**Description**

**Thera Bank - Loan Purchase Modeling**

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.

You are brought in as a consultant and your job is to build the best model which can classify the right customers who have a higher probability of purchasing the loan. You are expected to do the following:

* EDA of the data available. Showcase the results using appropriate graphs - **(10 Marks)**
* Apply appropriate clustering on the data and interpret the output(Thera Bank wants to understand what kind of customers exist in their database and hence we need to do customer segmentation) - **(10 Marks)**
* Build appropriate models on both the test and train data (CART & Random Forest). Interpret all the model outputs and do the necessary modifications wherever eligible (such as pruning) - **(20 Marks)**
* Check the performance of all the models that you have built (test and train). Use all the model performance measures you have learned so far. Share your remarks on which model performs the best. - **(20 Marks)**

**Hint :**split <- sample.split(Thera\_Bank$Personal Loan, SplitRatio = 0.7)  
#we are splitting the data such that we have 70% of the data is Train Data and 30% of the data is my Test Data  
  
train<- subset(Thera\_Bank, split == TRUE)  
test<- subset( Thera\_Bank, split == FALSE)

| **Criteria** | **Points** |
| --- | --- |
| 1. EDA - Basic data summary, Univariate, Bivariate analysis, graphs | 10 |
| 2.1 Apply Clustering algorithm < type, rationale> | 5 |
| 2.2 Clustering Output interpretation < number of clusters, remarks to make it meaningful to understand> | 5 |
| 3.1 Applying CART <plot the tree> | 5 |
| 3.2 Interpret the CART model output <pruning, remarks on pruning, plot the pruned tree> | 5 |
| 3.3 Applying Random Forests<plot the tree> | 5 |
| 3.4 Interpret the RF model output <with remarks, making it meaningful for everybody> | 5 |
| 4.1 Confusion matrix interpretation | 5 |
| 4.2 Interpretation of other Model Performance Measures <KS, AUC, GINI> | 10 |
| 4.3 Remarks on Model validation exercise <Which model performed the best> | 5 |
| Points | 60 |



Note: Over fitting problem in decision tree method.