Reporte Técnico

Proyecto Final – Concurso “DesafÍA la IA”

Predicción de la Rotación de Clientes (Churn) en una Empresa de Telecomunicaciones

Autor: Ing. Erik González Molina

Planteamiento del Problema

En el sector de las telecomunicaciones, la rotación de clientes —conocida como churn— representa una de las principales amenazas a la estabilidad financiera y operativa de las empresas. Captar nuevos clientes implica una inversión considerable en marketing y adquisición, mientras que mantener a los clientes actuales resulta mucho más rentable a largo plazo. En este contexto, la retención se convierte en una prioridad estratégica. No obstante, identificar de manera oportuna a los clientes con alta probabilidad de abandonar sigue siendo un desafío complejo debido a la diversidad de factores que influyen en esta decisión, como la calidad del servicio, las condiciones contractuales, el comportamiento de uso y la experiencia del cliente.

El problema que se busca abordar en este proyecto es cómo anticipar, con un alto grado de precisión, qué clientes podrían abandonar los servicios de una empresa de telecomunicaciones. Para ello, se propone utilizar técnicas de machine learning aplicadas sobre datos históricos que incluyen tanto características personales como información relacionada con el servicio contratado, patrones de pago y comportamiento de uso. La resolución de este problema permitiría a la organización optimizar sus estrategias de fidelización, focalizar sus campañas de retención y tomar decisiones informadas basadas en datos objetivos. Además, contribuiría a reducir los costos asociados a la pérdida de clientes, mejorando la eficiencia operativa y la rentabilidad general del negocio.

Desde el punto de vista técnico, este problema corresponde a una tarea de clasificación binaria, donde la variable objetivo es Churn, codificada como "Yes" (cliente que abandona) o "No" (cliente que permanece). La salida esperada del modelo es la probabilidad de churn para cada cliente o su respectiva clase predicha. Para evaluar el desempeño del modelo, se utilizarán métricas como el área bajo la curva ROC (AUC-ROC), la puntuación F1, la precisión (precision) y la sensibilidad (recall), las cuales permiten medir la capacidad del modelo para distinguir entre clientes propensos y no propensos al abandono.

Objetivos

Objetivo General

Desarrollar un modelo de aprendizaje automático que permita predecir la probabilidad de que un cliente abandone los servicios de una empresa de telecomunicaciones, a partir del análisis de sus características demográficas, patrones de uso, tipo de contrato y comportamiento de pago, con el fin de apoyar estrategias efectivas de retención.

Objetivos Específicos

Realizar un análisis exploratorio del conjunto de datos para detectar patrones relevantes, relaciones entre variables y comportamientos anómalos asociados al abandono de clientes.

Aplicar un proceso completo de preprocesamiento de datos que incluya limpieza de valores faltantes, transformación de variables categóricas, detección y tratamiento de valores atípicos, escalado de variables y codificación adecuada de la variable objetivo.

Diseñar, entrenar y comparar múltiples modelos de clasificación binaria, incluyendo regresión logística, árbol de decisión, random forest y XGBoost, con el fin de identificar el algoritmo con mejor desempeño.

Evaluar rigurosamente los modelos desarrollados utilizando métricas como AUC-ROC, F1-Score, precisión y recall, que permitan medir la calidad predictiva y la capacidad de generalización de cada enfoque.

Determinar las características más influyentes en la decisión de abandonar el servicio mediante técnicas de interpretabilidad y análisis de importancia de variables, con el propósito de extraer información valiosa para el área de negocio.

Presentar los hallazgos del estudio de forma clara y estructurada, proponiendo recomendaciones orientadas a fortalecer las acciones preventivas de churn, optimizar recursos y mejorar la experiencia del cliente. Metodología

El desarrollo del proyecto se llevó a cabo siguiendo un enfoque estructurado basado en buenas prácticas de ciencia de datos. En primer lugar, se realizó la importación de librerías necesarias para el manejo de datos, visualización, modelado predictivo, evaluación de modelos, optimización de hiperparámetros e interpretabilidad, lo cual permitió contar con un entorno robusto y flexible para el análisis. Posteriormente, se procedió con la carga de la configuración desde un archivo JSON, desde donde se extrajo la ruta del dataset y la variable objetivo (Churn). Esta configuración puede ser modificada dinámicamente a través de una interfaz gráfica. Una vez establecida la configuración, se cargaron y visualizaron los datos para una revisión inicial.

La etapa de exploración y limpieza de datos incluyó un resumen general del conjunto, analizando el número de registros, columnas, tipos de datos y valores

faltantes. Se calcularon estadísticas descriptivas como media, mediana, máximos, mínimos y desviación estándar, además de la distribución de frecuencias. Para asegurar la calidad de los datos, se reemplazaron valores nulos en variables categóricas por NaN y se eliminaron las filas con valores faltantes. Asimismo, se realizó una conversión automática de tipos de datos para transformar variables categóricas que pudieran ser interpretadas como numéricas o fechas.

En el análisis exploratorio, se examinó la distribución de la variable objetivo mediante gráficos circulares y se evaluó el balance de clases. Se eliminaron columnas que actuaban como identificadores únicos y que no aportaban información útil al modelo. Se clasificaron las variables en categóricas (tipo object o con pocos valores únicos) y numéricas (valores reales, excluyendo la variable objetivo y las binarias). Los valores atípicos se detectaron mediante el método del rango intercuartílico (IQR) y se procedió a su recorte. Se generó un resumen general de estos outliers. Para las variables categóricas, se utilizaron gráficos de barras y tablas de frecuencia para entender su distribución, mientras que las numéricas se analizaron a través de histogramas con curvas de densidad (KDE), boxplots y estadísticas detalladas.

La preparación de los datos comenzó con la transformación de la variable objetivo, donde los valores “Yes” y “No” se codificaron como 1 y 0, respectivamente. Se realizaron análisis bivariados y multivariados que incluyeron gráficos de barras para analizar la tasa de Churn por categoría y boxplots para examinar la distribución de variables numéricas según el estado de Churn. Se calculó una matriz de correlación utilizando el coeficiente de Pearson, la cual fue representada mediante un heatmap. Además, se identificaron pares de variables altamente correlacionadas para prevenir redundancia en el modelado. La codificación de variables categóricas se realizó aplicando LabelEncoder para aquellas binarias y OneHotEncoder (con eliminación de la primera categoría) para las variables multiclase.

Una vez preprocesados los datos, se realizó la división del conjunto en tres subconjuntos: entrenamiento (70%), validación (15%) y prueba (15%). Esta división se hizo de forma estratificada para mantener la proporción de clases en cada subconjunto, garantizando así una representación adecuada de la variable objetivo en todas las fases del modelado.

En cuanto al modelado, se entrenaron y evaluaron cuatro algoritmos de machine learning: regresión logística, árbol de decisión, random forest y XGBoost. La regresión logística se probó con y sin escalado de características (StandardScaler), y se evaluó mediante matriz de confusión, métricas de clasificación y curva AUC-ROC. Los modelos de árbol de decisión y random forest se entrenaron utilizando sus configuraciones base, aplicando posteriormente una evaluación basada en precisión, sensibilidad, F1 score y AUC-ROC. En el caso de random forest, se utilizaron 100 árboles como punto de partida. XGBoost se entrenó empleando la métrica de logloss y fue evaluado de igual manera. Los cuatro modelos fueron comparados inicialmente considerando sus métricas de desempeño con y sin escalado, para identificar posibles mejoras con normalización de datos.

Posteriormente, se procedió con el ajuste de hiperparámetros utilizando GridSearchCV para optimizar el rendimiento de cada modelo. En la regresión logística se evaluaron distintos valores de C, tipo de penalización y método de optimización, obteniéndose como mejor configuración: C=1, penalty='l2' y solver='lbfgs'. Para random forest, los parámetros optimizados incluyeron el número de árboles, profundidad máxima y condiciones de división, resultando en una configuración ideal de 200 árboles, max_depth=10, min_samples_split=5 y min_samples_leaf=1. XGBoost se optimizó variando el número de árboles, profundidad, tasa de aprendizaje y proporción de muestreo, alcanzando su mejor rendimiento con 200 árboles, max_depth=5, learning_rate=0.1 y subsample=1.0. En cuanto al árbol de decisión, los mejores parámetros fueron max_depth=10, min_samples_split=5, min_samples_leaf=2 y el criterio de impureza gini.

Con los modelos ajustados, se realizó una evaluación final utilizando las métricas previamente mencionadas. Los resultados fueron almacenados en un archivo .csv para futuras consultas y se seleccionó como modelo final aquel con el mejor desempeño en términos del área bajo la curva ROC (AUC-ROC).

Finalmente, se aplicó validación cruzada con cinco particiones (CV=5) a los modelos ajustados, incluyendo regresión logística, árbol de decisión, random forest y XGBoost. Este procedimiento permitió verificar la estabilidad y robustez de los modelos frente a diferentes subconjuntos de datos, asegurando que el rendimiento observado no dependiera exclusivamente de una sola partición.

Resultados

Tras la implementación del pipeline de modelado, se evaluaron diversos algoritmos de clasificación para predecir el abandono de clientes (churn) en una empresa de telecomunicaciones. En la fase inicial, se entrenaron y compararon modelos base tanto con escalado como sin él, destacando la Regresión Logística como el modelo con mejor rendimiento general. Específicamente, la Regresión Logística con escalado obtuvo un AUC-ROC de 0.8533, precisión de 0.6939, recall de 0.6050 y un F1 Score de 0.6464, superando en esta etapa a modelos más complejos como XGBoost y Random Forest. Esta primera evaluación sugiere que incluso un modelo lineal puede capturar relaciones significativas en los datos, proporcionando un equilibrio adecuado entre simplicidad e interpretabilidad.

Posteriormente, se realizó un ajuste de hiperparámetros mediante GridSearchCV para cada uno de los modelos. Nuevamente, la Regresión Logística ajustada mostró métricas sólidas, alcanzando un AUC-ROC de 0.8530 y un F1 Score de 0.6424. Sin embargo, se observó una notable mejora en el desempeño del modelo XGBoost tras el ajuste, logrando un AUC-ROC de 0.8516 y un F1 Score de 0.5825, lo que evidenció su capacidad de adaptarse mejor a los patrones complejos del dataset.

La evaluación final se llevó a cabo mediante validación cruzada (CV=5), con el objetivo de medir la estabilidad y robustez de los modelos ajustados. En esta etapa, XGBoost obtuvo el mayor AUC-ROC con un valor de 0.8494, además de un F1 Score de 0.5933, superando ligeramente a la Regresión Logística, que alcanzó un

AUC-ROC de 0.8482 y un F1 Score de 0.5953. Estas métricas reflejan un rendimiento muy similar entre ambos modelos, pero marcan una tendencia positiva para XGBoost en términos de escalabilidad y adaptabilidad.

Conclusiones y Recomendaciones

En conclusión, si bien la Regresión Logística ajustada se mantiene como el modelo con mayor balance entre precisión, recall y F1 Score, el modelo XGBoost demostró un mejor comportamiento progresivo a lo largo de cada ajuste. Esto sugiere que, en escenarios reales con mayor volumen y variabilidad de datos, XGBoost podría ofrecer una mayor capacidad de generalización y adaptación a distintos entornos. Por tanto, su uso puede considerarse estratégico cuando se busca combinar buen desempeño predictivo con flexibilidad operativa en ambientes de producción más exigentes.
La implementación de modelos de machine learning para predecir el abandono de clientes representa un avance estratégico significativo para las empresas del sector telecomunicaciones. En un entorno altamente competitivo y dinámico, la capacidad de anticiparse al comportamiento del cliente se convierte en una ventaja diferenciadora clave. Este proyecto demostró que las nuevas tecnologías basadas en inteligencia artificial, como los modelos de clasificación aplicados (Regresión Logística, Random Forest, XGBoost y Árbol de Decisión), son herramientas eficaces para afrontar este desafío. En particular, el modelo XGBoost, por su capacidad de adaptación a diferentes entornos de datos, se posiciona como una solución robusta para contextos empresariales reales.

El impacto directo de estos modelos sobre la organización radica en su capacidad de transformar datos históricos en acciones concretas. Gracias a las predicciones obtenidas, el área de marketing y atención al cliente puede identificar con anticipación a los clientes con mayor riesgo de abandono y actuar de manera proactiva. Entre las recomendaciones prácticas derivadas de los resultados, se sugiere el diseño de campañas de retención personalizadas que combinen incentivos económicos (descuentos, mejoras en el servicio) con intervenciones dirigidas a mejorar la experiencia del usuario (seguimiento, soporte dedicado). Además, la empresa podría implementar un sistema de alertas automatizadas basado en los puntajes de churn, lo cual facilitaría una respuesta más oportuna y eficaz.

Finalmente, se recomienda integrar este modelo predictivo dentro del flujo operativo habitual de la empresa, con actualizaciones periódicas del modelo y entrenamiento continuo con nuevos datos. De esta forma, no solo se mantiene la precisión del sistema, sino que también se garantiza su alineación con los cambios en el comportamiento del cliente. Adoptar este enfoque basado en datos permitirá a la organización no solo reducir las tasas de abandono, sino también fortalecer la fidelización, mejorar la rentabilidad y tomar decisiones estratégicas más informadas en el futuro.