Thera Bank - Loan Purchase Modeling

Submitted by,

Ajuna John

Objective

Thera Bank is interested in expanding the customer base of which majority are liability customers, to bring in more loan business and in the process, earn more through the interest on loans. A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. The objective of this project is to build the best model that will help Thera bank, to 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.

DataSet

Data	**Description**
ID	Customer ID
Age	Customer's age in years
Experience	Years of professional experience
Income	Annual income of the customer ($000)
ZIPCode	Home Address ZIP code
Family	Family size of the customer
CCAvg	Avg. spending on credit cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any (K)
Personal Loan	Did the customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by the bank?

The dataset has 5000 observations with 14 variables.

Assumption

- The data has one dependent variable and other response variables

Importing libraries

```
library(grid)
library(gridExtra)
library(lattice)
library(ModelMetrics)
library(randomForest)
library(corrplot)
library(ineq)
library(ROCR)
library(caret)
library(tidyverse)
library(readxl)
library(dplyr)
library(randomForest)
library(rpart)
library(ggplot2)
library(rpart.plot)
```

Analysis of Dataset

Personal Loan is considered as the Dependent variable and all other attributes as Independent variables.

The dataset has customer information like **Age, Experience, Income, zip code, family members, CCAvg and Education** which represent the customer behavior that needs to be considered.

The variables like **Mortgage, Securities Account, CD Account, online, credit card** helps us to understand the facilities availed by the customer which encourage them to take personal loan which needs to be considered too.

Here we should not consider the customer ID and Zip code as it does not help in model building.

Treatment of missing data:

Found missing values in 18 places in the 'Family members' column of the dataset. Since 18 observation rows having "NA" as family members are also having vital other information, we may replace NA with "median value of the column" to factor them instead of discarding them

Structure of the dataset

Dataframe – 5000 observations of 14 variables

```
 $ ID                   : num [1:5000] 1 2 3 4 5 6 7 8 9 10 ...
 $ Age (in years)       : num [1:5000] 25 45 39 35 35 37 53 50 35 34 ...
 $ Experience (in years): num [1:5000] 1 19 15 9 8 13 27 24 10 9 ...
 $ Income (in K/year)   : num [1:5000] 49 34 11 100 45 29 72 22 81 180 ...
 $ ZIP Code             : num [1:5000] 91107 90089 94720 94112 91330 ...
 $ Family members       : num [1:5000] 4 3 1 1 4 4 2 1 3 1 ...
 $ CCAvg                : num [1:5000] 1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
 $ Education            : num [1:5000] 1 1 1 2 2 2 2 3 2 3 ...
 $ Mortgage             : num [1:5000] 0 0 0 0 0 155 0 0 104 0 ...
 $ Personal Loan        : num [1:5000] 0 0 0 0 0 0 0 0 0 1 ...
 $ Securities Account   : num [1:5000] 1 1 0 0 0 0 0 0 0 0 ...
 $ CD Account           : num [1:5000] 0 0 0 0 0 0 0 0 0 0 ...
 $ Online               : num [1:5000] 0 0 0 0 0 1 1 0 1 0 ...
 $ CreditCard           : num [1:5000] 0 0 0 0 1 0 0 1 0 0 ...
```

All the observations are numeric

Experience has negative values. We will fix them with corresponding absolute values

```
Age_in_years `Experience(yea~ `Income(K/year)` Family_members CCAvg Education
       <dbl>            <dbl>            <dbl> <fct>          <dbl> <ord>
 1        25               -1              113 4                2.3  3
 2        24               -1               39 2                1.7  2
 3        24               -2               51 3                0.3  3
 4        28               -2               48 2                1.75 3
 5        24               -1               75 4                0.2  1
 6        25               -1               43 3                2.4  2
 7        25               -1              109 4                2.3  3
 8        25               -1               48 3                0.3  3
 9        24               -1               38 2                1.7  2
10        24               -2              125 2                7.2  1
# ... with 42 more rows, and 6 more variables: Mortgage <dbl>, Personal_loan <fct>,
#   Securities_Account <fct>, CD_Amount <fct>, Online <fct>, CreditCard <fct>
```

Columns like Personal Loan, Securities Account, CD Account, Online, Credit card etc are factor values with levels “0” and “1”. Education is ordered factor with 3 levels 1, 2 and 3

Education (in Years) is converted into ordered factors

```
summary(bankdata)
```

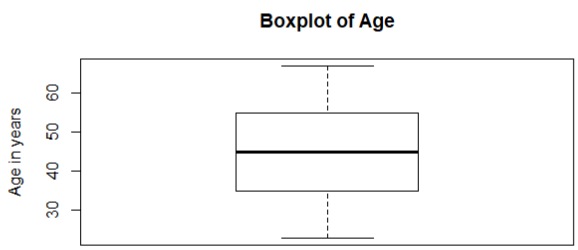
```
 Age (in years)  Experience (in years) Income (in K/year)
 Min.   :23.00   Min.   :-3.0         Min.   :  8.00
 1st Qu.:35.00   1st Qu.:10.0         1st Qu.: 39.00
 Median :45.00   Median :20.0         Median : 64.00
 Mean   :45.34   Mean   :20.1         Mean   : 73.77
 3rd Qu.:55.00   3rd Qu.:30.0         3rd Qu.: 98.00
 Max.   :67.00   Max.   :43.0         Max.   :224.00

  Family members      CCAvg         Education       Mortgage
  Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0
  1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0
  Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
  Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5
  3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
  Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0

 Personal Loan   Securities Account   CD Account        Online
 Min.   :0.000   Min.   :0.0000    Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.000   1st Qu.:0.0000    1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.000   Median :0.0000    Median :0.0000   Median :1.0000
 Mean   :0.096   Mean   :0.1044    Mean   :0.0604   Mean   :0.5968
 3rd Qu.:0.000   3rd Qu.:0.0000    3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :1.000   Max.   :1.0000    Max.   :1.0000   Max.   :1.0000

 CreditCard
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.294
 3rd Qu.:1.000
 Max.   :1.000
```

Personal loan is having mean of 0.096

Exploratory Data analysis on the dataset

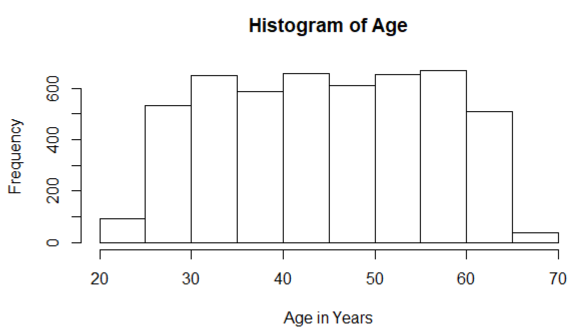
Univariate analysis

Boxplot of Age



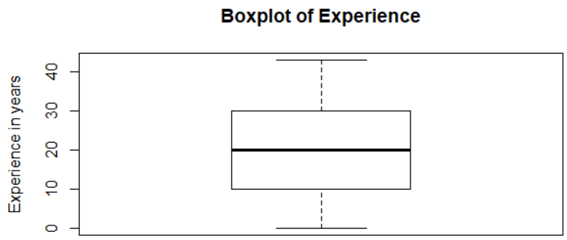
There is no outliers present in Age

Histogram of Age



We observe that Age is very close to the normal distribution

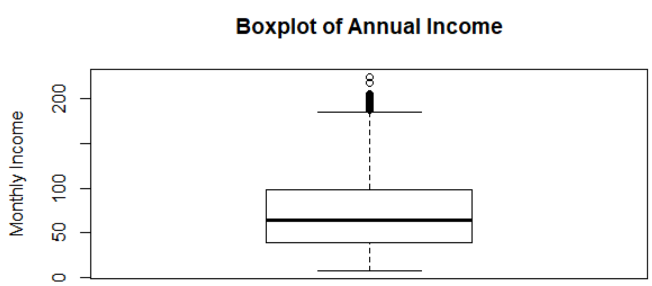
Boxplot of Experience in years



Histogram of Experience in years

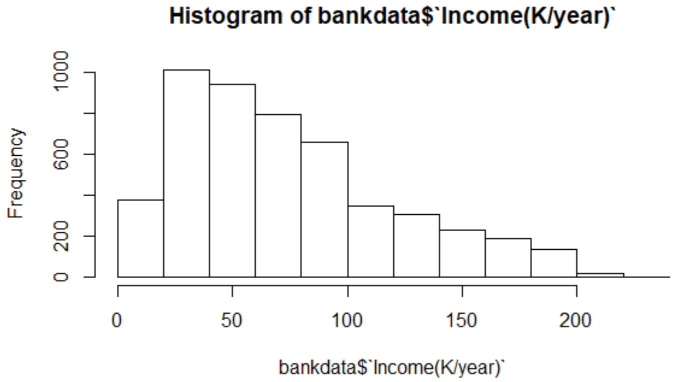
No outliers in Experience data

Boxplot of Annual Income

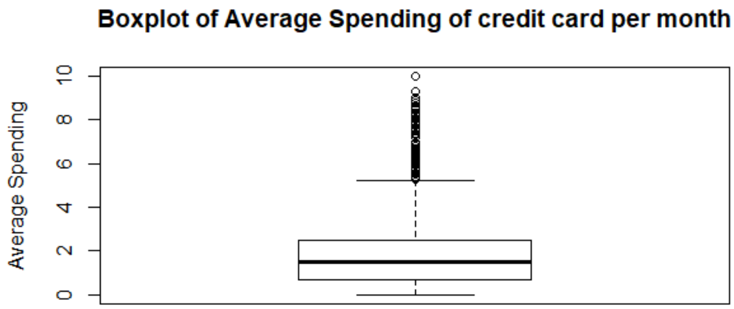


There are outliers in the Annual income data

Histogram of Annual Income

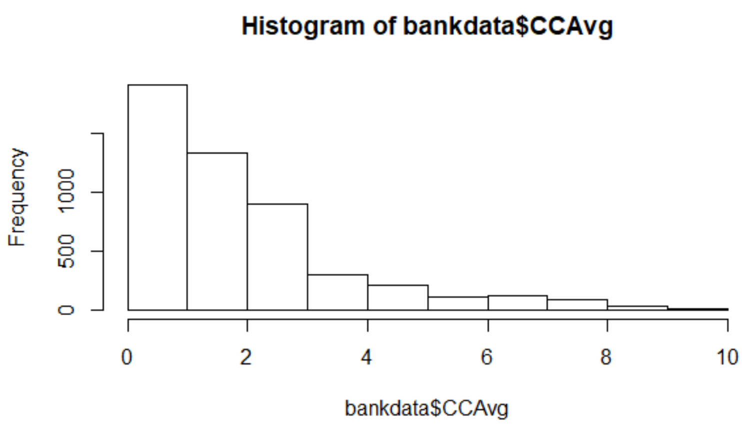


Boxplot of Average spending of credit card per month

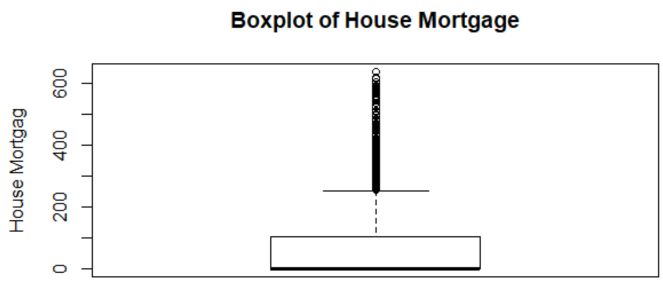


There are outliers in average spending of credit card per month

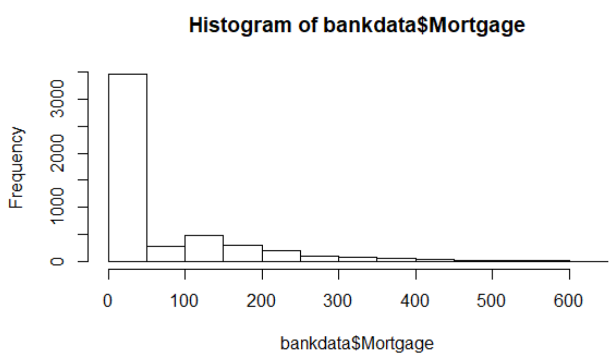
Histogram of Average spending of credit card per month



Boxplot of value of House mortgage

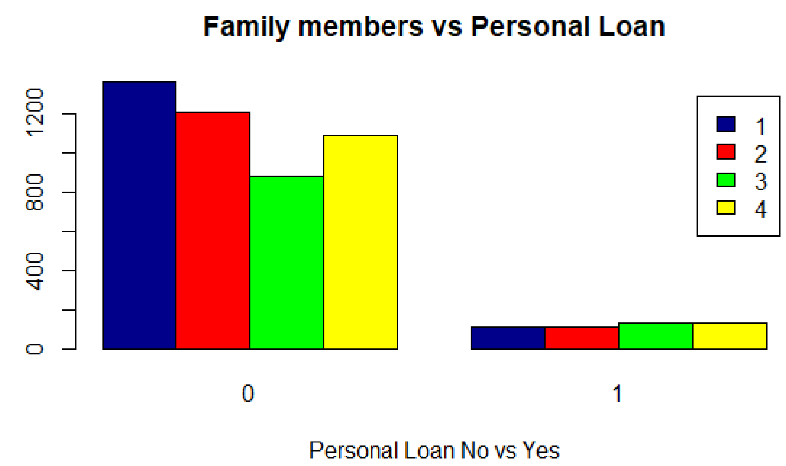


There are outliers in value of house mortgage



Multivariate analysis

Barplot – Number of family members Vs Personal Loan

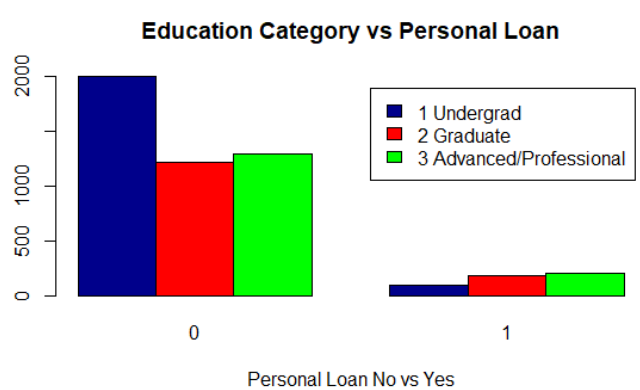


Families having more members have higher likelihood to take loan

Table to view relation between number of family members and personal loan

	0	1
1	1358	106
2	1202	108
3	876	133
4	1084	133

Barplot – Education Vs Personal Loan



Advanced/Professionals require loan

Table to view relation between Education category and Personal loan

Correlation between the numeric variables

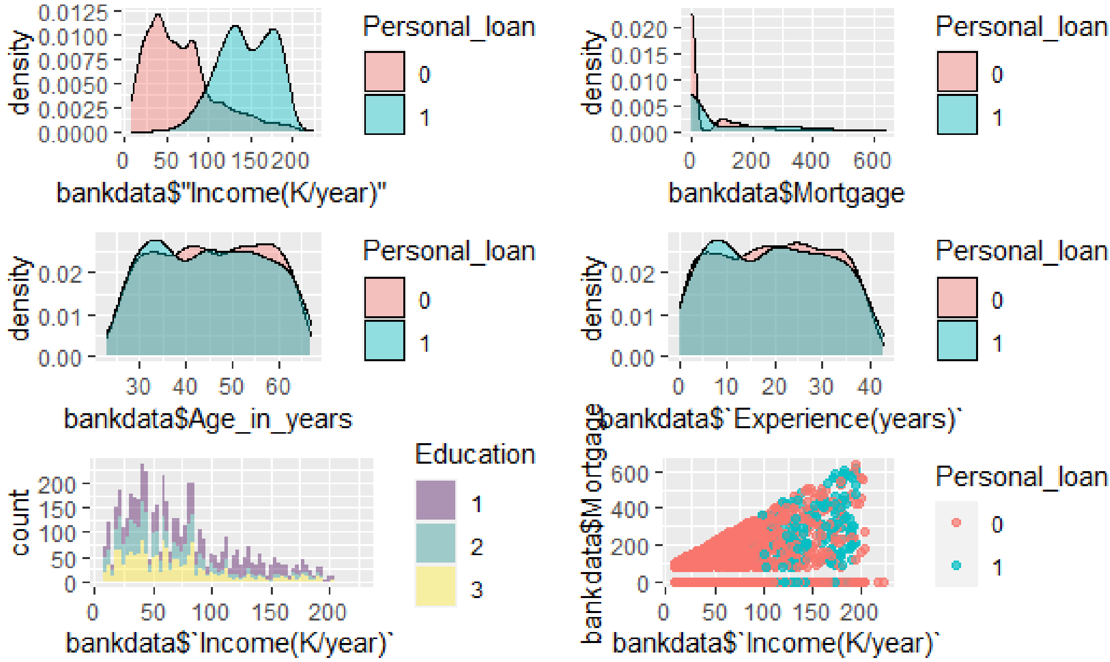
	Age_in_years	Experience(years)	Income(K/year)	CCAvg
Age_in_years	1.00	0.99	-0.06	-0.05
Experience(years)	0.99	1.00	-0.05	-0.05
Income(K/year)	-0.06	-0.05	1.00	0.65
CCAvg	-0.05	-0.05	0.65	1.00

- Age and Experience are highly positively correlated
- Monthly Income and Average credit card spend is also positively correlated

The customers who took personal loan vs no personal loan was 90.4% and 9.6% respectively

Following plots give us insight about how two categories of Personal Loan predictor are stacked across various other predictors

1. Income (density)
2. Mortgage (density)
3. Age (density)
4. Experience (density)
5. Income vs Education (histogram)
6. Income vs Mortgage (scatterplot)



Proportion of no-loan takers is very high across all three categories of Education - Undergrad, Grad, and Advanced Proffessionals

The customers who took personal loan vs no personal loan was 90.4% and 9.6% respectively

CART and Random Forest algorithm

Dataset is split into train (3500 observations and 12 variables) and test (1500 observations and 12 variables) data.

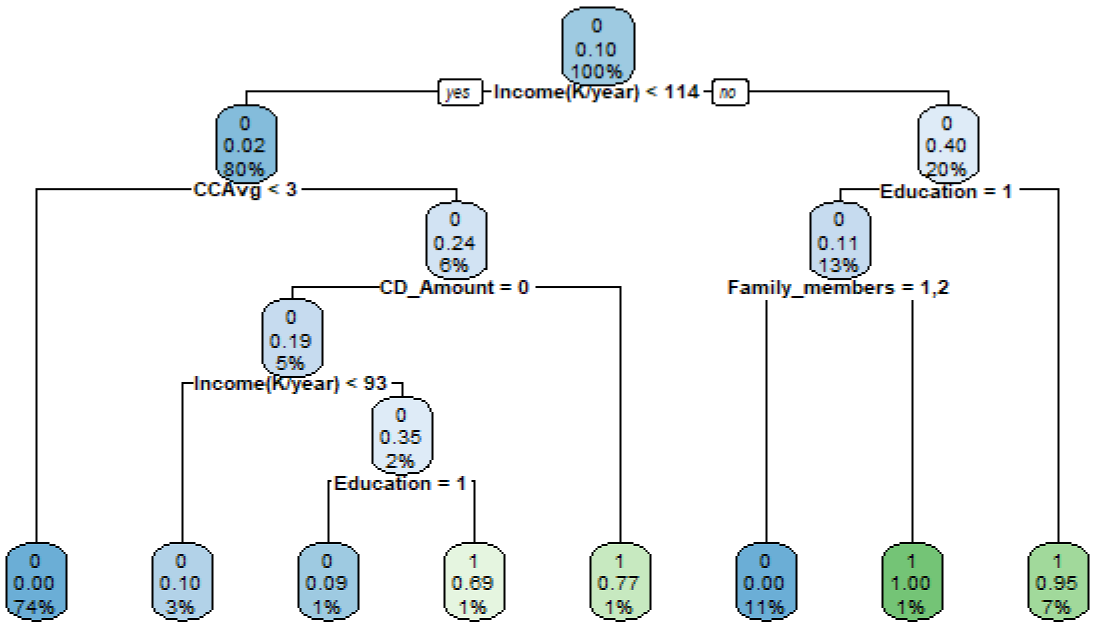
```
table(testdata$Personal_loan)

   0    1
1356  144


table(traindata$Personal_loan)

   0    1
3164  336
```

90.4% of train data says No and only 9.6% says Yes to personal loan. Similarly, 90.4% of test data says No and only 9.6% says Yes to personal loan. Hence, we need to balance the dataset. Both the train and test datasets are balanced with the help of certain functions.

CART Model

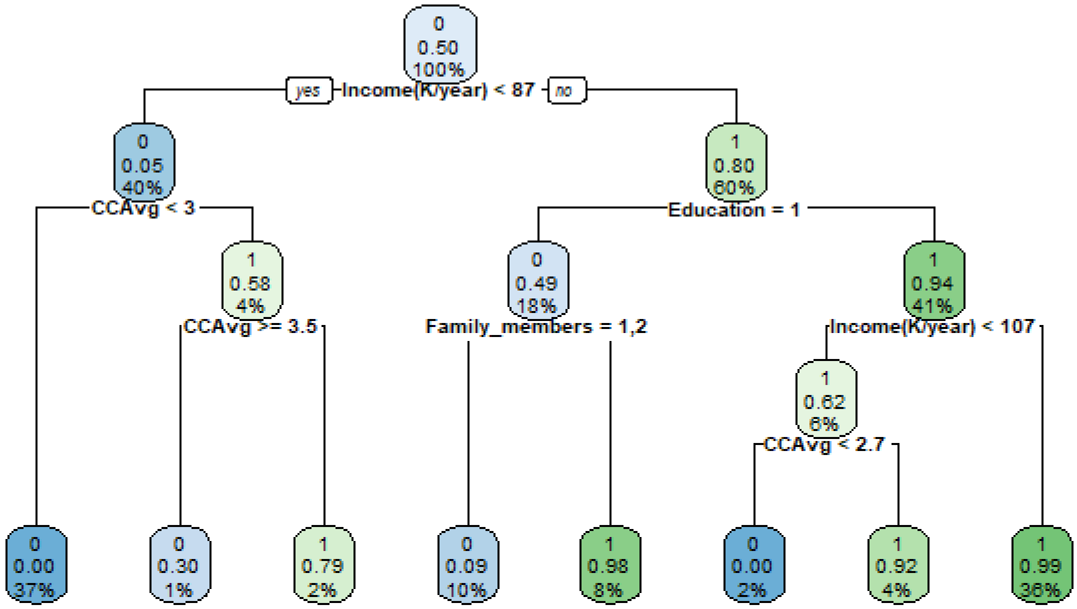
Using the below control parameters, the below CART model for the entire dataset is built.

minsplit = 20, minbucket = 10,xval = 5


0
0.10
100%
yes
Income(K/year) < 114
no
0
0.02
80%
CCAvg < 3
0
0.24
6%
CD_Amount = 0
0
0.19
5%
Income(K/year) < 93
0
0.35
2%
Education = 1
0
0.40
20%
Education = 1
0
0.11
13%
Family_members = 1,2
0
0.00
74%
0
0.10
3%
0
0.09
1%
1
0.69
1%
1
0.77
1%
0
0.00
11%
1
1.00
1%
1
0.95
7%

Using the below control parameters, the below CART model is built in the train dataset

minsplit = 20, minbucket = 10,xval = 5



---

n= 672

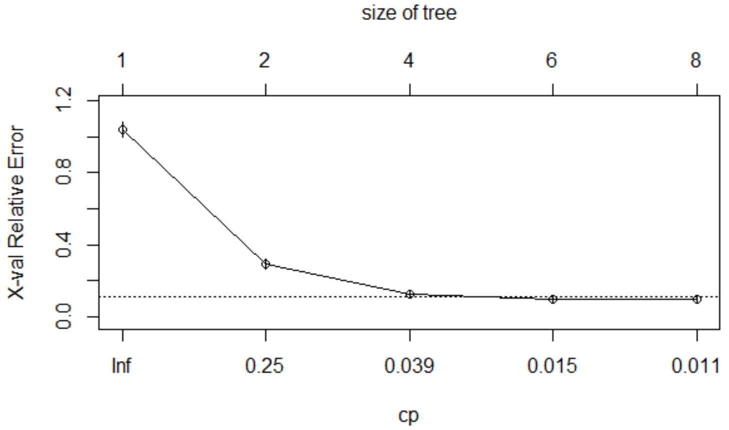
node), split, n, loss, yval, (yprob)
      * denotes terminal node

```
 1) root 672 336 0 (0.500000000 0.500000000)
   2) Income(K/year)< 87 271  14 0 (0.948339483 0.051660517)
     4) CCAvg< 2.95 247   0 0 (1.000000000 0.000000000) *
     5) CCAvg>=2.95 24  10 1 (0.416666667 0.583333333)
      10) CCAvg>=3.45 10   3 0 (0.700000000 0.300000000) *
      11) CCAvg< 3.45 14   3 1 (0.214285714 0.785714286) *
   3) Income(K/year)>=87 401  79 1 (0.197007481 0.802992519)
     6) Education=1 124  61 0 (0.508064516 0.491935484)
      12) Family_members=1,2 68   6 0 (0.911764706 0.088235294) *
      13) Family_members=3,4 56   1 1 (0.017857143 0.982142857) *
     7) Education=2,3 277  16 1 (0.057761733 0.942238267)
      14) Income(K/year)< 106.5 37  14 1 (0.378378378 0.621621622)
        28) CCAvg< 2.7 12   0 0 (1.000000000 0.000000000) *
        29) CCAvg>=2.7 25   2 1 (0.080000000 0.920000000) *
      15) Income(K/year)>=106.5 240   2 1 (0.008333333 0.991666667) *
```

<u>CP Table – Train dataset</u>

	CP	nsplit	rel error	xerror	xstd
1	0.72321429	0	1.00000000	1.03869048	0.03854695
2	0.08333333	1	0.27678571	0.29166667	0.02722984

3 0.01785714	3	0.11011905	0.12500000	0.01867545
4 0.01190476	5	0.07440476	0.09821429	0.01667184
5 0.01000000	7	0.05059524	0.09821429	0.01667184

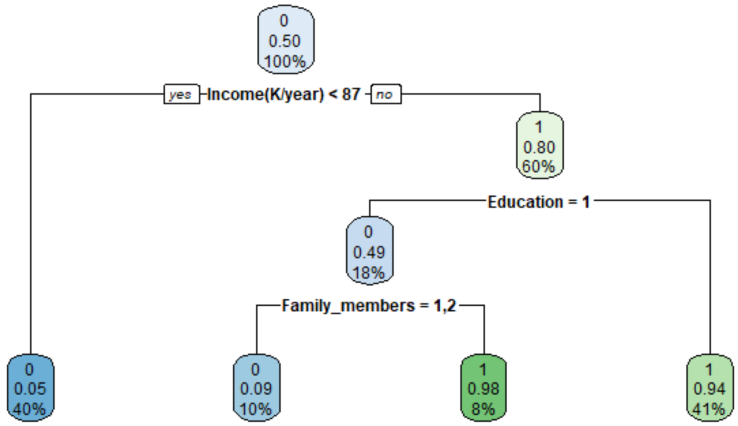


We will have to prune the train data considering .039 as the pruned parameter from the rpart plot.

After pruning the train data, we obtain the below tree

n= 672

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

 1) root 672 336 0 (0.50000000 0.50000000)
   2) Income(K/year)< 87 271  14 0 (0.94833948 0.05166052) *
   3) Income(K/year)>=87 401  79 1 (0.19700748 0.80299252)
     6) Education=1 124  61 0 (0.50806452 0.49193548)
      12) Family_members=1,2 68   6 0 (0.91176471 0.08823529) *
      13) Family_members=3,4 56   1 1 (0.01785714 0.98214286) *
     7) Education=2,3 277  16 1 (0.05776173 0.94223827) *
```



Cp table of the pruned data:

Classification tree:
rpart(formula = train_new$Personal_loan ~ ., data = train_new,
method = "class", control = r.ctrl)

Variables actually used in tree construction:
[1] Education Family_members Income(K/year)

Root node error: 336/672 = 0.5

n= 672

	CP	nsplit	rel error	xerror	xstd
1	0.723214	0	1.00000	1.03869	0.038547
2	0.083333	1	0.27679	0.29167	0.027230
3	0.039000	3	0.11012	0.12500	0.018675

Path of the pruned tree is:

node number: 1
root

node number: 2
root
Income(K/year)< 87

node number: 3
root
Income(K/year)>=87

node number: 6
root
Income(K/year)>=87

Education=1

node number: 7
root
Income(K/year)>=87
Education=2,3

node number: 12
root
Income(K/year)>=87
Education=1
Family_members=1,2

We shall now predict on test data and the confusion matrix we get is:

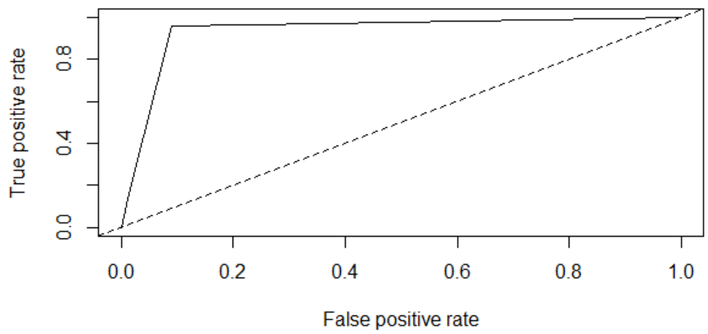
Confusion Matrix and Statistics

```
     0   1
  0 131  13
  1   6 138

               Accuracy : 0.934
                 95% CI : (0.8989, 0.9598)
    No Information Rate : 0.5243
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8681

 Mcnemar's Test P-Value : 0.1687

            Sensitivity : 0.9562
            Specificity : 0.9139
         Pos Pred Value : 0.9097
         Neg Pred Value : 0.9583
             Prevalence : 0.4757
         Detection Rate : 0.4549
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.9351

       'Positive' Class : 0
```

Accuracy is **93.4%**

Sensitivity : 95.62 %
Specificity : 91.39%

ROC



Area under the curve

Area under the curve is around **0.935**

The high values show that the model is built good and perform well

CART Model is close to 93.5% accurate in predicting personal loan on test data

Random Forest Model

672 samples
11 predictor
2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 672, 672, 672, 672, 672, 672, ...
Resampling results across tuning parameters:

mtry	Accuracy	Kappa
2	0.9506534	0.9012110
8	0.9708063	0.9415348
14	0.9651945	0.9303459

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 8.

Accuracy is **97.08%**

Prediction on test data

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	140	4
1	3	141

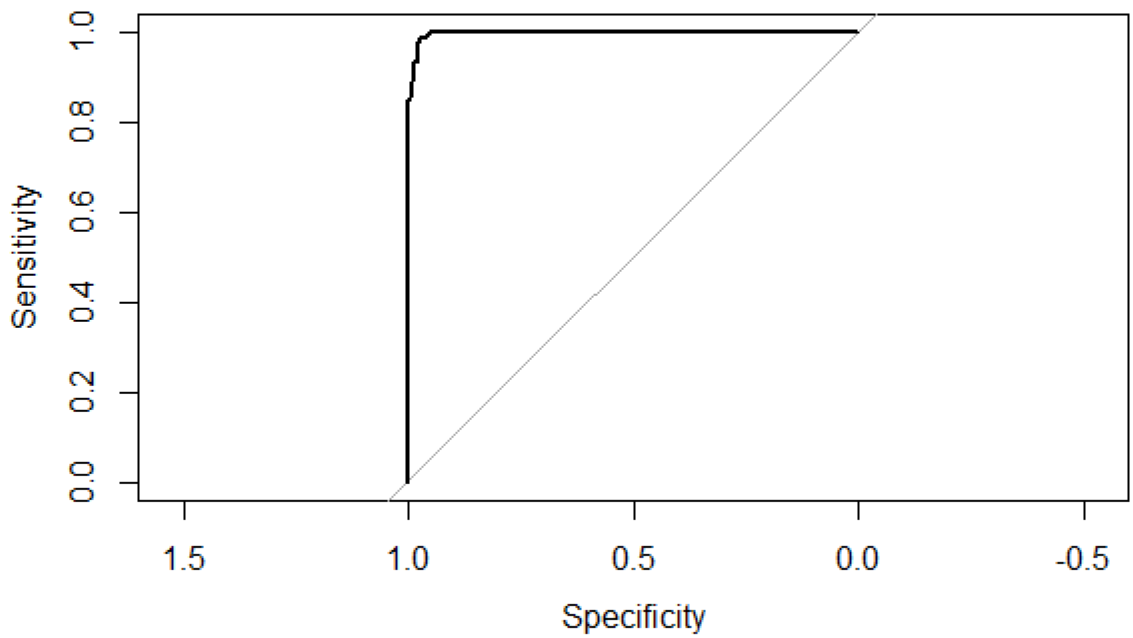
Accuracy : 0.9757
95% CI : (0.9506, 0.9902)
No Information Rate : 0.5035
P-Value [Acc > NIR] : <2e-16

Kappa : 0.9514

Mcnemar's Test P-Value : 1

Sensitivity : 0.9790
Specificity : 0.9724
Pos Pred Value : 0.9722
Neg Pred Value : 0.9792
Prevalence : 0.4965
Detection Rate : 0.4861
Detection Prevalence : 0.5000
Balanced Accuracy : 0.9757

'Positive' Class : 0



ROC is close to ideal one

Accuracy is 95.8%

Area under the curve: 0.9974

Monthly Income and Education is the most significant factor that decides personal loan