Modelo Predictivo de Precios de Apartamentos HabitAlpes

Análisis de Machine Learning

24 de noviembre de 2025

Índice

1. Introducción

HabitAlpes requiere una herramienta automatizada de valoración de apartamentos en Bogotá que permita a sus clientes obtener estimaciones competitivas basadas en características y ubicación. Este informe presenta el desarrollo de un modelo de machine learning para predecir precios de venta, evalúa su desempeño técnico y analiza su viabilidad económica considerando los costos operacionales actuales y potenciales ahorros.

2. Entendimiento y Preparación de Datos

El dataset original contenía 43,013 apartamentos con 26 variables incluyendo características físicas (área, habitaciones, baños), ubicación (coordenadas, sector), amenidades (gimnasio, piscina, parqueaderos) y precio de venta. Se eliminaron columnas redundantes (IDs, URLs, descripciones) y registros con precio de venta faltante. Las variables booleanas se imputaron con falso y la administración por mediana según estrato.

Se aplicó remoción de outliers mediante IQR (rango intercuartílico), eliminando 3.99 % de registros por área, 2.62 % por administración y 1.24 % por precio de venta. La variable precio de arriendo mostró 98.71 % de outliers y fue excluida. Se removió el percentil 99 de precios para eliminar propiedades de lujo extremo. El dataset final contiene aproximadamente 26,934 apartamentos.

Cabe señalar que, para el alcance de este taller, el análisis se centró exclusivamente en datos estructurados. El dataset suministrado no incluyó imágenes procesables (como fotos de fachada o interiores), por lo que no se realiza un análisis de características visuales en este reporte.

3. Entrenamiento de Modelos

Para garantizar una evaluación robusta, los datos se dividieron en **Entrenamiento (80 %)** y **Prueba (20 %)**. Se aplicó validación cruzada 5-fold sobre el conjunto de entrenamiento para la selección del mejor modelo, donde cada fold utilizó automáticamente 80 % para entrenamiento y 20 % para validación. El modelo final se reentrenó con todo el conjunto de entrenamiento y se evaluó en el conjunto de prueba reservado.

Las variables categóricas se codificaron con One-Hot Encoding y las numéricas se estandarizaron con StandardScaler. Los modelos evaluados fueron Regresión Lineal, Árbol de Decisión (profundidad = 8), Random Forest (300 árboles, profundidad = 12), XGBoost (300 árboles, learning rate = 0.1, profundidad = 8) y KNN (k = 5).

4. Análisis Cuantitativo de Resultados

Cuadro 1: Desempeño de modelos en conjunto de prueba

Modelo	R^2	MAE (M COP)	RMSE (M COP)
XGBoost	0.878	161.9	310.4
Random Forest	0.871	171.7	318.8
KNN	0.832	194.2	364.7
Árbol de Decisión	0.828	205.2	368.9
Regresión Lineal	0.683	306.3	500.1

XGBoost obtuvo el mejor desempeño con $R^2 = 0.878$ y MAE de 161.9 millones de pesos. El RMSE de 310.4 millones, siendo el doble del MAE, indica errores grandes en algunos apartamentos. Random Forest mostró desempeño similar ($R^2 = 0.871$, MAE = 171.7M). KNN obtuvo $R^2 = 0.832$ con MAE = 194.2M, y Árbol de Decisión $R^2 = 0.828$ con MAE = 205.2M. La Regresión Lineal tuvo el desempeño más bajo ($R^2 = 0.683$, MAE = 306.3M), confirmando que el problema requiere métodos no lineales.

El área es el predictor más importante (correlación 0.829). La administración (correlación 0.682) sirve como indicador de calidad y ubicación. La diferencia entre correlaciones de Pearson y Spearman para área (0.493 vs 0.829) y administración (0.393 vs 0.682) confirma relaciones no lineales, explicando por qué XGBoost y Random Forest obtienen mejores resultados que métodos lineales.

5. Análisis Cualitativo de Resultados

5.1. Interpretabilidad LIME

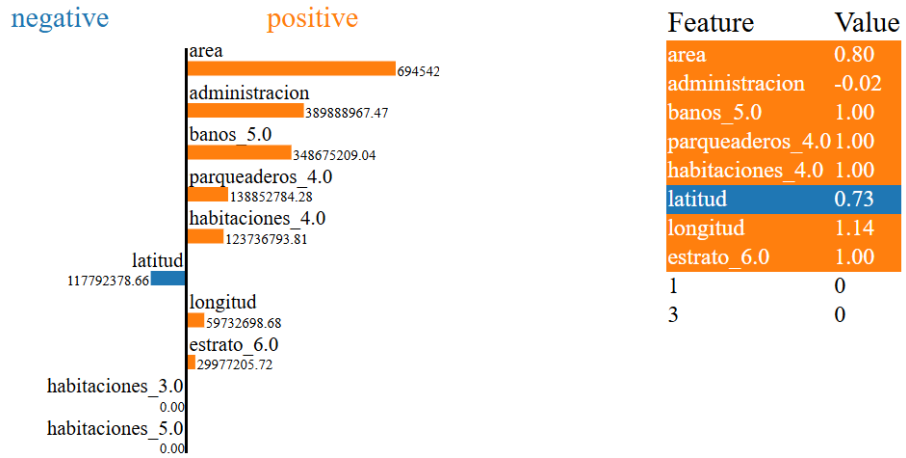

Figura 1: LIME Valores predicción local - valor predicho: 260935916.18

Para esta predicción local, podemos ver que el área tiene un efecto positivo muy fuerte: a mayor tamaño en metros cuadrados, mayor es el precio estimado del inmueble. De forma similar, la administración tiene un impacto positivo considerable, lo que indica que valores altos de administración suelen asociarse con propiedades de mayor valor. También, se observan efectos positivos importantes en variables categóricas como banos_5.0, parqueaderos_4.0 y habitaciones_4.0, es decir, que un apartamento con 5 baños, 4 parqueaderos o 4 habitaciones incrementa significativamente el precio. En contraste, la latitud presenta un efecto claramente negativo, lo que significa que, para esta ubicación específica, estar en esa coordenada geográfica reduce el valor del inmueble; la longitud, por su parte, muestra un efecto positivo muy pequeño. Algunas categorías, como tener 3 o 5 habitaciones, no modifican de manera significativa la predicción del modelo. Finalmente, el estrato 6 tampoco aparece como un factor determinante por sí solo, probablemente porque su influencia ya está capturada indirectamente por otras variables fuertemente correlacionadas, como áreas amplias y administraciones elevadas.

5.2. Interpretabilidad SHAPLEY

Por el lado local:

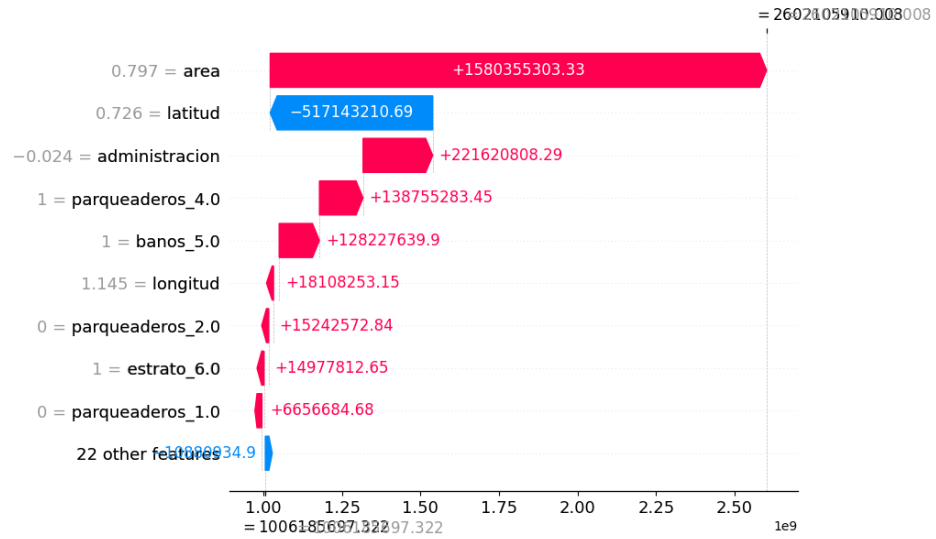

Figura 2: SHAPLEY Valores de predicción local

Para esta predicción local, podemos ver resultados similares a los vistos usando LIME. Por un lado, el área vuelve a ser la variable que ejrce el efecto positivo más fuerte. De forma similar, la administración también contribuye positivamente. Entre las variables categóricas, se destacan impactos positivos importantes como parqueaderos_4.0 y banos_5.0, lo que indica nuevamente que contar con cuatro parqueaderos o cinco baños aumenta significativamente el precio predicho. La diferencia con LIME es la aparición de variables categóricas como parqueaderos_2.0 o parqueaderos_1.0. Estas muestran contribuciones positivas moderadas, mientras que otras variables más pequeñas aparecen agrupadas y no modifican sustancialmente la predicción final.

Ahora, de forma global:

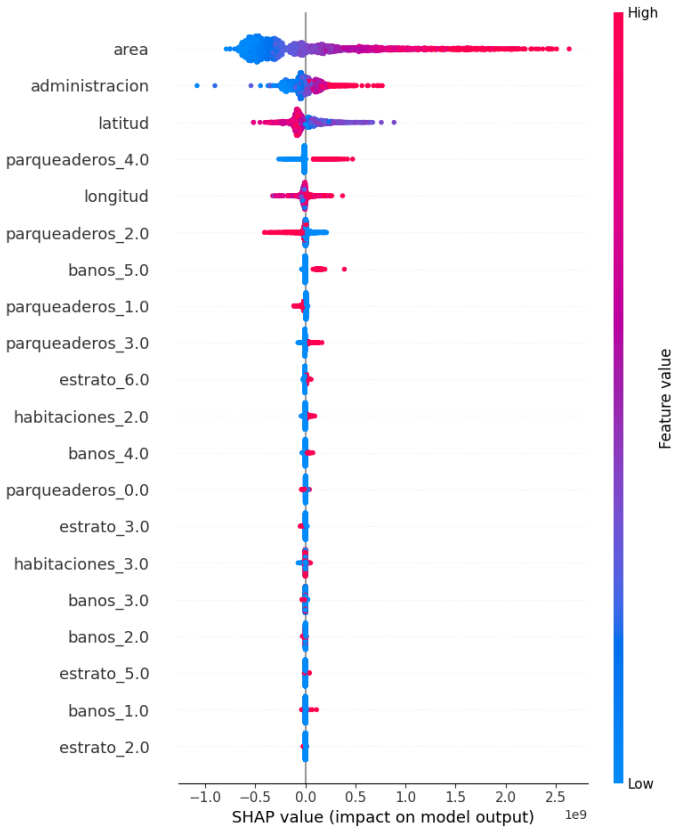

Figura 3: SHAPLEY Valores de predicción global

El análisis global con SHAPLEY confirma y amplía lo observado previamente con las explicaciones locales de LIME. En primer lugar, los valores altos de area tienen un impacto altamente positivo sobre la variable objetivo, evidenciando que el tamaño del inmueble es el componente principal de subida del precio en el conjunto completo de datos. De manera similar, valores altos de administracion se asocian con precios más altos, reflejando que los edificios con cuotas de administración elevadas suelen pertenecer a segmentos más caros. Por otro lado, valores altos de latitud tienden a reducir el precio, mientras que valores bajos lo aumentan, lo cual es coherente con la distribución real de Bogotá: zonas conocidamente caras como Rosales o Chapinero alto presentan latitudes más bajas que sectores del norte extendido. El patrón también muestra que tener 4 parqueaderos tiene un impacto positivo significativo, señal de que este atributo está correlacionado con inmuebles de alta gama. De igual forma, tener 5 baños aumenta el precio, aunque con una menor cantidad respecto a las variables más dominantes.

5.3. Cálculo de Costos y Ahorros

Para cuantificar el impacto, se aplicó la regla de negocio sobre el conjunto de validación: si el modelo subestima el valor real por más de $20 millones, se requiere una revisión presencial costosa. Las demás predicciones se consideran aceptables o éxitos operativos.

Cuadro 2: Matriz de Clasificación de Estimaciones (Proyección mensual 500 apts)

Categoría	**Regla (Error)**	**%**	**Cantidad**
Subestimación Crítica (Falla)	$E < -20\text{M}$	42.77 %	214
Acierto / Subestimación Leve	$-20\text{M} \leq E \leq 0$	20.59 %	103
Sobreestimación (No reportada)	$E > 0$	36.64 %	183

Escenario sin ML (actual):

$$\begin{aligned} \text{Costo por apartamento} &= 6 \text{ horas} \times \$9{,}500/\text{hora} = \$57{,}000 \\ \text{Costo mensual (500 apartamentos)} &= 500 \times \$57{,}000 = \$28{,}500{,}000 \\ \text{Costo anual} &= \$28{,}500{,}000 \times 12 = \$342{,}000{,}000 \end{aligned}$$

Escenario con ML:

Para cada categoría de predicción, el costo varía según si se genera ahorro o se requiere trabajo adicional:

Subestimaciones críticas (42.77 %, 214 apt/mes): Requieren avalúo presencial posterior. El modelo consume 1 hora ($9,500) más el avalúo completo adicional (6 horas, $57,000).

$$\begin{aligned} \text{Costo total por apartamento} &= \$9{,}500 + \$57{,}000 = \$66{,}500 \\ \text{Costo mensual} &= 214 \times \$66{,}500 = \$14{,}231{,}000 \end{aligned}$$

Subestimaciones aceptables y predicciones precisas (20.59 %, 103 apt/mes): El cliente acepta la estimación. Solo se requiere 1 hora del perito.

$$\begin{aligned} \text{Costo por apartamento} &= \$9{,}500 \\ \text{Costo mensual} &= 103 \times \$9{,}500 = \$978{,}500 \end{aligned}$$

Sobreestimaciones (36.64 %, 183 apt/mes): No reportadas por el cliente. Solo se requiere 1 hora del perito.

$$\begin{aligned} \text{Costo por apartamento} &= \$9{,}500 \\ \text{Costo mensual} &= 183 \times \$9{,}500 = \$1{,}738{,}500 \end{aligned}$$

Costos totales con ML:

$$\begin{aligned} \text{Costo mensual total} &= \$14{,}231{,}000 + \$978{,}500 + \$1{,}738{,}500 = \$16{,}948{,}000 \\ \text{Costo anual} &= \$16{,}948{,}000 \times 12 = \$203{,}376{,}000 \\ \text{Ahorro anual bruto} &= \$342{,}000{,}000 - \$203{,}376{,}000 = \$138{,}624{,}000 \end{aligned}$$

5.4. Costo de Desarrollo y Despliegue

Inversión inicial en desarrollo:

Los datos ya fueron recolectados y entregados por HabitAlpes, por lo que no se incluye costo de recolección. Los costos restantes se basan en tarifas del mercado colombiano para profesionales de tecnología. Según datos de 2025, un Data Scientist en Colombia gana en promedio $7.1M mensuales según Indeed [1], con tarifas horarias variables según experiencia. Se utilizan tarifas conservadoras de $50,000-$60,000/hora para este análisis.

Actividad	**Horas**	**Tarifa/hora**	**Costo**
Limpieza y preparación de datos	32	$50,000	$1,600,000
Entrenamiento y evaluación	60	$60,000	$3,600,000
Desarrollo de API e interfaz	100	$60,000	$6,000,000
Pruebas y validación	40	$50,000	$2,000,000
Documentación	20	$40,000	$800,000
Total inversión inicial			**$14,000,000**

Costos anuales de operación:

La infraestructura de nube se estima usando AWS región São Paulo (más cercana a Colombia). Se utiliza SageMaker para simplificar cálculos y administración, aunque esta opción suele ser más costosa que alternativas como EC2 con contenedores. Dado que el modelo no requiere estar activo constantemente, una infraestructura basada en EC2 con instancias on-demand o Lambda podría reducir costos significativamente. Sin embargo, para este análisis se utiliza SageMaker por simplicidad operacional. Se considera una instancia ml.m5.large para inference a $0.115/hora [2] (aproximadamente $82.8/mes o $380,000 COP/mes operando 24/7), aunque con uso intermitente los costos serían menores. Para estimación conservadora se asume $900,000/mes. Adicionalmente, almacenamiento S3 Standard para datos ($0.023/GB) [3] y base de datos RDS para logs y métricas [4].

Concepto	**Costo mensual**	**Costo anual**
SageMaker inference (ml.m5.large)	$900,000	$10,800,000
Almacenamiento S3 (100 GB)	$100,000	$1,200,000
RDS PostgreSQL (db.t3.micro)	$60,000	$720,000
Transferencia de datos y otros	$90,000	$1,080,000
Subtotal infraestructura	$1,150,000	$13,800,000
Mantenimiento (20 hrs/mes × $60,000)	$1,200,000	$14,400,000
Soporte técnico	$200,000	$2,400,000
Total operacional anual	$2,550,000	$30,600,000

5.5. Cálculo de ROI

$$\begin{aligned}\text{Ganancia anual neta} &= \text{Ahorro bruto} - \text{Costos operacionales} \\ &= \$138{,}624{,}000 - \$30{,}600{,}000 = \$108{,}024{,}000\end{aligned}$$

$$\begin{aligned}\text{ROI primer año} &= \frac{\text{Ganancia neta} - \text{Inversión inicial}}{\text{Inversión inicial}} \times 100 \\ &= \frac{\$108{,}024{,}000 - \$14{,}000{,}000}{\$14{,}000{,}000} \times 100 = \mathbf{672\,\%}\end{aligned}$$

$$\text{Ganancia mensual neta} = \frac{\$138{,}624{,}000 - \$30{,}600{,}000}{12} = \$9{,}002{,}000/\text{mes}$$

$$\begin{aligned}\text{Periodo de recuperación (break-even)} &= \frac{\text{Inversión inicial}}{\text{Ganancia mensual neta}} \\ &= \frac{\$14{,}000{,}000}{\$9{,}002{,}000} = \mathbf{1.56\ meses}\end{aligned}$$

El modelo alcanza el punto de equilibrio en **1.6 meses**, recuperando la inversión inicial. A partir del segundo mes, genera ganancias netas de **$9 millones mensuales**.

5.6. Análisis de Sensibilidad

El análisis económico depende del porcentaje de subestimaciones críticas. Se comparan tres escenarios basados en diferentes tasas de error del modelo. Para cada escenario se calculan los costos operacionales mensuales, el ahorro anual neto y el ROI considerando la inversión inicial de $14M y los costos operacionales anuales fijos de $30.6M.

Escenario Optimista (35 % subestimaciones críticas):
Con 500 apartamentos mensuales:

- Subestimaciones críticas: 35 % (175 apt) a $66,500 = $11,637,500
- Predicciones aceptables: 65 % (325 apt) a $9,500 = $3,087,500
- Costo mensual total: $14,725,000

$$\begin{aligned}\text{Costo anual operacional} &= \$14{,}725{,}000 \times 12 = \$176{,}700{,}000 \\ \text{Ahorro bruto anual} &= \$342{,}000{,}000 - \$176{,}700{,}000 = \$165{,}300{,}000 \\ \text{Ganancia neta} &= \$165{,}300{,}000 - \$30{,}600{,}000 = \$134{,}700{,}000 \\ \text{ROI} &= \frac{\$134{,}700{,}000 - \$14{,}000{,}000}{\$14{,}000{,}000} \times 100 = \mathbf{862\,\%}\end{aligned}$$

Escenario Base Real (42.77 % subestimaciones críticas):

Con 500 apartamentos mensuales:

- Subestimaciones críticas: 42.77 % (214 apt) a \$66,500 = \$14,231,000
- Predicciones aceptables: 57.23 % (286 apt) a \$9,500 = \$2,717,000
- Costo mensual total: \$16,948,000

$$\begin{aligned}
\text{Costo anual operacional} &= \$16{,}948{,}000 \times 12 = \$203{,}376{,}000 \\
\text{Ahorro bruto anual} &= \$342{,}000{,}000 - \$203{,}376{,}000 = \$138{,}624{,}000 \\
\text{Ganancia neta} &= \$138{,}624{,}000 - \$30{,}600{,}000 = \$108{,}024{,}000 \\
\text{ROI} &= \frac{\$108{,}024{,}000 - \$14{,}000{,}000}{\$14{,}000{,}000} \times 100 = \mathbf{672\,\%}
\end{aligned}$$

Escenario Pesimista (50 % subestimaciones críticas):

Con 500 apartamentos mensuales:

- Subestimaciones críticas: 50 % (250 apt) a \$66,500 = \$16,625,000
- Predicciones aceptables: 50 % (250 apt) a \$9,500 = \$2,375,000
- Costo mensual total: \$19,000,000

$$\begin{aligned}
\text{Costo anual operacional} &= \$19{,}000{,}000 \times 12 = \$228{,}000{,}000 \\
\text{Ahorro bruto anual} &= \$342{,}000{,}000 - \$228{,}000{,}000 = \$114{,}000{,}000 \\
\text{Ganancia neta} &= \$114{,}000{,}000 - \$30{,}600{,}000 = \$83{,}400{,}000 \\
\text{ROI} &= \frac{\$83{,}400{,}000 - \$14{,}000{,}000}{\$14{,}000{,}000} \times 100 = \mathbf{496\,\%}
\end{aligned}$$

Cuadro 3: Resumen del análisis de sensibilidad

Escenario	**Costo mensual ML**	**Ganancia neta anual**	**ROI**
Optimista (35 % subestimaciones)	\$14,725,000	\$134,700,000	862 %
Base real (42.77 % subestimaciones)	\$16,948,000	\$108,024,000	672 %
Pesimista (50 % subestimaciones)	\$19,000,000	\$83,400,000	496 %

El modelo es viable incluso en el escenario pesimista con 50 % de subestimaciones críticas (ROI de 496 %). El modelo real, con 42.77 % de subestimaciones críticas medidas en el conjunto de prueba, genera un ROI de 672 % en el primer año.

6. Insights y Recomendaciones

Hallazgos técnicos: XGBoost obtiene R^2 de 0.878 con MAE de 161.9 millones. Aunque el error absoluto parece elevado, es bajo considerando la alta variabilidad de los precios (desviación estándar de 890 millones) en propiedades predominantemente de estrato 6. El área es el predictor más importante (correlación 0.829), seguido por administración (0.682). Los métodos de ensamble obtienen mejores resultados que modelos lineales.

Hallazgos económicos: El modelo genera ROI de 672 % en el primer año con break-even en 1.6 meses. Aunque 42.77 % de predicciones requieren avalúo presencial, el ahorro neto anual es de $108 millones. Los costos operacionales de infraestructura AWS son de $30.6M anuales usando SageMaker; estos costos pueden reducirse con infraestructura alternativa.

Recomendación: Implementar el modelo XGBoost para HabitAlpes. El ROI de 672 % y break-even de 1.6 meses justifican la inversión. Monitorear la tasa de subestimaciones durante los primeros 3 meses y reentrenar el modelo trimestralmente. Para reducir la tasa de subestimaciones (42.77 %): (1) agregar variables como antigüedad y acabados, (2) usar modelos con intervalos de confianza para identificar predicciones de alto riesgo, y (3) desarrollar modelos por rango de precio o localidad.

Referencias

[1] Indeed Colombia. Sueldo de Data scientist en Colombia. https://co.indeed.com/career/data-scientist/salaries, 2025. Salario promedio: $7,119,949 COP/mes según 20 salarios publicados (agosto 2025).

[2] CloudForecast. AWS SageMaker Pricing Guide - Cost Breakdown & Optimization Tips. https://www.cloudforecast.io/blog/aws-sagemaker-pricing/, 2025. Menciona ml.m5.xlarge a $0.23/hora para training y ml.m5.large a $0.115/hora para inferencia.

[3] Amazon Web Services. Amazon S3 Pricing. https://aws.amazon.com/s3/pricing/, 2025. Precios de almacenamiento S3: $0.023/GB para S3 Standard en región US East.

[4] Amazon Web Services. Amazon RDS Pricing. https://aws.amazon.com/rds/pricing/, 2025. Página oficial de precios de AWS RDS para bases de datos administradas.