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?

9.3 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. (The example in Section 9.7 is a subset of this dataset).

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 9.6.

TABLE 9.6 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

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.