Estudio de la cartera de clientes de la compañía Telco NN mediante Machine Learning

Autor: Rahimi Vilchez, Aiad

Resumen

Este artículo va a estar conformado por dos partes: En primer lugar, un análisis exploratorio sobre distintos datos de una cartera de clientes, informados por la empresa Telco NN, con el fin de detectar quienes podrían dejar la compañía; Posteriormente, el objetivo es aplicar modelos de machine learning de aprendizaje supervisado como lo son…. a fines de predecir la variable CHURN, la cual indica, si el ha abandonado o no, la compañía.

1 INTRODUCCIÓN

El presente informe consiste en un análisis exhaustivo de la cartera de clientes de TELCO NN.

El objetivo es poder predecir cuáles de estos clientes podrían abandonar la compañía en un futuro próximo.

Partimos de un dataset que brinda TELCO NN sobre su propia cartera de clientes, y el cual se irá puliendo a fines de lograr los objetivos previamente mencionados, extrayendo toda información que no aporta valor a nuestro análisis.

Por último, con modelos de aprendizaje supervisado intentaremos predecir los potenciales clientes que abandonarían la empresa en cuestión.

2 DESCRIPCIÓN DEL DATASET

El Dataset utilizado para realizar el análisis es brindada por la empresa que realiza la consultoría.

Originalmente está conformado por 7043 samples y 21 features.

Éstas features son:

- **Customer ID**: Número único que sirve para identificar y enumerar a cada uno de los clientes.
- **Gender**: Señala el género del cliente (Masculino/Femenino).
- **Senior Citizien**: Denota si el cliente es un Senior Citizien o no.
- **Partner**: Indica si el cliente tiene un socio o no.
- **Dependents**: indica si el cliente tiene dependientes o no.
- **Tenure**: Identifica la antigüedad del cliente en la empresa.
- **Phone Service**: señala si el cliente tiene un servicio de teléfono o no.
- **Multiple lines**: Indica si el cliente tiene múltiples líneas o no.
- **Internet Service**: Señala el tipo de internet recibido, en caso de recibirlo.
- **Online Security**: Identifica si el cliente tiene un servicio de seguridad online o no.
- **Online Backup**: Indica si el cliente tiene un servicio de backup o no.
- **Device Protection**: Identifica si el cliente tiene un seguro del dispositivo o no.

- **Tech Support**: Indica si el cliente tiene soporte de tecnología o no.
- **Streaming TV**: Indica si el cliente tiene servicio de streaming para ver TV o no.
- **Streaming movies**: Señala si el cliente tiene un servicio de streaming para ver películas o no.
- **Contract**: Identifica el tipo de contrato del cliente
- **Paperless Billing**: Señala si el cliente recibe la factura en papel o no.
- **Payment Method**: Indica el tipo de pago del cliente.
- **Monthly Charges**: Denota el costo mensual.
- **Total Charges**: Denota los cargos totales.
- **Churn**: Indica si el cliente se fue de la compañía o no (variable a predecir).

3 ANÁLISIS EXPLORATORIO DE DATOS (EDA)

El objetivo de nuestro EDA es mostrar la relación entre ciertas variables del dataset con respecto a la cual queremos predecir, la variable “Churn” que nos indicará la posibilidad de que un cliente abandone la empresa.

Primeramente analizamos con qué tipo de dataset vamos a trabajar, viendo sus dimensiones y definiendo qué variables dejamos fuera del mismo. Procedemos entonces a una limpieza de las variables que no son relevantes para nuestro analisis "Unnamed: 0", "customerID","Contract", "PaperlessBilling","PaymentMethod", "MonthlyCharges", "OnlineBackup", "InternetService".

Con el dataset ya con un primer filtrado, pasamos a la instancia de revisión de NaN´s, y procedemos a eliminar las filas que tengan menos de 12 datos válidos (de las 14 variables que dejamos).

Empezamos entonces con nuestro primer análisis en el que realizamos la matriz de correlación entre las variables pre procesadas.

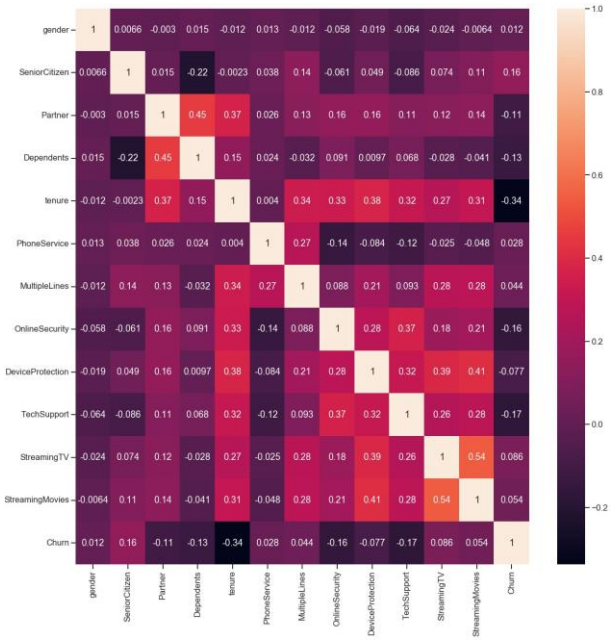

Figura 1

Posteriormente seguimos con el desarrollo de nuestro próximo análisis en el cual queremos entender cómo es la distribución de la variable Churn, que nos indica, si el cliente abandono o no, la compañía. Notamos en los resultados que el 26,5% de la cartera de clientes, son personas que abandonaron la compañía.

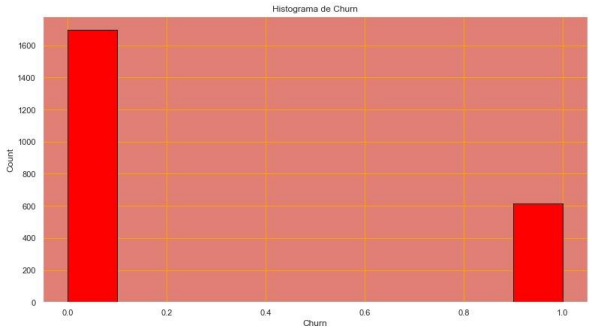

Figura 2

4 MODELO DE APRENDIZAJE

Una vez finalizado el EDA procedemos a aplicar 2 métodos de aprendizaje supervisado para poder predecir si una persona abandonará o no la compañía.

El tamaño de la muestra de entrenamiento definido es de un 25% y establecemos 15 vueltas de entrenamiento

En primera instancia utilizamos el método de Logistic Regression y obtuvimos un Score del modelo de 0.74, y un Accuracy de 0,765.

Luego, utilizamos otro de los métodos conocidos de clasificación, el de Support Vector Machine, el cual nos originó un score de 0,74, y un Accuracy de 0,775

Cabe mencionar que todos los modelos mencionados previamente fueron acompañados también con Grid Search Cross Validation para poder darle al modelo los hiperparámetros que mejor se ajustan a cada uno de los mismos.

En consecuencia, luego de evaluar ambos modelos que dieron igual score en Train, los utilizamos y probamos con el test, y el de mayor Accuracy fue el método de SVM.

	Modelo	accuracy	score
0	SVM	0.775087	0.74
1	Logistic Regression	0.764706	0.74

Figura 3

5 DISCUSIÓN Y CONCLUSIONES

Luego de los distintos análisis llevados a cabo con los modelos de aprendizaje podemos llegar a una conclusión:

Muchas de las muestras fueron eliminadas en la limpieza por poseer datos nulos, esto produjo una disminución del número de muestras. Agregado a esto, las variables poco correlacionadas no se tuvieron en cuenta para el modelo de machine learning.

El mismo aprendió de 4 variables que se consideraron las más importantes. A pesar de todo, el modelo obtuvo casi un 80% de probabilidad de acierto a la hora de predecir. Se necesitan nuevos datos de aprendizaje con menor cantidad de nulos.

A modo de cierre, podemos decir que el dataset utilizado no es el más óptimo para los modelos de predicción que establecimos o que conceptualmente no seleccionamos las variables indicadas dentro del mismo para este análisis.

6 REFERENCIAS

1. Pattern Recognition and Machine Learning - Christopher Bishop http://users.isr.ist.utl.pt/~wurmd/Livros/sc hool/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%2 0Springer%20%202006.pdf

2 .http://www.Stackoverflow.com

3. https://github.com/clusterai

4 .https://scikit-learn.org/stable/