

# Problem Set 3 – GRUPO 6

# Víctor Dulio Chique

# Víctor Iván Sánchez

# Natalia Castro

# HACIENDO DINERO CON ML: PRECIOS HEDONICOS

1. Introducción:

En el negocio de reventa de propiedades tener una predicción acertada de los precios y conocer las variables que influyen en los mismos permite realizar una mejor negociación de compra y evita pagar precios sobreestimados. De esta manera se maximizan los recursos de inversión así como el margen potencial que se puede obtener.

El objetivo de este trabajo es predecir los precios de la localidad de Chapinero en Bogotá donde se quiere adquirir la mayor cantidad de propiedades posibles para su futura reventa. La predicción de los precios se realizó utilizando diferentes modelos de aprendizaje automático que buscan obtener el menor error de predicción posible medido con el Mean Absolute Error (MAE) y el Mean Absolute Percentage Error (MAPE). Se exploraron modelos lineales, modelos de regularización, de árboles de decisión, clusters y super learners son el fin de entender mejor el espacio de predicción. Después de comparar la capacidad predictiva de los mismos se encontró que el modelo de *Super Learner* que incluyó *Random Forest* y el modelo lineal predecía mejor los precios (CONFIRMAR CON VICTOR MAE MAPE RESULTANTE).

La teoría de precios hedónicos plantea un problema en la economía del equilibrio espacial donde compradores y vendedores están siendo guiados por los precios implícitos existentes en el espacio de características del bien (Rosen (1974)) . Según la teoría, las estimaciones de precios de viviendas deben 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 sus atributos 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 propiedades 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 otros, 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.

En este trabajo, las variables utilizadas se obtuvieron de una muestra de 38.644 propiedades individuales de Bogotá recuperadas de https://www. properati.com.co las cuales contienen información de diferentes atributos propios de la vivienda. De igual manera, se crearon variables de distancia con el fin de capturar características del entorno que influyen en los precios. Finalmente, debido a que muchas de las observaciones no contaban con el dato de área total se utilizó Procesamiento de Lenguaje Natural en las descripciones para obtenerla.

2. Datos

2.1 Descripción de los Datos

Utilizamos los siguientes atributos de 38.644 propiedades de la ciudad de Bogotá obtenidos de https://www. properati.com.co: precio, numero de baños, superficie, habitaciones, cuartos, el tipo de propiedad (casa o apartamento), latitud, longitud y una descripción general del bien que proporcionan los vendedores de la propiedad. Las anteriores son características comúnmente evaluadas a la hora de comprar el bien, no obstante se precisa i) Completar de forma adecuada los datos faltantes en cada característica y ii) construir algunas variables espaciales de distancia a los parques, centros comerciales y universidades. Lo anterior nos permitirá tener un modelo lo suficientemente robusto en la labor de predicción. De igual manera contamos con una base de test con similares atributos pero sin la información del precio de las propiedades el cual es nuestro objetivo a predecir.

2.1.1 Limpieza de datos

El proceso para limpiar la base de datos y lograr obtener la muestra final de datos de tal manera que esta fuera completa y consistente, fue una labor retadora dada la cantidad de valores perdidos en variables tales como baños, habitaciones y en particular del área de las propiedades dado que solo un 22,1% de las propiedades tenían información de su área.

En general, teníamos el siguiente porcentaje de datos perdidos para cada variable: 79,6% de *superficie total*, 77.8% de datos perdidos de *superficie cubierta*, 26% de *cuartos* y 26% de *baños*. Respecto a los cuartos y baños optamos por usar el algoritmo de aprendizaje automático KNN “K-vecinos más cercanos” (en ingles “K- Nearest Neighboors”) que nos permite estimar los datos faltantes según la información de los vecinos más cercanos en el espacio de características o atributos.

En cuanto a los datos de superficie (tanto “total” como “cubierta”) consideramos que dado que la cantidad de valores perdidas era significativa, estimar los datos faltantes podría introducir sesgos al modelo, por lo tanto nuestro objetivo fue intentar recuperar la mayor cantidad de información de área desde la descripción, tarea que no fue sencilla dado que esta información es proporcionada por los vendedores y ninguna descripción es igual a la otra, no obstante aquellos vendedores que quieren indicar el área de su propiedad en la descripción lo hacen usando las mismas palabras ( p.e metros cuadrados, m2 o mts). Para lo anterior fue necesario hacer uso del “*Procesamiento del Lenguaje Natural”* el cual nos permite preprocesar texto de modo que este pueda ser usado como insumo para los modelos de aprendizaje de máquinas que vamos a utilizar en este trabajo. De igual forma, para poder identificar los patrones de área (metros cuadrados, mts o m2) en el texto de la descripción se hizo uso de *expresiones regulares* conocidas como *regex* o *regexp* que son secuencias de caracteres que nos ayudan a conformar patrones de busqueda[[1]](#footnote-1) Hecho lo anterior, logramos recuperar de la descripción, un 30,2% de los datos de área faltantes y complementando con KNN, se imputo el resto de los datos perdidos. Con lo anterior conseguimos completar la información faltante de la base datos en los atributos relevantes para nuestra predicción.

La variable precio no tuvo mayores inconvenientes salvo los *outliers*, que corresponden a valores que están fuera de los bigotes, es decir las líneas que se determinan como el tercer cuartil + 1.5 veces el rango intercuartílico y el primer cuartil -1.5 veces el rango intercuartílico.

Procedímos entonces a generar variables de distancia a lugares referentes como parques, centros comerciales, considerando que estos son atributos que pueden tener influencia en el precio de la propiedad, aquí utilizamos la distancia euclidiana De esta manera contamos con una base completa y variables relevantes para la predicción.

2.1.2 Variables espaciales de distancia:

Tomando en cuenta que la localización de las propiedades es una variable importante a 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 las 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.

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

Mapa 1: apartamentos y av principales

Map

Description automatically generated

Mapa 2: Apartamentos y parques

2.2 Estadísticas Descriptivas

Las estadísticas descriptivas de las variables incluidas en nuestro modelo se muestran en la siguiente tabla. El precio promedio de la vivienda (casa o apartamento) asciende a un poco más de $654 mil millones, la superficie o área promedio alcanza los 122 metros cuadrados, el número de habitaciones promedio es de 3, y tienen en promedio 3 baños. En cuanto a los atributos, la distancia promedio al parque mas cercano es de 160.7 metros, a una avenida principal es de 267 metros, a un centro comercial es de 295 metros y a la universidad llega a ser 1200 metros.

El cuadro 1 muestra las estadísticas descriptivas de las variables que se utilizaron en el modelo

Table

Description automatically generated

Cuadro 1: Estadísticas variables

En esta gráfica se muestra que el precio de las casas es más elevado respecto a los apartamentos. Y la correlación es baja frente a los atributos, siendo la *Distancia a parque* la variable con mayor correlación (0.082). La asociación con el parque es positiva y respecto a la avenida principal, a centro comercial y universidad es negativa, siendo este el esperado.

Diagram

Description automatically generated with low confidence

Gráfica 1: Precio y distancias

El cuanto a la relación con el tamaño de la vivienda o área, esta es positiva, lo que significa que a mayor área de la vivienda, mayor es el precio.

Text

Description automatically generated

Gráfica 2: Precio - Area

3. Modelos 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 el 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 nuestros modelos, el precio dependerá de un conjunto de características estructurales como número de habitaciones, área de la propiedad, número 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.

3.1 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.

3.2 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.

3.3 Modelos de Regularización

También se han planteado modelos de regularización para contrastarlo con el mejor modelo. El modelo que se planteo es el siguiente, con distintas especificaciones:

La variable dependiente es el precio de la vivienda (casa o apartamento) medido en pesos colombianos de la ciudad de Bogotá.

La primera variable predictora es el tipo de propiedad, es decir si es casa o apartamento, no se le hizo ningún tratamiento pues es una dummy; la otra variable es el área de la propiedad medido en metros cuadrados, estos datos fueron obtenidos de la descripción del anuncio, lográndose reducir el número de observaciones faltantes de la base train de 30079 a 8798, luego estos datos fueron imputados mediante el algoritmo K-Nearest Neighbors (KNN), considerando los seis vecinos más cercanos; la variable habitaciones resulta del valor máximo entre bedrooms y rooms; la variable baños tuvo 10071 valores faltantes y fue imputada mediante KNN con seis vecinos cercano, y se incluyen las variables de distancia, explicadas anteriormente.

Adicionalmente en la base train, mediante la función boxplot se detectaron y excluyeron los outliers de la variable dependiente y las variables independientes, considerando como todo valor que está fuera de los bigotes, es decir las líneas que se determinan como el tercer cuartil + 1.5 veces el rango intercuartílico y el primer cuartil -1.5 veces el rango intercuartílico.

En la base de datos test, con el propósito de no perder observaciones, en el caso del área los outliers fueron reemplazados por el promedio y dentro del rango 500 y 30 metros cuadrados, y los NAs fueron imputados con KNN.

En este grupo de modelos se plantean diversas especificaciones como lineal, semilogarítmico y polinómicos. Los resultados de predicción que mejor se ajustan provienen de una especificación lineal, cuyos coeficientes se muestran en la siguiente tabla:

Modelo de Regresión



El grado de penalización de los modelos de regularización está controlado por el hiperparámetro lambda, cuyos óptimos son los siguientes:

Hiperparámetros



El mejor modelo es el Lasso, porque dentro de la muestra tiene la menor métrica RMSE (161362733), frente al OLS (161555081), Ridge (161709161) y Elestic Net (161698822).

Los precios predichos son estimados con los predictores de la base test, cuya distribución es la siguiente:



El escore obtenido con dicha predicción del modelo Lasso en kaggle es un MAE de 282445492.70880, que es superado por el modelo Super Learner. Por lo tanto, en este caso, los modelos de regularización no superan las predicciones del modelo Super Learner.

Los otros modelos estimados de este grupo son los no lineales, es decir donde el precio está expresado en logaritmos y otro con los predictores en polinomios. Los resultados se muestran en los anexos. El modelo semi-logaritmico, si bien es lo recomendable para reducir la magnitud de la variable dependiente, el problema se da en las predicciones, pues llega a predecir por ejemplo precios por encima de $ 4 mil millones, lo cual empeora la métrica MAE. Por su parte, el modelo con el polinomio de las variables dependientes e interacciones entre ellas, arroja valores predichos del precio negativos para el test, lo que supone un MAE peor fuera de la muestra.

3. Conclusiones y Recomendaciones

4. Bibliografía

1. Lucas Antonio, Valentina e Ignacio. *“Tutorial de Procesamiento Natural de Texto”.*  [↑](#footnote-ref-1)