Universidad de Los Andes

Maestría en Inteligencia Analítica de Datos

Proyecto Final Aprendizaje no supervisado

Autores:

Jose Daniel Garcia Correa

Cristian Mauricio Romero Buitrago

Cesar Daniel Ramírez Cely

Departamento de Ingeniería Industrial

2024

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 que genera a traves de intereses y comisiones, sino también porque permite evaluar la capacidad de pago de los clientes, lo que a su vez facilita el conocimiento de los clientes y el ofrecimiento de otros productos financieros.

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.

Aunque parece que las tarjetas de crédito son una herramienta común, su uso sigue siendo limitado, ya que el 64% del cupo aprobado en el sistema financiero no es utilizado. Esto corresponde a 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 aprobado de 5,5 billones de pesos está en uso, lo que deja un potencial de uso para los clientes de esta entidad por un monto aproximado de 3,8 billones.

El análisis del mercado sugiere que existe una oportunidad importante para incrementar la cartera en el uso de las tarjetas de crédito por parte de 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 plantea el uso del análisis de datos mediante la implementación de algoritmos no supervisados de clusterización, que permitan identificar grupos de clientes con comportamientos de consumo específicos, y así diseñar campañas que se adapten a las necesidades de cada grupo de clientes, logrando una mayor efectividad que al implementar una campaña general. Los segmentos o clusters 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 clustering como K-medias o K-Medoides, que permiten agrupar clientes según sus patrones de consumo; otro posible metodo a implementar es el clustering jerárquico, que puede revelar la estructura de los segmentos de clientes. 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 bajo uso 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 que tengan tarjetas de crédito vigentes, implementando algoritmos de aprendizaje no supervisado enfocados en el clustering, de modo que se puedan identificar agrupaciones o segmentos de clientes que sean similares en sus patrones de consumo, y asi 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 de procedimiento y de los datos obtenidos

El código implementado tiene como objetivo procesar y analizar los datos de comportamiento de clientes bancarios. Se inicia cargando un archivo CSV que contiene los datos a utilizar, seguido de la verificación de valores nulos y la visualización de las primeras filas del DataFrame. A continuación, se procederá a limpiar la columna "Compras_Totales", eliminando los puntos y espacios en blanco, y convirtiendo su contenido de texto a un formato numérico adecuado para permitir un análisis cuantitativo preciso. De igual manera, las columnas que contienen porcentajes (relacionadas con el uso de franquicias de tarjetas y horarios de compra) se transforman en valores numéricos de punto flotante.

Además, se generó una codificación para las variables categóricas "Sitio_consumo_masfrecuente" y "grupo_de_cliente" mediante la función factorize, lo que facilitará el análisis de estas categorías asignandoles un valor numérico. Finalmente, se obtendrán estadísticas descriptivas de las variables numéricas clave, como el número de transacciones y valores monetarios, calculando también la desviación estándar y el rango intercuartílico (IQR). Esto permitirá una comprensión más detallada de la variabilidad y dispersión en los patrones de compra de los clientes.

Los datos analizados reflejan el comportamiento de compra de 47,871 clientes a través de una variedad de columnas que incluyen características numéricas como el número de transacciones y valores monetarios (promedio, mínimo y máximo). También se disponen de columnas categóricas y porcentuales que describen el uso de diferentes franquicias de tarjetas (Visa, Mastercard, entre otras) en transacciones nacionales e internacionales, así como el porcentaje de compras en distintos momentos del día y días de la semana. Las variables de porcentaje, inicialmente en formato de texto, requieren conversión a tipo numérico para facilitar el análisis tal como se realizó en el proceso descrito. La columna "Sitio_consumo_masfrecuente" proporciona información sobre los lugares de compra, mientras que "Compras_Totales" refleja el gasto total de cada cliente, crucial para la segmentación del valor del cliente.

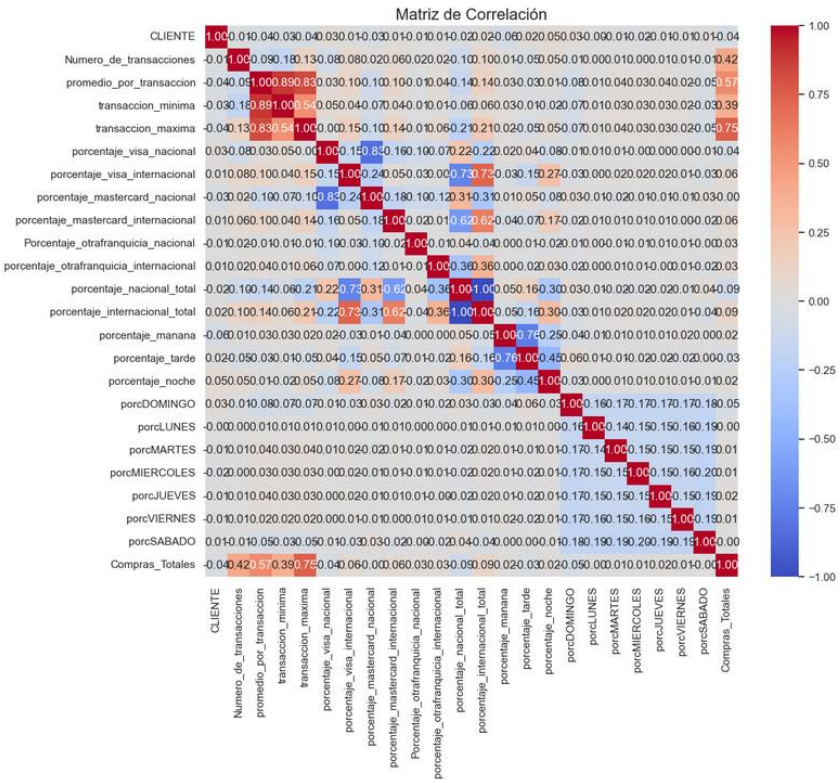
CLIENTE	porcDOMINGO	porcLUNES	porcMARTES	porcMIERCOLES	porcJUEVES	porcVIERNES	porcSABADO	Sitio_consumo_masfrecuente	Compras_Totales
count	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47.871.000.000	47871
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	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

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

El análisis revela que el número de transacciones por cliente oscila entre 1 y 142, con un promedio de 5.08, sugiriendo una heterogeneidad significativa en los patrones de compra. El promedio por transacción es de 371,602.69 COP, con un rango que va desde 0.04 COP hasta 11,040,000 COP, indicando variabilidad en el poder adquisitivo y preferencias de consumo. Las franquicias más utilizadas son Visa Nacional y Mastercard Nacional, posiblemente influenciadas por promociones en comercios específicos.

Adicionalmente, la matriz de correlación muestra relaciones moderadas y fuertes entre las variables transaccionales. En particular, se observa una alta correlación entre las variables de transacción mínima y máxima con la variable de promedio por transacción, así como entre los porcentajes de transacciones Visa/Mastercard nacionales e internacionales con los porcentajes totales nacionales/internacionales. Esto sugiere que las variables que agrupan los totales, así como los valores mínimos y máximos, pueden ser excluidas del modelo, lo que permitiría reducir la dimensionalidad sin perder información. A continuación, se presenta el detalle de las correlaciones entre las 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 teniendo en cuenta sus patrones de consumo. La finalidad es identificar grupos de clientes con comportamientos similares de gasto, con el fin de 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.

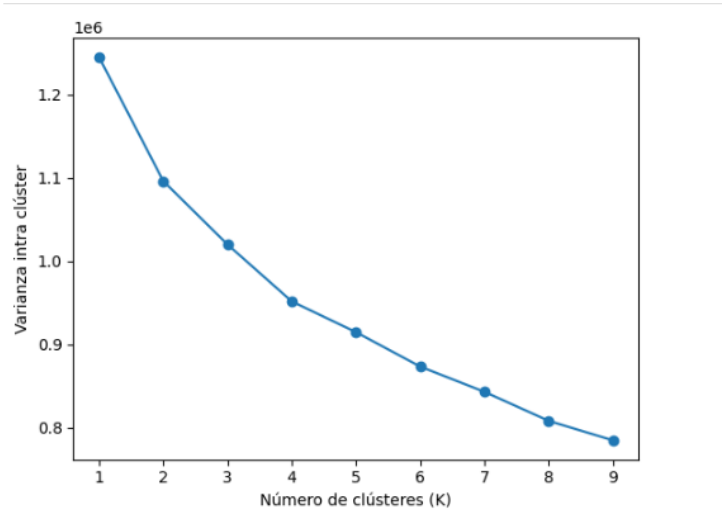
Razones para su Elección:

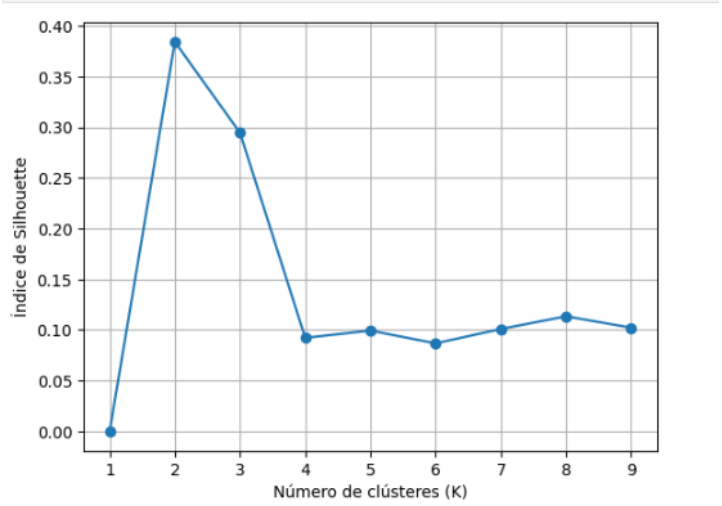
- Simplicidad y Eficiencia: K-medias es fácil de implementar y computationalmente eficiente, lo que lo hace adecuado para grandes volúmenes de datos típicos en el análisis de transacciones de tarjetas de crédito.
- Interpretabilidad: Los clusters generados por K-medias son fácilmente interpretables, facilitando la creación de estrategias 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.
- 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.

6. **Implementación y Evaluación de K-medias**

Para implementar el algoritmo de K-means, primero se estandarizaron los datos utilizando StandardScaler para evitar que las unidades de medida o la presencia de datos atípicos afectaran el resultado del modelo. Posteriormente se evaluó el número optimo de clusters, revisando la varianza intra cluster y el coeficiente de Silhouette.

Para la gráfica de la varianza intra cluster observamos un “codo” a partir de los 4 clusters, sin embargo, para este número de clusters el coeficiente de Silhouette es muy bajo, por esta razón se decide trabajar con 3 clusters, donde mantenemos un coeficiente de Silhouette alto y una varianza intra cluster similar al valor del “codo” encontrado. Los resultados se pueden observar a continuacion:

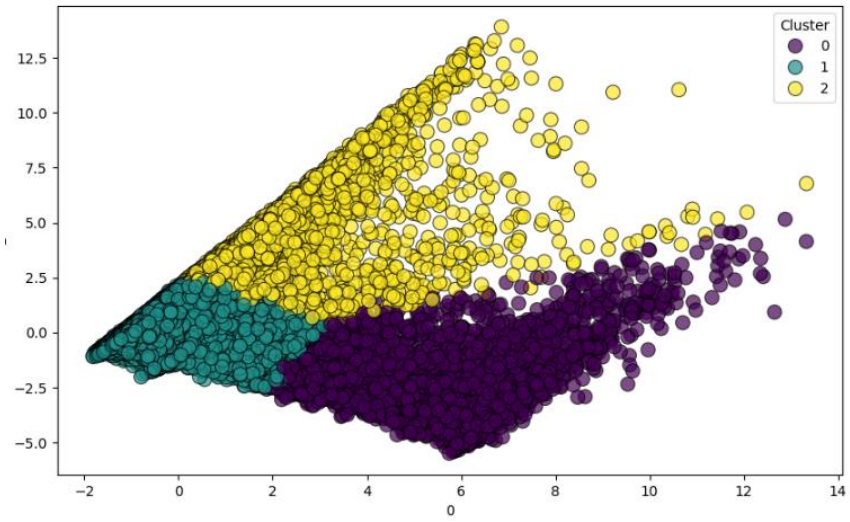




Para la selección del algoritmo se deciden realizar 3 implementaciones: K-Means contemplando todo el dataset, K-Means con reducción de dimensionalidad a través de PCA y DBSCAN. A continuación se presenta un resumen con los resultados obtenidos en cuanto a la varianza intra cluster y el coeficiente de silhouette:

Algoritmo	Varianza Intra cluster	Coeficiente de Silhouette
K-means	1019848,68	0,294
PCA K-Means	710601,77	0,345
DBSCAN	-	-0,0134

Teniendo en cuenta los resultados obtenidos, se selecciona el algoritmo de PCA K-Means, ya que presenta la varianza intracluster mas baja y el coeficiente de Silhouette más alto.



7. Descripción de clusters y estrategias recomendadas

Cluster 0: Este segmento se destaca por registrar el mayor promedio tanto en número de transacciones como en el monto de las mismas, lo que sugiere que son los clientes con mayor gasto en sus tarjetas de crédito, presumiblemente también con mayores ingresos ya que concentran la menor cantidad de clientes. Además, se caracterizan por realizar compras internacionales con diversas franquicias (no tiene preferencia por alguna franquicia) y mostrar una preferencia por realizar dichas compras durante la noche.

Cluster 1: Este segmento incluye a los clientes que, aunque realizan el menor número promedio de transacciones, presentan un alto valor promedio por operación, lo que indica que tienden a realizar compras de mayor valor. No muestran una preferencia específica por el horario de compra, ni volumen significativo en la compra internacional, pero sí presentan una clara inclinación por utilizar la franquicia Visa.

Cluster 2: Este segmento contiene a los clientes que realizan compras por menor monto con su tarjeta de crédito. Tienen preferencia por la franquicia MasterCard, no presentan volúmenes de compras internacionales y no evidencian un comportamiento fuerte de compra durante horas de la noche. Este Cluster contiene la mayor cantidad de clientes, lo que indica que se debe plantear una estrategia fuerte ya que son los que presentan menor usabilidad del cupo de su tarjeta de crédito.

Las estrategias a implementar se fundamentan en descuentos por compras en comercios aliados y aplicación de exoneraciones en la cuota de manejo, detalladamente para cada cluster serían las siguientes:

Cluster 0: Podría implementarse una campaña en la cual se otorgue una mejor tasa de cambio para las compras internacionales, esto con el fin de fidelizar a los clientes que hacen parte de este segmento, ya que demuestran un alto uso de sus tarjetas pero generaria un impacto significativo si optan por cambiar de banco.

Cluster 1: Para este cluster se plantea una estrategia de marketing en compañía con la franquicia Visa, puntualmente enfocada en descuentos con comercios aliados, estrategias de puntos o millas y beneficios de cobertura internacional, esto debido a su preferencia por esta franquicia, la oportunidad en el incremento del gasto de tarjeta de credito e incentivar el uso de compras internacionales.

Cluster 2: Para este segmento se plantea una estrategia de exoneración de cuotas de manejo a cambio de cierto nivel de gasto mensual con su tarjeta de crédito, ya que son los clientes que presentan menor uso de esta. Además, podrían pensarse estrategias en conjunto con la franquicia MasterCard para descuentos en comercios nacionales aliados, de acuerdo con la preferencia identificada.

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*.