Scientific Journal of Impact Factor (SJIF): 5.71

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

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

RISKY BANK LOANS PREDICTION THROUGH DATA MINING TECHNIQUES - AN EMPIRICAL COMPARATIVE STUDY

Nayani Sateesh[1]

[1]*Assistant Professor, IT Department CVR College of Engineering, Hyderabad - 501510*

Abstract —*nowadays, we are witnessing the financial crisis in the banking sector due to many factors and one amongst those most serious factors is the risky loans. Understanding the customer behavior is more crucial in this context. The volumes of the data being generated due to banking transactions are increasing day by day. To understand this huge volume of data, we need to adopt the data mining techniques to get the insights of the data and take the proper decisions. In this paper, we focus on developing a loan approval model which in turn helps in the prediction of the risky loans using various data mining techniques like Naïve Bayes, Decision Trees and Random Forest with implementation in R language and their prediction accuracies are compared further to select the best algorithm*

Keywords- *Data Mining, Classification, Prediction, Decision Tree, Naïve Bayes classification, Random Forest*

I. INTRODUCTION

The number of transactions in banking sector is rapidly growing and huge volumes of data are being generated day by day. Understanding the customers' behavior and the risks around loan is a major issue and need to be addressed. In this context, Data mining techniques are useful to get the insights of the data and take the proper decisions. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and data systems. The following figure illustrates the basic data mining tasks. Data mining tasks are broadly categorized into the following two categories:

- **Predictive :** induction on the present data for the predictions
- **Descriptive:** characterize properties of the data in a target set.

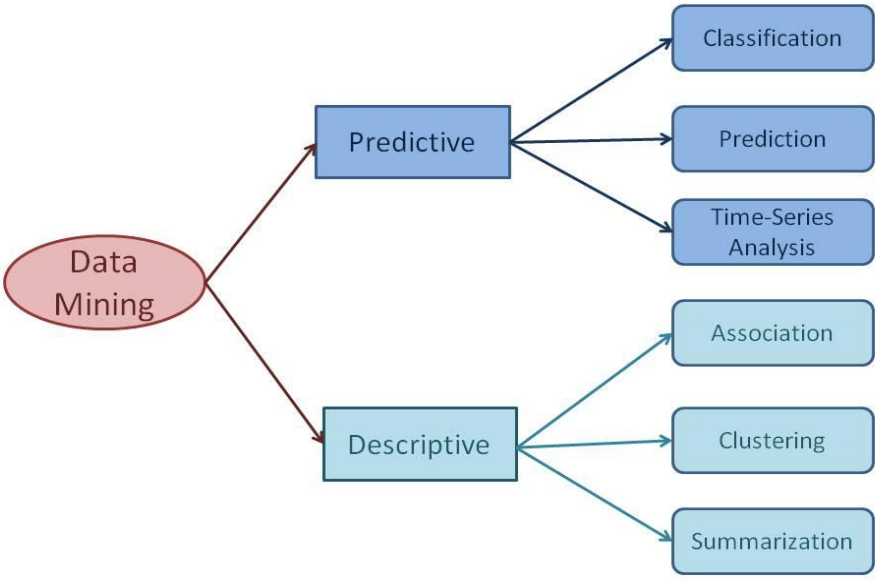


Figure 1: Data Mining Tasks (Source: wideskills.com)

In this paper, a study on loan approval model is done using data mining techniques with implementation in R language to predict the risky loans [1]. The decision of approving loan to a customer is identified through data mining techniques by considering the features of customers of a bank.

II. METHODOLOGY

The data mining techniques that are considered here are Naïve Bayes, Decision Trees and Random Forest [2,3, 4].

A. Naïve Bayes classification : It is based on Bayes' Theorem with an assumption of independence among Predictors.

Pseudo Code:

- Calculate probabilities for each attribute, conditional on the class value.
- Use the product rule to obtain a joint conditional probability for the attributes.
- Use Bayes rule to derive conditional probabilities for each class value.

Once this has been done for all class values, output the class value with the highest probability.

B. Decision Trees : Decision Tree is used to predict class or value of target variables by learning decision rules inferred from training data. It used the tree structure for the data representation.

Pseudo Code:

- Using attribute selection methods like Gini Index, Entropy etc., place the best attribute at the root of the tree.
- Split the training set into subsets in such a way that each subset contains data with the same value for an attribute.
- Repeat the above steps on each subset until we find leaf nodes in all the branches of the tree.

C. Random Forest: It is the widely used machine learning algorithm for classification and regression model with the goal of overcoming over fitting problem.

Pseudo Code: Each tree is grown as follows:

- **Random Record Selection:** Each tree is trained on 2/3rd (approx) of the total training data. Cases are drawn using simple random sampling with replacement. This sample will be the training set for growing the tree.
- **Random Variable Selection:** Some predictor variables (say, m) are selected at random out of all the predictor variables and the best split on this m is used to split the node. The value of m is held constant during the forest growing.
- For each tree, using the leftover data, calculate the misclassification rate - out of bag (OOB) error rate. Aggregate error from all trees to determine overall OOB error rate for the classification.
- Each tree gives a classification on leftover data (OOB), and we say the tree "votes" for that class.

The forest chooses the classification having the most votes over all the trees in it.

III. IMPLEMENTATION

Below are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends.

- Classification
- Prediction

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

The models that we consider in this paper are implemented using R language. The following are the prominent packages that we use for the model building and verifying the accuracy of the model being built.

- *library(c50)* - *Decision Trees*
- *library(e1071)* - *Naive Bayes*
- *library(randomForest)* - *Random Forest*
- *library(caret)* - *ConfusionMatrix*
- *library(gmodels)* - *CrossTable*

The basic steps involved in model building and analysis are:

A. Data Collection: In our study, we have taken the secondary data (source: *https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/credit.csv)* The following tables 1 dipicts the 21 feature variables which are taken in the dataset construction. ***default*** is the class variable

Attributes/Features	
checking_balance	savings_balance
months_loan_duration	employment_length
credit_history	installment_rate
purpose	personal_status
amount	other_debtors
residence_history	existing_credits
property	dependents
age	telephone
installment_plan	foreign_worker
housing	job
default (Class variable)	

Table 1: Dataset Attributes

B. Data Pre-processing: For better understanding and analysis of data, it should be processed in such a way that it increases the knowledge of the user.

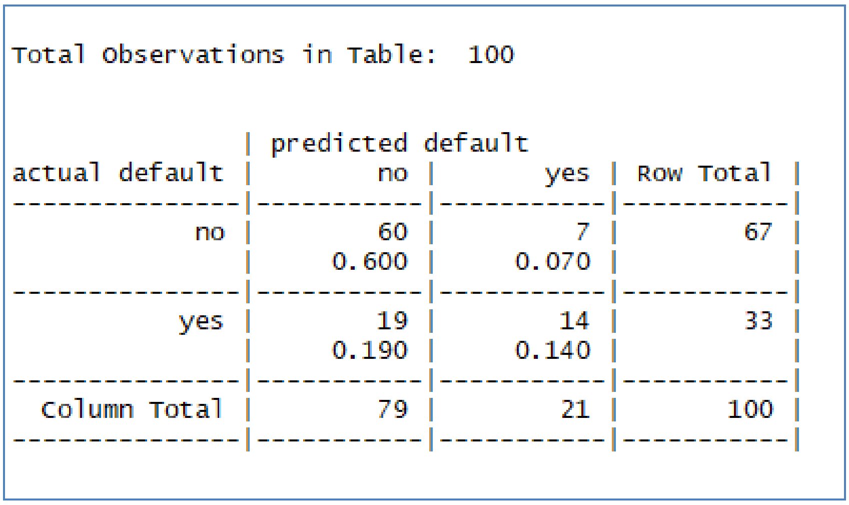
C. Data Splitting: Data set should be splitted into two disjoint sets such as training data to train/construct the model and test data to test the model. For data split, we use heuristic approach 80:20 or 90:10 ratios for training and test data respectively. In this study we considered 90:10.

D. Training the Model: We build the model using the training data in the above ratio and test the predictions based on the algorithms we use.

- C5.0() used for Decision tree model building
- naiveBayes() used for Naïve Bayes classifier building
- randomForest() used for Random forest building
- predict() is used to test the predictions.

E. Model Evaluation: To evaluate the accuracy of the algorithms, we use cross tabulation/confusion matrix and compute the F-measure. Since the data set size is 1000, the tested data is of size 100 due to 90:10 ratio.

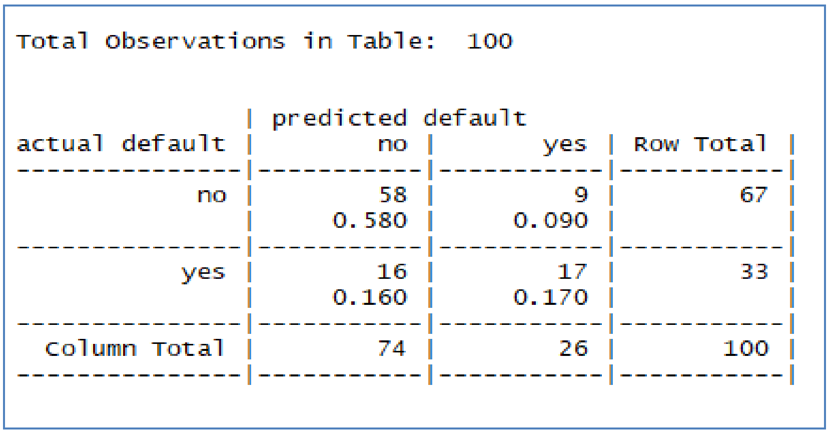
Confusion Matrix- Decision Tree



```
Total Observations in Table:  100


               | predicted default
actual default |        no |       yes | Row Total |
---------------|-----------|-----------|-----------|
            no |        60 |         7 |        67 |
               |     0.600 |     0.070 |           |
---------------|-----------|-----------|-----------|
           yes |        19 |        14 |        33 |
               |     0.190 |     0.140 |           |
---------------|-----------|-----------|-----------|
  Column Total |        79 |        21 |       100 |
---------------|-----------|-----------|-----------|
```

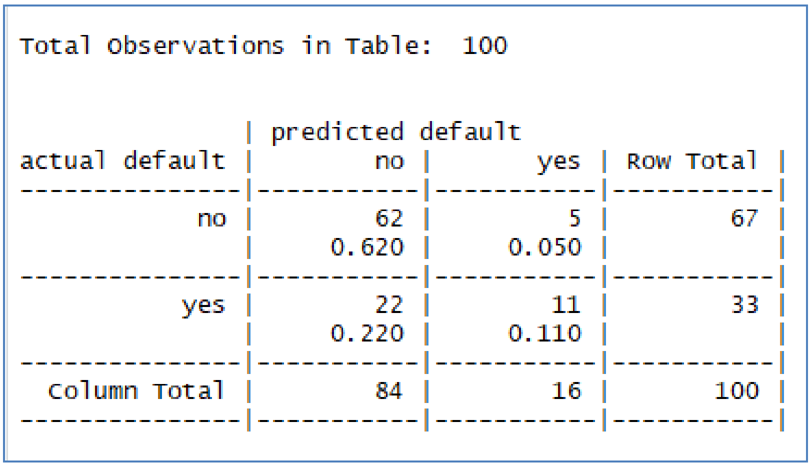
Confusion Matrix- Naïve Bayes



```
Total Observations in Table:  100


               | predicted default
actual default |        no |       yes | Row Total |
---------------|-----------|-----------|-----------|
            no |        58 |         9 |        67 |
               |     0.580 |     0.090 |           |
---------------|-----------|-----------|-----------|
           yes |        16 |        17 |        33 |
               |     0.160 |     0.170 |           |
---------------|-----------|-----------|-----------|
  Column Total |        74 |        26 |       100 |
---------------|-----------|-----------|-----------|
```

Confusion Matrix- Random Forest



```
Total Observations in Table:  100


               | predicted default
actual default |        no |       yes | Row Total |
---------------|-----------|-----------|-----------|
            no |        62 |         5 |        67 |
               |     0.620 |     0.050 |           |
---------------|-----------|-----------|-----------|
           yes |        22 |        11 |        33 |
               |     0.220 |     0.110 |           |
---------------|-----------|-----------|-----------|
  Column Total |        84 |        16 |       100 |
---------------|-----------|-----------|-----------|
```

IV. RESULTS AND DISCUSSION

The following table 2 summarizes the experimental results with respect to each algorithm.

	Accuracy	**Precision**	**Recall**	**F-measure**
Decision Tree	74.00%	89.55%	75.94%	82.21%
Naïve Bayes	75.00%	51.51%	78.38%	62.17%
Random Forest	73.00%	92.53%	73.80%	82.11%

Table 2: Summary of Empirical comparison

From the results, we can infer that the F-measure is good for the Decision Tree model. Hence, it is advisable to follow decision tree for the prediction/identification of risky loans. The accuracies of the prediction models change with the size of dataset, variables included in the analysis.

V. CONCLUSION

The objective of this paper is to study different machine learning techniques to predict risky bank loans and to compare them to find the best prediction model. Based on the results we obtained and the F-Measure, we infer that the Decision Trees model is the best model for the risky bank loan prediction.

REFERENCES

[1] G. Sudhamathy, C. Jothi Venkateswaran, "Analytics Using R for Predicting Credit Defaulters", Advances in Computer Applications (ICACA), IEEE International Conference, pp. 66-71, Oct. 2016.

[2] Rafik Khairul Amin, Indwiarti, Yuliant Sibaroni, "Implementation of Decision Tree Using C4.5 Algorithm in Decision Making of Loan Application by Debtor", 3rd International Conference on Information and Communication Technology (ICICT), IEEE International Conference, pp. 75-80, 2015.

[3] A. Abhijit, and P.M. Chawan, "Study of Data Mining Techniques used for Financial Data Analysis", International Journal of Engineering Science and Innovative Technology, vol. 2, no. 3, pp. 503-509, 2013

[4] Archana Gahlaut, Tushar, Prince Kumar Singh, "Prediction analysis of risky credit using Data mining classification models", Proc. Computing, Communication and Networking Technologies (ICCCNT), 8th International Conference, July 2017.