***PROBLEM SET 3: MAKING MONEY WITH ML***

BIG DATA AND MACHINE LEARNING

Jesús Cepeda - Arturo Trujillo - Juan Camilo Martínez - Miguel Contreras

Repositorio: [MiguelContreras1/Taller-3 (github.com)](https://github.com/MiguelContreras1/Taller-3)

**INTRODUCCIÓN**

Una de las muchas aplicaciones del machine learning se presenta en el sector inmobiliario, donde pueden resultar muy útiles los modelos para la predicción de precios de viviendas. En este trabajo se busca predecir precios de este tipo de inmuebles en Cali. Hay que recordar que estos inmuebles son bienes diferenciados, los cuales se ofertan y demandan en un mercado competitivo. De acuerdo con la teoría de precios hedónicos, dicho precio dependerá de atributos relevantes, es decir, características propias observables inherentes del inmueble, pero también hay otras características que no son propias de las viviendas, y que influyen en su precio, como las condiciones del entorno, las cuales deben tenerse en cuenta al elaborar modelos con el fin de mejorar la predicción de éstos. Teniendo esto en cuenta, a la información de las bases de datos proporcionadas (train y test), se le agregó información espacial proveniente de Open Street Map, para construir variables adicionales.

Como no se conoce la forma funcional exacta que determina los precios, resultan útiles los modelos que capturen no linealidades y puedan encontrar la mejor forma funcional, es decir, el patrón subyacente que mejora la predicción. Este documento expone los resultados de los modelos evaluados (ols cv, lasso, ridge, elastic net, tree, random forest, bagging, Xgboots, superlearners). Tras el proceso de depuración y entrenamiento de los modelos se evidenció que aquel con mejor desempeño fue el xxxxx, logrando tener el menor RMSE (xxxx).

**DATOS**

Se utilizaron datos provenientes de las bases de datos train (51.437 observaciones y 12 variables, de Bogotá y Medellín) y test (5.000 observaciones y 12 variables, de Cali). De estas se tomaron las variables de características inherentes a la vivienda, como los metros cuadrados de superficie, el número de habitaciones, el número de baños, el tipo de propiedad y la ciudad. Además, se usaron datos de fuentes externas como Open Street Map para la construcción de variables predictoras relevantes, por ejemplo, la cercanía a bienes públicos, como parques, centros comerciales, paraderos de buses, centros de salud, estaciones de policía, bancos y bares. También se utilizaron predictores que se obtuvieron a través de la descripción de los inmuebles, como el estrato, el piso en que está ubicado, si cuenta con ascensor y con parqueadero.

Tras la exploración inicial se pudo determinar que 37.985 viviendas están ubicadas en Bogotá, 13.452 en Medellín y 5.000 en Cali. En cuanto a los datos faltantes en la base de datos train se encontraron 39.044 en la variable surface\_total, 41.745 en surface\_covered, 24.915 en rooms, y 15.032 en bathrooms. Por su parte, en la base de datos test los missing values fueron 3.146 en surface\_total, 3.569 en surface\_covered, 1.837 en rooms, y 1.478 en bathrooms. Para recuperar información de estas variables relevantes se efectuó una captura de información la variable “description” en ambas bases. Con información de la descripción se crearon variables de estrato, piso, ascensor y parqueadero, además de lograr reducir los porcentajes de valores faltantes de las bases train y test, en las variables surface\_total (-xx % y -xx %), surface\_covered (-xx % y -xx %), romos (-xx % y -xx %), y bathrooms (-xx % y -xx %).

A pesar de la gran cantidad de información recuperada, aún persistía un número importante de valores faltantes en variables determinantes, lo que llevó a que fuera necesario realizar un tratamiento de estas, con el fin de darles un valor para no afectar las predicciones. Dicho tratamiento consistió en (xxxxx Imputar la media, o knn a los valores faltantes de superficie total\* o de superficie cubierta\*, también imputar en baños xxxxx). Vecinos espaciales. En este caso, se le asignaron valores a un total de xxx viviendas. Finalmente, la base train quedó conformada por xxxx propiedades, mientras que test por xxxx.

A grandes rasgos se puede observar que, tras la depuración las viviendas de Bogotá tienen precios entre $xxxx y $xxxx, en su mayoría son apartamentos (xx %), tienen una media de xx habitaciones y xx baños. Además, la mayoría de encuentran en estrato xx (xx %). En Medellín los precios van desde $xxxx hasta $xxxx. La mayoría son apartamentos (xx %), la media de habitaciones y baños es la misma que en Bogotá, el mayor porcentaje (xx %) de inmuebles pertenecen al estrato xx. Por su parte, en Cali se pudo observar que, el porcentaje de apartamentos si bien es mayoría (xx %), es menor respecto a las otras dos ciudades, mientras que la media de habitaciones (xx) y de baños (xx), es la misma. La mayoría de los inmuebles son de estrato xx (xx %).

En lo referente a la construcción de variables que indican cercanía de los inmuebles a bienes públicos que pueden incidir en su precio, se utilizó el Sistemas de Coordenadas Geográficas, crs=4326. Gracias a que se contaba con información de latitud y longitud de las casas y apartamentos, fue posible crear las variables dista\_park, dista\_bus, dista\_salud, dista\_policia, dista\_banco, y dista\_bar.

A continuación, se ve un cuadro resumen con las variables seleccionadas para los modelos, así como los mapas que muestran la ubicación de los inmuebles y variables espaciales.

2 tablas estadísticas descriptivas (de base train y test, depuradas y con las variables creadas)

3 mapas (uno por cada ciudad)

**MODELOS Y RESULTADOS**

Tras el tratamiento de los datos se estableció que el modelo base para determinar el precio tendría la siguiente estructura:

Posteriormente se corrieron los xx modelos descritos anteriormente, y el de mejor desempeño fue el modelo xxxx, el cual logró tener un MSE de xxxxxxx. Este tipo de modelos xxxx (colocar la descripción del modelo que sale en el libro an introduction of statistical learning)

Tabla con resumen de los 12 modelos (Modelo, RMSE, RSQUARED, MAE)

Es de resaltar que las variables que tienen mayor peso a la hora de predecir el precio, dado su nivel de significancia, en Cali son xxxxx, mientras que en Bogotá y Medellín son xxxxx.

**CONCLUSIONES Y RECOMENDACIONES**

Del ejercicio realizado se puede concluir que, las aplicaciones del machine learning son un instrumento muy útil en el campo inmobiliario. En especial, los modelos de regresión para predicción de precios de inmuebles, siempre que sean suficientemente precisos y tomen en cuenta variables relevantes, resultan particularmente útiles para lograr mayores rentabilidades. En este caso, para Cali se encontró un diferencial de $xxxxx entre el modelo más acertado y menos acertado. Este es un monto considerable que puede tener incidencia en la viabilidad de la operación de la empresa de compraventa de propiedades que nos contrató para el análisis.

También hay que resaltar que la teoría de los precios hedónicos se cumple en este ejercicio, pues las variables del entorno como xxxxxx resultaron muy relevantes a la hora de predecir los precios de las viviendas.

**BIBLIOGRAFÍA**

An Introduction of Statistical Learning

Open Street

Properaty

Dane

**ANEXOS**