# ML BASED LOAN PREDICTION

Prof. R.Yadagiri Rao[1], K.Srikanth[2],S. Ajay Kumar[3],S.Srinath[4], Y. Sandeep Kumar[5], B.Rajesh[6]
[1] Professor , Dept of CSE Sri Indu Institute of Engineering & Technology, Hyderabad
[2] Assistant Professor, Dept of ECE, Sri Indu Institute of Engineering & Technology, Hyderabad
[3-6] Student, Dept of ECE Sri Indu Institute of Engineering & Technology, Hyderabad.

**Abstract:** In the recent years, the number of people applying for the loans gets increased for various reasons. Loans can be taken from many sources. One of them is by using the online platforms that connects the investors and borrowers. As the investors involved in this platform are individual people not any organizations it is very important to make a wise decision in order to invest in a particular customer.The investor should be able to know whether is it safe to give loan to the borrower or not i.e., can the borrower be able to repay the loan. These platforms act as a connecting bridge for the people interested in investing their money to attain goodamount of interest and the people who wants to borrow money with low-interest rates compared to other sources. In this project, we are using Lending club data set to determine whether the loan is re- payed or charged-off., analyze the data using Exploratory Data Analysis and apply the machine learning algorithms like KNN Classifier, Random Forest Classifier, Decision Tree and Logistic Regression.

## I. INTRODUCTION

Peer-to-peer (P2P) lending, which is also known as Social lending, operates online trading platforms as a medium for lending money without the intrusion of traditional financial mediators, such as banks. Conducting business on peer platforms has recently become popular because it not only reduces financing costs but also has the potential for higher profitability for both investors and borrowers. Borrowers benefit from lower interest rates; investors receive a higher return than they would from a bank. However, evaluating the risk of investing is a common challenge in micro financing, where loans are typically unsecured. Further, P2P lending usually occurs in settings with a high level of information asymmetry – that is, settings where the investors do not have complete information about the borrowers' credit history. Even though when the information is available, lenders might not know how to extract useful data from the given data. Therefore, predicting a borrower's creditworthiness on whether or not to fund particular loans has emerged as a critical problem for online lending platforms. Hence machine learning algorithms on used on the available data in order to generate a model that helps to predict the repayment of loan.

Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e.the higher the interest rate, the riskier the borrower. We will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays theloan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment. Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default.

## II. LITERATURE REVIEW

In 2019, Vimala and Sharmili [1] proposed a loan prediction model usingNB 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 ofpredictions. 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 prediction. In 2019, Jency, Sumathi and Shiva Sri [2] 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 annual income versus loan purpose, customer‘s trust, loan tenure versus delinquent months, loan tenure versus credit category, loan tenure versus number of years in the current job, and 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 onavailing short-tenure loans rather than long-tenure loans.

“Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process” Author-
E. Chandra Blessie, R. Rekha - Year- 2019
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 thehold. This is achieved by feeding the past records of thecustomer 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 Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.

Logistic regression still has its limitations, andit requires a large sample of data for parameter estimation. Logistic regression also requires that thevariables be independent of each other otherwise the model tends to overweigh the importance of the dependent variables.

A solution to this multi collinearity problem among the categorical explanatory variables is the use of a categorical principal component analysis which can be seen used by Guilder and Ozlem on a case study for housing Loan approval data. The goal of Principal component analysis is to reduce the number of m variables where many of them would be highly correlated with each other, to a smaller set of n uncorrelated variables called principal components which account for the variances between the previous m variables. Methods such as PCA are known as dimension reduction of the data. It may be suitable for scaled continuous variables but it isn’t quite an appropriate method of dimension reduction for categorical variables. Thus, the authors here used a tweaked version of PCA for categorical data called CATPCA or categorical (nonlinear) principal components analysis which is specifically developed for where the dependent variables are a mix of nominal, ordinal, or numeric data which may not have linear relationships with each other. CATPCA works by using a scaling process optimized to convert the categorical variables into numeric variable.

## III. EXISTING SYSTEM

According to this methodology, the steps of research can bedescribed as follows:

1.Business understanding

Business understanding: It is the initial phase which focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

1. Data understanding.

The data understanding phase focuses on initial data collection, familiarization of data, identification of data quality problems, and interesting subsets to form hypotheses for hidden information etc.

2. Data preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modelling tool from the initial raw data). There is no prescribed order for data preparation tasks. Sometimes tasks are to be repeatedly performed, like selection of table, record and attribute as well as transformation and cleaning of data for modelling tools.

3. Modeling

Various modelling techniques are selected and applied in this phase. Typically, there are several techniques for the same data mining problem type. Since some techniques have specific requirements on the form of data, sometimes it needs to go back to the data preparation phase

4. Evaluation

This phase is to be covered before proceeding to the final deployment of the model, to be certain that business objectives are properly achieved. Consideration and successful implementation of all important business issues are to be confirmed. At the end of this phase, a decision on the use of the data mining results should be reached.

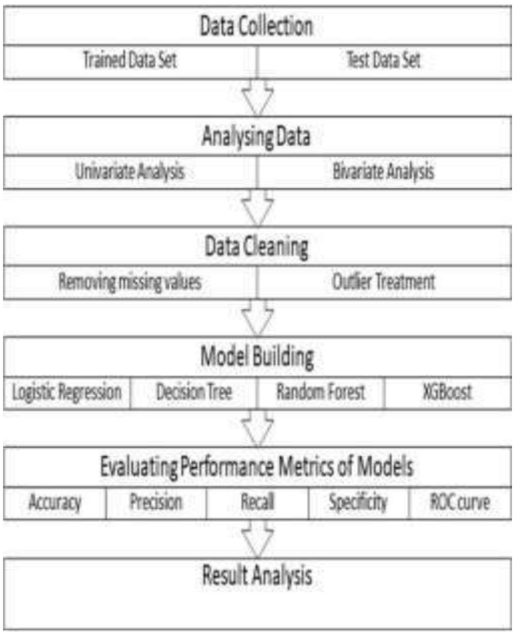
5. Deployment

Creation of the model is generally not the end of the project. The knowledge gained will have to be organized and presented in such a way that the customer can use it.

## IV.PROPOSED SYSTEM

### Machine Learning Predictive:

Machine Learning Predictive analytics is used to predict the data about future events. It includes many techniques such as data mining, machine learning [4, 9] and modelling. Machine learning is a type of artificial. intelligence that allows a software application to learn from the data & become more accurate in predicting outcomes without human intervention. Machine learning and deep learning help to design and develop such a machine that automatically learns and predicts your data and situation. Machine learning is often divided into different subcategories according to the type of problems being comes. Some ML type is as follows: 1) Supervised Learning Supervised learning is the point at which the model is getting prepared on a



labeled data

**FigNo.1:Methodology**

## 1.DECISION TREE

Classification is a two-step process, learning step and prediction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

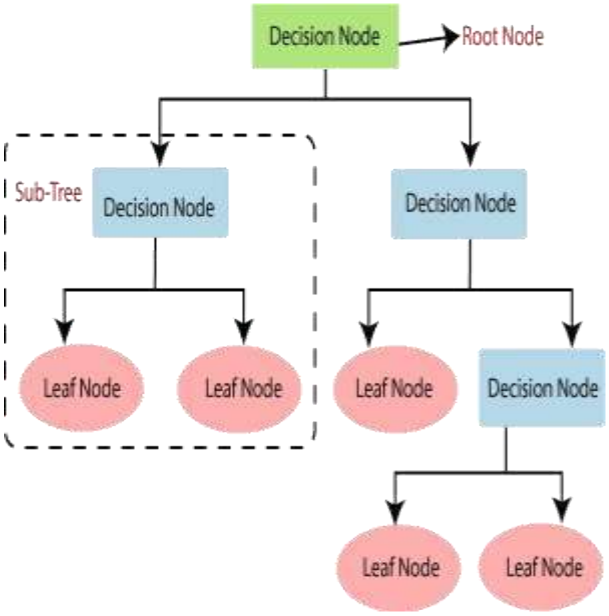


**Fig No .2 Decision Tree Flow Chart**

## 2. RANDOM FOREST

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Ensemble uses two types of methods:

1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

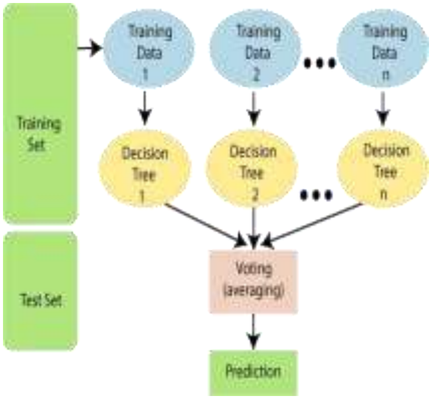


**Fig No.3 Random Forest Flow Chart**

## 3. XG BOOST

Basically, XGBoost is an algorithm. Also, it has recently been dominating applied machine learning. XGBoost is an implementation of gradient boosted decision trees. Although, it was designed for speed and performance. Basically, it is a type of software library. That you can download and install on your machine. Then have to access it from a variety of interfaces.

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models.

models, the most predominant usage has been with decision trees.The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical.

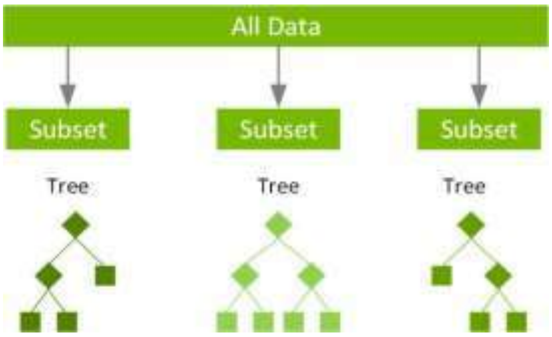


**Fig No.4 XG Boost**

## 4. LOGISTIC REGRESSION

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function.
- The dependent variable must be categorical in nature.
- The independent variable should not have multicollinearity.

## SYSTEM TESTING

Framework Testing is a kind of programming testing that is performed on a total incorporated framework to assess the consistency of the framework with the relating necessities. Framework testing recognizes absconds inside both the coordinated units and the entire framework. The consequence of framework testing is the noticed conduct of a part of a framework when it is tried.

Framework Testing is fundamentally performed by a testing group that is autonomous of the improvement group that assists with testing the nature of the framework unbiased. The steps of testing are involved for the proposed system as follow:

- Unit Testing
- Integration Testing
- Validation Testing.

### A. UNIT TESTING

Unit Testing, otherwise called Component Testing, is a degree of programming testing where individual units or parts of programming are tried. The reason for existing is to approve that every unit of the product proceeds as planned.

A unit is the littlest testable piece of any product. It for the most part has one or a couple of data sources and normally a solitary yield. Unit testing expands trust in changing/keeping up with code. It is worried about practical rightness of the independent modules.

### B. INTEGRATION TESTING

Joining testing takes as its feedback modules that have been unit tried, bunches them in bigger totals, applies tests characterized in a coordination test plan to those totals, and conveys as its yield the incorporated framework prepared for framework testing. It includes the mix of numerous modules which are firmly combined with one another. The primary capacity or objective of this testing is to test the interfaces between the units/modules. Blend testing can be started once the modules to be attempted are open. It doesn't need the other module to be finished for testing to be finished.

### C. VALIDATION TESTING

An approval test is utilized to test and check the ultimate result before conveying it to the client. This interaction includes understanding the target of the item with the goal that it would be not difficult to approve the result.

## Pre Processing

Data mining technique has been used in Pre- Processing for transforming raw data which is collect using online form into useful and efficient formats.There is a need to convert it in useful format because it may have some irrelevant,missing information and noisy data. To deal with this problem data cleaning technique has been used. Before data mining the data reduction techniques is used to deal with huge volume of data. So data analysis will become easier and it intends to get accurate results. So data storage capacity increase and cost to analysis of data reduces. The size of data can be reduced by encoding mechanisms. So it may be lossy or lossless. If the original data is obtained after reconstruction from compressed data, such reductions are called lossless reduction else it is called lossy reduction. Wavelet transforms and PCA (Principal Component Analysis) methods are effective for reduction.

ID 0

Sex 13
Married 3
No_Dependents 15
Qualification 0
In Service /

Self_Employed 32

Annual_Income_Applicant

0Annual_income_Coapplic

ant 0

Amount_Loan 22

Term 14

Credit_History _

Applicant 50 Assets 0

Status_Loan 0

### D. MODEL SELECTION

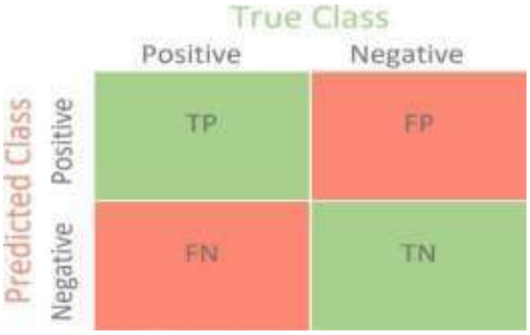
The process of selecting a final machine learning model from among a group of candidate machine learning models for a particular training dataset of Loan customer is called model selection. There are different types of model like logistic regression, SVM, KNN, etc. All these models have some merits and demerits for example predictive error gives the statistical noise in the data, the incompleteness of the sample data, and the limitations of each different model type. The chosen model meets the requirements and constraints of the stakeholders (Bank and Customers) project stakeholders. A model should have parameters like Skillful as compared to naive models. Skillful relative to other tested models. Skillful relative to the state-of-the-art.

### E. PROBLEM STATEMENT

Banks, Housing Finance Companies and some NBFC deal in various types of loans like housing loan, personal loan, business loan etc in all over the part of countries. These companies have existence in Rural, Semi Urban and Urban areas. After applying loan by customer these companies validates the eligibility of customers to get the loan or not. This paper provides a solution to automate this process by employing machine learning 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 algorithm. So the customer will fill an online loan application form. This form consist details like Sex, Marital Status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others. To automate this process by using machine learning algorithm, First the algorithm will identify those segments of the customers who are eligible to get loan amounts so bank can focus on these customers . Loan prediction is a very common real-life problem that every finance company faces in their lending operations. If the loan approval process is automated, it can save a lot of man hours and improve the speed of service to the customers. The increase in customer satisfaction and savings in operational costs are significant. However, the benefits can only be reaped if the bank has a robust model to accurately predict which customer's loan it should approve and which to reject, in order to minimize the risk of loan default

### F. MODEL EVALUATE

Model evaluation is technique which is used for the evaluating the performance of the model based on some constraints it should be kept in mind while evaluating the model that it can't underfoot or overfit the model. Various methods are present to evaluate the performance of the model such as Confusion metrics, Accuracy, Precision, Recall, F1 score etc.



1.ConfusionMetrics

**Fig No.6 Confusion Matrix**

1) 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.

2) 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 whole given dataset. Precision value tells the perfectness of our model. In our data set good precision value has been obtained.

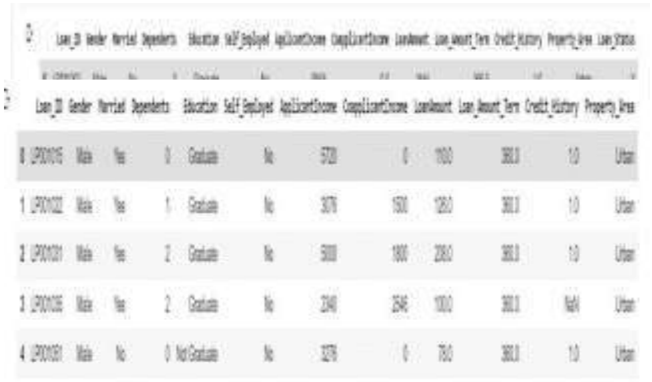
3) Recall: Percentage ratio of positive instances with actual total positive instances is recall value. Here denominator (TP + FN) shows the total number of positive instances which are present in whole dataset. As a result it has obtained 'how much extra right ones, the model will failed if it shows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

4) 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 valueand doesn't miss out on positives and predicts them negative (recall).

## I. SIMULATION RESULT

**Fig.7. Training dataset**

**Fig.8. Test dataset**



| | Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

| | Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | LP001015 | Male | Yes | 0 | Graduate | No | 5720 | 0 | 110.0 | 360.0 | 1.0 | Urban |
| 1 | LP001022 | Male | Yes | 1 | Graduate | No | 3076 | 1500 | 126.0 | 360.0 | 1.0 | Urban |
| 2 | LP001031 | Male | Yes | 2 | Graduate | No | 5000 | 1800 | 208.0 | 360.0 | 1.0 | Urban |
| 3 | LP001035 | Male | Yes | 2 | Graduate | No | 2340 | 2546 | 100.0 | 360.0 | NaN | Urban |
| 4 | LP001051 | Male | No | 0 | Not Graduate | No | 3276 | 0 | 78.0 | 360.0 | 1.0 | Urban |

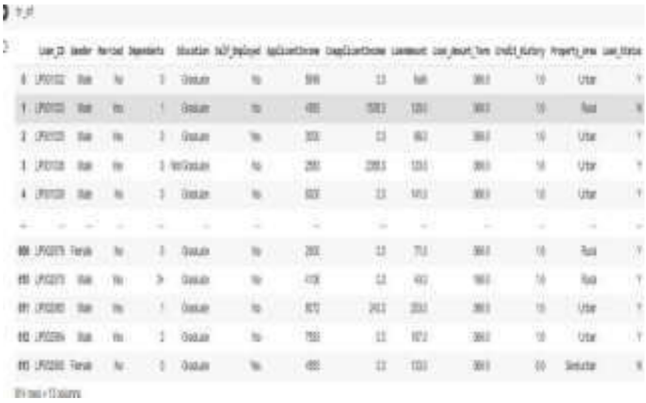


Fig.9. Missing Values

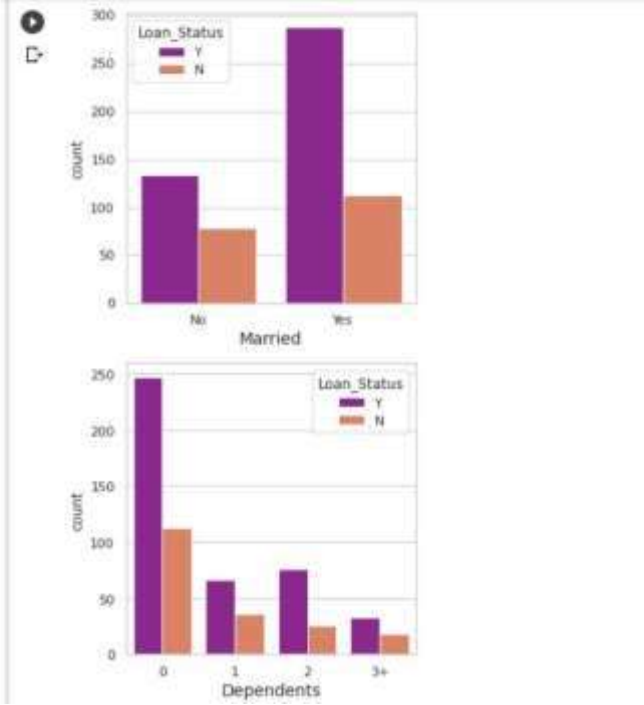


Fig.10. Married & Dependents

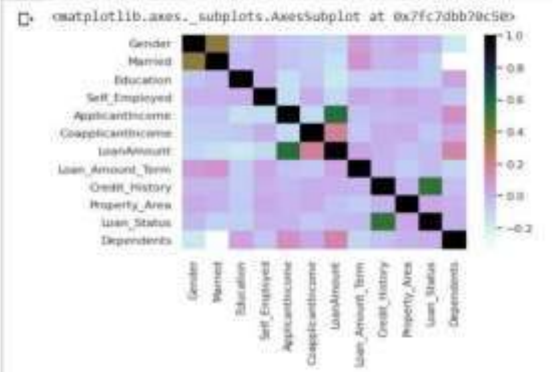
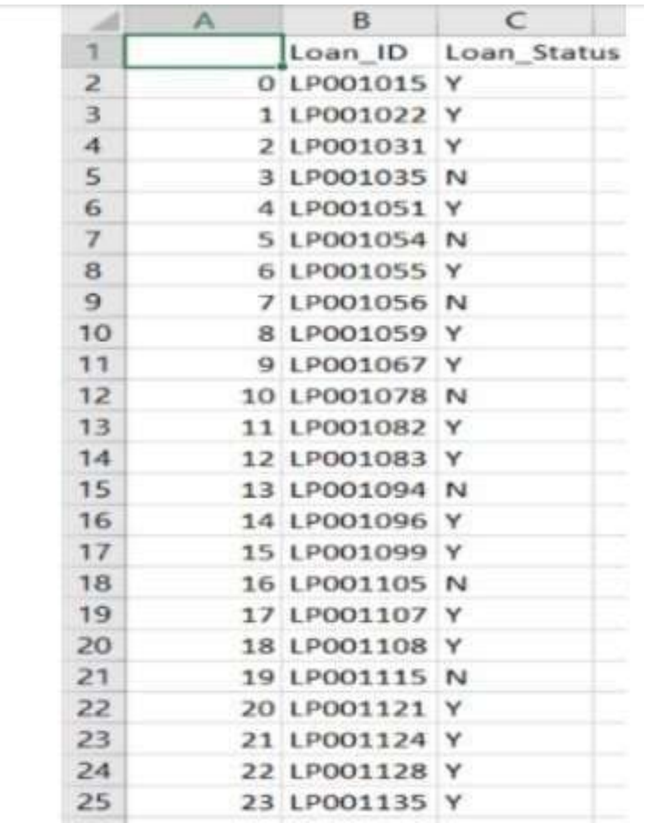


Fig.11. Correlation matrix



|  | A | B | C |
|---|---|---|---|
| 1 |  | Loan_ID | Loan_Status |
| 2 | 0 | LP001015 | Y |
| 3 | 1 | LP001022 | Y |
| 4 | 2 | LP001031 | Y |
| 5 | 3 | LP001035 | N |
| 6 | 4 | LP001051 | Y |
| 7 | 5 | LP001054 | N |
| 8 | 6 | LP001055 | Y |
| 9 | 7 | LP001056 | N |
| 10 | 8 | LP001059 | Y |
| 11 | 9 | LP001067 | Y |
| 12 | 10 | LP001078 | N |
| 13 | 11 | LP001082 | Y |
| 14 | 12 | LP001083 | Y |
| 15 | 13 | LP001094 | N |
| 16 | 14 | LP001096 | Y |
| 17 | 15 | LP001099 | Y |
| 18 | 16 | LP001105 | N |
| 19 | 17 | LP001107 | Y |
| 20 | 18 | LP001108 | Y |
| 21 | 19 | LP001115 | N |
| 22 | 20 | LP001121 | Y |
| 23 | 21 | LP001124 | Y |
| 24 | 22 | LP001128 | Y |
| 25 | 23 | LP001135 | Y |

Fig No.12. Result

## II. CONCLUSION

To detect the outliers the data is demonstrated visually and afterwards handled the outliers. When the outliers decisions visualized are of high precision and accurate. Percentiles is another mathematical method to detect outliers. In this method,it assumes a certain percentage of value from top or taken it from bottom as an outlier. The key point is here to set the percentage value once again, and this depends on the distribution of your data as mentioned earlier.

**FUTURE SCOPE:** The system is trained on old training dataset in future software can be made such that new testing data should also take part in training data after some fix time. In future, this model can be used to compare various machine learning algorithm generated prediction models and the model which will give higher accuracy will be chosen as the prediction model. This paper work can be extended to higher level in future. Predictive model for loans that uses machine learning algorithms, where the results from each graph of the paper can be taken as individual criteria for the machine learning algorithm.

## III. REFERENCES

[1]. 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 (2), 2013.

199 –201.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2]. J.H. Aboobyda, and M.A. Tarig, "Developing Prediction Model of Loan Risk in BanksUsing Data Mining", MachineLearning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016. K. Elissa, "Title of paper if known," unpublished.

[3]. A.B. Hussain, and F.K.E. Shorouq, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach", Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014. JAC: A JOURNAL OF COMPOSITION THEORY Volume XIII, Issue V, MAY 2020 ISSN: 0731-6755Page No: 324.

[4]. T. Harris, "Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions", Expert Systems with Applications, vol. 40, pp. 4404–4413, 2013.

[5] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma- "Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques - Volume 5 Issue 2, Mar-Apr 2019.

[6] Sudhamathi G.-"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.

[7] Aboobyda Jafar Hamid and Tarig Mohammed Ahmed - "Developing Prediction Model of Loan Risk in Banks using Data Mining".

[8] Anchal Goyal , Ranpreet Kaur-"Loan Prediction Using Ensemble Technique", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

[9] X. Francis Jency, V.P.Sumathi, Janani Shiva Sri- "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients".