UNAP
FINESI

IV Semestre

Modelos Discretos

Capacidad Predictiva del Modelo
Curva ROC

Mtr. Alcides Ramos Calcina

Capacidad Predictiva del Modelo

Mtr. Alcides Ramos Calcina

Introducción

- Para comprobar la efectividad de un modelo en la clasificación de observaciones, se puede construir una tabla de clasificación donde se cruza el verdadero valor de la observación (1 ó 0), con la predicción de la misma según algún modelo que se considere.

- La predicción se suele hacer con respecto a un valor de referencia arbitrario p_0:

$$\hat{y}_i = 1 \quad \text{si} \quad \hat{p}_i > p_0$$

mientras

$$\hat{y}_i = 0 \quad \text{si} \quad \hat{p}_i \leq p_0$$

donde, p_0 es el valor de corte y habitualmente suele tomar el valor de $p_0 = 0.5$.

1. Capacidad predictiva

- Una forma común de evaluar la calidad de un modelo de regresión logística es crear una matriz de confusión.

- La matriz tiene la siguiente estructura.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

1. Capacidad predictiva

- Se resume la capacidad predictiva de un modelo de regresión logística mediante los siguientes conceptos.

a) La exactitud (accuracy). Es la proporción de clasificaciones correctas.

$$Exactitud = \frac{VP + VN}{n}$$

b) Tasa de error (Misclassification Rate). En general, ¿qué proporción de los datos clasifica incorrectamente

$$Tasa\ de\ error = \frac{FP + FN}{n}$$

1. Capacidad predictiva

c) Sensibilidad, exhaustividad, tasa de verdaderos positivos (sensivity, true positive rate, recall or hit rate). Es la proporción de casos positivos observados que son correctamente clasificados:

$$Sensitividad = \frac{VP}{VP + FN}$$

d) Especificidad, tasa de verdaderos negativos (Especificidad, tasa de verdaderos negativos). Es la proporción de casos negativos observados que son correctamente clasificados:

$$Especificidad = \frac{VN}{FP + VN}$$

1. Capacidad predictiva

e) Tasa de falsos positivos (1- especificidad). Es la proporción de casos negativos observados que son incorrectamente clasificados:

$$1 - Especificidad = \frac{FP}{FP + VN}$$

f) Precisión o valor predictivo positivo VPP (Precision, positive predictive value). Es la proporción de casos positivos que son correctamente clasificados:

$$VPP = \frac{VP}{VP + VN}$$

1. Capacidad predictiva

g) Valor de predicción negativo VPN (Negative predictive value). Es la proporción de casos negativos que son correctamente clasificados:

$$VPP = \frac{VN}{FN + VN}$$

h) Prevalencia (Prevalence). Es la proporción de casos positivos observados:

$$Prevalencia = \frac{VP + FN}{n}$$

2. Curva ROC y AUC

- Una Curva ROC (*Receiver Operating Characteristic*, o *Característica Operativa del Receptor*) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación.

- El área bajo la curva ROC (AUC) nos permite comparar la capacidad de clasificación de distintos modelos. Se utiliza en el análisis de clasificación con el fin de determinar específicamente que modelos utilizados predice las clases mejor.

En general:

  - Si ROC ≤ 0.5 el modelo no ayuda a discriminar
  - Si $0.6 \leq$ ROC ≤ 0.8 el modelo discrimina de forma adecuada
  - Si $0.8 <$ ROC ≤ 0.9 el modelo discrimina de forma excelente
  - Si ROC > 0.9 el modelo discrimina de forma excepcional

2. Curva ROC y AUC

- Analizaremos la capacidad predictiva del modelo correspondiente a nuestro ejemplo de angina de pecho.

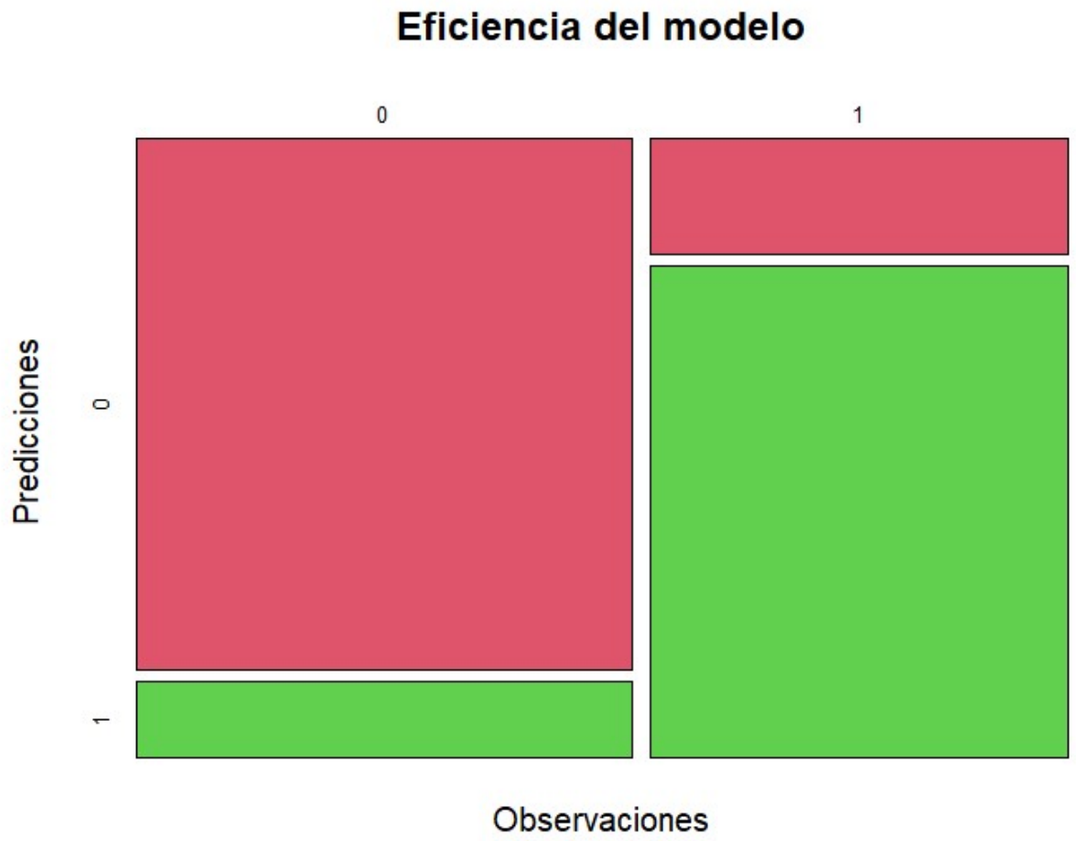
- Matriz de confusión utilizando un punto de corte o umbral (threshold) de 0.5.

```
pred <- ifelse(test = modelo2$fitted.values > 0.5, yes = 1, no = 0)
matriz_Conf <- table(modelo2$model$y, pred, dnn = c("Observaciones",
"Predicciones"))
matriz_Conf
             Predicciones
Observaciones  0  1
            0 71 10
            1 13 55
```

Para mayor claridad, visualizamos la matriz de confusión en forma de mosaico, el cual se muestra en la siguiente figura:

2. Curva ROC y AUC

```
mosaicplot(matriz_Conf, main = "Eficiencia del modelo", color = 2:7)
```



2. Curva ROC y AUC

- Para realizar una mejor interpretación del modelo, calculamos sus métricas, pero previamente es necesario instalar la librería "**caret**".

```
# install.packages("caret")
# library(caret)
observ <- as.factor(modelo2$model$y)
matriz<-confusionMatrix(observ, as.factor(pred))
matriz
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 71 10
         1 13 55

               Accuracy : 0.8456
                 95% CI : (0.7774, 0.8996)
    No Information Rate : 0.5638
    P-Value [Acc > NIR] : 1.813e-13

                  Kappa : 0.6878

 Mcnemar's Test P-Value : 0.6767

            Sensitivity : 0.8452
            Specificity : 0.8462
         Pos Pred Value : 0.8765
         Neg Pred Value : 0.8088
             Prevalence : 0.5638
         Detection Rate : 0.4765
   Detection Prevalence : 0.5436
      Balanced Accuracy : 0.8457

       'Positive' Class : 0
```

2. Curva ROC y AUC

- **Exactitud**: El valor obtenido para este modelo es de un 84.56%. No es muy bueno, pero podemos considerarlo aceptable).

- **Presición**: El valor obtenido para este modelo es de un 87.65%. Por tanto, nuestro modelo es más preciso que exacto.

- **Sensibilidad**: Representa, comose dijo anteriormente, la habilidad del modelo de detectar los casos relevantes. Un 84.52% es claramente un valor muy bueno para una métrica. Podemos decir que nuestro algoritmo de clasificación es muy sensible, es decir, “no se le escapan” muchos positivos.

- **Especificidad**: En este caso, la especificad también tiene un valor muy bueno de 84.62%. Esto significa que su capacidad de discriminar los casos negativos es muy buena. O lo que es lo mismo, es difícil obtener falsos positivos.

2. Curva ROC y AUC

- En este ejemplo, la sensibilidad es 84.52% y la especifidad es 84.62. Por tanto, este modelo esente relativamente más específico que sensible. Esta es la situación que nos interesa cuando nuestro objetivo es evitar a toda costa los falsos positivos.

- Con respecto a la curva de ROC y AUC, la graficamos haciendo uso de la librería "**pROC**"

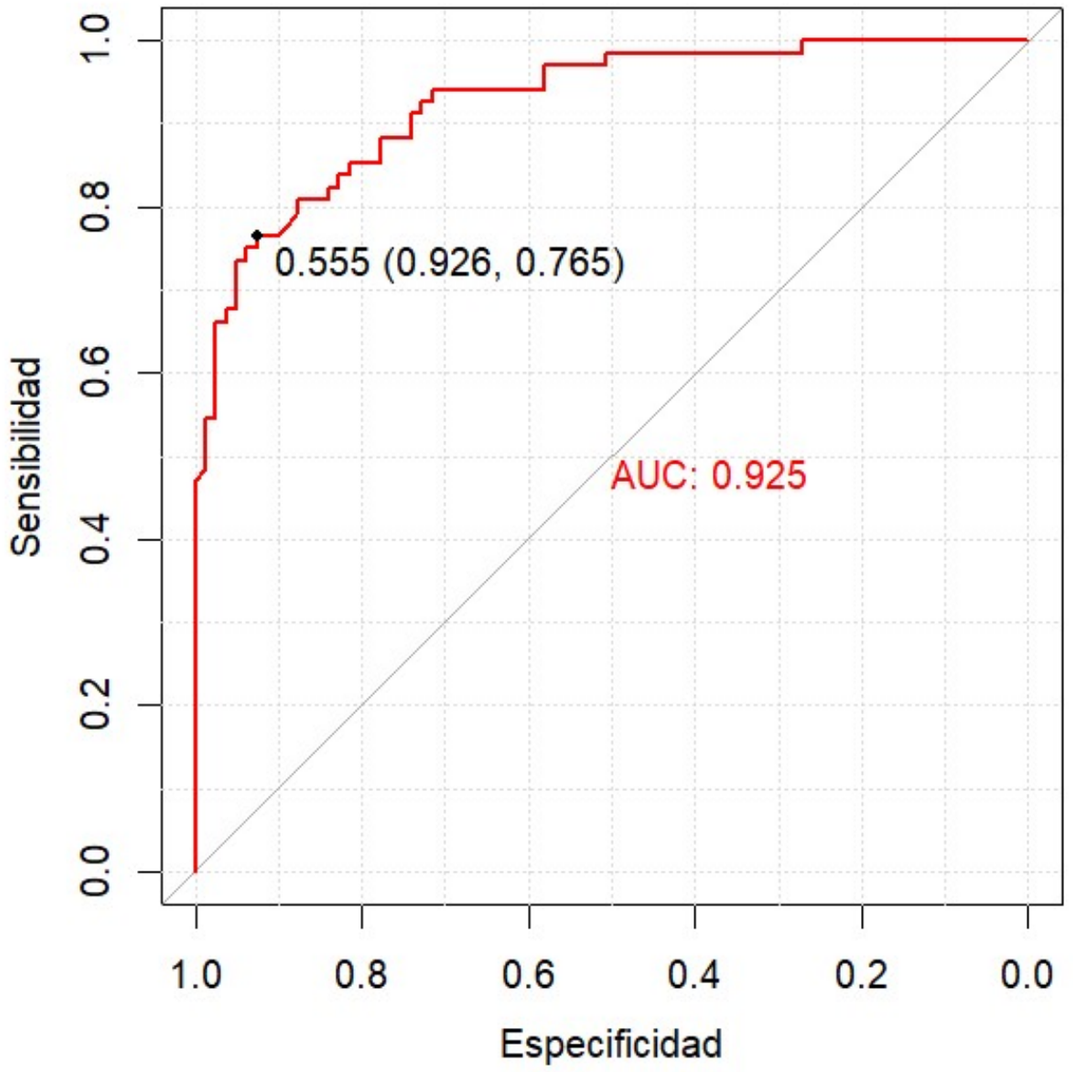
```
# install.packages("pROC")
# library(pROC)
par(pty = "s") # Hacer cuadrado el espacio ROC
predicciones <- modelo2$fitted.values
ROC_g <- roc(datos$y, predicciones, ci = F)
plot(ROC_g, print.auc=T, grid = T, print.thres = "best", col="red",
xlab="Especificidad", ylab="Sensibilidad")
```

Se muestra la siguiente figura:

2. Curva ROC y AUC

- La curva ROC para el modelo indica que este modelo tiene un buen poder discriminatorio, de hecho, el área bajo la curva (AUC) es 92.5%.

- En el punto de corte igual a 0.555, es decir que en 0.555 tendremos un balance entre las distintas métricas y por ello elegimos ese punto.



FINESI

Modelos Discretos

IV Semestre

https://aulavirtual2.unap.edu.pe/

GRACIAS