Homework 4

Peyton Kuhlers

10/15/2020

Question 1

In order to run a glm, I assigned male to 0 and female to 1. I also randomly assigned a training or test set label. I then performed a logistic regression with height and weight as predictors. The accuracy was 0.52 in the test set.

```
set.seed(111)
dat$Gender_num <- ifelse(dat$Gender == "Male", 0, 1)

subsamp <- c(rep("Train",450), rep("Test",50))
scramble <- sample(subsamp, size = 500, replace = F)

dat$set <- scramble

train <- dat[dat$set == "Train",]
test <- dat[dat$set == "Test",]

mod1 <- glm(formula = Gender_num ~ Height + Weight,
            family = "binomial",
            data = train)

summary(mod1)
```

```
##
## Call:
## glm(formula = Gender_num ~ Height + Weight, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.219  -1.192   1.140   1.161   1.188
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.0897359  1.0363879  -0.087    0.931
## Height       0.0010918  0.0057925   0.188    0.851
## Weight      -0.0005665  0.0029305  -0.193    0.847
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 623.69  on 449  degrees of freedom
## Residual deviance: 623.62  on 447  degrees of freedom
## AIC: 629.62
```

```
##
## Number of Fisher Scoring iterations: 3
```

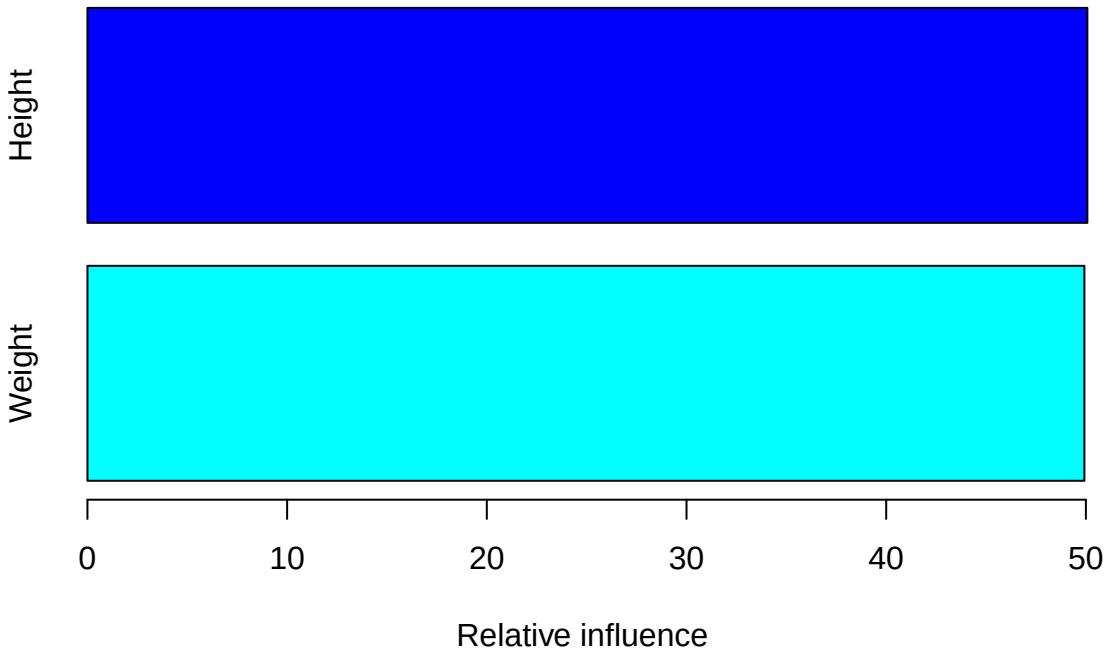
```
preds1 <- predict(object = mod1,
                  newdata = test,
                  type = "response")
sum((preds1 > 0.5) == test$Gender_num) /nrow(test)
```

```
## [1] 0.52
```

Question 2

I then used a GBM on the train set and then accuracy rose to 0.62 in the test set.

```
mod2 <- gbm(formula = Gender_num ~ Height + Weight,
            distribution = "bernoulli",
            data = train)
summary(mod2)
```




```
##           var  rel.inf
## Height Height 50.07467
## Weight Weight 49.92533
```

```
preds2 <- predict(object = mod2,
                  newdata = test,
                  type = "response")
```

```
## Using 100 trees...
```

```
sum((preds2 > 0.5) == test$Gender_num) / nrow(test)
```

```
## [1] 0.62
```

Question 3

Here I filtered the data to only contain 50 males (and all of the females). I then fit a logistic regression model to the new data. Every sample was classified as female – which actually caused an error in the F1 score calculation as all of the classifications were the same.

```
datMale <- dat %>%
  group_by(Gender) %>%
  filter(Gender == "Male") %>%
  slice_sample(n = 50, replace = F)

filtDat <- dat %>%
  group_by(Gender) %>%
  filter(Gender == "Female") %>%
  bind_rows(., datMale)

mod3 <- glm(Gender_num ~ Height + Weight,
            family = "binomial",
            data = filtDat)
summary(mod3)
```

```
##
## Call:
## glm(formula = Gender_num ~ Height + Weight, family = "binomial",
##     data = filtDat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0230   0.5562   0.5904   0.6157   0.6660
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.877619   1.719439   1.674   0.0942 .
## Height      -0.006380   0.009634  -0.662   0.5078
## Weight      -0.001472   0.004750  -0.310   0.7567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 272.14  on 304  degrees of freedom
## Residual deviance: 271.59  on 302  degrees of freedom
## AIC: 277.59
##
## Number of Fisher Scoring iterations: 4
```

```
preds3 <- predict(mod3, type = "response")
```

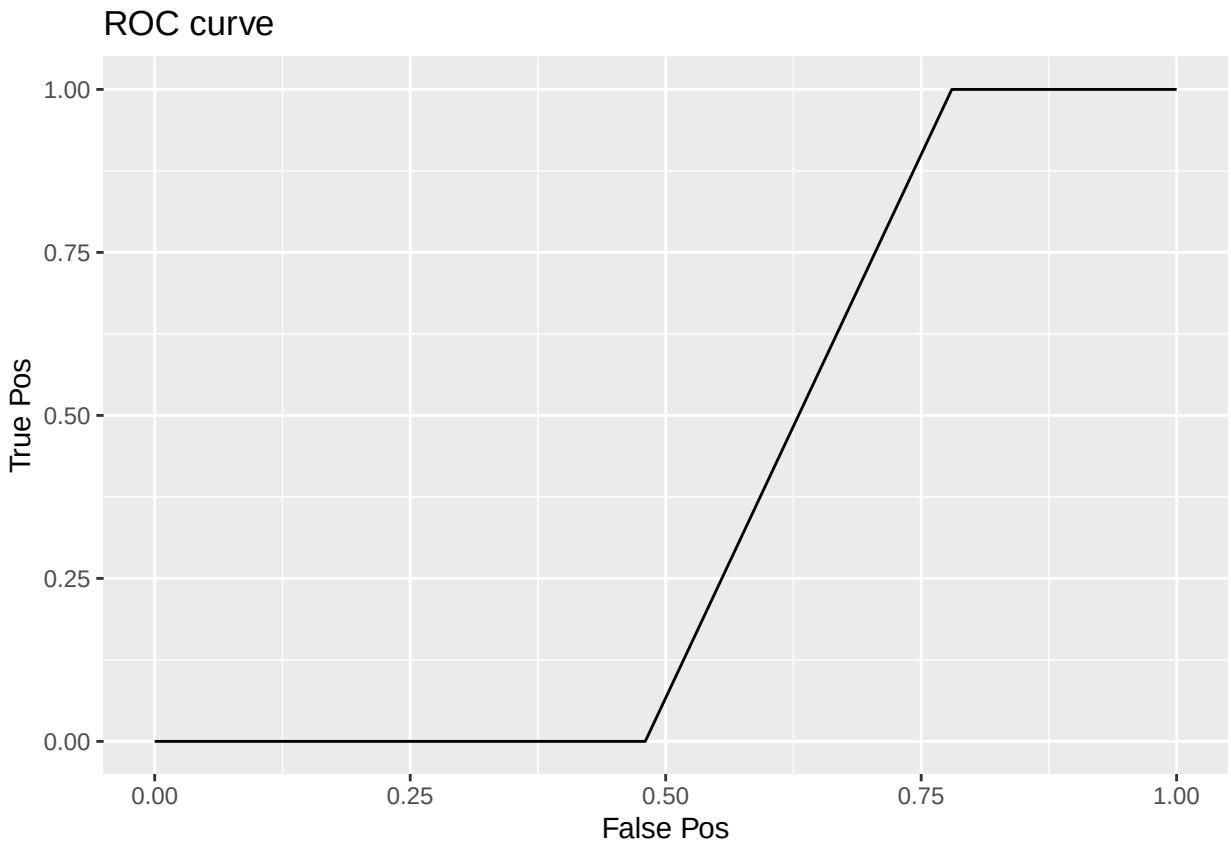
Question 4

The ROC curve shows that the best we can do with this model is about a 35% false positive rate, with a perfect true positive rate.

```
roc <- do.call(rbind, Map(function(threshold){
  p <- preds3 > threshold;
  tp <- sum(p[filtDat$Gender_num])/sum(filtDat$Gender_num);
```

```
  fp <- sum(p[!filtDat$Gender_num])/sum(!filtDat$Gender_num);
  tibble(threshold=threshold,
         tp=tp,
         fp=fp)
},seq(100)/100))

ggplot(data = roc, aes(x = fp, y = tp)) + geom_line() +
  xlim(0,1) + ylim(0,1) +
  xlab("False Pos") + ylab("True Pos") + labs(title = "ROC curve")
```

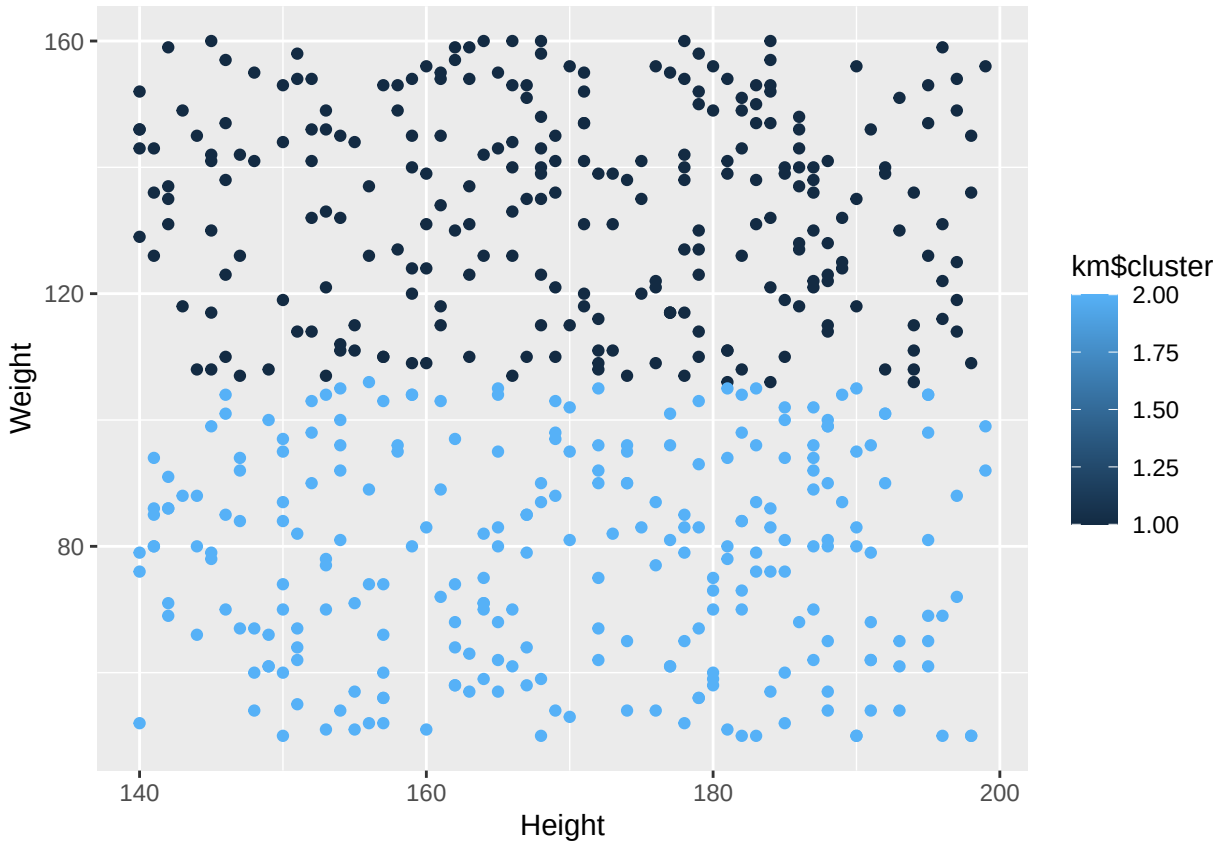


Question 5

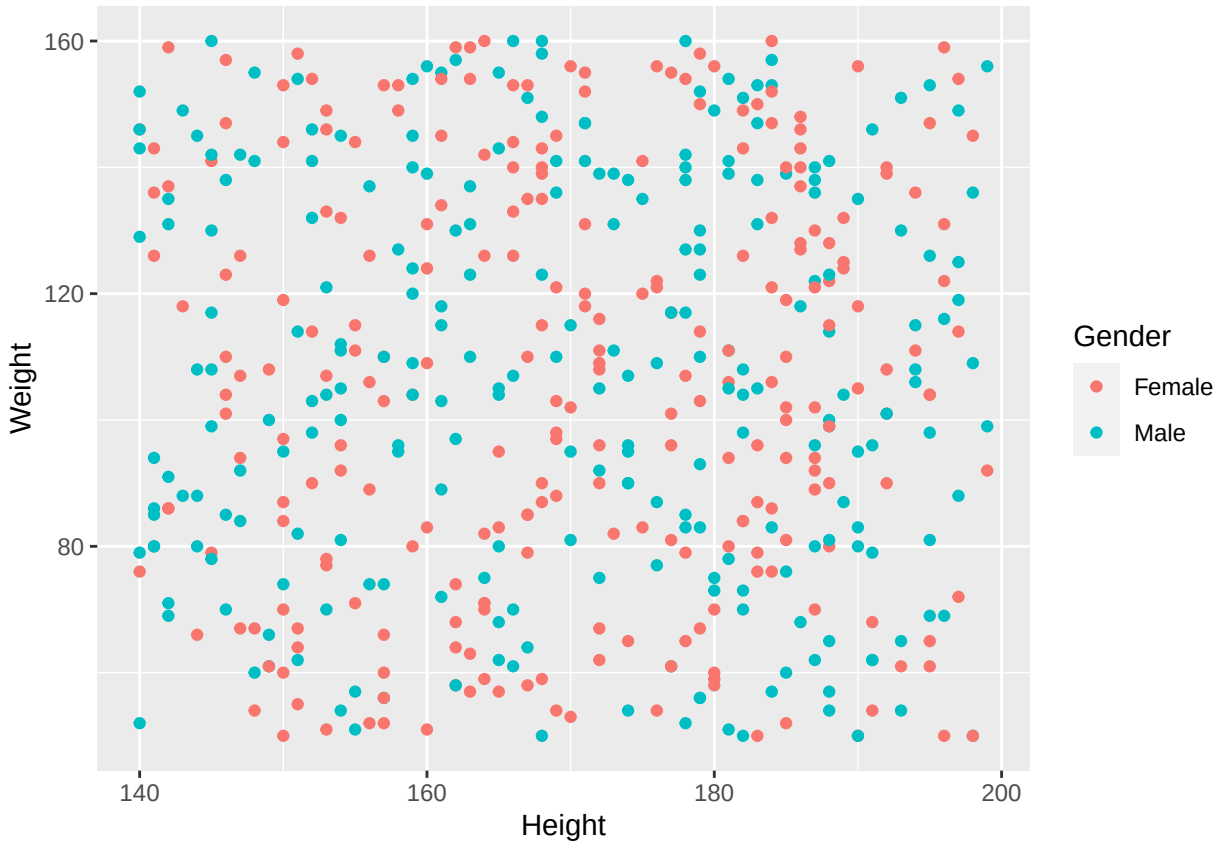
The k-means graph separates the upper half of the data from the lower half of the data. However, looking at a labeled graph of height vs weight it is clear that the two clusters do not represent clusters of males a females. They seem to just separate big and tall people from small and short people.

```
km <- kmeans(data.frame(dat$Height, dat$Weight),centers = 2)

ggplot(dat, aes(x = Height, y = Weight, col = km$cluster)) + geom_point()
```




```
ggplot(dat, aes(x = Height, y = Weight, col = Gender)) + geom_point()
```


160
120
80
Weight
140
160
180
200
Height
Gender
Female
Male