Classification model for heart attack

Siddharth Rastogi

2022-08-21

Heart-related diseases are among the most prevalent chronic diseases in the United States. Preventive identification of at-risk subjects and of the factors associated to hearth-related conditions is paramount for effective prevention of negative outcomes (like hearth attacks) and testing. The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related survey that collects state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The dataset data_heart_disease_BRFSS2015.csv contains records for 253,680 respondents of the survey. For each subject, information is available on whether the subject had a heart-related disease (HeartDiseaseorAttack), and additional information including general behavior, demographic characteristics, self-reported health status, and disease history. More information is available at the link 'https://www.kaggle.com/alexteboul/heart-disease-health-indicators- dataset'.

```
data <- read.csv("data_heart_disease_BRFSS2015.csv")
# make sure that binary/categorical variables are correctly encoded as factor
data[,c(1:4,6:14,18:19)] <- lapply( data[,c(1:4,6:14,18:19)], factor )
str(data)
```

```
## 'data.frame':    253680 obs. of  22 variables:
##  $ HeartDiseaseorAttack: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ HighBP              : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 1 ...
##  $ HighChol            : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 1 2 2 1 ...
##  $ CholCheck           : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
##  $ BMI                 : num  40 25 28 27 24 25 30 25 30 24 ...
##  $ Smoker              : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 2 2 1 ...
##  $ Stroke              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Diabetes            : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 3 1 ...
##  $ PhysActivity        : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 1 2 1 1 ...
##  $ Fruits              : Factor w/ 2 levels "0","1": 1 1 2 2 2 2 1 1 2 1 ...
##  $ Veggies             : Factor w/ 2 levels "0","1": 2 1 1 2 2 2 1 2 2 2 ...
##  $ HvyAlcoholConsump   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ AnyHealthcare       : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
##  $ NoDocbcCost         : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 1 1 ...
##  $ GenHlth             : num  5 3 5 2 2 2 3 3 5 2 ...
##  $ MentHlth            : num  18 0 30 0 3 0 0 0 30 0 ...
##  $ PhysHlth            : num  15 0 30 0 0 2 14 0 30 0 ...
##  $ DiffWalk            : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 2 2 1 ...
##  $ Sex                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 2 ...
##  $ Age                 : num  9 7 9 11 11 10 9 11 9 8 ...
##  $ Education           : num  4 6 4 3 5 6 6 4 5 4 ...
##  $ Income              : num  3 1 8 6 4 8 7 4 1 3 ...
```

Below code is used to check if classes are balanced or not, it shows that classes are highly imbalanced.

```
# classes are highly imbalanced
table(data$HeartDiseaseorAttack)
```

```
##
##      0      1
## 229787  23893
```

```
table(data$HeartDiseaseorAttack)/nrow(data)
```

```
##
##          0          1
## 0.90581441 0.09418559
```

```
fit <- glm(HeartDiseaseorAttack ~ ., data = data, family = "binomial")
```

```
summary(fit)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ ., family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2108  -0.4253  -0.2423  -0.1291   3.6480
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -7.9122683  0.1028048 -76.964  < 2e-16 ***
## HighBP1              0.5246260  0.0177529  29.552  < 2e-16 ***
## HighChol1            0.6120240  0.0164535  37.197  < 2e-16 ***
## CholCheck1           0.5255473  0.0662502   7.933 2.14e-15 ***
## BMI                  0.0010078  0.0012122   0.831   0.4058
## Smoker1              0.3629755  0.0157329  23.071  < 2e-16 ***
## Stroke1              0.9778703  0.0244359  40.018  < 2e-16 ***
## Diabetes1            0.0431510  0.0468276   0.921   0.3568
## Diabetes2            0.2955673  0.0179652  16.452  < 2e-16 ***
## PhysActivity1        0.0400679  0.0172009   2.329   0.0198 *
## Fruits1              0.0060530  0.0163281   0.371   0.7109
## Veggies1             0.0425865  0.0189370   2.249   0.0245 *
## HvyAlcoholConsump1 -0.2934763  0.0392871  -7.470 8.02e-14 ***
## AnyHealthcare1     -0.0074395  0.0412835  -0.180   0.8570
## NoDocbcCost1         0.2536366  0.0268931   9.431  < 2e-16 ***
## GenHlth              0.4906875  0.0095107  51.593  < 2e-16 ***
## MentHlth             0.0024885  0.0009779   2.545   0.0109 *
## PhysHlth             0.0010473  0.0008767   1.195   0.2322
## DiffWalk1            0.2947437  0.0193870  15.203  < 2e-16 ***
## Sex1                 0.7612181  0.0160334  47.477  < 2e-16 ***
## Age                  0.2557692  0.0036445  70.180  < 2e-16 ***
## Education            0.0108753  0.0081704   1.331   0.1832
## Income              -0.0432386  0.0042481 -10.178  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 158355  on 253679  degrees of freedom
## Residual deviance: 120928  on 253657  degrees of freedom
## AIC: 120974
```

```
##
## Number of Fisher Scoring iterations: 6
```

```
tau <- 0.5
p <- fitted(fit)
pred <- ifelse(p > tau, 1, 0)
```

```
# cross tabulation between observed and predicted
table(data$HeartDiseaseorAttack, pred)
```

```
##    pred
##         0      1
##   0 227210   2577
##   1  20845   3048
```

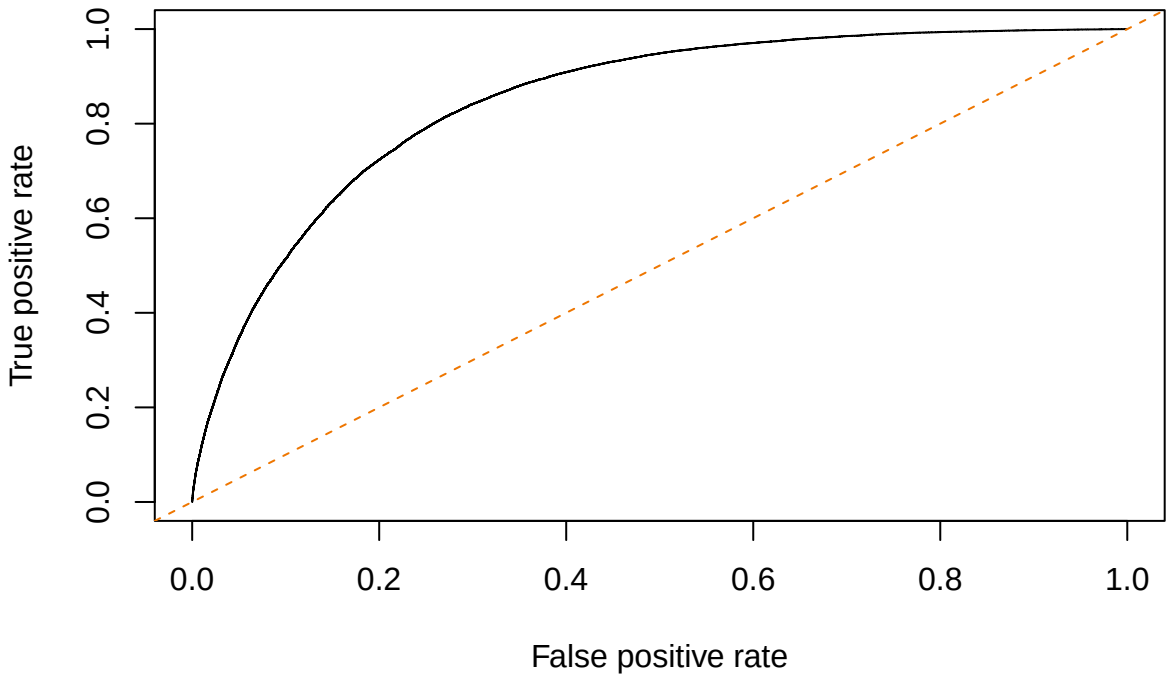
```
# compute accuracy for given tau
tab <- table(data$HeartDiseaseorAttack, pred)
sum(diag(tab))/sum(tab)
```

```
## [1] 0.9076711
```

Different measures for assessing the predictive performance of the logistic regression model can be computed for varying values of the discrimination threshold tau. Package ROCR can be used to calculate many performance measures. To use the functionalities of the package, we first need to create a prediction object, providing in input the estimated probabilities and the actual class values of the response variable.

```
library(ROCR)
pred_obj <- prediction(fitted(fit), data$HeartDiseaseorAttack)
```

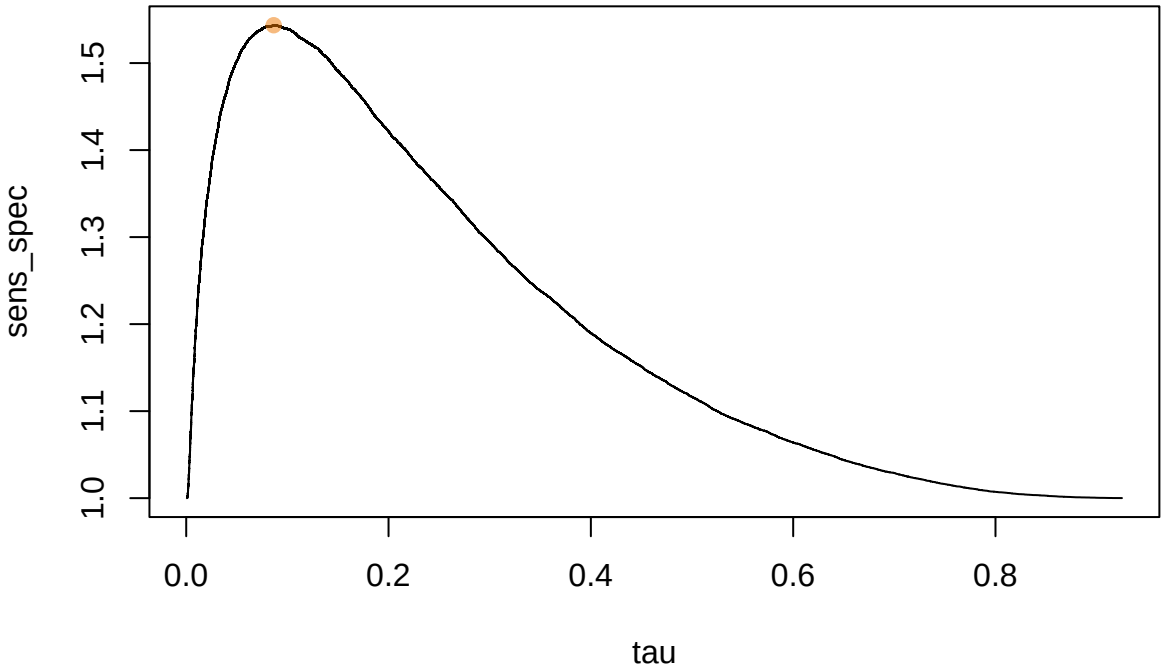
```
perf <- performance(pred_obj, "tpr", "fpr")
plot(perf)
abline(0,1, col = "darkorange2", lty = 2)
```




```
# compute the area under the ROC curve
auc <- performance(pred_obj, "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.8473533
```

```
sens <- performance(pred_obj, "sens")
spec <- performance(pred_obj, "spec")
tau <- sens@x.values[[1]]
sens_spec <- sens@y.values[[1]] + spec@y.values[[1]]
best <- which.max(sens_spec)
plot(tau, sens_spec, type = "l")
points(tau[best], sens_spec[best], pch = 19, col = adjustcolor("darkorange2", 0.5))
```




```
tau[best] # optimal tau
```

```
##     253181
## 0.08648205
```

```
# classification for optimal tau
pred <- ifelse(fitted(fit) > tau[best], 1, 0)
table(data$HeartDiseaseorAttack, pred)
```

```
##    pred
##          0      1
##   0 166371  63416
##   1   4315  19578
```

```
# accuracy for optimal tau
acc <- performance(pred_obj, "acc")
acc@y.values[[1]][best]
```

```
## [1] 0.7330101
```