Project – Personal Loan Campaign

Table of Content

1. Project Objective

The objective of this project work is building a Data Mining Model to predict whether the costumer will take the loan or not or will he/she responds to the loan campaign. The methods involved during the designing of the aforesaid model are CART (Classification and Regression Tree) and Random Forest. Various performance measures such as Rank Order Table, KS, AUC, GINI, Confusion Matrix, Concordance and Discordance are utilized to predict the accuracy of the build models.

2. Exploratory Data Analysis

- Exploratory data analysis tells us that the data set contain 39 independent variables and TARGET variable which is the dependent variable.
- The data set does not contain any missing value or NA values
- Summary of the data set is shown in Fig.1 below:

```
   CUST_ID          TARGET            AGE        GENDER       BALANCE          OCCUPATION    AGE_BKT          SCR
C1     :    1   Min.   :0.0000   Min.   :21.00   F: 5433   Min.   :      0   PROF    :5417   <25  :1753   Min.   :100.0
C10    :    1   1st Qu.:0.0000   1st Qu.:30.00   M:14376   1st Qu.:  64754   SAL     :5855   >50  :3035   1st Qu.:227.0
C100   :    1   Median :0.0000   Median :38.00   O:  191   Median : 231676   SELF-EMP:3568   26-30:3434   Median :364.0
C1000  :    1   Mean   :0.1256   Mean   :38.42             Mean   : 511362   SENP    :5160   31-35:3404   Mean   :440.2
C10000 :    1   3rd Qu.:0.0000   3rd Qu.:46.00             3rd Qu.: 653877                   36-40:2814   3rd Qu.:644.0
C10001 :    1   Max.   :1.0000   Max.   :55.00             Max.   :8360431                   41-45:3067   Max.   :999.0
(Other):19994                                                                                 46-50:2493
HOLDING_PERIOD  ACC_TYPE       ACC_OP_DATE    LEN_OF_RLTN_IN_MNTH NO_OF_L_CR_TXNS NO_OF_L_DR_TXNS TOT_NO_OF_L_TXNS
Min.   : 1.00   CA: 4241   11/16/2010:   24   Min.   : 29.0       Min.   : 0.00   Min.   : 0.000   Min.   :  0.00
1st Qu.: 7.00   SA:15759   04-03-09  :   23   1st Qu.: 79.0       1st Qu.: 6.00   1st Qu.: 2.000   1st Qu.:  9.00
Median :15.00              7/25/2010 :   22   Median :125.0       Median :10.00   Median : 5.000   Median : 14.00
Mean   :14.96              05-06-13  :   21   Mean   :125.2       Mean   :12.35   Mean   : 6.634   Mean   : 18.98
3rd Qu.:22.00              02-07-07  :   20   3rd Qu.:172.0       3rd Qu.:14.00   3rd Qu.: 7.000   3rd Qu.: 21.00
Max.   :31.00              8/24/2010 :   20   Max.   :221.0       Max.   :75.00   Max.   :74.000   Max.   :149.00
                           (Other)   :19870
NO_OF_BR_CSH_WDL_DR_TXNS NO_OF_ATM_DR_TXNS NO_OF_NET_DR_TXNS NO_OF_MOB_DR_TXNS NO_OF_CHQ_DR_TXNS   FLG_HAS_CC       AMT_ATM_DR
Min.   : 0.000           Min.   : 0.000    Min.   : 0.000    Min.   : 0.0000   Min.   : 0.000    Min.   :0.0000   Min.   :     0
1st Qu.: 1.000           1st Qu.: 0.000    1st Qu.: 0.000    1st Qu.: 0.0000   1st Qu.: 0.000    1st Qu.:0.0000   1st Qu.:     0
Median : 1.000           Median : 1.000    Median : 0.000    Median : 0.0000   Median : 2.000    Median :0.0000   Median :  6900
Mean   : 1.883           Mean   : 1.029    Mean   : 1.172    Mean   : 0.4118   Mean   : 2.138    Mean   :0.3054   Mean   : 10990
3rd Qu.: 2.000           3rd Qu.: 1.000    3rd Qu.: 1.000    3rd Qu.: 0.0000   3rd Qu.: 4.000    3rd Qu.:1.0000   3rd Qu.: 15800
Max.   :15.000           Max.   :25.000    Max.   :22.000    Max.   :25.0000   Max.   :15.000    Max.   :1.0000   Max.   :199300

AMT_BR_CSH_WDL_DR   AMT_CHQ_DR        AMT_NET_DR        AMT_MOB_DR        AMT_L_DR          FLG_HAS_ANY_CHGS AMT_OTH_BK_ATM_USG_CHGS
Min.   :      0     Min.   :      0   Min.   :      0   Min.   :     0    Min.   :      0    Min.   :0.0000   Min.   :  0.000
1st Qu.:   2990     1st Qu.:      0   1st Qu.:      0   1st Qu.:     0    1st Qu.: 237936    1st Qu.:0.0000   1st Qu.:  0.000
Median :340150     Median :  23840   Median :      0   Median :     0    Median : 695115    Median :0.0000   Median :  0.000
Mean   :378475     Mean   : 124520   Mean   :237308    Mean   : 22425    Mean   : 773717    Mean   :0.1106   Mean   :  1.099
3rd Qu.:674675     3rd Qu.:  72470   3rd Qu.:473971    3rd Qu.:     0    3rd Qu.:1078927    3rd Qu.:0.0000   3rd Qu.:  0.000
Max.   :999930     Max.   :4928640   Max.   :999854    Max.   :199667    Max.   :6514921    Max.   :1.0000   Max.   :250.000

AMT_MIN_BAL_NMC_CHGS NO_OF_IW_CHQ_BNC_TXNS NO_OF_OW_CHQ_BNC_TXNS AVG_AMT_PER_ATM_TXN AVG_AMT_PER_CSH_WDL_TXN AVG_AMT_PER_CHQ_TXN
Min.   :  0.000      Min.   :0.00000       Min.   :0.0000        Min.   :    0       Min.   :     0          Min.   :     0
1st Qu.:  0.000      1st Qu.:0.00000       1st Qu.:0.0000        1st Qu.:    0       1st Qu.:  1266          1st Qu.:     0
Median :  0.000      Median :0.00000       Median :0.0000        Median : 6000       Median :147095          Median :  8645
Mean   :  1.292      Mean   :0.04275       Mean   :0.0444        Mean   : 7409       Mean   :242237          Mean   : 25093
3rd Qu.:  0.000      3rd Qu.:0.00000       3rd Qu.:0.0000        3rd Qu.:13500       3rd Qu.:385000          3rd Qu.: 28605
Max.   :170.000      Max.   :2.00000       Max.   :2.0000        Max.   :25000       Max.   :999640          Max.   :537842

AVG_AMT_PER_NET_TXN AVG_AMT_PER_MOB_TXN FLG_HAS_NOMINEE  FLG_HAS_OLD_LOAN     random
Min.   :     0      Min.   :     0      Min.   :0.0000   Min.   :0.0000   Min.   :0.0000114
1st Qu.:     0      1st Qu.:     0      1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.2481866
Median :     0      Median :     0      Median :1.0000   Median :0.0000   Median :0.5061214
Mean   :179059      Mean   : 20304      Mean   :0.9012   Mean   :0.4929   Mean   :0.5019330
3rd Qu.:257699      3rd Qu.:     0      3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.7535712
Max.   :999854      Max.   :199667      Max.   :1.0000   Max.   :1.0000   Max.   :0.9999471
```

Fig.1: Summary of the data set

- Plot of Gender shows that there are more number of male candidates than the female candidates (Fig.2)

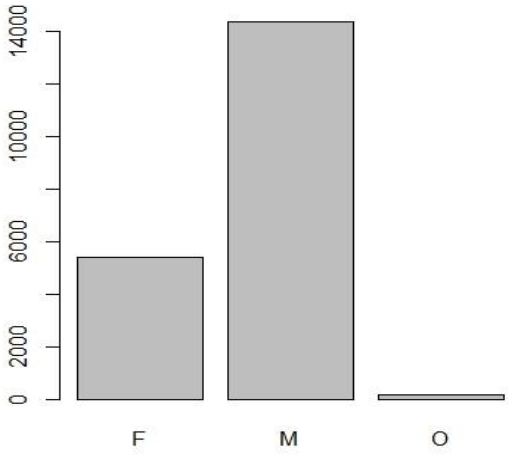

Fig.2: Gender

- Plot of account type shows that most of the accounts in the bank are savings account (Fig.3).

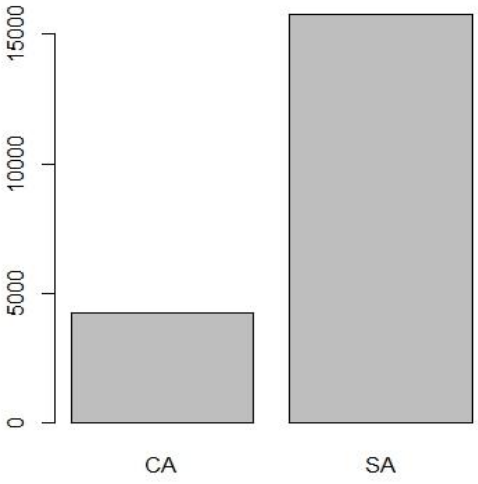

Fig.3: Account type

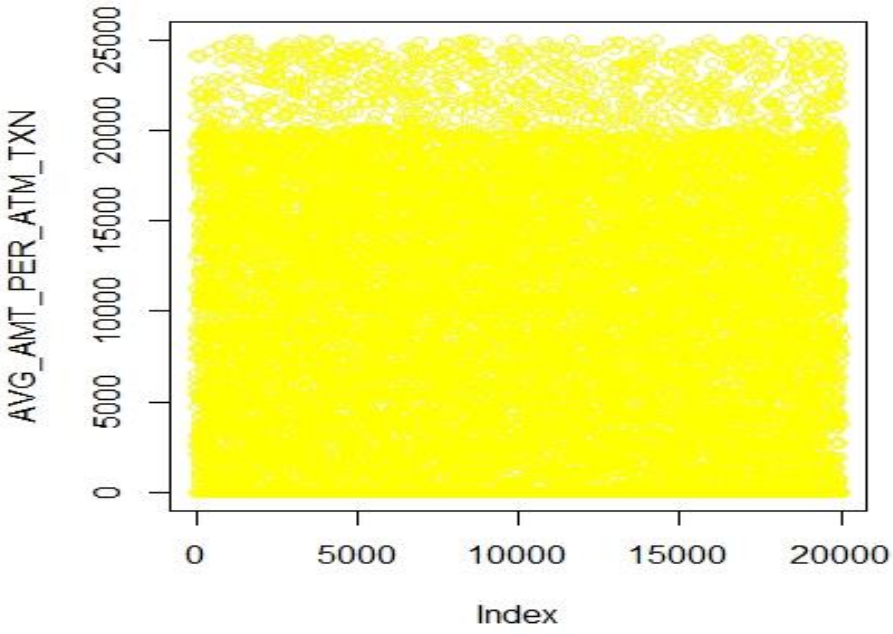

Fig.4: Average amount per ATM transaction

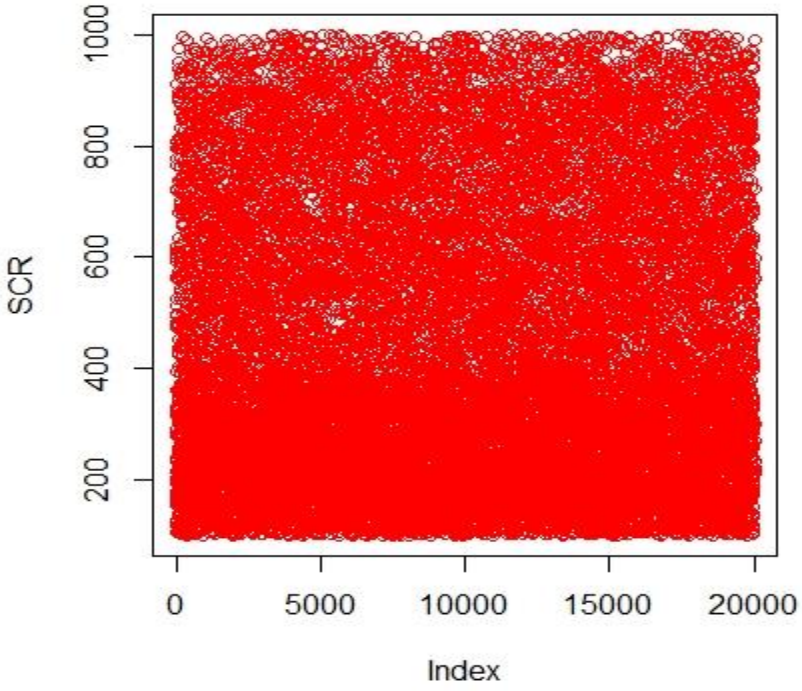

Fig.5: Generic Marketing Score

3. Clustering Analysis

Since the data set given is a large data set containing about 20,000 observations therefore we are choosing **kmeans clustering** to group the similar object into one particular group. The kmeans clustering, groups the similar objects by calculating the Euclidean distance of each variable with respect to the centroid. This process continues and continuously the centroids are updated until all the similar variables are grouped together.

While building the kmeans clustering model, only the most important 13 variables are considered like AGE, BALANCE, SCR, HOLDING PERIOD, LEN_OF_RLTN_IN_MNTH, TOT_NO_OF_L_TXNS, FLG_HAS_CC, AMT_L_DR, AVG_AMT_PER_ATM_TXN, AVG_AMT_PER_CSH_WDL_TXN, AVG_AMT_PER_CHQ_TXN, AVG_AMT_PER_NET_TXN, AVG_AMT_PER_MOB_TXN.

The Fig. below shows the elbow diagram and kmeans clustering plot. From the elbow diagram it is concluded that the **right number of clusters required for kmeans clustering is 6.**

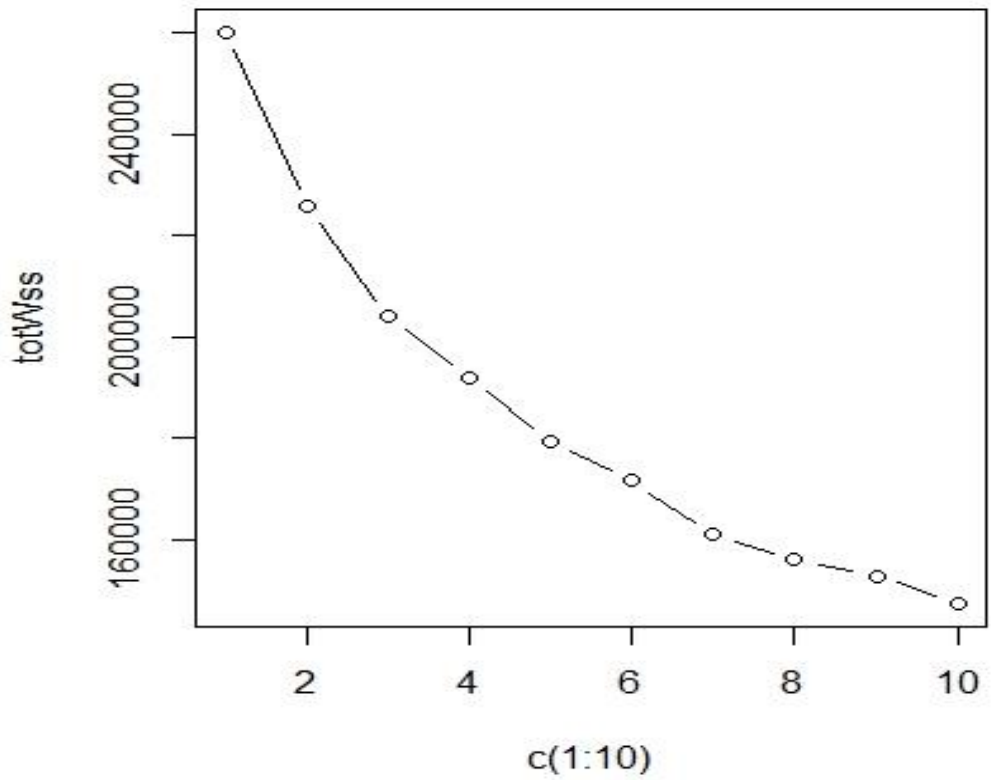

Fig.6: Elbow plot

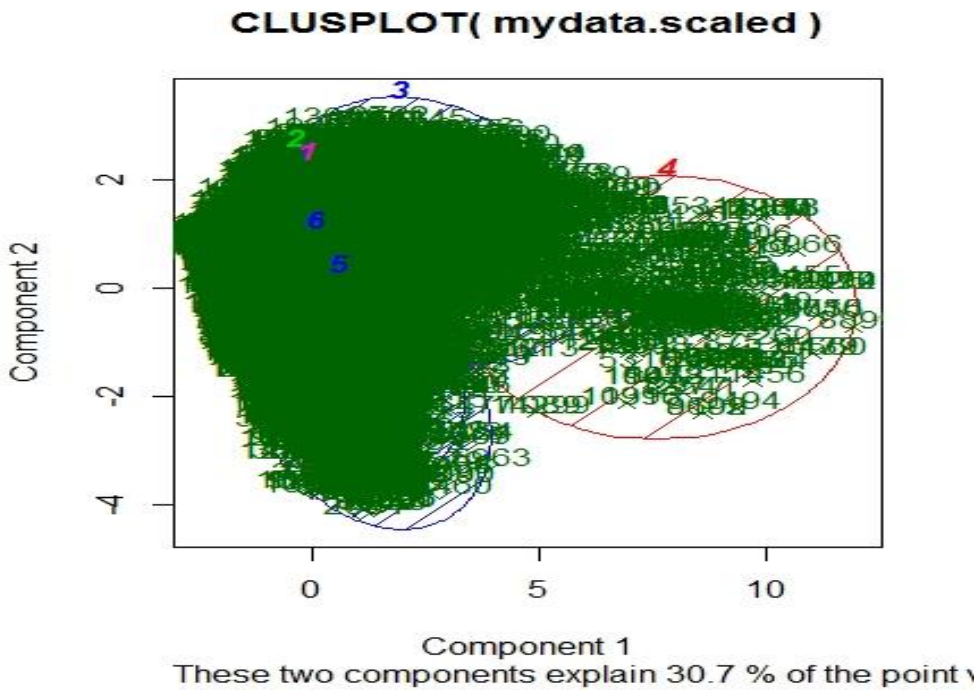

Fig.7: kmeans clustering

- Cluster-1 has total 3827 elements.
- Cluster-2 has total 5918 elements.
- Cluster-3 has total 3031 elements.
- Cluster-4 has total 302 elements.
- Cluster-5 has total 3673 elements.
- Cluster-6 has total 3249 elements.
- Ratio of between_ss and total_ss is 34%

4. CART (Clustering and Regression Tree) analysis

- In order to build the Decision tree, the data set provided is first split into train dataset and test data set.
- To build the Decision tree, the most important 13 variables are considered viz. AGE, BALANCE, SCR, HOLDING PERIOD, LEN_OF_RLTN_IN_MNTH, TOT_NO_OF_L_TXNS, FLG_HAS_CC,AMT_L_DR,AVG_AMT_PER_ATM_TXN,AVG_AMT_PER_CSH_WDL_TXN,AVG_AMT_PER_CHQ_TXN,AVG_AMT_PER_NET_TXN,AVG_AMT_PER_MOB_TXN.
- Train data set consist of 70 percent of the observations and is used to learn the CART model.
- Test data set contains 30 percent of the data and it is used to test the performance the build model.
- CART model was built on both the train data set and test data set.
- The relative error was observed for each of the build model and the complexity parameter was chosen accordingly where the relative error starts increasing.
- For train data set, the complexity parameter was found to be 0.00025.

- For test data set, the complexity parameter was found to be 0.0025.
- Pruning of the decision trees was done on the basis of the above computed complexity parameter for train and test data set.
- CART tree before pruning is show in Fig.8

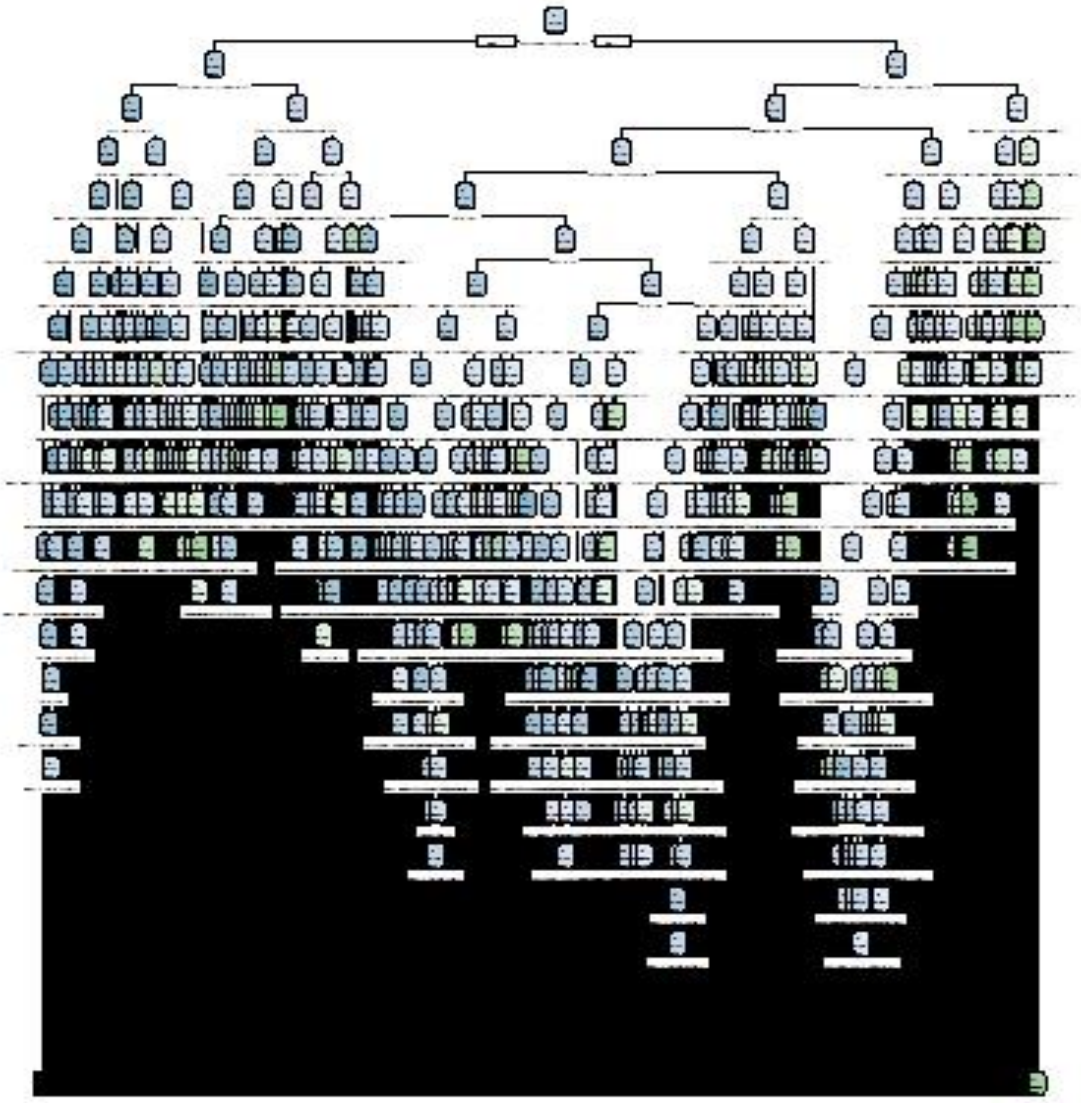

Fig.8: CART tree before pruning

- CART tree after pruning is shown in Fig.9

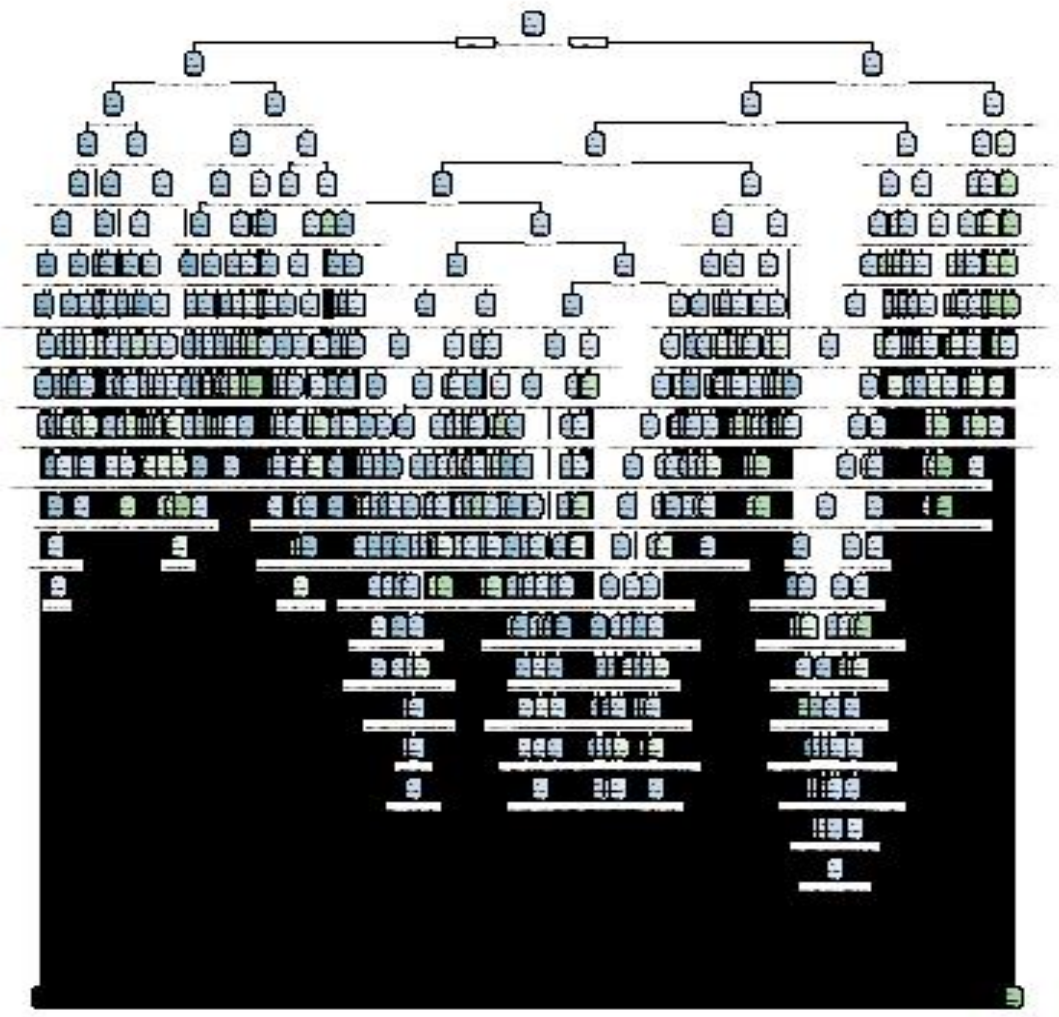

Fig.9: CART tree after pruning

- Various model performance measures were adopted to calculate the accuracy of the CART model.
- Model Performance Measures for **train data set** are as follows:
    - Accuracy as computed from confusion matrix is 96.67% (Table-1).

      Table-1: Confusion Matrix

	0	1
0	12048	194
1	271	1487

    - The maximum value of KS in this case was found to be 74.9% corresponding to the decile (0.0294, 0.0625]. The cumulative response rate corresponding to this decile was

found to be 94.71% meaning if we target this decile, 94.71% of the customers should respond to the loan.

- ROC plot for train data set is shown in Fig.10

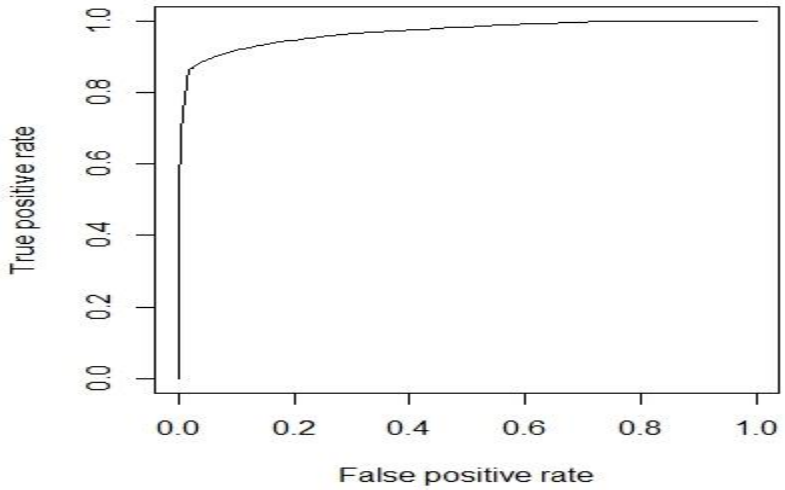

Fig.10: ROC plot for train data set

- The KS value for train data set is found to be 84.67% if the right hand values of the deciles is not included.
- Area under the curve was found to be 96.94%.
- Gini index was found to be 82.1%.
- Concordance is 96.78% meaning approximately 97% of the pairs have clear separation between the probabilities of TARGET=1 and TARGET=0.
- Discordance = 3.21%

- Model Performance Measures for **test data set** are as follows:
  - Accuracy as computed from confusion matrix is 92.13% (Table-2)

    Table-2: Confusion Matrix

	0	1
0	5209	37
1	435	319

  - The maximum value of KS in this case was found to be 45.54% corresponding to the decile (0.122, 0.179]. The cumulative response rate corresponding to this decile was found to be 59.02%% meaning if we target this decile, 59% of the customers should respond to the loan.
  - ROC plot for test data set is shown in Fig.11

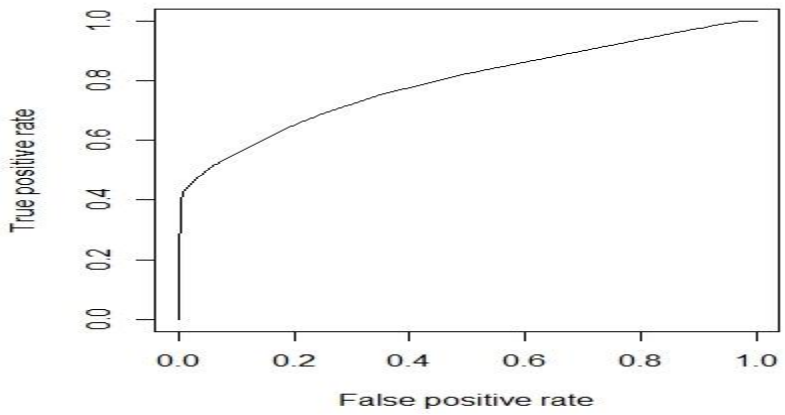

Fig.11: ROC plot for test data set

- The KS value for test data set is found to be 45.65% if right hand value of the deciles is not included.
- Area under the curve was found to be 79.22%.
- Gini index was measured to be 51.11%.
- Concordance is 75.72% meaning that approximately 76% of the pairs have clear separation between the probabilities of TARGET=1 and TARGET=0.
- Discordance = 24.27%

5. Random Forest Model

- The train is used to build the **Random Forest model** on 13 most important variables which are as follows: AGE, BALANCE, SCR, HOLDING PERIOD, LEN_OF_RLTN_IN_MNTH,TOT_NO_OF_L_TXNS,FLG_HAS_CC,AMT_L_DR,AVG_AMT_PER_ATM_TXN,AVG_AMT_PER_CSH_WDL_TXN,AVG_AMT_PER_CHQ_TXN,AVG_AMT_PER_NET_TXN,AVG_AMT_PER_MOB_TXN.
- Performance of the model is tested on the test data set.
- Train data set contains only 12% of the observations which has TARGET = 1.
- The Random Forest model is build for total number of trees to be equal to 501. The odd number of trees is selected in order that the majority rule satisfies.
- The Out of Bag estimate error was found to be 5.84% meaning that the Random Forest model build has a prediction error of approximately 6%.
- Confusion matrix is as shown (Table-3):

Table-3: Confusion Matrix

	0	1	Class.error
0	12231	11	0.000898546
1	806	952	0.458475540

- Importance suggests that the variables FLG_HAS_CC and ANG_AMT_PER_MOB_TXN are less important as compared to the variables of the model.

- The plot of the Random Forest is shown in Fig.12 from which it is conclude that the relative error decreases approximately up to 51 trees and then the relative error becomes almost constant.

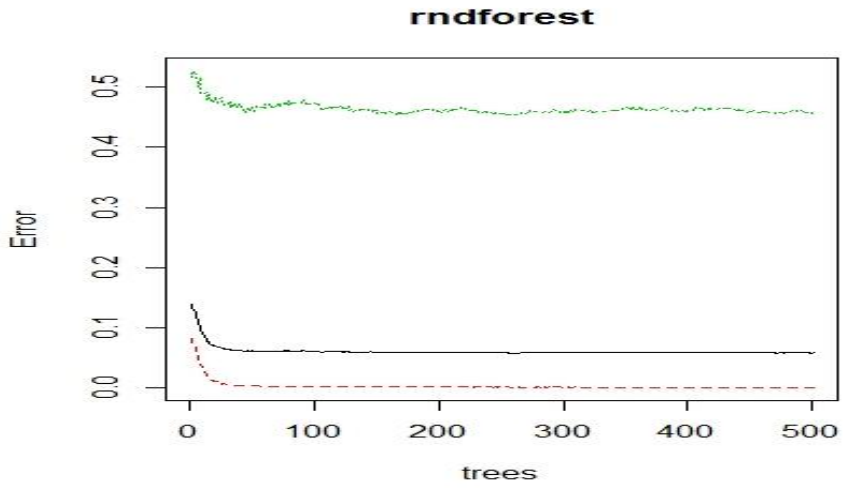

Fig.12: Random Forest number of trees

- Hence, the tuned Random Forest Model was tuned to 51 trees with the step factor of 1.5.
- From the tuned Random Forest model, it is observed that the Out of Bag estimate of error is reduced to 3.79% if we consider three variables for the building the model (Fig.13).

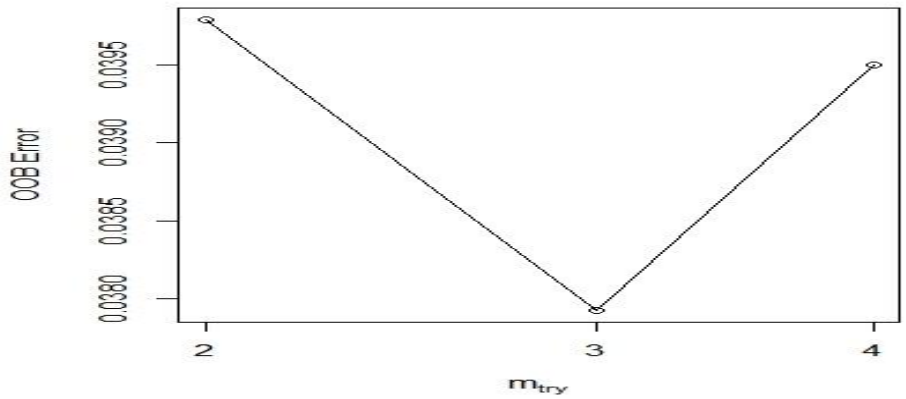

Fig.13: Tuned Random Forest (Number of variables)

- The probability of predicting whether the TARGET will respond or not is calculated.
- The prediction of error was found to be 0% on the train data set (Table-4).

Table-4: Confusion Matrix

	0	1
0	12242	0
1	0	1758

- However, when the same model is tested on the test data set, the error of mis-classification is found to be 3.51% (Table-5)

Table-5: Confusion Matrix

	0	1
0	5244	2
1	209	545

- Accuracy on test data set = 96.48%
- From Rank Order Table created for **train data set**, the maximum value of KS is found to be 92% against the decile of (0.066, 0.722] and the cumulative response rate is 100%.
- From test data set Rank Order Table, the maximum value of KS is found to be 64.55% against the decile of (0.066, 0.722] and the cumulative response rate is observed as 96% meaning if we target this decile there is a possibility that 96% of the costumers will respond to the loan or loan campaign initiated by the bank.
- The ROC curve for train data set is shown in Fig.14.

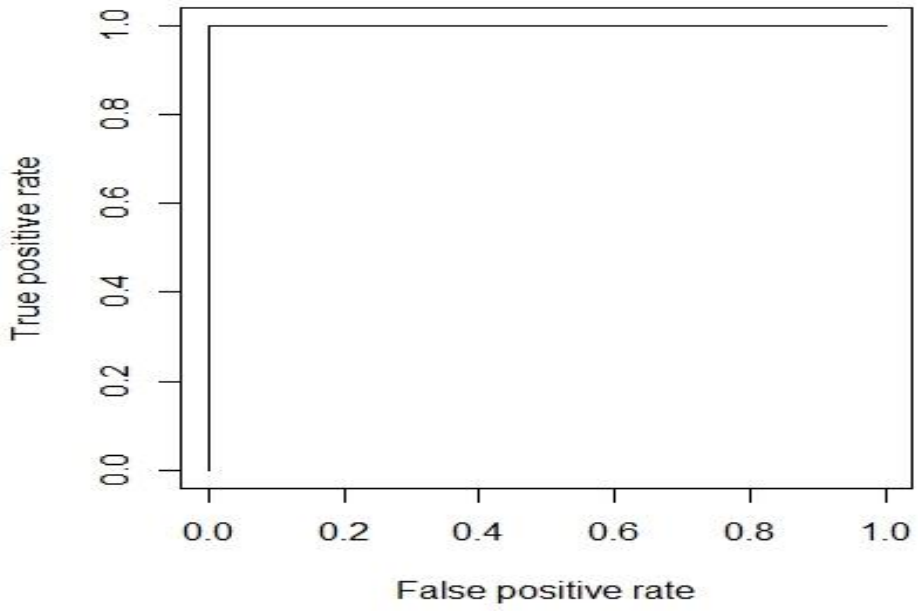

Fig.14: ROC curve for Random Forest on train data set

- The ROC curve for test data set is shown in Fig.15.

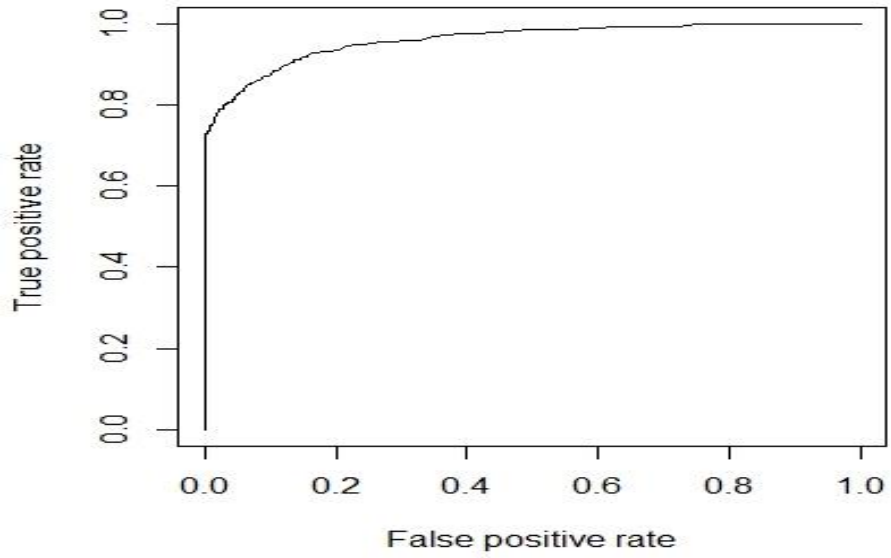

Fig.15: ROC curve for Random Forest on test data set

- The KS value calculated from Rank Order Table is obtained when the right hand is included while creating the deciles. If the right hand value is excluded then the KS value is found to be 100% for train data set.
- The KS value for test data set if right hand value is excluded was calculated as 78.36%
- For train data set, AUC = 1.
- For test data set, AUC = 96.12%
- For train data set:
    - Concordance is 1, meaning that all of the pairs have clear separation between the probabilities of TARGET=1 and TARGET=0.
    - Discordance = 0
- For test data set:
    - Concordance is 96.05%, meaning that approximately 96% of the pairs have clear separation between the probabilities of TARGET=1 and TARGET=0.
    - Discordance = 3.94%

6. Conclusion

The model performance measures were calculated for both CART model and Random Forest Model. The performance measures for test data set on both the models are tabulated in Table-6.

Table-6: Comparison of Models

Measures	Test Data Set	
	CART Model	**Random Forest Model**
Accuracy	92.13%	96.48%
KS	45.65%	78.36%
AUC	79.22%	96.12%
Concordance	75.72%	96.05%
Discordance	24.27%	3.94%
Gini	51.11%	-

From Table-6, it is clear that the Random Forest Model is much better then CART model. **Hence, it is concluded that the best model is Random Forest.**