Ejercicio de abandono (Churn) en compañía de telecomunicaciones (Clasificación Binaria)

1.- Introducción

Objetivo

Desarrollar un modelo predictivo que nos ayude a determinar cuáles son los clientes que van a abandonar la empresa (Churn).

Glosario

- **Abandono (Churn):** Acción de abandonar la empresa (dejar de utilizar los servicios).
- **Modelo predictivo:** Modelo matemático probabilístico que nos ayuda a predecir o clasificar cierto evento, es decir, otorgarle una probabilidad de ocurrencia a cierto evento con base a observaciones del pasado.
- **Clasificación Binaria:** Modelo predictivo que arroja un resultado de si o no, según sea el caso, es decir, la probabilidad de que un evento ocurra o no, en este caso, de que el cliente abandone.

Planteamiento del problema

Los modelos para detectar Churn son muy utilizados para ver si un cliente con ciertas características va a dejar de utilizar los servicios otorgados por la empresa o bien, seguirá utilizándooslos. Tener esta clase de información ayuda a las áreas de negocio a tomar acciones estratégicas para determinar la viabilidad, beneficios y desventajas de este evento, por ejemplo: Se podrías realizar campañas de marketing personalizadas para los clientes que fueron detectados con el evento de abandono, y así, incentivar a que sigan consumiendo los servicios ofrecidos por la empresa.

2.- Fuente de datos y estructura del proyecto

Los datos para realizar este ejercicio, se extrajeron de la plataforma de Kaggel, la cual es una plataforma que ofrece una diversidad de Data sets para realizar ejercicios de ciencia de datos (El link lo encontrarán en los anexos).

El Conjunto de datos se llama "Clientes de Telecomunicaciones", el cual viene en formato CSV. Las variables que vienen en el data set hacen referencia a características de los clientes, como pueden ser, cuantos servicios tenían contratados en la compañía, temporalidad de la adquisición de servicios del cliente, género del cliente, forma de pago, monto de los servicios, edad (rango), entre otras.

La estructura del proyecto se divide en 3 apartados, los cuales son:

1. **Archivos CSV:** Contiene el archivo CSV original y la salida de los datos preprocesados.

2. **Documentación:** Contiene los documentos necesarios para entender el problema y el desarrollo del mismo.
3. **Modelo:** Contiene los 3 modelos que se probaron y sus métricas de evaluación.

3.- Desarrollo del modelo

3.1 Carga y Limpieza de Datos

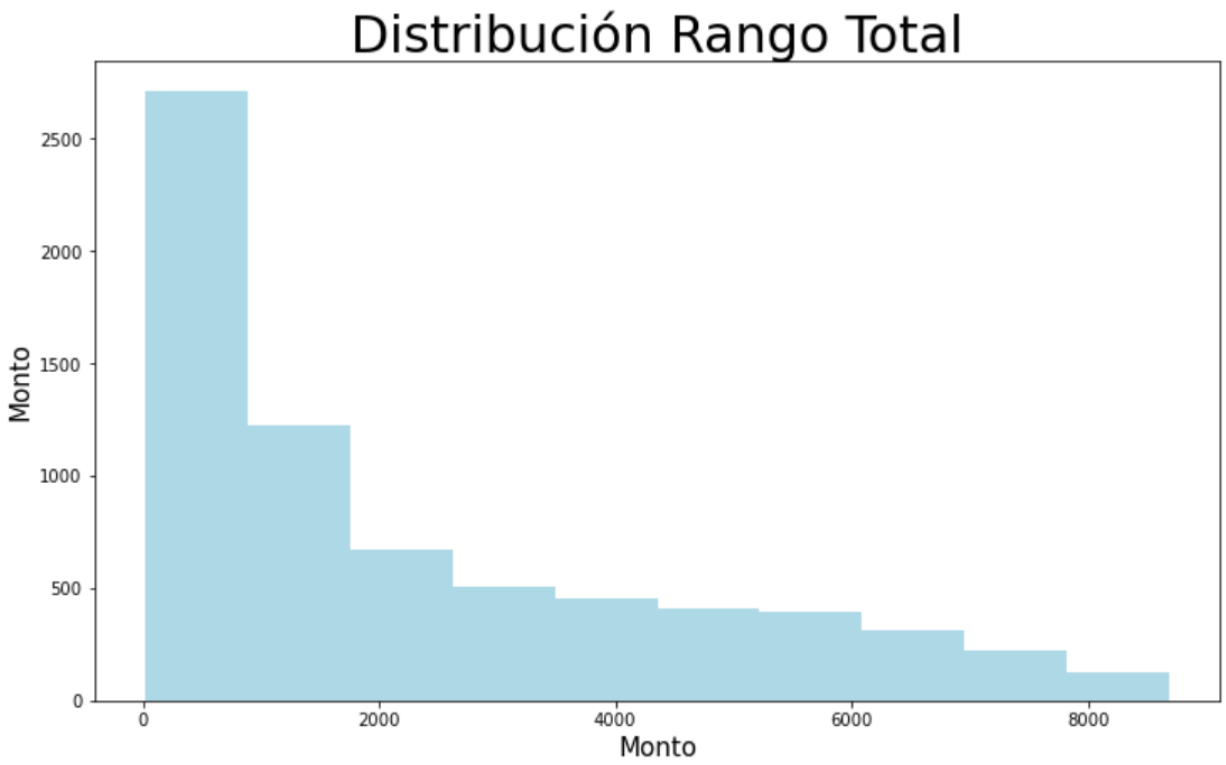
Se realizo el proceso de carga de datos como un CSV y se verificaron las dimensiones del Data Set, las cuales son de 7,043 filas (clientes) y 21 columnas.

En el proceso de limpieza, se verificaron los tipos de datos y si el data set tenía valores nulos, esto para garantizar la consistencia de los datos y así el modelo pudiera arrojar mejores resultados. En el proceso, se detectaron 11 valores nulos en la columna de cargo total, al ser tan pocos datos, se optó por eliminar estos datos.

3.2 Análisis exploratorio (Perfilamiento de datos)

En esta etapa se realizó un análisis exploratorio de los datos, viendo cuales son las variables que podrían influir más en el evento de abandono.

Figura 3.2.1. Histograma de monto total por cliente



En la gráfica anterior, podemos ver que la mayoría de clientes pagan menos de $2,000 por sus servicios de telecomunicaciones. Posteriormente, generamos una nueva columna por rangos para ver los números de estos clientes, los resultados son los siguientes:

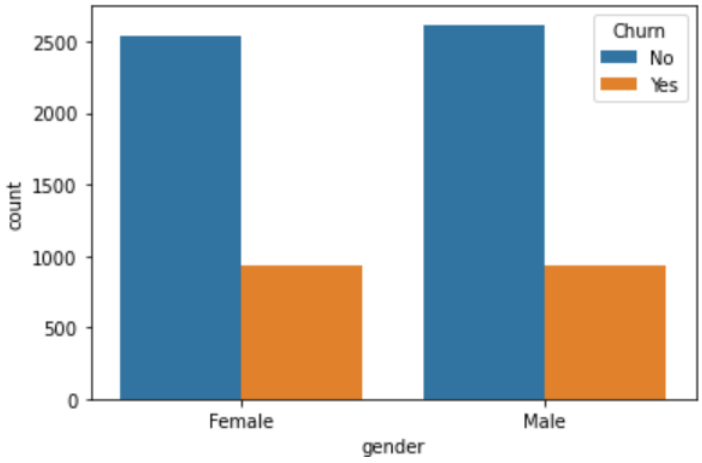
Tabla 3.2.1. Tabla de monto total por rango.

RANGOS_MONTO	CLIENTES	TotalCharges	%PROPORCION_CLIENTES	%PROPORCION_MONTO	MONTO_PROMEDIO
0 - 999	2893	1,037,347.40	41.14	6.46	358.57
1,000 - 1,999	1283	1,849,077.20	18.25	11.52	1,441.21
2,000 - 3,999	1208	3,544,243.75	17.18	22.07	2,933.98
4,000 - 5,999	956	4,753,709.25	13.59	29.61	4,972.50
6,000 o más	692	4,871,791.10	9.84	30.34	7,040.16

Con la tabla, podemos ver que aproximadamente el 60% de los clientes tienen un monto total menor a $2,000, pero que estos clientes, solamente componen aproximadamente el 18% de los ingresos total de la empresa, es decir, no son clientes muy rentables.

El siguiente análisis fue realizado con el genero de los clientes, pero con la variable de Churn, es decir, si hay algún género que determine más si realizará churn o no, la gráfica quedó de la siguiente forma:

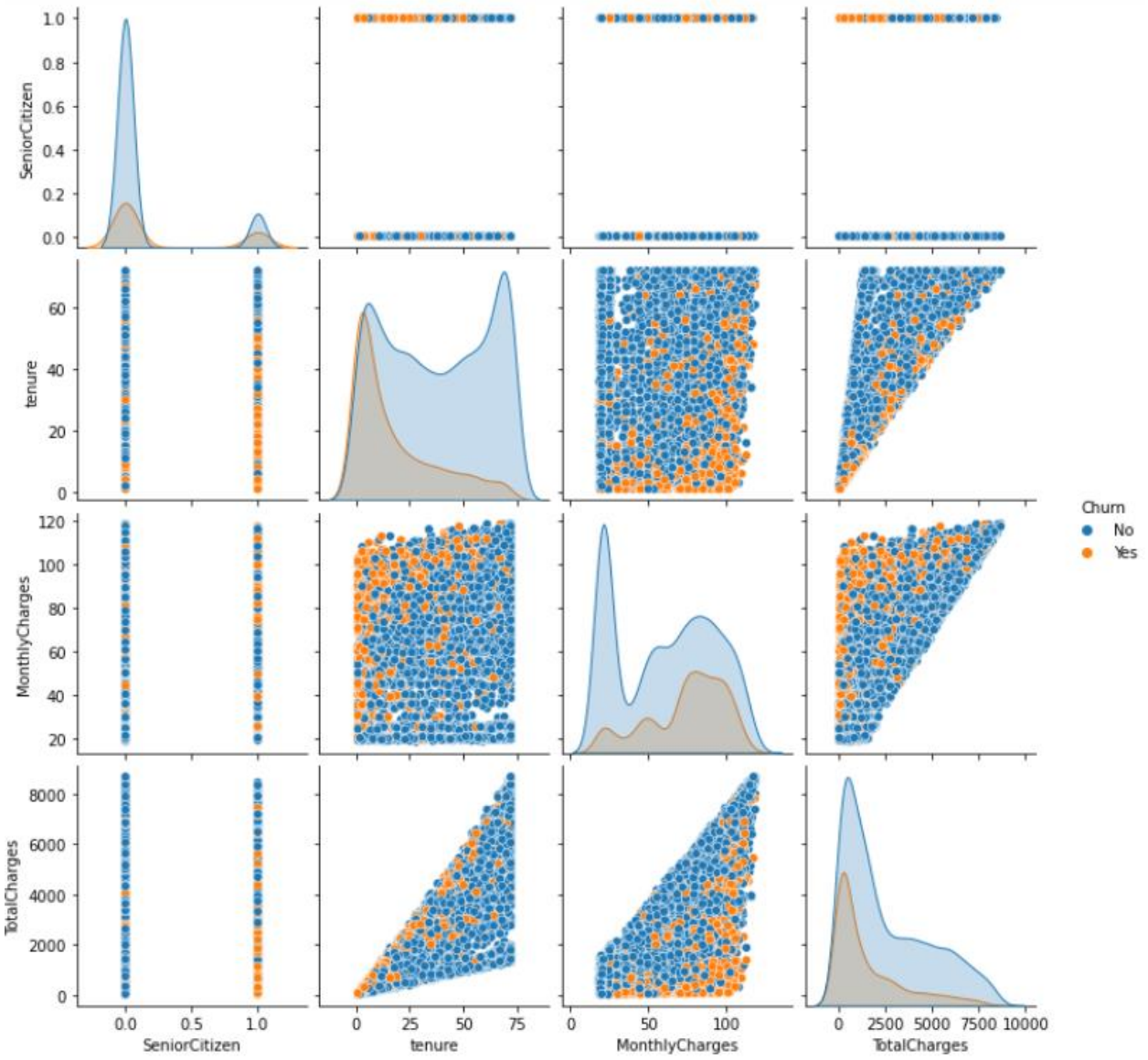
Figura 3.2.2. Churn por género



Con la gráfico podemos ver que es un data set equilibrado en cuestión de género y que no es relevante esta variable para determinar si un cliente abandonará o no.

Por último, realizamos gráficas para determinar la correlación entre las variables numéricas contra la variable de Churn.

Figura 3.2.3. Correlación entre variables



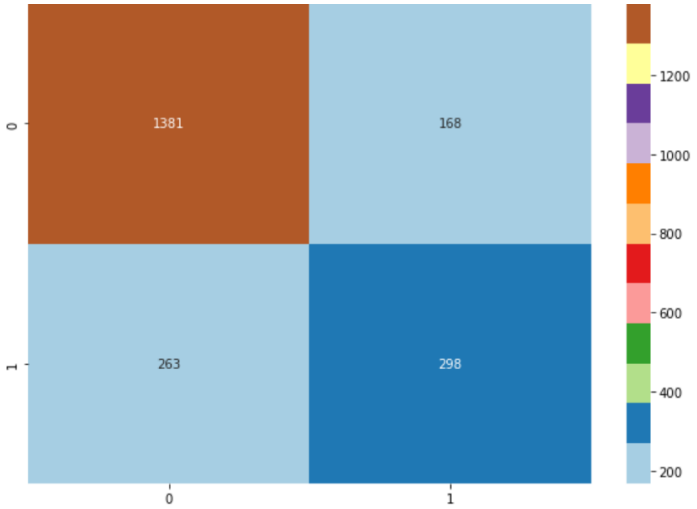
A simple vista, parece que las variables de monto mensual y monto total son las que más afectan al Churn, ya que a mayor sea la cantidad a pagar, más tendencia al Churn parece existir.

3.3 Modelado

En este apartado, se observarán las métricas de valuación de tres modelos diferentes, las características y resultados del modelo son las siguientes:

1) **Regresión Logística con todas las variables:** Se generó un modelo de regresión logística con los datos escalados y las variables categóricas convertidas a variables Dummy, las métricas de valuación son los siguientes:

Figura 3.3.1 Matriz de confusión modelo 1



La matriz de confusión nos muestra que clasifica muy bien a los que no van a realizar Churn, pero los que si abandonan, no los puede clasificar de manera óptima. Las métricas de Recall y F1 nos ayudarán a observar mejor lo antes escrito.

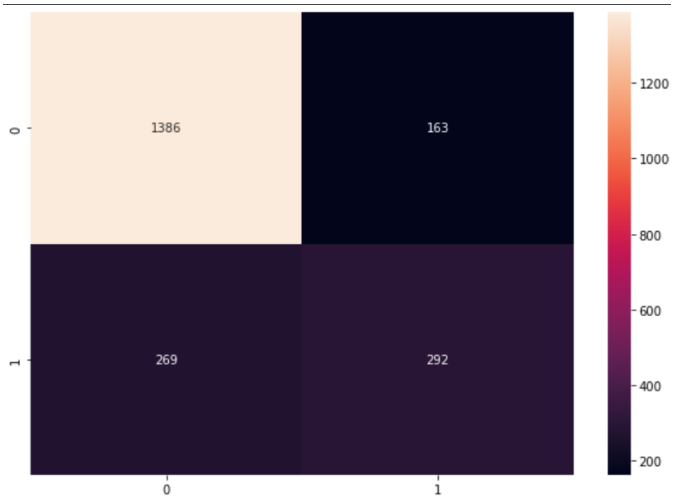
Tabla 3.3.1 Métricas de valuación modelo 1

	precision	recall	f1-score	support
0.0	0.84	0.89	0.87	1549
1.0	0.64	0.53	0.58	561
accuracy			0.80	2110
macro avg	0.74	0.71	0.72	2110
weighted avg	0.79	0.80	0.79	2110

Como bien mencionábamos, el Recall de los que si realizan Churn es del 53%.

2) **Regresión logística con las 11 variables más influyentes:** Se generó el mismo modelo con las 11 variables más significativas para el modelo, los resultados son los siguientes:

Figura 3.3.2 Matriz de confusión modelo 2



Los resultados son muy parecidos al primero modelo, con la matriz de confusión vemos que este modelo puede clasificar muy bien a los que no realizan Churn.
La ventaja de este modelo (a pesar de que es muy parecido al primero) es que al tener menos variables, el poder de procesamiento y los recursos computacionales son menores al primero.
Para observar mejor las métricas, se anexa la siguiente tabla:

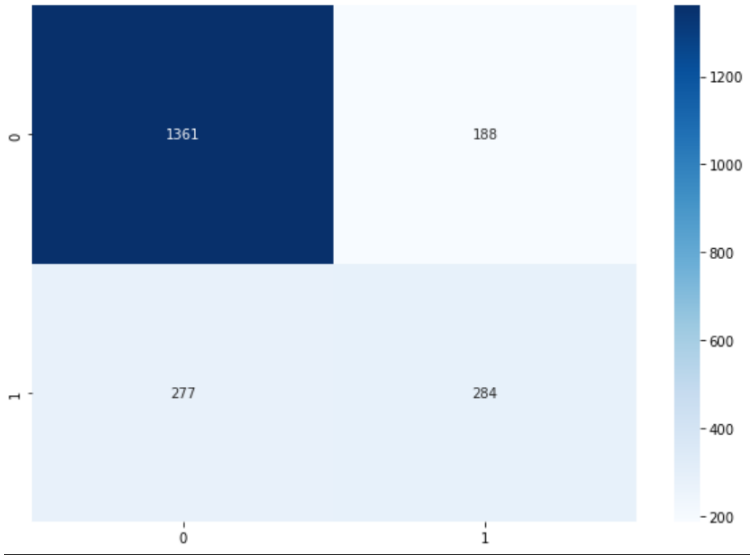
Tabla 3.3.2 Métricas de valuación modelo 2

	precision	recall	f1-score	support
0.0	0.84	0.89	0.87	1549
1.0	0.64	0.52	0.57	561
accuracy			0.80	2110
macro avg	0.74	0.71	0.72	2110
weighted avg	0.79	0.80	0.79	2110

El problema de clasificación de los que si realizan churn no mejoró, el 52% de las predicciones de este tipo fueron exitosas.

3) **XG BOOST:** Se generó un modelo de XG BOOST, el cual se caracteriza por ser un modelo de árboles de decisiones optimizados, esto significa que es más eficiente a la hora de detectar anomalías y en este caso, posibles abandonos, los resultados son los siguientes:

Figura 3.3.3 Matriz de confusión modelo 3



Como podemos ver, el problema de clasificar mejor a los clientes que realizaron Churn no mejoró mucho. Para cerciorarnos, se mostrarán las métricas de valuación:

Tabla 3.3.3 Métricas de valuación modelo 3

	precision	recall	f1-score	support
0.0	0.83	0.88	0.85	1549
1.0	0.60	0.51	0.55	561
accuracy			0.78	2110
macro avg	0.72	0.69	0.70	2110
weighted avg	0.77	0.78	0.77	2110

La métrica de Recall sigue siendo muy baja, por lo que el problema de clasificar correctamente a los clientes que harán Churn no se solucionó.

4.- Conclusiones

El problema de la mala clasificación de los clientes que van a hacer Churn se debe a que el Data Set no es equilibrado, es decir, hay muchos clientes que no han realizado Churn en comparación de los que si, por lo que es complejo para los modelos detectar patrones de los que abandonaron sus servicios de telecomunicaciones. La proporción de los que realizaron abandono con respecto a los que no, es la siguiente:

Etiqueta	Clientes	Proporción
No	5,163	73%
Yes	1,869	27%
TOTAL	**7,032**	**100%**

El 73% de los clientes en el Data Frame no realizaron Churn, por lo que es más fácil para el modelo predecir los que no se fueron.

5.- Anexos

- Repositorio en Git Hub: https://github.com/luis151294/Telecomunicaciones_ClasificacionBinaria
- Data Set en Kaggle: https://www.kaggle.com/datasets/blastchar/telco-customer-churn