AN EFFICIENT DATA PRE PROCESSING FRAME WORK FOR LOAN CREDIBILITY PREDICTION SYSTEM

Soni P M[1], Varghese Paul[2], M.Sudheep Elayidom[3]

[1]*Assistant Professor, Dept.of MCA, SNGIST, Manjali, Kerala, India*
[2]*Associate professor, Computer Engineering, CUSAT, Kochi, Kerala, India*
[3]*Associate Professor, Information Technology, CUSAT, Kochi, Kerala, India*

Abstract

In today's world data mining have increasingly become very interesting and popular in terms of all applications especially in the banking industry. We have too much data and too much technology but don't have useful information. This is why we need data mining process. The importance of data mining is increasing and studies have been done in many domains to solve tons of problems using various data mining techniques. The art of preparing data for data mining is the most important and time consuming phase. In developing countries like India, bankers should vigilant to fraudsters because they will create more problems to the banking organization. Applying data mining techniques, it is very effective to build a successful predictive model that helps the bankers to take the proper decision. This paper covers the set of techniques under the umbrella of data preprocessing based on a case study of bank loan transaction data. The proposed model will help to distinguish borrowers who repay loans promptly from those who do not. The frame work helps the organizations to implement better CRM by applying better prediction ability.

Keywords: *Data preprocessing, Customer behavior, Input columns, Outlier columns, Target column, Dataset, CRM*

----------***----------

1. INTRODUCTION

The areas in which Data mining Tools can be used in the banking industry are customer segmentation, Banking profitability, credit scoring and approval, Predicting payment from Customers, Marketing, detecting fraud transactions, Cash management and forecasting operations, optimising stock portfolios, and ranking investments [6] . Figure 1 depicts the stages of data mining process in a business application. The first phase called Business understanding understands about the domain for which the data mining has to be performed. Here the domain we considered is Predicting payment from customers in banking domain. Now days, there will be marvellous changes in the way the banking transactions are performed. It is very important to consider the customer relationship management of the enterprise to satisfy the customers as well as the entire business in the organization. The banking industry is widely recognizing the importance of the information it has about its customers [1] . The department wants data mining to find patterns that distinguish borrowers who repay promptly from those who don't. [7]. Data mining provides the technology to analyze huge volume of data and detect hidden patterns in data to convert raw data into valuable information. Data mining, in fact, helps to identify patterns and relationships in the data [1]. Data preprocessing is an often ignored but major step in the data mining process. Banking systems collect huge amounts of data on day to day basis, be it customer information, transaction details, risk profiles, credit card details, limit and collateral details, compliance and Antimony Laundering (AML) related information, trade finance data, SWIFT and telex messages[2].

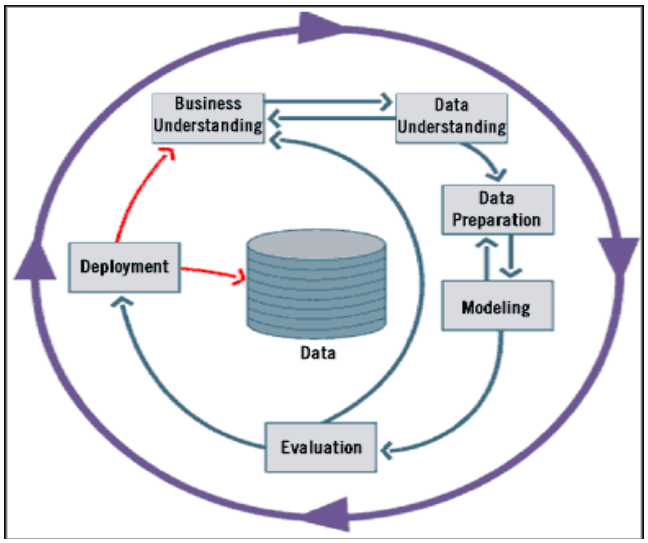

Fig-1: Data Mining Stages

The first part of the paper discuss about various concepts and data pre-processing methods that can be applied to the collected data for the purpose of data mining. The second phase of this paper deals with the problem statement which explains about the business domain understanding. Detailed system study was required for understanding about the business domain. Here the loan transactions are considered under the loan banking domain. The third phase of the stage is the data understanding which deals with the study of loan data corresponds to each customer from a premier recognized Cooperative Bank. The paper explained in detailed manner, about the data collected from the customer. The collected data may contain out of range values such as impossible data combinations, missing values, etc. This will

lead to produce misleading results. Thus, the fourth phase of this paper explains about how to preprocess the data based on the banking loan domain. Also the paper points about the technologies used for data preprocessing. The result expected after a reliable chaining of data preprocessing tasks is a final dataset, which can be considered correct and useful for further data mining algorithms. Kotsiantiset presents a well-known algorithm GALA for each step of data pre-processing [2]. Data mining can assist critical decision making processes in a bank [3]. This paper consists of describing a framework for data preprocessing and to retrieve a Boolean value that helps to decide whether or not a loan is sanctioned or not to the customer. As banking is in the service industry, the task of maintaining a strong and effective CRM is a critical issue[1].

Data mining technique helps to distinguish borrowers who repay loans promptly from those who don't. It also helps to predict when the borrower is at default, whether providing loan to a particular customer will result in bad loans etc [8]
Fig:2 depicts an effective CRM instrument in which customer insight is the main focus. Customer satisfaction is the key factor of a CRM frame work.

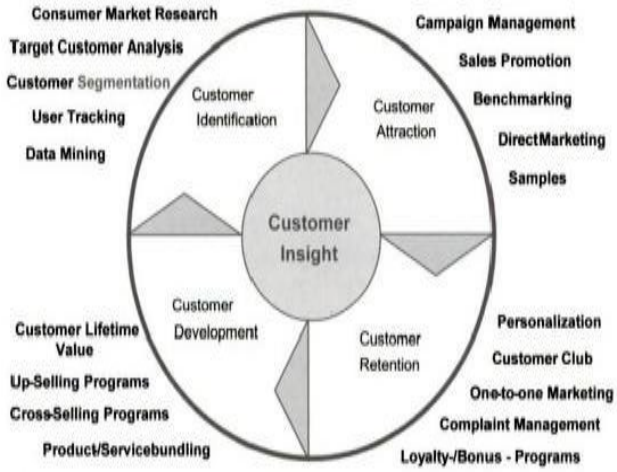

Fig-2: CRM instrument

The last stage of this paper is clearly explained about conclusions and future scope of the work.

2. PROBLEM DOMAIN

There are numerous areas in which data mining can be used in the banking industry to support customer relationships management.. Data mining technique will help to distinguish borrowers who repay loans promptly from those who do not. It also helps to predict the credit worthiness of borrower by analyzing the behavior and reliability of the customers. With data mining techniques, banks can do a thorough profiling and ranking of their branches with respect to loan fraud risk. In developing countries like India, Bankers face more problems with the fraudsters. Using data mining technique, it is simple to build a successful predictive model and visualize the report into meaningful information to the user[8]. Data mining can be applied to reduce the risk associated with lending due to fraud as well as find an appropriate solution to the borrower's need for funds, with proper assessment of risk and the inclusion of sufficient control systems to ensure repayment. Loan officers are tasked with entering borrower's credit data while the system does risk computation and sends the result to the database for approval decision by loan committee at head office. The decision is sent to loan database which is then assessed by the loan officer. The officer then informs the customer about the decision. Those banks and retailers that have realized the utility of data mining and are in the process of building a data mining environment for their decision making process will get immense benefit and advantages in future. The problem statement is "Propose a data mining methodology to analyse, design and test efficient data mining frame work for customer loan credibility prediction."

3. CONCEPTS USED

The main concepts used are data representation, different types of columns, attributes, data preprocessing methods and a frame work based on data preprocessing.

3.1 Structure of Data

The representation of data in a data mining process is normally as tabular form. The awareness of the structure of data representation is the most important step in preparing data for data mining. The action within the database is performed by considering the row as the unit of action. So the level of granularity within the database is a row that often corresponds to a separate customer transaction. Each column contains values. The range represents the set of allowable values for a column. It is also possible to find the minimum and maximum value of number data in a column. Some columns are referred as Unary valued columns and these are not used for distinguishing different rows.

3.2 Classification of Columns

The various classifications of columns are Input columns, Target column(s), Ignored columns, Identification columns, weight columns and Cost column. Input Column used as input into the model. Target column(s) are used for building predictive models. Columns that are not used are referred as ignored columns. Identification columns uniquely identify the data and these are ignored for data mining purposes. Weight column specifies a "weight" to be applied to this row. A record with a weight of three counts three times as much as a record with a weight of one. Cost column specifies a cost associated with a row. A customer's value can be considered as a cost.

3.3 Data Pre Processing

The data collected for mining process may contain missing values, noise or inconsistency. This leads to produce inconsistent information from the mining process. A data mining process with high quality of data will produce an efficient data mining results. To improve the quality of data and consequently the mining results, the collected data is to

be pre processed so as to improve the efficiency of data mining process. Data preprocessing is one of the critical step in data mining process which deals with preparation and transformation from the initial data set to the final data set. The following categories of data pre processing are applied to convert initial data set to final data set.

- Data cleaning
- Data integration
- Data transformation
- Data reduction

Data cleaning procedure is used to clean the data by filling the missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. If the user is believe that the data are dirty, and then they will not trust the results of the data mining process that has been applied to this data [4] . Data integration is process of combining data from different sources. The customer data may contain certain attribute that will take larger values[4]. The solution to this problem is normalization. Data reduction produces a reduced representation of the data set that is much smaller in volume and that should produce the same result [4].

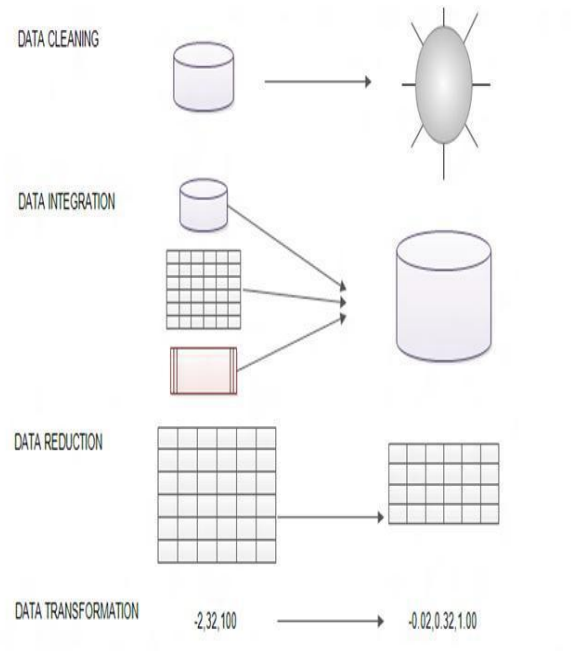

Fig - 3: Data preprocessing methods

Data preprocessing is the most time consuming phase of a data mining process. Data cleaning of loan data removed several attributes that has no significance about the behavior of a customer. Data integration, data reduction and data transformation are also to be applicable for loan data. For easy analysis, the data is reduced to some minimum amount of records. Data pre processing techniques use several statistical analyzing tools such as mean , mode, median, standard deviation , range , variance etc. Graphical tools are necessary to express and avoid outlier data. Examples of graphical tools are histograms , quantile plot , q-q plot, scalar plot etc. Binning method is suitable for smoothing the data. Clustering is a data mining technique to avoid outliers.

4. DATA USED

Data was collected from a premier cooperative bank that provides loans to individuals, business firms, etc so as to meet the requirements of all type of customers. Data collection was completed through procedures including on site observation and interview with the concerned authority. A detailed study about the loan processing and banking transactions are also made for the same. The data available consists of 2500 records of bank loan transaction data including 25 data fields. Some of the fields are removed directly by manual data preprocessing. The following are the data fields after removing the unnecessary data fields of manual data preprocessing.

Table-1: List of attributes

1	Loan Number
2	Loan Date
3	Due Date
4	Loan Amount
5	Opening
6	Payment
7	Receipt
8	Int_rcvd
9	Fine_rcvd
10	MemNo
11	Action
12	Secured
13	Loan Balance
14	Interest Rate
15	Category
16	Purpose
17	Gender
18	Occupation

5. PROPOSED ARCHITECTURE

The main focus of this paper is to propose a framework for analyzing behavior of bank customers and predict the credibility of repayment of loans. The behavior collected as inputs to the frame work and the decision such as whether a loan is to be sanctioned or not depends on the information retrieved from the classifier value of the frame work . The data collected are classified as in the table 2. The attributes are classified based on the type of attribute. It can be Input column, Target column and Ignored column. Only the input column attributes are given to the model as input. From literature review initially need to create the account for each customers in the bank and they should enter their personal details, income details, insurance details, loan details and the

account information of the corresponding customer in other banks[4] . Here when a loan is sanctioned, the customer details as well as the loan details are entered into the database. Loan number is recognized as the identification column that uniquely identifies the data. Generally this column is ignored for data mining process. This is because the unary valued columns do not contain any information that helps to distinguish between different rows. Category column also is removed in such a manner. The major input columns from the collection are Loan amount, Fine received, Action, Loan balance, Occupation, and Purpose. These values are considered as the input to the model. The target column is to be designed and is used when building predictive models. Here the target column specifies whether a loan is to be sanctioned or not depending on the input values to the model. Some columns are considered as ignored columns or outlier columns. These are not used in designing the model. We don't want the actual loan balance to process the model.

Table- 2: Classification of attributes

Data fields or attributes	Type of attribute	Remarks
Loan Number	Identification column	Not used for data mining process.
Loan amount, Fine received action, Loan balance, Interest Rate, Occupation, and purpose.	Input columns	These are the inputs to the model.
Loan Status	Target Column	Decides whether loan is to be sanctioned or not.
Loan Date, Due date,Opening,Payment , Receipt, Int-rcvd , MemNo,	Outlier Columns	These are not necessary to build the model.

The model requires either an amount is in the loan balance column or not to classify the persons who close the loan or not. In order to do those replace the value as 1 and 0. If there is a value in the column loan balance replace with 1, otherwise with 0. The missing values in the columns purpose and job are not considered because our data collection is very huge and these are considered as do nothing. These few missing values may not materially affect the models. The frame work for loan transaction processing system is described in fig 4. The framework consists of customer data as input and target column as output. The customer data is processed through Loan data preprocessing. It helps to distinguish the entire input data to input columns, identification columns and outlier columns. Outlier columns have no importance over the attributes of data mining process. Only the input columns are given to the model for loan data processing. The output obtained from the model is the Target column. This is actually a classifier data that decides whether or not the loan is to be sanctioned or not based on the customer behavior. In order to implement this, some more data mining techniques such as decision tree, SVM and neural network etc are to be designed and tested. Also we need to design an algorithm in data mining environment to check the truthfulness of this framework concept

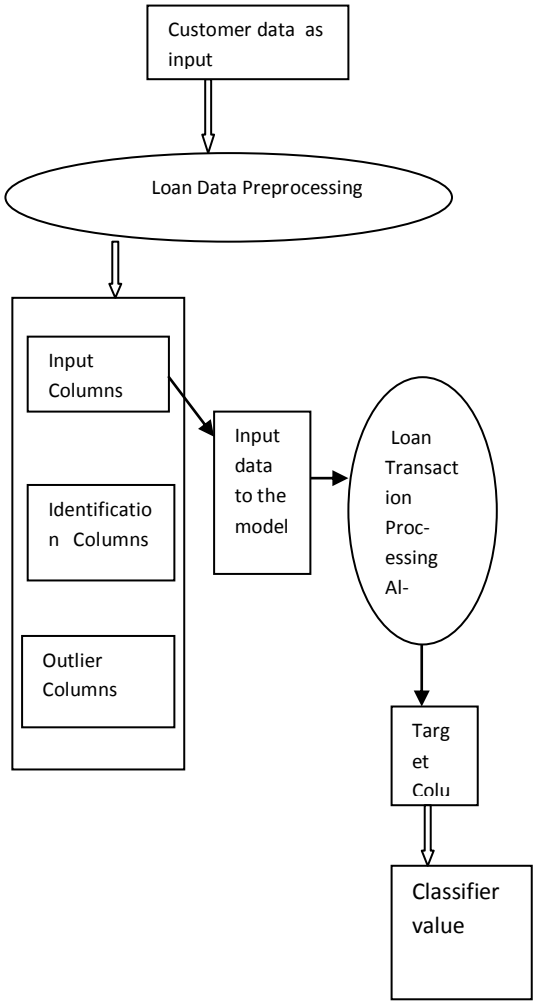

Fig 4: Framework for loan credibility prediction

5.1 Advantages

The advantage of using this frame work for banking organization is to enter the transactions of valid customers in to their data warehouse. The fraud customers are filtered by applying this data mining model at the time of first interaction with the concerned authorities. It allows solving many problems related with fraudsters in the banking sector. This will be an asset for developing nations like India to stable their financial records.

6. TECHNOLOGY USED

The Weka suite contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [5]. Weka is a powerful tool for data preprocessing, classification, clustering and Visualization. It

is freely available as well as platform-independent software. As is fig 5 Weka graphical user interface chooser consists of four applications such as Explorer, Experimenter, Knowledge Flow and simple CLI. Notepad helps to create the data set in .ARFF format and the file in .ARFF format is necessary to open in Weka. Weka can apply several algorithms and techniques in data mining and it is possible to compare the result of a process in different techniques. Microsoft Excel is a powerful tool to manage data in tabular form that is the most important format of data used by data mining algorithm. The initial data and final data can be easily represented in tabular form of Excel.

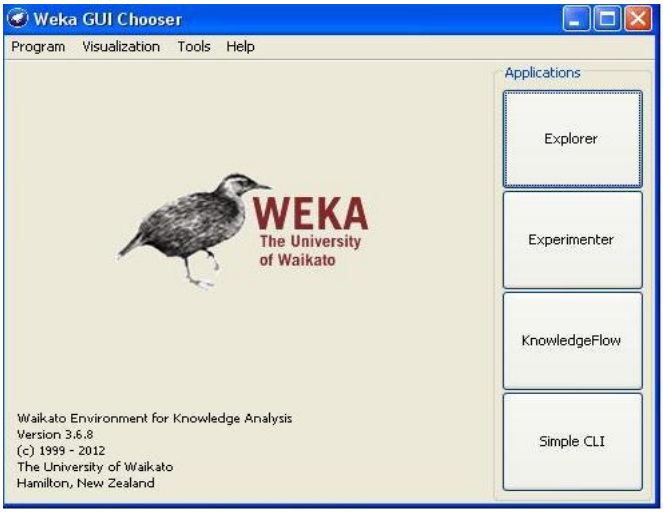

Fig 5: GUI of Weka

7. FUTURE SCOPE

The main focus of my work is the design of data mining models to predict the customers who repay loans promptly from those who do not. Models for SVM, Neural Network Analysis etc are to be designed and tested. The prediction is done by analyzing behavior and reliability of the customers using the prediction algorithm that take the input as the characteristics of the customer such as nature of job, past history, income and so many related fields. The output expected from the system is an indicator whether the customer is prospective or not. Also algorithms that outperform the performance of popular data mining models have to be developed and tested for the CRM domain.

8. CONCLUSION

In this work, a data preprocessing frame work for loan credibility prediction is proposed. The data was collected from a premier organization and the data were classified by applying some preprocessing techniques such as data cleaning, data integration and normalization etc. The behavior of customers after preprocessing is considered as the input to the model. The target column is a classifier that can decide whether or not the loan to be sanctioned. The framework is helpful to predict the loan repayment credibility of a customer by considering the behavior of customer .This will help the banking organization to avoid fraudsters entering into their transactions.

REFERENCES

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

[2]. 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.

[3]. Sreekumar Pulakkazhy and 2R.V.S. Balan "Data Mining in banking and its applications – a review " Journal of Computer Science 9 (10): 1252-1259, 2013 ,ISSN: 1549-3636 © 2013 Science Publications, doi:10.3844/jcssp.2013.1252.1259

[4]. Ms. Neethu Baby1, Mrs. Priyanka L.T "Customer Classification And Prediction Based On Data Mining Technique", International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 2, Issue 12, December 2012)

[5]. Swasti Singhal, Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering" , International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013

[6]. 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.

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

[8]. 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)