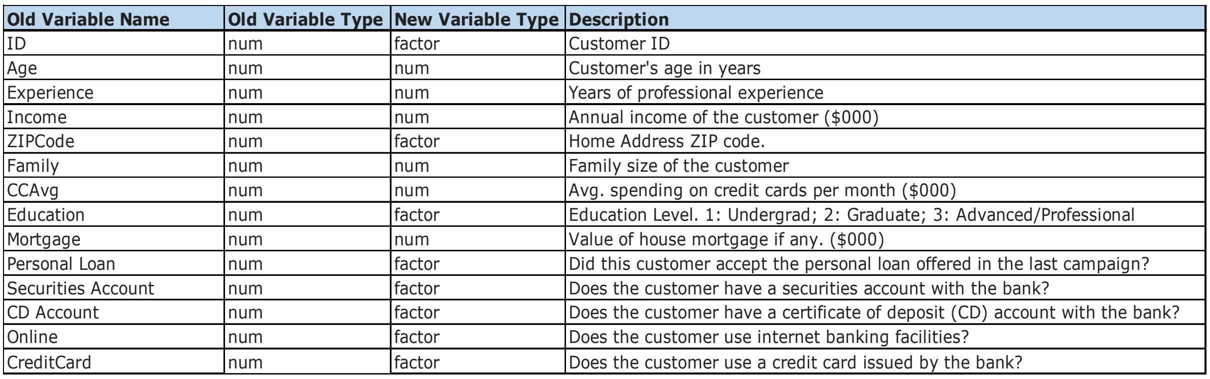
Kudakwashe Nyikadzino

1.) Exploratory Data Analysis

1.1) Variable and Missing Value Treatment

1.1.1) Variable Identification

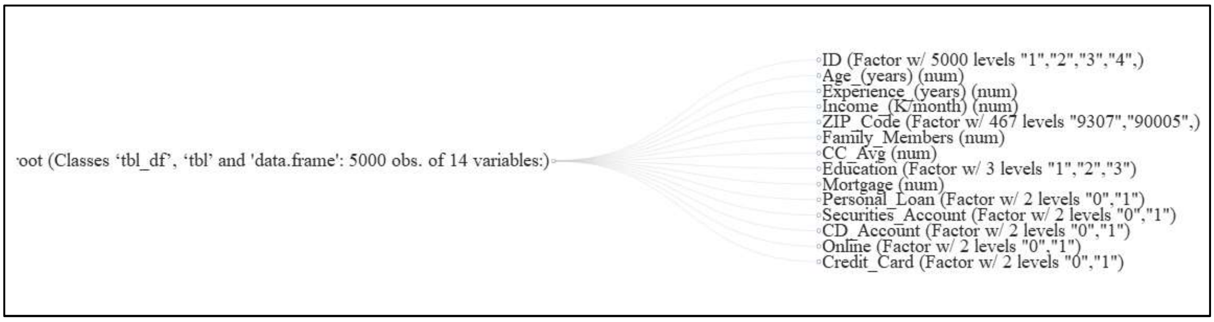
There are 14 variables and 5000 observations in the dataset. All were numeric initially and using the variable their descriptions below, some of the variable types have been converted to factors (*Table1.1.1a*).



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

Table1.1.1a

Variable names have also been changed putting them in the correct form for example ‘Personal Loan’ has been changed to ‘Personal_Loan’ for example (*Chart1.1b)*.



·oot (Classes ‘tbl_df’, ‘tbl’ and 'data.frame': 5000 obs. of 14 variables:)
ID (Factor w/ 5000 levels "1","2","3","4",)
Age_(years) (num)
Experience_(years) (num)
Income_(K/month) (num)
ZIP_Code (Factor w/ 467 levels "9307","90005",)
Family_Members (num)
CC_Avg (num)
Education (Factor w/ 3 levels "1","2","3")
Mortgage (num)
Personal_Loan (Factor w/ 2 levels "0","1")
Securities_Account (Factor w/ 2 levels "0","1")
CD_Account (Factor w/ 2 levels "0","1")
Online (Factor w/ 2 levels "0","1")
Credit_Card (Factor w/ 2 levels "0","1")

Chart1.1.1b

The “ID” variable has been removed because it does not provide any additional information for our analysis.

1.1.2) Missing Variable Identification

Family Members is the only variable with missing values in some observations (*Chart1.2a*). The treatment of the missing values applied is removing all the observations with missing values. These observations represent a low proportion of the entire dataset. Only 18 of the original 5000 removed from the data to be used for analysis which is 0.36% of the observations. We still have enough data points that can be used without introducing bias and he remaining observations are 4982.

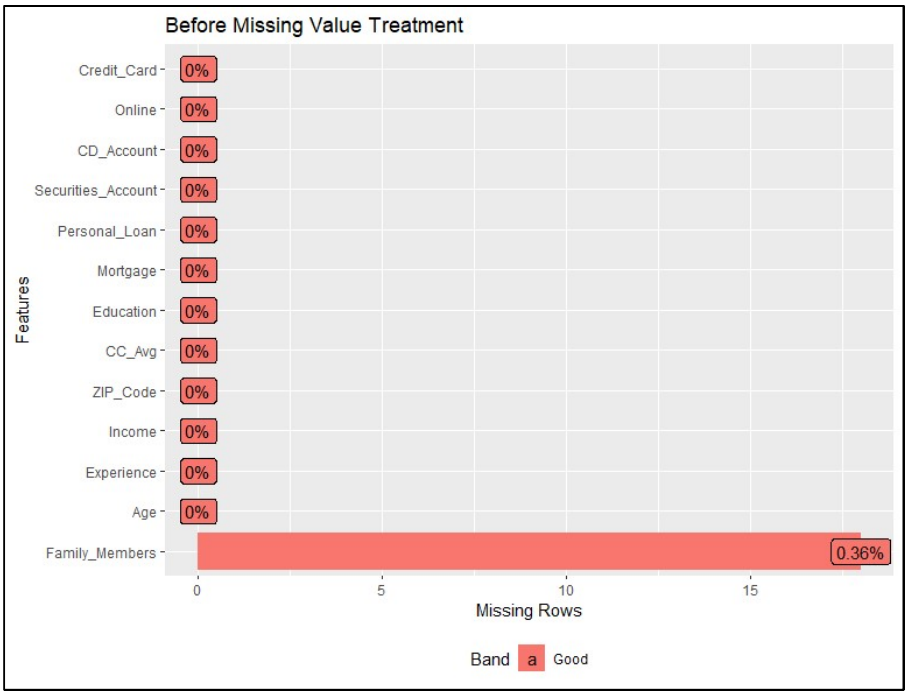


Chart1.1.2a

This is now the view of the dataset after the missing values have been removed (*Chart1.1.12a).*

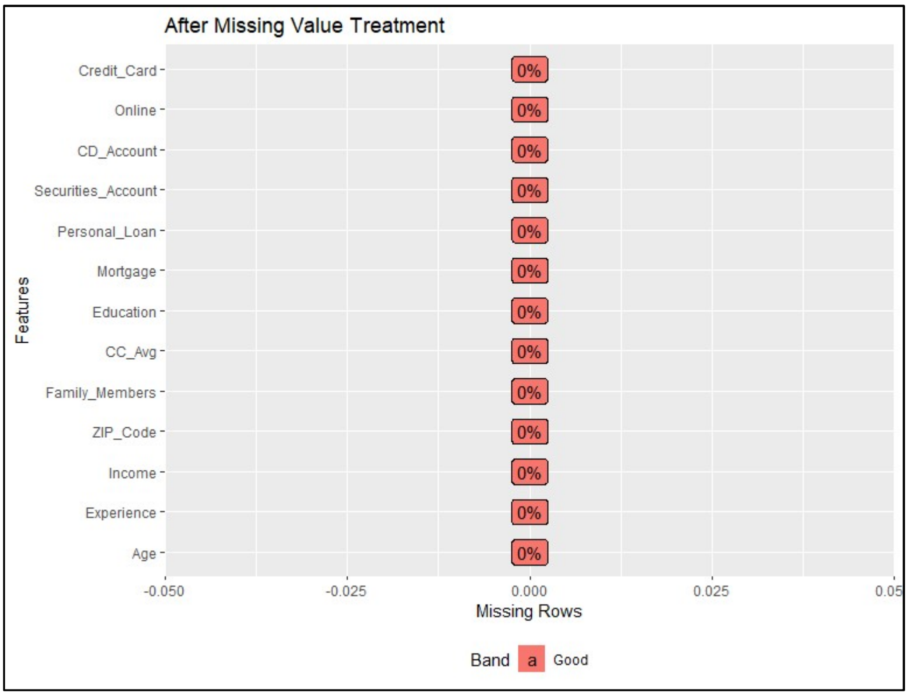


Chart1.1.2b

1.2) Univariate Analysis

1.2.1) Categorical Variables

There are 7 factor variables under consideration with the Zip Code having the most levels at 467. These are the home addresses of Thera Banks' customers by Zip Code. Education has 3 levels and the remaining 5 variables namely Personal Loan, Securities Account, CD

Account, Online and Credit Card have 2 levels. These two levels are either affirmative or negative states for each variable.

1.2.1.1) Education

Thera Bank's customers are grouped into three education levels namely undergraduate, graduate or advanced/professional. Most of the clients have an undergraduate level of education (2088) representing 41.9% of liability customers (*Chart1.2.1.1a)*. Customers with advanced/professional level of education (1495) representing 30.0% and graduate level (1399) at 28.1% liability customers.

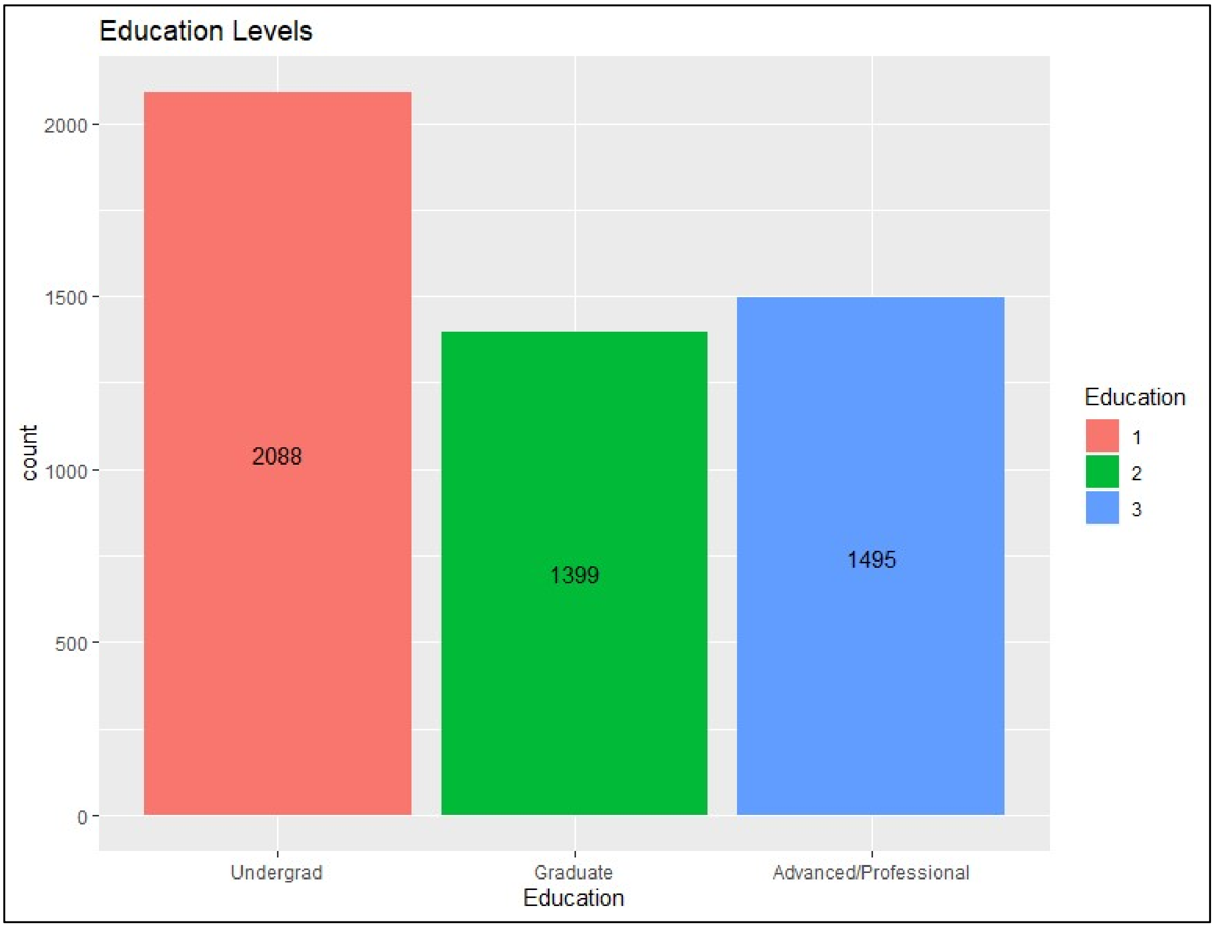


Chart1.2.1.1a

1.2.1.2) Zip Codes

The Zip Codes give an idea of where Thera Bank's liability customers are located. The Zip codes of home addresses are in *Chart1.2.1.2a* for the top 20 locations. With 168 customers located in Zip Code 94720.

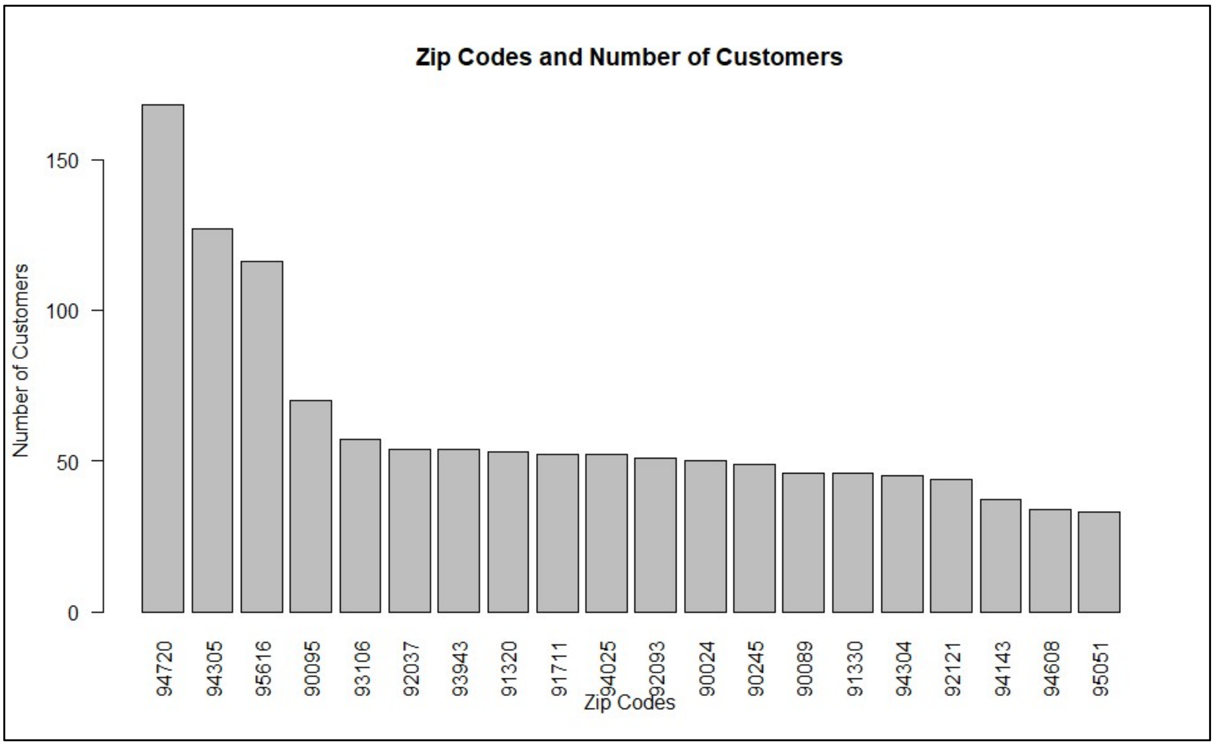


Chart1.2.1.2a

1.2.1.3) Personal Loan

Thera Bank ran a campaign to offer liability customers loans and 478 of these customers accepted the personal loans representing 9.6% of liability customers (*Chart1.2.1.3a*). Looking at features of customers that accepted the loan offer will enable us later to determine which customers to target for a loan offer.

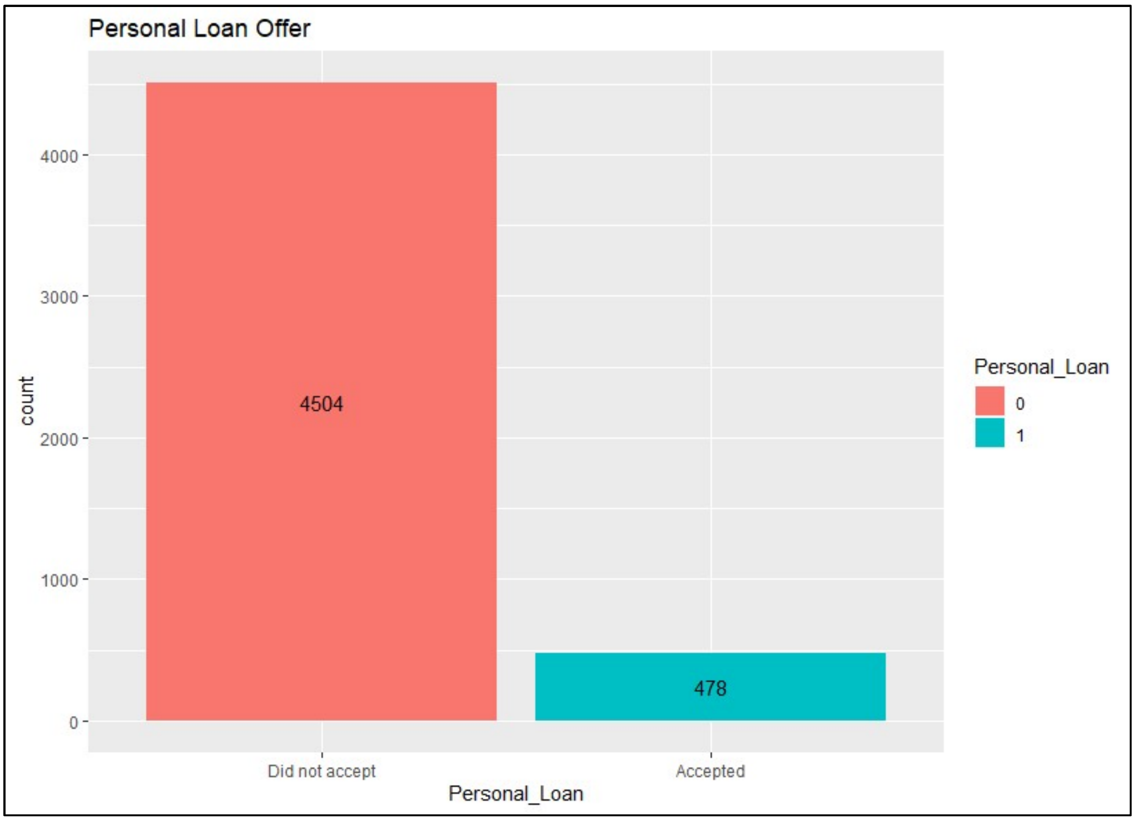
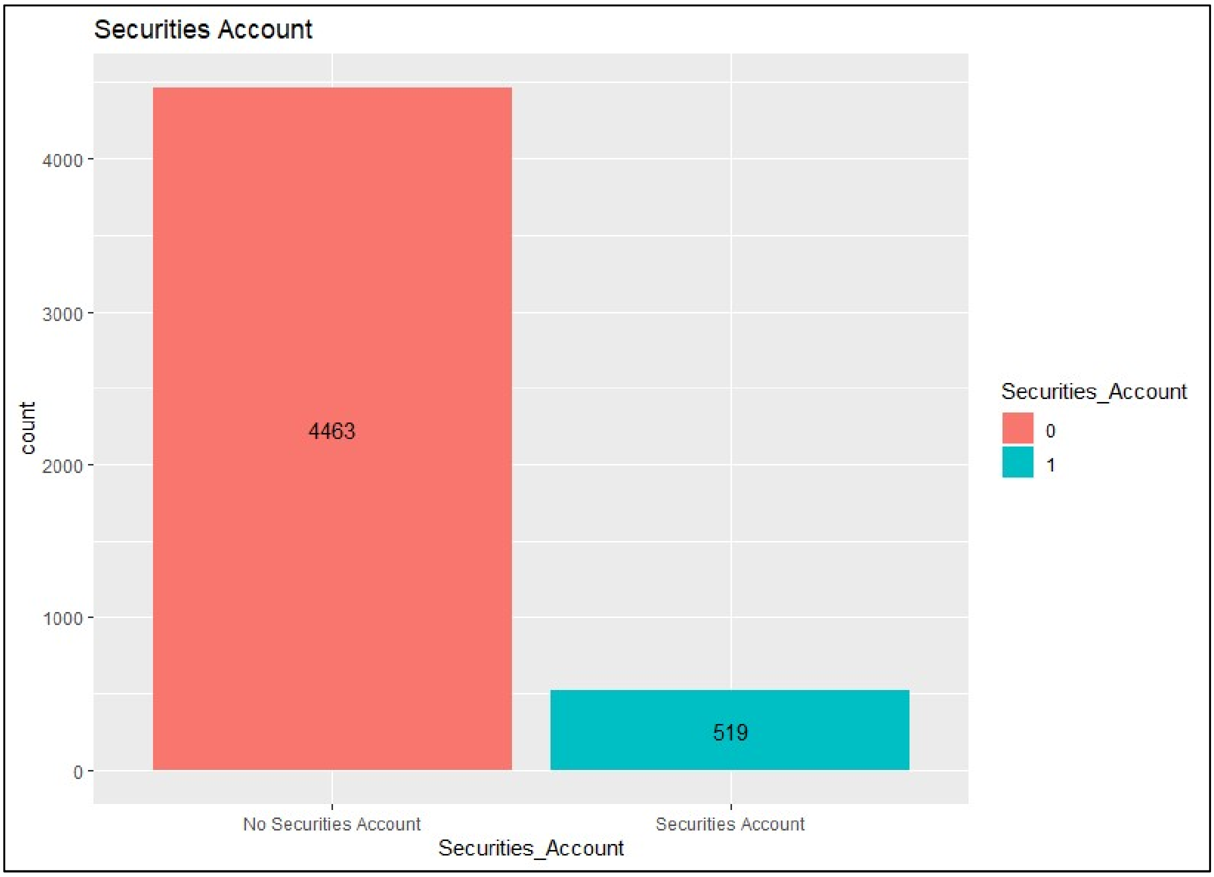


Chart1.2.1.2a

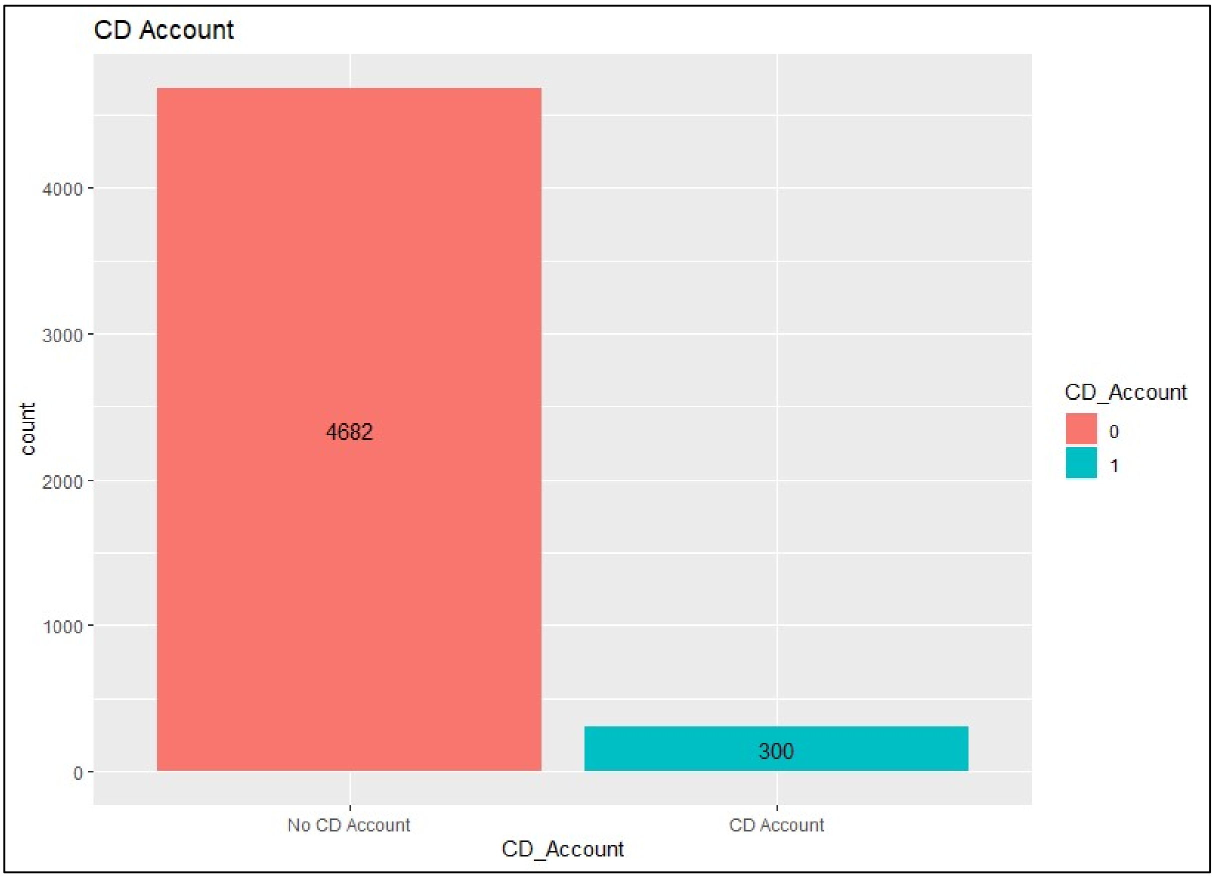
1.2.1.4) Securities Account

About 10.4% of the customers have a Securities Account with Thera Bank (*Chart1.2.1.4a*). This is a total of 519 liability customers. The proportions of those that accepted the Personal Loan offer and those that have Securities Account were almost similar at 9.6% and 10.4% respectively.



1.2.1.5) CD Account

A lower proportion of clients had a CD account (6.0%) compared to those with a Securities Account (10.4%) (*Chart1.2.1.5a*). Fewer customers having a CD Account compared to those with Securities Accounts may be as a result of the difference in product features for example returns or product restrictions.



1.2.1.6) Credit Card

A total of 1465 the liability customers have a credit card with Thera Bank (*Chart1.2.1.6a*). This 29.4% of the liability customers compared 9.6% of customers who accepted personal loans. The two products are different forms of borrowing and the difference between the two could be an indication of customers preferring credit cards to personal loans. It is also important to monitor whether increasing personal loan customers may reduce the number of credit card customers since both serve a similar need that is borrowing.

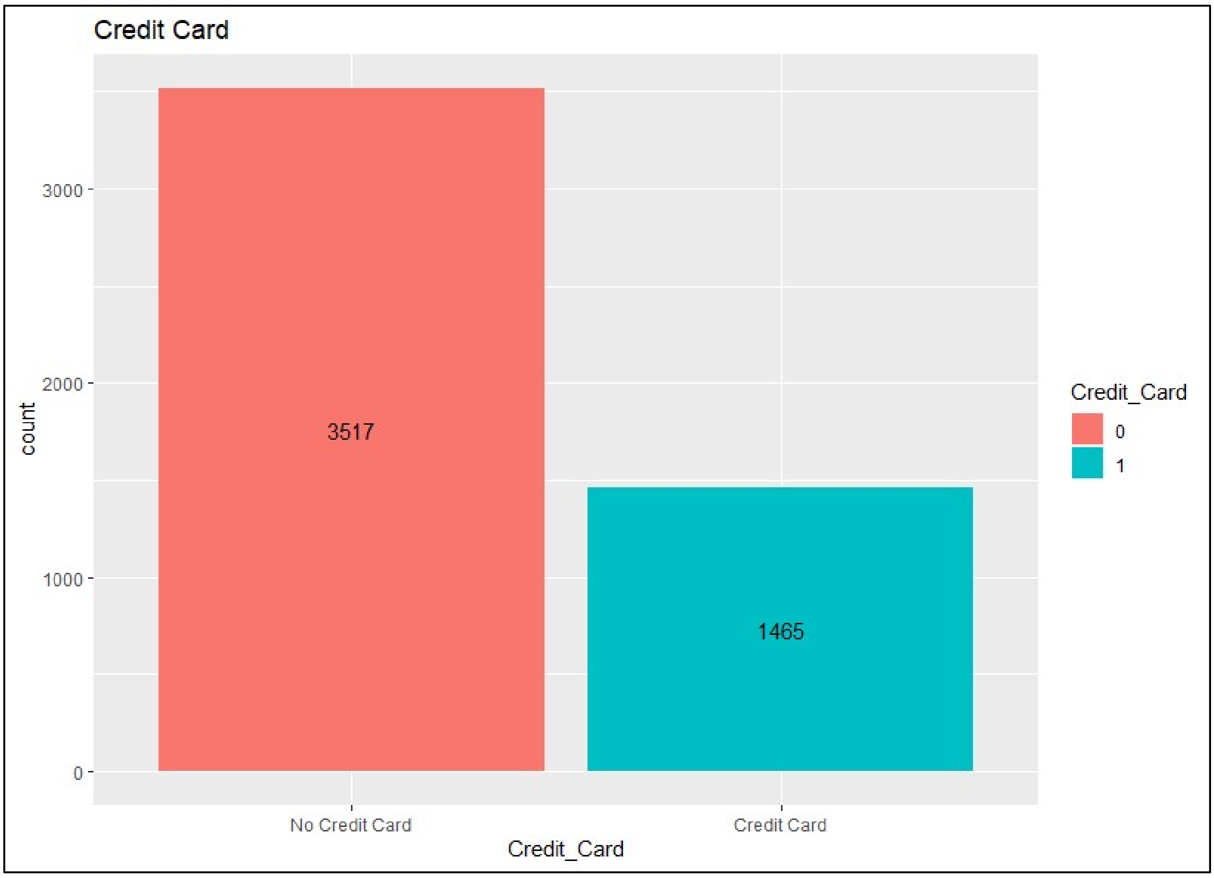


Chart1.2.1.6a

1.2.1.7) Online

More than half of the customers (59.6%) had internet banking facilities (*Chart1.2.1.7a)*. With such a high number, it may be helpful to identify whether personal loans are easily available online for customers to help with the conversion rate.

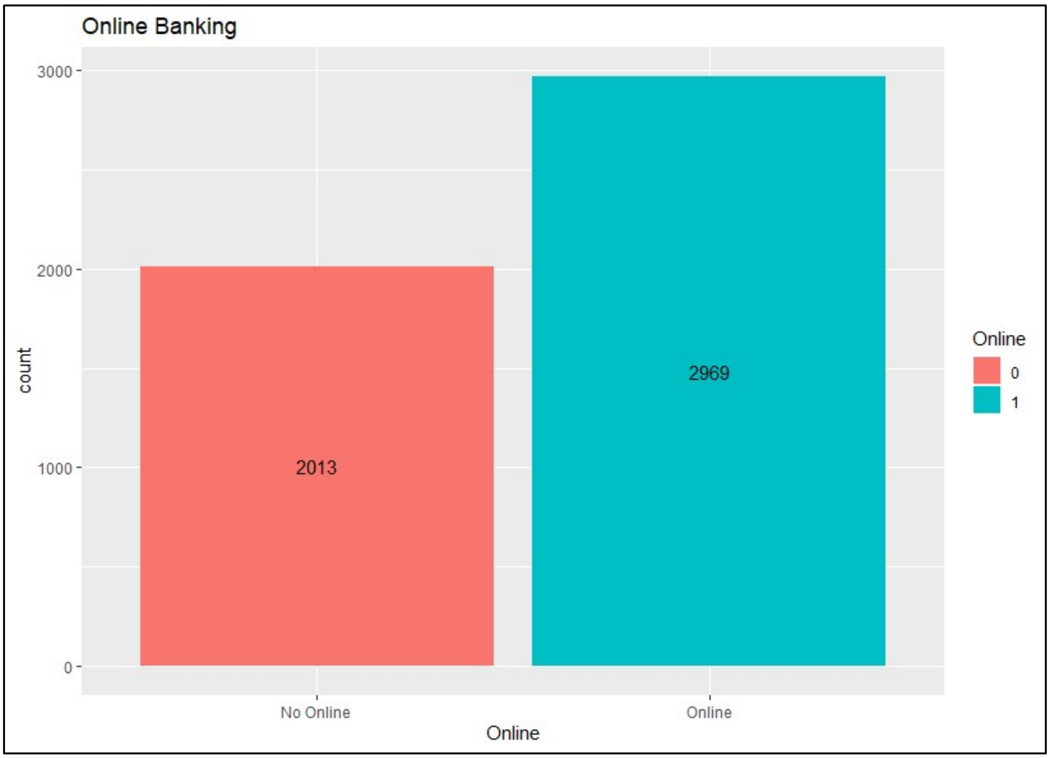


Chart1.2.1.7a

1.2.2) Numeric Variables

There are 6 numeric variables with Age and Education in years, Income, Credit Card average and Mortgage in ($000) and Family Size which is the number of family members.

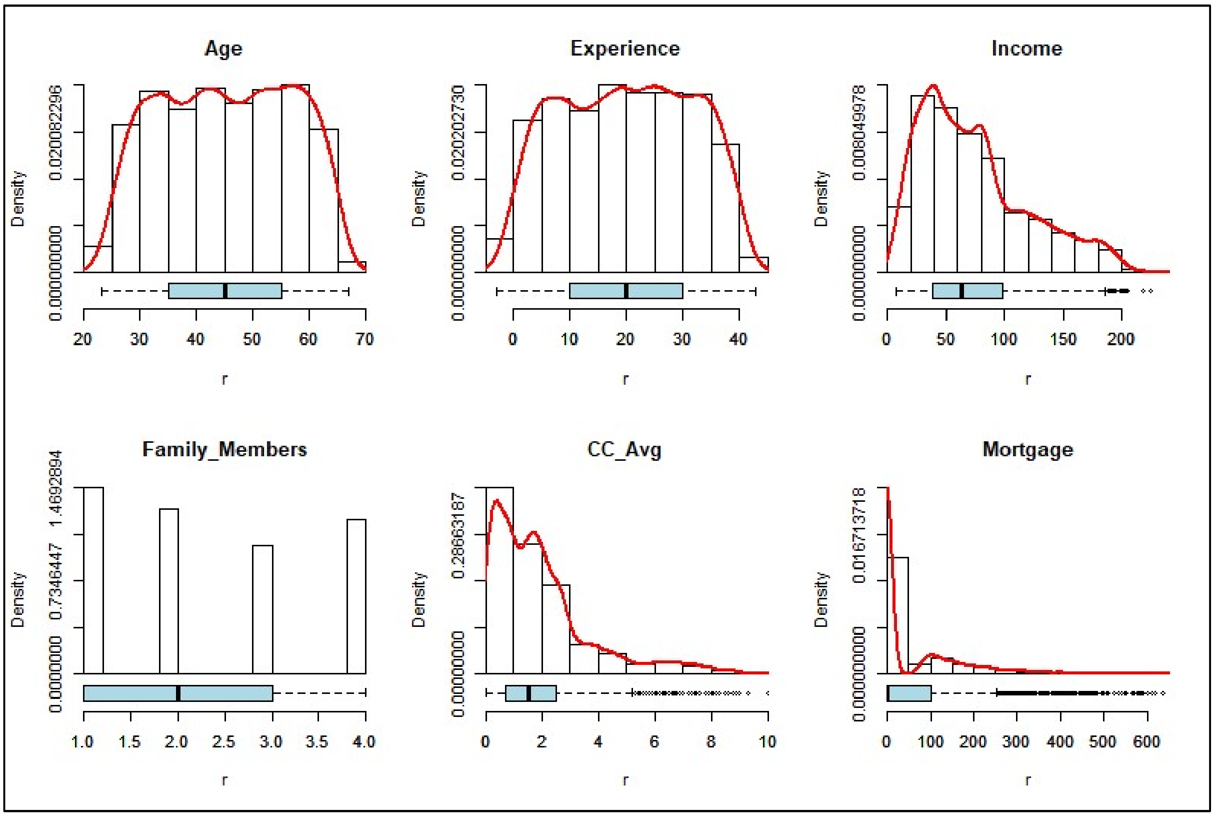
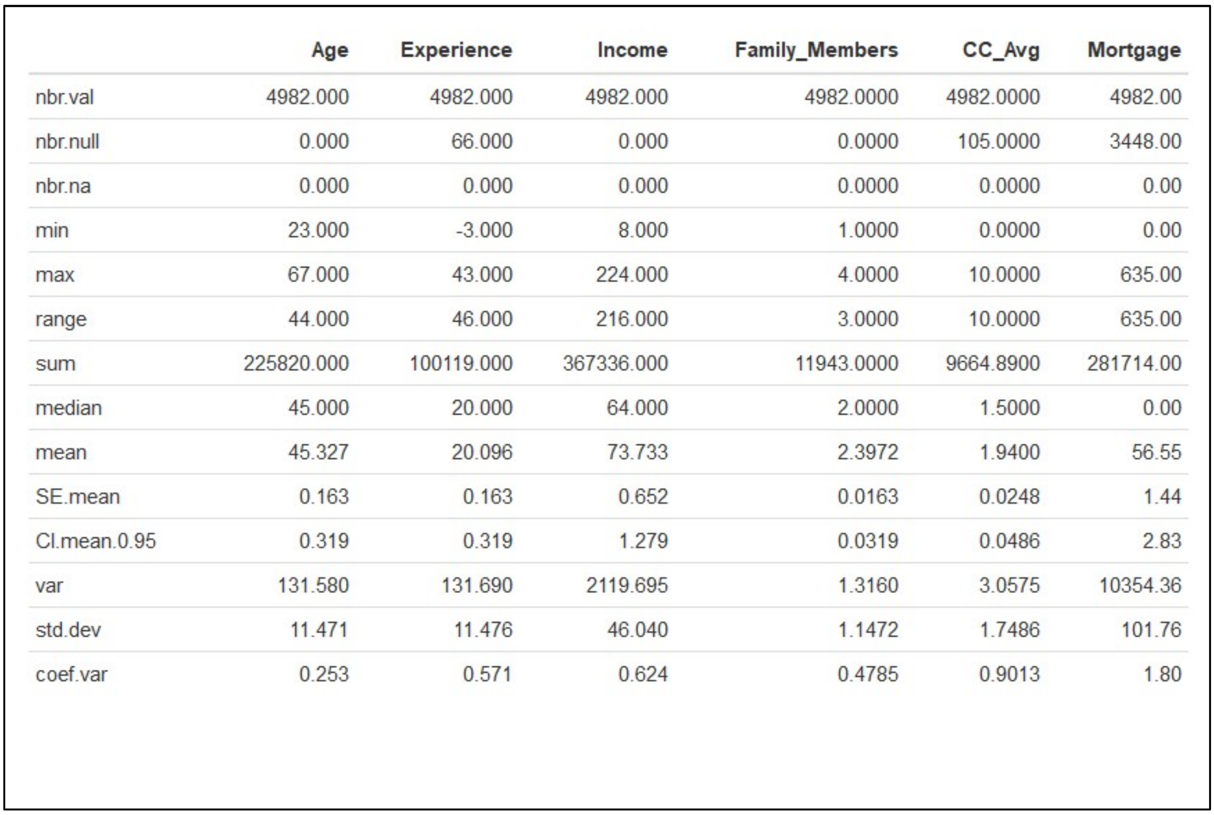


Chart1.2.2.a



	Age	Experience	Income	Family_Members	CC_Avg	Mortgage
nbr.val	4982.000	4982.000	4982.000	4982.0000	4982.0000	4982.00
nbr.null	0.000	66.000	0.000	0.0000	105.0000	3448.00
nbr.na	0.000	0.000	0.000	0.0000	0.0000	0.00
min	23.000	-3.000	8.000	1.0000	0.0000	0.00
max	67.000	43.000	224.000	4.0000	10.0000	635.00
range	44.000	46.000	216.000	3.0000	10.0000	635.00
sum	225820.000	100119.000	367336.000	11943.0000	9664.8900	281714.00
median	45.000	20.000	64.000	2.0000	1.5000	0.00
mean	45.327	20.096	73.733	2.3972	1.9400	56.55
SE.mean	0.163	0.163	0.652	0.0163	0.0248	1.44
CI.mean.0.95	0.319	0.319	1.279	0.0319	0.0486	2.83
var	131.580	131.690	2119.695	1.3160	3.0575	10354.36
std.dev	11.471	11.476	46.040	1.1472	1.7486	101.76
coef.var	0.253	0.571	0.624	0.4785	0.9013	1.80

Table1.2.2.b

1.2.2.1) Age and Experience
Both Age and Experience show very similar distributions as indicated by the histogram and density plots (*Chart1.2.2.1a)*. The two variables are closely related hence the similarity in distribution. The distributions are symmetric also shown by the boxplot with almost equal length whiskers, though the distributions do not resemble a normal distribution. They also do not have outliers in their distribution. The youngest customer is 23 years old and the oldest is 67 years old. The median and average age is about 45 years. Most of the clients lie between the ages of 25 and 65 years which is the usual age range where some starts and ends their career. In this age range most that have started their career are earning income and this means there is a greater potential for savings.

The minimum experience is -3 years and the maximum is 43 years. The negative experience is assumed to mean the number of years left for a customer's studies before they start their career. Both the median and mean values are similar just like the Age variable. The distribution of Age and Education deviate from the mean with similar values of 11.471 and 11.476 respectively as shown by their standard deviations.

1.2.2.2) Income
Distribution of income is right skewed and there are outliers on right whisker of the box plot which is longer than the left whisker contributing to the skew. The minimum income is $8 000 and the maximum of $224 000 with a fairly large difference between the lowest and highest earning customers with a difference of $216 000 between their incomes. The average income is about $73 733 and the median income is less than the mean at about $64 000.

1.2.2.3) Family Size
Most of the families have single members and the followed by those with two members. The average and median family size is similar at 2.4 and 2 respectively. The minimum and maximum family sizes are 1 and 4. From the boxplot, there is an indication that the family size distribution is not symmetric with a right skew present.

1.2.2.4) Credit Card Average
There are many outliers on the average monthly credit card spending resulting in a right skew. The maximum amount spent, and the range are the same at $10 000 because there are customers without credit cards and do not have monthly credit card spending. The average monthly spending is $1 940, and the median is $1 500 and standard deviation is $1 749.

1.2.2.5) Mortgage
The distribution of value of mortgages is right skewed with many outliers on the right of the distribution. The minimum mortgage value is 0 which could either mean a customer has not taken a mortgage yet or they have fully paid their mortgage. Like the credit card monthly average spending, the maximum value and the range of the mortgage are similar at $635 000. The average mortgage value is about $56 550 and the median is $0. This median means that most of the customers do not have a mortgage value.

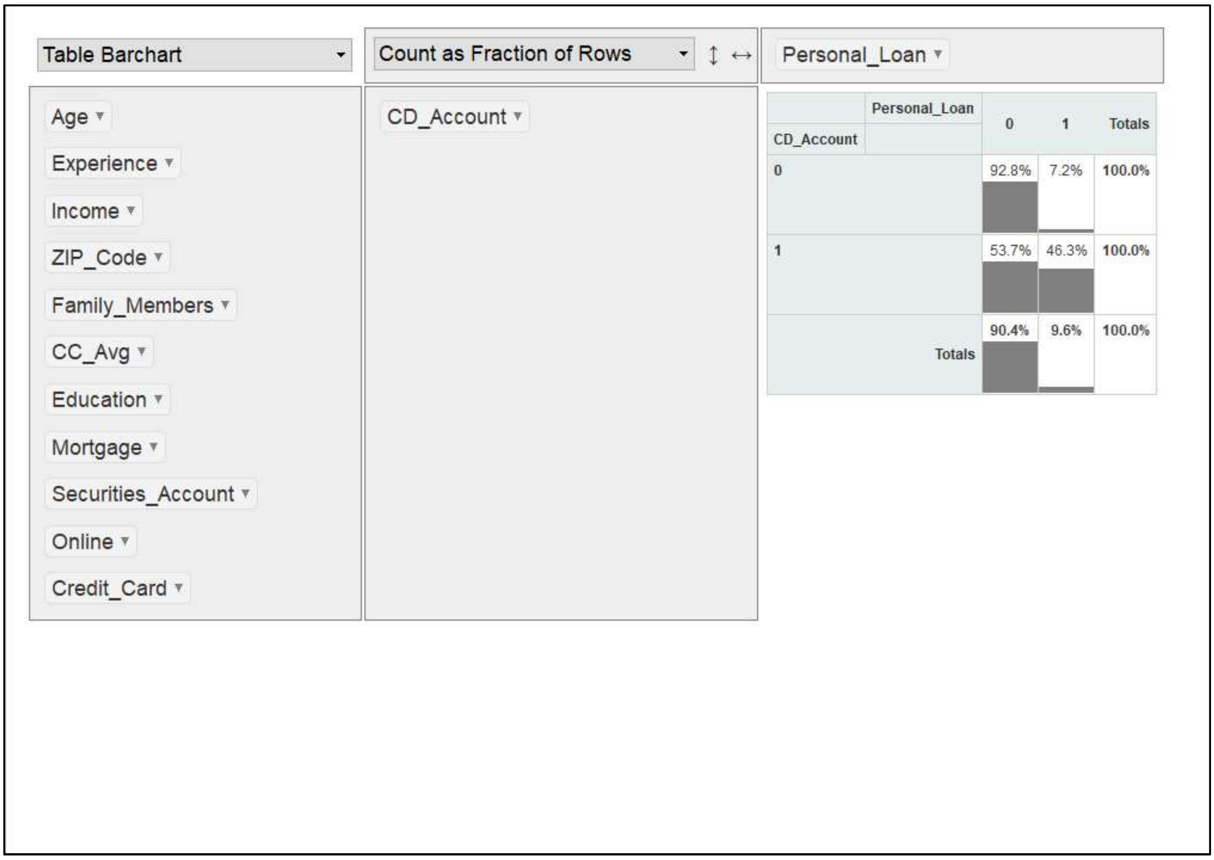
1.3) Bivariate Analysis

The bivariate analysis is based on clients that accepted the personal loan offer and other features that could help explain why they accepted the offer.

1.3.1) Bivariate Analysis – Categorical Variables

1.3.1.1) Personal Loan and CD Account
One interesting relationship is between customers that accepted a personal loan in the last campaign is with CD Accounts. A total of 46.3% of clients that had a CD Account accepted the loan offer compared to only 7.2% of clients that accepted the offer but did not have a CD

Account (*Chart1.3.1.1.a)*. This may indicate a customer is more likely to accept a personal loan if they have a CD account and this can be looked at further.

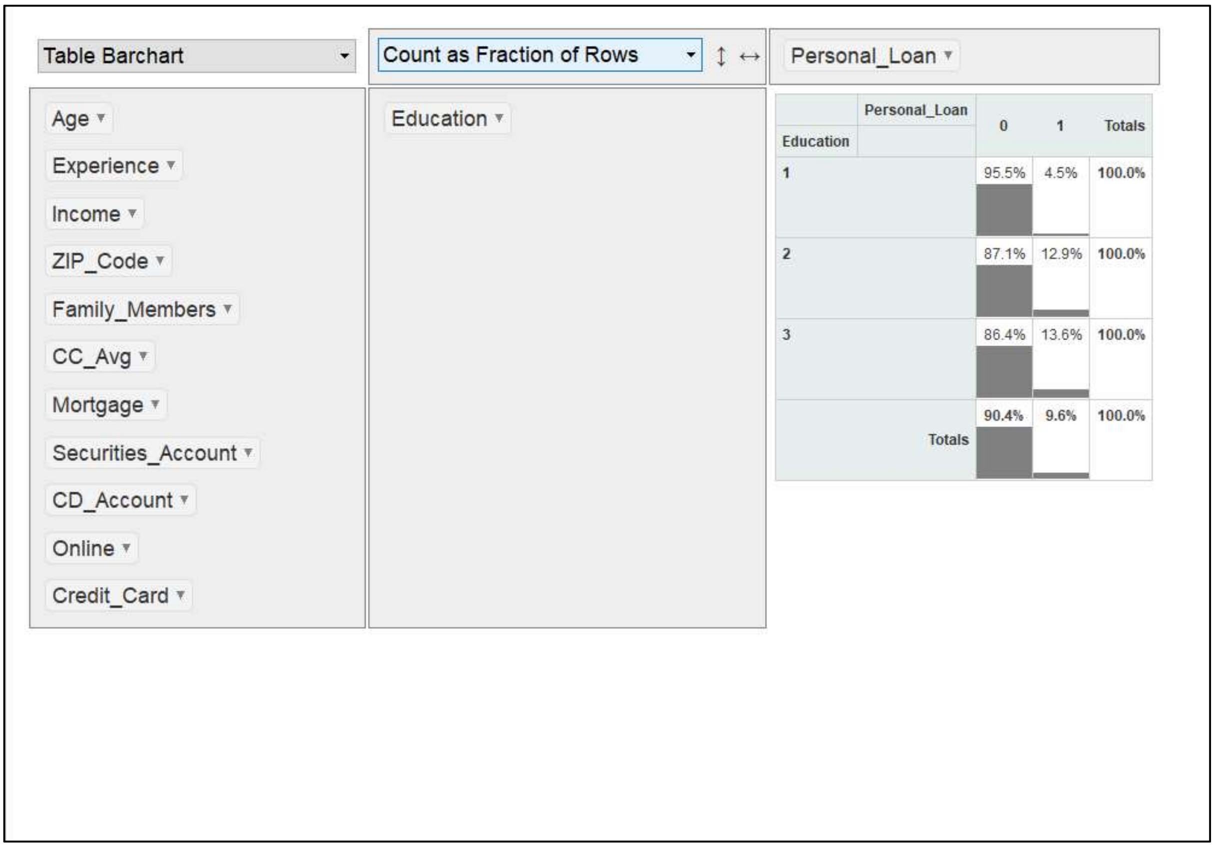


	Personal_Loan		
CD_Account	0	1	Totals
0	92.8%	7.2%	100.0%
1	53.7%	46.3%	100.0%
Totals	90.4%	9.6%	100.0%

Chart1.3.1.1.a

1.3.1.2) Personal Loan and Education

Customers that accepted the personal loan offer were more like to have either a graduate or advanced/professional level of education. Of those that have an undergraduate level of education only 4.5% accepted the loan offer compared to 12.9% and 13.6% that have a graduate or advanced/professional level of education respectively.



Education	Personal_Loan 0	1	Totals
1	95.5%	4.5%	100.0%
2	87.1%	12.9%	100.0%
3	86.4%	13.6%	100.0%
Totals	90.4%	9.6%	100.0%

Chart1.3.1.2.a

1.3.2) Bivariate Analysis – Numeric Variables

There is very high positive correlation between Age and Experience because experience is related to age. There is also a moderate positive correlation between the average monthly credit card spending and income. Usually, the size of your credit card limit depends on your income. Mortgage and income are also positively correlated as well as mortgage and average monthly credit card spending. The remaining correlations are negative and small.

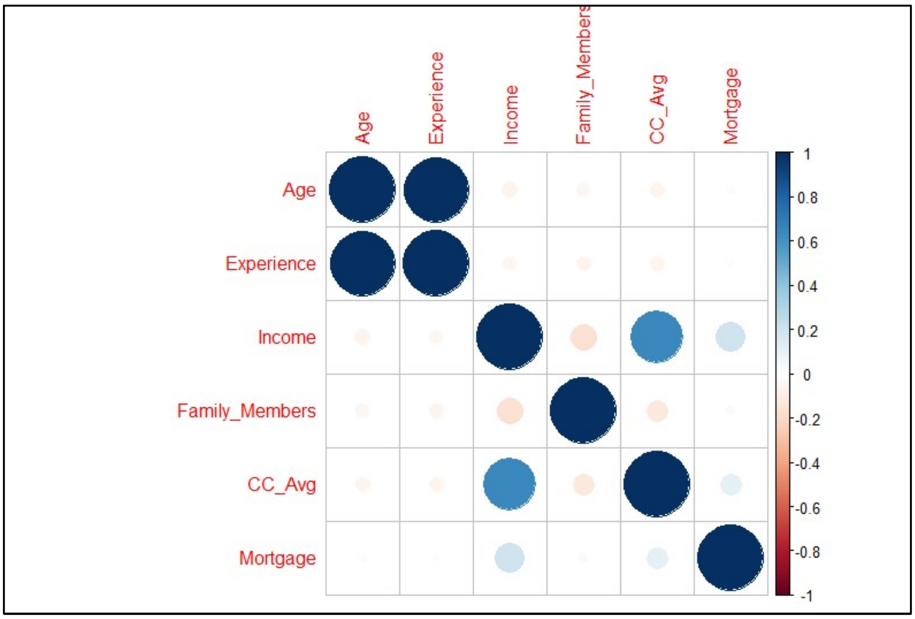


Chart1.3. 2.a

2.) Clustering

2.1) Clustering Algorithm and Rational

We are going to apply an unsupervised learning approach because we want to discover the type of liability customers Thera Bank has without making giving them prior labels. There is no pre-set business need stated by Thera Bank for the total number of customer segments that we must use. There is need to discover the clusters that would be applicable though. This means hierarchical clustering is appropriate because we do not need to determine the number of clusters upfront.

Hierarchical clustering gives the flexibility of producing many clusters and from this one can decide how to cluster the data. K-Means clustering does not have this flexibility as it requires a predetermined number of clusters to be given. Using hierarchical clustering in this case gives a lot more freedom in coming up with different customer segments. There is a cost associated with hierarchical clustering as it needs to calculate n(n+1)/2 distances. K-means clustering is linearly scaled requiring the number of chosen centroids multiplied by the number of points to get the number of distances that are needed. This is because hierarchical clustering uses a connectivity based algorithm which calculates the distance of each point

2.2) Hierarchical Clustering Algorithm Process, Output and Interpretation

The variables are scaled so that certain values do not dominate when distances between observations are calculated. The values are scaled to have a mean of 0 and a standard deviation of 1. We find the distance matrix which has distances between all points which allows us to find points that are near to each other. The Euclidean Method is used to calculate these distances. The average distance between clusters is then used to determine which clusters to combine.

The number of clusters is based on where we have the maximum vertical distance between clusters when you look across the graph. By inspection, 6 clusters seem to be appropriate (*Chart2.1 a*). The other drawback is that it is difficult to label the clusters from viewing the graph when there are so many points.

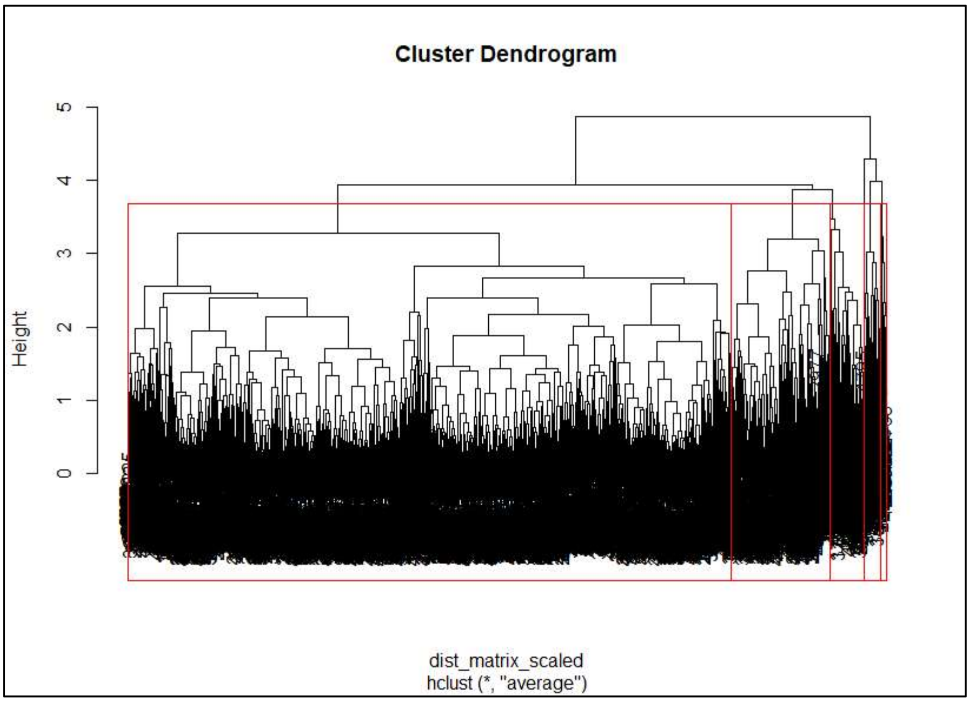
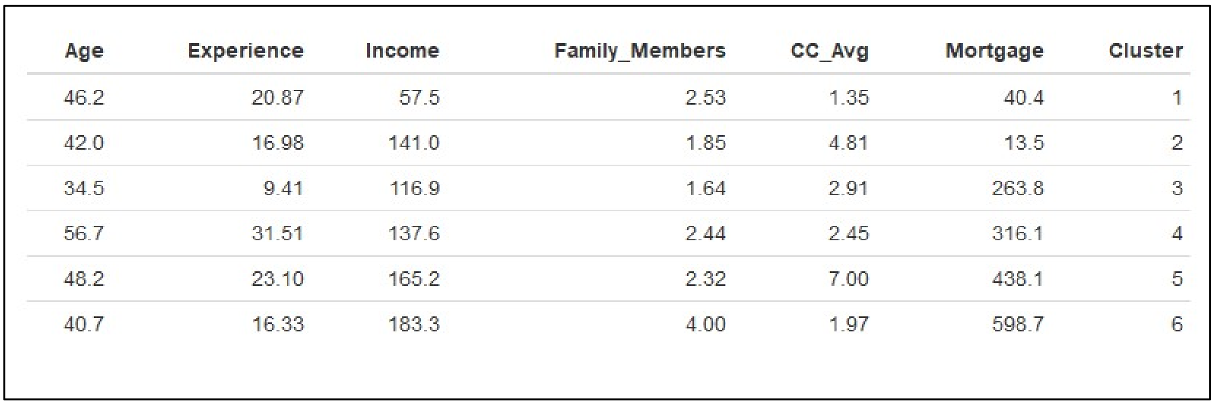


Chart2.2a

We then find the profile of the average customers in each group by assigning a cluster to each observation and finding the average for each variable in that cluster. Only numeric variables have been used to be able to calculate the distances and find the average.

From the table below (*Table2.2b)*, customers in cluster 6 are about 40.7 years old and have on average 16.3 years of experience. They have the highest average income of about $183 300 and the highest number of family members. They also have the highest average mortgage and their credit card average spending is the second lowest. Though customers in cluster 2 have an average age and experience close to customers in cluster 6 at 42 and about 17 years of experience, their income and average mortgage amounts are lower. They also have a lower number of family members at about 2 and their credit card spending is higher. Customers in cluster 3 are the youngest and have the least experience and interestingly their income is not the lowest. This indicates age and experience alone does not determine income. Adding the categorical variables will give a full picture of who these customers and potentially add to the reasons why they are in these segments.



Age	Experience	Income	Family_Members	CC_Avg	Mortgage	Cluster
46.2	20.87	57.5	2.53	1.35	40.4	1
42.0	16.98	141.0	1.85	4.81	13.5	2
34.5	9.41	116.9	1.64	2.91	263.8	3
56.7	31.51	137.6	2.44	2.45	316.1	4
48.2	23.10	165.2	2.32	7.00	438.1	5
40.7	16.33	183.3	4.00	1.97	598.7	6

Table2.2b

3.) Classification and Regression Trees (CART)

3.1) CART Algorithm and Rational

The CART algorithm is used to classify or predict using binary decision starting from a root node where similar observations are in the same partitions. The root node is split into two based on the variable that provides the most improvement in the gini impurity measure for example. The gini impurity is how often a randomly chosen observation is labelled incorrectly.

3.2) CART Decision Tree

I have removed the Zip Codes because the have many categories which results in a CART Tree that is difficult to derive conclusions from. The tree (*Chart3.2a)* has a minimum bucket size of 10 which is minimum number of customers that can be in any final bucket. The root node of the tree shows that most of the customers in the dataset did not accept the loan offer as indicated by the 0. It also shows that about 10% of the customers accepted the loan offer. The 100% means that the entire dataset is being considered at this point. The tree has a total of 8 splits with income as the variable that provides the most improvement in the gini impurity. The complexity parameter (explained below) is set to 0 which results in a complex tree that will overfitted the data.

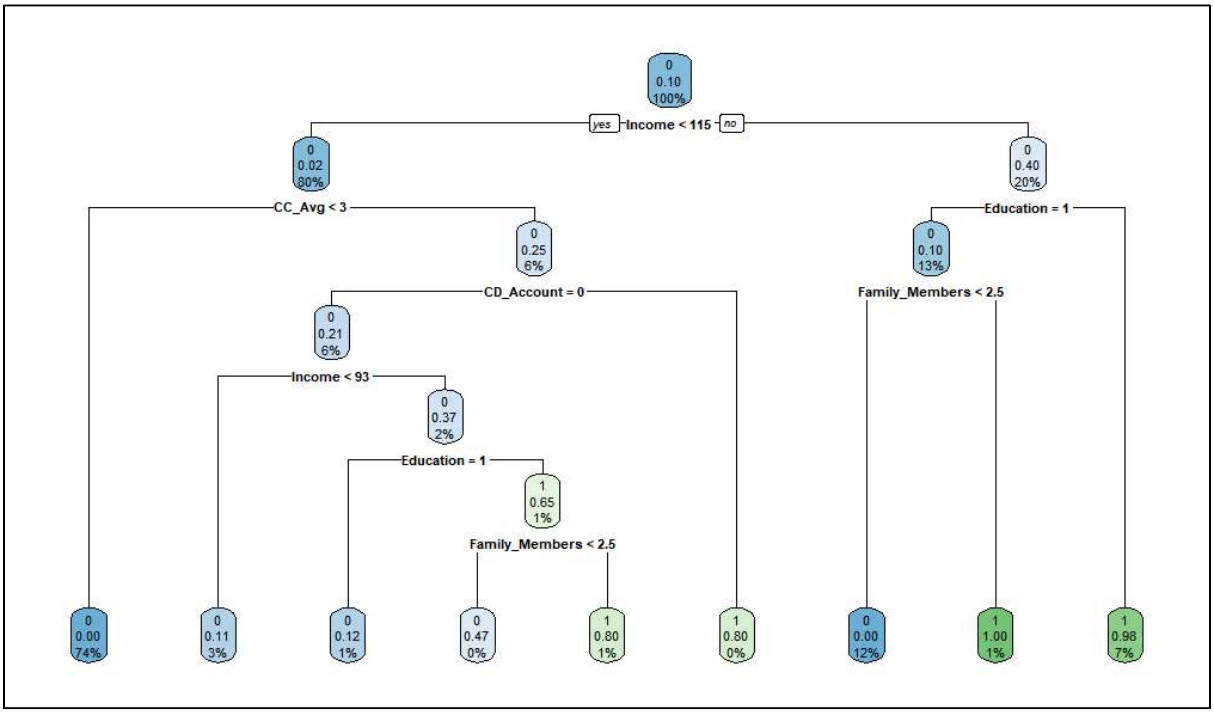
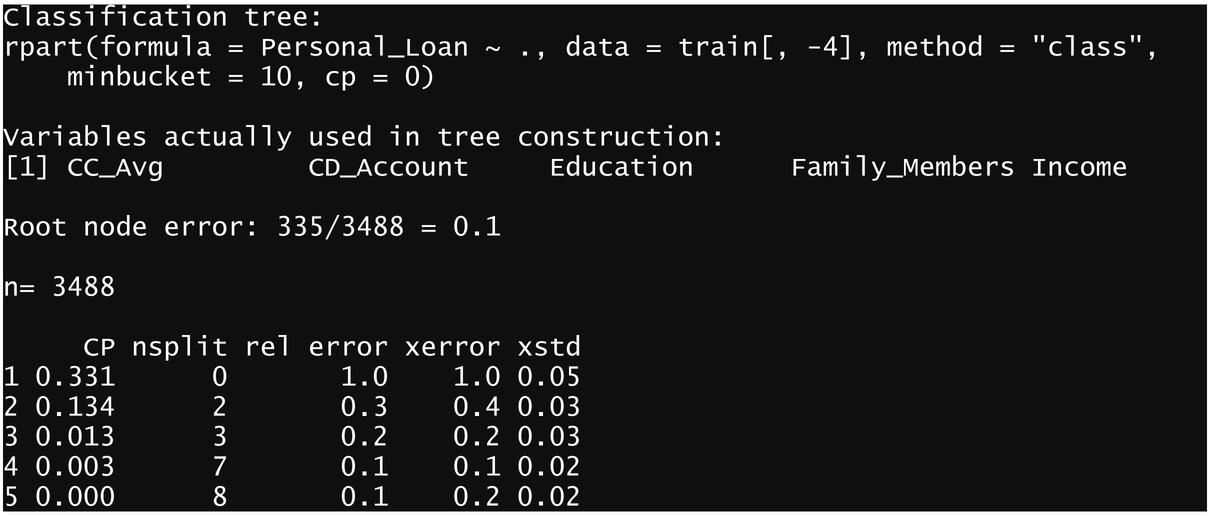


Chart3.2a

3.3) CART - Pruning

Pruning is used to reduce overfitting data by using trees with a lot of nodes which lowers their predictive ability. One approach is to use the cost complexity parameter to reduce overfitting. The cost complexity parameter requires each split to decrease the relative error by at least alpha. The cross-validation error represents the average error for each complexity of tree. The tree with the lowest cross validation error is used to determine the level of alpha that will be applied in the final algorithm to obtain the pruned tree.



```
Classification tree:
rpart(formula = Personal_Loan ~ ., data = train[, -4], method = "class",
    minbucket = 10, cp = 0)

Variables actually used in tree construction:
[1] CC_Avg         CD_Account     Education      Family_Members Income

Root node error: 335/3488 = 0.1

n= 3488

     CP nsplit rel error xerror xstd
1 0.331      0       1.0    1.0 0.05
2 0.134      2       0.3    0.4 0.03
3 0.013      3       0.2    0.2 0.03
4 0.003      7       0.1    0.1 0.02
5 0.000      8       0.1    0.2 0.02
```

Fig3.3a

From the output above (*Fig3.3a*), the cross-validation error is at its lowest at 0.1. This means we can select a cost complexity parameter great than 0.003 but less than 0.013. Using as complexity parameter of 0.1 will achieve the result of pruning the tree and reducing overfitting. This is also confirmed by the chart below, which shows the size of the tree between 4 and 7 based on out cost complexity parameter.

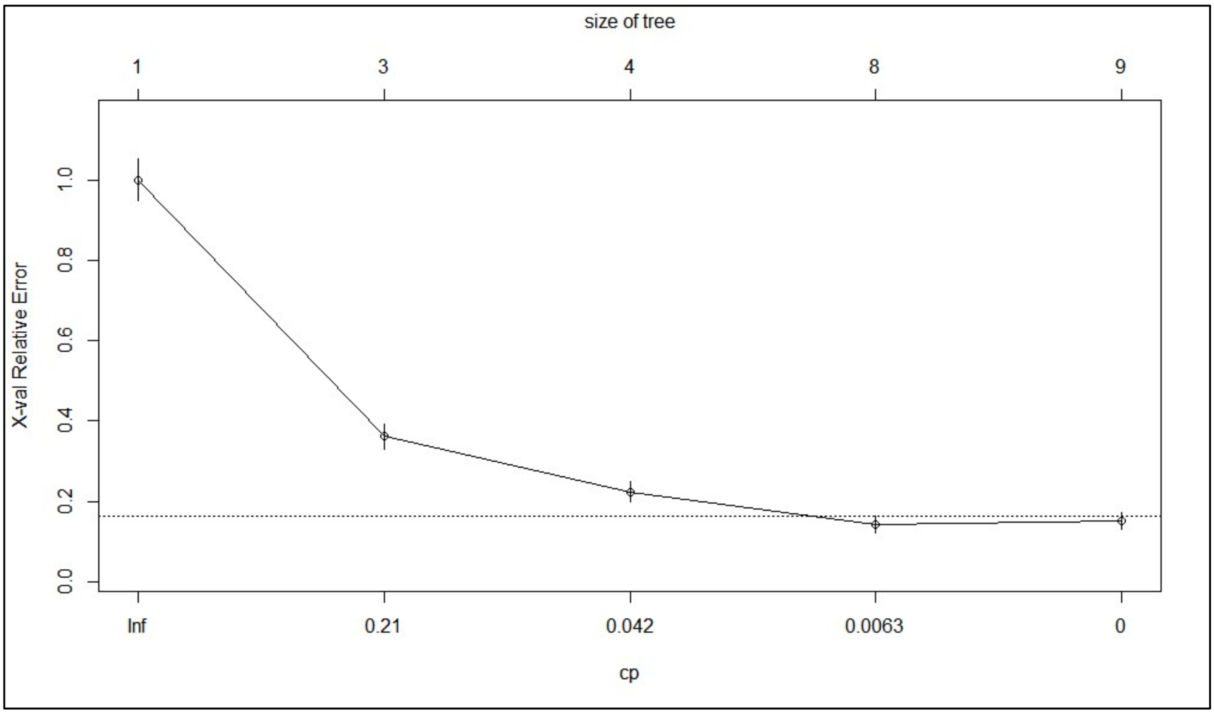


Fig3.3b

3.4) CART – Pruned Tree

The output of the pruned tree is below (*Chart3.4a)* with three splits. From this tree, income above $115 000 is the most important variable to look at when identifying the customers that are likely to accept the loan offer. The clients with an undergraduate level of education followed by households with an average family size above 2.5. The other group of customers likely to accept the loan offer has an income above $115 000 and undergraduate level of education.

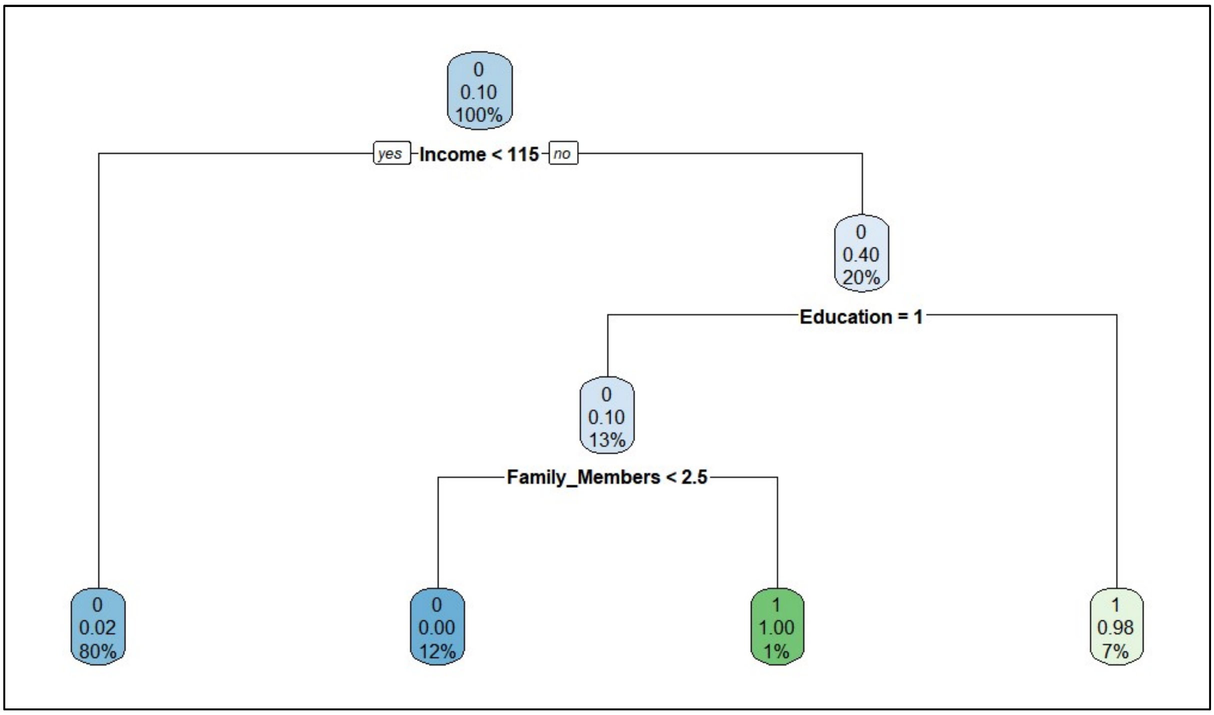


Chart3.4a

3.5) CART – Pruned Tree Model on Test Data

The output based on the model of the pruned tree is applied to the test dataset is below (*Chart3.5a)*. This tree still all the three variables i.e. income, education and family size with education and family size remaining the same while income is lower at $109 000. There are still two terminal nodes where customers are likely to accept the loan offer. The change is now that the test model indicates a slightly higher error at the education node of 42% compared to 40% in the training dataset. This means there more customers that could be considered for the loan offer. This the same with the family member node, which is now 14% vs 10% in the training dataset.

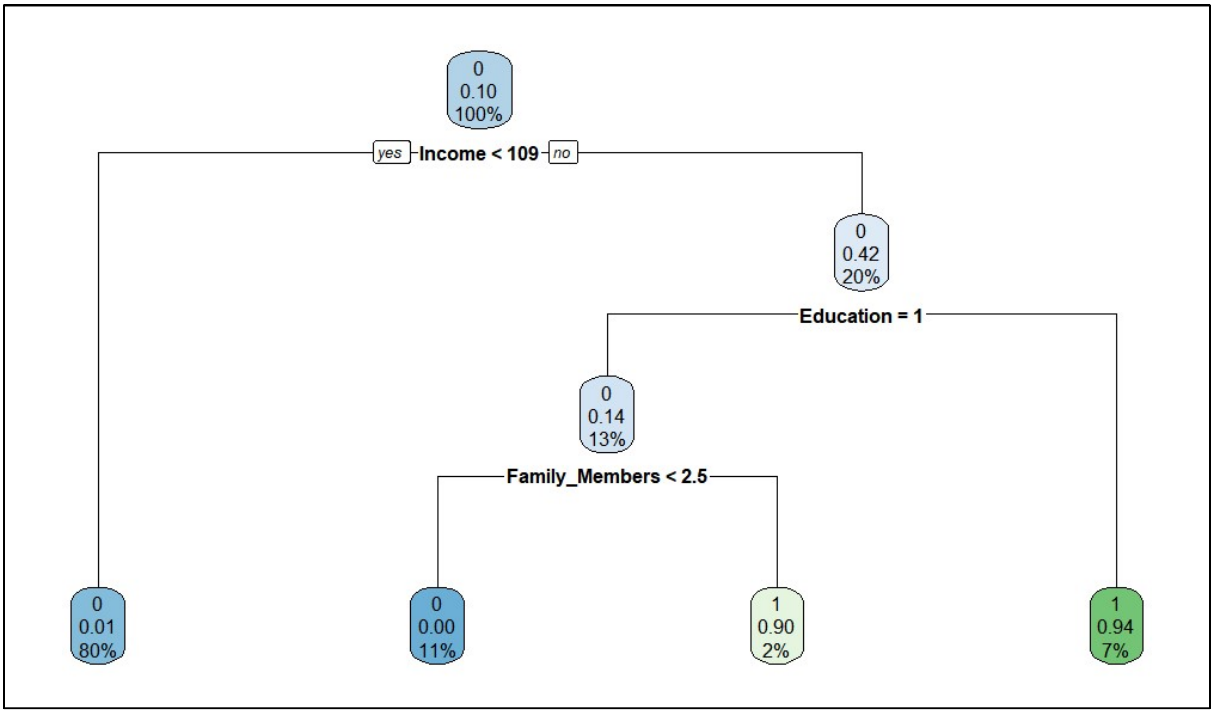


Chart3.5a

4.) Random Forests

4.1) Random Forest Algorithm and Rational

Random Forests are a more stable approach compared to decision trees because decision trees are sensitive to any slight changes in data. Using more trees instead one would make predictions more stable. To make sure we get the most value from these multiple trees, we apply bootstrapping which ensures that each tree is different from the next. This is done by using different samples of the data and using those samples to train data to get the trees. The samples are based on a random number of observations where observations can be repeated. A sample of the variable less than the total number of variables is also used. If the number of sample variables used is too high, then the trees become correlated. When the number of variables is small, then we are likely to miss the important variables and make poor predictions. For random forests, pruning is not required because of the use of different trees which only overfits on a portion of the data and not the entire data. From these trees, we will predict using modes since we are dealing with classification of customers likely to accept the personal loan offer.

4.2) Model Tuning

4.2.1) Number of trees

After generating an initial model, we can improve the model by tuning the different parameters. We start by determining an optimal number of trees where our OOB error rates no longer change with the increase in the number of trees. Out of Bag (OOB) error rates are comparisons of predictions and actuals for all observations where the predictions are made using trees that we did not use for certain observations. From the graph, the OOB error rates are constant where the number of trees is above 170 for example. The OOB error is 1.35% (*Fig4.2.1b).* We will then use an odd number of trees that is 171 to prevent ties since we require a majority result for predictions. The number of variables that are used to make each tree in the random forest is 6.

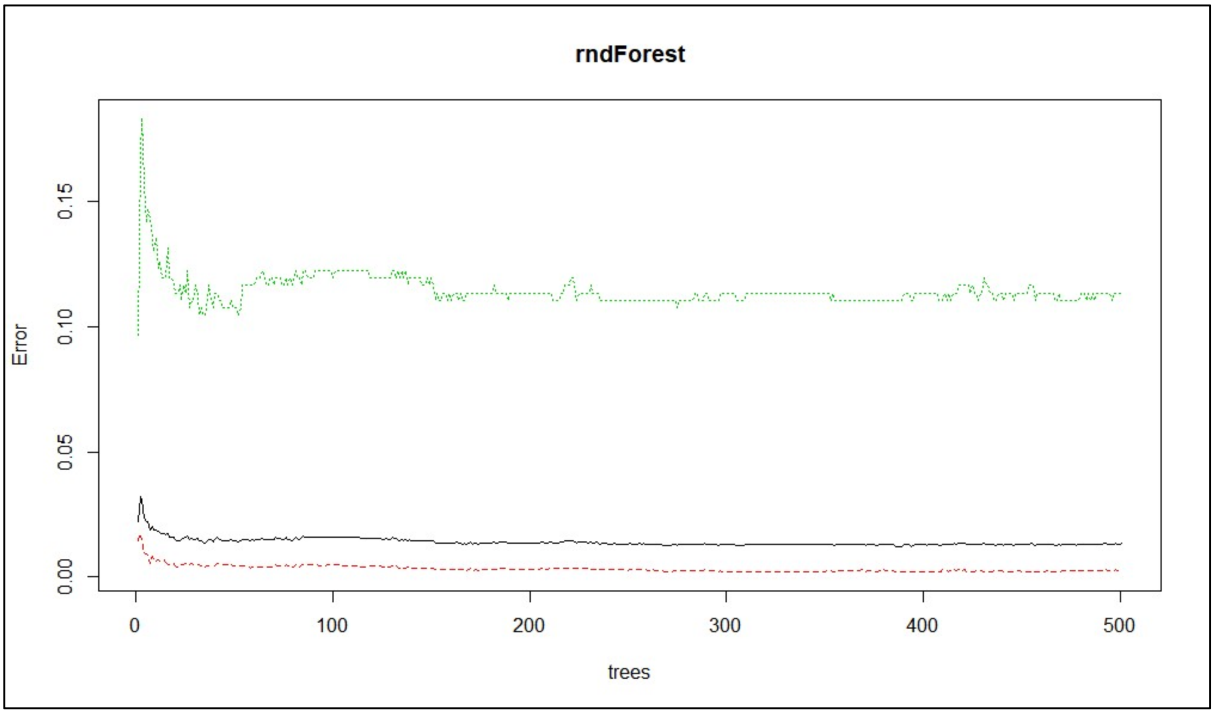
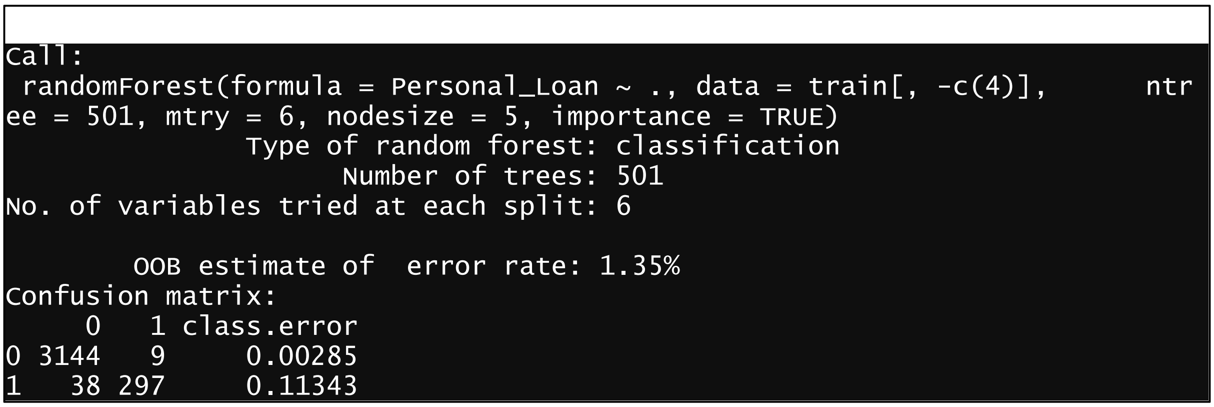


Chart4.2.1a



```
Call:
 randomForest(formula = Personal_Loan ~ ., data = train[, -c(4)],      ntr
ee = 501, mtry = 6, nodesize = 5, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 501
No. of variables tried at each split: 6

        OOB estimate of  error rate: 1.35%
Confusion matrix:
     0   1 class.error
0 3144   9     0.00285
1   38 297     0.11343
```

Fig4.2.1b

4.2.2) Number of variables to use

We can also tune our model to determine the optimal number of variables to use for the trees in the random forest model. We can use the graph output where the lowest OOB error rate occurs or identify it in out model output. These show that we need to use 7 variables at a time where the OOB error rate is 1.32% (*Fig4.2.2a*).

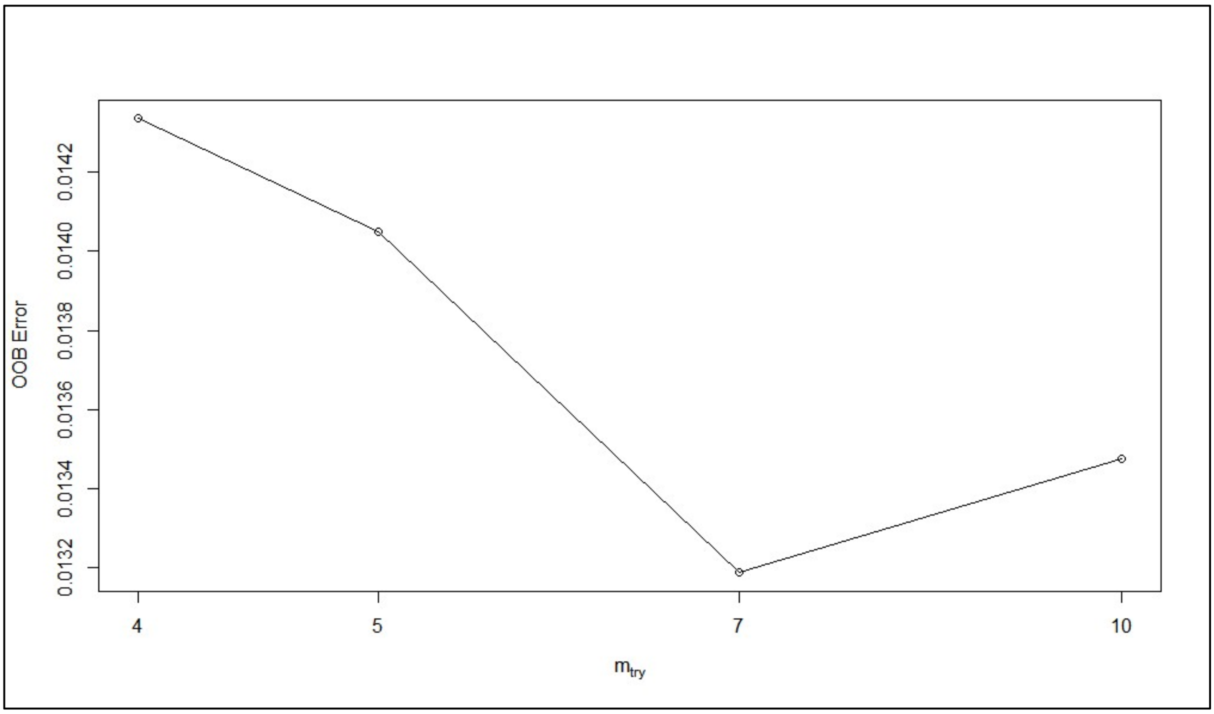
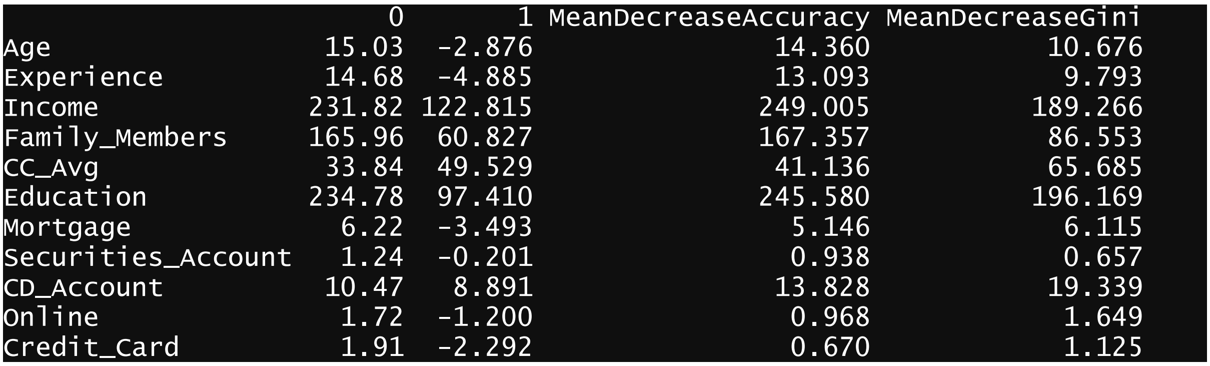


Fig4.2.1b

4.3) Random forest output interpretation

4.3.1) Variable importance

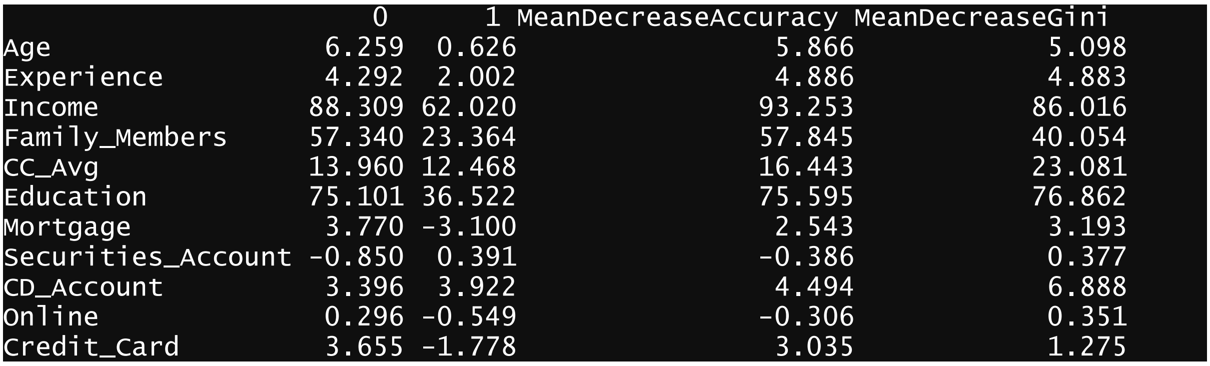
A variable's importance can be identified by the Mean Decreasing Accuracy or Mean Decrease Gini. The Mean Decreasing Accuracy shows how shuffling a variable and creating a random forest reduces the predictive power of that random forest. The larger the number, the more important the variable. The Mean Decrease Gini shows how much the Gini decreased on average across all trees where a tree was split based on a variable. The larger it is the more important the variable. Both measures indicate that Income is the most important variable in predicting whether a customer is likely to accept a loan offer (*Fig4.3.1a*). Education and the Family Members are the second and third most important variables based on both measures. Having a Securities Account has the least importance in determining whether a customer will accept a personal loan offer.



```
                       0       1 MeanDecreaseAccuracy MeanDecreaseGini
Age                15.03  -2.876               14.360           10.676
Experience         14.68  -4.885               13.093            9.793
Income            231.82 122.815              249.005          189.266
Family_Members    165.96  60.827              167.357           86.553
CC_Avg             33.84  49.529               41.136           65.685
Education         234.78  97.410              245.580          196.169
Mortgage            6.22  -3.493                5.146            6.115
Securities_Account  1.24  -0.201                0.938            0.657
CD_Account         10.47   8.891               13.828           19.339
Online              1.72  -1.200                0.968            1.649
Credit_Card         1.91  -2.292                0.670            1.125
```

Fig4.3.1a

The test dataset output(below) still show income, education and family size are the most important variables in that order.



```
                        0      1 MeanDecreaseAccuracy MeanDecreaseGini
Age                 6.259  0.626                5.866            5.098
Experience          4.292  2.002                4.886            4.883
Income             88.309 62.020               93.253           86.016
Family_Members     57.340 23.364               57.845           40.054
CC_Avg             13.960 12.468               16.443           23.081
Education          75.101 36.522               75.595           76.862
Mortgage            3.770 -3.100                2.543            3.193
Securities_Account -0.850  0.391               -0.386            0.377
CD_Account          3.396  3.922                4.494            6.888
Online              0.296 -0.549               -0.306            0.351
Credit_Card         3.655 -1.778                3.035            1.275
```

4.3.2) Error rate
Training dataset
The approximate error rate is now down to 0.00258 and our models seems to be able to predict all instances when someone will not accept an offer based on the training data set. We will focus on the top customers identified to accept a personal loan if they will accept the offer. We need to identify these customers and we use deciles to group them. Based on the quartile outputs, they show that the top ten probabilities of accepting a person loan offer are between 0.23772 and 1.000000. We can set a minimum probability of acceptance to decide who we send the loan offers based on their likelihood to accept it. Using a threshold of 0.23772, we will have 96% of customers that will likely accept the loan offer.

Out of sample data
For the test data set, our model's error rate is 0.00602% which is more than three times higher than the training data at. There is greater chance of incorrectly predicting customers that are likely to accept Thera bank's offer. The range of the probabilities of the top ten customers likely to accept the offer has narrowed and is from 0.290 to 1.000. based on this threshold, now only 92.6% of the customers are likely to accept the loan offer compared to 93.5% in the training data set.

Measure of model performance
Confusion Matrix
The confusion matrix gives the number of correct or incorrect predictions. It is made up of the following elements.

- TP -True Positive
- FP – False Positive
- FN – False Negative
- TN – True Negative

Classification error rate
This is number of times we are incorrectly predicting that customers would accept the loan offer when they would not and incorrectly predicting customers who would not accept a loan offer when they would. Based on our confusion matrix, the classification error rate is 1.9% for the CART training data and 1.6 percent on the test data.

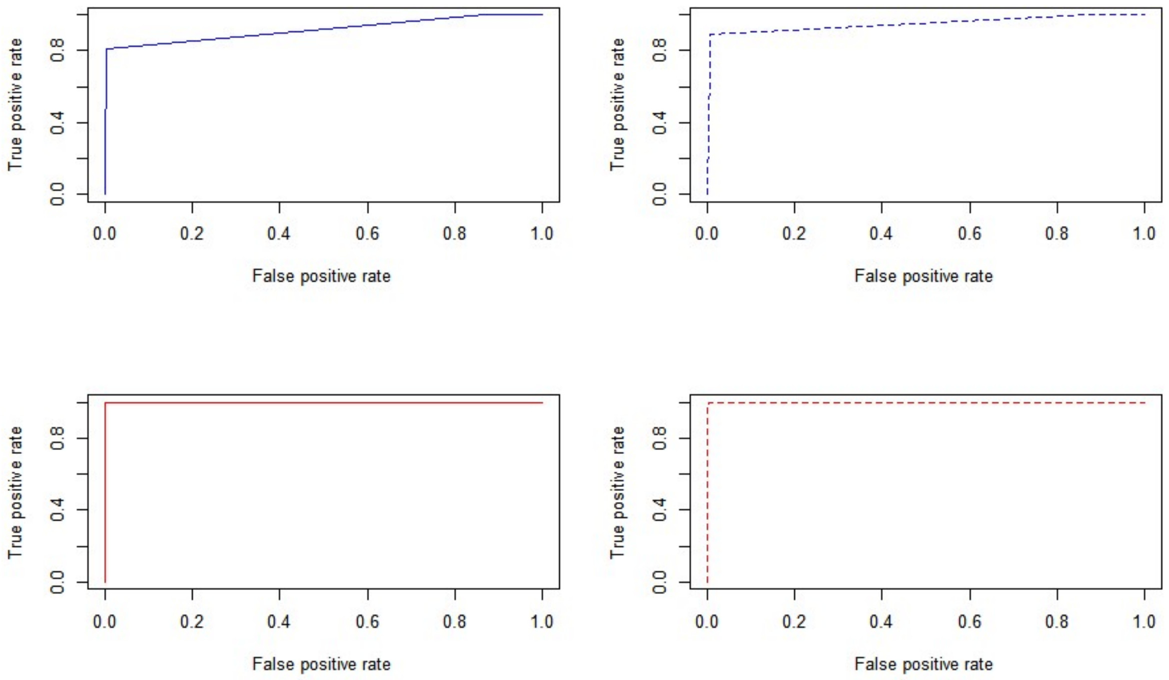
Sensitivity
This measures the fraction of customers that were predicted to accept the loan offer and actually accepted the loan (True Positives). The denominator is the total number of customers that actually meant to accept the offer including those that were misclassified when they would have accepted (True Positives + False Negatives).

Specificity

This is all the customers that didn’t not accept the loan offer, how many would have not accepted anyway. This is the TN/(TN + FP).

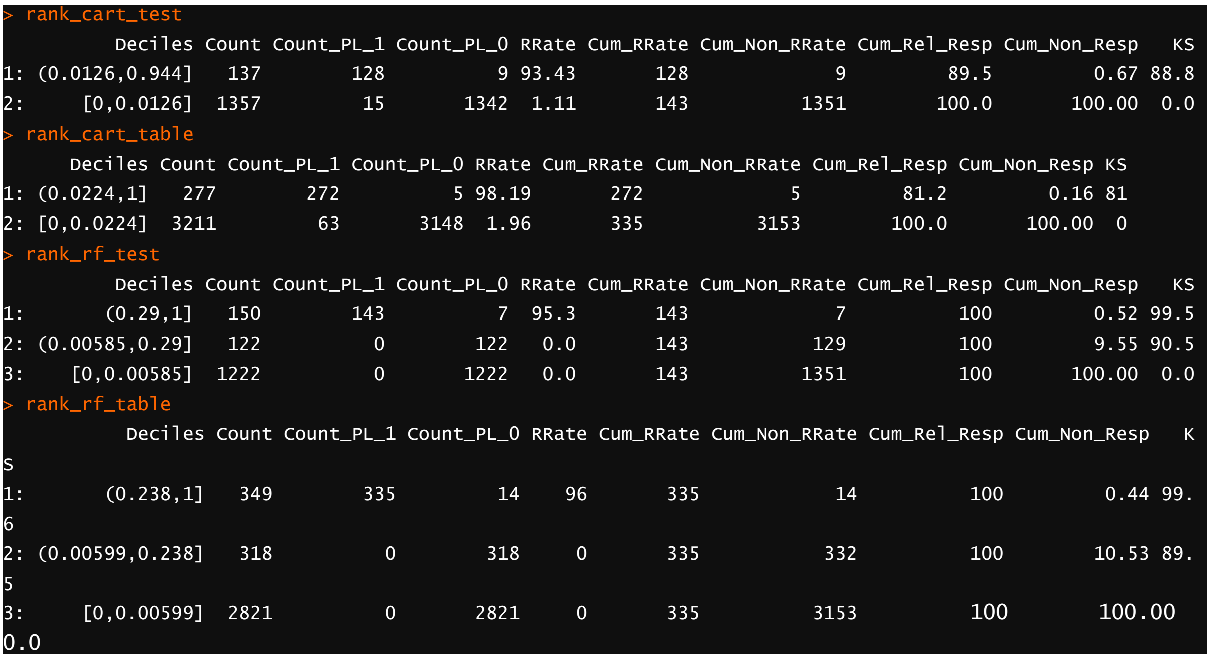
Based on the two measures, we would like to focus on Sensitivity because we want to increase acceptance on the loan offer. This means reducing the number of customers that are in the False Negative category thus increasing our sensitivity.

ROC Curve
The CART models in blue did not perform as well as the random forests in red. This is both training and test data. The random forests hard a larger area under the curve of close to one in both test and training datasets with the CART models at 0.95 and 0.92 respectively.



Gini Coefficient
The gini were 0.9 for the rndom forests for both test and training data. They were better than the CART at 0.82 and 0.75 respectively.

KS Chart Gain and Lift
If we target customers above a certain bucket how will we perform based on the KS Chart. The models all show the highest buckets contain the most instances where customers accepted loans. For example, the random forest test model shows that 99.5% of those who accepted the loan offer are in the highest deciles. This means our models are not random since all buckets are not the same. The CART model performed better using test data compared to training data. The random forest performed better when used in training data at 99.6%. Overall, the random forest performed better in both training and test data.



```
> rank_cart_test
          Deciles Count Count_PL_1 Count_PL_0 RRate Cum_RRate Cum_Non_RRate Cum_Rel_Resp Cum_Non_Resp   KS
1: (0.0126,0.944]   137        128          9 93.43       128             9         89.5         0.67 88.8
2:     [0,0.0126]  1357         15       1342  1.11       143          1351        100.0       100.00  0.0
> rank_cart_table
      Deciles Count Count_PL_1 Count_PL_0 RRate Cum_RRate Cum_Non_RRate Cum_Rel_Resp Cum_Non_Resp KS
1: (0.0224,1]   277        272          5 98.19       272             5         81.2         0.16 81
2: [0,0.0224]  3211         63       3148  1.96       335          3153        100.0       100.00  0
> rank_rf_test
          Deciles Count Count_PL_1 Count_PL_0 RRate Cum_RRate Cum_Non_RRate Cum_Rel_Resp Cum_Non_Resp   KS
1:       (0.29,1]   150        143          7  95.3       143             7          100         0.52 99.5
2: (0.00585,0.29]   122          0        122   0.0       143           129          100         9.55 90.5
3:    [0,0.00585]  1222          0       1222   0.0       143          1351          100       100.00  0.0
> rank_rf_table
           Deciles Count Count_PL_1 Count_PL_0 RRate Cum_RRate Cum_Non_RRate Cum_Rel_Resp Cum_Non_Resp   K
S
1:       (0.238,1]   349        335         14    96       335            14          100         0.44 99.
6
2: (0.00599,0.238]   318          0        318     0       335           332          100        10.53 89.
5
3:     [0,0.00599]  2821          0       2821     0       335          3153          100       100.00
0.0
```

Concordance -Discordance Ratio

This is derived from every possible pair of observations where one customer accepted the loan and the other did not accept the loan. This then looks at the associated probabilities of how likely the customer would accept or reject the loan offer. The total number of pairs where the probabilities aligned with accepting or rejecting the loan offer are then classified as concordance if they support what we predicted or discordance if they did not. This means for the CART model, the training data set shows that our model performed well 83.5% of the time and did not perform well 16.5% of the time. The random forest model performed better with 90% concordance

Conclusion

The random forest model performed better than the CART model. Thera bank would better predict the customers it wishes to target using the random forest model based on all measures of performance.