¿Ganar dinero con Machine Learning?

"Todo es cuestión de ubicación, ubicación, ubicación!!!"

Diana Yarith Higuera Cogua

Juan Sebastian Vallejo Triana

Lizbeth Karina Hernandez Noriega

Riky Andrés Carrillo Cadena

Problem Set 3

Universidad de los Andes
Facultad de Economía
Bogotá, Colombia
12 de marzo 2023

Introducción

La diferenciación de productos por medio de vectores de características c1,c2... cn permiten la formación de precios a partir de atributos particularmente ponderados para ese fin (Rosen, 1974); de acuerdo con la teoría de precios hedónicos, la heterogeneidad entre diferentes unidades de un mismo bien se refleja en el precio del mismo.

El dinamismo del mercado inmobiliario permite a los agentes que lo componen ajustar los bienes que comercializan a través de la oferta de productos diferenciados dentro de los segmentos que ofrecen, lo que les permite extraer el mayor beneficio económico posible y mitigar los riesgos del negocio (Uribe, 2022). Tanto compradores como vendedores buscarán maximizar valor, esto es, vender lo más alto que pueda y comprar al menor valor posible.

Bajo ese supuesto de maximización del valor, la Inteligencia Artificial provee herramientas de aprendizaje no supervisado que pueden ser muy útiles para predecir los precios óptimos que una propiedad pueda tener a partir de un vector de características previamente definidas. En este caso, el vector C=(c1,c2,c3,c4,c5,c6), se ha definido como un vector de atributos externos como área del inmueble, número de habitaciones, número de baños, recintos, etc.; y atributos que externos como localización, percepción de seguridad, entre otros. De esa manera se han determinado los siguientes componentes adicionales a las características internas: Balcón, Terraza o Patio, Parqueadero, Parques, Avalúo Catastral, Estaciones de Trasmilenio TM y hurto por UPZ.

Dichos componentes, obedecen a factores que se han tenido al desarrollar políticas de vivienda y que permiten obtener respuestas positivas de parte de los agentes que actúan en el mercado inmobiliario; de acuerdo con Bertaud, 2018, se debe tener en cuenta: el precio por metro cuadrado, el área del inmueble y la localización del mismo, donde está última variable permite incluir dentro de su análisis los atributos adicionales, que la vivienda por

sí sola no cobija y que pueden ser determinantes en la determinación de los precios, cercanía a parques, acceso a medios de transporte, entre otras.

El machine learning a través del aprendizaje no supervisado permite dar solución al problema de los precios hedónicos para el mercado de vivienda, a través de una reducción de las dimensiones de los datos sin perder la máxima variabilidad posible. Sin embargo, hay que tener en cuenta que aunque los métodos se apliquen correctamente, no hay manera de validar la respuesta del algoritmo, dado que no se conoce la respuesta a cada modelación y la segunda consideración a tener en cuenta es que estos algoritmos aprenden del futuro a través de información histórica lo que hace que los modelos puedan sobreestimar o subestimar las predicciones, por lo que a cambios de las tendencias del mercado se debe re-enseñar al algoritmo para que pueda funcionar de manera más acertada a la realidad.

Datos

La información empleada para predecir los precios de las viviendas en el barrio de Chapinero en Bogotá, Colombia, proviene de Properati https://www.properati.com.co que corresponde a un buscador de anuncios clasificados para venta y alquiler de inmuebles que funciona en 5 países de Sudamérica y contiene una muestra de datos para Bogotá.

Para este estudio, la base de datos se descargó de https://www.kaggle.com/competitions/uniandes-bdml-20231-ps3/data y se compone de dos bases de datos: 1) Entrenamiento y 2) Prueba. Las variables que componen las bases de datos corresponde a un identificado de la propiedad, la ciudad de ubicación, el tipo de propiedad, el tipo de operación, el título y la descripción del aviso publicitario, siendo estas variables de tipo factor, mientras que como variables numéricas tenemos el precio de la vivienda, mes y año de publicación de aviso publicitario, la superficie total, la superficie cubierta, el número de habitaciones, dormitorios y baño y las variables de localización que son latitud y longitud.

Limpieza, imputación y manipulación de las bases de datos

Teniendo en que tenemos dos bases, una de entrenamiento y otra de prueba, iniciamos realizando una inspección de los valores perdidos contenida en cada base de datos y que se muestra en el Cuadro 1. Podemos observar que tanto en la base de entrenamiento como de prueba la variable de superficie total y cubierta tiene un porcentaje alto de valores perdido y ronda el 80 % de las observaciones, en el caso de las habitaciones y los baños el porcentaje de valores perdidos ronda el 40 % y 25 % de las observaciones, en el caso de las variables de título y descripción el porcentaje de observaciones perdidas es bajo.

Cuadro 1. Frecuencia y porcentaje de valores perdidos en la base de entrenamiento y prueba

Variable	**Descripción**	**Train**		**Test**	
		Frecuencia	**% de valores perdidos**	**Frecuencia**	**% de valores perdidos**
ID_propiedad	Identificador de la propiedad	0	0 %	0	0 %
Ciudad	Ciudad de ubicación de la vivienda	0	0 %	0	0 %
Precio	Precio de la vivienda	0	0 %	10286	100 %
Mes	Mes de publicación	0	0 %	0	0 %
Año	Año de publicación	0	0 %	0	0 %
Superficie_total	Superficie total del inmueble	30790	80 %	8422	82 %
Superficie_cubierta	Superficie cubierta del inmueble	30790	80 %	7459	73 %
Habitaciones	Número de habitaciones	18260	47 %	4582	45 %
Dormitorios	Dormitorios	0	0 %	0	0 %
Baños	Baños	10071	26 %	2491	24 %
Tipo de vivienda	Tipo de vivienda (Casa-Apartamento)	0	0 %	0	0 %
Tipo de operación	Tipo de operación (Venta- Arriendo)	0	0 %	0	0 %
Latitud	Latitud	0	0 %	0	0 %
Longitud	Longitud	0	0 %	0	0 %
Titulo	Titulo del aviso	25	0 %	6	0 %
Descripción	Descripción del aviso	12	0 %	3	0 %

Tras inspeccionar se hicieron algunas imputaciones como la variable baños que se imputó haciendo uso de expresiones regulares a partir de la variable descripción, se obtuvo el

número de baños que se mencionaban en la descripción y se imputó a la variable baño predeterminada, igualmente se imputó la variable de superficie total a partir de la variable de superficie cubierta.

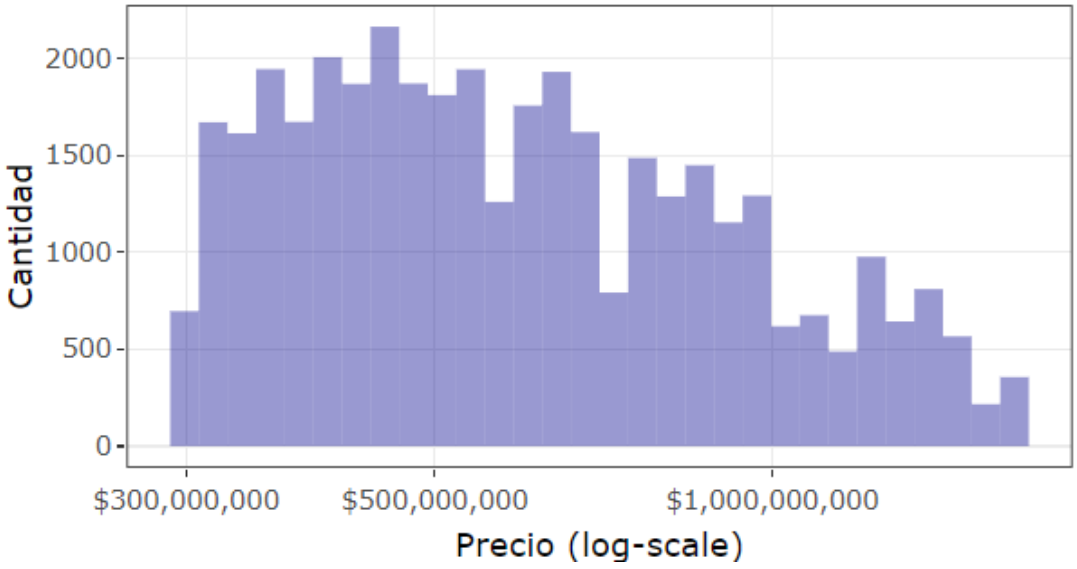
Finalmente, para completar el set de variables necesarias para la estimación de los modelos, acorde con la literatura estudiada, en la base de entrenamiento haciendo uso de la variable de descripción y utilizando la herramienta de expresiones regulares, se construyen dos variables binarias que son: 1) Balcón que es igual a 1 si el inmueble tiene registrado balcón en su descripción y cero de lo contrario, 2) Garaje que es también una variable binaria igual a 1 si el inmueble tiene registrado garaje en su descripción y cero de lo contrario. Estas variables nuevas corresponde a los predictores procedentes del título y la descripción de las propiedades.

Así mismo, se construyeron 6 predictores externos haciendo uso de Open Stree Map para Bogotá, esto permitió construir los predictores de distancia mínima del inmueble a: parques, colegios, hospitales, estación de bus, policía y supermercado. Este mismo proceso se repitió con la base de prueba.

Estadísticas descriptivas

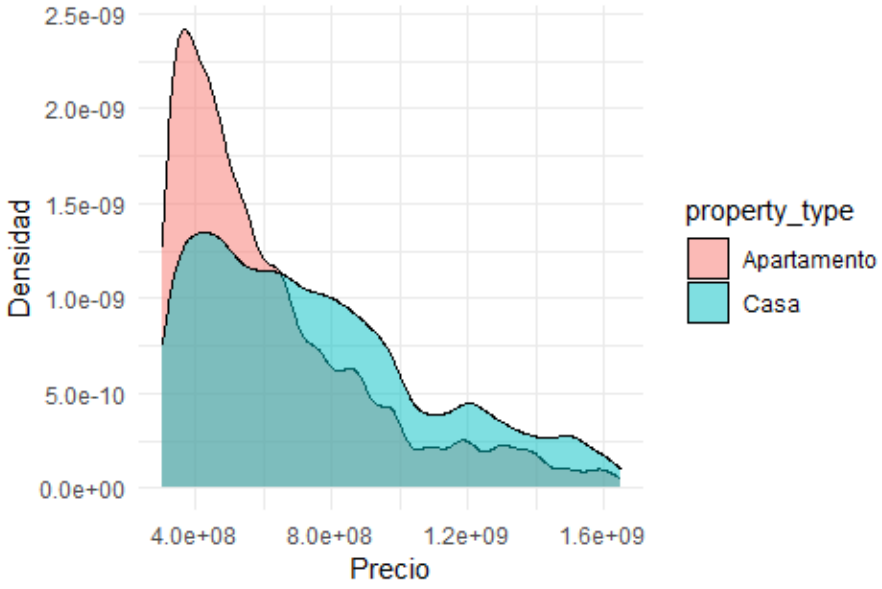
Iniciamos mostrando la distribución de inmuebles para la localidad de Chapinero en Bogotá según los precios de los inmuebles. La Figura 1 muestra que la mayoría de inmuebles de la localidad de chapinero tienen un valor de venta que ronda los 300 a 500 millones de pesos, no obstante, no es despreciable el número de inmuebles que están por encima de los 1000 millones de pesos. En la parte superior encontramos una menor concentración, pero son los inmuebles más costosos y probablemente en la mejor zona de la localidad.

Figura 1. Distribución de los inmuebles en Chapinero según el precio de venta



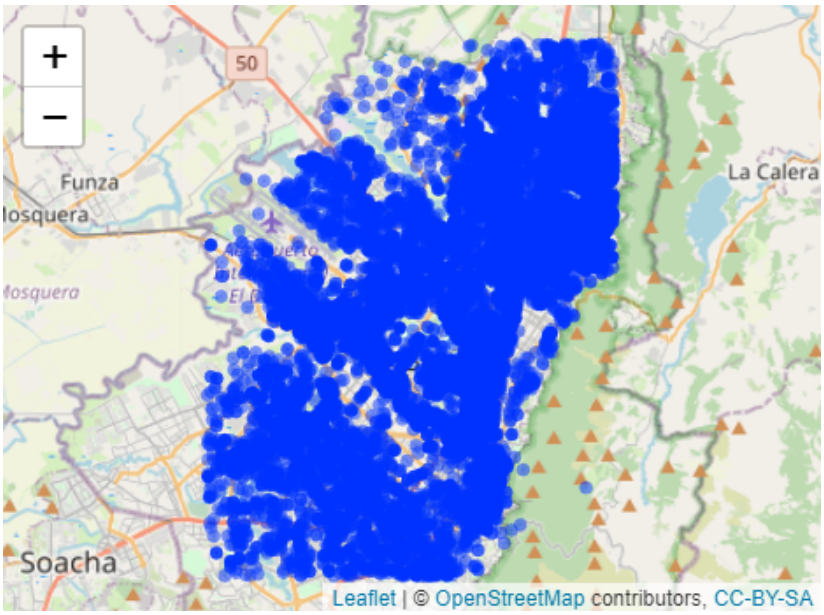
En la Figura 2 observamos la distribución del precio por tipo de inmueble, en este caso, vemos que que los precios de los apartamentos se agrupan en los precios más bajos, mientras que las casas tienen una distribución más alta en los precios medios y altos de los inmuebles en la localidad.

Figura 2. Distribución de precios por tipo de vivienda



Hacemos una visualización de los inmuebles que se encuentran dentro de la ciudad de Bogotá, con el objetivo de comprobar que nuestros datos son acordes para las predicciones que queremos realizar, esto se muestra en la Figura 3.

Figura 3. Mapa visual de inmuebles dentro de los límites de la ciudad de Bogotá



Igualmente, incluimos la visualización de dos mapas en la que se muestran dos de las variables externas que se tuvieron en cuenta para la modelacion de la predicción de los precios de la vivienda, en la Figura 4 y 5, las ubicaciones de parques y estaciones de bus de la ciudad de Bogotá

Figura 4. Mapa visual de inmuebles dentro de los límites de la ciudad de Bogotá

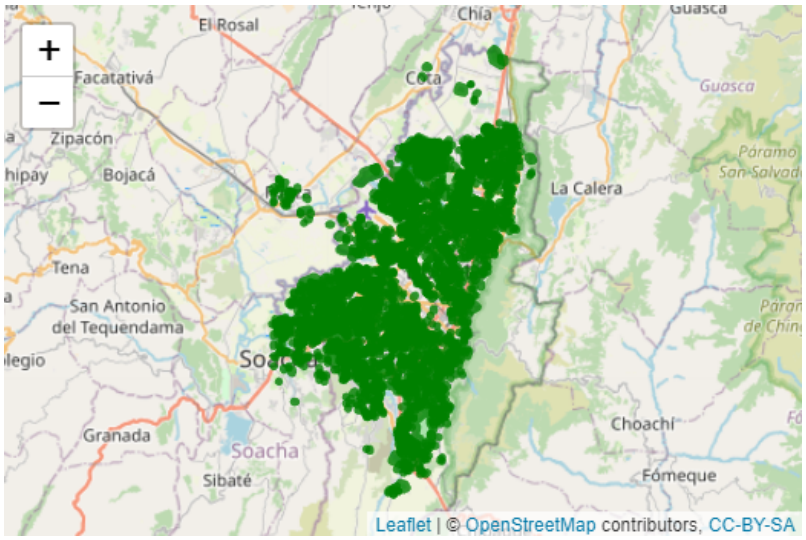
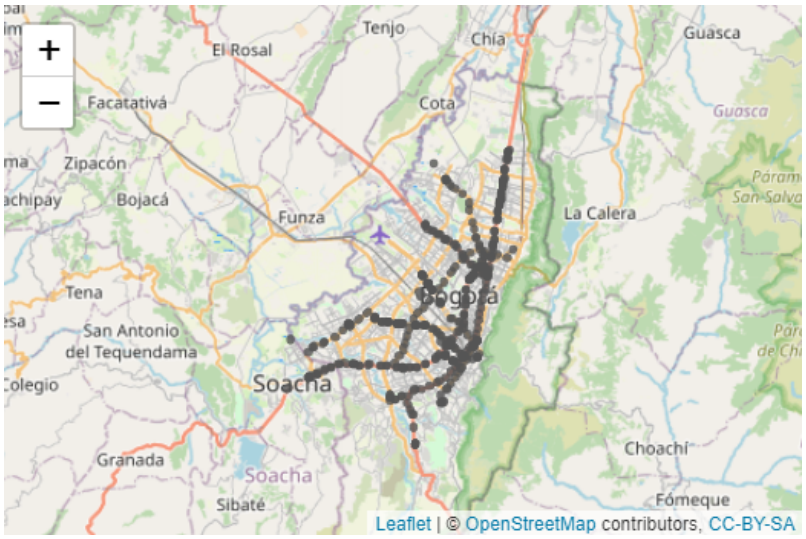


Figura 5. Mapa visual de inmuebles dentro de los límites de la ciudad de Bogotá



En el cuadro 2 se presentan los estadísticos descriptivos, el promedio y desviación

estándar de las variables de la base de datos de entrenamiento utilizadas para el modelo. El precio promedio de venta de los inmuebles es de 652.325.044 millones, en promedio los apartamentos tiene 3 habitaciones, 2 dormitorios y 2 baños. La distancia mínima al parque es de 160 metros en promedio, mientras que el área en metros del parque más cercano es de 7084 mts2, igualmente, la distancia mínima promedio al hospital es de 922 metros, mientras que la distancia mínima promedio a una estación de bus es de 949 metros, en este mismo sentido, la estación de policía esta ubica en promedio a 1021 metros y el supermercado en promedio a 1497 metros.

Cuadro 2. Estadísticas descriptivas generales

Variable	**Train**	
	Media	**Desv. Estandar**
Precio	652.325.044	308.381.860
Habitaciones	2,98	1,33
Dormitorios	2,97	1,32
Baños	2,55	1,09
Distancia minima al parque	160.673	102
Area del parque	7084.2	24.350
Distancia minima hospital	922.406	526
Distancia minima estación bus	949.702	680
Distancia minima policia	1.021.334	506
Distancia minima colegio	541.661	305
Distancia minima mercado	1497.09	788

Identificación

Para realizar una aproximación a la predicción de los precios de venta de los inmuebles ubicados en la localidad de Chapinero de Bogotá, este trabajo tiene como objetivo principal construir un modelo predictivo de precios de venta de las viviendas.

Obsérvese que un vector de características C

$$C = (c_1, c_2, ...c_)) \tag{1}$$

Describe un bien diferenciado, donde C es para una vivienda caracteristicas como atributos estructurales (e.g. númeor de dormitorios) , servicios publicos del vecindario (e.g calidad de la escuela local) y servicios locales (e.g. delicuencia, calidad del aire, etc).

Teniendo en cuenta esto, la estrategia empírica se basa en un modelo de predicción de preciso de venta P.

$$P_i = f(c_1, c_2, ...c_)) \tag{2}$$

Modelos y Resultados

Modelos de clasificación

Para la estimación de la variable de precio de la vivienda, se dividió la base de entrenamiento suministrada en dos submuestras de entrenamiento y testeo. Esto con dos objetivos. Por un lado, la submuestra de entrenamiento se utilizó, por medio de validación cruzada tipo K Fold con K igual a 10, para comparar las diferentes combinaciones de modelos propuestos y para hacer la optimización de los parámetros de los mejores modelos derivados de la comparación. Por otro lado, la base de testeo sirvió como último filtro de evaluación, permitiendo calcular la precisión de los modelos no solo por medio de los K Folds, sino con una última base de testeo que no se vio involucrada en el ajuste del modelo.

Dicha división se hizo por medio de la estratificación de la variable dependiente para mantener una misma proporción de dicha variable en las submuestras generadas.

Una vez definidas la base de entrenamiento, evaluación y validación cruzada, se corrió la

combinación de una receta[1] distintas que alimentaron 3 modelos. Para la receta utilizada, se utilizó la siguiente especificación derivada de la revisión de literatura y el análisis de estadísticas descriptivas presentado anteriormente.

$$\begin{aligned} Precio_i &= \beta_{i1} Superficie.Total + \beta_{i2} No.Habitaciones + \beta_{i3} No.Hab.Dormir + \\ &= \beta_{i4} No.Baños + \beta_{i5} Tipo.Propiedad + \beta_{i6} Balcon + \beta_{i7} Garaje + \\ &= \beta_{i8} Dist.Parque + \beta_{i9} Dist.Hospital + \beta_{i10} Dist.Estacion.Bus + \\ &= \beta_{i11} Dist.Policia + \beta_{i12} Dist.Colegio + \beta_{i13} Dist.Mercado + \\ &= \beta_{i14} Long + \beta_{i15} Lat + \beta_{i16} Num + \epsilon i \end{aligned} \tag{3}$$

La mitad de las variables incluidas se derivan directamente de la base de entrenamiento provista o fueron construídas con ayuda de datos externos. Dentro de estas últimas, se encuentran, las variables dicotómicas que identifican si una vivienda tiene o no balcón o garaje y las variables de distancia a determinadas localizaciones que, según la literatura afectan el precio de la vivienda (e.g. Distancia a hospitales, parques, estaciones de bus, colegios, etc).

En cuando a las metodologías de estimación contempladas para la predicción, se tuvieron en cuenta un modelo de regresión lineal, un modelo de random forest y un modelo de arboles secuenciales tipo Xgboost.

Se ajustaron y evaluaron 3 modelos derivados de la combinación de la receta con las metodologías de estimación por medio de una metodología de validación cruzada tipo K fold con K = 10. La métrica elegida para la evaluación y comparación de modelos fue el MAE, en el marco de la competencia de Kaggle.

[1]Se utiliza la terminología del paquete Tidy Models, el cual fue usado para el presente taller. Una receta hace referencia a la especificación usada en el modelo y a toda la preparación previa que se le aplica a la base de datos antes de correr el modelo

Cuadro 3. Modelos evaluados ordenados por su Error Absoluto Medio

	EAM	Error Estándar
(1) Random Forest	117.398.200	691.987
(2) Xgboost	126.943.400	958.053
(3) Regresión lineal	201.648.100	594.270

Como se puede observar, los mejores modelos fueron los basados en árboles de decisión, más específicamente el de bosques aleatorios. Por lo tanto, se tomó este modelo, para aplicarle un proceso de optimización de hiperparámetros con el objetivo de mejorar un poco más su EAM. Esto resulta relevante en cuanto a que, por carga de computo, todos los modelos en el paso anterior se corrieron con los valores predeterminados por los paquetes utilizados en el código. Por lo tanto, dicho valores pueden no ser los mejores para la muestra utilizada y la especificación definida.

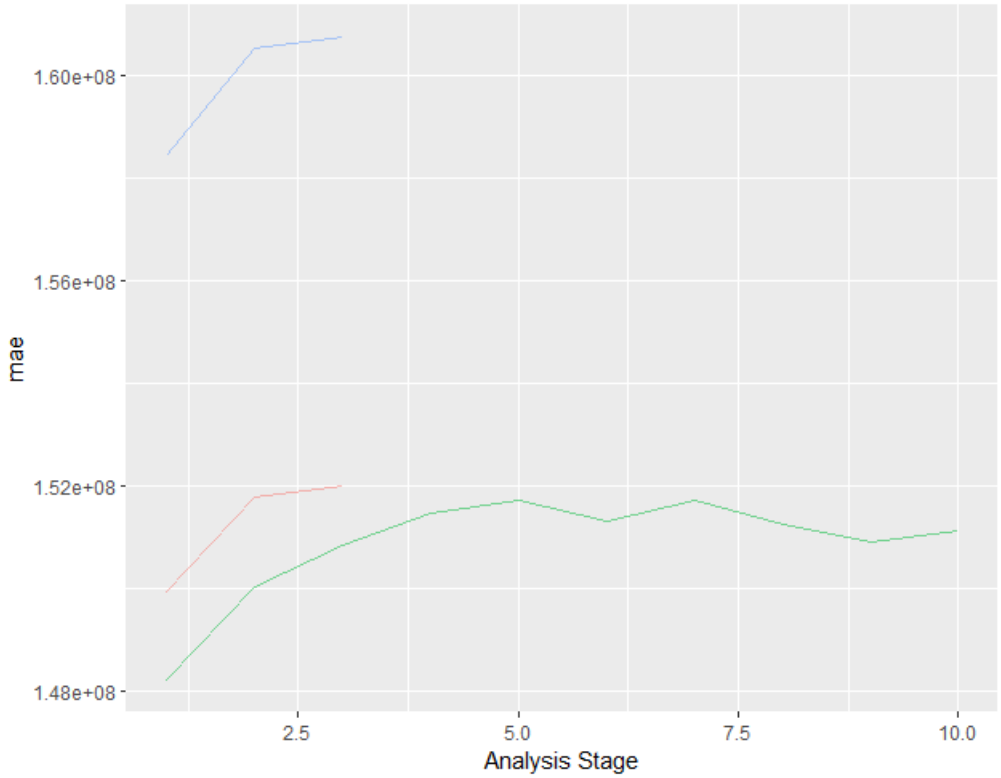
Se optimizaron todos los parámetros disponibles por medio del mismo proceso de validación cruzada tipo K Fold usado en el paso anterior. Para ello se definió una grilla con los valores del límite inferior y superior del rango que se quería evaluar para cada uno de los hiperparámetros en cuestión (ver Cuadro 4)

Cuadro 4. Rangos tenidos en cuenta para el proceso de optimización de parámetros

	Random Forest	
Parámetro	Límite inferior	Límite Superior
Número de Árboles	100	500
Número mínimo de observaciones por nodo	10	700
Número de predictores muestreados aleatoriamente	3	10

Para hacer mucho más eficiente el proceso de optimización por hiperparámetros se implementó un proceso de optimización por modelos ANOVA, que elimina las combinaciones de parámetros que son menos probables de arrojar los mejores resultados, reduciendo así los tiempo de cómputo de la optimización. En la Figura 2 se puede observar los combinaciones que fueron filtradas en los primeros resampleos de la optimización del Random Forest en función del EAM como métrica de referencia, mostrando los recursos y el tiempo ahorrados en no seguir sampleando dichas combinaciones en las submuestras aletorias restantes.

Figura 6. Proceso de optimización de parametros, proceso de eliminación de combinaciones no prometedoras

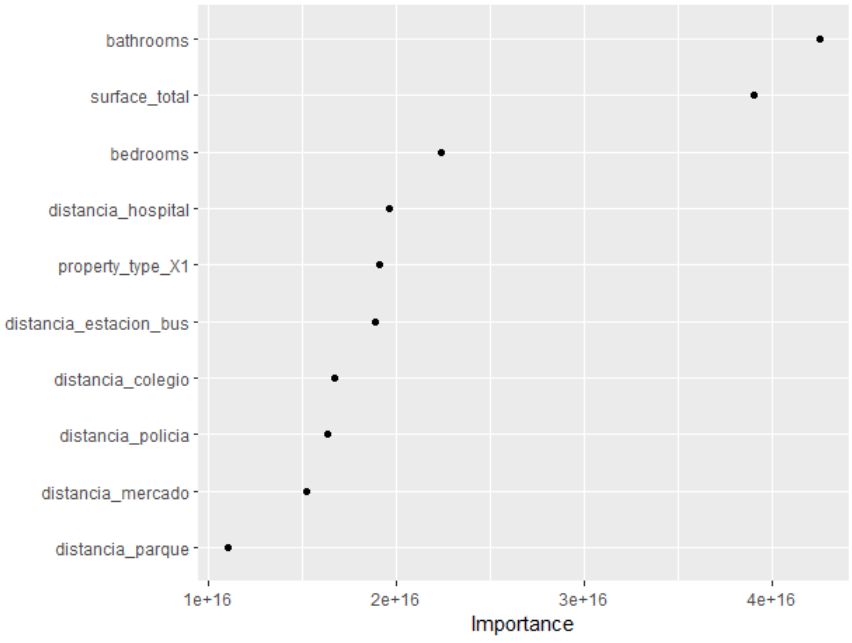


Alineado con lo anterior, los modelos que tuvieron un mejor desempeño en la predicción fuera de muestra, utilizando la base suministrada para la competencia, fueron los Random Forest con y sin optimización de parámetros. Estos modelos obtuvieron errores absolutos

medios cercanos a los 117 millones de pesos.

Cuando se toma dicho modelo y se evalúa la importancia de los predictores (ver Figura 3), se encuentra que la variable más importante es la de número de baños, seguida de la superficie total y el número de cuartos para dormir. Adicionalmente, se identifica que las variables de distancia construidas también resultan bastante importantes en la precisión del modelo, especialmente la distancia a establecimientos que prestan servicios públicos básicos de salud, transporte, educación y seguridad.

Figura 7. Importancia de los predictores Random Forest



Teniendo en cuenta el análisis anterior, el modelo seleccionado para las predicciones de kaggle fue un modelo tipo Random Forest con las siguientes características:

1) La especificación presentada en la ecuación dos

2) Un método de imputación por mediana para las variables continuas.

3) Un ajuste y evaluación del modelo por medio de validación cruzada tipo KFold con k = 10, y segunda evaluación fuera de muestra con una submuestra de evaluación equivalente al 25 % de la muestra original suministrada.

4) Un proceso de optimización de parámetros más eficiente con modelos de filtro y selección ANOVA y una grilla de optimización para los parámetros descritos anteriormente. La optimización también se llevo a cabo con las 10 submuestras del K Fold y se probó posteriormente el mejor modelo en la submuestra de evaluación previo a la carga de las predicciones a Kaggle.

Nota: Github

En el siguiente link https://github.com/Juansv24/Problem-Set-3-House-Prices se encuentran el acceso al repositorio del Problem set 3, donde se almacena tanto el documento en PDF, como el código y los gráficos.

Referencias

[1] Bertaud, A(2018)., *Order without Design: How Markets Shape Cities*The MIT Press.

[2] Uribe, J. P. (2019)., *Equilibrium Effects of Hpusing Subsidies: Evidence from a Policy Notch in Colombia*,Obtenido de https://juanpablo-uribe.github.io/Uribe_Brown_JMP_Housing_Subsidies_2021.pdf

[3] Rosen, S. ,(1974)., *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition* Journal of Political Economy, 82(1), 34–55. http://www.jstor.org/stable/1830899.