Tecnológico de Monterrey

Maestría en Inteligencia Artificial Aplicada

Avance 2

Ingeniería de Características

Proyecto Integrador (Gpo 10)

Alumnos:

Farid Krayem Pineda – A00506281

Gerardo Quiroga Nájera - A00967999

Kevin Alan García Macías - A01794867

12 de Mayo 2025

Índice

Conversión de datos en variables

Detalles y justificación en archivos de GIT

Métodos de filtrado

Detalles y justificación en archivos de GIT

Conclusiones de fase 1

Calidad y estructura de los datos:

Nuestro primer dataset se encuentra sin valores nulos en campos esenciales como Clics, Impresiones y Posición. Por ejemplo, la Fecha contiene 89 registros, todos completos, lo que indica una captura diaria estable. Esta calidad inicial es esencial para cualquier aplicación de Machine Learning, de acuerdo con Shearer, 2000.

Asimetría en distribuciones y presencia de outliers:

Las variables numéricas:

En C*onsultas*, el 75 % de las consultas tienen menos de 1 clic, con un promedio de 58%, pero el valor máximo es 34 clics. Esto significa una distribución de tipo "long tail" (Anderson, 2006), típica de entornos digitales donde pocos elementos concentran la atención. Este patrón nos hace pensar que requeriremos que nuestro modelo maneje datos desbalanceados.

CTR bajo e ineficiencia en visibilidad orgánica:

La mayoría de las páginas y consultas presentan un bajo CTR. En la variable de *Paginas*, la página con mayor número de impresiones (34 693) recibe 382 clics (CTR ≈ 1.10 %), mientras otras páginas tienen menos de 5 clics. Esto sugiere problemas de atractivo en los resultados presentados en Google. Según Fishkin (2015), mejorar metaetiquetas puede aumentar la tasa de clics.

Concentración de tráfico en pocos dispositivos y países:

En la variable de *Dispositivos*, 98.8 % de los clics provienen de dos dispositivos, lo que indica que la optimización debe centrarse en ellos. De forma similar, *Países* muestra que un solo país concentra 1,746 clics de un total de 1,880. Este tipo de concentración es clave para segmentar campañas o construir modelos por región. Brynjolfsson, E., Hu, Y. J., & Smith, M. D. (2006). From niches to riches: Anatomy of the long tail. Sloan Management Review, 47(4), 67–71.

Bibliografía

Davenport, T. H., & Bean, R. (2018, February 15). Big companies are embracing analytics, but most still don't have a data-driven culture. Harvard Business Review Digital Articles. https://hbr.org/2018/02/big-companies-are-embracing-analytics-but-most-still-dont-have-a-data-driven-culture.

Anderson, C. (2006). The Long Tail: Why the Future of Business is Selling Less of More. Hyperion.

Brynjolfsson, E., Hu, Y. J., & Smith, M. D. (2006). From niches to riches: Anatomy of the long tail. MIT Sloan Management Review, 47(4), 67–71.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Fishkin, R. (2015). How Google's CTR data may influence search rankings. Moz. https://moz.com/blog/google-ctr-data-influence-search-rankings

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.