CONTENTS

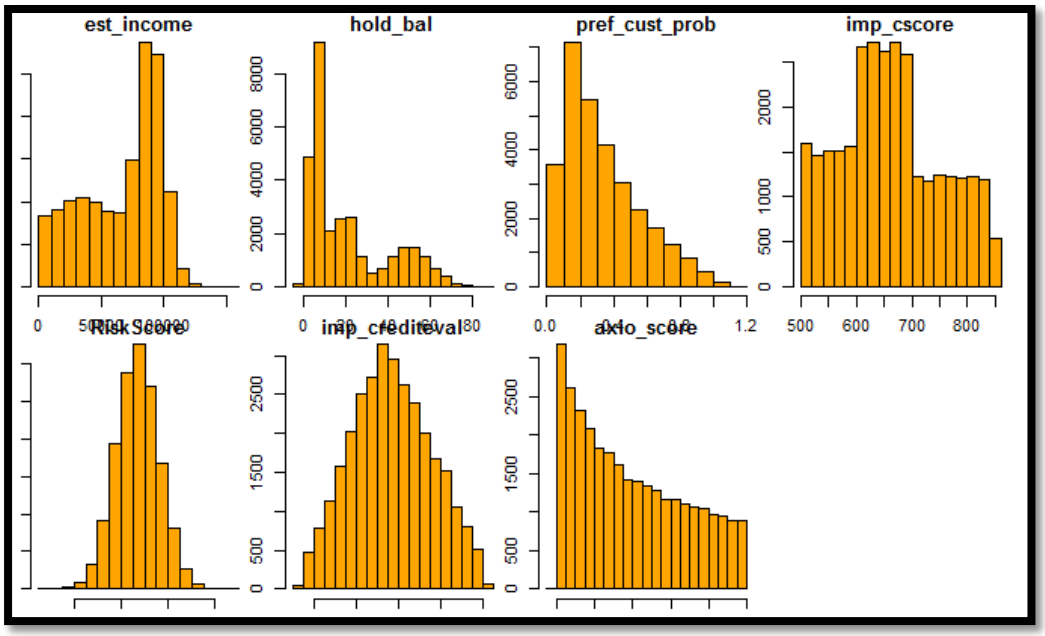
1. **Problem Statement**: Everyday huge number of customers submits online or offline application form for getting credit card from different bank. But it becomes very difficult for the banks to choose whom should they offer credit card and whom should not, Because there is huge risk of getting default or non-payments of bills if the customers are not chosen properly. So here we will build predictive models which can identify which customer to be offered with credit card to avoid any loss.
2. **Data used:** As provided there are 4 different data sources each contain 12 variables and 10000 observations. First 3 data sets are used for building models and the 4th data set has customer base of 10000, where identification of right customer for offering card has been done.

   Variables used

   - Customer_id
   - demographic_slice
   - country_reg
   - ad_exp
   - est_income
   - imp_cscore
   - imp_crediteval
   - axio_score
   - hold_bal
   - pref_cust_prob
   - RiskScore
   - card_offer(Dependent variable,Levels: FALSE,TRUE)

   ** Data dictionary was not provided

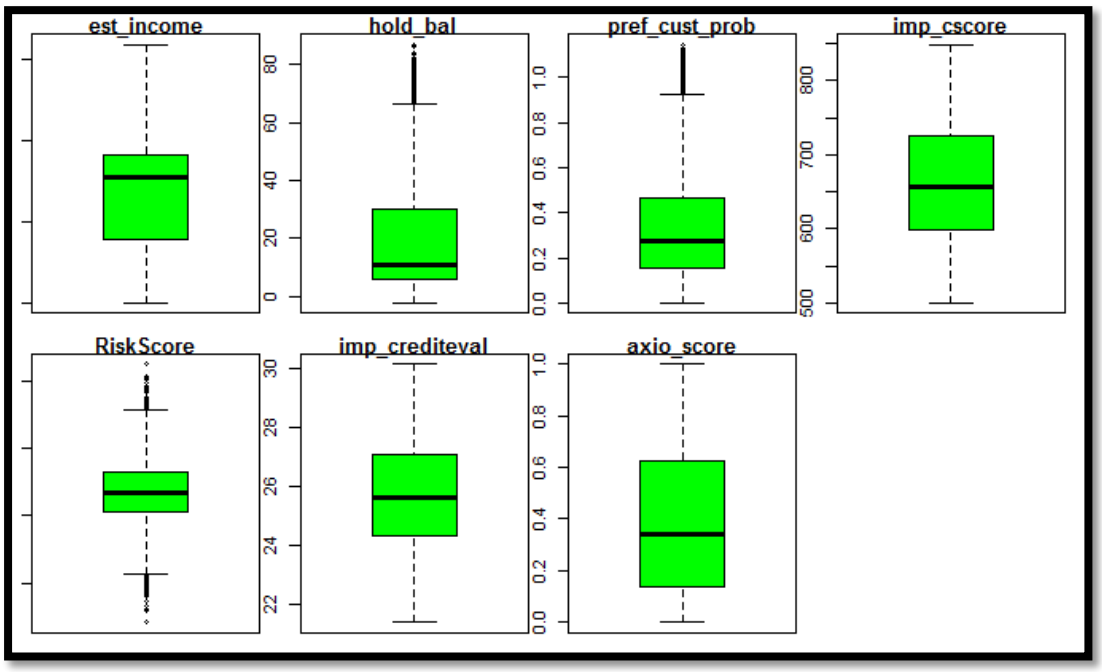
3. **Exploration of Numerical Variable**: After getting the data altogether it has dimension of 12 variables and 300000 observations. Out of 12 variables 7 were numerical variable and 5 were categorical variable, out of which one is dependent variable.
    - Checking the distribution of the variables:
      Histogram was plotted to check the distribution of the variables.



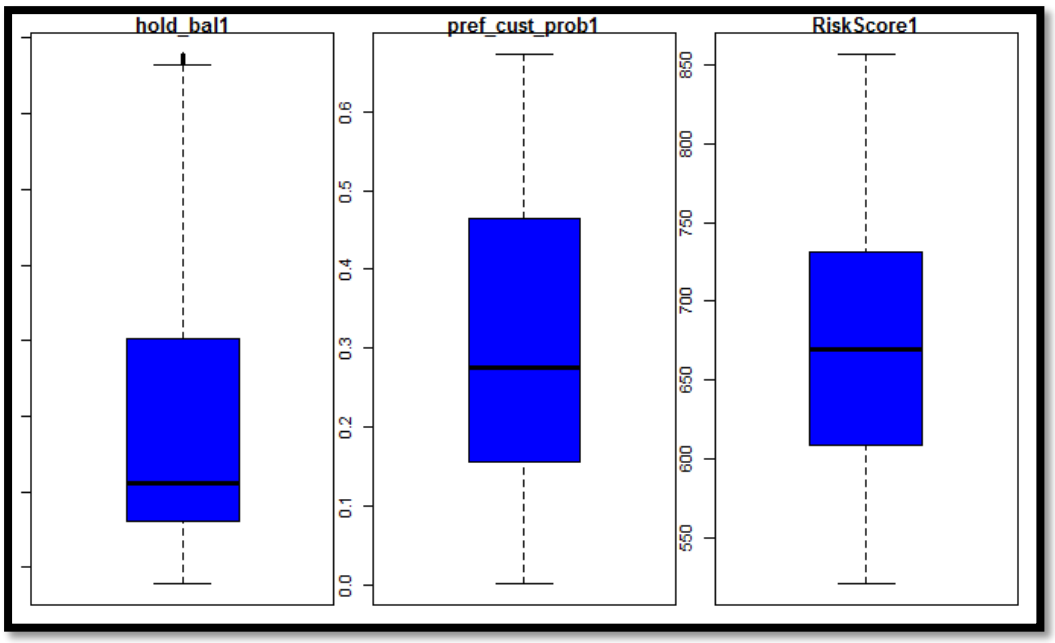
**RISK SCORE and imp_crediteval are quite normally distributed where as others are either left or right skewed

- Checking for missing value: There was no missing value in the data
- Checking for number of unique customers: Out of 30000 customer base 29562 were unique customers others have duplicate value. Though the customer_id was same but the observations were unique for them so they have been considered in the model.
- Checking for outliers: Boxplot was used to check the presence of outliers.

It is quite obvious from the box plot that hold_bal, pref_cust_prob and RiskScore has outliers



- Outlier Treatment: Winsorizing has been performed to cap the outlier at different percentile value. After capping ,three new variables were formed namely hold_bal1, pref_cust_prob1 and RiskScore1 and they have been investigated again for outliers.



** There is no presence of outliers detected

- Checking for multicollinearity: Correlation matrix has been plotted to check the presence of multicollinearity and it has been found that imp_crediteva and imp_cscore are highly correlated (92.68%)

	est_income	hold_bal	pref_cust_prob	imp_cscore	RiskScore	imp_crediteval	axio_score
est_income	1	0.007262397	-0.00037053	0.008151501	0.003054573	0.004528832	-0.01021144
hold_bal	0.007262397	1	-0.00171984	0.277269881	0.006093111	0.263531963	-0.001773238
pref_cust_prob	-0.00037053	-0.00171984	1	-0.004291352	-0.012527887	-0.005109821	-0.006460966
imp_cscore	0.008151501	0.277269881	-0.004291352	1	-0.001807692	0.926806661	-0.002696883
RiskScore	0.003054573	0.006093111	-0.012527887	-0.001807692	1	-0.001992242	0.005369174
imp_crediteval	0.004528832	0.263531963	-0.005109821	0.926806661	-0.001992242	1	-6.03926E-05
axio_score	-0.01021144	-0.001773238	-0.006460966	-0.002696883	0.005369174	-6.03926E-05	1

4. Exploration of Categorical Variable.

- Table of `demographic_slice`

  AX03efs BWEsk45 CARDIF2 DERS3w5
  24.96% 25.49% 24.63% 24.91%
- Table of country_reg

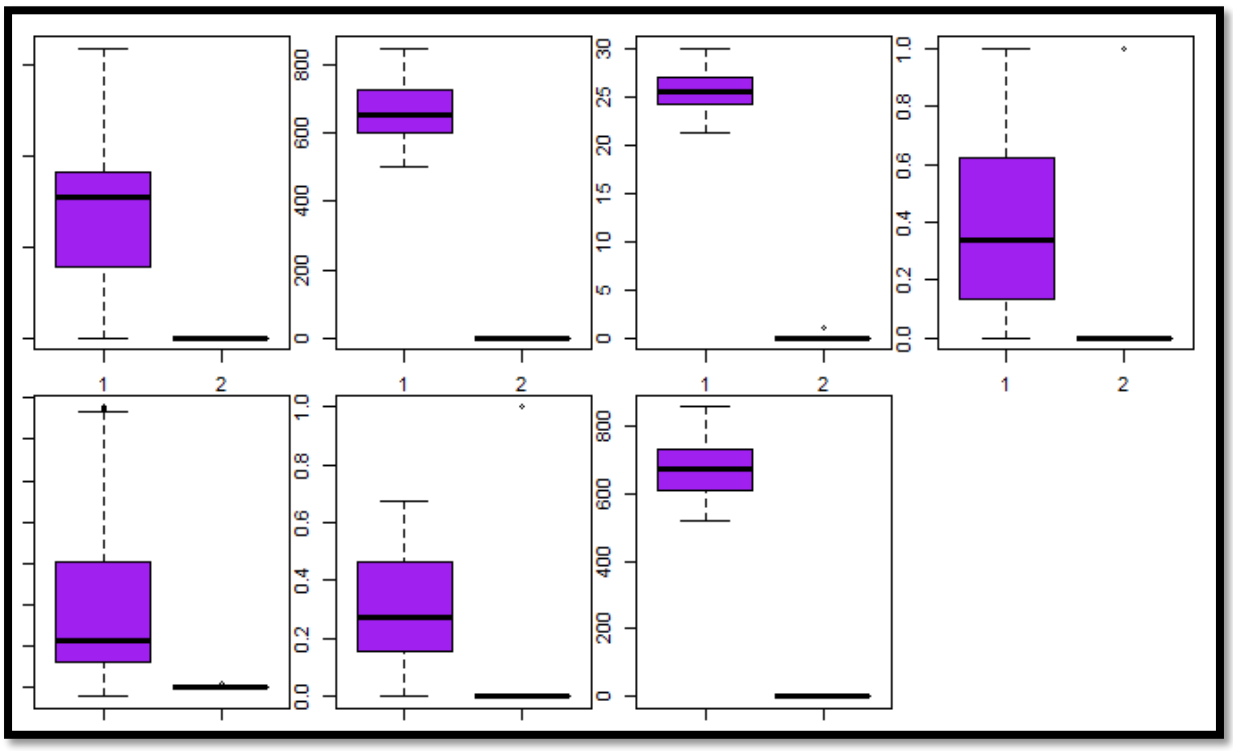
  E W
  49.71% 50.28%
- Table of ad_exp

  N Y
  50.17% 49.83 %
- Table of dependent variable card_offer

  FALSE TRUE
  85.06% 14.93%

  ***That means 85 % of the customers are not given the card while 15 % of the customers received it.

5. Boxplot of all the numerical variable with the dependent variable



est_income, imp_cscore, imp_crediteval, axio_score, hold_bal1, pref_cust_prob1, RiskScore1

It is quite obvious that there is a difference in the mean of the between FALSE and TRUE (that is not given and given) across all the numeric variables.

6. **Hypothesis Testing**

- T-TEST Result

Variable	**P_VALUE**
est_income	0.00000000000000000000
imp_cscore	0.00000006784177189400
imp_crediteval	0.00000203506708467040
hold_bal1	0.00000000000096397660
pref_cust_prob1	0.00000000000000000000

- CHI-SQUARE TEST Result

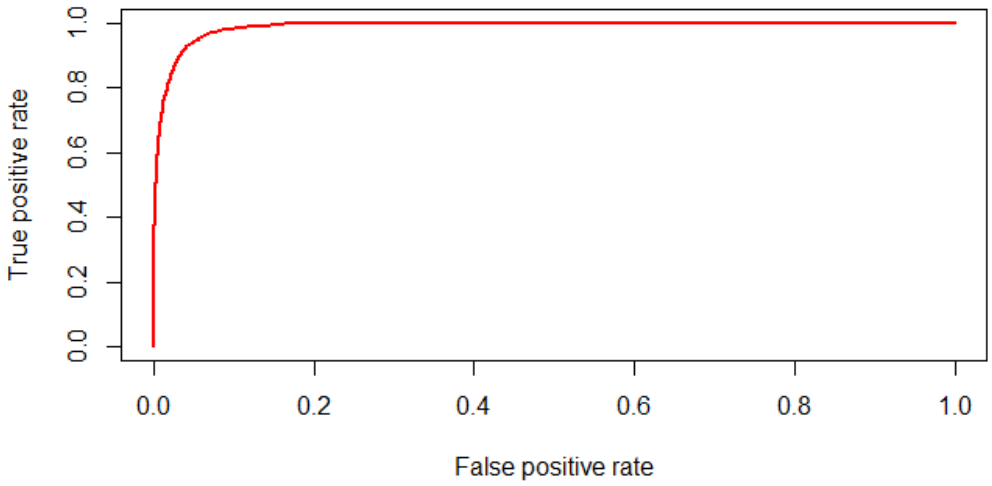
Variable	**P_VALUE**
demographic_slice	0.00000000000000022
country_reg	0.00000000000000022

7. Building Predictive Models

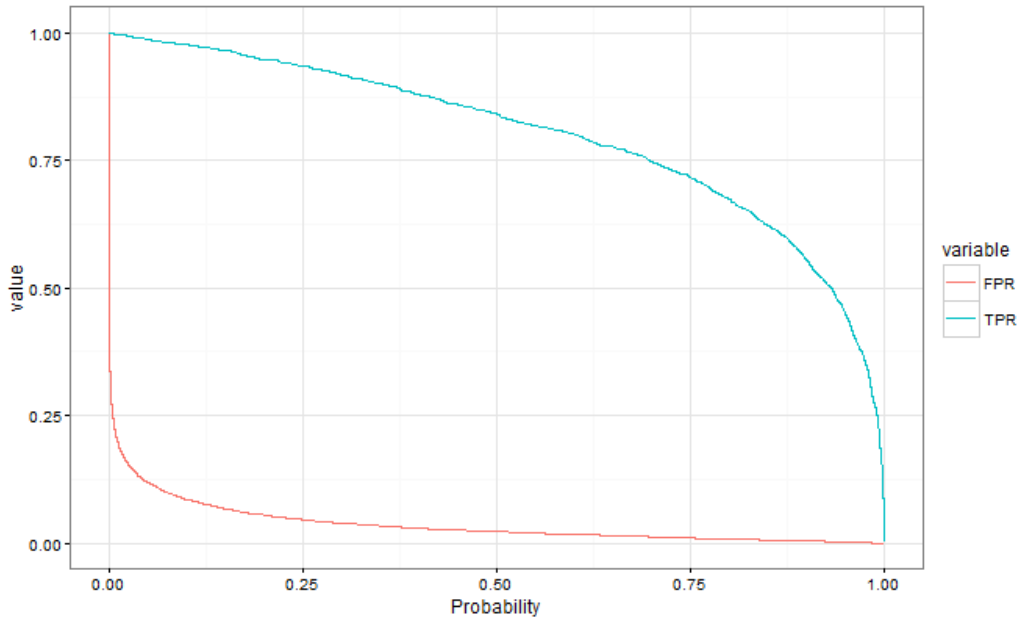
The data is divided into training and testing set in the proportion of 70:30(Class balanced is kept in mind)

- Logistic Regression Model.

  Model was build on training data and tested on test data. Confusion matrix, ROC, AUC, Accuracy has been checked. The model is 95 % accurate.



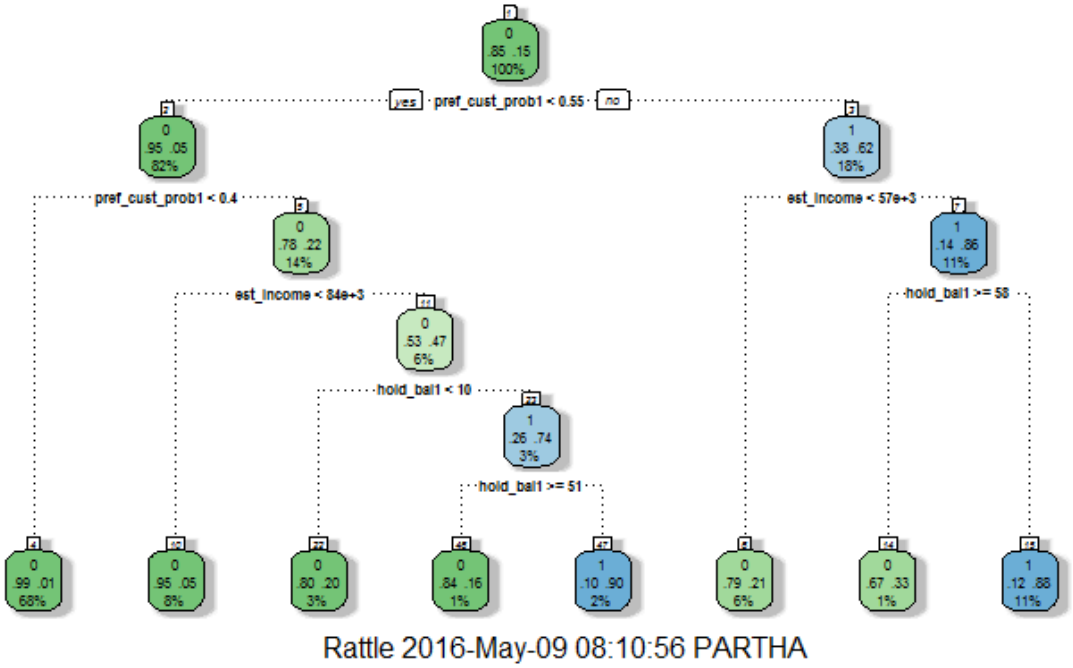
Thresh hold probability value is chosen at .50 as it was giving maximum accuracy



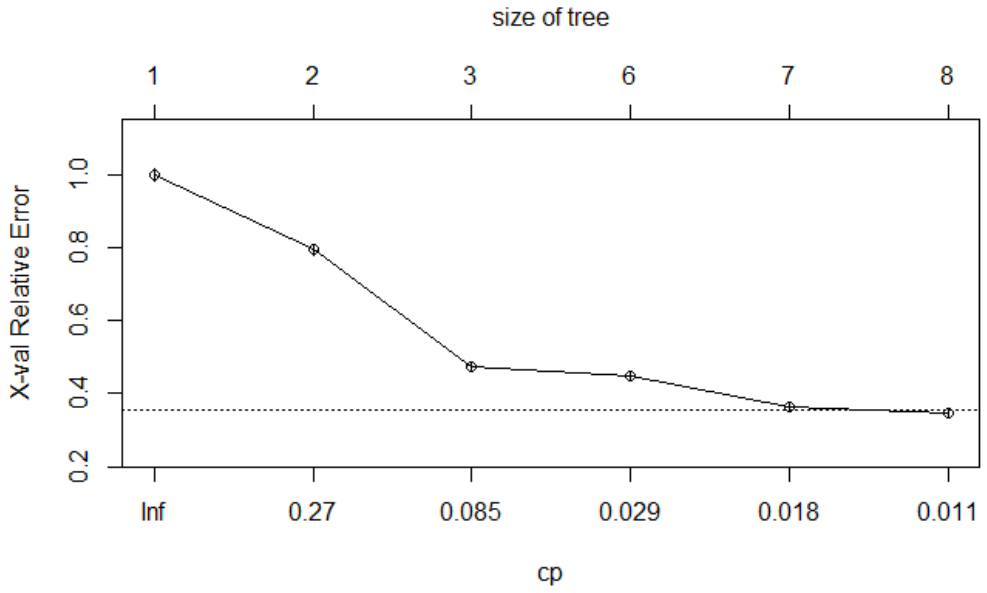
Confusion Matrix: 1 has been coded as TRUE and) for FALSE

	obs	
pred	0	1
0	7477	232
1	178	1112

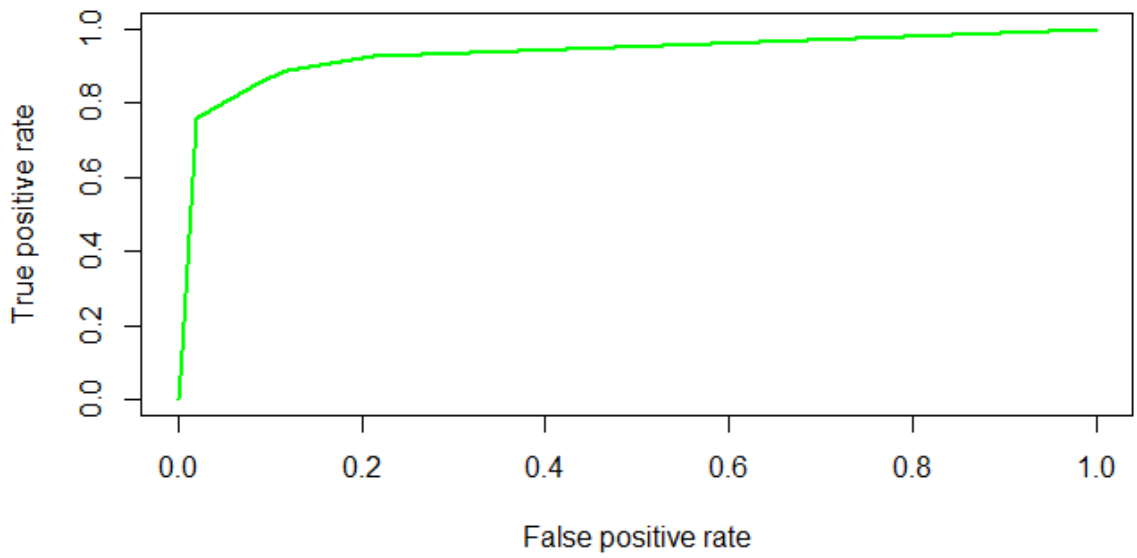
- CART Model
  Classification and Regression Tree has been build to predict the good set of customers eligible for getting card.



The tree has been pruned based on the best CP value



The RIC Curve of the model is found out.



The thresh hols probability value was set at .60.

Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7496  326
         1  159 1018

          Accuracy : 0.9461
            95% CI : (0.9412, 0.9507)
   No Information Rate : 0.8507
   P-Value [Acc > NIR] : < 0.00000000000000022

             Kappa : 0.7764
 Mcnemar's Test P-Value : 0.00000000000004784

        Sensitivity : 0.9792
        Specificity : 0.7574
     Pos Pred Value : 0.9583
     Neg Pred Value : 0.8649
         Prevalence : 0.8507
     Detection Rate : 0.8330
   Detection Prevalence : 0.8692
      Balanced Accuracy : 0.8683
```

- Neural Network:
  Neural Network Model is also build to find the eligible set of customers. Only one hidden layer has been used here.
  The data are normalized before feeding into NN.

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7639   16
         1   13 1331

               Accuracy : 0.9968
                 95% CI : (0.9954, 0.9978)
    No Information Rate : 0.8503
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.9873
 Mcnemar's Test P-Value : 0.7103

            Sensitivity : 0.9983
            Specificity : 0.9881
         Pos Pred Value : 0.9979
         Neg Pred Value : 0.9903
             Prevalence : 0.8503
         Detection Rate : 0.8489
   Detection Prevalence : 0.8507
      Balanced Accuracy : 0.9932
```

Conclusion:
Based on the predicted customers from three different models three sets of customers list has been provided.

_____THANK YOU_____