

Data Science & Machine Learning Challenge

Bienvenido al proceso de selección técnica. Este desafío está diseñado para evaluar tus habilidades en ciencia de datos, modelado estadístico y visión de negocio aplicada al sector de retail.

Buscamos identificar tu capacidad para estructurar problemas complejos, generar código de producción y comunicar hallazgos técnicos a stakeholders de negocio.

Objetivo General

Debes seleccionar **uno** de los siguientes tres casos de uso y desarrollar una solución "end-to-end". Asume el rol de Lead Data Scientist; tu solución debe ser técnicamente sólida y comercialmente viable.

Entregables Requeridos

El candidato debe proveer un enlace a un **Repositorio de GitHub público** que incluya:

1. Código Fuente:

- Jupyter Notebooks para análisis exploratorio y experimentación.
- Scripts modularizados (.py) para el preprocesamiento y entrenamiento final.
- Se valorará el uso de buenas prácticas de ingeniería de software (typing, docstrings, estructura de carpetas).

2. Documentación Técnica (README.md):

- Instrucciones claras de instalación y ejecución (ej. requirements.txt, environment.yml o Dockerfile).
- Breve explicación de la arquitectura de la solución.

Debes enviar el link al repositorio público en github.

3. Presentación Ejecutiva (PPT, PDF o similares):

- Resumen del problema y enfoque metodológico.
- Resultados del análisis, modelo y métricas de evaluación.
- Impacto de negocio estimado e interpretación de resultados.
- Conclusiones y próximos pasos.

Debes enviar la presentación.



Casos de Uso



Caso A: Optimización de Abastecimiento

Contexto

Predecir la demanda es solo la mitad de la batalla. Si pedimos menos, perdemos ventas (costo de oportunidad); si pedimos más, incurrimos en costos de almacenamiento y capital.

Objetivo

1. Desarrollar un modelo de pronóstico de demanda por SKU-Tienda para la siguiente semana.
2. Diseñar una **función de optimización simple** que determine la cantidad de pedido final, buscando minimizar el costo total esperado (balanceando el costo de *Stockout* vs. el costo de *Overstock*).

Datos Disponibles

```
ventas_historicas.csv: fecha, id_tienda, id_producto, unidades_vendidas.  
inventario_actual.csv: id_tienda, id_producto, stock_actual.  
catalogo_productos.csv: id_producto, nombre, costo_unitario, precio_venta,  
costo_almacenamiento_semanal.  
maestro_tiendas.csv: id_tienda, ciudad, tamaño_m2.
```

Consideraciones Técnicas

El candidato debe proponer cómo usar la incertidumbre del modelo (intervalos de confianza) junto con los márgenes del producto para decidir si ser agresivo o conservador en el pedido.



Caso B: Creación de Combos

Contexto

Buscamos incrementar el ticket promedio mediante estrategias de venta cruzada. Es necesario identificar patrones de compra no evidentes para generar recomendaciones

automáticas.

Objetivo

Implementar un algoritmo para identificar los Top 5 "Combos" (conjuntos de productos) con mayor potencial de venta (proponer el precio de cada combo) para diferentes clusters de tiendas, cuantificando el "lift" esperado.

Datos Disponibles

```
tickets.csv: id_ticket, fecha, id_tienda, id_cliente.  
detalle_tickets.csv: id_ticket, id_producto, cantidad, precio_unitario.  
catalogo_productos.csv: id_producto, nombre, categoria, subcategoria.
```

Consideraciones Técnicas

Uso de Reglas de Asociación o enfoques basados en grafos/embeddings. Estrategias para filtrar artículos de "ruido" (alta frecuencia global pero baja correlación específica).



Caso C: Modelado de Ticket Promedio (AOV Drivers)

Contexto

Existe una alta variabilidad en el Ticket Promedio (AOV) entre sucursales. Se requiere entender qué factores exógenos y endógenos influyen en este indicador.

Objetivo

1. Construir un modelo inferencial para determinar la importancia de las variables (feature importance) en el valor del ticket.
2. Desarrollar un modelo predictivo para estimar el gasto esperado de un cliente recurrente en su próxima visita.

Datos Disponibles

```
transacciones_resumen.csv: id_ticket, id_tienda, timestamp, total_venta,  
total_articulos, id_cliente.  
promociones_activas.csv: fecha_inicio, fecha_fin, id_producto, tipo_descuento,  
id_tienda.  
variables_exogenas.csv: fecha, id_tienda, clima, competitor_price_index,
```

```
indice_trafico.  
clientes_loyalty.csv: id_cliente, fecha_registro, edad, segmento.
```

Consideraciones Técnicas

Interpretabilidad del modelo, manejo de outliers y selección de métricas de error adecuadas para regresión.

Criterios de Evaluación

Se evaluará la madurez técnica y de negocio del candidato bajo los siguientes pilares:

1. **Análisis Exploratorio (EDA):** Profundidad en el entendimiento de los datos, detección de anomalías y generación de hipótesis.
2. **Rigor Metodológico:** Justificación en la selección de algoritmos y diseño correcto de la estrategia de validación (Cross-Validation, Time-Series Split).
3. **Calidad de Código:** Código limpio, reproducible y listo para escalar.
4. **Visión de Negocio:** Capacidad para traducir métricas técnicas (RMSE, AUC, Lift, etc) en KPIs de negocio (Ahorro de costos, Margen incremental, etc).