Trabajo Práctico 2

Tomás Gallo, Nadia Molina y Jazmín Sneider

Octubre 2024

1 Introducción

En este informe, haremos un recorrido por las ideas planteadas a lo largo de los diferentes submits, detallando los modelos que se eligieron para abordar el problema y el proceso de preparación de los datos necesario para ejecutar dichos modelos de manera eficaz.

2 Análisis exploratorio:

Es relevante destacar que contábamos con muy poca información clara sobre los datos, ya que gran parte de ellos estaban hasheados. No obstante, sospechábamos que algunas variables podrían ser especialmente significativas, como la edad de los usuarios y el dispositivo desde el que realizaban sus compras.

2.1 Variables de interés:

En el análisis de la variable edad, separamos los datos en rangos etarios, lo que nos permitió observar una tendencia coherente: la mayor cantidad de compras se concentraba en los grupos más jóvenes, específicamente entre los 15-24 y 25-34 años. También encontramos algunos datos atípicos, como usuarios mayores de 100 años o menores de 15, lo cual no tiene sentido en un contexto de compras, ya que estas personas, por lo general, no pueden realizar transacciones. Decidimos, por tanto, tratarlos como valores indefinidos, aunque la cantidad de estos casos fue muy reducida en comparación con el total de datos disponibles (solo 45). En cuanto al segundo análisis, relacionado con el dispositivo utilizado para la compra, partimos de la hipótesis de que este factor podía influir significativamente en la decisión de compra. Por ejemplo, es posible que comprar desde una computadora sea más cómodo que desde un teléfono móvil, o que algunos usuarios prefieran un sistema operativo sobre otro, como Linux frente a Mac o Windows, este último siendo el más común. Sin embargo, debido a que las identificaciones de los dispositivos estaban hasheadas, no pudimos confirmar a qué dispositivo específico correspondía la información que estábamos analizando. Lo que sí pudimos confirmar es que algunos dispositivos se utilizan considerablemente más que otros.

(Figure 1 y 2)

2.2 Correlación entre variables:

Quisimos ver la correlación entre las distintas variables con la label. Algunas de las variables más importantes en términos de correlación positiva incluyen **creative_height**, y **creative_categorical_2_count_encoded**, las cuales presentan correlaciones cercanas a 0.1, lo que sugiere que tienen una relación positiva, aunque moderada, con el objetivo. Por otro lado, algunas variables como **auction_categorical_2_count_encoded** y **creative_categorical_10_count_encoded** tienen una correlación negativa significativa, alrededor de -0.075. Esto indica que, cuando estas variables tienen valores altos, es menos probable que el evento de interés (por ejemplo, un clic en el anuncio) ocurra.

(Figure 3)

3 Ingeniería de Atributos:

3.1 Interacciones entre variables:

Para la ingeniería de atributos, implementamos varias interacciones y creamos nuevas características para enriquecer el conjunto de datos y mejorar los resultados de los modelos. Primero, trabajamos con interacciones numéricas. Por ejemplo, multiplicamos **auction_age** con el valor de **auction_bidfloor** y también con **auction_time**. Esto nos permitió ver cómo se relacionaban estas variables en conjunto. Además, creamos combinaciones cuadráticas, como elevar al cuadrado: **auction_age_squared** y **auction_bidfloor_squared**, para captar efectos que no se verían de manera lineal. También exploramos la relación entre el tamaño del anuncio, definido por **creative_height** y **creative_width**, y el valor mínimo de la puja **auction_bidfloor**. Esta interacción nos ayudó a analizar cómo el tamaño del anuncio podría influir en la cantidad mínima que los anunciantes estarían dispuestos a pagar. Para las variables categóricas, creamos nuevas combinaciones. Por ejemplo, juntamos el tipo de acción **(action_categorical_0)** con el tipo de creatividad **(creative_categorical_0)**, generando una nueva variable. De la misma forma, combinamos el tipo de dispositivo **(device_id_type)** con el género del usuario **(gender)** para obtener una variable que pudiera mostrar cómo estas dos características juntas influyen en el comportamiento de compra. También concatenamos los tres niveles de la jerarquía de acciones **(action_categorical_0, action_categorical_1 y action_categorical_2)** para crear una descripción más completa de las acciones que sucedían en cada subasta. Además, trabajamos con variables basadas en listas, donde calculamos la longitud de las listas de acciones **(action_list_0, action_list_1, action_list_2)**y sumamos estos valores para poder ver cómo se relacionan la cantidad de acciones de cada una de las variables por observación. También generamos ratios para entender

la proporción entre las diferentes listas de acciones. Finalmente, para las variables booleanas, combinamos varias de ellas, como **auction_boolean_0**, **auction_boolean_1** y **auction_boolean_2**, para crear una variable que indicara cuándo ciertos atributos estaban activos al mismo tiempo. También calculamos diferencias entre variables clave, como la diferencia entre el **auction_time** y **auction_bidfloor**, así como el ratio entre la altura y **creative_aspect_ratio**, que intuíamos que nos serviría para maximizar el potencial de los datos originales.

3.2 One-Hot Enconding y NA's

Dado que optamos por utilizar **XGBoost** para nuestro modelo, tuvimos que adaptar las variables categóricas, ya que este algoritmo no maneja bien ese tipo de datos por defecto. Para resolver esto, decidimos aplicar la técnica de *one-hot encoding*, que convierte cada categoría en una serie de columnas binarias, facilitando su procesamiento por el modelo. Al mismo tiempo, aprovechamos este paso para tratar los valores faltantes en las variables. En las variables numéricas, imputamos los valores ausentes usando la media, mientras que para las categóricas utilizamos el valor que más se repetía en cada columna.

4 Conjunto de validación:

Al inicio del proyecto, seleccionábamos uno de los archivos CSV disponibles para entrenar el modelo. Utilizábamos una estrategia de separación hold-out set, donde un 20% del archivo seleccionado se destinaba al conjunto de validación. Sin embargo, a medida que avanzábamos con los submits, sospechamos que entrenar con la mayor cantidad de datos posible iba a ser una mejor opción. Con esta nueva perspectiva, comenzamos a reducir progresivamente el porcentaje destinado a la validación, priorizando el uso de más datos para el entrenamiento. No obstante, al reducir demasiado este porcentaje, nos quedamos con un conjunto de validación bastante pequeño, por lo que optamos por usar datos de otros archivos CSV para validar. Con el mismo objetivo de maximizar el volumen de datos de entrenamiento, decidimos intentar entrenar utilizando más de un archivo CSV a la vez. Sin embargo, esto no mejoró los resultados. Incluso ajustamos los rangos de los hiperparámetros para evitar la caída en la performance, pero estos cambios tampoco fueron efectivos. Ante esto, dejamos de intentar esta opción. Finalmente, decidimos volver a entrenar con un solo archivo CSV y validar con otros. Esta estrategia de separación fue la que utilizamos en nuestra submit más exitosa. Sin embargo, el uso de múltiples archivos CSV para la validación sobrecargaba la capacidad de procesamiento de las computadoras, dado que implicaba validar millones de datos en varias dimensiones. Esto provocaba fallos, como la aparición de mensajes de "kernel crashed while executing" o incluso el apagado de las máquinas.

5 Modelo Elegido:

A lo largo de los submits, probamos algunos de los modelos que vimos en clase: Random Forest y Análisis de componentes principales, pero fue con **XGBoost** con el que obtuvimos los mejores resultados. Por esta razón, realizamos múltiples pruebas con este modelo, modificando hiperparámetros y hasta intentando combinarlo con otros modelos para mejorar su rendimiento, por ejemplo haciendo ensamble de 5 xgboosts modificando la semilla de cada uno y combinando pca con xgboost para manejar la multiplicidad de dimensiones (y evitar que el kernel muriera). Al no ver mejoras significativas con las pruebas manuales de hiperparámetros, decidimos implementar la técnica de **Random Search**. Aunque este proceso llevaba más tiempo en ejecutarse, terminó siendo la mejor forma de optimizar el modelo. Tuvimos que ser cuidadosos al definir el rango de búsqueda, ya que no todos los hiperparámetros funcionan bien con todos los valores, por lo que fue un proceso de prueba y error constante. Los hiperparámetros que seleccionamos para optimizar fueron los que conocíamos por haberlos trabajado en clase. Sin embargo, quizá hubiera sido interesante explorar otros menos familiares que podrían haber ofrecido mejores resultados.

6 Atributos del modelo final:

El gráfico, generado después de aplicar one-hot encoding a las variables categóricas, muestra que las primeras posiciones están ocupadas por categorías de las variables **auction_categorical_9**, **auction_categorical_7** y **creative_categorical_8.** Esto indica que las entidades relacionadas con subastas y aspectos creativos del anuncio influyen significativamente en la predicción. En particular, **auction_categorical_9** aparece dos veces en las primeras posiciones, sugiriendo que diferentes categorías dentro de esta variable son importantes.
Además, características creativas como **creative_categorical_8**, **creative_area_bidfloor** y **creative_aspect_ratio** también destacan, lo que resalta el papel crucial de la creatividad en la eficacia de los anuncios. Variables como **auction_bidfloor** y **bidfloor_creative_size** también son relevantes, lo que sugiere que afectan los resultados. Por último, las variables **action_categorical_0** y **action_categorical_1**, relacionadas con los primeros niveles de la jerarquía de acción, muestran un impacto notable. A medida que se avanza en la lista de características, su importancia disminuye.

(Figure 4)

Anexo

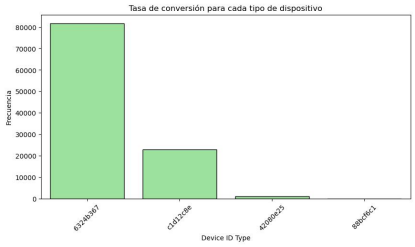

Figure 1: Análisis según dispositivo

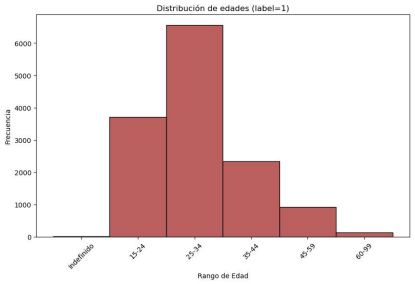

Figure 2: Análisis según edad

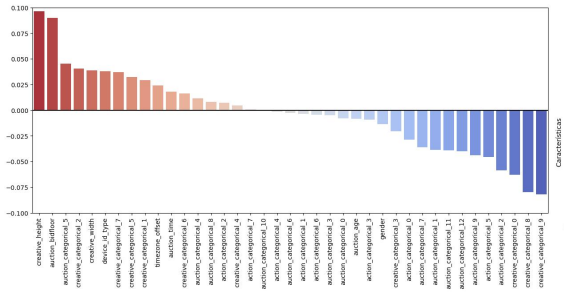

Figure 3: Correlación de Variables con label

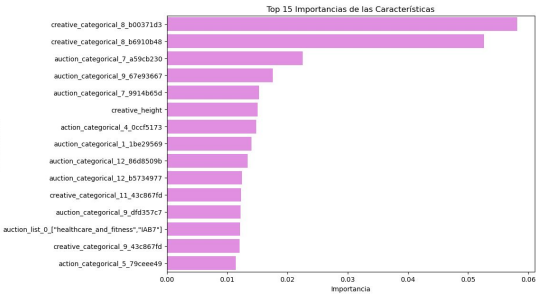

Figure 4: Atributos con más importancia