TRABAJO 2

INFORME DE ANÁLISIS NO SUPERVISADO

ESTIVEN GUTIÉRREZ
ALEJANDRO HERRERA
JUAN JOSE ISAZA
CAMILO LOAIZA

Introducción:

Hoy en día las decisiones de los clientes son más complejas y cambiantes, por lo tanto, realizar un seguimiento más preciso de aquellos gustos es de suma importancia para los bancos, si quieren plantear estrategias que contribuyan a que sus clientes se sientan satisfechos con los productos y servicios que estos ofrecen.

El conjunto de datos que se escogió para el proyecto presenta el comportamiento de uso de cerca de 9000 titulares de tarjetas de crédito, los cuales estuvieron activos durante los últimos 6 meses. La base de datos contiene 17 variables de comportamiento de los clientes. Lo que se busca en este trabajo es desarrollar una segmentación de clientes para definir la estrategia de marketing debida. En otras palabras, segmentar a los clientes es construir una ventaja competitiva ya que las empresas se dirigen hacia donde se encuentran sus clientes más valiosos, por eso los bancos tratan de asignar un capital y esfuerzo importante para generar la mayor cantidad de beneficios. Algunas preguntas claves a la hora de desarrollar este trabajo son:

¿Qué tan crucial se considera la segmentación de los clientes para el proyecto?
¿Cuáles variables tienen mayor influencia y por qué se puede considerar lógica?

Exploración de los datos

Inicialmente se cuenta con 17 variables, se decide eliminar la variable Id puesto que no brindaba información importante para el desarrollo del trabajo, quedado así, con las siguientes variables: 'saldo', 'frecuenciaactsaldo', 'comprastotales', 'montomaxcomprado', 'montoacuotas', 'anticipoefectivo', 'frecuenciacompras', 'freccomprasunavez', 'freccomprasplazo', 'frecpagoantefectivo', 'ntranscashinadv', 'ntransacciones', 'limtarjetacredito', 'pagos', 'montominpagos', '%totalpagado', 'antiguedad'.

Se continúa con una exploración acerca de la cantidad de nulos que se puedan presentar en cada una de las variables. Se logran observar dos variables con nulos, 'montominpagos' con 313 y 'limtarjetacredito' con tan solo una. Se procede a rellenar estos datos con la mediana, luego de haber visto el comportamiento que tenían los datos de estas variables por medio de un histograma y de un boxplot, y observar en este gran número de datos atípicos, por lo que se recomienda rellenarse de esta forma.

Luego se procede a realizar una prueba de normalidad para todas las variables. Esta se realizó con la shapiro_test, en donde se logra evidenciar una distribución normal en todas las variables, lo que indica qué tipos de modelo se pueden incursionar en el trabajo como por ejemplo el de Mezclas gaussianas.

En cuanto a la correlación de las variables, claramente se pudo observar alta correlación entre las variables presentes, como 'comprastotales' y 'montomaxcomprado', también entre 'frecuenciacompras' y 'freccomprasplazo'. Sin embargo, por términos prácticos se decide no eliminar dichas variables, aunque puedan llegar a presentar problemas de multicolinealidad en el trabajo, esto sobre todo porque en los primeros modelos ejecutados con las variables eliminadas, se pudo observar un agrupamiento de los datos que consideramos inconvenientes.

En el preprocesamiento se realiza un escalado de variables, puesto que los valores de los datos podrían presentar problemas a la hora de ser trabajados en su escala original. Luego de esto se comienza a analizar el comportamiento de los datos atípicos mediante gráficos como el histograma y la caja de bigotes, en donde en la mayoría de variables se logra observar un gran número de outliers. Por lo tanto, se procede a la eliminación de estos asignando un valor de threshold que es determinado por nosotros de acuerdo con la dispersión de los datos que se logra ver en el boxplot de cada una de las variables, el threshold que más se asignaba era de un valor de 3.5 el RIC, en donde se lograba en muchos casos eliminar más de 100 datos en total. La variable en la que más se eliminaron atípicos fue

'intranscashinadv' con un total de 199 datos eliminados, está también con un threshold asignado de 3.5 el RIC. También el threshold más alto fue el de 7 veces el RIC que fue asignado a la variable '%totalpagado'. En la depuración de los datos atípicos no se eligen threshold de 2 veces el RIC que suelen ser los más recomendados, puesto que eliminaban muchos datos y la idea era conservar buena cantidad de datos para poder dar con modelos más dicientes.

Reducción de la dimensionalidad

Se realizaron dos reducciones, primero se redujo a una varianza explicada del 80% y finalmente una varianza alrededor del 60% en solo tres componentes, como se obtuvieron mejores resultados con la última reducción de acá para adelante se trabajará solo con esa base. La varianza explicada por los 3 componentes principales son para la componente 1 el 29,31 %, para la 2 el 17,74% y para la 3 el 12,15%, con un total de 60% de varianza explicada.

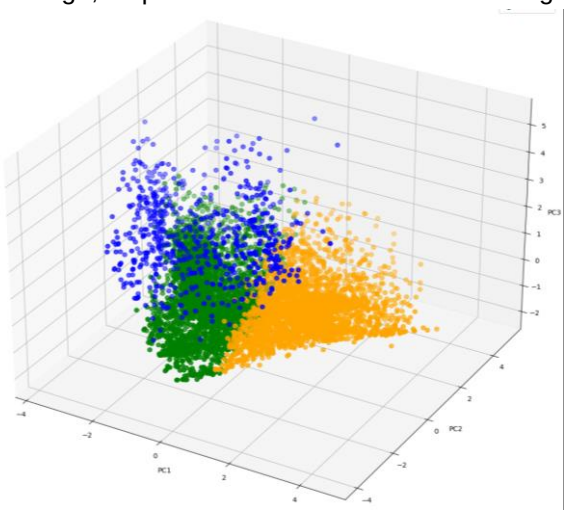
Modelos construidos

Se construyeron los siguientes modelos con las siguientes consideraciones, se realizó reducción de la dimensionalidad con una varianza explicada del 80%, además se realizó la reducción a solo 3 componentes que explicaban alrededor del 60% de la varianza, finalmente se obtuvieron 3 bases de datos, la original, la PCA al 80% y PCA con 3 componentes. De esta forma se construyó 4 modelos de K Means con las consideraciones anteriores, el último consideraba las 3 componentes, pero con un clúster de más, finalmente 4 modelos de gaussian mixture en donde el primero era con el PCA al 80% y los demás con PCA con 3 componentes variando el hiper parámetro del tipo de matriz de covarianza. Los resultados fueron los siguientes:

Modelo	Inertia	Silhouette Score	Calinski harabasz
Kmeans - Modelo base	32540,3	0,237	1622,91
Kmeans - PCA 80%	26927,84	0,296	2150,81
Kmeans - PCA (3 componentes - 3 clusters)	19862,29	0,397	3828,01
Kmeans - PCA (3 componentes - 4 clusters)	31699,27	0,243	2065,39
Gaussian mixture - PCA 80%	-	0,043	911,1
Gaussian mixture - PCA (3 componentes) full	-	0,115	1893,81
Gaussian mixture - PCA (3 componentes) tied	-	0,339	3471,86
Gaussian mixture - PCA (3 componentes) diag	-	0,289	3053,93

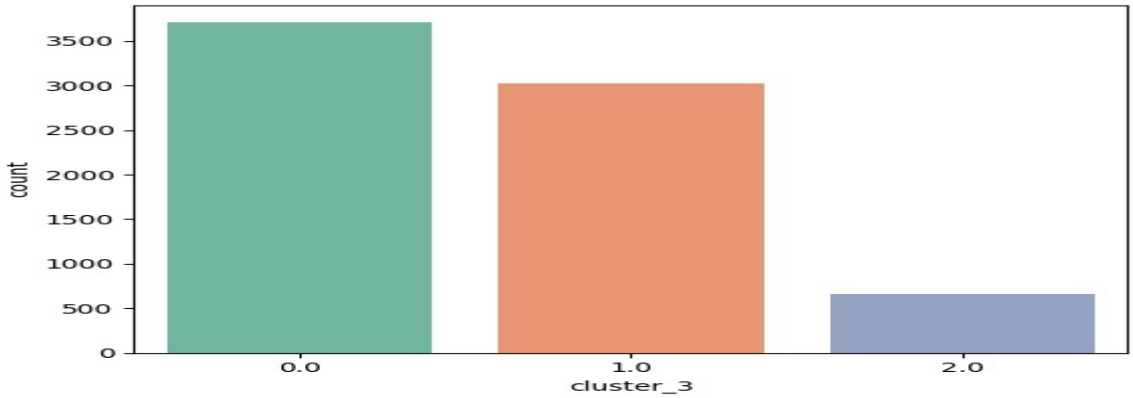
Tabla 1: Resumen de resultados

Luego del análisis se concluye que el mejor modelo construido es el de k-means con 3 componentes principales generados a partir del método de PCA, el segundo mejor modelo es el de mezclas gaussianas con 3 componentes y con la matriz de covarianza tipo tied. Es importante recordar que se pueden mejorar los modelos mejorando y optimizando algunos hiperparametros. Por lo anterior se escogió el modelo de k means descrito anteriormente, además se graficaron sus resultados del modelo en la figura, donde se puede observar que el clúster azul se traslapa en mayor medida con los demás clúster, sin embargo, se pueden identificar los tres clúster generados.



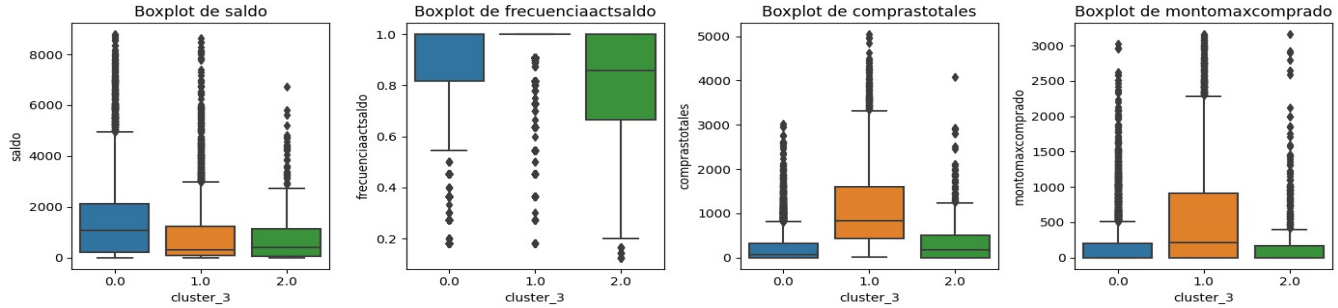
Gráfica 1: Distribución de los clústeres PCA con 3 componentes

Análisis de resultados



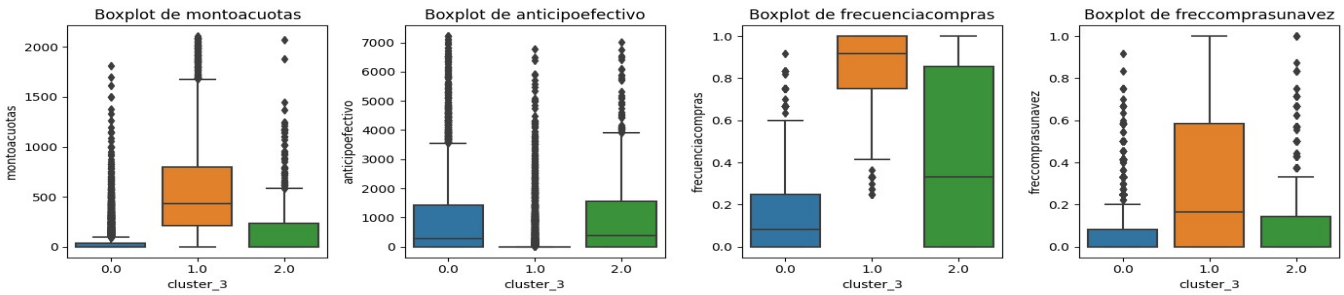
Gráfica 2: Gráfico de observaciones de clústeres

Los clústeres tienen observaciones diferentes, el clúster cero es el que mayor tiene observaciones, luego el uno y por último el clúster dos.



Gráfica 3: Boxplot variables saldo, frecuencia saldo, compras totales, monto máximo.

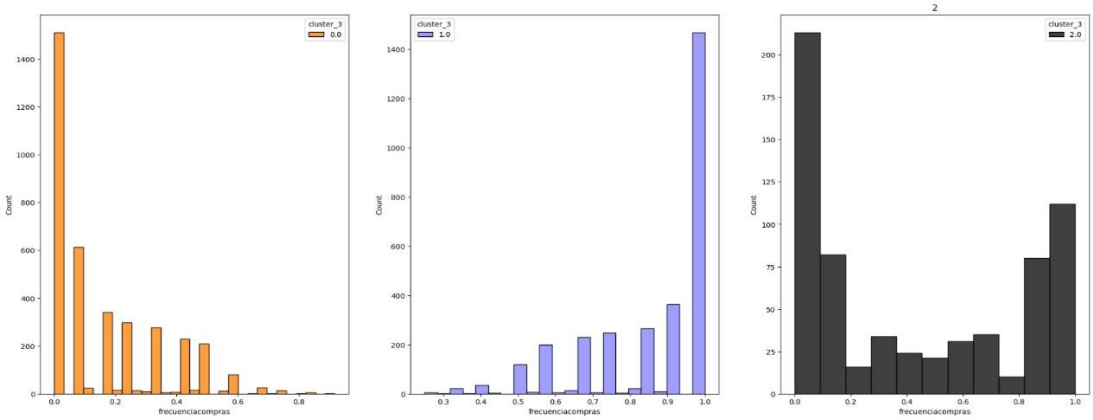
El clúster cero son los titulares que más saldo tienen en la cuenta, en los otros dos clúster el saldo se mantiene con cantidad promedio similar, los titulares del clúster 2 son a los que con menos frecuencia se les actualiza el saldo, luego son los del clúster cero y finalmente los del clúster uno, los titulares del clúster uno son los que han comprado con un mayor monto en la tarjeta de crédito, luego el clúster dos y finalmente el clúster cero, esto tiene sentido ya que a mayor número de compras se actualiza con más frecuencia el saldo. Finalmente, los titulares del clúster uno ha gastado un monto mayor en una sola compra a comparación de los otros dos clústeres con un promedio de 564 dólares, los otros dos clústeres el monto máximo comprado en una sola compra es muy similar, 199 dólares para clúster dos y 188 dólares para clúster dos.



Gráfica 4: Boxplot variables monto, anticipo, compras y compras una vez

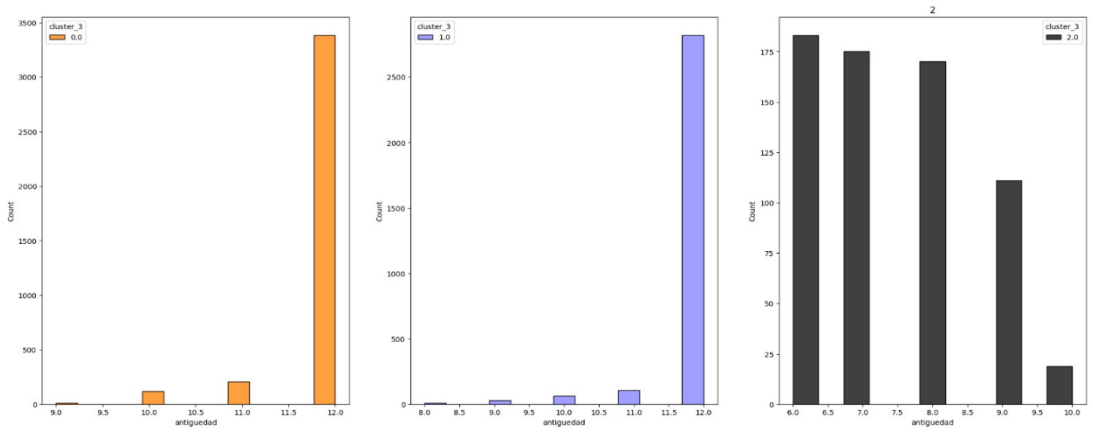
Los titulares del clúster uno son los que realizan más compras que se pagan con cuotas que los titulares del resto de clústeres, además los titulares del clúster dos realizan más compras que se pagan a cuotas que los titulares del clúster cero, por otro parte los titulares del clúster dos y cero han realizado un anticipo de efectivo similar alrededor de los 1000 dólares, los titulares del clúster uno son los que han anticipado menor dinero. Los titulares del clúster uno son los que compran frecuentemente con la tarjeta, luego los del clúster dos y finalmente los del clúster cero, además los titulares

del clúster uno se realiza con mayor frecuencia compras de un solo pago que el resto de clúster, los otros titulares de los clústeres tienen una menor frecuencia de compra de un solo pago.



Gráfica 5: Análisis de la variable de frecuencia de compras

Los titulares del clúster uno son los que realizan frecuentemente mayores compras, luego los del clúster dos y finalmente los del clúster cero. En el clúster cero existe gran variedad de clientes con diferentes frecuencias en compras, los titulares del clúster cero son los que compran con menor frecuencia.



Gráfica 6: Análisis de la variable antigüedad

Los titulares del clúster uno y cero tienen similitud en el comportamiento de la antigüedad de sus titulares, en mayor medida para ambos sus titulares tienen más de 10 años con la tarjeta. En el clúster dos sus titulares en su mayoría tienen menos de 10 años con la tarjeta y en mayor cantidad se encuentran entre 6 y 6.5 años.

Resumen

Clúster 0: mayor anticipo de efectivo, poca utilización de la tarjeta, poco pago.

Estos usuarios llevan muchos años con la tarjeta de crédito, pero casi no realizan compras, estos utilizan su tarjeta en mayor medida para realizar avances de dinero, además, no les gusta endeudarse con la tarjeta, pero tampoco tienen menor porcentaje de la deuda pagada por lo que tienen un monto mínimo a pagar más alto. Además, tienen buen límite en la tarjeta de crédito, lo que demuestra confianza por parte del banco. Por último, casi no compran a cuotas y tampoco de un solo pago.

Clúster 1: utilización mayor en compras en crédito y de un solo pago, poco anticipo de dinero.
Este tipo de titulares son los que más compras realizan con la tarjeta de crédito y por lo tanto tienen un saldo intermedio en sus tarjetas de créditos frente a los demás clústeres, ya que tienen un límite mayor en sus tarjetas de créditos, además, tiene una alta antigüedad lo que genera confianza al prestador de la tarjeta de crédito. Estos tienen un mayor porcentaje de su deuda pagada y por lo tanto tienen un monto mínimo de pago bajo, además, estos titulares son los que con un mayor monto y frecuentemente sacan sus compras a plazo, pero pocas veces realizan anticipo de efectivo.

Clúster 2: Menor capacidad adquisitiva, jóvenes para la empresa, moderadamente utilización de anticipo de dinero en efectivo.

Este tipo de titulares son los que menos saldo tiene en sus tarjetas, ya que tienen un límite de saldo menor y realizan compras de forma moderada, por lo tanto, el monto mínimo de pago requerido es menor y sacan sus compras a cuotas de forma mesurada. Estos titulares llevan menos tiempo con su tarjeta de crédito, con aproximadamente 7.8 años. Por último, son los titulares que anticipan moderadamente dinero en efectivo.

Estrategias de marketing

1. Para el clúster cero: A los titulares de este clúster se debe hacer campañas para estimular el uso de la tarjeta de crédito para realizar compras, además estudiar el impacto del uso intensivo de los anticipos de dinero para la empresa, por último, educación financiera que permita enseñar a realizar pagos adecuados y en el momento preciso.
2. Para el clúster uno: A estos titulares se les debe ofrecer una tarjeta premium que contenga servicios exclusivos y cupos más altos, ya que realizan muchas compras con su tarjeta y son buenos a la hora de abonar a su deuda. Finalmente se deben fidelizar a la empresa.
3. Para el clúster dos: A estos titulares se les debe de aumentar el límite del saldo para que puedan aumentar sus créditos, bajar las tasas de interés para que puedan realizar más compras a cuotas. Por último, educación financiera en especial sobre el uso de las tarjetas de créditos.

Conclusiones

El banco prestador del servicio de la tarjeta de crédito ahora podrá ofrecer nuevos servicios únicos para cada tipo de segmentos de clientes encontrados, satisfaciendo de esta forma a los clientes y generando beneficios para la empresa. Eliminar variables correlacionadas causó en el modelo construido a partir de los componentes principales determinados por el método PCA que disminuyeran sus métricas drásticamente, por lo tanto, se consideró dejar todas las variables. Los modelos de clustering construidos fueron sensibles a los datos atípicos, por lo tanto, al variar los criterios de eliminación de datos atípicos se modificaban en gran medida las métricas y los clustering generados.

Los clústeres generados por el modelo de k means el cual es uno de los modelos funcionales “simples” fueron diferentes y significativos, lo que permitió reconocer las características de cada grupo.

Se recomienda buscar nuevos hiperparametros, limpieza de los datos y nuevas exploraciones que permitan mejorar los resultados de los modelos.