

# Thera Bank Loan

## Project 4: Machine Learning

Saturday,  
7<sup>th</sup> December 2019



**Submitted by: Usman Tahir**

Course: Data Science & Business Analytics

Institute: McCombs – UT Austin and Greatlearning

Tutor: Sambath Margabandhu

## CONTENTS

- i. **Project Objective**
- ii. **Assumptions**
- iii. **Exploratory Analysis**
  - a. Environment Setup
  - b. Variable Identification
  - c. Missing Value Treatment
  - d. Outlier Treatment
  - e. Transformation
  - f. Univariate Analysis
  - g. Bivariate Analysis
- iv. **Clustering**
- v. **Decision Tree (CART)**
- vi. **Random Forest**
- vii. **Model Performance Evaluation**
- viii. **Conclusion**
- ix. **Source Code**





# Project Objective

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
- Apply appropriate clustering on the data and interpret the output (Thera Bank wants to understand what kind of customers exist in their database and hence we need to do customer segmentation)
- 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)
- 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.

# Assumptions

- Data is representative of the entire population and free from errors.
- Variables included in the data are significant representation to achieve project objective.

# 3 Exploratory Data Analysis



Exploratory journey usually has eight stages, which we are going to follow in this descriptive analysis expedition.

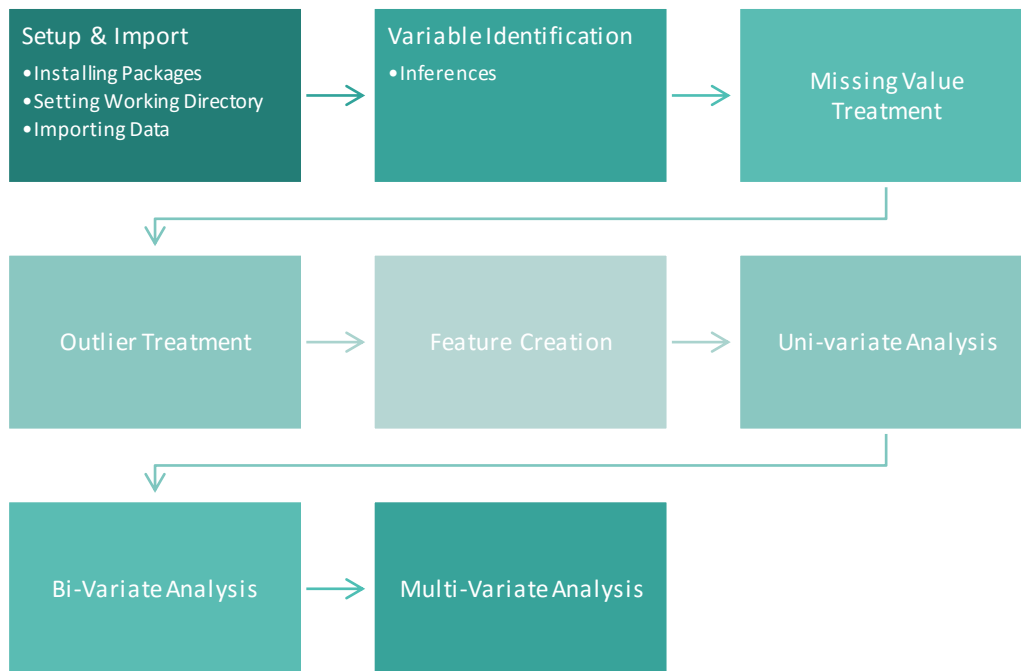


Figure 1: Exploratory Data Analysis Process

Data dictionary of the dataset is mentioned below:

- |                |  |
|----------------|--|
| 1. ID          | Customer ID  |
| 2. Age         | Customer's age in years  |
| 3. Exp         | Years of professional experience                                     |
| 4. Income      | Annual income of the customer (\$000)                                |
| 5. Zip         | Home Address ZIP code.   |
| 6. FMembers    | Family size of the customer  |
| 7. CCAvg       | Avg. spending on credit cards per month (\$000)                      |
| 8. Education   | Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional |
| 9. Mortgage    | Value of house mortgage if any. (\$000)                              |
| 10. PLoan      | Did this customer accept the loan offered in the last campaign?      |
| 11. Securities | Does the customer have a securities account with the bank?           |
| 12. CDAccount  | Does the customer have a certificate of deposit (CD) account?        |
| 13. Online     | Does the customer use internet banking facilities?                   |
| 14. CreditCard | Does the customer use a credit card issued by the bank?              |

## 3.1 Environment Setup and Data Import

This step is used for basically setting the stage for the analysis. Therefore, we will install packages, setting working directory and finally the most important step is to load the dataset:

### 3.1.1 Installing Packages & Invoking Libraries

We will require below packages for the purpose of analysis, please refer to Step-1 in the Appendix A-Source Code for further details:

LIBRARY	PURPOSE
PACMAN	Package Management – p_load function
READR	Reading CSV files
DATAEXPLORER	Exploring data structure of the dataset
MASS	
CATOOLS	
NBCLUST	Clustering Recommendations
FPC	
CLUSTER	CLUSTERING functions
RPART	CART Decision Tree Functions
RPART.PLOT	CART Visualization
TIDYVERSE	
RANDOMFOREST	Random Forest Functions
ROCR	Model Performance Functions
INEQ	Model Performance Functions
INFORMATIONVALUE	Concordance & Discordance
DATA.TABLE	Data table Manipulation functions

Table 1: Libraries Usage

### 3.1.2 Setting up working directory

We are setting the working directory to save all assets of analysis such as plots, tables and other explanatory documents at following path:

```
setwd("C:/DSBA_Course/Proper Learning/Module 4 [Machine Learning]/M4 W5/")
```

Snippet 1: Setting Working Directory

### 3.1.3 Importing and reading Dataset

Loading the dataset into “TheraBankData” dataframe through read.csv function:

```
TheraBankData <- read.csv("TheraBank.csv",header = TRUE)
```

Snippet 2: Loading CSV file into variables

## 3.2 Variable Identification

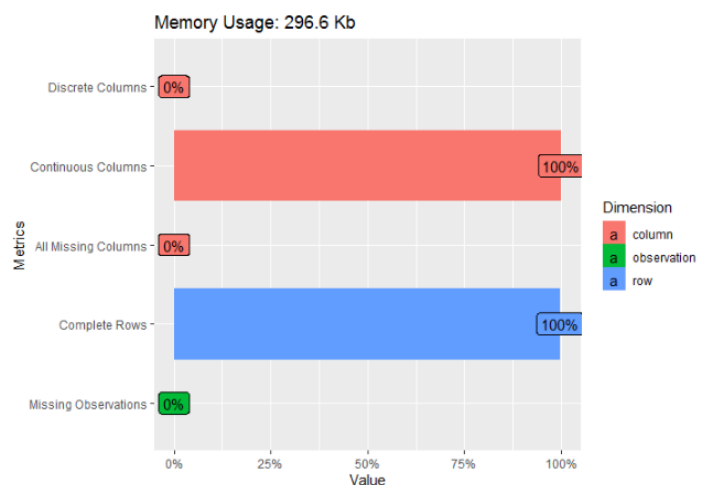
We will be using several R functions to analysis variables in the dataset for below mentioned purpose:

FUNCTION	PURPOSE
<b>DIM()</b>	Identify number of rows and columns in dataset
<b>PLOT_INTRO()</b>	Observations completeness in datasets
<b>NAMES()</b>	Identify if all column names are free from spaces in between
<b>PLOT_STR()</b>	Identify if there are empty values in dataset
<b>HEAD()</b>	Top 6 observations of the dataset
<b>PLOT_MISSING()</b>	Check Missing values
<b>SUMMARY()</b>	5 Number Summaries & Aggregation of variables

Table 2: Functions for Variable Identification

### 3.2.1 Inferences

- There are 5000 rows and 14 columns in the dataset.
- All variables are integer apart from CCAvg which is numeric.
- All variables are continuous with no categorical variable.
- There are no missing values and rows.



### DATASET STRUCTURE

```
$ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
$ Age     : int  25 45 39 35 35 37 53 50 35 34 ...
$ Exp     : int  1 19 15 9 8 13 27 24 10 9 ...
$ Income  : int  49 34 11 100 45 29 72 22 81 180 ...
$ Zip     : int  91107 90089 94720 94112 91330 92121 91711 93943 90089 93023 ..
$ FMembers : int  4 3 1 1 4 4 2 1 3 1 ...
$ CCAvg   : num  1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
$ Education : int  1 1 1 2 2 2 2 3 2 3 ...
$ Mortgage : int  0 0 0 0 0 155 0 0 104 0 ...
$ PLoan   : int  0 0 0 0 0 0 0 0 0 1 ...
$ Securities: int  1 1 0 0 0 0 0 0 0 0 ...
$ CDAmount : int  0 0 0 0 0 0 0 0 0 0 ...
$ Online   : int  0 0 0 0 0 1 1 0 1 0 ...
$ CreditCard: int  0 0 0 0 1 0 0 1 0 0 ...
```

For analysis we don't need variables such as ID and ZIP therefore we are going to remove them from the dataset. Variables found to be having wrong datatypes all variables should be changed to Numeric datatype as we don't intend to perform mathematical operations on them. Moreover, there are about 7 variables which needs to be changed to FACTORS or CATEGORICAL namely, CreditCard, Online, CD Amount, Securities, PLoan, Education, FMembers.

### 3.3 Missing Value Treatment

FMembers variable has NA values of about 36% which needs to be treated for further analysis.

<< EXCLUDED FROM PORTFOLIO ITEM >>

### 3.4 Outlier Treatment

<< EXCLUDED FROM PORTFOLIO ITEM >>

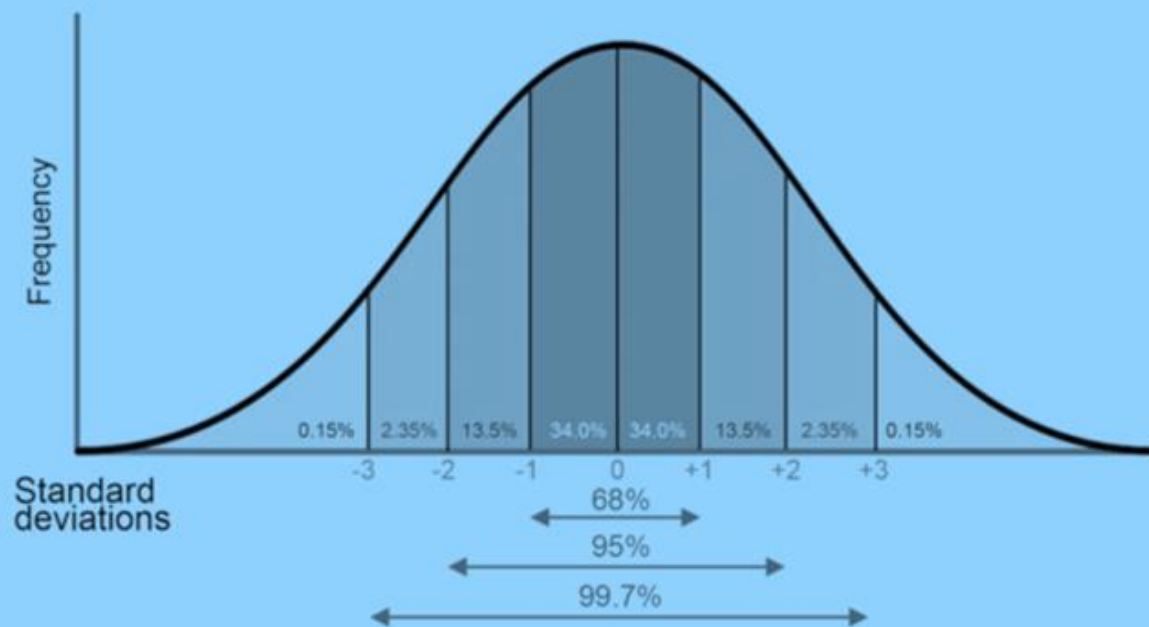
### 3.5 Variable Transformation/Feature Creation

We need to split data into two samples before applying K-FOLD for cross validation. Training and Testing Machine Learning Models.

<< EXCLUDED FROM PORTFOLIO ITEM >>

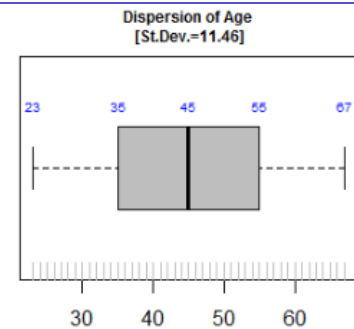
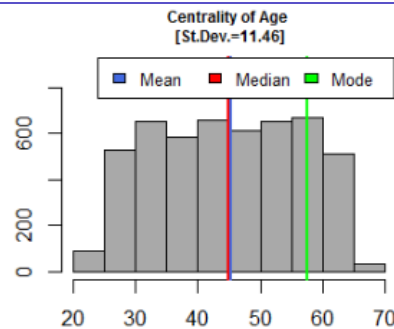


## 3.6 Univariate Analysis



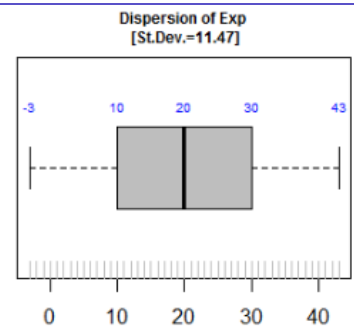
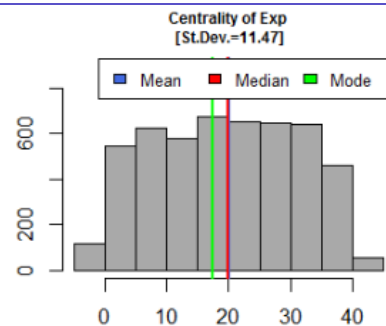
## Age

Multimodal distribution which indicates there are atleast three different groups of data. Very Wide dispersion of about 11.4 standard deviation. No outliers in this random variable.



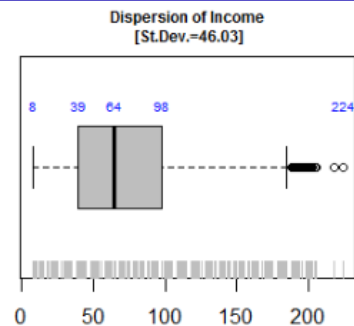
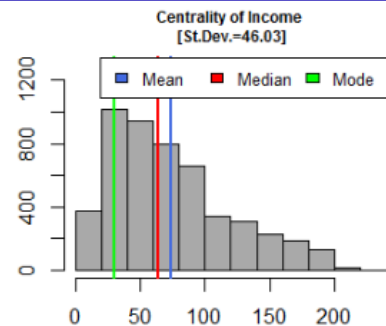
## Experience – Exp

Single Peak around 20 with very wide dispersion having 11.47 standard deviation. There are no outliers to be found. Almost symmetrical distribution.



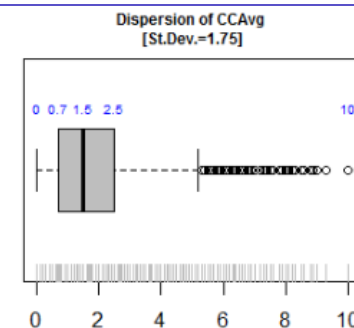
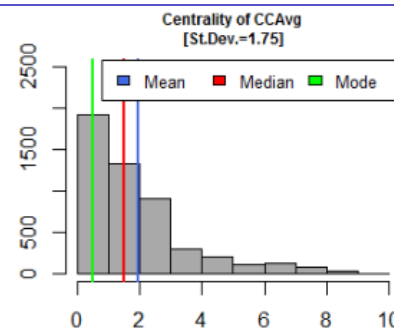
## Income

Right Skewed distribution having unimodal pattern. There are numerous outliers in the variables on upper end. Standard Deviation of 46.03.



## Credit Card Avg

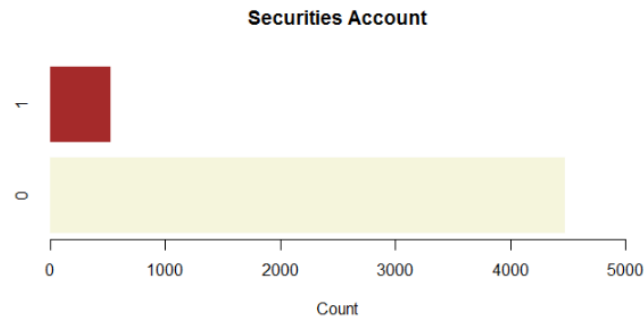
Heavily skewed to the right. With several outliers on the upper end of the distribution. Having narrow standard deviation of 1.75.



<< EXCLUDED FROM PORTFOLIO ITEM >>

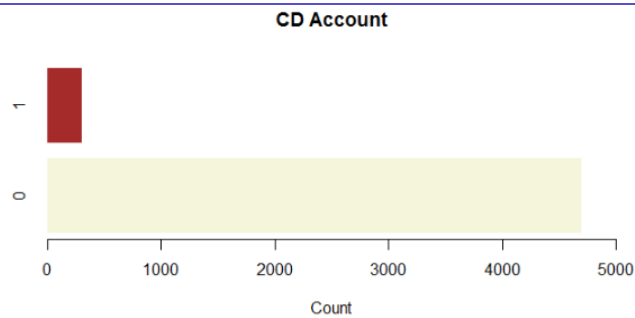
### Securities

Similar to PLoan about 90% observations with NO securities account associated with them. Only 9.5% does.



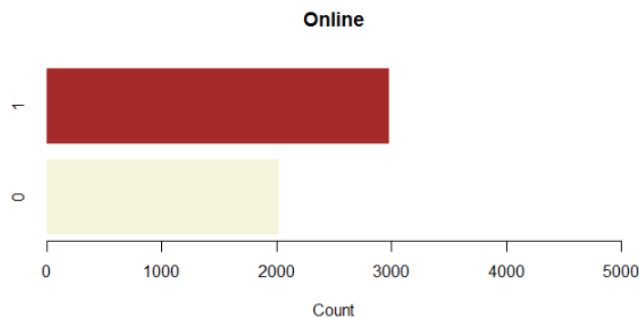
### CD

6% customers have certificate of deposit account with the bank whereas 94% does not.



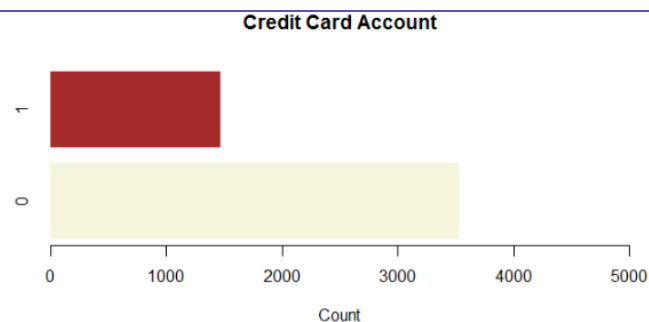
### Online

More than 60% observations of Online Account Settings whereas only 40% does not avail online services.

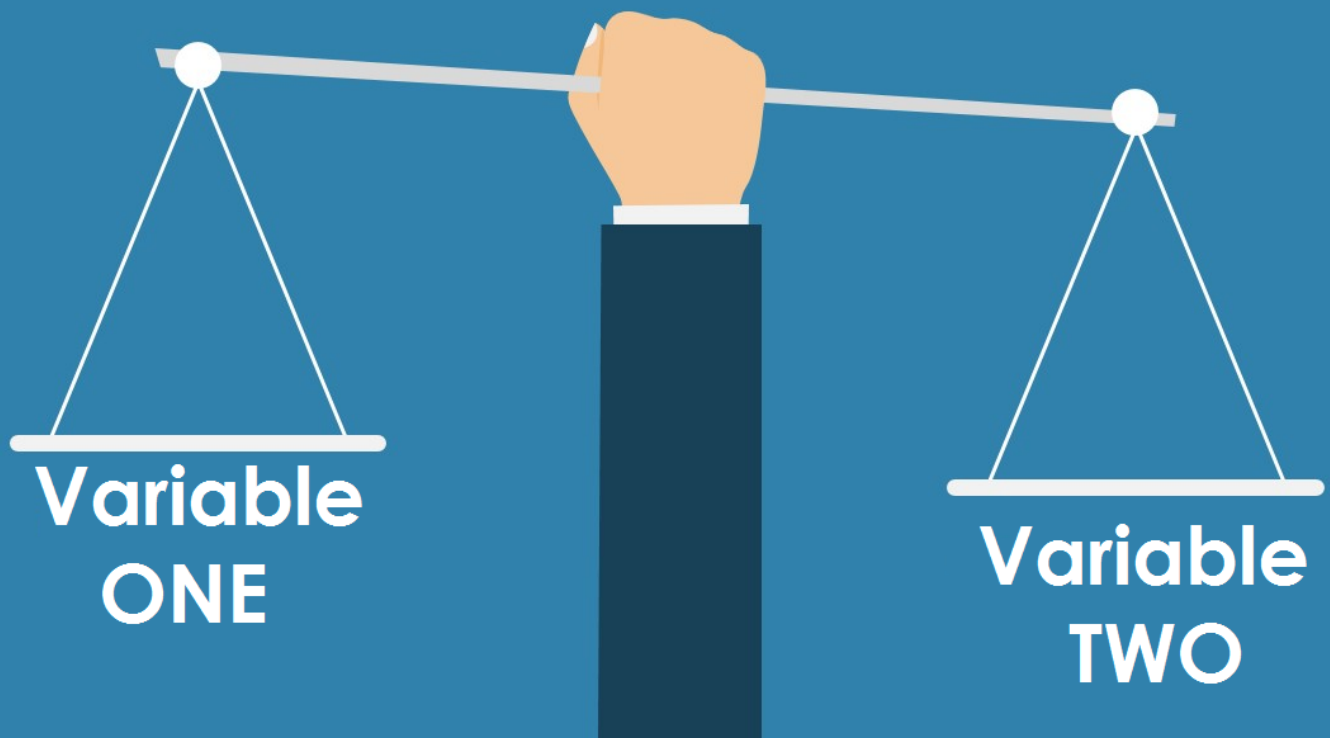


### Credit Card Account

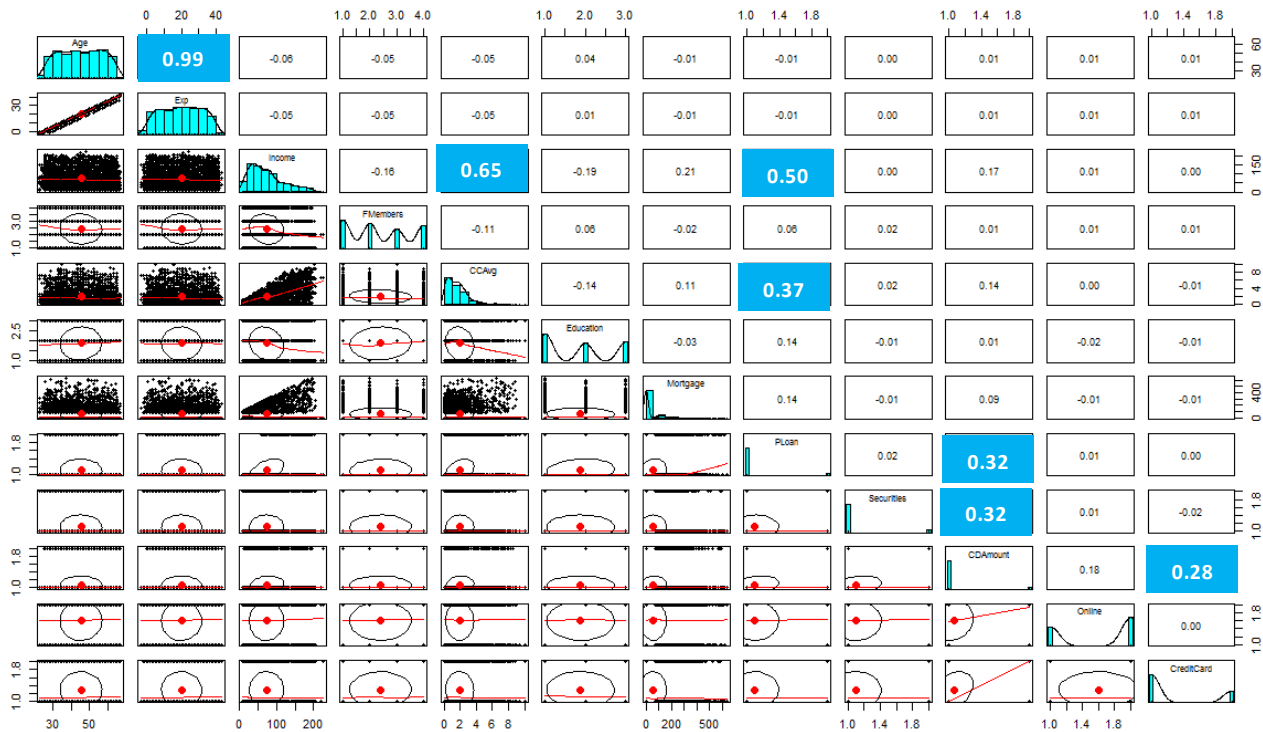
There are about 25% observations with Credit Cards whereas 75% does not have it.



## 3.7 Bivariate Analysis



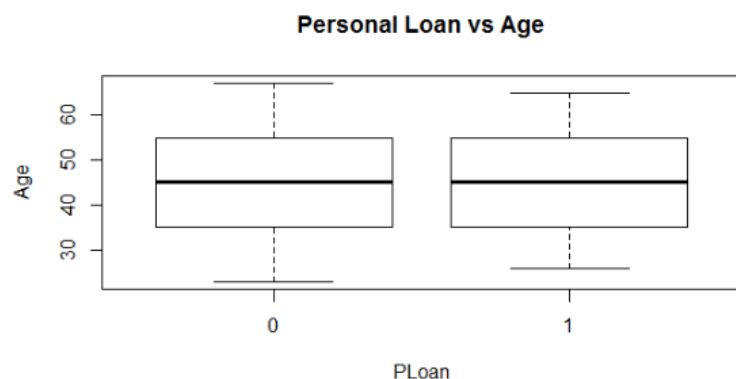
Correlation matrix is showing all the variables of the dataset and their corresponding strengths of correlation:



There are 72 bivariate analysis prospects out of which 7 seems to be having more than 25% correlation strength. However, we will conduct bivariate analysis on our area of interest “PLoan” and all its pairs, because there are no significant correlations apart from Income + Exp.

### PLoan vs Age

It is evident from plot that there is no variation in this pair due to each other. However, on close inspection you can see the Range of “1-PLoan” shrinks from both sides if compared with “0-PLoan”.

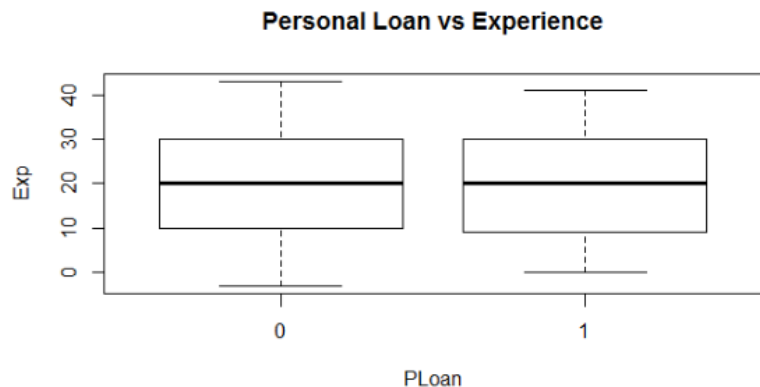






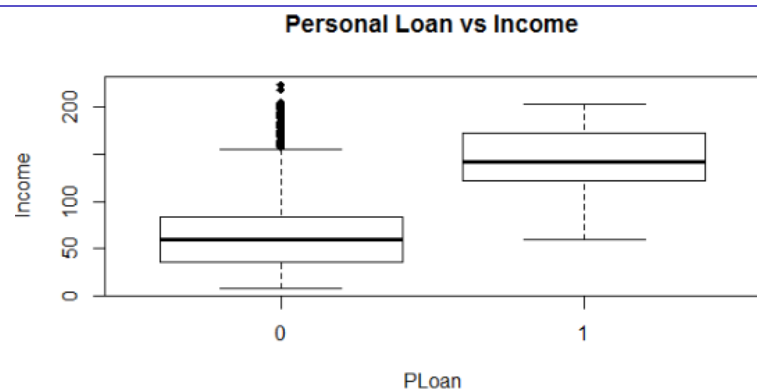
### PLoan vs Experience

Same behavior as PLoan+Age where “1” shrinks slightly from “0” and Mean of both are the same.



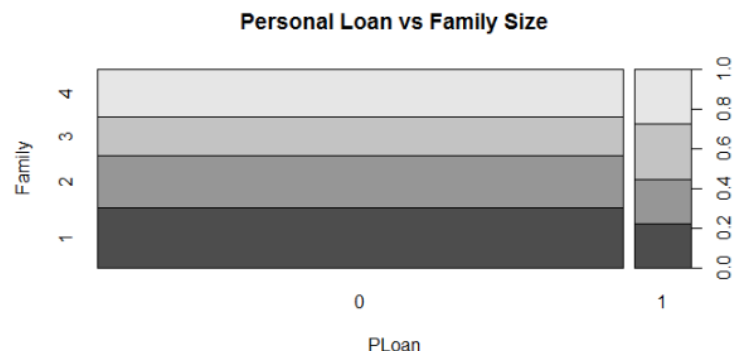
### PLoan vs Income

Distribution at “0-PLoan” is heavily right skewed with several outliers but it flips to left skewness with no outliers at “1-PLoan”.



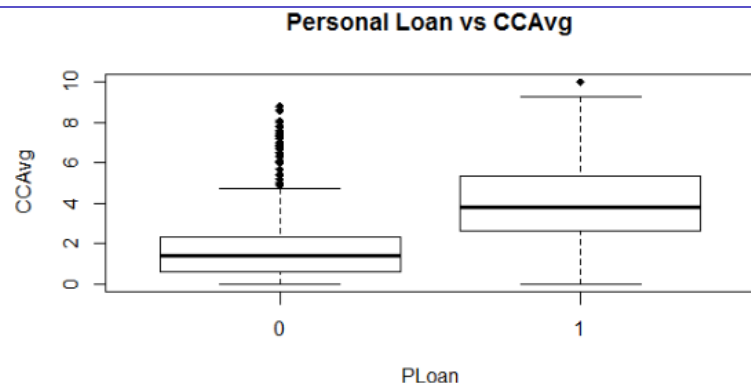
### PLoan vs FMembers

Looking at the plot it is clearly visible that in “0-PLoan” family size is uneven and irregular however at “1-PLoan” it gets somewhat uniform having sort of equal share around 25% of the sample.



### PLoan vs CCAvg

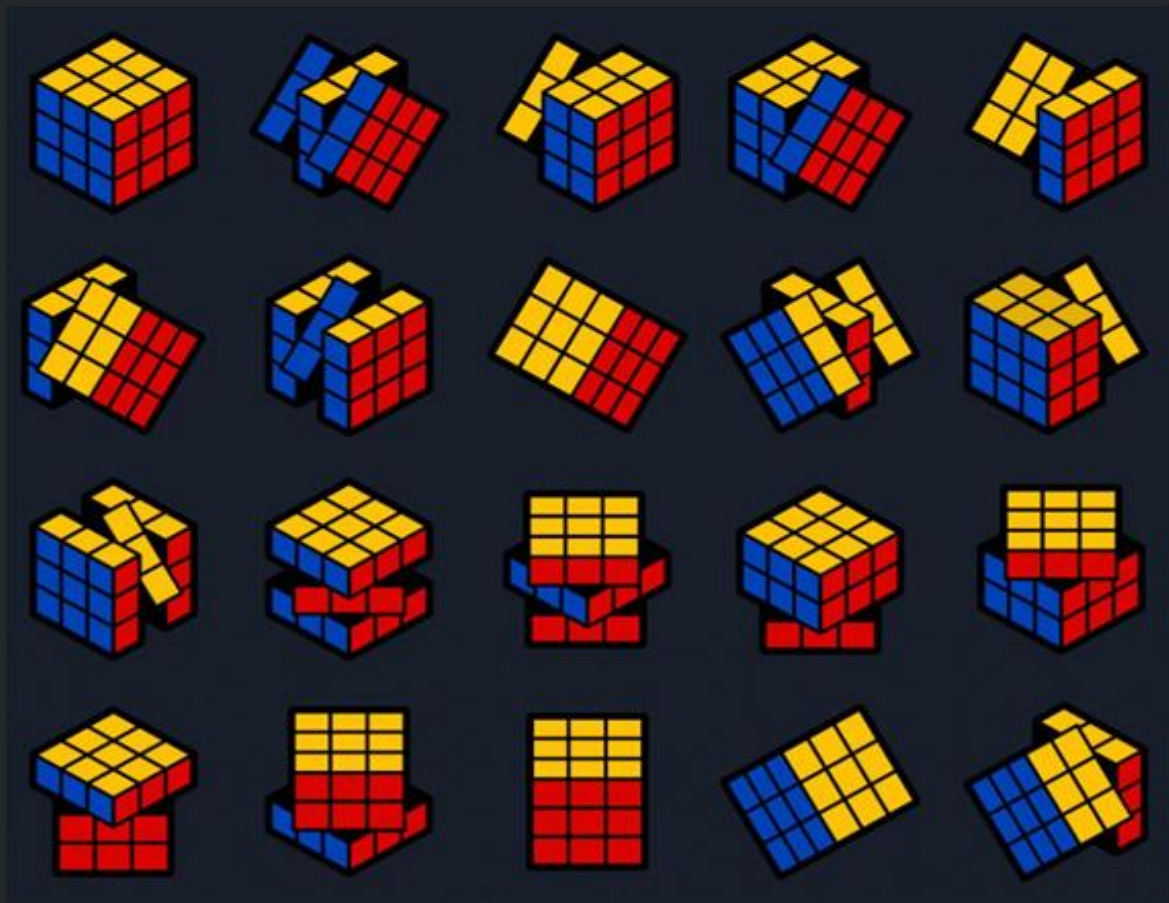
Range at “0-PLoan” is narrower with some outliers on the upper tail and heavily skewed to the right. However, at “1-PLoan” dispersion increases with single outlier and very slight right skewness.



---

<< EXCLUDED FROM PORTFOLIO ITEM >>

## 3.8 Clustering



Let's conduct unsupervised machine learning for customer segmentation. We will use KMEANS algorithm for our segmentation problem due to following reasons:

1. Customer segmentation is done on transactional data of customers and usually it can carry big data as well. Therefore, we need efficient algorithm which is future proof and can crunch mountain of data when transactions may go up-to millions.
2. Secondly, our clustering should work on hyper spherical data well. KMeans work well in circle shaped clusters.

## KMEANS CLUSTERING

We need to prepare data for clustering:

### Transforming to Numeric Variables:

Remove categorical variables which are not required for segmentation exercise. We need to convert all our factor variables to Numeric variables for the purpose of clustering. Our dataset structure will be like below:

```
$ Age      : int  25 45 39 35 35 37 53 50 35 34 ...
$ Exp      : int  1 19 15 9 8 13 27 24 10 9 ...
$ Income   : int  49 34 11 100 45 29 72 22 81 180 ...
$ FMembers : num  4 3 1 1 4 4 2 1 3 1 ...
$ CCAvg    : num  1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
$ Education : num  1 1 1 2 2 2 2 3 2 3 ...
$ Mortgage : int  0 0 0 0 0 155 0 0 104 0 ...
```

### Scaling the data:

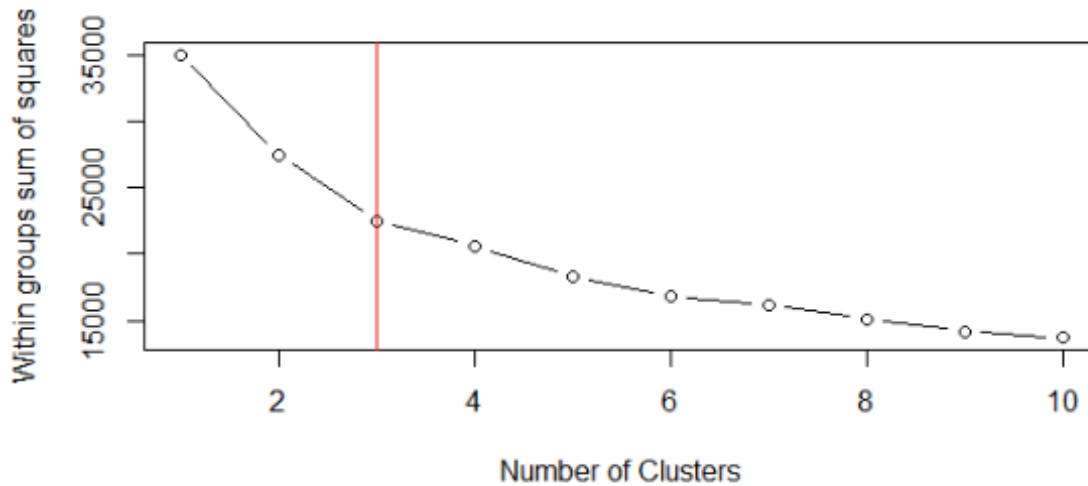
On inspecting the data, you can see that variables have varying data some are single digit numbers, some are with two digits and some are more than 3 digits. This means that since K-means work on Euclidean distance then variables having higher numbers gets preference in clustering which we don't want. Therefore, we scale the data so that all variables have equal opportunity for partitioning.

After scaling lets evaluate head of the data:

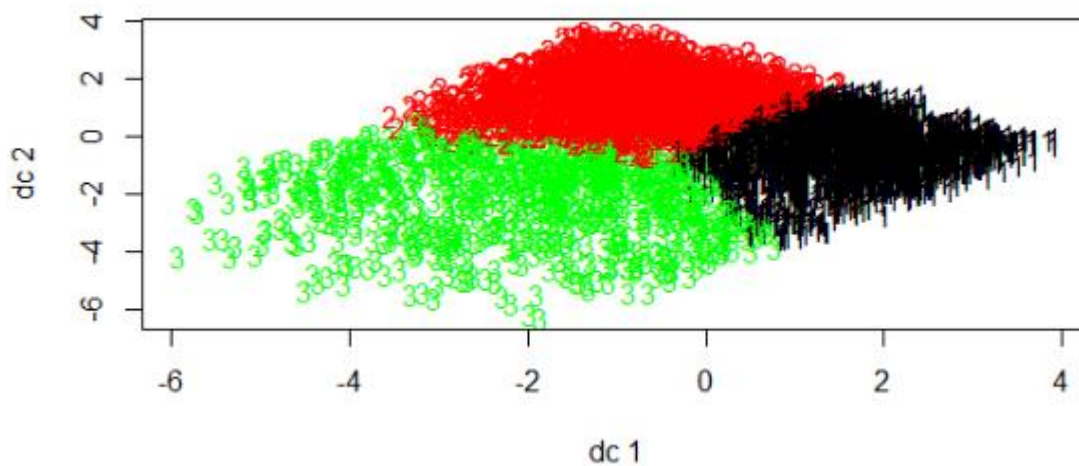
```
      Age    Exp Income FMembers CCAvg Education Mortgage
[1,] -1.77 -1.67  -0.54      1.40 -0.19    -1.05    -0.56
[2,] -0.03 -0.10  -0.86      0.53 -0.25    -1.05    -0.56
[3,] -0.55 -0.45  -1.36     -1.22 -0.54    -1.05    -0.56
[4,] -0.90 -0.97   0.57     -1.22  0.44     0.14    -0.56
[5,] -0.90 -1.06  -0.63      1.40 -0.54     0.14    -0.56
[6,] -0.73 -0.62  -0.97      1.40 -0.88     0.14     0.97
```



Elbow Method on WSS plot to find out about the optimum range of K-means clusters:



We will perform clustering with 3 number of clusters. Below is the plot of the clustering result:



### Customer Profile:

Cluster	Freq	Age	Exp	Income	FMembers	CCAvg	Education	Mortgage
1	2136	55.50515	30.201779	57.63764	2.399813	1.342683	1.963951	43.6264
2	2005	35.09875	9.844389	60.16808	2.600998	1.381661	1.953117	44.78105
3	859	43.95809	18.945285	145.65774	1.906868	4.716519	1.506403	115.85797

### Interpretation:

It seems that for Thera Bank CLUSTER-3 profile is the most suitable for Personal Loan, having higher Mortgage, CCAvg spendings:

### Path:

- Age around 43 years (Mid Age)
- Experience around 19 years (Mid Experience)
- Income around 145,000 (High Income)
- Family Size of 2 members (Small Family)

Cluster 3 yields higher response rates for cross-selling and having share in population of about 20%.

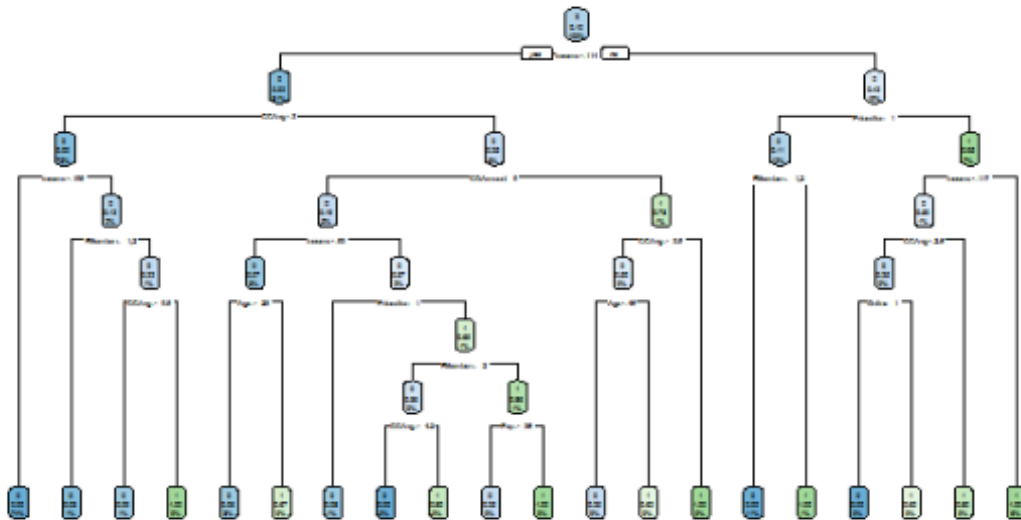
Cluster 1 and Cluster 2 are almost the same as far as their spending is concerned, though they have one thing which makes them distinct groups and that is “Age”. Cluster 2 is early age with Parents most of the people therefore their family size is bigger. However, cluster 1 are aged group of people having kids in their family. One interesting thing to note is “Low – Medium Income” in both groups. 80% customer belongs to these two groups.

## 4 CART – Decision Tree



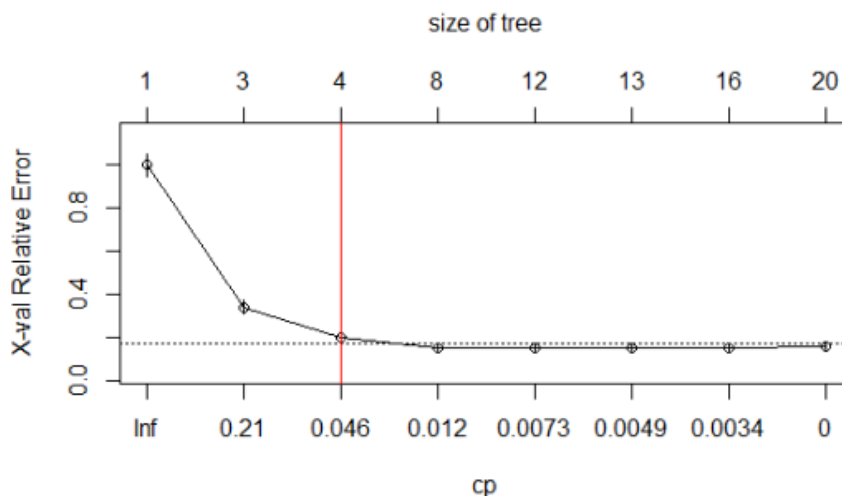
# Decision Tree Analysis

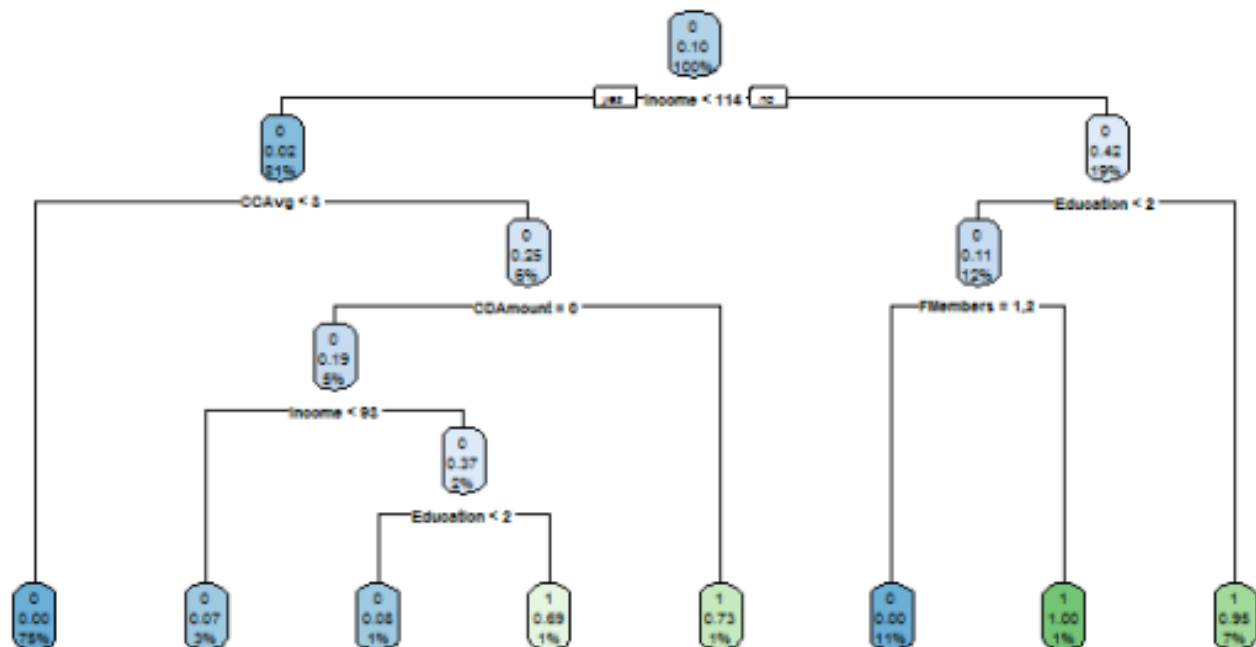
Our first attempt in CART Decision Tree model is going to be with maximum cost complexity without any consideration for overfitting. After making a tree from rpart function we get below tree:



	CP	nsplit	rel error	xerror	xstd
1	0.3303571	0	1.000000	1.00000	0.051870
2	0.1398810	2	0.339286	0.33929	0.031255
3	0.0148810	3	0.199405	0.19940	0.024127
4	0.0089286	7	0.133929	0.15179	0.021099
5	0.0059524	11	0.098214	0.15179	0.021099
6	0.0039683	12	0.092262	0.14881	0.020894
7	0.0029762	15	0.080357	0.14881	0.020894
8	0.0000000	19	0.068452	0.15476	0.021302

We set our threshold for pruning at 0.13 as Xerror becomes stagnant at this relative error.





	CP	nsplit	rel error	xerror	xstd
1	0.330357	0	1.00000	1.00000	0.051870
2	0.139881	2	0.33929	0.33929	0.031255
3	0.014881	3	0.19940	0.19940	0.024127
4	0.013000	7	0.13393	0.16071	0.021701

### Prune Tree:

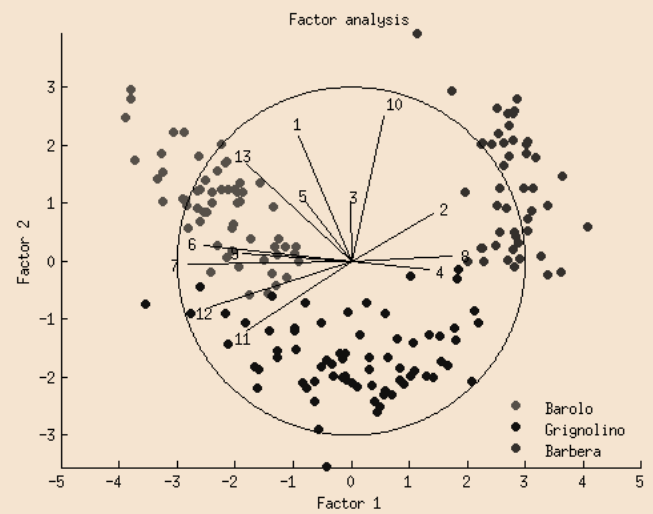
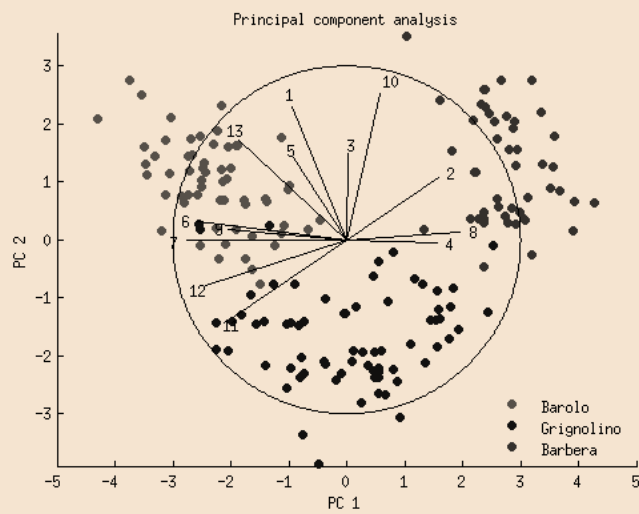
```

1) root 3500 336 0 (0.904000000 0.096000000)
2) Income< 113.5 2839 56 0 (0.980274745 0.019725255)
4) CCAvg< 2.95 2640 7 0 (0.997348485 0.002651515) *
5) CCAvg>=2.95 199 49 0 (0.753768844 0.246231156)
10) CDAmount< 0.5 177 33 0 (0.813559322 0.186440678)
20) Income< 92.5 109 8 0 (0.926605505 0.073394495) *
21) Income>=92.5 68 25 0 (0.632352941 0.367647059)
42) Education< 1.5 36 3 0 (0.916666667 0.083333333) *
43) Education>=1.5 32 10 1 (0.312500000 0.687500000) *
11) CDAmount>=0.5 22 6 1 (0.272727273 0.727272727) *
3) Income>=113.5 661 280 0 (0.576399395 0.423600605)
6) Education< 1.5 417 47 0 (0.887290168 0.112709832)
12) FMembers=1,2 370 0 0 (1.000000000 0.000000000) *
13) FMembers=3,4 47 0 1 (0.000000000 1.000000000) *
7) Education>=1.5 244 11 1 (0.045081967 0.954918033) *
  
```

Error Rate Train Sample	Error Rate Test Sample
0.0128	0.017



# 5 Random Forest

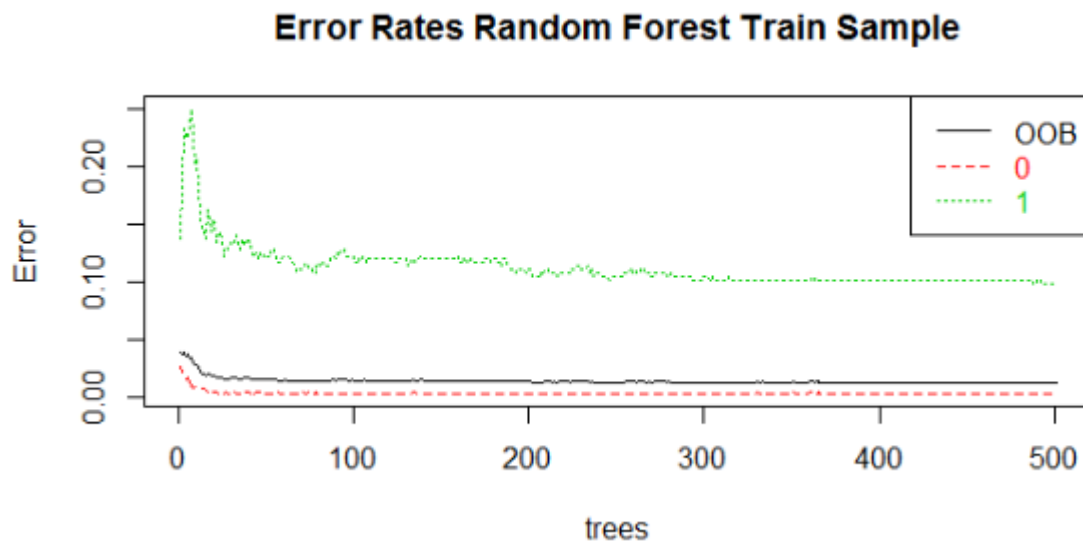


We start our random forest with 501 trees in the beginning and then we will tune it for optimum output.

Type of random forest: classification  
 Number of trees: 501  
 No. of variables tried at each split: 3  
 OOB estimate of error rate: 1.17%

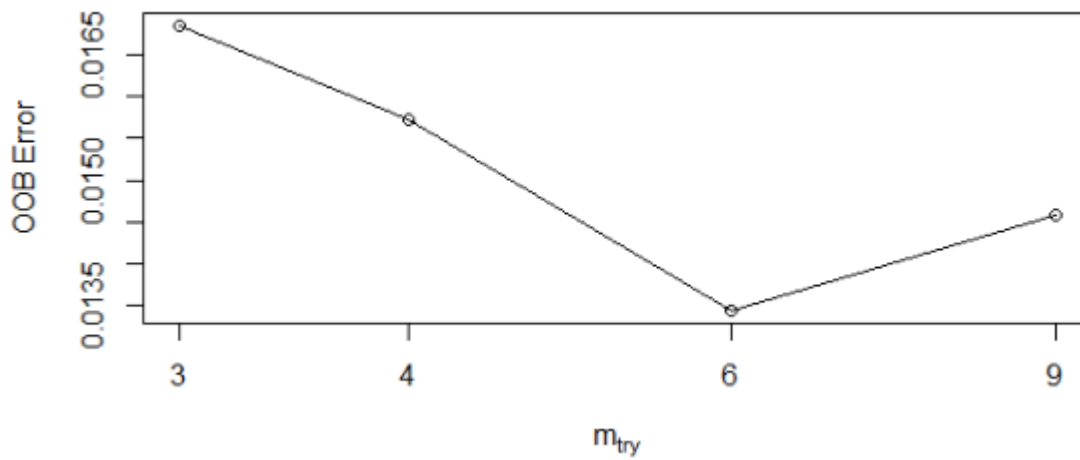
Confusion matrix:

	0	1	class.error
0	3156	8	0.002528445
1	33	303	0.098214286



Importance of random variables in the below table:

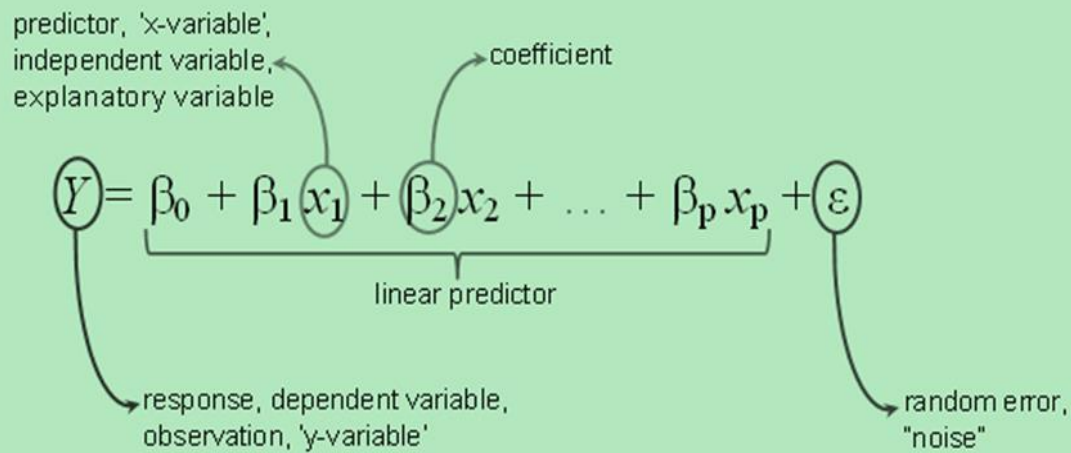
	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Age	5.742017e-03	0.0044125916	0.0056146374	12.986626
Exp	5.230618e-03	0.0021271887	0.0049408159	12.357312
Income	1.067789e-01	0.3623035935	0.1311825588	157.461434
FMembers	4.104341e-02	0.0506248972	0.0419638174	68.902630
CCAvg	3.096019e-02	0.0436876202	0.0321646118	63.922538
Education	5.674759e-02	0.1236132575	0.0631397007	119.715000
Mortgage	2.175284e-03	-0.0019505382	0.0017770994	15.183786
Securities	3.554892e-05	0.0031635272	0.0003317384	1.537400
CDAmount	3.357871e-03	0.0230781999	0.0052352997	30.385125
Online	1.553973e-04	0.0004721272	0.0001890109	1.618480
CreditCard	6.948328e-04	0.0028388338	0.0009046311	2.445369
Cluster	4.720398e-02	0.0458807700	0.0470802384	54.017611



	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Age	7.02202550	0.1004102	6.697624	4.0557989
Exp	5.33154974	1.0981812	5.580673	3.5703174
Income	123.38477563	78.3327083	143.289114	159.3313938
FMembers	85.28610642	45.8503726	84.666077	79.5514067
CCAvg	22.47408661	18.9013564	26.439952	47.1420802
Education	145.64431785	75.4987571	145.770892	166.0946931
Mortgage	5.22791107	-4.2325450	3.897591	4.3816170
Securities	-0.09646016	2.6080948	2.217230	0.3182155
CDAmount	9.02844814	10.2072496	11.928979	20.1327083
Online	2.18834413	1.1022228	2.431598	0.2437055
CreditCard	2.11115411	3.4861491	3.875219	0.7041286
cluster	15.47506037	5.0229575	15.874357	35.6306697

Error Rate Train Sample	Error Rate Test Sample
0.01	0.014

# 6 Model Performance Measures



### CONFUSION MATRIX

CART			RANDOM FOREST		
	0	1		0	1
0	1347	9	0	1353	3
1	17	127	1	18	126
ERROR: 0.017			ERROR: 0.014		
RANDOM FOREST is 3% better than CART as per the test sample provided as far as confusion matrix is concerned.					

### DECILING

CART			RANDOM FOREST			
[0, 0.00265)	[0.00265, 0.0833)	[0.0833, 1]	[0, 0.002)	[0.002, 0.016)	[0.016, 0.237)	[0.237, 1]
196	1148	156	1043	147	160	150

### RANK ORDER TABLE

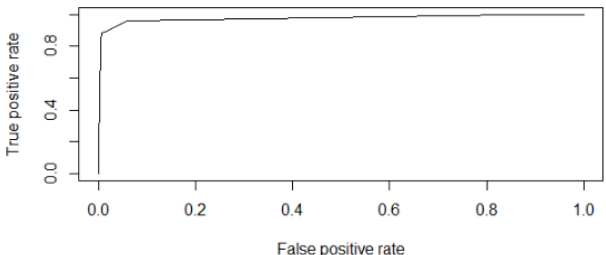
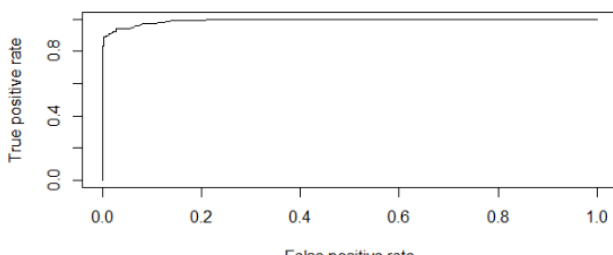
#### CART

Rank	Deciles	cnt	cnt_tar1	cnt_tar0	rrate	cum_resp	cum_non_resp	cum_rel_resp	cum_rel_non_resp	ks
1:00	[0.0833,1]	156	129	27	82.69	129	27	89.58	1.99	87.59
2:00	[0.00265,0.0833)	1148	15	1133	1.31	144	1160	100.00	85.55	14.45
3:00	[0,0.00265)	196	0	196	0.00	144	1356	100.00	100.00	0.00



## RANDOM FOREST

Rank	deciles	Cnt	cnt_tar1	cnt_tar0	rrate	cum_resp	cum_non_resp	cum_rel_resp	cum_rel_non_resp	ks
1	[0.237,1]	150	131	19	87.33	131	19	90.97	1.40	89.57
2	[0.016,0.237)	160	10	150	6.25	141	169	97.92	12.46	85.46
3	[0.002,0.016)	147	3	144	2.04	144	313	100.00	23.08	76.92
4	[0,0.002)	1043	0	1043	0.00	144	1356	100.00	100.00	0.00

CART	RANDOM FOREST
<b>ROC CURVE</b> 	<b>ROC CURVE</b> 
KS = 0.8971239	KS = 0.9094764
AUC = 0.9763474	AUC = 0.992505
GINI = 0.8856783	GINI = 0.8994366
CONCORDANCE = 0.9569199	CONCORDANCE = 0.9921849
DISCORDANCE = 0.04308014	DISCORDANCE = 0.007815061

A series of colorful, fan-like geometric shapes (triangles and quadrilaterals) in shades of blue, red, yellow, green, orange, and pink, arranged in a semi-circular arc at the top of the page.

# Conclusion

A series of colorful, fan-like geometric shapes (triangles and quadrilaterals) in shades of pink, orange, maroon, yellow, green, and blue, arranged in a semi-circular arc at the bottom of the page.

**We used clustering to not only develop customer profile for campaign costs savings but also to boost performance of Supervised Learning of classification (CART + RANDOM FOREST). Although there is not much difference in the performance measures of both CART and RANDOM FOREST models but with certainty, we can conclude that Random Forest model is shade better than CART because in all aspects of performance measurement this model performed better.**

<< EXCLUDED FROM PORTFOLIO ITEM >>