

TALLER 2 - CDA

Diego Felipe Carvajal Lombo - 201911910

Jesús Manuel Ospino Bernal - 201915195

En el siguiente informe, se presentan los hallazgos clave obtenidos del análisis de datos de la venta de bienes raíces.

1. Entendimiento y preparación de los datos

En primer lugar, se realizó una exploración de la calidad de los datos donde se corrigieron inconsistencias de los datos con la realidad del negocio. Se separó el conjunto de datos en 3 diferentes datasets. El primero para entrenar los modelos de regresión, el segundo para compararlos y poder seleccionar el mejor, y el tercero para poder realizar el análisis de resultados del modelo seleccionado.

Para poder hacer la exploración de los datos, primeramente, se utilizó la herramienta Pandas Profiling, la cual nos posibilita obtener un reporte detallado de cada columna, y así, obtener una primera intuición de los mismos.

Luego de esto, se revisó columna por columna, tanto las categóricas como las no categóricas, graficándolas y mirando su distribución y detalles, para ver si eran seleccionables para el modelo de regresión, requerían algún ajuste o se podían usar directamente. Dentro de estos hallazgos, se decidió solo utilizar los registros cuyo tipo de operación es Venta o Arriendo y venta, esto dada los requerimientos que necesita el negocio.

También, dado que el dataset estaba sesgado a estratos altos, se eligió juntar los estratos inferiores a 3 en una sola categoría. Igualmente, en la columna de antigüedad, se agruparon las categorías que no tenían antigüedad, dado que juntas no representaban ni un 3% de los datos. En tercer lugar, también se eliminaron los registros que presentaban datos no correspondientes con la vida real, por ejemplo, un número negativo de parqueaderos. Por último, también se filtran aquellos registros cuyo valor de precio venta era exorbitado, dado que primero no existen casas con valores de 12 dígitos en Colombia, o son outliers muy específicos y estos afectan el rendimiento del modelo. Los detalles de la escogencia o no de cada columna están en el notebook.

Al finalizar la tarea de entendimiento y preparación de los datos, seleccionamos únicamente las siguientes variables: área, habitaciones, baños, parqueaderos, localidad, estrato, antigüedad.

2. Entrenamiento del modelo de los Machine Learning

Para poder realizar el entrenamiento de los modelos de regresión, se creó un pipeline que aplica OneHotEncoding a todas las variables categóricas seleccionadas y a las numéricas realiza un SimpleImputer con el mean.

Se seleccionaron 2 modelos, el primero es una regresión lineal sencilla, y el segundo es un random forest regression. Para el segundo se hizo búsqueda de hiperparámetros para `n_estimators`, `max_depth` y `min_samples_split`, Usando un GridSearch, de los que se obtuvieron los siguientes resultados:

`max_depth: 12, min_samples_split: 5, n_estimators: 200`

3. Análisis cuantitativo de resultados del modelo

En primer lugar, es necesario dimensionar los valores en los que están los precios de venta de las casas de nuestro registro. Se sacó el promedio de venta de casa en el conjunto de prueba, el cual se obtuvo un valor de \$1'065.201.500 COP. El valor máximo era de: , mientras que el valor mínimo era de:

- Regresión Lineal

Se obtuvieron las siguientes métricas:

| MÉTRICA | TRAIN | TEST | VAL |
|----------------|-------------------|-------------------|-------------------|
| RMSE | \$ 475.786.983,71 | \$ 458.855.038,77 | \$ 484.594.304,67 |
| MAE | \$ 270.228.055,82 | \$ 267.328.055,43 | \$ 265.356.630,86 |
| R ² | 0,80 | 0,80 | 0,80 |

De esto, significa que para valores (inmuebles) nuevos, el promedio de error en la predicción de su valor de venta va a estar, en promedio, de 265 millones de pesos. Lo que, para la escala en la que estamos manejando, es un valor alto, cercano al 26% de error, sin embargo, vemos que el RMSE es considerablemente alto, lo que significa que hay errores grandes en los casos más atípicos (outliers), lo que para casas muy caras o baratas no predice muy bien.

Además de esto, vemos que el R² de 0.8, lo que significa que el 80% de la variabilidad de los precios es explicada. Es decir, el modelo puede capturar las diferencias y explicarlas con las variables.

Por último, vemos que las métricas no varían entre los conjuntos de train y test, por lo que su capacidad de generalización es buena.

- Random Forest Regression

| MÉTRICA | TRAIN | TEST | VAL |
|----------------|-------------------|-------------------|-------------------|
| RMSE | \$ 257.094.868,10 | \$ 396.773.939,74 | \$ 378.393.307,33 |
| MAE | \$ 136.596.230,76 | \$ 186.153.028,01 | \$ 177.053.381,49 |
| R ² | 0,94 | 85,00 | 0,87 |

De este modelo, podemos ver que mejoró considerablemente con respecto al anterior. En primer lugar, vemos que ahora el error promedio está en 177M, lo que significa un error

aproximado del 17% en la escala que manejamos, por lo que es mucho más acertado prediciendo.

En segundo lugar, el RMSE es de 378M, lo que significa que el modelo falla mucho menos en casos atípicos que antes, entonces los outliers afectan menos. También es notable la mejoría de la explicación de la variabilidad de los precios, mejorando a un 87%.

Sin embargo, vemos que ya hay una diferencia más significativa entre las métricas de train y los otros conjuntos, lo que me implica que, su capacidad de generalización es inferior. Aun así, este modelo es mucho mejor que el anterior, por lo que va a ser el seleccionado.

4. Análisis cualitativo de resultados del modelo

Ahora, desde una perspectiva por variable, de acuerdo con los valores Shapley, se evidencia una relación positiva entre la cantidad de metros cuadrados y el valor de avalúo del inmueble, de igual manera ser de la localidad de Chapinero influye positivamente en valor conjunto a si es estrato 6 el sector del inmueble. Por otro lado, se evidencia que, si un inmueble cuenta con más años de antigüedad su valor disminuye, mientras que si el inmueble está en un rango entre 0-5 años de antigüedad se encarece. Por último, el número de parqueaderos y baños también impactan positivamente al costo de la vivienda.

5. Generación de valor

Para el cálculo de generación de valor debemos tener en cuenta que un perito cobra \$9.500 la hora y una visita son de 6 horas, por lo tanto, cada inmueble que es revisado cuesta \$57.000. Como HabitAlpes es capaz de revisar hasta 500 al mes, el costo de un mes de peritos puede llegar a ser de: \$28.500.000. Ahora, con el algoritmo propuesto se prevé que el sólo se requerirá un peritaje de una hora, por lo que los costos mensuales podrán alcanzar el valor de \$4.750.000. Esto representa un ahorro de \$23.750.000 mensuales.

Por otro lado, si el algoritmo de Machine Learning estima 20M por debajo del valor del inmueble, el cliente solicita avalúo presencial, mientras que la sobreestimación no es reportada. Para nuestro caso específico, encontramos que, de los valores predichos por el algoritmo se reportarían cerca del 36% de predicciones, por lo cual se requeriría la presencia del perito en 180 de los casos de los 500; es decir, se incurriría en un gasto de $\$9.500 \times 5 \text{ horas} = \$47.500 \times 180 \text{ casos} = \$8.550.000$. Concluyendo que el ahorro real de la implementación del algoritmo de predicción sería de \$15.200.000 mensuales.

En otras palabras, para cada predicción el valor a ahorrar sería \$47.500, el porcentaje de reporte es del 36% y si un usuario reporta el costo asociado es de \$47.500, por lo tanto, el ahorro por predicción sería de $\$30.400 = \$47.500 - 0.36 \times \$47.500$. Esto equivale a un ahorro del 53.33% por visita del perito.

Ahora bien, los costos asociados a la implementación del modelo son los siguientes. Según el promedio de salario de un científico de datos en Colombia es de \$5.000.000, lo que diariamente equivale a \$28.410. Se asume que el tiempo de implementación es de dos semanas y que fueron necesarios 2 científicos de datos, por lo que el costo asociado al modelo es de \$795.480.

Tomando en cuenta lo anterior, el ROI calculado es de:

$$ROI = \frac{15.200.000 - 795.480}{795.480} \cdot 100 = 1810.79\%$$

Este modelo ofrece un ROI de 1810% con una precisión de 65%. Adicionalmente, nos interesa conocer el punto de equilibrio de la implementación del modelo; es decir, el momento en que el valor ahorrado supera los costos asociados al modelo.

$$\text{Punto equilibrio} = \frac{795.480}{15.200.000} = 0.052 \text{ meses}$$

Es decir, luego de dos días de aplicación del modelo se recuperaría la inversión.

6. Insights

Del modelo se puede encontrar que es una buena herramienta para hacer predicciones del precio de venta de las casas, dado que su margen de error no es alto, en la escala en la que se encuentran sesgadas las viviendas. Sin embargo, si es necesario tener atención con los casos de valores muy extremos (muy baratos o caros).

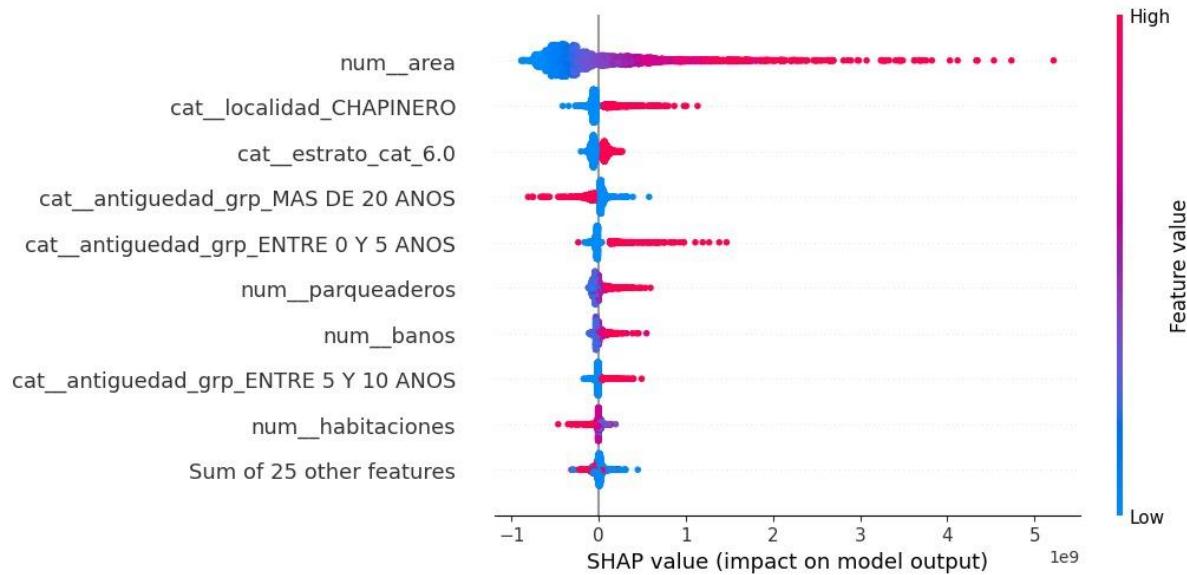
También es posible analizar que las variables que más afectan al modelo son:

1. Área
2. Localidad
3. Antigüedad.

De la primera variable, como es de esperarse, entre mayor sea el área, mayor es el valor del inmueble. De la localidad, podemos encontrar una peculiaridad y es que, si el inmueble está en Chapinero, afecta en una mayor medida que otras variables al aumento del avalúo, por lo que, puede que esta localidad esté sobrevalorada.

De la antigüedad, como es natural pensar, se pierde avalúo a medida que es más viejo la casa, se puede observar que, si se encuentra en la categoría de más de 20 años, influye negativamente al precio, de manera contraria a la categoría entre 0 y 5 años, que influye positivamente.

Por último, vemos que entre más número de parqueaderos y de baños aumenta el valor del bien raíz, como es de esperarse.



En base a lo anterior, se le recomienda a la empresa lo siguiente:

1. No comprar demasiados inmuebles en Chapinero, por lo que este sector tiene un sobreprecio de venta.
2. Comprar viviendas que tengan una antigüedad entre los 5 y 10 años, dado que esta categoría afecta a el precio de venta, pero en una menor medida que los más nuevos, y no afecta negativamente al precio como cuando son tan antiguados.