Análisis de Churn Aplicando la Metodología CRISP-DM

Profesor:
Facundo Oliva Cúneo

Alumnos:
Cristian Falco
Eduardo Figueroa
Sol Figueroa
Jorge Flores
Carlos Gimenez
Cinthya Yael Gomez
Monserrat Gutierrez
Walter Nieto
Jorgelina Tissera

ISPC
INSTITUTO SUPERIOR POLITÉCNICO CÓRDOBA

Índice

Análisis de Churn en una Empresa de Telecomunicaciones:

1. Comprensión del Negocio

En la actualidad, mantener la fidelidad de los clientes es un objetivo clave para las empresas. Este proyecto se centra en analizar un conjunto de datos para comprender mejor las dinámicas detrás de la rotación de clientes o "churn" y desarrollar estrategias efectivas para reducirlo y mejorar la retención de clientes.

Objetivos del Negocio:

El objetivo principal de este análisis es identificar los factores que contribuyen al abandono de clientes en una empresa de telecomunicaciones y desarrollar un modelo predictivo que permita anticipar este fenómeno. Comprender las razones detrás del churn permitirá a la empresa implementar estrategias efectivas para retener a los clientes y mejorar su satisfacción.

Criterios de Éxito de Negocio

Los criterios de éxito del proyecto de análisis de churn incluyen: reducir la tasa de churn en al menos un 20% durante el próximo año; retener al menos el 80% de los clientes en segmentos vulnerables; aumentar la satisfacción del cliente en un 15%; mejorar la retención de clientes en un 10% mediante una nueva estrategia de precios; y aumentar en un 25% la tasa de respuesta a las campañas de marketing dirigidas a clientes en riesgo de churn. La evaluación de estos criterios será llevada a cabo por el equipo de gestión, análisis de datos y marketing, mediante revisiones periódicas y el uso de métricas clave, con retroalimentación de los clientes y datos del mercado para validar el impacto de las iniciativas.

Inventario de recursos

El inventario de recursos incluye herramientas de minería de datos como Python, con bibliotecas de aprendizaje automático (Scikit-learn, TensorFlow) y visualización de datos (Matplotlib, Seaborn), además de un conjunto de datos históricos de clientes que contienen información demográfica y de uso de servicios. El proyecto necesita modelos predictivos precisos y soluciones técnicas para identificar y predecir el churn, con resultados claros y relevantes para la toma de decisiones estratégicas. Se presupone que los datos son representativos y suficientes, aunque el proyecto está limitado por restricciones de recursos humanos, financieros, de tiempo y consideraciones de privacidad y seguridad.

Objetivos de Minería de Datos

- Desarrollar un modelo predictivo de churn utilizando datos históricos de clientes, información demográfica y comportamientos de uso de servicios, con el fin de identificar los clientes propensos a abandonar el servicio en el futuro.
- Segmentar la base de clientes en grupos homogéneos basados en características demográficas, comportamientos de uso de servicios y patrones de churn históricos, con el fin de identificar los segmentos más propensos al churn y desarrollar estrategias específicas de retención para cada grupo

Criterios de Éxito de Minería de Datos

- Exactitud del Modelo: El modelo predictivo de churn debe tener una precisión mínima del 80% en la predicción de clientes que abandonarán el servicio.
- Interpretabilidad del Modelo: El modelo debe ser interpretable y capaz de proporcionar información sobre las características más influyentes en la predicción de churn.
- Funcionamiento y Complejidad: El modelo debe lograr un equilibrio entre un buen rendimiento predictivo y una complejidad razonable para facilitar su implementación y mantenimiento.

Plan del Proyecto

Objetivo del Proyecto:

Evaluar y mejorar los modelos de aprendizaje automático para la predicción del churn abandono de clientes y aplicar estrategias basadas en clustering para mejorar la retención de clientes y optimizar las campañas de marketing.

Duración del Proyecto:

- Fecha de inicio: 22 de abril
- Fecha de finalización: 24 de mayo

Equipo del Proyecto:

- Walter Nieto Project Manager (PM) / ML Engineer
- Sol Figueroa ML Engineer /Analista de BI
- Jorgelina Tissera Analista de Datos /Analista de B
- Eduardo Figueroa Analista de Datos
- Cristian Falco: Analista de Datos
- Carlos Giménez Analista de Datos
- Cinthya Gómez Analista de Datos/Calidad
- Monserrat Gutiérrez Analista de Datos/ Estadística

Entregables del Proyecto:

Inicio del Proyecto (Semana 1)

- Kick-off Meeting: Reunión inicial para definir objetivos, roles y responsabilidades.
- Plan de Proyecto Detallado: Documento que incluye el cronograma del proyecto, los recursos asignados y el plan de comunicación.
- Revisión del conjunto de Datos: Validación y limpieza inicial del conjunto de datos utilizado para la predicción del churn.

Análisis Exploratorio y Preprocesamiento (Semana 1 y 2)

- Análisis Exploratorio de Datos (EDA): Informe que detalla las características del conjunto de datos, identificando posibles problemas como datos faltantes y outliers.
- Preprocesamiento de Datos: Implementación de técnicas de normalización, imputación y codificación de datos, y generación de un informe de reprocesamiento.
- Desarrollo de Modelos de ML (Semana 2 y 3)
- Entrenamiento de Modelos: Desarrollar y entrenar modelos de ML como Extra Trees Classifier, LightGBM, XGBoost y Random Forest.
- Evaluación de Modelos: Informe con métricas de rendimiento (precisión, AUC, recall, F1-score, kappa y MCC) para cada modelo.
- Selección del Mejor Modelo: Justificación de la selección del modelo óptimo basado en las métricas obtenidas.

Implementación de Técnicas de Clustering (Semana 3)

- Aplicación de Clustering: Implementar técnicas de clustering para segmentar a los clientes.
- Informe de Segmentación: Descripción de los clústeres identificados y sus características.
- Validación y Optimización (Semana 3 y 4)
- Validación Cruzada: Realizar validación cruzada para asegurar la robustez de los modelos.
- Optimización de Hiperparámetros: Ajuste fino de los hiperparámetros de los modelos seleccionados.
- Despliegue y Recomendaciones (Semana 4)
- Implementación del Modelo en Producción: Despliegue del modelo de ML y del sistema de clustering en un entorno de producción.
- Estrategias de Retención y Marketing: Informe con recomendaciones de estrategias basadas en los resultados de la segmentación para mejorar la retención y optimizar campañas de marketing.

Cierre del Proyecto (Final de Semana 4)

- Informe Final del Proyecto: Documento que resume los hallazgos, el rendimiento de los modelos, las recomendaciones y los pasos siguientes.
- Reunión de Cierre: Presentación de los resultados del proyecto a las partes interesadas y discusión de los siguientes pasos.

2. Comprensión de los Datos

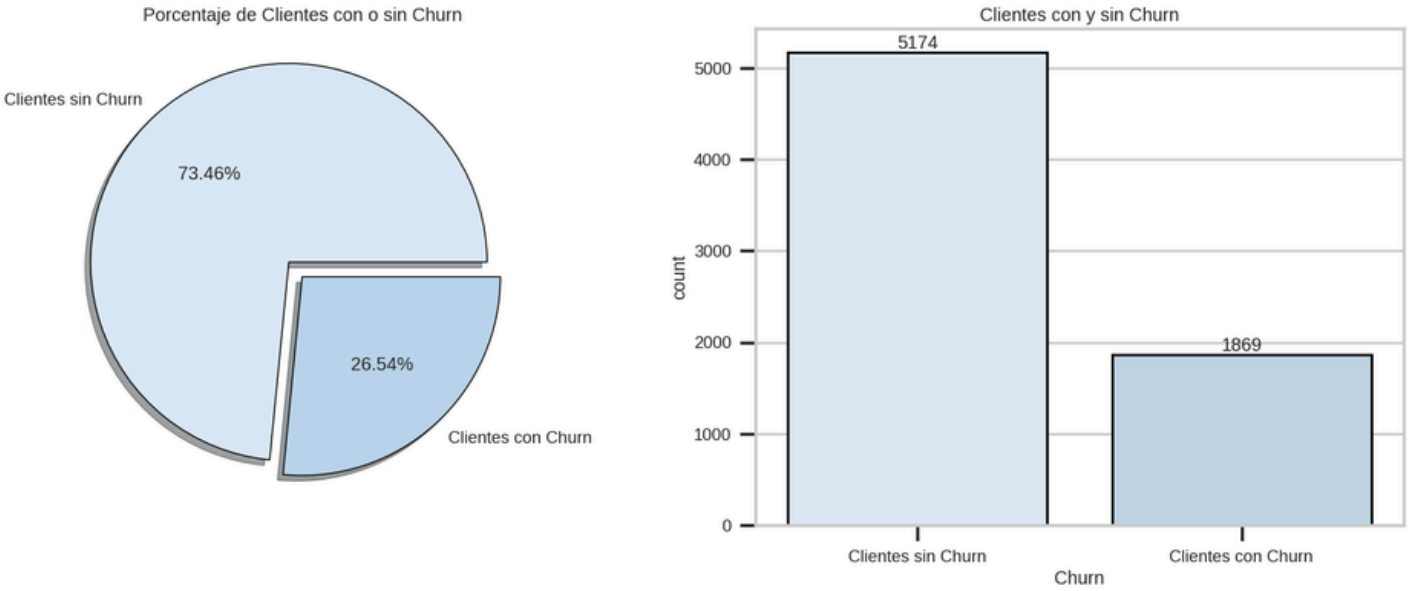
Descripción del Conjunto de Datos

El conjunto de datos utilizado contiene información de clientes, incluyendo variables demográficas, servicios contratados, facturación, y la etiqueta de Churn que indica si el cliente abandona o no la empresa.
El conjunto de datos consta de 7043 observaciones sin valores nulos, con las siguientes variables principales:

- customerID: Identificación única del cliente.
- gender: Género del cliente.
- SeniorCitizen: Indicador de si el cliente es una persona mayor.
- Partner: Indicador de si el cliente tiene pareja.
- Dependents: Indicador de si el cliente tiene dependientes.
- tenure: Antigüedad del cliente en meses.
- PhoneService: Indicador de si el cliente tiene servicio telefónico.
- MultipleLines: Indicador de si el cliente tiene múltiples líneas telefónicas.
- InternetService: Tipo de servicio de internet del cliente.
- OnlineSecurity: Indicador de si el cliente tiene servicio de seguridad en línea.
- OnlineBackup: Indicador de si el cliente tiene servicio de respaldo en línea.
- DeviceProtection: Indicador de si el cliente tiene protección de dispositivos.
- TechSupport: Indicador de si el cliente tiene soporte técnico.
- StreamingTV: Indicador de si el cliente tiene servicio de streaming de TV.
- StreamingMovies: Indicador de si el cliente tiene servicio de streaming de películas.
- Contract: Tipo de contrato del cliente.
- PaperlessBilling: Indicador de si el cliente tiene facturación electrónica.
- PaymentMethod: Método de pago del cliente.
- MonthlyCharges: Cargos mensuales.
- TotalCharges: Cargos totales.
- numAdminTickets: Número de tickets administrativos.
- numTechTickets: Número de tickets técnicos.
- Churn: Indicador de si el cliente abandonó el servicio.

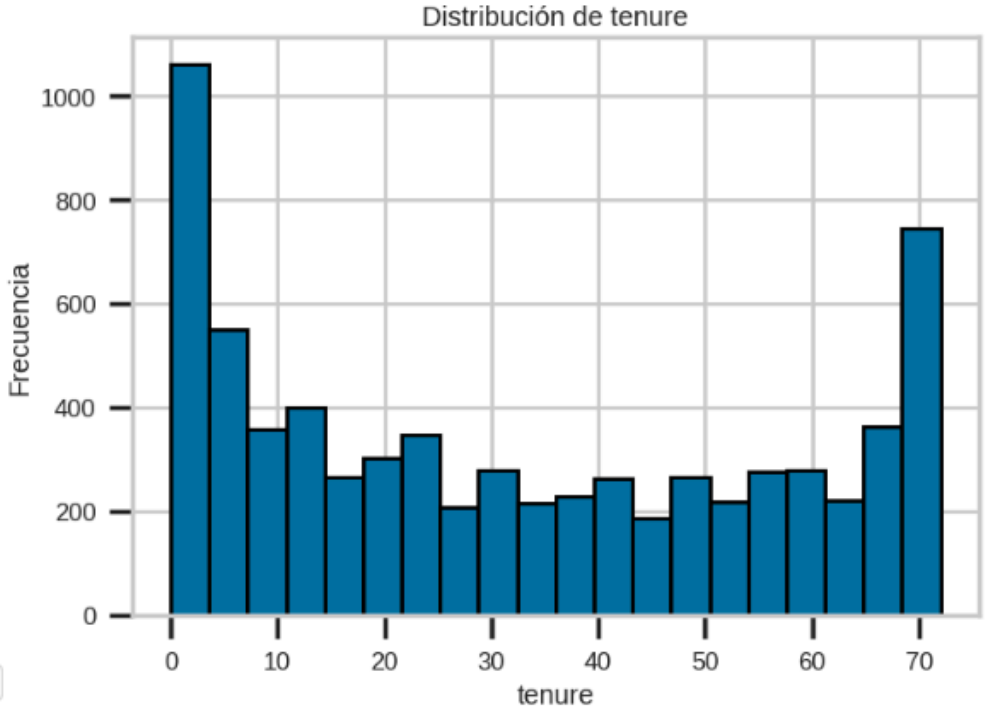
Se realizó un Análisis Exploratorio de Datos (EDA) para comprender las características de los datos, identificar valores faltantes, desequilibrio de clases, y las variables más relevantes para predecir el churn.

Gráfico 1. Porcentaje de clientes que abandonaron o no el servicio



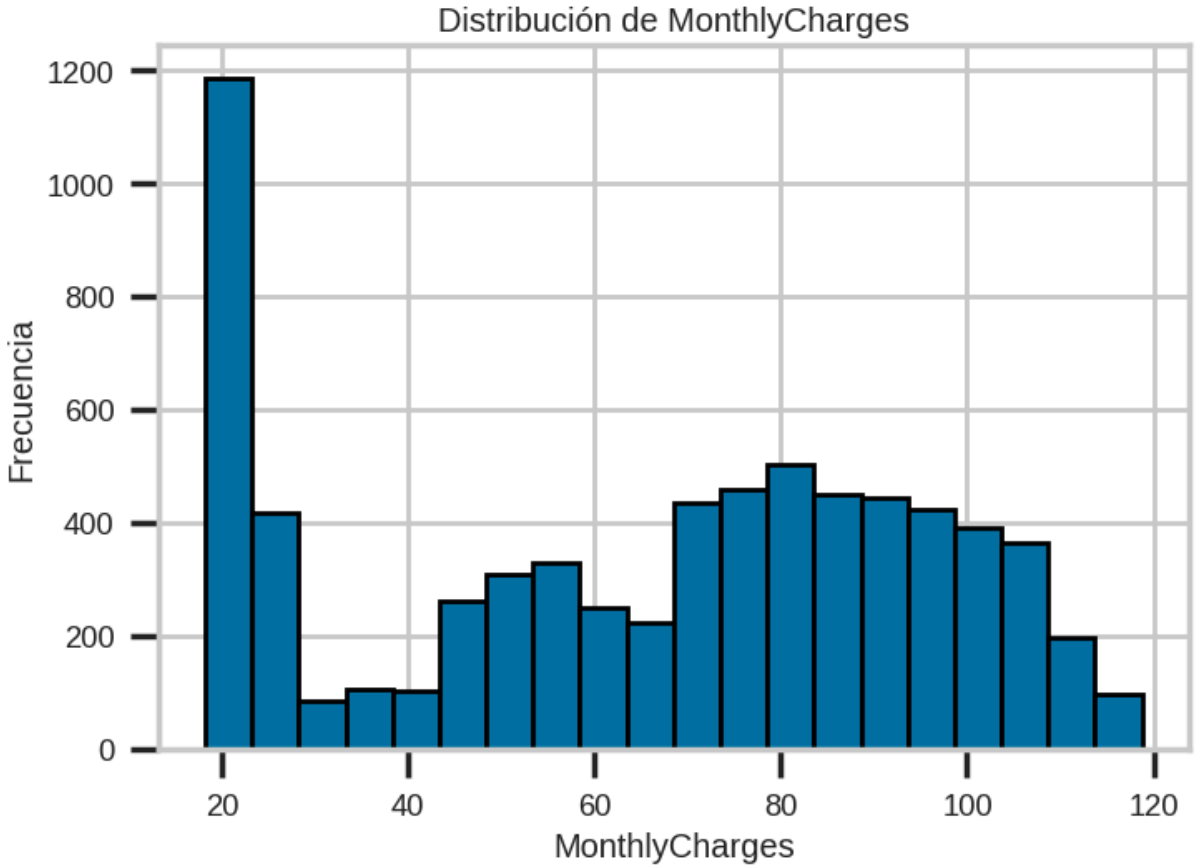
El gráfico circular muestra la distribución porcentual de clientes que han abandonado el servicio (churn) frente a aquellos que se han mantenido (sin churn). La mayoría de los clientes, el 73.46%, permanecen con la empresa, mientras que el 26.54% han abandonado. Sugiere que aproximadamente un cuarto de la base de clientes se pierde, lo cual es significativo y podría impactar negativamente en los ingresos y la estabilidad del negocio.

Gráfico 2. Distribución de la antigüedad de los clientes



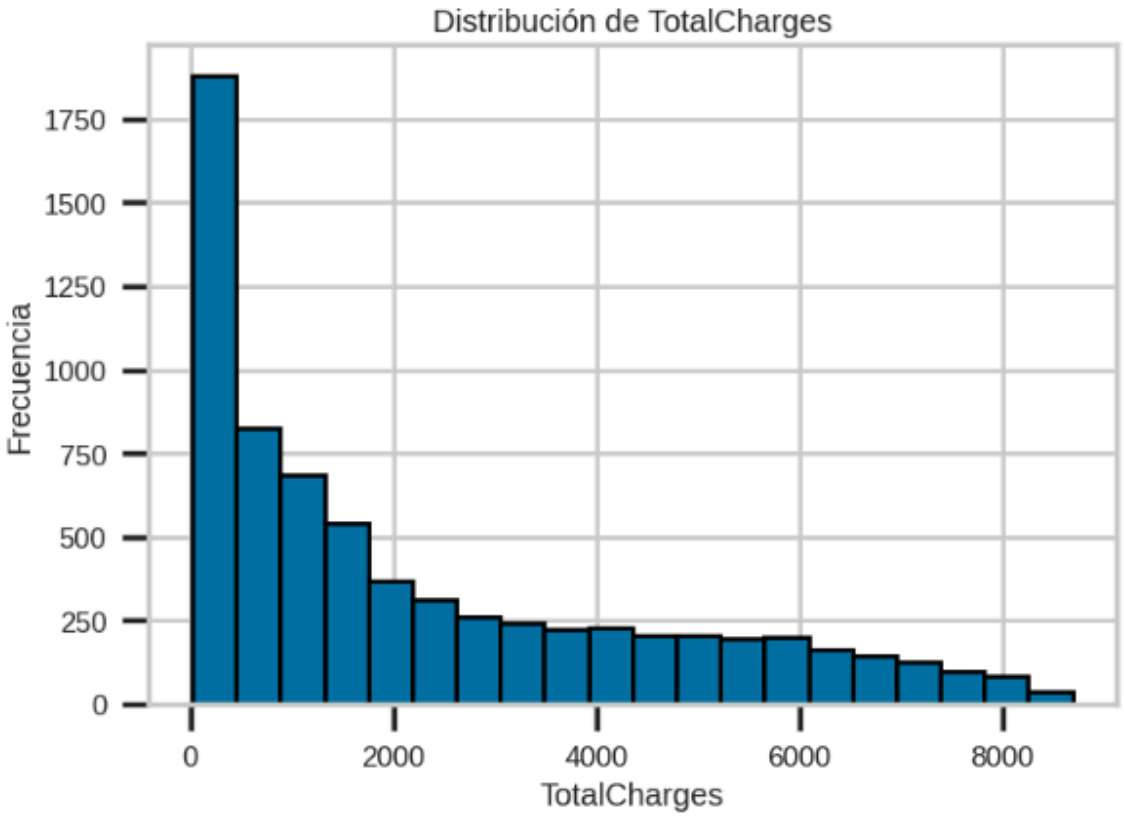
El histograma muestra la distribución de la variable 'tenure', que representa la duración de la suscripción de los clientes en meses. La asimetría de 0.24 indica que la distribución de la 'tenure' es ligeramente sesgada hacia la derecha. Implica que hay una mayor concentración de clientes con una duración de suscripción más corta, pero también hay un número considerable de clientes con una duración más larga.

Gráfico 3. Distribución de la variable cargos mensuales



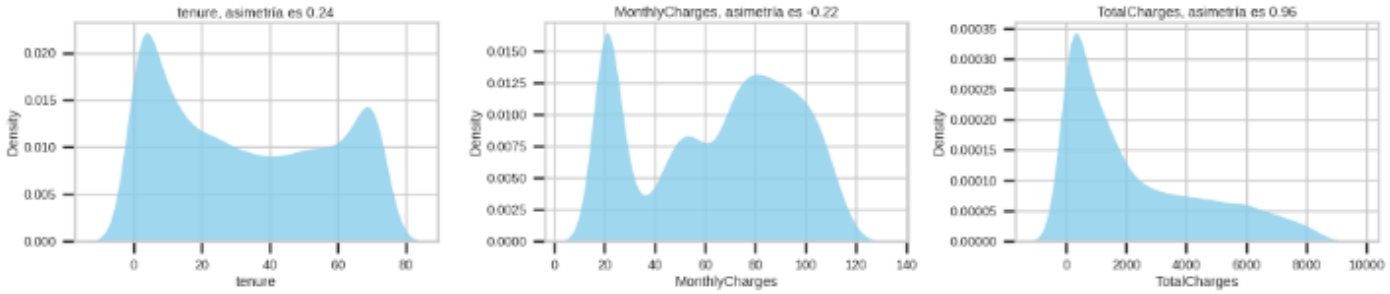
El histograma de la distribución de la variable 'MonthlyCharges', que representa los cargos mensuales que pagan los clientes. La asimetría de -0.22 indica que la distribución de 'MonthlyCharges' está ligeramente sesgada hacia la izquierda. Indica que hay una mayor concentración de clientes que pagan cargos mensuales más altos.

Gráfico 4. Distribución de la variable cargos totales



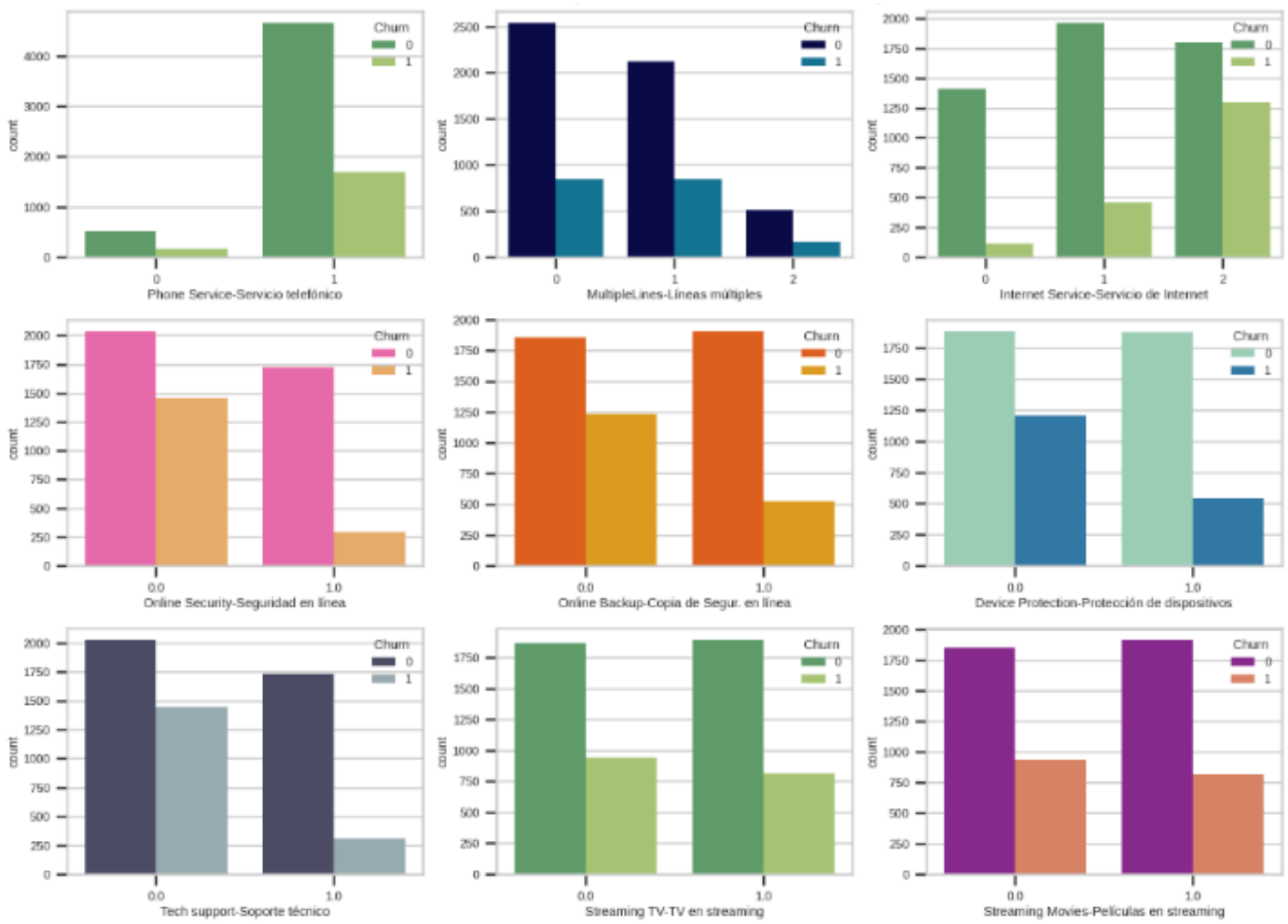
El histograma de la distribución de la variable 'TotalCharges', que representa el total acumulado de cargos pagados por los clientes. La asimetría de 0.96 indica que la distribución de 'TotalCharges' está sesgada hacia la derecha. Significa que hay una mayor concentración de clientes con cargos totales más bajos, mientras que una menor cantidad de clientes ha pagado cargos totales significativamente más altos.

Gráfico 5. Distribución de para las características.



Se observa que las variables no están distribuidas normalmente. Antigüedad y los cargos mensuales se asemejan a una distribución bimodal, mientras que el total de cargos está sesgada a la derecha.

Gráfico 6. Gráficos de conteo para características categóricas



Los gráficos de conteo para características categóricas muestran la distribución de clientes con y sin churn en relación con diversas características de los servicios ofrecidos por la empresa de telecomunicaciones.
Los gráficos revelan que la mayoría de los clientes tienen servicio telefónico y, entre ellos, hay una mayor proporción de clientes sin churn. Los clientes con múltiples líneas presentan una proporción similar de churn, aunque menos clientes con churn tienen múltiples líneas. La ausencia de servicio de Internet está asociada con una menor tasa de churn, mientras que aquellos con servicio de Internet, especialmente fibra óptica, muestran una mayor tasa de churn. Los servicios de seguridad en línea, copia de seguridad en línea, protección de dispositivos y soporte técnico están asociados con una menor tasa de churn. En contraste, los servicios de entretenimiento en streaming, como TV y películas, no muestran una relación significativa con la tasa de churn.

Verificación de la calidad de los datos

Se realizó un preprocesamiento exhaustivo de los datos para eliminar valores perdidos, verificar valores nulos, realizar imputaciones de valores faltantes, codificar variables categóricas y escalar las variables numéricas. Se puede afirmar que estos son completos. Los datos cubren los casos requeridos para la obtención de los resultados necesarios para cumplir los objetivos del proyecto.

3. Preparación de los Datos

El conjunto de datos utilizado para el análisis del churn comprende una variedad de variables relacionadas con los clientes de la empresa de telecomunicaciones con información detallada sobre las interacciones de los clientes con la empresa, sus características demográficas y su historial de servicios

Selección de los datos

Los campos seleccionados para el análisis son los siguientes:

- Variables relacionadas con el servicio: Se incluirán variables como PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling y PaymentMethod, ya que estas características están directamente relacionadas con los servicios que utiliza el cliente y pueden influir en su decisión de churn.

- Datos demográficos: Se incluirán variables como gender, SeniorCitizen, Partner y Dependents, ya que estas características pueden proporcionar información sobre el perfil del cliente y su propensión a churn.

- Historial de interacción del cliente: Se incluirán variables como tenure, MonthlyCharges, TotalCharges, numAdminTickets y numTechTickets, ya que estas características capturan la actividad y la historia del cliente con la empresa, lo que puede ser indicativo de su probabilidad de churn.

La selección de datos garantizará que el modelo se entrene y evalúe utilizando las características más relevantes y de mayor calidad, para mejorar la capacidad para predecir con precisión el churn de los clientes de la empresa de telecomunicaciones.
Las variables excluidas no proporcionan información relevante para predecir el churn y su inclusión no aportaría valor al modelo.

Limpieza de los datos

Se llevó a cabo un preprocesamiento de los datos, incluyendo la eliminación de valores perdidos, codificación de variables categóricas y escalado de variables numéricas.
Debido al desequilibrio de clases (mayor cantidad de clientes que no abandonaron), se aplicaron técnicas de sobremuestreo como SMOTE, ADASYN y SMOTEENN para balancear las clases antes del entrenamiento de los modelos de clasificación.
Para el análisis de clustering, se eliminó la variable Churn y se imputaron los valores faltantes en la variable TotalCharges con cero.

Construir datos

Se identificó un desbalance entre las clases de churn (abandono) y no churn (retención), lo que llevó a tomar medidas específicas. Se implementaron técnicas de sobremuestreo, como SMOTE (Synthetic Minority Oversampling Technique) o ADASYN (Adaptive Synthetic Sampling Approach), para balancear las clases antes del entrenamiento de los modelos.

Análisis de Correlaciones:

Tenure y TotalCharges: Hay una correlación positiva fuerte de 0.82. Significa que a medida que aumenta el tiempo que un cliente ha estado con la empresa (tenure), también tiende a aumentar el total de cargos acumulados por ese cliente (TotalCharges).
MonthlyCharges y TotalCharges: Existe una correlación positiva fuerte de 0.65. Esto indica que a medida que aumentan los cargos mensuales (MonthlyCharges) de un cliente, también aumenta el total de cargos acumulados (TotalCharges).

Formatear datos

Se creó un nuevo conjunto de datos denominado "ros_dataset" que integra las características aumentadas obtenidas después de aplicar la técnica de sobremuestreo SMOTE con su respectiva codificación de la variable objetivo 'Churn'. Esta combinación de datos representa una etapa esencial en la preparación de los datos, donde se formatean las características sintácticamente para adaptarse a los requisitos de las herramientas de modelado. Cabe destacar que estas transformaciones no alteraron el significado de los datos originales, pero fueron necesarias para asegurar la compatibilidad con las técnicas de modelado aplicadas posteriormente.

4. Modelado

Selección de la técnica de modelado

Clasificación: para abordar el problema de predicción de churn, se seleccionaron varias técnicas de modelado, incluyendo Extra Trees Classifier, Light Gradient Boosting Machine (LGBM), XGBoost y Random Forest Classifier. Estas técnicas fueron elegidas debido a su capacidad para manejar conjuntos de datos desbalanceados y su eficacia en la clasificación de casos minoritarios.

Clustering:Se aplicó el algoritmo de clustering K-Means para segmentar a los clientes en grupos homogéneos.

Generación de la prueba de diseño

El conjunto de datos se dividió en conjuntos de entrenamiento y prueba en una proporción de 80:20. Además, se utilizó una validación cruzada de 5 para evitar el sobreajuste y garantizar la generalización del modelo.

Construcción del modelo

Se procedió a construir los modelos utilizando las técnicas seleccionadas. Se ajustaron los parámetros de cada algoritmo de acuerdo con las mejores prácticas.

Evaluación del modelo

Los modelos fueron evaluados utilizando métricas como precisión, AUC, recall, F1-score, Kappa y MCC.
Se utilizaron técnicas como el método del codo (elbow method) y la métrica de silueta (silhouette score) para determinar el número óptimo de clústeres.
Después de aplicar técnicas de muestreo para abordar el desbalance de clases, se observó un rendimiento significativamente mejorado en términos de precisión y AUC.

Específicamente, los modelos XGBoost y LGBM, combinados con la técnica de muestreo SMOTEENN, demostraron ser los más efectivos, alcanzando una precisión superior al 96% y un AUC superior al 99%.

5. Evaluación

Evaluación de los resultados

Se realizó una revisión pormenorizada de los resultados obtenidos para verificar qué cumplen los objetivos iniciales del proyecto, centrándose en precisión, generalidad y otros aspectos relevantes.
Los modelos XGBoost y LGBM con SMOTEENN demostraron un rendimiento satisfactorio en términos de precisión y AUC, confirmando su capacidad para predecir el churn con alta precisión y su viabilidad para aplicaciones de producción. Las variables más importantes para predecir el churn fueron los cargos totales, los cargos mensuales y la antigüedad del cliente.
Inicialmente, se seleccionaron 3 clústeres basados en la interpretabilidad de los resultados y en métricas como el silhouette score y el índice de Calinski-Harabasz.
Los clústeres identificados presentaron diferencias significativas en características como edad, estado civil, dependientes, antigüedad, servicios utilizados, métodos de pago y cargos mensuales.

Proceso de revisión

Se analizaron matrices de confusión, validación cruzada, curvas de aprendizaje para evaluar el rendimiento de los modelos.

Decisión

Se decidió implementar el modelo de mejor rendimiento (LGBM con SMOTEENN) y realizar un seguimiento continuo para ajustar y mejorar el modelo según sea necesario.

6. Despliegue

Implementación

Se recomienda implementar el modelo de mejor rendimiento (LGBM con SMOTEENN) en un sistema de producción. Este modelo permitirá predecir el churn y tomar medidas preventivas para retener a los clientes.
Se sugiere realizar un análisis más profundo de las variables que influyen en el churn para identificar oportunidades de mejora en la atención al cliente y la retención de clientes.
Se recomienda utilizar los resultados del clustering para:

- Segmentar el mercado de manera efectiva y desarrollar estrategias de marketing personalizadas para cada grupo.
- Optimizar las campañas de marketing, enfocándose en los segmentos más propensos a utilizar ciertos servicios o realizar upselling/cross-selling.
- Mejorar la retención de clientes, implementando estrategias específicas para los segmentos con mayor riesgo de abandono.
- Desarrollar nuevos productos y servicios que satisfagan las necesidades de cada segmento identificado.

Se sugiere monitorear y ajustar continuamente las estrategias basadas en el clustering, integrando los resultados con otras herramientas de análisis de datos para obtener una visión completa del comportamiento del cliente.
Se destaca la importancia de abordar el desequilibrio de clases, la selección adecuada de modelos y técnicas de muestreo, y la integración de técnicas de clustering para obtener una comprensión más profunda de los clientes y desarrollar estrategias personalizadas.

Conclusiones y Recomendaciones

Conclusiones

Los modelos de aprendizaje automático son efectivos para predecir el churn. XGBoost y LGBM, combinados con SMOTEENN, son los modelos más adecuados para esta tarea.
Las variables relacionadas con los costos y la antigüedad del cliente son importantes para la predicción del churn.
El análisis de clústeres muestra que esta técnica es valiosa para segmentar el mercado de manera efectiva, optimizar campañas de marketing, mejorar la retención de clientes y desarrollar nuevos productos y servicios. Los clústeres identificados presentan diferencias significativas en características demográficas y de uso, permitiendo estrategias de marketing personalizadas. Por ejemplo, los clientes con mayor uso de servicios en línea y antigüedad en la empresa pueden ser objetivos ideales para campañas de upselling, mientras que los clientes con menor uso de servicios y más nuevos pueden beneficiarse de estrategias de adquisición y fidelización.

Recomendaciones

Se recomienda establecer el clustering como una práctica regular para mantener las estrategias de marketing actualizadas, utilizando sus resultados para crear perfiles detallados de clientes y diseñar campañas personalizadas.
Es esencial evaluar y ajustar continuamente estas campañas, integrando el clustering con otras herramientas de análisis y plataformas CRM para una visión integral del cliente. Además, se deben desarrollar estrategias específicas para retener a los clientes de alto riesgo de abandono y guiar el desarrollo de nuevos productos y servicios. Es fundamental capacitar al equipo en técnicas de clustering y realizar evaluaciones periódicas del impacto de estas estrategias para asegurar un retorno positivo sobre la inversión.

Anexo 1: Glosario

- CRISP-DM (Cross Industry Standard Process for Data Mining): Metodología estándar para la minería de datos que incluye seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.
- Churn: Abandono de clientes de un servicio o producto de una empresa. En este contexto, se refiere a los clientes que dejan de usar los servicios de la empresa de telecomunicaciones.
- SMOTE (Synthetic Minority Over-sampling Technique): Técnica de sobremuestreo utilizada para balancear clases desbalanceadas en un conjunto de datos generando nuevas muestras sintéticas.
- XGBoost: Algoritmo de aprendizaje automático basado en árboles de decisión que es altamente efectivo para problemas de clasificación y regresión.
- K-Means: Algoritmo de clustering que agrupa datos en un número determinado de clusters basados en características similares.
- AUC (Area Under the Curve): Métrica utilizada para evaluar la precisión de un modelo de clasificación. Específicamente, mide el área bajo la curva ROC (Receiver Operating Characteristic).
- LightGBM: Algoritmo de aprendizaje automático eficiente basado en árboles de decisión, optimizado para mayor velocidad y menor uso de memoria.
- Random Forest: Algoritmo de aprendizaje automático que crea un bosque de árboles de decisión para mejorar la precisión y evitar el sobreajuste.
- ADASYN (Adaptive Synthetic Sampling): Técnica de sobremuestreo que crea nuevas muestras sintéticas para la clase minoritaria de manera adaptativa, centrándose en las muestras que son difíciles de clasificar.
- Tenure: Tiempo que un cliente ha estado con la empresa.
- TotalCharges: Total de cargos acumulados por un cliente.
- MonthlyCharges: Cargos mensuales de un cliente.
- Recall: Métrica que mide la capacidad de un modelo para identificar todos los ejemplos positivos en un conjunto de datos.
- F1-Score: Métrica que combina la precisión y el recall en un solo valor, especialmente útil en casos de clases desbalanceadas.
- Kappa: Métrica que mide la precisión de un modelo teniendo en cuenta la posibilidad de que los resultados sean debidos al azar.
- MCC (Matthew's Correlation Coefficient): Métrica que evalúa la calidad de las clasificaciones binarias, teniendo en cuenta tanto los verdaderos como los falsos positivos y negativos.
- Elbow Method: Método utilizado para determinar el número óptimo de clusters en un algoritmo de clustering.
- Silhouette Score: Métrica utilizada para evaluar la calidad de los clusters formados por un algoritmo de clustering.
- Calinski-Harabasz Index: Métrica utilizada para evaluar la dispersión de los clusters formados por un algoritmo de clustering.
- Extra Trees Classifier: Algoritmo de aprendizaje automático basado en árboles de decisión, que es una variante del Random Forest, utilizando un conjunto de árboles extra para mejorar la precisión del modelo.

Anexo 2: Repositorio

Informe técnico
Análisis exploratorio de datos
Análisis y predicción
Plan de proyecto - Gantt