Modelos de Predicción de Precios de Vivienda - Machine Learning

*Repositorio: https://github.com/MiguelContreras1/Taller-3

1st Jesús Cepeda
Facultad de Economía
Universidad de los Andes
Bogotá, Colombia

2nd Juan Camilo Martínez
Facultad de Economía
Universidad de los Andes
Bogotá, Colombia

3rd Arturo Trujillo
Facultad de Economía
Universidad de los Andes
Bogotá, Colombia

4th Miguel Contreras
Facultad de Economía
Universidad de los Andes
Bogotá, Colombia

***Abstract*—Este documento muestra los resultados de una serie de modelos de Machine Lerning aplicados al sector inmobiliario de las tres ciudades màs pobladas de Colombia. Estos buscan predecir los precios de las viviendas de la manera más aceretada posible teniendo.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduccion

Una de las muchas aplicaciones del machine learning se presenta en el sector inmobiliario, donde pueden resultar muy útiles los modelos para la predicción de precios de viviendas. En este trabajo se busca predecir precios de este tipo de inmuebles en Cali. Hay que recordar que estos inmuebles son bienes diferenciados, los cuales se ofertan y demandan en un mercado competitivo. De acuerdo con la teoría de precios hedónicos, dicho precio dependerá de atributos relevantes, es decir, características propias observables inherentes del inmueble, pero también hay otras características que no son propias de las viviendas, y que influyen en su precio, como las condiciones del entorno, las cuales deben tenerse en cuenta al elaborar modelos con el fin de mejorar la predicción de éstos. Teniendo esto en cuenta, a la información de las bases de datos proporcionadas (train y test), se le agregó información espacial, para construir variables adicionales.

Como no se conoce la forma funcional exacta que determina los precios, resultan útiles los modelos que capturen no linealidades y puedan encontrar la mejor forma funcional, es decir, el patrón subyacente que mejora la predicción. Este documento expone los resultados del modelo evaluado (Xgboots, Random Forest, Riged, Lasso y Elastinet). Tras el proceso de depuración y entrenamiento de los modelos se evidenció que el Random Forest logrando tener un RMSE.

II. Datos

Se utilizaron datos provenientes de las bases de datos train (51.437 observaciones y 12 variables, de Bogotá y Medellín) y test (5.000 observaciones y 12 variables, de Cali). De estas se tomaron las variables de características inherentes a la vivienda, como los metros cuadrados de superficie, el número de habitaciones, el número de baños, el tipo de propiedad y la ciudad. Además, se usaron datos de fuentes externas para la construcción de variables predictoras relevantes, por ejemplo, la cercanía a bienes públicos, como parques, paraderos de buses, centros de salud y estaciones de policía.

Tras la exploración inicial se pudo determinar que 37.985 viviendas están ubicadas en Bogotá, 13.452 en Medellín y 5.000 en Cali. En cuanto a los datos faltantes en la base de datos train se encontraron 39.044 en la variable surface total, 41.745 en surface covered, 24.915 en rooms, y 15.032 en bathrooms. Por su parte, en la base de datos test los missing values fueron 3.146 en surface total, 3.569 en surface covered, 1.837 en rooms, y 1.478 en bathrooms. Para recuperar información de estas variables relevantes se efectuó una captura de información la variable description en ambas bases.

A grandes rasgos se puede observar que, las viviendas de Bogotá tienen precios entre $380y990$, en su mayoría son apartamentos (77 %), tienen una media de 3 habitaciones y 3 baños. En Medellín los precios van desde $320 hasta $730. La mayoría son apartamentos (80 %), la media de habitaciones y baños es la misma que en Bogotá. Por su parte, en Cali se pudo observar que, el porcentaje de apartamentos si bien es mayoría (56 %), es menor respecto a las otras dos ciudades, mientras que la media de habitaciones (3) y de baños (3), es la misma.

En lo referente a la construcción de variables que indican cercanía de los inmuebles a bienes públicos que pueden incidir en su precio, se utilizó el Sistemas de Coordenadas Geográficas, crs=4326. Gracias a que se contaba con información de latitud y longitud de las casas y apartamentos, fue posible crear las variables distancia parque, distancia estacion bus, distastancia hospital, distancia estación policia.

A continuación, se ve un resumen con las variables seleccionadas para el modelo.

III. Cotexto de Cali

La ciudad de Cali, es uno de los principales centros económicos e industriales de Colombia, además de ser el principal centro urbano, cultural, económico, industrial y agrario del suroccidente del país y el tercero a nivel nacional después

de Bogotá y Medellín. Tiene 2.545.682 habitantes en una desnidad de 4382,05 hab/km².

Segun empresas como habi, el valor de la mediana del metro cuadrado en Cali es $2.983.293 COP. Este es el menor valor del metro cuadrado en una ciudad principal de Colombia, si la comparamos con Bogotá, Medellín y Barranquilla.

Cali también es la ciudad principal en la que menos se han valorizado los predios. De acuerdo con el último informe de BBVA, sobre la situación inmobiliaria de Colombia, entre 2010 y 2020, Cali tuvo una valorización predial del 64,7%. Esto pone a la capital del Valle del Cauca por debajo de la media del país, que es 69,1%.

Es probable que por el valor de la mediana del metro cuadrado en la Sultana del Valle, el precio de compra y venta de las viviendas sea más económico que en Bogotá, Medellín y Barranquilla. Esto puede significar que, mientras en Bogotá, donde la mediana del valor del metro cuadrado es $4.166.666 COP, una casa de 60 m² puede valer aproximadamente $249.999.960 COP; en Cali una vivienda con el mismo metraje puede tener un valor aproximado de $178.997.580 COP. Ten en cuenta que estos valores son una referencia, por lo que podrías llegar a encontrar una casa con las mismas características, a un precio de venta mayor o menor. (HABI, 2022)

la zona urbana del Distrito de Santiago de Cali cuenta con una distribución espacial de las principales variables socioeconómicas marcada por la presencia de tres franjas gruesas en el sentido sur norte de la ciudad. En el eje central, en el cual se encuentran las comunas 2, 19 y 17 priorizadas para en el presente caso de estudio, cuenta con una concentración de mejores indicadores de calidad de vida; mientras que, en las franjas perimetrales de la ladera occidental y la zona oriental de la ciudad, se encuentran las zonas con los mayores desafíos en términos sociales y económicos.

IV. Teoría de Precios Hedónicos

La Teoría de Precios Hedónicos pretende explicar el valor de un bien raíz, entendido como un conjunto de atributos (superficie, aptitud de uso del suelo, calidad de la construcción, diseño interior y exterior, áreas verdes, ubicación, características del vecindario, etc.), en función de cada uno de ellos. (Wikipedia, 2022)

Como ejemplo, el caso más común dentro del método de los precios hedónicos (MPH) es el mercado de las propiedades residenciales. Entonces, en este mercado no sólo se transan las casas en juego, sino también, "implícitamente" se encuentra el mercado del paisaje del entorno de las casas, el mercado del clima de la zona residencial, el mercado de la ubicación residencial, etc. Por lo que siendo un poco más específico, el mercado implícito de refiere al mercado de las característica de los bienes transados. (Wikipedia, 2022)

Por otro lado, dentro de esta teoría se maneja el concepto de "amenidad" que se refiere al conjunto de características que potencia el valor de las características de los bienes en cuestión, en lenguaje sencillo, las buenas características. Inversamente, las "desamenidades" son las que afectan negativamente el valor de estos bienes. (Wikipedia, 2022)

V. MODELOS

Tras el tratamiento de los datos se estableció que el modelo base para determinar el precio tendría la siguiente estructura la cual tiene enfoque pseudohedónico con variables agregadas como distancia a estaciones de policias, hospitales, parques, numero de cuartos de la vivienda, munero de baños, los vecinos del sector entre otra.

$\text{Precio}_i = \beta_0 + \beta_1 * Surfacetotal2_i + \beta_2 * bedrooms_i + \beta_3 * bathrooms_i + \beta_4 * distancia_parque_i + \beta_5 * distancia_estacion_bus_i + \beta_6 * distancia_hospital_i + \beta_7 * distancia_estacion_policia_i$

A. Modelos de Predicion

Modelo Lasso El Método Lasso (Least Absolute Shrinkage and Selection Operator), introducido por Tibs- hirani (1996) es un método que combina un modelo de regresión con un procedimiento de con- tracción de algunos parámetros hacia cero y selección de variables, imponiendo una restricción o una penalización sobre los coeficientes.

Para cada modelo con la base de train completa, mostrado en los modelos anteriores, se realizaron los modelos de Ridge y Lasso, se obtuvo el MSE de las predicciones sobre training para este modelo contramos que los coeficientes que deja el modelo de lasso para el explicar el precio de la viiviend son los numero de cuartos con un 0,058, el numero de baños con un 0,1 la distancia a la parques 0,33 disttancia a estaciones de bus 0,67y distancia a la policta 0,68.

El modelo lasso nos arroja que elas personas en cali prefieren estar mas cerca de una estaciones de polica es decir pgar mas por tener seguridad y en segundo lugar por estar cerca a estaciones de buses por encia de estar la casa tiene un mayor numero de cuarto o baños. es decir seguridad y movidad son el primer criterio de deccion definir el precio de una vivienda.

B. Modelos de Clasificacion

Modelo Random Forest El algoritmo Random Forest (Breiman, 2001) es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado.

Posteriormente se corrieron los modelos descritos anteriormente, y el de mejor desempeño fue el modelo Random Forest, el cual logró tener el MSE mas bajo de todos los modelos. Este tipo de modelos es una implementación diseñada específicamente para un rendimiento óptimo y velocidad. Es de resaltar que las variables que tienen mayor peso a la hora de predecir el precio, son la superficie, el número de habitaciones y el número de baños.

Podemos notar que la especificidad del modelo de Randon Forest es mayor a la sensibilidad de los otros modelos. Esto quiere decir que el modelo tuvo una mayor capacidad de

identficar las variables que determinan el precio de la viendad Esto por fuera de muestra.

VI. RESULTADOS

Encontramos que en los modelos de clasificacion el mejor fue el Random Forest y las variables que mas tienen peso a la hora predecir el precio de las viviendas en fueron primer la superficie total de la vivenda, segunedo la numero de baños de la vivienda, tercero la distancia a hospitales, distancia a hopitales, cuarto la distiancia a estaciones de buses y el numero de cuartos.

Para el caso de los modelos de clasificaciones como Riged lasso y Elastinect encontrmos que las variables que mas explican el modelon son el numero de cuartos, nuermo de baños, distancia a parques, distancia a estaciones de bus al igual que los modelos de clasificaciones las vaiarbles los modelos coicindes y terminadan arrojando las mismas variables pero unos como menos o mayor niveles de error.

VII. CONCLUSIONES

Del ejercicio realizado se puede concluir que, las aplicaciones del machine learning son un instrumento muy útil en el campo inmobiliario. En especial, los modelos de regresión para predicción de precios de inmuebles, siempre que sean suficientemente precisos y tomen en cuenta variables relevantes, resultan particularmente útiles para lograr mayores rentabilidades. También hay que resaltar que la teoría de los precios hedónicos se cumple en este ejercicio, pues las variables del entorno como las distancias a estaciones de policta, buses o parques resultaron relevantes a la hora de predecir los precios de las viviendas.

REFERENCES

1* Datos de precios de Vivenda suminstradas por La Univerdiad de los Andes en el curso de Big Data.

2* Datos de carateristicas de Vivienda bajados por medio de la plataforma OpenStreetMap https://www.openstreetmap.org/map=5/4.632/-74.299

3* (HABI, 2022), https://habi.co/blog/cuanto-cuesta-el-metro-cuadrado-en-cali.

4* (Wikipedia, 2022), "Teoria de los precios hedonicos" https://www.google.com/search?sxsrf=ALiCzsaPUWv1d8FBoJQ

5° (Wikipedia, 2022), "Cidad de Santago de Cali" https://es.wikipedia.org/wiki/Cali.

	Coeficiente	Error Estandar	t	P Valor
Intercepto	0,0263	48,2050	33,0151	3,2844
bedrooms	0,0581	50,9077	37,6230	3,7574
bathrooms	0,1013	57,2363	39,2339	3,9959
distancia$_{p}arques$	0,3399	91,2507	39,2775	4,5896
distancia$_{b}us$	0,6022	128,2213	37,3970	4,8775
distancia$_{h}ospitales$	0,6754	129,9099	36,2468	4,8775
distancia$_{e}stacion_{p}olicia$	0,6847	134,3197	35,2439	4,8775

TABLE I
COFICIENTES DE LA REGRESION POR MCO CON PENALISATION LASSO

Modelo	RMSE	MEA
Random Forest	230523754	100805209
XGBoost	478135094	263805465
Lasso	496297766	268316530
Ridge	497961023	268854796
Elastic Net	498527651	276918658

TABLE II
DESEMPEÑO DE LOS MODELOS

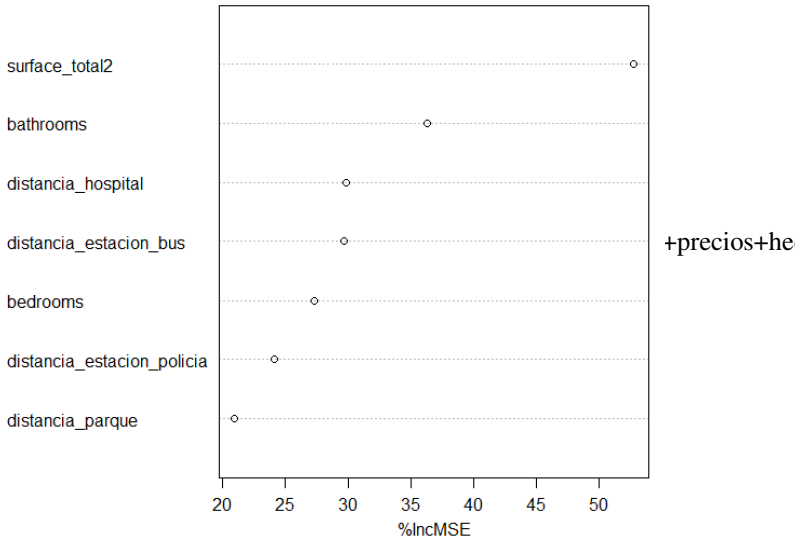

Fig. 1. Importancia variables