Trabajo Práctico Nº2

Mentoría Churn Prediction - Análisis Exploratorio y Curación de datos

Input:

https://github.com/yaninaiberra/DiploDatos2022MentoriaChurnCasoDeNegocio/blob/main/data/raw/mini_sparkify_event_data.zip

o el archivo input editado del TP Nº1, si hubiesen realizado alguna modificación.

Entregable:

- Se puede ir desarrollando cada punto en la misma notebook donde se escriba el código.
- Se debe subir el entregable a un repositorio GitHub o enviar el link a un Google Colab.
- Tener en cuenta que si bien pueden realizar diversos análisis y visualizaciones, se debe dejar en el entregable sólo aquello que sea relevante.
- Luego de cada análisis es importante poder obtener una conclusión de lo observado.

1. Valores nulos, outliers o erróneos

a. Analizar los valores nulos existentes y su impacto en el análisis del problema.
   i. ¿Qué porcentaje representan los registros nulos de cada variable con respecto al total?
   ii. En base a esas cantidades, decida si es posible eliminar los registros, imputarlos de alguna manera o predecirlos.
   iii. Verificar si existen registros duplicados que pudieran representar luego un problema al modelo de clasificación. ¿Qué acción tomaría sobre ellos?

b. Analizar si existen valores outliers en las variables numéricas, indagar si realmente son valores atípicos o son casos excepcionales que deben tenerse en cuenta. ¿Qué porcentaje representan? ¿Los eliminamos o mantenemos cierto porcentaje de ellos (percentiles)?
c. ¿Existen valores erróneos? Analizar cómo tratarlos, si dejarlos como están porque tienen cierto significado, eliminar las filas donde están esos valores erróneos o imputarlos de alguna manera.
d. Concluir luego de ésta “limpieza” cuántos registros hemos mantenido/eliminado. Con el fin de no quedarnos con muy pocos registros para avanzar más adelante con algún modelo de clasificación.

2. Transformación de datos existentes

a. Pensar en nuevas columnas que puede ser útil generar.
   Algunas ideas a nivel usuario:
   - Último “level” del usuario (‘paid’ o ‘free’).
   - Tiempo total de canciones, número de artistas y número de canciones escuchadas.
   - Media y desviación estándar del número de canciones escuchadas por artista.
   - Número de canciones escuchadas por sesión.
   - Número de “friends” (page = "Add Friend").
   - Cantidad de días entre la fecha de registro y la fecha de última interacción.
   - Sean creativos generando otras columnas.

b. Pasar las variables categóricas (strings) a numéricas. Analizar diferentes métodos para elegir el más adecuado (One hot encoding, categorías numéricas).
c. Transformar los features para que tengan distribuciones más cercanas a la normal (elegir qué método es más conveniente: escalar, normalizar, estandarizar).

d. Analizar la posibilidad de incluir nuevas columnas con el método PCA (éste paso debe realizarse luego de escalar o estandarizar las variables, para que en las componentes PCA las variables sean “pesadas” de manera similar, y no tenga alto impacto la varianza de las columnas originales).

3. Correlaciones

a. Verificar mediante una matriz de correlación la correlación entre features y entre cada variable y la columna target.
b. Eliminar los features fuertemente correlacionadas (una de cada par), ya que mantener columnas altamente correlacionadas, puede ocasionar un comportamiento no deseado en los modelos de clasificación.

4. Generar dataset “limpio”

Guardar en un nuevo csv el dataset “limpio”, con las transformaciones y nuevas columnas, ya que será el que utilizaremos en los siguientes TP para un modelo de clasificación.

Deadline de entrega: 17/06/2022.