Problem Set 3: Making Money with ML?

Big Data y Machine Learning para Economía Aplicada

Gustavo Adolfo Castillo Álvarez (201812166),
Alexander Almeida Ramírez (202225165),
Jorge Luis Congacha Yunda (201920042) y
Jaime Orlando Buitrago González (200612390)

2024-05-26

1 Introducción

En su *paper*, Rosen (1974) proporciona una aproximación metodológica (econométrica) para estimar precios hedónicos. Más específicamente, explora los equilibrios del mercado y las decisiones de los agentes y los efectos en el bienestar para evidenciar la relación entre los precios de los bienes y sus diferentes características. Esta fue una contribución importante para estimar los precios (Hedónicos) de los bienes a partir de las características o atributos (precios implícitos) a partir de la observación de los precios de bienes heterogéneos sus características asociadas.

Como uno de los *papers* más citados del *Journal of Political Economy* (Greenstone, 2017), se obtiene una aproximación al comportamiento de la oferta y demanda (equilibro) a partir del proceso generador de datos entre los bienes y sus características. Para el mercado inmobiliario -en el que los comportamientos de los agentes, oferta y demanda, y equilibrios no se pueden explicar en su totalidad con la teoría microeconómica convencional-, esta aproximación ha sido de singular importancia, más aún en el contexto de la ciudad de Bogotá.

Así por ejemplo, Perdomo (2011) utiliza los precios hedónicos cómo método econométrico para estimar el precio de las viviendas y posteriormente el de la cercanía a Transmileno; Carriazo et al. (2013) identifican cómo de no incluir la calidad del aire en las estimación a través de precio hedónicos conlleva un sesgo en los precios de las viviendas en Bogotá. Lozano & Anselin (2012) utilizan variables espaciales en los submercados de la viviendas para estimar los precios en Bogotá, identificando una gran capacidad predictiva, pero afectada (sobre estimación) por los estratos socioeconómicos. También se encuentran Medina et al. (2007) quienes demuestran como los subsidios aplicados a las tarifas de los servicios públicos se transfieren al precio de las viviendas.

Esta breve exploración bibliográfica refleja los usos y potencialidades de la contribución de Rosen (1974) al mercado inmobiliario con énfasis en la ciudad de Bogotá. Claro está, como la capital de Colombia, Bogotá es una ciudad heterogénea con dinámicas poblacionales, sociales y económicas distribuidas a por todo el territorio urbano, las cuales pueden afectar los precios del mercado inmobiliario. Más aun, en una localidad como Chapinero, la cual es una de las áreas centrales de la ciudad; con una gran catnidad de población flotante; donde se cuenta con vías que se conectan con el sistema de movilidad de la ciudad; se tienen usos del suelo asociados a actividades comerciales y residenciales; y se tienen parques, canales y corredores ecológicos (Decreto 468, 2006).

Por estas características, tal vez Chapinero sea la localidad con la mayor complejidad de formas de habitar el territorio de la ciudad del país, lo cual la hace un espacio geográfico donde convergen múltiples factores que afectan los precios de las viviendas. Por tal razón, construir uno o varios modelos predictivos de los precios es desafiante y una necesidad, ante una potencial inversión que compre la mayor cantidad de predios, gastando lo menos posible.

Ahora bien, para construir el o los modelos se utilizó una muestra de los anuncios publicados en la página web Properati. Este sitio web de origen argentino y con presencia en varios países de Latinoamérica, le permite a propietarios publicar un bien inmueble con la mayor cantidad de información (precio, características y amenidades, ubicación y descripción), y a personas interesadas en adquirir o arrendar un bien contactarse con el comprador. Es decir, se utilizó una página web que ofrece servicios de intermediación, pero que permite caracterizar la oferta de bienes en Bogotá.

De igual manera, se utilizó la información de Open Street Map (transporte público, comercio, equipamentos colectivos, y vías y trasnporte público), la cual permitió asociar información geográfica relevante a la muestra de Properati. Así como se utilizó también la información de la página web Infraestructura de Datos Espaciales de Bogotá (Ideca), un portal de libre acceso y consulta de la Unidad Administrativa Especial de Catastro Distrital (UAECD), de donde se utilizaron las bases de datos (capas) de las localidades de Bogotá

Con estas fuentes de información se construyeron diferentes modelos para predecir el precio de las viviendas en Chapinero. Estos modelos aplican los conceptos abordados en la clase *Big Data y Machine Learning para Economía Aplicada*, los cuales se describen con mayor detalle en las siguientes secciones. En las siguientes secciones adicionalmente se encuentran: la presentación de las bases de datos utilizadas, presentando a su vez estadísticas descriptivas que facilitan su comprensión; los resultados obtenidos luego de su impelmentación; y unas conclusiones de todo el ejercicio.

2 Datos y Métodos

Los datos principales consistieron en los archivos descargados de la página de datos de la Competencia en Kaggle. Esta base de datos contiene 48.930 observaciones, distribuidas en 10.286 observaciones para prueba (21%) y 38.644 para entrenamiento (79%). Cada observación representa un inmueble para venta publicado en Properati entre 2019 y 2021 en la ciudad de Bogotá. Dentro de las características de cada inmueble se tienen las áreas (construida y cubierta), el número de habitaciones y baños, el tipo de propiedad (casa o apartamento), y el título y descripciónn con el que se hizo la publicación en la página web y localización (latitud y longitud). Dada la estructura de la base de datos es muy probable que se haya construido a través de un ejercicio de *Web Scrapping* y luego se haya dispuesto para su descarga y uso en el Problem Set 3.

A partir de las descripciones, se utilizó el análisis de texto para constrastar el tipo de inmueble publicado con el que se encuentra en la descripción, extraer el número de pisos de las casas, e identificar el piso en el que se encuentran los apartamentos. Como se tienen todos los inmuebles para Bogotá, en función de su localización (latitud y longitud), se extrajeron aquellas casas y apartamentos dentro de la localidad de Chapinero. Este ejercio se realizó mediante la extracción los de puntos (capa construida con la latitud y longitud) dentro del polígono de Cahpinero (capa del Ideca).

Utilizando el mismo método de análisis espacial, se utilizaron las capas de Open Street Maps para medir las distancias del inmueble a gimnasios, estaciones de Transmilenio, bares, supermercados, colegios, hospitales, principales avenidas, centros comerciales y universidades.

Como resultado se encontraron 536 observaciones que difieren entre sus descripción (1,2% en la base de entrenamiento y 0,6% en la de prueba)

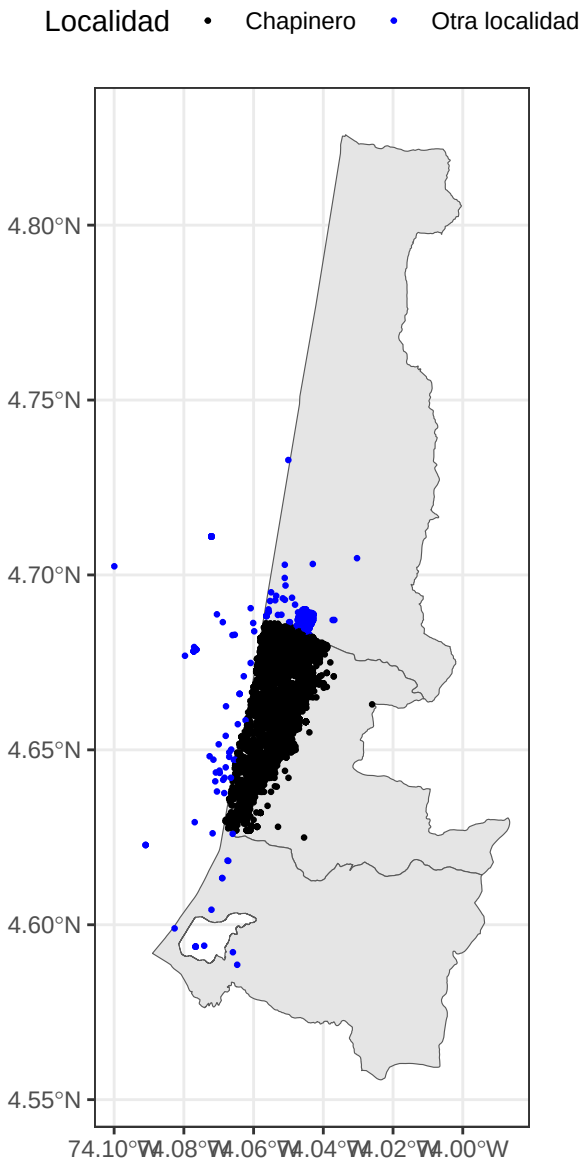
Variable	Tipo	Promedio	Desv. est.	Mín	Máx.
Precio (millones)	Apartamento	834.2	377.2	300.0	1600.0
Precio (millones)	Casa	899.3	368.6	358.0	1650.0
Área cubierta	Apartamento	127.1	44.6	31.0	505.0
Área cubierta	Casa	168.5	52.2	70.0	450.0
Área total	Apartamento	125.1	49.4	15.0	750.0
Área total	Casa	540.0	6293.1	32.0	108800.0
Habitaciones	Apartamento	2.4	0.8	1.0	10.0
Habitaciones	Casa	3.5	1.5	1.0	11.0
Baños	Apartamento	2.6	0.9	1.0	6.0
Baños	Casa	3.2	1.1	1.0	7.0
Dromitorios	Apartamento	2.3	0.9	0.0	10.0
Dromitorios	Casa	3.9	2.2	0.0	11.0
Número de pisos	Apartamento	1.0	0.0	1.0	2.0
Número de pisos	Casa	1.2	0.6	1.0	6.0
Piso	Apartamento	2.1	3.1	1.0	60.0
Piso	Casa	1.1	0.6	1.0	8.0
Dist. parque	Apartamento	158.9	93.4	4.7	1730.0
Dist. parque	Casa	192.2	144.9	10.0	1948.5
Dist. Transmilenio	Apartamento	805.5	477.3	0.3	2568.5
Dist. Transmilenio	Casa	658.6	421.9	21.5	2968.8
Dist. bar	Apartamento	1152.9	468.2	1.5	2297.6
Dist. bar	Casa	1026.1	513.9	35.5	1997.2
Dist. colegio	Apartamento	483.3	300.9	0.0	1619.2
Dist. colegio	Casa	415.0	285.2	0.0	2271.6
Dist. avenida	Apartamento	231.4	228.7	0.0	1826.6
Dist. avenida	Casa	194.1	217.7	1.2	2074.1
Dist. universidad	Apartamento	731.8	408.5	0.0	1696.1
Dist. universidad	Casa	516.3	458.0	4.6	2768.8
Dist. gimnasio	Apartamento	730.0	403.0	10.2	2417.1
Dist. gimnasio	Casa	734.1	384.5	19.5	2685.6
Dist. CAI	Apartamento	561.4	313.1	3.8	1680.2
Dist. CAI	Casa	523.4	293.6	26.8	1440.9
Dist. supermercados	Apartamento	534.3	289.4	0.0	2142.9
Dist. supermercados	Casa	438.9	279.7	16.8	1746.8
Dist. hospitales	Apartamento	835.9	414.0	0.0	2120.0
Dist. hospitales	Casa	644.8	428.3	0.0	3408.1
Dist. centro comercial	Apartamento	247.0	187.5	1.7	2260.3
Dist. centro comercial	Casa	245.3	222.0	1.1	2428.6

Cuadro 1: Estadísticas descriptivas

La información espacial de referencia se tomó del Mapa de Referencia para Bogotá D.C. del Ideca disponible aquí. Al cruzar la capa de los polígonos las localidades con la tabla de *test* observamos

que 593 propiedades se encuentran fuera de Chapinero (Figura 1), que es la localidad en la cuál se predecirán los precios de las viviendas.

Figura 1: Ubicación datos de validación



Características de la propiedad:

- Precio
- Fecha de publicación de clasificado
- Superficie total y superficie cubierta (80% y 76% valores faltantes)
- Número de habitaciones
- Número de baños
- Tipo de propiedad: apartamento o casa

- Ubicación georreferenciada

Spatial Ammenities: - parques - tiendas o centros comerciales - # Modelo y resultados

En esta sección se exponen los modelos que se utilizaron para predecir los precios de la vivienda en la localidad de Chapinero, las variables utilizadas en los modelos y la selección de hiperparámetros. En particular, se utilizaron los modelos de bosques, bosques aleatorios, regresión lineal, elasticnet y boosting para realizar la predicción. Finalmente, se presenta una comparación del rendimiento de estos modelos. El mejor resultado se obtuvo con el modelo de Random Forest utilizando como métrica el Mean Absolute Error metric (MAE).

2.1 Descripción de variables

Para todos los modelos utilizados para la predicción, se utilizó el siguiente modelo:

escribir modelo.

Las variables utilizadas para predecir el precio de las viviendas se pueden dividir en dos grupos. El primero, corresponde a variables que capturan atributos estructurales de las viviendas como el número de habitaciones, el número de baños, la superficie total y el tipo de propiedad. Estas variables El segundo grupo está conformado por aspectos del vencidario como

2.2 Descripción del modelo

2.3 Comparativa con otros modelos

3 Conclusiones

El presente estudio realizó un

4 Referencias bibliográficas

Alcaldía Mayor de Bogotá. (20 de noviembre de 2006). *Por el cual se reglamenta la UPZ 99, Chapinero, ubicada en la loclaidad Chapinero* (Decreto 468 de 2006, Ed.).

Carriazo, F., Ready, R., & Shortle, J. (2013). Using stochastic frontier models to mitigate omitted variable bias in hedonic pricing models: A case study for air quality in Bogotá, Colombia. *Ecological Economics*, *91*, 80-88.

Greenstone, M. (2017). The Continuing Impact of Sherwin Rosen's «Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition». *Journal of Political Economy*, *125*(6), pp. 1891-1902.

Lozano, n., & Anselin, L. (2012). Is the price right?: Assessing estimates of cadastral values for Bogotá, Colombia. *Regional Science. Policy & Practice*, *4*(4), 495-508.

Medina, C., Morales, L., Bernal, r., & Torero, M. (2007). Stratification and Public Utility Services in Colombia: Subsidies to Households or Distortion of Housing Prices? [with Comments]. *Economía*, *7*(2), 41-99.

Perdomo, J. (2011). A methodological proposal to estimate changes of residential property value: case study developed in Bogotá. *Applied Economics Letters*, *18*(16), 1577-1581.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34-55.

5 Repositorio

El respositorio se puede consultar en el siguiente enlace:

- Problem Set 3: Making Money with ML?