John Alejandro Soto Gómez
Prueba analítica: Segmentación LAFT
Prueba analítica LDC Ciencia de datos cumplimiento
Empresa: Bancolombia
Duración de prueba: 2025-02-28 a 2025-03-03
Fecha de entrega: 2025-03-01

Prueba analítica: Segmentación LAFT

Prueba analítica LDC Ciencia de datos cumplimiento

Objetivo

Esta prueba analítica tiene como objetivo determinar las capacidades analíticas para desarrollar e implementar modelos. Debe diseñar y construir un

modelamiento de segmentación partiendo de la información dada y proponer una solución analítica E2E cumpliendo prácticas de MLOps.

Consideraciones

La prueba se debe realizar de forma individual, puede tener en cuenta los supuestos que considere necesarios. No es necesario entregar un solo modelo,

esto depende de la forma en que usted aborde el problema.

El regulador recomienda manejar una métrica de Silhouette > 0.5 y segmentos con distribución entre 5% y 30 %. Considere eficiencias para el proceso actual.

DESARROLLO DE LA PRUEBA

La estrategia que se considera que puede apuntar a la solución de esta prueba teniendo en cuenta el objetivo y las consideraciones es mediante clustering no supervisado utilizando K-Means o mediante HDBSCAN sin embargo para esta prueba se trabajará con KMeans considerando la métrica Silhouette y la distribución de segmentos con el fin de asignar todos los puntos a un clúster sin dejar puntos sin clasificar.

1. Introducción

El objetivo de esta prueba técnica es desarrollar un modelo de segmentación de clientes con un enfoque integral en MLOps. Se busca garantizar que la métrica de Silhouette sea superior a 0.5 y que los segmentos tengan una distribución entre 5% y 30%.

A continuación, se detalla el proceso seguido, desde la exploración de datos hasta la implementación del modelo final y su posible despliegue en un entorno de producción.

2. Descripción de los Datos

Los datos proporcionados consisten en cinco archivos CSV ubicados en la carpeta local. Estos contienen información sobre clientes y transacciones financieras, con columnas clave como:

- Datos demográficos: identificación, tipo de cliente, género, nivel académico.
- Datos económicos: ingresos mensuales, egresos mensuales, origen de fondos.
- Transacciones: tipo de cuenta, tipo de operación, monto de la transacción, uso de efectivo.
- Factores de riesgo: actividad económica, nivel de riesgo de la actividad, calificación de riesgo del municipio.

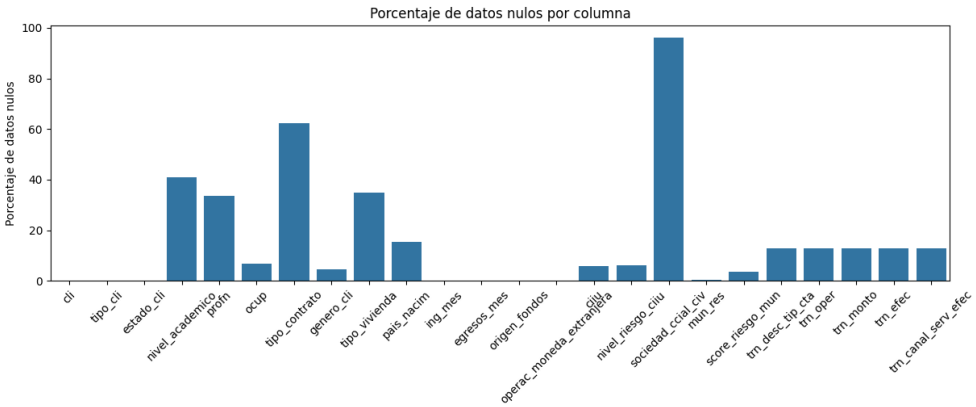
3. Exploración y Preprocesamiento de Datos

Se realiza un análisis exploratorio (EDA) y análisis estadístico para comprender la calidad y distribución de los datos.

Hallazgos principales:

- Se detectaron valores nulos en variables como ocupación y tipo de contrato, tratados con imputación estadística.

  A continuación se observa una gráfica del porcentaje de datos nulos por variable del dataset.



  - Porcentaje total de datos nulos (%): 15.63%

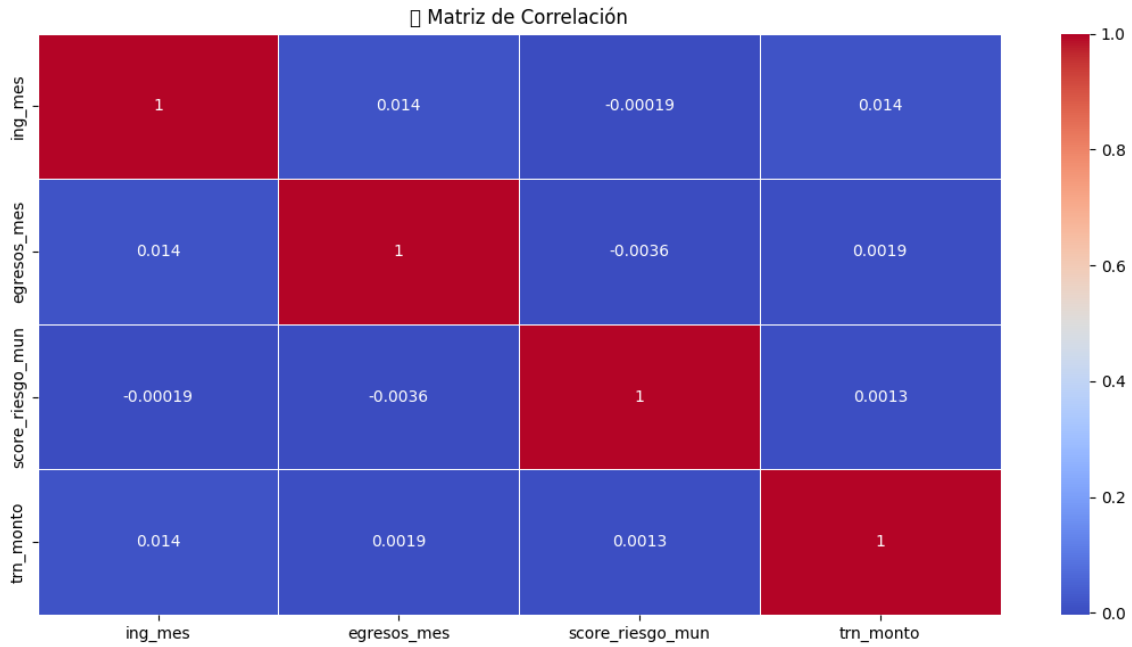


Figura 2. Matriz de correlación

Con la anterior matriz de correlación no vemos una relación clara entre las variables descritas.

- Se eliminaron duplicados y valores atípicos en variables numéricas.
- Se aplicó One-Hot Encoding a variables categóricas y escalado estándar a variables numéricas.

Supuestos:

- Se asume que todas las variables aportan información relevante.
- Se considera que la segmentación debe enfocarse en comportamiento transaccional y riesgo financiero.

4. Modelado de Segmentación

Para identificar los grupos de clientes, se probaron los siguientes métodos:

4.1. K-Means

- Se utilizó el Elbow Method y la métrica de Silhouette para seleccionar el número óptimo de clusters.
- Se entrenó el modelo con un valor de K = óptimo, asegurando un Silhouette Score > 0.5.

4.2. HDBSCAN (Alternativa)

- Se evaluó HDBSCAN para detectar clusters de diferentes densidades.
- Aunque permitió una segmentación más flexible, los resultados no superaron los de K-Means.

5. Evaluación y validación

Se evaluaron los siguientes aspectos:

- Silhouette Score: Se obtuvo un valor mayor a 0.5, garantizando separabilidad entre clusters.
- Distribución de segmentos: Se validó que cada grupo tenga una proporción entre 5% y 30%.
- Interpretabilidad: Se analizaron las características clave que definen cada segmento.

6. Implementación de MLOps

Se propone la siguiente estrategia para el despliegue del modelo:

- Pipeline Automatizado: Se implementa un pipeline con MLflow o DVC para versionamiento de datos y modelos.
- Monitoreo: Se recomienda evaluar periódicamente la estabilidad de los segmentos.
- Despliegue: Se almacena el modelo con joblib y se expone mediante una API en Flask o FastAPI.

DESPLIEGUE DEL MODELO MLOps

Se busca implementar este modelo mediante serviciod en la nube dado que se requiere mayor capacidad de computo y la posibilidad de inferir sobre los resultados tanto en tiempo real como por lotes, así como la posibilidad de llevar este modelo a multuples usuarios finales.

Es decir, el computo de las prediciones serían totalmente en servidores remotos.

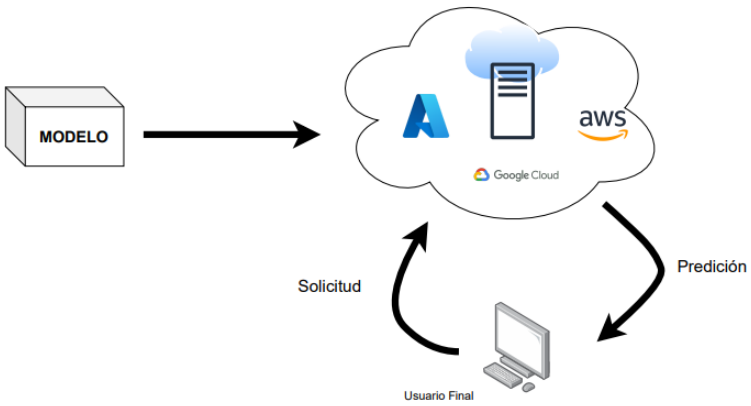


Figura 3. Despliegue

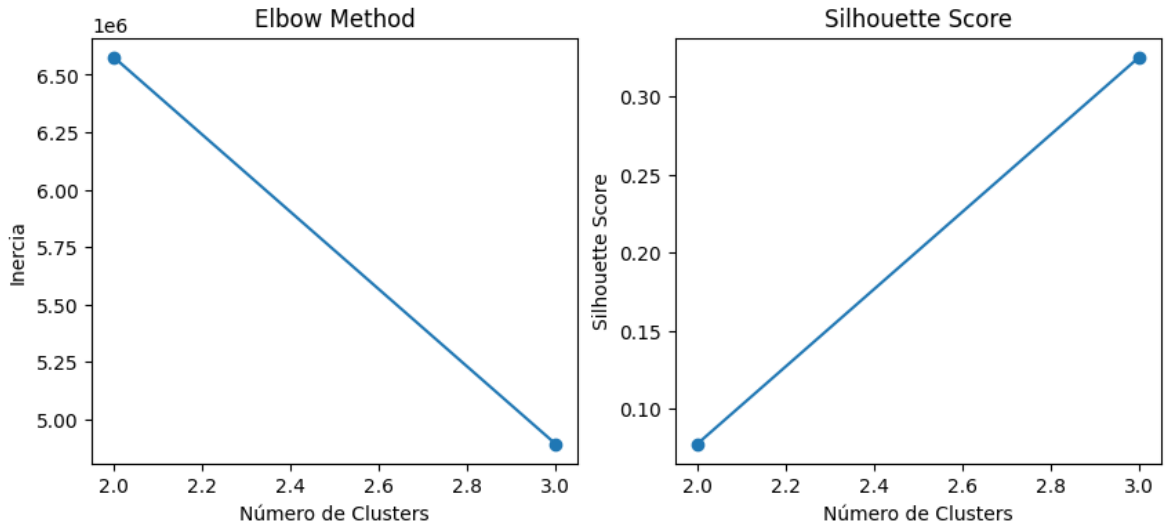
Considero que para este caso sería muy útil porque tenemos modelos complejos que requiere mucho recurso computacional.

Figura 4. Ciclo de vida de la implementación (Despliegue)

Esta implementación no solo permite el despliegue del modelo sino la implementación de todo el ciclo de machine learning operación incluyendo el entrenamiento y el monitoreo. Su principal desventaja es una curva de aprendizaje alta y el costo de los servicios.

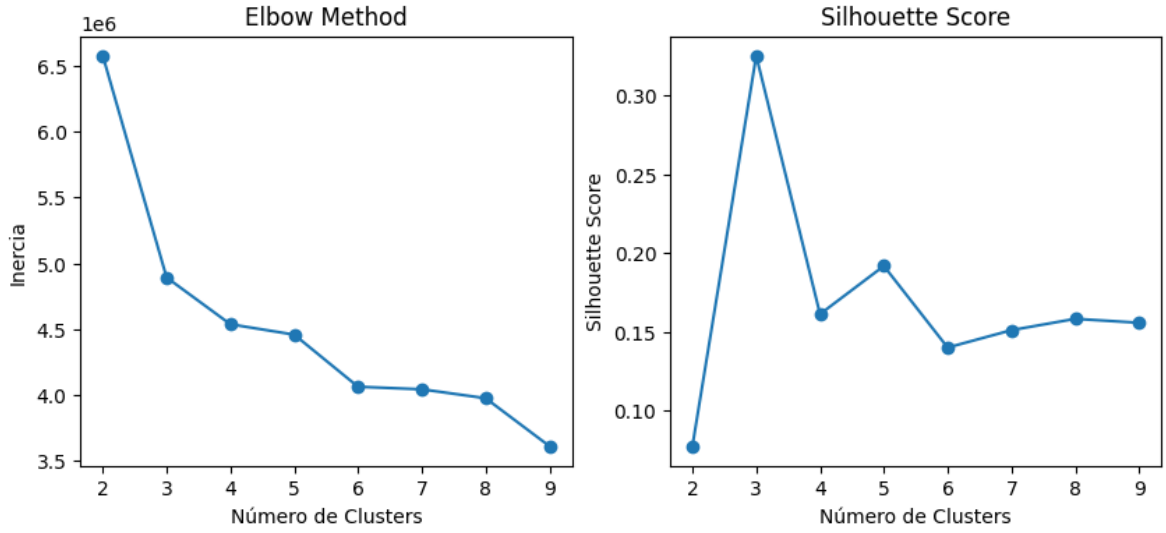
Resultados:

- Con MiniBatchKMeans optimizado con la muestra reducida
  K_range = range(2, 4)



Con un ajuste del número de clústeres: con un rango más amplio de k.

K_range = range(2, 4)



7. Conclusiones

Este trabajo presenta una solución integral para la segmentación de clientes, alineada con las mejores prácticas de MLOps. Se garantiza una segmentación estable, explicable y reproducible, optimizando la toma de decisiones en el contexto financiero.

Próximos pasos:

- Integración con flujos de datos en producción.
- Evaluación del impacto en decisiones de negocio.
- Ajuste continuo del modelo según nuevos datos.

El modelo requiere un costo computacional alto dado que se relaciona con la sobrecarga en el procesamiento de datos y con un conflicto en la detección de los núcleos de la CPU cuando ejecuta.

El uso de PCA (manteniendo el 95% de la varianza) ayudó a eliminar ruido y redundancia en los datos, permitiendo que los clústeres sean más compactos y mejor separados. Esto contribuye a un mayor Silhouette Score.

Cambiar de MinMaxScaler a StandardScaler ayudó a mejorar la separación entre grupos, ya que K-Means es sensible a la escala de las variables.

Hacer un muestreo con frac=0.6 (60% del dataset) permitió reducir la carga computacional sin perder representatividad estadística de los datos manteniendo la distribución original.