**Prueba técnica - Ciencia de Datos**

Cuestionario Solución

1. De las deficiencias en los datos, ¿cuáles y cómo las identificaste?

Algunas de las deficiencias identificadas fueron:  
Valores NaN dentro del dataset, esto fue manejado utilizando una función de limpieza que elimina filas con valores faltantes o infinitos, ya que muchas de estas variables requieren datos completos para tener sentido estadístico.

Analicé distribuciones con describe() y visualizaciones (boxplots, histogramas) para encontrar los principales insights del dataset.

2. De realizar creación de variables, explica cuales hiciste y por qué.

total\_delay, para capturar historial acumulado de morosidad.

pay\_ratio1, evaluar capacidad de pago relativa a la deuda.

3. De los modelos realizados, ¿cómo seleccionaste al mejor?

Para elegir el mejor modelo, se siguió un proceso basado en las necesidades específicas de cada problema(clasificación y regresión) priorizando que las métricas se alinearan a las necesidades del negocio.

La selección se basó en un equilibrio entre métricas técnicas (AUC-ROC, RMSE) y viabilidad operativa (velocidad, interpretabilidad)

4. ¿Qué desafíos encontraste y cómo los superaste?

Primeramente, el tiempo de desarrollo, ya que al tener otras actividades no me fue posible reintegrar mejoras en el modelo pero se identificaron como posibles mejoras en el modelo a entregar.

En cuanto a detalles técnicos:

El mayor reto fue trabajar con datos faltantes en fechas clave o inconsistencias entre inventario y ventas. Se podría resolver aplicando imputaciones condicionales y reglas de negocio (por ejemplo, asignar cero si no hay stock).

Otro reto sería la escalabilidad, al trabajar con millones de registros, la eficiencia fue crucial por lo que se puede optar por trabajar en entornos de mejor eficiencia como PySpark.

**Cuestionario general**

1. **Establece con tus propias palabras, algunas buenas prácticas y funciones recomendadas para optimizar operaciones de lectura, escritura y manipulación en Spark/PySpark.**

Primeramente, el proceso ETL trabajarlo como una petición a la API de donde se están obteniendo los datos y guardarlos en una tabla delta para su manipulación ya que es formato comprimido y permite una lectura selectiva a comparación de usar archivos json o csv que podrían ser más lentos y propensos a errores.

En cuanto a trabajar en entornos de big data, auxiliarse con queries de sql para un mejor manejo de los datos y de la base de datos en donde se encuentre.

1. **Indica las pruebas estadísticas que has utilizado como parte del desarrollo de una solución de ciencia de datos.**

Chi-cuadrado para analizar independencia entre variables categóricas.

Correlación de Pearson o Spearman para estudiar relaciones entre variables numéricas.

Z-test Para la detección de outliers.

T-test y ANOVA para comparar medias entre grupos y evaluar diferencias estadísticamente significativas.

1. **En el contexto de Machine Learning y Ciencia de datos, explica:**
   1. **No Free Lunch Theorem**

Este concepto nos indica que los modelos no necesariamente funcionan igual para todos los problemas planteados, esto depende del contexto, requerimientos del negocio, contexto general del negocio y sus diferentes variables.

* 1. **Occam’s Razor**Dado un problema que pueda resolverse con diferentes modelos, se sugiere usar el más simple.
  2. **Data Leakage**

Cuando en un dataset se filtra información; por ejemplo de datos futuros, ocasionando que se predicen datos buenos para el modelo pero para el negocio serían irreales.