Prueba Técnica: Data Scientist

Objetivo

Evaluar tus conocimientos técnicos en Ciencia de Datos, Machine Learning y programación orientada a objetos en Python.

Instrucciones generales

- Entrega el código en un repositorio git público (GitHub). Compartir el link en un email para su revisión.
- Adjunta un archivo README explicando claramente tu solución, el porqué de tus decisiones y cómo ejecutar tu código.
- Aplica buenas prácticas: comentarios claros, código legible, modularidad y estructura lógica.

Descripción

El siguiente dataset tiene 100.000 registros de ítems extraídos del marketplace en MercadoLibre, caracterizados a través de 26 diferentes columnas. El dataset está adjunto a este documento

A continuación, una descripción de las columnas:

Variable	**Descripción**
id	ID de la publicación
title	Título de la publicación
date_created	Fecha de creación de la publicación
base_price	Precio del producto en la publicación, sin descuento
price	Precio del producto en la publicación, con descuento
category_id	ID de categoría del producto
tags	Tags de la publicación
attributes	Atributos del producto publicado)
variations	Variaciones del producto publicado
pictures	Fotos del producto publicado
seller_id	ID del vendedor
seller_country	País de residencia del vendedor

seller_province	Provincia de residencia del vendedor
seller_city	Ciudad de residencia del vendedor
seller_loyalty	Loyalty o segmento del vendedor
buying_mode	Modo de compra especificado
shipping_mode	Modo de envío especificado
shipping_admits_pickup	Flag indicando si se puede retirar al domicilio
shipping_is_free	Flag indicando si el envío es gratis
status	Estado de la publicación
sub_status	Sub-estado de la publicación
warranty	Garantía del producto
is_new	Flag indicando si el producto es nuevo
initial_quantity	Stock inicial del producto
sold_quantity	Stock vendido del producto
available_quantity	Stock disponible del producto

Tareas a realizar

1. Análisis exploratorio de datos (EDA)

Se debe chequear la calidad del dataset para hacer una evaluación de qué tan apropiados son los datos para tareas de Data Science. Proponga un conjunto de correcciones en los datos de ser necesario.

Crea una clase llamada DataAnalyzer que incluya métodos para:

- Leer el dataset proporcionado.
- Proporcionar un resumen estadístico de los precios y cantidades vendidas.
- Identificar y gestionar posibles valores faltantes o inconsistentes.
- Detectar outliers en el precio de los productos.

2. Feature Engineering

Indicar posibles candidatos de features que podrían describir adecuadamente los items, tanto desde las columnas originales como desde transformaciones.

Implementa una clase llamada FeatureEngineer que incluya métodos para:

- Crear variables nuevas a partir de los datos existentes.

- Seleccionar variables relevantes mediante técnicas estadísticas o algoritmos específicos.
- Explicar claramente la elección y construcción de estas nuevas variables.

3. Modelado predictivo

Describa las posibles tareas de Machine Learning que podrían realizarse desde el dataset dado, que podrían ser valiosas en el dominio dado

Implementa una clase llamada ModelPredictor que cumpla con:

- División adecuada del dataset (train-test).
- Entrenamiento de al menos dos modelos predictivos diferentes (ej. RandomForest, XGBoost).
- Uso de validación cruzada para evaluar rendimiento.
  Selección del mejor modelo basado en métricas relevantes (explica tu elección).
- Predicción del target final de productos dado ciertos atributos (elige atributos relevantes).

4. Análisis de resultados

- Evalúa los resultados obtenidos y describe qué métricas utilizaste para decidir cuál es el mejor modelo.
- Proporciona un análisis crítico sobre los resultados obtenidos y posibles limitaciones.

5. Insights para Marketing y Negocio

- Extrae insights claros y accionables dirigidos a un equipo no-técnico de Marketing y Negocio.
- Describe cómo podrían utilizarse estos insights para mejorar decisiones estratégicas en la empresa.
- Describir posibles casos de usos a tratar con este dataset que podrían agregar valor al negocio dado, indicando métodos / técnicas y algoritmos por cada uno de ellos, así como justificando las decisiones tomadas.

Opcionales

- **Estrategia de Monitoreo:** Describe de manera libre y creativa cómo podrías monitorear el desempeño del modelo o los datos en producción.
- **Implementación técnica del pipeline:** Sugiere una forma programática de guardar el dataset final y el modelo entrenado en una base de datos (por ejemplo, BigQuery) y un Model Store (por ejemplo, MLflow).

ENTREGA

Tu solución final debe contener:

- Documentación clara del proyecto (README). Adicionalmente especificando comandos para ejecutar una predicción/inferencia.
- Breve reporte de resultados, métricas obtenidas y conclusiones.

Criterios de evaluación

- Correcta aplicación de conceptos de Data Science y Machine Learning.
- Claridad y calidad del código (estructura, legibilidad, modularidad).
- Capacidad analítica y crítica reflejada en conclusiones y decisiones documentadas.
- Creatividad y viabilidad técnica de la propuesta de integración y monitoreo (si aplica).