**Universidad Nacional Autónoma de México**

Facultad de Estudios Superiores Acatlán


Diplomado en Ciencia de Datos

GRUPO 11 | MÓDULO I

---

# Modelo de fuga aplicado a clientes de tarjetas de crédito

---

22 de septiembre de 2022

Varela López Ana Karen

anakarenvl99@gmail.com

# Índice general

# 1. Introducción

Las tarjetas de crédito se han convertido en un método de pago muy popular alrededor del mundo. Gracias a este instrumento financiero, podemos permitirnos la compra de ciertos productos sin tener un ahorro materealizado, con el compromiso de devolver el préstamo más una tasa de interés. Debido a su gran popularidad, el sector bancario a dirigido sus recursos al desarrollo de nuevo productos de crédito con beneficios atractivos para el cliente y que repercute en los ingresos del banco pues entre más préstamos seguros haga, mayor la tasa de retorno que ganará. No obstante, los clientes suelen abandonar los servicios de crédito aún sin terminar de pagar el préstamo, aumentando la tasa de pérdidas del banco, he aquí la importancia de controlar el riesgo de crédito.

## 1.1. Objetivo

El objetivo de este proyecto es encontrar uno o varios modelos que nos permitan predecir la cantidad de clientes buenos (continúa con el servicio) y malos (abandona el servicio) en la cartera con el propósito de controlar el riesgo de crédito de acuerdo al nivel de riesgo determinao por el Potafolio Manager[1].

Para el desarrollo del modelo he utilizado un conjunto de datos sobre las tarjetas de crédito del banco 'X' el cual puede consultar en la siguiente liga: https://www.kaggle.com/sakshigoyal7/credit-card-customers. El proyecto se presentará de acuerdo a las siguientes etapas:

- **Etapa I:** Preparación de tabla analítica de datos.
- **Etapa II:** Ajuste de modelo supervisado.
- **Etapa III:** Ajuste de modelo no supervisado.

[1]Equipo de analistas que determinan el nivel de riesgo objetivo de acuerdo a las políticas de la empresa y las reservas financieras.

## 1.2. Diccionario de datos

Los datos que utilizaron se componen de 23 variables y 10127 observaciones. De las 23 variables, el autor de los datos ha especificado que 2 no son de utilidad, por tanto la matriz queda reducida a 21 variable.

Antes de gestionar la calidad de datos, se reetiquetó cada variable de acuerdo al tipo de dato como sigue:

- Numérico ç_": discreteta y continua
- Categóricas "v_": nominal y ordinal
- ID ïd_"

| **Variable** | **Tipo** | **Descipción** |
|---|---|---|
| CLIENTNUM | id | Identificador único de cliente |
| Attrition_Flag | nominal | Estado de la cuenta |
| Customer_Age | discreto | Edad del cliente |
| Gender | nominal | Sexo del cliente (F = female, M = Male) |
| Dependent_count | ordinal | Número de dependientes económicos |
| Education_Level | ordinal | Nivel máximo de estudios |
| Marital_Status | ordinal | Estado civil del cliente |
| Income_Category | ordinal | Ingreso anual del cliente |
| Card_Category | nominal | Tipo de tarjeta |
| Months_on_book | discreta | Número de meses que el cliente tiene en el banco |
| Total_Relationship_Count | ordinal | Número de productos bancarios a nombre del cliente |
| Months_Inactive_12_mon | ordinal | Número de meses inactivo en los últimos 12 meses |
| Contacts_Count_12_mon | ordinal | Cantidad de veces que se ha utilizado la tarjeta en últimos 12 meses |
| Credit_Limit | continua | Límite de crédito de la tarjeta |
| Total_Revolving_Bal | continua | Saldo rotatorio total en la tarjeta |
| Avg_Open_To_Buy | continua | Línea de crédito abierta para comprar (promedio de los últimos 12 meses) |
| Total_Amt_Chng_Q4_Q1 | continua | Cambio en el monto de la transacción (Q4 sobre Q1) |
| Total_Trans_Amt | continua | Importe total de la transacción (últimos 12 meses) |
| Total_Trans_Ct | continua | Recuento total de transacciones (últimos 12 meses) |
| Total_Ct_Chng_Q4_Q1 | continua | Cambio en el recuento de transacciones (Q4 sobre Q1) |
| Avg_Utilization_Ratio | continua | Ratio de uso promedio de la tarjeta |

# 2. Calidad de datos

El propósito de esta sección es revisar y corregir las variables que tengan algún problema de duplicidad, completitud, orden y/o consistencia. Esta sección es escencial en al análisis de datos para poder obtener datos confiables para entrenar nuestro modelo.

## 2.1. Unicidad

Se revisó que no hubiera clientes repetidos por ID, siendo este un identificador único por cliente, se espera que no haya registros duplicados. Se encontró que todos los registros son únicos.

## 2.2. Completitud

Se calculó la cantidad de valores faltantes en cada variable tanto en casos como en porcentaje y se concluyó que todas las variables están completas al 100% por lo que no hay necesidad de imputar valores nulos.

| | columna | total | completitud(%) |
|---|---|---|---|
| 0 | id_Client_Num | 0 | 100.0 |
| 1 | c_Total_Trans_Ct | 0 | 100.0 |
| 2 | c_Total_Trans_Amt | 0 | 100.0 |
| 3 | c_Total_Amt_Chng_Q4_Q1 | 0 | 100.0 |
| 4 | c_Avg_Open_To_Buy | 0 | 100.0 |
| 5 | c_Total_Revolving_Bal | 0 | 100.0 |
| 6 | c_Credit_Limit | 0 | 100.0 |
| 7 | c_Contacts_Count_12_mon | 0 | 100.0 |
| 8 | c_Months_Inactive_12_mon | 0 | 100.0 |
| 9 | c_Total_Ct_Chng_Q4_Q1 | 0 | 100.0 |
| 10 | c_Total_Relationship_Count | 0 | 100.0 |
| 11 | v_Card_Category | 0 | 100.0 |
| 12 | v_Income_Category | 0 | 100.0 |
| 13 | v_Marital_Status | 0 | 100.0 |
| 14 | v_Education_Level | 0 | 100.0 |
| 15 | c_Dependent_count | 0 | 100.0 |
| 16 | v_Gender | 0 | 100.0 |
| 17 | c_Customer_Age | 0 | 100.0 |
| 18 | v_Attrition_Flag | 0 | 100.0 |
| 19 | c_Months_on_book | 0 | 100.0 |
| 20 | c_Avg_Utilization_Ratio | 0 | 100.0 |

**Figura 2.1:** Tabla de completitud

## 2.3. Orden

Se verificó que las variables numéricas tomaran valores válidos, es decir, que fueran positivos y siguieran la naturaleza de la variable. Adicionalemnte, se cambió el tipo de dato a float para asegurar que todos los valores sean del tipo numérico. En caso de haber strings, el cambio de tipo de dato marcaría error, sin embargo, no hubo dificultad alguna.

En el caso de las variables categóricas, se revisó que los valores que tomaran fueran del tipo string y se mantuvieran dentro de ciertas categorías, sin desviaciones. Adicionalmente se dió formato en minúsculas y sin espacios extra a cada registro. No se hayaron datos anómalos.

## 2.4. Consistencia

Primero se revisó que la cantidad de productos bancarios al nombre del cliente (c_Total- _Relationship_Count) no fuera cero pues si el cliente pertenece a estos registros es porque al menos tiene una tarjeta de crédito.

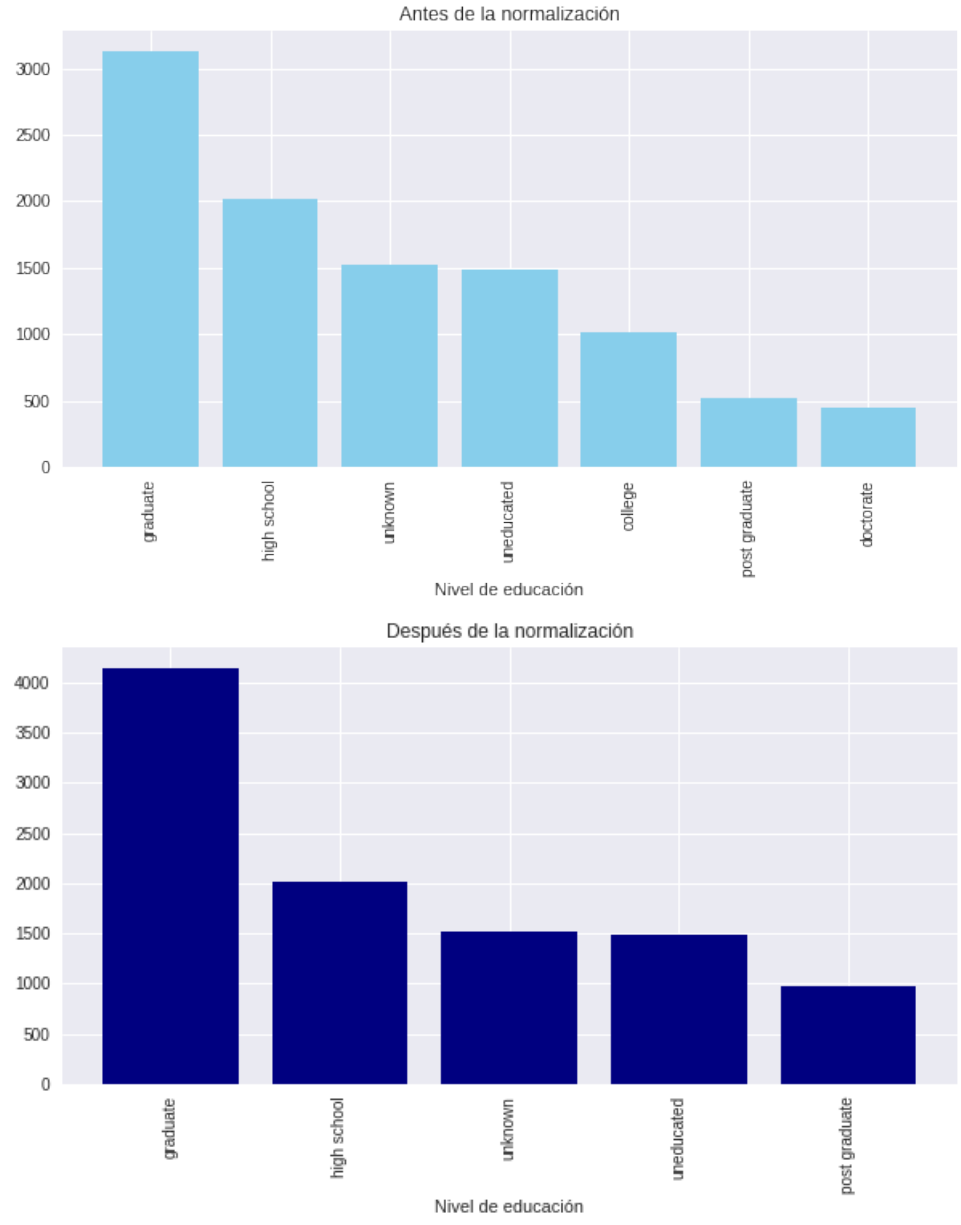
Posteriormente se verificó que el número de meses inactivo en el último año (c_Months_Inactive- _12_mon) no fuera mayor a 12, pues como lo indica la descripción de la variable, solo captura lo ocurrido en los últimos meses.

Así mismo se probó que ninguna de las variables tome valores negativos pues al ser préstamos o se debe o se ha saldado la deuda pero no se puede tener saldo negativo.

Finalmente se verificó que la edad de los clientes fuera mayor a 18 años, que es la edad mínima para tramitar una tarjeta de crédito, y que fuera menor a 100 años; claro que debemos revisar las políticas de la empresa para identificar la edad máxima para tramitar un crédito, sin embargo no tenemos acceso a esas políticas. En ninguno de los casos se encontraron anomalías.

## 2.5. Normalización de variables

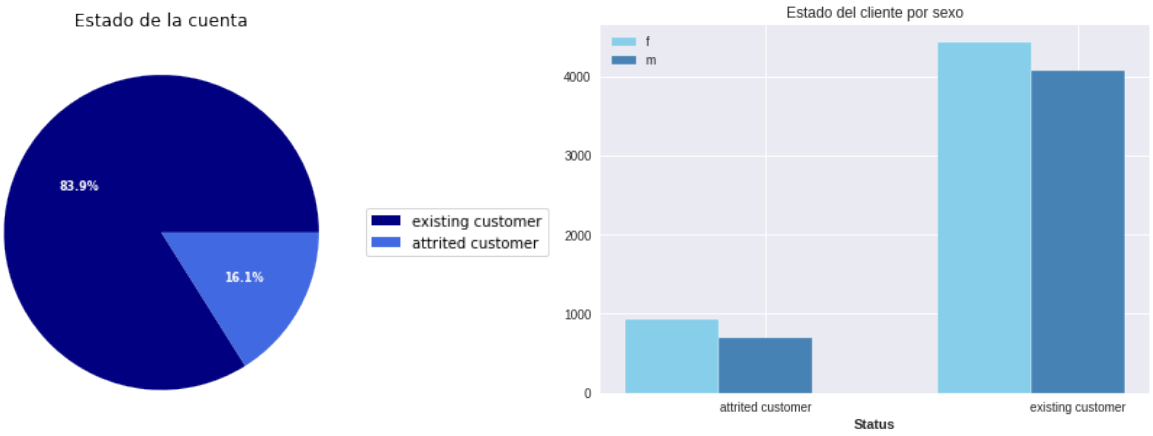
Se reetiquetaron ciertas categorías para reducir ambigüedades. La mayoría de variables tiene categorías precisas, a excepción de v_Education_Level la cual toma las categorías graduate y college, así como post graduate y doctorate, y ambas hacen referencia al mismo grado educacional, por ello fueron colapsadas en dos categorías, graduate y post graduate. A continuación se muestra la variable antes y después de la normalización:



# 3. Análisis exploratorio

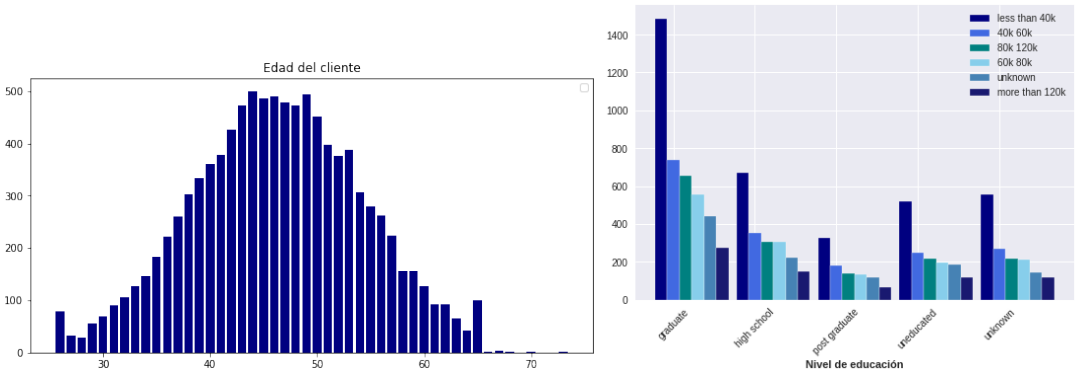
El propósito de esta sección es mostrar la información que podemos obetener de cada variable. Se utilizaron gráficos de barras, histogramas, boxplot, scatterplot y gráficos de pie para mostrar hechos relevantes.

## 3.1. Relación entre el estado de la cuenta y el sexo del cliente



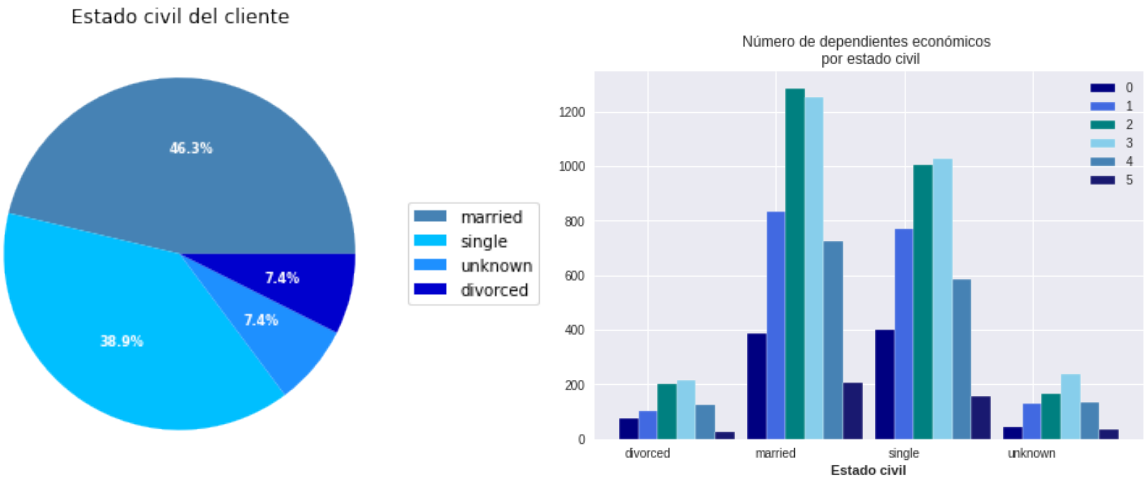

Del total de registros, el 83.9% continúan con el servicio y 16.1% han suspendido su tarjeta de crédito. De los clientes que continúan con el servicio, el 52.09% son mujeres y de los clientes que cancelaron su tarjeta representan el 57.17%; es decir, en general tenemos más clientes mujeres que varones.

## 3.2. Relación entre la edad, el nivel de estudios y los ingresos anuales del cliente



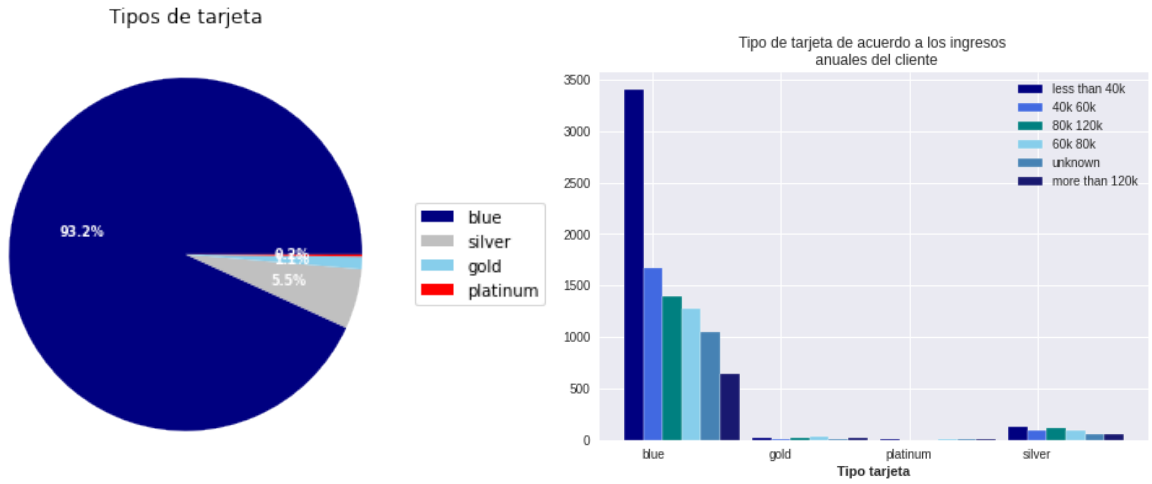

En promedio los clientes tienen 46 años, alcanzando un máximo de 73 y mínimo de 26 años. Por otro lado, la mayoría de usuarios estudió hasta la educación superior y perciben ingresos anuales menores a 40k, no obstante el 0.08% de los clientes sin educación recibe ingresos superiores a 120k, un porcentaje mayor a aquellos que estudiaron un posgrado y reciben los mismos ingresos, por tanto, no podemos asumir que un mayor nivel de estudios es sinónimo de mejores ingresos.

## 3.3. Relación entre el estado civil y el número de dependientes económicos del cliente



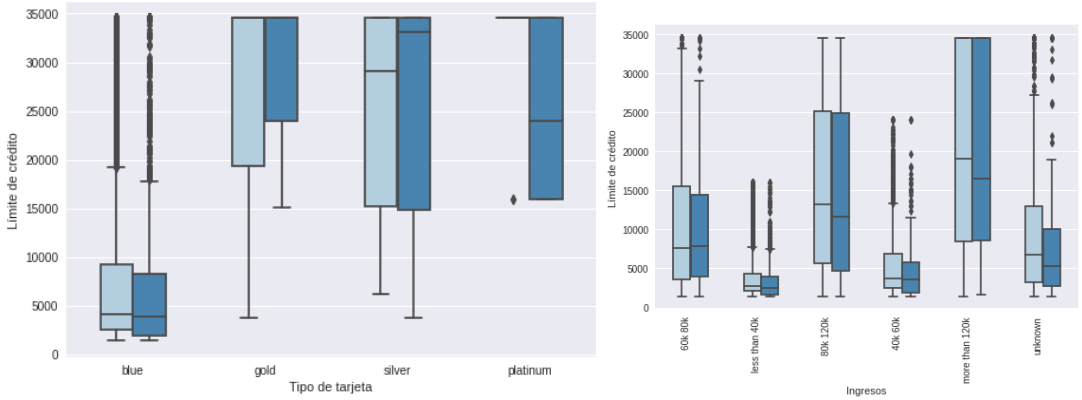

El 46.3% de la cartera está casada, mientras que el 38.9% está soltero, y el resto está divorciado o se deconoce el estado civil. En general, los clientes tienen en promedio de 2 a 3 dependientes económicos sin importar si están casados o no y tienen como máximo 5 dependientes; por tanto, no podemos suponer que una persona soltera tiene menos responabilidades que una casada.

## 3.4. Relación entre el tipo de tarjeta y los ingresos anuales del cliente



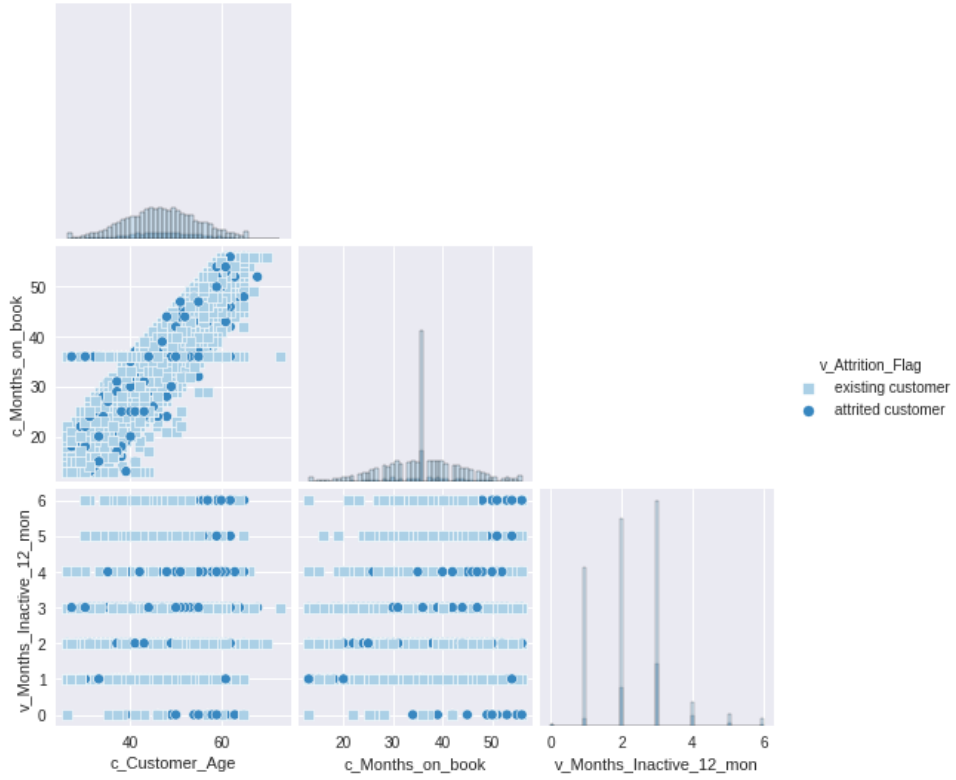

El tipo de tarjeta que predomina es la azul (blue), representando el 93.2% del total de datos, mientras que la menos solicitada es la platinum con el 0.19%. Como se muestra en el grafico de la derecha, los ingresos anuales que predominan son menores a 40k en clientes con tarjeta azul y plateada (silver), por otro lado, la mayoría de los usuarios con tarjeta oro (gold) reciben ingresos entre 60k y 80k. Así pues, podemos notar relación entre el tipo de tarjeta y los ingresos anuales. *Cabe destacar que la segunda imágen está compuesta por dos gráficos a diferente escala con el propósito de notar todos los datos.*

## 3.5. Relación entre el límite de crédito, el tipo de tarjeta y los ingresos anuales




Como se sospechaba, la tarjeta azul es básica ya que su limite de crédito promedio es de $7,363.8, mientras que el límite promedio de la tarjeta plateada es $25,277.84, de la oro es $28,416.37 y de la platinum es $30,283.45. Por otro lado, el límite de crédito está correlacionado con el monto de ingresos anuales pues aquellos clientes cuyo ingreso es menor a 40k tienen un límite de crédito menor que aquellos cuyos ingresos son mayores a 120k.

## 3.6. Relación entre los meses que el cliente lleva con el servicio, la edad y los meses inactivo en el último año




En este scatterplot podemos observar una clara correlación positiva entre la edad del cliente (c_Customer_Age) y los meses que el usuario ha mantenido relación con el banco (c_Months_on_Book), es decir, nuestros clientes más grandes han estado en nuestra cartera por más tiempo; por lo que podemos suponer que mantenemos a la mayoría de los usuarios a lo largo del tiempo.

Así mismo, observamos que la mayoría de los clientes que abandonan la cartera estuvieron inactivos por más de 3 meses en el último año. Además, los clientes más jóvenes tienden a estar inactivos de 1 a 4 meses.

Finalmente, se concluyó que en promedio los clientes permancen en la cartera por 3 años (35.9 meses) y en el útimo año han estado inactivos en promedio por 2.3 meses, alcanzando un máximo de hasta 6 meses.

En los siguientes capítulos se muestran 2 enfoques distintos para preparar los datos a ocupar en el modelo de fuga.

# 4. Enfoque I

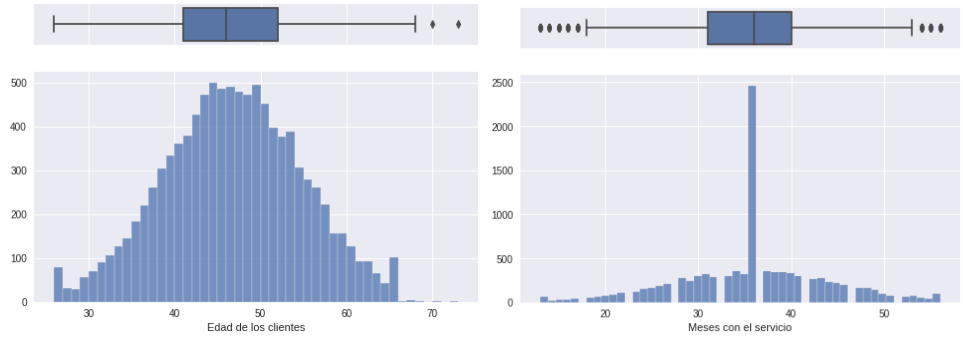
Este enfoque tratará a los datos de la manera "tradicional", es decir, se realizó el análisis de valores atípicos, ingeniería de variables,reducción de la dimensión de la matriz por dos métodos, filtro de alta correlación y multicolinealidad.

## 4.1. Outliers

Mejor conocidos como valores atípicos, son aquellas observaciones que no siguen el comportamiento natural de la variable. En esta sección solo se analizaron variables continuas pues las categóricas solo toman valores dentro de ciertas opciones que ya se corrigieron en la sección de normalización de variables.

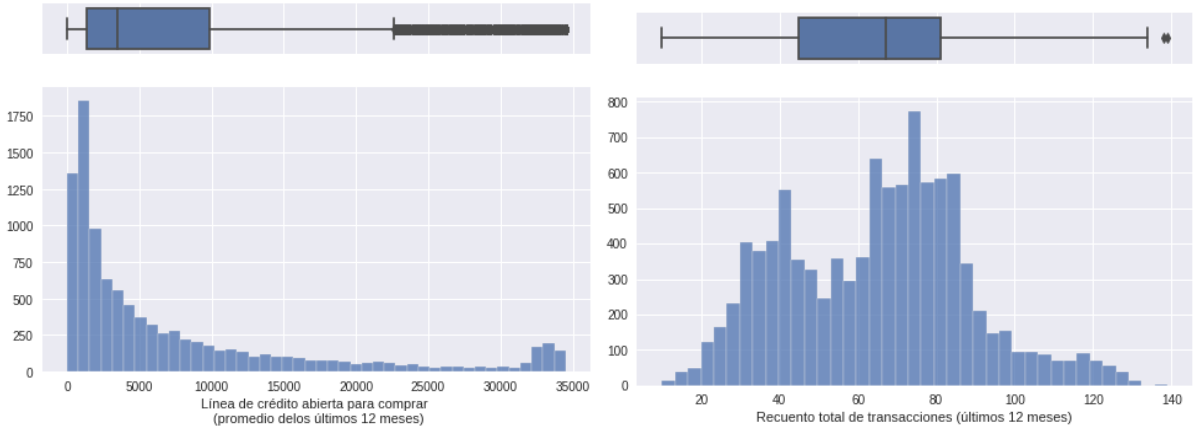
Utilizamos tres métodos para identificar outliers, los métodos univariados Inter Quantil Range (IQR) donde el intervalo de aceptación se basa en la fórmula (Q1 - 1.5IQR, Q3 + 1.5IQR) y Percentiles cuyo intervalo de aceptación es $(q_5, q_{95})$; y el método multivariado Ïsolation Forest".

De las 12 variables numéricas, 7 fueron identificadas con outliers, sin embargo, al graficar un boxplot e histograma se concluyó que no todos los valores son realmente atípicos. A continuación el debate:




La variable c_Customer_Age (edad del cliente), fue identificada con 2 outliers,

sin embargo, edades mayores a 70 años y menores a 80, son totalmente posibles, por ello se optó pot precerbar todos los datos. Similar ocurrió con la variable c_Months_on_Book (meses en el servicio), cuyas colas pesadas fueron identificadas como valores atípicos, sin embargo estos valores son totalmente posibles así que también se conservaron todos los datos.




La variable c_Avg_Open_To_Buy (Línea de crédito abierta para comprar, promedio delos últimos 12 meses) fue identificada con 507 outliers, sin embargo, al comparar con el límite de crédito concluímos que el máximo límite de crédito posible era de aproximadamente $35,000, por tanto es posible que la línea de crédito abierta promedio tome valores de hasta $35,000, así que se conservaron todos los datos, Respecto a la variable c_Total_Trans_Ct (número total de transacciones en el último año) se encontraron 2 datos atípicos que tomaban valores de hasta 140, pero al ser este un valor anual, el equivalente de transacciones por día es de 2 a 3 transacciones por semana que es completamente posible, por tanto se preservaron todas las observaciones.

Al hacer este análisis, determinamos que solo 3 variables presentan valores atípicos ya que tienen una varianza grande que podría afectar el ajuste del modelo. Si eliminábamos todos los outliers perdía el 11.71% de mis datos, y dado que los registros son limitados, decidimos imputarlos como se presenta en la siguiente sección.

### 4.1.1. Imputación y eliminación de outliers por el método univariado

Se identificaron como outliers aquellos valores que bajo las dos métricas, IQR y percentiles, fueron seleccinados como tal, de esta manera tenemos mayor certeza de que realemnte son valores atípicos.

Para imputar los outliers dividímos los datos en dos, el conjunto de prueba (X_test) con el 20% de los datos tomados de manera aleatoria, y el conjunto de entrenamiento (X_train) representando el 80%. Con el conjunto de entrenamiento probamos qué método de imputación es mejor entre media, mediana y moda de acuerdo a la prueba de bondad de ajuste Komolgorov-Smirnov, cuyas hipótesis son:

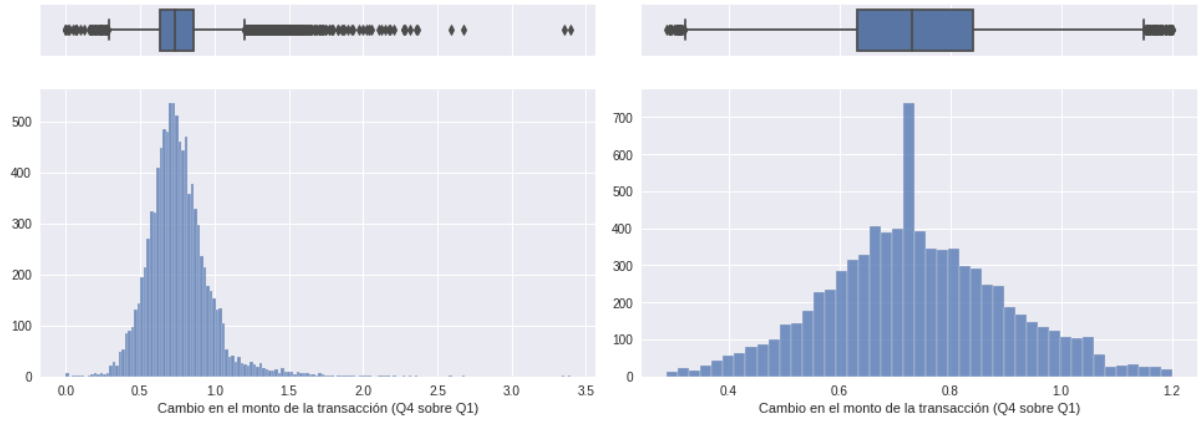
$$H_0 : Ambas\ muestras\ pertenecen\ a\ la\ misma\ distribución$$

$$H_1 : Las\ muestras\ provienen\ de\ distribuciones\ distintas$$

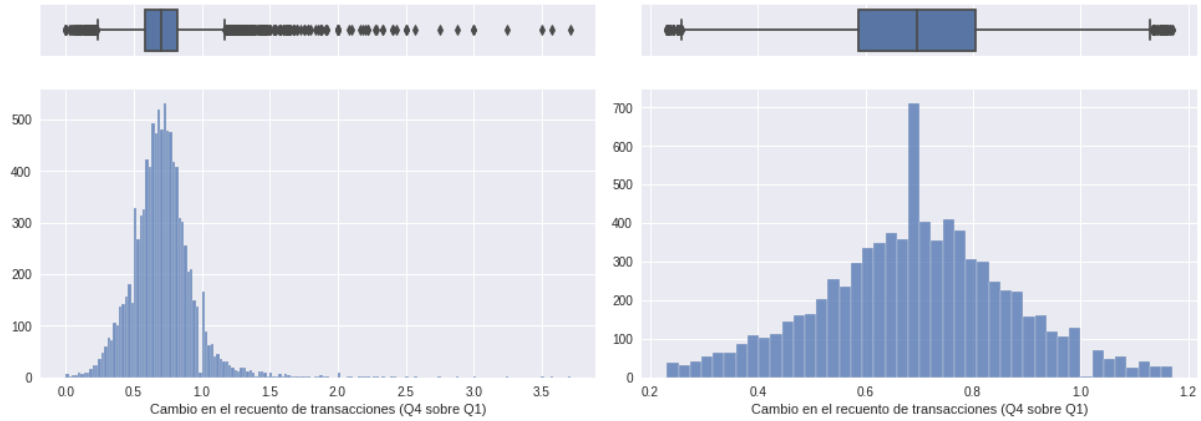
Para hacer esta prueba comparamos la distribución antes de imputar los datos y después de hacerlo. Encontramos que para las variables c_Total_Amt_Chng_Q4_Q1, c_Total_Ct_Chng_Q4_Q1 y c_Total_Trans_Amt cualquiera de los métodos es aceptado pues obtenemos un p-value=1, que es mayor al nivel de significancia de 5%, por tanto, no hay pruebas suficientes para rechazar $H_0$. Elegimos la media para todos los casos. El proceso de imputación fue:

1) Convertir outliers en datos nulos.
2) Imputar los nulos por los tres métodos, media, mediana y moda, y comparar con la muestra antes de imputar y sin valores nulos.
3) De acuerdo al p-value de la prueba K-S, determinar qué método utilizar e imputar los datos.

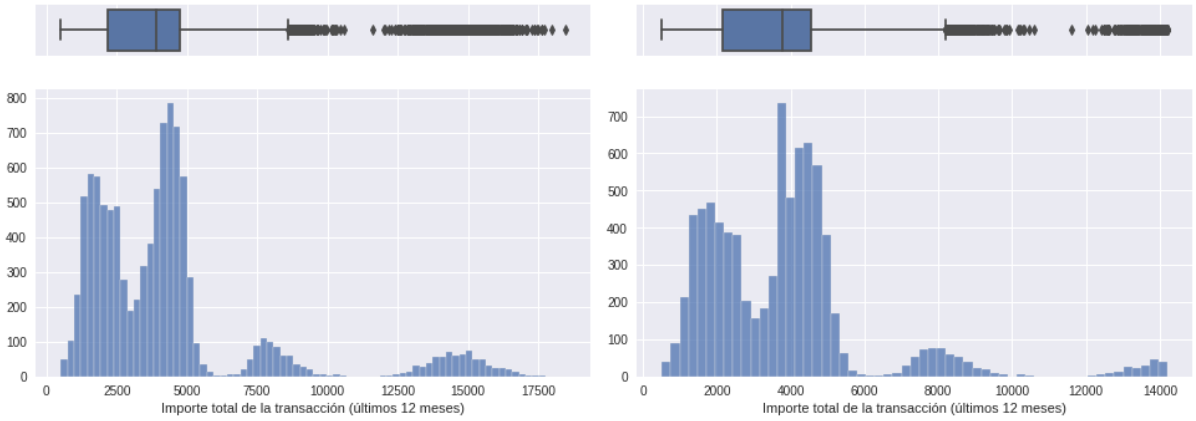
A continuación se muestran las variables antes de imputar (izquierda) y después de la imputación (derecha).




En la variable c_Total_Amt_Chng_Q4_Q1 se encontraron 395 valores atípicos que representan el 3.9% del total de datos, considerando los intervalos IQR = [0.29, 1.2] y Percentil = [0.46, 1.1]. Se asigno el valor de 0.7599 a los datos imputados,




La variable c_Total_Ct_Chng_Q4_Q1 presentaba 396 valores atípicos, que representaban el 3.91% de los datos. Consideré los intervalos IQR = [0.23, 1.17] y Percentil = [0.37, 1.07] para detectar outliers. Los datos imputados tomaron el valor 0.7122.

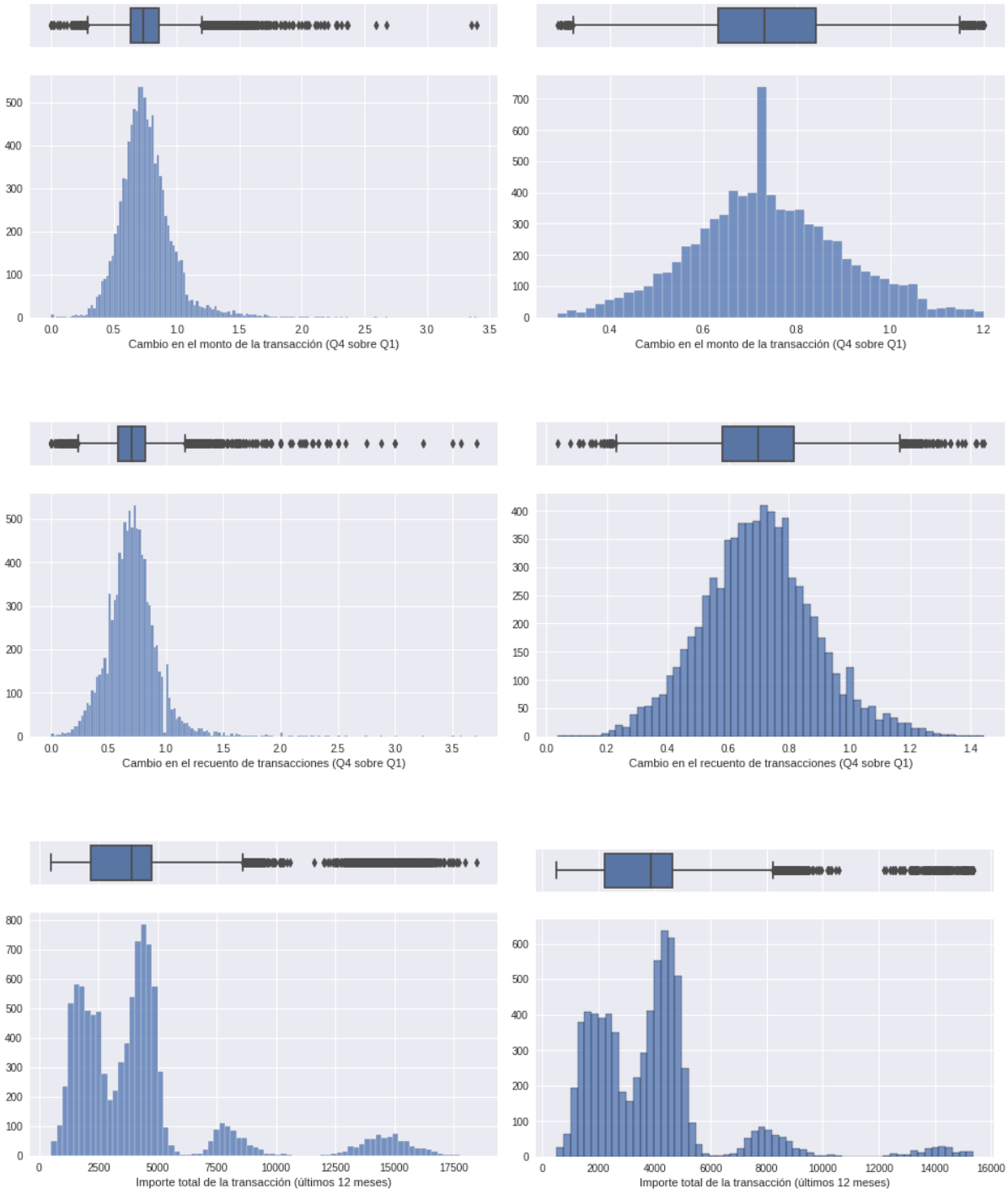



La variable c_Total_Trans_Amt presentó la mayor cantidad de valores atípicos con 506 datos, que representan el 5% del conjunto. Consideramos los intervalos IQR = [-1722.75, 8619.25] y Percentil = [1283.3, 14212] para identificar outliers, sin embargo dado que la variable no toma valores negativos, el intervalo de confianza se redujo a [0, 14212]. El valor a sustiituir fue 4404.086.

A partir de esta sección todas las moficaciones que haga se harán tanto al conjunto de entrenamiento como al conjunto de prueba.

### 4.1.2. Eliminación de outliers por el método multivariado

Este análisis solo se le aplicó a las 3 variables que se determinó como posible la existencia de valores atípicos. El método utilizado fue Isolation Forest que consiste en formar árboles de manera aleatoria hasta aislar la proporción de valores atípicos que se le indique al algorítmo mediante el parámtero de contaminación, que en nuestro análisis fue de 10%. A continuación se uestran las variables antes (imágen a la izquierda) y después (imágen a la derecha)de usar esta ténica:




Mediante esta técnica entrenamos el modelo para detectar valores atípicos con el conjunto de entrenamiento y se aplicó en ambas muestras, entrenamiento y prueba, sin embargo en este apartado solo se muestran los cambios sobre la muestra de entrenamiento. Finalmente decidimos ocupar este método ya

que si bien se eliminó un porcentaje mayor de datos, la detección de outliers es más precisa al considerar todas las variables en conjunto a la hora de detectar valores anómalos.

## 4.2. Ingenieria de variables categóricas

| One-hot Encoding | Codificación ordinal | Count Encoding |
|---|---|---|
| **v_Attrition_Flag**<br><br>1 existing customer<br>0 etrrited customer | **v_Income_Category**<br>0 unknown<br>1 less than 40k<br>2 40k-60k<br>3 60k-80k<br>4 80k-120k<br>5 more than 120k | **v_Education_Level**<br><br>3286 graduate<br>1618 high school<br>1223 unknown<br>1212 uneducated<br>762 post graduate |
| **v_Gender**<br><br>1 f<br>0 m | | **v_Marital_Status**<br>3757 married<br>3155 single<br>598 divorce<br>591 unknown |
| | | **v_Card_Category**<br>7537 blue<br>457 silver<br>91 gold<br>16 platinum |

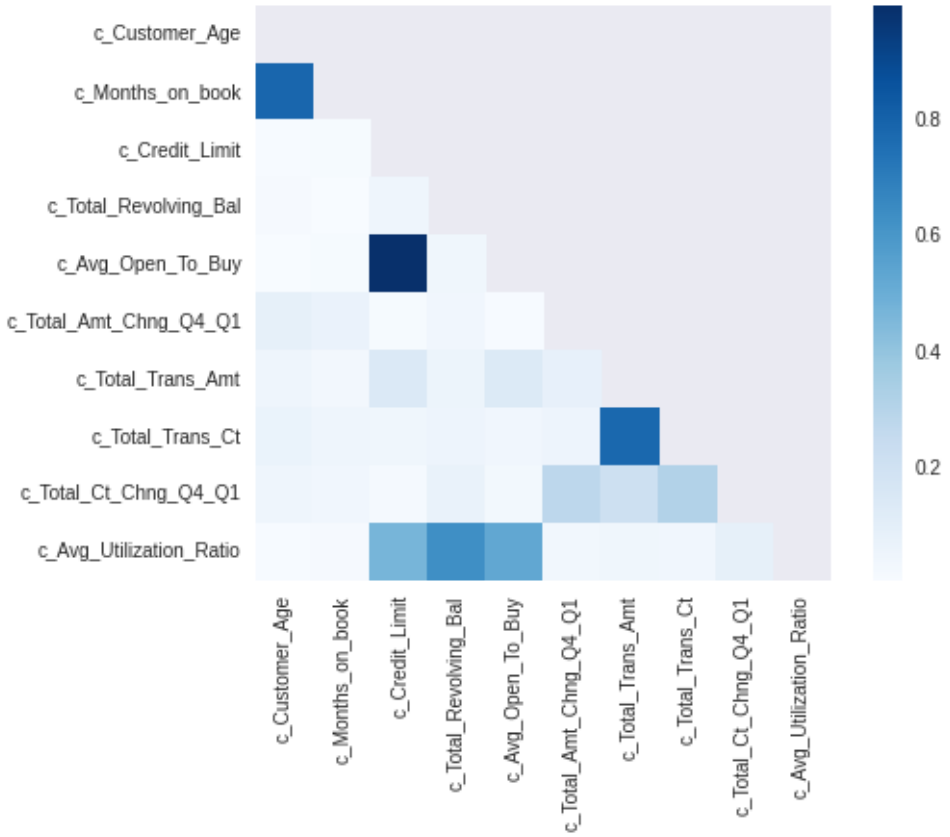
## 4.3. Reducción de dimensionalidad

El propósito de esta sección es identifcar las variables que están correlacionadas entre sí y no explican a la variable de interés v_Attrition_Flag. Se utilizaron dos métodos para daterminar las variables que tienen una estrecha

relación: filtro de alta correlación y multicolinealidad (VIF).

### 4.3.1. Filtro de alta correlación

En este método calculamos la matriz de correlación entre las variables numéricas, si su correlación en valor absoluto es mayor a 0.7, significa que ambas variables están capturando la misma información y es recomendable eliminar una de ellas. Para decidir qué variable omitir, se analizó la correlación de las variables con el resto de ellas y se eliminó aquella con mayor correlación.




La variable c_Customer_Age y c_Months_on_Book están correlacionadas con potencia de 0.7895, y dado que la variable que contiene la edad de los clientes, está más correlacionada con el resto de variables que el número de meses que tiene el usuario con el servicio, eliminé c_Customer_Age. Similar sucedió con c_Credit_Limit y c_Avg_Oopen_To_Buy pues su nivel de correlación

era de 0.9958, y dado que c_Avg_Oopen_To_Buy estaba más corralcionada con el resto de variables, se omitió del dataset. Finalmente la correlación entre c_Total_Trans_Amt y c_otal_Trans_Ct fue de 0.7824 y dado que la primera está más relacionada con el resto de variables, se decidió eliminarla.

### 4.3.2. Multicolinealidad

La multicolinealidad ocurre cuando dos variables o más están correlacionadas en un modelo de regresión, si esto sucede, estaríamos considerando más características de las necesarias para explicar a la variable dependiente, lo que hace que nuestro modelo tome más tiempo de entrenamiento y sea redundante.

Para datectar multicolinealidad utilizamos el factor de inflación de la varianza (o VIF por sus siglas en inglés) el cual mide hasta qué punto la varianza de un coeficiente de regresión se incementa a causa de la colinealidad. En el caso de las variables numéricas se considera alta colinealidad si el VIF es mayor a 10. El proceso para corregir la colinealidad es:

1) Calcular el VIF
2) Identificar la variable con VIF mayor, si es superior a 10, eliminar la variable
3) Repetir el paso 1 y 2 hasta que todas las variables tengan VIF menor a 10

| variables | VIF |
|---|---|
| c_Total_Ct_Chng_Q4_Q1 | 16.076419 |
| c_Total_Amt_Chng_Q4_Q1 | 13.745578 |
| c_Months_on_book | 11.795055 |

**(a)**

| variables | VIF |
|---|---|
| c_Total_Amt_Chng_Q4_Q1 | 12.165914 |
| c_Months_on_book | 11.997744 |

**(b)**

La primera vez que se calculó el VIF se dectaron 4 variables que presentaban multicolinealidad, sin ambargo, sólo se eliminó la de VIF mayor

y se volvió a realizar el proceso; finalmente solo se eliminaron 2 variables: c_Total_Amt_Chng_Q4_Q1 y c_Total_Ct_Chng_Q4_Q1.

En el caso de las variables categóricas se utilizó la métrica de VIF generalizado para detectar multicolinealidad. En el siguiente gráfico se muestra que los GVIF son menores a 5, por tanto ninguna fue considerada como redundante.

| factor | GVIF | Df | GVIF^(1/2Df) | VIF |
|---|---|---|---|---|
| v_Gender | 3.445391 | 1 | 1.856177 | 3.445391 |
| v_Dependent_count | 1.020390 | 1 | 1.010143 | 1.020390 |
| v_Education_Level | 1.008233 | 4 | 1.001025 | 1.002052 |
| v_Marital_Status | 1.012196 | 3 | 1.002022 | 1.004049 |
| v_Income_Category | 3.492036 | 5 | 1.133203 | 1.284150 |
| v_Card_Category | 1.027791 | 3 | 1.004579 | 1.009179 |
| v_Total_Relationship_Count | 1.014885 | 1 | 1.007415 | 1.014885 |
| v_Months_Inactive_12_mon | 1.003769 | 1 | 1.001883 | 1.003769 |
| v_Contacts_Count_12_mon | 1.009332 | 1 | 1.004655 | 1.009332 |

## 4.4. Datos finales

Al finalizar la gestión de calidad de datos, imputación de outliers, ingeniería de variables y reducción de dimensionalidad, concluímos con 2 muestras de datos, el conjunto de entrenamiento con 7311 observaciones (el 80% de los datos) y el conjunto de prueba con 1829 registros, ambas con 14 variables. Además todas las observaciones son del tipo numérico, no tienen valores nulos y la correlación entre ellas no es alta.

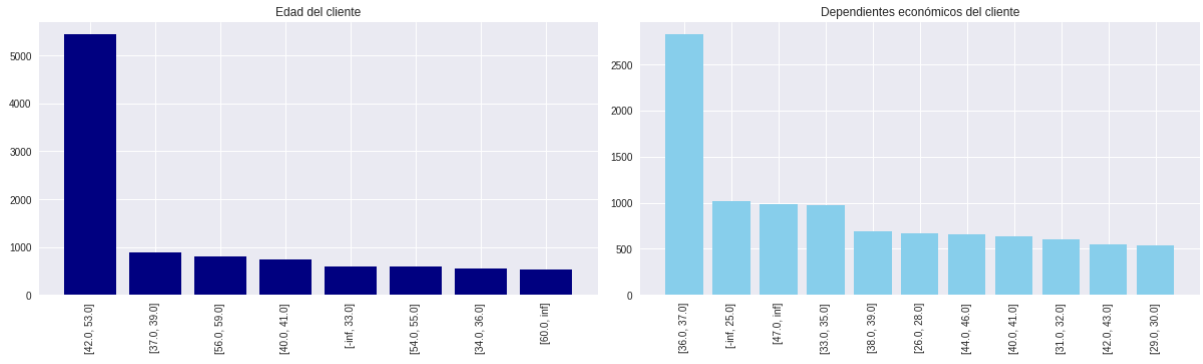
# 5. Enfoque 2

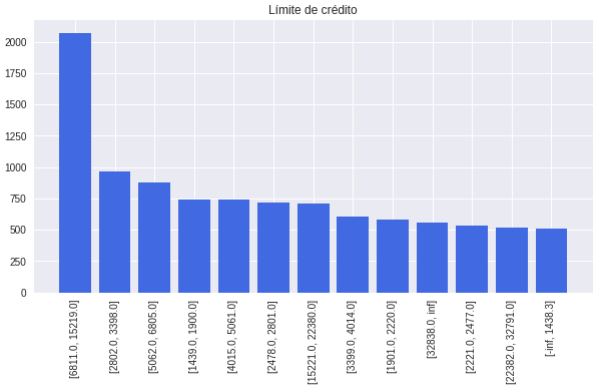
Dado que nuestra variable dependiente es dicotomica, podemos determinar qué variables son útiles con transformación entrópica; para ello convertiremos la variable objetivo en tipo dummy e imputaremos los datos considerando toda la muestra y no solo el conjunto de entrenamiento.

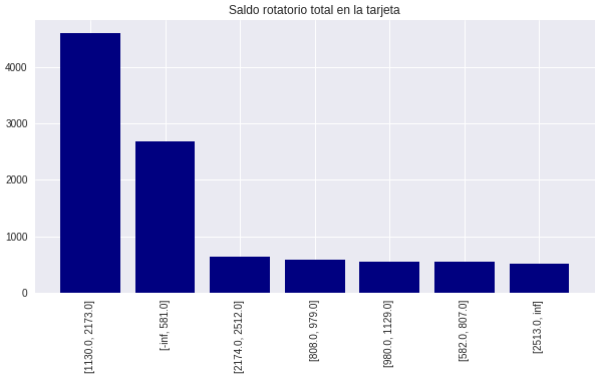
## 5.1. De continuas a categóricas

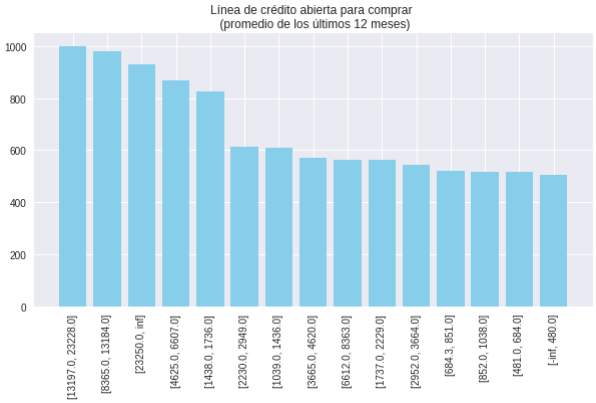
Para realizar transformación entrópica, es necesario que todas las variables sean del tipo categórico por ello convertimos las variables continuas en categóricas utilizando árbooles de decisión que nos ayudaran a definir el corte entre cada categoría.

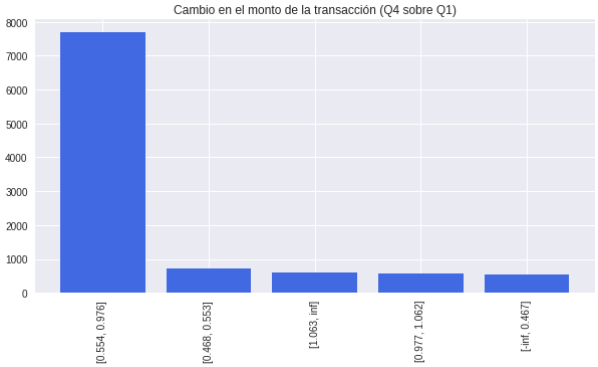
Con el objetivo que todas las categorías contengan al menos 5% de los datos y contengan tanto buenos (1= existing customer) como malos (0= attrited customer), utilizamos árboles de decisión que capturaran estos requerimientos. A continuación se muestran las primeras categorías que sugiere el árbol:

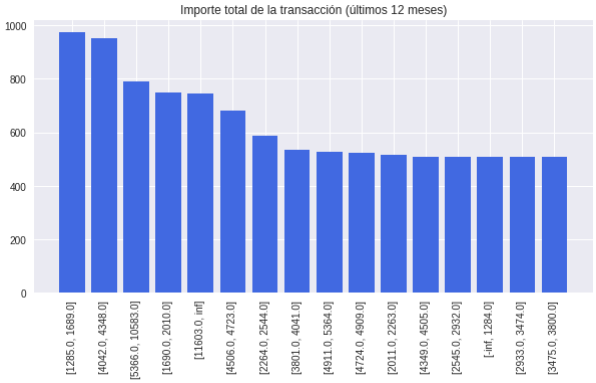


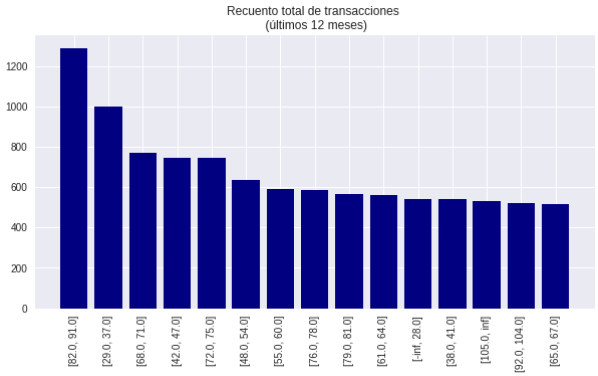

Límite de crédito
2000
1750
1500
1250
1000
750
500
250
0
[6811.0, 15219.0]
[2802.0, 3398.0]
[5062.0, 6805.0]
[1439.0, 1900.0]
[4015.0, 5061.0]
[2478.0, 2801.0]
[15221.0, 22380.0]
[3399.0, 4014.0]
[1901.0, 2220.0]
[32838.0, inf]
[2221.0, 2477.0]
[22382.0, 32791.0]
[-inf, 1438.3]

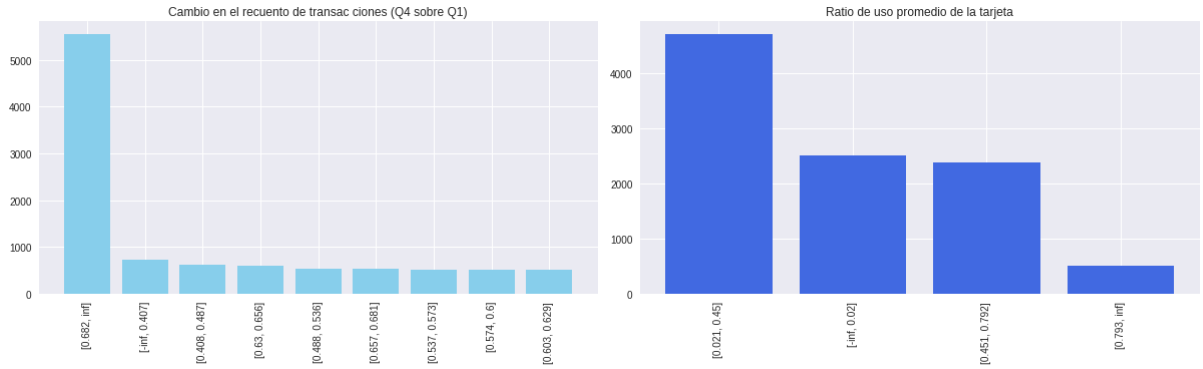

Saldo rotatorio total en la tarjeta
4000
3000
2000
1000
0
[1130.0, 2173.0]
[-inf, 581.0]
[2174.0, 2512.0]
[808.0, 979.0]
[980.0, 1129.0]
[582.0, 807.0]
[2513.0, inf]


Línea de crédito abierta para comprar
(promedio de los últimos 12 meses)
1000
800
600
400
200
0
[13197.0, 23228.0]
[8365.0, 13184.0]
[23250.0, inf]
[4625.0, 6607.0]
[1438.0, 1736.0]
[2230.0, 2949.0]
[1039.0, 1436.0]
[3665.0, 4620.0]
[6612.0, 8363.0]
[1737.0, 2229.0]
[2952.0, 3664.0]
[684.3, 851.0]
[852.0, 1038.0]
[481.0, 684.0]
[-inf, 480.0]


Cambio en el monto de la transacción (Q4 sobre Q1)
8000
7000
6000
5000
4000
3000
2000
1000
0
[0.554, 0.976]
[0.468, 0.553]
[1.063, inf]
[0.977, 1.062]
[-inf, 0.467]


Importe total de la transacción (últimos 12 meses)
1000
800
600
400
200
0
[1285.0, 1689.0]
[4042.0, 4348.0]
[5366.0, 10583.0]
[1690.0, 2010.0]
[11603.0, inf]
[4506.0, 4723.0]
[2264.0, 2544.0]
[3801.0, 4041.0]
[4911.0, 5364.0]
[4724.0, 4909.0]
[2011.0, 2263.0]
[4349.0, 4505.0]
[2545.0, 2932.0]
[-inf, 1284.0]
[2933.0, 3474.0]
[3475.0, 3800.0]


Recuento total de transacciones
(últimos 12 meses)
1200
1000
800
600
400
200
0
[82.0, 91.0]
[29.0, 37.0]
[68.0, 71.0]
[42.0, 47.0]
[72.0, 75.0]
[48.0, 54.0]
[55.0, 60.0]
[76.0, 78.0]
[79.0, 81.0]
[61.0, 64.0]
[-inf, 28.0]
[38.0, 41.0]
[105.0, inf]
[92.0, 104.0]
[65.0, 67.0]

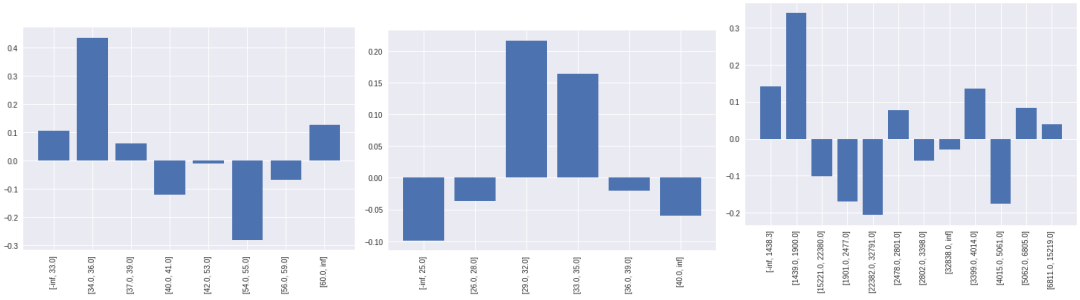



Posteriormente se dividió la muestra en conjunto de prueba (X_test1) y de entrenamiento(X_train1), estratificada por la variable objetivo (tgt_Attrited_Flag), y se calculó el peso de la evidencia (WoE por sus siglas en inglés) y valor de la información (IV por sus siglas en inglés) de acuerdo a las siguientes fórimulas:

$$\text{WoE} = \log\left(\frac{\%buenos}{\%malos}\right) \qquad IV \ \ (\%buenos - \%malos) * WoE$$

## 5.2. Corrección de categorías

Para identificar qué categorías podemos colapsar, comparamos los pesos esperando que los WoE sean diferentes. Si dos categorías tienen WoE similar, juntamos esa etiqueta con la más próxima. Las variables que modificamos fueron: v_tree_Months_on_book, v_tree_Credit_Limit, v_tree_Total_Revolving_Bal, v_tree_Total_Trans_Amt, v_tree_Total_Trans_Ct, v_Total_Relationship_Count y v_Contacts_Count_12_mon. A continuación se muestran los WoE de las variables categóricas una vez corregidas:




[-inf, 581.0]
[1130.0, 2512.0]
[2513.0, inf]
[582.0, 979.0]
[980.0, 1129.0]
[-inf, 481.0]
[481.0, 684.0]
[684.2, 851.0]
[852.0, 1038.0]
[1039.0, 1436.0]
[1438.0, 1736.0]
[1737.0, 2229.0]
[2230.0, 2949.0]
[2952.0, 3664.0]
[3665.0, 4620.0]
[4635.0, 6607.0]
[6612.0, 8363.0]
[8365.0, 13184.0]
[13197.0, 23228.0]
[23250.0, inf]
[-inf, 0.467]
[0.468, 0.553]
[0.554, 0.976]
[0.977, 1.062]
[1.063, inf]
[-inf, 2010.0]
[2011.0, 2544.0]
[2545.0, 4348.0]
[4349.0, 4505.0]
[4506.0, 4723.0]
[4724.0, 4909.0]
[4911.0, 5364.0]
[5366.0, inf]
[-inf, 28.0]
[29.0, 37.0]
[38.0, 60.0]
[61.0, 64.0]
[65.0, 67.0]
[68.0, 75.0]
[76.0, 78.0]
[79.0, 91.0]
[92.0, inf]
[-inf, 0.407]
[0.408, 0.487]
[0.488, 0.536]
[0.537, 0.573]
[0.574, 0.6]
[0.603, 0.629]
[0.63, 0.656]
[0.657, 0.681]
[0.682, inf]
[-inf, 0.02]
[0.021, 0.45]
[0.451, 0.792]
[0.793, inf]
f
m
0
1
2
3
4
5
graduate
high school
post graduate
uneducated
unknown
divorced
married
single
unknown
40k 60k
60k 80k
80k 120k
less than 40k
more than 120k
unknown
blue
gold
platinum
silver
0
1
2
3
4
5
6
1
3
4
6
0
1
2
3
4
5

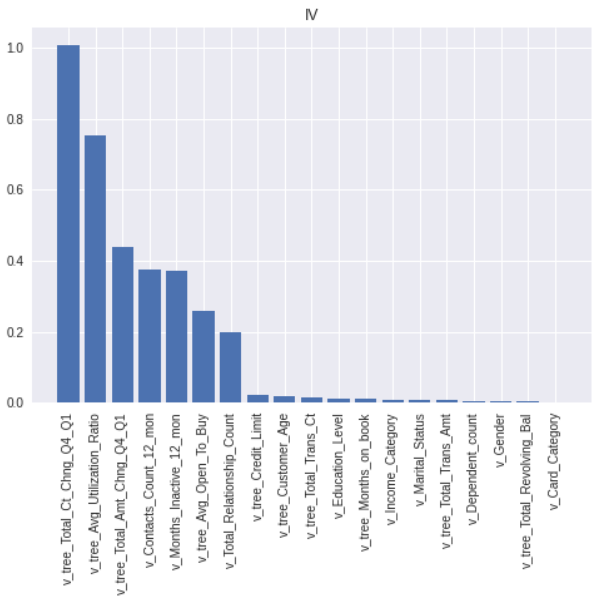
Una vez categorizadas y corregidas todas las variables, se aplicó el test Chi-cuadrada para determinar el tamaño del efecto, es decir, la asociación entre las variables; e interpretamos el valor de la información que se calculó. A continuación se muestra la interpretación y los hallazgos:

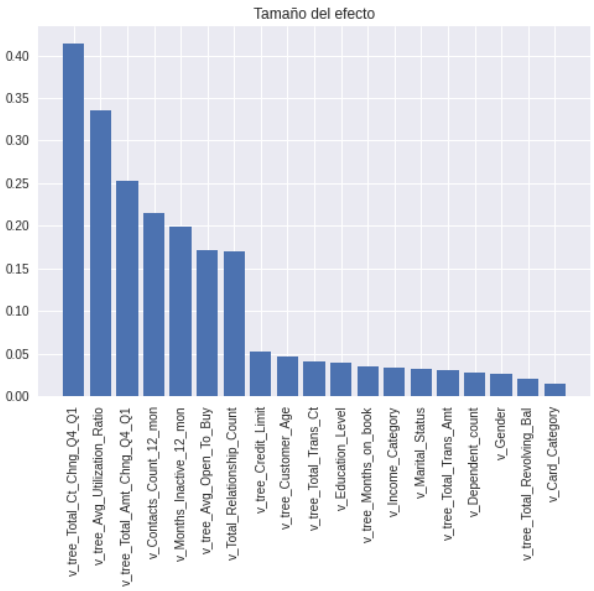
| Information Value \| Tamaño del efecto | | | |
|---|---|---|---|
| IV | Interpretación | V-Cramer | Interpretación |
| 0.5- inf | sospechoso | 0.4-inf | alto |
| 0.3- 0.5 | alto | 0.25-0.4 | medio |
| 0.1-0.3 | medio | 0.1-0.25 | bajo |
| 0.02- 0.1 | bajo | -inf-0.1 | insuficiente |
| -inf- 0.02 | insuficiente | | |

**Cuadro 5.1:** Interpretación de IV y V-Cramer

En una tabla se reunieron ambas interpretaciones y para seleccionar qué variables eliminar, revisamos que fuera identificada como ïnsuficiente"por ambas métricas. Observamos la correlación que hay entre el tamaño del efecto y el valor de la información pues si uno lo clasifica de una manera, es muy probable que el otro método lo identifique de igual forma.

| | predictor | p_value | tamaño_efecto | efe_interpre | iv | iv_status |
|---|---|---|---|---|---|---|
| 0 | v_Gender | 1.920821e-02 | 0.026867 | insuficiente | 0.005520 | insuficiente |
| 1 | v_Dependent_count | 3.251790e-01 | 0.027657 | insuficiente | 0.005716 | insuficiente |
| 2 | v_Education_Level | 1.599225e-02 | 0.040063 | insuficiente | 0.011490 | insuficiente |
| 3 | v_Marital_Status | 5.364123e-02 | 0.031753 | insuficiente | 0.007522 | insuficiente |
| 4 | v_Income_Category | 1.175673e-01 | 0.034027 | insuficiente | 0.008643 | insuficiente |
| 5 | v_Card_Category | 6.376019e-01 | 0.014948 | insuficiente | 0.001425 | insuficiente |
| 6 | v_Total_Relationship_Count | 1.736202e-47 | 0.170308 | bajo | 0.198746 | medio |
| 7 | v_Months_Inactive_12_mon | 7.770383e-62 | 0.198780 | bajo | 0.372580 | alto |
| 8 | v_Contacts_Count_12_mon | 1.163398e-73 | 0.214922 | bajo | 0.375476 | alto |
| 9 | v_tree_Customer_Age | 1.949629e-02 | 0.046880 | insuficiente | 0.017042 | insuficiente |
| 10 | v_tree_Months_on_book | 9.580175e-02 | 0.035091 | insuficiente | 0.009555 | insuficiente |
| 11 | v_tree_Credit_Limit | 3.979227e-02 | 0.051863 | insuficiente | 0.020422 | bajo |
| 12 | v_tree_Total_Revolving_Bal | 5.505246e-01 | 0.020019 | insuficiente | 0.003015 | insuficiente |
| 13 | v_tree_Avg_Open_To_Buy | 1.138873e-39 | 0.171295 | bajo | 0.258650 | medio |
| 14 | v_tree_Total_Amt_Chng_Q4_Q1 | 4.723787e-104 | 0.253177 | medio | 0.440204 | alto |
| 15 | v_tree_Total_Trans_Amt | 4.234967e-01 | 0.030471 | insuficiente | 0.007220 | insuficiente |
| 16 | v_tree_Total_Trans_Ct | 1.116347e-01 | 0.041382 | insuficiente | 0.014303 | insuficiente |
| 17 | v_tree_Total_Ct_Chng_Q4_Q1 | 8.770524e-275 | 0.413300 | alto | 1.005488 | sospechoso |
| 18 | v_tree_Avg_Utilization_Ratio | 1.819324e-184 | 0.335021 | medio | 0.753771 | sospechoso |



## 5.3. Datos finales

De 20 variables explicativas que tenía, se eliminaron 12, 11 que fueron señaladas como insuficientes debido a su relación con el objetivo y una que contenía el id del cliente y no aporta información a este modelo. Así terminamos el análisis con 2 muestras, el conjunto de entrenamiento con 8101 observaciones y el conjunto de prueba con 2026, ambas con 9 variables explicativas y una variiable objetivo.

# 6. Comparación de variables entre ambos enfoques

| Enfoque 1 | Enfoque 2 |
|---|---|
| 'tgt_Attrition_Flag' | |
| 'v_Gender' | |
| 'v_Dependent_count' | |
| 'v_Education_Level' | 'tgt_Attrition_Flag' |
| 'v_Marital_Status' | 'W_v_Total_Relationship_Count' |
| 'v_Income_Category' | 'W_v_Months_Inactive_12_mon' |
| 'v_Card_Category' | 'W_v_Contacts_Count_12_mon' |
| 'c_Months_on_book' | 'W_v_tree_Credit_Limit' |
| 'v_Total_Relationship_Count' | 'W_v_tree_Avg_Utilization_Ratio' |
| 'v_Months_Inactive_12_mon' | 'W_v_tree_Avg_Open_To_Buy' |
| 'v_Contacts_Count_12_mon' | 'W_v_tree_Total_Amt_Chng_Q4_Q1' |
| 'c_Credit_Limit' | 'W_v_tree_Total_Ct_Chng_Q4_Q1' |
| 'c_Total_Revolving_Bal' | |
| 'c_Total_Trans_Amt' | |
| 'c_Avg_Utilization_Ratio' | |

# 7. Reebalance de clases

Dado que la variable objetivo tiene un ligero sesgo hacia las observaciones que continúan con el servicio, se decidió por reebalancear las clases y comparar qué tan bueno es el ajuste con y sin reebalance. Se utilizó tanto la técnica de undefitting, la cual consiste en seleccionar solo algunas observaciones de la clase con mayor peso e igualarlas a la clase menor, y overfitting que, por el contrario, genera un complemento de la muestra menor para igualar la clase con mayor peso. Cabe mencionar que este reebalance solo se aplicó al enfoque I.

Para determinar qué técnica es mejor, ajusté una regresión logística con el nuevo balance y comparamos las métricas como se muestra a continuación:

multirow

| MÉTRICAS | UNDER | OVER (random) 1 | OVER (smote) |
|---|---|---|---|
| ROC | 0.8980 | 0.8974 | 0.8813 |
| F1 score | 0.8333 | 0.7985 | 0.8852 |

Dado que las ROC son muy similares y el F1 score del tercer método es mejor respecto a los demás, se consideró el reebalance por SMOTE en los análisis siguientes.

# 8. Modelos de aprendizaje supervisado

Dado que la variable objetivo (tgt_Attrition_Flag) es del tipo dicotómico, ajustaremos modelos de clasificación. Para determinar qué modelo y enfoque (en cuanto al tratamiento de los datos) es mejor, comparamos las métricas y se seleccionó aquel con mejor validación.

### 8.0.1. Regresión logística

Se aplicaron dos procesos, una regresión sin castigo y una con castigo (regresión lasso). Recordemos que se aplicó a 3 muestras; a continuación las métricas de cada modelo:

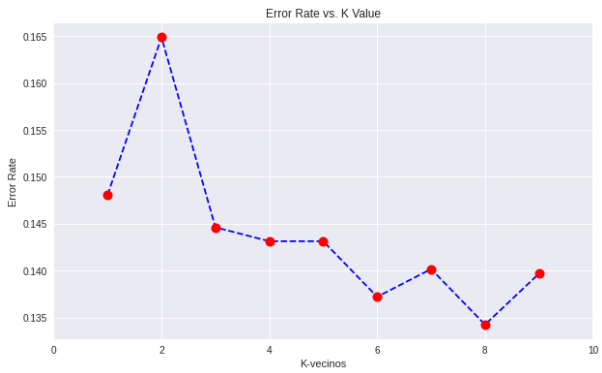
| Métricas | Enfoque 1 | | E1 (reebalance) | | Enfoque 2 | |
|---|---|---|---|---|---|---|
| | logística | lasso | logística | lasso | logística | lasso |
| ROC | 0.89523 | 0.89642 | 0.91746 | 0.91723 | 0.87326 | 0.87327 |
| Accuracy | 0.89143 | 0.86470 | 0.83911 | 0.83269 | 0.88648 | 0.88648 |
| F1 score | 0.93826 | 0.91826 | 0.83046 | 0.81571 | 0.93470 | 0.93470 |
| Media error | 0.10857 | 0.13530 | 0.26558 | 0.16731 | 0.11352 | 0.11352 |

Si bien, la ROC fue buena en el enfoque I con reebalance, la media del error, F1 score y Accuracy, empeoraron respecto a los otros enfoques, por ello se optó por solo considerar el balance inicial en futuros ajustes, considerando que este es buen predictor pues se encuentran buenas métricas y la predicción se distribuye entre buenos y malos.

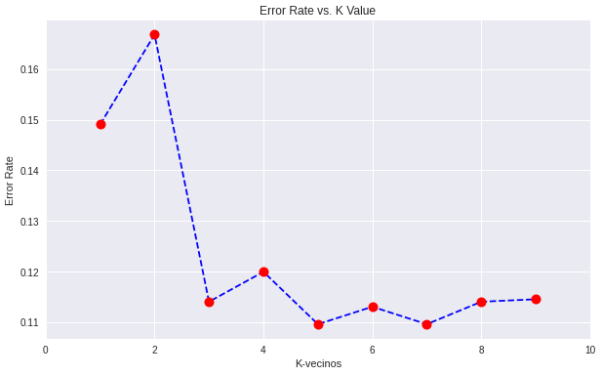
En conclusión, el modelo logístico con y sin castigo, considerando el enfoque I y enfoque II, resulta ser el mejor, sin embargo exploraremos otros modelos con la intención de encontrar mejores métricas.

### 8.0.2. K vecinos

Dado que éste algoritmo consume muchos recursos computacionales, escalamos las variables explicativas en ambas muestras por el método estándar. Por otro lado, para determinar con qué cantidad de nodos vecinos se encuentra el mejor ajuste, ajustamos 10 modelos considerando hasta 10 vecinos, y se comparó el error entre la predicción y lo observado, así se obtuvieron los siguientes gráficos:




**(a)** Enfoque 1




**(b)** Enfoque 2

Observando estos gráficos se detrminó que el mejor ajuste con el enfoque I se obtiene considerando 8 vecinos, mientras que con el enfoque II se obtiene con 5 vecinos. Al ajustar los modelos obtuve las siguientes métricas:

| ~~Métricas~~ | Enfoque I | Enfoque II |
|---|---|---|
| | K=8 | K=5 |
| ROC | 0.81890 | 0.81154 |
| Accuracy | 0.87725 | 0.89042 |
| F1 score | 0.93159 | 0.93715 |
| Media error | 0.12275 | 0.10958 |

Como observamos, con el enfoque II obtenemos mejores métricas además que el error de predicción disminuyó respecto al modelo logístico. No obstante, exploremos un último modelo de aprendizaje supervisado con la intención de encontrar mejores métricas.

### 8.0.3. Ensamble (XGBoost)

Este ajuste consiste en combinar el modelo de gradiente descendiente para minimizar la función de costos con el modelo de árboles de decisión. Para el ajuste se consideró una tasa de aprendizaje 0.1, un límite de hasta 190 árboles con profundidad máxima de 5 hojas. Las metricas que se obtuvieron son las siguientes:

| Métricas | Enfoque I | Enfoque II |
|---|---|---|
| ROC | 0.95990 | 0.94510 |
| Accuracy | 0.93290 | 0.92495 |
| F1 score | 0.96051 | 0.95639 |
| Media error | 0.06710 | 0.07505 |

Si bien los resultados por ambos enfoques son buenos, el enfoque I tiene menor error en la predicción, además, al comparar las métricas entre el conjunto de entrenamiento y el conjunto de validación, las métricas son muy similares por lo que se descarta sobreajuste del modelo; asi mismo, la predicción está balanceada entre buenos y malos por lo que descartamos un posible sesgo en las clases.

En conclusión, de los 3 modelos que se probaron con sus variantes, el modelo de aprendizaje supervisado que mejor se ajusta a nuestros datos considerando el enfoque I es un ensamble de tipo Extreme Gradient Boosting, incluso si consideramos el enfoque II las métricas bajo el XGBoost son buenas, sin embargo al comparar los indicadores de los conjuntos de entrenamiento y validación podrían sugerir algún problema de sobreajuste, por lo que se prefiere el enfoque O. Se obtuvieron las siguientes métricas al comparar los conjuntos:

| Métricas | Enfoque I | | Enfoque II | |
|---|---|---|---|---|
| | Train | Test | Train | Test |
| ROC | 0.99406 | 0.95990 | 0.94510 | 0.87122 |
| Accuracy | 0.97056 | 0.93290 | 0.92495 | 0.89339 |
| F1 score | 0.98254 | 0.96051 | 0.95639 | 0.93846 |
| Media error | 0.02944 | 0.06710 | 0.07505 | 0.10661 |

# 9. Modelos de aprendizaje no supervisado

El objetivo del aprendizaje no supervisado es encontrar patrones en los datos que nos ayuden a a identificar similitudes con el resto de observaciones. A diferencia del aprendizaje supervisado, este tipo de modelación no requiere de un conjunto de entrenamiento y uno prueba por lo que el tratamiento de ingeniería de variables, outliers y valores faltantes, que anteriormente se había trabajado, no es de utilidad en esta modelación.

El objetivo de esta etapa del proyecto es ajustar un modelo no supervisado que nos permita segmentar a los clientes que continuarán con el servicio y los que lo abandonarán, siguiendo así el objetivo general del proyecto. Para determinar qué modelo es mejor, analizaremos el perfilamiento de los cluster que cada modelo sugiera en complemento con la exploración visual de los cluster.

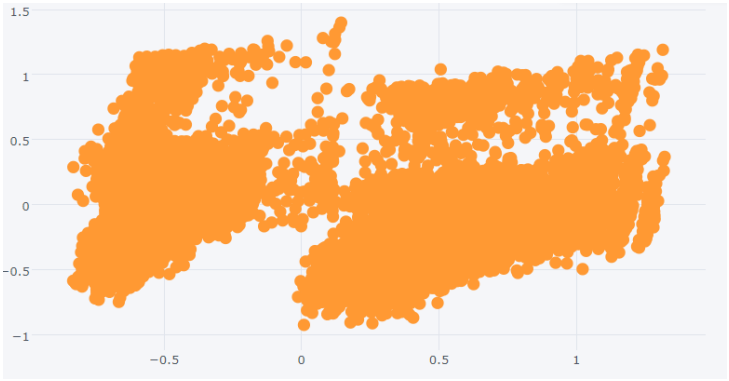
### 9.0.1. Ingenieria de variables

En aprendizaje supervisado utilicé tanto la muestra de entrenamiento como la de prueba para convertir variables categóricas en numéricas, en este caso, dado que los datos serán analizados en conjunto, se decidió realizar ingenieria de variables considerando todos los datos como a continuación se describe:

| One-hot Encoding | Codificación ordinal | Count Encoding |
|---|---|---|
| **v_Attrition_Flag** | **v_Income_Category** | **v_Education_Level** |
| 1 existing customer<br>0 atrrited customer | 0 unknown<br>1 less than 40k<br>2 40k-60k<br>3 60k-80k<br>4 80k-120k<br>5 more than 120k | 0.40890 graduate<br>0.1987 high school<br>0.1499 unknown<br>0.14683 uneducated<br>0.0954 post graduate |
| **v_Gender** | | **v_Marital_Status** |
| 1 f<br>0 m | | 0.4628 married<br>0.3893 single<br>0.0738 divorce<br>0.0739 unknown |
| | | **v_Card_Category** |
| | | 0.93176 blue<br>0.0548 silver<br>0.0114 gold<br>0.0019 platinum |

### 9.0.2. Visualización de los datos

Dado que nuestro conjunto de datos cuenta con 21 variables en diferentes unidades (decimas, centenas, etc), el primer paso fue escalar los datos por el método Min-Max con el objetivo de mantener la estructura de los datos, es decir, las observaciones lejanas que se sigan manteniendo lejanas y las cercanas, cerca. Posteriormente, se utilizó reducción de dimensionalidad por componentes principales (PCA) para poder observar la estructura de los datos. Cabe señalar que este método sugirió 10 componentes para explicar al menos el 90% de varianza, sin embargo, somos capaces de graficar a lo más en 3 dimensiones. A continuación se muestra el gráfico a 2 dimensiones y con escalamiento:




El grafico sugiere que encontraremos entre 2 y 4 cluster, sin embargo, dado que solo estamos visualizando 2 componentes, no podemos dar por hecho esta observación.

### 9.0.3. Multicolinealidad

La multicolinealidad es causada por la estrecha similitud entre dos o más variables, es decir, una explica a la otra por lo que se comportan igual o muy similar. Esto suele causar problemas de sobreajuste y por tanto las conclusiones pueden ser erroneas; adicionalmente, una gran cantidad de variables explicativas incrementa los recursos que la maquina usa para ajustar el modelo. No obstante, debemos tener encuenta que existen casos en los que el comportamiento de una o más variables es similar pero la información que capturan es totalmente diferente por lo que hay que analizar a las variables más allá de solo calcular su correlación.

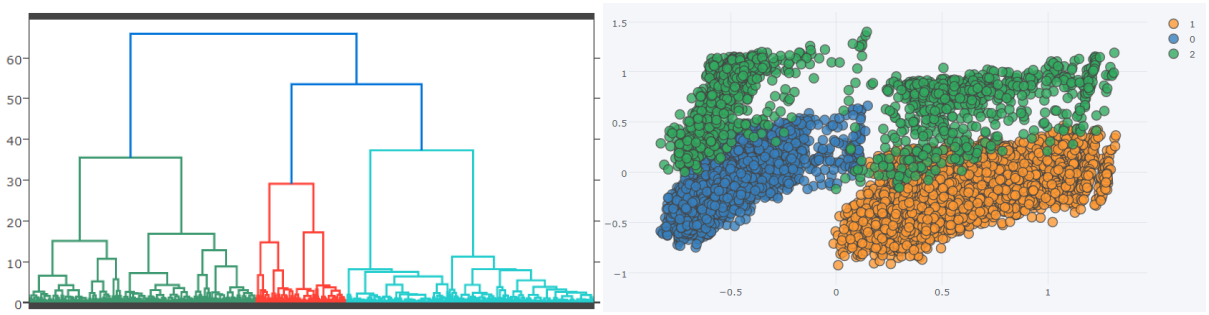
Considerando los puntos anteriores, consideramos adecuado aplicar clustering de variables con el objetivo de reducir la cantidad de características y conservar solo aquellas que realmente expliquen el comportamiento de los clientes. Después de analizar los cluster que sugiere el algoritmo, decidimos conservar 7 variables de acuerdo a su $R-square$, la cual se espera sea cercana a 1. Las variables que consideraré son:

| Cluster | Variable | RS_Own | RS_NC | RS_Ratio |
|---|---|---|---|---|
| 0 | c_Avg_Open_To_Buy | 0.928882 | 0.222085 | 0.091421 |
| 1 | c_Total_Trans_Amt | 0.870199 | 0.035171 | 0.134533 |
| 2 | c_Customer_Age | 0.876314 | 0.002426 | 0.123987 |
| 3 | v_Attrition_Flag | 0.625382 | 0.064059 | 0.400258 |
| 4 | v_Gender | 0.893304 | 0.144905 | 0.124777 |
| 5 | c_Total_Amt_Chng_Q4_Q1 | 0.692095 | 0.013582 | 0.312145 |
| 6 | v_Education_Level | 1.000000 | 0.000066 | 0.000000 |

A continuación se explica brevemente los hallazgos de cada modelo considerando todas las variables y solo las aquí planteadas con el objetivo de determinar si la reducción de variables afecta o no el resultado.

### 9.0.4. Clustering Jerárquico

Como se demostró a lo largo del curso, el enlace Ward y el algoritmo aglomerativo, son más eficientes en el ajuste de modelos, por ello se decidió ajustar un modelo de clustering jerárquico aglomerativo con métrica Ward de 4 clusters como lo sugiere el dendograma, sin embargo, uno de los grupos no contaba con observaciones suficientes (como se observa en el gráfico a la izquierda) por lo que finalmente se ajustó el modelo con 3 grupos y solo con las variables seleccionadas.




El perfilamiento de los clusters es bueno, pues segmentó a la población

que continuará en con el servicio (cluster 0 y 1) y a los que lo abandonarán (cluster 2), no obstante, no hace una segmentación clara entre el resto de características del cluster 0 y 1.

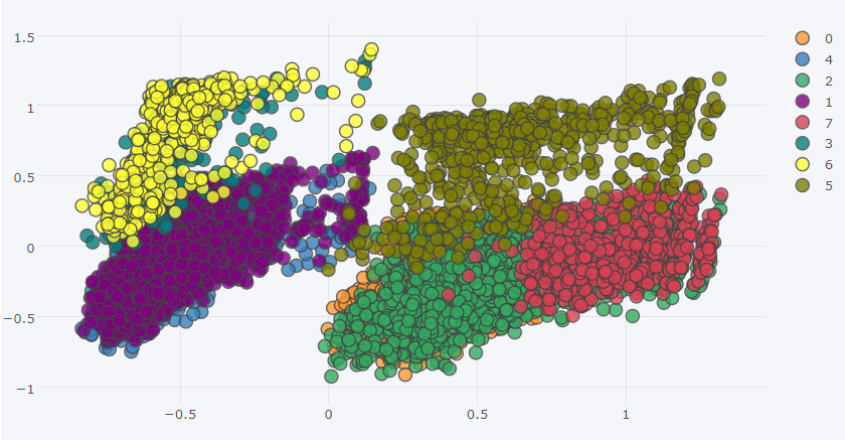
### 9.0.5. Clustering de optimización

Para este modelo se utilizó la media para determinar el centroide de cada cluster. Si bien, la media no es un estadístico recomendable pues es suceptible al sesgo por outliers además que el centroide puede no estar dentro del cluster, la convergencia del algoritmo se basa en la distancia euclideana, con la que estoy más familiarizada y se vió en clase.

Para determinar el número de clusters óptimo, utilizamos la métrica Davies Bouldin, donde se espera un ratio bajo, y la métrica Silhouette, que por el contrario, se espera sea alta. Así se ajustaron hasta 9 modelos con 2 a 10 cluster considerando las variables seleccionadas en el clustering de variables. Se obtuvieron los siguientes gráficos:


**(a)** Davies Bouldin Score **(b)** Silhouette Score

De acuerdo a la figura (a) el número óptimo de clusters es 8, por otro lado, la (b) sugiere 10 grupos si soy estricta y busco el mayor ratio, sin embargo 8 clusters es de los puntos máximos de la curva por lo que se ajustó un modelo de KMedias con 8 cluster:



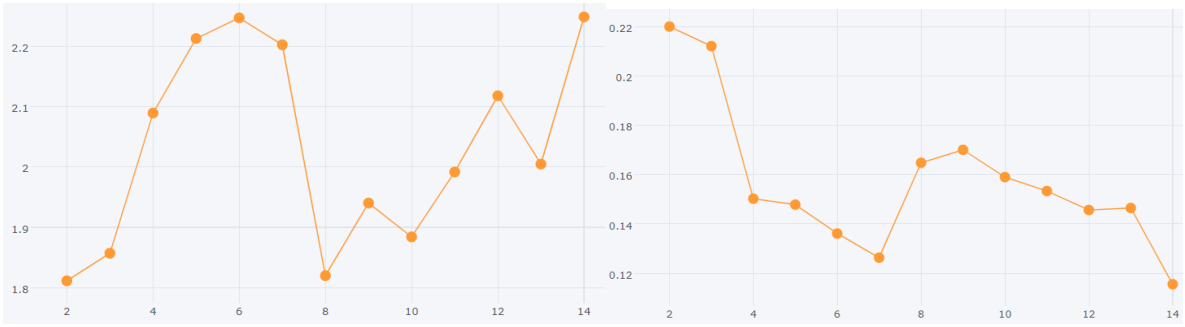

Como puede observarse, el cluster 6 y 3 son muy similares en sus características y son los que menos observaciones tienen, por ello se recomienda ajustar un modelo de 6 ó 7 clusters. No obstante, dado que este algoritmo supone que los clusters son redondos, se decidió ajustar un modelo más flexible que se ajuste a la estructura de los datos como es el algoritmo $GaaussianMixture$.

### 9.0.6. Clustering de densidad

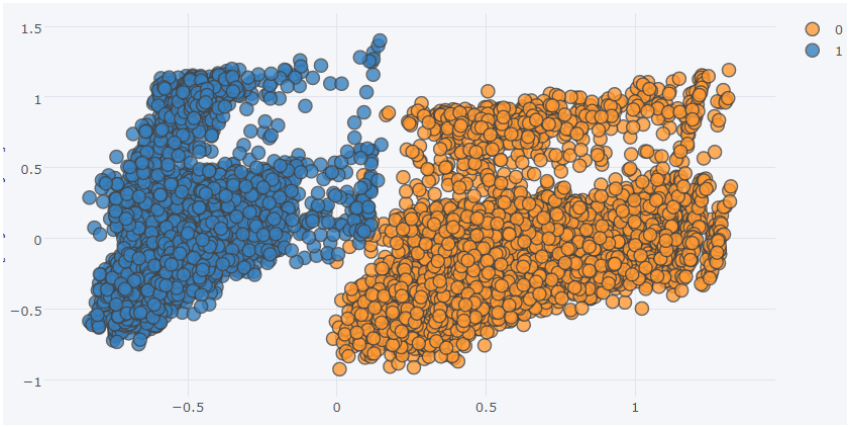
En este tipo de modelos se pueden encontrar varios algoritmos con supuestos distintos, en esta sección solo tratare con Gaussian Mixture y DBSCAN.

**Gaussian Mixture** Para contrastar los cambios en la segmentación del modelo considerando todas las variables contra el modelo con las variables relevantes, se ajustaron dos modelos.

Como se puede observar en el gráfico a y b, al considerar todas las variables, las métricas sugieren que el número de cluster óptimo es dos, sin embago, cuando ajusto el modelo, obtenemos que la principal variable que se considera para hacer los grupos es el género y no la de interés (attrition_flag), por lo que se sugiere analizar el modelo únicamente con las variables no colineales.

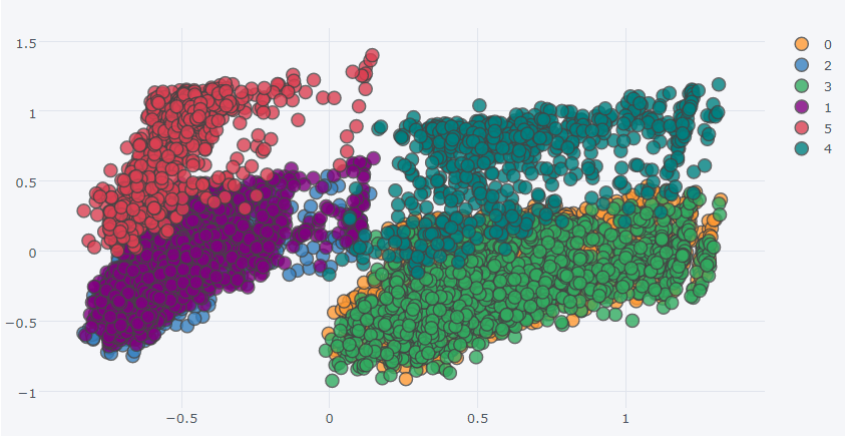



**(a)** Davies Bouldin Score **(b)** Silhouette Score




Cuando consideramos solo las variables preseleccionadas, las métricas sugieren que la cantidad de cluster óptimo es 6, justo lo que sugería el modelo de KMmedias.

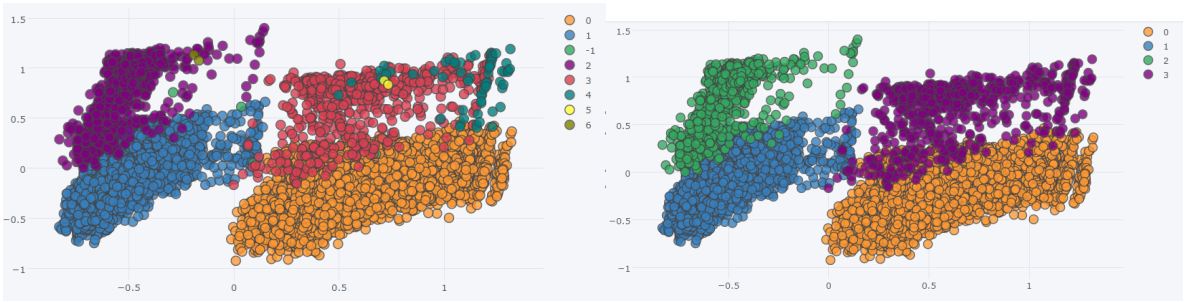

**(a)** Davies Bouldin Score **(b)** Silhouette Score




Tanto gráficamente como analíticamente, la segmentación de clientes que continuarán con el servicio y los que la abandonarán es clara. Además, cada cluster contiene suficientes observaciones que nos permiten perfilar cada grupo. No obstante, ajustamos 2 modelos más con el algoritmo DBSCAN ya que suele ser más eficiente en cuanto a tiempo máquina cuando tenemos numerosos datos y nos ayuda a detectar el ruido (observaciones raras).

**DBSCAN** En el gráfico (a) podemos observar el modelo considerando todas las variables, maxima distancia entre grupos de 1 unidad y con al menos 2 cluster. El algoritmo fue capaz de identificar 6 observaciones como ruido, y sugirió 3 clusters con menos de 100 observaciones. Dado que un con perfilamiento granular no se logra el objetivo, se ajusór otro modelo con las mismas características pero considerando solo las variables preseleccionadas, el resultado se observa en el gráfico (b).

En este segundo modelo, los grupos son más robustos y sus características son claras, no obstante, el perfilamiento de 4 grupos es muy general pues el cluster 0 y 1 representan más del 60% de la población total.



**(a)** **(b)**

### 9.0.7. Perfilamiento

Tomando en cuenta las observaciones en cada modelo, se consideró que el mejor ajuste que cumple con el objetivo es el modelo de Gaussian Mixture considerando solo las variables relevantes. A continuación se muestra el perfilamiento de cada cluster:

- Grupo 1
  - Edad: 26-73.
  - Género: masculino
  - Nivel de educación: preparatoria/desconocido
  - Ingresos: 60k-120k
  - Transacciones: bajas
  - Promedio de compra: bajo
- Grupo 2
  - Edad: 26-67.
  - Género: femenino
  - Nivel de educación: preparatoria/desconocido
  - Ingresos: menos de 40k o desconocido
  - Transacciones: altas
  - Promedio de compra: medio

- Grupo 3
  - Edad: 26-67.
  - Género: femenino
  - Nivel de educación: universitario
  - Ingresos: 0-60k
  - Transacciones: medio
  - Promedio de compra: alto
- Grupo 4
  - Edad: 26-68
  - Género: masculino
  - Nivel de educación: universitario
  - Ingresos: 60k-120k
  - Transacciones: medio
  - Promedio de compra: alto
- Grupo A
  - Edad: 26-68
  - Género: masculino
  - Nivel de educación: universitario/preparatoria
  - Ingresos: 60k o más
  - Transacciones: bajas
  - Promedio de compra: medio
- Grupo B
  - Edad: 26-68
  - Género: femenino
  - Nivel de educación: universitario/preparatoria
  - Ingresos: 60k o menos

- Transacciones: bajas
- Promedio de compra: bajo

Los grupos 1 al 4 son los clientes que continúan con su tarjeta de crédito y los grupos A y B son los que suspendieron el servicio. En el listado anterior solo se mencionan las características más sobresalientes, las demás se pueden identificar en el modelo.

# 10. Comentarios finales

En conclusión, el propósito de estos modelos es predecir, con cierto nivel de confianza, si un nuevo cliente mantendrá su tarjeta de crédito o abandonará el servicio de acuerdo a la cercanía de sus características con alguno de los grupos o bien, de acuerdo con la predicción resultado de la regresión ganadora, de esta manera el Banco podrá controlar el nivel de riesgo que está dispuesto a aceptar pues será capaz estimar la cantidad de clientes que saldrán de la cartera e implementar medidas control en caso de que se trate de clientes morosos.

Además, la institución podría dirigir parte de sus recursos a campañas para atracción de clientes con características de los grupos que continuarán en el Banco con el objetivo de aumentar la cartera de clientes buenos; y campañas para mantener a los clientes que se espera salgan de la cartera. Cabe señalar que los modelos presentados en este proyecto necesitan de monitoreo constante ya que la población suele cambiar su compotamiento y podría requerirse una calibración del modelo.

# 11. Anexos

### 11.0.1. Código del desarrollo

https://colab.research.google.com/drive/1rnDO9PVVe9Mg3ym6jupLBLVweEADDcgh?usp=sharing

### 11.0.2. Datos utilizados en el desarrollo

https://drive.google.com/file/d/1P9mDkE41frcfEuijD4qsVesZKJYbdxWr/view?usp=sharing