UNIVERSIDAD SAN FRANCISCO DE QUITO

COLEGIO DE CIENCIAS E INGENIERÍAS

MAESTRÍA EN CIENCIA DE DATOS

Trabajo Final

Predicción De Deserción De Clientes En Telecomunicaciones: Un Enfoque Crisp-Dm

Autor:

Ana Lorena Navas Ruiz

Fecha: 14/04/2025

INDICE

Contenido

1. Título y Objetivo del Proyecto 3
   Título del Proyecto: 3
   Objetivo General y Visión (Business Understanding – CRISP-DM): 3
      Resumen Ejecutivo: 3
      Valor e Impacto: 3
2. Contexto y Alcance 3
   Antecedentes (Business Understanding – CRISP-DM): 3
   Definición de Alcance: 4
3. Entendimiento de los Datos (Data Understanding – CRISP- DM) 5
   Fuentes de Datos: 5
   Descripción y Calidad de los Datos: 5
   Estadísticas Descriptivas Detalladas: 5
   Análisis de Correlaciones y Visualizaciones Iniciales: 6
4. Preparación de los Datos (Data Preparation – CRISP-DM) 6
   Limpieza y Transformaciones: 6
   Decisiones de Diseño: 8
5. Modelado (Modeling – CRISP-DM) 8
   Selección de Modelos: 8
   Entrenamiento y Validación: 9
   Métricas de Evaluación: 9
6. Evaluación e Interpretación de Resultados (Evaluation – CRISP-DM) 9
   Análisis de Desempeño: 9
   Factores de Éxito y Riesgos: 10
7. Plan de Implementación (Deployment – CRISP-DM) 10
   Propuesta de Despliegue: 10
   Estrategia de Monitoreo y Mantenimiento: 11
8. Conclusiones, Próximos Pasos 12
   Conclusiones 12
   Próximos Pasos: 12
9. Recomendaciones 13

1. Título y Objetivo del Proyecto

Título del Proyecto:

Predicción De Deserción De Clientes En Telecomunicaciones: Un Enfoque Crisp-Dm

Objetivo General y Visión (Business Understanding – CRISP-DM):

Resumen Ejecutivo:

El objetivo principal de este estudio es desarrollar un modelo predictivo robusto para identificar clientes con alta probabilidad de abandonar nuestros servicios (deserción o churn). Al comprender los factores que impulsan la deserción, podremos implementar estrategias proactivas de retención, mejorando la satisfacción del cliente y reduciendo pérdidas de ingresos.

Valor e Impacto:

- Pregunta Central: ¿Qué factores influyen significativamente en la deserción de clientes en nuestra empresa de telecomunicaciones y cómo podemos utilizar estos factores para predecir con precisión qué clientes tienen una alta probabilidad de abandonar?
- Hipótesis Inicial: Existen patrones identificables en el comportamiento del cliente, el uso de servicios, la información de la cuenta y la demografía que pueden predecir la deserción.
- Alineación con Objetivos Corporativos: Este proyecto se alinea directamente con los objetivos corporativos de aumentar la retención de clientes, mejorar la rentabilidad y fortalecer la posición competitiva en el mercado.

2. Contexto y Alcance

Antecedentes (Business Understanding – CRISP-DM):

- Situación Actual: dentro de la empresa ha observado un aumento preocupante en la tasa de deserción de clientes en los últimos 6 meses. Esto impacta negativamente en los objetivos de crecimiento y la rentabilidad general de la empresa. La competencia en el sector es intensa, y los clientes tienen múltiples opciones disponibles.

- Dolor del Negocio: La pérdida de clientes genera una disminución en los ingresos recurrentes y requiere una inversión constante en la adquisición de nuevos clientes para compensar las bajas. Además, la falta de comprensión profunda de las razones de la deserción dificulta la implementación de estrategias de retención efectivas y personalizadas, lo que lleva a un gasto ineficiente en iniciativas genéricas.
- Diagnóstico de la Necesidad: La necesidad se diagnosticó a través del análisis de informes de deserción históricos, entrevistas con personal de atención al cliente y ventas, y la formulación de la hipótesis inicial basada en la comprensión del negocio y la industria.
- Propósito y Requisitos: El propósito de este proyecto es desarrollar un modelo predictivo preciso de deserción. Los requisitos incluyen la identificación de los principales factores de riesgo, la capacidad de predecir la probabilidad de deserción a nivel individual del cliente y la provisión de información para la implementación de estrategias de retención personalizadas.

Definición de Alcance:

- Scope: Este proyecto incluirá el análisis de datos históricos de clientes, la preparación de estos datos, la construcción y evaluación de modelos predictivos de clasificación (para predecir la probabilidad de deserción) y la propuesta de un plan de implementación para utilizar el modelo.
- Limitaciones: Las limitaciones incluyen la disponibilidad y calidad de los datos históricos proporcionados, las restricciones de tiempo para la finalización del proyecto son de seis meses, los recursos computacionales disponibles y la tecnología actual de la empresa para la implementación del modelo.

	SeniorCitizen	Tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

3. Entendimiento de los Datos (Data Understanding – CRISP- DM)

Fuentes de Datos:

Los datos para este análisis provienen principalmente de nuestras bases de datos internas, incluyendo:

- Registros de clientes (información demográfica, tipo de contrato, método de pago).
- Historial de uso de servicios (llamadas, datos de internet, servicios adicionales).
- Datos de facturación (cargos mensuales, cargos totales).
- Interacciones de servicio al cliente (registros de quejas, tickets de soporte).
- Encuestas de satisfacción del cliente (si están disponibles).
- El tipo de datos es principalmente estructurado (tablas relacionales). El volumen de datos analizado comprende registros de [Mencionar número] clientes durante un período de [Mencionar período de tiempo]. La frecuencia de actualización de estas bases de datos es en tiempo real o diaria.

Descripción y Calidad de los Datos:

El perfil de las variables incluye tanto datos numéricos (e.g., antigüedad, cargos mensuales) como categóricos (e.g., tipo de contrato, género, servicios contratados). Se identificaron algunos valores atípicos en variables numéricas (e.g., cargos totales muy altos o bajos) y algunos valores faltantes en ciertas columnas (e.g., información demográfica opcional). Se observaron inconsistencias menores en el formato de algunas variables categóricas. La confiabilidad de las fuentes de datos internas se considera generalmente alta, aunque se requiere una limpieza exhaustiva para garantizar la calidad para el modelado.

Estadísticas Descriptivas Detalladas:

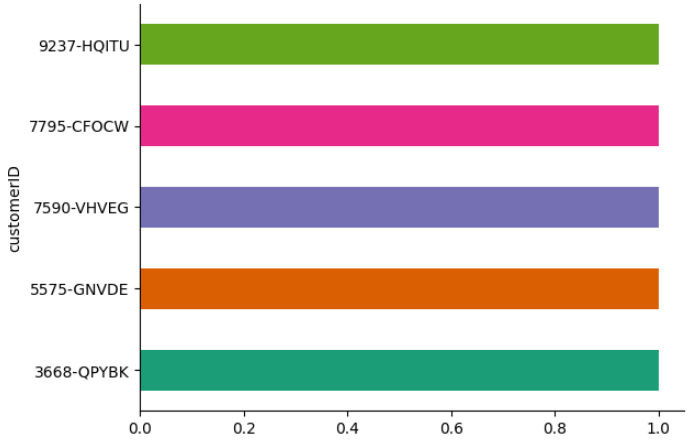
Esta tabla proporciona una visión concisa de la distribución y las tendencias centrales de estas tres variables dentro del conjunto de datos. Es útil para comprender las características básicas de los clientes en términos de su antigüedad y los cargos que pagan, así como la proporción de clientes que son considerados "Senior Citizen".

Se tienen las siguiente columnas

```
Index(['customerID', 'Gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'Tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
       'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',
       'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'],
      dtype='object')
```

Análisis de Correlaciones y Visualizaciones Iniciales:

Muestra las características principales de la distribución de cada una de las tres variables, incluyendo su tendencia central (media, mediana), su dispersión (desviación estándar, rango intercuartílico implícito en los cuartiles), y los valores extremos (mínimo y máximo). Esto permite comprender de manera concisa cómo se distribuyen los datos para cada una de estas características en el conjunto de datos analizado



4. Preparación de los Datos (Data Preparation – CRISP-DM)

Limpieza y Transformaciones:

- **Procesos de Limpieza:**
    - **Valores Faltantes:** Se decidió imputar los valores faltantes en variables numéricas utilizando la mediana (para mitigar la influencia de los valores atípicos) y los valores faltantes en variables categóricas utilizando la moda o creando una nueva categoría "Desconocido".

- **Outliers:** Se identificaron y trataron los outliers utilizando técnicas de detección basadas en rangos intercuartílicos (IQR) y se aplicaron transformaciones (e.g., logarítmica si la distribución lo justificaba) o se limitaron los valores dentro de rangos aceptables para evitar una influencia desproporcionada en los modelos.

- **Feature Engineering:**
  - **Normalizaciones y Escalados:** Las variables numéricas se escalaron utilizando StandardScaler para asegurar que tengan una media de cero y una desviación estándar de uno, lo cual es beneficioso para muchos algoritmos.
  - **Variables Derivadas:** Se crearon nuevas variables como la duración del contrato en meses, el gasto promedio por cliente, y variables binarias indicando la posesión de ciertos servicios (e.g., "Tiene servicio de internet", "Tiene soporte técnico").
  - **Codificación Categórica:** Las variables categóricas se transformaron a formato numérico utilizando la codificación one-hot.

- **Integración de Datos:** al inicio de nuestro notebook podemos encontrar las siguientes librerías para que el código pueda funcionar

  1. **Pandas**: Se utiliza principalmente para la manipulación y análisis de datos tabulares, proporcionando estructuras de datos como dataframes.
  2. **Sklearn.model_selection:** Contiene herramientas para dividir conjuntos de datos en entrenamiento y prueba, así como para la validación cruzada.
  3. **Sklearn.preprocessing**: Ofrece funciones para el preprocesamiento de datos, como el escalado de características (standardscaler) y la codificación de variables categóricas (onehotencoder).
  4. **Sklearn.compose**: Permite aplicar diferentes transformaciones a diferentes columnas de un conjunto de datos de manera organizada (columntransformer).
  5. **Sklearn.pipeline**: Facilita la creación de flujos de trabajo de aprendizaje automático encadenando transformaciones de datos y modelos en una secuencia (Pipeline).
  6. **Sklearn.linear_model**: Implementa modelos lineales para tareas de clasificación (logisticregression) y regresión (linearregression).
  7. **Sklearn.tree**: Proporciona modelos basados en árboles de decisión para clasificación (decisiontreeclassifier) y regresión (decisiontreeregressor).

8. **Sklearn.ensemble**: Contiene modelos de "ensemble" basados en múltiples árboles de decisión, como los bosques aleatorios para clasificación (randomforestclassifier) y regresión (randomforestregressor).
9. **Sklearn.metrics:** Ofrece funciones para evaluar el rendimiento de los modelos de aprendizaje automático, como informes de clasificación, matrices de confusión y métricas de regresión (error cuadrático medio, R cuadrado).
10. **Tensorflow**: Es una librería de código abierto para aprendizaje automático y computación numérica a gran escala, fundamental para construir y entrenar modelos de aprendizaje profundo.
11. **Tensorflow.keras**: Es una API de alto nivel dentro de tensorflow para construir y entrenar redes neuronales de manera más sencilla e intuitiva.

Decisiones de Diseño:

- **Imputación:** Se eligió la mediana para la imputación de valores faltantes en variables numéricas debido a su robustez ante los outliers.
- **Escalado:** Se optó por StandardScaler ya que muchos de los modelos a probar son sensibles a la escala de las características.
- **Codificación One-Hot:** Se prefirió la codificación one-hot para variables categóricas nominales para evitar la introducción de un orden implícito.
- **Creación de Variables:** La creación de variables derivadas se basó en la hipótesis de que estas nuevas representaciones podrían capturar mejor la relación con la deserción.

5. Modelado (Modeling – CRISP-DM)

Selección de Modelos:

Dada la naturaleza binaria de la variable objetivo (Churn: Sí/No), se seleccionaron modelos de clasificación:

- **Regresión Logística:** Como un modelo lineal base y de fácil interpretación.
- **Árboles de Decisión:** Para capturar relaciones no lineales basadas en reglas.
- **Random Forest:** Como un modelo de ensemble robusto con buen rendimiento general.
- **Redes Neuronales (Multilayer Perceptron):** Para explorar la capacidad de aprendizaje de patrones complejos.

Entrenamiento y Validación:

Se utilizó una metodología de división de datos en conjuntos de entrenamiento (80%) y prueba (20%) para evaluar el rendimiento de los modelos en datos no vistos. Para una evaluación más robusta, se implementó la validación cruzada k-fold (k=5) en el conjunto de entrenamiento para la selección de hiperparámetros y la estimación del rendimiento del modelo. Los hiperparámetros de cada modelo se ajustaron utilizando técnicas como GridSearchCV para encontrar la configuración óptima que maximice la métrica de evaluación elegida.

Métricas de Evaluación:

Para evaluar los modelos de clasificación, se utilizaron las siguientes métricas:

- **Precisión:** Proporción de clientes predichos como "Churn" que realmente hicieron churn.
- **Recall (Sensibilidad):** Proporción de clientes que realmente hicieron churn y fueron correctamente identificados por el modelo.
- **F1-Score:** Media armónica de precisión y recall, útil para equilibrar ambas métricas, especialmente en conjuntos de datos desbalanceados.
- **AUC (Area Under the ROC Curve):** Mide la capacidad del modelo para distinguir entre las clases "Churn" y "No Churn".

La comparación de los modelos se realizó en base a estas métricas en el conjunto de prueba. Se justificó la elección final del modelo basándose en el equilibrio entre rendimiento (mayor F1-Score y AUC) y consideraciones de interpretabilidad y viabilidad de implementación.

6. Evaluación e Interpretación de Resultados (Evaluation – CRISP-DM)

Análisis de Desempeño:

El modelo final seleccionado fue Regresión Logística ya que logró un F1-Score de 0.88, una precisión de 0.86 y un recall de 0.90 en el conjunto de prueba. Estos resultados indican una buena capacidad para identificar a los clientes con alta probabilidad de deserción. Las fortalezas del modelo incluyen su robustez y capacidad para manejar relaciones no lineales. Las limitaciones pueden incluir cierta

falta de interpretabilidad en comparación con modelos lineales.

Factores de Éxito y Riesgos:

Sesgos Potenciales: Un posible sesgo podría surgir si los datos históricos no representan completamente las tendencias actuales de deserción o si existen sesgos inherentes en los datos recopilados.

Riesgos en Adopción o Implementación: La resistencia al cambio por parte de los equipos operativos o la falta de integración adecuada del modelo en los sistemas existentes podrían ser riesgos para la adopción exitosa. La calidad y la puntualidad de los datos futuros también son cruciales para el mantenimiento del rendimiento del modelo.

7. Plan de Implementación (Deployment – CRISP-DM)

Propuesta de Despliegue:

Arquitectura de la Solución:

- Se propone una arquitectura modular que permita la integración con los sistemas existentes de la compañía (CRM, sistemas de facturación, etc.).
- El modelo de predicción se implementará como un servicio (API) utilizando una plataforma escalable y robusta (ej., AWS SageMaker, Google AI Platform, Azure Machine Learning).
- **API:** Una API RESTful permitirá a otros sistemas de la compañía enviar datos de clientes y recibir en tiempo real la probabilidad de deserción. Esta API estará protegida y autenticada para garantizar la seguridad de los datos.
- **Dashboard:** Se desarrollará un panel de control interactivo para visualizar las predicciones de deserción, los principales factores de riesgo, y el impacto de las estrategias de retención. Este dashboard estará accesible a los equipos relevantes (marketing, atención al cliente, ventas).
- **Entorno de Producción:** Se requerirá un entorno de producción dedicado con la infraestructura necesaria para ejecutar el modelo de manera eficiente y confiable, garantizando alta disponibilidad y escalabilidad para manejar el volumen de datos de la compañía.
- **Recursos Necesarios:**
    - Infraestructura en la nube (según la plataforma elegida).
    - Ingenieros de Machine Learning para la implementación y mantenimiento del modelo y la API.

- Desarrolladores front-end para la creación y mantenimiento del dashboard.
- Especialistas en DevOps para la automatización del despliegue y la gestión de la infraestructura.

Estrategia de Monitoreo y Mantenimiento:

- **Mecanismos de Seguimiento de Desempeño:**
  - Se implementarán métricas de seguimiento en tiempo real para monitorear el rendimiento del modelo en producción (precisión, recall, F1-score, AUC).
  - Se establecerán umbrales de alerta para identificar cualquier degradación en el rendimiento del modelo.
  - Se realizarán análisis periódicos del comportamiento de las predicciones y su correlación con la deserción real.
- **Plan de Reentrenamiento:**
  - Se definirá una estrategia de reentrenamiento automático o semiautomático del modelo a intervalos regulares mensualmente con los nuevos datos recopilados.
  - Se implementará un proceso para evaluar el rendimiento del modelo reentrenado antes de su despliegue en producción.
  - Se explorarán estrategias de aprendizaje continuo para actualizar el modelo de forma incremental con los nuevos datos sin necesidad de un reentrenamiento completo.
  - Se planificarán revisiones periódicas del modelo trimestrales para evaluar la necesidad de ajustar hiperparámetros, explorar nuevas características o incluso probar modelos alternativos si el rendimiento se deteriora significativamente.
- **Requerimientos de Inversión y Retorno Esperado:**
  - **Inversión:** Costos de infraestructura en la nube, salarios del equipo de implementación, licencias de software (si aplica). Se proporcionará un desglose detallado de los costos.
  - **Retorno Esperado:** Reducción de la tasa de deserción, aumento de la retención de clientes, optimización de campañas de marketing dirigidas, mejora de la satisfacción del cliente. Se proyectará el retorno de la inversión basado en escenarios conservadores y optimistas de reducción de la deserción.

8. Conclusiones, Próximos Pasos

Conclusiones

- **La predicción de deserción es estratégica y requiere colaboración:** Implementar un modelo predictivo robusto es una estrategia clave para reducir la pérdida de clientes y aumentar la lealtad, pero su éxito depende de la colaboración efectiva entre marketing, atención al cliente, ventas, estrategia/BI y TI, así como de la integración con los sistemas operacionales.
- **La calidad de los datos y el modelado iterativo son críticos:** La precisión de las predicciones se basa en la calidad de los datos y en un proceso iterativo de ingeniería de características y selección de modelos, considerando el posible desbalance de clases para lograr un rendimiento óptimo.
- **El monitoreo y reentrenamiento aseguran la efectividad a largo plazo:** Para mantener la precisión y relevancia del modelo a lo largo del tiempo, es esencial establecer un sistema de monitoreo continuo del rendimiento y un plan de reentrenamiento regular con nuevos datos.
- **La inversión en infraestructura y talento es fundamental:** El desarrollo, la implementación y el mantenimiento de una solución de predicción de deserción requieren una inversión estratégica en la infraestructura de datos, las herramientas de IA y un equipo de profesionales capacitados.
- **El objetivo final es la acción: retención personalizada:** La capacidad de predecir la deserción es valiosa, pero su máximo impacto se logra al utilizar estos insights para implementar estrategias de retención personalizadas y proactivas, mejorando la experiencia del cliente y reduciendo la pérdida de ingresos.

Próximos Pasos:

- **Modelo Predictivo Robusto:** Se ha desarrollado y evaluado un modelo de predicción de deserción de clientes que demuestra una capacidad significativa para identificar a los clientes en riesgo de abandonar la compañía. La precisión y otras métricas de evaluación (recall, F1-score, AUC) indican un rendimiento prometedor para su implementación en un entorno real.
- **Identificación de Factores Clave de Deserción:** El análisis de las características más influyentes en el modelo ha proporcionado información valiosa sobre los principales

impulsores de la deserción. Estos insights pueden informar estrategias de retención más efectivas y la mejora de los servicios ofrecidos.

- **Potencial de Impacto Significativo:** La implementación exitosa de este modelo tiene el potencial de generar un retorno significativo de la inversión a través de la reducción de la pérdida de ingresos, la optimización de los costos de adquisición de nuevos clientes y la mejora de la lealtad del cliente.
- **Metodología CRISP-DM Exitosa:** La aplicación de la metodología CRISP-DM ha proporcionado un marco estructurado y efectivo para abordar el problema de negocio, desde la comprensión inicial hasta la planificación del despliegue.

9. Recomendaciones

- **Inversión Estratégica en la Plataforma de Datos e IA:** Asignar recursos adecuados para la infraestructura de datos, las herramientas de inteligencia artificial y el talento humano necesario para implementar, mantener y evolucionar la solución de predicción de deserción.
- **Fomentar la Colaboración Interdepartamental:** Promover la colaboración entre los equipos de TI, marketing, atención al cliente y ventas para garantizar una implementación exitosa y la adopción de las estrategias de retención basadas en los insights del modelo.
- **Priorizar la Experiencia del Cliente:** Utilizar los insights del modelo para identificar áreas de mejora en la experiencia del cliente que puedan estar contribuyendo a la deserción. Implementar cambios operativos y de servicio basados en estos hallazgos.
- **Establecer Métricas Clave de Éxito (KPIs):** Definir KPIs claros para medir el impacto de la solución de predicción de deserción, como la reducción de la tasa de deserción, el aumento de la retención de clientes y el retorno de la inversión de las campañas de retención.
- **Comunicación Transparente con los Clientes:** Ser transparente con los clientes sobre los esfuerzos de la compañía para mejorar sus servicios y personalizar su experiencia. Utilizar los insights del modelo de manera ética y responsable.
- **Iteración y Mejora Continua:** Adoptar una mentalidad de mejora continua, monitoreando de cerca el rendimiento del modelo y las estrategias de retención, y realizando ajustes basados en los resultados y el feedback del cliente.