Ronny Cruz
Catalina Esmeral Flórez
Federico Ramirez
Juan Camilo Sánchez
Link al repositorio: GitHub
Marzo 12, 2023

Problem Set 3: Making Money with ML?

I. Introducción

La formación de precios de las viviendas en Bogotá es un proceso complejo que está influenciado por diversos factores, tales como la ubicación. Las zonas más exclusivas y cercanas al centro de la ciudad suelen tener precios más elevados que las áreas más alejadas de este. Otro factor determinante es la oferta y la demanda de vivienda. Cuando existe una alta demanda de vivienda y poca oferta, los precios suelen aumentar, mientras que en situaciones de sobreoferta, los precios tienden a disminuir. Otros factores que pueden afectar el precio de una vivienda son la calidad de la construcción, el tamaño de la propiedad, las comodidades que ofrece y el plan de ordenamiento territorial (POT) de la ciudad. A partir de lo anterior, este trabajo busca predecir los precios de las viviendas en Bogotá, a partir de la información provista por el mercado inmobiliario en la ciudad. Los datos utilizados están relacionados con el número de habitaciones por vivienda, el área construida, los servicios disponibles para cada barrio, transporte, niveles de seguridad y acceso a medios de transporte y vías.

Para evitar los errores en las predicciones de Zillow, este trabajo utiliza modelos de predicción priorizando la degradación de los precios en el mercado, a través del reentrenamiento de los modelos con nueva información. Esto es, asignar mayor relevancia a los datos recientes que a los datos históricos. En consecuencia, se prueban seis (6) modelos de predicción que utilizan las técnicas de regresión lineal, regresión Lasso, Ridge, Elastic Net, Random Forest y XG Boost.

Dentro de los resultados se encontró que variables como área, número de habitaciones y baños, distancia más cercana a ríos, universidad, estaciones de policía, parques y hospitales son variables que influyen de manera significativa en la predicción de los precios de las viviendas en Bogotá. En adición, se pudo concluir que los modelos de Regresión Lineal, Lasso y Ridge tienen un bajo poder de predicción con datos espaciales, a comparación del modelo de Random Forest (rf) y Elastic Net.

A continuación, se presenta una sección de datos que describe la obtención y limpieza de estos, estadísticas descriptivas que contribuyen a la comprensión del problema y la justificación de los datos utilizados en las muestras de entrenamiento y testeo; una sección siguiente que describe los modelos estimados y sus resultados, una sección final con conclusiones y recomendaciones que detallan las bondades de los hallazgos y sus principales limitaciones y una sección de apéndice donde se amplía la descripción de los datos.

II. Datos

Los datos utilizados provienen de diversas fuentes, tales como el portal web Properati, Open Street Map, Censo Nacional y geoportal del DANE 2018. De Properati se extrajo información del precio de venta de las viviendas, área construida, número de habitaciones y baños, tipo de vivienda, coordenadas de ubicación del inmueble y descripción básica. Para obtener datos cuantificables de esta última variable, se utilizó la herramienta de análisis de texto (expresiones regulares) y con ello, se obtuvieron características de servicios adicionales que ofrece el inmueble que pueden tener incidencia en el precio, tales como la existencia de balcón, terraza, piscina, gimnasio, entre otros.

Open Street Map, por su parte, ofrece información sobre el acceso y/o cercanía que tienen las viviendas a supermercados, centros comerciales, estaciones de servicio de combustible, estaciones de policía, centros médicos, centros de entretenimiento, fuentes hídricas, universidades y medios de transporte público. Finalmente, del DANE se obtuvo información del estrato socioeconómico de las viviendas con datos a nivel de barrio y manzana censal.

Si bien estas variables permitieron un mayor análisis a la información de las propiedades en Bogotá, en su mayoría tenían datos faltantes (NAs) que podrían perturbar la robustez de los modelos. Para solucionar esto, se imputaron características con base en los atributos de los vecinos en un radio de 150 metros mediante la estrategia “queen”, la cual define los vecinos como las unidades espaciales que comparten un borde o un vértice. Este algoritmo logró agregar información relevante de manera satisfactoria; no obstante, el costo computacional fue elevado.

Una vez conformada la base de datos, se dividieron los datos en muestra de testeo, la localidad de Chapinero y muestra de entrenamiento, la ciudad de Bogotá. El objetivo de este análisis es predecir el precio de las casas en la localidad de Chapinero y comparar los resultados con el resto de la ciudad.

Luego de limpiar la base de testeo, se encontró que algunas variables como el número de baños, las comodidades adicionales, la existencia de garajes y de luz natural tenían algunas observaciones vacías, para lo cual se llenaron estos espacios (no más de veinte observaciones por variable) con las modas de la base. Gracias a esto, se pudo contar con las 10286 observaciones requeridas para la base de testeo.

En este primer gráfico se muestra la agrupación de los datos en la localidad de Chapinero y se concluye que existe una alta densidad poblacional en esta zona. Sin embargo, existen algunos datos atípicos que refieren observaciones de otras localidades. Este fenómeno debe tenerse en cuenta a la hora de estimar los modelos de predicción, dado que pueden alterar la naturaleza de la muestra.

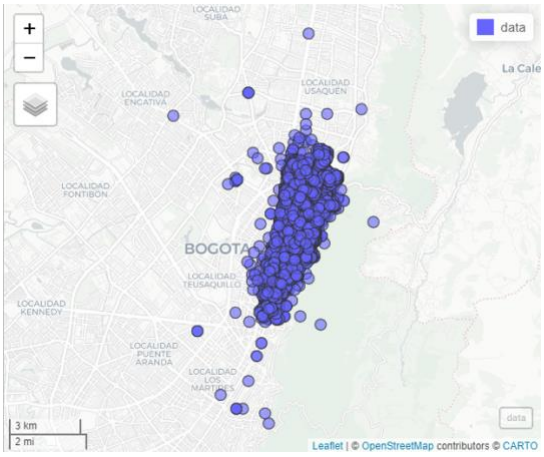


Ilustración 1. Mapa de elaboración propia con datos para Bogotá.

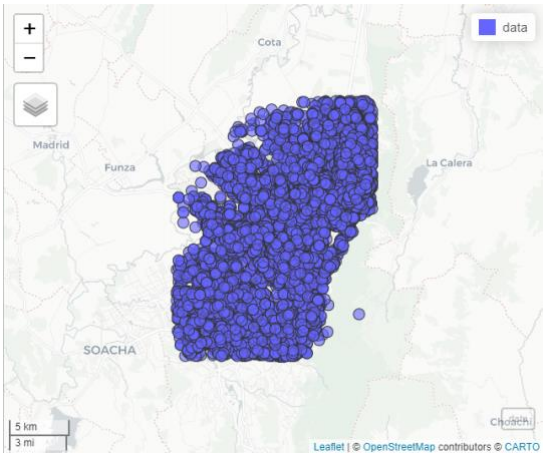


Ilustración 2. Mapa de elaboración propia con datos para Chapinero.

Por otro lado, en el gráfico dos se detalla la densidad habitacional de Bogotá, que presenta comportamientos similares a la muestra de testeo.

En la tabla 1 se detallan las principales medidas de tendencia central por variable para la muestra de entrenamiento y en la tabla 2 las medidas para la muestra de testeo. Algunos datos relevantes de ambos grupos son que el costo promedio de una vivienda en

Bogotá es de COP $654 millones, el promedio de habitaciones por vivienda en Bogotá es de 3, mientras que en Chapinero es 2. El promedio de área construida por vivienda es similar para ambos grupos de datos: 750 metros cuadrados, aproximadamente. La distancia promedio de las viviendas a una estación de policía es mayor en Bogotá que en la localidad de Chapinero; esta variable puede ser utilizada como *proxy* de acceso a la justicia en otro tipo de análisis y lo que nos permite concluir, en general, es que es menos accesible este nivel de "justicia" para el promedio de habitantes en Bogotá que en Chapinero.

Table 1: Estadísticas descriptivas datos vivienda Bogotá (train)

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
price	38644	654534675.291	311417886.954	3e+08	4.15e+08	8.1e+08	1.65e+09
month	38644	5.665	3.289	1	3	8	12
year	38644	2020.294	0.76	2019	2020	2021	2021
area M2	38548	795.73	2405.844	2	100	641.278	76610
Numero de habitaciones	38644	3.145	1.535	0	2	3	11
Numero de baños	38610	2.671	1.135	1	2	3	13
Tipo de vivienda (1 si es casa)	38644	0.245	0.43	0	0	0	1
Cuenta con sala o comedor	38624	0.953	0.212	0	1	1	1
Servicios interiores adicional	38616	0.905	0.293	0	1	1	1
Servicios exteriores adicional	38601	0.454	0.498	0	0	1	1
garage	38620	0.954	0.209	0	1	1	1
Luz natural	38593	0.298	0.458	0	0	1	1
estrato	38430	3.94	1.176	1	3	5	6
Numero de gasolineras a 500 mts	38644	1.025	1.194	0	0	2	8
Distancia gasolinería mas cercana	38644	484.49	278.354	0	272.038	650.789	3531.409
Numero de hospitales a 500 mts	38644	0.374	0.699	0	0	1	10
Distancia hospital mas cercano	38644	872.9	533.555	0	456.537	1208.114	3533.104
Numero de estaciones policia a 500 mts	38644	0.217	0.493	0	0	0	5
Distancia estacion policia mas cercana	38644	992.136	506.072	0	594.408	1361.11	2953.795
Numero de parques a 500 mts	38644	10.328	5.746	0	7	13	51
Distancia parque mas cercano	38644	112.946	94.731	0	44.378	156.849	2884.694
Numero de rios a 500 mts	38644	0.007	0.09	0	0	0	2
Distancia rio mas cercano	38644	5017.323	1825.721	22.366	3626.321	6482.354	8382.765
Numero de universidades a 500 mts	38644	0.428	1.458	0	0	0	20
Distancia universidad mas cercana	38644	963.17	561.907	0	549.045	1304.969	4328.717
Numero de estaciones transporte a 500 mts	38644	1.031	1.923	0	0	1	13
Distancia estacion transporte mas cercana	38644	942.975	680.655	0	409.064	1352.699	5945.504
Numero de Supermercado a 500 mts	38644	1.731	1.756	0	0	2	10
Distancia supermercado mas cercano	38644	395.917	257.098	0	214.165	521.846	3645.27
Numero de C.comercial a 500 mts	38644	1.074	1.508	0	0	2	14
Distancia C.Comercial mas cercano	38644	597.864	377.853	0	310.75	844.721	4601.649

Table 2: Estadísticas descriptivas datos vivienda Chapinero (test)

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
price	0						
... No	0	NaN%					
... Yes	0	NaN%					
month	10286	5.673	3.19	1	3	8	12
year	10286	2020.383	0.729	2019	2020	2021	2021
area M2	10286	760.025	2710.131	15	106	617.963	108800
Numero de habitaciones	10286	2.381	0.961	0	2	3	11
Numero de baños	10286	2.552	1.009	1	2	3	10
Tipo de vivienda (1 si es casa)	10286	0.027	0.161	0	0	0	1
Cuenta con sala o comedor	10286	0.988	0.111	0	1	1	1
Servicios interiores adicional	10286	0.982	0.132	0	1	1	1
Servicios exteriores adicional	10286	0.323	0.468	0	0	1	1
garage	10286	0.993	0.085	0	1	1	1
Luz natural	10286	0.365	0.481	0	0	1	1
estrato	10286	4.59	1.397	1	3	6	6
Numero de gasolineras a 500 mts	10286	0.616	0.967	0	0	1	5
Distancia gasolinería mas cercana	10286	568.984	265.746	4.006	361.181	761.86	2238.862
Numero de hospitales a 500 mts	10286	0.879	2.192	0	0	1	10
Distancia hospital mas cercano	10286	816.021	422.24	0	492.179	1159.655	3408.093
Numero de estaciones policia a 500 mts	10286	0.42	0.607	0	0	1	5
Distancia estacion policia mas cercana	10286	712.066	406.668	3.801	376.456	1018.766	1716.106
Numero de parques a 500 mts	10286	9.178	5.537	0	5	12	33
Distancia parque mas cercano	10286	123.051	91.135	0	55.645	174.953	1933.343
Numero de rios a 500 mts	10286	0	0	0	0	0	0
Distancia rio mas cercano	10286	6678.8	843.427	1310.319	6134.184	7339.469	9638.897
Numero de universidades a 500 mts	10286	1.137	2.174	0	0	1	20
Distancia universidad mas cercana	10286	543.736	323.856	0	275.825	782.409	2725.888
Numero de estaciones transporte a 500 mts	10286	0.734	1.471	0	0	1	8
Distancia estacion transporte mas cercana	10286	813.965	442.091	0	470.413	1169.893	2968.778
Numero de Supermercado a 500 mts	10286	0.866	1.308	0	0	1	11
Distancia supermercado mas cercano	10286	578.017	324.414	0	331.104	740.956	2142.917
Numero de C.comercial a 500 mts	10286	0.398	0.909	0	0	0	5
Distancia C.Comercial mas cercano	10286	738.187	337.804	0	505.615	937.093	2962.623

III. Modelos y resultados

Se diseñaron y calibraron seis (6) modelos distintos, dentro de estos se encuentran Regresión Lineal, Lasso, Ridge, Elastic Net, Random Forest y XG Boost.

Para medir el desempeño de los modelos se utilizó la medida del MAE, o promedio del error absoluto. Adicionalmente, antes de la estimación se realizó un proceso de estandarización de los precios de las propiedades para tener una mejor distribución de los datos. Al momento de realizar las predicciones, los precios predichos por los modelos se convirtieron a los valores originales mediante la media y la desviación estándar de la muestra.

A continuación, se detallan las medidas de este indicador por cada modelo estimado para su comparación.

Modelo	MAE
Modelo Lineal	0,6157
Lasso	0,5425
Ridge	0,5238
Elastic Net	0,4894

Random Forest	0,3321
XgBoost	-

Con base en lo anterior, se encontró que los modelos lineales, ridge y lasso tienen un tiempo de procesamiento significativamente menor en comparación con los otros dos modelos, pero su capacidad de predicción con datos espaciales es baja, con un error absoluto medio superior al 50%. En consecuencia, se concluye que variables como el área, el número de habitaciones y baños, la distancia a los ríos, universidades, estaciones de policía, parques y hospitales son factores clave para predecir el precio de la vivienda.

Entonces, en la tabla anterior se pueden ver los resultados de la métrica MAE para los mejores modelos en la base de train y su desempeño. En esta se puede observar que los modelos con mejores predicciones fueron elastic net con un MAE de 0,48 y bosques aleatorios con un MAE de 0.33, este modelo se calibró con 11 predictores al azar que se tomaron en cada iteración, 790 árboles y 3 en tamaño mínimo del último nodo. Por otro lado, no se pudo estimar el modelo de XGBoost por la velocidad de respuesta del servidor. Sin embargo, se cree que puede tener un MAE inferior al Random Forest, dada la naturaleza del modelo.

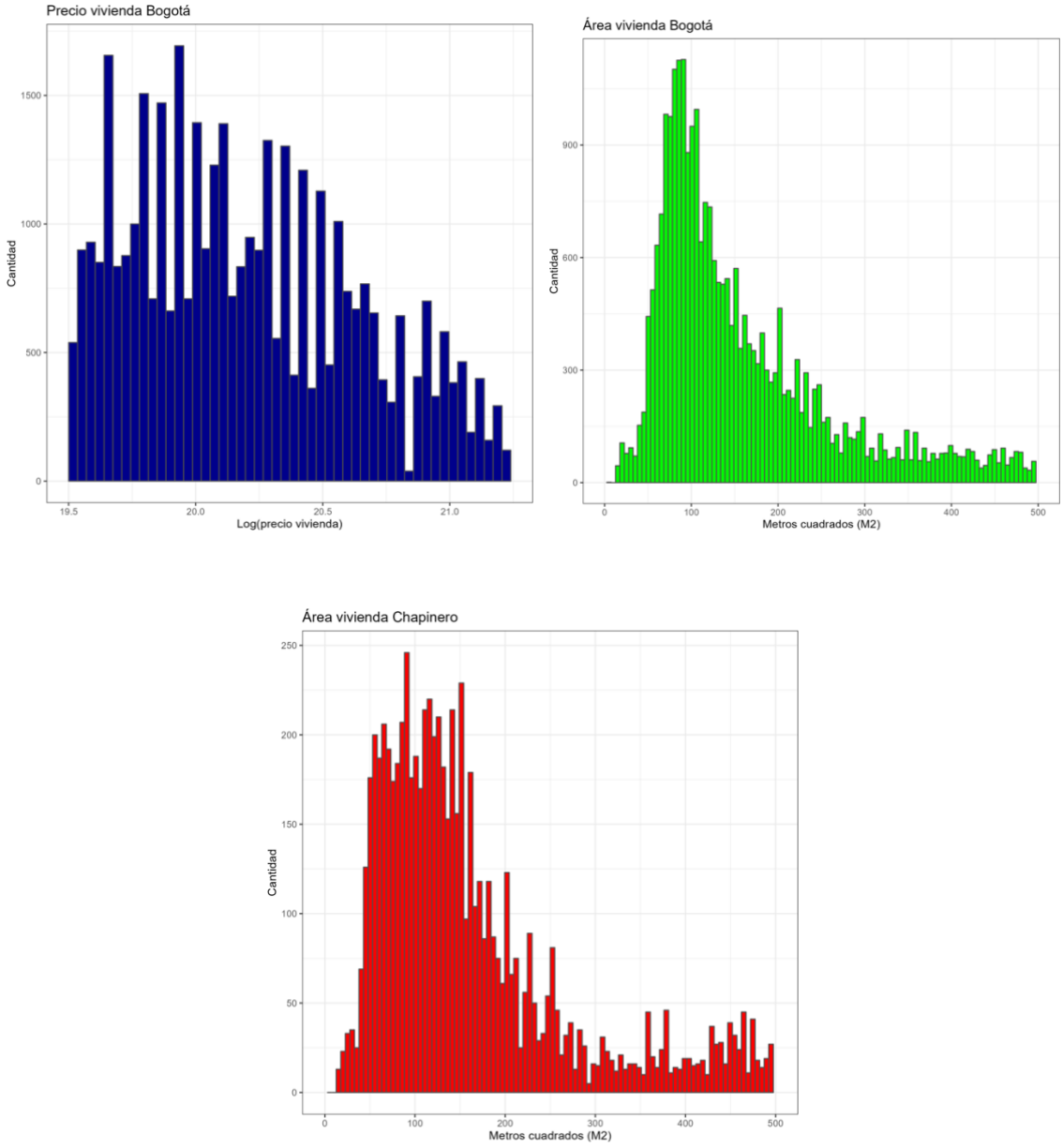
Ahora bien, parte del ejercicio consistía en testear el MAE de los modelos a especificados, relativo a otras entradas proporcionadas en el aplicativo de Kaggle. En este caso particular, se compararon las estimaciones del Elastic Net, entre otras, la cual obtuvo una calificación sobresaliente con respecto a otras entradas del mismo tipo.

IV. Conclusiones y recomendaciones

A partir de los resultados anteriores, se considera que los modelos especificados tienen un poder predictivo relativamente bajo, pues el MAE no llega al menos al 20%. Esto puede deberse a omisión de una o más variables que explican la fijación de los precios de la vivienda en Bogotá, pese a que se consolidó una base de datos con diversas fuentes de información. Las variables más relevantes de los modelos fueron área total, numero de baños y habitaciones, distancia al río, hospital, estación de policía y parques. Esto puede ser indicador de la relativa poca importancia de las demás variables, ya que, de manera preliminar al análisis se esperaba que variables como estrato o actividad económico tuvieran mayor peso.

El costo computacional de los modelos, las medidas de clusterización y la creación de funciones y parámetros de calibración fueron desafiantes para el desarrollo del ejercicio.

V. Apéndice



Referencias

DANE (2018). Censo nacional de población y vivienda. page np. Disponible en: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/

DANE (2021). Marco geoestadistico nacional (mgn). page np. Disponible en: https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/

EOG (2021). Earth observation group. Earth Observation Group. Disponible en: https://eogdata.mines.edu/products/vnl/.

OSM (2023). Openstreetmap. OpenStreetMap. Disponible en: https://wiki.openstreetmap.org/wiki/ES:P%C3%A1gina principal.

PROPERATI (2023). properati.com portal web. page np. Disponible en: https://www.properati.com.co/.