PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE CIENCIAS
INSTITUTO DE ESTADÍSTICA

DESAFÍO DE CLASIFICACIÓN: PREDICCIÓN DE LA FUGA DE CLIENTES EN UNA EMPRESA DE TELECOMUNICACIONES

GUSTAVO ALEXANDER ALCÁNTARA ARAVENA

Profesor: MARIO GUZMÁN BRIONES

Diplomado DATA SCIENCE, Machine Learning, Inteligencia Artificial, Deep Learning (versión 22)

Valparaíso - Chile
2025

RESUMEN

Este trabajo aborda el desafío de predecir la fuga de clientes (churn) en una empresa de telecomunicaciones utilizando técnicas de aprendizaje automático. A partir de un conjunto de datos desbalanceado, se realiza un análisis exploratorio para identificar patrones relevantes asociados al abandono, como la tenencia de planes internacionales, el uso del buzón de voz y la cantidad de llamadas al servicio al cliente. Se desarrollan e implementan modelos de clasificación utilizando *Random Forest* y *SVM*, evaluando tres escenarios: sin balanceo, con SMOTE en ambos modelos, y una combinación de *Random Forest* sin SMOTE y *SVM* con SMOTE. Los resultados muestran que *Random Forest* sin SMOTE ofrece el mejor desempeño general, con alta precisión y capacidad discriminativa, sin necesidad de técnicas adicionales de balanceo al comparar ambos modelos.

Tabla de contenido

Índice de tablas

Índice de Ilustraciones

CAPÍTULO 1. INTRODUCCIÓN

1.1 Objetivo

- Desarrollar un modelo de clasificación que anticipe la fuga de clientes en una empresa de telecomunicaciones analizando el *dataset* 'churn-analysis.csv', aplicando técnicas de aprendizaje automático para la identificación de clientes con mayor riesgo de abandono.

1.2 Planteamiento del Problema

En el sector de las telecomunicaciones, la retención de clientes es un factor crítico para la vida y sostenibilidad económica de la organización. Los clientes poseen libertad para cambiar de empresa proveedora de telecomunicaciones y además competencia del mercado es alta, entonces, la fuga de clientes *'churn'* representa una constante amenaza para la estabilidad económica y la proyección comercial de la organización.

La capacidad de anticipar la fuga de clientes permitiría potencialmente implementar estrategias de retención más efectivas en el quehacer de la organización, focalizar las acciones comerciales y mejorar la experiencia del usuario. En este contexto, surge la necesidad de desarrollar herramientas analíticas que permitan identificar, a partir de la información histórica y de uso contenida en el archivo ***churn-analysis.csv***, aquellos perfiles con mayor propensión a abandonar la compañía.

Se requiere maximizar la capacidad predictiva del modelo, evaluando su desempeño mediante métricas estándar en problemas de clasificación, tales como la exactitud *(accuracy)* y el área bajo la curva *ROC* (AUC). Estas métricas permiten cuantificar la habilidad del modelo para discriminar correctamente entre

clientes que permanecerán (clase 0) y aquellos que presentarán *'churn'* (clase 1), garantizando así la utilidad y relevancia del sistema en un entorno real de toma de decisiones empresariales.

CAPÍTULO 2. METODOLOGÍA

A continuación se detalla el enfoque metodológico adoptado para abordar el problema de clasificación en el contexto de la predicción de la fuga de clientes en una empresa de telecomunicaciones.

2.1 Exploración del *dataset*

Se procedió a subir el conjunto de datos a *Google Drive* para garantizar la replicabilidad y reproducibilidad de la implementación. Estando en *Colab*, se continuó con la manipulación y el análisis inicial de los datos gracias a la librería *pandas*, por su eficiencia en el procesamiento y gestión estructurada de datos en formato tabular. La inspección preliminar, se realizó mediante funciones como *head()*, *info()* y *describe()*, lo que proporcionó una vista integral de la estructura del *dataset*, se identificaron los tipos de variables presentes, se verificó la existencia de valores nulos y se analizó la distribución estadística de los atributos presentes. Esta etapa exploratoria es vital para el flujo de trabajo posterior, puesto que facilita la detección temprana de problemas de calidad o integridad de los datos, orienta la selección de variables más pertinentes para el análisis y revela la necesidad de tratamiento o adecuaciones específicas, especialmente en el caso de variables categóricas.

2.2 Descripción de las Variables y Dicionario de Datos

2.3 Análisis Exploratorio

Se examinó tanto la distribución de las variables numéricas como categóricas, evaluando la proporción de clientes que abandonan (churn) y explorando

la relación entre los distintos atributos y la probabilidad de fuga. Esto incluyó la identificación de patrones, detección de valores atípicos (outliers) y análisis de posibles correlaciones, lo que permitió anticipar desafíos como el desbalance de clases o la multicolinealidad. El análisis exploratorio permitió revelar relaciones no evidentes a simple vista, lo que ayudó a identificar sesgos inherentes en los datos y orientó la selección de variables relevantes.

2.4 Preprocesamiento de Datos

El preprocesamiento de datos representa una etapa crítica en la construcción de modelos de machine learning, ya que determina en gran medida la calidad y relevancia de la información que será utilizada por los algoritmos. En este proyecto, se implementaron técnicas de codificación para transformar variables categóricas en representaciones numéricas, utilizando métodos como one-hot encoding y label encoding, lo que permite que los modelos interpreten correctamente la información discreta y aprovechen su valor predictivo. Asimismo, se realizó una depuración del conjunto de datos, eliminando atributos irrelevantes como el número de teléfono, cuya presencia no solo carece de significado predictivo, sino que también puede introducir ruido y dificultar la generalización del modelo. La ausencia de valores nulos en el dataset permitió simplificar el flujo de trabajo, enfocando los esfuerzos directamente en la optimización del modelo. Un preprocesamiento riguroso y bien fundamentado resulta clave para maximizar el aprendizaje de los algoritmos, mitigar el riesgo de sobreajuste y garantizar que el modelo final sea robusto, interpretable y aplicable en entornos reales de negocio.

2.5 Ingeniería de Características

La ingeniería de características constituye una etapa fundamental en el proceso de modelado predictivo, ya que permite transformar los datos originales en representaciones más informativas y útiles para los algoritmos de aprendi-

zaje automático. En este proyecto, se realizaron diversas acciones orientadas a enriquecer el conjunto de datos y facilitar el aprendizaje de los modelos de clasificación.

En primer lugar, se eliminaron atributos que no aportaban valor predictivo o podían inducir ruido en los modelos, como el número telefónico y el identificador del cliente. Esta depuración inicial permitió enfocar el análisis en variables relevantes.

Posteriormente, se aplicaron técnicas de codificación para representar variables categóricas de manera numérica. Se utilizó *one-hot encoding* para atributos con múltiples categorías discretas, como el estado de residencia del cliente, y *label encoding* para variables binarias como “International Plan” y “Voice Mail Plan”. Este tratamiento fue esencial para que los modelos pudieran procesar información cualitativa sin perder su valor semántico.

Además, se creó una nueva variable categórica derivada denominada `customer_service_cat`, que agrupa la cantidad de llamadas al servicio al cliente en rangos significativos (0–1, 2–3, y 4 o más). Esta transformación se realizó a partir de los hallazgos del análisis exploratorio, donde se observó que los clientes que realizaron cuatro o más llamadas presentaban una alta tasa de abandono. Esta nueva variable permitió capturar patrones de comportamiento relacionados con la insatisfacción de manera más explícita.

Finalmente, todas las variables numéricas fueron escaladas mediante *StandardScaler*, asegurando que los algoritmos sensibles a la magnitud de las variables, como SVM, pudieran operar de manera óptima. Esta normalización también ayudó a mitigar el impacto de valores extremos detectados previamente, especialmente en las variables de cargo y minutos internacionales.

2.6 Selección y Entrenamiento de Modelos

La etapa de selección y entrenamiento de modelos es fundamental para garantizar que la solución propuesta sea tanto precisa como adecuada al contexto del problema. En este proyecto, se optó por implementar y comparar diversos algoritmos de clasificación, específicamente regresión logística, árbol de decisión y *random forest*, cada uno con ventajas conceptuales distintas. La regresión logística, por su naturaleza lineal y su alta interpretabilidad, resulta especialmente útil para problemas binarios y para comprender el peso de cada variable en la predicción. El árbol de decisión, en cambio, permite modelar relaciones no lineales y facilita la interpretación visual de las reglas de decisión, lo que puede ser valioso para la comunicación con áreas de negocio. Por su parte, el *random forest*, al ser un ensamble de múltiples árboles, ofrece una mayor robustez frente al sobreajuste y suele lograr un mejor desempeño predictivo en escenarios complejos. Para evaluar la capacidad de generalización de cada modelo, se dividió el conjunto de datos en subconjuntos de entrenamiento y prueba, siguiendo las mejores prácticas de la disciplina. Esta estrategia no solo previene el sobreajuste, sino que también permite estimar de forma objetiva el rendimiento del modelo frente a nuevos datos, facilitando la elección del enfoque más alineado con los objetivos y limitaciones del negocio.

2.7 Evaluación de Modelos

La evaluación del desempeño de los modelos constituye un pilar esencial en el ciclo de vida de cualquier proyecto de machine learning, ya que permite cuantificar objetivamente la capacidad predictiva y la utilidad práctica de las soluciones desarrolladas. En este caso, se emplearon métricas complementarias como la exactitud (accuracy) y el área bajo la curva ROC (AUC), lo que posibilita una valoración integral del modelo desde distintas perspectivas. Mientras que la

exactitud ofrece una medida global de la proporción de predicciones correctas, la AUC evalúa la habilidad del modelo para discriminar entre las clases, aspecto crucial en escenarios donde existe un desbalance significativo, como ocurre con la predicción de churn. Este enfoque multifacético es conceptualmente robusto, ya que evita depender de una sola métrica —la cual podría ser engañosa en contextos de clases desbalanceadas— y favorece la identificación precisa de la clase minoritaria, es decir, los clientes con mayor riesgo de abandono. Así, la utilización de múltiples métricas no solo enriquece el análisis comparativo entre modelos, sino que también asegura que la solución seleccionada sea la más adecuada para los objetivos estratégicos del negocio.

2.8 Selección del Modelo Final e Interpretación de Resultados

La etapa final del proceso consistió en la selección del modelo óptimo, fundamentada en un análisis riguroso de las métricas de desempeño obtenidas durante la evaluación. Se priorizó no solo la precisión global del modelo, sino también su capacidad para discriminar correctamente entre clientes que abandonan y los que permanecen, asegurando así la relevancia práctica de la solución. Este enfoque permitió identificar el modelo que mejor equilibra desempeño predictivo, interpretabilidad y aplicabilidad en el contexto real de la empresa. Posteriormente, se analizaron las implicancias de los resultados, destacando cómo el modelo seleccionado puede integrarse en los procesos de toma de decisiones estratégicas, facilitando la anticipación de la fuga de clientes y la implementación de acciones proactivas de retención. La selección del modelo final no solo se basó en criterios técnicos, sino también en su capacidad de ser interpretado fácilmente y alinearse con los objetivos del negocio. Esto garantiza que los resultados sean comprensibles y accionables para las áreas encargadas de la gestión de clientes. De este modo, el modelo se convierte en una herramienta valiosa para transformar datos en conocimiento y conocimiento en decisiones efectivas.

CAPÍTULO 3. DESARROLLO

3.1 Análisis Exploratorio

3.1.1. Análisis descriptivo de la base de datos.

El análisis exploratorio de datos (EDA) constituye la primera etapa práctica del proyecto, permitiendo obtener una visión general de la estructura y características del conjunto de datos. Se utilizó la librería `pandas` para cargar y explorar el archivo `churn-analysis.csv`. Las funciones `head()`, `info()` y `describe()` permitieron identificar el número de observaciones, determinar el tipo de variables (numéricas y categóricas) y obtener estadísticas descriptivas como media, mediana, desviación estándar, así como los valores mínimos y máximos de las variables numéricas. Se identificaron las siguientes variables principales:

- **CustomerID**: Identificador único del cliente (variable irrelevante para el modelo).
- **Gender**: Género del cliente (categórica).
- **SeniorCitizen**: Indica si el cliente es adulto mayor (binaria).
- **Tenure**: Meses de permanencia del cliente (numérica).
- **MonthlyCharges**: Cargo mensual (numérica).
- **TotalCharges**: Cargo total acumulado (numérica).
- **Churn**: Variable objetivo, indica si el cliente abandonó la empresa (binaria: 1 = sí, 0 = no).

El dataset contiene un total de **3.333** registros y **20** variables. La variable objetivo (**Churn**) presenta una proporción de **14,5 %** de clientes que han abandonado la empresa, lo que corresponde a **483** casos, frente a **2.850** clientes

que permanecen. Este desequilibrio indica un potencial **desbalance de clases**, que debe considerarse al momento de entrenar modelos de clasificación, ya que podría sesgar los resultados hacia la clase mayoritaria si no se aplican técnicas adecuadas de balanceo o ajuste.

3.1.2. Visualizaciones y tablas relevantes.

3.1.2.1. Distribución de la variable objetivo.

Para analizar la proporción de clientes que permanecen frente a los que abandonan la empresa, se generó un gráfico de barras de la variable objetivo (Churn). La Figura 3.1 se aprecia un desbalance de clases, ya que la mayoría de los clientes corresponden a la clase de no abandono, mientras que una menor proporción representa a los clientes que efectivamente abandonaron la compañía.

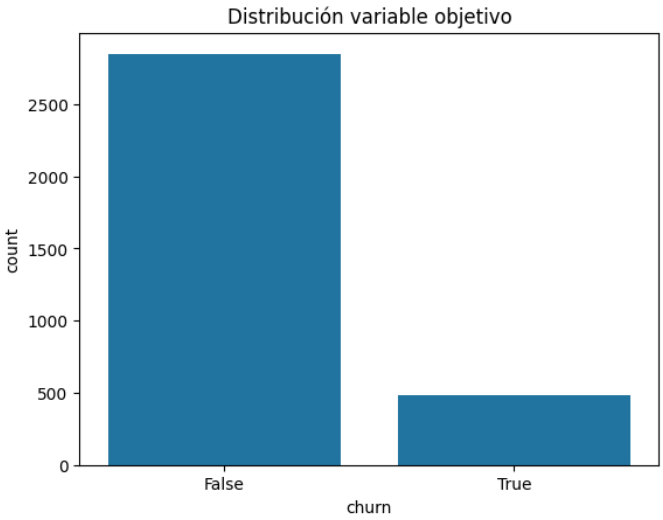

Figura 3.1: Distribución de la variable objetivo (Churn)

La Tabla 3.1 muestra la distribución de la variable objetivo *churn* en el conjunto de datos. Se observa un marcado desbalance de clases, ya que el 85.5 % de los clientes permanecen en la compañía, mientras que solo el 14.5 % han abandonado el servicio. Este desbalance es relevante para el análisis, ya que puede afectar

el desempeño de los modelos de clasificación y requiere considerar métricas de evaluación adecuadas y, en caso necesario, la aplicación de técnicas para el manejo de clases desbalanceadas.

Tabla 3.1: Proporción de clientes según la variable objetivo *churn*

Churn	**Proporción (%)**
No	85.51
Sí	14.49

3.1.2.2. Histogramas de variables numéricas.

Para analizar la distribución de las principales variables numéricas del conjunto de datos, se generaron histogramas para atributos como los minutos y cargos totales en distintos periodos del día (día, tarde, noche e internacional), el número de llamadas al servicio al cliente y la cantidad de mensajes de voz. Como se observa en la Figura 3.2, la mayoría de las variables presentan distribuciones concentradas en rangos específicos, aunque algunas muestran la presencia de valores atípicos o colas largas. Este análisis permitió identificar tendencias generales en el comportamiento de los clientes y detectar posibles anomalías que podrían influir en el desempeño de los modelos predictivos.

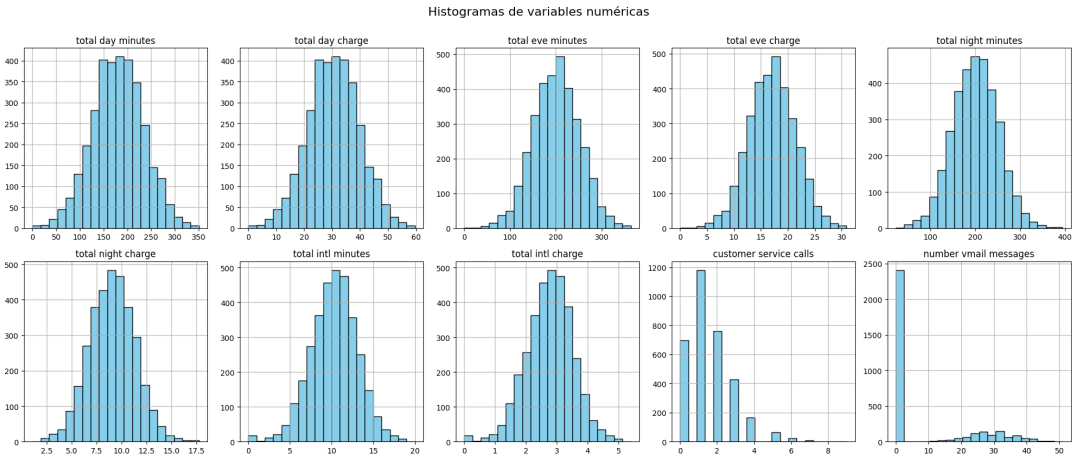

Figura 3.2: Histogramas de variables numéricas del dataset

3.1.2.3. Gráficos de cajas *(boxplots)*.

Para analizar la relación entre los cargos diarios, vespertinos y la probabilidad de abandono de clientes, se generaron *boxplots* de las variables ***'total day charge y total eve charge'*** diferenciados por la variable objetivo ***'churn'***. Como se aprecia en las Figuras 3.3 y 3.4, los clientes que abandonan tienden a presentar valores ligeramente superiores en los cargos diarios y vespertinos en comparación con quienes permanecen en la compañía. Además, los *boxplots* permiten identificar valores atípicos y observar la dispersión de los datos en ambos grupos, aportando información útil para detectar patrones relacionados con el *churn* y seleccionar variables predictivas.

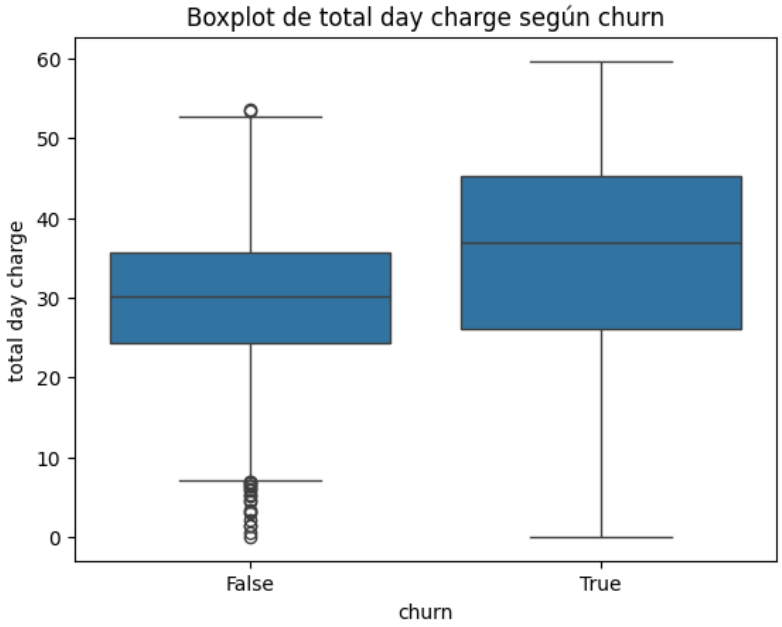

Figura 3.3: Boxplot de *'total day charge'* según *'churn'*

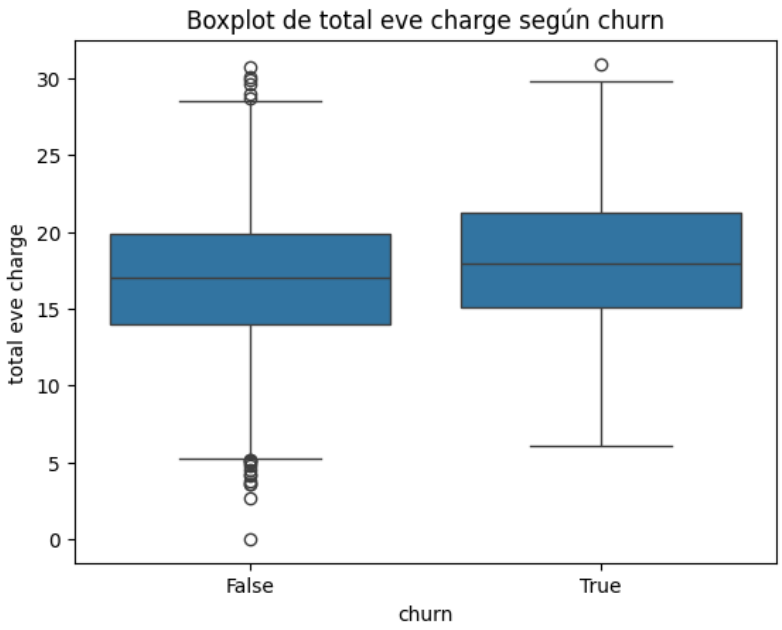

Figura 3.4: Boxplot de *'total eve charge'* según *'churn'*

3.1.2.4. Tablas de contingencia.

Se construyeron tablas de contingencia para analizar la relación entre diversas variables categóricas y la variable objetivo churn. Los resultados muestran que la proporción de clientes que abandonan la compañía es considerablemente mayor entre quienes poseen un plan internacional (42.4 %) en comparación con quienes no lo tienen (11.5 %), como se aprecia en la tabla 3,2. De manera similar, los clientes sin plan de correo de voz presentan una tasa de abandono superior (16.7 %) respecto a quienes sí cuentan con este servicio (8.7 %), según la tabla 3.3. Al analizar el número de llamadas al servicio al cliente, se observa que los clientes que han realizado cuatro o más llamadas tienen una probabilidad de abandono superior al 50 % (Tabla 3.4), lo que sugiere que la insatisfacción o problemas recurrentes pueden estar fuertemente asociados al churn. Por otro lado, la variable area code no muestra diferencias significativas en la tasa de abandono, mientras que la variable state revela cierta variabilidad regional, aunque en la

mayoría de los estados la proporción de churn se mantiene cercana al promedio general. Estos hallazgos permiten identificar factores de riesgo y segmentos de clientes con mayor propensión al abandono, lo que resulta fundamental para el diseño de estrategias de retención.

Debido a su tamaño, la tabla de contingencia completa por estado se adjunta en el Anexo 1 para consulta aún más detallada.

Tabla 3.2: Tabla de contingencia: International Plan vs Churn

International Plan	**No Churn**	**Churn**
No	0.885	0.115
Yes	0.576	0.424
Total	0.855	0.145

Tabla 3.3: Tabla de contingencia: Voice Mail Plan vs Churn

Voice Mail Plan	**No Churn**	**Churn**
No	0.833	0.167
Yes	0.913	0.087
Total	0.855	0.145

Tabla 3.4: Tabla de contingencia: Customer Service Calls (agrupado) vs Churn

Llamadas a Servicio	**No Churn**	**Churn**
0-1	0.886	0.114
2-3	0.890	0.110
4 o más	0.483	0.517
Total	0.855	0.145

La relación entre la tenencia de un plan internacional y la probabilidad de abandono se visualiza en la Figura 3.5, donde se observa que la proporción de clientes que abandonan es considerablemente mayor entre quienes poseen este tipo de plan. De manera similar, la Figura 3.6 muestra la relación entre el plan de

buzón de voz y el churn, evidenciando que los clientes sin este servicio presentan una mayor tasa de abandono.

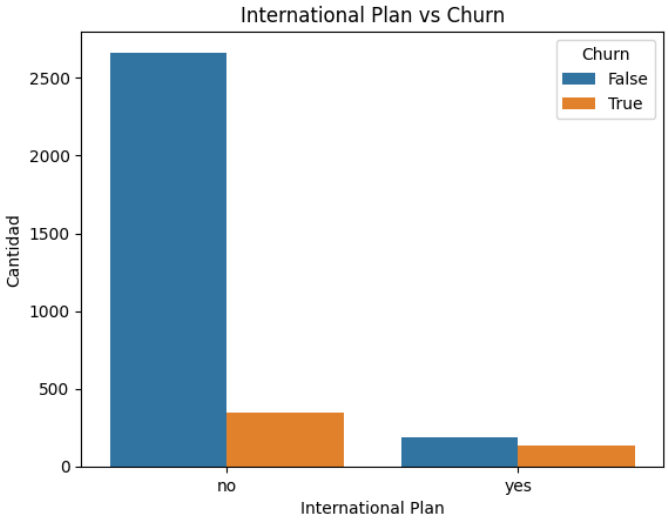

Figura 3.5: Relación entre plan internacional y churn

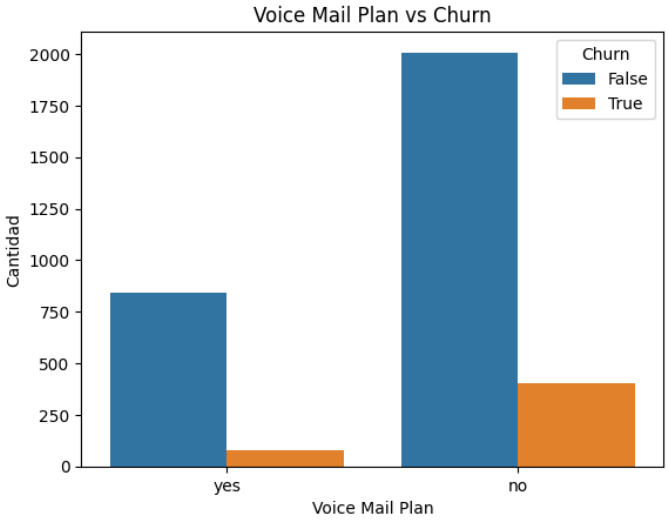

Figura 3.6: Relación entre plan de buzón de voz y churn

3.2 Evaluación de datos faltantes y/o atípicos

3.2.1. Revisión de datos faltantes y atípicos.

Para evaluar la calidad de los datos y la necesidad de aplicar técnicas de imputación, se realizó una revisión de valores faltantes en todas las variables del conjunto de datos. Como se observa en la Tabla 3.5, no se detectaron valores nulos en ninguna de las variables, lo que garantiza la integridad de la información y simplifica el proceso de análisis y modelado posterior.

Tabla 3.5: Cantidad de valores faltantes por variable

Variable	**Valores faltantes**
state	0
area code	0
phone number	0
international plan	0
voice mail plan	0
number vmail messages	0
total day minutes	0
total day calls	0
total day charge	0
total eve minutes	0
total eve calls	0
total eve charge	0
total night minutes	0
total night calls	0
total night charge	0
total intl minutes	0
total intl calls	0
total intl charge	0
customer service calls	0
churn	0
customer_service_cat	0

Dado que no se identificó valores faltantes en ninguna de las variables del dataset, no fue necesario aplicar técnicas de imputación ni eliminar registros

incompletos.

Para la detección de valores atípicos, se utilizó el rango intercuartílico (IQR) en las principales variables numéricas. El análisis identificó 25 outliers en total day charge, 24 en total eve charge, 30 en total night charge, 49 en total intl charge, 267 en customer service calls y 1 en number vmail messages.

En cuanto a los valores atípicos, se optó por mantenerlos en el análisis, ya que representan comportamientos reales de los clientes y pueden aportar información relevante para la predicción del churn. Además, la proporción de outliers detectados fue baja en la mayoría de las variables, por lo que no se consideró que afectaran negativamente la robustez de los modelos de machine learning.

La mayor cantidad de valores atípicos se observó en la variable customer service calls, lo que puede estar asociado a clientes con comportamientos inusuales o necesidades de atención recurrentes. En general, la proporción de outliers respecto al total de registros es baja, salvo en el caso de las llamadas al servicio al cliente, donde estos casos extremos pueden ser relevantes para la predicción del churn, como se resume en la tabla 3.6.

Tabla 3.6: Cantidad de valores atípicos detectados por variable numérica

Variable	**Cantidad de outliers**
total day charge	25
total eve charge	24
total night charge	30
total intl charge	49
customer service calls	267
number vmail messages	1

La presencia y distribución de estos valores atípicos puede observarse en los boxplots de las principales variables numéricas diferenciadas por la variable objetivo (Figuras 3.7-10).

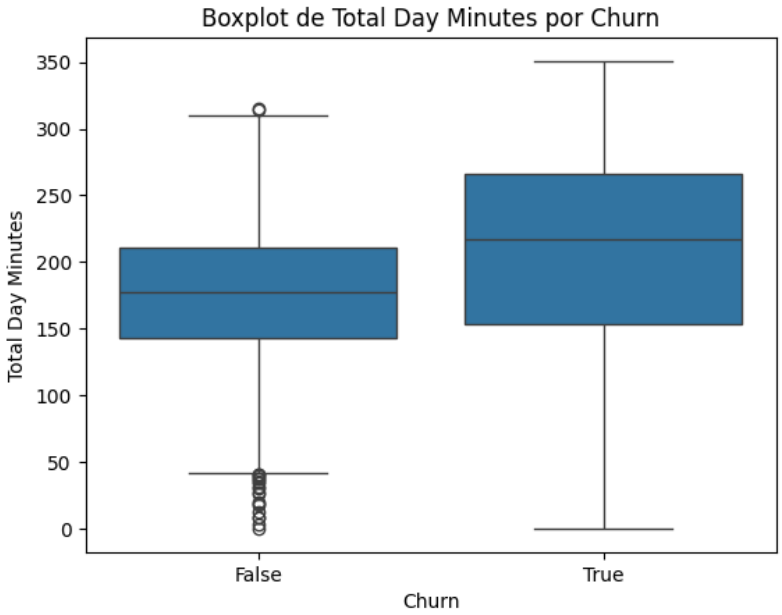

Figura 3.7: Boxplot de Total Day Minutes según churn

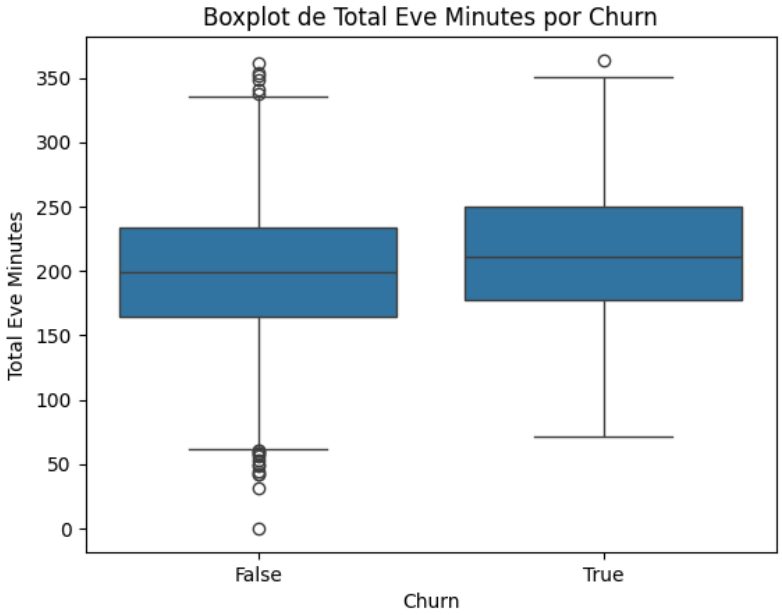

Figura 3.8: Boxplot de Total Eve Minutes según churn

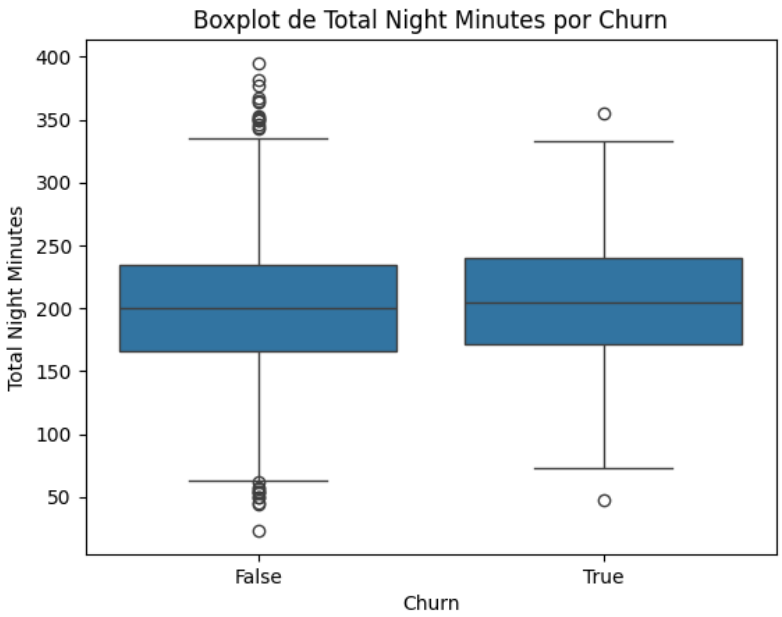

Figura 3.9: Boxplot de Total Night Minutes según churn

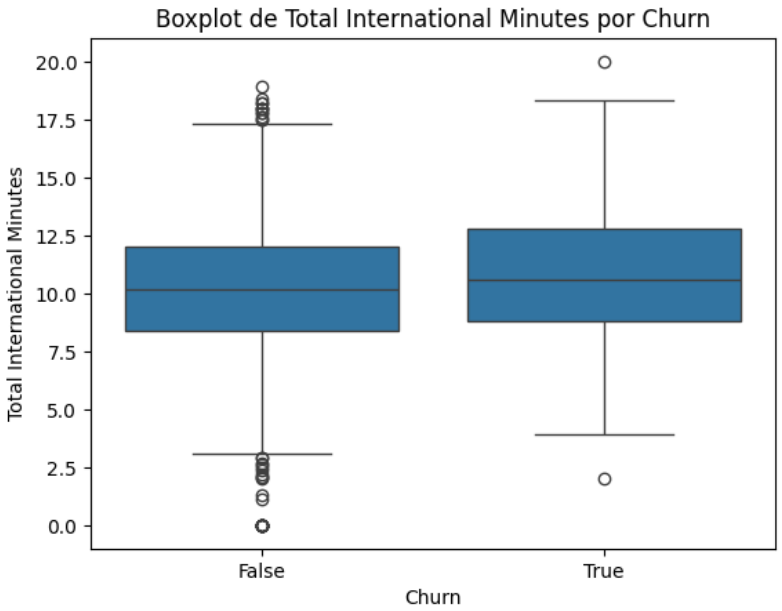

Figura 3.10: Boxplot de Total International Minutes según churn

Los boxplots presentados en las Figuras 3.7 a 3.10 permiten visualizar la distribución de las principales variables numéricas en función de la variable

objetivo *churn*. Se observa que, en general, los clientes que abandonan tienden a presentar valores ligeramente superiores en minutos utilizados, especialmente en las llamadas diurnas e internacionales. Además, la presencia de valores atípicos es más notoria en las variables relacionadas con cargos internacionales y llamadas al servicio al cliente, lo que sugiere la existencia de clientes con patrones de uso extremos o necesidades particulares. Estas diferencias en la distribución y la presencia de outliers pueden aportar información relevante para la predicción del churn y la identificación de segmentos de clientes con mayor riesgo de abandono.

3.3 Modelos implementados

En esta sección se presentan los modelos de clasificación desarrollados para predecir el abandono de clientes (*churn*) en una compañía de telecomunicaciones. Se evaluaron distintos enfoques combinando algoritmos y estrategias de balanceo, con el objetivo de comprender mejor el comportamiento del conjunto de datos desbalanceado y observar el impacto de aplicar la técnica SMOTE (Synthetic Minority Over-sampling Technique).

En particular, se implementaron los modelos ***Random Forest*** y **Support Vector Machine (*SVM*)** bajo tres enfoques distintos: (1) entrenamiento sin SMOTE, (2) entrenamiento con SMOTE aplicado a ambos modelos, y (3) un enfoque mixto en el cual se entrena *Random Forest* sin SMOTE y *SVM* con SMOTE. A continuación, se describen cada una de estas implementaciones de forma detallada.

3.3.1. Primera implementación: Sin SMOTE

3.3.1.1. Configuración y entrenamiento

Ambos modelos se entrenaron sobre los datos originales sin aplicar ninguna técnica de balanceo. Las variables fueron escaladas utilizando `StandardScaler`. Los hiperparámetros utilizados fueron:

- *Random Forest*: `n_estimators=100`, `random_state=42`
- *SVM*: `kernel='rbf'`, `C=1.0`, `gamma='scale'`, `probability=True`, `random_state=42`

3.3.1.2. Resultados

Tabla 3.7: Métricas sin SMOTE

Modelo	Accuracy	Precisión (churn)	Recall (churn)	F1 (churn)
Random Forest	0.95	0.92	0.70	0.80
SVM	0.92	0.88	0.50	0.63

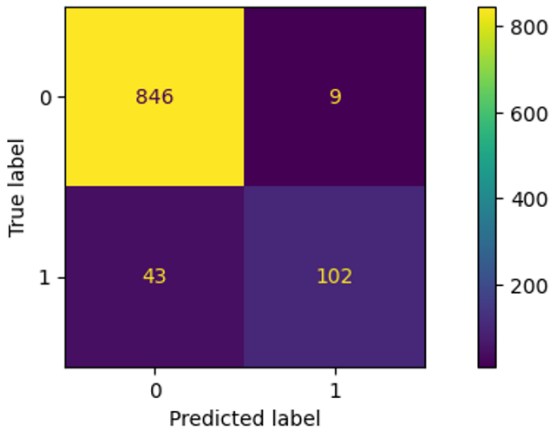

Figura 3.11: Matriz de confusión de *Random Forest* sin SMOTE

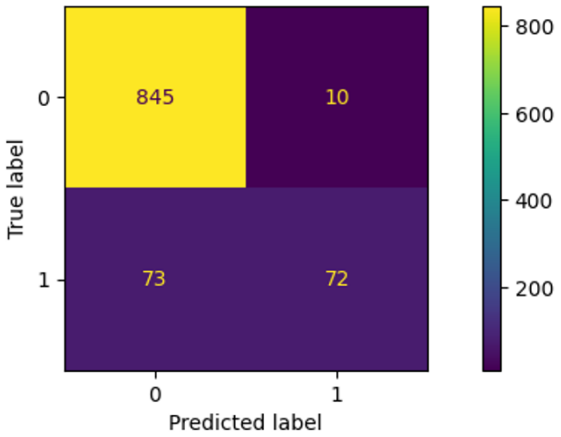

Figura 3.12: Matriz de confusión de *SVM* sin SMOTE

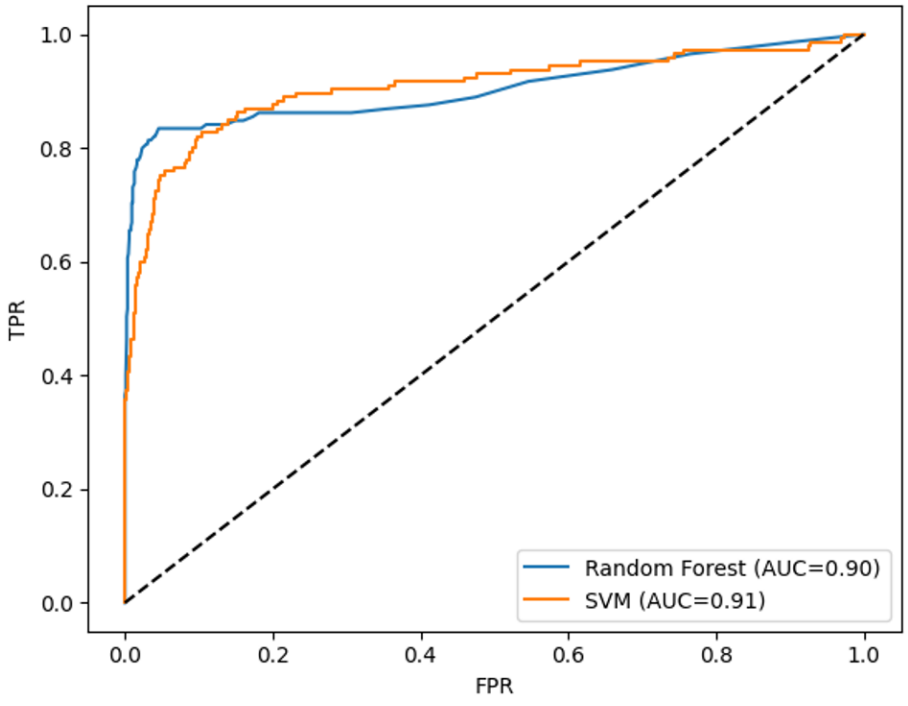

Figura 3.13: Curva ROC comparativa

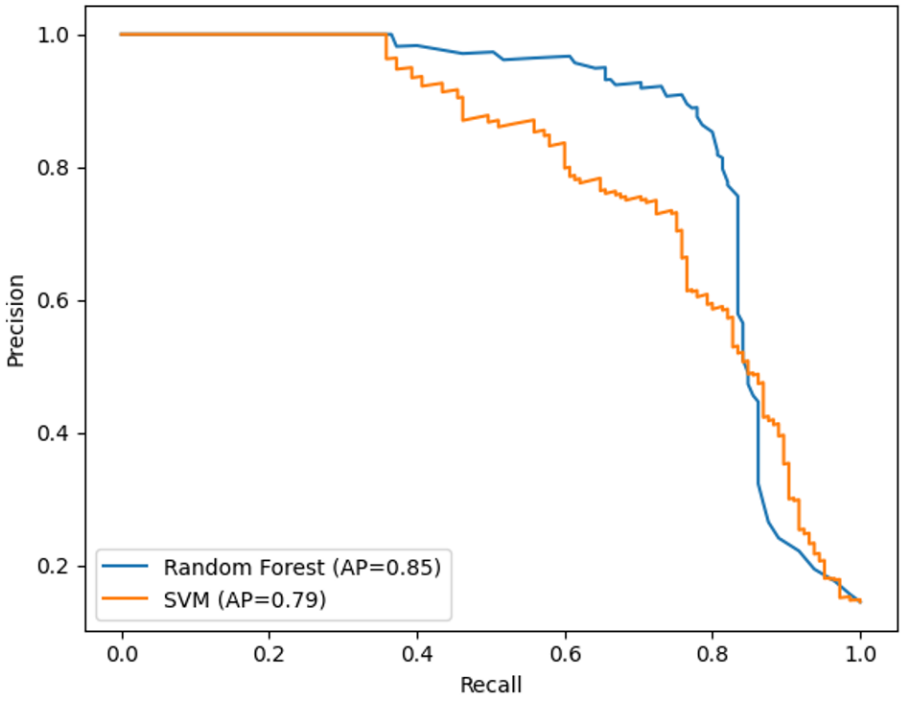

Figura 3.14: Curva Precision-Recall comparativa

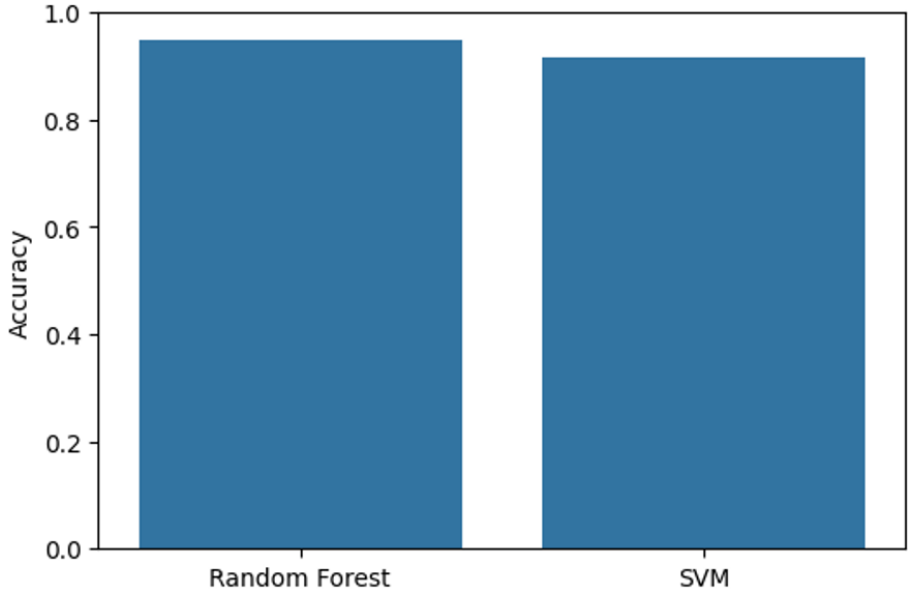

Figura 3.15: Comparación de Accuracy entre modelos

En la primera implementación, se entrenaron los modelos *Random Forest* y *Support Vector Machine (SVM)* sobre el conjunto de datos original sin aplicar técnicas de balanceo. La matriz de confusión del modelo *Random Forest* presente en la Figura 3.11 evidencia un desempeño robusto, con 846 verdaderos negativos y 102 verdaderos positivos, cometiendo solo 52 errores (43 falsos negativos y 9 falsos positivos). En contraste (figura 3.12), el modelo SVM presenta 845 verdaderos negativos y 72 verdaderos positivos, pero comete 83 errores (73 falsos negativos y 10 falsos positivos), lo que indica una menor capacidad para detectar la clase minoritaria (clientes que abandonan).

La curva ROC (Figura 3.13) muestra un área bajo la curva (AUC) de 0,90 para *Random Forest* y 0,91 para SVM, lo que sugiere que ambos modelos poseen buena capacidad discriminativa. Sin embargo, al observar la Figura 3.14, la curva *Precision-Recall*, *Random Forest* supera a SVM con un promedio de precisión (AP) de 0,85 frente a 0,79, lo que indica un mejor rendimiento al clasificar correctamente los casos de *churn*, especialmente en un contexto de clases desbalanceadas. La métrica de *accuracy* (Figura 3.15) confirma el mejor rendimiento de *Random Forest*, con un valor cercano al 95 %, frente al 92 % de *SVM*. Estos resultados refuerzan que, sin técnicas de balanceo, *Random Forest* logra una mayor eficacia y equilibrio en la detección de clientes propensos al abandono.

3.3.2. Segunda implementación: SMOTE aplicado a ambos modelos

3.3.2.1. Configuración y entrenamiento

En esta etapa, se aplicó SMOTE al conjunto de entrenamiento antes de ajustar ambos modelos. Se utilizó el mismo conjunto de hiperparámetros que en la primera implementación.

A continuación se resumen los hiperparámetros seleccionados para

cada modelo en la segunda implementación. Se mantuvieron constantes aquellos parámetros no modificados manualmente, confiando en los valores por defecto de las bibliotecas.

Segunda implementación (SMOTE aplicado a ambos modelos)

- ***Random Forest***:
  - `n_estimators=100`
  - `random_state=42`
  - Mismos valores que en la primera implementación
- **SVM**:
  - `kernel='rbf'`
  - `C=1.0`
  - `gamma='scale'`
  - `probability=True`
  - `random_state=42`

3.3.2.2. Resultados

Tabla 3.8: Métricas con SMOTE en ambos modelos

Modelo	**Accuracy**	**Precisión (churn)**	**Recall (churn)**	**F1 (churn)**
Random Forest	0.92	0.70	0.74	0.72
SVM	0.87	0.54	0.66	0.59

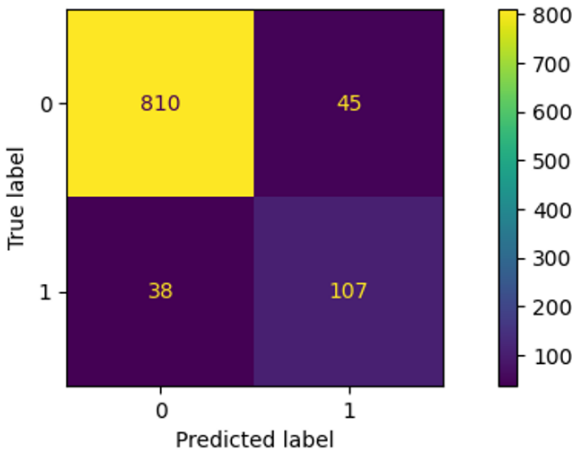

Figura 3.16: Matriz de Confusión - *Random Forest* (SMOTE)

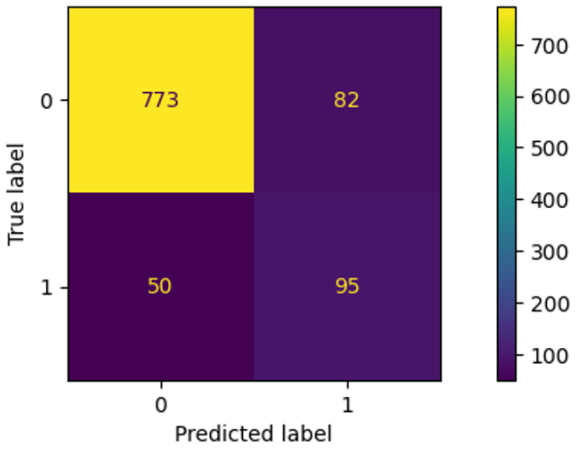

Figura 3.17: Matriz de Confusión - SVM (SMOTE)

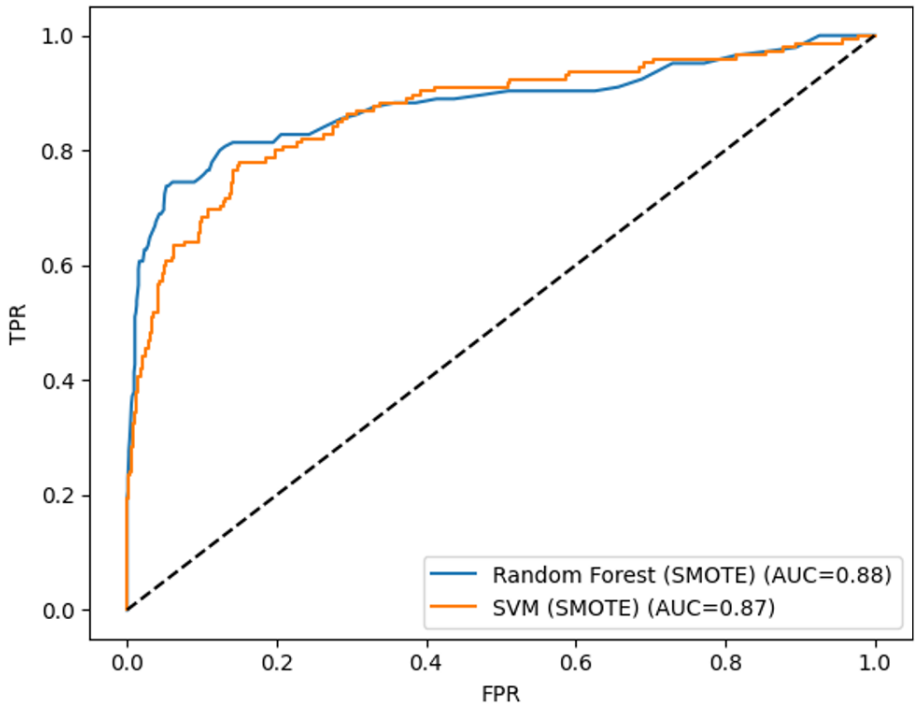

Figura 3.18: Curva ROC Comparativa (ambas con SMOTE)

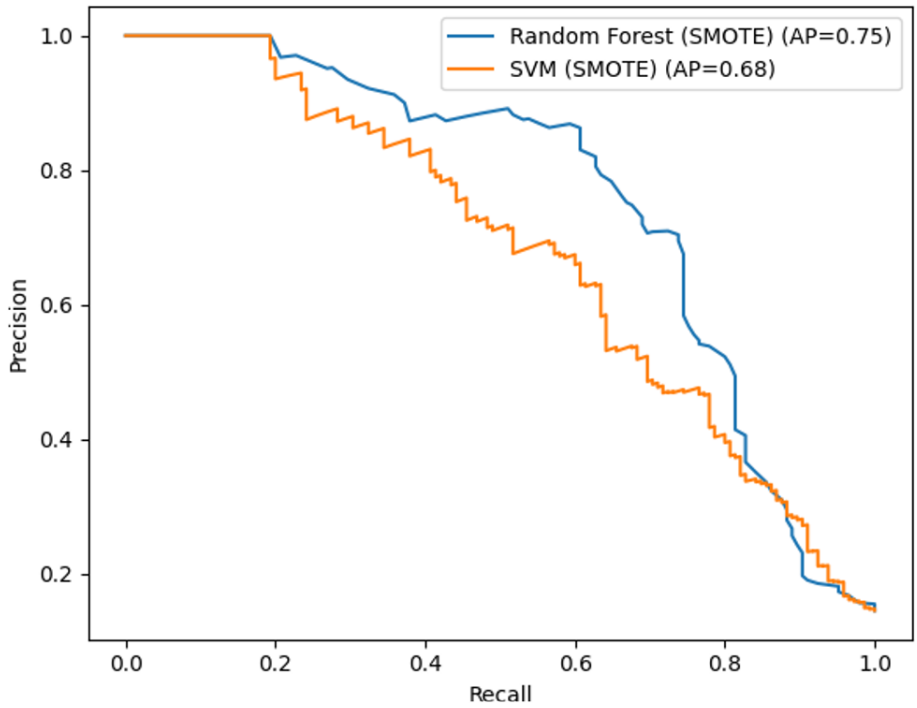

Figura 3.19: Curva Precision-Recall comparativa (de ambos modelos con SMOTE)

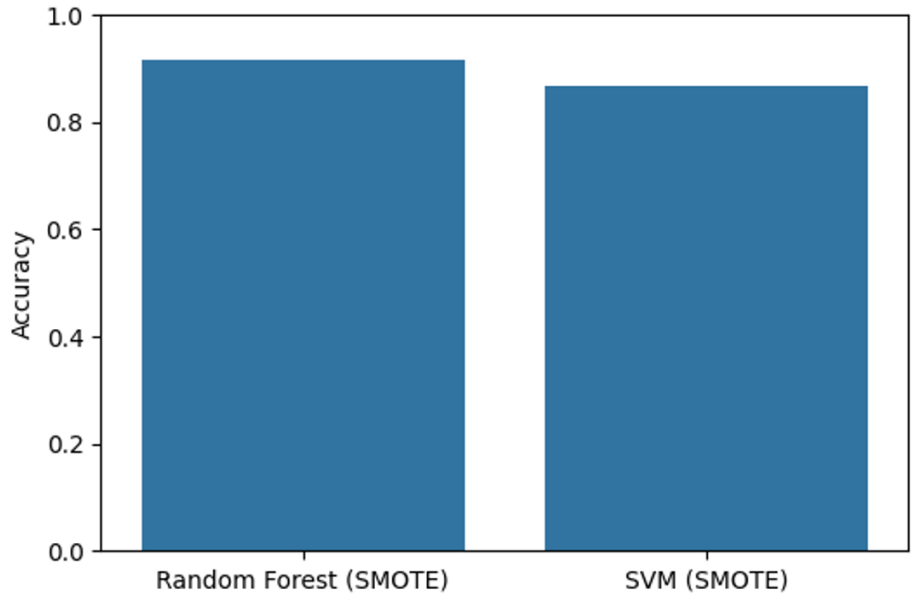

Figura 3.20: Accuracy de ambos modelos con SMOTE

En esta segunda etapa se aplicó la técnica de sobremuestreo *SMOTE* al conjunto de entrenamiento antes de ajustar los modelos *Random Forest* y *SVM*, con el objetivo de mitigar el impacto del desbalance de clases. A partir de las matrices de confusión (Figuras 3.16 y 3.17), se observa que *Random Forest* obtuvo 810 verdaderos negativos y 107 verdaderos positivos, mientras que SVM alcanzó 773 verdaderos negativos y 95 verdaderos positivos. No obstante, SVM también incurrió en un mayor número de falsos positivos (82) y falsos negativos (50) en comparación con *Random Forest* (45 y 38 respectivamente).

En cuanto al desempeño global, en base a la Figura 3.20 ambos modelos experimentaron una leve disminución en la métrica de *accuracy* respecto a la primera implementación. *Random Forest* mantuvo un valor alto (92 %), mientras que SVM descendió a un 87 %, según lo observado en el gráfico de barras comparativo. La curva *Precision-Recall* (Figura 3.19) muestra que el área promedio (AP) para *Random Forest* fue de 0,75, superando a SVM que alcanzó un 0,68, lo cual evidencia una mejor capacidad para identificar correctamente los casos positivos (clientes que abandonan), incluso tras aplicar SMOTE. En la curva *ROC*

(Figura 3.18), ambos modelos presentan un desempeño similar, con AUC de 0,88 para *Random Forest* y 0,87 para SVM.

En conjunto, los resultados indican que aunque la aplicación de SMOTE redujo levemente la exactitud general, permitió mejorar el *recall* y el equilibrio entre clases. Aun así, *Random Forest* mantiene su superioridad frente a SVM en todas las métricas clave, siendo más robusto frente al nuevo balance del conjunto de datos.

3.3.3. Tercera implementación: RF sin SMOTE / SVM con SMOTE

3.3.3.1. Configuración y entrenamiento

Esta implementación busca lo mejor de ambos enfoques: se entrena *Random Forest* sin aplicar SMOTE, aprovechando su robustez frente a clases desbalanceadas, mientras que SVM se entrena con SMOTE para mejorar su capacidad de detección de la clase minoritaria. Se mantienen los mismos hiperparámetros de las implementaciones anteriores.

A continuación se resumen los hiperparámetros seleccionados para cada modelo en la tercera implementación. Se mantuvieron constantes aquellos parámetros no modificados manualmente, confiando en los valores por defecto de las bibliotecas.

Tercera implementación (*Random Forest* sin SMOTE / SVM con SMOTE)

- ***Random Forest***:
  - `n_estimators=100`
  - `random_state=42`
  - Igual configuración que en las implementaciones anteriores

- **SVM (con SMOTE)**:
  - `kernel='rbf'`
  - `C=1.0`
  - `gamma='scale'`
  - `probability=True`
  - `random_state=42`

3.3.3.2. Resultados

Tabla 3.9: Métricas en implementación combinada

Modelo	Accuracy	Precisión (churn)	Recall (churn)	F1 (churn)
RF (sin SMOTE)	0.95	0.92	0.70	0.80
SVM (con SMOTE)	0.87	0.54	0.66	0.59

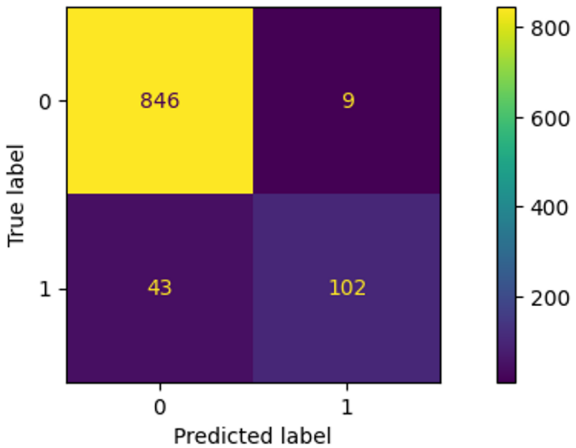

Figura 3.21: Matriz de confusión - *Random Forest* (sin SMOTE)

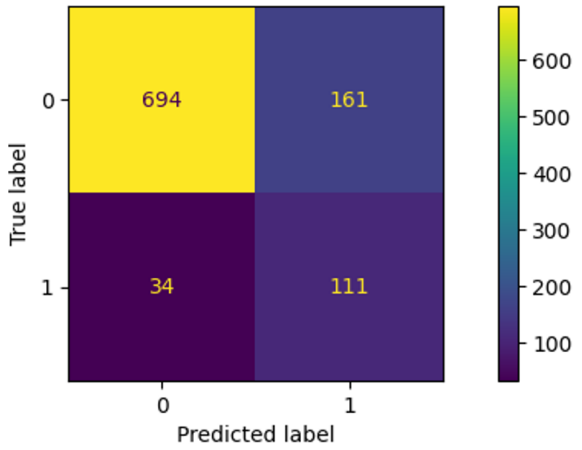

Figura 3.22: Matriz de confusión - SVM (con SMOTE)

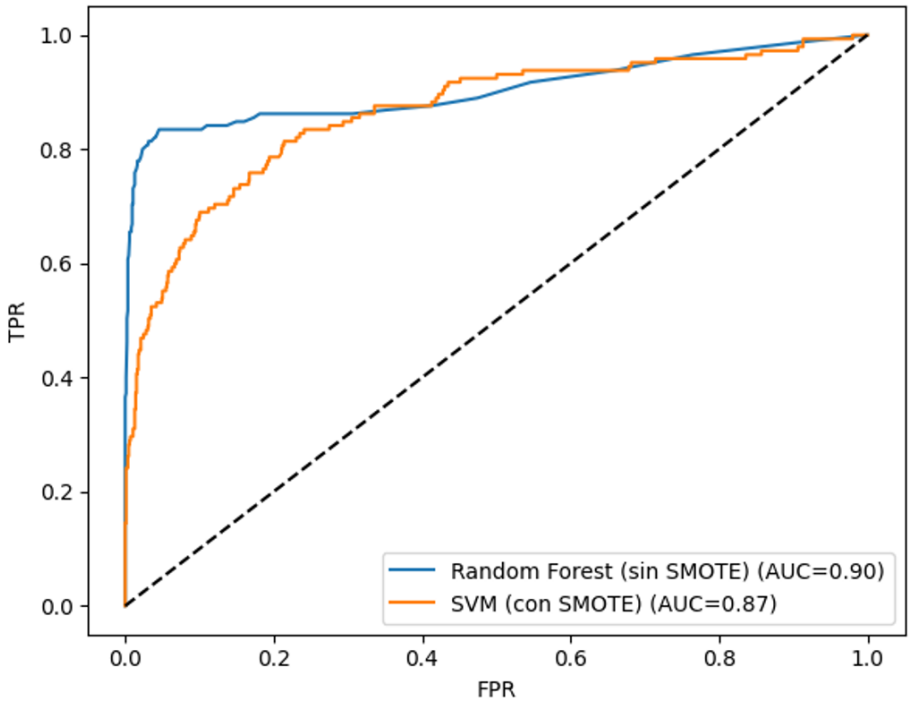

Figura 3.23: Curva ROC comparativa

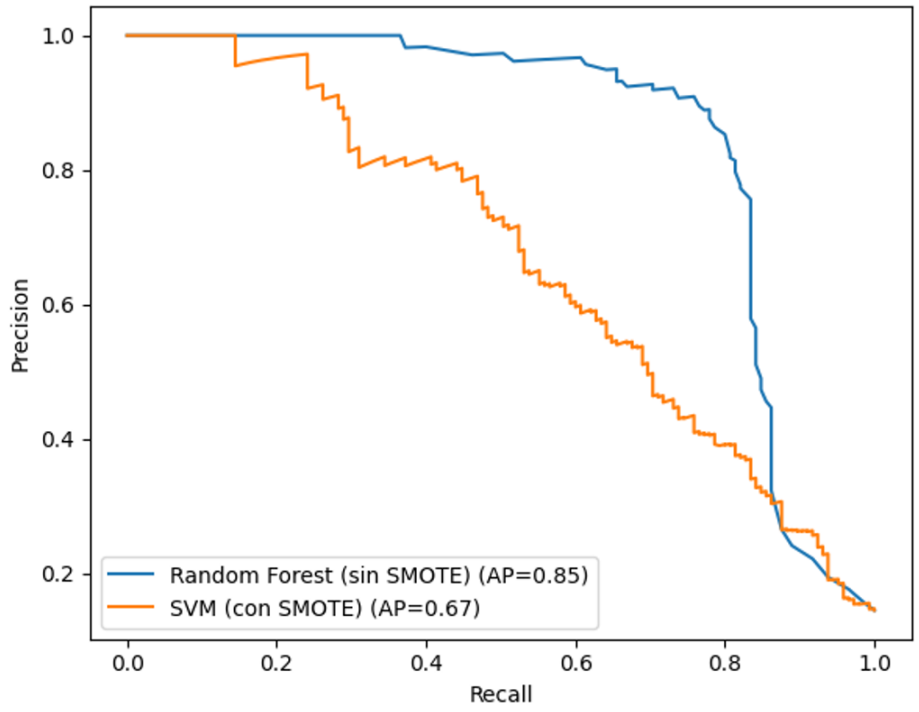

Figura 3.24: Curva Precision-Recall comparativa RF (sin SMOTE) y SVM (con SMOTE)

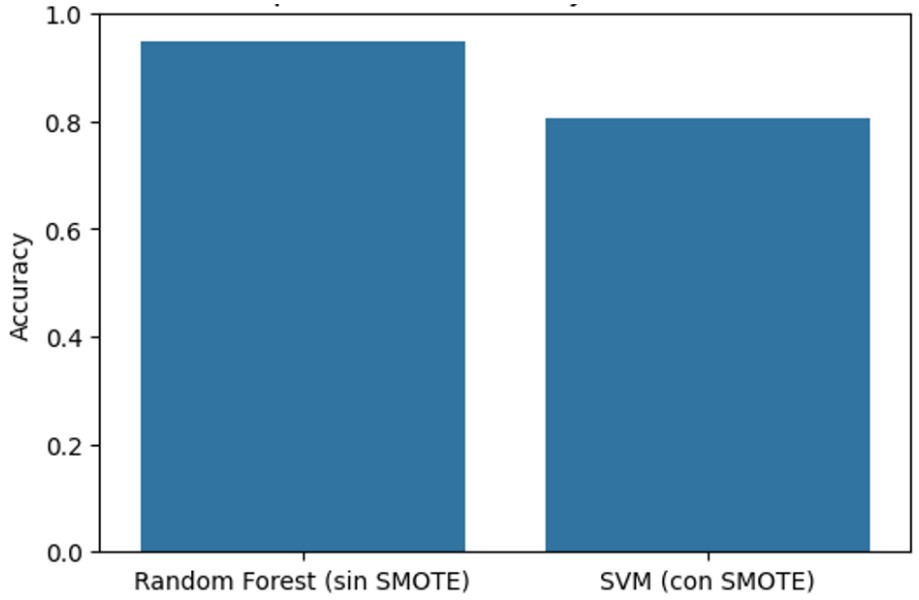

Figura 3.25: Comparación de Accuracy entre RF (sin SMOTE) y SVM (con SMOTE)

La tercera implementación se diseñó con un enfoque combinado: se entrenó el modelo *Random Forest* sin aplicar SMOTE, aprovechando su capacidad para manejar clases desbalanceadas, y se aplicó SMOTE exclusivamente al modelo *SVM* con el fin de mejorar su sensibilidad ante la clase minoritaria. Los resultados de esta configuración muestran que *Random Forest* obtuvo 846 verdaderos negativos (Figura 3.21) y 102 verdaderos positivos, con un total de 52 errores (43 falsos negativos y 9 falsos positivos), lo cual se traduce en un rendimiento robusto y equilibrado. Por otro lado, SVM, a pesar del uso de SMOTE, en la Figura 3.22 se aprecia que cometió un número significativo de falsos positivos (161) y falsos negativos (34), con solo 694 verdaderos negativos y 111 verdaderos positivos.

Las métricas gráficas refuerzan esta diferencia de desempeño: De acuerdo con la Figura 3.25 el modelo *Random Forest* mantiene un *accuracy* de 95 %, mientras que SVM desciende a aproximadamente 81 %. La curva *Precision-Recall* presente en la Figura 3.24 muestra un área promedio (AP) de 0,85 para *Random Forest*, muy superior al 0,67 alcanzado por SVM. Asimismo, la curva *ROC* revela un AUC de 0,90 para *Random Forest* frente a 0,87 para SVM, confirmando que el modelo sin SMOTE es más eficaz para discriminar correctamente entre clases.

En conclusión, esta implementación evidencia que el modelo *Random Forest* es suficientemente robusto para manejar el desbalance de clases sin la necesidad de aplicar técnicas de sobremuestreo, y que la aplicación aislada de SMOTE a SVM no logra superar el desempeño del enfoque tradicional con Random Forest. Por tanto, para este caso específico, *Random Forest* sin *SMOTE* se consolida como la estrategia más eficiente.

3.3.4. Comparación entre modelos

La comparación entre los distintos modelos implementados revela diferencias sustantivas en su capacidad para predecir correctamente la fuga de clientes. El modelo *Random Forest* entrenado sin aplicar técnicas de balanceo obtuvo el

mejor rendimiento general, alcanzando una exactitud de 95 % y un F1-score de 0.80 para la clase minoritaria (*churn*). Si bien la aplicación de SMOTE permitió mejorar el *recall* en algunos casos —especialmente en el modelo SVM—, esta técnica también implicó una reducción en la precisión general. El enfoque híbrido, que combinó *Random Forest* sin SMOTE con *SVM* con SMOTE, no logró superar el desempeño del modelo original de *Random Forest*, confirmando su robustez incluso frente a datos desbalanceados. Estos resultados sugieren que la elección del algoritmo tiene un impacto más determinante que la estrategia de balanceo, al menos en este conjunto de datos específico.

Tabla 3.10: Comparación de modelos implementados (indica uso de SMOTE)

Modelo	**Accuracy**	**Precision (churn)**	**Recall (churn)**	**F1 (churn)**	**SMOTE**
		Implementación N°1			
Random Forest	0.95	0.92	0.70	0.80	
SVM	0.92	0.88	0.50	0.63	
		Implementación N°2			
Random Forest	0.92	0.70	0.74	0.72	✓
SVM	0.87	0.54	0.66	0.59	✓
		Implementación N°3			
Random Forest	0.95	0.92	0.70	0.80	
SVM	0.87	0.54	0.66	0.59	✓

CAPÍTULO 4. CONCLUSIÓN

El presente proyecto ha abordado el problema de predicción de la fuga de clientes (churn) en una empresa de telecomunicaciones, aplicando técnicas de aprendizaje automático sobre un conjunto de datos real y desbalanceado. A través de una metodología sistemática y la implementación de modelos robustos, se logró identificar patrones relevantes y construir herramientas predictivas con aplicación directa en contextos empresariales.

4.1 Principales hallazgos

Se identificó un desbalance significativo en la variable objetivo (*churn*), con un 14.5 % de clientes que efectivamente abandonaron el servicio. Esta situación motivó el uso de técnicas de balanceo como SMOTE, permitiendo evaluar el impacto en la capacidad predictiva de los modelos.

El análisis exploratorio reveló que variables como la contratación de planes internacionales, la ausencia de buzón de voz y la cantidad de llamadas al servicio al cliente se correlacionan fuertemente con una mayor probabilidad de abandono. Estas relaciones fueron respaldadas por análisis gráficos (boxplots, histogramas) y tablas de contingencia.

En cuanto al rendimiento de los modelos, ***Random Forest* sin SMOTE** obtuvo los mejores resultados generales, con una *accuracy* de 95 % y un *F1-score* de 0.80 para la clase minoritaria. Aunque el uso de SMOTE permitió mejorar el *recall* en algunos modelos, también redujo la precisión general, evidenciando la necesidad de evaluar cuidadosamente el trade-off entre sensibilidad y especificidad.

4.2 Recomendaciones para el negocio

Se recomienda la adopción del modelo de ***Random Forest* sin SMOTE** como herramienta de apoyo a la toma de decisiones en estrategias de retención. Este modelo demostró una alta capacidad discriminativa sin necesidad de técnicas de sobremuestreo, lo que simplifica su integración operativa.

La compañía debería focalizar sus acciones en los segmentos identificados con mayor riesgo de abandono, especialmente aquellos con alta interacción con el servicio al cliente, usuarios de planes internacionales y quienes no utilizan buzón de voz. Se sugiere además monitorear continuamente estas variables, integrando el modelo en un pipeline automatizado que permita actualizaciones periódicas y predicciones en tiempo real.

4.3 Trabajo futuro

Como trabajo futuro, se propone incorporar técnicas de *feature selection* y *hyperparameter tuning* más avanzadas, tales como *Grid Search*, *Bayesian Optimization* o algoritmos evolutivos, que podrían mejorar aún más el rendimiento del modelo sin comprometer su interpretabilidad.

Asimismo, se plantea explorar enfoques más sofisticados para el manejo del desbalance, como *ensemble methods* especializados (e.g., *Balanced Random Forest, EasyEnsemble*), así como evaluar el uso de redes neuronales profundas y modelos de *gradient boosting* (como XGBoost o LightGBM) para comparar su desempeño frente a los modelos tradicionales.

CAPÍTULO 5. ANEXOS

5.1 Tabla de contingencia: State versus Churn

Tabla 5.1: Tabla de contingencia: State vs Churn

Estado	No Churn	Churn
AK	0.942	0.058
AL	0.900	0.100
AR	0.800	0.200
AZ	0.938	0.063
CA	0.735	0.265
CO	0.864	0.136
CT	0.838	0.162
DC	0.907	0.093
DE	0.852	0.148
FL	0.873	0.127
GA	0.852	0.148
HI	0.943	0.057
IA	0.932	0.068
ID	0.877	0.123
IL	0.914	0.086
IN	0.873	0.127
KS	0.814	0.186
KY	0.864	0.136

Continúa en la siguiente página

Tabla 5.1 – Continuación

Estado	No Churn	Churn
LA	0.922	0.078
MA	0.831	0.169
MD	0.757	0.243
ME	0.790	0.210
MI	0.781	0.219
MN	0.821	0.179
MO	0.889	0.111
MS	0.785	0.215
MT	0.794	0.206
NC	0.838	0.162
ND	0.903	0.097
NE	0.918	0.082
NH	0.839	0.161
NJ	0.735	0.265
NM	0.903	0.097
NV	0.788	0.212
NY	0.819	0.181
OH	0.872	0.128
OK	0.852	0.148
OR	0.859	0.141
PA	0.822	0.178
RI	0.908	0.092
SC	0.767	0.233
SD	0.867	0.133

Continúa en la siguiente página

Tabla 5.1 – Continuación

Estado	No Churn	Churn
TN	0.906	0.094
TX	0.750	0.250
UT	0.861	0.139
VA	0.935	0.065
VT	0.890	0.110
WA	0.788	0.212
WI	0.910	0.090
WV	0.906	0.094
WY	0.883	0.117
Total	0.855	0.145