Samuel Aranzola Monroy - A01786577
German Gedovius Dalloz - A01782383

Etapa 4a - Modelos predictivos

El objetivo de esta etapa fue construir y evaluar modelos predictivos iniciales que permitieran explicar y estimar el precio de los alojamientos de Airbnb, utilizando la base limpia obtenida en las etapas anteriores. Se trabajó con las variables más relevantes identificadas en el EDA, incluyendo características del alojamiento (capacidad, tipo de habitación, disponibilidad), desempeño económico (ocupación e ingresos estimados) y atributos del anfitrión (superhost, tasas de respuesta y aceptación). Se implementaron modelos de regresión lineal empleando un enfoque incremental para evaluar desempeño, supuestos y oportunidades de optimización.

1. Modelo inicial de regresión

El primer modelo utilizó price como variable objetivo. Las variables explicativas incluyeron: accommodates, minimum_nights, availability_365, estimated_occupancy_l365d, estimated_revenue_l365d, host_response_rate, host_acceptance_rate, host_is_superhost y variables categóricas transformadas mediante one-hot encoding para room_type y neighbourhood. El modelo obtuvo valores de desempeño moderados (R^2 en train y test entre ~0.45 -- 0.55, dependiendo de la base), reflejando que el precio es una variable altamente dispersa y con influencia de múltiples factores externos.

2. Análisis de residuales (numérico y gráfico)

Los residuales del modelo inicial mostraron dispersión heterogénea respecto a los valores ajustados, lo cual indica problemas de homocedasticidad. El histograma reveló asimetría y cola derecha pronunciada, evidenciando que el modelo sufría ruido elevado para precios altos. El QQ-plot confirmó desviaciones de la normalidad, especialmente en los extremos. Estos patrones justificaron la necesidad de una transformación de la variable objetivo.

3. Índices de correlación y potencia

La matriz de correlación mostró que las variables más relacionadas con el precio fueron: accommodates, estimated_revenue_l365d y ciertas categorías de room_type. Sin embargo, las correlaciones lineales fueron moderadas, reflejando que la relación precio-características no es puramente lineal. La potencia predictiva del modelo (evaluada por R^2 y RMSE) mostró diferencias entre el conjunto de entrenamiento y prueba, indicando un ajuste limitado del modelo lineal básico.

4. Normalidad de los datos y residuales

Las pruebas de normalidad (Shapiro-Wilk) aplicadas a los residuales del Modelo 1 arrojaron valores $p < 0.05$, confirmando que no seguían una distribución normal. Los gráficos QQ-plot reforzaron este hallazgo. Esto justificó formalmente la implementación de un modelo optimizado mediante transformación logarítmica.

5. Optimización del modelo

El Modelo 2 utilizó log(price) como variable objetivo. Esta transformación redujo la asimetría de la distribución y mejoró la normalidad de los residuales. El desempeño del modelo optimizado mostró mejoras en R^2 y reducciones en RMSE y MAE tanto en entrenamiento como en prueba. Los residuales fueron más simétricos y presentaron menor heterocedasticidad. El QQ-plot mostró un ajuste más cercano a la distribución normal.

Conclusión e interpretación final

El modelo inicial permitió identificar los factores que influyen en el precio, pero presentó problemas con la distribución asimétrica de la variable objetivo. Tras la optimización mediante el Modelo 2 (log(price)), se obtuvo un mejor ajuste y residuales más estables. La capacidad del alojamiento, el tipo de habitación, la ocupación estimada y el estatus de superhost mostraron ser predictores relevantes. Aunque el modelo no explica completamente la variabilidad del precio, proporciona un punto de partida sólido y funcional para el análisis predictivo que se ampliará y profundizará en la Etapa 4b por el equipo de Uniandes.