Homework 4

IE 7275: Data Mining in Engineering

Task 1: Tutorial

- Practice R models presented in "R Code for Textbook Examples in Chap 9 10.pdf." For your convenience, the data sets referenced in the document are included the homework folder.

Chapter 9: Classification and Regression Trees

Problem 9.1 [20 points]

Competitive Auctions on eBay.com. The file **eBayAuctions.csv** contains information on 1972 auctions that transacted on eBay.com during May–June 2004. The goal is to use these data to build a model that will classify auctions as competitive or noncompetitive. A competitive auction is defined as an auction with at least two bids placed on the item auctioned. The data include variables that describe the item (auction category), the seller (his/her eBay rating), and the auction terms that the seller selected (auction duration, opening price, currency, day-of-week of auction close). In addition, we have the price at which the auction closed. The task is to predict whether or not the auction will be competitive.

Data Preprocessing. Convert variable Duration into a categorical variable. Split the data into training (60%) and validation (40%) datasets.

a. Fit a classification tree using all predictors, using the best-pruned tree. To avoid overfitting, set the minimum number of records in a terminal node to 50 (in R: minbucket = 50). Also, set the maximum number of levels to be displayed at seven (in R: maxdepth = 7). Write down the results in terms of rules. (Note: If you had to slightly reduce the number of predictors due to software limitations, or for clarity of presentation, which would be a good variable to choose?)
b. Is this model practical for predicting the outcome of a new auction?
c. Describe the interesting and uninteresting information that these rules provide.
d. Fit another classification tree (using the best-pruned tree, with a minimum number of records per terminal node = 50 and maximum allowed number of displayed levels = 7), this time only with predictors that can be used for predicting the outcome of a new auction. Describe the resulting tree in terms of rules. Make sure to report the smallest set of rules required for classification.
e. Plot the resulting tree on a scatter plot: Use the two axes for the two best (quantitative) predictors. Each auction will appear as a point, with coordinates corresponding to its values on those two predictors. Use different colors or symbols to separate competitive

and noncompetitive auctions. Draw lines (you can sketch these by hand or use R) at the values that create splits. Does this splitting seem reasonable with respect to the meaning of the two predictors? Does it seem to do a good job of separating the two classes?

f. Examine the lift chart and the confusion matrix for the tree. What can you say about the predictive performance of this model?

g. Based on this last tree, what can you conclude from these data about the chances of an auction obtaining at least two bids and its relationship to the auction settings set by the seller (duration, opening price, ending day, currency)? What would you recommend for a seller as the strategy that will most likely lead to a competitive auction?

Problem 9.2 [20 points]

Predicting Delayed Flights. The file **FlightDelays.csv** contains information on all commercial flights departing the Washington, DC area and arriving at New York during January 2004. For each flight, there is information on the departure and arrival airports, the distance of the route, the scheduled time and date of the flight, and so on. The variable that we are trying to predict is whether or not a flight is delayed. A delay is defined as an arrival that is at least 15 minutes later than scheduled.

Data Preprocessing. Transform variable day of week (DAY_WEEK) info a categorical variable. Bin the scheduled departure time into eight bins (in R use function **cut**()). Use these and all other columns as predictors (excluding DAY_OF_MONTH). Partition the data into training and validation sets.

a. Fit a classification tree to the flight delay variable using all the relevant predictors. Do not include DEP_TIME (actual departure time) in the model because it is unknown at the time of prediction (unless we are generating our predictions of delays after the plane takes off, which is unlikely). Use a pruned tree with maximum of 8 levels, setting cp = 0.001. Express the resulting tree as a set of rules.

b. If you needed to fly between DCA and EWR on a Monday at 7:00 AM, would you be able to use this tree? What other information would you need? Is it available in practice? What information is redundant?

c. Fit the same tree as in (a), this time excluding the Weather predictor. Display both the pruned and unpruned tree. You will find that the pruned tree contains a single terminal node.

   i. How is the pruned tree used for classification? (What is the rule for classifying?)

   ii. To what is this rule equivalent?

   iii. Examine the unpruned tree. What are the top three predictors according to this tree?

   iv. Why, technically, does the pruned tree result in a single node?

   v. What is the disadvantage of using the top levels of the unpruned tree as opposed to the pruned tree?

vi. Compare this general result to that from logistic regression in the example in Chapter 10. What are possible reasons for the classification tree's failure to find a good predictive model?

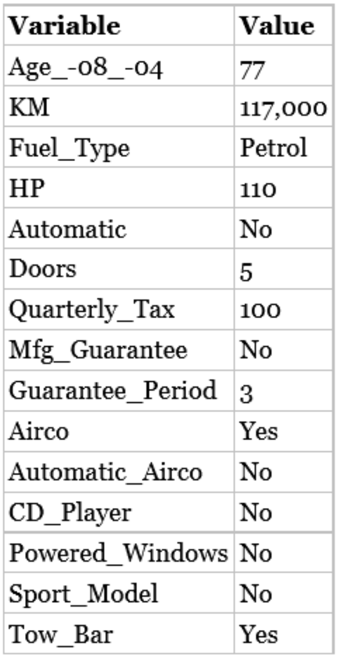
Problem 9.3 [15 points]

Predicting Prices of Used Cars (Regression Trees). The file **ToyotaCorolla.csv** 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. Split the data into training (60%), and validation (40%) datasets.

a. Run a regression tree (RT) with outcome variable Price and predictors 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. Keep the minimum number of records in a terminal node to 1, maximum number of tree levels to 100, and cp = 0.001, to make the run least restrictive.
   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 and validation sets by examining their RMS error and by plotting the two boxplots. What is happening with the training set predictions? How does the predictive performance of the validation set compare to the training set? Why does this occur?
   iii. How can we achieve predictions for the training set that are not equal to the actual prices?
   iv. Prune the full tree using the cross-validation error. Compared to the full tree, what is the predictive performance for the validation set?
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. Now repartition the data keeping Binned_Price instead of Price. Run a classification tree with the same set of input variables as in the RT, and with Binned_Price as the output variable. Keep the minimum number of records in a terminal node to 1.
   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 9.1.
   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.

Table 9.1. 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_Guarantee	No
Guarantee_Period	3
Airco	Yes
Automatic_Airco	No
CD_Player	No
Powered_Windows	No
Sport_Model	No
Tow_Bar	Yes

Chapter 10: Logistic Regression

Problem 10.1 [15 points]

Financial Condition of Banks. The file **Banks.csv** includes data on a sample of 20 banks. The "Financial Condition" column records the judgment of an expert on the financial condition of each bank. This outcome 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 is the ratio of total loans and leases to total assets and TotExp/Assets 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 (use R to do all the intermediate calculations; show your final

answers to four decimal places): the logit, the odds, the probability of being financially weak, and the classification of the bank (use cutoff = 0.5).

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 10.2 [15 points]

<u>Identifying Good System Administrators</u>. 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 **SystemAdministrators.csv**.

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

a. Create a scatter plot of Experience vs. Training using color or symbol to distinguish programmers who completed 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 completed the task, what is the percentage of programmers 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 0.5?

Problem 10.3 [15 points]

<u>Competitive Auctions on eBay.com</u>. The file **eBayAuctions.csv** contains information on 1972 auctions transacted on eBay.com during May–June 2004. The goal is to use these data to build a model that will distinguish competitive auctions from noncompetitive ones. A competitive auction is defined as an auction with at least two bids placed on the item being auctioned. The data include variables that describe the item (auction category), the seller (his or her eBay rating),

and the auction terms that the seller selected (auction duration, opening price, currency, day of week of auction close). In addition, we have the price at which the auction closed. The goal is to predict whether or not an auction of interest will be competitive.

Data preprocessing. Create dummy variables for the categorical predictors. These include Category (18 categories), Currency (USD, GBP, Euro), EndDay (Monday–Sunday), and Duration (1, 3, 5, 7, or 10 days).

a. Create pivot tables for the mean of the binary outcome (Competitive?) as a function of the various categorical variables (use the original variables, not the dummies). Use the information in the tables to reduce the number of dummies that will be used in the model. For example, categories that appear most similar with respect to the distribution of competitive auctions could be combined.
b. Split the data into training (60%) and validation (40%) datasets. Run a logistic model with all predictors with a cutoff of 0.5.
c. If we want to predict at the start of an auction whether it will be competitive, we cannot use the information on the closing price. Run a logistic model with all predictors as above, excluding price. How does this model compare to the full model with respect to predictive accuracy?
d. Interpret the meaning of the coefficient for closing price. Does closing price have a practical significance? Is it statistically significant for predicting competitiveness of auctions? (Use a 10% significance level.)
e. Use stepwise selection (use function **step**() in the stats package or function **stepAIC**() in the MASS package) and an exhaustive search (use function **glmulti**() in package **glmulti**) to find the model with the best fit to the training data. Which predictors are used?
f. Use stepwise selection and an exhaustive search to find the model with the lowest predictive error rate (use the validation data). Which predictors are used?
g. What is the danger of using the best predictive model that you found?
h. Explain why the best-fitting model and the best predictive models are the same or different.
i. If the major objective is accurate classification, what cutoff value should be used?
j. Based on these data, what auction settings set by the seller (duration, opening price, ending day, currency) would you recommend as being most likely to lead to a competitive auction?

Files Included in the Folder:

1. **Homework 6.pdf**
2. **R Code for Textbook Examples in Chap 9 10.pdf**

3. Banks.csv
4. eBayAuctions.csv
5. FlightDelays.csv
6. RidingMowers.csv
7. SystemAdministrators.csv
8. ToyotaCorolla.csv
9. UniversalBank.csv