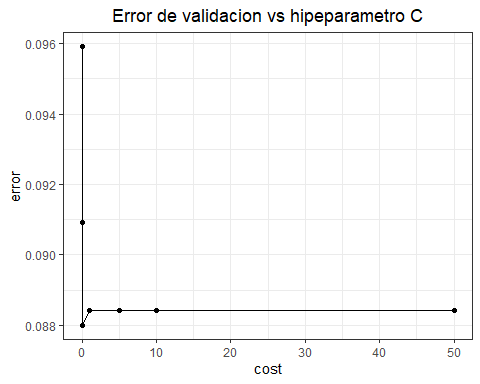
Modelos para la Clasificación y Segmentación

Dennis Peralta

2023-05-15

Se carga la base de datos de nacidos vivos, y se filtra información para la provincia de Manabí, también se elimina de la base de datos las observaciones que no tienen información, la variable peso se cambia a tipo facor debido a que es la variable de estudio, el peso del nacido vivo se codifica como “1” en el caso de que el peso sea mayor a 2500kg ya que que se considera como un peso adecuado del nacido vivo, y de las otras variables se realiza las transformaciones y categorías necesarias para que nuestro modelo pueda correr.

set.seed(1234)  
entrenamiento <- createDataPartition(nuevadata$peso,  
 p=0.1,list=F)  
modelo.tuneado <- tune(svm,  
 peso ~.,  
 data=nuevadata[entrenamiento,],  
 ranges = list(cost=c(0.001,0.01,0.1,1,5,10,50)),  
 kernel="linear",  
 scale=T,  
 probability=TRUE)  
  
ggplot(data=modelo.tuneado$performances,  
 aes(x=cost,y=error))+  
 geom\_line()+  
 geom\_point()+  
 labs(title="Error de validacion vs hipeparametro C")+  
 theme\_bw()+  
 theme(plot.title = element\_text(hjust = 0.5))

 Se observa en el gráfico que la taza de error cae de forma drástica a medida que el costo va aumentando, sin embargo el proceso de cross-validation muestra que existe un costo que consigue un error muy bajo.

mejor.modelo <- modelo.tuneado$best.model  
summary(mejor.modelo)

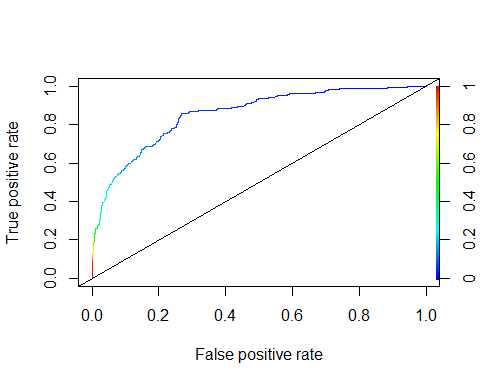
Call:  
best.tune(METHOD = svm, train.x = peso ~ ., data = nuevadata[entrenamiento,   
 ], ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 50)), kernel = "linear",   
 scale = T, probability = TRUE)  
  
  
Parameters:  
 SVM-Type: C-classification   
 SVM-Kernel: linear   
 cost: 0.1   
  
Number of Support Vectors: 457  
  
 ( 230 227 )  
  
  
Number of Classes: 2   
  
Levels:   
 no.adecuado adecuado

El mejor modelo que minimiza el error sería el que tenga un costo de 0.1 con 457 vectores de soporte clasificado en dos clases “adecuado” y “no adecuado”.

ajustados.mejor.modelo <- predict(mejor.modelo,  
 nuevadata[entrenamiento,],  
 type="prob",  
 probability = T)  
  
confusionMatrix(ajustados.mejor.modelo,  
 nuevadata$peso[entrenamiento],  
 positive = levels(nuevadata$peso)[2])

Confusion Matrix and Statistics  
  
 Reference  
Prediction no.adecuado adecuado  
 no.adecuado 54 18  
 adecuado 176 2150  
   
 Accuracy : 0.9191   
 95% CI : (0.9075, 0.9297)  
 No Information Rate : 0.9041   
 P-Value [Acc > NIR] : 0.005987   
   
 Kappa : 0.3268   
   
 Mcnemar's Test P-Value : < 2.2e-16   
   
 Sensitivity : 0.9917   
 Specificity : 0.2348   
 Pos Pred Value : 0.9243   
 Neg Pred Value : 0.7500   
 Prevalence : 0.9041   
 Detection Rate : 0.8966   
 Detection Prevalence : 0.9700   
 Balanced Accuracy : 0.6132   
   
 'Positive' Class : adecuado

pred <- prediction(attr(ajustados.mejor.modelo,  
 "probabilities")[,2],  
 nuevadata$peso[entrenamiento])  
  
perf <- performance(pred,"tpr","fpr")  
  
  
  
plot(perf,colorize=T,lty=3)  
abline(0,1,col="black")



aucmodelo1 <- performance(pred,measure = "auc")  
aucmodelo1 <- aucmodelo1@y.values[[1]]  
aucmodelo1

[1] 0.8548723

El mejor modelo tiene una precisión de clasificación muy bueno ya que su valor está muy cercano a 1, por otra parte clasifica con una probabilidad de 0.9243 el peso del nacido vivo cuando es el adecuado y con una probabilidad de 0.75 cuando el peso no es el adecuado, también la curva ROC es muy buena ya que la curva esta muy cerca de la parte superior.

Sin embargo la sensitividad es muy alta y la especificidad es muy baja lo que quizás se debe a un problema de desproporcionalidad muestral, lo que se va a corroborar más adelante o podría deberse aun problema del umbral de discriminación que es por defecto 0,5.

max.accuracy <- performance(pred,measure = "acc")  
indice <- which.max(slot(max.accuracy,"y.values")[[1]])  
acc <- slot(max.accuracy,"y.values")[[1]][indice]  
cutoff <- slot(max.accuracy,"x.values")[[1]][indice]  
  
print(c(accuracy=acc,  
 cutoff=cutoff))

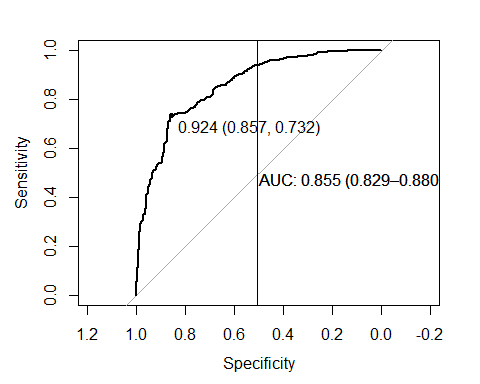
accuracy cutoff.10430   
 0.9195163 0.5068994

Según los resultados de la parte superior el umbral óptimo sería de 0.568994, lo que se espera es que el umbral óptimo mejore las predicciones, la especificidad, la sensitividada, lo que permitiría clasificar de una mejor forma cuando el peso de un nacido vivo en la provincia de Manabí es el adecuado y cuando no.

umbral <- as.numeric(cutoff)  
  
prediccionescutoff <- attr(ajustados.mejor.modelo,  
 "probabilities")[,1]  
  
prediccionescutoff <- as.numeric(prediccionescutoff)  
  
  
predCut <- factor(ifelse(prediccionescutoff>umbral,1,0))  
  
  
matrizpuntocorte <- data.frame(real=nuevadata$peso[entrenamiento],  
 predicho=predCut)  
  
matrizpuntocorte <- matrizpuntocorte %>% mutate(predicho=recode\_factor(predicho,  
 `0`="no.adecuado",  
 `1`="adecuado"))  
  
  
confusionMatrix(matrizpuntocorte$predicho,  
 matrizpuntocorte$real,  
 positive = "adecuado")

Confusion Matrix and Statistics  
  
 Reference  
Prediction no.adecuado adecuado  
 no.adecuado 55 20  
 adecuado 175 2148  
   
 Accuracy : 0.9187   
 95% CI : (0.907, 0.9293)  
 No Information Rate : 0.9041   
 P-Value [Acc > NIR] : 0.007342   
   
 Kappa : 0.329   
   
 Mcnemar's Test P-Value : < 2.2e-16   
   
 Sensitivity : 0.9908   
 Specificity : 0.2391   
 Pos Pred Value : 0.9247   
 Neg Pred Value : 0.7333   
 Prevalence : 0.9041   
 Detection Rate : 0.8957   
 Detection Prevalence : 0.9687   
 Balanced Accuracy : 0.6150   
   
 'Positive' Class : adecuado

curvaroc <- plot.roc(nuevadata$peso[entrenamiento],  
 as.vector(prediccionescutoff),  
 precent=TRUE,  
 ci=TRUE,  
 print.auc=TRUE,  
 threholds="best",  
 print.thres="best")  
abline(v=cutoff)

 El modelo con el punto de corte óptimo no ha mejorado significativamente en la precisión, ni en la especificidad y tampoco tiene mejoras en la sensitividad, de igual manera la curva ROC que en el punto de corte óptimo mi modelo tiene una sensitivad de aproximadamente 0.84, sin embargo la especificidad no ha mejorado en gran manera, se sospecha que hay una desproporcionalidad muestral para clasificar los “1” y los “0”.

newdata2 <- data.frame(talla=45,  
 sem\_gest=38,  
 sexo=1,  
 edad\_mad=30,  
 sabe\_leer=1,  
 con\_pren=1,  
 edad2=900)  
  
pronostico1 <- predict(mejor.modelo, newdata2,probability = TRUE)  
pronostico1

1   
adecuado   
attr(,"probabilities")  
 adecuado no.adecuado  
1 0.7328316 0.2671684  
Levels: no.adecuado adecuado

pronostico2 <- ifelse(attr(pronostico1,"probabilities")[1]>0.924,1,0)  
pronostico2

[1] 0

La clasificación para el modelo con un corte umbral de 0.5 por defecto clasifica según el valor de las variables de “newdata2” como un niño que nacerá con un peso adecuado sin embargo con un umbral de 0.924 el niño se clasifica como “0”, es decir un niño que nacera con un peso no adecuado.

A continuación se va a evaluar el modelo con el valor de umbral que está mas distante a la línea de 45°

umbral1 <- 0.924  
  
pred.umbral1 <- attr(ajustados.mejor.modelo,  
 "probabilities")[,1]  
  
pred.umbral1 <- as.numeric(pred.umbral1)  
  
  
pred.cut.umbral1 <- factor(ifelse(pred.umbral1>umbral1,1,0))  
  
  
matriz.corte <- data.frame(real=nuevadata$peso[entrenamiento],  
 predicho=pred.cut.umbral1)  
  
matriz.corte <- matriz.corte %>% mutate(predicho=recode\_factor(predicho, `0`="no.adecuado",  
 `1`="adecuado"))  
  
  
  
confusionMatrix(matriz.corte$predicho,  
 matriz.corte$real,  
 positive = "adecuado")

Confusion Matrix and Statistics  
  
 Reference  
Prediction no.adecuado adecuado  
 no.adecuado 197 587  
 adecuado 33 1581  
   
 Accuracy : 0.7415   
 95% CI : (0.7234, 0.7589)  
 No Information Rate : 0.9041   
 P-Value [Acc > NIR] : 1   
   
 Kappa : 0.2821   
   
 Mcnemar's Test P-Value : <2e-16   
   
 Sensitivity : 0.7292   
 Specificity : 0.8565   
 Pos Pred Value : 0.9796   
 Neg Pred Value : 0.2513   
 Prevalence : 0.9041   
 Detection Rate : 0.6593   
 Detection Prevalence : 0.6731   
 Balanced Accuracy : 0.7929   
   
 'Positive' Class : adecuado

El modelo evaluado con el punto de corte de “0.924” tiene un valor menor del accuracy en comparación con el modelo evaluado del umbral de “0.5” por defecto y con el valor del cutoff, sin embargo la sensitividad y la sensitivad son buenos y clasifica los adecuados con una probabilidad de 0.9796, pero el valor de probabilidad de clasificación de los nacidos vivos con un peso adecuado es insignificante.

train\_data <- nuevadata[entrenamiento, ]  
roses <- ROSE(peso ~.,  
 data = train\_data,seed = 1)$data  
  
modelo.rose <- tune(svm, peso ~ .,  
 data=roses,  
 ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 50)),  
 kernel = "linear",  
 scale=T,  
 probability = TRUE)  
  
mejor.modelo.rose <- modelo.rose$best.model  
  
ajustadosrose <- predict(mejor.modelo.rose,  
 roses, type="prob",probability=TRUE)  
  
confusionMatrix(roses$peso,ajustadosrose,  
 dnn = c("Actuales", "Predichos"),  
 levels(ajustadosrose)[1])

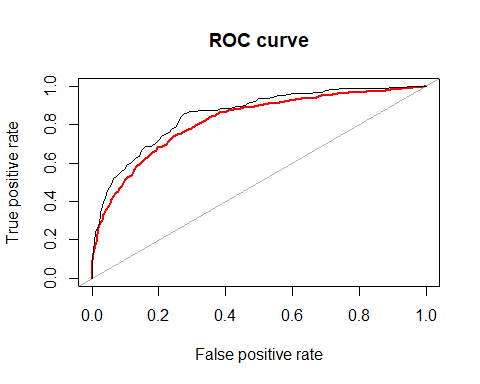
Confusion Matrix and Statistics  
  
 Predichos  
Actuales adecuado no.adecuado  
 adecuado 977 272  
 no.adecuado 352 797  
   
 Accuracy : 0.7398   
 95% CI : (0.7217, 0.7573)  
 No Information Rate : 0.5542   
 P-Value [Acc > NIR] : < 2.2e-16   
   
 Kappa : 0.4772   
   
 Mcnemar's Test P-Value : 0.001564   
   
 Sensitivity : 0.7351   
 Specificity : 0.7456   
 Pos Pred Value : 0.7822   
 Neg Pred Value : 0.6936   
 Prevalence : 0.5542   
 Detection Rate : 0.4074   
 Detection Prevalence : 0.5209   
 Balanced Accuracy : 0.7403   
   
 'Positive' Class : adecuado

La matriz de confusión tiene un accuracy casi del mismo valor que el modelo evaluado con 0.924, al igual que la sensitividad y especificidad, lo mejor del modelo con remuestreo ROSE es que mantiene la probabilidad para predecir si un nacido vivo nace con un peso “adecuado” y mejora la probabilidad para predecir si peso del nacido es no “adecuado”, en resumen había un problema de desproporcionalidad muestral, ya que se podría decir que el modelo con remuestreo es el mejor de todos los anteriores modelados.

predrose <- prediction(attr(ajustadosrose,  
 "probabilities")[,2],  
 roses$peso)  
  
roc.curve(roses$peso, attr(ajustadosrose,  
 "probabilities")[,2], col="red")

Area under the curve (AUC): 0.820

roc.curve(nuevadata$peso[entrenamiento], attr(ajustados.mejor.modelo,  
 "probabilities")[,2], col="black",  
 add.roc = T)



Area under the curve (AUC): 0.854

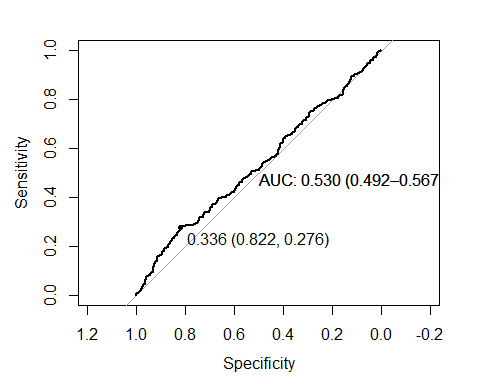
pred1 <- prediction(attr(ajustadosrose,  
 "probabilities")[,2],  
 nuevadata$peso[entrenamiento])  
  
max.accuracy1 <- performance(pred1,measure = "acc")  
indice <- which.max(slot(max.accuracy1,"y.values")[[1]])  
acc <- slot(max.accuracy1,"y.values")[[1]][indice]  
cutoff1 <- slot(max.accuracy1,"x.values")[[1]][indice]  
  
print(c(accuracy=acc,  
 cutoff1=cutoff))

accuracy cutoff1.10430   
 0.9040867 0.5068994

puntocorte <- as.numeric(cutoff1)  
  
prediccionescut <- attr(ajustadosrose,  
 "probabilities")[,1]  
  
prediccionescut <- as.numeric(prediccionescut)  
  
  
predCutt <- factor(ifelse(prediccionescut>puntocorte,1,0))  
  
  
matriz <- data.frame(real=nuevadata$peso[entrenamiento],  
 predicho=predCutt)  
  
matriz <- matriz %>% mutate(predicho=recode\_factor(predicho,  
 `0`="no.adecuado",  
 `1`="adecuado"))  
  
  
confusionMatrix(matriz$predicho,  
 matriz$real,  
 positive = "adecuado")

Confusion Matrix and Statistics  
  
 Reference  
Prediction no.adecuado adecuado  
 no.adecuado 230 2168  
 adecuado 0 0  
   
 Accuracy : 0.0959   
 95% CI : (0.0844, 0.1084)  
 No Information Rate : 0.9041   
 P-Value [Acc > NIR] : 1   
   
 Kappa : 0   
   
 Mcnemar's Test P-Value : <2e-16   
   
 Sensitivity : 0.00000   
 Specificity : 1.00000   
 Pos Pred Value : NaN   
 Neg Pred Value : 0.09591   
 Prevalence : 0.90409   
 Detection Rate : 0.00000   
 Detection Prevalence : 0.00000   
 Balanced Accuracy : 0.50000   
   
 'Positive' Class : adecuado

curvaroc1 <- plot.roc(nuevadata$peso[entrenamiento],  
 as.vector(prediccionescut),  
 precent=TRUE,  
 ci=TRUE,  
 print.auc=TRUE,  
 threholds="best",  
 print.thres="best")



newdata3 <- data.frame(talla=45,  
 sem\_gest=38,  
 sexo=1,  
 edad\_mad=30,  
 sabe\_leer=1,  
 con\_pren=1,  
 edad2=900)  
  
pronostico3 <- predict(mejor.modelo, newdata3,probability = TRUE)  
pronostico3

1   
adecuado   
attr(,"probabilities")  
 adecuado no.adecuado  
1 0.7328316 0.2671684  
Levels: no.adecuado adecuado

pronostico4 <- predict(mejor.modelo.rose, newdata3,probability = TRUE)  
pronostico4

1   
no.adecuado   
attr(,"probabilities")  
 adecuado no.adecuado  
1 0.2722259 0.7277741  
Levels: adecuado no.adecuado

Para los datos ingresados en las variables, el pronóstico con el mejor modelo sin muestreo lo clasifica como un peso “adecuado”, pero el pronóstico del modelo con remuestreo lo clasifica como un peso “no adecuado”, esto se debe a que la clasificación para pesos no adecuado era mala para el modelo tuneado, sin embargo esa probabilidad para clasificación mejoró con el remuestreo. Esto quiere decir que mi modelo no se encontaba clasificando de buena manera el peso de un nacido vivo que es menor a los 2500 kg.

predicciones <- attr(ajustados.mejor.modelo,  
 "probabilities")[,1]  
predicciones <- as.numeric(predicciones)  
predd <- factor(ifelse(predicciones>0.924,1,0))  
  
predictt <- attr(ajustadosrose,  
 "probabilities")[,1]  
predictt <- as.numeric(predictt)  
preddi <- factor(ifelse(predictt>0.824,1,0))  
  
matriz.comparacion <- data.frame(pronóstico.mejor.modelo.punto.corte=predd,  
 pronostico.remuestreo.punto.corte=preddi)  
head(matriz.comparacion,30)

pronóstico.mejor.modelo.punto.corte pronostico.remuestreo.punto.corte  
1 1 0  
2 0 0  
3 0 0  
4 0 0  
5 1 1  
6 1 0  
7 0 0  
8 1 0  
9 1 0  
10 1 0  
11 1 1  
12 1 0  
13 0 0  
14 1 0  
15 1 0  
16 1 0  
17 0 0  
18 1 1  
19 0 0  
20 1 0  
21 1 0  
22 1 0  
23 1 0  
24 1 0  
25 1 0  
26 1 0  
27 0 0  
28 0 1  
29 1 0  
30 0 0

Si comparamos el mejor modelo evaluado con el punto de corte óptimo y el modelo con el remuestreo por el método ROSE con el punto de corte óptimo podemos concluir lo siguiente:

1. En cuanto a la curva ROC en los dos modelos se podría decir que están clasificando bien.
2. El Accuracy para los modelos no difiere significativamente, solo existe entre ellos una diferencia de 0.0017.
3. La sensitividad entre los dos modelos difiere en 0.0059 no hay una significancia grande.
4. El modelo tuneado con el punto de corte óptimo clasifica con un valor de “0.9796” cuando el peso del nacido vivo es el adecuado y con el remuestreo y el punto de corte de óptimo clasifica co un valor de “0.7822”, esto quiere decir que el modelo sin remuestreo está pronosticando mejor cuando un niño va a nacer con un peso adecuado, sin embargo entre los dos modelos no hay una diferencia que sea tan significativa.
5. Hay una diferencia muy significativa para clasificar cuando un nacido vivo nacerá con un peso no adecuado, de casi 0.5, en este caso el modelo con remuestreo está pronósticando de mejor manera cuando un nacido vivo nacerá con un peso menor a 2500 kg.

En resumen por lo antes expuestos el mejor modelo en mi caso es la del método de remuestreo.