procesalab

PRUEBA TÉCNICA CIENTÍFICO DE DATOS

Introducción:

La siguiente prueba tiene como finalidad medir sus capacidades en diferentes aspectos fundamentales cómo científico de datos que va desempeñar si es seleccionado. Los aspectos que se van a medir son los siguientes:

Capacidad Analítica: Capacidad de transformar datos con el fin de poder responder preguntas de negocio.
Capacidad de Programación y Modelación: Capacidad de representar mediante algoritmos soluciones frente a situaciones o requerimientos analíticos y poder transformarlos en lenguaje de programación python .
Entendimiento de negocio: Capacidad de comprender los diferentes aspectos y conceptos del negocio para poder responder y comprender los datos y requerimientos manejados

Le solicitamos que responda estas preguntas a conciencia y como a usted considere que puede ser la manera de llegar a una solución efectiva y oportuna, dado que eventualmente se le pedirá que justifique sus respuestas de manera argumentativa y técnica con sus respectivos entregables.

La prueba 5 retos de los cuales se deberán elegir y responder minimo tres, cada uno tiene un participación porcentual distinta la cual nos reservamos mencionar por lo que **existen dos retos que sí contesta adicionalmente servirán de bonus para la nota final**. Los restos se basan en responder preguntas abierta dando el entendimiento del problema, opción múltiple, elaboración de querys y algoritmos de tratamiento y modelación de datos

procesalab

Retos:

1. Entendimiento de negocio y procesamiento de información
2. Entendimiento técnico Marketing Mix Modeling
3. Ejercicio de Clustering
4. Ejercicio Predictivo o Forecasting
5. Test SQL

Habilidades a evaluar:

- ★ Diseño, entendimiento e implementación de Modelación
- ★ Programación en Python con buena metodología y prácticas (PEP8, POO), entendimiento de lenguaje R y construcción de querys en SQL
- ★ Conocimientos en Marketing mix modeling o marketing digital (atribución, conversión y activación)
- ★ Conocimientos en técnicas de evaluación y validación de modelos, métricas de desempeño, ingeniería de variables, optimización de hiperparamétricos de modelos.
- ★ Fundamentos en procesamiento de datos y arquitecturas en la nube

procesalab

1. Entendimiento de negocio y procesamiento de información

Una Empresa de Retail, que tiene aproximadamente 50.000 clientes registrados en sus bases de datos y unas ventas mensuales de aproximadamente $10 mil millones, la cual realiza ventas tanto de manera presencial como en su comercio electrónico, debe presentar a la junta directiva cada mes todos los indicadores que muestren cómo van funcionando las áreas de la compañía, por eso hasta hace poco el área de TI enviaba una sábana de datos que sacan directamente desde el ERP enviarlos a cada área para que saquen sus indicadores de gestión.

Actualmente dicha sábana de datos e indicadores se están quedando limitados para tomar decisiones óptimas dentro del área, dichas decisiones no están ajustadas a la realidad del negocio, ya que por las estrategias realizadas por mercadeo, los clientes se vienen incrementando al igual que las ventas, por eso las decisiones tomadas son tardías y no dan abasto con los pedidos hechos ya que el inventario se les agota muy rápido.

Tenga en cuenta las siguientes aclaraciones:

El ERP es una solución a la medida que por debajo tiene una base de datos DB2

La descarga de datos depende completamente de una persona de soporte, como el informe es para la junta directiva, deben hacerlo muy rápido, por lo que busca un modelo análitico que responda con efectividad y eficiencia

La compañía está interesada en contratar a una persona que tenga conocimientos analíticos como científico de datos que les pueda ayudar diseñando una solución de modelación predictiva que les permita ser costo eficiente por medio de tomar decisiones con prontitud y proyección.

Cada área es la encargada de moler los datos y sacar sus propias estadísticas e indicadores, por lo que les ha tocado buscar dentro de su personal a empleados que tengan conocimientos en procesamiento de datos y gestión de indicadores.

Según la situación presentada, por favor responda las siguientes preguntas.

- ¿Cómo explicaría la problemática actual de la compañía?
- ¿Qué solución propone para el problema que enfrenta actualmente la compañía que permita tener un entorno analítico escalable y administrable? (Debe detallar los siguientes elementos: Procesamiento de la información, ¿Qué técnicas de modelación utilizaría para la construcción de dicha solución?, automatización en la nube, ¿Qué tecnologías utilizará: lenguajes, servicios de IA en la nube, etc ? y ¿Cómo lo visualizará?)
- ¿Qué metodología utilizaría para la construcción y validación de la solución?
- Represente de manera gráfica cómo sería la arquitectura y modelación de la solución

procesalab

2. Entendimiento técnico Marketing Mix Modeling

Los equipos de pauta siempre necesitan herramientas o modelos analíticos que mejoren sus toma de decisiones de inversión. Actualmente existen diferentes medios tanto digitales (Facebook Ads, Google Ads) cómo tradicionales (TV, radio, prensa, etc) para invertir en pauta y alcanzar los objetivos establecidos en el área de mercadeo por lo que se hace necesario un modelo análitico que pueda recogerlos todos optimizar los presupuesto gastado todos los medios publicitarios. En este contexto para responder a esta necesidad nacen los MMM (Marketing Mix Modeling). Uno de lo más utilizados es Robyn el cual tiene la siguientes características:

Definición:

Es un modelo automatizado de Marketing Mix que funciona por medio de técnicas econométricas y Machine Learning. Se utiliza con lenguaje de programación R descargando su librería respectiva

Objetivos:

- Toma de decisiones procesables proporcionando un asignador de presupuesto y curvas de rendimientos
- Reducción del sesgo humano a través de un proceso de optimización automatizado

Características:

- Respetuoso con la privacidad no requiere datos de nivel de registro individual
- No depende de las cookies o los datos de píxeles
- Modelos basados en metodologías de medición directa y optimización multiobjetivo

procesalab

Revisando la documentación oficial sobe la tecnica de modelación (Marketing Mix Modeling (https://facebookexperimental.github.io/Robyn/docs/analysts-guide-to-MMM#modeling-techniques-)del modelo de marketing mix modeling desarrollado por el equipo de científico de datos de Facebook- Meta responda los siguientes preguntas:

- ¿Qué tipo de técnicas se aplican en la modelación?
- ¿Qué entendimiento le genera las ecuaciones del modelo?
- ¿Cómo debería ser el proceso de calibración de este Modelo?
- ¿Cómo mejorar y personalizar el modelo ?

procesalab

3. Ejercicio de Clustering

El marketing es fundamental para el crecimiento y la sostenibilidad de cualquier negocio. Los especialistas en marketing pueden ayudar a desarrollar la marca de la empresa, atraer clientes, aumentar los ingresos y aumentar las ventas.

Uno de los puntos críticos para los especialistas en marketing es conocer a sus clientes e identificar sus necesidades. Al comprender al cliente, los especialistas en marketing pueden lanzar una campaña de marketing dirigida que se adapte a necesidades específicas. Si los datos sobre los clientes están disponibles, la ciencia de datos se puede aplicar para realizar la segmentación del mercado.

En este reto, nos han contratado como expertos en data science para una empresa de minorista. La empresa tiene muchos datos sobre el comportamiento de sus clientes en su e-commerce. Se nos encomienda la tarea de crear campañas de marketing enfocadas a los clientes, dividiéndolos para ello en por lo menos 3 segmentos diferentes encontrando patrones en los datos de 150 días del comportamiento de los clientes en el sitio web.

Los datos se pueden encontrar en el siguiente link: https://drive.google.com/file/d/11_kFS0-czY989uFFwhoZ4yZQRntqTiFA/view?usp=sharing

La definición de las variables son las siguientes:

- fullVisitorID: Identificador del cliente
- channelgrouping: Medio por el cual llegó al sitio web
- date: dia
- OS: Tipo de dispositivo con el que entro al sitio web
- Apparel, Office, Electronics, LimitedSupply, Accessories, ShopByBrand, Bags: Categorias de los productos comprados
- totalSpent_USD: Total gastado

La solución debe tener lo siguiente:

- Tres clusters o segmentos con explicación de sus características
- Técnica de selección de número de clusters (En código)
- Técnica de clusterización (En código)
- Posiblemente técnica de reducción de dimensionalidad (En código)
- Determinación de hiperparametros (En código)
- Lenguaje: **Python**

Entregables:

- Link de notebook en colab (https://colab.research.google.com/) o archivo jupyter notebook con salidas de ejecución de código preferiblemente para no tener que volver ejecutar por que se puede generar error
- Informe con tablas y gráficos de resultados de clusters con su respectivo análisis (Corto y concreto, solo una pagina)
- PDF con resultados de ejecución del notebook por si existe algún problema con el notebook enviado

procesalab

4. Ejercicio Predictivo o Forecasting

El marketing digital actual nos provee data innumerable por lo que para obtener mejores resultados se necesitan hacer predicciones con el uso de técnicas de modelación predictiva y de pronóstico.

En la publicidad tipo SEM se determinan inversiones según unas palabras clave para aparecer en las primera posiciones de la búsquedas, en donde el hecho de que una de sus palabras clave sea rentable en torno a que las personas den click en la publicidad no significa que será así en el futuro, como también palabras que hoy no parecen rentables pueden serlo en futuro.

Por lo anterior debido a la naturaleza dinámica de las palabras clave se deben agregar pronósticos regulares a la toma de decisiones sobre la inversión en anuncios o publicidad digital en esas palabras.

Según la base de datos del siguiente link: https://drive.google.com/file/d/1DpbbHQIC-zpHQ4IbKehEKdBw3QunpND_/view?usp=sharing, elabore por medio de una técnica predictiva o de pronóstico un modelo que responda las siguientes preguntas:

- ¿Qué palabras serían más rentables en el futuro?
- ¿Qué palabras son más rentables por temporalidad en el futuro?. Defina la temporalidad como prefiera.
- ¿En cuáles palabras se debería dejar de invertir a futuro?

Nota:
Cost: inversion en la campaña de anuncios (CampaingID)
CPC: Costo por click

La solución debe tener lo siguiente:

- Técnica de modelación (En código)
- Técnica de evaluación del modelo (En código)
- Determinación de hiperparametros si es necesario (En código)
- Lenguaje: **Python**

Entregables:

- Link de notebook en colab (https://colab.research.google.com/) o archivo jupyter notebook con salidas de ejecución de código preferiblemente para no tener que volver ejecutar por que se puede generar error
- Informe con tablas y gráficos de resultados de clusters con su respectivo análisis (Corto y concreto, solo una pagina)
- PDF con resultados de ejecución del notebook por si existe algún problema con el notebook enviado

procesalab

5. Test SQL

Select the code which shows all abi_email which starts with 'john' or 'phillip'.

- [x] SELECT abi_email FROM usa_web_form WHERE abi_email LIKE 'john%' OR abi_email LIKE 'phillip%'
- [ ] SELECT abi_email FROM usa_web_form WHERE abi_email LIKE 'john%' OR 'phillip%'
- [ ] SELECT abi_email FROM usa_web_form WHERE abi_email LIKE 'john%' AND abi_email LIKE 'phillip%'
- [ ] SELECT abi_email FROM usa_web_form WHERE abi_email LIKE 'john%' AND 'phillip%'
- [ ] SELECT abi_email FROM usa_web_form WHERE abi_email LIKE '%john' OR abi_email LIKE '%phillip

Select the code that shows, in a single row, how many unique abi_email were obtained in td_host www.budweiser.com with abi_age between 21 and 34

- [x] SELECT COUNT(DISTINCT abi_email) FROM usa_web_form WHERE td_host = 'www.budweiser.com' AND abi_age BETWEEN 21 AND 34
- [ ] SELECT COUNT(DISTINCT abi_email) FROM usa_web_form WHERE td_host = 'www.budweiser.com' AND BETWEEN (21, 34)
- [ ] SELECT COUNT(abi_email) FROM usa_web_form WHERE td_host = 'www.budweiser.com' AND abi_age BETWEEN 21 AND 34
- [ ] SELECT DISTINCT abi_email FROM usa_web_form WHERE td_host = 'www.budweiser.com' AND abi_age BETWEEN (21, 34)
- [ ] SELECT abi_email FROM usa_web_form WHERE td_host = 'www.budweiser.com' AND abi_age BETWEEN (21, 34)

Given the table usa_web_form and respective data content, what does the following SQL code retrieves?

```
 SELECT *
FROM usa_web_form
WHERE
abi_first_name = 'John' AND
 abi_last_name = 'Doe' AND
```

procesalab

abi_last_name = 'Foo

id	abi_first_name	abi_last_name	abi_email
1	Maria	Doe	maria.doe@gmail.com
2	John	Foo	jfoo@gmail.com
3	John	Doe	john.doe@gmail.com
4	Maria	Foo	maria.foo@gmail.com

- [X] Will not retrieve anything because some WHERE statements are conflicting
- [ ] Will retrieve 2 records. IDs 2, 3
- [ ] Will retrieve 1 record. ID 3
- [ ] Will retrieve all 4 records. IDs 1, 2, 3, 4

Considering the tables below, please select the correct SQL statement that will retrieve the amount of unique visited_pages from each consumer identified by abi_email that have an entry in usa_web_form, by matching via cookie_id.

usa_web_form					
cookie_id	timestamp	abi_email	abi_first_name	abi_last_name	abi_brand
12543	2021-05-02 23:22:21	jdoe@gmail.com	John	Doe	Budweiser
39823	2021-05-14 13:11:20	jane@gmail.com	Jane	Doe	Corona
59485	2021-05-16 01:54:11	maria@gmail.com	Maria	Dane	Corona
98243	2021-05-16 20:00:14	jdoe@gmail.com	John	Doe	Michelob
...					

usa_page_views			
cookie_id	timestamp	visited_page	abi_brand
12543	2021-05-02 23:22:21	www.budweiser.com	Budweiser
59485	2021-05-16 01:54:11	www.corona.com	Corona
98243	2021-05-16 20:00:14	www.michelobultra.com	Michelob
12543	2021-05-20 17:32:12	www.becks.com	Becks
...			

- [X] SELECT COUNT(DISTINCT v.visited_page), f.abi_email FROM usa_web_form f JOIN usa_page_views v ON f.cookie_id = v.cookie_id GROUP BY f.abi_email
- [ ] SELECT COUNT(DISTINCT v.visited_page), f.abi_email FROM usa_web_form f JOIN usa_page_views v ON f.cookie_id = v.cookie_id
- [ ] SELECT COUNT(v.visited_page), f.abi_email FROM usa_web_form f JOIN usa_page_views v ON f.cookie_id = v.cookie_id GROUP BY f.abi_email

procesalab

- ☐ SELECT DISTINCT v.visited_page, f.abi_email FROM usa_web_form f JOIN usa_page_views v ON f.visited_page = v.visited_page
- ☐ SELECT AMOUNT(DISTINCT v.visited_page), f.abi_email FROM usa_web_form f, usa_page_views v WHERE f.abi_email = v.abi_email

Considering the table below, what SQL statement should be used to standardize the field abi_gender to 'M' for male, 'F' for female and 'O' for other?

usa_web_form			
cookie_id	**timestamp**	**abi_email**	**abi_gender**
340714	2021-11-20 16:25:15	jdoe@gmail.com	male
302190	2020-11-12 15:24:39	jane@gmail.com	F
887607	2021-12-14 18:25:14	pfoo@gmail.com	MEN
696492	2021-11-24 20:14:44	abc@gmail.com	female
994959	2021-10-14 15:13:44	e-mail@aol.com	other
602243	2021-12-24 14:16:55	maria@netscape.com	girl
796312	2020-10-11 22:25:39	phill@outlook.com	male

- ☒ SELECT cookie_id, timestamp, abi_email CASE WHEN lower(abi_gender) IN ('male', 'men') THEN 'M' WHEN lower(abi_gender) IN ('f', 'female', 'girl') THEN 'F' WHEN lower(abi_gender) IN ('other') THEN 'O' END AS abi_gender FROM usa_web_form
- ☐ SELECT cookie_id, timestamp, abi_email, NORMALIZE(abi_gender, 'M', 'F', 'O') FROM usa_web_form HAVING abi_gender in ('male', 'MEN', 'female', 'girl', 'other')
- ☐ SELECT cookie_id, timestamp, abi_email, NORMALIZE(abi_gender AS 'M' WHEN 'male' OR 'men', 'F' WHEN 'girl' OR 'female', 'O' WHEN 'other') FROM usa_web_form
- ☐ SELECT cookie_id, timestamp, abi_email, CASE abi_gender IN ('male', 'MEN', 'female', 'girl', 'other') THEN 'M', 'M', 'F', 'F', 'O' FROM usa_web_form

This test have been built around a dummy scenario around consumer consents and have the description as it follows:

BUSINESS RULES

procesalab

Before using any consumer data, we need their consent for Terms & Conditions (TC-PP) and Marketing Activation (MARKETING-ACTIVATION) for each brand. For each consent accepted, our table will receive one row, with an id for the consumer, a brand to which this consent was collected and the consent itself. Considering this, we want you to work with the consent table data (abi_consents), according to these BUSINESS RULES:

1. The consumer consent will be classified as Opted-In if she/he accepted both Terms & Conditions and Marketing Activation.
2. The consumer consent will be classified as Not Given if she/he accepted only Terms & Conditions
3. For any other case, the consumer consent will be classified as Unknown.

Example:

- If the table contains a row with TC-PP and another row with MARKETING-ACTIVATION for a specific consumer and specific brand, the consumer consent have to be classified as "Opted- In" for that brand.
- If the table contains only a row with TC-PP for a specific consumer and brand, the consent have to be classified as "Not Given".

Go to https://sqliteonline.com/ and use the DDL below to create the environment for your exercise:

```
CREATE TABLE abi_consents (
client_id VARCHAR(4),
brand_name VARCHAR(15),
consent VARCHAR(20)
);


INSERT INTO abi_consents VALUES ('AF32', 'BRAHMA', 'TC-PP');
INSERT INTO abi_consents VALUES ('AF32', 'BRAHMA', 'MARKETING-ACTIVATION');
INSERT INTO abi_consents VALUES ('YD71', 'BRAHMA', 'TC-PP');
INSERT INTO abi_consents VALUES ('ODA2', 'BRAHMA', null);
```

INSERT INTO abi_consents VALUES ('LA94', 'BRAHMA', 'MARKETING-ACTIVATION');
INSERT INTO abi_consents VALUES ('JA13', 'BRAHMA', 'MARKETING-ANALYTICS');
INSERT INTO abi_consents VALUES ('JA13', 'BRAHMA', 'TC-PP');
INSERT INTO abi_consents VALUES ('YD71', 'SKOL', 'TC-PP');
INSERT INTO abi_consents VALUES ('YD71', 'SKOL', 'MARKETING-ACTIVATION');
INSERT INTO abi_consents VALUES ('KD81', 'SKOL', 'TC-PP');
INSERT INTO abi_consents VALUES ('KD81', 'SKOL', 'MARKETING-ACTIVATION');
INSERT INTO abi_consents VALUES ('OSW1', 'BRAHMA', 'TC-PP');
INSERT INTO abi_consents VALUES ('KD81', 'SKOL', null);

Build a SQL statement that retrieves the unique amount of consumers (client_id) that provided the TC-PP consent for each brand.

Please write your query in the open field below.

```
SELECT brand_name, COUNT(DISTINCT client_id) FROM abi_consents
WHERE consent = 'TC-PP'
GROUP BY brand_name;
```