GSE\_BigData\_Classification\_Assignment3

Add your name

2023-06-19

### Set the working directory and read in the dataset (PISA2018MS\_KOR\_BQ.rdata)

# infile <- outfile <- ""  
# setwd(infile)  
# getwd()  
# load("")  
  
infile <- outfile <- getwd()  
setwd(infile)  
load("PISA2018MS\_KOR\_BQ.rdata")  
  
data = PISA2018MS\_KOR[ , c("PV1MATH", "EMOSUPS", "ST004D01T", "EC154Q02IA")]

### Classification using the logistic regression

#### Make the table of the outcome variable EC154Q02IA

* EC154Q02IA: attending additional instruction in mathematics
* Make sure to to include if there are any missing cases (useNA='always')

table(data$EC154Q02IA, useNA='always')

##   
## 1 2 <NA>   
## 2586 3843 221

#### Treatment of outcome variable

* Remove the missing cases of the outcome variable
* Check the dimension
* Recode 1 -> 0 & 2 -> 1
* Create the table again to check if the recoding is done successfully

dataL1 = subset(data, ! is.na(EC154Q02IA))  
dim(data)

## [1] 6650 4

# Recode  
dataL1$EC154Q02IA[dataL1$EC154Q02IA == 1] = 0  
dataL1$EC154Q02IA[dataL1$EC154Q02IA == 2] = 1  
  
# Check recoding  
table(dataL1$EC154Q02IA, useNA='always')

##   
## 0 1 <NA>   
## 2586 3843 0

### Three ways to compute proportions: outcome variable by gender

#### Create three types of tables

# 전체  
prop.table(table(dataL1$EC154Q02IA))

##   
## 0 1   
## 0.4022399 0.5977601

# 남성  
prop.table(table(dataL1$EC154Q02IA[dataL1$ST004D01T == 2]))

##   
## 0 1   
## 0.4102564 0.5897436

# 여성  
prop.table(table(dataL1$EC154Q02IA[dataL1$ST004D01T == 1]))

##   
## 0 1   
## 0.3937058 0.6062942

#### Interpret the proportions of each table

* Out of total observations,
* Among the students who do not attend additional instruction in math (conditional on math == 2),
* Among female students (conditional on gender == 1),

### Logistic regression

#### Fit the following three logistic regressions (no interpretation is required)

* M1: EC154Q02IA by ST004D01T (gender)
* M2: EC154Q02IA by explanatory variable: PV1MATH, ST004D01T (gender)
* M3: EC154Q02IA by explanatory variables: PV1MATH, ST004D01T (gender), interaction between PV1MATH and ST004D01T (gender)

dataL2 = subset(dataL1, ST004D01T == 1)  
  
M1 = glm(EC154Q02IA ~ ST004D01T, data=dataL2, family=binomial())  
M2 = glm(EC154Q02IA ~ PV1MATH + ST004D01T, data=dataL2, family=binomial())  
M3 = glm(EC154Q02IA ~ PV1MATH + ST004D01T + PV1MATH \* ST004D01T, data=dataL2, family=binomial())

### [EXTRA] Model Evaluation

#### Predict the probabilities and values (either 0 or 1) from each model

* Keep get\_logistic\_pred function
* You need to generate three sets of predicted values based on each model (M1, M2, M3)

get\_logistic\_pred = function(mod, data, res = "y", pos = 1, neg = 0, cut = 0.5) {  
 probs = predict(mod, newdata = data, type = "response")  
 ifelse(probs >= cut, pos, neg)  
}  
  
dataL2$prdM1 = get\_logistic\_pred(M1, dataL2)  
dataL2$prdM2 = get\_logistic\_pred(M2, dataL2)  
dataL2$prdM3 = get\_logistic\_pred(M3, dataL2)

#### Evaluate models (M2 & M3) based on the following quantities

* Report Accuracy, Specificity, Subjectivity, F1 score from M2 & M3
* Use confusion matrix

library(caret)

## Warning: 패키지 'caret'는 R 버전 4.3.1에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: ggplot2

## 필요한 패키지를 로딩중입니다: lattice

confusionMatrix(as.factor(dataL2$prdM2), as.factor(dataL2$EC154Q02IA))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 424 281  
## 1 802 1607  
##   
## Accuracy : 0.6522   
## 95% CI : (0.6352, 0.669)  
## No Information Rate : 0.6063   
## P-Value [Acc > NIR] : 6.995e-08   
##   
## Kappa : 0.2129   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.3458   
## Specificity : 0.8512   
## Pos Pred Value : 0.6014   
## Neg Pred Value : 0.6671   
## Prevalence : 0.3937   
## Detection Rate : 0.1362   
## Detection Prevalence : 0.2264   
## Balanced Accuracy : 0.5985   
##   
## 'Positive' Class : 0   
##

confusionMatrix(as.factor(dataL2$prdM3), as.factor(dataL2$EC154Q02IA))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 424 281  
## 1 802 1607  
##   
## Accuracy : 0.6522   
## 95% CI : (0.6352, 0.669)  
## No Information Rate : 0.6063   
## P-Value [Acc > NIR] : 6.995e-08   
##   
## Kappa : 0.2129   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.3458   
## Specificity : 0.8512   
## Pos Pred Value : 0.6014   
## Neg Pred Value : 0.6671   
## Prevalence : 0.3937   
## Detection Rate : 0.1362   
## Detection Prevalence : 0.2264   
## Balanced Accuracy : 0.5985   
##   
## 'Positive' Class : 0   
##