

# 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

2.2 Estadísticas Descriptivas

Correlaciones

Shape

Description automatically generated with low confidence

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, surface\_covered, rooms, tipo, 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 4 predictores aleatorios no disminuye mucho más y para 8 predictores (el caso de *bagging*) inclusive se incrementa.

El mejor número de nodos son 2. El cuadro XX muestra el MEA y MAPE dentro y fuera de muestra.

Chart, line chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | MAPE | MAE |
| Entrenamiento | 0.087 | 50591918 |
| Testeo | 0.2 | 114573674 |

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. 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 | 0.345 | 195434630 |
| Testeo | 0.341 | 193418098 |

El modelo escoge la variable “*bedrooms*” como la más relevante. Para apartamentos con menos de 2.5 cuartos predice un precio de $499´000.000. Para las propiedades con más de 2.5 baños se debe tener en cuenta el área construida: menos de 156m2 tendrá un costo de $659´000.000 y las que tienen más de esa área un costo de $947´000.000.

Aunque este es un modelo muy sencillo puede resultar práctico para tener una visión general de los precios de acuerdo a dos variables relativamente fáciles de obtener. Adicionalmente, el MAPE dentro y fuera de muestra es bastante similar lo cual indica un modelo estable. Sin embargo, tiene un error de predicción mayor que el modelo de *Random Forest*. Aunque este último no es claro en especificar los predictores más importantes pues es escoge 4 de los 8 posibles de forma aleatoria.

Regresión lineal (8 predictores):

La estimación con el modelo de regresión lineal con los predictores: bedrooms, surface\_covered, rooms, tipo, distancia\_parque, distancia\_comercial, distancia\_avenida\_principal, distancia\_universidad, se obtuvo con *cross validation* k=10 y escalando las variables. Los siguientes fueron los parámetros estimados:

Table

Description automatically generated

En modelo de regresión lineal predice coeficientes muy grandes por lo que es difícil distinguir su importancia. Los más significativos son aquellos relacionados con las características de la casa: bedrooms, surface\_covered, rooms y tipo de casa. Entra las variables espaciales la distancia a las áreas comerciales es la más significativa.

3. Conclusiones y Recomendaciones

4. Bibliografía