El desafío consta de 5 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

Algunas reglas y recomendaciones:

1. Si bien son 5 ejercicios, los 3 y 5 son obligatorios, y luego uno a elección.
2. Un número recomendado a resolver son 3 de los 5 desafíos, pero sentite libre de resolver la cantidad deseada.
3. 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. Deben ser fácilmente reproducibles en caso que se requiera.
4. No dejes de hacernos preguntas sobre cualquier duda con los enunciados

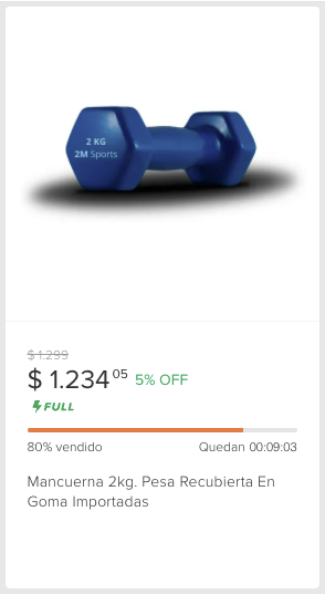
El desafío se analiza de acuerdo al seniority del postulante y teniendo en cuenta también las necesidades particulares de la posición.

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. ¿Cuán rápido puedes ordenar estos productos?

Descripción

La seccion de ofertas https://www.mercadolibre.com.ar/ofertas#nav-header es una sección que agrupa las mejores ofertas de MELI, y a grandes rasgos es un listado de productos en oferta ordenados por un score de ML y distintas reglas de negocio.

Anexado al desafío, se encuentra un archivo “ordenamiento.csv” el cual tiene un listado de productos, con su score y categorías para este desafío.

Un sample de este csv se ve de la siguiente manera:

item_id	vertical	category	domain	score
632383659	HOME & INDUSTRY	HOME&DECOR	MLC-STOOLS	0.1186
479170683	CE	ELECTRONICS	MLC-WASHING_MACHINES	0.9216
492059986	HOME & INDUSTRY	HOME&DECOR	MLC-BATHROOM_SUPPLIES	0.5662
566362328	ACC	AUTOPARTS	MLC-VEHICLE_PARTS	0.1652
601750109	CE	COMPUTERS	MLC-EXTERNAL_LAPTOP_COOLERS	0.7694
579746719	OTHERS	PET PRODUCTS	MLC-FLEA_AND_TICK_TREATMENTS	0.7020
591927686	APP & SPORTS	SPORTS	MLC-BICYCLE_CHILD_SEATS	0.5956

El score determina cuán bueno es un item_id, siendo 1 el mejor valor posible.

De no existir reglas de negocio que ponen restricciones sobre el ordenamiento, la solución ideal sería ordenar nuestro dataset de mayor a menor por score, obteniendo un resultado de la siguiente forma:

ITE_ITEM_ID	VERTICAL	DOM_DOMAIN_AGG1	DOMAIN_ID	SCORE
590602034	CE	ELECTRONICS	MLC-GAME_CONSOLES	0.9998
609438042	CE	ELECTRONICS	MLC-GAME_CONSOLES	0.9996
634352041	CE	MOBILE	MLC-CELLPHONES	0.9994
615879515	CE	MOBILE	MLC-CELLPHONES	0.9992
631654974	CE	MOBILE	MLC-CELLPHONES	0.9990

El problema es que este ordenamiento debe reflejar variedad de productos y categorías permitiendo el discovery de distintos tipos de productos por lo cual tenemos en producción reglas como las siguientes:

1. El “domain_id” no se puede repetir en 4 posiciones consecutivas.
2. El “vertical” no se puede repetir en 1 posición consecutiva.
3. De existir el id 641416750 en el listado debe estar en la posición 3 siendo esta regla más fuerte que las demás.
4. De existir el id 22351223 en el listado debe estar en la posición 6 siendo esta regla más fuerte que las demás.
5. Las posiciones 9,10,11 deben tener sí o sí items de la categoría “HOME&DECO” siendo esta regla más fuerte que la 1 y 2.
6. Cumpliendo estas condiciones, el ordenamiento debe respetar un ordenamiento de mayor score a menor.

El desafío es diseñar un algoritmo que, dado el dataset y estas restricciones, devuelva el listado final ordenado de ítems. El algoritmo debe estar diseñado para escalar eficientemente con el número de ítems, y contemplar los casos en que no se pueden cumplir las restricciones. **¡El tiempo de ejecución es el factor clave!**

Entregable

Notebook con el algoritmo para generar el listado ordenado y su tiempo de ejecución.

3. 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.

4. Series de tiempo

Descripción

Pronosticar las ventas de un producto o de una categoría es un desafío recurrente para cualquier ecommerce.

En este caso, el desafío va a ser pronosticar la cantidad de unidades diarias que van a vender 3 categorías distintas de MELI.

El dataset “series.csv” tiene las ventas en unidades diarias de 3 categorías que poseen un id único y su fecha de venta.

CATEGORY	DATE	UNITS_SOLD
CATEG-3	2020-11-16	3666
CATEG-1	2021-04-30	55
CATEG-3	2019-07-27	393
CATEG-3	2020-05-28	2349
CATEG-2	2020-01-03	6
CATEG-1	2019-11-11	32
CATEG-3	2019-10-28	634
CATEG-3	2020-01-20	734
CATEG-1	2020-11-13	152
CATEG-1	2020-03-05	33

Entregable

El objetivo de este desafío es construir un modelo de forecast que permita estimar las ventas de 3 semanas a nivel diario utilizando la historia de ventas de la categoría. Es decir, predecir las ventas de los siguientes 21 días. Las métricas y la medición de la performance del forecast son un punto clave de este desafío.

TIP: Dividir el dataset en entrenamiento, testing y validación correctamente es muy importante en problemas de forecasting!

5. Data Pipeline

Descripción

En machine learning es importante poder desarrollar procesos que forman parte del punta a punta de las tareas que implica.

En este caso, debes desarrollar un código claro y con el diseño apropiado para generar un dataset de ítems a partir de la API de Mercado Libre.

El endpoint a utilizar será el de search:

https://api.mercadolibre.com/sites/MLA/search?q=tv%2

04k

La tarea consiste en desarrollar un código en un archivo llamado build_dataset.py que al ejecutarse de la siguiente manera:

python build_dataset.py MLA

Se ejecute y devuelva como resultado en el mismo directorio un archivo llamado dataset.csv con los siguientes campos:

ITEM_ID, TITLE, PRICE, DOMAIN_ID, BRAND

Los ítems del dataset resultante deben pertenecer a la condición de nuevos (condition=new). El site (MLA, MLB o MLM) debe ser un parámetro que tome el script como argumento posicional.

Hint: La marca se encuentra dentro de attributes con el nombre de ID=BRAND.

Entregable

El script build_dataset.py y el archivo dataset.csv (con +- 5 mil registros)