Diana Olano - A00369468
Laura Martínez - A00365187

Predicción de precios de propiedades

Proceso de investigación

En este apartado, se presenta un resumen del análisis CRISP-DM aplicado al proyecto de predicción de precios de propiedades en una agencia de bienes raíces.

El objetivo de este proyecto es convertir a la empresa de bienes raíces en líder del mercado mediante ventas y adquisiciones exitosas de propiedades, utilizando una estrategia precisa de evaluación de precios. Además, se busca implementar un sistema de predicción de precios de propiedades que mejore la ventaja competitiva de la empresa y aumente la eficiencia en la venta de propiedades.

El procesamiento de datos se divide en dos objetivos principales: identificar los factores clave que influyen en el precio de una propiedad y predecir dicho precio considerando estos factores. Se establecen indicadores de éxito relacionados con la precisión del modelo, su confiabilidad, capacidad de generalización y tasa de falsos positivos/negativos.

En cuanto a las necesidades del negocio, se destaca la importancia de predecir las tendencias del mercado para tomar decisiones informadas a largo plazo, así como ahorrar tiempo y recursos en la evaluación de precios. Además, se menciona la necesidad de identificar oportunidades de inversión en el sector inmobiliario.

A raíz de lo anterior surge la siguiente pregunta de interés: ¿Cuáles son los factores con mayor influencia al establecer los precios de propiedades en Colombia durante el periodo de 2023?

Finalmente, se mencionan los usuarios potenciales de la solución desplegada, como agentes inmobiliarios, inversores, propietarios y compradores que buscan información precisa y valiosa sobre los precios de las propiedades.

Desarrollo

El proyecto se desarrolló en Google Colab, en el que se usaron librerías de análisis, manipulación y visualización de datos.
Inicialmente, se realizó el análisis exploratorio y la limpieza. Aquí, se eliminan las columnas que, en su mayoría, tienen valores vacíos, también aquellas que más del 90% de sus filas correspondan al mismo valor.
Asimismo, se eliminaron las filas que no tuvieran datos en la columna “Precio”, dado que cualquier instancia sin precio no puede ser utilizada porque no será útil para aprender de ella.
Por otra parte, fue necesario utilizar una herramienta de Geolocalización, en este caso se usó OpenStreetMap, con el objetivo de encontrar los valores faltantes de las columnas “city” y “departments” por medio de las coordenadas de latitud y longitud (reverse geocoding).

Posteriormente, se desarrollaron Pipelines. Primero, se implementó una clase que agrupa aquellos elementos cuyos tamaños sean menores que 1000 y se ubican en la categoría “Other”. También, se creó una clase para eliminar columnas, otra para realizar el Feature Selection y finalmente, una clase que personalizaba el One Hot Encoder. Con estas implementaciones se realizó un Feature Union que maneja el preprocesamiento de las variables tanto numéricas como categóricas.

Validación

En la evaluación de los modelos para determinar cuál es el más óptimo para el proyecto, se llevó a cabo una implementación y aplicación de medidas de evaluación de rendimiento en diferentes modelos de regresión. Estos modelos abarcan desde Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression hasta Gradient Boosting Regressor.

Primero, se importa la librería Pycaret, y se utiliza la función compare_models para comparar y seleccionar automáticamente el mejor modelo de regresión entre una variedad de opciones disponibles. Esta función evalúa varios modelos utilizando validación cruzada y devuelve el modelo que muestra el mejor rendimiento según una métrica predefinida, como el coeficiente de determinación (R2) o el error cuadrático medio (RMSE).

La segunda forma utilizada para evaluar los modelos fue algo más manual, para ello se obtuvieron los mejores valores de Alpha para Lasso y para Ridge. Posteriormente, se evaluaron 5 modelos, esto se realizó con la creación de un Pipeline que contiene el preprocesamiento de los datos, el estándar scaler y un model, este último varía de acuerdo al modelo que se esté evaluando. Para este proceso se utilizaron diferentes medidas de evaluación de rendimiento, como el error absoluto medio (MAE), el error cuadrático medio (RMSE) y el coeficiente de determinación (R2)

Luego, se selecciona el mejor modelo, Light Gradient Boosting Machine, y se utiliza la función create_model para crear una instancia del modelo LightGBM, que es un algoritmo de refuerzo de gradientes. Este modelo se crea utilizando la configuración óptima determinada por PyCaret durante la comparación de modelos.

Para finalizar, el proceso de evaluación que se realizó proporciona una visión completa de la implementación matemática y aplicación de medidas de evaluación de rendimiento para diferentes modelos de regresión. Esto permite una comparación precisa del rendimiento de cada modelo y ayuda a seleccionar el modelo más adecuado para un problema específico.

Despliegue de la solución analítica

Para desplegar el modelo se utilizó el framework Flask. En la clase app.py se carga el modelo y se especifican las rutas y columnas necesarias para que funcione el modelo.
En la página desplegada, se encuentran las 6 columnas, y al lado un cuadro para presentar el resultado de la predicción.