Análisis y Predicción de Deserción de Clientes (Churn) en Telecomunicaciones

Proyecto de Fin de Semestre - Asignatura: Aplicaciones de la matemática en ingeniería

Antonio Aguilar, Nicolás Lara, Diego Sandoval,
Rodrigo Montecinos, Cristopher Martínez

30 de noviembre de 2025

Universidad Técnica Federico Santa María
Departamento de Matemática

Definición del Problema

Deserción de Clientes (Churn)

La alta rotación de clientes genera un problema estratégico para las empresas de telecomunicaciones:

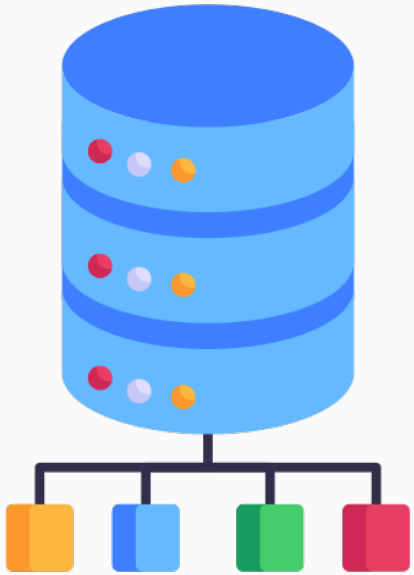
- **Costo Alto:** Retener es entre $\times 5$ y $\times 7$ más barato que adquirir un nuevo cliente.
- **Impacto Directo:** El churn reduce la rentabilidad y el valor de vida del cliente.
- **Necesidad Clave:** Anticipar el abandono permite focalizar acciones de retención.

Definición Técnica del Problema

Naturaleza del Problema

El churn es una **variable binaria**:

Churn $= 1$ (abandona) Churn $= 0$ (permanece)



El dataset *Telco Customer Churn* incluye:

- **7.043 clientes** y **21 variables**.
- **Variables mixtas**: numéricas y categóricas.
- **Desbalanceo de clases**: mayoría de clientes permanecen.

Objetivo General del Proyecto

Propósito Central

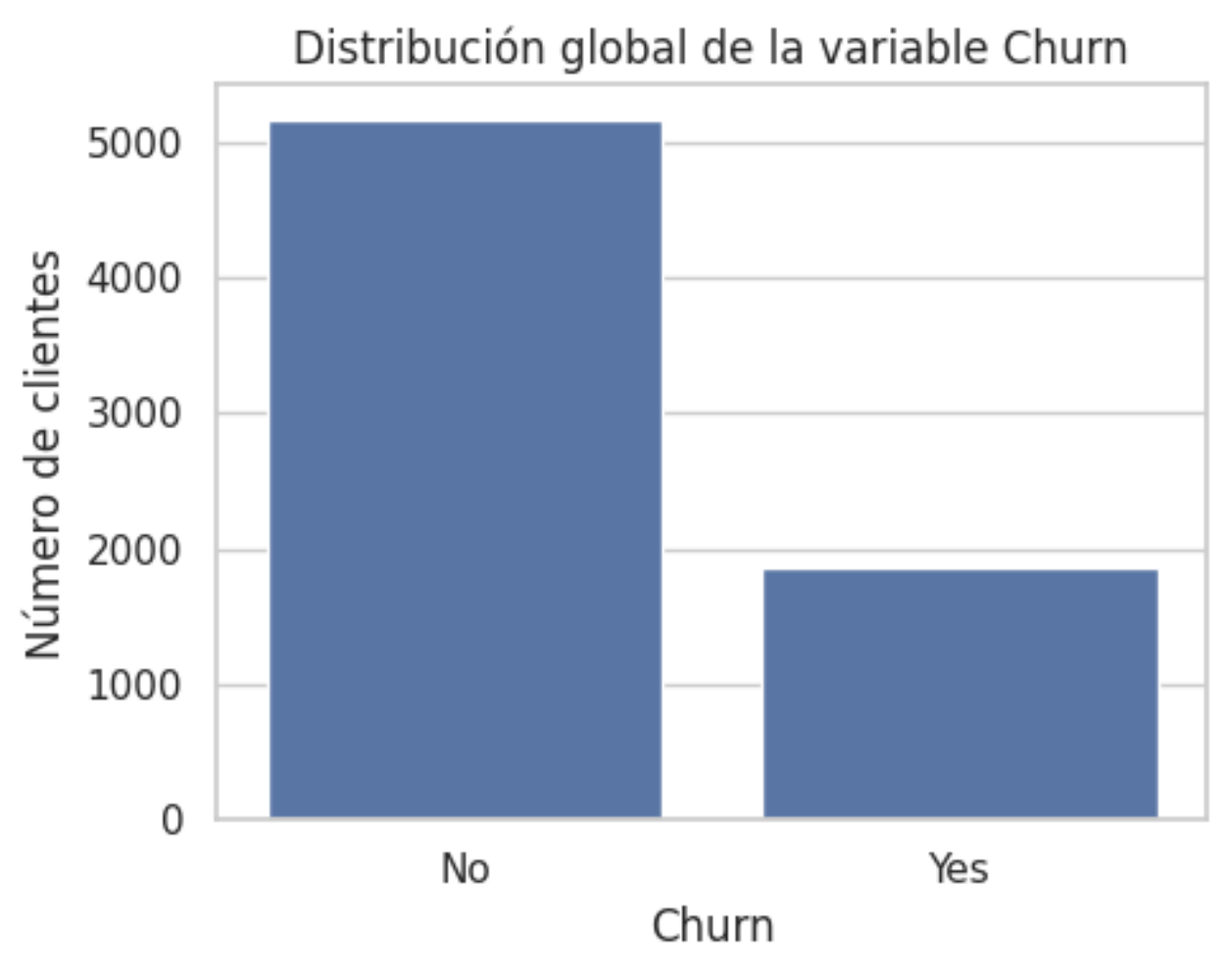
Desarrollar un modelo predictivo capaz de identificar clientes con alta probabilidad de abandono (*churn*), utilizando el dataset *Telco Customer Churn*, para apoyar decisiones estratégicas de retención y mejorar la rentabilidad del negocio.

Objetivos Específicos

Metas del Proyecto

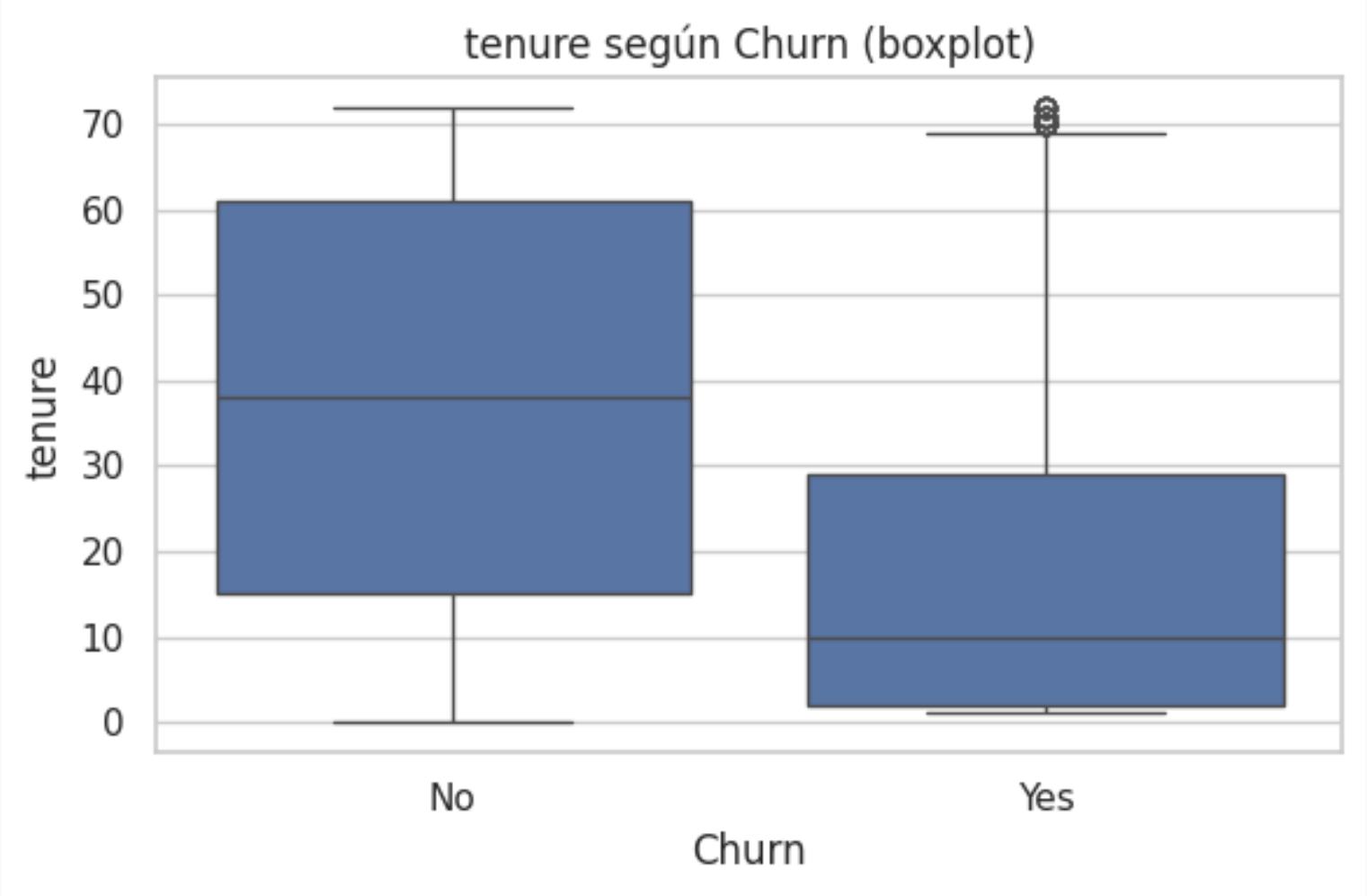
- **Analizar y preprocesar** el dataset (EDA, limpieza, codificación).
- **Entrenar y comparar modelos** de clasificación binaria.
- **Interpretar resultados** para identificar factores que influyen en el churn.

El Desafío: Desbalance de Clases



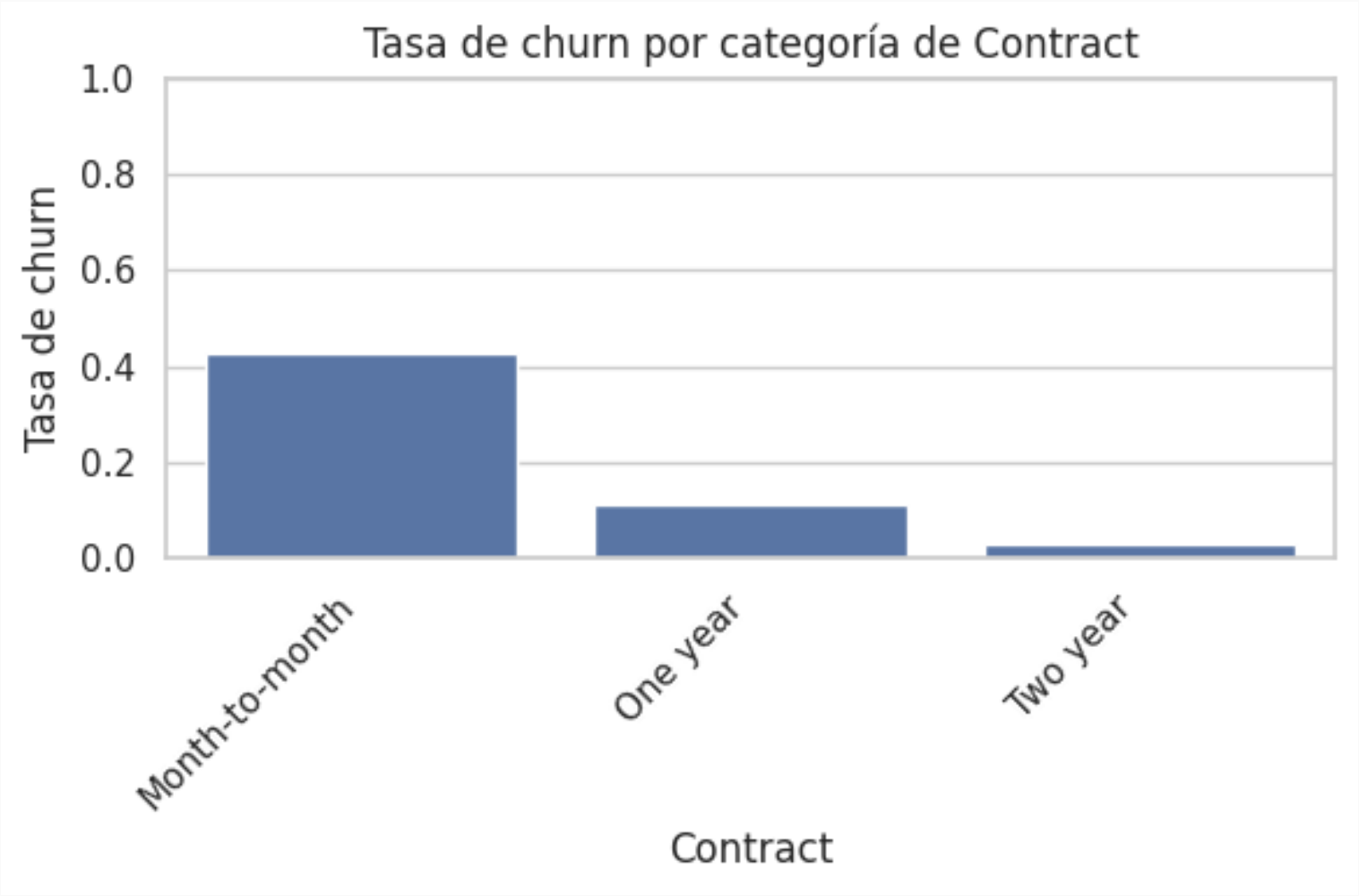
El 26.5 % de los clientes abandona la compañía.

El Factor Tiempo: ¿Cuándo se van?



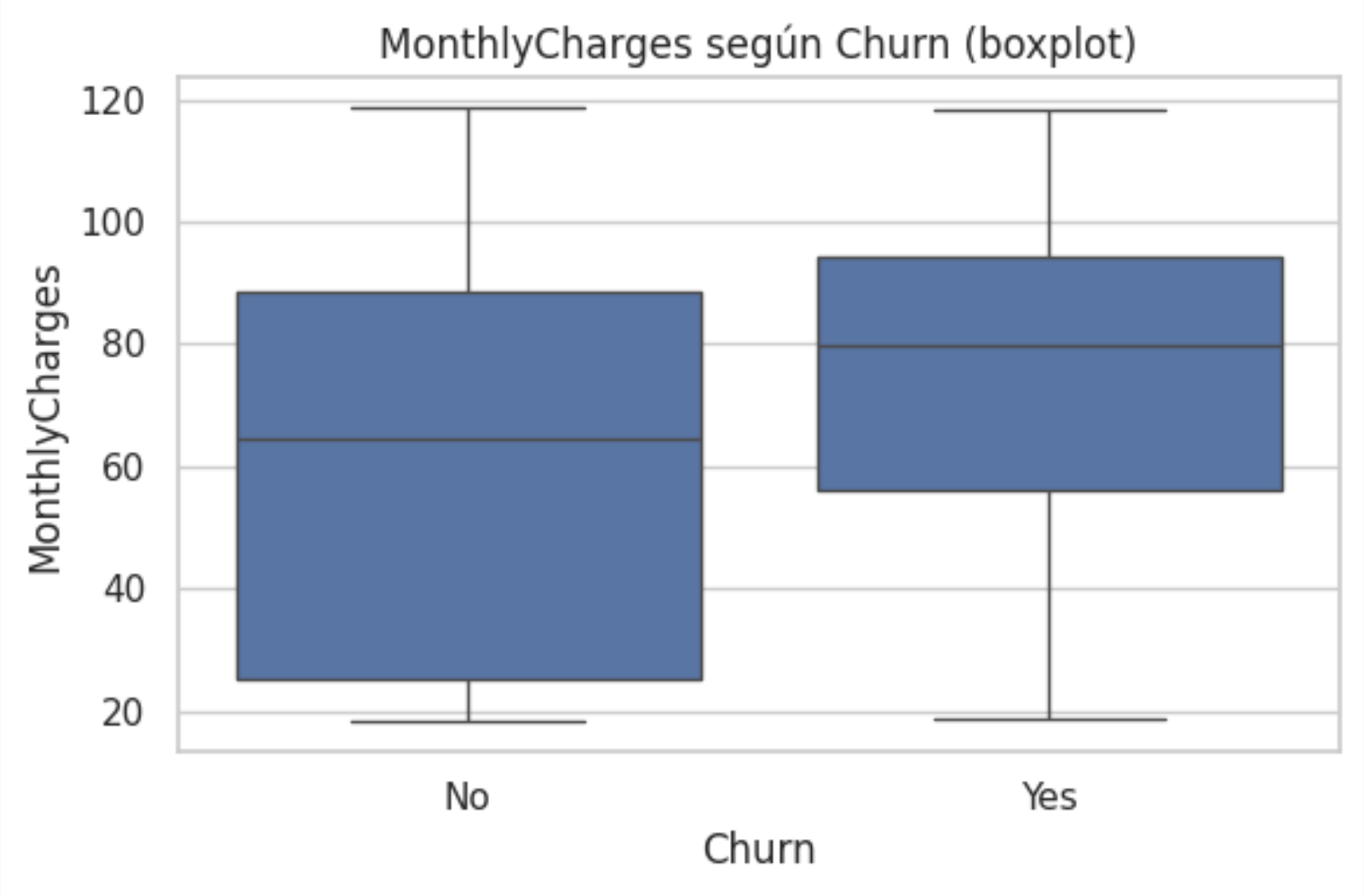
El riesgo es crítico en los primeros meses.

El Compromiso: Tipo de Contrato



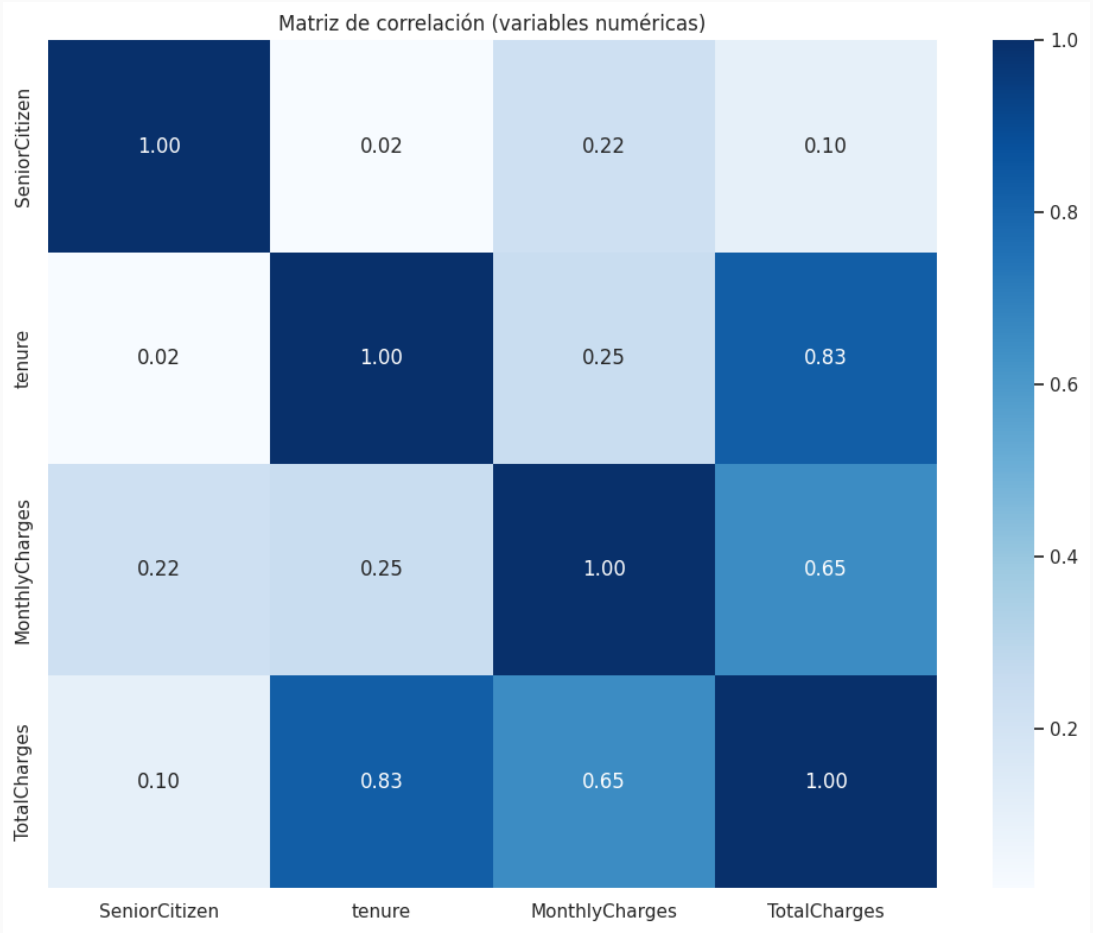
Contratos mensuales = Alto Riesgo (40 % Churn).

El Factor Económico: Sensibilidad al Precio



A mayor cargo mensual, mayor probabilidad de fuga.

Estadística y Preprocesado: Correlaciones



- **Alta colinealidad detectada** (Tenure vs TotalCharges).
- **Estrategia de Preprocesado:** Selección de features + SMOTE (Balanceo).

Metodología de Evaluación Rigurosa

Objetivo
Garantizar la robustez de las predicciones y evitar el sobreajuste (overfitting).

Métricas Clave:

- **ROC-AUC:** Métrica principal para balancear sensibilidad y especificidad.
- Matriz de Confusión, F1-Score, Accuracy.

Validación Cruzada:

- *Stratified K-Fold* ($k = 5$).
- Entrenamiento con datos balanceados (**SMOTE**).

Optimización: Modelos Base

Se implementaron modelos clásicos utilizando búsqueda exhaustiva de hiperparámetros (**Grid Search**).

Modelo	Configuración Clave
Regresión Logística	Penalización L2, Solvers optimizados.
SVM	Kernel RBF, ajuste de Gamma y C.
KNN	Distancia Minkowski, ponderación por distancia.

Nota Técnica

Se priorizó la probabilidad (‘predict proba‘) para el cálculo preciso de la curva ROC.

Optimización: Modelos de Ensamble

Para capturar patrones complejos no lineales, escalamos a modelos basados en árboles.

Random Forest

- **Estrategia:** Grid Search.
- **Foco:** Profundidad máxima y número de estimadores.
- Reducción de varianza mediante *bagging*.

XGBoost

- **Estrategia: Randomized Search**.
- **Foco:** Learning rate y submuestreo.
- Eficiencia computacional mediante *boosting*.

Resultados: Resumen Cuantitativo

Evaluación sobre el conjunto de prueba (Test Set) tras la optimización.

Modelo	Accuracy	Recall	F1-Score	ROC-AUC
Logistic Regression	0.778	**0.636**	**0.603**	**0.824**
XGBoost	**0.782**	0.607	0.597	0.818
Random Forest	0.769	0.588	0.575	0.807
SVM	0.776	0.588	0.583	0.805
KNN	0.743	0.626	0.565	0.786

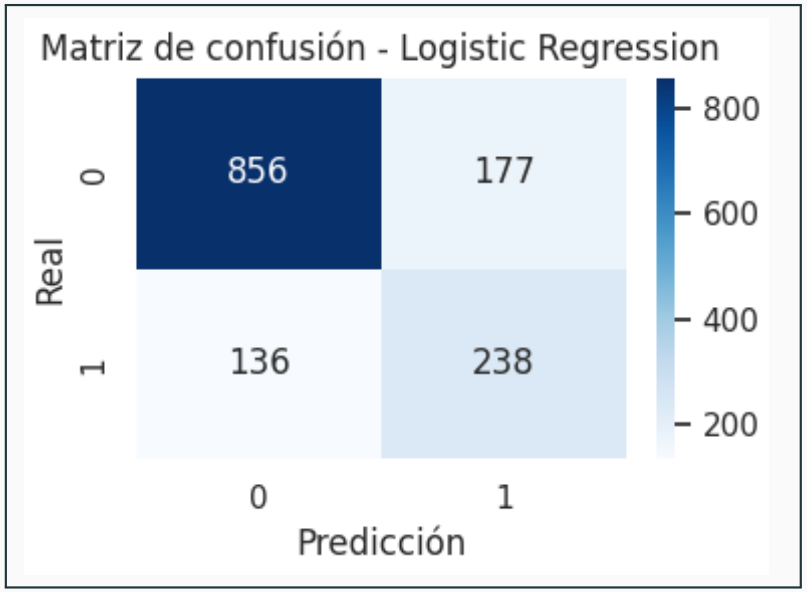
Hallazgo Clave

Aunque **XGBoost** tiene la mayor exactitud global, la **Regresión Logística** domina en capacidad de detección (Recall y AUC).

Análisis Visual: El ”Trade-off

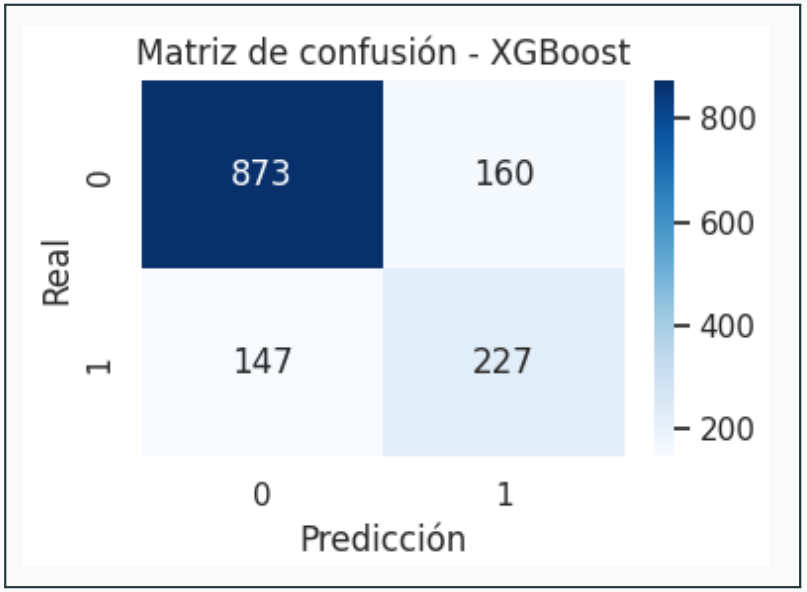
Comparación de curvas ROC y Matrices de Confusión.

Logistic Regression



Detecta mejor los positivos (Churners).

XGBoost



Menos falsos positivos, pero pierde churners.

Selección del Modelo y Justificación

Criterio de Negocio: Minimizar Falsos Negativos

En un contexto de *Churn* o Fraude, es más costoso perder un cliente (o no detectar un fraude) que revisar una falsa alarma.

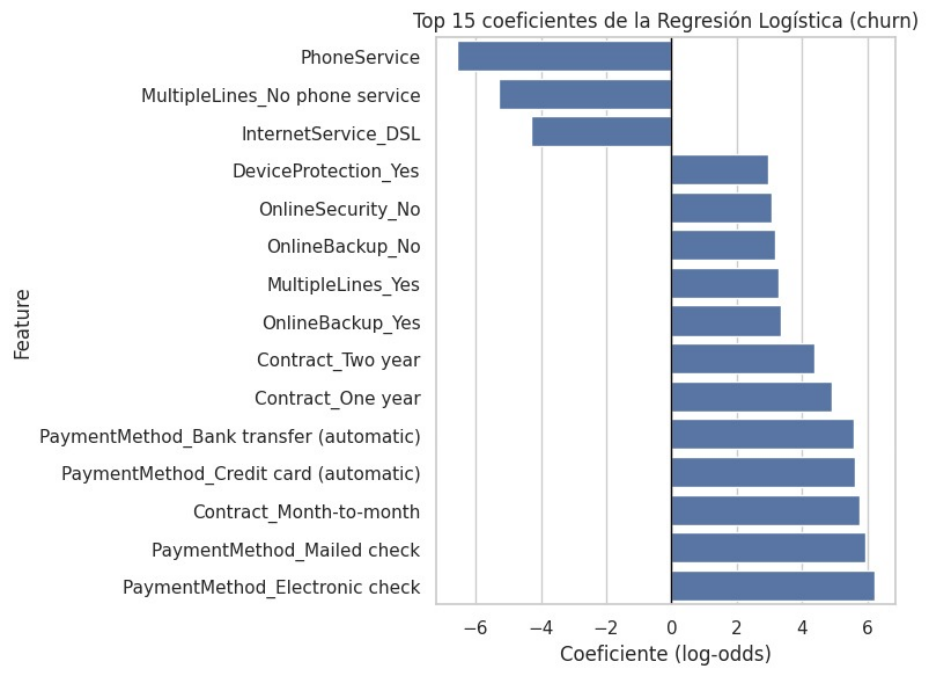
¿Por qué gana Logistic Regression?

- **Mejor Recall (63.6 %):** Atrapamos más casos reales.
- **Mejor ROC-AUC (0.824):** Mejor separación global de clases.
- **Interpretabilidad:** Permite entender qué variables influyen en la decisión.

✓

Modelo Robusto

Resultados: Interpretación de Coeficientes



Factores de Alto Riesgo (+):

- Pago Electrónico: El predictor más fuerte de abandono.
- Contrato Mes a Mes: La inestabilidad contractual aumenta el riesgo.
- Falta de Seguridad/Backup Online.

Factores de Retención (-):

- Servicio DSL: Menor riesgo vs Fibra Óptica.

**Nota Técnica: La negatividad en PhoneService sugiere colinealidad con MultipleLines (redundancia de información).*

Validación y Refinamiento: Valores SHAP

Resolviendo la Colinealidad:

- SHAP distribuye el impacto real entre variables correlacionadas.
- **Resultado:** *PhoneService* desaparece del top de importancia, confirmando que era redundante.

Nuevo Hallazgo (MonthlyCharges):

- Los puntos rojos (Cargos Altos) están a la **izquierda**.
- *Interpretación:* Clientes que pagan más tienen menor riesgo de fuga (efecto ”amarre”por tener más servicios contratados).

Conclusiones del Proyecto

- Problema: predicción de **churn** en una empresa de telecomunicaciones (dataset Telco Customer Churn).
- Tarea de **clasificación binaria** supervisada: estimar la probabilidad de abandono para cada cliente.
- Objetivo doble: **buen desempeño predictivo** + interpretabilidad.

Hallazgos Clave de Datos y Modelado

- Dataset **desbalanceado** $\Rightarrow$ uso de **SMOTE** en entrenamiento.
- Patrones por categoría:
  - Contratos **Month-to-month** y pagos **Electronic check** $\Rightarrow$ churn alto.
  - Pagos **automáticos** y contratos de **1–2 años** $\Rightarrow$ menor churn.
  - **Falta** de servicios aumenta el riesgo.
- Comparación de modelos: Logistic Regression obtiene las mejores métricas y se elige como **modelo campeón**.

Factores de Riesgo y Segmentos Clave

- Análisis de coeficientes + **SHAP**:
  - Segmento de **alto riesgo**: clientes nuevos, con contrato mensual, sin servicios adicionales y que pagan manualmente.
  - Segmento de **bajo riesgo**: clientes antiguos, con contratos de largo plazo, servicios extra y pagos automáticos.
- El modelo captura efectos condicionados que no siempre coinciden con las correlaciones simples vistas en el EDA.

Recomendaciones y Líneas Futuras

- Campañas **focalizadas**
- Estrategia de contratos **a largo plazo**
- Pagos **automáticos** y servicios **adicionales**
- **Fidelización** y cross-selling
- **Trabajo futuro**: más *feature engineering*, modelos avanzados y mejor tratamiento de colinealidades.

¡Gracias por su atención!