75.06/95.58 Organización de Datos
Segundo Cuatrimestre de 2018
Trabajo Práctico 1: Enunciado

Realizaremos un análisis de datos sobre un conjunto de eventos de web analytics de usuarios que visitaron www.trocafone.com, su plataforma de ecommerce de Brasil. Trocafone es un side to side Marketplace para la compra y venta de dispositivos electrónicos que se encuentra actualmente operando en Brasil y Argentina.

La empresa realiza distintas actividades que van desde la implementación de plataformas de trade-in (conocidos en la Argentina como Plan Canje), logística directa y reversa, reparación y recertificación de dispositivos (refurbishing) y venta de productos recertificados por múltiples canales (ecommerce, marketplace y tiendas físicas).

Para conocer más de su modelo de negocio, pueden visitar el siguiente artículo:
https://medium.com/trocafone/el-maravilloso-mundo-de-trocafone-5bdc5761856b

Para acotar el alcance del trabajo práctico analizaremos un subconjunto de eventos de web analytics que Trocafone tiene disponible para una cantidad representativa de usuarios que visitaron su plataforma de ecommerce.

El link para obtener el set de datos para el TP1 es el siguiente:
https://drive.google.com/file/d/1gUddcLLujjFfwZslypUv1LESTM6KiwJn/view?usp=sharing

En el link pueden encontrar el archivo events.csv el cual contiene las siguientes columnas:

- **timestamp:** Fecha y hora cuando ocurrió el evento. (considerar BRT/ART).
- **event:** Tipo de evento
- **person:** Identificador de cliente que realizó el evento.
- **url:** Url visitada por el usuario.
- **sku:** Identificador de producto relacionado al evento.
- **model:** Nombre descriptivo del producto incluyendo marca y modelo.
- **condition:** Condición de venta del producto
- **storage:** Cantidad de almacenamiento del producto.
- **color:** Color del producto
- **skus:** Identificadores de productos visualizados en el evento.
- **search_term:** Términos de búsqueda utilizados en el evento.
- **staticpage:** Identificador de página estática visitada
- **campaign_source:** Origen de campaña, si el tráfico se originó de una campaña de marketing
- **search_engine:** Motor de búsqueda desde donde se originó el evento, si aplica.
- **channel:** Tipo de canal desde donde se originó el evento.

- **new_vs_returning:** Indicador de si el evento fue generado por un usuario nuevo (New) o por un usuario que previamente había visitado el sitio (Returning) según el motor de analytics.
- **city:** Ciudad desde donde se originó el evento
- **region:** Región desde donde se originó el evento.
- **country:** País desde donde se originó el evento.
- **device_type:** Tipo de dispositivo desde donde se genero el evento.
- **screen_resolution:** Resolución de pantalla que se está utilizando en el dispositivo desde donde se genero el evento.
- **operating_system_version:** Version de sistema operativo desde donde se origino el evento.
- **browser_version:** Versión del browser utilizado en el evento

Por otro lado, los siguientes tipos de eventos se encuentran disponibles (en el campo event) sobre los cuales se brinda una breve descripción:

- **“viewed product”:** El usuario visita una página de producto.
- **“brand listing”:** El usuario visita un listado específico de una marca viendo un conjunto de productos.
- **“visited site”:** El usuario ingresa al sitio a una determinada url.
- **“ad campaign hit”:** El usuario ingresa al sitio mediante una campana de marketing online.
- **“generic listing”:** El usuario visita la homepage.
- **“searched products”:** El usuario realiza una búsqueda de productos en la interfaz de búsqueda del site.
- **“search engine hit”:** El usuario ingresa al sitio mediante un motor de búsqueda web.
- **“checkout”:** El usuario ingresa al checkout de compra de un producto.
- **“staticpage”:** El usuario visita una página
- **“conversion”:** El usuario realiza una conversión, comprando un producto.
- **“lead”:** El usuario se registra para recibir una notificación de disponibilidad de stock, para un producto que no se encontraba disponible en ese momento.

Algo a tener en cuenta es que no todos los datos descritos en las columnas corresponde a todos los tipos de eventos.

Por otro lado es importante tener en cuenta que el carácter temporal de los eventos para un usuario nos indican una progresión de los mismos en el tiempo, por lo cual es factible analizarlos de esa forma.

Por último, deberian tener en cuenta que dado que los usuarios están navegando catálogos de productos hacia un flujo de compra en una página de producto es posible derivar información entre eventos para poder enriquecer unos y otros.

El objetivo del primer TP es realizar un análisis exploratorio del set de datos del TP. Queremos ver qué cosas podemos descubrir sobre los datos que puedan resultar interesantes. Los requisitos de la primera entrega son los siguientes:

- El análisis debe estar hecho en Python Pandas o R.
- El análisis debe entregarse en formato papel en una carpeta en donde se incluya el reporte completo y todas las visualizaciones generadas. Es altamente recomendable que las visualizaciones se impriman en color.
- Informar el link a un repositorio Github en donde pueda bajarse el código completo para generar el análisis.
- Agregar en Kaggle un kernel con el análisis exploratorio realizado.

La evaluación del TP se realizará en base al siguiente criterio:

- Originalidad del análisis exploratorio.
- Calidad del reporte. ¿Está bien escrito? ¿Es claro y preciso?
- Calidad del análisis exploratorio: qué tipo de preguntas se hacen y de qué forma se responden, ¿es la respuesta clara y concisa con respecto a la pregunta formulada?
- Calidad de las visualizaciones presentadas.
    - ¿Tienen todos los ejes su rótulo?
    - ¿Tiene cada visualización un título?
    - ¿Es entendible la visualización sin tener que leer la explicación?
    - ¿El tipo de plot elegido es adecuado para lo que se quiere visualizar?
    - ¿Es una visualización interesante?
    - ¿El uso del color es adecuado?
    - ¿Hay un exceso o falta de elementos visuales en la visualización elegida?
    - ¿La visualización es consistente con los datos?
- Conclusiones presentadas.
    - ¿Presenta el grupo un listado de "insights" aprendidos sobre los datos en base al análisis realizado? ¿Es interesante?
    - ¿Pudieron descubrir features en el campo **‘model’**? ¿Cuales fueron?
    - ¿Identificaron patrones o funnels de usuarios que realizan checkouts/conversiones en Trocafone?
        - ¿Se comportan de forma distinta dependiendo del tipo de dispositivo desde el cual acceden?
        - ¿Se comportan de forma distinta dependiendo del tipo de fuente de tráfico al que pertenecen?
    - ¿Realizaron algún análisis sobre búsquedas que realizan los usuarios y las keywords que utilizan apoyándose en algún tipo de visualización?
    - ¿Realizaron algún análisis de lugar donde se originan las visitas de los usuarios de Trocafone (a nivel país, regiones más importantes o ciudades más importantes) apoyándose en algún tipo de visualización?

- ¿Pudieron descubrir features jerarquizando información de alguno de los campos (por ejemplo “screen_resolution”)?
- ¿El análisis realiza un aporte a Trocafone?

El grupo que realice el mejor análisis exploratorio obtendrá 10 puntos para cada uno de sus integrantes que podrán ser usados en el parcial además de ser publicado en el repositorio de la materia como ejemplo para los siguientes cuatrimestres.