Loan Prediction using Decision Tree and Random Forest

Yash Mistry D19DIT081, Kushagra Patel 18DIT053

BTech student, Dept. of IT, Devang Patel Institute of Advance Technology and Research

---***---

Abstract - *In India, the amount of individuals or organization applying for loan gets multiplied once a year. The bank staff have to be compelled to place throughout heaps of labor to analyse or predict whether or not the client pays back the loan quantity or not (defaulter or non-defaulter) within the given time. The aim of this paper is to hunt out the character or background or quality of the shopper that is applying for the loan. We tend to use preliminary information analysis technique to have an effect on the matter of approving or rejecting the loan request or in brief loan prediction. The most focus of this paper is to figure out whether or not the loan given to a particular person or a company shall be approved or not.*

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

1. INTRODUCTION

The term banking area unit typically mentioned as receiving associate degreed protective cash that is deposited by a personal or an entity. It additionally includes disposition cash to folks and businesses that has to be paid back among the given quantity of some time while not failing. Banking could also be a sector that is regulated in most of the countries as a result of it's an important believe determinant the money stability of the country. The prime goal in banking sector is to require an edge their assets in safe hands wherever there area unit less probabilities of failure. these days several banks and money firms approve loan when a nerve-wracking, long and weary method of verification however still there is not any surety whether or not the chosen applier is credible or not or in alternative words if he is able to come back the number with interest among the given time. the aim of the loan area unit typically something supported the client desires. Loans area unit generally divided as open over and close-ended loans. Examples of open-end loans are credit cards and a home equity line of credit (HELOC).

With each interest, the amount owed on a closed-ended loan decreases. After an installment, the amount is decreased.

In other words, it's a legal concept that the creditor can't alter. Closed-ended loans include personal loans, mortgages, car payments, EMIs, and student loans, to name a few.

A secured loan, also known as a collateral loan, is one that is secured by an asset. Houses, automobiles, and other types of property

2. DATA SET

The banking sector provides a range of data. Weka accepts the ARFF (Attribute-Relation File Format) format for the data collection. Tags in an ARFF file include the name, types of attributes, values, and the data itself. We are using 12 attributes in this article, such as gender, legal status, qualification, income, and so on.

The information collection that we used is shown in the table below:

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, we add the training data set to the machine learning model, and the model is trained with known examples during this data set. The most recent applicants' entries will serve as evaluation data to be filled in when submitting the application. Following the completion of such tests, the model will decide if the loan granted to the individual is safe or not, based on the authorization on the loan.

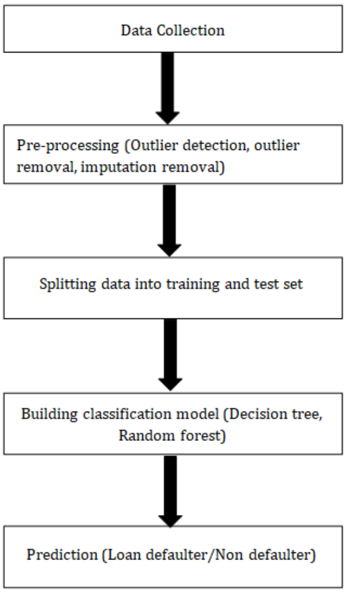

Data Collection
Pre-processing (Outlier detection, outlier removal, imputation removal)
Splitting data into training and test set
Building classification model (Decision tree, Random forest)
Prediction (Loan defaulter/Non defaulter)

Fig-1: Chronology of Data

The diagram above shows how data is used in this machine learning model or method.

It's divided into four parts, each of which uses data to predict the outcome of the entire operation. To begin, we coach our model with a training data set. We evaluate the model with unknown examples from an analogous scenario after it has been conditioned.

Data pre-processing is another procedure we use before analysing and training data. "We have a propensity to exclude all types of values that can trigger a slip-up in information pre-processing, such as redundant values, incomplete values, missing information, and so on."

There are two types of feature selection methods: supervised and unsupervised. Wrapper, filter, and intrinsic are the three sections of the supervised process. We use the goal variable in the supervised approach to eliminate data inconsistencies. The goal variable is not included in the unsupervised approach to eliminate inconsistencies. The correlation mechanism is used in the unsupervised system.

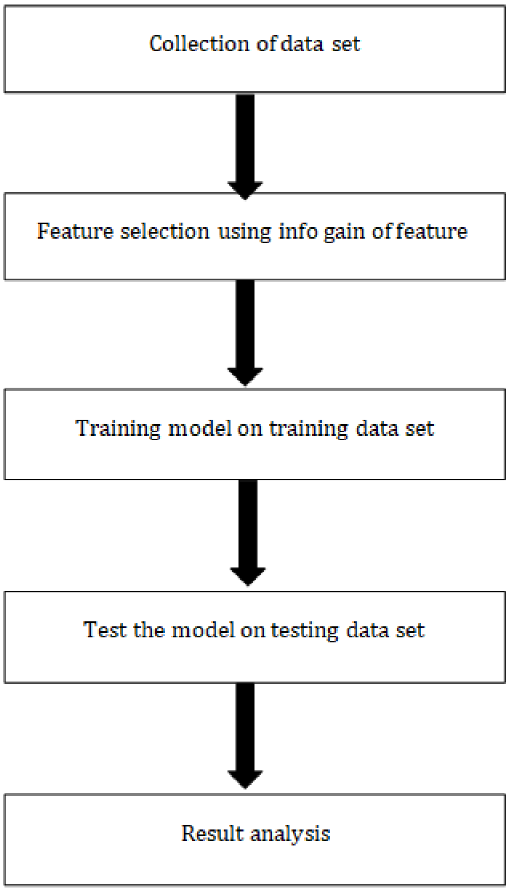

Collection of data set
Feature selection using info gain of feature
Training model on training data set
Test the model on testing data set
Result analysis

Fig-2: Loan Prediction Methodology

3. EXPLORATORY DATA ANALYSIS

1. The person with the higher salary has a better chance of being approved.
2. Those who have completed their education have a higher chance of being approved.
3. For permission, married people would have a stronger hand than unmarried people.
4. The person with the smallest number of dependents has a high chance of being approved.
5. The higher the chances of getting a loan, the smaller the loan sum.

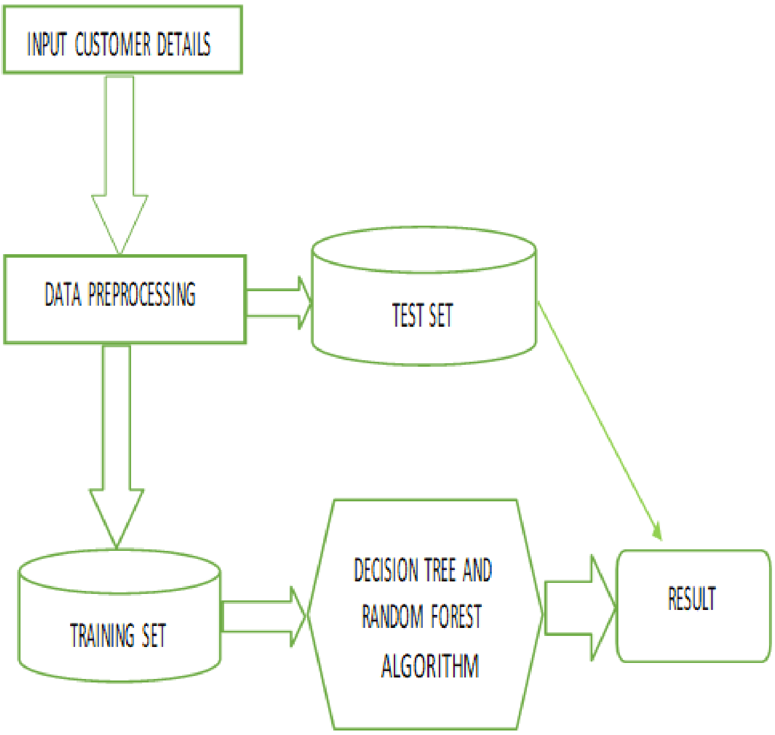


Fig- 4: Training and testing model

7. MACHINE LEARNING METHODS

For the estimation of applications that will be used in Android applications, two machine learning classification models are used. The open source software R, which is licenced under the GNU GPL, can also be used to access these models. The following is a brief summary of each model.

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., if a coin toss comes up heads or tails), each branch represents the outcome of the test, and every leaf node represents a category mark (decision made after all attributes have been computed). C4.5 classification algorithms are extended in this model. "We played with the J48 call Tree classifier, which is a C4.5 call Tree implementation." The lower the arrogance factor, the more pruning is completed in this classifier. For this, we used various confidence factors and analysed them with higher confidence factors, with the accuracy increasing in each case as the confidence factor increased. The easiest precision is 78 percent with a confidence factor of 0.77. It means that as the amount of pruning done decreases, the accuracy increases.

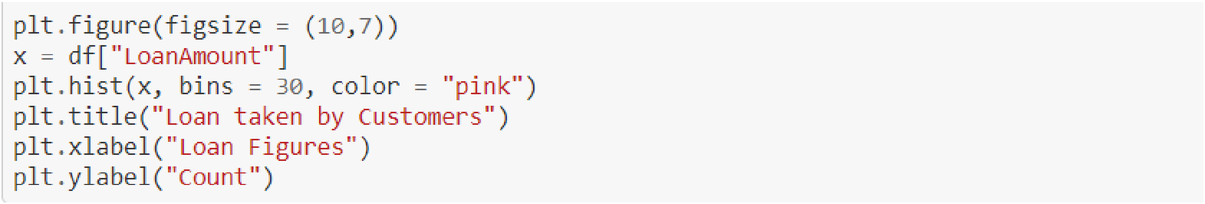
7.2 Random forest

Random forest or random call forests are an ensemble learning technique for classification, regression, and alternative tasks that works by building a large number of call trees at training time and then outputting the class that is the mode of the categories or mean prediction of the individual trees. The ultimate call is created supported the bulk of the trees and is chosen by the random forest.

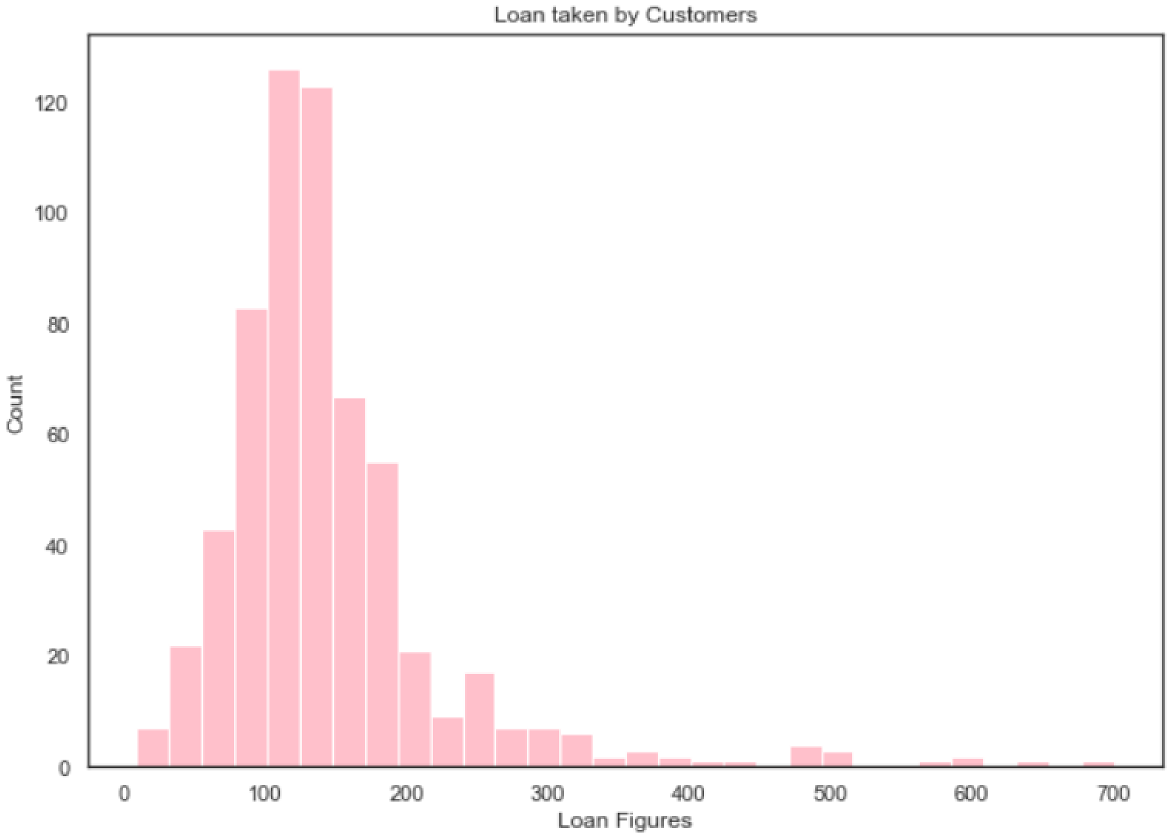
We conducted several trials using Random Forest, including executions with supervised and unsupervised discretization's (equal-frequency and equal-width), and with all attributes. The simplest result in the experiments without attribute selection was 80.50 percent.

8. IMPLEMENTATION

8.1 HISTOGRAM REPRESENTATION

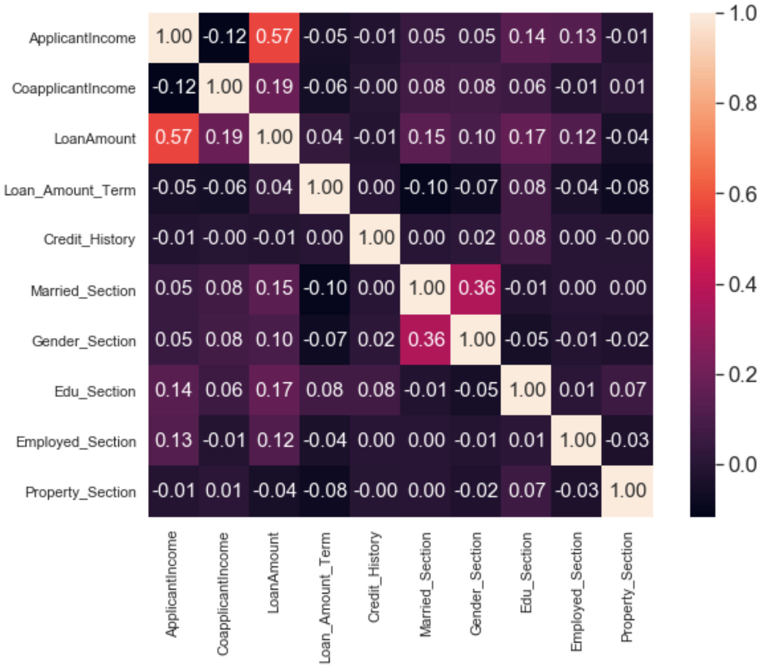


```
plt.figure(figsize = (10,7))
x = df["LoanAmount"]
plt.hist(x, bins = 30, color = "pink")
plt.title("Loan taken by Customers")
plt.xlabel("Loan Figures")
plt.ylabel("Count")
```



8.2 CORRELATION MATRIX

A matrix may be a table showing correlation coefficients between variables. every cell within the table shows the correlation between 2 variables. A matrix is employed to summarize information, as AN input into a a lot of advanced analysis, and as a diagnostic for advanced analyses.

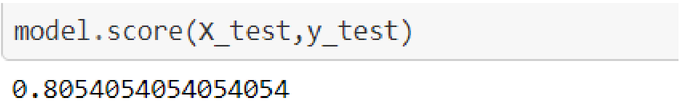


	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Married_Section	Gender_Section	Edu_Section	Employed_Section	Property_Section
ApplicantIncome	1.00	-0.12	0.57	-0.05	-0.01	0.05	0.05	0.14	0.13	-0.01
CoapplicantIncome	-0.12	1.00	0.19	-0.06	-0.00	0.08	0.08	0.06	-0.01	0.01
LoanAmount	0.57	0.19	1.00	0.04	-0.01	0.15	0.10	0.17	0.12	-0.04
Loan_Amount_Term	-0.05	-0.06	0.04	1.00	0.00	-0.10	-0.07	0.08	-0.04	-0.08
Credit_History	-0.01	-0.00	-0.01	0.00	1.00	0.00	0.02	0.08	0.00	-0.00
Married_Section	0.05	0.08	0.15	-0.10	0.00	1.00	0.36	-0.01	0.00	0.00
Gender_Section	0.05	0.08	0.10	-0.07	0.02	0.36	1.00	-0.05	-0.01	-0.02
Edu_Section	0.14	0.06	0.17	0.08	0.08	-0.01	-0.05	1.00	0.01	0.07
Employed_Section	0.13	-0.01	0.12	-0.04	0.00	0.00	-0.01	0.01	1.00	-0.03
Property_Section	-0.01	0.01	-0.04	-0.08	-0.00	0.00	-0.02	0.07	-0.03	1.00

Fig- 4: Correlation Matrix

8.3 RANDOM FOREST IMPLEMENTATION

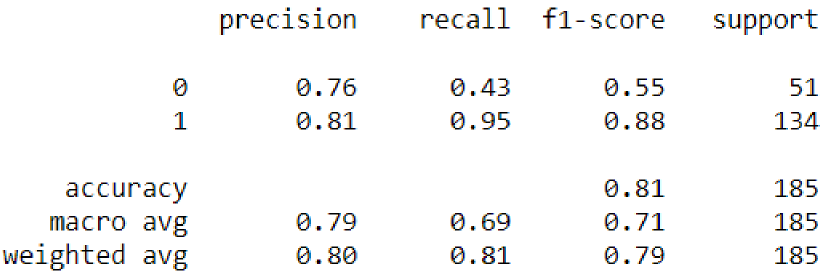
Accuracy in the testing model.



```
model.score(X_test,y_test)
```

```
0.8054054054054054
```

Report



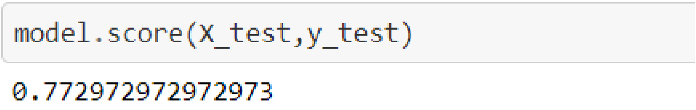
```
              precision    recall  f1-score   support

           0       0.76      0.43      0.55        51
           1       0.81      0.95      0.88       134

    accuracy                           0.81       185
   macro avg       0.79      0.69      0.71       185
weighted avg       0.80      0.81      0.79       185
```

8.4 DECISION TREE IMPLEMENTATION

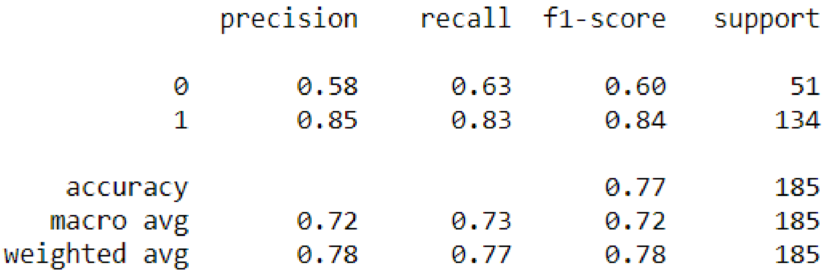
Accuracy in the testing model.



```
model.score(X_test,y_test)

0.772972972972973
```

Report



```
              precision    recall  f1-score   support

           0       0.58      0.63      0.60        51
           1       0.85      0.83      0.84       134

    accuracy                           0.77       185
   macro avg       0.72      0.73      0.72       185
weighted avg       0.78      0.77      0.78       185
```

9. CONCLUSIONS

The paper's main goal is to identify and evaluate the loan candidates' personalities. Based on a thorough examination of available data and banking sector constraints, it is possible to conclude that, with safety in mind, this product is highly efficient or cost-effective. This application performs well and meets all of Banker's primary requirements. Despite the fact that the application is adaptable to a variety of systems and can be effectively blocked.

This paper work could be expanded to a higher level in the future, and the software package could be updated to make it more accurate, stable, and right. As a result, the system is trained with gift data sets that might become older in the future, allowing it to engage in new testing to be created, such as passing new test cases.

There have been several instances of laptop malfunctions, content failures, and the most important weight of choices is anchored in a machine-driven prediction framework. So, in the not-too-distant future, a software package for safer, more efficient, and dynamic weight adjustment may be created. This prediction module could be combined with the machine-driven process system module in the near future.

REFERENCES

[1] https://www.myperfectwords.com/blog/research-paper-guide/research-paper-example

[2] https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012042/pdf

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

[4] https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees