Loan Status Prediction

Sumanth Prasad - PES2201800092

Gokul K M - PES2201800517

Y R Pavan Sai - PES2201800484

ABSTRACT:- Loan prediction analysis is one such analysis which banks opt to sort which customers would be able to repay the loan if provided to them. So, the banking company should be able to find out the key criteria that are needed for granting the loan. Customers first apply for a loan which can be of any type. After that the banking company validates the customer eligibility for the loan. We have classified the customers based on the Loan_Status variable present in the dataset.

I. Introduction

This analysis is used to uncover the underlying structure of a relatively larger set of variables present in the dataset.. Almost all banks deploy this analysis scheme to ensure guaranteed return of the loan given. This also avoids any scams related to loans and ensures safe transactions for customers demanding the loan. The dataset comprises multiple attributes such as loan id, gender, married, education, loan amount.etc. which would directly affect the income of a person in 1994, as explained in our report which says that married people earned more than single people during the year 1994. We have used different methods such as logistic regression, K nearest neighbours and decision tree classifiers to compare the results

II. Dataset

We have used the **customer loan dataset** for our problem statement. We use 2 datasets here: one for training and one for testing. There are altogether 25 columns in our dataset. We use 13 columns for the training dataset and 12 columns for the test dataset.

1) Loan ID -> This is a unique number given to every applicant.

2) Gender -> These are the “sex” of various applicants.

3) Married -> A married applicant is represented as Y and a non married applicant as N in the dataset.

4) Dependents -> These are the other people on whom the customer is dependent on.

5) Education -> This is based on whether the applicant is a non graduate or a graduate.

6) Self_Employed -> A self-employed applicant is written as Y and a non self-employed as N in the dataset

7) Applicant Income -> This is the income obtained by the applicant.

8) Co Applicant income -> Some applicants have co-applicants and this column refers to the co-applicant’s income.

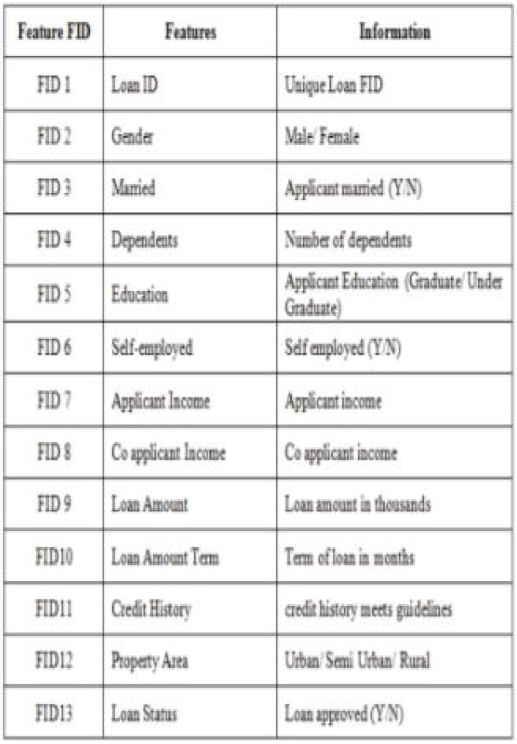
9) Loan Amount -> This is the amount of loan which the applicant desires.

10) Loan_Amount_Term -> This is the time given for the applicant to repay his loan.

11) Credit_History -> A value =1 tells us that the applicant has a decent history of repayments of loans if he has taken

12) Property_Area -> This refers to the region from where the applicant comes from.

13) Loan_Status -> This is the target variable for our problem statement. A “no” means the applicant is not granted the loan and a “yes” means the applicant is granted the loan.



Feature FID	Features	Information
FID 1	Loan ID	Unique Loan FID
FID 2	Gender	Male/ Female
FID 3	Married	Applicant married (Y/N)
FID 4	Dependents	Number of dependents
FID 5	Education	Applicant Education (Graduate/ Under Graduate)
FID 6	Self-employed	Self employed (Y/N)
FID 7	Applicant Income	Applicant income
FID 8	Co applicant Income	Co applicant income
FID 9	Loan Amount	Loan amount in thousands
FID10	Loan Amount Term	Term of loan in months
FID11	Credit History	credit history meets guidelines
FID12	Property Area	Urban/ Semi Urban/ Rural
FID13	Loan Status	Loan approved (Y/N)

Figure 1: Columns in the dataset

III. Single variable analysis

Now, we perform exploratory data analysis using a single variable.

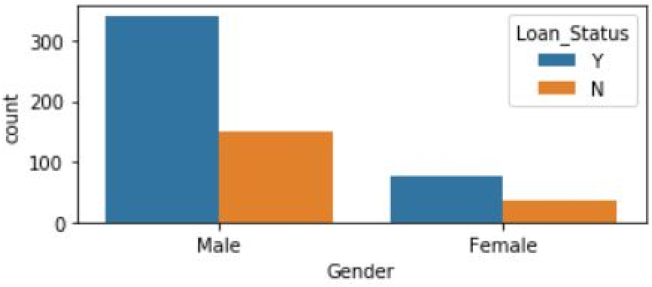


Figure 2: For variable "Gender"

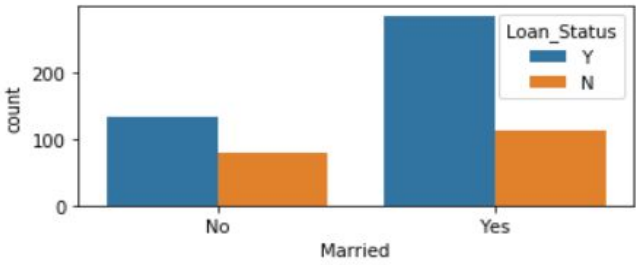


Figure 3: For variable "Married"

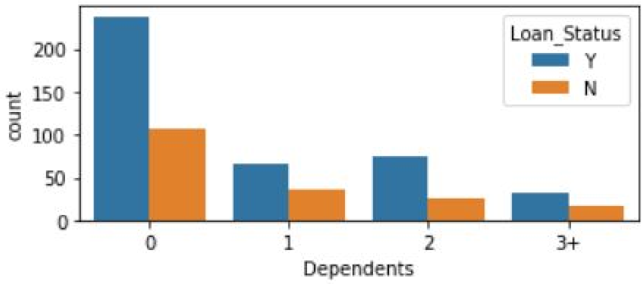


Figure 4: For variable "Dependents"

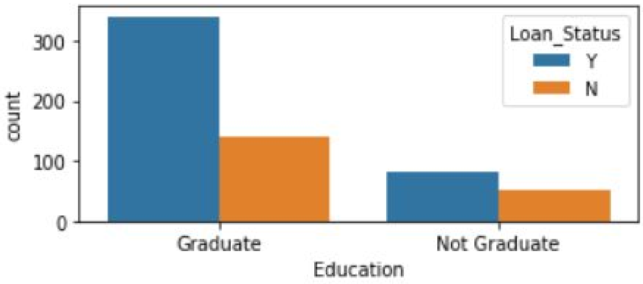


Figure 5: For variable "Education"

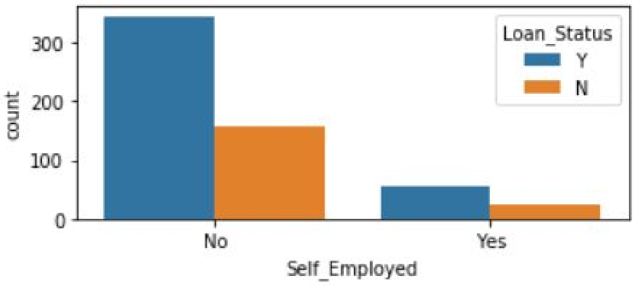


Figure 6: For variable "Self_Employed"

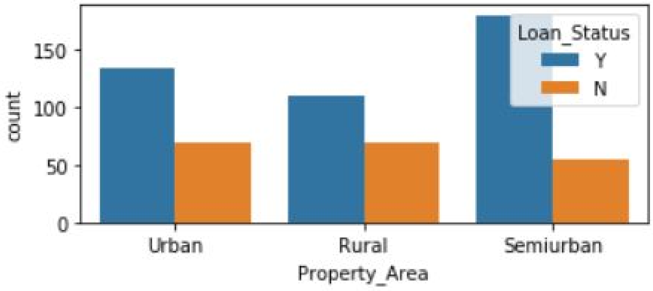


Figure 7: For variable "Property Area"

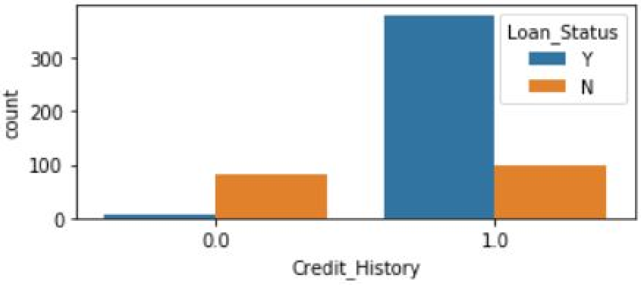


Figure 8: For variable "Credit History"

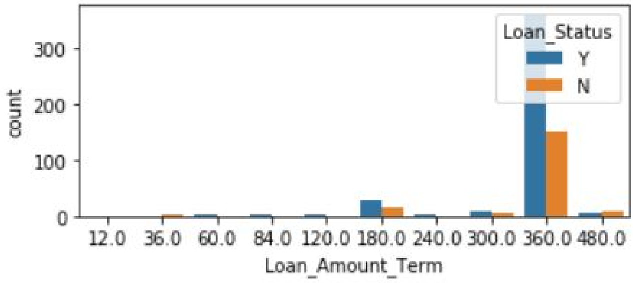


Figure 9: For "Loan Amount Term"

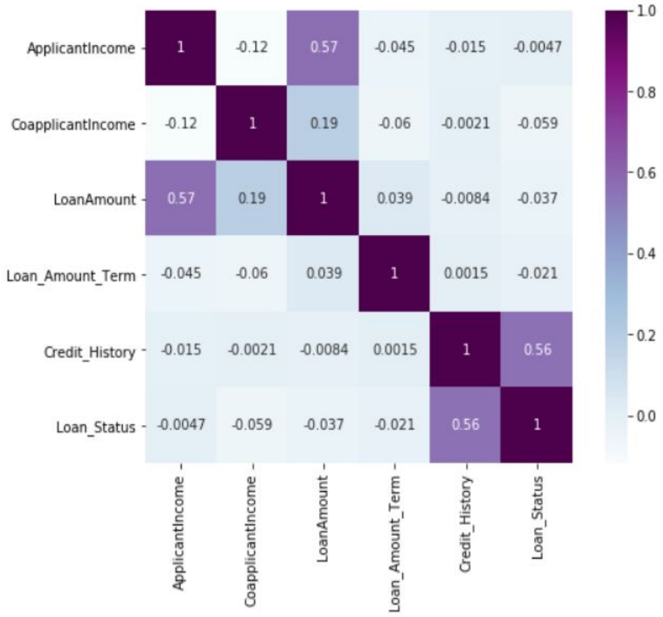


Figure 10: Correlation among numeric values

From the analysis done, we can come to the following conclusions:

1) Approximately 80% are Male and 20% are female applicants.
2) Percentage of applicants having no dependents is higher.
3) There are more graduate applicants than non-graduates.
4) Semi Urban people are slightly higher than Urban people among the applicants.
5) The percentage of people having a good credit history is high.
6) The percentage of people that the loan has been approved for has been higher rather than the percentage of applicants for which the loan has been declined.
7) According to the correlation analysis:

ApplicantIncome - LoanAmount has correlation coefficient of 0.57
Credit_History - Loan_Status has correlation coefficient of 0.56
LoanAmount is also correlated with CoapplicantIncome with correlation coefficient of 0.19.

IV. Preprocessing

Steps included for preprocessing:

1) The basic preprocessing step is to check for any NaN and missing values.. We use the isnull( ) function to check for missing values.
2) The missing values were replaced with median for continuous variables and mode for categorical variables.
3) As we know that a loan id must be unique to all customers, we check for the uniqueness as well for the Loan_Id variable using the unique( ) function.
4) For modelling, we have to convert any categorical variables to numeric ones. Here, we apply the get dummies ( ) function using the regular one-hot encoding method.

These preprocessing steps are enforced on both, the training and the test data sets.

As we are interested in the Loan_Status variable, we analyze numerical variables by plotting boxplots against the Loan_Status variables.

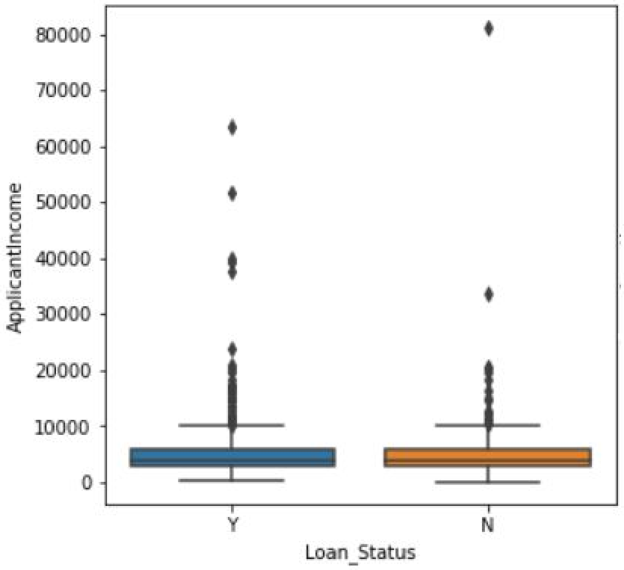


Figure 10: Loan_Status vs Applicant Income

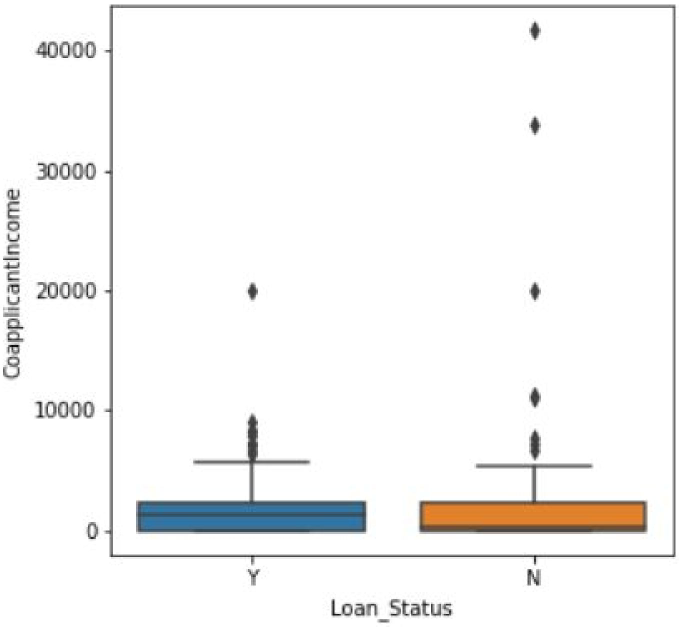


Figure 11: Loan_Status vs CoApplicant Income

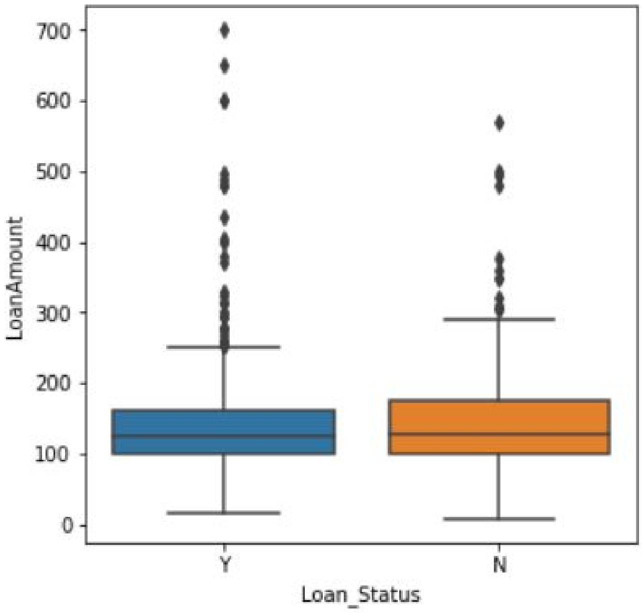


Figure 12: Loan_Status vs Loan_Amount

From the plots, we see that there is no significant relation to the Loan_Status variable.

V. Model Building:

As our problem statement is a classification problem, there can be several models put to use. Our team has used 3 models to check which fits the best for our problem statement.

A) Logistic Regression

- Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.
- Logistic regression is an estimation of Logit function. Logit function is simply a log of odds in favor of the event.
- This function creates a s-shaped curve with the probability estimate.
- Logistic Regression takes only numeric values as input

For our problem statement, we will make use of dummy variables to deal with categorical variables. For the “Gender” variable, it has two classes, Male and Female. Once we apply dummies to this variable, it will convert the “Gender” variable into two variables (Gender_Male and Gender_Female), one for each class, i.e. Male and Female. Gender_Male will have a value of 0 if the gender is Female and a value of 1 if the gender is Male

Now, we train the model using the training data set and make predictions using the test data set.

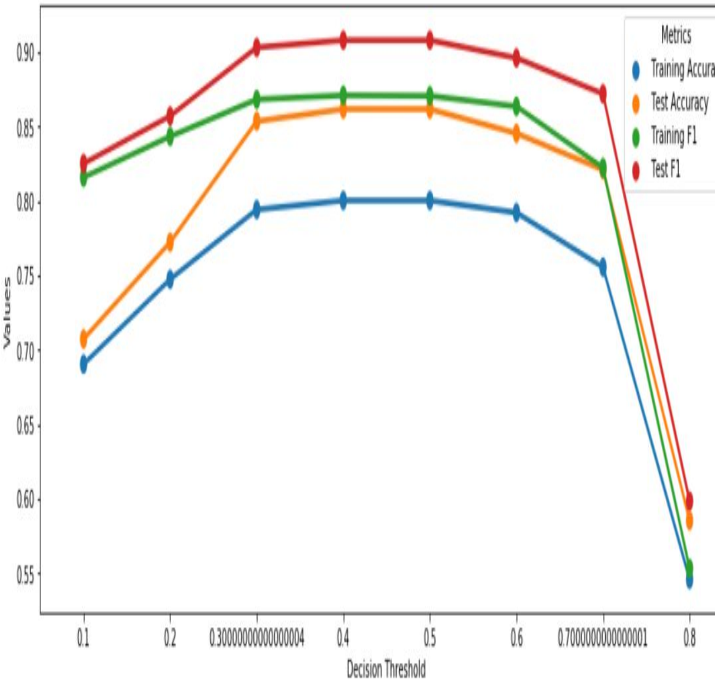
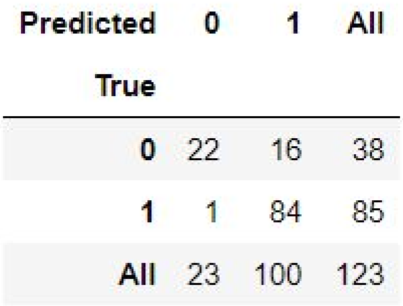


Figure 13: Test-Train Curve 1

From the above test-train curve, we could keep a min threshold of 0.4.



Predicted	0	1	All
True			
0	22	16	38
1	1	84	85
All	23	100	123

Figure 14: Confusion Matrix 1

Test Accuracy: 0.8617886178861789
Test F1 Score: 0.9081081081081082

Based on the above 2 metrics, we find that logistic regression gave an approximate **86%** accuracy.

B) Decision Tree Classifier

Decision tree is a supervised learning algorithm that is mostly used in classification problems. Here, we split the population into two or more homogeneous sets based on the most significant splitter in input variables. Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. We use Gini-Index to find the best node to split.

We get the following results after training the model and testing the predictions made.

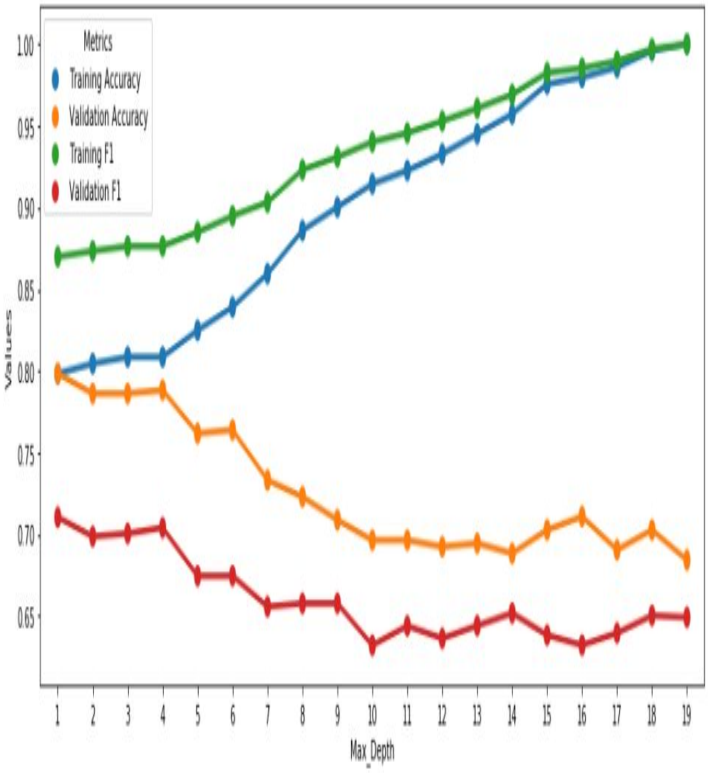


Figure 14: Max Depth Plot

From the above graph, we can conclude that keeping 'Max_Depth' = 3 will yield optimum Test accuracy and F1 score Optimum Test Accuracy ~ 0.805; Optimum F1 Score: ~0.7.

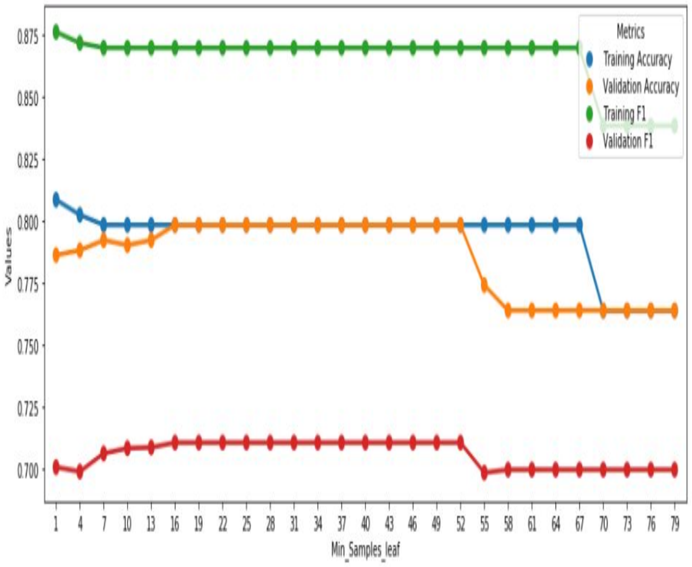
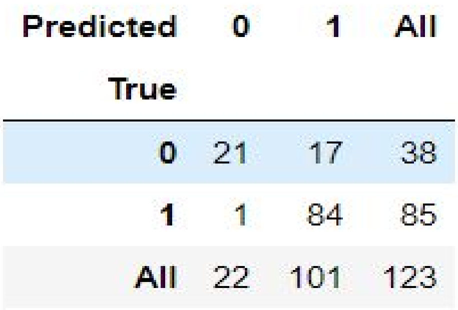


Figure 15: Min-Samples leaf plot

From the above plot, we can consider the minimum number of leaf samples = 35 to improve test accuracy.



Predicted	0	1	All
True			
0	21	17	38
1	1	84	85
All	22	101	123

Figure 16: Confusion matrix 2

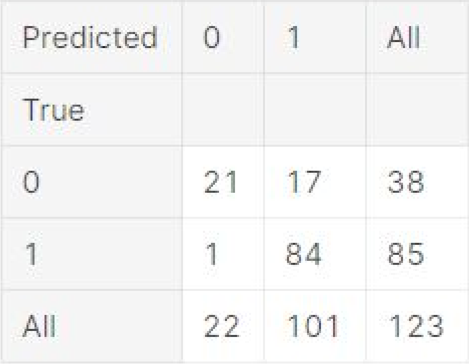
Test Accuracy: 0.8536585365853658
Test F1 Score: 0.903225806451613

Although we got a good test accuracy, from the confusion matrix, we see that there have been misclassifications. Some customers are granted loans but actually, they shouldn't.

C) Random Forest Classifier

- RandomForest is a bootstrap algorithm implementation where a number of decision trees are combined to make a powerful model.

- For every learner, a sample of rows and a few randomly chosen variables are used to build a decision tree.
- The end predicted value can be that of all the predictions made by the individual learners.
- Usually the bigger the forest the better, there is a small chance of overfitting here. The more estimators you give it, the better it will do.



Predicted	0	1	All
True			
0	21	17	38
1	1	84	85
All	22	101	123

Figure 17: Confusion matrix 3

Test Accuracy: 0.8536585365853658
Test F1 Score: 0.903225806451613

Here, we almost got the same accuracy as that of the decision tree classifier.

VI. Experimental Results

Looking at the accuracy values and the least number of misclassifications, we have used logistic regression as our proposed model for prediction. But we must make not of some of the assumptions made by the model :

1) The third category of gender is not included.
2) The probability of clearing the loan amount would be higher if the applicant is a graduate.
3) If co applicant income is higher , the probability of being eligible would be higher.
4) If Loan amount is higher , the probability of repaying would be lesser and vice versa.

VII. Conclusion

Hence, we conclude that the loan prediction analysis is done based on the accuracy level of different models. As doing this manually takes a lot of time, we have automated the loan eligibility process (real time) based on customer information. There are many more attributes which can be taken into account, but we have considered the most important and widely used attributes. All banks try their best to segregate customers based on whether they can repay the loan or not, but there are still many cases where the customers do not repay the loan but satisfy all requirements for the banks.

VIII. References

1) https://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%203/4.%2018-21.pdf?id=7557
2) https://medium.com/@vishnumbaprof/case-study-loan-prediction-ac035f3ec9e4
3) http://cloudstechnologies.in/cloudtech-admin/basepaperfiles/1593149297loan%20approval%20prediction%20using%20decision%20tree%20in%20python.pdf