Activity Sheet

Learning outcomes

After solving these exercises, you will understand the following:

1. **Applying the Decision Trees using C5.0 and CART algorithms to solve classification and regression problems respectively**
2. **Understand and interpret the results generated from each algorithm in R**
3. **Comparison of the model performance in terms of Accuracy for Classification**
4. **Comparison of the model performance in terms of Mean square error/ Root Mean square error for regression**

Dataset and Target variables

Use the data set "Part1.csv"and "Part2.csv" to solve these exercises.

- Take the variable "RESPONSE" as target variable for classification and apply C5. 0 algorithm to classify whether a customer is good/bad customer
- Take the variable "AMOUNT" as target variable for regression and apply CART algorithm to predict how much a customer is eligible for

Note: *We have provided r code wherever it is necessary. This code is for your reference only. Do not copy and paste code in R console. To use this code, please ensure you change the data frame names and the variable names as per your working environment.*

Steps to follow for Classification

1. Go through 'Data_Descripion.csv' file to understand each variable and identify the data type (Numeric/categorical)
2. Import both the 'part1' and 'part2' csv files into R
3. Merge all two data sets by "obs" and remove the missing values
4. Separate the numerical and categorical data

5. Type conversion for each of the attribute- All attributes in data_Cat should be factor and data_Num should be numeric
6. Discretizing the numeric data using equal width or equal frequency method

   library(infotheo)

7. Let us construct a new data frame that contains **all** the variables in appropriate type. We need to create a dataframe with all the categorical and adding discretized numeric variables
8. Look at the summary of the data frame to check whether all the variables are categorical or not
9. Split the data into Training (60% of the total records) and Testing (40% of the records). Use below R code.

10. Let us the start building our first model. You need to install the below library.
    install.packages("C50")

11. Apply C.50 model on the training dataset. Use the below R code.
    library(C50)

    DT_C50=C5.0(response~.,data=TrainData,rules=T)

    #To get the important attributes

    C5imp(DT_C50,pct=T)

12. Understand the summary of the rules generated and the important variables
    summary(dtC50)

13. To validate our algorithm, let us apply on the test data and get the results.
    a=table(TrainData$response, predict(DT_C50, newdata=TrainData, type="class"))

    a

    b=table(TestData$response, predict(DT_C50, newdata=TestData, type="class"))

    accTrain = sum(diag(a))/sum(a)

    accTest = sum(diag(b))/sum(b)

Steps to follow for Regression problem

Our objective is to build a prediction model for the target variable 'Amount'. Let us reuse some of r code we had already written in the above exercise.

1. Take the final data you had built earlier and remove the amount variable (because this is in the binned format!).
2. Get the 'amount' variable, either from initial data frame or from the numeric data frame you generated in at step-4 in the above activity. Add this variable to the data frame you got in the step-1
3. Split the data into Training (60% of the total records) and Testing (40% of the records). Use R code at step-8 in the previous activity.
4. You need to install 'rpart' library to build the prediction model.
5. Use below r code to build the model

```
library(rpart)

DT_rpart<-rpart(amount~.,data=TrainData1,method="anova")

plot(DT_rpart,main="Regression tree for Amount",margin=0.15,uniform=T)
```

```
text(DT_rpart,use.n=T)

DT_rpart_PredTrain=predict(DT_rpart,newdata=TrainData1, type="vector")

A<-data.frame(TrainData1[28],DT_rpart_PredTrain)

DT_rpart_PredTest=predict(DT_rpart, newdata=TestData1,type="vector")

B<-data.frame(TestData1[28],DT_rpart_PredTest)

library(DMwR)

regr.eval(TrainData1[28],DT_rpart_PredTrain)
```