PREDICTION SYSTEM FOR BANK LOAN CREDIBILITY USING R

[1]Kalyani R. Rawate, [2]Dr. D. M. Dakhane [3]Prof. P. A. Tijare
[1]ME Student, [2]Professor, [3]Associate Professor
Computer Science & Engineering Department
[1]Sipna College of Engineering and Technology, Amravati, India

***Abstract*— Bank monitor several loan related attributes for tracking the condition and quality of their financial portfolios. If the trend of loan related status is understood well, the bank would be able to proactively take actions to avoid the prolonged delinquency and loan defaults. If an early warning system is available to predict the risk with a loan well-ahead of time, the bank can potentially take corrective measures to prevent the loan from defaulting. Estimating the probability that an individual would default on their loan, is useful for banks to decide whether to sanction a loan to the individual or not. But this process is more difficult to bank as they holds the huge volume of customer data from which they are unable to arrive at a judgment if an applicant can be defaulter or not. This problem can be overcome by using the Data Mining technique; it is a promising area of data analysis which aims to extract useful knowledge from tremendous amount of complex data sets. We introduce an effective prediction technique that helps the banker to predict the credit risk for customers who have applied for loan. The different models uses in this application i.e. Decision Tree and Random Forest model and analyses the credit risk for optimum result. This application can be used by the organizations in making the right decision to approve or reject the loan request of the customers.**

***Keywords*— Random forest; Decision tree; Prediction; Attribute selection; R**

I. INTRODUCTION

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. Bank monitor several loan related attributes for tracking the condition and quality of their financial portfolios. In every country there banks are facing a bad loan problem for example. In our country all the leading banks including government and private sector are facing huge bad loan problems and day by day its severity going to increase. It's very difficult for bank to predict the future of loan which they are going to sanction in initial days. Almost all banks are facing NPA (Non Performing assets) problem and after every day this trend is on increasing. If banks are not able to fix this trend early then on one day this will result in the total collapse of the financial system of the country. Also this process is extremely difficult for banks to predict the future of sanctioned loan (i.e. whether it may be the part of NPA or not) as they holds the huge volume of customer behaviour related data.

In this case our application will helps banks whether to approve a loan or reject it in initial loan approval days using the data mining techniques. Using data mining techniques we can identify the bad loan patterns and depends on that bank can decide whether to sanction the loan or reject it. Two different models used in this application i.e. Decision Tree and Random Forest model and analyses the credit risk for optimum result. Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. When the sample size is large enough, study data can be divided into training and validation datasets. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model. 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.

Several R functions and packages are used to build the classification model and predict the customer loan request depends on that banks can arrive on a particular conclusion. To save the time of both banks and customers, it is required to develop a system which will collect the data from customer and will provide the result after processing with different data mining techniques. After knowing the loan request result, customer can visit a bank and bank can process the request based on loan number which was sent to customer. As we are using the more sophisticated data mining approach, our model gives us the most accurate result. This is a whole automate process it saves the both Bank and individual/institution time and also help to bank in identifying the future of loan request (i.e. Bad loan or good one) and depends on that bank can take the action. Our proposed model helps bank to know the future of loan request more accurately in advance and also helps them to reduce the bad loan problem. To save the time of both banks and customers, it is required to develop a system which will collect the data from customer and will provide the result after processing with different data mining techniques. After knowing the loan request result, customer can visit a bank and bank can process the request based on loan number which was sent to customer.

II. LITERARTURE SURVEY

Data mining is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology. There are different types of data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc.[1] The work in[2]the model proposed 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 work in [3] 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.

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 [4]. Customer acquisition and retention are very important concerns of any industry, especially the banking industry 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 analyse 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.

The proposed model in [5] proposed ensemble classifier is constructed by incorporating several data mining techniques, that involves optimal associate binning, discretize continuous values, neural network, support vector machine, and Bayesian network are used. The data driven nature of the proposed system distinguishes it from existing credit scoring systems. The aim of the study in [6] is to introduce a discrete survival model to study the risk of default and to provide the experimental evidence using the Italian banking system. The work in [7] proposed to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The work in [8] introduces 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 to predict the attributes relevant for credibility. The work in[9] proposed various ensemble algorithms like bagging, boosting and stacking are implemented and their efficiency and accuracy is compared. The robust predictive models are able to predict the default with high degree of accuracy. Its attempts to build robust data mining models to predict the defaulters using data obtained from one of finance company

III. IMPLEMENTATION

Using J2EE features we have developed an application which captures all the required information from the customer, then all the captured information sent to the classification model which is developed using different R functions and packages. Classification model analyses the data and provides the result, based on this J2EE application returns the loan request status to customer and also persists the same in bank repository for future use. As we are using the more sophisticated data mining approach, our model gives us the most accurate result. This is a whole automate process it saves the both Bank and individual/institution time and also help to bank in identifying the future of loan request (i.e. Bad loan or good one) and depends on that bank can take the action. Our proposed model helps bank to know the future of loan request more accurately in advance and also helps them to reduce the bad loan problem. The steps involved in this model building methodology are represented by 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

Figure no. 1 shows loan prediction flow of bank in which user login into the system by using, username and password. User fills the loan application form and submits these submitted data goes to the bank by using data selection method. The dataset has many missing and imputed data which is replaced by using pre-processing method. Using pre-processing data are in correct form that is used for prediction. On the basis of prediction result the bank passed or failed the loan request.

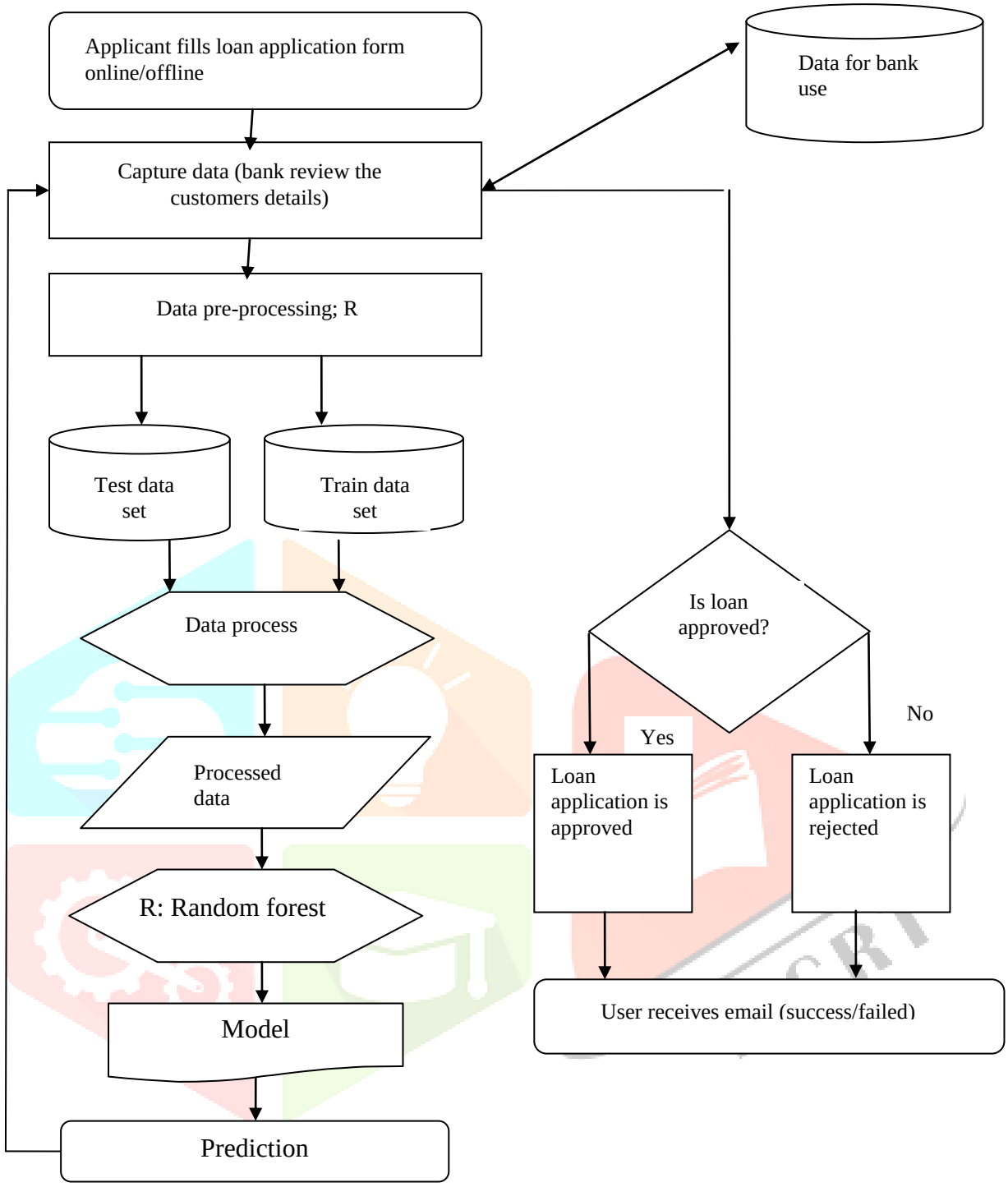

Fig. 1 Bank loan prediction flow

IV. EXPERIMENTAL RESULT

The result of the experimental analysis in predicting the loan repayment capacity presented in this section. We have implanted our proposed model in J2EE application. An existing bank dataset has been used for the prediction.

A. **Dataset Selection**

TABLE 1 Dataset Attribute Types

Gender	Married	Dependents	Education	Self- Employed	Applicant Income

Co-Applicant Income	Loan amount	Loan amount in term	Credit history	Property area	Loan status

B. Pre-Processing

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.

1. Outlier Detection: Outliers in data can distort predictions and affect the accuracy, if you don't detect and handle them appropriately especially in regression models.

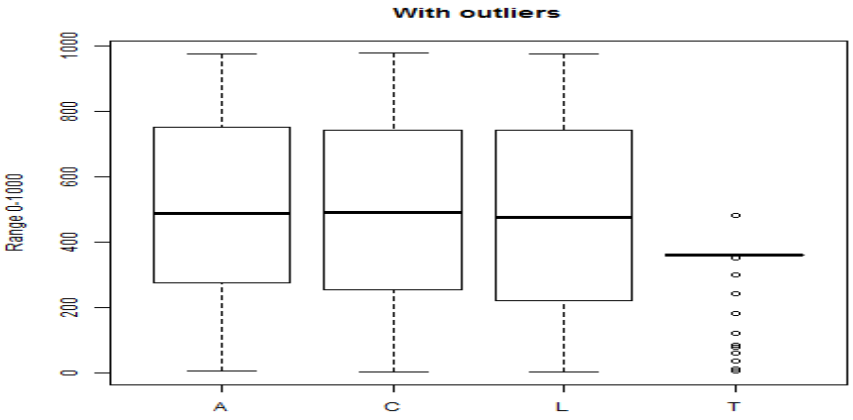

Fig. 2 Outlier detection for all numeric features

2. Outlier Ranking: The agglomerative hierarchical clustering algorithm chosen for ranking the outlier is less complex and easy to understand. The ranking is obtained on the basis of the path each case follows within the merging steps of an agglomerative hierarchical clustering method.

3. Outlier Removal: The observations which are out of range (based on ranking) are removed. Remove the outlier i.e. foreign data for better model building.

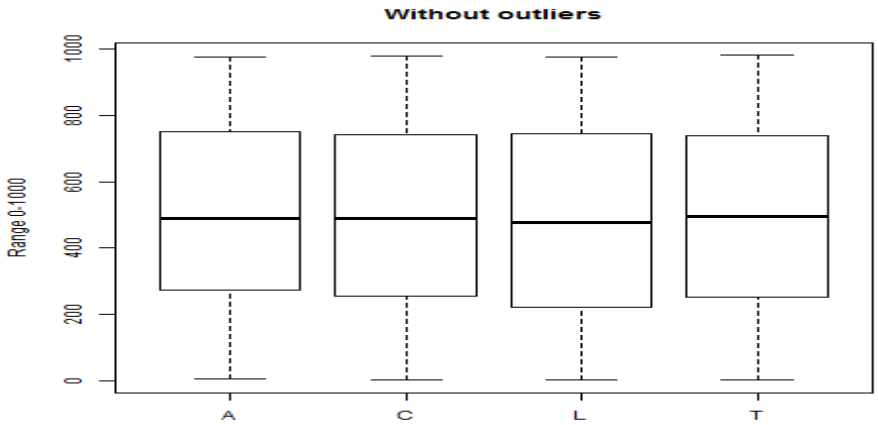

Fig. 3 Outlier removal for all numeric features

4. Imputations Removal: Imputation is the process of replacing missing data with substituted values. The method used for null values removal is multiple imputation method. There were no null values for the attributes in the dataset we have chosen and hence the number of records remains unchanged after this step.

5. Splitting training and Test Datasets: Before proceeding to further steps, the dataset has to be split into training and test dataset so that the model can be built using the training dataset.

6. Balancing Training Dataset: A balanced data set is a set that contains all elements observed in all time frames. The Generalized Linear model handles unbalanced classification problems and it generates the new dataset that addresses the unbalanced class problem.

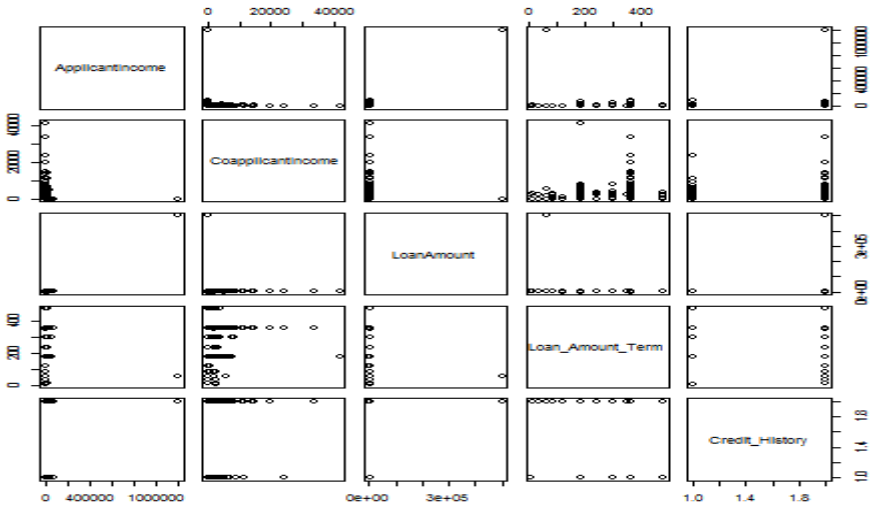

Fig. 4 Data distribution before balancing

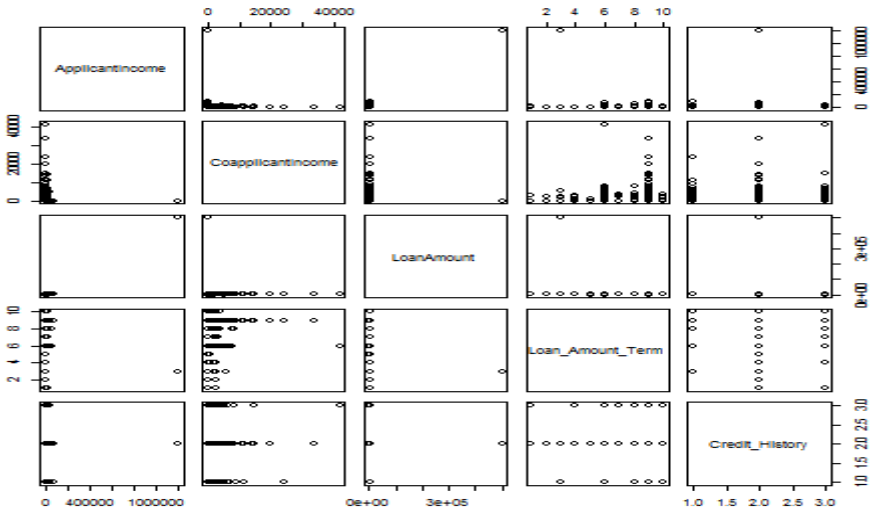

Fig. 5 Data distribution after balancing

C. Features Selection

1. Correlation analysis of features: Dataset may contain irrelevant or redundant feature which might make the model more complicated. Hence removing such redundant features will speed up the model. Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight). This particular type of analysis is useful when a researcher wants to establish if there are possible connections between variables.

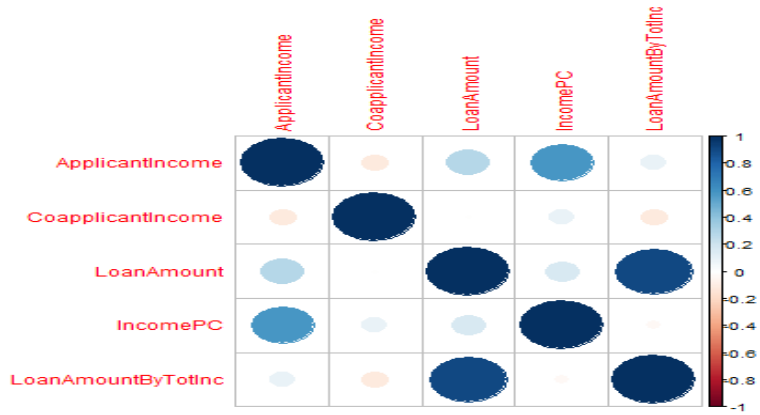

Fig. 6 Correlation analysis of features

2. Ranking Features: The aim of this step is to find the subset of feature that will be really relevant for the analysis as irrelevant feature causes drawbacks like increased runtime, complex patterns etc. this resultant subset of features should give the same results as that of the original dataset. In table 2 we are finding the important features which have huge impact on loan approval.

TABLE 2 Importances of Features

	N	Y	MeanDecreaseAccuracy	MeanDecreaseGini
Gender	-1.33764632947594	1.4210020890507	1.00240096384017	0.061562737901801
Married	0.18746836801288	1.90926101308462	2.31137279161465	0.33695883106517
Dependents	-0.73867411060769	0.141642717062641	-0.2226271874418	0.555641092628831
Education	-0.768132995032143	-0.0505691088800489	-0.667893678176774	0.328809235924787
Self_Employed	-1.34487456235782	-0.0555811130229936	-0.915220514495456	0.17330934988943
ApplicantIncome	-5.23154892182091	5.46984211042205	4.94229907178482	3.23266267993515
CoapplicantIncome	0.0736430881905734	2.81563508799967	2.90134950739121	2.16911028572888
LoanAmount	-2.93355584649627	3.93121794677794	3.63482862072485	2.31902987926741
Loan_Amount_Term	0.516780011686052	3.03778116550568	2.51408026026735	4.00117217386115
Credit_History	56.7714146178043	58.7399576363075	60.2274670592035	70.4026208179577
Property_Area	0.369676381935068	0.543622140555037	0.658393486640546	2.84972174578796
TotalIncome	1.12680955255331	7.23546047676469	7.8820910816435	7.47135403984594
FamilySize	-2.90899457552823	2.51984182704391	2.16414194930226	0.33209311904596
IncomePC	-3.30251152119075	4.57451302274271	4.45794147270076	2.44463337137723
LoanAmountPC	-3.06134135120483	4.74540409818781	4.76597023086524	3.63690897129238
LogApplicantIncome	-3.0357159000745	4.64393080135164	4.80371938072773	2.39552258649861
LogCoapplicantIncome	-1.2571757347456	3.37196541514641	3.52779545164769	1.98505467968776
LogLoanAmount	-2.66308353044698	3.90801099765315	3.57539058379471	2.17435837265116
LogLoanAmountPC	-3.74795443107923	5.03309827016859	4.57329629996999	3.12054798321396
LogLoanPerMOnth	-2.61614531570583	4.04709192440055	3.86735922527691	2.55470184381927
LogLoanPerMOnthPC	-4.88629394294528	5.12068217559174	4.79890678659572	2.61902680940991

3. Ranking of features and Features selection: After finding the important features plotting them highest to lowest impact on loan approval decision. Selecting the important feature on which loan can be approved.

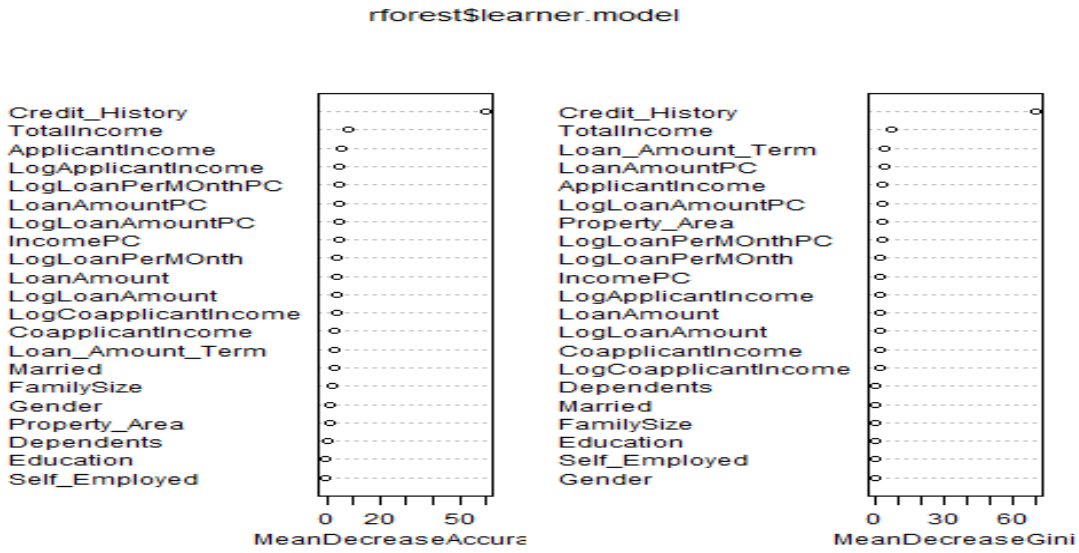

Fig. 7 Ranking of features and Features selection

D. 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.

E. Predicting Class Labels of Test Dataset:

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.

F. Evaluating Prediction:

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

V. CONCLUSIONS

The banking industry is highly competitive. It is sensitive to political and economic conditions in their domestic countries and all over the world. Because of a lot of risks, a key strategy for many banks is to improve their performance by reducing costs and increasing revenues. 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. This application helps the organizations in making the right decision to approve or reject the loan request of the customer. This will definitely help the banking industry to open up efficient delivery channels and avoid the huge financial losses. This model is built using the data mining functions available in the R package. For data selection we are using a web based application which is developed using the J2ee features. After selecting the data, the most important and time consuming step in data model building is the data pre-processing. Classification techniques in R were used to make the data ready for further use. Several R functions and packages are used to build the classification model and predict the customer loan request depends on that banks can arrive on a particular conclusion. 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.

REFERENCES

[1] 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.

[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] Sudhamathy G and Jothi Venkateswaran C. “Analytics Using R for Predicting Credit Defaulters”, IEEE International Conference on Advances in Computer Applications (ICACA), 978-1-5090-3770-4, 2016.

[4] Rajanish Dass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad. Hafiz Alaka, Lukumon O. Oyedele, Muhammad Bilal, Olugbenga O Akinade, Hakeem Owolabi, Saheed Ajayi "Bankruptcy Prediction of Construction Businesses: Towards a Big Data Analytics Approach", IEEE 2015.

[5] Hsieh N.C., and Hung L.P., “A data driven ensemble classifier for credit scoring analysis”, Expert Systems with Applications, vol. 37, pp. 534– 545, 2010.

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

[7] Kumar Arun, Garg Ishan, Kaur Sanmeet, “ Loan Approval Prediction based on Machine Learning Approach”, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 79-81.

[8] Sivasree M S, Rekha Sunny T, “Loan Credibility Prediction System Based on Decision Tree Algorithm”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV4IS090708 Vol. 4 Issue 09, September-2015.

[9] Pramod s. patil, Dr. j. v. Aghav and vikram sareem “An Overview of Classification Algorithm and Ensemble methods in personal credit scoreing”, International journal of Computer Science and Technology (IJCST) ISSN: 09768491 IJCST vol. 7, Issue 2, April- June 2016.