Módulo 2 - Práctica 1

Descubrimientos

Uno de los puntos principales de la práctica era el poder seleccionar una variable continua y modelarla haciendo una regresión lineal. En nuestro caso, pudimos ver que sólo tenemos 3 variables continuas:

- Tenure.(antigüedad del cliente en la empresa).
- Monthly Charges.
- Total Charges.

Nosotros elegimos ***MonthlyCharges***, ya que, en un negocio tradicional nos interesarí saber cuánto se le va a cobrar a un cliente mes tras mes, sin importar mucho de sus años de antigüedad.

Al confirmar los data types de las variables continuas y discretas, hubo un problema con ***TotalCharges***,. Dado que nos los contaba como objeto en vez de variable flotante.

Había 11 elementos con espacios como valor, en nuestro caso los eliminamos ya que no era representativo para nuestro data set.

```
In [16]: #There are missing values in the total charges variable, we can delete them as they will no affect our data
         df[df["TotalCharges"] == ' '].shape, int(len(df[df["TotalCharges"] == ' '])/len(df))
Out[16]: ((11, 21), 0)
```

Variables Discretas

Al explorar las variables discretas, nos dimos cuenta que no había valores con frecuencias pequeñas, las categorías “más pequeñas” se encontraban en las variables *PhoneService* y *MultipleLines* en las que representaban cerca de un 9.7%.

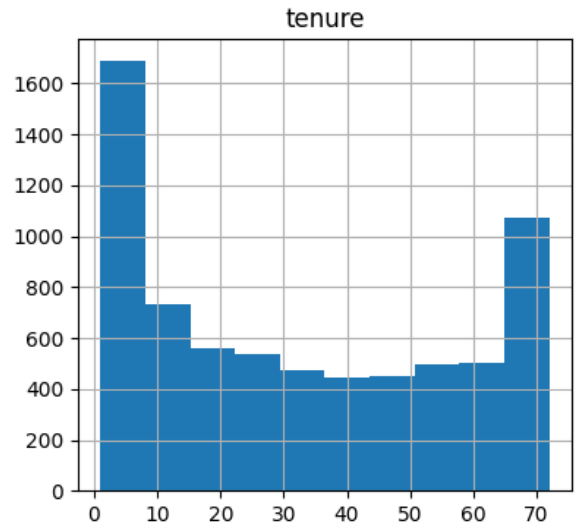
Variables Continuas

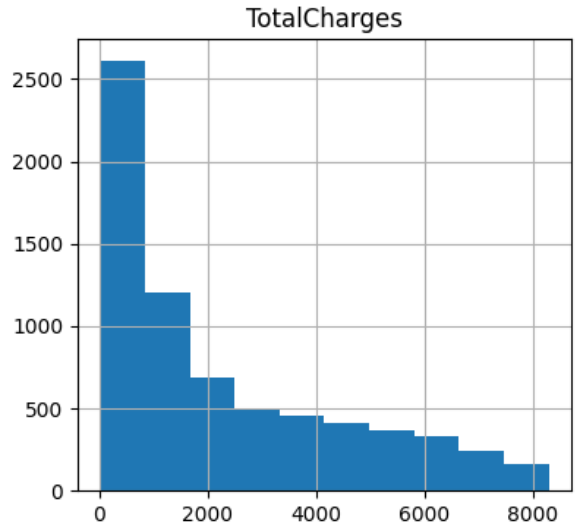
Para las variables Continuas, no encontramos missing values; y, para temas de Outliers, realizamos dos análisis:

- Cerca Intercuartil.
- Cerca Percentil.

En el caso de la cerca percentil no se encontraron valores extremos; pero, si tomamos la cerca percentil con 0.005 y 0.995 de valores referencia, tuvimos 1% de nuestros

datos como outliers, los cuales eliminamos. Nuestras distribuciones quedaron de la siguiente manera:





Al realizar el análisis de correlación, encontramos que estaban relacionadas en un 0.8236, a lo que procedimos al análisis de multicolinealidad para encontrar la mejor variable.

De acuerdo a nuestro análisis con VarCluShi, nos arrojó que la variable ***Tenure*** era la mejor para nuestro procedimiento.

Solución

En esta práctica se nos pidió realizar dos modelos:

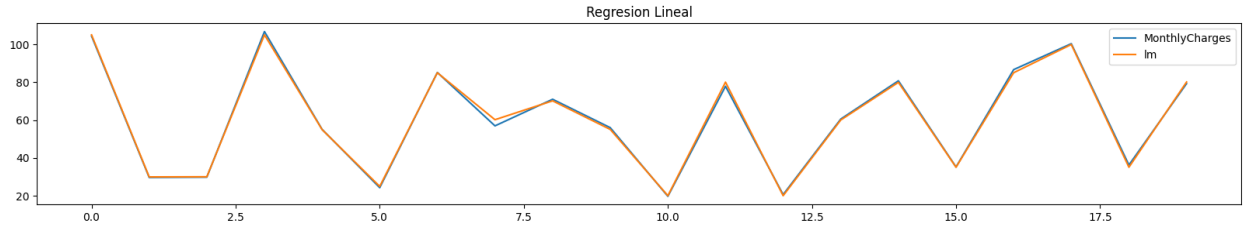
- Modelo de Regresión Lineal.
- Modelo de Clasificación de Churn.

A lo que ocupamos dos diferentes direcciones, utilizar dummies y usar las transformación WoE.

Regresión Lineal

Para poder realizar este modelado, utilizamos el *approach* de las variables dummies con las variables discretas. Al terminar este proceso, nuestra matriz X terminó con 46 variables continuas.

Utilizamos la clase de ***LinearRegression*** para poder observar cómo se comportaban los datos. Obtuvimos un error absoluto de ~0.778 en el 30% de nuestros datos (test split). Al graficar el comportamiento de los valores reales vs valores predichos, terminamos con una gráfica muy parecida.



A lo que paramos este modelado aquí.

Regresión Logística

Para el modelo de clasificación, utilizamos la transformación WoE para las variables discretas. Pero antes, tuvimos que:

- Renombrar nuestras variables continuas, target y quitar *Churn* de las variables discretas.
- Rehacer el análisis bivariado.

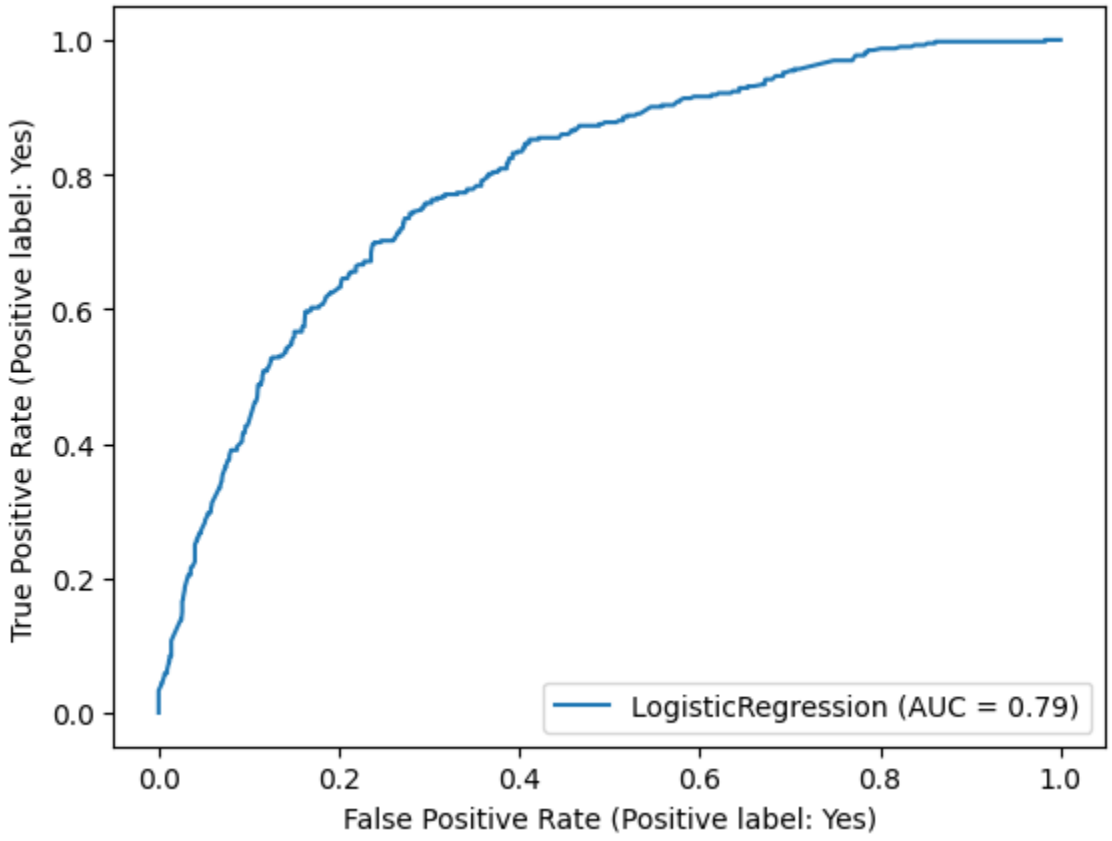
Para poder elegir la mejor variable continua, nuestro análisis de VarCluShi nos arrojó a ***TotalCharges*** como la mejor opción.

Luego, discretizamos nuestra variable continua con hasta 5 bins distintos, a lo que nuestra función IV nos concluyó que la variable discreta con 5 bins era mejor que menos bins.

Al elegir las mejores variables discretas y hacer el mapa WoE, se terminó con 9 variables totales para nuestra TAD.

Al comparar métricas de las variables training vs variables testing, nos dimos cuenta que nuestro modelo generalizaba.

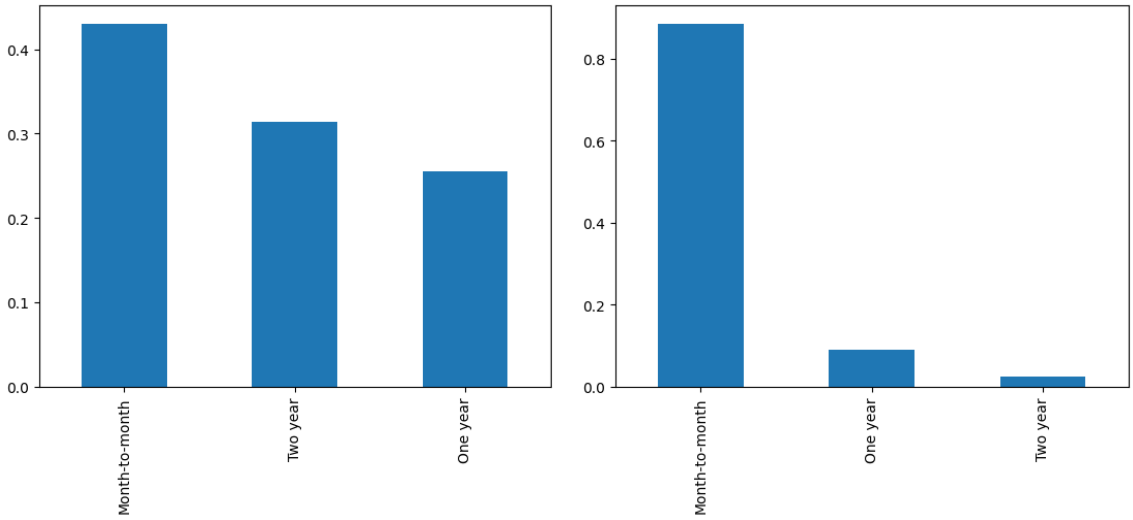
```
In [99]: metrics_Logistic(Log_Reg, Xt, yt)

Roc Validate: 0.803
Acc Validate: 0.780
Matrix Conf Validate:
[[2292  208]
[ 543  367]]

In [100]: metrics_Logistic(Log_Reg, Xv, yv)

Roc Validate: 0.795
Acc Validate: 0.777
Matrix Conf Validate:
[[986  84]
[242 150]]
```



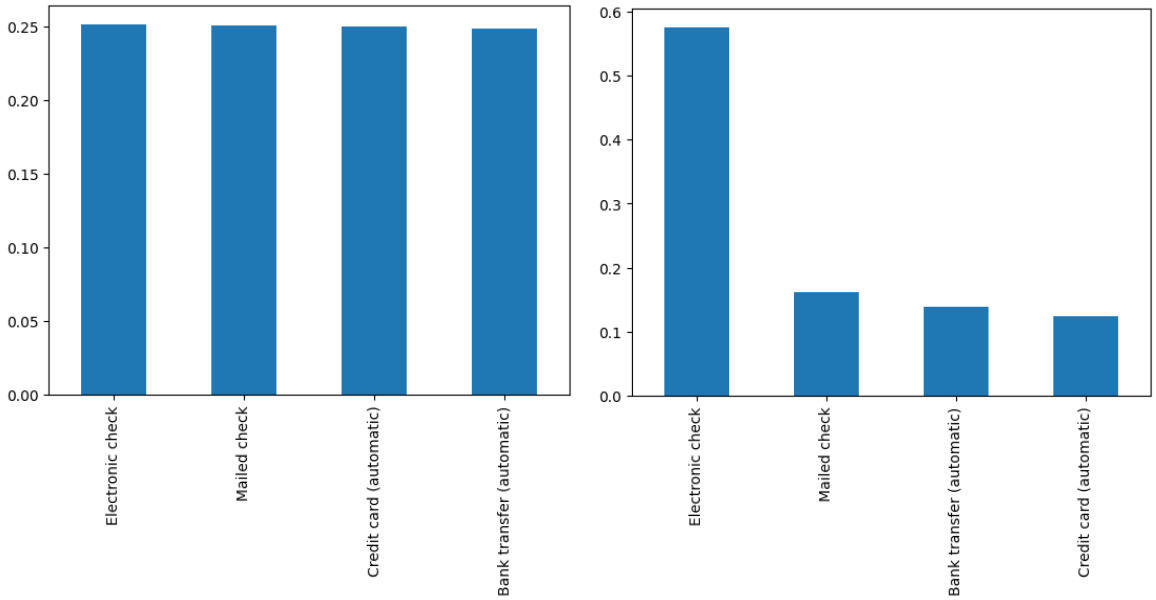
Acciones del negocio

Una vez hecho el modelaje, procedimos a ver qué podía hacer el negocio para bajar la tasa de churn. A lo que procedimos a realiza unas gráficas de usuarios churn vs non-Churned.

- Algo que nos dimos cuenta, es que 80%+ de usuario quemados tienen contratos mes-a-mes; mientras que, en los usuarios no quemados, este tipo de contrato solo representa poco más del 40% en los usuarios, el resto está distribuído de manera muy similar entre contratos de 1 y 2 años (img izq).



- Otro ejemplo similar, es el método de pago de los usuarios, en los usuarios actuales (img izq), el metodo de pago esta igualmente distribuído, mientras que, en los churned users, .el cheque electrónico es dominante con ~60%.



Recomendaciones

- Lo que el negocio puede hacer, es buscar hacer el upselling de los contratos y pasarlos de mes-a-mes a por lo menos, contratos de 1 año.
- Otra cosa a mejorar es la distribución de métodos de pago, buscar el poder mejorar el cobro automático a través de campañas de marketing y/o promociones.

En cuanto a nuestra regresión lineal, el análisis nos arrojó que la mejor variable que ayuda a predecir el cobro mensual es la antigüedad de nuestro cliente. Lo siguiente a mejorar es poder filtrar las variables discretas a través de la WoE, y, en un futuro, buscar obtener más variables continuas para mejorar el modelo.