Individual_assignment7

Yuhan_Xu_474154

2019/10/25

Prefix

In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

```
library(ISLR)
attach(Carseats)
```

(a)

Q: Split the data set into a training set and a test set.
A:

```
set.seed(1)
train = sample(nrow(Carseats),nrow(Carseats)*0.7)
Carseats.train = Carseats[train,]
Carseats.test = Carseats[-train,]
Sales.test = Sales[-train]
```

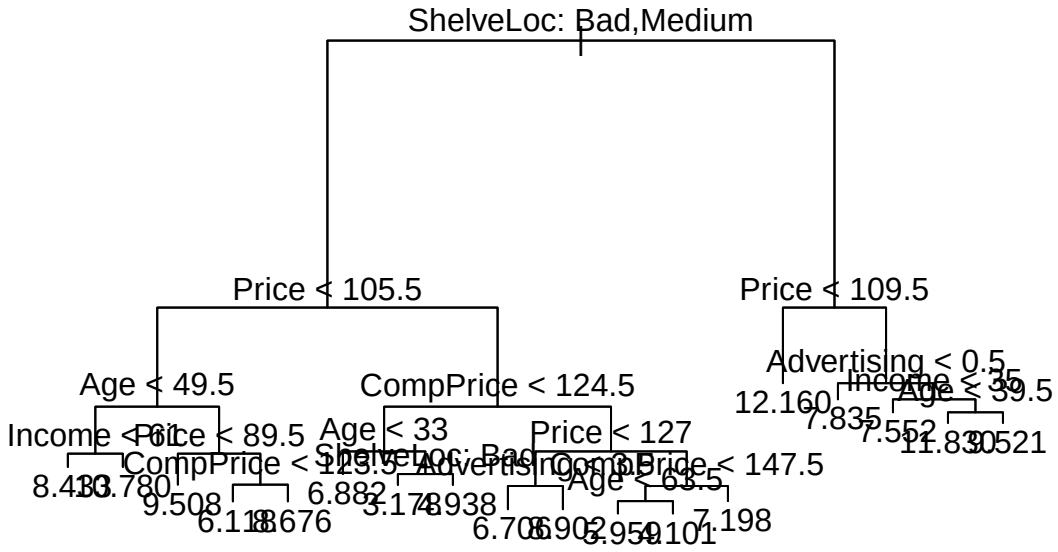
(b)

Q: Fit a regression tree to the training set. Plot the tree, and interpret the results.
A:

```
library(tree)
tree.Carseats = tree(Sales~., Carseats, subset = train)
summary(tree.Carseats)
```

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = Carseats, subset = train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "Income"      "CompPrice"
## [6] "Advertising"
## Number of terminal nodes:  18
## Residual mean deviance:  2.409 = 631.1 / 262
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.77800 -0.96100 -0.08865  0.00000  1.01800  4.14100
```

```
plot(tree.Carseats)
text(tree.Carseats, pretty = 0)
```



The tree has 18 terminal nodes.

6 variables are used in tree constructions. They are:
-"ShelveLoc", a factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site
-"Price", the price company charges for car seats at each site
-"Age", average age of the local population
-"Income", community income level (in thousands of dollars)
-"CompPrice", the price charged by competitor at each location
-"Advertising", local advertising budget for company at each location (in thousands of dollars)

When ShelveLoc, that is the quality of the shelving location for the car seats at each site, is bad or medium, we need to look at the price company charges for car seats at each site to make the prediction.
1. When that price is lower than 105.5, we need to look at average age of the local population.
(i) When average age is lower than 49.5, we will turn to community income level:
When income level is lower than $61,000, the prediction of sale will be 8433.
When income level is higher than $61,000, the prediction will be 10780.
(ii) when average age is more than 49.5, we will turn to price charged for car seats again:
When price is lower than 89.5, the prediction will be 9508.
When price is higher than 89.5, we need to look at price charged by competitor:
When competitor's price is lower than 123.5, the prediction will be 6118.
When competitor's price is higher than 123.5, the prediction will be 8676.
2. When price is higher than 105.5, we need to look at price charged by competitor first.
(i) When competitor's price is lower than 124.5. We will turn to average age of the local population.
When average age is lower than 33, the prediction will be 6882.
When average age is higher than 33, if the quality of the shelving location for the car seats is bad, the prediction will be 3178; If the quality is medium, then the prediction will be 4938.
(ii) When competitor's price is lower than 124.5. We turn to car seats price.
When the price is lower than 127, we will look at the local advertising budget for company:
When the budget is is less than $3500, the prediction will be 6706.
When the budget is is more than $3500, the prediction will be 8902.
When the price is higher than 127, we first look at competitor's price.
When competitor's price is lower than 147.5, we will look at average age of the local population:
When average age is lower than 63.5, the prediction will be 5959.
When average age is higher than 63.5, the prediction will be 4101.
When competitor's price is higher than 147.5, the prediction will be 7198.

When the quality of the shelving location for the car seats at each site is good, we need to look at car seats price to determine the prediction.
1. When price is lower than 109.5, the prediction will be 12160.
2. When price is higher than 109.5, we turn to local advertising budget.
When advertising budget is less than 500, the prediction will be 7835.
When advertising budget is higher than 500, we will turn to community income level.
(i) When income level is lower than $35,000, the prediction will be 7552.
(ii) When income level is higher than $35,000, we turn to average age of local population.
When average age is lower than 39.5, the prediction will be 11830. When average age is higher than 39.5, the prediction will be 9521.

Q: What test MSE do you obtain?
A: The test MSE is 4.208383.

```
yhat = predict(tree.Carseats, Carseats.test)
mean((Sales.test - yhat)^2)
```

```
## [1] 4.208383
```

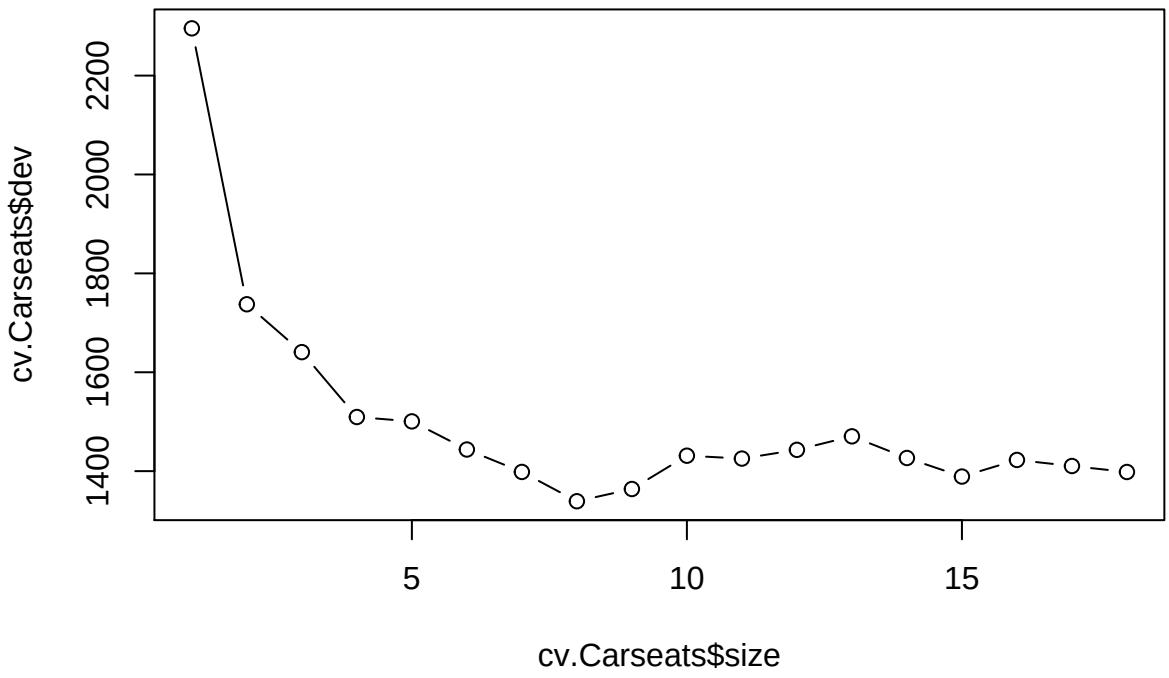
(c)

Q: Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

```
set.seed(2)
cv.Carseats = cv.tree(tree.Carseats)
cv.Carseats
```

```
## $size
##  [1] 18 17 16 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1
##
## $dev
##  [1] 1398.250 1410.354 1422.633 1389.014 1426.590 1470.396 1443.051
##  [8] 1425.404 1431.216 1363.823 1339.107 1398.344 1443.745 1500.622
## [15] 1509.476 1640.675 1737.490 2295.575
##
## $k
##  [1]      -Inf  22.92183  23.41401  26.69354  27.92150  28.73544  32.58363
##  [8]  34.84105  36.07741  48.85081  50.17274  66.89667  74.67562  96.50399
## [15] 101.36487 163.18945 213.84387 576.01474
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
plot(cv.Carseats$size, cv.Carseats$dev, type = "b")
```



A: From the graph, we can see that the tree with 8 terminal nodes has the lowest cross-validated deviation. However, pruning the tree doesn't result in lower test MSE. On the contrary, pruning the tree to 8 terminal nodes increases the test MSE to 4.579256.

```
prune.Carseats = prune.tree(tree.Carseats, best = 8)
summary(prune.Carseats)
```

```
##
## Regression tree:
## snip.tree(tree = tree.Carseats, nodes = c(8L, 23L, 7L, 19L, 10L,
## 22L))
## Variables actually used in tree construction:
## [1] "ShelveLoc" "Price"     "Age"       "CompPrice"
## Number of terminal nodes:  8
## Residual mean deviance:  3.541 = 963.3 / 272
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5.8980 -1.3460 -0.0245  0.0000  1.3030  4.3790
```

```
yhat1 = predict(prune.Carseats, Carseats.test)
mean((Sales.test - yhat1)^2)
```

```
## [1] 4.579256
```