Problem Set 2 – Ganancias con ML

Julio 26,2022

Andrea Margarita Beleño, María Valeria Gaona Guevara

*Abstract*—El precio de las viviendas cuenta con diferentes factores que se tienen en consideración para establecer un valor. Por lo tanto, contar con un modelo en el que se identifiquen dichas características principales es fundamental para poder generar una predicción lo más óptima posible. Por lo tanto, en el siguiente documento se realizará el modelo de predicción para los precios de las viviendas por medio del algoritmo Super Learner. El link al Github del presente taller, se encuentra en el siguiente enlace: https://github.com/mvgaona/Problem-Set-3

# Introducción

El precio de una vivienda está dado por diferentes factores como el área total, el número de baños, número de habitaciones, la distancia a puntos importantes de la ciudad, entre otros elementos. Además, cada ciudad cuenta con diferentes características que aportan a que los vendedores decidan establecer un precio. Sin embargo, se puede generar un modelo en donde se pueda observar el efecto de cada una de esas características esenciales que da el valor de un inmueble y así, tanto el vendedor como el comprador puedan realizar futuras compras y/o ventas de inmuebles.

Por otra parte, generar una predicción de estos valores es una herramienta para conocer el comportamiento económico de una sociedad y además, cada individuo pueda obtener información valiosa para la toma de decisiones a mediano o largo plazo. De acuerdo con lo anterior, en el presente documento se presenta el proceso de la limpieza de datos pertinente para realizar la caracterización de las variables fundamentales y así, realizar el modelo de predicción de los precios de las viviendas en la localidad de Chapinero en la ciudad de Bogotá y en el área del poblado en Medellín en Colombia.

# Datos

El precio de una vivienda puede estar dado por diferentes factores, tanto económicos como sociales. En el caso Colombiano, estos precios pueden estar influenciados por elementos propios de la vivienda y a su vez, por elementos geoespaciales, como la distancia a diversos lugares públicos (Bares, transporte público, etc). Para realizar un modelo de predicción de precios de la vivienda, es necesario contar con las variables determinantes y relevantes del precio, para que este modelo sea robusto, pero no se incurran en gastos que entorpezcan esta labor. En el presente modelo de predicción se tomó en cuenta solo los datos de las viviendas ubicadas en las localidades objetivo (Chapinero y Poblado) y el total de datos (Base de entrenamiento y base de testeo) es de 24843 viviendas.

Dentro del proceso de recolección de datos puede ocurrir que no se encuentre toda la información necesaria, ya sea porque la base de datos no provee dicha información o, por el contrario, no se pudo recolectar toda la información de cada vivienda. Es por eso que es necesario realizar un proceso de limpieza y recuperación de datos. Se realizaron diversos procesos para poder contar con toda la información:

1. Se cuenta con variables las cuales ya tienen la información completa, por lo tanto no es necesario realizar ningún tratamiento:

* Ubicación: La ubicación del inmueble en el modelo será esencial para el análisis ya que se analizarán dos localidades de dos ciudades diferentes y los precios de las ciudades si bien están dados por características similares, el costo de vida por ciudad también influye en el valor de la vivienda. Esta variable es categórica, y se tiene que la mayoría de datos son de la ciudad de Bogotá (ver Cuadro 1).

|  |  |
| --- | --- |
| **Ubicación** | |
| **Bogotá D.C** | **Medellín** |
| 14244 (57,34%) | 10599 (42,7%) |

* Tipo de propiedad: Si la propiedad es un apartamento o si es una casa influye sustancialmente en el precio, ya que el área de una casa suele ser más grande, cuenta con más oportunidad de reformas y espacios más amplios de esparcimiento. Esta variable categórica cuenta con una mayor proporción de apartamentos (Ver Cuadro 2).

|  |  |
| --- | --- |
| **Tipo de vivienda** | |
| **Casa** | **Apartamento** |
| 2038 (8,2%) | 22805 (91,8%) |

1. Para algunas variables existentes en la base de datos, se generó la comparación entre dicha información y los valores hallados por medio de la inspección de la descripción de cada inmueble, generando las siguientes variables sin información faltante:

* Habitaciones: El número de habitaciones de la vivienda es determinante en el precio de la misma, ya que se puede contar con un aproximado del espacio y de cuántos individuos pueden vivir, es decir, entre más habitaciones, el precio del inmueble tiende a incrementarse. Por lo tanto, dentro del análisis descriptivo de este predictor se encuentra que es una variable numérica, con media de 3 habitaciones (Ver Cuadro 3).

|  |  |  |  |
| --- | --- | --- | --- |
| **Habitaciones** | | | |
| **Min** | **Media** | **Max** | **Moda** |
| 1 | 3 | 11 | 3 |

* Cantidad de baños: Este es otro factor de decisión importante en el precio del inmueble, ya que sin baños, los individuos no pueden satisfacer las necesidades básicas de aseo. Por lo tanto es necesario contar con al menos un baño y con ello, se puede identificar que a medida que aumentan la cantidad de baños, el precio de igual manera se verá afectado e incrementará. De acuerdo con lo anterior, se observa que en la variable numérica, la media fue de 3 baños (ver Cuadro 4).

|  |  |  |  |
| --- | --- | --- | --- |
| **No. Baños** | | | |
| **Min** | **Media** | **Max** | **Moda** |
| 1 | 3 | 13 | 2 |

1. De acuerdo con la información proporcionada por la descripción de cada inmueble, se generaron las siguientes variables:

* Ascensor: Si el inmueble cuenta con ascensor o no es determinante en el caso de los apartamentos de pisos altos, es por eso que se considera un predictor importante del modelo, ya que quienes viven en pisos más altos, buscan apartamentos donde tenga ascensor y a su vez, el precio aumenta si el inmueble cuenta con este. Dentro del análisis descriptivo de la variable, es posible identificar que es una variable dicótoma, donde 1 corresponde a si el inmueble cuenta con ascensor y 0 si no (Ver Cuadro 5)

|  |  |
| --- | --- |
| **Ascensor** | |
| **Tiene ascensor** | **No tiene ascensor** |
| 19388 (78,04%) | 10599 (21,96%) |

Los inmuebles que no cuentan con ascensor también pueden ser apartamentos que no cuentan con uno y casas que no lo necesitan.

Parqueadero: En la actualidad, gran parte de los individuos cuentan con uno o más carros, por lo tanto, es necesario que el inmueble cuente con al menos un parqueadero, por lo tanto, si el inmueble incluye al menos un garaje, el precio de la vivienda tenderá a aumentar su valor. Por otra parte, de acuerdo con el análisis descriptivo, se identifica que esta es una variable dicótoma, en donde 1 hace referencia a que la casa o el apartamento cuenta con al menos un parqueadero y 0 que no lo hace. La proporción de inmuebles que cuentan con al menos un parqueadero es que la mayoría de inmuebles de la base no cuenta con parqueaderos (ver cuadro 6) .

|  |  |
| --- | --- |
| **Parqueadero** | |
| **Tiene parqueadero** | **No tiene parqueadero** |
| 7885 (31,74%) | 16958 (68,26%) |

Esta proporción puede estar dada debido a que algunos vendedores no ofrecieron dicha información, por ejemplo.

1. De acuerdo con la ubicación geoespacial, se pueden recuperar datos por medio de la búsqueda de características similares de otras viviendas y dicho promedio, adjuntarlo al inmueble que no cuenta con información, ya que al estar en un perímetro cercano, los inmuebles suelen presentar características similares.

* Área: El área de un inmueble es fundamental para conocer el precio del mismo, ya que a medida que el área aumente, también el precio. Esto se da debido a que un espacio amplio permite contar con un mayor número de habitaciones, baños, una cocina más amplia y demás espacios de esparcimiento. Además, permite más posibilidades de remodelación y de inversión. Por lo tanto, ante la variable numérica presentada, se encuentran que se tiene una media de 286 m2. Para más información (ver Cuadro 7)

|  |  |  |  |
| --- | --- | --- | --- |
| **Área (Mts2)** | | | |
| **Min** | **Media** | **Max** | **Moda** |
| 74,08 | 286,75 | 3940.14 | 193.7414 |

* Por medio de la ubicación geoespacial, las variables adicionales que se encontraron y que consideramos relevantes para la predicción, fueron las siguientes: Transporte público: Contar con al menos un transporte público cerca de la vivienda es fundamental para poder desplazarse con facilidad. Por lo tanto, esta variable presenta la distancia mínima que tiene el inmueble con al menos un medio de transporte público, ya que a medida que esta distancia sea más corta, el precio puede incrementarse. Para esta variable se cuenta con una media de 1,5km (Ver Cuadro 8)

|  |  |  |  |
| --- | --- | --- | --- |
| **Distancia TP (Mts2)** | | | |
| **Min** | **Media** | **Max** | **Moda** |
| 4,397 | 1519,834 | 4472,144 | 2168,091 |

* Bares: La distancia mínima a bares es otro factor a tener en cuenta, ya que, en general, este tipo de establecimiento hace que pueda generar incomodidad a los habitantes en las horas de la noche. Por lo tanto, al estar cada vez más alejados de un bar, el comprador puede estar más interesado en el inmueble. De acuerdo con la variable numérica, se presenta que la media de distancia se encuentra en 694m (Ver Cuadro 9).

|  |  |  |  |
| --- | --- | --- | --- |
| **Distancia bares (Mts2)** | | | |
| **Min** | **Media** | **Max** | **Moda** |
| 2,101 | 694,033 | 3062,209 | 487,8808 |

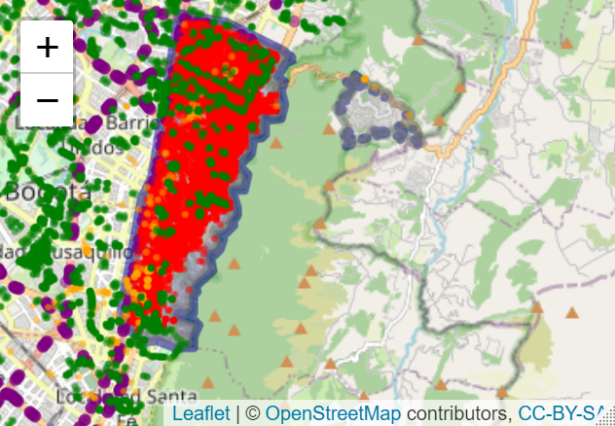
* Parques: La distancia mínima a por lo menos un parque es otro predictor relevante dentro del modelo, ya que, en general, los individuos buscan tener un espacio verde y de recreación cerca, ya sea porque dentro del hogar tienen niños los cuales puedan entretenerse, mascotas que necesiten espacio libre para derrochar su energía o cumplir sus necesidades fisiológicas o porque los compradores también desean un espacio al aire libre para su propio entrenamiento. Por lo tanto, a medida que esta distancia va disminuyendo, el precio del inmueble tiende a incrementar. La siguiente variable categórica presenta la siguiente descripción estadística:

|  |  |  |  |
| --- | --- | --- | --- |
| **Distancia parques (Mts2)** | | | |
| **Min** | **Media** | **Max** | **Moda** |
| 0,4975 | 230,5497 | 1567,7500 | 478,1448 |

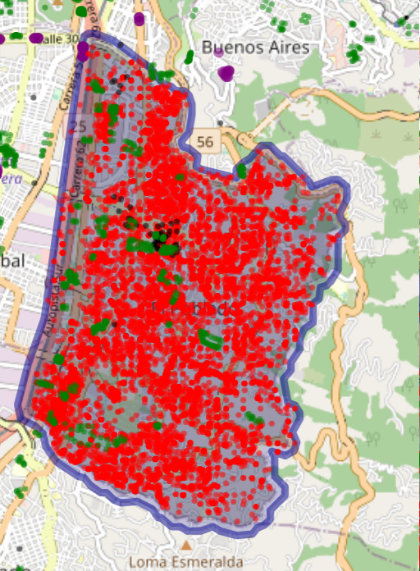
# MAPAS

Los siguientes mapas de la localidad de Chapinero en Bogotá y El Poblado en Medellín, evidencian la distribución de bares, estaciones de buses, y parques que se encuentran en cada uno de los sectores, junto con las viviendas que se encuentran en las localidades anteriormente mencionadas. Esto permite contar con un análisis gráfico de cada vivienda y así, un comprador, por ejemplo, puede contar con información gráfica que le permite tomar decisiones de manera informada.

Localidad Chapinero, Bogotá:



Comuna 14- El Poblado, Medellín:



En donde el polígono década localidad está limitado por medio del color azul oscuro, las estaciones de buses en círculos morados, los apartamentos en círculos rojos, parques en círculos verdes y finalmente, los bares en anaranjado en Bogotá y negro en Medellín.

# MODELO Y RESULTADOS

# Conclusiones y recomendaciones

# REFERENCIAS

XXXXXXX

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**