

# Problem Set 3 – GRUPO 6

# Víctor Dulio Chique

# Víctor Iván Sánchez

# Natalia Castro

# HACIENDO DINERO CON ML: PRECIOS HEDONICOS

1. Introducción:

Las estimaciones de precios de viviendas están basadas en el modelo de precios hedónicos, que consiste en establecer la relación entre el precio de un bien mercadeable, en este caso viviendas urbanas, y las características del bien. Las viviendas son productos diferenciados desde la perspectiva del consumidor, pues los atributos o características de las viviendas les proporcionan utilidad. Por su parte, los productores de estos bienes incurren en costos que dependen de los atributos asignados a estas viviendas. Entonces, la interacción en este mercado entre consumidores y productores determinan la senda de equilibrio del precio de la vivienda.

Las viviendas pueden ser descritas por un conjunto de características estructurales como tamaño, número de habitaciones, numero de baños, número de garajes, tipo de vivienda (casa o apartamento), antigüedad, entre otro, y adicionalmente por un conjunto de características del entorno, que pueden acceso a bienes y servicios públicos como parques, centros comerciales, universidades, transporte público, entre otros.

2. Datos

2.1 Descripción de los Datos

Variables espaciales

Tomando en cuenta que la localización de las propiedades es una variable importante de incorporar en los modelos de predicción de precios, se crearon 4 predictores que reflejan las cercanía de las propiedades a las áreas comerciales, las universidades, los parques y las avenidas principales. Se utilizó Open Street Map y se estimaron las distancias de los apartamentos a cada uno de esos lugares:

*Distancia\_parques*: se identificaron los polígonos de los parques en Bogotá y sus centroides. Se calcularon las distancias desde los apartamentos hasta estos puntos centrales.

*Distancia\_comercial*: de la librería “*building*” se escogió “*commercial*” y se calculó la distancia de los apartamentos a los áreas comerciales: centros comerciales y supermercados.

*Distancia\_ universidad*: distancia de cada apartamento a la universidad más cercana.

*Distancia\_ avenida\_principal*: Corresponden a avenidas como la Circunvalar, Carrera 11, Calle 19, Calle 72 o la Calle 92 que se encontraron utilizando la librería “highway” y la opción “secondary”. Se calculó la menor distancia de cada apartamento a esas avenidas.

Bogotá es una ciudad bastante densa y las propiedades más costosas se encuentran en la parte Nor-Este de la ciudad. A continuación se muestran los mapas con las principales avenidas en color rojo y los polígonos que representan los parques en color verde.

Map

Description automatically generated

Map

Description automatically generated

2.2 Estadísticas Descriptivas

Correlaciones

3. Modelo y Resultados

El precio de la vivienda (P) es una función de las características estructurales y los atributos del entorno. Así, estas se representan por un vector de características y atributos , donde denota una de las característica o atributos de la vivienda. Y cada vivienda es denotada por , donde superíndice indica una propiedad distinta con distinto vector de características y atributos . Por lo tanto, el precio de venta de una vivienda es función de las características y atributos de la misma.

En nuestro modelo, el precio dependerá de un conjunto de características estructurales como número de habitaciones, numero de baños y tipo de vivienda (casa o apartamento), y adicionalmente de un conjunto de características del entorno como la distancia al parque, distancia al centro comercial, distancia a la universidad y cercanía a avenida principal.

Random forest (8 predictores):

Este tipo de modelos busca realizar las predicciones a través de varios árboles de decisión que se construyen a partir de un subconjunto aleatorio de un número determinado de variables predictoras. En este ejercicio de utilizaron 8 de estas variables: bedrooms, bathrooms, property\_type, superficie, distancia\_parque, distancia\_comercial, distancia\_avenida\_principal, distancia\_universidad.

Adicionalmente, el modelo de *Random Forest* se entrena sobre un subconjunto de los datos de la variable de interés en este caso: el precio de las casas y apartamentos de Bogotá. El subconjunto de datos de entrenamiento se obtuvo a través de *cross validation* con k=10.

Finalmente, se construyó una grilla con 2, 3, 4 ,5 y 8(bagging) predictores aleatorios. La regla de corte fué “varianza” y los diferentes tamaños de nodos: 1,2,3 y 6.

Como se puede ver en la gráfica XX a medida que se incrementa el número de predictores disminuye el RMSE, sin embargo después de 3 predictores aleatorios comienza a incrementarse.

El mejor número de nodos es 1. El cuadro XX muestra el MEA y MAPE dentro y fuera de muestra.

Chart, line chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | MAPE | MAE |
| Entrenamiento | 0.051 | 29403419 |
| Testeo | 0.121 | 68925784 |

Aunque el MAPE dentro de muestra es bajo no logra mantener el mismo porcentaje en el conjunto de evaluación. Esto se puede deber a *overfitting*. Sin embargo, aunque la variación entre los dos subconjuntos es evidente, una predicción en testeo con un MAPE del 20% es razonable.

Arboles de decisión (8 predictores):

Los árboles de decisión dividen los datos en subconjuntos utilizando particiones recursivas binarias. Se utilizaron 8 variables: bedrooms, bathrooms, property\_type, superficie, distancia\_parque, distancia\_comercial, distancia\_avenida\_principal, distancia\_universidad. Para el conjunto de entrenamiento y evaluación se utilizó *cross validation* con k=10. Se obtuvo el siguiente modelo:

Diagram

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | MAPE | MAE |
| Entrenamiento | 0.217 | 122791972 |
| Testeo | 0.218 | 122787377 |

El modelo escoge la variable “Superficie” como la más relevante. Para apartamentos con menos de 97m2 predice un precio de $441´000.000. Para las propiedades con un área entre 97m2 y 122 m2 $580´000.000 y las que tienen más de esa área un costo de $824´000.000.

Este es un modelo muy sencillo que puede resultar práctico para tener una visión general de los precios de acuerdo a una variable relativamente fácil de obtener pero que tiene un error de predicción mayor que el modelo de *Random Forest*. Este último sin embargo no es claro en especificar los predictores más importantes pues es escoge 3 de los 8 posibles de forma aleatoria.

3. Conclusiones y Recomendaciones

4. Bibliografía