UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

DEPARTAMENTO
DE MATEMÁTICA

Predicción y Retención de Clientes: Estrategias de Negocio Basadas en el Análisis de Churn
Mat281

Estudiantes: Benjamín Benimelis - Joel Figueroa - Tomás Garrido - Marcelo Lector - Marcelo Ramírez - Diego Maldonado
Profesor: Francisco Alfaro

Domingo 30 de Noviembre, 2025

Contenido

Definición del problema

- Para las empresas de telecomunicaciones es fundamental retener a los clientes y, al mismo tiempo, atraer nuevos. La cancelación del servicio por parte del cliente (Churn) genera un costo económico elevado, ya que adquirir nuevos clientes es más caro que retenerlos.

Justificación del problema

- Un aumento en la tasa de churn implica:
  - Menores ingresos recurrentes.
  - Mayor gasto en marketing y captación de nuevos clientes.
- Poder **predecir quiénes están en riesgo de churn** permite:
  - Aplicar campañas de retención más focalizadas.
  - Ofrecer descuentos, upgrades o soporte adicional a los clientes adecuados.
  - Tomar decisiones basadas en datos.

Objetivos del análisis

- **Objetivo principal:**
  - Constuir un modelo predictivo que permita anticipar qué clientes podrían dejar la empresa, además de detectar variables que expliquen el por qué de esta decisión. Esto a través de la variable objetivo **Churn**
- **Objetivos secundarios:**
  - Detectar patrones y variables clave asociadas al abandono.
  - Evaluar distintos modelos de Machine Learning y comparar su desempeño.
  - Traducir los resultados a **recomendaciones de negocio** para retención de clientes.

Descripción del dataset

- Cada fila representa un **cliente**.
- Algunas variables relevantes:
  - `gender`, `SeniorCitizen`, `Partner`, `Dependents`.
  - `tenure`: meses que lleva con la empresa.
  - `PhoneService`, `InternetService`, `OnlineSecurity`, `TechSupport`, `StreamingTV` `StreamingMovies`.
  - `Contract`: mes a mes, 1 año, 2 años.
  - `PaymentMethod`, `PaperlessBilling`.
  - `MonthlyCharges`, `TotalCharges`.
  - `Churn`: variable objetivo (**Yes/No**).

Análisis exploratorio: vista general

- El dataset contiene:
  - 7043 clientes.
  - 21 columnas (3 numéricas, 18 categóricas).

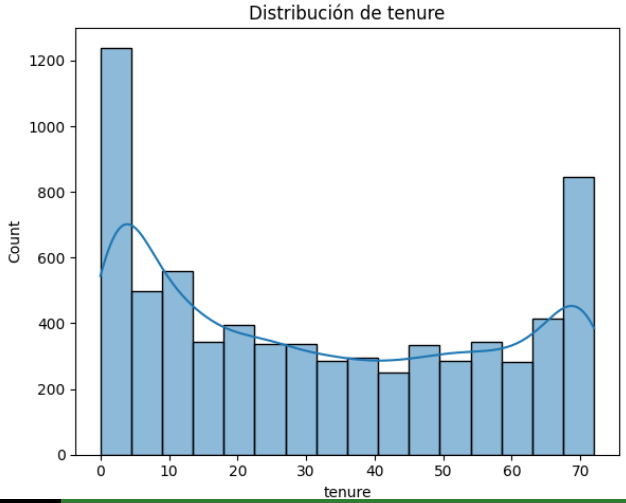
Estadísticas descriptivas generales

- Variables numéricas:
  - `tenure` varía entre 0 y 72 meses, con una media cercana a 32 meses.
  - `MonthlyCharges` muestra gran variabilidad, con tarifas bajas y tarifas altas asociadas a servicios más completos.
- Variables categóricas:
  - Mayoría de clientes con contrato `Month-to-month`.
  - `Fiber optic` es el tipo de internet más frecuente.
  - Muchos clientes no contratan servicios adicionales como `OnlineSecurity` o `TechSupport`.
- La variable objetivo `Churn`:
  - Aproximadamente **73 %** de clientes que se quedan.
  - Aproximadamente **27 %** de clientes que abandonan.

.

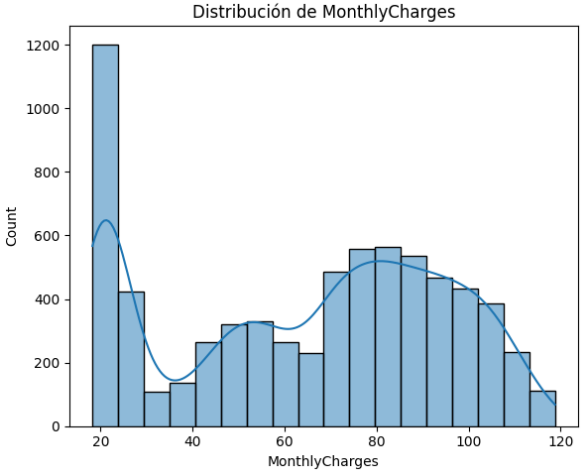
Distribución de variables numéricas: `tenure`

- La mayoría de los clientes tiene poca antigüedad en la empresa.
- Esto sugiere que los primeros meses y años son críticos para la retención.



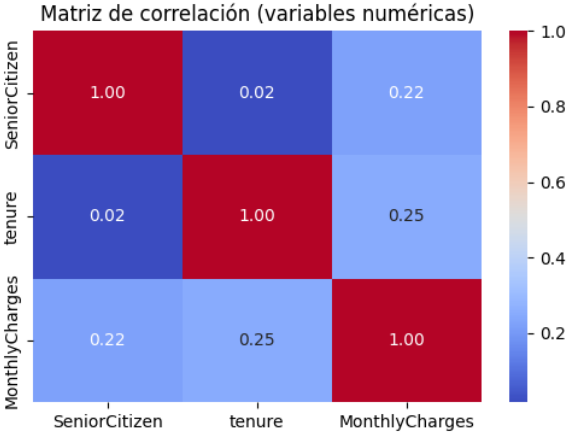
Distribución de variables numéricas: `MonthlyCharges`

- Existen clientes de bajo, medio y alto costo mensual.
- Muchos clientes pagando cerca de 70-90 USD al mes.



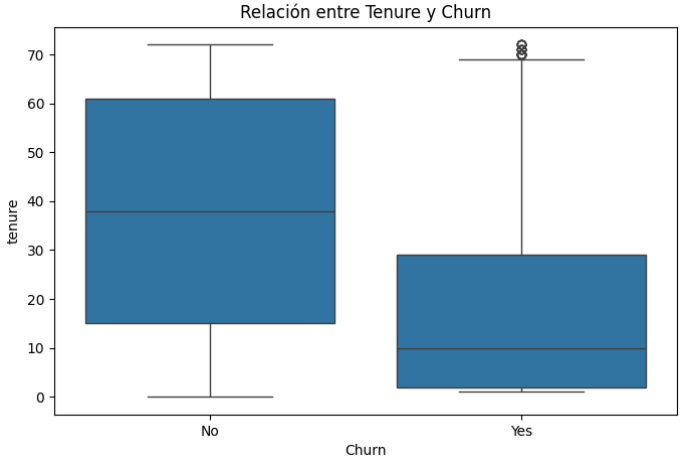
Correlación entre variables numéricas

- `MonthlyCharges` se correlaciona con:
  - `tenure`: mientras más tiempo en la empresa, tiende a aumentar el pago mensual.
  - `SeniorCitizen`: los clientes mayores pagan pagan cuotas un poco más altas.
- `SeniorCitizen vs tenure` tiene una correlación de 0.02, practicamente nula.



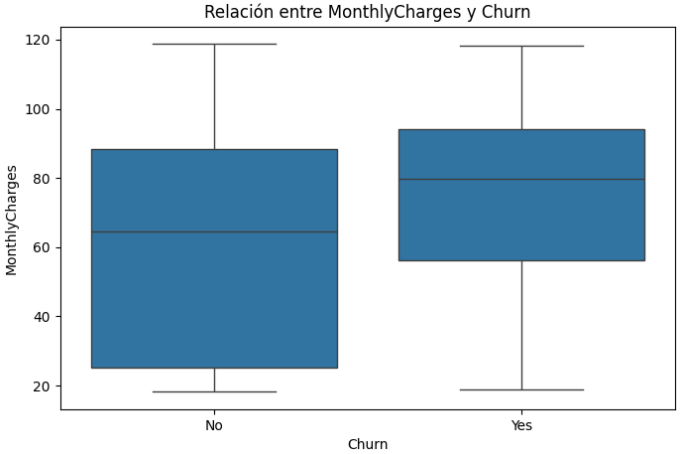
Relación entre tenure y Churn

- Los clientes que abandonan suelen tener poca permanencia.
- Los clientes con más de 3-4 años presentan churn mucho menor.



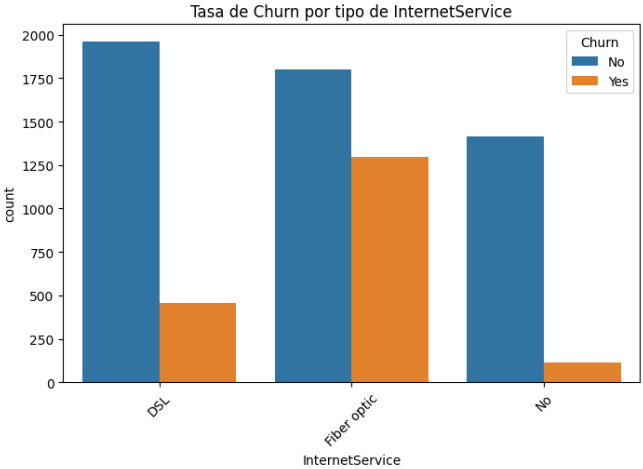
Relación entre MonthlyCharges y Churn

- Los clientes con cargos mensuales altos tienen más probabilidad de churn.
- Esto sugiere sensibilidad al precio o insatisfacción con servicios costosos.
- Los clientes de bajo costo abandonan menos.



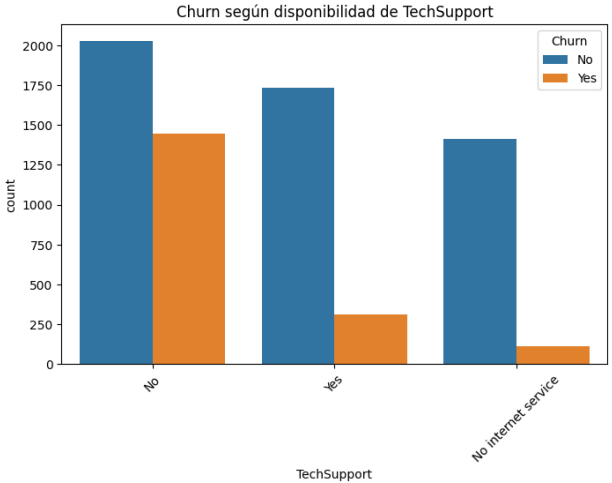
Churn según tipo de Internet

- Los clientes con **Fiber optic** muestran la mayor tasa de churn.
- Esto puede deberse a:
  - precios más altos,
  - expectativas de calidad más exigentes.
- Los clientes sin internet casi no abandonan.



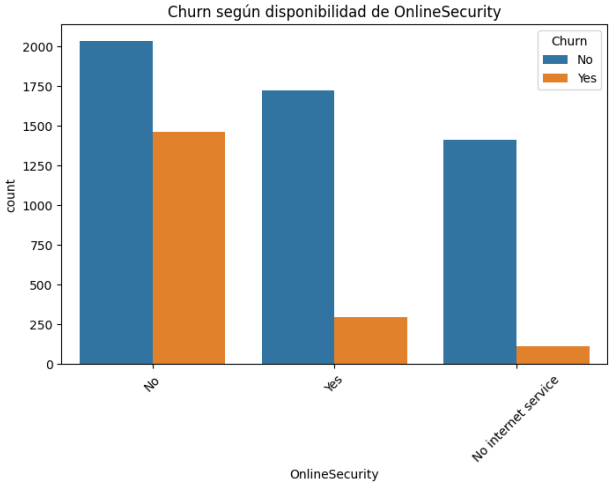
Churn y TechSupport

- Clientes con **TechSupport** presentan menor churn.
- Este servicio funciona como mecanismos de fidelización.



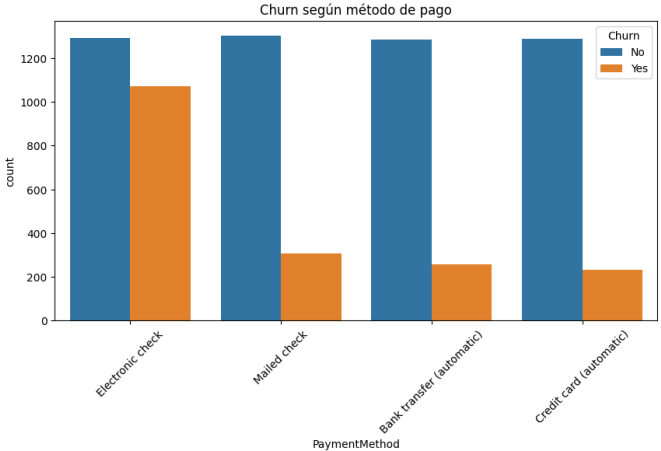
Churn y OnlineSecurity

- Al igual que el anterior, los clientes con **OnlineSecurity** presentan menor churn



Churn según método de pago

- El pago mediante **Electronic Check** tiene la mayor tasa de churn.
- Los métodos automáticos (banco o tarjeta) muestran menor abandono.
- Esto indica que los métodos manuales aumentan el riesgo de cancelación.



Limpieza y Transformación de Variables

1. Limpieza de Datos

- **Eliminación de Ruido:** Se eliminó la variable `customerID`. Al ser un identificador único aleatorio, no aporta valor predictivo y genera ruido.
- **Corrección de Tipos:** Se corrigió `TotalCharges` (de texto a numérico). Los valores vacíos en clientes nuevos se imputaron con 0.

2. Encoding (Texto a Números)

- **Label Encoding:** Para variables binarias (ej. `Partner`, `PhoneService`), transformándolas a 0 y 1.
- **One-Hot Encoding:** Para variables nominales (ej. `InternetService`). *Importancia:* Se crean columnas independientes para cada categoría, evitando que el modelo asuma un falso orden jerárquico (ej. $1 < 2 < 3$).

Preparación Estratégica para el Modelo

3. Escalamiento (MinMax)

- Se normalizaron las variables numéricas (`tenure`, `MonthlyCharges`) al rango $[0, 1]$.
- **Objetivo:** Evitar que variables con magnitudes grandes (como `TotalCharges` ≈ 8000) dominen injustamente sobre variables pequeñas (como `tenure` ≈ 70) en el cálculo de distancias.

4. División y Balanceo

- **División (Split):** 80 % Entrenamiento / 20 % Prueba para asegurar una validación honesta.
- **Balanceo (SMOTE):** Se generaron datos sintéticos de la clase minoritaria (*Churn*).
- El balanceo se aplicó exclusivamente al conjunto (`X_train`) para mantener la realidad de los datos de prueba.

Objetivo de la Etapa de Modelado

La Problemática

¿Cómo identificar eficazmente a los clientes con alta probabilidad de fuga (Churn) cuando la mayoría de los usuarios son fieles (Desbalance de clases)?

Estrategia de Selección:

- Se sometieron a prueba **5 algoritmos** de distinta naturaleza matemática.
- Objetivo: Encontrar el equilibrio entre *detección de riesgo* (Recall) y *estabilidad* (No Overfitting).
- Métrica clave: **ROC-AUC** (debido al desbalance de clases).

Metodología de Entrenamiento

El flujo de trabajo implementado en Python se dividió en tres fases automatizadas:

1. Candidatos

- Logistic Regression
- KNN
- SVM
- Random Forest
- XGBoost

2. Optimización

- Ajuste de hiperparámetros con `GridSearchCV` y `RandomizedSearch`.
- Sintonización fina para evitar memorización de datos.

3. Validación

- **Cross-Validation (CV=3):** Evaluar estabilidad.
- **Test Set:** Evaluación final con datos no vistos.
- **SMOTE:** Entrenamiento con datos balanceados.

Resultados de la Comparación

Tras el entrenamiento, se obtuvo el siguiente ranking ordenado por capacidad de generalización:

Modelo	CV ROC-AUC	Test ROC-AUC	Test Recall	Tiempo (s)
Logistic Reg.	0.85	**0.84**	**0.80**	**6.7s**
XGBoost	0.93	0.84	0.61	20.1s
Random Forest	0.92	0.82	0.59	28.3s
SVM	0.91	0.80	0.67	459.5s
KNN	0.86	0.77	0.70	1.0s

**Nota: CV = Cross Validation (Entrenamiento). Test = Datos reales.*

Análisis: ¿Por qué ganó la Regresión Logística?

Problema con Modelos Complejos

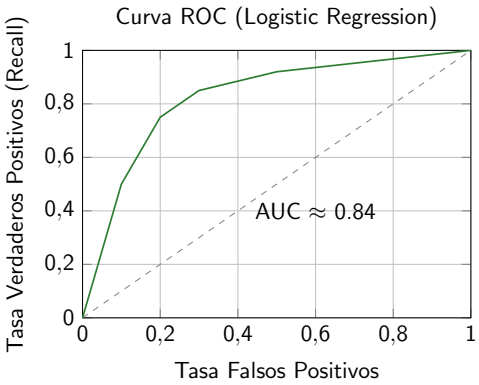
XGBoost y Random Forest mostraron **Overfitting**. Tuvieron puntajes casi perfectos en entrenamiento (0.93), pero cayeron drásticamente en Test, fallando en detectar el Churn real (Recall bajo 0.60).

Ventaja del Modelo Ganador

La **Regresión Logística**:

1. **Alto Recall (0.80):** Identifica a 8 de cada 10 clientes que se van a ir.
2. **Estabilidad:** Mantiene su rendimiento entre entrenamiento y prueba.
3. **Interpretabilidad:** Permite explicar el ≪por qué≫ al negocio.

Evaluación Visual del Modelo Seleccionado



Interpretación:

- La curva se aleja significativamente de la línea diagonal (azar).
- Un área de **0.84** indica una excelente capacidad para distinguir entre un cliente que se queda y uno que se va.

Conclusión de la Selección de Modelos

- Se descartan modelos de ≪Caja Negra≫ (SVM, XGBoost) por su alto costo computacional o tendencia al sobreajuste en este dataset específico.
- **Modelo Seleccionado: Regresión Logística.**

Implicancia para el Negocio:

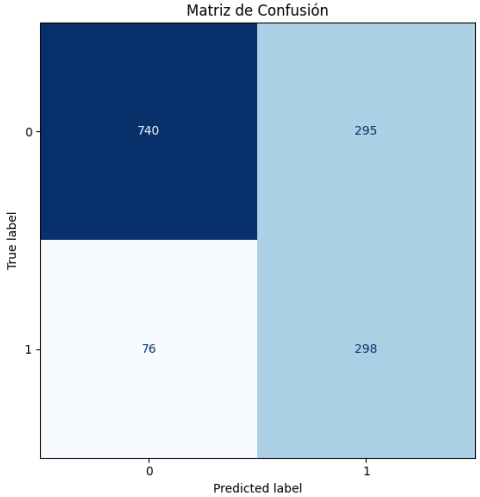
Al priorizar el **Recall**, el modelo prefiere "pecar de precavido": detectará a la gran mayoría de los clientes en riesgo, permitiendo a la empresa actuar proactivamente antes de que se produzca la baja.

Evaluación del Modelo Seleccionado

El modelo de **Regresión Logística** fue evaluado en el conjunto de prueba (20 % de los datos), mostrando un desempeño alineado con el objetivo de negocio: **minimizar la fuga no detectada**.

Métrica	Valor	Interpretación
ROC-AUC	**0.84**	Excelente capacidad de distinción global.
Recall (Churn)	**0.80**	*Prioridad:* Detectamos al 80 % de los fugados.
Accuracy	0.74	Rendimiento general aceptable.
Precision (Churn)	0.50	Aceptamos Falsos Positivos para ganar Recall.

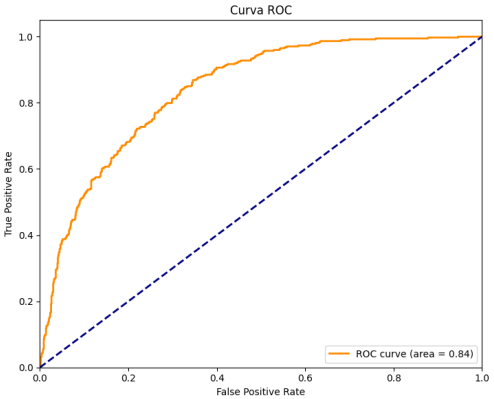
Análisis de Errores: Matriz de Confusión



Desglose de Predicciones:

- **Verdaderos Positivos (298):** Clientes en riesgo detectados correctamente.
- **Falsos Negativos (76):** *El dato crítico.* Solo perdimos de vista a un 20 % de los fugados reales.
- **Falsos Positivos (295):** Clientes fieles marcados como riesgo. Es el çosto.ªceptado de la estrategia preventiva.

Capacidad de Discriminación: Curva ROC



Análisis de la Curva:

- El modelo se aleja significativamente del azar (línea punteada).
- Un **AUC de 0.84** valida que el modelo asigna probabilidades altas a los casos de fuga real de manera consistente.

Conclusión Técnica

El modelo es **robusto** y cumple con la estrategia de maximizar la detección (Recall) manteniendo una precisión operativa viable para el negocio.

8.1 Recordatorio del modelo seleccionado

- Objetivo: predecir `Churn` (1 = se da de baja, 0 = se mantiene).
- Pipeline:
  - Estandarización: `tenure`, `MonthlyCharges`, `TotalCharges`.
  - One-hot encoding para variables categóricas.
  - Balanceo con **SMOTE** (solo en entrenamiento).
- Modelos probados: LR, KNN, SVM, RF, XGBoost.
- Modelo elegido: **Regresión Logística**.

Métrica	Accuracy	Recall (churn)	ROC-AUC
Valor	$\sim 0{,}74$	$\sim 0{,}80$	$\sim 0{,}84$

8.2 Interpretación de las métricas

Valores

- Accuracy $\approx 0{,}74$
- Recall (churn) $\approx 0{,}80$
- Precision (churn) $\approx 0{,}50$
- ROC-AUC $\approx 0{,}84$

Lectura

- Buen desempeño global.
- Detecta $\sim 80\,\%$ de los clientes que se van.
- Acepta falsos positivos para no perder churn.
- Buena separación entre clientes que se van y se quedan.

8.3 Interpretación de las variables

- **Regresión Logística**:
  - Coeficiente $> 0 \Rightarrow$ aumenta riesgo de churn.
  - Coeficiente $< 0 \Rightarrow$ disminuye riesgo de churn.
- **Tenure & cargos mensuales**
  - `tenure` bajo $\Rightarrow$ más riesgo.
  - `MonthlyCharges` altos $\Rightarrow$ más riesgo.
- **Servicios complementarios**
  - Sin `TechSupport` / `OnlineSecurity` $\Rightarrow$ más churn.
  - Tenerlos actúa como factor protector.
- **Método de pago**
  - `Electronic check` se asocia a mayor churn.
  - Pagos automáticos $\Rightarrow$ menor riesgo.

8.4 Implicancias para la toma de decisiones

- **Ranking de riesgo**: usar probabilidad de churn para priorizar clientes.
- **Segmentos clave de retención**:
  - Clientes nuevos, con cargos altos, sin soporte/seguridad.
  - Clientes que pagan con `Electronic check`.
- **Umbral de decisión**:
  - Ajustar según costo de perder un cliente vs. costo de la campaña.
- **Monitoreo**:
  - Actualizar periódicamente las probabilidades.
  - Evaluar impacto de las campañas sobre la tasa de churn.

9 Conclusiones

- Se realizó un examen del dataset "Telco Customer Churn"
- Hubo preprocesamiento de datos
- Múltiples modelos chequeados y evaluados

- Modelo de regresión logística obtuvo mejores resultados:
  - Accuracy: approx 0.74
  - Recall: approx 0.80
  - Precision: 0.50
  - ROC-AUC: approx 0.85

El modelo es bueno para distinguir entre clientes que se quedan y clientes que se van.

9 Recomendaciones

- Se recomienda programas de onboarding y acompañamiento proactivo durante los primeros meses.
- Agregar o reforzar servicios como TechSupport (soporte técnico) y OnlineSecurity (seguridad en línea)
- Monitoreo continuo y acciones proactivas.