Diplomado en Análisis de Datos |
Diplomado en Machine Learning

Curso: Aprendizaje Supervisado
Prof. Rolando de la Cruz

Tarea 2

Desarrollo obligatorio en grupos de 2, 3 o 4 integrantes.

Fecha de entrega: hasta las 23.59 hrs del 02/12/25 vía webcursos. Sólo 1 integrante del grupo debe subir la tarea.

DESCRIPCIÓN DEL PROBLEMA

Un vendedor de revistas está tratando de decidir qué revistas comercializar a los clientes. En los "viejos tiempos", esto podría haber implicado tratar de decidir a qué clientes enviar anuncios a través del correo regular. En el contexto de hoy y de la "web", esto podría implicar decidir qué recomendaciones hacer a un cliente que ve una página web sobre otros artículos en los que el cliente podría estar interesado y por lo tanto querer comprar. Los dos problemas son esencialmente los mismos.

En este caso, el sitio web NoExisto.com (nombre ficticio) quiere decidir qué revistas incluir en los correos electrónicos a los clientes como parte de una campaña de comercialización por correo electrónico. Todos los e-mails que se enviarán irán a clientes que hayan comprado previamente una suscripción a la revista en NoExisto.com y que no hayan optado por no recibir e-mails.

Las revistas anunciadas en cada correo electrónico se seleccionarán automáticamente de forma específica para cada cliente cuando se genere el correo electrónico, a fin de maximizar la probabilidad de que el cliente compre. NoExisto.com solo incluirá anuncios de tres revistas en cada correo electrónico en una fila en la parte superior del mensaje porque la administración cree que incluir más anuncios es ineficaz. NoExisto.com también cree que incluir solo tres anuncios hace mucho más probable que los anuncios aparezcan en la vista previa del correo electrónico del destinatario y por lo tanto sean realmente vistos (sin que el destinatario tenga que abrir el correo electrónico).

Debido a que todos los destinatarios de los correos electrónicos han hecho previamente una compra en NoExisto.com, la compañía puede cotejar los datos recopilados cuando el cliente hizo su compra anterior con los datos de terceros (que pueden ser comprados en fuentes de datos como las agencias de calificación de crédito), por lo que tienen bastante información sobre cada cliente. Por ejemplo, tienen datos como los ingresos, el número de

personas en el hogar, etc. Cada vez es más común este tipo de fusión de datos procedentes de múltiples fuentes para reunir un "perfil" notablemente rico de cada cliente.

Aquí están las variables que NoExisto.com tiene sobre cada cliente de fuentes de terceros:

- Ingreso familiar (Income; redondeado a los $1,000.00 más cercanos)
- Sexo (IsFemale = 1 si la persona es mujer, 0 en caso contrario)
- Estado civil (IsMarried = 1 si está casado, 0 en caso contrario)
- Educado en la universidad (HasCollege = 1 si tiene uno o más años de educación universitaria, 0 en caso contrario)
- Empleado en una profesión (IsProfessional = 1 si está empleado en una profesión, 0 en caso contrario)
- Jubilado (IsRetired = 1 si está jubilado, 0 en caso contrario)
- No empleado (Unemployed = 1 si no está empleado, 0 en caso contrario)
- Duración de la residencia en la ciudad actual (ResLength; en años)
- Doble Ingreso si está casado (Dual = 1 si es doble ingreso, 0 en caso contrario)
- Niños (Minors = 1 si hay niños menores de 18 años en el hogar, 0 en caso contrario)
- Propiedad de la vivienda (Own = 1 si es residencia propia, 0 en caso contrario)
- Tipo de residente (House = 1 si la residencia es una casa unifamiliar, 0 en caso contrario)
- Raza (White = 1 si la raza es blanca, 0 en caso contrario)
- Idioma (English = 1 es el idioma principal en el hogar es el Inglés, 0 de lo contrario)

¿Cómo puede NoExisto.com decidir qué revistas comercializar para cada persona, es decir, qué anuncios poner en cada e-mail? Una forma sería desarrollar una ecuación (aquí es donde entra en juego la regresión logística multivariante) que prediga la probabilidad de que un cliente compre una revista en particular basándose en los datos que la compañía tiene sobre el cliente. Tal ecuación se desarrollaría para cada revista que la compañía vende.

Si NoExisto.com tiene un modelo de este tipo para cada revista que vende, puede calcular la probabilidad de que el cliente compre para cada una de las revistas que ofrece. Luego pueden poner las tres revistas principales en el correo electrónico (es decir, las tres que el modelo predice que el cliente tiene más probabilidades de comprar).

Nota: NoExisto.com puede hacer cosas más complicadas que sólo mirar las probabilidades predichas (como mirar la ganancia esperada de la venta), pero para simplificar, supongamos que el objetivo es poner anuncios en el correo electrónico para las tres revistas que el cliente tiene más probabilidades de comprar.

Para poder desarrollar una ecuación que prediga la probabilidad de que un cliente compre una revista en particular, la empresa deberá realizar un experimento para recoger datos sobre el comportamiento de compra del cliente. Una forma de hacerlo es seleccionar al azar algunos clientes de la base de datos de clientes y luego enviarles correos electrónicos con anuncios seleccionados al azar. El hecho de que estos clientes compren o no las revistas

anunciadas puede proporcionar los datos necesarios para estimar las ecuaciones que se utilizarán para predecir la probabilidad de que un cliente compre una revista en particular.

Si se dispone de un gran número de revistas que se venden, puede ser necesario enviar un gran número de correos electrónicos para obtener ecuaciones de predicción útiles. Asegurarse de que se dispone de suficientes datos para que cada revista termine con una ecuación útil para predecir la probabilidad de compra puede ser un poco complicado (y requiere un gran número de correos electrónicos en el experimento), pero no se abordarán estos temas en este caso.

Así pues, el problema de decidir qué anuncios de revista colocar en cada correo electrónico se reduce a desarrollar una ecuación para cada revista que prediga la probabilidad de que un cliente compre. Vamos a centrarnos ahora en la cuestión de desarrollar dicha ecuación para una revista ("Niños Creativos") cuyo público objetivo son los niños de entre 9 y 12 años. En el proceso de envío de los e-mails "experimentales", el anuncio de " Niños Creativos" fue mostrado en 673 e-mails a los clientes y se registró el comportamiento de compra.

Además de las variables para cada cliente enumeradas anteriormente (las obtenidas de fuentes de terceros), NoExisto.com tiene las siguientes variables de sus propias bases de datos:

- Previamente compró una revista para padres (PrevParent = 1 si previamente compró una revista para padres, 0 en caso contrario).
- Previamente compró una revista infantil (PrevChild = 1 si previamente compró una revista infantil, 0 en caso contrario).

La variable dependiente Y proviene del "experimento", es decir, de los 673 correos electrónicos a los clientes que contienen el anuncio de "Niños Creativos" y si el cliente compró o no la revista. Es decir, la variable dependiente es:

- Compró "Niños Creativos" (Buy = 1 si compró "Niños Creativos", 0 en caso contrario)

El archivo “NiniosCreativos.csv” contiene la data descrita.

a) Divida la data en una muestra de entrenamiento (80%) y una muestra de validación (20%). Realice un análisis exploratorio de los datos que revelen cosas de interés para lograr el desarrollo del modelo de regresión logística.
b) Ajuste un modelo de regresión logística multivariado usando la data de entrenamiento considerando todas los atributos para predecir P(Buy=1|atributos). Comente resultados que sean interesantes destacar.
c) Usando técnicas de selección de variables encuentre un modelo más reducido. Comente los resultados.
d) Verifique si los atributos seleccionados en el modelo de la parte c) son significativos o no. Comente sus resultados.

e) Usando el modelo seleccionado en la parte d) calcule métricas de calidad predictiva (AUC, K-S, Sensibilidad, Especificidad, etc.) en la muestra de entrenamiento y en la muestra de validación. Comente y compare resultados.
f) Haga un gráfico de las curvas ROC para la muestras de entrenamiento y de validación. Comente resultados.
g) Ahora ajuste una regresión logística con regularización (ridge, lasso, elestic net) a los datos de entrenamiento. Calcule métricas de calidad predictiva (AUC, K-S, Sensibilidad, Especificidad, etc.) en la muestra de entrenamiento y en la muestra de validación. Comente y compare los resultados obtenidos con los obtenidos en la parte e).
h) De todos los modelos considerados, ¿Cuál es el mejor? ¿Por qué?
i) Discuta cómo usaría el modelo para cumplir los objetivos de NoExisto.com.

Subir a la plataforma un Notebook de Python (en Colab) con la solución de la tarea. Al momento de seleccionar las muestras de entrenamiento y testing use una semilla (con la función apropiada en Python) y deje la semilla utilizada en su Notebook.