MEMORIA CIENCIA DE DATOS 2

CLAUDIA CORONA

Analizamos el contenido del trabajo y hacemos un primer plan de acción:   
Tenemos tres DataFrames que requieren limpieza y un análisis exploratorio exhaustivo. El objetivo es preparar los datos para obtener insights clave que nos permitan identificar patrones relacionados con el churn de clientes.  
  
En esta primera etapa, comenzaremos con la limpieza y el análisis exploratorio del primer DataFrame. Esto incluirá:  
1. Identificación y manejo de duplicados.  
2. Tratamiento de valores nulos.  
3. Verificación de consistencia de datos.  
4. Creación de nuevas variables relevantes.  
5. Exploración inicial para entender las características principales del conjunto de datos.

PRIMER DATAFRAME

Una vez tratado con los duplicados, al empezar con el tratamiento de valores nulos, cree una nueva columna útil para el df: **`is\_churner`**  
  
La columna `churn\_date` indica la fecha en la que un cliente abandonó el servicio. Los valores nulos en esta columna (es decir, `NaT`) representan a clientes que no han abandonado el servicio y que, por lo tanto, todavía están activos.  
  
Para facilitar el análisis y la segmentación, podemos derivar una nueva columna llamada `is\_churner` que clasifique a los clientes en dos grupos:  
1. Churners: Clientes que tienen una fecha registrada en `churn\_date`.  
2. No churners: Clientes con `NaT` en `churn\_date`.  
  
La nueva columna simplifica el análisis porque convierte la información de `churn\_date` (una fecha) en una variable categórica más manejable. Esto permite segmentar y analizar los grupos de churners y no churners de manera más directa, sin manipular las fechas originales.  
  
  
La columna `is\_churner` permite analizar rápidamente el grupo de churners y no churners sin manipular directamente las fechas. Esto mejora la claridad y eficiencia del análisis, especialmente para realizar segmentaciones y explorar tendencias.

* Analizaremos las columnas con valores faltantes y las completaremos según el contexto.
  + gender, paymentmethod, zipcode: Completar con "Desconocido".
  + contract\_channel: Decidir si eliminar o mantener.
* Revisar la columna contract\_date para evaluar cómo manejar sus valores faltantes.

*Manejo de valores faltantes en `contract\_date`*  
Existen registros con valores faltantes en la columna `contract\_date`, lo que impide calcular métricas clave como la antigüedad del cliente. Estos registros corresponden a clientes activos (sin fecha de `churn\_date`), por lo que es importante mantenerlos en el análisis.  
Dado que estos registros son clientes activos y la columna `contract\_date` es fundamental para cálculos importantes, se recomienda:  
- Imputar la fecha promedio de contrato del resto del DataFrame.  
- Este enfoque permite asignar un valor razonable basado en los datos disponibles.  
- Es una solución consistente que evita perder información relevante al eliminar registros.  
Con esta imputación, los registros serán integrados en el análisis sin introducir sesgos significativos en las métricas calculadas.

El siguiente paso del análisis

**Revisar tipos de datos:**

* He de asegurarnos de que los tipos de datos sean correctos (por ejemplo, fechas en formato datetime, categóricas optimizadas como category).

**Crear nuevas columnas útiles:**

* **Edad:** Calcularla a partir de birth\_date.
* **Duración del contrato:** Calcular la diferencia entre contract\_date y churn\_date (o la fecha actual para clientes activos).

**DESPUES DE LA LIMPIEZA COMENZAMOS CON EDA**

Pasos que seguir

**1. Análisis Univariado** Examinaremos cada variable individualmente para entender su distribución:

- Variables categóricas: `gender`, `married`, `paymentmethod`, etc.

- Usaremos gráficos de barras para visualizar frecuencias.

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

- Variables numéricas: `age`, `contract\_duration\_days`, etc.

- Usaremos histogramas y boxplots para analizar distribuciones, outliers y rangos.

En cuanto a **outliers,** no se encontraron ningún dato resaltante, pero en términos de edad, decidí poner un mínimo de 15 años ya que clientes más jóvenes no los veía muy acorde con el análisis

**2. Análisis Bivariado** Exploraremos cómo las variables se relacionan con el churn (`is\_churner`):

- Compararemos distribuciones entre churners y no churners.

- Visualizaremos proporciones y distribuciones usando gráficos de barras apiladas, boxplots, e histogramas segmentados. Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

Conclusión Estado civil es un factor relevante para el churn.  
Los clientes no casados son más propensos a abandonar el servicio, mientras que los casados tienen una mayor retención.

Insight clave: Los métodos de pago automatizados parecen actuar como un factor protector contra el churn.  
Acción sugerida: La empresa podría fomentar el uso de métodos automáticos mediante incentivos para mejorar la retención

3. Análisis de Outliers :

Detección en variables clave:

- Identificar valores extremos en variables como `age`, `contract\_duration\_days`.

-Revisar si los outliers son errores o valores legítimos que deben conservarse.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

No parece haber una relación fuerte entre la edad (age) y el churn (is\_churner).  
Esto indica que el comportamiento de churn no depende directamente de la edad del cliente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Relación clara: Los churners tienden a tener contratos más cortos que los no churners.  
Insight potencial: La duración del contrato podría ser una métrica importante para predecir el churn.

Gráfico

Descripción generada automáticamente

Relación débil pero notable: Los churners suelen ser clientes sin hijos, mientras que los no churners incluyen una proporción mayor de clientes con hijos.  
Este patrón podría sugerir que los clientes con responsabilidades familiares son menos propensos a abandonar el servicio.

4. Análisis Multivariado Investigaremos interacciones entre varias variables clave, como:

- Combinación de `paymentmethod` y `married` en relación con el churn.

- Usaremos tablas cruzadas o gráficos de calor (heatmaps).

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

**insight principal:**

* Los métodos automáticos (transferencia bancaria o tarjeta de crédito) combinados con clientes casados son el grupo más estable.
* El cheque electrónico es una señal clara de mayor churn, especialmente entre los no casados.
* La categoría Desconocido debe analizarse más profundamente, ya que muestra altas tasas de churn.

Gráfico

Descripción generada automáticamente

**Insight clave:** La combinación de paymentmethod y contract\_duration\_days es un fuerte predictor del churn.

**4. Análisis Temporal**

**Objetivo:**

* Identificar patrones temporales en el churn y en la contratación de clientes.
* Explorar si ciertos períodos (meses o años) están asociados con mayor churn o contratación.

**Pasos:**

1. Distribución de churn por períodos (mes, año):
2. Tendencias de contratación (contract\_date):

Gráfico

Descripción generada automáticamente

**Gráfico : Churn por Mes**

* **Observación clave:**
  + El churn se distribuye de manera bastante uniforme entre los meses de marzo a septiembre.
  + No se observan picos significativos en ningún mes, lo que indica que no hay una estacionalidad clara en el churn.
* **Interpretación:**
  + La ausencia de una tendencia mensual específica sugiere que el churn puede depender más de factores individuales del cliente (como duración del contrato o método de pago) que de la temporalidad.

Interfaz de usuario gráfica, Gráfico, Aplicación

Descripción generada automáticamente

**Gráfico: Contrataciones por Año**

1. **Observaciones:**
   * El número de contrataciones crece de manera significativa en 2022, con un pico notable en comparación con los años anteriores.
   * En 2023, hay muy pocas contrataciones registradas, probablemente debido a que los datos están incompletos o a un rango temporal limitado para este año.
2. **Interpretación:**
   * El crecimiento en contrataciones en 2022 podría estar relacionado con campañas de marketing, nuevos servicios o promociones que atrajeron a nuevos clientes.
   * La menor cantidad de contrataciones en años anteriores sugiere que el crecimiento en la base de clientes es un fenómeno reciente.

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Gráfico: Proporción de Churn por Año de Contratación**

1. **Observaciones:**
   * Los clientes contratados en años más recientes (2022) tienen una proporción significativamente mayor de churn (más del 40%).
   * En contraste, los clientes contratados en años anteriores, como 2017 y 2018, muestran tasas de churn mucho más bajas (por debajo del 20%).
2. **Interpretación:**
   * Los clientes más recientes tienden a ser menos leales o a tener contratos más cortos, lo que los hace más propensos al churn.
   * Esto podría indicar un problema con la retención de clientes adquiridos recientemente o un cambio en la calidad del servicio o las expectativas del cliente.

SEGUNDO DF

**Plan Completo: Limpieza y EDA del DF2**

**1. Limpieza y Preparación de los Datos**

1. **Revisión de duplicados y nulos:**

Verificamos duplicados en todo el DataFrame y combinaciones de `customerid` + `billing\_date`.

- Resultado: 0 duplicados encontrados.

He de asegurarnos de que customerid y billing\_date sean únicos para cada registro.

1. **Valores nulos y faltantes:**
   * **Columnas críticas:**
     + monthlycharges: Analizar valores nulos (si los hay) y decidir si imputar con la mediana o la moda.
     + internetservice y servicios relacionados: Verificar si hay categorías desconocidas o nulas, y decidir cómo tratarlas.

Identificamos valores nulos en columnas clave:

- `monthlycharges` (273 registros). - `internetservice`, `onlinesecurity`, `onlinebackup`, `deviceprotection`, `techsupport` (varias columnas relacionadas con servicios).

- Decidimos eliminar los registros con valores nulos, dado que imputarlos podría comprometer la calidad de los datos.

2.**Consistencia de datos:**

* + Validar que las fechas en billing\_date sean cronológicamente válidas.
  + Comprobar la coherencia de los servicios contratados (Yes/No) y sus relaciones con los cargos mensuales.

-Confirmamos que la columna `billing\_date` tiene un rango válido entre enero y octubre de 2023.

- Detectamos facturación \*\*irregular\*\* en algunos clientes, lo que sugiere períodos no consecutivos de facturación.

**Relación entre servicios y cargos mensuales:**

- Creamos una nueva columna llamada `total\_services` para contar los servicios activos por cliente.

- Analizamos la relación entre `monthlycharges` y `total\_services`, identificando anomalías:

- Cargos altos con 0 servicios activos.

- Cargos bajos con múltiples servicios activos.

**Identificación y manejo de registros inconsistentes:**

- Creamos una bandera (`is\_inconsistent`) para marcar registros problemáticos.

- Resultados:

- 5,809 registros fueron marcados como inconsistentes (6% del total). Ejemplo de inconsistencia: Clientes con `total\_services = 0` pero con cargos mensuales mayores a 50. 5.

**Acción tomada**

- Guardamos los registros inconsistentes en un archivo CSV para análisis futuro.

- Eliminamos estos registros del DataFrame principal para garantizar un análisis confiable.

1. **Conversión de tipos:**
   * Convertir billing\_date a formato datetime.
   * Convertir las columnas categóricas (internetservice, onlinesecurity, etc.) a tipo category para optimización.
2. **Creación de nuevas variables útiles:**
   * **Número total de servicios activos:** Sumar las columnas onlinesecurity, onlinebackup, deviceprotection, etc.
   * **Cambio en cargos mensuales:** Identificar fluctuaciones en monthlycharges para detectar posibles anomalías.
   * **Mes/Año de facturación:** Extraer estas características de billing\_date para análisis temporal.

**2. Exploración de Datos (EDA)**

1. **Análisis univariado:**
   * Distribución de monthlycharges.
   * Frecuencia de cada categoría en internetservice y servicios relacionados.

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamenteImagen que contiene Gráfico

Descripción generada automáticamenteImagen que contiene Gráfico de barras

Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

**Conclusiones Generales**  
  
1. Cargos Mensuales (`monthlycharges`):  
 - Existe una gran variabilidad en los cargos mensuales, que parece estar estrechamente relacionada con el número de servicios contratados por cada cliente.  
  
2. Servicios Adicionales (`onlinesecurity`, `onlinebackup`, etc.):  
 - La mayoría de los clientes no utiliza estos servicios adicionales.   
 - Este dato podría ser clave para analizar el churn, ya que los clientes con menos servicios contratados suelen ser más propensos a abandonar el servicio.  
  
3. Servicio de Internet:  
 - "Fiber optic" es el servicio de internet más contratado por los clientes, destacándose como líder en el mercado.  
 - Sin embargo, también existe una proporción considerable de clientes que no tienen servicio de internet, lo que podría reflejar una segmentación diferente dentro de la base de clientes.

1. **Análisis bivariado:**

Pasos principales:

1. Relación entre `monthlycharges` y `total\_services`:

- Visualizar con un boxplot cómo varían los cargos mensuales según el número de servicios activos.

2. Comparación de `monthlycharges` por servicios específicos:

- Gráficos para analizar diferencias de cargos entre clientes con o sin servicios como `onlinesecurity`, `onlinebackup`, y `streaming`.

3. Relación entre `internetservice` y `monthlycharges`:

- Examinar cómo el tipo de servicio de internet afecta los cargos mensuales.

4. Impacto de `is\_irregular` en `monthlycharges`:

- Analizar si la facturación irregular influye en los cargos.

5. Patrones en combinación de servicios:

- Crear tablas cruzadas para identificar relaciones entre servicios contratados y el total de servicios activos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Interfaz de usuario gráfica, Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Número de Servicios Activos:**

* A medida que aumenta total\_services, la activación de servicios adicionales como onlinesecurity, onlinebackup, streaming, deviceprotection y techsupport es mayor.

**Segmentación de Clientes:**

* Los clientes con total\_services = 0 son los menos valiosos, sin servicios adicionales contratados.
* Los clientes con total\_services = 5 son los más valiosos, con una activación completa de servicios adicionales.

**Implicaciones:**

* Servicios como streaming, onlinesecurity y onlinebackup son indicadores clave de clientes premium.
* La inclusión de servicios adicionales está fuertemente asociada al número total de servicios contratados, destacando su importancia en estrategias de upselling.

**ANALISIS DE SERVICIOS Y CARGOS MENSUALES**

* **Número de servicios activos:**
  + A mayor cantidad de servicios contratados, mayores son los cargos mensuales.
  + Los clientes con 5 servicios activos generan los ingresos más altos.
* **Onlinesecurity:**
  + Los clientes con este servicio pagan más, indicando que tiene un costo adicional.
* **Onlinebackup:**
  + Similar a onlinesecurity, los clientes con este servicio tienen cargos más elevados, posiblemente como parte de un paquete premium.
* **Streaming:**
  + Los servicios de streaming están asociados a los cargos más altos, siendo un indicador clave de clientes de alto valor.
* **Facturación irregular:**
  + No se observan diferencias significativas en los cargos entre clientes con facturación regular e irregular.
* **Tipo de internet:**
  + Los clientes con fibra óptica (Fiber optic) tienen los cargos más elevados, seguidos por DSL. Los clientes sin internet tienen los costos más bajos.

**Conclusión:**

Los servicios contratados y el tipo de internet son factores clave que influyen significativamente en los cargos mensuales, lo que ayuda a segmentar y entender el valor de los clientes.

**COMBINACIONES DE SERVICIOS Y PATRONES AVANZADOS.**

1. **Identificar combinaciones frecuentes de servicios:**
2. **Relación de combinaciones de servicios con los cargos mensuales:**
3. **Segmentación de clientes según combinaciones de servicios:**

**Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente**

Conclusión del Análisis de Combinaciones de Servicios  
Clientes sin servicios adicionales (00000):  
Este segmento representa la mayoría de los clientes, pero es el menos rentable.  
Es crucial implementar estrategias de upselling para motivar a estos clientes a contratar servicios adicionales, aumentando así su valor para la empresa.  
Clientes con todos los servicios contratados (11111):  
Aunque son menos frecuentes, estos clientes generan los ingresos más altos.  
La retención de este segmento debe ser una prioridad, asegurando que permanezcan con la empresa a través de programas de fidelización o beneficios exclusivos.  
Combinaciones intermedias:  
Los patrones indican que ciertos servicios, como streaming y onlinesecurity, son más populares y accesibles.  
Focalizar las estrategias de marketing en estos servicios podría ayudar a convertir a los clientes de combinaciones parciales en clientes con configuraciones más completas, maximizando el ingreso promedio por cliente.  
Implicaciones Estratégicas:  
Diseñar campañas específicas de upselling dirigidas al segmento 00000 para aumentar su valor.  
Implementar programas de fidelización para retener a los clientes premium (11111).  
Aprovechar la popularidad de servicios clave (como streaming) para promover configuraciones intermedias que conduzcan a una mayor contratación de servicios.

TERCER DF

**Plan de análisis para el DF3**

1.Revisión de duplicados:

- Confirmamos que no existen registros duplicados en el DataFrame, ni combinaciones duplicadas de `customerid` y `date`.

- Esto asegura que cada registro es único y representa un evento válido.

2. Validación de valores nulos:

- Ninguna columna tiene valores nulos.

- Esto elimina la necesidad de imputación y asegura la integridad de los datos.

3. Detección de inconsistencias:

- Identificamos registros donde `tenure\_months` no coincide con el número de registros históricos por cliente.

- Esto permitió detectar posibles anomalías en la duración del contrato.

4. Identificación y manejo de outliers:

- Analizamos la columna `monthly\_penalty\_rate` y detectamos 266 registros con valores fuera del rango esperado.

- Estos registros fueron eliminados para evitar que distorsionen el análisis posterior.

5. Creación de nuevas variables:

- `calculated\_tenure`: Verifica la coherencia de `tenure\_months`.

- `remaining\_tenure`: Estima el tiempo restante de contrato.

- `monthly\_penalty\_rate`: Proporciona una tasa de penalización por mes para análisis adicional.

**Razón de estas acciones**:

- Garantizar la calidad y validez de los datos para análisis exploratorios y modelado.

- Eliminar valores extremos que puedan influir negativamente en los resultados.

- Crear variables derivadas que faciliten la interpretación y el análisis de los datos.

**2. Exploración de Datos (EDA)**

1. **Análisis univariado:**
   * Distribución de tenure\_months y tenure\_penalty.

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de líneas, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente Conclusiones Generales de las Gráficas

1. Distribución de Tenure (Meses)

Conclusión: La duración del contrato (tenure) muestra una tendencia ascendente. A medida que aumenta el tiempo de permanencia, más clientes alcanzan mayores tenures, lo que podría reflejar una buena retención de clientes en general.

2. Distribución de Penalizaciones Totales

- Conclusión: La mayoría de las penalizaciones totales se concentran en valores bajos, indicando que muchos clientes están en etapas iniciales de sus contratos o tienen menor acumulación de penalizaciones

3. Distribución de Penalización Mensual

- Conclusión: La penalización mensual se encuentra consistentemente entre 10 y 25 euros. Esto refleja una política uniforme en los costos asociados a la penalización. Estos datos pueden servir para analizar la relación entre la penalización mensual y el churn, así como para identificar posibles ajustes en las políticas de penalización.

UNIÓN DE LOS DF 🡪 MERGED\_DF

***Por qué es útil este análisis***

El análisis del DataFrame combinado permite consolidar toda la información relevante de los clientes en un único conjunto de datos. Esto nos ayuda a:

- Conectar patrones: Relacionar la información contractual, de facturación, servicios y duración del contrato para obtener una visión más completa de los clientes.

- Resaltar insights clave: Explorar cómo las decisiones relacionadas con los servicios y métodos de pago impactan en el churn.

- Identificar segmentos: Descubrir grupos de clientes vulnerables al churn o aquellos de alto valor que requieren estrategias específicas de retención o upselling.

**Por qué el trabajo previo no ha sido en vano**

Aunque hemos fusionado los datos en un único DataFrame, las limpiezas y análisis previos fueron fundamentales para garantizar:

1. Integridad de los datos: Cada uno de los DataFrames (DF1, DF2 y DF3) fue revisado para eliminar duplicados, valores inconsistentes y outliers.

2. Preparación para el merge: Las columnas clave, como `customerid`, se limpiaron y estandarizaron, asegurando que el merge fuera preciso.

3. Construcción de variables útiles: Creamos características nuevas (como `total\_services`, `monthly\_penalty\_rate`) que son esenciales para el análisis del DataFrame combinado.

**Pasos a seguir para completar el merged\_df**

1. Validación de características clave

- Revisar las columnas clave para asegurar que están completas y correctamente definidas.

- Analizar los registros con valores nulos para decidir si imputar, excluir o mantenerlos.

Texto

Descripción generada automáticamente

**-Validacion y ajuste del df combinado**

* Identificación de Tarifas Base

- Acción: Detectamos registros con `total\_services = 0` pero con valores positivos en `monthlycharges`.

- Estos registros representan clientes que pagan una \*\*tarifa base\*\* independientemente de los servicios adicionales.

- Identificar estos casos nos permitirá diferenciarlos del resto y entender su comportamiento único.

- Resultado:

- El análisis muestra que las tarifas base tienen un promedio de \*\*23.77\*\* con un rango concentrado entre \*\*18.25\*\* y \*\*24.80\*\*.

- Creamos una nueva columna, `base\_fee`, para etiquetar estos registros como "Sí" (tarifa base) o "No".

* Resolución de Inconsistencias en `tenure\_months`

- Acción: Revisamos y corregimos registros con inconsistencias en esta columna.

- `tenure\_months` es clave para entender la antigüedad y la penalización de los clientes, por lo que debe ser coherente con otras métricas como `remaining\_tenure` y `monthly\_penalty\_rate`.

- Resultado:

- Todas las inconsistencias en esta columna se resolvieron, dejando un DataFrame limpio para el análisis.

* Preparación para el Análisis

- Acción: Validamos los datos restantes y generamos una columna adicional que identifica los registros con tarifas base.

- Esta columna permitirá segmentar clientes de manera más precisa en futuros análisis.

- Resultado:

- Los registros están ahora listos para ser utilizados en análisis detallados sin que las inconsistencias afecten los resultados.

2. Análisis exploratorio avanzado (EDA)

**1. Análisis Univariado:**

* Objetivo: Explorar cada variable de forma independiente para identificar distribuciones, valores dominantes y posibles outliers.
* Acciones:
  + Distribución de variables numéricas (monthlycharges, tenure\_months, total\_services, etc.).

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente con confianza media

ANALIZAMOS LOS RESULTADOS DE LAS GRAFICAS:

**1. Distribución de monthlycharges**

**Observación:** Hay un pico significativo en torno a valores bajos (~20–30), que corresponde a clientes con servicios mínimos o tarifas base.

El resto de la distribución muestra una tendencia más dispersa, con varios picos en tarifas medias y altas.

No parece haber outliers severos.

* + Podemos segmentar clientes en grupos según rangos de tarifas (low, medium, high) para análisis posteriores.

**2. Distribución de tenure\_months**

**Observación:** La mayoría de los clientes tiene una duración del contrato cercana a 0, lo que indica clientes nuevos o recientes.

Hay algunos puntos extremos hacia valores altos (~8–9 meses).

Validar si los puntos extremos son relevantes o si representan datos inusuales.

* + Considerar crear una variable categórica de short-tenure, medium-tenure, y long-tenure.

**3. Distribución de total\_services**

**Observación:** Los clientes sin servicios contratados dominan la distribución.

Hay una distribución multimodal (picos claros en valores 1, 2, 3, etc.), reflejando combinaciones frecuentes de servicios.

Analizar combinaciones de servicios más comunes para identificar patrones de consumo.

* + No se observan outliers significativos.

**4. Distribución de remaining\_tenure**

**Observación:** La mayoría de los clientes tienen poco tiempo restante en su contrato (cercano a 0 meses).

Hay algunos valores dispersos hacia periodos más largos (>6 meses), que podrían corresponder a contratos a largo plazo.

Validar si los valores altos de remaining\_tenure se relacionan con clientes premium o con contratos antiguos.

* + Considerar dividir esta variable en categorías para facilitar el análisis.

Una vez analizado todas las graficas, el plan de acción que vamos a tomar es:

Segmentación por Rangos:

Vamos a crear categorías para las variables clave basándonos en sus distribuciones:

* **monthlycharges**: Clasificación en rangos low, medium, y high.
* **tenure\_months**: Clasificación en short-tenure, medium-tenure, y long-tenure.
* **remaining\_tenure**: Clasificación similal.

Validación de Outliers:

Identificar registros con valores extremos en las variables clave para confirmar su relevancia o ajustar en caso necesario.

- Revisar registros donde `tenure\_months` sea mayor a 8.

- Revisar registros donde `remaining\_tenure` sea mayor a 6.

- Analizar ejemplos concretos de estos valores extremos.

COMO VAMOS A TRATAR LOS OUTLIERS:

**Outliers en tenure\_months:**

Dado que tenure\_months está relacionado con la duración del contrato y es una variable crucial para el análisis, recalcularemos sus valores usando contract\_duration\_days si encontramos inconsistencias.

* + Si tenure\_months no coincide con contract\_duration\_days, recalcularemos el valor dividiendo contract\_duration\_days por 30 (redondeado al entero más cercano).
  + Si el valor no puede ser corregido lógicamente, lo reemplazaremos por la mediana de los valores consistentes.

**Outliers en remaining\_tenure:**

remaining\_tenure representa el tiempo restante en meses y debe ser consistente con la duración del contrato y tenure\_months.

* + Recalcular este valor como contract\_duration\_days / 30 - tenure\_months.
  + Si el resultado no es lógico (e.g., negativo), lo ajustaremos a 0 (indicando que ya terminó el contrato).

**Validación Posterior:**

* Revisaremos si después de los ajustes, los valores recalculados están en rangos razonables.
* Verificaremos si las distribuciones generales de las variables han cambiado significativamente tras los ajustes.

**2. Análisis Bivariado**

* Explorar relaciones entre pares de variables y cómo estas afectan al churn o a las métricas de negocio.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Una captura de pantalla de un videojuego

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Conclusión General de las Gráficas**

1. **Total Services y Monthly Charges:**

Existe una clara relación positiva: a mayor número de servicios contratados, mayores son los cargos mensuales.

Sin embargo, los clientes con menos servicios contratados presentan mayor variabilidad en los costos, lo que puede deberse a personalización de planes o promociones.

1. **Remaining Tenure y Monthly Charges:**

El tiempo restante del contrato no parece influir significativamente en los cargos mensuales.

Esto sugiere que los costos están más relacionados con el tipo y cantidad de servicios contratados que con la duración restante del contrato.

1. **Tenure Category y Monthly Charges:**

La duración del contrato (tenure) tampoco afecta significativamente los cargos mensuales.

Esto indica que los precios son consistentes independientemente de la antigüedad del cliente.

1. **Internet Service y Monthly Charges:**

Los clientes con fibra óptica tienen cargos mensuales más altos, destacándose como el servicio premium.

Los clientes sin internet o con DSL tienen cargos menores, reflejando niveles de servicio más básicos.

**3.Análisis Multivariado:**

comparaciones estadísticas entre churners y no churners,

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Relación entre Monthly Charges y Churn  
Los clientes que abandonan el servicio (churn) tienden a tener una mediana de cargos mensuales más alta en comparación con aquellos que permanecen.  
Existe mayor variabilidad en los cargos mensuales de clientes que no hacen churn.  
Relación entre Total Services y Churn  
Los clientes con churn tienen menos servicios contratados en promedio.  
Los clientes con más servicios son menos propensos a hacer churn, lo que sugiere una relación positiva entre el número de servicios contratados y la retención.

Ahora vamos a explorar combinaciones de servicios específicos y su impacto en el churn:

**Objetivo**:

* Identificar combinaciones de servicios contratados que estén asociadas con una mayor o menor probabilidad de churn.

**Pasos**:

* **Paso 1**: Crear combinaciones de servicios como columnas binarias o en un formato compacto (ej. cadena binaria 01011 que indique activación de servicios específicos).
* **Paso 2**: Calcular la frecuencia de las combinaciones de servicios y su proporción en clientes con churn (is\_churner=1) y sin churn (is\_churner=0).
* **Paso 3**: Visualizar combinaciones de servicios más frecuentes para clientes con churn y sin churn.
* **Paso 4**: Extraer insights clave para sugerir combinaciones de servicios que podrían mejorar la retención.

Gráfico

Descripción generada automáticamente con confianza media

**Conclusión:**

**Clientes más vulnerables al churn:**

* + Clientes con pocos servicios contratados o únicamente servicios básicos.
  + Usuarios sin servicio de internet o con una configuración limitada de servicios adicionales.

**4.Análisis Temporal:**

Identificar patrones en el churn a lo largo del tiempo para determinar si hay momentos específicos donde la probabilidad de churn es mayor.

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Tendencia Observada:**

* El churn muestra un incremento constante a lo largo de los meses.
* Marzo de 2023 tuvo el menor número de churners, mientras que septiembre de 2023 tuvo el mayor.

**Posibles Explicaciones:**

* Es posible que los contratos a corto plazo (e.g., 6 meses) estén expirando en estos períodos, incrementando el churn.
* Cambios en precios, calidad del servicio o actividades de la competencia pueden haber impactado negativamente en los clientes.

En este punto del trabajo, quería incluir un modelado predictivo, el plan inicial para ejecutarle era:

**Preparación de los datos**:

Separar las variables predictoras (X) y la variable objetivo (y).

Dividir los datos en conjuntos de entrenamiento y prueba.

Asegurarnos de que las variables categóricas y numéricas estén listas para ser procesadas.

**Selección del modelo**:

* + Elegir modelos adecuados para problemas de clasificación binaria:
    - Regresión logística.
    - Árboles de decisión o Random Forest.
    - Gradient Boosting (XGBoost, LightGBM).
  + Comparar el rendimiento de los modelos.

**Entrenamiento y evaluación**:

* + Entrenar los modelos con el conjunto de entrenamiento.
  + Evaluar el desempeño utilizando métricas como:
    - Precisión (accuracy).
    - Recall y F1-Score (importantes en churn para detectar abandonos).
    - Curva ROC-AUC.

**Interpretación del modelo**:

* + Identificar las características más importantes para predecir el churn.
  + Visualizar el impacto de estas variables.

**Generación de recomendaciones**:

* + Basar las estrategias de retención en los insights del modelo.
  + Diseñar intervenciones específicas para los clientes en riesgo.

Esos eran los pasos que estaba previsto seguir, pero porq fallo esta idea, bien el modelado que quería crear, me obligaba a forzar algunos tipos de datos, es decir teníamos que hacer lo siguiente:

**1. Preprocesamiento de Datos**

El preprocesamiento es esencial para asegurarnos de que los datos estén en un formato adecuado para el modelo.

**Uniformización de Tipos de Datos**

* **Qué hicimos:** Convertimos las columnas categóricas a str, ya que los modelos como OneHotEncoder necesitan datos uniformes para funcionar.
* **Por qué:** Los modelos y herramientas como scikit-learn no aceptan mezclas de tipos (por ejemplo, float y str) en una misma columna. La falta de uniformidad puede causar errores durante la transformación.

**Manejo de Valores Faltantes**

* **Qué hicimos:** Rellenamos los valores faltantes en las columnas categóricas con 'unknown'.
* **Por qué:** Los valores faltantes (NaN) son problemáticos para los codificadores y modelos. Al rellenarlos, evitamos que el modelo rechace estos datos.

**2. ColumnTransformer**

El ColumnTransformer nos permite aplicar diferentes transformaciones a distintos tipos de columnas dentro de un mismo conjunto de datos. Esto es útil porque:

* **Qué hace:**
  + Aplica **escalado** (StandardScaler) a las columnas numéricas para normalizar sus valores.
  + Aplica **codificación one-hot** (OneHotEncoder) a las columnas categóricas para convertirlas en variables binarias (0 y 1).
* **Por qué:** Los modelos basados en números no entienden las cadenas de texto directamente, por lo que las columnas categóricas necesitan convertirse en un formato numérico.

**3. Pipeline**

Un pipeline es una herramienta para encadenar pasos consecutivos de preprocesamiento y modelado. Esto asegura que:

* Todos los pasos se ejecuten en el orden correcto.
* Se eviten errores manuales al transformar datos fuera de orden.
* Se facilite la reproducibilidad del proceso.

**Qué hicimos:**

* Creamos un pipeline que:
  1. Aplica el ColumnTransformer para preprocesar los datos.
  2. Entrena un modelo (RandomForestClassifier en este caso) usando los datos transformados.

**4. Entrenamiento y Evaluación**

* Entrenamos el modelo con los datos de entrenamiento (X\_train y y\_train).
* Evaluamos el modelo en un conjunto de prueba (X\_test) para medir su rendimiento.

Pero al tener que transformar y cambiar los tipos de datos, no vi muy factible realizar este proceso, para la finalidad final del trabajo, pudiendo alterar posiblemente la calidad de los resultados esperados.

Bien, una vez retomada la idea inicial del proyecto, ya tenemos el df final limpio y analizado, vamos a empezar a responder a preguntas, dudad, inquietudes que han ido surgiendo durante el análisis, como hemos ido añadiendo varias columnas extras a los distintos df, ha quedado un df demasiado extenso asiq antes de nada, vamos a hacer una ultima limpieza de las columnas que ya sean irrelevantes:

**Columnas a eliminar:**

* customerid (si no es necesario para trazabilidad).
* birth\_date (redundante con age).
* contract\_date (redundante con tenure\_months).
* contract\_duration\_days (redundante con tenure\_months).
* billing\_date (si no se requiere para análisis específicos).
* calculated\_tenure (redundante con tenure\_months).

**Columnas mantenidas y su utilidad**

1. `gender`

- Caracterización sociodemográfica.

2. `paperlessbilling`

- Relación con churn por comodidad o insatisfacción tecnológica.

3. `paymentmethod`

- Análisis de métodos de pago asociados con churn.

4. `children`

- Factor sociodemográfico importante.

5. `married`

- Caracterización sociodemográfica.

6. `churn\_date`

- Crucial para análisis temporales y económicos del churn .

7. `country`

- Caracterización sociodemográfica y segmentación.

8. `is\_churner`

- Variable objetivo clave en el análisis.

9. `age`

- Caracterización sociodemográfica para .

10. `contract\_end`

- Complementa análisis de renovaciones y churn.

11. `churn\_month`

- Análisis temporal del churn.

12. `contract\_year`, `contract\_month`

-Agrupaciones y análisis temporales de contratos.

13. `monthlycharges`

- Clave para analizar impacto económico del churn.

14. `internetservice`

- Identificar servicios relacionados con churn.

15. Servicios contratados (`onlinesecurity`, `onlinebackup`, `deviceprotection`, `techsupport`, `phone\_lines`, `streaming`)

- Análisis de combinaciones de servicios y su impacto en churn.

16. `is\_irregular`, `is\_inconsistent`

- Detectar anomalías relacionadas con churn.

17. `tenure\_months`

- Clave para segmentar clientes por antigüedad.

18. `tenure\_penalty`

- Relación entre penalizaciones y churn.

19. `remaining\_tenure`, `monthly\_penalty\_rate`

- Relevantes para campañas de retención (pregunta 3).

20. `base\_fee`

- Segmentación de clientes.

21. Columnas categóricas (`monthlycharges\_category`, `tenure\_category`, `remaining\_tenure\_category`)

- Facilitan análisis categóricos.

22. `service\_combination`

- Detectar combinaciones de servicios más asociadas con churn.

23. ´churn\_month\_year`

- Análisis temporal del churn.

EL PROXIMO PASO ES RESPONDER A PREGUNTAS CLAVE PARA COMPLETAR EL ANALISIS

* **¿Qué características sociodemográficas tienen en común los churners? ¿Cuántos clientes activos están en riesgo en base a esta caracterización?**

Para resolverlo, primero vamos a **seleccionar columnas** relevantes:

gender: Género del cliente.

age: Edad del cliente.

country: País del cliente.

children: Número de hijos.

married: Estado civil.

is\_churner: Indicador de churn.

Después haremos el **análisis descriptivo**:

Distribución por género:

Comparar la proporción de churners (is\_churner == 1) entre hombres y mujeres.

Distribución por edad:

Identificar si los churners tienen una edad promedio distinta.

Detectar grupos de edad con mayor churn.

Distribución por país:

Verificar si el churn está concentrado en un país específico.

Estado civil y número de hijos:

Relacionar estado civil (married) y número de hijos (children) con churn.

Y por último identificación de clientes activos en riesgo

Identificar patrones comunes entre los churners:

Ejemplo: género, rango de edad, estado civil.

Aplicar esos patrones para segmentar clientes activos (is\_churner == 0) que compartan esas características.

Gráfico, Histograma

Descripción generada automáticamente

### Estado Civil

Distribución de estado civil entre churners: married

No 0.618985

Yes 0.377811

Desconocido 0.003203

### Distribución por País

Distribución de país entre churners: country

ES 0.999673

PT 0.000327

### Número de Hijos

Distribución de número de hijos entre churners: children

0 12468

1 1428

2 1400

## Identificación de Clientes Activos en Riesgo

Número de clientes activos en riesgo basado en patrones: 11671

Gráfico, Histograma

Descripción generada automáticamente

## Conclusiones

1. Los churners se concentran principalmente en el rango de edad de 30 a 50 años, con una media alrededor de 40.

2. Predominan los clientes no casados (61.9%) y sin hijos.

3. Los géneros masculino y femenino tienen distribución similar (49.8%).

4. Los churners son mayoritariamente de España (99.97%).

5. Basándonos en estas características, identificamos 11671 clientes activos en riesgo de churn.

Conclusión:

En este análisis, hemos identificado las características principales que definen a los clientes churners y estimado el número de clientes activos en riesgo basándonos en estas características.

1. Características sociodemográficas comunes entre los churners:

- Género: La proporción de churners entre hombres y mujeres es prácticamente igual (aproximadamente 50%), sin una diferencia significativa. Sin embargo, hay un pequeño grupo de clientes con género desconocido que podría indicar un problema de calidad en los datos.

- Estado Civil: La mayoría de los churners no están casados (61.9%), lo que podría indicar que el estado civil está relacionado con el comportamiento de churn.

- Edad: La distribución de edad muestra que los churners están concentrados principalmente entre los 30 y 50 años. Clientes más jóvenes y mayores tienen menor representación en el grupo de churners.

- País: El 99.97% de los churners pertenecen a España, lo que hace que este sea el país principal de análisis. Portugal tiene una representación insignificante.

2. Servicios asociados con los churners:

- Internet Service: El 64% de los churners tienen fibra óptica, mientras que el 26% usan DSL. Solo un 7% no tiene servicio de internet.

- Seguridad en línea: Más del 71% de los churners no tienen contratado este servicio, lo que podría asociarse a una menor percepción de valor del servicio.

- Método de Pago: El 55% de los churners utilizan cheques electrónicos, el método de pago con mayor proporción. Métodos automáticos como transferencia bancaria (14%) y tarjeta de crédito (13%) tienen menor representación.

3. Estimación de clientes en riesgo:

- Basándonos en patrones de género, estado civil y edad, se han identificado **11,671 clientes activos en riesgo de churn.**

- Al incorporar características relacionadas con servicios y método de pago, el número de clientes activos en riesgo ajustado asciende a 2,250 clientes

4. Implicaciones para la compañía:

- La mayoría de los churners no están casados, tienen entre 30 y 50 años, utilizan fibra óptica y prefieren métodos de pago no automáticos como el cheque electrónico.

- Los clientes activos en riesgo pueden ser objetivos prioritarios para campañas de retención personalizadas que destaquen el valor de los servicios de internet, seguridad en línea y beneficios de métodos de pago automáticos.

* **¿Cuál consideras que es la principal razón que hace a un cliente irse a la competencia? Cuantifica el impacto económico de los churners durante el año del dataset.**

**Identificar la razón principal del churn**:

* + Analizaremos patrones en los churners basándonos en:
    - Servicios contratados (ej. internetservice, onlinesecurity, etc.).
    - Métodos de pago (paymentmethod).
    - Facturación (monthlycharges, monthly\_penalty\_rate, etc.).
  + Observaremos correlaciones o relaciones significativas que puedan indicar el motivo principal.

**Cuantificar el impacto económico del churn**:

* + Calcularemos el costo económico de los churners observando:
    - **Cargos mensuales promedio (monthlycharges)**.
    - **Duración promedio del contrato antes del churn**.
    - Estimaremos las pérdidas basadas en estos datos.

Texto

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Razones Principales para el Churn**  
El análisis destaca que el churn está fuertemente influido por los siguientes factores:

**Métodos de Pago:**

* + Los clientes que usan el método de "Electronic Check" tienen la mayor proporción de churners (55.8%). Este método puede estar relacionado con problemas de comodidad o confiabilidad percibida.

**Combinaciones de Servicios:**

* + Las combinaciones de servicios sin seguridad o con funcionalidades básicas (por ejemplo, "No Security-Streaming") son más comunes entre los churners. Esto sugiere que los clientes con un paquete básico de servicios son más propensos a abandonar.

**Cargos Mensuales y Duración del Contrato:**

* + Los churners pagan un cargo mensual promedio de **75.80**.
  + La duración promedio antes de abandonar el servicio es **1.12 meses**, indicando un churn temprano y frecuente.

**Impacto Económico del Churn**  
La pérdida económica para la compañía debido a churners es significativa:

**Pérdida Promedio por Cliente:**

* + Cada cliente churner representa una pérdida estimada de **84.60**.

**Pérdida Total Estimada:**

* + Considerando un total de **15,296 churners**, la pérdida económica total asciende a **1,294,032.43**.

con tdo esto, lo mejor va a ser:

* **Optimizar los métodos de pago:** Implementar estrategias que promuevan métodos automáticos como transferencias bancarias o tarjetas de crédito.
* **Revisar las combinaciones de servicios:** Incentivar la contratación de paquetes con más servicios o beneficios adicionales.
* **Reducir el churn temprano:** Detectar y mitigar factores que influyen en la insatisfacción durante los primeros meses del contrato.

Para terminar, vamos a: Analizar la relación entre cargos mensuales y combinaciones de servicios, ya que es directamente relevante para entender las razones principales detrás del churn.

Gráfico

Descripción generada automáticamente

Patrones Temporales del Churn:

Gráfico, Gráfico de barras

Descripción generada automáticamente

Validar Factores Sociodemográficos con Impacto Económico

Texto

Descripción generada automáticamente

Conclusiones Generales:

- Servicios contratados: Las combinaciones más avanzadas son responsables de los mayores impactos económicos. Esto refuerza la importancia de estrategias específicas para retener clientes de alto valor.

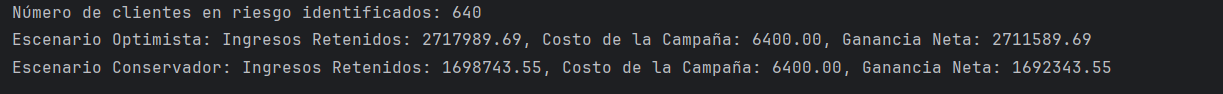
- Temporalidad: La creciente tendencia del churn a lo largo del tiempo indica la necesidad de intervenir con campañas de retención antes de los picos mensuales.

- Diferencias geográficas: Los clientes en Portugal representan un segmento crítico debido a su mayor impacto económico, lo que sugiere priorizar este mercado en estrategias de retención.

Para dar solución a todo esto, vamos a crear na campaña de retención de clientes:

**Fase 1: Diseño de la campaña de retención**  
Identificar a los clientes en riesgo:  
  
Basándonos en las características de los churners identificadas en las preguntas anteriores, seleccionamos a los clientes activos que coincidan con estos patrones:  
Género predominante: Masculino o femenino.  
Servicios contratados: Combinaciones avanzadas como "YesYesYesYesYes".  
Cargos mensuales altos (por encima de la media de los churners: $75.80 USD).  
Segmentos geográficos relevantes (priorizando clientes de Portugal, dado su impacto económico más alto).

Diseñar una oferta de retención:  
Descuento en los cargos mensuales: Ofrecer un descuento temporal del 10-20% en los cargos mensuales a los clientes en riesgo.  
Paquetes personalizados: Incluir servicios gratuitos (como seguridad en línea o respaldo en línea) durante los primeros 6 meses para incentivar la retención.  
Métodos de pago simplificados: Incentivar el cambio a métodos automáticos, como tarjeta de crédito o transferencia bancaria, que tienen menor churn.

**Fase 2: Análisis del impacto económico**  
Calcular el número de clientes en riesgo retenidos:  
  
Estimar la tasa de éxito de la campaña (por ejemplo, asumimos que el 30% de los clientes en riesgo aceptarían la oferta).  
Estimar el impacto en la facturación:  
  
Ingresos retenidos: Calcular los ingresos generados por los clientes retenidos durante los próximos 12 meses.  
Costo de la campaña: Evaluar el costo de los descuentos y beneficios adicionales ofrecidos.  
Simular escenarios:  
  
Escenario optimista: Tasa de retención del 50%.  
Escenario conservador: Tasa de retención del 20%.  
Comparar el impacto financiero en ambos escenarios.  
  


**Conclusiones:**

1. **Impacto potencial de la campaña**:
   * En el mejor de los casos, la campaña podría retener ingresos por más de **2.7 millones** con una inversión mínima de 6,400 unidades monetarias.
   * Incluso en un escenario conservador, se generarían ganancias netas significativas superiores a **1.69 millones**.
2. **Clientes prioritarios**:
   * El grupo identificado representa una oportunidad estratégica para maximizar ingresos retenidos y minimizar churn.
3. **Estrategia recomendada**:
   * Focalizar esfuerzos de retención en los **640 clientes de mayor riesgo**, priorizando aquellos con altos cargos mensuales y métodos de pago problemáticos.

Por último, responderemos a preguntas/ inquietudes que han ido surgiendo durante el análisis:

- ¿Qué combinaciones de servicios generan mayores ingresos para la compañía?

¿Existen combinaciones que incrementen el riesgo de churn?

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Imagen que contiene Gráfico

Descripción generada automáticamente

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

1. **Top 10 combinaciones de servicios con mayores cargos mensuales**:
   * Las combinaciones más caras incluyen configuraciones complejas o premium de servicios, destacando clientes con múltiples servicios contratados simultáneamente.
   * Por ejemplo, combinaciones como "No internet serviceYesYesNoYes" y "No internet serviceYesNoYesYes" muestran cargos mensuales promedio superiores a 110 unidades monetarias.
2. **Comparación de cargos mensuales entre churners y no churners por combinación de servicios**:
   * Los clientes con combinaciones de servicios más complejas tienden a tener mayores cargos mensuales en ambos grupos.
   * Sin embargo, los churners con estas combinaciones presentan valores ligeramente superiores, lo que podría indicar una relación entre los altos costos y la decisión de abandonar el servicio.
   * Combinaciones como "YesYesNo internet serviceYesYes" y "NoYesNoNo internet serviceYes" reflejan esta tendencia.

**Conclusiones**

1. **Servicios premium y churn**:
   * Las combinaciones de servicios premium tienen una mayor propensión al churn, lo que sugiere la importancia de enfocarse en este segmento para estrategias de retención.
2. **Relación entre costo y abandono**:
   * Los altos cargos mensuales son un factor clave en la decisión de churn, como se observa en las diferencias promedio entre churners y no churners.
3. **Recomendaciones**:
   * Implementar descuentos o beneficios adicionales para clientes con configuraciones premium, especialmente aquellos en riesgo de churn.
   * Focalizar las campañas de retención en los servicios y combinaciones más costosos para maximizar el impacto económico de las estrategias implementadas.

Para concluir con el trabajo, me hubiese gustado poder revisar el Impacto de las Conclusiones que hemos ido extrayendo, pero a falta de tiempo, creo que están todos los putos correctamente explicados.