Universidad de Los Andes

Maestría en Inteligencia Analítica de Datos

Proyecto Final Aprendizaje no supervisado – entregable semana 4

Autores:

Jose Daniel Garcia Correa

Cristian Mauricio Romero Buitrago

Cesar Daniel Ramírez Cely

Departamento de Ingeniería Industrial

1. Resumen

Las tarjetas de crédito son el mecanismo de deuda más común en las personas, desempeñando un papel crucial en la construcción del historial crediticio, ya que suele ser el primer instrumento financiero de muchas personas. Para los bancos, este rubro es fundamental no solo por los ingresos generados a través de los intereses y comisiones, sino también porque permite evaluar la capacidad de pago de los clientes, lo que a su vez facilita la oferta de otros productos financieros llevando a la fidelización de los mismos.

Según el reporte de la Superintendencia Financiera de Colombia hasta mayo de 2024, los bancos en el país cuentan con 14,2 millones de tarjetas de crédito vigentes, concentradas principalmente en 8 de las 26 entidades financieras presentes en el reporte. Bancolombia y Scotiabank Colpatria lideran el mercado, con participaciones del 16,8% y 13,92% respectivamente. La mayoría de estas tarjetas están asociadas a las franquicias de Visa y MasterCard, que controlan el 88% del mercado con más de 12 millones de tarjetas.

Sin embargo, a pesar de que parece que las tarjetas de crédito son una herramienta de uso común, su uso sigue siendo limitado. De acuerdo con la información de la Superfinanciera de Colombia con corte mayo 2024, el 64% del cupo aprobado en el sistema financiero para este rubro no es utilizado, con un monto aproximado de 72,5 billones de pesos aún sin utilizarse. En el caso de uno de los bancos más importantes del suroccidente colombiano, objeto del presente estudio, solo el 31% de un cupo total de 5,5 billones de pesos está en uso, lo que deja un potencial de profundización aproximado de 3,8 billones.

El análisis del mercado sugiere que existe una oportunidad importante para incrementar la cartera en la profundización del uso de las tarjetas de crédito de los clientes vinculados. Para lograrlo, es esencial implementar estrategias de promoción que incentiven un mayor uso de las tarjetas. El éxito de estas estrategias se debe medir por el aumento en el número de transacciones y el valor promedio por transacción. Para esto se sugiere el uso del análisis de datos mediante la implementación de algoritmos no supervisados, que permitan identificar grupos con comportamientos de consumo específicos, para diseñar campañas efectivas. Los segmentos encontrados deben ser claramente diferenciables, alcanzables y con un significado interpretativo claro para el negocio, minimizando la diferencia entre los individuos de cada grupo y maximizando las diferencias entre los grupos, con el fin de impulsar un uso más efectivo y rentable de las tarjetas de crédito.

Para realizar una óptima segmentación de clientes, son aplicables métodos de aprendizaje no supervisado como el clustering con K-medias para agrupar clientes según sus patrones de consumo, el clustering jerárquico puede revelar la estructura de los segmentos de clientes, mientras que los algoritmos de detección de anomalías pueden identificar comportamientos atípicos. Cabe mencionar que se

debe realizar un proceso de estandarización de los datos, de modo que los resultados no se vean afectados por las unidades de medida ni por los datos atípicos, ya que pueden existir clientes que tengan un comportamiento inusual en el uso de sus tarjetas de crédito.

2. Introducción

En el contexto actual del sistema financiero colombiano, el uso de las tarjetas de crédito representa tanto un desafío como una oportunidad significativa para las entidades bancarias. Con más de 14,2 millones de tarjetas de crédito vigentes en el país, existe un aprovechamiento subóptimo de los cupos disponibles, ya que un 64% de ellos permanece sin utilizar. Este escenario plantea una pregunta clave: ¿cómo pueden los bancos incentivar un mayor uso de las tarjetas de crédito entre sus clientes para incrementar su cartera y la rentabilidad? La solución propuesta a este problema se enfoca en desarrollar una segmentación de los clientes basado en los patrones de consumo mediante la implementación de algoritmos de aprendizaje no supervisado, de modo que se puedan identificar agrupaciones diferenciables y desarrollar estrategias de profundización personalizadas que no solo permitan aumentar el uso de las tarjetas, sino que también fortalecer la relación entre los clientes y las entidades financieras.

3. Revisión preliminar de antecedentes en la literatura

“Credit card customer segmentation and target marketing based on data mining” (Li et al., 2010):

En este artículo se aborda la implementación del algoritmo K-Means para segmentar a los clientes con base en el comportamiento del uso de tarjetas de crédito, teniendo en cuenta variables como la frecuencia de uso, el monto promedio por transacción, el límite o cupo de las tarjetas, historial de pagos y características demográficas. Previo a la implementación del algoritmo se realizó un proceso de estandarización (crucial en la implementación de K-Means) y se realizó un proceso de reducción de dimensionalidad mediante PCA. Como resultado se obtuvieron diferentes segmentos de clientes con características distintivas en su comportamiento financiero, que permiten a las entidades financieras no solo personalizar sus estrategias de marketing basándose en las necesidades de cada grupo, sino también identificar aquellos clientes con mayor riesgo de incumplimiento de pago basados en su comportamiento de gasto.

Barragán Garnica, D. (2022). Patrones de comportamiento de clientes con tarjetas de crédito de consumo con deterioro de calificación por riesgo utilizando K-Means:

Similar al artículo anterior, se pudo observar la implementación del algoritmo de K-Means para la agrupación de clientes con base en su comportamiento del uso de tarjetas de crédito. Sin embargo, esta investigación tiene un enfoque en la administración del riesgo crediticio, donde se tienen en cuenta otras variables como la calificación de riesgo y otras variables de riesgo créditicio para identificar aquellos grupos con mayor riesgo, de modo que se generen planes de acción para la administración del riesgo e incumplimientos focalizado a grupos de clientes específicos.

"Unsupervised Learning in Marketing: A Case Study of Financial Institutions." (Lee, S., & Kim, H.,2019):

Este artículo presenta un caso de estudio en el que un banco mediano en Asia utilizó modelos de aprendizaje no supervisado, específicamente algoritmos de clustering como DBSCAN, para mejorar la segmentación de clientes. La segmentación tradicional, basada en datos demográficos y transaccionales, no estaba produciendo campañas de marketing efectivas. Con el uso de aprendizaje no supervisado permitió al banco diseñar estrategias de marketing más alineadas con las necesidades de los clientes, demostrando la eficacia de estas técnicas en la segmentación de clientes en el sector financiero.

4. Descripción detallada de los datos

Los datos que se utilizarán contienen una amplia variedad de columnas que describen el comportamiento de compra de los clientes. Las variables incluyen características numéricas como el número de transacciones, el promedio, mínimo y máximo de las transacciones. Además, hay múltiples columnas categóricas y porcentuales que detallan el uso de diferentes franquicias de tarjetas (Visa, Mastercard, y otras), diferenciadas entre transacciones nacionales e internacionales, así como el porcentaje de transacciones realizadas en distintos momentos del día y días de la semana.

Las variables de tipo porcentaje están almacenadas como cadenas de texto, lo que sugiere que podrían haber sido formateadas para incluir un símbolo de porcentaje. Esto implica que puede ser necesario convertirlas a un tipo numérico para realizar un análisis cuantitativo preciso. La variable "Sitio_consumo_masfrecuente" identifica los lugares donde los clientes realizan la mayor parte de sus compras, lo que proporciona información adicional sobre sus preferencias. Finalmente, la columna "Compras_Totales", aunque representada como texto, indica el monto total gastado por cada cliente, un dato crucial para segmentar y analizar el valor de los clientes en función de sus patrones de consumo.

Variable	Tipo de Dato	Explicación
CLIENTE	int64	Identificador del cliente (anonimizado)
grupo_de_cliente	object	Clasificación del cliente en la segmentación del banco. Se desconocen los detalles.
Numero_de_transacciones	int64	Número de transacciones en el último mes
promedio_por_transaccion	float64	Promedio por transacción en el último mes
transaccion_minima	float64	Valor de transacción mínima en el último mes
transaccion_maxima	float64	Valor de transacción máxima en el último mes
desviacion_estandar_por_transaccion	float64	Desviación estándar del valor de las transacciones del último mes
porcentaje_visa_nacional	object	Porcentajes de uso de Visa en el último mes por consumo nacional
porcentaje_visa_internacional	object	Porcentajes de uso de Visa en el último mes por consumo internacional
porcentaje_mastercard_nacional	object	Porcentajes de uso de Mastercard en el último mes por consumo nacional
porcentaje_mastercard_internacional	object	Porcentajes de uso de Mastercard en el último mes por consumo internacional
Porcentaje_otrafranquicia_nacional	object	Porcentajes de uso de otra franquicia en el último mes por consumo nacional
porcentaje_otrafranquicia_internacional	object	Porcentajes de uso de otra franquicia en el último mes por consumo internacional
porcentaje_nacional_total	object	Porcentajes de uso en el último mes por consumo nacional
porcentaje_internacional_total	object	Porcentajes de uso en el último mes por consumo internacional
porcentaje_manana	object	Porcentajes de uso de tarjeta en el último mes por bloque del día: mañana (6-12 a.m.)
porcentaje_tarde	object	Porcentajes de uso de tarjeta en el último mes por bloque del día: tarde (12 a.m.-6 p.m.)
porcentaje_noche	object	Porcentajes de uso de tarjeta en el último mes por bloque del día: noche (6 p.m.-6 a.m.)
porcDOMINGO	object	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes
porcLUNES	object	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes
porcMARTES	object	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes
porcMIERCOLES	object	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes
porcJUEVES	object	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes
porcVIERNES	object	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes
porcSABADO	object	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes
Sitio_consumo_masfrecuente	object	Clasificación MCC del grupo de sitios de consumo más frecuente
Compras_Totales	object	Monto total de las compras realizadas por el cliente

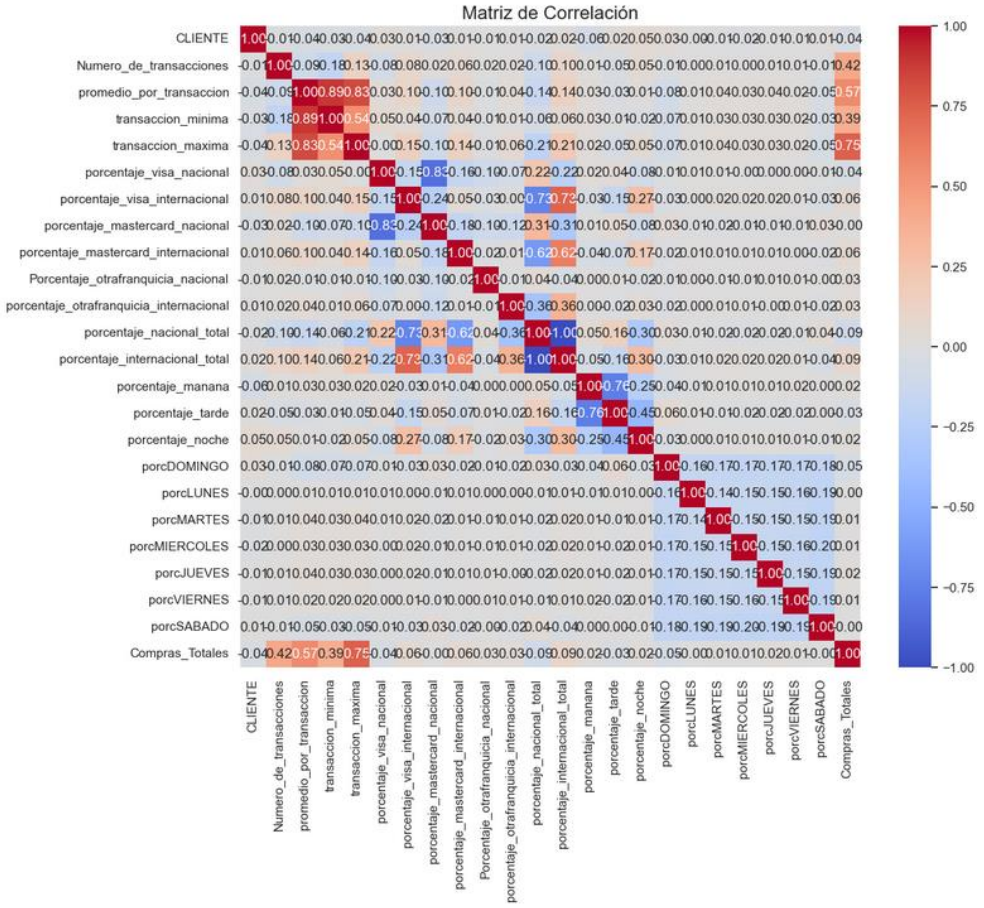
A continuación, se muestran las estadísticas descriptivas obtenidas

CLIENTE	grupo_de_cliente	Numero_de_transacciones	promedio_por_transaccion	transaccion_minima	transaccion_maxima
count	47.871.000.000	47871	47.871.000.000	4,79E+10	4,79E+10
unique	NaN	5	NaN	NaN	NaN
top	NaN	A	NaN	NaN	NaN
freq	NaN	42679	NaN	NaN	NaN
mean	23.936.000.000	NaN	5.083.161	3,72E+11	2,53E+11
std	13.819.311.705	NaN	8.483.558	5,80E+11	5,18E+11
min	1.000.000	NaN	1.000.000	1,00E+06	4,00E+04
25%	11.968.500.000	NaN	1.000.000	8,10E+10	3,07E+10
50%	23.936.000.000	NaN	2.000.000	1,67E+11	8,00E+10
75%	35.903.500.000	NaN	5.000.000	3,84E+11	2,18E+11
max	47.871.000.000	NaN	142.000.000	6,26E+12	6,15E+12

CLIENTE	porcentaje_visa_nacional	porcentaje_visa_internacional	porcentaje_mastercard_nacional	porcentaje_mastercard_internacional
count	4,79E+10	47.871.000.000	47.871.000.000	47.871.000.000
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN
mean	5,81E+11	37.487.507	3.645.922	53.932.185
std	8,85E+11	41.346.952	16.041.107	42.501.955
min	1,00E+06	0.000000	0.000000	0.000000
25%	1,14E+11	0.000000	0.000000	0.000000
50%	2,58E+11	20.000.000	0.000000	57.140.000
75%	6,27E+11	82.760.000	0.000000	100.000.000
max	1,10E+13	100.000.000	100.000.000	100.000.000

CLIENTE	porcDOMINGO	porcLUNES	porcMARTES	porcMIERCOLES	porcJUEVES	porcVIERNES	porcSABADO	Sitio_consumo_masfrecuente	Compras_Totales
count	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47871
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	109
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	SUPERMERCADOS / TIENDAS EXPRESS
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	9204
mean	12.820.821	13.932.558	13.173.200	13.274.230	13.947.279	13.622.064	14.152.377	17.898.014	NaN
std	26.446.157	28.016.771	26.533.006	26.583.531	27.222.813	26.860.637	27.316.936	30.778.624	NaN
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	NaN
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	NaN
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	NaN
75%	11.760.000	14.290.000	14.290.000	14.290.000	16.670.000	16.670.000	16.670.000	25.000.000	NaN
max	100.000.000	100.000.000	100.000.000	100.000.000	100.000.000	100.000.000	100.000.000	100.000.000	NaN

El análisis los datos muestra información sobre 47,871 clientes, con variables que incluyen el número de transacciones, el promedio por transacción, y los porcentajes asociados a diferentes tipos de tarjetas y franquicias. La columna Numero_de_transacciones varía entre 1 y 142, con una media de 5.08 transacciones. Los valores monetarios muestran una amplia gama, con un promedio de promedio_por_transaccion de 371,602.70 y un rango que va desde 1 hasta más de 6 millones. Los porcentajes de uso de diferentes tipos de tarjetas muestran que Visa Internacional y Mastercard Nacional son predominantes, aunque hay una amplia variabilidad. Los porcentajes de consumo por hora del día y días de la semana también presentan una alta dispersión, con una concentración notable en ciertos días y horarios. El sitio de consumo más frecuente es "SUPERMERCADOS / TIENDAS EXPRESS", y las compras totales oscilan entre 1 y casi 1,000.



La matriz de correlación muestra que las variables relacionadas con el tipo y la cantidad de transacciones tienen algunas correlaciones moderadas y fuertes entre sí, como entre promedio_por_transaccion, transaccion_minima, y transaccion_maxima. También revela que las variables de porcentaje de tarjetas tienen correlaciones significativas, como la fuerte correlación negativa entre porcentaje_visa_nacional y porcentaje_mastercard_nacional. Además, las variables de tiempo del día tienen correlaciones bajas entre ellas. En general, hay una variabilidad notable en las relaciones entre las diferentes variables.

5. Propuesta metodológica

Para abordar el problema de cómo incentivar un mayor uso de las tarjetas de crédito, se propone una metodología centrada en la segmentación de clientes basada en patrones de consumo. La finalidad es identificar grupos de clientes con comportamientos similares para diseñar estrategias personalizadas que aumenten el uso de las tarjetas y mejoren la rentabilidad.

Algoritmo Principal: K-medias (K-means)

K-medias es un algoritmo de clustering ampliamente utilizado que agrupa datos en k clusters distintos, minimizando la varianza dentro de cada grupo. Es efectivo para identificar patrones de consumo en grandes conjuntos de datos y se adapta bien a datos numéricos y continuos.

Razones para su Elección:

- Simplicidad y Eficiencia: K-medias es fácil de implementar y computationalmente eficiente, lo que lo hace adecuado para grandes volúmenes de datos típicos en el análisis de transacciones de tarjetas de crédito.
- Interpretabilidad: Los clusters generados por K-medias son fácilmente interpretables, facilitando la creación de estrategias personalizadas basadas en los perfiles de los segmentos identificados.
- Escalabilidad: K-medias maneja bien el aumento en el tamaño del dataset y es escalable para grandes volúmenes de datos, lo cual es crucial dado el tamaño potencial de las bases de datos de clientes.

¿Por qué K-medias?

- Contexto del Problema y Organización: Dado que el problema involucra grandes volúmenes de datos de transacciones de tarjetas de crédito, K-medias ofrece una solución eficiente y escalable. La capacidad de K-medias para manejar datos numéricos y su interpretabilidad son ventajas clave para diseñar estrategias basadas en los patrones de consumo de los clientes.
- Objetivo de Segmentación: La simplicidad y eficiencia de K-medias en la identificación de clusters bien definidos permiten desarrollar estrategias de marketing personalizadas basadas en características concretas de cada

grupo, alineándose con el objetivo de incrementar el uso de tarjetas y fortalecer la relación con los clientes.

Implementación y Evaluación de K-medias

- Preprocesamiento de Datos: Normalización de datos y manejo de valores faltantes.
- Determinación de K: Utilización del método del codo y análisis de la variación intra-cluster para seleccionar el número óptimo de clusters.
- Evaluación del Modelo: Evaluación de la cohesión de clusters y utilidad de los perfiles generados para las estrategias de marketing.

Si bien se tiene una propuesta inicial donde se evalúen los datos con el modelo K-medias, existen modelos que se podrían utilizar y aplicar para el caso de estudio.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise):
  - Descripción: DBSCAN identifica clusters de forma densa y puede manejar ruido y outliers, a diferencia de K-medias.
  - Consideraciones: Aunque DBSCAN es útil para detectar clusters de forma arbitraria y manejar outliers, puede ser menos adecuado para datos de alta dimensión o grandes volúmenes debido a su complejidad computacional.
- Algoritmos de Clustering Jerárquico:
  - Descripción: Los algoritmos jerárquicos construyen un árbol de clusters, permitiendo una visión detallada de las agrupaciones.
  - Consideraciones: Son útiles para comprender las jerarquías entre los clusters, pero pueden ser menos escalables que K-medias para grandes conjuntos de datos.

Bibliografía

Li, X., Wang, Y., & Liu, Y. (2010). *Credit card customer segmentation and target marketing based on data mining*.

Barragán Garnica, D. (2022). *Patrones de comportamiento de clientes con tarjetas de crédito de consumo con deterioro de calificación por riesgo utilizando K-Means*.

Lee, S., & Kim, H. (2019). *Unsupervised Learning in Marketing: A Case Study of Financial Institutions*.