Application on CART and Logistic Regression: Earnings

Goals : -Become familiar with data splits  
 -Become familiar with the Logistic Regression method  
 - Become familiar with the Tree Classification method  
 - Make Predictions

Part 1 Logistic Regression

1. Load train data “census.csv” using read.csv( ) function. Name your variable census   
   *Our data contains 1994 census data for 31978 individuals in the US*.
2. Explore your data using str( ), summary( )
3. Load caTools library ( use install.packages("caTools") if it is not already installed)
4. Copy: set.seed(2000) (this will make all of us have the same sequence of random values)
5. Split your data into training and testing sets (based on over50k variable) using sample.split() and subset functions.
6. Create ur logistic regression model using glm( dep ~ indep1 +indep2+.. , data=..,family=binomial), name it modlog.
7. Explore it using summary()
8. Make predictions using predict(model,newdata=test,type=”response”).
9. Using the table function, give the binomial predictions by presenting the confusion matrix
10. Compare this method’s accuracy to the baseline one
11. Calculate AUC and present the ROCR curve by copying  
    install.packages("ROCR")

library(ROCR)

ROCRpred = prediction(predictions, test$over50k)

ROCRperf= performance(ROCRpred,"tpr","fpr")

plot(ROCRperf)

plot(ROCRperf,colorize=TRUE)

plot(ROCRperf,colorize=TRUE,print.cutoffs.at=seq(0,1,0.1),text.adj=c(-0.2,1.7))

as.numeric(performance(ROCRpred, "auc")@y.values)

Part 2 Classification Tree

1. Install and load rpart package
2. Create a CART model using rpart(over50k ~ . - over50k, data=train, method="class") , call it CART50
3. Illustrate the tree using the prp() function
4. Make predictions using the predict function (type=”class” for Classification)
5. Show the confusion matrix
6. Compare Accuracies.