RESEARCH ARTICLE OPEN ACCESS

Loan Prediction using Machine Learning Algorithms

Sanket Bhattad [1], Sumit Bawane [2], Shweta Agrawal [3], Unnati Ramteke [4],
Dr. P. B. Ambhore [5]

[1],[2],[3,[4] B.Tech student, Dept. of IT, Gossvernment college of Engineering, Amravati
[5] Assistant Professor, Dept. of IT, Government college of Engineering, Amravati - Maharashtra

ABSTRACT
In India, the number of people or organization applying for loan is 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, 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 focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.

Keywords: **-** Loan, Prediction, Machine Learning, Training**.**

I. INTRODUCTION

Distribution of the loans is the core business part of almost every bank. The main portion the bank's asset is directly came from the profit earned from the loans distributed by the banks. 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. 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.

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 ornot. 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 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.

II. 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.

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.

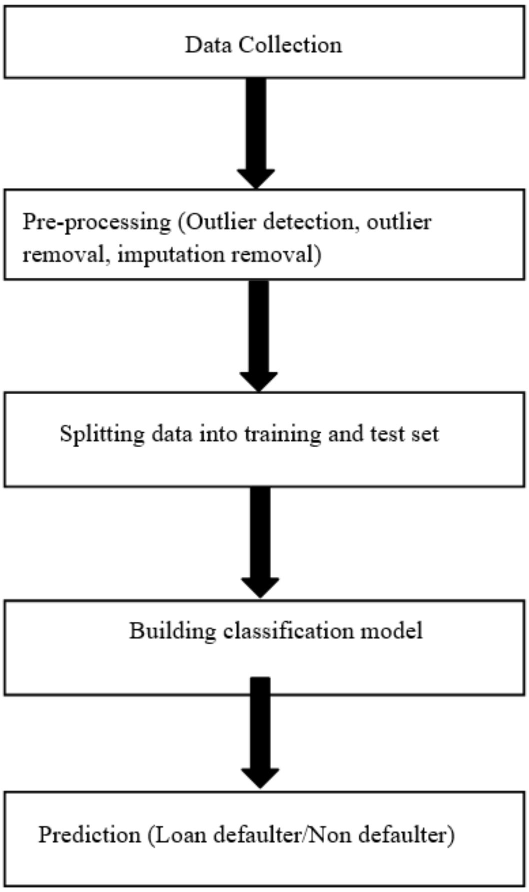
Data Collection
Pre-processing (Outlier detection, outlier removal, imputation removal)
Splitting data into training and test set
Building classification model
Prediction (Loan defaulter/Non defaulter)

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.

III. LOANPREDICTION 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

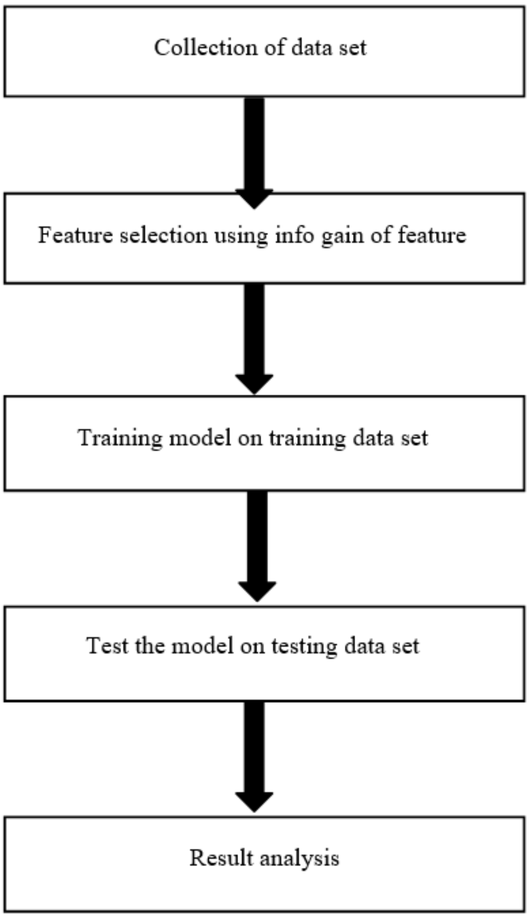
Collection of data set
Feature selection using info gain of feature
Training model on training data set
Test the model on testing data set
Result analysis

Fig-2: Loan Prediction Methodology

IV. EXPLORATORY DATA ANALYSIS

1. 80% applicants are male 20% are female.

2. 80% are not self-employed.

3. 60% are married and 85% have repaid their debts.

4. Most of the Applicants have no dependents.

5. Around 80 % Applicants are graduates.

6. Majority of Applicants are from Semi urban area.

7. Distribution of Applicant income is towards left which means it is not normal distribution. This can be attributed to

Income Disparity in society Driven by the fact that People have different education levels.

8. Proportion of male and female applicants is the same for approved as well as unapproved loans.

9. Proportion of married applicant is more for approved loans.

10. Distribution for applicants having 1 to 3+ dependents is same across both the categories.

11. If co-applicant income is less then less chances of loan approval.

12. More chances of approval for low and average loan amount as compared to high loan amount.

13. the most correlated variables are applicant income and loan amount & credit History and Loan status.

V. MODEL USED FOR TRAINING AND TESTING

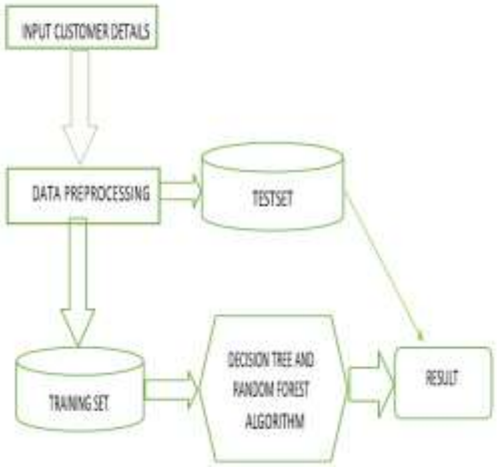

Fig- 3: Training and testing

VI. MACHINE LEARNING METHODS

Three machine learning classification models are used for the prediction of application that can be used in android applications. The brief description of each model is explained below.

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).

Decision tree is a type of supervised learning algorithm having predefined target variable that is mostly used in classification problems. In this technique we spilt the population or sample into two or more homogenous sets based on the most significant splitter/differentiator in the input variables

Decision tree uses multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words the purity of node increases with respect to the target variable.

The accuracy of this algorithm is 77%.

2. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many extensions that are more complex exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). With the confidence factor of c=1.0 the best accuracy is 78.91%

3. 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 80.20%.

VII. CONCLUSION

The main purpose of the paper is to classify and analyse 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 it 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 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.

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

[6] https://machinelearningmastery.com/types-of-classification-in-machine-learning/