Contact Prediction Challenge 2023

Properati (https://www.properati.com/) es un portal inmobiliario que ofrece una propuesta novedosa para inmobiliarias o agentes particulares que quieran vender o alquilar un inmueble. Actualmente, Properati opera en Argentina, Colombia, Ecuador, Perú y Uruguay, y en todos los países realiza acuerdos con las inmobiliarias, agentes y constructoras más importantes para publicar sus propiedades. Su modelo de negocios se basa en entregar **contactos** de calidad. Al pagar sólo por contacto recibido y no por banners o pop-ups, los incentivos entre el vendedor, el usuario y Properati quedan alineados, dando por resultado un sitio limpio y con una interfaz amigable.

El objetivo de la competencia es desarrollar un modelo que prediga, para aquellas publicaciones creadas en julio, agosto y septiembre de 2022, si tendrán al menos tres contactos durante los primeros quince días de publicación. Para ello cuentan con información de las características de todas las publicaciones efectuadas durante 2020, 2021 y parte de 2022 (hasta 2022-09-15 inclusive). Además, sólo para aquellas publicaciones con fecha de creación anterior a 2022-06-16, cuentan con el detalle de cuántos contactos tuvieron en sus primeros quince días de publicación.

Noten que poder predecir la cantidad de contactos de un aviso podría ayudar a la empresa de diversas formas, tales como: pronosticar sus ingresos esperados, poder segmentar sus publicaciones mejor, comunicar este número a las agencias con el fin de promocionar su servicio, entre otras.

Datos

Para entrenar y evaluar sus modelos cuentan con distintos conjuntos de datos. A continuación se detalla cada uno de ellos:

- **Ads data** (132 archivos). Contienen datos de los anuncios de propiedades publicadas para Argentina, Colombia, Ecuador y Perú durante los años 2020, 2021 y parte de 2022 (hasta 2022-09-15 inclusive). Cada archivo contiene datos para un mes y un país dado, y dentro de cada archivo, cada registro corresponde a un anuncio. Estos archivos contienen variables predictoras que deberían ser utilizadas para elaborar sus predicciones, pero **NO cuenta con la variable *contacts***.

- **Contacts data** (1 archivo). Contiene el id de cada anuncio y la cantidad de contactos que tuvo dicho ad_id en los primeros quince días de publicación.

  Importante:

  1 - Este archivo sólo contiene información de contacto de publicaciones cuyo valor de *created_on* es menor al 16 de junio de 2022.

  2 - Ustedes deberán unir el dataset de anuncios con el de contactos a los fines de generar el dataset de entrenamiento. Para hacer esto, deberán utilizar la columna *ad_id* que aparece en ambos datasets y es única para cada publicación.

  3 - Si un ad_id de ads data no se encuentra en contacts data, esto implica que dicho anuncio no tuvo contactos (i.e., contacts debe valor 0).

- **Sample submission** (1 archivo). Es un ejemplo de la estructura que deben tener los archivos que se suban a la plataforma de Kaggle. Noten que debe tener dos columnas. La primera llamada *ad_id* (con el valor de id de cada anuncio de evaluación que se predice, este valor no debe tener duplicados ni decimales) y otra llamada *contacts* (con

la probabilidad predicha de que un anuncio tenga al menos tres contactos en sus primeros 15 días de ser publicado). El archivo se encuentra delimitado por comas y tiene nombres de columnas.

Variables predictoras contenidas en ads data

Cada conjunto de datos tiene 26 variables que pueden usarse para predecir la cantidad de contactos. A continuación se copia la información que Properati dio respecto a cada una de ellas:

- *ad_id*: ID de cada anuncio.
- *operation*: motivo de publicación (venta, alquiler, desarrollo inmobiliario, etc.).
- *place_l1*: país.
- *Place_l2*: provincia.
- *place_l3*: barrio.
- *place_l4*: información más precisa de la ubicación.
- *place_l5*: información aún más precisa de la ubicación.
- *place_l6*: información aún más precisa de la ubicación.
- *lat*: latitud de la propiedad.
- *lon*: longitud de la propiedad.
- *price*: precio publicado.
- *currency_id*: moneda del precio publicado.
- *price_usd*: precio publicado en dólares.
- *rooms*: cantidad de ambientes (variable opcional).
- *bedrooms*: cantidad de dormitorios (variable opcional).
- *bathrooms*: cantidad de baños (variable opcional).
- *surface_total*: metros cuadrados totales de la propiedad.
- *surface_covered*: metros cuadrados cubiertos de la propiedad.
- *title*: título de la publicación.
- *description*: descripción del anuncio.
- *property_type*: tipo de propiedad.
- *created_on*: fecha de publicación.
- *development_name*: nombre del desarrollo inmobiliario. Representa el desarrollo completo, es decir, el edificio/complejo entero. No representa un inmueble a alquilar o vender.
- *current_state*: condición del desarrollo inmobiliario.
- *short_description*: descripción corta del anuncio.
- *property_is_development*: indica si la propiedad corresponde o no a un desarrollo inmobiliario. Son los departamentos/propiedades en sí que provienen de un desarrollo.

Política de privacidad de los datos

Los datos aquí provistos son de uso exclusivo para los alumnos de la materia Minería de Datos del Máster in Management and Analytics de la Universidad Torcuato Di Tella, dentro del marco de la competencia "*Contact Prediction Challenge*". Siendo expresamente prohibida su copia, reproducción y difusión, así como el aprovechamiento y comunicación de su contenido para otros fines.