Trabajo Práctico Nº1

Mentoría Churn Prediction - Análisis y visualización de datos

Input:

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

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. **Familiarización con el dataset**

a. Importar y analizar el dataset para ver con qué datos contamos, en cantidad y calidad: indagar no sólo el tipo de cada dato (categóricos, numéricos) si no su naturaleza (si son demográficos, económicos, información personal de los clientes, etc).
b. ¿Tenemos forma de identificar de manera única a cada cliente? ¿Cuántos clientes únicos tenemos en el dataset?
c. ¿Qué periodos de fechas tenemos en el dataset?
d. ¿Tenemos datos nulos? Pensando en el histórico de datos, ¿tenemos datos "suficientes" para pensar en realizar un modelo predictivo?

e. ¿Tenemos la columna target (necesaria en problemas de aprendizaje supervisado de clasificación)? ¿Cómo podemos definir "Churn" para éste dataset?
   Hint: Tenemos dos enfoques diferentes posibles (o una combinación de ambos):
   I. Predecir cuándo un usuario abandonará el servicio, definido mediante la acción 'Cancellation Confirmation' de un usuario determinado.
   II. Predecir cuándo un usuario pago bajará de categoría al servicio gratuito, definido por la acción 'Submit Downgrade'.
   III. Una combinación de ambas acciones.

   Estas acciones pueden encontrarse en la columna ‘page’.
   Se puede decidir en definir una nueva columna “churn” (con valores 1 ó 0, en caso de que haya hecho “churn” o no.) cuando se da la ocurrencia de ambas opciones.
   O definir churn sólo en base a la acción ‘Cancellation Confirmation'.
   O definir churn sólo en base a la acción 'Submit Downgrade'.
   Generalmente ésta definición es dada por el “cliente” dueño de los datos.
   Pero resulta interesante entender que muchas veces no existe una única definición de “churn” en un conjunto de datos y que es importante definirlo y tenerlo claro durante el trabajo de ciencia de datos sobre el dataset.

2. Análisis y Visualización de los datos

a. Para los datos numéricos, realizar una breve descripción de resumen estadísticos, para comprender los rangos (mín, max, media, moda, quantiles).
b. Visualizar la distribución de los features, identificar valores atípicos o datos faltantes.
c. Una vez definido “Churn”, ¿las clases están balanceadas? (¿Hay más usuarios que hacen churn que los que no o viceversa?).

d. Comparar el comportamiento de los “usuarios churn” vs “usuarios no churn” en términos de:

   i. Uso a diferentes horas del día
   ii. Uso en diferentes días de la semana
   iii. Nivel de usuario (gratuito o pago)
   iv. Tipos de eventos (por ejemplo, añadir un amigo, publicidad, pulgares arriba)
   v. Dispositivo utilizado (por ejemplo, Mac, Windows, iPhone)
   vi. Ubicación del usuario (por ejemplo, Nueva Inglaterra, Pacífico)
   vii. Tiempo transcurrido desde la baja (downgrade) hasta el abandono (Cancellation Confirmation).
   viii. ¿Cuál es la distribución de los usuarios por género?
   ix. ¿Cuáles usuarios escucharon más canciones?
   x. ¿Dónde se encuentra el mayor número de usuarios del servicio?
   xi. ¿Cuántos artistas y canciones únicos tiene el conjunto de datos?
   xii. ¿Cuáles son las canciones y artistas más populares?
   xiii. ¿Qué usuarios han escuchado más canciones?
   xiv. Página más visitada
   xv. Analice las interacciones con la página “Help” por “userId” y por ‘sessionId’.
   xvi. ¿Cómo es la relación de ambos grupos (churn vs no churn) con la page == 'Error'?

e. En base a los puntos anteriores analice la correlación entre las variables y la variable target.
f. Analice el tiempo que pasa entre que un usuario se registra y es considerado “churn” (eventos 'Cancellation Confirmation' o 'Submit Downgrade'), calcule la media y mediana en cantidad de días.
g. ¿Cuáles columnas resultan “interesantes” para ser incluídas en un futuro modelo predictivo de churn? (Debieran ser aquellas que presentan cierto comportamiento diferente para los usuarios churn o no churn).

Deadline de entrega: 20/05/2022.