International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Loan Prediction using Decision Tree and Random Forest

Kshitiz Gautam[1], Arun Pratap Singh[2], Keshav Tyagi[3], Mr. Suresh Kumar[4]

[1-3]BTech student, Dept. of IT, Galgotias College of Engineering and Technology, Greater Noida, U.P
[4]Assistant Professor, Dept. of IT, Galgotias College of Engineering and Technology, Greater Noida, U.P

Abstract - *In India, the number of people or organization applying for loan gets increased every year. The bank employees have to put in a lot of work to analyse or predict whether the customer can pay back the loan amount or not (defaulter or non-defaulter) in the given time. The aim of this paper is to find the nature or background or credibility of the client that is applying for the loan. We use exploratory data analysis technique to deal with the problem of approving or rejecting the loan request or in short loan prediction. The main focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.*

Key Words: Loan, Prediction, Machine Learning, Training, Testing.

1. INTRODUCTION

The term banking can be referred to as receiving and protecting money that is deposited by an individual or an entity. It also includes lending money to people and businesses which has to be paid back within the given amount of time without failing. Banking is a sector that is regulated in most of the countries as it is an important factor in determining the financial stability of the country. The prime goal in banking sector is to invest their assets in safe hands where there are less chances of failure. Today many banks and financial companies approve loan after a stressful, long and weary process of verification but still there is no surety whether the chosen applicant is credible or not or in other words if he is able to return the amount with interest in the given time. The purpose of the loan can be anything based on the customer needs. Loans are broadly divided as open ended and close-ended loans.

Examples of open-end loans are credit cards and a home equity line of credit (HELOC).

Close-ended loans decreases with each payment. It means the amount is reduced after an instalment.

In other words, it is a legal term that cannot be modified by the borrower. Personal loans, mortgages, auto payments, EMI and student loans are the most common examples of close-ended loans.

Secured or collateral loan are those loans that are protected by an asset. Houses, Vehicles, Savings accounts are the personal properties used to secure the loan.

2. DATA SET

A collection of data is taken from the banking sector. The Data set is in ARFF (Attribute-Relation File Format) format that is acceptable by Weka. ARFF file is composed of tags that include the name, types of attributes, values and data itself. For this paper we are using 12 attributes like gender, marital status, qualification, income, etc.

The table below represents the data set that we have used:

Table-1: Data set variables along with description and type

Variable Name	Description	Type
Loan_ID	Unique ID	Integer
Gender	Male/Female	Character
Marital_Status	Applicant married(Y/N)	Character
Dependents	Number of Dependents	Integer
Education_Qualification	Graduate/Under Gradute	String
Self_Employed	Self-employed(Y/N)	Character
Applicant_Income	Applicant income	Integer
Co_Applicant_Income	Co-applicant income	Integer
Loan_Amount	Loan amount in thousands	Integer
Loan_Amount_Term	Term of loan in months	Integer
Credit_History	Credit history meets guidelines	Integer
Property_Area	Urban/Semi urban/Rural	String
Loan_Status	Loan Approved(Y/N)	Character

Now in machine learning model, we first apply the training data set, in this data set the model is trained with known examples. The entries of new applicants will act as a test data which are to be filled at the time of submitting the application. After performing such tests, model can determine whether the loan approved to the person is safe or not basically about the loan approval on the basis of the various training data sets.

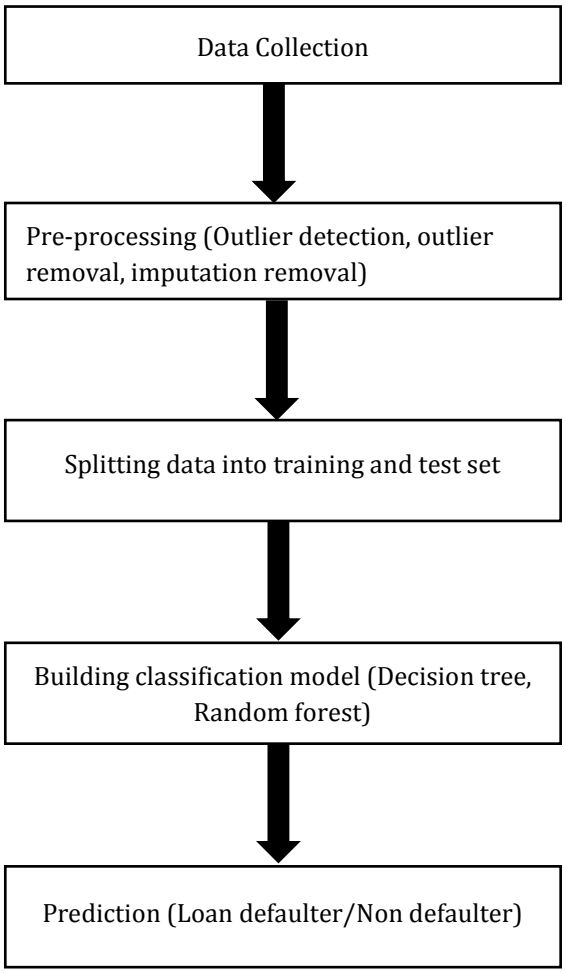

Fig-1: Chronology of Data

The diagram above gives us an outline on how data is used in this machine learning process or model.

Basically, it is divided into four parts in which we use data to predict the outcome of the whole process. First, we use training data set to train our model. After the model is trained, then we test it with unknown examples from the same scenario.

Another process that we use before testing and training data is data pre-processing. In data pre-processing we remove all sorts of values that can cause an error like redundant values, incomplete values, missing data, etc.

3. LOAN PREDICTION METHODOLOGY

The diagram 2 represents the working of our model. It basically gives us a rough idea on how the loan prediction system works. After collecting data, we use feature selection process on data. Feature selection can be defined as a process of reducing number of input variables when we develop a predictive model.

Feature selection is divided into two parts i.e. supervised method and unsupervised method. Supervised method is divided into three parts which are wrapper, filter and intrinsic. In supervised method we use target variable to remove discrepancies in data. While in unsupervised method we do not use target variable to remove discrepancies. Unsupervised method uses the process of correlation.

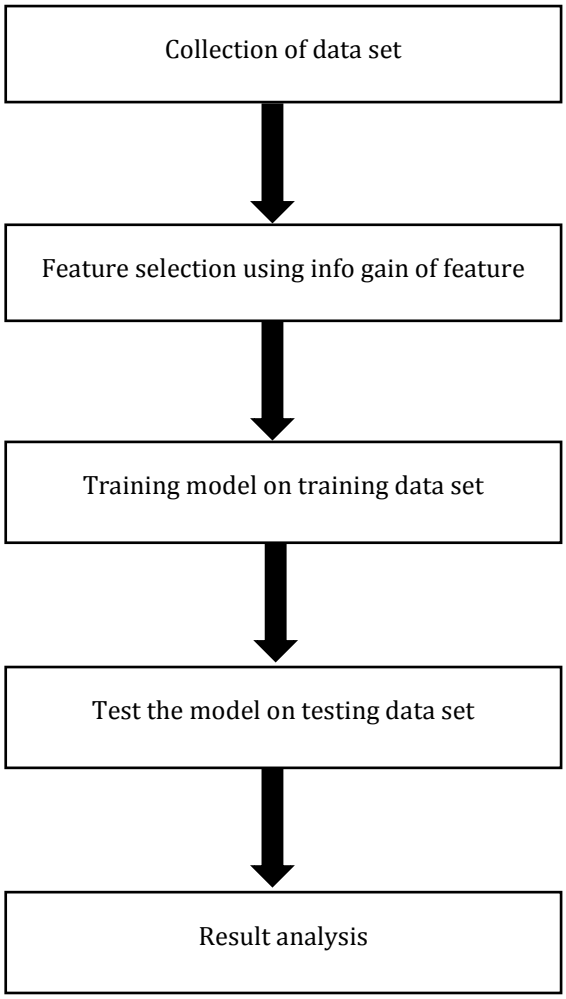

Fig-2: Loan Prediction Methodology

4. WORKING OF THE MODEL

We have represented the working of the model through a use case diagram. The figure below represents the attributes, process of the model that we have built.

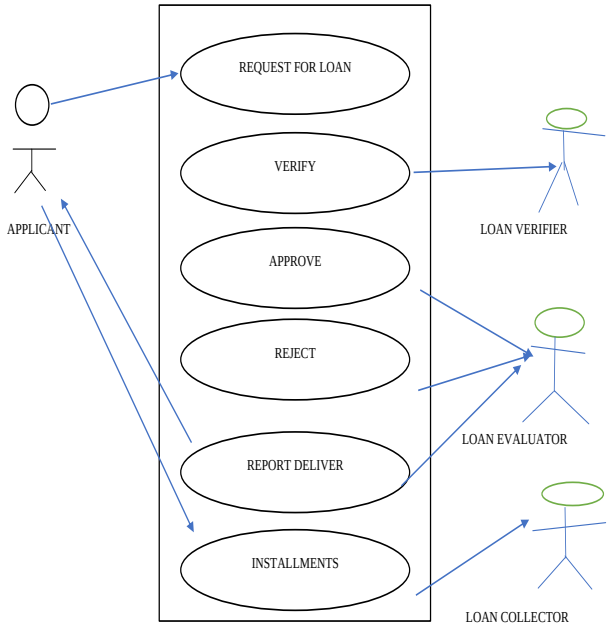

Fig-3: Use case diagram

Table-2: Use case diagram variable and description

Actor	Applicant, Loan Verifier, Loan Evaluator, Loan Collector
Description	An applicant requests for a loan. After request loan verifier verified its document and transfer to loan evaluator may approve or reject the loan.
Data	Applicant personal information and its documents.
Stimulus	User command issue by online loan and application.
Response	Loan may be approved or may be rejected.
Comments	Improve installment policy.

5. EXPLORATORY DATA ANALYSIS

1. The one whose salary is more can have a greater chance of loan approval.
2. The one who is graduate has a better chance of loan approval.
3. Married people would have an upper hand than unmarried people for loan approval.
4. The applicant who has a smaller number of dependents have a high probability for loan approval.
5. The lesser the loan amounts the higher the chance for getting loan.

6. Model used for training and testing

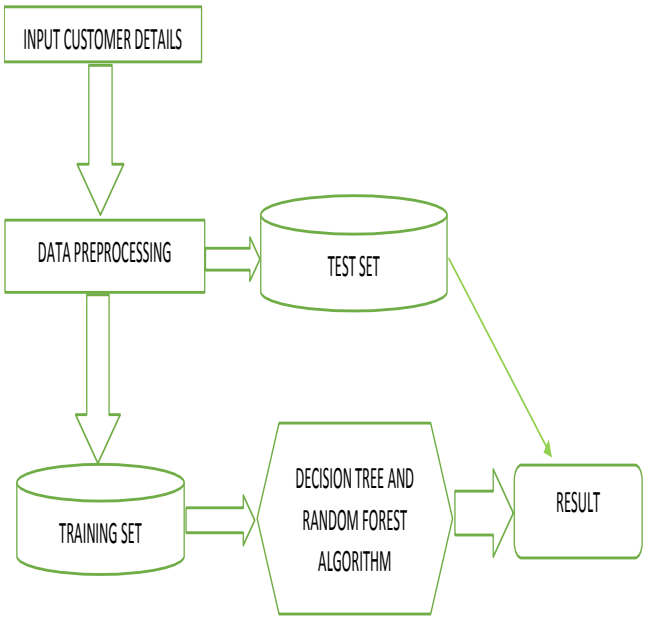

Fig- 4: Training and testing model

7. MACHINE LEARNING METHODS

Two machine learning classification models are used for the prediction of application that can be used in android applications. These models can also be accessed in the open source software R, which is licensed under GNU GPL. The brief description of each model is explained below.

7.1 Decision tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin toss comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

This model is an extension of C4.5 classification algorithms. We experimented with J48 Decision Tree classifier which is an implementation of C4.5 Decision Tree. In case of this classifier, the lower the confidence factor, the more pruning is done. For this we have used different confidence factors and analysed them with higher confidence factor and with the increase of confidence factor the accuracy has increased in each case. With the confidence factor of 0.15 the best accuracy is 62.12% and with a confidence factor of 0.25 it is 63.39%. It means that when less pruning is done the accuracy improves.

7.2 Random forest

Random forest or random decision forests are an ensemble learning method used for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

We have done several trials with Random Forest with different parameters: executions with supervised and unsupervised discretization's (equal-frequency and equal-width), with all attributes. In the experiments without attribute selection the best result was 85.75% and it was achieved with unsupervised equal-frequency 5 bins discretization with 450 trees and seed equal to 4.

Table-3: Parameter setting for machine learning models

Model	Parameter Setting
Decision Tree	Min Split=20, Max Depth=30, Min Bucket=7
Random Forest	Number of trees=450, number of variables=8

8. CONCLUSIONS

The main purpose of the paper is to classify and analyze the nature of the loan applicants. From a proper analysis of available data and constraints of the banking sector, it can be concluded that by keeping safety in mind that this product is much effective or highly efficient. This application is operating efficiently and fulfilling all the major requirements of Banker. Although the application is flexible with various systems and can be plugged effectively.

This paper work can be extended to higher level in future so the software could have some better changes to make it more reliable, secure, and accurate. Thus, the system is trained with a present data sets which may be older in future so it can also take part in new testing to be made such as to pass new test cases.

There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system. So, in the near future the so – called software could be made more secure, reliable and dynamic weight adjustment. In near future this module of prediction can be integrated with the module of automated processing system.

REFERENCES

[1] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[2] A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.

[3] G. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".

[4] A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".

[5] Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. Expert systems with Applications, 37(1), 534-545.

[6] https://en.wikipedia.org/wiki/Exploratory_data_analysis

[7] https://www.experian.com/blogs/ask-experian/credit-education/score- basics/what-is-a-good-credit-score/