[30 DE NOVIEMBRE 2023]

[Problem Set #2]

Por: [Andrés Chavarro, Ricardo Silva, Federico Camacho & Juan David Vélez Pérez]

[Introducción]

En este trabajo, nos basamos en las regresiones hedónicas y buscamos mejorar la precisión de las predicciones en nuevas bases de datos. Nuestro objetivo principal es ajustar los precios de viviendas en Bogotá, para el barrio de Chapinero, utilizando diversas metodologías predictivas. Exploramos características observables de las viviendas, que van desde sus atributos físicos hasta sus condiciones contextuales y situacionales. Según la literatura sobre los determinantes de los precios de viviendas, sabemos que el precio de una vivienda depende tanto de sus características internas como del entorno en el que se encuentra. La regresión hedónica, introducida por Rider y Henning en 1997 en St. Louis, Estados Unidos, sirve como base para la idea de explorar otros factores del entorno y el vecindario, incluyendo la contaminación del aire, como determinantes del precio de las viviendas.

Metodología

Dado que no teníamos datos sobre la contaminación del aire, optamos por incorporar los efectos de la ubicación y el entorno mediante el uso de datos espaciales. Creamos variables de distancia a diversas instituciones privadas y bienes públicos para comprender mejor la estructura de los precios de las viviendas en Bogotá. Utilizamos datos de OpenMap Street y la biblioteca de R para calcular las distancias a puntos de interés como los Comandos de Atención Inmediata (CAI), estaciones de autobús, parques, bares, entre otros. Nuestra hipótesis fue que la proximidad a estos lugares podría influir en el precio de las viviendas. Por ejemplo, esperábamos que una mayor cercanía a los CAI se relacionara con precios más altos debido a la mayor seguridad percibida. Del mismo modo, esperábamos que la proximidad a las estaciones de autobús influyera en los precios debido a la facilidad de transporte.

También consideramos atributos específicos de las viviendas, como el número de habitaciones, baños, si se trata de una casa o un apartamento, la presencia de parqueadero, terraza, entre otros. Estos atributos se extrajeron del título y las descripciones de las observaciones y se utilizaron en un conjunto de datos de entrenamiento con 38,644 observaciones. Posteriormente, probamos nuestros modelos predictivos en un conjunto de datos de prueba que no contenía información de precios, pero incluía todas las demás características para predecir los precios de las viviendas.

Modelos Predictivos:

Utilizamos varias metodologías diferentes en nuestros modelos predictivos. Comenzamos con métodos de regularización, incluyendo regresiones con penalizaciones tipo Ridge, Lasso y Elastic-net. Exploramos diferentes valores de lambda para evaluar la sensibilidad de nuestras variables independientes.

En el caso de los árboles de regresión, experimentamos con distintas profundidades de árbol, con rangos que iban desde 10 hasta 200. A su vez, ajustamos la cantidad mínima de nodos permitidos en cada árbol, que varió entre 2 y 50. La búsqueda de los mejores hiper-parámetros se realizó mediante validación cruzada y validación cruzada espacial, dividiendo la base de entrenamiento en 5 folds y utilizando el error absoluto medio (MAE) como parámetro para elegir el mejor modelo. En cuanto a los bosques aleatorios, definimos hiper-parámetros para la cantidad de árboles, con valores que oscilaron entre 100 y 500, y la cantidad mínima de nodos u observaciones permitidas en cada árbol. Tratamos de evitar configuraciones con muy pocos árboles para no sobrecargar el modelo. Además, determinamos cuántos predictores debían considerarse en la construcción de los árboles, generalmente en el rango de 2 a 4, dependiendo del modelo de regresión utilizado. Por último, en el Boosting probamos distintas tasas de

aprendizaje (learning rates) que variaron desde 0,001 hasta 0,2. Los demás hiper-parámetros se mantuvieron similares a los de los otros modelos que utilizamos.

Selección del Mejor Modelo:

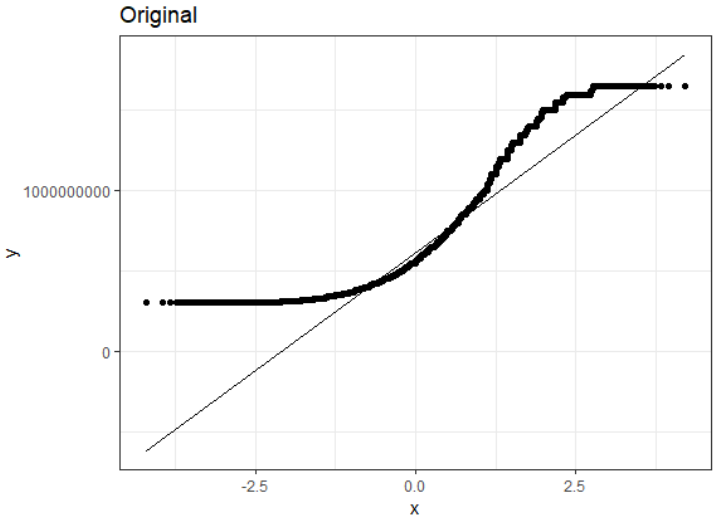
Luego de explorar estos rangos de búsqueda de hiper-parámetros, que comenzaron a ser computacionalmente exigentes, seleccionamos el mejor modelo en función del MAE más bajo. Este modelo se utilizó para realizar predicciones en el conjunto de prueba, y evaluamos su rendimiento en función de su capacidad para predecir los precios de las viviendas en Bogotá.

El mejor modelo que logramos predecir en Kaggle se hizo con la metodología de random forest utilizo 24 predictores en donde se encontraban 12 de nivel espacial en distancias a distintos bienes públicos e instituciones privadas, 9 dicótomas de características de hogar y las variables de conteo de cantidad de baños y cuartos y una aproximación a metros cuadrados que se hizo por medio de una imitación de datos con las pocas observaciones que teníamos. Este random forest además utilizo validación cruzada espacial para evitar problemas de autocorrelación espacial. Los hiper-parámetros terminaron siendo 8 variables aleatorias, 106 árboles y 11 observaciones como mínimo para poder hacer separaciones. Este término siendo aquel con el mejor resultado en la plataforma con un error a los valores reales de 253 millones COP.

Por otro lado, para la replicabilidad de los modelos y datos, se puede encontrar todo en el siguiente repositorio de GitHub: https://github.com/Randresil/BDML_PS2

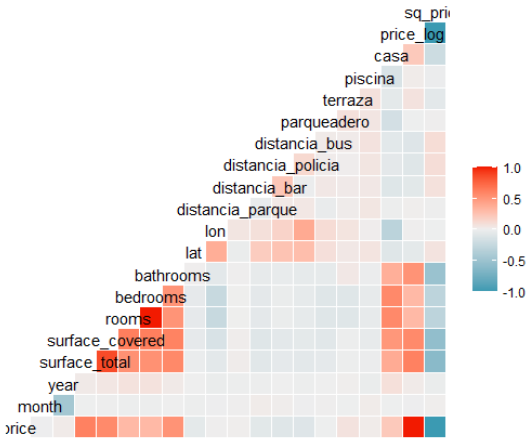
[Descripción de los datos]

En primer lugar, decidimos explorar un poco la variable a predecir, en primer lugar, exploramos la normalidad de la distribución en la variable, para ver si se puede ajustar en un paso más a allá a las regresiones lineales, y, además exploramos algunas de sus transformaciones:



Gráfica 1. Precio Original

Como parece ser evidente los precios en su estatus normal no siguen una distribución normal, pues no se comporta para nada como la línea. En los otros casos, tanto en la transformación logarítmica como en la transformación de raíz, se comporta de una mejor manera- en el sentido que los puntos se sobre ponen a la línea-. Como segundo paso decidimos hacer una revisión de la correlación de las variables, no tan solo para conocer posibles pronosticadores, sino también para poder evitar algunos problemas de correlación entre las variables predictoras. Para que el análisis sea robusto decidimos hacerlo con dos comandos distintos.



Gráfica 2: Correlaciones de variables

En este caso nos podemos fijar en dos cuestiones importantes. En primer lugar, no hay correlaciones que parezcan ser lo suficientemente grandes para poder eliminarlas. Como una segunda cuestión, vemos que las variables como área cubierta, cuartos, baños tienen una relación importante con el precio. A continuación, pueden encontrar la tabla con las estadísticas descriptivas de cada una de las variables de tipo binario, categórico y numérico que se emplearon:

Estadísticas Descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Price	38,644	654,534,675.0	311,417,887.0	300,000,000	1,650,000,000
Month	48,930	5.7	3.3	1	12
Year	48,930	2,020.3	0.8	2,019	2,021
Surface Total	9,718	160.5	1,129.9	15	108,800
Surface Covered	11,392	130.5	72.4	2	1,336
Rooms	48,930	2.9	1.0	1	11
Bedrooms	48,930	3.0	1.5	0	11
Bathrooms	48,930	2.9	0.9	1	13
Latitude	48,930	4.7	0.04	4.6	4.8
Longitude	48,930	-74.1	0.03	-74.2	-74.0
Parqueadero	48,930	0.7	0.5	0	1
Terraza	48,930	0.5	0.5	0	1
Piscina	48,930	0.1	0.3	0	1
Conjunto	48,930	0.2	0.4	0	1
Apartaestudio	48,930	0.02	0.2	0	1
Número Piso	48,930	1.8	2.1	1	20
Cantidad Baños	48,930	1.0	0.1	1	7
Cantidad Cuartos	48,930	3.0	0.5	1	11
Metros	48,930	106.6	31.2	1	300
Duplex	48,930	0.1	0.3	0	1
Vista	48,930	0.3	0.4	0	1
Penthouse	48,930	0.01	0.1	0	1
Casa	48,930	0.2	0.4	0	1
Distancia Parque	48,930	160.6	99.8	1.0	3,344.6
Distancia Estadio	48,930	3,686.9	1,677.4	9.2	7,211.3
Distancia Banco	48,930	623.6	474.5	2.4	4,724.6
Distancia Bus	48,930	918.6	648.7	3.6	6,296.8
Distancia College	48,930	2,128.5	1,116.8	4.6	5,762.1
Distancia Hospital	48,930	911.0	510.3	9.7	3,565.8
Distancia Policía	48,930	925.4	505.9	2.4	2,957.1
Distancia Universidad	48,930	961.2	558.1	1.3	4,338.0
Distancia Pub	48,930	1,518.2	989.7	13.7	7,226.4
Distancia Veterinaria	48,930	1,293.7	769.2	11.4	6,974.3
Distancia Mall	48,930	684.8	379.3	0.6	4,706.6
Distancia Reserva	48,930	4,232.4	2,121.3	404.6	11,817.0

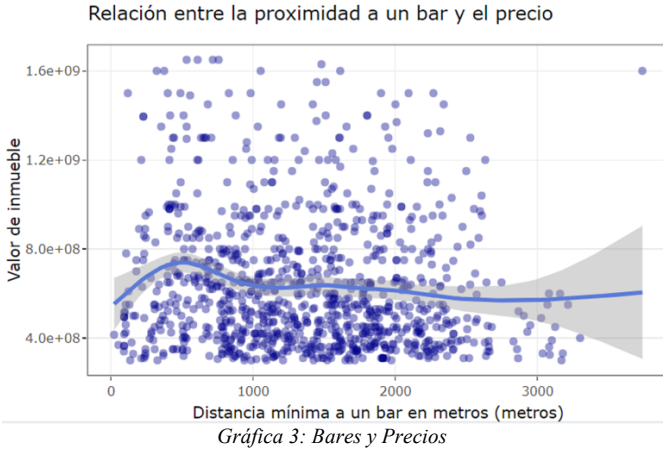
Tabla 1: Estadísticas Descriptivas

Aquellas variables que nosotros contábamos desde un inicio fueron las siguientes: Property_id, city, price, month, year, Surface_total, Surface_covered, rooms, bedrooms, bathrooms, property_type, operation_type, latitude, longitude, title y description. Aquellas que se construyeron con OpenStreetMap fueron todas aquellas que tienen "Distancia" en su nombre. En este caso a parques, estadio, banco, parada de bus, college, hospitales, estaciones de policía, universidades, bares, veterinarias, centros comerciales y reservas naturales. Finalmente, aquellas variables como parqueadero, terraza, piscina, conjunto, apartaestudio, número de piso, cantidad de baños, cantidad de cuartos, metros, dúplex, vista, penthouse y casa fueron todas extraídas de los textos en el título y en las descripciones de cada una de las observaciones. Para todas aquellas variables de tipo númerico que no quedaron con información, se utilizaron la mediana y la media como forma de imputar un dato aproximado y no perder la información recolectada por medio del manejo de textos.

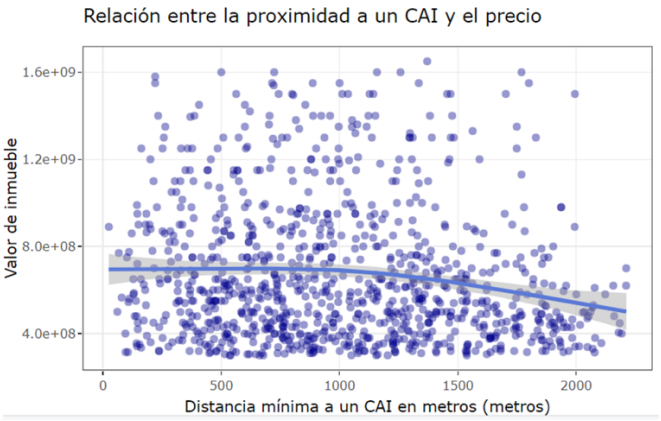
Respecto a las variables que fueron generadas por todo este proceso de manejo de textos y OpenStreetMap, encontramos una relación positiva en el caso de la proximidad a CAI y estaciones de autobús, con pendientes no muy pronunciadas. Esto sugiere que las viviendas cercanas a estos centros tienden a tener precios más altos, y a medida que la distancia aumenta, los precios tienden a disminuir. Esta relación se respalda con argumentos teóricos, ya que la accesibilidad al transporte público en el caso de las estaciones de autobús y la sensación de seguridad relacionada con los CAIs pueden influir en los precios. Claro está, la seguridad no necesariamente tiene que ser mayor pero la percepción ciudadana puede influir en el precio.

Por otro lado, en el caso de la distancia a bares, la relación no fue tan clara y la línea de regresión mostró un patrón no lineal. Lo que sí observamos es que las viviendas más costosas, en el rango de 1,200 a 1,600 millones de pesos, se encontraban a una distancia de 0 a 2,000 metros de algún bar. Mientras que en la región de 2,000 metros o más, solo encontramos viviendas de menor precio, en el rango de 400 a 800 millones. En el rango intermedio de distancias, encontramos una concentración de ambos rangos de precios, lo que puede explicar la falta de una relación clara.

En otros casos, no pudimos identificar a simple vista una relación clara entre las distancias y los precios, como sucedió con la variable de distancia a parques. Esto podría deberse a la distribución de muchos parques en la muestra, lo que dificulta la identificación de patrones claros. Por otro lado, hubo situaciones inquietantes como con la relación de precios y distancia a universidad donde no parecía ser una relación lineal y en la mitad del rango de distancia se encontraban precios bajos, y en valores lejanos estaban precios más altos de vivienda.



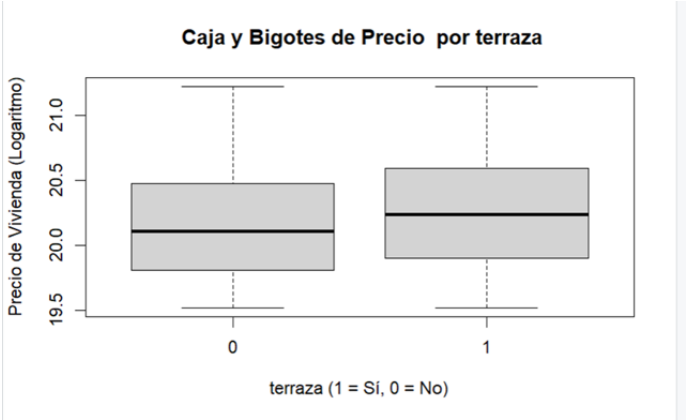
Gráfica 3: Bares y Precios



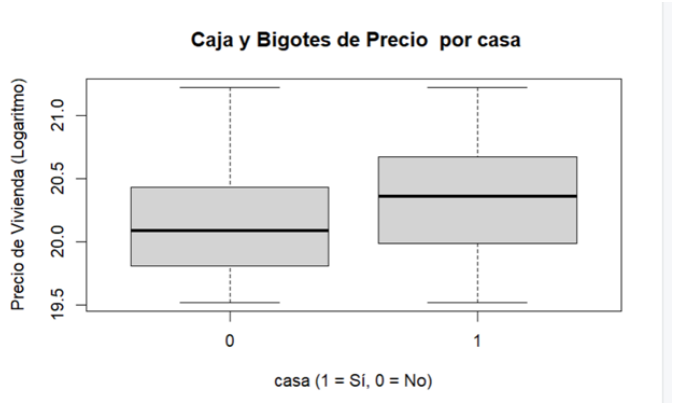
Gráfica 4: Policía y Precios

Por otro lado, al analizar las variables adicionales creadas a partir de la información proporcionada en la base de datos, examinamos el comportamiento de las variables dummy relacionadas con características como casa, terraza, parqueadero, piscina, penthouse, apartaestudio en relación con la distribución del precio de las propiedades, utilizando medidas como la media, mediana y cuartiles. Por otro lado, notamos que las casas tienen precios significativamente más altos que los apartamentos en Bogotá. Esta diferencia es aún más evidente, ya que desplaza todo el diagrama de caja y bigotes hacia arriba. La mediana del precio de las viviendas unifamiliares es aproximadamente 0.5 niveles logarítmicos mayor, y el primer cuartil de las casas se sitúa casi al mismo nivel que la mediana de los apartamentos, lo que demuestra una clara diferenciación de precios entre casas y apartamentos.

Por otro lado, las otras dos variables, piscina y parqueadero, arrojaron resultados sorprendentes. En el caso de los parqueaderos, el diagrama de caja y bigotes muestra que las viviendas sin parqueadero son incluso más caras que las que disponen de parqueadero, lo cual es un resultado inesperado. La mediana es más alta, y el tercer cuartil también. Aunque la diferencia entre las medianas no es considerable, este hallazgo fue inesperado. Respecto a la variable piscina, se había hipotetizado que las viviendas con piscina tendrían un rango de precios considerablemente más elevado que las que no la tienen. También se consideró que esta variable podría ser indicativa de otras áreas comunes en conjuntos residenciales, como clubes con piscina, que podrían influir en los precios. Sin embargo, la diferencia en precios resultó ser bastante pequeña, y es aún más evidente que características como terraza y ser una casa tienen un impacto mayor en la diferenciación de precios. Es probable que estas variables sean de mayor utilidad al incluirlas en análisis de regresión, junto con incluirlas en regresiones que incluyen arboles de decisión y random forest, ya que estas variables pueden contribuir a una segmentación más precisa de las categorías y a la identificación de patrones de precios.



Gráfica 5: Diagrama de caja para terrazas y precios



Gráfica 6: Diagrama de caja para casas y precios

[Modelos Predictivos]

Frente a las estimaciones de los modelos, las predicciones menos precisas se obtuvieron utilizando los modelos de regularización Ridge, Lasso y Elastic Net. Estas estimaciones arrojaron distancias de predicción que oscilaron entre 330 y 310 millones de error. Es posible que los errores se hayan debido a una elección inadecuada de los parámetros de penalización. Dado que nuestras estimaciones se encuentran en millones, es posible que la penalización aplicada a los coeficientes no haya sido lo suficientemente fuerte como para evitar el sobreajuste.

En términos generales, para la penalización de los cuadrados y el valor absoluto de los coeficientes en Ridge y Lasso, respectivamente, utilizamos una búsqueda en una cuadrícula de valores que varió desde 0 hasta 200. Adicionalmente, se supone que estos modelos son útiles para identificar las variables más relevantes para nuestras predicciones, ya que aquellas que no son importantes tienden a acercarse a cero, pero estas regresiones para los modelos no fueron eficientes en reconocer este comportamiento reiterativo o ineficaz en las variables predictivas. El valor óptimo para el valor del penalizador terminó siendo de valores como 0.001.

Por otro lado, y para suplir la falta de utilidad de Ridge y Lasso para elegir parámetros, utilizamos el comando r-part y los árboles de regresión en su versión sencilla para aproximarnos a la importancia de variables, al graficarlos para tener una observación visual se miraba que las variables de numero de baños y de cuartos eran las más importantes para hacer las particiones. También al incluir la distancia a universidades esta parece ser una variable predictiva que es un determinante importante del precio de los inmuebles.

Otra de las predicciones destacadas se obtiene mediante un modelo Random Forest. En este enfoque, se incorporan más variables espaciales para la predicción, incluyendo la distancia a centros comerciales y universidades. También se incluyen variables dummy que indican la presencia de un conjunto, así como una variable que se imputó, específicamente, el número de piso de las viviendas.

El proceso de limpieza y manejo de datos implicó derivar, a partir de la descripción de las propiedades, la información sobre el número de pisos en las viviendas. Para aquellos casos en los que esta información no estaba disponible, se realizó una imputación utilizando el valor medio. La especificación del modelo fue:

$$\begin{aligned} Price_i = {} & \beta 0 + \beta 1 \cdot numero\ de\ baños_i + \beta 2 \cdot numero\ de\ cuartos_i + \beta 3 \\ & \cdot piso_i + \beta 4 \cdot conjunto_i + \beta 5 \cdot Penthouse_i + \beta 6 \cdot casa_i \\ & + \beta 7 \cdot terraza_i + \beta 8 \cdot distancia\ policia_i + B9 \\ & \cdot distancia\ universidad_i \\ & + \beta 10\ distancia\ bus_i + \beta 11\ distancia\ centro\ comercial_i \\ & + \beta 12\ distancia\ parque_i + E_i \end{aligned}$$

Este nuevo modelo se llevó a cabo una búsqueda exhaustiva de hiper-parámetros, que incluyó la varianza como regla de partición. Se evaluaron opciones para el número de predictores a

considerar en la partición, variando entre 1 y 3, y se exploraron diferentes valores para el número mínimo de hojas por árbol, a saber, 5, 10 o 15. Un número mínimo mayor simplifica el modelo, mientras que uno menor aumenta su complejidad al generar más árboles con estructuras de hojas diferentes.

La selección de los mejores parámetros se basó en un proceso de validación cruzada con 5 conjuntos (k-folds). Los datos se dividieron en 5 grupos aleatorios, con 7,728 observaciones como conjunto de prueba, y el resto, 30,905 observaciones, se utilizó como conjunto de entrenamiento. Esta división se realizó de manera iterativa. La métrica utilizada para seleccionar los mejores parámetros fue el MAE (Error Absoluto Medio). Durante el entrenamiento del modelo, se obtuvo un MAE de 128 millones en la muestra de entrenamiento. Sin embargo, al aplicar el modelo a la base de datos ciega, se observó un error de 280 millones.

```
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 30914, 30917, 30914, 30915, 30916
Resampling results across tuning parameters:

  mtry  min.node.size  RMSE       Rsquared   MAE
  1      5             249219703  0.4945445  190347098
  1     10             248654687  0.4939095  189754935
  1     15             249722748  0.4896152  190679367
  2      5             193571440  0.6389264  136746323
  2     10             196044945  0.6293901  139117099
  2     15             197842370  0.6219923  140726824
  3      5             178030420  0.6796002  117531509
  3     10             180919005  0.6698729  121024735
  3     15             183590842  0.6606426  124158179

Tuning parameter 'splitrule' was held constant at a value of variance
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mtry = 3, splitrule = variance and
 min.node.size = 5.
```

El modelo resultante estableció un número mínimo de nodos igual a 5 y optó por utilizar 3 predictores para la partición. El bosque aleatorio incluyó aproximadamente 500 árboles en su construcción. Este modelo se caracteriza por su complejidad, ya que estableció un número mínimo de nodos relativamente bajo y consideró el máximo número posible de predictores para la partición.

Por otro lado, una de las metodologías predictivas que resultó ser decepcionante en términos de precisión de predicción fue Boosting. A pesar de nuestros esfuerzos en la experimentación con diversos conjuntos de hiperparámetros y tasas de aprendizaje que variaban desde 0.001 hasta 0.2, este modelo generaba predicciones con valores notoriamente alejados, en algunos casos superando los 350 millones de MAE en el conjunto de prueba de Kaggle. Incluso cuando aplicamos la misma especificación de modelo utilizada en el árbol anterior, que consistía en los mismos 13 predictores, el mejor modelo que obtuvimos tenía un MAE de 179 millones, siendo este el valor más bajo alcanzado. No obstante, aún estaba 58 millones por encima del MAE obtenido utilizando la misma especificación con el modelo Random Forest.

En nuestros esfuerzos por mejorar el rendimiento de Boosting, implementamos la validación cruzada y ajustamos la tasa de aprendizaje en comparación con el árbol previamente construido. Utilizamos la biblioteca 'caret' para permitir la construcción de procesos de árboles que aprendían de su predecesor en un rango de entre 300, 400 y 500 iteraciones.

```
38644 samples
   13 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 30915, 30916, 30915, 30915, 30915
Resampling results across tuning parameters:

  maxdepth  nu     mstop  RMSE       Rsquared   MAE
  1         0.001  300    298245698  0.2505257  235671136
  1         0.001  400    294807222  0.2523609  232519950
  1         0.001  500    291756894  0.2531921  229672527
  1         0.010  300    264759882  0.2988333  202170715
  1         0.010  400    262027381  0.3071553  199104006
  1         0.010  500    260078594  0.3156392  196965407
  2         0.001  300    291616922  0.2730348  229624666
  2         0.001  400    287070254  0.2731515  225415571
  2         0.001  500    283298393  0.2747623  221856002
  2         0.010  300    254775724  0.3460167  192713077
  2         0.010  400    251220300  0.3609119  188962782
  2         0.010  500    248838418  0.3704667  186500925
  3         0.001  300    288603115  0.3147221  226877193
  3         0.001  400    283303741  0.3161530  221933313
  3         0.001  500    278844472  0.3182759  217700320
  3         0.010  300    247152991  0.3827771  185301242
  3         0.010  400    243767511  0.3961925  181955822
  3         0.010  500    241332992  0.4063939  179607822

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mstop = 500, maxdepth = 3 and nu = 0.01.
```

Dadas dichas complicaciones con los anteriores modelos de Lasso, Ridge, Elastic Net, Arbol de regresión, Random Forest y Boosting, definimos una nueva ecuación que iba a ser evaluada por medio de Random Forest; dado que este arrojó los mejores resultados. Dicha ecuación siendo:

$$price\_i = distancia\_park + distancia\_stadium + bank + bus\_station + college + hospital + police + university + pub + veterinary + mall + nature\_reserve + parqueadero + terraza + piscina + conjunto + apartaestudio + duplex + vista + penthouse + casa + habitaciones\_numerico + bano\_numerico + metros\_num$$

Para el modelaje del que fue nuestro mejor resultado, tomamos una grilla para hiper-parámetros que contemplara variables aleatorias de 2-10, un número mínimo de observaciones por hoja de 2-12 y un total de 100-150 árboles y un proceso de validación cruzada espacial. Los valores óptimos de estos terminaron siendo 8 variables, 11 observaciones por hoja como máximo y 106 árboles. Este terminó arrojando un error con respecto a los valores reales dados por Kaggle de 253 millones de COP.

[Conclusiones]

En cuanto a las metodologías de predicción, observamos que los árboles y los Random Forest se destacaron como las mejores opciones. Es importante señalar que estos modelos tendieron a seleccionar estructuras más profundas y complejas. A lo largo del proceso, afinamos los parámetros utilizando la validación cruzada, y notamos que a medida que incluíamos un mayor número de predictores, los modelos lograban una predicción más precisa, reflejada en un menor Error Absoluto Medio (MAE). Sin embargo, esta mayor inclusión de predictores también se traducía en modelos más complejos, caracterizados por un gran número de árboles y la elección de un mayor número de predictores para las particiones. En contraste, la metodología de boosting no resultó tan efectiva, lo que sugiere un posible sobreajuste debido al número significativo de iteraciones de los árboles que permitimos en las grillas y una tasa de aprendizaje lenta.

Por otro lado, los modelos Ridge, Lasso y Elastic Net arrojaron predicciones menos precisas, posiblemente debido a desafíos en la elección de los valores de lambda y las penalidades. Variables como 'metros' y 'cantidad de baños' demostraron ser importantes predictores, y las variables categóricas 'casas' y 'penthouse' desempeñaron un papel relevante en las predicciones. En cuanto a las variables de distancia, destacaron 'distancia a universidades' y 'distancia a la policía'. Para mejorar aún más la precisión de las predicciones, implementamos una metodología de validación cruzada espacial en el modelo de mejor ajuste. Esta estrategia resultó fundamental, ya que contribuyó a mitigar lo que termina siendo la autocorrelación espacial entre los barrios de la ciudad de Bogotá. El github se encuentra en el siguiente link.