Proyecto de Introducción a la Inteligencia Artificial

Entrega 2

Cristian Camilo Julio Mejía

Universidad de Antioquia

Introducción a la Inteligencia Artificial

Raúl Ramos Pollan

Octubre

2023

1. Exploración de Datos

En este proyecto se desea averiguar si un cliente abandonará o no la compañía de servicio de telecomunicaciones, por ende, la variable objetivo en este caso sería el valor de abanando (Churn Value).

Inicialmente para este item, se importaron las librerías necesarias para la implementación en colab del proyecto, seguidamente se hizo la conexión con la base de datos que se encontraba en Drive, para posteriormente cargar dicho elemento al entorno virtual. Luego de ello lo que se hizo fue revisar el Dataset con los comandos head, info y descibre posteriores a "data" que fue el nombre que se le asignó a la basa de datos.

1.1. Análisis de la variable objetivo

En este paso solo se realizó una gráfica en donde se mostró la mayor concentración que tenía la variable, la cual con "1" indica si el cliente abandonó la compañía, y "cero" si permaneció.

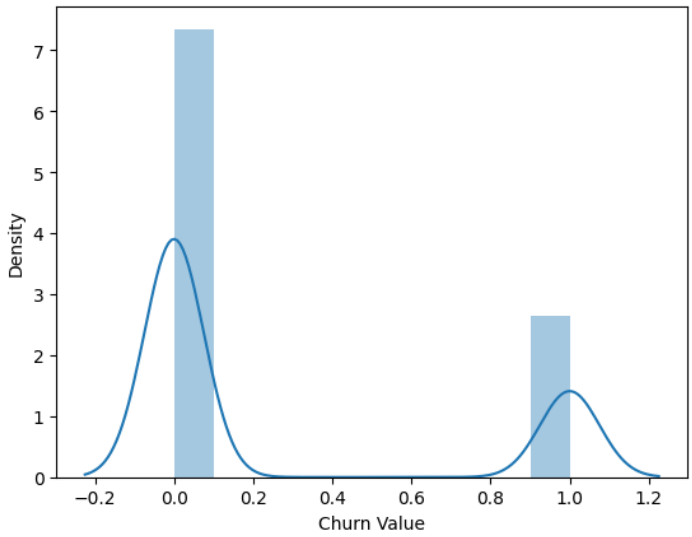

Fig. 1.0. Distribución de la variable objetivo

1.2. Exploración de variables

En este punto se realizaron 4 gráficas de barras en donde se determinó el comportamiento de la variable objetivo (Churn Value) con respecto a otras; como, por ejemplo, el género, el tipo de contracto que adquirió el cliente la forma en como pagó la persona, el tipo de servicio contratado, etc.

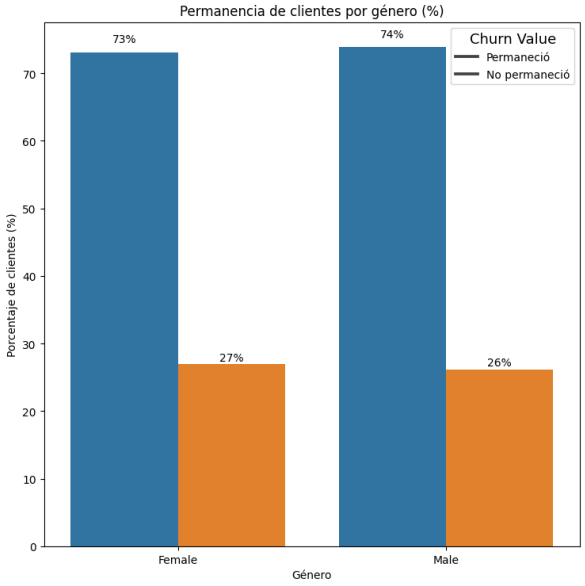

Fig. 2. Distribución de la variable objetivo por género

1.3. Conversión de Variables Categóricas a Numéricas

Cuando se cargó la base de datos al entorno virtual de Colaboraty se evidenció que habían algunas variables categóricas (Sexo, Tipo de internet, tipo de contrato, forma de pago, Servicio de teléfono, etc.) que podrían ser convertidas a valores numéricos para realizar un mejor tratamiento de los datos posteriormente; por ende, de la librearía sklearn se importó LabelEncoder, para que realizara dicho proceso.

9	Gender	7043 non-null	object
10	Senior Citizen	7043 non-null	object
11	Partner	7043 non-null	object
12	Dependents	7043 non-null	object
13	Tenure Months	7043 non-null	int64
14	Phone Service	7043 non-null	object
15	Multiple Lines	7043 non-null	object
16	Internet Service	7043 non-null	object
17	Online Security	7043 non-null	object
18	Online Backup	7043 non-null	object
19	Device Protection	7043 non-null	object
20	Tech Support	7043 non-null	object
21	Streaming TV	7043 non-null	object
22	Streaming Movies	7043 non-null	object
23	Contract	7043 non-null	object

Fig. 3. Tratamiento de Variables Categóricas

6	Contract	7043 non-null	int64
7	Dependents	7043 non-null	int64
8	Device Protection	7043 non-null	int64
9	Gender	7043 non-null	int64
10	Internet Service	7043 non-null	int64
11	Monthly Charges	7043 non-null	float64
12	Multiple Lines	7043 non-null	int64
13	Online Backup	7043 non-null	int64
14	Online Security	7043 non-null	int64
15	Paperless Billing	7043 non-null	int64
16	Partner	7043 non-null	int64
17	Payment Method	7043 non-null	int64

Fig. 4. Tratamiento de Variables Categóricas

1.4. Manejo de Datos Faltantes

En primera instancia, se evidenció que los valores faltantes estaban solo determinados por las razones o motivos de abandono (Churn Reason). Algunas personas no escribieron porque dejaron la compañía y dichos espacios permanecieron sin diligenciar, pero la tasa de dichos valores seguía siendo baja, por ende, se tuvieron que simular dichos valores generando indicies aleatorios y reemplazándolos con NaN en las columnas correspondientes a las copias de seguridad en línea, servicio de streaming para televisión y servicio telefónico. Luego se realizó la limpieza de los datos, teniendo presente técnicas de limpieza, como: eliminación de filas con valores faltantes, la interpolación y eliminación de datos duplicados.

```
Valores faltantes por columna:
CLTV                   0
Churn Label            0
Churn Reason        5174
Churn Score            0
Churn Value            0
```

Fig. 5. Tratamiento de Variables Faltantes

```
Online Backup        689
Online Security        0
Paperless Billing      0
Partner                0
Payment Method         0
Phone Service        352
Senior Citizen         0
Streaming Movies       0
Streaming TV         352
```

Fig. 6. Tratamiento de Variables Faltantes

Online Backup	0
Online Security	0
Paperless Billing	0
Partner	0
Payment Method	0
Phone Service	0
Senior Citizen	0
Streaming Movies	0
Streaming TV	0

Fig. 7. Tratamiento de Datos Limpiados

Finalmente, después de simular los valores faltantes y aplicar su limpieza se realizó una exploración inicial de los datos, haciendo histogramas de las variables numéricas y un análisis de correlación con respecto al valor y puntaje de abandono (Churn Value, Vs Churn Score)

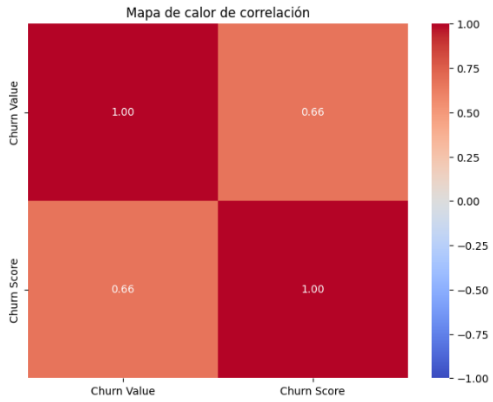

Fig. 8.0 Correlación de Valores

2. Referencia

IBM. (23 de Mayo de 2020). *Kaggle*. Obtenido de Kaggle: https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset