Predict behavior to retain customers with IBM Sample Data Sets

Martina Blanco Juan Bustos Diego Mallea
Emilia Pardo Felipe Pérez

Universidad Técnica Federico Santa María
Departamento de Matemática

28 de Noviembre, 2025

Contenido

Definición del problema

En el mundo de las telecomunicaciones hay dos elementos super importantes para sobrevivir, retener y atraer clientes nuevos.

Usando un conjunto de datos proporcionado por la empresa "GUAU"de telecomunicaciones, construiremos un modelo predictivo para anticipar qué clientes podrían dejar la empresa, y algunas recomendaciones para mantenerlos y atraer nuevos.

Definir el Dataset (variables y tamaño aprox.)

- TotalCharges: total de dinero que el cliente ha pagado históricamente a la compañía hasta la fecha.

Consideramos que esta variable es predictiva. Notemos que posee una pequeña cantidad de elementos vacios que rellenaremos,
Aproximación para el rellenado:

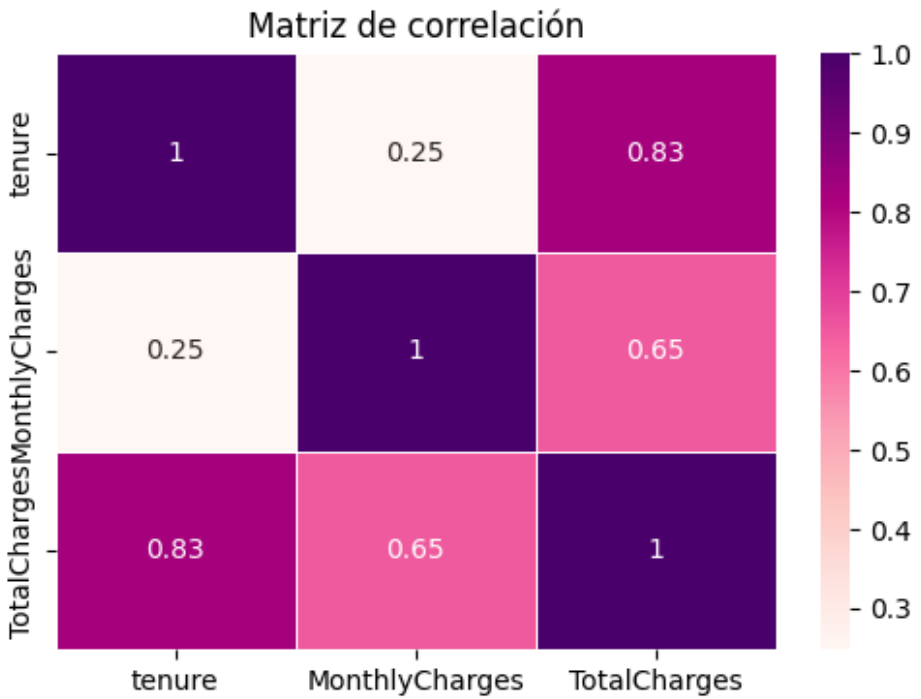
$$\text{TotalCharges} = \text{tenure} * \text{MonthlyCharges}$$

Donde:

- MonthlyCharges: monto que el cliente paga cada mes por los servicios que tiene contratados.
- tenure: número de meses que el cliente lleva en la compañía.

Eliminamos la variable customerID ya que no reprentan datos significativos.

Matriz de Correlación

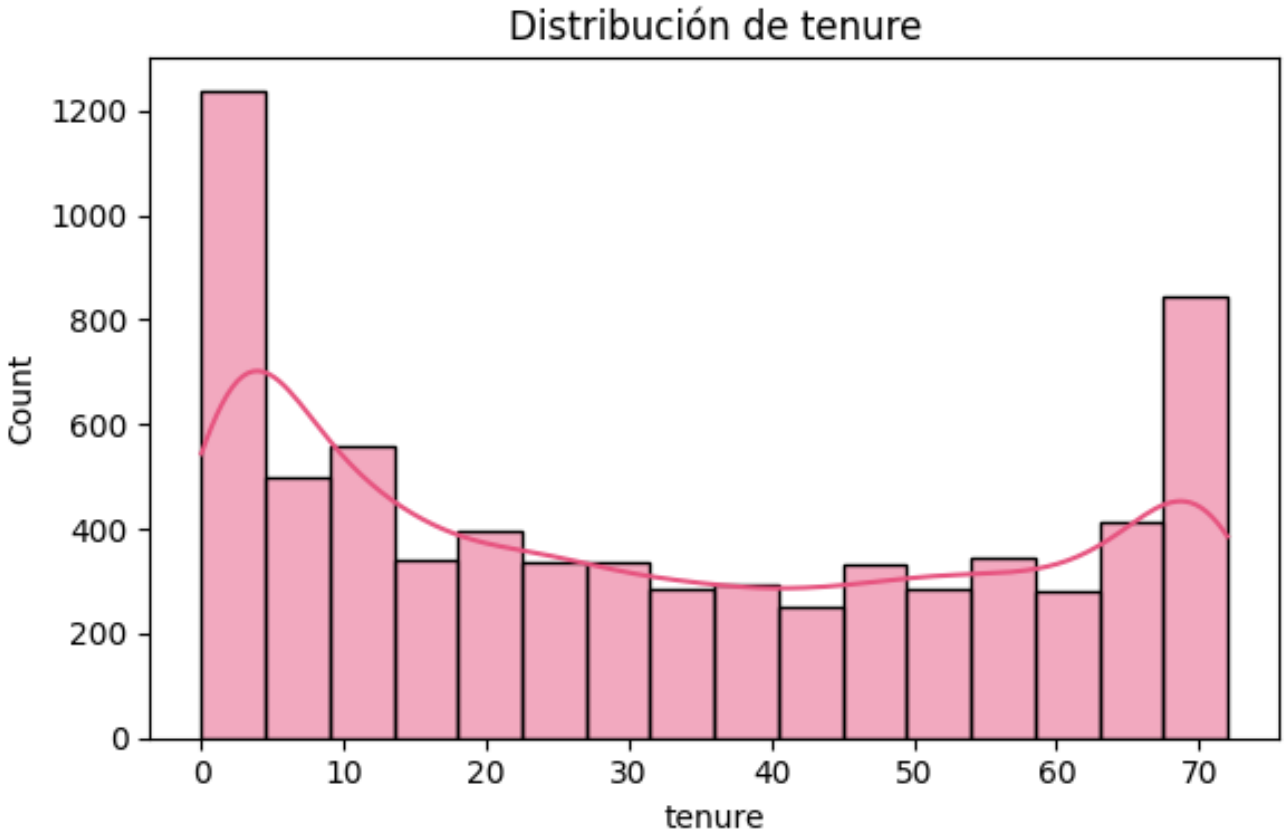


Matriz de Correlación

La matriz muestra la relación entre las variables numéricas tenure, MonthlyCharges y TotalCharges. Podemos obtener de estas relaciones que:

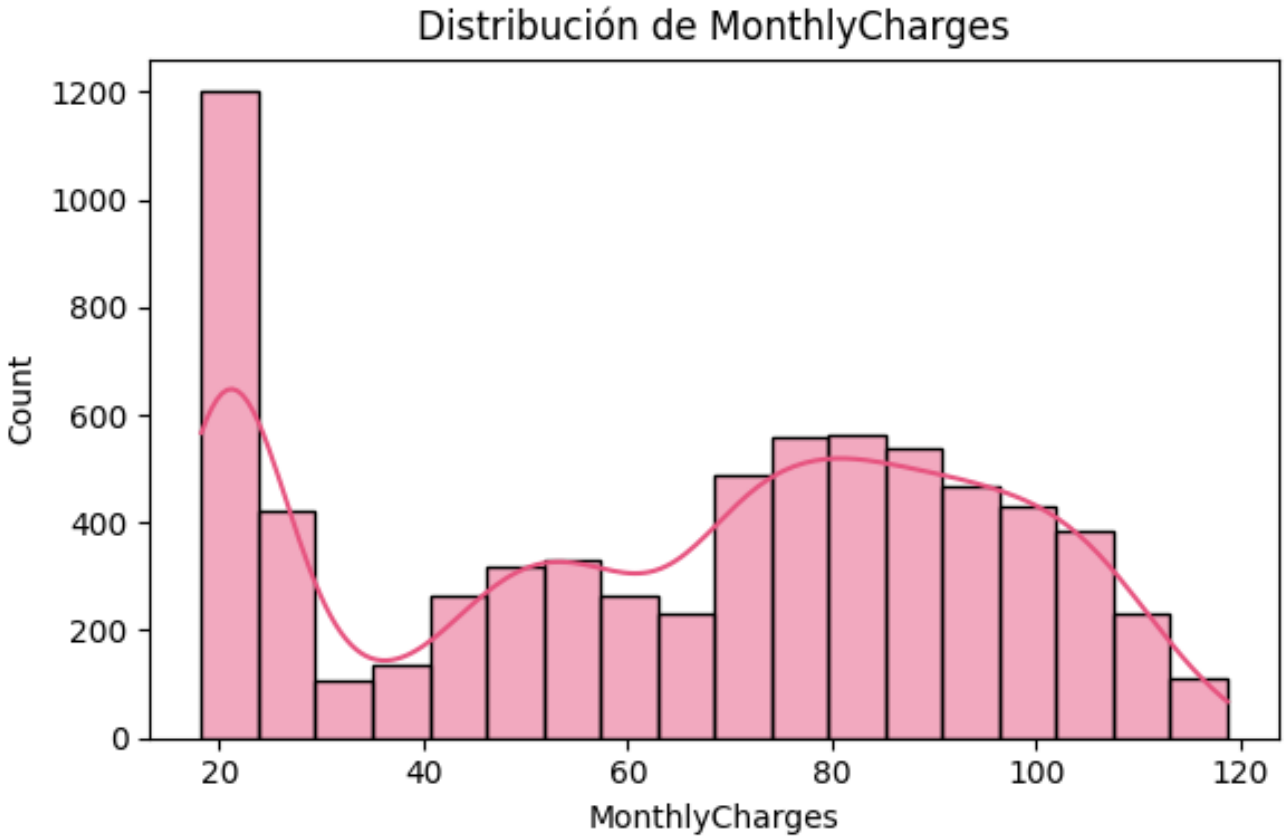
- Mientras más tiempo lleva un cliente en la empresa, mayor es el monto total que ha pagado.
- Los clientes con planes mensuales más costosos tienden a acumular mayores cargos totales.
- La antigüedad de un cliente no está relacionada con cuánto paga mensualmente.

Distribución de tenure



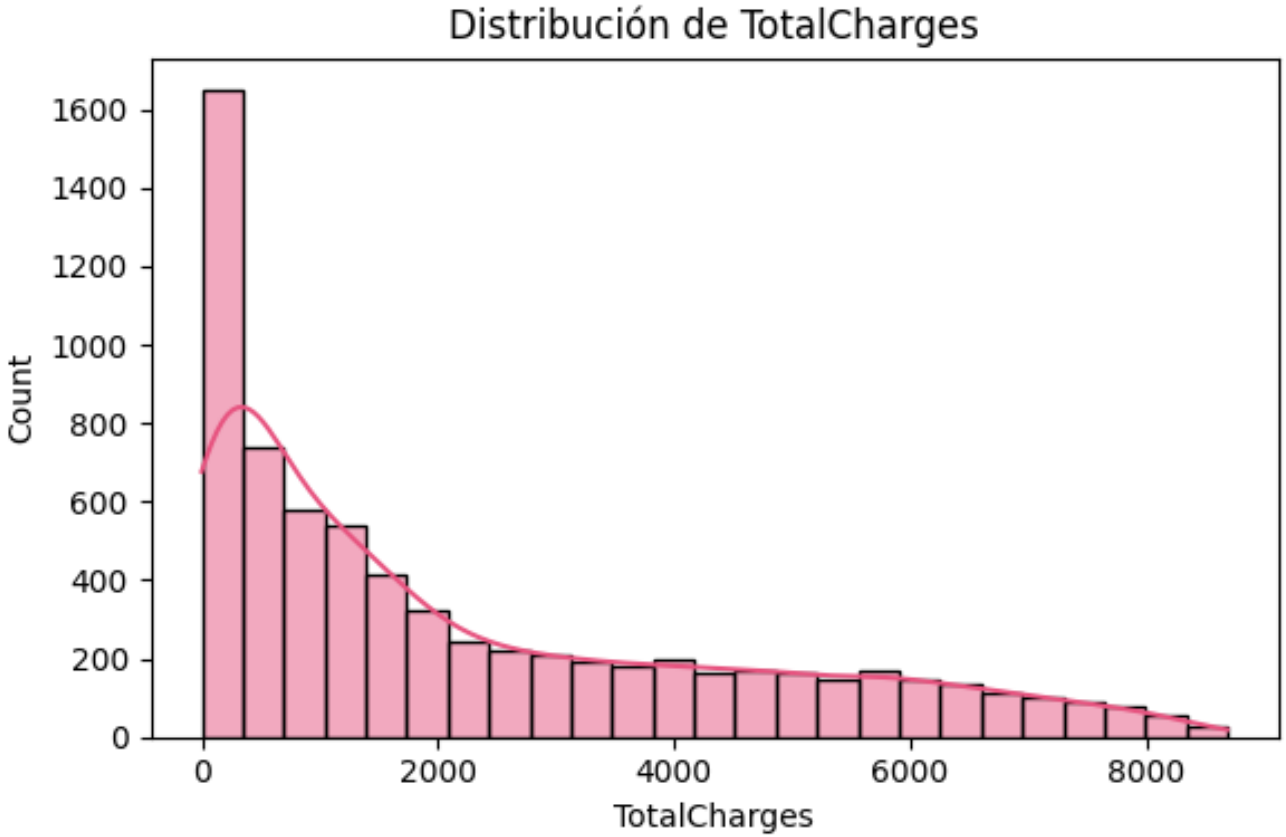
Esta distribución muestra que hay muchos clientes muy nuevos (0 a 5 meses) y muy antiguos (alrededor de 70 meses). Esto indica que la empresa tiene problemas para retener a los clientes nuevos, pero que quienes permanecen más tiempo tienden a quedarse por años.

Distribución MonthlyCharges



La distribución muestra dos grupos claros: uno de clientes con planes baratos (entre 20 y 30 dólares) y otro grupo con planes más caros (70 a 90 dólares). Esto sugiere segmentación de productos. Los clientes con planes caros tienden a tener mayor riesgo de churn.

Distribución TotalCharges

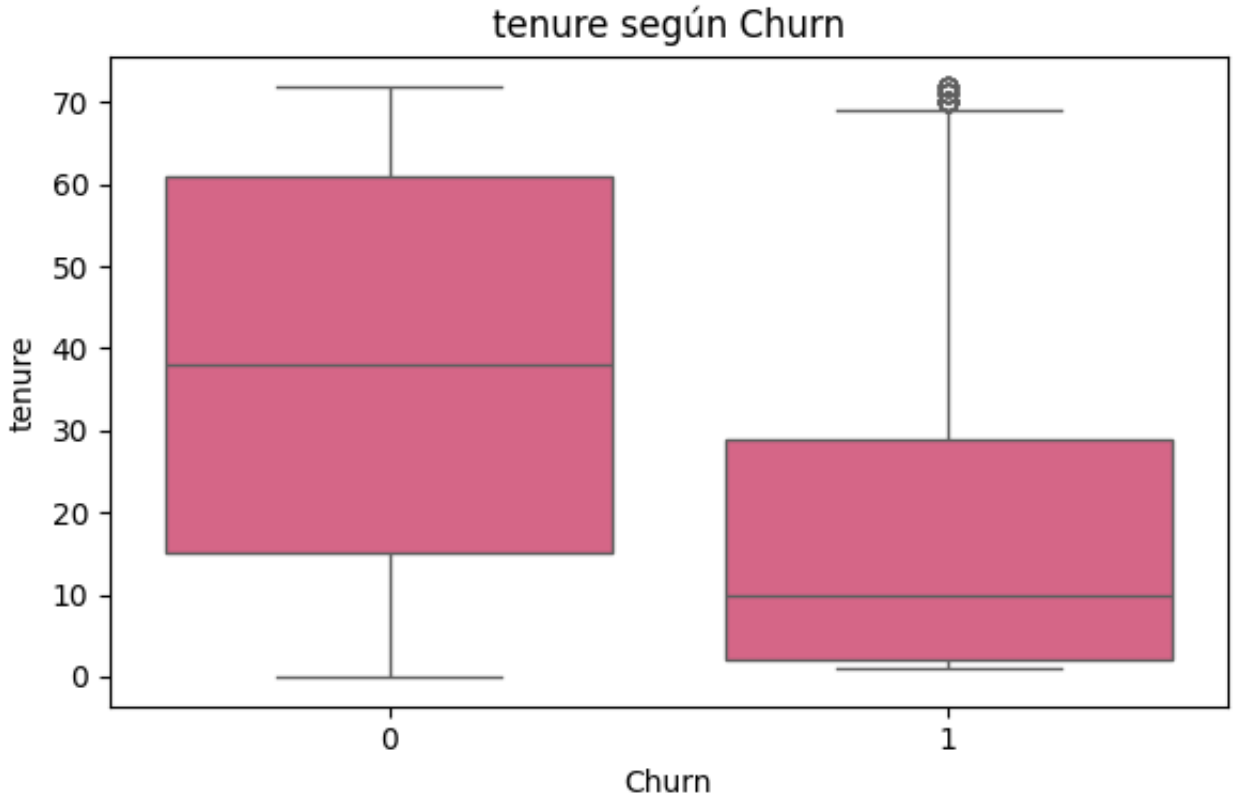


La distribución indica que muchos clientes han pagado poco, es decir, son nuevos, y a medida que el total pagado aumenta, hay menos clientes. Los que han pagado más suelen ser antiguos y tienen menor churn.

Distribución

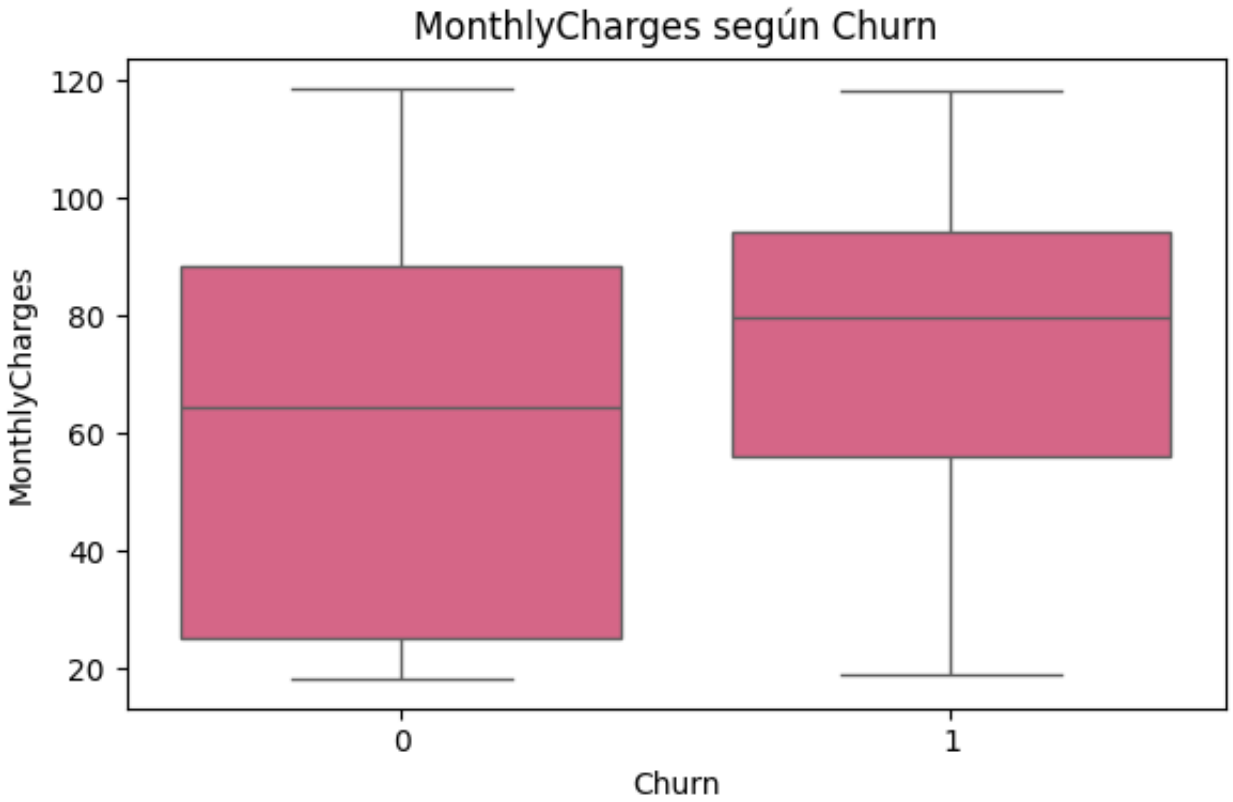
En conjunto, los tres gráficos muestran que la empresa tiene muchos clientes nuevos, lo cual aumenta el churn, mientras que los clientes antiguos presentan mayor estabilidad.
También se observan dos niveles claros de precios mensuales, lo que refleja distintos segmentos de clientes.

Variable tenure



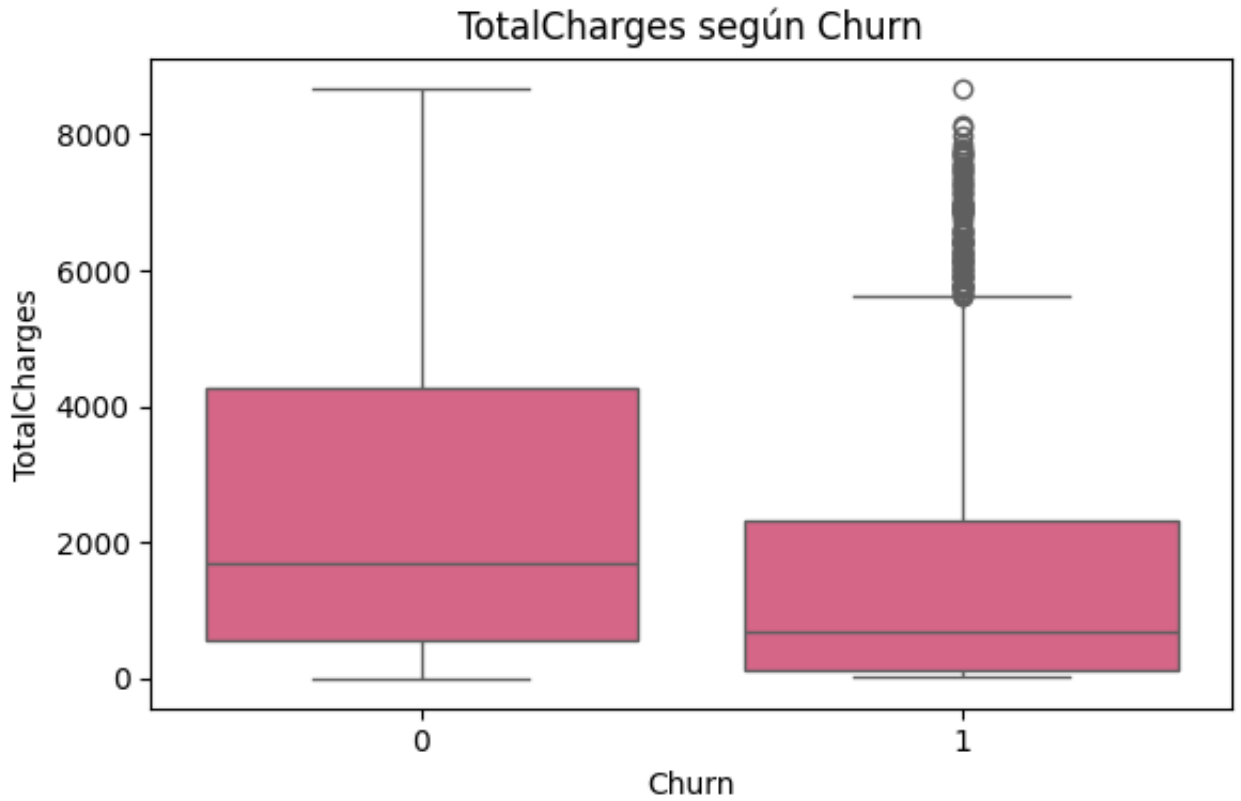
En la variable tenure se observa que los clientes que no hacen churn tienen mayor antigüedad en la compañía. Por el contrario, los clientes que churnean tienen un tenure muy bajo, esto indica que los clientes nuevos son los que más se van, mientras que los clientes que llevan más tiempo tienden a permanecer.

Variable MonthlyCharges



En la variable MonthlyCharges se observa que los clientes que churnean pagan planes más costosos. Esto sugiere que los planes de mayor precio, especialmente los asociados a fibra óptica o servicios adicionales, presentan una mayor probabilidad de churn.

Variable TotalCharges

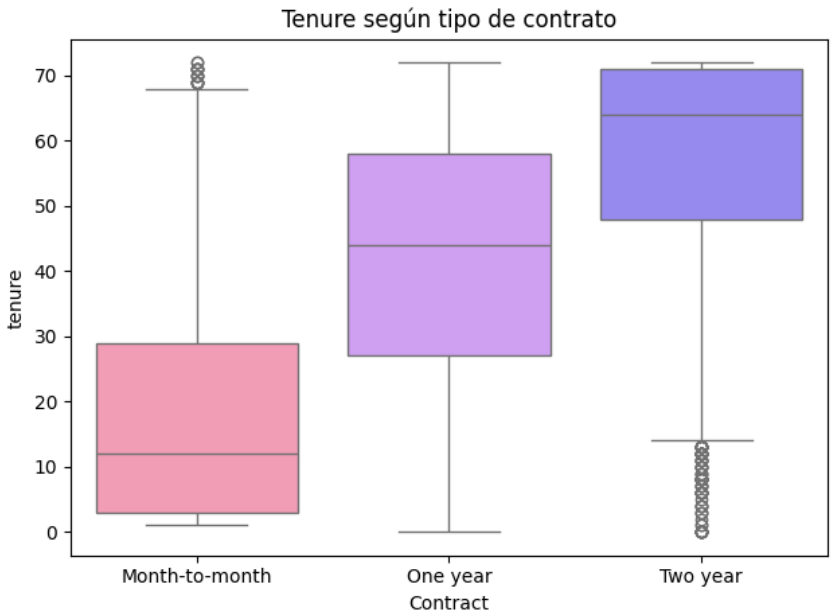


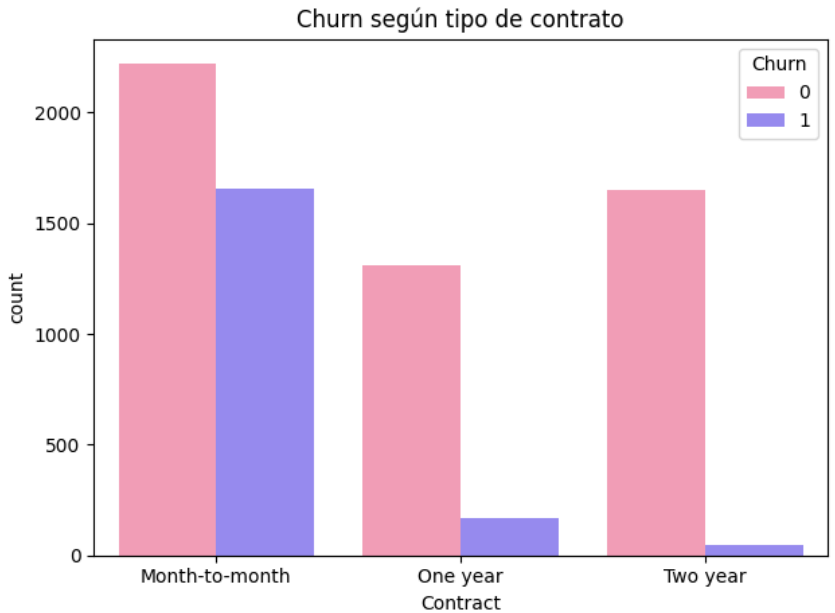
En la variable TotalCharges, los clientes que no churnean muestran un total pagado significativamente mayor. Los clientes que churnean presentan valores mucho menores. Esto confirma que los clientes que abandonan suelen ser nuevos, ya que han pagado poco tiempo en la empresa.

Clientes y churn

En conjunto, los tres gráficos indican que el churn ocurre principalmente en clientes nuevos que pagan planes más caros. Esto define un perfil claro de riesgo: **clientes con poco tiempo en la compañía y altos costos mensuales.**

Análisis detallado de tenure y contrato





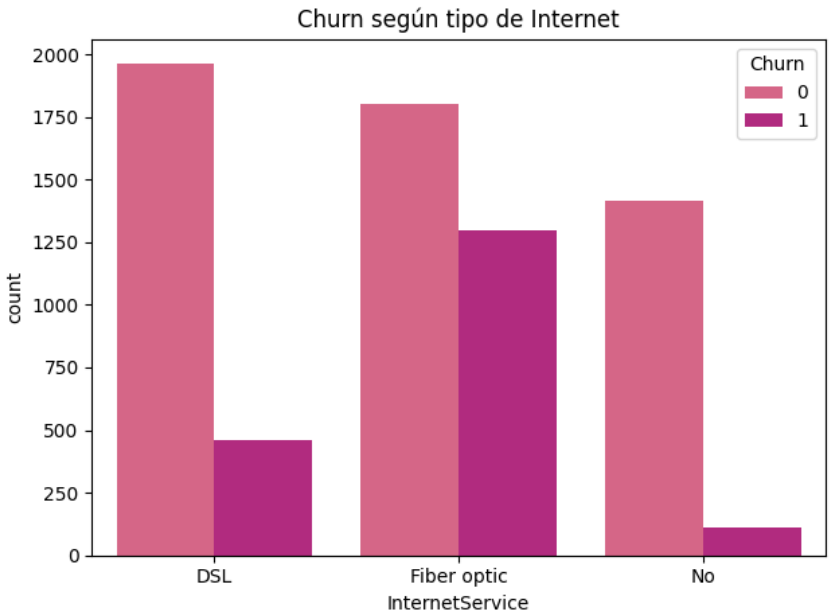
Análisis detallado de tenure y contrato

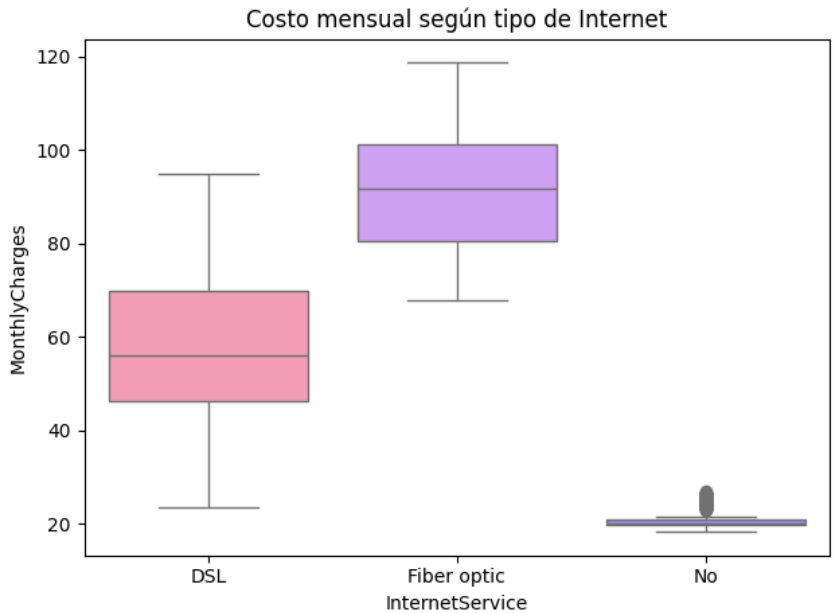
Los gráficos muestran una relación muy fuerte entre el tipo de contrato y el churn:

- Los clientes con contrato month-to-month presentan un tenure bajo y abandonan más el servicio.
- Los clientes con contratos de un año muestran un tenure significativamente mayor y una tasa de churn mucho más baja.
- Los clientes con contratos de dos años son los más estables: tienen el mayor tenure y casi no presentan churn.

Con esto podemos decir que el tipo de contrato es uno de los factores más importantes para predecir churn.

Análisis detallado de servicios de internet





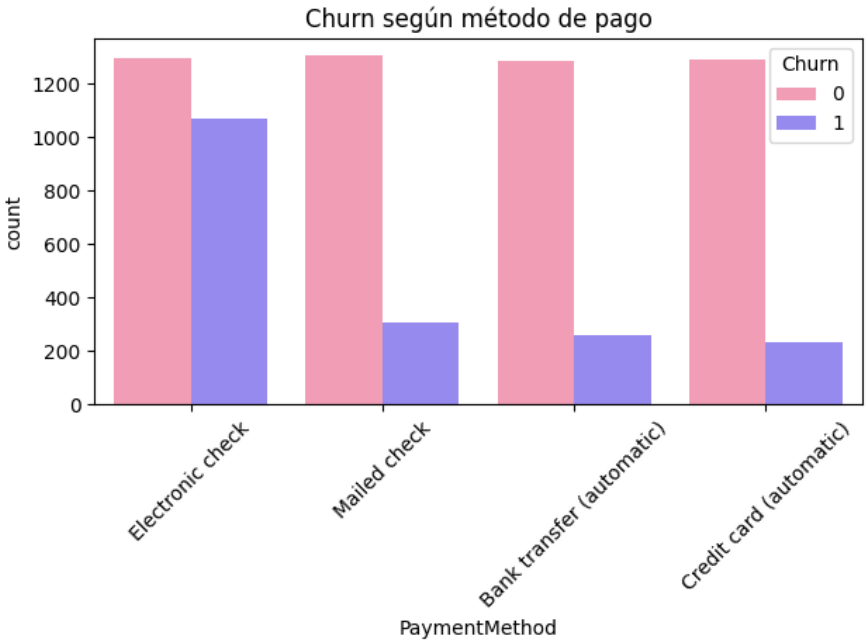
Análisis detallado de servicios de internet

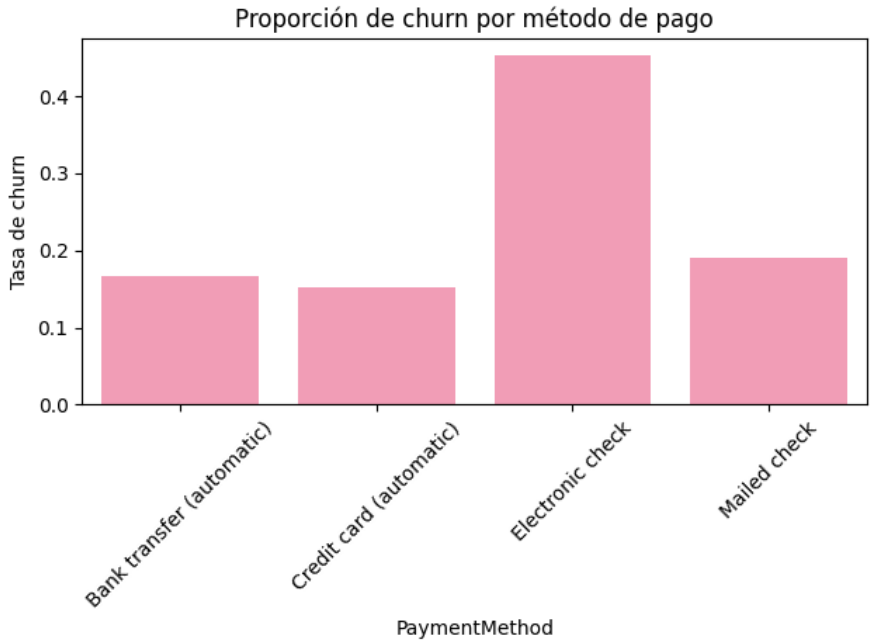
Los gráficos muestran que el tipo de servicio de Internet tiene una relación clara con el churn:

- Fiber optic presenta la mayor cantidad de abandonos (costo mensual más alto).
- DSL tiene un churn (y costo) considerablemente menor.
- Los clientes sin servicio de internet prácticamente no abandonan.

Esto sugiere que el costo elevado de la fibra óptica aumenta la probabilidad de churn, mientras que los servicios más económicos retienen mejor a los clientes.

Análisis detallado de método de pago





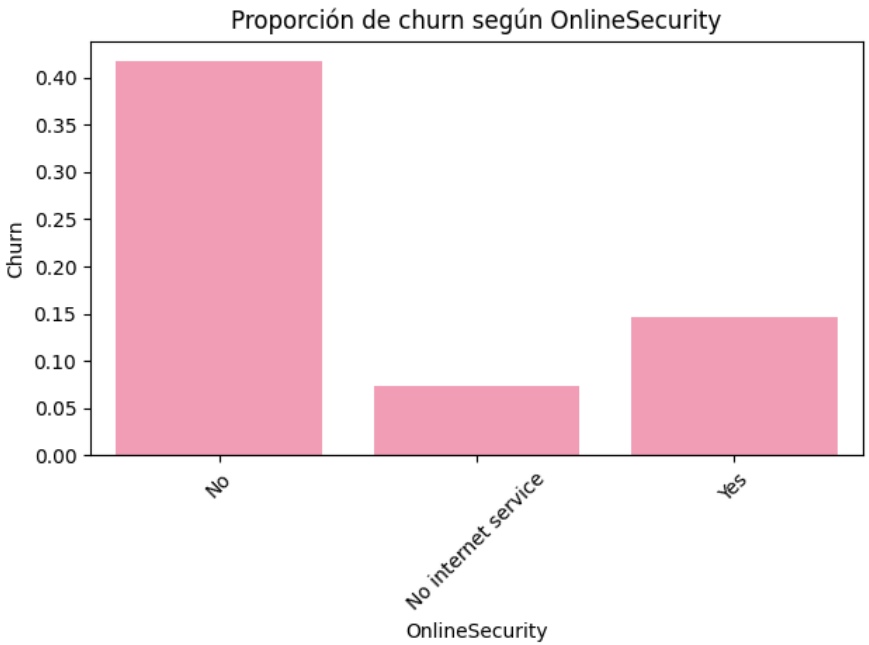
Análisis detallado de método de pago

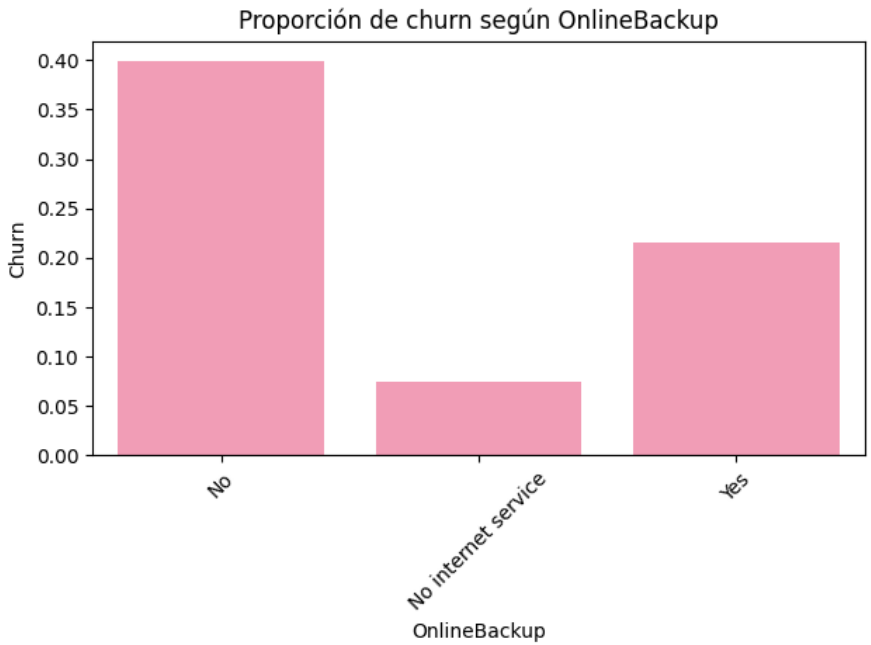
Los gráficos muestran que el método de pago tiene una relación directa con el churn:

- Electronic Check es el método más riesgoso, donde la proporción de clientes que abandonan es significativamente mayor que en los demás métodos.
- Bank Transfer y Credit Card presentan las tasas de churn más bajas.
- Mailed Check también muestra un churn relativamente bajo.

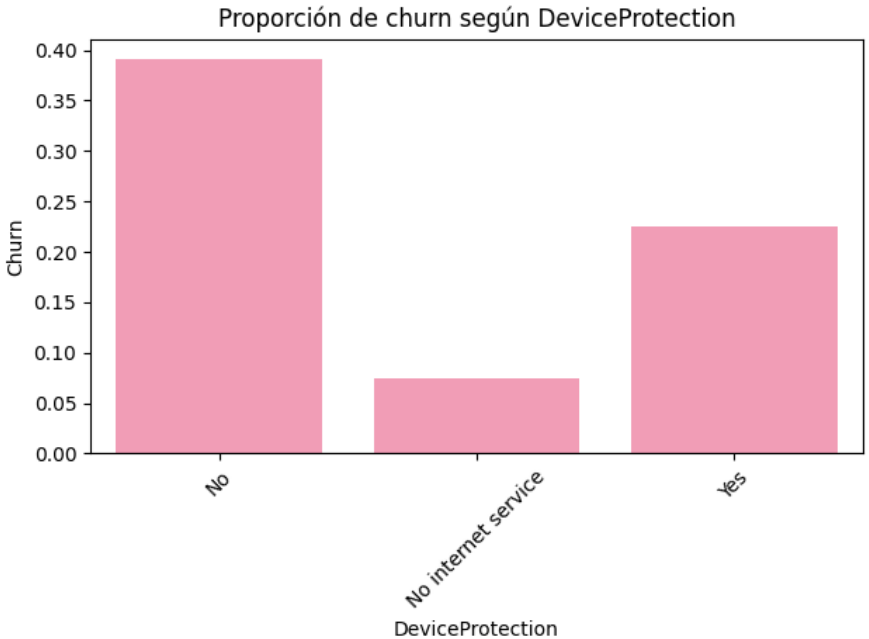
En general, el método de pago es un fuerte predictor del comportamiento de churn, destacando que los clientes que utilizan Electronic Check constituyen un segmento crítico para la retención.

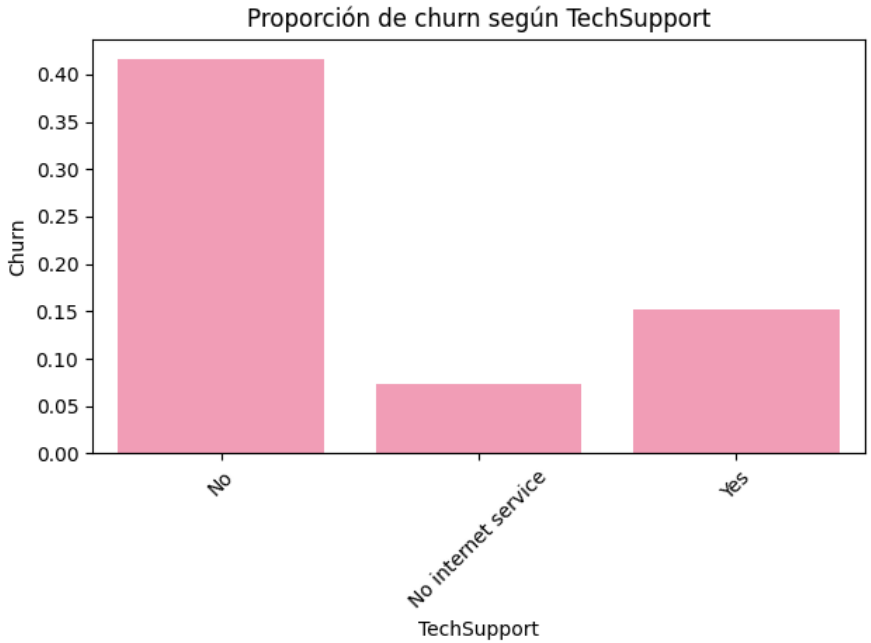
Análisis detallado de servicios de protección





Análisis detallado de servicios de protección





Análisis detallado de servicios de protección

Los gráficos permiten observar la proporción de churn según cuatro servicios adicionales ofrecidos por la compañía: OnlineSecurity, OnlineBackup, DeviceProtection y TechSupport.

- Los clientes que no cuentan con estos servicios presentan las tasas de churn más elevadas, cercanas al 40 %.
- Los clientes clasificados como "No internet service" exhiben tasas de churn considerablemente bajas, alrededor del 7–8 %.
- Los clientes que sí cuentan con los servicios adicionales presentan tasas de churn intermedias, aproximadamente entre 14 % y 22 %.

Esto sugiere que servicios como seguridad en línea, respaldo remoto, protección de dispositivos o soporte técnico contribuyen a aumentar la retención de usuarios, fortaleciendo la relación del cliente con la empresa.

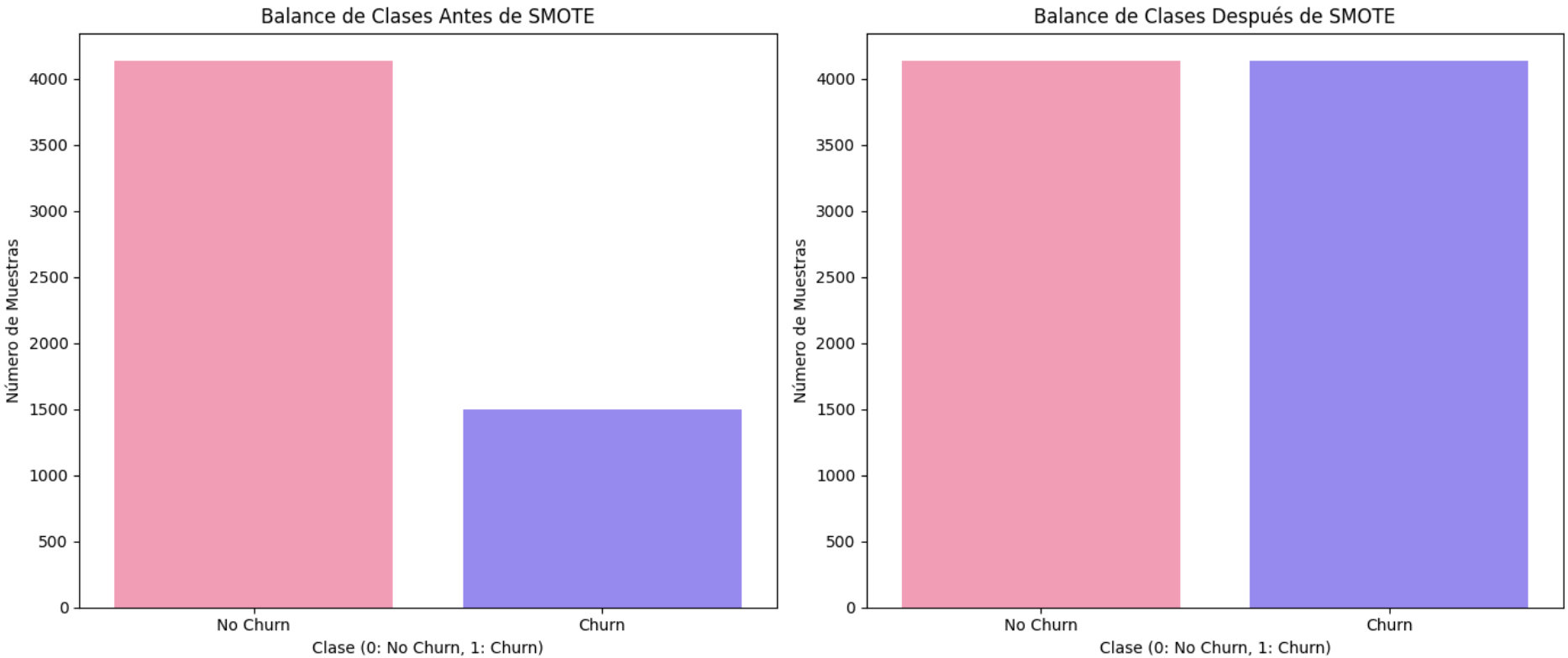
Preprocesamiento de Datos para la Predicción de Churn

En esta etapa crucial, transformamos los datos brutos en un formato adecuado para el entrenamiento del modelo.

Primero, la variable objetivo 'Churn' se convirtió de categórica ('Yes'/'No') a numérica (1/0). Luego, el conjunto de datos se dividió en entrenamiento y prueba, asegurando una estratificación adecuada para mantener la proporción de clientes con y sin churn. Las columnas

categóricas se manejaron mediante One-Hot Encoding para convertirlas en un formato numérico que los modelos pudieran interpretar, y las columnas 'tenure', 'MonthlyCharges' y 'TotalCharges' se normalizaron usando MinMaxScaler para escalar sus valores a un rango común.

Gráfico

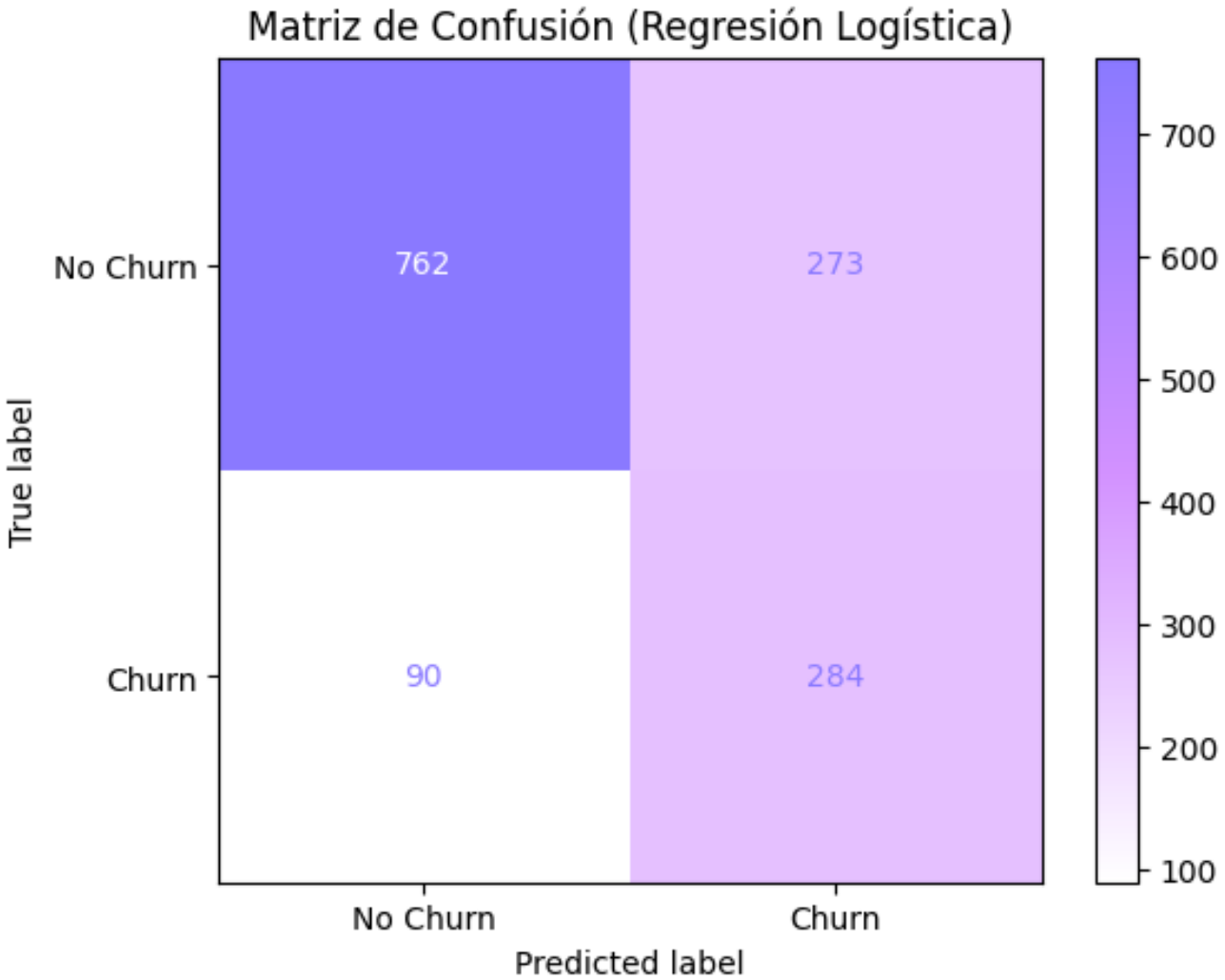


Finalmente, para abordar el desequilibrio de clases, aplicamos la técnica SMOTE. Esta generó nuevas muestras sintéticas de la clase minoritaria 'Churn', equilibrando el dataset y permitiendo que el modelo aprenda de manera más efectiva los patrones asociados al abandono de clientes.

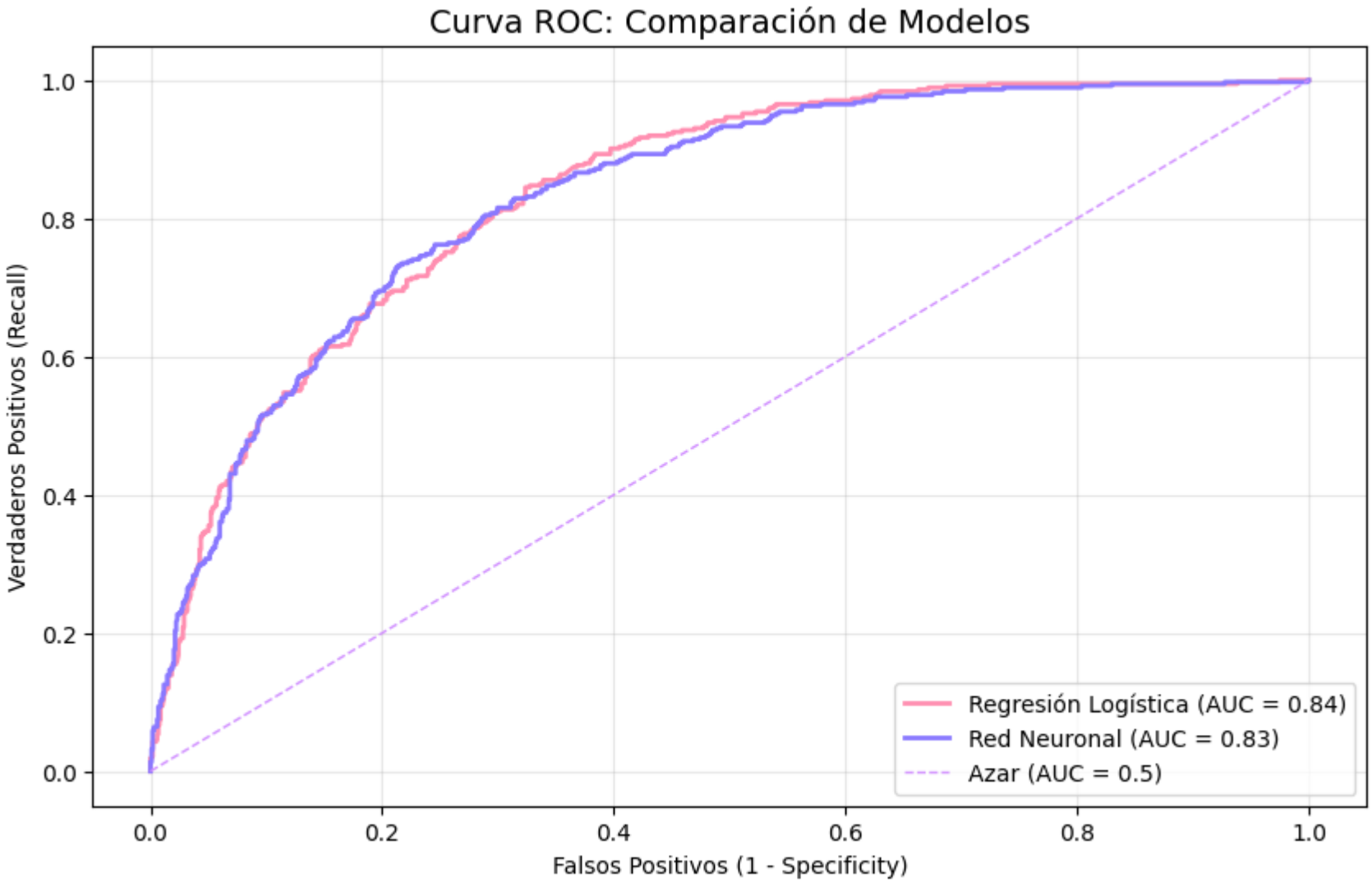
Evaluaciones de resultados (Uso de métricas).

	ROC-AUC	Precision (Churn)	Recall (Churn)	F1-Score (Churn)
LogisticRegression	0.840810	0.505942	0.796791	0.618899
PyTorch NN	0.833579	0.509338	0.802139	0.623053
XGBoost	0.829643	0.580402	0.617647	0.598446
RandomForest	0.827181	0.581218	0.612299	0.596354

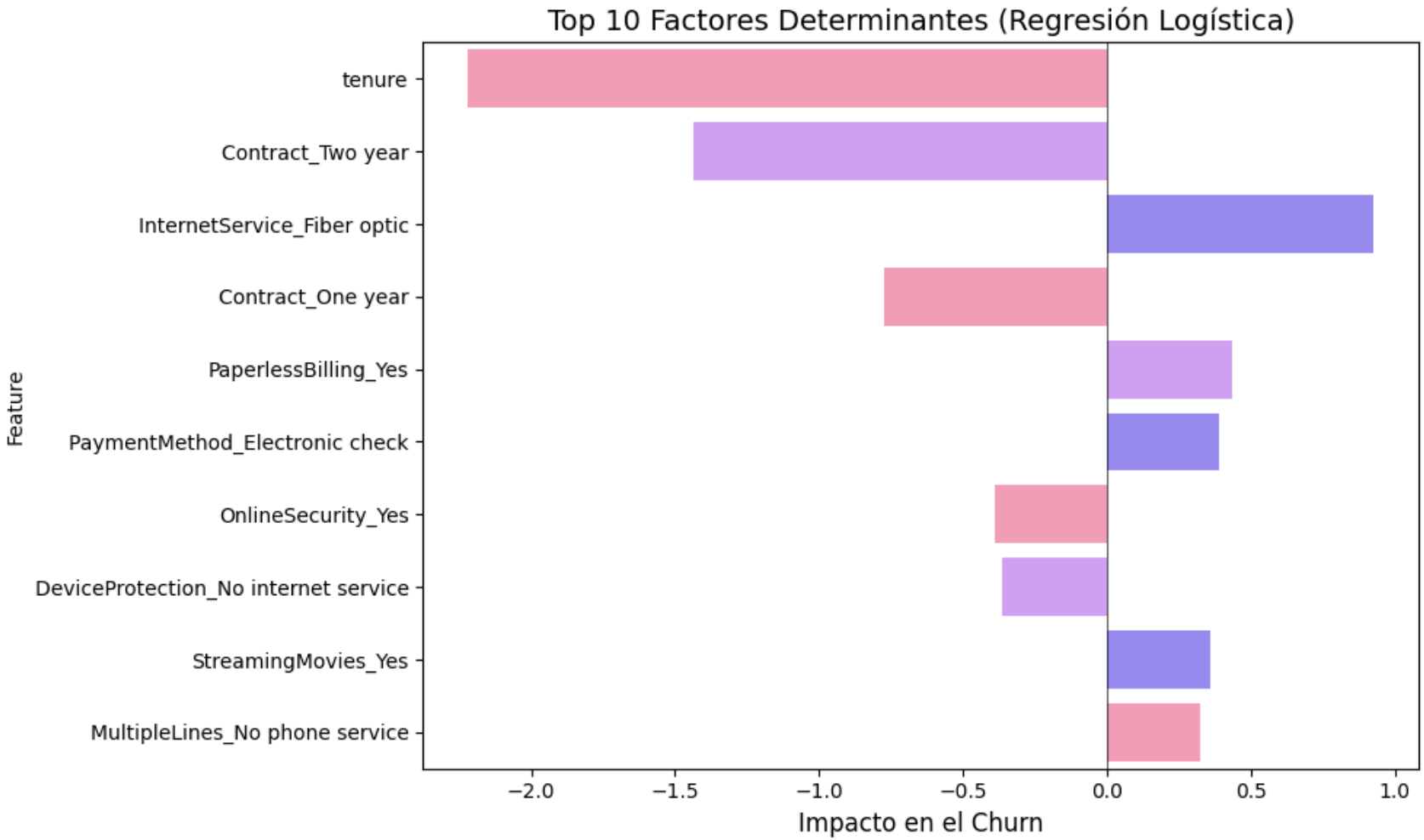
Matriz de Confusión (Regresión Logística)



Curva ROC


Curva ROC: Comparación de Modelos
Verdaderos Positivos (Recall)
Falsos Positivos (1 - Specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Regresión Logística (AUC = 0.84)
Red Neuronal (AUC = 0.83)
Azar (AUC = 0.5)

Top 10 factores determinantes (Regresión Logística)



Interpretación del modelo

Los modelos de Regresión Logística y Red Neuronal obtuvieron los mejores resultados de ROC-AUC (aproximadamente 0.84 y 0.83, respectivamente) y priorizan un alto recall para la clase Churn. Esto significa que son efectivos para identificar a la mayoría de los clientes propensos a abandonar el servicio, incluso si esto conlleva algunos falsos positivos.
La Regresión Logística destaca por su interpretabilidad, mostrando que una mayor antigüedad y contratos a largo plazo reducen el churn, mientras que servicios como fibra óptica, facturación electrónica y ciertos métodos de pago aumentan el riesgo de abandono.

Conclusiones y recomendaciones.

El modelo desarrollado presenta un rendimiento sólido, destacando la Regresión Logística por su buen desempeño e interpretabilidad. Los resultados indican que los clientes nuevos con contratos mensuales y ciertos servicios digitales tienen mayor riesgo de abandono.
Se recomienda a la empresa enfocar esfuerzos de retención en estos perfiles de riesgo, ofreciendo renovaciones de contrato, revisando la experiencia con fibra óptica o promoviendo métodos de pago más estables.
El modelo actual ofrece información valiosa, pero puede mejorarse con nuevas variables o ajustes.

Referencias

- **Dataset oficial:** BlastChar. (2018). *Telco Customer Churn*. Recuperado de Kaggle: https://www.kaggle.com/datasets/blastchar/telco-customer-churn%7D