Presentado por Yilmer Palacios, Betina Cortés, Lida Jimena Rivera, Nelson Fabián López

Problem Set 3: Making Money with ML?

I. Introducción

Este estudio se basa en un modelo predictivo que contribuye al mercado inmobiliario, particularmente a la compraventa de propiedades, cuyo objetivo es comprar la mayor cantidad de propiedades en el barrio de Chapinero en Bogotá, Colombia, invirtiendo lo menos posible. Los datos para este análisis se toman de una muestra de datos de propiedades individuales en Bogotá, los cuales se extraen de https://www.properati.com.co. Con este modelo de predicción se pretende generar evidencia para tomar las mejores decisiones de compra de inmuebles basándose en las variables que mayor influencia tienen en el precio. En este caso, se consideran variables del entorno de la vivienda, como la tasa de homicidio, tasa de hurto a residencias, número de colegios, hospitales y parques en la UPZ y el estrato, que depende de atributos como vías de acceso, puntos de transporte e infraestructura social. Por otro lado, se incluyen variables inherentes a la vivienda, tales como el área en metros cuadrados, si tiene parqueadero, terraza, depósito o patio.

En el presente documento se consideran cuatro (4) modelos predictivos y se profundiza en el que presenta mejor desempeño. Se evalúan una regresión lineal simple, regresiones lineales regularizadas de Lasso y Ridge y, finalmente, un modelo Random Forest. Como resultado del ejercicio, se obtiene que el modelo Random Forest presenta el mejor desempeño en la predicción de precios de vivienda en la localidad de Chapinero, con un RMSE (Root Mean Squared Error) de COP $ 239.012.001,98.

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/Nelson1802/Repositorio_taller3.git

Contexto

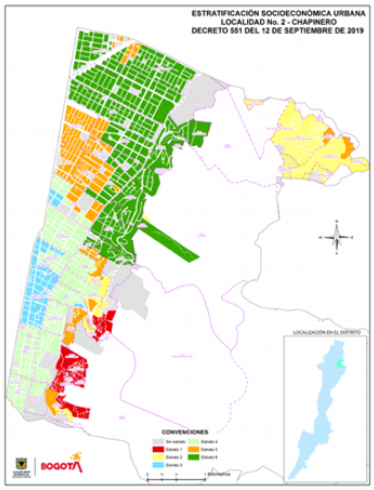
En Colombia, en el primer semestre de 2022 se vendieron en promedio 127.218 viviendas nuevas, representando un crecimiento de 2,5% en comparación con el mismo periodo del año inmediatamente anterior (Camacol, 2022). Adicionalmente, 7 de cada 10 viviendas vendidas fueron de interés social y las regiones que generaron más ventas fueron: Bogotá con 27 mil unidades, Valle con 18 mil, y Atlántico con 15 mil. Lo anterior representa 27,1 billones de inversión en vivienda en lo corrido del año (Camacol, 2022). Por otro lado, informes de la Cámara Colombiana de Construcción demuestran que el sector inmobiliario terminó el 2022 con un balance positivo, donde las ventas fueron 66.368 viviendas No VIS y 168.224 unidades de VIS (Semana, 2022). El promedio de las ventas ascendió en un 28% con relación a los últimos años y se estima la construcción de alrededor de 360.000 viviendas en construcción para el futuro.

Ahora bien, para entender el mercado inmobiliario en Bogotá es importante conocer algunos aspectos que pueden definir las preferencias de los consumidores a la hora de adquirir inmuebles. Para el caso objeto de análisis, cabe mencionar que Chapinero surge a finales del siglo XIX producto de la migración de las élites capitalinas hacia el norte de lo que en la época era la zona urbana. Es una localidad que conforma el centro extendido de la capital y es el centro financiero, cultural y gastronómico de Bogotá. Tiene cercanía con importantes lugares tanto para el trabajo como para el estudio y el esparcimiento. Se caracteriza por el contraste entre edificios modernos y la arquitectura europea de mediados del siglo XX (Properati, 2021).

Chapinero, en la época colonial fue un lugar de tránsito entre Santa Fe y los municipios aledaños del norte. Este sector se configuró como una alternativa lejana de la zona urbana de Bogotá y de todos los problemas de higiene y hacinamiento que se presentaron en el cambio de siglo. Tiempo después, y tras el Bogotazo en 1948, este sector fue epicentro de las clases privilegiadas que se mudaron del centro de la ciudad (Properati, 2021). Esta localidad está ubicada en el centro norte de la ciudad: entre las Avenida Caracas – Autopista norte y los Cerros Orientales; y entre la Avenida 39 y la Calle 100 al norte y contempla los siguientes barrios por estrato:

Estrato 1: El Paraíso, Siberia Urbano, Siberia II, Siberia Central, Siberia Rural, Pardo Rubio I e Ingemar Oriental. **Estrato 2:** El Paraíso, Ingemar I y Oriental, Juan XXIII, Maria Cristina, Páramo Urbano, San Luis Altos del Cabo Rural I, San Luis Altos del Cabo Rural II, La Esperanza, San Isidro Rural, San Isidro Rural II y Páramo Rural V. **Estrato 3:** Chapinero Central, Chapinero Norte y Porciúncula. **Estrato 4:** Sucre, Cataluña, Marly, Pardo Rubio, Bosque Calderón, La Salle, María Cristina, Granada, Chapinero Norte, Quinta Camacho y Porciúncula. **Estrato 5:** El Paraíso, Ingemar, María Cristina, Granada, Emaús, Porciúncula, El Nogal, Espartillal, Lago Gaitán, El Retiro, Antiguo Country y La Cabrera. **Estrato 6:** Las Acacias, Emaús, Bellavista, El Bagazal, Los Rosales, El Nogal, El Retiro, La Cabrera, El Refugio, El Refugio I, El Refugio II, Páramo Rural, El Chicó, Chicó Norte, Chicó Norte II Sector y Chicó Norte III Sector.

Grafica No. 1. Estratificación socioeconómica urbana Localidad Chapinero



Fuente: Secretaría Distrital de Planeación de Bogotá

II. Datos

a. Descripción de las fuentes de datos

Para el desarrollo de este Problem Set se utilizarán los siguientes datos: por un lado, la tasa de homicidios y tasa de hurto a residencias por UPZ (Unidad de Planeación Zonal), entendida como una división administrativa de la ciudad que es más pequeña que las localidades, pero más grande que los barrios. Tanto la tasa de homicidios como la de hurto a residencias son tomadas de la Secretaría Distrital de Seguridad, Convivencia y Justicia y, las estimaciones de población de las UPZ de 2018 a 2021 se extrajeron de la Secretaría Distrital de Planeación. Se procede a dividir la ocurrencia del crimen entre la población por 100 mil habitantes y se obtiene la tasa por 100 mil habitantes.

Por otro lado, para identificar la cercanía de parques, hospitales (IPS's) y colegios, entre otros, se utilizaron datos provenientes del IDECA (Infraestructura de Datos Espaciales para el Distrito Capital) de la Unidad Administrativa Especial de Catastro Distrital. Partiendo de que los datos son georreferenciados se cruzan con los polígonos de UPZ de Bogotá, así para cada UPZ se halló el número de parques, hospitales (IPS's) y colegios para cada una de estas. Para las fuentes que se derivan de los textos, se realiza una búsqueda de las descripciones de cada una de las viviendas, buscando por palabras clave como parqueadero, terraza, depósito y patio.

Adicionalmente, se utilizó el paquete sf (simple features) para trabajar con datos georreferenciados que permiten cargar la longitud y la latitud que aparece en el dataset de Kaggle, que a su vez, permite crear un objeto espacial que muestre en un mapa dónde están las viviendas, y luego, con la función step-join se identifican los polígonos donde están los puntos de las viviendas para asignar a cada casa el valor de tasa de homicidio, hurto a residencias, número de hospitales, número de parques y número de colegios que aparecen en cada UPZ.

b. Análisis descriptivo de los datos (estadísticas descriptivas)

El precio de una vivienda puede estar determinado por diversos factores tanto económicos como sociales. Para el caso de estudio, se analizan factores internos (propios de las viviendas) como externos (elementos geoespaciales como distancia a hospitales, parques y colegios, etc.). En este sentido, es necesario tener información de variables asociadas al precio, con el propósito de obtener un modelo robusto. Para el análisis de los modelos de predicción se utiliza una muestra tomada de Properati sobre los precios de venta y características de inmuebles ubicados en la localidad de Chapinero en Bogotá, lo que arroja una muestra representativa de entrenamiento y testeo del precio de venta de las viviendas en esta zona, así como de los principales atributos que definen el precio de mercado de esta.

Seguido de esto, se define tomar solamente información de esta zona porque es en ella en donde se realiza la predicción y, por lo tanto, es relevante entrenar los modelos con información de la misma zona para no alterar la predicción de los precios, porque el precio de las viviendas puede ser muy diferente dependiendo de la zona en donde esté ubicada. Los polígonos de análisis se obtienen de las UPZ de IDECA (Infraestructura de Datos Espaciales para el Distrito Capital) y se selecciona la información de la base de datos que corresponde a estos polígonos.

Por medio de la descripción de las viviendas en venta, se recopiló información sobre atributos adicionales y que influyen en el precio de los inmuebles, se obtuvo variables adicionales de características físicas de si cuenta o no con terraza, patio, parqueadero y depósito, de igual forma, para las variables existentes que presentaban missing values, se procedió a imputar información rescatada del texto sobre número de baños, número de habitaciones y área total. Adicionalmente, se consideraron variables de distancia mínima entre los inmuebles y las zonas comerciales, y de esparcimiento (parques, hospitales y colegios por UPZ), esta información se obtuvo también de IDECA (Infraestructura de Datos Espaciales para el Distrito Capital) de la Unidad Administrativa Especial de Catastro Distrital.

Con lo anterior, se buscó obtener la información total de las variables mencionadas anteriormente, al considerarlas importantes para este estudio. Según Rosen (1974), las características de los bienes describen a los bienes diferenciados, lo que quiere decir que estas características pueden explicar qué tan diferente es el bien y cómo éstas pueden ser un factor determinante para impactar el precio de las viviendas. Así las cosas, el modelo de predicción toma en cuenta datos de las viviendas ubicadas en la localidad objetivo (Chapinero) y

el total de datos (base de entrenamiento = 38.644 y base de testeo = 10.286) para un total de 49.930 viviendas. A continuación, se describen las variables objeto de estudio:

Ubicación: Es esencial para el análisis ya que se concentra en la localidad de Chapinero. Los precios de las localidades difieren, si bien están dados por características similares, el costo de vida por localidad también influye en el valor de la vivienda. Esta variable es categórica.

Tipo de propiedad: Esta variable describe si la propiedad es un apartamento o si es una casa, pues esta connotación influye sustancialmente en el precio, ya que el área de una casa, por lo general, suele ser más grande y contar con amplios espacios de esparcimiento. Esta variable categórica.

Habitaciones: El número de habitaciones de la vivienda es determinante en el precio de esta, ya que se puede tener un aproximado del espacio y de cuantos individuos pueden vivir, es decir, entre más habitaciones, el precio del inmueble tiende a incrementarse (directamente proporcional). Esta es una variable numérica.

Parqueadero: Si el inmueble incluye al menos un garaje, el precio de la vivienda tenderá a aumentar su valor. Por otra parte, de acuerdo con el análisis descriptivo, se identifica que esta es una variable booleana, en donde V hace referencia a que la casa o el apartamento cuenta con al menos un parqueadero y F que no lo tiene.

Terraza: Es un lugar que permite realizar una serie de actividades al aire libre y se puede apreciar de una buena vista de la ciudad. Una terraza puede convertirse en el corazón de la casa y aumentarle el valor. Es una variable booleana, en donde V hace referencia a que tiene terraza y F que no la tiene.

Patio: Aquella parte de una construcción que se destina a la recreación al aire libre. Permite también hacer uso de un espacio abierto en cuanto a su diseño, pero privado en cuanto al acceso. Es una variable booleana, en donde V hace referencia a que tiene terraza y F que no la tiene.

Depósito: Este espacio garantiza una opción para almacenar diferentes tipos de objetos como elementos de aseo, herramientas y enseres. Es una variable booleana, en donde V hace referencia a que tiene terraza y F que no la tiene.

Por medio de la ubicación geoespacial, las variables adicionales que se encontraron y que consideramos relevantes para la predicción, fueron las siguientes:

Distancia a parques, colegios y hospitales (IPS's): La distancia mínima a por lo menos un parque, un colegio o un hospital, lo cual es predictor relevante dentro del modelo, ya que, en general, los individuos buscan tener fácil acceso a estos escenarios, y de esta manera, tienen mayor disposición a pagar por viviendas que se encuentren más cercanas. Esta es una variable numérica.

Las estadísticas descriptivas muestran que, en promedio, en Chapinero, los inmuebles cuestan alrededor de $654.534.675 millones de pesos, cuentan en promedio con 3 habitaciones. Por otra parte, cuentan con una tasa de 6 homicidios por cada 100 mil habitantes, tasa de hurto a residencias de 159 por cada 100 mil habitantes, cercanía en promedio a 16 colegios, 63 parques y 94 IPS's por UPZ dependiendo a la que pertenece cada vivienda (Ver gráfica No. 1).

Tabla No. 1. Estadísticas descriptivas variables numéricas

Estadísticas de resumen

Dataset de características de viviendas en Bogotá

	Entrenamiento				Prueba			
Variable	Mean	Min	Max	SD	Mean	Min	Max	SD
Precio	654,534,675.29	300,000,000.00	1,650,000,000.00	311,417,886.95	NA	NA	NA	NA
Cuartos	3.14	0.00	11.00	1.53	2.38	0.00	11.00	0.96
Latidud	4.69	4.58	4.77	0.04	4.67	4.59	4.73	0.01
Longitud	−74.06	−74.17	−74.03	0.03	−74.05	−74.10	−74.03	0.01
Tasa de Homicidio	6.68	0.00	236.97	15.09	5.13	0.00	122.43	4.68
Tasa de hurto a residencias	159.03	0.00	583.28	62.53	212.39	70.54	435.32	58.39
Número de colegios	16.39	0.00	100.00	12.69	6.81	0.00	41.00	3.69
Número de parques	63.18	0.00	156.00	30.60	40.90	0.00	156.00	13.56
Número de hospitales	94.23	0.00	364.00	70.85	173.50	0.00	364.00	165.16

Nota: No se incluye información sobre las variables de habitaciones, area total, cubierta y baños porque no se usaron en el entrenamiento.

Fuente: R Studio

Partiendo de que las variables terraza, patio, parqueadero y depósito son booleanas (verdadero/falso), la proporción en promedio de todas las observaciones que cumplen con esta condición se muestra en la Tabla No. 2. La predicción del modelo indica que en promedio el 30% de las casas/apartamentos del conjunto de entrenamiento tienen terraza, el 11% tienen patio, el 43% tienen parqueadero y el 34% tienen depósito.

Tabla No. 2. Estadísticas descriptivas variables lógicas

Estadísticas de resumen

Variables lógicas

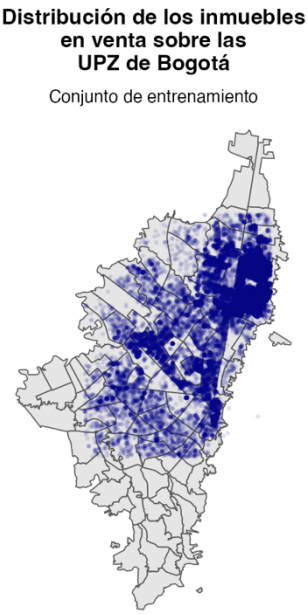
	Proporción	
Variable	Entrenamiento	Prueba
Tiene terraza	0.30	0.32
Tiene patio	0.11	0.03
Tiene parqueadero	0.43	0.45
Tiene depósito	0.34	0.34

Nota: No se incluye información sobre las variables de habitaciones, area total, cubierta y baños porque no se usaron en el entrenamiento.

Fuente: R Studio

Los mapas de la localidad de Chapinero en Bogotá presentados en las Gráficas 3 y 4, evidencian la distribución de inmúteles en venta sobre las UPZ de Bogotá, tanto para el conjunto de tratamiento como para el de prueba. Esto permite contar con un análisis gráfico de cada vivienda y así un comprador, podría contar con información gráfica que le permita tomar decisiones de manera informada.

Grafica No. 3. Conjunto de entrenamiento



Fuente: R Studio

Grafica No. 4. Conjunto de prueba

Fuente: R Studio

III. Modelos y resultados

Para el ejercicio de predicción de precios de las viviendas en Chapinero se realizaron cuatro modelos. Los enfoques de pronóstico usados fueron regresión lineal, pasando a su uso con regularización de Lasso y Ridge, y finalmente se utilizó Random Forest.

Las variables seleccionadas para la predicción se muestran en la ecuación siguiente. Sin embargo, estas se describieron de manera detallada en la sección anterior.

$$precio = f\begin{pmatrix} año, mes, habitaciones, tipodepropiedad, latitud, longitud, \\ tasadehomicidios, tasadehurtosresidencias, \\ númerodeescuelas, númerodeparques, númerodeIPS, terraza, parqueadero, patio, deposito \end{pmatrix}$$

Antes de describir las técnicas utilizadas para el pronóstico, es importante destacar que en este proyecto se optó por utilizar la totalidad de los datos para entrenar los modelos predictivos y no dividir la base de datos "train" en submuestras de entrenamiento y prueba. Esto se debe a que reducir la cantidad de datos de entrenamiento disponibles podría afectar la precisión del modelo. Sin embargo, en caso de requerirse una evaluación aislada de los modelos, se podría hacer una "data partition" y proceder a encontrar el mejor modelo en función de las distintas métricas de evaluación y comparación de la partición de testing (como la métrica RMSE).

En cuanto a las técnicas de modelado utilizadas, primero se empleó un modelo de regresión lineal sin herramientas adicionales. Luego, se usó la técnica de regularización con Lasso y Ridge, utilizando el paquete "gmlnet" que optimiza la regularización a través de la técnica "coordinate descent". Esta técnica ajusta un coeficiente predictor a la vez manteniendo los demás fijos en cada iteración, hasta que se minimiza la función de pérdida. En ambos casos, se consideró un penalty de 0.001.

Por otro lado, para el enfoque de Random Forest, se utilizó el paquete "ranger" para la optimización. "Ranger" es similar al algoritmo original de Random Forest, pero realiza optimización de los cálculos en memoria e implementación paralela (utilizando varios núcleos del procesador), lo que le permite ser más rápido. Adicionalmente, emplea la técnica de "importancia de variables basada en permutación" para seleccionar las variables aleatorias predictoras, evaluando cómo cambia el desempeño del modelo si se aleatoriza el valor de una variable predictora mientras las demás permanecen constantes. Por último, "ranger" realiza la selección de hiperparámetros basado en cross validation y random search, evaluando distintas combinaciones de los valores de los hiperparámetros para seleccionar la mejor combinación.

Una vez corrido el modelo bajo los distintos enfoque de pronóstico, la métrica usada para la selección del mejor modelo fue el RMSE *(Root Mean Squared Error),* una medida de la diferencia entre los valores reales y los valores predichos. El RMSE representa la raíz cuadrada de la media de los errores al cuadrado entre los valores predichos y los valores reales, de esta forma, cuanto menor sea, mejor será el rendimiento del modelo en la predicción. La siguiente tabla muestra los resultados obtenidos:

Tabla No. 3 Resultados del RMSE por modelo

Modelo predictivo	RMSE
Regresión lineal	$ 314.449.267,97
Lasso	$ 314.755.575,65
Ridge	$ 314.449.267,97
Random Forest	$ 239.012.001,98

Fuente: Elaboración propia

Así, se evidencia que el modelo con mejor capacidad predictiva del precio de las viviendas en Chapinero es el Random Forest y, por lo tanto, fue el seleccionado como resultado del ejercicio.

IV. Conclusión

En conclusión, el estudio ofrece recomendaciones útiles para los inversores inmobiliarios que desean adquirir propiedades en Chapinero y sugiere que el modelo Random Forest es el más preciso para predecir los precios de la vivienda en esta localidad. Particularmente, consideramos que el modelo predictivo desarrollado puede traer los siguientes beneficios a los actores del mercado inmobiliario:

1. Ayudar en la toma de decisiones de inversión: El modelo de predicción de precios de viviendas en Chapinero puede ayudar a la startup inmobiliaria a determinar qué propiedades tienen un mayor potencial de ganancias y, por lo tanto, en cuáles debería invertir.
2. Establecer precios competitivos: Al conocer los precios previstos para el mercado, la startup puede establecer precios que se ajusten al mercado y sean atractivos para los compradores.
3. Identificar oportunidades de mercado: Por ejemplo, si el modelo predice que los precios de las propiedades en una determinada área están a punto de aumentar, la startup puede considerar invertir en esa área.
4. Optimizar la gestión de inventario: Al conocer los precios previstos para el mercado, la startup puede ajustar su inventario para satisfacer la demanda de los compradores.

V. Bibliografía

Camacol., (2022). Indicadores ventas de vivienda primer semestre 2022. Recuperado: https://camacol.co/actualidad/noticias/indicadores-ventas-de-vivienda-primer-semestre-2022#:~:text=El%20primer%20semestre%20del%202022,lo%20revelan%20cifras%20de%20%23CoordenadaUrbana.

Ciencuendras., (s.f.). Guía de barrio Chapinero en Bogotá. Recuperado de: https://www.ciencuadras.com/blog/guia-de-barrio-chapinero-bogota

DANE. (2022). Obtenido de https://www.dane.gov.co/index.php?option=com_content&task=category&sectionid=101&id=604 &Itemid=1183

Kuhn, M. (2012). The caret package. R Foundation for Statistical Computing, Vienna, Austria.

Properti., (2021). ¿Por qué vivir en Chapinero? Recuperado de: https://blog.properati.com.co/por-que-vivir-en-chapinero-recomendaciones/

Rosen, S. (1974), 'Hedonic prices and implicit markets: Product diferentiation in pure competition', Journal of Political Economy.

Semana., (2022). Sector inmobiliario cierra el 2022 con balance positivo; ventas aumentaron en un 28%. Recuperado: https://www.semana.com/economia/macroeconomia/articulo/sector-inmobiliario-cierra-el-2022-con-balance-positivo-ventas-aumentaron-en-un-28/202210/