



Regresión logística

Integrantes:

- Yuliana Casanova López
- Darina Jocelyn Espinosa Tlatelpa
- Dylan Ibrahim Marín Alcalá

Descripción de los casos de correlación logística

Caso	Variable dependiente	Variables independientes	Justificación
1	host_is_superhost	<ul style="list-style-type: none"> • number_of_reviews • review_scores_rating • review_scores_accuracy • review_scores_cleanliness • review_scores_checkin • review_scores_communication • review_scores_location • review_scores_value • host_response_rate • host_acceptance_rate • availability_365 	Se eligió host_is_superhost como variable dependiente por ser un indicador de calidad del anfitrión. Las variables independientes seleccionadas reflejan la experiencia del huésped el comportamiento del host y la disponibilidad del alojamiento, todos factores clave que pueden influir en ser o no superhost.
2	host_identity_verified	<ul style="list-style-type: none"> • number_of_reviews • review_scores_rating • host_acceptance_rate • availability_365 	Se eligió host_identity_verified como variable dependiente por su importancia en la confianza del usuario. Las variables independientes seleccionadas reflejan la actividad del host, la calidad del alojamiento y su comportamiento ante solicitudes, factores que pueden estar relacionados con la verificación de identidad.
3	has_availability	<ul style="list-style-type: none"> • availability_30 • availability_90 • availability_365 	Se eligió has_availability como variable dependiente porque indica si un alojamiento está activo. Las variables independientes reflejan la disponibilidad real en distintos periodos, por lo que están directamente relacionadas con el estado de disponibilidad general del anuncio.
4	instant_bookable	<ul style="list-style-type: none"> • number_of_reviews • review_scores_rating • host_listings_count 	Se eligió instant_bookable como variable dependiente por su impacto en la

		<ul style="list-style-type: none"> • host_acceptance_rate • availability_365 • minimum_nights • accommodates • 	<p>facilidad de reserva. Las variables independientes consideran la experiencia del anfitrión, la calidad y actividad del anuncio y características del alojamiento que pueden influir en permitir reservas instantáneas.</p>
5	review_scores_rating	<ul style="list-style-type: none"> • number_of_reviews • review_scores_accuracy • review_scores_cleanliness • review_scores_checkin • review_scores_communication • review_scores_location • review_scores_value • accommodates • minimum_nights • maximum_nights • host_listings_count • host_total_listings_count 	<p>Se eligió review_scores_rating como variable dependiente por ser una medida global de satisfacción del huésped. Las variables independientes incluyen aspectos específicos del servicio, características del alojamiento y la experiencia del anfitrión, todos factores que pueden influir en la calificación general.</p>
6	number_of_reviews	<ul style="list-style-type: none"> • availability_30 • availability_90 • availability_365 • host_listings_count • host_total_listings_count • accommodates • minimum_nights • maximum_nights • price 	<p>Se eligió number_of_reviews como variable dependiente por reflejar la actividad y visibilidad del anuncio. Las variables independientes incluyen disponibilidad, características del alojamiento y nivel de experiencia del anfitrión, factores que pueden influir en la cantidad de reseñas recibidas.</p>
7	price	<ul style="list-style-type: none"> • availability_30 • availability_90 • availability_365 • accommodates • minimum_nights • maximum_nights • bedrooms • beds • host_listings_count 	<p>Se eligió price como variable dependiente por ser clave en la estrategia del anfitrión. Las variables independientes consideran la disponibilidad, capacidad, duración mínima y características físicas del alojamiento, así como la experiencia del anfitrión,</p>

			todos elementos que pueden influir en el precio por noche.
8	host_response_rate	<ul style="list-style-type: none"> • availability_30 • availability_90 • availability_365 • host_listings_count • host_total_listings_count • host_acceptance_rate • price 	Se eligió host_response_rate como variable dependiente por su relación con la atención al cliente. Las variables independientes reflejan la carga de trabajo del anfitrión, su disponibilidad, comportamiento de aceptación y precio, factores que pueden influir en su rapidez para responder.
9	host_acceptance_rate	<ul style="list-style-type: none"> • availability_30 • availability_90 • availability_365 • host_listings_count • host_total_listings_count • host_response_rate • price 	Se eligió host_acceptance_rate como variable dependiente por su importancia en la experiencia de reserva. Las variables independientes consideran la disponibilidad, la carga de anuncios, la tasa de respuesta y el precio, factores que pueden afectar la disposición del anfitrión a aceptar solicitudes.
10	availability_30	<ul style="list-style-type: none"> • availability_365 • availability_90 • host_listings_count • host_total_listings_count • host_response_rate • price 	Se eligió availability_30 como variable dependiente por representar la disponibilidad inmediata del alojamiento. Las variables independientes reflejan la disponibilidad general, la carga de anuncios del anfitrión, su nivel de respuesta y el precio, todos elementos que pueden influir en la disponibilidad a corto plazo.

Análisis de los coeficientes de precisión, exactitud y sensibilidad por país

Milán				
Caso	Precisión	Exactitud	Sensibilidad	Justificación
1	0.7148846960167715	0.7823397075365579	0.9472254559565386	Buen equilibrio entre precisión y sensibilidad. Datos balanceados y bien representados.
2	0.9399606299212598	0.9399606299212598	0.0	Alta precisión pero sensibilidad nula. Posible desbalance: el modelo no detecta casos positivos.
3	0.999437570303712	0.999437570303712	1.0	Resultados casi perfectos. Dataset bien balanceado y features altamente predictivos.
4	0.7638062871707731	0.7989313835770528	0.781618224666143	Desempeño sólido y balanceado. Modelo capta patrones relevantes en los datos.
5	0.9888086131180054	0.9855174353205849	0.26851851851851855	Alta precisión, baja sensibilidad. Modelo conservador: solo predice positivos si está muy seguro.
6	0.5230769230769231	0.702334083239595	0.9937900641025641	Sensibilidad alta, pero baja precisión. Modelo agresivo: predice muchos positivos
7	0.7101620526671168	0.6936164229471317	0.29717560376586166	Desempeño general bajo. Posible desbalance o falta de variables predictivas relevantes.

8	0.0	0.9800337457817773	1.0	Precisión 0.0 indica que el modelo no predijo positivos. Clásico caso de dataset desbalanceado.
9	0.4318181818181818	0.8979190101237345	0.9960888610763454	Alta sensibilidad, baja precisión. Modelo exagera al predecir positivos, generando muchos errores.
10	0.9452221545952526	0.9247750281214848	0.876599634369287	Buen desempeño general. Datos de calidad y variables predictivas bien seleccionadas.

CDMX				
Caso	Precisión	Exactitud	Sensibilidad	Justificación
1	0.6485270049099836	0.7296551724137931	0.8313371293932849	Buen equilibrio general. Sensibilidad alta, precisión algo baja, pero el modelo detecta bien los positivos.
2	0.957115987460815	0.957115987460815	0.0	Alta precisión y exactitud, pero sensibilidad 0. El modelo no detectó ningún positivo, posible
3	0.9989968652037617	0.998996865203761	1.0	Desempeño perfecto. El modelo clasifica correctamente todas las instancias. Puede deberse a datos simples o posible sobreajuste.

4	0.690234375	0.7306583072100313	0.8365959200494539	Modelo balanceado. Precisión y sensibilidad razonables, muestra aprendizaje adecuado.
5	0.9957340025094102	0.9953605015673981	0.05555555555555555	Alta precisión, pero sensibilidad muy baja. Modelo conservador que evita predecir positivos.
6	0.5879458794587946	0.6761128526645768	0.936178319679939	Alta sensibilidad, baja precisión. Modelo sobrepredice positivos, generando falsos positivos.
7	0.7636212624584717	0.7436990595611285	0.48385926731955026	Desempeño aceptable, pero sensibilidad baja. El modelo se pierde muchos positivos reales.
8	0.0	0.9796865203761755	1.0	Alta sensibilidad pero precisión nula. El modelo detecta todos los positivos reales, pero se equivoca en cada predicción positiva que hace, lo que indica un exceso de falsos positivos y mala capacidad para discriminar entre clases
9	0.3333333333333333	0.9410658307210031	0.9970729111229377	Muy alta sensibilidad con

				baja precisión. El modelo prioriza no perder positivos, pero comete muchos falsos positivos.
10	0.8908843537414965	0.8846394984326019	0.9041128646580583	Buen equilibrio entre precisión, exactitud y sensibilidad. El modelo generaliza bien y es confiable para ambas clases.

Hong Kong				
Caso	Precisión	Exactitud	Sensibilidad	Justificación
1	0.6097560975609756	0.8874007936507936	0.9727582292849035	Alta sensibilidad con precisión moderada. El modelo detecta bien los positivos, pero comete algunos
2	0.966765873015873	0.966765873015873	0.0	Alta precisión y exactitud, pero sensibilidad nula. No detecta positivos, lo que indica un problema de desbalance.
3	0.9945436507936508	0.9945436507936508	1.0	Desempeño perfecto. Clasifica correctamente todos los casos. Posible sobreajuste o datos fáciles de separar.

4	0.6765799256505576	0.8035714285714286	0.9429508196721311	Buen balance general. Precisión y sensibilidad elevadas, modelo confiable y estable.
5	0.9898167006109979	0.9851190476190477	0.6774193548387096	Muy buena precisión y exactitud, aunque sensibilidad algo limitada. El modelo es conservador al detectar positivos.
6	0.674074074074074	0.8308531746031746	0.972972972972973	Alta sensibilidad con precisión media. El modelo busca no perder positivos, pero con ciertos falsos
7	0.7496671105193076	0.751984126984127	0.5091383812010444	Buen desempeño, aunque con sensibilidad baja. Puede estar perdiendo varios positivos verdaderos.
8	0.4791666666666667	0.9156746031746031	0.9864718614718615	Alta sensibilidad pero precisión baja. El modelo predice muchos positivos, algunos incorrectos.
9	0.856020942408377	0.8249007936507936	0.9016979445933869	Buen equilibrio. Alta sensibilidad y precisión aceptable, con exactitud moderada.
10	0.9375	0.9226190476190477	0.9634146341463414	Desempeño general alto. Buen balance entre todas las

				métricas. Modelo confiable.
--	--	--	--	-----------------------------

Amsterdam				
Caso	Precisión	Exactitud	Sensibilidad	Justificación
1	0.5970873786407767	0.8492063492063492	0.9657166460140438	Alta sensibilidad con precisión baja. El modelo detecta la mayoría de positivos pero comete muchos
2	0.9685990338164251	0.9685990338164251	0.0	Alta precisión y exactitud, pero sensibilidad nula. El modelo no detectó positivos, probablemente por desbalance.
3	0.961352657004831	0.961352657004831	1.0	Resultados excelentes. Alta precisión, exactitud y sensibilidad. Modelo equilibrado y confiable.
4	0.779874213836478	0.8668046928916494	0.9855550969872059	Precisión y sensibilidad altas. El modelo clasifica de forma efectiva ambas clases.
5	0.8978217821782178	0.8702553485162181	0.49707602339181284	Buena precisión, pero baja sensibilidad. Modelo conservador que pierde muchos positivos.
6	0.0	0.7060041407867494	1.0	Sensibilidad perfecta pero

				precisión nula. El modelo predice todos los positivos, pero falla en todos ellos.
7	0.0	0.9989648033126294	1.0	Sensibilidad perfecta pero precisión nula. El modelo predice todos los positivos, pero falla en todos ellos.
8	0.7652068126520681	0.805728088336784	0.8983675618746709	Desempeño sólido. Alta sensibilidad y precisión aceptable. El modelo generaliza bien.
9	0.6118769883351007	0.8371290545203589	0.834762979683973	Buen balance general. Sensibilidad y precisión moderadas con buena exactitud.
10	0.9758454106280193	0.9758454106280193	0.0	Muy buena precisión y exactitud, pero no detecta positivos. Probable desbalance en el dataset.

Conclusiones Generales

La regresión logística es una técnica estadística utilizada para modelar variables categóricas binarias en función de variables independientes. En el análisis realizado, se implementó para predecir variables como *host_is_superhost* y *instant_bookable*, entre otras, utilizando factores como número de reseñas, calificación promedio y disponibilidad.

Se observaron resultados variados según la variable dependiente y el conjunto de datos (por ejemplo, diferencias entre Milán, CDMX, Hong Kong y Ámsterdam).

- *Equilibrio en el modelo:* Algunos casos muestran un buen balance entre precisión, exactitud y sensibilidad, lo que indica que el modelo clasifica de forma consistente ambas clases. Esto sugiere que las variables independientes seleccionadas son predictivas y que los datos están razonablemente balanceados.
- *Problemas de desbalance:* En otros casos se reporta una alta precisión y exactitud junto a una sensibilidad nula o muy baja. Esto significa que, aunque el modelo acierta al clasificar la clase mayoritaria, falla en identificar los casos positivos, lo que suele indicar un desbalance en los datos.