**Project Description**

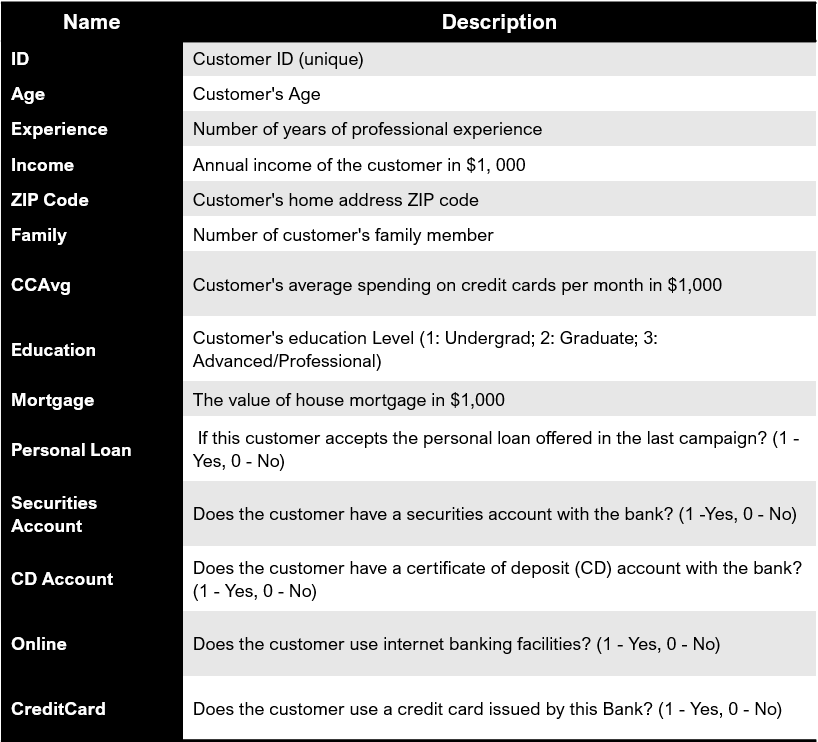
As a service industry, banking uses customer insights to learn more about their customers to save costs and increase profit. By analyzing customers’ personal data and needs, banks can understand their current customers better and expect their future ones. In detail, these techniques help banks to improve their business by increasing sales targeting and effectiveness based on personalized business strategy.

In this project, our team is going to predict whether the bank can successfully offer a personal loan to its customer. By exploring the potential customers, the bank can make appropriate marketing strategy and achieve better results.

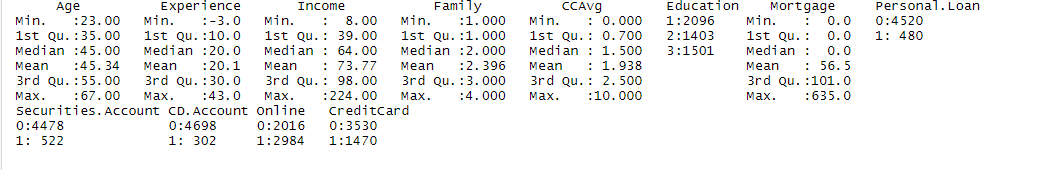
**Data Exploration and Preprocessing**

In this dataset, we have 5,000 entries, 14 columns in total. In this project, we focus on identifying potential customers for personal loan. Hence, our output variable will be personal loan. About the input variables, we will have to consider different aspects before deciding which ones are critical in our project.

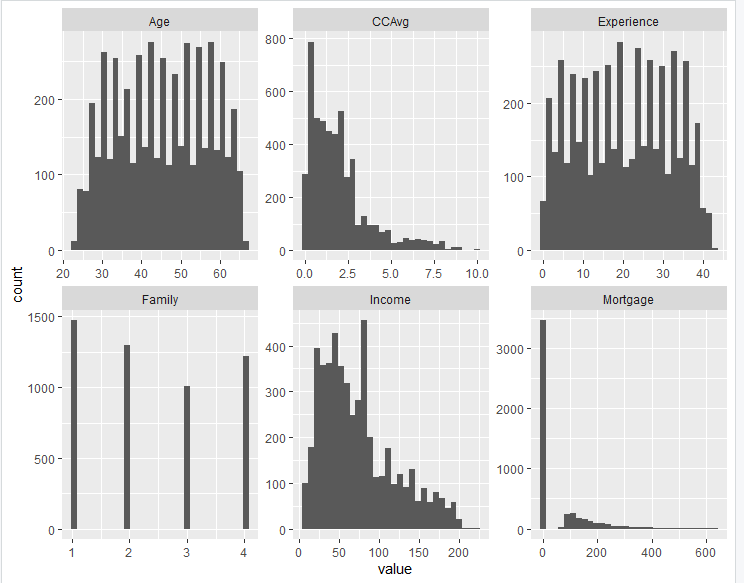
Below is the data dictionary:



Firstly, we can drop ID and Zip.Code Columns because they do help in identifying potential customers. Then, we factor categorical variables including "Education", "Personal.Loan", "Securities.Account", "CD.Account", "Online", "CreditCard”. Then, we check for outliers and data entry errors by looking at statistical numbers.

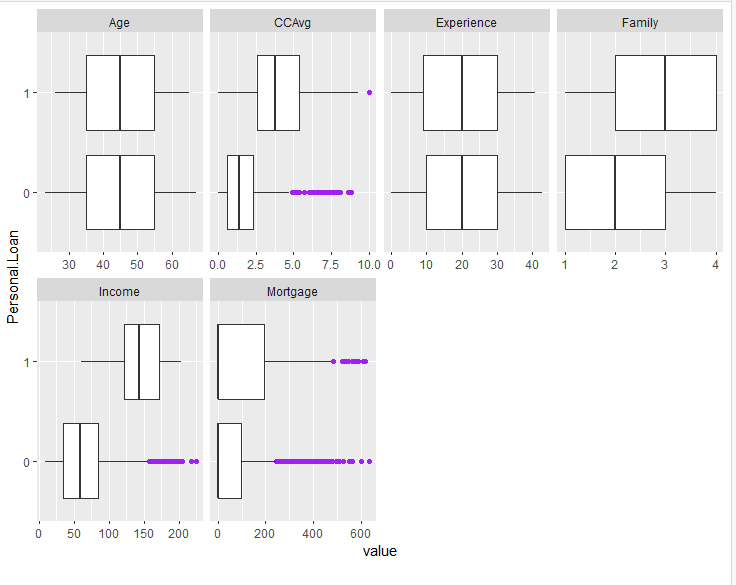


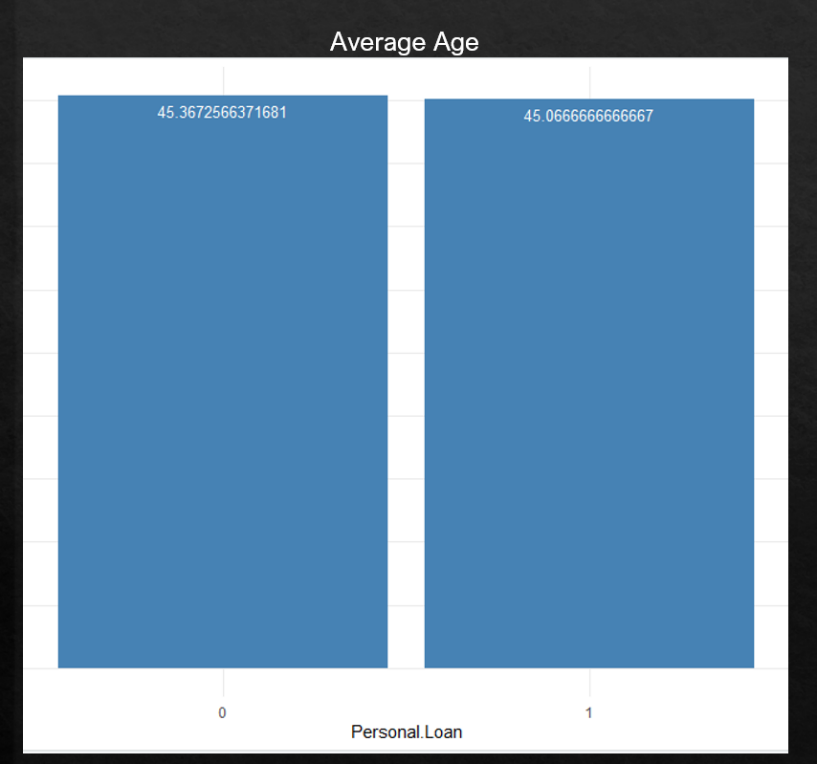
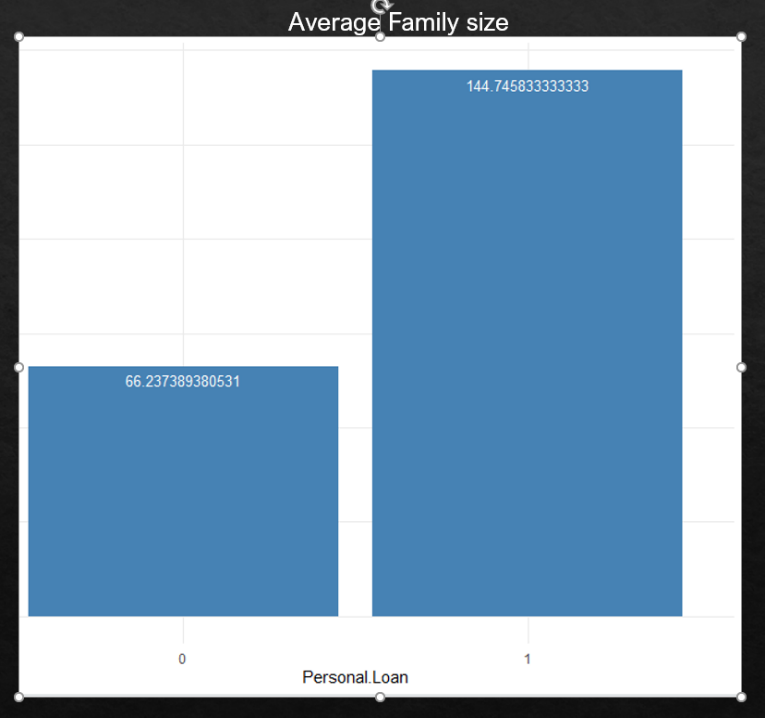
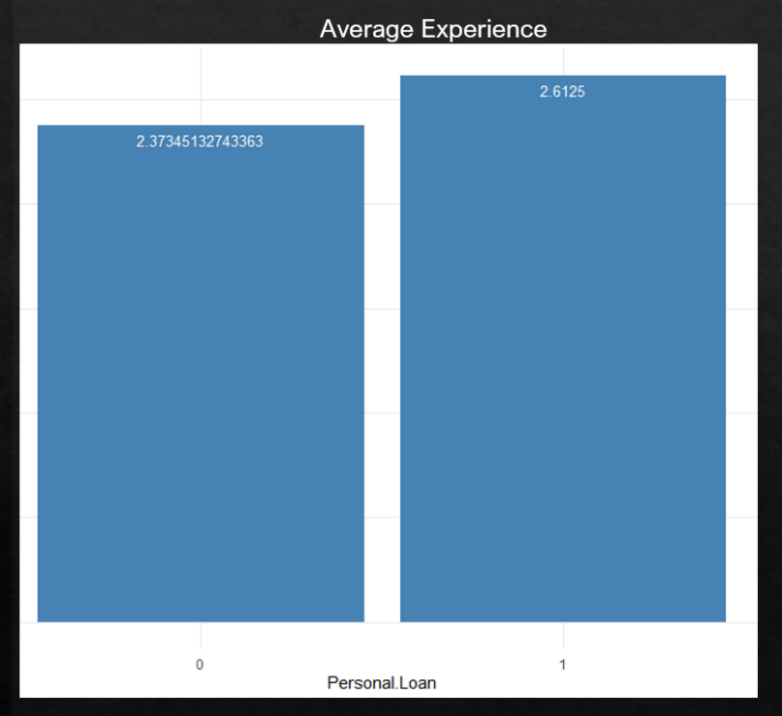
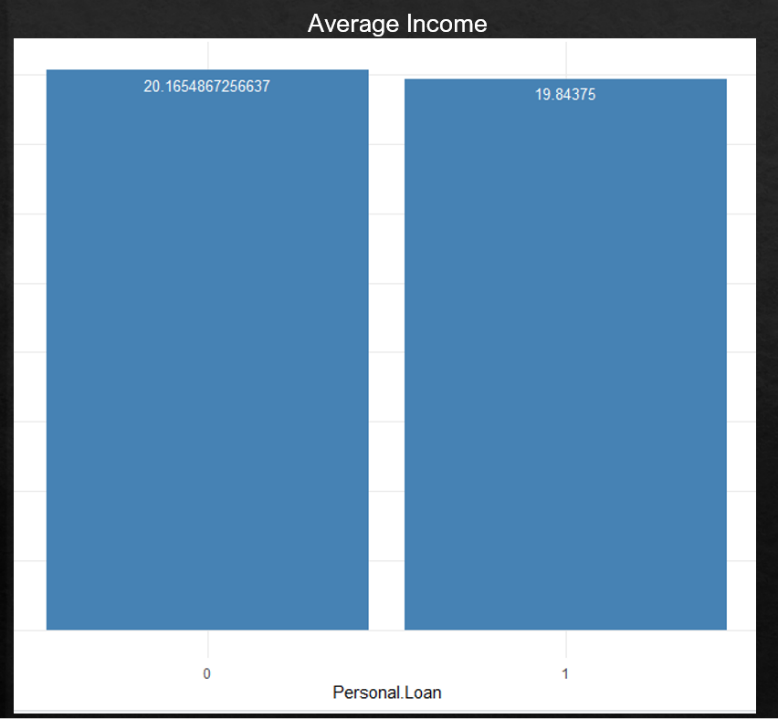
There are negative values in the “Education” column. They are data entry errors. Thus, we applied absolute function for that column to fix those errors.

Now, we explore data deeper by looking at different types of graphs. 

The above graphs show us the distribution in age, average spending on credit card, professional experience, family size, annual income, and the value of house mortgage in the dataset. The dataset includes a wide range of age, experience, family size, and income. However, the dataset includes mostly people with no mortgage value and the average credit card spending between 0 and $2,500. We will have to pay attention on those variables when reporting the results of the models and making recommendations.

The boxplots below tell us the similar thing. There are outliers in average spending on credit card, and mortgage values. In this project, we decide to keep those outliers because we believe those outliers can help us understand customers better.



Then, we compare the average age, experience, income, and family size of people who accept personal loan and people who do not. The graphs illustrate that while those two kinds of people have very similar age, experience, income, people who accept personal loan have bigger families than who do not.



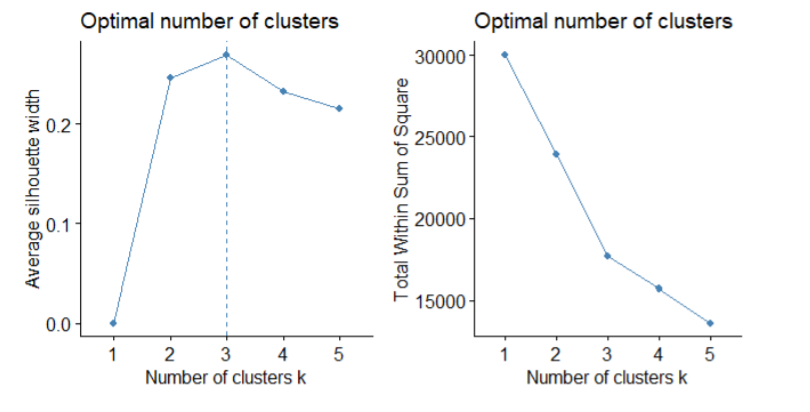


Density Plots and Frequency Plots are generated across various attributes and by Personal Loan availed or not, to get more information about the data distribution.

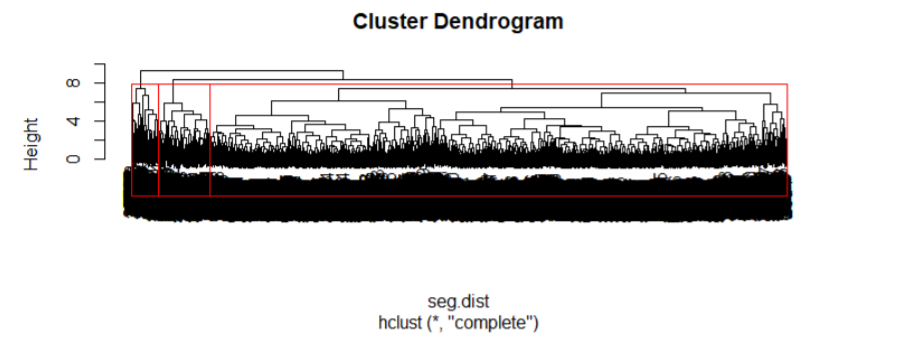
**Modelling**

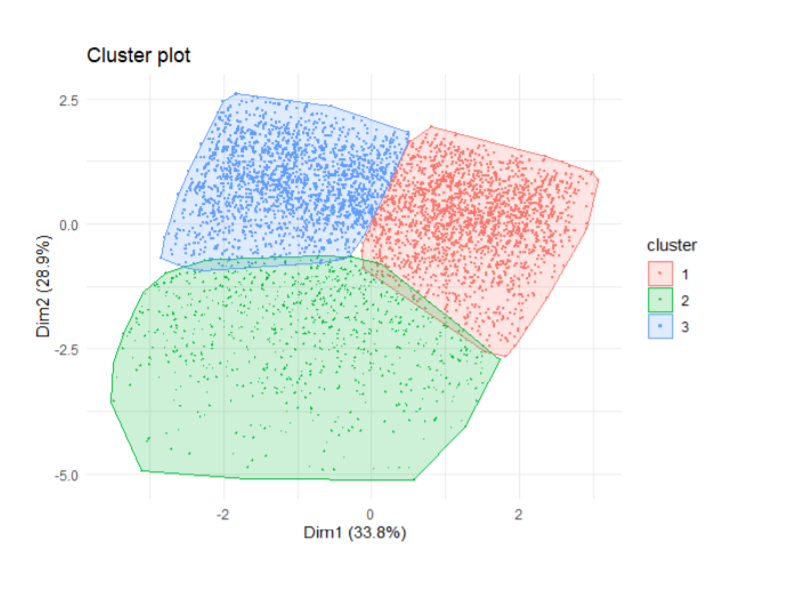
1. **K-Means Clustering for Consumer Segmentation**

Firstly, numerical columns are scaled, then package “factoextra” is installed, it provides some easy-to-use functions to extract and visualize the output of multivariate data analyses". To determine and visualize the optimal number of clusters, function “fviz\_nbclust()” is used. Optimal number of clusters were found using two different methods: within cluster sums of squares and average silhouette. Visualization can be seen below for both the methods:



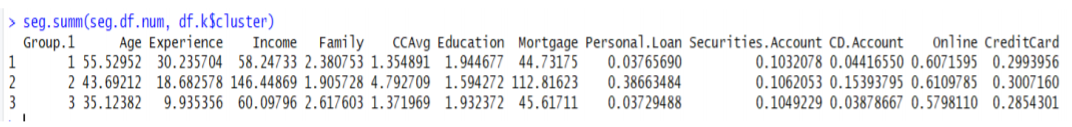
Both methods point to 3 being the optimal number of clusters. As an additional check, hierarchical clustering is also performed and the resulting Dendrogram is checked for 3 clusters. Visualization can be seen below:





Next, k-means clustering is implemented by specifying to generate 3 clusters . Resulting clustering plot can be visualized as seen in the left:

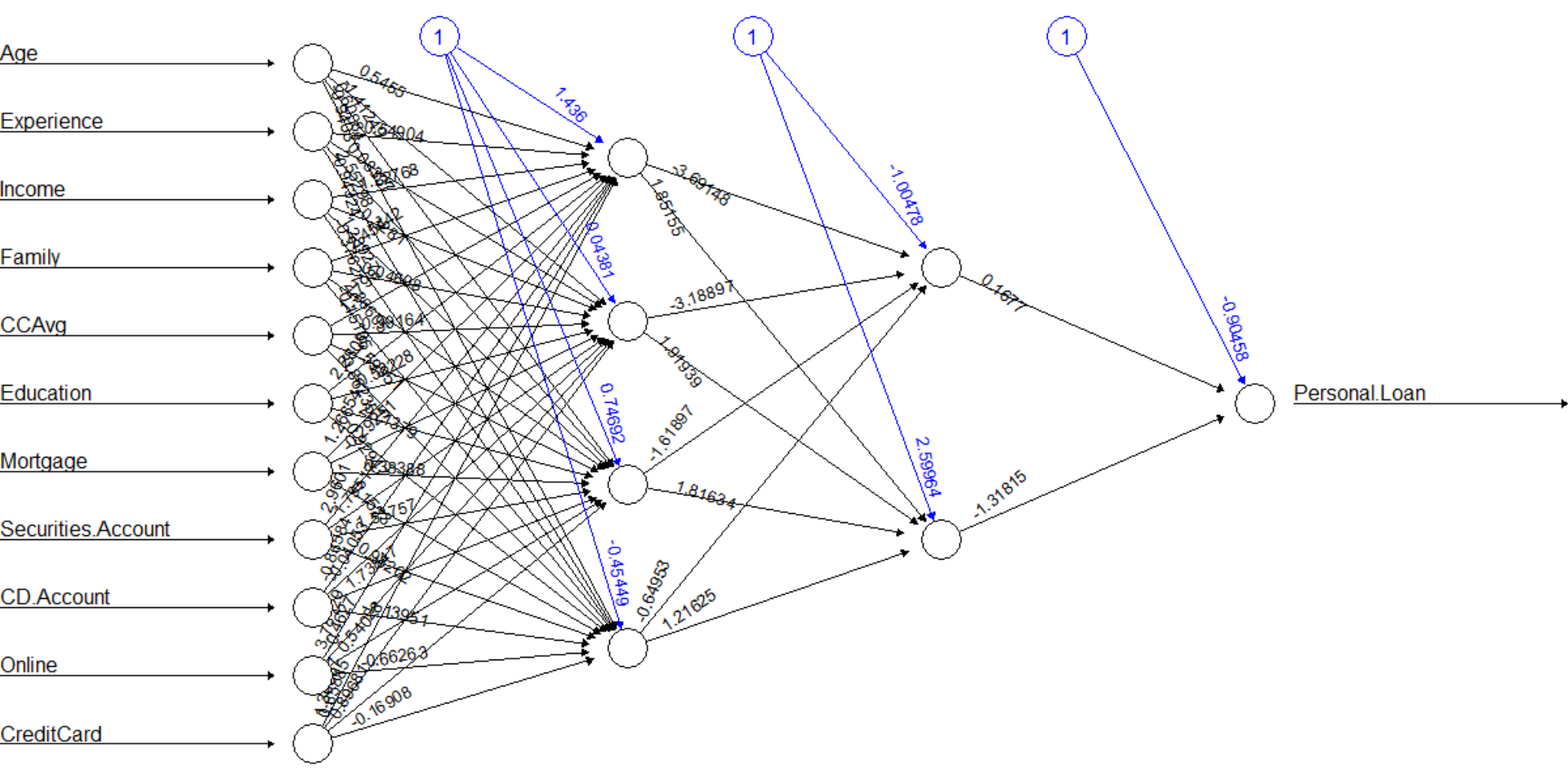
After aggregating all the attributes with respect to each cluster, following results are seen below:



Cluster 2 seems to be the segment which is most likely to avail a personal loan. This segment belongs to middle-aged people, belonging to high income category, small family size, high credit card spending, mortgage value of house being high and high probability of having Deposit account(CD.Account)

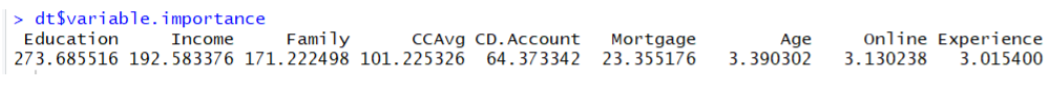
1. **Neural Networks**

Neural Nets has been fitted on a 80-20 split train-test data with the help of ‘neuralnet’, ‘nnet’ & ‘caret’ libraries and a network as shown below has been generated with hidden layers. The fitted model gave a specificity of 0.84 and an accuracy of 0.9

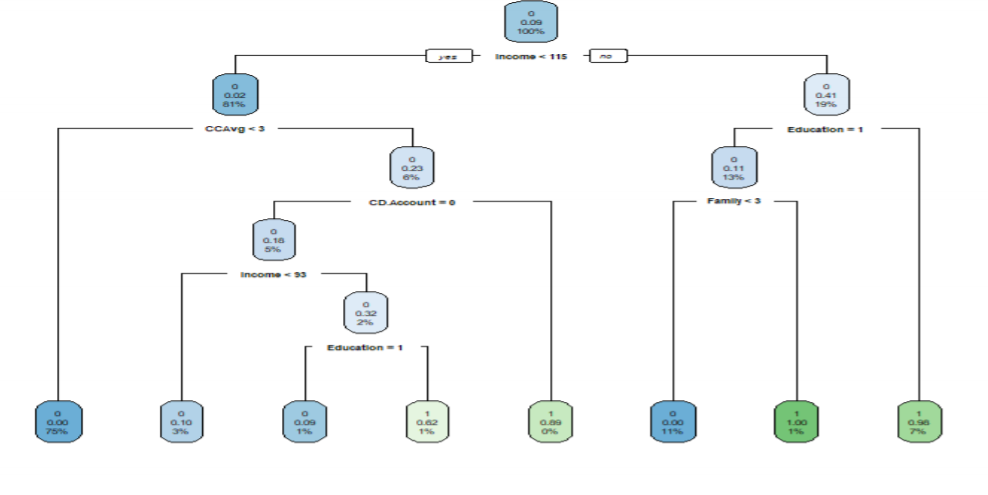


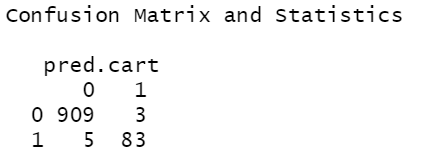
1. **Classification Methods: Decision Tree, Random Forest, Logistic Regression**
   1. **Decision Tree (CART Model):**

To start with, the dataset is split randomly into training dataset and test dataset with 80% for training and 20% left for testing. Next a decision tree model is fitted on the training dataset, using “rpart” package. Variable importance is checked for and as seen below, Education appears to be the most important, followed by income, family size and credit card spending



The tree plot can be seen below:

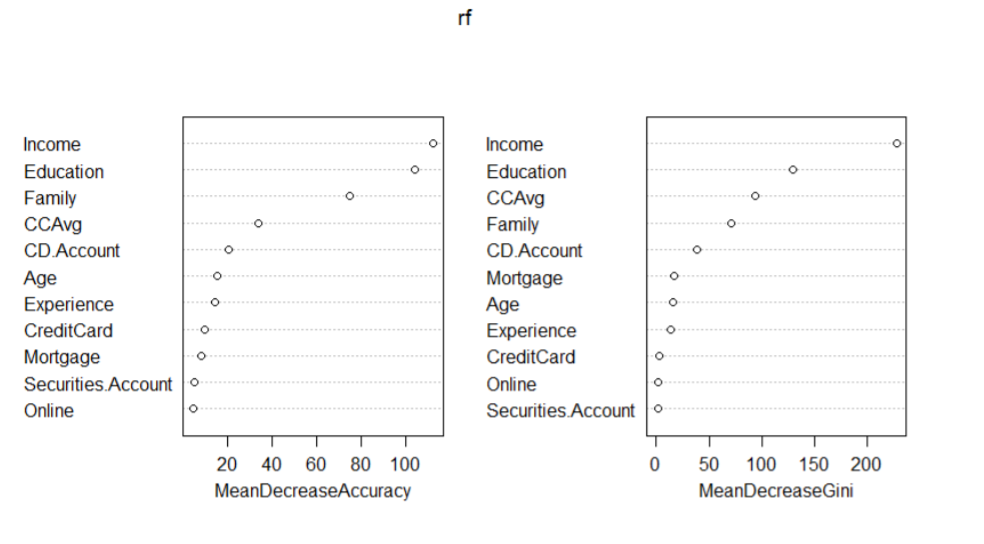


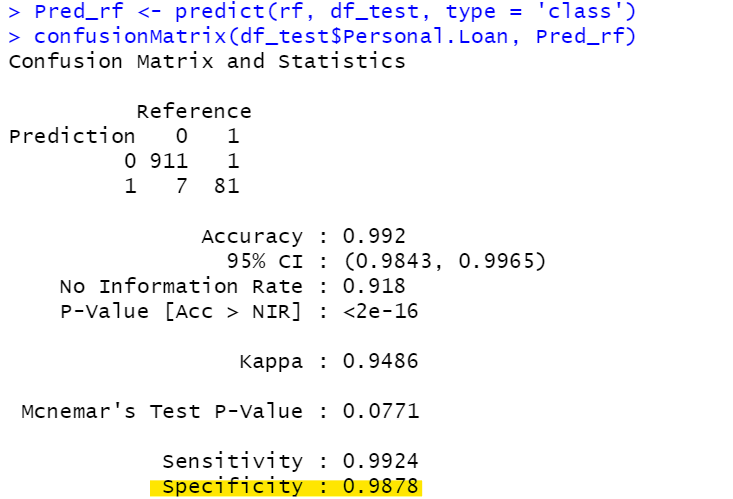


After using the fitted model on test data set, a Specificity of 0.96 is achieved. The confusion matrix is displayed to the side.

* 1. **Random Forest**

To see if better results, Random Forest method has been deployed where 100 trees are utilized for the bootstrapping purpose with minimum size of terminal nodes to be 10 and the importance of each variable has also been checked as seen below:

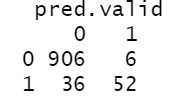




After using the RF model on test data set, a Specificity of 0.98 is achieved. The confusion matrix is displayed to the side.

* 1. **Logistic Regression**

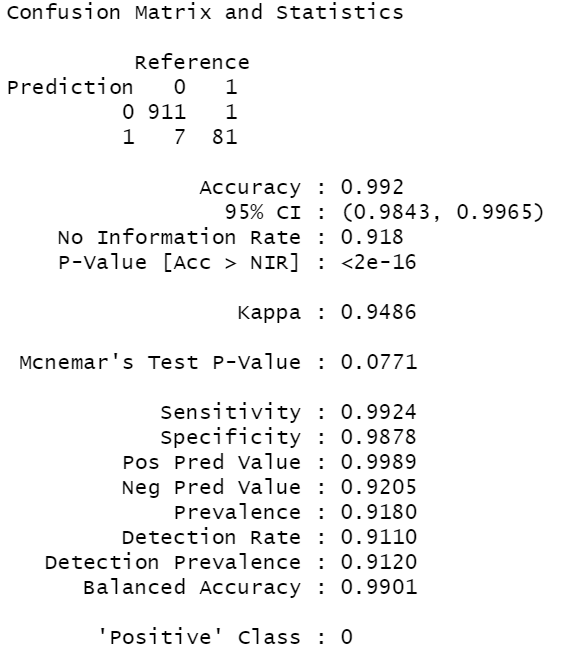
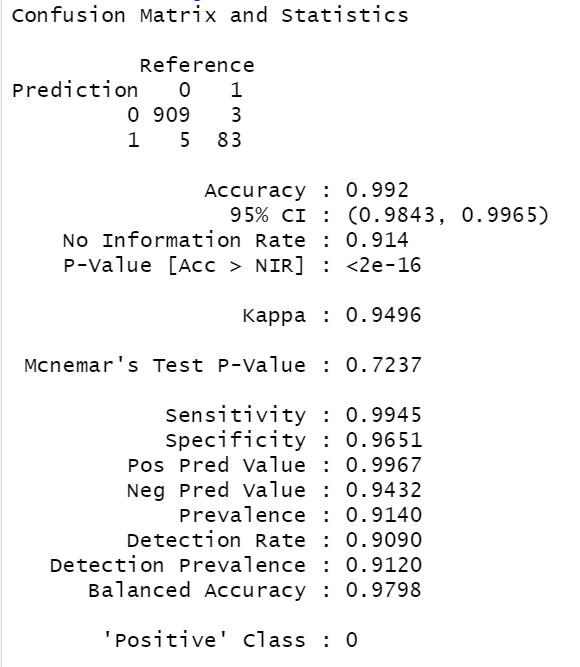
Logistic Regression model has also been fitted to check the specificity and accuracy. The model yielded a specificity and accuracy of 0.96 as shown below:



1. **Ensemble Modeling**

Finally, Ensemble technique is adopted wherein, Majority Voting method is used for CART, Random Forest and Logistic Regression values and resulted in a Specificity of 0.98. Then, the Ensemble technique is adopted again, wherein, Weighted Average method is used which resulted in a specificity of 0.96.

The results are as below:

****

**Results:**

The Ensemble model using weighted average method gave a specificity of 98% and has been chosen as the desired model.

Also, based on segmentation done using clustering, the bank can effectively target the segment, which is well educated, earning high income, with strong credit card spending history for selling more personal loans. The bank should think of more innovative schemes to attract elderly and young people with not so high incomes and spending history. These are the people who earn between 40000$ to $100,000 with credit card spending less than 3000$ per month. Bank may try to deliver loan schemes with low interest to attract these people, and in the case of elderly have dedicated staff to help senior citizens avail a loan account in a smooth manner. In the case of young people, bank should think of potential spending on attractive marketing campaigns online, specifically targeting online places where young people are likely to engage mostly. Banks can also engage HR teams of companies and promote the loan schemes to young people working in the company.

**Summary:**

Data Analytics is helpful and useful for digital transformation in Banking. In this report we explored how banks can utilize the customer data they possess to come up with effective market segmentation and predict who is likely to avail a personal loan based on certain attributes of the individual person. After implementing machine learning models related to clustering, we have come up 3 segments which Bank can target. High income and well-educated people with good credit card spending are the main customer segment that Bank should go after and Bank should design better loan schemes for elderly and young people.