Homework 5
IE 7275: Data Mining in Engineering

Read the material "Tutorial on CART with R.pdf". Read the book chapter on "Logistic and Poisson Regression with R.pdf." Ignore the Poisson Regression part of the tutorial.

Problem 1 (Predicting Price of Used Car, CART) [35 points]

The file **ToyotaCorolla.xlsx** contains the data on used cars (Toyota Corolla) on sale during late summer of 2004 in The Netherlands. It has 1436 records containing details on 38 attributes, including *Price, Age, Kilometers, HP*, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications.

Data Preprocessing: Create dummy variables for the categorical predictors (Fuel Type and Color). Split the data into training (50%), validation (30%), and test (20%) datasets.

a. Run a regression tree (RT) with the output variable Price and input variables *Age_08_04, KM, Fuel_Type, HP, Automatic, Doors, Quarterly_Tax, Mfg_Guarantee, Guarantee_Period, Airco, Automatic_Airco, CD Player, Powered_Windows, Sport_Model,* and *Tow_Bar*.

   i. Which appear to be the three or four most important car specifications for predicting the car's price?

   ii. Compare the prediction errors of the training, validation, and test sets by examining their RMS error and by plotting the three boxplots. What is happening with the training set predictions? How does the predictive performance of the test set compare to the other two? Why does this occur?

   iv. If we used the full tree instead of the best pruned tree to score the validation set, how would this affect the predictive performance for the validation set? (Hint: Does the full tree use the validation data?)

b. Let us see the effect of turning the price variable into a categorical variable. First, create a new variable that categorizes price into 20 bins of equal counts. Now repartition the data keeping Binned Price instead of Price. Run a classification tree (CT) with the same set of input variables as in the RT, and with Binned Price as the output variable.

i. Compare the tree generated by the CT with the one generated by the RT. Are they different? (Look at structure, the top predictors, size of tree, etc.) Why?

ii. Predict the price, using the RT and the CT, of a used Toyota Corolla with the specifications listed in Table below.

Table: Specifications for a particular Toyota Corolla

Variable	Value
Age_08_04	77
KM	117,000
Fuel_Type	Petrol
HP	110
Automatic	No
Doors	5
Quarterly_Tax	100
Mfg_Garantee	No
Guarantee_Period	3
Airco	Yes
Automatic_Airco	No
CD_Player	No
Powered_Windows	No
Sport_Model	No
Tow_Bar	Yes

iii. Compare the predictions in terms of the predictors that were used, the magnitude of the difference between the two predictions, and the advantages and disadvantages of the two methods.

Problem 2 (Financial condition of banks, Logistic Regression) [30 points]

The file **Banks.xlsx** includes data on a sample of 20 banks. The *Financial Condition* (*Y*) column records the judgment of an expert on the financial condition of each bank. This dependent variable takes one of two possible values -- *weak* or *strong* -- according to the financial condition of the bank. The predictors are two ratios used in the financial analysis of banks: *TotLns&Lses/Assets* (X_1) is the ratio of total loans and leases to total assets and *TotExp/Assets* (X_2) is the ratio of total expenses to total assets. The target is to use the two ratios for classifying the financial condition of a new bank.

Run a logistic regression model (on the entire dataset) that models the status of a bank as a function of the two financial measures provided. Specify the success class as *weak*

(this is similar to creating a dummy that is 1 for financially *weak* banks and 0 otherwise), and use the default cutoff value of 0.5.

a. Write the estimated equation that associates the financial condition of a bank with its two predictors in three formats:
   i. The logit as a function of the predictors
   ii. The odds as a function of the predictors
   iii. The probability as a function of the predictors
b. Consider a new bank whose total loans and leases/assets ratio = 0.6 and total expenses/assets ratio = 0.11. From your logistic regression model, estimate the following four quantities for this bank: the logit, the odds, the probability of being financially weak, and the classification of the bank.
c. The cutoff value of 0.5 is used in conjunction with the probability of being financially weak. Compute the threshold that should be used if we want to make a classification based on the odds of being financially weak, and the threshold for the corresponding logit.
d. Interpret the estimated coefficient for the total loans & leases to total assets ratio (TotLns&Lses/Assets) in terms of the odds of being financially weak.
e. When a bank that is in poor financial condition is misclassified as financially strong, the misclassification cost is much higher than when a financially strong bank is misclassified as weak. To minimize the expected cost of misclassification, should the cutoff value for classification (which is currently at 0.5) be increased or decreased?

Problem 3 (Identifying good system administrators, Logistic Regression) [35 points]

A management consultant is studying the roles played by experience and training in a system administrator's ability to complete a set of tasks in a specified amount of time. In particular, she is interested in discriminating between administrators who are able to complete given tasks within a specified time and those who are not. Data are collected on the performance of 75 randomly selected administrators. They are stored in the file **`System Administrators.xlsx`**.

The variable *Experience* (X_1) measures months of full-time system administrator experience, while *Training* (X_2) measures the number of relevant training credits. The dependent variable *Completed* (Y) is either *Yes* or *No*, according to whether or not the administrator completed the tasks.

a. Create a scatterplot of *Experience* versus *Training* using color or symbol to differentiate programmers who complete the task from those who did not complete it. Which predictor(s) appear(s) potentially useful for classifying task

completion?

b. Run a logistic regression model with both predictors using the entire dataset as training data. Among those who complete the task, what is the percentage of programmers who are incorrectly classified as failing to complete the task?
c. To decrease the percentage in part (b), should the cutoff probability be increased or decreased?
d. How much experience must be accumulated by a programmer with 4 years of training before his or her estimated probability of completing the task exceeds 50%?

Files Included in the Folder:

1. **Homework 5.pdf**
2. **Tutorial on CART with R.pdf**
3. **Logistic and Poisson Regression with R.pdf**
4. **ToyotaCorolla.xlsx**
5. **Banks.xlsx**
6. **System Administrators.xlsx**