Reporte modelo predictivo del Churn

Introducción

El presente documento contiene el análisis de datos realizado y las manipulaciones de datos hechas para así implementar un modelo predictivo de regresión logística.

Objetivos

El objetivo de este proyecto es predecir con la mayor precisión posible la variable Churn utilizando la base de datos provista por la empresa Telco NN.

Descripción del dataset

El dataset contiene 7.043 instancias del historial de clientes y 20 dimensiones que sirven para caracterizar a cada cliente.

EDA

Con la finalidad de estudiar la base de datos utilizada se calculó el porcentaje de nulos por cada variable y también el porcentaje de instancias de la tabla que poseen algún nulo. Como resultado se obtuvo que el 86% de las instancias tienen al menos un valor nulo, y la distribución porcentual de nulos para cada variable es la siguiente:

```
Porcentaje de nulos por variable:
PaperlessBilling    16.995598
Contract            16.995598
MonthlyCharges      16.995598
PaymentMethod       16.995598
OnlineBackup        13.005821
Dependents          13.005821
Partner             13.005821
gender              13.005821
tenure              13.005821
PhoneService        13.005821
MultipleLines       13.005821
InternetService     13.005821
OnlineSecurity      13.005821
TotalCharges         0.156183
SeniorCitizen        0.000000
DeviceProtection     0.000000
TechSupport          0.000000
StreamingMovies      0.000000
StreamingTV          0.000000
Churn                0.000000
```

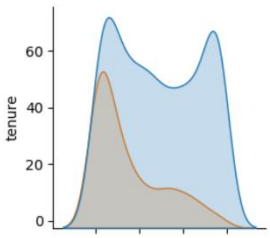
Las variables se pueden clasificar de la siguiente manera:

- **Descripción personal del cliente**:
    - Gender: Genero del cliente
    - seniorCitizen: Si el cliente es jubilado
    - Partner: Si el cliente tiene socio
    - Dependents: Si el cliente tiene personas a cargo
- **Servicios contratados**:
    - PhoneService: Si el cliente tiene un servicio de telefono o no
    - MultipleLines: Si el cliente tiene multiples lineas o no
    - InternetService: Tipo de servicio de internet que recibe. Si es que recibe
    - Otros

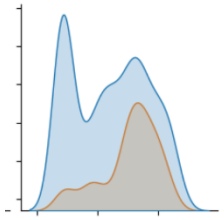
- **Pagos**:
    - Tenure: Antigüedad del cliente
    - monthlyCharges: Costo mensual
    - totalCharges: Cargos totales
    - paymentMethod: Tipo de pago del cliente
    - contract: Tipo de contrato del cliente
- **Churn**: Si el cliente se fue de la compañía o no
- **IDs**: CustomerID

Para poder entender la distribución de las variables en función de nuestra variable churn, se realizaron dos visualizaciones. Para visualizar los valores numéricos se realizó un pairplot coloreado según churn:

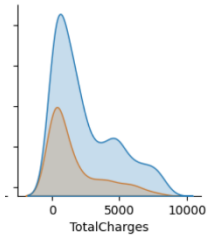
Analizando los histogramas:



Tenure: Podemos visualizar que el churn sucede principalmente en los primeros meses de antigüedad.

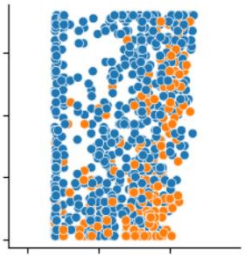


- **Monthly charges:** Podemos visualizar que los clientes perdidos, son principalmente aquellos que se encuentran en el rango de gastos mensuales mas alto. También podemos encontrar que de los clientes activos, hay una gran proporción que se encuentra en el tramo de gastos mensuales bajos.



- **Total charges:** No se encuentra una gran diferencia entre los clientes perdidos y los activos.

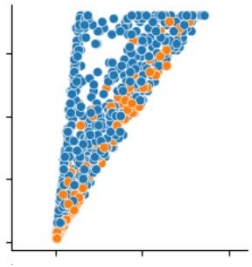
Analizando los scatterplots (Eje Y / Eje X):



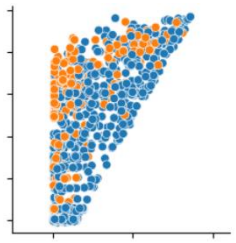
- **Tenure / MonthlyCharges:** En esta visualización podemos encontrar que se suelen perder clientes en los primeros meses de antigüedad, con una mayor tendencia en el tramo de gastos mensuales más alto. A medida que aumenta la antigüedad, esta

tendencia se vuelve más fuerte, perdiéndose casi exclusivamente a clientes con gastos mensuales altos.

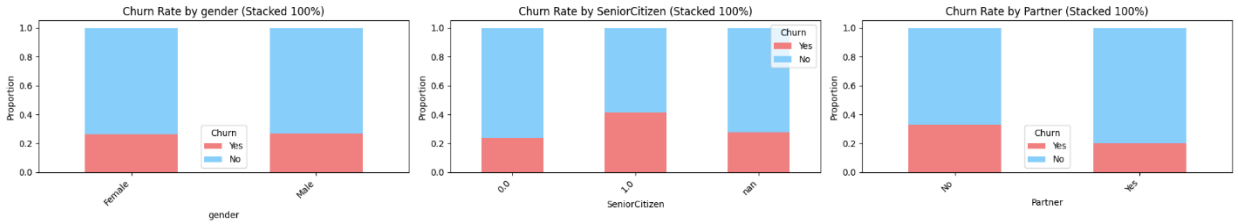
- **<u>Tenure / Total charges:</u>** En este gráfico podemos visualizar una figura triangular. En la parte inferior se encuentran aquellos clientes que tienen la mayor cantidad de gastos totales en el menor tiempo de antigüedad. Esta línea coincide con los clientes que se perdieron. Por otro lado tenemos la línea casi vertical, que indica aquellos clientes que tienen la mayor antigüedad con la menor cantidad de gastos totales. Estos se encuentran casi totalmente activos.



- **Monthly charges / Total chrages:** En este gráfico podemos visualizar la misma figura triangular. En este caso podemos ver que los clientes activos con mayor cantidad de gasto total tienen una mayor tendencia a quedarse.



Para visualizar la distribución de churn en las variables categóricas se realizó una serie de gráficos de barras al 100% en donde cada columna representa a una de las posibles categorías por variable.



Al ver la distribución de churn en cada una de las categorías podemos decir que hay variables mas relevantes que otras para determinar el churn. Por ejemplo, podemos ver que la proporción de churn se mantiene constante según el género del cliente (no es relevante), pero los senior citizens tienen considerablemente mayor churn.

Materiales y métodos

Debido a la alta proporción de instancias con valores nulos (86%), no es viable eliminar las instancias. Es por esto que se optó por un método de simple imputer, utilizando los valores más frecuentes y las medianas para llenar los valores nulos.

Luego, las variables booleanas (Yes/No) se transformaron en números (1/0) para que puedan ser procesados por el modelo predictivo. Las variables categóricas fueron

tranformadas utilizando el onehotencoder. Y por último, las variables numéricas fueron normalizadas utilizando el método de minmaxscaler.

Una vez completado el pipeline de datos, se decidió utilizar el modelo logistic regression [i]como algoritmo para predecir los parámetros. A la hora de entrenar el modelo, se aplicó gridsearch para encontrar los mejores hiperparámetros: C = 100, solver = 'liblinear'.

La principal métrica utilizada para medir los resultados del modelo es el ROC AUC. Utilizamos esta métrica debido a que es una métrica mucho mas consistente estadísticamente y mas discriminante que el Accuracy[ii]. Lo cual es específicamente importante ante conjuntos de datos desequilibrados, como el conjunto utilizado.

El resultado obtenido de ROC AUC es igual a 0.8129, lo cual es muy positivo para un modelo predictivo.

Experimentos y resultados

Con la finalidad de buscar mejorar los resultados del modelo, aplicamos la técnica de feature engineering denominada PCA[iii]. En esta técnica, se busca capturar la mayor cantidad de dispersión dentro del modelo de datos, disminuyendo su dimensionalidad hasta un valor definido.

En este caso, disminuimos la dimensionalidad del modelo procesado que era de 27 dimensiones, a un modelo de 12 dimensiones.

Los resultados obtenidos con esta reducción de dimensionalidad son marginalmente inferiores al modelo previo. Con un resultado de ROC AUC igual a 0.8087 (-0.042).

Conclusiones

Como conclusión podemos decir que el modelo predictivo de logistic regression es un modelo apto para lograr predecir el churn de la compañía Telco NN. Por lo cual será clave para lograr entender y predecir los ingresos a futuro de la empresa.

Referencias

[i] ISLP – Capítulo 4 – Página 144: Aplicación de modelos de regresión logística para predicciones cualitativas binarias.
[ii] AUC: a Statistically Consistent and more Discriminating Measure than Accuracy
[iii] ISLP – Capítulo 6.3.1: Utilización de método de PCA para la reducción de la dimensionalidad de un modelo.