Trabajo Práctico 2

Juan Manuel Basso
Emiliana Verdun
Mariana Zunino

Tecnología Digital VI

Universidad Torcuato Di Tella

7 de octubre de 2024

1 Análisis exploratorio de los datos

1.1 Descripción de las variables principales

El conjunto de datos proporcionado incluye características referidas al usuario como así de la subasta. Las principales variables son:

- **Label**: Es la variable objetivo y representa si un usuario hizo click (1) o no (0) en una publicidad.
- **auction_bidfloor**: Es una variable numérica y puede influir en la propensión de los usuarios a interactuar con anuncios de alto o bajo costo.
- **auction_time**: Representa la hora en que se realizó la subasta. Esta variable es importante ya que el comportamiento del usuario puede variar a lo largo del día y de la semana.
- Categorías de negocio (**action_categorical_0 a action_categorical_7**): Estas variables categóricas identifican diferentes aspectos del negocio y pueden ser claves para identificar patrones de comportamiento en ciertos segmentos de usuarios o tipos de anuncios.
- **auction_age**: Representa la edad del usuario al momento de la subasta, lo que podría estar relacionado con las tendencias de interacción con las publicidades.

1.2 Análisis y Visualización de Patrones

1.2.1 Relación entre *auction_time* y *Label*

La hora del día en que ocurre la subasta puede influir en el comportamiento del usuario. A continuación, se presenta la distribución de impresiones por hora del día, seguido por la distribución de clicks (*Label = 1*) por hora.

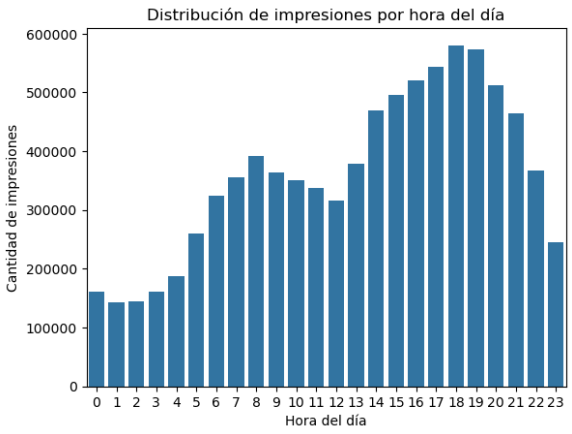

Figure 1: Cantidad de impresiones por hora del día.

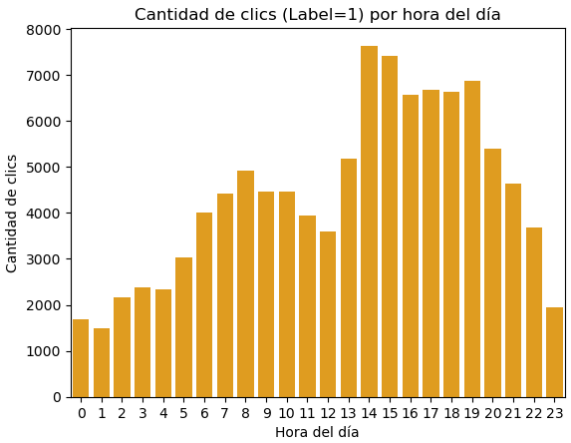

Figure 2: Cantidad de clicks (*Label = 1*) por hora del día.

1.2.2 Correlación de Pearson entre las Variables y *Label*

Para identificar las variables numéricas que parecen estar más relacionadas con los clicks, se calculó el coeficiente de correlación de Pearson entre las principales variables numéricas y la variable *Label*. En el siguiente gráfico se muestran las 8 variables con mayor correlación.

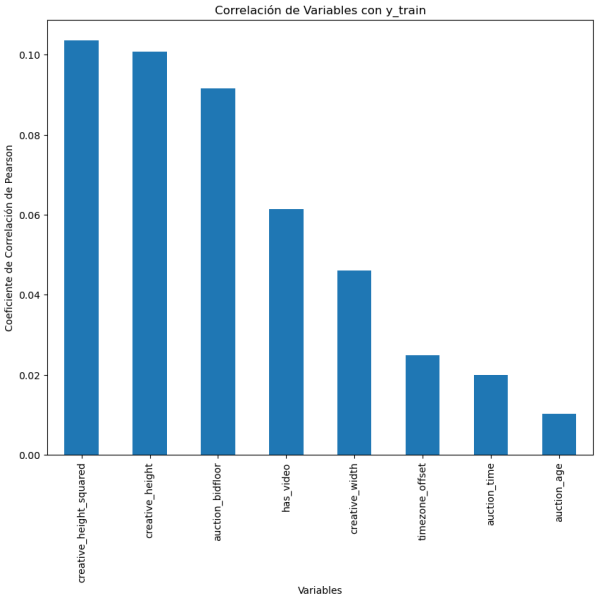

Figure 3: Correlación de Pearson de las variables numéricas con *Label*.

2 Creación de Variables Adicionales

Durante el análisis, se crearon varias variables adicionales con el fin de mejorar la capacidad predictiva del modelo. A continuación, se describen las principales:

- **Interacción entre *auction_bidfloor* y *auction_age***: Se generó la variable *bidfloor_age_interaction*, que representa el producto entre el valor mínimo de oferta en una subasta (*auction_bidfloor*) y la edad del usuario al momento de la subasta (*auction_age*). La idea detrás de esta interacción es capturar posibles efectos conjuntos entre el costo de los anuncios y la edad del usuario en la probabilidad de hacer click.
- **Cuadrado de la altura del anuncio (*creative_height_squared*)**: Se generó una nueva variable a partir de elevar al cuadrado la altura original del anuncio (*creative_height*). Esto permite capturar posibles relaciones no lineales entre el tamaño del anuncio y la probabilidad de que un usuario haga click.
- **Hora del día (*hour*)**: A partir de la variable *auction_time*, que representa el momento exacto en que ocurrió la subasta, se extrajo la hora del día. Esto nos permite analizar patrones de comportamiento de los usuarios en función de las horas del día, ya que la actividad de los usuarios puede variar considerablemente a lo largo del día.
- **Día de la semana (*day_of_week*)**: De forma similar, se generó una variable que indica el día de la semana en que ocurrió la subasta. Esto puede ser útil para identificar patrones semanales en el comportamiento de clicks.

- **Indicador de fin de semana (*is_weekend*)**: Esta variable booleana indica si la subasta ocurrió durante el fin de semana. La hipótesis es que el comportamiento de los usuarios podría ser diferente entre los días de semana y los fines de semana.

Estas nuevas variables fueron diseñadas para capturar más información sobre el comportamiento de los usuarios y mejorar el rendimiento predictivo del modelo.

3 Conjunto de Validación

Para armar el conjunto de validación, se utilizó la técnica de *train-test split*, dividiendo los datos en un 80% para entrenamiento y un 20% para validación.

La proporción 80/20 fue seleccionada porque ofrece una cantidad suficiente de datos para entrenar el modelo, al mismo tiempo que permite evaluar su rendimiento en un conjunto de validación no visto durante el entrenamiento. La cantidad de observaciones de validación con esta proporción es similar a la vez con la cantidad de observaciones del conjunto de testing.

En resumen, esta estrategia de validación asegura que el modelo se entrene de manera adecuada y se evalúe su rendimiento antes de realizar predicciones sobre el conjunto de prueba final.

4 Modelos Predictivos y Búsqueda de Hiperparámetros

En este trabajo se utilizaron dos modelos principales de aprendizaje automático para predecir si un usuario haría click en una publicidad: **XGBoost** y **LightGBM**. A continuación, se explica el proceso de optimización de cada modelo y el desempeño obtenido.

4.1 XGBoost

El modelo XGBoost se seleccionó debido a su capacidad para manejar grandes volúmenes de datos y su buen rendimiento en problemas de clasificación binaria. Para optimizar los hiperparámetros, se realizó una búsqueda mediante el método de *Randomized Search*, explorando los siguientes parámetros: `max_depth`, `learning_rate`, `n_estimators`, `gamma`, `subsample` y `colsample_bytree`.

Esta búsqueda permitió probar diferentes combinaciones de estos parámetros, y tras múltiples iteraciones, se identificaron los hiperparámetros que ofrecieron el mejor rendimiento para este modelo.

4.2 LightGBM

El modelo LightGBM, por su parte, fue seleccionado debido a su eficiencia en términos de tiempo de entrenamiento, su capacidad de trabajar con categóricas y su capacidad para manejar grandes conjuntos de datos. En este caso también se realizó un *Randomized Search* de los siguientes hiperparámetros: `max_depth`, `learning_rate`, `n_estimators`, `min_child_sample` y `colsample_bytree`. Si bien está búsqueda fue lo más costoso temporalmente, su realización permitió llegar a muy buenos resultados y fue más eficiente que

la realizada con XGBoost. Se eligió nuevamente Random Search, con 10 combinaciones posibles, porque realizar Grid Search era muy costoso en términos de tiempo.

Este ajuste permitió optimizar el rendimiento del modelo LightGBM de manera eficiente, resultando en una mejor capacidad predictiva en comparación con XGBoost.

4.2.1 Comparación de Rendimientos

Tras realizar las pruebas con ambos modelos, **LightGBM** fue el que obtuvo los mejores resultados en términos de AUC (0.827), lo que lo convirtió en el modelo final seleccionado. Aunque XGBoost también mostró un buen rendimiento (0.79), LightGBM se destacó por su velocidad y eficiencia en el manejo del conjunto de datos, además este modelo tiene la capacidad de manejar variables categóricas cosa para lo cual xgboost necesita transformaciones, como one-hot encoding, que al ser tan grande el conjunto de datos tardaban demasiado tiempo.

5 Importancia de Atributos en el Modelo Final

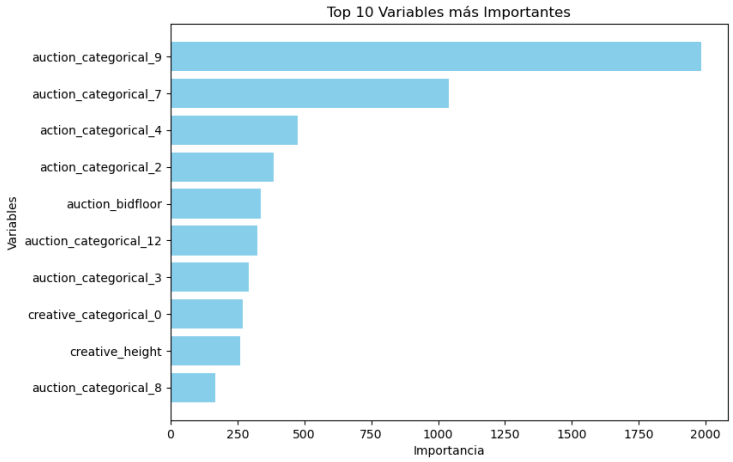

Figure 4: Atributos más importantes *Label*.

En nuestro modelo final, el que usa LightGBM, se ve que en las variables con más importancia la mayoría son justamente las variables categóricas.