Universidad de los Andes
Facultad de Economía

Predicción de Precios
Profesor: Ignacio Martin Sarmiento Barbieri
Repositorio en Github

Kevin Alberth Martínez Macías – 202214432
Juan Camilo Obando Martínez – 201515899
Julián David Rojas Aguilar – 201924937
David Javier Zegarra Florez – 202213173

1. Introducción

La formación de precios en el mercado inmobiliario es compleja. Sin embargo, múltiples estudios investigan sobre los determinantes de los precios. Los primeros estudios se enfocaron en modelos simples que buscaban explicar las preferencias de ubicación en una determinada ciudad. En particular, el modelo Alonso-Muth-Mills asume que:

1) las oportunidades laborales se ubican en el centro de la ciudad.

2) los costos de transporte[1] son una función de la distancia desde la vivienda hasta el centro de la ciudad, denominado *Central Business District* (CBD) –y entendido como el centro comercial y de negocios de una ciudad–.

Dicho modelo es monocéntrico y predice que los precios de las vivienda disminuyen conforme aumenta la distancia con respecto al CBD. Lo anterior replica una regularidad empírica conocida como "gradientes de precios", donde las viviendas más cercanas al CBD tienen los precios más altos. Esto se debe a que las ciudades requieren de un alto grado de interacción entre sus ciudadanos para el correcto funcionamiento de la actividad económica y, para facilitarla, las personas buscan ubicarse cerca de las zonas laborales (fuerza centrípeta), pero la tierra es escasa y, por tanto, en el mercado inmobiliario se compite (en precios) por las localizaciones con menores costos de transporte asociados (fuerza centrífuga). Así, los modelos exhiben el *trade-off* entre el pago de altas rentas por la vivienda o el pago de altos costos de transporte.

Avances posteriores permitieron modelar las ciudades como policéntricas, donde el precio no es función únicamente de la distancia con respecto al CBD, sino que existen, a su vez, distritos secundarios que también impactan los precios. Sin embargo, el enfoque pasó de los modelos de localización intraciudad hacia los modelos de valoración hedónica, donde se estima el valor de los diferentes inmuebles como función tanto de las características de la vivienda como de la provisión de bienes públicos cerca a dicha localización. En estos modelos, el objetivo es determinar la utilidad percibida por las personas de viviendas con ciertas características o provisión de bienes públicos en la cercanía, y funcionalmente pueden describirse como:

$$P_i = f(H_{i,j}) + g(L_{i,k}) + \varepsilon_i \quad \forall j = 1, \ldots, J, k = 1, \ldots, K, \text{ donde:} \tag{1}$$

- P_i representa el precio del i–ésimo inmueble.
- f, g son funciones generales.

[1] En estos modelos, los costos de transporte son incorporados como un costo de oportunidad. Por ejemplo, si bien tomar bus o transmilenio cuesta lo mismo a pesar de que el individuo se baje en cinco o diez estaciones más adelante, aquel que se baje en la décima estación habrá destinado una mayor cantidad de su tiempo al transporte. Tiempo que pudo haber dedicado a otra actividad. Así mismo, tomar un taxi para diez cuadras también es más costoso que para cinco cuadras, y la diferencia monetaria es dinero que pudo haber dedicado a su consumo.

- H y L denotan características del hogar y de la localización, respectivamente, tal que se cuenta con J variables explicativas relacionadas con las características del hogar (como lo son los metros cuadrados, la cantidad de dormitorios, de baños, y demás) y con K variables explicativas asociadas a las provisiones públicas cercanas a la localización del inmueble (facilidad de transporte, centros comerciales, universidades, y demás).

En este contexto, nuestro objetivo es desarrollar un modelo predictivo de los precios de propiedades en la localidad de Chapinero, Bogotá – Colombia, con observaciones de todas las demás localidades de Bogotá excepto Chapinero[2]. Una estimación adecuada nos permitiría adquirir la mayor cantidad de propiedades con una inversión mínima, en tanto sobreestimaciones nos llevarían a gastar dinero innecesariamente, y subestimaciones no nos permitirían cerrar negocios. Para llevar a cabo este proyecto, hemos explorado un conjunto de datos que contiene información sobre propiedades en Bogotá, accesible a través del portal *properati* y provisto por Ignacio.

A manera de resumen, empleamos diferentes modelos. Por un lado, modelos que capturaran relaciones lineales, tales como la *regresión lineal*, *ridge*, *lasso*, y *elastic net*. Por otro lado, modelos que generaran particiones endógenas de las variables para modelar no-linealidades que no necesariamente nosotros conocemos. En este subconjunto de modelos utilizamos *bagging*, *random forest*, *boosting* tradicional, y exploramos alternativas como *XGBoost* y *LightGBM*. Finalmente, para aprovechar las bondades de ambas familias de modelos, usamos un *ensemble* estimado por mínimos cuadrados no-lineales, que es una combinación convexa de los pronósticos realizados por la totalidad de los modelos[3]. Las familias de modelos anteriores fueron entrenadas a partir de hiperparámetros elegidos con validación cruzada de cinco pliegues elegidos con bloques espaciales para evitar problemas de sobreajuste por características espaciales, pues la predicción debe ser robusta ante las particularidades de la localidad de Chapinero. El modelo que tuvo mejor estimación puntual en Kaggle fue el XGBoost, teniendo un Error Absoluto Medio (MAE) de $ 206.6 M COP, seguido del *ensemble*[4] con un MAE de $ 206.9 M COP.

En este sentido, el presente informe se divide en cinco secciones. En primer lugar, la actual introducción. En segundo lugar, los datos. En esta sección, en particular, brindamos una descripción de los datos y sus características, donde los resultados de la estadística descriptiva están asociados con los valores de las variables *post*-imputación, pues por limitaciones en la extensión del documento es preferible incluir solo una tabla –y, además, consideramos que la distribución final es más relevante–. Así mismo, detallamos el proceso de limpieza de los datos y la expansión de la información mediante fuentes de datos externas. Durante esta etapa, hemos empleado una variedad de técnicas y estrategias para garantizar la calidad y la integridad de los datos. Por ejemplo, el tratamiento de datos atípicos que se asocian a errores en la digitación. En la tercera sección listamos los modelos empleados en la estrategia predictiva, y en la cuarta sección subrayamos los resultados alcanzados por cada modelo en el 20 % de la muestra de evaluación utilizada por Kaggle. Finalmente, la última sección concluye los resultados del ejercicio.

2. Datos

A continuación detallamos los datos que empleamos para la estimación de los diferentes modelos. Por un lado, contamos con información extraída directamente de properati. En esta, existen campos con texto (título de la publicación y descripción del inmueble), que podemos explotar mediante expresiones

[2] En términos estrictos, sí contamos con observaciones de propiedades en Chapinero. Sin embargo, el precio de las viviendas para el grueso de las observaciones en Chapinero no lo observamos, sino solo una parte.

[3] En otras palabras, una regresión sin intercepto donde el parámetro asociado a cada variable debe ser no-negativo y, además, la suma de los coeficientes es igual a uno.

[4] Los modelos dominantes en la combinación lineal son el *XGBoost* y *lasso*.

regulares[5]. En particular, antes de imputar datos faltantes mediante medianas condicionales, validamos si la información se encuentra en la descripción para añadir observaciones más certeras. Por otro lado, en términos de información de fuentes externas, contamos con datos provenientes de *OpenStreetMap*, TransMilenio, y datos abiertos.

Cuadro 1. Variables empleadas en el análisis.

Variable	Descripción
Variable de interés. Precio de oferta de las propiedades en el barrio de Chapinero en Bogotá, Colombia.	P_i representa el precio de oferta de la i-ésima propiedad, con el objetivo principal de construir un modelo predictivo que pueda estimar con precisión estos precios de oferta.
Variable de identificación. Estación de Transmilenio.	Variable nominal. Identifica cada estación en la red de transporte de TransMilenio. Permite determinar estaciones de alta demanda en horas de la mañana o en la tarde, pues propiedades cercanas a estaciones con hora pico en la mañana pueden ser zonas residenciales, mientras que inmuebles cercanos a estaciones con hora pico en la tarde representan zonas laborales.
Variable de identificación. Identificador del hogar.	Identificador único de cada propiedad. Permite asociar las predicciones en Kaggle.
Variable predictora. Año.	Variable categórica. Indica el año en que se publicó el anuncio. Captura fenómenos propios de cada año, como lo son políticas de vivienda o efectos inflacionarios.
Variable predictora. Número de metros cuadrados.	Variable numérica. Indica los metros cuadrados que tiene la propiedad. A más metros cuadrados, más costoso el inmueble.
Variable predictora. Número de dormitorios.	Variable numérica. Describe cuántos dormitorios tiene una propiedad. Influye positivamente dado que con más dormitorios el valor del inmueble es mayor.
Variable predictora. Número de baños.	Variable numérica. Representa la cantidad de baños en una propiedad, ya que influye en su valor y en el nivel de comodidad. Entre mayor número de Baños, mas atractivo es el inmueble y mayor su valor.
Variable predictora. Número de parqueaderos.	Variable numérica. Indica cuántos parqueaderos tiene una propiedad. Influye positivamente dado que contar con parqueadero propio evita incurrir en costos de parqueo.
Variable predictora. Número de piso.	Variable numérica. Indica el número del piso. Típicamente se evita apartamentos en el primer piso por temas de seguridad, pero por temas de mudanzas o compras de electrodomésticos es preferible que el apartamento no se encuentre en un piso alto.
Variable predictora. Indicadora que toma el valor de uno si el inmueble es una casa y cero de lo contrario.	Variable binaria. Los apartamentos típicamente tienen seguridad privada, por lo que puede existir una diferencia en precio dada la falta de seguridad adicional en las casas.
Variable predictora. Indicadora que toma el valor de uno si el inmueble se ubica en una zona residencial y cero en otro caso.	Variable binaria. Toma el valor de uno si el número de validaciones en estaciones de TransMilenio cercanas tiene pico en horas de la mañana y cero en otro caso.
Variable predictora. Indicadora que toma el valor de uno si el inmueble se ubica en una zona laboral y cero en otro caso.	Variable binaria. Toma el valor de uno si el número de validaciones en estaciones de TransMilenio cercanas tiene pico en horas de la tarde y cero en otro caso. Inmuebles en zonas laborales tienen precios más altos en tanto disminuyen los costos de transporte asociados para actividades laborales.
Variable predictora. Indicadora que toma el valor de uno si el inmueble cuenta con parqueadero y cero en otro caso.	Variable binaria. Toma el valor de uno si, a partir de expresiones regulares usadas sobre los campos del título o la descripción del anuncio, encuentra información relacionada con parqueaderos o garajes, y cero en otro caso.

Continúa en la siguiente página.

[5]Exploramos la antigüedad de la construcción. Sin embargo, hay pocas observaciones que incluyen esta información en la descripción y, por tanto, la excluimos del análisis.

Cuadro 1. (Continuación).

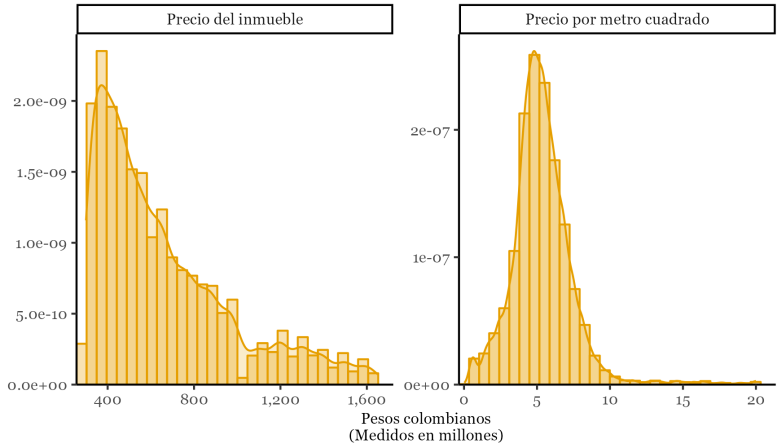
Variable	Descripción
Variable predictora. Indicadora que toma el valor de uno si el inmueble cuenta con ascensor y cero en otro caso.	Variable binaria. Toma el valor de uno si, a partir de expresiones regulares usadas sobre los campos del título o la descripción del anuncio, encuentra información relacionada con ascensores o elevadores, y cero en otro caso. Existe una mayor demanda de inmuebles con ascensor por facilidades en mudanzas, personas con discapacidades motoras, y demás.
Variable predictora. Indicadora que toma el valor de uno si el inmueble fue remodelado recientemente y cero en otro caso.	Variable binaria. Toma el valor de uno si, a partir de expresiones regulares usadas sobre los campos del título o la descripción del anuncio, encuentra información relacionada con remodelaciones recientes, y cero en otro caso. Es esperable que inmuebles recientemente remodelados tengan un costo más alto, pues evitan que el comprador deba remodelarlo una vez adquirido.
Variable predictora. Indicadora que toma el valor de uno si el inmueble cuenta con un cuarto con *walking closet* y cero en otro caso.	Variable binaria. Toma el valor de uno si, a partir de expresiones regulares usadas sobre los campos del título o la descripción del anuncio, encuentra información relacionada con *walking closests*, y cero en otro caso. Dado que es una moda reciente, puede servir como *proxy* de la antigüedad del apartamento, así como de los metros cuadrados asociados a la propiedad –pues un inmueble pequeño no desperdiciaría espacio en un *walking closet*–.
Variable predictora. Distancia hacia la estación de TransMilenio más cercana.	Variable numérica. Mide la distancia mínima con respecto a las diferentes estaciones de TransMilenio. Influye positivamente en el precio, pues representa una facilidad en el transporte. Para su cálculo, empleamos la información de datos abiertos.
Variable predictora.. Distancia hacia la universidad más cercana.	Variable numérica. Mide la distancia mínima con respecto a la universidad más cercana. Influye positivamente en el precio, pues representa una reducción en costos de transporte. Para su cálculo, empleamos la información de *OpenStreetMap*.
Variable predictora. Distancia hacia el centro comercial más cercano.	Variable numérica. Mide la distancia mínima respecto al centro comercial mas cercano. La proximidad a centros comerciales es relevante en la valoración de las propiedades. Para su cálculo, empleamos la información de *OpenStreetMap*.
Variable predictora. Distancia hacia el parque más cercano.	Variable numérica. Mide la distancia hacia el parque más cercano. Medimos, a su vez, su área. Para su cálculo, empleamos la información de *OpenStreetMap*.
Variable predictora. Distancia hacia el Comando de Atención Inmediata (CAI) más cercano.	Variable numérica. Mide la distancia hacia el CAI más cercano. Influye positivamente en el precio en tanto funciona como *proxy* de seguridad en el vecindario.
Variable predictora. Distancia hacia la cicloruta más cercana.	Variable numérica. Mide la distancia hacia la cicloruta más cercana. Influye positivamente en el precio en tanto facilita el transporte por cicla.
Variable predictora. Estrato.	Variable Categórica. Representa la clasificación socioeconómica de las diferentes propiedades. Sirve como *proxy* del nivel de ingresos. Para su cálculo, empleamos la información de datos abiertos en caso tal que no aparezca información en el campo de descripción.
Variable predictora. Número de parqueaderos en 1.5 kilómetros a la redonda.	Variable numérica. Zonas laborales suelen tener una mayor cantidad de parqueaderos a la redonda.
Variable predictora. Número de lugares de ocio en 1.5 kilómetros a la redonda.	Variable numérica. A mayor nivel de ingresos tenga el vecindario, es probable que haya una mayor cantidad de lugares de ocio, como lo son restaurantes, cafés, bares, y demás. Para su cálculo, empleamos la información de datos abiertos.

Fuente. Elaboración propia, con base en los datos de properati, *OpenStreetMap*, TransMilenio, y datos abiertos.
Nota. Las variables en **morado** fueron generadas a partir de expresiones regulares mientras que las **rosadas** fueron generadas a partir de bases externas con datos georeferenciados. Variables sin color se encuentran en la base de datos original.

A continuación, la Figura 1 presenta la distribución de los precios. Como puede verse en el panel de la izquierda, la distribución de los precios por inmueble es asimétrica y, típicamente, concentrada

en los menores valores, con varios valores atípicos. En el panel de la derecha, por el contrario, se relaciona el precio por metro cuadrado, el cual tiene una distribución más simétrica, pero con datos atípicos. Graficamos únicamente esta variable al ser la variable objetivo, y en la siguiente sección brindamos la estadística descriptiva de todas las demás.

Figura 1. Distribución de los precios de inmuebles en todas las localidades menos Chapinero.



Fuente. Elaboración propia, con base en los datos de properati.

2.1. Estadística descriptiva

A continuación, se presenta la estadística descriptiva de las variables provenientes de la base de datos original, así como de las nuevas variables que agregamos a utilizar en las predicciones. Cabe destacar que estas estadísticas ya incluyen la imputación de los datos a las variables que presentaban *missing values*. Para ello, empleamos la mediana a nivel grupo dadas las características de: *i.)* si es casa o apartamento, *ii.)* localidad, *iii.)* el estrato predominante de la manzana donde se ubica la propiedad.

Cuadro 2. Estadísticos descriptivos

Variable	N = 38,644
Precio del inmueble	
Promedio (Desviación std)	654,534,675 (311,417,887)
Mínimo y máximo	(300,000,000, 1,650,000,000)
Número de dormitorios	
Promedio (Desviación std)	3 (2)
Mínimo y máximo	(0, 11)
Número de baños	
Promedio (Desviación std)	2.87 (1.20)
Mínimo y máximo	(1.00, 35.00)
Número de metros cuadrados	
Promedio (Desviación std)	132 (147)
Mínimo y máximo	(1, 17,137)
Distancia a universidades	
Promedio (Desviación std)	1,060 (566)
Mínimo y máximo	(3, 4,338)
Número de lugares de ocio	
Promedio (Desviación std)	30 (21)
Mínimo y máximo	(0, 166)
Distancia a estación de Transmilenio	
Promedio (Desviación std)	1,155 (784)
Mínimo y máximo	(1, 5,453)
Distancia al mall	
Promedio (Desviación std)	660 (384)

Continúa en la siguiente página.

Cuadro 2. (Continuación).

Variable	N = 38,644
Mínimo y máximo	(1, 4,707)
Distancia al parque	
Promedio (Desviación std)	161 (101)
Mínimo y máximo	(1, 3,345)
Área de parques	
Promedio (Desviación std)	7,075 (25,332)
Mínimo y máximo	(20, 975,008)
Distancia al CAI	
Promedio (Desviación std)	723 (369)
Mínimo y máximo	(2, 2,165)
Distancia a ciclovía	
Promedio (Desviación std)	561 (568)
Mínimo y máximo	(0, 5,868)
Número de parqueaderos	
Promedio (Desviación std)	30 (21)
Mínimo y máximo	(0, 166)
Número de piso	
Promedio (Desviación std)	2 (2)
Mínimo y máximo	(1, 20)
Año,	
2019	7,149 (18 %)
2020	12,983 (34 %)
2021	18,512 (48 %)
Estrato socioeconómico,	
Seis	446 (1.2 %)
Cinco	1,138 (2.9 %)
Cuatro	2,590 (6.7 %)
Tres	10,983 (28 %)
Dos	13,320 (34 %)
Uno	5,558 (14 %)
Sin estrato	4,609 (12 %)
Dummy de casa,	9,252 (24 %)
Dummy de zona residencial,	9,274 (24 %)
Dummy de zona laboral,	26,919 (70 %)
Dummy de parqueadero,	26,738 (69 %)
Dummy de elevador,	6,933 (18 %)
Dummy de remodelada,	4,674 (12 %)
Dummy de walking closet,	1,600 (4.1 %)

Fuente. Elaboración propia.

Nota. La base de datos tiene un total de 38,644 observaciones. Si la variable es categórica, presentamos el número de observaciones en dicha categoría, así como el porcentaje de observaciones en dicha categoría con respecto al total de observaciones entre paréntesis. Si la variable es numérica (precio o número de dormitorios), presentamos el valor promedio, la desviación estándar, y los valores mínimos y máximos.

Respecto a los datos en el cuadro anterior, observamos que en promedio los inmuebles cuestan 655 millones de pesos colombianos, donde el inmueble más barato cuesta 300 M y el más caro vale 1,650 M. Adicionalmente, observamos tres variables muy relacionadas con el valor económico del inmueble. Primero, el número de dormitorios en promedio es de 3 con una desviación estándar de 2. Además, hay propiedades con 0 dormitorios y otras con un máximo de 11. Segundo, el número de baños en promedio es 2.87 y una desviación estándar de 1,2. Asimismo, hay un mínimo de un baño, pero se llega hasta las 35 unidades de este por propiedad. Tercero, el total de metros cuadrados, en promedio son 132, pero su desviación estándar es mayor (147), esto indica que hay una gran variación del tamaño de estos inmuebles considerando que hay tanto casas como apartamentos. La propiedad más pequeña en venta es de 1 metro cuadrado, mientras la más grande tiene 17,137 metros cuadrados.

Por otro lado, de las variables agregadas, la distancia a universidades en promedio es de 1,060 metros, pero hay propiedades que están muy cerca a estas (a 3 metros) y las más lejanas lo están a 4,338 metros. Respecto a los lugares de ocio, en promedio existen 30 de estos en un radio de 1,5

kilómetros, pero es posible que no lugares de ocio cercanos, o por el contrario, el máximo de este en dicho radio sera de 166. Además, la distancia la estación de Transmilenio en promedio es de 1,155 metros con una desviación estándar de 784. Sin embargo la distancia más cercana de una propiedad es de un metro, mientras la distancia más lejana a una estación es de 5,453 metros.

Adicionalmente, las distancias al mall o al parque más cercano son en promedio 660 y 161 metros, respectivamente. En cuanto al primero, la distancia mínimca es de 1 metro, mientras la más lejana es de 4,707 metros. Por el lado del segundo, la distancia mínima de una propiedad al parque más cercano es de 1 metro, mientras la más larga es de 3,345 metros. También, consideramos el área de los parques más cercanos, donde en promedio es de 7,075 metros cuadrados, con una desviación estándar de 25,332. Esto indica una gran variación en los parques, pues hay un mínimo de 20 y un máximo de 975,008 metros cuadrados.

Asimismo, la distancia promedio al CAI y la ciclovía más cercana es de 723 y 561 metros con una desviación estándar de 369 y 568, respectivamente. Además, la distancia más corta a un CAI es de 2 metros, mientras la más lejana es de 2,165 metros. Mientras que la distancia más corta de una propiedad a una ciclovía es de 0 metros, es decir, justo tiene una ciclovía al frente, mientras la distancia máxima es de 5,868 metros. En cuanto al número de parqueaderos en un radio de 1.5km, el promedio es de 30, pero el mínimo de estos es 0 y hay un máximo de 166 de dichos espacios. Por otro lado, el número de piso donde se encuentra el inmueble en promedio es el segundo, pero el rango de pisos va desde el primero hasta el vigésimo nivel.

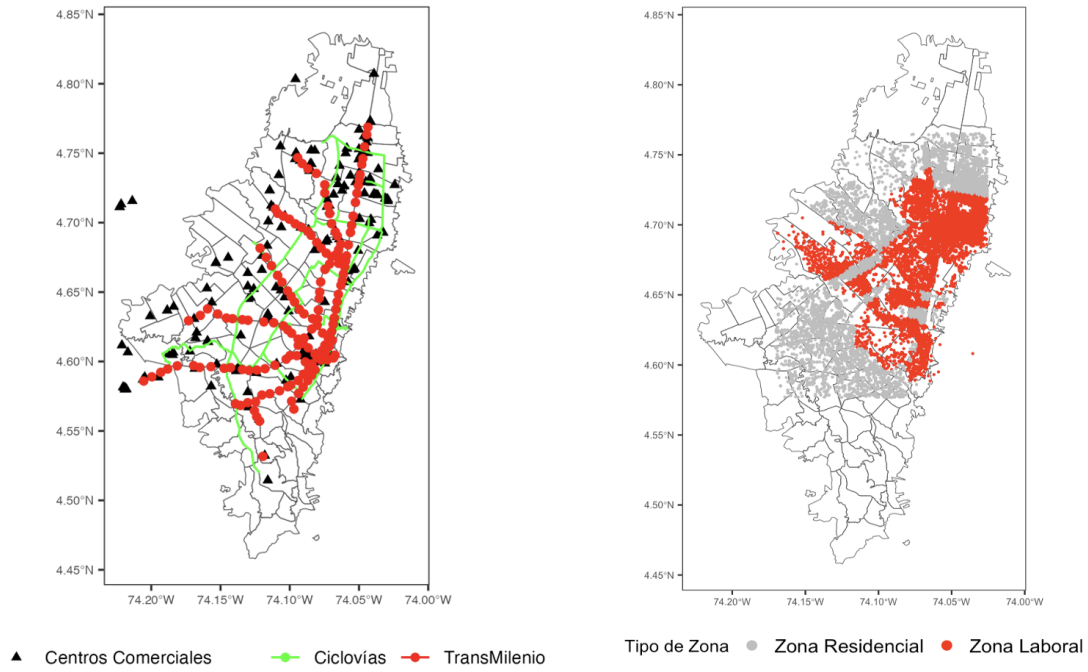
También, las observaciones de la base de datos original tienen diferente distribución a lo largo de los años, siendo la mayoría del 2021 (48 %), la segunda del 2020 (34 %) y el resto (18 %) del 2019. En cuanto a la distribución de las propiedades según estrato socieconómico, la mayoría se encuentra en el estrato dos (34 %), seguido del tres (28 %), el uno (14 %), sin estrato (12 %), el cuatro (6.7 %), el cinco (2.9 %) y el seis (1.2 %). Por último, analizamos las variables dummy de las siguientes observaciones. Por un lado, el 24 % de las propiedades son casas (en oposición a los apartamentos), el 24 % se encuentra en una zona residencial, mientras el 70 % de viviendas está en una zona laboral. Adicionalmente, las propiedades que tienen parqueadero son un 69 %, las que tienen elevador son un 18 %, los inmuebles remodelados representan un 12 %, y los que tienen un walking closet son solo el 4.1 %.

2.2. Información espacial

En la Figura 2 ubicamos, geográficamente, algunos lugares en los que nos basamos para la construcción de algunas variables que alimentan nuestros modelos. El mapa a la izquierda muestra los centros comerciales, de los que sospechamos valorizan los inmuebles cercanos. En verde están las ciclovias, pues la facilidad en el transporte mediante ciclas también disminuye los costos de transporte y por último, las estaciones de transmilenio sin incluir las estaciones de SITP, ya que hay muchas más pero la mayoría de las veces su función es servir como alimentadores del sistema de TransMilenio. Con estas se construyeron las variables que asignan a cada inmueble la distancia más corta a cada una de las 3 categorías de lugares. Las demás categorías construidas con esta misma lógica (distancia más cercana) no fueron añadidas porque cargan de manera excesiva de información el mapa.

Por otro lado, el mapa de la derecha muestra a todos los inmuebles del *trainset* con un color que representa la variable categórica binaria a la que pertenece: rojo si pertenece a una zona laboral y gris si el inmueble está en una zona residencial. Esto considerando que, en la zonas laborales, los inmuebles pueden tener mayores costos debido a una mayor demanda (pues las personas prefieren vivir más cerca a su trabajo), mientras que en las zonas residenciales los precios pueden ser menores por los costos de transporten y tiempo en que incurren las personas por temas de desplazamiento. La definición de zona residencial o laboral se realiza a partir del número de validaciones en estaciones de TransMilenio cercanas, pues zonas residenciales tienen el pico de sus validaciones en horas pico de la mañana, mientras que zonas laborales tienen el pico de sus validaciones en horas pico de la tarde.

Figura 2. Mapas de Bogotá con algunas variables espaciales empleadas en la predicción.



Fuente. Elaboración propia, con base en los datos de *OpenStreetMap*, TransMilenio, y datos abiertos.

3. Modelos

En todos los modelos, para encontrar los hiperparámetros, tenemos el problema de que ubicaciones cercanas son dependientes como consecuencia de la autocorrelación espacial. En términos del modelo de precios hedónicos, es equivalente a afirmar que propiedades cercanas se benefician de la misma provisión de bienes públicos. Luego, las características de la localización donde está incrustada la propiedad son importantes para explicar los precios de la vivienda, pero la base de datos con que contamos originalmente no cuenta con ninguna de esas características. Es decir, tenemos un sesgo por variable omitida que repercutirá, negativamente, en el pronóstico.

Ello implica que la validación cruzada tradicional no sea adecuada para determinar los hiperparámetros óptimos y, por el contrario, tengamos que usar la validación cruzada por bloques. En este sentido, generamos cinco pliegues o *folds* para la validación de los hiperparámetros pseudo-fuera de muestra, de forma tal que los datos "no observados" (porque se encuentran en el k–ésimo pliegue con que no se entrenó los datos) están agrupados por bloques, tal que los modelos aprenden de precios sin capturar las características no observables presentes en la localización.

Una vez encontrados los hiperparámetros que minimizan el MAE de la validación cruzada por bloques, estimamos nuevamente el modelo con la totalidad de la muestra y pronosticamos los precios de las propiedades ubicadas en Chapinero. En particular, los modelos explorados son:

1) **Regresión lineal**. Modelo lineal que estima los parámetros mediante mínimos cuadrados insesgados. En ejercicios de predicción no es eficiente en tanto minimiza el sesgo a costa de alta varianza.

2) **Ridge**. Modelo lineal que minimiza el cuadrado de los errores más una penalidad cuadrática que, aunque sesga el estimador, genera mejores resultados de pronóstico por fuera de muestra. En este modelo hemos probado con 100 valores diferentes de penalización que van desde un

valor de $\lambda = 0{,}001$ hasta $\lambda = 1000$. Además, asumiendo que para algunas variables la relación con el precio no sigue una relación lineal, hemos considerado:

- Relaciones cuadráticas para la distancia con calles principales y centros comerciales, en tanto puede existir una relación cóncava. Intuitivamente, no es totalmente bueno vivir encima de una calle o avenida principal, principalmente por temas de contaminación auditiva y visual. Así, el máximo incremento en los precios de la vivienda se da, típicamente, a unas cuadras de las avenidas principales y no sobre estas.
- Interacciones entre la variable binaria de apartamento y el parqueadero. Intuitivamente, dada la densidad poblacional de Bogotá, nuevos apartamentos están siendo construidos sin parqueadero, lo cual valoriza todos aquellos apartamentos que tienen uno incluído a diferencia de los demás.

3) **Lasso**. Modelo lineal que minimiza el cuadrado de los errores mediante una penalidad L1 (valor absoluto) que, además de sesgar los estimadores y disminuir la varianza, selecciona las variables más importantes para el ejercicio de pronóstico al volver algunos coeficientes cero. Para la grilla de hiperparámetros, tomamos los mismos valores que en *Ridge* y adoptamos la misma forma funcional.

4) **Elastic Net**. Modelo lineal que combina los tipos de penalidad de *lasso* y *ridge*. Al igual que antes, usamos 100 valores diferentes de penalización que van desde un valor de $\lambda = 0{,}001$ hasta $\lambda = 1000$, pero adicionalmente incluimos diferentes valores de proporciones entre lasso y ridge, usando 20 valores diferentes que van desde $\alpha = 0$ –es equivalente al modelo de *ridge*– hasta $\alpha = 1$ –que es igual al modelo de *lasso*–. Como en casos anteriores, la forma funcional es la misma.

5) **Bagging**. *Bagging*, estrictamente, corresponde a la agregación de las predicciones de un mismo modelo estimado sobre conjuntos de observaciones diferentes y obtenidas a partir de *boostrap* con reemplazo. En este caso en particular, utilizamos como modelo un árbol de regresión. Para su calibración, utilizamos 25 árboles[6], y calibramos hiperparámetros relacionados con el número mínimo de observaciones en los nodos finales (desde 100 hasta 2,000 propiedades) y el costo de complejidad (desde 0.001 hasta 1,000).

6) **Random Forest.** En *bagging*, si existe un predictor fuerte, los distintos árboles son muy similares entre sí. Por tanto, la agregación de los pronósticos individuales es el promedio de pronósticos muy similares y puede no haber ganancia en la capacidad predictiva. En este sentido, elegir un subconjunto aleatorio de los predictores antes de estimar el modelo disminuye la correlación entre los diferentes modelos. En particular, con respecto a los hiperparámetros, evaluamos una selección desde tres hasta 12 variables independientes (de un total de 25), observaciones mínimas en las hojas finales desde 100 viviendas hasta 2,000 propiedades, y una estimación desde 100 hasta 2,000 árboles.

7) **Boosting.** En *boosting*, ajustamos árboles que aprenden de los errores del árbol inmediatamente anterior –dado que es un modelo secuencial e iterativo–. En particular, con respecto a los hiperparámetros, evaluamos observaciones mínimas en las hojas finales desde 100 viviendas hasta 2,000 propiedades, una estimación desde 100 hasta 2,000 árboles, y una tasa de aprendizaje desde 0.001 hasta 0.1.

8) **XGBoost**. Es una extensión de *boosting* explicado previamente, pero donde es posible definir una función objetivo diferente a la pérdida cuadrática. Específicamente, permite seleccionar los árboles más relevantes durante la estimación a través una penalización tipo *lasso* y, adicionalmente, no todo árbol aporta de la misma forma al pronóstico final. En particular, cada nuevo árbol aporta menos a la predicción final. Los hiperparámetros asociados son la

[6] Si bien sería provechoso utilizar más árboles, por capacidad computacional no fue posible.

profundidad del árbol (medido mediante el número de ramificaciones), el número de árboles, la tasa de aprendizaje, y el número de variables elegidas aleatoriamente en cada iteración.

9) **LightGBM.** Implementación desarrollada por Microsoft que agiliza el procedimiento de XGBoost.

10) **Ensemble**. Modelo lineal estimado por mínimos cuadrados no-lineales (dadas unas restricciones impuestas sobre los parámetros), que es una combinación convexa de los pronósticos realizados por la totalidad de los modelos. En otros términos, es una regresión sin intercepto donde el parámetro asociado a cada variable debe ser no-negativo y, además, la suma de los coeficientes es igual a uno. Es estimado tomando como variables independientes no las variables listadas en el Cuadro 1, sino los pronósticos realizados por los demás modelos. Esto le permite aprovechar las bondades de los diferentes modelos. Intuitivamente, esperamos que el aumento en el sesgo pero la disminución de la varianza de unir modelos como *elastic net* a modelos como *xgboost* lleve a mejores predicciones.

4. Resultados

A continuación, presentamos los resultados de los distintos modelos que probamos en las predicciones de Kaggle (con el 20 % de la muestra de evaluación correspondiente al 20 % de los precios observados para Chapinero) e incluimos sus respectivos MAE.

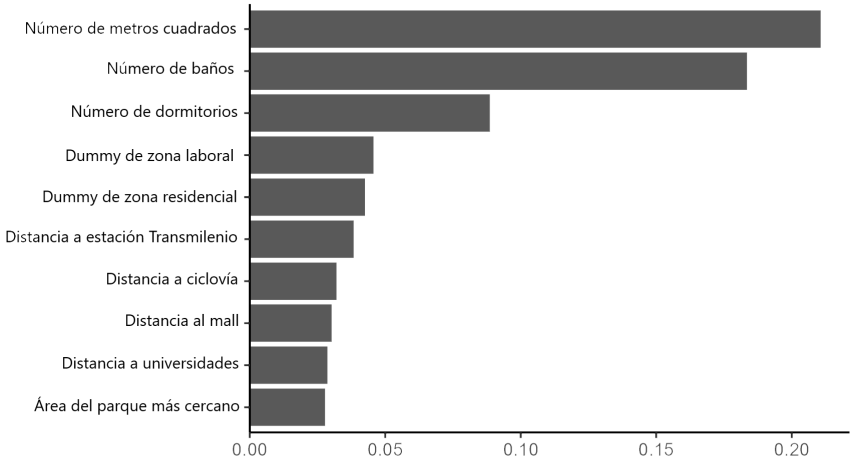
Cuadro 3. Resultados de modelos por fuera de muestra.

Modelo	MAE
Regresiónl lineal	271,137,231.16623
Lasso	270,977,106.41008
Ridge	269,486,393.29093
Elastic net	269,315,149.66851
Bagging	264,721,052.68464
Random Forest	253,060,617.54235
Boost	219,231,700.3719
Light GBM	223,151,130.47151
XG Boost	206,630,416.61886
Ensemble	206,956,506.15776

Merece la pena señalar que el modelo con mejor predicción por fuera de muestra es el XGBoost. Sin embargo, al mirar los resultados de la validación cruzada, si bien XGBoost tiene la menor estimación puntual del MAE (y por eso es preferido), la desviación estándar en el MAE (dado que contamos con la estimación para los cinco *folds* de la validación cruzada) es menor en el *ensemble* relativo al XGBoost. En particular, XGBoost, tiene una varianza de 7.2 millones, mientras que *ensemble* de 5.5 millones. Lo anterior, tal y como mencionamos previamente, se debe a la disminución en la varianza como consecuencia de incorporar modelos lineales menos complejos.

Así mismo, si bien, los modelos de *boosting* pierden interpretación en relación a un modelo lineal o CART típico (es decir, de un único árbol), todavía es posible ver qué variables aparecen en la mayoría de los árboles del *boosting*. La Figura 3, muestra que los predictores más fuertes son el número de metros cuadrados, el número de baños y de dormitorios, el tipo de zona –siendo esta residencial o laboral–, la distancia con respecto a la estación de transmilenio más cercana, y edificios como lo pueden ser centros comerciales o universidades.

Figura 3. Variables más utilizadas en los modelos de *boosting* para la predicción del precio.



Fuente. Elaboración propia.
Nota. El eje vertical señala las variables por orden de importancia. El eje horizontal denota la cantidad de veces que apareció la variable en los distintos modelos estimados. Las variables son, en su orden: metros cuadrados del inmueble, número de baños, número de dormitorios, variables binarias que indican si la propiedad está ubicada en una zona laboral o residencial, la distancia mínima hacia una estación de transmilenio, la ciclovía, un centro comercial, y universidades, y el área del parque más cercano.

En este sentido, el mejor modelo predictor, según el MAE de 206 millones de pesos colombianos, es el XGBoost. Este fue elegido para la competencia en la plataforma Kaggle.

5. Conclusiones

Dados los resultados, podemos observar una gran ventaja en las predicciones de los modelos no lineales (*xgboost*, *lightgbm*, *bagging*, *random forest*, y *boosting*), sobre los resultados sobre los modelos lineales (regresión lineal, *ridge*, *lasso* y *elastic net*). Si bien resulta extraño que el modelo *ensemble* no se beneficie de unir las bondades de *xgboost* y *lasso* (los mejores predictores de cada familia de modelos), puede ser consecuencia de aleatoriedad en la muestra. En particular, la varianza del MAE asociada al modelo de *ensemble* fue menor y, por tanto, puede que se desempeñe mejor que el XGBoost evaluado sobre la totalidad de la muestra de evaluación y no solo el 20 % de dicha muestra.

No obstante, y a manera de conclusión, un error de más de 200 millones de pesos es demasiado alto. Es necesario explorar más variables u obtener más observaciones para mejorar la capacidad predictiva, pues errar en un precio de compra en promedio por 200 millones de pesos puede traer consecuencias económicas negativas para la firma inversora. Un error de 40 millones, aproximadamente 10,000 USD, parece un objetivo razonable.