HW2 #6

Philip Lee (phl6cw)

9/22/2019

##Problem 6

#6a

```
library(ISLR)
Auto<-Auto
attach(Auto)
medmpg<-median(mpg)  ##gets median mpg
hillow<-ifelse(Auto$mpg > medmpg,1,0) ##makes binary variable hillow
Auto$hilow<-as.factor(hillow) ##converts hillow to hilow and into factor
```

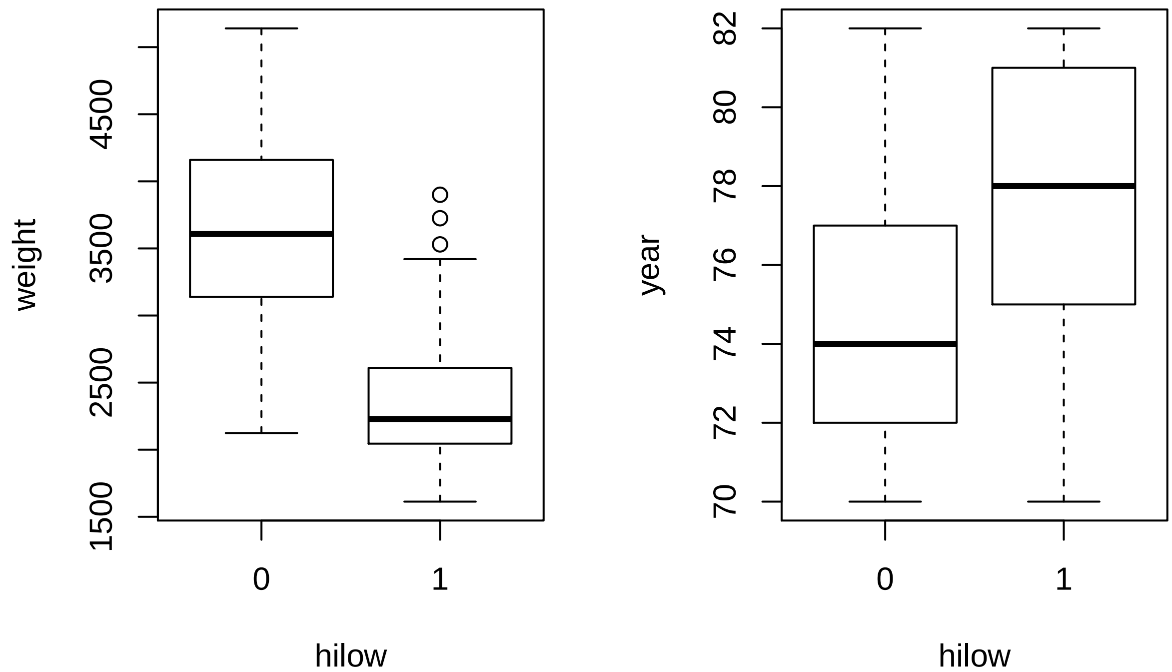
#6b We want to remove the variable mpg as a predictor since the response variable is based on the predictor mpg, which will result in a perfect correlation if it is left in the model.

#6c

```
attach(Auto)
```

```
## The following objects are masked from Auto (pos = 3):
##
##     acceleration, cylinders, displacement, horsepower, mpg, name,
##     origin, weight, year
```

```
par(mfrow=c(1,2))
boxplot(weight~hilow)
boxplot(year~hilow)
```



The median weight for the automobiles with high gas mileage (automobiles with mpg greater than the median mpg) is much smaller than the automobiles with low gas mileage (automobiles with mpg less than the median mpg). The range for the weight is greater for low gas mileage vehicles. The five-number summary for the low gas mileage vehicles are greater than the high gas mileage's for all of the values. There are three outliers for the weight of high gas mileage vehicles, but these outliers appear to be close in value of each other. In essence, the weight boxplot tells us that the lighter vehicles have higher gas mileage. The boxplot of the year tells us that the newer vehicles tend to have higher gas mileage.

#6d

```
set.seed(18735)
sample.data<-sample.int(nrow(Auto), floor(.50*nrow(Auto)), replace = F) ##splits data into two
train<-Auto[sample.data, ] ##makes train data
test<-Auto[-sample.data, ]  ##makes test data
result_train<-glm(hilow~weight+year, family=binomial, data=train) ##fits logistic regression based on t
preds<-predict(result_train,newdata=test, type="response") ##gets predicted values
```

#6e

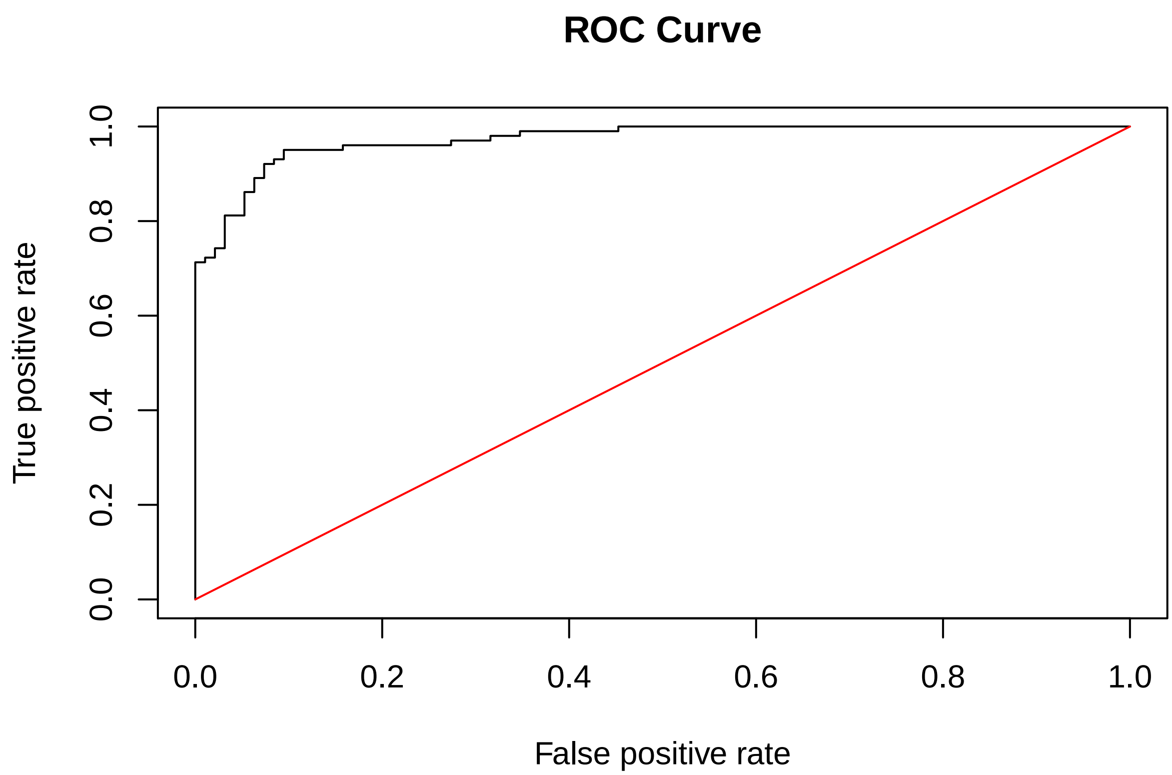
```
library(ROCR)
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
rates<-prediction(preds, test$hilow)  ##produce the numbers associated with classification table
roc_result<-performance(rates,measure="tpr", x.measure="fpr") ##store the true positive and false posi
plot(roc_result, main="ROC Curve") ##plot ROC curve and overlay the diagonal line for random guessing
lines(x = c(0,1), y = c(0,1), col="red")
```




```r
auc<-performance(rates, measure = "auc") ##gets AUC value
print(auc@y.values)
```

```
## [[1]]
## [1] 0.9725899
```

Since the ROC curve is very far away from the diagonal and is close to (0,1), it indicates that the model has a very good predictive ability and performs better than random guessing as it also has AUC of 0.9726.

#6f

```r
confusion.mat<-table(test$hilow,preds > 0.5)  ##creates confusion matrix
overall.error<- (confusion.mat[1,2] + confusion.mat[2,1]) /sum(confusion.mat)  ##gets overall error
print(overall.error)
```

```
## [1] 0.08163265
```

Overall error rate for the test data is 0.0816.

#6g

```r
false_pos_num<-confusion.mat[1,2]     ##numerator for false positive rate
false_pos_den<-sum(confusion.mat[1,])  ##denominator for false positive rate
false_pos_rate<-false_pos_num/false_pos_den  ##false positive rate
```

```r
false_neg_num<-confusion.mat[2,1]  ##numerator for false negative rate
false_neg_den<-sum(confusion.mat[2,])  ##denominator for false negative rate
false_neg_rate<-false_neg_num/false_neg_den ##false negative rate
print(false_pos_num)
```

```
## [1] 9
```

```r
print(false_pos_den)
```

```
## [1] 95
```

```r
print(false_pos_rate)
```

```
## [1] 0.09473684
```

```r
print(false_neg_num)
```

```
## [1] 7
```

```r
print(false_neg_den)
```

```
## [1] 101
```

```r
print(false_neg_rate)
```

```
## [1] 0.06930693
```

The false positive rate for test data is 0.0947. The false negative rate for teste data is 0.06931. The numbers that were used for getting the rates is commented along with the code.

#6h The threshold should be lowered. Since we are concerned with failing to identify cars that have high gas mileage, we would want the classification to have lower false negatives. Reducing the threshold reduces the number of classifications that would fall into false negatives.

#6i If the threshold is lowered, then the false negative rate will decrease while the false positive rate will increase.