**Problem Set 3. Predicción Precios de viviendas en Chapinero (Bogotá) y el Poblado (Medellín).**

**Grupo 5. Presentado por: Carlos Avilán y Francisco Ortiz.**

**Enlace GitHub:** [**https://github.com/faor10/Problem-set-3/**](https://github.com/faor10/Problem-set-3/)

**Introducción**

La industria de la finca raíz es una de las más importantes en la estabilidad y crecimiento económico de los países por la creación de trabajo que supone y por el flujo de dinero en las transacciones. De acuerdo con lo dicho con anterioridad, para las compañías en esta industria es esencial poder realizar predicciones de los precios de las viviendas en busca de generar unas mejores ventas y de no infravalorar estos precios. Unas predicciones acertadas de los precios de las viviendas implican unas mejores ventas y un crecimiento de la industria más acelerado. En este ejercicio se buscó predecir de la mejor manera los precios de viviendas en Chapinero (Bogotá) y el Poblado (Medellín) donde se encontró que el modelo Superlearner (glmnet, ranger, Lm y Mean) es el mejor modelo de predicción al tener la varianza más baja de todos los modelos construidos.

**Datos**

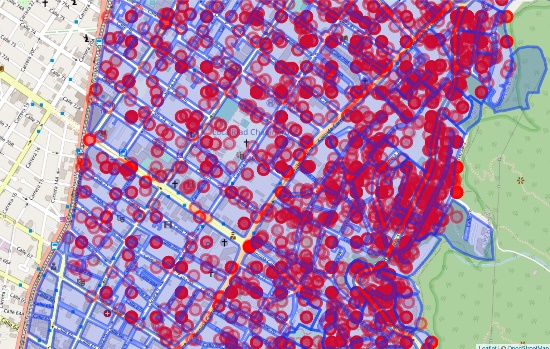
En la tabla 1 se presentan las estadísticas descriptivas de los datos utilizados para entrenar nuestros modelos.

Tabla

Descripción generada automáticamente

Tabla 1. Estadísticas descriptivas Train

En las gráficas 1 y 2, los puntos rojos indican las viviendas utilizadas para el entrenamiento. Este se realizó con 13473 observaciones para Bogotá (solo Chapinero) y 21356 para Medellín. En el mapa de la gráfica 1 se pueden observar las manzanas que se utilizaron para predecir las superficies de algunos apartamentos. Algunas variables se obtuvieron directamente de la base de datos, otras fueron estimadas a partir de datos espaciales Open Street Maps (min\_dist\_bus y min\_dist market) y otras de la descripción de cada una de las propiedades (new\_surface y balcon\_terr).

 Mapa

Descripción generada automáticamente

Gráfica . Mapa de Chapinero (datos train). Gráfica 2. Mapa de Medellín (datos train).

Las variables utilizadas para entrenar nuestro modelo fueron las siguientes:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Detalle** | **Descripción** |
| Price | Precio de los apartamentos | En el caso de Bogotá podemos observar que el precio promedio de los apartamentos es mayor (1300 millones COP para Chapinero) que para Medellín es (405 millones COP) |
| Bedrooms | Número de cuartos | En Bogotá el promedio de cuartos es 2.6 y en Medellín 3.6. |
| New\_surface | Esta variable fue estimada a partir de las variables surface\_total y surface\_covered. Y en aquellos donde existían *missing values* se estimaron a partir de la descripción de las viviendas y después del promedio de los inmuebles vecinos. | En Bogotá el promedio fue 164 m2 y en Medellín 134 m2 |
| Property\_type | Tipo de vivienda: Casa o Apartamento | En Bogotá en su mayor21ía son apartamentos, 96 % del total. En Medellín los apartamentos constituyen el 77%. |
| Min\_dist\_bus | Mínima distancia a una estación de bus. Variable estimada a partir de la estación de bus más cercana. | En Bogotá está distancia es de 869 m en promedio y en Medellín es de 994 m. |
| Min\_dist\_market | Mínima distancia a un supermercado. Variable estimada a partir del supermercado más cercano al inmueble. | Tanto en Bogotá como en Medellín el promedio al supermercado mas cercano es de 1 km. |
| Balcon\_terr | Variable dummy que indica si el inmueble cuenta con terraza o con balcón. Esta variable fue estimada a partir de la descripción del inmueble. | La proporción de apartemos con balcón o terraza |
| Ciudad | Ciudad donde se encuentran los inmuebles. | 13473 inmuebles en Bogotá y 21356 para Medellín. |

Tabla 2. Variables utilizadas para entrenar nuestros modelos.

**Modelos y resultados**

Para poder realizar la predicción de los precios de las viviendas el primer procedimiento realizado fue **entrenar nuestro modelo con** **la siguiente forma funcional:**

(***Esta misma forma funcional se utilizó para los 5 modelos realizados en el procedimiento).***

Se decidió realizar **5 diferentes modelos de entrenamiento** para así encontrar el mejor modelo predictivo, estos cinco modelos fueron: **OLS, Ridge, Lasso, Elastic Net y Superlearner**. Los **modelos Ridge y Lasso** se construyeron a partir del **Lambda mínimo óptimo** y el **modelo Superlearner** se decidió construirlo en base a **Random Forest, Elastic Net**.

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelos** | | **MSE** | **Lambda Óptimo** |
| Modelo 1 | Estimación con OLS | 3.014304e+17 | No aplica |
| Modelo 2 | Estimación con Ridge – Alpha=0 | 3.023e+17 | 45016048 |
| Modelo 3 | Estimación con Lasso – Alpha=1 | 3.014e+17 | 1407074 |
| Modelo 4 | Estimación con Elastic Net | 3.011238e+17 | 900321 |
| Modelo 5 | Superlearner (inluye glmnet, ranger, Lm, mean) | 1.612468e+17 | No aplica |

Tabla 3. Modelos utilizados en el problema de clasificación.

Como se observa en la 5abla 3, el mejor modelo de predicción del precio de las viviendas en Chapinero (Bogotá) y el Poblado (Medellín) es a partir del **Modelo 5 – Superlearner (incluye glmnet, ranger, Lm, mean**). Este es el mejor modelo ya que, **es el modelo con menor MSE (1.612468e+17)** lo cual se puede interpretar como el modelo que mejor predice el precio y el que menor variación presenta dentro de los datos (menor varianza).

|  |  |  |
| --- | --- | --- |
| **Superlearner** | **Riesgo** | **Coeficiente** |
| SL.mean | 6.142570e+17 | 0 |
| SL.lm | 3.014159e+17 | 0 |
| SL.ranger | 1.620275e+17 | 1 |
| SL.glmnet | 3.014292e+17 | 0 |

Tabla 4. Modelo Superlearner

En la Tabla 4 podemos apreciar los resultados del modelo Superlearner, encontrando que el modelo ranger (Random Forest) es el modelo que mejor predice los precios de las viviendas con un coeficiente agregado de 1 mientras que los demás modelos (mean, lm y glmnet) obtuvieron un coeficiente de 0, es decir, de poca importancia en el modelo Superlearner. Un análisis valioso al realizar este modelo es que, al ponderar los algoritmos individuales, se encontró que el único algoritmo relevante para esta predicción del precio es el proveniente de Random Forest.

Gráfico, Gráfico de líneas, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfica 3. Boxplot de las predicciones por ciudad Gráfica 4. Coeficientes del Superlearner

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mínimo** | **1st cuantil** | **Mediana** | **Promedio** | **3rd cuantil** | **Máximo** |
| 1.219e+08 | 3.738e+08 | 5.152e+08 | 6.309e+08 | 7.413e+08 | 3.272e+09 |

Tabla 5. Estadísticas descriptivas predicciones del precio

La gráfica 3 nos muestra un boxplot de las predicciones del modelo por ciudad analizándose que los precios de las viviendas en Chapinero (Bogotá) y las del Poblado (Medellín) son muy similares pues el precio de la mayoría de estas está por debajo de los 1,000,000,000 COP, pero siendo la mediana de los precios mayor en Chapinero (Bogotá). Ahora bien, este gráfico nos ayuda a determinar que en el Poblado (Medellín) es donde se encuentran las viviendas más baratas y caras.

La tabla 5 nos muestra las estadísticas descriptivas de las predicciones de los precios de las viviendas pudiéndose determinar de que el precio mínimo predicho es de 121,900,000 COP mientras que el precio máximo predicho es de 3,272,000,000 COP. El promedio de los precios predichos es de 630,900,000 COP lo cual se encuentra acorde con el precio promedio de las viviendas en nuestra base train (763,300,000 COP).

**Conclusiones y recomendaciones**

Para realizar estas predicciones se utilizaron nuevos métodos para poder determinar algunas de las variables predictoras como fue el uso de datos espaciales y el uso de texto como datos. De esta manera confirmamos que Big Data no solo se refiere a grandes volúmenes de datos sino a datos de alta complejidad. Posteriormente el uso de técnicas de predicción avanzadas como Superlearnes con el cual obtuvimos una mejora significativa en las predicciones en comparación a los métodos comunes.