Facultad de Estudios Superiores de Acatlán

Práctica 1:
Regresión Logística y Lineal

Diplomado en Ciencia de Datos

Módulo II:
Modelado Supervisado

Alumno:
Francisco Roman Peña de la Rosa

Profesora:
Act. Lorena Pineda Rodríguez

Ciudad de México Agosto, 2021

Regresión Logística

1. **Calidad de los datos**
   1.1. **Inspección del dataset**

   La primera etapa consistió en analizar algunas métricas importantes del dataset como: su dimensión, las variables y tipo de variables que lo integran son algunos ejemplos.
   - Tenemos un total de 21 variables y 7,043 registros.
   - La variable objetivo (y) es '**churn**'.
   - Se dio formato al dataframe para que los encabezados de todas las columnas fueran convertidos a minúsculas, y de esta forma se facilitará el llamado a cada variable.

   1.2. **Etiquetado de variables**

Variables discretas:

- gender
- partner
- dependents
- phoneservice
- multiplelines
- internetservice
- onlinesecurity
- onlinebackup
- deviceprotection
- techsupport
- streamingtv
- streamingmovies
- contract
- paperlessbiling
- paymentmethod
- seniorcitizen

Variables continuas:

- tenure
- monthlycharges
- totalcharges

Target

- churn

1.3. Registros duplicados y nulos

- **Registros nulos**

En un primer acercamiento en la exploración del dataset, observamos que no tenemos valores nulos:

```
data.isnull().sum()

customerid          0
gender              0
seniorcitizen       0
partner             0
dependents          0
tenure              0
phoneservice        0
multiplelines       0
internetservice     0
onlinesecurity      0
onlinebackup        0
deviceprotection    0
techsupport         0
streamingtv         0
streamingmovies     0
contract            0
paperlessbilling    0
paymentmethod       0
monthlycharges      0
totalcharges        0
churn               0
dtype: int64
```

- **Registros duplicados**

Por otra parte, en la inspección de valores duplicados

1.4. Limpieza y normalización de variables

Para esta sección no se realizó mucho trabajo ya todas las variables del dataset no presentaron comportamientos anómalos dentro de su estructura y formato. Al momento de realizar un '**describe**' para las variables continuas, '**totalcharges**' fue la que no estaba siendo incluida correctamente.

De lo anterior se realizó un análisis por separado de dicha variable, se tuvieron los siguientes hallazgos:

- 11 valores ausentes.
- La variable no estaba reconocida como tipo numérico, sino como tipo objeto. Por esta razón no estaba siendo incluida en el '**describe**' aplicado a continuas.

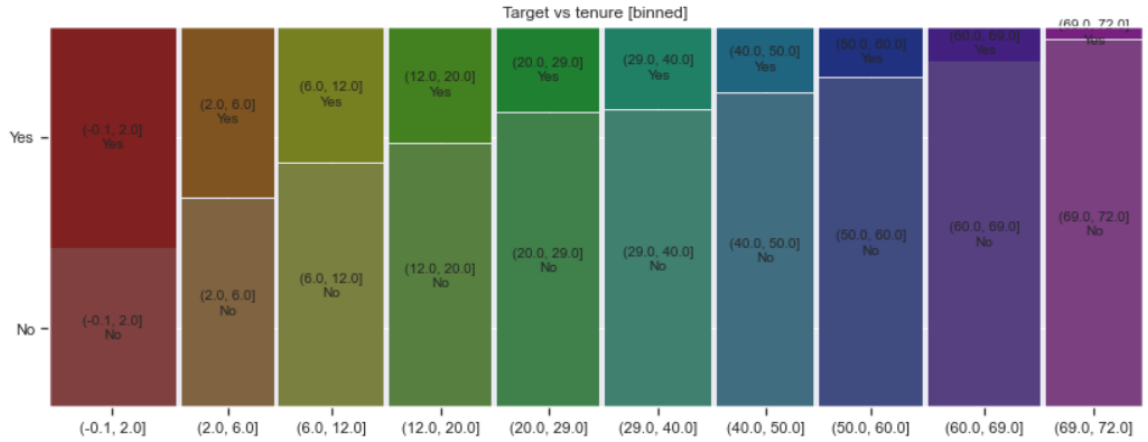
- Para resolver este problema, se reemplazaron los valores ausentes por '**NaN**'**.**
- La variable fue transformada a numérica.
- En este punto, podemos afirmar que tenemos **11 valores ausentes**, que no fueron reconocidos en la etapa anterior.

2. Análisis exploratorio de los datos (EDA)

El EDA se trabajó a través de cuantiles, y se comparó cada una de las variables continuas y discretas con la variable objetivo:

2.1. Variables continuas

Figura.1. '**Target**' vs '**tenure**'



3. Figura.2. '**Target**' vs '**monthlycharges**'

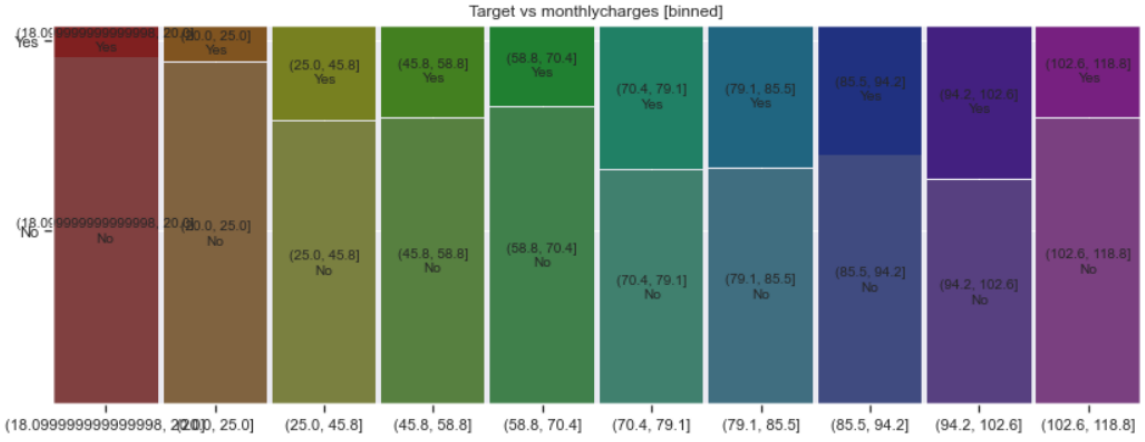
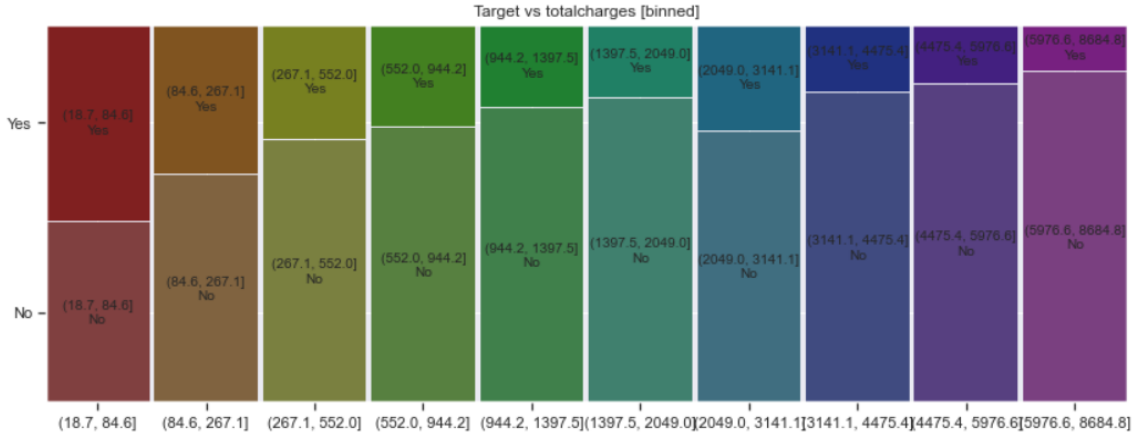


Figura.3. ‘**Target**’ vs ‘**totalcharges**’



3.1.1. Interpretación de resultados

- La probabilidad de '**churn**' disminuye a medida que se incrementa '**tenure**', sin embargo, en el **caso de 'monthlycharges**' el riesgo se incrementa.

3.2. Variables discretas

Figura.4. ‘**Target**’ vs ‘**gender**’

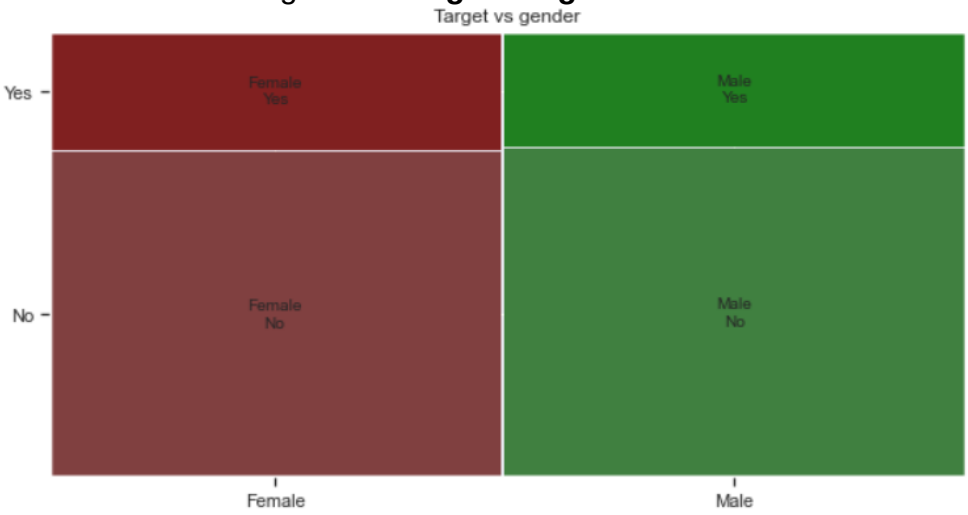


Figura.5. ‘**Target**’ vs ‘**partner**’

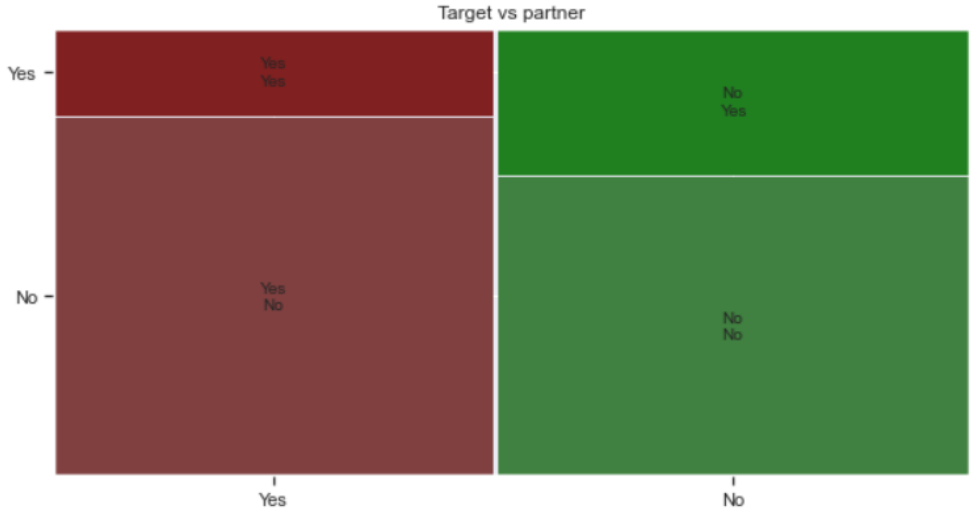


Figura.6. ‘**Target**’ vs ‘**dependents**’

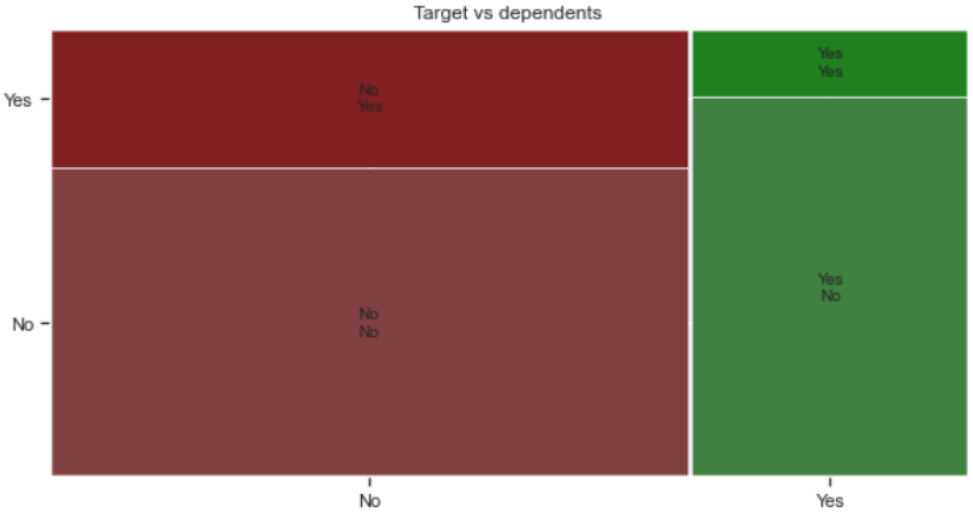


Figura.7. ‘**Target**’ vs ‘**dependents**’

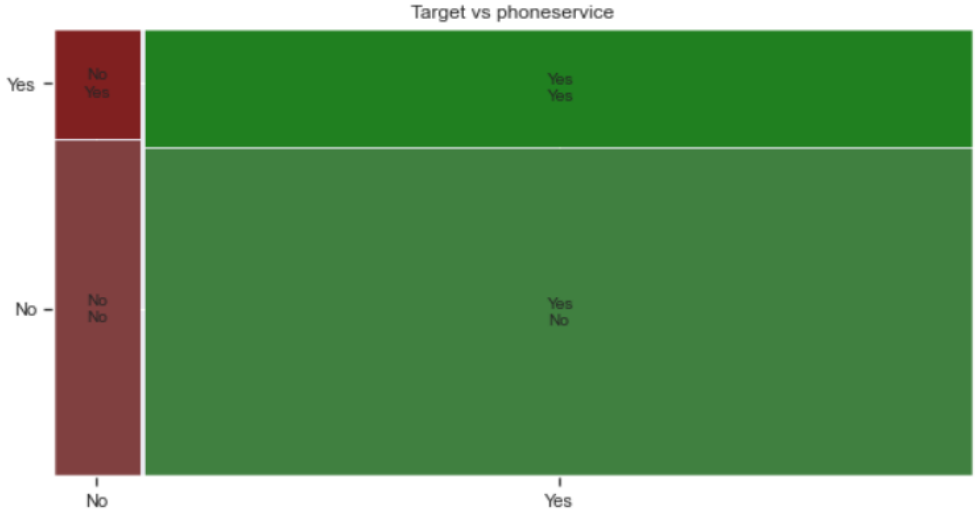


Figura.8. ‘**Target**’ vs ‘**multiplelines**’

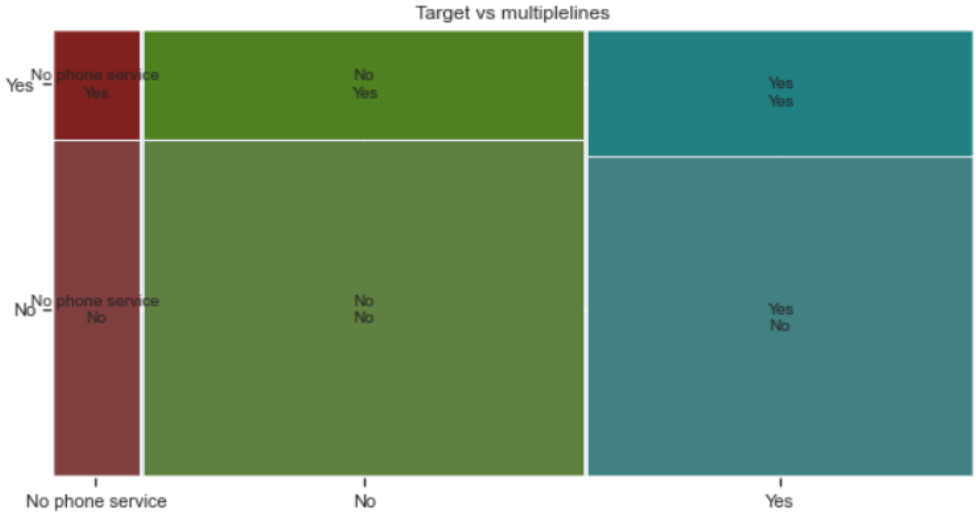


Figura.9. ‘**Target**’ vs ‘**internetservice**’

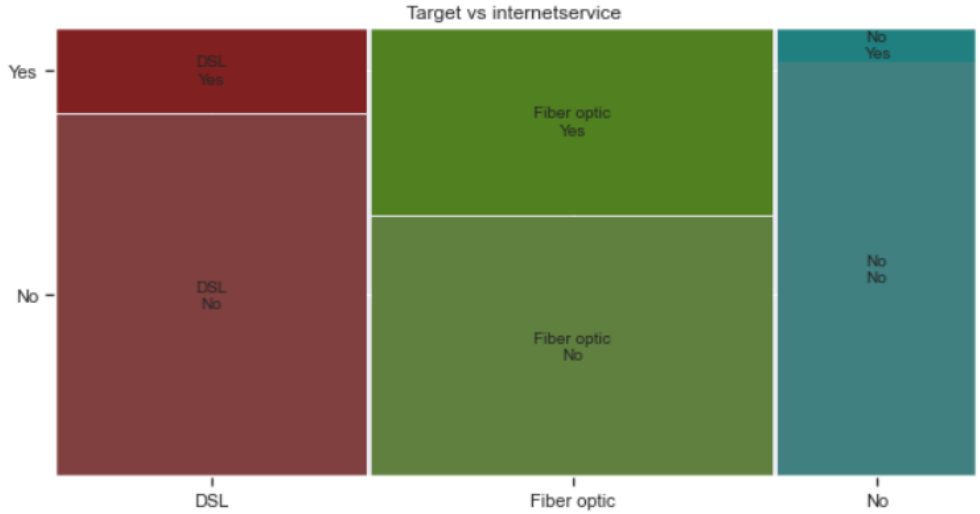


Figura.10. ‘**Target**’ vs ‘**onlinesecurity**’

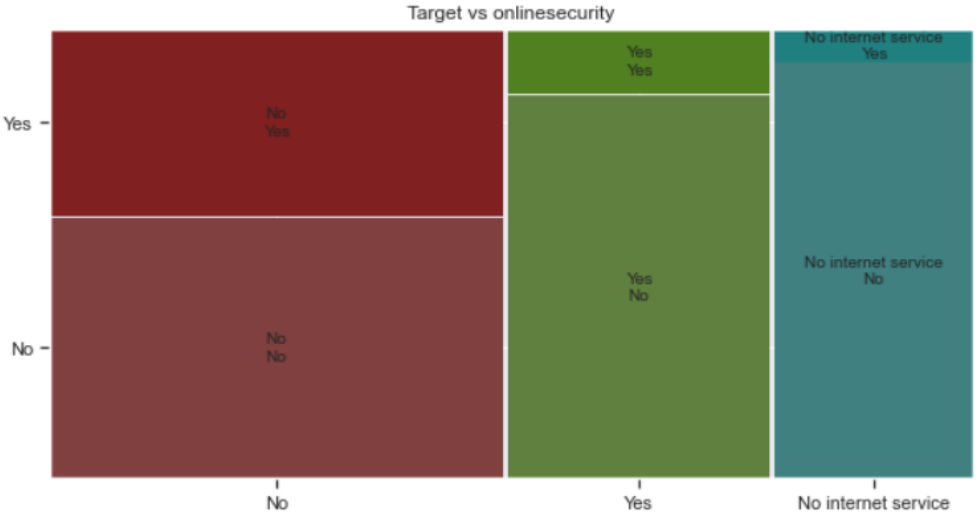


Figura.11. ‘**Target**’ vs ‘**onlinebackup**’

Figura.12. ‘**Target**’ vs ‘**deviceprotection**’

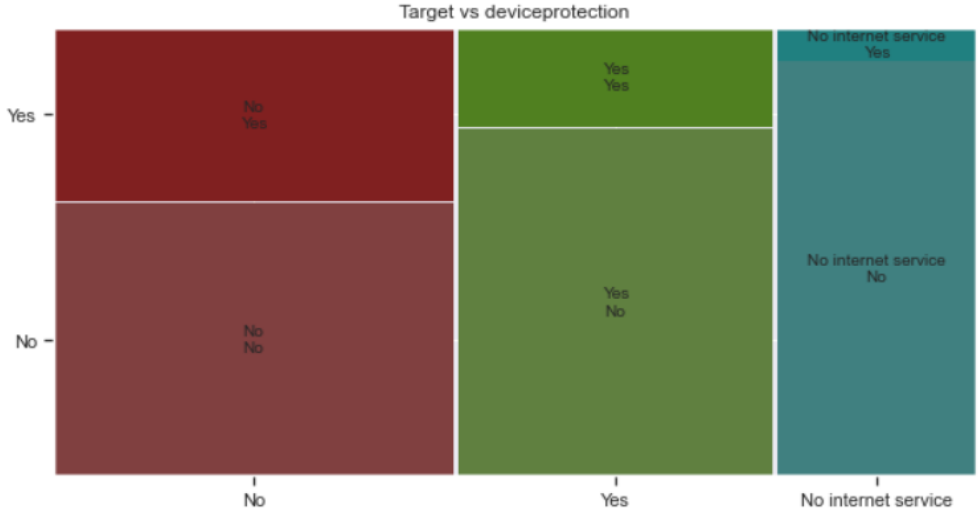


Figura.13. ‘**Target**’ vs ‘**techsupport**’

Figura.14. ‘**Target**’ vs ‘**streamingmovies**’

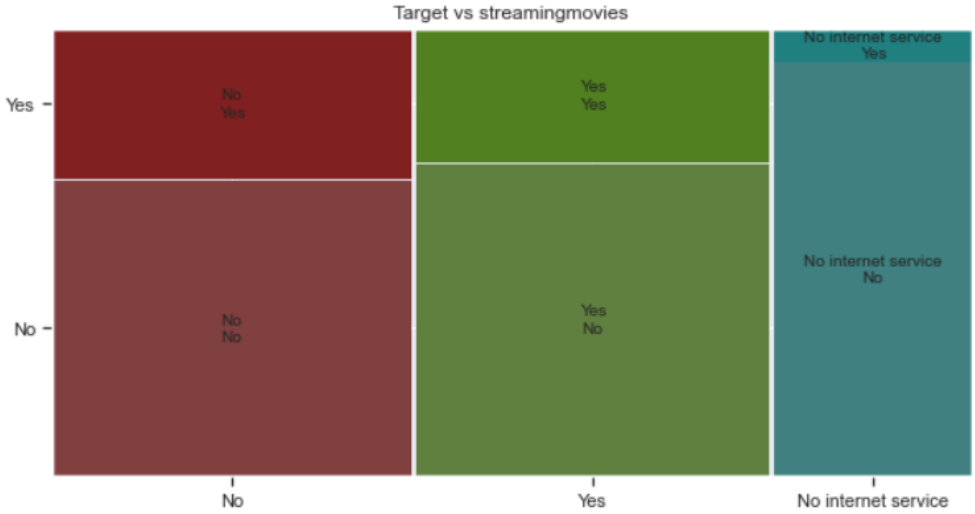
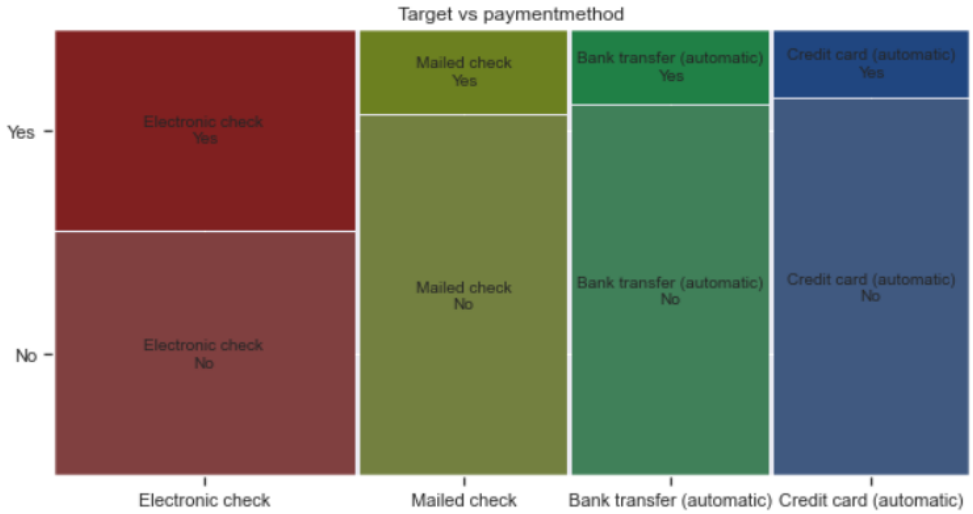


Figura.15. ‘**Target**’ vs ‘**paymentmethod**’



3.2.1. Interpretación de resultados

Algunos de los hallazgos fueron:

- Observamos que ‘**gender**’ no representa un gran impacto en el incremento/decremento de la probabilidad de ‘**churn**’.
- Por otro lado, también vemos que ‘**partner**’, ‘**dependents**’ y ‘**seniorcitizen**’ representan **menor** probabilidad de ‘**churn**’.
- Del gráfico de ‘**no_internetservice**’ también es posible interpretar menor riesgo de ‘**churn**’.
- Otra observación importante corresponde a las variables: ‘**month-to-month**’ y ‘**paymentmethod**’, las cuales muestran un alto riesgo de ‘**churn**’.

4. Ingeniería de variables

Las variables creadas para efectos del análisis fueron:

- ***Sección de Visualización***: contraste de variables continuas y categóricas VS target.
  - bin_tenure

    - bin_monthlycharges
    - bin_totalcharges

- ***Sección de Modelado:*** se obtienen las variables dummy para las categóricas correspondientes.
    - contract_Month-to-month
    - contract_One year
    - contract_Two year
    - paymentmethod_Bank transfer (automatic)
    - paymentmethod_Credit card (automatic)
    - paymentmethod_Electronic check
    - paymentmethod_Mailed check
    - gender_Female
    - gender_Male
    - internetservice_DSL
    - internetservice_Fiber optic
    - multiplelines_No
    - multiplelines_No phone service
    - multiplelines_Yes
    - onlinesecurity_No
    - onlinesecurity_Yes
    - onlinebackup_No
    - onlinebackup_Yes
    - deviceprotection_No
    - deviceprotection_Yes
    - techsupport_No
    - techsupport_Yes
    - streamingtv_No
    - streamingtv_Yes
    - streamingmovies_No
    - streamingmovies_Yes

5. **Análisis de correlación**
   En la **Fig. 16.** Se muestra la **Matriz de Correlación** entre las variables del dataset que tienen un coeficiente de correlación mayor a **0.7**. Dentro de estas variables se encuentran:
    - internetservice_No
    - onlinesecurity_No internet service
    - onlinebackup_No internet service
    - deviceprotection_No internet service
    - techsupport_No internet service
    - streamingtv_No internet service
    - streamingmovies_No internet service

Las variables arriba mencionadas tienen un valor de correlación igual a 1, por esta razón fueron eliminadas de los dataframes: **X_train** y **X_test**.

Figura.16. **Matriz de Correlación – Antes**

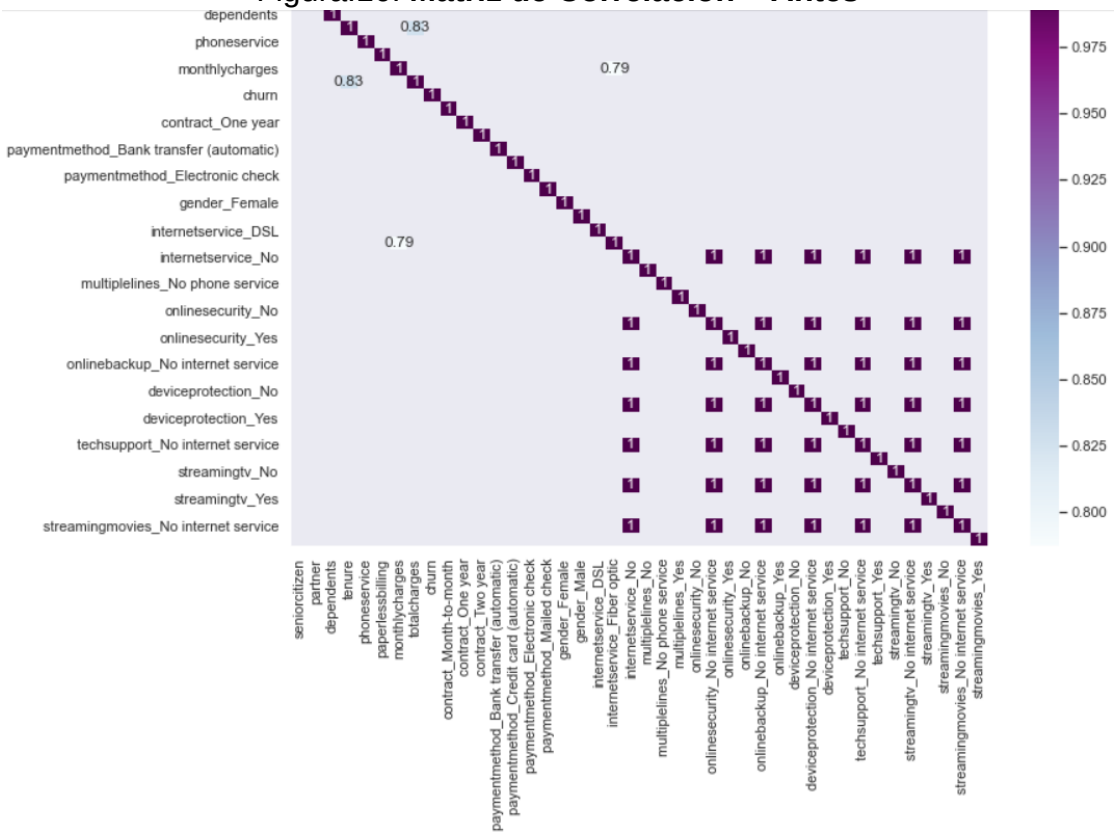
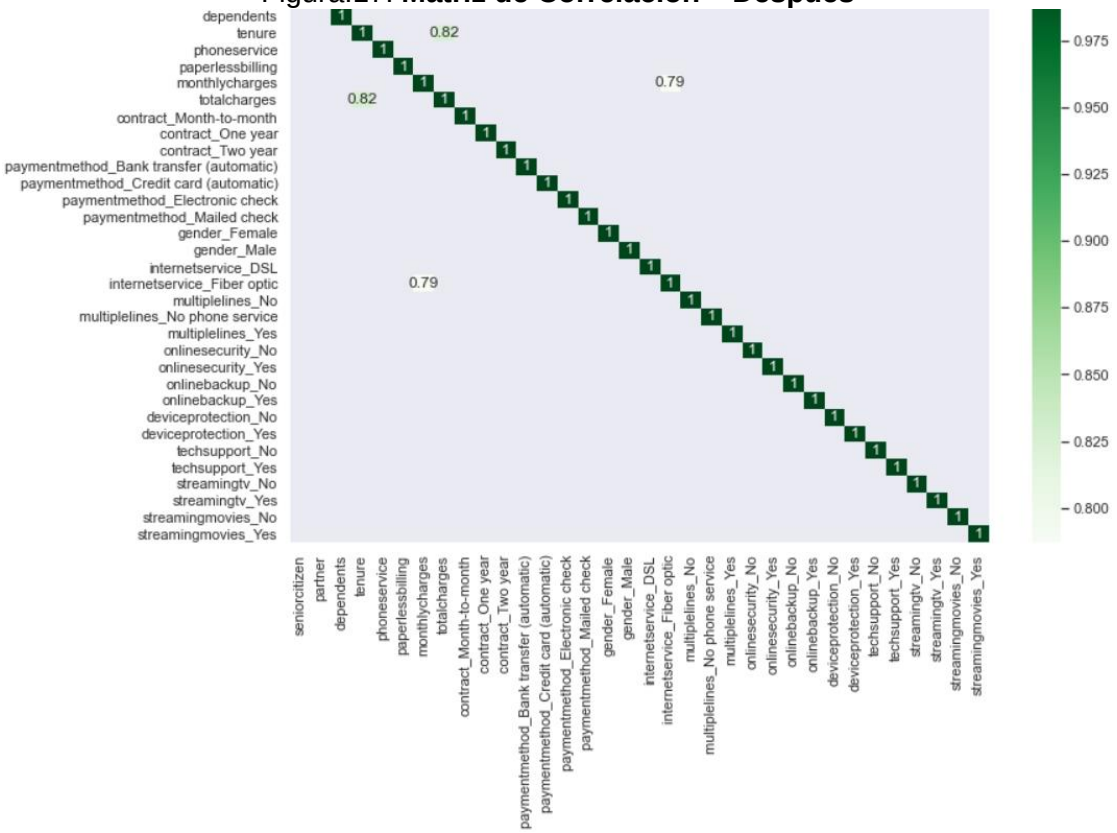


Figura.17. **Matriz de Correlación – Después**

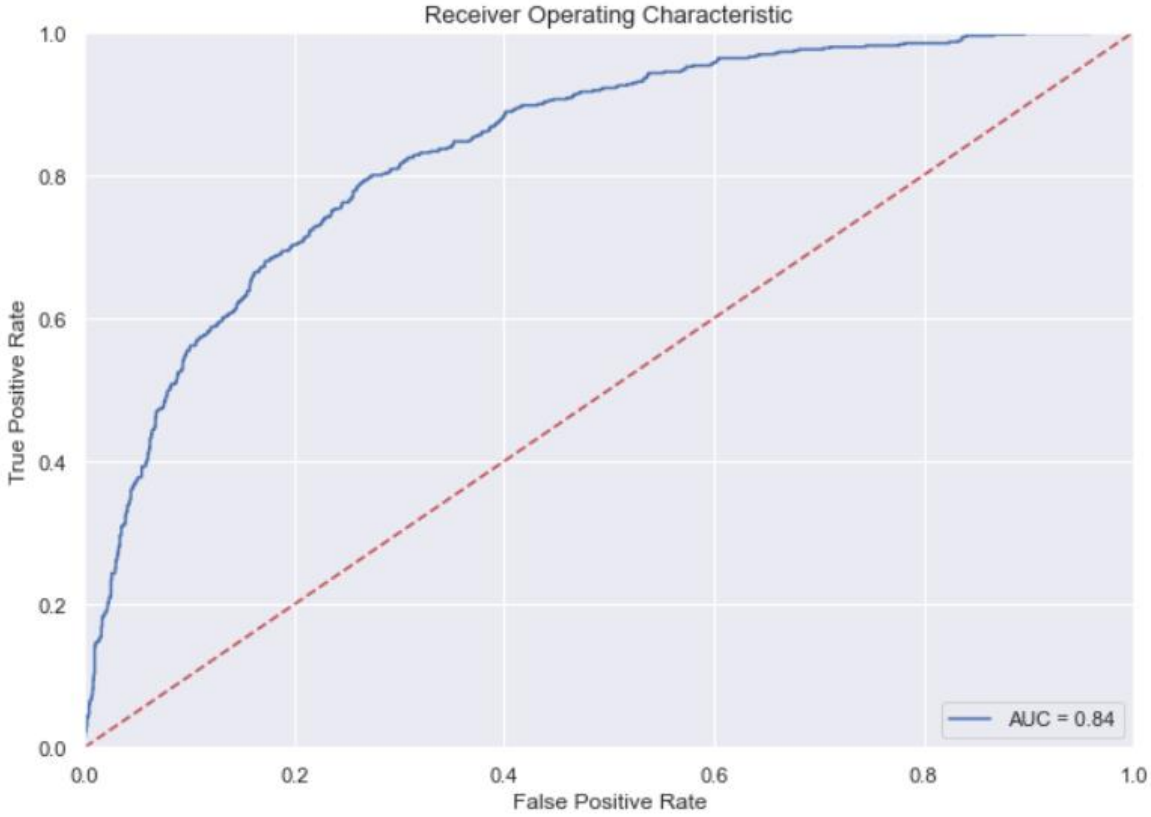


6. Resultados

Las métricas obtenidas al evaluar el desempeño de nuestro modelo de regresión logística fueron las siguientes:

```
Roc Validate: 0.837
Acc Validate: 0.806
Matrix Conf Validate:
[[1380  155]
[ 254  321]]
```

Figura.18. **Curva ROC – Área Bajo la Curva= 0.84**



7. Interpretación de Resultados

De lo anterior concluimos que el modelo es aceptable para ser implementado en la predicción de churn de nuestros clientes.

Regresión Lineal

- ***Justificación***

Durante el análisis de **'churn'**, vimos el escenario en donde un cliente con tipo de contrato **'mes-a-mes'**, **'facturación electrónica'**, y **'tenencia'** con valores de **0-12** representa un mayor riesgo de dejar la compañía.

También durante el EDA se observó que dentro de las principales variables que mejor describen el aumento en la probabilidad de churn son: **tipo de contrato**, **forma de pago**, **cargo mensual** y **si se trata de una persona pensionada**. Además de estas variables, para efectos del análisis se toman en cuenta: **tenencia, género** y la variable objetivo **cargos totales.**

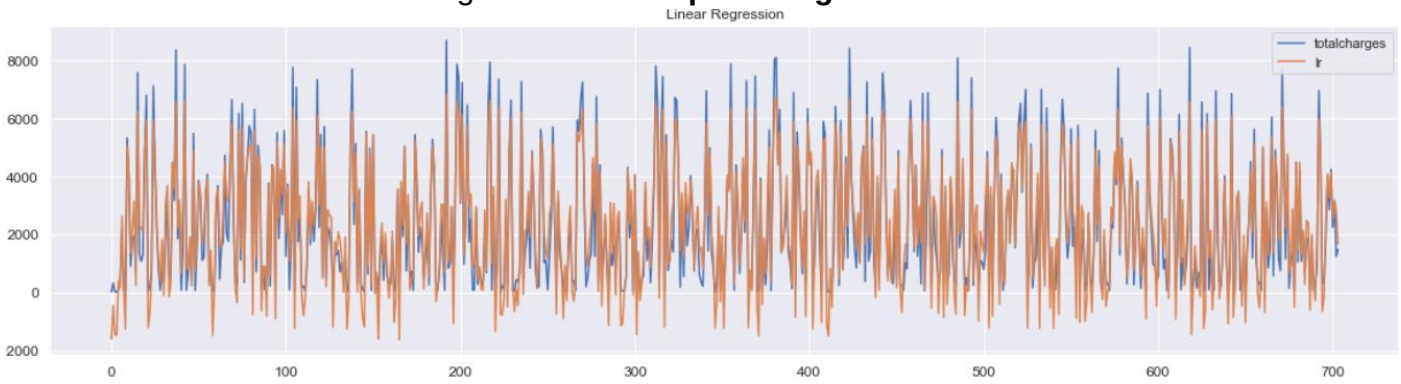
El tratamiento de las variables que fue realizado en el modelo de churn es retomado, y verificamos que no se tengan registros nulos o vacíos en el dataset.

- ***Métricas de Desempeño de los Modelos***

1. **Métricas de la Regresión Lineal**

```
El r2 score es 0.8940419125591534
El error cuadrático medio es 537922.5704229756
El error medio absoluto es 586.5291361767629
```

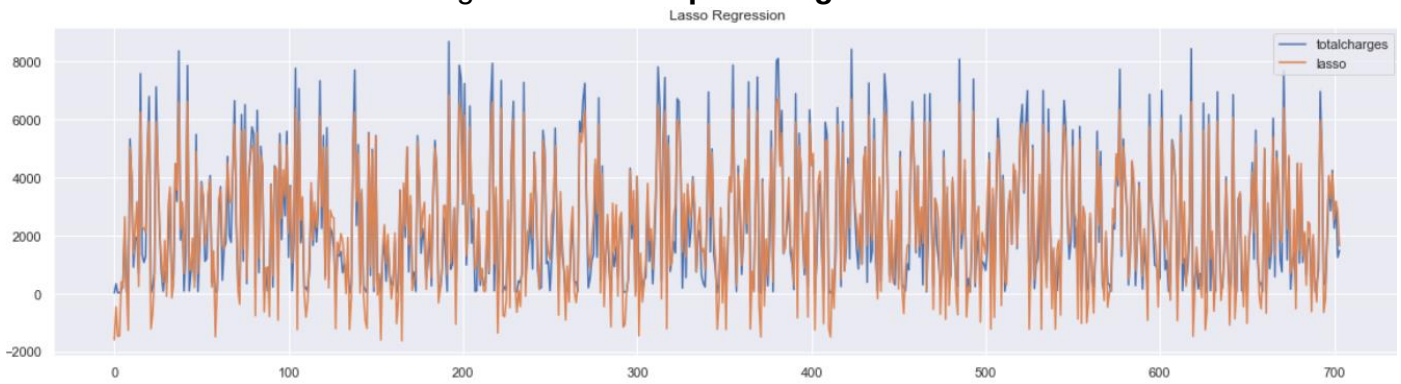
Figura.19. **Desempeño Regresión Lineal**



2. **Métricas de la Regresión Lasso**

```
El r2 score es 0.8940748164214757
El error cuadrático medio es 537755.5257865089
El error medio absoluto es 586.2432813381228
```

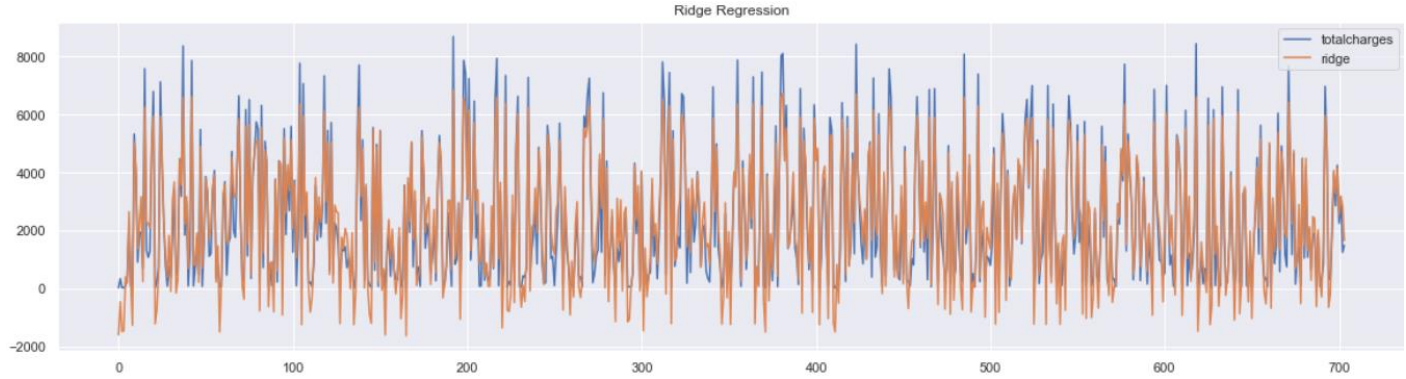
Figura.20. **Desempeño Regresión Lasso**



3. Métricas de la Regresión Ridge

```
El r2 score es 0.8940304391941787
El error cuadrático medio es 537980.8178123697
El error medio absoluto es 586.3729601221135
```

Figura.21. **Desempeño Regresión Ridge**



4. Métricas de la Regresión Elástica

```
El r2 score es 0.8470570259929616
El error cuadrático medio es 776453.0267869384
El error medio absoluto es 694.6792285453965
```

Figura.22. **Desempeño Regresión Elástica**

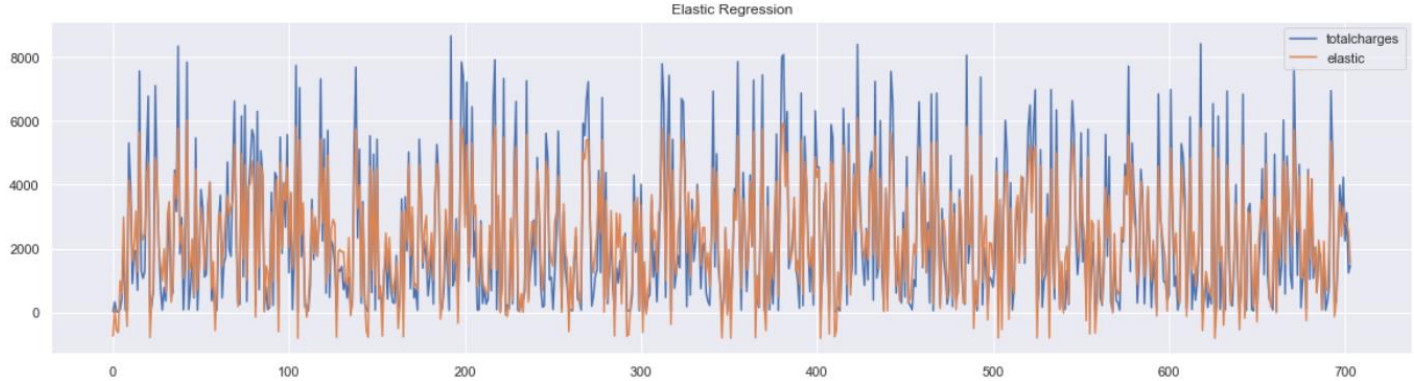


Figura.23. **Desempeño Regresión Elástica**

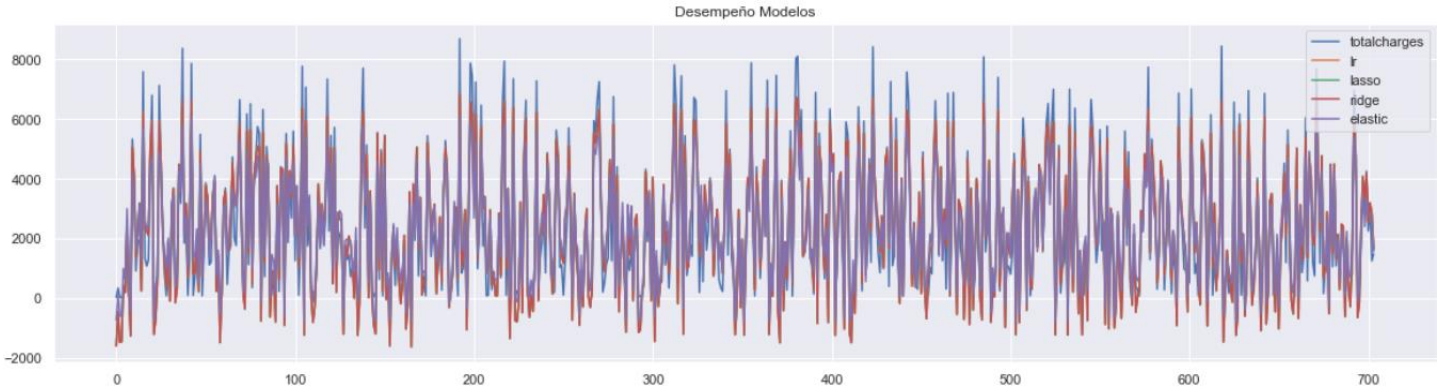


Tabla 1. **Métricas Modelos de Regresión.**

Métrica	Regresión Lineal	Regresión Lasso	Regresión Ridge	Regresión Elástica
R2 Score	0.894041910	**0.894074816**	0.894030439	0.847057026
Error Cuadrático Medio	537922.570422970	537755.525786500	537980.81.78123697	776453.026786938
Error Medio Absoluto	586.529136170	586.243281338	586.372960122	694.679228545

- ***Interpretación de Resultados***

De la Tabla 1., vemos que el modelo de mejor desempeño fue la **Regresión Lasso** tuvo el mejor valor **R2**, con base a ello se toma la decisión de elegirla para generar el archivo **pickle** correspondiente.
Algunas acciones que podría tomar la compañía son:

- Las personas jubiladas con un tipo de contrato con tipo de pago mes a mes tienen mayor riesgo de dejar la compañía, por esta ello podría aplicarse algún tipo de promoción:
  - Brindar un descuento inicial determinado en la contratación del servicio de internet por 1 o 2 años (según sea el caso), independientemente de la forma de pago. Pues vimos que el tipo de pago con cheque electrónico presenta mayor riesgo de churn.
  - En otro supuesto a este mismo grupo de clientes se puede adicionar un servicio como: películas streaming durante 6 meses más el descuento inicial del punto anterior para contratos de 1año y beneficios extra en el segundo año.
- Por otro lado, los clientes que tienen dependientes y se encuentran en el mismo esquema de pago mensual, pueden ser candidatos a recibir algunos de los beneficios de las personas pensionadas, sin embargo, dichos beneficios estarían sujetos a otras variables explicativas que no se encuentran dentro del dataset: edad, número de dependientes, ingreso mensual son algunos ejemplos.