Segmentación de clientes basado en consumo bancario

Métodos y aplicaciones de analítica I

Fabián Alexis Pallares Jaimes

Entendimiento del negocio

Contexto de negocio: Con motivo de su nuevo sistema de gestión de beneficios basado en analítica de clientes, el Banco X planea encontrar promociones que permitan a los clientes disfrutar de sus actividades preferidas de compra con el uso de tarjetas y garantizar un mayor beneficio económico al incentivar el uso de estas.

El Banco X solicita el presente informe para observar segmentaciones naturales en sus clientes con motivo de generar una toma de decisiones basada en datos para el desarrollo de promociones diferenciadas por el comportamiento de dichos segmentos.

Para esto, el Banco X provee una base de datos de clientes en formato xlsx con información de transacciones anonimizadas, de acuerdo con la ley de protección de datos personales, de 47.871 de sus clientes, correspondientes a aquellos que han adquirido/usado sus tarjetas. Cabe mencionar que los tarjetahabientes del Banco X pueden adquirir tarjetas para uso tanto nacional como internacional de las franquicias VISA, Mastercard, y otra provista por el banco.

Objetivos del negocio y criterios de éxito

El banco X planea obtener una segmentación de clientes que permita representar a menos 5 grupos de clientes a los cuales enfocar las diferentes promociones, así como las características principales de los mismos, en especial en lo correspondiente a:

- Tipo de tarjeta de uso (Nacional o internacional)
- Franquicia de preferencia
- Consumo promedio: categorizado en bajo, medio y alto

Para una mayor confianza en el éxito de las promociones a lanzar, el Banco solicita que el modelo de segmentación de clientes sea debidamente evaluado y presente al menos 5 *clusters*.

Exploración y verificación de calidad de los datos

La base de datos que se usará para este análisis corresponde a los datos transaccionales de los clientes poseedores de tarjetas del Banco X en el mes de agosto de 2020.

El banco recopiló información de valores por transacción, así como el promedio, mínimo, máximo y desviación estándar de estos valores para cada cliente. Asimismo, se tienen los datos del porcentaje de uso de las tarjetas (VISA, Mastercard, otra), así como su uso en porcentaje por días y momento del día.

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

Tabla 1. Descripción y tipo de los datos

Como podemos observar en la tabla 1, nos encontramos con 26 variables en la base de datos, de las cuales tenemos dos categóricas, correspondientes a una clasificación no detallada por el banco y al sitio de consumo más frecuente por el cliente. Ambas variables categóricas son prescindibles para la segmentación, ya que se trabajará con variables numéricas, las cuales corresponden a las otras 24 variables del set de datos.

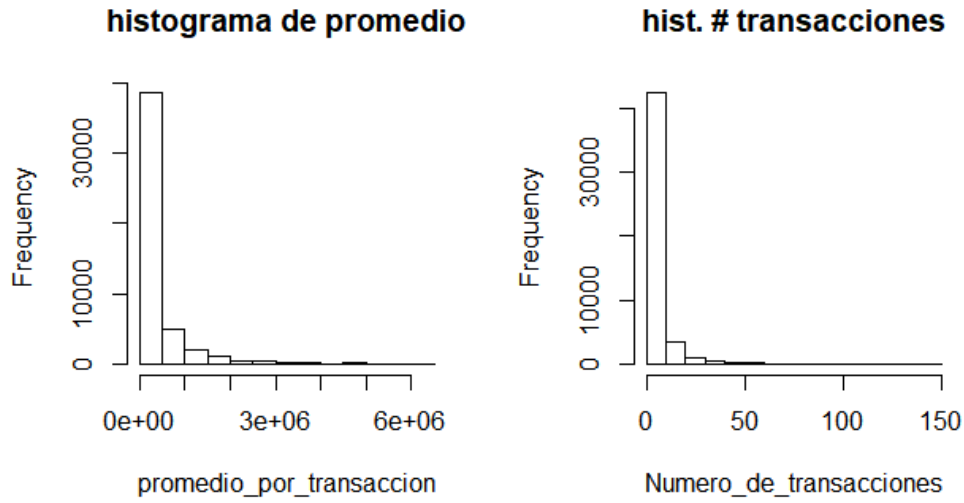
Estadísticamente, podemos observar algunas variables clave como el promedio gastado por transacción, el número de transacciones por cliente y los porcentajes totales de uso

nacional e internacional de los clientes para conocer, a grandes rasgos, su comportamiento de consumo.

Variable	Min	Max	Media	Mediana	Asimetría	Curtosis
Total de transacciones	1	142	5.8	2	5.58	47.21
Promedio por transacción	1	6262025	371602.69	167173	3.79	18.97
Total porcentaje consumo nacional	0	1	0.93	1	-3.28	9.55
Total porcentaje consumo internacional	0	1	0.07	0	3.28	9.55

Tabla 2. Algunas medidas estadísticas de variables de interés

Tal como se ve en la tabla 2, presentamos una asimetría considerable (>1) en nuestras variables de interés a la hora de conocer el comportamiento de los clientes. Podemos observar, gracias a la mediana, que más de la mitad de los clientes del Banco X realizaron únicamente 2 o menos transacciones durante el mes de agosto, valor que se contrasta con una media superior de 5.8 transacciones. Sobre el total de transacciones también podemos decir que puede presentar valores atípicos debido al alto valor de su curtosis.

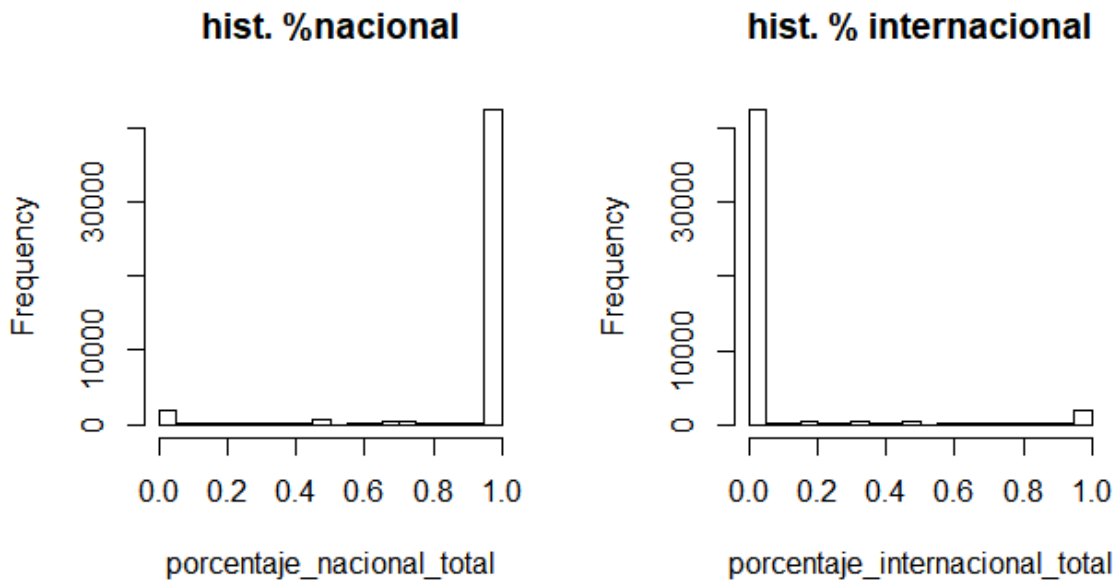


Figuras 1 y 2. Histogramas de promedio de valor de transacciones por cliente (izquierda) y número de transacciones por cliente (derecha)

En lo que respecta al promedio por transacción, podemos observar el mismo fenómeno de una mediana menor a la media y una curtosis alta, lo que puede indicarnos valores que, de manera atípica, están afectando el resultado de la media.

Estos fenómenos de concentración de los clientes al extremo izquierdo (asimetría positiva) podemos comprobarlo gracias a las gráficas presentadas en las figuras 1 y 2, donde vemos que una gran parte de los clientes no realizaron ninguna transacción en el mes de agosto (figura 2), lo cual también aporta a la asímetría del promedio, siendo que su promedio por transacciones para a ser cero (figura 1).

En el contraste de porcentaje de consumo nacional e internacional, podemos ver que los tarjetahabientes del Banco X están concentrados en el consumo nacional, siendo este representado en promedio por el 93% del consumo en general, llegando además a tener una mediana de 1. Para el caso del consumo internacional, en cambio, solo representan un consumo del 7%, siendo además indicado por la mediana, que más de la mitad de los clientes no presentan ningún consumo de tipo internacional. Para las variables de porcentaje de consumo presentamos también una curtosis alta.



Figuras 3 y 4. Histogramas del porcentaje de uso total de tarjetas nacionales (izquierda) e internacionales (derecha)

En este caso, vemos un comportamiento que verifica lo mencionado en el párrafo anterior. En donde la concentración en el uso nacional se da a la derecha (asimetría negativa), y en el uso internacional se presenta a la izquierda (asimetría positiva).

En lo que respecta a la verificación de calidad de los datos, se encontró que ninguna de las 26 variables contiene *missing values.*

Sin embargo, sí es posible que se presenten datos atípicos, tal como se vio en las variables de interés descritas anteriormente. El tratamiento de datos atípicos se tendrá en consideración en la fase de preparación de los datos para la construcción de la vista minable.

Preparación de la vista minable

Lo primero a realizar para obtener una vista minable será generar nuevas variables, las cuales se consideraron partiendo de agrupaciones (generalmente sumas) de las variables existentes:

- Porcentaje_visa = suma de los porcentajes de las variables que representan el consumo nacional e internacional de tarjetas de la franquicia VISA
- Porcentaje_mc = suma de los porcentajes de las variables que representan el consumo nacional e internacional de tarjetas de la franquicia MasterCard
- Porcentaje_otra = suma de los porcentajes de las variables que representan el consumo nacional e internacional de tarjetas de la otra franquicia manejada por el banco
- Porcentaje_fds = suma de los porcentajes de las variables que representan el porcentaje de consumo de los días viernes, sábado y domingo
- Porcentaje_seman = suma de los porcentajes de las variables que representan el porcentaje de consumo de los días lunes, martes, miércoles y jueves.

Asimismo, se decidió omitir variables que, de acuerdo con el contexto de negocio utilizado, no aportan a nuestro modelo de segmentación:

- CLIENTE
- grupo_de_cliente
- transacción_minima
- transaccion_maxima
- desviacion_estandar_por_transaccion
- Sitio_consumo_masfrecuente

Para la reducción de los datos atípicos, se realiza de nuevo una observación a los valores de curtosis de las variables, siendo que las variables con curtosis altas requieren de un tratamiento de atípicos.

Para reducir el sesgo generado por los datos atípicos aplicamos un logaritmo natural de (1+x), el cual nos sirve en este caso para la reducción ya que tiene en cuenta los ceros o valores menores que cero.

Finalmente, estandarizamos los datos para tener una vista minable que nos permita una correcta ejecución del modelo.

Pruebas iniciales para el desarrollo del modelo – selección de K

Para encontrar el k adecuado en el modelo a desarrollar, es posible guiarse por distintos criterios. En este caso se piensa analizar este número con los criterios de suma de cuadrados, criterio de Silhouette y la estadística GAP.

Criterio de suma de cuadrados

Para el análisis en este método se usa un método gráfico, que consiste en analizar un cambio de codo en la pendiente, siendo que en esos cambios las reducciones de cuadrados internos comienzan a ser mínimas y puede escogerse tal valor como el número de clusters o segmentos.

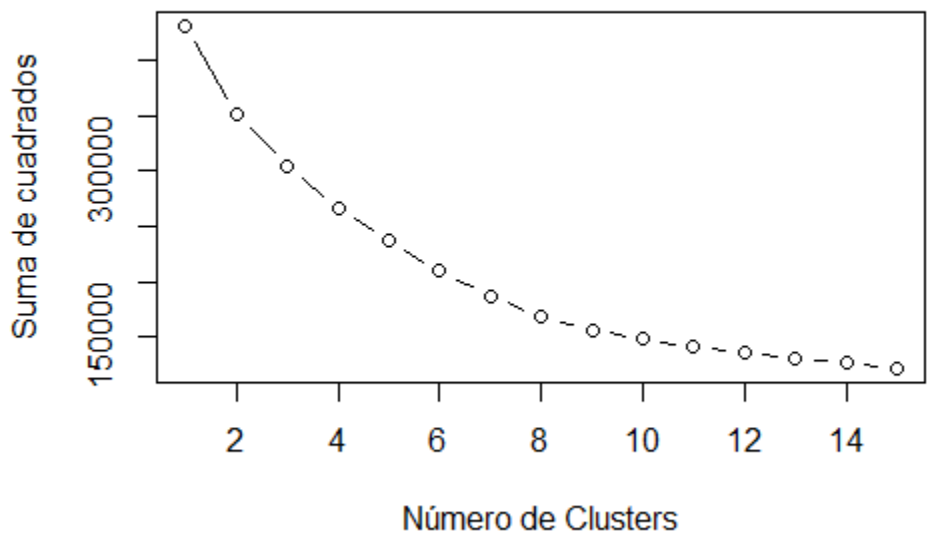


Figura 5. Criterio de suma de cuadrados

En este caso, mediante el método de suma de cuadrados, podemos ver un cambio en la pendiente *entre 7 y 9 clusters*. Sin embargo, este valor no es exacto y difícilmente concluyente.

Criterio de Silhouette

Nos habla de cuánto es la perdida al mover un registro de un cluster a otro. Para este criterio puede ser necesario realizar un muestreo de acuerdo con la cantidad de datos a procesar.

# Cluster	Criterio Silhouette
2 clusters	0.192
3 clusters	0.191
4 clusters	0.205
5 clusters	0.225
6 clusters	0.231
7 clusters	0.261
8 clusters	0.280
9 clusters	0.289
10 clusters	0.296
11 clusters	0.293

Tabla 3. Criterio Silhouette

Sugerencia: *10 Clusters*

Criterio de estadística de GAP

Por último, probamos con el criterio de estadística de GAP, en el cual, al igual que en el criterio de Silhouette, se tuvo que hacer un muestreo para realizar la ejecución.

Sugerencia de GAP: *7 Clusters*

Modelado

Realizamos un modelo de segmentación (*clustering*) escogiendo 7 como k (número de *clusters*).

El modelo fue realizado gracias a la función kmeans, puesto que la segmentación, para este caso, es desarrollada con el proceso k-medias.

$$bcluster < -kmeans(banca, centers = 7, nstart = 10, iter.max = 20)$$

$$bcluster\$size$$

$$bcluster\$iter$$

$$bcluster\$centers$$

Script 1. Cálculo de *Clusters*

Tamaño de los Clusters:

- *Cluster 1:* 9991
- *Cluster 2:* 5505
- *Cluster 3:* 8607
- *Cluster 4:* 3777
- *Cluster 5:* 9023
- *Cluster 6:* 6522
- *Cluster 7:* 4446

Iteraciones requeridas: 4

Heatmap de los grupos:

Figura 6. Heatmap de Clusters

- **Descripción grupo 1:** Se caracteriza por un alto número de transacciones en el mes, el uso de tarjeta de la franquicia mastercard y transacciones de bajas cantidades.

- **Descripción grupo 2:** Se caracteriza por su horario de compras, siendo que las realiza comúnmente en las tardes, su medio de compras principal es la tarjeta de la franquicia VISA y realiza sus compras los fines de semana.

- **Descripción grupo 3:** Caracterizado por el uso de la tarjeta de franquicia VISA y compras comúnmente entre semana.

- **Descripción grupo 4:** Grupo de clientes de compras internacionales, usa comúnmente su tarjeta de franquicia VISA y tiene un alto número tanto de transacciones como de cantidades movidas en estas.

- **Descripción grupo 5:** Caracterizado por compras con tarjeta MasterCard, compras comúnmente entre semana.

- **Descripción grupo 6:** Caracterizado por el uso de tarjetas MasterCard y compras los fines de semana y en la tarde.

- **Descripción grupo 7:** Compras comúnmente en la mañana y los fines de semana.

Propuestas y recomendaciones de negocio

Es recomendable fijarse en las variables que más cambian en los grupos y generar promociones respecto a ellas, publicitándolas a los diferentes segmentos via correo electrónico.

- A los grupos 2, 3 y 4 es recomendable generar promociones por el uso continuo de las tarjetas VISA. Esto puede ser mediante un sistema de puntos basado en porcentajes del monto de la transacción. Esto ayuda a incentivar el uso de la tarjeta y las compras con un mayor monto.

- Del mismo modo, se puede realizar algo similar con el grupo 1, el cual, a pesar de tener un muy buen uso de su tarjeta MasterCard, realiza compras de pocas cantidades, cosa que se puede incentivar con beneficios por compras superiores a $150.000, siendo este un valor cercano a la mediana del promedio por transacción.

- Los grupos 5, 6 y 7 se caracterizan por sus horarios y días de compras, por lo cual puede ser posible crear promociones (por ejemplo, devolver un porcentaje del dinero por compra, para compras mayores a $100.000) en los días u horarios que más funcionen para cada tipo de cliente:
    - Entre semana y en la tarde para el grupo 5
    - Fines de semana y en la tarde para el grupo 6
    - Fines de semana y en la mañana para el grupo 7

  Estas promociones se pueden aplicar para tarjetas MasterCard en el caso de los grupos 5 y 6; y sin distinción por tarjeta para el grupo 7.