diabetes-models

David Riina

2023-05-09

```r
library(readr)
df <- read_csv("C:/Users/david/Downloads/archive (5).zip")
```

```
## Rows: 100000 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): gender, smoking_history
## dbl (7): age, hypertension, heart_disease, bmi, HbA1c_level, blood_glucose_l...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
View(df)
#correct mistakes in the smoking_history column
df$smoking_history <- ifelse(df$smoking_history == "ever", "never", df$smoking_history)

#data preprocessing
df$gender <- as.factor(df$gender)
df$hypertension <- as.factor(df$hypertension)
df$heart_disease <- as.factor(df$heart_disease)
df$smoking_history <- as.factor(df$smoking_history)
df$diabetes <- as.factor(df$diabetes)


summary(df)
```

```
##     gender           age        hypertension heart_disease    smoking_history
##  Female:58552   Min.   : 0.08   0:92515      0:96058       current    : 9286
##  Male  :41430   1st Qu.:24.00   1: 7485      1: 3942       former     : 9352
##  Other :   18   Median :43.00                              never      :39099
##                 Mean   :41.89                              No Info    :35816
##                 3rd Qu.:60.00                              not current: 6447
##                 Max.   :80.00
##       bmi         HbA1c_level    blood_glucose_level diabetes
##  Min.   :10.01   Min.   :3.500   Min.   : 80.0       0:91500
##  1st Qu.:23.63   1st Qu.:4.800   1st Qu.:100.0       1: 8500
##  Median :27.32   Median :5.800   Median :140.0
##  Mean   :27.32   Mean   :5.528   Mean   :138.1
##  3rd Qu.:29.58   3rd Qu.:6.200   3rd Qu.:159.0
##  Max.   :95.69   Max.   :9.000   Max.   :300.0
```

```r
#split into training and testing
set.seed(123)
inTrain <- sample(nrow(df), 0.7*nrow(df))

dftrain <- df[inTrain,] # with 70% of the data
dftest <- df[-inTrain,] # with 30% of the data
```

```r
# NAIVE BAYES MODEL
#import library for naive bayes
library(e1071)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
#create model and make class predictions
model <- naiveBayes(diabetes~., data = dftrain)
prediction <- predict(model, newdata = dftest)
#confusion matrix
print(cm1 <- table(dftest$diabetes, prediction, dnn = list('actual','predicted')))
```

```
##       predicted
## actual     0     1
##      0 27013   509
##      1   879  1599
```

```r
acc1 <- sum(cm1[1,1],cm1[2,2])/sum(cm1)
sen1 <- cm1[2,2]/sum(cm1[2,1],cm1[2,2])
spec1 <- cm1[1,1]/sum(cm1[1,2], cm1[1,1])

cat("Accuracy:", acc1,"\n")
```

```
## Accuracy: 0.9537333
```

```r
cat("Sensitivity:", sen1,"\n")
```

```
## Sensitivity: 0.6452785
```

```r
cat("Specificity:", spec1)
```

```
## Specificity: 0.9815057
```

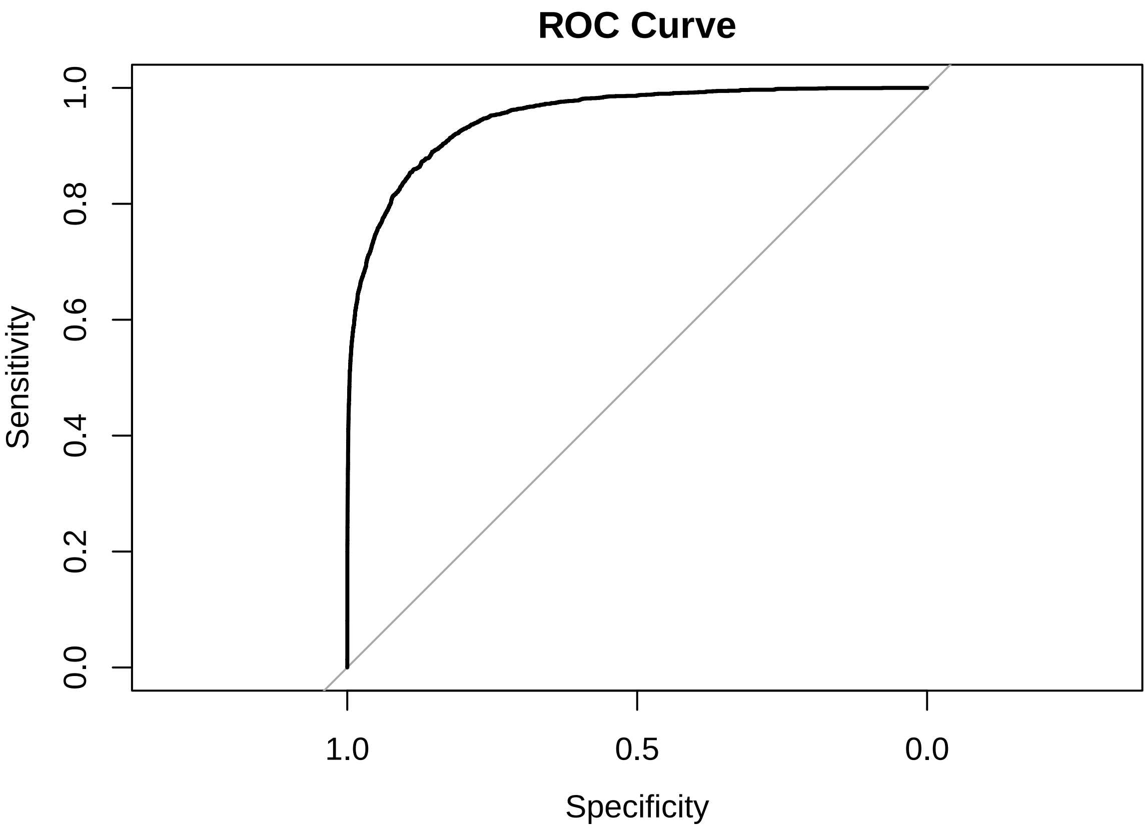
```
#get class probabilities
class_prob<- predict(model, newdata = dftest[,-9], type="raw")

#create curve with actual vs predicted probabilities of success case
roc_curve <- roc(dftest$diabetes, class_prob[,2])
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
# Plotting the ROC curve
plot(roc_curve, main = "ROC Curve", xlim = c(1,0))
```



```
library(rpart)
library(rpart.plot)

#create decision tree model
model <- rpart(diabetes ~ ., data = dftrain, method = "class")

#make predictions on the test set
predictions <- predict(model, newdata = dftest, type = "class")

#calculate the confusion matrix and error rate
print(cm2 <- table(predictions, dftest$diabetes))
```

```
##
## predictions     0     1
##           0 27522   793
##           1     0  1685
```

```
acc2 <- sum(cm2[1,1],cm2[2,2])/sum(cm2)
sen2 <- cm2[2,2]/sum(cm2[2,1],cm2[2,2])
spec2 <- cm2[1,1]/sum(cm2[1,2], cm2[1,1])

cat("Accuracy:", acc2,"\n")
```

```
## Accuracy: 0.9735667
```

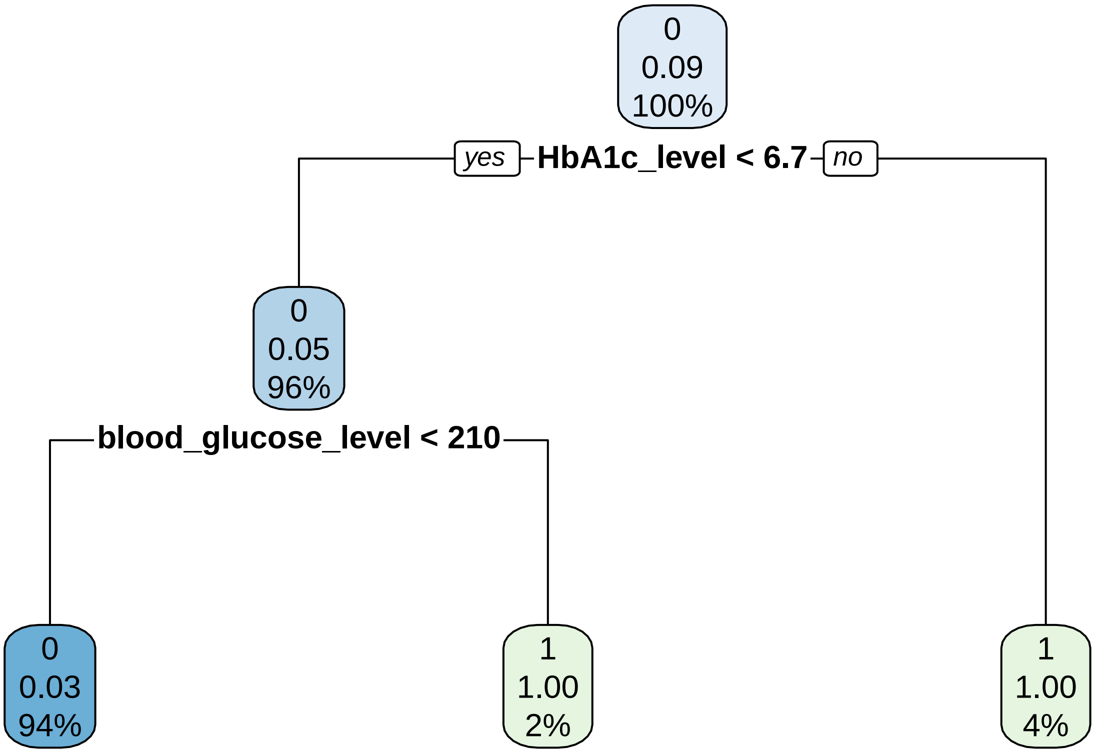
```
cat("Sensitivity:", sen2,"\n")
```

```
## Sensitivity: 1
```

```
cat("Specificity:", spec2)
```

```
## Specificity: 0.9719936
```

```
#plot the decision tree
rpart.plot(model)
```



```r
#LOGISTIC MODEL
logmodel <- glm(diabetes~., family = "binomial", data = dftrain)
summary(logmodel)
```

```
##
## Call:
## glm(formula = diabetes ~ ., family = "binomial", data = dftrain)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                -2.715e+01  3.491e-01 -77.776  < 2e-16 ***
## genderMale                  2.946e-01  4.296e-02   6.857 7.03e-12 ***
## genderOther                -9.738e+00  1.190e+02  -0.082   0.9348    
## age                         4.625e-02  1.343e-03  34.436  < 2e-16 ***
## hypertension1               6.917e-01  5.662e-02  12.216  < 2e-16 ***
## heart_disease1              8.135e-01  7.196e-02  11.306  < 2e-16 ***
## smoking_historyformer      -1.203e-01  8.347e-02  -1.441   0.1495    
## smoking_historynever       -1.451e-01  7.110e-02  -2.041   0.0413 *  
## smoking_historyNo Info     -7.341e-01  7.931e-02  -9.257  < 2e-16 ***
## smoking_historynot current -1.989e-01  9.852e-02  -2.019   0.0435 *  
## bmi                         9.178e-02  3.085e-03  29.751  < 2e-16 ***
## HbA1c_level                 2.345e+00  4.255e-02  55.106  < 2e-16 ***
## blood_glucose_level         3.314e-02  5.747e-04  57.656  < 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: 41055  on 69999  degrees of freedom
## Residual deviance: 15945  on 69987  degrees of freedom
## AIC: 15971
##
## Number of Fisher Scoring iterations: 12
```

```r
#make predictions and create confusion matrix
test.predictedprob<- predict(logmodel, newdata= dftest, type = "response")
test.predict<- ifelse(test.predictedprob >= 0.5, 1, 0)
cm3<- table(dftest$diabetes, test.predict, dnn = c("actual", "predicted"))
print(cm3)
```

```
##       predicted
## actual     0     1
##      0 27251   271
##      1   907  1571
```

```r
acc3 <- sum(cm3[1,1],cm3[2,2])/sum(cm3)
sen3 <- cm3[2,2]/sum(cm3[2,1],cm3[2,2])
spec3 <- cm3[1,1]/sum(cm3[1,2], cm3[1,1])

cat("Accuracy:", acc3,"\n")
```

```
## Accuracy: 0.9607333
```

```
cat("Sensitivity:", sen3,"\n")
```

```
## Sensitivity: 0.633979
```

```
cat("Specificity:", spec3)
```

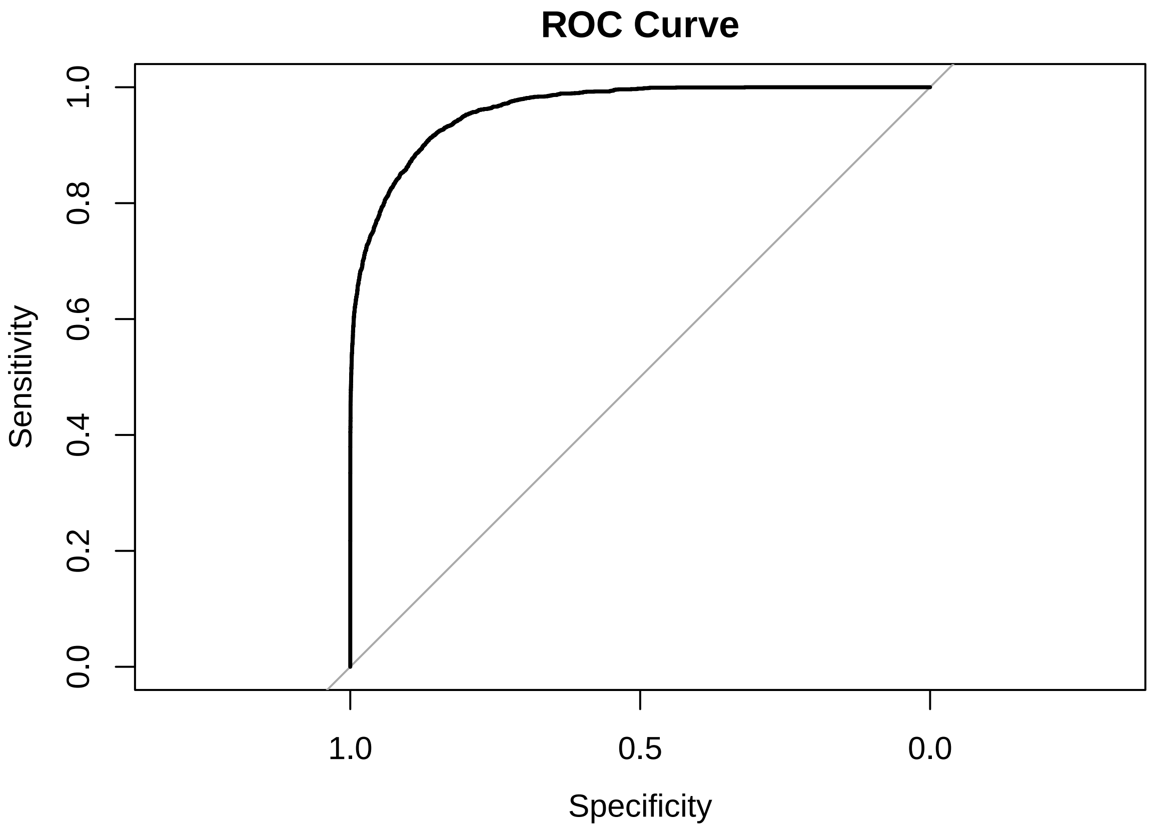
```
## Specificity: 0.9901533
```

```
#create ROC curve
roc_curve3 <- roc(dftest$diabetes, test.predictedprob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
# Plotting the ROC curve
plot(roc_curve3, main = "ROC Curve")
```


ROC Curve
Sensitivity
Specificity