Benemérita Universidad Autónoma de Puebla

Curso: Ingeniería de Software I

Actividad 7 (Regresión Logística)

Docente: PhD Alfredo García Suárez

Alumnos:

Jahir Flores Zaragoza
Mario Aldair Flores Hernández
José Daniel Legy Espinosa

Fecha de Entrega: 31/03/25

Introducción:

En el ámbito del análisis de datos y la inteligencia de negocios, la regresión logística se destaca como una herramienta estadística fundamental para modelar relaciones entre variables categóricas y predictores numéricos o categóricos. Este método es especialmente útil para predecir la probabilidad de que un evento ocurra como, por ejemplo, si un anfitrión de Airbnb es superhost o no, en función de variables independientes.

En este reporte, aplicaremos técnicas de regresión logística para analizar patrones en bases de datos de Airbnb correspondientes a México y otras dos ciudades seleccionadas. El objetivo es identificar relaciones significativas entre distintas variables, como precio, número de reseñas, tasa de respuesta del anfitrión, entre otras, y fenómenos de interés como la verificación de identidad del anfitrión, estatus de superhost, etc.

Objetivo:

El objetivo principal es construir y evaluar modelos predictivos que permitan clasificar correctamente variables dependientes relevantes del entorno de Airbnb, utilizando diferentes combinaciones de características. Este análisis busca responder preguntas como: ¿Es posible predecir si un anfitrión es superhost basándonos en el número de reseñas, precio y número de camas? ¿Qué tan bien se puede anticipar la disponibilidad o el tipo de habitación ofrecida con base en otras variables?

A partir de estas predicciones, se pretende identificar qué variables tienen mayor poder predictivo y cómo varía su efectividad en las diferentes ciudades como México, Chicago, Sydney y Atenas.

Variables dependientes seleccionadas:

1. **host_is_superhost**: Ser superhost representa un sello de calidad para los usuarios. Predecir este atributo ayuda a entender qué factores contribuyen a este estatus, lo cual puede ser valioso.

2. **instant_bookable:** La posibilidad de reservar de forma instantánea es un elemento importante para muchos viajeros. Predecirlo permite entender si ciertos precios o puntuaciones influyen en que un anuncio sea más accesible.

3. **has_availability**: Saber si un alojamiento tiene disponibilidad es fundamental para los usuarios. Evaluar su predictibilidad en función de la disponibilidad anual y el mínimo de noches ayuda a detectar posibles inconsistencias o patrones de oferta.

4. **host_identity_verified**: La verificación de identidad del anfitrión es un factor de confianza para los huéspedes. Predecir este atributo permite explorar si ciertos comportamientos o características de los anfitriones están asociados con este estatus.

5. **host_has_profile_pic**: La presencia de una foto de perfil está asociada a la confiabilidad del host. Evaluar su predictibilidad aporta a la comprensión de qué tan profesional o involucrado es el anfitrión.

6. **room_type**: El tipo de habitación ofrecida (privada, compartida, etc.) influye directamente en la experiencia del huésped. Poder predecirlo en función de precio y número de dormitorios es útil para recomendaciones y segmentación de mercado.

7. **host_response_time**: El tiempo de respuesta de un anfitrión impacta la satisfacción del usuario. Determinar si es predecible ayuda a identificar qué factores influyen en una buena atención.

Tablas de análisis:

host_is_superhost VS number_of_reviews, price, beds				
	MEXICO	**CHICAGO**	**SYDNEY**	**ATENAS**
Precisión T	0.5538	0.6628	0.4673	0.6132
Precisión F	0.6787	0.6267	0.7027	0.6327
Sensibilidad T	0.2671	0.4081	0.1452	0.1715
Sensibilidad F	0.8779	0.8271	0.9242	0.9295
Exactitud	0.6569	0.6368	0.6798	0.6305

instant_bookable VS review_scores_rating, price				
	MEXICO	**CHICAGO**	**SYDNEY**	**ATENAS**
Precisión T	0.574	0.4052	0.55	0.6584
Precisión F	0.622	0.6468	0.7102	0
Sensibilidad T	0.1384	0.086	0.0382	1
Sensibilidad F	0.9324	0.92	0.9868	0
Exactitud	0.6174	0.6283	0.7069	0.6584

has_availability VS availability_365, minimum_nights				
	MEXICO	**CHICAGO**	**SYDNEY**	**ATENAS**
Precisión T	0.9591	0.9931	0.9905	0.9957
Precisión F	0	0	0	0
Sensibilidad T	1	1	1	1
Sensibilidad F	0	0	0	0
Exactitud	0.9591	0.9931	0.9905	0.9957

host_identity_verified VS host_response_rate, host_total_listings_count				
	MEXICO	**CHICAGO**	**SYDNEY**	**ATENAS**
Precisión T	0.9534	0.904	0.9546	0.9806
Precisión F	0	0	0	0
Sensibilidad T	0.9998	1	1	1
Sensibilidad F	0	0	0	0
Exactitud	95.33	0.904	0.9546	0.9806

host_has_profile_pic VS review_scores_communication, beds				
	MEXICO	**CHICAGO**	**SYDNEY**	**ATENAS**
Precisión T	0.9814	0.9786	0.9763	0.9596
Precisión F	0	0	0	0
Sensibilidad T	1	1	1	1
Sensibilidad F	0	0	0	0
Exactitud	0.981	0.9786	0.9763	0.9596

room_type VS price, bedrooms				
	MEXICO	CHICAGO	SYDNEY	ATENAS
Precisión T	0.6998	0.8559	0.8767	0.921
Precisión F	0.6128	0.6693	0.8442	0
Sensibilidad T	0.9096	0.935	0.9751	1
Sensibilidad F	0.2683	0.454	0.4951	0
Exactitud	0.6866	0.8274	0.8726	0.921

host_identity_verified VS calculated_host_listings_count, price				
	MEXICO	CHICAGO	SYDNEY	ATENAS
Precisión 1	0.9579	0.9068	0.9558	0.9797
Precisión 0	0	0	0	0
Sensibilidad 1	1	1	1	1
Sensibilidad 0	0	0	0	0
Exactitud	0.9579	0.9068	0.9558	0.9797

has_availability VS review_scores_cleanliness, review_scores_location				
	MEXICO	CHICAGO	SYDNEY	ATENAS
Precisión T	0.9621	0.9967	0.9876	0.9959
Precisión F	0	0	0	0
Sensibilidad T	1	1	1	1
Sensibilidad F	0	0	0	0
Exactitud	0.9621	0.9967	0.9876	0.9959

host_response_time VS price, number_of_reviews, bedrooms				
	MEXICO	CHICAGO	SYDNEY	ATENAS
Precisión 1	0	0	0	0
Precisión 0	0.9549	0.9653	0.6837	0.9783
Sensibilidad 1	0	0	0	0
Sensibilidad 0	1	1	1	0.9783
Exactitud	0.9549	0.9653	0.6837	0.9783

host_is_superhost VS room_type, bedrooms				
	MEXICO	CHICAGO	SYDNEY	ATENAS
Precisión T	0	0.4915	0	0
Precisión F	0.6249	0.5674	0.6837	0.6086
Sensibilidad T	0	0.159	0	0
Sensibilidad F	1	0.8702	1	1
Exactitud	0.6249	0.5566	0.6837	0.6086

Conclusiones:

- En Host_is_superhost vs number_of_reviews, price, beds la predicción de si un host es superhost basada en reseñas, precio y número de camas tiene un rendimiento aceptable en general, destacando México y Chicago.
- En instant_bookable vs review_scores_rating, price En Atenas, sorprendentemente, se logra una sensibilidad perfecta para la clase verdadera, aunque con precisión baja para la clase falso.
- En has_availability vs availability_365, minimum_nights el modelo muestra una exactitud y sensibilidad perfectas para la clase positiva en todas las ciudades.
- En host_identity_verified vs host_response_rate, host_total_listings_count el modelo identifica casi perfectamente a los hosts verificados, con alta precisión y sensibilidad.
- En host_has_profile_pic vs review_scores_communication, beds el rendimiento del modelo es excelente en todas las ciudades, alta precisión, sensibilidad perfecta para la clase positiva y exactitud elevada.
- En room_type vs price, bedrooms los modelos muestran buen rendimiento general, especialmente en Sydney y Atenas, con alta exactitud y sensibilidad.
- En host_identity_verified vs calculated_host_listings_count, price el modelo predice de forma perfecta los casos positivos, pero no logra identificar los negativos.
- En has_availability vs review_scores_cleanliness, review_scores_location hay una altísima precisión y sensibilidad para los casos positivos en todas las ciudades
- En host_response_time vs price, number_of_reviews, bedrooms la precisión y exactitud solo son altas debido al posible desbalance a favor de una clase.
- En host_is_superhost vs room_type, bedrooms la predicción de superhost basada en tipo de habitación y número de dormitorios tiene bajo desempeño en general, con sensibilidad casi nula.

Los resultados del análisis de regresión logística muestran la efectividad de diferentes variables para predecir atributos clave dentro de la plataforma Airbnb. Se encontró que algunas relaciones, como la disponibilidad de un alojamiento en función

del número de noches mínimas y la disponibilidad anual, se pueden predecir con gran precisión y sensibilidad. Otras, como la verificación de identidad del anfitrión y la presencia de una foto de perfil, también presentan un alto grado de predictibilidad en todas las ciudades analizadas. Sin embargo, en ciertas comparaciones, como la predicción del estatus de superhost basado en el tipo de habitación y número de dormitorios, los modelos no lograron un buen desempeño, con sensibilidades muy bajas. Esto sugiere que algunas variables tienen un mayor impacto en ciertos atributos que en otros, lo que subraya la importancia de elegir adecuadamente los predictores para cada caso de estudio. En general, el análisis demuestra la utilidad de la regresión logística para identificar patrones y factores influyentes en la plataforma, aunque con limitaciones en escenarios donde los datos están desbalanceados o donde las relaciones entre variables son débiles.