Repeat Buyers Prediction

Minería de Datos

9 de septiembre de 2015

1. Introducción

Este trabajo tiene como objetivo promover la aplicación de técnicas avanzadas de investigación en Inteligencia Artificial a los problemas del mundo real. Se tendrá acceso a una gran cantidad de datos proporcionados por Tmall.com, la plataforma B2C más grande de China. El objetivo es aplicar técnicas avanzadas de aprendizaje automático y minería de datos para predecir que compradores repetirán su acción después de una promoción de ventas en Tmall.com.

2. Definición del Problema

Los comerciantes a veces hacen grandes promociones (por ejemplo, descuentos o cupones de efectivo) en fechas particulares (como "Boxing-day Sales", "Black Friday" o "Double 11") con el fin de atraer a un gran número de nuevos compradores. Desafortunadamente, muchos de los compradores atraídos son cazadores de solo una oferta, y estas promociones pueden tener poco impacto duradero en las ventas. Para aliviar este problema, es importante para los comerciantes lograr identificar quienes pueden convertirse en compradores reiterados. Al enfocarse en estos potenciales leales clientes, los comerciantes pueden reducir en gran medida el costo de promoción y mejorar el costo de la inversión. Es bien sabido en el campo de la publicidad online, que encontrar el cliente "objetivo" es extremadamente difícil, especialmente para los compradores nuevos. Sin embargo, con el registro de comportamiento de los usuarios acumulado a largo plazo por Tmall.com, puede que ayude a resolver este problema.

En este desafío, se dispone de un conjunto de comerciantes y sus correspondientes nuevos compradores adquiridos durante la promoción en el "Double 11 day". La tarea consiste en predecir cuál de los nuevos compradores dado los comerciantes se convertirán en clientes habituales en el futuro. En otras palabras, es necesario predecir la probabilidad de que estos nuevos compradores compren artículos de los mismos comerciantes de nuevo dentro de 6 meses.

3. Conjunto de Datos

El conjunto de datos contiene registros anónimos de compras de usuarios en los últimos 6 meses antes y en el "Double 11 day" además de la información que indica si son compradores repetidos.
Existen dos formatos para el conjunto, los cuales se detallan a continuación:

- Formato 1.
  - Registro del Comportamiento del Usuario.

Atributo	**Definición**
user_id	Un id único para el comprador.
item_id	Un id único para cada ítem.
cat_id	Un id único para la categoría donde el ítem pertenece.
merchant_id	Un id único para el comerciante.
brand_id	Un id único para la marca del ítem.
time_stamp	Fecha en que la acción se realiza (formato: mmdd).
action_type	Tipo enumerado {0,1,2,3}, donde 0 es para click, 1 es para añadir al carro, 2 es por comprar y 3 añadir a favorito.

  - Perfil de Usuario.

Atributo	**Definición**
user_id	Un id único para el comprador.
age_range	Rango de edad de los usuarios: 1 para <18; 2 para [18,24]; 3 para [25,29]; 4 para [30,34]; 5 para[35,39]; 6 para [40,49]; 7 y 8 para $\geq$ 50; 0 y NULL para desconocido.
gender	Sexo del usuario: 0 para femenino, 1 para masculino, 2 y NULL para desconocido.

  - Datos de Entrenamiento y de Prueba.

Atributo	**Definición**
user_id	Un id único para el comprador.
merchant_id	Un id único para el comerciante.
label	Es un tipo enumerado {0,1}, donde 1 es un comprador repetido, y 0 para el caso contrario. Este campo está vacío para los datos de prueba.

- Formato 2.

Atributo	**Definición**
user_id	Un id único para el comprador.
age_range	Rango de edad de los usuarios: 1 para <18; 2 para [18,24]; 3 para [25,29]; 4 para [30,34]; 5 para[35,39]; 6 para [40,49]; 7 y 8 para $\geq$ 50; 0 y NULL para desconocido.
gender	Sexo del usuario: 0 para femenino, 1 para masculino, 2 y NULL para desconocido.
merchant_id	Un id único para el comerciante.
label	Valores entre {0,1,-1,NULL}, 1 denota que el user_id es un comprador reiterado para el merchant_id, mientras que 0 es lo contrario. -1 representa que el user_id no es un nuevo cliente del comerciante, por tanto, está fuera de la predicción. Sin embargo, estos registros pueden proporcionar información adicional. NULL solo se da en los datos de prueba , indicando que es un par a predecir.
activity_log	Conjunto de registros de interacción entre {user_id, merchant_id}, donde cada registro es una acción representada como {item_id:category_id:brand_id:time_stamp:action_type}. # Se utiliza para separar dos elementos vecinos. Los registros no están ordenados en un orden particular.

4. Entrega de datos

El formato para la entrega de datos es el siguiente:

Atributo	**Definición**
user_id	Un id único para el comprador.
merchant_id	Un id único para el comerciante.
prob	Probabilidad predecida del usuario de convertirse en un comprador reiterado del comerciante dado. El valor debe estar entre 0 y 1.