Daniel Casas Bautista

Código: **202120803**

Lucia Fillippo Aguillón

Código: **202213187**

Irina Andrea Vélez López

Código:

Miguel Ángel Victoria Simbaqueva

Código:

**Problem Set 2**

Big Data & Machine Learning

**Fecha de entrega: 15 de julio de 2023**

**Resumen:** El presente informe presenta la solución al Problem Set 2 de la clase Big Data & Machine Learning, en donde se aplicaron diversos conceptos y herramientas para la limpieza de bases de datos y el desarrollo de un modelo de predicción del precio de las viviendas en la localidad de Chapinero en Bogotá.

En el repositorio GitHub que abajo se referencia contiene el presente informe junto a los scripts de RStudio donde se generaron todos los resultados. Este repositorio se encuentra en: <https://github.com/irivelez/PS2_Making_Money_with_ML.git>

1. **Introducción**

***The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.***

El precio de una vivienda está determinado por una serie de factores que influyen en su valor, dentro de estos determinantes se encuentra el tamaño total del inmueble, el número de baños, la proximidad a bienes públicos y lugares importantes dentro de la ciudad, entre otros aspectos relevantes. Estos elementos se constituyen como variables predictivas que permiten explicar el precio de la propiedad. Cada ciudad tiene sus propias características distintivas que influyen en la determinación de precios por parte de los oferentes del mercado; sin embargo, es posible desarrollar un modelo de aprendizaje de máquinas que permita, a través de la inclusión de estas variables predictivas, proporcionar a los compradores y vendedores del mercado inmobiliario de vivienda una comprensión más precisa de cómo influyen estos factores en el precio de mercado del sector inmobiliario de vivienda.

De esta manera, será posible predecir el precio de uno de los mercados más relevantes en el desempeño económico de la ciudad y del país y, además, de cara al análisis de las políticas públicas, determinar el precio de mercado de uno de los bienes utilizados como política contra cíclica ante los ciclos macroeconómicos.

Considerando lo anterior, el objetivo principal de este documento es construir un modelo predictivo de los precios de la vivienda, a partir del siguiente modelo, donde la función representa un vector C con las características que explican el precio de la vivienda:

Para lograr el anterior objetivo se utilizaron datos de Properati sobre características de las viviendas y de Datos Abiertos Bogotá[[1]](#footnote-2), donde se logró extraer información de mapas para la creación de las mejores variables predictivas; sin embargo, para lograr lo anterior fue necesario realizar una limpieza de las bases de datos para lograr tener un grupo de muestra y uno de entrenamiento, mediante el cual se evaluaron los distintos modelos de predicción, para posteriormente elegir el de mejor error de predicción y, por ende, el que mejor ha predicho el precio de la vivienda.

1. **Datos**

***In this problem set, you are required to add expand the variables in your data (remember to expand the training and testing data), at a minimum you have to add six extra variables:***

* ***At least 4 predictors coming from external sources; these can be from open street maps***
* ***At least 2 predictors coming from the title or description of the properties***

***Describe the data, it's suitability for the problem, and the sample construction process, including how the data was cleaned, combined, and how new variables were created.***

***Include a descriptive analysis of the data. At a minimum, you should include a descriptive statistics table and two maps with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a \dry" list of ingredients.***

Se ha tomado una muestra de la página *Properati,* la cual contiene características de las viviendas como ciudad, precio, número de cuartos, número de baños, precio de la vivienda, así como algunas características más específicas, con el objetivo de limpiar y consolidar una muestra de prueba y una de entrenamiento que permita evaluar distintos modelos de predicción del precio de la vivienda; lo anterior, para la localidad de Bogotá en Chapinero.

Dentro de la información que tiene la base de datos se encuentran las siguientes variables relevantes.

* **Ubicación:** Esta variable puede influenciar el precio de la vivienda, pues, en caso de estar más cercano al centro de la ciudad o a una zona empresarial, puede variar el precio de la vivienda. En la base de datos se trata de una variable que da las coordenadas geométricas.
* **Número de cuartos:** Esta variable es relevante en el análisis precisamente porque influye en el tamaño de la propiedad y en la cantidad de personas que pueden habitarla.
* **Número de baños:** Esta variable es relevante en el análisis precisamente porque influye en el tamaño de la propiedad y en la cantidad de personas que pueden habitarla.
* **Tipo de propiedad**: Variable dummy que indica el tipo de propiedad si es apartamento o casa. El tipo de propiedad afecta el precio de la vivienda precisamente por distintos motivos como el tamaño, el número de amenidades y espacios de recreación privados, entre otros.

Las siguientes variables, provienen de la descripción de los apartamentos y/o casas en observación. Es decir, se complementaron a partir de la información descriptiva y de la segunda variable en adelante, todas están dumificadas[[2]](#footnote-3).

* **Metros cuadrados (*Surface covered)***: Referencia al número de metros cuadrados que componen la propiedad.
* **Baño privado:** Variable dummy que toma el valor de 1 si tiene baño privado y cero en caso contrario.
* **Ascensor:** Variable dummy que toma el valor de 1 si tiene ascensor y cero en caso contrario.
* **Balcón:** Variable dummy que toma el valor de 1 si tiene balcón y cero en caso contrario.
* **Vista:** Variable dummy que toma el valor de 1 si tiene vista y cero en caso contrario.
* **Remodelado:** Variable dummy que toma el valor de 1 si la propiedad está remodelada y cero en caso contrario.

Adicionalmente, en cuanto a los predictores provenientes de bases de datos externos, se ha tomado las siguientes variables de referencia:

* **Cercanías a parques y zonas de recreación:** Distancia desde las observaciones hasta los parques.
* **Cercanía a centros deportivos:** Distancia desde las observaciones hasta los centros deportivos (Ej. Gimnasios).
* **Área del parque más cercano:** Busca medir el tamaño del parque más cercana a la observación.
* **Cercanía a las piscinas:** Distancia desde las observaciones hasta las piscinas.

Tras limpiar las bases de datos y crear los polígonos de análisis, los cuales se obtienen de Open Street Map, se selecciona la información de la base de datos correspondiente a estas zonas. En la Figura 1 se muestran la información de muestra y entrenamiento para la zona.

Figura 1. Polígono de estudio

Mapa

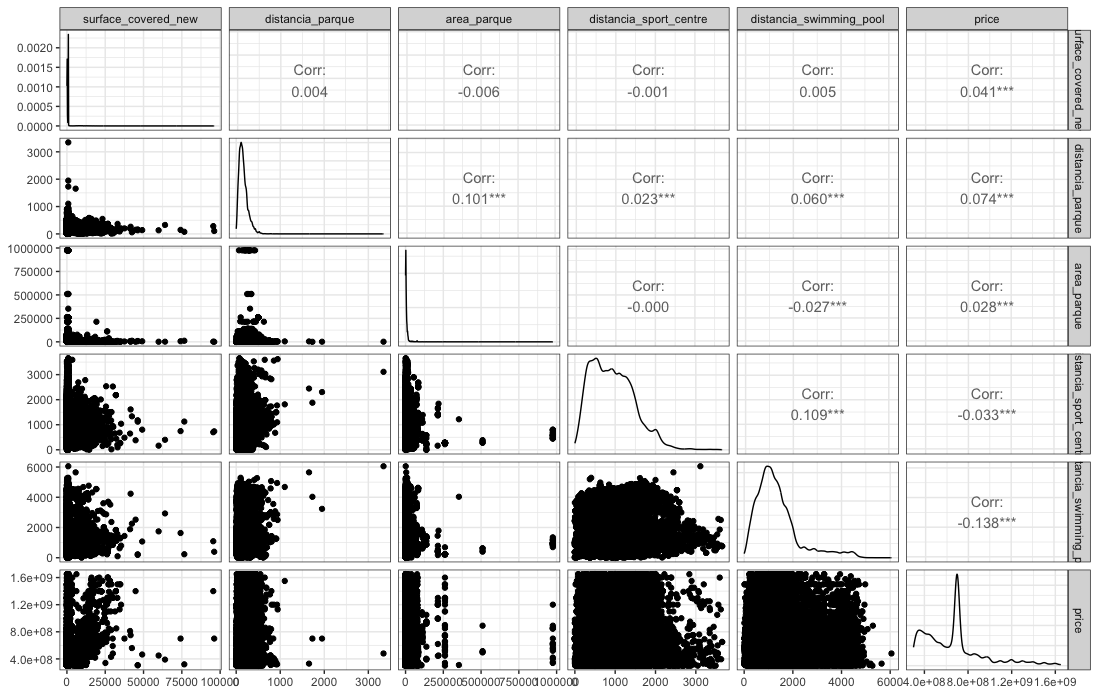
Descripción generada automáticamente

Fuente: Elaboración propia mediante Rstudio con datos de Properati

* 1. **Estadísticas descriptivas**

A continuación, se presentan las estadísticas descriptivas de las diferentes variables que componen nuestra base de datos tanto para predictores provinientes de la base de datos inicial como de fuentes externas.

En ese sentido, la siguiente ilustración describe el compotamiento de las variables continuas de la base de datos.



**Fuente: Elaboración propia en Rstudio con base de datos realizada a partir de**

**Properati y Datos Abiertos de la Alcaldía de Bogotá**

A partir de la anterior ilustración, obtenemos la distribución de los datos entre sí, sus gráficas de dispersión y finalmente, sus coeficientes de correlación. Sobre estos resultados, vale la pena resaltar, que la distribución de los datos de la mayoría de las variables está centrada sobre valores pequeños, que existe una concentración inicial en los valores pequeños y, adicionalmente, que ninguna de las variables se comporta de manera normal.

Analizando la variable de interés, ***Price*** el coeficiente de correlación más alto es el referente al ***Area\_parque*** que referencia el tamaño del área del parque. Esto, es contrastable y verificable con la realidad en ejemplos como el valor de las propiedades cercanas al Parque del Virrey o el Museo del Chicó, que de acuerdo a Metro Cuadrado, los barrios en los que se ubican, son los más costosos de la ciudad y por ende de la localidad[[3]](#footnote-4).

El coeficiente de correlación más alto fue aquel entre ***distancia\_swimming\_pool*** que referencia la distancia entre las observaciones y la piscina más cercanay ***distancia\_sport\_center*** que explica la distancia entre las observaciones y el centro deportivo más cercano, esto puede deberse precisamente a que muchas piscinas pueden estar incluidas en los centros deportivos, o son interdependientes entre sí.

1. **Modelo y resultados**

***This section presents the model(s) submitted for evaluation. When writing this section up, include:***

* ***An explanation of the variables used to train this model, remember to use the variables you added in the previous section.***
* ***A detailed explanation on how it was trained, the selection of hyper-parameters, and any other relevant information.***
* ***A comparison to at least 4 other specifications submitted to Kaggle.***

A continuación presentamos los modelos elegidos para evaluación:

1. **LM model: Contiene variable Y = Logaritmo natural del Precio (Ln\_Price) contra la variable de metros cuadrados.**

Se implementó un modelo de regresión lineal simple (LM) como estimación primeriza y preliminar predecir el precio de las viviendas de Chapinero. Para ello, se utilizó una variable dependiente Y, que representa el logaritmo natural del precio de las viviendas (Ln\_Price), y una variable independiente, que corresponde a los metros cuadrados de las viviendas. La elección del logaritmo natural del precio (Ln\_Price) como variable dependiente puede deberse a varios motivos, como la necesidad de reducir la asimetría en la distribución de los precios o la estabilización de la varianza en los residuos del modelo.

El modelo de regresión lineal simple busca establecer una relación lineal entre el logaritmo natural del precio y los metros cuadrados, lo que implica que se supone una relación directamente proporcional entre ambas variables. En ese sentido, la idea es que al conocer el valor de los metros cuadrados de una vivienda, el modelo pueda proporcionar una estimación del precio considerando una normalización por el logaritmo natural, que luego podría ser transformado nuevamente para obtener una estimación del precio real de la vivienda.

1. **LOOCV Model: El mismo modelo al anterior pero realizando varias iteraciones.**
2. **LOOCV Model II: Contiene variable Y = Logaritmo natural del Precio contra el resto de variables creadas en la base de datos (Variables provenientes de la descripción, variables ya existentes en la base de datos y las variables derivadas de fuentes externas).**
3. **RPart Model: Primer modelo de árbol con los mismos parámetros.**
4. **Ranger Model I: Modelo de árbol con los hiperparámetros:**
   1. **MIN.NODE.SIZE con un rango desde diez (10) a cien (100).**
   2. **MTRY de tres (3), (5) y (10).**
5. **Ranger Model II: Modelo de árbol con los hiperparámetros:**
   1. **MIN.NODE.SIZE con un rango desde cinco (5) a cincuenta (50).**
   2. **MTRY de tres (3), (5) y (10).**
6. **Boosting Model:**
7. **Conclusiones y recomendaciones**

***In this section, you briefly state the main takeaways of your work***

1. Alcaldía de Bogotá (2023). Mapa de referencia para Bogotá D.C. Datos abiertos. Disponible en: <https://datosabiertos.bogota.gov.co/dataset/mapa-de-referencia> (Consultado el 12 de julio de 2023). [↑](#footnote-ref-2)
2. Termino acuñado a una variable dummy que toma los valores de cero y uno. [↑](#footnote-ref-3)
3. Metro Cuadrado (2023). Los cinco barrios más caros de Bogotá. Tomado de: <https://www.metrocuadrado.com/noticias/actualidad/los-cinco-barrios-mas-caros-de-bogota-1248/> (Consultado el 15 de julio de 2023). [↑](#footnote-ref-4)