Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Ingeniería en Tecnologías de la Información

Reporte regresión logística

Elaborado por:

Adylene Baylon Cuahtlapantzi
Karla Irais De Florencio Romero
Elias Granillo Castro

31 de marzo de 2025

Contenido

INTRODUCCIÓN

Airbnb maneja grandes volúmenes de datos sobre sus alojamientos y anfitriones, pero sabemos que no cuenta con herramientas que le permitan analizarlos. Por esta razón, hemos estado trabajando en analizar estos datos, en este caso realizamos un análisis de Regresión Logística con el objetivo de identificar patrones que ayuden a reconocer posibles áreas de mejora dentro de la plataforma. En particular, buscamos determinar qué factores influyen en ciertas variables clave y cómo estos pueden optimizar la toma de decisiones.

La regresión logística es una técnica de análisis de datos que utiliza modelos matemáticos para encontrar relaciones entre variables. A partir de estas relaciones, permite estimar la probabilidad de que ocurra un determinado evento en función de un conjunto de variables independientes. Generalmente, se emplea en problemas de clasificación donde la variable objetivo tiene un número finito de categorías, como responder si un alojamiento será reservado o no, o si un anfitrión alcanzará el estatus de "Superhost".

La regresión logística es especialmente útil cuando se trabaja con variables categóricas y dicotómicas, es decir, aquellas que solo pueden tomar dos valores, como "sí" o "no". Sin embargo, también tiene ciertas limitaciones. No es recomendable utilizarlo cuando se busca predecir valores continuos, como el precio de una propiedad, ya que en estos casos es más adecuado aplicar un modelo de regresión lineal. Además, si la relación entre las variables es altamente no lineal o si las clases están muy desbalanceadas, otros modelos, como los árboles de decisión o las redes neuronales, pueden ofrecer mejores resultados.

RESUMEN

Dado que la regresión logística es un modelo relativamente simple e interpretable, resulta una herramienta útil para analizar qué factores influyen en los resultados y tomar decisiones basadas en datos. En este análisis, lo aplicaremos en el contexto de Airbnb y evaluaremos su utilidad para identificar patrones clave dentro de la plataforma.

Cabe destacar que, en el presente reporte, se presentarán los pasos y procedimientos que se siguieron para obtener un modelo bien entrenado, para cada ciudad: México, Albany, Valencia y Bankoog. Asimismo, al final se expondrán los resultados obtenidos y una breve comparación entre estos.

- ## Variables objetivo

DEPENDIENTES

En este análisis, se seleccionaron **10 variables dependientes** para evaluar la correlación logística entre diferentes factores de nuestra base de datos, aplicando la técnica de **Regresión Logística**.

El proceso de selección se realizó a partir de un **análisis comparativo** sobre un conjunto inicial de 50 variables, eligiendo aquellas que mostraban mayor relevancia y significado lógico dentro del contexto del estudio.

Los criterios principales para la selección fueron:

- **Calidad de los datos**: Variables sin valores nulos excesivos ni sesgos significativos.
- **Adecuación al modelo logístico**: Variables binarias o categóricas con dos clases.

Teniendo en cuenta lo anterior, las 10 variables dependientes seleccionadas fueron:

1. host_is_superhot
2. host_identify_verified
3. host_acceptance_rate
4. has_availability
5. instant_bookable
6. room_type
7. accommodates
8. availability_365
9. host_response_time
10. host_has_profile_pic

Las 10 variables dependientes seleccionadas fueron utilizadas en distintos modelos de regresión logística, con el objetivo de identificar patrones y relaciones significativas en los datos.

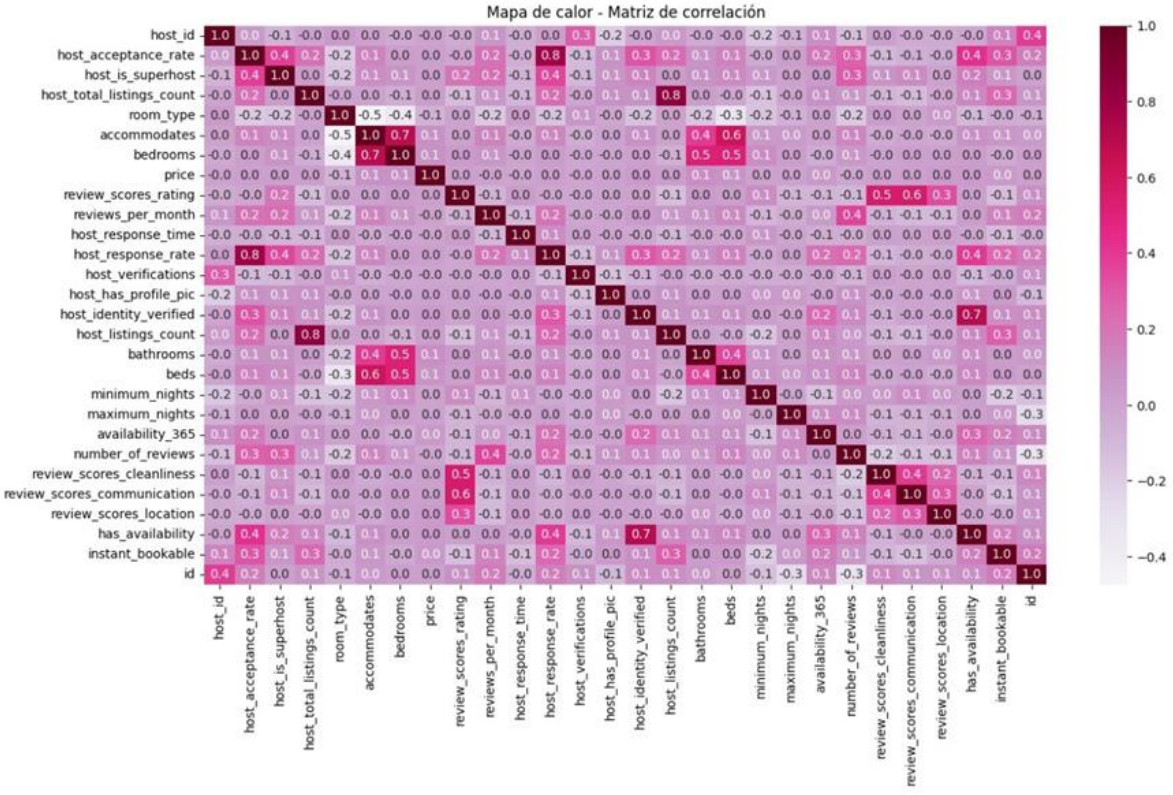
INDEPENDIENTES

Después de haber seleccionado las variables dependientes, seleccionamos las variables independientes con mayor relevancia en la predicción de la variable

dependiente; priorizando aquellas con mayor impacto en la relación lógica y predictiva del modelo.

El proceso de selección se realizó considerando los siguientes criterios:

- **Correlación con la variable dependiente:** Se analizaron coeficientes de correlación para identificar las variables más influyentes, tomando en cuenta el siguiente mapa de calor:



- **Procedimiento**

Sabiendo que primero debemos importar las siguientes librerías:

- ✓ **pandas** para manipulardatos en tablas tipo DataFrame.
- ✓ **numpy** para hacer operaciones numéricas y matrices.
- ✓ **matplotlib.pyplot** para elaborar gráficos y visualización.
- ✓ **scipy.special** para funciones matemáticas avanzadas.
- ✓ **scipy.optimize.curve_fit** nos ayuda con el ajuste de curvas.
- ✓ **seaborn** para gráficos estadísticos avanzados.
- ✓ **sklearn.metrics.r2_score**: métrica para evaluar modelos de regresión.

- ✓ **sklearn.model_selection.train_test_split**: división de datos en entrenamiento y prueba.
- ✓ **sklearn.preprocessing.StandardScaler**: normalización de datos.

Asimismo, debemos cargar el CSV de la base de datos del país correspondiente, en Python para comenzar a hacer el análisis usando la Regresión lógistica.

REEMPLAZAR VALORES

➔ Teniendo en cuenta las variables independientes:

Sabiendo que las variables independientes deben ser numéricas para meterlas al modelo, se convirtieron TODAS las variables independientes seleccionadas que no eran numéricas a numéricas, ya fuera a enteros o flotantes, respectivamente.

➔ Teniendo en cuenta las variables dependientes:

Ya que algunas variables dependientes que fueron seleccionadas para el modelo, no eran dicotómicas, se tuvieron que reemplazar valores por solo dos de ellos: claro está, que se hizo un análisis previo para saber qué datos codificados eran los mejores para reemplazar.

Por ejemplo:

En una primera instancia, se pensaba poner valores de tipo string, sin embargo, considerando que esas variables dicotómicas se pudieran usar en como variables independientes en otro modelo, se reemplazaron por valores binarios de 0 y 1.

host_acceptance_rate 0 aceptacion baja 1 aceptacion alta	Room_type 0 alojamiento compartido 1 alojamiento privado	availability_365 1 disponible todo el año 0 no disponible todo el año	host_response_time 1 within an hour 0 within more than an hour

REGRESIÓN LOGÍSTICA

A continuación, se muestra el procedimiento que se siguió para realizar los modelos de regresión logística, teniendo en cuenta que todas las variables, tanto dependientes (dicotómicas) como independientes (numéricas) ya han pasado por un proceso previo.

Sabiendo lo anterior, analizaremos sólo una variable: HOST_IS_SUPERHOST

1. Declaramos las variables dependientes e independientes para la regresión logística.
2. Redefinimos las variables como X mayúscula e y minúscula.
3. Dividimos el conjunto de datos en la parte de entrenamiento y prueba.
4. Se escalan todos los datos.
5. Realizamos el escalamiento de las variables “X” tanto de entrenamiento como de prueba.
6. Definimos el algoritmo a utilizar (LogisticRegression).
7. Entrenamos el modelo.
8. Realizamos una predicción de los valores de la variable dependiente usando el modelo de regresión logística previamente entrenado.
9. Evaluamos el rendimiento del modelo de regresión logística utilizando la matriz de confusión.

¿Cómo interpretar la matriz de confusión?

[[Verdaderos Negativos Falsos Positivos]

[Falsos negativos Verdaderos positivos]]

10. Calculamos la precisión del modelo de regresión logística, que mide cuántos de los casos que el modelo predijo como positivos realmente lo son.

La precisión se calcula con la fórmula:

$$\text{Precisión} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}}$$

11. Calculamos la exactitud del modelo de regresión logística usando accuracy_score de scikit-learn. Esta función se usa para calcular la **exactitud**, que mide el porcentaje de predicciones correctas sobre el total de casos.

¿Cómo se calcula la exactitud?

$$\text{Exactitud} = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Verdaderos\ Positivos + Verdaderos\ Negativos + Falsos\ Positivos + Falsos\ Negativos}$$

12. Calculamos la sensibilidad del modelo de regresión logística usando recall_score de scikit-learn. Esta función se usa para calcular la **sensibilidad o recall**, mide qué tan bien el modelo detecta los casos positivos.

¿Cómo obtenemos la sensibilidad?

$$\text{Sensibilidad} = \frac{Verdaeros\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

- ## Resultados

Después de hacer y entrenar los diez modelos de regresión logística, aplicando los criterios y procedimiento anteriormente explicados; se obtuvieron los siguientes datos para México, Albany, Valencia y Bankoog.

TABLA COMPARATIVA

CIUDAD	MATRIZ DE CONFUSIÓN	PRECISIÓN DEL MODELO	EXACTITUD DEL MODELO	SENSIBILIDAD DEL MODELO
host_is_superhost				
México	[[7576 2271] [2434 3669]]	0.6176767676767677	0.7050156739811912	0.7693713821468467
Albany	[[34 82] [23 114]]	0.5816326530612245	0.5849802371541502	0.29310344827586204
Valencia	[[3851 241] [1381 196]]	0.448512585812357	0.713882518962778	0.9411045943304008
Bankoog	[[10092 685] [3473 798]]	0.5380984490896831	0.7236842105263158	0.9364387120720052
host_identify_verified				

México	[[241 135] [83 7516]]	0.9823552476800418	0.9726645768025078	0.6409574468085106
Albany	[[0 13] [0 114]]	0.8976377952755905	0.8976377952755905	0.0
Valencia	[[0 69] [0 2766]]	0.9756613756613757	0.9756613756613757	0.0
Bankoog	[[673 64] [46 6741]]	0.9905951506245407	0.9853801169590644	0.9131614654002713
host_acceptance_rate				
México	[[0 1931] [0 6044]]	0.7578683385579937	0.7578683385579937	0.0
Albany	[[0 21] [0 106]]	0.8346456692913385	0.8346456692913385	0.0
Valencia	[[0 71] [0 2764]]	0.9749559082892416	0.9749559082892416	0.0
Bankoog	[[1022 1592] [617 4293]]	0.7294817332200509	0.7064061669324827	0.3909716908951798
has_availability				
México	[[150 55] [13 5099]]	0.9893286767559177	0.9872108331766033	0.7317073170731707
Albany	[[85 90] [60 75]]	1.0	1.0	0.0
Valencia	[[0 43] [0 1847]]	0.9772486772486773	0.9772486772486773	0.0
Bankoog	[[385 40] [11 4580]]	0.9913419913419913	0.9898325358851675	0.9058823529411765
instant_bookable				
México	[[4843 0] [3132 0]]	0.0	0.6072727272727273	1.0
Albany	[[94 0] [33 0]]	0.0	0.7401574803149606	1.0

Valencia	[[1342 271] [844 378]]	0.5824345146379045	0.6067019400352733	0.8319900805951643
Bankoog	[[4203 133] [3017 171]]	0.5625	0.5813397129186603	0.9693265682656826
room_type				
México	[[2731 0] [0 5244]]	1.0	1.0	1.0
Albany	[[37 0] [0 90]]	1.0	1.0	1.0
Valencia	[[851 0] [0 1984]]	1.0	1.0	1.0
Bankoog	[[2634 0] [0 4890]]	1.0	1.0	1.0
accommodates				
México	[[2450 2093] [305 3127]]	0.5990421455938697	0.6993103448275862	0.5392912172573319
Albany	[[37 28] [2 60]]	0.6818181818181818	0.7637795275590551	0.5692307692307692
Valencia	[[800 626] [36 1373]]	0.6868434217108554	0.7664902998236331	0.5610098176718092
Bankoog	[[2436 2068] [309 3162]]	0.9744782666489432	0.7019435736677117	0.5408525754884547
availability_365				
México	[[7612 0] [363 0]]	0.0	0.9544827586206897	1.0
Albany	[[118 1] [8 0]]	0.0	0.9291338582677166	0.9915966386554622
Valencia	[[2795 0] [40 0]]	0.0	0.9858906525573192	1.0
Bankoog	[[0 192] [1 7331]]	0.9744782666489432	0.9743487506645402	0.0

host_response_time				
México	[[3 1444] [15 6513]]	0.8185245695613925	0.8170532915360501	0.002073255010366275
Albany	[[9 16] [0 102]]	0.8644067796661017	0.8740157480314961	0.36
Valencia	[[0 621] [0 2214]]	0.7809523809523381	0.7809523809523381	0.0
Bankoog	[[0 1476] [0 6048]]	0.8038277511961722	0.8038277511961722	0.0
host_has_profile_pic				
México	[[1 157] [8 7809]]	0.9802912377604821	0.9793103448275862	0.006329113924050633
Albany	[[0 5] [0 122]]	0.9606299212598425	0.9606299212598425	0.0
Valencia	[[1 134] [3 2697]]	0.9526669021547156	0.9516754850088184	0.007407407407407408
Bankoog	[[0 192] [1 7331]]	0.9744782666489432	0.9743487506645402	0.0

CONCLUSIÓN

Una vez que nosotros observamos detalladamente los resultados obtenidos podemos identificar patrones clave que pueden ser aprovechados para optimizar el desempeño de los anfitriones y las estrategias de crecimiento en la plataforma de Airbnb en las ciudades evaluadas (México, Valencia, Bankoog y Albany). Ya que gracias a este tipo de modelo podemos identificar variables con un impacto positivo o que pueden representar una enorme oportunidad de crecimiento como lo son las siguientes:

Room_type: Los anfitriones pueden priorizar el ofrecer cierto tipo de habitaciones como casa/departamento completo. Haciendo que airbnb priorice este tipo de resultados en sus algoritmos de búsqueda y recomendación, ya que es altamente confiable y demandante para los usuarios.

Has_Availability: Permite el promover el uso de herramientas de gestión de calendarios para los anfitriones, asegurando que la disponibilidad esté siempre actualizada permitiendo incluir esta variable como un filtro prioritario para los viajeros, ya que se pueden predecir correctamente los listados activos.

Host_is_Superhost: Esta variable nos permite identificar y replicar ciertos factores que hacen exitosos a los Super_Hosts en México como por ejemplo las reviews positivos, alta disponibilidad, etc. Y poder ofrecer programas de mentoría o una pequeña guía para los anfitriones en ciudades con bajo desempeño (Albany).

Accommodates: Se puede animar a los anfitriones a actualizar la capacidad real de sus alojamientos, evitando discrepancias con las expectativas de los huéspedes. Mediante campañas de promoción, preferencias en los resultados de búsqueda, etc.

Por lo que podemos ver el uso de la regresión logística en este análisis ha sido fundamental y puede ser una gran herramienta para airbnb para identificar patrones clave y tomar decisiones basadas en datos dentro de la plataforma. A continuación, se destacan los aspectos más relevantes que demuestran su importancia de una manera general:

Priorizar variables como room_type y has_availability para impulsar el crecimiento, dado su alto poder predictivo en el modelo.

Descartar o reformular variables no útiles como lo pueden ser: host_has_profile_pic evitando invertir recursos en características irrelevantes. Mejorando además la recopilación de datos

Clasificar con precisión variables como room_type y has_availability, donde el modelo alcanza un 100% de exactitud o porcentajes considerablemente altos. Esto permitiría confirmar que características en específico son altamente predictivas de manera correcta y poder priorizarlas en estrategias de optimización de listados.

Como podemos observar no solo se trata de impulsar el crecimiento de la plataforma en términos de números o ingresos, sino de crear un círculo virtuoso donde todos ganan: los anfitriones, los huéspedes y el propio Airbnb.

Ya que, para los anfitriones, estos comportamientos/señales son como una brújula. como, por ejemplo, si el modelo revela que actualizar constantemente la disponibilidad (has_availability) aumenta las reservas, un anfitrión en Valencia podría dedicar solo 5 minutos al día a ajustar su calendario y ver resultados más agradables. O si descubre que su tasa de aceptación (host_acceptance_rate) es baja, recibiría recomendaciones automáticas para mejorar, como activar respuestas rápidas o ajustar sus políticas de reserva. Es decir, ya no están adivinando si funciona de manera adecuada; tienen herramientas concretas para tomar el control de su éxito.

Para la plataforma, esto significa un ecosistema más eficiente y confiable. Al identificar variables clave como room_type (que tiene un 100% de precisión), Airbnb puede priorizar estas características en sus búsquedas y algoritmos de recomendación, asegurando que los huéspedes encuentren exactamente lo que buscan. Y finalmente para los huéspedes, la experiencia se vuelve más transparente y satisfactoria.