PROBLEM SET 3- MAKING MONEY WITH ML?

Camila Alejandra Velasco Contreras

Jairo Alexander Torres Preciado

Enlace repositorio: camivel/ProblemSet3 (github.com)

1. Introducción

En la literatura se encuentran varias aplicaciones de modelos de machine learning para la predicción de precios de vivienda en Colombia. En particular, se destacan los trabajos de Correa et al. (2020), Pérez-Rave et al. (2020) para apartamentos y Pérez-Rave et al. (2019) para todo tipo de viviendas. Adicionalmente, se destaca una aplicación de tipo regional en Antioquia por parte de Gutiérrez y Parra (2022).

En el presente ejercicio se utilizó una base de datos proveniente de Properati para predecir los precios de la vivienda en Bogotá y Medellín, la cual se enriqueció con información del DANE y de OSM. En el caso de Bogotá, se utilizaron datos correspondientes a toda la ciudad, con el fin de predecir precios en la localidad de Chapinero. De forma similar, en Medellín se utilizaron datos generales de la ciudad para predecir precios de la comuna de El Poblado. Se entrenaron diversos modelos para los datos conjuntos de las dos ciudades y se encontró que el modelo que mejor predice los precios de la vivienda es el random forest, el cual combina 20 árboles y arroja un MSE de 0.1190.

2. Datos

Limpieza de datos y creación nuevas variables. Para empezar, juntamos las bases *train* y *test* en una sola con el propósito de que las nuevas variables se creen para las dos bases, para poder identificarlas y separarlas posteriormente agregamos una nueva columna que identificaba si la observación hace parte de la base *train* o *test*. Esta nueva base tiene en total 118,717 observaciones. Observamos que la base tiene gran cantidad de missings values, por ejemplo, para variables fundamentales como área de superficie total y baños el 74% (88,936) y 29% (34,343) de las observaciones no contienen información, respectivamente. Para tratar de rescatar la mayor cantidad de observaciones utilizamos varias estrategias que se describen a continuación.

Extracción información descripción. Usando la descripción de la propiedad en venta rescatamos información del número de baños y el área de la propiedad. Primero, pasamos todo el texto a minúscula e inspeccionamos los datos para identificar los patrones que más se repetían para dar información respecto a la superficie y el número de baños de la propiedad, identificamos 10 patrones distintos para área y 3 para baños. Logramos recuperar más de 40 mil observaciones para superficie y para baño cerca de 26 mil. Cuando extrajimos los patrones limpiamos las nuevas variables de tal manera que quedaran sólo en números y todas las observaciones con el mismo separador de decimales. Imputamos a las variables de interés (superficie total y baños) los valores de las nuevas variables. También imputamos la información de la variable superficie cubierta a superficie total, para aquellas observaciones con missings values .

Imputación información de área por manzana. Continuamos teniendo cerca de 50 mil missing values para la variable superficie total por lo cual vamos a imputar el promedio del área de las propiedades en la manzana. También recuperamos información del número de cuartos y estrato promedio por manzana. Para hacer esto fue necesario bajar del geoportal del DANE la información del Marco Geoestadístico Nacional (MGN) del año 2017 los datos por manzana de Bogotá y Antioquia. De los datos de Antioquia nos quedamos solo con los de Medellín, filtrando por el código de municipio. Utilizamos los datos de hogares, vivienda y manzanas, pegándolas todas en una sola base para cada ciudad. Pegamos estas nuevas bases con

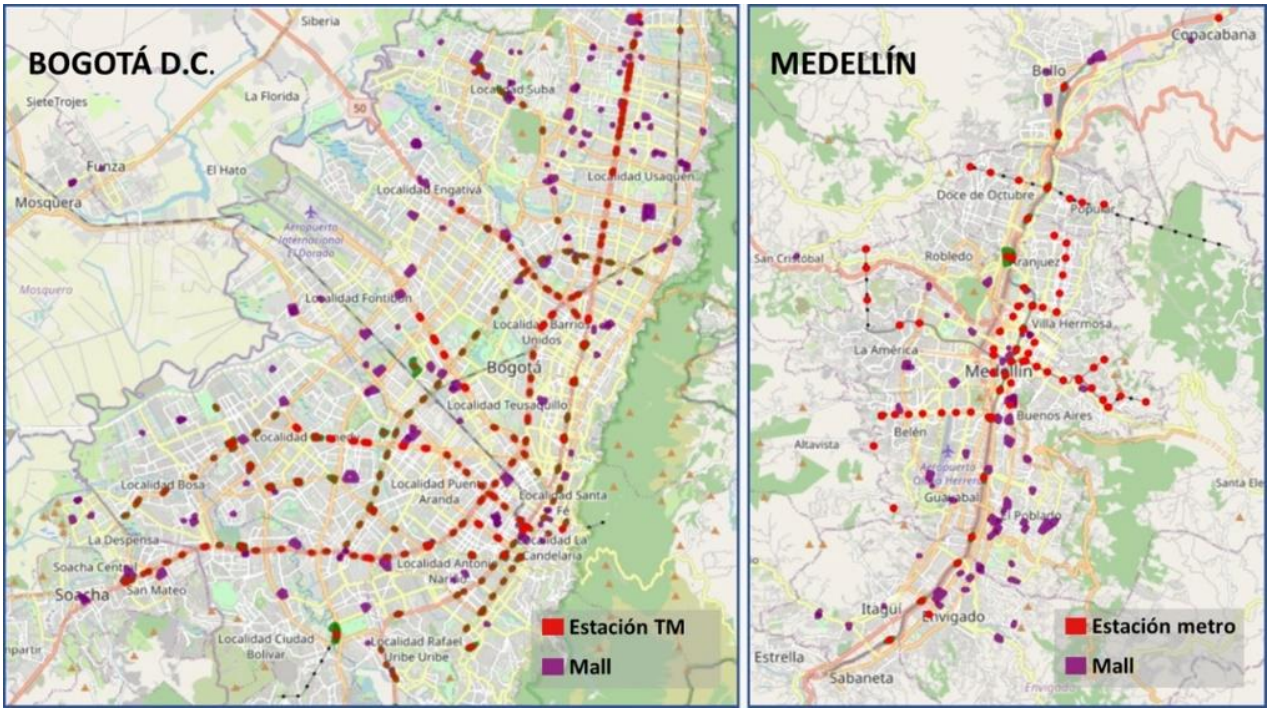
información por manzana a nuestra base original. Creamos una nueva variable con el área promedio por manzana, estos datos los imputamos a las observaciones con missings de nuestra variable de superficie total. Ahora pasamos de tener 50 mil missings a 10 mil.

Creamos variable estrato. Usando la información proveniente del DANE creamos una variable de estrato, imputando la media del estrato de la manzana a las propiedades ubicadas dentro de ella. Inspeccionamos las variables y encontramos que hay decimales y los filtramos. Logramos imputar el estrato a cerca de 74 mil observaciones.

K-vecinos cercanos para imputar missings. Seguimos teniendo varios missings values para superficie (10 mil), baños (24 mil) y estrato (45 mil). Para estas propiedades imputamos la observación más cercana (k=1), usando las coordenadas de la propiedad. Haciendo esto quedamos con cero missing values para las tres variables mencionadas.

***Creación variables geoespaciales*.** Usando la librería osmdata obtenemos la geolocalización de las estaciones de metro/ transmilenio y los centros comerciales de ambas ciudades.Creamos dos nuevas variables, la distancia mínima a un centro comercial y a una estación de transmilenio/metro a partir de esto. Con esto ya tenemos la base completa sin missings con nuestras variables que creemos pueden influir en el precio de la vivienda, separamos nuevamente las bases train y test. Le agregamos la variable de precio a nuestra variable de train.

Figura 1. Mapa de Bogotá y Medellín con centros comerciales y estaciones de metro y TM



***Análisis valores atípicos*.** Hicimos una inspección de nuestros datos y quitamos observaciones atípicas. Quitamos las propiedades con precios superiores a los 3 mil millones y menores a los 100 millones. Esto nos reduce las observaciones de 107,567 a 104,665. Graficamos la distribución de la variable precio tanto para la base train original y la cortada, dado que está sesgada a la izquierda aplicamos logaritmo. (Anexo 1)

Nos quedamos con las propiedades que tienen menos de 9 baños y por lo menos un cuarto, quedamos con 103.059 observaciones. También quitamos observaciones con superficies superiores a los 1000 metros cuadrados, quedamos con 102,721 observaciones.

Estadísticas descriptivas

Primero inspeccionamos la relación entre el logaritmo del precio y la ciudad, el estrato, número de baños y número de habitaciones. Se observa que el precio promedio de una propiedad es superior en Bogotá que en Medellín, vemos una relación creciente en el estrato y el número de baños y cuartos. (Anexo 2)

En el Anexo 3 se presentan las estadísticas descriptivas del precio y las variables explicativas por ciudad. En el Anexo 4 presentamos el t-test para las variables continuas, observamos que la diferencia de precios es estadísticamente significativamente entre ciudades pese a que la diferencia entre el área promedio de la propiedad entre ciudades no es estadísticamente significativa. También la diferencia entre la distancia mínima a un centro comercial y a una estación de metro/Transmilenio es estadísticamente significativa.

3. Modelos y Resultados

Se implementaron 5 modelos diferentes para determinar cuál de todos predice mejor el precio de las viviendas en Bogotá y Medellín en la base de entrenamiento. El primero de ellos fue un árbol de regresión sencillo, el cual eligió dos variables, la cantidad de baños y el estrato como predictores (Ver Anexo 6). En segundo lugar, se implementó un random forest con 20 árboles para simplificar el tiempo de procesamiento de los datos. En el Anexo 7 se observa que las variables más importantes son estrato, la distancia a centros comerciales, la distancia estaciones de metro o Transmilenio y el número de habitaciones. Los resultados de los tres modelos restantes se pueden ver en el Anexo 8. Se trata de regresiones por mínimos cuadrados ordinarios (OLS) con diferentes especificaciones. La primera regresión incluye a todos los predictores que componen la base de datos. Dado que las variables de baño y habitaciones parecen tener un comportamiento no lineal, se incluyeron versiones elevadas al cuadrado de estas en la segunda regresión, descartando otros posibles predictores. Por último, se corrió la regresión incluyendo interacciones con la variable de ciudad para controlar por las diferencias en el nivel de precios entre Bogotá y Medellín.

En cuanto al desempeño de los modelos, se tomó el error cuadrático medio (MSE) en la base de prueba como medida de comparación. El Anexo 9 muestra los MSE para los 5 modelos estimados. El mejor desempeño de predicción lo obtuvo el random forest con un MSE de 0.1190, seguido por la regresión OLS con interacciones de ciudad, el cual tiene un MSE de 0.2025. Por su parte, las peores predicciones las arrojó el modelo con las dos variables explicativas no lineales, el cual tuvo un MSE en la muestra de prueba de 0.3634. Los anteriores resultados evidencian la ventaja de los bosques al incluir varios árboles que no están correlacionados entre sí, lo cual hace que el promedio ente los árboles sea menos variable y, por lo tanto, haya más confiabilidad en las predicciones.

4. Conclusiones

El presente trabajo consistió en entrenar varios modelos para predecir el precio de la vivienda en la localidad de Chapinero en Bogotá y en la comuna de El Poblado en Medellín. Se utilizaron datos con información de las viviendas y del entorno en donde están ubicadas para crear modelos de precios hedónicos que permitieran hacer la mejor predicción. Se encontraron diferencias significativas entre los precios de Medellín y Bogotá, así como en las distancias mínimas de las viviendas a los centros comerciales en cada ciudad y las estaciones del metro y Transmilenio. El modelo entrenado que arrojó el mejor resultado de predicción de precios fue el random forest, el cual tuvo un MSE de 0.1190.

5. Referencias

Correa, M., Becerra, O., Otero, D., Laniado, H., Mateus, R. y Romero, D. (2020). Housing-Price Prediction in Colombia using Machine Learning. OSF Preprints W85Z2, Center for Open Science.

Gutiérrez, C. y Parra, D. (2022). Predicción del Precio de Vivienda en Antioquia. [Trabajo de grado especialización]. Universidad de Antioquia, Medellín.

Pérez-Rave, J.I, González-Echavarría, F. and Correa-Morales, J.C, (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. Journal of Property Research, Vol. 36, No. 1, 59–96

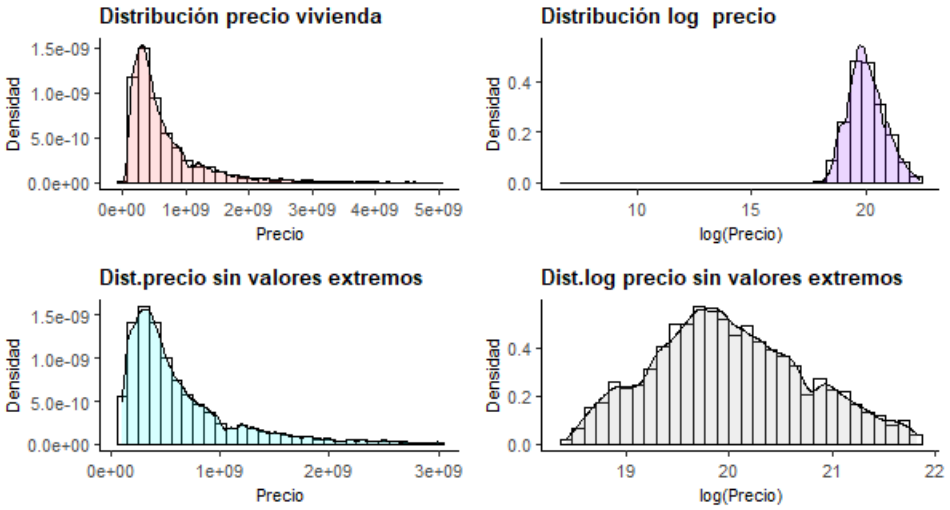
Pérez-Rave, J.I, González-Echavarría, F. and Correa-Morales, J.C, (2020). Modeling of apartment prices in a Colombian context from a machine learning approach with stable-important attributes. DYNA, 87(212), pp. 63-72.

6. Anexos

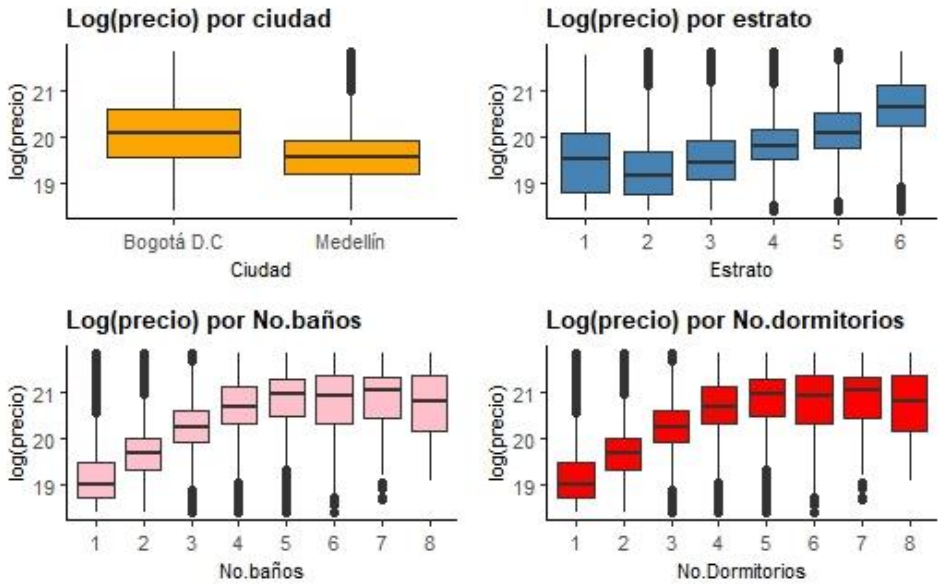
Anexo 1. Resumen Missing values

	Base original (train+ test)	Con imputación de información proveniente del texto	Imputando datos promedio por manzana	Imputación de k-vecinos cercanos
Área total	88.936	50.912	10.054	0
Baños	34.343	23.709		0
Estrato	-	-	45.192	0
Total observaciones	118.717	118.717	118.717	118.717

Anexo 2. Distribuciones de precio y log(precio)



Anexo 3. Boxplots de Log(precio) por variables



Anexo 4. Estadisticas descriptivas por ciudad

Característica	Bogotá D.C, N = 82,828[1]	Medellín, N = 20,231[1]
Precio vivienda	702,241,958 (558,668,076)	394,777,996 (319,592,696)
Tipo de propiedad		
Apartamento	63,026 (76%)	15,563 (77%)
Casa	19,802 (24%)	4,668 (23%)
Dormitorios	3.08 (1.43)	3.09 (1.06)
Baños	2.75 (1.16)	2.33 (0.96)
Estrato		
1	446 (0.5%)	202 (1.0%)
2	7,609 (9.2%)	1,251 (6.2%)
3	17,737 (21%)	5,139 (25%)
4	18,588 (22%)	6,250 (31%)
5	15,944 (19%)	6,714 (33%)
6	22,504 (27%)	675 (3.3%)
Superficie (m2)	132 (150)	127 (840)
Dist. min a una estación Transmilenio/Metro	1,095 (1,079)	976 (766)
Dist. min a un CC	733 (763)	922 (643)

[1]Promedio (Est.Desv)/ Frecuencia(%)

Anexo 5 . T-test entre ciudades

Característica	Bogotá D.C, N = 82,828[1]	Medellín, N = 20,231[1]	Difference[2]	95% CI[23]	p-value[2]
Precio vivienda	702,241,958 (558,668,076)	394,777,996 (319,592,696)	307,463,962	301,644,082, 313,283,843	<0.001
Dormitorios	3.08 (1.43)	3.09 (1.06)	-0.01	-0.03, 0.01	0.4
Baños	2.75 (1.16)	2.33 (0.96)	0.42	0.40, 0.43	<0.001
Superficie (m2)	132 (150)	127 (840)	5.4	-6.2, 17	0.4
Dist. min a una estación Transmilenio/Metro	1,095 (1,079)	976 (766)	118	105, 131	<0.001
Dit. min a un CC	733 (763)	922 (643)	-189	-200, -179	<0.001

[1]Promedio (Est.Desv)

[2]Welch Two Sample t-test

[3]CI = Intervalo de confianza

Anexo 6. Árbol de decisión

```
Regression tree:
tree(formula    =    log_price~l3+bedrooms+property_type+iluminado+banos_new+
estrato_new+surface_new+dist_estacion+distancia_mall,data=training)         --
------------------------------------------------------------------------

Variables actually used in tree construction:
[1] "baños"   "estrato"
Number of terminal nodes:  8
------------------------------------------------------------------------
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.45200 -0.30880 -0.02791  0.00000  0.28680  2.82100
```

Anexo 7. Importancia de variables Random Forest

Variable	%IncMSE
Ciudad (l3)	36.75342
Dormitorios	50.05109
Tipo propiedad	33.34058
Iluminado	10.99061
Baños	67.03175
Estrato	72.19015
Superficie	26.67070
Dist estacion	53.14239
Dist mall	54.26497

Anexo 7. Resultados modelos de regresión

Precio de vivienda

	Dependent variable:		
	log_price		
	(OLS)	(OLS2)	(OLS3)
l3Medellín	-0.284***		-0.072
	(0.005)		(0.049)
bedrooms	0.043***		0.039***
	(0.002)		(0.002)
property_typeCasa	0.098***		0.100***
	(0.005)		(0.005)
iluminado1	0.0002		0.0005
	(0.008)		(0.008)
banos_new	0.279***		0.274***
	(0.002)		(0.002)
estrato_new2	-0.267***		-0.252***
	(0.022)		(0.026)
estrato_new3	0.017		0.067***
	(0.022)		(0.026)
estrato_new4	0.321***		0.402***
	(0.022)		(0.026)
estrato_new5	0.547***		0.636***
	(0.022)		(0.026)
estrato_new6	0.885***		0.951***
	(0.022)		(0.026)
surface_new	0.0001***		0.0003***
	(0.00000)		(0.00001)
dist_estacion	-0.00002***		-0.00002***
	(0.00000)		(0.00000)
distancia_mall	0.00002***		0.00003***
	(0.00000)		(0.00000)
I(banos_new2)		0.059***	
		(0.0003)	

I(bedrooms2)		-0.007***	
		(0.0002)	
l3Medellín:bedrooms			0.066***
			(0.005)
l3Medellín:property_typeCasa			-0.039***
			(0.012)
l3Medellín:iluminado1			-0.036
			(0.025)
l3Medellín:banos_new			-0.008
			(0.005)
l3Medellín:estrato_new2			0.134***
			(0.049)
l3Medellín:estrato_new3			-0.192***
			(0.047)
l3Medellín:estrato_new4			-0.351***
			(0.047)
l3Medellín:estrato_new5			-0.397***
			(0.047)
l3Medellín:estrato_new6			-0.569***
			(0.052)
l3Medellín:surface_new			-0.0003***
			(0.00001)
l3Medellín:dist_estacion			0.00004***
			(0.00001)
l3Medellín:distancia_mall			-0.0001***
			(0.00001)
Constant	18.776***	19.578***	18.705***
	(0.022)	(0.003)	(0.026)
Observations	72,142	72,142	72,142
R2	0.621	0.326	0.630
Adjusted R2	0.621	0.326	0.630
Residual Std. Error	0.455 (df = 72128)	0.607 (df = 72139)	0.450 (df = 72116)
F Statistic	9,098.241*** (df = 13; 72128)	17,475.810*** (df = 2; 72139)	4,918.346*** (df = 25; 72116)
Note:			*p<0.1; **p<0.05; ***p<0.01

Anexo 9. Comparación de modelos

Modelo	Regression Tree	Random forest	OLS (todas las variables)	OLS (predictor no lineal)	OLS (interacciones)
MSE	0.2371623	0.119052	0.2090798	0.3634746	0.2025432