Predicción de continuidad y abandono de clientes – Telco

José Ignacio Jonte & Sol Mazzeo

Tomando como base de datos el archivo 'telco_churn_clusterai.csv' que contiene información sobre los clientes de una empresa de telefonía, se buscara predecir mediante un modelo de aprendizaje supervisado si el cliente abandonara el servicio o no el servicio.

Glosario:

Churn(Y/N): Expresa si el cliente dejo o no la compañía.

Tenure: Antigüedad en meses del cliente.

Introducción y objetivos:

En el siguiente informe, desarrollaremos el pedido hecho por La Telco NN para predecir que clientes dejarán la compañía. Para ello nos presentaron un dataset de una cartera de clientes que muestran algunas características de los clientes en la empresa.

El objetivo del informe es resolver dicha predicción a través de la herramienta Python. Desarrollaremos un pipeline de Machine Learning para predecir la a futuro si un cliente dejara o no la compañía.

Descripción del Dataset

La empresa posee una base de datos con información de 7043 clientes que abonaron el servicio durante los últimos meses. Cada uno de estos samples posee 21 categorías o features, de las cuales solo 3 de ellas son variables numéricas (tenure, MonthlyCharge y TotalCharge). El esto de las variables corresponden a particularidades del cliente y el servicio que contrato.

Esta información esta previamente etiquetada en la variable de decisión 'churn', por lo que el tipo de aprendizaje estadístico que será es supervisado. Buscaremos entonces predecir la variable 'churn', es decir si un cliente dejara o no la compañía.

Antes de realizar en EDA, se realizo un preprocesamiento de los datos para poder completar la mayor cantidad de 'nulls' presentes en el dataset. Borrar todos estos datos nulos no era una opción ya que estaban repartidos en diferentes columnas lo que hacia que al intentar eliminarlos quede un dataset únicamente de 800 filas.

Para poder completar los datos categóricos faltantes se utilizó el método de '*ffil()*' y '*bfill()*' ya que esta información no tenía una correlación marcada entre variables.

El tipo de contrato observamos que estaba relacionado con la variable numérica 'Total Charge' por lo que utilizamos esta dependencia para llenar los valores faltantes. Los 'TotalCharge' <1500 eran mayoritariamente un tipo de contrato 'Month to Month' mientras que aquellos valores >3000 correspondían a un contrato 'Two Years'.

En las variables numéricas si se encuentra una relación lineal entre ellas que pudimos ver gracias al EDA. Esta relación es la siguiente:

$$tenure = TotalCharges/MonthlyCost$$

Relacionando las 3 variables pudimos completar entonces los 'nulls' faltantes de las columnas numéricas.

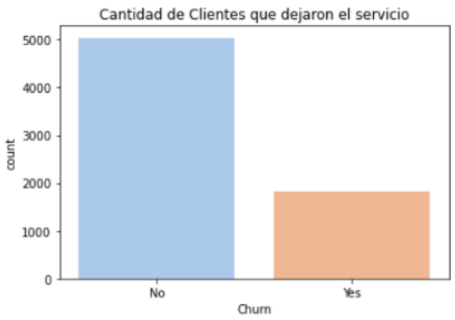
Por último, la variable TotalCharges tenia 11 valores vacíos (no nullos) los cuales fueron identificados por el tipo de dato de la columna. Estos 11 valores fueron pasados a null para ser procesados.

Luego de este preprocesamiento, el dataset que utilizaremos quedo con una cantidad total de samples de 6872 y se mantuvieron las 21 features originales.

Análisis Exploratorio de Datos (EDA)

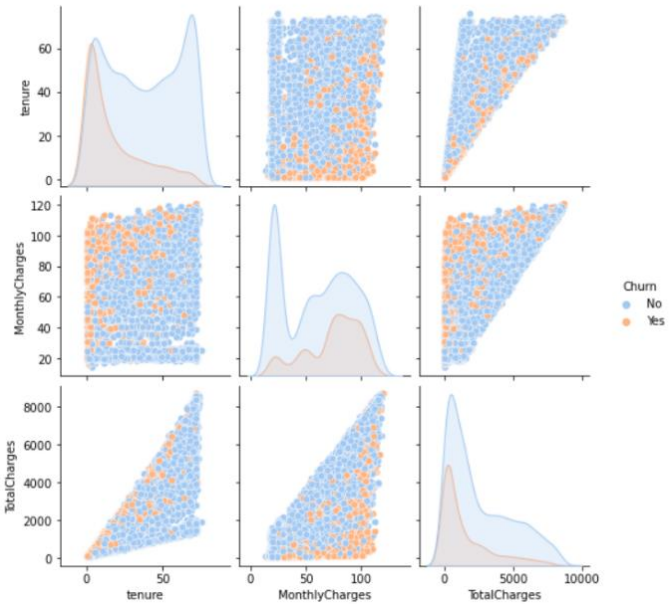
Buscamos primeramente analizar las variables numéricas con el objetivo de entender si hay relación entre las mismas y como se relacionan con el 'churn'.

Resulta fundamental para esto entender primeramente que magnitud de clientes es la que abandona la empresa. Para eso se realizo un grafico de barras o 'count plot' que nos permite entender fácilmente la relación entre ambas variables.

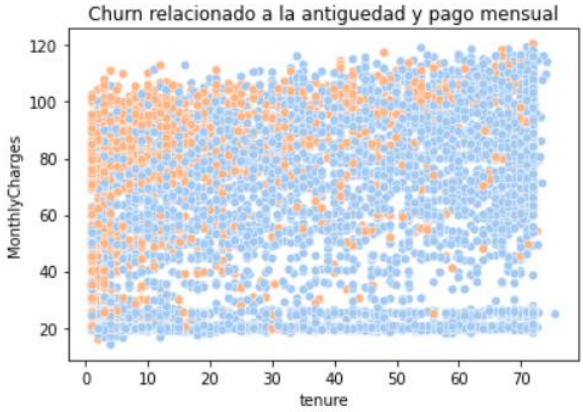


Luego mediante una visualización tipo scaterplott[1], podemos entender cómo se relaciona cada una de las variables numéricas agrupando (‘hue) las variables en el grupo ‘churn’.

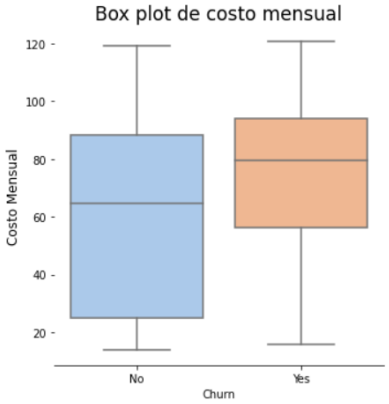
A partir de este grafico pudimos observar la relación entre las variables numéricas que nos permitió completar la mayoría de los valores nulos presentes.



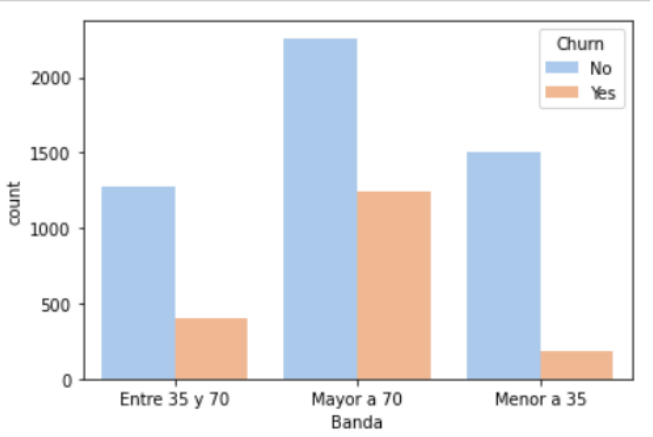
De este grafico también se desprende otra conclusión importante sobre el comportamiento de los usuarios. La mayor concentración de ‘churn’ esta en aquellos usuarios con pagos mensuales altos y bajo tenure.



Podemos analizar también los valores de ‘MonthlyCharge’ por separado y encontrar esta misma relación independientemente del tenure. Para visualizar rápidamente esto (es decir, en 1 dimensión) utilizamos un boxplot divido en cuartiles. En este observamos que el valor medio en los clientes que abandonaron el servicio es mayor y se concentran en un rango de precio mas acotado que el resto de los clientes.

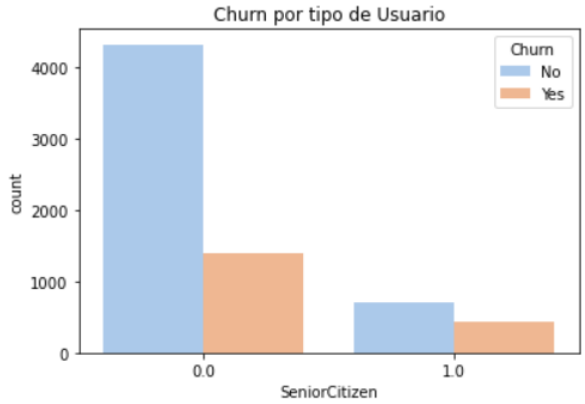


Para poder entender aun mas el impacto del ‘MonthlyCharge’ en el ‘churn’ de los clientes se establecieron 3 rangos de precios en donde se ve claramente que el ratio de abandono aumenta considerablemente en las suscripciones de mayor costo, así como también se reduce en las de menor.



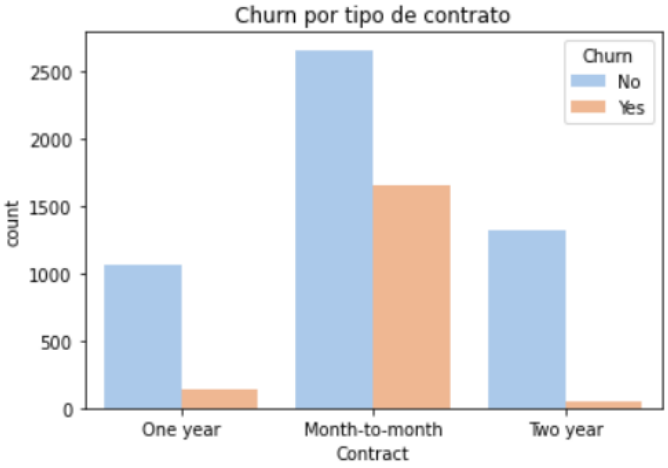
Centrándonos ahora en el análisis de variables categóricas, vemos que los tipos de clientes ‘Senior Citizen’ tienen un ratio mas elevado de abandono.

[1] https://seaborn.pydata.org/generated/seaborn.scatterplot.html



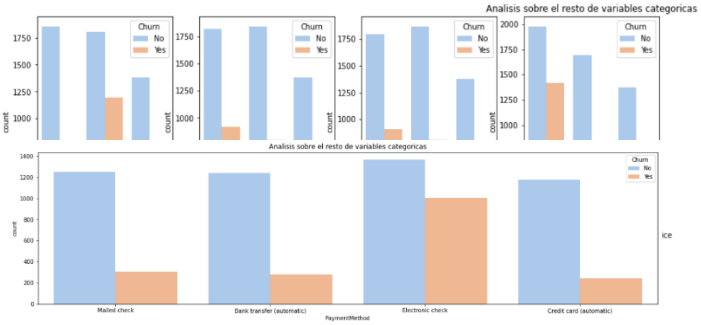
Esta ratio más elevando probablemente relacionado a que poseen un costo mensual medio más alto que los clientes 'comunes'.

Otra de las categorías fundamentales a analizar es el tipo de contrato. La pregunta disparadora de este análisis fue entender si los clientes se fidelizaban con el servicio a mayor tiempo de contrato.



El grafico de barras muestra que, hay una fuerte tendencia a continuar con el pago del servicio en aquellos clientes que lo contrataron por uno o dos años, mientras que el mayor ratio de abandono esta en los tipos de clientes que abonan todos los meses. De este análisis podría surgir por ejemplo, algún tipo de estrategia comercial para incentivar las suscripciones anuales por sobre las mensuales.

Para finalizar con el análisis de las variables categóricas se realizó un análisis similar al anterior sobre el resto de las variables categóricas.



Se observa que, los clientes con pago tipo 'Electronic Check' y aquellos que tienen un tipo de internet de 'Fibra Optica' están fuertemente relacionados con el abandono del cliente. Esto podría servir también para entender la relación del cliente con nuestros servicios y entender su nivel de satisfacción con los productos de la empresa.

De este EDA entonces se desprender algunas conclusiones importantes sobre nuestras distribución de datos:

- Nos permitió encontrar la relación entre Total Charges y Tenure para poder completar los valores nulls de Monthly Charge

- Se observa una relación entre los clientes que abandonan la suscripción y
- El precio mensual: A menor precio menor probabilidad de abandono
- La antigüedad del cliente: A mayor antigüedad menor probabilidad abandono.

- Dentro de las variables categóricas 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.

- El tipo de internet service con fibra óptica resulta el servicio con mayor % de abandono

- Por último, el tipo de pago con cheque electrónico está relacionado a la mayor tasa de abandono.

Materiales y métodos (algoritmos utilizados)

Para la resolución de este problema utilizaremos dos modelos de aprendizaje supervisado. Por un lado, el modelo Support Vector Machine y por el otro una Logistic Regression (ya que el resultado buscado es binario, Y/N). Lo que buscaremos en ambos casos es una función $f(x)=y$ que minimice el error del problema.

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

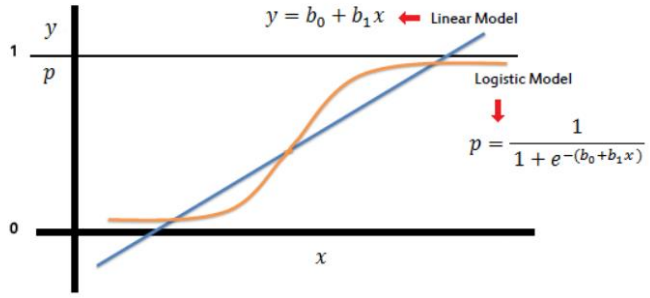
Luego, evaluaremos los mismos modelos luego de realizar un Principal Component Analysis (PCA), en donde buscamos reducir la dimensión de las variables numéricas sin perder la variabilidad intrínseca de nuestros datos con el objetivo de optimizar el resultado y performance de nuestro modelo.

En el SVM buscaremos el hiperplano separador que maximice la distancia entre nuestras clases, es decir que las muestras más cercanas al hiperplano determinaran el posicionamiento del mismo. Estas distancias más próximas al hiperplano son las que reciben el nombre de 'vectores soporte', dando nombre al modelo.

Buscaremos entonces Maximizar la distancia ‘M’ (entre puntos e hiperplano separador) aprendiendo pesos o parámetros ‘w’ de nuestros datos

$$yi(w0 + w1x1 + w2x2 + \cdots) \geq M$$

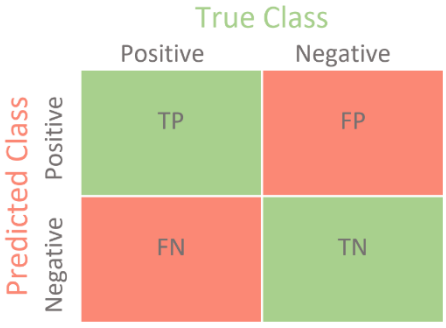
Por otro lado, el modelo de Logistic Regression es un clasificador lineal, que posee una función de activación tipo ‘sigmoide’ que genera una salida binaria para cada valor de input. Según donde mapee cada input, corresponderá a una clase u otra.



* https://www.saedsayad.com/logistic_regression.htm

En este caso, el modelo aprende los pesos ‘b’ de la imagen 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.

Para evaluar el modelo, utilizaremos principalmente la matriz de confusión. La misma posee en sus diagonales los valores que fueron predichos de forma correcta (True Positive y True Negatives). Las filas de esta matriz representan lo que el modelo predijo, mientras que las columnas serán nuestros ‘y_test’. Es decir, aquellos valores previamente categorizados por una persona que fueron excluidos del entrenamiento de nuestro modelo.



Esta matriz se complementa con el Accuracy , que representa que tan bien clasifico nuestro modelo de forma general (no puedo ver que tipo de errores tiene, como si puedo verlo en la matriz).

$$Acc = \frac{True\ Positives + True\ Negatives}{Total}$$

El PCA fue sumado a nuestro Pipeline para reducir la dimensionalidad de nuestras variables numéricas. Esto lo logra mediante una combinación lineal de los parámetros, al descomponer la matriz de covarianza en autovectores y autovalores. De esta manera determina las variables que mas variabilidad explican de mi modelo. En este caso al no tener un gran número de variables numéricas (solo 3) para realizar PCA no resultaría indispensable para poder graficar mis datos, pero como veremos mas adelante ayudo a la optimización del modelo.

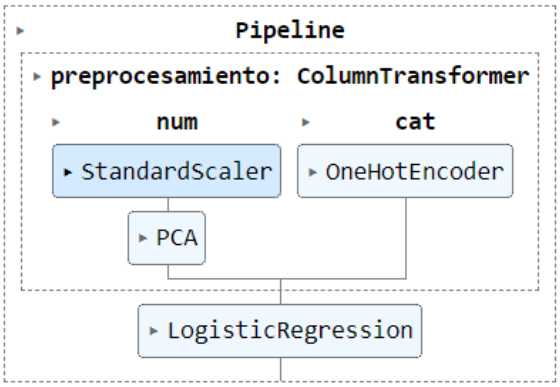
EXPERIMENTOS Y RESULTADOS

Para poder obtener los mejores hiperparametros (aquellas configuraciones de cada modelo elegidas por el usuario) se realizo un Grid Search Cross Validation (GSCV). Este método combina los hiperparametros de un modelo, luego realiza Cross Validation y obtiene el promedio del error. Luego de comparar todos los modelos, se queda con aquel que haya tenido mejor performance y recalcula los parámetros ahora con todos los datos disponibles de entrenamiento.

El Cross Validation, divide a nuestros datos de entrenamiento en ‘k’ folds o grupos, de los cuales utiliza k-1 folds como train y el grupo restante como dato de validación. De esta manera entreno al modelo con mayor cantidad de información para una misma selección de hiperparametros, sincerando al modelo.

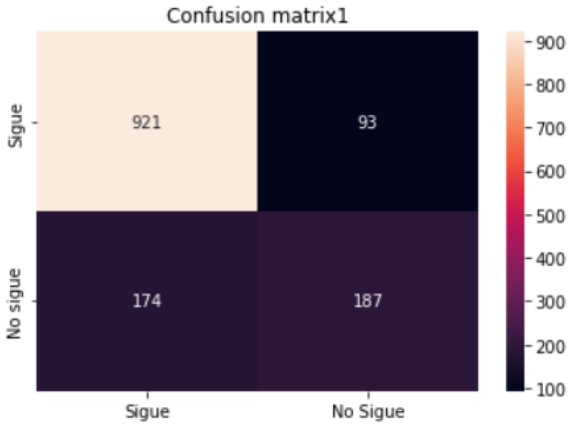
Los datos numéricos fueron previamente estandarizados, para evitar diferencias entre los rangos (por ejemplo, el valor máximo de meses era de 80 mientras que de total charge superaba los 7000) y por ser también requisito necesario del PCA. Para los categóricos se generaron ‘dummies’ mediante OneHotEncoder.

El Pipeline de nuestro modelo final quedaría entonces de la siguiente manera:



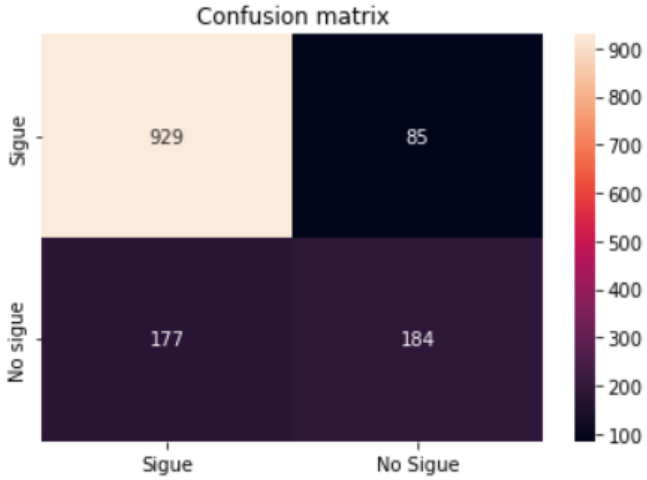
El modelo de LogisticRegression es el estimador inicial, pero luego cambia en las iteraciones del GSCV.

Como primer resultado obtuvimos la siguiente Confussion Matrix:



El modelo seleccionado fue la Logistic Regression. Con un Accuracy del 0.8058. Observamos también que el modelo tiene una tendencia a falsos negativos.

Cuando se realizó la optimización con PCA los resultados cambiaron sensiblemente:



El Accuracy de este modelo es ahora del 0.8094 gracias a la optimización del PCA.

DISCUSIONES Y CONCLUSIONES

Los modelos empleados permiten predecir de manera aceptable si un cliente abandonara el pago del servicio o no. Esta información resultaría de vital importancia para poder anticiparnos a este posible cliente que abandonaría la compañía para poder, por un lado, entender cual es el motivo por el cual se abandona la suscripción y por otro, intentar retener al cliente en el largo plazo.

Para poder comprender al cliente e intentar retenerlo, la información que se explico durante el EDA es muy importante por lo que ambas partes del informe se complementan.

BIBLIOGRAFIA

Introduction to Statistical Learning- Gareth James, Daniela Witten, Trevor Hastie,Robert Tibshirani

Scikit-learn – Informacion de librerías utilizadas.

Python Data Science Handbook - Jake VanderPlas

Deep Learning Book – Part 1