International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process

E. Chandra Blessie, R. Rekha

Abstract*: Extending credits to corporates and individuals for the smooth functioning of growing economies like India is inevitable. As increasing number of customers apply for loans in the banks and non- banking financial companies (NBFC), it is really challenging for banks and NBFCs with limited capital to device a standard resolution and safe procedure to lend money to its borrowers for their financial needs. In addition, in recent times NBFC inventories have suffered a significant downfall in terms of the stock price. It has contributed to a contagion that has also spread to other financial stocks, adversely affecting the benchmark in recent times. In this paper, an attempt is made to condense the risk involved in selecting the suitable person who could repay the loan on time thereby keeping the bank's non-performing assets (NPA) on the hold. This is achieved by feeding the past records of the customer who acquired loans from the bank into a trained machine learning model which could yield an accurate result. The prime focus of the paper is to determine whether or not it will be safe to allocate the loan to a particular person. This paper has the following sections (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.

***Keywords* : Loan Prediction, Big data, Machine Learning, Logistic Regression, SVM, Decision Tree, Naïve Bayes.**

I. INTRODUCTION

Finance raising and lending for real estate, consumer, mortgage and companies' loans is the central part of almost every bank's business model. Lending money to inappropriate customers forms the major source of credit risk. The major share of the bank's assets comes directly from the profit derived from the bank's loans. The banking companies' face, however dual challenge to distinguish the possible deliberate defaulters from the applicants and the biased nature of few bank employees who have been at the instigation of developers of defaulting companies for many years. The primary goal of the banking community is to safely invest their capital. In the current scenario, many NBFCs and banks approve loans after a clear verification and authentication process, however, it remains uncertain whether the candidate selected is the worthy correct of all the applicants. Through this method, we can predict whether or not that particular applicant is secure and the machine learning technique automates the entire process of authentication. The major disadvantage of this model lies in the fact that more importance is given in assigning weightages to each factor, but, in real-time a loan can be sanctioned solely on the basis of a single strong factor, which is not feasible through this method. This paper seeks to ensure that the deserving customers can be quickly selected with ease which offers various benefits to the bank itself. This method will measure the weight automatically of each criterion that participates in the loan processing and process the same with regards to the associated weight of the new test data. The borrower can set a time to check whether or not the loan is sanctioned. This system can skip the sequential verification process and could jump to a specific point to be verified on the basis of priority. This system of loan prediction is solely for the bank officials and completely foolproof that the private players and investors can not alter the data. Results on a particular loan ID may be sent to various bank departments to take adequate action and carry out other formalities.

II. PROCEDURE FOR PAPER SUBMISSION

Evaluating the risk associated with a loan application is one of the most important concerns for the sustainability and profitability of the highly competitive market. On a daily basis, these banks receive multiple applications for loans from their clients and others. Not all of them are sanctioned. Many banks use their own credit scoring and risk analysis methods to review the request for loan and make credit approval decisions. Nonetheless, there are many instances occurring each year where borrowers do not repay the sums of the loan and they default, as a result of which these financial institutions incur tremendous losses.

The following section depicts the existing loan prediction methods.

Aboobyda Jafar Hamid and Tarig Mohammed Ahmed [1] presented a loan risk prediction model based on the data mining techniques, such as Decision Tree (J48), Naïve Bayes (NB) and BayseNet approaches. The procedure followed was (1) training set preparation, (2) building the model, (3) Applying the model and finally (4) Evaluating the accuracy.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Dr.E.ChandraBlessie, Department of MCA, Nehru College of Management, Coimbatore, Tamilnadu, India
R.Rekha*, Department of MCA, Nehru College of Management, Coimbatore, Tamilnadu, India, rekhaoct18@gmail.com

This approach was implemented using Weka Tool and considered a dataset with eight attributes, namely, gender, job, age, credit amount, credit history, purpose, housing, and class. Evaluating these models on the dataset, experimental results concluded that, J48 based loan prediction approach resulted in better accuracy than the other methods.

Vimala and Sharmili [2] proposed a loan prediction model using NB and Support Vector Machines (SVM) methods. Naïve Bayes, an independent speculation approach, encompasses probability theory regarding the data classification. On the other hand, SVM uses statistical learning model for classification of predictions. Dataset from UCI repository with 21 attributes was adopted to evaluate the proposed method. Experimentations concluded that, rather than individual performances of classifiers (NB and SVM), the integration of NB and SVM resulted in an efficient classification of loan predictions.

Kacheria, Shivakumar, Sawkar and Gupta [3] suggested a loan sanctioning prediction procedure basded on NB approach integrated with K-Nearest Neighbor (KNN) and binning algorithms. The seven parameters considered were income, age, profession, existing loan with its tenure, amount and approval status. The sub-processes include, (1) Pre-processing (handling the missing values with KNN and data refinement using binning algorithm), (2) Classification using NB approach and (3) Updating the dataset frequently results in appropriate improvement in the loan prediction process. Experimentation put-forth the conclusion that, integration of KNN and binning algorithm with NB resulted in improved prediction of loan sanctioning process.

Jency, Sumathi and Shiva Sri [4] proposed a Exploratory Data Analysis (EDA) regarding the loan prediction procedure based on the client's nature and their requirements. The major factors concentrated during the data analysis were (1) annual income versus loan purpose, (2) customer's trust, (3) loan tenure versus delinquent months, (4) loan tenure versus credit category, (5) loan tenure versus number of years in the current job, and (6) chances for loan repayment versus the house ownership. Finally, the outcome of the present work was to infer the constraints on the customer who are applying for the loan followed by the prediction regarding the repayment. Further, results showed that, the customers were interested more on availing short-tenure loans rather than long-tenure loans.

Goyal and Kaur [5] suggested an ensemble technique based loan prediction procedure for the customers. The sub-processes in the present method includes, (1) data collection, (2) filtering the data, (3) feature extraction, (4) applying the model, and finally (5) analysis the results. The various loan prediction procedures implemented in the present method were Random Forest (RF), SVM and Tree model with Genetic Algorithm (TGA). The parameters considered for evaluating the models were (1) accuracy, (2) Gini Coefficient, (3) Area Under Curve (AUC), (4) Receiver Operating Curve (ROC), (5) Kolmogorov - Smirnov (KS) Chart, (6) Minimum Cost - Weighted Error Rate, (7) Minimum Error Rate, and (8) K-Fold Cross Validation parameters. Experimentation outcome concluded that the integration of three methods (RF, SVM and TGA) resulted in improved loan - prediction results rather than individual method's prediction.

Goyal and Kaur [6] presented a loan prediction model using several Machine Learning (ML) algorithms. The dataset with features, namely, gender, marital status, education, number of dependents, employment status, income, co-applicant's income, loan amount, loan tenure, credit history, existing loan status, and property area, are used for determining the loan eligibility regarding the loan sanctioning process. Various ML models adopted in the present method includes, (1) Linear model, (2) Decision Tree (DT), (3) Neural Network (NN), (4) Random Forest (RF), (5) SVM, (6) Extreme learning machines, (7) Model tree, (8) Multivariate Adaptive Regression Splines, (9) Bagged Cart Model, (10), NB and (11) TGA. When evaluated these models using R Environment in five runs, TGA resulted in better loan forecasting performance than the other methods.

Sudhamathy [7] suggested a risk analysis method in sanctioning a loan for the customers using R package. The various modules include data selection, pre-processing, feature extraction and selection, building the model, prediction followed by the evaluation. The dataset used for evaluation in this method was adopted from UCI repository. To fine tune the prediction accuracy, the pre-processing operation includes the following sub-processes: detection, ranking and removal of outliers, removal of imputation, and balancing of dataset by proportional bifurcation regarding testing and training process. Further, feature selection process improves the prediction accuracy. When evaluated, the DT model resulted in 94.3% prediction accuracy.

Supriya, Pavani, Saisushma, Vimala Kumari and Vikas [8] presented a ML based loan prediction model. The modules in the present approach were data collection and pre-processing, applying the ML models, training followed by testing the data. During the pre-processing stage, the detection and removal of outliers and imputation removal processing were carried out. In the present method, SVM, DT, KNN and gradient boosting models were employed to predict the possibilities of current status regarding the loan approval process. The conventional 80:20 rule was adopted to split the dataset into training and testing processes. Experimentation concluded that, DT has significantly higher loan prediction accuracy than the other models.

Arun, Ishan and Sanmeet [9] suggested a load prediction procedure using ML models. The sub-processes include data collection, feature selection, training, testing and analyzing the performance of the present model. The dataset with 10 features were employed for observation and loan prediction process. Various ML approaches used in the present method includes LM, DT, RF, SVM, NN and Adaboost methods. Further, authors suggested few significant parameters that plays a major role in loan prediction process for various ML models, such that, it helps to bankers in approval of loans to the customers based on their requirements.

III. PROPOSED METHOD

The architecture of the proposed model is shown in flow chart Fig.1.

The major objective of this project is to derive patterns from the datasets which are used for the loan sanctioning process and create a model based on the patterns derived in the previous step. Classification data mining algorithms are used to filter out the probable loan defaulters from the list. For analysis purposes, essential inputs like gender, age, marital status, residential status, job, income, loan expectation, existing client, account balance, total debt, etc., are collected and used to find the appropriate attributes.

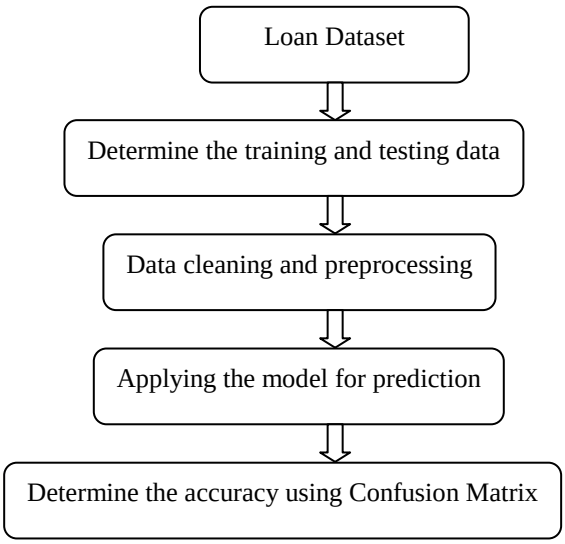


Fig. 1.Architecture of the Proposed Loan Prediction Model

In the present research work, four models are implemented for loan predictions. They are, (1) Logistic Regression (LR), (2) Decision Tree (DT) (3) Support Vector Machines (SVM) and (4) Naïve Bayes (NB) method. The following section deals with the brief description of loan prediction algorithms used in the proposed method.

A. LR Model

In general, linear Regression model was used to predict the functionalities of a continuous variable, say for example "Y". If the variable "Y" is categorical, instead of continuous, then the LR method is adopted. The output of LR model is dichotomous i.e., binary possibilities, used for prediction of loan sanction possibilities. Properties of LR include (1) dependent variable(s) follows Bernoulli distribution and (2) Maximum likelihood is used for estimation. Further, the function f(g) is a logistic function, referred as Sigmoidal function.

B. Decision Tree (DT) Model

DT (Shown in Fig.2) is a supervised learning algorithm used to solve classification and regression problems too. Here, DT uses tree representation to solve the prediction problem, i.e., external node and leaf node in a tree represents attribute and class labels respectively. The pseudo code for DT model is depicted in the following section:

Step 1: Best attribute is chosen as the tree's root.

Step 2: Training set is divided into subsets, such that, each subset comprises similar value for an attribute.

Step 3: Step 1 and Step 2 are repeated for all subsets until all the leaf nodes are traversed in a tree.

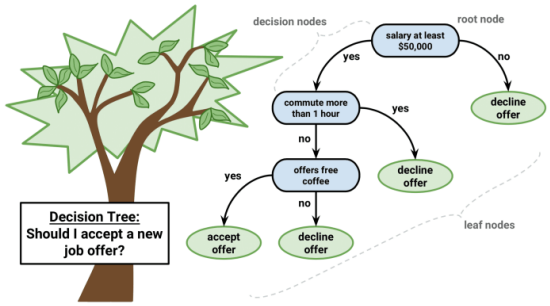


Fig. 2.shows the example in implementation of DT based approach regarding the new job offer processing the depicted form [10].

C. SVM

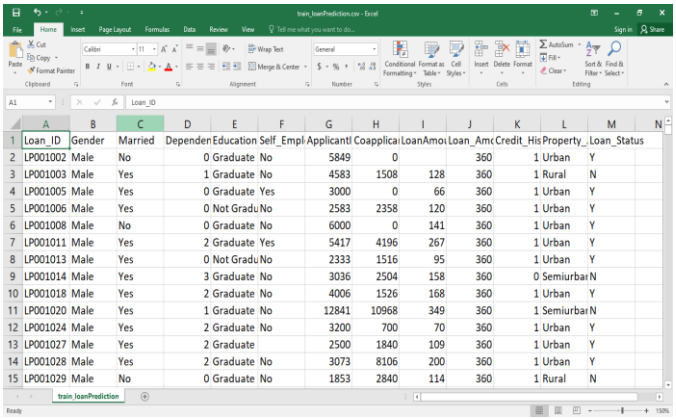
In this approach, each data item is plotted in a n-dimensional space, where 'n' represents the number of features with each feature represented in a corresponding co-ordinates. A hyperplane is determined to distinguish the classes (possibly two) based on their features.

D. Naïve Bayes (NB) Model

The basis for NB model is Bayes Theorem (BT), where events are mutually exclusive similar to rolling a die. Moreover, the BT presumes that the input features also referred as predictors are independent in nature. Similarly, NB also presumes that the input features are independent in nature. But, this is impossible in the realistic procedures. Since this assumption leads to naïve, this algorithm is termed as Naïve Bayes algorithm. Thus, NB is a probabilistic algorithm, where the conditional probability is determined regarding the input features. On the other hand, during the dependent input features scenario, conditional probability is calculated twice resulting in improper results. Hence, for better prediction results with respect to NB model, independent input features are selected and processed [11]. The following shows the pseudo code for the proposed loan prediction method.

1. Load the data.
2. Determine the training and testing data.
3. Data cleaning and preprocessing.
   a) Fill the missing values with mean values regarding numerical values.
   b) Fill the missing values with mode values regarding categorical variables.
   c) Outlier treatment.
4. Apply the modeling for prediction.
   a) Removing the load identifier.
   b) Create the target variable (based on the requirement). In this approach, target variable is loan-status.
   c) Create a dummy variable for categorical variable (if required) and split the training and testing data for validation.
   d) Apply the model
      - LR method
      - DT method
      - RF method
      - SVM method
5. Determine the accuracy followed by confusion matrix.

IV. RESULTS AND DISCUSSIONS



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	CoapplicaI	LoanAmou	Loan_Amo	Credit_His	Property_	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Gradu	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Gradu	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3	Graduate	No	3036	2504	158	360	0	Semiurbar	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurbar	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N

Fig. 3.Input Loan Prediction Dataset from Kaggle [12]

Figure 3 shows the details of dataset collected from Kaggle source [12]. The feature in the dataset includes;

1. Loan_Id
2. Gender
3. Marital Status
4. Number of dependents
5. Educational Profile
6. Employment Status
7. Applicant's Income
8. Co-Applicant's Income
9. Loan Amount
10. Loan Tenure
11. Credit History
12. Property Area
13. Loan Status

A. Exploratory Data Analysis (EDA)

At the start, the dataset was cleaned. Then exploratory data analysis and feature engineering were performed. Then a model was created which predicted whether the applicant would repay the loan or not. Whenever the bank makes decision to give loan to any customers then it automatically exposes itself to several financial risks. It is necessary for the bank to be aware of the clients applying for the loan. This problem motivates to do an EDA on the given dataset and thus analyzing the nature of the customer. The dataset that uses EDA undergoes the process of normalization, missing value treatment, choosing essential columns using filtering, deriving new columns, identifying the target variables and visualizing the data in the graphical format. Python is used for easy and efficient processing of data. This paper used the pandas library available in Python to process and extract information from the given dataset. The processed data is converted into appropriate graphs for better visualization of the results and for better understanding. For obtaining the graph Matplot library is used. The Figs 4a to 4i are shown the graphical presentation of different data set.

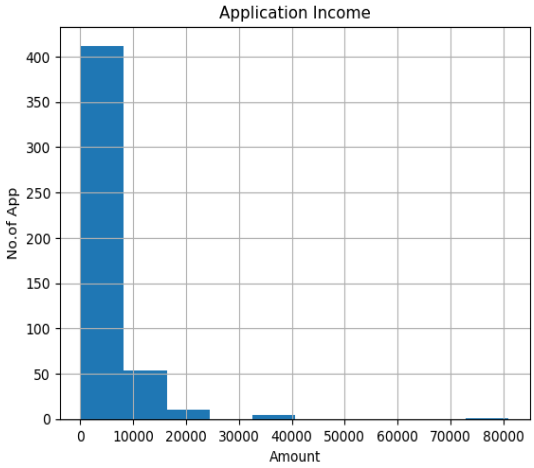


Fig. 4.a. Application Income

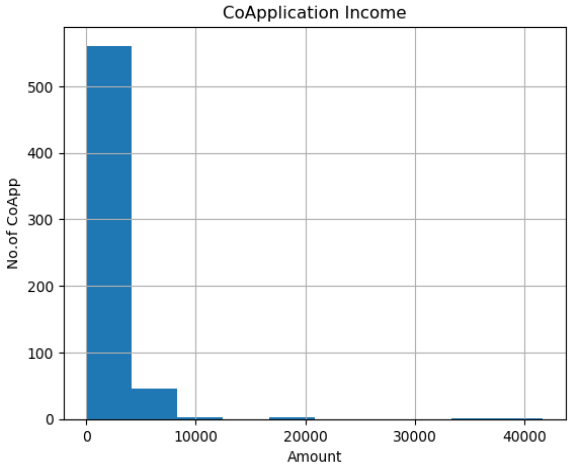


Fig. 4b. Co-Application Income

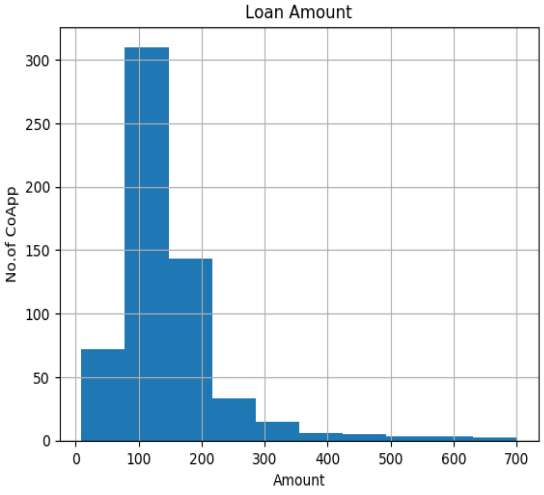


Fig 4c. Loan Amount

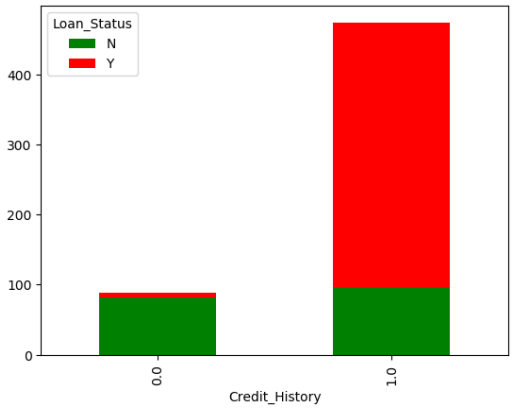


Fig.4d Credit History

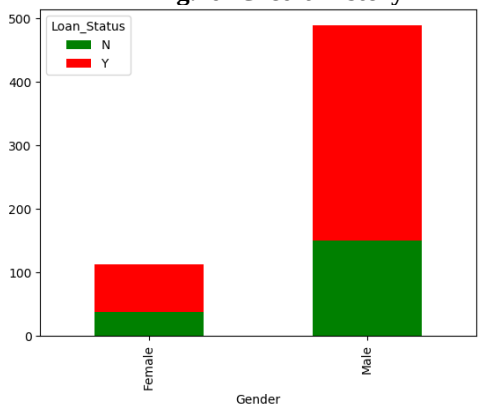


Fig. 4e. Gender Loan Status

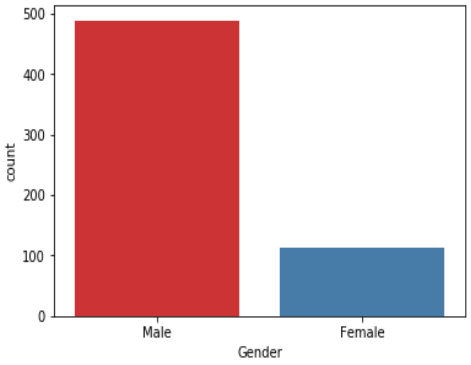


Fig.4f Graph for Gender

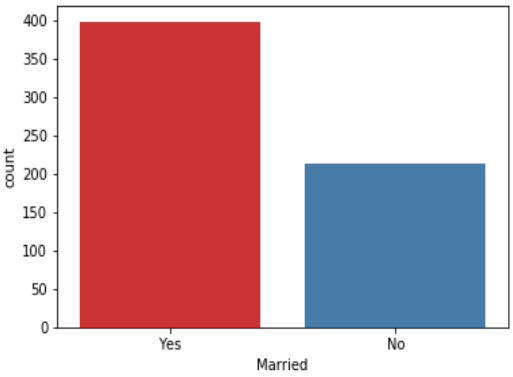


Fig .4g Graph for Relation Status

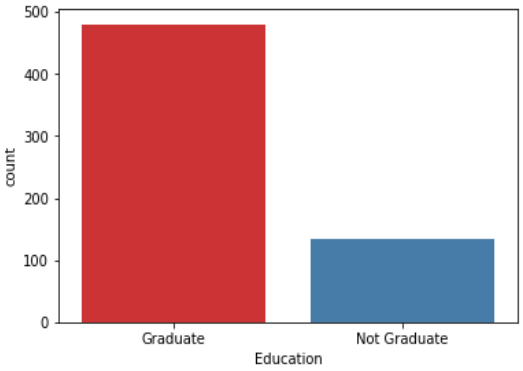


Fig.4h Graph for Graduate and not graduate

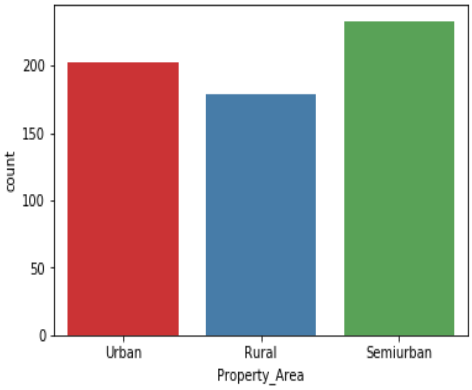


Fig.4i. Graph for Rural or Urban or Semi urban

As represented in the pseudo code of the proposed loan prediction model, after loading the data from the dataset, the training and testing data are determined followed by the data cleaning and preprocessing procedures. Later the prediction models are applied and the performance is determined using the Confusion Matrix (CM) method.

The Confusion Matrix (CM) is used to analyze and determine the performance of the proposed loan prediction model (Shown in Table I and II). Figure 5 shows the CM parameters summarized from [13-14]. The interpretation in the CM is as follows:

- True Positive (TP), when both the actual and predicted values are positive (1)
- True Negative (TN), when both the actual and predicted values are negative (0)
- False Positive (FP), when the actual value is negative and the predicted value is positive (1)
- False Negative (FN), when the actual value is positive (1) and the predicted value is negative (0)

Table I Predicted Value

		Predicted Values	
		Negative (0)	***Positive (1)***
Actual Values	Negative (0)	TN	FP
	Positive (1)	FN	TP

Table II Performance Comparison

Parameter	Loan Prediction Models			
	LR	*DT*	*SVM*	*NB*
Loan Prediction Accuracy	78.91	71.92	65.27	80.42

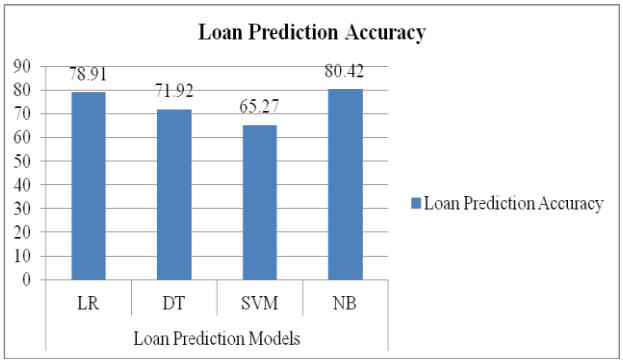


Fig.5. Comparison of Prediction Accuracy

V. CONCLUSION

By properly analyzing positive qualities and constraints, it can be concluded with confidence that the Naïve Bayes model is extremely efficient and gives a better result when compared to other models. It works correctly and fulfills all requirements of bankers and can be connected to many other systems. There were multiple malfunctions in the computers, content errors and fixing of weight in computerized prediction systems. In the near term, the banking software could be more reliable, accurate, and dynamic in nature and can be fit in with an automated processing unit. The old data sets are initially fed into the system for training purpose and then the new data sets. Machine learning helps to understand the factors which affect the specific outcomes most. Other models like neutral network and discriminate analysis can be used individually or combined for enhancing reliability and accuracy prediction.

REFERENCES

1. Aboobyda Jafar Hamid and Tarig Mohammed Ahmed, "Developing Prediction Model of Loan Risk in Banks using Data Mining", Machine Learning and Applications: An International Journal (MLAIJ), Vol.3, No.1, pp. 1-9, March 2016.
2. S. Vimala, K.C. Sharmili, "Prediction of Loan Risk using NB and Support Vector Machine", International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018.
3. Aditi Kacheria, Nidhi Shivakumar, Shreya Sawkar, Archana Gupta, "Loan Sanctioning Prediction System", International Journal of Soft Computing and Engineering (IJSCE), vol. 6, no. 4, pp. 50-53, 2016.
4. X. Francis Jency, V.P.Sumathi, Janani Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients", International Journal of Recent Technology and Engineering (IJRTE), Vol. 7, No. 48, pp. 176-179, 2018.
5. Anchal Goyal, Ranpreet Kaur, "Loan Prediction Using Ensemble Technique", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 3, pp. 523 – 526, March 2016.
6. Anchal Goyal, Ranpreet Kaur, "Accuracy Prediction for Loan Risk using Machine Learning Models", International Journal of Computer Science Trends and Technology (I JCST), Vol. 4, Issue 1, pp. 52-57, Jan - Feb 2016
7. Sudhamathy, "Credit Risk Analysis and Prediction Modelling of Bank Loans using R", International Journal of Engineering and Technology (IJET), Vol. 8, No. 5, pp. 1954-1966, Oct-Nov 2016.
8. Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, K. Vikas, "Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques, Vol. 5, Issue 2, pp. 144-148, Mar-Apr 2019.
9. Kumar Arun, Garg Ishan, Kaur Sanmeet, "Loan Approval Prediction based on Machine Learning Approach", IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue 3, pp. 79-81, Ver. I (May-Jun. 2016).
10. How Decision tree algorithm works, https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/
11. Naive Bayes — Probabilistic Algorithm, https://medium.com/datadriveninvestor/naive-bayes-probabilistic-algorithm-68d5a6647738
12. Loan Prediction Dataset Source, https://www.kaggle.com/uttam96/loanpred.
13. Supervised Learning with Python, https://www.datascience.com/blog/supervised-learning-python.
14. K. Kavitha, "Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp.162–166, 2016.