TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

GREAT LAKES
EXECUTIVE LEARNING

GROUP ASSIGNMENT – DATA MINING

Data Mining

Post-graduation in Business Analytics and Business Intelligence
Group - IX

By
Adith Bharath Ram
Avni Tandon
Jithesh Janardhan
Manikantan R
Shyamsundar

Contents

1. Introduction 2
2. Assumptions and Approach 2
3. Problem Statement 2
4. Feature Engineering and EDA 3
4.1 Univariate Analysis 3
4.2 Bivariate Analysis 4
4.3 Model Building 5
4.4 Model performance with Manual pruning 7
4.5 Model with the CP criteria 9
4.6 Variable importance table 10
4.7 ROC Curve & AUC 11
Inferences 12
5. Model using the Random Forest algorithm 13
5.1 Steps to building random forest 13
5.2 Cutoff decision 14
5.3 Train and Test the model 14
5.4 ROC Curve and AUC value 15
5.5 Variable Importance Table 15
5.6 Inferences from the model 16
5.7 Selection of model 16

1. Introduction

The given Thera dataset contains few demographic information (age, income, etc.), details related to the relationship with the client bank (mortgage, securities account, etc. and their response towards a personal loan campaign held last year of their 5000 liability customers (Depositors). Majority of their business comes through the deposits.

Now the bank aim to establish their loan business through their current customers by running a personal loan campaign. A similar engagement done last year gave encouraging results for them. Close to 10% of the customers accepted their personal loan offer. Considering this they want to widen their business line though better targeting the probable customers through another campaign.

The management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). The client wants to build the best model which can classify the right customers who have a higher probability of purchasing the loan.

2. Assumptions and Approach

- Few of the experience details were negative which are assumed to be zero as all the records fall under the age of 30.
- Making a model based on the assumption that the probability of a customer availing a personal loan increase or decrease based on other variable values in the model (Predictor variables)
- The approach is to build and CART and Random Forest models to see which one works better based on the performance measures

3. Problem Statement

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.

4. Feature Engineering and EDA

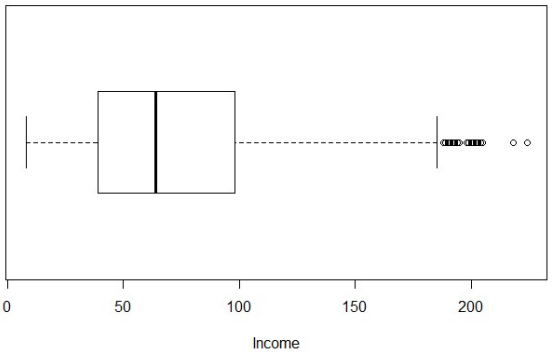
- Categorical variables are converted to factors
- The summary of the dataset gives an idea on the distribution and range as given below
- Excluding ID and Zip Code from the analysis

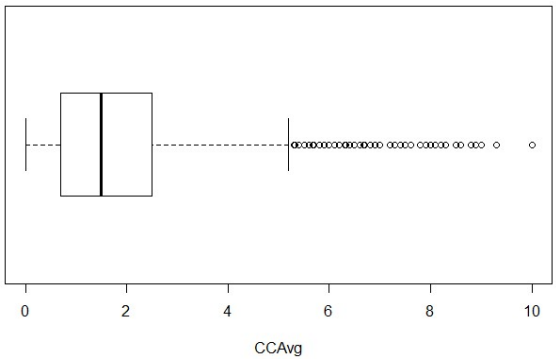
```
      Age            Experience       Income       Family_members      CCAvg          Education
 Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   :1.000   Min.   : 0.000   Min.   :1.000
 1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000
 Median :45.00   Median :20.0   Median : 64.00   Median :2.000   Median : 1.500   Median :2.000
 Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :2.397   Mean   : 1.938   Mean   :1.881
 3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000
 Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :4.000   Max.   :10.000   Max.   :3.000
                                                 NA's   :18
    Mortgage      Personal_Loan Securities_Account CD_Account Online    CreditCard
 Min.   :  0.0   0:4520        0:4478             0:4698     0:2016    0:3530
 1st Qu.:  0.0   1: 480        1: 522             1: 302     1:2984    1:1470
 Median :  0.0
 Mean   : 56.5
 3rd Qu.:101.0
 Max.   :635.0
```

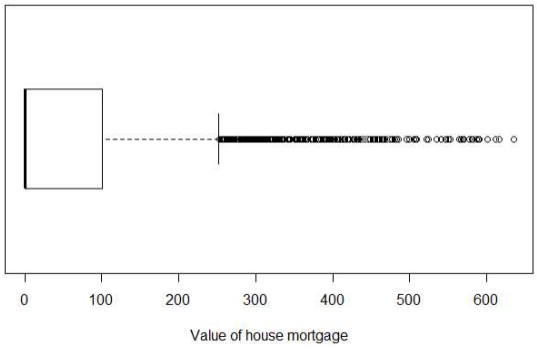
- We have 7 continuous variables and 5 categorical variables
- Since education is specified in the increased rating scale from 1-3, we are considering it as a numerical variable.
- We have some negative values in the experience column which is assumed to be all zero experience as all the customers with negative experience listed fall under 30 years of age.
- There are 18 missing values in the family members variable. There are imputed using KNN method of imputation which uses K nearest neighbors to predict the value

4.1 Univariate Analysis

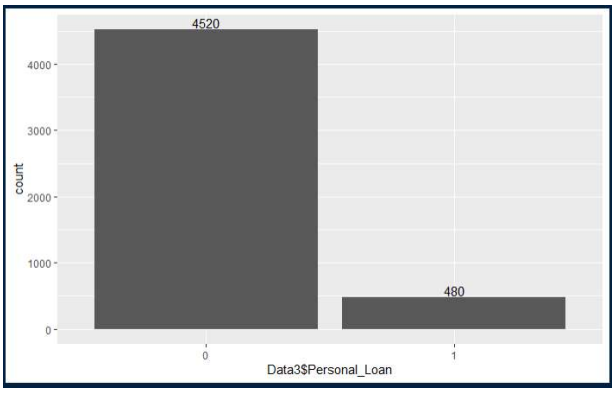
- Income, average spend on the credit card, and Mortgage distributions are right skewed with lot of outliers in the upper side has been observed.

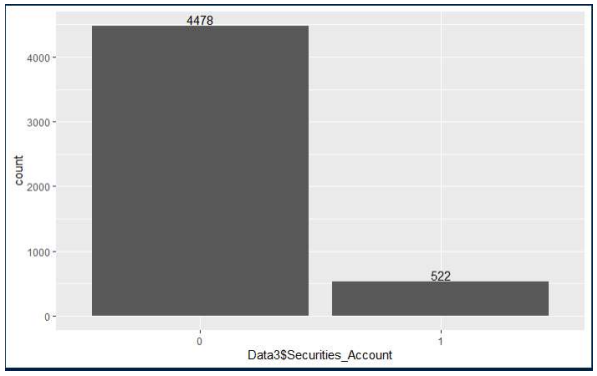


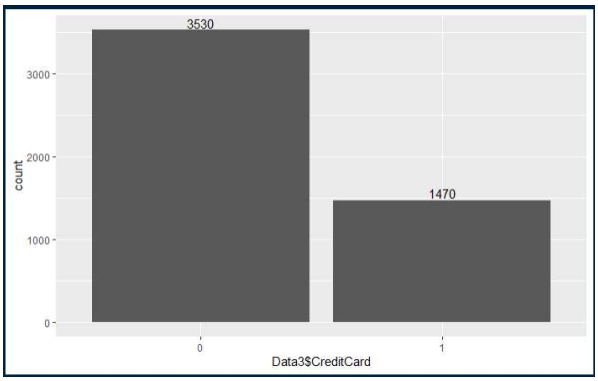


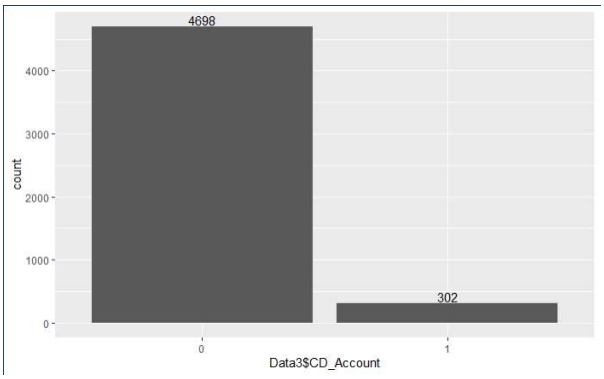


- Except for education and Online variables, all the other binary variables have class imbalance.





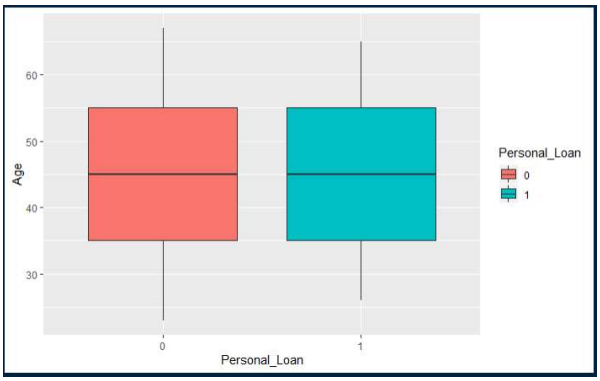


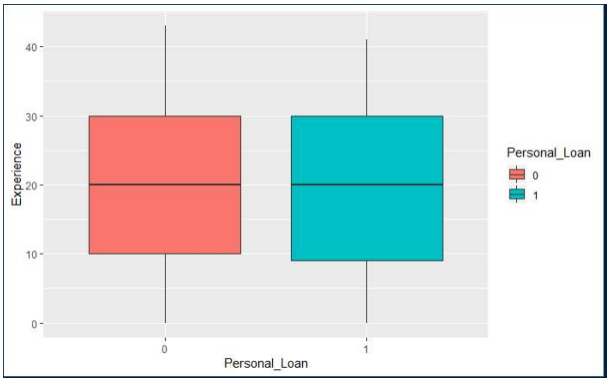


4.2 Bivariate Analysis

Age and experience don't differentiate the personal loan affinity class much as implicit from the box plot below. This is in line with the general notion that experience increases with age.

The correlation between age and experience is almost 1. We can safely assume even before the modelling that these are weak predictors of personal loan affinity

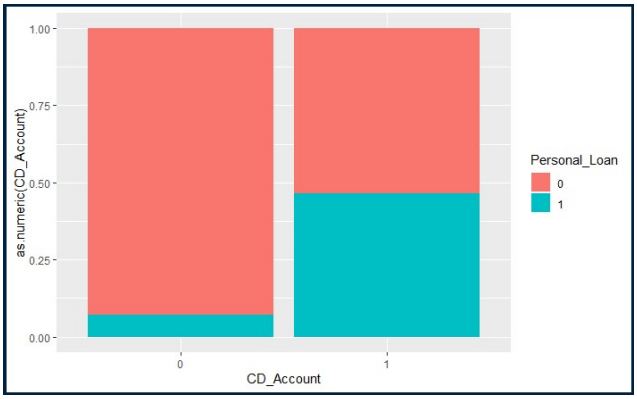


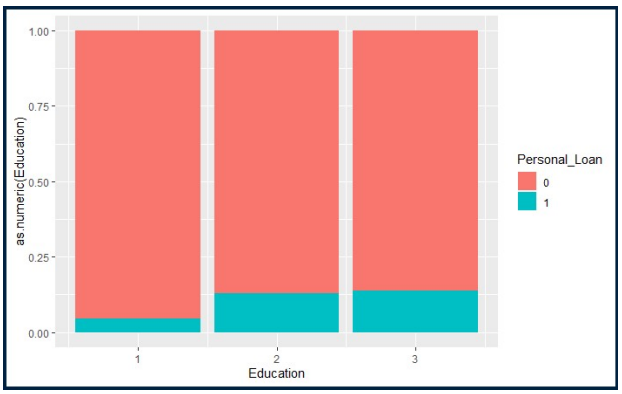


Income, number of family members and average spend on credit card differentiates the Personal loan affinity class by a large extent as we can see from the boxplots below. This will subsequently be reflected in the variable importance table after the model building as we would expect

Amongst the categorical variables, except for CD account where those who hold a CD account with the bank have higher proportion opting for the personal loan, other categorical variables don't have much impact on the personal loan affinity as one can understand from the below charts

Also, there is slightly higher proportion of people amongst graduates and advanced professional who have opted for a personal loan than amongst undergrads





4.3 Model Building

- The data is first split into train and test in the ratio 80:20 in order to build the model on the train data and test the model performance on the test data
- We build a fully-grown tree on the train data (4000 samples) with CP=-1, Minsplit = 2 and minbucket =1, criteria.

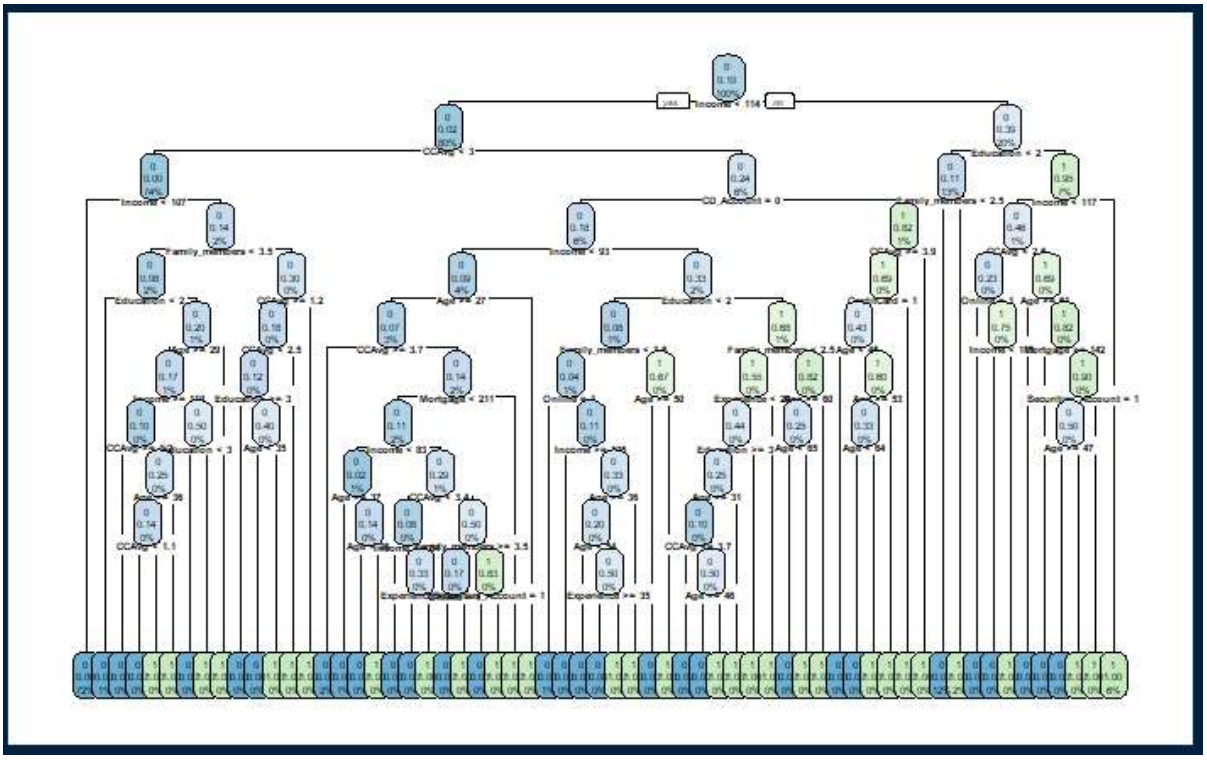
- We get an accuracy, sensitivity and specificity of each 1 as expected from the un-pruned tree. The tree diagram, confusion matrix and the performance measures after doing a prediction using the model in train and test data are given below:

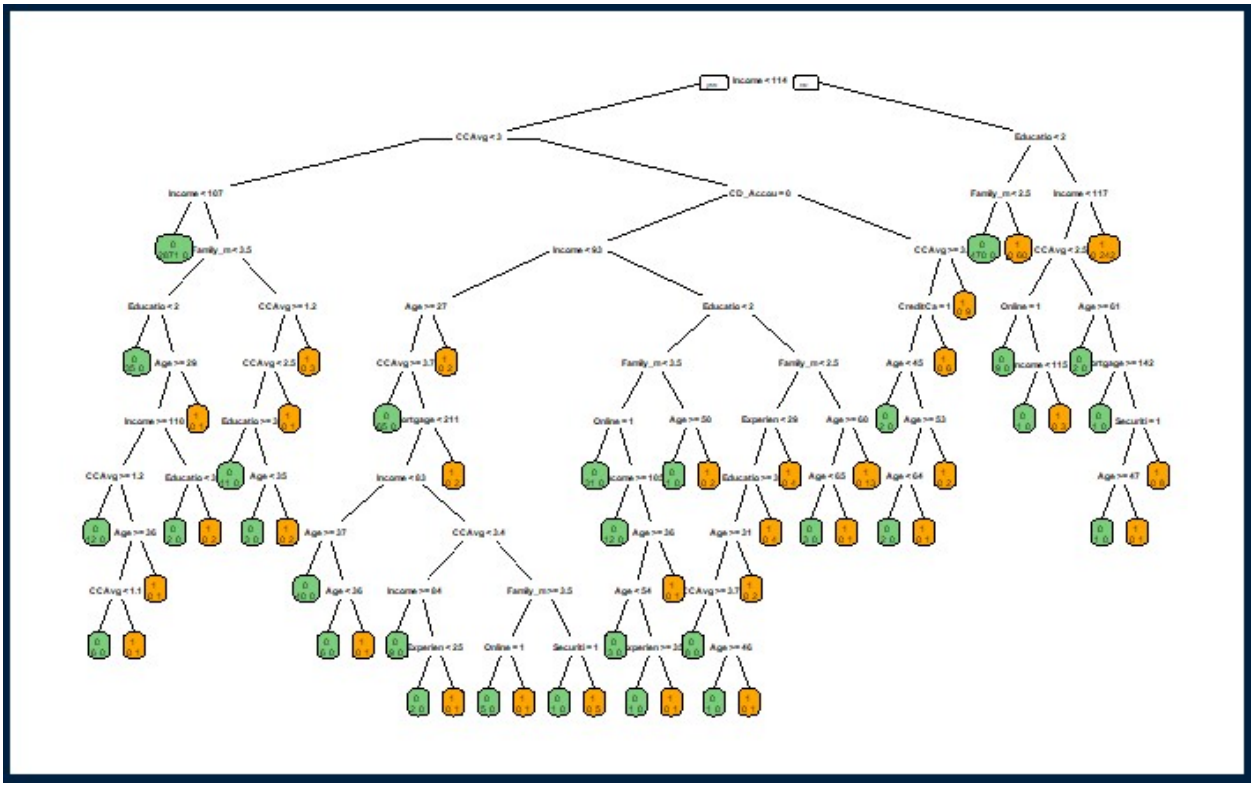
Fully grown train tree model measures

```
                predict.class
Personal_Loan    0    1
            0 3616    0
            1    0  384
[1] 1
[1] 1
[1] 1
```

Fully grown test tree model measures

```
                predict.class
Personal_Loan    0    1
            0 896    8
            1   6   90
[1] 0.986
[1] 0.9375
[1] 0.9911504
```



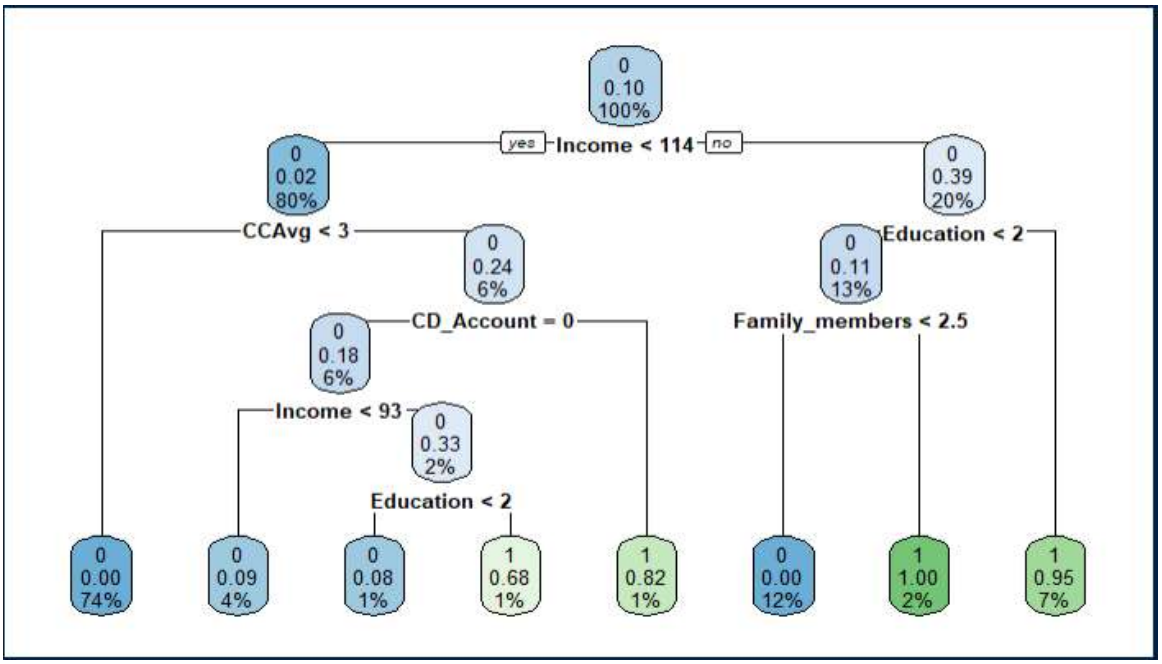


- We get a high Accuracy, sensitivity and specificity values for the fully-grown tree as given below in the table
- Income, CCAvg and Education are the top differentiating variables for prediction as we could see from the exploratory data analysis done in the initial stage.

	ACC <dbl>	SENS <dbl>	SPEC <dbl>
tree_full_train	1.000	1.0000	1.0000000
tree_full_test	0.986	0.9375	0.9911504

4.4 Model performance with Manual pruning

Minsplit = 50, minbucket = 15



Manually pruned train tree model measures

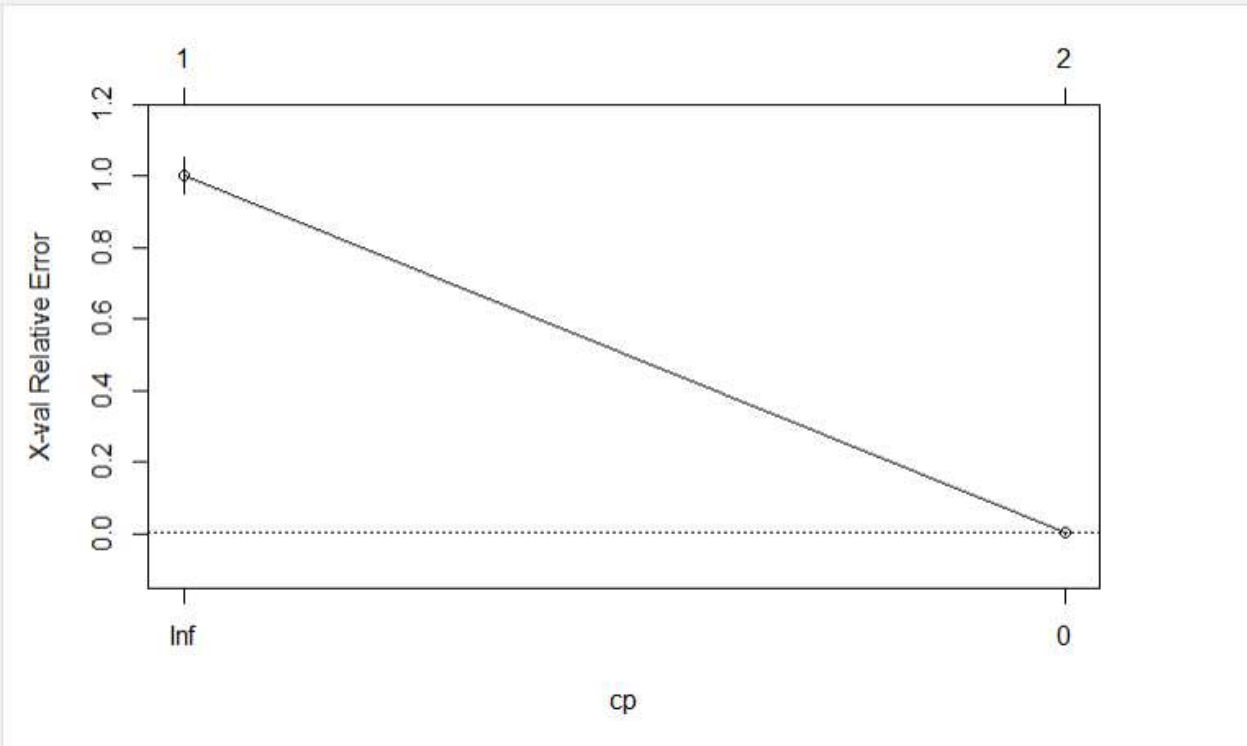
```
             predict.class
Personal_Loan    0    1
            0 3586   30
            1   27  357
[1] 0.98575
[1] 0.9296875
[1] 0.9917035
```

Manually pruned test tree model measures

```
             predict.class
Personal_Loan   0   1
            0 897   7
            1   8  88
[1] 0.985
[1] 0.9166667
[1] 0.9922566
```

Comparison of model performance

	ACC <dbl>	SENS <dbl>	SPEC <dbl>
tree_full_train	1.00000	1.0000000	1.0000000
tree_full_test	0.99000	0.9166667	0.9977876
MPruned tree_train	0.98575	0.9296875	0.9917035
Mpruned tree_test	0.98500	0.9166667	0.9922566



4.5 Model with the CP criteria

A simultaneous model is built with the best Complexity parameter obtained and the model performance is obtained as below

Best CP train tree model measures

```
             predict.class
Personal_Loan    0    1
            0 3611    5
            1   10  374
[1] 0.99625
[1] 0.9739583
[1] 0.9986173
```

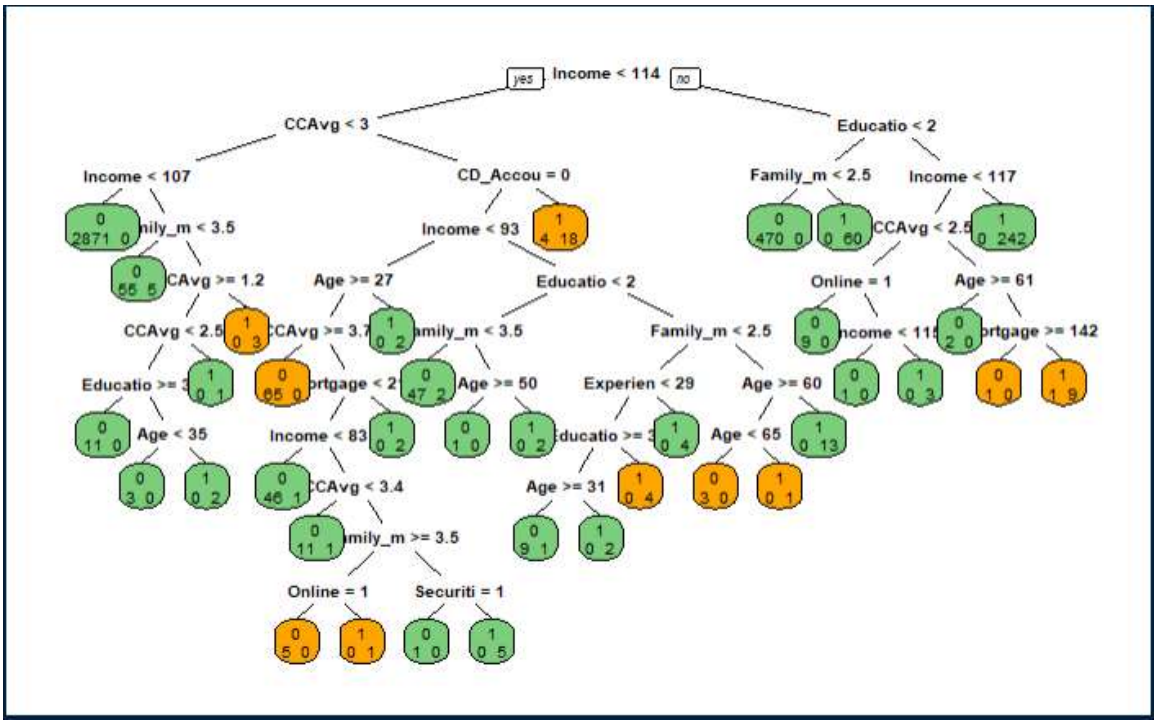
Best CP test tree model measures

```
             predict.class
Personal_Loan    0    1
            0 898    6
            1   8   88
[1] 0.986
[1] 0.9166667
[1] 0.9933628
```

- We then compare the overall performance measures of the 3 models in the below table to identify which one gives best results.

	ACC <dbl>	SENS <dbl>	SPEC <dbl>
tree_full_train	1.00	1.00	1.00
tree_full_test	0.99	0.94	0.99
MPruned tree_train	0.99	0.93	0.99
MPruned tree_test	0.98	0.92	0.99
Best CP_tree_train	1.00	0.97	1.00
Best CP_ptree_test	0.99	0.92	0.99

- From the table above, we can infer that there is not much difference between different models in terms of performance and the predictor variables are fully able to explain the variances in the probability of a customer availing personal loan.
- We have very high accuracy, sensitivity and specificity for the model in all the 3 cases we did
- The best tree can be plotted as below



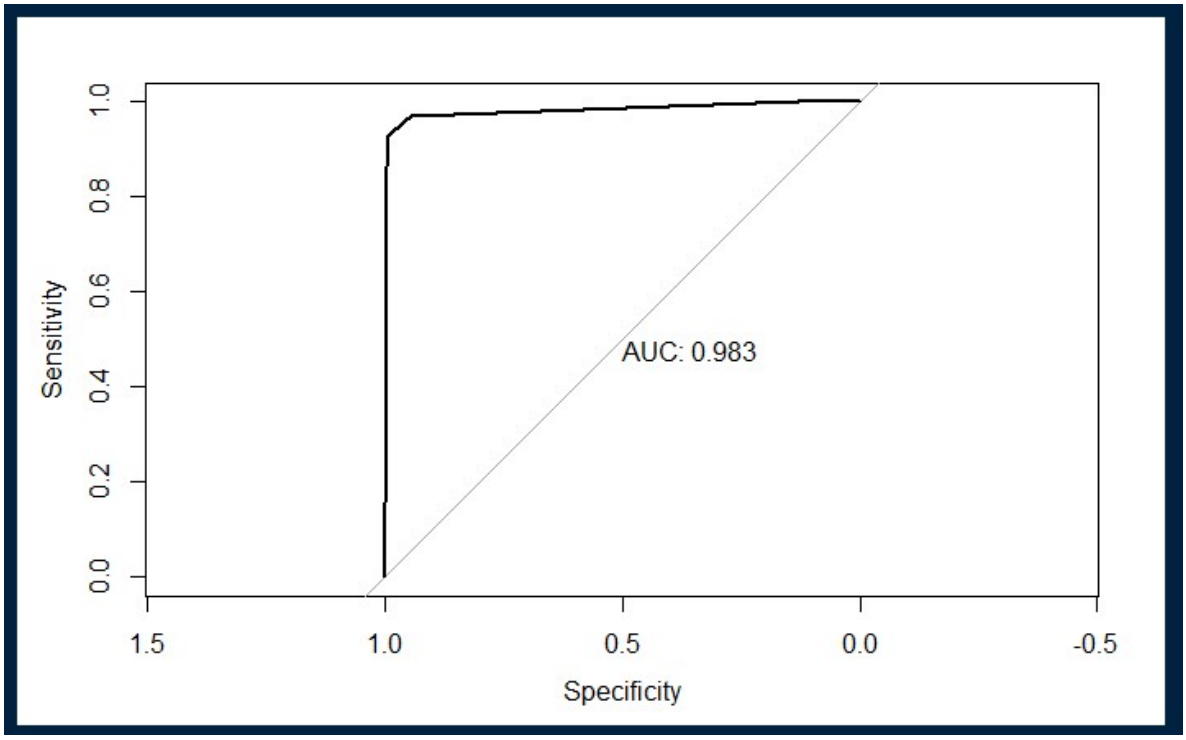
4.6 Variable importance table

- We run a variable importance function to see the order of variable importance in the model
- A we had inferred from the initial EDA, Education the demographic numeric variables are turning out to be most important predictor variables in the model
- The relationship oriented binary variables do not impact the dependent variable much as the table below suggests

	round.ptree.variable.importance..2. <dbl>
Education	271.50
Income	214.76
Family_members	182.81
CCAvg	111.01
CD_Account	53.63
Mortgage	28.73
Age	28.12
Experience	16.76
Online	6.56
Securities_Account	2.92

4.7 ROC Curve & AUC

- We obtain a high ROC for the model which is equal to 98.3%
- This essentially tells that 98.3% of the time, the model separates the positive class from the negative class accurately



Inferences

- The CART model we built obtain high performance and the variables in the model can accurately predict the outcome variable with high precision and sensitivity.
- The model separates the positive class from the negative class almost every time we run the model (AUC = 98.3%)
- The model also tells us that, since there is a high AUC value, we can infer that there is high amount of separation between the predicted classes as the inequality between these

  GINI = 2AUC- 1 = (2*98.3 – 1) = 96.6

- The model is highly reliable to predict and target the sample where there is a high probability of availing the model as we infer from the model performance measures
- We do a further modelling using the random forest algorithm to check which algorithm gives us the best possible results

5. Model using the Random Forest algorithm

5.1 Steps to building random forest

1. Read the file
2. Feature engineering and exploratory data analysis (Already done and explained in the beginning)
3. Check the class imbalance
4. Split into train and test data
5. Build a random forest model
6. Predict for train and test dataset
7. Derive and analyze the performance parameters
8. Plot the variable table
9. Compare the results with the CART model

In the feature engineering part, after loading the file into R, we do the following steps before running the random forest function in the data

- *Converting the categorical variables into factors*
- *Imputing the missing values in the Family members column by kNN imputation method*
- *Setting all the negative values in the experience column to zero (Reasoning explained in the beginning)*

Split the data into train and test in the same ratio as we did for the CART methodology (80:20)

Deciding on the value of ntree (Number of trees to build) and mtry (Number of variables to consider at a time)

- Setting ntree = 1500(Roughly 40% of the train data set) and mtry = 4 initially(sqrt(12)), for 12 predictor variables in the model
- We obtain the following output

```
                Number of trees: 1500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 1.27%
Confusion matrix:
     0   1 class.error
0 3609   7 0.001935841
1   44 340 0.114583333
```

- We obtain OOB error rate of 1.27%
- Now we build another model changing the cut off value of the probability.So, we set ntree = 1500 and mtry =3 for further analysis, with changing the cutoff value

5.2 Cutoff decision

- Regarding this problem, the most important thing here is to decide the cutoff probability, above which we target a customer with a personal loan campaign
- Here since we don't have the costs associated with the campaign/per customer, we assume that the cost of losing a customer should be reduced, which effectively means that we should reduce the false negatives to the maximum extent
- Since the aim of the bank is to maximize their business through personal loan, they shouldn't lose any customer by not approaching them with the campaign
- So, after multiple iterations in the model, we take .25 as the cutoff value below which a customer will not be approached with a personal loan campaign.
- This is the value at which the specificity of the model is 1 and there are no false negatives for the current train and test data set. The model results are as given below:

ntree = 1500, mtry = 3, cutoff = (0.25,0.75)

```
        OOB estimate of  error rate: 2.3%
Confusion matrix:
     0   1 class.error
0 3616   0   0.0000000
1   92 292   0.2395833
```

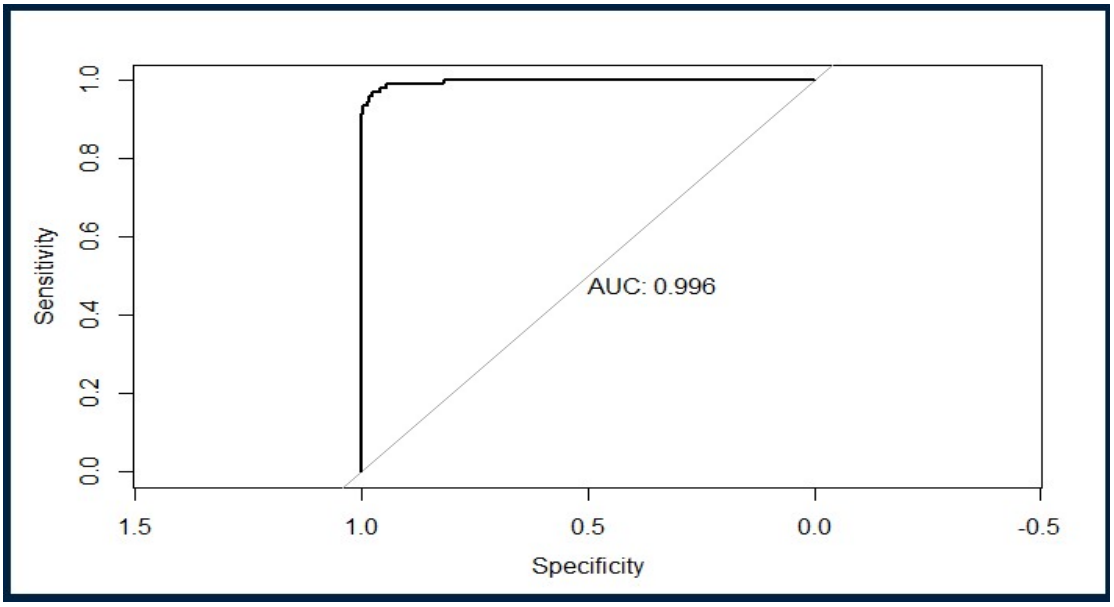
- The classification error rate is increased here because we have a higher sensitivity in the model when setting the cutoff probability to .25
- But we can let that go as we are trying to maximize the specificity by reducing the number of false positives

5.3 Train and Test the model

After training and testing the model with the above values we obtain a below table of performance measures

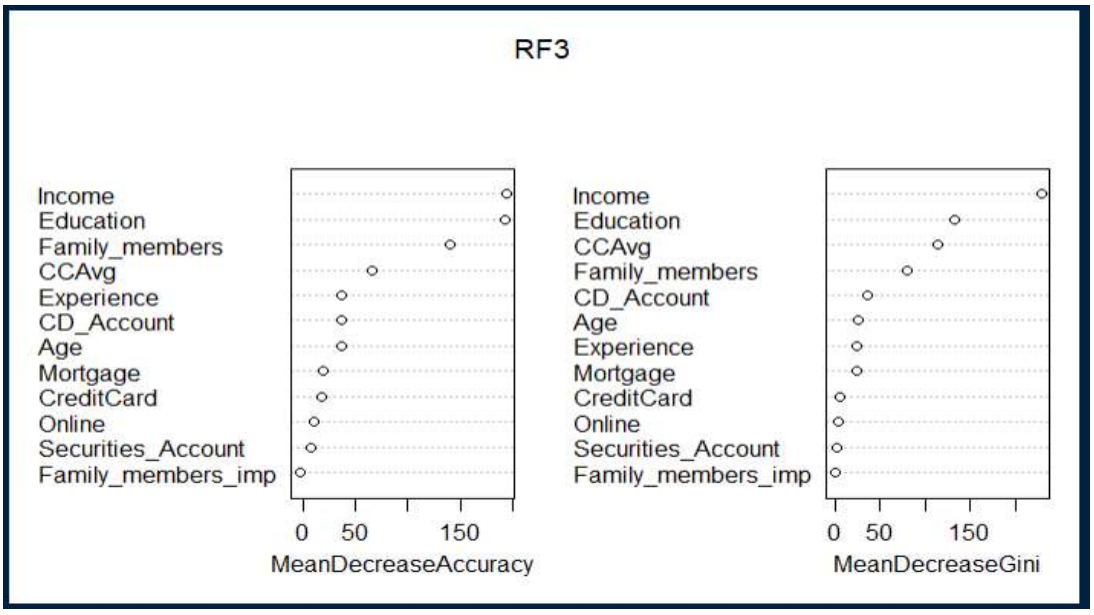
	ACC <dbl>	SENS <dbl>	SPEC <dbl>
tree_full_train	0.99275	0.9244792	1
tree_full_test	0.97900	0.7812500	1

5.4 ROC Curve and AUC value



- We obtain a very high AUC value of 0.99 with the model, which is the best AUC value we obtained out of all the models so far

5.5 Variable Importance Table



- We obtain similar variable importance to the CART model which validates the model as well

5.6 Inferences from the model

- From the model using the random forest algorithm, we built a model after several iterations of ntree, mtree and cutoff probability values.
- The model gives high accuracy and specificity but moderate sensitivity when setting the cut off to .25
- There is a drop in sensitivity when predicting in the test data, but accuracy and specificity remains high.
- We obtain the best AUC and specificity values out of all the models we built using the Random forest model with above mentioned parameters

5.7 Selection of model

- As a consultant to the bank, we select the Random Forest model with cut off value of 0.25 because of the high specificity and AUC values. Essentially this means that when making a prediction using this model, those customers who fall in the probability above .25 prediction of the predicted personal loan dependent variable value will be targeted for the campaign
- If we had a cost matrix of False negatives and False positives, based on the previous campaign, we could have ideally calculated the total cost of misclassification and arrived at a final cutoff value
- We are assuming that he cost of missing a probable customer is way too higher than those the cost of campaign/customer
- Based on the above assumption, we have tried to nullify the false negatives in the model