media.monks

JR DATA SCIENTIST - TECHNICAL CHALLENGE

INTRODUCCIÓN

El ejercicio de evaluación para la posición de JR *data scientist* consta de 2 partes. En la primera evaluaremos tu conocimiento en procesos básicos de la ciencia de datos a partir del lenguaje que te resulte más cómodo entre Python y R. En el segundo, haremos un chequeo rápido de tu práctica con el lenguaje SQL para consulta de bases de datos. Para este ejercicio (Python / R) necesitarás estas 2 cosas:

Dataset para Python o R

Google Form para llenar las respuestas

Una vez que completes el Google Form, serás contactado/a en el transcurso de los siguientes días para comunicarte los resultados y próximos pasos.

EJERCICIO

Introducción

Para esta evaluación vamos a trabajar con un dataset de Google Analytics. Esta es una herramienta basada en *1st party cookies* que permite realizar un seguimiento del comportamiento de los usuarios en un sitio (páginas visitadas, acciones realizadas) así como también la procedencia de los mismos.

En particular, vamos a trabajar con un dataset disponible públicamente que proviene de las mediciones hechas en el sitio Merchandise Store de Google. Esta información fue anonimizada para preservar la identidad de los usuarios. Acá tenés información sobre el esquema del *dataset* aunque para la evaluación vamos a utilizar solamente un *subset* del mismo al cual vas a poder acceder en formato .csv con el siguiente link:

DATASET

Los datos

El *subset* con el que vas a trabajar está compuesto por 6 meses de información de los siguientes campos:

fullVisitorId	STRING	An identifier for each visitor (user).
visitNumber	INTEGER	The session number for this user. If this is the first session, then this is set to 1.
date	STRING	The date of the session in YYYYMMDD format.
bounces	INTEGER	For a bounced session (a session without interactions from the user), the value is 1, otherwise it is null.
hits	INTEGER	Total number of hits (actions taken by the user, either visiting pages or clicking buttons) within the session.
pageviews	INTEGER	Total number of pageviews (pages visited by the user) within the session.
timeOnSite	INTEGER	Total time of the session expressed in seconds.
transactionRevenue	INTEGER	Total transaction revenue
transactions	INTEGER	Total number of ecommerce transactions within the session.
source	STRING	Could be the name of the search engine, the referring hostname, or a value of the utm_source URL parameter.
channelGrouping	STRING	Group sources into channels according to general rules.
browser	STRING	The browser used (e.g., "Chrome" or "Firefox").
deviceCategory	STRING	The type of device (Mobile, Tablet, Desktop).
country	STRING	The country of the session, based on IP address.
city	STRING	The city, based on IP addresses or Geographical IDs.

El objetivo

Con este *dataset* suponemos que como parte del equipo de Data Science te asignan un nuevo cliente. En la reunión de inicio del proyecto, el cliente te plantea:

> *"Tengo estos datos que me pasó el equipo de BI. Vienen de Google Analytics y muestran cómo interactúan los usuarios con nuestro sitio. Yo no tengo mucha experiencia con estos datos pero desde la Gerencia me están pidiendo que para la semana que viene entregue un reporte que explique quiénes son los usuarios, qué hacen en el sitio y cualquier otro dato relevante que pueda obtener. Yo se que a la Gerencia hay cosas que le interesan particularmente. Por ejemplo, poder predecir cuántas ventas va a haber en el sitio el próximo mes para saber si estamos camino a cumplir con los objetivos del trimestre o no. O poder predecir si un usuario va a convertir o no para incluirlo o excluirlo de las campañas de marketing que realiza nuestro equipo. Otra cosa que se que les interesa es clasificar a los usuarios en grupos en base al comportamiento que hayan tenido en el sitio, para poder segmentar mejor las campañas. ¿Ustedes podrán ayudarme con esto?"*

Las consignas

Entonces, a partir del *dataset*, tenés **7 días** para resolver las siguientes consignas:

1) **EDA**

Realizar un análisis exploratorio de los datos. Deberás mostrar como máximo 5 *charts* (gráficos o tablas), los que te parezcan más relevantes para describir los datos.

2) **MODELO**

Elegir uno de los siguientes problemas a resolver o contestar con los datos:

a) Predecir la cantidad de transacciones del mes siguiente

b) Estimar la probabilidad de que un usuario compre (haga una transacción)

c) Agrupar a los usuarios de acuerdo a sus características

Para la opción elegida, crear un modelo que sea adecuado para tratar con ese tipo de problema. Generar algunas variaciones del modelo y comparar sus resultados. Se tendrán en cuenta: la elección del modelo, sus parámetros, hiper parámetros, las métricas utilizadas para evaluar la perfomance y los mecanismos de validación de los resultados obtenidos.

3) **REPORTE**

Generar un informe de no más de 3 párrafos, en un lenguaje no-técnico que permita explicar a una audiencia más general cómo son los datos, en qué consiste el modelo utilizado, cuáles son los resultados más relevantes y cómo podría continuar el proyecto.

Entrega

El ejercicio podrás realizarlo en R o Python (según tu preferencia), utilizando las librerías que quieras y con un máximo de 50 líneas de código (sin contar la importación de librerías). Antes de la fecha límite (**7 días** desde el envío de este correo) deberás completar el siguiente Google Form

GOOGLE FORM

con el notebook que utilizaste y el reporte de 3 párrafos. Evaluaremos tanto el código como el reporte, así como también la presentación que puedas hacer del mismo.