Introduction:

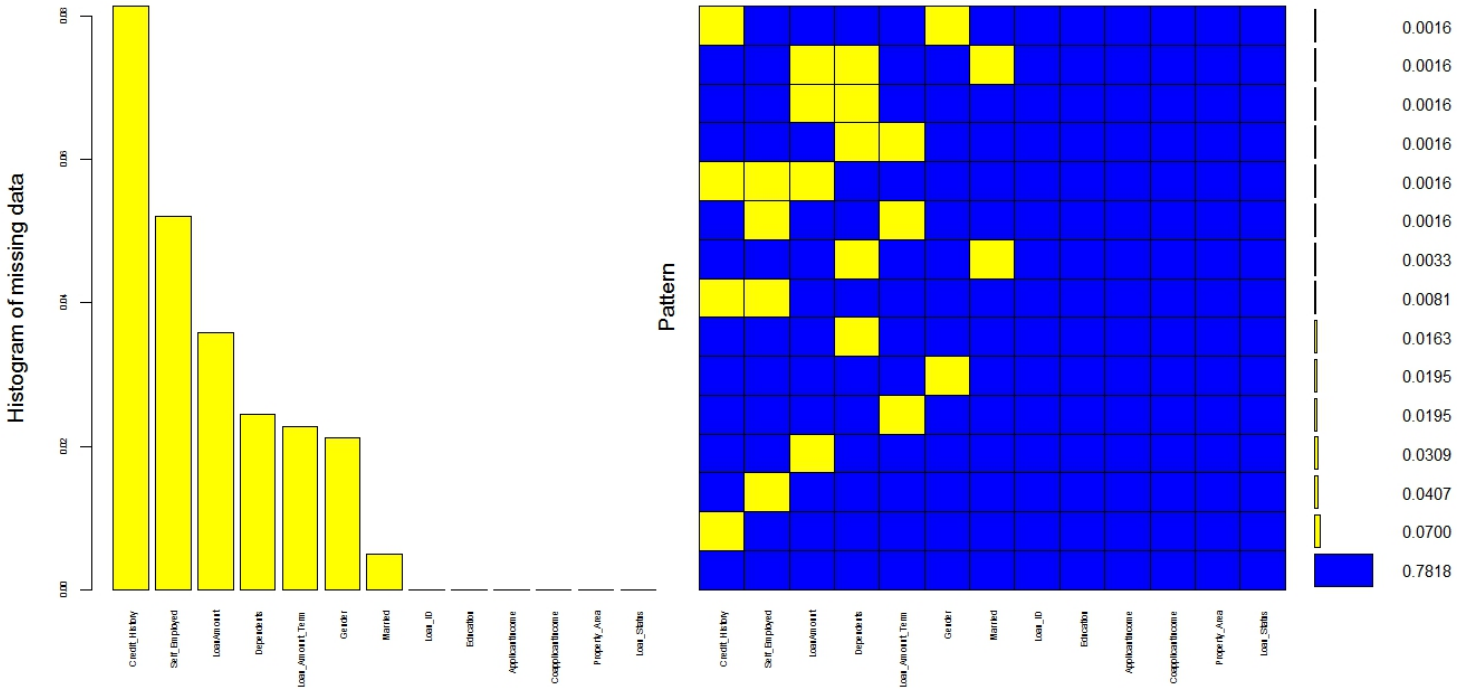
The file is a brief summary of the loan prediction problem. The outputs as well as the essential findings have been mentioned.

The entire analysis is carried out in three steps:

1. Exploratory Data Analysis
2. Implementing the Machine Learning Model
3. Studying the ROC Curves

Exploratory Data Analysis

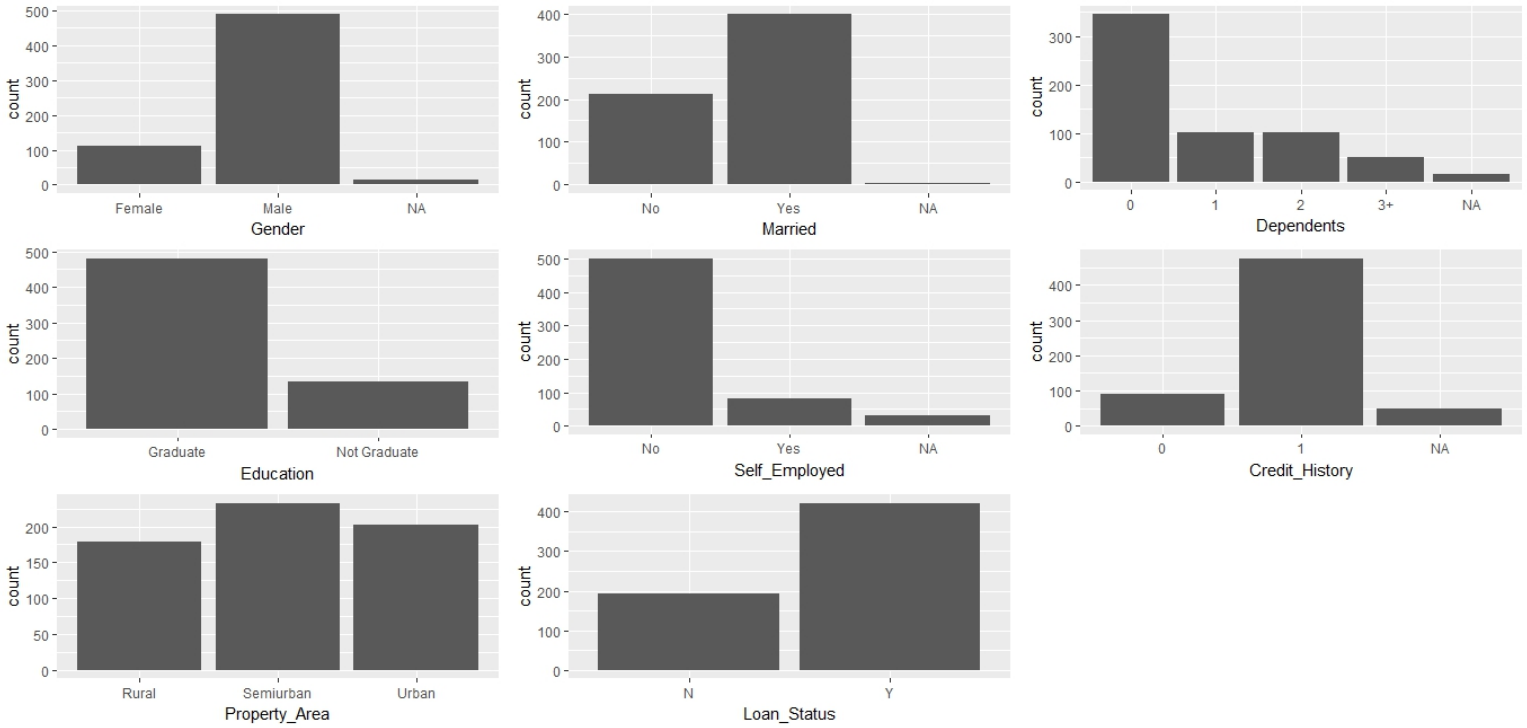
The required packages are installed and loaded in R. The file is then read keeping in mind that some NA are coded as empty strings, hence it is converted to NA during the process. The missing data is then analyzed using the aggregate function in R which results in the following plot:



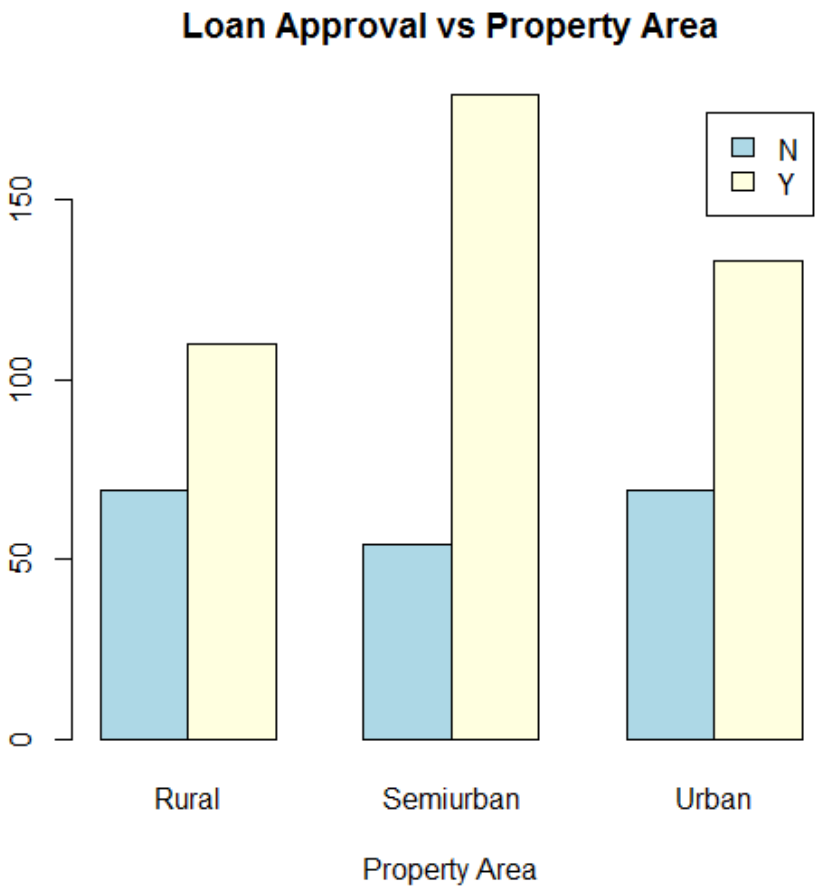
The plot shows the distribution of NA depicting that Credit_History has the highest number of missing fields. It also shows that 78.18 % of the data has no NA value.

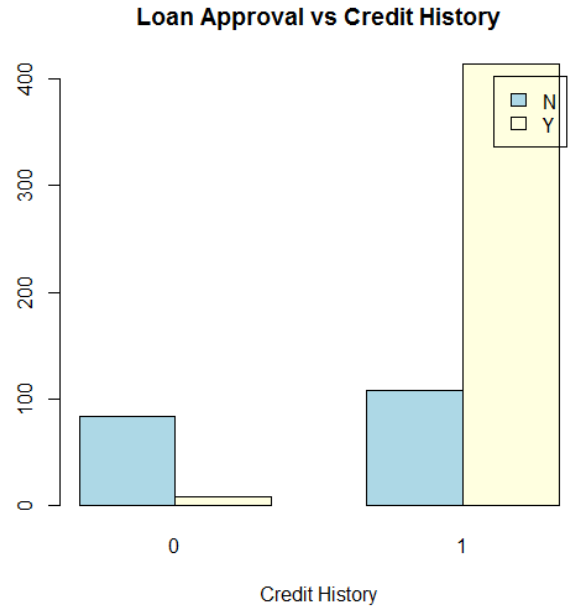
A Univariate analysis is done for the attributes which can be seen in the next figure . Following inferences can be drawn out of it:

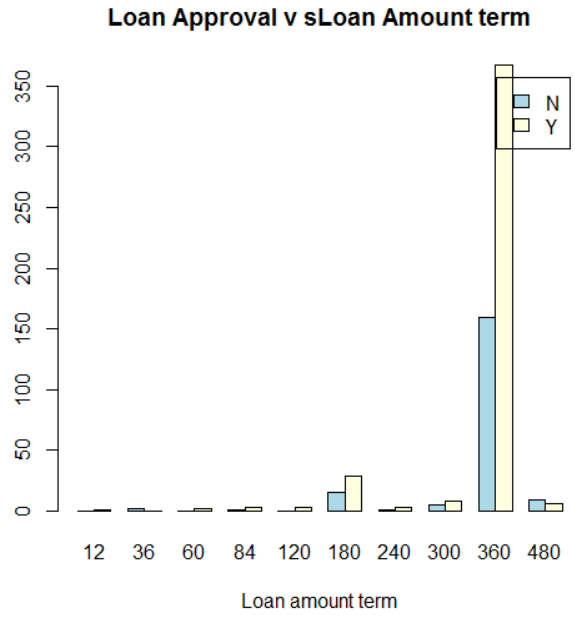
- The male population forms a huge ratio of the total applicants.
- Married population is almost double of that of single people.
- People with no dependants, people who are graduate and the ones who are not self employed form the majority.
- Applicants with a credit history of 0 are less in number.
- People having property In all the three areas are abundant with semi urban being higher than the other two.
- The data set has two-third loan applicants whose loan is sanctioned.

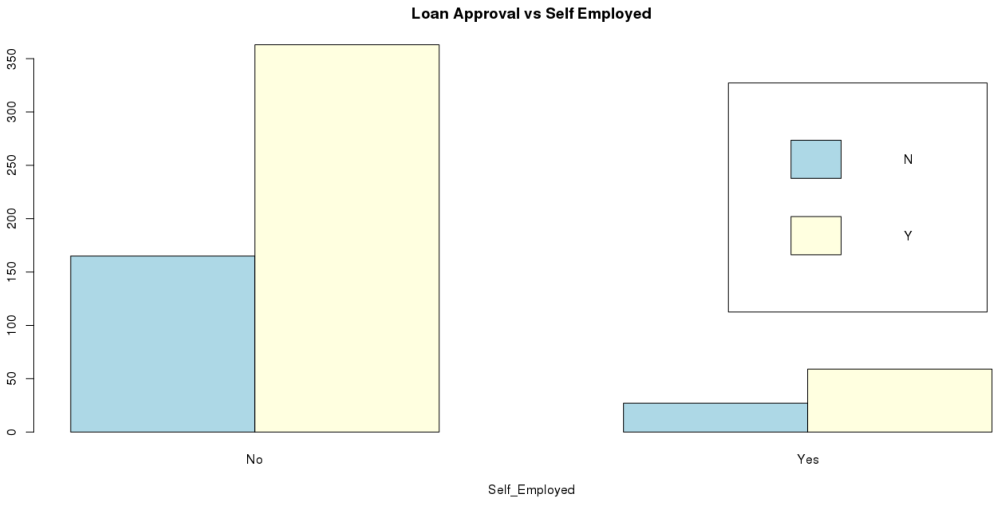


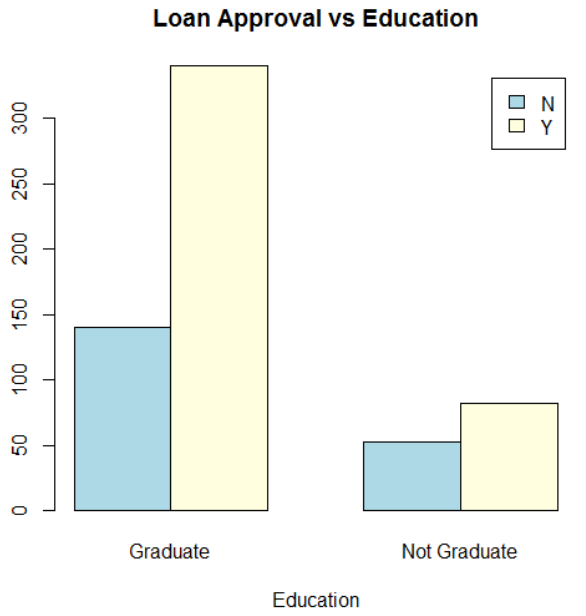
The next task is to perform imputation to fill in the missing values. This has been done using the MICE package. Several plot tables have then been plotted to study the relation between loan approval and given factors. Following plots are obtained:

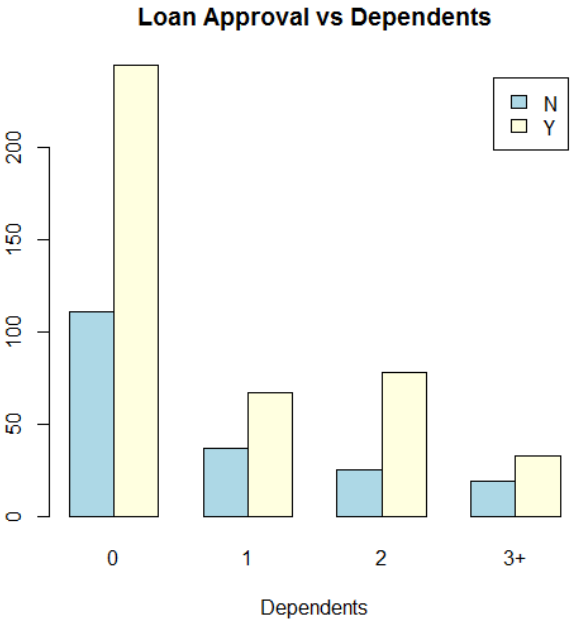


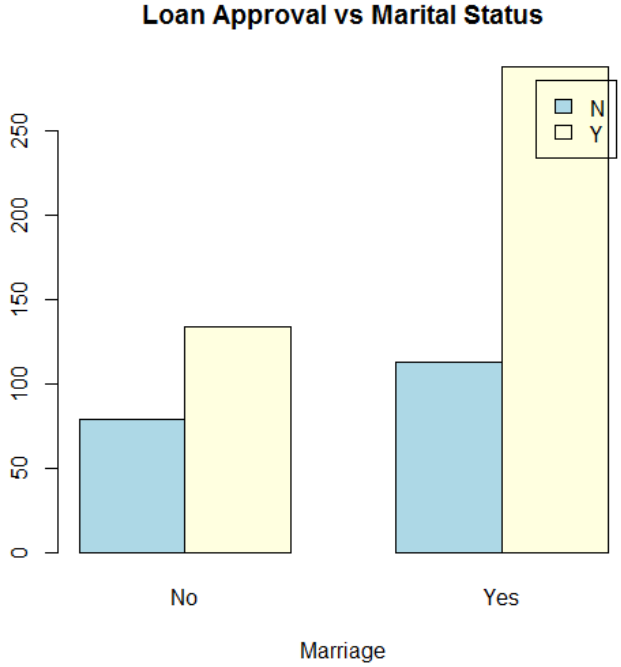

Loan Approval vs Credit History
400
300
200
100
0
N
Y
0
1
Credit History

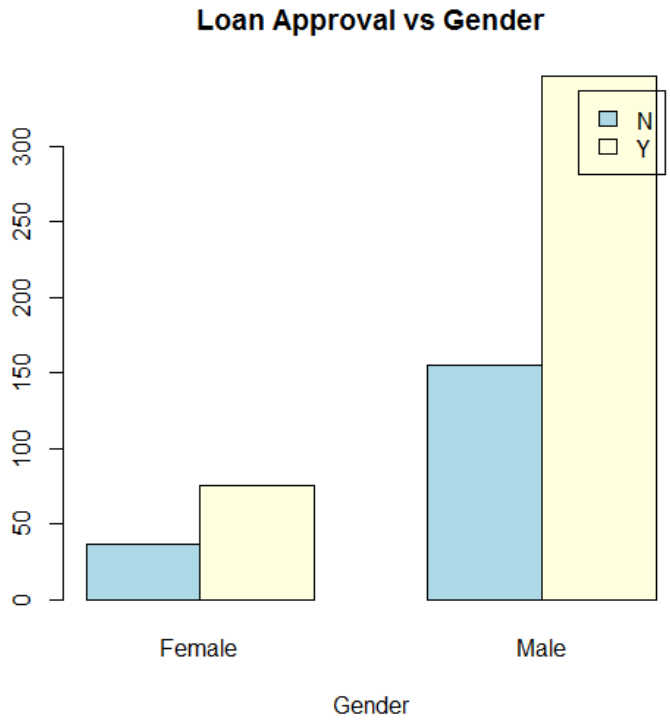

Loan Approval v sLoan Amount term
350
300
250
200
150
100
50
0
N
Y
12
36
60
84
120
180
240
300
360
480
Loan amount term


Loan Approval vs Self Employed
350
300
250
200
150
100
50
0
N
Y
No
Yes
Self_Employed


Loan Approval vs Education
N
Y
0
50
100
150
200
250
300
Graduate
Not Graduate
Education


Loan Approval vs Dependents
N
Y
0
50
100
150
200
0
1
2
3+
Dependents

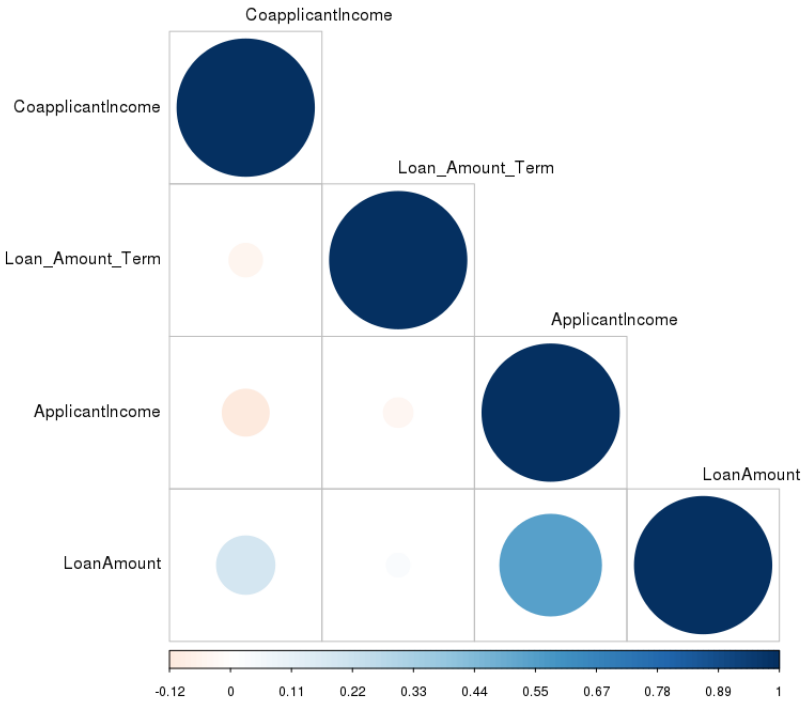

Loan Approval vs Marital Status
N
Y
0
50
100
150
200
250
No
Yes
Marriage



Following inferences can be drawn from the above plots:

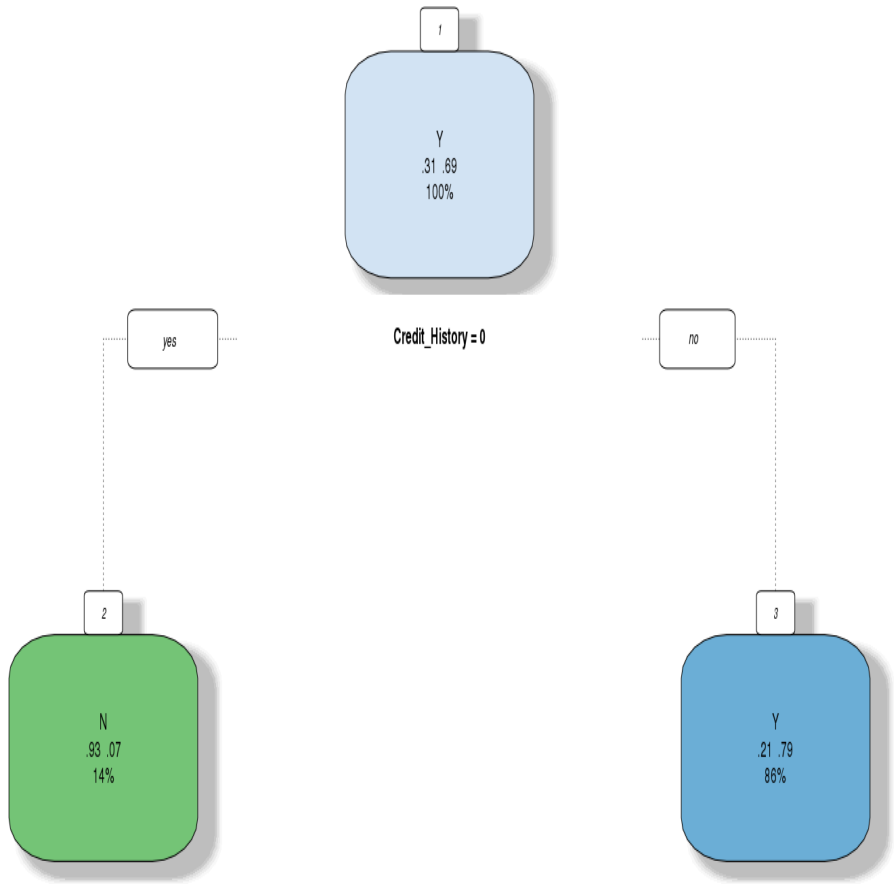
- The ratio of the approved loans is higher with semi urban property.
- Almost no loan has been sanctioned to applicants with 0 credit history.
- Most of the approved loans are for 1.5 years or 3 years duration.
- Graduates and non self-employed have been granted loans in majority.
- Male population, married people and people with no dependants are in abundance in the loan approved category.

Correlation for numeric attributes is then checked. The figure has been obtained using the corrplot function. A correlation is observed between ApplicantIncome and LoanAmount. This issue is later dealt with while constructing the model.



Implementing the Machine Learning Model

- The data is split into training and testing set in the ratio of 7:3. The first model is built using GLM. An accuracy of 81.08 % is obtained on the baseline criteria.
- A decision tree model is trained choosing the best value of complexity parameter which comes out to be .37 in this case. The following plot is obtained through this model :



- The decision tree model gives an accuracy of 81.08% .
- An accuracy of 81.62 % is obtained using the Random Forest.

Studying the ROC Curves

Area under the curve is calculated for the three models. The output obtained is as follows:

```
"AUC of Logistic Regression: 0.787"
"AUC of Decision Tree: 0.717"
"AUC of Random Forest: 0.735"
```

ROC curves have then been plotted for a visual interpretation: