Ejercicio regresión logística

Henry Navarro

1. Credit-scoring

Cuando los bancos quieren dar un crédito a un cliente, han de estimar primero el riesgo de impago por parte del mismo. El objetivo de los sistemas de credit-scoring es modelizar o predecir la probabilidad de pago o impago por parte de un cliente con cierto factores de riesgo. Los métodos de credit scoring se han convertido en una técnica estándar no sólo para bancos, sino también para otras entidades financieras y grandes almacenes.

Uno de los primeros usos del credit scoring fue en el ámbito de las tarjetas de crédito y de ahí se extendio a los créditos personales y a las hipotecas. La mayoría de los problemas del credit scoring son más de tipo técnico que teórico. Para empezar. El primer problema se encuentra en los datos necesarios para el credit-scoring: deberían estar disponibles la mayor cantidad posible de variables relevantes, pero el conseguir estas bases de datos es bastante caro, por lo que es necesario alcanzar un equilibrio entre el coste de los datos y los errores que se pueden cometer si la información de la que se dispone no es suficiente. Los bancos consiguen la información de sus propias fuentes internas (por ejemplo, de las solicitudes de crédito anteriores), de fuentes externas (cuestionarios y entrevistas personales) y de otras fuentes. Para conocer la situación del solicitante, normalmente se recoge la siguiente información: edad, sexo, estado civil, nacionalidad, nivel educativo, número de hijos, salario, gastos, etc. Las variables que entran en el modelo utilizado para el credit-scoring deben ser elegidas con cuidado ya que la cantidad de datos puede ser elevada y computacionalmente compleja.

2. Descripción de los datos

Los datos corresponden a 800 clientes de un banco. Se dispone de 20 variables explicativas que pueden influir a la hora de conceder un crédito, de las cuales he seleccionado 7.

Descripción de las variables

- Y: es una variable dicotómica: 'bueno', 'malo'que indica si el cliente se considera un buen o mal pagador.
- X_1: es una variable categórica con los niveles 'no', 'cuenta.buena', 'cuenta.mala', que corresponde al estado del crédito de la cuenta actual del cliente.
- X_2: variable numérica que indica el número de meses de duración del crédito.
- X_3: variable categórica con niveles 'buen pagador previo', 'mal pagador previo', si el cliente ha sido buen o mal pagador con anterioridad.

- X_4: variable categórica con niveles 'personal', 'profesional', indicando la intención de uso del préstamo.
- X_5: variable numérica que indica la cantidad del préstamo (en una cierta unidad mo- netaria).
- X_6: variable categórica con niveles 'mujer', 'hombre', indicando el sexo del cliente
- X_7: variable categórica con niveles 'solo', no 'solo', que indica si la persona tiene pareja o no.

La base de datos debe dividirse en dos partes. Un conjunto de entrenamiento y otro de test.

3. Ejercicios

1. Ajusta 7 modelos de regresión logística (uno para cada una de las variables explicativas) y comenta los resultados.
2. Ajusta el modelo con todas las variables (ten en cuenta que la calidad de la cuenta existente del cliente y el comportamiento previo al pagar créditos anteriores pueden interaccional con otras variables). Utiliza el test de la razón de verosimilitud para determinar las variables que deben estar o no el modelo y compara los resultados lo que obtendría si utilizaras el criterio AIC.
3. Basado en el modelo final, obtén intervalos de confianza al 90% para el odds ratio de las variable que corresponden al tipo de uso del préstamo y al estado civil.
4. Interpreta los coeficientes del modelo obtenido en términos del odds ratio.
5. Realiza la predicción para tu modelo final con los datos de entrenamiento.
6. Calcula la matriz de confusión para los cutoff's siguientes: 0.3, 0.5, 0.8. Analiza los resultados. ¿Cuál cutoff seleccionarías? ¿Por qué? Sugerencia: utiliza una función
7. Grafica la curva ROC y analiza los resultados.
8. Halla un valor de cutoff óptimo para los datos de entrenamiento y grafica la curva ROC. Analice muy bien los resultados, ¿qué observa para el TPR y TNR?
9. Repita el ejercicio realizando una Cross-Validation, balanceando los datos y utilizando los datos de test para calcular todas las medidas de rendimiento. Para esto siga los pasos: a) balancee la muestra, b) cree un modelo con todas las variables, c) realice la predicción con el conjunto de test, d) Halle el valor de cutoff óptimo, e) aplique cross-validation.
10. ¿Podría mejorarse esta Cross-Validation?
11. ¡A divertirse con Machine Learning! ¿Eres buen pagador? Utiliza el modelo anterior para determinar si eres elegible para un préstamo del monto que quieras, a pagar en el tiempo que desees. Analiza los resultados.
12. Entrene el modelo con todos los datos. ¿Son los coeficientes estimados similares a los obtenidos sólo con la muestra de entrenamiento?, ¿deberían serlo?