**Project on ‘*Home Loan Predictions’***

1. **Business Problem**

Housing Finance company is a company which provide home loans for the houses which were present across all urban, semi-urban and rural areas for their valued customers.

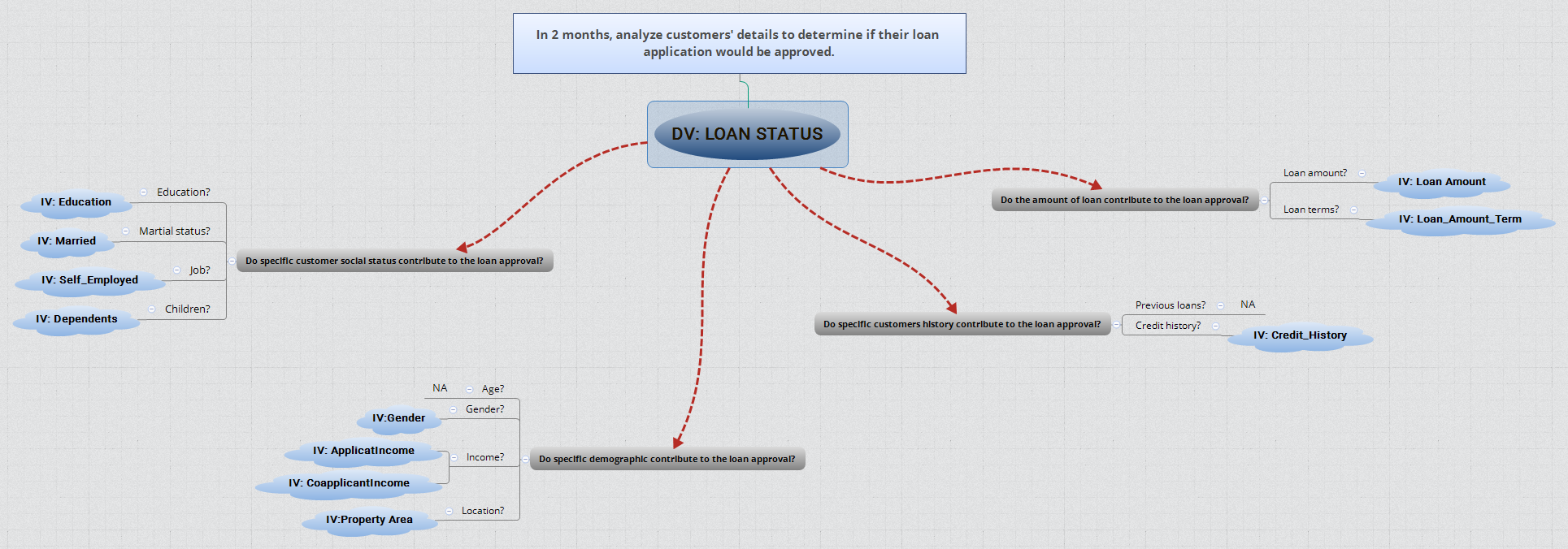
The company validates the eligibility of loan after customer applies for the loan. However, it consumes lot of time for the manual validation of eligibility process.

Hence, the company wants to automate the loan eligibility process based on the customer information and identify the factors/customer segments who are eligible for taking the loan.

As banks would give loans to only those customers that are eligible so that they can be assured of getting the money back.

Hence, the more accurate we are in predicting the eligible customers the more beneficial it would be for the company.

1. **Problem Statement**



The first thing we need to do and before jumping to analyse the data is to understand the problem statement and create a S.M.A.R.T objective. The next step is to identify our independent variables and our dependent variable. The above mind map illustrates the process I have conducted to structure plan the project.

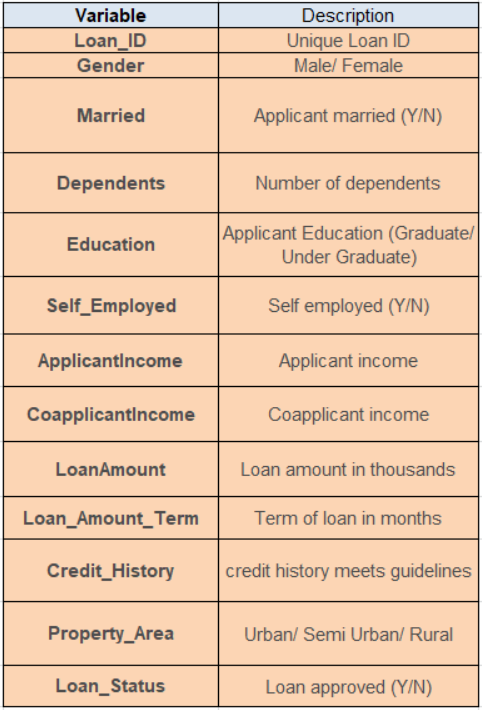
The Problem statement here is to automate the process for faster approval or rejection of home loan by analysing the various variables.

The two most critical questions in the lending industry are: 1) How risky is the borrower? 2) Given the borrower’s risk, should we lend him/her? The answer to the first question determines the interest rate the borrower would have. Interest rate measures among other things (such as time value of money) the risk of the borrower, i.e. the riskier the borrower, the higher the interest rate. With interest rate in mind, we can then determine if the borrower is eligible for the loan.

Investors (lenders) provide loans to borrowers in exchange for the promise of repayment with interest. That means the lender only makes profit (interest) if the borrower pays off the loan. However, if he/she doesn’t repay the loan, then the lender loses money.

1. **Brief Description of the data**

It’s very useful to know about the data columns before getting in to the actual problem for avoiding confusion at a later state. Now let us understand the data columns (that has been already given by the company itself) first so that we will get a glance.



There are altogether 13 columns in our data set. Of them Loan\_Status is the response variable and rest all are the variables /factors that decide the approval of the loan or not.

Now let us look in to each variable and can make some assumptions. (It’s just assumptions right, there is no harm in just assuming few statements)

**Loan ID** -> As the name suggests each person should have a unique loan ID.

**Gender** -> In general it is male or female. No offence for not including the third gender.

**Married** -> Applicant who is married is represented by Y and not married is represented as N. The information regarding whether the applicant who is married is divorced or not has not been provided. So, we don’t need to worry regarding all these.

**Dependents** -> the number of people dependent on the applicant who has taken loan has been provided.

**Education** -> It is either non -graduate or graduate. The assumption I can make is “The probability of clearing the loan amount would be higher if the applicant is a graduate”.

**Self\_Employed** -> As the name suggests Self Employed means, he/she is employed for himself/herself only. So, freelancer or having an own business might come in this category. An applicant who is self-employed is represented by Y and the one who is not is represented by N.

**Applicant Income** -> Applicant Income suggests the income by Applicant. So, the general assumption that i can make would be “The one who earns more have a high probability of clearing loan amount and would be highly eligible for loan”

**Co Applicant income** -> this represents the income of co-applicant. I can also assume that “If co applicant income is higher, the probability of being eligible would be higher “

**Loan Amount** -> This amount represents the loan amount in thousands. One assumption I can make is that “If Loan amount is higher, the probability of repaying would be lesser and vice versa”

**Loan\_Amount\_Term** -> This represents the number of months required to repay the loan.

**Credit\_History** -> A credit history is a record of a borrower’s responsible repayment of debts. It suggests → 1 denotes that the credit history is good and 0 otherwise.

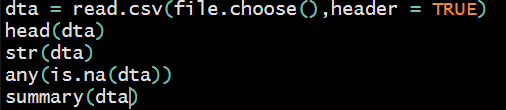
**Property\_Area** -> The area where they belong to is my general assumption as nothing more is told. Here it can be three types. Urban or Semi Urban or Rural

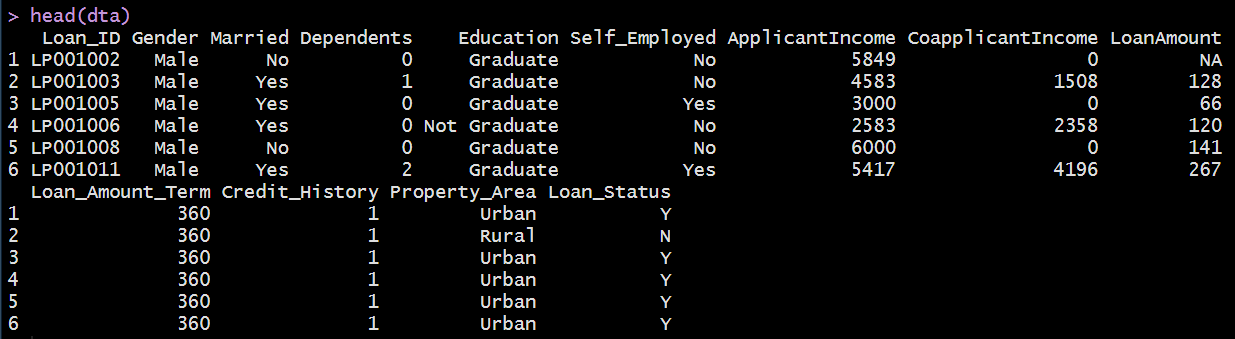
**Loan\_Status** -> If the applicant is eligible for loan it’s yes represented by Y else it’s no represented by N.

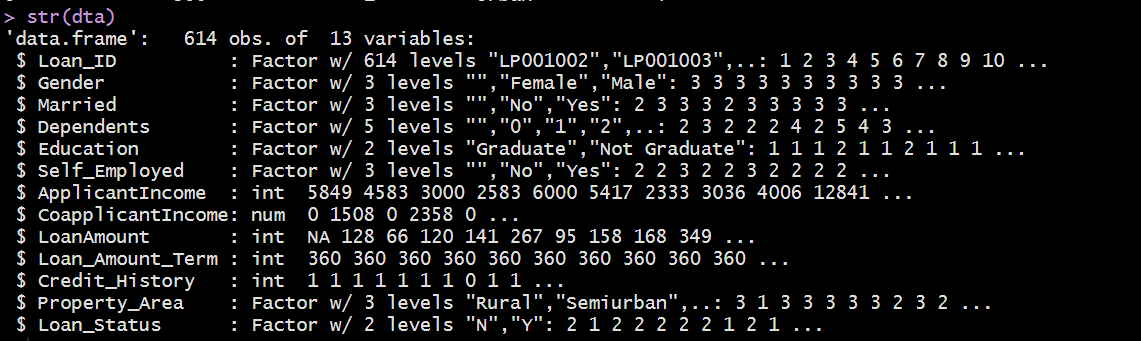
1. **Data Analysis**

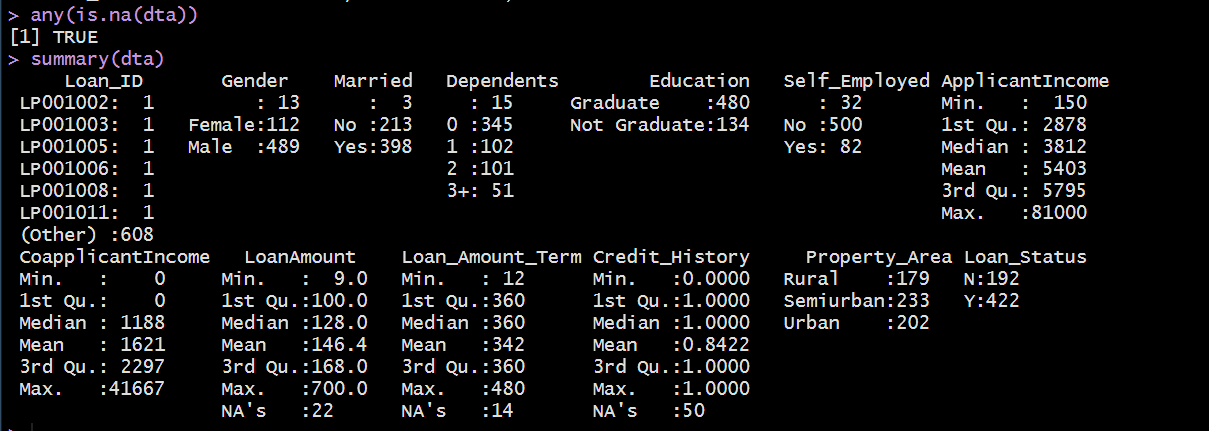
Let’s check if our initial analysis proves correct and also if we are able to predict accurately or not.

**Loading the data:**



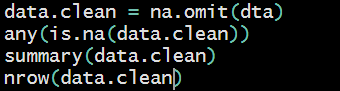


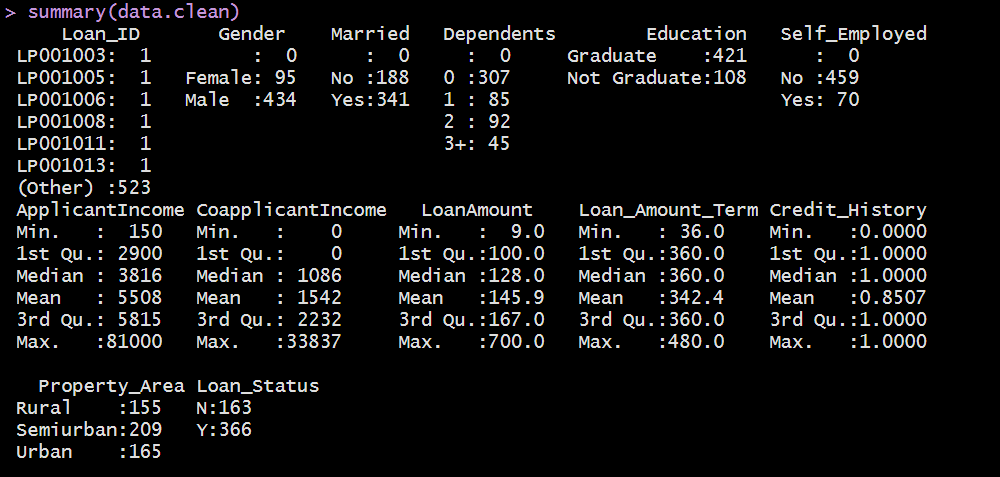




So, we can see that we have some NA values in our dataset, so before proceeding for further analysis, we should clean our data.

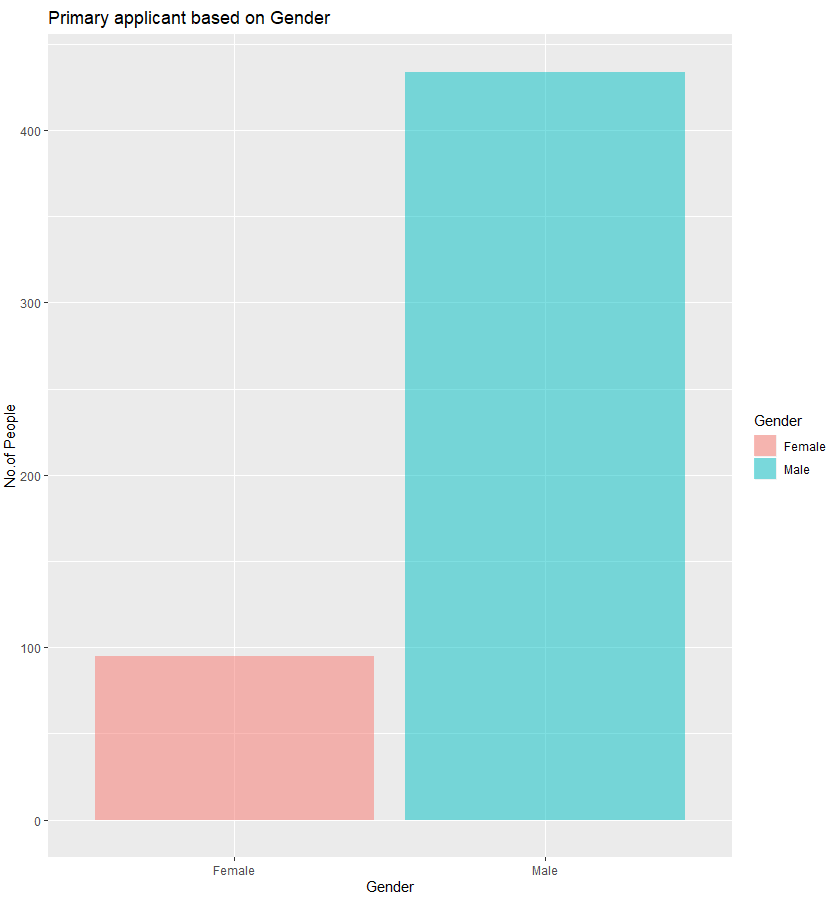
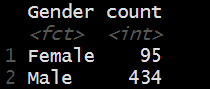
**Data Cleaning:**



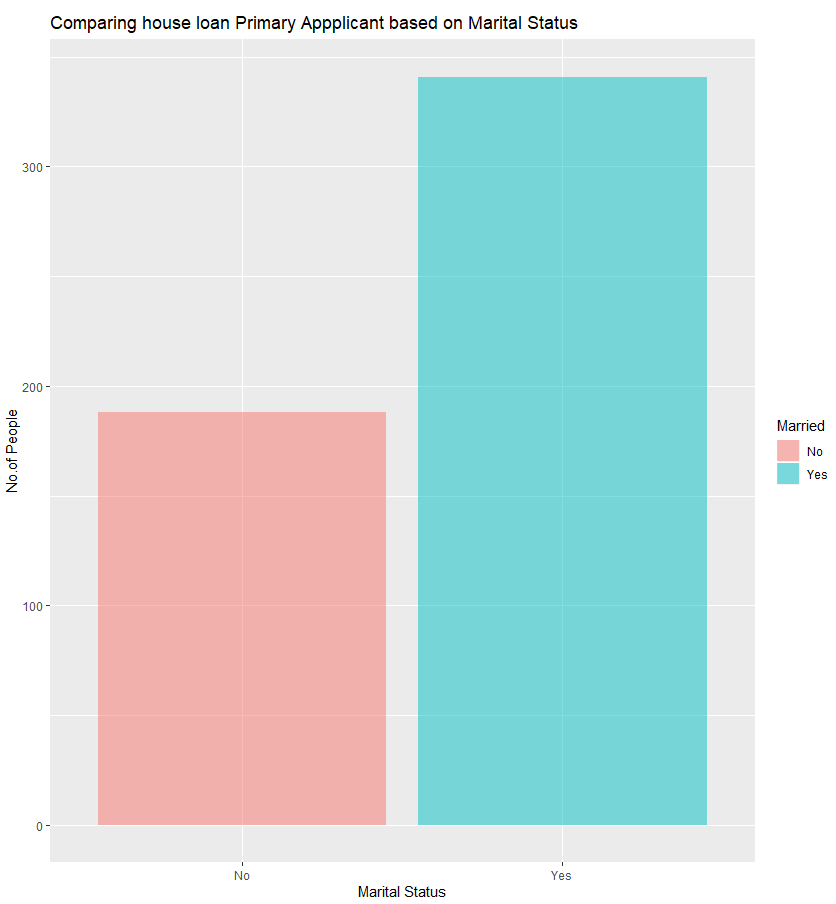
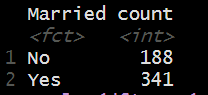


**Exploratory Data Analysis:**

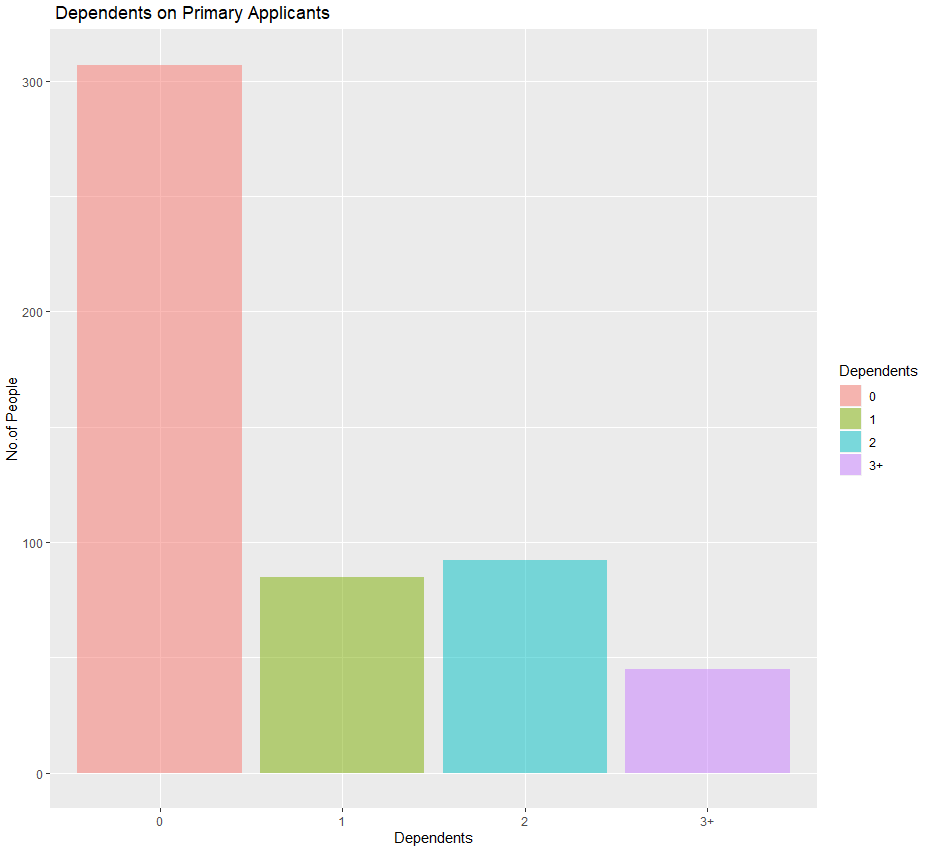
* By looking at the column’s description in the above paragraph, we can make many assumptions like
* The one whose salary is more can have a greater chance of loan approval.
* The one who is graduate has a better chance of loan approval.
* Married people would have an upper hand than unmarried people for loan approval.
* The applicant who has a smaller number of dependents have a high probability for loan approval.
* The lesser the loan amount the higher the chance for getting loan.

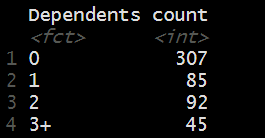


So, from this we can see that there are more primary applicants who are Male than that of females.



We can see that People who have applied for housing loan a greater number of people are married. Let’s see the number of dependents which is also a very big factor in loan approval.

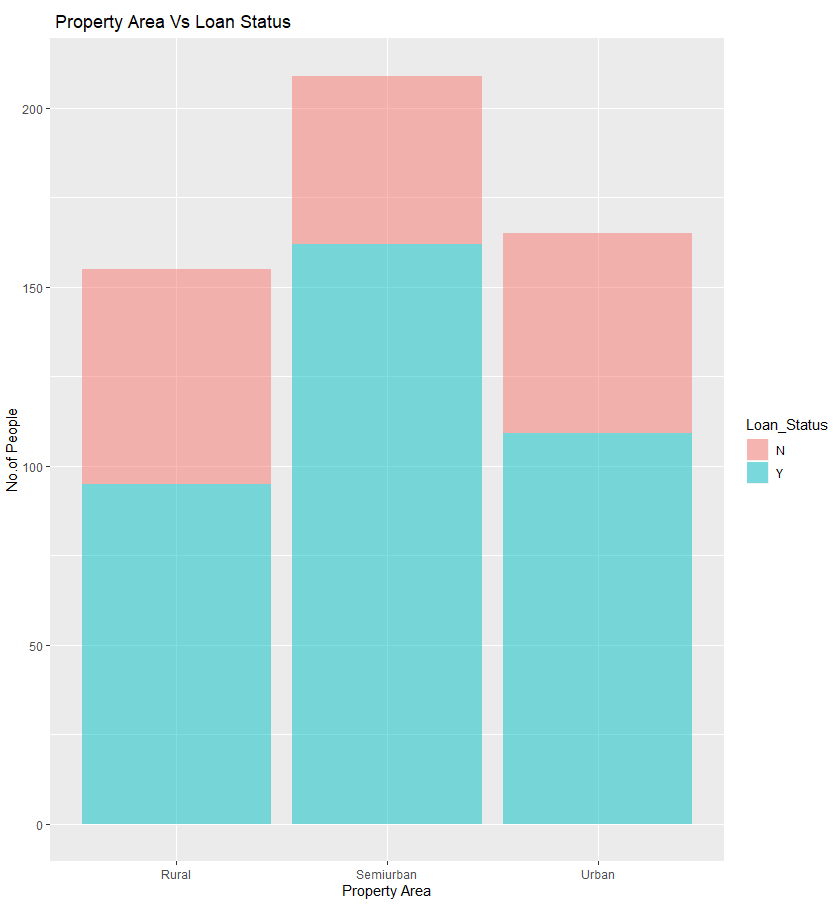
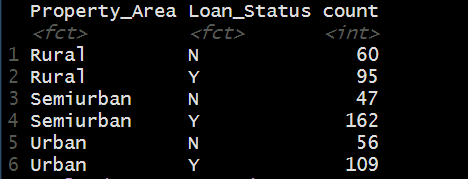




According to initial analysis, it seems like people who are not having any dependents or a smaller number of dependents. However, we will find these factors later in our analysis.

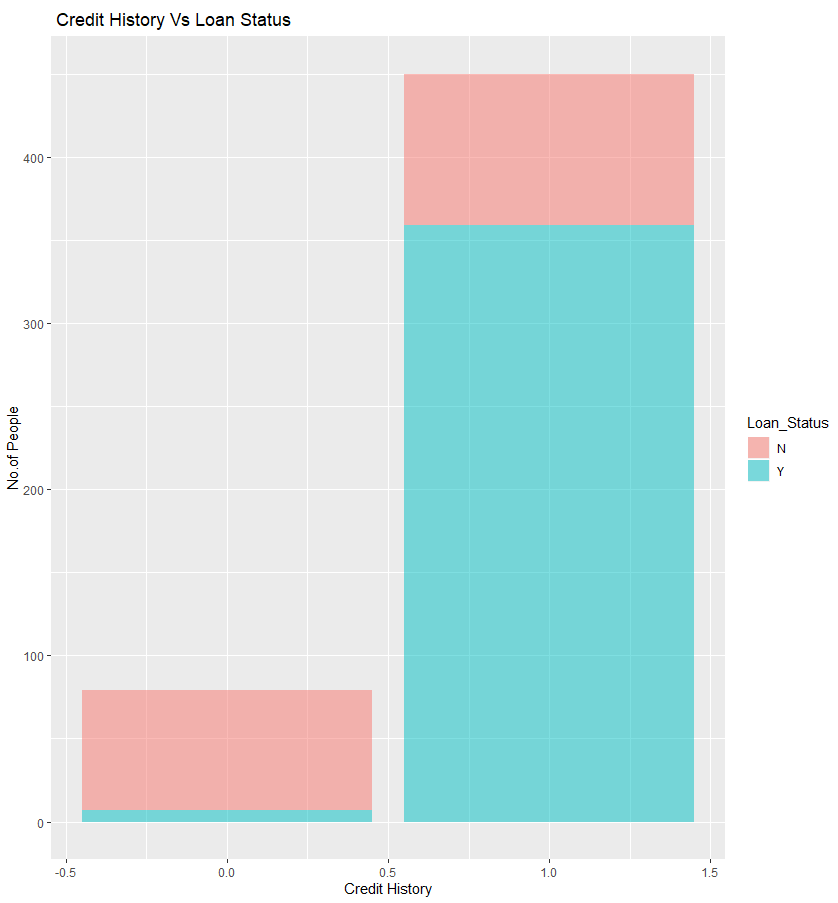
Now, with our exploratory analysis, we will try to find out the relationship of each variable in Loan status.

**Property Area Vs Loan Status**



With this we can see thar People who are living in Semi-Urban are having more chances of getting the loan approved.

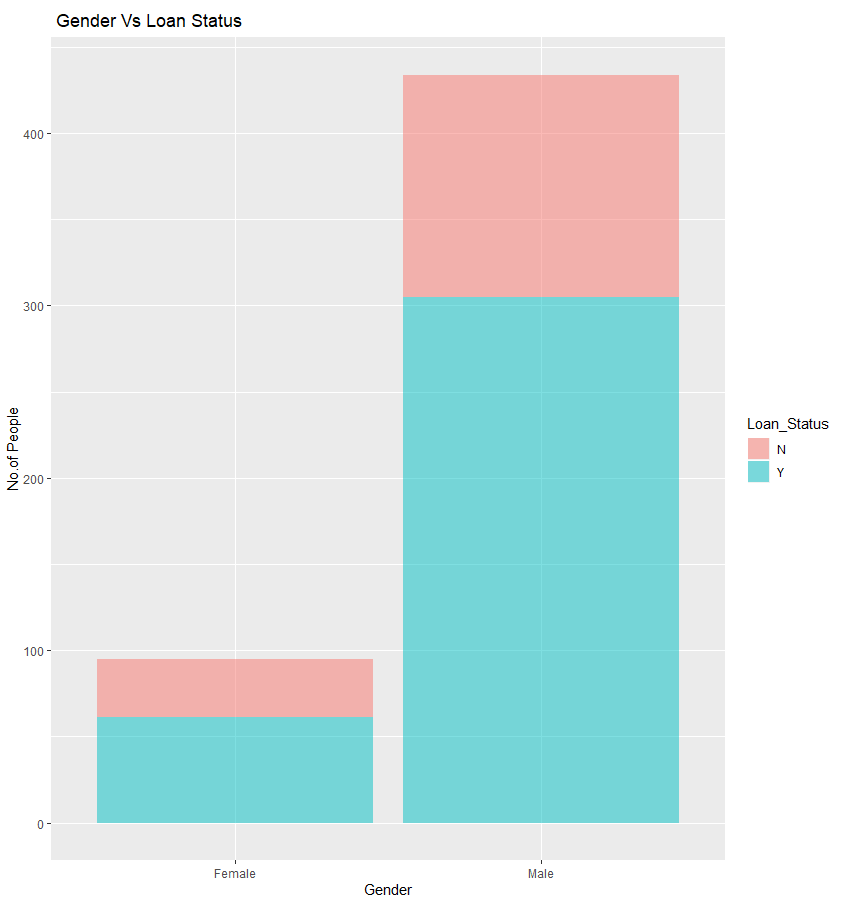
**Credit History Vs Loan Status**



So, we can see that the applicants who are having good credit score are having more chances of getting the loan approved.

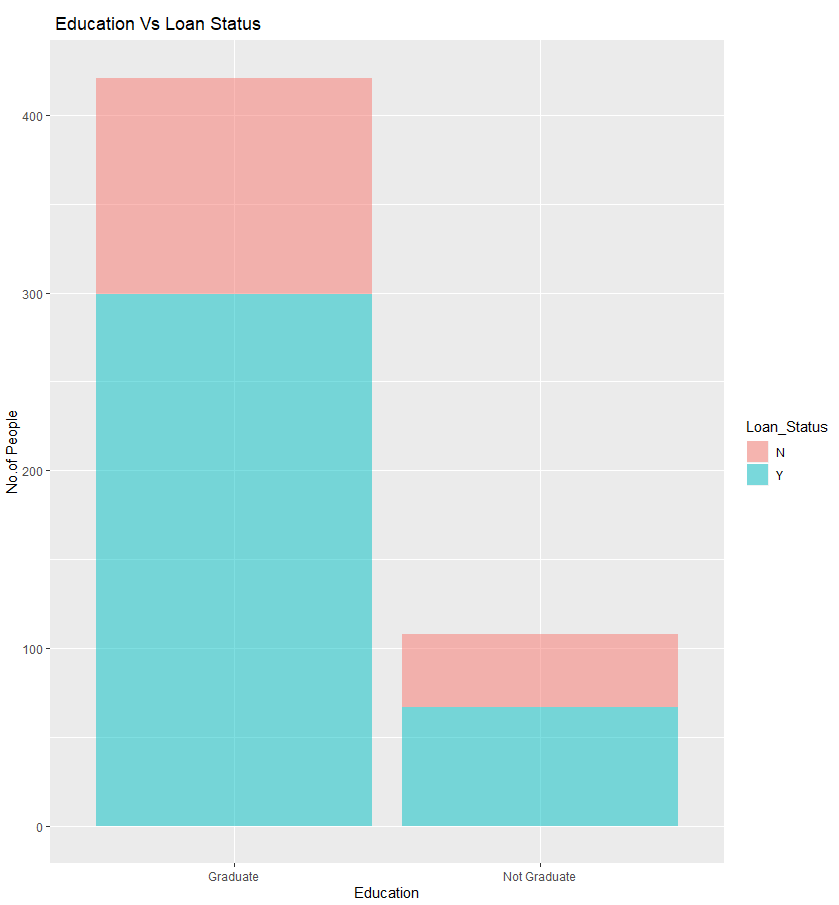
Let’s see whether Gender is having any role to play in Loan approval or not.

**Gender Vs Loan\_Status**



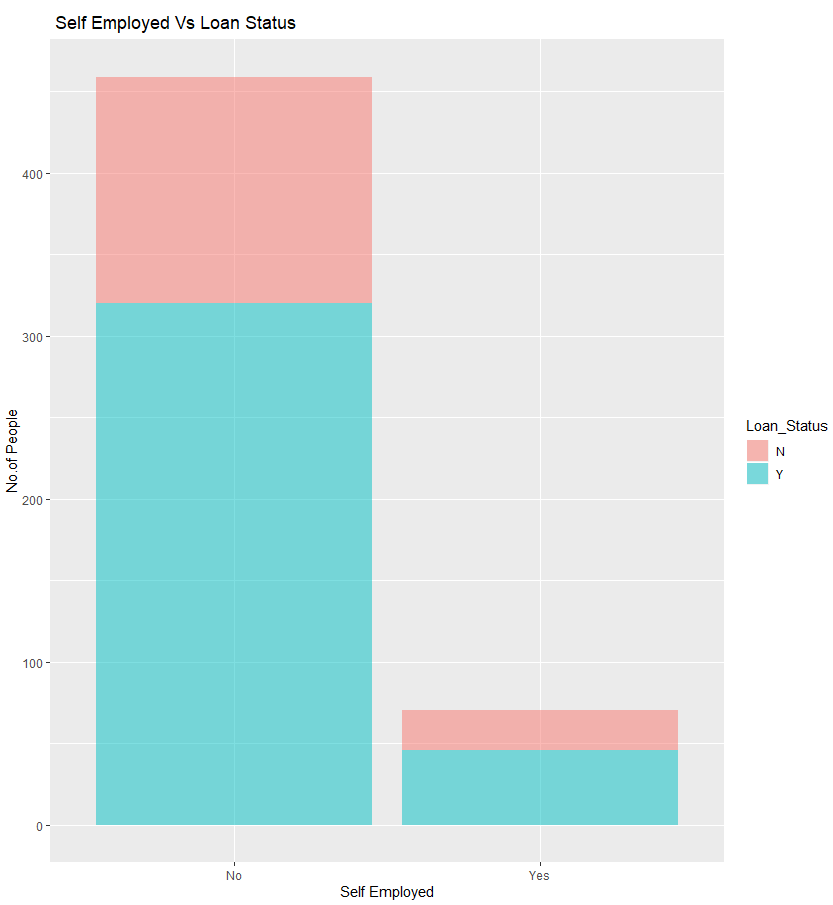
We can see when Males are Primary applicant, chances of getting the loan approved are more.

**Education Vs Loan Status**



As per Education level, People who are graduated are having more chances of getting the loan approved.

**Self Employed Vs Loan Status**



People who are not self employed and doing some fixed job are having greater chances of getting the loan approved.

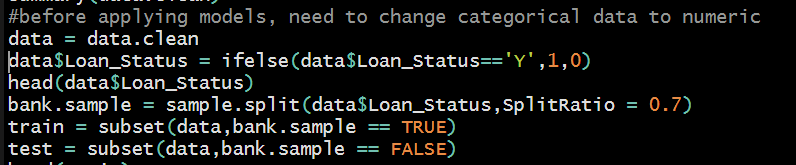
1. **Modelling and Classification**

After getting an overview of our data, we will build models to analyse the accuracy of our models.

Before proceeding for our Modelling, we need to do some more data cleaning and splitting the data into Train and Test and performing further Analysis.

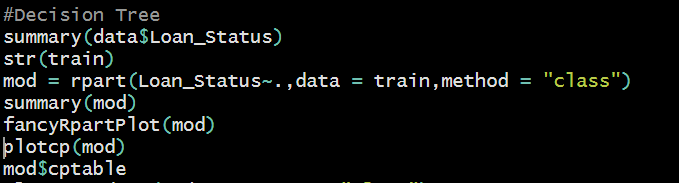
We will apply three models and Test our Prediction.

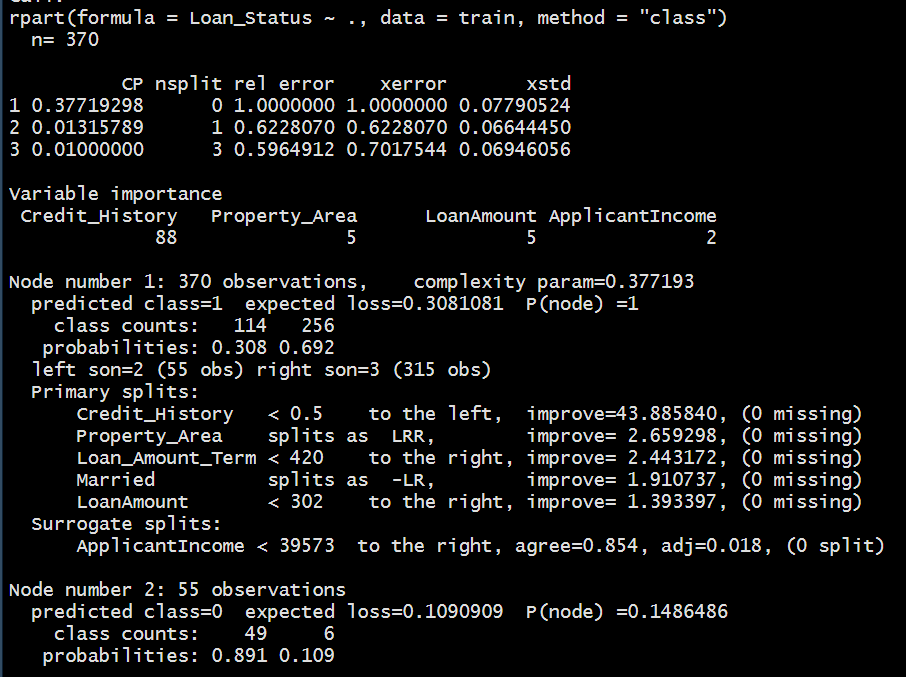
* **DECISION TREE**
* **RANDOM FOREST**
* **NEURAL NETWORK**

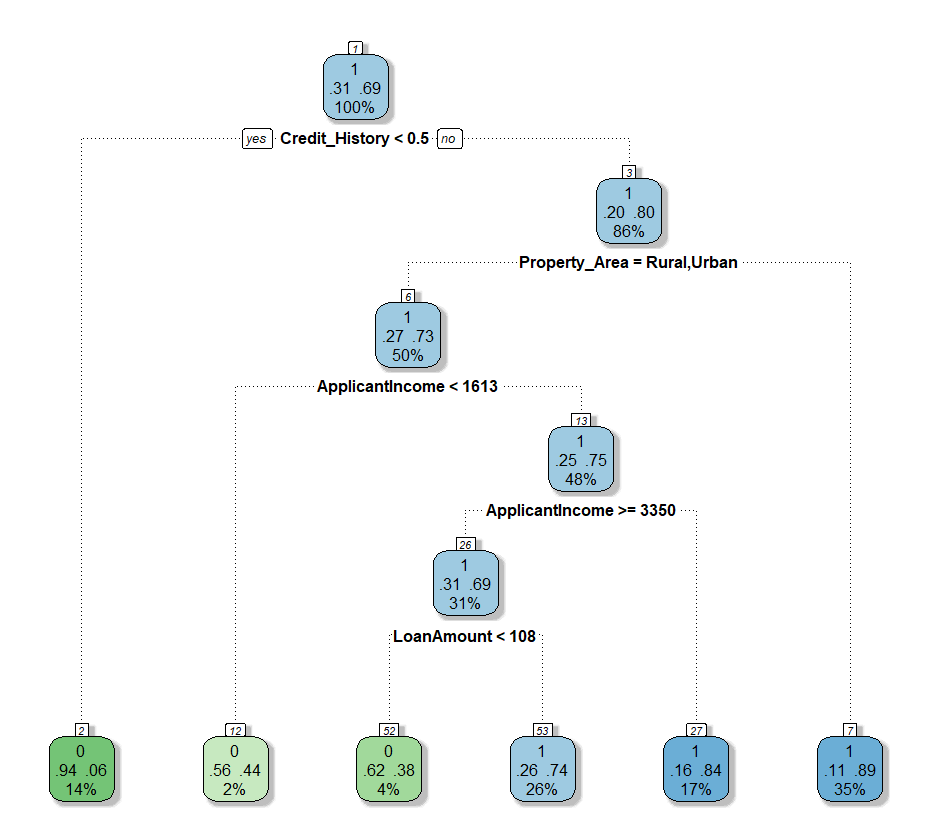


**DECISION TREE**

We will apply our rpart as a part of Decision Tree Modelling. With summary of

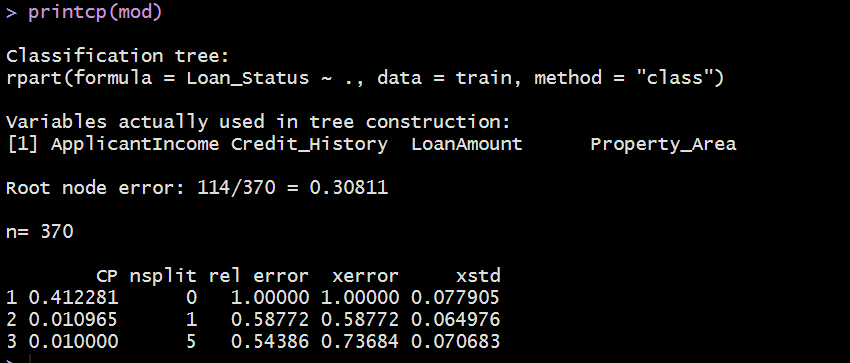






So, here with a graphical representation of Decision Tree we see that at the top of the node we have an approximate percentage of people for home loan got granted and for whom it is not granted. So, we see that for 69% of people Loan got granted whereas for rest 31% of people Loan did get granted. Then after comparing Credit History, People who are having credit history less than 0.5 only 15% chances are there for the loan to get approved. For people who have credit history credit history more than 0.5, for them Property Area is being compared and found out that only 11% chances are there for the people who are staying in rural areas that to for the people who have good Credit score. From this we can find out that Credit History is the first criteria for approval or rejection of approving the loan. Because the Bank or finance company just wants the money back and doesn’t take any risk.

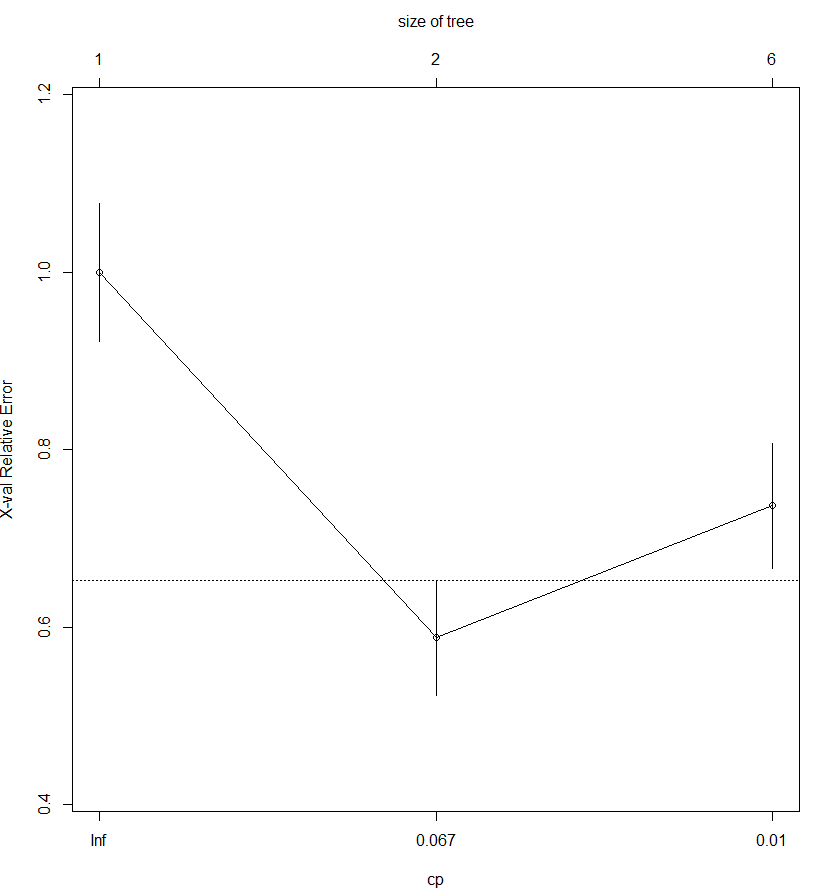
After checking the classification of decision tree, we check if this is the desired decision tree which we need, just to avoid overfitting the model. For that we check CP (Complexity Parameter) of Decision Tree.



We see that we have almost 30% root node error in our classification tree. We need to find optimum number of CP value in order to avoid overfitting the model.

From the above-mentioned list of cp values, we can select the one having the least cross-validated error and use it to prune the tree.

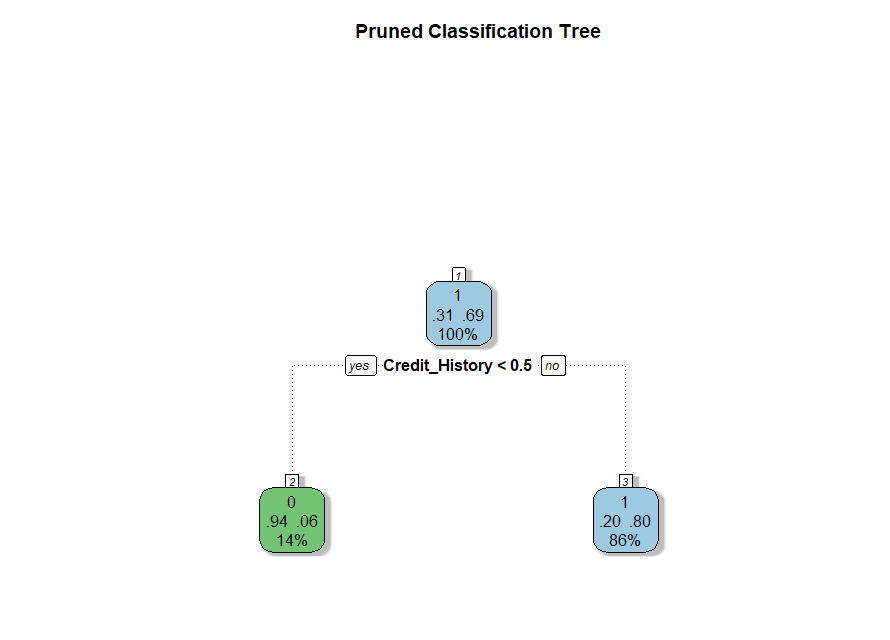
Let’s first plot our CP values:



Plotcp() provides a graphical representation to the cross validated error summary. The cp values are plotted against the geometric mean to depict the deviation until the minimum value is reached.

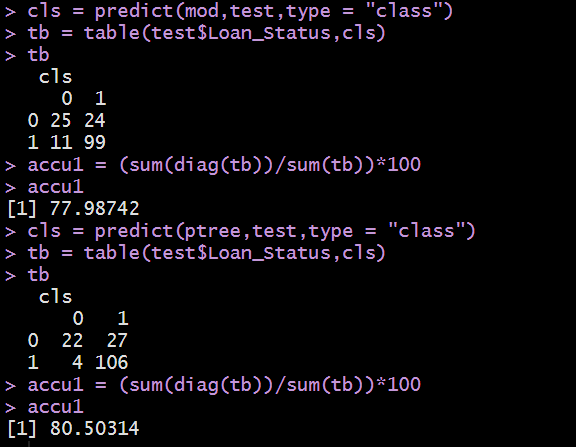
Prune the tree to create an optimal decision tree:





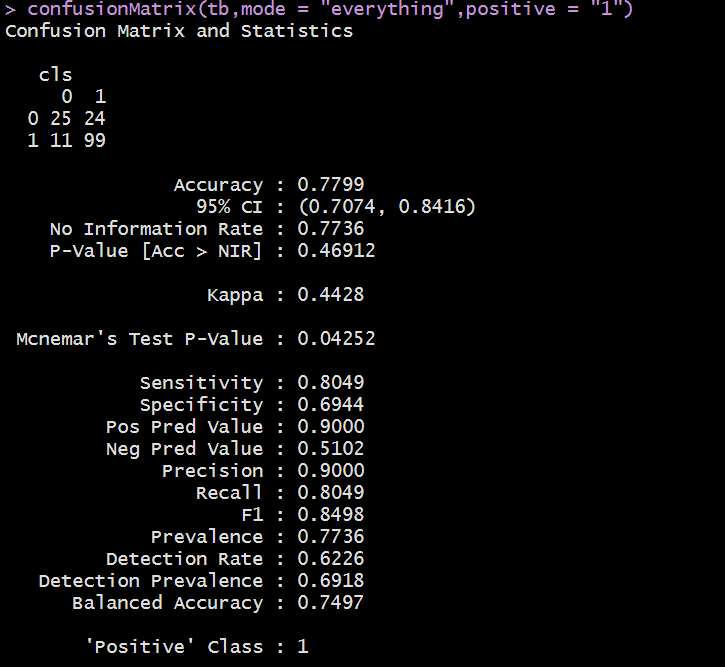
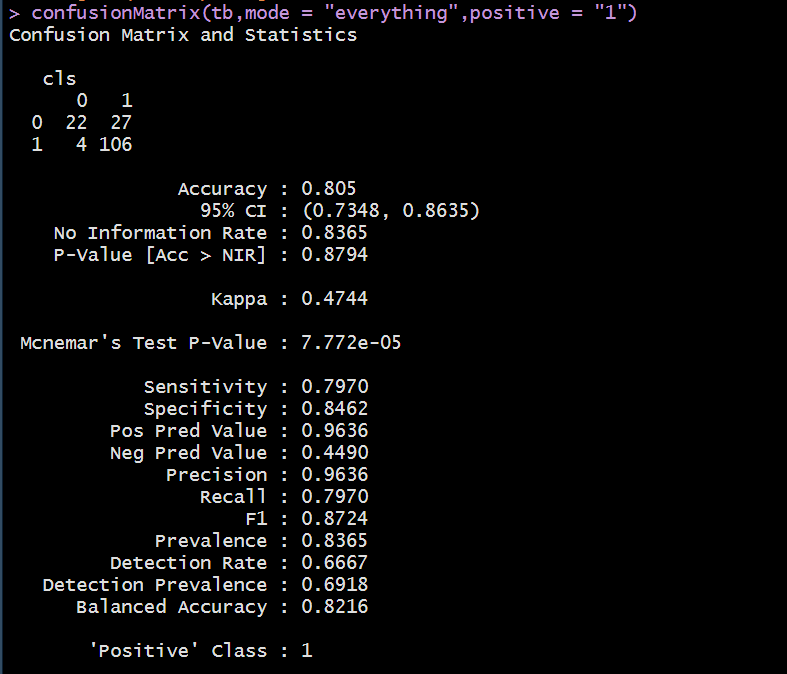
As concluded, in earlier decision tree, that credit history is most important factor which was only deciding the initial stage if the loan is getting approved or not. So, here it shows that people maintaining good credit history have 86% chance of the loan getting approved.

Now, we will check how much our accuracy improves by using a pruned tree instead of our previous model.



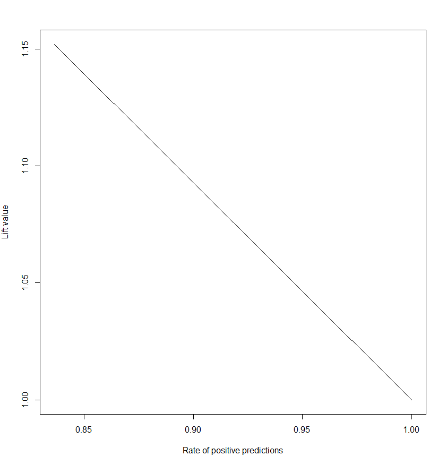
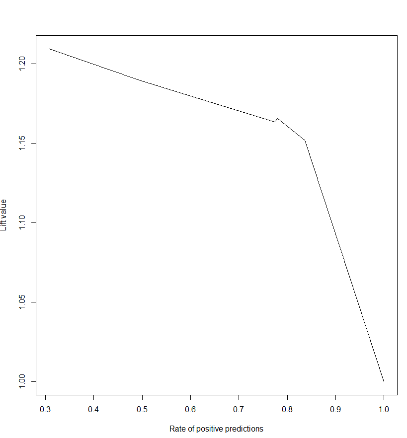
So, we can see that pruning the tree improves our accuracy rate from 77% to 81% (approx.). Our main motive is to get higher accuracy rate to automate the model to be used by housing finance company.

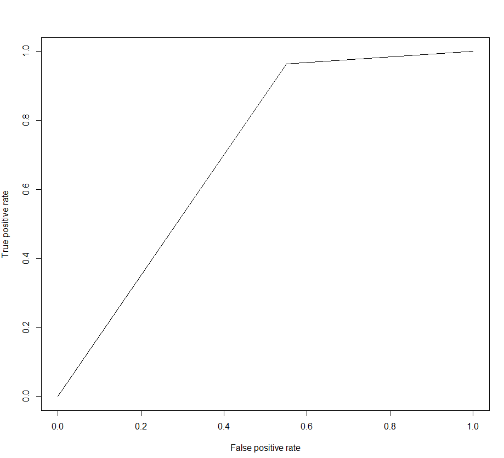
Let’s check the precision of our prediction which is more important than being accurate.

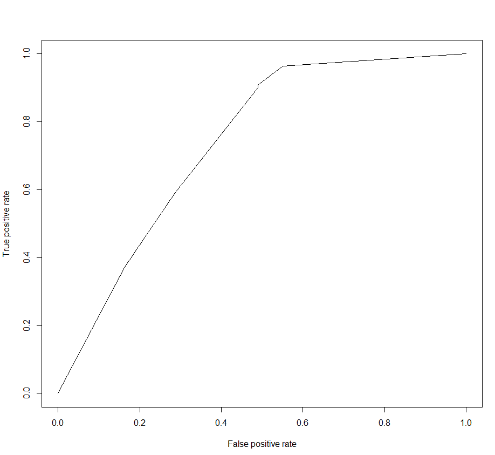


We see that our Precision without pruning is 90% whereas with Pruning it is 96%.

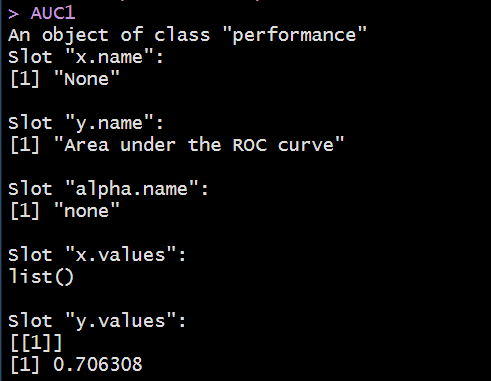
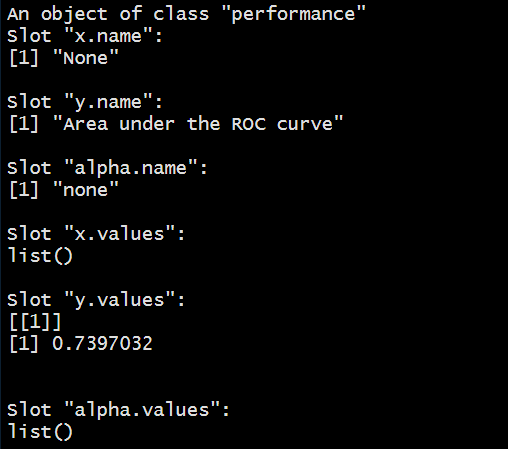
Lift curve without Pruning Lift curve with Pruning



 ROC Curve without Pruning ROC Curve with Pruning



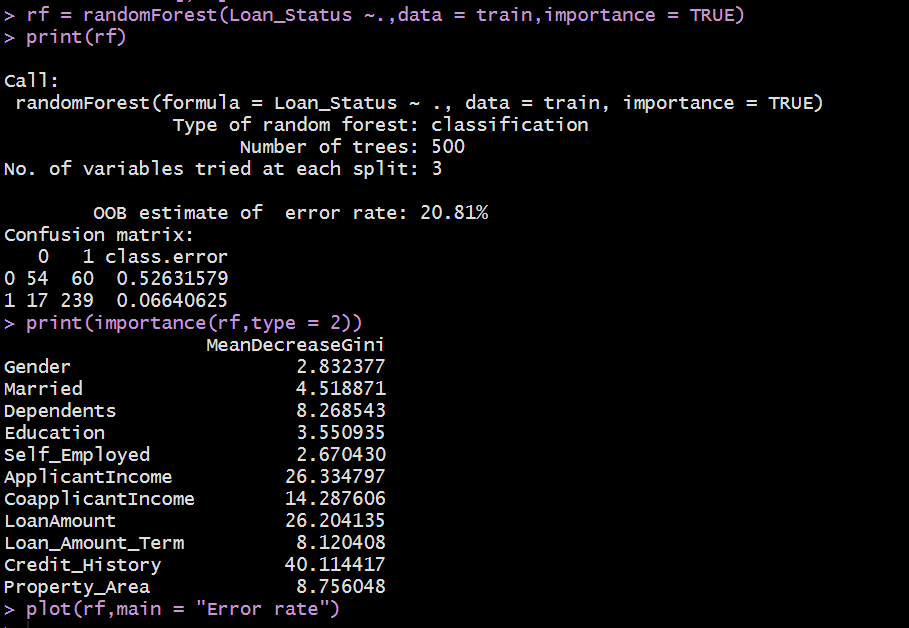
AUC Value without Pruning AUC value with Pruning



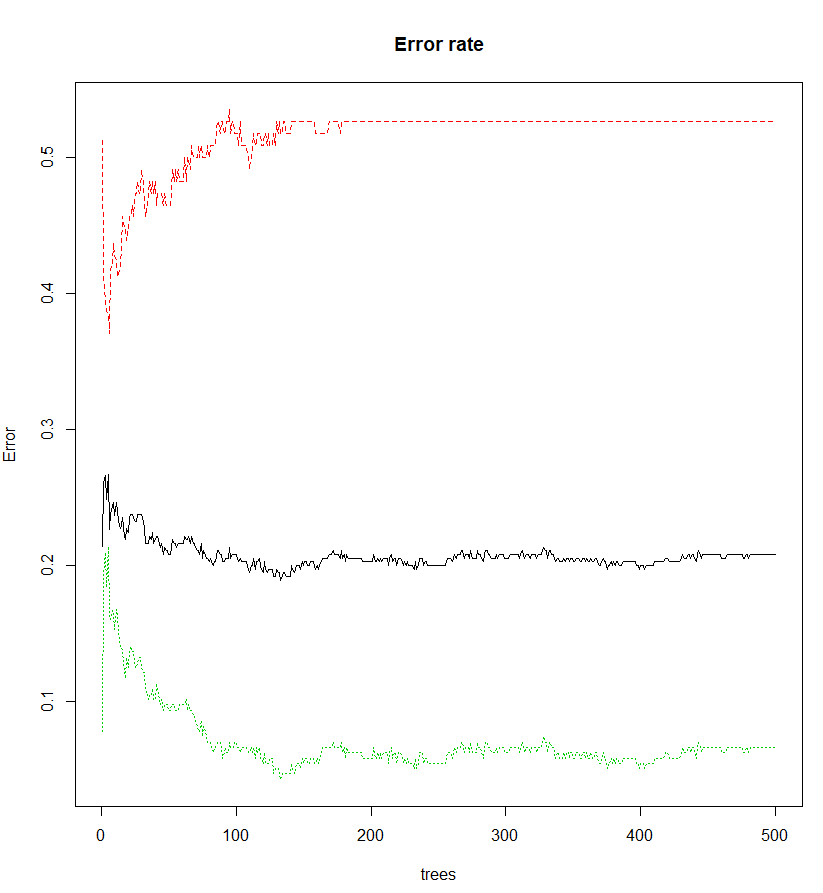
AUC value of a Pruned tree degrades a bit but can be managed.

Now, we will try another model which is an advanced form of decision tree, random forest is a collection of 500 trees. The more the number of trees, more less the error rate. Let’s try and check if random forest accuracy is same as decision tree or more.

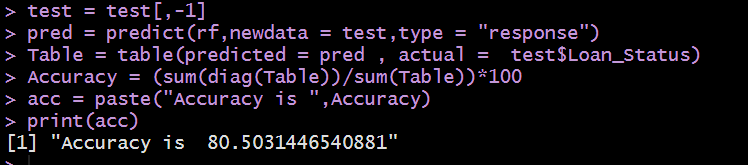
**RANDOM FOREST**



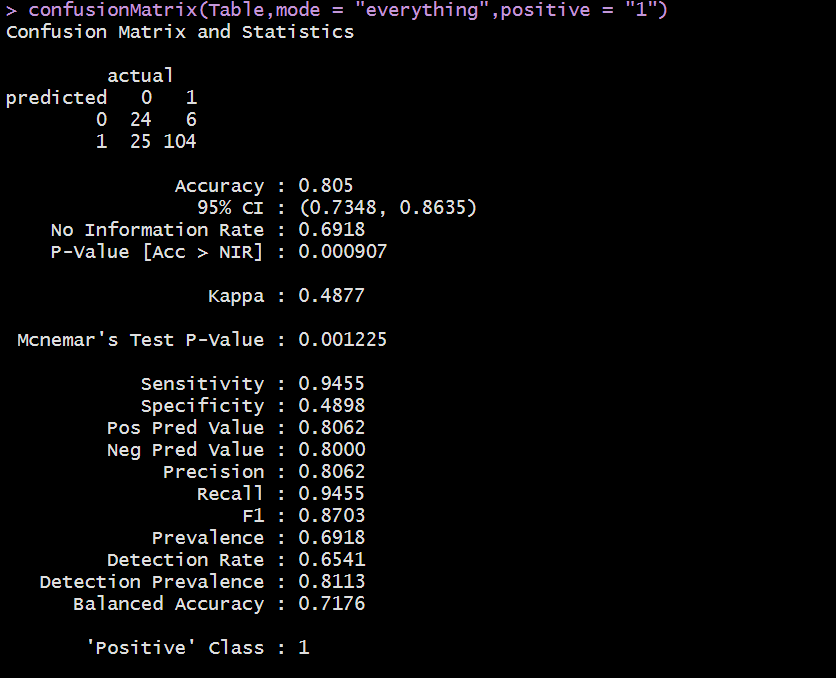
Random forest collection of 500 trees which eliminates the use of pruning the tree as we can see that our accuracy will be approx. 80% which we got in decision tree after Pruning.

We can also get the importance of each variable in our model, where we can see that Credit history is having the highest importance, which means that it plays a vital role in deciding whether loan will be approved or not.

So, we can see here that how with the increasing number of trees, the error keeps decreasing but at a particular point after which the error almost remains the same. As here we can see that after about 150 trees our error remains constant which means that instead of 500 Trees if we use 150 trees also in this dataset, our accuracy will be almost same.



Our Prediction accuracy is 80.50% which is similar to Decision Tree after pruning. Let’s check the Precision of our model which is also an important aspect in prediction.



Our Precision, Recall and F1 score are very good.

So, we can conclude that we will get almost same accuracy or even more whether we use Decision tree with Pruning and Random forest. Hence, we can replace Decision Tree with Random Forest.

Let’s check another model which is Neural Network and validate, how does it perform.

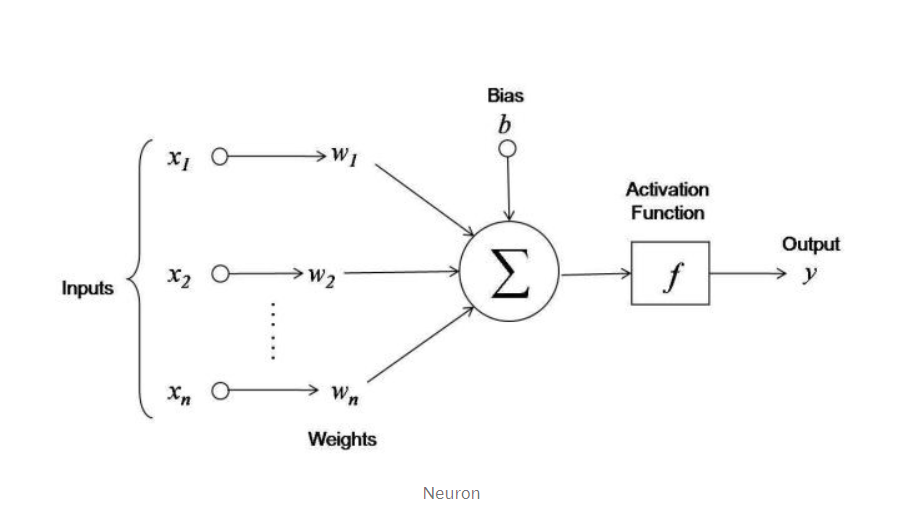
**NEURAL NETWORK**

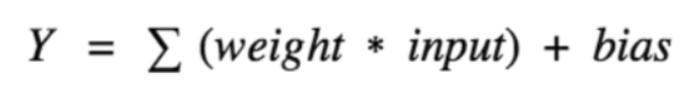
Before proceeding with our classification using Neural Network. We will have a brief overview of about Neural Network and its working.

Neural Network (or Artificial Neural Network) has the ability to learn by examples. ANN is an information processing model inspired by the biological neuron system. It is composed of a large number of highly interconnected processing elements known as the neuron to solve problems. It follows the non-linear path and process information in parallel throughout the nodes. A neural network is a complex adaptive system. Adaptive means it has the ability to change its internal structure by adjusting weights of inputs. The neural network was designed to solve problems which are easy for humans and difficult for machines such as identifying pictures of cats and dogs, identifying numbered pictures. These problems are often referred to as pattern recognition. Its application ranges from optical character recognition to object detection.

Like, in Human body Dendrites receive signals from other neurons. Cell body sums all the inputs signals to generate output. Axon through output When the sum reaches to a threshold. Synapses is a point of interaction neurons. It transmits electrical or chemical signals to another neuron. Synapse is derived from the Greek word which means conjunction.

Similarly, in Neural Network also act similarly,





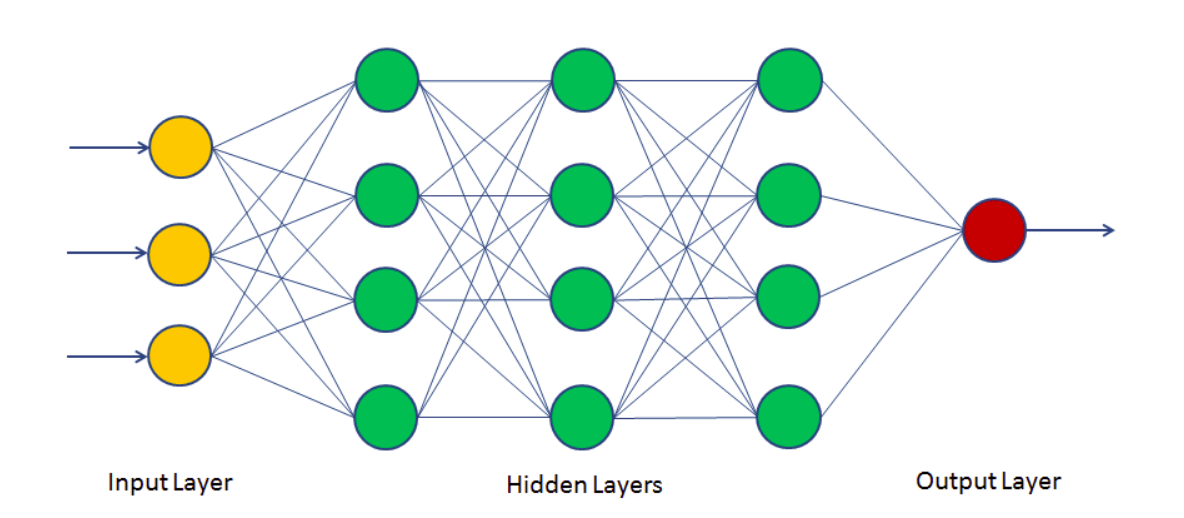
Here, x1, x2....xn are input variables. w1, w2....wn are weights of respective inputs. b is the bias, which is summed with the weighted inputs to form the net inputs. Bias and weights are both adjustable parameters of the neuron. Parameters are adjusted using some learning rules. The output of a neuron can range from -inf to +inf. The neuron doesn’t know the boundary. So, we need a mapping mechanism between the input and output of the neuron. This mechanism of mapping inputs to output is known as Activation Function.

We have two types of Neural Network:

**Feedforward and Feedback Artificial Neural Networks**

There are two main types of artificial neural networks: Feedforward and feedback artificial neural networks. Feedforward neural network is a network which is not recursive. Neurons in this layer were only connected to neurons in the next layer, and they are don't form a cycle. In Feedforward signals travel in only one direction towards the output layer.

Feedback neural networks contain cycles. Signals travel in both directions by introducing loops in the network. The feedback cycles can cause the network's behaviour change over time based on its input. Feedback neural network also known as recurrent neural networks.



**Activation Functions**

Activation function defines the output of a neuron in terms of a local induced field. Activation functions are a single line of code that gives the neural nets non-linearity and expressiveness. There are many activation functions. Some of them are as follows:

**Identity function** is a function that maps input to the same output value. It is a linear operator in vector space. Also, known straight line function where activation is proportional to the input.

In **Binary Step Function**, if the value of Y is above a certain value known as the threshold, the output is True (or activated), and if it’s less than the threshold, then the output is false (or not activated). It is very useful in the classifier.

**Sigmoid Function** called S-shaped functions. Logistic and hyperbolic tangent functions are commonly used sigmoid functions. There are two types of sigmoid functions.

**Binary Sigmoid Function** is a logistic function where the output values are either binary or vary from 0 to 1.

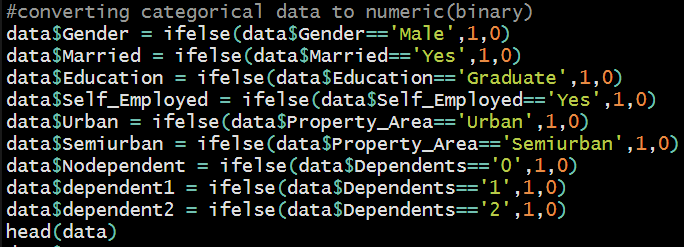
**Bipolar Sigmoid Function** is a logistic function where the output value varies from -1 to 1. Also known as Hyperbolic Tangent Function or tanh.

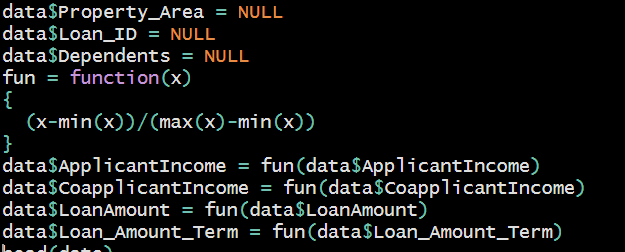
**Ramp Function:** The name of the ramp function is derived from the appearance of its graph. It maps negative inputs to 0 and positive inputs to the same output.

**ReLu** stands for the rectified linear unit (ReLU). It is the most used activation function in the world. It output 0 for negative values of x.

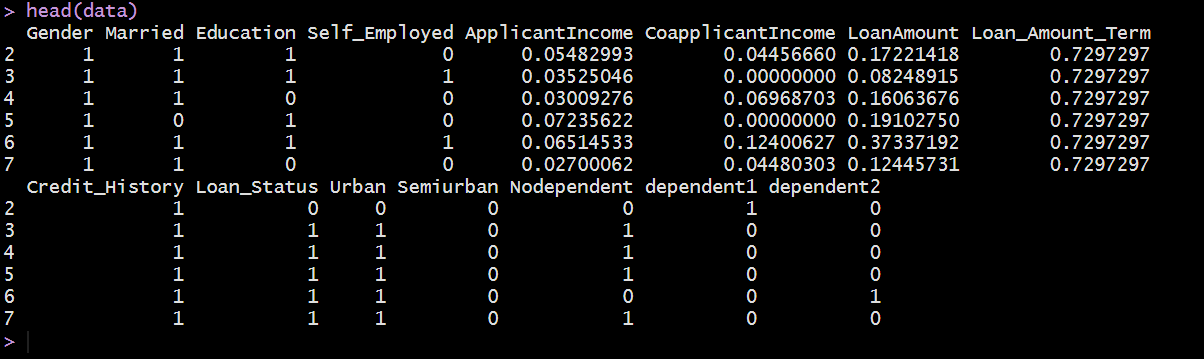
After a brief introduction, we will analyse our Neural network model:

Before, applying Neural Network we need to change our data to numeric data:

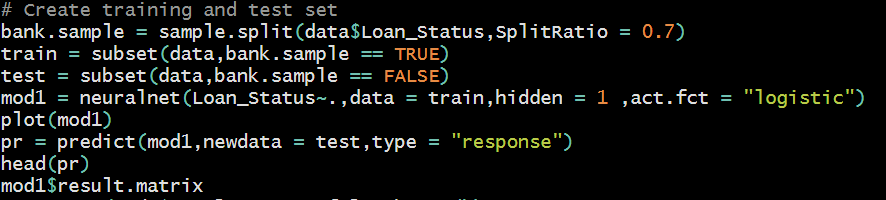


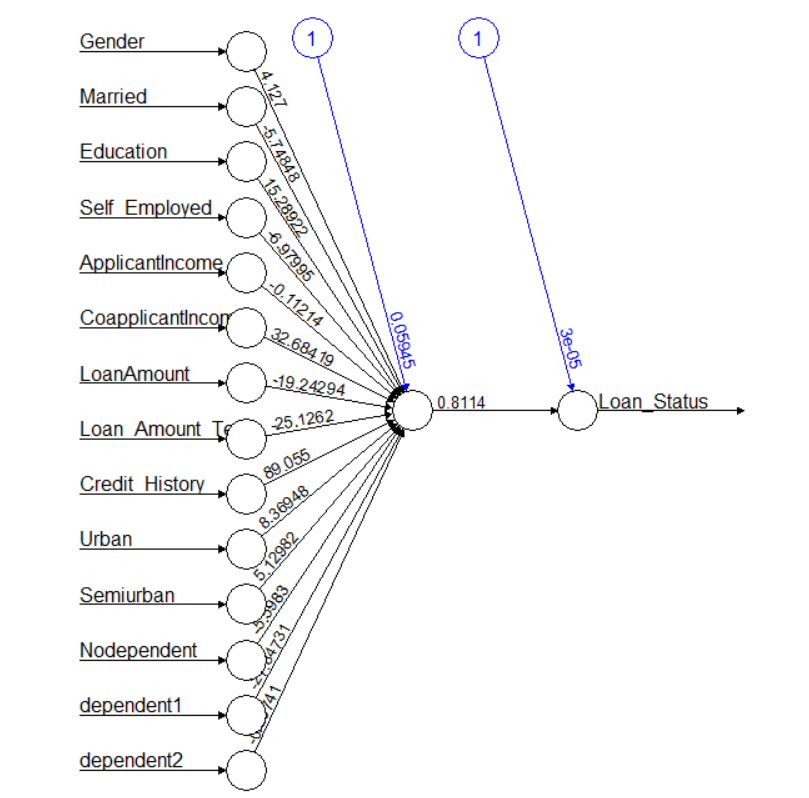


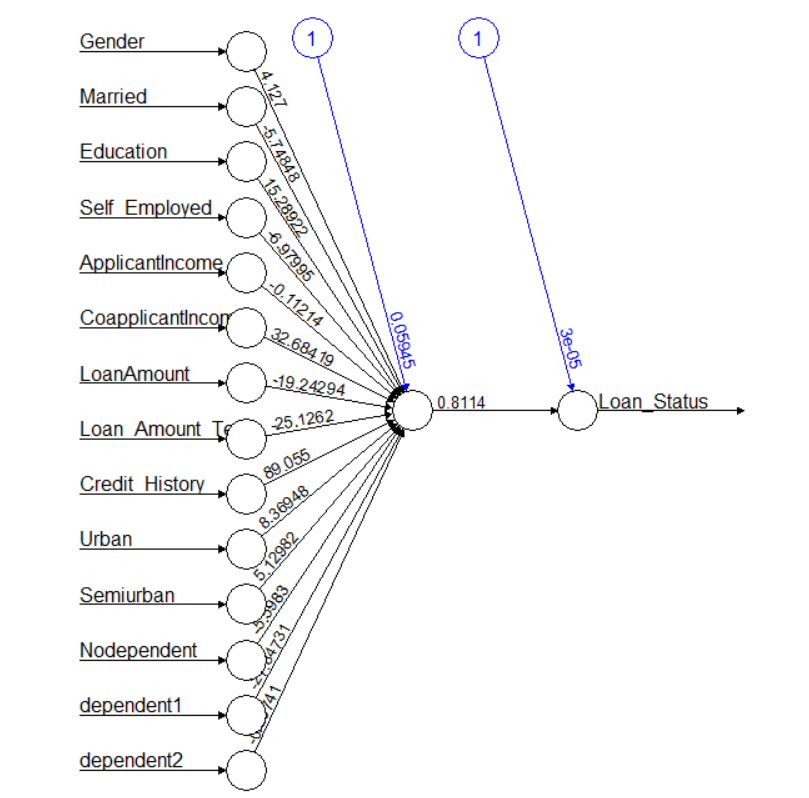
After changing our data to Numeric data, need to scale other numeric data.

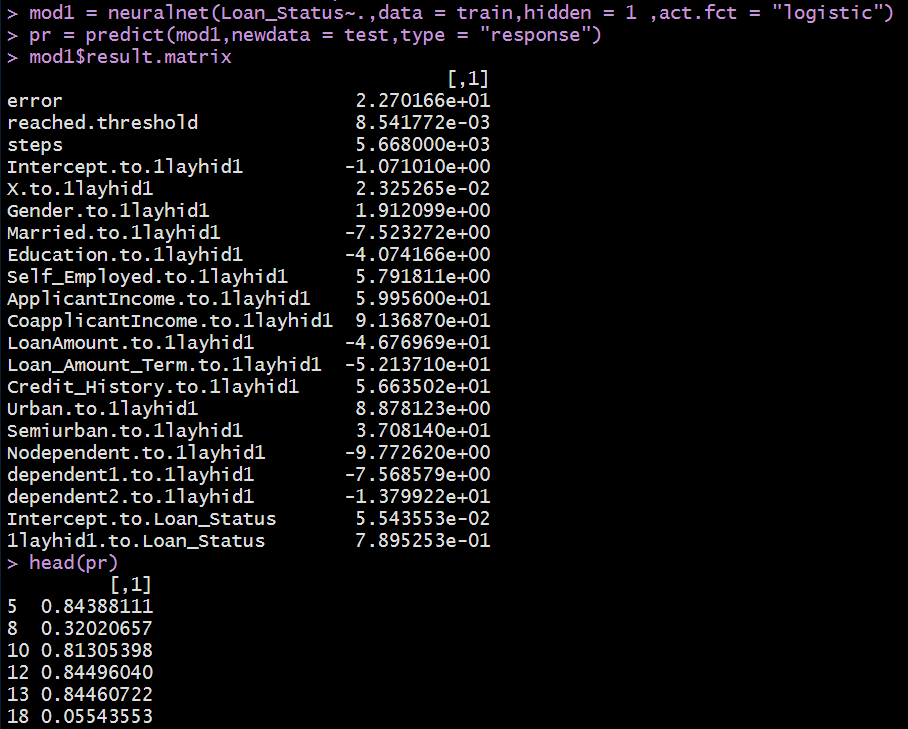


After preparing the data, need to split our data into training and testing .



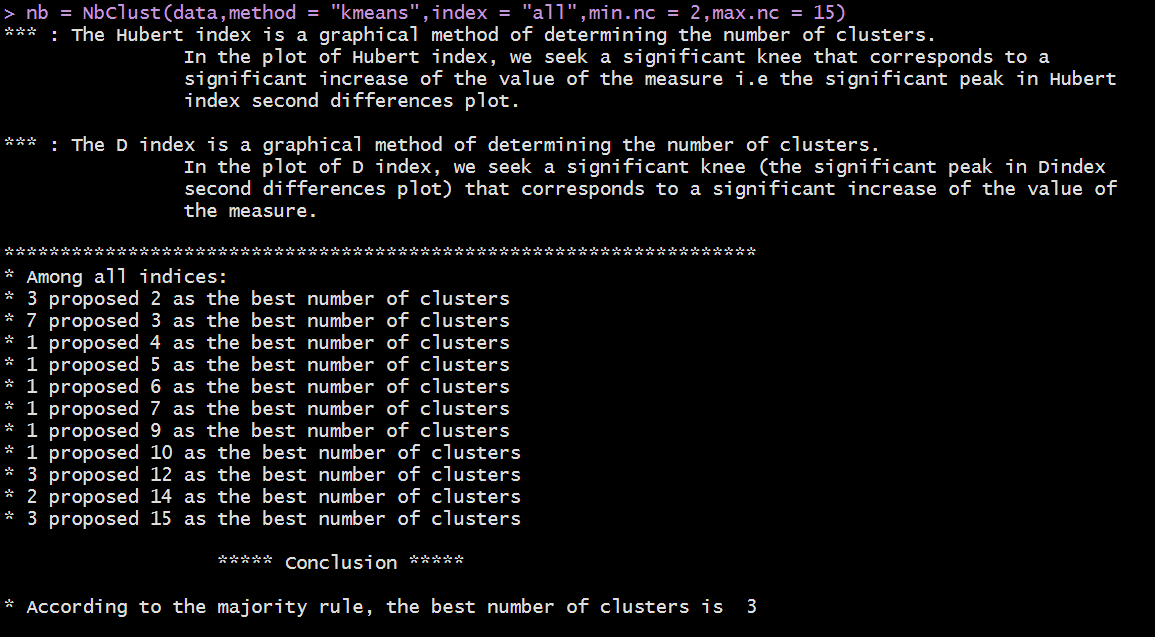
In our Neural network, we choose our target variable, and use train data. We should choose hidden node = 1 and our activation function we choose as logistic. After running the model, let’s check our model:

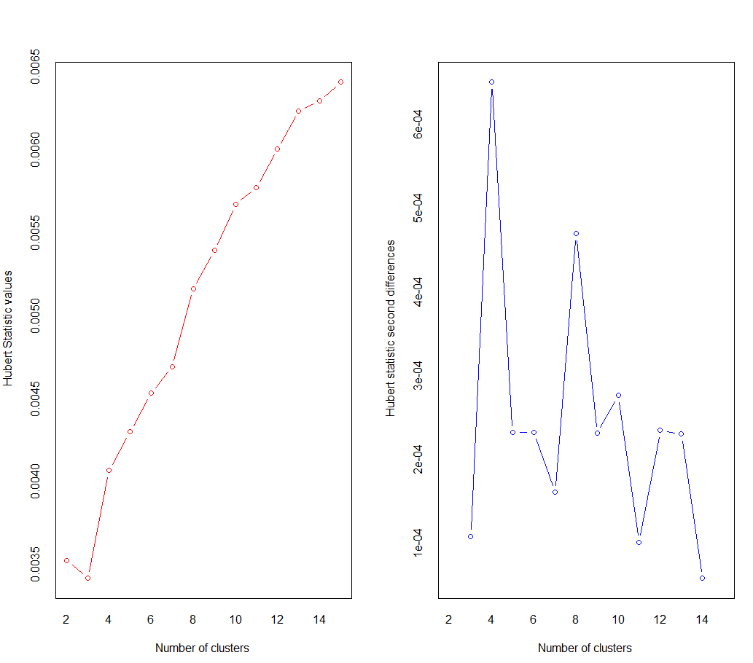


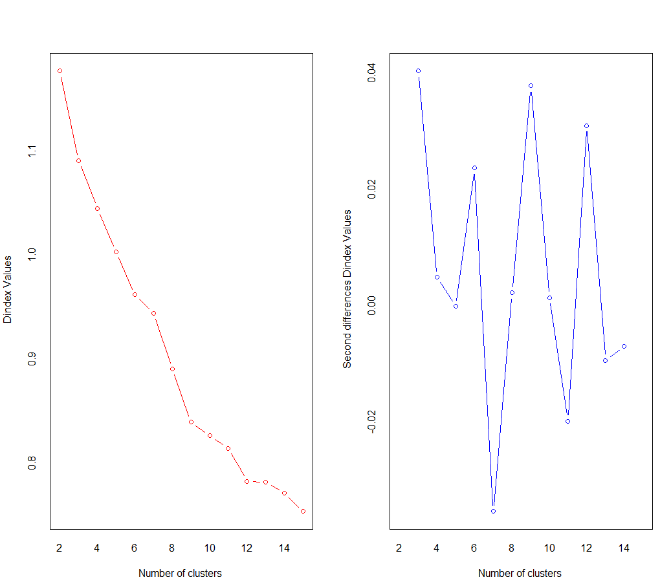


So, we try to predict the accuracy: We see that we have 22.70% error rate, that means 78% accuracy.

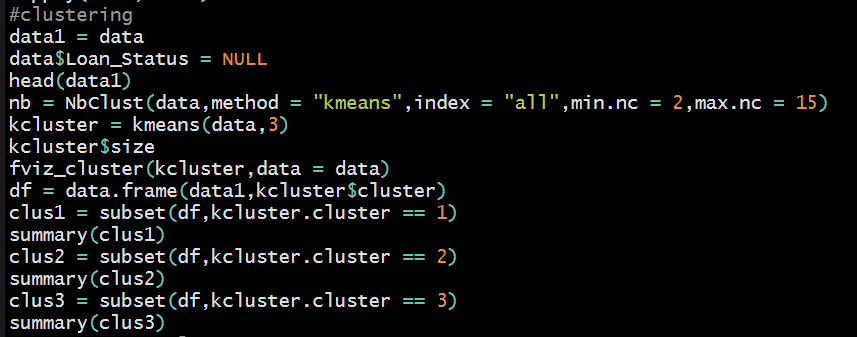
1. **Clustering and segmentation**



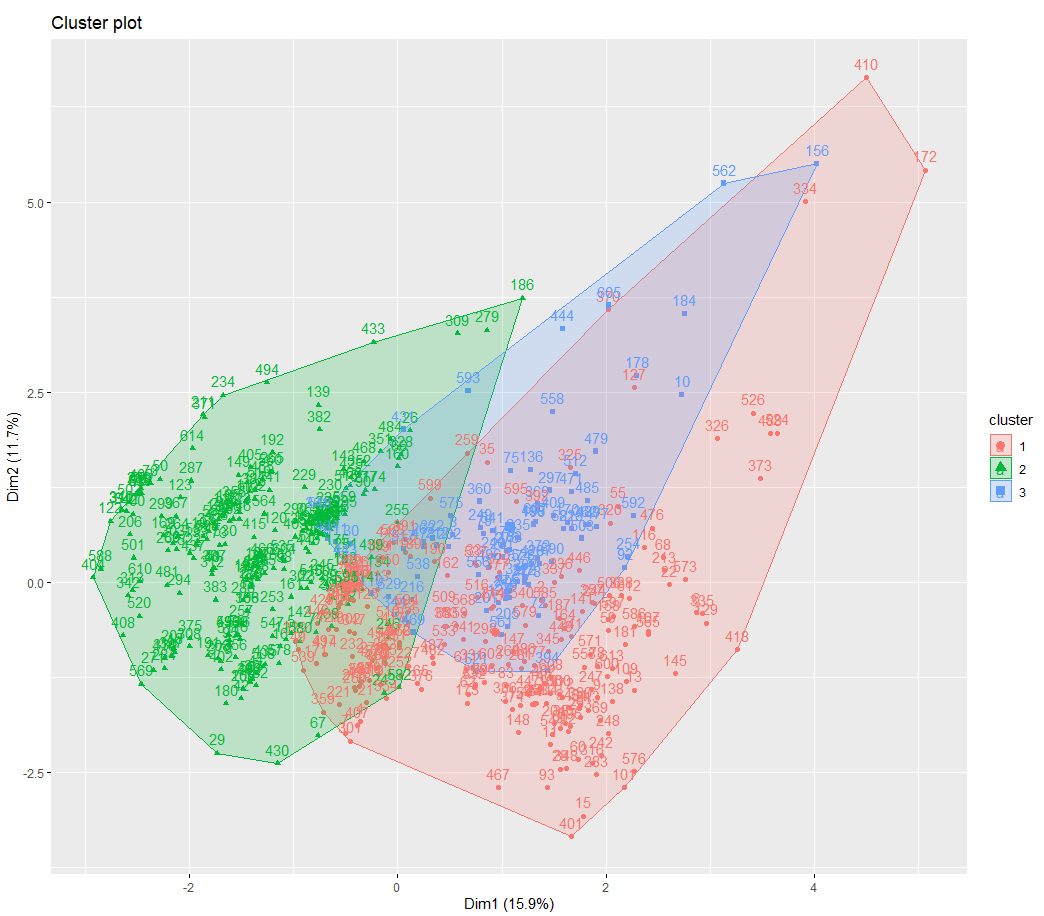
Hubert Dindex



So, after analysing Nbclust we find that our optimal number of clusters are 3.

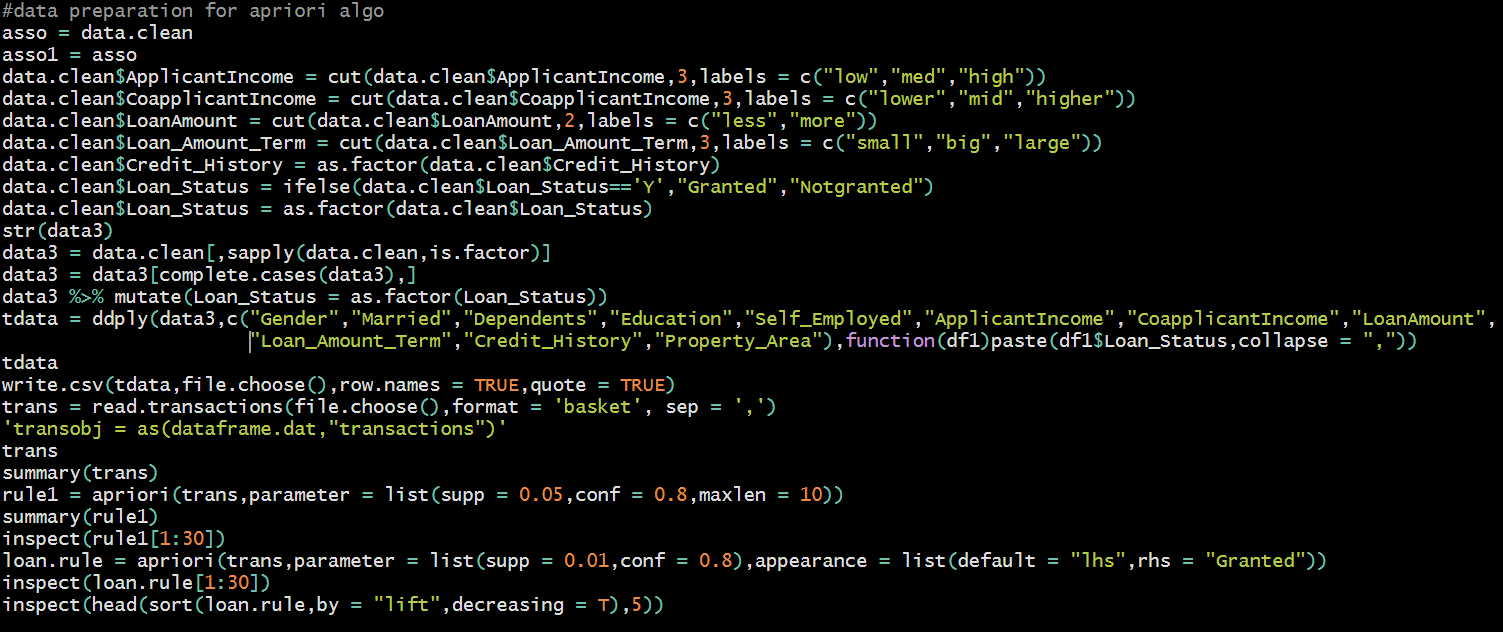


Now, using Kmeans we cluster in our data and also analyse the cluster

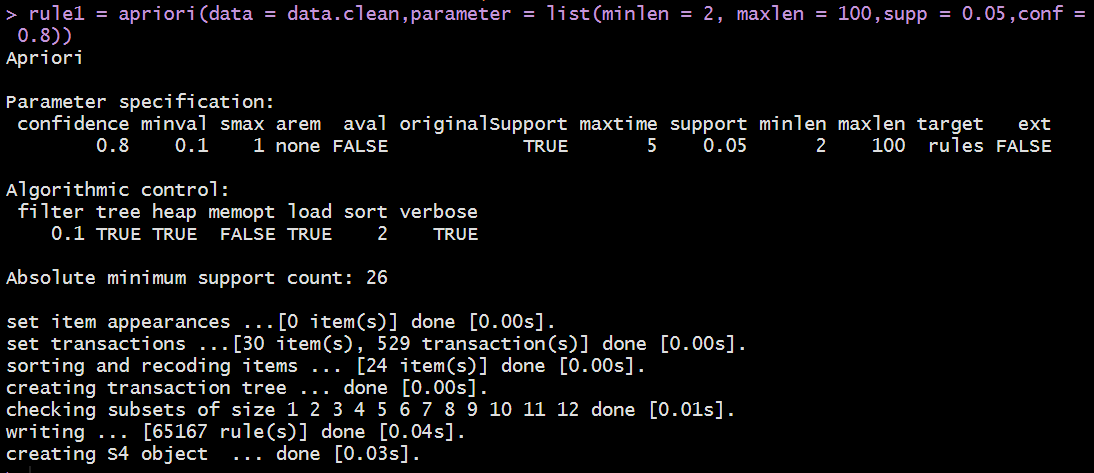


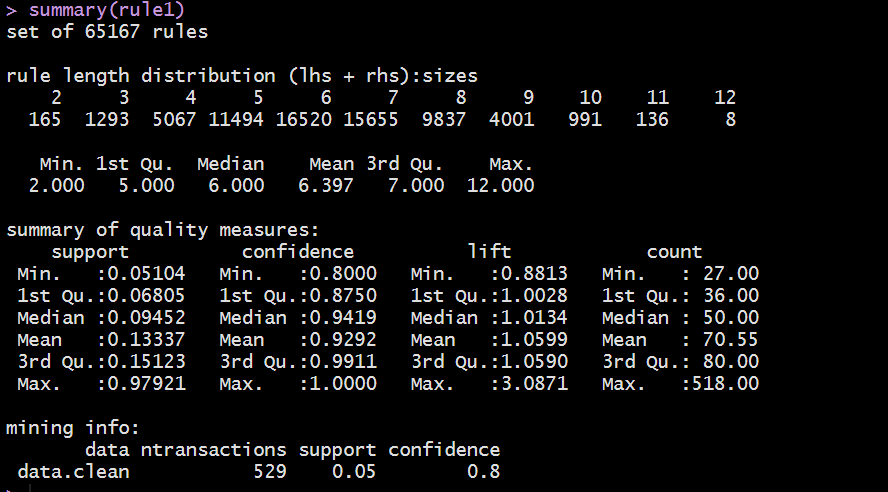
To, analyse each cluster first we need to build association between each variable, so that we can find that which cluster will have most chances of getting the loan approved.

Before building association rules, we need to prepare our data as factors in order to apply association.

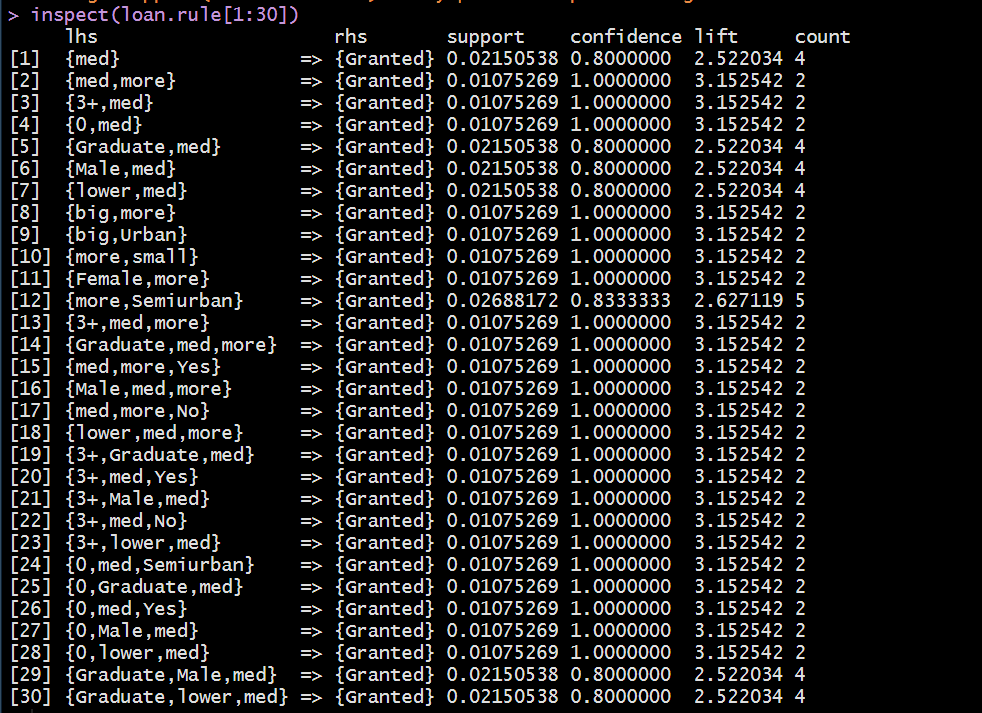


After building rules:





Let’s check for which rules Loan has been granted,



So, we will analyse our rules on the basis of confidence:

We have arranged our data as follows:

**Applicant Income = Low, med, high**

**CoapplicantIncome = Lower, mid, higher**

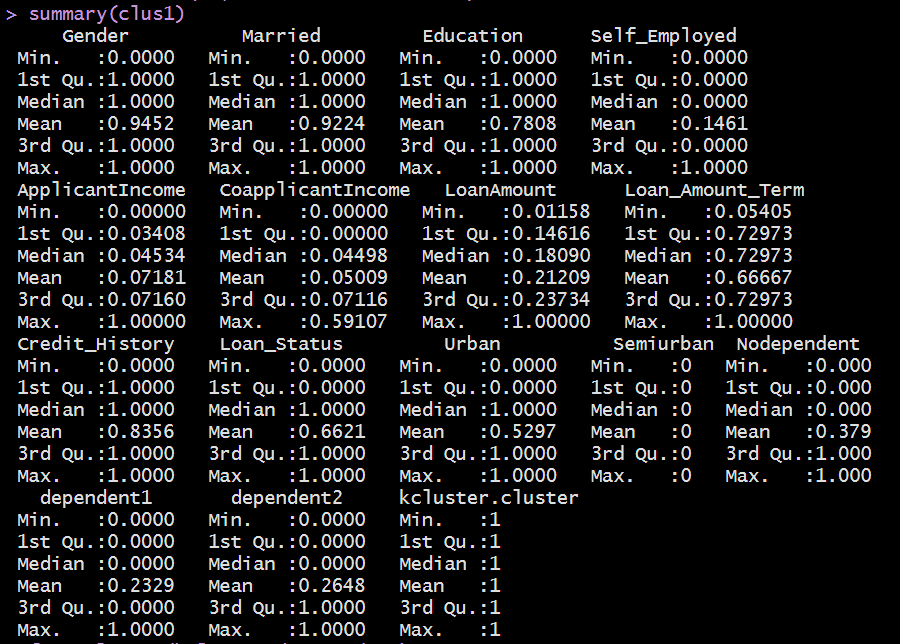
**Loan Amount = Less, more**

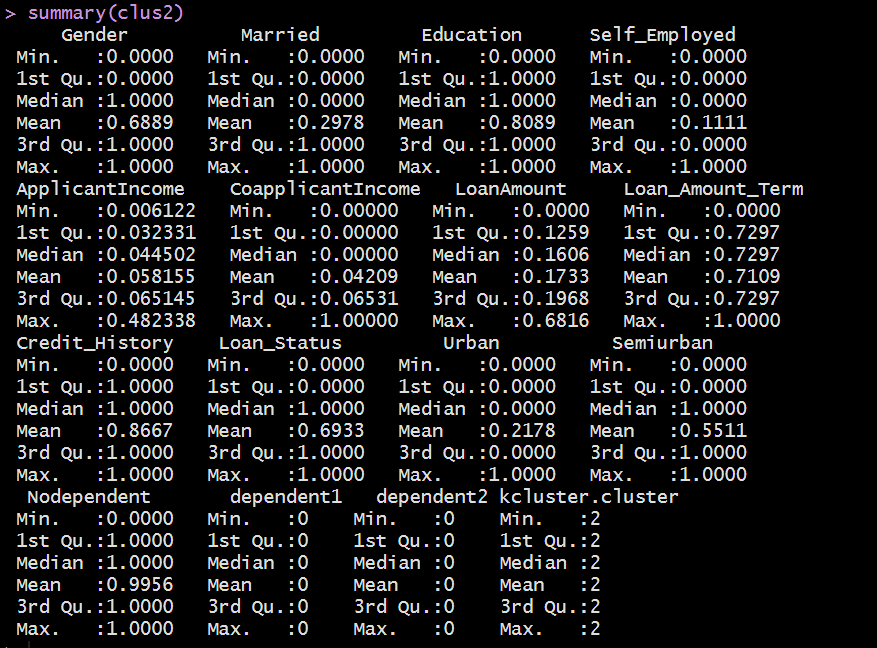
**Loan\_Amount\_Term = Small, big, large**

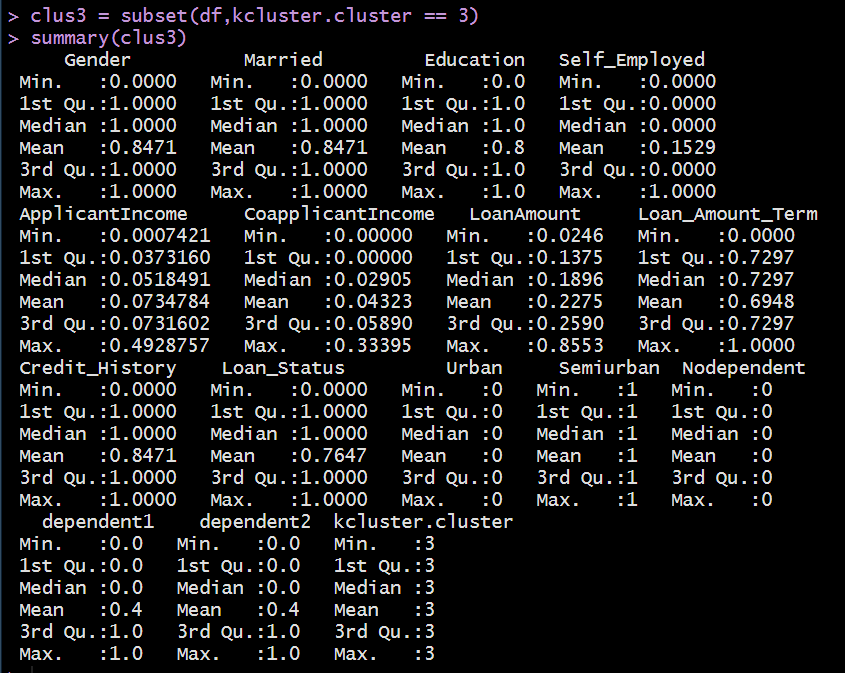
Accordingly, we see, people who are having

* no dependents, are graduate and primary applicant income is medium range then loan is getting granted.
* For the people who are staying in Semi urban areas are also having higher chance of getting the loan
* we see people who are not having any dependents are having very high chance of getting the loan approved
* If Primary applicant income ranges between medium to high and also Co applicant income ranges from low to medium have higher chance of getting loan approved as compared to Co applicant salary is zero.
* Credit history is good also is important for loan approval
* Males are having more probability of getting the loan approved
* People who are not self-employed are having chances

Now by keeping these few points analysing each cluster:







Moving step by step:

* **Credit History: 1st: 0.83 2nd: 0.86 3rd: 0.84 >> Cluster 2 is having good chances due to good credit score**
* **Education: 1st:0.78 2nd:0.81 3rd:0.8 >> Cluster 2 is having good chances due to good education background**
* **Dependents (no dependents): 1st:0.379 2nd:0.99 3rd:0 >> Cluster 2 is having good chances due to no dependents as per our study**
* **Property area (Semi Urban): 1st: 0 2nd: 0.55 3rd: 1 >> Cluster 3 is having good chances on the basis of Property**
* **Gender: 1st:0.94 2nd:0.68 3rd:0.84 >> Cluster 1 is having good chances**
* **Applicant Income: 1st: 0.071 2nd:0.058 3rd:0.073 >> Cluster 3 is having good chances**
* **Self Employed: 1st: 0.14 2nd: 0.11 3rd: 0.15 >> Cluster 3 is having good chances**
* **Co Applicant Income: 1st: 0.045 2nd: 0.042 3rd: 0.043 >> Cluster 1 is having good chances**

Hence, as Credit History is having more weightage and is the primary decision factor in deciding if the person is getting loan approval or not.

**CLUSTER 2** is a set of people who are having very high probability of getting the loan approved by fulfilling almost all-important criteria of loan approval which is being followed in this study.

**TARGETTED CLUSTER** should be **CLUSTER 3** as they are also fulling a good number of criteria only slightly lagging behind in credit history.

1. **Conclusion**

* The main purpose of this project is to classify and analyse the nature of the loan applications.
* From a proper analysis of the data set and constraints of the banking sector, different graphs were generated and visualized.
* From data analysis, many conclusions have been made and information were inferred such as short-term loan was preferred by majority of the loan applicants and the client’s majority apply loan for debt consolidation.
* From predictive analysis, Random Forest in simple terms predicts the probability of occurrence of an event by fitting data. We generated a confusion matrix with accuracy, precision, recall score and F1 score of 80.50%, 80.62% and 94.55% , 87.03 % for the model.