Problem Set 3: Making Money with ML?

Universidad de los Andes
Maestría en Economía Aplicada
Andres Felipe Martinez, Angela Paola Morales Guio y Oscar Cortes

Repositorio: https://github.com/paolamguio/Problem_Set_3_G16

I. Introducción

El modelo planteado para estimación de precios de la vivienda se basa en las variables más determinantes y utilizadas en Colombia, como el área medida en metros cuadrados, el estrato puesto que de acuerdo con la metodología definida por el DANE esta clasificación depende del tipo de vías de acceso, puntos de transportes e infraestructura social como colegios, comercio y parques. Después de probar varios modelos se seleccionó XGBoost reduciendo alrededor de un 30% el RMSE frente a otros modelos como OLS, Lasso, Ridge y Elasticnet.

Los resultados de nuestro modelo de predicción de precios de vivienda han despertado el interés de Habi (empresa nueva que se especializa en la comprar y venta de vivienda en Colombia) por la simplicidad, actualización continúa y precisión en la estimación del costo de la vivienda a partir de características de esta, su ubicación, cercanía a vías de acceso, medios de transportes y de esparcimiento como parques. A continuación, presentamos las principales ventajas de nuestro modelo de predicción y las limitaciones:

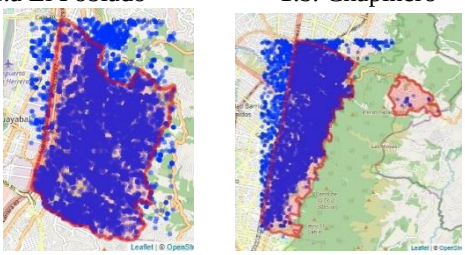
II. Datos

Por medio de una muestra tomada de Properati sobre los precios de venta y características de inmuebles ubicados en la localidad de Chapinero en Bogotá y en la Comuna 14 El Poblado en Medellín, se tiene una muestra representativa de entrenamiento y testeo del precio de venta de las viviendas en estas zonas, así como de los principales atributos que definen el precio de mercado de las mismas. Se define tomar solamente información de estas zonas porque es en ellas en que se va a enfocar la predicción y, por lo tanto, es relevante entrenar los modelos con información de las mismas zonas para no alterar la predicción de los precios porque el precio de las viviendas puede ser muy diferente dependiendo de la zona en donde esté ubicada.

Los polígonos de análisis se obtienen de Open Street Map y se selecciona la información de la base de datos que corresponde a estos polígonos, la Gráfica 1 muestra la información para las observaciones train y test por cada zona.

Gráfica 1. Polígonos de estudio

1.a El Poblado 1.b. Chapinero



Fuente: Elaboración propia

Por medio de la descripción de las viviendas en venta, se recopiló información sobre atributos adicionales y que influyen en el precio de los inmuebles, se obtuvo variables adicionales de características físicas de si cuenta o no con garaje, ascensor, terraza y balcón, de igual forma, para las variables existentes que presentaban missing, se procedió a imputar información rescatada del texto sobre número de baños, número de habitaciones y área total. Adicionalmente, se obtuvo el estrato medio de las viviendas por zona extraído del Censo Nacional de Población y Vivienda - CNPV - 2018, y, se consideraron variables de distancia mínima entre los inmuebles y las zonas comerciales, y de esparcimiento (bares, restaurantes, zonas de parqueo y parques), zonas de estudio y de estaciones de bus, esta información se obtuvo de Open Street Map.

No obstante lo anterior, se continuó con missing en las variables de interés por lo que se obtuvo información del DANE de las manzanas de estas ciudades las cuales se incluyeron en el análisis de datos como polígonos, con el fin de calcular la

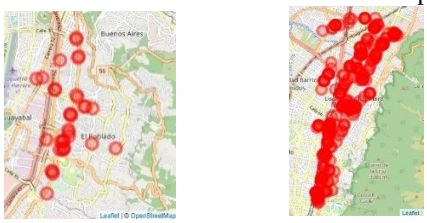
mediana de las variables área metros cuadrados, número de baños y estrato, e imputar los valores missing. Adicionalmente, se realizaron buffers por distancia para calcular la mediana de estas variables para imputar algunos missing values finales que quedaron en la base.

Se buscó obtener la información total de las variables mencionadas anteriormente, al considerarlas importantes para este estudio pues como menciona Rosen (1974), las características de los bienes describen a los bienes diferenciados, lo que quiere decir que estas características pueden explicar que tan diferente es el bien, para el caso del precio de las viviendas, estas variables pueden considerarse como un factor relevante que impacta este precio.

La Tabla 1 refleja un análisis descriptivo de las diferencias de los principales atributos entre las opciones de vivienda en Chapinero y en el Poblado, mostrando que, en promedio, en El Poblado los inmuebles cuentan con un área total mayor que en Chapinero, lo que coincide con un mayor número de habitaciones promedio en El Poblado, así mismo, en este último se cuenta con una mayor proporción de casas que en Chapinero. En promedio, en Chapinero existe menor distancia entre las unidades habitaciones al servicio de transporte público respecto a El Poblado, lo que se observa de forma más detallada en el Gráfico 2, esto, se da debido a que en Medellín el principal servicio de transporte público es el metro el cual tiene una infraestructura lineal específica, adicional a que El Poblado no cuenta con un sistema integrado de transporte lo que lleva a que no existan paraderos de transporte público definidos, mientras que en Bogotá es el Sistema Integrado de Transporte Público (SITP), lo que lleva a que existan múltiples paraderos definidos de este sistema.

Gráfico 2. Densidad de servicio de transporte público.

2.a Densidad El Poblado 2.b. Densidad Chapinero



Fuente: Elaboración propia

III. Modelos y resultados

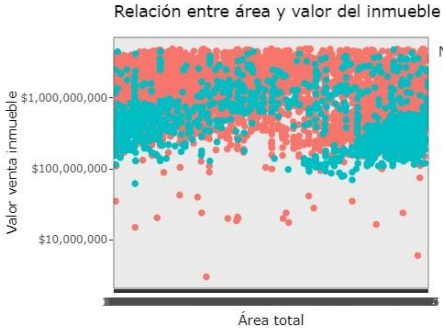
El modelo seleccionado para la predicción de los precios de la vivienda fue el XGBoost, cuyo RMSE fue 30% menor que los otros modelos que se evaluaron para la predicción como lo fueron OLS, Lasso, Ridge y Elasticnet.

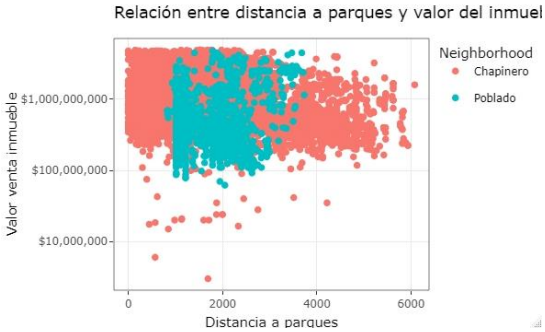
Las variables que se utilizaron para el entrenamiento del modelo principalmente fueron las de área en metros cuadrados del apartamento/casa y el estrato donde está ubicado el inmueble por ser las variables más relevantes para definir el precio en el país, dado que un referente de las ventas de inmuebles es el valor por metro cuadrado. Adicionalmente el DANE establece los lineamientos para definir las zonas de las ciudades por estratos de acuerdo con características de la zona como acceso de vías, infraestructura social como parques, colegios, comercios y zonas de diversión como bares. También se incluyeron otras variables para mejorar la predicción del precio de las viviendas como el número de habitaciones, número de baños, distancia a parques y bares.

Se utilizó una grilla de 100 por 250 en el modelo XGboost, con lo cual se obtuvieron predicciones rápidas y precisas. El valor de gama para la regularización, prever la sobreestimación, fue de 0.1. Se utilizaron un vector de profundidades de 4, 6 y 8 para que se mejorar el aprendizaje con los datos recolectados. Finalmente, el peso para los nodos para la clasificación y división fue de 10, 25 y 50.

Para comprobar empíricamente los modelos de predicción del precio de las viviendas, se utilizó una muestra de testeo con un total de 15160 observaciones para Chapinero y 1659 para El Poblado, con 16 atributos, se desarrollaron modelos OLS, Lasso, Ridge, Elasticnet y XGBoost, observando que el modelo de predicción de precios de viviendas que menor RMSE obtuvo fue el XGBoost con un valor de 486,466,506.

En este modelo indica que los atributos como: área total del inmueble, estrato y distancia a parques son las de mayor incidencia sobre el precio de las viviendas, relación que se refleja en los siguientes gráficos.


Relación entre área y valor del inmueble
Valor venta inmueble
$1,000,000,000
$100,000,000
$10,000,000
Área total


Relación entre distancia a parques y valor del inmuel
Valor venta inmueble
$1,000,000,000
$100,000,000
$10,000,000
0
2000
4000
6000
Distancia a parques
Neighborhood
Chapinero
Poblado

Anexo

Tabla 1. Estadísticas descriptivas generales

Descriptivas principales	Test Chapinero	Test Poblado	Train Chapinero	Train Poblado
N	7931	10357	15160	1659
Características inmuebles				
No. Habitaciones	1.91 (1.27)	3.02 (0.91)	2.67 (1.16)	3.10 (1.06)
No. Baños	1.86 (0.89)	3.23 (1.08)	3.04 (1.13)	2.65 (1.30)
Área Total	78 (69)	222 (3,092)	151 (890)	178 (311)
Parqueadero	471 (59%)	7,046 (68%)	10,311 (68%)	1,112 (67%)
Ascensor	266 (34%)	1,974 (19%)	3,437 (23%)	464 (28%)
Balcón	151 (19%)	4,857 (47%)	4,079 (27%)	889 (54%)
Terraza	321 (40%)	1,779 (17%)	5,281 (35%)	262 (16%)
Remodelado	66 (8.3%)	554 (5.3%)	2,122 (14%)	32 (1.9%)
Tipo de inmueble				
Apartamento	735 (93%)	8,923 (86%)	14,177 (94%)	1,176 (71%)
Casa	58 (7.3%)	1,434 (14%)	983 (6.5%)	483 (29%)
Distancia cercana a:				
Bares	99 (78)	932 (618)	529 (304)	739 (596)
Estaciones de bus	292 (146)	3,086 (1,279)	785 (474)	1,125 (1,159)
Bancos	84 (55)	560 (437)	300 (238)	892 (415)
Restaurantes	38 (44)	434 (325)	212 (179)	580 (429)
Colegios	243 (118)	470 (277)	436 (335)	1,017 (486)
Parques	4,156 (833)	1,534 (832)	1,515 (1,103)	1,833 (593)
Estrato				
1	0 (0%)	0 (0%)	58 (0.4%)	0 (0%)
2	6 (0.8%)	104 (1.0%)	172 (1.1%)	71 (4.3%)
3	445 (56%)	218 (2.1%)	678 (4.5%)	297 (18%)
4	342 (43%)	442 (4.3%)	1,880 (12%)	533 (32%)
5	0 (0%)	1,212 (12%)	1,962 (13%)	90 (5.4%)
6	0 (0%)	8,381 (81%)	10,410 (69%)	668 (40%)
Precio	NA (NA)	NA (NA)	1,285,603,208 (899,054,576)	666,671,055 (746,542,318)

1 Mean (SD); n (%)

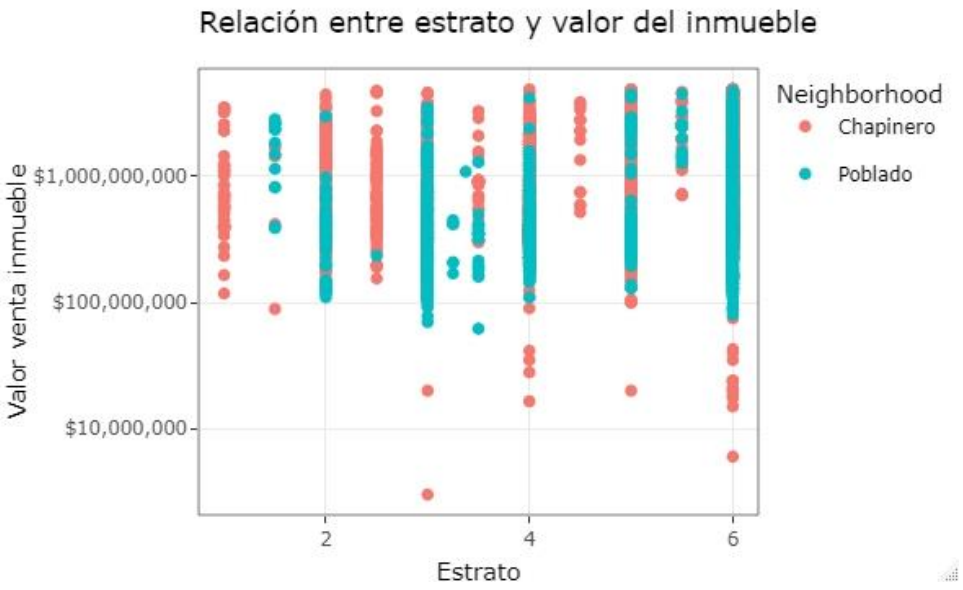


Tabla 2. Comparación RMSE de los modelos de predicción

Modelo	**RMSE**
OLS	672,664,869
Lasso	672,940,134
Ridge	672,754,758
Elasticnet	723,316,417
XGBoost	486,466,506

Tabla 3. Número de observaciones por sector

Bogotá D.C Chapinero	Medellín El Poblado
15160	1659

Tabla 4. Número de observaciones por estrato

1	2	3	4	5	6
58	353	1638	3197	3264	1949

Tabla 5. Estaditicos modelo XGBoost

Statistic	N	Mean	St. Dev.	Min	Max
.outcome	11,774	1,220,432,071.000	897,572,757.000	3,000,000	4,929,000,000
bedrooms	11,774	2.719	1.151	0	11
bathrooms	11,774	2.997	1.139	1.000	13.000
surface_total	11,774	155.968	1,012.379	0.000	108,800.000
dist_bar	11,774	547.804	345.441	0.000	3,270.333
dist_bus_station	11,774	814.797	580.991	0.000	5,600.505
dist_school	11,774	493.860	392.820	0.000	2,074.414
dist_park	11,774	1,542.229	1,069.473	0.000	5,936.106
dist_parks_total	11,774	181.361	175.752	0.000	2,983.775
parking	11,774	0.681	0.466	0	1
ascensor	11,774	0.235	0.424	0	1
balcon	11,774	0.294	0.456	0	1
terraza	11,774	0.331	0.470	0	1
remodelado	11,774	0.129	0.335	0	1
estrato	11,774	5.344	1.055	1.000	6.000

Tabla 6. Resultados modelo XGBoost

eta	max_depth	gamma	colsample_bytree	min_child_weight	subsample	nrounds	RMSE
0.3	8	1	0.7	25	0.6	250	494130166.1
0.3	8	0	0.7	10	0.6	100	494330104.1
0.3	8	0	0.7	25	0.6	250	495140748.3
0.3	8	0	0.7	10	0.6	250	495930971.4
0.3	6	1	0.7	10	0.6	250	496243604.7
0.3	6	0	0.7	10	0.6	250	496444647
0.3	8	1	0.7	50	0.6	250	496961887.1
0.3	6	1	0.7	25	0.6	250	497044548
0.3	8	1	0.7	25	0.6	100	497129193.7
0.3	8	1	0.7	10	0.6	100	497893073.5