Trabajo Final de Aprendizaje No Supervisado

Fecha: 15/11/2024

Tema: Segmentación de Clientes de Tarjetas de Crédito mediante Clustering K-Means

Integrantes: Federico Libertun

Adaia Montaño

Sol Guerreiro

1. Introducción

El dataset utilizado en este trabajo es una muestra de una cartera de clientes con tarjetas de crédito, que incluye información sobre el comportamiento de 8.950 usuarios activos durante un período de 6 meses. Este conjunto de datos contiene 18 variables que reflejan diferentes aspectos del comportamiento financiero de los clientes. Entre ellas, se encuentran indicadores relacionados con el saldo disponible, el límite de crédito, los pagos mínimos realizados, el porcentaje de pago completo y el monto total de compras, los cuales ayudan a evaluar la capacidad de gasto y el compromiso de los usuarios con sus obligaciones financieras.

Asimismo, el dataset incluye variables como la frecuencia de compras y la frecuencia de adelantos en efectivo, así como el pago de esos adelantos, que pueden ser útiles para identificar perfiles de clientes con comportamientos de riesgo financiero. Por otro lado, la antigüedad del servicio actúa como un indicador del historial de uso, proporcionando una perspectiva sobre la lealtad y estabilidad de los clientes en su relación con la entidad financiera.

El objetivo principal de este análisis es aplicar técnicas de aprendizaje no supervisado, específicamente el algoritmo K-Means, para identificar patrones de comportamiento y segmentar a los clientes en grupos homogéneos. Este enfoque permitirá explorar características comunes entre los grupos, lo que puede ser valioso para diseñar estrategias personalizadas, optimizar la gestión de la cartera de clientes y predecir comportamientos futuros.

2. Breve Descripción del método de clustering utilizado y su justificación

K-Means es un algoritmo de clustering ampliamente utilizado en aprendizaje no supervisado que particiona un conjunto de datos en K grupos, minimizando la variabilidad dentro de cada cluster. Requiere variables cuantitativas, un número de clusters definido previamente, y utiliza la distancia euclidiana para medir similitud. El proceso implica la inicialización de los centroides, el cálculo iterativo de estos para cada cluster y la reasignación de observaciones al cluster más cercano, repitiendo hasta alcanzar estabilidad

Elegimos K-Means porque los centroides generados por este método permiten interpretar fácilmente el perfil promedio de cada cluster, lo cual es valioso para identificar patrones de comportamiento entre los clientes y facilita la toma de decisiones comerciales. K-Means es particularmente adecuado para segmentar clientes, ya que busca minimizar la variabilidad dentro de cada grupo, resultando en clusters bien definidos y diferenciados entre sí.

Además, K-Means es conocido por su rapidez y eficiencia en problemas de clustering. Dado que K-Means opera con variables numéricas y utiliza la distancia euclidiana como medida, el dataset cumple con los supuestos necesarios para aplicar el modelo de forma adecuada, asegurando un análisis óptimo y relevante para nuestros objetivos.

3. Exploración y Limpieza de los Datos

El dataset original contiene información de 8.950 usuarios de tarjetas de crédito con 18 variables que describen su comportamiento financiero. Como primer paso en el análisis exploratorio, se utilizó el método describe() para observar estadísticas básicas como medias, desviaciones estándar y cuartiles de cada variable. Un punto llamativo fue la variable TENURE, que representa la antigüedad en el servicio de tarjeta de crédito: aunque su media es 11, todos sus cuartiles tienen un valor de 12. Esto sugiere que la mayoría de los usuarios del dataset tienen la misma antigüedad, lo cual podría indicar que se trata de clientes que comenzaron a usar el servicio en un periodo similar.

Al analizar los valores nulos, encontramos que la variable MINIMUM_PAYMENTS tiene 313 registros sin datos (aproximadamente el 3.5% del total de registros), mientras que CREDIT_LIMIT presenta un único registro nulo. En lugar de imputar valores, decidimos eliminar estos registros, considerando que estas variables deberían tener datos completos debido a la naturaleza de las operaciones con tarjetas de crédito. Por otro lado, se eliminó la variable CUST_ID, ya que solo representa una identificación única para cada cliente y no aporta información relevante para el análisis de comportamiento.

Por otro lado, en la búsqueda de inconsistencias, identificamos casos en los que la columna PAYMENTS (pagos realizados por el usuario) presenta valores más bajos que la columna MINIMUM_PAYMENTS (pago mínimo requerido). Aunque podría parecer un error en los datos, asumimos que estos registros son correctos, ya que reflejan una situación común en el uso de tarjetas de crédito. Es posible que, al momento del vencimiento, se habilite únicamente el pago mínimo para evitar retrasos, pero posteriormente el cliente realice un pago adicional, no total, que sea menor al monto mínimo previamente facturado. Este comportamiento es consistente con la práctica financiera de algunos usuarios y no se considera una anomalía en el dataset.

Para identificar y visualizar los valores atípicos, se generaron gráficos de box plot para cada variable. Este análisis reveló que la variable CASH_ADVANCE_FREQUENCY, que debería oscilar entre 0 y 1, tenía 8 registros con valores superiores a 1, probablemente debido a errores de carga en la base de datos. Estos registros fueron eliminados por no cumplir con los rangos esperados.

Posteriormente, se calculó la cantidad de outliers en cada variable utilizando el rango intercuartílico (IQR). Esto mostró que algunas variables tienen un número considerable de valores extremos, lo cual es común en datos financieros, donde las distribuciones suelen ser asimétricas. Sin embargo, en lugar de eliminar todos los outliers, decidimos agrupar algunos valores en rangos categóricos para reducir el impacto de estos extremos en el análisis. Por ejemplo, se crearon nuevas variables de rangos para BALANCE, PURCHASES, ONEOFF_PURCHASES, INSTALLMENTS_PURCHASES, CASH_ADVANCE, CREDIT_LIMIT, PAYMENTS y MINIMUM_PAYMENTS.

Realizamos también un análisis de correlación entre las variables para identificar relaciones lineales significativas. Se observaron algunas correlaciones altas, como entre PURCHASES y ONEOFF_PURCHASES (correlación de 0.92), y entre CASH_ADVANCE_TRX y

CASH_ADVANCE_FREQUENCY (correlación de 0.8). Estos hallazgos son importantes, ya que variables altamente correlacionadas podrían aportar información redundante, lo cual es relevante para el análisis posterior, pero al realizar el modelado con sin ellas no encontramos diferencias significativas, y si nos genera valor tener en el caso por ejemplo de PURCHASES y ONEOFF_PURCHASES estos valores a la hora de sacar y reforzar conclusiones de los comportamientos.

4. Selección del Número de Clusters

Para determinar el número óptimo de clusters en el análisis de K-Means, empleamos dos métodos de validación: el Método del Codo y el Índice de Silhouette. Ambos enfoques ayudan a evaluar la estructura de los clusters y a seleccionar un valor de K que maximice la cohesión interna de los clusters y la separación entre ellos.

El Método del Codo evalúa la suma de las distancias cuadradas dentro de los clusters (WCSS) para diferentes valores de K. A medida que aumenta el número de clusters, el valor de WCSS disminuye, ya que cada cluster contiene menos puntos. El objetivo es identificar el punto donde la reducción de WCSS se vuelve menos pronunciada, formando un "codo" en la gráfica, lo cual indica el número de clusters óptimo. En nuestro caso, al observar la gráfica generada, el codo aparece claramente en K=3, lo que sugiere que tres clusters es el número óptimo para segmentar los datos.

El Índice de Silhouette mide la cohesión y separación de los clusters, proporcionando una evaluación del número de clusters más adecuado. Este índice varía entre -1 y 1, donde un valor cercano a 1 indica clusters bien definidos y separados. Calculamos el índice de Silhouette para valores de K entre 2 y 10, y seleccionamos el número de clusters que maximiza este valor que para nuestra gráfica nos indica que es 2 clusters.

Aunque el Índice de Silueta sugiere K=2 como el número óptimo de clusters, optamos por utilizar K=3 basado en el resultado del Método del Codo. Esta decisión se fundamenta en una consideración de negocio: contar con tres grupos permite una segmentación más detallada del comportamiento de los clientes, lo cual resulta en una mayor capacidad para personalizar estrategias comerciales y para adaptar productos y servicios a subgrupos específicos dentro de la cartera de clientes

5. Selección de Muestra y Estandarización de las Variables

Para facilitar el análisis y reducir el tiempo de procesamiento, seleccionamos una muestra aleatoria de 1.000 registros del dataset que quedó de la limpieza. Esta muestra permite mantener una representación suficiente de los patrones generales sin comprometer la capacidad de ejecución y análisis del modelo. Utilizamos una semilla fija para asegurar la reproducibilidad de los resultados.

Antes de aplicar el algoritmo de K-Means, estandarizamos todas las variables de la muestra seleccionada utilizando el método StandardScaler de sklearn, el cual transforma cada característica para que tenga una media de 0 y una desviación estándar de 1. Este paso es

fundamental para el análisis de clustering, ya que K-Means utiliza la distancia euclidiana para medir similitud, y las variables en diferentes escalas podrían afectar los resultados.

La estandarización asegura que todas las variables tengan la misma importancia en el cálculo de las distancias, evitando que aquellas con valores numéricos más grandes dominen el proceso de clustering. Esto es especialmente importante en nuestro caso, donde las variables del dataset original, como BALANCE y CREDIT_LIMIT, pueden tener valores significativamente mayores que variables de frecuencia, como PURCHASES_FREQUENCY o CASH_ADVANCE_FREQUENCY.

6. Aplicación de K-Means y Análisis de Clusters

A continuación, estaremos detallando 2 tipos de aplicaciones de K-Means, una con los 3 clusters que nos sugirió el método de codos y otra con 4 clusters dado que es un valor cercano a ese que nos puede aportar información de algún grupo adicional con característica que valen la pena diferenciar.

Primero, optamos por utilizar 3 clusters, siguiendo la recomendación del Método del Codo. Para hacer el análisis de los clusters utilizamos las variables que consideramos principales a la hora de hablar de comportamiento del usuario. Considerando esas variables de comportamiento encontramos que:

- Cluster 0: Este grupo presenta una frecuencia alta de compras en cuotas y balances actualizados frecuentemente. Además, tienen un límite de crédito alto y realizan un número considerable de compras, pero tienden a evitar los adelantos en efectivo. Estos usuarios son consistentes en mantener su saldo actualizado y muestran un comportamiento de consumo moderado y planificado, probablemente debido a su preferencia por el uso de cuotas. Son clientes que podrían considerarse de riesgo bajo debido a su aversión a los adelantos en efectivo y su capacidad para mantener un saldo controlado.
- Cluster 1: Los usuarios en este cluster tienen una frecuencia baja de compras en cuotas y utilizan adelantos de efectivo con frecuencia, además de tener una capacidad limitada para realizar el pago completo de sus deudas. Su límite de crédito es más bajo en comparación con el Cluster 0, y muestran un comportamiento más riesgoso, al depender de adelantos de efectivo. Estos clientes podrían beneficiarse de estrategias de educación financiera para optimizar su uso de la tarjeta y evitar el endeudamiento a través de adelantos de efectivo.
- Cluster 2: Este cluster está caracterizado por una frecuencia de balance actualizada baja y un porcentaje relativamente alto de pagos completos. Tienen límites de crédito moderados y no realizan adelantos de efectivo frecuentemente. Este grupo parece incluir clientes con un comportamiento financiero más estable, ya que son capaces de realizar pagos completos de sus deudas y mantienen un bajo nivel de endeudamiento. Podrían ser usuarios con buenos hábitos de pago, aunque con una menor actividad en su cuenta.

En resumen, la segmentación en 3 clusters permite identificar distintos perfiles de clientes según su comportamiento financiero: el Cluster 0 representa a clientes de bajo riesgo, con un

consumo moderado y preferencia por compras en cuotas; el Cluster 1 incluye usuarios con alta dependencia de adelantos de efectivo y capacidad limitada para realizar pagos completos, lo que los sitúa en un perfil de riesgo; y el Cluster 2 agrupa a clientes con comportamiento financiero estable y una tendencia a evitar el endeudamiento.Si bien el Método del Codo sugirió 3 clusters como el número óptimo para segmentar los datos, decidimos explorar una división en 4 clusters, ya que es un número cercano al recomendado por el método y podría ofrecer una mayor granularidad en la segmentación.

Por otro lado, en la aplicación del modelo de K-Means con 4 clusters obtuvimos los siguientes perfiles:

- Cluster 0: Clientes de bajo riesgo con un consumo planificado, alta frecuencia de compras en cuotas y bajo uso de adelantos de efectivo.
- Cluster 1: Usuarios de alto riesgo, con alta frecuencia de adelantos de efectivo y baja capacidad para realizar pagos completos.
- Cluster 2: Clientes estables y de bajo riesgo, con baja actividad general y un bajo nivel de endeudamiento.
- Cluster 3: Usuarios de riesgo moderado a alto, con bajo límite de crédito y alta dependencia de adelantos de efectivo para mantener su saldo.

6. Conclusiones

Consideramos que la elección entre 3 o 4 clusters depende de las necesidades específicas que buscamos en el negocio:

- Opción de modelado con 3 Clusters: permiten una segmentación más simple y clara, con un grupo de bajo riesgo, uno de riesgo alto y otro de estabilidad media. Esta opción facilita una estrategia de manejo de clientes basada en tres perfiles generales, ideal para una implementación rápida y sin gran complejidad.

- Opción de modelado con 4 Clusters: proporcionan una segmentación más detallada, identificando usuarios de riesgo moderado en el Cluster 3 y separando a los usuarios de bajo riesgo en Clusters 0 y 2 con comportamientos de consumo distintos. Esta opción puede ser útil para estrategias más específicas, como campañas personalizadas o ajustes en políticas de crédito para usuarios con distintas necesidades de financiamiento.

Creemos que si el objetivo es optimizar la gestión del riesgo y las estrategias de marketing en un esquema más detallado, 4 clusters permiten una segmentación con mayor precisión. Sin embargo, si se busca una estrategia simplificada y fácil de implementar, 3 clusters serían suficientes para identificar los perfiles principales y aplicar políticas generales de manejo de clientes.