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.
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

Problem Statement:

A large child education toy company which sells its products online as well as in retail stores wants to improve its business. They would like to categorize the customers based of revenue so that they can make business decisions accordingly.

(a) Classify which customers are Regular and Premium based on Revenues.
(b) Predict the Revenues the customer is likely to contribute.

Pre-processing the data:

1. Understand the problem statement
2. Load the "CustomerData.csv" data into R.
3. Understand the data and identify the pre-processing steps to be applied on the data.
4. Apply the below pre-processing steps.
    a. Remove the attribute "CustomerID"
    b. Check for the missing values and impute it using knnImputation
    c. Convert the categorical attributes in to factor using as.factor()
    d. Bin the numeric attributes if you think it is good for analysis.
5. Custom bin the target as follow
    a. Revenue with less than $150 as "Regular" customers, and
    b. Revenue with greater than $150 as "Premium" customers.
    c. Convert "Revenue" into factor
6. Split the data into train and test data (70:30 ratio).

Classifying using C50:

7. Build model C50 using "Revenue" as target attribute

```
Library(C50)
DT_C50 <- C5.0(Revenue~.,data=train)
```

8. Predict on the train and test data sets.

```
Predict "Revenue" for train and test datasets
pred_Train = predict(DT_C50,newdata=train, type="class")
pred_Test = predict(DT_C50, newdata=test, type="class")
```

9. Generate confusion matrix.
10. Calculate accuracy on the train and test data.

11. Importance of the attributes
    C5imp(DT_C50, pct=TRUE)

Classifying using rpart:

12. Build model rpart using “Revenue” as target attribute
    Library(rpart)
    DT_rpart_class<-rpart(Revenue~., data=train, method="class")

13. Predict on the train and test data sets.
    #b. Predict "Revenue" for train and test datasets
    pred_Train = predict(DT_rpart_class,newdata= train, type="class")
    pred_Test = predict(DT_rpart_class, newdata=test, type="class")

14. Generate confusion matrix.

15. Calculate accuracy on the train and test data.

Regression Problem using rpart:

1. Apply the pre-processing steps from 1-4 listed above.

2. Use the original revenue attribute present in the data.

3. Split the dataset into train and test (70:30 ratio)

4. Build model rpart using “Revenue” as target attribute
   library(rpart)
   DT_rpart <- rpart(Revenue~.,data= train, method="anova")

5. Plot the tree
   library(rpart.plot)
   rpart.plot(DT_rpart,type=3,extra=101,fallen.leaves = FALSE)

6. Predict the Revenue for train and test datasets
   pred_Train=predict(DT_rpart,newdata= train, type="vector")
   pred_Test=predict(DT_rpart, newdata= test, type="vector")

7. Check the evaluation metrics on train data
   regr.eval(train$Revenue,pred_Train)

8. Check the evaluation metrics on test data
   regr.eval(test$Revenue,pred_Test)