International Journal of Engineering & Technology, 7 (2.21) (2018) 339-341

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Exploratory analysis on prediction of loan privilege for customers using random forest

K. Ulaga Priya[1*], S. Pushpa[2], K. Kalaivani[3], A. Sartiha[4]

[1]Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.
[2]Department of Computer Science and Engineering, St.Peters University, Chennai, India.
[3]Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.
[4]Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.
**Corresponding author E-mail:upriya.se@velsuniv.ac.in*

Abstract

In Banking Industry loan Processing is a tedious task in identifying the default customers. Manual prediction of default customers might turn into a bad loan in future. Banks possess huge volume of behavioral data from which they are unable to make a judgement about prediction of loan defaulters. Modern techniques like Machine Learning will help to do analytical processing using Supervised Learning and Unsupervised Learning Technique. A data model for predicting default customers using Random forest Technique has been proposed. Data model Evaluation is done on training set and based on the performance parameters final prediction is done on the Test set. This is an evident that Random Forest technique will help the bank to predict the loan Defaulters with utmost accuracy.

Keywords: *Machine learning, random forest, prediction, R.*

1. Introduction

Credit Risk plays a vital role in Banking Domain. The success of a Bank mainly depends upon the evaluation of credit Risk. Before sanctioning a loan, the officials should be able to predict whether the customer is a defaulter or Non Defaulter. The aim of this paper is to apply machine learning technique on dataset which has 1000 cases and 7 numerical and 6 categorical attributes. The creditability of a customer for sanctioning loan depend on several parameters, such as credit history, Installment etc. This paper is summarized as follows. Section 2 discusses about the Supervised learning Random Forest Technique. In Section 3 the Methodology adopted for predicting the default customers is described. In section 4 Experimental results and the performance measures were discussed. In Section 5, the conclusion of data analysis was discussed.

2. Machine learning: random forest

Random Forest is an ensemble machine learning method which efficiently performs both classification and regression tasks. It creates multiple decision 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).

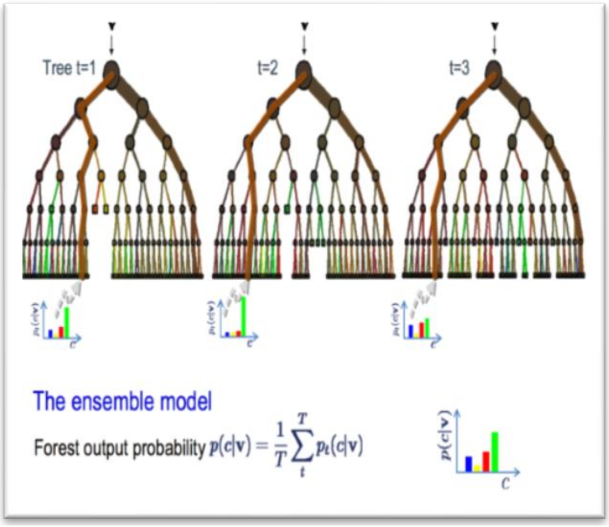

Tree t=1
t=2
t=3
The ensemble model
Forest output probability $p(c|v) = \frac{1}{T}\sum_{t}^{T} p_t(c|v)$
p(c|v)
c

Fig. 1: Random forest

3. Methodology

The methodology adopted for predicting loan Defaulters using Random Forest Technique is derived using a flow diagram. The steps involved in Building the data model is depicted below:

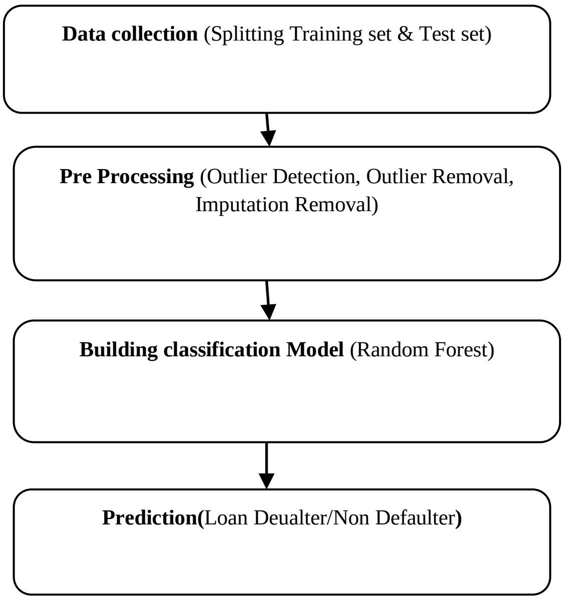

Data collection (Splitting Training set & Test set)
Pre Processing (Outlier Detection, Outlier Removal, Imputation Removal)
Building classification Model (Random Forest)
Prediction(Loan Deualter/Non Defaulter**)**

Fig. 2: Machine learning methodology

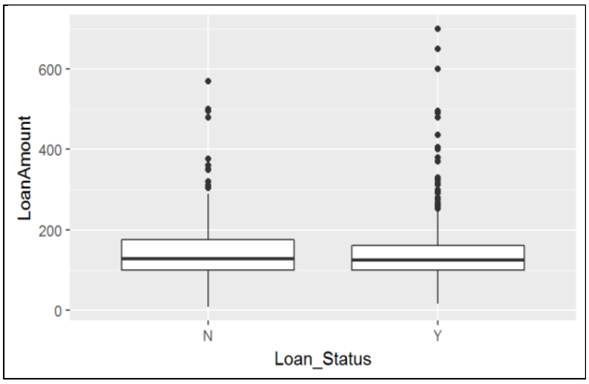
Data collection

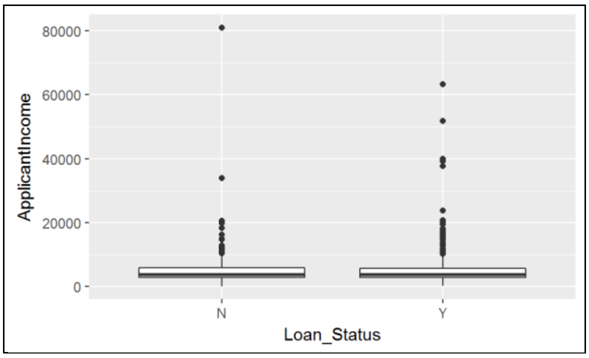
The data set collected for predicting loan default customers is splited into Training set and Test set. Generally 80:20 ratio is applied to split the Training set and Test set. The Data Model which was created using Random Forest is applied on the Training set and based on the test result accuracy, Test set prediction i done. Following are the attributes

Atrribute Name	Category
Loan_ID	Qualitative
Gender	Categorical
Married	Categorical
Dependents	Qualitative
Education	Categorical
Self_Employed	Categorical
ApplicantIncome	Qualitative
CoapplicantIncome	Qualitative
LoanAmount	Qualitative
Loan_Amount_Term	Qualitative
Credit_History	Qualitative
Property_Area	Categorical
Loan_Status	Categorical

Pre processing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers has to be removed and also variable conversion need to be done.





Correlation among attributes

Based on the correlation among attributes it was observed that attributes that are significant individually include property area, education, loan amount, and lastly credit history, which is the strongest among all. Some variables such as applicant income and co applicant income are not significant alone, which is strange since by intuition it is considered as important. The correlation among attributes can be identified using corplot and boxplot in R Platform.

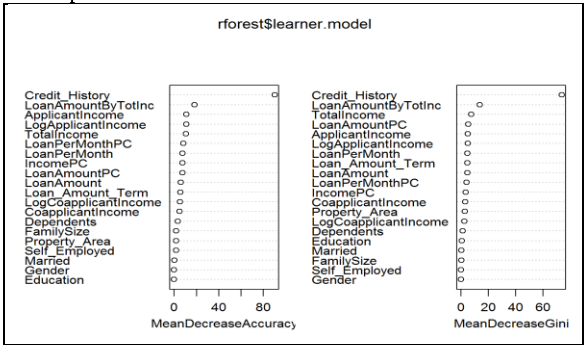
Building the classification model using random forest

For predicting loan defaulters and non defaulter's problem, Random Forest prediction model is effective because of the following reasons:

- It provides better results in classification problem.
- It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.
- It produces out of bag estimate error which has proven to be unbiased in many tests.
- It is relatively easy to tune with.

For evaluation method, the performance parameter "Accuracy" is chosen since the aim is to correctly predict as many cases as possible. Finally, using OOB error estimate as evaluation, we will tune the two parameters below of the random forest model, since they are most likely to have the biggest effect on our final accuracy.

The data model was built using Random Forest in R and the importance of variable in the dataset was also derived. It is observed that credit History is identified as the most important variable. The random forest also finds that most of the variables are not important.



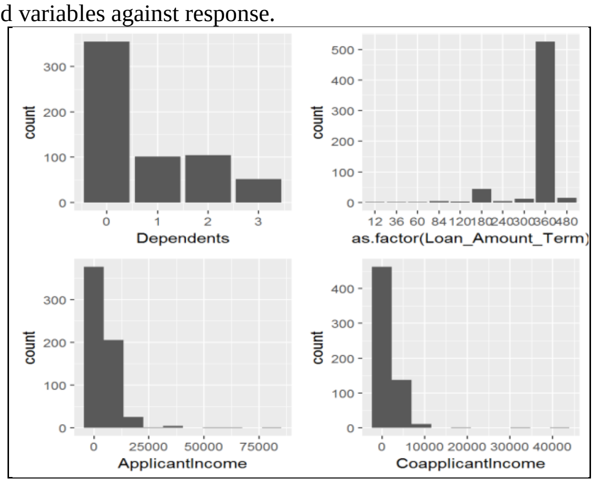
Predicting default customers

```
predictions<-unname(predict(final_rf,newtest[]))
solution<-data.frame(Loan_ID=test[1],Loan_Status=predictions
## 0.8110553
```

We noticed that 299 cases in the test set are predicted as "Y", which is more than 81%, whereas in the training set only about 69% had this status.

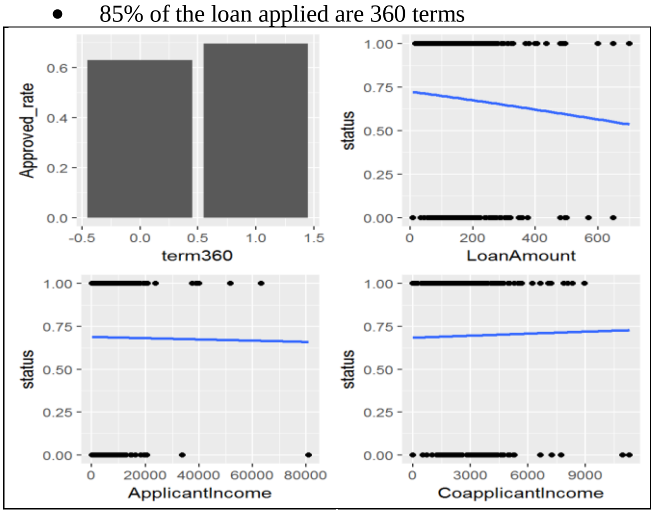
4. Experimental results

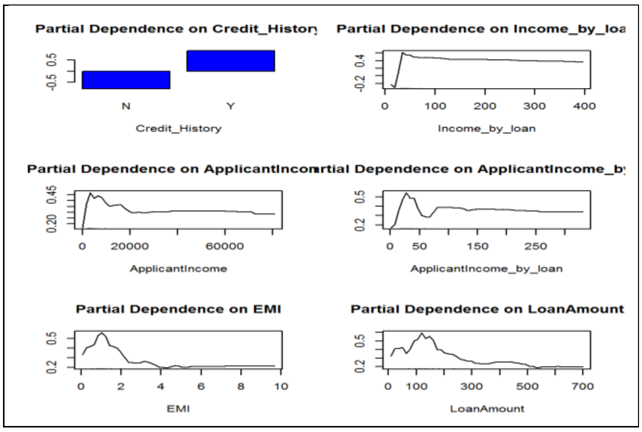
This section is categorized into two parts, univariate visualizatio and variables against response.



Some of the observation of plots is

- Most of the applicants do not have dependents
- The applicant income and coapplicant income has a similiar extremely left-skewed distribution
- 85% of the loan applied are 360 terms





5. Conclusion

The analytical process started from data cleaning and processing, missing value imputation with mice package, then exploratory analysis and finally model building and evaluation. The best accuracy on public test set is 0.811. This brings some of the following insights about loan approval.

- Applicants with credit history not passing guidelines mostly fails to get approved, probably because that they have a higher probability of not paying back.
- Most of the time, applicants with high income, sanctioning low amount is more likely to get approved, which makes sense, those applicants are more likely to pay back their loans.
- Some basic characteristic such as gender and the status of marriage seems not to be taken into consideration by the company

References

[1] Sudhamathy G & Venkateswaran J, “Analytics Using R for Predicting Credit Defaulters”, *IEEE international conference on advances in computer applications*, (2016).

[2] Jina Y & Zhua Y, “A data-driven approach to predict default risk of loan for online Peer-to-Peer(P2P) lending”, *Fifth International Conference on Communication Systems and Network Technologies*, (2015).

[3] Odeh O, Koduru P, Featherstone A, Das S & Welch SM, “A multi-objective approach for the prediction of loan defaults”, *Elsevier/ Expert Systems with Applications*, Vol.38, (2011), pp.8850–8857

[4] Aboobyda JH & Tarig MA, “Developing Prediction Model Of Loan Risk In Banks Using Data Mining”, *Machine Learning and Applications: An International Journal (MLAIJ)*, Vol.3, No1, (2016), pp. 1–9.

[5] Tsai MC, Lin SP, Cheng CC & YP Lin, “The consumer loan default predicting model–An application of DEA–DA and neural network”, *Elsevier Expert Systems with Applications*, Vol.36, (2009), pp.11682–11690