Exercise 1

JAS

## Instructions for Exercise 1

R has a built-in dataset from the US National Health And Nutrition Examination Study (NHANES). Using that dataset, you will try to predict Diabetes using a subset of variables and logistic regression.

Implement the following steps:

1. Restrict the NHANES data to the list of 11 variables below. “Age”, “Race1”, “Education”, “HHIncome”, “Weight”, “Height”, “Pulse”, “Diabetes”, “BMI”, “PhysActive”, “Smoke100”
2. Remove any duplicate observations using the unique() function.
3. Remove any observations with missing data using the na.omit() function.
4. Partition your data into training and testing using a 70/30 split.
5. Set up your control settings to run 5-fold cross-validation
6. Within the training data, fit a logistic regression model using the glm method within caret.
7. Examine the average accuracy metrics across folds using confusionMatrix() function.
8. Apply your final model to the test set and obtain final evaluation metrics using confusion.Matrix()

### Chunk 1

This code chunk loads packages and performs the initial data cleaning steps. This includes subsetting to only the needed columns, removing duplicates and removing individuals with missing data.

library(lattice)  
library(NHANES)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'

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

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

library(caret)

## Loading required package: ggplot2

data ("NHANES")  
  
data.NHANES<-NHANES  
  
#Remove duplicates  
NHANES.unique<-unique(data.NHANES)  
  
keep.var <- names(NHANES.unique) %in% c("Age", "Race1", "Education", "HHIncome", "Weight", "Height", "Pulse", "Diabetes", "BMI", "PhysActive", "Smoke100")  
  
NHANES.subset<-NHANES.unique[keep.var]  
  
#Examine data  
str(NHANES.subset)

## tibble [7,832 × 11] (S3: tbl\_df/tbl/data.frame)  
## $ Age : int [1:7832] 34 4 49 9 8 45 66 58 54 10 ...  
## $ Race1 : Factor w/ 5 levels "Black","Hispanic",..: 4 5 4 4 4 4 4 4 4 4 ...  
## $ Education : Factor w/ 5 levels "8th Grade","9 - 11th Grade",..: 3 NA 4 NA NA 5 4 5 2 NA ...  
## $ HHIncome : Factor w/ 12 levels " 0-4999"," 5000-9999",..: 6 5 7 11 9 11 6 12 10 NA ...  
## $ Weight : num [1:7832] 87.4 17 86.7 29.8 35.2 75.7 68 78.4 74.7 38.6 ...  
## $ Height : num [1:7832] 165 105 168 133 131 ...  
## $ BMI : num [1:7832] 32.2 15.3 30.6 16.8 20.6 ...  
## $ Pulse : int [1:7832] 70 NA 86 82 72 62 60 62 76 80 ...  
## $ Diabetes : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ PhysActive: Factor w/ 2 levels "No","Yes": 1 NA 1 NA NA 2 2 2 2 NA ...  
## $ Smoke100 : Factor w/ 2 levels "No","Yes": 2 NA 2 NA NA 1 2 1 1 NA ...

summary(NHANES.subset)

## Age Race1 Education HHIncome   
## Min. : 0.00 Black :1073 8th Grade : 397 more 99999 :1588   
## 1st Qu.:16.00 Hispanic: 538 9 - 11th Grade: 712 75000-99999: 809   
## Median :35.00 Mexican : 920 High School :1131 25000-34999: 784   
## Mean :35.97 White :4626 Some College :1695 35000-44999: 691   
## 3rd Qu.:54.00 Other : 675 College Grad :1534 45000-54999: 596   
## Max. :80.00 NA's :2363 (Other) :2711   
## NA's : 653   
## Weight Height BMI Pulse Diabetes   
## Min. : 2.80 Min. : 83.6 Min. :12.88 Min. : 40.00 No :7077   
## 1st Qu.: 54.35 1st Qu.:155.8 1st Qu.:21.40 1st Qu.: 66.00 Yes : 622   
## Median : 71.60 Median :165.4 Median :25.83 Median : 72.00 NA's: 133   
## Mean : 69.67 Mean :160.9 Mean :26.52 Mean : 73.58   
## 3rd Qu.: 87.90 3rd Qu.:173.9 3rd Qu.:30.64 3rd Qu.: 82.00   
## Max. :230.70 Max. :200.4 Max. :81.25 Max. :136.00   
## NA's :61 NA's :317 NA's :327 NA's :1236   
## PhysActive Smoke100   
## No :2853 No :3055   
## Yes :3521 Yes :2423   
## NA's:1458 NA's:2354   
##   
##   
##   
##

#Remove missings  
NHANES.subset<-na.omit(NHANES.subset)

### Chunk 2

This chunk will partition data into training and testing using a 70/30 split

#Remember to set seed for random actions  
set.seed(123)  
training.data <- NHANES.subset$Diabetes %>% createDataPartition(p=0.7, list=F)  
  
#OR  
#training.data<-createDataPartition(NHANES.subset$Diabetes, p=0.7, list=FALSE)  
  
train.data<-NHANES.subset[training.data, ]  
test.data<-NHANES.subset[-training.data, ]

### Chunk 3

This chunk will set up the training procedure in caret. This includes setting up 5-fold cross validation and constructing the logistic regression model.

set.seed(123)  
  
train.control.settings<-trainControl(method="cv",   
 number=5)  
logit.caret<-train(Diabetes ~ .,  
 data=train.data,   
 method="glm",   
 family="binomial",   
 trControl=train.control.settings)  
  
logit.caret$results

## parameter Accuracy Kappa AccuracySD KappaSD  
## 1 none 0.8799977 0.0929398 0.006083958 0.03154778

confusionMatrix(logit.caret)

## Cross-Validated (5 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction No Yes  
## No 87.1 10.4  
## Yes 1.6 0.9  
##   
## Accuracy (average) : 0.88

### Chunk 4

This chunk applies the “best” model within the test set and obtains a final confusion matrix.

set.seed(123)  
pred.nhanes.test<-predict(logit.caret, test.data)  
  
eval.results.4<-confusionMatrix(pred.nhanes.test, test.data$Diabetes, positive = "Yes")  
print(eval.results.4)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1251 146  
## Yes 22 16  
##   
## Accuracy : 0.8829   
## 95% CI : (0.8652, 0.8991)  
## No Information Rate : 0.8871   
## P-Value [Acc > NIR] : 0.7088   
##   
## Kappa : 0.1223   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.09877   
## Specificity : 0.98272   
## Pos Pred Value : 0.42105   
## Neg Pred Value : 0.89549   
## Prevalence : 0.11289   
## Detection Rate : 0.01115   
## Detection Prevalence : 0.02648   
## Balanced Accuracy : 0.54074   
##   
## 'Positive' Class : Yes   
##

#Examine coefficients for metric of variable importance  
coef(logit.caret$finalModel)

## (Intercept) Age Race1Hispanic   
## -19.040165895 0.067767276 -0.333604715   
## Race1Mexican Race1White Race1Other   
## -0.044725451 -0.846670116 0.308293836   
## `Education9 - 11th Grade` `EducationHigh School` `EducationSome College`   
## -0.326922122 -0.395041665 -0.229886337   
## `EducationCollege Grad` `HHIncome 5000-9999` `HHIncome10000-14999`   
## -0.411426420 0.459674817 0.462407329   
## `HHIncome15000-19999` `HHIncome20000-24999` `HHIncome25000-34999`   
## -0.120993315 0.410450051 0.097577311   
## `HHIncome35000-44999` `HHIncome45000-54999` `HHIncome55000-64999`   
## 0.191921346 -0.064970401 0.348671274   
## `HHIncome65000-74999` `HHIncome75000-99999` `HHIncomemore 99999`   
## -0.201939240 -0.098666166 -0.366669839   
## Weight Height BMI   
## -0.044912151 0.061951110 0.207374672   
## Pulse PhysActiveYes Smoke100Yes   
## 0.015592825 0.006000855 0.233530380

set.seed(123)  
  
train.control.settings<-trainControl(method="cv",   
 number=5, sampling="up")  
logit.caret<-train(Diabetes ~ .,  
 data=train.data,   
 method="glm",   
 family="binomial",   
 trControl=train.control.settings)  
  
logit.caret$results

## parameter Accuracy Kappa AccuracySD KappaSD  
## 1 none 0.7319276 0.2678175 0.01911008 0.03497403

confusionMatrix(logit.caret)

## Cross-Validated (5 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction No Yes  
## No 64.6 2.7  
## Yes 24.1 8.6  
##   
## Accuracy (average) : 0.7319