Problem Set 2: Making Money with ML?

“It’s all about location location location!!!”

**Integrantes:** Juan Diego Duarte[[1]](#footnote-1); Erick Julian Villabon[[2]](#footnote-2)

## **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.

La precisión en la modelación predictiva de los precios de vivienda es de suma importancia en un contexto donde se puede sobrevalorar los precios de las propiedad y resultar en perdidas importantes, por lo que, en este documento se explora diferentes estrategias y metodologías para superar estos obstáculos y evitar estimaciones lejanas como el caso de Zillow, un mercado en línea de bienes raíces el cual presento problemas que surgía de los algoritmos de compre de propiedad de la empresa, dichos algoritmos resultaron en una sobrevaloración significativa de los inmuebles adquiridos, con un impacto económico que incluyó pérdidas de USD 304 millones. Así se está en la búsqueda de obtener la mejor precisión posible de los precios de las viviendas en Chapinero.

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 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.

## **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%.

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.

## **Modelo y resultados**

**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.

1. Código: 202011999 [↑](#footnote-ref-1)
2. Código: 201815677 [↑](#footnote-ref-2)