Análisis de variables y predicción de continuidad de clientes en compañía de telecomunicaciones

Carolina Abritta & Simon Sittner

INTRODUCCION Y OBJETIVOS

En el presente informe se trabajará con un dataset proporcionado por la empresa Telco NN de telecomunicaciones, dentro de cual se muestran algunas características de su cartera de su catera de clientes.

El objetivo consiste resolver el pedido hecho por La Telco NN para predecir que clientes dejarán la compañía. Para lograrlo, se deberán aplicar conceptos del ámbito de la Ciencia de Datos y Aprendizaje Estadístico, utilizando Python para procesar los algoritmos y a través de Jupyter Notebooks, desarrollaremos un pipeline de Machine Learning para predice cuando los clientes se van o no de la empresa.

DESCRIPCION DEL DATASET

La empresa dispone de una base de datos que contiene información sobre 7043 clientes que han pagado por servicios en los últimos meses. Cada una de las 7043 muestra (samples) tiene 21 categorías (features), sólo 3 de ellas son variables numéricas (tenure, MonthlyCharge y TotalCharge). EL resto corresponde a variables categóricas, correspondientes a características de la persona y de los servicios contratados y por último una variable identificatoria. A continuación, se muestra el diccionario de variables (en color la variable a predecir).

Diccionario dataset del Telco churn		
Variable	Significado	Dtype
Customer ID	Valor identificador de clientes	object
gender	Género del cliente	object
SeniorCitizien	Si el cliente es un SeniorCitizien o no	float64
Partner	Si el cliente tiene un socio o no	object
Dependents	Si el cliente tiene dependientes o no	object
tenure	Antigüedad del cliente	float64
PhoneService	Si el cliente tiene un servicio de telefono o no	object
MultipleLines	Si el cliente tiene multiples lineas o no	object
InternetService	Tipo de servicio de internet que recibe. Si es que recibe	object
OnlineSecurity	Si el cliente tiene un servicio de seguridad online o no	object
OnlineBackup	Si el cliente tiene un servicio de backup o no.	object
DeviceProtection	Si el cliente tiene un seguro del dispositivo o no	object
TechSupport	Si el cliente tiene soporte de tecnología o no.	object
StreamingTV	Si el cliente tiene servicio de streaming o no	object
StreamingMovies	Si el cliente tiene servicios de streaming de peliculas o no	object
Contract	Tipo de contrato del cliente	object
PaperlessBilling	Si el cliente recibe la factura en papel o no.	object
PaymentMethod	Tipo de pago del cliente	object
MonthlyCharges	Costo mensual	float64
TotalCharges	Cargos totales	object
Churn	Si el cliente se fue de la compañía o no	object

↖ Variable a predecir

La variable ‘Churn’, (a predecir), esta previamente etiquetada, por lo que infiere que el tipo de aprendizaje estadístico a implementar será del tipo supervisado.

Previo a la realización del Análisis Exploratorio de Datos (EDA), se realizó un preprocesamiento de los datos ya que el dataset contaba con una gran cantidad de nulos (NaN), es decir valores vacíos no computables que deben ser tratados en el conjunto de datos antes de trabajar con ellos.

En el primer intento, se borraron todas las filas que contenían nulos, pero nos encontramos con el problema de la pérdida de más del 90% del dataset, ya que los nulos estaban repartidos a lo largo de diferentes columnas. Así pues, se decidió completar la mayor cantidad de datos con diferentes métodos.

Resumen preprocesamiento

- Se indica que ‘Unamed: 0’ se debe utilizar como índice, aplicando index_col=0 al importar el dataset.
- Se modifica el tipo de variable “TotalCharges” ya que aparece inicialmente como objeto, con esta acción apareciron 11 nuevos nulos.
- Se elimina la columna "customerID" del conjunto de datos.
- Se rellenan los registros nulos de las variables "TotalCharges", "tenure" y "MonthlyCharges" en base a su relación ⍰ tenure=TotalCharges/MonthlyCost.
- Del análisis surge que no se espera un poder predictivo significativo la variable género se llena con el metodo 'filla', además que las variables ‘SeniorCitiezen', 'Dependents' tienen ciertos valores muy predominantes por lo que se llena también con el método 'filla' pero aplicano la moda (.mode()).
- La variable ‘Contract', que indica la duración de contrato, se relaciona con ‘TotalCharge’, se evalúa las apariciones en Bandas de precio y se le aplican los más frecuentes
- Según los valores únicos vistos, en 'MultipleLines' se indica 'No phone service', cuando no se tiene servicio, se procede a colocar 'No' a la columna de correspondiente de 'PhoneService'. Además, como la persona que contrata el servicio de telefonía puede o no tener multiples lineas, se procede a rellenar en un principio con "yes" en "PhoneService" si tiene "MultipleLines".

'MultipleLines'= "yes"→ 'PhoneService'= "yes".
'MultipleLines'= "No Phone service" → 'PhoneService'= "No"
'PhoneService'= "No" → 'MultipleLines'= "No"

- Según los valores únicos vistos, en los servicios extras (OnlineSecurity, OnlineBackup, etc.) se indica 'No internet service', se procede a colocar 'No' a la columna 'InternetService'.

‘XXX’ (servicio extra) = 'No internet service" → 'InternetService'= "No"

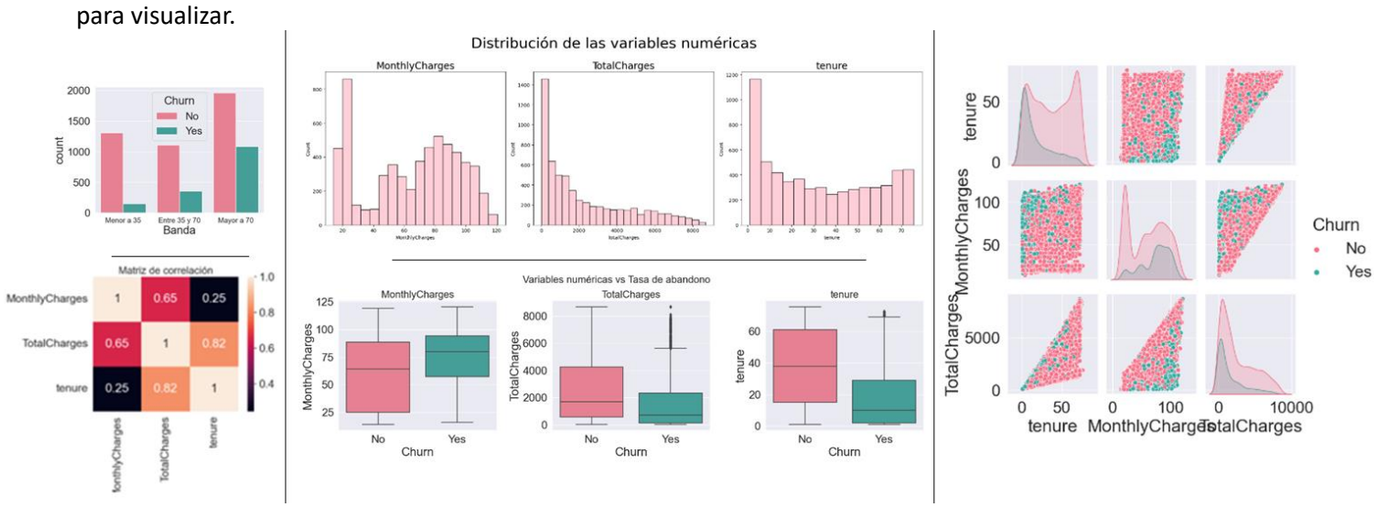
- Luego se llenan nulos de manera aleatoria con base en su distribución de las variables ('Partner', 'PhoneService', 'PaperlessBilling', 'PaymentMethod', 'InternetService', 'PhoneService', 'OnlineSecurity', 'OnlineBackup'), se vuelve a imputar la condición en 'MultipleLines' y se llenan con aleatorio.
- Se eliminan los últimos nulos que quedaron (157), perdiendo as la menor cantidad posible.

ANALISIS EXPLORATORIO DE DATOS (EDA)

Se observa la tasa de abandono de la empresa en forma numérica y grafica. Y luego se realiza el análisis de los dos grupos de variables identificadas (numéricas y categóricas por separado). En busca de alguna relación entre estas y con la variable a predecir ‘Churn’.

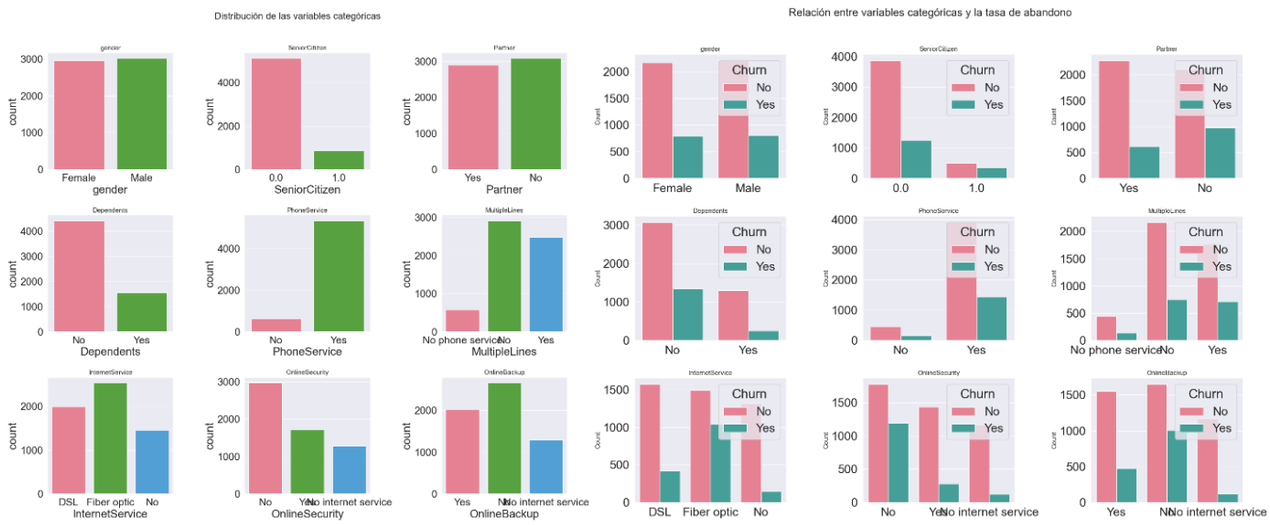
Variables Numéricas

Generamos histogramas de las variables numéricas y una matriz de correlación, un pairplot y boxplots en base a ‘Churn’. Se hizo foco en las variables tenure y MonthlyCharges, generando bandas de precios y un Countplot para visualizar.



Variables Categóricas

Generamos Countplot, para visualizar la distribución de dichas variables por si solas y otro con la tasa de abandono. Algunos ejemplos de las visualizaciones:



CONCLUSIONES DEL EDA

- Se observó que la tasa de abandono en el conjunto de datos es de aproximadamente el 26,5%. Además, se encontró que la duración media del contrato es de 32 meses, y que la mayoría de los clientes tienen facturación electrónica y servicios de telefonía fija e Internet.
- El análisis permitió encontrar la relación entre Total Charges y Tenure, dicha relacion fue utilizada para completar nulos.
- Encontramos outliers en Churn según los meses y MonthlyCharge según Seniorcitizen.

- Se observa una relación entre los clientes que abandonan con el precio mensual (a menor precio menor probabilidad de abandono) y con La antigüedad del cliente (a mayor antigüedad menor probabilidad de abandono).
- El tipo de contrato resulta una de las que más relación con el abandono de clientes tiene. A mayor plazo de contrato se reduce drásticamente la tasa de abandono.
- Se observó que la mayoría de los clientes que abandonan la compañía tienen servicios de telefonía y múltiples líneas contratadas. Además, se encontró una relación entre el tipo de servicio y la tasa de abandono, siendo la fibra óptica, la que posee mayor cantidad de abandono.
- Por último, el tipo de pago con cheque electrónico está relacionado a la mayor tasa de abandono.

METODOS Y MATERIALES
(algoritmos)

Para la resolución del problema utilizaremos dos modelos de aprendizaje supervisado, como hemos mencionado, la variable a predecir esta etiquetada ('Yes' y 'No'), al ser un resultado binario serán utilizados, el modelo Support Vector Machine y una Logistic Regression. Ambos, buscan una función $f(x) = y$, que explica la relación entre 'x' (input) e 'y' (output). Como no se conoce la verdadera $f(x)$, se intenta aproximarla, es en esta aproximación que surge el error, (diferencia entre la "y" dada y la aprendida).

Este error se mide por la función $\boldsymbol{L(y, y')}$ de Costo o Pérdida (Loss function) que tomará valores altos cuando y sea muy distinto de y'. En otras palabras, el aprendizaje supervisado puede plantearse como un problema de optimización llamado Empirical Risk Minimization.

Los modelos planteados serán evaluados por el Accuracy junto con una confussion matrix, que permitirá ver si el modelo está orientado a un tipo de error (Falsos positivos o negativos).

Luego, evaluaremos los mismos modelos, después de aplicar algún método de la reducción de la dimensionalidad, en este sentido se aplicará PCA (Principal Component Analysis), este método busca reducir la dimensión de las variables numéricas sin perder la variabilidad intrínseca de los datos con el objetivo de optimizar el resultado y performance de nuestro modelo.

Modelo Support Vector Machine

En el modelo Support Vector Machine se busca el hiperplano separador que maximice el margen entre clases (equivalente a minimizar la norma al cuadrado de los pesos w sujeto a restricciones en la función lineal). Dicho margen separador queda definido por S muestras, llamadas support vectors, dándole el nombre al modelo.

Modelo Support Logistic Regression

Logistic regression es un clasificador lineal, posee una función de activación tipo 'sigmoid' que genera que el output sea binario y no continuo como una regresión normal. A cada muestra clasificada, le asigna una probabilidad de pertenecer a cada clase existente. Si la probabilidad es mayor a cierto threshold (0.5) entonces pertenece a una clase y viceversa.

El regresor logístico debe aprender un parámetro interno por cada dimensión del vector de entrada (vector W). Para eso calculará el gradiente del error de clasificación y tratará de minimizarlo.

El modelo aprende los pesos 'W' que minimicen el error de nuestro modelo. El error de estará dado por la Función 'Cross Entropy' que penaliza el modelo al momento de falla una predicción y no penaliza cuando la misma es acertada.

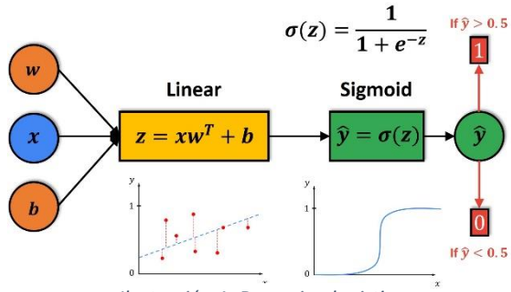


Ilustración 1: Regresion logistica

Para evaluar el modelo, utilizaremos la matriz de confusión. En cada posición se cuentan los TruePositive, TrueNegative, FalsePositive, FalseNegative. Luego se obtiene el coeficiente de Accuracy, que representa que tan bien clasifico nuestro modelo de forma general.

observación \ predicción	+	-
+	TruePositive TP	FalseNegative FN
-	FalsePositive FP	TrueNegative TN

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{total}$$

Método PCA

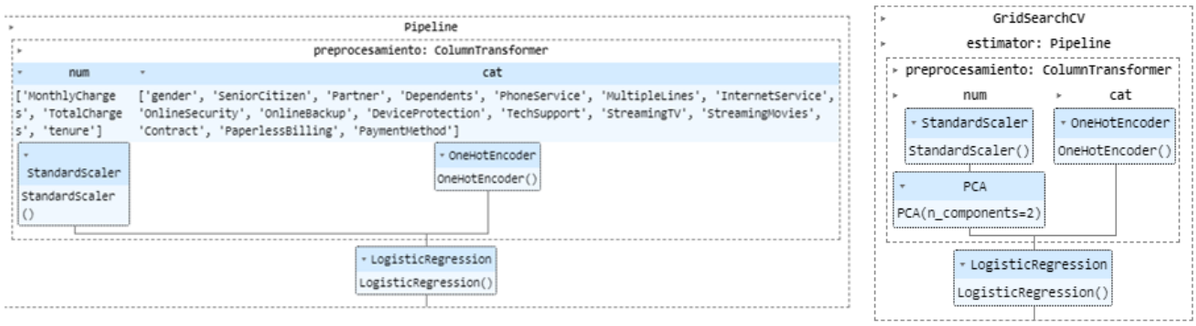
Se suma al Pipeline, por PCA transformaremos nuestra matriz de datos Xnd a una matriz de datos Znp de menor dimensión de las variables numéricas. Crea nuevas features llamadas "Componentes Principales" mediante la descomposición espectral (autovalores y auto vectores). Lo que busca el PCA es proyectar en una dimensión menor nuestros datos tratando de retener la mayor cantidad de información (variabilidad) posible, es decir determinar las variables que más variabilidad explican de mi modelo.

EXPERIMENTO Y RESULTADOS

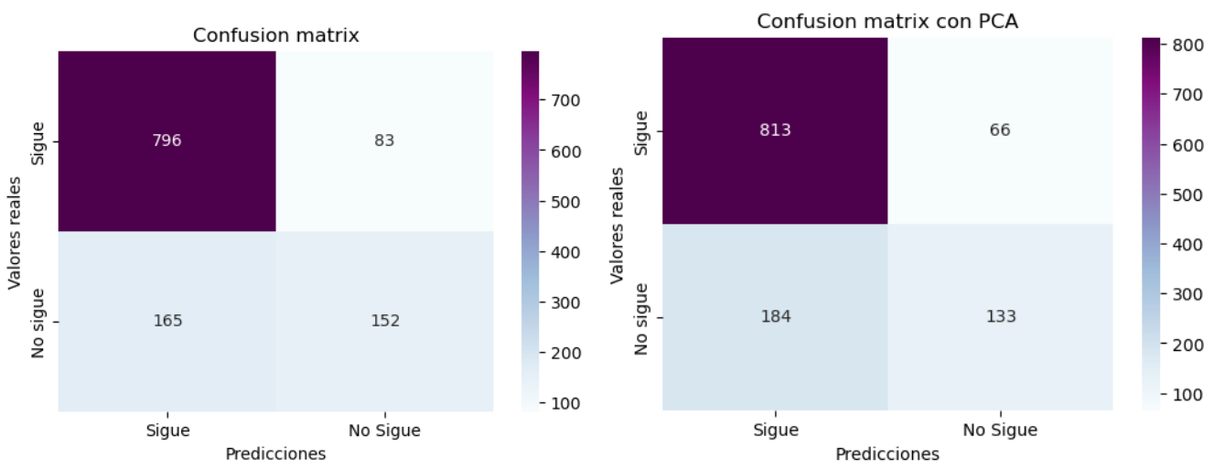
Para poder obtener los mejores hiperparámetros se realizó un GridSearch - CrossValidation .Este método combina los hiperparametros de un modelo, luego realiza Cross Validation y obtiene el promedio del error, compara todos los modelos y se queda con aquel que haya tenido mejor performance y recalcula los parámetros ahora con todos los datos disponibles de entrenamiento.

- El **Cross validation** (CV) se realiza con las muestras de entrenamiento. Consiste en dividir nuestro training set en K 'folds' e iterar K veces. En cada iteración, una porción se utiliza como validación independiente y el resto como train.
- En el Grid-serch, para saber qué hiperparámetros seleccionar, se genera una lista de estos y se prueba todas las combinaciones posibles de ellos.

Los datos numéricos fueron previamente estandarizados, para evitar diferencias entre los rangos y también por ser requisito necesario del PCA. Por último, para los datos categóricos se generaron 'dummies' mediante OneHotEncoder. El Pipeline de nuestro modelo final quedaría entonces de la siguiente manera:



El modelo de LogisticRegression es el estimador inicial, pero luego cambia en las iteraciones del GSCV. Como primer resultado obtuvimos la Confussion Matrix de la izquierda, el modelo seleccionado fue Logistic Regression, con un Accuracy del 0.7926. Cuando se realizó la optimización con PCA como resultados obtuvimos la Confussion Matrix de la derecha con un Accuracy ahora del 0.7909.



Podemos ver una distribución similar de los valores en los diversos cuadrantes, Teniendo un valor final de Accuracy del 79,10%. Resultando este valor ligeramente inferior al original, lo que indica un menor grado de precisión en la clasificación, comparando las matrices podemos ver que el clasificador con PCA arroja una mayor cantidad de falsos positivos, pero una menor de falsos negativos.

CONCLUSIONES

El objetivo que persigue el análisis es por un lado la predicción de la continuidad o no de los clientes en la compañía, pero a su vez nos permite conocer la influencia de las distintas variables conocidas en este resultado, pudiendo de esta manera trabajar sobre ellas para maximizar la permanencia de los clientes. Desde este lado es posible dividir las conclusiones en dos partes, primeramente, la detección de variables relevantes, por otro lado, el análisis de efectividad de los modelos predictivos. Del EDA surge la evidente importancia de los cargos mensuales cobrados a los clientes, como podría haber sido esperable, pero adicionalmente podemos mencionar otras como la antigüedad, permitiendo por ejemplo realizar campañas orientadas a la retención temprana de clientes siendo este periodo vital; otra variable de importancia resultó la duración del contrato. Disponer de esta información permite poder orientar políticas empresariales de manera más efectiva.

Por otro lado, se llegó a un modelo que permite la predicción del abandono o no de los clientes con una exactitud de prácticamente el 79%, acertando casi en 4 de cada 5 casos analizados. Permitiendo de esta manera poder realizar estimaciones de cartera de clientes en el futuro, información altamente relevante para estimaciones de resultados de la compañía o flujos de caja.

BIBLIOGRAFIA

- ILUSTRACION1: https://datahacker.rs/004-machine-learning-logistic-regression-model/
- Introduction to Statistical Learning- Gareth James, Daniela Witten, Trevor Hastie,Robert Tibshirani
- Scikit-learn – https://scikit-learn.org/stable/
- Deep Learning Book – Part 1