

Reporte Final de Proyecto de Ciencia de Datos

Predicción de Deserción de Clientes Bancarios (*Churn*)

Gabriela Mariel Lopez Armenta

Curso: Introducción a la Ciencia de Datos

Noviembre 2025

1. Resumen Ejecutivo

Este proyecto desarrolló y evaluó modelos de *Machine Learning* para la predicción de *Churn* en el sector bancario. De tres modelos probados, el **XGBoost Classifier** fue seleccionado como el modelo ganador, alcanzando un **Recall de 0.55** y una *Accuracy* de 0.87. Los resultados confirman que la **Edad** es el factor individual más crítico, seguido por variables de compromiso financiero. Se concluye que el modelo cumple con los criterios de éxito y está listo para integrarse en estrategias de retención proactivas.

2. 1. Introducción y Planteamiento del Problema

La deserción de clientes (*Churn*) representa una pérdida significativa de ingresos futuros. El objetivo es proporcionar al banco una herramienta predictiva que permita la intervención temprana.

2.1. 1.1. Pregunta de Investigación (Q.I.)

¿Cuáles son los factores demográficos, financieros y de compromiso que tienen el mayor impacto en la probabilidad de deserción de un cliente bancario, y cuál es el modelo de *Machine Learning* más eficiente (medido por *AUC* y *Recall*) para identificar a los clientes de alto riesgo?

2.2. 1.2. Hipótesis

La deserción de clientes está impulsada principalmente por un **bajo nivel de compromiso financiero** con la entidad. Específicamente, se postula que la combinación de un Saldo bajo, una baja Puntuación Crediticia y un número reducido de productos será el predictor más robusto de *Churn*.

3. 2. Metodología y Análisis de Datos

El proyecto utilizó un *dataset* público de 10,000 clientes bancarios. La metodología se centró en la **Clasificación Binaria** debido al desbalance de clases (79.6 % No-Churn vs. 20.4 % Churn).

3.1. 2.1. Preprocesamiento Clave

- **Limpieza:** Eliminación de identificadores (*CustomerID*, *Surname*).
- **Codificación:** Variables categóricas (*Geography*, *Gender*) transformadas mediante *One-Hot Encoding*.
- **Escalado:** Variables numéricas escaladas con *StandardScaler* para estandarizar su impacto.

4. 3. Resultados del Modelado y Evaluación

Se comparó el rendimiento de la Regresión Logística (modelo base), Random Forest y XGBoost.

4.1. 3.1. Criterios de Éxito

El éxito se definió por un **Recall** para la clase *Churn* de al menos **0.50** y un **AUC** superior a 0,85.

Cuadro 1: Tabla Comparativa de Métricas de Rendimiento (Clase 1: Churn)

Modelo	Accuracy	Recall	Precision	AUC Score
Regresión Logística	0.81	0.20	0.55	0.7789
Random Forest	0.87	0.47	0.76	0.8653
XGBoost (Ganador)	0.87	0.55	0.72	0.8502

4.2. 3.2. Conclusión del Modelo Ganador

El modelo **XGBoost** es seleccionado por ser el **único** en superar el criterio de **Recall** con un valor de ****0,55****. Esto significa que es capaz de identificar correctamente al 55 % de los clientes que realmente van a desertar, cumpliendo el objetivo primordial del proyecto.

5. 4. Hallazgos y Validación de Hipótesis

5.1. 4.1. Factores de Riesgo (Feature Importance)

El análisis de importancia de características del modelo XGBoost identificó los siguientes factores como los más influyentes:

- **Dominante: Edad** (El factor individual con mayor peso).

- **Secundarios:** Balance, NumOfProducts y CreditScore.

5.2. 4.2. Validación de la Hipótesis

La hipótesis es **parcialmente validada**. Si bien los factores de compromiso (Balance y NumOfProducts) son críticos, la **Edad** se reveló como el factor de riesgo más importante, superando a la combinación de variables financieras. Esto sugiere que las estrategias de retención deben considerar fuertemente el ciclo de vida del cliente.

6. 5. Conclusiones y Recomendaciones de Negocio

El proyecto ha resultado en un modelo de alta calidad, listo para la implementación.

1. **Implementación:** Se recomienda integrar el modelo **XGBoost** para la puntuación diaria de riesgo de *churn*.
2. **Segmentación de Retención:** Las campañas deben enfocarse proactivamente en clientes de **Edad avanzada** que también presenten **bajo Saldo** y **pocos Productos**, ya que este es el segmento de mayor riesgo.