AVANCES PROYECTO- SEMANA 3 – GRUPO 21

Santiago Pulido, Liliana Briceño, Cristhian Barbosa, Diego Peñaloza

Enlace GitHub: https://github.com/spulidov91/DSA_churn.git

Problema que abordarán y su contexto.

De cara a una entidad financiera la retención de sus clientes es uno de los puntos más importantes para su funcionamiento, pues de ello depende la estabilidad y sanidad de su cartera. Por ello el proyecto busca desarrollar un modelo predictivo capaz de identificar a los clientes de un banco con mayor probabilidad de abandonar sus servicios de tarjeta de crédito. Para ello, se utilizará un conjunto de datos que incluye la información demográfica y financiera de aproximadamente10,000 clientes. A través del uso de técnicas de Machine Learning, se busca identificar patrones en los datos de los clientes que permitan predecir la probabilidad de que dejen de utilizar el producto (probabilidad de Churn). El resultado otorgará al banco un sistema que le permita implementar acciones efectivas destinadas a retener los clientes en riesgo de abandono, lo que contribuiría a aumentar la tasa de retención, mejorar la satisfacción del cliente y mantener la cartera estable.

Pregunta de negocio y alcance del proyecto.

¿Qué clientes están en mayor riesgo de abandonar el servicio de tarjeta de crédito, como podemos identificar características claves para implementar estrategias de retención de clientes efectivas y cuál es el valor esperado asociado a la perdida por abandono de clientes?

El proyecto se centrará en el desarrollo y evaluación de un modelo predictivo para identificar clientes con alta probabilidad de abandono en el servicio de tarjetas de crédito de un banco. Para ello, se procesará y analizará un conjunto de datos de aproximadamente 10,000 clientes, con información demográfica y financiera, se estimará la probabilidad de abandono y se identificaran las principales causales de abandono. Más específicamente nos enfocaremos en los siguientes puntos:

- Estimar que características del cliente lo hacen más o menos propenso a abandonar los productos del banco.
- Estimar en la cohorte de datos cuales clientes tienen mayor probabilidad de abandonar sus productos.
- Estimar el valor esperado de las perdidas por abandono de clientes.
- Reportar estos tres puntos anteriores en un dashboard en línea a los stakeholders.

Conjuntos de datos a emplear.

Se utilizarán datos existentes proporcionados por el banco, que incluyen 18 características por cliente, como edad, ingresos, estado civil, límites de crédito, historial de transacciones y métricas

de uso de la tarjeta. No se planea recolectar nuevos datos. Además, se generarán nuevas variables derivadas de las ya existentes, como razones de cambio de comportamiento entre trimestres, tasas de utilización de crédito, y agrupaciones basadas en categorías.

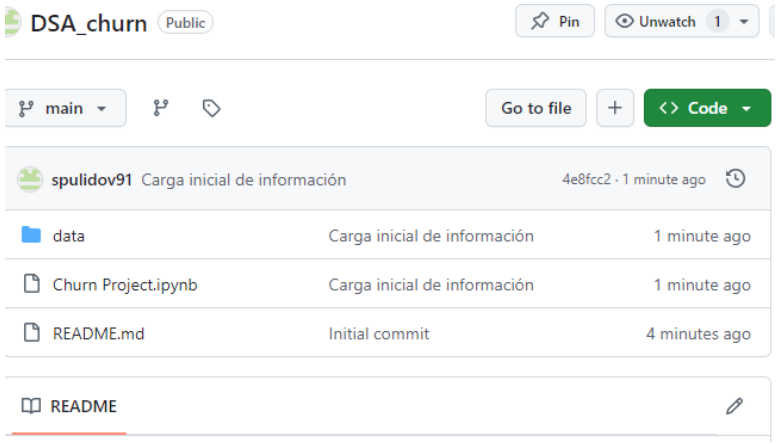
El conjunto de datos presenta 21 variables, las cuales se describen a continuación:

Tabla 1. Variables del conjunto de datos

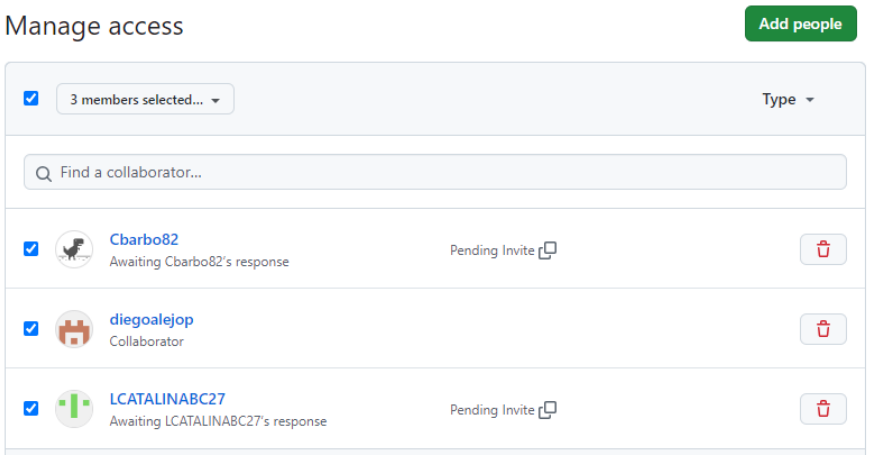
Variable	Tipo de Variables	Descripción
CLIENTNUM	Numérica	Identificador único del cliente titular de la cuenta.
Attrition_Flag	Categórica	Si la cuenta está cerrada entonces 1 de lo contrario 0
Customer_Age	Numérica	Edad del cliente en años
Gender	Categórica	M=Hombre, F=Mujer
Dependent_count	Numérica	Número de dependientes
Education_Level	Categórica	Título educativo del titular de la cuenta (ejemplo: bachillerato, título universitario, etc.)
Marital_Status	Categórica	Casado, Soltero, Divorciado, Desconocido
Income_Category	Categórica	Categoría de ingresos anuales del titular de la cuenta (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, >)
Card_Category	Categórica	Tipo de tarjeta (Azul, Plata, Oro, Platino)
Months_on_book	Integer	Periodo de relación con el banco
Total_Relationship_Count	Numérica	Número total de productos que posee el cliente
Months_Inactive_12_mon	Numérica	Número de meses inactivos en los últimos 12 meses
Contacts_Count_12_mon	Numérica	Número de contactos en los últimos 12 meses
Credit_Limit	Decimal	Límite de crédito en la tarjeta de crédito
Total_Revolving_Bal	Decimal	Saldo rotatorio total de la tarjeta de crédito
Avg_Open_To_Buy	Decimal	Línea de crédito abierta para comprar (promedio de los últimos 12 meses)
Total_Amt_Chng_Q4_Q1	Decimal	Cambio en el monto de las transacciones (cuarto trimestre respecto del primer trimestre)
Total_Trans_Amt	Decimal	Monto total de la transacción (últimos 12 meses)
Total_Trans_Ct	Numérica	Recuento total de transacciones (últimos 12 meses)
Total_Ct_Chng_Q4_Q1	Decimal	Cambio en el recuento de transacciones (cuarto trimestre con respecto al primer trimestre)
Avg_Utilization_Ratio	Decimal	Tasa de utilización promedio de la tarjeta

Repositorio Git en uso para el código.

Se crea un repositorio abierto en GitHub donde se alojan los archivos del proyecto:
https://github.com/spulidov91/DSA_churn.git



Se adicionan colaboradores al repositorio:



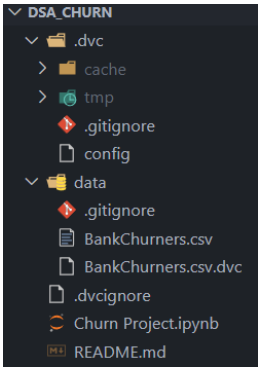
Repositorio DVC en uso para los datos.

Se crea repositorio DVC para el uso de los datos -> data/ BankChurners.csv.

```
dapen@Diego MINGW64 ~/OneDrive - Universidad de los andes/MIAD/Despliegue de soluciones analiticas/Proyecto (master)
$ dvc add DSA_churn/data/BankChurners.csv
100% Adding...|████████████████████|1/1 [00:00,  1.12file/s]

To track the changes with git, run:

        git add 'DSA_churn\data\BankChurners.csv.dvc' 'DSA_churn\data\.gitignore'

To enable auto staging, run:

        dvc config core.autostage true
(venv)
dapen@Diego MINGW64 ~/OneDrive - Universidad de los andes/MIAD/Despliegue de soluciones analiticas/Proyecto (master)
$ dvc push
Collecting         |1.00 [00:00,  200entry/s]
Pushing
1 file pushed
(venv)
```

Este se incluye en Git y apunta a una carpeta en la máquina local de uno de los miembros del equipo, el cual servirá como repositorio base de los datos.

Creación de archivos . dvc y .dvcignore:



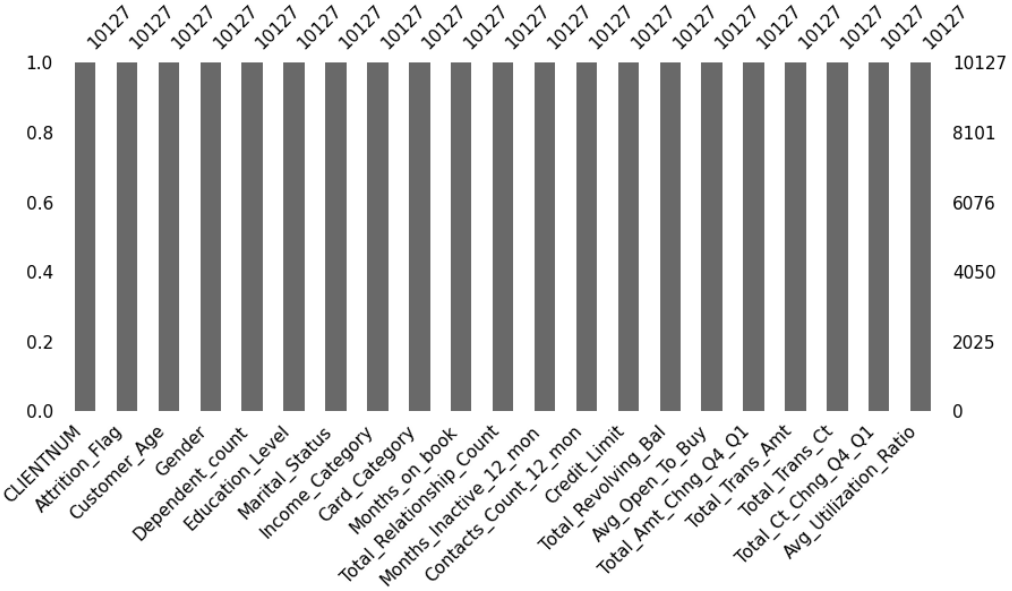
Pruebas de funcionamiento de DVC

```
daper@Diego MINGW64 ~/OneDrive - Universidad de los andes/MIAD/Despliegue de soluciones analiticas/DSA_churn/DSA_churn (main)
$ rm -rf .dvc/cache/

daper@Diego MINGW64 ~/OneDrive - Universidad de los andes/MIAD/Despliegue de soluciones analiticas/DSA_churn/DSA_churn (main)
$ rm -f data/BankChurners.csv

daper@Diego MINGW64 ~/OneDrive - Universidad de los andes/MIAD/Despliegue de soluciones analiticas/DSA_churn/DSA_churn (main)
$ ls data
BankChurners.csv.dvc

daper@Diego MINGW64 ~/OneDrive - Universidad de los andes/MIAD/Despliegue de soluciones analiticas/DSA_churn/DSA_churn (main)
$ dvc pull
Collecting
Fetching
Building workspace index
Comparing indexes
Applying changes
A       data\BankChurners.csv
1 file added and 1 file fetched

daper@Diego MINGW64 ~/OneDrive - Universidad de los andes/MIAD/Despliegue de soluciones analiticas/DSA_churn/DSA_churn (main)
$ ls data
BankChurners.csv  BankChurners.csv.dvc
```

Entendimiento de los datos:

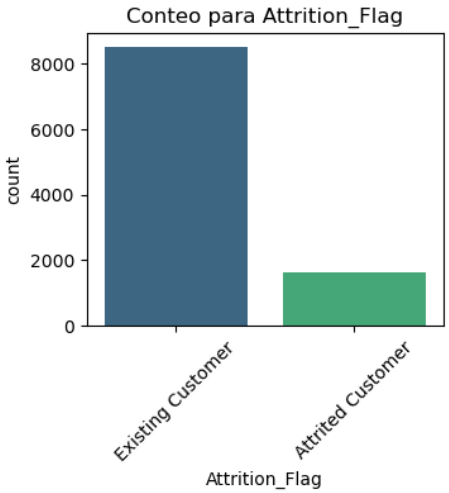
Dimensionalidad: El conjunto de datos cuenta con un total de 10.127 filas y 21 variables.

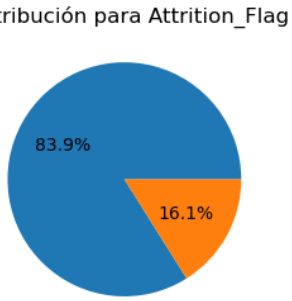
Valores ausentes: No se evidencian valores ausentes dentro del conjunto de datos.



Valores duplicados: No existen clientes duplicados, con base en la columna **CLIENTNUM** (Identificador único de cliente).

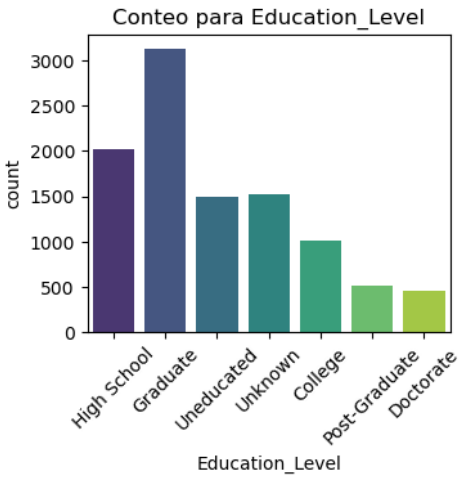
Variables categóricas:

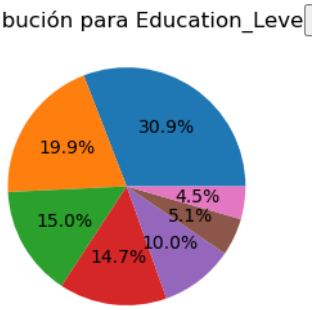




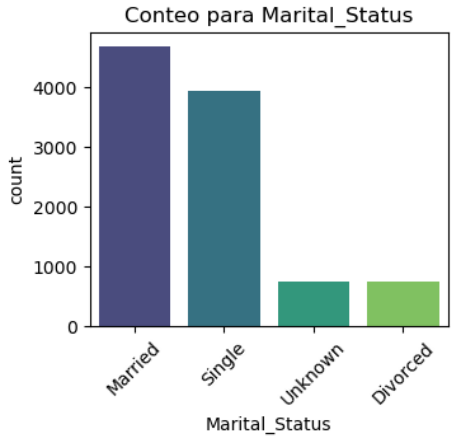
La variable dependiente, que es un flag que marca si el cliente abandonó el banco o no, tiene un desbalanceo en tanto que solo el 16.1% de los clientes abandonaron los servicios del banco.

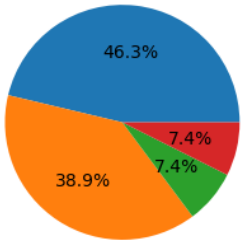
Un 52% de los clientes del banco son del género femenino.



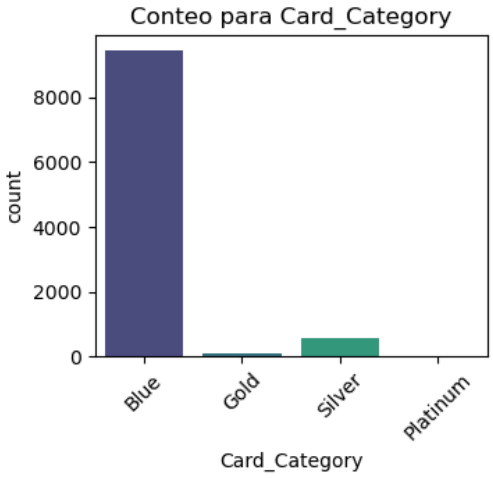


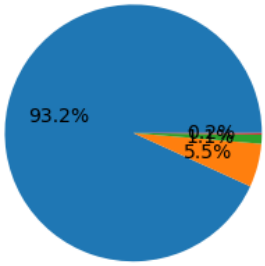
La mayoría de los clientes tienen estudios de educación básica primaria y secundaria (High School 19.9% y Graduate de 30.9%). Un 15% no cuenta con educación formal y menos del 20% tiene estudios de educación superior.



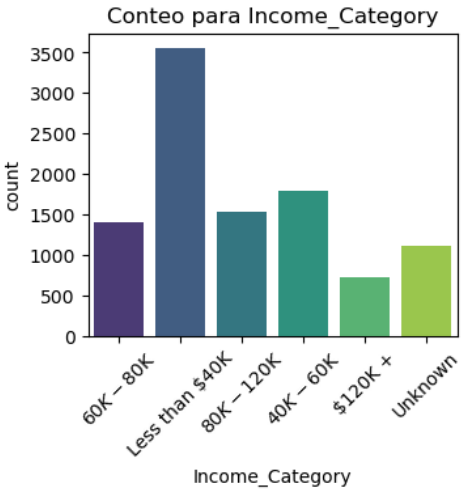


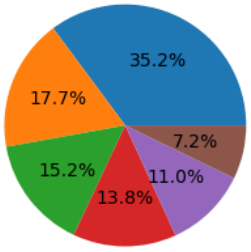
En cuanto a estado marital la mayoría de clientes (46.3%) se encuentra casado, mientras que un 38.9% se encuentra soltero.





Por otra parte, es importante ver que la gran mayoría, un 93.2%, de los clientes tiene una tarjeta tipo blue, seguidos de un 5.5% de clientes con tarjeta Silver y menos de un 2% de clientes tienen tarjeta tipo gold o platinum.





Finalmente, en cuanto a categorías de ingreso, se ve que la mayoría, un 35.2%, de clientes tiene ingresos anuales menores a 40 mil dólares, seguidos del grupo de clientes con ingresos entre 40 y 60 mil dólares (17.2% del total). Se destaca que un 11% de los clientes no reporta sus ingresos.

Variables continuas

VARIABLE	COUNT	MEAN	STD	MIN	MAX
CLIENTNUM	10127	7.39e+08	3.69e+07	7.08e+08	8.28e+08
Customer_Age	10127	46.32	8.01	26	73
Dependent_count	10127	2.34	1.29	0	5
Months_on_book	10127	35.92	7.98	13	56
Total_Relationship_Count	10127	3.81	1.55	1	6
Months_Inactive_12_mon	10127	2.34	1.01	0	6
Contacts_Count_12_mon	10127	2.455317	1.106225	0	6
Credit_Limit	10127	8631.95	9088.77	1438.3	34516
Total_Revolving_Bal	10127	1162.81	814.98	0	2517
Avg_Open_To_Buy	10127	7469.13	9090.68	3	34516
Total_Amt_Chng_Q4_Q1	10127	0.75	0.21	0	3.397
Total_Trans_Amt	10127	4404.08	3397.12	510	18484
Total_Trans_Ct	10127	64.85	23.47	10	139
Total_Ct_Chng_Q4_Q1	10127	0.71	0.23	0	3.714
Avg_Utilization_Ratio	10127	0.27	0.27	0	0.999

- **Customer_Age:** La mayoría de los clientes tienen entre 30 y 60 años, lo cual representa una población adulta. Este grupo etario puede reflejar una mayor estabilidad financiera, pero también algunas variaciones en el comportamiento según las diferentes etapas de vida.

- **Dependent_count:** Con una media de 2.34 dependientes, la mayoría de los clientes tiene entre 1 y 4 dependientes. La cantidad de dependientes puede influir en el uso y manejo de crédito, ya que los clientes con más dependientes podrían tener más necesidades financieras que los demás.

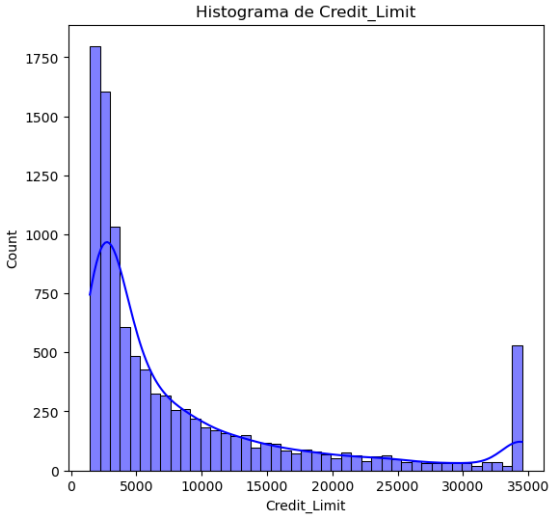
- **Months_on_book:** Los clientes han tenido sus tarjetas en promedio por casi tres años (36 meses), lo que indica una relación prolongada. La antigüedad en el banco podría estar relacionada con la probabilidad de retención o abandono, ya que los clientes con más tiempo podrían mostrar más lealtad o una relación financiera más estable.

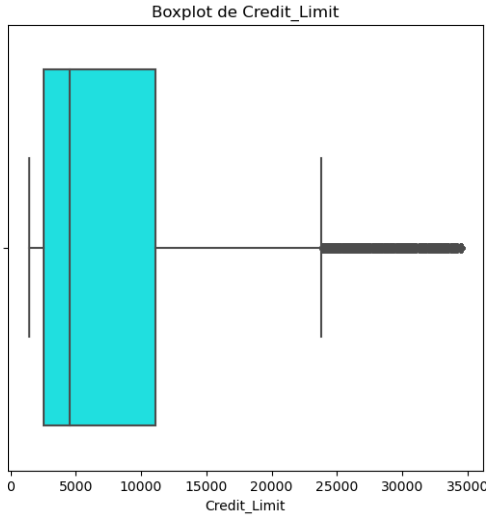
- **Total_Relationship_Count:** Los clientes tienen, en promedio, más de tres productos con el banco, lo que sugiere un grado de dependencia de la mayoría. Un mayor número de productos puede indicar una mayor retención.

- **Months_Inactive_12_mon:** Se tiene un promedio de 2.34 meses de inactividad en un año. La inactividad llega a ser un indicador clave del riesgo de abandono, ya que clientes inactivos por más tiempo podrían estar en riesgo de dejar de usar los productos del banco.

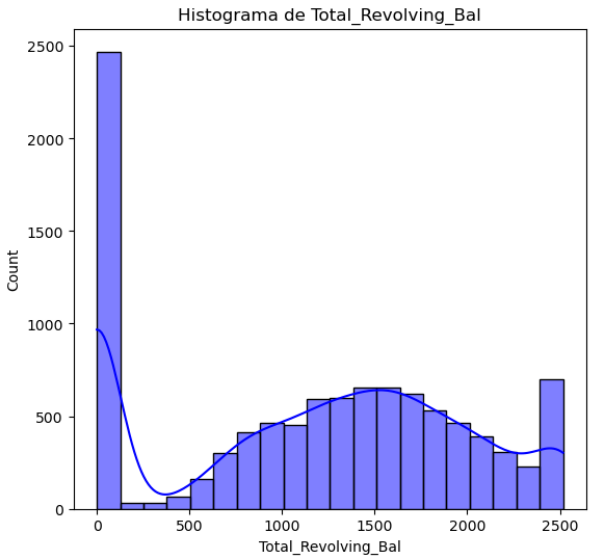
- **Contacts_Count_12_mon:** La frecuencia de contacto es un aspecto importante, ya que los clientes con bajo contacto podrían sentir menos compromiso, mientras que en un alto número de contactos podría sugerir una buena gestión de la relación con el cliente. En promedio, los clientes han tenido cerca de 2.5 contactos con el banco en el último año

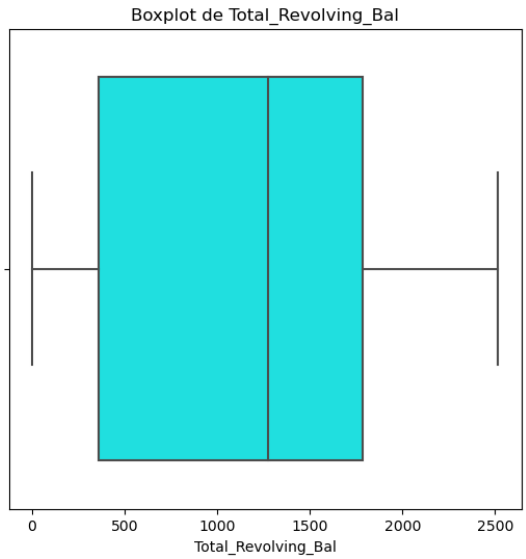
- **Credit_Limit:** Existe una gran variabilidad en los límites de crédito. La distribución de esta variable es inclinada hacia la izquierda con un segmento de clientes con límites de crédito excepcionalmente altos. Los clientes con un mayor límite pueden tener perfiles financieros más estables.





- **Total_Revolving_Bal:** Para esta variable podemos ver según su distribución que la mayoría de los clientes cuenta con balances totales iguales a cero, mientras que aquellos que tienen balances negativos rondan los 1500 dólares. Esto puede ser relevante ya que el saldo de crédito rotativo podría implicar dependencia en la tarjeta de crédito o una mayor probabilidad de riesgo financiero, lo cual se puede relacionar con la probabilidad de abandono.

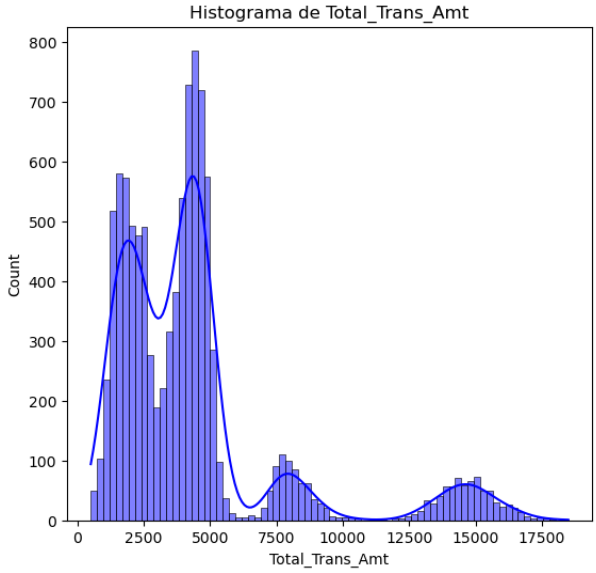


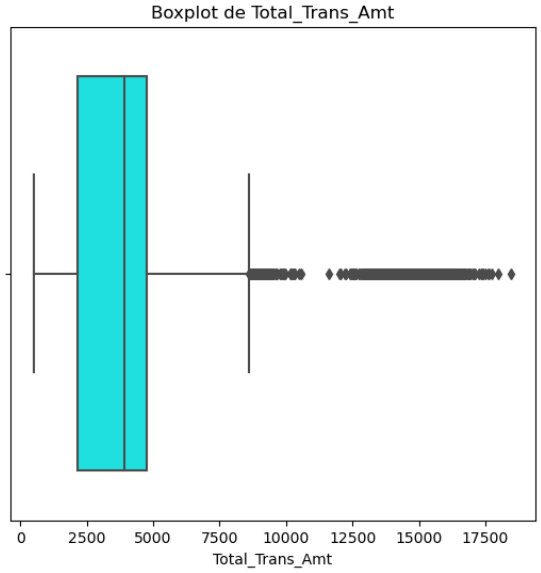


- **Avg_Open_To_Buy:** El monto disponible para gastar refleja la capacidad de endeudamiento restante del cliente. Aquellos con menos disponible podrían estar más endeudados, lo cual puede incidir en el abandono del servicio.

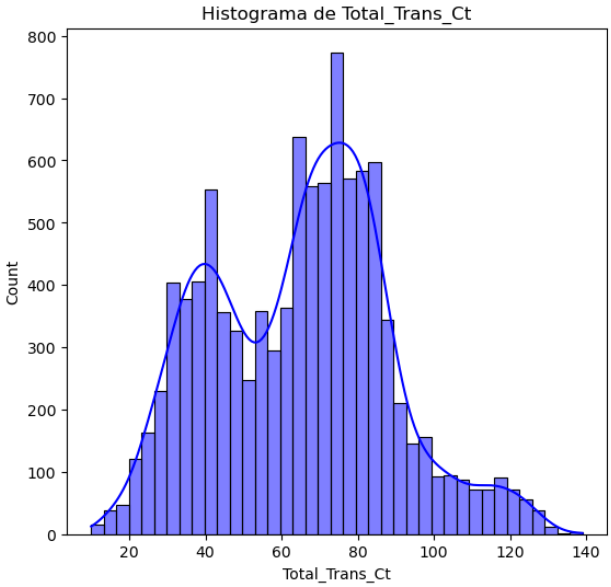
- **Total_Amt_Chng_Q4_Q1:** Este cambio en el monto de gasto entre el cuarto y primer trimestre tiene una media de 0.75, lo que relfeja refleja el dinamismo del cliente en el uso de su tarjeta. Un cambio bajo podría indicar una disminución en la utilización de la tarjeta.

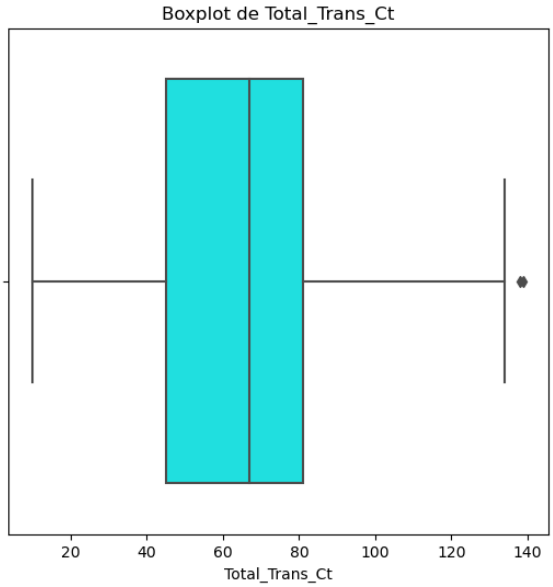
- **Total_Trans_Amt:** Al revisar el monto de las transacciones de los clientes vemos que no hay una distribución centralizada, sino que se presentan 4 modas en los datos.





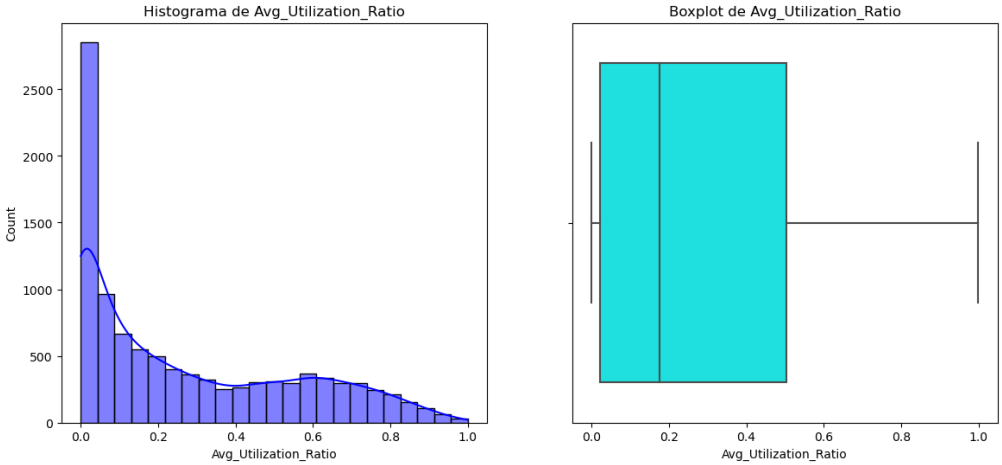
- **Total_Trans_Ct:** La frecuencia de uso de la tarjeta es importante, ya que los clientes con baja frecuencia pueden considerar alternativas financieras. Por ello revisamos el conteo de transacciones por cliente y encontramos una distribución bimodal que ronda las 40 y las 80 transacciones por cliente.



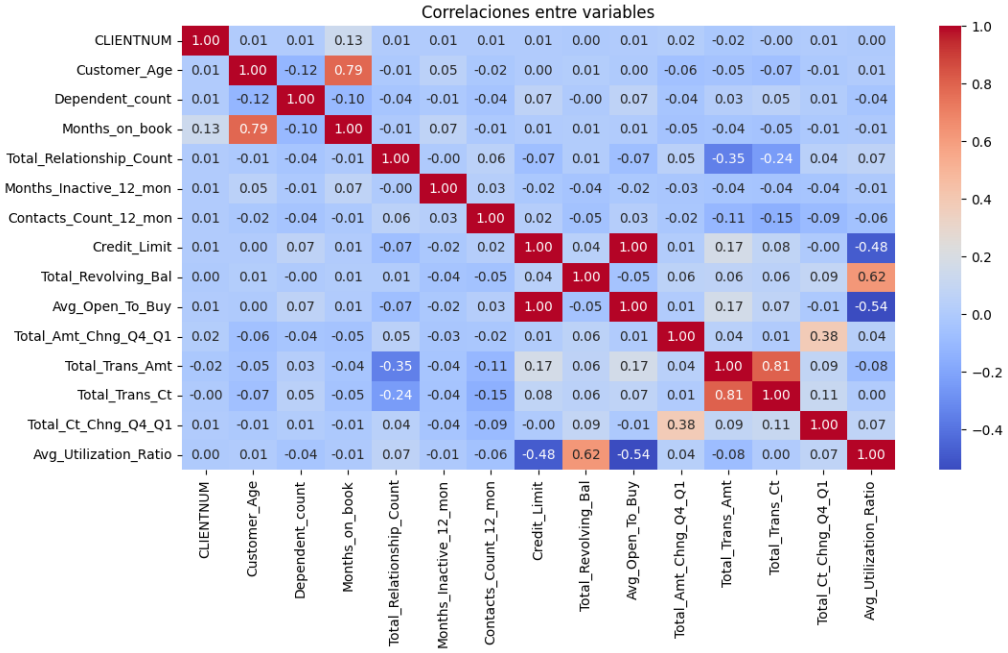


- **Total_Ct_Chng_Q4_Q:** Este cambio en el recuento de transacciones trimestral tiene un valor medio de 0.71, lo cual sugiere cambios en los hábitos de los clientes, donde una reducción podría ser una señal de desinterés en el producto

- **Avg_Utilization_Ratio:** La utilización promedio de crédito es un indicador clave de comportamiento financiero. Una alta utilización puede ser indicio de una dependencia en la tarjeta, pero a su vez un riesgo financiero. Para este caso en específico vemos que la distribución de la variable es completamente sesgada hacia el 0, lo que nos indica que la mayoría de los clientes tiende a darle baja utilización a su cupo de tarjeta de crédito.



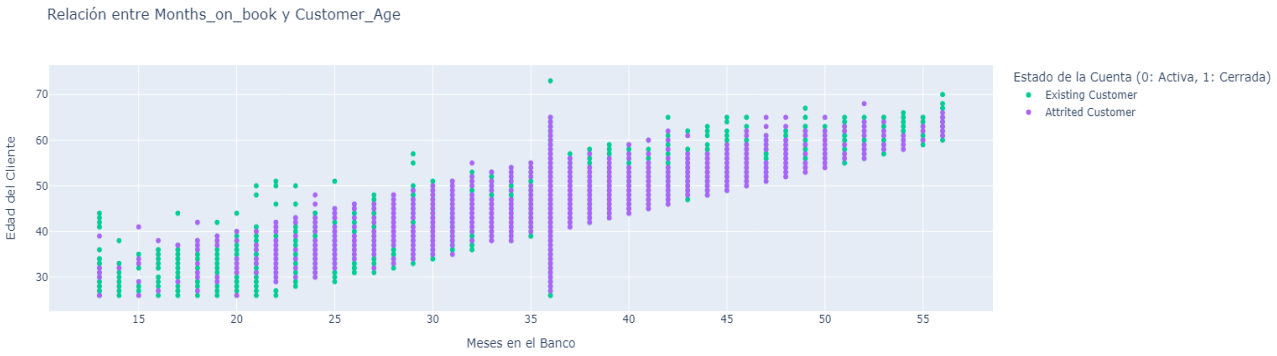
Correlación entre variables



La matriz de correlación revela una fuerte relación positiva entre 'Credit_Limit' y 'Avg_Open_To_Buy', lo cual es esperable, ya que un mayor límite implica más crédito disponible. También muestra una correlación negativa levemente moderada entre 'Avg_Utilization_Ratio' y 'Credit_Limit'-'Avg_Open_To_Buy', sugiriendo que algunos de los clientes con límites de crédito altos tienden a usar una menor proporción de su crédito. Por último, se evidencia una correlación positiva entre 'Avg_Utilization_Ratio' y 'Total_Revolving_Bal'', mostrando que las tarjetas de crédito cuando se usan con más frecuencia tienden a aumentar el saldo rotatorio.

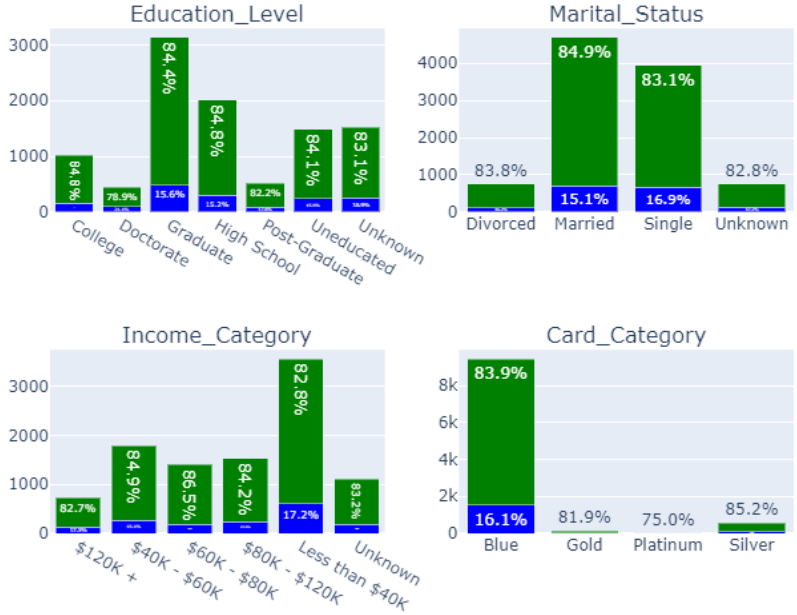
Análisis por variable objetivo

Con el objetivo de evaluar si existen patrones entre las variables para los grupos de clientes que abandonan el banco y los que no, realizamos un análisis descriptivo separando los clientes en estas dos categorías.

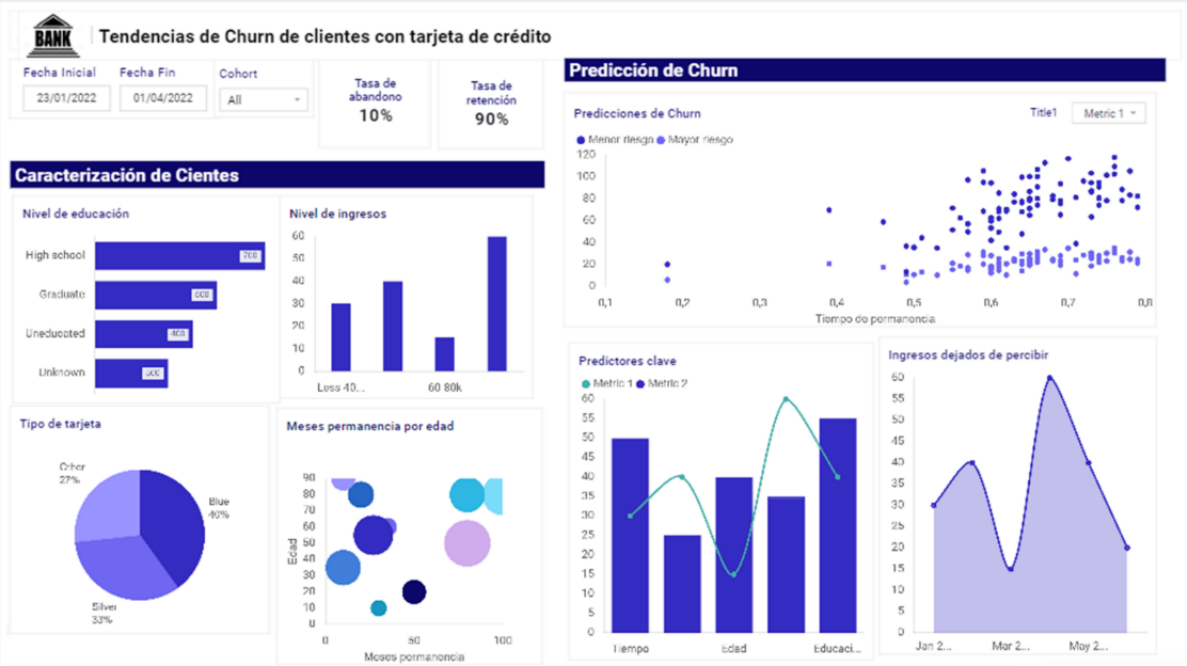


Las variables 'Months_on_book' y 'Customer_Age' reflejan una de las relaciones más altas, dado que parece natural que conforme mayor es la edad del cliente mayor sea número de periodos de relación con el banco. Por otro lado, al evaluar como esta relación se distribuye respecto a la variable dependiente se observa que un mayor número de clientes parecen cerrar el producto con mayor frecuencia al ser más jóvenes y tener un menor número de periodos de relación con el banco.

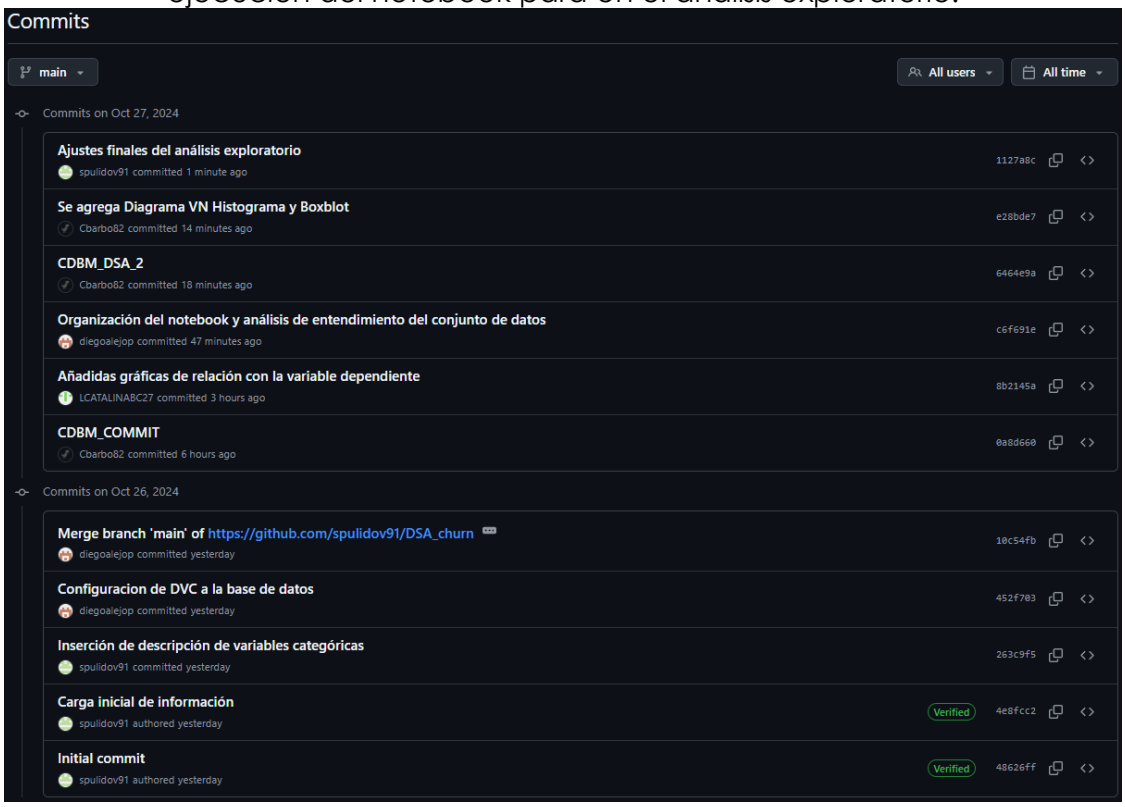
A su vez, al revisar la distribución de variables categóricas respecto a las variables dependientes se observa que en la mayoría de los casos la participación promedio entre categorías de clientes que abandonan el producto es similar. Algunos comportamientos destacados señalan que: hay un mayor desistimiento en las mujeres que en los hombres, con un 17.4% de casos respecto a un 14.6% respectivamente, las personas con doctorado y posgrado reflejan un porcentaje más alto de retiro con 21.1% y 17.8%, los clientes con ingresos menores a $40K revelan una proporción más alta de retiro (17.2%) y el tipo de tarjeta 'Platinum' pese a tener la menor participación de clientes cuenta con la mayor proporción de clientes que desisten de la tarjeta (25%).



Maqueta del prototipo.



Evidencia de11 commits por parte del equipo de trabajo para la creación de archivos y ejecución del notebook para en el análisis exploratorio.



En general se ve que el aporte de todos los miembros del equipo al trabajo fue parejo y equitativo.

Reporte de trabajo en equipo.

Actividad	S. Pulido	L. Briceño	C. Barbosa	D. Peñaloza
Escoger problema	X	X	X	X
Pregunta de negocio	X	X	X	X
Creación de repositorio en git	X			
Creación de DVC				X
EDA Entendimiento de los datos				X
EDA Variables numéricas	x		X	
EDA Variables categóricas	X			
Maqueta del prototipo		X		