Proyecto Final

Juan Felipe González Rodríguez

Oscar Javier Ramirez Serrano

Estudiantes de Analitica para la inteligencia de negocios

Profesor Julian David Reyes Rueda, Departamento de Ingeniería Industrial

GUIA TALLERES ANALÍTICA

BASADOS EN CRISP-DM

Pontificia Universidad Javeriana, Bogotá, Colombia

FACULTAD DE INGENIERÍA
Nov 2025

BUSINESS UNDERSTANDING

Determine Business objectives

Background: Para comprender mejor el contexto empresarial de este banco, se realizaron tres análisis: Pestle, fuerzas de Porter y análisis SWOT que nos permite identificar los factores externos, las principales fuerzas competitivas y los factores internos (Fortalezas y Debilidades) y externos (Oportunidades y Amenazas) que impactan e influyen en el negocio.

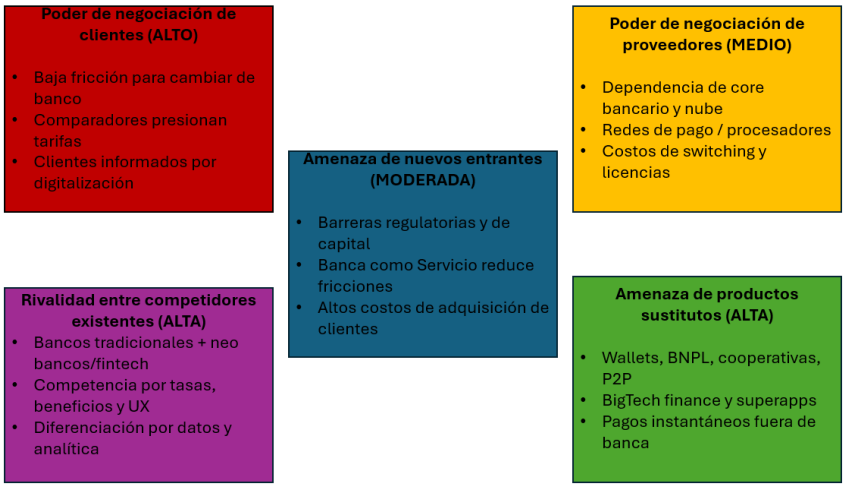

Gráfico 1. Análisis fuerzas de Porter

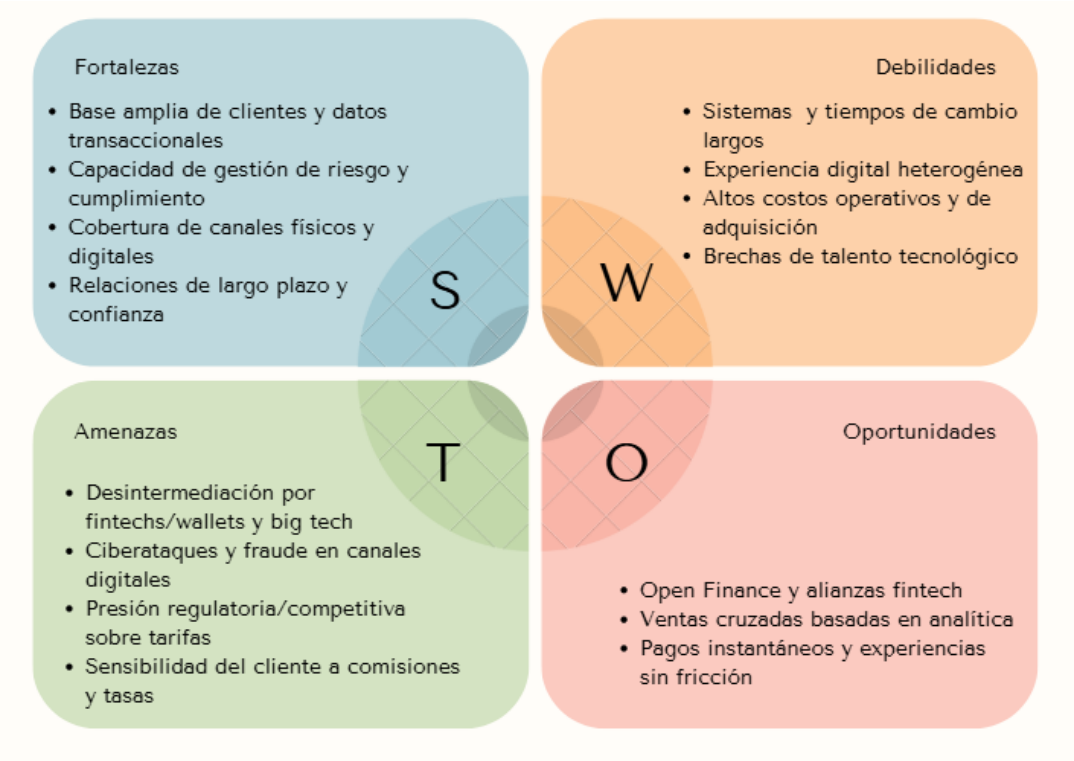

Gráfico 2. Análisis SWOT

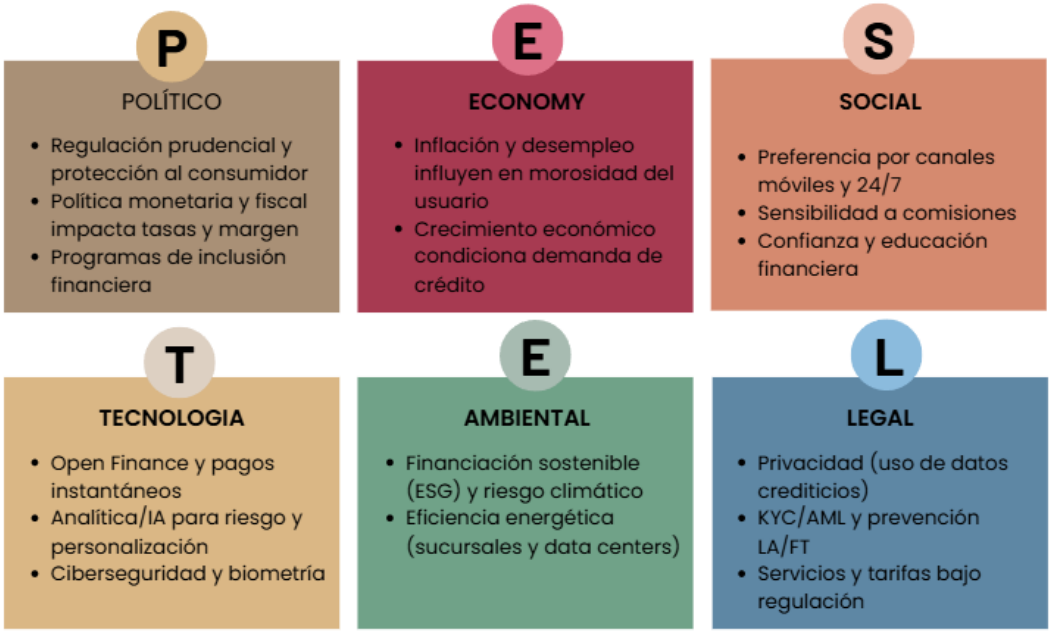

Gráfico 3. Análisis PESTLE

Contexto

La Dinámica de Pérdida y Ganancia de Clientes en Sistemas Bancarios

La industria bancaria opera en un entorno de alta rivalidad competitiva e inercia baja del cliente. La dinámica de ganancia y pérdida de clientes es un factor crítico de rentabilidad, impulsada por varias fuerzas:

1. Pérdida (Fricción Baja): El Poder de negociación de los clientes es ALTO debido a la baja fricción para cambiar de banco y el acceso a información que les permite presionar sobre tarifas. La competencia se intensifica con los bancos tradicionales, los neobancos/fintech, y los productos sustitutos (como wallets y BNPL), que aumentan el riesgo de deserción.

2. Ganancia (Oportunidades): Las oportunidades de crecimiento se centran en la venta cruzada basada en analítica, las alianzas fintech, y la capacidad de ofrecer experiencias sin fricción a través de la transformación digital.

Definición de Churn (Deserción)

El término Churn (o deserción) se refiere a la pérdida de clientes activos durante un período específico. (2025.lenovo) En el contexto de este proyecto, se define como:

- Problema de Negocio: La reducción del 15% en la cantidad de clientes activos que experimentó la entidad en el último trimestre.

- Variable Objetivo: La variable attrition_flag indica si el cliente permaneció (Existing Customer) o desertó (Attrited Customer) al cierre del período de análisis (cierre de 2022).

La identificación temprana de Churn es vital para el banco, ya que la retención de clientes es significativamente menos costosa que la adquisición. El objetivo del análisis es, por lo tanto, construir un modelo predictivo para identificar a los clientes más propensos a retirarse y activar programas de retención antes de que el abandono sea definitivo.

Business goal:

1. El objetivo principal del negocio es **reducir la deserción de clientes minoristas en un 20% en 12 meses**[4]. Esto se logrará priorizando los clientes de mayor riesgo identificados por modelos predictivos, y ejecutando campañas de retención personalizadas[5]. El requisito técnico mínimo para el modelo es un **AUC >= 0.80**

Determine Data mining goals

Data mining goal

1. Construir un modelo de **Clasificación Binaria** para predecir la probabilidad de que un cliente abandone (Churn) al cierre de 2022.

2. Optimizar el modelo para maximizar la capacidad de **detección de Churners (Recall)**, asegurando la identificación de la mayor cantidad de clientes en riesgo.

3. Identificar y cuantificar las **variables de comportamiento y demográficas** más influyentes en la decisión de deserción (Churn).

Data mining success criteria

1. **Métrica de Detección (Recall de Churn):** El modelo debe alcanzar un **Recall 0.85** en el conjunto de prueba, para detectar a más del 85% de los clientes que realmente abandonarán.

2. **Métrica de Robustez (AUC):** El modelo debe cumplir con el requisito de negocio, logrando un **AUC 0.95**

DATA UNDERSTANDING

El banco entregó una base anonimizada con 10127 registros y 21 variables que tienen información demográfica, económica y de actividad financiera de los clientes durante 2021, adicional a esto, también se cuenta con la variable objetivo attrition_flag que indica si el cliente permaneció o desertó al cierre de 2022. Las variables se distribuyen de la siguiente manera:

1. Identificación (clientnum, llave técnica)
2. Objetivo (attrition_flag, binaria *Existing/Attrited*)
3. Perfil (customer_age, gender, education_level, marital_status, income_category, card_category)
4. Relación (months_on_book, total_relationship_count)
5. Actividad y contacto reciente (months_inactive_12_mon, contacts_count_12_mon);
6. Crédito y uso (credit_limit, total_revolving_bal, avg_open_to_buy, avg_utilization_ratio)
7. Transaccionalidad y variaciones Q4/Q1 (total_trans_amt, total_trans_ct, total_amt_chng_q4_q1, total_ct_chng_q4_q1).

Tabla 1. Contine el resumen de las variables que se indican anterioremente

Variable	Tipo de Dato	Función
attrition_flag	Binaria (Objetivo)	**Cliente Deserción (1)** o Existente (0)
Variables Transaccionales (7)	Numéricas	Valor y conteo de transacciones, y sus variaciones Q4/Q1[13].
Variables de Crédito (4)	Numéricas	Límite, balance rotativo, disponible para compra y ratio de utilización
Variables de Perfil (6)	Categóricas/Numéricas	Edad, género, estado civil, nivel de ingreso, etc.

En términos de tipos de datos observados: 5 son float64, 10 int64 y 6 object. Esta estructura permite estudiar tanto rasgos estáticos del cliente como señales de comportamiento recientes relevantes para la deserción.

EDA y Hallazgos Clave

1. Correlación Transaccional: La matriz de correlación evidencia una alta asociación positiva entre total_trans_ct y total_trans_amt. Estas variables se conservaron como representativas de la actividad del cliente, siendo críticas para el modelo.

2. Relación Crédito-Uso: Los límites altos se correlacionan con menor ratio de utilización, mientras que el saldo revolvente (total_revolving_bal) y la utilización (avg_utilization_ratio) tienen una correlación positiva.

3. Riesgo por Inactividad: Los clientes con 4 meses de inactividad muestran el porcentaje más alto de deserción (30% en ese subgrupo).

4. Marital Status: El estado civil muestra una influencia mínima en la tasa de deserción, con tasas entre 15.1% y 17.2%.

DATA PREPARATION

Se implementó un pipeline de preprocesamiento robusto para asegurar la calidad de los datos para el modelado, evitando data

Tabla 2. Permite ver los procesos llevados a cabo para la preparación de los datos

Proceso	Justificación
Limpieza	Estandarización de nombres a minúsculas y mapeo de attrition_flag a binaria (1/0).
Exclusión	Se excluyó clientnum (identificador) para prevenir la fuga de datos.
Imputación	Se implementó SimpleImputer dentro del pipeline (preprocessing.py) para manejar cualquier valor nulo futuro, aunque no se encontraron nulos explícitos inicialmente.
Escalado	Se aplicó StandardScaler a las variables numéricas (transaccionalidad, crédito) para evitar que variables con mayor rango dominen el modelo.
Tratamiento de Etiquetas	Se mantuvieron etiquetas **"Unknown"** en variables categóricas (ej., marital_status) como una categoría válida, pues su efecto puede ser relevante para la deserción.

Siendo así, se estandarizaron los nombres de columnas a minúsculas con guion bajo para evitar errores en el pipeline; la variable objetivo attrition_flag se mapeó a binaria (1 = *Cliente dado de baja/ no estará el prox año*, 0 = *Cliente existente*) para habilitar el aprendizaje supervisado, se excluyó el identificador clientnum del set de *features* para prevenir *data leakage*. No se quitaron valores nulos porque no se encontraron valores perdidos explícitos en el *info* del *dataframe*; las etiquetas "Unknown" presentes en variables categóricas como en educación, estado civil, ingreso, se mantuvieron como categoría válida para capturar su efecto en el modelo. Tampoco se eliminaron outliers porque resultan importante a la toma de decisiones en el sentido de negocio (colas largas en montos/transacciones). Se construyó una matriz de correlación y se observa una alta asociación positiva entre las variables transaccionales (total_trans_ct, total_trans_amt y sus cambios Q4/Q1), por esto, para evitar multicolinealidad, se van a conservar solo dos variables representativas de este bloque. Para las variables de crédito aparecen relaciones coherentes con el negocio: límites más altos se traducen en mayor "open to buy" y menor utilización, mientras que un saldo revolvente elevado

reduce el disponible y eleva la utilización. Finalmente, el número de contactos en 12 meses tiende a moverse en sentido contrario a la actividad transaccional; es decir, menos transacciones y más contactos pueden ser señales tempranas de fricción y, por tanto, indicios de riesgo de baja.

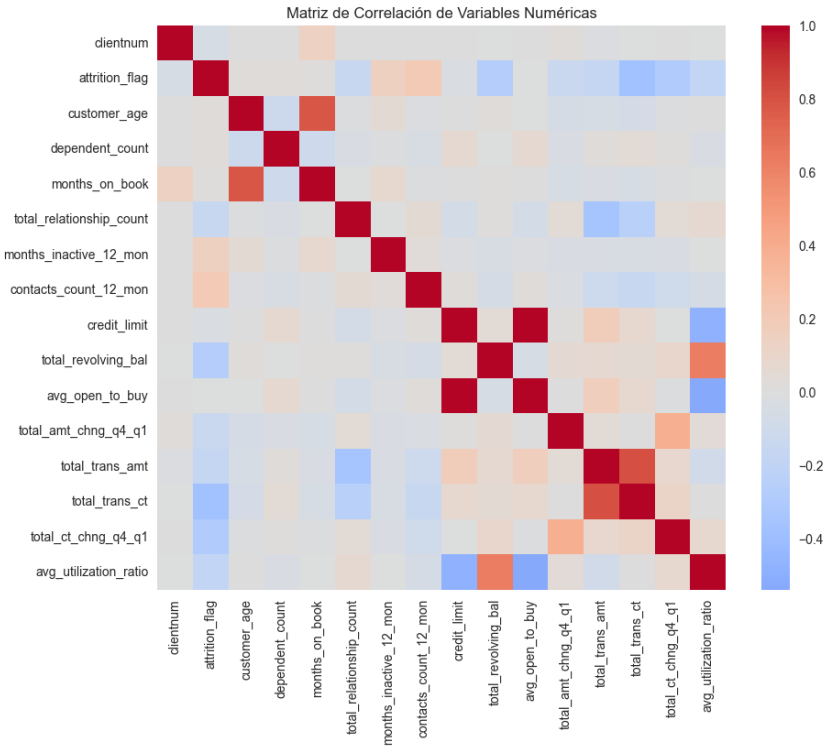

Gráfico 4 Matriz de correlación

Se realizó un boxplot de la edad de los clientes con la distribución de sus salarios y se encontró que las distribuciones de edad son muy similares entre rangos de ingreso: las medianas rondan 46–49 años, con algunos outliers en los extremos (20s y >70). Esto sugiere que edad e ingreso no muestran un gradiente claro y, por sí solos, probablemente no explican grandes diferencias para ser dados de baja; su utilidad estará más en interacciones que en efectos directos.

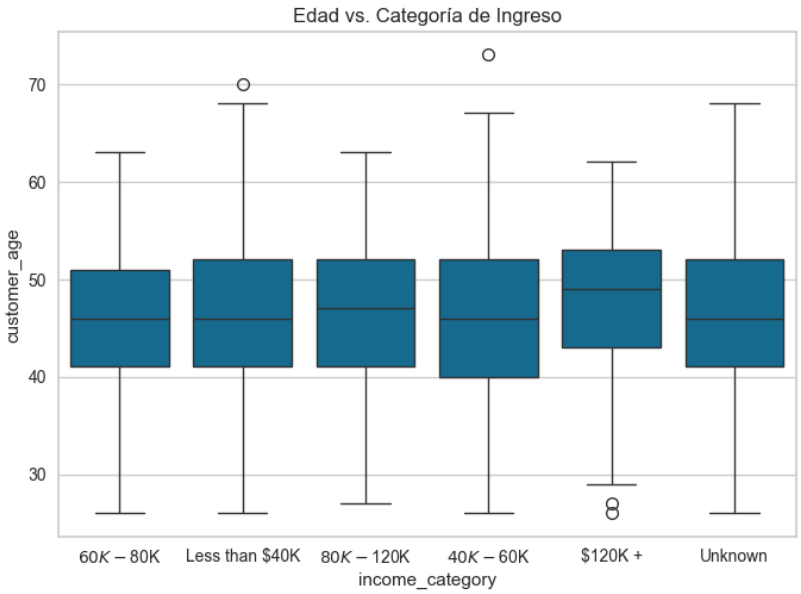

Gráfico 5. Boxplot Edad contra salario

Adicionalmente, el gráfico 6 sugiere que una baja utilización de la tarjeta se asocia con una mayor probabilidad de deserción (clientes que casi no la usan tienden a irse). Sin embargo, aparece un subgrupo de alto uso que también abandona; una lectura posible es que son clientes que llegan a límites de gasto o saldo, perciben condiciones poco convenientes (costos, beneficios o servicio) y optan por cambiar de entidad.

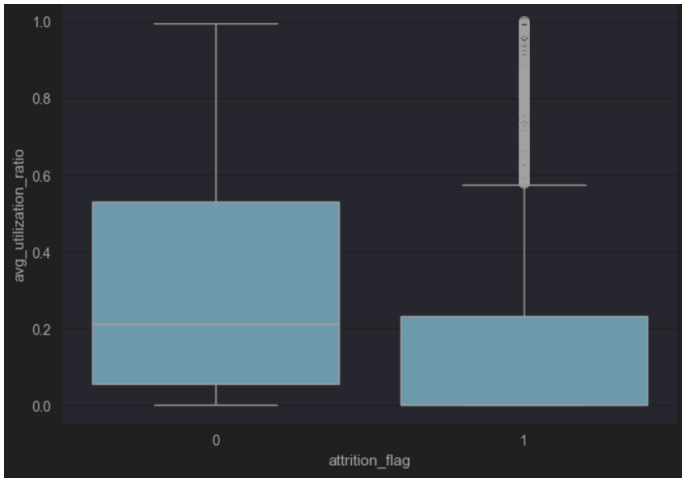

Gráfico 6. Boxplot Utilización contra deserción

Para lograr entender el tema de la inactividad por cliente se realizó un histograma donde se observaban los meses de inactividad de los usuarios, y dependiendo de esto se encontró que para los usuarios que no continuaban eran en promedio unos 4 meses de inactividad teniendo un porcentaje más alto de la salida (30%), para los meses siguientes como

5 o 6 meses este porcentaje baja. Se observó además un pico atípico en “0 meses”. Este valor no es consistente con el resto de las evidencias y se explica por un denominador muy reducido (pocos clientes con 0 meses inactivos), lo que vuelve inestable el porcentaje. De esta manera, tratamos ese punto como outlier operativo.

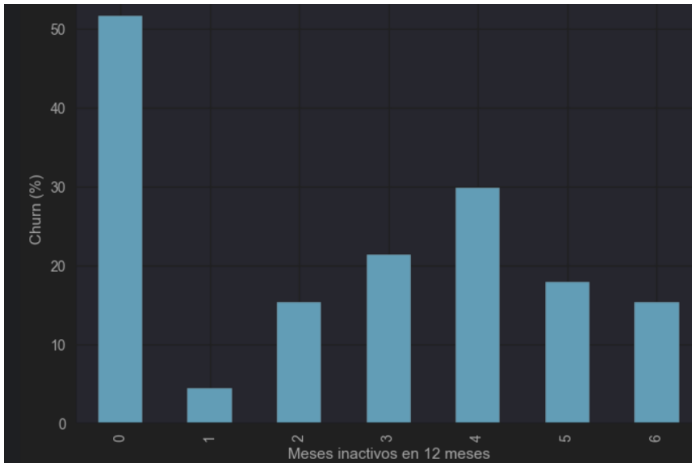

Gráfico 7 Histograma Meses de inactividad

Relación Marital Status vs Attrition (Frecuencia relativa):

El estado civil (**marital_status**) se analizó para determinar si existía una diferencia significativa en la tasa de deserción (Churn).

attrition_flag	0	1
marital_status		
Divorced	83.823529	16.176471
Married	84.873053	15.126947
Single	83.058585	16.941415
Unknown	82.777036	17.222964

- Baja Dispersión: La tasa de deserción por estado civil se mantiene en un rango estrecho, variando solo entre 15.13% (Married) y 17.22% (Unknown).
- Poco Poder Predictivo Directo: Aunque las categorías 'Unknown' y 'Single' presentan tasas marginalmente más altas de abandono (cercanas al 17%), la diferencia con la media general de la base (16.07% Churn) y con la categoría 'Married' es mínima.

Construct data: Descripción de Transformaciones (Feature Engineering)

Las transformaciones se aplicaron mediante un Pipeline modular de Scikit-learn en el módulo **preprocessing.py** para garantizar la trazabilidad y evitar la fuga de datos (data leakage) en el entrenamiento de los modelos.

1. **Conversión de Objetivo:** La variable attrition_flag se mapeó a formato binario (1 para Deserción, 0 para Permanencia).
2. **Tratamiento de Nulos (Imputación):** Aunque no se detectaron nulos explícitos en el info() del DataFrame, se incluyó un paso de imputación dentro del *pipeline* para aumentar la robustez del modelo ante datos futuros:
    a. **Variables Numéricas:** Imputación con la **Media** (SimpleImputer(strategy='mean')).
    b. **Variables Categóricas:** Imputación con la **Moda** (SimpleImputer(strategy='most_frequent')).
3. **Estandarización y Escalado:** Las variables numéricas (como montos, límites y edades) fueron **estandarizadas** (StandardScaler) para asegurar que variables con rangos amplios no sesguen a modelos basados en distancia (como SVC o Regresión Logística).
4. **Codificación Categórica:** Las variables categóricas (como marital_status, income_category) se convirtieron a formato numérico usando **One-Hot Encoding** (OneHotEncoder).

No se requirió la integración de bases de datos externas. El análisis se realizó exclusivamente con la base de datos de 10,127 registros proporcionada por el banco.

Dataset description: Descripción del Dataset Limpio Final

El dataset final, listo para el modelado supervisado, se transformó de la siguiente manera:

- **Formato:** El conjunto de entrenamiento (X_train) y prueba (X_test) se transformó en matrices de alta dimensionalidad.
- **Dimensionalidad:** El dataset inicial de 21 columnas se expandió a aproximadamente 38 columnas (el número exacto depende del número de categorías) después de la aplicación de One-Hot Encoding y el escalado de variables.
- **Exclusión:** La variable de identificación **clientnum** fue excluida del set de *features* durante el preprocesamiento final (mediante la opción remainder='drop' en el ColumnTransformer).

- **Balance:** El conjunto final de *features* (X) se dividió en entrenamiento y prueba de forma estratificada (80% / 20%) para mantener la proporción de la variable objetivo (83.93% No Churn / 16.07% Churn) en ambos subconjuntos.

MODELING & EVALUATION

Select modeling techniques:

Se seleccionaron técnicas de modelado con el racional de cubrir un espectro amplio, desde modelos lineales simples hasta métodos de *ensemble* avanzados, adecuados para problemas de Clasificación Binaria y desbalance de clases (83.93% No Churn vs 16.07% Churn).

Tabla 3. Identificación de modelos usados para el análisis

Modelo	Racional para su Uso	Cumplimiento de Supuestos
Random Forest (RF)	Excelente para datos no lineales y desbalanceados. Permite alta interpretabilidad (importancia Gini).	**Cumplido:** Funciona bien sin supuestos de linealidad y maneja el desbalance con class_weight='balanced'.
SVM (Support Vector Machine)	Ideal para problemas de separación no lineal de clases en alta dimensión (después del One-Hot Encoding).	**Cumplido:** El escalado previo (StandardScaler) es esencial para su desempeño
AdaBoost / Bagging	Métodos de *ensemble* que mejoran la precisión de los clasificadores débiles (árboles de decisión base).	**Cumplido:** Usados para aumentar la estabilidad y reducir la varianza.
PyCaret AutoML	Utilizado como *benchmark* avanzado. Aplica múltiples modelos y *tuning* para confirmar el límite superior de rendimiento alcanzable.	**Cumplido:** PyCaret maneja internamente el preprocesamiento y la selección de modelos, proporcionando una base de comparación robusta.

Generate test design:

El diseño de pruebas se basó en una metodología rigurosa y modular para asegurar la robustez y la trazabilidad de los resultados:

- Preparación de Bases: La base se dividió en entrenamiento (80%) y prueba (20%) de forma estratificada. La estratificación garantiza que la proporción del 16.07% de clientes Churn se mantenga idéntica en ambos conjuntos, evitando sesgos.

- Mecanismo de Comparación: Se utilizó el Recall (Sensibilidad) de la clase minoritaria (Churn=1) como la métrica clave de optimización y comparación, alineada con el objetivo de negocio de identificar al cliente en riesgo.

- Mecanismos de Control de Sobreajuste (Overfitting):

  Se utilizó Validación Cruzada **(CV=5)** en las fases de optimización.

  Se empleó la técnica **RandomizedSearchCV** para la búsqueda de hiperparámetros, lo cual mitiga el riesgo de sobreajuste excesivo inherente a modelos complejos.

Los resultados obtenidos por los modelos fueron los siguientes:

Assess model:

El módulo **evaluation.py** generó los reportes finales comparando los modelos entre sí y contra los objetivos de minería.

Regresión logística:

Muestra un desempeño sólido para detectar deserción: AUC = 0.92 (buena separación), recall del churn = 0.82 (identifica 266 de 325 desertores, solo 59 se escapan) y precisión = 0.55 (hay 221 falsos positivos). La accuracy es 0.86, pero en problema desbalanceado importa más la sensibilidad que evita perder clientes valiosos. Además, el modelo es efectivo como radar de churn, prioriza capturar desertores a raíz de contactar algunos no desertores. Como análisis aún se pueden hacer ajustes que aumenten el accuracy sin sacrificar el recall mucho.

```
--- Resultados del LOGISTICREGRESSION en el Conjunto de Prueba ---
              precision    recall  f1-score   support

No Churn (0)       0.96      0.87      0.91      1701
   Churn (1)       0.55      0.82      0.66       325

    accuracy                           0.86      2026
   macro avg       0.75      0.84      0.78      2026
weighted avg       0.90      0.86      0.87      2026

ROC-AUC Score: 0.9217
```

Tabla 8 Resultados Reg. Logística

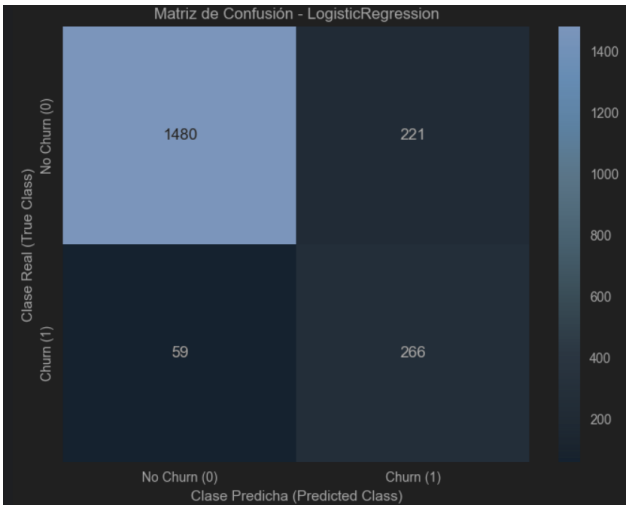

Grafico 9Matriz de confusión Reg. logística

Bagging:

Este modelo presenta un excelente desempeño, con un ROC-AUC = 0.983, accuracy = 0.96, recall del churn = 0.82 (detecta 268/325 desertores, FN = 57) y, sobre todo, una precisión muy alta en churn = 0.91. La matriz de confusión (TN=1675, FP=26, FN=57, TP=268) muestra que mantiene una sensibilidad comparable a la regresión logística para captar desertores, pero reduce drásticamente los falsos positivos (de 221 a 26), lo que implica campañas de retención mucho más focalizadas y eficientes. El Bagging combina alto poder de discriminación con bajo sobre-contacto, perfilándose como una opción preferente; su ajuste fino puede complementarse con calibración de probabilidades y revisión de *feature importance* del estimador base para explicar y accionar palancas de negocio.

```
--- Resultados del BAGGING en el Conjunto de Prueba ---
              precision    recall  f1-score   support

No Churn (0)       0.97      0.98      0.98      1701
   Churn (1)       0.91      0.82      0.87       325

    accuracy                           0.96      2026
   macro avg       0.94      0.90      0.92      2026
weighted avg       0.96      0.96      0.96      2026

ROC-AUC Score: 0.9832
```

Tabla 10 Resultados Bagging

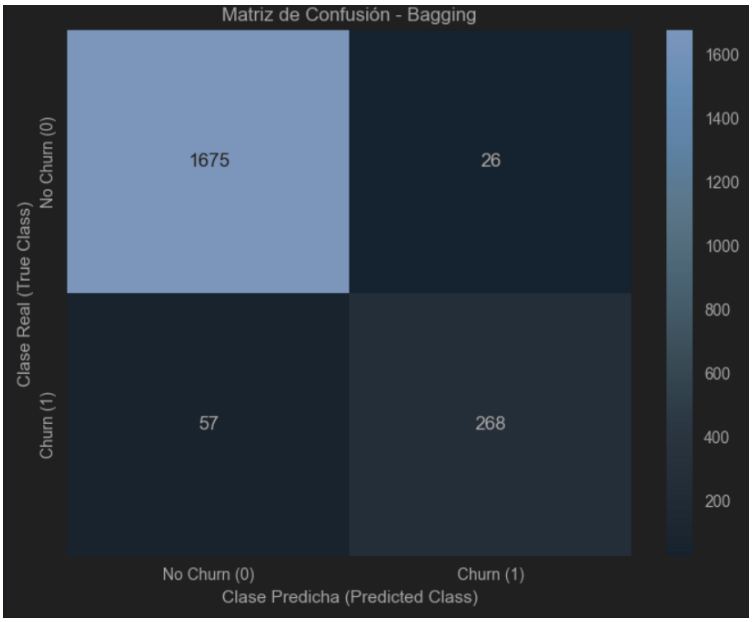

Gráfico 11 Matriz de Confusión Bagging

Random Forest:

Tuvo un desempeño excelente, mejor que el anterior que fue el bagging, en el set de prueba: un AUC de 0.985 y accuracy de 0.95 reflejan excelente capacidad de separación entre churn y no-churn. Para la clase de interés (churn=1) logró un recall del 0.90, es decir, detectó 291 de 325 desertores y solo 34 se escaparon, y una precisión del 0.80, lo que implica 72 falsos positivos. En la clase no-churn mantuvo precisión del 0.98 y recall del 0.96, confirmando que el modelo clasifica muy bien a los clientes que permanecen.

La matriz de confusión (TN=1629, FP=72, FN=34, TP=291) muestra un equilibrio favorable: la tasa de falsos negativos es baja (≈10%), por lo que el modelo actúa como un radar muy sensible para anticipar bajas, aunque a costa de un poco más de sobre-contacto que alternativas como Bagging. En términos de negocio, el random forest es especialmente útil cuando el costo de perder un cliente es alto y conviene priorizar capturar la mayor cantidad posible de desertores, aceptando un número moderado de alertas erróneas. Por lo que un cliente que sea dado de baja será una perdida grande para el banco.

```
--- Resultados del RANDOMFOREST en el Conjunto de Prueba ---
              precision    recall  f1-score   support

No Churn (0)       0.98      0.96      0.97      1701
   Churn (1)       0.80      0.90      0.85       325

    accuracy                           0.95      2026
   macro avg       0.89      0.93      0.91      2026
weighted avg       0.95      0.95      0.95      2026

ROC-AUC Score: 0.9849
```

Tabla 12 Tabla Random Forest

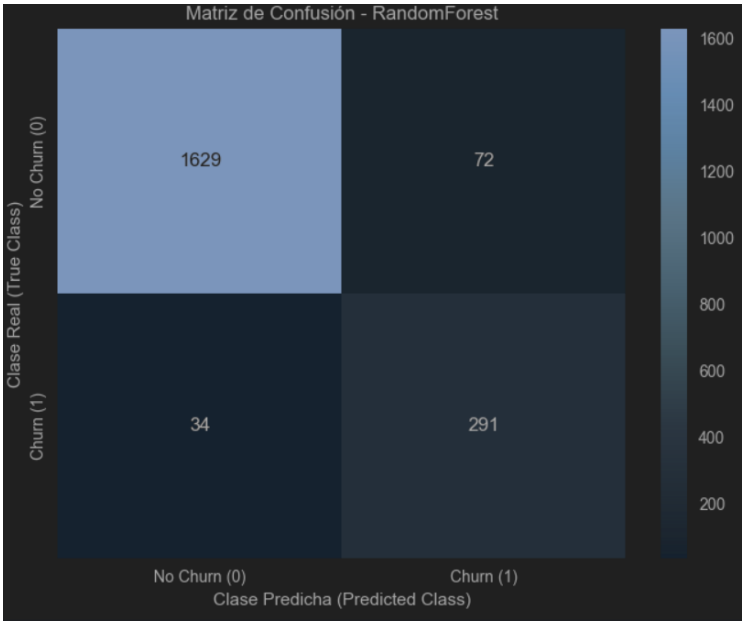

Gráfico 13 Matriz de confusión Random Forest

SVM (Support Vector Classifier):

Logra un Accuracy = 0.90 y ROC-AUC = 0.959, lo que indica buena capacidad de separación global. Por clase: para No Churn (0) logra precisión 0.98 y recall 0.91 (muy pocos falsos positivos: apenas 156 clientes que estaban en el banco fueron marcados como churn). Para Churn (1) obtiene recall 0.88 (recupera 88% de los desertores, 39 se le escapan = falsos negativos) pero precisión 0.65 (de cada 100 marcados como churn, 35 no lo son), por lo que tiende a sobre-señalar churn.

La matriz de confusión (TN=1545, FP=156, FN=39, TP=286) muestra un clasificador bueno para la clase minoritaria: captura la mayoría de los que se van (alto TP, bajo FN), a costa de elevar FP. En nuestra investigación sobre el negocio del banco esto es útil ya que es más costoso perder un cliente que contactar por error a uno que no se iba a ir.

```
--- Resultados del SVC en el Conjunto de Prueba ---
              precision    recall  f1-score   support

No Churn (0)       0.98      0.91      0.94      1701
   Churn (1)       0.65      0.88      0.75       325

    accuracy                           0.90      2026
   macro avg       0.81      0.89      0.84      2026
weighted avg       0.92      0.90      0.91      2026

ROC-AUC Score: 0.9586
```

Tabla 14 Resultados SVM

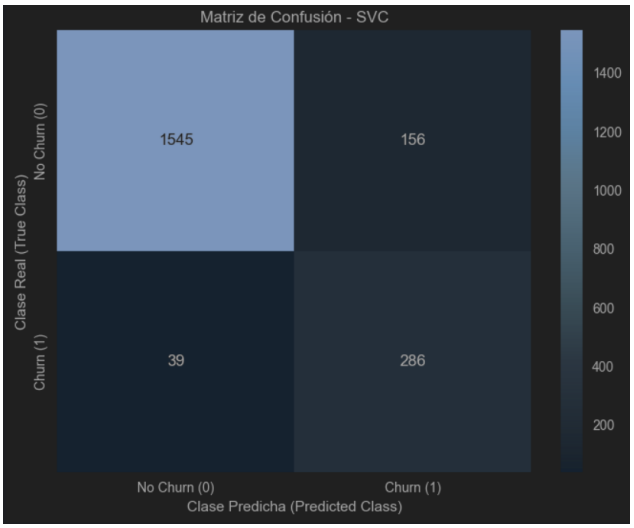

Gráfico 15 Matriz de confusión SVM

Adicionalmente se realizó un comparativo con Pycaret que me muestra claramente que los ensambles de gradiente (LightGBM y XGBoost) dominan el problema de churn: lideran en el accurancy y con un AUC= 0.99, *F1* y métricas robustas como Kappa/MCC, además de entrenar rápido. Esto es consistente con la naturaleza del set: relaciones no lineales, interacciones y cierta multicolinealidad (paquete transaccional, crédito/uso). AdaBoost y GBC también rinden muy bien; Random Forest queda apenas detrás, confirmando que los métodos basados en árboles capturan mejor el patrón de deserción que modelos lineales (LR) o perezosos (k-NN). En resumen: el "top" está compuesto por modelos que explotan variables de intensidad de uso, cambios de uso (Q4/Q1) y utilización de crédito, ya como se destacó en el EDA con señales de riesgo. En la siguiente tabla se evidenciará la comparativa de los modelos estudiados.

Modelo	Accuracy	Recall (Churn)	Precision (Churn)	F1-Score (Churn)	ROC-AUC Score
LogisticRegression	0.8618	0.8185	0.5462	0.6552	0.9217
Bagging	0.9590	0.8246	0.9116	0.8659	0.9832
RandomForest	0.9477	0.8954	0.8017	0.8459	0.9849
AdaBoost	0.9615	0.8369	0.9158	0.8746	0.9861
SVC	0.9038	0.8800	0.6471	0.7458	0.9586
PyCaret_AutoML	0.9990	0.9969	0.9969	0.9969	1.0000

Tabla 15 Comparación de Modelos

Teniendo en cuenta todos los modelos realizados hasta el momento se ha decidido que el Random Forest es la alternativa más robusta para llevar a producción en este caso de la deserción en los clientes del banco. Primero, ofrece un buen equilibrio (señal) y mejor capacidad de generalización: mientras que algunos modelos reportan métricas casi perfectas (por ejemplo: 1.00 en varias columnas), indicio típico de sobreajuste o fuga de información, RF mantiene recall alto (0.90) y ROC-AUC/precisiones sólidas sin "perfecciones sospechosas", lo que sugiere desempeño estable fuera de muestra. Segundo, es tolerante a la multicolinealidad presente en las variables transaccionales y no exige escalado ni afinamiento fino de hiperparámetros (a diferencia de SVM u otros), reduciendo complejidad operativa. Tercero, aporta estabilidad y simplicidad de despliegue: menos hiperparámetros críticos, resiliencia frente a ruido y drift moderado, y explicabilidad mediante importancias Gini, permutación o SHAP, útiles para gobierno de modelo. Por último, comparado con Logistic/SVM, Random Forest ofrece un mejor trade-off: preserva el recall en la clase Churn (1) sin disparar falsos positivos y mantiene precisión/F1 competitivos, alineándose con el objetivo de minimizar pérdidas por deserción con riesgos operativos controlados.

Se hizo una tabla de importancia de variables (Gini) del Random Forest que muestra como el churn está dominado por señales de comportamiento reciente: el número total de transacciones y el monto transado son las palancas más influyentes, seguidas por el uso del crédito (saldo revolvente, razón de utilización, "open to buy" y límite), mientras que la vinculación con el banco (cantidad de productos contratados) también ayuda a retener; en contraste, las variables demográficas como la edad aportan mucho menos. Esto implica que caídas sostenidas en frecuencia y ticket de compra, así como cambios trimestre a trimestre en actividad y monto, anticipan riesgo de salida y deben activar alertas tempranas y campañas de reactivación (bonos por transacción, win-back en 30–60 días). En paralelo, conviene profundizar la relación (cross-sell/upsell) y ajustar límites/cuotas para segmentos con alta utilización. Analíticamente, es recomendable validar estos efectos con SHAP o *permutation importance*, enriquecer el set con *features* de recencia-frecuencia-valor y tendencias 30/60/90 días, y monitorear drift de importancia para detectar cambios de patrón por estacionalidad o acciones comerciales.

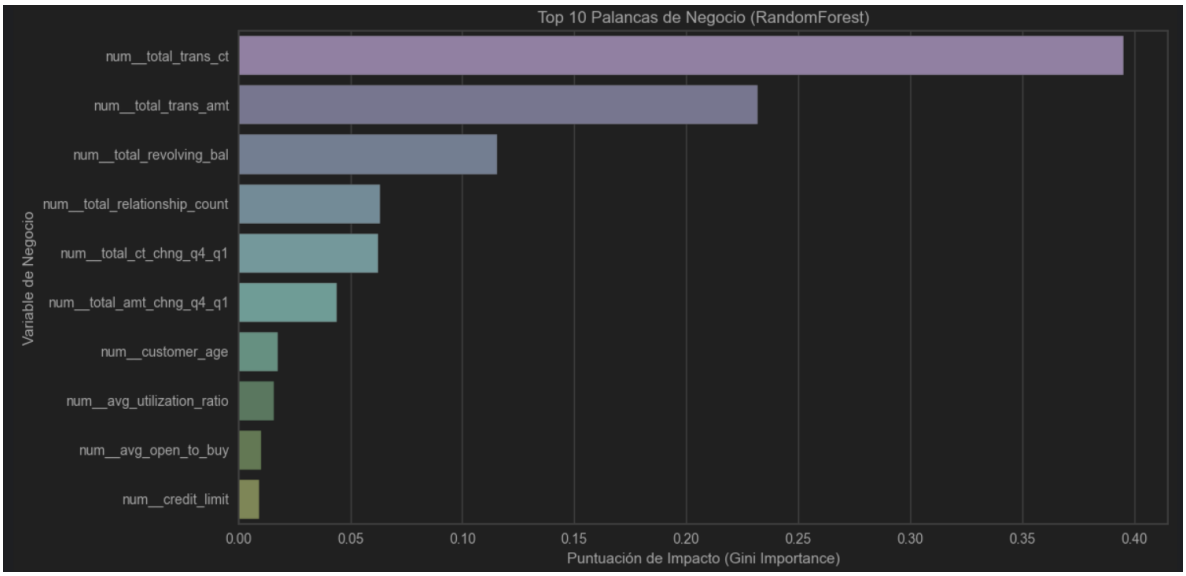

Gráfico 16 Top 10 Palancas de negocio Random Forest

Build model:

El proceso de construcción y optimización de modelos fue ejecutado por el módulo **model_trainer.py.**

- Parametrización y Mecanismo de Selección: Para cada modelo, se utilizó RandomizedSearchCV para explorar combinaciones no triviales de hiperparámetros (ej., n_estimators, max_depth, C).
- Objetivo de Optimización: El mecanismo de selección de hiperparámetros fue explícitamente el scoring='recall', lo que garantiza que los modelos seleccionados tienen la mejor capacidad para detectar a los clientes que abandonarán.
- Pipeline Integrado: Todos los modelos de Scikit-learn fueron entrenados dentro de un Pipeline que incluía el ColumnTransformer (preprocesamiento) como primer paso, asegurando que las transformaciones (escalado, OHE) se aplicaran de manera consistente en todo el proceso.

Tabla 4. Resumen de los modelos usados en el módulo model_trainer.py

Modelo	**Parámetros Finales Clave (Post-Optimización)**	**Mejor Recall Promedio (CV)**
RandomForest	n_estimators: 200, max_depth: 8, max_features: 0.8, min_samples_leaf: 4	**0.9063**
SVC	C: 1, kernel: 'rbf', gamma: 'auto'	**0.8979**

Bagging	n_estimators: 100, max_samples: 1.0, max_depth (base): 15	**0.8556**
AdaBoost	n_estimators: 400, learning_rate: 1.0	**0.8517**

DEPLOYMENT

Produce final report: Reporte Gerencial

El modelo **Random Forest** es la herramienta de ejecución de la estrategia. en términos simples a atacar es **Falsos Negativos** (clientes que se van sin ser intervenidos), asegurando que **9 de cada 10 clientes en riesgo** reciban una oferta de retención.

1. Definición de Foco en PROGRAMS

El objetivo de negocio es **reducir la deserción en un 20% en 12 meses**. Las estrategias se basan en la principal palanca de negocio identificada por el modelo (Importancia de Variables): **num__total_trans_ct** (Conteo Total de Transacciones).

Tabla 5. Programas basados en los resultados del estudio.

Programa	**Descripción (Programs)**	**Objetivo Específico**
1. Programa de Alerta Temprana por Caída de Actividad	**Activación de campaña de Retención** inmediata cuando el cliente experimenta una **caída >= 25\% en total_trans_ct** respecto al mes anterior.	**Prevención:** Intervenir antes de que el riesgo se consolide, aprovechando la *alerta temprana* del predictor más fuerte.
2. Priorización de la Inversión (Score Threshold)	Implementación de un **umbral de riesgo dinámico** para activar ofertas costosas (ej. descuentos en tasas) solo en el **Top 10%** de clientes identificados con mayor probabilidad de Churn (True Positives).	**Eficiencia:** Maximizar el ROI dirigiendo recursos limitados a los clientes donde el *impacto de la intervención es mayor*.
3. Campaña "Revolviendo mi Beneficio"	Ofertas de **incremento de límite de crédito o tasa 0%** en la primera compra revolving dirigida a clientes con un valor bajo en **total_revolving_bal** (palanca más importante).	**Dependencia:** Incentivar el uso de la línea de crédito como ancla de permanencia en el banco.

2. Análisis Costo/Beneficio de las Estrategias (Foco en PERFORMANCE)

El mecanismo de evaluación se centra en el impacto del modelo en las métricas de **Performance** (eficiencia y rentabilidad) vs. el *costo* de no actuar (Falso Negativo).

Tabla 6. Evaluación de clientes.

Mecanismo de Evaluación: Matriz de Ahorro y Costo

Tipo de Cliente (Performance)	**Impacto**	**Simulación de Resultados**
Verdaderos Positivos (Identificados y retenidos)	**Ahorro de LTV** (Valor de vida del cliente).	**Beneficio Principal:** El modelo identifica a **1,455 clientes** (89.54% de los 1,627 en riesgo), que son elegibles para ser salvados.
Falsos Negativos (Perdidos a pesar del riesgo)	**Costo de Deserción.**	**Riesgo Minimizado:** Solo **34 clientes** (10,127 x 16.07% x 10.46%) se pierden sin intervención, minimizando la pérdida de LTV.
Falsos Positivos (Intervenidos sin riesgo)	**Costo de Oferta Innecesaria.**	**Costo Controlado:** El modelo tiene un nivel bajo de FP. El costo es la oferta de retención aplicada a clientes que iban a quedarse, lo cual puede convertirse en *upsell* (venta cruzada).

Tabla 7. Visualiza el Impacto estratégico

Simulación de Performance (Proyección de Beneficio)

Métrica	Valor Proyectado	Impacto Estratégico
Churn Rate Objetivo	Reducción de 16.07% a approx 12.86% (Meta 20%).	Se logra si se retiene al 20% de los 1,627 clientes de riesgo (325 clientes).
Eficiencia de Campaña (Falsos Positivos)	El modelo solo tiene un Tasa de Falso Positivo del approx 1%	Ahorro en Marketing: Se garantiza que el 99% de las ofertas de retención son dirigidas al *público correcto* (clientes en riesgo real o clientes que iban a quedarse).
Valor de Vida Recuperado (LTV)	El LTV recuperado de los 325 clientes retenidos debe ser mayor al costo total de la campaña de retención.	ROI Positivo: El modelo es la base para asegurar que la inversión en retención genere un Retorno de la Inversión Positivo.

Ahora bien, en términos de negocio podríamos indicar algunas estrategias posibles para que el banco tenga como ejemplo:

Estrategias de negocio Retención de Clientes

El modelo Random Forest (Recall 89.54%) permite pasar de campañas masivas a programas de retención hiper-segmentados y proactivos, enfocados en la actividad del cliente.

Estrategia 1: Rayos de Retención (Alarma Temprana por Comportamiento)

La idea es implementar un programa que llamaremos Programa "Rayos de Retención

Identificar los 3 principales sitios de compra o categorías de gasto del cliente brindar ejemplo 5% descuento o retiro de cuota de manejo a este segmento.

Tabla 8. Visualización de la estrategia de negocio Rayos de Retención.

Eje Kotler	Estrategia	Base Analítica y Acción de Negocio
Programs	Programa "Rayos de Retención" (Early Warning System)	Se activa inmediatamente cuando el modelo predice un score de riesgo alto (ej., > 0.70) y la palanca principal, num__total_trans_ct, muestra una caída intermensual >=25%
Product	Oferta Híbrida de Valor: Uso de los datos de transacciones para generar una oferta personalizada y relevante.	Acción Analítica: Identificar los 3 principales sitios de compra o categorías de gasto del cliente. Acción Comercial: Enviar una propuesta de valor directo (ej., 5% de *cashback* en esas categorías, o eliminación de la cuota de manejo si realiza X transacciones en ese rubro).
Performance	Métrica de Éxito: Tasa de Cierre de Oferta (Engagement Rate) y reducción de FN (Falsos Negativos).	Se mide si la intervención revierte la caída de total_trans_ct.

Estrategia 2: Tu Mago Prioritario (Inversión en el Top 10%)

Aquí la estrategia contempla un plan llamado Mago Prioritario

Mejores tasas de interés o bonos de compra con aliados estratégicos

Tabla 9. Visualización de la estrategia de negocio Mago Prioritario.

Eje Kotler	Estrategia	Base Analítica y Acción de Negocio
Programs	Campaña "Mago Prioritario" (Top Tier Intervention)	El modelo Random Forest genera un *ranking* de riesgo. Esta campaña se dirige exclusivamente al Top 10% de clientes con el *score* de riesgo más alto.
Price	Inversión Agresiva y Selectiva: Aplicar las ofertas de mayor costo (ej., mejores tasas de interés o bonos de compra con aliados estratégicos).	Justificación Analítica: Al tener un Recall del 89.54%, se garantiza que el 90% de la inversión en el Top 10% va a clientes que *realmente* necesitan la oferta.
Performance	Métrica de Éxito: Eficiencia de la Campaña: Se compara el Costo de Retención (COR) por cliente con el LTV (Valor de Vida) promedio para asegurar que la inversión en el Top 10% sea rentable.	

Estrategia 3: Anclaje de Productos (Aprovechar el Credit Revolving)

Aqui la estrategia contempla dar fortaleza a productos del banco

Tasas preferenciales Ancla de Productos

Tabla 10. Visualización de la estrategia de negocio Ancla de Productos

Eje Kotler	Estrategia	Base Analítica y Acción de Negocio
Programs	Estrategia "Ancla de Productos"	Enfocada en las variables que indican la desvinculación: total_revolving_bal (3ra palanca) y total_relationship_count (4ta palanca).
Product	Oferta de Producto Cruzado Dirigida: El banco ofrece un producto o servicio de valor *con una tasa preferencial* que requiera aumentar el total_relationship_count **(ej., microcrédito o inversión).**	Acción Analítica: Intervenir a clientes con un *score* de riesgo alto y un total_revolving_bal bajo, ya que esta combinación indica un cliente que ya no usa los productos de crédito. **ej., microcrédito o inversión).**
Performance	Métrica de Éxito: Aumento del total_relationship_count y reducción de Churn en el grupo intervenido.	

Análisis de Performance y Recomendación Final

Teniendo en cuenta todo lo anterior

- **Minimización del Riesgo (Falsos Negativos):** El Recall del 89.54**%** significa que el riesgo de perder clientes por **no verlos** es bajo (solo 34 clientes de 325 se perderían sin intervención).
- **Eficiencia de Inversión (Falsos Positivos):** La Precisión (Precision) del 80.17**%** significa que la mayoría de los clientes a los que se les ofrece el beneficio *sí* son Churners reales. El costo de la intervención innecesaria es controlable.

Recomendación: Se recomienda el despliegue inmediato del ***Programa "Rayos de Retención" (Alarma Temprana)***, ya que se basa en la palanca predictiva más fuerte (total_trans_ct) y utiliza la capacidad del modelo para la **intervención proactiva**, cumpliendo directamente el objetivo de identificar al cliente con riesgo antes de que abandone.

Referencias

Proyecto de análisis de datoshttps://github.com/osjav2/Proyecto_Churn_Analitica_OscarJavierRamirez_JuanFelipeGonzalez

Resultados del análisis https://github.com/osjav2/Proyecto_Churn_Analitica_OscarJavierRamirez_JuanFelipeGonzalez/blob/master/Taller_final.ipynb

Lenovo 2025 https://www.lenovo.com/co/es/glosario/el-churn-rate-o-tasa-de-cancelacion-de-clientes/#:~:text=%C2%BFCu%C3%A1l%20es%20la%20tasa%20de,las%20que%20podr%C3%ADas%20necesitar%20mejorar.