UNIVERSIDAD NACIONAL DE INGENIERIA

FACULTAD DE INGENIERÍA ECONÓMICA, ESTADÍSTICA Y CIENCIAS SOCIALES

XVI PROGRAMA DE ESPECIALIZACION EN BUSINESS INTELLIGENCE & BUSINESS ANALYTICS

Análisis de la deserción de clientes de tarjetas de crédito mediante técnicas de Business Intelligence y Business Analytics

TRABAJO DE FIN DE PROGRAMA

AUTOR(ES):

Lino Moreno, Smith Eusebio

Campos Nuñez, Nicole Margarita

LINK DE EXPOSICIÓN

https://youtu.be/rfe7eyk0naE

Lima, 13 de septiembre del 2025

ÍNDICE

1. Descripción del caso de uso

1.1. Contexto del negocio y sector financiero

El negocio de tarjetas de crédito se desarrolla en un ecosistema financiero altamente competitivo, donde la gestión eficiente de clientes es clave para la rentabilidad sostenida. Las entidades emisoras enfrentan presiones crecientes por ofrecer productos diferenciados, adaptarse a la digitalización del consumo y mantener márgenes en medio de regulaciones cambiantes. En América Latina, el avance de las billeteras digitales y neobancos ha transformado el panorama competitivo, obligando a los emisores tradicionales a replantear sus estrategias de fidelización y riesgo (Portafolio, 2023). Además, estudios recientes advierten que el incremento de tasas de interés, la regulación sobre comisiones y el aumento del fraude cibernético han generado una menor confianza en el producto por parte de los usuarios (Barrantes Caballero, 2025). En el Perú, el Banco Central ha reportado una reestructuración del perfil crediticio de los clientes, priorizando aquellos con mejor comportamiento de pago, como respuesta al riesgo de sobreendeudamiento y morosidad elevada en este segmento (BCRP, 2022).

1.2. Fundamentación del problema

La Superintendencia de Banca, Seguros y AFP informa que después de una contracción de 0.4 % en 2023 la economía peruana creció 3.3 % en 2024 y se proyecta una expansión de 3.2% para 2025. Este impulso macroeconómico ha fortalecido la solidez del sistema financiero con niveles de solvencia que pasaron de 14.7 % en diciembre de 2019 a 16.8 % en marzo de 2025. La tasa de morosidad alcanzó un pico de 4.97 % en mayo de 2024 y desde entonces exhibe una trayectoria descendente, mientras que los ratios de incumplimiento muestran una mejora gradual tras los niveles históricos del año anterior (SBS, 2025).

De acuerdo con el reporte de inclusión financiera de la SBS el número de deudores de tarjetas de crédito creció 3.34 % respecto a 2023 y la proporción de tarjeta-habientes sobre población adulta se redujo en 0.68 %. El análisis por género revela que la participación de mujeres deudoras de la población económicamente activa cayó 0.49 puntos porcentaje mientras la de los hombres aumentó 0.51 % (SBS, 2024).

La industria bancaria global opera en un entorno dinámico y altamente competitivo, donde la retención de clientes se ha consolidado como un pilar crítico para la rentabilidad y la sostenibilidad financiera (Viswadhanush, 2025). La creciente competencia en la industria financiera y la diversificación de opciones para los consumidores han reducido la lealtad, obligando a los bancos a implementar modelos predictivos para anticipar comportamientos y retener a los usuarios existentes (Li & Yan, 2025). El costo de adquirir un nuevo cliente es significativamente mayor que el de retener uno existente, lo que convierte a la gestión del churn de clientes en una prioridad estratégica fundamental (AL-Najjar et al., 2022a). Estudios destacan que la retención de clientes es esencial, dado que el churn reduce las ganancias del sistema bancario, y el análisis de datos reales

permite identificar variables clave, como el número de transacciones y saldos pendientes, para mitigar este riesgo (Nie et al., 2009).

Para predecir la deserción de clientes de tarjetas de crédito, Miao y Wang (2022) compararon Random Forest, regresión lineal y KNN y hallaron que Random Forest afinado por grid search alcanzó una exactitud de 96,25 %, identificando como variables más influyentes el importe total de transacciones, el número de operaciones y el saldo revolvente. Peng, Peng y Li (2023) emplearon un ensamble GA-XGBoost potenciado con SMOTEENN para corregir el desbalance de clases y consiguieron un F1-score de 0,90 junto a un AUC de 0,99, lo que confirma la eficacia de combinar boosting con sobremuestreo para aislar clientes en riesgo de abandono. AL-Najjar, Al-Rousan y AL-Najjar (2022) fusionaron múltiples variables categóricas en un solo atributo y evaluaron cinco clasificadores, de los cuales el árbol C5 obtuvo 96,4 % de precisión usando el conteo de transacciones, el saldo revolvente y la variación en la frecuencia de uso como predictores clave. Tran Hoang Hai, Vu Van Thieu y Doan Minh Hieu (2024) validaron KNN, Random Forest, AdaBoost y CNN-1D con SMOTE y reportaron precisiones de test superiores al 80 %. Chang, Hall, Gao y Uchenna (2022) añadieron una dimensión temporal a AdaBoost y Random Forest para modelar la evolución de saldos y de transacciones, incrementando en más de un 10 % la detección temprana de churners. Panduro-Ramírez et al. (2022) aplicaron SMOTE y ajustaron CatBoost mediante grid search para alcanzar un 97,85 % de exactitud en test. Chen (2024) comparó nueve clasificadores e identificó a XGBoost como el más robusto, con 97 % de acierto en test y un F1-score de 0,92. Ram Kumar et al. (2023) construyeron híbridos de regresión logística con KNN y con árboles de decisión, logrando F1-scores de 0,90 y 0,928 respectivamente. Sin embargo, los métodos tradicionales de machine learning resultan insuficientes para una gestión efectiva, ya que se necesita mayor interpretabilidad en los modelos para comprender las correlaciones y causalidades detrás de la deserción (Demirberk, 2021). Por ejemplo, investigaciones que aplican técnicas como regresión logística y árboles de decisión subrayan que el churn en tarjetas de crédito es un problema relevante en el sector bancario, donde perder clientes es costoso y adquirir nuevos resulta aún más oneroso, promoviendo el uso de data mining para pronosticar y prevenir la deserción (Gupta et al., 2022). Asimismo, el uso de clasificadores avanzados como XGBoost ha demostrado alta precisión al identificar clientes propensos al churn, permitiendo a los bancos ofrecer servicios personalizados que reduzcan pérdidas corporativas (Jovanovic et al., 2023). Finalmente, la dinámica de un mercado competitivo, con múltiples proveedores alternos y tasas de interés bajas, fomenta la migración de clientes, lo que resalta la importancia de seleccionar características relevantes en modelos predictivos para optimizar estrategias de retención (Ahmad et al., 2019).

1.3. Objetivos del caso

1.3.1. Objetivo general

Diseñar un modelo predictivo que permita anticipar la deserción de clientes de tarjetas de crédito en el sector bancario, identificando con alta precisión a los usuarios en riesgo de abandono y facilitando la adopción temprana de estrategias de retención.

1.3.2. Objetivos específicos

- Realizar un análisis exploratorio de los determinantes transaccionales, demográficos y de uso de tarjetas, con el fin de seleccionar las variables más informativas para el modelo.
- Comparar el desempeño de distintos algoritmos de clasificación , evaluando métricas clave como precisión, recall y AUC.
- Aplicar técnicas de re-muestreo para corregir el desbalance en la proporción de clientes churners versus no churners y optimizar la capacidad predictiva de los modelos.
- Incorporar métodos de interpretabilidad que expliquen la contribución de cada variable al riesgo de churn, de modo que se puedan priorizar las intervenciones.
- Proponer recomendaciones tácticas de fidelización basadas en los hallazgos del modelo, dirigidas a los segmentos de clientes identificados como de alto riesgo de abandono.

2. Descripción del conjunto de datos

Para el desarrollo del caso de estudio se utilizó el conjunto de datos "*Credit Card Customers*", disponible en la plataforma Kaggle. Este dataset contiene información de 10,127 clientes de un banco, incluyendo variables demográficas, socioeconómicas y de comportamiento financiero, lo que lo convierte en una fuente adecuada para el análisis de deserción de clientes.

Enlace de la base de datos:
https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

A continuación, se presenta una tabla que detalla las variables incluidas en el conjunto de datos, junto con sus respectivas descripciones:

Tabla 1

Variables del dataset Credit Card Customers

Variable	Descripción
CLIENTNUM	Identificador único del cliente
Attrition_Flag	Evento interno: 1 si la cuenta está cerrada, 0 si no
Customer_Age	Edad del cliente en años
Gender	Género: M=Masculino, F=Femenino
Dependent_count	Número de dependientes
Education_Level	Nivel educativo del titular de la cuenta
Marital_Status	Estado civil: Casado, Soltero, Divorciado, Desconocido
Income_Category	Categoría de ingresos anuales del titular
Card_Category	Tipo de tarjeta: Blue, Silver, Gold, Platinum

Months_on_book	Período de relación con el banco
Total_Relationship_Count	Número total de productos que tiene el cliente
Months_Inactive_12_mon	Número de meses inactivo en los últimos 12 meses
Contacts_Count_12_mon	Número de contactos en los últimos 12 meses
Credit_Limit	Límite de crédito de la tarjeta
Total_Revolving_Bal	Saldo rotativo total en la tarjeta de crédito
Avg_Open_To_Buy	Línea de crédito disponible (promedio de los últimos 12 meses)
Total_Amt_Chng_Q4_Q1	Cambio en el monto de transacciones (Q4 sobre Q1)
Total_Trans_Amt	Monto total de transacciones (últimos 12 meses)
Total_Trans_Ct	Cantidad total de transacciones (últimos 12 meses)
Total_Ct_Chng_Q4_Q1	Cambio en la cantidad de transacciones (Q4 sobre Q1)
Avg_Utilization_Ratio	Tasa promedio de utilización de la tarjeta

3. Análisis exploratorio de los datos (EDA)

3.1. Inspección inicial

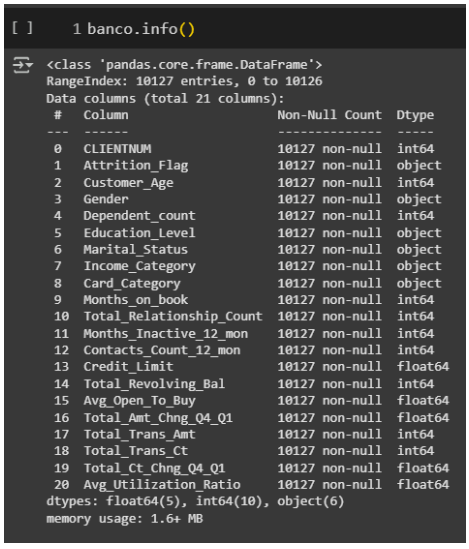
El conjunto de datos “Credit Card Customers” contiene inicialmente 23 variables, de las cuales se eliminaron dos, relacionadas con un modelo de Naive Bayes Classifier, por no ser relevantes para el presente caso de estudio. Las 21 variables restantes se dividen en:

- Categóricas: Attrition_Flag, Gender, Education_Level, Marital_Status, Income_Category, Card_Category.
- Númericas: Customer_Age, Dependent_count, Months_on_book, Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon, Credit_Limit, Total_Revolving_Bal, Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Ct_Chng_Q4_Q1, Total_Trans_Ct, entre otras.

En la **Figura 1** se detallan las variables utilizadas y sus descripciones:

Figura 1

Variables categóricas y numéricas

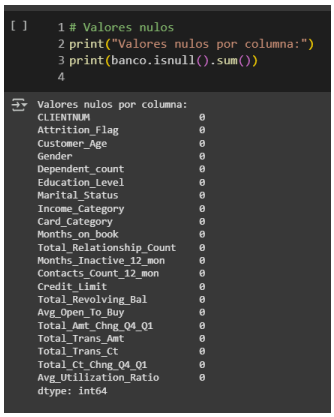


```
[ ]    1 banco.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  object
 6   Marital_Status            10127 non-null  object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
```

Como parte de la inspección inicial, se verificó la presencia de valores nulos, se analizó la existencia de valores faltantes en cada columna utilizando el método "*isnull().sum()*" en Python.

Figura 2

Valores nulos

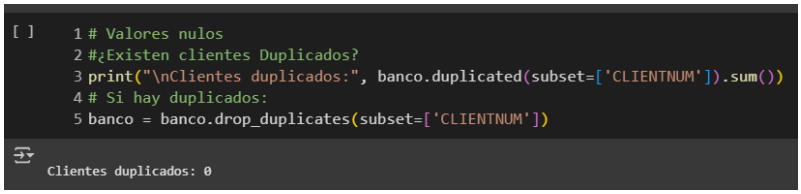


Los resultados muestran que ninguna de las 21 variables presenta valores nulos, lo que indica que el dataset está completo y no requiere imputación de datos.

Se comprobó la unicidad de los registros mediante la variable CLIENTNUM, que actúa como identificador único.

Figura 2

Valores duplicados



No se encontraron clientes duplicados, confirmando que cada registro corresponde a un cliente único. Por lo tanto, el conjunto de datos cuenta con 10,127 clientes únicos, de los cuales se analizan 6 variables categóricas y 15 variables numéricas

3.2. Preprocesamiento

Para preparar el conjunto de datos para el modelado, se realizaron los siguientes pasos de preprocesamiento:

1. Análisis de Variables Categóricas: Se evaluó la cantidad de categorías y valores únicos en las variables categóricas:

- Attrition_Flag: 2 valores únicos (Existing Customer, Attrited Customer).
- Gender: 2 valores únicos (M, F).
- Education_Level: 7 valores únicos (High School, Graduate, Uneducated, Unknown, College, Post-Graduate, Doctorate).
- Marital_Status: 4 valores únicos (Married, Single, Unknown, Divorced).
- Income_Category: 6 valores únicos ($60K - $80K, Less than $40K, $80K - $120K, $40K - $60K, $120K +, Unknown).
- Card_Category: 4 valores únicos (Blue, Gold, Silver, Platinum).

Donde, se encontró como categoria al termino “Unknown” como categoría en varias variables, esta misma será imputada.

Figura 3

Suma de unknown

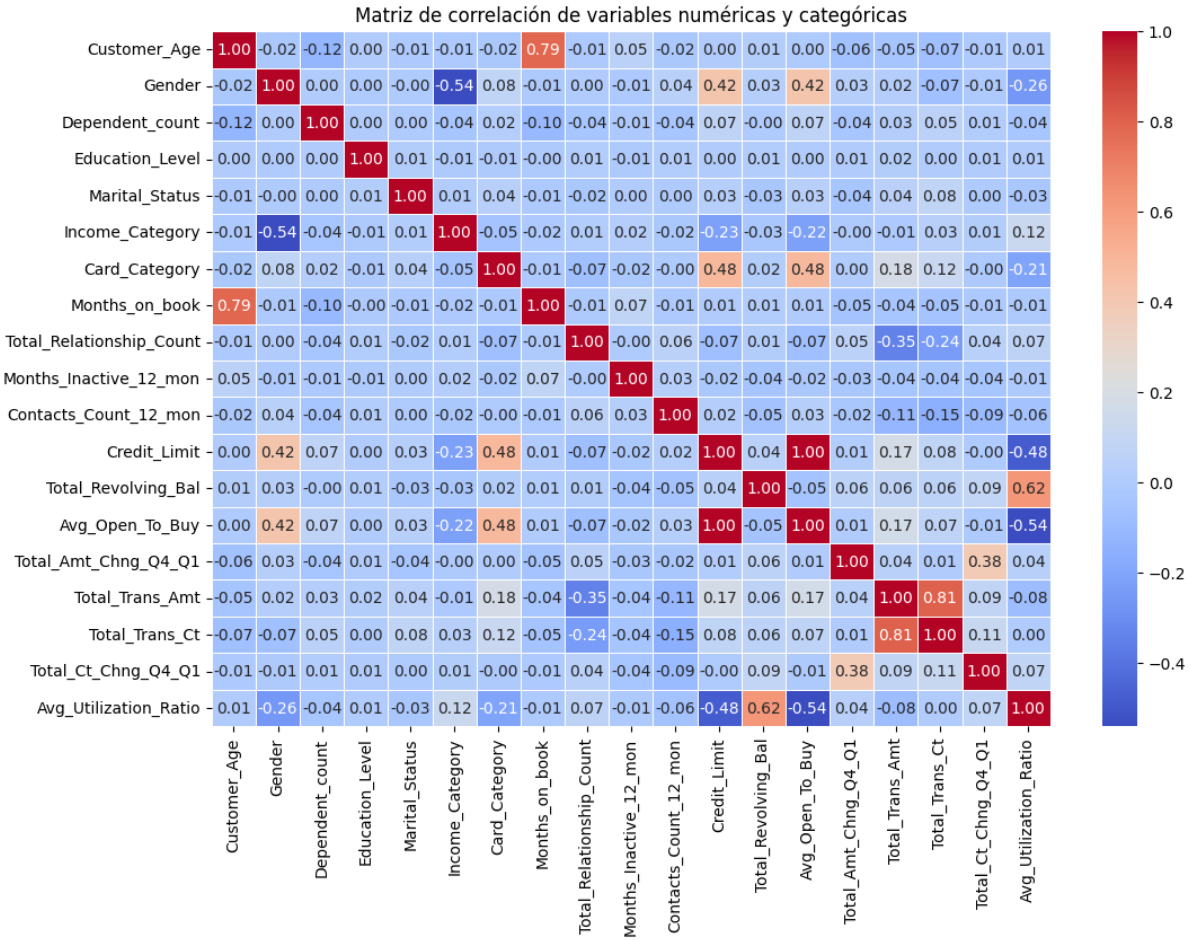
```
# Revision de valores imputables de las variables

# Conteo de 'Unknown' en todas las columnas
unknown_counts = (banco == "Unknown").sum()

# Filtrar solo las columnas donde existe al menos un 'Unknown'
unknown_counts = unknown_counts[unknown_counts > 0]

print(unknown_counts)
```

```
Education_Level    1519
Marital_Status      749
Income_Category    1112
dtype: int64
```

Para a imputación se realizo una matriz de correlación de variables, para conocer la relación directa o inversa de las variables a imputar con las demás variables del dataset.

Figura 4

Matriz de correlación de variables numéricas y categóricas



A continuación, se detallan las variables que tienen una correlación significativa con las variables desconocidas (Education_Code, Marital_Code e Income_Code), que podrían ser útiles para realizar una imputación adecuada de los valores faltantes. Las correlaciones directas son positivas, mientras que las inversas son negativas.

Tabla 2

Variables correlacionadas con Education_Code

Variable	Correlación
Total_Trans_Amt	0.0153 (directa)
Marital_Code	0.0147 (directa)
Total_Relationship_Count	0.0096 (directa)
Contacts_Count_12_mon	0.0085 (directa)
Total_Revolving_Bal	0.0080 (directa)
Total_Ct_Chng_Q4_Q1	0.0073 (directa)
Avg_Utilization_Ratio	0.0065 (directa)
Total_Amt_Chng_Q4_Q1	0.0055 (directa)
Months_Inactive_12_mon	-0.0081 (inversa)
Income_Code	-0.0104 (inversa)

Tabla 3

Variables correlacionadas con Income_Code

Variable	Correlación
Total_Trans_Ct	0.0759 (directa)
Total_Trans_Amt	0.0446 (directa)
Avg_Open_To_Buy	0.0336 (directa)
Credit_Limit	0.0313 (directa)
Education_Code	0.0147 (directa)
Income_Code	0.0097 (directa)
Customer_Age	-0.0113 (inversa)
Months_on_book	-0.0121 (inversa)
Total_Relationship_Count	-0.0214 (inversa)
Total_Revolving_Bal	-0.0254 (inversa)
Avg_Utilization_Ratio	-0.0275 (inversa)
Total_Amt_Chng_Q4_Q1	-0.0362 (inversa)

Tabla 4

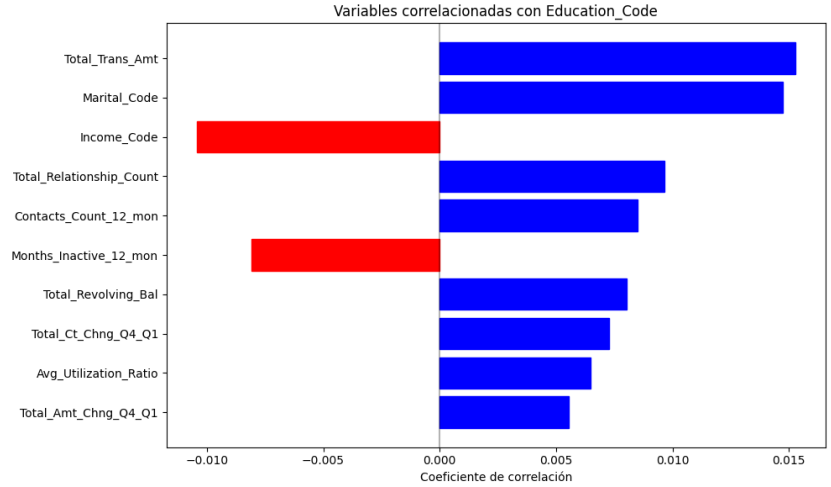
Variables correlacionadas con Education_Code

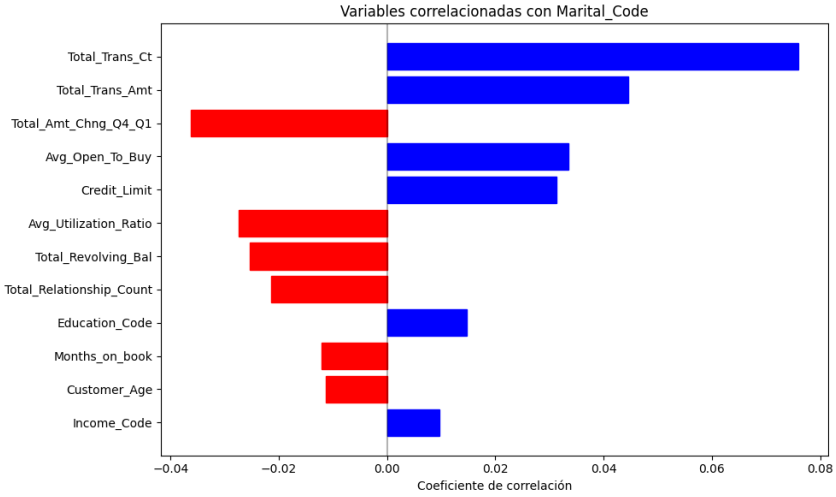
Variable	Correlación
Gender_Code	0.5397 (directa)
Avg_Utilization_Ratio	0.1233 (directa)
Total_Trans_Ct	0.0335 (directa)
Months_Inactive_12_mon	0.0240 (directa)
Total_Ct_Chng_Q4_Q1	0.0149 (directa)
Marital_Code	0.0097 (directa)
Total_Relationship_Count	0.0081 (directa)
Education_Code	-0.0104 (inversa)
Customer_Age	-0.0135 (inversa)
Total_Trans_Amt	-0.0147 (inversa)
Months_on_book	-0.0164 (inversa)
Contacts_Count_12_mon	-0.0184 (inversa)
Total_Revolving_Bal	-0.0258 (inversa)
Dependent_count	-0.0354 (inversa)
Avg_Open_To_Buy	-0.2230 (inversa)
Credit_Limit	-0.2254 (inversa)

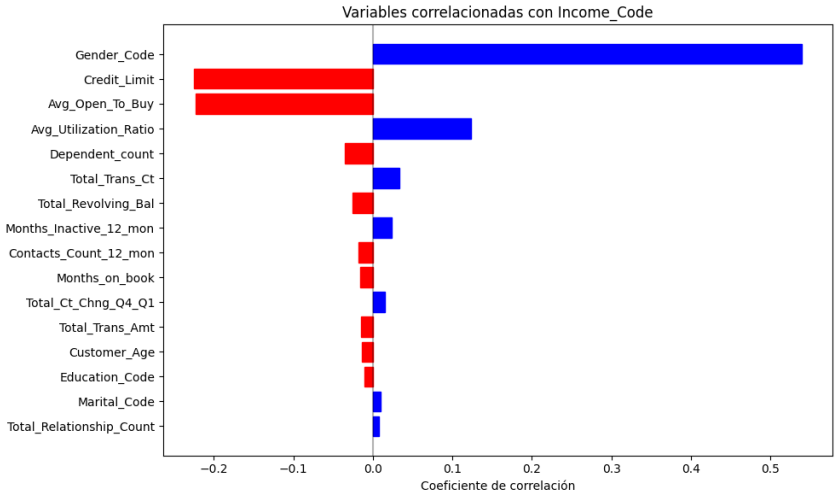
Estas relaciones ayudan a identificar qué variables podrían ser útiles para imputar valores faltantes, dado que las correlaciones directas indican una posible relación lineal, mientras que las inversas sugieren que las variables podrían estar asociadas de manera opuesta.

Figura 5

Correlación entre variables y variables objetivos



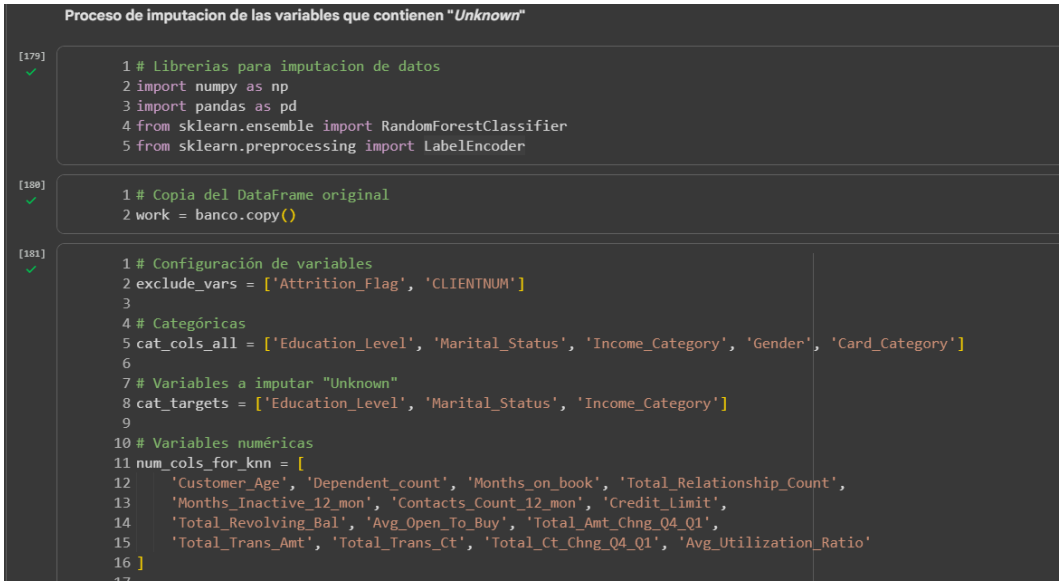




El proceso de imputación de las variables con "Unknown" se llevó a cabo utilizando un enfoque supervisado basado en RandomForestClassifier. Primero, se reemplazaron los valores "Unknown" por NaN en las variables Education_Level, Marital_Status e Income_Category. Luego, se seleccionaron las variables más correlacionadas como predictores para cada objetivo. Para cada variable a imputar, se entrenó un modelo de Random Forest con los registros completos y se utilizaron para predecir los valores faltantes en los registros con "Unknown". Los valores imputados fueron asignados de nuevo a las filas correspondientes y se guardaron los modelos entrenados para asegurar la consistencia y eficiencia en futuras imputaciones, logrando una imputación precisa y fiable de las variables categóricas.

Figura 6

Proceso de imputacion de las variables que contienen "Unknown"



El proceso de codificación de las variables categóricas se realizó utilizando la técnica Label Encoding, en la que cada categoría se convierte en un valor numérico. Este proceso es útil para convertir variables categóricas en un formato que pueda ser procesado por modelos de machine learning, especialmente aquellos que requieren entradas numéricas. A continuación, se detallan las variables categóricas codificadas y su mapeo:

Tabla 5

Codificación de Variables Categóricas

Variable	Categoría	Valor Codificado
Attrition_Flag	Attrited Customer	0
	Existing Customer	1
Gender	F (Femenino)	1
	M (Masculino)	0

Education_Level	College	0
	Doctorate	1
	Graduate	2
	High School	3
	Post-Graduate	4
	Uneducated	5
Marital_Status	Divorced	0
	Married	1
	Single	2
Income_Category	$120K +	0
	$40K - $60K	1
	$60K - $80K	2
	$80K - $120K	3
	Less than $40K	4
Card_Category	Blue	0
	Gold	1
	Platinum	2
	Silver	3

El proceso de binning o discretización se ha utilizado para transformar variables numéricas continuas en categorías (bins) que representan rangos de valores. Este enfoque permite que los modelos de machine learning trabajen con categorías en lugar de valores continuos, lo que puede ayudar a mejorar el rendimiento del modelo y hacer las predicciones más interpretables. A continuación se describen las variables numéricas y sus correspondientes categorías:

Tabla 6

Binning de Variables Numéricas

Variable	Bin 1	Bin 2	Bin 3	Bin 4
Customer_Age	0–19 (Joven)	20–38 (Adulto joven)	39–57 (Adulto)	≥58 (Adulto mayor)
Dependent_count	0–2 (Sin/pocos dep.)	3–4 (Familia pequeña)	5–6 (Familia mediana)	≥7 (Familia numerosa)
Months_on_book	0–15 (Cliente nuevo)	16–30 (Relación media)	31–45 (Cliente consolidado)	≥46 (Cliente histórico)
Total_Relationship_Count	0–2 (Poca vinculación)	3–4 (Relación moderada)	5–6 (Relación consolidada)	≥7 (Relación intensiva / full)
Months_Inactive_12_mon	0–2 (Inactividad baja)	3–4 (Inactividad moderada)	5–6 (Inactividad alta)	≥7 (Inactividad crítica)

Contacts_Count_12_mon	0–2 (Seguimiento bajo)	3–4 (Seguimiento moderado)	5–6 (Seguimiento alto)	≥7 (Seguimiento intensivo)
Credit_Limit	0–8750 (Límite bajo)	8751–17500 (Límite medio)	17501–26250 (Límite alto)	≥26251 (Límite premium)
Total_Revolving_Bal	0–630 (Saldo bajo)	631–1260 (Saldo medio)	1261–1890 (Saldo alto)	≥1891 (Saldo elevado)
Avg_Open_To_Buy	0–8750 (Capacidad baja)	8751–17500 (Capacidad media)	17501–26250 (Capacidad alta)	≥26251 (Capacidad premium)
Total_Amt_Chng_Q4_Q1	0–0.85 (Cambio leve)	0.86–1.7 (Cambio moderado)	1.8–2.55 (Cambio alto)	≥2.56 (Cambio muy alto)
Total_Trans_Amt	0–5000 (Volumen bajo)	5001–10000 (Volumen medio)	10001–15000 (Volumen alto)	≥15001 (Volumen muy alto)
Total_Trans_Ct	0–35 (Pocas transacciones)	36–70 (Frecuencia media)	71–105 (Frecuencia alta)	≥106 (Frecuencia intensiva)
Total_Ct_Chng_Q4_Q1	0–0.95 (Cambio leve)	0.96–1.9 (Cambio moderado)	2.0–2.85 (Cambio alto)	≥2.86 (Cambio muy alto)
Avg_Utilization_Ratio	0–0.25 (Baja utilización)	0.26–0.5 (Uso moderado)	0.51–0.76 (Uso alto)	≥0.77 (Uso intensivo)

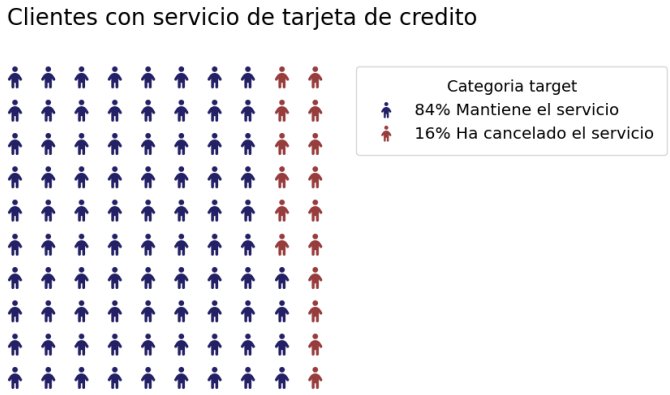
3.3. Análisis univariado

El análisis univariado se realizó para explorar la distribución de las variables categóricas y numéricas del conjunto de datos, identificando patrones y desbalances.

La variable Attrition_Flag indica si un cliente ha cancelado su servicio de tarjeta de crédito o si lo mantiene activo. A continuación, se presenta la distribución de esta variable.

Figura 7

Distribución de la variable Attrition_Flag

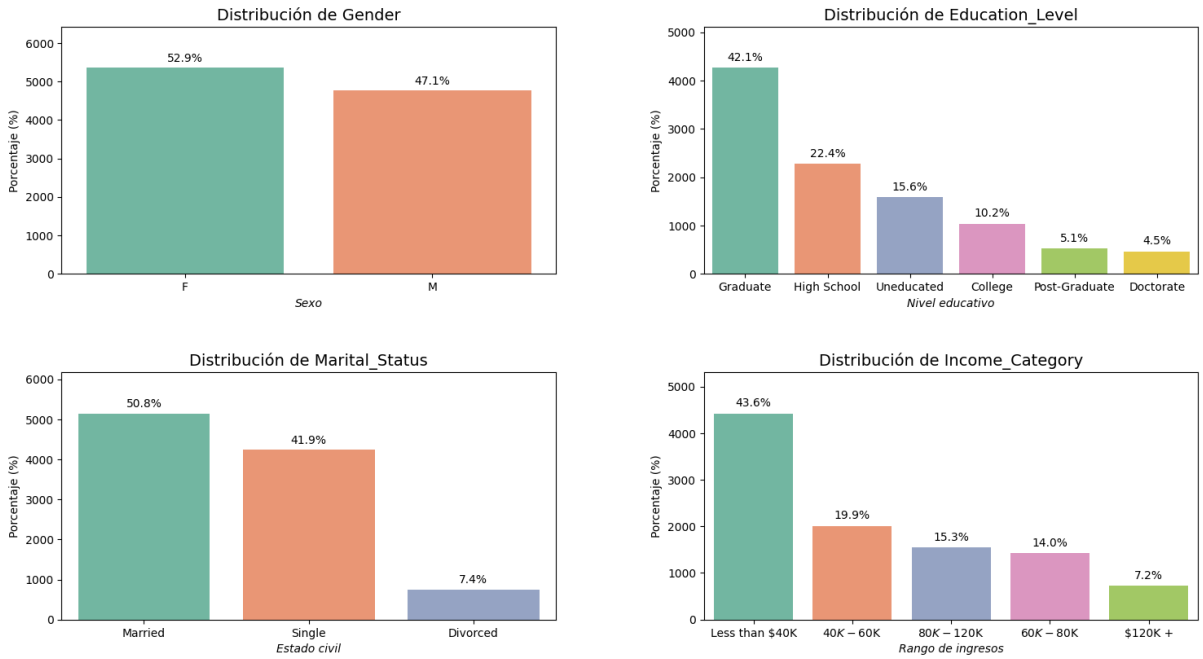


En la **Figura 7** se muestra la distribución de los clientes con servicio de tarjeta de crédito. Se observa que un 84% de los clientes mantienen el servicio (representados por iconos azules), mientras que el 16% ha cancelado el servicio (representados por iconos rojos). Esta distribución desbalanceada es importante para el modelado, ya que el modelo podría estar sesgado hacia la categoría mayoritaria (clientes activos).

El análisis de las variables categóricas permite comprender las características demográficas y socioeconómicas afectan al comportamiento de los clientes.

Figura 8

Distribución de variables categóricas



La Figura 8 ilustra las distribuciones de las variables categóricas:

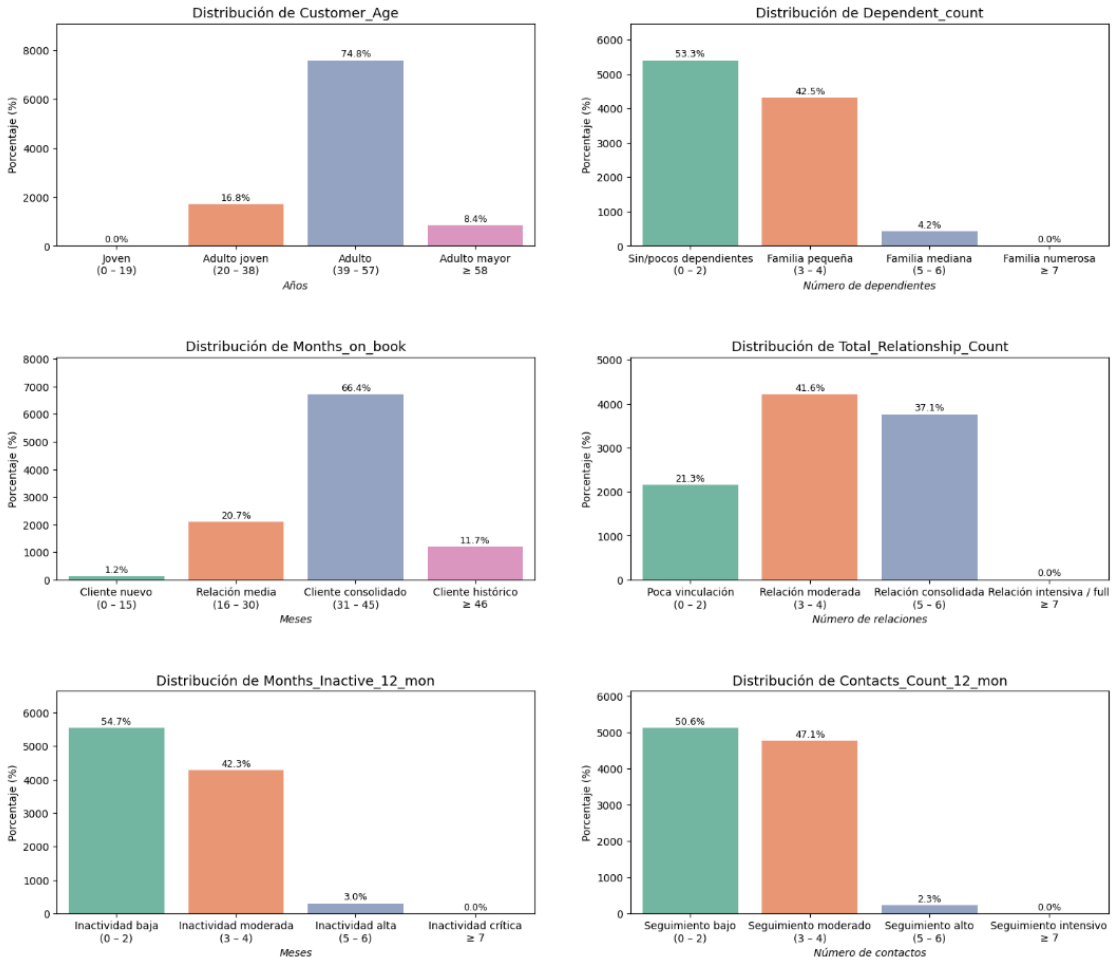
- En Gender, se observa una distribución equilibrada, con 52.9% de mujeres y 47.1% de hombres.

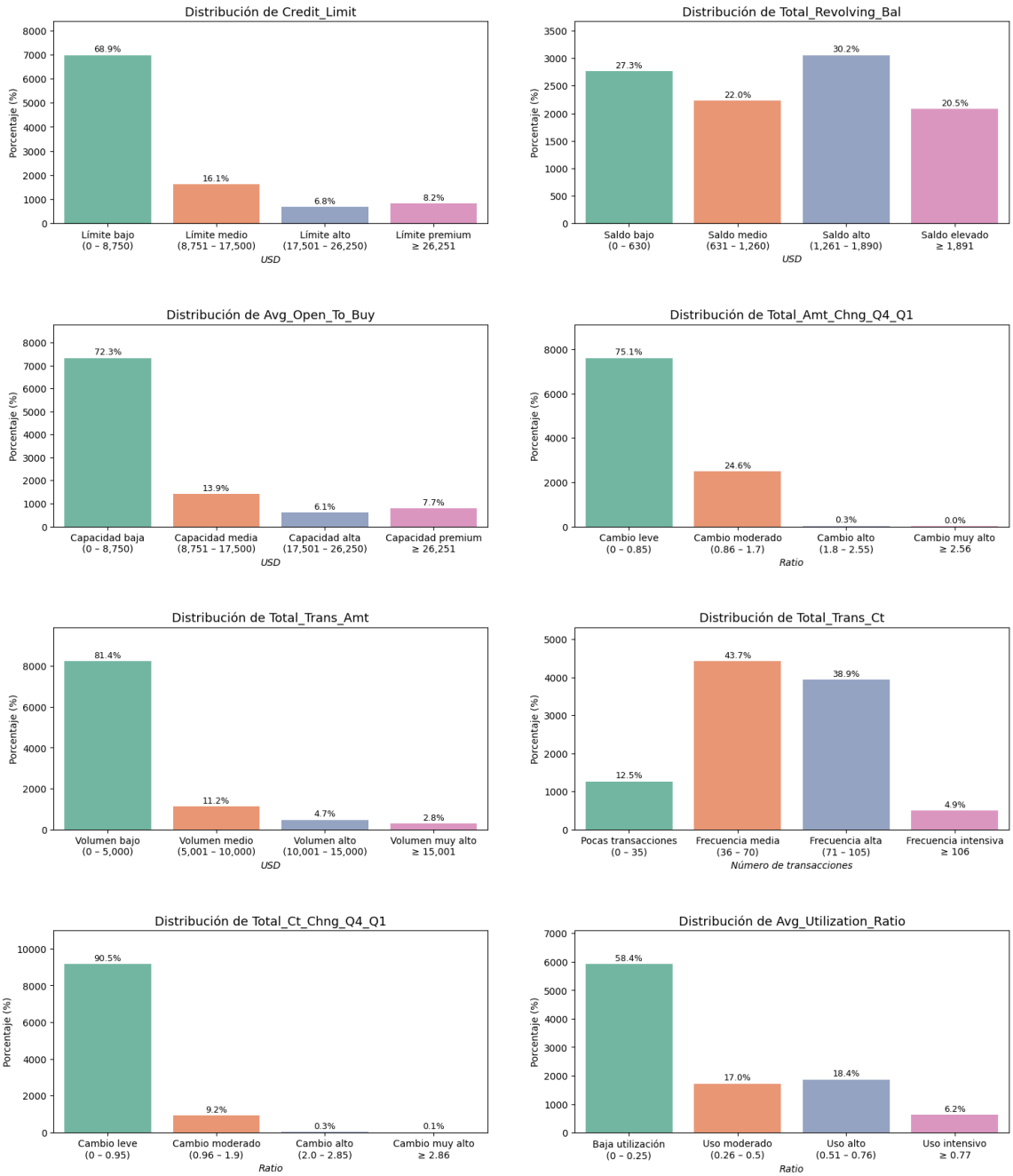
- Respecto a Education_Level, la categoría Graduate es la más común, con 42.1% de los clientes, seguida de High School con 22.4% y Uneducated con 15.6%.
- En la variable Marital_Status, el 50.8% de los clientes están casados, el 41.9% son solteros, y solo el 7.4% están divorciados.
- Por último, en Income_Category, el 43.6% de los clientes tiene ingresos inferiores a $40K, seguido por el 19.9% en el rango de 40K–60K.

El análisis de las variables numéricas ofrece información crucial sobre las características cuantitativas de los clientes, como la edad, el límite de crédito, el total de transacciones y el número de relaciones. Este análisis es fundamental para segmentar y comprender mejor los patrones de comportamiento de los clientes.

Figura 9

Distribución de variables numéricas





En la **Figura 9** se muestran las distribuciones de variables numéricas como Customer_Age, Dependent_count, Months_on_book, Total_Relationship_Count, entre otras.

- En Customer_Age, la mayoría de los clientes se encuentran en el rango de Adulto (39-57), con un 74.8% de los clientes, lo que indica que la base de clientes está predominantemente en edad adulta.
- En Dependent_count, el 53.3% de los clientes no tiene dependientes o tiene pocos, mientras que 42.5% tiene una familia pequeña (3-4 dependientes).

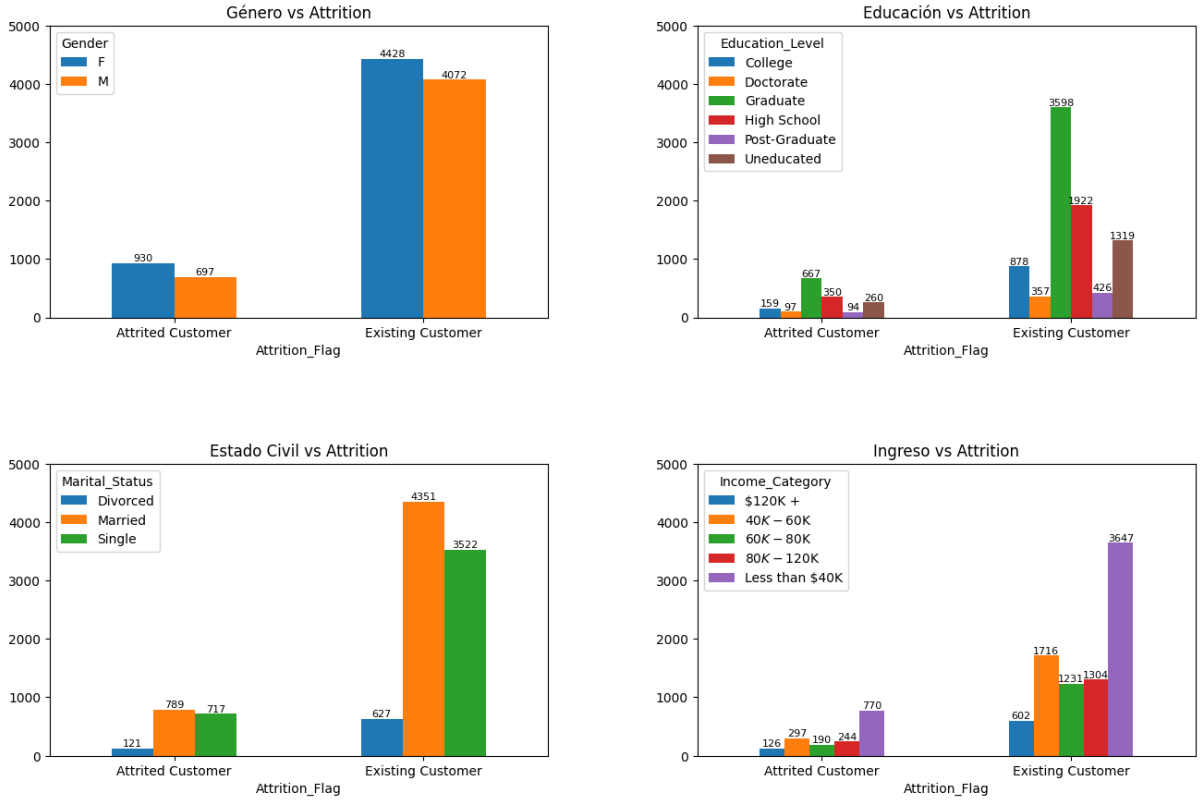
- En Months_on_book, el 66.4% de los clientes se encuentra en una relación consolidada (31-45 meses), lo que refleja que la mayoría de los clientes tienen una relación estable con la entidad financiera.
- En Total_Relationship_Count, el 41.6% tiene una relación moderada (3-4 relaciones), seguido por el 37.1% con una relación consolidada (5-6 relaciones).
- Finalmente, en Credit_Limit, el 68.9% de los clientes tiene un límite bajo (0–8,750 USD), lo que indica que la mayoría de los clientes cuentan con un límite de crédito limitado.

3.4. Análisis bivariado

El análisis bivariado permite estudiar las relaciones entre las variables y su impacto sobre la variable objetivo, Attrition_Flag. Este análisis es crucial para identificar patrones y relaciones que puedan ser útiles para mejorar la precisión del modelo. A continuación, se presentan tres conjuntos de gráficos que muestran cómo diferentes características demográficas, de comportamiento y transaccionales se correlacionan con la variable objetivo.

Figura 10

Distribución de Variables Demográficas vs Attrition



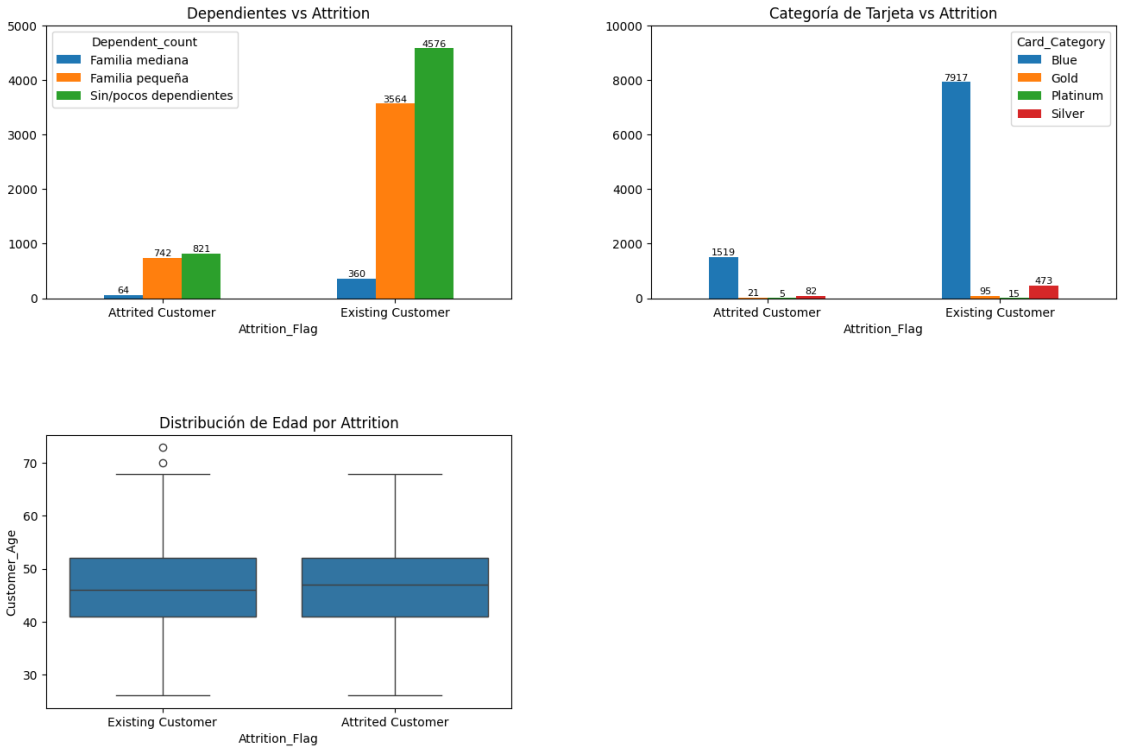
La **Figura 10** muestra la relación entre variables demográficas, como Gender, Education_Level, Marital_Status e Income_Category, con la variable objetivo Attrition_Flag. En los gráficos, se observa cómo las categorías de género, nivel

educativo, estado civil e ingresos se distribuyen entre los clientes que han cancelado el servicio y los que lo mantienen.

- Gender muestra que la mayoría de los Existing Customers son mujeres (con un 42% de mujeres y 41% de hombres).
- En Education_Level, se ve una clara diferencia en el nivel educativo entre los clientes que han cancelado el servicio y aquellos que lo mantienen.
- En Marital_Status, el 50% de los clientes activos están casados, mientras que un porcentaje menor de los que cancelaron el servicio están en esta categoría.
- Para Income_Category, la mayor parte de los clientes que han cancelado el servicio están en el rango de ingresos menos de $40K.

Figura 11

Distribución de edad, dependientes y attrition



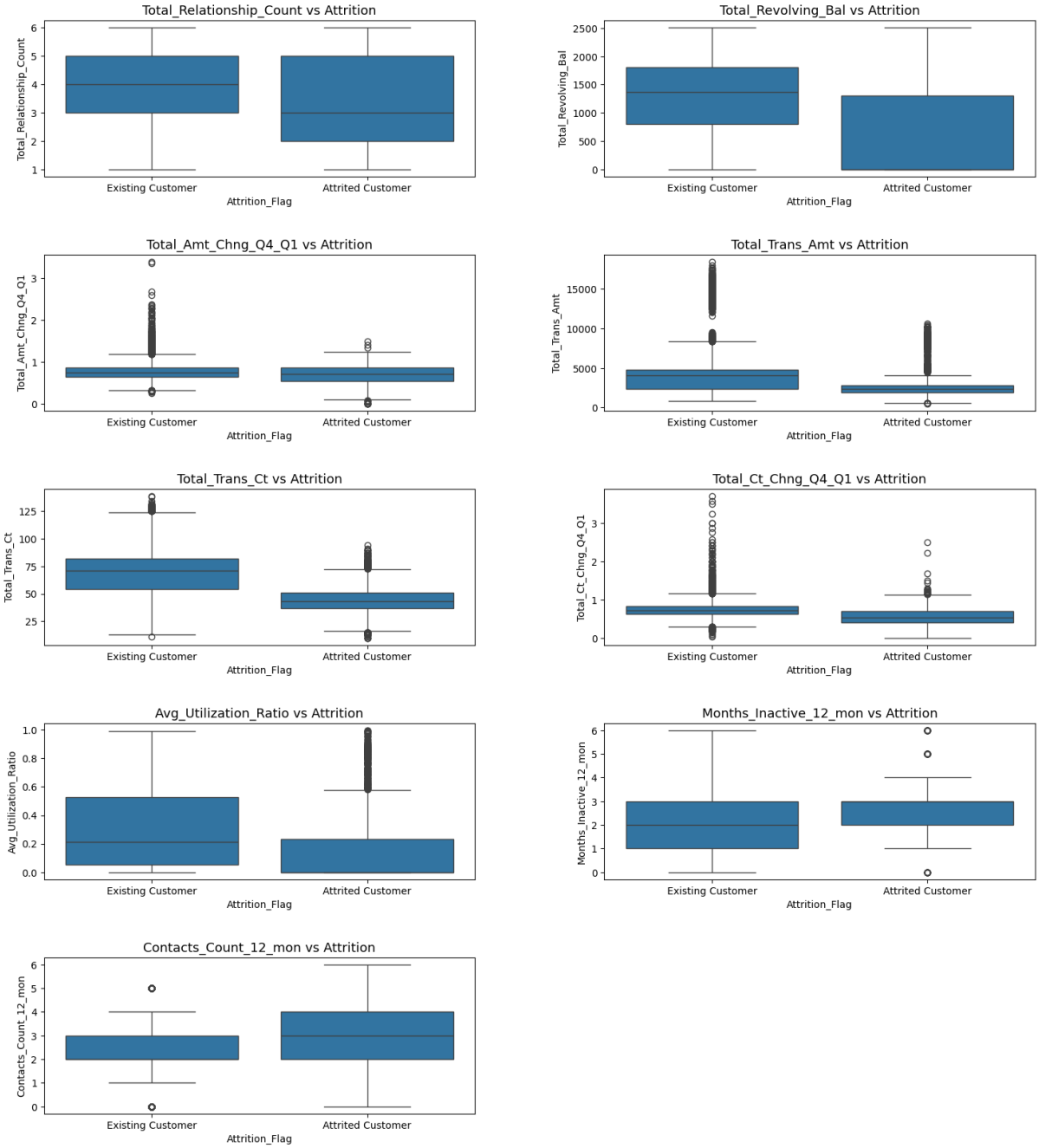
La **Figura 11** muestra la relación de Dependents, Age y Card Category con Attrition_Flag. Se observa que la mayoría de los clientes con más dependientes y con edades más altas tienden a mantener el servicio.

- Dependents muestra que los clientes sin dependientes o con pocos dependientes son más propensos a mantener el servicio.
- En Edad, no parece haber una gran diferencia en cuanto a la edad entre los que mantienen el servicio y los que lo han cancelado.

- Card_Category revela que la mayoría de los clientes que mantienen el servicio tienen tarjetas Blue, mientras que los clientes que han cancelado el servicio tienen más diversidad en las categorías de tarjetas.

Figura 12

Comportamiento transaccional por Attrition



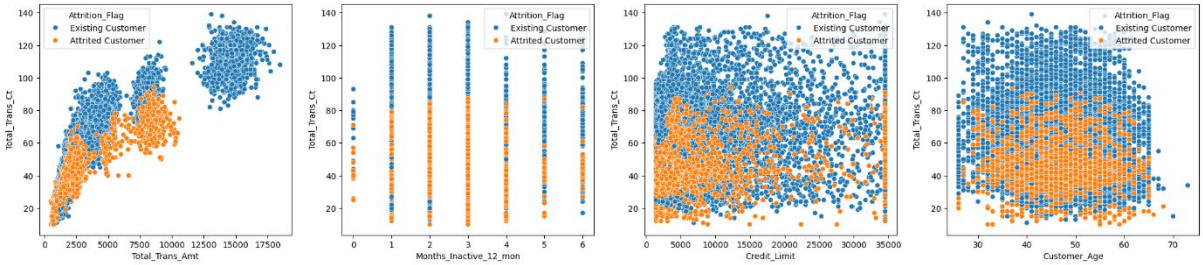
Finalmente, la Figura 12 ilustra cómo las variables relacionadas con el comportamiento transaccional, como el Total_Trans_Ct, Total_Revolving_Bal, y el Avg_Utilization_Ratio, afectan la cancelación del servicio. Los clientes que cancelaron el servicio tienden a tener un comportamiento transaccional distinto, como se observa en las distribuciones de estas variables.

- Total_Trans_Ct muestra que los clientes con mayor número de transacciones tienden a mantener el servicio.
- Total_Revolving_Bal indica que los clientes con balances rotatorios más altos tienen más probabilidades de cancelar el servicio.
- En Avg_Utilization_Ratio, se observa que los clientes que han cancelado el servicio tienen una utilización más alta de su crédito.

Para identificar los patrones que diferencias el comportamiento de los clientes, se analizó la relación entre la frecuencia transaccional (Total_Trans_Ct) y monto total de transacciones (Total_Trans_Amt), los meses inactivos (Months_Inactive_12_mon), el límite de crédito (Credit_Limit) y la edad del cliente (Customer_Age).

Figura 13

Frecuencia transaccional vs Attrition_Flag



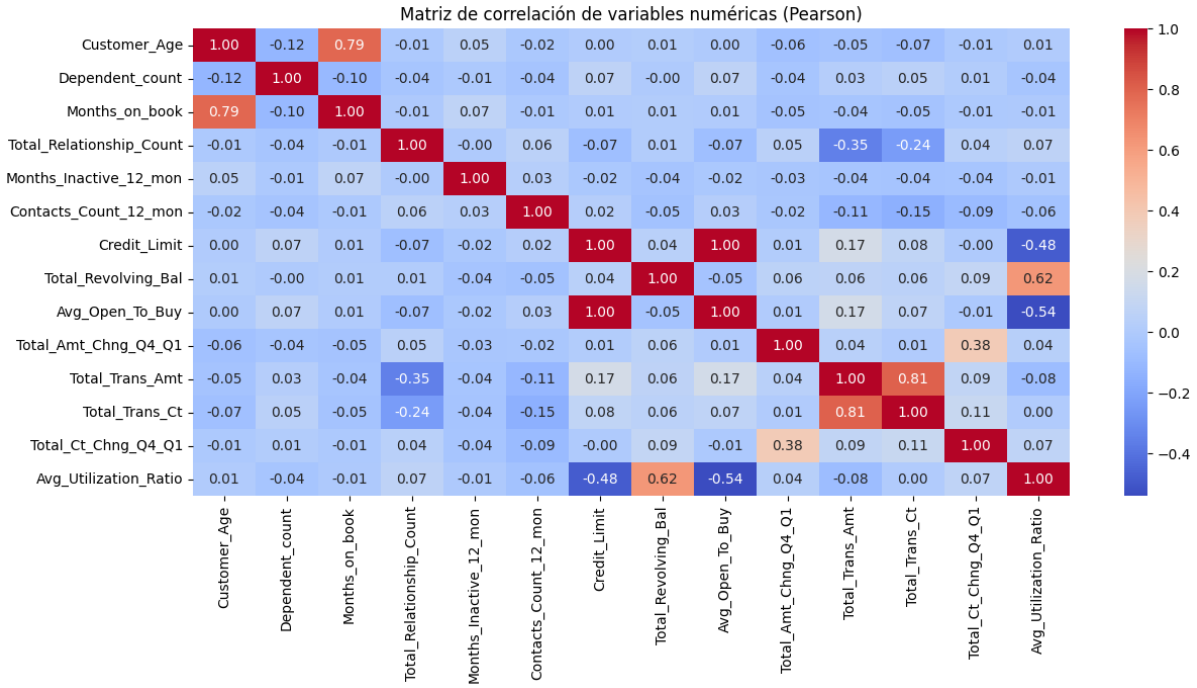
En la **Figura 13**, se observa una tendencia positiva entre Total_Trans_Ct y Total_Trans_Amt, donde los clientes activos tienden a concentrarse en áreas de mayor frecuencia y monto, mientras que los que desertan muestran valores más bajos en ambas dimensiones. Respecto a Months_Inactive_12_mon, los clientes que desertan se agrupan en zonas de mayor inactividad y menor frecuencia, sugiriendo que la inactividad prolongada reduce el engagement. En cuanto a Credit_Limit, los clientes activos exhiben una distribución más amplia hacia límites altos con frecuencias moderadas, mientras que los que desertan se concentran en límites bajos con menor actividad. Finalmente, la relación con Customer_Age indica que los clientes más jóvenes y mayores que desertan tienen frecuencias más bajas, mientras que los de edad media activa muestran mayor consistencia, destacando la edad como un factor relevante en el comportamiento transaccional.

3.5. Visualización de datos

En esta sección, se realiza un análisis exploratorio para identificar relaciones entre las variables, para las variables numéricas se utilizará Pearson y para categóricas Cramér V.

Figura 14

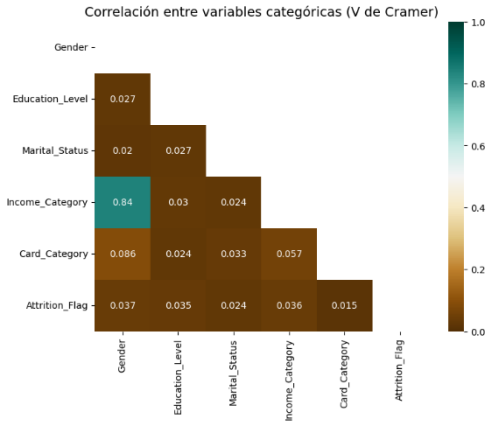
Matriz de correlación de variables numéricas



En la **Figura 14** se encontró que existen relaciones significativas entre varias variables financieras y de comportamiento del cliente. Destaca la fuerte correlación positiva entre Customer_Age y Months_on_book (0.79), así como entre Total_Trans_Amt y Total_Trans_Ct (0.81), lo que refleja patrones consistentes de antigüedad y actividad transaccional. Asimismo, se evidenció una correlación casi perfecta entre Credit_Limit y Avg_Open_To_Buy (0.99), indicando redundancia entre estas variables. También se observan correlaciones negativas relevantes, como la relación entre Avg_Utilization_Ratio y Avg_Open_To_Buy (-0.54), y entre Avg_Utilization_Ratio y Credit_Limit (-0.48), lo que sugiere que a mayor uso de la tarjeta, menor es la disponibilidad de crédito y que los clientes con límites más altos tienden a presentar una menor proporción de utilización. Además, variables como Dependent_count, Months_Inactive_12_mon y Contacts_Count_12_mon no muestran correlaciones fuertes con el resto, lo cual indica que aportan información independiente dentro del análisis.

Figura 15

Matriz de correlación de variables categóricas

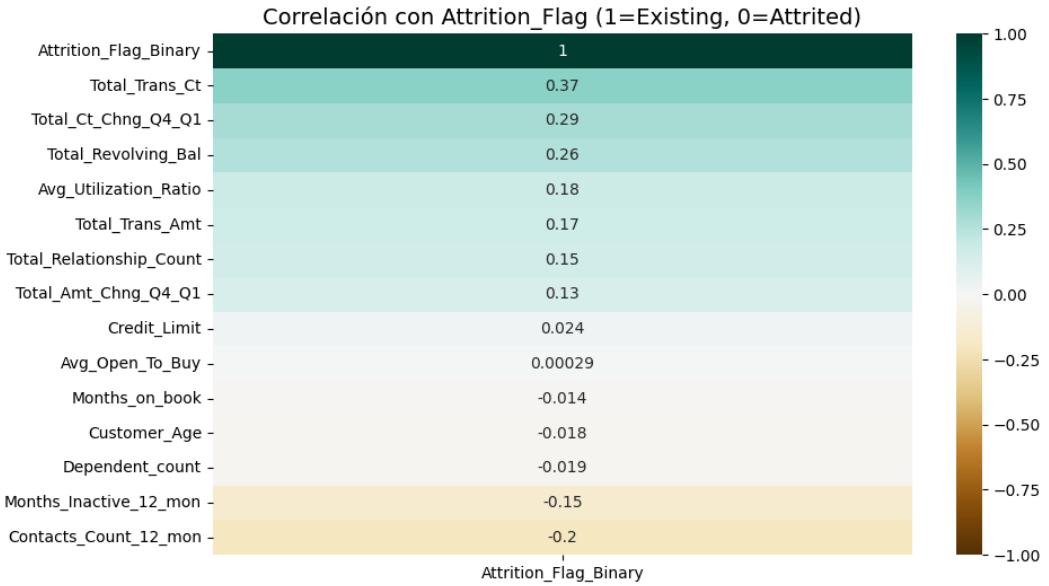


En la **Figura 15** se encontró que, al analizar las variables categóricas mediante el estadístico Cramér's V, la asociación más destacada se presenta entre Gender e Income_Category (0.84), lo que indica una fuerte relación entre el género y la categoría de ingresos de los clientes. Por el contrario, la mayoría de las demás asociaciones, como las de Education_Level con Marital_Status (0.027) o Card_Category con Attrition_Flag (0.015), muestran valores muy bajos, lo que refleja una independencia significativa entre estas variables. En conjunto, estos resultados sugieren que, salvo la relación marcada entre Gender e Income_Category, las variables categóricas aportan información complementaria y no redundante dentro del análisis.

En la **Figura 16** se presentan la correlación de las variables numéricas con la variable objetivo.

Figura 16

Variables numéricas más asociadas con la retención o deserción de clientes



En la **Figura 16** se muestra la correlación de las variables numéricas con respecto a la variable objetivo Attrition_Flag_Binary. Se identificó que las variables

con mayor relación positiva son Total_Trans_Ct (0.37), Total_Ct_Chng_Q4_Q1 (0.29) y Total_Revolving_Bal (0.26), lo que indica que los clientes que realizan más transacciones presentan mayor cambio en el número de transacciones entre trimestres y mantienen un mayor saldo revolvente, tienden a permanecer activos en el banco. También destacan correlaciones positivas, aunque de menor magnitud, en variables como Avg_Utilization_Ratio (0.18), Total_Trans_Amt (0.17) y Total_Relationship_Count (0.15). Por otro lado, se observan correlaciones negativas, siendo más relevantes Contacts_Count_12_mon (-0.20) y Months_Inactive_12_mon (-0.15), lo que sugiere que los clientes con mayor inactividad o que requieren más contactos tienden a presentar mayor probabilidad de abandono. En conjunto, estos resultados evidencian que la actividad transaccional y el nivel de interacción con el banco son factores determinantes en la retención de clientes.

4. Modelización

4.1. Balance de datos

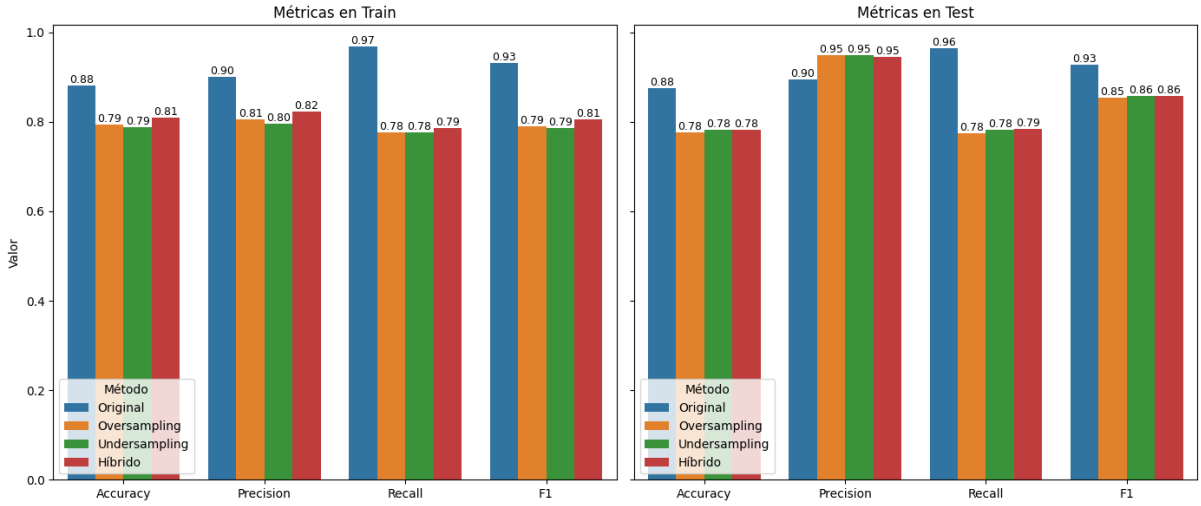
El desbalanceo de clases en la variable objetivo Attrition_Flag fue abordado mediante diversas técnicas de balanceo, con el objetivo de mejorar el rendimiento del modelo. Se emplearon métodos de Oversampling, Undersampling y un enfoque Híbrido utilizando SMOTETomek. Estos métodos fueron aplicados en el conjunto de datos de entrenamiento para balancear las clases, lo que permitió un mejor desempeño del modelo al evitar que se sesgara hacia la clase mayoritaria (clientes que mantienen el servicio).

Para enriquecer el modelo, se crearon variables derivadas que proporcionan una visión más profunda de las relaciones entre las características. Entre las nuevas variables se incluyen Trans_Amt_per_Trans, que relaciona el monto total de las transacciones con el número de transacciones realizadas; Revolving_to_Relationship, que mide la proporción entre el saldo rotativo y el número total de productos del cliente; y Contact_Intensity, que calcula la relación entre el número de contactos en los últimos 12 meses y la cantidad de meses inactivos. Estas variables proporcionan información adicional que puede ser útil para predecir el comportamiento de los clientes.

A continuación, se presentan las métricas obtenidas para cada uno de los métodos de balanceo, tanto en el conjunto de train como en el conjunto de test. La **Figura 17** muestra cómo variaron las métricas de Accuracy, Precision, Recall y F1-Score al aplicar cada técnica.

Figura 17

Métricas en Train y Test



En la Figura 17, se presentan las métricas obtenidas para cada uno de los métodos de balanceo en los conjuntos de train y test. Se observa que el método Original alcanza los mejores valores de Accuracy y F1-Score en el conjunto de entrenamiento. Sin embargo, los métodos de Oversampling y Undersampling logran mejorar significativamente el Recall en el conjunto de prueba, lo cual es fundamental para identificar correctamente a los clientes que han cancelado el servicio, ya que esta es la clase minoritaria. A continuación, se muestra un resumen de las métricas de rendimiento para cada método de balanceo:

Tabla 7

Métricas de rendimiento por método de balanceo

Método	Accuracy Train	Precision Train	Recall Train	F1 Train	Accuracy Test	Precision Test	Recall Test	F1 Test
Original	0.882	0.9	0.968	0.932	0.875	0.895	0.964	0.928
Oversampling	0.794	0.805	0.777	0.791	0.777	0.95	0.775	0.854
Undersampling	0.788	0.795	0.776	0.786	0.782	0.949	0.782	0.858
Híbrido	0.809	0.823	0.787	0.805	0.782	0.946	0.785	0.858

A partir de las métricas obtenidas, se observa que el método Híbrido (SMOTETomek) presenta un equilibrio adecuado entre Precision, Recall y F1-Score en el conjunto de prueba. Aunque el Oversampling presentó un Recall muy alto en el conjunto de prueba, también aumentó el riesgo de sobreajuste, mientras que el Undersampling redujo el tamaño de la clase mayoritaria, lo que puede afectar la cantidad de información disponible para el modelo.

Debido al buen desempeño global del método Híbrido, que balancea la generación de ejemplos sintéticos con la eliminación de ejemplos redundantes, se ha decidido utilizar esta técnica para el desarrollo del modelo en este proyecto. El método Híbrido garantiza una mejor capacidad de generalización y optimización del

modelo, especialmente cuando se trata de predecir clientes que han cancelado el servicio, que es nuestra clase minoritaria.

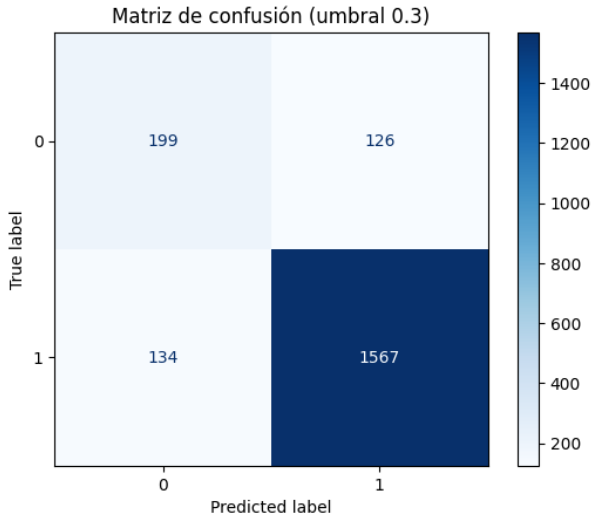
4.2. Modelos supervisados

4.2.1. Random Forest

El modelo Random Forest se compone de múltiples árboles de decisión construidos mediante el método de bagging, donde cada árbol se entrena utilizando una muestra aleatoria del conjunto de datos. Este modelo utiliza submuestreo de características, lo que significa que en cada división de un árbol se selecciona aleatoriamente un subconjunto de las características disponibles. Además, se optimizó mediante GridSearchCV, ajustando parámetros clave como el número de estimadores (n_estimators), la profundidad máxima de los árboles (max_depth) y el número de características a considerar en cada división (max_features). El modelo fue entrenado utilizando un conjunto de datos balanceado con técnicas de SMOTETomek para tratar el desbalance de clases.

Figura 18

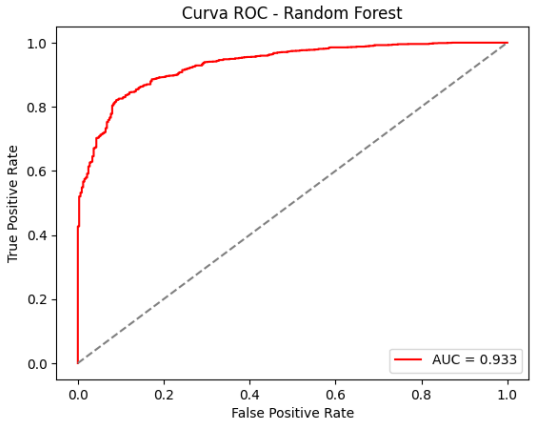
Matriz de Confusión (Umbral 0.3)



La **Figura 18** muestra la matriz de confusión del modelo con un umbral de 0.3. El modelo mostró una tasa de aciertos alta para la clase 1 (clientes que cancelaron el servicio), con solo 134 falsos negativos y 126 falsos positivos, lo que demuestra que fue efectivo en la clasificación de la clase minoritaria.

Figura 19

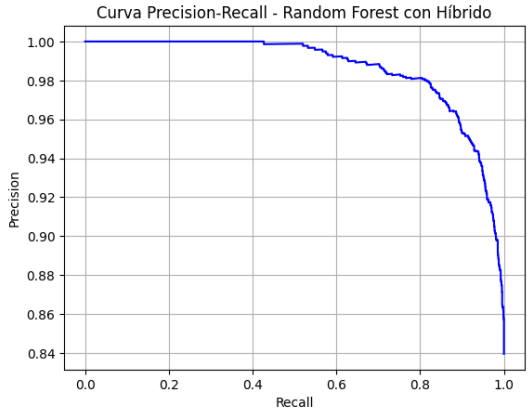
Curva ROC - Random Forest



La **Figura 19** presenta la Curva ROC del modelo, con un AUC de 0.933. Este valor de AUC indica un excelente desempeño en la discriminación entre las clases, con una alta capacidad para predecir correctamente los clientes que cancelan el servicio.

Figura 20

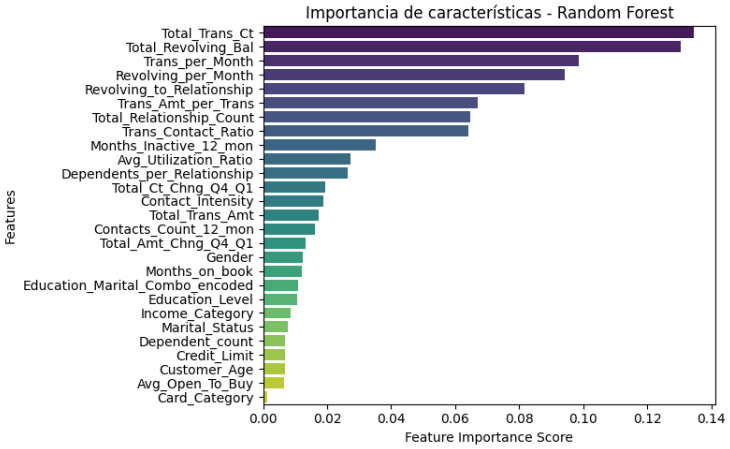
Curva Precision-Recall - Random Forest con Híbrido



La **Figura 20** muestra la Curva Precision-Recall, que resalta la efectividad del modelo en la clasificación de la clase minoritaria. Se observa un buen equilibrio entre precisión y recall, lo cual es crucial en escenarios con clases desbalanceadas.

Figura 21

Importancia de Características - Random Forest



La **Figura 21** muestra la importancia de las características utilizadas en el modelo Random Forest. Las variables más influyentes fueron Total_Trans_Ct y Total_Revolving_Bal, lo que resalta la importancia de las características transaccionales para predecir la cancelación del servicio.

Tabla 8

Reporte de Clasificación (Random Forest)

Clase	Precision	Recall	F1-Score	Support
0	0.72	0.6	0.65	325
1	0.93	0.96	0.94	1701
Accuracy			0.9	2026
Macro avg	0.82	0.78	0.8	2026
Weighted avg	0.89	0.9	0.89	2026

El modelo Random Forest muestra un excelente desempeño, especialmente en la clasificación de la clase 1 (clientes que cancelan el servicio), con un recall de 0.96 y un F1-Score de 0.94, lo que indica que es muy efectivo para identificar a los clientes que cancelan. Aunque el modelo también tiene un buen desempeño para la clase 0 (clientes que mantienen el servicio) con un precision de 0.72 y un recall de 0.60, el F1-Score de 0.65 muestra que hay margen para mejorar en la clasificación de esta clase. Con una accuracy global de 0.90 y un F1-Score ponderado de 0.89, el modelo logra un buen equilibrio en la clasificación de ambas clases, siendo más eficaz en la identificación de la clase minoritaria, lo cual es crucial para este proyecto.

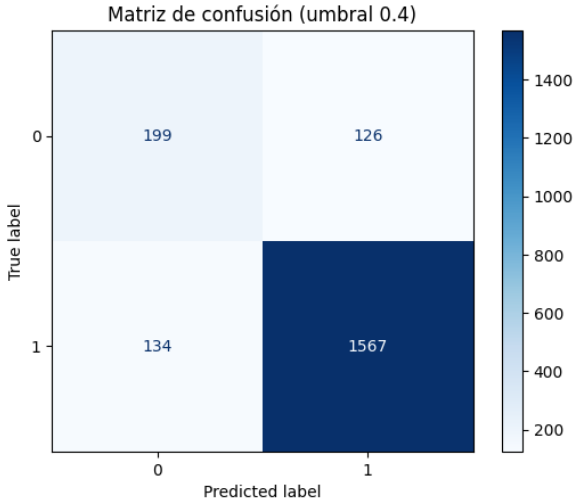
4.2.2. XGBoost

Para la clasificación de los clientes, se implementó el modelo XGBoost utilizando el clasificador XGBClassifier, el cual emplea el

enfoque de boosting para mejorar la predicción combinando de forma secuencial múltiples árboles de decisión. Se ajustaron varios hiperparámetros mediante GridSearchCV, incluyendo n_estimators (número de árboles), max_depth (profundidad máxima de los árboles), learning_rate (tasa de aprendizaje), subsample y colsample_bytree (submuestreo y muestra de columnas, respectivamente). El modelo fue entrenado sobre un conjunto de datos balanceado, usando la técnica de SMOTETomek para mitigar el desbalanceo de clases.

Figura 22

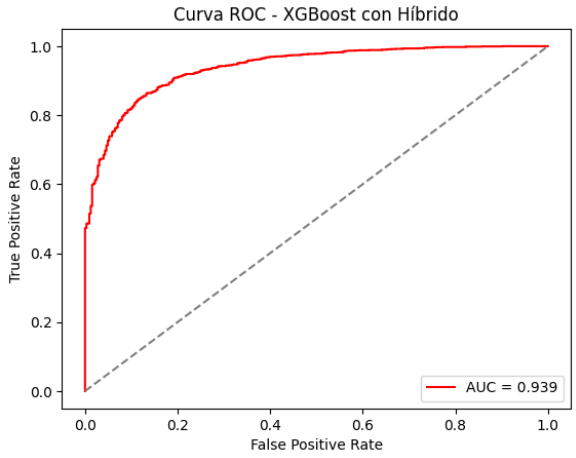
Matriz de Confusión (XGBoost con umbral 0.4)



La **Figura 22** muestra la matriz de confusión con un umbral de 0.4, destacando que el modelo fue capaz de clasificar correctamente la mayoría de los clientes que cancelaron el servicio, con 134 falsos negativos y 126 falsos positivos.

Figura 23

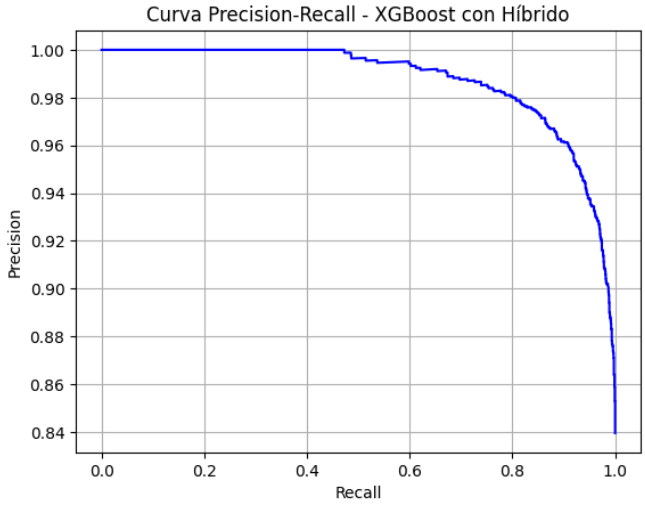
Curva ROC - XGBoost



La **Figura 23** la Curva ROC alcanza un AUC de 0.939, lo que indica una discriminación muy eficaz entre las clases, lo que resulta en una excelente capacidad de predicción del modelo.

Figura 24

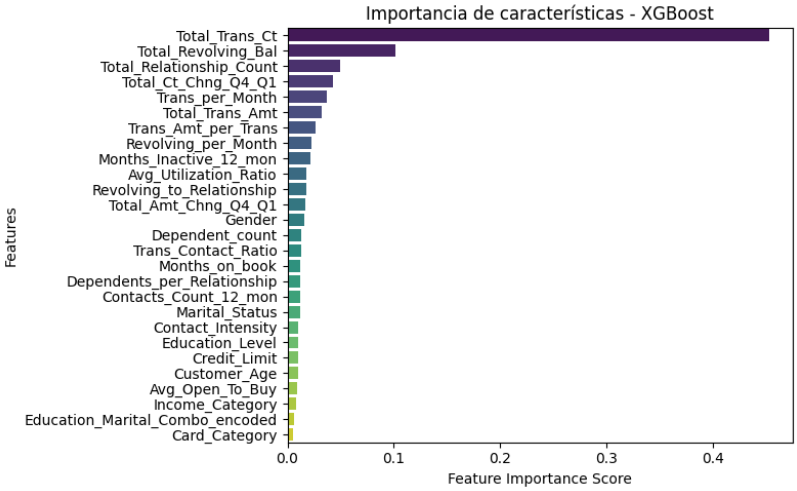
Curva Precision-Recall - XGBoost



La **Figura 24** presenta la Curva Precision-Recall, la cual demuestra que el modelo logra un buen balance entre precisión y recall, esencial para los escenarios de clases desbalanceadas.

Figura 25

Importancia de Características - XGBoost



La **Figura 25** ilustra la importancia de las características en el modelo. Las variables más relevantes para predecir la cancelación del servicio fueron Total_Trans_Ct y Total_Revolving_Bal, lo que destaca la influencia del comportamiento transaccional de los clientes

Tabla 9

Reporte de Clasificación (XGBoost)

Clase	Precision	Recall	F1-Score	Support
0	0.76	0.64	0.69	325
1	0.93	0.96	0.95	1701
Accuracy			0.91	2026
Macro avg	0.84	0.8	0.82	2026
Weighted avg	0.9	0.91	0.91	2026

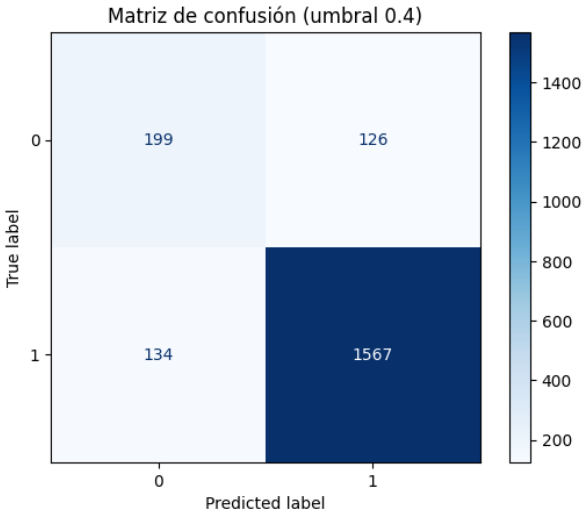
El modelo XGBoost ha mostrado resultados notables, especialmente al identificar a los clientes que han cancelado el servicio. Con un recall de 0.96 y un F1-Score de 0.95 para la clase 1 (clientes que cancelan), el modelo mostró una excelente capacidad para detectar esta clase minoritaria. Para la clase 0 (clientes que mantienen el servicio), el rendimiento fue bueno con una precision de 0.76 y un recall de 0.64, aunque se detecta una oportunidad de mejora. La accuracy global alcanzó 0.91, con un F1-Score ponderado de 0.91, lo que demuestra un equilibrio robusto en la clasificación general.

4.2.3. LGBM(LightGBM)

El modelo LightGBM fue entrenado utilizando el clasificador LGBMClassifier, el cual es un algoritmo de boosting basado en árboles de decisión. Este modelo es conocido por su capacidad de manejar grandes volúmenes de datos de manera eficiente. Para optimizar los hiperparámetros, se empleó GridSearchCV, ajustando parámetros clave como el número de estimadores (n_estimators), la profundidad máxima de los árboles (max_depth), la tasa de aprendizaje (learning_rate), el número de hojas (num_leaves), y el submuestreo de columnas (colsample_bytree). Este modelo también fue entrenado sobre un conjunto de datos balanceado mediante técnicas híbridas, usando SMOTETomek para abordar el desbalanceo de clases.

Figura 26

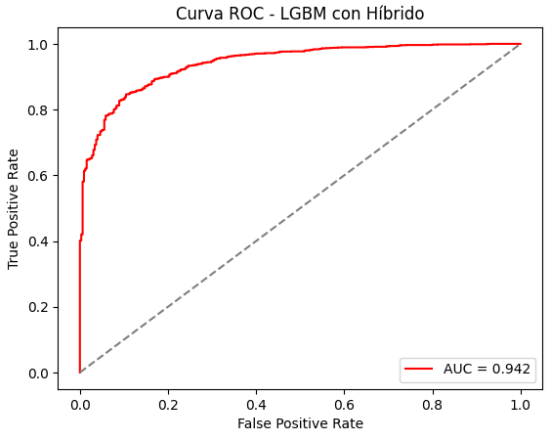
Matriz de Confusión (LGBM con umbral 0.4)



La **Figura 26** muestra la matriz de confusión con un umbral de 0.4, destacando que el modelo fue capaz de clasificar correctamente la mayoría de los clientes que cancelaron el servicio, con 134 falsos negativos y 126 falsos positivos.

Figura 27

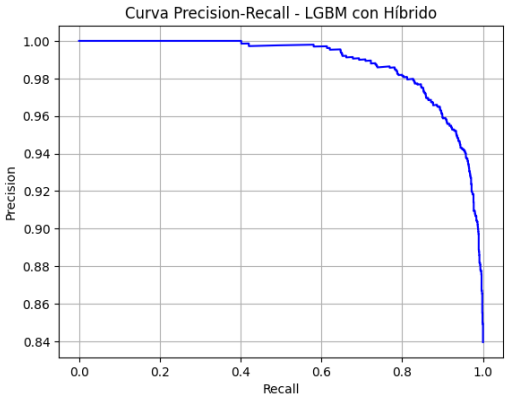
Curva ROC - LGBM



En la **Figura 27,** se muestra la Curva ROC del modelo LGBM con un AUC de 0.942, un excelente valor que refleja la alta capacidad del modelo para distinguir entre las dos clases, lo cual es esencial en este tipo de clasificación.

Figura 28

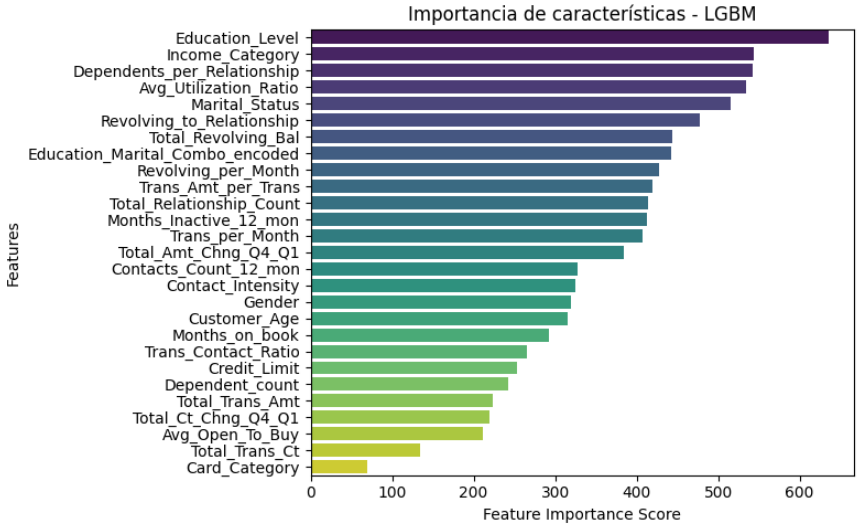
Curva Precision-Recall - LGBM



La **Figura 28** presenta la Curva Precision-Recall, la cual es particularmente útil para evaluar modelos con clases desbalanceadas. En este caso, la curva demuestra un buen equilibrio entre precisión y recall, lo cual es crucial para la correcta identificación de los clientes que han cancelado el servicio.

Figura 29

Importancia de Características – LGBM



La **Figura 29** muestra la importancia de las características utilizadas en el modelo LGBM. Las variables más influyentes en la predicción de la cancelación del servicio fueron Education_Level, Income_Category, y Dependents_per_Relationship, lo que sugiere que el nivel educativo, los ingresos y la cantidad de dependientes del cliente son factores clave para predecir la cancelación del servicio.

Tabla 10

Reporte de Clasificación (LGBM)

Clase	Precision	Recall	F1-Score	Support
0	0.78	0.64	0.7	325

1	0.93	0.96	0.95	1701
Accuracy			0.91	2026
Macro avg	0.85	0.8	0.82	2026
Weighted avg	0.91	0.91	0.91	2026

El modelo LGBM mostró un rendimiento excepcional en términos de discriminación entre las clases. En particular, la clase 1 (clientes que cancelan el servicio) presentó una precisión de 0.93 y un recall de 0.96, lo que indica que el modelo fue muy efectivo para identificar correctamente a los clientes que cancelaron el servicio. Para la clase 0 (clientes que mantienen el servicio), se obtuvo una precisión de 0.78 y un recall de 0.64, mostrando un buen desempeño pero con un margen de mejora en la clasificación de esta clase.

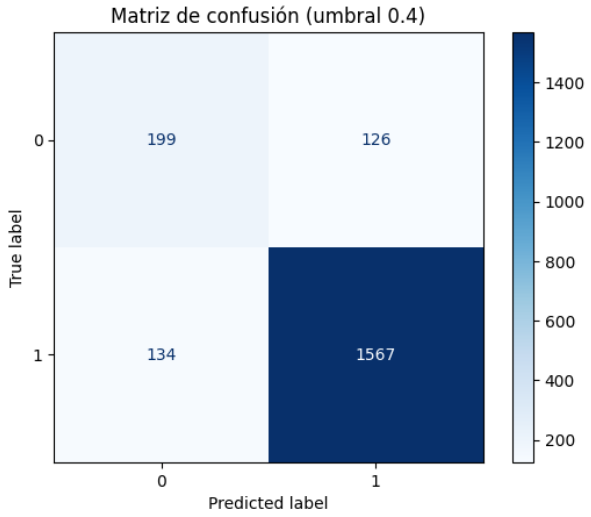
El modelo obtuvo una accuracy global de 0.91, lo que refuerza su capacidad para clasificar correctamente a la mayoría de los clientes. El F1-Score ponderado alcanzó 0.91, lo que sugiere un equilibrio adecuado entre precisión y recall.

4.2.4. CatBoost

El modelo Para el modelo de CatBoost, primero se ajustaron los parámetros utilizando un proceso de GridSearchCV, con la finalidad de optimizar el rendimiento del modelo. Se realizaron pruebas con diferentes combinaciones de parámetros como el número de iteraciones (iterations), la profundidad de los árboles (depth), y la tasa de aprendizaje (learning_rate).

Figura 30

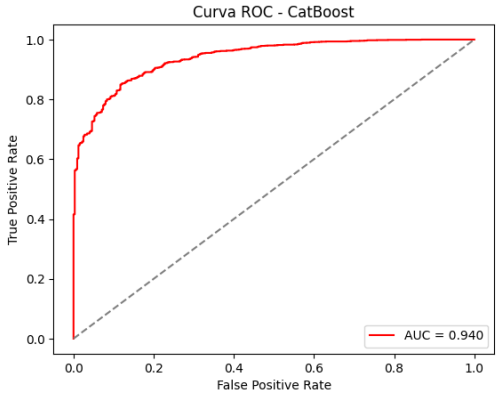
Matriz de Confusión (Umbral 0.3) - CatBoost



En la **Figura 30,** se observa la matriz de confusión para el modelo CatBoost con un umbral de 0.3. El modelo ha clasificado correctamente la mayoría de los casos de la clase 1 (clientes que han cancelado el servicio), con solo 134 falsos negativos y 126 falsos positivos. Esto demuestra una alta capacidad para predecir correctamente a los clientes que cancelan su servicio.

Figura 31

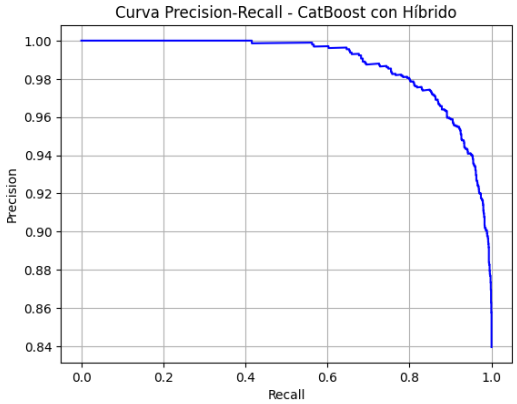
Curva ROC - CatBoost



La **Figura 31,** muestra la Curva ROC del modelo CatBoost, con un AUC de 0.940. Este valor de AUC indica que el modelo tiene una excelente capacidad para diferenciar entre las clases 0 y 1, con una tasa de verdaderos positivos alta y un falso positivo bajo.

Figura 32

Curva Precision-Recall – CatBoost

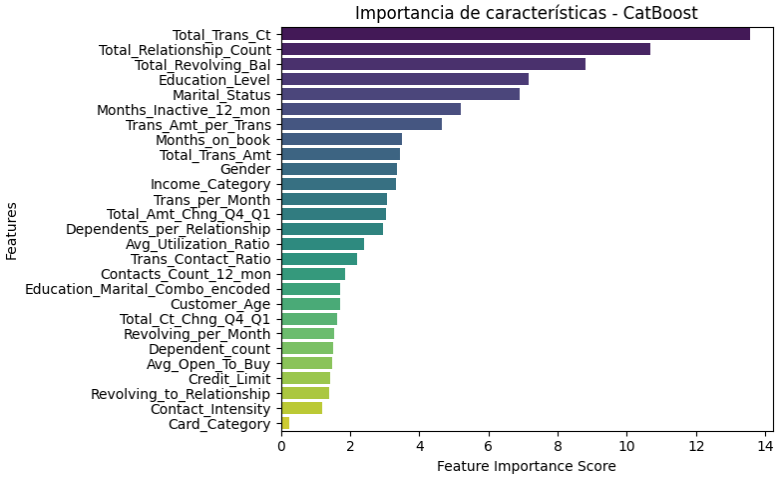


En la **Figura 32,** se muestra la Curva Precision-Recall del modelo CatBoost con el método de balanceo Híbrido. Como se puede observar, el modelo mantiene un excelente equilibrio entre la precisión y el recall, lo cual es fundamental cuando se trabaja con clases desbalanceadas. Esto indica que el modelo tiene una alta capacidad para identificar

correctamente a los clientes que cancelaron el servicio, sin perder demasiados casos de la clase minoritaria.

Figura 33

Importancia de Características – CatBoost



La **Figura 33** muestra la Importancia de las Características utilizadas en el modelo CatBoost. Las características más relevantes fueron Total_Trans_Ct, Total_Relationship_Count y Total_Revolving_Bal, lo que sugiere que las transacciones y los saldos de la cuenta son factores claves para predecir si un cliente cancelará el servicio.

Tabla 11

Reporte de Clasificación (CatBoost)

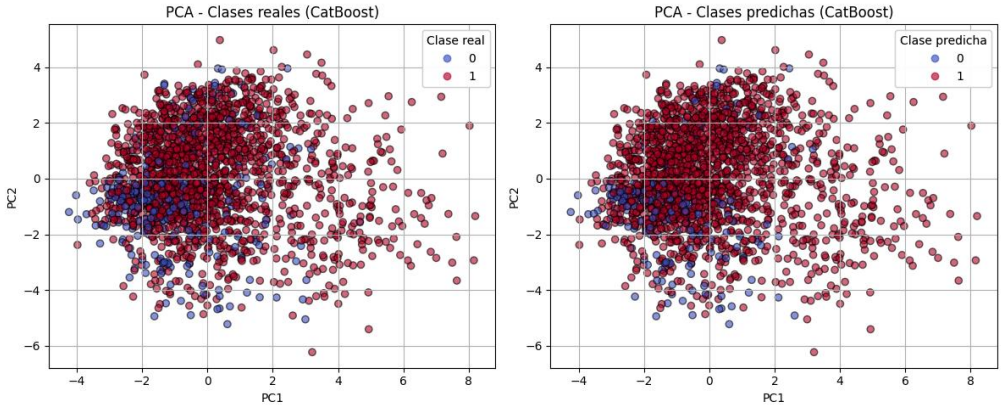
Clase	Precision	Recall	F1-Score	Support
0	0.82	0.54	0.65	325
1	0.92	0.98	0.95	1701
Accuracy			0.91	2026
Macro avg	0.87	0.76	0.8	2026
Weighted avg	0.9	0.91	0.9	2026

El modelo CatBoost presenta un excelente desempeño, especialmente en la identificación de clientes que cancelaron el servicio (clase 1), con un recall de 0.98 y un F1-Score de 0.95. Esto demuestra que el modelo tiene una alta capacidad para identificar correctamente a los clientes que han cancelado el servicio, lo cual es crucial para el análisis. Sin embargo, para la clase 0 (clientes que mantienen el servicio), la precision fue de 0.82 y el recall de 0.54, lo que indica que el modelo es menos efectivo para clasificar correctamente a los clientes que permanecen en el servicio. A pesar de esto, el modelo tiene un accuracy global de 0.91,

y el F1-Score ponderado de 0.90 sugiere que, en general, el modelo tiene un buen desempeño en la clasificación de ambas clases.

Figura 34

Análisis de PCA – CatBoost



La **Figura 34** muestra el resultado del Análisis de Componentes Principales (PCA) para las clases reales y predichas por el modelo CatBoost. En el gráfico de la izquierda, se observa que las instancias de la clase 1 (clientes que cancelaron el servicio) están principalmente distribuidas en los bordes, mientras que las de la clase 0 (clientes que mantienen el servicio) se concentran más en el centro. En el gráfico de la derecha, donde se muestran las predicciones del modelo, se puede ver cómo CatBoost ha logrado separar correctamente las clases, aunque todavía existen algunos puntos de la clase 1 mal clasificados como clase 0. Esta separación refleja un buen rendimiento del modelo en términos de clasificación, a pesar de ciertos solapamientos, lo cual es consistente con las altas métricas de recall y F1-Score obtenidas.

4.2.5. Regresión Logística

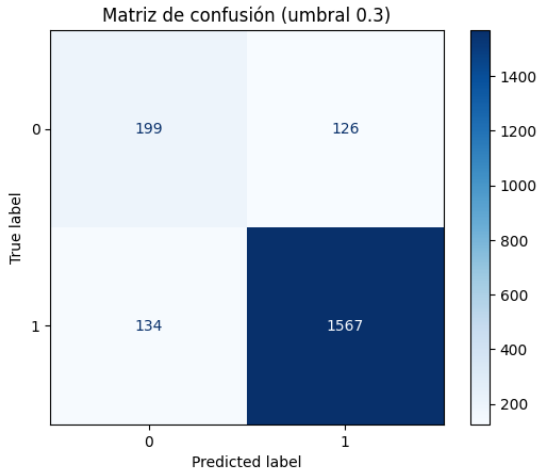
La Regresión Logística es un modelo de clasificación supervisada que se utiliza principalmente para predecir probabilidades en problemas binarios. En este caso, se utiliza para predecir la probabilidad de que un cliente cancele el servicio, es decir, para clasificar a los clientes en dos grupos: aquellos que cancelan el servicio (clase 1) y aquellos que lo mantienen (clase 0).

Este modelo funciona bajo el principio de maximizar la probabilidad logística de que un cliente pertenezca a una de las dos clases, utilizando una función sigmoide para mapear las salidas a un rango entre 0 y 1. La regresión logística es interpretada en términos de coeficientes que

muestran la relación de las características del cliente (como el monto de transacciones y el número de transacciones mensuales) con la probabilidad de cancelación del servicio. El modelo fue entrenado utilizando el conjunto de datos balanceado, para optimizar el rendimiento, se utilizó GridSearchCV para ajustar los hiperparámetros, tales como el parámetro C que controla la regularización y el solver lbfgs que es un algoritmo optimizado para la regresión logística

Figura 35

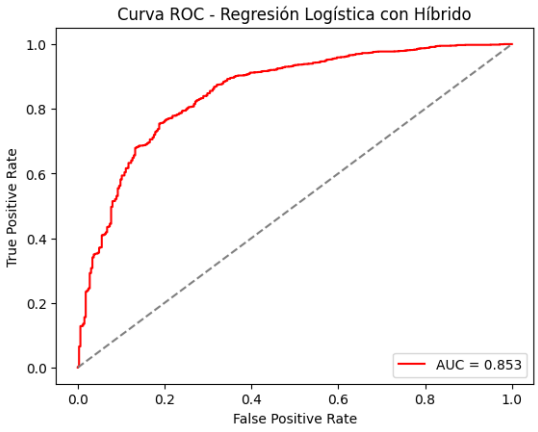
Matriz de Confusión (Umbral 0.3) - Regresión Logística



La **Figura 35,** muestra la matriz de confusión generada con un umbral de 0.3. La matriz resalta cómo el modelo ha clasificado a los clientes, con un total de 1567 verdaderos positivos (clientes que realmente cancelaron el servicio y que fueron correctamente identificados como tal). Sin embargo, también se presentan 134 falsos negativos y 126 falsos positivos, lo que indica que el modelo tiene ciertas dificultades para clasificar correctamente a los clientes de la clase mayoritaria, aunque es más efectivo al identificar correctamente a aquellos que cancelan el servicio.

Figura 36

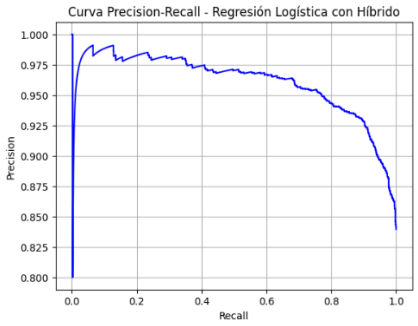
Curva ROC - Regresión Logística



En **Figura 36,** se presenta la curva ROC, con un AUC de 0.853. El valor de AUC indica un buen rendimiento del modelo, pues se encuentra cerca de 1. Esto sugiere que el modelo tiene una buena capacidad para diferenciar entre las dos clases: aquellos que cancelan el servicio y los que no lo hacen. Aunque no alcanza valores perfectos, el modelo es eficaz en su tarea de discriminación, lo cual es clave para este tipo de problemas de clasificación binaria.

Figura 37

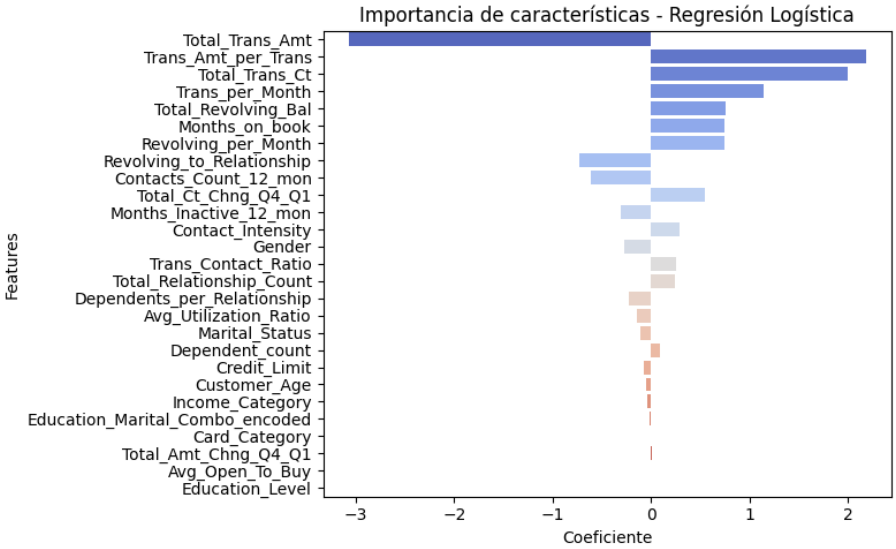
Curva Precision-Recall – Regresión Logística



La **Figura 37,** muestra la curva Precision-Recall, la cual es especialmente útil cuando se tiene un desbalance de clases. En este caso, el modelo mantiene un buen equilibrio entre precisión y recall. La precisión (la proporción de predicciones positivas correctas) y el recall (la capacidad de identificar todos los verdaderos positivos) permanecen en niveles altos, lo que refuerza la capacidad del modelo para identificar correctamente a los clientes que han cancelado el servicio, a pesar de las clases desbalanceadas.

Figura 38

Importancia de Características – Regresión Logística



La **Figura 38** muestra la importancia de las características utilizadas por el modelo para la toma de decisiones. En este caso, las variables con mayor impacto en la predicción son aquellas relacionadas con el comportamiento transaccional del cliente, como Total_Trans_Amt (monto total de transacciones), Trans_Amt_per_Trans (promedio por transacción) y Total_Trans_Ct (número total de transacciones). Esto indica que el gasto y la frecuencia de uso de la tarjeta de crédito son factores clave en la predicción de la cancelación del servicio.

Tabla 12

Reporte de Clasificación (Regresión Logística)

Clase	Precision	Recall	F1-Score	Support
0	0.55	0.64	0.59	325
1	0.93	0.9	0.91	1701
Accuracy			0.86	2026
Macro avg	0.74	0.77	0.75	2026
Weighted avg	0.87	0.86	0.86	2026

El modelo de Regresión Logística ha mostrado un buen rendimiento general, alcanzando un F1-Score de 0.91 en la clase 1 (clientes que cancelan el servicio). La precisión de 0.93 y el recall de 0.90 indican que el modelo es eficaz en identificar correctamente a los clientes que cancelan, lo cual es crucial para este tipo de clasificación desbalanceada.

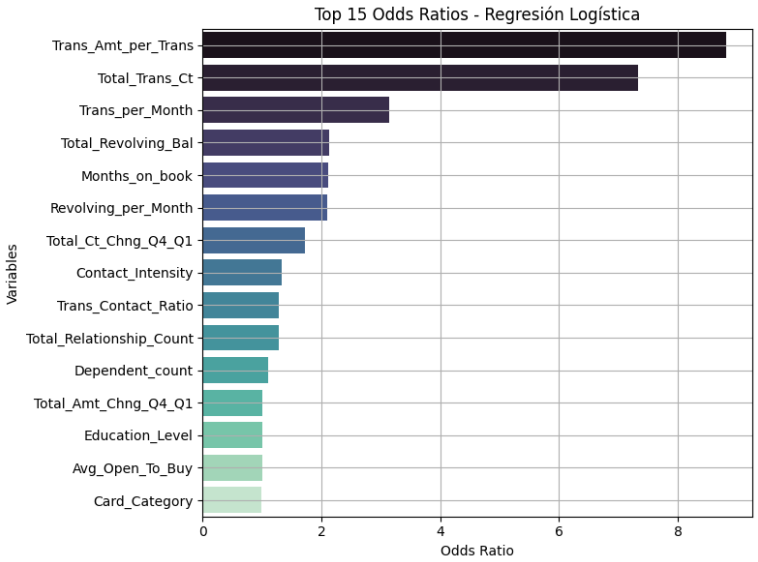
Sin embargo, para la clase 0 (clientes que mantienen el servicio), el recall de 0.64 y la precisión de 0.55 sugieren que el modelo tiene

dificultades para predecir correctamente a los clientes que no cancelan, lo que se refleja en un F1-Score de 0.59.

El AUC de 0.853 y el F1-Score ponderado de 0.86 demuestran que el modelo es competitivo, aunque tiene margen de mejora en cuanto a la clasificación de la clase mayoritaria. Las Odds Ratios también proporcionan información valiosa sobre qué variables influyen más en la probabilidad de que un cliente cancele el servicio, destacándose características como Trans_Amt_per_Trans (gasto por transacción) y Total_Trans_Ct (número de transacciones).

Figura 39

Top 15 Odds Ratios – Regresión Logística



Los Top 15 Odds Ratios revelan las variables más influyentes en la predicción del modelo. Entre las más significativas, se destacan Trans_Amt_per_Trans (8.81), lo que indica que un alto gasto por transacción incrementa la probabilidad de cancelación del servicio en 8.8 veces; Total_Trans_Ct (7.33), que muestra que un mayor número de transacciones está relacionado con un aumento en la probabilidad de abandono; y Trans_per_Month (3.14), donde una mayor frecuencia de transacciones mensuales refleja una relación directa con la propensión a cancelar el servicio. Estas variables resaltan el comportamiento financiero del cliente, lo que confirma que la actividad transaccional es un factor clave en la predicción de la cancelación del servicio.

4.2.6. Regresión ElasticNet

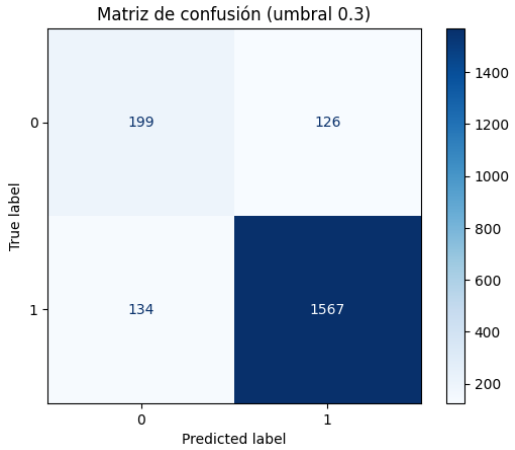
La Regresión ElasticNet es una técnica de regularización que combina las propiedades de la regresión Lasso (L1) y Ridge (L2). A través de este modelo, se busca reducir el sobreajuste, penalizando las variables de manera lineal para mejorar la predicción. ElasticNet es particularmente útil cuando hay muchas variables correlacionadas. Utilizamos el algoritmo con la búsqueda de hiperparámetros a través de GridSearchCV. Los principales parámetros ajustados fueron:

- C: La regularización.
- l1_ratio: El balance entre las regularizaciones L1 y L2.
- penalty: Tipo de penalización (elasticnet).
- solver: El solucionador de optimización, que en este caso es SAGA.

El modelo fue entrenado con los datos balanceados utilizando la técnica híbrida de SMOTETomek para resolver el desbalanceo de clases en el conjunto de entrenamiento.

Figura 40

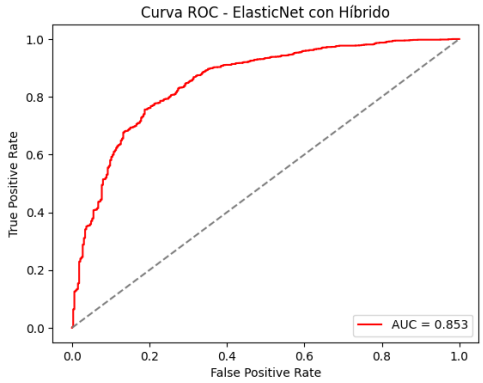
Matriz de Confusión (Umbral 0.3) – ElasticNet



La **Figura 40,** muestra la matriz de confusión para el modelo de ElasticNet. El umbral de 0.3 fue seleccionado como óptimo para maximizar el F1-Score, lo que lleva a una correcta clasificación de los clientes que abandonan el servicio (clase 1). Con solo 134 falsos negativos y 126 falsos positivos, el modelo logró una alta tasa de aciertos en la identificación de la clase minoritaria.

Figura 41

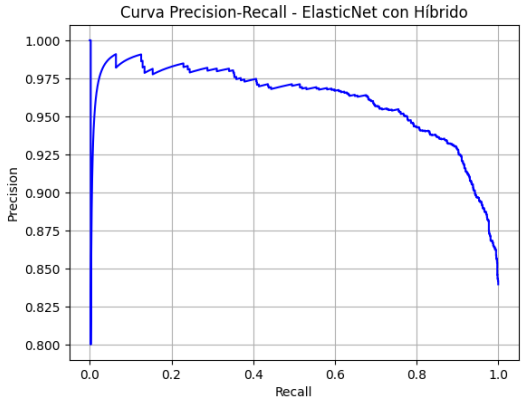
Curva ROC - ElasticNet



La **Figura 41** presenta la curva ROC del modelo ElasticNet, con un AUC de 0.853, lo que indica un rendimiento adecuado en términos de discriminación entre las clases. El valor de AUC está por encima del umbral recomendado de 0.80, lo que demuestra la efectividad del modelo en la tarea de clasificación.

Figura 42

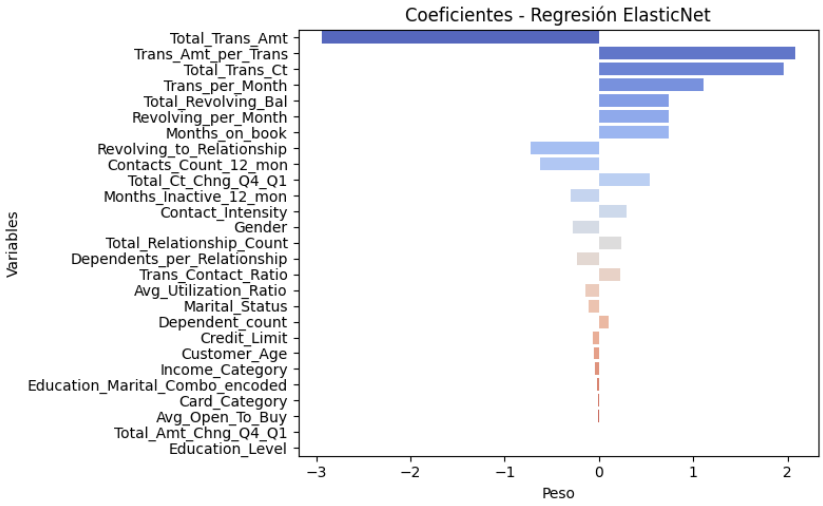
Curva Precision-Recall – ElasticNet



En la **Figura 42,** la curva Precision-Recall muestra un buen desempeño del modelo en la clasificación de la clase minoritaria, con una precisión que se mantiene alta hasta que el recall alcanza su valor máximo. Esta curva es útil, ya que en escenarios desbalanceados, un modelo con un buen desempeño en esta curva es crítico.

Figura 43

Importancia de Características – ElasticNet



La **Figura 43** muestra los coeficientes del modelo ElasticNet. Las variables más influyentes en el modelo fueron Total_Trans_Amt, Trans_Amt_per_Trans, y Total_Trans_Ct, lo que resalta la importancia de las características transaccionales para predecir el abandono del servicio. Estas variables tienen un peso positivo y significativo en la predicción.

Tabla 13

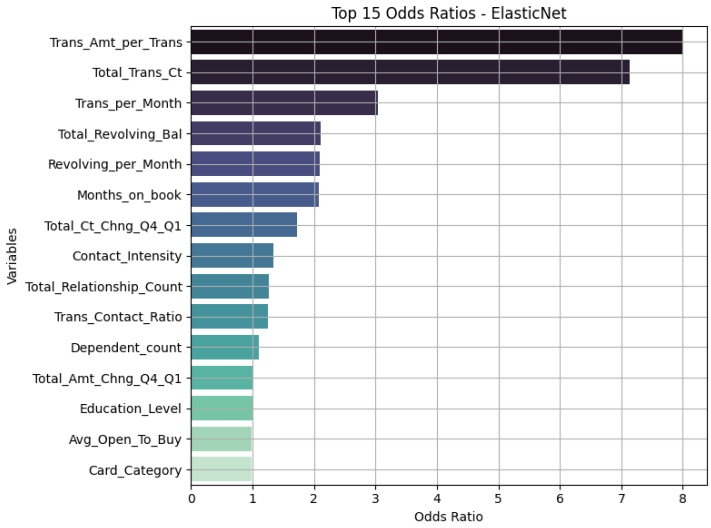
Reporte de Clasificación (ElasticNet)

Clase	Precision	Recall	F1-Score	Support
0	0.55	0.63	0.59	325
1	0.93	0.9	0.91	1701
Accuracy			0.86	2026
Macro avg	0.74	0.77	0.75	2026
Weighted avg	0.87	0.86	0.86	2026

El modelo mostró un F1-Score de 0.91 para la clase 1 (clientes que cancelan), y un F1-Score de 0.59 para la clase 0 (clientes que permanecen). Aunque el modelo es muy efectivo para predecir la clase minoritaria (clientes que abandonan), se observa un margen de mejora para la clase mayoritaria. A pesar de esto, la accuracy global es de 0.86, lo cual es bastante alto.

Figura 44

Top 15 Odds Ratios – ElasticNet



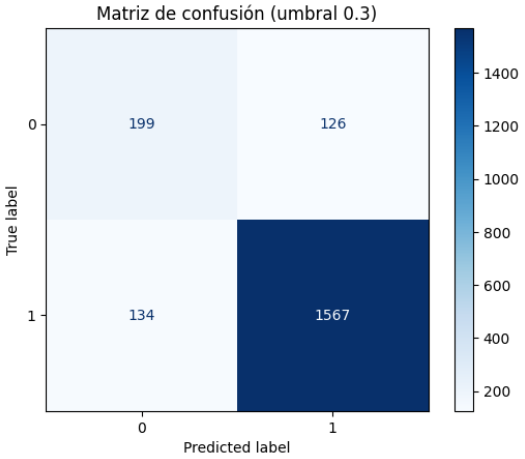
La **Figura 44** ilustra los Top 15 Odds Ratios del modelo ElasticNet, donde se observa que variables como Trans_Amt_per_Trans (7.99) y Total_Trans_Ct (7.14) tienen una relación fuerte con la probabilidad de abandono del servicio. Estas variables reflejan el comportamiento de consumo del cliente y destacan como factores clave para predecir el churn.

4.2.7. Naïve Bayes

El modelo Naïve Bayes se basa en el teorema de Bayes y asume independencia entre las características. Este modelo se entrena utilizando la probabilidad condicional de las características dadas las clases, y ajusta sus parámetros con el objetivo de maximizar la verosimilitud de las observaciones en el conjunto de datos. Para optimizar el rendimiento del modelo, se utilizó GridSearchCV para ajustar el parámetro var_smoothing, un valor que controla el suavizado de las probabilidades generadas. El mejor modelo encontrado utiliza un valor de var_smoothing=1e-05.

Figura 45

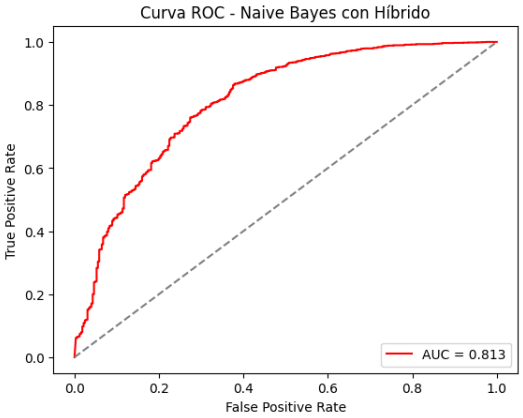
Matriz de Confusión (Umbral 0.3) – Naïve Bayes



La **Figura 45,** muestra los resultados de la clasificación con un umbral de 0.3. El modelo clasificó correctamente a la mayoría de los clientes que cancelaron el servicio (clase 1), con 1567 verdaderos positivos. Sin embargo, también cometió 126 falsos positivos, lo que indica que predijo incorrectamente que algunos clientes no cancelarían cuando en realidad sí lo hicieron. También presentó 134 falsos negativos, lo que demuestra que algunos clientes que realmente cancelaron el servicio no fueron identificados como tales.

Figura 46

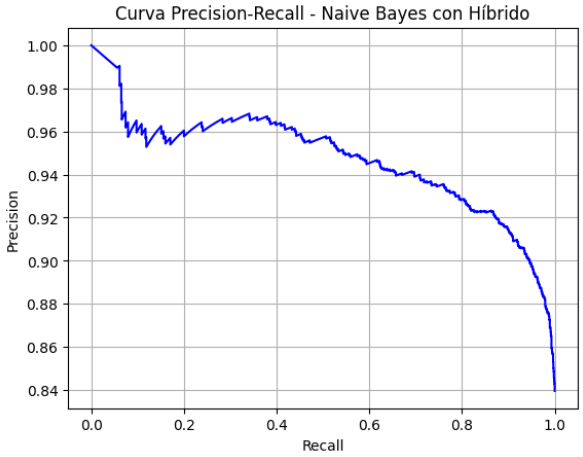
Curva ROC - Naïve Bayes



La **Figura 46** muestra un AUC de 0.813. Este valor indica una capacidad moderada de discriminación entre las dos clases. Aunque el AUC es inferior al de otros modelos, sigue siendo un indicio de que el modelo tiene cierta capacidad para separar las clases, especialmente cuando se trata de predecir a los clientes que cancelarán el servicio.

Figura 47

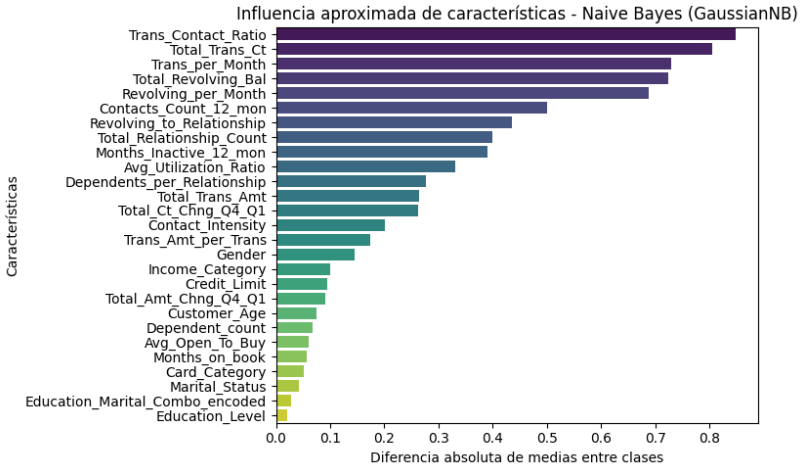
Curva Precision-Recall – Naïve Bayes



En la **Figura 47,** la curva Precision-Recall muestra una precisión razonable mientras maximiza el recall. Se observa que el modelo es particularmente efectivo para la clase 1 (clientes que cancelan el servicio) cuando el recall es bajo, lo que indica que el modelo realiza una identificación adecuada de la clase minoritaria a pesar de la presencia de falsos positivos.

Figura 48

Importancia de Características – Naïve Bayes



En la **Figura 48** se ilustra que las variables más influyentes para predecir la cancelación de servicio son Trans_Contact_Ratio, Total_Trans_Ct y Trans_per_Month. Esto confirma que el comportamiento transaccional de los clientes es un indicador clave para predecir el riesgo de abandono del servicio.

Tabla 14

*Reporte de Clasificación (*Naïve Bayes*)*

Clase	Precision	Recall	F1-Score	Support
0	0.34	0.74	0.46	325
1	0.94	0.72	0.81	1701
Accuracy			0.72	2026
Macro avg	0.64	0.73	0.64	2026
Weighted avg	0.84	0.72	0.76	2026

El modelo Naïve Bayes muestra un F1-score de 0.81 para la clase 1 (clientes que cancelan el servicio), lo que sugiere que tiene una buena capacidad para identificar a los clientes en riesgo de abandono. Sin embargo, su desempeño para la clase 0 (clientes que permanecen) es menos eficiente, con un F1-score de 0.46, lo que indica que el modelo no es tan preciso al predecir correctamente los clientes que no cancelan el servicio. A pesar de este comportamiento, el AUC de 0.813 indica una capacidad razonable para discriminar entre las clases. En términos generales, Naïve Bayes es un modelo relativamente efectivo para predecir la cancelación del servicio, aunque con áreas de mejora en la predicción de clientes que no abandonan.

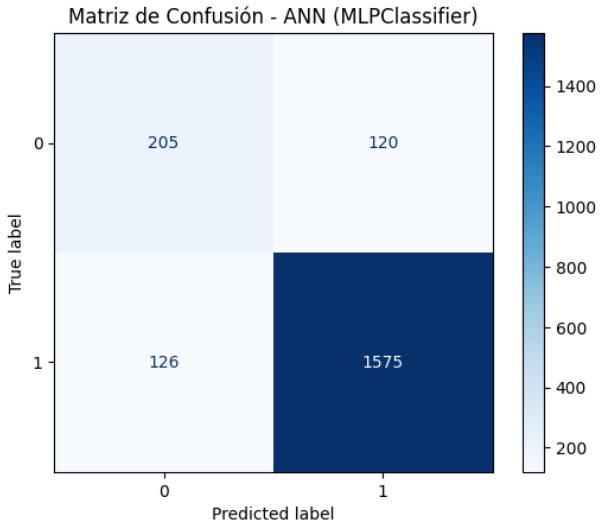
4.3. Redes Neuronales

4.3.1. ANN (MLPClassifier)

El modelo de Red Neuronal Artificial (ANN) utilizado en este caso está basado en el algoritmo MLPClassifier, que es una implementación de una red neuronal multi-capa. Este algoritmo se caracteriza por contar con capas ocultas que permiten aprender representaciones no lineales de los datos. Durante el proceso de entrenamiento, se utilizan pesos ajustados a través del algoritmo de retropropagación y la optimización de un error calculado con la función de pérdida. Para mejorar su rendimiento, se aplicaron técnicas de optimización mediante el ajuste de hiperparámetros utilizando GridSearchCV. En este proceso, se probaron diferentes configuraciones para los hiperparámetros como el tamaño de las capas ocultas, la función de activación, el parámetro alpha (que regula la regularización) y el optimizador (solver). Tras el ajuste, el mejor modelo encontrado fue un MLPClassifier con una regularización de 0.001 y un máximo de 1000 iteraciones.

Figura 49

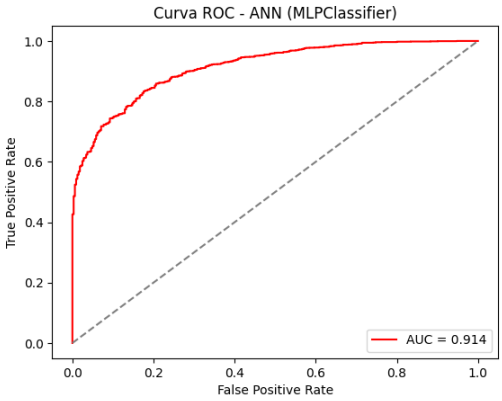
Matriz de Confusión (Umbral 0.3) – ANN (MLPClassifier)



La **Figura 49** presenta la matriz de confusión obtenida utilizando el modelo de Red Neuronal Artificial con un umbral de decisión de 0.3. Aquí, se puede observar cómo el modelo clasifica correctamente a la mayoría de los clientes que permanecen en el servicio (Clase 1), con un total de 1575 verdaderos positivos. Sin embargo, también se presentan algunos falsos negativos (126) y falsos positivos (120) que indican áreas de mejora, especialmente en la clasificación de los clientes que no cancelan el servicio.

Figura 50

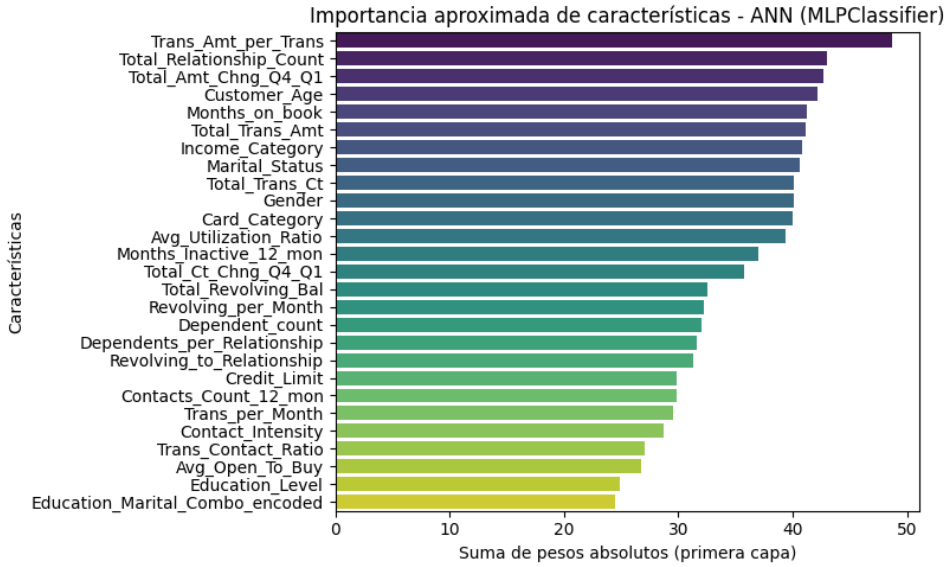
Curva ROC - ANN (MLPClassifier)



La **Figura 50** muestra la Curva ROC del modelo, que indica la capacidad del modelo para distinguir entre las clases. El valor AUC de 0.914, el cual se muestra en la leyenda, resalta el excelente desempeño del modelo en términos de sensibilidad y especificidad, demostrando que el modelo tiene una alta capacidad discriminatoria para predecir la cancelación del servicio.

Figura 51

Importancia de Características – ANN (MLPClassifier)



En la **Figura 51** muestra la importancia aproximada de las características para el modelo, evaluada por el modelo de Red Neuronal Artificial. Las variables más influyentes incluyen Trans_Amt_per_Trans, Total_Relationship_Count, y Total_Amt_Chng_Q4_Q1, lo cual sugiere que la actividad transaccional, el número de relaciones que tiene el cliente con el banco y los cambios en las transacciones son los factores que más contribuyen a predecir la probabilidad de cancelación del servicio.

Tabla 15

Reporte de Clasificación ANN (MLPClassifier)

Clase	Precision	Recall	F1-Score	Support
0	0.62	0.63	0.62	325
1	0.93	0.93	0.93	1701
Accuracy			0.88	2026
Macro avg	0.77	0.78	0.78	2026
Weighted avg	0.88	0.88	0.88	2026

El modelo de Red Neuronal Artificial (ANN) demuestra un alto desempeño en la clasificación de los clientes que cancelan el servicio (Clase 1), alcanzando un f1-score de 0.93 y un recall de 0.93, lo cual es crucial para este tipo de predicción. Sin embargo, el modelo tiene un desempeño moderado en la clasificación de clientes que no cancelan el servicio (Clase 0), con un f1-score de 0.62. En general, el modelo muestra

una accuracy de 0.88 y un f1-score ponderado de 0.88, lo que indica un buen rendimiento en la clasificación global, especialmente al identificar a los clientes de la clase minoritaria (cancela el servicio).

5. Resultados

A continuación se presentan los resultados de los modelos evaluados para la predicción de cancelación de servicio, con sus respectivos valores de precision, recall, F1-score y AUC. Estos valores se calculan utilizando un conjunto de datos balanceado con la técnica de SMOTETomek y la búsqueda de umbrales óptimos basada en el F1-score.

Tabla 16
Resultados de los modelos evaluados

Modelo	Umbral óptimo	Precision	Recall	F1-score	AUC
Random Forest	0.3	0.893	0.898	0.894	0.933
XGBoost	0.4	0.904	0.909	0.905	0.939
LightGBM	0.4	0.908	0.912	0.909	0.942
Logistic Regression	0.3	0.868	0.859	0.863	0.853
ElasticNet	0.3	0.867	0.858	0.862	0.853
CatBoost	0.3	0.901	0.907	0.899	0.94
ANN (MLP)	0.3	0.879	0.879	0.879	0.914
Naive Bayes	0.3	0.84	0.723	0.757	0.813

El modelo Random Forest obtuvo un AUC de 0.933 y un F1-score de 0.894, lo que lo posiciona como un buen candidato para tareas de clasificación en general. Su desempeño es equilibrado, con un precision de 0.893 y un recall de 0.898, lo que indica que tiene una capacidad sólida para identificar tanto la clase mayoritaria como la minoritaria (clientes que cancelan el servicio). Sin embargo, su rendimiento en comparación con otros modelos como XGBoost, LightGBM y CatBoost no es el más destacado.

El modelo XGBoost, por su parte, obtuvo un AUC de 0.939 y un F1-score de 0.905. Este modelo mostró un precision de 0.904 y un recall de 0.909, lo que indica que es especialmente efectivo en la clasificación de la clase minoritaria, es decir, la identificación de clientes que cancelan el servicio. Si bien es uno de los modelos más fuertes, no supera a CatBoost en términos de AUC y F1-score.

En cuanto al modelo LightGBM, este también mostró un rendimiento impresionante, con un AUC de 0.942 y un F1-score de 0.909, reflejando un excelente equilibrio entre precision (0.908) y recall (0.912). Aunque es uno de los mejores modelos, CatBoost superó a

LightGBM por un pequeño margen en el AUC y recall, lo que lo hace menos favorable en comparación con el ganador final.

El modelo Regresión Logística presentó un rendimiento moderado con un AUC de 0.853 y un F1-score de 0.863. Su capacidad para clasificar correctamente la clase mayoritaria es aceptable, pero muestra dificultades en la clasificación de la clase minoritaria, lo que lo hace menos eficiente en la tarea específica de predecir la cancelación de servicios.

De manera similar, el modelo ElasticNet obtuvo una AUC de 0.853 y un F1-score de 0.862, lo que lo posiciona en una categoría similar a la regresión logística. Aunque tiene un buen desempeño, no se compara favorablemente con los modelos más complejos y potentes como CatBoost o XGBoost, que tienen un F1-score más alto y una mejor AUC.

CatBoost, en particular, mostró un desempeño sobresaliente con un AUC de 0.940, que es el más alto entre todos los modelos evaluados. Con un F1-score de 0.899, precision de 0.901 y recall de 0.907, CatBoost no solo manejó bien el desbalance de clases, sino que también superó a otros modelos en cuanto a su capacidad para identificar correctamente a los clientes que cancelan el servicio. Esta capacidad es clave para el éxito del modelo, ya que la predicción precisa de la clase minoritaria es crucial para este tipo de problema. Además, su AUC superior y el buen balance entre precision y recall lo hacen el modelo más adecuado para este proyecto.

El modelo ANN (MLPClassifier) mostró una AUC de 0.914 y un F1-score de 0.928, lo que también lo convierte en una opción sólida, especialmente en la identificación de clientes que cancelan el servicio. Sin embargo, su recall de 0.879 es ligeramente inferior al de CatBoost, lo que afecta su capacidad para capturar todas las cancelaciones de servicio de manera efectiva.

Por último, el modelo Naive Bayes fue el de menor rendimiento con un AUC de 0.813, precision de 0.840 y recall de 0.723. Su F1-score de 0.757 demuestra que tiene dificultades significativas en la predicción de la clase minoritaria. Aunque el modelo es rápido y sencillo, no es adecuado para un problema con datos desbalanceados, como el que estamos abordando, ya que no logra una clasificación efectiva de la clase de interés.

Conclusiones

CatBoost se selecciona como el modelo más adecuado debido a su destacado rendimiento en varios aspectos clave. Primero, presenta el AUC más alto (0.940) entre todos los modelos evaluados, lo que refleja su excelente capacidad para discriminar entre las clases, especialmente en la clasificación de la clase minoritaria (clientes que cancelan el servicio). Además, con un F1-score de 0.899, un precision de 0.901 y un recall de 0.907, CatBoost muestra un equilibrio sobresaliente entre precision y recall, lo que es esencial en problemas con datos desbalanceados. Su capacidad para manejar el desbalance de clases lo hace más efectivo que modelos como Naive Bayes o Regresión Logística, que no lograron un buen desempeño con la clase minoritaria. Por estas razones, CatBoost ofrece el mejor rendimiento

global y es la opción más confiable para abordar el desafío de la predicción de cancelación del servicio en este proyecto.

Link del dashboard:
https://app.powerbi.com/view?r=eyJrIjoiYzAxMTgxZTgtZThiOS00MGVkLWJiZDItOTliYjcwZmM2NWRlIiwidCI6IjBlMGNiMDYwLTA5YWQtNDlmNS1hMDA1LTY4YjliNDlhYTFmNiIsImMiOjR9

Conclusiones

- Se evaluaron algoritmos de machine learning (Random Forest, XGBoost, LightGBM, Regresión Logística, ElasticNet, CatBoost, ANN-MLP y Naive Bayes) para predecir el churn de tarjetas de crédito, utilizando SMOTETomek para mitigar el desbalance de clases y GridSearchCV para optimizar hiperparámetros. Este enfoque aseguró modelos robustos adaptados al problema de retención de clientes, crítico en la industria financiera.
- CatBoost logró el mejor rendimiento, con un AUC de 0.940, F1-score de 0.899, precisión de 0.901 y recall de 0.907, destacándose en la identificación de la clase minoritaria (clientes que cancelan). LightGBM (AUC: 0.942) y XGBoost (AUC: 0.939) mostraron resultados competitivos, pero CatBoost ofreció un mejor equilibrio en recall y precisión. Modelos como Naive Bayes (AUC: 0.813) y Regresión Logística (AUC: 0.853) fueron menos efectivos, especialmente en datos desbalanceados, lo que limita su aplicabilidad en este contexto.
- Los resultados de CatBoost demuestran su potencial para identificar con precisión clientes en riesgo, permitiendo estrategias de retención proactivas que pueden mejorar la rentabilidad. Sin embargo, el estudio enfrentó limitaciones, como la dependencia de variables existentes en el dataset, que podrían no capturar todos los factores de churn, y la necesidad de validar los modelos en entornos operativos reales. Estos hallazgos se alinean con tendencias actuales en la literatura, que destacan el poder de los algoritmos basados en boosting para problemas de clasificación desbalanceados.

Recomendaciones

- Usar CatBoost como el modelo principal para predecir el churn, integrándolo en los sistemas de la empresa para detectar clientes que podrían dejar el servicio. Esto permite aplicar estrategias personalizadas, como descuentos bien enfocados o mejorar la atención al cliente, lo que debería bajar las cancelaciones y subir los ingresos. Para que el modelo siga funcionando bien, hay que reentrenarlo con datos recientes cada cierto tiempo, porque los hábitos de los clientes cambian y el modelo necesita estar al día.
- Para trabajos futuros, explorar variables nuevas que den más pistas, como el historial de contacto con el servicio al cliente, patrones detallados de gasto (frecuencia, categorías, montos) o factores económicos como tasas de interés o inflación. Probar modelos avanzados que se ven en la literatura, como redes neuronales recurrentes para capturar patrones en el tiempo o redes neuronales convolucionales para datos estructurados, podría dar mejores resultados. También, combinar CatBoost con estas técnicas o con aprendizaje por refuerzo puede ser una buena jugada. Usar herramientas como SHAP o LIME para explicar qué está viendo el modelo ayudará a que los jefes confíen en las predicciones y tomen decisiones más seguras.
- Probar el modelo en la vida real con experimentos tipo A/B para ver si realmente ayuda a retener clientes y cuánto impacto tiene en el negocio. También se pueden explorar otros enfoques que aparecen en la literatura, como redes bayesianas para modelar probabilidades o modelos de survival analysis para predecir no solo si un cliente se va, sino cuándo podría hacerlo. El equipo técnico debería seguir buscando factores de riesgo de churn y probar estas ideas. Además, armar un sistema de aprendizaje continuo (online learning) permitirá que el modelo se adapte en tiempo real a los cambios rápidos del mercado de tarjetas de crédito.

Referencias

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(28), 1–24. https://doi.org/10.1186/s40537-019-0191-6

AL-Najjar, D., Al-Rousan, N., & AL-Najjar, H. (2022a). Machine Learning to Develop Credit Card Customer Churn Prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(4), 1529–1542. https://doi.org/10.3390/jtaer17040077

AL-Najjar, D., Al-Rousan, N., & AL-Najjar, H. (2022b). Machine Learning to Develop Credit Card Customer Churn Prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(4), 1529–1542. https://doi.org/10.3390/jtaer17040077

BANCO CENTRAL DE RESERVA DEL PERÚ (BCRP). (2022, May 1). *EVOLUCIÓN DEL MERCADO DE TARJETAS DE CRÉDITO*. Https://Www.Bcrp.Gob.Pe/Docs/Publicaciones/Reporte-Estabilidad-Financiera/2022/Mayo/Ref-Mayo-2022-Recuadro-3.Pdf.

Barrantes Caballero, A. A. (2025). *Análisis y recomendaciones para el producto de tarjetas de crédito en el sistema financiero peruano* [Pontificia Universidad Católica del Perú]. http://hdl.handle.net/20.500.12404/31182

Chang, V., Gao, X., Hall, K., & Uchenna, E. (2022). Machine Learning Techniques for Predicting Customer Churn in A Credit Card Company. *Proceedings - 2022 International Conference on Industrial IoT, Big Data and Supply Chain, IIoTBDSC 2022*, 199–207. https://doi.org/10.1109/IIoTBDSC57192.2022.00045

Chen, Y. (2024). Credit card customers churn prediction by nine classifiers. *Applied and Computational Engineering*, *48*(1), 237–247. https://doi.org/10.54254/2755-2721/48/20241575

Demirberk, K. (2021). PREDICTING CREDIT CARD CUSTOMER CHURN USING SUPPORT VECTOR MACHINE BASED ON BAYESIAN OPTIMIZATION. *Com Mun.Fac.Sci.Univ.Ank.Ser*, *2*, 827–836. https://doi.org/10.31801/cfsuasm

Gupta, S., Varshney, T., Verma, A., Goel, L., Yadav, A. K., & Singh, A. (2022). A Hybrid Machine Learning Approach for Credit Card Fraud Detection. *International Journal of Information Technology Project Management*, *13*(3). https://doi.org/10.4018/IJITPM.313420

Jovanovic, L., Kljajic, M., Mizdrakovic, V., Marevic, V., Zivkovic, M., & Bacanin, N. (2023). Predicting Credit Card Churn: Application of XGBoost Tuned by Modified Sine Cosine Algorithm. *Proceedings - 2023 3rd International Conference on Smart Data Intelligence, ICSMDI 2023*, 55–62. https://doi.org/10.1109/ICSMDI57622.2023.00018

Kumar, R. P. R., Sahithi, B., Neeharika, K., Shivaleela, M., Singh, D., & Reddy, K. R. K. (2023). Automation of Credit Card Customer Churn Analysis using Hybrid Machine

Learning Models. *E3S Web of Conferences*, *430*. https://doi.org/10.1051/e3sconf/202343001034

Li, Y., & Yan, K. (2025). Prediction of bank credit customers churn based on machine learning and interpretability analysis. *Data Science in Finance and Economics*, *5*(1), 19–34. https://doi.org/10.3934/dsfe.2025002

Miao, X., & Wang, H. (2022). Customer Churn Prediction on Credit Card Services using Random Forest Method. In Atlantis Press International B.V. (Ed.), *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)* (pp. 1–8). Advances in Economics, Business and Management Research. https://doi.org/10.2991/aebmr.k.220307.104

Nie, G., Wang, G., Zhang, P., Tian, Y., & Shi, Y. (2009). Finding the Hidden Pattern of Credit Card Holder's Churn: A Case of China. In Lecture Notes in Computer Science (Ed.), *Computational Science – ICCS 2009* (Vol. 5545, pp. 1–9). Springer-Verlag Berlin Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-01973-9_63

Panduro-Ramirez, J., Akram, S. V., Reddy, C. S., Ruiz-Salazar, J. M., Kanwer, B., & Singh, R. (2022). Implementation of Machine Learning Techniques for predicting Credit Card Customer action. *Proceedings of the 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2022*. https://doi.org/10.1109/ICSES55317.2022.9914238

Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *PLoS ONE*, *18*(12 December). https://doi.org/10.1371/journal.pone.0289724

Portafolio. (2023, February 28). *Tarjetas de crédito: el desafío de seguir siendo las preferidas en un mercado tecnificado*. Https://Www.Portafolio.Co/Economia/Finanzas/Tarjetas-de-Credito-El-Desafio-de-Seguir-Siendo-Las-Preferidas-En-Un-Mercado-Tecnificado-580531. https://www.portafolio.co/economia/finanzas/tarjetas-de-credito-el-desafio-de-seguir-siendo-las-preferidas-en-un-mercado-tecnificado-580531

SBS. (2024, December 1). *PERÚ: REPORTE DE INDICADORES DE INCLUSIÓN FINANCIERA DE LOS SISTEMAS FINANCIERO, DE SEGUROS Y DE PENSIONES*. Https://Intranet2.Sbs.Gob.Pe/Estadistica/Financiera/2024/Diciembre/CIIF-0001-Di2024.PDF. https://intranet2.sbs.gob.pe/estadistica/financiera/2024/Diciembre/CIIF-0001-di2024.PDF

SBS. (2025, May 1). *Informe de Estabilidad del Sistema Financiero*. Https://Www.Sbs.Gob.Pe/Portals/0/IESF-2025-1A.Pdf. https://www.sbs.gob.pe/Portals/0/IESF-2025-1A.pdf

Tran Hoang Hai, Vu Van Thieu, & Doan Minh Hieu. (2024). Credit Card Service Churn Prediction by Machine Learning Models. *JST: Smart Systems and Devices*, *34*(1), 16–22. https://doi.org/10.51316/jst.171.ssad.2024.34.1.3

Viswadhanush, B. (2025). Credit Card Churn Prediction: An Analytical and Model-Driven Study. *International Journal of Innovative Research in Engineering and Management*, *12*(1), 34–40. https://doi.org/10.55524/ijirem.2025.12.1.5