Mini Project – Thera Bank Project

Model Report

Table of Contents

1. [Project Objective 3](#_bookmark0)
2. [Exploratory Data Analysis – Step by step approach 3](#_bookmark2)
   1. [Environment Set up and Data Import 3](#_bookmark3)
   2. [Variable Identification](#_bookmark7) 3
      1. [Variable Identification – Inferences](#_bookmark8) 3
   3. [Missing Value Identification](#_bookmark11) 4
   4. [Univariate Analysis](#_bookmark9) 4-5

3 Bivariate Analysis………………………………………………………………………………………………………………… 6

4 Cart Model……………………………………………………………………………………………………………………… 7-10

5 Random Forest…………………………………………………………………………………………… 11-13

6 Comparison …………………………………………………………………………………………. 14

7 [Conclusion……………………………………………………………………………………………………………..](#_bookmark14) 15

**Project Objectives**

The main objective of this project is to build a decision tree model to identify which group of customers have a higher chance of taking a personal loan. This is further divided into small objectives which are as follows:

1. To perform exploratory data analysis.
2. To build a cart tree and prune it by tuning appropriate parameters like minbucket, minsplit, complexity parameter,etc.
3. To build a random forest model and tune appropriate parameters like nodesize, mtry, etc.
4. Selecting the best model by comparing different model performance measures like accuracy, KS stat, gini coefficient, concordance ratio, etc.

**Environment set up and Data Import**

Setting up of the working directory help in accessing the dataset easily. Different packages like “tidyverse”, “car”, “InformationValue”, “randomForest”, “caTools”, “caret”, “rpart.plot”, “gini”, “ROCR” and “rpart.plot” are installed to make the analysis process slightly easier.

The data file is in “.xlsx” format. To import the dataset, read\_excel function(available in readxl package) is used. The dataset consists of 5000 rows and 14 columns.

**Variable Identification**

**“head”** function is used to show top six rows of the dataset.

**“tail”** function is used to show last six rows of the dataset.

**“str”** function is used to identify type of the variables. In this dataset, all the variables are numeric. Initially all the variables are numerical. Some of the variables like Personal Loan, Credit card, Experience, CD account, Security account and Online are converted into factors using as.factor() function.

**“summary”** function is use to show the descriptive statistics of all the variables

**Missing Values & Negative values**

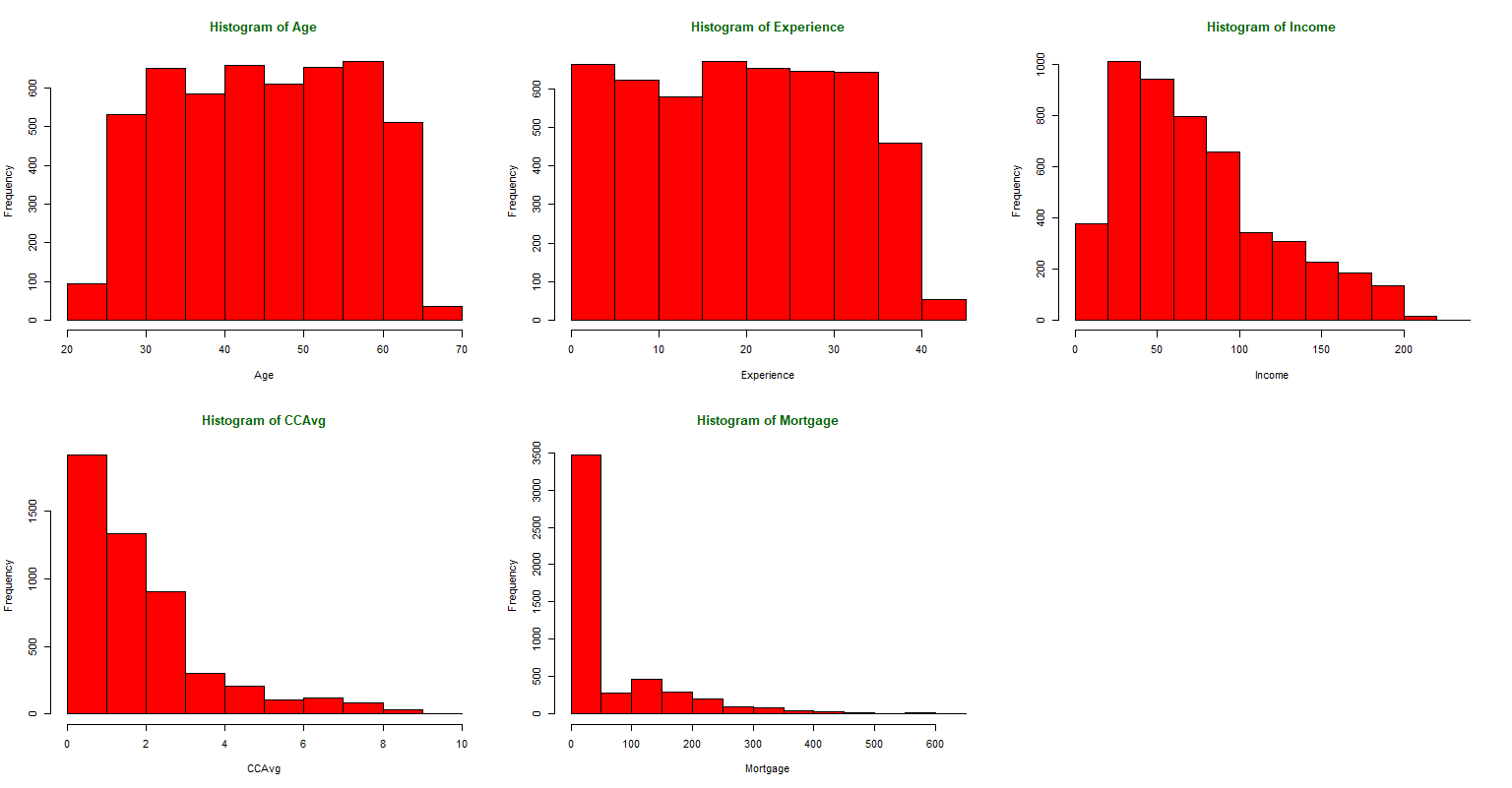
There are 18 missing values in variable family members. All the 18 values are removed to improve the effectiveness of the model.

There is presence of 52 negative values in Experience variable. The experience of the individual can’t be negative and thus need to be treated. All the negative values are converted into positive values using absolute function(abs()).

**Univariate Analysis**

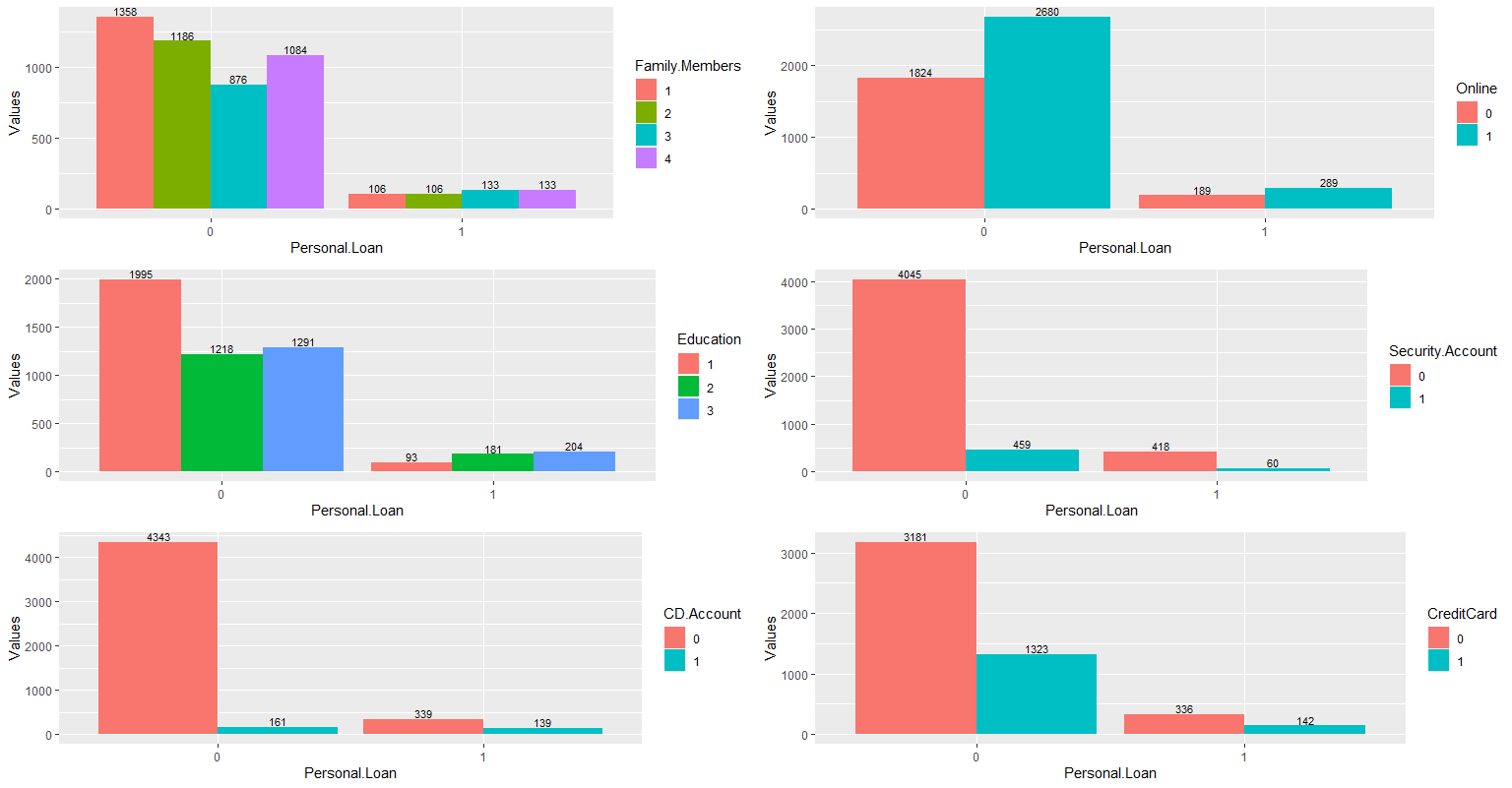
1. **Histograms**

The histogram is used to signify the normality and skewness of the data. It also signifies central location, shape and size of the data.



From above figure we can infer that:

1. Age and Experience variable is normally distributed.
2. Variables like income, Mortgage and average spending on credit card shows sign have right skewness.
3. **Barplots**

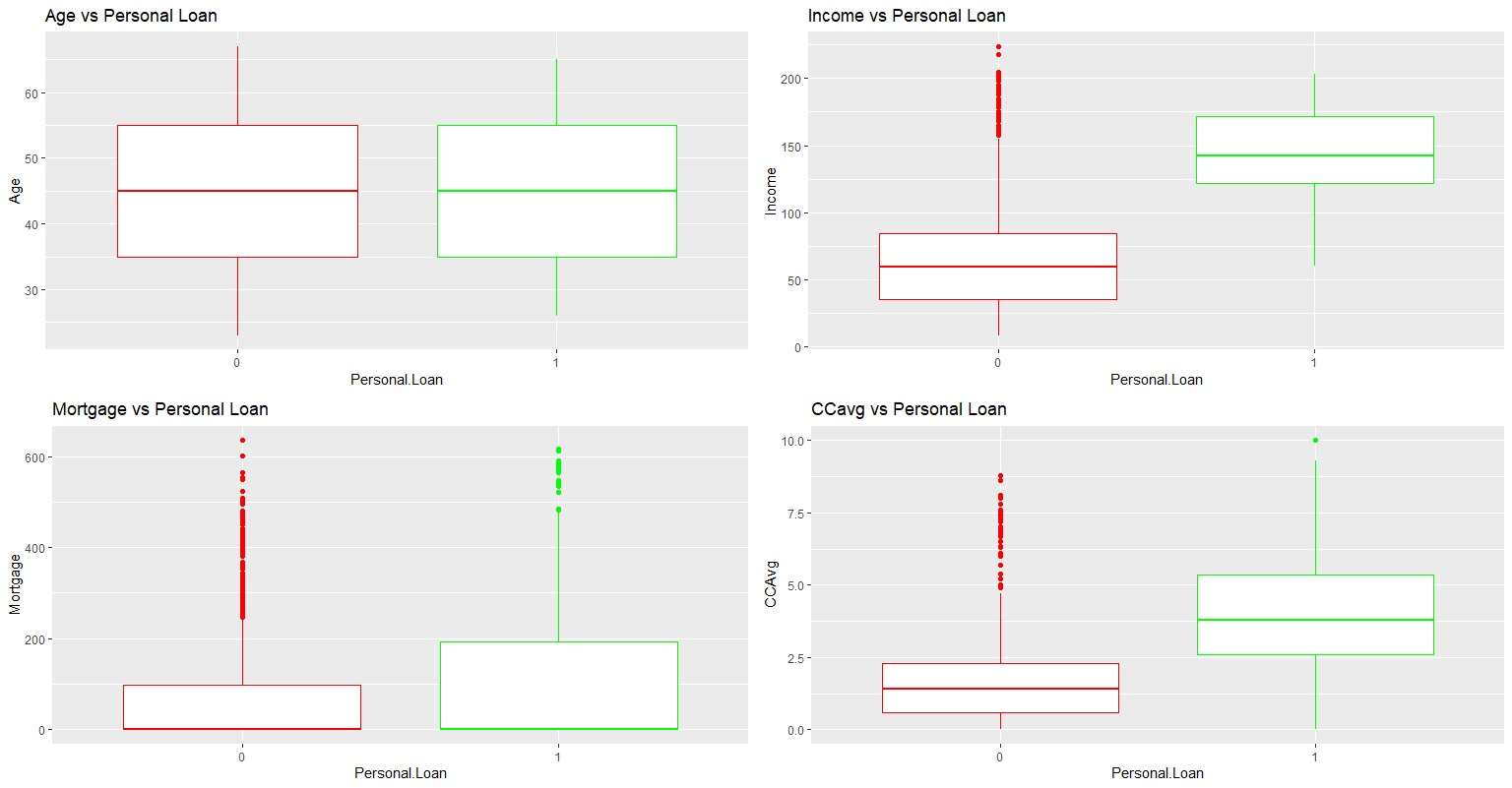


From above figure we can infer that:

1. Out of total customers who have taken personal loan, only 189 customers don’t use internet banking and 289 customers use internet banking.
2. Out of total customers who have taken personal loan, 336 customers don’t have credit cards and only 142 customers have credit cards.
3. Out of total customers who have taken personal loan, 418 customers don’t have security account and only 60 customers have security account.
4. Out of total customers who have taken personal loan, 339 customers don’t have CD account and only 139 customers have CD account.
5. Out of total customers who have taken personal loan, 93 customers are undergraduate, 181 are post-graduate and 204 customers are professionals.
6. Out of total customers who have taken personal loan, 106 customers have only 1 family member,106 customers have only 2 family members, 133 have 3 family members and 103 have 4 family members.

**Bivariate Analysis**

1. **Boxplot**



It can be easily seen that only age variable do not have any outliers. Mortgage have highest outliers followed by CCavg and Income.

**CART (Classification and Regression Tree)**

Cart decision tree is a machine learning algorithm used in predictive analysis. It can be used to solve both regression and classification problems. Under this, the dataset is divided into branches and then into sub – branches, based on some decision rules, which leads to formation of a tree – like structure.

The cart algorithm uses recursive binary splitting technique to choose a variable which leads best split at each step. The **“best split variable”** can be chosen using different metrics like Gini Impurity, Entropy, Variance, etc. We will emphasis on the Gini impurity metrics.

**Gini Impurity**

It is a measure which show that how many randomly chosen data points if labelled randomly are incorrectly labelled. The formula used to calculate the Gini impurity is given by:

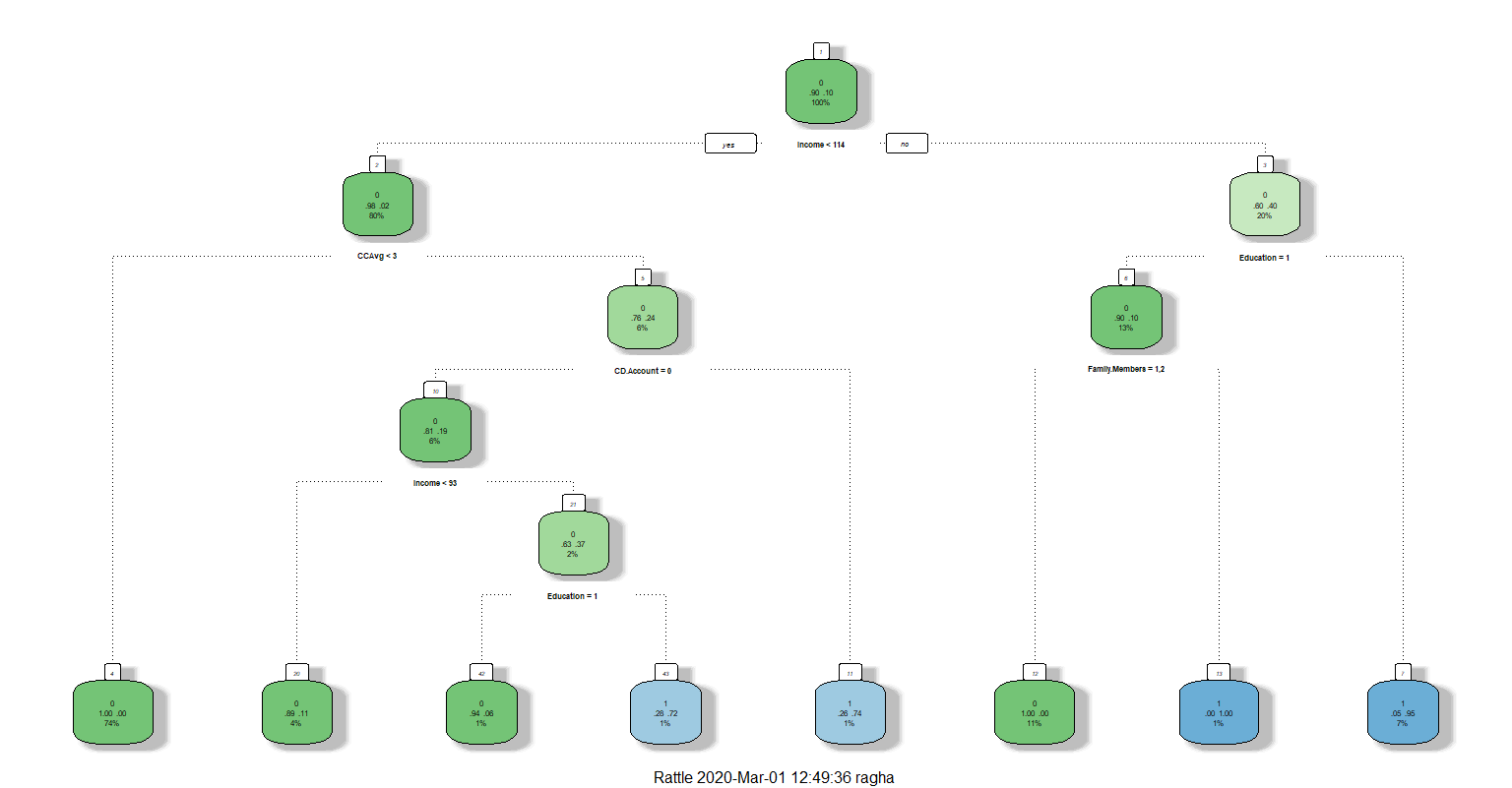
**Gini impurity = 1 – sum[(pi)^2]** where pi is the probability of item with label i to be chosen.

If Gini impurity = 0 or 1, the dataset is completely pure.

If Gini impurity = 0.5, the dataset is highly impure.

**Gini gain** = Gini impurity at root node – Weighted average of Gini impurity of nodes at each split.

The variable (groups in a variable) which will bring highest Gini gain at the time of each split should be considered as best split variable. It can be seen in the figure below that income variable with split (< =144, > 144) leads to highest Gini gain and thus used as first splitting mechanism.



**Pruning of the tree**

Due to use of recursive binary splitting mechanism by cart algorithm, sometimes data is divided into very smaller groups which leads to overfitting of the data. Thus, to avoid overfitting of the data, pruning is used to reduce the size of the tree by selecting a tolerance value of different parameters.

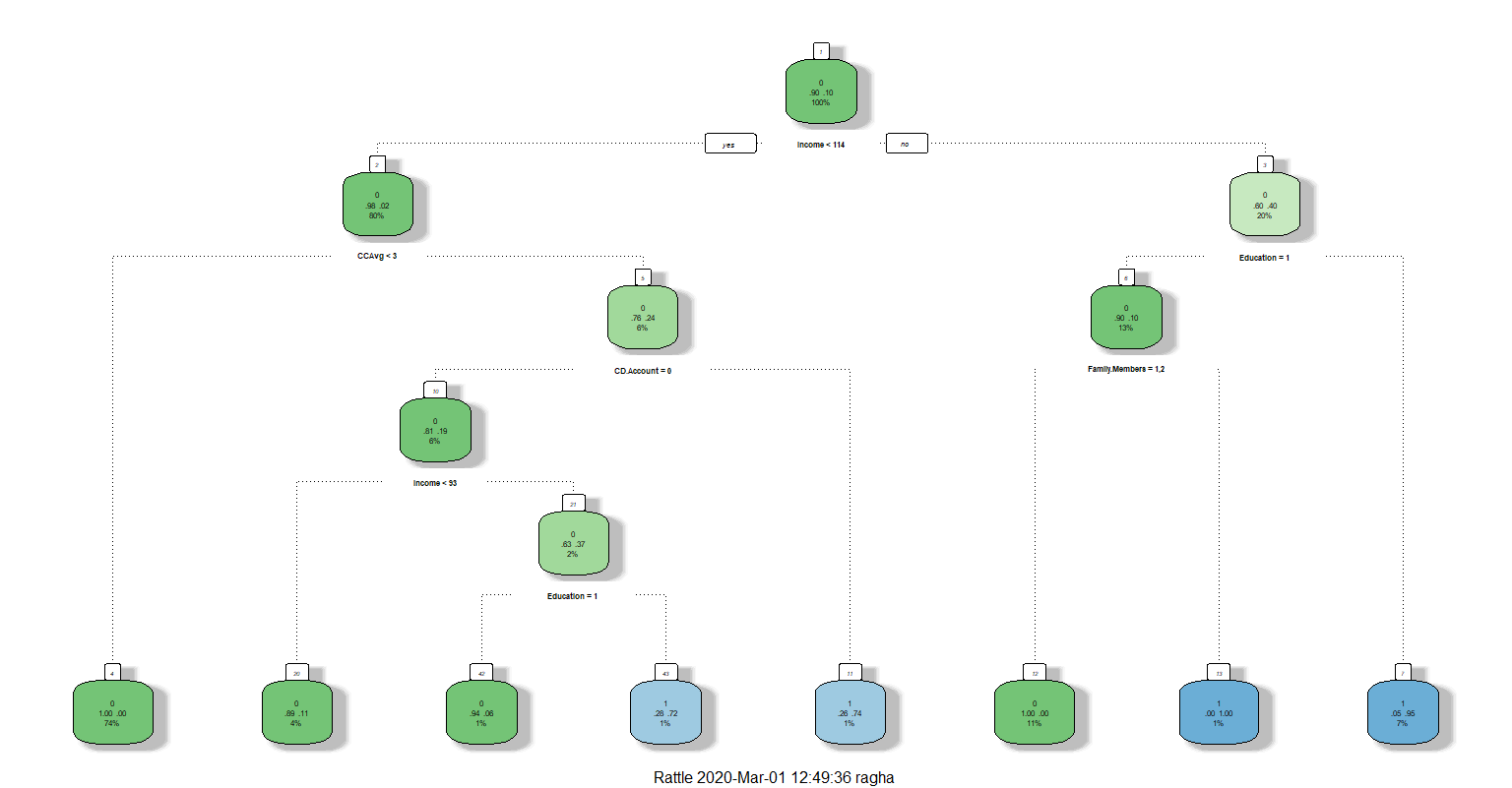
The most common pruning is known as complexity parameter pruning. This type of pruning implies that each split in a tree will depends upon the level of decrease in the relative error. If the decrease in the relative error is more than the pre-defined threshold complexity parameter value (alpha), then the split will take place. Otherwise, there will be no splitting.

To select the appropriate threshold of the complexity parameter (alpha value) is another task to be solved. K-fold cross validation is used to select the best cp value. The cp value which is having least weighted average cross validation error.

|  |  |  |  |
| --- | --- | --- | --- |
| CP | nsplit | rel error | xerror |
| 0.325893 | 0 | 1.00000 | 1.00000 |
| 0.136905 | 2 | 0.3482 | 0.36905 |
| 0.013393 | 3 | 0.21131 | 0.23810 |
| 0.010000 | 7 | 0.14583 | 0.18155 |

In the given dataset, it can be seen from above table that minimum weighted average cross validation error is 0.14583. So, corresponding cp = 0.01 is the best cp value.

After pruning the tree, the final result we get is as under:



From the above tree we can draw following conclusions:

1. An undergraduate customer whose income is above 144k and whose family size is 3 or 4 has taken a personal loan.
2. A customer whose income is above 144k, having a postgraduate or a professional degree should be given personal loan.
3. A customer whose income is less than 144k, who have average credit card spending more than 3k and who has certificate of deposit (CD) account should be given a personal loan.
4. A postgraduate or a professional customer whose income is between 30k to 144k, even though his/her average credit card spending is more than 3k and he/she doesn’t have certificate of deposit (CD) account should be given a personal loan.

**Model performance**

The dataset was split randomly into test set and training set with a split ratio of 7/10. The cart analysis is run on the training data set and the trained model is used to make prediction of test data set. This process is called cross validation of the model.

Different key performance measures like accuracy, sensitivity, specificity, Area under curve, KS – stat, Gini inequality measure, etc., are used to check the efficiency and effectiveness of the model. The results drawn from performance measure are as follows:

Train Test Difference

Accuracy 0.9860000 0.98533333 0.0006666667

Classification error 0.0140000 0.01466667 -0.0006666667

Sensitivity 0.9924123 0.99188791 0.0005243612

Specificity 0.9258160 0.92361111 0.0022049126

KS stat 0.9206700 0.91549902 0.0051710212

AUC 0.9603350 0.95774951 0.0025855106

Gini 0.8780017 0.87930184 -0.0013001438

Concordance 0.9736404 0.96470420 0.0089362011

It can be easily seen from above table that all the performance measure signifies that the model is extremely good and effective. Measures not only performed well on the training dataset but also on the testing dataset. This implies there is no overfitting of the data.

**Random Forest**

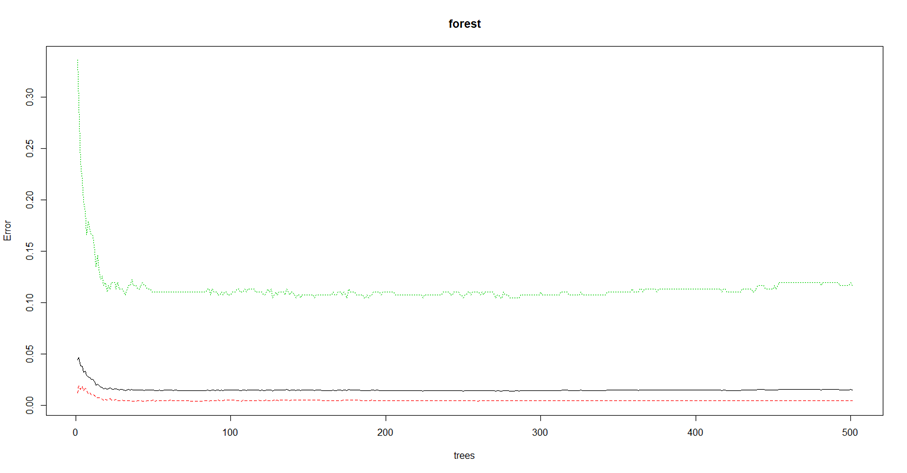
Random forest is an ensemble machine learning technique used to solve regression and classification problems. Random forest is an improvement over the cart algorithm analysis because instead of building a single tree from the dataset, the dataset set is divided into smaller subsets and then multiple trees are formed. The average value of all the trees in taken as the final value. In case of regression, we use mean of regression trees and in case of classification we use mode of regression trees.

Steps involved in a Random Forest:

1. Bagging/Bootstrap aggregating: It is a method of taking out n number of samples from the dataset by sampling with replacement.
2. Selecting the n independent variables out of the total variables(N) so as to avoid problem of multicollinearity among the variables.
3. Building decision tree model on each n samples and using cross- validation to obtain the required results.
4. Take average of the all n results obtained from n decision trees.

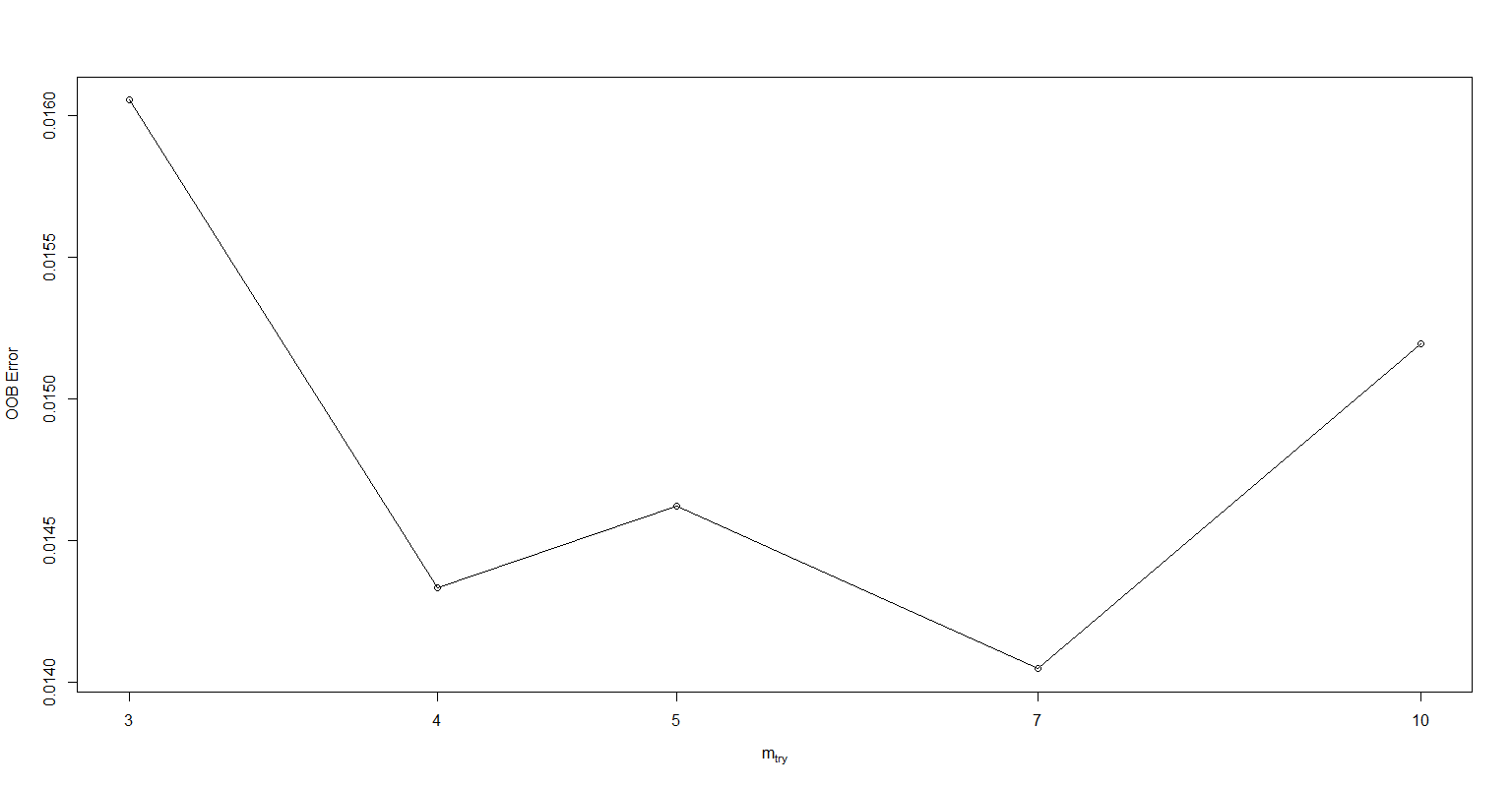
**Model building**

Initially, we started building the random forest model with hyperparameters like no. of individual variables(mtry) = 5, node size = 10 and number of tress = 501. The out of bag error rate (OOB) came out to be 1.52%. Out of bag error rate is the mean prediction error on the training samples using the trees that are not included in bootstrap samples. Lower the OOB, better the model.



**Tuning the model**

To build the best model with least OOB error, we used tuneRF() function to tune the hyperparameter. After tuning the model, we conclude that the number of independent variables(mtry) should be increased to 7 and number of trees should be reduced to 301. After tuning the model, the OOB error significantly reduced to 1.46%.



|  |  |  |
| --- | --- | --- |
|  | Mean Decrease Accuracy | Mean Decrease Gini |
| Income | 245.242553 | 191.286524 |
| Education | 221.207697 | 180.558172 |
| Family.Members | 153.756541 | 88.537596 |
| CCAvg | 38.213438 | 64.704602 |
| CD.Account | 14.089213 | 22.169334 |
| Age | 16.658784 | 9.363847 |
| Mortgage | 3.120347 | 4.348757 |
| Experience | 10.121139 | 6.744241 |
| CreditCard | 4.017780 | 1.116254 |
| Online | 3.964811 | 1.181976 |

From the above table we can see that, only variables like Income, Education, Family Members, CCAvg and CD. Account are only significant variables as the mean decrease in Gini (Gini Gain) is high for these variables. The variable (groups in a variable) which will bring highest Gini gain at the time of each split should be considered as best split variable. It can be seen in the figure below that income variable leads to highest Gini gain and thus used as first splitting mechanism, followed by Education, Family Members, CCAvg and CD.Account.

**Model performance**

Different key performance measures like accuracy, sensitivity, specificity, Area under curve, KS – stat, Gini inequality measure, etc., are used to check the efficiency and effectiveness of the model. The results drawn from performance measure are as follows:

TrainRF TestRF DifferenceRF

Accuracy 0.98881881 0.98393574 0.004883064

Classification error 0.01118119 0.01606426 -0.004883064

Sensitivity 0.98778195 0.98608059 0.001701369

Specificity 1.00000000 0.96124031 0.038759690

KS stat 0.88358209 0.86343190 0.020150185

AUC 0.94179104 0.93171595 0.010075092

Gini 0.90079491 0.89320595 0.007588961

Concordance 0.99978225 0.99816764 0.001614614

It can be easily seen from above table that all the performance measure signifies that the model is extremely good and effective. Measures not only performed well on the training dataset but also on the testing dataset. This implies there is no overfitting of the data.

**Comparison of Random Forest and Cart**

Since we are focussed on identifying the customers opting for personal loan. The key measure among all the given measure is **Sensitivity.** Sensitivity refer to proportion of positive results out of the number of samples which were actually positive.

It can be easily seen that cart model predicted higher sensitivity i.e. 0.9924(Train) and 0.99188(test) as compared to random forest model i.e 0.9850(train) and 0.98107(test). Also under cart model, the difference between train sensitivity measure and test sensitivity measure is least i.e. 0.000524 as compared to random forest where difference is 0.00392.

**Thus, we can conclude that in this case cart model performance is much better than the random forest model.**

**Conclusion**

1. While performing exploratory analysis, we came to know that there are 18 missing value in the family members variables. There was also presence of the outliers in the dataset. It was found that mortgage variable has highest outliers.
2. The best cp value which is used for pruning the tree is 0.01 with least weighted average cross validation error equal to 0.18155
3. After tuning the random forest model, the optimal number of independent variables increased from 5 to 7 and optimal number of tree equal reduced from 501 to 301. The OOB error reduced from 1.52% to 1.46%
4. Key variables in both the models are Income, Education, Family Members, average credit card spending and CD account.
5. Key measure out of all the performance measure is the Sensitivity.
6. The cart model is better than the random forest in this case.

Using cart analysis, we can infer that emphasis should be on only four groups of people which are as follows:

1. An undergraduate customer whose income is above 144k and whose family size is 3 or 4 has taken a personal loan.
2. A customer whose income is above 144k, having a postgraduate or a professional degree should be given personal loan.
3. A customer whose income is less than 144k, who have average credit card spending more than 3k and who has certificate of deposit (CD) account should be given a personal loan.
4. A postgraduate or a professional customer whose income is between 30k to 144k, even though his/her average credit card spending is more than 3k and he/she doesn’t have certificate of deposit (CD) account should be given a personal loan.