PROJECT 3: THERA BANK

By: Omar Ossama

Contents

1. EDA

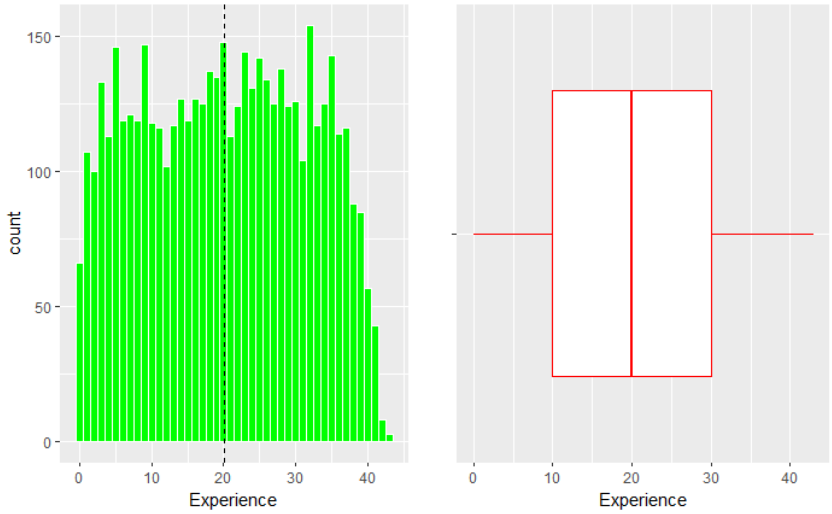
1.1. Univariate analysis:

1.1.1. Age in years:

Minimum age of the customers is 23 years.
Maximum age is 67 years.
Average age is around 45 years.
No outliers in the data.
Data slightly skewed to the right

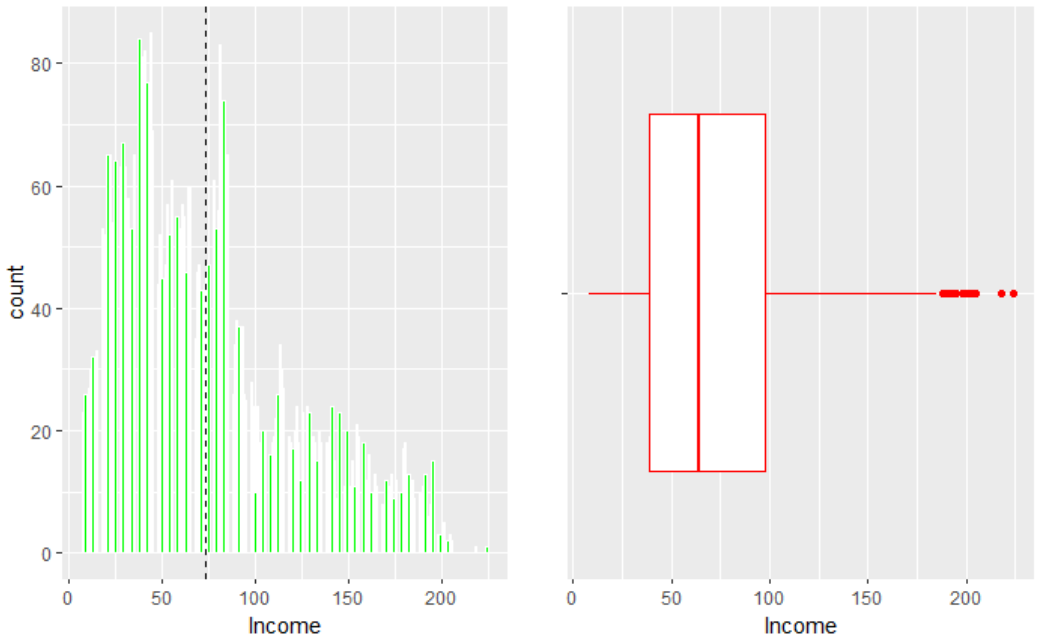
1.1.2. Customer professional experience:

Some customers have no professional experience.
Maximum years of experience are 43 years.
The average is around 20 years.
No outliers in the data.
Data slightly skewed to the right.



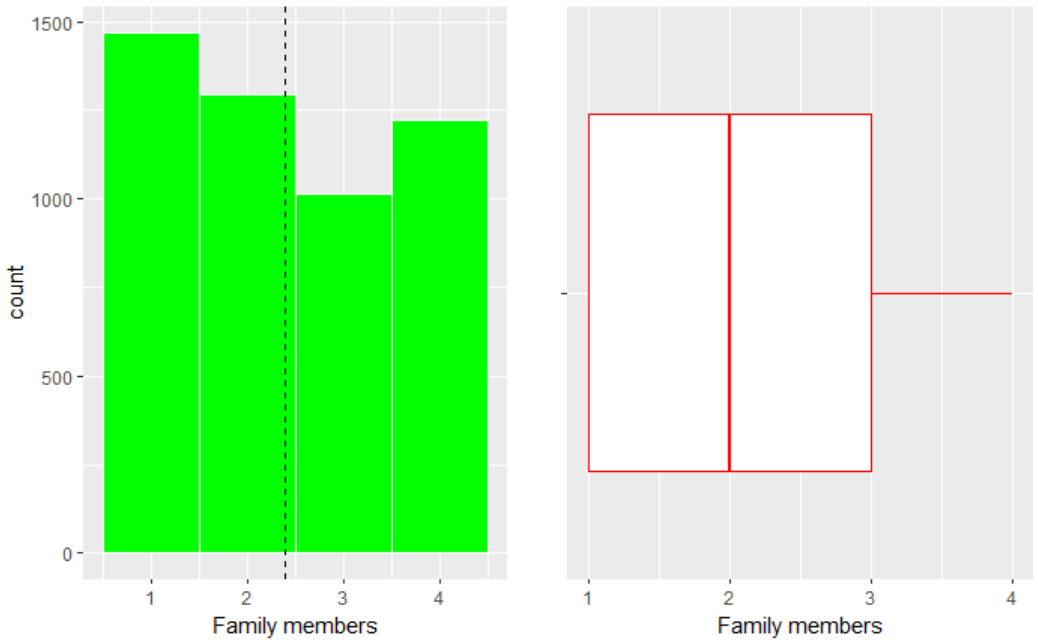
1.1.3. Annual income of the customer:

Minimum income of customers is $8,000/year
Maximum is $224,000/year
Average is $73,770/year with a median of $64,000/year
Data has outliers present above the upper limit.
Data skewed to the right.



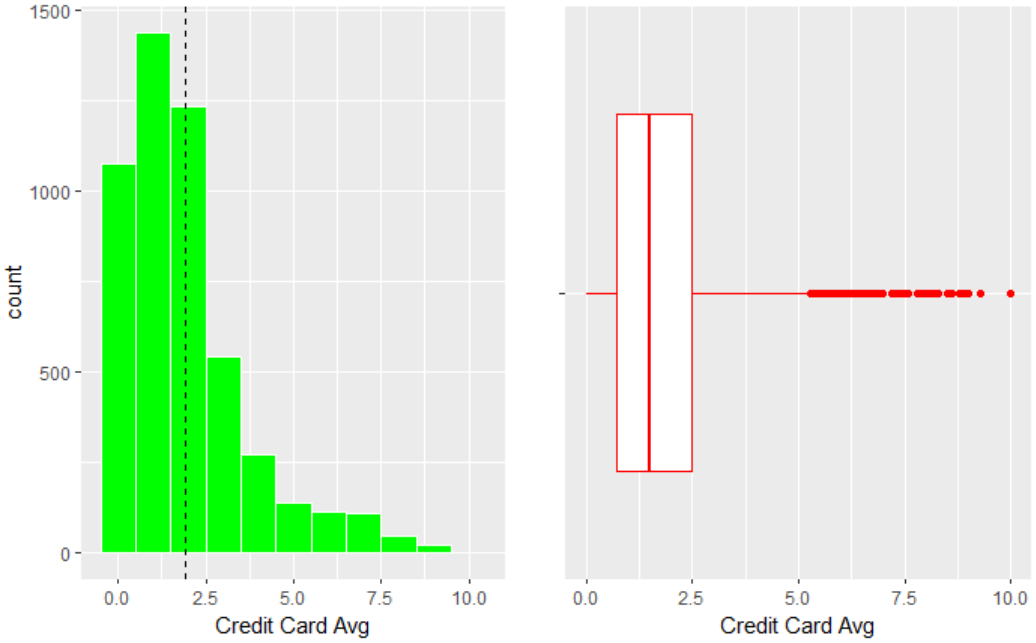
1.1.4. Family size:

Customers have a family size of between 1 and 4 members.
With an average of 2 family members.
No outliers present in data.
Data slightly skewed to the right.



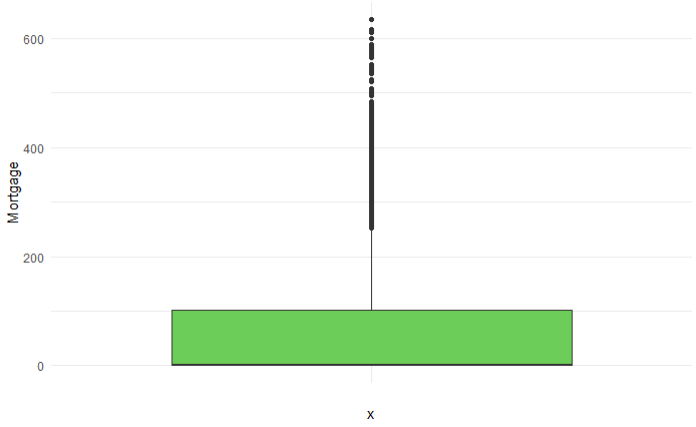
1.1.5. Average spending on credit cards/month:

A large number of customers do not spend money using their credit cards.
The maximum credit card spend/month is $10,000/month.
Average is $1,938/month with a median of $1,500/month.
Outliers present due to the fact that little to no customers spend using their credit cards.
Data skewed to the right.



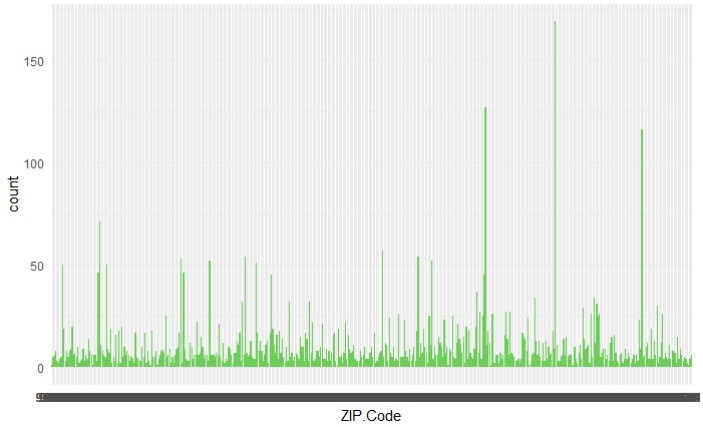
1.1.6. Value of house mortgage (if any):

Most customers have no mortgage.
Maximum mortgage is $635,000.
Average mortgage is $1,938 with a median of $0.
Many outliers exist as most customers do not have mortgage.
Data is skewed to the right.



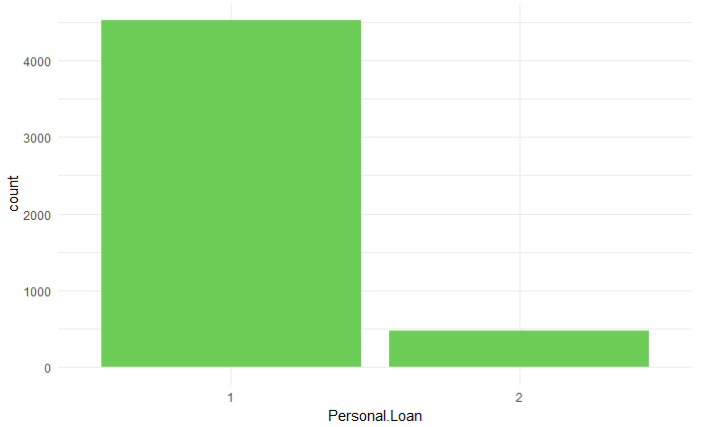
1.1.7. Zip Code:

Customers are spread along a very wide area.



1.1.8. Personal loan acceptance:

The acceptance rate of a personal loan is 90.4% did not accept to 9.6% who accepted.

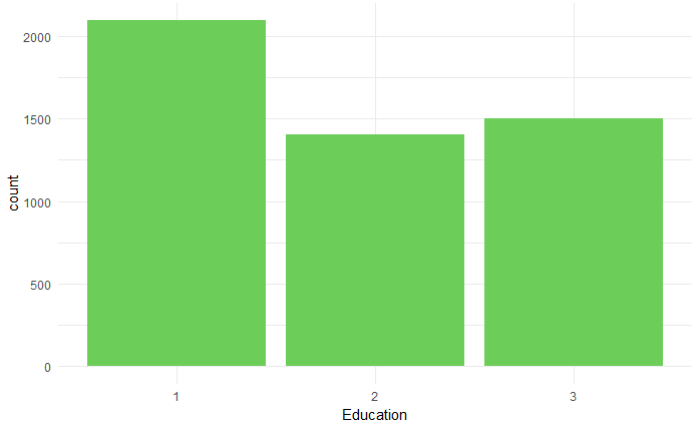


1.1.9. Education level:

41.92% of the customers are undergraduates.

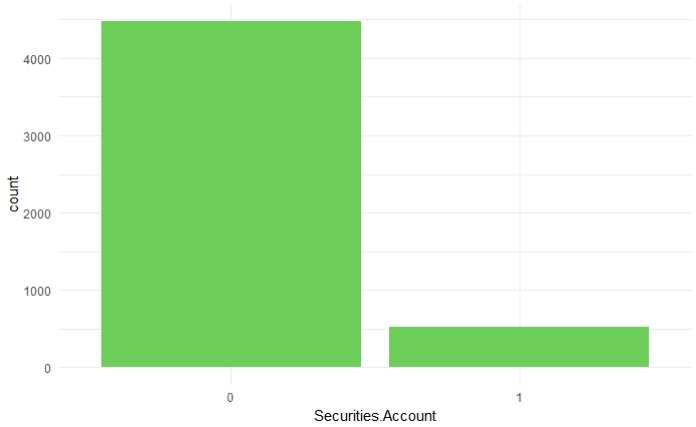
28.06% are graduates.

30.02% are advanced/professionals.



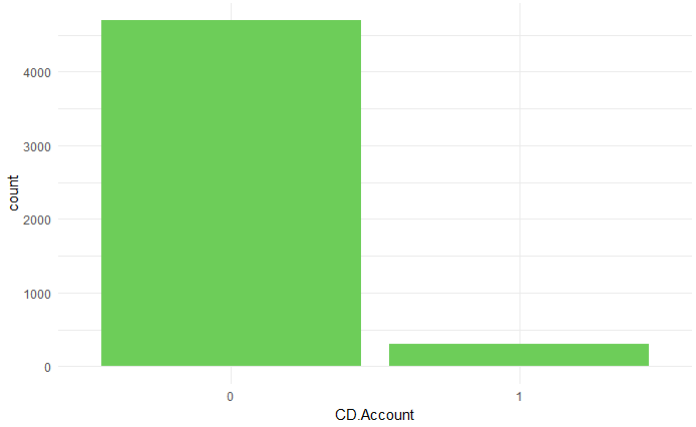
1.1.10. Securities account ownership:

89.56% of the customers do not have a securities account while 10.44% have one.



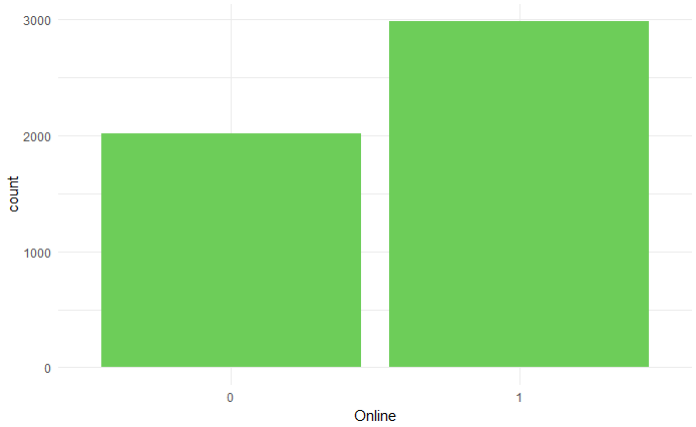
1.1.11. Certificate of deposit account ownership:

93.96% of the customers do not have a certificate of deposit account, while 6.04% do.



1.1.12. Use of online banking:

59.68% of the customers use the online banking service, while 40.32% do not.



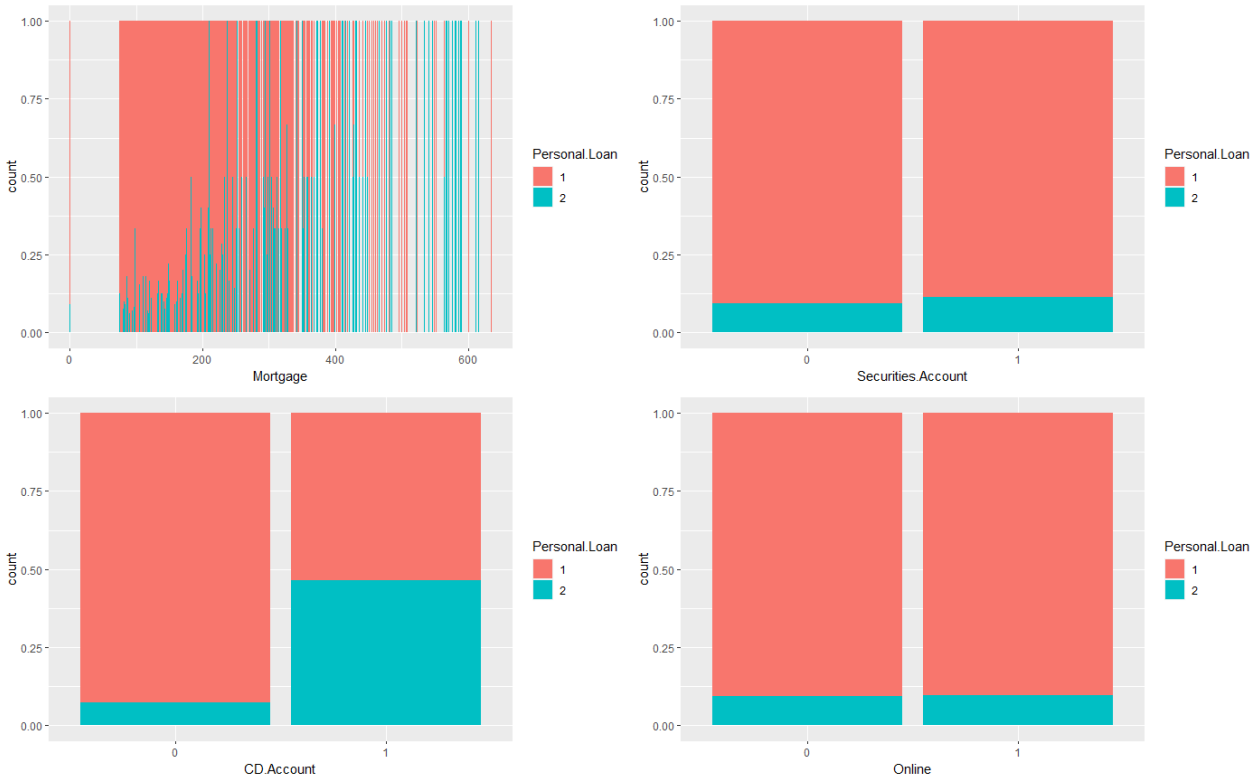
1.1.13. Credit Card ownership:

70.6% of customers do not own a credit card, while 29.4% of customers own one.

1.2. Bivariate analysis:

1.2.1. Personal loan vs the other variables:

Variables that appear to have a relation with the personal loan acceptance variable are the customers' income, mortgage value, education and the ownership of a credit deposit account.

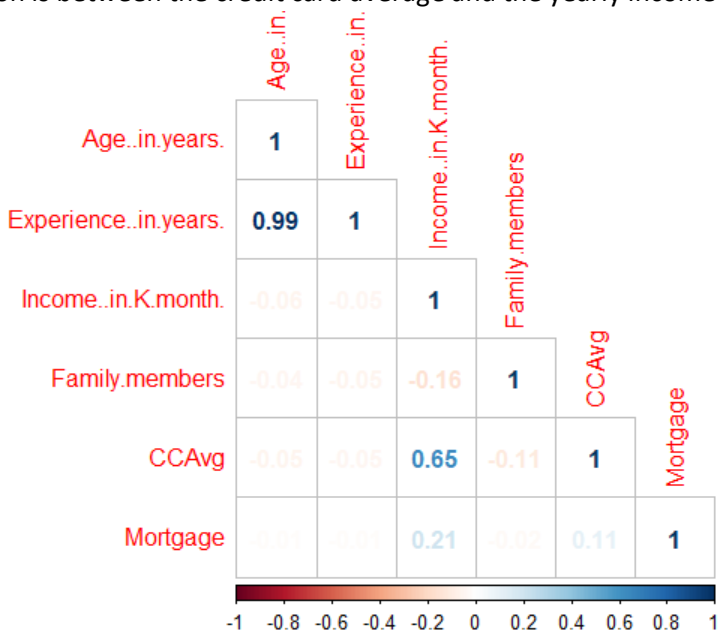


1.2.2. Correlation between the numerical variables:

Only two relations can be found.

The first is between Age of the customers and their years of experience.

The second relation is between the credit card average and the yearly income of the clients.

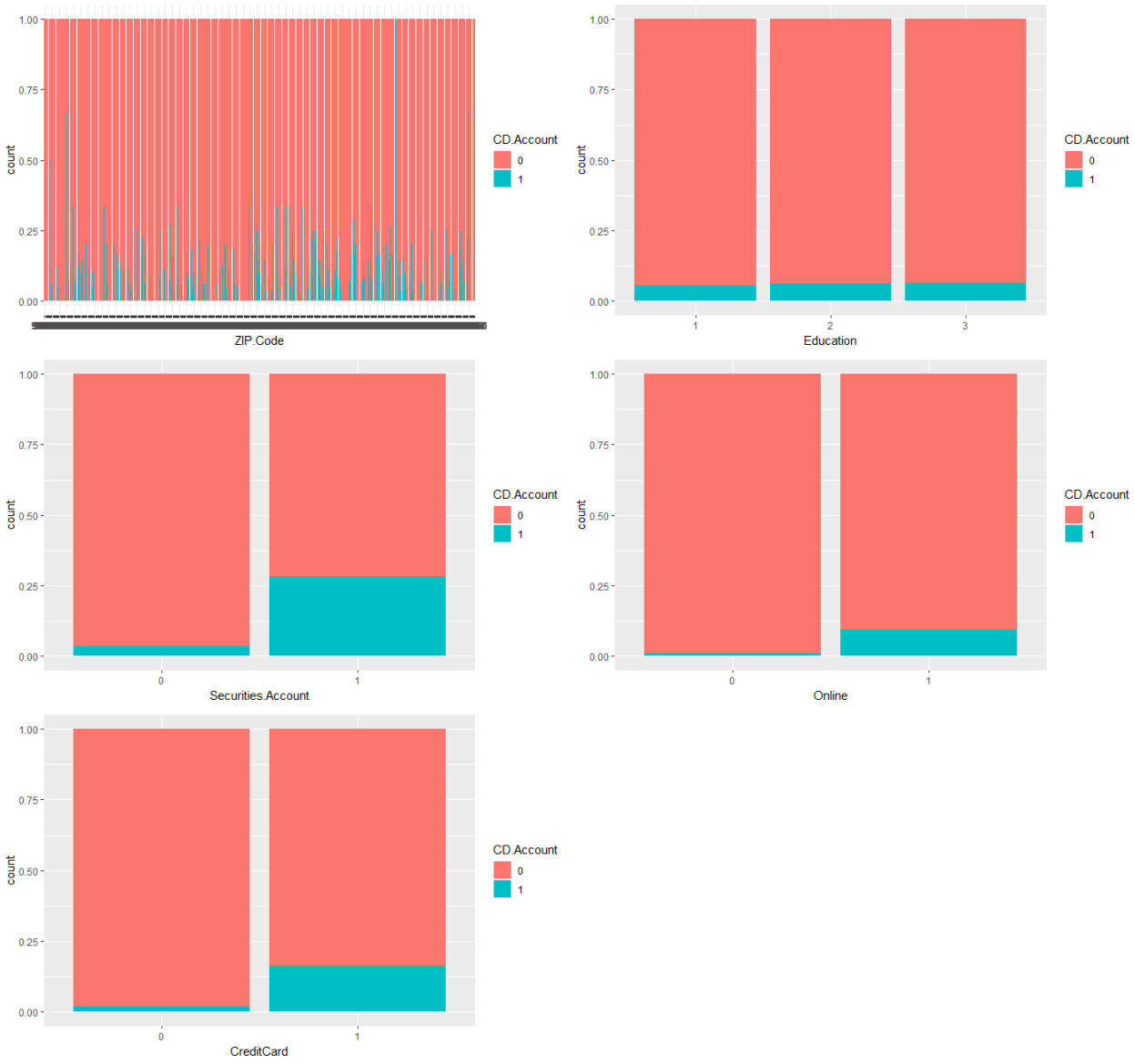


1.2.3. Correlation between the ownership of a credit deposit account and the other factor variables:

With a p-value of < 2.2e-16, there is a very high correlation between the ownership of a credit deposit account and:

1) Ownership of a securities account.
2) Use of the online banking services.
3) Ownership of a credit card.

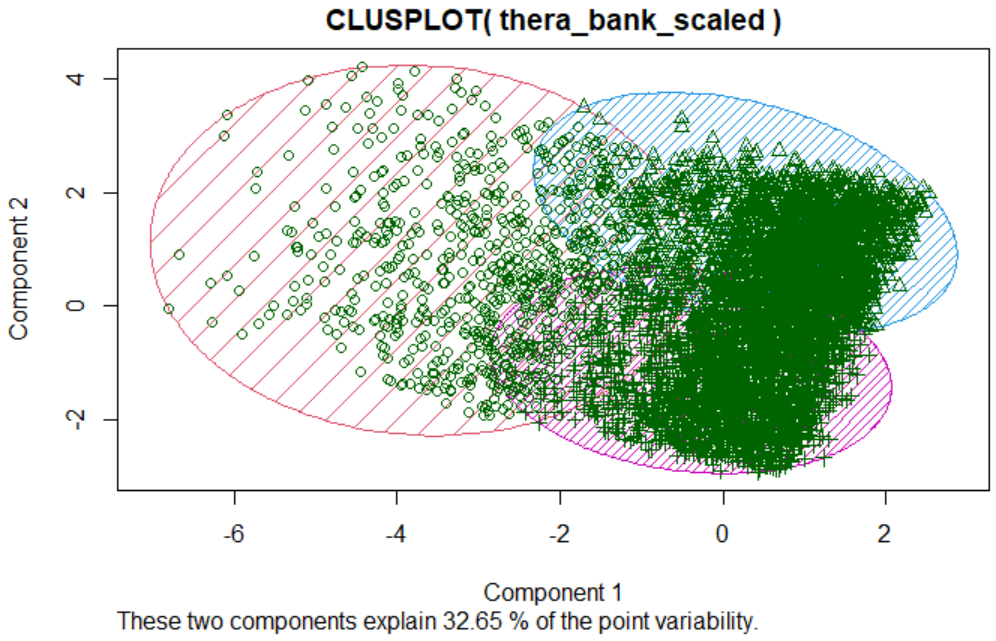
Other variables have no impact on the ownership of a credit deposit account



2. Clustering

Clusters were created the K Means Clustering method because it is the best method to reach the numbers of centroids that best fit the data.

Customers were divided into 3 clusters. The characteristics of each cluster was as follows:



2.1. Cluster1:

Age: Medium

Experience: Medium

Income: High

Family members: Low

Credit card average: High

Education: Low

Mortgage: High

Personal loan: High

Securities account: High

Credit deposit account: High

Online use: High

Credit card ownership: High

2.2. Cluster 2:

Age: High

Experience: High

Income: Low

Family members: Medium

Credit card average: Low

Education: High

Mortgage: Medium

Personal loan: Low

Securities account: Low

Credit deposit account: Medium

Online use: Medium

Credit card ownership: Medium

2.3. Cluster 3:

Age: Low

Experience: Low

Income: Medium

Family members: High

Credit card average: Medium

Education: Medium

Mortgage: Low

Personal loan: Low

Securities account: Medium

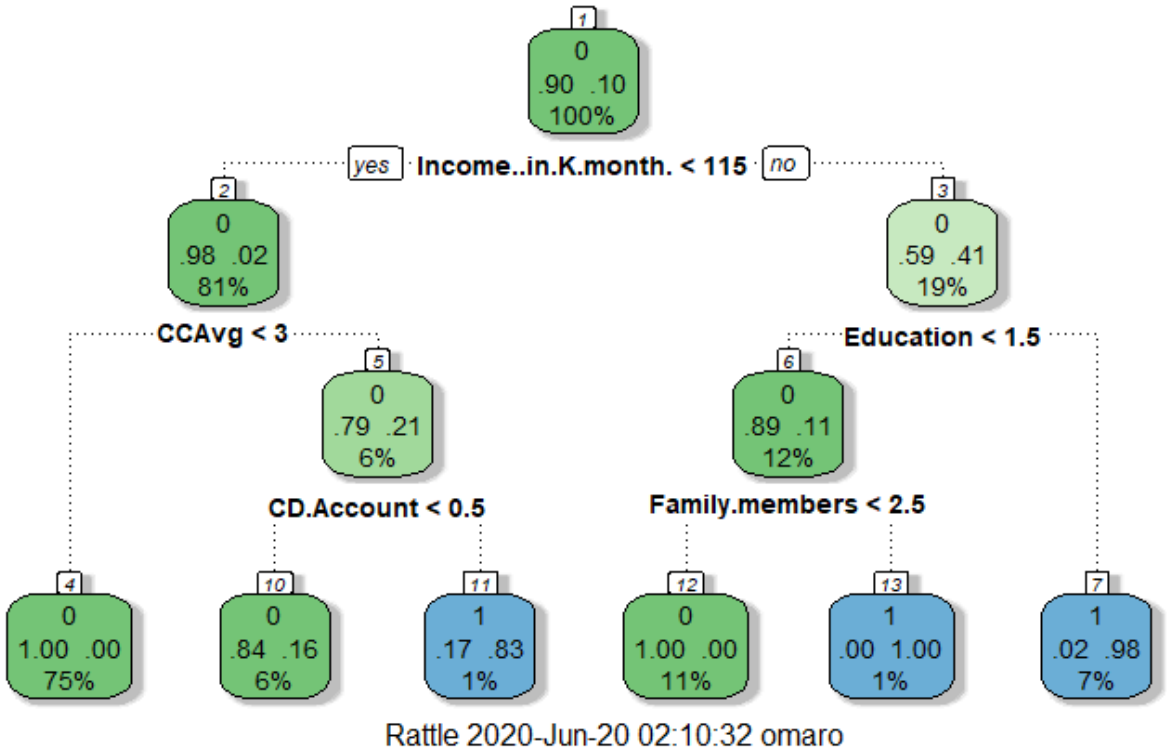
Credit deposit account: Low

Online use: Low

Credit card ownership: Low

3. Predictive Models

3.1. CART model



According to the CART model:

- Customers with an income of less than $115,000/year, credit card average more than or equal to 3 and have a credit deposit account have an 83% chance of accepting a personal loan. They make up 1% of the total customers.
- Customers with income more than $115,000/year, are undergraduates and have 3 or more family members have a 100% chance of accepting a personal loan. They make up 1% of the total customers.
- Customers with income more than $115,000/year, are graduates or advanced/professionals have a 98% of accepting a personal loan. They make up 7% of the total customers.
- 75% of the customers of Thera Bank have an income less than $115,000/year and a credit card average of less than 3. These customers have a 100% to refuse a personal loan.

3.1.1. Model Performance:

Train data:

Accuracy = 98.57%

Sensitivity = 98.69 %

Specificity = 97.35%

KS value = 0.9126

AUC = 0.9815

Gini = 0.8705

Concordance = 96.5%

Discordance = 3.5%

Test data:

Accuracy = 97.87%

Sensitivity = 97.97%

Specificity = 96.67%

KS value = 0.9243

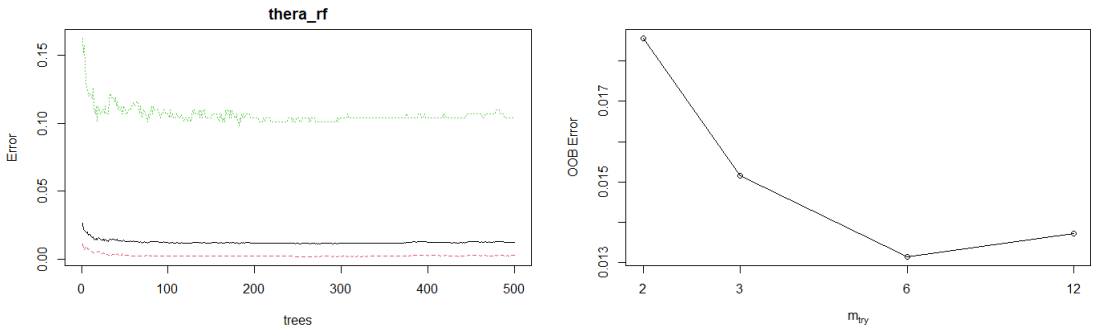
AUC = 0.9826

Gini = 0.8739

Concordance = 96.7%

Discordance = 3.3%

3.2. Random Forest



OBB = 1.26%

3.2.1. Model Performance:

Train data:

Accuracy: 99.34%

Sensitivity: 99.34%

Specificity: 99.37%

KS value = 0.9943

AUC = 0.9999

Gini = 0.8987

Concordance = 99.85%

Discordance = 0.015%

Test data:

Accuracy: 98.2%

Sensitivity: 98.33%

Specificity: 96.8%

KS value = 0.9662

AUC = 0.9979

Gini = 0.8991

Concordance = 99.78%

Discordance = 0.22%

3.2. Conclusion

Seeing as the bank is trying to find potential customers who would accept a personal loan. It is the most important model performance measure is the sensitivity (rate of true positives) as they need to reach the most customers who have a potential to accept the loan.

With a baseline accuracy of 90.4%. The two models should be considered to have performed very well.

But the Random Forest model performed better. Which is expected seeing as the random forest is a more robust model as it is resistant to over-fitting and under-fitting