Desafío 3. Predicción de Clicks

Introducción

La subasta de avisos en tiempo real (*Real-Time Bidding* en inglés) es la técnica más relevante de los últimos años en cuanto a publicidad en línea, cualquiera sea el dispositivo. Miles de millones de impresiones de anucios se compran diariamente en subastas públicas llevadas a cabo por martilleros virtuales. La subasta es por cada impresión, independientemente, y todo el proceso ocurre en menos de 100 milisegundos. En este contexto, plataformas de demanda como Jampp, tienen la tarea de ayudar a sus clientes a administrar y optimizar las campañas en estas subastas. Para esto, es central poder predecir la probabilidad de que un anuncio sea activado y poder así asignarle un valor en la subasta.

Objetivos:

- En este caso vamos a trabajar sobre una semana de datos para construir un modelo capaz de predecir si un usuario hará *click* en un determinado aviso.
- La probabilidad de click estimada es importante, no sólo si el usuario clickea o no, porque determina cuánto pagar por ese anuncio.

Requisitos

Los materiales deberán ser entregados en un Notebook Jupyter que satisfaga los requerimientos del proyecto. El notebook deberá estar debidamente comentado. Además los grupos deberán crear un repositorio para el proyecto (anonimizado) en Github. Para la presentación en clase se deben armar algunos slides no técnicos para una presentación en no más de 10 minutos.

Material a entregar

Un notebook con el código que genera los estadísticos y los gráficos debidamente comentado. El código básico y una guía de pasos fue diseñado en formato de notebook Jupyter. Pueden usar éste notebook como guía pero presentar los análisis y modelos realizados, junto con los principales resultados en un informe estructurado (ppt o google slides). El mismo debe constar en una introducción (planteo del problema, la pregunta, la descripción del dataset, etc.), un desarrollo de los análisis realizados (análisis descriptivo, análisis de correlaciones preliminares, visualizaciones preliminares) y una exposición de los principales resultados y conclusiones.

Fecha de entrega

- El material deberá entregarse en la **clase 41** del curso.

Dataset

Las variables categóricas fueron transformadas vía *hash* para anonimizarlas. Notar que puede haber, además, archivos faltantes.
Entrenamiento. Los archivos de entrenamiento ('ctr_n.csv' files) consisten en una porción de los datos de *clicks* de Jampp en el transcurso de una semana. Estos registros no están necesariamente en orden cronológico. Sin embargo, hay una columna que refiere al tiempo en el que ocurren los mismos.
Test. En el conjunto de prueba ('ctr_test.csv') se tiene una muestra del tráfico recibido por Jampp en la siguiente semana, obtenidos de forma similar al conjunto de entrenamiento. No se tiene para estos la columna "Label". Sobre estos registros es que se deberá predecir el valor de esta columna faltante. Contiene además una columna "id" que identifica el registro.

La información de cada registro que incluye el dataset es la siguiente:

- Label. Variable objetivo que indica si dicho aviso fue *clickeado* (1) o no (0).
- action_categorical_0: Identificador de unidad de Negocio, nivel 1. A cada unidad de nivel uno puede corresponderles varias unidades de nivel 2 (pero no al revés).
- action_categorical_1: Identificador de unidad de Negocio, nivel 2.
- action_categorical_2: Identificador de unidad de Negocio, nivel 3.
- action_categorical_3: Identificador de unidad de Negocio, nivel 4.
- action_categorical_4: Identificador de unidad de Negocio, nivel 5.
- action_categorical_5: Una variable categórica.
- action_categorical_6: Una variable categórica.
- action_categorical_7: Una variable categórica.
- action_list_0: Lista de categorías relacionadas con la subasta.
- action_list_1: Lista de categorías relacionadas con la subasta.
- action_list_2: Lista de categorías relacionadas con la subasta.
- auction_time: Tiempo en el que ocurrió la subasta. El tiempo está en formato *unix* (o sea, *epoch time*).
- auction_age: Edad del usuario/a
- auction_bidfloor: Mínimo valor de entrada a la subasta.
- auction_boolean_0: Atributo de la subasta, codificado en una variable binaria.
- auction_boolean_1: Atributo de la subasta, codificado en una variable binaria.
- auction_boolean_2: Atributo de la subasta, codificado en una variable binaria.
- auction_categorical_0: El identificador de una entidad relacionada con la subasta.
- auction_categorical_1: El identificador de una entidad relacionada con la subasta.
- auction_categorical_2: Una variable categórica.
- auction_categorical_3: Una variable categórica.
- auction_categorical_4: Una variable categórica.
- auction_categorical_5: Una variable categórica.
- auction_categorical_6: Una variable categórica.
- auction_categorical_7: El identificador de una entidad relacionada con la subasta.
- auction_categorical_8: El identificador de una entidad relacionada con la subasta.
- auction_categorical_9: El identificador de una entidad relacionada con la subasta.
- auction_categorical_10: Una variable categórica.
- auction_categorical_11: El identificador de una entidad relacionada con la subasta.
- auction_categorical_12: Una variable categórica.
- auction_list_0: Lista de categorías relacionadas con la subasta.
- creative_categorical_0: Unidad de negocio.
- creative_categorical_1: Una variable categórica.
- creative_categorical_10: Una variable categórica.

- creative_categorical_11: Una variable categórica.
- creative_categorical_12: Una variable categórica.
- creative_categorical_2: Una variable categórica.
- creative_categorical_3: Una variable categórica.
- creative_categorical_4: Una variable categórica.
- creative_categorical_5: Unidad de negocio.
- creative_categorical_6: Una variable categórica.
- creative_categorical_7: Una variable categórica.
- creative_categorical_8: Una variable categórica.
- creative_categorical_9: Una variable categórica.
- creative_height: Altura (en pixels) del espacio del aviso.
- creative_width: Ancho (en pixels) del espacio del aviso.
- device_id: Identificador único (o casi) del dispositivo.
- device_id_type: Tipo de identificador de dispositivos. Hay muchos tipos de identificadores e incluso un dispositivo puede tener varios tipos distintos.
- gender: género.
- has_video: Una variable que indica si el *banner* contiene un video.
- timezone_offset: Diferencia horaria (timezone offset) en horas respecto al país y región de la subasta.

¿Cómo empezar? Sugerencias

En la presentación de los resultados tengan en cuenta que es altamente probable que la audiencia no tenga un nivel técnico así que mantengan el lenguaje en un nivel accesible.

En términos generales, recuerden las siguientes sugerencias:

- Escribir un pseudocódigo antes de empezar a codear. Suele ser muy útil para darle un esquema y una lógica generales al análisis.
- Leer la documentación de cualquier tecnología o herramienta de análisis que uses. A veces no hay tutoriales para todo y los documentos y las ayudas son fundamentales para entender el funcionamiento de las herramientas utilizadas.
- Documentar todos los pasos, transformaciones, comandos y análisis que realices.

Recursos útiles

- Hashing Trick
- Hashing Trick Scikit
- Paper Google Ad Click Prediction