Big Data for Applied Economics – Problem Set 3

Juan Diego López (201425104), María Paula Nieto (201612379), Daniela Jaime (201531520)

Introducción

Los mercados de vivienda alrededor del mundo son complejos. Al respecto, la economía urbana ha buscado identificar los determinantes del precio de las viviendas (Alonso,1964; Rosen,1979; Roback,1982). Sin embargo, en los últimos años se han desarrollado modelos de predicción más complejos que han buscado determinar los precios a partir de redes neuronales, *random forest* e incluso imágenes satelitales, y modelos hedónicos basados en los atributos y las *amenities* de cada vivienda (Liu & Wu, 2020; Wang et al., 2016; Bency et al., 2017).

Con esto en mente, en este trabajo se presentará un modelo de predicción para los precios de las viviendas en los barrios de Chapinero, en Bogotá, y el Poblado, en Medellín. Para esto, se explicará de qué manera se realizó la limpieza y el manejo de los datos, cómo se definieron los hiperparámetros en los modelos y las principales conclusiones del modelo elegido. La información que definió y alimentó el modelo elegido proviene del portal "Properati" y es complementada por información georreferenciada de las plataformas de Datos Abiertos de Bogotá y Medellín, así como por la información en Open Street Map. El script, figuras y bases utilizadas se podrán encontrar en el siguiente repositorio: https://github.com/marianieto198/Problem-Set-3.git

Así, se eligió un modelo de *XGBoost* para las dos ciudades que permitió comprar el 68.6% de las propiedades en la base. Aunque este tipo de modelos permiten construir un modelo óptimo a partir de predictores débiles (un poco mejores que un *random guess*) a través de la construcción de árboles, puede generar problemas de sobreajuste por lo que sus hiperparámetros deben ser elegidos de manera que disminuyan esta problemática.

Manejo de datos

La base de datos original contenía información sobre propiedades en el país con una serie de variables acerca de sus *amenities*: cuartos, baños, superficie total y cubierta, entre otros. En general, se observa que el 11.11% de las variables tienen *missing values*. Específicamente, el análisis gráfico muestra que las variables con mayor cantidad de valores faltantes eran: cuartos, baños, superficie total y superficie cubierta. Posteriormente, se aplican los filtros necesarios para delimitar la base, ya convertida en un shape file, para enfocar el análisis sobre las zonas de interés.

Ahora bien, con el propósito de determinar la mayor cantidad de características de las propiedades a partir de la información original y disminuir los *missing values* en variables como la superficie total, se busca extraer información de la descripción. Por lo tanto, se establecen patrones[1] que permitan extraer datos sobre la superficie total a partir de la descripción. Así, se logra disminuir el número de valores faltantes en la superficie total. Además, dada la relevancia de la superficie de la vivienda en este ejercicio y el hecho de que para algunas propiedades se tiene información sobre solo una de las variables de superficie cubierta y total, se reemplaza el valor de la variable de la superficie total por el máximo valor entre la variable superficie cubierta y la superficie total.

El proceso de derivación de información a partir de la descripción se repite con los pisos, el estrato, y los parqueaderos. De las variables creadas, las que menos valores faltantes tienen son las de superficie y los parqueaderos entonces extraemos los números de esas variables. Finalmente, se observa que los *missing values* disminuyen significativamente. Por otro lado, para complementar la información se utiliza

[1] Para la variable de superficie total se definieron 12 patrones diferentes.

el método de vecinos cercanos a nivel de manzana. Después de esto, se logra que la variable de superficie total no tenga ningún missing value.

Adicionalmente, usando información de *OpenStreetMap* se generan nuevas variables predictivas. Estas variables miden la distancia de las propiedades a diferentes lugares como estaciones de bus, centros comerciales, centros de negocios como la Plaza de Bolívar en Bogotá o la Plaza Botero en Medellín, y bares. Lo anterior, con el propósito de identificar si la cercanía a estos lugares genera un valor agregado (o no) sobre el precio del inmueble. Por el otro lado, usando bases de datos externas, se adiciona información relacionada con las dinámicas del crimen, contaminación y distancia a colegios. Los puntos referenciados y los polígonos que definieron el análisis se pueden observar en el Anexo 1. Para el caso de Bogotá se encontró más información en la página de datos abiertos de la administración municipal como los colegios, la ciclovía, la contaminación, la ubicación de los centros comerciales y el crimen (hurto a personas). Sin embargo, para Medellín solo se utilizó información de la ciclovía y los colegios, por la disponibilidad de los datos.

A partir de las estadísticas descriptivas de las variables, es posible identificar que las variables *rooms*,*surface covered*, estrato, parqueadero y pisos tienen un alto número de *missing values*, por lo que en los modelos se incluirán variables con mayor número de observaciones y con las que estén correlacionadas. Además, es posible identificar que la variable de precio no tiene una distribución simétrica y que tiene un mayor peso a la derecha, lo que se confirma con un *skewness* de 2.4. Esto también ocurre con la variable de superficie total (Anexo 2). En consecuencia, se realizan diferentes transformaciones para encontrar aquella más cercana a cero (Anexo 3); en este caso, las distribuciones Box-Cox y logaritmo ofrecen los valores más cercanos a cero. Por eso, dentro de los modelos que se evalúen, algunos tendrán la variable precio transformada con logaritmo. En cuanto a los valores atípicos, los diagramas de cajas y bigotes permiten identificar que existen valores atípicos en las siguientes variables: precio, superficie total, el número de cuartos, sin embargo, teniendo en cuenta las características de las zonas de interés se puede considerar razonable la existencia de estos valores.

Modelos de predicción y resultados

A partir de lo anterior, se estimaron una serie de modelos predictivos dentro de los que se incluyeron: OLS, Ridge, Lasso, *Elastic Net* y *XGBoost*. Adicionalmente, se incluyó la transformación logarítmica del precio en algunos, y se estimaron modelos separados para Bogotá y Medellín. En cuanto al criterio de evaluación de los modelos, se utilizó el RMSE para determinar cuál era más adecuado. En ese sentido, se encontró que el modelo con el menor RMSE fue el *XGBoost* para las dos ciudades. Al respecto, se debe destacar que, si bien algunos modelos para Medellín o Bogotá tenían un error menor, también tenían pocas observaciones. Además, al comparar los errores entre el modelo general y los modelos específicos para cada ciudad, se identificó que la diferencia en el error no era amplia por lo que la pérdida en observaciones no era justificable.

Ahora bien, el modelo de predicción elegido fue un *XGBoost* estimado para las dos ciudades. En cuanto a las variables seleccionadas, se incluyó información proveniente de la descripción de la propiedad (bedrooms y surface_total). Igualmente, se incluyeron variables calculadas con fuentes de datos externas como *Open Street Map* y las bases de datos abiertas tanto de Bogotá como de Medellín. Como se mencionó anteriormente, estas variables fueron seleccionadas teniendo en cuenta que tienen un valor menor de missing values y que parecen estar correlacionadas con el precio de la propiedad. En este modelo, la variable dependiente de precio no contiene la transformación logarítmica por cuanto los MSE y, por consiguiente, los RMSE obtenidos sin esta transformación eran menores.

$$\hat{Y} = \sum_{i=1}^{N} f(\text{Bedrooms} + \text{Surface\_total} + \text{dist\_bar} + \text{dist\_bus} + \text{Ciclovia} + \text{parques} + \text{colegios} + \text{dist\_CBC} + \text{ciudad} + \text{property\_type})$$

Específicamente, este modelo de XGBoost fue estimado con los siguientes hiperparamétros encontrados a través de un método de validación cruzada: i). 250 a 500 rondas; ii). *learning rate* (eta) igual a 0.05; iii). max_depth = 6; iv). Regularización de 0.01; v). un mínimo peso de la hoja de los árboles de 10 y vi). subsample = 0.6. En cuanto a la selección de los anteriores hiperparámetros, resulta pertinente mencionar que estos fueron determinados usando la validación cruzada como un método de selección óptima.

No obstante, se debe destacar que cada uno de los hiperparámetros seleccionados puede contribuir a disminuir la posibilidad de un *overfitting* del modelo XGBoost y que suele caracterizar a los métodos de *gradient boosting*. Frente al *learning rate*, Khun & Johnson (2016), destacan que los valores menores a 0.01 suelen tener un mejor desempeño, pero con un gran costo computacional. En ese sentido, el *learning rate* determinado para este modelo de 0.05 parece ajustarse a esta disyuntiva entre un adecuado comportamiento del parámetro y el costo computacional requerido para determinar el modelo óptimo. Por el otro lado, frente al parámetro de regularización, el método de validación cruzada determinó que sería de 0.01 lo que determina la fracción del valor predicho actual que se adiciona a las iteraciones previas del valor predicho con el fin de evitar un sobreadición de valores previos y generar *overfitting*. Finalmente, en términos del número de rondas, se determinó con validación cruzada que estuviera entre 250 y 500, escogiendo 500. Lo anterior, debido a que un número mayor de interacciones puede ser para reducir el error, pero si este valor es muy grande puede generar problemáticas de sobreajuste debido a que el error se vuelve arbitrariamente pequeño.

Si bien se corrió un modelo para toda la base de *train,* es importante mencionar que se platean también dos modelos diferentes para cada ciudad. Es relevante hacer dos modelos diferentes teniendo en cuenta que para la ciudad de Bogotá se cuenta con una mayor cantidad de información georreferenciada que influye en los precios de las viviendas como lo son estrato, crimen, cercanía a los centros comerciales y cercanía a los CAI de la Policía Nacional. Asimismo, los mercados de la vivienda pueden contar con diferentes dinámicas dependiendo del centro urbano que se esté analizando y las características de la ciudad (Glaeser et al., 2014). Sin embargo, como se destacó anteriormente, la disminución en el error no es suficiente para compensar la pérdida en observaciones por lo que se elige el modelo para las dos ciudades.

En términos de la capacidad predictiva de los modelos, se encuentra que el modelo de *XGBoost* general tiene un menor MSE y, por lo tanto, RMSE. A comparación de los modelos de Ridge y Lasso estimados, este modelo de *XGBoost* donde la variable dependiente no tiene la transformación logarítmica tiene una mejor capacidad predictiva. Por otro lado, en términos de la utilidad del modelo en relación con la predicción de los precios de las propiedades, el Anexo 6 muestra que tiene un buen desempeño. Lo anterior, debido a que permite comprar el 68.6% de las propiedades de la base con el 65.2% del valor total de los inmuebles, por lo que es posible afirmar que genera ahorro.

Conclusiones y recomendaciones

A lo largo de este escrito, se buscó identificar los determinantes del precio de inmuebles en Bogotá y Medellín. A partir de una base de datos, se buscó generar nueva información que permitiera predecir adecuadamente el precio de lo mismo. Con esto en mente, se derivó información de las descripciones de los anuncios de venta, lo que permitió imputar información clave acerca del tamaño y cantidad de habitaciones. Igualmente, se usó *OpenStreetMap* para generar variables que midieran la distancia de las propiedades a lugares clave como estaciones de bus, centros de negocios y bares. Así mismo, usando los datos abiertos se complementó esta información con tendencias de delitos, contaminación y la cercanía a establecimientos educativos. Así, se encontró que el modelo con mejor ajuste predictivo fue *XGBoost* que permitió comprar el 68.6% de las propiedades en la base. Si bien este modelo ofrece la ventaja de constituir interacciones para determinar un modelo óptimo a partir de clasificadores o predictores débiles, puede tener problemas de overfitting.

Referencias

Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge, MA.

A. J. Bency, S. Rallapalli, R. K. Ganti, M. Srivatsa and B. S. Manjunath, "Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 320-329, doi: 10.1109/WACV.2017.42.

Glaeser, E., Gyourko, J., Morales, E. and Nathanson, C., 2014. Housing dynamics: An urban approach. *Journal of Urban Economics*, 81, pp.45-56.

Liu, L. and Wu, L., 2020. Predicting housing prices in China based on modified Holt's exponential smoothing incorporating whale optimization algorithm. *Socio-Economic Planning Sciences*, 72, p.100916.

Roback, J., 1982. Wages, rents, and the quality of life. Journal of Political Economy 90 (4), 1257–1278.

Rosen, S., 1979. Wage-based indexes of urban quality of life. In: Mieszkowski, P., Straszheim, M. (Eds.), Current Issues in Urban Economics. Johns Hopkins Univerity Press, Baltimore.

L. Wang, F. F. Chan, Y. Wang and Q. Chang, "Predicting public housing prices using delayed neural networks," *2016 IEEE Region 10 Conference (TENCON)*, 2016, pp. 3589-3592, doi: 10.1109/TENCON.2016.7848726.

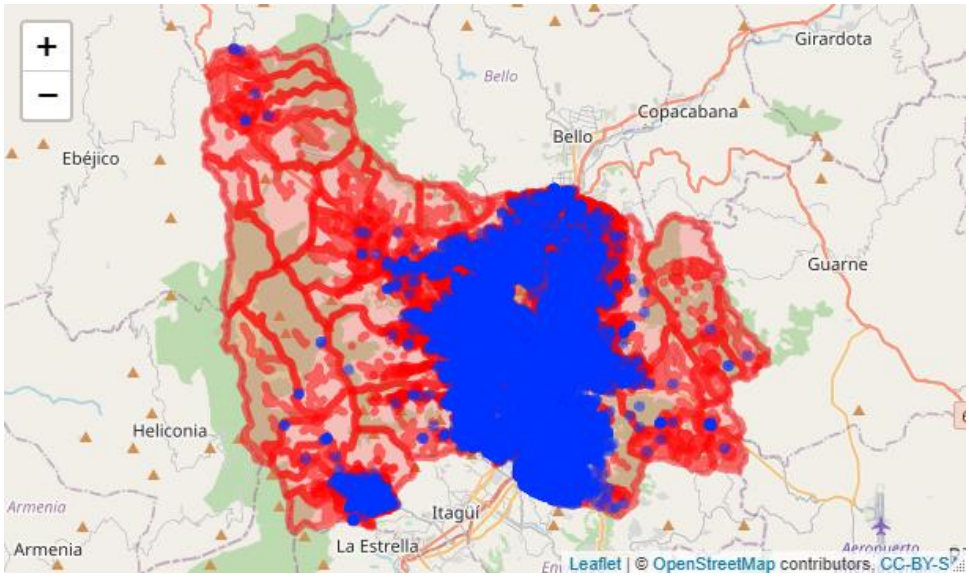
Kuhn, M., & Johnson, K. (2016). Applied predictive modeling (Vol. 26, p. 13). New York: Springer

Anexos

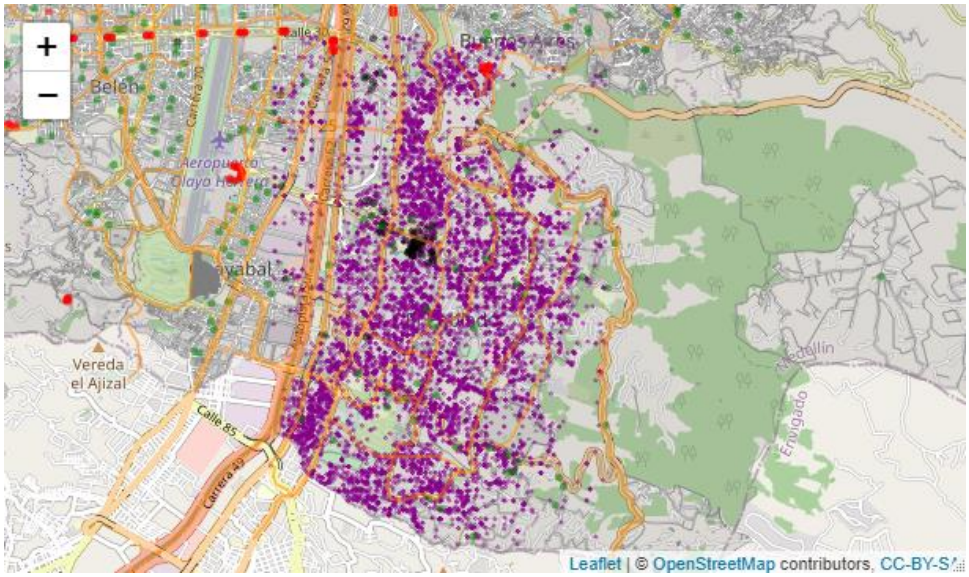
Anexo 1. Mapas de Medellín y Bogotá para las variables de interés

1.1 Manzanas y viviendas en venta

Medellín



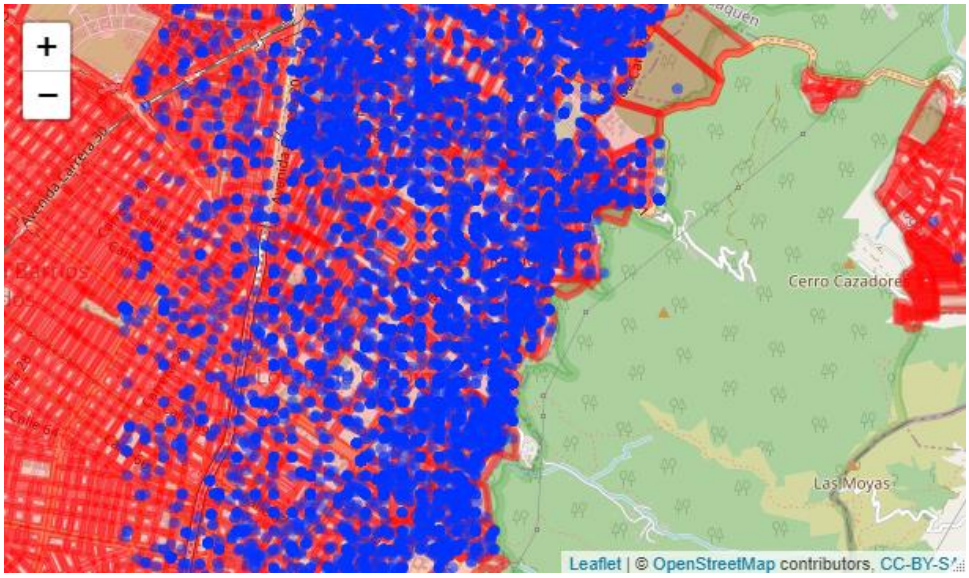
Convenciones: Rojo: manzanas; Azul: viviendas en venta



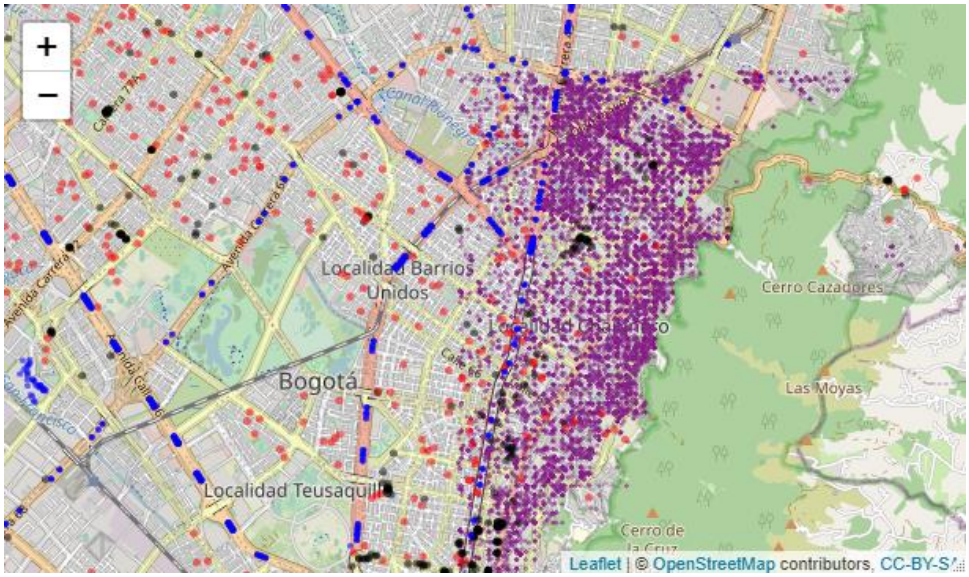
Convenciones:

Morado: Viviendas en venta; Negro: bares; Rojo: estaciones de transporte; Verde: Colegios; Naranja: ciclovías.

Bogotá-Chapinero



Convenciones: Rojo: manzanas; Azul: viviendas en venta



Convenciones:

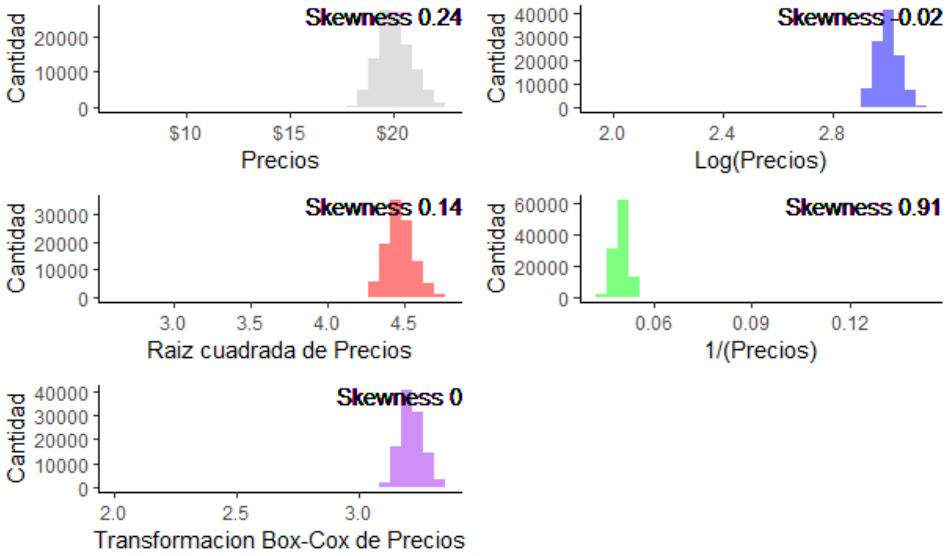
Morado: Viviendas en venta en chapinero; Negro: bares; Rojo: colegios; Azul: estaciones de transporte.

Anexo 2. Tablas estadísticas descriptivas

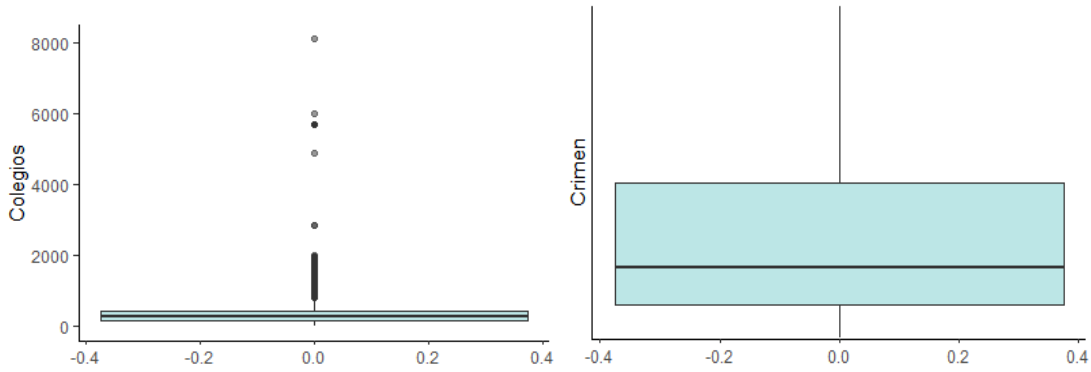
Variable	n	mean	sd	median	min	max	skew
rooms	53961	2,976557143	1,238792191	3	1	11	2,109642259
bedrooms	107567	3,080870527	1,384925212	3	0	11	2,106236801
bathrooms	107458	2,788801206	0,815152522	3	1	13	0,894084076
surface_total	107567	86,1662522	452,1934758	52,3	0	108800	188,7164615
surface_covered	20199	148,1715927	176,1324606	110	1	11680	28,28876504
price	107567	693106705,8	677656641,8	4,60E+08	1100	4,99E+09	2,595423507
estrato	72673	4,091629629	1,716372027	4	0	6	-0,801999761
parqueaderos	18189	2,157402826	1,143496823	2	0	9	1,774998998

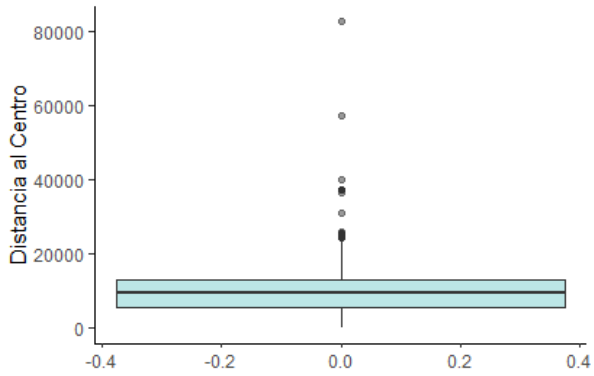
pisos	7409	3,445809151	1,827262805	3	1	9	0,837120461
dist_bar	107567	699,3209077	634,8547342	572,250997	1,996460872	47040,8728	11,92394733
dist_bus	107567	951,9451368	796,9983248	800,3750599	2,194368114	31298,32088	6,228637969
ciclovia	107567	56219,8182	28348,9499	69572,26385	0,015290021	82406,97532	-1,405269003
parques	107567	187670,0589	93226,41594	233748,1608	0	237381,8513	-1,511111796
crimen	86209	290,6576344	251,7999035	197	0	2761	1,012031118
colegios	107567	318,5902593	241,8594347	265,5916386	1,004765824	8104,945953	4,562160033
CAI	86211	329,2900926	250,0250845	274,1413561	2,298351044	8104,945953	4,930684736
centrocomercial	86211	329,2900926	250,0250845	274,1413561	2,298351044	8104,945953	4,930684736
distancia_CBC	107567	9967,428353	4502,866157	10331,28331	16,15186703	82599,57538	0,130428091

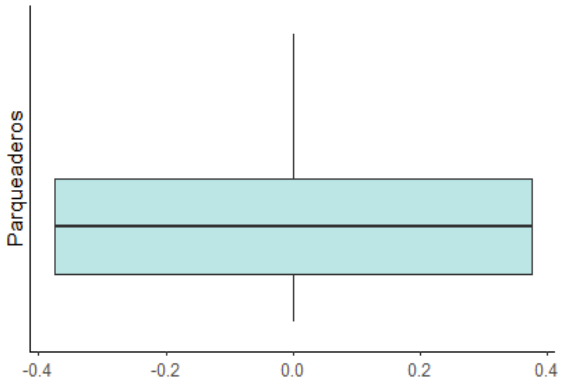
Anexo 3. Distribución de los precios con diferentes transformaciones

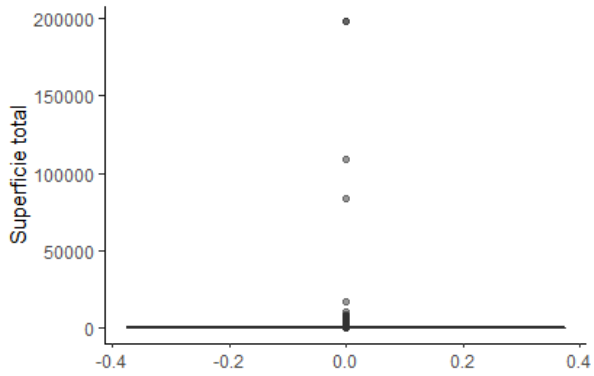


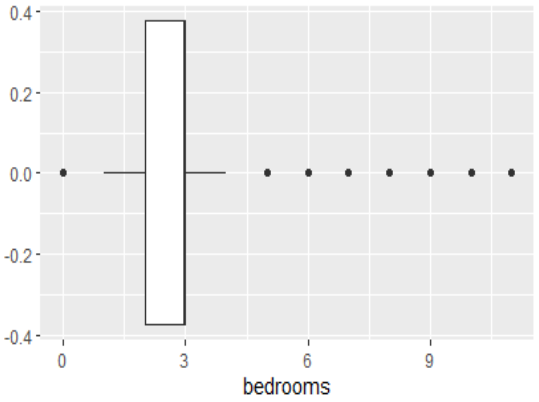
Anexo 4 . Análisis de datos atípicos y distribución de las variables de interés

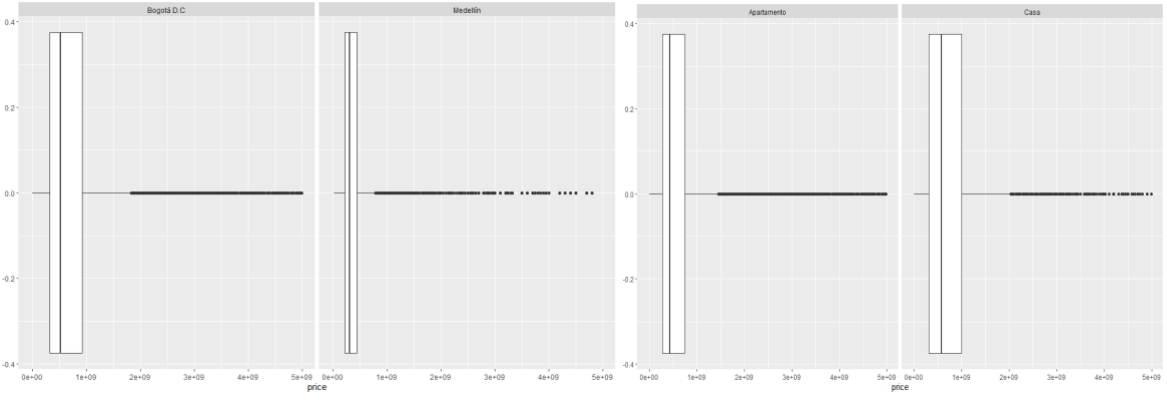


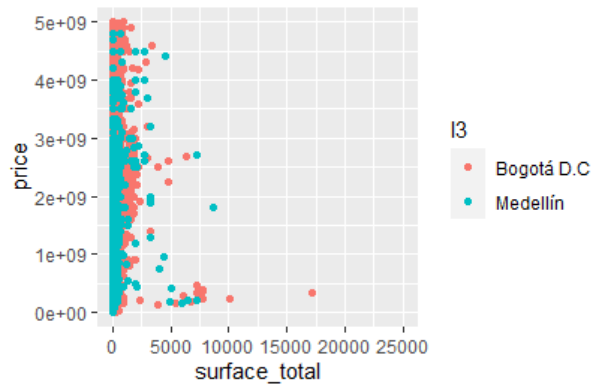

Distancia al Centro
80000
60000
40000
20000
0
-0.4
-0.2
0.0
0.2
0.4

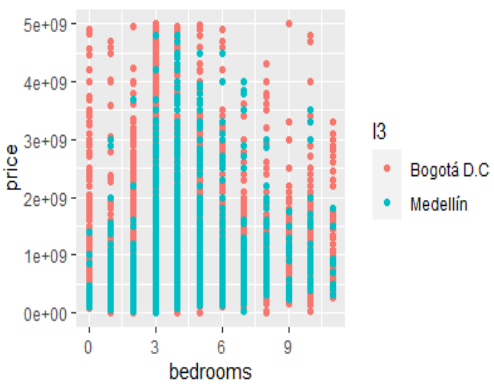

Parqueaderos
-0.4
-0.2
0.0
0.2
0.4


Superficie total
200000
150000
100000
50000
0
-0.4
-0.2
0.0
0.2
0.4


0.4
0.2
0.0
-0.2
-0.4
0
3
6
9
bedrooms


Bogotá D.C
Medellín
Apartamento
Casa
price


price
5e+09
4e+09
3e+09
2e+09
1e+09
0e+00
0
5000
10000
15000
20000
25000
surface_total
l3
Bogotá D.C
Medellín


price
5e+09
4e+09
3e+09
2e+09
1e+09
0e+00
0
3
6
9
bedrooms
l3
Bogotá D.C
Medellín

Anexo 5. Tabla Resultado Modelos

ModeloNombre	MSEValores	RMSEValores
Modelo OLS	1,84E+35	42.898.106.877.031
Modelo OLS sin Logaritmo	3,58339E+17	598.614.044.881.627
Modelo Ridge	4,22702E+17	650.155.475.541.091
Modelo Lasso	3,5835E+17	598.623.790.390.438
Modelo XGBoost	1,64307E+17	405.348.540.721.670
Modelo XGBoost sin logaritmo	1,41112E+17	375.649.109.611.085
Elastic Net	3,584E+17	598.664.805.720.749
Modelo RidgeEducacionYActividad	3,78018E+17	614.831.517.713.333

Anexo 6. Tabla resultados modelo XGBoost

Resumen de Compra con el Modelo XGBoost	
Número de inmuebles totales	**Valor total de los inmuebles**
32.270	$ 22.462.684.890.797,00
Número de Inmuebles comprados	**Valor Total de Compra**
22.122	$ 14.635.586.046.492,00
Proporción de Inmuebles comprados	**Proporción del valor comprado sobre el total**
68,6%	65,2%