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 que explican dicho valor. 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 conoce cada uno de estos impactos para realizar futuras compras y/o ventas de inmuebles. Por otra parte, generar una predicción de estos valores es una herramienta óptima para conocer el comportamiento económico de una sociedad y además, cada individuo puede 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 continuar con 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, ambas ciudades Colombianas.

II. Datos

El precio de una vivienda puede estar dados 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 la investigación. 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 encuentren 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:

1. 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, la cual presenta la proporción de cada ciudad dentro de la base de datos, presentada en el 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 la proporción presentada en el Cuadro II.

2. De acuerdo con los valores proporcionados de cada variable, 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 puedes vivir con la mayor calidad de vida posible, 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, la cual presenta los valores respecto al número de habitaciones consignados en el 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 se evidencian los valores respecto a el análisis de todos los inmuebles consignados en el Cuadro IV.

3. 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 esta herramienta de desplazamiento dentro del edificio. Dentro del análisis descriptivo de la variable, es posible identificar que es una variable categórica de 2 niveles: 1 y 0, donde 1 corresponde a si el inmueble cuenta con ascensor y 0 si no. Esta variable cuenta con la proporción presentada en el Cuadro Cuadro V. Los inmuebles que no cuentan con ascensor también contar con apartamentos que no cuentan con uno y casas que no necesitan, las cuales 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 para que el individuo no incurra en gastos adicionales en encontrar un lugar seguro para su(s) carro(s), 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 categórica de dos niveles, 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 se presenta en el Cuadro VI. La proporción puede estar dada debido a que algunos vendedores no ofrecieron dicha información, por ejemplo.
4. De acuerdo con la ubicación geoespacial, se pueden recuperar datos por medio el hallazgo 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, este también aumenta. 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, en el Cuadro VII se encuentran los valores respecto al análisis de las viviendas.
5. Por medio de la ubicación geoespacial se pueden hallar diferentes variables que pueden predecir los precios de las viviendas, ya que existen puntos importantes de las ciudades, las cuales entre menos distancia exista entre el inmueble y el punto, el valor del inmueble aumenta
   - *Transporte público*: Contar con al menos un transporte público cerca de la vivienda es fundamental para poder analizar la facilidad vial y que tan alejado puede estar del resto de la ciudad. 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. De acuerdo al análisis, se evidencian las distancias mínimas presentadas en el Cuadro VIII.
   -*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 presentan las distancias mínimas en el Cuadro IX.
   - *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, ya sea porque dentro del hogar tienen niños los cuales puedan entretenerse, mascotas que necesiten espacio libre para derrochar su energía o cumplir sus necesidades fisiológicas 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. La variable numérica presenta la descripción estadística en el Cuadro XI.

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, por ejemplo, puede contar con información gráfica que le permite tomar decisiones de manera informada.

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* 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, al utilizar incluso menos datos que la base *test* principal, por lo cual, se decidió

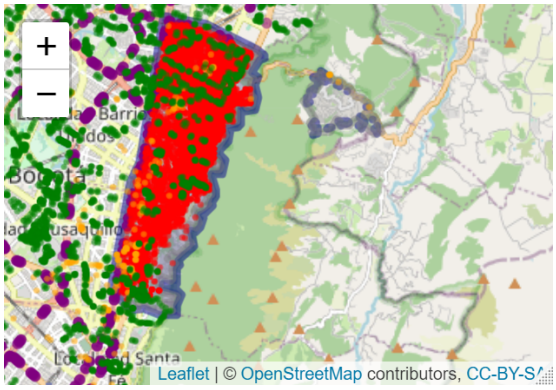

Figura 1. Mapa Chapinero, Bogotá- *Polígono color azul, apartamentos en rojo, transp. público en morado, parques en verde y naranja para bares*

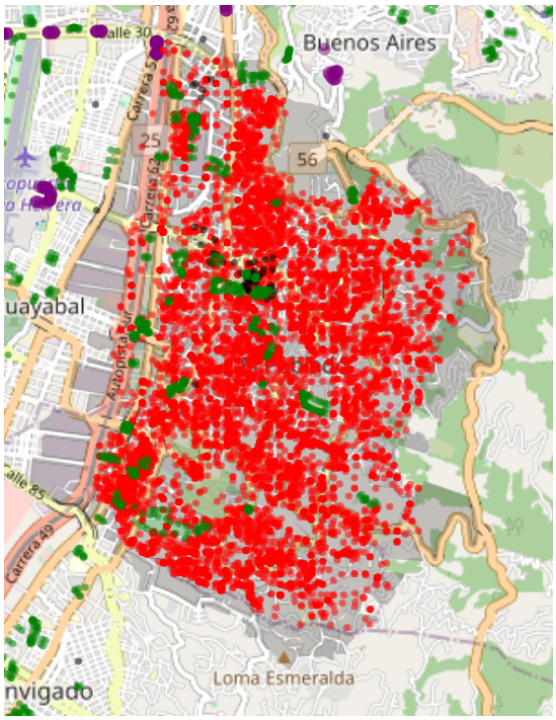

Figura 2. Mapa El Poblado, Medellín- *Polígono color azul, apartamentos en rojo, transp. público en morado, parques en verde y negro para bares*

realizar implementar los modelos sobre la base *train* reducida. En primera instancia, se realizó un modelo de regresión con *OLS* (modelo 1), tomando las 10 variables presentadas anteriormente, luego, se realizó la regularización de Lasso para la regresión OLS U (modelo 2). Adicionalmente, se realizaron los modelos *Random Forest* (modelo 3) y *XGBoost* (modelo 4), teniendo en cuenta las variables mencionadas anteriormente. A manera ilustrativa, se presentará el modelo OLS utilizado contemplando las 10 variables.

$$Price = \beta_0 + \beta_1 Ubicacion + \beta_2 Tipo{-}vivienda + \beta_3 Habitac. + \beta_4 Banios + \beta_5 Ascensor + \beta_6 Parqueadero + \beta_7 Area + \beta_8 Dist.Transp.Pub. + \beta_9 Dist.bares + \beta_{10} Dist.parques + u$$

Luego de correr los modelos, la significancia de las variables variaba de un modelo a otro, por ejemplo, para *OLS* existían 5 variables más significativas para el modelo vs las que se obtuvieron en *Random Forest* que fueron 6 con mayor peso para el modelo. A continuación se presentarán las variables relevantes obtenidas en *OLS*:

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

A continuación se presentarán las variables relevantes obtenidas en *Random Forest*:

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

Utilizando las variables presentadas anteriormente, se corrieron los modelos mencionados incialmente (*OLS*, *Lasso*, *Random Forest*, *XGBoost*, es decir, se contó con 16 modelos. Al hacer la comparación de la raíz cuadrada del MSE, se obtuvo que para los modelos incluyendo las 10 variables se presentó el menor MSE. El modelo de *Random Forest* fue entrenado variando el número de árboles, para el cual se obtuvo un valor de $n.trees = 1000$.

En la Figura 3 , se presenta la comparación de los primeros 4 modelos con las 10 variables en términos de la raíz del MSE para tener una primera aproximación de cuál modelo escoger. Ahora bien, existen otros temas relevantes 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, que se escoge el de menor ratio, porque para esos casos se compran más viviendas a un costo 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 el Cuadro **??**. Lo anterior, para concluir que el mejor modelo para realizar la predicción fue el de *Random Forest* con 10 variables. Finalmente, la predicción obtenida se guardó en el archivo .csv solicitado.

V. Conclusiones y recomendaciones

- Fue un reto bastante interesante obtener los datos geográficos para utilizarlos como predictores, teniendo en cuenta
- Se encontraron diferencias significativas en cuanto a la predicción por clasificación y por regresión a través del ingreso. Los parámetros utilizados para realizar la clasificación de Pobre o No Pobre, puede incidir.
- La regresión del Ingreso presentó un MSE alto para todos los modelos, motivo por el cual, se puede estar dando este resultado en donde hay pobres clasificados como no pobres.

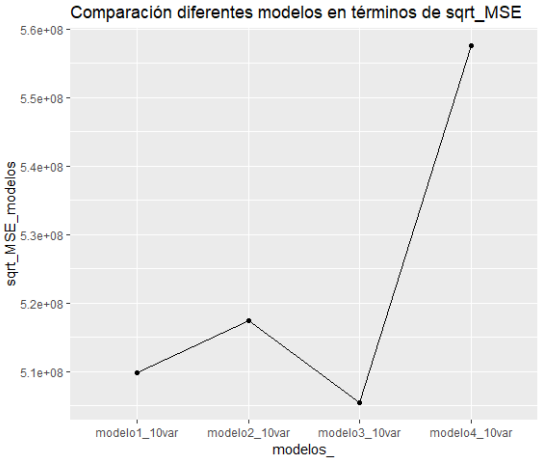

Figura 3. Comparación MSE modelos para obtener *price*

- Los resultados más aceptables para el ejercicio fueron los que tuvieron en cuenta mayor cantidad de variables para obtener la predicción (variable dependiente).

APÉNDICE
CUADROS DE VARIABLES DESCRIPTIVAS

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 (m2)			
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 VARIABLES 10

Distancia parques (m)			
OLS	**Lasso**	**RF, n=5**	**RF, n=1000**
0,4975	230,5497	1.567,7500	478,1448

Cuadro XII

	Dependent variable:
	price
factor(Medellin)1	−787,636,375.000***
	(61,083,213.000)
factor(Apto)1	207,203,199.000***
	(34,086,495.000)
factor(parqueaderoT)1	−63,801,814.000***
	(13,453,959.000)
factor(ascensorT)1	124,196,586.000***
	(14,752,857.000)
bathrooms	153,238,037.000***
	(5,325,096.000)
habitaciones	317,659,483.000***
	(7,138,710.000)
min_dist_bar_	−13,904.660
	(25,102.360)
min_dist_transp_	526,386.800***
	(16,935.530)
min_dist_park	−720,497.400***
	(62,258.320)
surface_new_3	123,250.800***
	(36,015.530)
Constant	−463,068,798.000***
	(44,297,272.000)
Observations	13,693
R^2	0.370
Adjusted R^2	0.369
Residual Std. Error	720,926,630.000 (df = 13682)
F Statistic	802.327*** (df = 10; 13682)
Note:	*p<0.1; **p<0.05; ***p<0.01