Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Inteligencia de Negocios

Reporte Regresión Logística

Cesar Serafín Sampallo Amador

Diego Ricardo Rodríguez Gonzalez

Introducción

La regresión logística es un método de análisis estadístico ampliamente utilizado para modelar relaciones entre una variable dependiente categórica y una o más variables independientes. A diferencia de la regresión lineal, la regresión logística se emplea cuando el resultado a predecir es discreto, como la clasificación de datos en distintas categorías.

En esta actividad se implementó un modelo de regresión logística en Python utilizando los datos de Airbnb de: Cdmx, Tasmania y Paris. El objetico fue identificar patrones e insights valiosos, así como obtener la exactitud, precisión y sensibilidad del modelo, así para determinar cual fue el mejor modelo.

Desarrollo

Para esta parte se seleccionaron 10 variables (columnas) para realizar la regresión logística. A continuación, se detalla el proceso y los resultados obtenidos para las variables analizadas.

Regresión Logística: "host_is_superhost"

En este caso, la variable dependiente es "host_is_superhost", que indica si un anfitrión tiene el estatus de Superhost en la plataforma. Como variables independientes se seleccionaron:

- price (precio del alojamiento)
- host_total_listings_count (número total de alojamientos del anfitrión)
- review_scores_communication (puntaje de comunicación en las reseñas)

Regresión Logística: “instant_bookable”

En este caso, la variable dependiente es " instant_bookable", que indica si el Airbnb se puede reservar instantáneamente en la plataforma. Como variables independientes se seleccionaron:

- number_of_reviews (número de reviews)
- host_total_listings_count (número total de alojamientos del anfitrión)
- review_scores_checkin (puntaje al checking de la habitación)

Regresión Logística: “host_response_time”

En este caso, la variable dependiente es " host_response_time", que indica el tiempo de respuesta del anfitrión en la plataforma. Esta variable se convirtió en dicotómica*, 'a few days or more' 'within a day' y 'within a few hours'* se convirtieron a “*NO within an hour”,* así para analizar correctamente a *within an hour.*

Como variables independientes se seleccionaron:

- number_of_reviews (número de reviews)
- host_total_listings_count (número total de alojamientos del anfitrión)
- review_scores_communication (puntaje a la comunicación del anfitrión)

Regresión Logística: “room_type”

En este caso, la variable dependiente es " room_type", que indica el tipo de cuarto en la plataforma. Esta variable se convirtió en dicotómica*, 'Entire home/apt','Hotel room', 'Shared room'* se convirtieron a “*NO Private room”,* así para analizar correctamente a *Private room.*

Como variables independientes se seleccionaron:

- bedrooms (número de habitaciones)
- price (precio de renta)
- beds (número de camas)

Regresión Logística: “review_scores_cleanliness”

En este caso, la variable dependiente es **"review_scores_cleanliness"**, que representa la puntuación de limpieza otorgada por los huéspedes. Esta variable fue transformada en una variable dicotómica, donde se asignó el valor **1 si la limpieza es mayor o igual a 4.8**, y **0 si es menor**, con el objetivo de analizar los factores que influyen en obtener una calificación alta en limpieza.

Como variables independientes se seleccionaron:

- **room_type** (tipo de habitación ofrecida)
- **price** (precio de la propiedad)
- **host_is_superhost** (si el anfitrión tiene la distinción de “superhost”)

Regresión Logística: “host_identity_verified”

En este caso, la variable dependiente es **"host_identity_verified"**, que indica si la identidad del anfitrión fue verificada por la plataforma. Esta variable fue transformada en dicotómica, donde **1 representa que el anfitrión sí está verificado (“t”)** y **0 que no está verificado (“f” o “No Hay”)**, permitiendo analizar los factores que influyen en esta verificación.

Como variables independientes se seleccionaron:

- **price** (precio de la propiedad)
- **number_of_reviews** (cantidad de reseñas recibidas)
- **review_scores_communication** (calificación otorgada al anfitrión en su comunicación con los huéspedes)

Regresión Logística: “has_availability”

En este caso, la variable dependiente es **"has_availability"**, que indica si una propiedad tiene disponibilidad en la plataforma. Esta variable se codificó de forma binaria para representar **1 si hay disponibilidad** y **0 si no la hay**, con el objetivo de predecir la disponibilidad en función de otros factores.

Como variables independientes se seleccionaron:

- **number_of_reviews** (número total de reseñas recibidas por el anuncio)
- **host_total_listings_count** (cantidad total de propiedades del anfitrión)
- **review_scores_checkin** (calificación promedio que recibe el anfitrión en el proceso de check-in)

Regresión Logística: “review_scores_communication”

En este caso, la variable dependiente es **"review_scores_communication"**, que representa la calificación promedio que los huéspedes otorgan a los anfitriones en cuanto a su capacidad de comunicación. Esta variable fue transformada en binaria, asignando **1 si la puntuación es mayor o igual a 4.8**, y **0 si es menor**, con el propósito de predecir qué factores influyen en una alta valoración en comunicación.

Como variables independientes se seleccionaron:

- **host_response_rate** (tasa de respuesta del anfitrión ante solicitudes)
- **number_of_reviews** (número de reseñas recibidas por el anuncio)
- **price** (precio de la propiedad)

Regresión Logística: “accommodates”

En este caso, la variable dependiente es **"accommodates"**, que indica la cantidad de personas que una propiedad puede alojar. Esta variable fue transformada en binaria, asignando **1 si la propiedad puede alojar a 3 o más personas**, y **0 si puede alojar a menos de 3**, con el objetivo de predecir qué factores están asociados a una mayor capacidad de alojamiento.

Como variables independientes se seleccionaron:

- **bathrooms** (número de baños disponibles en la propiedad)
- **price** (precio de renta de la propiedad)
- **beds** (número de camas)

Regresión Logística: “minimum_nights”

En este caso, la variable dependiente es **"minimum_nights"**, que indica el número mínimo de noches requerido para reservar una propiedad. Esta variable fue transformada en binaria, asignando **1 si el mínimo es mayor o igual a 2 noches**, y **0 si es menor**, con el objetivo de predecir qué factores influyen en establecer un mínimo de estadía más largo.

Como variables independientes se seleccionaron:

- **price** (precio de la propiedad)
- **number_of_reviews** (cantidad de reseñas recibidas)
- **accommodates** (capacidad de alojamiento)

Tablas de resultados CDMX

Regresión logística	Precisión del modelo	Exactitud del modelo	Sensibilidad del modelo	Matriz de confusión
host_is_superhost	0.62	0.60	0.80	[[3739 913] [2237 1086]]
instant_bookable	0.63	0.62	0.90	[[4362 438] [2517 658]]
host_response_time	0.84	0.84	1.0	[[0 1241] [0 6734]]
room_type	0.69	0.79	0.68	[[4516 798] [837 1824]]
review_scores_cleanlines	0.79	0.54	0.47	[[1731 698] [2909 2637]]
host_identity_verified	0.12	0.73	0.78	[[285 78] [1996 5616]]
has_availability	0.13	0.76	0.85	[[274 48] [1825 5828]]
review_scores_communication	0.74	0.66	0.79	[[915 1514] [1128 4418]]
accommodates	0.81	0.82	0.77	[[3839 630] [776 2730]]
minimum_nights	0.47	0.58	0.51	[[3052 1790] [1533 1600]]

Tablas de resultados Tasmania

Regresión logística	Precisión del modelo	Exactitud del modelo	Sensibilidad del modelo	Matriz de confusión
host_is_superhost	0.38	0.53	0.02	[[24 821] [39 985]]
instant_bookable	0.55	0.60	0.70	[[607 248] [492 522]]
host_response_time	0.75	0.75	1.0	[[0 451] [0 1418]]
room_type	0.06	0.85	0.01	[[1602 43] [221 3]]
review_scores_cleanlines	0.93	0.62	0.63	[[83 75] [626 1085]]
host_identity_verified	0.090	0.53	0.80	[[85 20] [854 910]]
has_availability	0.021	0.66	0.73	[[13 2] [594 1260]]
review_scores_communication	0.92	0.65	0.67	[[69 89] [562 1149]]
accommodates	0.90	0.79	0.78	[[446 108] [277 1038]]
minimum_nights	0.69	0.57	0.54	[[471 268] [519 611]]

Tablas de resultados Paris

Regresión logística	Precisión del modelo	Exactitud del modelo	Sensibilidad del modelo	Matriz de confusión
host_is_superhost	0.90	0.6047	0.5777	[[13505 9870] [1449 3815]]
instant_bookable	0.7683	0.7634	0.9799	[[21045 430] [6345 819]]
host_response_time	0.7951	0.7950	0.9999	[[0 5867] [2 22770]]
room_type	0.1217	0.6470	.4600	[[17327 8693] [1414 1205]]
review_scores_cleanlines	0.5695	0.6723	0.3005	[[16211 2301] [7083 3044]]
host_identity_verified	0.1192	0.3797	0.8388	[[2343 450] [17312 8534]]
has_availability	0.0983	0.5485	0.9051	[[1394 146] [12783 14316]]
review_scores_communication	0.4331	0.5841	0.5706	[[10950 7562] [4348 5779]]
accommodates	0.6862	0.7354	0.8922	[[7953 5994] [1583 13109]]
minimum_nights	0.7569	0.5426	0.5843	[[2852 4075] [9024 12688]]

Conclusión

En este análisis, se implementó la regresión logística para modelar distintos aspectos de los alojamientos de Airbnb en Ciudad de México, Tasmania y París. A través de la selección de variables clave, se evaluó la capacidad predictiva del modelo en diferentes escenarios, obteniendo métricas como precisión, exactitud y sensibilidad para cada caso.

Los resultados indican que algunas variables presentan un desempeño predictivo notablemente superior a otras. Por ejemplo, "host_response_time" alcanzó una precisión y exactitud del 84% y una sensibilidad del 100%, lo que sugiere que los factores seleccionados explican en gran medida el tiempo de respuesta del anfitrión. Asimismo, "accommodates" y "room_type" mostraron valores elevados de precisión y exactitud, reflejando una fuerte relación entre las variables predictoras y la capacidad de alojamiento o el tipo de habitación.