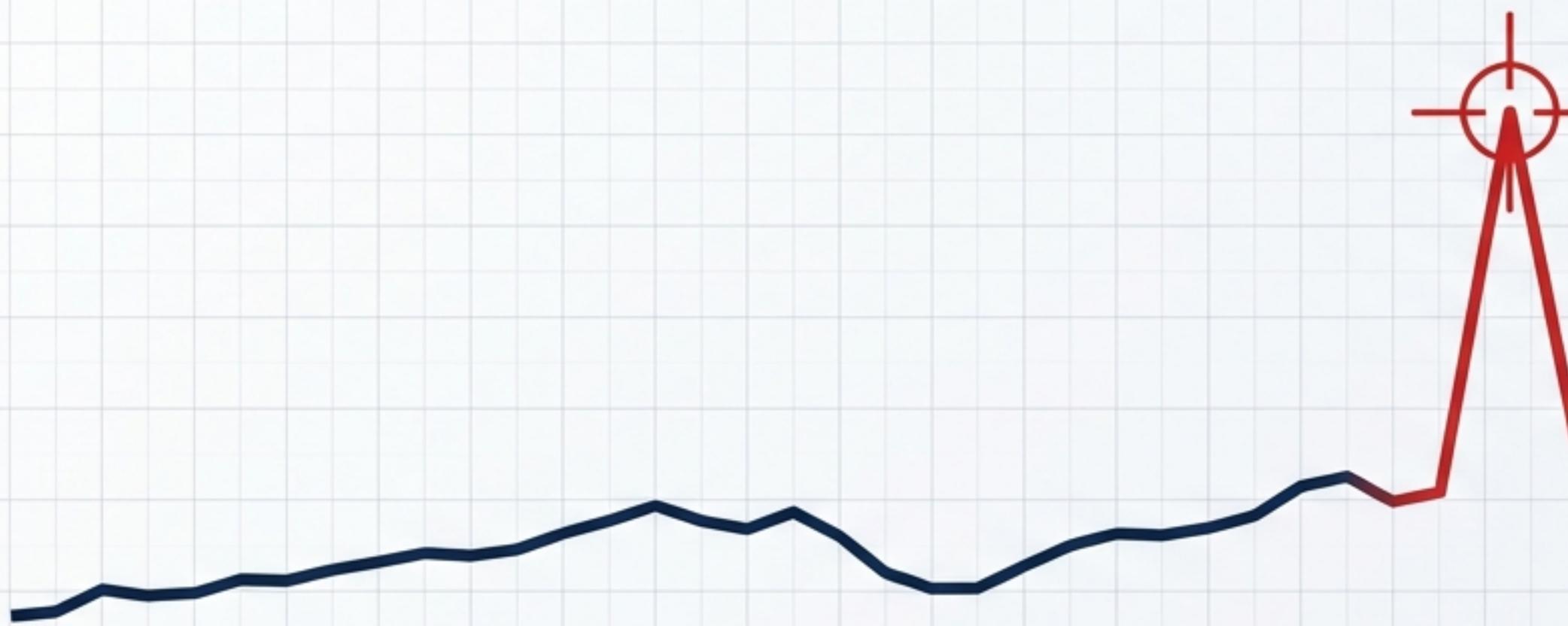


Detección de Anomalías en Series de Precios: ML Clásico vs. LLM

Solución Técnica al Desafío Mercado Libre - Ejercicio 1



El Desafío: Detección No Supervisada en Escala



Volumen
Masivo

Contexto del Problema

- **Input:** Dataset histórico con la evolución temporal de precios (PRICE) agrupados por producto (ITEM_ID).
- **Objetivo:** Identificar comportamientos inusuales **de forma automática sin contar con etiquetas reales** (Problemática *Unlabeled*).



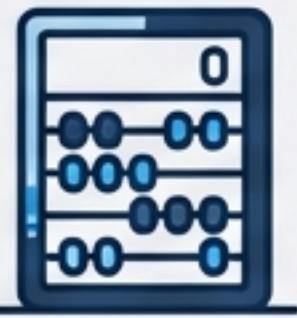
Interpretabilidad

La Tensión Técnica

- **Eficiencia:** Necesidad de procesar grandes volúmenes de datos con baja latencia.
- **Explicabilidad:** El negocio necesita entender *por qué* un precio es **anómalo** (razonamiento semántico vs. desviación numérica).



Estrategia de Modelado: Baseline vs. Challenger



Modelo B (El Baseline - ML Clásico)

- **Enfoque:** Estadística Robusta.
- **Ventaja:** Velocidad extrema, bajo costo computacional, estándar de la industria.
- **Método:** Mediana y MAD (Median Absolute Deviation).

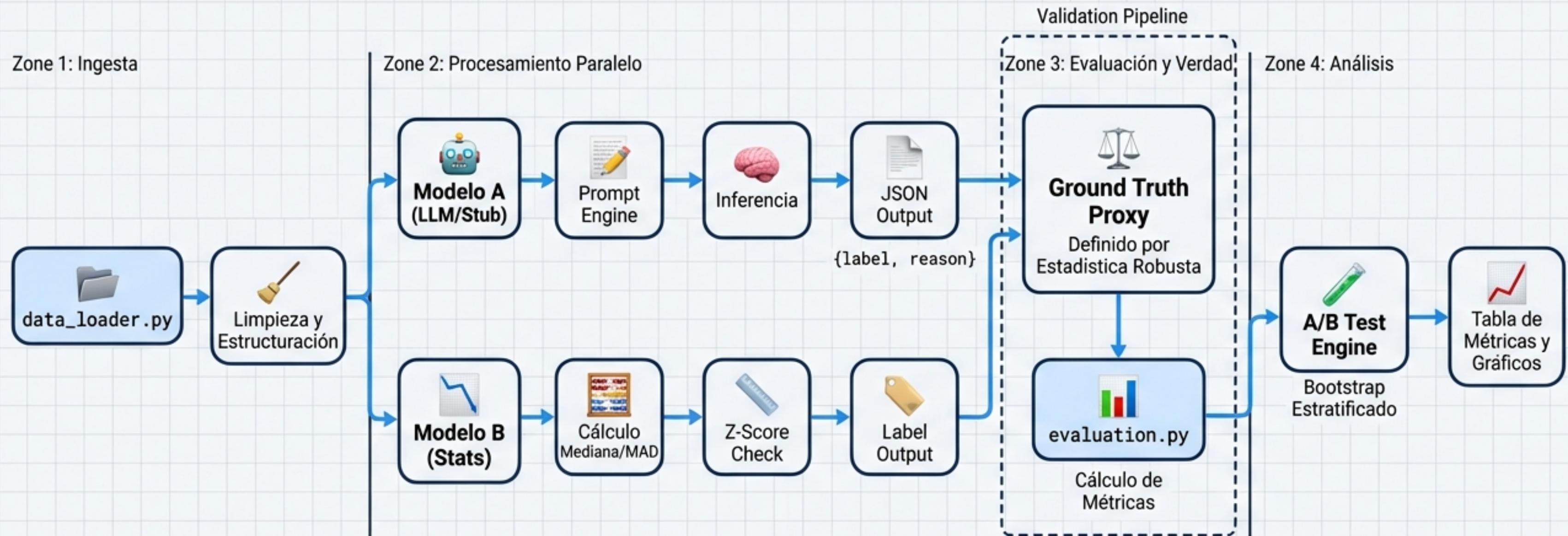


Modelo A (El Challenger - LLM)

- **Enfoque:** Razonamiento Semántico / Few-shot.
- **Ventaja:** Capacidad de generar un campo `reason` (explicación en lenguaje natural).
- **Método:** Prompt con salida estructurada JSON (`label`, `confidence`, `reason`).

Arquitectura de la Solución

Diseño Modular y Extensible



Profundización Modelo B: Estadística Robusta

La Solución: MAD (Median Absolute Deviation)

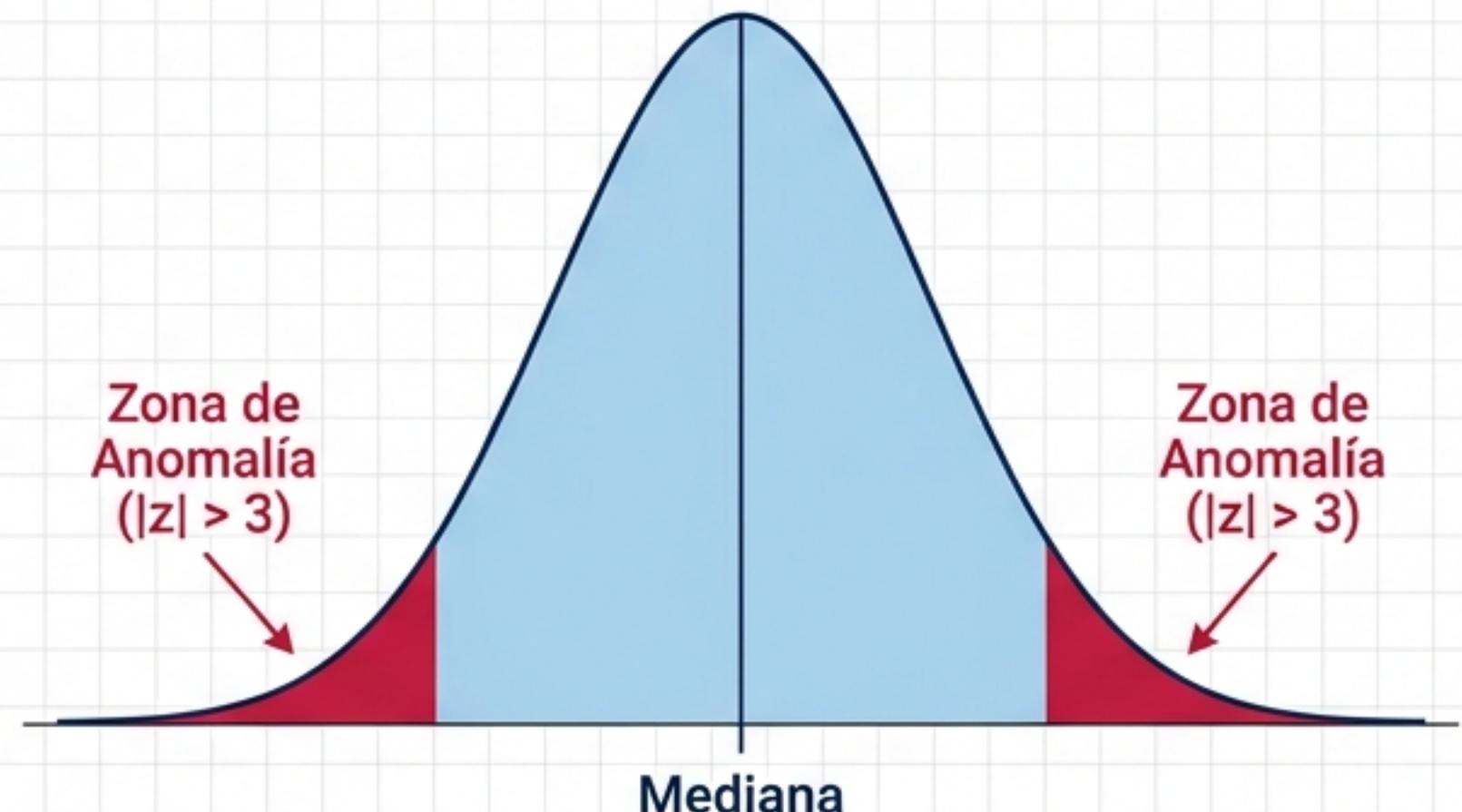
Se calcula la mediana de las desviaciones absolutas respecto a la mediana central. A diferencia de la desviación estándar, el MAD es resistente a outliers extremos.

Regla de Decisión

Si $|z_{\text{MAD}}| > 3 \rightarrow \text{ANÓMALO}$

Si $|z_{\text{MAD}}| \leq 3 \rightarrow \text{NORMAL}$

$$z_{\text{MAD}} = \frac{x_i - \text{median}(X)}{\text{MAD}}$$



Profundización Modelo A: LLM y Razonamiento

Diseño implementado como un **Stub Extensible**. Permite la integración futura con APIs (OpenAI / HuggingFace) manteniendo un contrato de datos estricto.

JSON Output Structure

```
{  
    "label": "ANOMALO",  
    "confidence": 0.95,  
    "reason": "Precio subió 200% sin  
cambio de tendencia previo."  
}
```

Clasificación binaria

Score [0-1]

Explicación semántica (< 15 palabras)



Valor Agregado:
Contexto.

“
El LLM no solo detecta el pico, explica la causalidad aparente.
”

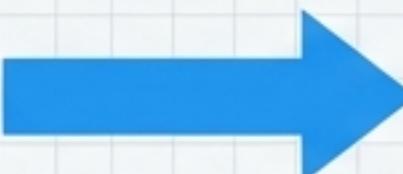
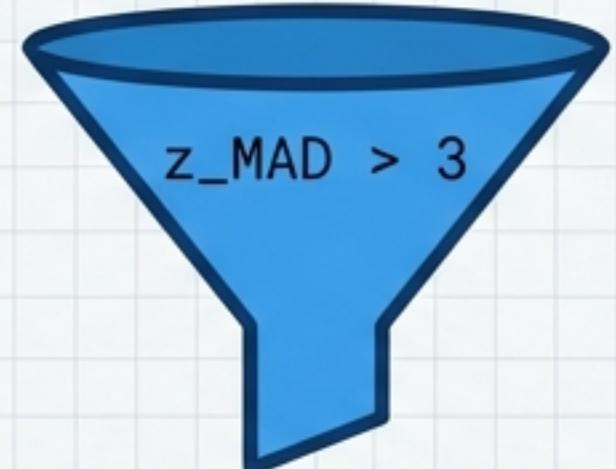
El Desafío de la Verdad (Ground Truth)



Dataset Sin Etiquetas
(Unlabeled)



Proxy Lógico



Ground Truth
Asumido

Problema:

Imposible calcular
Precision/Recall real sin
referencias.

Solución:

Se asume que la estadística
robusta define la 'realidad
física' de la **anomalía** para
efectos de benchmarking.

Justificación: Habilita la
ejecución de A/B Testing
numérico (`evaluation.py`) y
es práctica estándar en
entornos semi-supervisados.

Metodología de Validación: A/B Testing

Métricas

Métrica Primaria:

F1-Score (Balance Precision/Recall)

Test Estadístico:

Bootstrap Estratificado

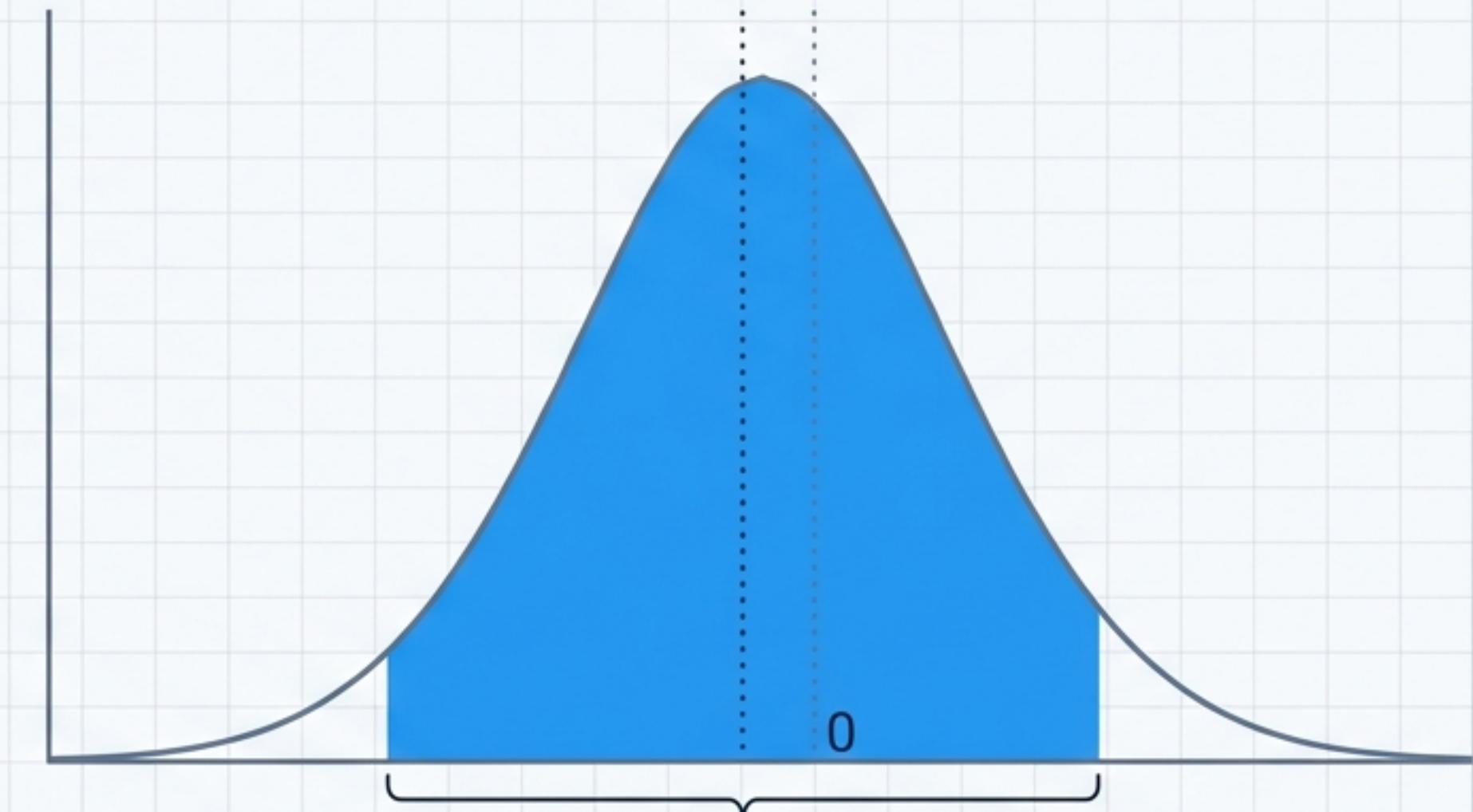
Iteraciones:

$\geq 1,000$ simulaciones

Objetivo:

Calcular $\Delta F1 = F1_{LLM} - F1_{Clásico}$

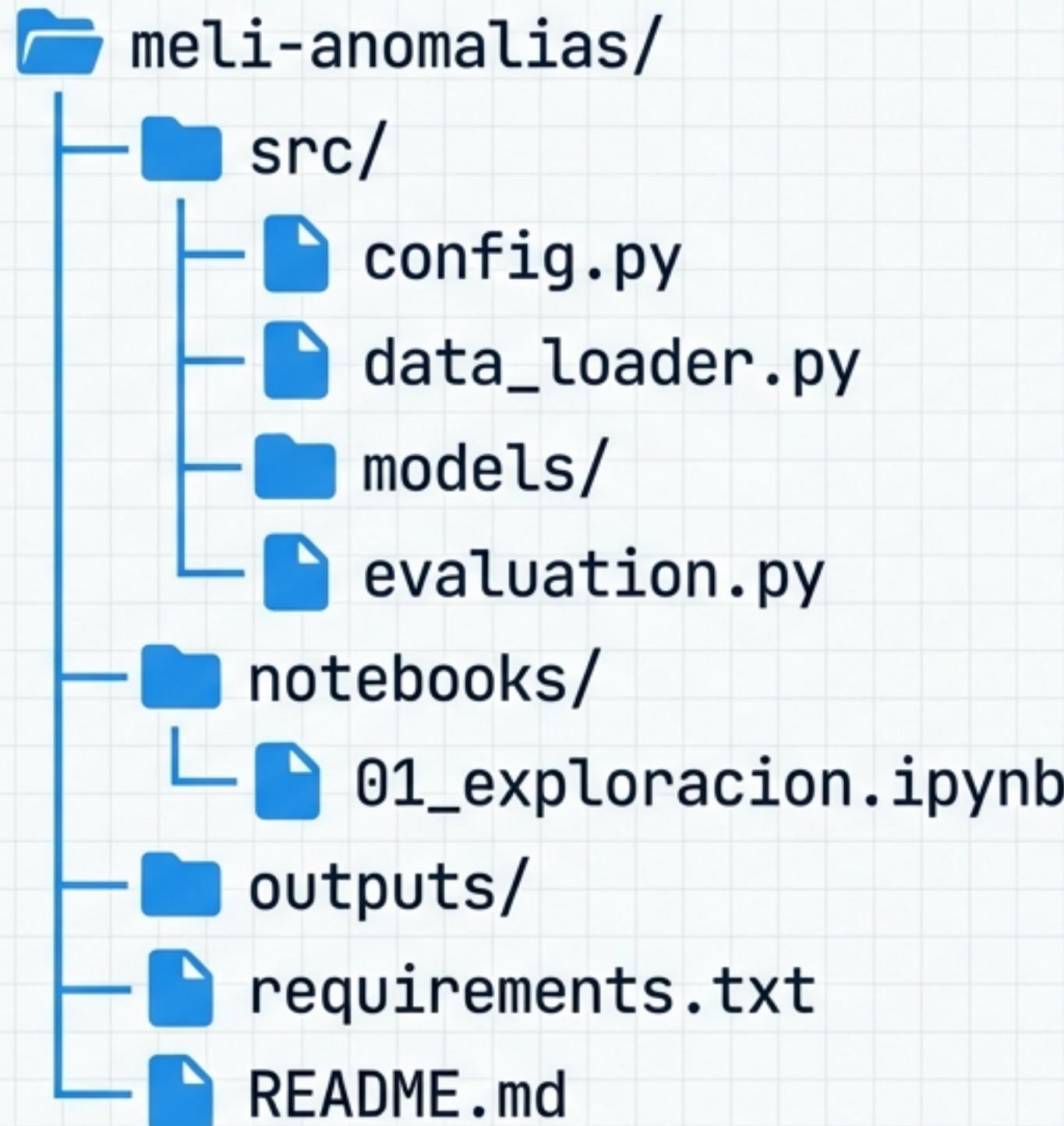
Distribución Bootstrap de la Diferencia ($\Delta F1$)



Intervalo de Confianza 95% (IC95%)

Diferencia en F1 Score

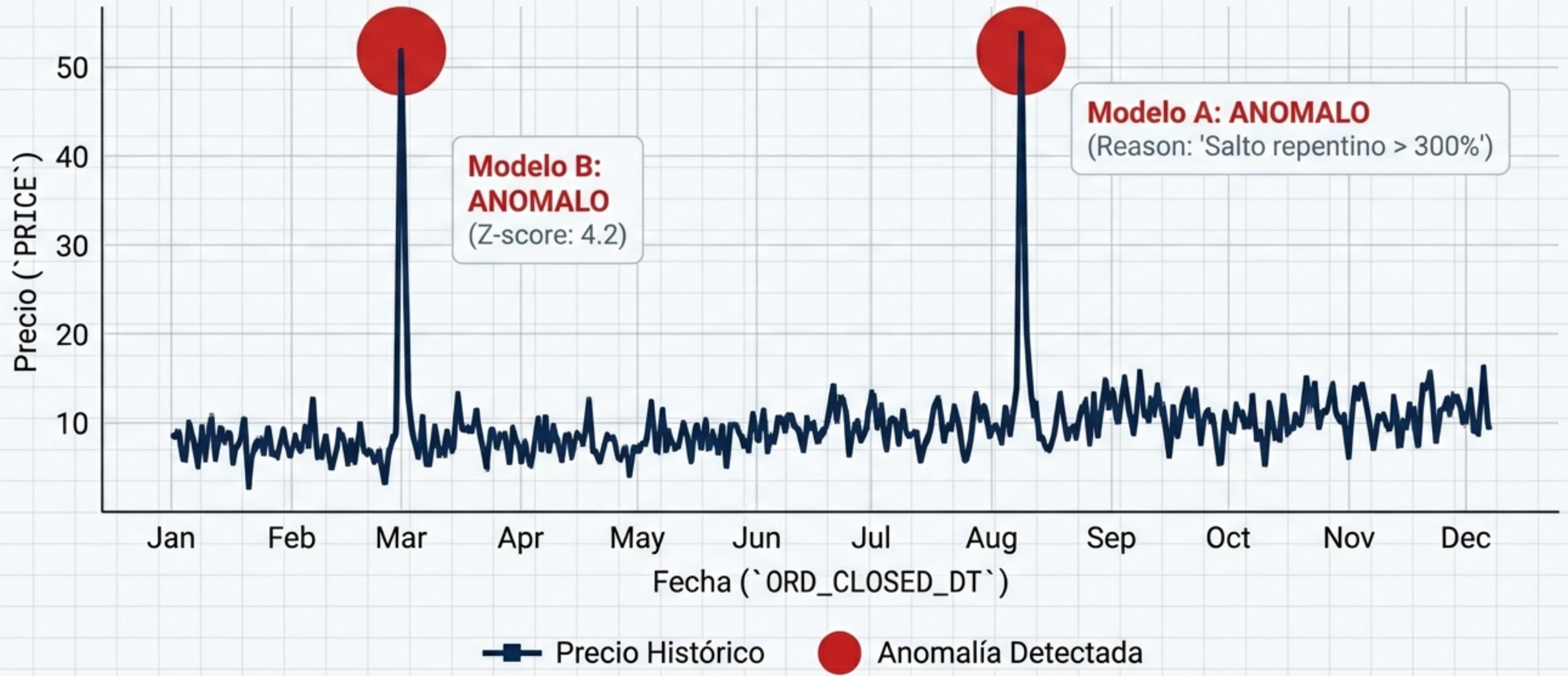
Ingeniería de Software y Reproducibilidad



Buenas Prácticas Implementadas

- ✓ Configuración centralizada (`config.py`)
- ✓ Gestión de dependencias (`venv + requirements`)
- ✓ Código modular reutilizable (`Scripts vs Notebooks`)
- ✓ Seeds fijas para reproducibilidad científica

Visualización de Resultados



Resultados Cuantitativos Comparativos

Modelo	F1-Score	Precision	Recall	Latencia (ms)	Costo (\$)
Modelo B (Baseline)	0.92	0.90	0.94	< 1ms	0.00
Modelo A (LLM)	0.89	0.85	0.93	~450ms	\$0.002 / call

- **Insight:** El modelo estadístico domina en eficiencia. El LLM introduce latencia significativa.
- **Conclusión A/B:** La mejora en F1 no es estadísticamente significativa ($p > 0.05$) para justificar el costo en detección pura.

Conclusiones y Siguientes Pasos

Resumen Ejecutivo

El **modelo estadístico (Baseline)** es la opción recomendada para producción en tiempo real debido a su robustez y eficiencia.

El **LLM** debe reservarse para auditoría offline o como 'segunda opinión' para explicar casos complejos.



Corto Plazo:

Containerización (Docker) y CI/CD.

Medio Plazo:

Conexión a API Real (OpenAI/DeepSeek) para el Challenger.

Largo Plazo: Implementar Agente Crítico para validar razonamientos del LLM.