Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 12, December -2017

REVIEW ON PREDICTION SYSTEM FOR BANK LOAN CREDIBILITY

Kalyani R. Rawate[1], Prof. P. A. Tijare[2]

[1]Department Of Computer Science and Engineering, Sipna COET Amravati
[2]Department Of Computer Science and Engineering, Sipna COET Amravati

Abstract- *In today's world there are many risks involved in bank loans, so as to reduce their capital loss; banks should perform the risk and assessment analysis of the individual before sanctioning loan. In the absence of this process there are many chances that this loan may turn in to bad loan in near future. Banks hold huge volumes of customer behavior related data from which they are unable to arrive at a decision point i.e. if an applicant can be defaulter or not. This can be achieved using the data mining techniques. Data analysis can be done using the data mining techniques. Here customers data sets compared with the trained data sets and depend on that comparison final prediction can be done. Data Mining is a promising area of data analysis which aims to extract useful knowledge from tremendous amount of complex data sets. We are going to implement a model for the bankers that help them predict the credible customers who have applied for loan. Random Forest Data Mining Algorithm is applied to predict the attributes relevant for credibility. This model can be used by the organizations in making the right decision to approve or reject the loan request of the customers.*

Keywords: *Random forest; Credit risk assessment; Prediction; Attribute selection; R*

I. INTRODUCTION

Bank plays a vital role in market economy. The success or failure of organization largely depends on the industry's ability to evaluate credit risk. Before giving the credit loan to borrowers, bank decides whether the borrower is bad (defaulter) or good (non defaulter).The prediction of borrower status i.e. in future borrower will be defaulter or non defaulter is a challenging task for any organization or bank. Basically the loan defaulter prediction is a binary classification problem Loan amount; costumer's history governs his credit ability for receiving loan. The problem is to classify borrower as defaulter or non defaulter. However developing such a model is a very challenging task due to increasing in demands for loans. Prototypes of the model which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers. This work includes the construction of an ensemble model by combining three different machine learning models.

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 [6]. Credit Risk assessment is a crucial issue faced by Banks nowadays which helps them to evaluate if a loan applicant can be a defaulter at a later stage so that they can go ahead and grant the loan or not. This helps the banks to minimize the possible losses and can increase the volume of credits. The result of this credit risk assessment will be the prediction of Probability of Default (PD) of an applicant. Hence, it becomes important to build a model that will consider the various aspects of the applicant and produces an assessment of the Probability of Default of the applicant. R Package is an excellent statistical and data mining tool that can handle any volume of structured as well as unstructured data and provide the results in a fast manner and presents

the results in both text and graphical manners. This enables the decision maker to make better predictions and analysis of the findings. The objective of this research is to develop a data mining model using R for predicting PD for new loan applicants of a Bank. The data used for analysis contains many inconsistencies like missing values, outliers and inconsistencies and they have to be handled before being used to build the model. To classify if the applicant is a defaulter or not, the best data mining approach is the classification modeling using Decision Tree. The above said steps are integrated into a single model and prediction is done based on this model. 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.

II. LITERATURE REVIEW AND RELATED WORK

In [1] the author introduces a framework to effectively identify the Probability of Default of a Bank Loan applicant. The metrics derived from the predictions reveal the high accuracy and precision of the built model. The model proposed in [2] an effective prediction model for predicting the credible customers who have applied for bank loan. Decision Tree is applied to predict the attributes relevant for credibility. This prototype model can be used to sanction the loan request of the customers or not. The model proposed in [3] has been built using data from banking sector to predict the status of loans. This model uses three classification algorithms namely j48, bayes Net and naïve Bayes. The model is implemented and verified using Weka. The best algorithm j48 was selected based on accuracy. An improved Risk prediction clustering Algorithm that is Multi-dimensional is implemented in [4] to determine bad loan applicants. In this work, the Primary and Secondary Levels of Risk assessments are used and to avoid redundancy, Association Rule is integrated. In [5] a decision tree model was used as a classifier and for feature selection genetic algorithm is used. The model was tested using Weka. The work in [6] developed two data mining models for credit scoring that helps in decision making of giving loans for the banks in Jordan. Considering the rate of accuracy, the regression model is found to perform better than radial function model.

The work in [7] develops many credit scoring models that are based on the multilayer approach. The work proves its performance than the other models that uses logistic regression techniques. The results show that the neural network model performs better than the other three techniques. The work in [8] compares support vector machine based models for credit-scoring developed using the various default definitions. The work concluded that the broad definition models are better than the narrow definition models in their performance. Financial data analysis is done in [9] susing the techniques such as Decision Tree, Random forest, Boosting, Bayes classification, Bagging algorithm and others. Support Vector Machine, Decision Tree, Logistic Regression, Neural Network, Perceptron model, all these techniques are combined in this model. The effectiveness of applying the above techniques on credit scoring is studied. The analysis results show the performance is outstanding based on accuracy. The aim of the study in [10] is to introduce a discrete survival model to study the risk of default and to provide the experimental evidence using the Italian banking system.

2.1. 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. 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[3]. 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. 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.

2.2. 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.

III. PROPOSED WORK

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. On the output from the classifier, decision on whether to approve or reject the customer request can be made. Using different data analytics tools loan prediction and there severity can be forecasted. In this process it is required to train the data using different algorithms and then compare user data with trained data to predict the nature of loan. Several R functions and packages were used to prepare the data and to build the classification model. The work proves that the R package is an efficient visualizing tool that applies data mining techniques. Using R Package, customer's data analysis can be done and depends on that bank can sanction or reject the loan. In real time customers data sets may have many missing and imputed data which needs to be replaced with valid data generated by making use of the available completed data. The dataset has many attributes that define the credibility of the customers seeking for several types of loan. The values for these attributes can have outliers that do not fit into the regular range of data. Hence, it is required to remove the outliers before the dataset is used for further modeling. This can be achieved using the different R Package libraries.

For ranking the features, ***randomForest()*** function of the Random Forest package is used. The steps involved in model building methodology are represented as below.

Step 1 – Data Selection

Step 2 – Data Pre-Processing

Step 2.1 – Outlier Detection

Step 2.2 – Outlier Ranking

Step 2.3 – Outlier Removal

Step 2.4 – Imputations Removal

Step 2.5 – Splitting Training & Test Datasets
Step 2.6 – Balancing Training Dataset
Step 3 – Features Selection
Step 3.1 – Correlation Analysis of Features
Step 3.2 – Ranking Features
Step 3.3 – Feature Selection
Step 4 – Building Classification Model
Step 5 – Predicting Class Labels of Test Dataset
Step 6 – Evaluating Predictions

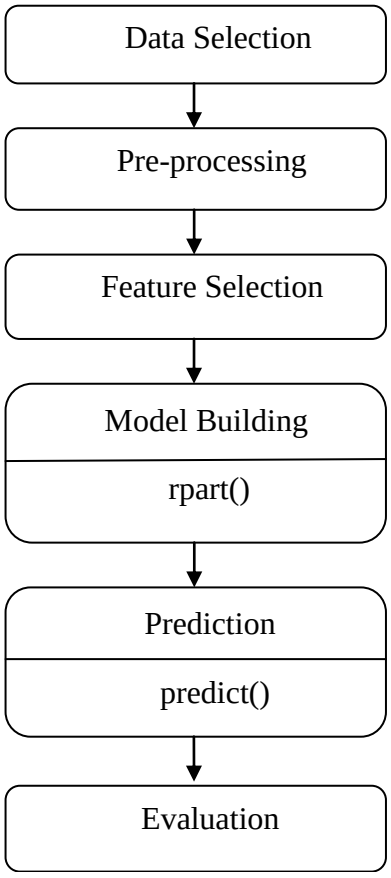

Fig.1. Major steps of the credit risk analysis and prediction modeling using R

3.1. Dataset selection

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. The dataset after selecting and understanding is loaded into R software.

3.2. Data preprocessing

The dataset has many missing and imputed data which is replaced in this step. 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. 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. 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.

3.3. Feature selection and building classification model

It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known. Using the Random Forest algorithm feature selection can be achieved and the targeted learner model can be build.

3.4. Prediction

It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data. The model is tested using the test dataset by using the predict() function.

3.5. Evaluation

In the final stage, the designed system is tested with test set and the performance is assured. Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time. Common metrics calculated from the confusion matrix are Precision; Accuracy .The calculations for the same are listed below.

$$\text{Precision} = \frac{\text{True Defaults}}{\text{True Defaults + False Defaults}}$$

$$\text{Accuracy} = \frac{\text{True Defaults + True Nondefaults}}{\text{Total Testset}}$$

IV. RANDOM FOREST ALGORITHM

Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

4.1 How does it work?

Assume that the user knows about the construction of single classification trees. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

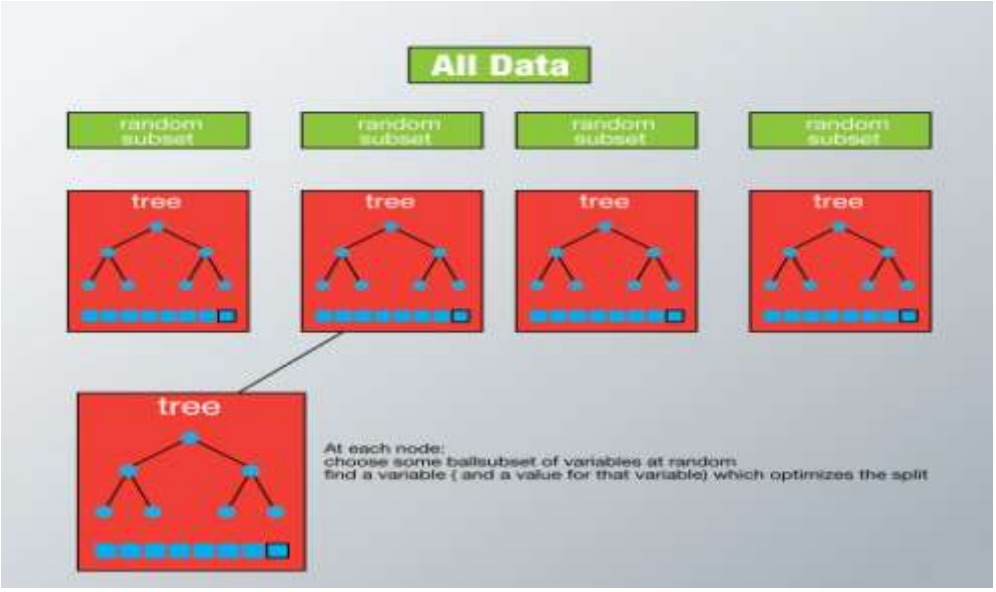

Fig.2. Random forest

It works in the following manner. Each tree is planted & grown as follows:

1. Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.
3. Each tree is grown to the largest extent possible and there is no pruning.
4. Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

4.2 Advantages of random forest

This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts. One of benefits of Random forest which excites most is the power of handle large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

- It has methods for balancing errors in data sets where classes are imbalanced.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
- Random Forest involves sampling of the input data with replacement called as bootstrap sampling. Here one third of the data is not used for training and can be used to testing. These are called the out of bag samples. Error estimated on these out of bag samples is known as out of bag error. Study of error

estimates by Out of bag, gives evidence to show that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. Therefore, using the out-of-bag error estimate removes the need for a set aside test set.

V. APPLICATION FLOW

Loan prediction by bank: Banks can verify the customer's eligibility for loan using this application.
Online loan eligibility check for individual: User can check loan eligibility online by providing required information and in response he will received email whether he is eligible for loan or not. Here it is not required to visit person in bank so it will save more time of the individual and helps in better user experience.

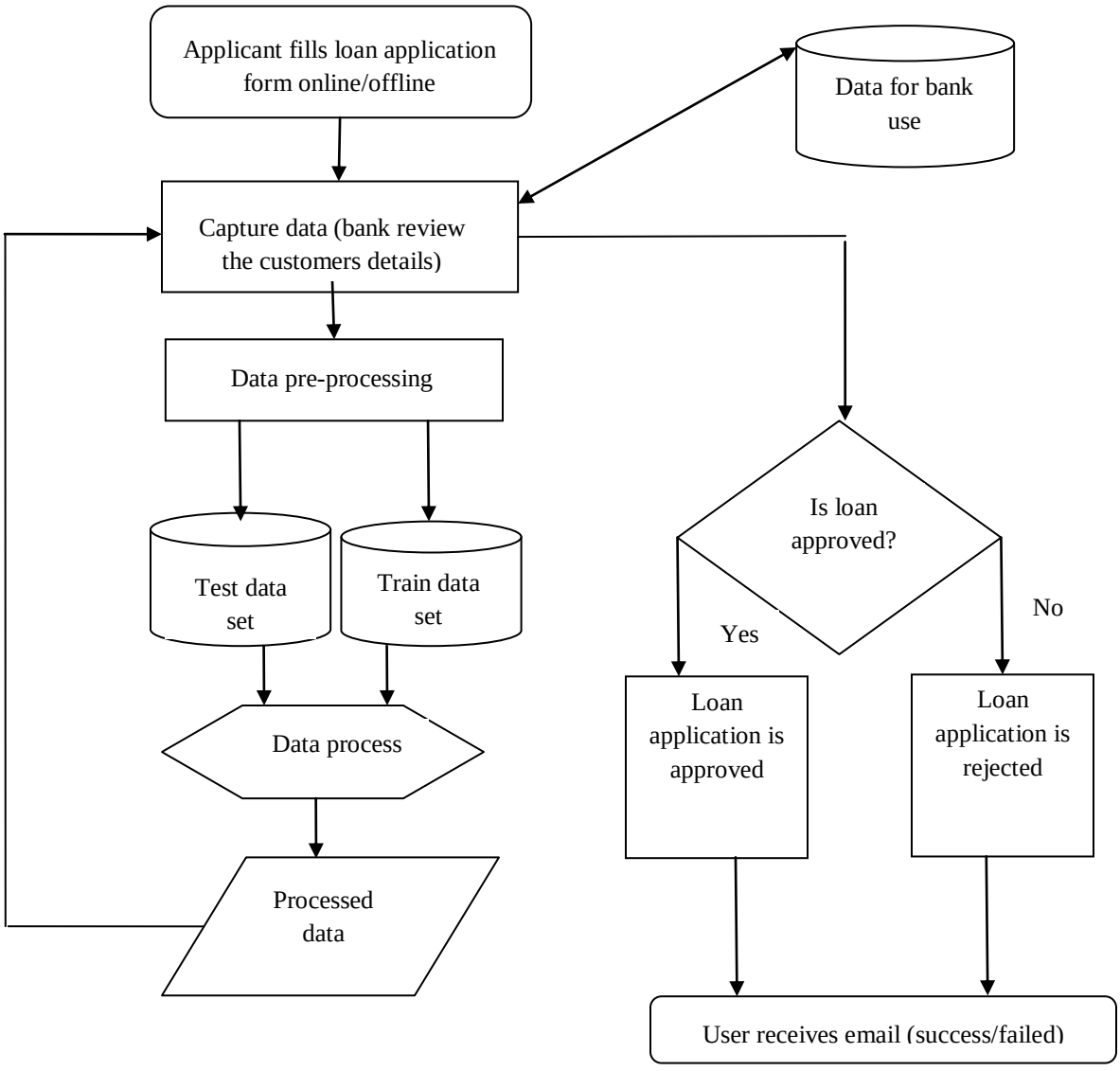

Fig.3. Bank loan prediction flow

VI. CONCLUSION

This application can help banks in predicting the future of loan and its status and depends on that they can take action in initial days of loan. Using this application banks can reduce the number of bad loans and from incurring sever losses. Several R functions and packages were used to prepare the data and to build the classification model. R Package libraries help in successful data analysis and feature selection. Using this methodology bank can easily identify the required information from huge amount of data sets and helps in successful loan prediction to reduce the number of bad loan problems. Data Mining techniques are very useful to the banking sector for better targeting and acquiring new customers, most valuable customer retention, automatic credit approval which is used for fraud prevention, fraud detection in real time, providing segment based products, analysis of the customers, transaction patterns over time for better retention and relationship, risk management and marketing.

REFERENCES

[1]. Sudhamathy G and Jothi Venkateswaran " Analytics Using R for Predicting Credit Defaulters",IEEE international conference on advances in computer applications (ICACA), 978-1-5090-3770-4, 2016.

[2]. M. Sudhakar, and C.V.K. Reddy, "Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5, no.3, pp. 705-718, 2016.

[3]. J.H. Aboobyda, and M.A. Tarig, "Developing Prediction Model Of Loan Risk In Banks Using Data Mining", Machine Learning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016.

[4]. K. Kavitha, "Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp. 162–166, 2016.

[5]. Z. Somayyeh, and M. Abdolkarim,"Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran", Jurnal UMP Social Sciences and Technology Management, vol. 3, no. 2, pp. 307–316, 2015.

[6]. A.B. Hussain, and F.K.E. Shorouq, "Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach", Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014.

[7]. A. Blanco, R. Mejias, J. Lara, and S. Rayo, "Credit scoring models for the microfinance industry using neural networks: evidence from Peru", Expert Systems with Applications, vol. 40, pp. 356–364, 2013.

[8]. T. Harris, "Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions", Expert Systems with Applications, vol. 40, pp. 4404–4413, 2013.

[9]. 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.

[10] G. Francesca, "A Discrete-Time Hazard Model for Loans: Some Evidence from Italian Banking System", American Journal of Applied Sciences,vol. 9, no.9), pp. 1337–1346, 2012.

