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**

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**

La constitución del modelo de predicción incorpora factores económicos y sociales que han sido recopilados de la página de *Properati* y de Datos Abiertos Bogotá. La primera base de datos contenía información de identificación de la vivienda, ciudad, precio, número de cuartos, número de baños, precio y valores de latitud y longitud; además, incluía una descripción de la vivienda, mediante la cual se permitió extraer de dicho texto información estratégica para crear algunas características más específicas. Para lograr lo anterior fue necesaria una limpieza de la base de datos, eliminando aspectos como tildes, mayúsculas, espacios extra y caracteres especiales.

Por otro lado, de Datos Abiertos se lograron extraer mapas que pudieron adherirse a la anterior base de datos, mediante sus características geoespaciales. A través de este proceso se logró consolidar una única base de datos que posteriormente permitiría obtener una muestra de prueba y una de entrenamiento, las cuales serían fundamentales para la evaluación de los distintos modelos de predicción del precio de la vivienda. Es importante precisar que lo anterior se acotó a la localidad de Chapinero en Bogotá.

Dicho lo anterior, de la primera base de datos construida con información de *Properati*, sin haber extraído la información de la variable descripción, se incluyen las siguientes variables:

* **Número de cuartos:** Esta variable es numérica y se considera relevante porque influye en el tamaño de la propiedad, en la cantidad de personas que pueden habitarla y, por ende, en su potencial valor; es decir, entre más habitaciones tenga la vivienda, mayor sería su valor;
* **Número de baños:** Esta variable es numérica e incide en el valor que podría tener una propiedad, dado que, si un hogar no tiene baños, los miembros del hogar no podrían satisfacer sus necesidades básicas de aseo, razón por la cual, en caso de contar con más baños, el precio de la vivienda podría ser mayor;
* **Tipo de propiedad**: Esta es una variable dummy que toma el valor de 1 si la propiedad es un apartamento y 0 si el valor es una casa. Esto es relevante para determinar el valor de la vivienda, pues una casa podría tener un mayor tamaño y otro número de amenidades y espacios de recreación privados, entre otros.

Ahora bien, producto de haber extraído nuevas variables de la base de datos obtenida de *Properati* y depurada, se crearon las siguientes variables adicionales que complementaron la base anterior.

* **Metros cuadrados (*Surface covered)***: Esta variable es numérica y referencia el número de metros cuadrados que componen la propiedad. Esta variable puede determinar su precio, en la medida en que un mayor tamaño de la vivienda puede representar un mayor precio;
* **Baño privado:** Esta es una variable dummy que toma el valor de 1 si la vivienda tiene baño privado y 0 en caso contrario. En la medida en que una vivienda cuente con baño privado, se considera que puede incrementar su valor. El 10.8% de las viviendas tiene baño privado;
* **Ascensor:** Esta es una variable dummy que toma el valor de 1 si la vivienda tiene ascensor y 0 en caso contrario. Tener un ascensor podría facilitar el acceso a personas con movilidad reducida, personas mayores o familias con niños pequeños. Estas características pueden ser un factor importante que podría aumentar la demanda de viviendas con ascensor. El 18.7% de las viviendas tiene ascensor;
* **Balcón:** Esta es una variable dummy que toma el valor de 1 si la vivienda tiene balcón y 0 en caso contrario. La presencia de un balcón puede influir en el precio de una vivienda dado el espacio al aire libre que brinda, la vistas, luz natural y atractivo para aquellos que valoran un estilo de vida al aire libre. Por dicha razón, podría incrementar el valor de la vivienda. El 43.1% de las viviendas tiene balcón;
* **Parqueadero:** Esta es una variable dummy que toma el valor de 1 si la vivienda tiene parqueadero y 0 en caso contrario. Contar con parqueadero puede influir en el precio de la vivienda debido a la comodidad, demanda y seguridad que proporciona. Un parqueadero propio incrementa el valor de reventa y atrae a compradores interesados en evitar problemas de estacionamiento en áreas urbanas congestionadas o con restricciones de estacionamiento. El 69.5% de las viviendas tiene parqueadero;
* **Vista:** Esta es una variable dummy que toma el valor de 1 si la vivienda tiene vista a un sitio importante de la ciudad y 0 en caso contrario. Se considera que esta variable puede ser importante, dado que una vivienda con una ubicación que permita una vista a un gran sitio puede tener un mayor precio final. El 25.6% de las viviendas tiene buena vista;
* **Remodelado:** Esta es una variable dummy que toma el valor de 1 si la propiedad está remodelada y 0 en caso contrario. Se considera que una vivienda remodelada puede tener un valor mayor, respecto de una que no haya surtido tal proceso. El 41.6% de las viviendas están remodeladas;

Tras el anterior proceso de depuración se procedió a utilizar la información extraída de Datos Abiertos Bogotá para identificar la cercanía de cada vivienda a sitios estratégicos como el centro de la ciudad, a bienes públicos o a una zona empresarial; lo anterior, mediante las coordenadas geométricas de la base de datos de *Properati*. Dicho lo anterior, a continuación se muestran las variables identificadas:

* **Distancia a parques y zonas de recreación:** Esta es una variable continua que describe la distancia de la vivienda al sitio referenciado;
* **Distancia a centros deportivos:** Esta es una variable continua que describe la distancia de la vivienda al sitio referenciado (ejemplo: gimnasios);
* **Distancia a las piscinas:** Esta es una variable continua que describe la distancia de la vivienda al sitio referenciado;
* **Área del parque más cercano:** Esta es una variable continua que describe el tamaño del parque más cercano a la vivienda;

De las anteriores variables creadas, debe tenerse en cuenta que en los modelos a utilizar se utilizarán únicamente dos de las variables obtenidas de *Properati*; por otro lado, se utilizarán las cuatro variables obtenidas de fuentes externas. Finalmente, y tras limpiar las bases de datos y obtener las respectivas variables predictoras, a continuación se muestran los polígonos de análisis, los cuales se obtienen de Open Street Map mediante el software Rstudio. Estos polígonos corresponden a la localidad de Chapinero, como se puede ver en la Figura 1, donde se incluye la información de muestra y entrenamiento.

Figura 1. Polígono de estudio

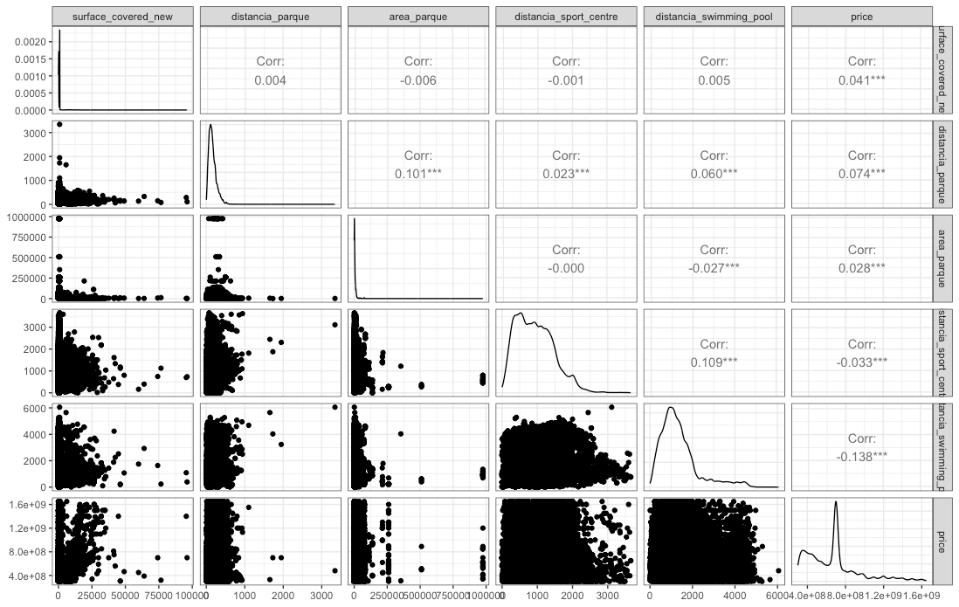
Mapa

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, esto es, el precio (***Price****)* el coeficiente de correlación más alto es el referente al área del parque más cercano a la vivienda (***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 con Metro Cuadrado, los barrios en los que se ubican son los más costosos de la ciudad y, por ende, de la localidad[[2]](#footnote-3). Por otro lado, el coeficiente de correlación más alto fue aquel entre la distancia a la piscina más cercana (***distancia\_swimming\_pool***) y la distancia al centro deportivo o gimnasio más cercano (***distancia\_sport\_center***), esto puede deberse a que muchas piscinas pueden estar incluidas en los centros deportivos, o son interdependientes entre sí.

1. **Modelo y resultados**

A continuación se presentan los modelos elegidos para evaluación; no obstante, es pertinente aclarar que el primer modelo es netamente preliminar y se ha utilizado como referencia, con el objetivo de mostrar la manera en que el error de predicción va mejorando en la medida en que se hace uso de un modelo más robusto.

1. **Linear model (LM)**

Este modelo contiene como variable Y el Logaritmo natural del Precio (Ln\_Price) y como variable predictora el tamaño de la vivienda en metros cuadrados. Este primer resultado implementó un modelo de regresión lineal simple (LM) como estimación preliminar para predecir el precio de las viviendas de Chapinero. La elección del logaritmo natural del precio como variable dependiente se debe a la necesidad de reducir la asimetría en la distribución de los precios o la estabilización de la varianza en los residuos del modelo. En ese sentido, la idea es que al conocer el valor del área 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. En este ejercicio inicial, el Mean Abolute Error – MAE fue de: 0.1336067 (87521698 COP)

1. **Leave One Out Cross Validation (LOOCV)**

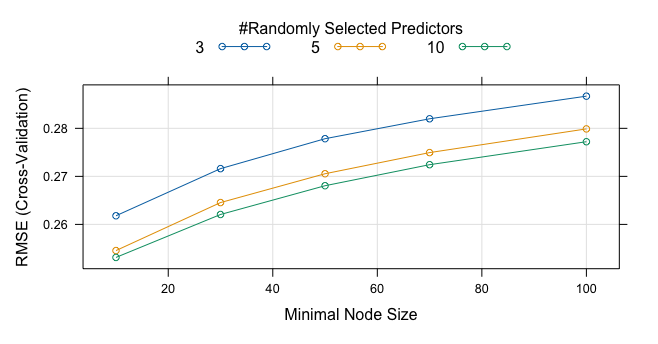
En esta ocasión se tomó el mismo modelo al anterior; no obstante, por la naturaleza del modelo se implementaron varias iteraciones para la información de muestra y de entrenamiento. En este caso el MAE fue de: 0.1336067 (87521698 COP)

1. **Leave One Out Cross Validation (LOOCV) – Modelo II**

Este modelo contiene como variable Y el Logaritmo natural del Precio (Ln\_Price) y como variables predictoras tiene: área en metros cuadrados, Número cuartos, Número de baños, Parqueadero, Ascensor, Baño privado; Balcón; Vista; Remodelado; Tipo de propiedad; Cercanía a parques; Área del parque más cercano; Cercanía a centros deportivos y Cercanía a piscinas. En este caso el MAE fue de: 0.1609716 (88643254 COP)

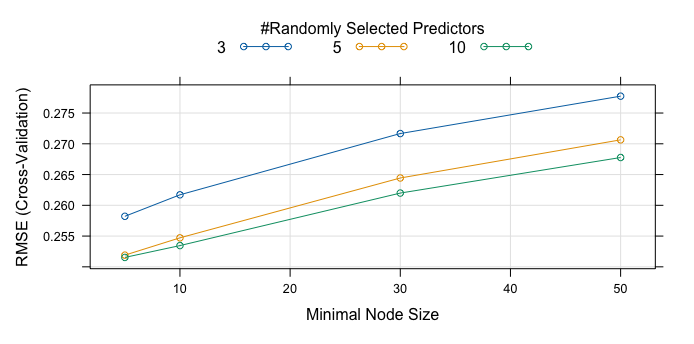
1. **Primer modelo de bosques son hiper parámetros (Ranger)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo; no obstante, se adicionaron como hiper parámetros un MIN.NODE.SIZE con un rango desde (10) a (100), y un MTRY de (3), (5) y (10). Los resultados del RMSE, a la luz de la inclusión de los hiper parámetros, se observa en la siguiente gráfica, donde se muestra que el nodo en 10 es el de menor RMSE, mientras que el de 3 tiene mayor RMSE. Producto de lo anterior, en este caso el MAE fue de: 0.08934838 (59397009 COP)

****

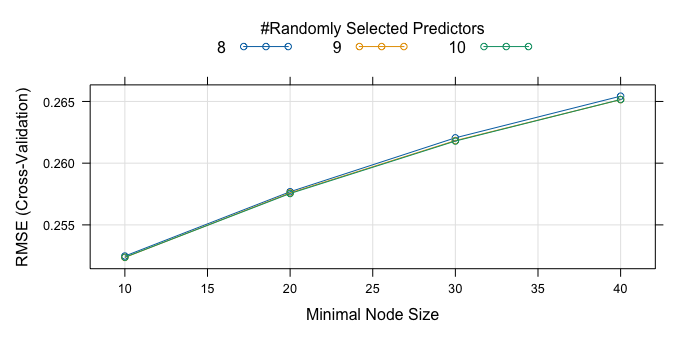
1. **Segundo modelo de bosques son hiperparámetros (Ranger)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a los hiper parámetros se tiene un MIN.NODE.SIZE con un rango de (5), (10), (30) y (50) y un MTRY de (3), (5) y (10). Los resultados del RMSE, como se observa en la gráfica, muestra que el nodo en 10 sigue siendo el de menor RMSE y varía en la medida en que se incrementa el tamaño del nodo. En este caso el MAE fue de: 0.0826849 (54991319 COP)

****

1. **Tercer modelo de bosques son hiperparámetros (Ranger)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a los hiper parámetros se tiene un MIN.NODE.SIZE con un rango de (10), (20), (30) y (40) y un MTRY de (8), (9) y (10). Los resultados del RMSE, como se observa en la gráfica, muestra que el nodo en 10 sigue siendo el de menor RMSE, incluso, si se sube a más de 10, el RMSE se incrementa. En este caso el MAE fue de: 0.09132121 (60677661 COP)

****

1. **Primer modelo con árboles (RPart)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En este caso el MAE fue de: 0.1944964 (125098502 COP)

1. **Boosting Model I**

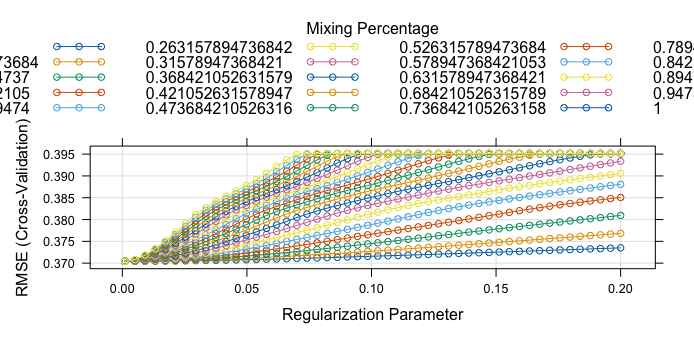
Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a hiper parámetros se tiene una learn\_rate de (0.001); un ntrees de (100); un max\_depth de (6); un min\_rows de (70) y un col\_sample\_rate de (0.2). En este caso el MAE fue de: ­­0.1324681 (88168556 COP)

1. **Boosting Model II**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a hiper parámetros se tiene una learn\_rate de (0.1), (0.01) y (0.001); un ntrees de (50), (100) y (500); un max\_depth de (10); un min\_rows de (70) y un col\_sample\_rate de (0.2). En este caso el MAE fue de: 0.1315953 (84721894 COP)

1. **Elastic net**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. El model es LMNET y el MAE fue de: ­­0.1324681 (88168556 COP)



A continuación se muestra el resumen de los datos obtenidos para cada modelo:

|  |  |  |
| --- | --- | --- |
| Modelo | MAE (logaritmo) | MAE (COP) |
| Linear model (LM) | 0.1336067 | 87521698 |
| Leave One Out Cross Validation (LOOCV) | 0.1336067 | 87521698 |
| Leave One Out Cross Validation (LOOCV) – II | 0.1609716 | 88643254 |
| Ranger I | 0.08934838 | 59397009 |
| Ranger II | 0.0826849 | 54991319 |
| Ranger III | 0.09132121 | 60677661 |
| RPart | 0.1944964 | 125098502 |
| Boosting – I | 0.1337077 | 88910529 |
| Boosting – II | 0.1315953 | 84721894 |
| Elastic net | 0.1324006 | 84836876 |

Si bien el menor MAE en la presente tabla aparece en el modelo Ranger II, al comparar los datos en la lista en Kaggle, el modelo que mejor se aproximó a los datos reales es el Boosting I.

1. **Conclusiones y recomendaciones**

* Haber utilizado datos geográficos como predictores ha sido útil para imputar datos a la base de datos que permite dar robustez al modelo y a los predictores elegidos;
* El modelo de Ranger ha sido el que menor MAE arrojó; no obstante, eso no se traduce en una mejor predicción, de hecho, ha sido el Boosting el modelo que, al compararlo con los datos reales en la plataforma Kaggle, tuvo un menor error de predicción;
* Para los modelos Boosting hay algunas ajustes que permiten mejorar la capacidad predictiva de un modelo, esto es, aumentar el número de árboles, permitir que el modelo tenga más capacidad para capturar patrones complejos en los datos reducir la tasa de aprendizaje, ajustar la profundidad máxima de los árboles;
* El resultado de las predicciones se pudo haber visto afectado por la heterogeneidad que existe en el sector de Chapinero en Bogotá; lo anterior, dado que se trata de una localidad que tiene muchos contrastes, al tener barrios de diferentes estratos socioeconómicos. Esto podría dificultar a los modelos identificar patrones correctamente en los datos.

Bla bla bla

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. 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-3)