BENEMERITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Inteligencia de Negocios

Reporte: Regresión logística

Docente: Alfredo García Suarez

Integrantes del equipo:

Ángel Gabriel López Alvarado

Angelica Rodríguez Vallejo

Scarlett Itzel Xochicale Flores

Primavera 2025

Fecha: 31 de marzo de 2025

Índice

Introducción

El presente reporte tiene como objetico analizar los datos obtenidos tras realizar un análisis de regresión logística en los datos brindados de Airbnb, principalmente se comparan los datos de las siguientes cuatro ciudades: México, Brasil, Berlín y Nápoles. De esta forma se busca identificar patrones y relaciones entre las variables que influyen en la experiencia del usuario y en la confiabilidad de los anfitriones.

Como se ha visto en las clases anteriores, la regresión logística es una herramienta estadística para modelar relaciones entre una variable dependiente dicotómica y una o varias variables independientes. Por ende, en este estudio se aplicará para predecir características relacionadas con los anfitriones, el tipo de propiedad, así como la disponibilidad de los alojamientos. De esta forma es como se seleccionaron 10 variables claves como dependientes y de las cuales se seleccionaron sus respectivas variables predictoras. Como muestra a continuación:

10 variables generales analizadas

1. host_response_time
2. host_is_superhost
3. beds
4. host_has_profile_pic
5. host_identity_verified
6. room_type
7. has_availability
8. instant_bookable
9. bedrooms
10. minimum_nights

Y en el siguiente segmento se enlistan las variables analizadas con sus variables predictoras:

Variables generales con sus variables predictoras

1. host_response_time (Tiempo de respuesta del anfitrión)
   - host_acceptance_rate
   - Host_is_superhost
   - Number_of_reviews

2. host_is_superhost (Si el anfitrión es Superhost o no)
   - host_response_rate
   - host_acceptance_rate
   - host_total_listings_count

3. beds (Cantidad de camas disponibles)
   - Accommodates
   - Bedrooms
   - Room_type
4. host_has_profile_pic (Si el anfitrión tiene foto de perfil)

- host_listings_count
- host_verifications
- Host_is_superhost

5. host_identity_verified (Si la identidad del anfitrión está verificada)
   - Host_is_superhost
   - review_scores_communication
   - review_scores_rating

6. room_type (Tipo de habitación)
   - property_type
   - accommodates
   - price

7. has_availability (Disponibilidad del alojamiento)
   - price
   - availability_365
   - number_of_reviews

8. instant_bookable (Si la propiedad es reservable de inmediato)
   - maximum_nights
   - minimum_nights
   - host_response_time

9. bedrooms (Cantidad de habitaciones)
   - Room_type
   - Bathrooms
   - Beds

10. minimum_nights (Número mínimo de noches de estancia)
    - price
    - room_type
    - availability_365

A lo largo del reporte, se presentarán los resultados obtenidos para cada una de estas variables mediante modelos de regresión logística mediante una tabla comparativa de la cual se presentan los tres principales coeficientes de análisis: precisión, exactitud y sensibilidad. De esta forma se analizará la influencia de las variables predictoras y se evaluaran los coeficientes comparados de los modelos para cada país. Esto permitirá identificar diferencias y similitudes en la dinámica del mercado de Airbnb en distintas regiones.

Desarrollo

A continuación, se presenta la tabla que muestra los resultados obtenidos para cada una de las variables analizadas en los cuatro países: México, Brasil, Berlín y Nápoles. Para cada variable se muestran los coeficientes de precisión, exactitud y sensibilidad, permitiendo una comparación detallada de los modelos de regresión logística que se han llevado a cabo.

Cada variable representa una característica de los alojamientos de los tres distintos países seleccionados y México en la plataforma Airbnb, y cada una se predice con base en ciertas variables predictoras.

- Precisión: Indica que tan confiable es el modelo al identificar correctamente los casos positivos. Un valor alto significa que el modelo comete pocos errores cuando predice algo como positivo, si la precisión es baja significa que el modelo hace muchas predicciones erróneas.
- Exactitud: Mide que tan bien el modelo clasifica en general. Un valor alto indica que el modelo clasifica correctamente la mayoría de los casos, tantos positivos como negativos, si la exactitud es baja, el modelo falla en general.
- Sensibilidad: Muestra cuántos casos positivos reales el modelo es capaz de detectar. Un valor alto el modelo detecta correctamente la mayoría de los casos positivos, si es baja al modelo se le escapan ejemplos reales.

Tabla comparativa de correlación logística por país

Caso	Coeficientes	México	Brasil	Berlín	Nápoles
host_response_time	Precisión	0.7865830721003135	0.6964035964035964	0.820855614973262	0.8199558173784978
	Exactitud	0.7865830721003135	0.6662158506897484	0.7369181844463718	0.7923844061650045
	Sensibilidad	1.0	0.9131516898087503	0.5650306748466257	0.9183505154639175
host_is_superhost	Precisión	0.6318892591643168	0.7126513490323195	0.7728852838933952	0.7364762768207918
	Exactitud	0.6285893416927 9	0.7043548823 3703	0.7727666955767563	0.7364762768207918
	Sensibilidad	0.9820717131474104	0.9657622739018088	0.9992509363295881	1.0
beds	Precisión	0.8261194029850746	0.8782526115859449	0.0	0.9740102750075551
	Exactitud	0.7936050156739812	0.8649355333153007	0.6204105232726221	0.9740102750075551
	Sensibilidad	0.8611435239206534	0.9771766694843618	0.0	1.0
host_has_profile_pic	Precisión	0.9811912225705329	0.9685330448111081	0.9852558542931483	0.9079624583711777
	Exactitud	0.9811912225705329	0.9685330448111081	0.9852558542931483	0.9066183136899365

	Sensibilidad	1.0	1.0	1.0	0.998335552596538
host_identity_verified	Precisión	0.9543573667711599	0.8363538003786855	0.9095114194854004	0.9271683288002418
	Exactitud	0.9543573667711599	0.8363538003786855	0.9095114194854004	0.9271683288002418
	Sensibilidad	1.0	1.0	1.0	1.0
room_type	Precisión	0.7941983535868287	0.8172757475083057	0.7356495468277946	0.6242578456318915
	Exactitud	0.8425078369905956	0.8716076097736903	0.7421220005782018	0.7428226050166213
	Sensibilidad	0.7348567283278926	0.5170239596469105	0.40448504983388706	0.6433566433566433
has_availability	Precisión	0.9606269592476488	0.9917951492200884	0.9985544955189362	0.9972801450589301
	Exactitud	0.9606269592476488	0.9917951492200884	0.9985544955189362	0.9972801450589301
	Sensibilidad	1.0	1.0	1.0	1.0
instant_bookable	Precisión	0.7487684729064039	0.7718871156793797	0.7831363781539444	0.5788770053475936
	Exactitud	0.6978056426332289	0.7718871156793797	0.7591789534547557	0.6053188274403143
	Sensibilidad	0.7546545304095986	1.0	0.9409056024558711	0.30407303370786515
bedrooms	Precisión	0.8825248392752776	0.8459923428430283	-	0.7917511832319134
	Exactitud	0.8278369905956113	0.8462717518708863	-	0.9821698398307646
	Sensibilidad	0.8547169811320755	0.9373289545703338	-	0.9641827912721285
minimum_nights	Precisión	0.6377277599142551	0.8699846722567848	0.9855449551893611	0.7148014440433214
	Exactitud	0.6171786833855799	0.8699846722567848	0.9855449551893611	0.6119673617407072
	Sensibilidad	0.8584089035449299	1.0	1.0	0.7237762237762237

Análisis de los resultados

De acuerdo con los datos anteriormente agrupados, se presenta a continuación las observaciones generales por cada una de las variables:

1. **Host response time**

Se puede observar que la precisión es alta en los cuatro países, con valores superiores al 0.69. De forma especifica se tiene que Berlín y Nápoles presentan las mejores precisiones con 0.82 y 0.81 respectivamente. En cuanto la sensibilidad México es de 1.0, lo que indica que el modelo

es capaz de identificar todos los casos positivos en ente país. Mientras que en Nápoles también presenta una alta sensibilidad (0.91), sin embargo, Berlín presenta una menor sensibilidad (0.56), lo que sugiere que en este país el modelo tiene dificultades para detectar los casos positivos.

2. **Host is super host**

Para este caso son Berlín y Nápoles quienes tienen los mejores calores de precisión (0.77 y 0.73), mientras que México presenta la menor precisión, indicando que es confiable al momento de identificar correctamente los casos positivamente. La sensibilidad es extremadamente alta en todos los países (superior a 0.96), lo que indica que los modelos son muy buenos detectando casos positivos.

3. **Beds**

Se observa una gran variabilidad en los resultados, por ejemplo, México y Brasil tienen buenas precisiones (0.82 y 0.87 respectivamente), mientras que en Berlín el modelo no logró predecir correctamente esta variable (precisión de 0.0), caso contario de Nápoles que presenta altos valores en todos los coeficientes. Además, la sensibilidad en Brasil es la más alta (0.97), así se tiene que en Nápoles nos da un valor de 1.0, indicando que todos los casos positivos fueron correctamente detectados.

4. **Host has profile pic**

Para este caso todos los países presentan valores de precisión y exactitud muy altos (arriba de 0.90), lo que indica que las variables predictoras demuestran un buen desempeño en la regresión logística para estos cuatro países. Mientras que la sensibilidad es de 1.0 en todos los países a excepción de Nápoles, donde se obtiene un valor cercano con 0.99.

5. **Host identity verified**

En este caso, se presentan valores de precisión y exactitud muy altos en los cuatro países (superiores a 0.83), lo que indica una fuerte relación entre las variables predictoras y la variable objetivo. La sensibilidad es de 1.0 en todos los países, lo que confirma que el modelo logra identificar correctamente todos los casos positivos.

6. **Room type**

México y Brasil presentan valores de precisión de aproximadamente 0.79 y 0.81 respectivamente, mientras que en Berlín y Nápoles los valores son menores (0.73 y 0.62 respectivamente), esto muestra la variabilidad del desempeño esperado del modelo. Y se tiene que la sensibilidad en Berlín es bastante baja (0.40), lo que sugiere que el modelo tiene problemas para identificar correctamente los casos positivos en este país.

7. **Has availability**

Los coeficientes presentan valores extremadamente altos en todos los países, con precisiones y exactitudes superiores a 0.96. La sensibilidad es de 1.0 en todos los países, no ha indicado un desempeño perfecto del modelo en esta variable de acuerdo con sus variables predictoras.

8. **Instant Bookable**

Para el caso de México y Brasil tienen precisiones cercanas a 0.75 y 0.77 respectivamente, pero para el caso que se observa en Nápoles, se tiene que la precisión es significativamente menor (0.57). De igual forma se muestra que la sensibilidad en Nápoles es la más baja con un valor de 0.30, lo que sugiere que el modelo tiene dificultades para predecir correctamente los casos positivos en esta ciudad.

9. **Bedroms**

Para el caso de Berlín no cuenta con datos suficientes para esta variable, debido a que los datos obtenidos en su dataset muestran solo valores de 1, por ende, al momento de realizar dicho análisis no constituye la característica básica (ser dicotómica) para realizar en análisis de regresión logística.

Sin embargo, observamos que, en México, Brasil y Nápoles, las precisiones y exactitudes son altas, con valores superiores a 0.79. Y que la sensibilidad en Brasil y Nápoles es de 0.93 y 0.96 respectivamente, lo que indica un buen desempeño en la identificación de casos positivos.

10. **Minimum Nigths**

Para esta última variable, tenemos que Berlín presenta la mayor precisión y exactitud (0.98), lo que sugiere que el modelo es altamente eficiente en esta ciudad, mientras que en México se tiene la menor precisión (0.63), lo que indica que el modelo en este país no es tan preciso en la predicción de esta variable. Por último, la sensibilidad en todos los países es muy alta, destacando valores de 1.0 en Brasil y Berlín.

Conclusión

Los resultados obtenidos evidencian diferencias significativas en el desempeño de los modelos de regresión logística entre México, Brasil, Berlín y Nápoles. En particular, variables como "Has availability" y "Host identity verified" muestran un rendimiento óptimo en todos los países, con valores de precisión, exactitud y sensibilidad cercanos a 1.0, lo que indica una alta capacidad del modelo para predecir correctamente estas categorías.

Por otro lado, variables como "Beds" y "Instant bookable" presentan inconsistencias en su desempeño. En Berlín, por ejemplo, el modelo no logra predecir correctamente la variable "Beds", reflejando una precisión de 0.0 y una exactitud de solo 0.62. Asimismo, la sensibilidad de "Instant bookable" en Nápoles es particularmente baja (0.30), lo que sugiere que el modelo no está capturando adecuadamente los casos positivos en esa región.

Estas diferencias pueden atribuirse a múltiples factores, como la estructura del mercado de Airbnb en cada país, la distribución de las variables predictoras o incluso la calidad y cantidad de datos disponibles para el entrenamiento del modelo, tal es el caso que se observa en la ciudad de Berlín en el análisis de “Bedroms”. De igual forma se establece que un aspecto clave a considerar es el posible desbalance en las clases, debido a que se observó a lo largo del análisis la importancia del balance de las clases que están sobrerrepresentadas respecto a otras, afectando así la sensibilidad del modelo.

En términos generales, los modelos presentan una capacidad predictiva aceptable, pero en algunos casos se observan deficiencias en la sensibilidad, lo que sugiere que podrían explorarse ajustes adicionales. Entre las estrategias para mejorar estos modelos se encuentran la mejora en el balance de clases, así como la incorporación de nuevas variables predictoras que podrían mejorar la complejidad de las relaciones entre las variables presentadas para cada caso.