Predicción de churn en clientes de telecomunicaciones

Ahmed Darwiche
Tomás Moreno
Nicolas Muraro
Marcelo Valdivia

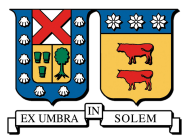
MAT281
Universidad Técnica Federico Santa María

Noviembre 2025

Contenido

Contexto del problema

- En telecomunicaciones, retener clientes suele ser más barato que adquirir nuevos.
- Cuando un cliente hace churn (abandona el servicio), se pierde ingreso futuro y se deben invertir recursos para reemplazarlo.
- Un modelo de predicción de churn permite:
  - Identificar clientes en riesgo.
  - Focalizar campañas de retención.
  - Usar mejor el presupuesto comercial.

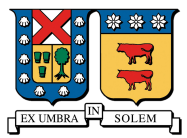


Objetivo del proyecto

Problema a resolver

Construir un modelo de clasificación binaria que estime la probabilidad de que un cliente haga churn.

- Dataset: Telco Customer Churn (Kaggle).
- Cada fila representa un cliente, con información sobre:
  - Datos demográficos.
  - Servicios contratados (internet, seguridad, soporte, streaming).
  - Tipo de contrato y método de pago.
  - Cargos mensuales y totales.
- Variable objetivo: `Churn` (Yes / No).



Metodología general

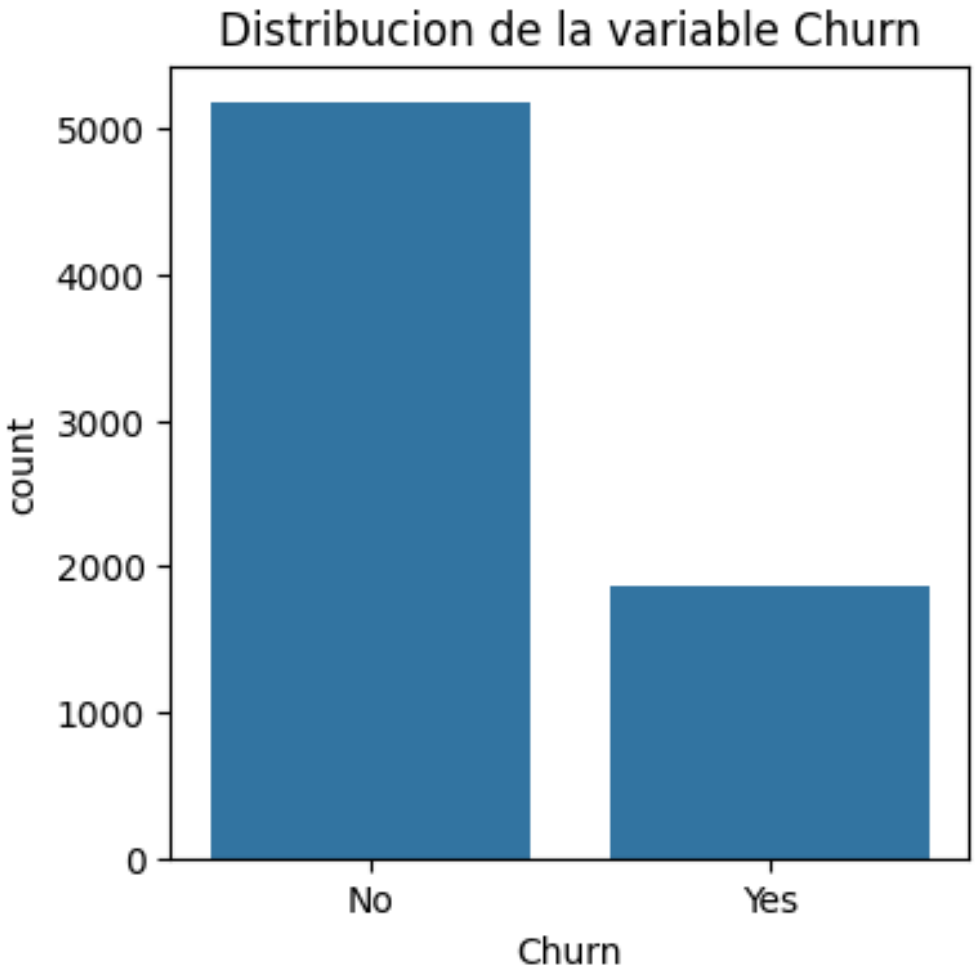
1. Definición del problema y objetivo analítico.
2. Análisis exploratorio de datos y visualizaciones.
3. Preprocesamiento:
   - Limpieza, transformaciones y codificación.
   - Manejo del desbalance (SMOTE).
4. Entrenamiento y comparación de modelos.
5. Evaluación con métricas de clasificación.
6. Interpretación de variables y conclusiones para el negocio.

Descripción del dataset

- Aproximadamente 7043 clientes en el dataset original.
- Tras limpiar `TotalCharges`, quedan alrededor de 7032 registros.
- Variables principales:
  - Numéricas: `tenure`, `MonthlyCharges`, `TotalCharges`, `SeniorCitizen`.
  - Categóricas: `Contract`, `InternetService`, `PaymentMethod`, servicios adicionales, etc.
- Variable objetivo:
  - `Churn` (Yes / No) y versión binaria `Churn_binaria` (1 / 0).

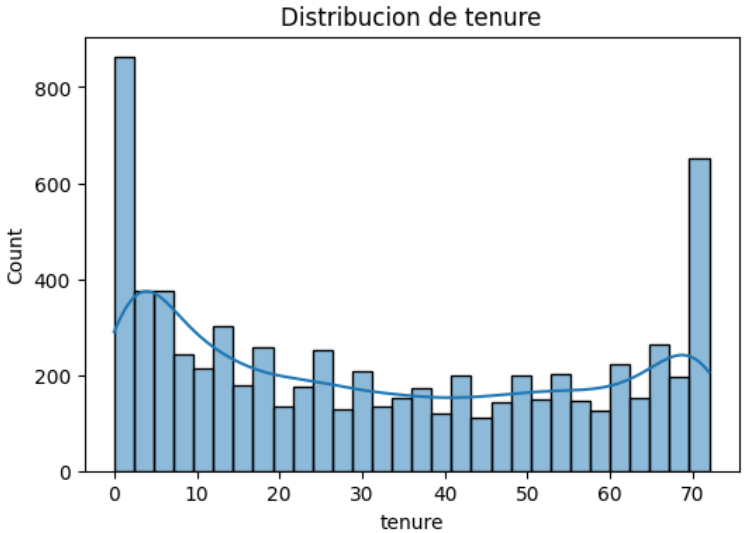
Distribución de la variable objetivo

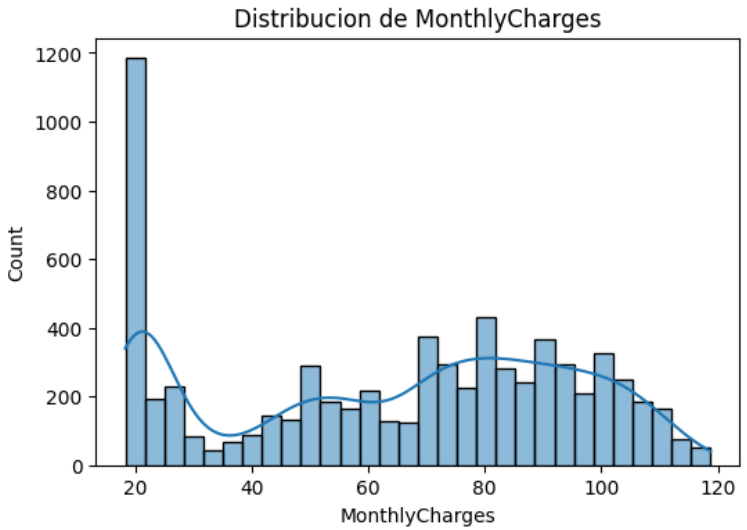
- Proporciones aproximadas:
  - 73,5 % de clientes que no hacen churn.
  - 26,5 % de clientes que sí hacen churn.
- Existe un desbalance moderado de clases.



Distribucion de variables numericas

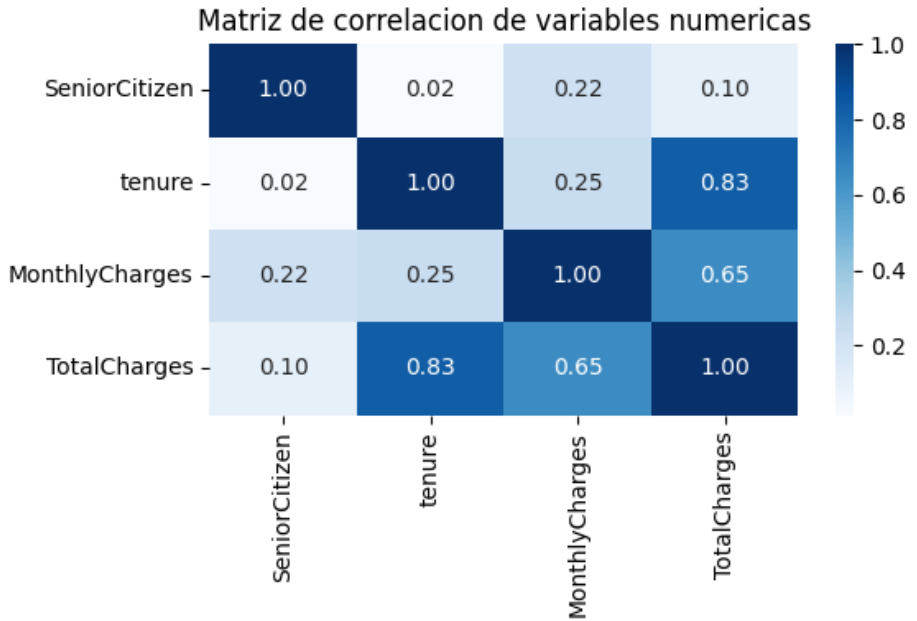
- Histograma de tenure: muestra la distribucion de meses como cliente.
- Histograma de MonthlyCharges: muestra la distribucion de los cargos mensuales.





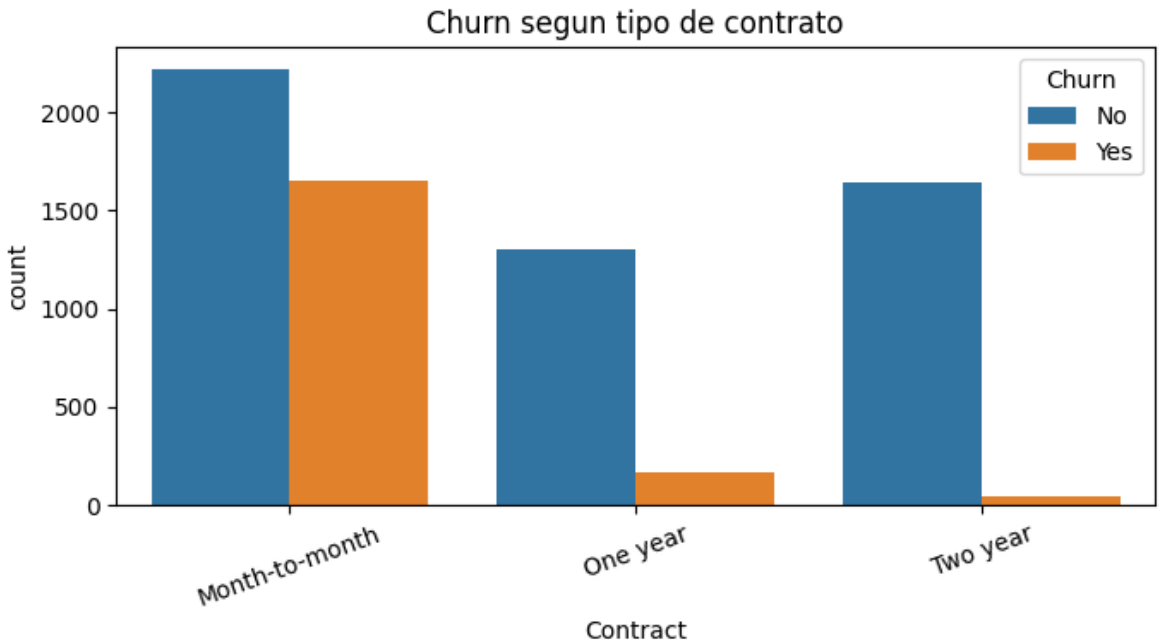
Correlaciones entre variables numéricas

- Se analizó la matriz de correlación entre:
  - SeniorCitizen, tenure, MonthlyCharges y TotalCharges.
- Se observa:
  - Relación moderada entre MonthlyCharges y TotalCharges.
  - Relación moderada entre tenure y TotalCharges.
  - No se ven correlaciones excesivamente altas.



Churn según tipo de contrato

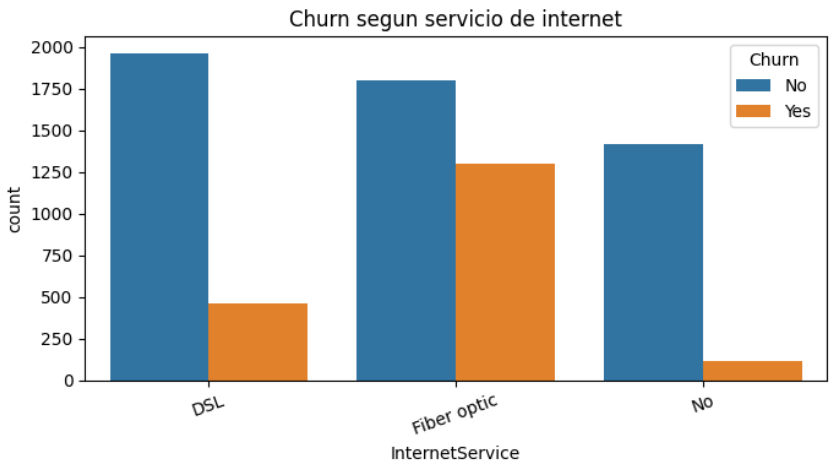
- Variable Contract:
  - Month-to-month.
  - One year.
  - Two year.
- Los clientes con contrato mensual muestran tasas de churn más altas.



Churn según internet y método de pago

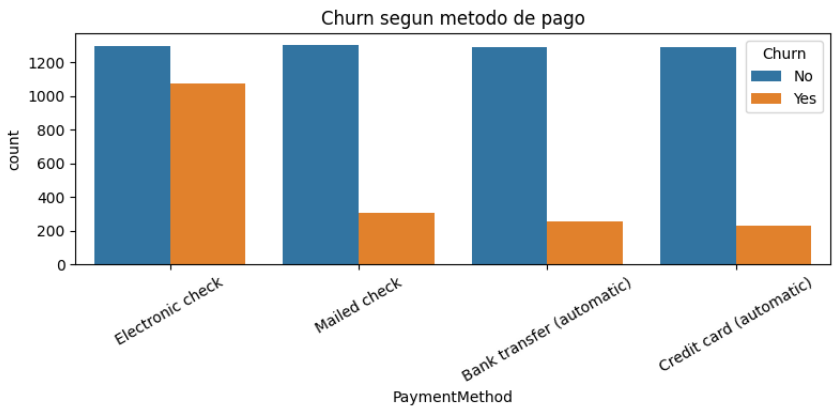
InternetService

- Clientes con Fiber optic presentan mayor churn.



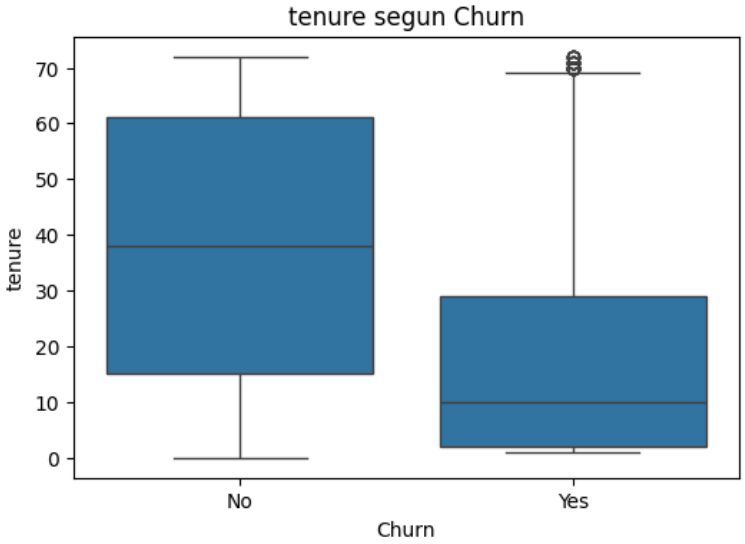
PaymentMethod

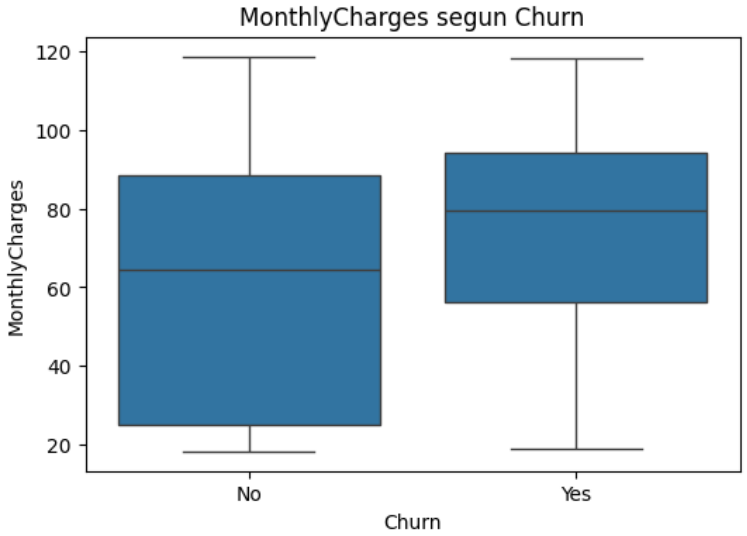
- El método Electronic check se asocia a un churn más alto que otros métodos.



Variables numéricas según churn

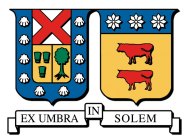
- Tenure:
  - Los clientes que hacen churn suelen tener menor tiempo de permanencia.
- MonthlyCharges:
  - Hay diferencias en la distribución de cargos mensuales entre clientes con y sin churn.





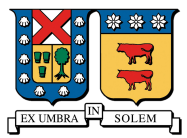
Limpieza y transformaciones

- TotalCharges:
  - Originalmente almacenado como texto, con algunos valores vacíos.
  - Se convirtió a numérico; se eliminaron pocas filas con valores faltantes.
- customerID:
  - Identificador sin valor predictivo directo.
  - Se eliminó del dataset de modelado.
- Variable objetivo:
  - Se creó Churn_binaria (0 para No, 1 para Yes).



Codificación y escalado

- Separación de columnas:
  - Numéricas: tenure, MonthlyCharges, TotalCharges, etc.
  - Categóricas: Contract, InternetService, PaymentMethod, servicios adicionales.
- Se utilizó un ColumnTransformer con:
  - StandardScaler para las variables numéricas.
  - OneHotEncoder para las variables categóricas.
- División en conjuntos:
  - 80 % entrenamiento, 20 % prueba.
  - División estratificada según la variable objetivo.

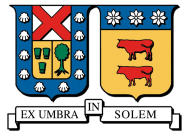


Desbalance y uso de SMOTE

- La clase churn (1) representa alrededor del 26,5 % de los registros.
- Para reducir el efecto del desbalance:
  - Se aplicó SMOTE en el conjunto de entrenamiento.
  - SMOTE genera ejemplos sintéticos de la clase minoritaria.
- SMOTE se integró en los pipelines junto con el preprocesamiento y los modelos.

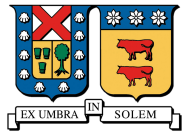
Modelos considerados

- Se construyeron pipelines con:
  - Preprocesamiento (escalado + one-hot).
  - SMOTE para balancear la clase minoritaria.
  - Clasificador supervisado.
- Modelos entrenados:
  1. Regresión logística.
  2. Random Forest.
  3. XGBoost.
  4. K-Nearest Neighbors (KNN).
- Se utilizaron GridSearchCV o RandomizedSearchCV para ajustar hiperparámetros.



Métricas de evaluación

- Accuracy: proporción de predicciones correctas.
- Precision: de los clientes predichos como churn, cuántos realmente lo son.
- Recall: de los clientes que hacen churn, cuántos detecta el modelo.
- F1-score: promedio armónico entre precision y recall.
- ROC-AUC: calidad de separación entre clases a distintos umbrales.
- En churn interesa especialmente tener buen recall en la clase positiva, sin perder demasiada precision.

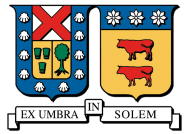


Resultados de los modelos

- Resultados aproximados en el conjunto de prueba:

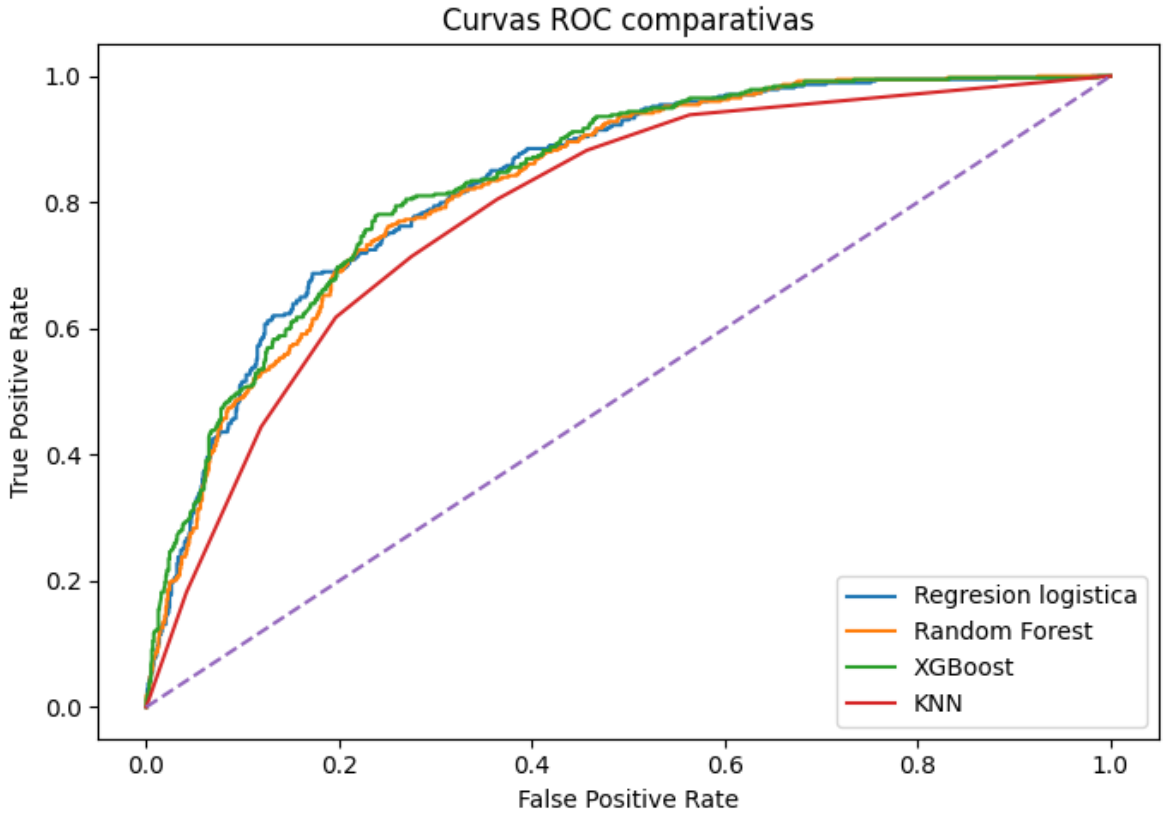
Modelo	F1	ROC-AUC
Regresión logística	$\approx 0{,}61$	$\approx 0{,}83$
Random Forest	$\approx 0{,}62$	$\approx 0{,}83$
XGBoost	$\approx 0{,}63$	$\approx 0{,}84$
KNN	$\approx 0{,}58$	$\approx 0{,}79$

- XGBoost obtiene el mejor F1 y el mayor ROC-AUC.
- Random Forest y regresión logística tienen desempeño cercano.



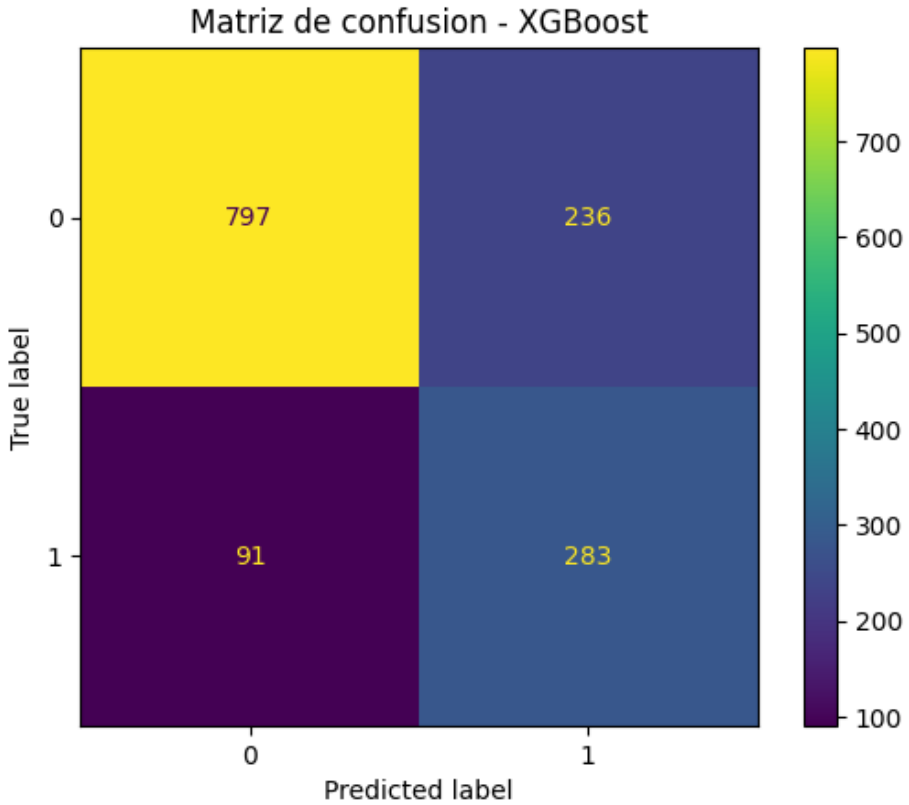
Curvas ROC comparativas

- Se graficaron las curvas ROC de los cuatro modelos.
- XGBoost alcanza el área bajo la curva más alta, seguido de cerca por Random Forest y regresión logística.



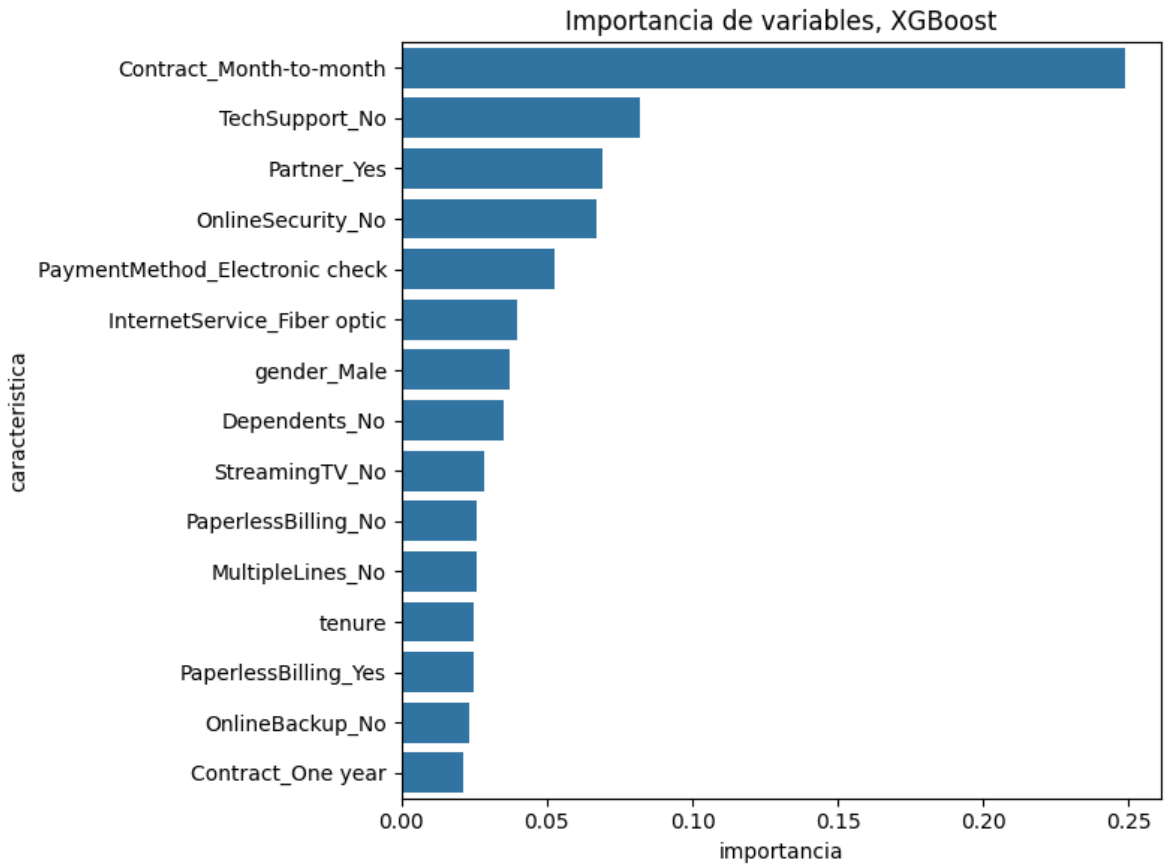
Matriz de confusión del mejor modelo

- Se toma XGBoost como modelo principal para evaluar el comportamiento por clase.
- La matriz de confusión muestra:
  - Buen número de verdaderos negativos (no churn correctamente clasificados).
  - Proporción razonable de verdaderos positivos (churn detectado).
  - Todavía hay falsos positivos y falsos negativos, lo que deja espacio para mejoras.



Importancia de variables según XGBoost

- Se analizó la importancia de variables del modelo XGBoost.
- Entre las variables con mayor peso aparecen:
  - Contract_Month-to-month.
  - OnlineSecurity_No y TechSupport_No.
  - PaymentMethod_Electronic check.
  - Contract_Two year y Contract_One year.
  - Tenure, MonthlyCharges y TotalCharges.
  - InternetService_Fiber optic y algunos servicios adicionales.

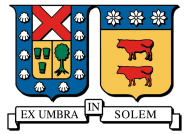


Lectura de los factores clave

- Contratos Month-to-month:
  - Mayor flexibilidad para dejar el servicio, asociado a mayor churn.
- Falta de servicios de seguridad y soporte:
  - Clientes sin OnlineSecurity ni TechSupport tienden a abandonar más.
- Método Electronic check:
  - Se relaciona con una tasa de churn más alta que otros métodos de pago.
- Permanencia y cargos:
  - Clientes con poco tenure y ciertos niveles de MonthlyCharges presentan mayor riesgo de abandono.

Conclusiones principales

- El dataset presenta un churn cercano al 26,5 %, con desbalance moderado.
- XGBoost fue el modelo con mejor equilibrio entre F1 y ROC-AUC:
  - F1 en torno a 0,63.
  - ROC-AUC alrededor de 0,84.
- Los factores más asociados al churn son:
  - Tipo de contrato.
  - Servicios de seguridad y soporte.
  - Método de pago.
  - Permanencia y nivel de cargos.



Recomendaciones para el negocio

- Clientes con contrato Month-to-month:
  - Diseñar campañas para migrarlos a contratos de uno o dos años.
  - Ofrecer beneficios especiales en los primeros meses.
- Servicios de valor agregado:
  - Promover paquetes que incluyan OnlineSecurity y TechSupport a clientes que no los tienen.
- Método Electronic check:
  - Revisar la experiencia de este método y considerar incentivos para otros medios de pago.