IE 425

Homework 1 (due April 5, Tuesday @23:59)

1. Consider the dataset given in the file “Financialdistress-cat.csv”. The output attribute to be predicted is the Financial.Distress attribute, which is zero if the company is in a healthy condition, one otherwise. Use a seed value of 500 whereever you need a seed.

a) Partition the dataset into training and test sets where 75% of goes into the training set and 25% goes into the test set. What is the percentage of companies in distress in the overall, training, and test sets?

b) Using the rpart package and training set, determine the best size of the best tree in terms of cross validation error. How many leaf nodes do exist in the tree?

c) Make predictions in the test set and report the error rate, sensitivity, specificity, and precision using the confusionMatrix function of the caret package.

d) Using the tree package find the size of the tree which makes the cost complexity (measured by the deviance in the tree package) the smallest? How many leaf nodes does it have?.

e) Make predictions in the test set and report the error rate, sensitivity, specificity, and precision using the confusionMatrix function of the caret package. Compare the result with part (c).

2. Consider the dataset given in the file “ToyotaCorolla.csv”. The output attribute to be predicted is the Price attribute. Use a seed value of 500 whereever you need a seed.

a) Partition the dataset into training and test sets where 80% of goes into the training set and 20% goes into the test set.

b) Using the rpart package and training set, determine the best size of the best tree in terms of cross validation error. How many leaf nodes do exist in the tree?

c) Make predictions in the test set and report the root mean square error rate and mean absolute error using the functions in the Metrics package.