Prueba Técnica - Modelador Junior

Contexto

Se te proporciona un archivo Excel con datos de precios de commodities y predicciones de múltiples modelos para crear un metamodelo que prediga la dirección (Subida o bajada) de los precios cuando la diferencia entre `date_requested` y `date_prediction` es de 4 semanas.
El archivo Excel (.xlsx) contiene dos hojas:

Hoja 1: "Real"

- **id_commodity**: ID único del producto
- **type**: Tipo de producto (ej: metales, agrícolas, energía)
- **incoterm**: Términos Internacionales de Comercio
- **origin**: Origen de los datos de precios
- **publication**: Medio de publicación de los precios
- **date**: Fecha de publicación
- **value**: Precio del producto

Hoja 2: "Predicted"

- **id_commodity**: ID único del producto
- **type**: Tipo de producto
- **incoterm**: Términos Internacionales de Comercio
- **origin**: Origen de los datos de precios
- **publication**: Medio de publicación
- **model**: Modelo que realizó la predicción
- **date_requested**: Fecha en que se solicitó la predicción
- **date_prediction**: Fecha objetivo de la predicción
- **prediction**: Precio predicho por el modelo

Objetivo Principal

Crear un modelo que prediga la **dirección del precio** (subida/bajada) utilizando las predicciones de múltiples modelos como features, específicamente para horizontes de **4 semanas**.

Entregables Esperados

1. Análisis de Datos

- **Inspección básica**: `.head()`, `.info()`, `.describe()` de ambas hojas
- **Identificación de problemas**: valores nulos, duplicados evidentes
- **Conteo por categorías**: commodities por tipo, modelos disponibles
- **1-2 visualizaciones clave**: distribución de precios, serie temporal simple

2. Preprocesamiento:

- **Filtrado de datos**: Solo registros con diferencia de 4 semanas
- **Creación de variable objetivo**:
  - `direccion = 1` si precio real > predicción
  - `direccion = 0` si precio real ≤ predicción
- **Merge de datos**: Unir precios históricos con predicciones
- **Limpieza básica**: Eliminar nulos en variables clave
- **Feature simple**: Error de predicción por modelo

3. Modelado:

- **Split de datos**: 70% train, 30% test (sin validación cruzada)
- **Features principales**:
  - Predicción de cada modelo
  - Error promedio histórico por modelo
  - Consenso entre modelos (promedio, desviación estándar)
  - Tipo de commodity (encoded)
- **Modelo único**: Logistic Regression o Random Forest
- **Evaluación**: Accuracy, Precision, Recall, F1-score, R2
- **Feature importance**: Top 5 variables más importantes

4. Interpretación y Conclusiones:

- **Resumen en 3-4 bullets**: ¿Qué modelos predicen mejor la dirección?
- **Limitaciones identificadas**: ¿Qué no se pudo hacer por tiempo?
- **Próximos pasos**: ¿Qué haría con más tiempo?