**Taller Grupal No. 3**

Problem Set No. 3 – Making Money with ML?

# Introducción:

Por medio de este documento, se presentan los resultados al estudio y desarrollo de un modelo predictivo tendiente a la definición del precio de inmuebles en la localidad de Chapinero, Bogotá D.C., y su aplicación en la estrategia de compra de dichos bienes por el menor valor posible. Todo lo anterior, en el marco del Problem Set No. 3 del curso Big Data de la Facultad de Economía de la Universidad de Los Andes.

## El mercado inmobiliario:

El mercado inmobiliario es uno de los mercados fundamentales para el desarrollo económico de una nación, toda vez que permite la generación de espacios para vivienda, comercio e industria a través de actividades de construcción que generan impacto en la producción, la prestación de servicios y el uso de instrumentos financieros, entre otros aspectos[[1]](#footnote-2). Es así, que el mercado inmobiliario tiene una relación importante con los consumidores y las firmas de inversión, ya que estos agentes acuden a los inmuebles, con recursos de ahorro o por medio del acceso a crédito, como una forma de inversión, bien sea para uso directo o a manera de renta[[2]](#footnote-3). En el contexto colombiano, podemos encontrar las siguientes cifras:

* Según el centro de estudios económicos del Banco BBVA, el sector de la construcción y venta de inmuebles representa el 13,5% del PIB nacional, y el 7,3% del empleo en el país.[[3]](#footnote-4)
* En materia de vivienda para 2021, entre el 50% y 70% de la compra de este tipo de bienes, tuvo una vocación de inversión para renta, y no para uso propio. En materia de inmuebles comerciales y de industria, se encontró una disminución en su comercialización entre 2020 y 2021. Para 2022, inicia un crecimiento de las transacciones de venta y arriendo. [[4]](#footnote-5)
* Según la Asobancaria, para 2022, se realizaron 128.200 operaciones de financiación para la compra de inmuebles, cuyo valor fue aproximadamente de 15.8 billones de pesos. Igualmente, aumentó el uso de instrumentos como el leasing habitacional y el crédito hipotecario[[5]](#footnote-6).

Todo esto, genera una necesidad importante: ¿Cómo definir el precio real de un inmueble? Al respecto, se debe considerar, no solo el avaluó comercial del bien, generalmente provisto por las autoridades territoriales, sino también las utilidades o servicios a los que puede acceder el poseedor del inmueble dentro de una zona determinada. Es así como, por ejemplo, el valor que le puede dar una pareja con hijos a un inmueble que se ubica en una zona con oferta de colegios es distinta a la de unas personas sin hijos respecto al mismo bien. Se resalta dentro de las investigaciones realizadas, la del autor Sherwin Rosen frente al uso de precios hedónicos en mercados implícitos y productos diferenciados[[6]](#footnote-7).

En este marco, es que un modelo predictivo que considere estas variables es importante para conocer las dinámicas del mercado y que consumidores y firmas conozcan de la mejor manera posible el verdadero precio del bien a ser adquirido o arrendado.

## Estructura del documento:

Para llegar a lo anterior, y considerando lo expuesto, el documento se dividirá en los siguientes acápites, a saber: Primero, presentación del enlace al repositorio de GitHub; segundo, indicaciones frente al origen y método de procesamiento de los datos, así como las correspondientes estadísticas descriptivas; tercero, análisis de los modelos realizados y sus resultados y; cuarto, conclusiones.

# Enlace a repositorio de GitHub:

El enlace de Github donde podrá encontrarse el repositorio con las respuestas del taller es el siguiente:

|  |
| --- |
| **Enlace al repositorio en GitHub** |
| <https://github.com/Carlosvergara1995/Problem_Set_3_Making_Money_with_ML.git> |

# Descripción de los datos:

A continuación, se presenta información sobre los datos utilizados dentro del ejercicio, así como las acciones tomadas para su limpieza y procesamiento.

## Origen de los datos:

Los datos utilizados para efectos de este análisis fueron sustraídos de la página web propertati.com.co, con información para 2021[[7]](#footnote-8). En este bloque de datos, se encuentran variables relacionados con la fecha de publicación del inmueble para venta y arriendo, su superficie, precio, el título, la descripción del inmueble y otros datos suministrados por el usuario. La información entregada consistió en dos (2) bases de datos: Una base de testeo (test) y una base de entrenamiento (training).

La información, a su vez fue complementada con datos del aplicativo OpenStreetMap para la ciudad de Bogotá y las Unidades de Planeación Zonales y localidad de Chapinero[[8]](#footnote-9).

## Proceso de limpieza y preparación de datos:

El proceso de limpieza de datos se dio con fundamento en las instrucciones del Taller, pero también buscando complementar la información que se encontraba dentro de las bases de datos provenientes del portal Propertati. Así, y dada el alto número de *missing values* se tomaron los datos, tanto de la descripción como del título de los inmuebles para completar los datos faltantes.

Adicionalmente, y considerando que la base de datos contaba con información de carácter geográfico, particularmente la latitud y longitud, se tomaron dichas bases, para convertirlas en data frames a través de funciones para la obtención de datos espaciales, convirtiendo las variables de latitud y longitud en una variable de coordenadas (coords=c(“lon”,”lat”) bajo el sistema de referencia WGS84. Con ello, se pudo delimitar el espacio de la variable de entrenamiento a la localidad de Chapinero, Bogotá D.C. con la correspondiente ubicación de los inmuebles.

En particular, utiliza la biblioteca “leaflet” para crear el mapa y “addTiles” para agregar un fondo de mapa. Luego, utiliza la función “addCircles” para agregar círculos a cada ubicación de la casa en el mapa, con el tamaño del círculo determinado por la variable especificada en el argumento “radius” y el color determinado por la variable especificada en el argumento “color”. En este caso, se utiliza la función “addCircles” en combinación con un objeto de datos llamado “house”, que se supone que es un objeto de datos espaciales que contiene la ubicación de las casas en las variables “lat” (latitud) y “lon” (longitud).

| **Extracto de los mapas interactivos de la Localidad de Chapinero, Bogotá D.C.** | | |
| --- | --- | --- |
| **Definición de la localidad** | **Ubicación de los inmuebles** | |
|  |  |
| **Fuente**: Elaboración propia. | |

Posteriormente, y para complementar la información de las bases de datos, se buscaron características relevantes (“amenities”), tanto para Bogotá D.C., como con Chapinero a través del código R de gCentroid para calcular los crentroides espaciales contenidos en la columna de coordenadas. Para efectos del ejercicio, esto se utilizó para calcular las distancias de cada inmueble con bares, escuelas, parques, restaurante y bancos. Con ello, se tienen las variables definitivas, previa imputación y complemento de NAs. Cabe aclarar que se mantienen separadas las bases de datos de testeo y entrenamiento, finalmente se realizó un cluster con 10 grupos, con base en la cercanía de la georreferenciación.

Una vez se tienen las variables finales, se ubican los *missing values* y se procede a complementar la información, con fundamento en la descripción y título de los bienes. Para estos efectos, se procesan las variables de baños (bathrooms) y superficie total (surface\_total). Para los datos restantes, se usó k vecinos cercanos, respecto a la imputación con descripción se encontré algunos datos atípicos como apartamentos de 2 m2 y con más baños que habitaciones, por lo cual se eliminaron estos valores y se los trato como NA, también las lecturas demasiado grandes como apartamentos. Igualmente, se agregan unas variables nuevas como parqueadero (parking), terraza (Terraza) y Garaje (garaje), a partir de la descripción y título de los bienes.

Finalmente, se dumifican las variables de parqueadero (parking), terraza (Terraza) y Garaje (garaje) y se realiza un análisis de colinealidad para ambas bases. Se advierte, que las bases finales son, para entrenamiento = training\_s y para testeo = test\_s.

## Estadísticas descriptivas:

A continuación, se presentan las estadísticas descriptivas.

### Frente a las bases de datos de entrenamiento (trainign):

En relación con las variables en las bases de datos de entrenamiento, se presentan las siguientes estadísticas descriptivas:

|  |  |
| --- | --- |
| **Estadísticas descriptivas entrenamiento** | |
|  | |
| **Fuente**: Anexo descrip\_train\_s | |

De la anterior tabla se puede apreciar lo siguiente:

Primero, existen, 38644 observaciones con 13 variables, las cuales representan la información obtenida de la base de datos de entrenamiento.

En segundo lugar, por ejemplo, frente a la relación de casas y apartamentos (property\_type), un total de 9,467 (24%) del total de las observaciones son clasificados como casa y el restante son apartamentos.

En tercer lugar, 16,474 inmuebles, correspondientes al 43% tienen parqueadero; 12,175, correspondientes al 32% tienen garaje y 11,568, 30% tienen terraza.

En cuarto lugar, el valor promedio de los inmuebles es de COP $654,534,675.

Adicionalmente, se encuentran las siguientes gráficas:

|  |  |
| --- | --- |
| **Gráficas de estadísticas descriptivas** | |
| **Relación entre el número de inmuebles y su valor** | |
|  | |
| **Fuente:** Anexo P | |

| **Distancia entre parques y valor de los inmuebles** | |
| --- | --- |
|  | |
| **Fuente:** Anexo p2 | |

|  |  |
| --- | --- |
| **Relación entre el valor del inmueble y distancia a colegios** | |
|  | |
| **Fuente:** Anexo p3 | |

|  |  |
| --- | --- |
| **Relación entre distancia a restaurantes y el valor del inmueble** | |
|  | |
| **Fuente:** Anexo p4 | |

| **Relación entre distancia a bancos y el valor del inmueble** | |
| --- | --- |
|  |
| **Fuente:** Anexo p5 |

|  |  |
| --- | --- |
| **Relación entre distancia a estaciones de bus y el valor del inmueble** | |
|  | |
| **Fuente:** Anexo p6 | |

### Frente a las bases de datos de testeo (test):

En relación con las bases de datos de testeo, se presentan las siguientes estadísticas descriptivas:

|  |
| --- |
| **Estadísticas descriptivas testeo** |
|  |
| **Fuente**: Anexo descrip\_test\_s |

De la table anterior, se pueden sustraer los siguientes puntos:

Primero, existen la base de datos solo tiene 10286 observaciones, divididas en 13 variables. De esta información, tan solo 274 observaciones, esto es, el 2,7% de los datos corresponden a inmuebles tipo casa.

Segundo, del total de observaciones el 32% cuenta con terraza, el 45% con parqueadero y el 31% con garaje de algún tipo.

Tercero, no se cuentan con datos de la variable precio, dado que es la variable dependiente, que será evaluada por medio de los modelos.

# Modelos:

A continuación, se presentan los modelos realizados dentro del ejercicio:

## Modelo de clasificación:

## Análisis de los árboles:

Para nuestro análisis por medio de modelos de clasificación hacemos uso de cross validation con el fin de nuestros modelos se generalicen bien con los datos fuera de muestra, para esto utilizamos subdivisiones de 8 y 10 las cuales entrenan el modelo en el subconjunto de entrenamiento y evalúan su rendimiento en el subconjunto de validación, de tal forma que nos ayudan a identificar los mejores hiperparametros de manera robusta y precisa.

El resultado de nuestro primer árbol nos indica lo siguiente:

|  |
| --- |
| **Arboles** |
|  |

El resultado nos indica que la variable bathrooms, la cual representa el número de baños, estima de manera significante el precio de las propiedades.

Procedemos a generar un segundo modelo incorporando una grilla para crear una cuadrícula de parámetros que se utilizarán en un modelo de bosque aleatorio, random forest, con divisiones de nodos dadas por la varianza, un número mínimo de observaciones para las siguientes divisiones de valores posibles de 10, 30 y 100, aplicando ramdom serch y un número de características a considerar en cada modelo de 4, 7 y 12, esto teniendo en cuenta el número de variables sometidas y su raíz cuadrada.

|  |
| --- |
| **Arboles realizados con las bases de datos de Test y Train** |
|  |
| **Fuente**: Anexo descrip\_test\_s |

El parámetro "min.node.size" es el número mínimo de observaciones permitidas en una hoja del árbol de regresión que se construye en cada iteración del algoritmo. Si se establece un valor alto para este parámetro, el árbol tendrá menos hojas y será más generalizado, mientras que un valor bajo permitirá que el árbol tenga más hojas y sea más específico para los datos de entrenamiento.

En el gráfico, se muestra cómo varía el error de validación cruzada (RMSE) al ajustar diferentes valores del parámetro "min.node.size". Cada línea representa una combinación diferente de valores de "mtry" (número de predictores seleccionados al azar) y muestra cómo cambia el RMSE a medida que se varía el valor de "min.node.size". Los puntos en la línea indican el valor del RMSE para un valor específico de "min.node.size" y "mtry".

En nuestro caso encontramos que el modelo óptimo tiene una cantidad óptima de 4 variables con un aproximado de 10 nodos.

## Modelo de regresiones:

Para el modelo de regresión se usaron todas las variables, de área, grupo de georreferencia, baños, habitaciones, año y las dummies mencionadas anteriormente. Para realizar la selección de hiperparametros utilizamos gridsearch variando "alpha" y "lambda". El "alpha" controla la mezcla de las regularizaciones L1 (Lasso) y L2 (Ridge) y puede tomar valores entre 0 y 1, con intervalos de 0,05, para Alpha se seleccionó los valores de 0, 0.001, 0.01, 0.1, 1, 10. Dado que no se cuenta con los datos de test para la validación en producción del modelo, se decidió aumentar el número de folder en la validación cruzada a 10, lo que permite una mejor estimación.

|  |
| --- |
| **Regresión lineal variando Alpla** |
|  |
| **Fuente**: Elaboración propia |

El modelo seleccionado es un elastic net con Alpha de 0.26 y los valores de lamda no afectan en gran medida los valores de RMSE. Al cargar la información e kaggle se obtiene un score de 288792156.36494 el cual es más alto que el obtenido por árboles, lo cual nos permite concluir que el mejor modelo para esta data es de árboles.

## Modelo final:

Dado que no se obtuvo valores bajos de RSME en ambos casos anteriores se realizó la prueba con superlearner, utilizando regresión lineal, random forest y rager, obteniendo como resultado lo siguiente:

|  |  |  |
| --- | --- | --- |
| **Resultado de los modelos** | | |
| Modelo | RISK | COEF |
| SL.randomForest\_all | 2.305177e+14 | 0 |
| Sl.lm\_ALL | 4.898973e-10 | 1 |
| SL.ranger\_1\_ALL | 5.215924e+14 | 0 |
| SL.ranger\_2\_ALL | 2.635153e+16 | 0 |
| SL.ranger\_3\_ALL | 4.502402e+15 | 0 |

Observamos como los valores de RISK son muy altos y únicamente un coeficiente está actuando, dado que esta clase de modelos tienen alto costo computación y no se pueden correr varias veces, se decide elegir uno de los modelos anteriormente planteado.

El modelo seleccionado fue modelo de ramdom forest el cual presento un ajuste fuera de muestra de 253963963.86393, este modelo como se indicó anteriormente fue ajustado usando cross validation con un valor de 8 y una grilla que integra divisiones de nodos dados por la varianza, un número mínimo de observaciones para las siguientes divisiones de valores posibles de 10, 30 y 100, aplicando ramdom serch y un número de características a considerar en cada modelo de 4, 7 y 12, esto teniendo en cuenta el número de variables sometidas y su raíz cuadrada.

# Conclusiones:

Se puede concluir que, variables como área total del inmueble, grupo de georeferencia, número de baño o habitaciones son las de mayor incidencia sobre el precio de las viviendas. Lo anterior se concluyó utilizando una validación cruzada de 10-fold, teniendo en cuenta que no se contaba con datos de prueba para evaluar la capacidad predictiva del modelo.

Respecto a la diferencia entre los valores de test, se destaca que se requiere otras variables para mejorar la predicción como la antigüedad de las viviendas, dado que se encontró que no existe una relación lineal entre los datos.

Finalmente, y en aras de mejorar el modelo, se utilizó el método superlearner y el modelo seleccionado pero debido a los recursos de máquina que requiere, no fue posible obtener los resultados y se dejará como enseñanza para futuros modelos.

1. Al respecto, el sub-director del FMI para 2014 indicó, particularmente frente al mercado de inmuebles para vivienda: “*First, housing is an essential sector of the economy but also one that has been the source of vulnerabilities and crises. Hence, while the recent recovery in global housing markets is a welcome development, we need to guard against another unsustainable boom. Second, detecting over-valuation in housing markets is still more of an art than a science. Broad measures, such as house price to rent ratios, provide a first pass. But detailed analysis and judgment are needed to make a call about overvaluation. Third, the policy toolkit to manage housing booms is still under construction. A variety of tools have been used and the evidence suggests some short-run success. But more analysis and sharing of experience are needed on what works and what doesn’t. Conferences of this kind are useful in adding to our stock of knowledge*.” En este orden de ideas, la importancia del mercado inmobiliario, no solo tiene efectos financieros, sino también sociales y de bienestar, dadas las barreras de acceso, y especialmente la vivienda como parte de una garantía fundamental en el bienestar de los agentes. FMI. *Housing Markets, Financial Stability and the Economy.* (11 de junio de 2014). Disponible en: <https://www.imf.org/en/News/Articles/2015/09/28/04/53/sp060514>. [↑](#footnote-ref-2)
2. Frente a la relación del mercado de inmuebles y el gasto de los consumidores ver: Alan Carruth y Andrew Henley. *The Housing Market and Consumers Spending.* Fiscal Studies, vol. 11, No. 3. (1990). [↑](#footnote-ref-3)
3. BBVA Research. *Colombia. Situación Inmobiliaria.* (enero de 2023).Disponible en: <https://www.bbvaresearch.com/publicaciones/colombia-situacion-inmobiliaria-2022/> [↑](#footnote-ref-4)
4. Fondo Inmobiliario Colombia. *Panorama Inmobiliario 2022.* (marzo de 2022). Disponible en: <https://fondoinmobiliariocolombia.com/panorama-inmobiliario-2022/1794/> [↑](#footnote-ref-5)
5. InHaus Colombia. *El sector inmobiliario en cifras: del 2021 al 2022.* (noviembre de 2022). Disponible en: <https://www.lahaus.com/blog/tendencias-del-mercado/sector-inmobiliario-cifras-2021-2022>. [↑](#footnote-ref-6)
6. Sherwin Rosen. *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.* Journal of Political Economy, vol. 82, No. 1, (1974). [↑](#footnote-ref-7)
7. Properati es un portal inmobiliario para la presentación y búsqueda de anuncios clasificados para la venta y alquiler de bienes inmuebles, el cual tiene presencia sur-américa. El portal ofrece datos abiertos por medio de la licencia Creative Commons 3.0. Ver: <https://cloud.google.com/customers/properati?hl=es-419>; Storybench. *How an Argentine real estate platform is building visualizations for Latin American newsrooms.* (5 de julio de 2017), disponible en: <https://www.storybench.org/argentine-real-estate-platform-building-visualizations-latin-american-newsrooms/>. [↑](#footnote-ref-8)
8. OpenStreenMpas es una base de con datos geográficos de carácter colaborativo, la cual cuenta con licencias de uso abierto. El aplicativo fue creado en 2004 y se fundamenta en la metodología WSG 84 para la ubicación de latitudes y longitudes. Disponible en: <https://www.openstreetmap.org/about>. [↑](#footnote-ref-9)