LOAN PREDICTION PROBLEM

Ashik Pal
(ashikpal531@gmail.com, Lovely Professional University, India)

Abstract

Loan prediction is a very common real-life problem that each bank faces at least once in its lifetime. If done correctly, it can save a lot of man hours at the end of a retail bank. Banks have presence across all urban, semi urban and rural areas. Customers first apply for home loan then the bank validates the customer eligibility for loan. Banks wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. It is a classification problem where we have to predict whether a loan would be approved or not. The motive of this paper is to classify a customer on the basis of his previous history as well as the previous history of the bank in order to know whether the person is eligible for a loan approval or not. The classification is done on the basis of several parameters according to our datasets like age, gender, education, previous loan debts, income, EMI, number of years, credit history and other similar kind of parameters. Here I've used an effective prediction technique that helps the banks to predict the credit risk for customers who have applied for loan. In this paper I've used different models on the training and testing datasets in order to get the highest accuracy. A theoretical approach is described in the paper which can be used for making the accurate or correct decision to approve or reject the request for loan of the customers.

Keywords

Classification, decision tree, logistic regression, machine learning, prediction, train, test data

1. Introduction

The main work of a bank among others is to provide monetary help to its customers in form of loans. The primary or the main source of income or profit of a bank lies through these loans. They charge interest on the money given to the customer and that interest acts as the profit for a bank. Also a major objective of a bank is to invest their assets in safe hands where it is. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. [3] Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasizes different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system.[11] Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each features taking part in loan

processing and on new test data same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance Company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various departments of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities. Other than predicting loans banks and other businesses in the financial industry use machine learning technology for other purposes also like to identify important insights in data, and prevent fraud. The insights can identify investment opportunities, or help investors know when to trade. Data mining can also identify clients with high-risk profiles, or use cyber surveillance to pinpoint warning signs of fraud. [4]

2. Literature Review

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.[5] Loan Prediction is a classification problem where we have to predict whether a loan would be approved or not. In a classification problem, we have to predict discrete values based on a given set of independent variable(s). [8] Classification can be of two types: **Binary Classification**: In this classification we have to predict either of the two given classes. For example: classifying the gender as male or female, predicting the result as win or loss, etc. **Multiclass Classification**: Here we have to classify the data into three or more classes. For example: classifying a movie's genre as comedy, action or romantic, classify fruits as oranges, apples, or pears, etc. We will list out all the possible factors that can affect the outcome. This process is also known as **Hypothesis Generation**.[10] This is a very important stage in any data science/machine learning pipeline. It involves understanding the problem in detail by brainstorming as many factors as possible which can impact the outcome. It is done by understanding the problem statement thoroughly and before looking at the data. Below are some of the factors which I think can affect the Loan Approval (dependent variable for this loan prediction problem):

- Salary: Applicants with high income should have more chances of loan approval.
- Previous history: Applicants who have repaid their previous debts should have higher chances of loan approval.
- Loan amount: Loan approval should also depend on the loan amount. If the loan amount is less, chances of loan approval should be high.
- Loan term: Loan for less time period and less amount should have higher chances of approval.
- EMI: Lesser the amount to be paid monthly to repay the loan, higher the chances of loan approval.

3. Proposed Model

In the dataset with which I'm working I've used three machine learning models logistic regression, decision trees and random forests.
Before we talk about the models lets first know the data. We have two datasets, train and test. We have 12 independent variables and 1 target variable, i.e. Loan_Status in the train dataset. We have similar features in the test dataset as the train dataset except the Loan_Status. We will predict the Loan_Status using the model built using the train data. We have 614 rows and 13 columns in the train dataset and 367 rows and 12 columns in test dataset.

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	Credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (Y/N)

3.1 Logistic Regression

Logistic Regression, in simple terms, predicts the probability of occurrence of an event by fitting data to a logit function. Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.[9] This type of models is part of a larger class of algorithms known as Generalized Linear Model or GLM.

3.2 Decision Tree

Decision trees create a set of binary splits on the predictor variables in order to create a tree that can be used to classify new observations into one of two groups.[9] Here, we will be using classical trees. The algorithm of this model is the following:

- Choose the predictor variable that best splits the data into two groups;
- Separate the data into these two groups;
- Repeat these steps until a subgroup contains fewer than a minimum number of observations;

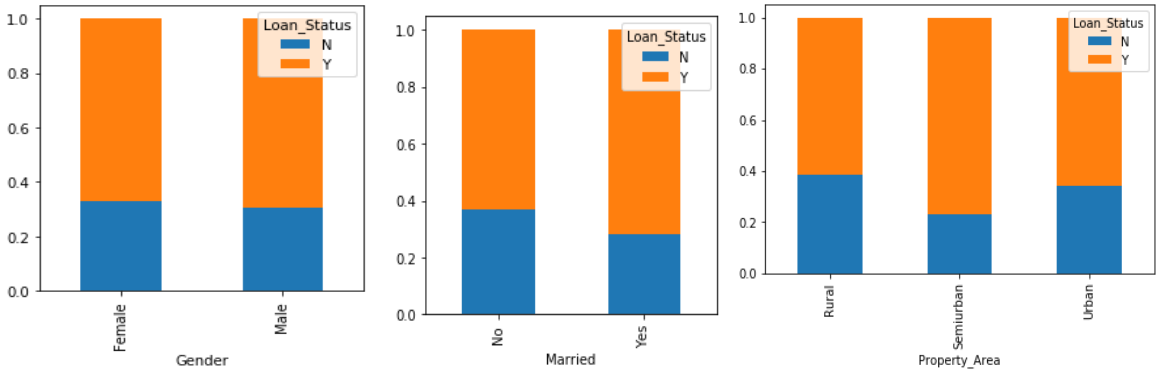
3.3 Random Forest

A random forest is an ensemble learning approach to supervised learning.[13] This approach develops multiple predictive models, and the results are aggregated to improve classification. The algorithm is as follows:

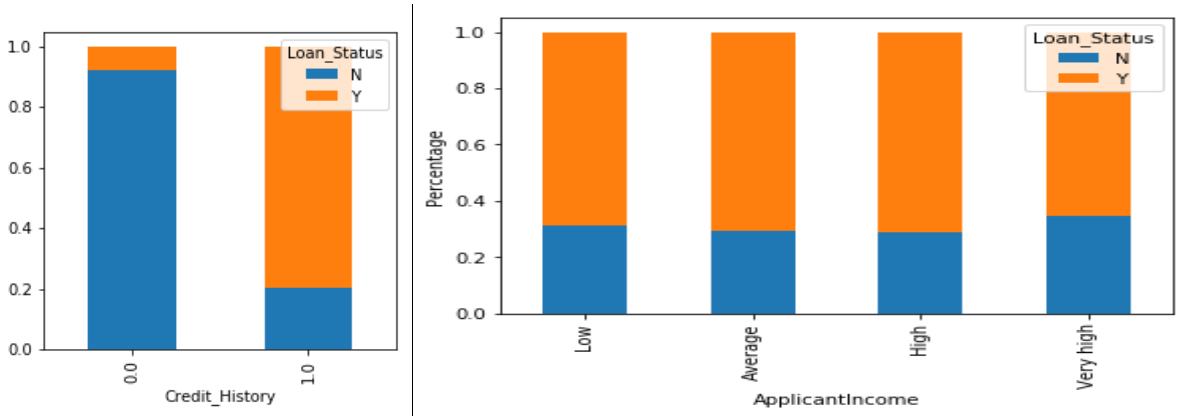
- Grow many decision trees by sampling;
- Sample m < M variables at each node;
- Grow each tree fully without pruning;
- Terminal nodes are assigned to a class based on the mode of cases in that node;
- Classify new cases by sending them down all the trees and taking a vote.

4. Result & Discussion

A. It can be inferred that the proportion of male and female applicants is more or less same for both approved and unapproved loans.
B. Proportion of married applicants is higher for the approved loans.
C. Proportion of loans getting approved in semi urban area is higher as compared to that in rural or urban areas.

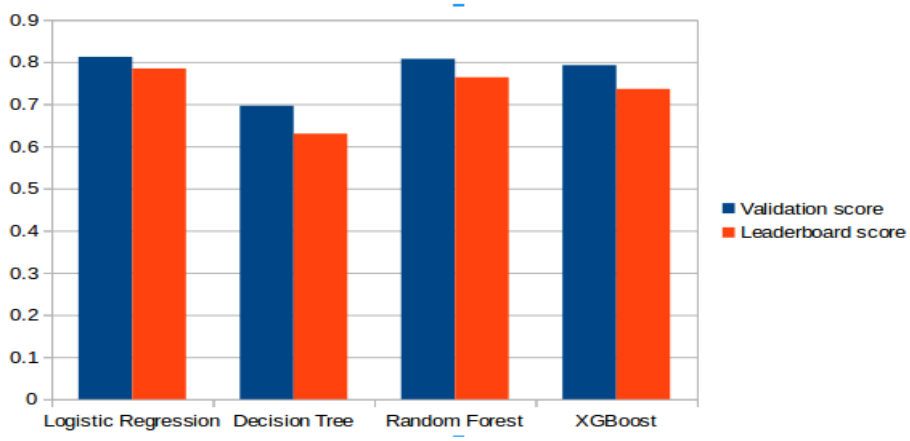


D. People with a good credit history are more likely to get their loans approved.
E. the proportion of approved loans is higher for Low and Average Loan Amount as compared to that of High Loan Amount



5. Conclusion

The analytical process started from understanding the data then analyzing it and then treating the missing value by using imputation, then exploratory analysis and finally model building and evaluation. The best accuracy on test set is 0.78. This brings some of the following insights about loan eligibility. We can see that Credit History is the most important feature followed by Balance Income, Total Income, and Loan Amount. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which makes sense, more likely to pays back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the bank. After trying and testing four different algorithms, the best accuracy on the public leaderboard is achieved by Logistic Regression (0.7847), followed by Random Forest (0.7638).



6. References

[1] Sarwesh Site, Dr. Sadhna K. Mishra, " A Review of Ensemble Technique for Improving Majority Voting for Classifier", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 1, pp. 177- 180,January 2013.

[2] A.R.Ghatge, P.P.Halkarnikar, "Ensemble Neural Network Strategy for Predicting Credit Default Evaluation", International Journal of Engineering and Innovative Technology, Vol. 2, Issue 7, pp. 223-225, January 2013.

[3] Maher Ala'raj and Maysam Abbod,"A systematic credit scoring model based on heterogeneous classifier ensembles", Innovations in Intelligent Systems and Applications (INISTA), pp. 1-7, September 2015.

[4] Marc Claesen, Frank De Smet, Johan A.K. Suykens and Bart De Moor, "A Library for Ensemble Learning Using Support Vector Machines", Journal of Machine Learning Research 15, pp. 141-145 , January 2014.

[5] Gang Wang, Jian Ma, "Study of corporate credit risk prediction based on integrating boosting and random subspace", Elsevier Expert Systems with Applications, Vol. 38, Issue 11, pp. 13871–13878,October 2011.

[6] Wo- Chiang Lee, "Genetic Programming Decision Tree for Bankruptcy Prediction", Joint Conference on Information Science, October 2006.

[7] M. Yaghini , T. Zhiyan , and M. Fallahi, “A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks”, Int'l Conf. Data Mining , 2011.
[8] Kumar Arun, Garg Ishan, Kaur Sanmeet, May- Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)
[9] Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques, Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.
[10] A. Goyal and R. Kaur, “Loan Prediction Using Ensemble Technique.,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no. 3, pp. 523–526, 2016.
[11] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.
[12] Mean Decrease Accuracy https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.hml
[13] Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei University of Technology, Hefei 23009, China.
[14] Kumar Arun, Garg Ishan, Kaur Sanmeet, May- Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)
[15] M. Yaghini , T. Zhiyan , and M. Fallahi, “A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks”, Int'l Conf. Data Mining , 2011.