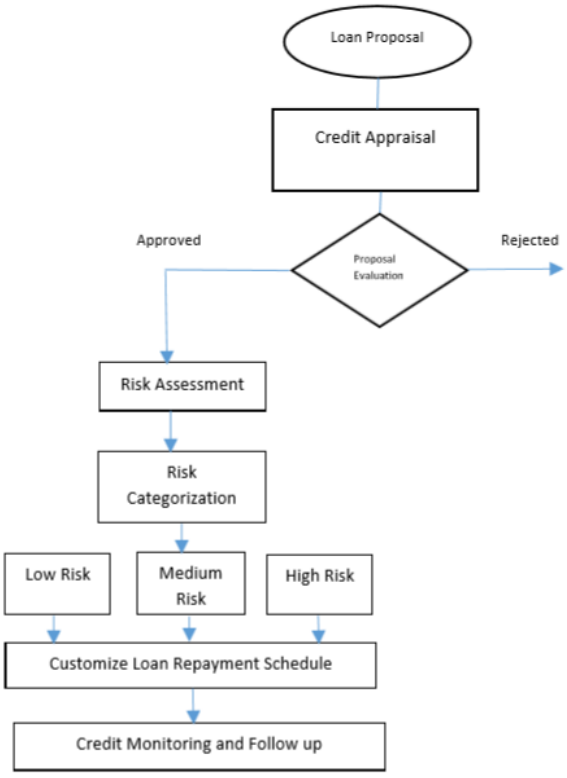


Fig. 1. Proposed Risk Categorization Process

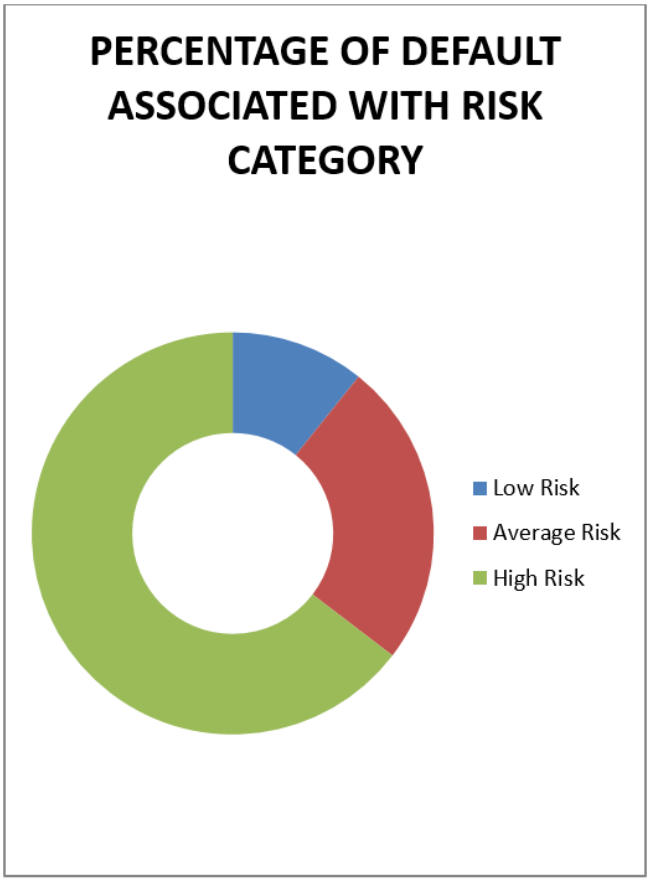
The methodology uses three algorithms SVM, CART and j 48 for classification and regression of customer data [6][4].Decision trees are used to decide the outcome of the loan application after appraisal of loan whether loan request should be recommended for sanction or rejection [8]. The decision tree will use variables in table 1.1 to decide the outcome. For the loans recommended for sanction, second step is the risk assessment that will be done by calculating risk indicator associated with loan requested by customer by quantifying variables in Table 1.2. At third step data mining classification algorithm, SVM,CART and mainly j48 algorithm is used to classify prospective loan customer to fall into one of the three risk categories i.e; high risk ,low risk and medium risk category[10]. Categorization of loan customer will help in loan repayment prediction and categorization will be used for customizing repayment schedule of the customer and will suggest follow up procedure for the prospective loan customer [7].

Table 1

Bank Savings/Deposits	Monthly income	Source of income regular	Income From other sources	Total monthly income	No of dependents	Nature of loan	Purpose of loan	Total liabilities	Total assets	Contingent liabilities	Status of credit products availed	Loan amount requested
70000	30000	yes	9000	39000	4	secured	car	200000	30000	500000	satisfactory	350000
15000	20000	no	15000	25000	2	secured	house	150000	200000	300000	satisfactory	500000
200000	40000	yes	0	40000	3	unsecured	Personal Consumption	350000	150000	250000	Not satisfactory	600000

Table 2

Outstanding Loan amount	History of default in previous/ current loans	Overflow amount	Underflow amount	Installments pending in other loans availed	Premature repayment history	Un availed credit amount (If applicable)	NPA history
17000	NO	3000	0	0	YES	NO	NO
30000	NO	0	2000	2000	NO	NO	NO
27000	YES	0	7000	7000	NO	75000	NO



VII. CONCLUSION

In this paper, we have proposed two stage predictive model one for deciding the outcome of a loan application and other for risk assessment of loan for categorization of loan based on the risk perceived. We are using decision trees in SVM algorithm for stage 1 and j48, CART algorithm in combination with an SVM algorithm for classification of loans based on risk indicator field. This model is useful in deciding whether the loan should be recommended for sanction or not and then if sanctioned in which risk category loan will fall so that customized repayment schedule is designed for loan and risk indicator will also suggest a monitoring procedure and schedule of the loan.

VIII. REFERENCES

[1] Bharat Deshmukh, Ajay S. Patil & B.V. Pawar IJCSIT International Journal of Computer Science and Information Technology, Vol. 4, No. 2, December 2011, pp. 85-90 Comparison of Classification Algorithms using WEKA on various Datasets

[2] Aboobyda Jafar Hamid and Tarig Mohammed Ahmed Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1, March 2016 DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING

[3] Sudhakar M, Dr. C. V. K Reddy Global Journal of Computer Science and Technology: C Software & Data Engineering Volume 14 Issue 5 Version 1.0 Year 2014, TWO STEP CREDIT RISK ASSESMENT MODEL FOR RETAIL BANK LOAN APPLICATIONS USING DECISION TREE DATA MINING TECHNIQUE

[4] Bharat Deshmukh, Ajay S. Patil & B.V. Pawar Comparison of Classification Algorithms using WEKA on Various Datasets IJCSIT International Journal of Computer Science and Information Technology, Vol. 4, No. 2, December 2011, pp. 85-90

[5] Daniel T. Larose, "Data Mining Methods and Models",John Wiley & Sons, INC Publication, Hoboken, New Jersey(2006).

[6] Sudhakar M, Dr. C. V. Krishna Reddy. CREDIT EVALUATION MODEL OF LOAN PROPOSALS FOR BANKS USING DATA MINING TECHNIQUES International journal of latest research in science and technology volume 3, issue 4: page no 126-131 july-august 2014

[7] 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)

[8] 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

[9] Jiawei Han, MichelineKamber, "Data Mining Concepts and Technique", 2nd edition

[10] Gaganjot Kaur,Amit Chhabra, Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Applications (0975 – 8887)Volume 98 – No.22, July 2014

[11] Prerna Kapoor, ReenaRani Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning International Journal of Engineering Research and General Science Volume 3, Issue 3, May-June, 2015

[12] Lipo Wang Support vector machines: theory and applications Springer Science & Business Media, 2005

[13] https://www.datasciencecentral.com/

[14] Sudhakar M& Dr. C.V.K Reddy, Application areas of data mining in indian retail sector Global Journal of computer science and technology C software & data engineering volume 14 issue 5 version 1.0 year 2014