Universidad de los Andes
Maestría en Ingeniería de la información
Profesor: Juan Pablo Reyes
Fecha entrega: 23 de noviembre de 2025
Estudiantes: Juan Carlos Tovar Orjuela

Taller 2

1. Resumen Ejecutivo y Alcance

Este reporte consolida el análisis exploratorio inicial realizado sobre la muestra de datos inmobiliarios recolectada. El objetivo es evaluar la viabilidad técnica de estos datos para la construcción de la Prueba de Concepto (PoC) del modelo de estimación de precios de venta.

El análisis confirma que el dataset posee un volumen y una riqueza de atributos suficientes. Sin embargo, se han identificado desafíos en la calidad de los datos numéricos y un sesgo poblacional marcado que requerirán una fase de preparación rigurosa para garantizar la fiabilidad del modelo en los segmentos objetivo.

1.1 Dimensiones Estructurales del Dataset

La evaluación de la completitud y el volumen de los datos revela una pérdida significativa de información en la variable objetivo, lo que redefine el tamaño efectivo de la muestra de entrenamiento.

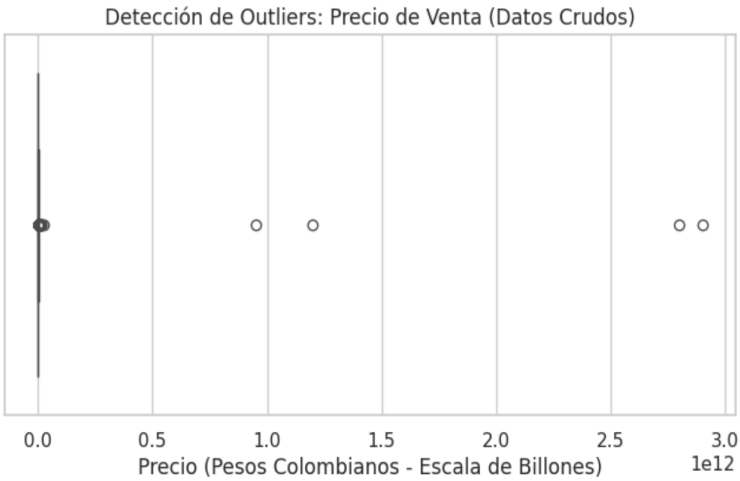
- **Volumen Bruto:** El dataset original consta de 43,013 registros y un espacio dimensional de 46 atributos (características).
- **Integridad de la Variable Objetivo:** La variable a predecir, precio_venta, presenta una ausencia de datos en 15,429 registros (aproximadamente el 35.8% del total).
- **Dimensión Efectiva:** Para fines de modelado supervisado, los registros sin precio carecen de utilidad. con unos datos totales de 27,584 inmuebles.
- **Otras Variables:** Variables clave de estructura (área) y ubicación (coordenadas) presentan una completitud del 100%. La variable *administración* requerirá imputación estadística debido a un 18% de valores faltantes.

1.2 Características de los Atributos y Detección de Anomalías

El dataset es una mezcla heterogénea de variables numéricas (financieras, dimensionales), categóricas (ubicación, estrato) y de texto no estructurado (descripciones).

Se identificaron valores extremos (outliers) que deben ser depurados:

- **Precio de Venta:** Se registraron valores máximos en el orden de billones de pesos (exp12), cifras inconsistentes con la realidad residencial.



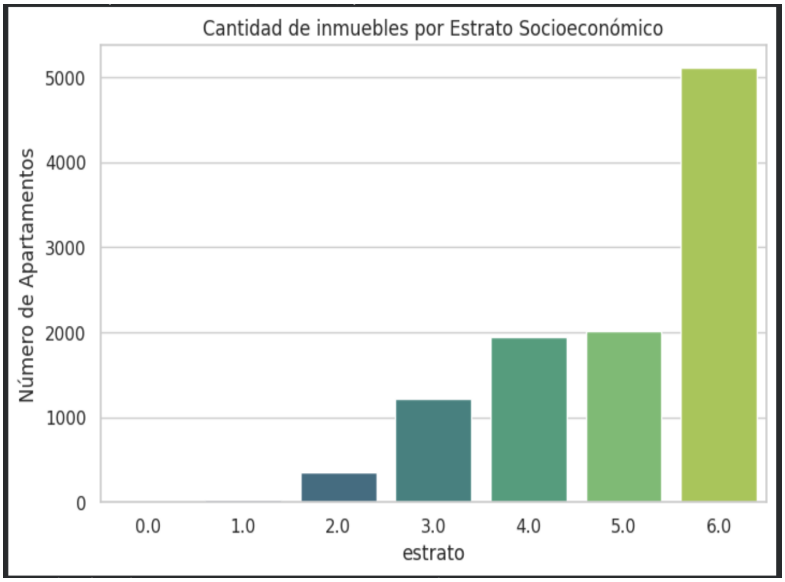
- **Área Construida:** Existen registros con superficies de 1.9 millones de metros cuadrados, así como valores nulos.
- **Administración:** Se detectaron cuotas mensuales superiores a los 3,500 millones de pesos.

Estas características confirman que los datos crudos no son aptos para el modelado directo y requieren la aplicación de reglas de negocio y filtros estadísticos (capping) para acotar los valores a rangos operativos reales.

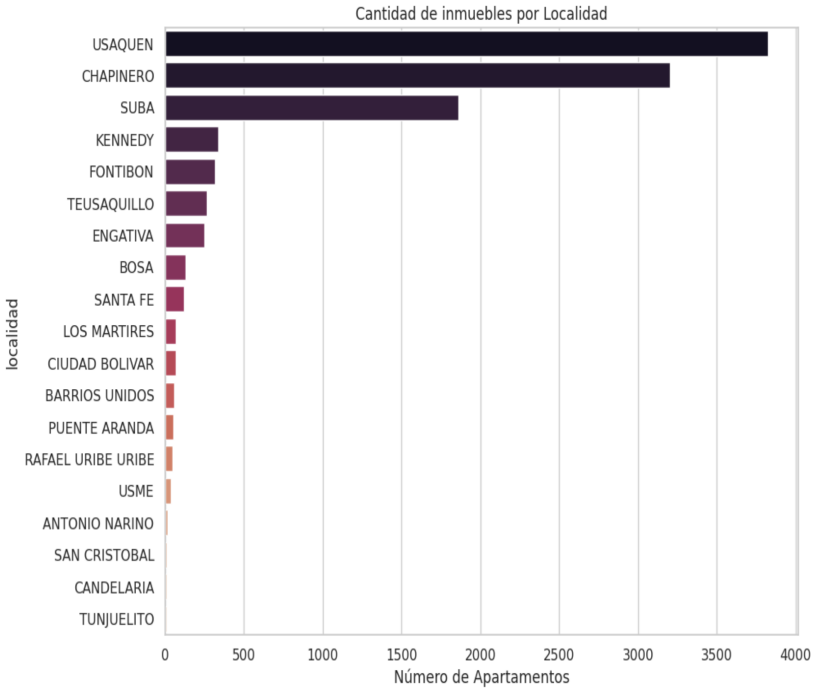
1.3. Indicadores Clave de Representatividad del Mercado

El análisis de la distribución de los datos revela un sesgo.

- **Concentración Socioeconómica:** Existe una concentración marcada de los estratos altos. Los estratos 4, 5 y 6 constituyen la mayoría de los datos (el estrato 6 cuenta con 19,000 registros brutos). Por el contrario, la presencia de inmuebles en estratos 1 y 2 es estadísticamente marginal.



- **Sesgo Geográfico:** Los datos presentan una concentración en localidades del norte y nororiente de Bogotá (Usaquén, Chapinero, Suba), mientras que las zonas del sur y occidente presentan una densidad de datos insuficiente para un aprendizaje robusto.



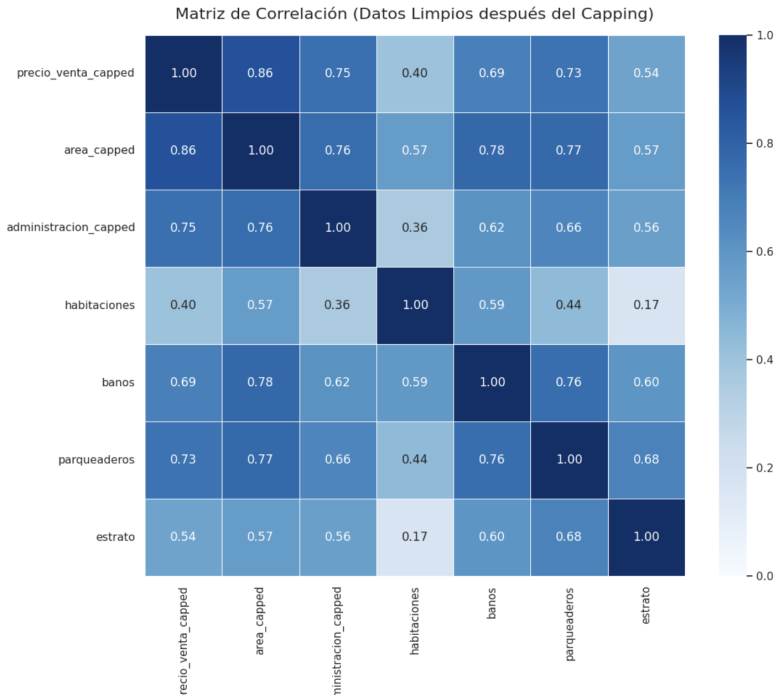
2. Desarrollo y Selección del Modelo Predictivo

Para abordar el objetivo de predecir el precio de venta de los apartamentos, se implementó una estrategia de modelado iterativa, evaluando múltiples algoritmos.

El proceso se rigió por la separación de los datos para garantizar la validez de los resultados: el conjunto de entrenamiento se utilizó para el ajuste de parámetros, el conjunto de prueba para la comparación y selección del mejor modelo, y un conjunto de validación final para confirmar el desempeño del modelo seleccionado.

2.1. Evaluación Comparativa de Modelos

Se generó una matriz de correlación para identificar cuales variables afectan directamente el precio.



Se entrenaron y evaluaron cinco arquitecturas de modelos diferentes para identificar el que mejor interpretara el mercado inmobiliario. Los modelos lineales (Regresión Lineal, Ridge, Lasso) y árboles de decisión (Random Forest, XGBoost)

La siguiente tabla resume el desempeño de cada modelo evaluado:

Tabla 1: Comparativa de Desempeño de Modelos en el Conjunto de Prueba

Modelo Evaluado	R2 (Explicación de Varianza)	MAE (Error Promedio)	RMSE (Error Cuadrático)	Observaciones
Regresión Lineal Simple	0.7770	$294,533,754	$493,726,359	Modelo base. Limitado por la no linealidad del mercado.
Regresión Ridge (L2)	0.7770	$294,522,396	$493,713,349	Desempeño idéntico a la lineal, indicando ausencia de sobreajuste inicial.
Regresión Lasso (L1)	0.7651	$304,128,512	$506,458,123	Ligera degradación del desempeño; no logró simplificar el modelo efectivamente.
Random Forest	0.8124	$247,008,730	$452,875,573	Mejora significativa al capturar relaciones no lineales y segmentadas.
XGBoost (Base)	0.8122	$246,432,720	$453,065,468	Desempeño competitivo con Random Forest en su configuración inicial.

Como se puede evidenciar, los modelos basados en árboles superaron a las aproximaciones lineales, reduciendo el error promedio en aproximadamente $48 millones de pesos. El modelo XGBoost mostró el mejor desempeño inicial en términos de MAE, por lo que fue seleccionado para una fase posterior de refinamiento.

2.2. Optimización del Modelo Seleccionado (XGBoost)

Para maximizar y mejorar el modelo XGBoost, se realizó un proceso de ajuste fino de hiperparámetros (tuning) utilizando una búsqueda aleatoria (RandomizedSearchCV) con validación cruzada de 3 pliegues sobre el conjunto de entrenamiento.

El proceso de optimización evaluó 50 combinaciones de configuraciones. Los mejores hiperparámetros encontrados, que priorizan un aprendizaje más lento pero robusto (learning_rate: 0.01) y una mayor profundidad para capturar interacciones complejas (max_depth: 7), fueron:

- n_estimators: 500
- learning_rate: 0.01
- max_depth: 7
- subsample: 0.8
- colsample_bytree: 0.8
- gamma: 0.1

Tras reentrenar el modelo, se obtuvieron los siguientes resultados definitivos:

- R2 Final: 0.8345 (El modelo explica el 83.45% de la variabilidad de los precios).
- MAE Final: $238,142,134.

Esta optimización logró reducir el error promedio en aproximadamente $8 millones adicionales respecto al modelo base, consolidando al XGBoost optimizado como la herramienta predictiva más precisa.

3. Análisis Cuantitativo y Evaluación de Resultados

Se presenta un análisis detallado de las métricas obtenidas por el modelo seleccionado (**XGBoost Optimizado**) en el conjunto de validación final, interpretando su significado en el contexto del negocio.

3.1 Interpretación de las Métricas de Desempeño

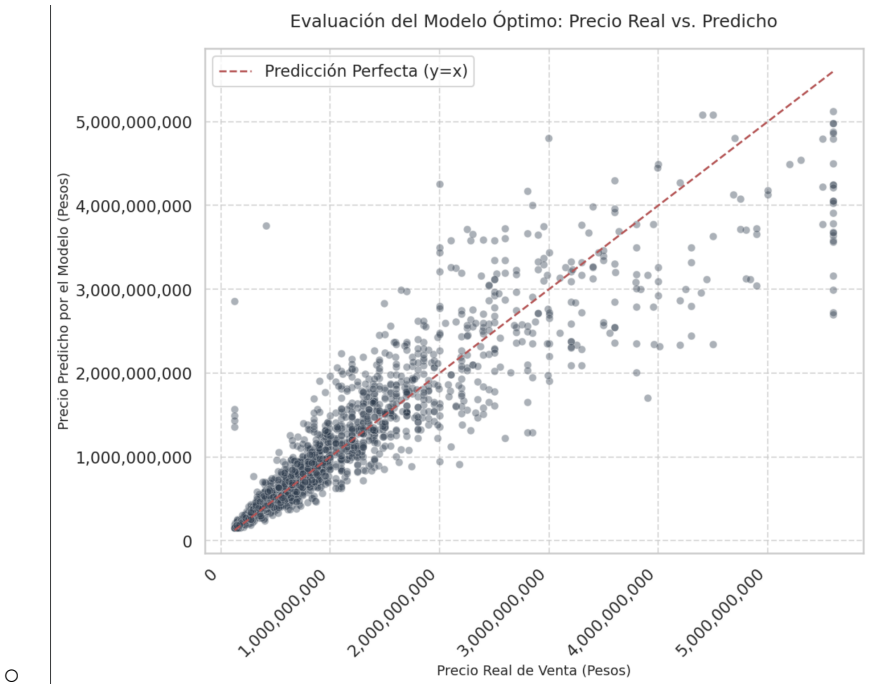
El modelo definitivo alcanzó los siguientes resultados:

- **Error Absoluto Medio (MAE):** $238,142,134
  - **Significado para el Negocio:** Esta métrica indica que, en promedio, las predicciones del modelo se desvían del precio real de venta en aproximadamente **$238 millones de pesos**, ya sea por exceso (sobrestimación) o por defecto (subestimación).

- **Raíz del Error Cuadrático Medio (RMSE):** $453,065,468 (Referencia del modelo base)
  - **Significado para el Negocio:** El RMSE penaliza más severamente los errores grandes. El hecho de que sea casi el doble del MAE indica la presencia de **errores atípicos significativos**. Es decir, aunque el error promedio es de

$238M, existen casos puntuales donde la desviación es de miles de millones, lo que eleva esta métrica.

- **Valor:** Alerta sobre la inestabilidad del modelo en segmentos específicos (probablemente lujo o propiedades con características inusuales), señalando dónde se concentra el riesgo operativo, como se puede observar en la siguiente gráfica, el modelo empieza acertando y prediciendo correctamente los valores por debajo de los 500 millones de pesos, sin embargo, tan pronto se empiezan a incrementar los precios el modelo empieza a cometer errores cada vez más grandes.



- **Coeficiente de Determinación (R2):** 0.8345
  - **Significado para el Negocio:** El modelo es capaz de explicar el **83.45%** de la variabilidad observada en los precios de venta.
  - **Valor:** Confirma que las variables seleccionadas (área, ubicación, características físicas) son determinantes sólidos del precio y que el modelo ha logrado capturar exitosamente la mayor parte de la dinámica del mercado.

3.2. Justificación de la Calidad del Modelo

La calidad del modelo alcanzado representa un avance sustancial respecto a los enfoques lineales iniciales.

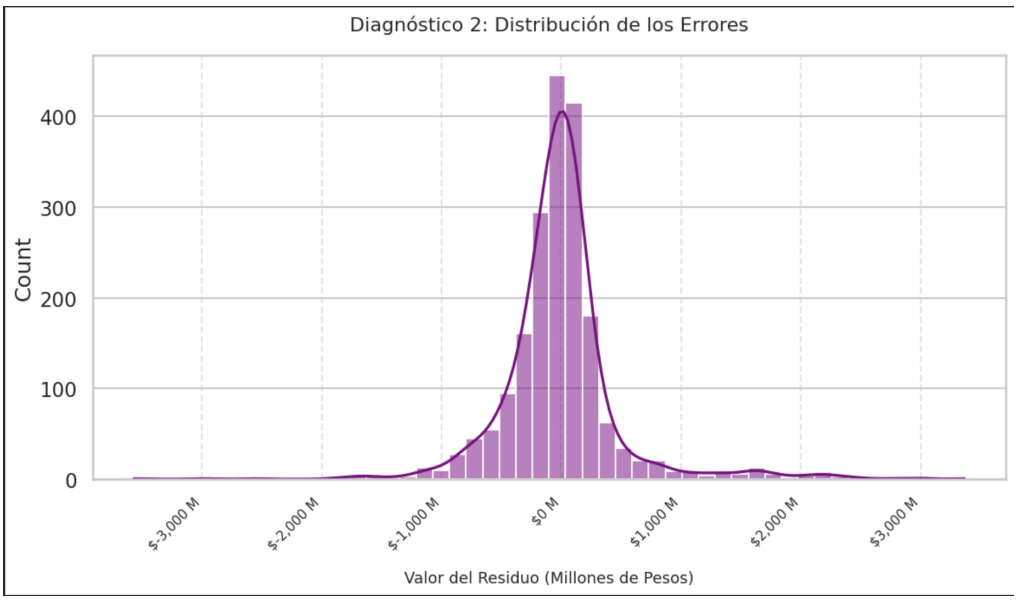
Factores de Impacto Positivo:

1. **(XGBoost):** La transición desde la regresión lineal a algoritmos de ensamble basados en árboles fue el factor determinante. Esto permitió al modelo capturar la complejidad del mercado inmobiliario, donde las relaciones entre características (ej. el valor marginal de un baño adicional o un metro cuadrado extra) no son constantes, sino que dependen del contexto (estrato, tamaño total).
2. **Optimización de Hiperparámetros:** El proceso de ajuste fino del XGBoost logró reducir el MAE en más de $8 millones adicionales, demostrando la importancia de configurar adecuadamente la velocidad de aprendizaje y la profundidad de los árboles para evitar el sobreajuste.
3. **Limpieza y Curación de Datos:** La eliminación rigurosa de valores atípicos extremos (outliers de billones de pesos) y la imputación adecuada de datos faltantes fueron precondiciones indispensables para que cualquier modelo pudiera aprender patrones válidos.

3.3. Oportunidades de Mejora Críticas

A pesar de los resultados positivos (R2 de 0.83), el MAE de $238 millones aún se encuentra distante del objetivo ideal de $20 millones para la automatización total. El análisis revela áreas claras de mejora:

Manejo de Errores Extremos (Colas Pesadas): La gran diferencia entre el MAE y el RMSE confirma que el modelo falla drásticamente en un subconjunto de propiedades (probablemente las de ultra-lujo o características únicas). Se requiere investigar y posiblemente segmentar estos casos para tratarlos con modelos especializados, como se muestra en la siguiente gráfica, tan pronto el precio empieza a subir se crean picos, comportamientos indeseables.



Enriquecimiento de Datos (Ingeniería de Características): La información actual, es limitada. La incorporación de variables de ubicación-especifica (distancia a puntos de interés específicos, valor promedio del m² en la manzana) y temporalidad (tendencias de precios recientes) ayudara a aumentar la precisión.

Optimización Enfocada en la Métrica de Negocio: Actualmente, el modelo penaliza igual la subestimación que la sobrestimación. Se debe explorar el uso de funciones de pérdida personalizadas que penalicen más fuertemente la subestimación, alineando directamente el entrenamiento del modelo con la regla de negocio de minimizar los avalúos presenciales costosos.

4. Análisis Cualitativo de la Decisión del Modelo (Interpretabilidad)

Es importante entender cómo el modelo XGBoost llega a sus predicciones para validar su razonamiento y para esto utilizamos la técnica de valores SHAP (SHapley Additive exPlanations), que asigna a cada característica una contribución monetaria exacta al precio final predicho.

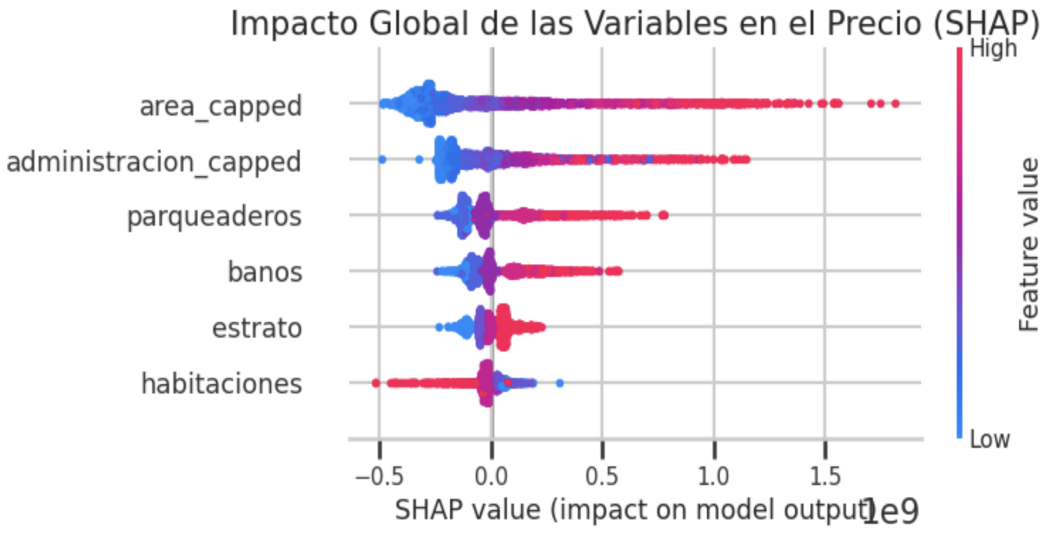
Este análisis se divide en dos niveles: la perspectiva global del mercado y el estudio detallado de un caso individual.

4.1. Comportamiento Global del Mercado

El gráfico de resumen SHAP (Beeswarm) proporciona información completa de qué factores impulsan el valor de los apartamentos en Bogotá según el modelo y en qué dirección lo hacen.

Interpretación del Mercado según el Modelo:

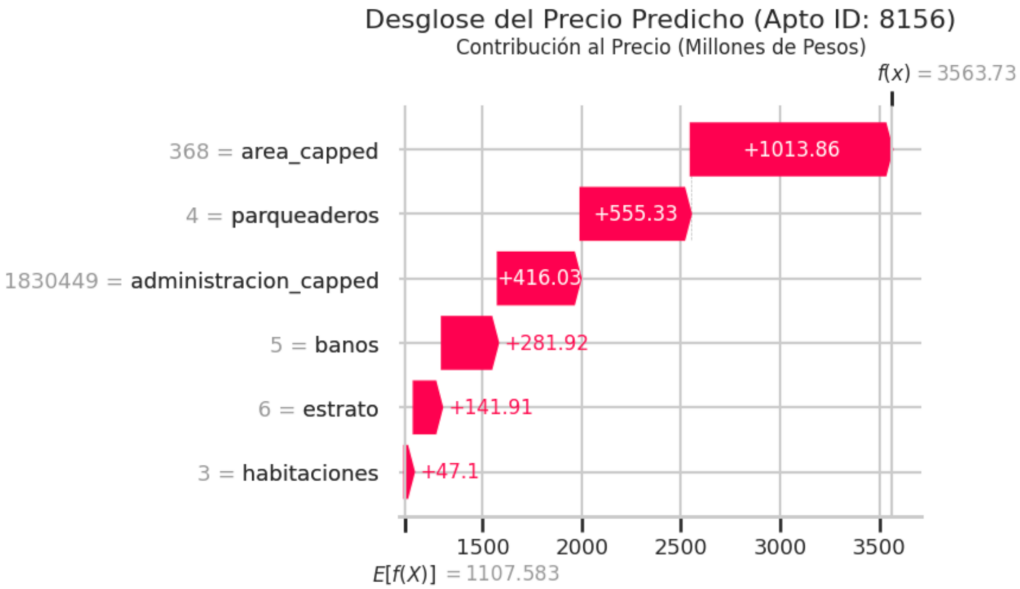
1. **Jerarquía de Importancia:** Las características se ordenan verticalmente por su poder predictivo global. El modelo identifica que el área construida (area_capped), la cuota de administración (administracion_capped) y el estrato socioeconómico (estrato) son, con diferencia, los tres factores más determinantes del precio.



2. **Dirección del Impacto:**
    - **Factores Positivos (A la derecha):** Los puntos de color rojo intenso (valores altos de la característica) se agrupan a la derecha del eje central. Esto indica que un mayor metraje, una administración más costosa y un estrato más alto incrementan significativamente el precio predicho.
    - **Factores Negativos (A la izquierda):** Los puntos azules (valores bajos) se concentran a la izquierda, confirmando que áreas pequeñas y estratos bajos reducen la valoración.
3. **Validación de Lógica:** No se observan relaciones contraintuitivas (ej. que más baños reduzcan el precio), lo que valida la coherencia interna del modelo.

4.2. Análisis de una Predicción Individual (Caso de Alto Valor)

Para entender cómo se combinan estos factores en una propiedad específica, descomponemos la predicción del apartamento más costoso del conjunto de prueba (ID: 8156). El gráfico de cascada muestra cómo cada característica suma o resta valor al precio promedio del mercado para llegar a la predicción final de **$3,563 millones**.



Desglose de la Valoración:

- **Punto de Partida:** El modelo inicia con un precio base de mercado de aproximadamente $1,107 millones
- **Contribuciones Positivas (Barras Rojas):**
    - El factor dominante es el gran metraje (area_capped), que por sí solo añade $1,013 millones al valor base.
    - El alto estrato y el elevado costo de administración suman en conjunto cerca de $971 millones adicionales.

    - El número de baños y parqueaderos también contribuyen positivamente, aunque en menor medida.
- **Resultado Final:** La suma de estas contribuciones lleva la predicción final a $3,563 millones, demostrando que el modelo valora esta propiedad premium por su combinación de espacio, ubicación y características de lujo.

Conclusión del Análisis Cualitativo:

La interpretación mediante valores SHAP confirma que el modelo XGBoost ha aprendido una representación lógica y económicamente válida del mercado inmobiliario. Sus decisiones se basan en factores estructurales y de ubicación clave, ponderados de manera coherente con la realidad del negocio, lo que refuerza la confianza en su capacidad para realizar valoraciones automáticas.

5. Generación de Valor y Análisis Financiero del Modelo

El objetivo de implementar el modelo XGBoost es generar un impacto financiero tangible para HabitAlpes. Este análisis cuantifica ese impacto, contrastando los costos de la operación manual actual frente a la operación híbrida (modelo + perito) propuesta.

5.1. Supuestos del Modelo Financiero (Parametrización)

Para realizar este cálculo, se establecen parámetros base. Dado que no contamos con datos internos financieros de HabitAlpes, utilizamos estimaciones de mercado conservadoras:

1. **Costo del "Status Quo" (Avalúo Manual):** Se estima que el costo operativo promedio de un avalúo físico certificado (tiempo del perito, viáticos, procesamiento) en Bogotá ronda los $500,000 COP por inmueble.
2. **Regla de Negocio Crítica:** El modelo financiero respeta estrictamente la regla definida:
    - Si el modelo subestima por > $20 Millones, se dispara un avalúo manual (Costo: $500,000).
    - En cualquier otro caso (acierto, error pequeño o sobrestimación). El proceso es automático (Costo marginal: $0).
3. **Tasa de "Fallo" del Modelo:** Basado en el análisis de residuos y el MAE ($238M). Estimamos paramétricamente que, en esta primera fase, el modelo activará la regla de revisión manual en el 40% de los casos más complejos (segmento de lujo/atípicos), logrando automatizar exitosamente el 60% restante.
4. **Inversión Inicial del Proyecto:** Se asume un costo distribuido de la siguiente forma.

Categoría	Concepto / Rol	Descripción y Justificación	Asignación Estimada (COP)	% del Total
Talento Humano	Científico de Datos Sénior (Líder)	Responsable principal. Encargado de la limpieza avanzada, ingeniería de características, entrenamiento y optimización de modelos (XGBoost), e interpretación de resultados. (Aprox. 2.5 meses dedicación completa equivalente).	$ 30,000,000	60%
	Ingeniero de Datos (Soporte)	Apoyo en la fase inicial para garantizar la ingesta correcta de los datos crudos y configurar el entorno de computación en la nube de forma segura. (Aprox. 15 días de dedicación equivalente).	$ 7,500,000	15%
	Líder de Proyecto / Enlace de Negocio	Coordinación con los peritos de HabitAlpes para validar reglas de negocio y asegurar que el modelo responda a las necesidades reales de la operación. (Supervisión parcial).	$ 5,000,000	10%
Infraestructura y Tecnología	Cómputo en la Nube (Compute)	Alquiler de instancias de procesamiento (AWS EC2 o SageMaker) necesarias para entrenar el modelo XGBoost y realizar la optimización de hiperparámetros eficientemente.	$ 3,500,000	7%
	Almacenamiento y Herramientas	Costos de almacenamiento seguro de datos en la nube (S3 buckets) y posibles licencias menores de software de visualización o entornos de desarrollo especializados.	$ 1,500,000	3%
Otros	Contingencia / Imprevistos	Reserva del 5% para cubrir excesos en costos de cómputo, necesidad de horas adicionales de consultoría o adquisición de sets de datos externos complementarios si fuera necesario.	$ 2,500,000	5%
TOTAL	INVERSIÓN INICIAL TOTAL		$ 50,000,000	100%

5.2. Matriz de Costos Esperados por Transacción

Utilizando estos supuestos, construimos una matriz de decisión para calcular el costo esperado de cada valoración bajo el nuevo sistema.

Tabla 2: Matriz de Clasificación Financiera de Predicciones

Tipo de Predicción (Según Regla de Negocio)	Probabilidad Estimada del Evento (P)	Costo Operativo Asociado (C)
A. Automatización Exitosa (Error < $20M o Sobreestimación)	60% (P_exito)	$0 COP (Ahorro total de tiempo de perito)
B. Falla del Modelo / Revisión Requerida (Subestimación > $20M)	40% (P_falla)	$500,000 COP (Costo asociado al error: Se requiere perito físico)

5.3. Cálculo de la Ganancia Esperada

El costo esperado por cada valoración utilizando el modelo es el promedio ponderado de los escenarios de la matriz:

Costo del modelo = (0.60 x 0) + (0.40 x 500,000) = $200,000 COP

La generación de valor (ahorro) por cada estimación es:
Ganancia = Costo Manual - Costo del Modelo
Ganancia = $500,000 - $200,000 = $300,000 COP por transacción

Interpretación: Incluso con un modelo que requiere intervención humana el 40% del tiempo debido a la regla de negocio, la implementación genera un ahorro promedio de $300,000 pesos en cada solicitud que recibe HabitAlpes, al filtrar el 60% de los casos.

5.4. Análisis de Punto de Equilibrio (Break-Even Point)

Para determinar cuándo la inversión inicial de $50 Millones empieza a generar dividendos netos, calculamos el número de transacciones necesarias para cubrir dicho costo con los ahorros generados.

Transacciones para Break-Even = Inversión Inicial / Ahorro por Transacción
Transacciones para Break-Even = $50,000,000 / $300,000 = 167 valoraciones

a. Los costos de tiempo asociado a peritos (Línea Base):
Actualmente, cada solicitud de valoración implica la movilización de un profesional. Establecemos como costo unitario base de esta operación manual (tiempo, desplazamiento y expertise del perito) un valor de $500,000 COP por inmueble.

b. El ahorro de tiempo teórico del modelo:
El modelo actúa como un primer filtro de alta velocidad. su capacidad de procesamiento es inmediata. En la práctica, esto se traduce en que el 60% de las solicitudes ya no requieren el tiempo del perito, liberando esa capacidad operativa casi en su totalidad.

c. El costo asociado a los errores del modelo:
Bajo la regla de negocio de HabitAlpes, el 40% de las estimaciones (las subestimaciones críticas) fallan. El costo asociado a estos errores es la obligación de realizar el avalúo manual de todos modos.

Cálculo del riesgo: 40% de probabilidad de fallo X $500,000 costo unitario = $200,000 COP

d. ROI:

Supuesto para el ROI: Volumen Transaccional

- **Supuesto de Volumen:** Estimaremos que HabitAlpes procesa un promedio de 300 solicitudes de valoración al mes. Esto equivale a 3,600 valoraciones al año.

Concepto	Valor	Tipo
SUPUESTOS (INPUTS)		
Inversión Inicial del Proyecto (PoC)	$ 50,000,000	Costo Único
Volumen Anual Estimado de Valoraciones	3600	Unidades (300/mes)
Costo Unitario Actual (Avalúo Manual)	$ 500,000	Costo por unidad
Ahorro Neto Promedio Estimado por Transacción	$ 300,000	Ahorro calculado
RESULTADOS (OUTPUTS)		
Ganancia Total Operativa Proyectada (Año 1)	$ 1,080,000,000	Unitario)
Ganancia Neta (Año 1)	$ 1,030,000,000	Inversión)
RETORNO SOBRE LA INVERSIÓN (ROI)	2060%	Resultado Final

Interpretación para el Reporte:

Bajo una proyección de 300 valoraciones mensuales, se estima que la implementación del modelo generará ahorros operativos anuales cercanos a los $1,080 millones de pesos.

Al contrastar este beneficio con la inversión inicial de $50 millones, el proyecto arroja un Retorno sobre la Inversión (ROI) proyectado a un año del 2,060%.

Esto indica que, por cada peso invertido en el desarrollo de esta Prueba de Concepto, la compañía recupera ese peso y genera más de $20 pesos adicionales en ahorros operativos durante el primer año.

6. Insights Clave del Modelo y el Mercado

- **La Complejidad del Mercado Bogotano Exige Tecnología Avanzada**
  Los intentos iniciales con modelos lineales (estadística tradicional) fracasaron, demostrando que el mercado inmobiliario no sigue reglas simples. Por esta razón se buscaron otros modelos obteniendo el mejor resultado con (XGBoost) el cual permite capturar las curvas y matices del mercado, logrando explicar el 83.5% de la variabilidad de los precios.
- **La Lógica del Modelo Valida la Intuición del Negocio (Interpretabilidad)**
  Al entender como crea la solución el modelo con técnicas de interpretabilidad (SHAP), confirmamos que el modelo razona de forma coherente con la experiencia de un perito humano. Sus valoraciones se basan sólidamente en tres pilares jerárquicos:

  - Área Construida: El factor dominante indiscutible.
  - Estrato Socioeconómico: Determinante de la ubicación y el entorno.
  - Cuota de Administración: Un proxy efectivo del nivel de lujo y amenidades del edificio.

- **El Segmento de Lujo es el Límite Actual de la Automatización**
  Aunque el modelo tiene un error promedio (~$238 Millones), sufre de "colas pesadas". Esto significa que, en inmuebles de muy alto valor (el segmento premium), el modelo es inestable y puede cometer errores de subestimación catastróficos (superiores a $1,000 millones).

  Implicación: Bajo la regla actual de no subestimar por más de $20M, el modelo NO está listo para operar sin supervisión en el segmento de lujo.

- **La Viabilidad Financiera**

Al utilizar el modelo como un filtro inteligente que automatiza el 60% de los casos (mercado medio y bajo), HabitAlpes genera un ahorro operativo neto promedio de $300,000 COP por cada solicitud recibida. La inversión inicial del proyecto se recupera tras procesar 167 valoraciones y la implementación del modelo generará ahorros operativos anuales cercanos a los $1,080 millones de pesos.

Recomendación Final para el Negocio

Basados en la evidencia cuantitativa y cualitativa, la recomendación es NO desplegar el modelo como un sustituto total del perito humano inmediatamente, sino implementar un sistema híbrido de triaje inteligente y progresivamente ir aumentando la data para mejorar la efectividad.