Homework 8

Ben Buzzee

November 20, 2019

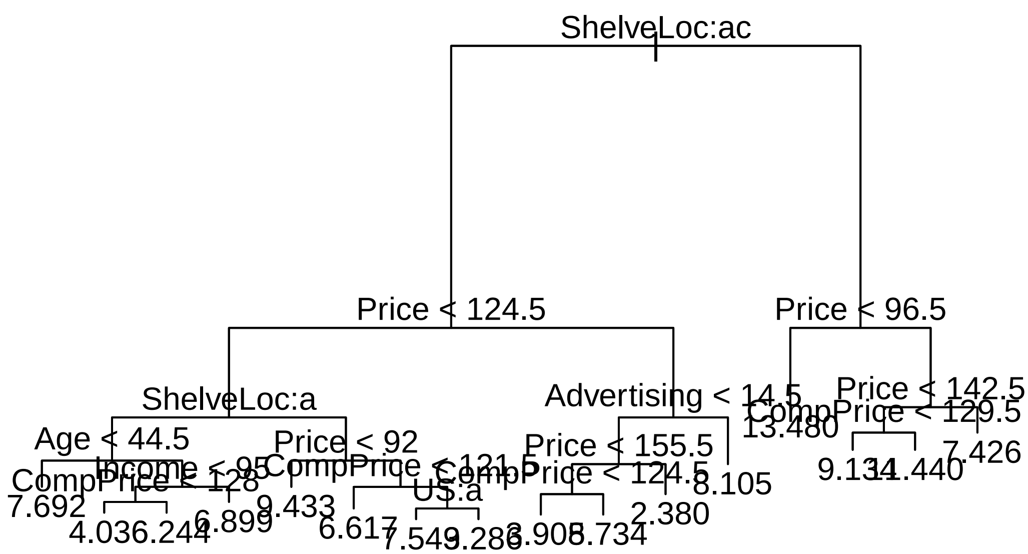
1.

(a) Fit Tree

First we will split the data in half and fit an unrestricted tree on the training set.

```
train_sample <- sample(1:400, 200)
cars.train <- Carseats[train_sample,]
cars.test <- Carseats[-train_sample,]
```

```
tree.cars <- tree(Sales~., data = cars.train)
plot(tree.cars)
text(tree.cars)
```




```
summary(tree.cars)
```

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = cars.train)
## Variables actually used in tree construction:
```

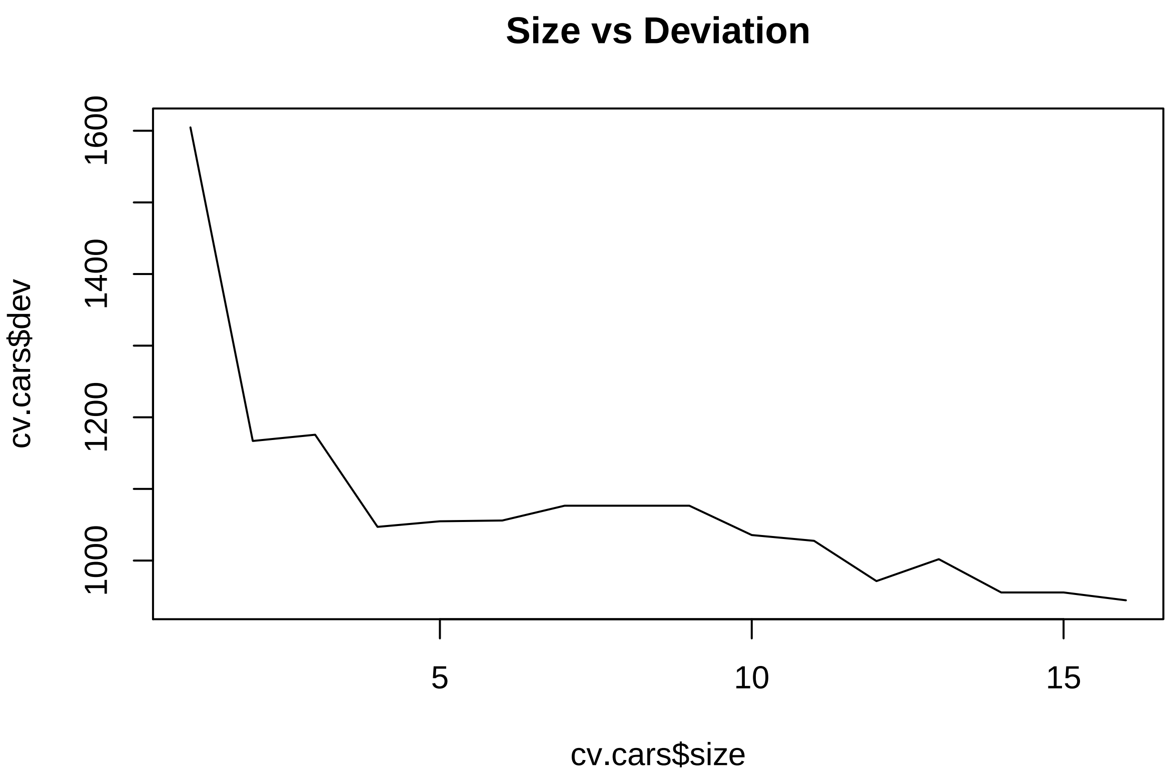
```
## [1] "ShelveLoc"   "Price"       "Age"         "Income"      "CompPrice"
## [6] "US"          "Advertising"
## Number of terminal nodes:  16
## Residual mean deviance:  2.085 = 383.7 / 184
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.7480 -1.0130  0.0290  0.0000  0.9575  3.9270
```

We see that our residual mean deviance is 2.015. There are also 19 terminal nodes which might be indicative of overfitting.

(b) Size by CV

Next we'll find the optimal tree size using cross validation.

```r
cv.cars <- cv.tree(tree.cars)
plot(cv.cars$size, cv.cars$dev, main = "Size vs Deviation", type = "l")
```



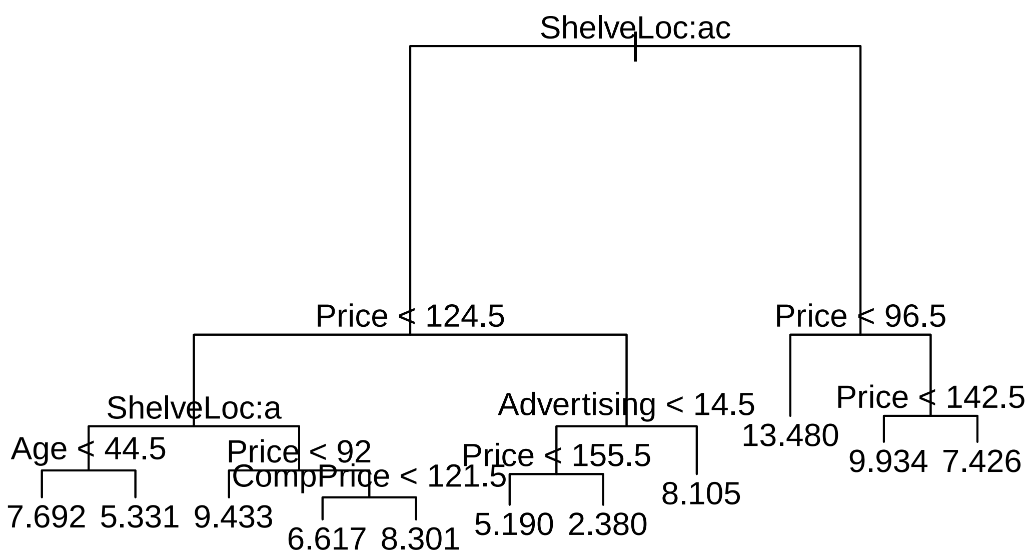
Our CV error seems to decrease as size increases all the way up to a size of 19. So pruning is probably not necessary, but we will try it.

(c) Pruning

```r
prune.cars <- prune.tree(tree.cars, best = 11)
summary(prune.cars)
```

```
##
## Regression tree:
## snip.tree(tree = tree.cars, nodes = c(39L, 17L, 14L, 20L))
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "CompPrice"   "Advertising"
## Number of terminal nodes:  11
## Residual mean deviance:  2.66 = 502.8 / 189
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -5.03000 -1.02600 -0.06841  0.00000  0.89240  3.92700
```

```
plot(prune.cars)
text(prune.cars)
```

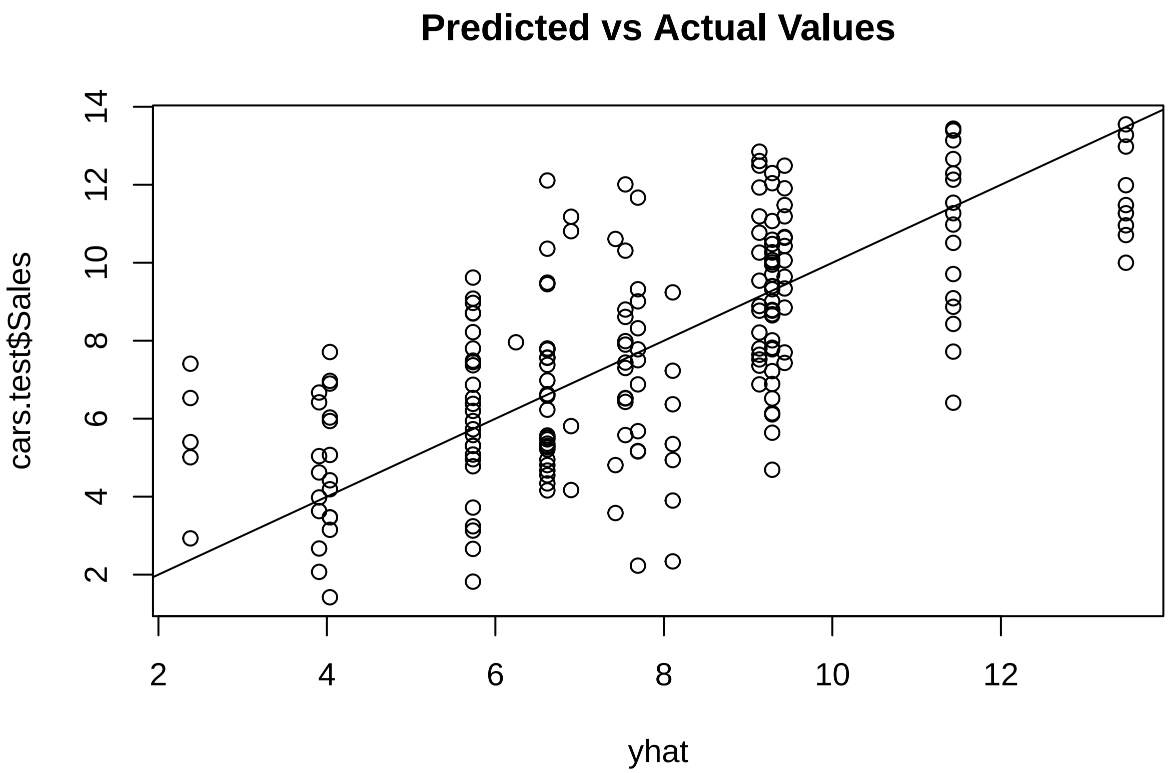


We see that pruning does not improve the MSE of our model. It does however make our model simpler and more interpretable.

(d) Predicted Sales

Next we will use our fitted model to predict sales in the test dataset.

```
yhat = predict(tree.cars, newdata=cars.test)
plot(yhat, cars.test$Sales, main = "Predicted vs Actual Values")
abline(0,1)
```



And our prediction MSE is:

```
mean((yhat-cars.test$Sales)^2)
```

```
## [1] 4.663227
```

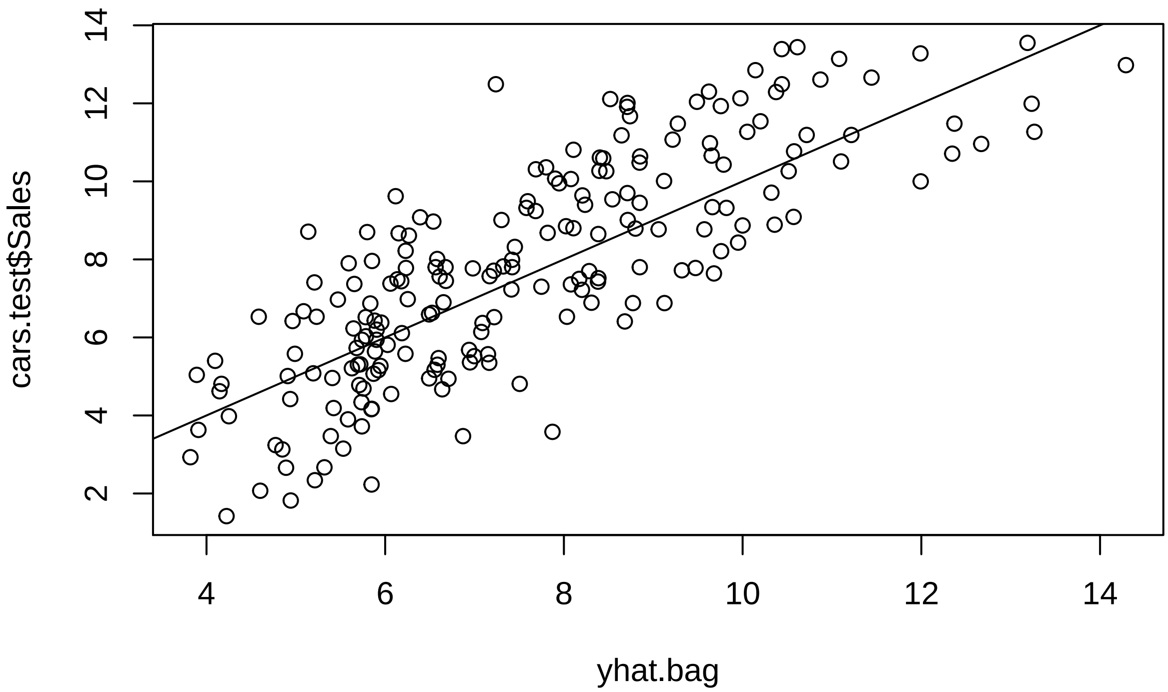
Which is roughly double our test MSE.

(e) Bagging

```
bag.cars <- randomForest(Sales~., data = cars.train, mtry = 10, importance = TRUE)
bag.cars
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = cars.train, mtry = 10,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 2.781318
##                     % Var explained: 64.74
```

```
yhat.bag <-  predict(bag.cars, newdata=cars.test)
plot(yhat.bag, cars.test$Sales)
abline(0,1)
```




```r
mean((yhat.bag-cars.test$Sales)^2)
```

```
## [1] 2.868836
```

With a prediction MSE of 2.9, bagging is a significant improvement over the unrestricted regression tree. The % of variance explained is 64.31 which leaves room for improvement.

```r
round(importance(bag.cars), 2)
```

```
##             %IncMSE IncNodePurity
## CompPrice     20.75        138.13
## Income         6.77         72.43
## Advertising   18.79         96.46
## Population     0.58         50.35
## Price         56.79        501.04
## ShelveLoc     60.54        500.14
## Age           17.42        147.69
## Education     -2.22         34.60
## Urban         -0.71          4.73
## US             1.93          4.55
```

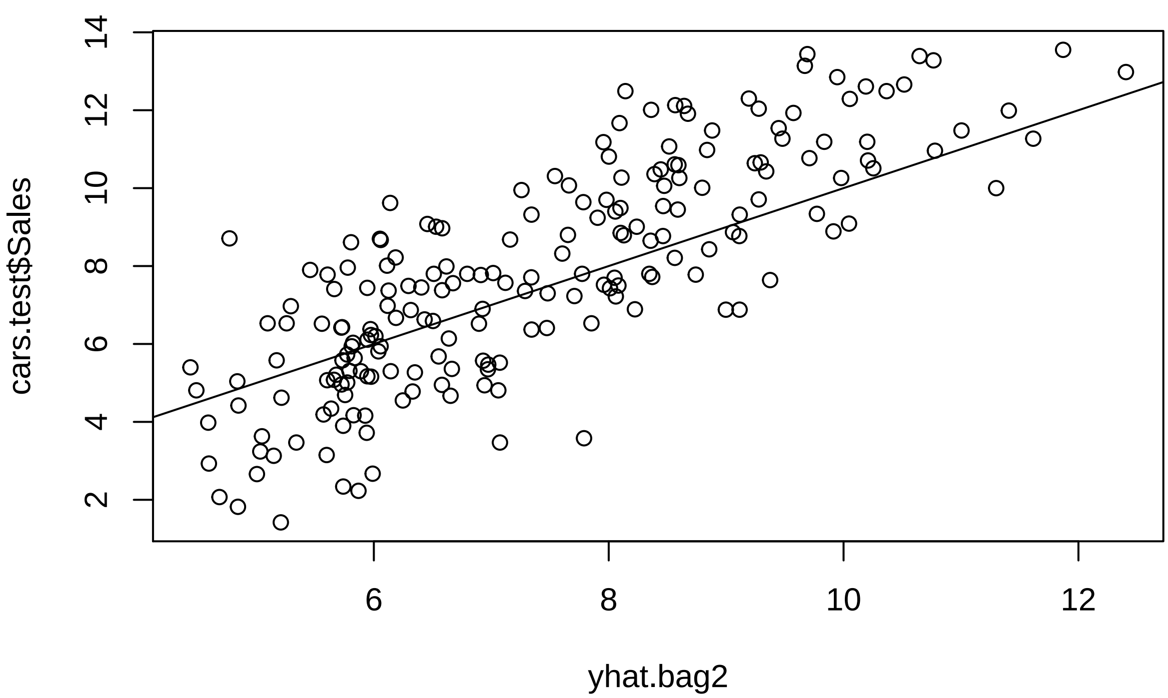
From the above importance metrics we see that Price and ShelveLoc are the most important predictors of unit Sales.

(f) Random Forest

```
bag.cars2 <- randomForest(Sales~., data = cars.train, mtry = 4, importance = TRUE)
bag.cars2
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = cars.train, mtry = 4,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 3.020153
##                     % Var explained: 61.72
```

```
yhat.bag2 <-  predict(bag.cars2, newdata=cars.test)
plot(yhat.bag2, cars.test$Sales)
abline(0,1)
```




```
mean((yhat.bag2-cars.test$Sales)^2)
```

```
## [1] 3.143459
```

With a prediction MSE of 3.3 and 61.6 % of the variance explained, our random forest predictions are not an improvement over bagging.

Finally, we still find that price and shelving location are the most important predictors of unit sales.

```r
round(importance(bag.cars2), 2)
```

```
##             %IncMSE IncNodePurity
## CompPrice     10.41        120.76
## Income         2.47        118.95
## Advertising   14.24        124.91
## Population    -0.38         85.83
## Price         38.79        407.19
## ShelveLoc     44.27        398.00
## Age           12.49        172.63
## Education     -1.75         58.71
## Urban         -1.90          9.81
## US            -0.08         12.41
```