Predication Approval for Bank Loan Using Random Forest Algorithm

Rutvik Vanara [1], Piyush Wani[1], Sagar Pawar[1], Punitkumar More[1], Priyanka Patil[1]

[1]Student, Department of Computer Engineering, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, India.

SSBT's Collage of Engineering and Technology, Bambhori, Jalgaon, Maharashtra, India.

Corresponding Author: rutvikvanara123@gmail.com

Abstract: - Banking Industry always needs a more accurate predictive modeling system for many issues. Predicting credit defaulters is a difficult task for the banking industry. The loan status is one of the quality indicators of the loan. It does not show everything immediately, but it is a first step of the loan lending process. The loan status is used for creating a credit-scoring model. The credit-scoring model is used for accurate analysis of credit data to find defaulters and valid customers. The objective of this is to create a credit-scoring model for credit data. Various machine-learning techniques are used to develop the financial credit-scoring model. For this classification we use the 'Random Forest Algorithm'. This proposed provides the important information with the highest accuracy. It is used to predict the loan status in commercial banks using machine-learning classifier.

Key Words:— ***Loan Prediction, ML, Random Forest Algorithm.***

I. INTRODUCTION

Now a day, Banks struggle a lot to get an upper hand over each other to enhance overall business due to tight competition. Banks have realized that retaining the customers and preventing fraud must be the strategy tool for healthy competition. Availability of the huge quantity of data, creation of knowledge base and efficient utilization of the same have helped banks to open up efficient delivery channels. Business decisions can be optimized through data mining. Customer segmentation, banking profitability, credit scoring and approval, predicting payment from customers, marketing, detecting fraud transactions, cash management and forecasting operations, optimizing stock portfolios and ranking investments are some of the areas where data mining techniques can be used in the banking industry.

Credit risks, which account for the risk of loss and loan defaults, are the major source of risk encountered by banking industry. Data mining techniques like classification and prediction can be applied to overcome this to a great extent. There are mainly two objectives that is to be achieved through these techniques. These are as follows,

- Identification of the relevant attributes that signal the capacity of borrowers to pay back the loan, and
- Determining the best model(s) to evaluate credit risk. Random Forest Algorithm is one of the best techniques to achieve this objective. The model thus developed will provide a better credit risk assessment, which will potentially lead to a better allocation of the bank's capital.

In this regard, a study is conducted and an efficient prediction model, which helps to reduce the proportion of unsafe borrowers, is introduced herewith. Due to the significance of credit risk analysis, this study helps banking industry by providing additional information to the loan decision making process, potentially decreases the cost and time of loan applications appraisal, and decreases the level of uncertainty for loan officers by providing knowledge extracted from previous loans. Random Forest Induction Algorithm used in this model is the data mining technique for predicting credible customers.

Data Set: -A collection of data is taken from the banking sector. The Data set is in csv format. CSV 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,

Manuscript revised July 26, 2021; accepted July 27, 2021.
Date of publication July 29, 2021.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898

qualification, income, etc. The table below represents the data set that we have used:

Table-1: Data set Attributes for Loan

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/ no Graduate	String
Self-employed	Self-employed(Y/N)	Character
Applicant Income	Applicant income	Integer
CoapplicantIncome	Co-applicant income	Integer
Loan Amount	Loan amount	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	Yes/No	Integer

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.

II. BACKGROUND

Prediction of loan status is the process of identifying whether the customers can pay the loan amount or not. Machine learning provides various algorithm for the prediction. In this System, we are using the Random Forest algorithm for prediction. This algorithm helps to predict whether the customer is defaulter or not on the basis of classifying the attributes of the customer.

III. METHODOLOGY

The diagram represents the working of our model. It gives us an 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.

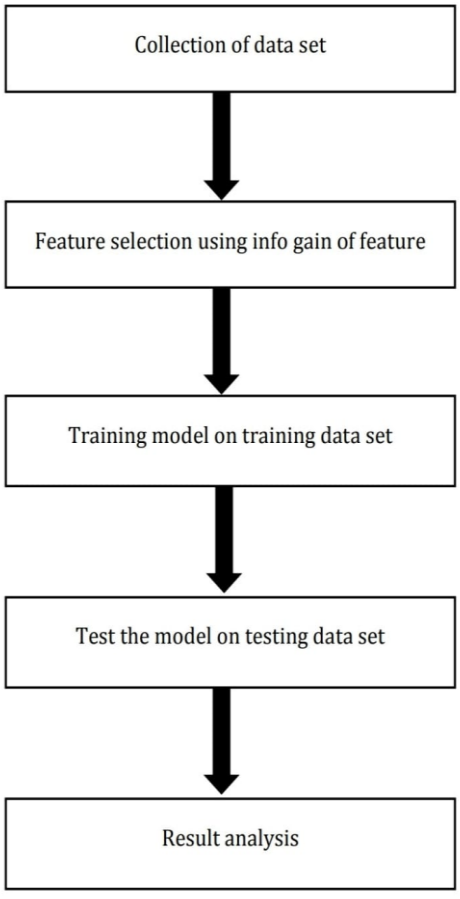

Fig.1. Loan Prediction Methodology

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.

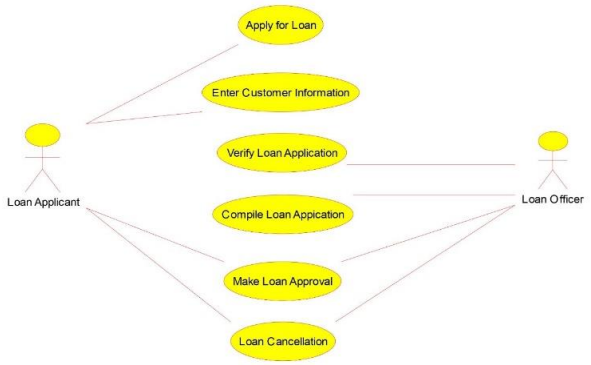

Fig.2. Use case diagram

Table.2. Use case diagram variable and description

Actor	Loan Applicant, Loan Officer
Description	An applicant requests for a loan. After request loan Machine Learning verified its Information and loan evaluator may approve or reject the loan.
Data	Applicant personal information and its documents.
Response	Loan may be approved or may be rejected.

The proposed model focuses on predicting the credibility of customers for loan repayment by analyzing their behavior. The input to the model is the customer behavior collected. Based on the output from the classifier, decision on whether to approve or reject the customer request can be made. Random Forest Algorithm is used to generate the relevant attributes and also make the decision in the model.

Analysis:

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

IV. SYSTEM ARCHITECTURE

The proposed model focuses on predicting the credibility of customers for loan repayment by analyzing their behavior. The input to the model is the customer behavior collected. Based on the output from the classifier, decision on whether to approve or reject the customer request can be made. Random Forest Algorithm is used to generate the relevant attributes and also make the decision in the model.

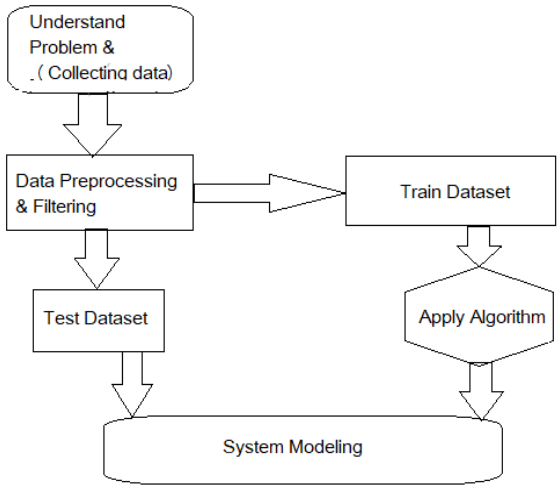

Fig.3. System Architecture of Bank Loan

Problem Understanding: - The data-mining model is initiated with collection of details regarding the banking sector and the existing loan processing procedures. The challenges and the main risks associated with the loan approval/rejection in banking sector are thus better understood.

Data Understanding: - In Data Understanding phase, the bank data set of customer details, which is required for data mining, is collected and got familiarized with. Various attributes needed are also studied.

Data Filtering: - The attributes in the bank data set are filtered and the relevant attributes needed for prediction are selected. After those, the incomplete and noisy records in the data set are removed and prepared for mining.

System Modelling: - In this stage, the system is developed in an efficient and user-friendly manner so that even those users with less technical knowledge can also use it comfortably. The system provides the most relevant attributes that helping determining whether to approve or reject the loan application. This aids in predicting the credibility of future customers. [8]
System Evaluation: - In the final stage, the designed system is tested with test set and the performance is assured.

A. Random Forest Algorithm

An efficient Random Forest is formulated with Random Forest Induction Algorithm. It produces a model with the most relevant 11 attributes. Random Forest with different parameters: executions with supervised and unsupervised discretization (equal-frequency and equal-width), with all attributes. In the experiments without attribute selection, the best result was 85.75 Percentage and it was achieved with unsupervised equal-frequency 5 bins discretization with 450 trees and seed equal to 4. The main benefit of applying Data Mining is that we can always rely on the result of the algorithm to accept or reject the loan application.

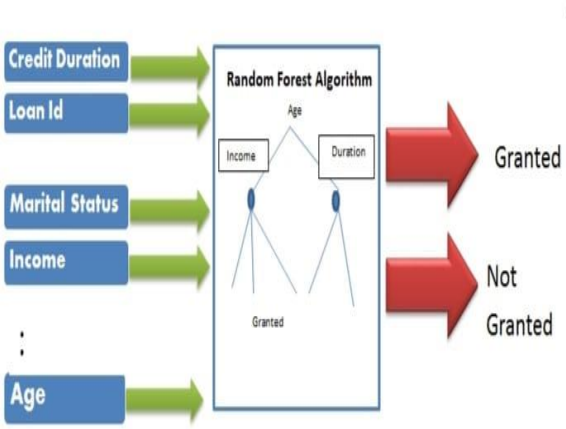

Fig.4. Random Forest Algorithm

Gini (D) = 1 − Xn i=1 p 2

Gini split (D) = |D1| Gini(D1) + |D2| Gini(D2).
------ ------
|D| |D|

Accuracy is 77.27272727272727

V. TESTING

Testing is the process of evaluating a system or its component with the intent to find whether it satisfied requirements or not. Testing is executing a system in order to identify any gaps, errors, or missing requirements in contrary to the actual requirement.
White Box Testing: White box Testing strategy deals with the internal logic and structure of the code. White box testing is also called as glass, structural, open box or clear box testing. The tests written based on the white box testing strategy incorporate coverage of the code written, branches, paths, statements and internal logic of the code etc. In order to implement white box testing, the tester has to deal with the code and hence is needed to possess knowledge of coding and logic i.e., internal working of the code. White box test also needs the tester to look into the code and find out which unit/statement/chunk of the code is malfunctioning.
Black Box Testing: Black Box Testing is not a type of testing; it instead is a testing strategy, which does not need any knowledge of internal design or code etc. As the name" black box" suggests, no knowledge of internal logic or code structure is required. The types of testing under this strategy are totally based/focused on the testing for requirements and functionality of the work product/software application. The base of the Black box testing strategy lies in the selection of appropriate data as per functionality and testing it against the functional specifications in order to check for normal and abnormal behavior of the system.
Manual Testing: It is the oldest and rigorous type of testing. Human sitting in front of computer carefully going through application screens, typing various usage, performs it and input combinations, comparing the results to thee expected behavior and recording their observations about project. There are certain ways of manual testing first test cases are return then they are executed and then report is generated according to test case result. ˜cite1

Table.3. Test Cases

Test Case	**Test Case Expected Result**	**Test Case Actual Result**	**Test Case Pass / Fail**
Prediction	Applicant Information Tested with	Matched the Information with Patterns	PASS

	Patterns of Dataset	of Trained Dataset	
Prediction	Applicant Information Tested with Patterns of Dataset	Information does not match with Pattern of Trained Dataset	FAIL

Table.4. Test Cases for Bank Loan Prediction

Loan_ID	Gender	Married	Dependents	Education
LP001002	Male	No	0	Graduate
LP001003	Male	Yes	1	Graduate
LP001005	Male	Yes	0	Graduate
LP001006	Male	Yes	0	Not Graduate
LP001008	Male	No	0	Graduate
LP001011	Female	Yes	2	Graduate
LP001013	Male	Yes	0	Not Graduate
LP001014	Male	Yes	3	Graduate
LP001018	Male	Yes	2	Graduate
LP001020	Male	Yes	1	Graduate

Table.5. Test Cases for Bank Loan Prediction

Self-employed	Applicant Income	CoapplicantIncome	Loan Amount	Loan_Amount_Term
No	5849	0	523	360
No	4583	1508	128	360
Yes	3000	0	66	360
No	2583	2358	120	360
No	6000	0	141	360
Yes	5417	4196	267	360
No	2333	1516	95	360
No	3036	2504	158	360
No	4006	1526	168	360
No	12841	10968	349	360

Table.6. Test Cases for Bank Loan Prediction

Credit History	Property Area	Loan Status	Test P/F
1	Urban	Yes	Pass
1	Rural	No	Pass
1	Urban	Yes	Pass
1	Urban	Yes	Pass
1	Urban	Yes	Pass
1	Urban	Yes	Pass
0	Urban	Yes	Fail
0	Semiurban	No	Pass
1	Urban	Yes	Pass
1	Semiurban	No	Pass

A. Model Evaluate

Accuracy: Accuracy of the model has been measured by predefined metrics. In a balance class model shows high accuracy but in the case of unbalanced class the accuracy is very less.

Precision: Percentage ratio of positive instances and total predicted positive instances gives precision value. In the below equation denominator represents the model positive prediction done from the completely given dataset. Precision value tells the perfectness of our model. In our data set, good precision value has been obtained.

Precision Score = TP / (FP + TP)

```
Precision: 0.765625
```

Recall: Percentage ratio of positive instances with actual total positive instances is recall value. Here denominator (TP + FN) shows the total number of positive instances which are present in whole dataset. As a result, it has obtained 'how much extra right ones, the model will be failed if it shows maximum right ones. .

Recall Score = TP / (FN + TP)

```
Recall:  0.98
```

F1 Score: The harmonic mean (HM) of precision and recall values is called F1 Score. Model will be best performer if it shows maximum F1 Score. Numerator shows the product of precision and recall if one goes low either precision or recall, the final F1 score goes down significantly. So, a model does well in F1 score if the positive predicted (precision) having positive value and doesn't miss out on positives and predicts them negative (recall).

F1 Score = 2 *Precision Score* Recall Score/ (Precision Score + Recall Score/)

```
F1 Score: 0.8596491228070174
```

VI. CONCLUSION

The loan status prediction system that helps the organizations in making the right decision to approve or reject the loan request of the customers. This will definitely help the banking industry to open up efficient delivery channels. Random Forest Induction Algorithm is used for the prediction. Further, the comparison study has been made with different levels of iterations. This model can be used to avoid the huge loss of commercial banks.

REFERENCES

[1]. Hidayti, Ery.(2003). Sistem Pendukunug Keputusan Berbasis Logika Fuzzy Untuk Analisis Kelaykan Kredit. Fakultas Matematika Dan Iimu Pengetahuan Alam: Institut Tenknologi Sepuluhnopembe.

[2]. Rafik Khairul Amin, Yuliant Sibaroni.(2015).Implementation of Decision Tree Using C4.5 Using algorithm in decision making of loan aaplication by debartor (Case study :Bank Pasar of Yogyakarta Special Region) 3rd International Conference on Information and Communication Technology (ICOICT).

[3]. Santosa, Budi. (2007) Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis .Graha llMU:Yogyakarta.

[4]. Dileep B. Desai, Dr. R.V.Kulkarni A Review: Application of Data Mining Tools in CRM for Selected Banks, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4, 2013, 199 201

[5]. Dr. K. Chitra1, B. Subashini , Aug(2013) Data Mining Techniques and its Applications in Banking Sector , International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8.Dr. Madan Lal Bhasin, Data Mining:

[6]. Kazi Imran Moin, Dr. Qazi Baseer Ahmed Use of data minning in Banking, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, vol. 2, Issue 2, Mar-Apr 2012, pp.738-742 738.

[7]. Vivek Bhambri Application of Data Mining in Banking Sector, International Journal of Computer Science and Technology Vol. 2, Issue 2, pp.342-344 June 2011.

[8]. Proposals for Banks Using Data Mining, International Journal of Latest Research in Science and Technology, pp. 126-131, July 2014.

[9]. Nikhil Madane,Siddhart Nanda,December (2019),"Loan Predication Using Decision Tree ", Journal of the Gujarat Research History, Volume 21 Issue 14's,December 2019.

[10]. Andy Liaw, Matthew Wiener, November (2002), Classficaton and Regression by RandomForest.R news.

[11]. J.M Chambers, Octomber, Computational method for data analysis. Applied Statistics, Wiley, 1(2):1-10-1077.

[12]. Kusrini, Hadi,(2013),Data Mining. ANDI:Yogyakarta.