HW2

Duc Le

11/11/2020

```
library(foreign)
library(ROSE)
```

Loading prerequisite libraries

```
## Loaded ROSE 0.0-3
```

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data = read.dta("wcgs.dta")

data$chd69 <- ifelse(data$chd69=="Yes", 1, 0)

x = data %>%
  select(-c(typchd69, id, t1, agec, wghtcat))

x = na.omit(x)
attach(x)
```

Data Processing + Cleaning

```
base = glm(chd69~., data = x, family = "binomial")
final1 = stepAIC(base, trace = 0)
```

```r
# Model with interactions
final2 = stepAIC(base, ~.^2, trace = 0)

# Acquiring the predicted probabilities for each model
pp1 = predict(final1, type = "response")
pp2 = predict(final2, type = "response")
```

Running the base logistic model I personally prefer this ROC package because the x-axis (False Positive Rate) is displayed better & not in reverse. However RMarkdown isn't printing the ROC curves produced by this package therefore I won't include them in this RMarkdown.

```r
hoslem.test(chd69, fitted(final1))
```

Goodness of Fit Test w/ Hosmer-Limeshaw

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  chd69, fitted(final1)
## X-squared = 23.005, df = 8, p-value = 0.003358
```

The p.value is less than 0.05 thus indicating that this model did not pass the goodness of fit test.

```r
hoslem.test(chd69, fitted(final2))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  chd69, fitted(final2)
## X-squared = 13.26, df = 8, p-value = 0.1032
```

The p.value is larger than 0.05 thus we do not reject the null.

```r
theta = seq(0,1,0.01)
sens1 = rep(0,length(theta))
spec1 = rep(0,length(theta))
sens2 = rep(0,length(theta))
spec2 = rep(0,length(theta))
dist1 = rep(0,length(theta))
dist2 = rep(0,length(theta))

#want to replace values in pps w/ 0 or 1 depending on > threshold

for (i in 1:length(theta)){
  low_thresh1 = as.numeric(pp1 > theta[i])
  low_thresh2 = as.numeric(pp2 > theta[i])
  sens1[i] = sum(chd69*low_thresh1)/sum(chd69)
  sens2[i] = sum(chd69*low_thresh2)/sum(chd69)
  spec1[i] = sum((1-chd69)*(1-low_thresh1))/(length(chd69)-sum(chd69))
  spec2[i] = sum((1-chd69)*(1-low_thresh2))/(length(chd69)-sum(chd69))
  dist1[i] = (1-spec1[i])^2+(1-sens1[i])^2
  dist2[i] = (1-spec2[i])^2+(1-sens2[i])^2
}
which.min(dist1)
```

Setting thresholds between [0,1] & picking the optimal threshold that yields the best TPR-FPR ratio.

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

```
which.min(dist2)
```

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

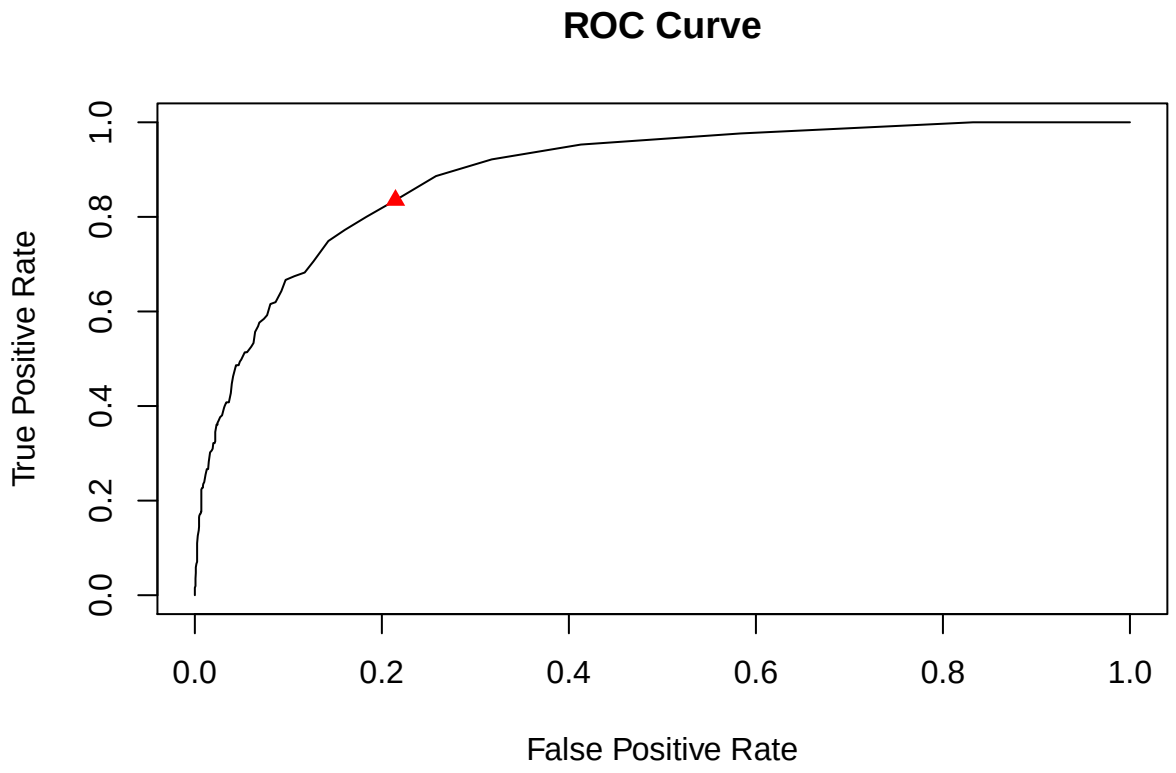
```
theta[which.min(dist1)]
```

```
## [1] 0.06
```

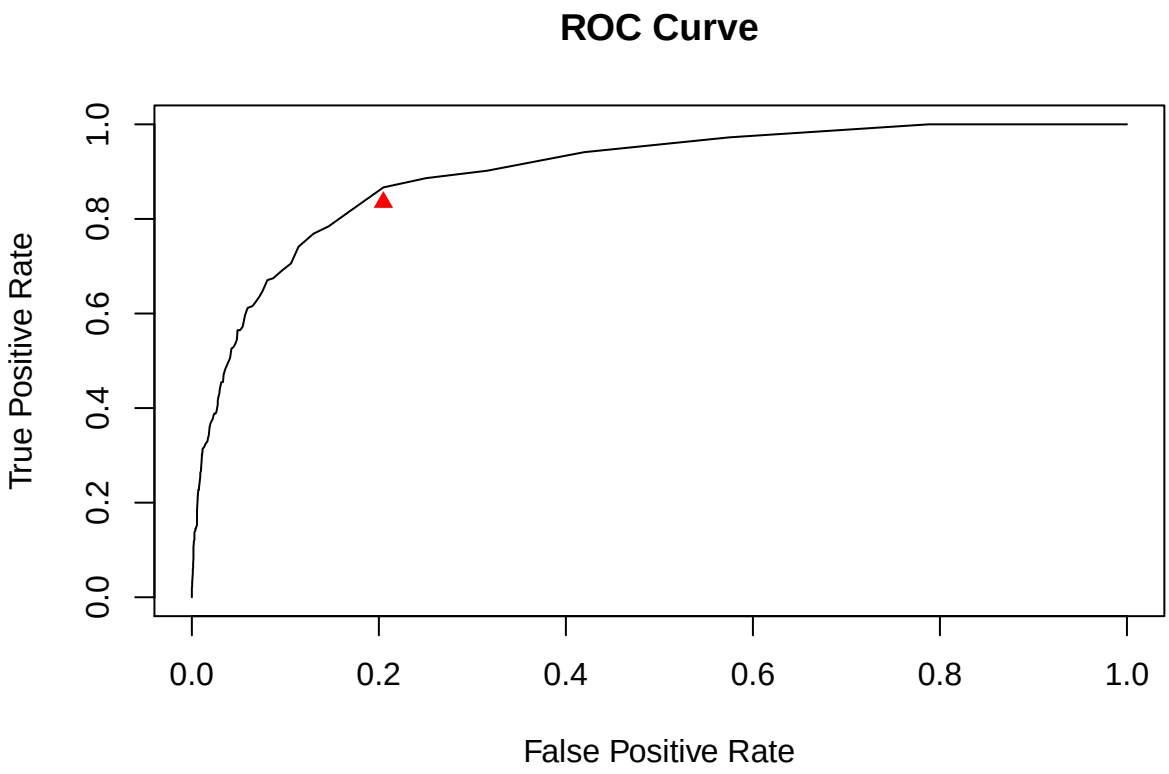
```
theta[which.min(dist2)]
```

```
## [1] 0.06
```

```
plot(1 - spec1, sens1, type="l", main = "ROC Curve",
     xlab = "False Positive Rate", ylab = "True Positive Rate")
points(1 - spec1[which.min(dist1)], sens1[which.min(dist1)], pch = 17, col = "red")
```




```
plot(1 - spec2,sens2,type = "l", main = "ROC Curve",
     xlab = "False Positive Rate", ylab = "True Positive Rate")
points(1 - spec2[which.min(dist2)], sens1[which.min(dist1)], pch = 17, col = "red")
```



Model Summary Above I constructed 2 logistic models, one included the interactions between variables & one did not. The 2nd model that included interactions performed much better as it yields a higher AUC score (~0.9) vs. the 1st model (~0.893). But most importantly, the 2nd model satisfied the Hosmer-Limeshaw GoF Test, thus it is probably a safer bet if we had to classify a new dataset. My guess on why the interactions enchanced the accuracy of this model is it may possibly have something to do with the dependence of some of our features on one another. Intuitively, it would make sense that high chloesterol level may be worse for a patient of higher age or one who has a smoking history.