Actividad 2.

Regresión No Lineal.

Gestión de proyectos de plataformas tecnológicas
Gpo 201

06 de Octubre, 2025
ITESM Puebla

Valeria Becerril Hernandez | A01736860

Introducción.

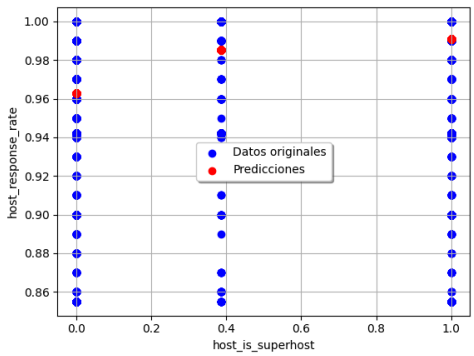
La regresión no lineal es una herramienta analítica esencial que se utiliza cuando la simple regresión lineal (la línea recta) resulta inadecuada para describir la relación entre variables. En el contexto de negocios como Airbnb, muchas dinámicas económicas y de rendimiento no siguen patrones lineales, sino que exhiben curvas de saturación, crecimiento exponencial o rendimientos decrecientes a partir de ciertos umbrales. Utilizar modelos no lineales (como el logarítmico o el cuadrático) es crucial porque permite capturar estas complejas dinámicas, proporcionando un ajuste más preciso y revelando cómo cambia el impacto de una variable predictora a lo largo de su rango, un nivel de conocimiento que es vital para la optimización y gestión de ingresos.

Resumen de la actividad anterior con hallazgos clave.

En la Actividad 1 de regresiones lineales (simple y múltiple), realizando la limpieza de datos se establecieron la base para esta segunda entrega, revelando dónde la asunción de una relación lineal es inadecuada y debe ser sustituida por modelos no lineales. El proceso inicial incluyó una limpieza robusta de una base de datos de 26,401 filas, transformando variables clave (como tasas de respuesta y aceptación) y gestionando outliers mediante el método IQR, lo que garantiza un conjunto de datos estable. El análisis lineal arrojó dos conclusiones cruciales: primero, la presencia de una multicolinealidad extrema en variables redundantes (como métricas de inventario y subpuntuaciones de calidad) llevó a R^2 artificialmente altos, un problema que debe resolverse. Segundo, y más importante para esta actividad, es que, la regresión lineal demostró un poder predictivo débil (R^2 bajo, en el rango de 0.20 a 0.30) en el modelado de variables clave de negocio, como la Tasa de Aceptación del Anfitrión (host_acceptance_rate) usando variables de gestión, a pesar de que la Tasa de Respuesta (host_response_rate) se confirmó como el predictor lineal más fuerte. Este fallo en la linealidad, junto con la evidencia de dinámicas segmentadas (donde la relación entre precio y aceptación es nula en casas completas, pero negativa en hoteles), indica que las relaciones subyacentes son probablemente curvilíneas. Por lo tanto, el foco de esta segunda entrega es aplicar regresiones no lineales para capturar estos umbrales y aceleraciones, y así mejorar el R^2 del modelo y la comprensión de cómo el valor y la gestión impactan las métricas de rendimiento más allá de una simple línea recta.

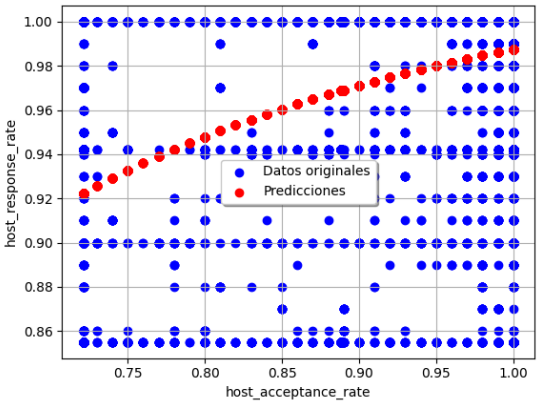
Regresión No lineal

Bloque 1 'host_response_rate'



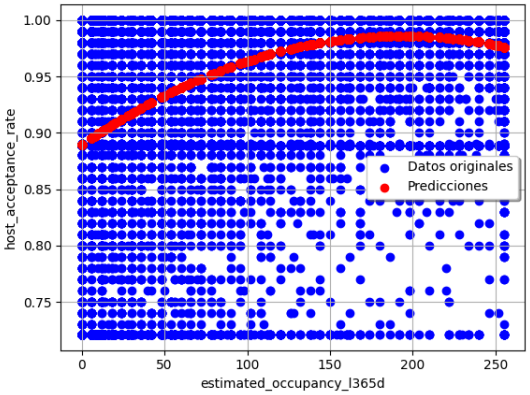
El Modelo 1, que intenta modelar la Tasa de Respuesta (host_response_rate) utilizando el estatus Superhost (*host_is_superhost*, una variable binaria 0/1), resultó ser un modelo deficiente. Su R^2 de 0.0983 (9.83%) y una correlación (R) de 0.3135 indican un poder explicativo muy bajo. La regresión no lineal falló porque el predictor (X) sólo toma dos valores (0 o 1), lo que hace que la compleja curva

polinomial se colapse ineficazmente sobre solo dos grupos de datos. Como se observa en el gráfico, la curva de predicción roja es casi plana y apenas se distingue de una regresión lineal simple que ajustaría la media para cada grupo. Esto confirma que el modelo no lineal es inapropiado para una variable binaria y que el estatus de Superhost por sí solo tiene una relación lineal muy débil con la Tasa de Respuesta.

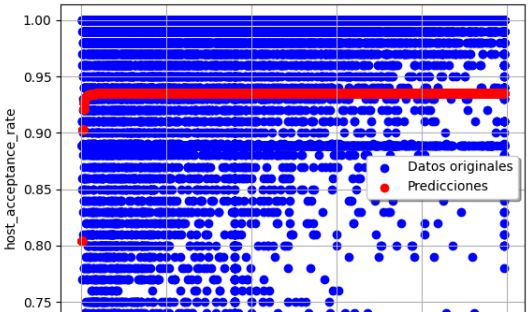


El Modelo 2, que ajusta la Tasa de Respuesta (host_response_rate) en función de la Tasa de Aceptación (*host_acceptance_rate)* mediante una curva cuadrática, demostró un poder explicativo modesto. Con un R^2 de 0.2413 (24.13%), este modelo logra explicar una porción considerablemente mayor de la variabilidad que el Modelo 2, y la correlación (R) de 0.4912 indica una relación positiva moderada. Gráficamente, la línea de predicción roja (Predicciones) sigue una clara tendencia ascendente, dibujando una curva cóncava o de rendimientos decrecientes, sugiriendo que la Tasa de Respuesta aumenta con la Tasa de Aceptación, pero a un ritmo que se desacelera a medida que ambas métricas se acercan al 100%. Los coeficientes reflejan esto: el término cuadrático (a=0.2224) es positivo, pero la influencia del intercepto y el término lineal también moldean la curva, mostrando que la relación entre ambas métricas de rendimiento es consistente con la lógica de negocio, aunque el 76% de la variabilidad todavía no se explica.

Bloque 2 'host_acceptance_rate'



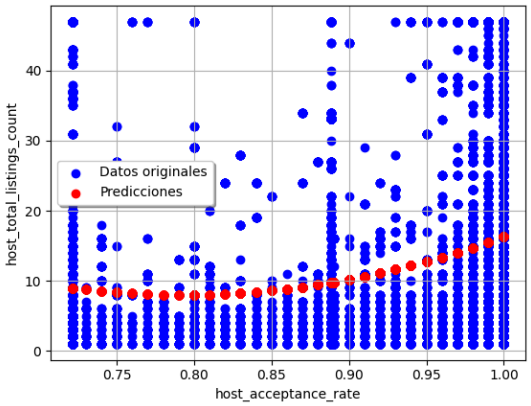
El Modelo 1, que relaciona la Tasa de Aceptación (host_acceptance_rate) con la Ocupación Estimada Anual (*estimated_occupancy_l365d*) mediante una función logarítmica o similar, es el mejor de los dos. Aunque su R^2 de 0.1671 (16.71%) es bajo, la correlación (R) de 0.4088 indica una relación positiva débil a moderada, pero con una clara forma funcional. Gráficamente, la curva de predicción roja (Predicciones) dibuja una clásica curva de crecimiento con saturación (logarítmica): la Tasa de Aceptación sube rápidamente a medida que la Ocupación Estimada aumenta de 0 a 100 días, pero luego el crecimiento se desacelera fuertemente (saturación o "rendimientos decrecientes") y la curva se aplana, acercándose a un límite asintótico (cercano a 1.00)



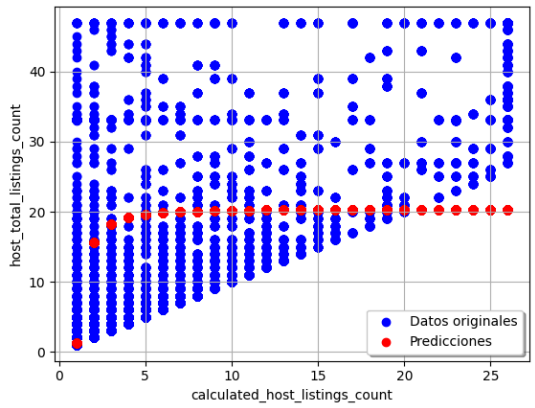
El Modelo 2, que relaciona la Tasa de Aceptación con la Frecuencia de Reseñas por Mes (*reviews_per_month*), resultó ser

extremadamente deficiente. Con un R^2 de 0.0094 (menos del 1%) y una correlación (R) de 0.0967, el modelo es esencialmente inútil para predecir la Tasa de Aceptación. Gráficamente, la línea de predicción roja es casi perfectamente horizontal, lo que indica que, para este modelo, la Tasa de Aceptación es constante, independientemente de cuántas reseñas reciba el listado al mes.

Bloque 3 'host_total_listings_count'



El Modelo 1, que utiliza la *Tasa de Aceptación* para predecir el Conteo Total de Anuncios, es muy débil. Con un R^2 de 0.0402 (solo 4.02%) y una correlación (R) de 0.2006, este modelo tiene un poder explicativo insignificante. La curva de predicción roja dibuja una ligera forma de 'U', sugiriendo que los anfitriones con tasas de aceptación muy bajas o muy altas tienden a tener más listados que los anfitriones de rango medio. Sin embargo, dado el R^2 tan bajo, esta tendencia es irrelevante estadísticamente. El amplio rango de datos azules en el eje Y (conteo de listados) muestra que el inventario es casi constante en todo el rango de la Tasa de Aceptación, reafirmando que el comportamiento de aceptación no es un predictor útil del tamaño de la operación de un anfitrión.

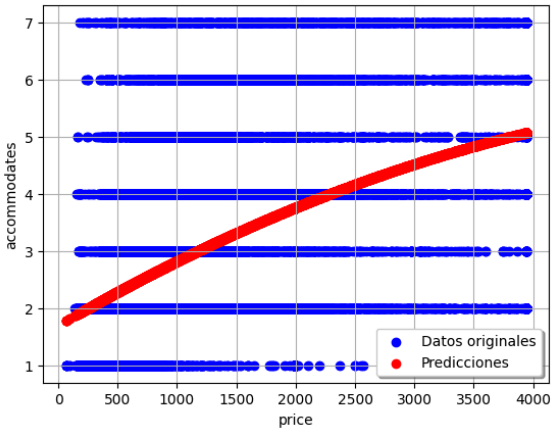


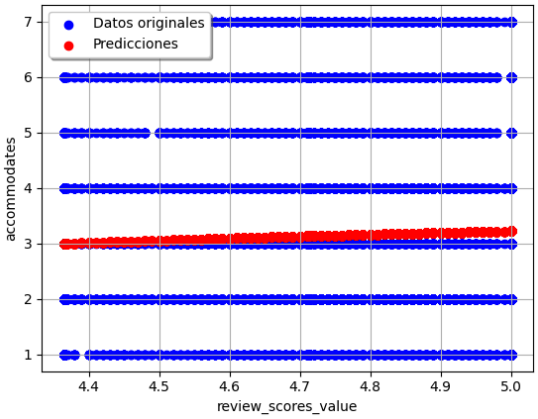
El Modelo 2, que intenta predecir el Conteo Total de Anuncios en función del *Conteo Calculado de Anuncios*, logra un R^2 de 0.2883 (28.83%) y una correlación (R) de 0.5369, indicando una relación positiva moderada. A pesar de tener el R^2 más alto, el modelo fracasa en su objetivo debido a la multicolinealidad inherente. Ambas variables miden casi el mismo concepto, por lo que el alto R^2 es esperado. Sin embargo, el gráfico muestra que la curva de predicción roja es casi horizontal y se aplana rápidamente alrededor de 20, lo que indica que el modelo no puede seguir la vasta dispersión de los datos originales (puntos azules), que alcanzan más de 40. La predicción es una subestimación constante para la mayoría de los anfitriones de gran escala, lo que lo hace inútil para predecir inventarios altos.

Bloque 4 'accommodates'

El Modelo 1, que utiliza una función logarítmica o potencial (inferida por la forma de la curva) para modelar la relación entre Capacidad de Huéspedes y *Precio*, es el mejor modelo. El R^2 es de 0.2267 (22.67%), y la correlación (R) es de 0.4761. Esto indica una

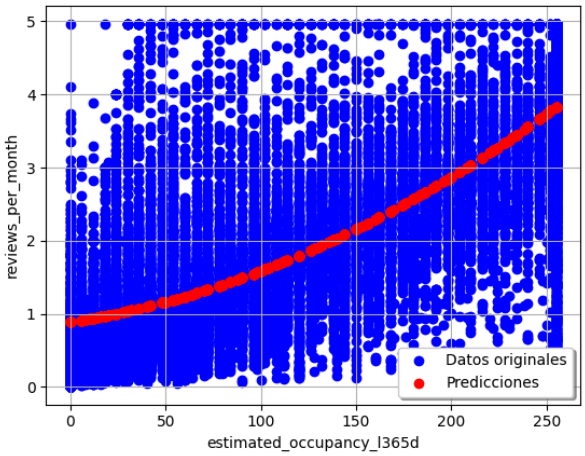
relación positiva moderada. El modelo explica más del 22% de la variabilidad, lo cual es significativo considerando que se está usando una métrica económica (precio) para predecir una métrica estructural (capacidad). Gráficamente, la curva de predicción roja es ascendente y cóncava, dibujando una curva de rendimientos decrecientes. Esto significa que para precios bajos (hasta ≈$1500), la capacidad de huéspedes aumenta linealmente con el precio y para precios altos (por encima de ≈$3000), la curva se aplana. Esto sugiere que, a partir de cierto precio elevado, el anfitrión ya no puede aumentar significativamente el número de huéspedes que aloja; es decir, el precio adicional está pagando por lujo o ubicación, no por un aumento sustancial en el tamaño físico o capacidad de camas.





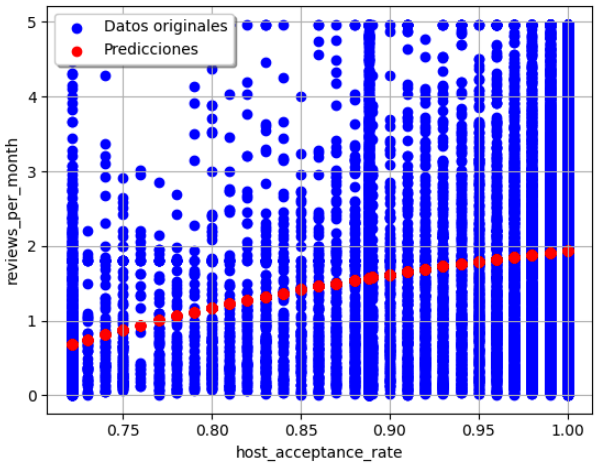
El Modelo 2 intenta predecir la Capacidad de Huéspedes (accommodates) en función de la Puntuación de Valor de la Reseña *(review_scores_value)* demostró ser completamente inútil y carente de poder predictivo. El hallazgo principal es la ausencia total de relación funcional entre una métrica de feedback del huésped y una métrica estructural de la propiedad. Esto se confirma con un R^2 de solo 0.0014 (0.14%), lo que significa que el modelo no explica prácticamente nada de la variabilidad en la capacidad de alojamiento. La correlación (R) de 0.0376 también es insignificante. Gráficamente, el fracaso del modelo es evidente: la curva de predicción roja se presenta como una línea horizontal casi perfecta alrededor de accommodates≈3.1, incapaz de capturar la estructura de los datos azules que se encuentran en niveles discretos (1, 2, 3, 4, etc.). Este resultado reafirma un principio fundamental que ya se observó en la regresión lineal: las variables estructurales (accommodates, bedrooms) no pueden ser predichas por variables de servicio o calidad (review_scores_value), independientemente de si el modelo es lineal o no lineal.

Bloque 5 'reviews_per_month'



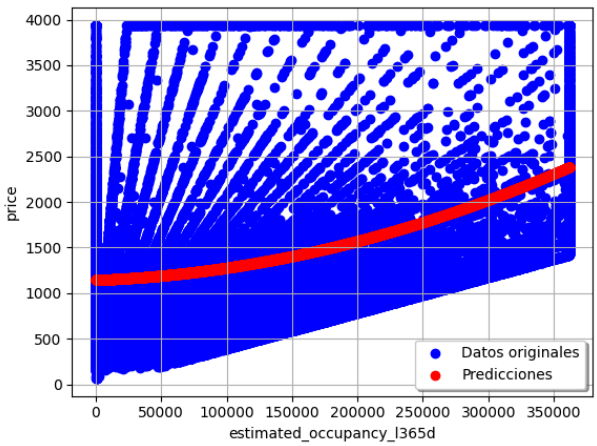
El Modelo 1, que utiliza la Ocupación Estimada Anual *(estimated_occupancy_l365d)* como predictor, es el modelo más exitoso que has construido hasta ahora. Con un impresionante R^2 de 0.5683 (56.83%), este modelo logra explicar más de la mitad

de la variabilidad en la frecuencia de reseñas, y la fuerte correlación (R) de 0.7539 confirma una relación positiva y robusta. El hallazgo principal es que la relación es marcadamente no lineal y exponencial/cuadrática, tal como lo muestra la curva de predicción roja. Gráficamente, la curva asciende de forma acelerada (convexo o en forma de 'U'), lo que significa que el crecimiento en la frecuencia de reseñas es cada vez más rápido a medida que la ocupación estimada aumenta. Esto sugiere que no solo la ocupación alta genera más reseñas, sino que los anfitriones con ocupación muy alta (>150 días) se benefician de un efecto de red o de volumen que dispara el índice de reseñas por mes.



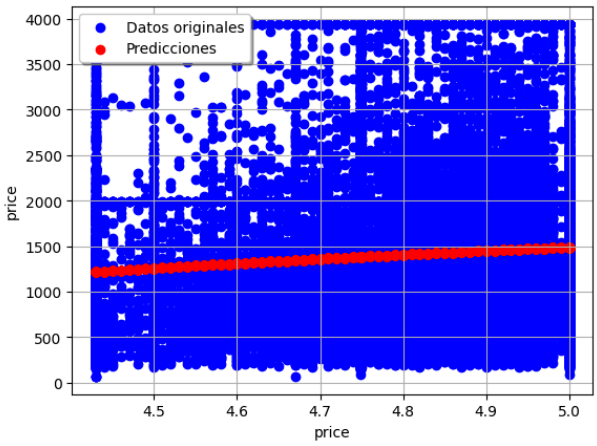
Por otro lado, el Modelo 2, que utiliza la Tasa de Aceptación del Anfitrión (*host_acceptance_rate*) como predictor, es muy débil. Con un R^2 de 0.0888 (8.88%) y una correlación (R) de 0.2980, este modelo apenas explica la variabilidad. El hallazgo aquí es la ausencia de relación funcional, ya que la curva de predicción roja es casi lineal y muy plana. Esto confirma que la frecuencia con la que se reciben reseñas está impulsada casi exclusivamente por la demanda y la ocupación (Modelo 1) y es, en gran medida, independiente del comportamiento de aceptación o servicio marginal del anfitrión (Modelo 2). En resumen, la alta demanda predice altas reseñas, mientras que un anfitrión proactivo en aceptar solicitudes no se traduce directamente en un aumento significativo de reseñas mensuales.

Bloque 6 'price'



El Modelo 1, que predice el Precio (price) en función de la Ocupación Estimada Anual (*estimated_occupancy_l365d*) mediante una función no lineal (potencial o exponencial), logró un R^2 de 0.1581 (15.81%). Esto significa que el modelo explica cerca del 16% de la variabilidad del precio, y la correlación (R) de 0.3977 indica una relación positiva débil a moderada. El hallazgo principal es la relación no lineal de aceleración del precio por demanda. Gráficamente, la curva de predicción roja es ascendente y cóncava hacia arriba, sugiriendo que el precio aumenta lentamente a medida que la ocupación sube en el rango bajo, pero experimenta una aceleración más marcada en el rango de ocupación alta (más allá de los 250 días). Esto valida la lógica de gestión de ingresos: los anfitriones con mayor demanda histórica pueden permitirse un precio base más alto, y el impacto del precio se vuelve más

significativo a medida que la ocupación se acerca al máximo, lo que no sería capturado por un modelo lineal.



Por otro lado, el Modelo 2, que intenta modelar el Precio utilizando la Puntuación de Valor de la Reseña (aunque la etiqueta del eje X en el gráfico es incorrecta y debería ser *review_scores_value*), resultó ser completamente fallido. Con un R^2 de 0.0073 (menos del 1%), el modelo no tiene poder predictivo. La curva de predicción roja es esencialmente una línea horizontal plana, lo que confirma que la calidad percibida (Puntuación de Valor) por sí sola es un predictor irrelevante para el precio de lista. El precio está determinado principalmente por factores estructurales y la demanda, no por el feedback directo de valor.

Tabla de determinación y correlación.

Variable Dependiente (Y)	Variable Independiente (X)	R^2	R	Hallazgo Principal
host_response_rate	host_acceptance_rate	0.2413	0.4912	Moderado. Captura la curva de rendimiento entre las tasas del anfitrión.
	host_is_superhost	0.0983	0.3135	Débil. Inapropiado para variable binaria. Bajo poder predictivo.
reviews_per_month	estimated_occupancy_l365d	0.5683	0.7539	Fuerte. Mejor modelo. Relación de aceleración de reseñas impulsada por la alta demanda.
	host_acceptance_rate	0.0888	0.2980	Débil. La aceptación no predice la frecuencia de reseñas.
host_total_listings_count	calculated_host_listings_count	0.2883	0.5369	Moderado. Afectado por multicolinealidad. Fracasa en predecir inventario alto.

	host_acceptance_rate	0.0402	0.2006	Muy Débil. La aceptación no predice el tamaño de la operación.
accommodates	price	0.2267	0.4761	Moderado. Muestra la saturación: la capacidad deja de crecer rápidamente a precios altos.
	review_scores_value	0.0014	0.0376	Nulo. Falla conceptual: la calidad no predice el tamaño físico.
price	estimated_occupancy_l365d	0.1581	0.3977	Débil a Moderado. Precio con tendencia de aceleración en rangos de ocupación alta.
	review_scores_value	0.0073	0.0854	Nulo. La calidad percibida no predice el precio.
host_acceptance_rate	reviews_per_month	0.0888	0.2980	Débil. La voluntad de aceptar reservas tiene muy poca influencia en la frecuencia de reseñas, que es impulsada por la demanda (ocupación).
	estimated_occupancy_l365d	0.1671	0.4088	Débil a Moderado. La aceptación se satura (se aplana) en los niveles de ocupación más altos.

Conclusiones generales.

El análisis se inició con una Actividad 1 de regresiones lineales que, tras una robusta limpieza de datos, reveló dos problemas cruciales: multicolinealidad extrema en variables redundantes (como métricas de inventario) y un poder predictivo generalmente débil (R^2 entre 0.20 y 0.30) al modelar variables clave de negocio. Este fallo en la linealidad, sumado a la evidencia de relaciones segmentadas, estableció la necesidad de aplicar regresiones no lineales en la Actividad 2 para capturar umbrales y aceleraciones.

Los modelos no lineales demostraron ser cruciales para validar la naturaleza de estas relaciones:

- *Éxito con la Demanda:* El mejor modelo fue la predicción de Frecuencia de Reseñas por Mes (reviews_per_month) a partir de la Ocupación Estimada Anual (estimated_occupancy_l365d), logrando un alto R^2 de 0.5683. La curva fue marcadamente exponencial, revelando un efecto de aceleración donde las reseñas crecen cada vez más rápido a medida que la ocupación aumenta (más allá de los 150 días).
- *Saturación y Rendimiento*: Se confirmaron dinámicas de saturación (rendimientos decrecientes) en dos modelos clave:

    - La Tasa de Aceptación (host_acceptance_rate) se aplana cerca del 100% cuando se relaciona con la Tasa de Respuesta (R^2 =0.2413) o la Ocupación Estimada (R^2 =0.1671).
    - La Capacidad (accommodates) también se aplana a precios altos (R^2 =0.2267), indicando que el precio adicional en el rango superior ya no paga por más capacidad, sino por lujo o ubicación.
- Irrelevancia de Calidad: Los modelos que utilizaron la Puntuación de Valor (review_scores_value) como predictor resultaron ser completamente inútiles (R^2 cerca de 0.00), validando que las métricas de feedback no son funcionales para predecir variables estructurales o de mercado como la capacidad o el precio.

En conclusión general, aunque la Ocupación Estimada fue el factor individual más poderoso para explicar el engagement y el precio, la mayoría de los R^2 se mantuvieron bajos, lo que lleva a la conclusión fundamental de que ninguna métrica de rendimiento puede ser predicha con alta precisión por una única variable, incluso con un modelo no lineal, pues las complejas dinámicas de mercado y gestión requieren la inclusión de múltiples factores. Por lo tanto, se recomienda enfocar los futuros análisis en la construcción de Modelos de Regresión Múltiple No Lineal que combinen los mejores predictores (ej. estimated_occupancy_l365d y host_acceptance_rate). También es crucial validar la segmentación de precios analizando la curva accommodates vs. price por rango de precio para aplicar una gestión de ingresos más precisa. Analíticamente, se debe descartar el uso de variables de puntuación de reseña como predictoras de precio y capacidad, y considerar modelos especializados en conteo o transformación logarítmica para la Reingeniería del Inventario (host_total_listings_count).