Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Profesor: Alfredo García Suárez

Materia: Inteligencia de Negocios

Equipo 6:

Arely Corona Morales | 202032443

José Armando Ramos Ramos | 202154423

Giovani Jimenez Bonilla | 202128781

Primavera 2025

Introducción

La regresión logística es un modelo estadístico ampliamente utilizado para predecir probabilidades de ocurrencia de eventos binarios (dicotómicos), es decir, variables dependientes con dos posibles resultados, como "sí/no", "0/1" o "verdadero/falso". A diferencia de la regresión lineal, que modela relaciones lineales entre variables, la regresión logística utiliza una función sigmoide (o logística) para garantizar que las predicciones sean valores dentro del rango de probabilidad de 0 a 1.

En el siguiente reporte presentamos un análisis acerca de regresión logística, para explorar las características y tendencias en datos de Airbnb, comparando resultados entre cuatro países seleccionados. Para poder identificar patrones que permitan predecir comportamientos clave, como la probabilidad de reserva en función de variables como la limpieza, el tipo de alojamiento, entre otros.

La metodología empleada involucra la evaluación de múltiples variables categóricas y numéricas, seleccionadas dado si tenemos dicotómicas y en caso de que queramos ocupar una variable la podemos pasar a esta misma, ayudándonos a entender mejores aspectos de los datos y así poder realizar las operaciones. Los resultados obtenidos nos permiten no solo comprender diferencias y similitudes en el comportamiento de los usuarios entre los países estudiados, sino también extraer conclusiones que pueden ser aplicadas para optimizar la oferta y mejorar la experiencia del cliente en el sector de alojamiento temporal.

Este análisis es particularmente relevante para desarrollar estrategias basadas en datos que respondan a las demandas específicas de cada mercado y reflejen las variaciones culturales y operativas de los países incluidos en el estudio.

Se ha practicado en un data set inicial para poder entender de manera practica como debemos hacer estas, para después poder usarlas en los data sets propios antes mencionados.

Tabla de relaciones de confusión

Variable dicotómica dependiente “host_is_superhost”

Variables independientes: review_scores_cleanliness, host_response_rate y host_acceptance_rate

País	Sensibilidad	Exactitud	Precisión
Múnich	0.888888	0.503946	0.217096
Portland	0.798733	0.443289	0.220034
Londres	0.425270	0.482343	0.202047
México	0.701278	0.665579	0.600695

Variable dicotómica dependiente “host_response_rate”

Variables independientes: host_acceptance_rate, calculated_host_listings_count y review_scores_communication

País	Sensibilidad	Exactitud	Precisión
Múnich	0.885643	0.707104	0.732271
Portland	0.998723	0.743213	0.773121
Londres	0.936730	0.783963	0.804326
México	0.968876	0.812789	0.828035

Variable dicotómica dependiente “host_acceptance_rate”

Variables independientes: host_response_rate, calculated_host_listings_count y review_scores_communication

País	Sensibilidad	Exactitud	Precisión
Múnich	0.902766	0.651433	0.614639
Portland	0.870221	0.724334	0.614651
Londres	0.976886	0.726630	0.697234
México	0.985929	0.891285	0.899292

Variable dicotómica dependiente “property_type”

Variables independientes: room_type, review_scores_cleanliness y review_scores_communication.

País	Sensibilidad	Exactitud	Precisión
Múnich	1.0	0.997507	0.996252
Portland	0.079253	0.733160	0.059468
Londres	0.738970	0.882010	0.822832
México	0.557390	0.724388	0.577875

Variable dicotómica dependiente “review_scores_cleanliness”

Variables independientes: property_type, room_type y number_of_reviews

País	Sensibilidad	Exactitud	Precisión
Múnich	0.695790	0.600332	0.602061
Portland	0.0	0.474093	0.0
Londres	0.0	0.888658	0.888658
México	0.0	0.968025	0.968025

Variable dicotómica dependiente “review_scores_communication”

Variables independientes: number_of_reviews, reviews_per_month y host_verifications_count

País	Sensibilidad	Exactitud	Precisión
Múnich	0.881921	0.790195	0.801212
Portland	0.788191	0.791289	0.789321
Londres	0.0	0.939343	0.939343
México	0.105244	0.648150	0.661870

Variable dicotómica dependiente “calculated_host_listings_count”

Variables independientes: host_is_superhost, host_response_rate y host_acceptance_rate

País	Sensibilidad	Exactitud	Precisión
Múnich	0.921222	0.687577	0.717162
Portland	0.773321	0.987111	0.439211
Londres	0.079608	0.697964	0.711218
México	1.0	0.862194	0.0

Variable dicotómica dependiente “reviews_per_month”

Variables independientes: host_is_superhost, host_response_rate y host_acceptance_rate

País	Sensibilidad	Exactitud	Precisión
Múnich	0.206963	0.602825	0.611428
Portland	0.0	0.432817	0.342199
Londres	0.966550	0.597839	0.52
México	1.0	0.776677	0.0

Variable dicotómica dependiente “neighbourhood_cleansed”

Variables independientes: property_type, minimum_nights y review_scores_location

País	Sensibilidad	Exactitud	Precisión
Múnich	1.0	0.846281	0.0
Portland	0.0	0.0	0.984232
Londres	1.0	0.832571	0.0
México	1.0	0.874608	0.0

Variable dicotómica dependiente “host_response_time”

Variables independientes: host_response_rate, review_scores_communication y host_acceptance_rate

País	Sensibilidad	Exactitud	Precisión
Múnich	0.764593	0.703780	0.568604
Portland	0.789423	0.943221	0.897312
Londres	0.979349	0.677191	0.663265
México	0.379339	0.788338	0.783799

Conclusión

El análisis muestra que la capacidad predictiva de las variables depende significativamente del país y del tipo de variable dicotómica evaluada. Variables como “host_response_rate” y “host_acceptance_rate” presentan alto desempeño en la mayoría de los países, especialmente en México, lo que indica modelos sólidos. En contraste, variables como “review_scores_cleanliness” y “reviews_per_month” presentan bajos niveles de sensibilidad y precisión, sugiriendo limitaciones en su predicción. Además, se evidencian diferencias marcadas por país, como el excelente rendimiento en Múnich para “property_type” frente al bajo en Portland. Estos resultados destacan la importancia de ajustar los modelos según el contexto geográfico y la variable analizada.

En algunos resultados como la precisión, sensibilidad nos daba cero y uno, por decir un ejemplo host_is_superhost; esto debido a que la precisión (precision_score) mide cuántas de las predicciones positivas ("t") son realmente correctas. Si el modelo dice que todo es "t", pero hay muchos "f" reales que no predice correctamente, entonces la precisión cae a 0.0. Si la precisión (precision_score) es 0.0, pero la sensibilidad (recall_score) es 1.0, el modelo no es útil porque está clasificando todo como "f" (Superhost). Por ende, en los data set la mayoría son f, más del ochenta por ciento, a eso se debe esos resultados. Para poder corregirlos hicimos un balance (en algunos). Dejarlos así es detectar a todos los Superhosts ("t") sin importar los falsos positivos, entonces la alta sensibilidad puede ser útil.