UNIVERSIDAD TÉCNICA NACIONAL

SEDE SAN CARLOS

CARRERA

Ingeniería del Software

Minería de Datos

Laboratorio #4

Elaborado por

José Julián Vargas Cordero

Kendall Aaron Angulo Chaves

Keisy Nicole Angulo Chaves

Tabla de Contenido

Resumen de hallazgos

1. **Carga y comprensión inicial de los datos**

   El dataset analizado consta de 5000 filas y 18 columnas, lo que lo clasifica como un conjunto de datos relativamente amplio. A continuación, se describen las principales características. Los datos se cargaron correctamente, sin valores nulos, y los tipos de datos de las columnas son coherentes con sus valores. Además se revisaron las primeras 5 filas, confirmando la estructura y coherencia general de los datos.

2. **Limpieza y preparación de los datos**

   - Valores nulos: No se encontraron valores nulos en ninguna columna.
   - Valores duplicados: No se encontraron filas duplicadas.
   - Valores atípicos: Varias columnas presentaron valores atípicos, principalmente en variables numéricas relacionadas con el uso de los servicios (por ejemplo, 'totaldayminutes', 'numbervmailmessages', 'totalintlminutes', 'totalintlcalls'). Estos valores atípicos fueron tratados, ya sea ajustándolos a rangos válidos o transformándolos.
   - Conversión de tipos de datos: Se ajustaron los tipos de algunas columnas, como la conversión de variables categóricas (e.g., "churn", "internationalplan") a valores numéricos (0, 1).
   - Variables categóricas: Se codificaron adecuadamente las variables categóricas para su análisis posterior.
   - Tras la limpieza, el dataset no contiene valores nulos ni duplicados, y las variables atípicas fueron corregidas.
   - El dataset está listo para su análisis posterior o modelado, con las variables categóricas ya codificadas y los valores numéricos ajustados.

3. **Análisis univariado**

   <u>Estadísticas descriptivas:</u>

Las estadísticas descriptivas proporcionan una visión general de las variables numéricas:

- churn:
  Representa clientes que abandonan el servicio. La media es baja (0.14), indicando un bajo porcentaje de abandono. La asimetría (2.06) y curtosis (2.24) confirman que la mayoría de los valores están concentrados en el 0, con pocos casos en 1 (alto sesgo positivo).
- accountlength:
  El tiempo medio como cliente es de 100 meses, con un rango amplio (1-208). La asimetría cercana a 0 (0.07) y curtosis (-0.21) sugieren una distribución simétrica y ligeramente plana.
- internationalplan y voicemailplan:
  Son binarias (0 y 1). Las medias son 0.09 y 0.26 respectivamente, indicando que pocos clientes optan por estos servicios. La alta asimetría (2.77 y 1.07) refleja esta concentración en 0.
- numbervmailmessages:
  La mayoría tiene 0 mensajes de voz (mediana = 0), pero hay clientes con valores altos hasta 42.5. La asimetría (1.33) y curtosis (0.08) reflejan valores extremos en la cola derecha.
- Otras variables numéricas (totaldayminutes, totaleveminutes, etc.):
  Estas siguen distribuciones típicas de uso. Por ejemplo, totaldayminutes tiene una media de 180.31 con un rango de 34.95 a 324.95, y su distribución es simétrica (asimetría ≈ 0, curtosis ≈ 0). Variables como totalintlcalls y numbercustomerservicecalls muestran ligeros sesgos positivos.

Histograma y distribución de datos:

Histograma:

Con base en las descripciones:

- Distribuciones "sesgadas" o atípicas:

    - churn, internationalplan, voicemailplan y numbervmailmessages: Los histogramas presentan una barra alta a la izquierda (cero) y barras muy pequeñas hacia la derecha. Esto confirma la concentración en un solo valor (mayoría de clientes sin plan o con bajo número de mensajes).
    - numbercustomerservicecalls y totalintlcalls: Histogramas con barras altas separadas reflejan valores discretos y un rango limitado, confirmando patrones de llamadas.
- Distribuciones "normales" o más balanceadas: Variables como totaldayminutes, totaleveminutes, y totalnightminutes muestran histogramas con barras juntas, creando una forma de campana típica de distribuciones normales.

Boxplots:

Según la descripción, los boxplots destacan:

- Sin cajas visibles (churn, internationalplan): Solo aparece una línea y un círculo pequeño (valor atípico), reflejando la concentración en valores bajos.
- Cajas amplias o poco comunes (voicemailplan, numbervmailmessages):
    - voicemailplan: ocupa casi todo el gráfico, confirmando la distribución uniforme entre los valores 0 y 1.
    - numbervmailmessages: tiene una caja que abarca casi la mitad del rango, reflejando mayor dispersión.
- Distribuciones normales: Variables como totaldayminutes, totaleveminutes, etc., presentan cajas estándar con bigotes simétricos y líneas centrales cercanas al medio del rango.

Comprobación de asimetría y normalidad:

- Variables muy sesgadas (asimetría alta):

    - churn, internationalplan, voicemailplan: Concentradas en valores bajos con pocas observaciones en el otro extremo.
    - numbervmailmessages: Aunque no es binaria, presenta un comportamiento similar con valores acumulados en 0.
- Variables normales o cercanas a la simetría (asimetría ≈ 0):
    - totaldayminutes, totaleveminutes, totalnightminutes: Distribuciones bien balanceadas.
- Curtosis:
    - Valores como voicemailplan y churn presentan curtosis alta (>2), indicando concentraciones en los extremos.

4. Análisis bivariado

Correlación

- Matriz de correlación y mapa de calor(este refuerza visualmente las correlaciones):
  A partir de la matriz de correlación, se identificaron las siguientes relaciones más relevantes:
    - churn y internationalplan (0.259): Correlación positiva moderada, indicando que los clientes con un plan internacional tienden a hacer churn con mayor frecuencia.
    - churn y numbercustomerservicecalls (0.159): Correlación positiva moderada, sugiriendo que más llamadas al servicio al cliente están asociadas con una mayor probabilidad de churn.
    - totalintlminutes y totalintlcharge (0.999): Correlación casi perfecta debido a que ambas variables están directamente relacionadas.

Gráficos de dispersión

- La mayoría de los gráficos de dispersión presentan una tendencia diagonal de izquierda abajo a derecha arriba, lo que indica correlaciones lineales positivas débiles o moderadas.

- Destacado: numbervmailmessages vs voicemailplan muestra una fuerte correlación, con una línea en la parte superior del gráfico y un único punto atípico fuera de ella.

Boxplots y análisis por categorías

- churn y totaldayminutes: Los clientes con churn presentan menos minutos totales durante el día, lo que podría indicar que los usuarios con menor uso tienden a abandonar el servicio.
- internationalplan y totalintlminutes: La distribución de minutos internacionales es similar entre quienes tienen y no tienen plan internacional, lo que sugiere que esta categoría no afecta significativamente el uso.
- voicemailplan y numbervmailmessages: Los clientes con un plan de buzón de voz tienen más mensajes, aunque en general el número es bajo, lo que sugiere que esta función no es ampliamente utilizada.

5. Análisis de Distribuciones y Sesgos

Comprobación de Sesgo en los Datos

- Variable churn

En el boxplot de churn, no se muestra una caja completa; únicamente se observan puntos dispersos en el lado izquierdo del gráfico y un punto aislado en el extremo derecho, cerca del valor 1.

Esto sugiere que la variable está altamente desbalanceada, con la mayoría de los valores concentrados cerca de 0 y pocos en 1. El punto aislado podría representar un caso extremo que merece atención, posiblemente una categoría con menos frecuencia. El conteo de outliers detectados mediante IQR en esta variable es **aproximadamente 700**.

- Variable accountlength

El boxplot de esta variable presenta una caja central bien definida y equilibrada, sin valores extremos o puntos fuera de los límites del rango intercuartílico.

Esto indica una distribución uniforme y sin sesgo significativo en los datos.

- Variable internationalplan

El gráfico de esta variable es similar al de churn, mostrando puntos dispersos en un lado y un punto aislado en el otro extremo, sugiriendo un comportamiento similar.

Los outliers detectados son **aproximadamente 470** según el método IQR y también confirmados con Z-scores. Esto indica que esta variable tiene una distribución desbalanceada, con posibles valores extremos.

- Variable voicemailplan

El boxplot de esta variable se asemeja al de accountlength, con una caja de bigotes más pequeña y bien centrada en el gráfico.

No se identificaron valores atípicos en esta variable.

- Variable totalnightminutes

Este gráfico muestra una distribución equilibrada y consistente, con una línea central que indica simetría en los datos.

No se detectaron valores atípicos en esta variable, reflejando una distribución uniforme.

<u>Transformaciones de Variables</u>

- Para variables con sesgo significativo, como churn e internationalplan, las transformaciones matemáticas, como aplicar logaritmos o raíces cuadradas, pueden ayudar a normalizar las distribuciones.
- Estas técnicas permiten reducir el impacto de valores extremos y mejorar la calidad de los datos para análisis posteriores. Sin embargo, en

variables balanceadas como accountlength o totalnightminutes, estas transformaciones no son necesarias.

6. Detección de valores atípicos

Boxplots

Los diagramas de caja fueron utilizados para detectar valores atípicos.

- En la variable churn, el boxplot muestra un punto aislado cerca del valor 1 y puntos concentrados en la parte izquierda, indicando un desbalance significativo. Se detectaron aproximadamente 700 outliers según IQR.
- En la variable internationalplan, el boxplot es similar al de churn, con aproximadamente 470 outliers identificados por IQR.
- Para accountlength, voicemailplan y totalnightminutes, los boxplots muestran distribuciones equilibradas sin valores atípicos.

Análisis visual de gráficos de dispersión

Los gráficos de dispersión permiten observar si existen puntos que se alejan significativamente de la tendencia general. En el análisis realizado:

- Variables como churn e internationalplan muestran puntos que se apartan del resto, confirmando la presencia de outliers.
- Otras variables, como accountlength y totalnightminutes, presentan una distribución uniforme y sin patrones de valores extremos.

Z-scores o IQR

- Rango intercuartílico (IQR):

Se utilizó para identificar valores fuera de los límites establecidos por 1.5 veces el rango intercuartílico. Las variables churn e internationalplan mostraron altos niveles de outliers (700 y 470 respectivamente).

- Z-scores:

Aplicado para detectar valores que se alejan más de 3 desviaciones estándar de la media. Los resultados confirman los outliers identificados en variables desbalanceadas, como churn e internationalplan, mientras que otras variables no mostraron valores significativos.

7. Visualización de los datos

Gráficos de Barras

- Churn: Se observa que la mayoría de los clientes no abandonaron el servicio (Churn 0: 4293) en comparación con una minoría que sí lo hizo (Churn 1: 707). Esto es positivo para la retención de clientes.
- Internationalplan: La mayoría no cuenta con este plan (Internationalplan 0: 4527) mientras que una pequeña proporción sí lo tiene (Internationalplan 1: 473). Se necesita explorar qué ventajas o desventajas presenta este plan para los clientes.
- Voicemailplan: Un mayor número de clientes no utiliza este plan (Voicemailplan 0: 3677), pero una proporción significativa sí lo emplea (Voicemailplan 1: 1323). Podría ser útil analizar la correlación entre este plan y la satisfacción del cliente.

Gráficos de Lineas

Account length: Se identificaron tendencias interesantes:

- Totalevecharge y totaleveminutes muestran valores elevados en comparación con otras variables, pero tienen una caída abrupta alrededor del punto 166.
- Totalnightminutes y totalnightcharge tienen valores bajos en general, pero aumentan al final, cerca del punto 190.
  Las demás variables no presentan tendencias notables, mostrando fluctuaciones normales.

Diagramas de Dispersion

- DayCharge vs DayMinutes: Existe una clara relación lineal directa, representada por una línea diagonal que sube de izquierda a derecha.

- DayMinutes y Eves: Hay una alta dispersión, formando casi un patrón circular que cubre gran parte del gráfico.
- Numbercustomerservicecalls vs DayMinutes: Se observan cinco líneas horizontales que indican posibles grupos o niveles de atención al cliente.
- Numbervmailmessages vs DayMinutes: La dispersión es mayor en la parte superior, sugiriendo que los mensajes de voz son más comunes con altos minutos diurnos, mientras que hay una línea horizontal en la parte inferior.
- Numbervmailmessages vs Numbercustomerservicecalls: Se identificaron cinco líneas verticales no continuas, con dos de ellas destacándose hacia el final.

Gráficos de Densidad

La mayoría de los histogramas tienen distribuciones normales y el nivel de densidad es mayor en comparación con las barras de los histogramas.

- Totaldaycharge: Presenta una única barra en el histograma, y su nivel de densidad es significativamente bajo.
- TotalintIcalls y Numbercustomerservicecalls: Tienen varias barras muy separadas y delgadas, con niveles de densidad bajos que suben y bajan de manera notable.

8. Normalización de datos

Para este análisis, se optó por realizar transformaciones en los datos utilizando normalización y escalado. Estas técnicas permiten homogeneizar las variables para que sean comparables, además de mejorar la eficiencia de los algoritmos de modelado. La selección de las técnicas se justificó considerando las características de las variables:

- Estandarización (StandardScaler): Aplicada a variables continuas, como minutos y cargos totales, que tienen distribuciones aproximadamente

normales y diferentes escalas. Esta técnica es útil para algoritmos sensibles a la escala.

- Escalado Robusto (RobustScaler): Implementado en variables discretas, como conteos, que pueden contener valores atípicos. Este método maneja de manera efectiva distribuciones no normales y outliers.

Hallazgos Claves

Relación entre variables

- Las variables de tiempo y cargo diurno, vespertino, nocturno e internacional mostraron una fuerte correlación directa, confirmada por la reducción de desviación estándar tras la estandarización (transformed_std = 1). Esto sugiere que el consumo en minutos está directamente relacionado con los cargos en cada segmento horario.

Distribuciones inusuales:

- Número de mensajes de voz: La mediana original era 0, con una dispersión significativa (IQR = 17). Tras el escalado robusto, se homogenizó (IQR = 1), lo que indica que una proporción considerable de clientes no utiliza este servicio.
- Llamadas al servicio al cliente: Aunque la mediana inicial era 1, el IQR reducido tras el escalado (de 1 a 1) resalta que la mayoría de los clientes realiza pocas llamadas, pero algunos casos extremos requieren mayor atención.

Calidad de los datos

- No se detectaron valores faltantes ni errores evidentes, pero las variables relacionadas con cargos y minutos presentaban escalas heterogéneas, justificando la necesidad de normalización.

Variables más importantes

- Las variables relacionadas con el uso de minutos y cargos (totaldayminutes, totaldaycharge, totalintlminutes, etc.) destacaron como claves para describir el comportamiento del cliente, dado su impacto directo en los ingresos.
- El número de llamadas al servicio al cliente y mensajes de voz podría ser relevante para predecir el abandono (churn), especialmente debido a la presencia de outliers en estos atributos.

Estos hallazgos reflejan la importancia de realizar un adecuado preprocesamiento para obtener información clara y significativa de los datos.

Preparación para el modelado

Tras un exhaustivo análisis exploratorio de datos (EDA) sobre el conjunto de datos de telecomunicaciones, se han identificado diversos patrones y características que resultan fundamentales para la siguiente fase de modelado. La preparación adecuada de los datos es crucial para garantizar que los modelos predictivos capturen eficazmente los patrones subyacentes que conducen al abandono de clientes. A continuación, se detallan las decisiones y pasos necesarios para preparar los datos para la fase de modelado.

Estrategia de Selección de Variables

La selección de variables se ha realizado considerando tanto las correlaciones observadas como la relevancia práctica de cada variable:

Variables a Mantener

- Variables de uso y facturación: Se conservarán todas las métricas relacionadas con minutos y cargos, dado que han demostrado patrones de comportamiento significativos en el análisis bivariado.
- Llamadas al servicio al cliente (numbercustomerservicecalls): Esta variable mostró una correlación moderada con el abandono, sugiriendo su importancia como predictor.

- Plan internacional (internationalplan): Su correlación positiva con el abandono (0.259) la convierte en un indicador relevante.
- Antigüedad del cliente (accountlength): Se mantiene como potencial indicador de la lealtad del cliente.
- Variables de buzón de voz: Tanto voicemailplan como numbervmailmessages se conservan por su posible relación con la satisfacción del cliente.

Variables a Excluir

Se eliminarán variables con correlación casi perfecta entre sí (por ejemplo, totalintlminutes y totalintlcharge), manteniendo solo una de cada par para evitar problemas de multicolinealidad.

Estrategia de Transformación de Datos

- Normalización de Variables Numéricas
    - Aplicación de StandardScaler para variables continuas (minutos y cargos) que mostraron distribuciones aproximadamente normales.
    - Implementación de RobustScaler para variables con presencia significativa de valores atípicos, específicamente numbercustomerservicecalls y numbervmailmessages.

Tratamiento de Variables Categóricas

Las variables categóricas (internationalplan, voicemailplan) ya se encuentran codificadas en formato binario (0/1), por lo que no requieren transformación adicional.

Manejo del Desbalanceo de Clases

El análisis reveló un desbalanceo significativo en la variable objetivo (churn), con una proporción de 14% vs 86%. Para abordar esta situación, se proponen las siguientes estrategias:

- Submuestreo de la clase mayoritaria
- Aplicación de técnicas SMOTE para sobremuestrear la clase minoritaria

- Evaluación comparativa de ambos enfoques para determinar la estrategia óptima

Estrategia de Validación

Considerando el tamaño del conjunto de datos (5000 registros), se implementará:

- División inicial: 70% entrenamiento, 30% prueba
- Validación cruzada con k=5 folds en el conjunto de entrenamiento
- Estratificación por churn para mantener las proporciones de clases en cada fold

Métricas de Evaluación

Dada la naturaleza desbalanceada del problema, se utilizarán:

- AUC-ROC como métrica principal
- F1-score como métrica complementaria
- Precision y recall para análisis detallado de falsos positivos/negativos
- Matrices de confusión para visualización de resultados

Ingeniería de Características

Se propone la creación de nuevas variables para capturar patrones más complejos:

- Ratios de uso entre diferentes períodos del día
- Indicadores de variabilidad en el uso del servicio
- Métricas de intensidad de uso del servicio al cliente
- Variables de interacción entre predictores relevantes

Tratamiento de Valores Atípicos

La estrategia para el manejo de outliers será:

1. Mantener inicialmente los valores atípicos identificados
2. Utilizar RobustScaler para minimizar su impacto
3. Crear una versión alternativa del dataset sin outliers para comparación

Pipeline de Modelado

Se implementará un pipeline estructurado que incluya:

1. Fase de preprocesamiento con los scalers correspondientes
2. Implementación de la estrategia de balance de clases seleccionada
3. Proceso de selección de características
4. Implementación del modelo base

Esta estructura permitirá mantener la consistencia en las transformaciones, evitar el data leakage y facilitar la comparación entre diferentes aproximaciones metodológicas.

La implementación de estas estrategias se realizará de manera iterativa, permitiendo ajustes basados en los resultados preliminares y el rendimiento observado en cada fase del proceso de modelado.