Caso de estudio BankCard

La entidad financiera BankCard está buscando mejorar su estrategia de marketing para los titulares de tarjetas de crédito. Actualmente, carecen de una comprensión clara del comportamiento de sus clientes y desean segmentarlos en grupos distintos para adaptar mejor sus estrategias de marketing a las necesidades y preferencias de cada grupo.

El objetivo es segmentar a los titulares de tarjetas de crédito en grupos distintos en función de su comportamiento de uso de la tarjeta. Al hacerlo, BankCard espera identificar patrones comunes entre los grupos y desarrollar estrategias de marketing específicas para cada segmento. Esto les permitirá ofrecer productos y servicios más personalizados, mejorar la retención de clientes y aumentar la satisfacción del cliente.

El conjunto de datos de muestra resume el comportamiento de uso de aproximadamente 9000 titulares activos de tarjetas de crédito durante los últimos 6 meses. El archivo está a nivel de cliente con 18 variables de comportamiento.

A continuación, se muestra el Diccionario de datos para el conjunto de datos de tarjetas de crédito:

Nombre de la Variable	Descripción
BALANCE	Monto de saldo restante en su cuenta para hacer compras
BALANCE_FREQUENCY	Valores entre 0 y 1 1 = Actualización de saldo frecuente 0 = No Actualiza el saldo con frecuencia
CASH_ADVANCE	Efectivo anticipado en su tarjeta de crédito
CASH_ADVANCE_FREQUENCY	Valores entre 0 y 1 1 = Paga frecuentemente el efectivo anticipado 0 = No paga frecuentemente el efectivo anticipado.
CASH_ADVANCETRX	Número de transacciones realizadas con "Efectivo anticipado"
CREDIT_LIMIT	Límite de la tarjeta de crédito para el usuario
CUST_ID	Identificación del titular de la tarjeta de crédito (Categórico)
INSTALLMENTS_PURCHASES	Monto de compra realizada en cuotas
MINIMUM_PAYMENTS	Monto mínimo de pagos realizados por el usuario
ONEOFF_PURCHASES	Monto máximo de compra realizada de una vez
ONEOFF_PURCHASES_FREQUENCY	Valores entre 0 y 1 1 = Compras frecuentes 0 = Compras no frecuentes
PAYMENTS	Monto de pago realizado por el usuario
PRC_FULL_PAYMENT	Porcentaje de pago completo realizado por el usuario
PURCHASES	Monto de compras realizadas desde la cuenta
PURCHASES_FREQUENCY	Con qué frecuencia se realizan las compras, puntaje entre 0 y 1 (1 = compras frecuentes, 0 = compras no frecuentes)
PURCHASES_TRX	Número de transacciones de compra realizadas
PURCHASESINSTALLMENTSFREQUENCY	Con qué frecuencia se realizan compras a plazos (1 = frecuentemente realizadas, 0 = no frecuentemente realizadas)
TENURE	Duración del servicio de tarjeta de crédito para el usuario

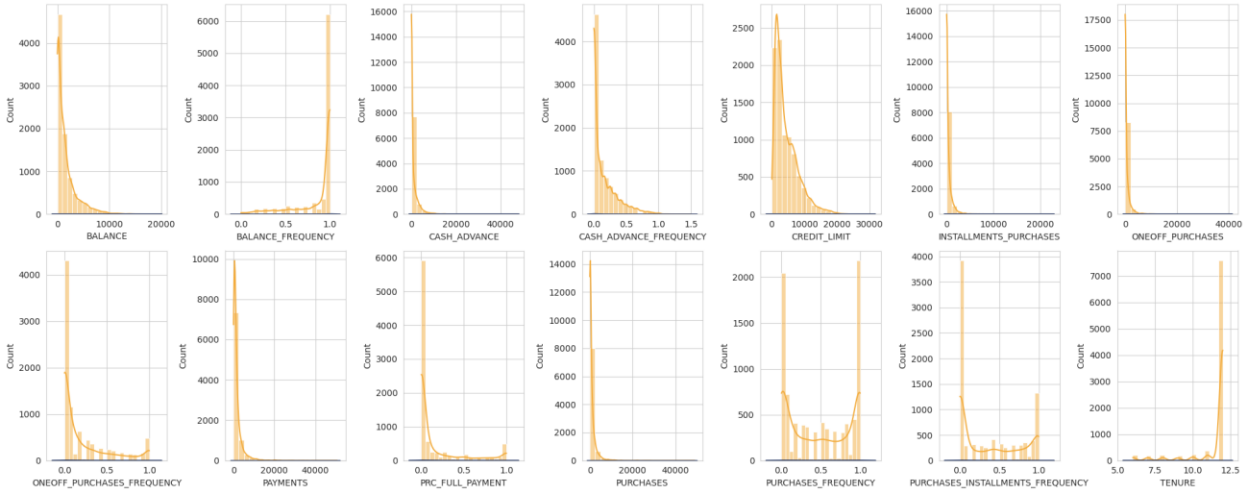
***Dataset* inicial** (df)

De acuerdo con la tabla presentada anteriormente el *dataset* cuenta inicialmente con 8.950 observaciones divididas en 18 columnas, sin embargo, para el tratamiento de los datos, solo se conservarán las variables numéricas, por lo que solo será necesario eliminar el CUST_ID. Todas las variables son flotantes, exceptuando CASH_ADVANCE_TRX, PURCHASES_TRX y TENURE, que son variables enteras.

De las variables mencionadas anteriormente, las características relacionadas con TRX y MINIMUM_PAYMENTS contienen información que ya está representada de alguna manera en otras variables del conjunto de datos, por lo que se pueden eliminar de la base de datos para tener un mejor tratamiento de los mismos.

EDA

La **distribución** de las variables muestra una evidente presencia de datos atípicos, y distribuciones tienen un sesgo significativo representado hacia la derecha, representando valores extremos o mayores en características específicas que serán tratados posteriormente.



En cuanto a la **correlación**, hay ciertas características que cuentan con una correlación superior al 0.5 (que se tomó como valor de referencia moderado). En esta matriz destacan dos valores por encima del 0.85 los cuales se consideraron altos y deberían ser tratados, el primer par de variables es PURCHASES y ONEOFF_PRUCHASES, y el segundo es PURCHASES_INSTALLMENTS_FRECUENCY y PURCHASES_FRECUENCY con valores asociados de 0.92 y 0.86, respectivamente. Sin embargo, se consideró que el segundo par de variables, presentan información valiosa para el tratamiento, por lo cual, se decide conservar ambas, contrariamente, entre el otro par, se decide eliminar la variable PURCHASES ya que tiene similitud con otras variables del mismo *dataset*.

Esto nos lleva a reducir el *dataset* a un total de 13 variables para el caso de estudio.

Como se mencionaba anteriormente, es necesario hacer un **tratamiento de *outliers***, ya que los modelos de segmentación son sensibles a los datos atípicos, y debido a la alta cantidad de observaciones atípicas, se decide eliminar aquellos que estén por encima del último cuartil y, teniendo en cuenta, el rango intercuartílico con un multiplicador de 3 para ampliar más el límite superior. Al realizar dicho tratamiento, se logra reducir el dataset a una cantidad de 7.496 observaciones (83,75% del *dataset* original).

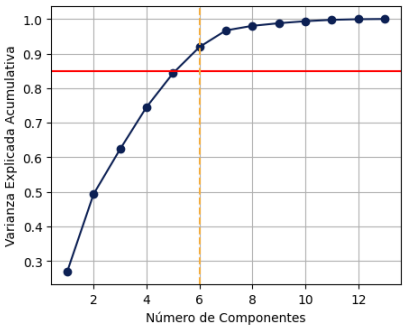
Otro de los tratamientos que se hizo es la eliminación de frecuencias que sobrepasaban el valor de 1, ya que este era el límite, y la variable CASH_ADVANCE_FREQUENCY presentaba 6 observaciones por encima de dicho límite, consiguiendo entonces un dataset nuevo de 7.490 observaciones.

Con este tratamiento se logró una mejor distribución de las variables, sin embargo, se seguían presentando valores muy por encima de del último cuartil, por lo que se decidió hacer un **ajuste *threshold***, con esto lo que se busca un valor predefinido que se utiliza como límite para identificar observaciones que se desvían significativamente del comportamiento general de un conjunto de datos, posteriormente se llevan todos estos datos atípicos al límite establecido buscando conseguir una distribución más normal.

En este proceso se redimensionaron aproximadamente 10% de las observaciones de algunas características, logrando una distribución más normal en las variables. Y de acuerdo con la información de las variables resultantes, se presentó un dato nulo el cual fue eliminado para tener un nuevo dataset nombrado data_clean, con un total de 7.489 observaciones.

Finalmente, para trabajar todas las variables en la misma escala, se utilizó el método StandarScaler() para realizar una **estandarización de variables** y continuar con la construcción de modelos de segmentación.

Reducción dimensionalidad



Mediante el uso de la función PCA(), muestra como resultado 6 componentes, dichos componentes pueden explicar el 85% de la variabilidad en los datos escalados, lo que sugiere que se puede reducir la dimensionalidad de los datos manteniendo la mayoría de su información importante. Este resultado ayudará en la selección de características para el modelo de segmentación de los clientes.

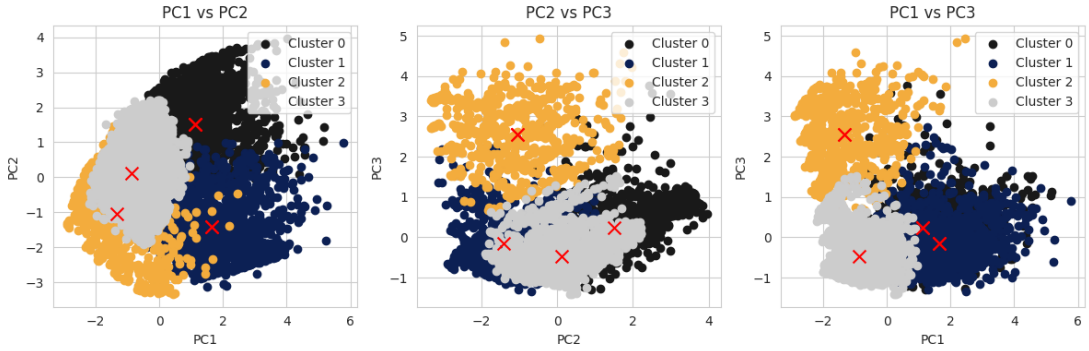
PCA Teniendo en cuenta estos resultados, utilizando las 3 primeras componentes, la reducción de la dimensionalidad explica al rededor del 62% de la varianza. El PC1 explica alrededor del 26,8% de la varianza total, lo cual sugiere que captura una cantidad significativa de la variabilidad en los datos. Los PC siguientes explican gradualmente menos varianza, con el PC2 explicando aproximadamente un poco menos que el PC1 (22,5%), y así sucesivamente.

Modelo K-MEANS: Con este modelo, se puede observar que a medida que aumenta el número de clústeres, el valor de la inercia tiende a disminuir. Sin embargo, hay un punto en el gráfico donde la disminución en la inercia comienza a ser menos significativa. Este punto se conoce como "codo" y es donde añadir más clústeres no produce una reducción sustancial en la inercia. En este caso, parece que el codo está alrededor de 3 o 4 clústeres, lo que sugiere que estos valores podrían ser opciones razonables para el número de clústeres en el análisis de *K-Means*.

Los resultados de este modelo, segmentan a los clientes en 4 grupos diferentes, distribuidos de la siguiente manera; **Clúster 0:** 1469, **Clúster 1:** 3812, **Clúster 2:** 1566 y **Clúster 3:** 642.

También se trató el modelo con reducción de dimensionalidad, en donde se comportaba igual al modelo anterior, sin embargo, se mejoraban las estadísticas presentadas, la distribución de los clústeres varió de la siguiente manera: **Clúster 0:** 652, **Clúster 1:** 3799, **Clúster 2:** 1473 y **Clúster 3:** 1565.

En la representación gráfica de los clústeres, se pueden identificar los grupos separados, sin embargo, hay varios datos que se ven mezclados y esto puede representar un poco de ruido para la interpretación de los clústeres.



Modelo Jerárquico: El dendrograma muestra cuatro uniones significativas donde la distancia entre los grupos es mayor, lo que sugiere que se pueden formar cuatro grupos distintos en los datos, como el modelo representa una misma cantidad de grupos, al igual que K-means, se estudian las estadísticas de cada modelo para validar cual tiene mejores métricas (se muestran los resultados más adelante).

Adicionalmente, se estudiaron otros algoritmos como ***DBSCAN*** sin embargo no es uno de los más pertinentes debido a que divide los clústeres en un mismo grupo separado de los atípicos, y otro de los que se trató fue el ***Gaussian Mixture***, pero, como se mencionaba anteriormente los datos no cuentan con una distribución normal, por lo que no sería adecuado adaptarlo a este tipo de algoritmos, a pesar de que se podrían realizar transformaciones. Sin embrago, el modelo *K-means*, cuenta con información suficiente para segmentar los grupos.

Resultados de las métricas:

	CANTIDAD DE CLÚSTERES					
	Escalado		Reducido		Escalado	Reducido
Clústeres	**3**	**4**	**3**	**4**	**4**	**1**
Inercia	38302	32582	33768	28060		
Silueta	0,25	0,28	0,27	0,3	0,18	0,33
Calinski	1781	1833	2016	2124	1414	112
	K-Means				***Jerárquico***	***DBSCAN***

Basándonos en los resultados de las métricas se puede evidenciar como el trabajar con un dataset reducido utilizando el método de K-Means se pueden obtener mejores resultados en cuanto a la segmentación de clientes.

Teniendo en cuenta esto se hace entonces una nueva reducción de variables teniendo en cuenta solo los 3 principales componentes identificados anteriormente y con esto se busca mejorar las métricas.

Al usar solamente los principales componentes, y aplicando nuevamente las el algoritmo de K-means, se identifican nuevamente 4 grupos, corroborando la cantidad de grupos que se deberían usar, y dando como resultado las siguientes métricas; Inercia: 28060, Silueta: 0.42 y Calinski: 4411, mejorando significativamente los resultados de las métricas. Ya teniendo entonces identificados estos clústeres, se hace la interpretación final, basada en 4 tipos de clientes.

Clúster 0: Conservadores

Los clientes en este clúster tienen un comportamiento financiero conservador pero moderado. Mantienen un saldo restante en su cuenta para hacer compras y actualizan su saldo con una frecuencia moderada. Aunque tienen acceso a efectivo anticipado y un límite de crédito promedio, tienden a realizar compras de forma poco frecuente y prefieren no pagar anticipadamente el efectivo. Sus compras suelen ser de bajo a moderado valor, tanto en cuotas como en compras únicas. Además, muestran un porcentaje de pago completo y una duración de servicio moderados.

Clúster 1: Estándar

Este clúster está compuesto por clientes que actualizan frecuentemente su saldo y mantienen un saldo restante bajo en su cuenta para hacer compras. A pesar de tener acceso a efectivo anticipado y un límite de crédito promedio, tienden a realizar compras de bajo valor y no muestran un patrón de compra frecuente. Sin embargo, muestran un comportamiento de pago moderado y una duración de servicio alta, lo que indica una relación estable con la institución financiera.

Clúster 2: Premium

Los clientes en este clúster son aquellos que realizan compras frecuentes y de alto valor. Mantienen un saldo restante alto en su cuenta y actualizan su saldo con frecuencia. Aunque tienen acceso a efectivo anticipado y un límite de crédito alto, prefieren realizar compras únicas y en cuotas de alto valor. Además, muestran un porcentaje de pago completo y una duración de servicio alta, lo que sugiere una relación sólida y rentable para la institución financiera.

Clúster 3: Alta liquidez

Este clúster está conformado por clientes que utilizan frecuentemente el efectivo anticipado y mantienen un saldo restante muy alto en su cuenta. Aunque tienen un límite de crédito alto, muestran un comportamiento de pago moderado y tienden a realizar compras de alto valor, tanto en compras únicas como en cuotas. Sin embargo, su frecuencia de compra es baja, lo que indica una preferencia por el efectivo anticipado en lugar de compras frecuentes. Su duración de servicio es alta, lo que sugiere una relación estable con la institución financiera.

Estrategias propuestas:

1. **Clúster 0: Conservadores**

- Resaltar la seguridad y estabilidad de las transacciones, promoviendo la tranquilidad de mantener un saldo moderado y no utilizar el efectivo anticipado.
- Ofrecer recursos y consejos sobre cómo utilizar de manera eficiente el crédito y maximizar los beneficios de sus tarjetas de crédito.
- Ofrecer incentivos para aumentar la frecuencia de compras, como descuentos exclusivos o puntos de recompensa por compras frecuentes.

2. **Clúster 1: Estándar**

- Ofrecer promociones y ofertas personalizadas basadas en su historial de compras, centrándose en productos o servicios de bajo valor pero de alta relevancia para ellos.
- Fomentar la lealtad a través de programas de recompensas o beneficios exclusivos para clientes habituales, incentivándolos a seguir utilizando su tarjeta de crédito.
- Implementar opciones de pago flexibles y recordatorios de pago para garantizar que se mantengan al día con sus pagos y mantener una buena relación con el cliente.

3. **Clúster 2: Premium**

- Ofrecer experiencias exclusivas, beneficios VIP y acceso anticipado a productos o eventos especiales para resaltar su estatus premium.
- Proporcionar atención al cliente de alta calidad y asesoramiento financiero personalizado para satisfacer sus necesidades específicas.
- Promocionar productos o servicios de lujo que se alineen con su comportamiento de compra de alto valor, ofreciendo incentivos adicionales para compras frecuentes.

4. **Clúster 3: Alta liquidez**

- Presentar opciones de inversión y productos financieros que les permitan maximizar el rendimiento de su saldo alto y su preferencia por la liquidez.
- Ofrecer incentivos para utilizar menos efectivo anticipado y realizar más compras con tarjeta, como bonificaciones o recompensas exclusivas.
- Proporcionar servicios adicionales para la gestión de su alto saldo, como líneas de crédito flexibles o planes de ahorro personalizados.

Recomendaciones:

Para aprovechar al máximo el análisis de datos realizado y contribuir al logro de los objetivos de crecimiento y excelencia en el servicio al cliente de BankCard, se formulan las siguientes recomendaciones:

- Se podrían implementar estrategias o algoritmos adicionales para el tratamiento de datos atípicos debido a que esta información puede ser valiosa para la segmentación de clientes.
- Se deberían incluir algunas variables categóricas para comprender un poco más la situación de los clientes, variables como el género o la edad podrían ser valiosas para una interpretación posterior.
- Explorar el comportamiento de variables aplicando transformaciones, como la logarítmica para usar otros algoritmos, como, por ejemplo, el algoritmo de mezclas gaussianas.

Es esencial mejorar la experiencia del cliente mediante la omnicanalidad y una atención proactiva, lo que podría lograrse mediante la implementación de sistemas integrados y un servicio al cliente más anticipado.