Problem Set 2: Making Money with ML?

"It's all about location location location!!!"

Nota: *la base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: https://github.com/stivenperalta/Problem_set_2*

Presentado por Stiven Peralta, Jazmine Galdos, Andrea Clavijo, Sergio Jiménez y Nicolás Barragán

I. Introducción

Este trabajo busca construir un modelo predictivo para pronosticar precios de vivienda. La principal motivación para construir estos modelos se basa en los postulados de la teoría de los precios hedónicos en la que se incluyen atributos inherentes a las características de la vivienda para determinar su valor comercial. Particularmente, en este *problem set* se busca predecir los precios de los inmuebles en Chapinero a partir de sus características estructurales (área, tipo de inmueble, número de habitaciones, baños, amenities, entre otros), así como características del entorno rescatadas del análisis geoespacial (estrato socioeconómico, tasas de criminalidad, índice de espacio público efectivo, densidad poblacional, cercanía a parques, colegios, centros comerciales, sitios turísticos, transporte público, entre otros).

Con este objetivo en mente, se elaboraron diferentes modelos predictivos, con metodologías como lasso, ridge, elastic net, random forest, entre otros, de los cuales se profundiza en los 5 mejores modelos predictivos con mejor desempeño, medidos por el criterio del error absoluto medio (MAE). Como resultado, se obtiene que Boosting con validación cruzada por barrio presenta el mejor desempeño en la predicción de precios de vivienda en la localidad de Chapinero, con un MAE fuera de muestra de COP 214.332.282. No obstante, los modelos dentro de muestra tienen a sobreajustarse aun cuando se valoran métodos de validación cruzada que corrigen correlación espacial.

II. Antecedentes

En Colombia, de acuerdo con cifras reportadas por Camacol a mayo de 2023, aunque la oferta de vivienda en Bogotá incrementó en 10,7% (5.700 viviendas) comparado con 2022, la compra disminuyó en un 30% (26.900 viviendas). Aunque los efectos del ahorro acumulado como consecuencia de las medidas tomadas para enfrentar la pandemia habían generado un incentivo para el sector de la construcción, y con ellos la compra de vivienda, el sector volvió a mostrar un comportamiento decreciente. La justificación radica en un contexto macroeconómico caracterizado por una desaceleración de la actividad económica, una alta inflación y tasas de interés elevadas (Banco de la República, 2023). Esta coyuntura ejerce presión sobre la confianza de los consumidores a la hora de comprar vivienda; por tanto, tienen especial relevancia en el análisis.

Algunos estudios que han analizado la formación de precios en el mercado de vivienda en Colombia encontraron que el estrato, la condición de entrega y el estado constructivo afectan el precio como también el área, y las distancias a parques, vías y estaciones de Transmilenio. Adicionalmente, que los incrementos de precios son más altos hacia el nororiente de la ciudad (Toloza Delgado, Melo Martinez, & Azcarate Romero, 2021). Estos son algunos de los aspectos que pueden definir las preferencias de los consumidores a la hora de adquirir vivienda en Bogotá. Para nuestro objeto de estudio, cabe mencionar que Chapinero es una de las localidades más grandes de Bogotá y está compuesta de tres grandes centros urbanos: Chapinero barrio, El Chicó y El Lago. Además, es considerada, como localidad, el centro de la nueva metrópolis al concentrar conglomerados financieros, novedosos lugares culturales y gastronómicos de Bogotá (Alcaldía Local de Chapinero, s.f.). Dichas características ejercen influencia a la hora de considerar comprar vivienda en esta localidad, por ello es interés de este trabajo conocer con mayor profundidad las variables que determinan sus precios.

III. Datos

Para el desarrollo del modelo predictivo, se tomaron los datos de ofertas inmobiliarias en la ciudad de Bogotá publicados en el portal web properati[1] entre 2019 y 2021, los cuales se descargaron de la página web de la competencia[2]. En total, la base contó con 48.930 observaciones y 16 variables con características propias de la oferta inmobiliaria (tipo de inmueble, área, número de habitaciones, número de baños, entre otras). Al ser una base de datos obtenida de publicaciones generadas por diferentes usuarios mediante webscrapping, una parte ardua del trabajo se concentró en la revisión, limpieza y corrección de esta base[3], la cual se detalla en el Anexo 1.

Al analizar el mercado inmobiliario es importante evaluar tanto las características estructurales de los inmuebles (i.e. información reportada en el portal inmobiliario) como las características del entorno en que se encuentran (Herart & Maier, 2010). De acuerdo con lo anterior, se complementó la base con información espacial extraída de las bases de datos de la administración distrital y nacional, obteniendo la identificación de barrios, unidades de planeación zonal (UPZ) y localidad para cada una de las ofertas, la cual permite analizar la correlación espacial de las diferentes observaciones. Así mismo, se identificaron datos catastrales (estrato socioeconómico, valores de referencia del suelo), indicadores de turismo (densidad de establecimientos de gastronomía y alojamiento turístico), indicadores de crimen (incidentes reportados y delitos de alto impacto), población (densidad poblacional, número de habitantes y número de viviendas). Finalmente, también se identificó la distancia de cada una de las ofertas a servicios relevantes como colegios, parques, centros comerciales y estaciones de transporte público (en el Anexo 1 se detallan las variables extraídas de fuentes externas).

Una vez realizada la limpieza de datos (detallada en el Anexo 1) se cuenta con una base de 47.149 observaciones distribuidas en una muestra de entrenamiento (***train***) con 36.863 observaciones (78%) y otra de prueba (***test***) con 10.286 observaciones (22%). Los datos de la base de prueba corresponden a los precios a predecir a partir del modelo, por lo que los valores de precio para la base de prueba no se encuentran registrados. Respecto a la base de entrenamiento, el precio de los inmuebles oscila entre 300 y 1.650 millones de pesos con un valor promedio de 644,4 millones. En la Tabla 1 se resume el análisis descriptivo de las características de los inmuebles y su entorno para las submuestras analizadas.

Tabla 1. Estadísticas descriptivas de variables continuas

	Test				**Train**			
	Min.	**Max.**	**Promedio**	**Desv. Est.**	**Min.**	**Max.**	**Promedio**	**Desv. Est.**
Características internas								
Precio (millones COP)	NA	NA	NA	NA	300,0	1.650,0	644,4	303,7
Área (m2)	15	108.800	161,5	1.083,5	2	416	134,0	63,8
Habitaciones	0	11	2,4	1,0	0	11	3,1	1,4
Baños	1	10	2,7	1,0	1	13	2,8	1,0
Características del entorno								
Indicador de espacio público efectivo	0,0	37,5	7,4	2,3	0,0	938,9	7,5	30,6
Incidentes delictivos	0	2.896	844	531	0	7.340	1.270	770

[1] https://www.properati.com.co

[2] https://www.kaggle.com/competitions/uniandes-bdml-202313-ps2

[3] Todos los análisis presentados se realizaron en R versión 4.3.1 (R Core Team, 2023).

Delitos de alto impacto	0	7.669	1.748	1.430	0	7.669	1.531	908
Valor de referencia del suelo (Millones COP/m2)	0,80	13,29	6,13	1,84	0,03	20,33	3,85	1,42
Densidad poblacional (Hab/m2)	0,000	0,203	0,024	0,018	0,000	0,392	0,031	0,025
Distancia a parques (m)	0	723	99	74	0	1.782	88	76
Distancia a colegios (m)	2	1.612	492	10.286	3	3.074	316	36.863
Distancia a centros comerciales (m)	0	2.950	754	374	0	4.598	850	479
Distancia a estaciones de Transporte público (m)	2	3.845	994	524	1	5.453	1.164	785

	Test		**Train**	
	N. obs.	**Porcentaje**	**N. obs.**	**Porcentaje**
Tipo de inmueble				
Apartamento	10.012	97,3%	28.648	77,7%
Casa	274	2,7%	8.215	22,3%
Amenities				
Estudio	3.620	35,2%	13.756	37,3%
Parqueadero	4.638	45,1%	16.184	43,9%
Balcón	2.782	27,0%	9.956	27,0%
Chimenea	3.086	30,0%	8.803	23,9%
Ascensor	2.158	21,0%	6.888	18,7%
BBQ	1.557	15,1%	6.837	18,5%
Gimnasio	2.322	22,6%	8.180	22,2%
Vigilancia	1.332	12,9%	5.526	15,0%
Jardín	279	2,7%	2.400	6,5%
Cuarto de servicio	349	3,4%	1.401	3,8%
Conjunto cerrado	90	0,9%	2.040	5,5%
Estrato socioeconómico				
NA	404	0,2%	2.579	0,9%
1	35	0,7%	9	0,0%
2	80	0,7%	552	3,1%
3	363	4,6%	5.449	16,9%
4	1.197	16,7%	10.291	35,6%
5	1.450	13,4%	9.216	22,3%
6	6.757	63,9%	8.767	21,1%

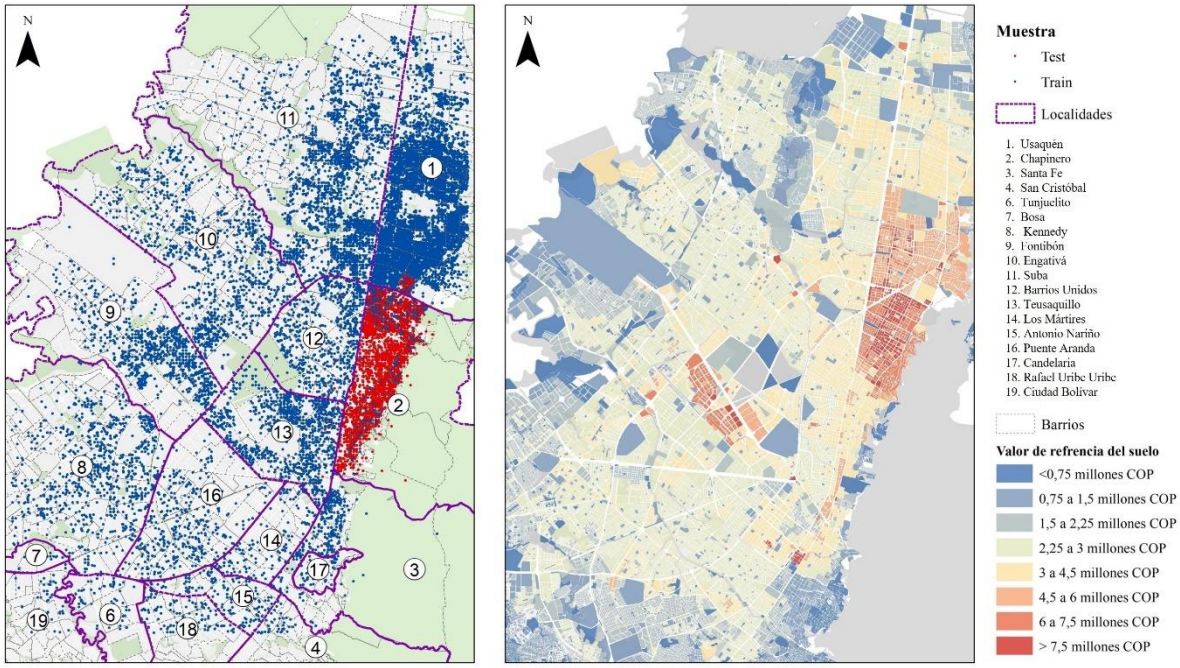
Pese a que la muestra ***train*** contiene más casas que la muestra ***test*** (22.3% vs 2.7%), en promedio las viviendas en ***test*** tienen un área mayor a las de ***train*** por un 20.5%. Sin embargo, cabe resaltar que la desviación estándar de los valores en ***test*** es mucho mayor. Este no parece ser el caso para el numero de baños, donde los valores de ambas muestras se asemejan mucho. Hay cierta diferencia en la cantidad de habitaciones, en la muestra ***train*** se evidencia una mayor cantidad y varianza en la distribución. Con respecto a las *amenities* (características adicionales

de los inmuebles) se identifica que, en general, hay una proporción similar para ambas muestras con excepción de la presencia de viviendas en conjunto cerrado y con jardín, las cuales tienen mayor participación en la muestra ***train***.

Es importante resaltar que las principales diferencias en las muestras corresponden a la concentración espacial de las mismas. En el panel izquierdo del Mapa 1**¡Error! No se encuentra el origen de la referencia.**, se evidencia la distribución de las observaciones en las localidades de la ciudad, los datos de test corresponden en su mayoría a observaciones para la localidad de Chapinero, mientras que los datos de ***train*** se encuentran distribuidos en las demás localidades de la ciudad, con muy poca participación en dicha localidad y una mayor concentración en las del norte de la ciudad. Esto se justifica en que son principalmente los inmuebles en estratos medio-alto los que se ofertan por portales web, en donde los estratos 1 y 2 tienen una participación marginal en las muestras (0,9%). Así mismo, se evidencia una mayor representación de estratos altos (5 y 6) en la muestra ***test*** (77,3%) respecto a la muestra ***train*** (43,4%). Por otro lado, se evidencia que esta última presenta en promedio una mayor densidad poblacional (0,031) que la muestra ***test*** (0,024) lo cual puede estar correlacionado con mayores precios en la muestra a predecir[4].

En el panel derecho del Mapa 1 se evidencian los valores de referencia del suelo establecidos por la Unidad Administrativa Especial de Catastro Distrital (UAECD), los cuales sirven de referencia para entender la variación de los precios en los diferentes sectores de la ciudad. Llama la atención la concentración de valores más altos en la zona norte de la localidad de Chapinero, cuyos precios se busca estimar. De igual forma, se evidencia que los datos en ***train*** tienen en promedio un valor de referencia más bajo (63%) que los valores en ***test,*** lo que representa una limitación en la capacidad predictiva ya que no se cuenta con suficiente información del precio de los inmuebles en la localidad de Chapinero, la cual tiene condiciones diferenciales al analizar la información catastral.

Mapa 1. Distribución de la muestra en las localidades - Valores de referencia del suelo

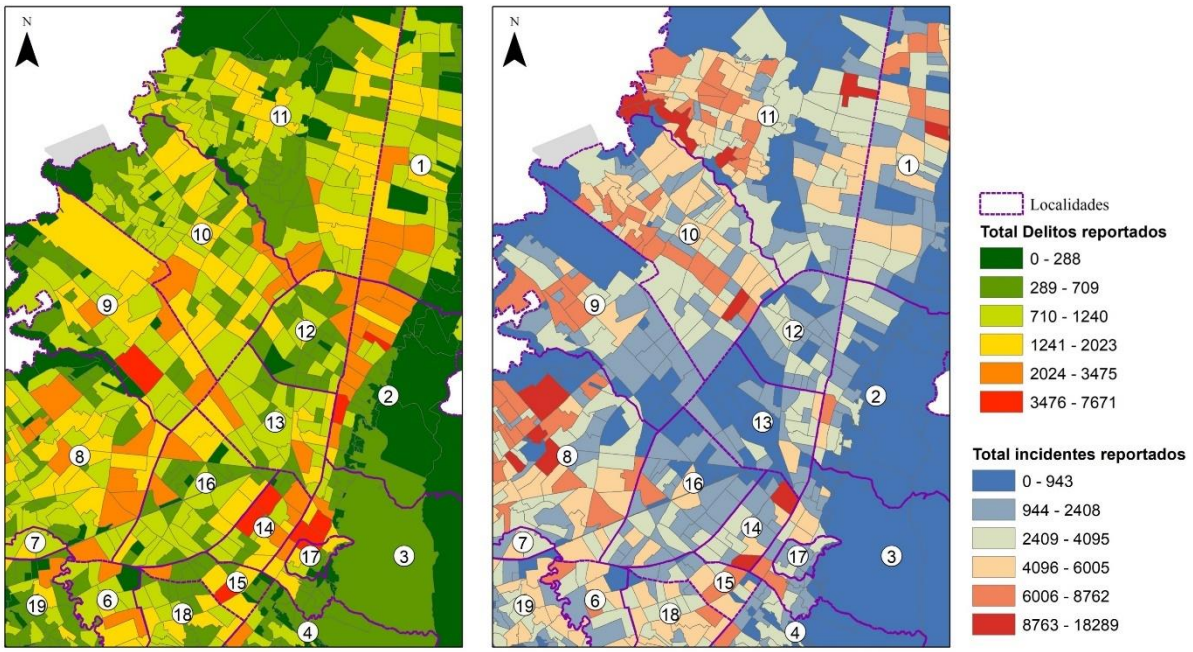


Por otro lado, resulta relevante analizar los indicadores de criminalidad que pueden tener impactos significativos en los precios del mercado inmobiliario (Gibbons & Machin, 2008).

[4] En el Mapa 4 del Anexo 2 se visualiza la distribución de los estratos socioeconómicos y densidad poblacional en las observaciones analizadas.

En ellos Mapa se identifican los valores totalizados de indicadores de criminalidad para cada barrio (Mapa 2). Se evidencia que los datos de la muestra a predecir ***test*** tienen un número menor de reportes de incidentes delictivos (riñas, maltrato, orden público, etc.). No obstante, estos corresponden a las zonas con mayor presencia de delitos de alto impacto (homicidios, hurto, violencia sexual, etc.). Dadas estas condiciones aparentemente contrarias, resulta relevante identificar en los modelos la relevancia de estas variables.

Mapa 2.Indicadores de criminalidad.



Finalmente, se hace importante analizar las condiciones de espacio público y la cercanía a servicios de los diferentes inmuebles, las cuales suponen componentes relevantes del precio (Herart & Maier, 2010) (Cardenas, Gallego, & Urrutia, 2023). Con respecto al índice de espacio público efectivo, el cual mide la cantidad de espacio público por habitante, se cuenta con valores muy similares para ambas muestras, lo cual puede deberse a la escala de agregación a nivel de UPZ. Así mismo, los inmuebles en ambas muestras se encuentran en promedio a la misma distancia de los parques aledaños. Con respecto a la cercanía a servicios como colegios y centros comerciales, se evidencian valores diferenciales en ambas muestras, en ***test*** los inmuebles están en promedio 55% más lejos de colegios, mientras que, en ***train*** se alejan un 12% más de los centros comerciales. Con respecto a la cercanía al transporte público resulta relevante que los inmuebles a predecir se encuentran en promedio 15% más cerca de estaciones de transporte público, lo cual puede implicar unos mayores precios (Vergel Tovar, Rodriguez, & Camargo, 2016).

IV. Modelo y resultados

Para el ejercicio de predicción de precios de las viviendas en Chapinero se realizaron diferentes modelos. Los enfoques de pronóstico se concentraron primero en modelos regularizados: Lasso, Ridge y Elastic Net; posteriormente, especificaciones más complejas de árboles como Random Forest y Boosting. Entre los modelos estimados, aquel que minimizó el MAE en la predicción de los valores de los inmuebles en Chapinero fue un Boosting validación cruzada por barrio, para controlar por correlación espacial, seguido de un Random Forest, Lasso y Elastic Net de acuerdo con el criterio de selección MAE (Tabla 2).

Tabla 2.Resultados MAE por modelo

Modelo	MAE
Boosting	214.332.282
Random Forest	217.267.151
Lasso	238.633.590
Elastic Net	244.093.112
Elastic Net	246.713.352

Las variables seleccionadas para la predicción varían dependiendo del modelo especificado, y las características de cada uno se describen brevemente a continuación:

- **Modelo 1. Spacial boosting**

Boosting es un modelo que emplea un método de aprendizaje que permite mejorar la predicción de los mediante el crecimiento secuencial de árboles que van creciendo usando información del error en los árboles previos, lo cual permite que el error vaya disminuyendo progresivamente. El modelo de precios de vivienda se realizó en función de los siguientes atributos:

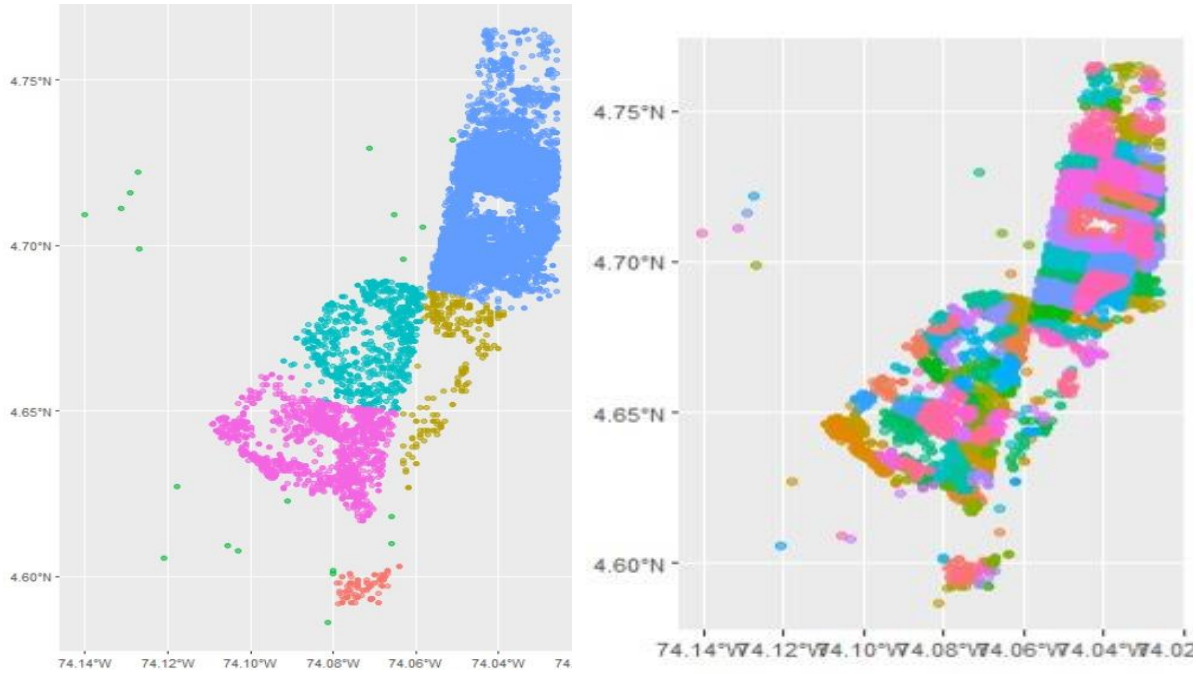
$$\boldsymbol{price = f(X)}^{5}$$

Este modelo incluye en la predicción no solo los atributos inherentes a la vivienda sino también los relacionados con las características espaciales y las variables de criminalidad que mejoraron notablemente el ajuste del modelo comparado con Lasso. En cuanto a la selección de los hiperparámetros, se empezó corriendo un Boosting con los parámetros elegidos aleatoriamente. Para empezar, se fijó la cantidad de variables a seleccionar aleatoriamente, tomando como base la raíz cuadrada del total de variables, así, se definieron valores cercanos de 7, 21 y 28. En cuanto a las divisiones de cada árbol, usamos como regla de partición (splitrule) *extratrees*. Para los nodos, aquel que minimizó el error medio absoluto fue de 135 valores seleccionados aleatoriamente. Una vez se tuvo el resultado de ese modelo y la combinación de dichos parámetros que minimizaban el error, se realizaron combinaciones de los rangos para minimizar el MAE dentro de muestra. Sumado a esto, se realizó validación cruzada seleccionando como folds 1) las localidades (MAE dentro de muestra = 195362746) y los 2) barrios (MAE dentro de muestra = 184881244). Ambos modelos mejoraron significativamente la predicción dentro de muestra. No obstante, al evaluar el MAE fuera de muestra, se encontró que aplicar validación cruzada por barrio (MAE fuera de muestra = 214332282) mejoró marginalmente la predicción modelo en comparación con la validación cruzada por localidad ((MAE fuera de muestra = 214487848). Como se puede observar, aunque la validación cruzada por barrio mejoró la predicción, también aumentó el sobreajuste del modelo y la mejora no fue significativa en la predicción fuera de muestra.

El Mapa 3 representa visualmente la validación cruzada espacial por localidad y barrio:

Mapa 3.Analisis de correlación espacial por localidad (izquierda) y barrio (derecha).

[5] Donde **X** corresponde a un vector que Incluye las variables: *habitaciones*, *área*, *baños*, *Indicador de espacio público efectivo*, *Incidentes delictivos*, *Delitos de alto impacto* (maltrato, homicidios, lesiones, hurtos a residencias, comercios, autos, motos, bicicletas y celulares), *Valor de referencia del suelo*, *Densidad poblacional*, *Distancia a parques*, *Distancia a colegios*, *Distancia a centros comerciales*, *Distancia a estaciones de Transporte público*, *Tipo de inmueble*, *Estudio*, *Parqueadero*, *Balcón*, *Chimenea*, *Ascensor*, *BBQ*, *Gimnasio*, *Vigilancia*, *Jardín*, *Cuarto de servicio* y *Conjunto cerrado*.



- **Modelo 2. Random Forest**

El enfoque de Random Forest emplea la técnica de combinar árboles de decisión junto con bagging, con la particularidad de aleatorizar las variables predictoras, evaluando cómo cambia el desempeño del modelo. Esto permite mejorar el performance ya que cada árbol se entrena con distintas muestras de datos para un mismo problema. De esta manera, al combinar sus resultados, unos errores se compensan con otros y se logra una predicción que generaliza mejor en particular.
En el modelo estimado se incluyó la base con las mismas variables empleadas en el modelo de Boosting. Al predecir dentro de muestra, se obtuvo que el MAE fue de 196.304.294 mientras que fuera de muestra el MAE fue mayor que los obtenidos con boosting (217.267.151).
La selección de los hiperparámetros se basó en la definición de una grilla que fijara las combinaciones de estos, donde la indicación dada en el código fue la de probar con diferentes tipos de nodos terminales del árbol, la cantidad de variables a seleccionar aleatoriamente en cada división del árbol (25 y 26: raíz cuadrada de las variables iniciales del dataset), y el criterio de variance que permite validar la reducción de la varianza en cada nodo del árbol. Se alternó, por ejemplo, en la cantidad de nodos del árbol, iniciando con un valor de 12.000 y se fue disminuyendo hasta llegar a 130 y 140. Al realizar las estimaciones, se obtuvo que, la combinación que minimizaba el MAE dentro de muestra fue emplear 22 variables y un mínimo de nodos de 100.

- **Modelo 3. Lasso con validation test**

Este modelo de aprendizaje se fundamentó en la siguiente forma funcional:

$$\boldsymbol{precio\_inmueble} = \boldsymbol{X\beta} + \boldsymbol{u}$$

Donde *precio* es la variable dependiente sin aplicarle logaritmo, X es la matriz de variables explicativas, β es el vector de coeficientes del modelo y u es el término de error.
Comparado con la especificación de Boosting y Random Forest, este modelo no incluyó variables de criminalidad como los incidentes de maltrato, narcóticos, riñas, robos y de orden

público porque tenían valores missing, así que se optó por excluirlas de este análisis. No obstante, en este modelo se incluyeron, a diferencia de Boosting y Random Forest, las variables de *Upz, Tegb* (densidad de establecimientos de gastronomía y bar), nivel UPZ, *Teat* (densidad de establecimientos de alojamiento turístico) y *tipología_ZIT (*zona de interés turístico). La motivación detrás de la inclusión de estas nuevas variables comparado con los dos primeros modelos fue la de evaluar especificaciones con otro tipo de variables que intuitivamente pudieran estar relacionadas con la formación de precios.
Como método de validación se procedió a correr el modelo dividiendo el set de entrenamiento en 2 submuestras, primero se predijo en una de ellas y se evaluó en otra; con esto se buscó validar los hiperparámetros que dieran el mejor desempeño. Este modelo tiene una desventaja comparada con los anteriores y es que no tiene en cuenta la autocorrelación espacial de los datos, teniendo en cuenta que en la base de entrenamiento solo hay 299 observaciones de Chapinero.
En este modelo, el MAE dentro de muestra fue de 161.471.077, mientras que fuera de muestra fue de 238.633.590. En el Anexo 2 se pueden observar los coeficientes de las variables más importantes del modelo en función de la regularización, como área, baños, estrato, estudio, habitaciones, entre otras. Como se puede observar, este modelo fue uno de los que mayor sobreajuste obtuvo.

- **Modelos 4 y 5. Elastic Net**

En la estimación de los modelos de Elastic Net se realizaron diferentes combinaciones del alpha (para definir el porcentaje de ridge y lasso que obtendría el modelo), además del lambda como hiperparámetro de regularización. Estos modelos se emplearon con las mismas variables enunciadas en el modelo de Random Forest y Boosting. Para fijar los hiperparámetros, primero se estimó el modelo sin grilla y se evaluó la combinación de alpha y lambda que minimizaban el MAE dentro de muestra. Con eso valores de referencia se creo una secuencia de 7 valores cercanos para alpha y 3 valores cercanos para lambda y, posteriormente, se evaluó cuál de todos esos reflejaba el menor MAE dentro de muestra. Así las cosas, con un lambda de 28.423.157 y un alpha de 0,2 se minimizó el error dentro de muestra (MAE dentro de muestra = 205442049). Con esta estimación el MAE fuera de muestra fue de (244093112). Como se puede observar, dado el alpha que minimiza el MAE, en el modelo, ridge tiene un mayor peso que lasso. Adicionalmente, se observa que estos modelos sobreajustaron la predicción[6].
Si bien estos modelos mostraron un buen performance, al realizar la predicción con otras especificaciones se lograron mejores resultados. De hecho, constituyó una buena base para construir a partir de ahí los modelos de árboles.

V. Conclusiones y recomendaciones

De acuerdo con la descripción y análisis de los datos, el área y el número de habitaciones fueron las variables con mayor correlacion sobre el precio de los inmuebles en Chapinero. Asimismo, otras variables como la cantidad de baños también son importantes para explicar el precio de un inmueble en la muestra utilizada como también variables geoespaciales tales como delitos de alto impacto y cercanía a colegios y centros comerciales.
El trabajo realizado muestra las diferentes especificaciones utilizadas para predecir un modelo adecuado que explique la formación de precios en el mercado inmobiliario. Todos los modelos presentados mostraron sobreajuste, sin embargo, el uso de validación cruzada espacial mejoró sustancialmente la estimación de estos dentro y fuera de muestra. Aunque el costo

[6] Es importante resaltar que también se estimó el modelo con un alpha = 0 y alpha = 1 para evaluar el ajuste de ridge y lasso respectivamente. En ambos casos, la estimación del MAE dentro y fuera de muestra fue peor que la obtenida al emplear elastic net.

computacional de mejorar la predicción en el método de boosting fue muy alto, este no mostró mejoras significativas.

Bibliografía

Banco de la República (2023). Análisis de la cartera y del mercado inmobiliario en Colombia. Informe Especial

Herart, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research. *Institute for Regional Development and Environment*.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in pure competition. *Journal of Political Economy, 82*(1).

Gibbons, S., & Machin, S. (2008). Valuing school quality, better transport, and lower crime: evidence from house prices. *Oxford Review of Economic Policy, 24*(1), 99-119.

Cardenas, J., Gallego, J. M., & Urrutia, M. A. (2023). Announcement of the first metro line and its impact on housing prices in Bogotá. *Case Studies on Transport Policy, 11*.

Toloza Delgado, J., Melo Martinez, O., & Azcarate Romero, J. (2021). Determinantes del precio de la vivienda nueva en Bogotá para el año 2019: una aproximación a través de un modelo semiparam{etrico de regresión espacial. *Ingenieria y Ciencia, 17*(34), 23 - 52.

Vergel Tovar, E., Rodriguez, D., & Camargo, W. (2016). Land development impacts of BRT in a sample of stops in Quito and Bogotá. *Transport Policy, 15*.

Anexo 1 - Procesamiento de la base de datos

A continuación, se presenta la descripción del procesamiento de datos realizado para obtener la base de datos descrita en el numeral 3 del documento:

1. Exploración de datos

Se unieron las bases ***train*** y ***test*** descargados de la página web de la competencia en Kagel[7] para luego poder hacer filtros por ubicación, en caso de que fuera necesario, y se creó la variable "*sample"* para identificar la fuente de cada una de las bases.

En una exploración inicial a los datos, se observa que hay una gran cantidad de valores omitidos (NA) en las variables *"surface_total", "surface covered", "rooms", "bathrooms"*, además de algunas en *"title"* y "*description*". Además, más del 50% de las viviendas (casas y apartamentos) tienen un precio menor a 6 millones de pesos. Por otro lado, al visualizar los precios de las viviendas respecto al área, podemos notar muchos valores atípicos, esto ya que, en términos generales, uno esperaría que los precios de las viviendas sean mayores a medida que incrementa el área. Sin embargo, esto también se puede deber a la heterogeneidad de las viviendas, por la ubicación y otras características. Otra característica importante es que el 76% de la base ***train*** corresponde a información de Apartamentos, mientras que en la base ***test*** es el 97%. Finalmente, analizamos la variable "*description*", donde encontramos información valiosa en forma de texto (chr), a partir de la cual se extrajo información a partir del procesamiento del lenguaje natural.

2. Creación de variables

Para poder obtener las características físicas de las viviendas, se optó por extraer la información de la variable *"description"*, siguiendo los siguientes pasos:

1. **Stopwords**: se eliminaron las *stopwords* de la variable *"description"*.
2. **Tokenización**: para 1, 2 y 3 palabras. Se mantuvieron las palabras repetidas en los tokens de 1.
3. **Lista de variables de interés:** se hizo una lista de variables que podrían afectar el precio de las viviendas. Para esto, nos basamos en características relevantes comunes en portales inmobiliarios y en la bibliografía de referencia (Herart & Maier, 2010). Creamos un *loop* que busca cada una de estas variables de interés en los tokens de cada observación. Si encuentra esa variable en los tokens, rompe el *loop* y crea una dummy con un valor de 1.

3. Imputación de Datos

Para reducir los valores omitidos de las variables de área y baños, se aplicaron *loops* en los tokens de 2 palabras, se capturaron los números que estuvieran en los tokens con las palabras de interés (e.g. bano, bao, baos, mts, mts, m2…). Además, se pusieron rangos máximos como filtros para evitar errores. Finalmente, en el caso de baños, para las observaciones donde no se encontró valores numéricos se aplicó un *loop* en los tokens de 1 palabra y se realizó un conteo de las veces que la palabra de interés estuviera presente (bano, banos, bao, baos).
Una vez se capturaron esos datos, se pudo reducir los valores faltantes de ambas variables (de 30,790 a 21,099 para área y de 10,071 a 3,826 para baños). Para el resto de los valores, se realizó una imputación de vecino más cercano, utilizando el paquete *mice.*

[7] https://www.kaggle.com/competitions/uniandes-bdml-202313-ps2

4. Tratamiento de valores atípicos de las variables

Con el objetivo de evaluar los valores atípicos (outliers), se utilizó una gráfica de bigotes y una prueba de outliers estandarizada para las variables *"price", "bedrooms", "banos" y "area"*. Se hizo una exploración de estos datos atípicos capturados, teniendo en cuenta otras características de las propiedades incluyendo *"property_type", "localidad" y "description"*. Se encontró que las observaciones con muchas habitaciones y baños eran aquellas que eran casas comerciales, con múltiples habitaciones y en localidades como Suba, Barrios Unidos y Engativá. Por otro lado, los *outliers* en precios se presume son debido a falta de puntos (e.g. un apartamento de 17350 m2 se esperaría que fuera en realidad 173.50m2). Sin embargo, debido a que no es posible realizar estas presunciones, se decidió eliminar todos estos valores.

5. Incorporación de datos del entorno a partir de la unión de datos espaciales

Para complementar la base de datos se realizó la identificación de variables relevantes en las bases de datos de la administración distrital y nacional, entre las que se destacan:

- Identificación de áreas de planeamiento y división espacial (Localidades, UPZ, Barrios).
- Datos catastrales (Estrato socioeconómico, Valores de referencia del suelo).
- Indicadores de turismo (Densidad de establecimientos de Gastronomía y Bar, Densidad de establecimientos de alojamiento turístico, identificación de zonas de interés Turístico).
- Indicadores de criminalidad (Numero de incidentes delictivos y delitos de alto impacto reportados en diferentes categorías).
- Indicador de espacio público efectivo.
- Datos poblacionales (Densidad poblacional, Numero de habitantes y Numero de viviendas en la manzana).
- Ubicación de servicios relevantes (Parques, Colegios, Centros Comerciales, Estaciones de transporte publico).

Una vez descargados estos datos, se realizó la unión espacial de los datos en R usando la librería *sf*, de manera tal que los puntos geográficos tomaran los valores del sector catastral, upz o manzana en la que se encuentran según la escala de agregación disponible. Para identificar la distancia de los inmuebles a los servicios relevantes se utilizó nuevamente la librería *sf* para calcular la distancia de cada inmueble con el punto de interés mas cercano en cada caso.

En la Tabla 3 se detallan las bases utilizadas, así como las variables de interés extraídas, su escala de agregación, fuentes y enlaces de descarga de los archivos utilizados los cuales también se pueden descargar en este enlace: https://drive.google.com/file/d/1pvOZqn-tUOfV-S8v01EGpBHa63mAj9E/view?usp=sharing

Tabla 3.Identificación de Fuentes externas

Nombre	Variables de interés	Escala de agregación	Fuente	Url descarga	Nombre archivo
Sectores catastrales (barrios)	Código de sector (ID) Nombre del Barrio	Sector catastral	UAECD	https://datosabiertos.bogota.gov.co/dataset/sector-catastral	SECTOR.geojson
Valor de referencia comercial m2 de terreno	valor del suelo (precio)	Manzana	UAECD	https://datosabiertos.bogota.gov.co/dataset/valor-de-referencia-por-metro-cuadrado-de-terreno	valor_ref_2023.geojson
Estratos	Estrato Socioeconómico	Manzana	UAECD	https://datosabiertos.bogota.gov.co/dataset/estratificacion-para-bogota	manzanaestratificacion.zip ManzanaEstratificacion.shp
Establecimiento de Gastronomía y Bar	Densidad de establecimientos de gastronomía y bar	UPZ	Instituto Distrital de Turismo	https://datosabiertos.bogota.gov.co/dataset/establecimiento-de-gastronomia-y-bar-bogota-d-c	establecimientos gastronomia y bar.geojson
Establecimiento de Alojamiento Turístico	Densidad de establecimientos de alojamiento turístico	UPZ	Instituto Distrital de Turismo	https://datosabiertos.bogota.gov.co/dataset/establecimiento-de-alojamiento-turistico-bogota-d-c#	establecimientos alojamiento turistico.geojson
Colegios	Ubicación de colegios	coordenadas	Secretaria Distrital de Educación	https://datosabiertos.bogota.gov.co/dataset/colegios-bogota-d-c	colegios.geojson
incidentes delictivos	Número de incidentes delictivos reportados en las categorías: - Riñas -Narcóticos -Orden Público -Maltrato	Sector catastral	Secretaría Distrital de Seguridad, Convivencia y Justicia	https://datosabiertos.bogota.gov.co/dataset/incidente-reportado-bogota-d-c	IRSCAT.geojeson
delitos de alto impacto	Número de delitos de alto impacto reportados en las categorías: - Homicidio -Hurto -Violencia sexual -Violencia intrafamiliar	Sector catastral	Secretaría Distrital de Seguridad, Convivencia y Justicia	https://datosabiertos.bogota.gov.co/dataset/delito-de-alto-impacto-bogota-d-c	DAISCAT.geojeson

Nombre	Variables de interés	Escala de agregación	Fuente	Url descarga	Nombre archivo
Zonas de interés turístico	Identificación de inmuebles ubicados en zonas de interés turístico por tipología	polígonos	Instituto Distrital de Turismo	https://datosabiertos.bogota.gov.co/dataset/zonas-interes-turistico-bogota-d-c#	zitu.geojson
parques	Ubicación de parques	polígonos	Secretaria distrital de planeación	https://datosabiertos.bogota.gov.co/dataset/parque-bogota-d-c	parques.zip Parque.shp
espacio público efectivo	indicador de espacio público efectivo (m2 de espacio público por habitante)	UPZ	DADEP	https://datosabiertos.bogota.gov.co/dataset/espacio-publico-efectivo-upz-2021	EPE_UPZ.shp
estaciones TM	ubicación estaciones TM	coordenadas	Transmilenio S.A	https://datosabiertos-transmilenio.hub.arcgis.com/datasets/estaciones-troncales-de-transmilenio/explore	Estaciones_TM.geojson
Densidad poblacional	densidad poblacional en la manzana Número de habitantes en la manzana Número de viviendas en la Manzana	Manzana	DANE	https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/visor-descarga-geovisores/#gsc.tab=0	MGN_ANM_MANZANA.geojson
localidad	localidad (ID) Nombre de la localidad	localidad	SDP	https://datosabiertos.bogota.gov.co/dataset/localidad-bogota-d-c	localidad.zip localidad.shp
centros comerciales	ubicación centros comerciales	coordenadas	SDP	https://datosabiertos.bogota.gov.co/dataset/gran-centro-comercial-bogota-d-c	Grandes_centros_comerciales.zip Grandes_centros_comerciales.shp

Anexo 2 -Ilustraciones Adicionales

Mapa 4. Distribución de la densidad poblacional (Izquierda) y estratos (derecha).

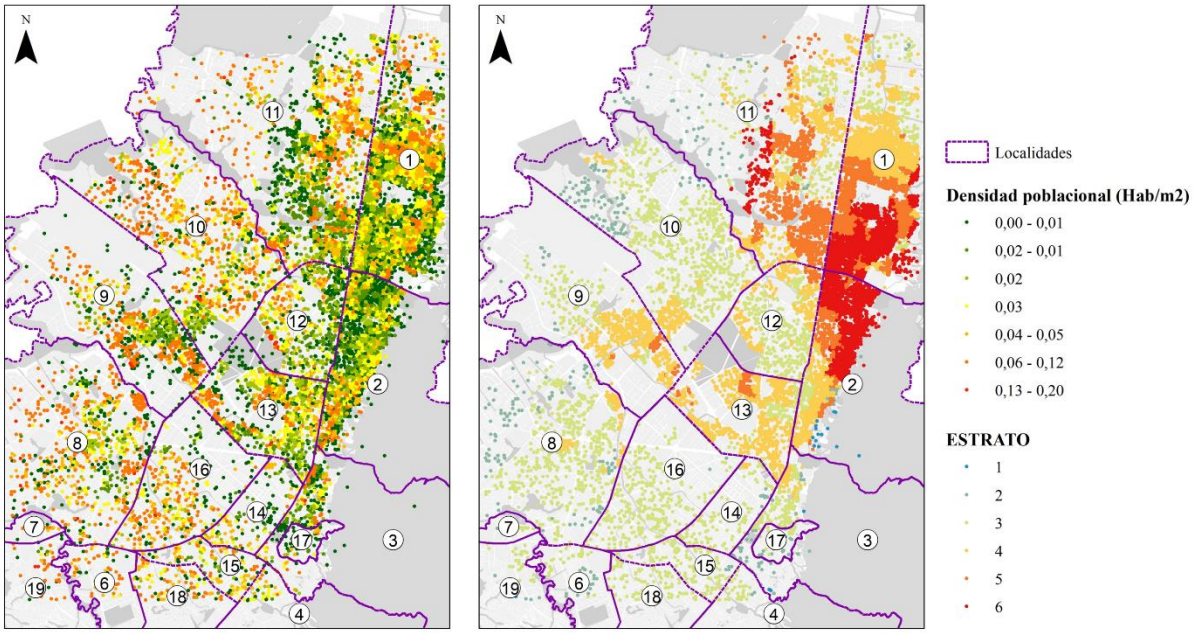


Ilustración 1.Coeficientes Lasso en función de la regularización