# Predicciones precios de viviendas en Colombia

**Mateo Contreras**
**Carlos Ayala**
**Federico Meneses**

***Link de Repositorio:*** ***https://github.com/Mateocontrerass/Price-Prediction-With-Spacial-Data***

## Introducción:

En 2019, la empresa inmobiliaria Zillow había creado un modelo cuya aplicación permitía predecir los precios para diferentes inmuebles en la ciudad. Sin embargo, debido a la pandemia próxima a acontecer para el año 2020 y el cambio en la oferta y demanda de las viviendas, los cambios en los precios fueron diferentes a los que el modelo de Zillow había predicho, es decir, el modelo perdió su poder predictivo ante cambios externos inesperados (Sarnoff, 2022). Las consecuencias para Zillow fueron devastadoras económicamente, por lo que tuvo que despedir casi 25% de su mano de obra. Este acontecimiento muestra la dificultad de realizar modelos para la predicción de inmuebles y la facilidad con la que puede presentarse casos de sobreajuste para estos modelos. Dicho esto, es necesario traer a discusión modelos que puedan reducir el sobreajuste y que funcionen con variables que no sean cambiantes en el tiempo.

Para poner a prueba un modelo que cumpla estas características, el problema que este texto busca solucionar radica en predecir los precios de una ciudad en base a otras dos ciudades que son diferentes a esta en varios aspectos. Para este caso, se busca predecir los precios de las viviendas en Cali (Colombia), en base de las características de inmuebles de Bogotá y Medellín (Colombia). El objetivo entonces del modelo va a ser tomar de estas dos ciudades características inmobiliarias que puedan afectar los precios de las viviendas reduciendo en lo más posible el ruido blanco que se capture del modelo y que las estimaciones estén ligadas, en lo más posible, a factores o variables relacionados a las características del inmueble.

Para ello, se busca seleccionar un modelo que pueda predecir con el menor ruido blanco posible, es decir, nuestro criterio de evaluación se basa en el error cuadrático medio del modelo (MSE). Un menor MSE implica menor varianza de los datos, es decir, un menor ruido blanco presente en el modelo. Por lo tanto, se considera utilizar un Extreme Gradient Boost (XGBoost). Este modelo tiende a reducir considerablemente el MSE y reducir el sobreajuste del modelo de predicción. Ahora bien, es necesario tener en cuenta el número de iteraciones de este modelo a la hora de hablar de reducción del sobreajuste y asertividad en las predicciones, por lo que se utilizará la metodología de Cross Validation (CV) para decidir esto.

## Datos

Uno de los principales insumos utilizados son las distancias hacia diferentes servicios que son ofrecidos por terceros y se considera que tiene tener injerencia en el precio final del inmueble. Algunos de estos servicios son: hospitales, escuelas, policía, zonas de juego, centros comerciales, entre otros (…). Después de esto se utiliza información del inmueble

como área total, baños, si está remodelada la vivienda, que, sumado a su ubicación, permiten una mejor aproximación hacia el costo del lugar.

**Procesamiento**

Para las bases de datos tomadas del ejercicio, donde la base Train incluye los precios de las casas de Bogotá y Medellín además de las características del inmueble junto a su descripción. Test incluye estas mismas variables, pero únicamente par inmuebles de Cali. Para los datos se les agrega las distancias espaciales de una gran selección de Ammenities por ejemplo cafés, policía, estaciones de Transmilenio, centros comerciales; demás servicios que puedan encontrarse en el conjunto del inmueble como parques o gimnasios; el sector que se encuentra tipo industrial o residencial; y tiendas cercanas como tiendas de barrios, cosméticos, o cigarrerías. Por otra parte también fue necesario completar algunos datos en base de las descripciones de los inmuebles o agregar algunas otras variables, entre estas que se rellenaron están la de baños, habitaciones y superficie total; mientras que algunas variables nuevas de características del hogar, por ejemplo, si la vivienda está remodelada, si es apartamento, si tiene parqueadero, si tiene lavandería, si tiene terraza, si tiene cocina integral, si tiene ascensor, si tiene vigilancia, si tiene iluminación, o si tiene piscina. en el procesamiento de los datos fue la imputación de valores nulos.

Por otra parte, para este procedimiento se utilizó un algoritmo de imputación múltiple basado en XGBoost, Boostraping y Predictive Mean Matching con la librería mixgb. Se utilizó este método dado que se buscó un modelo más complejo que la imputación de las medias de la variable en caso de baja Skewness y mediana para alta Skewness. Se omitió la implementación de vecinos espaciales como forma de imputación debido a que para este proyecto requería una alta capacidad computacional y no se pudo llevar a cabo.

**Estadísticas descriptivas**

En la *tabla.1* se presenta la tabla de estadísticas descriptivas para las variables que resultan después del preprocesamiento de los datos. En total son 38 variables, 7 de estas son tipo carácter porque son Dummies y están relacionadas con algunas características con las que cuenta el inmueble, obtenidas por medio de expresiones regulares que se presentaron anteriormente en procesamiento; los 31 restantes son atribuidas a las distancias espaciales obtenidas de OpenStreetMaps o de otras características restantes del hogar, como número de habitaciones, baños o superficie total.

Primero, es necesario analizar la variable precios para observar el valor de los inmuebles en ambas ciudades. Se observa que estos valores son promedio de 800 millones de pesos en promedio con una alta desviación estándar de aproximadamente 200 millones de pesos, esto puede indicar que los precios son altamente variables de su media. Al pasar ahora con las características no dummies de los inmuebles, es posible observar que superficie total y habitaciones tiene una gran proporción de missing values, por lo que los datos pueden no ser los más acertados hasta que se haga la imputación por medio del XGBoost. Por otra parte, la variable baños está completa y contiene una media de 1.30 con una desviación estándar cercana a 0, lo que indica baja variabilidad en la cantidad de baños, similar al de habitaciones.

Al pasar ahora con las variables tipo carácter, estas no tienen relevancia para las predicciones. Al analizar las variables de distancias, es posible observar que, metros, el promedio de cada

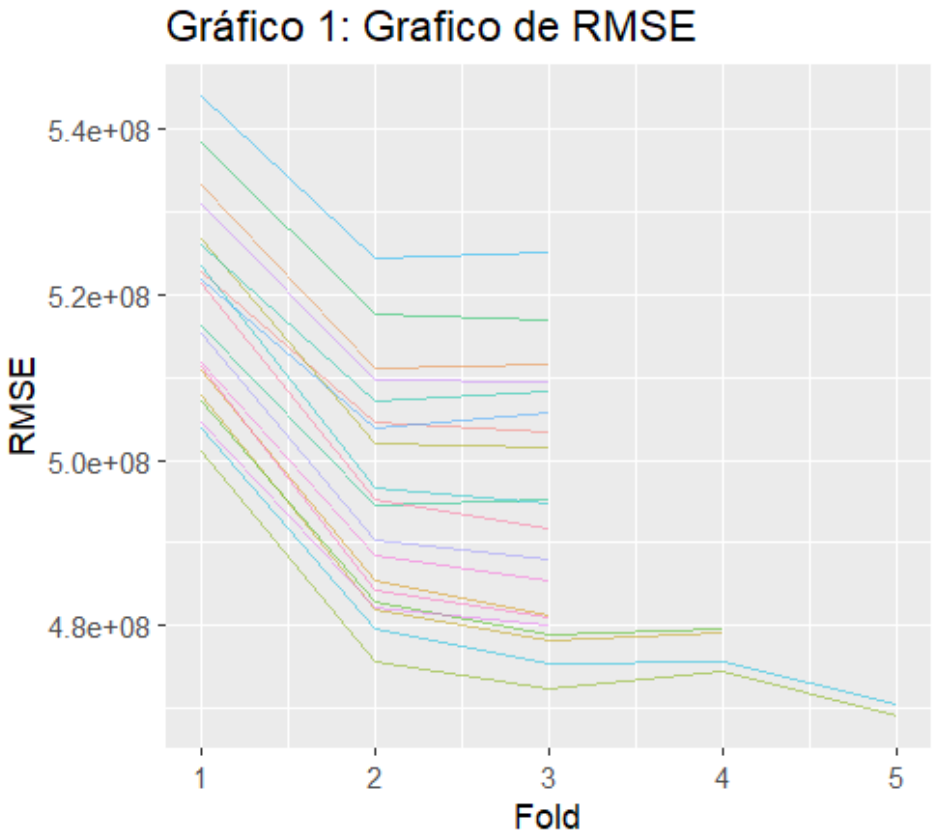
una de estas distancias es similar para la mayoría de las variables. Esto puede indicar que, en promedio, lugares como bancos, hospitales, escuelas o supermercados son sitios más importantes para los hogares o que así mismo para estos sitios es importante estar más cerca de los hogares, por lo que la mayoría de los hogares prefiere tener estos lugares cerca, es decir, no tenerlos puede que tenga afectaciones negativas en el precio. Por parte de aquellos lugares con una distancia promedio alta, como cinemas, restaurantes, cafeterías o joyerías, puede que no tengan la misma relevancia por lo que no muchos hogares las tienen cerca, sin embargo, también existe la posibilidad que tener cerca una de esta locación pueda aumentar el precio de la casa, ya que no muchas casas deben tener estos sitios cerca.

**Implementación del modelo**

Para la ejecución del modelo, se hizo uso de un XGBoost, con el objetivo de reducir el MSE por medio de la prueba iterativa del árbol, tomando los valores correctos que se obtenían de cada iteración, y usar los mismo para seguir iterando. Como resultado, se obtienen 20 modelos diferentes, de los cuales se escoge el que tiene menor RMSE para hacer la predicción. La gráfica muestra el comportamiento del RMSE con las diferentes especificaciones, durante las diferentes iteraciones. Así, se observa que hay 2 en específico que convergen a un menor RMSE, y se escoge la de menor valor.

Para ajustar los valores de algunos parámetros del modelo se utilizó el método de búsqueda en grilla para determinar que valores se ajustaban mejor al comportamiento de la información obtenida. La selección de los rangos de estos parámetros es arbitraria.

El modelo se implementó con datos para inmuebles (casas y apartamientos) para las ciudades de Bogotá y Medellín. Esto, se hizo con el objetivo de evitar sobre ajustar los parámetros a las condiciones específicas de una ciudad en particular, y tener problemas para predecir sobre una ciudad diferente. Así, se utilizan datos de distancias a zonas relevantes como lo son vías principales, zonas comerciales, zonas escolares y zonas residenciales. Adicionalmente, se hace uso de información relevante tomada de la descripción del inmueble ofrecida por el vendedor, donde se toman características como el área, cantidad de habitaciones, si está remodelado, si tiene parqueadero, si tiene lavandería, si tiene terraza, entre otras.




**Conclusiones**

Si bien por el tipo de modelo no es posible encontrar la contribución de cada factor a las predicciones, es necesario recalcar que no era el objetivo de este documento encontrar contribución o causalidad, sino experimentar un modelo que permitiese reducir en lo más posible el SCE en base a unos factores y características fijas de los inmuebles para un momento en el tiempo. Se considera que la implementación del XGBoost mediante Cross Validation pudo reducir significativamente el RMSE y se observa para la prueba con la base test, que lo valores predichos son cercanos a su contraparte en las predicciones, por lo que, teniendo en cuenta que se obtuvieron características de dos ciudades totalmente distintas para predecir los precios de otra ciudad, es un buen modelo.

## Anexos:

### *Tabla 1: Estadísticas descriptivas 1*

```
Group variables             None

── Variable type: character ──────────────────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable  n_missing complete_rate min  max empty n_unique whitespace
1 property_id            0             1  24   24     0    51437          0
2 city                   0             1   8   10     0        2          0
3 title                 15             1.00  3  152     0    29210          0
4 description           36             0.999 1 7945     0    45199          2
5 property_type          0             1   4   11     0        2          0
6 operation_type         0             1   5    5     0        1          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────────────────────────────────────────────
   skim_variable       n_missing complete_rate         mean           sd          p0         p25         p50         p75         p100 hist
 1 price                       0             1     809472258.   842469412.   200000000   360000000   550000000   900000000   2000000000
 2 surface_total           34600             0.327       191.        2167.           2          81         122         206         198000
 3 rooms                   24915             0.516         3.01        1.24          1           2           3           3           11
 4 bathrooms                   0             1             2.80        1.30          0           2           3           3           30
 5 dist_via                    0             1         56825.       94985.         0         151.        369.     211913.      223842.
 6 dist_pub                    0             1           541.         696.         0         136.        276.        584.        6076.
 7 dist_restaurant             0             1          1116.         878.         0.513     394.        907.       1627.        4083.
 8 dist_college                0             1          1152.         660.         2.45      647.       1031.       1581.        4258.
 9 dist_library                0             1           609.         589.         1.40      229.        428.        741.        4570.
10 dist_school                 0             1           748.         538.         0         340.        625.       1019.        4329.
11 dist_university             0             1           670.         541.         0.669     312.        533.        817.        3959.
12 dist_fuel                   0             1           585.         384.         3.56      309.        502.        762.        3561.
13 dist_atm                    0             1           450.         337.         0.722     213.        375.        597.        4075.
14 dist_bank                   0             1           603.         463.         0.615     271.        503.        839.        5286.
15 dist_clinic                 0             1           793.         585.         0.610     365.        661.       1041.        5734.
16 dist_hospital               0             1           409.         370.         0.154     164.        296.        524.        4142.
17 dist_pharmacy               0             1           865.         562.         0         452.        781.       1162.        5517.
18 dist_cinema                 0             1          1217.         718.         4.88      682.       1119.       1601.        6453.
19 dist_nightclub              0             1           778.         480.         1.53      426.        687.       1040.        4284.
20 dist_police                 0             1           915.         610.         4.42      426.        783.       1310.        4114.
21 dist_bus_station            0             1           918.         723.         1.29      441.        750.       1131.        5087.
22 dist_commercial             0             1          6651.        3995.         2.38     1964.       7681.      10082.       15293.
23 dist_industrial             0             1           746.         580.         3.28      305.        593.       1036.        4633.
24 dist_fitness_centre         0             1           500.         440.         0.180     209.        370.        622.        3986.
25 dist_playground             0             1           466.         366.         0.180     206.        370.        617.        2781.
26 dist_alcohol                0             1          1006.         685.         2.66      516.        847.       1311.        5017.
27 dist_coffee                 0             1          1643.        1039.        17.6       790.       1437.       2358.        6993.
28 dist_mall                   0             1           591.         377.         0.937     310.        530.        826.        4602.
29 dist_supermarket            0             1           296.         217.         0         151.        252.        397.        3734.
30 dist_jewelry                0             1          2648.        2783.         5.35      749.       1304.       4009.       12395.
31 dist_cosmetics              0             1          2154.        1612.        14.2      1073.       1663.       2631.        9139.
> |
```