Proyecto Final

Erica Marcela Martínez Silva [a,c]
Maria Fernanda Izquierdo Aparicio [a,c]
Raul Eduardo Vasquez Duarte [a,c]
Sebastian Amaya Porras [a,c],
Julian David Reyes Rueda[b,c]

[a]*Estudiante de Maestría de Analitica para la inteligencia de Negocios,*
[b]*Profesor, Departamento de Ingeniería Industrial*
[c]*Pontificia Universidad Javeriana, Bogotá, Colombia*

Abstract

En el mercado financiero colombiano, las entidades compiten por retener clientes y cumplir sus expectativas y necesidades. Recientemente, se observó que una entidad financiera presentó una disminución del 15% en clientes activos, dato que fue confirmado por Datacrédito. Para abordar esta situación, el equipo a cargo empleó cinco modelos de aprendizaje automático como KNN, SVM, Adaboost, RandomForest y redes neuronales. Estos modelos se validaron utilizando técnicas como validación cruzada y optimización de hiperparámetros, con el objetivo de mejorar las estrategias de retención identificando clientes propensos a cambiar de entidad y mitigar así esta disminución de clientes activos.

1. BUSINESS UNDERSTANDING

1.1. Determine Business objectives.

1.1.1. Background

Las entidades bancarias a nivel nacional se encuentran catalogadas dentro del sistema financiero colombiano como establecimientos de crédito (EC) y actualmente están vigiladas por la Superintendencia Financiera de Colombia (SFC). Estas entidades tienen como función principal canalizar recursos entre los agentes superavitarios de la economía hacia los deficitarios; lo cual se entiende como un proceso de oferta y demanda, en el cual los oferentes ponen a disposición su exceso de recursos y los demandantes solicitan el préstamo de estos recursos; está operación se lleva cabo a través de un proceso inicial de captación de fondos para posteriormente llevar a cabo la colocación en el mercado.

En Colombia las entidades que componen el sistema financiero colombiano para julio de 2023 presentaron activos por COP2.780 billones, de los cuales COP1.152 billones corresponden a los establecimientos de crédito representando el 75% del PIB del país" (Banrep, 2023). El listado de entidades vigiladas suministrado por la SFC deja ver que actualmente se encuentran registrados un total de 29 entidades bancarias (SFC, 2023), las cuales tienen a disposición una amplia gama de productos financieros, como cuentas de ahorro, corrientes y AFC, los CDT, la disposición de tarjetas de crédito y créditos. Este sector se encuentra liderado por Bancolombia con más de 30 millones de clientes, seguido por Banco de Bogotá con más de 24.6 millones de clientes y en tercer lugar Davivienda con 23.6 millones de clientes.

Uno de los productos de mayor demanda son las tarjetas de crédito, las cuales representan una proporción significativa del valor de transacción en los pagos con tarjeta en Colombia, sin embargo, el mercado es liderado por las tarjetas de débito (Passport, s.f.). Las tarjetas de crédito a nivel operacional son un medio

de pago que permite llevar a cabo compras fraccionadas a cuotas, las cuales se cancelaran periódicamente con un costo adicional definido por la tasa de interés.

De acuerdo con el reporte presentado por la SFC con corte a mayo de 2023, se evidencia un total de 13.762.751 tarjetas vigentes emitidas y se relaciona un total de 33.675.545 compras efectuadas a nivel nacional por un monto de COP$ 6.602.250 millones y a nivel internacional un total de 4.889.431 compras por COP$ 1.368.138 millones. Así mismo, cifras relacionadas en Passport indican que el monto transado mediante tarjetas de crédito para el año 2022 incremento un 6% a un total de COP 71,1 billones, el número de tarjetas en circulación creció un 3% para un total de 17 millones y la emisión de tarjetas de crédito en Colombia sigue estando dominada por Tuya SA, Bancolombia SA, Red Multibanca Colpatria y Banco Falabella SA, dominando MasterCard y Visa (Passport, s.f.).

La competencia en el sector se ve marcada por los beneficios que ofrece cada uno de los productos y como estos se contrastan y difieren de los de la competencia; por otra parte, la llegada de nuevos competidores el mercado como lo es la banca digital marco un nuevo un nuevo ángulo a nivel de competencia. Cada uno de estos aspectos, interfiere en la decisión del cliente al momento de optar por permanecer en el banco actual o trasladar sus productos a otro banco.

Por lo tanto, las instituciones financieras enfrentan un desafío crítico al presentar una marcada rotación en la base de clientes, marcado por la movilidad de clientes a otras instituciones debido a mejores productos o beneficios que suplen las necesidades actuales.

Para el entendimiento del negocio a nivel interno y externo se llevó a cabo 3 metodologías diferentes PESTEL, Fuerzas de Porter y SWOT, con el fin de identificar la posición de la compañía en el mercado.

FORTALEZAS	DEBILIDADES
- Amplia base de clientes consolidada. - Tecnología avanzada y capacidad digital. - Marca reconocida y confiable. - Red de sucursales y cajeros automáticos - Capacidades de análisis de datos avanzadas para personalizar ofertas y mejorar la gestión de relaciones con los clientes.	- Dependencia de la economía nacional. - Rigidez en estructuras organizativas. - Falta de diversificación de ingresos. - Dependencia de fuentes de financiamiento tradicionales - Limitaciones en la oferta de productos y servicios especializados - Exposición al riesgo crediticio debido a una cartera de préstamos concentrada en ciertos sectores o clientes.
OPORTUNIDADES	**AMENAZAS**
- Crecimiento de la banca digital. - Alianzas estratégicas con fintech u otras instituciones financieras. - Desarrollo de productos y servicios innovadores, como programas de educación financiera, para captar nuevos segmentos de mercado. - Participación en programas de financiamiento gubernamental o internacional para proyectos de infraestructura y desarrollo sostenible.	- Competencia intensa de bancos tradicionales y fintech. - Riesgos regulatorios en seguridad cibernética y protección al consumidor. - Impacto de crisis económicas y eventos financieros globales. - Vulnerabilidad a ciberataques y brechas de seguridad - Cambios repentinos en las políticas regulatorias - Impacto negativo en la reputación

Tabla 1. SWOT Compañía Bancaría

POLÍTICOS	ECONÓMICOS	SOCIAL
* **Regulaciones financieras:** Normativas sobre préstamos, tasas de interés, gestión de riesgos y protección al consumidor. * **Estabilidad política:** Cambios políticos y estabilidad gubernamental que afectan la percepción de riesgo. * **Estabilidad regulatoria:** La consistencia y previsibilidad en las regulaciones financieras. * **Políticas fiscales:** Cambios en impuestos y subsidios que impactan la rentabilidad y la demanda de servicios financieros.	* **Ciclos económicos:** Impacto de recesiones o crecimiento en la demanda de servicios bancarios. * **Tasas de interés:** Variaciones que afectan la rentabilidad de préstamos y depósitos. * **Inflación y tipo de cambio:** Fluctuaciones que afectan la capacidad de los clientes para ahorrar, invertir y pagar deudas. * **Desigualdad económica:** El nivel de desigualdad influye en la demanda de productos financieros y en las estrategias de inclusión financiera del banco.	* **Cambios demográficos:** Envejecimiento de la población, migración urbana y nuevas tendencias de consumo. * **Cambios en el comportamiento financiero:** Adopción de tecnología, preferencias de pago y conciencia ambiental. * **Educación financiera:** Niveles de conocimiento y educación financiera.
TECNOLÓGICO	**ENTORNO**	**LEGAL**
* **Innovaciones financieras:** Adopción de banca móvil, blockchain e inteligencia artificial. * **Seguridad cibernética:** Protección de datos y prevención de ciberataques. * **Automatización y eficiencia operativa:** Implementación de tecnologías para mejorar procesos internos y reducir costos operativos. * **Fintech y competencia digital: A**parición de empresas fintech y la competencia en servicios digitales.	* **Sostenibilidad:** Presiones para prácticas sostenibles en operaciones y financiamiento. * **Riesgos ambientales:** Evaluación de riesgos asociados al cambio climático y problemas ambientales. * **Responsabilidad social corporativa:** Prácticas sostenibles y compromiso con el medio ambiente que pueden influir en la percepción de la marca y la lealtad de los clientes.	* **Cumplimiento normativo:** Regulaciones que impactan gestión de riesgos y reporte financiero. * **Protección al consumidor:** Leyes contra prácticas abusivas y políticas de transparencia.

Tabla 2. PESTEL Compañía Bancaría

Rivalidad entre competidores existentes	En el sector bancario, la rivalidad entre competidores es alta debido a la cantidad de bancos tradicionales y fintech que compiten por los mismos clientes y productos. La competencia se intensifica por la oferta de tasas de interés, comisiones y servicios innovadores, lo que puede presionar los márgenes de ganancia y requerir constantes mejoras en la oferta de valor.
Poder de negociación de los proveedores	El poder de negociación de los proveedores (tecnología, seguridad, servicios complementarios, etc.) puede variar dependiendo de la disponibilidad de alternativas, la importancia de los productos o servicios suministrados y la capacidad de integración vertical del banco.
Poder de negociación de los clientes	Los clientes de un banco tienen un poder de negociación significativo, especialmente en entornos donde hay muchas opciones disponibles y la lealtad del cliente es baja. Los clientes pueden comparar fácilmente los servicios bancarios en términos de tasas, tarifas, conveniencia y calidad, lo que obliga al banco a ofrecer productos y servicios diferenciados y atractivos.
Amenaza de nuevos entrantes	La entrada al sector bancario puede ser desafiante debido a las barreras de entrada como requisitos regulatorios, altos costos de capital, necesidad de infraestructura tecnológica y la necesidad de construir una base de clientes sólida.
Amenaza de productos o servicios sustitutos	La amenaza de sustitución puede ser alta si estos servicios ofrecen ventajas competitivas significativas en términos de conveniencia, costos y eficiencia en comparación con los servicios bancarios tradicionales.

Tabla 3. Fuerzas de Porter Compañía Bancaría

1.1.2. Business goal:

- Reducir la tasa de abandono de clientes del banco y mejorar la retención de los clientes activos mediante la identificación de aquellos con mayor propensión a retirarse y la implementación de estrategias de retención personalizadas.

- Diseñar estrategias de mercado enfocadas en los clientes, a partir del entendimiento del comportamiento y de la identificación de aquellos clientes que tienen posibilidad de cambiar de entidad; a fin de mejorar la fidelidad, experiencia y preferencia de los clientes del banco.

1.1.3. Business success criteria:

- Incrementar la retención de clientes activos en al menos un 10% en el próximo trimestre.
- Mejorar la satisfacción del cliente, medida a través de KPIs como Net Promoter Score (NPS) y tasas de lealtad, en línea con la estrategia de experiencia del cliente y calidad de servicio.
- Incrementar el uso y la preferencia de los productos ofrecidos por la entidad bancaria.

1.2. Determine Data mining goals.

1.2.1. Data mining goal:

- Contrastar el rendimiento de diferentes modelos de predicción, determinando las variables relevantes para lograr una identificación precisa de los clientes propensos a abandonar el banco

1.2.2. Data mining success criteria:

- Comparar y seleccionar el modelo de predicción con el mejor rendimiento, respaldado por métricas como F1, Accuracy y Recall, para una identificación efectiva de clientes en riesgo de churn.

2. DATA UNDERTANDING

2.1. Describe data:

Este es el conjunto de datos del banco que detallan 10,127 clientes, el dataset incluye una serie de variables demográficas, económicas y de actividad financiera que abarcan desde el año 2021 hasta el cierre de 2022. Entre las variables claves se encuentran **clientnum**, que es un identificador único por cliente; **attrition_flag,** la variable objetivo que indica si el cliente ha dejado el banco; junto con otras variables importantes como la edad del cliente (**customer_age**), sexo (**gender**), número de dependientes (**dependent_count**), nivel educativo (**education_level**), estado civil **(marital_status**), categoría de ingresos (**income_category**), y varias métricas relacionadas con la actividad bancaria del cliente como la antigüedad **(months_on_book**), productos bancarios totales (**total_relationship_count**), actividad de transacciones (**total_trans_amt**), y utilización de crédito (**avg_utilization_ratio**).

Variable	Descripción	Tipo de Dato
clientnum	Identificador del cliente.	Entero
attrition_flag	Variable Objetivo. Indica si el cliente se fue a otro banco o no.	Booleano
customer_age	Edad del cliente.	Entero
gender	Sexo del cliente.	String
dependent_count	Número de personas económicamente dependientes del cliente.	Entero
education_level	Nivel de educación del cliente.	String
marital_status	Estado civil del cliente.	String
income_category	Categoría de ingresos del cliente.	String
card_category	Categoría de tarjeta del cliente.	String

months_on_book	Antigüedad del cliente.	Entero
total_relationship_count	Total, de productos que tiene el cliente con el banco.	Entero
months_inactive_12_mon	Meses en que el cliente estuvo inactivo en el último año.	Entero
contacts_count_12_mon	Número de contactos con el cliente en el último año.	Entero
credit_limit	Cupo de crédito del cliente.	Entero
total_revolving_bal	Balance de crédito rotativo del cliente.	Entero
avg_open_to_buy	Promedio de cupo disponible en tarjetas de crédito.	Entero
total_amt_chng_q4_q1	Cambio en el valor total de transacciones entre Q4 y Q1.	Entero
total_trans_amt	Valor total de transacciones.	Entero
total_trans_ct	Cantidad total de transacciones.	Entero
total_ct_chng_q4_q1	Cambio en la cantidad total de transacciones entre Q4 y Q1.	Entero
avg_utilization_ratio	Razón de utilización de la tarjeta.	Entero

Tabla 4. Descripción de Variables

2.2. Explore data

- **Clientnum:** representa el identificador único de cada cliente. Al ser un identificador único, no tendrá un valor predictivo para el modelo, ya que no encierra información generalizable sobre patrones de comportamiento de los clientes, por lo que validamos solo si los valores son únicos y confirmamos que cada registro representa un cliente diferente.

- **attrition_flag:** Es la variable objetivo en el análisis de churn, es decir, indica si un cliente ha dejado el banco o se ha quedado. Esta es una variable categórica binaria. Al graficarlo los resultados muestran claramente que hay un desbalance significativo en las clases de la variable objetivo, con aproximadamente el 84% de los clientes clasificados como "Existing Customer" y solo el 16% como "Attrited Customer" y hay que considerar estrategias para reequilibrar los datos.

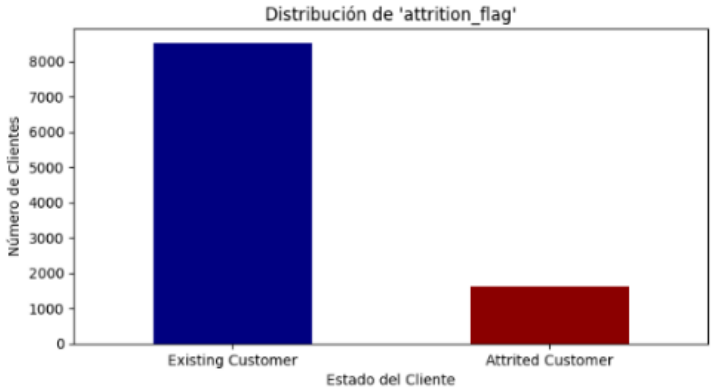

Figura 1. Distribución attrition_flag

- **customer_age**: La edad de los clientes del banco se distribuye con una media de aproximadamente 46 años y una desviación estándar de 8 años, con edades que oscilan entre 26 y 73 años. El histograma muestra una distribución aproximadamente normal, pero ligeramente sesgada hacia las edades más jóvenes, con una concentración más alta alrededor de los 40 a 50 años.

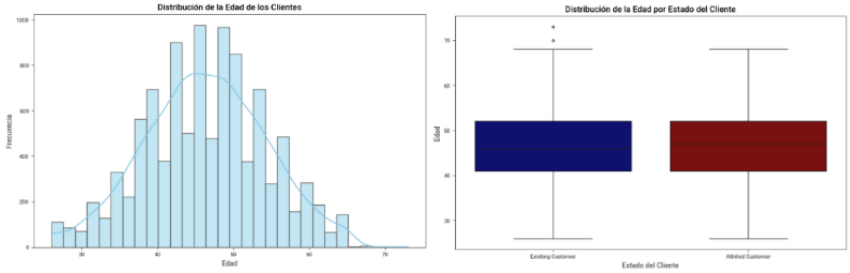

Figura 2. Distribución customer_age

- **Gender:** Eepresenta el sexo del cliente y es una variable categórica con categorías "Masculino" (M) y "Femenino" (F). El dataset muestra una distribución casi equitativa entre géneros, con un ligero predominio de mujeres (52.9%) sobre hombres (47.1%) y podemos ver que las mujeres tienen una tasa de churn ligeramente superior en comparación con los hombres en este conjunto de datos.

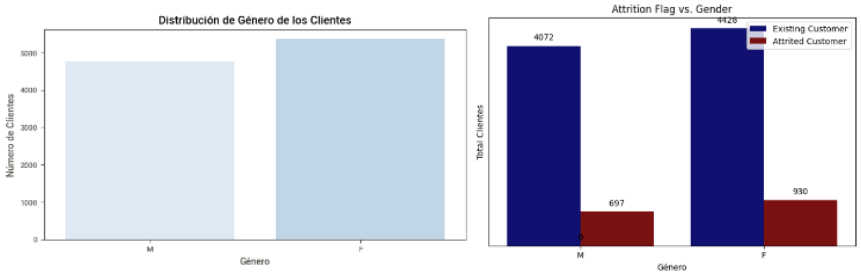

Figura 3. Distribución Gender

- **dependent_count**: Indica el número de personas económicamente dependientes del cliente, las gráficas muestran que la mayoría de los clientes tienen entre 1 y 4 dependientes, con una concentración especialmente alta en aquellos con 2 y 3 dependientes. Esto es indicativo de que muchos clientes del banco tienen responsabilidades familiares.

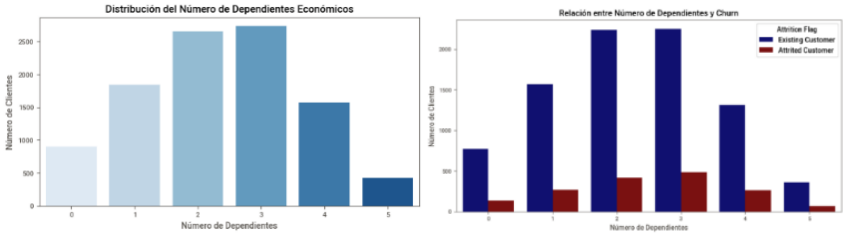

Figura 4. Distribución dependent_count

- **education_level:** Refleja el nivel de educación alcanzado por los clientes. El gráfico de distribución muestra que la mayoría de los clientes del banco son graduados universitarios, seguidos por aquellos con educación secundaria y luego por aquellos catalogados como 'Unknown'. Las categorías con menor representación son las de educación postgraduada y doctorados.

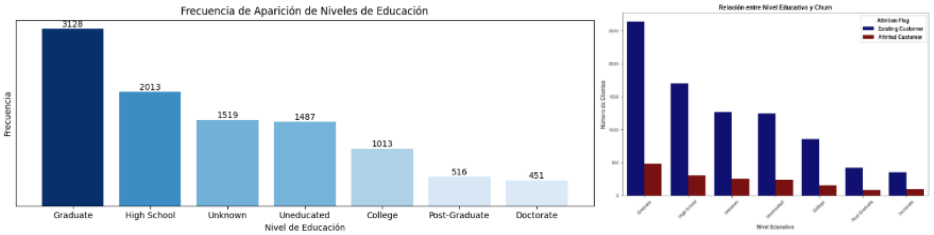

Figura 5. Distribución education_level

- **marital_status:** Refleja el estado civil de los clientes, los datos muestran que los clientes casados son el grupo más grande, representando aproximadamente el 46% de la base de clientes, seguidos por los solteros con casi el 39%. Los divorciados y los de estado civil desconocido constituyen cada uno alrededor del 7%. Esta distribución indica que la mayoría de los clientes del banco tienen compromisos familiares significativos, lo que potencialmente afecta sus necesidades financieras y decisiones.

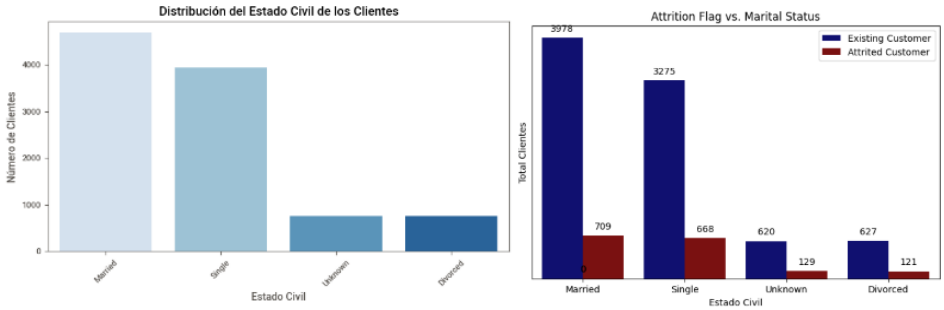

Figura 6. Distribución marital_status

- **income_category** refleja las categorías de ingresos de los clientes del banco, la categoría de ingresos "Less than $40K" es la más representativa entre los clientes del banco, constituyendo más del 35% del total, lo que sugiere que una gran parte de la clientela del banco puede ser más sensible a los productos y servicios financieros asequibles esta misma categoría tiene la mayor cantidad de churn absoluto (612 de 3561), lo cual podría indicar que los clientes en este segmento pueden estar enfrentando dificultades financieras o podrían estar menos satisfechos con los productos que no se ajustan bien a sus necesidades.

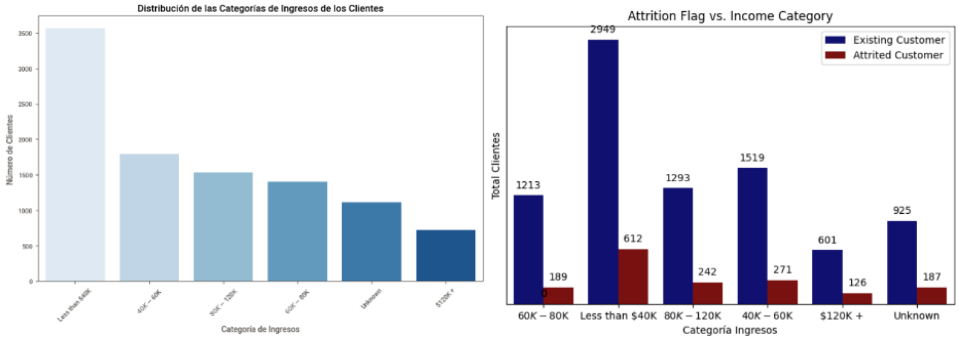

Figura 7. Distribución income_category

- **card_category**: representa las categorías de tarjetas de crédito ofrecidas por el banco a sus clientes, como tarjetas estándar, oro, platino, etc., la categoría "Blue" domina con más del 93% de los clientes. Las tarjetas "Silver" representan aproximadamente el 5.5%, mientras que las categorías "Gold" y "Platinum" son mucho menos comunes, con el 1.15% y 0.2% respectivamente. Esto sugiere que la mayoría de los clientes están utilizando el nivel de entrada o productos básicos de tarjetas de crédito, además vemos que la tarjeta "Blue", es la más común y tiene el número más alto de churn en términos absolutos (1519 de 9436), pero cuando se observa en términos relativos, su tasa de churn es comparativamente menor que algunas de las tarjetas más exclusivas.

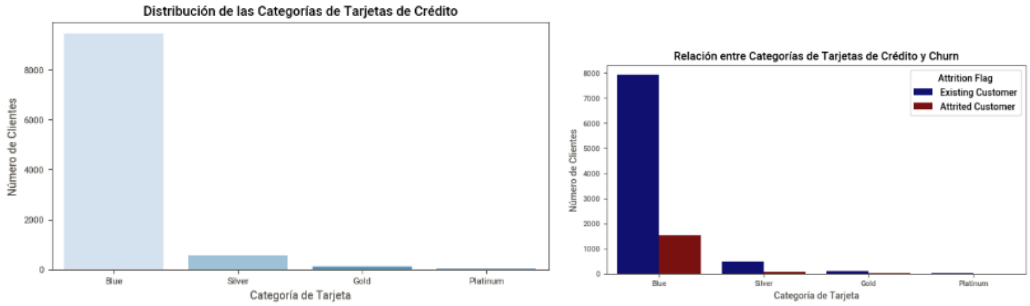

Figura 8. Distribución card_category

- **months_on_book**: Indica la antigüedad del cliente con el banco en meses, el histograma muestra una distribución que se concentra principalmente alrededor del rango medio de antigüedad, con un pico significativo alrededor de los 36 meses.

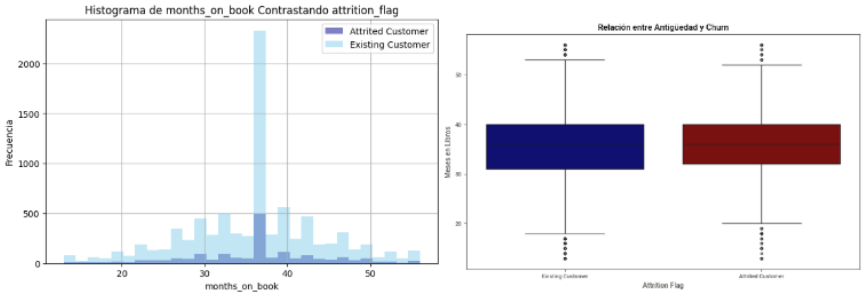

Figura 9. Comportamiento months_on_book

- **total_relationship_count**: refleja el número total de productos que el cliente tiene con el banco, incluyendo cuentas corrientes, de ahorros, préstamos, tarjetas de crédito, entre otros, los datos muestran que la mayoría de los clientes tienen entre tres y seis productos con el banco, además vemos que los clientes con solo un producto tienen una tasa de churn significativamente más alta (233 de 910), lo que indica que los clientes con menos productos son más propensos a dejar el banco.

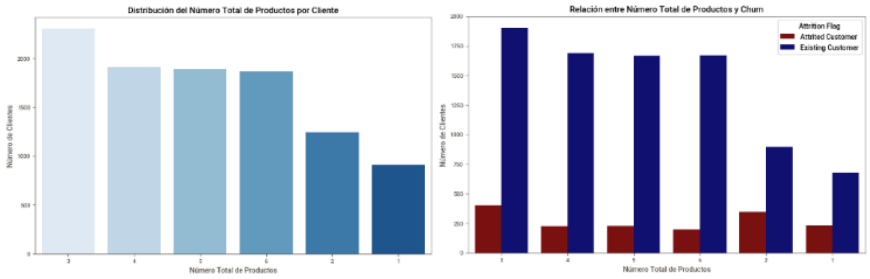

Figura 10. Distribución total_relationship_count

- **months_inactive_12_mon**: Refleja el número de meses en que un cliente ha estado inactivo durante el último año, la distribución de los meses de inactividad revela que la mayoría de los clientes tienen entre dos y tres meses de inactividad en el último año, con un pico prominente en tres meses. Esto sugiere que un período breve de inactividad es común y posiblemente normal dentro del ciclo de vida del cliente.

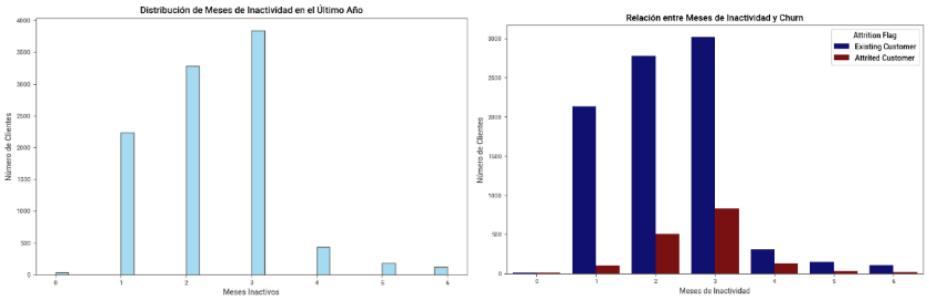

Figura 11. Distribución months_inactive_12_mon

- **contacts_count_12_mon:** Representa el número de veces que un cliente ha sido contactado por el banco en los últimos doce meses, la distribución del número de contactos muestra que la mayoría de los clientes han sido contactados 2 o 3 veces en el último año, los contactos excesivos (4 veces o más), hace que la proporción de clientes que abandonan el banco aumente. Esto podría sugerir que un exceso de contactos podría ser percibido como intrusivo o molesto, llevando a una experiencia negativa del cliente.

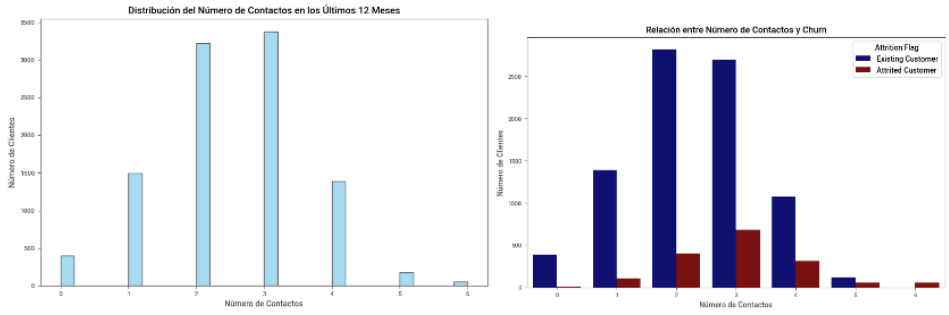

Figura 12. Distribución contacts_count_12_mon

- **credit_limit**: representa el límite de crédito total asignado a un cliente los datos muestran una concentración notable de clientes con límites en el rango inferior, y una disminución progresiva a medida que el límite de crédito aumenta. La mayoría de los clientes tienen límites de crédito por debajo de 15,000 unidades, con picos significativos alrededor de los valores mínimos. También se observa un pequeño pico hacia el extremo superior del rango de límites de crédito, lo que indica un número limitado de clientes con límites muy altos.

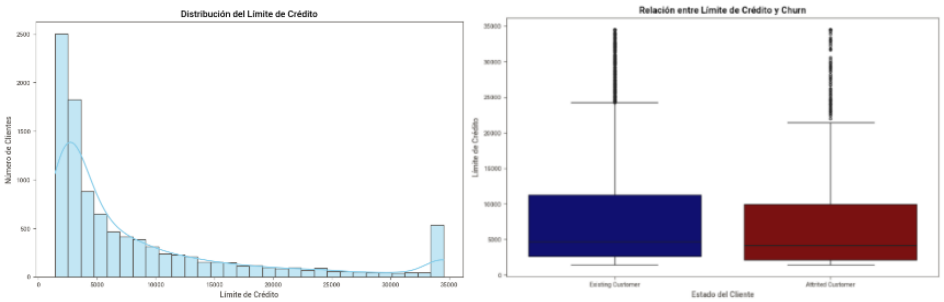

Figura 13. Distribución contacts_credit_limit

- **total_revolving_bal**: representa el saldo rotativo total que un cliente mantiene en sus cuentas de crédito con el banco, la distribución es bastante concentrada en valores más bajos, con una gran cantidad de clientes que tienen saldos cercanos a cero, lo cual indica que una proporción significativa de clientes no mantiene balances rotativos altos en sus tarjetas de crédito o líneas de crédito. Esto puede sugerir un uso conservador del crédito por parte de estos clientes.

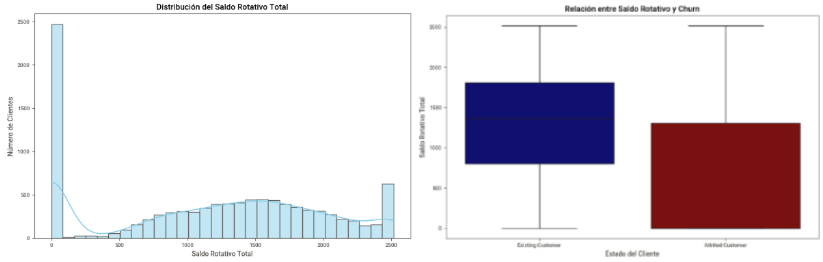

Figura 14. Distribución total_revolving_bal

- **avg_open_to_buy**: Representa el crédito promedio disponible que un cliente tiene en sus cuentas de tarjetas de crédito durante un período determinado, muestra una concentración principal en valores más bajos, con una gran cantidad de clientes que tienen relativamente poco crédito disponible.

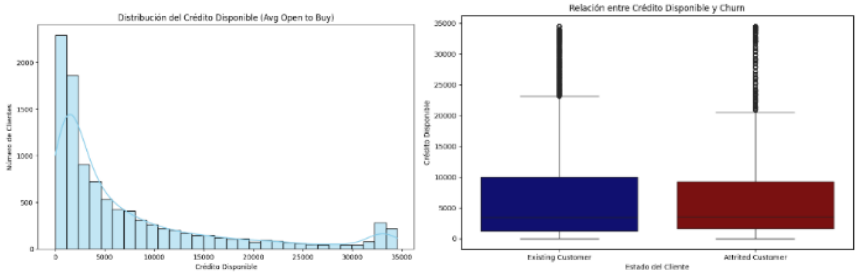

Figura 15. Distribución avg_open_to_buy

- **total_amt_chng_q4_q1:** Muestra el cambio en el monto total de transacciones de un cliente entre el cuarto trimestre y el primer trimestre. El histograma revela que la mayoría de los cambios en el monto de las transacciones están centrados alrededor de un factor de cambio de aproximadamente 0.75, con la distribución mostrando una forma bastante simétrica alrededor de esta mediana. Esto sugiere que la mayoría de los clientes tienen cambios moderados en sus patrones de gasto de un trimestre a otro.

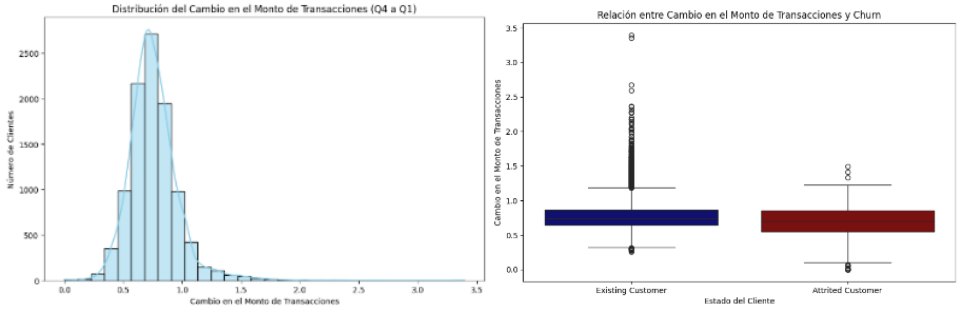

Figura 16. Distribución total_amt_chng_q4_q1

- **total_trans_amt**: Muestra el monto total de transacciones realizadas por los clientes, el histograma indica que la mayoría de las transacciones se concentran en montos menores, con picos significativos en ciertos intervalos. La distribución muestra una caída progresiva a medida que aumenta el monto de las transacciones, pero con algunos aumentos notorios en los montos más altos. Esto sugiere que mientras la mayoría de los clientes realizan transacciones menores, hay un grupo significativo que maneja montos más altos.

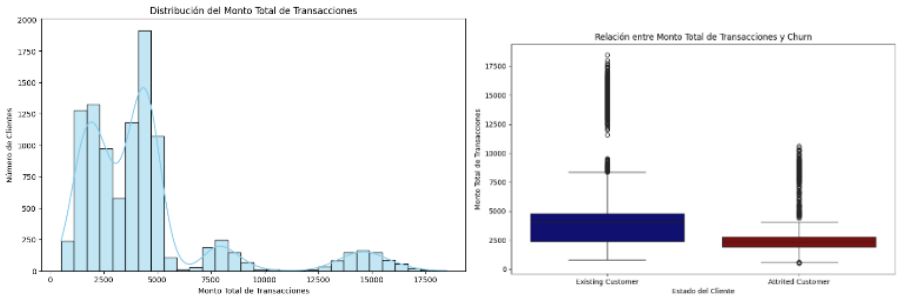

Figura 17. Distribución total_trans_amt

- **total_trans_ct:** Representa el conteo total de transacciones por cliente, el histograma muestra una distribución bimodal donde hay dos picos en torno a los 60 y cerca de las 80 transacciones. Esto puede indicar dos grupos distintos de comportamiento de transacciones entre los clientes.

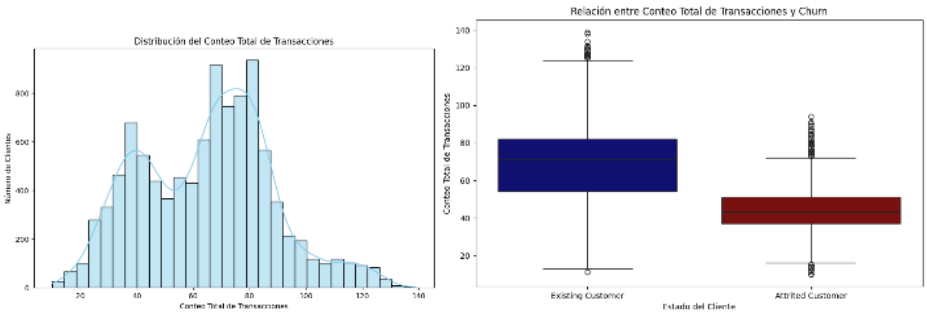

Figura 17. Distribución total_trans_ct

- **total_ct_chng_q4_q1**: Muestra el cambio en el número de transacciones de un cliente desde el cuarto trimestre hasta el primer trimestre. La distribución tiene una forma aproximadamente normal, pero con algunos valores extremos, lo que puede indicar casos específicos de cambios significativos en el comportamiento del cliente.

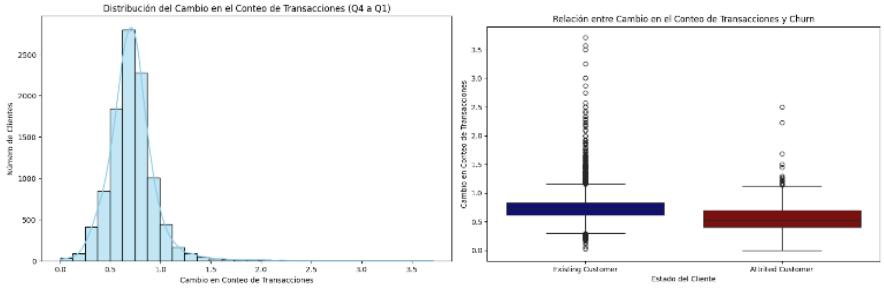

Figura 18. Distribución total_ct_chng_q4_q1

- **total_trans_ct**: Representa la cantidad total de transacciones realizadas por cada cliente, el histograma muestra un comportamiento bimodal, donde hay dos picos principales que podrían representar diferentes segmentos de clientes.

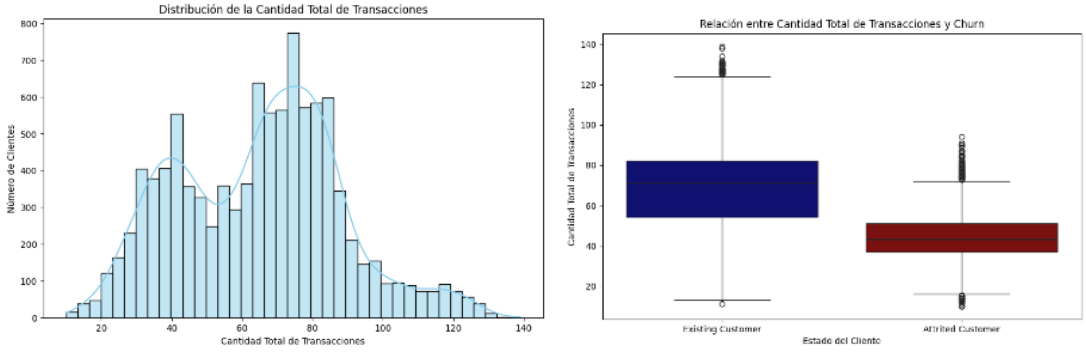

Figura 19. Distribución total_trans_ct

La entropía es una medida estadística que cuantifica la incertidumbre o el desorden en una distribución de datos. A continuación, se analizan los valores de entropía para cada variable categórica:

attrition_flag (0.636): Esta es la variable objetivo con una entropía relativamente baja, lo que indica que la distribución de los clientes que se quedan versus los que se van no es completamente uniforme.

gender (0.998): El género del cliente muestra una entropía casi máxima, sugiriendo una distribución bastante equilibrada entre géneros.

education_level (2.555): Con la entropía más alta entre las variables analizadas, el nivel educativo presenta una gran diversidad en sus categorías. Esto implica que es una variable muy rica en información y podría ser clave para entender los patrones de churn, dado que el nivel educativo puede influir en las decisiones financieras y la lealtad al banco.

marital_status (1.600): El estado civil también muestra una entropía significativa, lo que indica una buena variedad en esta variable. El estado civil puede afectar cómo los individuos utilizan los servicios bancarios, lo que a su vez podría influir en su decisión de permanecer en el banco o irse.

income_category (2.402): Esta variable, que categoriza los ingresos de los clientes, también muestra una alta entropía y es probable que sea muy relevante para la predicción de churn. Los ingresos afectan directamente la capacidad de un cliente para participar en productos financieros, lo que puede correlacionarse con su lealtad al banco.

card_category (0.416): La categoría de tarjeta de crédito tiene la entropía más baja, lo que sugiere que hay menos diversidad en esta variable. Aunque tiene baja entropía, la categoría de tarjeta puede ser un indicador del nivel de compromiso y la relación del cliente con el banco.

2.3. Análisis Correlación

La correlación de las variables deja ver que existen correlaciones positivas fuertes entre las variables “months:on_book” y “customer_age” (0.79), “avg_open_day_to_buy” y “credit_limit” (1.00) y “total_trns_ct” y “total_trns_amt” (0.81). Por lo tanto, estas correlaciones representan un comportamiento similar entre estas variables y puede implicar problemas de multicolinealidad al momento de modelar.

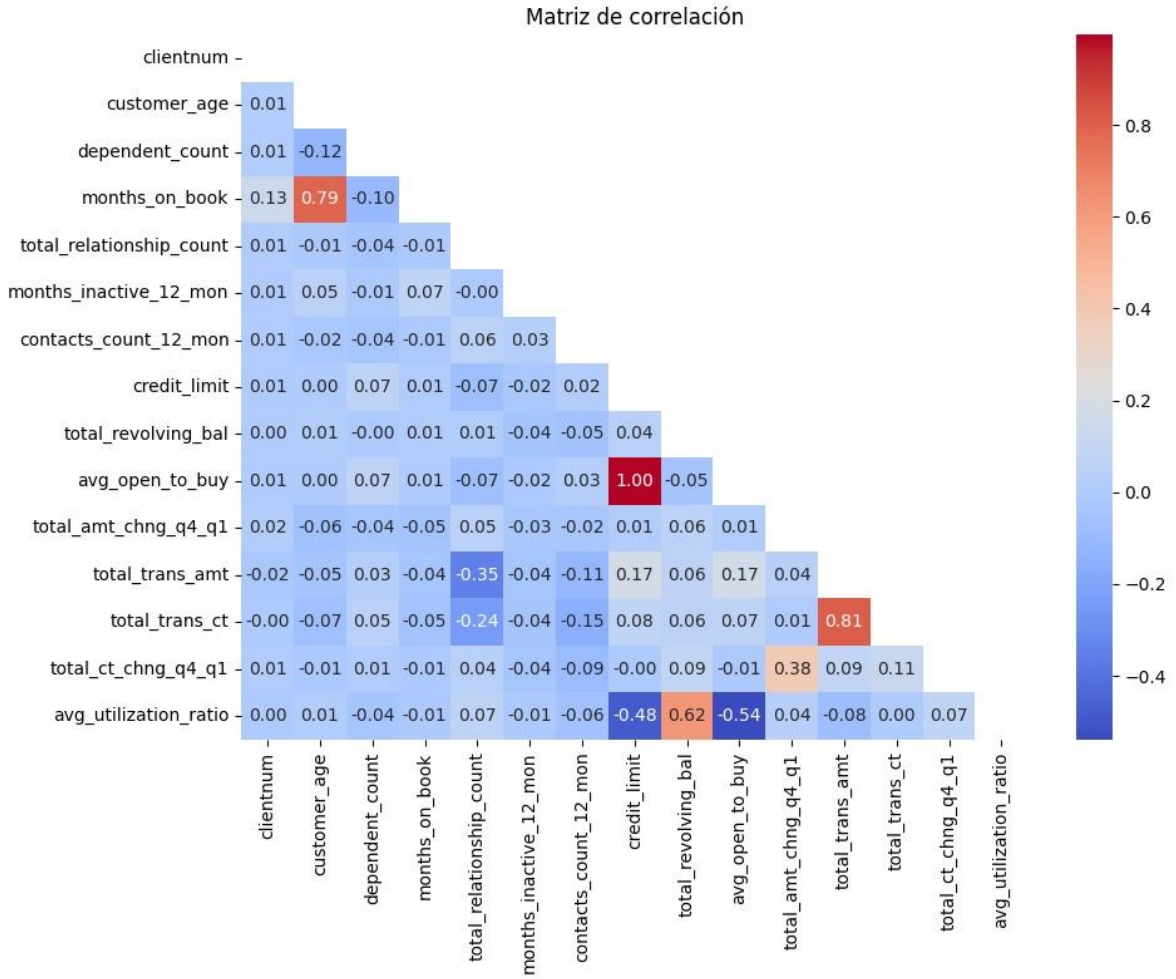

Figura 20. Correlación

2.4. Verify Data Quality

La evaluación de la calidad de los datos se llevó a cabo sobre el data set suministrado, bajo lo cual no se evidenciaron valores duplicados o datos faltantes que requirieran tratamiento alguno. Así mismo, se evidenciaron algunos datos atípicos descritos en el apartado anterior, sin embargo, de acuerdo con el entendimiento del negocio, es coherente presentarse este tipo de escenarios, teniendo en cuenta el mercado y la entidad sobre la cual se desarrolla el ejercicio.

3. Data Preparation

3.1. Clean data

Para generar el conjunto de datos adecuado para la aplicación de modelos de machine learning, se llevaron a cabo varios pasos de limpieza y preparación. Inicialmente, se creó la función CombineRareCategories para agrupar las categorías poco frecuentes bajo una nueva categoría denominada 'Other'. Este proceso ayuda a reducir la dimensionalidad de las características categóricas, evitando así el sobreajuste en los

modelos de aprendizaje. Posteriormente, se transformó la columna clientnum de tipo numérico a cadena para prevenir su tratamiento como variable continua en análisis futuros.

El siguiente paso consistió en la imputación de datos, donde se llenaron los valores faltantes en el DataFrame, excluyendo las columnas attrition_flag y clientnum. Se aplicaron diferentes estrategias de imputación dependiendo del tipo de dato (numérico o categórico), asegurando que todas las variables estén completas para el análisis. Luego, se procedió a la selección de columnas, eliminando aquellas que no eran relevantes después de las transformaciones e imputaciones realizadas, lo que permitió focalizar el análisis en los datos más significativos.

Además, se dumificaron las variables categóricas, transformándolas en 'dummies', excluyendo attrition_flag y clientnum, preparándolas para ser utilizadas en modelos de machine learning que requieren entradas numéricas exclusivamente.

Finalmente, se eliminó la columna clientnum, ya que no aportaba información relevante para los modelos. Estos fueron los pasos utilizados para asegurar que el conjunto de datos final estuviera bien estructurado y listo para la implementación efectiva de modelos de machine learning.

3.2. Construct data

La fase de construcción de la data se desarrolló a partir del analisis y la limpieza realizada previamente, por lo tanto, en un principio de llevó a cabo el proceso de dummificación de las variables categóricas, posteriormente para evitar problemas de multicolinealidad se procedió con la eliminación de una categoría por variable.

Posteriormente, se estandarizaron las variables numéricas, teniendo en cuenta que estas presentan diferentes escalas de datos lo cual conlleva a posibles sesgos en el modelo por los pesos de las mismas.

Así mismo, se evidenció desbalanceo en los datos de la variable a predecir con un 83.93% de los datos clasificados como existing customer y 16.07% de los datos clasificados como attrited customer; sin embargo, frente a ello no se llevó a cabo ningún proceso de balanceo, teniendo en cuenta que es relevante que el modelo aprende de dicha dispersión de datos.

Finalmente, se llevó a cano la partición de los datos en train con el 80% y test con el 20%; para ello se uso una semilla con el objetivo de hacer comparables los modelos. Para la medición de los resultados se utilizarán métricas que tengan en cuenta el desequilibrio de clases como precision, recall o F1-score.

3.3. Dataset description

El conjunto de datos resultante, tras un proceso de limpieza y transformación, está preparado para ser usado en diferentes modelos predictivos. En este dataset, las variables categóricas han sido convertidas en variables dummy, resultando en nuevas columnas que representan categorías como rangos de ingresos, niveles educativos, estado marital y tipo de tarjeta. Estas transformaciones han dejado valores de 1 y 0, facilitando la ejecución de modelos que requieren entradas numéricas.

Se ha mantenido una selección de variables cruciales, tales como avg_utilization_ratio, total_trans_amt, y total_revolving_bal, enfocadas en indicadores financieros clave. La estrategia de la imputación de datos se evidencia en el dataset final, ya que el número de registros en cada columna se ha conservado

íntegramente, eliminando la preocupación por valores faltantes que podrían afectar negativamente la validez de futuros modelos.

Nombre Columna	Tipo de Dato
avg_utilization_ratio	float
total_trans_amt	float
total_ct_chng_q4_q1	float
total_revolving_bal	float
months_inactive_12_mon	float
total_trans_ct	float
dependent_count	float
contacts_count_12_mon	float
customer_age	float
avg_open_to_buy	float
credit_limit	float
attrition_flag	Category
total_amt_chng_q4_q1	float
total_relationship_count	float
months_on_book	float
income_category_$40K - $60K	bool
income_category_$60K - $80K	bool

Nombre Columna	Tipo de Dato
income_category_$80K - $120K	bool
income_category_Less than $40K	bool
income_category_Unknown	bool
education_level_Doctorate	bool
education_level_Graduate	bool
education_level_High School	bool
education_level_Post-Graduate	bool
education_level_Uneducated	bool
education_level_Unknown	bool
marital_status_Married	bool
marital_status_Single	bool
marital_status_Unknown	bool
gender_M	bool
card_category_Gold	bool
card_category_Platinum	bool
card_category_Silver	bool

Tabla 5. Descripción de Variables

4. MODELING

4.1. Select modeling techniques:

El modelado de datos para abordar la problemática de predicción de abandono de clientes en la entidad financiera se ha llevado a cabo mediante diversas técnicas de aprendizaje automático. En este caso, se ha decidido enfocarse en clasificadores de aprendizaje automático debido a que son herramientas con amplio campo de aplicación para la predicción. Las principales razones detrás de la elección de éstos métodos son:

- ***Identificación de patrones en grandes conjuntos de datos:*** Los modelos de aprendizaje automático tienen la capacidad de analizar grandes volúmenes de datos de clientes y detectar patrones sutiles que pueden indicar el riesgo de abandono. Esto permite a las entidades financieras anticiparse a las necesidades y comportamientos de los clientes, tomando medidas proactivas para retenerlos.
- ***Toma de decisiones informadas:*** Al categorizar y clasificar los datos de clientes, los modelos de aprendizaje automático proporcionan a las organizaciones información valiosa para tomar decisiones informadas. Esto incluye identificar segmentos de clientes con mayor probabilidad de abandonar, lo que permite implementar estrategias específicas de retención y fidelización.
- ***Capacidad de adaptación y mejora continua:*** Los modelos de aprendizaje automático pueden aprender y mejorar con el tiempo a medida que se alimentan con nuevos datos. Esto es crucial en un entorno dinámico como el sector financiero, donde los comportamientos y preferencias de los

clientes pueden cambiar rápidamente. La capacidad de adaptación de estos modelos garantiza una predicción más precisa y actualizada del abandono de clientes.

Ahora, las técnicas utilizadas en el proyecto incluyen K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Redes Neuronales (RNN), Random Forest y AdaBoost, cada una ajustada para maximizar la capacidad predictiva y generalización del modelo. Cabe destacar que durante el proceso de modelado se han evaluado los supuestos y racionales de cada técnica para garantizar su aplicabilidad y rendimiento óptimo en la predicción de abandono de clientes en entidades financieras.

A continuación, se mencionan puntos clave de cada uno de los métodos de clasificación utilizados y cómo se han evaluado en relación con los supuestos y racionales para su aplicación:

- **K-Nearest Neighbors (KNN):**
  - **Técnica y Parámetros Utilizados:** KNN se implementó como modelo base. Se entrenó utilizando validación cruzada con cinco folds, variando el número de vecinos en un rango específico (1, 5, 10, 15, 20, 25, 30) para determinar la configuración óptima.
  - **Supuestos:** KNN supone que los puntos cercanos en el espacio de características tienen etiquetas similares, lo que implica que los clientes con características similares tenderán a tener comportamientos similares respecto al abandono.
  - **Evaluación de Supuestos:** KNN asume que los puntos cercanos en el espacio de características tienen etiquetas similares. Al variar el número de vecinos, se evalúa cómo afecta esto a la capacidad del modelo para capturar la estructura subyacente de los datos y evitar el sobreajuste.
  - **Razonamiento para su uso:** Dado que KNN no hace suposiciones específicas sobre la distribución de los datos, es útil cuando la relación entre las características y la variable objetivo puede no ser lineal o tener una estructura compleja.

- **Support Vector Machine (SVM):**
  - **Técnica y Parámetros Utilizados:** Se empleó SVM con validación cruzada de cinco folds, ajustando los parámetros de regularización C con valores de 0.1, 1y 10. Además de utilizaron los kernels RBF y POLY.
  - **Supuestos:** SVM asume que los datos son linealmente separables en un espacio de características de alta dimensión o que pueden ser transformados en un espacio donde lo son. Además, asume que la separación entre clases es clara y definida.
  - **Evaluación de Supuestos:** SVM busca encontrar el hiperplano que mejor separa las clases en el espacio de características. Al variar los parámetros de complejidad y adaptabilidad, se examina cómo se ajusta el modelo a la complejidad del conjunto de datos y su capacidad para generalizar a nuevos datos.
  - **Razonamiento para su uso:** SVM es útil cuando se busca una separación clara entre los clientes que abandonan y los que no, lo que puede ser el caso en ciertos segmentos de clientes con comportamientos claramente diferenciados.

- **Redes Neuronales (RNN):**
  - **Técnica y Parámetros Utilizados:** Se diseñaron dos tipologías de redes neuronales y se seleccionó la mejor configuración basada en la métrica obtenida por época. Se aplicó dropout y early stopping para prevenir el sobreajuste.

- ***Red Neuronal 1:*** Esta red neuronal tiene cinco capas ocultas. Cada capa oculta tiene un número específico de neuronas, respectivamente: 1000, 500, 250, 75 y 25. Utiliza la función de activación ReLU (Rectified Linear Unit) en todas sus capas ocultas. ReLU es una función que devuelve cero si el valor de entrada es negativo y el mismo valor si es positivo, lo que la hace eficaz en la activación de neuronas.
- ***Red Neuronal 2:*** Esta red neuronal también tiene cinco capas ocultas. Las neuronas en estas capas están configuradas con 500, 250, 125 y 75 unidades respectivamente. En contraste con la Red 1, aquí se utiliza la función de activación Leaky ReLU en todas las capas ocultas. La función Leaky ReLU es similar a ReLU, pero en lugar de devolver cero para valores negativos, devuelve una fracción pequeña del valor negativo, lo que ayuda a mitigar el problema de "neuronas muertas" que puede surgir en ReLU cuando las neuronas quedan inactivas.

  - **Supuestos:** Las redes neuronales suponen que los datos tienen patrones complejos y no lineales que pueden ser aprendidos mediante conexiones ponderadas entre neuronas. Además, asumen que una representación jerárquica de características puede ser aprendida automáticamente.
  - **Evaluación de Supuestos:** Las redes neuronales son poderosas en la captura de patrones complejos en datos. La comparación de diferentes arquitecturas y la aplicación de técnicas como dropout y early stopping ayudan a controlar la complejidad del modelo y mejorar su generalización.
  - **Razonamiento para su uso:** Las redes neuronales son adecuadas cuando se trabaja con conjuntos de datos complejos y no lineales. Su capacidad para aprender representaciones de características de manera jerárquica puede capturar relaciones sutiles y no lineales.

- **Random Forest:**
  - **Técnica y Parámetros Utilizados:** Esta técnica se ajustó mediante validación cruzada con cinco folds, variando parámetros clave. Los parámetros variados incluyeron n_estimators (10, 50, 100), max_depth (1, 10, 20) y min_samples_split (2, 5, 10).
  - **Supuestos:** Random Forest supone que la combinación de múltiples modelos débiles puede mejorar el rendimiento predictivo global. Además, supone que los modelos individuales sean diversificados para que no estén altamente correlacionados.
  - **Evaluación de Supuestos:** Random Forest es un método de ensemble que combina múltiples modelos para mejorar el rendimiento predictivo. Al ajustar los parámetros, se evalúa cómo afecta esto a la capacidad del ensemble para reducir el sesgo y la varianza del modelo.
  - **Razonamiento para su uso:** Este método es útil cuando se quiere mejorar la robustez y la estabilidad del modelo predictivo. Al combinar múltiples modelos, se reducen los riesgos de sobreajuste y se puede capturar una mayor variedad de patrones en los datos de clientes.

- **AdaBoost:**
  - **Técnica y Parámetros Utilizados:** Esta técnica se ajustó mediante validación cruzada con cinco folds, variando parámetros clave como: n_estimators (10, 50, 100), max_depth (1, 10, 20) y min_samples_split (2, 5, 10).
  - **Supuestos:** AdaBoost supone que la combinación de múltiples modelos débiles puede mejorar el rendimiento predictivo global. Además, supone que los modelos individuales sean diversificados para que no estén altamente correlacionados

- **Evaluación de Supuestos:** AdaBoost es un método de ensemble que combina múltiples modelos para mejorar el rendimiento predictivo. Al ajustar los parámetros, se evalúa cómo afecta esto a la capacidad del ensemble para reducir el sesgo y la varianza del modelo.
- **Razonamiento para su uso:** Este método es útil cuando se quiere mejorar la robustez y la estabilidad del modelo predictivo. Al combinar múltiples modelos, se reducen los riesgos de sobreajuste y se puede capturar una mayor variedad de patrones en los datos de clientes.

4.2. Generate test design:

Con el fin de evaluar y comparar el rendimiento de las diferentes técnicas de aprendizaje automático (K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Redes Neuronales (RNN), Random Forest y AdaBoost) que se describieron en el punto anterior, y con el objetivo de predecir el abandono de clientes en un banco, se desarrollaron los siguientes puntos:

- **Mecanismos de Comparación de Técnicas:**
  - **Validación Cruzada:** Se utilizó esta técnica para evaluar el rendimiento general de cada modelo y mitigar el riesgo de sobreajuste. La validación cruzada se implementará en la mayoría de los modelos.
  - **Comparación de Modelos:** Se utilizó la librería PyCaret para comparar los modelos desarrollados manualmente con modelos generados automáticamente, evaluando su rendimiento y eficacia.
- **Variaciones de las Técnicas que se Compararon:**
  - **K-Nearest Neighbors (KNN):** Se empleó KNN como modelo base y se cambió el número de vecinos (1, 5, 10, 15, 20, 25, 30) para determinar la configuración óptima.
  - **Support Vector Machine (SVM):** Se utilizó SVM con validación cruzada de cinco folds, ajustando los parámetros de regularización C (0.1, 1, 10) y utilizando los kernels RBF y POLY.
  - **Redes Neuronales (RNN):** Se diseñaron dos tipologías de redes neuronales y se seleccionará la mejor configuración basada en la métrica obtenida por época. Se aplicará dropout y early stopping para prevenir el sobreajuste.
  - **Random Forest y AdaBoost:** Se ajustaron estas técnicas mediante validación cruzada con cinco folds, variando parámetros clave como el número de estimadores y la profundidad máxima para Random Forest, y el número de estimadores y la tasa de aprendizaje para AdaBoost.

- **Preparación de Bases de Entrenamiento, Validación y/o Prueba:**
  - **División de Conjunto de Datos:** Se dividió el conjunto de datos en conjuntos de entrenamiento, validación y prueba para evaluar el rendimiento de cada modelo.
  - **Preprocesamiento de Datos:** Se realizó el preprocesamiento de datos para asegurar que los conjuntos de entrenamiento, validación y prueba estén normalizados y sean adecuados para el análisis.

- **Mecanismos de Control del Sobreajuste o Subajuste (Overfitting/Underfitting):**
  - **Early Stopping:** Se utilizó esta técnica para interrumpir el entrenamiento de los modelos neuronales cuando el rendimiento en la validación comienza a disminuir, evitando el sobreajuste.

- **Dropout:** Se aplicó *dropout* en las redes neuronales para reducir la complejidad del modelo y prevenir el sobreajuste.

- **Mecanismos de Selección de Variables (Feature Selection):**
  - **GridSearchCV:** Se utilizó esta técnica para ajustar hiperparámetros críticos y realizar la selección de características, optimizando los modelos para alcanzar el mejor rendimiento posible.

- **Evaluación del Rendimiento:**
  - **Métricas de Evaluación:** Se utilizaron las métricas de precisión, recall, F1-score y área bajo la curva ROC para evaluar el rendimiento de cada modelo.
  - **Comparación de Resultados:** Se compararán los resultados de cada modelo para determinar el que mejor se adapta a la predicción de abandono de clientes en el banco.

4.3. Build model:

Parametrización y entrenamiento: Se definieron y ajustaron los parámetros iniciales para cada modelo, seguido de entrenamientos utilizando conjuntos de datos preparados y técnicas de estandarización de datos.

Optimización y evaluación: La selección de hiperparámetros y la validación cruzada ayudaron a determinar las configuraciones óptimas para cada modelo, asegurando que cada uno operara con la máxima eficacia.

Prevención de sobreajuste: Se implementaron técnicas específicas como dropout en las RNN y early stopping, para controlar el sobreajuste y mejorar la generalización de los modelos en datos no vistos.

5. EVALUATION

Buscando cumplir con el objetivo de predecir el abandono de clientes de la entidad financiera, utilizando técnicas de clasificación avanzadas. Las técnicas evaluadas incluyen K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Redes Neuronales (RNN), Random Forest y AdaBoost, seleccionadas por su capacidad para identificar patrones complejos en grandes volúmenes de datos, lo cual es crucial en el dinámico sector financiero.

K-Nearest Neighbors (KNN): mostró un desempeño moderado, con un número óptimo de vecinos identificado en 5 a través de GridSearch. La exactitud en el conjunto de prueba no mejoró significativamente después de este valor, lo que sugiere que modelos más simples eran adecuados para evitar el sobreajuste.

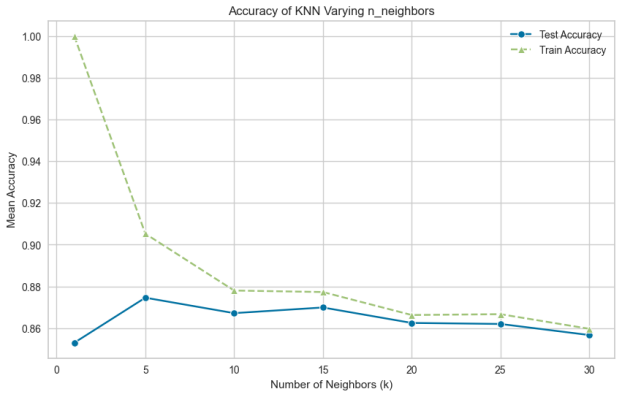

Figura 21. Comportamiento Modelo KNN

Support Vector Machine (SVM): Demostró una mejora continua en la exactitud a medida que se incrementaba el parámetro de regularización C, especialmente con el kernel RBF, alcanzando su mejor desempeño con C=10

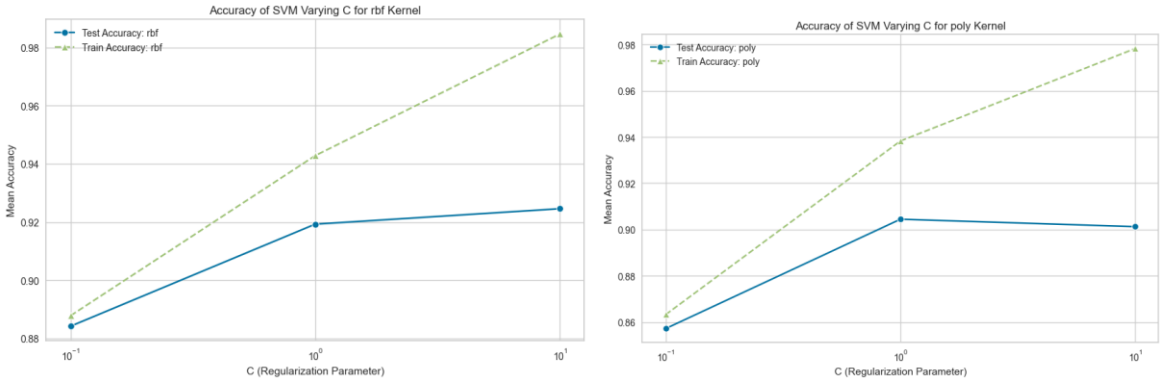

Figura 22. Comportamiento Modelo SVM

Redes Neuronales (RNN): Mostraron una mejora significativa en la exactitud con el aumento de épocas, apoyadas por técnicas como el dropout y el early stopping para mitigar el sobreajuste. La topología 1 se identificó como la más eficaz.

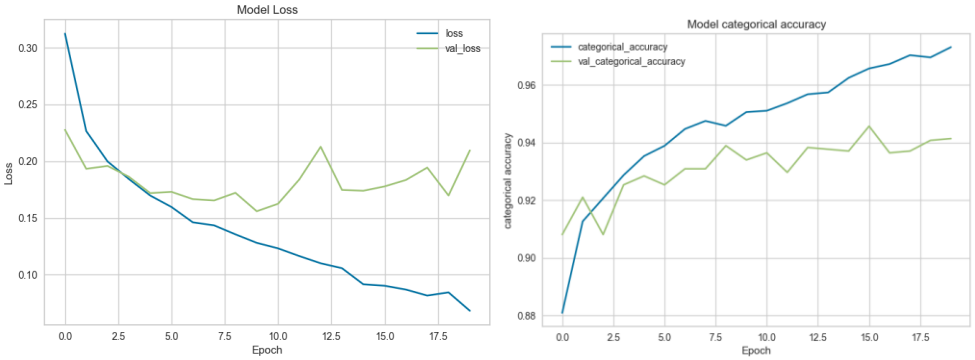

Figura 23. Comportamiento Modelo RNN

Random Forest: *Obtuvo resultados óptimos con una profundidad máxima de 20, un número de estimadores de 100 y un mínimo de muestras por división de 2, destacando en la reducción del sesgo y la varianza.*

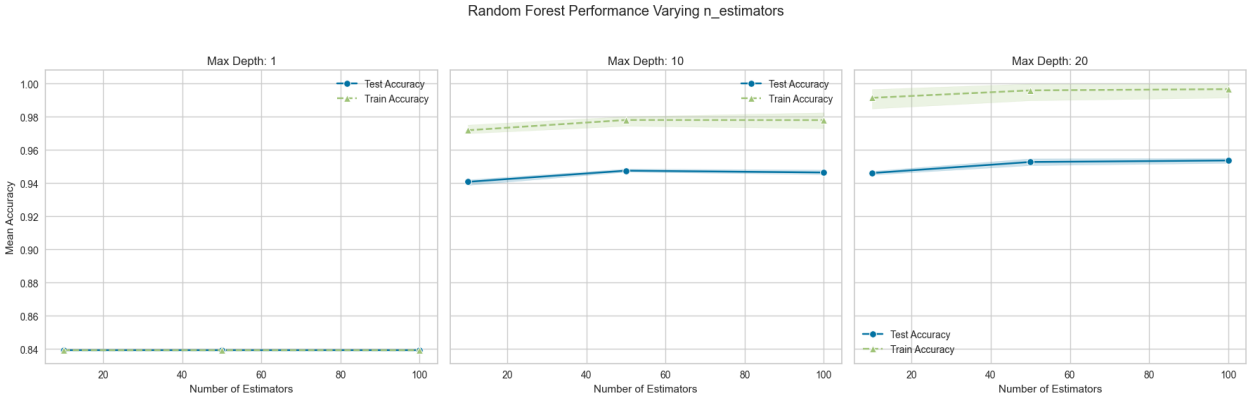

Figura 24. Comportamiento Modelo Random Forest

AdaBoost: Sobresalió con una tasa de aprendizaje de 1.0 y 100 estimadores, mostrando un alto nivel de precisión y robustez, y fue especialmente efectivo en el manejo de características diversas y complejas.

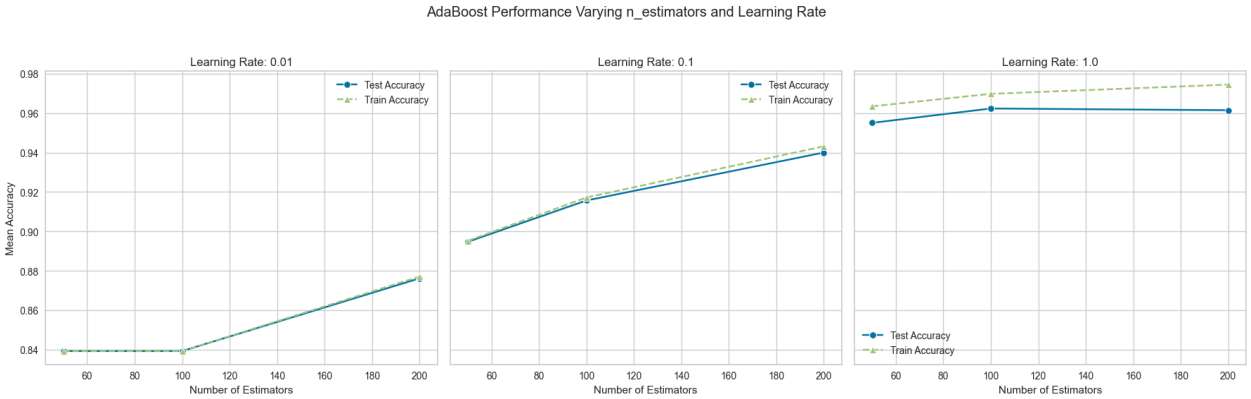

Figura 25. Comportamiento Modelo AdaBoot

La evaluación exhaustiva de los modelos para predecir la deserción de clientes en entidades financieras, particularmente usando el F1 Score como métrica clave debido al desbalance inicial de clases, destaca al modelo AdaBoost como la elección superior. Este modelo no solo mostró la puntuación F1 más alta (0.8625), sino que también registró valores de AUC (0.9879) y precisión (0.9566) excepcionalmente altos, reforzando su idoneidad para la implementación en la predicción de churn. Estos resultados subrayan la capacidad de AdaBoost para manejar de manera eficiente tanto la precisión como la exhaustividad, proporcionando un equilibrio óptimo crucial para la toma de decisiones informadas en el dinámico sector financiero.

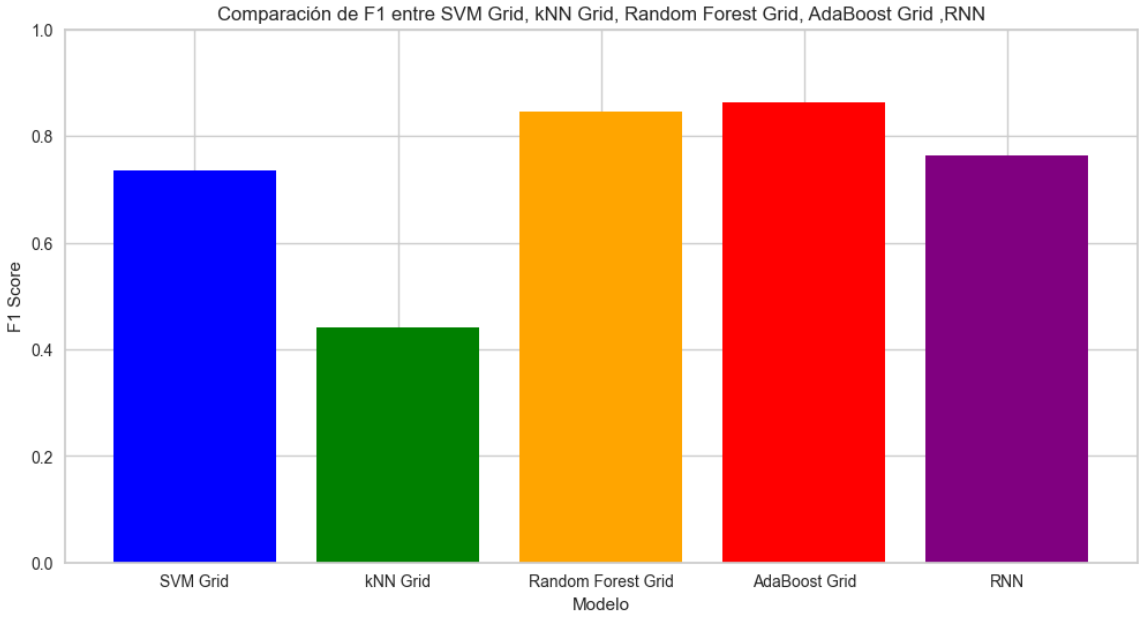

Figura 26. Comparación F1

Modelo	**F1**	**AUC**	**Precisión**	**Sensibilidad**	**Especificidad**	**Positive Rate**	**Negative Rate**
Support Vector Machine (SVM)	0.735	0.943	0.921	0.683	0.966	0.796	0.941
K-Nearest Neighbors (KNN):	0.440	0.812	0.873	0.311	0.981	0.754	0.882
Random Forest	0.847	0.988	0.955	0.782	0.988	0.924	0.959
AdaBoost	0.863	0.988	0.957	0.849	0.977	0.876	0.971
Redes Neuronales (RNN)	0.764	0.962	0.927	0.738	0.963	0.792	0.951

Tabla 6. Métricas de medición

Adicionalmente, la robustez de AdaBoost fue confirmada mediante comparaciones automatizadas usando la librería PyCaret, donde, aunque el Light Gradient Boosting Machine fue automáticamente seleccionado como el mejor método, AdaBoost sigue siendo destacado entre los mejores modelos. Esto valida la selección de AdaBoost dada su notable consistencia en desempeño a través de diversas métricas y su capacidad para adaptarse a nuevos datos, lo cual es esencial para responder eficazmente a las rápidas variaciones en comportamientos y preferencias de los clientes.

Model	**Accuracy**	**AUC**	**Recall**	**Prec.**	**F1**	**Kappa**	**MCC**	**TT (Sec)**	
lightgbm	Light Gradient	0.970	0.993	0.883	0.929	0.905	0.887	0.8881	0.108

	Boosting Machine								
gbc	Gradient Boosting Classifier	0.963	0.989	0.838	0.924	0.878	0.856	0.8579	0.180
ada	Ada Boost Classifier	0.956	0.984	0.830	0.887	0.857	0.831	0.8318	0.052
rf	Random Forest Classifier	0.953	0.987	0.774	0.922	0.841	0.813	0.8182	0.095
dt	Decision Tree Classifier	0.938	0.879	0.794	0.814	0.803	0.766	0.7663	0.012
et	Extra Trees Classifier	0.9244	0.9754	0.5764	0.9252	0.7087	0.6681	0.6937	0.068
lda	Linear Discriminant Analysis	0.905	0.922	0.606	0.757	0.672	0.617	0.623	0.009
qda	Quadratic Discriminant Analysis	0.860	0.925	0.814	0.547	0.651	0.568	0.5878	0.009
nb	Naive Bayes	0.896	0.873	0.596	0.710	0.647	0.587	0.5904	0.008
lr	Logistic Regression	0.897	0.911	0.524	0.762	0.620	0.563	0.5766	0.076
knn	K Neighbors Classifier	0.888	0.877	0.536	0.699	0.606	0.542	0.5489	0.024
ridge	Ridge Classifier	0.9008	0.9218	0.4726	0.8425	0.6038	0.5523	0.5834	0.008
svm	SVM - Linear Kernel	0.815	0.654	0.210	0.407	0.210	0.149	0.1853	0.016
dummy	Dummy Classifier	0.839	0.500	0.000	0.000	0.000	0.000	0	0.010

Tabla 7. Resultados PyCaret

Dado su desempeño superior, se propone a AdaBoost como el modelo principal para la estrategia de retención de clientes. Además, se recomienda implementar una estrategia de 'champion challenger' donde AdaBoost, como modelo 'champion', sea continuamente comparado en desempeño contra nuevos modelos emergentes desarrollados a través de PyCaret, asegurando así una mejora continua y la adopción de las técnicas más efectivas y actualizadas para la predicción de churn. Esta aproximación no solo reafirma la elección de AdaBoost basado en su actual superioridad, sino que también fomenta la innovación constante y la adaptabilidad de los modelos predictivos generados dentro de la entidad financiera.

6. Produce Final Report

En el mercado financiero colombiano, las entidades bancarias enfrentan desafíos significativos en la retención de clientes y en la mejora de su fidelidad debido a, por un lado, la alta competencia y, por el otro, a la creciente presencia de fintechs y banca digital. La presente propuesta de estrategias se enfoca en abordar estos desafíos mediante el uso de análisis avanzado de datos para identificar clientes con alta propensión a abandonar el banco y la implementación de programas personalizados que mejoren la experiencia del cliente resultando en un aumento de su lealtad. Las estrategias detalladas a continuación no solo buscan reducir la tasa de abandono de clientes, sino también mejorar la experiencia y preferencia de los clientes hacia los productos y servicios del banco.

Estrategia 1: Programa de Fidelización Personalizado

Implementar un programa de fidelización que ofrezca beneficios personalizados basados en el análisis del comportamiento de cada cliente. Este programa puede incluir recompensas, descuentos y servicios exclusivos adaptados a las preferencias y necesidades individuales de los clientes.

Componentes Clave:

- Análisis de Datos del Cliente: Utilizar análisis de big data y machine learning para entender el comportamiento de los clientes.
- Beneficios Personalizados: Crear una variedad de recompensas, descuentos y servicios exclusivos que se ajusten a las preferencias y necesidades específicas de cada cliente.
- Plataforma de Gestión de Fidelización: Desarrollar una plataforma digital que gestione y rastree la participación de los clientes en el programa de fidelización.

Métodos de Implementación:

1. Recopilación y Análisis de Datos: Implementar herramientas de análisis de big data para recopilar y analizar datos de clientes (transacciones, interacciones con el banco, uso de productos y servicios, etc.).
2. Diseño del Programa de Fidelización: Crear estructuras de recompensas y beneficios basados en los segmentos de clientes identificados en el análisis de datos.
3. Lanzamiento de la Plataforma de Fidelización: Desarrollar y lanzar una plataforma digital integrada (aplicación móvil y web) que permita a los clientes acceder y gestionar sus beneficios.

Posibles Costos Asociados:

- Desarrollo de Software y Herramientas de Análisis
- Implementación y Mantenimiento de la Plataforma
- Costos de Recompensas y Beneficios
- Marketing y Comunicación

Beneficios Esperados:

- Reducción de la Tasa de Abandono
- Incremento en la Lealtad del Cliente
- Aumento en el Uso de Productos y Servicios
- Mayor Satisfacción del Cliente: Mejora en el Net Promoter Score (NPS)

Pasos Específicos a Seguir:

1. Identificar población a retener: Usando el modelo de ML seleccionado (Adaboost), identificar a la población propensa a abandonar el banco.
2. Diseño de Ofertas Personalizadas: Crear ofertas y beneficios específicos para este grupo de clientes a partir de un análisis adicional de características (análisis descriptivo).
3. Comunicación del Programa: Lanzar una campaña de marketing para informar a los clientes sobre el programa de fidelización y cómo participar.
4. Monitorización y Ajustes: Monitorizar continuamente la participación y satisfacción del cliente, ajustando el programa según sea necesario para maximizar su efectividad.

Estrategia 2: Segmentación de Clientes y Ofertas Personalizadas

Utilizar técnicas avanzadas de segmentación de clientes para diseñar campañas de marketing personalizadas que aborden las necesidades específicas de cada segmento de clientes. Esto incluirá correos electrónicos personalizados, notificaciones push y promociones dirigidas a través de la aplicación del banco.

Componentes Clave:

- Segmentación de Clientes: Uso de análisis de datos para segmentar a los clientes en grupos basados en comportamientos, preferencias y valor potencial.
- Campañas Personalizadas: Diseño de campañas de marketing dirigidas específicamente a cada segmento.
- Automatización de Marketing: Implementación de sistemas de automatización para gestionar y ejecutar campañas de marketing personalizadas en múltiples canales.

Métodos de Implementación:

1. Identificar población a retener: Usando el modelo de ML seleccionado (Adaboost), identificar a la población propensa a abandonar el banco.
2. Análisis y Segmentación de Datos: Implementar herramientas de análisis para segmentar la base de clientes identificados en grupos homogéneos (técnicas de clustering)
3. Diseño de Contenido Personalizado: Crear contenido y ofertas personalizadas para cada segmento de cliente identificado.
4. Automatización y Ejecución de Campañas: Utilizar plataformas de automatización de marketing para enviar mensajes personalizados a través de correos electrónicos, notificaciones push y otros canales digitales.

Posibles Costos Asociados:

- Desarrollo de Contenido Personalizado
- Implementación de Automatización de Marketing
- Marketing y Promociones

Beneficios Esperados:

- Mejora en la Tasa de Conversión de Campañas
- Aumento en el Engagement de Clientes
- Incremento en la Fidelidad del Cliente
- Aumento en la Adopción de Productos y Servicios Adicionales

Pasos Específicos a Seguir:

1. Identificar población a retener: Usando el modelo de ML seleccionado (Adaboost), identificar a la población propensa a abandonar el banco.
2. Recopilación de Datos del Cliente: Recopilar datos detallados sobre el comportamiento y las preferencias de éstos clientes.
3. Análisis y Segmentación: Utilizar técnicas de análisis avanzado para crear segmentos de clientes.
4. Desarrollo de Contenido Personalizado: Diseñar mensajes y ofertas específicas para cada segmento.
5. Lanzamiento de Campañas: Implementar las campañas personalizadas utilizando herramientas de automatización de marketing.
6. Evaluación y Ajustes: Monitorizar el rendimiento de las campañas y ajustar las estrategias basadas en los resultados obtenidos.

Estrategia 3: Mejora de la Experiencia del Cliente

Implementar mejoras continuas en la experiencia del cliente tanto en sucursales físicas como en plataformas digitales. Esto incluye la formación continua del personal en atención al cliente, así como la optimización de la interfaz de usuario en aplicaciones y sitios web.

Componentes Clave:

- Optimización de Canales Digitales: Mejora de aplicaciones móviles, servicios de banca en línea y otros canales digitales para hacerlos más intuitivos y fáciles de usar.
- Formación del Personal: Programas de formación continua para mejorar las habilidades de atención al cliente del personal del banco.
- Feedback del Cliente: Sistemas para recopilar y analizar el feedback del cliente en tiempo real.

Métodos de Implementación:

1. Auditoría de la Experiencia del Cliente: Realizar una auditoría detallada de la experiencia del cliente en todos los puntos de contacto (digitales y físicos).
2. Optimización de Plataformas Digitales: Rediseñar la interfaz de usuario y mejorar las funcionalidades de las aplicaciones móviles y la banca en línea.
3. Formación y Desarrollo del Personal: Implementar programas de formación continua para el personal en atención al cliente y en el uso de nuevas tecnologías.

Posibles Costos Asociados:

- Auditoría y Consultoría de Experiencia del Cliente
- Desarrollo y Optimización de Plataformas Digitales
- Formación y Desarrollo del Personal
- Sistemas de Feedback del Cliente

Beneficios Esperados:

- Incremento en la Satisfacción del Cliente: Mejora en el NPS
- Reducción en el Tiempo de Resolución de Problemas
- Aumento en el Uso de Canales Digitales
- Mejora en la Eficiencia Operativa: Reducción de costos operativos

Pasos Específicos a Seguir:

1. Diagnóstico Inicial: Realizar un diagnóstico completo de la experiencia actual del cliente en sucursales y plataformas digitales.
2. Desarrollo de un Plan de Mejora: Crear un plan detallado para optimizar los canales digitales y mejorar la formación del personal.
3. Implementación de Mejoras: Implementar las mejoras planificadas en las plataformas digitales y realizar programas de formación para el personal.
4. Recopilación y Análisis de Feedback: Utilizar herramientas de análisis para recopilar feedback del cliente y medir la satisfacción.
5. Ajustes Continuos: Realizar ajustes continuos basados en el feedback del cliente y los datos de uso para asegurar una experiencia del cliente óptima.

Estas estrategias detalladas, junto con los costos y beneficios esperados, ofrecen un marco claro para implementar mejoras significativas en la retención, fidelidad y experiencia del cliente en el banco.

Referencias

Bancolombia. (s/f). Uno de cada cuatro colombianos es cliente de banco. Recuperado de https://www.bancolombia.com/acerca-de/sala-prensa/noticias/productos-servicios/uno-de-cada-cuatro-colombianos-es-cliente-de-banco

Banco de Bogotá. (2021). Presentación corporativa marzo 2021. Recuperado de https://www.bancodebogota.com/wps/themes/html/banco-de-bogota/pdf/relacion-con-el-inversionista/sobre-el-banco/presentacion/presentacion-corporativa-marzo-2021.pdf

Banco de la República de Colombia. (2023). Reporte de estabilidad financiera segundo semestre 2023. Recuperado de https://www.banrep.gov.co/es/publicaciones-investigaciones/reporte-estabilidad-financiera/segundo-semestre-2023

Davivienda. (s/f). Sobre nosotros. Recuperado de https://www.davivienda.com/wps/portal/personas/nuevo/personas/quienes_somos/sobre_nosotros#:~:text=Somos%20el%20Banco%20exclusivo%20en,m%C3%A1s%20de%202.840%20cajeros%20autom%C3%A1ticos.

Portafolio. (s/f). BCG analizó el comportamiento de los usuarios bancarios en Colombia. Recuperado de https://www.portafolio.co/mis-finanzas/bcg-analizo-el-comportamiento-de-los-usuarios-bancarios-en-colombia-589212

Passport. (s.f.). *Credit Cards in Colombia*. Obtenido de https://www.portal.euromonitor.com/?keTF4RUVfyRc7iCR3dTtY4TCYoMWOQZHwKgAzkqYcc5YznxWgsZqCw%3d%3d

Semana. (s/f). Bancos: conoce todos los productos que tu banco te ofrece. Recuperado de https://www.semana.com/consumo-inteligente/articulo/bancos-conoce-todos-los-productos-que-su-banco-le-ofrece/80419/

Superintendencia Financiera de Colombia, SFC. (2023). Lista general de entidades vigiladas por la Superintendencia Financiera de Colombia. Obtenido de https://www.superfinanciera.gov.co/inicio/industrias-supervisadas/entidades-vigiladas-por-la-superintendencia-financiera-de-colombia/lista-general-de-entidades-vigiladas-por-la-superintendencia-financiera-de-colombia-61694

Superintendencia Financiera de Colombia, SFC. (2023). Informe de tarjetas de crédito y débito. Obtenido de https://www.superfinanciera.gov.co/inicio/informes-y-cifras/cifras/establecimientos-de-credito/informacion-periodica/mensual/informe-de-tarjetas-de-credito-y-debito-60952