75.06/95.58 Organización de Datos
Primer Cuatrimestre de 2019
Trabajo Práctico 2: Enunciado

El segundo trabajo práctico es una competencia de Machine Learning en donde cada grupo debe intentar determinar, para cada dispositivo presentado por Jampp, el tiempo que transcurrirá hasta que el mismo aparezca nuevamente en una subasta, y el tiempo hasta que el usuario del mismo decida instalar una nueva aplicación.

La competencia se desarrollará en la plataforma de Kaggle. En el siguiente link se provee la siguiente información para la competencia: http://www.kaggle.com/c/jampp-at-fiuba

Los datos a analizar estan disponibles en
https://drive.google.com/open?id=1-HKn0Pw4irUVrK2rrYFsxkjcsG5C3YNa

Glosario

Convertir: el objetivo de mostrar publicidad es que un dispositivo instale una aplicación, a ese evento se le llama conversión.
Dispositivo: *entidad con un id de publicidad asociado. Por ejemplo: un celular Samsung J6 con Android tiene un id único, un Apple iPhone tiene un identificador único.*
Evento: cualquier tipo de acción categorizada dentro de una aplicación. *Por ejemplo, en una aplicación de e-commerce un funnel de eventos muy común puede ser del estilo "abrir_app" → "buscar_producto" → "revisar_catalogo" → "agregar_a_carrito" → "efectuar_compra". Cada uno de estos pasos es un* ***evento****.*
Subasta: en el momento que una aplicación quiere mostrar una publicidad, ese espacio se vende en una subasta (generalmente de segundo precio) donde todos los interesados en mostrar una publicidad ofertan un precio y gana quién más ofrece.

Planteo del problema

Al momento de apostar en subastas RTB, suele resultar útil conocer los hábitos de un dispositivo para saber cuánto (y si) conviene apostar en la misma. Teniendo una estimación de cuándo será la

próxima vez que se vea un dispositivo y una estimación de la próxima vez que el dispositivo convierta orgánicamente, se pueden elaborar estrategias de apuestas, con mayor énfasis en aquellos dispositivos cercanos a convertir y menor en aquellos que les falta mucho tiempo para convertir. Una estrategia posible, por ejemplo, sería apostar en subastas dónde sabemos que al dispositivo en cuestión le falta poco para convertir pero que no lo veremos de vuelta antes de ese evento. De modo que si ganamos la subasta se nos atribuye la instalación.

El problema involucra dos partes:

- En un instante dado, estimar $S_t(d)$ el tiempo hasta que un dispositivo d aparezca de vuelta en una subasta RTB
- En un instante dado, estimar $S_c(d)$ el tiempo hasta que un dispositivo d convierta

Cada una de estas partes puede verse como un problema de survival analysis. Es importante notar que ambos valores deben ser no negativos.

TL;DR: *Se dice survival analysis a un conjunto de técnicas para analizar datos donde la variable objetivo es el tiempo hasta que ocurra un evento de interés.*

Ejemplo

Martín tiene un Moto G6 cuyo android_advertising_id es `75244996-bc3b-4374-9249-57a6f19e4091`. En los últimos 7 días se le han mostrado publicidades dentro de App1, App2 y App3. A Jampp le han llegado esas subastas y tenemos datos del comportamiento de Martín dentro de las aplicaciones que son clientes de Jampp y Martín se ha instalado. Sabemos a partir de cuáles publicidades que le hemos mostrado se instaló nuevas aplicaciones. Lo que queremos es, con toda esta data, en un momento dado poder responder las siguientes preguntas:

- Cuándo vamos a ver a Martín de nuevo en una subasta? Es decir, cuando se le podrá intentar mostrar una publicidad?
    - Es el equivalente a $S_t($`75244996-bc3b-4374-9249-57a6f19e4091`$)$
- Cuánto tiempo falta para que Martín esté dispuesto a instalar una nueva aplicación?
    - Es el equivalente a $S_c($`75244996-bc3b-4374-9249-57a6f19e4091`$)$

Set de entrenamiento

- Todos los datasets corresponden al período del 2019-02-08 al 2019-03-09 (incluidos)
- Hay un total de 4 datasets
    - Dataset de clicks de publicidades mostradas
    - Dataset de instalaciones de aplicaciones a partir de publicidades mostradas
    - Dataset de eventos intra-app

- Dataset de subastas RTB

Salida esperada

El archivo de salida que se espera es un csv con las siguientes columnas:

- Device_id [_st, _sc]
- Predicted_Value

Para construir este archivo, tendran que tomar los device_id dados, y para cada uno calcular:

- $S_t(d)$
- $S_c(d)$

Luego, armar un archivo con el device_id y cada uno de esos valores (para que el device_id sea clave se le concatenará _st o _sc segun sea el caso).

Los tiempos $S_t(d)$ y $S_c(d)$ se miden desde las 0:00 del 2019-03-09 en segundos. Se provee un archivo con la lista de devices_id objetivo, para los cuales se debe calcular S_t y S_c. Los valores objetivo se calcularon con un máximo de 7 días. Los device_id pueden obtenerlos del archivo target_competencia.csv