IT ACADEMY

Barcelona Activa

Predicción del precio de una vivienda

*Modelo Híbrido CNN

California

Verónica Sánchez

Proyecto IT Academy Bootcamp Data Science
Marzo 2023
https://github.com/vsm-data-science

Introducción

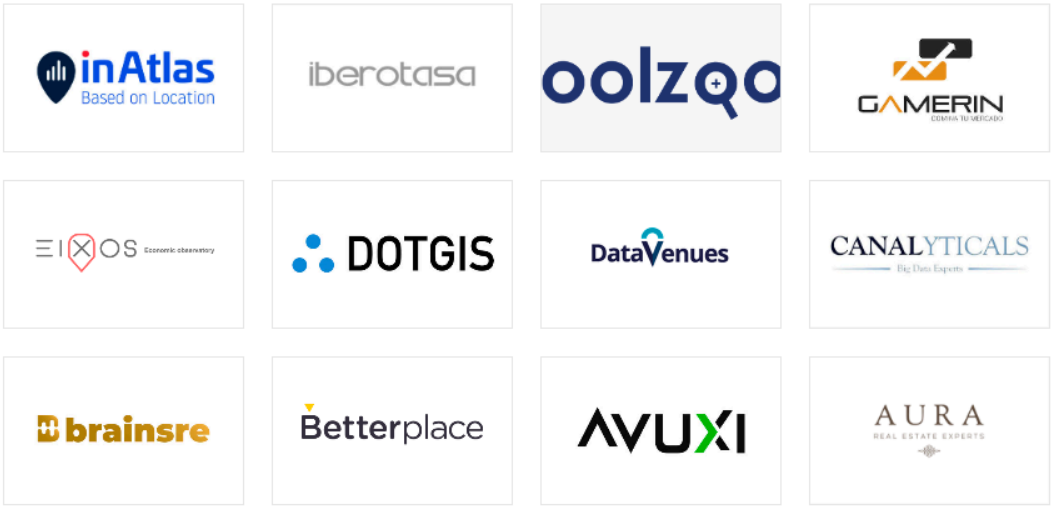
En el competitivo mercado inmobiliario español, diversas herramientas y plataformas ofrecen servicios de valoración de viviendas, cada una con su metodología y enfoque. Según se puede explorar en el Mapa Proptech de España (http://mapaproptech.com/mapa/), aunque existe una variedad de soluciones tecnológicas enfocadas en el sector inmobiliario, son pocas las que integran de manera efectiva la inteligencia artificial para analizar datos históricos de ventas reales y el estado actual de las propiedades mediante imágenes. La principal barrera para incorporar estos elementos es la dificultad para acceder a datos fiables de transacciones pasadas, los cuales suelen estar restringidos y son accesibles principalmente a través del Registro de la Propiedad.

VALORADORES DE VIVIENDAS ESPAÑA

Fuente: http://mapaproptech.com/mapa/

La mayoría de las herramientas disponibles tienden a enfocarse en la oferta y demanda actual, así como en las características físicas y de ubicación de las propiedades, para determinar su valor en el mercado. Esta aproximación, aunque útil, puede resultar en valoraciones que no reflejan completamente la realidad del mercado ni el verdadero valor de las viviendas. La falta de un análisis profundo y detallado que incluya datos históricos de ventas y una evaluación visual del estado actual de la propiedad, puede conducir a estimaciones de precio que, en ocasiones, distorsionan la percepción del valor real de las viviendas.

VALORADORES DE VIVIENDAS ESPAÑA



Fuente: http://mapaproptech.com/mapa/

Dicha limitación subraya la importancia de que estos procesos de valoración sean supervisados y revisados por profesionales del sector inmobiliario, quienes pueden aportar su conocimiento y experiencia para ajustar las estimaciones a la realidad del mercado. Sin embargo, esta dependencia también destaca la necesidad de innovar y buscar métodos más avanzados y precisos para la valoración de propiedades, que integren la inteligencia artificial y el análisis de datos de forma más efectiva.

En este contexto, el proyecto que se desarrolla para el mercado inmobiliario de California representa un esfuerzo por superar estas limitaciones. Al hacer uso de técnicas de web scraping para recopilar datos históricos de ventas reales y combinar esta información con el análisis de imágenes de las propiedades, se busca proporcionar una herramienta de valoración más precisa y confiable. Este enfoque no solo tiene el potencial de mejorar la exactitud de las valoraciones inmobiliarias, sino también de ofrecer una perspectiva más completa y ajustada a la realidad del mercado, al incorporar un análisis visual detallado del estado actual de las viviendas.

Presentación del Conjunto de Datos

El proyecto final se centra en la predicción de precios de viviendas en el estado de California, Estados Unidos. Para llevar a cabo este análisis, se ha escogido realizar web scraping en el portal inmobiliario Zillow. Esta decisión se tomó debido a la accesibilidad de datos históricos de viviendas vendidas en EE.UU., lo cual no es común en España.

Además de obtener una base de datos a través de web scraping, se utilizan imágenes de cada una de las viviendas, complementadas con un dataset adicional extraído de la web Realtor.com que facilita datos de mercado. Otro dataset utilizado proviene de la página Open Data California, incluyendo información sobre ciudades y condados de California.

Este enfoque multidimensional permite un análisis exhaustivo y detallado del mercado inmobiliario en California.

Características Generales del Conjunto de Datos

El conjunto de datos utilizado para este proyecto se distingue por su multidimensionalidad y diversidad de fuentes, cada una aportando perspectivas únicas y valiosas sobre el mercado inmobiliario en California. La compilación de datos abarca distintos aspectos relevantes para una análisis inmobiliario comprensivo:

- **Información textual y características físicas:** Incluye detalles sobre propiedades vendidas, como la ubicación, número de habitaciones y baños, y el área en metros cuadrados. Esta información fundamental proporciona una base sólida para la valoración de las propiedades.
- **Imágenes de las propiedades:** Las fotografías de las viviendas ofrecen una dimensión visual crítica, permitiendo análisis más profundos sobre el estado y atractivo de las propiedades, lo que puede influir significativamente en su valoración.
- **Datos de mercado:** Se incorporan datos adicionales sobre el mercado inmobiliario, como el precio medio por condado, lo que enriquece el análisis y ayuda a contextualizar las valoraciones en el panorama más amplio del mercado.
- **Lista de ciudades y condados de California:** Este dataset sirve como puente entre las distintas fuentes de datos, permitiendo correlacionar la información sobre las propiedades con datos de mercado relevantes a nivel de condado.

La elección de integrar el dataset de Realtor.com fue estratégica, motivada por la limitada cantidad de listings obtenidos a través del web scraping en Zillow, que resultó en un total de 324 propiedades. Al enfrentar este volumen bajo de datos, la inclusión del dataset de Realtor.com, que ofrece el precio medio por condado, se convirtió en una solución clave para complementar y enriquecer la información disponible, especialmente en aquellos condados con un menor número de propiedades listadas.

Adicionalmente, el desafío de correlacionar los datos del web scraping, que detallaban principalmente la ciudad sin especificar el condado, con la información de precios medios por condado de Realtor.com, llevó a la incorporación de otro dataset crucial: el de ciudades y condados de California.

Este dataset funcionó como un eslabón esencial, permitiendo la conexión entre los datos de las propiedades y los precios medios por condado a través de la variable "ciudad", facilitando así una integración fluida y efectiva de las distintas fuentes de datos.

Esta meticulosa selección y combinación de datasets subrayan la complejidad y la riqueza del conjunto de datos compilado para el proyecto. Al abordar los desafíos de integración de datos de diversas fuentes, el proyecto no solo logra una comprensión más profunda y detallada del mercado inmobiliario en California, sino que también establece una base sólida para la predicción precisa de precios de viviendas, aprovechando la riqueza de información numérica, categórica y visual.

Presentación de los Objetivos

El proyecto tiene como objetivo integrar datos de viviendas vendidas con un análisis visual para desarrollar un modelo predictivo que incorpore tanto características cuantitativas como cualitativas de las propiedades. Esto incluye:

Desarrollar un modelo predictivo: Que considere no solo los datos numéricos y categóricos tradicionales, sino también el análisis de imágenes para una valoración más precisa del precio de las viviendas.

Precisión en la predicción del valor de la vivienda: Mejorar la precisión de las estimaciones de precios mediante la incorporación de un análisis detallado del estado visual de las propiedades.

Definición de las Variables

Las principales variables contenidas en el conjunto de datos incluyen:

- **Address**: Ubicación exacta de la vivienda.
- **Price**: Valor de venta de la propiedad.
- **Beds**: Número de habitaciones en la vivienda.
- **Baths**: Número de baños en la propiedad.
- **Square Feet**: Metros cuadrados de la propiedad.
- **Sold Date**: Cuándo se vendió la propiedad.
- **Image**: Fotografías de las fachadas y otros aspectos relevantes de las viviendas.
- **average_listing_price**: Información adicional sobre el mercado inmobiliario, como demanda y oferta en áreas específicas.
- **City & County**: Información geográfica detallada sobre la ubicación de las propiedades.

Variables que pertenecen al conjunto de datos pero que finalmente no se han utilizado:

- **month_date_yyyymm:** La fecha en el formato AAAAMM que representa el mes y el año.
- **state**: El nombre del estado.
- **state_id:** Un identificador único para cada estado.
- **median_listing_price:** El precio medio de venta de las viviendas en el estado especificado.
- **median_listing_price_mm:** La variación porcentual intermensual del precio medio de venta.
- **median_listing_price_yy:** La variación porcentual interanual del precio medio de venta.
- **active_listing_count:** El recuento de anuncios activos en el estado especificado.
- **active_listing_count_mm:** El cambio porcentual intermensual en el recuento de anuncios activos.
- **active_listing_count_yy**: Variación porcentual interanual del número de anuncios activos.
- **median_days_on_market**: El número medio de días que una propiedad está en el mercado antes de ser vendida.
- **median_days_on_market_mm:** Variación porcentual intermensual de la mediana de días en el mercado.
- **median_days_on_market_yy:** La variación porcentual interanual de la mediana de días en el mercado.
- **new_listing_count:** El recuento de nuevos listados en el estado especificado.
- **new_listing_count_mm**: El cambio porcentual intermensual en el recuento de nuevos anuncios.
- **new_listing_count_yy:** La variación porcentual interanual en el recuento de nuevos anuncios.
- **price_increased_count:** El recuento de listados en los que el precio ha aumentado.
- **price_increased_count_mm:** El cambio porcentual mes a mes en el recuento de precios incrementados. price_increased_count_yy: La variación porcentual interanual en el recuento de precios incrementados.
- **price_reduced_count:** El recuento de listados en los que se ha reducido el precio.
- **price_reduced_count_mm**: El cambio porcentual mes a mes en el recuento de precios reducidos.
- **price_reduced_count_yy:** La variación porcentual interanual en el recuento de precios reducidos.
- **pending_listing_count**: El recuento de listados que están pendientes (bajo contrato).
- **pending_listing_count_mm:** La variación porcentual intermensual en el recuento de anuncios pendientes.
- **pending_listing_count_yy:** Variación porcentual interanual del número de anuncios pendientes.
- **median_listing_price_per_square_foot**: El precio medio de venta por pie cuadrado.

- **median_listing_price_per_square_foot_mm:** La variación porcentual intermensual del precio medio de venta por pie cuadrado.
- **median_listing_price_per_square_foot_yy:** La variación porcentual interanual del precio medio de venta por pie cuadrado.
- **median_square_feet:** El tamaño medio de las viviendas anunciadas.
- **median_square_feet_mm:** La variación porcentual intermensual de la mediana de pies cuadrados.
- **median_square_feet_yy:** La variación porcentual interanual de la mediana de pies cuadrados.
- **average_listing_price:** El precio medio de venta de las viviendas.
- **average_listing_price_mm:** El cambio porcentual en el precio promedio de cotización con respecto al mes anterior.
- **average_listing_price_yy**: El cambio porcentual en el precio promedio de cotización con respecto al mismo mes del año anterior.
- **total_listing_count**: El total de listados activos y pendientes dentro de la geografía especificada durante el mes especificado. Esta es una medida instantánea de cuántos listados totales se pueden esperar en un día determinado del mes especificado.