International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 09, September-2015

Loan Credibility Prediction System Based on Decision Tree Algorithm

Sivasree M S
P.G Scholar
SCMS School of Technology and Management
Cochin, Kerala, India

Rekha Sunny T
Asst. Professor
SCMS School of Technology and Management
Cochin, Kerala, 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. Techniques of data mining are implemented in various domains such as retail industry, telecommunication industry, biological data analysis, intrusion detection and other scientific applications. Data mining techniques can also be used in the banking industry which help them compete in the market well equipped. 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— Decision Tree; Credit Risk Assessment; Classification; Prediction; Attribute Selection

I. INTRODUCTION

Nowadays, Banks struggle a lot to get an upper hand over each other to enhance overall business due to tight competition. Banks have realized that retaining the customers and preventing fraud must be the strategy tool for healthy competition [5]. Availability of the huge quantity of data, creation of knowledge base and efficient utilization of the same have helped banks to open up efficient delivery channels. Business decisions can be optimized through data mining [3]. Customer segmentation, banking profitability, credit scoring and approval, predicting payment from customers, marketing, detecting fraud transactions, cash management and forecasting operations, optimizing stock portfolios and ranking investments are some of the areas where data mining techniques can be used in the banking industry [1].

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(s) to evaluate credit risk.

Decision Tree Induction Algorithm is one of the best technique 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 herewith. 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, a brief review of some of the related works is presented. Research Methodology, Proposed model and the Architecture of Proposed Model are described in Sections 3, 4 and 5 respectively. The experimental results and the prototype for prediction are given in Section 5. The conclusion and future directions are summed up in Section 6.

II. LITERATURE REVIEW

A. Data Mining

Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it. It is the core of knowledge discovery process. The various steps involved in extracting knowledge from raw data as depicted in figure-1. Different 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]. 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.

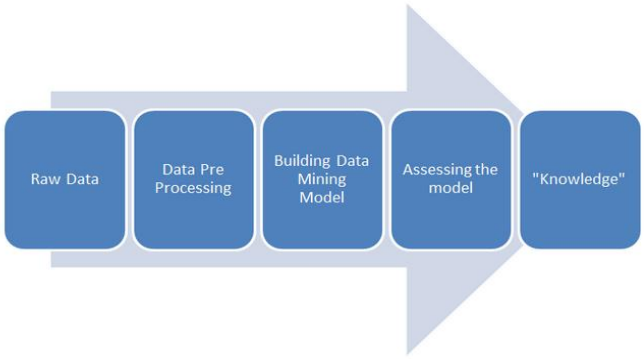

Fig. 1. Steps in Knowledge Extraction

A Decision Tree 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 figure2.

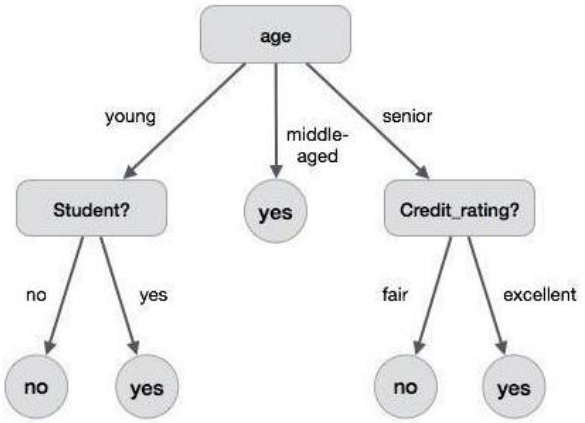

Fig. 2. Decision Tree Induction

B. 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 marketing, risk management and customer relationship management.

Marketing: It is one of the most widely used areas of data mining in the banking industry. The consumer behavior with reference to product, price and distribution channel can be analyzed by the marketing department. The reaction of the customers to the existing and new products can also be known. This information can be used by the banks to promote the products, improve quality of products and services, and gain competitive advantages. Bank analysts can also analyze the past trends, determine the present demands and forecast the customer behavior of various products and services, in order to grab more business opportunities [10].

Risk Management: It is widely used for managing risks in the banking industry. Bank executives need to know the credibility of customers they are dealing with. Offering new customers credit cards, extending existing customers' lines of credit, and approving loans can be risky decisions for banks, if they do not know anything about their customers [4]. Banks provide loans to their customers by verifying the various details relating to the loan, such as amount of loan, lending rate, repayment period etc. Even though, banks are cautious while providing loan, there are chances of loan repaying defaults by customers. Data mining technique helps to distinguish borrowers who repay loans promptly from those who default.

Customer Relationship Management: Data mining can be useful in all the three phases of a customer relationship cycle such as customer acquisition, increasing value of the customer and customer retention [11]. Customer acquisition and retention are very important concerns of any industry, especially the banking industry [4]. Banks have to cater the needs of the customers by providing the services they prefer. This will ultimately lead to customer loyalty and customer retention. Data mining techniques help to analyze the customers who are loyal from those who shift to other banks for better services. If the customer is shifting from his bank to another, reasons for such shifting and the last transaction performed before shifting can be known, and this will help the banks to perform better and retain their customers [10].

III. RESEARCH METHODOLOGY

The reference model for our work is cross-industry standard process for data mining (CRISP-DM) fig 3 [2] which is well-known to develop Data Mining projects.

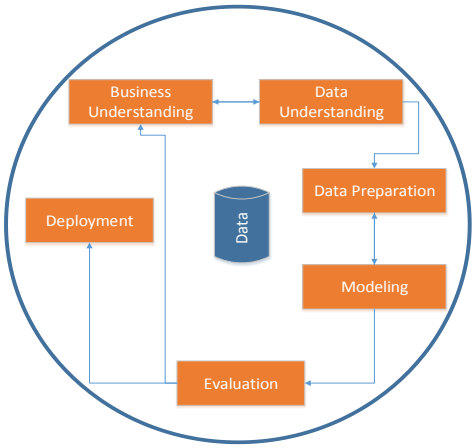

Fig. 3. CRISP-DM framework

According to this methodology, the steps of research can be described as follows:

A. *Business understanding*

Business understanding: It is the initial phase which focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

B. *Data understanding*.

The data understanding phase focuses on initial data collection, familiarization of data, identification of data quality problems, and interesting subsets to form hypotheses for hidden information etc.

C. *Data preparation*

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modelling tool from the initial raw data). There is no prescribed order for data preparation tasks. Sometimes tasks are to be repeatedly performed, like selection of table, record and attribute as well as transformation and cleaning of data for modelling tools.

D. *Modeling*

Various modelling techniques are selected and applied in this phase. Typically, there are several techniques for the same data mining problem type. Since some techniques have specific requirements on the form of data, sometimes it needs to go back to the data preparation phase.

E. *Evaluation*

This phase is to be covered before proceeding to the final deployment of the model, to be certain that business objectives are properly achieved. Consideration and successful implementation of all important business issues are to be confirmed. At the end of this phase, a decision on the use of the data mining results should be reached.

F. *Deployment*

Creation of the model is generally not the end of the project. The knowledge gained will have to be organized and presented in such a way that the customer can use it [1].

IV. PROPOSED MODEL

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.

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.

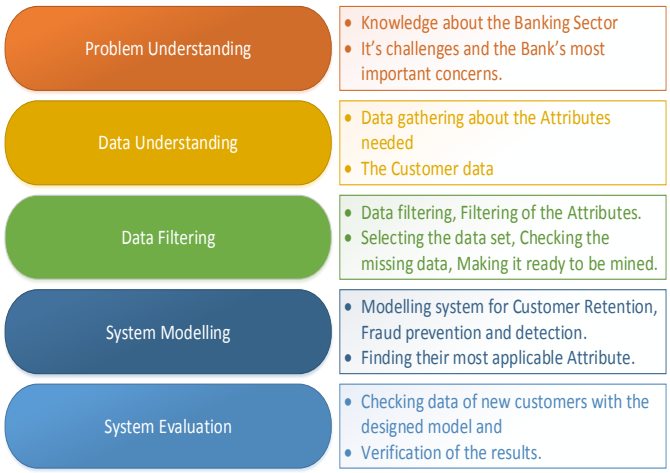

Fig. 4. Proposed Model

B. *Data Understanding*

In Data Understanding phase, the bank dataset of customer details, which is required for data mining, is 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.

V. ARCHITECTURE OF PROPOSED MODEL

Architecture of the proposed model is as shown in the figure5.

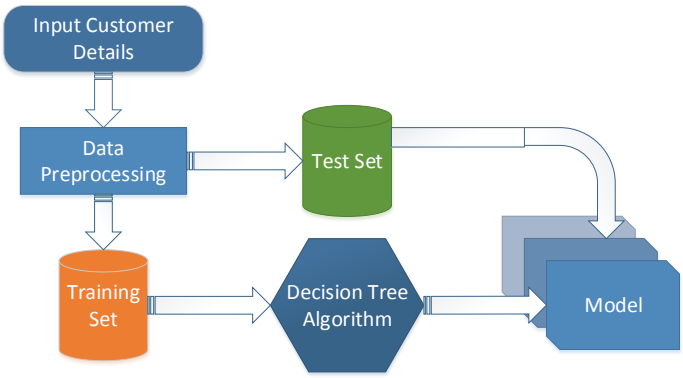

Fig. 5. Architecture of Proposed Model

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 4520 records and 17 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-6 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 possess 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.

VI. EXPERIMENTAL RESULTS

The results of the experimental analysis in predicting the loan repayment capacity are presented in this section. We have implemented our proposed model in ASP.NET-MVC5. An existing bank dataset has been used for the prediction. We have used a bank dataset of moderate size (4520) for the experimental analysis. After the pre-processing phase where dimensionality reduction was done manually and the dataset was reduced to a size of 3271. Ranks of the attributes are found out by manually adding and applying Information Gain as attribute evaluator and Ranker as search.

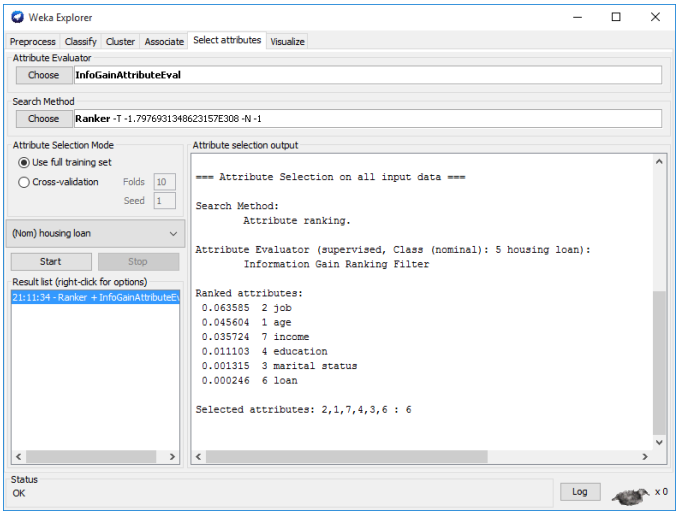

Fig. 6. Attribute Ranking

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

Rank	Attribute	Description
1	Job	Occupation of the Applicant
2	age	Age of the Applicant
3	Income	Monthly Income of the Applicant
4	Education	Education Qualification of the Applicant
5	Marital Status	Marital Status of the Applicant
6	Existing Loan	Whether the Applicant have an existing EMI or not.

Fig. 7. Relavant attributes along with rank (Table 1)

The decision tree thus generated is as given in figure 7.

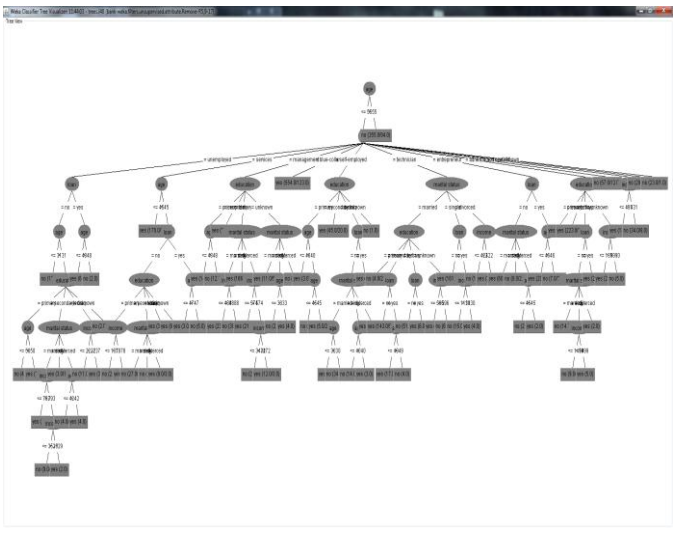

Fig. 8. Decision Tree

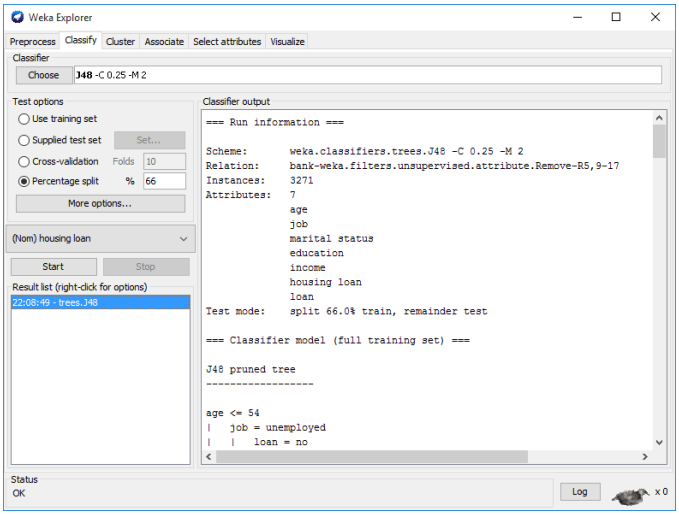

Fig. 9. Decision Tree generation

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

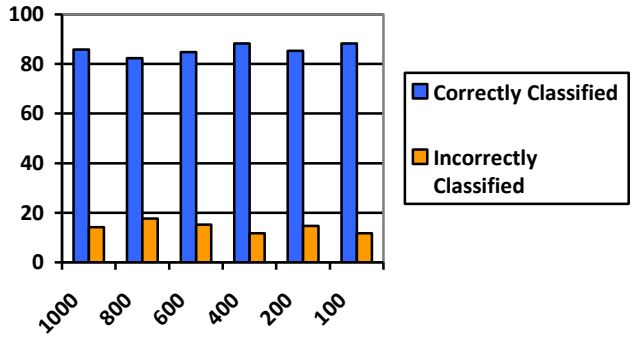

Fig. 10. Accuracy vs. number of customers

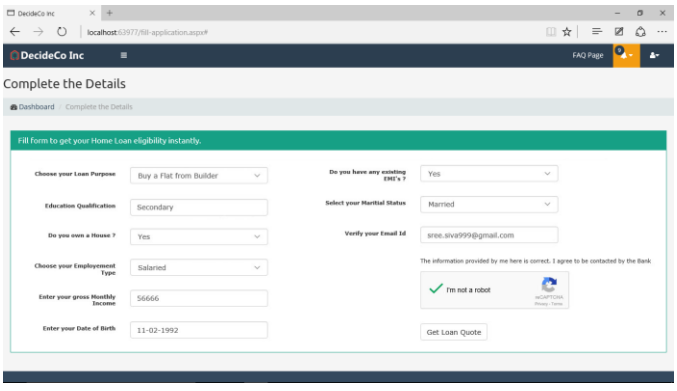

Fig. 11. Prediction (success scenario)

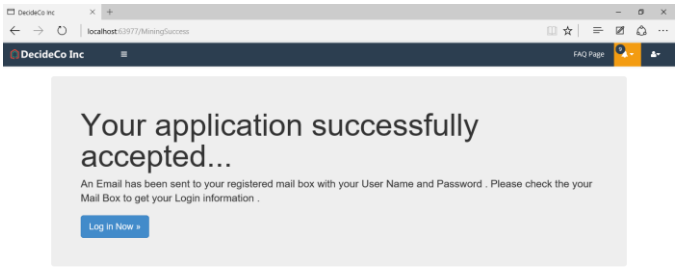

Fig. 12. Prediction (passed)

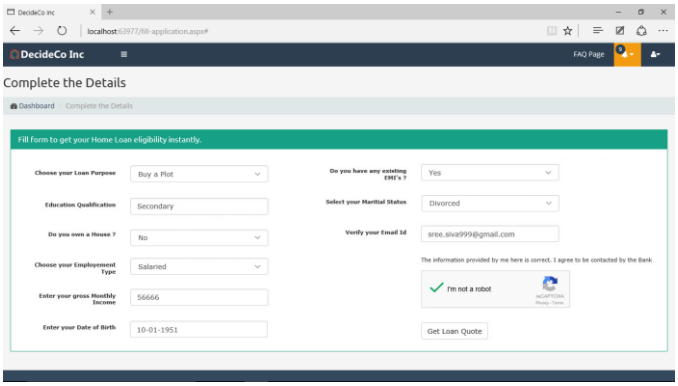

Fig. 13. Prediction (failure scenario)

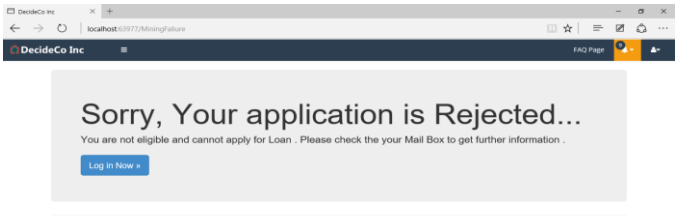

Fig. 14. Prediction result (failed)

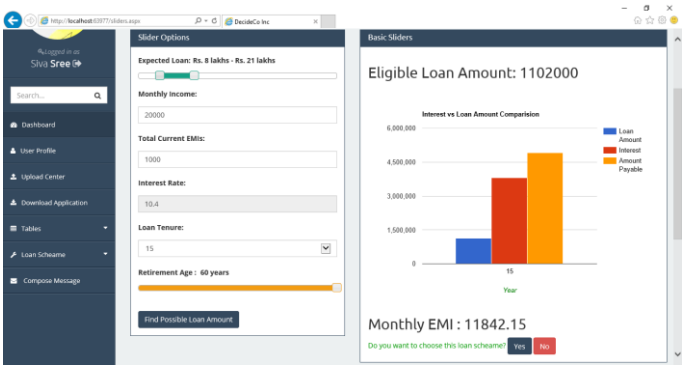

Fig. 15. Finding eligible loan amount

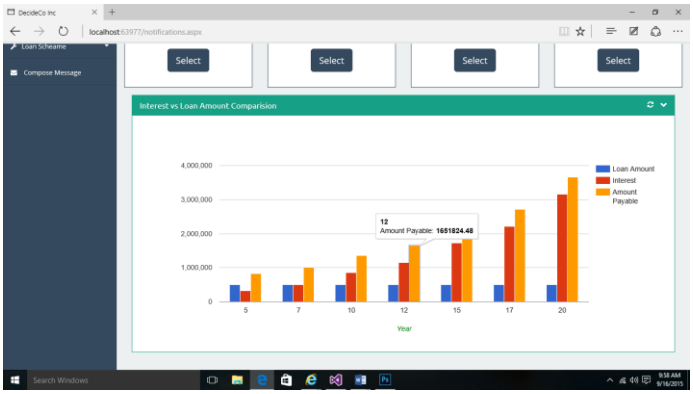

Fig. 16. Comparison among different loan scheme.

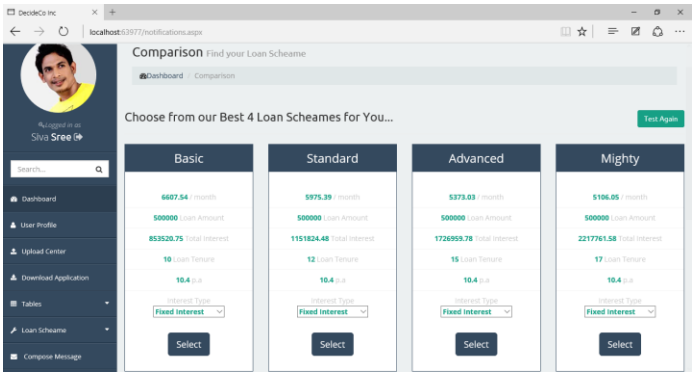

Fig. 17. Best 4 loan scheme chosen by the system

VII. Conclusion and Future Directions

In this paper, we have presented a 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 have to be implemented and tested for the domain.

Acknowledgment

The authors gratefully acknowledge the insights of all the supporters and reviewers of this paper.

References

[1] 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) , 2013, 199 – 201.

[2] Rob Gerritsen, "Loan Risks: A Data Mining Case Study".

[3] Dr. K. Chitra1, 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)

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

[5] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, 13(3):57.

[6] 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.

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

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

[9] 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.

[10] 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, Issue 2,Mar-Apr 2012, pp.738-742 738

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

[12] Hamid Eslami Nosratabadi, Sanaz Pourdarab 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

[13] E.Chapman and et.al., "CRISP-DM 1.0 Step-by- Step Data Mining Guide, SPSS