Abandono de clientes en telecomunicaciones: Un enfoque de clasificación

Abigail Zepeda Castillo[1]; Kiara Castro Sánchez[1]; Kristel Acuña Valverde[1]
abigail.zepeda@ucr.ac.cr, kiara.castro@ucr.ac.cr, kristel.acuna@ucr.ac.cr

Resumen

En los últimos años las empresas de telecomunicaciones han aumentado considerablemente y esto conlleva a diferentes servicios ofrecidos los cuales producen pérdidas o aumento de la clientela según las preferencias del consumidor. Es por esto que el objetivo general de esta investigación es determinar cuál es la mejor técnica para clasificar a los clientes en si abandonarán o no el servicio. Adicionalmente se desea conocer los factores claves que influyen en la decisión de los clientes de permanecer en un servicio de telecomunicaciones. El análisis se enfoca en la aplicación de varios métodos de clasificación y con una base de datos de 7043 observaciones relacionadas con el campo de las telecomunicaciones. Se emplearon las técnicas de: modelo logístico, k-vecinos más cercanos, árboles de decisión, bosques aleatorios, bagging y boosting. Con el fin de determinar la técnica más adecuada para realizar dicha predicción, se llevaron a cabo calibraciones y validaciones cruzadas utilizando las distintas técnicas mencionadas para obtener los mejores indicadores de desempeño. Los resultados obtenidos revelan que el método de bosques aleatorios demostró ser la técnica más efectiva para predecir y clasificar a los clientes, con un error de 19.55% y un área bajo la curva de 70.85%, además se identificaron que las variables Contrato, Permanencia del cliente (tenure), Cargos totales, Cargos mensuales, Seguridad en línea, Servicio de internet y Soporte técnico, son las que más influyen en la decisión de los clientes sobre abandonar o no el servicio. Para futuras investigaciones se recomienda el uso de datos actuales y reales ya que estos pueden brindar mejores resultados, además se recomienda agregar otras variables que podrían explicar mejor.

PALABRAS CLAVES: Modelo logístico, k-vecinos más cercanos, bosques aleatorios, ensambles, predicciones.

Introducción

En la actual era digital, las compañías de telecomunicaciones desempeñan un papel fundamental en nuestras vidas. Ya sea a través de servicios telefónicos, internet o televisión, estas empresas nos conectan con el mundo y nos mantienen comunicados. Sin embargo, en un mercado altamente competitivo, mantener a los clientes satisfechos con la calidad de los servicios brindados, se ha convertido en un desafío constante para estas compañías. Cuando se menciona la satisfacción de los clientes puede ser un tema amplio, ya que es un aspecto fundamental para poder explicar el comportamiento de los consumidores.

Las empresas de telecomunicaciones luchan constantemente por poder sobrevivir en un mercado muy competitivo, ya que los clientes tienen la libertad para elegir entre diferentes proveedores. Es por esta razón que Falla (2021) menciona que existen 3 estrategias para generar más

ingresos: adquirir nuevos [1]clientes, aumentar las ventas de los clientes existentes y aumentar el periodo de retención de los clientes. Se ha demostrado que retener a un cliente existente es mucho más rentable, ya que cuesta mucho menos.

Es por esta razón que las empresas deciden utilizar un indicador que les permita detectar con anticipación si un cliente está a punto de abandonarlos o no, este indicador corresponde a la tasa de abandono de clientes, también conocida como "*Churn*" para el idioma inglés. Lozano (2015) lo define cómo aquellos clientes que dejan la compañía o al proveedor de un servicio durante un período de tiempo determinado, los períodos más utilizados para el cálculo de esta son: mensual, trimestral y anual.

Bocho Moreno (2018) realizó su trabajo final de graduación titulado "Modelo de clasificación para la reducción de la tasa de abandono (Churn Rate) en una empresa del sector de distribución minorista". El objetivo de este trabajo era exponer el proceso para llegar a discutir estas reglas que permiten anticipar la baja de un cliente. Como parte de los resultados de este trabajo se menciona que después de analizar el modelo se decide que el árbol de decisión es mejor para clasificar los datos.

Rodríguez Garzón (2020) realizó un trabajo final titulado "Análisis de Predicción de Churn en una Empresa de Telecomunicaciones" que tenía como objetivo predecir los posibles clientes que podrían cancelar la suscripción en una empresa de telecomunicaciones. Para el análisis de este trabajo se aplicaron técnicas de algoritmos de predicción como regresión logística, redes neuronales, random forest, gradient boosting, entre otros. Así mismo, entre las variables utilizadas se encuentran la antigüedad del cliente, tipo de contrato mes a mes, cargos totales, clientes con seguridad en línea, servicio de internet con fibra óptica y facturación electrónica.

De ahí que surge el deseo de analizar algunas de las variables que contribuyen a la decisión de un cliente de permanecer en un servicio de telecomunicaciones, sabiendo que diferentes aspectos influyen en la experiencia del usuario y pueden determinar si optan por continuar o abandonar el servicio. En un mundo en constante evolución tecnológica, las compañías de comunicación deben mantenerse al tanto de las últimas tendencias, servicios y ofrecerlos actualizados para satisfacer las necesidades cambiantes de sus clientes, los cuales tienen múltiples características y demandas. Por consiguiente, el objetivo general de este trabajo es determinar cuál es la mejor técnica para clasificar a los clientes en si abandonarán o no el servicio (Churn). Adicionalmente se desea conocer los factores claves que influyen en la decisión de los clientes de permanecer en un servicio de telecomunicaciones.

Metodología

Para llevar a cabo el análisis, se utilizó una base de datos obtenida de una plataforma en línea que se centra en la comunidad de ciencia de datos y aprendizaje automático llamada Kaggle. Esta base contiene datos referentes al abandono de clientes de una compañía ficticia llamada Telco en el año 2018. La base de datos contiene información relacionada con los servicios ofrecidos por la compañía y abarca un total de 7043 clientes (BlastChar, 2018). Este conjunto de datos cuenta con información demográfica de los clientes, servicios a los que cada cliente se ha suscrito, información de la cuenta

[1] Estudiantes de Bachillerato en Estadística de la Universidad de Costa Rica

del cliente e información sobre clientes que abandonaron el servicio en el último mes (Churn), es importante aclarar que esta última es la variable respuesta con la que se trabajará, (Tabla 1, sección Anexos).

Como primer paso, se realizó un análisis de la base de datos para identificar la presencia de valores faltantes (NAs), mediante un resumen de esta se confirmó su existencia y dichos valores se eliminaron obteniendo así una base actualizada con un total de 7032 observaciones. Posteriormente, se generó un análisis descriptivo de las variables para observar más de cerca su comportamiento.

Siguiendo el análisis, se utilizó la técnica de regresión logística debido a que la variable respuesta era dicotómica, es decir, tenía dos categorías posibles: "Sí" y "No". Para aplicar esta técnica, se empleó un modelo lineal generalizado (GLM por sus siglas en inglés) que permite modelar la relación entre las variables predictoras y la variable respuesta.

Se evaluaron diferentes modelos, uno para cada combinación de variables predictoras, con el propósito de determinar cuáles variables tenían una influencia significativa en el modelo de regresión logística. Inicialmente, se utilizó un modelo que incluía todas las 20 variables predictoras y de manera iterativa, se eliminó una variable a la vez para evaluar su impacto en los valores de falsos positivos (FP) y falsos negativos (FN). Estos valores son indicadores clave del rendimiento del modelo en términos de errores de clasificación. Aquellas variables que no produjeron cambios sustanciales en los valores de FP y FN fueron consideradas como no aportantes y se eliminaron en un enfoque acumulativo. Es decir que, en cada etapa, se creó un nuevo modelo excluyendo una de estas variables hasta llegar a eliminar las 6 variables consideradas como no influyentes.

Posteriormente, se validaron dichos modelos resultantes utilizando una técnica de validación cruzada con 10 pliegues. La base de datos se dividió en 10 partes y se realizaron 10 iteraciones, entrenando el modelo con nueve partes de datos y evaluándose con la parte restante. Esto permitió obtener medidas de desempeño, como el error de clasificación, los falsos positivos, los falsos negativos, la presión positiva, presión negativa, área bajo la curva ROC (AUC, siglas en inglés) y estadístico KS. Se repitió este proceso de validación cruzada para cada uno de los 6 modelos generados después de eliminar las variables no aportantes. Al finalizar, se calculó el promedio de los indicadores de desempeño obtenidos en los 10 pliegues para cada modelo. Esto proporcionó una medida más sólida y generalizada del rendimiento de cada modelo, al considerar diferentes combinaciones de datos de entrenamiento y prueba.

Como segundo método se utilizó k vecinos más cercanos (Knn, siglas en inglés), junto con la validación cruzada, para realizar la clasificación. Dado que la base de datos contenía variables tanto categóricas como continuas, se aplicó la distancia de Gower para calcular la similitud entre los datos de entrenamiento y validación. Se probó un rango de valores impares para k (número de vecinos más cercanos) desde 1 hasta 15. Para cada valor de k, se realizó una validación cruzada de la misma forma que se explicó anteriormente. Estas medidas permitieron evaluar la capacidad predictiva y la eficacia del modelo de KNN en cada valor de k. Luego, se calculó el promedio de los indicadores de desempeño para cada vecino, lo que facilitó la selección del valor óptimo de k con el mejor rendimiento en términos de precisión y capacidad de generalización.

La tercera técnica utilizada; fue árboles de decisión, su algoritmo de aprendizaje supervisado funciona tanto para variables continuas como para categóricas y permite tomar decisiones basadas en condiciones y reglas lógicas, esto quiere decir que cuando se decide dividir recursivamente el conjunto de datos en ramas basadas en las características o atributos, se crea una estructura de árbol donde cada nodo representa una condición. Los árboles de decisión son fáciles de interpretar y visualizar, lo que los convierte en una opción popular para el análisis de datos, además de que pueden capturar relaciones no lineales entre variables. No obstante, pueden ser propensos a que suceda un sobreajuste si no se controla su crecimiento, esto es que se ajustan demasiado a los datos de entrenamiento y no generalizan bien los nuevos datos (Rajesh S, 2018).

Para abordar el sobreajuste y mejorar la precisión de la clasificación, se utilizó una variante de los árboles de decisión llamada Bosques aleatorios (Random Forests en inglés). De acuerdo con Kulkarni & Sinha (2013) los bosques aleatorios generan un conjunto de árboles de decisión. Para asegurar la diversidad entre los árboles, se utiliza una aleatorización con reemplazo. En lugar de construir un solo árbol, se generan varios árboles de decisión utilizando diferentes subconjuntos de datos y características seleccionadas al azar. Luego, se realiza la clasificación mediante el voto mayoritario de los árboles individuales. Los bosques aleatorios reducen el sobreajuste y mejoran la precisión de la clasificación, además de proporcionar una medida de la importancia de las variables en el proceso de clasificación.

Antes de aplicar los bosques aleatorios, se realizó una calibración para dos parámetros del modelo de árboles de decisiones, para el cp (parámetro de complejidad) y para el maxdepth (profundidad máxima del árbol). En ambos casos, los parámetros se compararon a distintos niveles de valores, evaluando su impacto en el rendimiento del modelo. Para ello, se utilizó la validación cruzada para obtener una estimación confiable del rendimiento del modelo con cada combinación de valores de cp y maxdepth. Seguidamente, se seleccionó el modelo final de árboles de decisiones con los valores óptimos para cada parámetro, donde se incluye el parámetro minbucket con valor de 2%*n y el minisplit de 5%*n, siendo (n) el tamaño de muestra de la base de entrenamiento y se realizó la validación cruzada para dicho árbol final obteniendo la media de esta técnica.

La siguiente técnica utilizada fue la de Bagging (Bootstrap AGGregatING en inglés) seguida de la de Boosting, ambos utilizados en clasificadores de árboles de decisión. Estas técnicas (al igual que el bosque aleatorio) son conocidas como métodos de ensamblaje, que buscan mejorar el rendimiento del modelo mediante la combinación de múltiples clasificadores. En el caso del Bagging, se utilizan muestras de entrenamiento obtenidas mediante el muestreo bootstrap con reemplazo. Cada clasificador se entrena de forma independiente con una muestra diferente y luego se combinan utilizando métodos de promediado, como el promedio ponderado o la mayoría de los votos. Esto ayuda a reducir la varianza y mejorar la generalización del modelo (Jafarzadeh et al., 2021).

Por otro lado, el Boosting se construye secuencialmente. Se entrena para corregir los errores cometidos por las clasificaciones anteriores. De esta manera, el modelo va aprendiendo y mejorando su capacidad de clasificación a medida que se construyen más clasificaciones. En el Boosting, se presta especial atención a los casos mal clasificados en cada iteración, asignándoles un mayor peso para enfocarse en ellos y corregir los errores (Jafarzadeh et al., 2021).

Para ambas técnicas, se realizó la validación cruzada para evaluar el rendimiento del modelo. En el caso del Bagging, se utilizaron las reglas duras definidas en el árbol de decisión para la clasificación. Mientras que, para el Boosting, se calibraron los parámetros iter (número de árboles) y un (parámetro de contracción) mediante la validación cruzada para obtener los mejores valores que maximizaran los indicadores de desempeño del modelo.

El programa estadístico utilizado para los análisis fue R (poner la versión de R) con interfaz en RStudio (versión 2023.03.01) y los paquetes empleados fueron: cluster (Maechler, Rousseeuw, Struyf & Hubert, 2022), caret (Kuhn, 2023), MASS (Ripley, 2023), DT (Xie, Cheng, Tan, 2023), class (Ripley, 2023), kknn (Schliep, Hechenbichler & Lizee, 2016), e1071 (Meyer, Dimitriadou, Hornik, Weingessel, Leisch, Chang, Lin, 2023), adabag (Alfaro, Gámez & García 2023), randomForest (Liaw, A. & Wiener, M, 2023), modeest (Poncet, 2019), tidyverse (Wickham, 2023), nnet (Ripley, 2023), rpart (Therneau & Atkinson, 2022), rattle (Williams, 2022) y readr (Wickham, et al, 2023).

Resultados

Para iniciar con el análisis de la base de datos se realiza la corrección de esta, donde se decide eliminar la variable del identificador del cliente, se convierten las variables a factor o a numéricas decisivamente, se observa la relación de la variable respuesta con cada una de las variables y la cantidad de clientes en cada una. Al contar en su mayoría con variables categóricas se decide crear un gráfico de barras horizontal para cada variable para así observar más fácilmente la distribución de los datos. En la siguiente figura, (Ver figura 1, sección de Anexos), se observa algunos resultados de la base en general como lo son el tipo de Contrato, si el cliente utiliza el servicio de internet para transmitir películas de un proveedor externo, el tipo de servicio de internet que tenía y los diferentes métodos de pago, todo en relación con la variable respuesta de si abandono o no el servicio

Para la variable Contract la cantidad de personas que no abandonaron el servicio y contaban con el contrato por dos años es mayor que las personas que sí abandonaron lo mismo suceden para la cantidad de personas que no abandonaron el servicio y contaban con un contrato de un año, caso contrario para el periodo mensual, ya que se puede ver que la cantidad de personas que abandonan el servicio es proporcional a la cantidad de personas que no abandonaron. Con respecto a la variable StreamingMovies se puede ver que la cantidad de personas que no abandonaron el servicio es casi la misma ya sea si usan y no usan los servicios de internet para transmitir películas de un proveedor externo, por otro lado, lo que sucede con las personas que no cuentan con servicio a internet tiene sentido ya que, si no cuentan con este servicio, no lo van a abandonar.

Para la variable InternetService la cantidad de personas que cuentan con fibra óptica (que utiliza un cable conocido por ser de vidrio) y no abandonaron el servicio es similar a las personas que sí abandonaron, mientras que para las personas que cuentan con servicio de internet de DSL (cable de cobre) la cantidad de personas que no abandonaron el servicio es mayor a las que sí abandonan, las personas que no abandonaron el servicio y no cuentan con internet es mayor que las que sí eligieron abandonar, lo cual tiene sentido. Con respecto a la última variable mostrada en la figura 1, esta hace referencia a las múltiples opciones de pago que tienen los clientes, donde se puede ver que en su mayoría el porcentaje de clientes que no abandona es mayor para cualquier tipo de pago. Las demás variables se distribuían parecido por lo que no fueron agregadas a la figura 1. De los gráficos de barras

anteriores se puede observar que difiere mucho la categoría de la variable con respecto a si el cliente continuó con el servicio o no.

Como parte del primer escenario de resultados se decide dividir la base para su entrenamiento y su validación, el 80% (5625 observaciones) de los datos se escogen como base de entrenamiento y el 20% (1407 observaciones) restante para validación. Una vez que se obtiene esta división, se evaluaron diferentes modelos, uno para cada combinación de variables predictoras, eliminando iterativamente cada una de las variables con el propósito de determinar cuáles variables presentan una baja influencia significativa en el modelo de regresión logística. Entre las variables que presentaron en ambos casos falsos positivos y falsos negativos bajos son las siguientes: Dependents, OnlineBackup, Deviceprotection, StreamingTV, Payment Method y MonthlyCharges. Esto indica que estas variables no están aportando información relevante al modelo de regresión logística y, por lo tanto, su exclusión no afectaría significativamente las clasificaciones incorrectas ya que, al eliminar estas variables del modelo, se espera que no se produzca un aumento en los errores de clasificación, ya que su presencia o ausencia no tiene un impacto sustancial en la capacidad del modelo para predecir con precisión si el cliente abandona o no el servicio.

Seguido de esto se procedió a crear 6 modelos GLM, el primer modelo con todas las variables, pero sin contar Dependent, el segundo modelo con todas las variables sin contar Dependents ni OnlineBackup y así sucesivamente con las 6 variables recientemente mencionadas, esto con el objetivo de seleccionar el mejor modelo. Para cada uno de estos 6 modelos se calcularon los indicadores de desempeño: Error de Clasificación, Falsos Positivos, Falsos Negativos, Precisión Positiva, Área bajo la curva (AUC por sus siglas en inglés) y prueba de Kolmogorov-Smirnov (por sus siglas KS). Como paso final se realizó una validación cruzada con 10 pliegues para cada uno de estos 6 modelos calculados y obteniendo los mismos indicadores de desempeño en cada caso, los resultados obtenidos se observan en la Tabla 2.

Tabla 2

Indicadores de desempeño para la validación cruzada del modelo logístico

Modelos	Indicadores					
	e	FP	FN	PP	AUC	KS
Modelo 1	19.525	10.470	44.474	55.526	72.528	45.056
Modelo 2	19.681	10.515	45.071	54.929	72.207	44.414
Modelo 3	19.667	10.549	44.906	55.094	72.272	44.545
Modelo 4	19.653	11.011	43.532	56.468	72.729	45.457
Modelo 5	19.796	10.468	45.648	54.352	71.942	43.884
Modelo 6	19.554	10.355	44.938	55.061	72.353	44.707

A partir de esta tabla es que se realiza la comparación de los valores para decidir cuál es el mejor modelo, que en este caso se escogió el modelo 4 ya que su valor de AUC era uno de los mayores, así como el valor de KS, aunque en general los valores eran bastante parecidos para todos los modelos, en este caso para el 4to modelo se aumentaba la precisión positiva y en el caso de los falsos negativos era uno de los menores entre su categoría, lo cual es deseable para ambos casos.

Continuando con el análisis, se recuerda que en este caso para el algoritmo de Knn este se tiene que realizar con la distancia de Gower ya que la base de datos cuenta con variables tanto numéricas como categóricas. Se realizan iteraciones de validación cruzada en el que se vuelven a crear 10 pliegues, pero en este caso con una secuencia de la cantidad de vecinos más cercanos con un mínimo de 1 hasta 15, saltando de 2 en 2 ya que se recomienda utilizar valores impares en los K vecinos. Recordando que al final se obtiene el promedio de los 10 pliegues para cada indicador de desempeño como se muestra en la siguiente tabla:

Tabla 3

Indicadores de desempeño para validación cruzada de k-vecinos más cercanos

K vecinos más cercanos	Indicadores					
	e	FP	FN	PP	AUC	KS
k = 1	28.258	20.421	49.853	50.147	64.863	29.726
k = 3	24.616	8.039	70.379	29.621	60.791	21.582
k = 5	25.071	16.693	48.228	51.772	67.539	35.079
k = 7	22.554	9.006	59.987	40.013	65.503	31.006
k = 9	23.265	14.636	47.086	52.914	69.139	38.278
k = 11	22.284	9.697	57.028	42.971	66.637	33.274
k = 13	23.293	14.369	47.918	52.082	68.857	37.713
k = 15	22.312	10.339	55.327	44.673	67.167	34.334

De acuerdo con las medias obtenidas para cada vecino, se observa que primeramente que el AUC mayor es para el k = 9 con un 69.139% esto nos dice que tiene una capacidad moderada de clasificación, lo que está bien ya que el AUC debería estar en un rango de 0.6 y 0.9, menor a ese rango se podría decir que el modelo podría estar clasificando aleatoriamente. En cuanto al KS, se observa que para el mismo vecino de k = 9 presenta el valor más alto, dicho valor nos indica que tiene un mejor rendimiento del modelo.

Seguido a esto se aplicaron árboles de decisión, no sin antes observar los resultados del modelo sin calibrar y su respectivo árbol. Para aplicar poco a poco la calibración de modelo se realizaron los siguientes pasos:

- Validación cruzada con 10 pliegues para el valor del parámetro de complejidad (cp por sus siglas en inglés) donde se aplicaron 5 diferentes valores de cp con valor mínimo de 0.0005 y valor máximo de 0.0009.
- Validación cruzada con 10 pliegues para el valor de la profundidad máxima del árbol (maxdepth definido en inglés) con una secuencia desde 1 hasta 15 recorriendo cada uno de los resultados (de 1 en 1).
- Se obtiene el valor único correspondiente al número mínimo de observaciones para que un nodo se pueda dividir (minsplit en inglés) obteniendolo 5% * 5625 (cantidad de datos en la base de entrenamiento), con un resultado de 281.
- Se obtiene el número mínimo de observaciones que debe tener un nodo para ser considerado como terminal (minbucket en inglés) como 2% * 5625 con un resultado de 112.

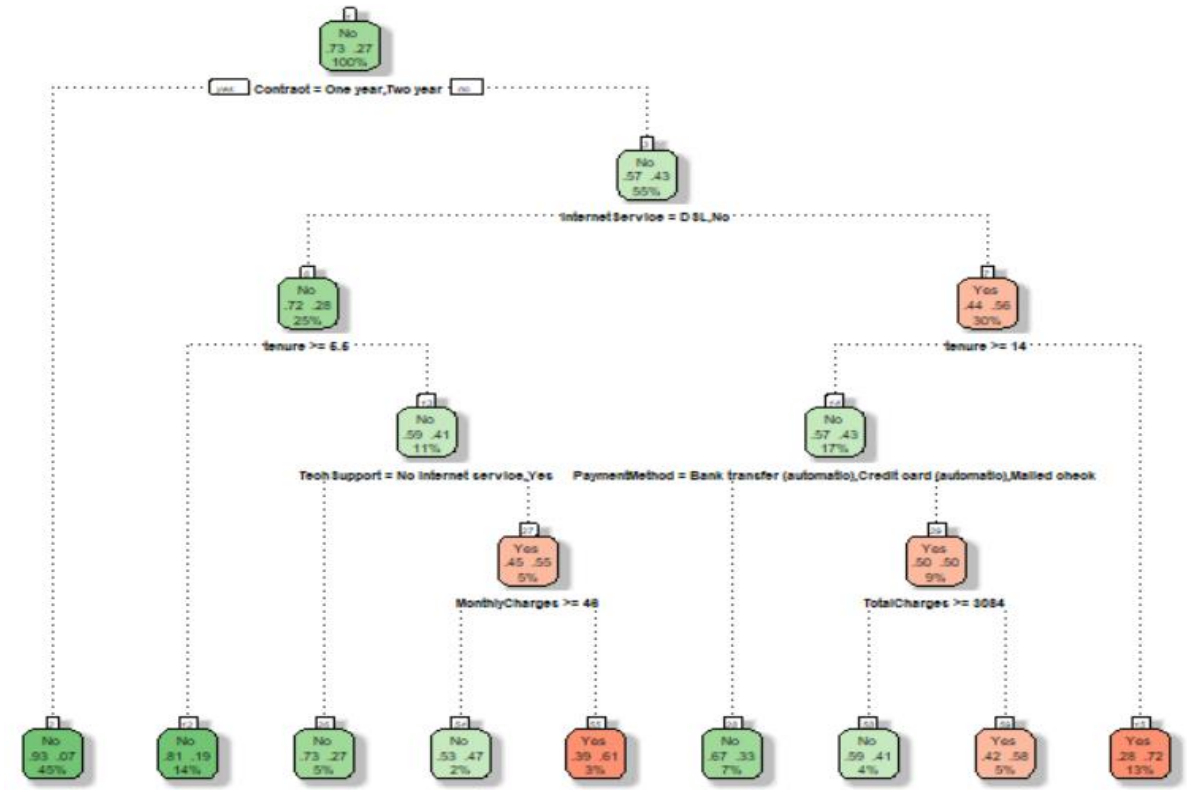
Una vez realizadas las validaciones se obtuvieron los indicadores de desempeño para hacer su comparación, en el caso del cp los mejores valores de indicadores se obtuvieron para el valor de

0.0009, cabe mencionar que dichas medidas de desempeños fueron muy similares para los 4 modelos, no obstante, para el modelo con el cp más alto obtuvo uno de los AUC y KS más altos, así mismo sus FP y FN fueron moderadamente bajos. Lo mismo sucedió en el caso de la profundidad máxima del árbol quedando con una cantidad de 5 por ser uno de los que presentaban los indicadores más altos. Con estos datos se obtiene la calibración para el nuevo modelo al cuál se le realiza validación cruzada y se obtienen los indicadores de desempeño nuevamente, de acuerdo con el árbol de decisión se obtuvo que las variables que aportan más al modelo fueron: el tipo de contrato, el tipo de servicio de internet, la permanencia del cliente al servicio, si adquiere soporte técnico, el método de pago, cantidad de pago mensual y los cargos totales. En la figura 2 se muestra la forma del árbol de decisión con las variables mencionadas anteriormente.

Cabe destacar que 2 de estas variables importantes según el árbol de decisiones (métodos de pago y pagos mensuales) son variables que de acuerdo con el primer análisis con el modelo logístico se podrían eliminar, sin embargo, al escoger el modelo más apto en el modelo logístico, dichas variables se dejaron en el modelo, lo cual es interesante que en ambos casos se mantuvieran.

Figura 2

Árbol de decisión con el modelo calibrado



Como parte del ensamble de modelos se realizaron bosques aleatorios (Random forest por su nombre en inglés), agregación de bootstrap (Bagging) y método de potencialización (Boosting). En el caso de los bosques aleatorios, se creó un modelo con un valor de ntree = 200 ya que al crear la validación cruzada con diferentes valores de ntree (número de árboles) los promedios de los valores

de los indicadores de desempeño se mantenían muy similares, esto nos dice que indiferentemente del número (100 a 600 con saltos de 100) que se le proporciona al parámetro, los resultados serán similares. Seguidamente también se realizó una validación cruzada para el parámetro de mtry (número de variables aleatorias del árbol) usando una secuencia de 1 a 19 variables predictoras, donde se destacó que el modelo con 7 variables tenía los mejores indicadores de desempeño, sin embargo, cabe destacar que entre cada modelo los indicadores no reflejaron muchas diferencias.

De acuerdo al modelo de bosques aleatorios, a medida que incrementa el valor de importancia de una variable, más contribuye está en la precisión del modelo (esto al realizar las divisiones de los árboles de decisión), quiere decir que estas variables, (Ver figura 3, sección de Anexos), tienen un mayor impacto en la predicción final

Con respecto al resultado de Bagging, este no se pudo realizar con la fórmula directa, se realizó paso a paso con iteraciones para árboles de decisión, obteniendo en esta etapa la moda que brinda la clasificación para cada uno de los individuos para conocer si se clasificaron correctamente. Por otro lado, para el resultado de Boosting se tuvo que aplicar la técnica paso a paso, realizando la calibración del número de árboles (iter) y del parámetro de contracción (un) respectivamente y realizando su validación cruzada para obtener los mejores indicadores de desempeño.

Para finalizar el estudio de las técnicas descritas se presenta una comparación de cada uno de los indicadores de desempeño para cada técnica utilizada, (Ver tabla 4, sección de Anexos), se decide escoger el método de ensamble de bosques aleatorios ya que se cree que es el que mejor ayuda a predecir si un cliente va a abandonar o no el servicio en el último mes.

Los indicadores resultantes obtenidos con la técnica de bosques aleatorios fueron los siguientes; para el error de clasificación alrededor del 19% de las predicciones fueron erróneas. Para los falsos positivos (FP) un 8.7% de los casos, se predijo erróneamente que estos clientes no abandonarían el servicio. Por otro lado, hubo 49.6% de falsos negativos (FN), en los que se predijo erróneamente que los clientes abandonarían el servicio. En cuanto a la precisión positiva (PP) fue más del 90%, lo que indica que el modelo identificó correctamente los casos de clientes que abandonaron el servicio. El Área bajo la curva (AUC) fue de 70.846, lo que indica una capacidad moderada para distinguir entre los dos grupos de clientes. La Estadística KS fue cerca del 42%, lo que sugiere que el modelo tiene una buena capacidad para clasificar a los clientes en términos de si abandonarán o no el servicio en el último mes.

Conclusiones

En este estudio, se llevó a cabo una comparación de seis técnicas de clasificación para determinar cuál de ellas era la más adecuada para predecir si los clientes abandonarían o no el servicio (Churn). Las técnicas evaluadas fueron el modelo logístico, k-vecinos más cercanos, árboles de decisión, bosques aleatorios, bagging y boosting. Después de aplicar cada técnica y calcular los indicadores de desempeño correspondientes, se encontró que el método de bosques aleatorios mostró el mejor rendimiento para clasificar a los clientes en términos de si abandonarán o no el servicio.

Como se observó en los resultados, los falsos positivos obtenidos con el método de bosques aleatorios representan el 8.6%, este dato significa que se predijo incorrectamente que un cliente no abandonaría el servicio cuando en realidad, sí lo hizo. Es importante destacar este punto, ya que en el caso de que una empresa asigne erróneamente servicios a clientes con alta probabilidad de abandono podría generar pérdidas para la misma. Lo mismo sucede con los falsos negativos, donde se predijo que el cliente abandonaría el servicio cuando en realidad no lo hizo. En este caso, el porcentaje obtenido fue de 49.6%, esto implica que muchos clientes que abandonaron el servicio podrían haberse beneficiado de acciones de retención, pero no fueron identificados como en riesgo de abandono. Reducir los falsos negativos es esencial para evitar la pérdida de clientes valiosos. Es importante destacar que el modelo logró clasificar correctamente más del 90% de los casos en los que el cliente realmente abandonó el servicio, lo cual es muy bueno. Una alta precisión positiva proporciona confianza en la correcta clasificación de los clientes y permite generar a las empresas estrategias efectivas de retención. Además, es necesario resaltar que estos resultados no solo se deben a la elección de un modelo con buenos indicadores, sino también al uso de la validación cruzada. Esta técnica permite evaluar el rendimiento del modelo y evitar el sobreajuste, proporcionando una estimación más realista al evaluar su capacidad en múltiples divisiones de los datos de entrenamiento y validación.

Por otro lado, era de interés conocer los factores claves que influyen en la decisión de los clientes de permanecer en un servicio de telecomunicaciones, puesto que se obtuvieron 200 de árboles de manera aleatoria y cada uno con diferentes variables, se eligieron las variables que se mostraron con más frecuencia en los árboles generados, tomando de esta forma un modelo con 7 variables finales, y estas resultaron ser: Contrato, Permanencia del cliente (tenure), Cargos totales, Cargos mensuales, Seguridad en línea, Servicio de internet y Soporte técnico.

Según el estudio de Rodríguez (2020) cuyo objetivo principal era poder encontrar el modelo que predijera los clientes que pueden cancelar la suscripción en una empresa de telecomunicaciones, utilizando un indicador Churn (tasa de cancelación), se aplicaron seis diferentes técnicas de clasificación y de los modelos que fueron seleccionados como ganadores se tomaron las variables más frecuentes: Antigüedad del cliente, tipo de contrato mes a mes, cargos totales, clientes con seguridad en línea, servicio de internet con fibra óptica y facturación electrónica. Esto concuerda, no sólo con el método que se utilizó en el presente análisis para obtener las variables que mayor influencia tienen, sino que variables como Cargos totales, Contrato, Permanencia del cliente y Servicios de Internet influyen en la decisión de los clientes de permanecer en un servicio de telecomunicaciones, en ambos estudios.

Como parte de las recomendaciones finales se encuentra el poner en práctica este tipo de análisis para datos reales, ya que, al estar trabajando con datos ficticios, los resultados obtenidos podrían no ser tan fiables, dado que existe el riesgo de estar alejándose de la realidad. Además, se recomienda la inclusión de otras variables, como las utilizadas por Bocho Moreno (2018), por ejemplo: Intervalo de edad que se encuentra el cliente en el momento de extraer los datos, Nivel cultural, Ciudad y Descuento en precios e incentivos monetarios. Sin embargo, es necesario tomar en cuenta que una gran cantidad de variables subjetivas podría dificultar el análisis de las técnicas.

Bibliografía

Alfaro, E., Gámez, M. & García, N. (2023). adabag: Aplica Multiclase AdaBoost.M1, SAMME y Bagging . (s/f). Red integral de archivos R (CRAN). Recuperado de https://cran.r-project.org/web/packages/adabag/index.html

BlastChar. (2018). Churn de clientes de telecomunicaciones [Conjunto de datos].Recuperado de: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Bocho Moreno, C. (2018). Modelo de clasificación para la reducción de la tasa de abandono (Churn Rate) en una empresa del sector de distribución minorista. Recuperado de: https://uvadoc.uva.es/bitstream/handle/10324/31957/TFG-G2956.pdf?sequence=1&isAllowed=y

Falla, J. D. (2021). Predicción de abandono de clientes en telecomunicaciones mediante el aprendizaje automático. Recuperado de: http://hdl.handle.net/20.500.12010/22247.

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation. *Remote Sensing*, *13*(21), 4405.https://espace.inrs.ca/id/eprint/12218/1/P4039.pdf

Kuhn, M. (2023). Entrenamiento de clasificación y regresión [R package caret versión 6.0-94] . https://cran.r-project.org/web/packages/caret/index.html

Kulkarni, V. Y., & Sinha, P. K. (2013). Random forest classifiers: a survey and future research directions. Int.J. Adv. Comput, 36(1), 1144-1153. https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Random-Forest-Classifiers_A-Survey-and-Future.pdf

Liaw, A. & Wiener, M. (2023). Classification and Regression by randomForest. https://cran.rproject.org/web/packages/randomForest/randomForest.pdf

Lozano, D. (2015). MODELOS PREDICTIVOS DEL CHURN–ABANDONO DE CLIENTES–PARA OPERADORES DE TELECOMUNICACIONES. Recuperado de: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1269.pdf

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. (2022). CRAN - Package cluster. CRAN. https://svn.r-project.org/R-packages/trunk/cluster/

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., Lin, C. (2023). Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [R package e1071 version 1.7-13]. Recuperado de https://cran.r-project.org/web/packages/e1071/index.html

Poncet, P. (2019). Mode Estimation [R package modeest version 2.4.0]. https://cran.r-project.org/web/packages/modeest/index.html

Rajesh S., B. (2018) Decision trees - A simple way to visualize a decision. Medium: GreyAtom. https://medium.com/greyatom/decision-tree-intuition-a38669005cb7

Ripley, B. (2023). Support Functions and Datasets for Venables and Ripley's MASS [R package MASS version 7.3-60]. https://cran.r-project.org/web/packages/MASS/index.html

Ripley, B. (2023). Functions for Classification [R package class version 7.3-22]. https://cran.r-project.org/web/packages/class/index.html

Ripley, B. (2023). Feed-Forward Neural Networks and Multinomial Log-Linear Models [R package nnet version 7.3-19]. https://cran.r-project.org/web/packages/nnet/index.html

Rodríguez Garzón, M. A. (2020). Análisis de predicción de CHURN en una empresa de telecomunicaciones. Recuperado de: https://docta.ucm.es/rest/api/core/bitstreams/96ec5700-9fa5-4de0-b279-6983736aae95/content

Schliep,K., Hechenbichler,K & Lizee, A (2016). kknn: Weighted k-Nearest Neighbors. R package version 1.3.1. Recuperado de https://CRAN.R-project.org/package=kknn

Therneau,T & Atkinson,B. (2022). Recursive Partitioning and Regression Trees [R package rpart version 4.1.19]. https://cran.r-project.org/package=rpart

Wickham, H. (2023). Easily Install and Load the "Tidyverse" [R package tidyverse version 2.0.0]. https://cran.r-project.org/web/packages/tidyverse/index.html

Williams, G. (2022). Graphical User Interface for Data Science in R [R package rattle version 5.5.1]. https://cran.r-project.org/web/packages/rattle/index.html

Xie, Y., Cheng, J., Tan, X. (2023). Un contenedor de la biblioteca de JavaScript "DataTables" [R paquete DT versión 0.28] . https://cran.r-project.org/web/packages/DT/index.html

Anexos

Figura 1

Distribución de variables categóricas

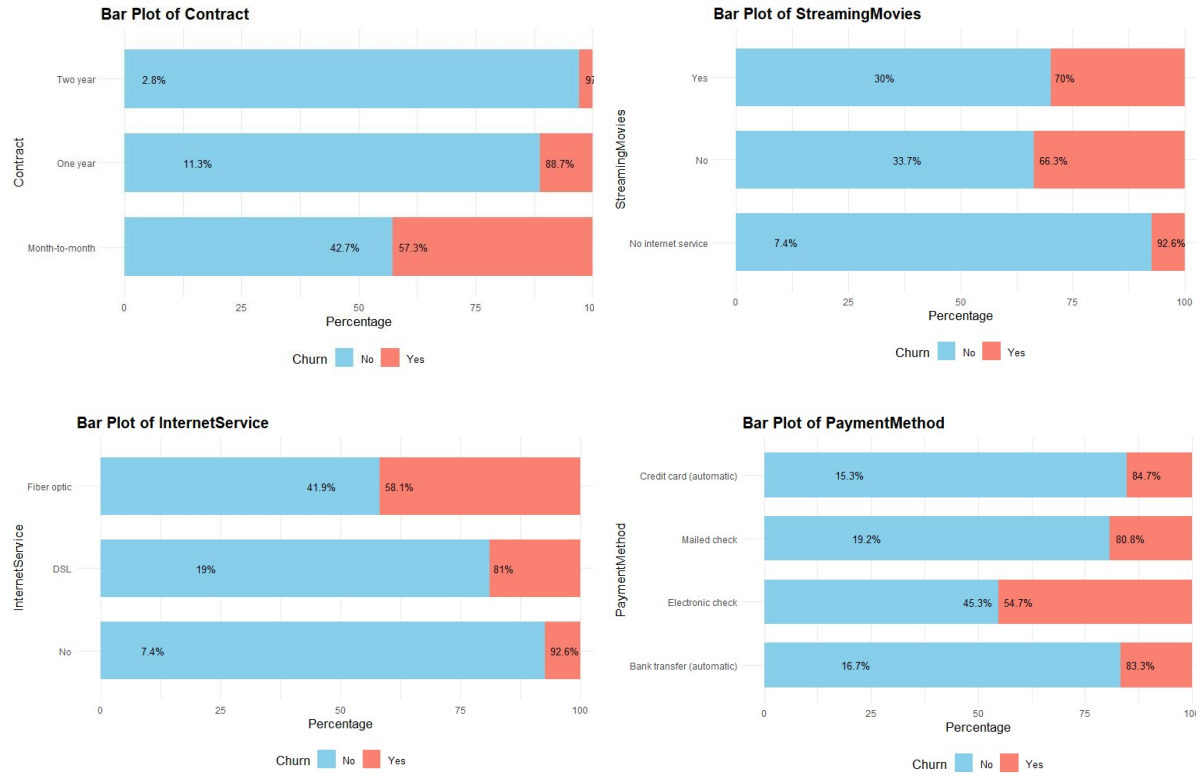


Tabla 1

Descripción de las variables utilizadas en el estudio

Variables demográficas

Variable	***Tipo***	***Definición***
Gender	Categórica	Género (Femenino o Masculino)
SeniorCitizen	Numérica	Si el cliente es adulto mayor o no
Partner	Categórica	Si el cliente tiene pareja o no

Variables sobre los servicios a los que cada cliente se ha suscrito

Variable	***Tipo***	***Definición***
PhoneService	Categórica	Si el cliente tiene el servicio de telefonía o no.

MultipleLines	Categórica	Si el cliente tiene más de una línea telefónica, no tiene más de una línea o no tiene el servicio de telefonía).
InternetService	Categórica	Qué tipo de servicio de internet tiene (DSL, Fibra óptica, no tiene).
OnlineSecurity	Categórica	Si el cliente tiene servicio de ciber seguridad, si no lo tiene o no tiene el servicio de internet del todo.
OnlineBackup	Categórica	Si el cliente tiene un respaldo en línea, si no lo tiene, si no tiene el servicio de internet del todo.
DeviceProtection	Categórica	Si el cliente tiene protección del dispositivo de internet, si no tiene, si no tiene el servicio de internet del todo.
TechSupport	Categórica	Si el cliente adquiere soporte técnico, si no lo adquiere o si no tiene el servicio de internet del todo.
Streaming TV	Categórica	Si el cliente usa su servicio de internet para transmitir programación de televisión de un proveedor externo, si no lo utiliza y si no tiene el servicio de internet del todo.
StreamingMovies	Categórica	Si el cliente utiliza su servicio de Internet para transmitir películas de un proveedor externo, si no lo utiliza y si no tiene el servicio de internet del todo.

Variables de las cuentas del cliente

Variable	***Tipo***	***Definición***
Tenure	Numérica	Número de meses que el cliente se ha mantenido con el servicio (desde 1 hasta 72 meses).
Contract	Categórica	Si tiene un contrato con el servicio mensual, por 1 año, por 2 años.
PaperlessBilling	Categórica	Si la persona tiene facturación electrónica o no.
PaymentMethod	Categórica	Si el cliente paga con cheque electrónico, cheque enviado por correo, transferencia bancaria (automática), tarjeta de crédito (automática).
MonthlyCharges	Numérica	El cobro mensual total que se le hace al cliente por todos los servicios que tiene (desde $18 hasta $119).
TotalCharges	Numérica	Cargas totales del cliente calculados al final del trimestre (desde $18 hasta $8685)

Figura 3

Importancia de variables de acuerdo con el modelo de bosques aleatorios

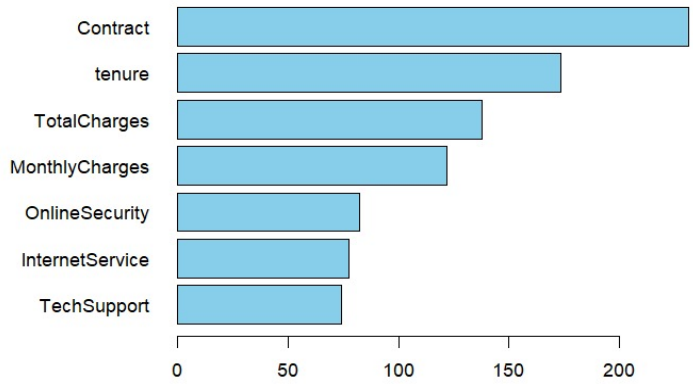


Tabla 4

Indicadores de desempeño para cada una de las técnicas

Técnicas de decisión	Indicadores					
	e	FP	FN	PP	AUC	KS
Modelo logístico	19.653	11.011	43.532	56.468	72.729	45.457
K-vecinos más cercanos	23.265	14.636	47.086	52.914	69.139	38.278
Árbol de decisión	20.520	10.734	47.675	52.325	70.796	41.591
Bosques aleatórios	19.554	8.661	49.645	91.338	70.846	41.693
Bagging	21.545	13.808	42.887	57.113	71.652	43.305
Boosting	19.397	10.290	44.641	55.358	72.534	45.068