Homework 5
IE 7275: Data Mining in Engineering

Problem 1 (Predicting Delayed Flights, CART) [50 points]

The file **FlightDelays.xlsx** contains information on all commercial flights departing the Washington, D.C., 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: Create dummies for day of week, carrier, departure airport, and arrival airport. This will give you 17 dummies. Bin the scheduled departure time into 2-hour bins. After binning DEP_TIME into 8 bins, this new variable should be broken down into 7 dummies (because the effect will not be linear due to the morning and afternoon rush hours). This will avoid treating the departure time as a continuous predictor because it is reasonable that delays are related to rush-hour times. 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 doing our predicting of delays after the plane takes off, which is unlikely). In the third step of the classification tree restrict maximum number of levels to be displayed to 6. Use the best pruned tree without a limitation on the minimum number of observations in the final nodes. Express the resulting tree as a set of rules.

b. If you needed to fly between DCA and EWR on a Monday at 7 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 another tree, this time excluding the day-of-month predictor. (Why?) Select the option of seeing both the full tree and the best pruned tree. You will find that the best pruned tree contains a single terminal node.

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

    ii. To what is this rule equivalent?

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

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

v. What is the disadvantage of using the top levels of the full tree as opposed to the best 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 2 (Predicting Price of Used Car, CART) [50 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) using the Prediction menu in XLMiner with the outputvariable 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*. Keep the minimum number of records in a terminal node to 1, maximum number of tree levels to 100, and the scoring option to Full Tree, 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, 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. Use *Transform* → *Bin continuous data* to categorize Price into 20 bins of equal counts (leave all other options at their default). Now repartition the data keeping Binned Price instead of Price. Run

a classification tree (CT) using the *Classification* menu of XLMiner 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 and uncheck the *P rune Tree* option, to make the run least restrictive.

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.

Files Included in the Folder:

1. **Homework 5.pdf**
2. **Tutorial on CART with R.pdf**
3. **FlightDelays.xlsx**
4. **ToyotaCorolla.xlsx**