Técnicas de Clasificación

PRÁCTICA 2: Regresión logística.

V. Miguel Sempere Navarro

En la presente práctica se va a llevar a cabo un modelo de regresión logística estudiado con anterioridad en clase que tiene como objetivo clasificar si una familia tiene riesgo de pobreza o no.

Para ello conviene recordar que una regresión logística es un modelo lineal para explicar una variable dependiente en base a unas variables explicativas independientes que son categóricas. En este caso, la variable a explicar es “Hogar en riesgo de pobreza”.

En primer lugar, se comenzará el desarrollo de la práctica abriendo el dataset que se ha proporcionado y realizando una primera limpieza de variables que no se consideran adecuadas para explicar el modelo. Una de ellas será la variable “Renta”, tal y como se ha sugerido, entre otras. Se pasa de trabajar con 18 variables a 12 y se transforman todos los datos a tipo factor.

Posteriormente, se separa el dataset en dos muestras del 60% y 40% siendo la primera para entrenar el modelo y la segunda para testearlo. A partir de ahora, se trabaja con el modelo de “training”, es decir, con la muestra que supone el 60% de la población total y, realizando un modelo de regresión lineal y, seguidamente, un ANOVA, obtenemos como conclusión que las variables que mejor explican el modelo son: “Ayuda Familias”, “VacacionesOutdoor”, “CapacidadAfrontar”, “LlegarFinMes”, “Miembros”, “HogaresSemanales” y “ActMayor”.

Una vez seleccionadas las variables significativas se llevará a cabo el análisis de bondad de ajuste para observar la discrepancia entre los valores observados y los valores esperados en el modelo de estudio. El contraste de McFadden representa lo que se reduce la desviación, la cual es proporcional al aumento de verosimilitud del modelo. Nos arroja un resultado de 0.3789. Este estadístico indica que, cuanto más se acerque a 1, mejor ajustado estará el modelo.

A continuación, se realiza exactamente lo mismo, pero con la otra muestra (testing). Gracias a ella se comprobará si el modelo responde de manera óptima. Para ello, se

establece un umbral del 68% para la predicción, de tal forma que valores superiores a este umbral tomarán el valor 1 y 0 en caso contrario.

El modelo logit tiene la ventaja sobre los modelos de probabilidad, de que las probabilidades calculadas siempre están comprendidas entre 0 y 1, con lo cual se evita el tener que hacer aproximaciones a 0,01 cuando las probabilidades son negativas, o a 0,99 cuando son mayores a 1.

Para finalizar, es necesario crear una matriz de confusión que nos diga cuánto se aproxima el modelo o, mejor dicho, cuántas veces del total es capaz de predecir correctamente si una familia está en riesgo de pobreza o no.

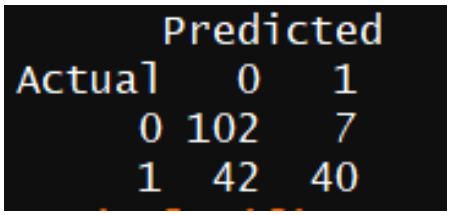

Ilustración 1. Matriz de confusión. Fuente: elaboración propia.

Se puede observar que el error es muy bajo, pues el modelo tan solo falla en 7 y 42 de las observaciones; frente a más de 100 y 40 de las mismas que sí acierta.

Una vez calculada la matriz de confusión, se observa la precisión del modelo realizando un “accuracy” que consiste en obtener la proporción entre las predicciones correctas que ha generado nuestro modelo y el total de las predicciones. Tras realizarlo, se obtiene una proporción de aproximadamente un 74%, lo que termina de validar el modelo realizado.