Formas de gasto: una segmentación bancaria con Análisis Topológico de Datos

Daniel Eduardo Arana Bodart, A01741202

Mayo 2025

Abstract

El uso de Análisis Topológico de Datos (TDA) en el ámbito financiero representa un enfoque novedoso para descubrir estructuras complejas que no son evidentes mediante técnicas tradicionales de análisis. En este proyecto, se aplicó TDA a un conjunto de datos transaccionales de tarjetas de crédito utilizadas en la India, con el objetivo de identificar y clasificar a los usuarios según su comportamiento de gasto, y evaluar si estas agrupaciones pueden servir como base para estrategias comerciales diferenciadas. La hipótesis plantea que los usuarios pueden ser segmentados en grupos diferenciables con estructuras topológicas persistentes. Para evaluar esta hipótesis, se implementaron técnicas de clusterización, análisis de homología persistente y un modelo de clasificación supervisada. Los resultados muestran que, si bien es posible identificar un grupo principal de comportamiento común, no se hallaron diferencias suficientemente significativas entre los supuestos subgrupos, lo que lleva a rechazar la hipótesis inicial. A pesar de ello, el estudio demuestra el potencial del TDA como herramienta exploratoria en contextos financieros complejos.

1 Introducción y Objetivo

El Análisis Topológico de Datos no es una estrategia desconocida en el mundo de las finanzas, pues en la actualidad existen diversos proyectos como este que emplean TDA para sacarle el máximo provecho a la información, como lo fue el proyecto de Sourav Majumdar y Arnab Kumar Laha denominado *Clustering and classification of time series using topological data analysis with applications to finance* en donde proponen métodos de clasificación y clusterización de series de tiempo basado en técnicas de Análisis Topológico de Datos.

El análisis del comportamiento financiero de usuarios a partir de sus transacciones con tarjetas de crédito ofrece un campo fértil para la exploración de patrones ocultos, útiles en estrategias comerciales, gestión de riesgo y segmentación. Este proyecto utiliza un conjunto de datos recopilado por Sadat Akash, que incluye información transaccional de usuarios en la India: ciudad, fecha, tipo de tarjeta, tipo de gasto, género del usuario y monto gastado.

El objetivo central es analizar esta información mediante técnicas de Análisis Topológico de Datos (TDA), una metodología emergente que permite detectar agrupamientos, ciclos y estructuras de alta dimensión que no son visibles con herramientas estadísticas convencionales.

A partir de este enfoque, se plantean las siguientes preguntas de investigación:

- ¿Qué grupos de clientes existen según su comportamiento de gasto con tarjetas de crédito?
- ¿Existen patrones o comportamientos repetitivos en el uso de las tarjetas?
- ¿Qué tan conectados están los clientes entre sí según su perfil financiero?
- ¿Qué tan diferenciables son los grupos encontrados en el análisis?
- ¿Qué tan estables son los patrones topológicos identificados?

- ¿Es posible predecir a qué grupo pertenece un nuevo cliente con base en sus características?

Estas preguntas se abordan mediante Mapper, homología persistente y otras herramientas del TDA, con el objetivo de evaluar la viabilidad de segmentaciones financieras robustas a partir de estructuras topológicas.

2 Justificación e Hipótesis

El sector financiero representa una de las áreas con mayor aplicación de análisis de datos, y al mismo tiempo, plantea retos significativos para el descubrimiento de patrones no lineales en el comportamiento de los usuarios. Dado el interés profesional del autor en este ámbito, se consideró pertinente explorar un enfoque topológico para la segmentación de clientes, basado en sus hábitos de gasto.

> ***Hipótesis:*** *Los usuarios de tarjetas de crédito en la India pueden ser agrupados en segmentos diferenciables según su comportamiento de gasto, y estos segmentos presentan estructuras topológicas persistentes que permiten tanto su análisis como su predicción mediante técnicas del Análisis Topológico de Datos).*

3 Metodología

El flujo metodológico seguido se estructuró en cuatro etapas principales: (1) preprocesamiento y codificación de los datos, (2) análisis estadístico exploratorio, (3) aplicación de técnicas de Análisis Topológico de Datos (TDA), y (4) entrenamiento de un modelo de clasificación supervisada para evaluación complementaria.

3.1 Preprocesamiento de los datos

Se trabajó con un conjunto de 26,052 registros de transacciones realizadas con tarjetas de crédito en diversas ciudades de la India. Las variables categóricas (como ciudad, tipo de tarjeta y tipo de gasto) fueron transformadas mediante codificación *One-Hot*, agrupando las ciudades con pocos registros en una categoría común (`Other`) para reducir la dimensión. La variable `Amount` fue normalizada con `RobustScaler`, debido a su resistencia frente a valores atípicos.

3.2 Análisis estadístico preliminar

Se calculó un conjunto de estadísticas descriptivas (media, mediana, varianza, conteo por categoría) con el objetivo de caracterizar la base de datos, identificar posibles desequilibrios en la distribución de las variables y guiar la selección de métodos topológicos y modelos de evaluación.

3.3 Técnicas topológicas empleadas

- **Mapper:** se aplicó utilizando UMAP como función de filtro y DBSCAN como algoritmo de agrupamiento. Los parámetros fueron afinados empíricamente hasta obtener una cobertura y superposición adecuadas.
- **Homología Persistente:** se construyeron complejos de Rips sobre submuestras aleatorias del conjunto completo, dada la alta complejidad computacional. Se extrajeron los grupos de homología H_0 y H_1 mediante la biblioteca `ripser` y se visualizaron mediante diagramas de persistencia y *barcodes*.

- **Persistence Landscapes:** se calcularon sobre cinco submuestras independientes utilizando funciones triangulares aproximadas, con el fin de analizar la estabilidad de los patrones topológicos detectados.

3.4 Clasificación supervisada

Se empleó un modelo de *Random Forest* para intentar predecir los nodos asignados por Mapper como etiquetas generadas no supervisadamente. El modelo fue entrenado y evaluado con una división estándar $70\% - 30\%$ entre datos de entrenamiento y prueba, y se evaluó mediante métricas de clasificación sin ajuste de hiperparámetros.

4 Resultados

Nota: Para interpretaciones más detalladas y visualizar todas las gráficas generadas, se recomienda consultar el Notebook anexo.

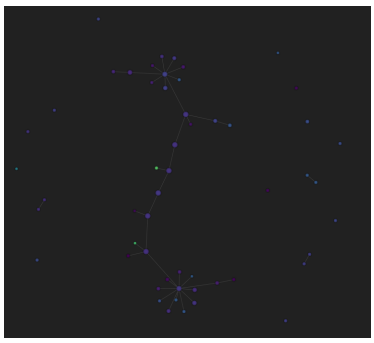

Figura 1. Mapper de los datos.

En el Mapper se observa un solo cluster general que contiene la mayoría de los nodos y varios puntos sin agrupación.

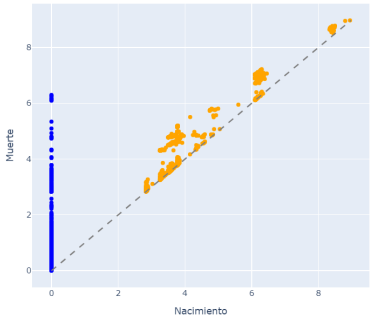

Figura 2. Diagrama de Persistencia.

El Diagrama de Persistencia muestra un comportamiento similar al Mapper, pues se ve una sola componente conexa persistente, así como varios 1-huecos que son medianamente persistentes.

	Grupo 1	Grupo 2	Bottleneck H_1	Wasserstein H_1
2	Gold	Silver	0.135958	13.077494
5	Signature	Silver	0.138953	11.336258
0	Gold	Platinum	0.124724	10.473221
3	Platinum	Signature	0.124293	8.826137
4	Platinum	Silver	0.104475	6.950988
1	Gold	Signature	0.074627	6.208132

Tabla 1. Distancia Bottleneck y Wasserstein entre grupos de diferentes tarjetas de crédito.

Las distancias entre estos grupos es bastante pequeña, lo que sugiere que no hay tanta diferencia en sus propiedades topológicas, esto coincide con ambos resultados previos.

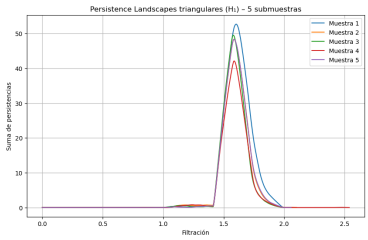

Figura 3. Landscapes de muestras aleatorias de los datos.

En los Landscapes se observa que las cinco muestras siguen el mismo comportamiento, por lo que conservan las propiedades topológicas.

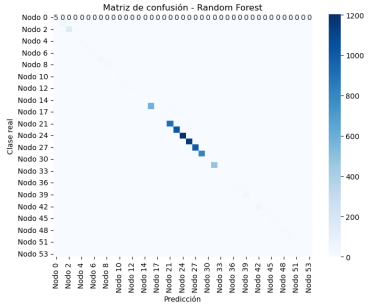

Figura 4. Matriz de confusión del modelo de clasificación.

Se observa que la matriz de confusión prácticamente no presenta errores y la gran mayoría de las predicciones se encuentran en los nodos centrales.

5 Conclusiones

Nota: Para una conclusión más detallada y la explicación de porqué se llegó a esa conclusión revisar el Notebook anexado. En base a los resultados obtenidos se pueden contestar las preguntas de investigación:

- ¿Qué grupos de clientes existen según su comportamiento de gasto con tarjetas de crédito? A través del Mapper y comparando el resto de resultados se llegó a la conclusión de que solamente existe un grupo principal de clientes en donde se engloba la gran mayoría de estos, por lo que no hay tanta diferencia topológica entre los registros de este grupo, a excepción de uno cuantos grupos muy pequeños que son prácticamente intrascendentes para el análisis global.
- ¿Existen patrones o comportamientos repetitivos en el uso de las tarjetas? En este caso con ayuda de los H_1 más persistentes se determinó que hay comportamientos similares no lineales entre diversos usuarios, aunque estos no son totalmente persistentes.
- ¿Qué tan conectados están los clientes entre sí según su perfil financiero? Para esta pregunta se toman en cuenta los H_0 persistentes, y como en este caso el más persistente es una sola componente conexa se puede contestar que la gran mayoría de clientes se encuentran conglomerados en un perfil financiero similar, con algunas excepciones que también son ligeramente persistentes.
- ¿Qué tan diferenciables son los grupos encontrados en el análisis? Al menos al separar los grupos por género (revisar anexo) y por tipo de tarjeta, no son para nada diferentes, lo que tiene sentido al encontrar que solo hay un grupo para separar los datos.
- ¿Qué tan estables son los patrones topológicos identificados? Los patrones son bastante estables, ya que para todos los subgrupos que se crearon se obtuvo un comportamiento bastante similar, lo que indica que las propiedades topológicas dentro de los grupos persisten, quizá en mayor o menor escala, pero tienen propiedades similares.
- ¿Es posible predecir a qué grupo pertenece un nuevo cliente con base en sus características? Si, se puede predecir y con bastante precisión a qué grupo pertenecerá un usuario nuevo, pero esto no es porque el modelo sea excelente; sino más bien porque solamente existe un grupo significativo al que el uso de una tarjeta de crédito podría pertenecer.

La hipótesis original de este proyecto planteaba que los clientes de tarjetas de crédito en la India podían agruparse en segmentos claramente diferenciables según su comportamiento de gasto, y que dichas agrupaciones revelarían estructuras topológicas persistentes, útiles tanto para su análisis como para su predicción mediante técnicas de Análisis Topológico de Datos (TDA).

Sin embargo, los resultados obtenidos no respaldan esta hipótesis. Aunque se lograron agrupar las transacciones mediante Mapper y se detectaron patrones cíclicos mediante homología H_1, la gran mayoría de las observaciones se concentraron en un único grupo dominante, sin una diferenciación significativa respecto a los demás. Esta conclusión fue consistente en todas las etapas del análisis, incluyendo la clusterización, los diagramas de persistencia y el intento de clasificación supervisada.

A pesar de ello, el ejercicio permitió aplicar de manera efectiva las herramientas del TDA, validar su capacidad descriptiva y revelar una estructura relevante en el conjunto de datos, aunque no fuera la esperada. Este hallazgo resalta tanto los límites como el valor del TDA en escenarios reales donde la segmentación no es evidente.

6 Anexos

Nota: A pesar de ser un Google Drive, se puede verificar que todos los anexos están subidos antes de la hora límite de entrega.

Repositorio del proyecto: `https://drive.google.com/drive/folders/18y8eEKzNlYgTUEgfbORL4XxV4MVaJB
usp=sharing`

References

Ucán A. N7_mapper.ipynb, 2025a. Notebook base utilizado para construir la visualización Mapper.

Ucán A. N4_pdmetrics.ipynb, 2025b. Notebook base para métricas entre diagramas de persistencia.

Ucán A. N2_persistenthomology.ipynb, 2025c. Notebook base para diagramas de persistencia con Ripser y Gudhi.

Ucán A. N6_topologicallydimensionalityreduction.ipynb, 2025d. Notebook base para técnicas de reducción dimensional topológica.

Sadat Akash. Analyzing credit card spending habits in india. `https://www.kaggle.com/datasets/thedevastator/analyzing-credit-card-spending-habits-in-india`, 2020. Base de datos utilizada en este estudio.

Kumar Laha A. Majumdar S. Topological data analysis: A new perspective on big data. *Expert Systems with Applications*, 149, 2020. ISSN 0957-4174. doi: 10.1016/j.eswa.2020.113211. URL `https://www.sciencedirect.com/science/article/abs/pii/S095741742030676X`. Lectura de referencia teórica sobre TDA aplicado en finanzas.

OpenAI. Chatgpt, 2020. Asistente utilizado para resolver cuellos de botella computacionales en diagramas de persistencia y barcodes.

scikit-learn developers. Robustscaler — scikit-learn documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html`, 2024. Referencia del método de normalización utilizado.