Data Science Challenge

Data & Analytics Team

El desafío consta de 4 ejercicios independientes que van desde análisis exploratorio, machine learning o el diseño de una solución de data science.

¿Qué evaluamos?

El desafío busca evaluar distintos aspectos como:

- Capacidad analitica y exploración de datos
- Visualización de resultados
- Conocimientos de técnicas de generación de features y modelado ● Análisis de performance
- Buenas prácticas de desarrollo
- Diseño e implementación de Machine learning en producción
- Diseño e implementación de flujos end to end usando LLMs
- Habilidades de prompt engineering

Algunas reglas y recomendaciones:

1. La mayoría de los ejercicios se piden resolver en Jupyter notebooks y te recomendamos subirlas a un repositorio de GitHub público para compartir los resultados.
2. No dejes de hacernos preguntas sobre cualquier duda con los enunciados.

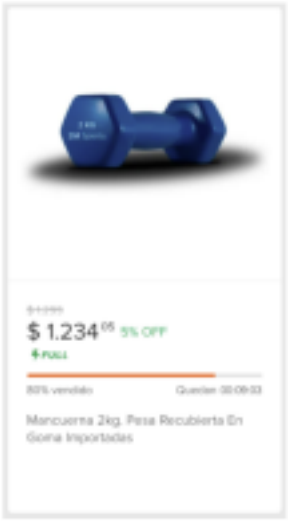
El desafío se analiza de acuerdo al seniority del postulante y teniendo en cuenta también las necesidades particulares de la posición.

1. Explorar las ofertas relámpago, ¿qué insights puedes generar?

Descripción

En conjunto con el desafío te compartimos un archivo llamado ofertas_relampago.csv el cual posee información de los resultados de ofertas del tipo relampago para un periodo de tiempo y un país determinado.

Estas ofertas en mercadolibre se pueden ver de la siguiente manera:



Es decir, son ofertas que tienen una duración definida de algunas horas y un porcentaje de unidades (stock) comprometidas.

El objetivo de este desafío es hacer un EDA sobre estos datos buscando insights sobre este tipo de ofertas.

Las columnas del dataset son autoexplicativas pero puedes preguntarnos cualquier duda.

Entregable

El entregable de este desafío es una Jupyter notebook con el EDA.

2. Similitud entre productos

Descripción

Un desafío constante en MELI es el de poder agrupar productos similares utilizando algunos atributos de estos como pueden ser el título, la descripción o su imagen.

Para este desafío tenemos un dataset “items_titles.csv” que tiene títulos de 30 mil productos de 3 categorías diferentes de Mercado Libre Brasil

ITE_ITEM_TITLE
Tênis Yate Masculino Estilo Pleno Fr6500 Cinza
Tenis De Academia Super Confortavel Pisada Neu...
Tênis Olympikus Voa Feminino Rosa
Tenis Masculino Under Armour Amortecimento Con...
Tênis 36 Branco Star Universe Original Novo Al...
Cinta Trava Microtex Presilha Sapatilha Fizik ...
Bicicleta Absolute Aro 29 12v Preto/azul - Tam...

Entregable

El objetivo del desafío es poder generar una Jupyter notebook que determine cuán similares son dos títulos del dataset “item_titles_test.csv” generando como output un listado de la forma

ITE_ITEM_TITLE	ITE_ITEM_TITLE	Score Similitud (0,1)
Zapatillas Nike	Zapatillas Adidas	0.5
Zapatillas Nike	Zapatillas Nike	1

donde ordenando por score de similitud podamos encontrar los pares de productos más similares en nuestro dataset de test.

3. Previsión de falla

Descripción

Los galpones de Full de mercado libre cuentan con una flota de dispositivos que transmiten diariamente telemetría agregada en varios atributos.

Las técnicas de mantenimiento predictivo están diseñadas para ayudar a determinar la condición del equipo de mantenimiento en servicio para predecir cuándo se debe realizar el mantenimiento. Este enfoque promete ahorros de costos sobre el mantenimiento preventivo de rutina o basado en el tiempo porque las tareas se realizan solo cuando están justificadas.

Entregable

Tiene la tarea de generar una Jupyter notebook con un modelo predictivo para predecir la probabilidad de falla del dispositivo con el objetivo de bajar los costos del proceso. Como una referencia, una falla de un dispositivo tiene un costo de 1 mientras el costo de un mantenimiento es 0,5. El archivo "full_devices.csv" tiene los valores diários para los 9 atributos de los dispositivos y la columna que está tratando de predecir se llama 'failure' con el valor binario 0 para no fallar y 1 para fallar.

4. Diseño de un asistente Q&A para publicaciones de Mercado Libre

Descripción

Semanalmente se reciben aproximadamente 4.3 millones de preguntas en la plataforma de Mercado Libre. Se quiere diseñar un proceso que automatice la generación de respuestas haciendo uso de modelos de lenguaje.

Para esta prueba se proporciona el dataset "*preguntas_mercadolibre.xlsx*", el cual contiene información en formato texto sobre distintas publicaciones de Mercado Libre. El archivo incluye los siguientes campos:

- SITE: país de origen de la publicación
- PREGUNTA: pregunta realizada por el comprador
- PUBLICACIÓN: información de la publicación en formato texto
-

El objetivo es diseñar un flujo con un LLM que permita responder preguntas de clientes sobre productos, utilizando únicamente la información disponible en la publicación.

Actividades a realizar

1. Diseño de prompts: crear uno o varios prompts adecuados para la tarea asignada.

2. Selección del modelo: elegir el LLM a utilizar y justificar la decisión en términos de costo y desempeño.

3. Manejo de limitaciones: explicar cómo la solución propuesta previene o gestiona problemas como alucinaciones, información contradictoria, ausencia de datos, u otros inconvenientes identificados.

4. Evaluación: diseñar una métrica de desempeño para medir la efectividad del flujo de respuestas.

NOTA: puede usar apis de llms de consumo gratuito como gemini.

Entregables

- Jupyter Notebook con la implementación completa del flujo, incluyendo los prompts finales y las estrategias aplicadas para la mitigación de inconvenientes.

- Archivo CSV con los resultados de prueba en el formato: *site, pregunta, publicación, respuesta generada*.

- Reporte de métrica de desempeño obtenida.