Universidad de los Andes Colombia | Departamento de Ingeniería Eléctrica y Electrónica

Ciencia de datos aplicada
Taller 2. Punto 5: Generación de valor para HabitAlpes

Francisco Alejandro Santamaría
Dept. de Ingeniería Eléctrica y Electrónica
Universidad de los Andes
Bogotá D.C, Colombia
f.santmaria@uniandes.edu.co

I. Preámbulo. Contexto de negocio de HabitAlpes.

Como consultor científico de datos para HabitAlpes, el objetivo de este entregable es traducir el desempeño de un modelo de estimación de precios inmobiliarios en métricas concretas de negocio. En particular, se busca cuantificar cuánto dinero puede ahorrar la empresa en cada avalúo, cuál es el costo asociado a los errores del modelo y a partir de cuántas estimaciones la inversión en ciencia de datos empieza a generar dividendos netos.

A diferencia de un análisis puramente técnico (centrado en métricas como *RMSE* o R^2), aquí se construye un marco paramétrico que conecta dichas métricas con variables económicas. El resultado final se resume en:

1. Costos de tiempo y dinero asociados a peritos.
2. Ahorro de tiempo teórico al introducir el modelo.
3. Costo esperado de los errores de predicción.
4. Ahorro neto por estimación correcta.
5. Retorno de inversión (ROI) del modelo, incluyendo desarrollo y despliegue.

II. Supuestos de contexto y parámetros de referencia.

Antes de entrar en fórmulas, es necesario fijar algunos supuestos. Estos mezclan información obtenida del *notebook* (a partir del dataset de apartamentos) con datos externos de mercado. El objetivo no es obtener cifras exactas, sino órdenes de magnitud razonables para dimensionar el valor del modelo.

En el *notebook* se exploró la distribución de `precio_venta` y se construyó la variable `precio_m2 = precio_venta/area`. A partir de histogramas y estadísticas descriptivas se observó que:

- La mayoría de inmuebles se concentran en estratos 4–6.
- Los valores de `precio_venta` para estos estratos se ubican, en general, entre 400 y 800 millones de pesos.
- Los valores de `precio_m2` se encuentran típicamente entre 5 y 7 millones de pesos por metro cuadrado.

Estos rangos son coherentes con reportes de mercado para zonas residenciales de Bogotá. Por esta razón se toma como referencia un apartamento típico de aproximadamente 90 m^2 y un valor:

$$V_{\text{apt}} \approx 600\,\text{M COP}.$$

En cuanto al avalúo tradicional, empresas de avalúos y entidades financieras en Colombia suelen cobrar:

- Un porcentaje del valor del inmueble (del orden del 0,1 %), o
- Tarifas fijas en el rango de 450k–650k COP para inmuebles entre 400 y 800 millones.

Tomando un valor intermedio coherente con el rango de precios considerado, se fija el siguiente supuesto:

$$C_{\text{avalúo}} \approx 600{,}000\,\text{COP por inmueble}.$$

Finalmente, para capturar el costo de tiempo del perito se combinan salarios de referencia con un supuesto operativo. Asumiendo un salario de 2,5M COP/mes y 176 horas laborales mensuales, el costo por hora se aproxima como:

$$C_{\text{hora}} \approx 15{,}000\,\text{COP/hora}.$$

Si un avalúo tradicional tarda unas 4 horas efectivas, y se incluye un factor de overhead de 2 (prestaciones, administración), se obtiene:

$$C_{\text{tiempo perito}} \approx 4 \times 15{,}000 \times 2 \approx 120{,}000\,\text{COP}.$$

III. Comparación de procesos y ahorro operativo.

Una vez definidos los parámetros, se comparan dos formas de trabajo: el proceso tradicional sin modelo y el proceso asistido por el modelo de *machine learning*.

III-A. Escenario sin modelo.

En el proceso actual, para cada nuevo apartamento HabitAlpes debe solicitar un avalúo completo a un perito externo. El costo por estimación se resume como:

$$C_{\text{actual}} \approx C_{\text{avalúo}} = 600{,}000\,\text{COP}.$$

III-B. Escenario con modelo de estimación.

En el nuevo escenario, el modelo entrenado en el *notebook* genera una primera estimación de precio a partir de variables como `area`, `estrato`, `localidad`, `distancia_estacion_tm_m`, `distancia_parque_m` y amenidades. El perito pasa a trabajar como revisor especializado.

Se asume que:

- El tiempo promedio del perito baja de 4 horas a 1 hora:

$$C_{\text{tiempo con modelo}} \approx 30{,}000\,\text{COP}.$$

- El costo de operar el modelo (infraestructura, monitoreo, almacenamiento) es:

$$C_{\text{infra}} \approx 20{,}000\,\text{COP por estimación}.$$

El costo total por estimación en este escenario es:

$$C_{\text{con ML}} = C_{\text{tiempo con modelo}} + C_{\text{infra}} \approx 50{,}000\,\text{COP}.$$

III-C. Ahorro teórico por estimación.

El ahorro bruto por avalúo al pasar de un proceso tradicional a uno asistido por el modelo se calcula como:

$$A_{\text{bruto}} = C_{\text{actual}} - C_{\text{con ML}}.$$

Reemplazando:

$$A_{\text{bruto}} \approx 600{,}000 - 50{,}000 = 550{,}000\,\text{COP por inmueble}.$$

Si HabitAlpes realiza N avalúos al año, el ahorro bruto anual es:

$$A_{\text{bruto anual}} = N \times A_{\text{bruto}}.$$

Por ejemplo, para $N = 1{,}000$ estimaciones anuales:

$$A_{\text{bruto anual}} \approx 1{,}000 \times 550{,}000 = 550\,\text{M COP/año}.$$

IV. Costo esperado de los errores del modelo.

El modelo reduce tiempos, pero introduce un riesgo: los errores de predicción. Algunos son pequeños y fáciles de corregir; otros pueden ser mayores y afectar la fijación de precios si se pasan por alto.

En el conjunto de *validation* se dispone de pares $(y, \hat{y})$ de precio real y predicho. A partir de estos datos se calcula el error relativo:

$$\text{err\_rel} = \frac{\hat{y} - y}{y},$$

y se clasifica cada predicción en:

- **Buena:** $|\text{err\_rel}| \leq 10\,\%$.
- **Moderada:** $10\,\% < |\text{err\_rel}| \leq 20\,\%$.
- **Grande:** $|\text{err\_rel}| > 20\,\%$.

A partir de la distribución de *validation* se pueden estimar las proporciones P_{buena}, P_{mod} y P_{grande} (con $P_{\text{buena}} + P_{\text{mod}} + P_{\text{grande}} = 1$). Además, se modela la capacidad del perito para corregir errores mediante:

- corr_{mod}: porcentaje de errores moderados corregidos.
- $\text{corr}_{\text{grande}}$: porcentaje de errores grandes corregidos.

Para cuantificar el impacto económico se asocia a cada banda un costo potencial:

$$C_{\text{mod}} = \alpha_{\text{mod}} V_{\text{apt}}, \quad C_{\text{grande}} = \alpha_{\text{grande}} V_{\text{apt}},$$

donde α_{mod} y α_{grande} son porcentajes promedio de pérdida en los casos no corregidos.

El costo esperado por tipo de error se resume en la Tabla I.

Cuadro I: Matriz de costos esperados por tipo de error del modelo

Tipo	Prob.	Costo potencial	% corregido	Costo esperado
Buena	P_{buena}	≈ 0	≈ 1	≈ 0
Moderada	P_{mod}	C_{mod}	corr_{mod}	$P_{\text{mod}}(1-\text{corr}_{\text{mod}})C_{\text{mod}}$
Grande	P_{grande}	C_{grande}	$\text{corr}_{\text{grande}}$	$P_{\text{grande}}(1-\text{corr}_{\text{grande}})C_{\text{grande}}$

A partir de la matriz, el costo esperado de error por estimación se expresa como:

$$C_{\text{error}} = P_{\text{mod}}(1-\text{corr}_{\text{mod}})C_{\text{mod}} + P_{\text{grande}}(1-\text{corr}_{\text{grande}})C_{\text{grande}}.$$

En el marco del taller, y buscando un valor razonable que capture frecuencia y gravedad de errores, se adopta como referencia:

$$C_{\text{error}} \approx 50{,}000\,\text{COP por estimación}.$$

V. Ganancia neta, punto de equilibrio y ROI.

La ganancia neta por estimación se define como:

$$G_{\text{neto}} = A_{\text{bruto}} - C_{\text{error}}.$$

Con los valores de referencia:

$$G_{\text{neto}} \approx 550{,}000 - 50{,}000 = 500{,}000\,\text{COP por estimación}.$$

Para estimar el punto de equilibrio se fija una inversión inicial asociada al desarrollo y despliegue del modelo:

- Desarrollo del modelo (científico de datos): $I_{\text{DS}} \approx 36\,\text{M COP}$.
- Ingeniería de datos / MLOps: $I_{\text{MLOps}} \approx 24\,\text{M COP}$.
- Infraestructura inicial, licencias y soporte: $I_{\text{infra}} \approx 10\,\text{M COP}$.

Por tanto:

$$I_{\text{total}} = I_{\text{DS}} + I_{\text{MLOps}} + I_{\text{infra}} \approx 70\,\text{M COP}.$$

El número de estimaciones necesario para recuperar la inversión es:

$$N_{\text{BE}} = \frac{I_{\text{total}}}{G_{\text{neto}}} \approx \frac{70\,\text{M}}{500{,}000} \approx 140 \text{ estimaciones}.$$

Si HabitAlpes realiza N estimaciones al año, la ganancia neta anual es:

$$G_{\text{anual}} = N \times G_{\text{neto}}.$$

El ROI se calcula como:

$$\text{ROI} = \frac{G_{\text{anual}} - I_{\text{total}}}{I_{\text{total}}}.$$

Por ejemplo, para $N = 1{,}000$:

$$G_{\text{anual}} \approx 1{,}000 \times 500{,}000 = 500\,\text{M COP},$$

$$\text{ROI} \approx \frac{500\,\text{M} - 70\,\text{M}}{70\,\text{M}} \approx 6{,}14,$$

lo que equivale a un retorno aproximado del 614 % anual bajo los supuestos utilizados.

VI. Cierre.

Con el deseo de conectar el trabajo de modelado con decisiones de negocio, este documento presenta un esquema de generación de valor que traduce el desempeño del modelo en ahorros económicos directos para HabitAlpes. A partir de los supuestos fijados se concluye que:

1. El modelo permite reducir de forma significativa el costo por avalúo, al pasar de un proceso tradicional de 600k COP a uno asistido de aproximadamente 50k COP.
2. Incluso incorporando el costo esperado de errores, la ganancia neta por estimación se mantiene en torno a 500k COP.
3. La inversión inicial en desarrollo y despliegue se recupera después de un número relativamente pequeño de estimaciones (alrededor de 140), haciendo atractivo el proyecto desde el punto de vista financiero.

Aunque los resultados numéricos dependen de los supuestos adoptados, el marco paramétrico propuesto permite actualizarlos fácilmente con datos reales y recalcular la ganancia neta, el punto de equilibrio y el ROI, manteniendo siempre una lectura alineada con las necesidades del negocio.

Referencias

[1] Cathy O'Neil and Rachel Schutt. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Inc. 2013.

[2] Daniel Vaughan. Data Science: The Hard Parts. O'Reilly Media, Inc. 2024.

[3] Foster Provost and Tom Fawcett. Data Science for Business. O'Reilly Media, Inc., 2013.

[4] Doug Rose. Data Science: Create Teams That Ask the Right Questions and Deliver Real Value. APress. 2016.