

Modelo Predictivo del Precio de Oferta de Inmuebles Residenciales en México a Partir de Variables Estructurales, Geográficas y Textuales de Anuncios Inmobiliarios

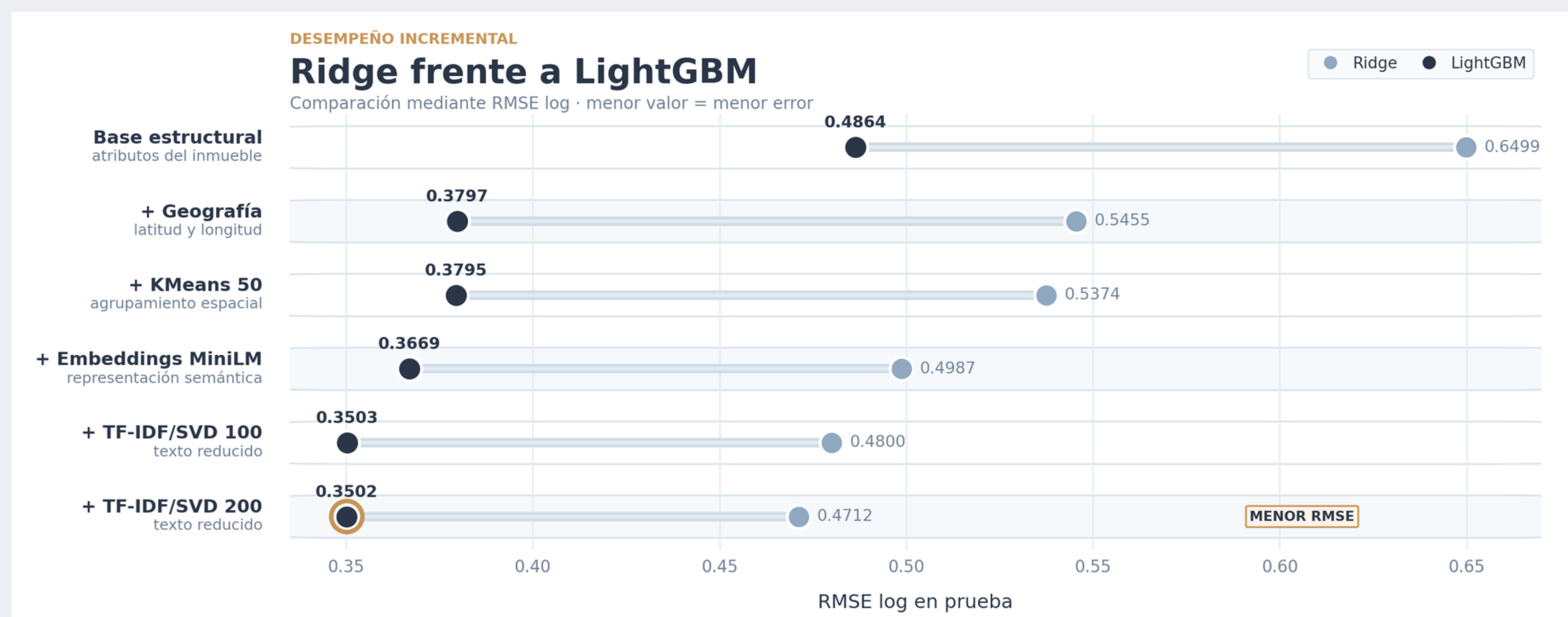
1 Introducción

Los anuncios inmobiliarios concentran información sobre superficies, recámaras, amenidades, ubicación y descripciones que intentan destacar aquello que vuelve atractiva a una propiedad. Esa combinación convierte el anuncio en una fuente valiosa para analizar el precio de oferta de los inmuebles.

Desde el enfoque hedónico de precios, una vivienda puede entenderse como un conjunto de atributos que el mercado valora y que se reflejan en su precio (Lancaster, 1966; Rosen, 1974). A partir de esta idea, el proyecto desarrolló un modelo predictivo incremental para estimar el precio de oferta de inmuebles residenciales en México y comparar el aporte de tres tipos de información: variables estructurales, geográficas y textuales.

3 Resultados

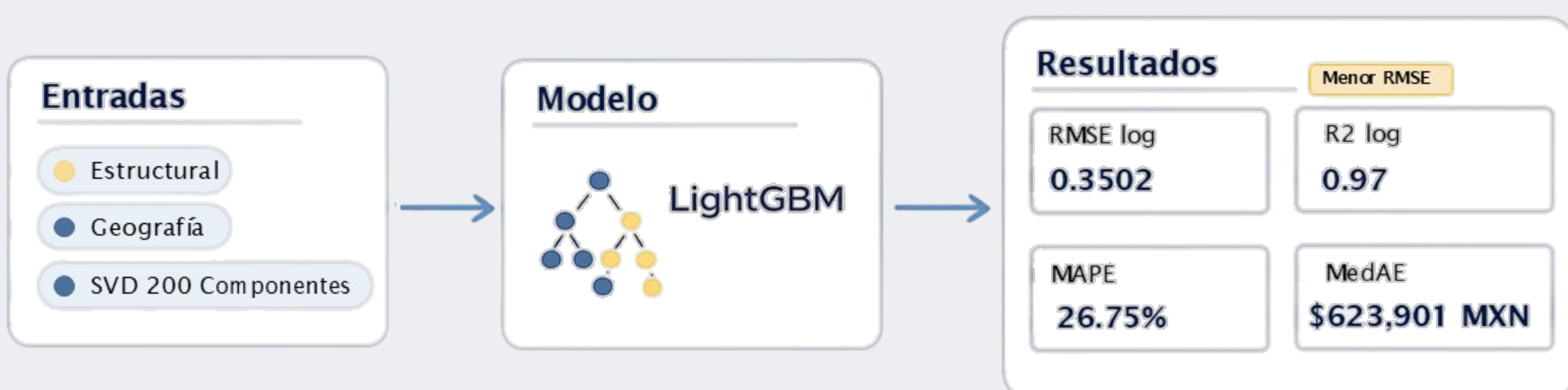
La comparación incremental mostró que el desempeño cambió de forma distinta al incorporar cada bloque de información.



Geografía: La ubicación produjo la mayor reducción del error respecto a la configuración estructural. K-means tuvo un efecto marginal, por lo que no formó parte del modelo final seleccionado.

Texto: Las dos representaciones textuales mejoraron el desempeño frente al modelo geográfico. MiniLM alcanzó un RMSE de 0.3669, mientras que TF-IDF/SVD obtuvo 0.3503 con 100 componentes y 0.3502 con 200 componentes. La diferencia entre estas dos últimas configuraciones fue mínima.

Las descripciones añadieron información predictiva y una representación léxica reducida capturó mejor esas señales que los embeddings.



5 Referencias

- Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74(2), 132-157.
- Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55.

2 Metodología

Preparación de datos

Se consolidaron 125,861 anuncios inmobiliarios residenciales, integrando el precio oferta, características del inmueble, ubicación y descripciones publicadas.



Construcción incremental del modelo

El modelo comenzó con las características estructurales del inmueble. Después se incorporó la información geográfica y el agrupamiento espacial. En la última etapa se probaron dos representaciones textuales.



Evaluación y selección

Cada configuración se entrenó por separado con Ridge y LightGBM. Después se comparó su desempeño para identificar qué combinación de variables y modelo producían los mejores resultado.



4 Conclusiones

Las variables estructurales establecieron el punto de partida; la incorporación de la geografía produjo la mayor reducción del error, mientras que el texto añadió una mejora posterior.

LightGBM obtuvo menor error que Ridge en todas las configuraciones comparadas. TF-IDF/SVD superó a los embeddings MiniLM, aunque las versiones de 100 y 200 componentes presentaron resultados prácticamente equivalentes. SVD 200 se seleccionó por registrar el menor RMSE log, con 0.3502, frente a 0.3503 de SVD 100