Adsmovil
MOBILE ADVERTISING SOLUTIONS

Prueba Técnica - Científico de Datos

Generalidades

Bienvenidos a la Prueba Técnica para aspirar al rol de Científico de Datos. En el sector de marketing digital, las impresiones (publicaciones de banners publicitarios en páginas o aplicaciones web), son dadas por subastas, lo cual naturalmente lleva al siguiente problema de negocio.

Requerimiento:
Al momento de presentar un anuncio publicitario en un dispositivo móvil, se realiza una subasta. Se desea ofertar al **menor precio posible**, con la condición de asegurar **ganar las subastas con mayor probabilidad de click**.

La base de datos en formato .parquet contiene información de tres campañas publicitarias diferenciadas por el campo *campaign_id.* Cada fila indica la información de una solicitud de impresión, ya sea que se haya ganado o no. En caso de haberse ganado, la columna *bid_won* tendrá valor 1.

Los datos nulos pueden venir tanto como un campo `NULL`, como un campo `UNKNOWN`, `POSITION_UNKNOWN` (para ad_position), o `-1` en caso que el campo sea numérico. Considere esto al momento de hacer limpieza de datos.

Descripción de los datos

Las características a usar en el modelo son:

campaign_id	Número de identificación de la campaña publicitaria
ad_position	Posición en el sitio web o App.
ad_size	Tamaño del creativo, expresado como <width>x<height>
app_bundle	Paquete de aplicación o nombre del paquete (ejemplo, com.foo.mygame)
day_of_the_week_utc	Día de la semana, entre 1 (domingo) y 7 (sábado) en zona horaria utc
display_manager	Nombre de la tecnología o player responsable de la renderización del ad
domain	dominio web
hour_of_day_utc	Hora del día entre 0 y 23
inventory_interstitial	La impresión acepta creativos intersticiales; 1 indica sí, 0 indica no
inventory_source	Nombre del Exchange
placement_type	Si el espacio fue designado para banner, video, o ambos.
platform_bandwidth	Si el navegador está usando wifi o un operador para establecer la conexión
environment_type	Tipo de entorno (APP o WEB)
platform_carrier	Si el dispositivo está utilizando un operador de telefonía móvil

Adsmovil
MOBILE ADVERTISING SOLUTIONS

platform_browser	Nombre del navegador
platform_os	Sistema Operativo del dispositivo
platform_js	¿Admite JavaScript el navegador? 1 para sí, 0 para no.
content_language	Lenguaje del contenido (como lo declara el publisher)
platform_os_version	Versión del Sistema Operativo del navegador
bid_price	Precio ofertado por la impresión (en dólares)
bid_date	Fecha de la puja
bid_won	La puja fue ganada? 1 indica sí, 0 indica no.
clicks	Cantidad de clicks en la impresión. En caso de no ganarse la impresión,el campo es nulo

Variable Objetivo: **Click**

Nota: Durante el bid request, **solo es posible usar las columnas 2 a 18**, es decir que su modelo únicamente podrá usar estas columnas para predecir:

```
ad_position , ad_size, app_bundle, day_of_the_week_utc,
display_manager, domain, hour_of_day_utc,
inventory_interstitial, inventory_source, placement_type,
platform_bandwidth, environment_type, platform_carrier,
platform_browser, platform_os, platform_js, content_language,
platform_os_version
```

1. EDA

Inicia con un análisis exploratorio de los datos. Da una descripción estadística y responde las siguientes preguntas:

¿Existe alguna asociación o relación entre variables?
¿Qué diferencias estadísticas existen entre las campañas?
¿Es adecuado aplicar modelos entrenados con datos de una campaña para predecir sobre otra campaña distinta?

2. Ingeniería de Variables

Si realiza transformaciones, descarta variables, o imputa valores en algún campo, describa el proceso.
¿De las 18 columnas cuáles son las características más importantes?
¿Cuántas variables sugiere usar? Explica el método de selección usado.

3. Modelo de probabilidad de click

Tienes como objetivo crear un modelo predictivo que permita predecir la **probabilidad de click**. El modelo para utilizar es libre y puedes usar todas las campañas o filtrar por alguna; Sin embargo, ten presente que valoramos

la comparación entre distintos caminos.
Nota: En este caso, es adecuado filtrar solo por pujas ganadas, ya que no hay información de click en las pujas pérdidas.

4. Desempeño de predicción y CTR

Agrupa los datos usando las características seleccionadas en el modelo. Crea una columna llamada CTR de acuerdo a la fórmula

$$CTR = \left(\frac{Total\ de\ Clicks}{Total\ de\ Impresiones}\right) \times 100$$

Compara las predicciones entre el modelo y el CTR. Si son distintas, ¿es preferible usar la data histórica o el modelo entrenado?

5. Precio mínimo ofertado

Usando la información de la columna Bid_Price, y su modelo, determine para las impresiones, cuáles podrían disminuir el precio de puja, y cuáles aumentar. Un estándar de la industria para este análisis es usar el win_rate

$$Win\ Rate = \left(\frac{\#\ de\ impresiones\ ganadas}{\#\ total\ de\ Impresiones}\right) \times 100$$

6. NLP: Nombre de las aplicaciones (Opcional)

Si usaste las columnas de `app_bundle` o `domain`, en caso de verlo necesario ¿usaste algún método de reducción de cardinalidad? ¿usaste alguna metodología para agrupar nombres similares?

¿Qué evaluamos?

El desafío busca evaluar distintos aspectos como:

- Capacidad analítica y de exploración de datos.
- Visualización de resultados.
- Conocimientos en técnicas de generación de features y modelado.
- Análisis de performance del modelo.
- Buenas prácticas de desarrollo, uso de funciones, módulos, código reproducible, etc.

La solución se enviará a más tardar el 21 de febrero al correo juan.munoz@adsmovil.com Adjuntando el código utilizado, así como la interpretación de los resultados. Se agendará una sesión virtual para conversar sobre el desarrollo.

Bienvenida toda tu creatividad.

Adsmovil
MOBILE ADVERTISING SOLUTIONS

¡Éxitos!