Predicción de la deserción de clientes en empresas de telecomunicaciones mediante un modelo de regresión logística

Leydi Tamara Ariza Cabrejo

Director

Julio Eduardo Mejia Manzano

Asesor

Jorge Luis Quintero Lopez

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básica Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Agradecimientos

Agradezco profundamente al Dr. Freddy Torres Payoma, por su generosidad, su rigor académico y disposición que fueron fundamentales para superar los obstáculos técnicos de este trabajo, a pesar de no ser mi tutor formal, fue el pilar que permitió llevar este proyecto a término con excelencia.

De igual manera, manifiesto mi gratitud a Eduardo Malagón. Gracias por ser el pilar emocional que me sostuvo en los momentos de duda y por creer siempre en mi potencial; su motivación constante fue clave para continuar mi formación. Le guardo un profundo afecto.

Resumen

En la industria de telecomunicaciones la cantidad de clientes genera diariamente datos masivos, por tanto, resulta más viable retener los clientes actuales ya que una base de datos nueva implica altos costos, por ese motivo, se propone implementar un modelo predictivo de regresión logística para anticipar la deserción de clientes en estas empresas. Se realizará una clasificación de variables que ayudarán a determinar los factores que más influyen en la deserción y probabilidad de churn de cada usuario tales como variables sociodemográficas, patrones de consumo y datos de facturación. Los resultados permitirán a las empresas diseñar estrategias de retención más efectivas, reducir los costos asociados a la pérdida de clientes y fortalecer su competitividad en el mercado.

***Palabras clave*:** telecomunicación, retención, análisis de datos, datos estadísticos, Análisis de regresión

Abstract

In the telecommunications industry, the number of customers generates large-scale data daily. In this case, retaining current customers is feasible since a new database entails high costs. Therefore, a predictive logistic regression model is proposed to anticipate customer churn in these companies. A classification of variables will help determine the factors that most influence churn and the probability of churn for each user, such as sociodemographic variables, consumption patterns, and billing data. The results will allow companies to design more effective retention strategies, reduce the costs associated with customer loss, and strengthen their competitiveness in the market.

Keywords**:** telecommunications, retention, data analysis, statistical data, regression analysis

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

En el sector de las telecomunicaciones la alta competencia y la facilidad de migración de los clientes entre operadores ha generado que las compañías creen diferentes estrategias de retención convirtiéndolo en un desafío constante. La pérdida de usuarios o churn representa una disminución directa en los ingresos y también un incremento en los costos asociados a la atracción de nuevos clientes afectando la rentabilidad y sostenibilidad de las compañías.

Teniendo en cuenta esa situación las organizaciones han adoptado medidas con ayuda de la analítica de datos adoptando diferentes modelos predictivos para anticipar el riesgo de abandono y diseñar estrategias de fidelización.

En el desarrollo de este trabajo se busca implementar un modelo predictivo basado en regresión logística que permite estimar la probabilidad de deserción de clientes en empresas de telecomunicaciones usando un conjunto de datos de 7043 registros que incluye variables como tenencia, métodos de pago, cargos mensuales, tipo de servicio y churn. Este enfoque permite analizar los factores que más influyen en la decisión de abandono y evaluar el desempeño del modelo mediante métricas estadísticas como accuracy, precision, recall, F1-score y ROC-AUC, proponiendo estrategias orientadas a mejorar la retención de clientes.

De esta manera, este proyecto aborda un problema relevante para la industria proporcionando a la empresa herramientas para tomar decisiones estratégicas basadas en datos, reducir pérdidas económicas y fortalecer su posición competitiva en un mercado dinámico.

Planteamiento del Problema

En el sector de las telecomunicaciones, la competencia entre empresas es cada vez más fuerte, Por eso, retener a los clientes se ha vuelto crucial para asegurar la sostenibilidad y rentabilidad de las compañías. Adquirir nuevos usuarios suele ser mucho más caro que mantener a los que ya se tienen. Sin embargo, muchas empresas todavía tienen dificultades para identificar a tiempo a los clientes que están a punto de dejar el servicio, lo que provoca pérdidas en los ingresos, reducción de la cartera activa y, en última instancia, una caída en la rentabilidad.

Aunque cuentan con grandes cantidades de datos, como información demográfica, contractuales, de uso e interacción, no siempre se aprovechan de manera efectiva para predecir el comportamiento de deserción. Esta limitación se agrava porque la tasa de churn suele ser baja y las bases de datos tienden a estar desbalanceadas, lo que dificulta el desarrollo de modelos predictivos estables y confiables. Como resultado, las estrategias de retención suelen ser reactivas, implementadas cuando el cliente ya ha tomado la decisión de cambiar de proveedor. Esta situación no solo afecta la rentabilidad del negocio, sino también la calidad del servicio y la satisfacción del cliente, factores que son clave para la fidelización y la ventaja competitiva en el sector.

La problemática central que motiva esta investigación puede sintetizarse en la siguiente pregunta: ¿Es posible anticipar qué clientes están en riesgo de abandonar la compañía?, Esta pregunta plantea la necesidad de identificar a los clientes que presentan mayor probabilidad de churn, utilizando información histórica y actual sobre su comportamiento y características, busca comprender los patrones de abandono, los factores que influyen en la decisión de los clientes y cómo estos pueden ser modelados mediante técnicas de análisis de datos y machine learning.

Responder esta pregunta permite a la empresa tomar decisiones proactivas y estratégicas, busca generar conocimiento accionable sobre los factores de riesgo, fortaleciendo la capacidad de la empresa para anticiparse a la deserción y mejorar la experiencia del cliente. Así se conecta directamente con los objetivos del proyecto y con la importancia de la investigación para el sector de telecomunicaciones, donde la retención de clientes es un indicador crítico de sostenibilidad y ventaja competitiva.

Justificación

Para las empresas de telecomunicaciones la deserción de clientes es un problema estratégico y financiero debido a la fuerte competencia, los cambios de tarifas y la facilidad de migración a otros operadores. Su justificación se basó en tres áreas principales que son financiero y de negocio, ventaja competitiva y personalización y optimización de recursos con toma de decisiones. Contar con un modelo predictivo permite anticipar la probabilidad de abandono ofreciendo a la empresa la oportunidad de diseñar estrategias personalizadas de retención, optimización de recursos y mejora de la experiencia del cliente.

La elección de la regresión logística como modelo principal se fundamenta por razones estratégicas y técnicas. La regresión logística es uno de los modelos más utilizados en investigaciones sobre predicción de churn, debido a su capacidad para estimar probabilidades de manera directa y cuantificar la influencia de cada variable en la decisión de los clientes. Esto resulta especialmente valioso para la interpretación de resultados y la toma de decisiones, permitiendo a los responsables de negocio comprender qué factores incrementan el riesgo de abandono y cómo intervenir sobre ellos. A diferencia de modelos más complejos, como Random Forest o redes neuronales, la regresión logística facilita la explicación de los hallazgos aumentando la confianza en la aplicación de las estrategias derivadas del modelo.

El proyecto busca responder a la necesidad de identificar de manera temprana los clientes con mayor probabilidad de churn, optimizando la asignación de recursos y permitiendo diseñar intervenciones personalizadas, como promociones específicas, mejoras en los servicios o incentivos. Este enfoque permite a la empresa pasar de una estrategia reactiva a una proactiva, reduciendo costos asociados al reemplazo de clientes y aumentando la rentabilidad a largo plazo. Los beneficiarios directos de este proyecto son las áreas de marketing, ventas y gestión de

clientes, que podrán priorizar sus esfuerzos en segmentos de alto riesgo, mientras que los clientes se benefician de un servicio más personalizado y adaptado a sus necesidades.

Además, la regresión logística ofrece una base sólida para la evaluación de desempeño del modelo, mediante métricas confiables como accuracy, precision, recall, F1-score y ROC-AUC pudiendo comparar su rendimiento frente a otras técnicas de clasificación. Durante la fase experimental, el modelo demostró resultados satisfactorios tanto en capacidad predictiva como en interpretabilidad, consolidando su idoneidad para el proyecto. Aunque otros modelos como árboles de decisión o Random Forest pueden ofrecer mayor precisión, su menor interpretabilidad dificulta explicar de manera estratégica los factores que motivan la deserción.

Finalmente, el proyecto es pertinente y viable porque combina metodologías de análisis de datos y modelado predictivo que se ajustan a los recursos disponibles y a la información histórica de clientes. La implementación de este modelo no solo aborda un problema empresarial relevante, sino que también genera beneficios tangibles en términos de optimización de recursos, mejora de la experiencia del cliente y ventaja competitiva en un sector altamente dinámico.

Objetivos

Objetivo General

Desarrollar un modelo predictivo basado en regresión logística para estimar la probabilidad de deserción de clientes en empresas de telecomunicaciones.

Objetivos Específicos

Realizar la recolección, exploración y preprocesamiento de los datos, mediante técnicas de limpieza, transformación y análisis descriptivo, para construir un conjunto de variables para el análisis predictivo.

Diseñar y entrenar un modelo de regresión logística que permita estimar la probabilidad de deserción de clientes en el sector de telecomunicaciones, a partir de las variables seleccionadas.

Evaluar el desempeño del modelo predictivo mediante métricas estadísticas como precisión, especificidad, AUC y matriz de confusión, e identificar las variables más influyentes en la deserción, con el fin de proponer recomendaciones estratégicas orientadas a la retención y fidelización de clientes.

Marco de Referencia

Estado del Arte

La predicción de churn se ha convertido en un tema casi inevitable dentro de la industria de las telecomunicaciones. Y es de esperarse debido a la competencia ya que atraer nuevos usuarios suele ser costoso, lento. En este contexto, diversos estudios han puesto de relieve cómo el análisis de datos y las técnicas de machine learning se han vuelto aliados esenciales para reconocer patrones de abandono y, en consecuencia, fortalecer las estrategias de retención (De Caigny, 2018)

Por ejemplo, (Jain, 2020) experimentaron con una combinación de regresión logística y Logit Boost. El resultado fue bastante prometedor ya que se logró un equilibrio entre precisión y recall, incluso cuando trabajaban con datasets desbalanceados, esos que suelen complicar más de la cuenta cualquier modelo. Otros autores han apostado por enfoques híbridos, mezclando árboles de decisión, Random Forest y SVM para mejorar la capacidad predictiva y reducir los errores de clasificación tanto como sea posible (Olle, 2014); (Nurtriana, 2024)

Muchas empresas de telecomunicaciones ya han pasado de la teoría a la acción. Hoy utilizan estos modelos para segmentar a sus clientes de forma más inteligente y diseñar estrategias de fidelización que sean acertadas para cada necesidad del cliente como planes personalizados y, en general, propuestas que se ajustan mejor a lo que cada usuario necesita (L. F. Khalid, 2021); (Krishna, 2024); (Wagh, 2024). Este enfoque impulsado por machine learning no solo ayuda a tomar decisiones más informadas, sino que también permite anticiparse a la pérdida de clientes y usar los recursos con mayor eficiencia.

Marco Contextual

El sector de telecomunicaciones en la actualidad se caracteriza por una alta competitividad y rapidez en la evolución tecnológica, obligandoa las empresas a ofrecer servicios diferenciados y de alta calidad, por tanto, la retención de clientes se convierte en un factor importante, dado que el abandono de un cliente representa no solo la pérdida de ingresos recurrentes, sino también un impacto en la rentabilidad de la empresa debido a los altos costos de adquisición de nuevos usuarios.

El fenómeno de churn en telecomunicaciones está influenciado por diversos factores relacionados con la experiencia del cliente, como la duración del contrato, el costo mensual de los servicios, la modalidad de pago y la percepción de valor recibida. Comprender estos factores permite a las empresas enfocar acciones de fidelización y diseñar intervenciones personalizadas que reduzcan la probabilidad de abandono, garantizando la sostenibilidad del negocio en un mercado dinámico.

Marco Teórico

El presente marco teórico se centra en los conceptos fundamentales para abordar la implementación de el modelo de aprendizaje supervisado, regresión logística, para predecir la tasa de churn. En el sector de las telecomunicaciones, la ciencia de datos permite analizar grandes volúmenes de información generada por sus sistemas que facilita el análisis y comprensión más acertada del comportamiento de los clientes.

Aprendizaje Supervisado (Machine Learning)

Los algoritmos de aprendizaje supervisado permiten que los modelos aprendan patrones a partir de datos históricos y realicen predicciones sobre nuevos clientes (Atay, 2025) usando conjuntos de datos de entrada y salida etiquetados.

El aprendizaje supervisado contiene modelos de clasificación y de regresión. Los modelos de clasificación se caracterizan por aprender de los datos para prever un resultado o evento en el futuro y, si el resultado aban una clasificación binaria, es decir, una respuesta de dos valores posibles como si o no / verdadero o falso.

Regresión Logística

Entre estas técnicas de machine Learning supervisado, la regresión logística es un algoritmo de clasificación que destaca por su facilidad de implementación, su simplicidad, robustez y su capacidad de interpretar la influencia de cada variable sobre la probabilidad de churn. No solo predice, sino que deja ver la dirección de asociación (positiva o negativa) de cada variable y esa cualidad la convierte en una elección, precisa, puede decirse, para investigaciones que requieren claridad, trazabilidad e interpretación de los resultados como este proyecto.

Una de las desventajas de la regresión logística es la suposición de linealidad entre la variable dependiente y las variables independientes, los problemas no lineales no se pueden resolver con regresión logística porque tiene una superficie de decisión lineal. La regresión logística requiere multicolinealidada media o nula entre variables independientes

Árboles de Decisión

Un árbol de decisiones es un modelo de predicción utilizado para clasificación y regresión. Utiliza una estructura similar a la de un árbol para describir un conjunto de datos y las soluciones se pueden visualizar siguiendo diferentes rutas a través del árbol.

Es un conjunto jerárquico de reglas que explican la forma en que un gran conjunto de datos se puede dividir en particiones de datos más pequeñas. Cada vez que se produce una división, los componentes de las particiones resultantes se vuelven cada vez más similares entre sí con respecto al objetivo.

Ventajas:

Son simples de entender y de interpretar.

No requiere una preparación de los datos demasiado exigente.

Se puede trabajar tanto con variables cuantitativas como cualitativas

Desventajas:

Los árboles de decisión tienden al sobre entrenamiento, especialmente cuando el número de características predictivas es alto.

Son inestables: cualquier pequeño cambio en los datos de entrada puede suponer un árbol de decisión completamente diferente.

No se puede garantizar que el árbol generado sea el óptimo.

Si hay clases dominantes es fácil que los árboles se generen sesgados, por lo que se recomienda balancear el conjunto de datos antes de entrenar el modelo.

Random Forest (Bosques Aleatorios)

Es un algoritmo de aprendizaje automático que combina múltiples clasificadores en forma de árboles de decisión. Cada árbol en el bosque se construye de manera independiente a partir de un conjunto de datos muestreado mediante bootstrap y utilizando un subconjunto aleatorio de características en cada división de nodo. Esto introduce variabilidad en los árboles individuales, reduciendo la correlación entre ellos y mejorando el desempeño del modelo.

Dentro de sus principales ventajas se tiene que, al implementar un promedio de varios árboles, reduce la posibilidad de sobreajuste en comparación con un solo árbol de decisión y, es capaz de manejar grandes conjuntos de datos con muchas características y detectar interacciones complejas entre ellas.

Si bien, la combinación de arboles resulta ser una ventaja, también llega a ser complejo entender cómo el modelo llega a sus decisiones, necesita más tiempo y recursos para ser entrenado y realizar las predicciones.

Tasa de Abandono (Churn-Rate)

La tasa de abandono o churn-rate, es una métrica que mide el porcentaje de clientes que dejan de utilizar los servicios de una empresa en un período determinado. Generalmente, lo evalúan empresas que ofrecen servicios por suscripción, ya que esas organizaciones necesitan construir una relación duradera y contractual con sus clientes. En este contexto, es una tasa esencial para identificar problemas de satisfacción del cliente, calidad del servicio o alta competencia en el mercado. Para empresas de servicios como los proveedores de internet, el churn-rate es esencial para evaluar la estabilidad financiera y el crecimiento sostenible.

Reducirlo ayuda a mantener la base de clientes, reducir costos asociados con la adquisición de nuevos suscriptores y mejorar la rentabilidad. Este valor se calcula dividiendo la cantidad de clientes perdidos en un periodo de tiempo entre la cantidad de clientes totales al inicio del mismo periodo y multiplicarlo por 100.

$$ChurnRate = \frac{cantidad\ de\ clientes\ perdidos\ en\ un\ periodo\ de\ tiempo}{Cantidad\ de\ clientes\ totales\ al\ inicio\ del\ mismo\ periodo}\ x\ 100$$

Teoría del Comportamiento del Cliente

La teoría del comportamiento del cliente nos recuerda que detrás de cada dato hay una emoción percepciones y pequeñas decisiones cotidianas que influyen en la decisión y la forma en que el usuario evalúa el servicio, cuánto siente que paga y cuán cómodo se siente con su contrato influyen en su deseo de permanecer o abandonar la empresa. Variables como los costos mensuales, el tipo de contrato o la duración de la relación con el proveedor reflejan esta dinámica y se correlacionan, a veces de forma sorprendente, con la probabilidad de churn

(Manďák, 2019). Aquí se hace evidente que la retención no es solo un asunto técnico, sino también emocional y experiencial.

En problemas de churn, los clientes que abandonan representan generalmente una minoría de la población, generando desbalance de clases. La teoría sugiere el uso de técnicas de balanceo de clases y métricas como precision, recall, F1-score y ROC-AUC para garantizar que el modelo aprenda de manera equitativa de ambas clases y pueda identificar correctamente a los clientes de riesgo (Fundación Universitaria Konrad Lorenz, 2022); (Nurtriana, 2024)

Finalmente, este marco teórico justifica la aplicación práctica del modelo optimizado para segmentar clientes, priorizar estrategias de fidelización y reducir pérdidas financieras, generando valor tangible para la empresa, en la literatura demuestra que modelos predictivos basados en machine learning y regresión logística son efectivos para anticipar el abandono de clientes en telecomunicaciones (L. F. Khalid, 2021); (Wagh, 2024)

Los antecedentes empíricos muestran que integrar variables como tipo de contrato, tenure, pagos y cargos mensuales mejora la capacidad de predicción y permite acciones de retención más precisas (Olle, 2014); (Nurtriana, 2024).

El uso de métricas de evaluación adaptadas a problemas desbalanceados, como F1-score y ROC-AUC, garantiza que el modelo no solo prediga correctamente la clase mayoritaria, sino que identifique de manera confiable a los clientes en riesgo.

Marco Conceptual

Se establecen los siguientes términos y definiciones que facilitan la interpretación de los resultados y el uso del modelo desarrollado.

Churn

Se define como la pérdida de clientes de un servicio o empresa durante un periodo determinado. En este proyecto es la variable objetivo del modelo predictivo y permite identificar clientes con mayor riesgo de abandono.

Cliente de Alto Riesgo

Cliente que presenta una probabilidad elevada de churn según el modelo predictivo. Este concepto es clave para la priorización de estrategias de retención.

Regresión Logística

Modelo estadístico supervisado utilizado para predecir variables binarias. En el contexto de churn, estima la probabilidad de que un cliente abandone la empresa en función de características como tenure, tipo de contrato y cargos mensuales.

Precision

Medida que indica qué proporción de las predicciones positivas realizadas por el modelo son correctas, se usa para evaluar la efectividad en identificar clientes que realmente abandonarán el servicio.

Recall

Métrica que indica qué proporción de los clientes que en realidad abandonan fueron correctamente identificados por el modelo, permitiendo evaluar la capacidad de captura de la clase minoritaria.

F1-Score

Es una métrica que se usa para evaluar la precisión de un modelo de clasificación, ya que combina en un solo valor las medidas de precision y exhaustividad (recall), se considera una

medida equilibrada entre ambas, especialmente útil cuando los datos están desbalanceados o cuando el costo de los errores es elevado

ROC-AUC

Área bajo la curva ROC, que mide la capacidad del modelo para distinguir entre clientes que abandonan y clientes que permanecen, representando la eficiencia global del modelo.

Overfitting (Sobreajuste)

Situación en la que un modelo se ajusta demasiado a los datos de entrenamiento, capturando ruido y patrones específicos, lo que reduce su capacidad de generalización a nuevos datos.

Balanceo de Clases

Estrategia utilizada para ajustar el aprendizaje del modelo cuando las clases no están equilibradas, asegurando que los clientes minoritarios (los que abandonan) sean correctamente aprendidos por el algoritmo.

Segmentación de Clientes

Proceso de clasificación de clientes según características y comportamiento de riesgo, permitiendo diseñar estrategias de fidelización y retención personalizadas.

Predicción Individual

Capacidad del modelo para asignar a cada cliente una probabilidad específica de churn, lo que facilita la toma de decisiones proactivas sobre estrategias de retención y asignación de recursos.

Metodología

Tipo de Investigación

Esta investigación adopta un enfoque cuantitativo, analítico y predictivo, basado en técnicas de ciencia de datos y analítica predictiva orientado a la construcción de un modelo de machine learning capaz de estimar la probabilidad de abandono (churn) en clientes del sector de telecomunicaciones. El desarrollo se basó en los principios del ciclo CRISP-DM e incluye etapas de comprensión del negocio, recolección y preparación de datos, modelado, evaluación e implementación. Se implementaron técnicas de imputación, transformación y depuración de datos como correlación para seleccionar variables relevantes. Posteriormente, se entrenaron y evaluaron distintos modelos de clasificación, tales como regresión logística, árboles de decisión y random forest, utilizando métricas de desempeño como sensibilidad, precisión, F1- score y área bajo la curva ROC (AUC) para seleccionar el modelo con mejor rendimiento

Este proyecto adopta un enfoque cuantitativo, ya que su análisis se fundamenta en variables numéricas y categóricas provenientes del comportamiento real de clientes. El estudio es también analítico, ya que evalúa patrones y relaciones entre características como cargos mensuales, antigüedad, métodos de pago o tipo de servicio, y es predictivo, pues el objetivo es anticipar la ocurrencia de churn mediante algoritmos de aprendizaje supervisado.

Este tipo de enfoque ha sido utilizado en estudios recientes de predicción de Churn en telecomunicaciones, como (Atay, 2025) y (Wagh, 2024), quienes destacan que los métodos cuantitativos permiten identificar patrones robustos y comparables entre modelos.

Método

El proceso metodológico siguió las fases del estándar CRISP-DM (Cross Industry Standard Process for Data Mining), el cual es considerado uno de los marcos más utilizados en investigaciones científicas y prácticas empresariales.

Comprensión del Negocio

Se definió como objetivo principal predecir qué clientes presentan mayor probabilidad de abandonar la compañía, con el fin de anticipar comportamientos de churn y apoyar la toma de decisiones estratégicas orientadas a la retención. Con este proceso se busca identificar segmentos críticos, estimar el impacto financiero de la pérdida de clientes y priorizar acciones como campañas personalizadas, mejoras de servicio o ajustes en la oferta comercial.

Comprensión de los Datos

Se exploró la estructura del dataset, distribuciones, valores atípicos, patrones de churn y correlaciones. Esta fase permitió detectar posibles inconsistencias y diferencias estructurales entre clientes que permanecen y los que abandonan. Como resalta (Nurtriana, 2024)el análisis exploratorio es fundamental para una correcta modelación permitiendo formular hipótesis, seleccionar variables significativas y orientar las características, reduciendo así errores en etapas posteriores.

Preparación de Datos

La preparación del dataset incluyó tareas de limpieza para corregir valores faltantes e inconsistencias, imputación de datos para evitar pérdida de información, normalización mediante Z-score y codificación de variables categóricas a través de One-Hot Encoding. Adicionalmente, se abordó el problema del desbalance de clases mediante la aplicación de class weights. Este

enfoque ha sido validado por (Manďák, 2019) en implementación de regresión logística para Churn, demostrando mejoras significativas en métricas sensibles a la clase minoritaria.

Modelado

En esta fase, se exploran y desarrollan diferentes algoritmos de machine learning para para predecir el riesgo de abandono de clientes. Se aplicarán varias técnicas para determinar qué modelo es el más adecuado

Se probó inicialmente un modelo base para establecer un punto de referencia y posteriormente se aplicó RandomizedSearchCV para optimizar los hiperparámetros. Este enfoque permitió explorar de manera eficiente un amplio espacio de configuraciones sin incurrir en los altos costos computacionales de métodos exhaustivos como GridSearchCV, además de ser difícil de ejecutar cuando el ordenador no cumple con los recursos suficientes. Esta metodología ha sido destacada en investigaciones como la de (Adeniran, 2024), quienes recomiendan técnicas de búsqueda aleatorias para problemas de telecomunicaciones debido a su equilibrio entre precisión y eficiencia operativa.

Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., Abhulimen, A. O., & OneAdvanced, U. (2024). Implementing machine learning techniques for customer retention and churn prediction in telecommunications. *Computer Science & IT Research Journal*, *5*(8), 2011-2025.

Evaluación

Se evalúan los modelos desarrollados en función de las métricas establecidas, determinando cuál es el más adecuado para la predicción del churn-rate en función de su desempeño.

Para evaluar el desempeño del modelo se utilizaron métricas como F1-score, Recall y AUC-ROC, debido a que resultan más informativas con clases desbalanceadas. El Recall permite

identificar correctamente la clase minoritaria, siendo importante cuando se busca minimizar la omisión de eventos relevantes, por otra parte, el F1-score equilibra dicho indicador con la precisión. Se usó el AUC-ROC que ofrece una visión general de la capacidad discriminativa del modelo. Estas métricas son recomendadas en estudios como los de (Atay, 2025) y (Manďák, 2019), quienes evidencian su utilidad donde la detección de eventos críticos es prioritaria.

Técnicas Utilizadas

Para el desarrollo del modelo predictivo se empleó la regresión logística, esta es una técnica estadística ampliamente utilizada en problemas de clasificación binaria debido a su capacidad para modelar la probabilidad de ocurrencia de un evento. Esta técnica es útil cuando se busca interpretar el impacto que tienen las variables independientes sobre la variable objetivo, permite estimar odds ratios y analizar cómo cambios pequeños en un predictor influyen en la probabilidad final.

Entre sus principales ventajas destacan su interpretabilidad, ya que facilita la comprensión del efecto de cada predictor mediante los coeficientes y los odds ratios, su robustez frente a datos ruidosos, y su eficiencia computacional, que la hace adecuada incluso con recursos limitados. Además, la regresión logística presenta un buen desempeño cuando las relaciones entre las variables son aproximadamente lineales en el logit, lo que ayuda a la estabilidad y generalización del modelo.

Adicionalmente, se aplicaron técnicas complementarias como la evaluación mediante curvas ROC, análisis de importancia de variables, gráficos de distribución de probabilidades y herramientas de desempeño como el Lift Chart y el Gains Chart, que permitieron medir la capacidad discriminativa del modelo, validar su utilidad práctica y determinar qué tan bien logra clasificar a los casos positivos frente a métodos aleatorios. Estas técnicas en conjunto fortalecen

la calidad del análisis al ofrecer una visión integral del rendimiento y utilidad del modelo en escenarios reales.

Recolección de Datos

Se utilizó la base de datos Telco Churn 7k, disponible en la plataforma Hugging Face: https://huggingface.co/datasets/mnemoraorg/telco-churn-7k). Licencia: ECL-2.0.

Este dataset contiene 7043 clientes de una empresa de telecomunicaciones, para analizar los factores que influyen en la deserción, cuenta con 18 variables categóricas, 2 númericas y 1 identificador. Estas variables incluyen datos relacionados al tipo de servicio contratado, medios de pago y meses contratados.

A continuación, se presenta un diccionario de datos que describe el significado de cada variable y su tipo:

Tabla 1

Diccionario de Datos de las Variables

Variable	Descripción	Tipo
customerID	Identificador único del cliente	Identificador
gender	Género del cliente	Categórica
SeniorCitizen	Identifica si el cliente es adulto mayor (1= sí, 0=no)	numérica binaria
Partner	Si el cliente tiene pareja (Yes/No)	Categórica
Dependents	Si el cliente tiene dependientes (Yes/No)	Categórica
tenure	Meses que el cliente ha permanecido con la empresa.	Categórica

Variable	Descripción	Tipo
PhoneService	Disponibilidad de servicio telefónico	Categórica
MultipleLines	Tiene múltiples líneas telefónicas	Categórica
InternetService	Tipo de servicio de internet	Categórica
OnlineSecurity	Servicio de seguridad en línea contratado	Categórica
DeviceProtection	Protección de dispositivo contratada	Categórica
TechSupport	Soporte técnico disponible	Categórica
StreamingTV	Servicio de streaming de TV contratado	Categórica
StreamingMovies	Servicio de streaming de películas contratado	Categórica
Contract	Tipo de contrato (Month-to-month, One year…)	Categórica
PaperlessBilling	Uso de facturación electrónica	Categórica
PaymentMethod	Método de pago (Electronic check, Credit card)	Categórica
MonthlyCharges	Cargos mensuales del cliente	Numérica

Variable	Descripción	Tipo
TotalCharges	Total de cargos acumulados durante el servicio	Numérica
Churn	Variable objetivo: indica si el cliente desero.	Categórica

Limpieza y Calidad de los Datos

Durante la etapa de preparación de datos, la variable TotalCharges se encontraba en formato object debido a la presencia de registros vacíos, por tanto, los valores en blanco fueron reemplazados por NaN y posteriormente la variable fue convertida al tipo float64.

Esta transformación permitió integrar el campo en los análisis estadísticos y en los modelos supervisados como una medida numérica continua.

Se realiza la descripción de las variables numéricas presentes en la base de datos con el objetivo de identificar su comportamiento general, evaluar posibles sesgos, rangos de valores y presencia de valores atípicos. Esta revisión permite detectar distribuciones no normales y orientar decisiones posteriores en cuanto a transformaciones o imputaciones necesarias. A continuación, se presentan las principales estadísticas descriptivas para estas variables.

Tabla 2

Estadísticas de las Variables Numéricas

Variable	Cantidad	Media	Desviación estándar	Mínimo	Máximo
Tenure	7<,043	-2.42e-17	1.000071	-1.318165	1.613701

Variable	Cantidad	Media	Desviación estándar	Mínimo	Máximo
MonthlyCharges	7,043	-6.41e-17	1.000071	1.000071	1.794352
TotalCharges	7,043	4.54e-18	1.000071	-1.000507	2.824678

Como parte de la validación de consistencias se ajustaron las variables relacionadas con servicios de Internet y telefonía. Para los clientes sin servicio de Internet, se reemplazaron todos los valores distintos de “No” en las variables OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV y StreamingMovies. También se estandarizó la variable MultipleLines, garantizando que no se asignaran líneas múltiples a clientes sin servicio telefónico.

Finalmente, se verifica la ausencia de registros duplicados y se revisaron las variables categóricas para asegurar uniformidad en los valores, evitando inconsistencias de escritura o formato. Esto permite obtener un conjunto de datos limpio y coherente para la etapa de preparación y análisis, esto garantiza la confiabilidad del proceso de modelado.

Resultados

Análisis Exploratorio de Datos (EDA)

En esta fase se realizó el análisis Exploratorio de datos con el objetivo de comprender la distribución de las variables, identificar patrones relevantes y detectar posibles problemas de calidad de los datos que pudieran afectar el modelado predictivo. Se utilizaron técnicas de visualización como histogramas, gráficos de dispersión y mapas de calor (heatmaps) para observar correlaciones y distribuciones.

Principales hallazgos:

La variable objetivo Churn muestra un desbalance, con una mayor proporción de clientes que no abandonan el servicio (No) frente a los que sí lo hacen (Yes). Esto indica que el modelo deberá considerar este desbalance para mejorar la predicción de clientes que abandonan.

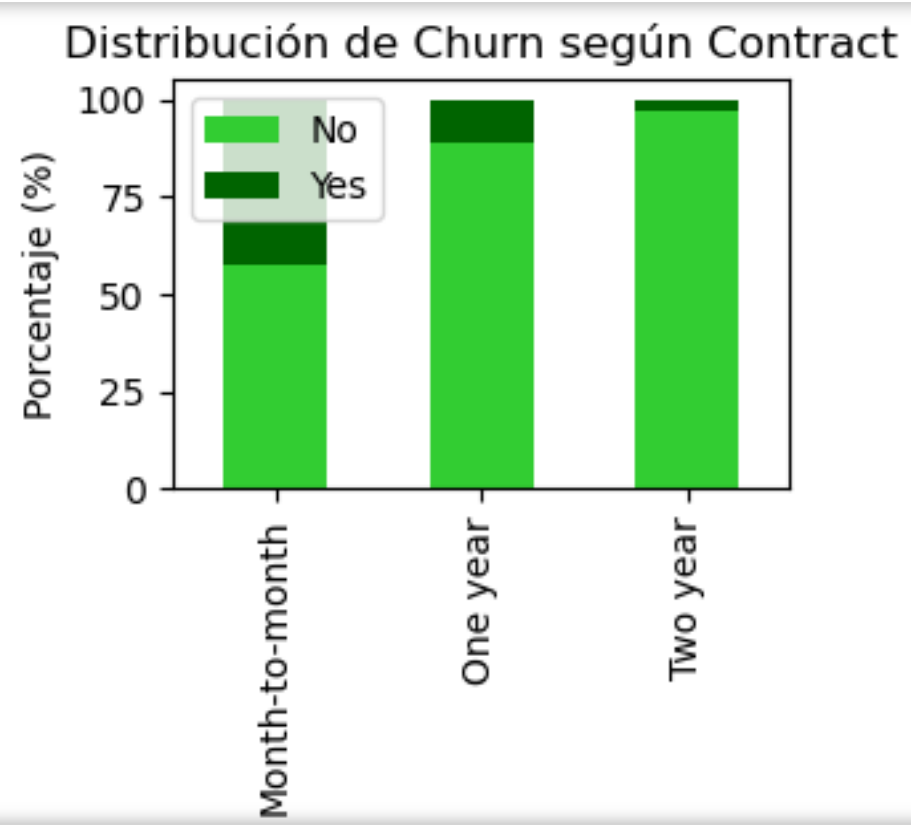
Variables categóricas como Contract, InternetService, PaymentMethod, OnlineSecurity y TechSupport muestran diferencias en la tasa de churn, como los clientes con contratos de mes a mes o que no tienen servicios de soporte técnico o seguridad onlin**e** presentan mayores probabilidades de churn.

Distribución de Churn Según la Variable Contract

En la *Figura 1* se observa el comportamiento en este tipo de servicios, los contratos a mayor tiempo suponen un mayor compromiso para las clientes y normalmente, viene con incentivos que hacen que mantengan la permanencia que supone tener el servicio. Desde la perspectiva del negocio, esto sugiere que las estrategias de retención deberían enfocarse en convertir contratos mensuales en anuales o bianuales.

Figura 1

Distribución de Churn por Contrato



Month-to-month: 42.7% de los clientes con contratos mensuales abandonan el servicio.

One year: 11.3% de los clientes con contrato anual abandonan el servicio.

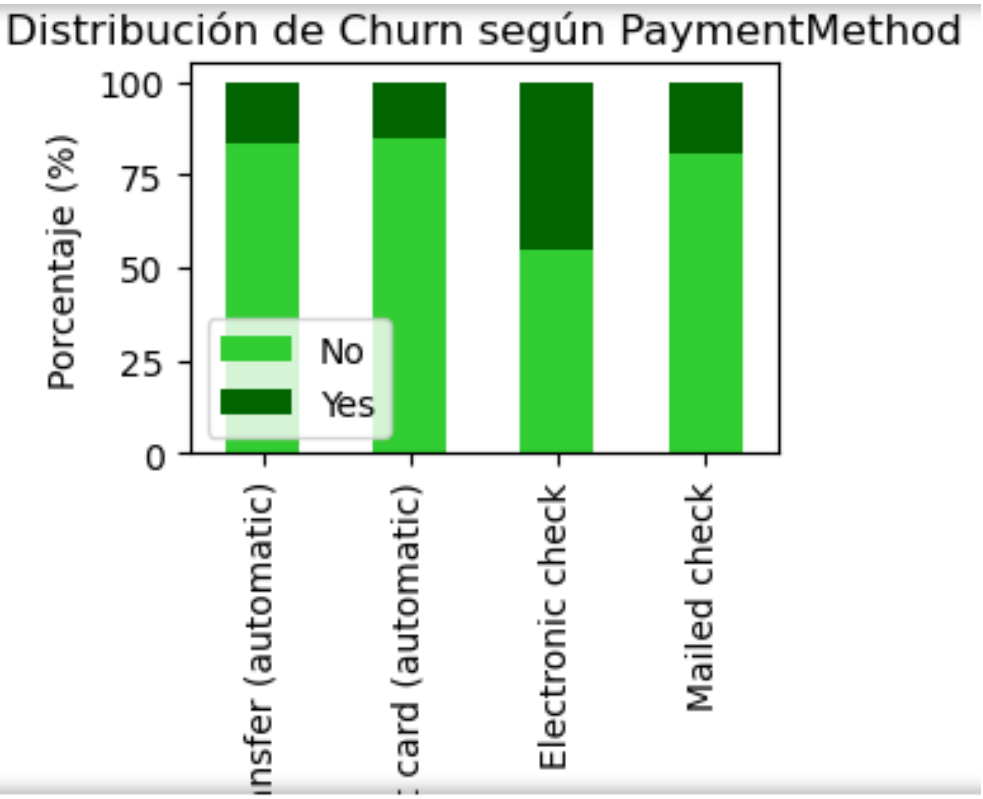
Two year: Solo 2.8% de los clientes con contrato de dos años presentan churn.

Distribución de Churn Según la Variable PaymentMethod

En la *Figura 2* figura 2 se puede observar en los resultados que los clientes que utilizan métodos de pago electrónicos no automatizados presentan tasas de abandono más altas a aquellos con pagos automáticos.

Figura 2

Distribución de Churn por Método de Pago



Distribución de Churn Según la variable IntenetService

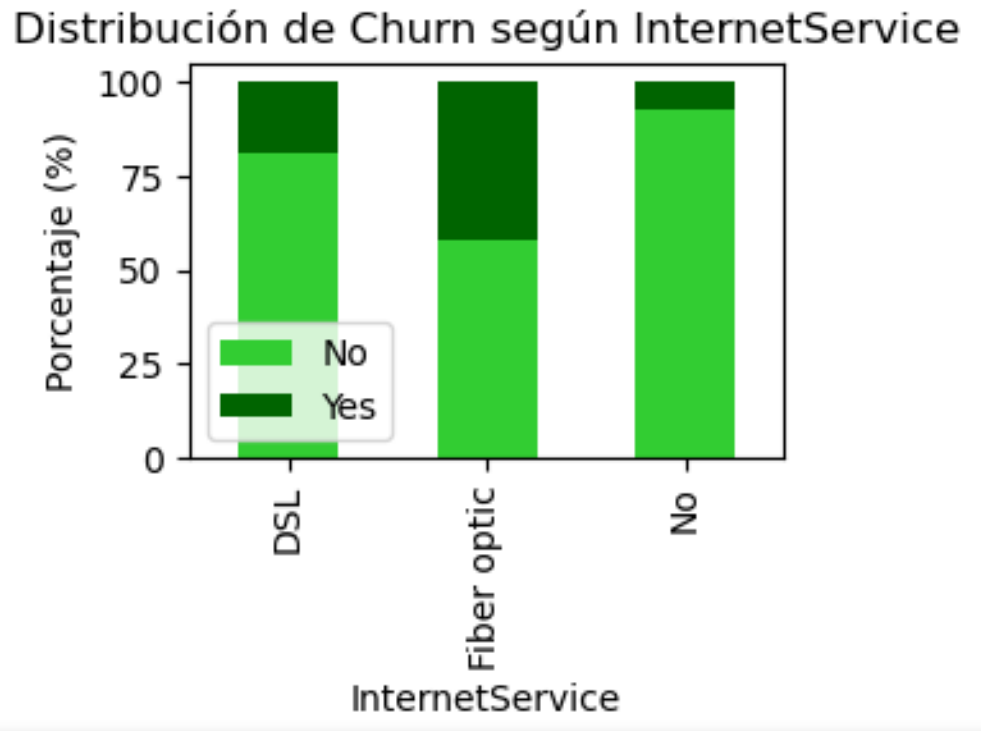
En la .

Figura 3 se observa que la probabilidad de abandono es mayor en clientes que poseen servicios de internet más complejos y costosos, lo que podría estar relacionado con expectativas de servicio o insatisfacción con la calidad. También, puede darse que, al mantener servicios de

solo telefonía o televisión, tienen menos exposición a ofertas competitivas y menores expectativas.

Figura 3

Distribución de Churn por Tipo de Servicio de Internet



Fiber optic: 41.9% de los clientes con fibra óptica presentan churn.

DSL: 19.0% de los clientes con DSL abandonan.

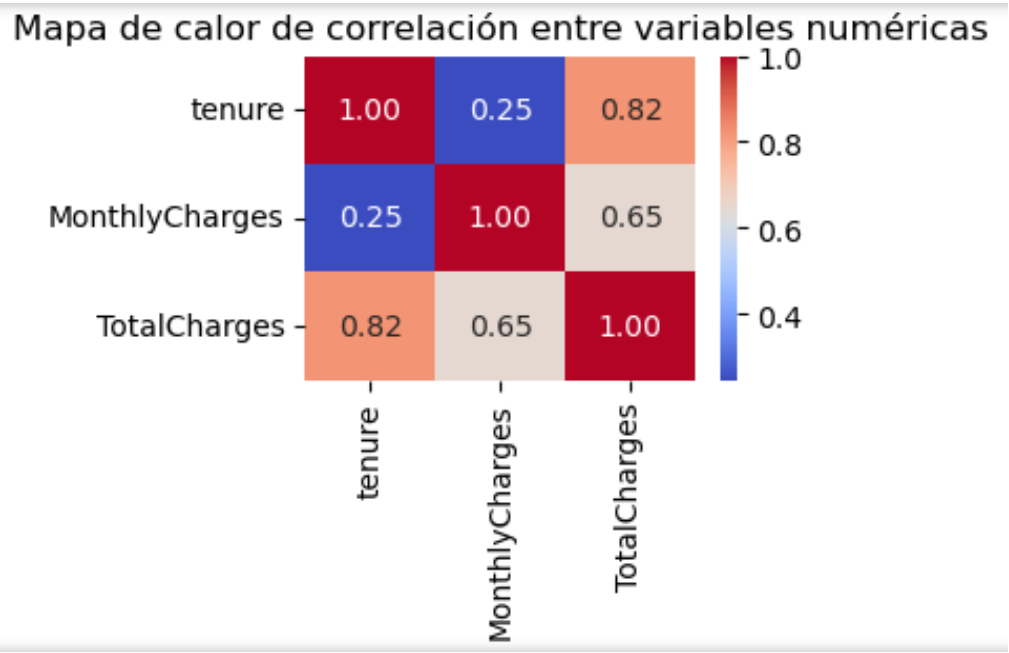
No internet: Solo 7.4% de los clientes sin servicio de internet presentan churn.

Entre las variables numéricas, tenure, MonthlyCharges y TotalCharges presentan correlaciones con churn. Los clientes con menor antigüedad (tenure baja) y mayores cargos

mensuales (MonthlyCharges) tienden a abandonar más el servicio. Para comprender las relaciones de estas variables se construyó un mapa de calor que permite visualizar las correlaciones en la *Figura 4*.

Figura 4

Heatmap de Correlacion Entre Variables Numéricas



Tenure vs TotalCharges: Correlación positiva con un coeficiente de aproximadamente 0.83, muestra una relación casi fuerte, mientras más tiempo permanezca un cliente con la empresa, mayor será el monto total que habrá pagado.

En este caso, un cliente antiguo con servicios básicos tendrá un tenure alto, pero TotalCharges moderado, mientras que un cliente relativamente nuevo, pero con Múltiples servicios premium presentará el patrón inverso.

MonthlyCharges vs TotalCharges: Correlación moderada (0.65), mostrando que cargos mensuales más altos contribuyen a facturación total más elevada, pero no necesariamente afectan directamente el churn.

Sugiere que existen múltiples trayectorias hacia una facturación total elevada ya que algunos clientes llegan mediante cargos altos mensuales, pero permanencias cortas, mientras que otros lo logran con cargos moderados pero gran antigüedad. Esta heterogeneidad es importante para el objetivo predictivo, pues indica que MonthlyCharges aporta información adicional más allá de lo que ya se conoce por tenure y TotalCharges.

Tenure vs MonthlyCharges: Tienen una correlación baja, lo que indica que la duración del cliente no depende de sus cargos mensuales, pero la combinación de ambas variables es útil para predecir churn.

Al combinar ambas variables en el modelo predictivo, cada una aporta información única y no redundante. Un cliente puede tener alta antigüedad con bajos cargos mensuales, o viceversa, y cada combinación puede asociarse con diferentes niveles de riesgo de abandono.

La correlación entre las variables numéricas muestra relaciones moderadas, por ejemplo, entre tenure y TotalCharges (clientes con más tiempo tienden a tener mayor facturación acumulada), lo cual ayuda en la selección de variables para el modelado predictivo.

Modelado Predictivo

Una vez realizado el análisis Exploratorio y el análisis de las variables se procede a la fase de análisis predictivo. Antes de ello, se llevó a cabo la preparación de los datos para garantizar la calidad de los resultados.

Para los algoritmos de aprendizaje automático se requiere que toda la información este en forma numérica, por tanto, algunas variables categóricas es necesario transformarlas a

numéricas. Para la transformación de las variables se aplica la técnica de One-Hot-Encoding ya que este método crea una columna binaria por cada categoría posible donde cada cliente tendrá un 1 en la columna correspondiente a su tipo de contrato y un 0 en las demás.

Posteriormente, se dividió el dataset en 80% para entrenamiento y 20% para prueba. Por un lado, se busca que el modelo tenga acceso a la mayor cantidad de datos posible durante su entrenamiento, ya que más datos generalmente conducen a un mejor aprendizaje de patrones. Por otro lado, se necesita que el modelo nunca haya visto, para poder evaluar su capacidad de generalización a nuevos casos.

Para la división del dataset se tuvo en cuenta la estratificación sobre la variable “stratify = y”, ya que la variable churn tuvo un desbalance, por tanto, la estratificación asegura que ambos conjuntos mantengan la misma proporción de clientes que abandonan y clientes que permanecen, asegurando que el modelo entrene y se evalúe bajo condiciones comparables.

Se realizó normalización de variables numéricas (tenure, MonthlyCharges, TotalCharges) usando StandardScaler, ya que las variables presentan disparidad en escaladas que puede afectar la predicción en el modelo de regresión logística.

Esta técnica hace que cada variable tenga media 0 y desviación estándar de 1, poniendo las variables en una escala común, permitiendo que el modelo las trate igual y, da la ventaja que al estar normalizadas los coeficientes estimados del modelo se vuelven directamente comparables, permitiéndonos identificar qué variables ejercen mayor influencia sobre la predicción.

Entrenamiento de los Modelos Base

Posterior a l preparación de los datos, se realizó el entrenamiento de tres modelos de clasificación diferentes, los modelos seleccionados fueron:

Regresión Logística: Un modelo lineal que a pesar de su simplicidad obtiene resultados sólidos y tiene la ventaja de ser altamente interpretable.

Árbol de Decisión: este modelo aprende reglas de decisión jerárquicas, capturando relaciones no lineales de forma natural.

Random Forest: usa múltiples árboles de decisión que combina sus predicciones para obtener resultados más robustos.

Los resultados de las métricas iniciales fueron:

Tabla 3

Métricas Iniciales de los 3 Modelos

Modelo	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.801	0.647	0.553	0.597	0.841
Decision Tree	0.774	0.588	0.503	0.542	0.688
Random Forest	0.796	0.669	0.455	0.541	0.840

Los resultados muestran que la Regresión Logística es el modelo más equilibrado y efectivo para implementar en el poryecto. Con una puntuación F1 de 0,597, tiene la mejor combinación entre precisión (64,7%) y recuperación (55,3%). Su ROC-AUC de 0.841 también es el más alto junto con Random Forest, indicando una capacidad robusta para distinguir entre clientes que abandonarán y los que permanecerán.

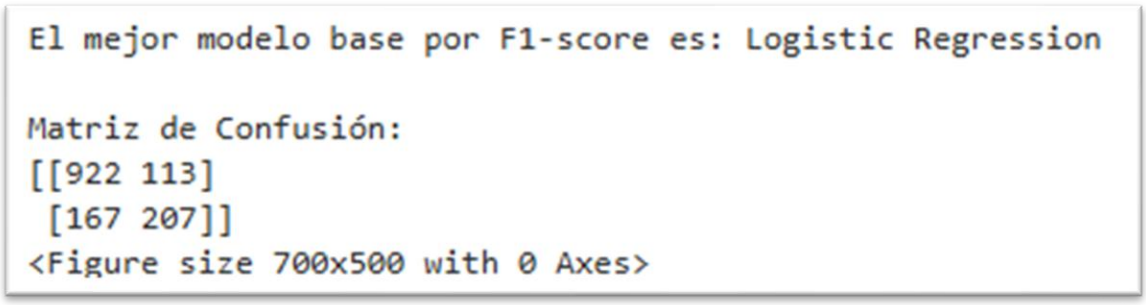
El Árbol de Decisión tiene un desempeño moderado en casi todas las métricas. Su ROC-AUC de 0.688 es inferior al de los otros modelos, indicando menor capacidad discriminatoria, puede deberse a que los árboles individuales tienden a sobreajustarse a los datos de entrenamiento.

Por otro lado, el Random Forest , aunque tiene la precisión más alta con 66,9%, presenta el recuerdo más bajo con 45,5%, indicando que cuando el modelo predice que un cliente hará abandono, acierta con mayor frecuencia, pero a costa de dejar escapar muchos casos reales. Este patrón refleja un modelo que solo hace predicciones positivas cuando está muy seguro, lo que puede afectar la sensibilidad.

A continuación, se observa en la *Figura 5* la matriz de confusión de la regresión logística donde se observarán los aciertos y los errores del modelo.

Figura 5

Matriz de Confusión del Modelo Base de Regresión Logística



922 verdaderos Negativos: Clientes que no hicieron churn y el modelo correctamente predijo que no lo harían.

207 verdaderos Positivos: Clientes que sí hicieron churn y el modelo correctamente los identificó.

167 falsos Negativos: Clientes que hicieron abandono, pero el modelo no los detectó.

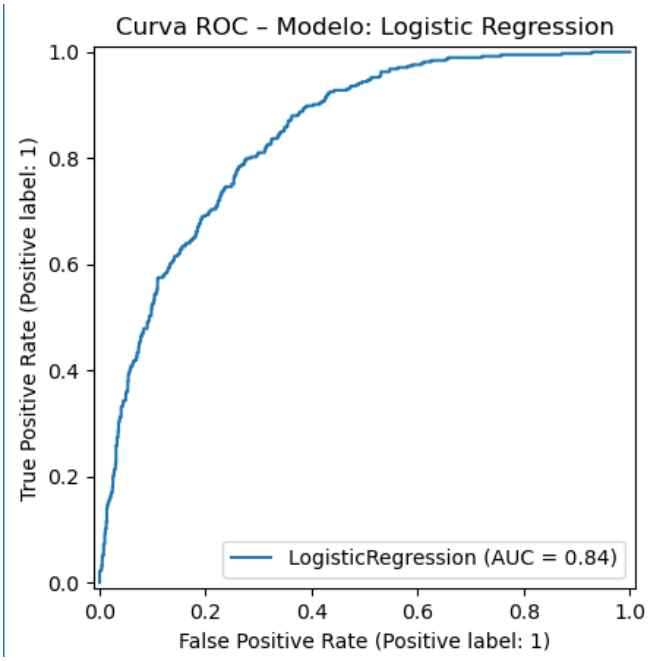
113 Falsos Positivos: Clientes que no hicieron abandono, pero el modelo predijo que sí.

Los clientes que están en riesgo de hacer Churn son los 167 falsos positivos que el modelo no logró identificar, esto indica que debe realizar optimización en este punto.

Para visualizar la capacidad de discriminación del modelo se muestra la **Figura *6*** de la curva de ROC que deja ver el rendimiento del modelo a través de las posibles clasificaciones.

Figura 6

Curva ROC del Modelo Base de Regresión Logística



La regresión logística obtuvo un **ROC-AUC de 0.841**, es decir, si tomamos aleatoriamente un cliente que hizo abandono y otro que no, existe un 84.1% de probabilidad de que el modelo asigne una puntuación de riesgo más alto al primero. De otra forma, el modelo posee una capacidad notable para ordenar a los clientes según su riesgo real de abandono correcta.

Optimización del Modelo con RandomizedSearchCV

Para mejorar el desempeño del modelo base seleccionado de Regresión Logística, se implementó RandomizedSearchCV. Es una técnica de optimización de hiperparámetros que funciona seleccionando combinaciones aleatorias de valores dentro de un rango definido, en lugar de probar todas las combinaciones posibles, prueba solo algunas, pero de manera inteligente y eficiente, es más rápido, y escalable. Se eligió sobre GridSearchCV debido a su menor tiempo de cómputo y capacidad de cubrir más combinaciones posibles, era demasiado pesado para el equipo y no lograba completarse.

Entrenar el Modelo Final Optimizado

Los hiperparamétros utilizados fueron:

C: parámetro de regularización que controla la magnitud de los coeficientes. Se exploró en un rango continuo de 0.001 a 10 utilizando una distribución logarítmica (loguniform). Esto permite evaluar tanto regularización fuerte (valores pequeños de C) como débil (valores grandes).

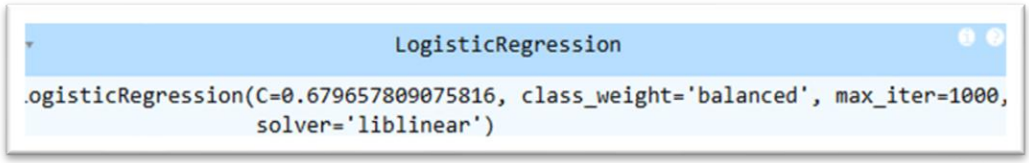
class_weight: se probó con None (sin balance) y balanced, indica que el modelo ajusta automáticamente los pesos de las clases, yaque presentó un desbalance.

La búsqueda se realizó con 10 combinaciones aleatorias y 3 validaciones cruzadas para reducir el costo computacional, utilizando F1-score como métrica principal debido al desbalance de clases y a la necesidad de equilibrar precisión y recall.

Resultados Obtenidos

Figura 7

Resultados de la Optimización del Modelo Final de la Regresión Logística



Esto indica que una regularización moderada y el ajuste del peso de las clases permiten mejorar la detección de clientes que abandonan el servicio, reduciendo los falsos negativos.

Métricas del Modelo Final

Tras la optimización del modelo se obtuvieron las siguientes métricas:

El Recall elevado (75.9%) indica que el modelo es capaz de identificar correctamente a la mayoría de los clientes que abandonan el servicio, lo cual es crítico para estrategias de retención.

La Precision moderada (52.2%) muestra que algunos clientes predichos como churn realmente no abandonan, un resultado aceptable dado que priorizamos capturar clientes en riesgo.

El F1-score (61.9%) refleja un balance adecuado entre precisión y recall, y la ROC-AUC (0.841) confirma que el modelo discrimina bien entre clientes churn y no churn.

Tabla 4

Métricas del Modelo Final de Regresión Logística

Métrica	Valor
Accuracy	0.752
Precision	0.522

Recall	0.759
F1-score	0.619
ROC-AUC	0.841

Interpretación en Contexto del EDA

Los hallazgos del análisis exploratorio se reflejan en el modelo optimizado con el siguiente análisis:

Los clientes con contratos month-to-month presentan una tasa de churn significativamente mayor (42.7%), comparado con clientes con contratos anuales o de dos años. Se infiere que los clientes con compromisos a corto plazo tienen menos fidelidad, probablemente porque perciben mayor flexibilidad y facilidad de cambio de proveedor.

El servicio de internet Fiber optic también se asocia con mayores tasas de churn (41.9%), posiblemente debido a expectativas de calidad más altas o costos asociados.

Los clientes que utilizan pagos electrónicos, especialmente mediante cheques electrónicos, muestran mayores tasas de abandono (45.3%). Esto puede reflejar que la falta de automatización o la percepción de menor control financiero contribuye a la insatisfacción.

La variable tenure indica la duración del cliente en la empresa, sigue siendo un factor determinante. Clientes recientes (tenure baja) presentan mayor riesgo de churn, especialmente si tienen MonthlyCharges elevados. Esto sugiere que los clientes nuevos pueden sentirse insatisfechos si perciben que el costo del servicio no corresponde a los beneficios recibidos.

Esta relación fue consistente con el heatmap de correlación, allí se observó que tenure y TotalCharges están altamente correlacionados, y que una combinación de tenure baja y cargos

mensuales altos incrementa la probabilidad de abandono. El modelo refleja esta relación, ajustando la probabilidad de churn según la interacción entre estas variables.

La variable objetivo Churn presenta un desbalance natural con aproximadamente 26% de clientes que abandonan frente a 74% que permanecen. Al ajustar class_weight='balanced', el modelo asigna mayor peso a la clase minoritaria durante el entrenamiento dejando que el modelo aprenda patrones más representativos para los clientes en riesgo.

Esto se refleja en el Recall elevado (75.9%), lo que indica que el modelo es capaz de identificar correctamente a la mayoría de los clientes que churn, reduciendo falsos negativos y facilitando la toma de decisiones estratégicas para retención.

Simulación de una Predicción

Para demostrar la aplicabilidad práctica del modelo optimizado, se realizó una simulación de predicción de churn para clientes individuales. Esta simulación permite identificar a los clientes con mayor probabilidad de abandono y diseñar estrategias de retención específicas. Seleccionamos múltiples clientes.

Se obtienen tanto la probabilidad de churn como la predicción de la clase para cada cliente generando una lista de 10 clientes.

Figura 8

Lista de Clientes con Probabilidad de Churn 1

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	PaperlessBilling	MonthlyCharges	TotalCharges	customerID_0003-MKNFE	...	TechSupport_Yes	Strea
437	1	0	1	1	72	1	1	114.05	8468.20	False	...	True	
2280	0	1	0	0	8	1	1	100.15	908.55	False	...	True	
2235	0	0	1	1	41	1	1	78.35	3211.20	False	...	False	
4460	1	0	1	0	18	1	0	78.20	1468.75	False	...	True	
3761	0	0	1	0	72	1	1	82.65	5919.35	False	...	False	
5748	0	0	0	0	21	1	1	99.85	1992.55	False	...	False	
3568	0	0	0	0	21	1	1	99.15	1956.40	False	...	True	
2976	1	0	0	1	19	1	1	24.10	439.20	False	...	False	
5928	0	0	1	1	61	1	0	19.75	1311.60	False	...	False	
1639	1	1	0	0	17	1	1	45.05	770.60	False	...	False	

Figura 9

Lista de Clientes con Probabilidad de Churn 2

StreamingTV_Yes	StreamingMovies_Yes	Contract_One year	Contract_Two year	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check	Prob_Churn	Predicción
True	True	False	True	True	False	False	0.099974	0
True	True	False	False	True	False	False	0.843363	1
True	False	True	False	True	False	False	0.140944	0
False	False	False	False	False	True	False	0.617263	1
True	True	False	True	True	False	False	0.054281	0
True	True	False	False	True	False	False	0.790312	1
True	True	False	False	False	False	False	0.664236	1
False	False	False	False	True	False	False	0.261840	0
False	False	False	True	True	False	False	0.008960	0
False	False	False	False	False	True	False	0.627600	1

Se observan los siguientes patrones:

Clientes como 2280, 4460, 5748 y 1639 muestran alta probabilidad de churn (>60%), confirmando que clientes recientes con contratos de corta duración son los más vulnerables.

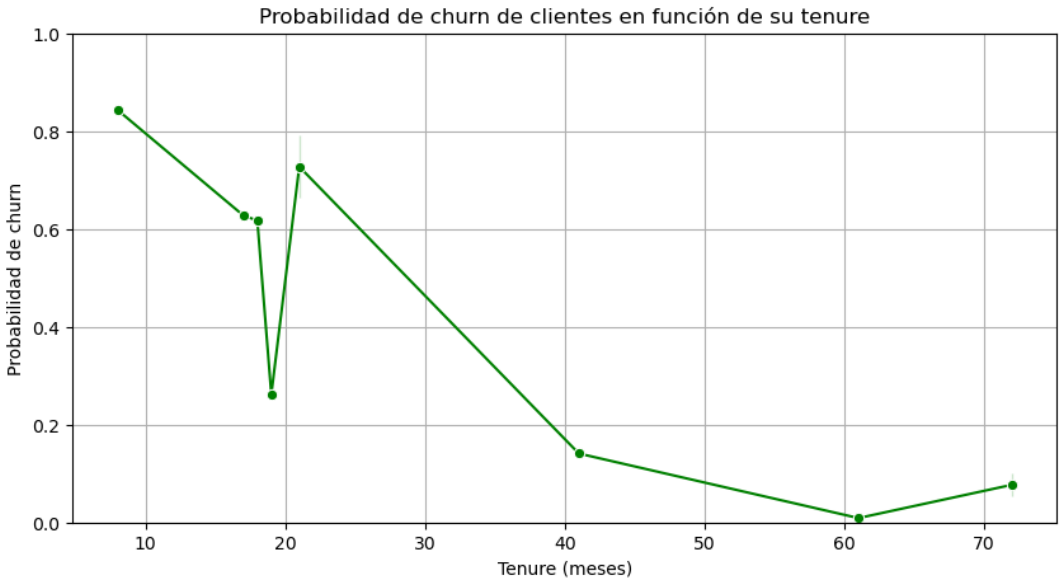
Los clientes con cargos altos (2280: 100.15, 5748: 99.85) presentan mayores probabilidades de abandono, reforzando la relación observada en el EDA.

Los clientes 437, 3761 y 5928 tienen tenure alto y contratos de 2 años, y sus probabilidades de churn son muy bajas (<10%), mostrando la fidelización asociada a contratos prolongados.

Clientes que usan electronic check muestran mayor riesgo en combinación con contratos cortos y tenure bajo (2280: 0.84, 1639: 0.63), mientras que pagos automáticos (credit card) y contratos largos reducen el riesgo.

Figura 10

Probabilidad de Churn Según Tenencia



Clientes con tenure bajo (<25 meses) tienen alta probabilidad de churn).

Clientes con tenure alto (>60 meses) muestran baja probabilidad de abandono.

Se confirma que tenure y tipo de contrato son variables determinantes para predecir churn.Conclusiones

El análisis exploratorio de datos permitió identificar que los clientes con contratos month-to-month, servicio de internet por fibra óptica y pagos mediante electronic check presentan la mayor probabilidad de abandono. Además, la variable tenure mostró ser un factor crítico ya que clientes recientes con altos cargos mensuales tienen mayor riesgo de churn. Estos hallazgos aportan un conocimiento concreto sobre los patrones de deserción y permiten que la empresa enfoque sus estrategias de retención en los segmentos de mayor riesgo, optimizando recursos y esfuerzos de marketing.

La regresión logística optimizada mediante RandomizedSearchCV demostró ser efectiva para anticipar el riesgo de abandono, alcanzando un F1-score de 0.6187 y un ROC-AUC de 0.8413, evidenciando su capacidad para discriminar correctamente entre clientes que abandonan y los que permanecen. Comparado con modelos base como Decision Tree y Random Forest, la regresión logística ofrece un balance adecuado entre interpretabilidad y precisión, lo que facilita entender cómo cada variable influye en la probabilidad de churn y permite una aplicación práctica de los resultados en la toma de decisiones estratégicas.

La simulación de clientes y la predicción individual muestran que es posible anticipar el abandono de manera confiable, permitiendo diseñar intervenciones proactivas como promociones de fidelización, ajustes en contratos o incentivos personalizados. Esto representa un aporte directo al problema empresarial, al transformar la gestión de clientes de una estrategia reactiva a una estrategia basada en datos, reduciendo costos asociados al churn y mejorando la rentabilidad a largo plazo.

Recomendaciones

Se recomienda que la organización utilice el modelo de regresión logística desarrollado como una herramienta operativa para la predicción del churn. Esto permitirá identificar clientes de alto riesgo y priorizar estrategias de retención, optimizando la asignación de recursos y reduciendo pérdidas económicas. Las áreas de marketing y atención al cliente pueden beneficiarse de esta información para diseñar promociones personalizadas, incentivos por fidelidad y planes contractuales a largo plazo.

El comportamiento de los clientes y las condiciones del mercado cambian constantemente, por lo que es recomendable actualizar periódicamente los datos y reentrenar el modelo. Esto garantizará que las predicciones se mantengan precisas y reflejen nuevas tendencias, cambios en tarifas, servicios o patrones de uso de los clientes.

Para mejorar la capacidad predictiva y la comprensión del churn, se sugiere incorporar nuevas variables que puedan influir en la deserción, como interacciones con el servicio de atención al cliente, historial de quejas, satisfacción del cliente o uso de servicios digitales. La integración de estas fuentes de datos permitirá un análisis más completo y aumentará la efectividad de las estrategias de retención.

Los estudios posteriores podrían profundizar en la segmentación de clientes según riesgo de abandono, análisis de cohortes y simulaciones de impacto de diferentes estrategias de retención. Asimismo, se puede investigar el uso de modelos de aprendizaje no supervisado para identificar patrones ocultos en los clientes que podrían no ser evidentes mediante modelos supervisados.

Referencias Bibliográficas

Adeniran, I. A. (2024). *Implementing machine learning techniques for customer retention.* doi:10.51594/csitrj.v5i8.1489

Atay, M. T. (2025). *Analysis of customer churn prediction using logistic regression, k-nearest neighbors, decision tree and random forest algorithms.* doi:10.17654/0972361725008

De Caigny, A. C. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European journal of operational research.* doi:10.1016/j.ejor.2018.02.009

Fundación Universitaria Konrad Lorenz, 2. (2022). *Modelo análisis predictivo para el cálculo de tasa de deserción en una empresa aseguradora*. Bogotá DC: Fundación Universitaria Konrad Lorenz, 2022.

Jain, H. K. (2020). Churn prediction in telecommunication using logistic regression and logiy boost. *Procedia Computer Science*. doi:10.1016/j.procs.2020.03.187

Kau, F. M. (2017). Service ... Proceddings, 100. doi:10.1000/21

Krishna, R. J. (2024). *Application of machine learning techniques for churn prediction in the telecom business. Results in Engineering*. doi:10.1016/j.rineng.2024.103165

L. F. Khalid, A. M. (2021). Customer Churn Prediction in Telecomunications Industry Based on Data Mining. *IEEE Symposium on Industrial Electronics & Applications*. doi:10.1109/ISIEA51897.2021.9509988

Manďák, J. &. (2019). Use of Logistic Regression for Understanding and Prediction of Customer Churn in telecomunications. Statistika: *Statistics & Economy Journal*. doi:10084/138783

Nurtriana, A. R. (2024). *Churn prediction analysis of telecom customers usaing svm, random forest and logistic regression models using orange data mining tools*. E3S Web of Conferences. doi:10.1051/e3sconf/202450102012

Olle, G. D. (2014). A hybrid churn prediction model in mobile telecomunication industry. *International Journal of e-Education, e-Business, e-Management and e-Learning*. doi:10.7763/IJEEEE.2014.V4.302

Wagh, S. K. (2024). *Customer churn prediction in telecom sector using machine learning techniques*. Results in Control and Optimization. doi:10.1016/j.rico.2023.100342