Prueba data scientist

2020

BCN MAD VLC SCL BOG MDE LIM MEX MIA SFO

Introducción

La presente prueba persigue ofrecer a l@s candidat@s al puesto de data scientist la posibilidad de demostrar sus distintas habilidades en dos enunciados que simulan posibles tareas del día a día en esta posición.

Para la misma se propone una dedicación de entre 8 y 15 horas en total, aunque evidentemente no llevaremos a cabo ningún control sobre el tiempo de ejecución. Se agradecerá que l@s candidat@s compartan el tiempo que le han dedicado junto con la entrega de resultados para una mejor evaluación del trabajo realizado.

Al final de cada enunciado se detalla lo que se debe entregar en cada caso.

Enunciado 1

Se entiende como **conversion rate** el ratio de transacciones realizadas por sesión en un periodo de tiempo x. Esta métrica es una métrica muy utilizada en los e-commerce para hacer una valoración del funcionamiento del negocio. Sobre el valor de esta métrica se toman decisiones de diseño, de optimización y a veces de inversión. El objetivo principal es tener una CR tan alto como sea posible y a su vez procurar que no haya caídas. Cuando se produce una caída de CR puede ser producida por diferentes factores: menos usuarios que acuden a nuestro site, creación de campañas sobre audiencias con baja propensión de compra, aumento del precio, etc..

El caso que nos ocupa se focaliza sobre las audiencias. Se requiere contestar a la siguiente pregunta: ¿qué tipo de impacto tiene el aumento o reducción de un tipo de audiencia sobre el CR?.

En este test adjuntamos un dataset con datos históricos a nivel diario desde enero del 2019. A continuación detallamos los metadatos del mismo:

Variable	**Tipo**	**Descripción**
date	String	Fecha en formato YYYYMMDD
channelGrouping	String	Agrupación de tráfico por canal de entrada
userAgeBracket	String	Rango de edad
userType	String	Tipología de usuario por recurrencia: New Visitor / Returning Visitor
sessions	int	Número de sesiones
transactions	int	Número de transacciones

Así pues viendo la cabecera de los datos tenemos la siguiente información:

```
date,channelGrouping,userAgeBracket,userType,sessions,transactions
20190101,(Other),25-34,New Visitor,17,0
20190101,(Other),25-34,Returning Visitor,30,2
20190101,(Other),35-44,New Visitor,11,0
20190101,(Other),35-44,Returning Visitor,71,2
20190101,(Other),45-54,Returning Visitor,11,0
20190101,(Other),55-64,Returning Visitor,10,0
20190101,Direct - Non Paid,18-24,New Visitor,48,0
20190101,Direct - Non Paid,18-24,Returning Visitor,90,2
20190101,Direct - Non Paid,25-34,New Visitor,288,3
20190101,Direct - Non Paid,25-34,Returning Visitor,651,14
20190101,Direct - Non Paid,35-44,New Visitor,310,2
20190101,Direct - Non Paid,35-44,Returning Visitor,881,13
20190101,Direct - Non Paid,45-54,New Visitor,182,2
20190101,Direct - Non Paid,45-54,Returning Visitor,562,5
20190101,Direct - Non Paid,55-64,New Visitor,110,0
```

Esto se lee de la siguiente manera:

> *fila1: 20190101,(Other),25-34,New Visitor,17,0*
>
> La fila 1 nos indica que en el día 1 de enero del 2019 hubo 17 sesiones y 0 transacciones de gente entre 25-34 años que entraron por el canal Other y era su primera visita al web.
>
> *fila 15: 20190101,Direct - Non Paid,55-64,New Visitor,110,0*
>
> La fila 15 nos indica que en el día 1 de enero del 2019 hubo 110 sesiones y 0 transacciones de gente entre *55-64* años que entraron por el canal Direct - Non Paid y era su primera visita al web.

Se requiere de **una solución que nos ayude a explicar cómo los cambios de sesiones en función de la audiencia afectan al convertion rate del día**, es decir, poder crear unos insights que de información relevante como: por cada mil sesiones que se pierde del grupo de Returning Visitors entre 45-54 años que entran por newsletter se pierde x décimas de CR o bien por cada mil sesiones recurrentes de 45-54 años se pierden y décimas de CR.

Para facilitar la compresión de la solución, en el caso de canal, se requiere que se agregue la información por el primer ítem que aparece a la izquierda en la descripción, de manera que consideremos sólo el primer nivel de descripción: (Other), Direct, Email, Social….

En este caso se deberá entregar el código R o Python debidamente comentado en detalle más una pequeña presentación autoexplicativa de los resultados en el formato que desees y realizada con la herramienta con la que te sientas más cómod@.

Enunciado 2

En un conjunto de 100 tiendas físicas de retail tenemos instalada nuestra tecnología de digital sigange y audience analytics. La instalación, en cada una de las tiendas, consta de 10 pantallas ubicadas en distintas zonas de la tienda (3 escaparates, entrada, cola caja, entrada probadores, zona mujeres, zona hombres, zona niños, zona complementos) de las cuales se controla la emisión de la totalidad del contenido mediante nuestro sistema, midiéndose por completo. Cada una de estas pantallas está acompañada de una cámara frontal que permite medir las visualizaciones de contenido y características de las personas que realizan dichas visualizaciones. Además, la instalación contiene cámaras cenitales de otra tecnología en el techo de la tienda que permiten medir, usuario a usuario y con un ID para cada uno, las visitas y su movimiento por el espacio correspondiente a toda la superficie de cada una de las tiendas.

Los metadatos de las BBDD que disponemos de las distintas fuentes de información son estos:

- Contenido emitido por las pantallas y cámaras frontales ubicadas en las pantallas:

Variable	**Descripción**
Store ID	Identificador único de tienda
Display ID	Identificador único de pantalla
Content ID	Identificador único de contenido emitido
Content tags	Metainformación conceptual del contenido
Start ts	Time stamp del inicio de impresión del contenido
End ts	Time stamp del final de impresión del contenido
Age	Rango de edad estimado por la cámara frontal ubicada encima de la pantalla(<15, 15-25, 26-45, 46-60, >60)
Gender	Género estimado por la cámara frontal ubicada encima de la pantalla (masculino o femenino)
Expression	Expresión detectada estimado por la cámara frontal ubicada encima de la pantalla (sonrisa, neutra, seria o disgusto)
Views	Número de miradas detectadas a esa emisión concreta de contenido en esa pantalla concreta realizadas por personas con esas características de edad, género y expresión

- Cámaras cenitales:

Variable	**Descripción**
Store ID	Identificador único de tienda
Zone ID	Identificador único de zona de la tienda (podemos definir cuantas queramos y de la forma que queramos en el momento de la instalación del sistema)
User ID	Identificador único del usuario que visita la tienda. Este identificador se mantiene a través de las distintas zonas, desde que el usuario entra por la puerta hasta que sale de la tienda
Entrance ts	Time stamp del momento en el que el usuario entra a una zona concreta
Exit ts	Time stamp del momento en el que el usuario sale a una zona concreta
Gender	Género estimado estimado por la tecnología de las cámaras cenitales (masculino o femenino)

Supón que en el equipo de Data Science vamos a hacer un ejercicio de brainstorming de posibles soluciones avanzadas (estadística, Machine Learning, Inteligencia Artificial, etc.) que podríamos añadir a nuestro producto dados estos datos y otros que podamos obtener de fuentes o APIS públicas.
¿Qué propuestas harías tú? Explícalas a nivel conceptual y nombra la técnica o técnicas estadísticas que utilizarías en cada caso.

METRIPLICA HEADQUARTERS

barcelona@metriplica.com

www.metriplica.com