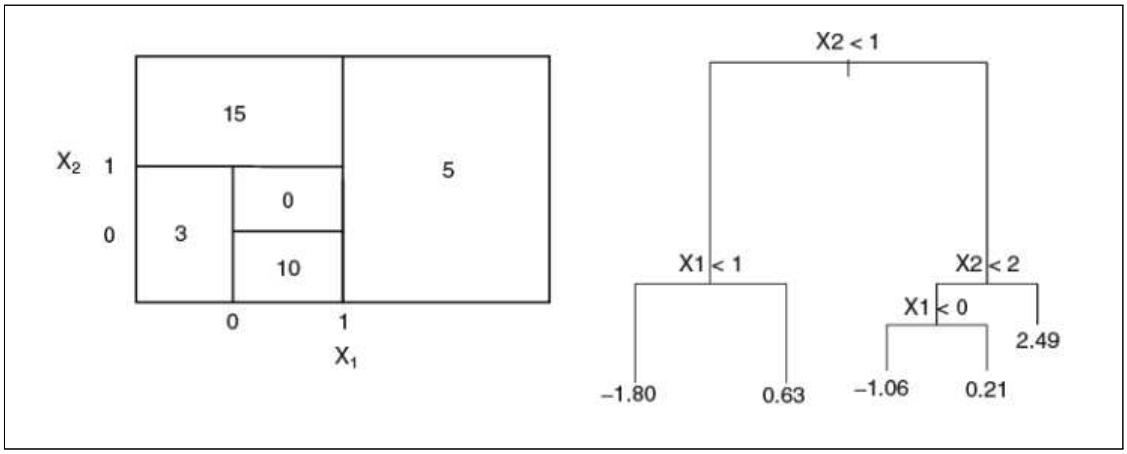
Data Mining HW03

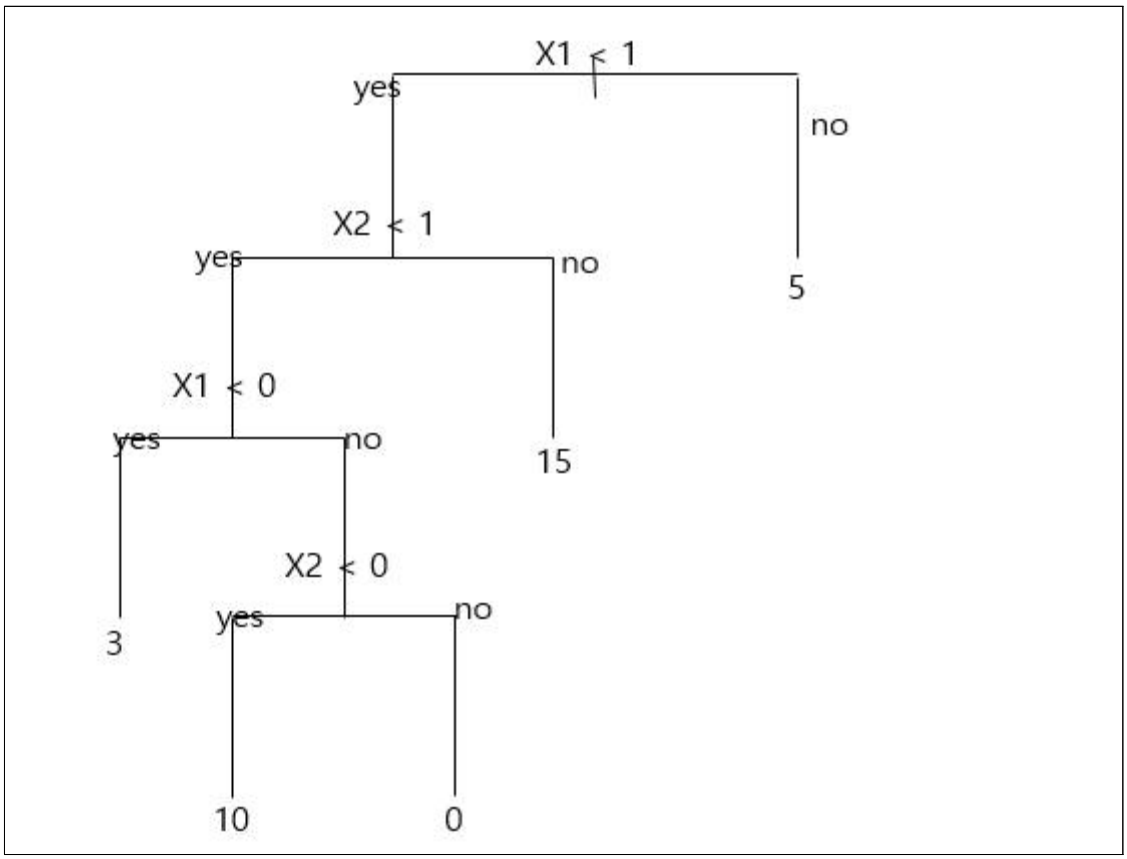
202055364 황 성 윤

Exercises for Tree

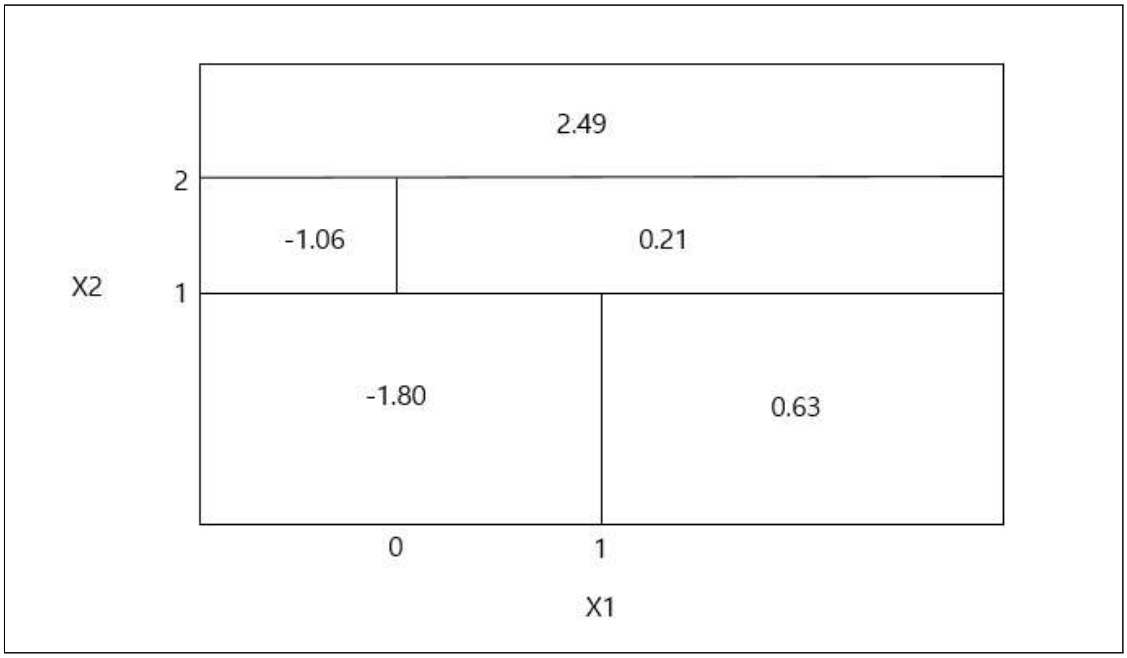
1. 다음 그림을 보고, 아래의 물음에 답하여라.



(a) 왼쪽 그림을 보고, 오른쪽과 같은 tree를 그려라. 왼쪽 그림의 박스 안의 숫자는 각 영역 내의 Y의 평균이다.



(b) 오른쪽 그림을 보고 왼쪽과 유사한 다이어그램을 생성하여라. 예측 변수 공간을 올바른 영역으로 나누고, 각 영역의 평균을 표시해야 함.



2. Tree 강의노트 p.22의 데이터를 이용하여, 불순도 측도로 Gini 지수를 사용했을 때, 첫번째 분리 규칙을 찾아라.

	age	income	student	credit_rating	buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	exellent	no
3	31...40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	exellent	no
7	31...40	low	yes	exellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	exellent	yes
12	31...40	medium	no	exellent	yes
13	31...40	high	yes	fair	yes
14	>40	medium	no	exellent	no

위의 데이터의 목적은 고객의 컴퓨터 구매 여부를 판단하고자 하는 것이다. 불순도 측도인 지니계수와 분리규칙 A에 의해서 분리된 경우에 대한 지니계수는 다음과 같이 정의되며 분리 이후 산출되는 지니계수를 가장 작게 만들어주는 분리규칙을 첫 번째 분리규칙으로 정하면 된다.

$$G(D) = 1 - \sum_{k=1}^{K} p_k^2$$

$$G_A(D) = \frac{|D_1|}{|D|} G(D_1) + \frac{|D_2|}{|D|} G(D_2)$$

각 설명변수에 따라 분리규칙을 정하고 지니계수를 구하면 다음과 같다.

1) age

$$G(age > 30) = \frac{5}{14}\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) + \frac{9}{14}\left(1 - \left(\frac{2}{9}\right)^2 - \left(\frac{7}{9}\right)^2\right) = 0.394$$

$$G(age > 40) = \frac{9}{14}\left(1 - \left(\frac{3}{9}\right)^2 - \left(\frac{6}{9}\right)^2\right) + \frac{5}{14}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) = 0.457$$

$$G(30 < age \le 40) = \frac{10}{14}\left(1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2\right) = 0.357$$

2) income

$$G(income = low) = \frac{4}{14}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{10}{14}\left(1-\left(\frac{4}{10}\right)^2-\left(\frac{6}{10}\right)^2\right)= 0.450$$

$$G(income = medium) = \frac{6}{14}\left(1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2\right)+\frac{8}{14}\left(1-\left(\frac{3}{8}\right)^2-\left(\frac{5}{8}\right)^2\right)= 0.458$$

$$G(income = high) = \frac{4}{14}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)+\frac{10}{14}\left(1-\left(\frac{3}{10}\right)^2-\left(\frac{7}{10}\right)^2\right)= 0.443$$

3) student

$$G(student = yes) = \frac{7}{14}\left(1-\left(\frac{1}{7}\right)^2-\left(\frac{6}{7}\right)^2\right)+\frac{7}{14}\left(1-\left(\frac{4}{7}\right)^2-\left(\frac{3}{7}\right)^2\right)= 0.367$$

4) credit_rating

$$G(credit\,rating = excellent) = \frac{6}{14}\left(1-\left(\frac{3}{6}\right)^2-\left(\frac{3}{6}\right)^2\right)+\frac{8}{14}\left(1-\left(\frac{2}{8}\right)^2-\left(\frac{6}{8}\right)^2\right)= 0.429$$

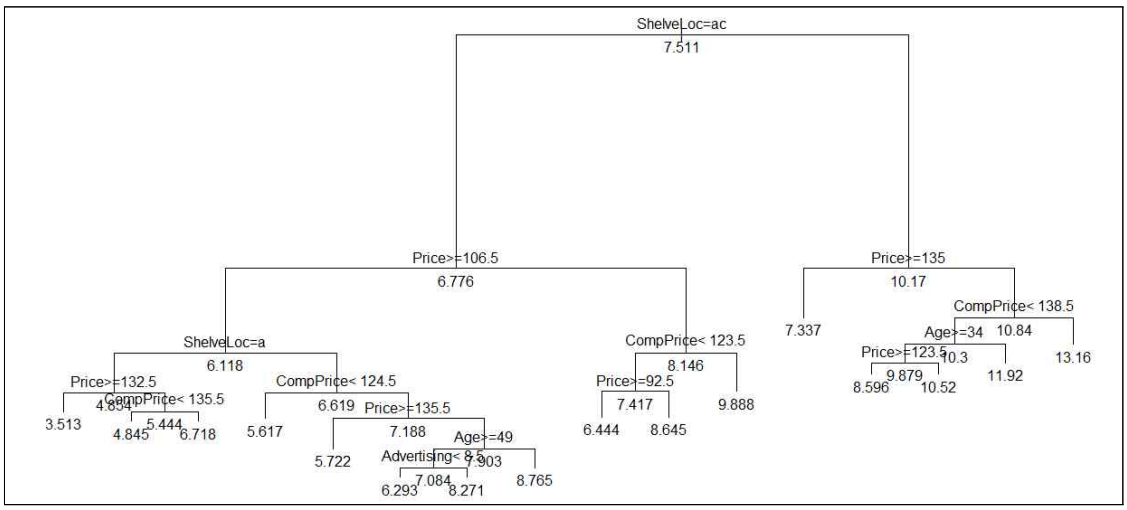
-> 결과적으로 지니계수의 값을 가장 작게 산출해주는 규칙인 $30 < age \leq 40$을 첫 번째 분리 규칙으로 정하면 된다.

3. 'Carseats.csv' 데이터를 이용하여 Sales(매출액)을 예측하려고 한다. 다음 물음에 답하여라. (변수설명 : ISLR2 패키지의 Carseats 데이터 설명 참고)

(a) 데이터를 training data (60%)와 testing data(40%)로 나누어라.
solve) 다음과 같은 R code를 이용하여 데이터를 train과 test로 나눌 수 있다. 단, 결과값이 일정하게 나오도록 하기 위해 set.seed()를 임의로 정하였음을 밝혀둔다.

```
carseats <- read.csv("C:/Users/HSY/Desktop/Carseats.csv",sep=",",header=T)
set.seed(55364)
train_id <- sample(1:nrow(carseats), nrow(carseats)*0.6)
train_dt <- carseats[train_id,]
test_dt <- carseats[-train_id,]
```

(b) training data를 이용하여 회귀트리를 적합하여라. 그림을 그리고, 결과를 설명하여라. test MSE는 얼마인가?
solve) Carseats 데이터는 서로 다른 400곳의 상점에서 어린이용 카시트가 어느 정도의 가격대에서 판매되는지를 살펴보고자 하는 것이 목적이다. 이에 따라 연속형 변수 Sales를 반응변수로 두고 회귀트리를 적합하면 다음과 같은 그림을 그릴 수 있다.

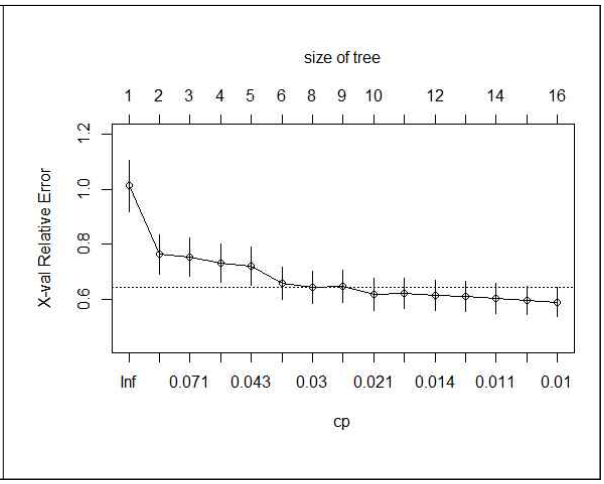


위의 그림을 통해서 봤을 때 첫 번째 분리규칙은 변수 ShelveLoc(A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site)가 Bad 또는 Medium 인지의 여부이다. (a=Bad, b=Good, c=Medium) 이를 시작으로 하여 MSE의 값이 작아지도록 하는 분리규칙에 따라 나뭇가지가 형성되었고 총 16개의 노드가 만들어졌음을 확인할 수 있다. 하지만, 그 구조가 한눈에 알아보기에는 복잡하게 구성되어 있기 때문에 가지치기를 고려해볼 필요가 있다고 판단된다. 이 회귀트리를 통하여 산출되는 test MSE의 값은 약 4.284이다.

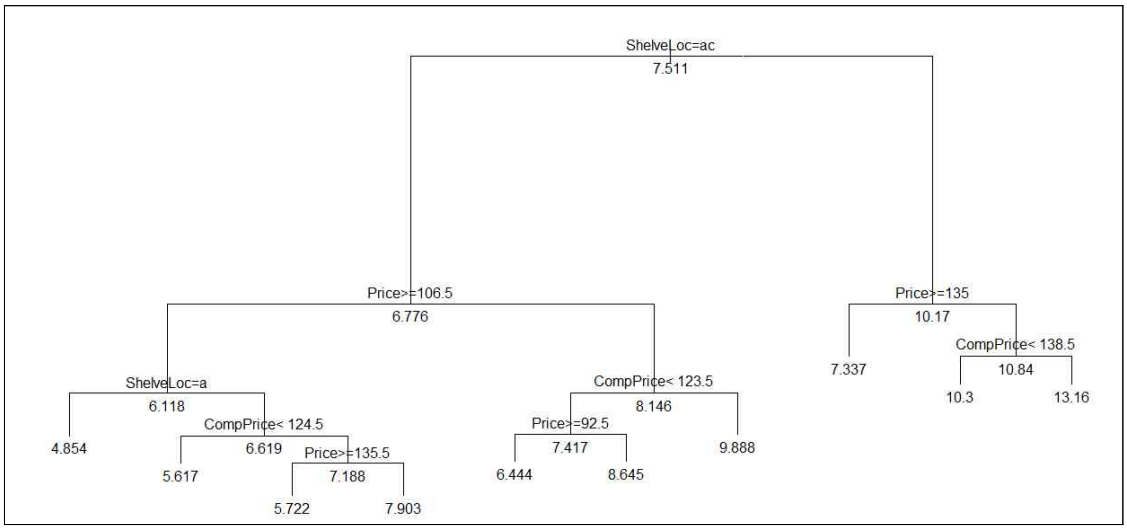
(c) tree의 complexity를 고려하여 가지치기를 시행하여라. 가지치기 시행으로 test MSE는 향상되었는가?

solve) (b)에서 적합한 tree에 대해 cptable을 작성하고 이에 대한 그래프를 그리면 다음과 같다.

	CP	nsplit	rel error	xerror	xstd
1	0.25524728	0	1.0000000	1.0116665	0.09262961
2	0.09222337	1	0.7447527	0.7642330	0.07140336
3	0.05405639	2	0.6525294	0.7552449	0.06989580
4	0.04372102	3	0.5984730	0.7331746	0.06955104
5	0.04218150	4	0.5547519	0.7218198	0.06953125
6	0.03069070	5	0.5125704	0.6594287	0.05955322
7	0.02888221	7	0.4511890	0.6450079	0.05799473
8	0.02797230	8	0.4223068	0.6473801	0.05789332
9	0.01551854	9	0.3943345	0.6183973	0.05871301
10	0.01497661	10	0.3788160	0.6225472	0.05616663
11	0.01256036	11	0.3638394	0.6162781	0.05485280
12	0.01209806	12	0.3512790	0.6124529	0.05478971
13	0.01038149	13	0.3391810	0.6042131	0.05363320
14	0.01022377	14	0.3287995	0.5967215	0.05229306
15	0.01000000	15	0.3185757	0.5914457	0.05181439



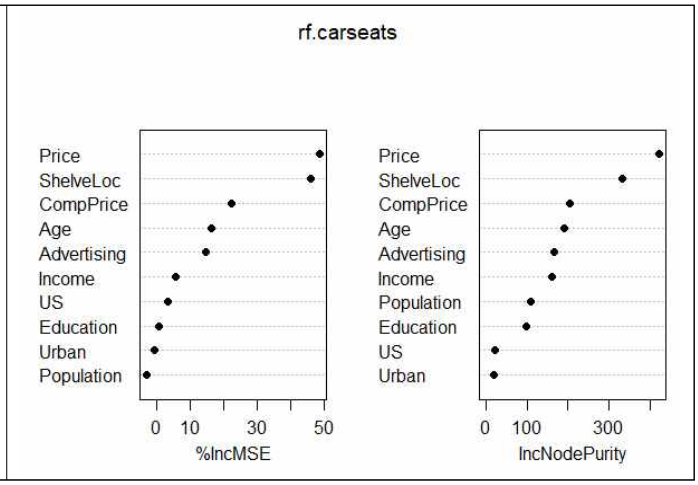
위의 결과를 바탕으로 가지치기를 위한 적절한 cp는 0.021과 0.025 사이의 약 0.023 정도로 둘 수 있겠다. 이를 이용하여 가지치기를 시행하면 다음과 같은 형태로 tree가 간소화됨을 확인할 수 있다. 그리고 이 결과를 통해서 계산되는 test MSE의 값은 약 4.363으로서 가지치기를 하기 전의 결과와 비교했을 때 향상되지는 않았지만 거의 비슷한 값임을 확인할 수 있다.



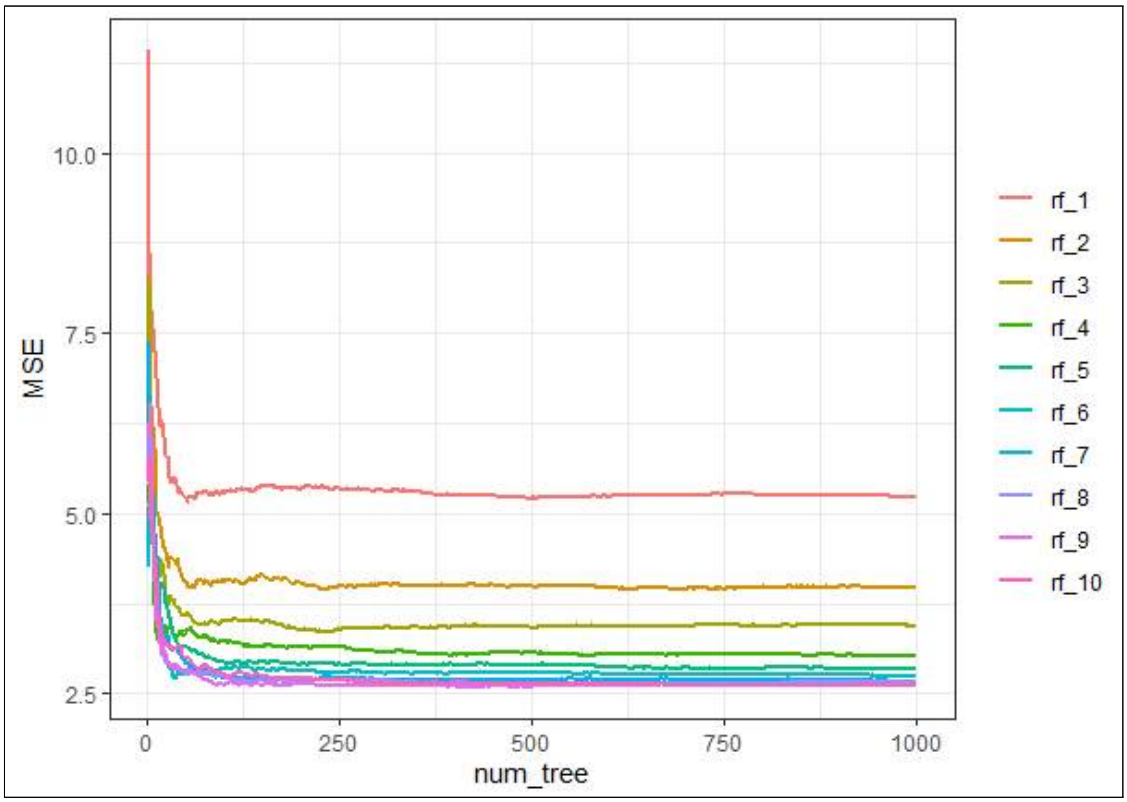
(d) Bagging을 시행하여라. test MSE는 얼마인가?
solve) train data를 이용하여 총 1000개의 bootstrap sample을 추출하고 out-of-bag 방법을 통하여 Bagging을 실시하고 이에 대하여 test MSE를 계산하면 약 2.800이다.

(e) Random Forest를 시행하여라. test MSE는 얼마인가? 변수중요도를 구하여라. $m = 1,10$을 포함하여 m값을 변화시키면서 test MSE를 비교하고, 이 때 m의 효과를 설명하여라.
solve) train data를 이용하여 총 1000개의 bootstrap sample을 추출하고 각 sample마다 임의로 3개의 설명변수를 선택한 뒤 out-of-bag 방법을 통하여 Random Forest를 실시하고 이에 대하여 test MSE를 계산하면 약 3.094이다. 변수중요도를 계산하고 이를 그래프로 표현하면 다음과 같다.

	%IncMSE	IncNodePurity
CompPrice	22.2012803	204.61976
Income	5.8185013	160.71919
Advertising	14.7926121	165.90464
Population	-2.8468776	110.65859
Price	48.5660725	423.31730
ShelveLoc	45.7920126	332.65570
Age	16.3912221	191.90495
Education	0.9131834	97.54457
Urban	-0.4082852	18.73053
US	3.3323909	21.55804



위의 결과를 통해 변수 ShelveLoc과 Price가 변수중요도가 높다는 것을 확인할 수 있다. 그리고 각 bootstrap sample마다 임의로 선택하는 설명변수의 개수와 생성하는 tree의 개수에 따라 test MSE를 계산하고 이를 그래프로 표현하면 다음과 같다.



위의 그래프를 통해서 bootstrap sample에서 몇 개의 설명변수를 선택하는지에 따라 test MSE의 값이 달라짐을 확인할 수 있다. 그러므로 Random Forest를 시행할 때 뽑히는 설명변수의 개수를 적절하게 설정할 필요가 있다고 할 수 있겠다. 다만, 본 예제에서는 특이하게 모든 설명변수를 다 사용하는 Bagging이 Random Forest보다 test MSE의 값이 더 작게 산출되었다. 하지만, 일반적으로는 임의의 설명변수를 선택하는 Random Forest가 Bagging에 비해 bootstrap sample의 연관성을 줄여주기 때문에 Random Forest의 성능이 더 좋게 나오게 된다.

4. 'OJ.csv' 데이터를 이용하여, 다음 물음에 답하여라. (변수설명 : ISLR2 패키지의 OJ 데이터 설명 참고)

(a) 800개의 관측값으로 구성된 training data를 만들고, 나머지를 포함하는 testing data를 구성하여라.
solve) 다음과 같은 R code를 이용하여 데이터를 train과 test로 나눌 수 있다. 단, 결과값이 일정하게 나오도록 하기 위해 set.seed()를 임의로 정하였음을 밝혀둔다. OJ 데이터는 오렌지 주스와 관련한 데이터로서 Citrus Hill와 Minute Maid 중 어떤 오렌지주스를 선택하는지 알아보는 것이 주된 목적이다.

```
oj <- read.csv("C:/Users/HSY/Desktop/OJ.csv",sep=",",header=T)
set.seed(55364)
train_id <- sample(1:nrow(oj), 800)
train_dt <- oj[train_id,]
test_dt <- oj[-train_id,]
```

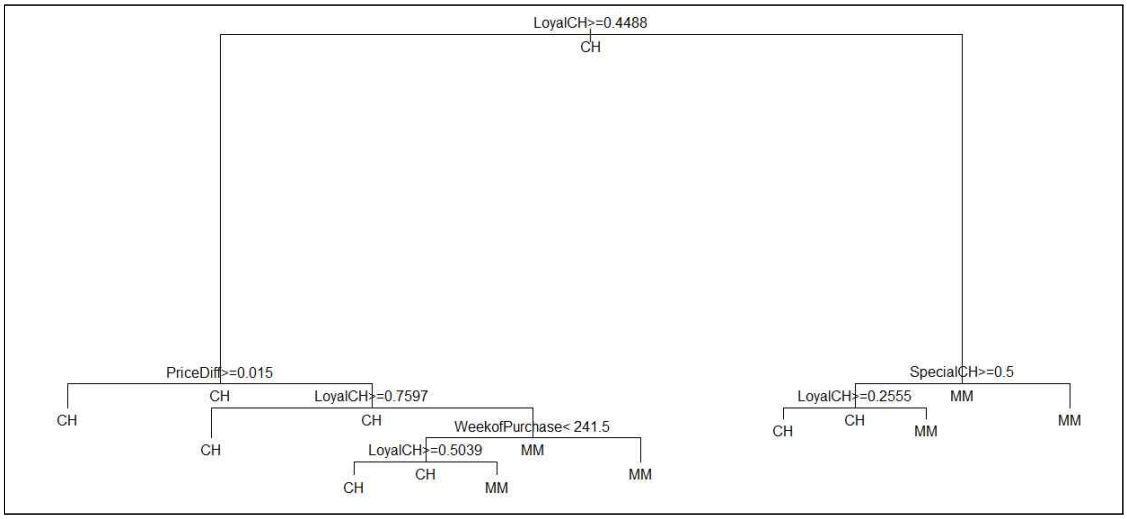
(b) training data를 이용하여 반응변수를 Purchase로 하는 tree를 적합하여라. 결과를 설명하여라. training error는 얼마인가? terminal node의 갯수는 몇 개인가?
solve) training data를 이용하여 tree를 적합하면 다음과 같은 결과를 얻을 수 있다.

```
n= 800

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 800 310 CH (0.61250000 0.38750000)
   2) LoyalCH>=0.4488 517  86 CH (0.83365571 0.16634429)
     4) PriceDiff>=0.015 384  34 CH (0.91145833 0.08854167) *
     5) PriceDiff< 0.015 133  52 CH (0.60902256 0.39097744)
      10) LoyalCH>=0.7596895 49   3 CH (0.93877551 0.06122449) *
      11) LoyalCH< 0.7596895 84  35 MM (0.41666667 0.58333333)
        22) WeekofPurchase< 241.5 43  18 CH (0.58139535 0.41860465)
          44) LoyalCH>=0.5039495 29   8 CH (0.72413793 0.27586207) *
          45) LoyalCH< 0.5039495 14   4 MM (0.28571429 0.71428571) *
        23) WeekofPurchase>=241.5 41  10 MM (0.24390244 0.75609756) *
   3) LoyalCH< 0.4488 283  59 MM (0.20848057 0.79151943)
     6) SpecialCH>=0.5 23   9 CH (0.60869565 0.39130435)
      12) LoyalCH>=0.2554975 14   2 CH (0.85714286 0.14285714) *
      13) LoyalCH< 0.2554975 9   2 MM (0.22222222 0.77777778) *
     7) SpecialCH< 0.5 260  45 MM (0.17307692 0.82692308) *
```

결과적으로 총 4개의 설명변수가 tree를 적합하는 데 사용되었으며, terminal node의 개수는 8이다. 예측이 아닌 분류와 관련한 문제이기 때문에 training error는 결국 training 오분류율을 의미한다. 이 tree를 사용하여 training data에 대한 confusion matrix를 다음과 같이 적을 수 있고, 오분류율의 값은 약 0.135이다.

```
yhat  CH  MM
  CH 429  47
  MM  61 263
```

(c) tree의 그림을 그리고 결과를 설명하여라.
solve) (b)에서 적합한 tree를 그림으로 표현하면 다음과 같다.



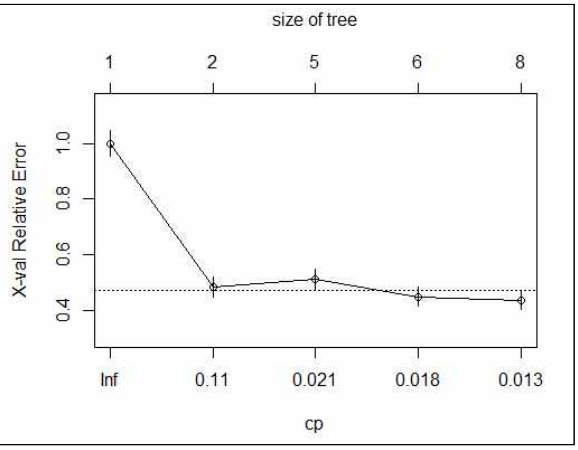
결과적으로 불순도 측도가 얼마나 줄어드는지를 바탕으로 하여 변수 LoyalCH의 값이 0.4488 이상인지의 여부를 첫 번째 분류규칙으로 두고 tree가 생성되고 있음을 확인할 수 있다.

(d) testing data를 이용하여 예측을 하고, confusion matrix를 생성하여라. test error은 얼마인가?
solve) testing data를 이용하여 (b)에서 생성한 tree를 이용하여 예측을 하고 confusion matrix를 적으면 다음과 같다. 이에 대한 test error, 측 test 오분류율은 약 0.207이다.

yhat	CH	MM
CH	136	29
MM	27	78

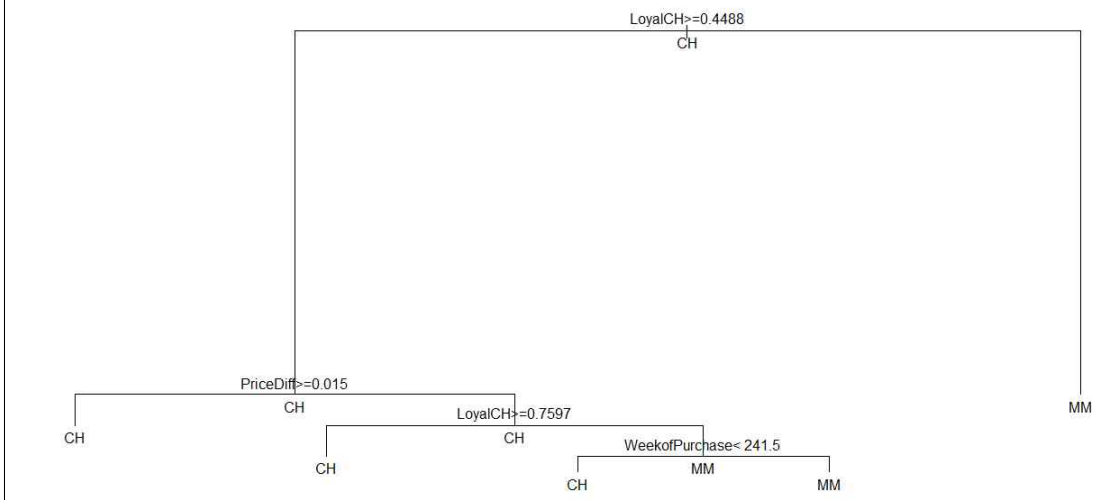
(e) cptable 및 그림을 이용하여 최적의 tree를 구하여라.
solve) (b)에서 적합한 tree에 대하여 cptable과 이에 대한 그림을 살펴보면 다음과 같다.

	CP	nsplit	rel error	xerror	xstd
1	0.532258	0	1.00000	1.00000	0.044450
2	0.022581	1	0.46774	0.48387	0.035612
3	0.019355	4	0.40000	0.50968	0.036324
4	0.016129	5	0.38065	0.44839	0.034570
5	0.010000	7	0.34839	0.43548	0.034172



위의 결과를 바탕으로 가지치기를 위한 적절한 cp의 값은 0.018과 0.021 사이에 있는 약 0.02 정도로 설정하도록 한다. 이를 바탕으로 새로운 tree를 적합하면 다음과 같다.

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 800 310 CH (0.61250000 0.38750000)
   2) LoyalCH>=0.4488 517  86 CH (0.83365571 0.16634429)
     4) PriceDiff>=0.015 384  34 CH (0.91145833 0.08854167) *
     5) PriceDiff< 0.015 133  52 CH (0.60902256 0.39097744)
      10) LoyalCH>=0.7596895 49   3 CH (0.93877551 0.06122449) *
      11) LoyalCH< 0.7596895 84  35 MM (0.41666667 0.58333333)
        22) WeekofPurchase< 241.5 43  18 CH (0.58139535 0.41860465) *
        23) WeekofPurchase>=241.5 41  10 MM (0.24390244 0.75609756) *
   3) LoyalCH< 0.4488 283  59 MM (0.20848057 0.79151943) *
```



결과적으로 가지치기를 통해 terminal node의 개수가 5인 tree가 적합되었다.

(f) 가지치기 전/후 tree의 training error를 비교하여라.
solve)

	가지치기 전	가지치기 후
confusion matrix	yhat CH MM CH 429 47 MM 61 263	yhat CH MM CH 421 55 MM 69 255
misclassification rate	0.135	0.155

training error의 경우는 가지치기 후의 경우가 전보다 더 높음을 알 수 있다. 하지만 이것은 실제 tree를 적합한 training data에 대해 예측을 한 경우이므로 과적합의 오류가 발생할 수도 있는 만큼 test data에 대해서는 어떤 결과가 나오는 지 살펴볼 필요가 있다.

(g) 가지치기 전/후 tree의 test error를 비교하여라.
solve)

	가지치기 전	가지치기 후
confusion matrix	yhat CH MM CH 136 29 MM 27 78	yhat CH MM CH 136 29 MM 27 78
misclassification rate	0.207	0.207

test error를 비교하면 가지치기 전과 후의 결과가 동일하다. 다시 말해서 복잡한 tree를 적합할 필요 없이 가지치기를 통해 단순화된 tree를 가지고 보다 더 효율적으로 예측이 가능하다는 것을 말해주는 예시라고 할 수 있겠다.