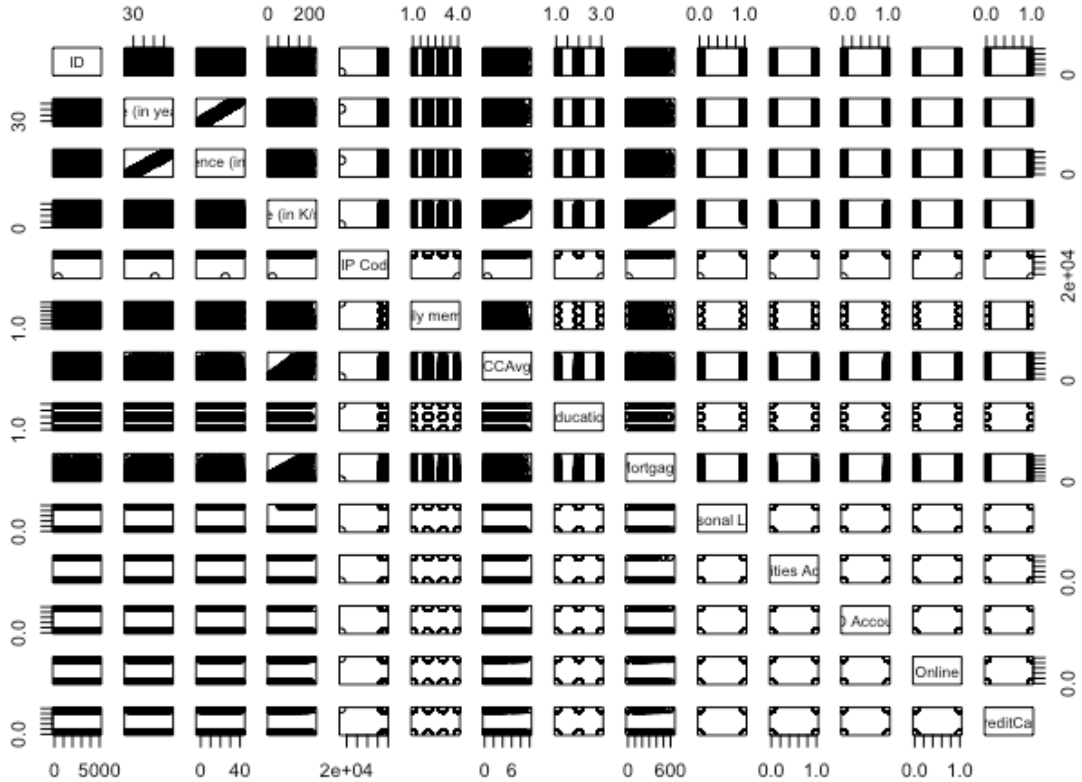
This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with a minimal budget. The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign. The dataset has data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

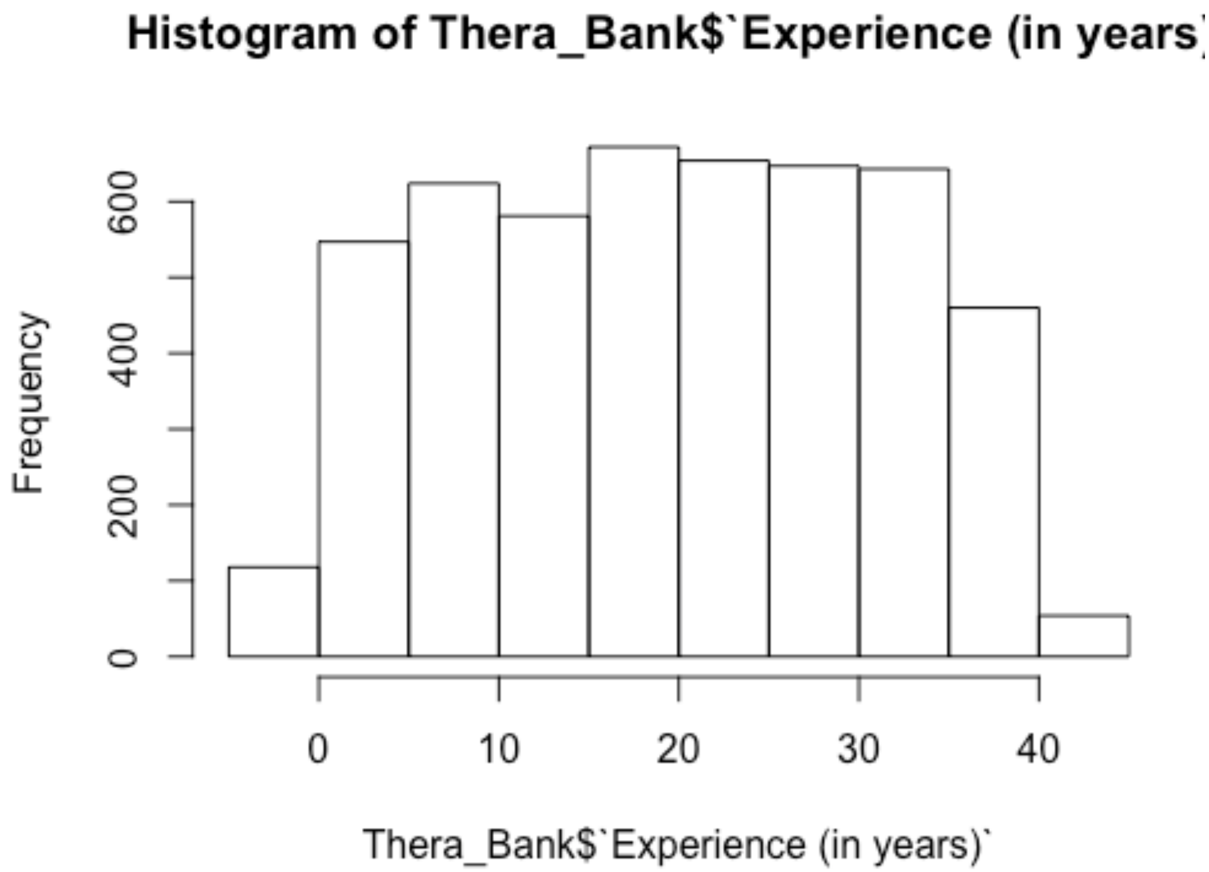
You are brought in as a consultant and your job is to build the best model which can classify the right customers who have a higher probability of purchasing the loan. You are expected to do the following:

- EDA of the data available. Showcase the results using appropriate graphs - **(10 Marks)**
- Apply appropriate clustering on the data and interpret the output - **(10 Marks)**
- Build appropriate models on both the test and train data (CART & Random Forest). Interpret all the model outputs and do the necessary modifications wherever eligible (such as pruning) - **(20 Marks)**
- Check the performance of all the models that you have built (test and train). Use all the model performance measures you have learned so far. Share your remarks on which model performs the best. - **(20 Marks)**

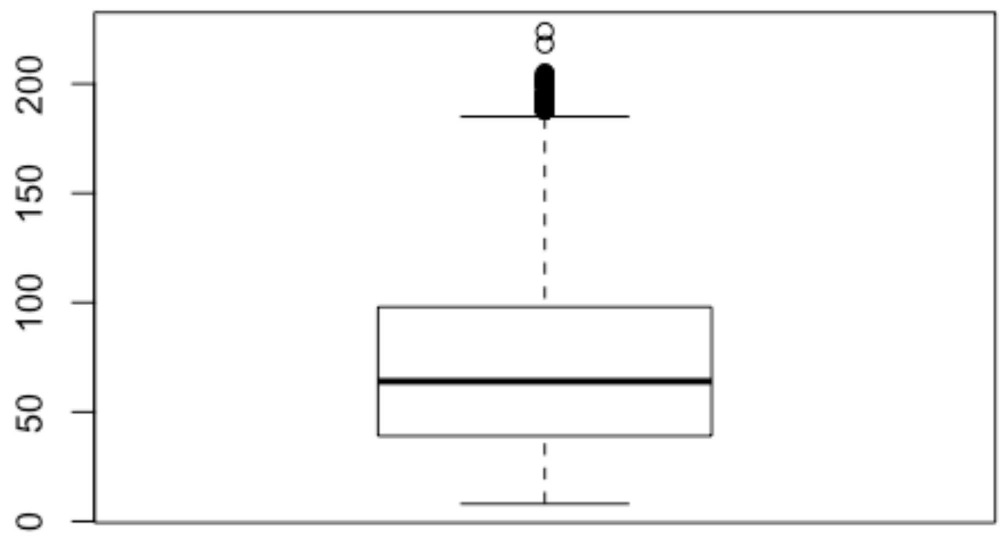
Exploratory Data Analysis includes the following Analysis:

```
head(Thera_Bank)
tail(Thera_Bank)
summary(Thera_Bank)
plot(Thera_Bank$`Experience (in years)`)
plot(Thera_Bank)
hist(Thera_Bank$`Experience (in years)`)
boxplot(Thera_Bank$`Income (in K/month)`)
```



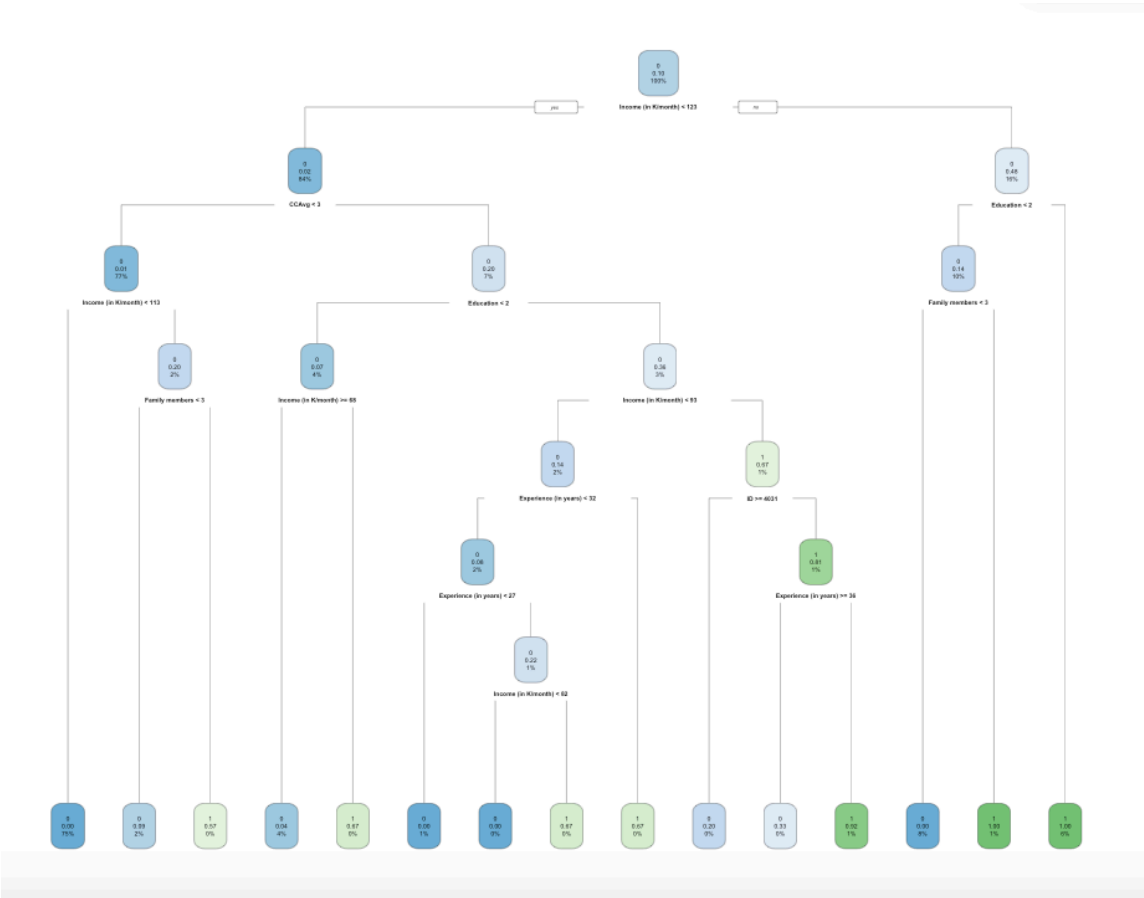


Histogram of Thera_Bank$`Experience (in years)
Frequency
0
200
400
600
0
10
20
30
40
Thera_Bank$`Experience (in years)`



0
50
100
150
200

CART MODEL



Classification tree:

Classification tree:

rpart(formula = `Personal Loan` ~ ., data = trainThera_Bank,

    method = "class", cp = 0, minbucket = 3)

Variables actually used in tree construction:

[1] Age (in years)   CCAvg   CD Account

[4] CreditCard   Education   Family members

[7] ID   Income (in K/month) ZIP Code

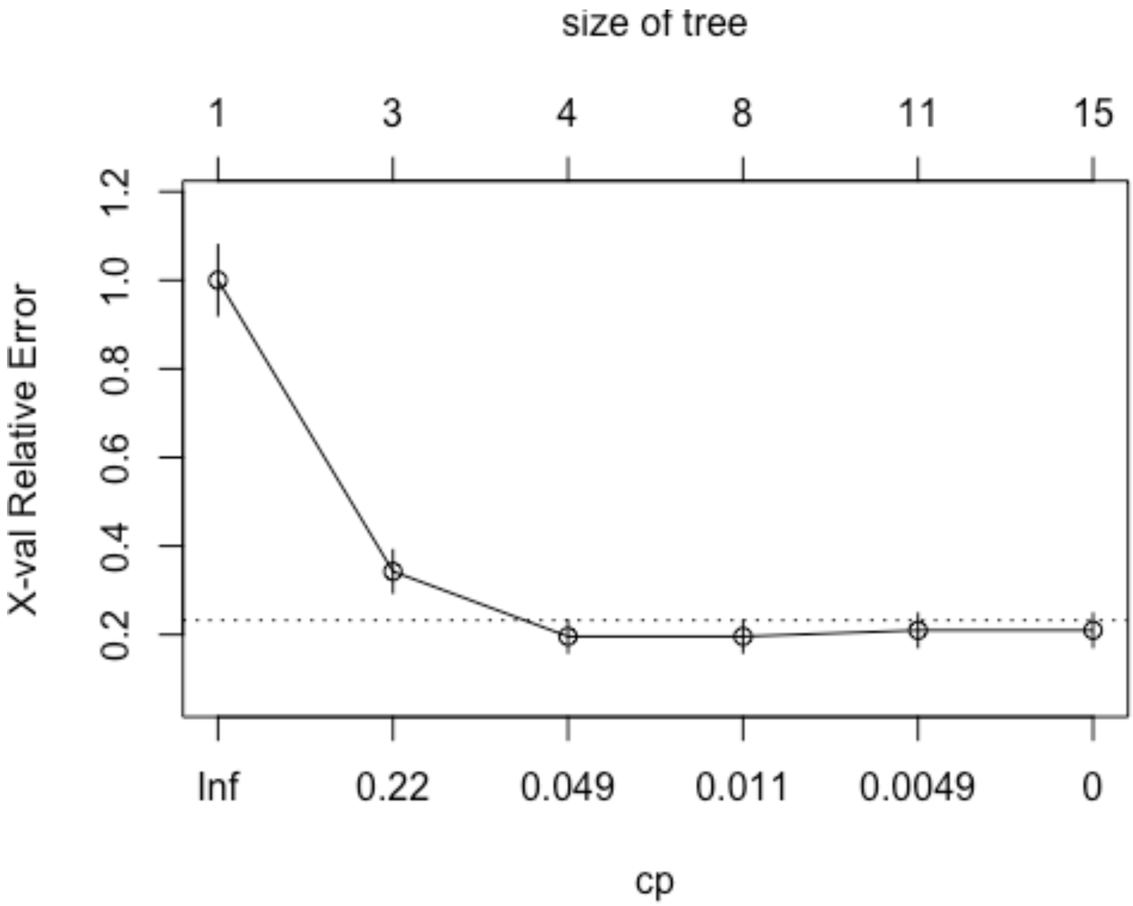
Root node error: 335/3488 = 0.096044

```
n= 3488

          CP nsplit rel error  xerror     xstd
1  0.3014925      0  1.000000 1.00000 0.051946
2  0.1223881      2  0.397015 0.43881 0.035421
3  0.0388060      3  0.274627 0.28955 0.028988
4  0.0358209      4  0.235821 0.25373 0.027184
5  0.0164179      5  0.200000 0.21791 0.025236
6  0.0149254      9  0.134328 0.20597 0.024549
7  0.0074627     10  0.119403 0.16119 0.021765
8  0.0059701     12  0.104478 0.15522 0.021365
9  0.0039801     13  0.098507 0.16119 0.021765
10 0.0029851     16  0.086567 0.16418 0.021963
11 0.0017910     18  0.080597 0.19403 0.023841
12 0.0000000     24  0.068657 0.21791 0.025236
> plotcp(tree)
> tree = prune(tree, cp = 0.3014925)
> tree
n= 3488

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 3488 335 0 (0.90395642 0.09604358)
  2) Income (in K/month)< 106.5 2706  45 0 (0.98337029 0.01662971) *
```

3) Income (in K/month)>=106.5 782 290 0 (0.62915601 0.37084399)

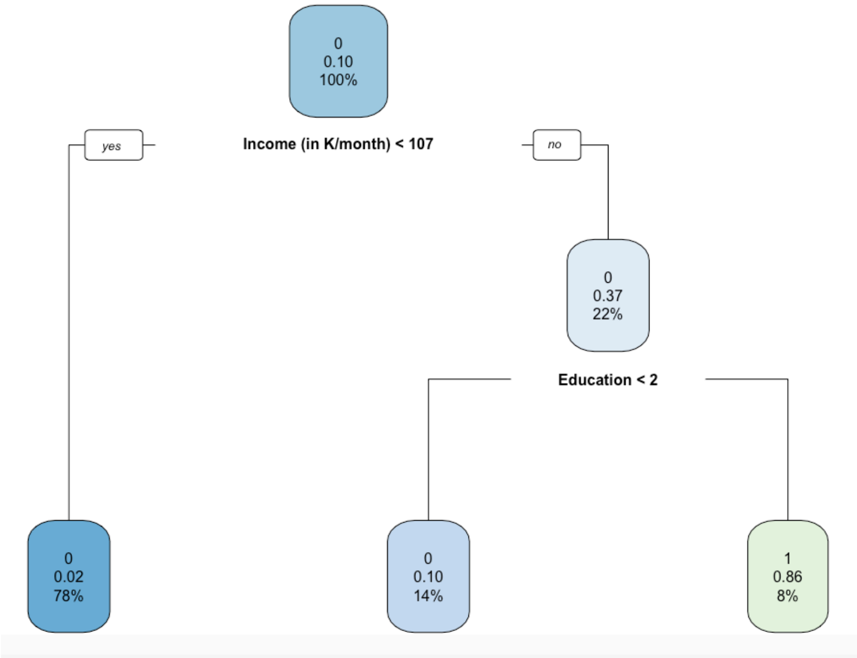
6) Education< 1.5 504 50 0 (0.90079365 0.09920635) *

7) Education>=1.5 278 38 1 (0.13669065 0.86330935) *



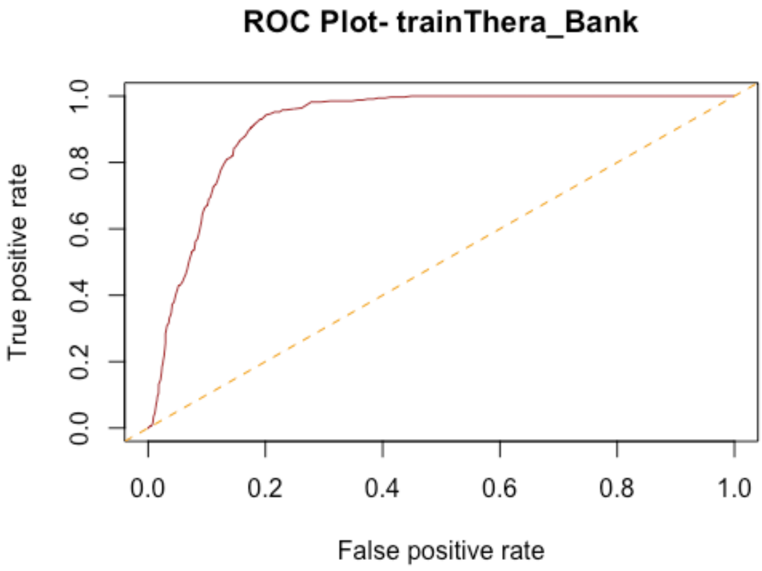
Pruning:

```
tree = prune(tree, cp = 0.3014925)
tree
rpart.plot(tree)
printcp(tree)
trainThera_Bank$prediction = predict(tree, trainThera_Bank, type ="class" )
trainThera_Bank$prediction = predict(tree, trainThera_Bank, type = "prob")
View(trainThera_Bank)
trainThera_Bank = with(trainThera_Bank, table(`Age (in years)`, trainThera_Bank$prediction))
trainThera_Bank
trainThera_Bank[1,1]
```



ROC Curve

```
library(ROCR)
predict = prediction(trainThera_Bank$`Income (in K/month)`,trainThera_Bank$`Personal Loan`)
perf = performance(predict,"tpr", "fpr")
plot(perf, col ="Brown", main = "ROC Plot- trainThera_Bank")
abline(0,1, lty =8, col = "orange")
```



cart.trainThera_Bank.auc = performance(predict,"auc")
cart.trainThera_Bank.auc = cart.trainThera_Bank.auc@y.values
cart.trainThera_Bank.auc

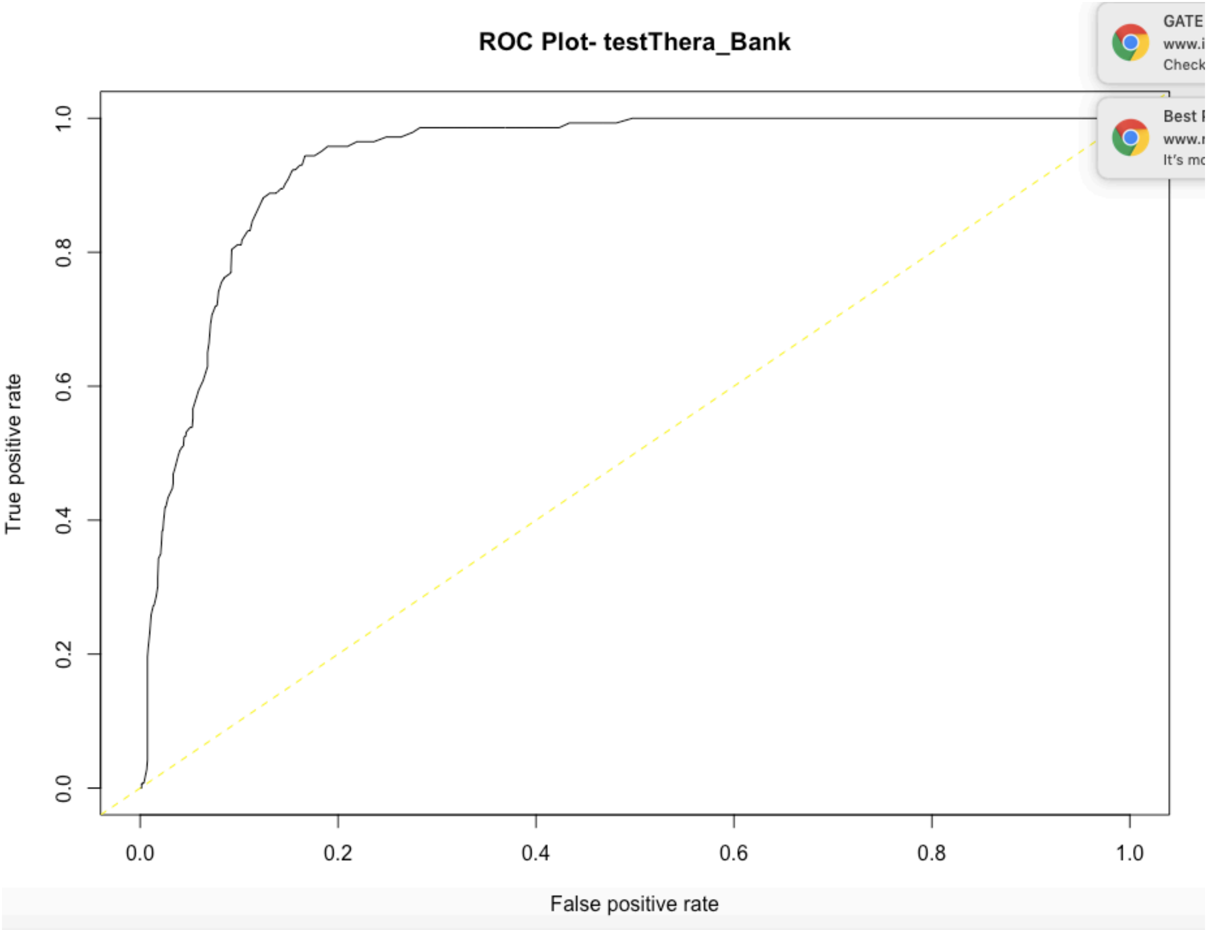
Value is 91.51

cart.trainThera_Bank.KS = max(perf@y.values[[1]]-perf@x.values[[1]])
> cart.trainThera_Bank.KS

Value is 74.22

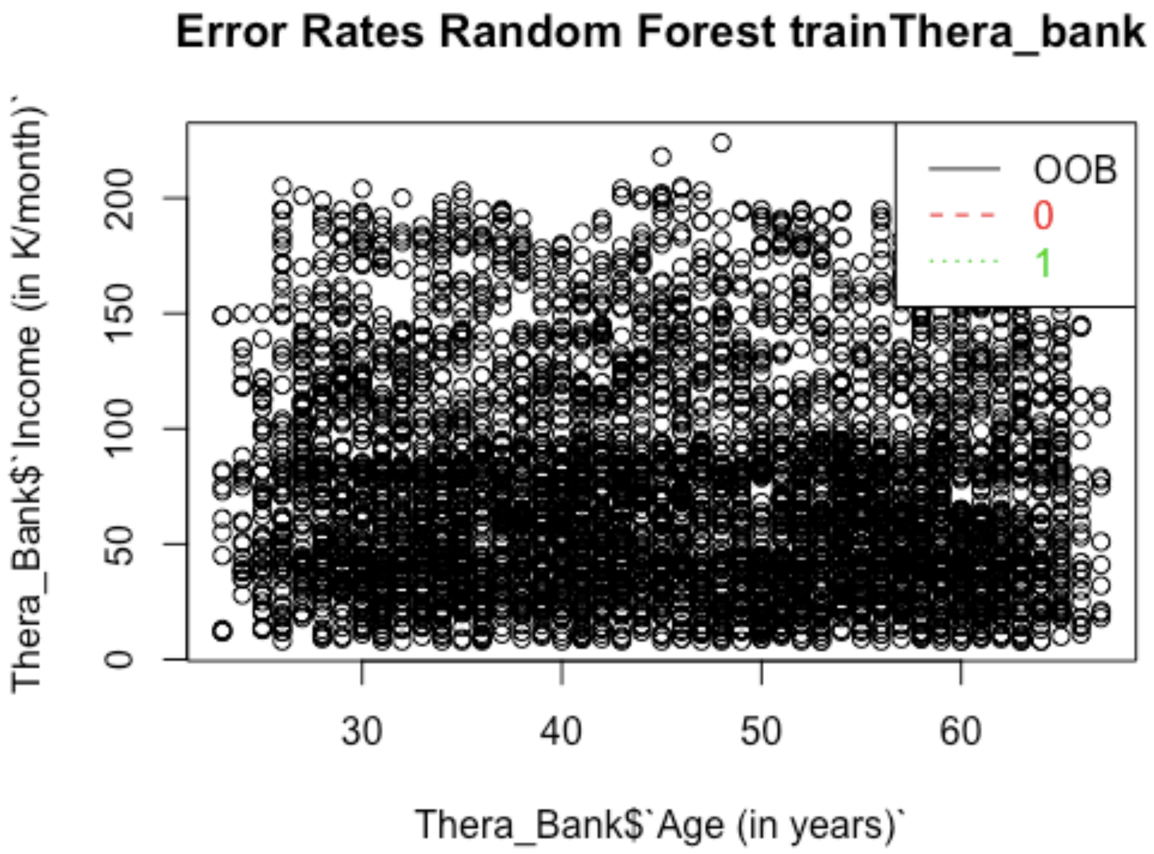
Roc Curve for Test Dataset:

predict1 = prediction(testThera_Bank$`Income (in K/month)`,testThera_Bank$`Personal Loan`)
perf1 = performance(predict1, "tpr", "fpr")
plot(perf1, col ="Black", main = "ROC Plot- testThera_Bank")
abline(0,1, lty =8, col = "yellow")


ROC Plot- testThera_Bank
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
GATE
www.i
Check
Best I
www.r
It's mo

Random Forest:

```
head(Thera_Bank)
tail(Thera_Bank)
summary(Thera_Bank)
plot(Thera_Bank$`Experience (in years)`)
plot(Thera_Bank)
hist(Thera_Bank$`Experience (in years)`)
boxplot(Thera_Bank$`Income (in K/month)`)
boxplot(Thera_Bank)
View(Thera_Bank)
dim(Thera_Bank)
attach(Thera_Bank)
complete.cases(Thera_Bank)
Thera_Bank = Thera_Bank[complete.cases(Thera_Bank),]
dim(Thera_Bank)
table(Thera_Bank$`Personal Loan`)
prop.table(table(Thera_Bank$`Personal Loan`))
round(prop.table(table(Thera_Bank$`Personal Loan`)),3)
plot(Thera_Bank$`Age (in years)`, Thera_Bank$`Income (in K/month)`)
points(Thera_Bank$`Age (in years)`[Thera_Bank$`Experience (in years)`=="Good"],Thera_Bank$`Income (in K/month)`[Thera_Bank$`Experience (in years)`=="Good"],col="green",pch=19)
points(Thera_Bank$`Age (in years)`[Thera_Bank$`Experience (in years)`=="Bad"],Thera_Bank$`Income (in K/month)`[Thera_Bank$`Experience (in years)`=="Bad"],col="Yellow",pch=19)
library(caTools)
set.seed(123)
split = sample.split(Thera_Bank$`Personal Loan`,SplitRatio = 0.70)
trainThera_Bank= subset(Thera_Bank, split == TRUE)
testThera_Bank = subset(Thera_Bank, split == FALSE)
dim(trainThera_Bank)
dim(testThera_Bank)
round(prop.table(table(trainThera_Bank$`Personal Loan`)),3)
round(prop.table(table(testThera_Bank$`Personal Loan`)),3)
install.packages("randomForest")
library(randomForest)
View(trainThera_Bank)
mtry = floor(sqrt(ncol(trainThera_Bank)))
mtry
set.seed(123)
randomForest = randomForest(trainThera_Bank~ ., data = trainThera_Bank[-1], ntree = 50, mtry = 3, nodesize = 10, importance = TRUE)
print(randomForest)
randomForest$err.rate
plot(randomForest, main= "")
legend("topright", c("OOB", "0", "1"), text.col = 1:6, lty = 1:3, col=1:3)
title(main = "Error Rates Random Forest trainThera_bank")
```


Error Rates Random Forest trainThera_bank
OOB
0
1
Thera_Bank$`Income (in K/month)`
0
50
100
150
200
30
40
50
60
Thera_Bank$`Age (in years)`

- Check the performance of all the models that you have built (test and train). Use all the model performance measures you have learned so far. Share your remarks on which model performs the best. –

Model		Accuracy	Specificity	Sensitivity	Ks	Auc
CART	Train	91.35%	98.56	37.65	94	99
	Test	89.45	98.2	12.6	39	77.74
Random Forest	Train	90	95	33	95	86
	Test	88	64.37	20	83.56	82

According to the above interpretation I can say that CART model works the best in this case as compared to Random Forest.

Precision Recall Curve or compared to ROC Curve when compared with both the models, cart model works the best.

Also by applying the various clustering techniques, I can K means clustering plays a major role in determining the eligibility of individual.

Also I can conclude by saying that CART model has various features as compared to Random Forest wherein we are able to determine a lot of factors which plays a key role in determining the eligibility of the customer.

The various factors include customers previous details where we can check everything and decide if we can give loan or not.