Problem Set 3: Making Money with ML?

Paula Ramos, Karen Uribe-Chaves y Juan D. Urquijo

July 26, 2022
Repository Link:Github

1 Introduction

El precio de la vivienda está fuertemente correlacionado con otros factores como la ubicación, el área, la población, y la distancia a puntos de interés del comprador. Este trabajo explora la predicción del precio de viviendas ubicadas en la localidad de Chapinero en Bogotá y en el barrio El Poblado en Medellín, Colombia. Al realizar el análisis de modelos para cada ubicación, se encuentra que la mejor predicción es la realizada por OLS para Chapinero y XGBoost para El Poblado.

En particular, los modelos de regresión *OLS* describen una relación lineal entre la variable dependiente y las variables predictoras independientes. Este tipo de modelos son los de menor complejidad y menor carga computacional a la hora de predecir una variable. Después de probar varios modelos más complejos para la localidad de Chapinero, el modelo de regresión lineal OLS obtuvo el mejor desempeño en el valor promedio por unidad comprada. Esto nos da a entender que la clave para predecir el precio de la vivienda, teniendo en cuenta este parámetro en Chapinero no está del lado de proponer modelos complejos y estructurados, sino del lado de entender bien la dinámica que se quiere explicar escogiendo de mejor manera sus predictores.

Entre las desventajas de utilizar este tipo de modelos está su sensibilidad ante valores atípicos y la necesidad de muestras con un número de observaciones considerable para que sus estimaciones sean robustas, además, necesitan que la muestra de datos utilizados tengan un comportamiento lineal. Se espera entonces que las predicciones de este modelo sean acertadas según nuestro criterio a la hora de escoger las variables explicativas, así como el trabajo de análisis y la transformación de los datos.

Por su parte *XGBoost* es un algoritmo predictivo supervisado que utiliza el principio de *boosting*. A partir del procesamiento en paralelo, realiza la poda de árboles, y la regularización (penaliza la complejidad de los modelos) para evitar el *overfitting*. Sin embargo, no funciona tan bien con datos dispersos y no estructurados, así como en presencia de valores atípicos lo que podría reducir la precisión de sus predicciones. Se espera que su aplicación en El Poblado, logre una precisión alta debido al trabajo de limpieza de datos que se realizó y el análisis de las variables independientes escogidas, y a que el modelo funciona muy bien en datos medianos y pequeños con subgrupos y conjuntos de datos estructurados. Lo anterior se ve reflejado en un mejor desempeño en el valor promedio por unidad comprada, así como en los menores errores de predicción obtenidos en la base de entrenamiento comparados con otros modelos.

2 Data

Autores como Lancaster(1966), Rosen (1974) y Limsombunchai (2004), definen que los atributos de una vivienda (el número de habitaciones, el número de baños, el número de chimeneas, el aparcamiento, la superficie habitable y el tamaño del terreno) se incorporan implícitamente a los bienes y a sus precios de mercado observados, lo que se conoce como la teoría de los precios hedónicos. Basados en esta teoría, para predecir el precio de las viviendas, se utilizaron las siguientes fuentes de datos y variables:

- *Properti*:La base principal procede de *www.properti.com*. Esta contiene algunas características de los inmuebles como: *área total*, *tipo de propiedad*, *cantidad de habitaciones*, *dormitorios* y *baños*, así como la variable a predecir *precio de venta*. Adicionalmente, se cuenta con una columna de *descripción*, un texto breve con características adicionales del inmueble. A partir de esta última se crearon dos variables: *(1)Piso* bajo la intuición de que el precio de las viviendas se puede ver afectado por: qué tan alto se encuentra un apartamento, o cuántos pisos tiene una casa; *(2) Estrato* considerando que el estrato socioeconómico también se relaciona con la valoración de los inmuebles, ya que está asociado al barrio, el valor de los servicios públicos, entre otros aspectos.

- *Open Street Map*: Una vez definidos los polígonos del Poblado (Medellín) y Chapinero (Bogotá), se buscan las características disponibles en *Open Street Maps Features* y se complementa la base de datos agregando la distancia mínima a una *estación de policía*, un *bar*, un *parque*, una *estación de bus* y un *banco* para cada propiedad, tomando en consideración temas como la percepción de seguridad, recreación, movilidad y acceso a servicios financieros.

Dado que habían missing values, en la limpieza de datos de las bases de entrenamiento y prueba se realizó imputación de datos a través de: la búsqueda de texto (Ej: Superficie), la mediana a nivel de manzana de los datos originales (Ej: Piso) y la información del DANE a nivel de manzana censal (Ej: Estrato y número de habitaciones). Cabe resaltar que, en los casos que los datos no se lograron imputar o no coincidieron con el criterio experto (Ej: Pisos muy altos, superficie errada, etc), se eliminaron esas observaciones.

Respecto a las estadísticas descriptivas de la base de entrenamiento (Table 1), es evidente que el promedio del precio en Chapinero ($COP 1,000 millones) es considerablemente alto en relación con el Poblado ($COP 380 millones). En general, todas las distancias calculadas son más cortas en Chapinero, la diferencia más alta entre ambos lugares es la distancia a una estación de policía, a favor de Chapinero. El número de habitaciones promedio es igual (3) en ambas zonas y las viviendas son de estratos altos, el 91% de las viviendas en Chapinero son estratos 5 y 6, mientras que en el Poblado, el 84% son estratos 4, 5 y 6. En ambos lugares predominan los apartamentos, especialmente en Chapinero que representan el 95%.

Table 1: Estadísticas descrptivas - Base Entrenamiento

Characteristic	**Chapinero**, N = 11,886	**Poblado**, N = 1,409
Precio	1,000,000,000 (630,000,000, 1,680,000,000)	380,000,000 (265,000,000, 700,000,000)
Piso	4.00 (4.00, 5.00)	3.00 (2.00, 4.00)
Cuartos	3.00 (3.00, 3.00)	3.00 (3.00, 3.00)
Superficie Total	147 (105, 192)	101 (70, 146)
Distancia Bares	505 (333, 760)	584 (444, 686)
Distancia Parques	105 (55, 168)	335 (258, 550)
Distancia Bancos	269 (152, 421)	335 (258, 550)
Distancia Estaciones Bus	766 (436, 1,232)	868 (338, 1,195)
Distancia Policía	431 (278, 594)	1,889 (1,316, 2,163)
Baños	3.00 (2.00, 4.00)	2.00 (2.00, 3.00)
Estrato		
1	11 (¡0.1%)	0 (0%)
2	50 (0.4%)	49 (3.5%)
3	146 (1.2%)	173 (12%)
4	760 (6.4%)	368 (26%)
5	881 (7.4%)	162 (11%)
6	10,038 (84%)	657 (47%)
Tipo de Propiedad		
Apartamento	11,288 (95%)	1,019 (72%)
Casa	598 (5.0%)	390 (28%)

[1] Median (IQR); n (%)

Adicionalmente, para complementar el análisis se graficaron los mapas de ambos lugares. En la Figure 1 de la localidad de Chapinero (área de 38 km2), se evidencia que las viviendas ubicas en el norte y hacia los cerros orientales son más costosas (Puntos de color azul oscuro), cuando se avanza hacia el centro el costo disminuye. Los bares se encuentran ubicados cerca a las estaciones de bus, en zonas donde la cantidad de viviendas disminuye. En la Figure 2, el mapa del Poblado (área de 23 km2), se puede observar que las viviendas de más superficie (Puntos rojo oscuro) se ubican hacia el este de la comunidad, adicionalmente, las estaciones de bus (Puntos de color negro) se encuentran al norte y son pocas en relación a la cantidad de viviendas. Los bares (Puntos color azul) se encuentran concentrados en un área de la comuna.

Figure 1: Mapa Localidad Chapinero - Bogotá

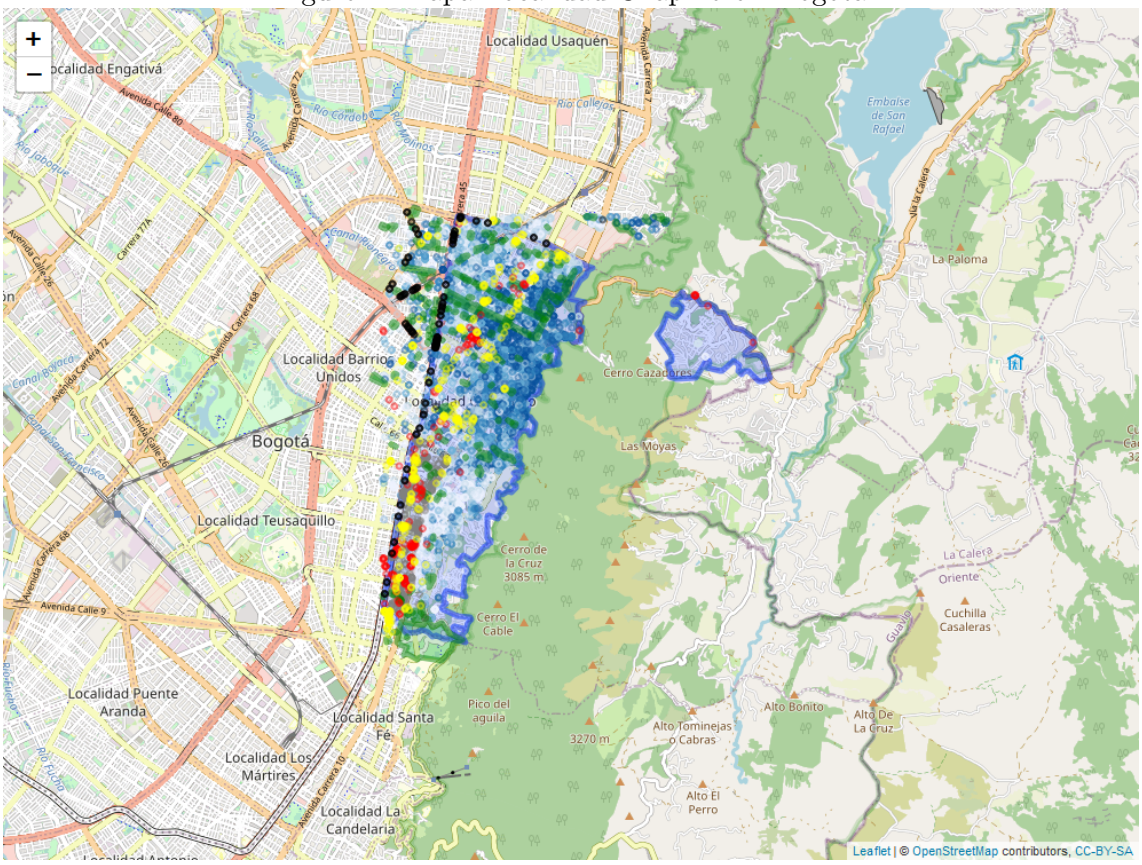
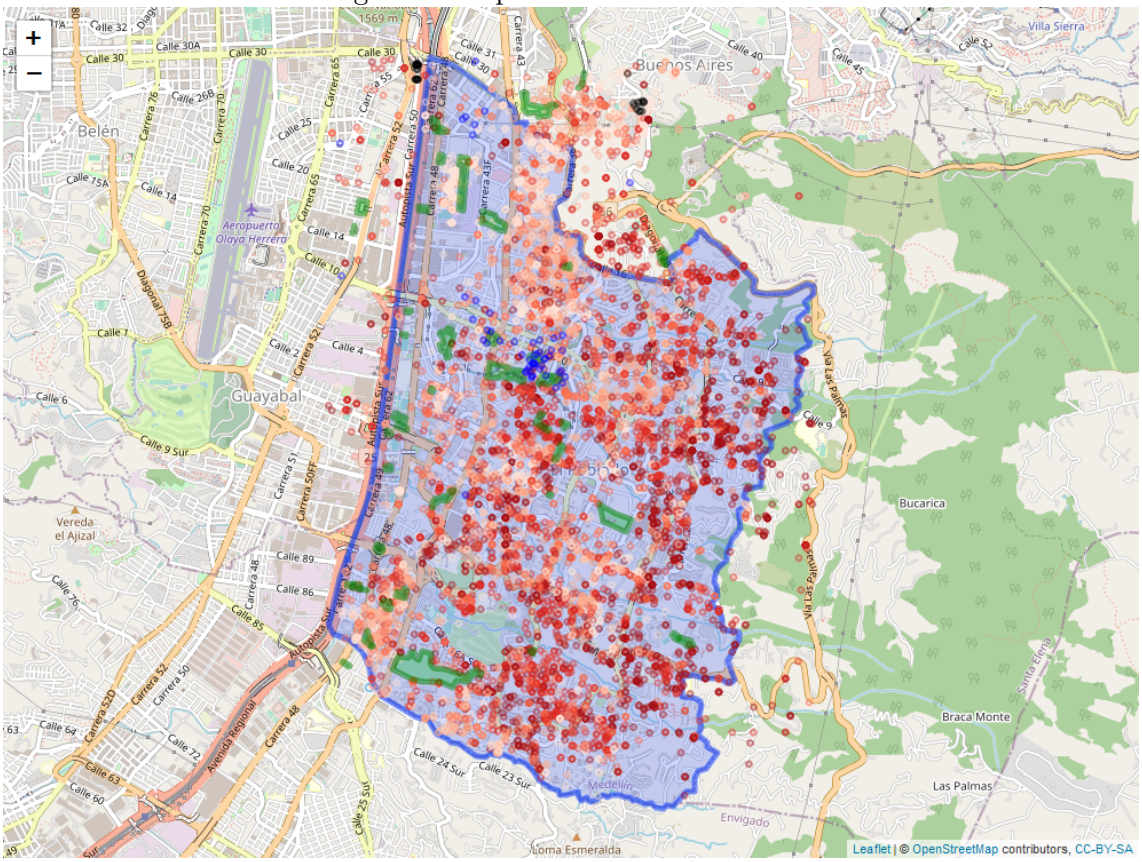


Figure 2: Mapa El Poblado - Medellín



3 Model and Results

Se decidió realizar predicciones separadas para cada área (Chapinero y El Poblado) dado que consideramos que pueden existir diferentes factores en cada una de ellas que afectan el precio de la vivienda. Esto se verá reflejado en las variables con mayor importancia en cada uno de los modelos que se presentan a continuación.

Con el fin de realizar las predicciones se utilizaron los modelos: OLS, Ridge, Lasso, Árbol, Bagging y XGBoost como entrenamiento. Las variables utilizadas en los modelos, se explican en la sección 2 de este trabajo. Es importante mencionar que se escaló la variable dependiente de "Precio" en cada modelo con el fin de tener mejores distribuciones (i.e Transformación OLS en la Figura A.1)

$$Price = Piso + Estrato + Cuartos + Baños + SuperficieTotal + DistanciaBares + DistanciaParques + DistanciaBancos + DistanciaEstaciónBus + DistanciaPolicía + u$$

La comparación de los modelos se observa en la Table A.1 y A.2 del anexo de este trabajo. El criterio de selección del modelo se centró en darle mayor relevancia a aquel modelo que tenía la menor relación de gasto promedio por propiedad comprada en cada localidad. Sin embargo, también se analizó el MAE y el RMSE para cada modelo predictivo. Los resultados para cada localidad se presentan a continuación:

3.1 *OLS Chapinero*

Después de validar los modelos mencionados, se encontró que para Chapinero el mejor modelo es el *OLS* (Ver resultados: A.3). Nuestras métrica de interés son el R^2, el MAE, RMSE y el ratio entre gasto total y número de propiedades compradas.

- *Variables*: Se encuentra que las variables relevantes son el Piso, el número de cuartos, la superficie total, la distancia a parques, bancos, estación de buses, policía, Número de baños y el tipo de propiedad.
- *Ratio*: Para la localidad de Chapinero el mejor modelo predictivo es el OLS, con un gasto promedio por vivienda comprada de $1.186 millones como resultado de la compra de 7.077 viviendas. Este ratio mejora la media del precio de la base de entrenamiento de Chapinero que alcanza $1,286.8 millones.
- R^2: El modelo tiene una bondad de ajuste de 0.540, lo que nos indica que el modelo tiene un poder de explicación medio en relación con la variabilidad de los datos.
- *MAE (Mean Absolut Error) y RMSE (Root Mean Squared Error)*: El MAE del modelo es de $426.3 millones, lo que se considera alto, sin embargo es el menor después de XGboost de los modelos aplicados. El RMSE del modelo fue de $648.7 millones, siendo mayor que el MAE lo que nos indica una alta varianza en los errores individuales de la muestra.

3.2 *XGBoost El Poblado*

En el análisis del mejor modelo para El Poblado - Medellín, encontramos que bajo el criterio de selección del gasto promedio, el mejor modelo predictivo es el *XGBoost (Extreme Gradient Boosting)*. De igual forma, este modelo es el que arroja mejores resultados en el MAE y el RMSE (Ver A.2).

Este modelo se estimó utilizando un rango de rondas 250 a 500. Una profundidad del árbol que ajusta el modelo de 4, 6 y 8 nodos. Un rango de la tasa de aprendizaje del modelo entre 0.01, 0.3 y 0.5. Un rango de observaciones en la región final del árbol de entre 10, 25 y 50. Dando como resultado:

- *Hiper-parámetros del mejor árbol estimado*: Tasa de aprendizaje del modelo de 0.01; profundidad del árbol 8 nodos; penalización por particiones del árbol 1; porcentaje de subsampleo de columnas para el árbol 0.7; número de observaciones en la región final del árbol de 10; y porcentaje de subsampleo de observaciones para el árbol 0.6. El número de rondas de árboles fue 500.
- *Importancia Variables*: En el árbol estimado, se encuentra que las variables más importantes según la estimación son: i) Superficie Total, ii) Distancia Bares, iii) Distancia Estación de Buses y iv) Número de Baños. El ranking completo se encuentra en la Figura A.2 del anexo de este trabajo.
- *Ratio*: Para El Poblado el mejor modelo predictivo es el *XGBoost*, con un gasto promedio por vivienda comprada de $COP 535.3 millones como resultado de la compra de 1.023 viviendas. Este ratio mejora la media del precio de la base de entrenamiento de El Poblado que alcanza $646 millones.

- *MAE (Mean Absolut Error) y RMSE (Root Mean Squared Error)*: Se tuvo un RMSE mínimo de $358,691,318 y un MAE de $177,684,837.

4 Conclusions and Recommendations

Este artículo investiga diferentes modelos para la predicción del precio de la vivienda en Chapinero, Bogotá y El Poblado, Medellín. Se exploran 6 tipos diferentes de métodos de Machine Learning que incluyen OLS, Ridge, Lasso, Árboles, Bagging y XGBoost. El método XGBoost tiene el error más bajo en el conjunto de entrenamiento para ambas bases, pero es propenso a sobreajustarse, por tanto el criterio de selección es el ratio de valor promedio por vivienda comprada. El resultado de las estimaciones arrojó que los mejores modelos según este criterio de selección son OLS para Chapinero y XGBoost para El Poblado.

Las variables y su importancia en cada modelo, brindan evidencia a favor de realizar dos estimaciones por aparte, dado que la relevancia varía en cada área analizada. A manera de ilustración, la variable piso es de alta relevancia para Chapinero, contrario a El Poblado. Sin embargo, el Estrato es relevante en el Poblado, mientras que para Chapinero no es significativa, intuitivamente por la concentración en los estratos 5 y 6 de la muestra. Las predicciones finales sobre la base de prueba, arrojan la compra de un total de 9,458 propiedades entre ambas áreas (Chapinero: 506 propiedades y El Poblado: 8,953) por un valor total de $9,288,170 millones (Chapinero: $249,432 millones y El Poblado: $9,038,737 millones).

Para futuros trabajos, se recomienda la aplicación de modelos de Deep Learning y su combinación con Machine Learning que permitan ampliar las estimaciones, así como el uso de técnicas de comparación que permitan escoger soluciones óptimas. Además sería interesante integrar en los modelos de predicción variables de percepción de los ciudadanos en relación a su entorno a nivel de manzana,que permitan tener una medida de satisfacción de las personas asociadas a la seguridad, costo de vida, ruido, entre otros; sin embargo, esta última idea requeriría contar con bases de datos robustas y costosas para los tomadores de decisiones.

5 References

- Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy, 82(1), 34-55. https://doi.org/10.1086/260169
- Lancaster, K. J. (1966). A new approach to consumer theory. Journal of Political Economy, 74
- Limsombunchai (2004) House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. American Journal of Applied Sciences 1 (3): 193-201, 2004

A *Appendix*

1. Resultados comparación de Modelos

Table A.1: Resultados Modelos Chapinero

	MAE	RMSE	Gasto	Compras	Ratio
OLS	$426.3	$648.7	$8,390,124.9	7,077.0	$1,185.5
Ridge	$450.7	$645.7	$9,573,808.4	7,893.0	$1,212.9
Lasso	$448.7	$645.9	$9,533,505.5	7,770.0	$1,227.0
Árbol	$445.8	$645.6	$9,439,211.4	7,488.0	$1,260.6
Bagging	$442.3	$639.9	$9,482,381.3	7,639.0	$1,241.3
XGBoost	$291.6	$486.1	$9,682,280.3	7,971.0	$1,214.7

Note: Cifras en millones COP. Ratio calculado como Gasto total sobre No.de propiedades compradas.

Table A.2: Resultados Modelos El Poblado

	MAE	RMSE	Gasto	Compras	Ratio
OLS	$227.2	$423.5	$511,011.1	905.0	$564.7
Ridge	$263.9	$447.3	$572,527.2	1,026.0	$558.0
Lasso	$259.5	$446.9	$592,791.8	1,051.0	$564.0
Árbol	$219.2	$390.6	$542,145.3	907.0	$597.7
Bagging	$220.9	$399.1	$561,354.7	937.0	$599.1
XGBoost	$177.7	$358.7	$535,289.4	1,023.0	$523.3

Note: Cifras en millones COP. Ratio calculado como Gasto total sobre No.de propiedades compradas.

Figure A.1: Distribución $\sqrt{Price}$ Viviendas - Localidad Chapinero

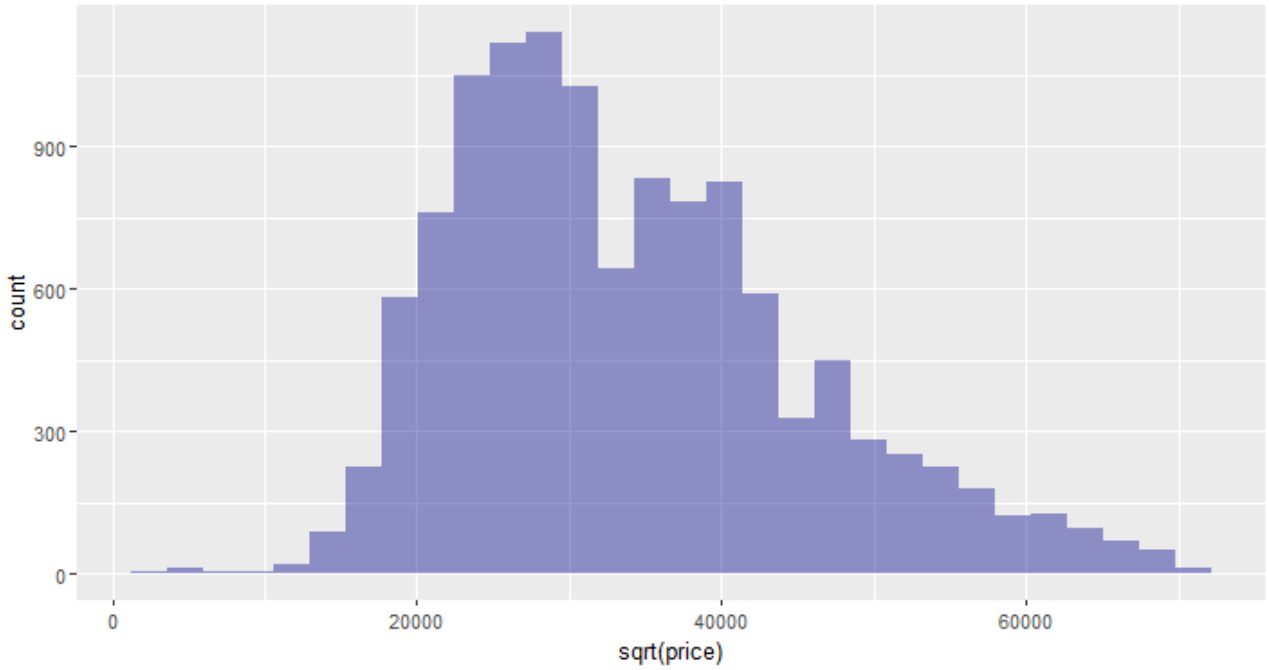


Table A.3: Resultados Modelo OLS - Chapinero

	Dependent variable:
	$\sqrt{Price}$
Piso	180.519***
	(42.949)
Estrato 2	−1,139.912
	(2,613.553)
Estrato 3	−3,035.611
	(2,439.000)
Estrato 4	−2,505.624
	(2,370.808)
Estrato 5	881.175
	(2,368.941)
Estrato 6	352.653
	(2,353.345)
No. de Cuartos	296.542***
	(40.380)
Superficie Total (m^2)	36.944***
	(0.922)
Distancia Bares	−0.377
	(0.295)
Distancia Parques	−11.957***
	(0.713)
Distancia Banco	−0.947**
	(0.435)
Distancia Estación de Bus	6.564***
	(0.193)
Distancia Policía	3.206***
	(0.358)
No. Baños	4,518.972***
	(77.424)
Apartamento = 1	962.136***
	(344.911)
Constante	6,849.048***
	(2,400.017)
Observations	11,886
R^2	0.540
Adjusted R^2	0.540
Residual Std. Error	7,782.644 (df = 11870)
F Statistic	929.521*** (df = 15; 11870)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure A.2: Importancia de las Variables XGBoost - El Poblado