Problem Set 2: Making Money with ML?
"It's all about location location location!!!"

Integrantes: Juan Diego Duarte[1]; Erick Julian Villabon[2]

1. Introducción

La predicción de precios de vivienda a través de modelos analíticos se ha convertido en una herramienta esencial en el ámbito de bienes raíces, proporcionando valiosas perspectivas para aquellos involucrados en la toma de decisiones relacionadas con la compra y venta de propiedades, la cual presenta una compleja relación entre los precios de los bienes y las valoraciones subjetivas de sus características específicas en las cuales en un mercado competitivo, las firmas diferencian sus productos en función de estas características y los consumidores están dispuestos a pagar por activos que tienen características deseables Rosen (1974). En este contexto, se exploran diversas formas de aprovechar estas predicciones como herramientas estratégicas de aproximación al precio de las viviendas en la localidad de Chapinero en Bogotá, donde factores como la ubicación, cercanía a estaciones de transporte masivo o características como el número de habitaciones o baños deben ser considerados en conjunto para contar una buena aproximación del precio de dichas viviendas.

Con el fin de obtener una predicción precisa de los precios de las viviendas en Chapinero, se han aplicado diversos modelos predictivos. En este contexto, hemos seleccionado variables que reflejan características que podrían influenciar en el precio de las viviendas, estas variables las dividimos en dos grupos importantes: características de la ubicación a través de variables espaciales, debido a que la ubicación es uno de los factores más significativos en la determinación del precio de una vivienda como la proximidad a transporte, parques, universidades, teatros, estaciones de policías, áreas comerciales, zonas verdes, entre otras; y características propias de la viviendas, como el tamaño de la vivienda, el número de habitaciones, baños, el estrato socioeconómico de la vivienda, entre otros factores importantes a considerar (Maté de Dios, 2022; Grajales, 2019).

En este orden de ideas, contamos con dos bases de datos –test– donde se encuentran las viviendas ubicadas en la localidad de Chapinero a las cuales vamos a predecir el precio; –train– base donde se están las demás viviendas con las cuales vamos a entrenar nuestros modelos, cada base de datos

[1] Código: 202011999
[2] Código: 201815677

contiene información como la identificación y coordenadas de las viviendas, título y descripción de las viviendas, y algunas variables propias de caracterización, adicionalmente, se crean variables espaciales como la cercanía a estaciones de Transmilenio, universidades, parques, entre otras. A partir de estos datos utilizamos modelos de regresión lineal, Ridge, Lasso, Elastic Net, CART, Random forest y Boosting para realizar nuestras predicciones de los precios de viviendas.

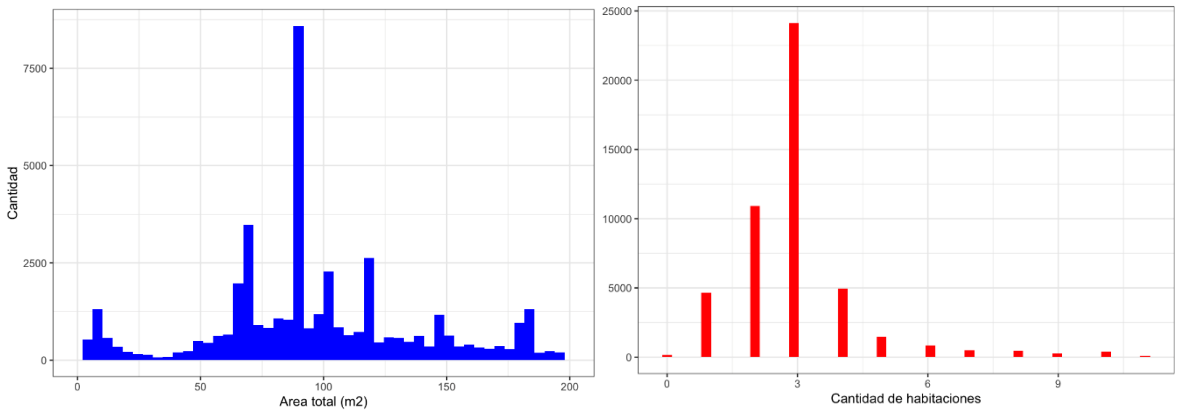
El siguiente trabajo se desarrolla de la siguiente manera: la sección 2 proporciona información sobre la descripción y el tratamiento de los datos. Posteriormente, la sección 3 presenta el modelo con mejor puntuación presentado para evaluación. Por último, la sección 4 concluye.

2. Datos

Los datos corresponden parten de dos bases de datos, la base de entrenamiento del modelo "train" y la base de prueba "test", ambas bases contando con las mismas variables y tratamiento de estas las cuales fueron seleccionadas después de una revisión de literatura y accesibilidad que se tenía. Al iniciar, las bases contaban con variables como área total; habitaciones; cuartos; baños; tipo de vivienda; locación a través de coordenadas de latitud y longitud; título y descripción de cada inmueble, a los cuales al revisarlos cuidadosamente se podrán encontrar valores faltantes.

Debido a la importancia de tener variables que cuenten con la totalidad de los valores y no perder observaciones importantes al momento de realizar predicciones, se hizo una serie de pasos para completar los datos de cada variable, iniciando por la imputación de acuerdo a valores tomados de la descripción de cada vivienda, para este paso se realizó la normalización del texto evitando inconvenientes al momento de realizar la busque dentro del mismo, seguido de tener en cuenta que podrían haber palabras mal escritas o sinónimos de las mismas se procedió a buscar de acuerdo a una lista de posibles formas en que podría estar la información de las variables, si al realizar este paso las variables seguían presentando datos faltantes, se proseguía a imputar la mediana de acuerdo a ciertas divisiones que en la mayoría de casos se tomó la variable de dormitorios –bedrooms– y debido a que es la única variable que contiene el 100% de sus datos, se realizó la agrupación de viviendas dependiendo del número de dormitorios para imputar la mediana dependiendo del grupo a los datos faltantes de las viviendas. Por último, cabe mencionar que antes de la imputación de dichos valores se verifico la existencia de datos fuera de lo común para tratarlos, en la mayoría de los casos al 1% más alto se le imputo la mediana que resultaba al momento de omitir ese 1%, el resultado parcial de estos datos podrá verse en la figura 1.

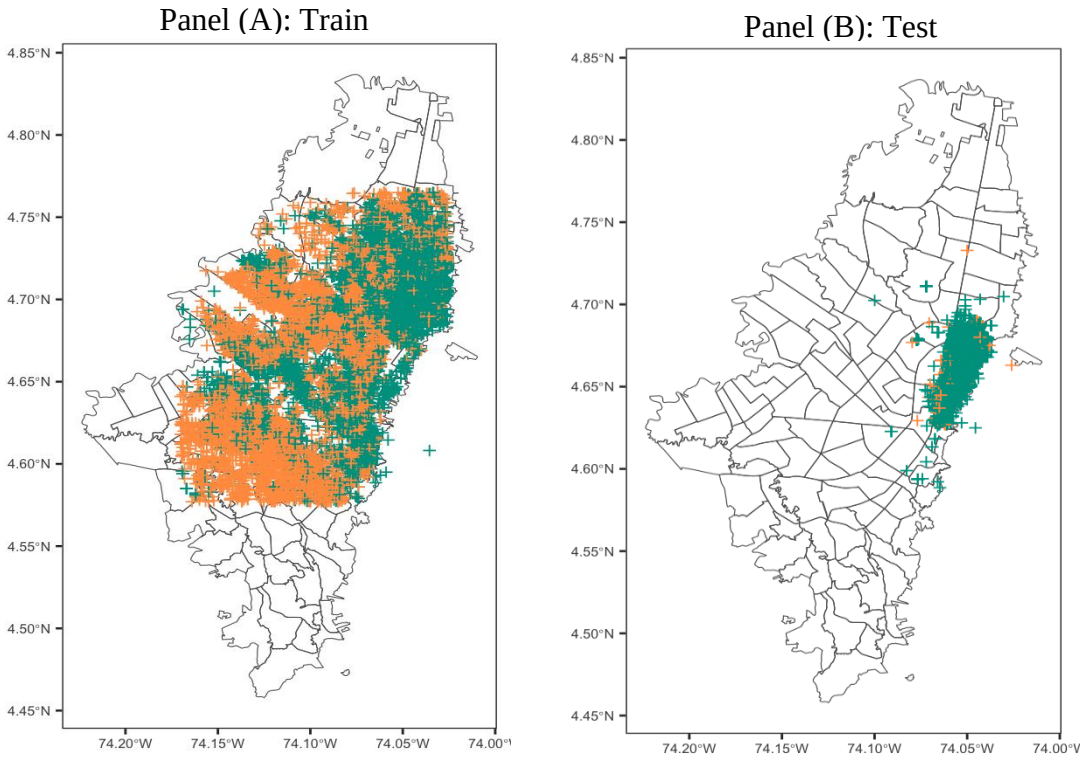
Figura 1. Área total y cantidad de habitaciones después de imputación.



Al momento de realizar este trabajo con la descripción se puede observar el potencial para obtener más variables extraídas desde esta, por ello, se buscó obtener variables indicadoras para conocer si la vivienda cuenta con parqueadero; si está ubicado en un conjunto o residencia; y si es un lugar nuevo, recién construido o para estrenar. En estos casos la variable toma el valor de 1 cuando el inmueble cuenta con la característica buscada o 0 de lo contrario.

Como bien lo explica la literatura, la ubicación de un inmueble y su cercanía o lejanía a ciertos puntos determinan fuertemente el precio de venta. Dentro de lo que cabe, la vivienda responde a dinámicas de mercado complejas que están asociadas tanto a las características innatas del inmueble, como también a la ubicación espacial en donde se encuentra con respecto al resto de la ciudad. Básicamente, si un inmueble se encuentra en un barrio de estrato alto, su precio de venta va a ser más elevado. De igual forma, un inmueble que tiene acceso a diferentes áreas genera un valor agregado con respecto a las viviendas que no tienen acceso o cercanía a un hospital, un centro comercial, estación de Transmilenio, universidad, entre todos. No obstante, la cercanía a vías principales, el acceso a servicios públicos, la cercanía a otros municipios, entre otros, son elementos que determinan ese valor de venta. En últimas, la ubicación espacial establece el valor agregado que posee un inmueble dentro del mercado inmobiliario. Analizando la muestra de manera espacial nos encontramos que en las bases de entrenamiento –train– y base de prueba –test– se distribuye las casas y apartamentos como se observa en la figura 2.

Figura 2. División del tipo de vivienda por base de datos.



Nota: Las formas verdes pertenecen a apartamentos y las formas naranjas pertenecen a casas.

Como puede observarse, la muestra –test– se limita a la localidad de Chapinero, mientras que la muestra –train– se distribuye por toda la ciudad. Así mismo, se evidencia que en el suroccidente de la ciudad se encuentra un mayor número de casas que apartamentos, mientras que en el nororiente de la ciudad la distribución es contraria encontrando más apartamentos que casas.

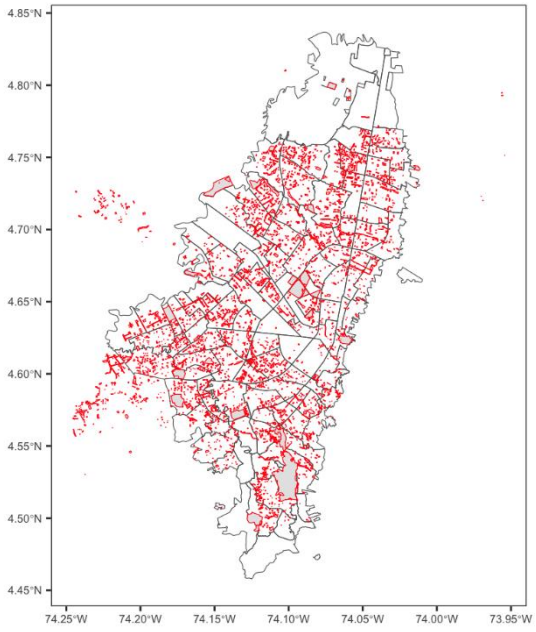
Debido a la relación entre el aspecto espacial y el valor de la vivienda, se decidió explorar y establecer multitud de variables espaciales, para así determinar las características de las viviendas que están asociadas a su ubicación dentro de la ciudad. Para esto, se determinó la cercanía que tenía cada inmueble con respecto a diferentes lugares y espacios que están distribuidos por la ciudad. Para lo cual, fue importante obtener la ubicación de cada sitio a través de OpenStreetmap y con base a los polígonos allí guardados, se calcularon las distancias entre cada inmueble y el lugar o espacio más cercano según corresponda. Con estos polígonos en mente se obtuvo el área del parque más cercano a cada inmueble[3] como lo muestra la figura 3. Este proceso se repitió para obtener los centros comerciales, teatros, estaciones de policía, concesionarios, bancos, estaciones de gasolina, talleres

[3] Se realizó bajo el supuesto de que estar cerca de un parque muy grande va a aumentar el precio más de lo que sería un parque pequeño, por lo que, se controló este efecto.

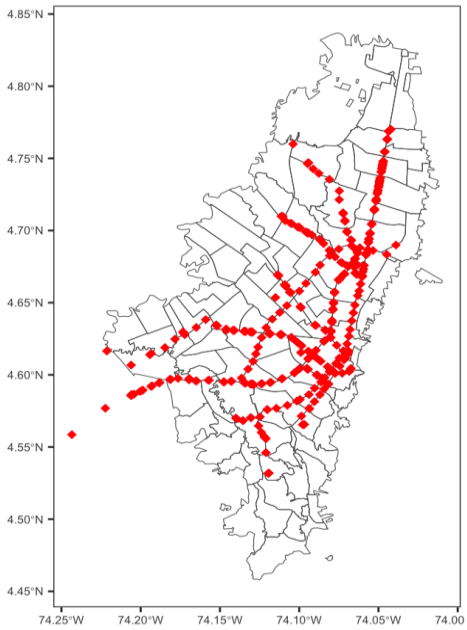
automotrices y universidades más cercanas, al igual que acceso a medios de transporte como elementos para determinar el valor de un predio.

Figura 3. Variables espaciales

Panel (A): Ubicación de parques

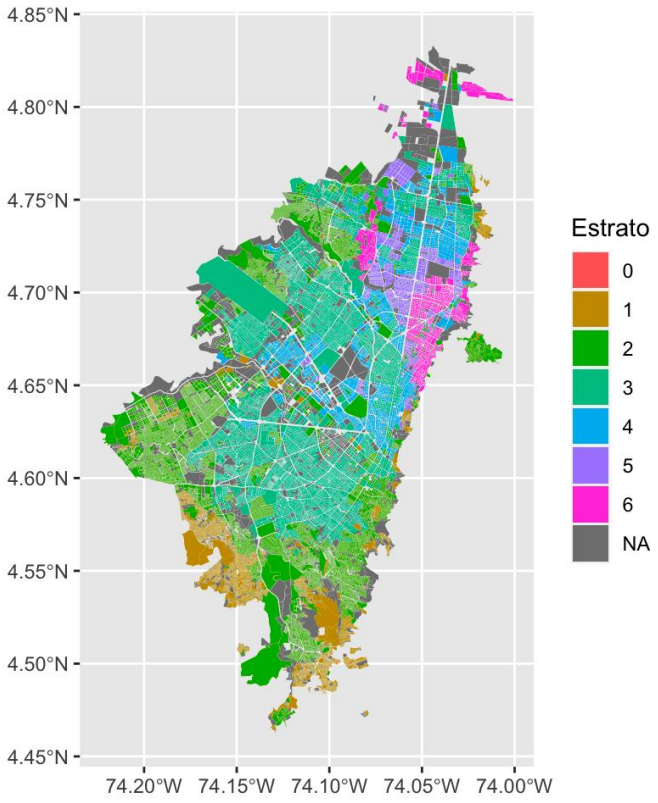


Panel (B): Estaciones de Transmilenio



En Colombia estas las zonas de precios se toman en cuenta de manera institucional para el cobro de ciertos servicios e impuestos, estas zonificaciones son llamadas estratos. Por lo tanto, incluir esta variable es fundamental para entender la determinación de precios a nivel barrial en Bogotá. Para obtener estos datos se usó el censo nacional del 2018 y la división por manzanas del Marco Geoestadístico Nacional de Colombia para Bogotá. En este orden de ideas, se extrajeron los datos de todas las respuestas del censo nacional para Bogotá, donde se tomó el estrato que reporta cada persona y ese estrato se asoció a la ubicación en donde vive la persona que respondió la encuesta. De esta manera se cruzan ambos datos y cada polígono de Bogotá tiene un estrato asociado como lo muestra la figura 4. Con estos polígonos en mente se salta a clasificar cada uno de nuestros inmuebles dentro de un estrato determinado. Imputando el valor del estrato más cercano de ese polígono.

Figura 4. Estratos socioeconómicos en Bogotá



3. Modelo y resultados

En base a las variables recolectadas, se procedió a la construcción de diversos modelos con el propósito de evaluar su desempeño en relación con cada especificación. Como resultado de este proceso, de desarrolló el modelo presentado en la Ecuación 1. Posteriormente, se procedió a entrenar el modelo utilizando la técnica de Boosting Trees. Este enfoque es un método de aprendizaje automático que se basa en la construcción secuencial de árboles de decisión, donde cada nuevo árbol se enfoca en corregir los errores del modelo anterior. Este proceso iterativo permite que el modelo se adapte de manera más preciosa y eficaz a los datos, mejorando su capacidad de generalización. El ajuste de los hiperparámetros específicos del Boosting Trees, como la tasa de aprendizaje y el número de árboles base, se seleccionaron a medida que iba mejorando la precisión de las predicciones.

$$Precio\ de\ vivientas = cuartos + baños + habitaciones + parquerado + tipo\ de\ vivienda + estrato\ socioeconomico + area\ comercial + cuarto * estrato + distancia\ a\ paredero\ de\ bus + distancia\ a\ paredero\ de\ bus^2 + nuevo + area\ parques * distancia\ parques + superficie\ total + distancia\ a\ universidades + distancia\ a\ gasolineras + ditancia\ comercial \quad [1]$$

Una vez seleccionada las variables a utilizar los hiperparámetros fueron ajustados de manera exhaustiva en un proceso de búsqueda aleatoria, generando una cuadricula de combinaciones posibles para determinar la configuración óptima del modelo, ajustando así solo el número de árboles en el rango de 400 a 800, el mínimo de observaciones por nodo en el rango de 1 a 8 y la tasa de aprendizaje en el rango de 0.001 a 0.1. Los valores y el modelo óptimo se seleccionaron después de realizar una serie de 53 predicciones con diferentes variables, métodos y parámetros, los cuales se podrá detallar los 5 mejores resultados en la tabla 2.

Tabla 2. Calificación de los cinco mejores modelos presentados

Modelo	Calificación en Kaggle	Diferencia
Boosting Trees	251.811.698	-
Boosting Trees	252.031.791	0,1%
Random Forest	258.430.583	2,6%
Árbol de regresión	264.403.715	5,0%
Elastic net	268.809.217	6,8%

Podemos observar una diferencia amplia entre utilizar solo árboles e implementarlos Boosting Trees o Random Forest, estos dos resultan ser los mejores modelos de predicción, ambos métodos de aprendizaje mejoraron significativamente el rendimiento de los árboles de decisión. Aunque en los modelos de aprendizaje son lo de mejores resultados, presentan diferencias en su enfoque y funcionamiento. En el caso de Random Forest, se construyen múltiples árboles de decisión en paralelo, utilizando subconjuntos aleatorios de los datos de entrenamiento con reemplazo y realizando la selección de características aleatoria en cada nodo de decisión. Por otro lado, Boosting Trees construye árboles secuencialmente, centrándose en los errores de los árboles previos y asignando pesos para adaptar las observaciones según el desempeño de cada árbol, sin embargo, tiende a ser propenso al sobreajuste. Sumado a esto al evaluar modelos como Ridge, Lasso y Elastic Net podemos ver una diferencia más marcada con respecto al mejor modelo, a pesar de utilizar las mismas variables para aproximar el precio de las viviendas en Chapinero la diferencia es en promedio de 5.5% para esto modelos y el método más alejado es al utilizar regresión lineal con una diferencia del 8% en el mejor modelo, no obstante, para llegar a este se incluyeron más variables que los demás modelos.

4. Conclusiones y recomendaciones

En este estudio, se llevaron a cabo un total de 53 modelos de predicción de precios de viviendas, incluyendo modelos de regresión lineal, Ridge, Lasso, Elastic Net, CART, Random Forest y Boosting. Un hallazgo significativo fue la influencia positiva de la inclusión de variables adicionales, más allá de las características propias de las viviendas, en las predicciones de precios. Factores como

la proximidad a estaciones de Transmilenio, área de parques, comercio o universidades demostraron tener un impacto notorio en la precisión de las estimaciones. Asimismo, se constató la relevancia del estrato socioeconómico como una variable crucial para mejorar la calidad de las predicciones.

Entre los modelos evaluados, se destacaron Random Forest y Boosting como los más sobresalientes, obteniendo las calificaciones más favorables en la competencia de Kaggle. No obstante, es importante resaltar que Boosting, por su naturaleza secuencial, mostró cierta susceptibilidad al sobreajuste, lo que enfatiza la necesidad de equilibrar apropiadamente el sesgo y la varianza en el diseño del modelo.

Al comparar estos enfoques con los métodos de regularización, como Ridge, Lasso y Elastic Net, se evidenció una diferencia marcada en el rendimiento, con una perdida en promedio del 5.5% en relación con el mejor resultado, y la regresión lineal presentando las predicciones menos acertadas. Estos resultados destacan la importancia de la elección del método de predicción, el manejo del conjunto de datos y el tratamiento de las variables para lograr una buena predicción de precios de viviendas en la localidad de Chapinero.

Referencias

Grajales Alzate, Y. V. (2019). Modelo de predicción de precios de viviendas en el municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz. Tesis de Maestría, Escuela de Ingenierías.

Maté de Dios, A. (2022). Estudio sobre las variables que impactan los precios de vivienda nueva en Bogotá en sus diferentes segmentos económicos.

Github:

https://github.com/Erick-Villabon/Problem_Set_2