

Problem Set 3: Ganancias con Machine Learning

Andrea Margarita Beleño

200620739

E-mail: a.beleno@uniandes.edu.co

María Valeria Gaona

202214418

E-mail: mv.gaona@uniandes.edu.co

Resumen—El precio de las viviendas cuenta con diferentes factores que se tienen en consideración para establecer un valor. Por lo tanto, contar con un modelo en el que se identifiquen dichas características principales es fundamental para poder generar una predicción lo más óptima posible. Por lo tanto, en el siguiente documento se realizará el modelo de predicción para los precios de las viviendas por medio de Random Forest. El link al Github del presente taller, se encuentra en el siguiente enlace: <https://github.com/mvgaona/Problem-Set-3>

I. INTRODUCCIÓN

El precio de una vivienda está dado por diferentes factores como el área total, el número de baños, número de habitaciones, la distancia a puntos importantes de la ciudad, entre otros elementos. Además, cada ciudad cuenta con diferentes características que aportan a que los vendedores decidan establecer un precio. Sin embargo, se puede generar un modelo en donde se pueda observar el efecto de cada una de esas características esenciales que da el valor de un inmueble y así, tanto el vendedor como el comprador puedan realizar futuras compras y/o ventas de inmuebles. Por otra parte, generar una predicción de estos valores es una herramienta para conocer el comportamiento económico de una sociedad y además, cada individuo pueda obtener información valiosa para la toma de decisiones a mediano o largo plazo. De acuerdo con lo anterior, en el presente documento se presenta el proceso de la limpieza de datos pertinente para realizar la caracterización de las variables fundamentales y así, realizar el modelo de predicción de los precios de las viviendas en la localidad de Chapinero en la ciudad de Bogotá y en el área del poblado en Medellín en Colombia.

II. DATOS

El precio de una vivienda puede estar dado por diferentes factores, tanto económicos como sociales. En el caso Colombiano, estos precios pueden estar influenciados por elementos propios de la vivienda y a su vez, por elementos geoespaciales, como la distancia a diversos lugares públicos (Bares, transporte público, etc). Para realizar un modelo de predicción de precios de la vivienda, es necesario contar con las variables determinantes y relevantes del precio, para que este modelo sea robusto, pero no se incurran en gastos que entorpezcan esta labor. En el presente modelo de predicción se tomó en cuenta solo los datos de las viviendas ubicadas en las localidades objetivo (Chapinero y Poblado) y el total de datos (Base de entrenamiento y base de testeo) es de 24.843 viviendas. Dentro del proceso de recolección de datos puede

ocurrir que no se encuentre toda la información necesaria, ya sea porque la base de datos no provee dicha información o, por el contrario, no se pudo recolectar toda la información de cada vivienda. Es por eso que es necesario realizar un proceso de limpieza y recuperación de datos. Se realizaron diversos procesos para poder contar con toda la información:

-Se cuenta con variables las cuales ya tienen la información completa, por lo tanto no es necesario realizar ningún tratamiento:

- **Ubicación:** La ubicación del inmueble en el modelo será esencial para el análisis ya que se analizarán dos localidades de dos ciudades diferentes y los precios de las ciudades si bien están dados por características similares, el costo de vida por ciudad también influye en el valor de la vivienda. Esta variable es categórica, y se tiene que la mayoría de datos son de la ciudad de Bogotá (ver Cuadro I).
- **Tipo de propiedad:** Si la propiedad es un apartamento o si es una casa influye sustancialmente en el precio, ya que el área de una casa suele ser más grande, cuenta con más oportunidad de reformas y espacios más amplios de esparcimiento. Esta variable categórica cuenta con una mayor proporción de apartamentos (Ver Cuadro II).

-Para algunas variables existentes en la base de datos, se generó la comparación entre dicha información y los valores hallados por medio de la inspección de la descripción de cada inmueble, generando las siguientes variables sin información faltante:

- **Habitaciones:** El número de habitaciones de la vivienda es determinante en el precio de la misma, ya que se puede contar con un aproximado del espacio y de cuántos individuos pueden vivir, es decir, entre más habitaciones, el precio del inmueble tiende a incrementarse. Por lo tanto, dentro del análisis descriptivo de este predictor se encuentra que es una variable numérica, con media de 3 habitaciones (Ver Cuadro III).
- **Cantidad de baños:** Este es otro factor de decisión importante en el precio del inmueble, ya que sin baños, los individuos no pueden satisfacer las necesidades básicas de aseo. Por lo tanto es necesario contar con al menos un baño y con ello, se puede identificar que a medida que aumentan la cantidad de baños, el precio de igual manera se verá afectado e incrementará. De acuerdo con lo anterior, se observa que en la variable numérica, la media fue de 3 baños (ver Cuadro IV).

-De acuerdo con la información proporcionada por la descripción de cada inmueble, se generaron las siguientes variables:

- **Ascensor:** Si el inmueble cuenta con ascensor o no es determinante en el caso de los apartamentos de pisos altos, es por eso que se considera un predictor importante del modelo, ya que quienes viven en pisos más altos, buscan apartamentos donde tenga ascensor y a su vez, el precio aumenta si el inmueble cuenta con este. Dentro del análisis descriptivo de la variable, es posible identificar que es una variable dicótoma, donde 1 corresponde a si el inmueble cuenta con ascensor y 0 si no (Ver Cuadro V). Los inmuebles que no cuentan con ascensor también pueden ser apartamentos que no cuentan con uno y casas que no lo necesitan.
- **Parqueadero:** En la actualidad, gran parte de los individuos cuentan con uno o más carros, por lo tanto, es necesario que el inmueble cuente con al menos un parqueadero, por lo tanto, 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 dicótoma, en donde 1 hace referencia a que la casa o el apartamento cuenta con al menos un parqueadero y 0 que no lo hace. La proporción de inmuebles que cuentan con al menos un parqueadero es que la mayoría de inmuebles de la base no cuenta con parqueaderos (ver cuadro VI). La proporción puede estar dada debido a que algunos vendedores no ofrecieron dicha información, por ejemplo.

-De acuerdo con la ubicación geoespacial, se pueden recuperar datos por medio de la búsqueda de características similares de otras viviendas y dicho promedio, adjuntarlo al inmueble que no cuenta con información, ya que al estar en un perímetro cercano, los inmuebles suelen presentar características similares.

- **Área:** El área de un inmueble es fundamental para conocer el precio del mismo, ya que a medida que el área aumente, también el precio. Esto se da debido a que un espacio amplio permite contar con un mayor número de habitaciones, baños, una cocina más amplia y demás espacios de esparcimiento. Además, permite más posibilidades de remodelación y de inversión. Por lo tanto, ante la variable numérica presentada, se encuentran que se tiene una media de 286 m². Para más información (ver Cuadro VII)

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

- **Distancia a Transporte público:** Contar con al menos un transporte público cerca de la vivienda es fundamental para poder desplazarse con facilidad. Por lo tanto, esta variable presenta la distancia mínima que tiene el inmueble con al menos un medio de transporte público, ya que a medida que esta distancia sea más corta, el precio

puede incrementarse. Para esta variable se cuenta con una media de 1,5km (Ver Cuadro VIII)

- **Distancia a Bares:** La distancia mínima a bares es otro factor a tener en cuenta, ya que, en general, este tipo de establecimiento hace que pueda generar incomodidad a los habitantes en las horas de la noche. Por lo tanto, al estar cada vez más alejados de un bar, el comprador puede estar más interesado en el inmueble. De acuerdo con la variable numérica, se presenta que la media de distancia se encuentra en 694m (Ver Cuadro IX).
- **Distancia a Parques:** La distancia mínima a por lo menos un parque es otro predictor relevante dentro del modelo, ya que, en general, los individuos buscan tener un espacio verde y de recreación cerca, por tema de hijos o mascotas, o porque los compradores también desean un espacio al aire libre para su propio entrenamiento. Por lo tanto, a medida que esta distancia va disminuyendo, el precio del inmueble tiende a incrementar. Esta variable numérica tiene una media de 230m (Ver Cuadro X).

III. MAPAS

Los mapas de la localidad de Chapinero en Bogotá y El Poblado en Medellín, presentados en las Figuras 1 y 2, evidencian la distribución de bares, estaciones de buses, y parques que se encuentran en cada uno de los sectores, junto con las viviendas que se encuentran en las localidades anteriormente mencionadas. Esto permite contar con un análisis gráfico de cada vivienda y así, un comprador, podría contar con información gráfica que le permite tomar decisiones de manera informada.

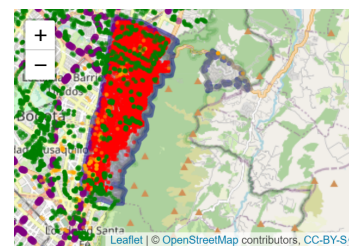


Figura 1. Mapa Chapinero, Bogotá- Polígono color azul, apartamentos en rojo, transp. público en morado, parques en verde y naranja para bares

IV. MODELO Y RESULTADOS

Con base en las variables presentadas en la sección II, se procedió a realizar varios modelos para determinar cuál era el más apropiado para realizar la predicción de precio de viviendas. Se analizó la posibilidad de partir la base *train* (con 13.693 datos) en un *subtrain* y *subtest*, sin embargo, teniendo en cuenta que la base de *test* original cuenta con 11.150 observaciones, realizar la subdivisión de *train* en un menor número de muestras, podría ir en perjuicio de la predicción por contar con un menor número de datos para entrenar, por lo cual, se decidió implementar los modelos sobre la base *train*. En primera instancia, se realizó una prueba utilizando las 10 variables mencionadas en la sección

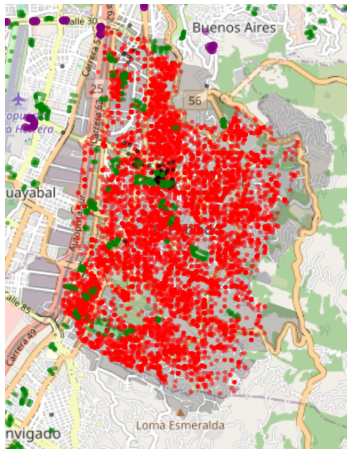


Figura 2. Mapa El Poblado, Medellín- Polígono color azul, apartamentos en rojo, transp. público en morado, parques en verde y negro para bares

II y se realizó un modelo de regresión con *OLS* (modelo 1), luego, se realizó la regularización de Lasso para la regresión *OLS* (modelo 2). Adicionalmente, se realizaron los modelos *Random Forest* (modelo 3) y *XGBoost* (modelo 4), teniendo en cuenta las variables mencionadas anteriormente, para explorar si estas alternativas para datos no lineales podían dar un mejor resultado. Las 10 variables utilizadas para realizar los modelos mencionados anteriormente son las siguientes:

- *Ubicación, Tipo de vivienda, Habitaciones, Baños, Área, Distancia a bares, Distancia a transporte público, As-censor, Parqueaderos, Distancia a parques.*

Luego de correr los modelos, se decidió disminuir el número de variables para realizar las regresiones para los 4 modelos mencionados anteriormente, utilizando 5 variables que se mencionan a continuación:

- *Habitaciones, Baños, Área, Distancia a bares, Distancia a transporte público*

Se obtuvieron las variables relevantes en el modelo *Random Forest*, que son las 6 que se presentan a continuación :

- *Habitaciones, Baños, Área, Distancia a bares, Distancia a transporte público, Distancia a parques*

Utilizando las variables presentadas anteriormente, se corrieron los modelos mencionados inicialmente (*OLS*, *Lasso*, *Random Forest*, *XGBoost*, para un total de 12 modelos. El parámetro de comparación entre modelos fue el RMSE. Se observó que al disminuir la cantidad de variables, el RMSE aumentó, por lo cual, se decidió realizar la comparación con los 4 modelos descritos anteriormente en términos de RMSE. Al hacer la comparación, se obtuvo que el modelo de *Random Forest* con un número de árboles igual a 5, fue el de menor RMSE (este resultado se presenta en la Figura 3).

Ahora bien, existen un tema relevante a tener en cuenta para decidir cuál modelo se debe escoger para realizar la predicción, como lo es la relación de dinero para compra (\$) vs. la cantidad de viviendas compradas (*Viv*) (no solo en términos de RMSE), en donde se escoge el modelo con el que se obtenga la menor relación, teniendo en cuenta que entre más casas compradas y

menor valor gastado, la relación $\$/Viv$ será menor. Se realizó la comparación de cada set de modelos, con las 10, 5 y 6 variables, de lo cual se obtuvo la información presentada en los Cuadros XI, XII y XIII. Lo anterior, para concluir que el mejor modelo para realizar la predicción fue el de *Random Forest* con 10 variables. Cabe anotar que la cantidad de árboles contemplada para el modelo fue de 5, por lo cual, se decidió entrenar el modelo variando el número de árboles, aumentándolo para disminuir el RMSE; después de intentar varios valores, se decidió fijar el número de árboles en 1.000 (el número de predictores por default fue el \sqrt{p}); para este caso, se obtuvo un menor RMSE, y al realizar el cálculo de relación dinero gastado vs viviendas compradas, también fue el más eficiente. Finalmente, la predicción obtenida se guardó en el archivo .csv solicitado.

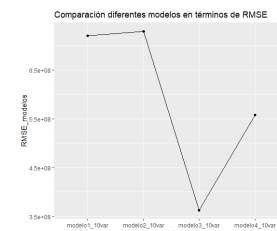


Figura 3. Comparación RMSE modelos para obtener price

V. CONCLUSIONES Y RECOMENDACIONES

- La adquisición de información del texto de descripción es muy dependiente de patrones fijos que permitan extraer los datos de interés fácilmente, por ello, es posible que haya faltado información para tener la información completa de los datos.
- Fue un bastante interesante obtener los datos geográficos para utilizarlos como predictores o para imputar datos de variables existentes, fue en la parte del PS3 que más se demoró para obtener y que representó un gran reto y esfuerzo, sin embargo, consideramos la información obtenida muy relevante y que puede servir para futuras aplicaciones.
- Los resultados más aceptables para el ejercicio fueron los que tuvieron en cuenta mayor cantidad de variables para obtener la predicción.
- Para el presente trabajo, el modelo con menor RMSE y que también presentó una menor relación precio vs vivienda, fue el escogido para realizar la predicción y fue el que se entrenó posteriormente para obtener la predicción.
- Como recomendación, con mayor cantidad de tiempo, se hubiese podido realizar otros modelos variando los hiperparámetros del modelo *XGBoost*.

APÉNDICE

CUADROS DE VARIABLES DESCRIPTIVAS

Se presentan los Cuadros donde se consignan las estadísticas descriptivas de las variables explicadas en la sección II.

Cuadro I
CUADRO DESCRIPTIVO *Ubicación*

Ubicación	
Bogotá D.C.	Medellín
14.244	10.599
(57,34 %)	(42,7 %)

Cuadro II
CUADRO DESCRIPTIVO *Tipo de vivienda*

Tipo de Vivienda	
Casa	Apartamento
2.038	22.805
(8,2 %)	(91,8 %)

Cuadro III
CUADRO DESCRIPTIVO *Habitaciones*

Habitaciones			
Mín	Media	Máx.	Moda
1	3	11	3

Cuadro IV
CUADRO DESCRIPTIVO *Número de baños*

Número de baños			
Mín	Media	Máx.	Moda
1	3	13	2

Cuadro V
CUADRO DESCRIPTIVO *Ascensor*

Ascensor	
Tiene ascensor	No tiene ascensor
19.388 (78,04 %)	5.455(21,96 %)

Cuadro VI
CUADRO DESCRIPTIVO *Parqueadero*

Parqueadero	
Tiene parqueadero	No tiene parqueadero
7.885(31,74 %)	16.958(68,26 %)

Cuadro VII
CUADRO DESCRIPTIVO *Área*

Área (m ²)			
Mín	Media	Máx.	Moda
74,06	286,75	3.940,14	193,7414

Cuadro VIII
CUADRO DESCRIPTIVO *Transporte Público-TP*

Distancia TP (m)			
Mín	Media	Máx.	Moda
4,397	1.519,834	4.472,144	2.168,091

Cuadro IX
CUADRO DESCRIPTIVO *Distancia bares*

Distancia bares (m)			
Mín	Media	Máx.	Moda
2,101	1.694,033	3.062,209	487,8808

Cuadro X
CUADRO DESCRIPTIVO *Distancia parques*

Distancia parques (m)			
Mín	Media	Máx.	Moda
0,4975	230,5497	1.567,7500	478,1448

Cuadro XI
EVALUACIÓN MODELOS CON 10 VARIABLES

	Evaluación Modelos Variables 10				
	<i>OLS</i>	<i>Lasso</i>	<i>RF5</i>	<i>RF1000</i>	<i>XGBoost</i>
Tot.\$ gastado	1,14e+13	1,14+13	1,11e+13	1,08e+13	1,09e+13
Cant.Viv.Compradas	8.657	8.994	9.396	9.605	8.517
Rel.\$/Viv.compradas	1,315e+9	1,267e+9	1,182e+9	1,128e+9	1,279e+9

Cuadro XII
EVALUACIÓN MODELOS CON 6 VARIABLES

	Evaluación Modelos Variables 6			
	<i>OLS</i>	<i>Lasso</i>	<i>RF5</i>	<i>XGBoost</i>
Tot.\$ gastado	1,14e+13	1,14+13	1,08e+13	1,09e+13
Cant.Viv.Compradas	8.737	9.045	9.204	8.614
Rel.\$/Viv.compradas	1,31e+9	1,26e+9	1,181e+9	1,272e+9

Cuadro XIII
EVALUACIÓN MODELOS CON 5 VARIABLES

	Evaluación Modelos Variables 5			
	<i>OLS</i>	<i>Lasso</i>	<i>RF5</i>	<i>XGBoost</i>
Tot.\$ gastado	1,15e+13	1,13+13	1,10e+13	1,10e+13
Cant.Viv.Compradas	8.816	9.062	9.213	8.697
Rel.\$/Viv.compradas	1,31e+9	1,26e+9	1,197e+9	1,269e+9