Imagen que contiene Texto

Descripción generada automáticamente

Modelo de marketing para entidad bancaria con metodología crisp-dm

Por:

Daniel Esteban González Zuluaga

Profesor a cargo:

Jorge Andrés Alvarado Valencia

Pontificia Universidad Javeriana

Maestría de Analítica para la Inteligencia de Negocios

Métodos y aplicaciones de analítica1

2024

**1. Business Understanding**

**1.1 Resumen Ejecutivo**

El banco “Daniel González” es destacado como una de las principales entidades financieras del país, clasificándose como uno de los 10 mejores a nivel nacional. Con una amplia gama de productos financieros y así buscando fidelizar cada vez más a sus clientes y poder catalogarse como el mejor banco nacional.

En este momento se ha llevado un análisis exhaustivo en el manejo de datos efectuando consigo una base de datos enfocada a las transacciones bancarias por crédito de los clientes en el mes pasado, proporcionando una visión completa de las actividades financieras.

Este análisis de datos y modelamiento se utilizará para diseñar estrategias de marketing que se alineen con las necesidades y comportamientos de los clientes, con el objetivo de aumentar su fidelidad. Además, ayudará a la empresa a comprender mejor su desempeño actual y a tomar decisiones informadas basadas en datos.

**1.2 Background**

En el ámbito bancario, entender los factores externos e internos es clave. La estabilidad política, las regulaciones gubernamentales y la tecnología son aspectos que influyen en las oportunidades y desafíos del sector. La competencia entre bancos, tanto locales como internacionales, es intensa, llevando a una lucha por clientes y recursos de mercado. Los bancos buscan destacarse con servicios de calidad y tecnología innovadora. Entre las fortalezas se encuentran la experiencia acumulada y la infraestructura establecida, mientras que las debilidades incluyen la dependencia tecnológica y la necesidad de capacitación del personal. Las oportunidades radican en la personalización de servicios basada en datos, pero se deben enfrentar amenazas como la competencia agresiva y la gestión adecuada de datos.

**1.3 Business goal**

1. Incrementar las ventas y fomentar el reconocimiento mediante promociones, a través de la aplicación de estrategias de marketing personalizadas que se enfoquen en las características y preferencias específicas de cada segmento de clientes, identificados mediante análisis de datos.

**1.4 Business success criteria**

1. Observar el incremento en el valor de transacción promedio observando que estrategia obtiene un mayor incremento según la segmentación.
2. Índice de satisfacción al cliente cuando se muestre la promoción personalizada, a través de la aplicación con una pregunta de satisfacción.

**1.5 Data mining goal**

1. Utilizar el algoritmo K-Means para agrupar a los clientes en segmentos según su comportamiento de compra, con el objetivo de personalizar estrategias de marketing y mejorar la retención.
2. Emplear un análisis de componentes principales para reducir la dimensionalidad de los datos y extraer características significativas, facilitando la comprensión del modelo.

**1.6 Data mining success criteria**

1. Evaluar la separación de los grupos generados por el algoritmo K-Means, mediante heat map y box plot permitiendo determinar la calidad de la segmentación.

Evaluar los clusters con el algoritmo de clusterboot para determinar la estabilidad de los clusters, y por último utilizar el coeficiente de silueta para determinar la calidad de agrupamiento.

1. Evaluar los componentes principales según el número de variable que se lograron reducir.

**2. Data Understanding**

**2.1 Describe data**

En la base de datos se identificaron 47871 clientes donde tenemos variables categóricas y numéricas.

**Identificadores de cliente**: Hay 2 variables que ayudan a identificar el id del cliente y además una segmentación previa que se hizo con la base de datos.

**Número de transacciones**: Esta variable indica la cantidad total de transacciones realizadas en el mes, en este caso se identifica cuantas veces cada cliente hace una transacción de tipo crédito.

**Datos estadísticos sobre las transacciones**: Hay 4 variables que proveen datos estadísticos sobre la transacción indicando el monto promedio de dinero transferido en el mes, monto mínimo de las transacciones en el mes, monto máximo de las transacciones en el mes y por último la desviación estándar entre las transacciones del mes indicando cuanto varían los montos de las transacciones con respecto al promedio de transacciones.

**Porcentaje de uso de red de pago por compañía y localización:** Tenemos 8 variables que indican que tipo de red de pago se usó, para la transacción mostrando si es MasterCard, Visa u otra franquicia del banco,además de señalar si la transacción fue internacional o nacional.

**Momento de la transacción:** La base de datos posee 10 variables que indican en qué momento del día fue la transacción: mañana, tarde, noche y también cuáles días de la semana se hicieron las transacciones.

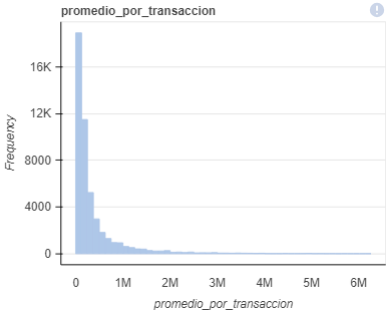
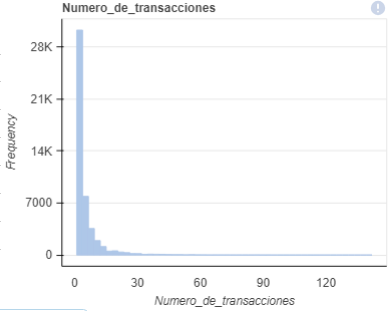
**Sitio de consumo frecuente:** Por último, hay una variable que muestra donde es el sitio de consumo más frecuente para el cliente, en esta variable podemos encontrar una gran cantidad de categorías.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

**2.2 Explore data**

En la base de datos podemos observar 4 comportamientos de las variables numéricas, donde el número de transacciones indica un sesgo hacia la derecha y curtosis alta con valores discretos, los datos estadísticos sobre las transacciones cuentan con el mismo sesgo y curtosis a diferencia que cuentan con datos continuos lo que hace más fácil de trabajar esta variable, por último, algunas variables porcentuales con amplia distribución tienen un sesgo bajo y curtosis cercana a 0, para las variables poco distribuidas como lo es con compras internacionales encontramos un sesgo negativo o positivo dependiendo de que tan común sea, estos valores aumentan mientras más uniforme sea el valor. En esta gráfica se puede observar el comportamiento general de algunas variables.



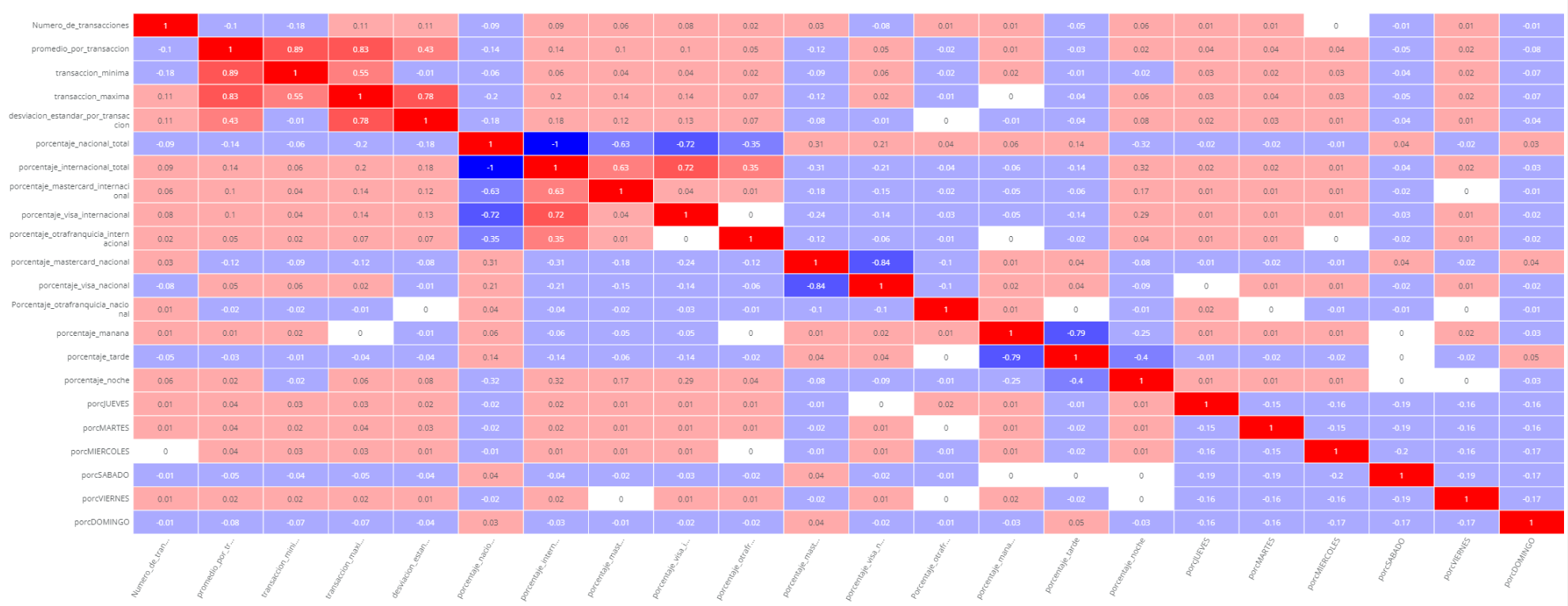
**Gráfico

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente**

*Gráfico 2. Comportamiento general de las variables.*

Para poder observar el comportamiento de todas las variables, observamos en la gráfica 3 una matriz de correlación de todos los datos numéricos. En esta matriz, destacamos que las variables de monto de transacción están altamente correlacionadas entre sí. Además, notamos una correlación negativa de -1 entre variables que representan lo opuesto, como por ejemplo "internacional" y "nacional", lo cual indica que, si una ocurre, la otra no sucede. Este hallazgo nos lleva a considerar la conveniencia de transformar nuestros datos de porcentaje a entero sabiendo que variables con una correlación demasiado alta puede afectar nuestro modelo de clustering. En la gráfica 4, se muestra con más detalle el comportamiento de correlación entre las variables "internacional\_total" y "nacional\_total" según su tipo de dato.

*  
Gráfico 3. Correlación entre todas las variables.*

*Imagen que contiene Forma

Descripción generada automáticamente*Imagen que contiene Forma

Descripción generada automáticamente

*Gráfico 4. Correlación entre transacciones nacionales e internacionales (Izq=porcentaje, Der=Entero).*

**2.3 Verify data quality**

El proceso de verificar la calidad de los datos es fundamental en cualquier análisis o investigación que se base en información cuantitativa. En este contexto, es imprescindible identificar y abordar posibles datos atípicos, tanto aquellos que son posibles dentro del dominio del problema como aquellos que son completamente improbables o imposibles. Por consiguiente, se analizarán los datos que podrían afectar nuestro modelo:

**Transacciones de 1$**: Hay transacciones de 1 peso que son generadas y pueden afectar nuestro análisis y más aún cuando es la única transacción que se hizo.

**Transferencias con desviación estándar 0$**: La inclusión de clientes que realizaron una sola transacción, o múltiples transacciones del mismo monto, puede sesgar nuestros datos, ya que varios clientes podrían caer en esta categoría, lo que a su vez podría etiquetar incorrectamente como atípicos a aquellos clientes cuyas transacciones no tienen una desviación estándar de cero. En la grafica 3 se puede mostrar como después de un proceso de transformación la desviación estándar esta muy sesgada por la desviación estándar 0$.

**Porcentajes**: Los porcentajes más que ser un valor atípico puede conducir a una perdida significativa de información sobre la magnitud de las variables, para una mejor ilustración consideremos un cliente con una transacción tiene el 100% de transacciones los viernes, pero un cliente con 100 transacciones tiene el 80% de transacciones los viernes. A simple vista, el clustering podría sugerir que el primer cliente tiene una tendencia mucho más marcada a realizar transacciones los viernes, cuando en realidad esto sería erróneo.

**Otra franquicia:** Esta variable tiene muy pocos datos con respecto a mastercard y visa por lo que produce un sesgo extremadamente alto positivo y una curtosis alta.

**Sitio de consumo frecuente:** Esta variable categórica puede plantear problemas de interpretación y análisis ya que contiene 109 valores únicos.

Gráfico, Histograma

Descripción generada automáticamente

*Gráfico 3. Sesgo en desviación estandar después de transformación.*

**3. Data preparation**

**3.1 Clean data**

El proceso de garantizar la calidad de los datos es fundamental para cualquier análisis. En el apartado de verificación de la calidad de los datos, se llevó a cabo un análisis para identificar posibles problemas y deficiencias en los datos recopilados. Este análisis dio lugar a una serie de procesos de limpieza de datos, los cuales se realizaron con el objetivo de remover, o estudiar los datos según fuera necesario para mejorar su calidad y confiabilidad. Estos procesos se llevaron a cabo, asegurando que cada acción realizada estuviera respaldada por una sólida fundamentación. Sin embargo, en algunos casos, se determinó que los datos no requerían ningún proceso de limpieza adicional.

**Limpieza de transacciones de 1$**: Con el propósito de mejorar la calidad de los datos y optimizar el modelo de negocio, se decidió implementar un proceso de limpieza de clientes basado en el valor de sus transacciones. En este sentido, se identificó que aquellos clientes que realizaron transacciones de 1$ podrían no ser representativos del perfil objetivo de la empresa. Por lo tanto, se determinó eliminar únicamente a aquellos clientes cuyas únicas transacciones registradas fueran de 1$. En cambio, los clientes que realizaron más de una transacción teniendo algún monto de 1$, fueron conservados en el modelo de negocio, garantizando así, que también pudieran beneficiarse de las promociones y ofertas pertinentes.

**3.2 Construct data**

En esta sección, exploraremos cómo hemos transformado nuestras variables existentes y creado nuevas características en nuestra base de datos final. Estas modificaciones son esenciales para mejorar la precisión y utilidad de nuestros modelos predictivos. A través de una descripción detallada de cada paso y su justificación, buscamos potenciar la capacidad de nuestro análisis para extraer información relevante y tomar decisiones fundamentadas.

**Porcentajes**: Como bien se ha dicho estos valores se prefieren como número entero para que se pueda tener una información más acertada de la magnitud de las variables, por lo tanto, se realiza una simple operación en cada variable que tenga datos porcentuales.

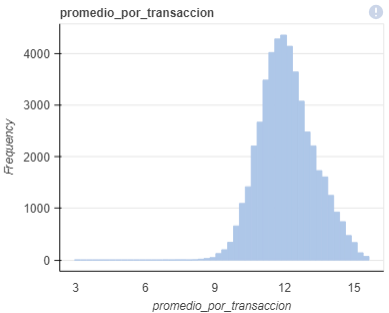
**Agrupación de días y redes de pago:** Para una mejor interpretabilidad de los datos, reducción de complejidad y observar patrones de comportamiento diferentes se decidió agrupar los días entre semana y fin de semana en estas dos variables, así igualmente con las redes de pago para tratar de disminuir un poco el sesgo y curtosis en las variables de otra franquicia.

**Categorización de sitio frecuente:** La columna que indica el sitio más frecuente de transacción contiene 109 valores únicos por lo que se optó por categorizar esos variables en estas categorías:

|  |  |  |
| --- | --- | --- |
| Salud y Bienestar | Hogar y Construcción | Tecnología |
| Comercio | Alimentación y Restaurantes | Agropecuario |
| Finanzas | Entretenimiento | Eventos |
| Otra | Educación | Arte y Antigüedades |
| Moda | Servicios | Mascotas |
| Automotriz | Deporte y Recreación | Seguridad |
| Viajes y Transporte | Industria |  |

**Log1p:** El logaritmo sirve para ayudar a transformar datos muy sesgados en una distribución más normalizada, ya que los clusters funcionan mejor con esta distribución, este logaritmo solo va a ser usado para los montos de transacciones ya que es una variable con datos continuos, a diferencia de las variables que cuentan número de transacciones ya que tienen datos discretos. En la gráfica 4 se puede observar el cambio usando el logaritmo en la variable transacción promedio.

**Gráfico, Histograma

Descripción generada automáticamente**

*Gráfico 4. Transformación de distribución después de usar logaritmo.*

Por último, log1p se usó para evitar con la desviación estándar 0, un error de cálculo.

**Estandarización Z:** La estandarización de variables numéricas es esencial para abordar las disparidades en las escalas de las variables, lo cual puede sesgar el análisis, especialmente en técnicas como el clustering. Por ejemplo, si consideramos variables como los montos de transacciones, que varían de 100k a 10M, y el número de transacciones, que oscila entre 1 y 200, las diferencias de escala pueden influir significativamente en los resultados del clustering. Sin estandarización, las variables con valores más grandes ejercen un peso desproporcionado en el análisis debido a sus rangos más amplios. Al estandarizar las variables, se alinean todas en una escala común. Esto garantiza que todas las variables contribuyan de manera equitativa al proceso de clustering, permitiendo que el algoritmo se base en las relaciones y similitudes entre las variables en lugar de sus magnitudes.

**3.3 Dataset description**

El conjunto de datos final para realizar el modelo de negocio se puede observar en la tabla 2. Cabe aclarar que no todos estos datos van a ser usados para el modelamiento de clustering más sin embargo si se usarán para observar patrones dentro el resultado de los clusters. Por ejemplo, ver si hay patrones de compra en los clusters con la categoría de sitio frecuente.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

**4. Modeling**

**4.1 Select modeling techniques**

Antes de iniciar el modelado de clustering, decidimos reducir la complejidad de nuestros datos mediante el uso de análisis de componentes principales. Se enfocó en un conjunto específico de variables relevantes para nuestro negocio, que incluyen *promedio\_de\_transacciones, nacional\_total, internacional\_total, mañana, tarde* y *noche*. Este enfoque permite obtener una representación más compacta de los datos y una menor complejidad al hacer análisis. Posteriormente, se aplicó la rotación Varimax a los componentes principales obtenidos. Esta técnica de rotación nos ayudó a mejorar la interpretación de los pesos de nuestras nuevas variables, permitiéndonos identificar patrones más claros y significativos en nuestros datos.

Finalmente, utilizamos el algoritmo de clustering K-means para agrupar nuestros datos en clusters coherentes. Dado que nuestras variables se limitan a datos numéricos, K-means resultó ser la elección más adecuada. Este método de clustering se adapta bien a este escenario, ya que utiliza la distancia euclidiana para calcular las similitudes entre observaciones, lo cual es una restricción para trabajar únicamente con variables numéricas.

**4.2 Generate test design**

El método de rotación que se usó se escogió gracias a un mejor desempeño que PCA, esto se puede comprobar ya que se busca una ortogonalidad entre las nuevas variables, es decir que tengan correlación 0 idealmente, en el gráfico 5 podemos observar la matriz de correlación usando ***none*** y ***varimax***.

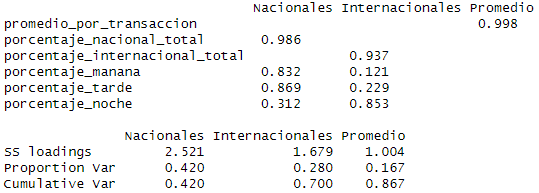
Forma, Cuadrado

Descripción generada automáticamenteForma, Cuadrado

Descripción generada automáticamente

*Gráfico 5. Matriz de correlación entre las nuevas variables*

Después de usar componentes principales y factorial nuestras variables quedan representadas así:



Como podemos observar escogimos 3 variables ya que en el *cumulative var* obtenemos el 80% de todas nuestras variables, también podríamos concluir que de pronto existe una alta correlacionalidad en hacer compras nacionales entre mañana y tarde, y hacer compras internacionales en la noche.

Ya que tenemos nuestras nuevas variables procedemos a usar el modelo de clustering k-means, esto debido a que solo vamos a trabajar con datos numéricos. Ahora bien, vamos a usar la técnica de ASW para calcular el número de clusters recomendado, donde el número más cercano a 1 indica el mejor número de clusters.

Texto

Descripción generada automáticamente

Por ende, usamos el parámetro de 5 clusters.

**4.3 Build model**

Imagen que contiene Texto

Descripción generada automáticamenteBasándonos en el análisis realizado utilizando la técnica de ASW, se determinó que el número óptimo de clusters es 5. Por lo tanto, se procede a configurar el algoritmo k-means con el parámetro centers igual a 5, lo que implica que se buscarán cinco clusters en los datos. Además, se establece el número de iteraciones en 10, lo que significa que el algoritmo se ejecutará en 10 inicializaciones aleatorias. Esto se hace para mitigar cualquier influencia que una inicialización no óptima pueda tener en el resultado final. También para garantizar que el modelo no se ejecute indefinidamente, se establece la iteración máxima en 20. Esto asegura que el algoritmo se detenga después de un máximo de 20 iteraciones, incluso si no ha convergido completamente. Esta configuración permite un equilibrio entre la búsqueda exhaustiva de soluciones y un tiempo de ejecución razonable. Dándonos como resultados estos clusters con su estimación de estabilidad clusterboot.

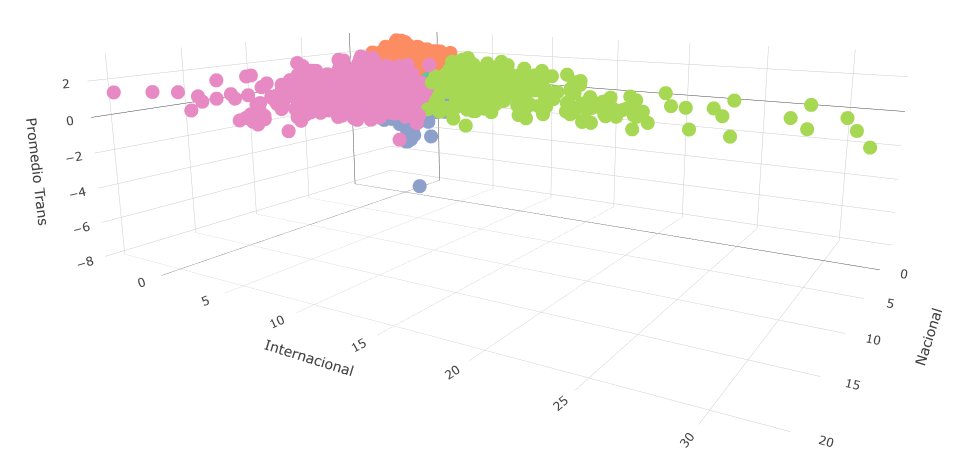
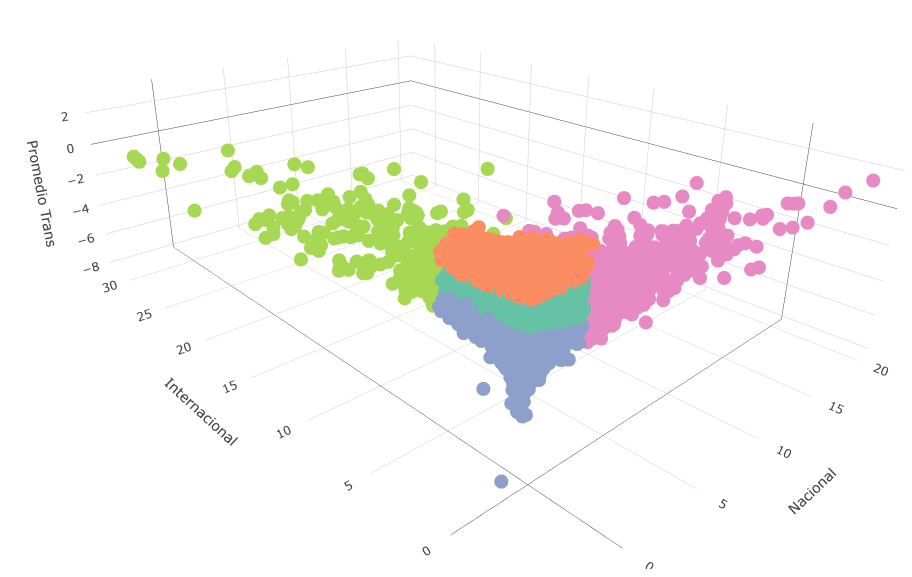
Texto

Descripción generada automáticamente

**5. Evaluation**

**5.1 Assess model**

Antes de finalizar con los resultados, se realizó una evaluación satisfactoria de la reducción de variables. Además, se empleó el algoritmo de silueta para calcular la calidad de los clusters generados, obteniendo un valor de 0.34. Aunque idealmente este valor sería 1, una medida de 0.34 se considera aceptable, indicando una separación adecuada entre los clusters y una asignación apropiada de los puntos a sus respectivos clusters. Es importante señalar que, a pesar de esta medida aceptable, podría existir cierta superposición entre los clusters, lo cual es común en algunos conjuntos de datos. Para visualizar el comportamiento de los clusters, se puede consultar el gráfico 6.

Gráfico

Descripción generada automáticamente con confianza media

*Gráfico 6. Perspectiva visual del clustering.*

Realizando un análisis de los clusters en el gráfico 7 observamos que la primera variable está muy correlacionada al número de transacciones por lo que podemos obtener más información para declarar nuestros clusters:

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

*Gráfico 7. Número de transacciones por cluster.*

También es importante destacar la magnitud de los pesos en la matriz de correlaciones, representada en el gráfico 8. Esta visualización nos proporciona información sobre en qué clusters la magnitud de la variable es más alta. Es un aspecto relevante para comprender la distribución y la importancia de las variables dentro de cada cluster.

*Gráfico, Gráfico de rectángulos

Descripción generada automáticamente*

*Gráfico 8. Matriz de correlación de las variables.*

Finalmente, con el objetivo de ofrecer una promoción más enfocada y personalizada, se identifican las categorías en las que los clusters gastaron más. Esto nos permite dirigir nuestros esfuerzos promocionales hacia estas áreas específicas, maximizando así el impacto de nuestras iniciativas. Para obtener más detalles, se puede observar el gráfico 9.

*Gráfico, Gráfico de barras

Descripción generada automáticamente*

*Gráfico 9. Observación de número de clientes por categoría y clustering.*

Segmentación de clientes:

* **Compradores ocasionales** (Cluster 1): Estos usuarios realizan transacciones de manera ocasional, con un gasto moderado y principalmente en transacciones locales. Son personas que utilizan su tarjeta de crédito de forma esporádica para compras pequeñas y cotidianas.
* **Compradores de lujo** (Cluster 2): Este grupo de usuarios realiza menos transacciones, pero tiende a gastar más en cada una. Prefieren productos y servicios de alta calidad y están dispuestos a pagar un precio más alto por ellos
* **Ahorradores** (Cluster 3): Estos usuarios realizan pocas transacciones y gastan poco en cada una. Son personas que tienden a ser conservadoras con sus finanzas y prefieren ahorrar en lugar de gastar.
* **Consumidores locales activos** (Cluster 4): Este segmento realiza un gran número de transacciones locales, aunque el gasto individual puede ser bajo. Son usuarios que participan activamente en la economía local, apoyando negocios y servicios locales con transacciones frecuentes.
* **Consumidores internacionales activos** (Cluster 5): Estos usuarios realizan un alto volumen de transacciones internacionales, ya sea por viajes frecuentes o por compras en línea en mercados extranjeros.

**5.2 Produce final report**

Este informe gerencial tiene como objetivo presentar estrategias sugeridas para cada uno de los segmentos de clientes identificados, con un enfoque en la mejora de la participación del cliente y el crecimiento del negocio. Se analizará el costo-beneficio de estas estrategias y se propondrán posibles pasos adicionales para un análisis más detallado.

**Compradores ocasionales**

* **Estrategia 1:** Descuento en la primera compra con la tarjeta de crédito.
* **Estrategia 2:** Programa de recompensas por uso frecuente de la tarjeta.
* **Estrategia 3:** Crear programas de fidelización que recompensen específicamente las compras en comercios, moda y tecnología. Por ejemplo, acumular puntos de recompensa más rápidamente cuando se realizan compras en estos sectores, o recibir bonificaciones especiales.

**Compradores de lujo**

* **Estrategia 1:** Personalización de servicios bancarios premium.
* **Estrategia 2:** Ofrecer acceso exclusivo a experiencias de viaje de lujo, como excursiones privadas, tours VIP, acceso a eventos culturales exclusivos, entre otros. Ya que tienen varios gastos en viajes, transportes y comercio.
* **Estrategia 3:** Proporcionar servicios personalizados que ayuden a los clientes a planificar y reservar sus viajes de lujo, ofreciendo recomendaciones personalizadas, asistencia en la planificación de itinerarios.

**Ahorradores**

* **Estrategia 1:** Asesoramiento financiero personalizado para maximizar el ahorro.
* **Estrategia 2:** Programa de recompensas por ahorro, donde según sus metas de ahorro se puedan acumular puntos para hacer gastos ya sea en comercio, moda o servicios.

**Consumidores locales activos**

* **Estrategia 1:** Promoción de tarjetas de crédito con beneficios locales.
* **Estrategia 2:** Descuentos en comercios locales.

**Consumidores internacionales activos**

* **Estrategia 1:** Tarifa de cambio de divisa preferencial.
* **Estrategia 2:** Eliminación de tarifas por transacciones internacionales.
* **Estrategia 3:** Descuento en productos con convenio con empresas de tecnología.

El análisis costo/beneficio de las estrategias propuestas implica evaluar los costos asociados con la implementación y mantenimiento de cada estrategia, así como los beneficios esperados en términos de adquisición y retención de clientes, aumento de ingresos y mejora de la experiencia del cliente. Para ello, se deben considerar aspectos como los costos de promoción, los descuentos ofrecidos y los recursos necesarios para personalizar servicios y programas de recompensas. Los siguientes pasos de análisis incluyen realizar un estudio continuo de mercado detallado para comprender las necesidades y preferencias de cada segmento de clientes, y así analizar si los clientes siguen con la misma tendencia o hay cambios con respecto al tiempo, por último, realizar pruebas piloto de las estrategias seleccionadas y monitorear continuamente los resultados para realizar ajustes según sea necesario y maximizar el retorno de la inversión.

**Bibliografía**

**1.** Kotler-Keller, (2015) Marketing Management.. Pearson.