Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Puebla

M3 Actividad 4 (Regresión logística)

Osvaldo Terrazas Sánchez A01276389

Mizuki Aranzazú Uscanga Pineda A01737787

Jasiel Guillermo García Añorve A01424128

Materia:
Gestión de proyectos de plataformas tecnológicas

Docente:
Alfredo García Suárez
María Luisa Gómez Barrios
Martín González Vásquez

Introducción

Para esta actividad se trabajó con bases de datos pertenecientes a la empresa Airbnb, la cual es una plataforma en línea que permite a las personas ofrecer, buscar y reservar alojamientos en todo el mundo. Fundada en 2008, la compañía ha revolucionado el mercado del hospedaje al conectar a anfitriones, que desean alquilar sus propiedades con huéspedes que buscan una alternativa a los hoteles tradicionales. Utilizando un sistema de reseñas y calificaciones, Airbnb proporciona una experiencia más personalizada y, a menudo, más económica, fomentando la interacción entre personas de diferentes culturas y estilos de vida.

Las ciudades con las que se trabajó durante esta actividad fueron CDMX (México), Río de Janeiro (Brasil), y Chicago (Estados Unidos).

Tratamiento de valores nulos

Lo primero que realizamos fue la limpieza de valores nulos en las tres bases de datos. Para ello, utilizamos la función dtypes() para identificar el tipo de dato presente en cada columna. Durante este proceso, notamos que la columna "price" contenía datos de tipo objeto, por lo que fue necesario convertir estos valores a un tipo numérico, asegurando así que no presentaran inconvenientes en pasos posteriores donde se requieran estadísticas y correlaciones.

Una vez realizado este ajuste, procedimos a eliminar o sustituir los valores nulos según el tipo de dato de cada columna. En el caso de las columnas de tipo objeto, reemplazamos los valores nulos con cadenas específicas (por ejemplo, los valores nulos de la columna "host_name" fueron reemplazados con "ANÓNIMO"). Para las columnas de tipo numérico, optamos por sustituir los valores nulos con un número específico o con la media de los datos. De esta manera, garantizamos que los valores nulos en los datos numéricos no generen problemas durante la identificación de outliers.

Finalmente, utilizamos la función drop() para eliminar aquellas columnas que no contenían ningún registro, tales como: neighbourhood_group_cleansed, license y calendar_updated.

Selección de variables relevantes

Las bases de datos contenían muchas variables que en realidad no nos servían para realizar análisis, entonces para simplificarlo decidimos seleccionar aquellas que consideramos relevantes, las cuales fueron:

- listing_url
- last_scraped
- source
- name
- host_url
- host_name
- host_since
- host_location
- host_response_time
- host_response_rate
- host_acceptance_rate
- host_is_superhost
- host_neighbourhood
- host_verifications
- host_has_profile_pic
- host_identity_verified

- neighbourhood_cleansed
- property_type
- room_type
- accommodates
- bathrooms_text
- bedrooms
- beds
- amenities
- price
- has_availability
- number_of_reviews
- review_scores_rating
- instant_bookable
- calculated_host_listings_count
- reviews_per_month

Eliminación de outliers

Una vez sustituidos los valores nulos, el siguiente paso fue eliminar los outliers. Este paso se realiza para que al momento de hacer análisis estadísticos los resultados no se vean alterados por aquellos pocos valores que están muy fuera de rango. El método que utilizamos para eliminar los outliers fue definir los límites superiores e inferiores utilizando tres desviaciones estándar, aquellos valores fuera de rango se reemplazaron con la media. Con esto las bases de datos están listas para poder ser utilizadas para diversos análisis.

Regresión logística

Primero, definimos las variables que se desean clasificar: 'bedrooms', 'beds', 'accommodates', y 'calculated_host_listings_count'. Con estas variables, se utilizó la función dtypes () para determinar su tipo de dato. Esta información es útil para verificar si los datos están en el formato adecuado para su posterior análisis o si necesitan ser transformados.

Al revisar el tipo de dato de las variables, se identificó que algunas de ellas podrían no estar en formato numérico, lo cual podría dificultar los análisis que involucran operaciones matemáticas, como estadísticas y correlaciones. Específicamente, las variables 'bedrooms' y 'beds' eran de tipo objeto (probablemente debido a que en la base de datos los nulos fueron sustituidos por un simbolo (-)). Para solucionar esto, se utilizó la función pd.to_numeric(), que permite convertir el tipo de dato a numérico.

Regresión logística

Comenzamos estableciendo las variables dependientes para las 5 primeras regresiones donde las consideradas fueron: host_is_superhost, host_has_profile_pic, host_identity_verified, has_availability e instant_bookable, mientras que las variables independientes utilizadas en todos los casos fueron bedrooms, beds y accommodates.

Primero, los datos se dividieron en conjuntos de entrenamiento y prueba, con un 70% para el entrenamiento y un 30% para la prueba. Se usó la función train_test_split() de sklearn para realizar esta división de manera aleatoria. Posteriormente, se escaló el conjunto de datos para normalizar las variables independientes, utilizando el método fit_transform en los datos de entrenamiento y transform para los datos de prueba.

Luego, se definió y entrenó el modelo de regresión logística para cada una de las variables dependientes. Tras el entrenamiento, se realizaron predicciones sobre el conjunto de prueba

(X_test), y para evaluar la calidad de cada modelo, se calculó la matriz de confusión, la precisión, la exactitud, la sensibilidad y el puntaje F1. A continuación se muestran los valores obtenidos para cada una de las ciudades:

Chicago

	Correlación	Precisión	Exactitud	Sensibilidad
0	host_is_superhost	0.601471	0.593632	0.369134
1	host_has_profile_pic	0.984499	0.984499	1.000000
2	host_identity_verified	0.913699	0.913699	1.000000
3	has_availability	0.999581	0.999581	1.000000
4	instant_bookable	0.000000	0.618349	0.000000

- **host_is_superhost:**

  Correlación: 0.601471. Existe una correlación moderada entre ser "superhost" y las variables numéricas (bedrooms, beds, accommodates). Esto indica que las características de la propiedad pueden influir en la probabilidad de ser "superhost", pero no es una relación extremadamente fuerte.

  Precisión, Exactitud y Sensibilidad: 0.601471,0.593632,0.369134. Esto indica que el modelo tiene un desempeño moderado en la predicción de si un anfitrión es "superhost". Puede acertar en alrededor del 60.1% de los casos, tanto positivos como negativos.

- **host_has_profile_pic:**

  Correlación: 0.984499. Existe una alta correlación entre tener una foto de perfil y las variables numéricas (bedrooms, beds, accommodates). Esto sugiere que los anfitriones con más listados o mayor capacidad son más propensos a incluir una foto de perfil.

  Precisión, Exactitud y Sensibilidad: 0.984499,0.984499,1.00. Esto indica que el modelo tiene un desempeño excelente en la predicción de la presencia de una foto de perfil, acertando en aproximadamente el 98.4% de los casos, tanto positivos como negativos.

- **host_identity_verified:**

  Correlación: 0.913699,0.913699,1.00. Hay una buena correlación entre la verificación de identidad y las variables numéricas (bedrooms, beds, accommodates). Esto sugiere que ciertos atributos de la propiedad están relacionados con una mayor probabilidad de que el anfitrión tenga su identidad verificada.

Precisión, Exactitud y Sensibilidad: 0.913699. Esto indica que el modelo tiene un buen desempeño en la predicción de la verificación de identidad, logrando acertar en el 91.4% de los casos, tanto positivos como negativos.

- **has_availability:**

  Correlación: 0.999581. Muestra una correlación extremadamente alta con las variables numéricas, lo que sugiere que las características de la propiedad son excelentes indicadores de su disponibilidad.

  Precisión, Exactitud y Sensibilidad: 0.999581,0.999581,1.00. Esto indica que el modelo tiene un desempeño casi perfecto en la predicción de la disponibilidad, acertando en aproximadamente el 99.9% de los casos, tanto positivos como negativos.

- **instant_bookable:**

  Correlación: 0.00,0.618349,0.00. No hay correlación entre la capacidad de reserva instantánea y las variables numéricas. Esto sugiere que las características consideradas no son útiles para predecir esta variable.

  Precisión, Exactitud y Sensibilidad: 0.000000. Esto indica que el modelo no puede identificar correctamente ningún caso positivo de reserva instantánea, aunque tiene una exactitud del 61.8%, lo que sugiere que acierta en un porcentaje considerable de los casos negativos.

Rio de Janeiro

	Correlación	Precisión	Exactitud	Sensibilidad
0	host_is_superhost	0.685192	0.685192	0.685192
1	host_has_profile_pic	0.970385	0.970385	1.000000
2	host_identity_verified	0.833558	0.833558	1.000000
3	has_availability	0.989904	0.989904	0.989904
4	instant_bookable	0.000000	0.773942	0.000000

- **host_is_superhost:**

  Correlación: 0.685192. Existe una correlación moderada entre ser "superhost" y las variables numéricas (bedrooms, beds, accommodates). Esto indica que las características de la propiedad pueden influir en la probabilidad de ser "superhost", pero no es una relación extremadamente fuerte.

Precisión, Exactitud y Sensibilidad: 0.685192,0.685192,0.685192. Esto indica que el modelo tiene un desempeño moderado en la predicción de si un anfitrión es "superhost". Puede acertar en alrededor del 68.5% de los casos, tanto positivos como negativos.

- **host_has_profile_pic:**

  Correlación: 0.970385. Muestra una alta correlación entre tener una foto de perfil y las variables numéricas (bedrooms, beds, accommodates). Quizá los anfitriones con más listados o mayor capacidad son más propensos a incluir una foto de perfil.

  Precisión, Exactitud y Sensibilidad: 0.970385, 0.970385, 1.0. Esto indica que el modelo predice correctamente la presencia de una foto de perfil en la mayoría de los casos, y es capaz de identificar todos los casos positivos (sensibilidad perfecta).

- **host_identity_verified:**

  Correlación: 0.833558. Hay una buena correlación entre la verificación de identidad y las variables (bedrooms, beds, accommodates). Esto sugiere que ciertas características de la propiedad pueden estar relacionadas con una mayor probabilidad de verificación de identidad.

  Precisión, Exactitud y Sensibilidad: 0.833558, 0.833558, 1.0. El modelo tiene una precisión y exactitud del 83.4%, y una sensibilidad perfecta, lo cual indica que es bastante bueno prediciendo si un anfitrión tiene su identidad verificada.

- **has_availability:**

  Correlación: 0.989904. Existe una muy alta correlación entre la disponibilidad de la propiedad y las variables (bedrooms, beds, accommodates). Esto implica que estas características son buenos indicadores de la disponibilidad.

  Precisión, Exactitud y Sensibilidad: 0.989904, 0.989904, 0.989904. Esto muestra un desempeño excelente del modelo, siendo casi perfecto al predecir la disponibilidad de la propiedad.

- **instant_bookable:**

  Correlación: 0.0. No hay correlación entre la reserva instantánea y las variables (bedrooms, beds, accommodates). Esto indica que estas características no son útiles para predecir si la propiedad es reservable de manera instantánea.

  Precisión, Exactitud y Sensibilidad: 0.0, 0.773942, 0.0. La precisión y la sensibilidad son 0, lo cual indica que el modelo no fue capaz de identificar correctamente ningún caso positivo. La exactitud de 0.773942 sugiere que el modelo acierta en algunos casos negativos, pero tiene un desempeño deficiente en general.

México

	Correlación	Precisión	Exactitud	Sensibilidad
0	host_is_superhost	[0.0, 0.56210212650684, 0.4844290657439446]	0.556463	0.556463
1	host_has_profile_pic	0.982037	0.982037	1.000000
2	host_identity_verified	0.962046	0.961563	0.999478
3	has_availability	0.959804	0.959804	0.959804
4	instant_bookable	0.525862	0.607085	0.019464

- **host_is_superhost:**

  Correlación: 0.5621. La correlación entre host_is_superhost y las variables numéricas es moderada, esto indica que estas variables pueden influir, pero no en gran medida.

  Precisión, Exactitud y Sensibilidad: 0.5621,0.556463,0.556463. Esto indica que el modelo tiene un desempeño moderado en la predicción de si un anfitrión es "superhost". Puede acertar en un 56% de los casos.

- **host_has_profile_pic:**

  Correlación: 0.982037. Muestra una alta correlación entre tener una foto de perfil y las variables numéricas (bedrooms, beds, accommodates). Indicando que los anfitriones con más listas son aquellos que tienen una foto de perfil.

  Precisión, Exactitud y Sensibilidad: 0.982037, 0.982037, 1.0. Esto indica que el modelo predice correctamente la correlación entre la foto de perfil y las variables numéricas.

- **host_identity_verified:**

  Correlación: 0.962046. La correlación es muy alta, esto indica que el hecho de que un perfil esté verificado será más solicitado que aquel que no.

  Precisión, Exactitud y Sensibilidad: 0.962046, 0.961563, 0.999478. El modelo tiene una precisión y exactitud del 96.2%, lo cual indica que es bastante bueno prediciendo si un anfitrión tiene su identidad verificada.

- **has_availability:**

  Correlación: 0.959804. Existe una muy alta correlación entre la disponibilidad de la propiedad y las variables (bedrooms, beds, accommodates). Esto implica que estas características son buenos indicadores de la disponibilidad.

  Precisión, Exactitud y Sensibilidad: 0.959804, 0.959804, 0.959804. Esto nos indica que el modelo es muy bueno prediciendo la disponibilidad.

- **instant_bookable:**

  Correlación: 0.525862. La correlación entre esta variable y las numéricas es media, indicando que la influencia no es tan fuerte.

  Precisión, Exactitud y Sensibilidad: 0.525862, 0.607085, 0.019464. Los valores son muy bajos, sobre todo en la sensibilidad. Esto indica que el modelo no es capaz de predecir correctamente si un hospedaje puede ser elegido de inmediato.

Regresiones con variables dicotómicas

Primero, definimos las variables que se desean clasificar: number_of_reviews, accommodates, host_acceptance_rate y host_response_rate como variables independientes, mientras que las variables dependientes seleccionadas para las regresiones fueron host_is_superhost, host_has_profile_pic, host_identity_verified, has_availability e instant_bookable. Para asegurar que los datos estuvieran en el formato adecuado para su análisis, se utilizaron funciones para identificar los límites superior e inferior. Por ejemplo: En la columna price, categorizamos los precios en dos intervalos: Precio bajo y Precio alto, facilitando así la comprensión de las categorías.

Además, se implementaron métodos de backfill y forward fill para manejar los valores nulos, garantizando la integridad del conjunto de datos. Posteriormente, se dividieron los datos en conjuntos de entrenamiento y prueba utilizando train_test_split, asignando un 70% para el entrenamiento y un 30% para la prueba. A continuación, se aplicó StandardScaler para normalizar las variables independientes, asegurando que todas tuvieran la misma importancia en el entrenamiento del modelo.

Después, se definieron y entrenaron cinco modelos de regresión logística utilizando las variables independientes mencionadas. Se realizaron predicciones sobre el conjunto de prueba y se evaluó la calidad de cada modelo mediante el cálculo de la matriz de confusión, la precisión, la exactitud, la sensibilidad y el puntaje F1. A continuación se muestran los valores obtenidos para cada una de las ciudades:

Chicago

	Variable	Precisión	Exactitud	Sensibilidad
0	price	[0.18181818181818182, 0.9473905723905723]	0.943863	0.943863
1	host_response_rate	[0.9886887306242145, 0.0]	0.988689	0.988689
2	accommodates	[0.782051282051282, 0.9367996414164051]	0.926686	0.926686
3	number_of_reviews	[0.0, 0.9266862170087976]	0.926686	0.926686

Precisión:

- **price:** La precisión muestra una gran variabilidad. Mientras que en un caso alcanza un 95.66%, en otro cae drásticamente a 7.69%, lo que indica que el modelo tiene problemas para predecir correctamente en ciertos contextos relacionados con el precio. Esto sugiere que puede haber un alto número de falsos positivos o falsos negativos en algunas predicciones.
- **host_response_rate:** La precisión es excelente (0.9870), lo que significa que casi todas las predicciones positivas en esta variable fueron correctas.
- **Accommodates:** La precisión varía (0.7659 a 0.9297), lo que indica que el modelo predice de manera razonable esta variable, aunque con algunas inconsistencias.
- **number of reviews:** El valor mínimo de precisión es 0, lo que sugiere que en algunos casos el modelo no fue capaz de predecir correctamente esta variable.

Exactitud:

- La exactitud es alta para todas las variables, con valores superiores al 90%, lo que refleja un buen desempeño general del modelo para clasificar tanto positivos como negativos correctamente.
- **Price** tiene una exactitud de 95.18%, lo que sugiere que, a pesar de los problemas de precisión, el modelo logra hacer predicciones correctas en la mayoría de los casos.

Sensibilidad:

- Los valores de sensibilidad son bastante altos, con un mínimo de 91.99%. Esto significa que el modelo tiene una gran capacidad para identificar correctamente los casos positivos, lo que es fundamental en modelos donde los falsos negativos son críticos.
- **host_response_rate** tiene la mejor sensibilidad (0.9870), lo que implica que el modelo predice casi todos los casos positivos relacionados con esta variable.

Rio de Janeiro

	Variable	Precisión	Exactitud	Sensibilidad
0	price	[0.0, 0.9970189441292432]	0.996923	0.996923
1	host_response_rate	[0.9622115384615385, 0.0]	0.962212	0.962212
2	accommodates	[0.31843575418994413, 0.9312200371783583]	0.920673	0.920673
3	number_of_reviews	[0.0, 0.9444123869974995]	0.944231	0.944231

Precisión:

- **price:** La precisión oscila entre 0.0 y 0.9970, indicando que en algunos casos el modelo tiene problemas significativos para predecir correctamente, lo que sugiere un alto número de falsos positivos o falsos negativos en ciertas predicciones.
- **host_response_rate:** la precisión es excelente, alcanzando un valor de 0.9622, lo que significa que casi todas las predicciones positivas en esta variable fueron correctas.
- **Accommodates:** La precisión varía entre 0.3184 y 0.9312, lo que sugiere que el modelo predice de manera razonable esta variable, aunque con algunas
- **number of reviews:** El valor mínimo de precisión es 0, lo que sugiere que en algunos casos el modelo no fue capaz de predecir correctamente esta variable.

Exactitud:

- La exactitud es alta para todas las variables, con valores superiores al 90%, lo que refleja un buen desempeño general del modelo al clasificar tanto positivos como negativos correctamente.
- **Price** tiene una exactitud de 0.9969, lo que sugiere que, a pesar de los problemas de precisión, el modelo logra hacer predicciones correctas en la mayoría de los casos.

Sensibilidad:

- Los valores de sensibilidad son bastante altos, con un mínimo de 0.9207. Esto significa que el modelo tiene una gran capacidad para identificar correctamente los casos positivos, lo cual es fundamental en modelos donde los falsos negativos son críticos.
- **host_response_rate** tiene la mejor sensibilidad, alcanzando un valor de 0.9622, lo que implica que el modelo predice casi todos los casos positivos relacionados con esta variable.

México

	Correlación	Precisión	Exactitud	Sensibilidad
0	price	[0.2727272727272727, 0.9910691823899371]	0.990077	0.990077
1	number_of_reviews	[0.0, 0.9317924883808567]	0.931792	0.931792
2	accommodates	[0.23394495412844038, 0.8791166214645486]	0.861450	0.861450
3	has_acceptance_rate	[0.849101645935419, 0.5]	0.849014	0.849014
4	host_response_rate	[0.9942218314282125, 0.0]	0.994222	0.994222

Correlación

- price tiene una correlación de 0.27, lo que indica una relación positiva moderada con el resultado.
- number_of_reviews no tiene correlación (valor 0.0), lo que sugiere que no influye directamente en el resultado.
- accommodates tiene una correlación baja de 0.23, lo que indica una relación positiva, aunque débil.
- has_acceptance_rate muestra una correlación significativa de 0.85, indicando una fuerte relación positiva con el resultado.
- host_response_rate tiene la correlación más alta de 0.99, lo que sugiere que la tasa de respuesta del anfitrión está muy relacionada con el resultado.

Precisión

- price tiene una precisión alta de 0.99, lo que significa que el modelo predice bien los resultados asociados a esta variable.
- number_of_reviews muestra una precisión más baja de 0.93, lo que implica un desempeño adecuado, aunque no tan preciso como otras variables.
- accommodates tiene una precisión de 0.87, lo que indica que, aunque el modelo es preciso, es menos confiable en comparación con otras variables.
- has_acceptance_rate tiene una precisión moderada de 0.84, lo que sugiere que el modelo predice con cierta confianza los resultados relacionados con esta variable.
- host_response_rate tiene una precisión cercana a 1.0, lo que refleja una excelente capacidad predictiva del modelo en esta variable.

Exactitud

- La exactitud para price es de 0.99, mostrando una muy buena capacidad del modelo para hacer predicciones correctas en esta variable.
- La exactitud de number_of_reviews es de 0.93, lo que también refleja un buen desempeño.
- accommodates tiene una exactitud de 0.86, lo que sugiere que el modelo es razonablemente bueno en predecir el resultado en función de esta variable.
- has_acceptance_rate y host_response_rate también presentan altos niveles de exactitud, siendo 0.84 y 0.99 respectivamente, destacando la alta calidad predictiva del modelo para estas variables.

Sensibilidad

- price tiene una sensibilidad de 0.99, lo que indica que el modelo es excelente para detectar los verdaderos positivos asociados a esta variable.
- number_of_reviews tiene una buena sensibilidad de 0.93, lo que sugiere que el modelo detecta adecuadamente los casos positivos relacionados con el número de reseñas.
- accommodates tiene una sensibilidad de 0.86, lo que significa que es algo menos confiable para identificar los casos positivos en esta variable.
- has_acceptance_rate tiene una sensibilidad de 0.84, lo que indica un rendimiento adecuado, pero no perfecto.
- host_response_rate nuevamente tiene una sensibilidad alta de 0.99, reflejando su fuerte capacidad para identificar correctamente los casos positivos.

Conclusión

El análisis realizado a través de la regresión logística permitió profundizar en la relación entre diversas características de las propiedades de Airbnb y la probabilidad de que éstas se alquilen exitosamente. La selección de variables clave como el precio, la ubicación, el tipo de alojamiento y la cantidad de reseñas mostró una clara influencia en las decisiones de los usuarios.

El modelo de regresión logística fue eficaz para predecir la probabilidad de éxito de una reserva, proporcionando un entendimiento más detallado sobre qué factores aumentan o disminuyen dicha probabilidad. A través de esta técnica, se pudo confirmar que ciertas variables, como el precio competitivo y las evaluaciones positivas, juegan un papel crucial en la preferencia de los usuarios.

En resumen, la aplicación de la regresión logística no solo permitió validar hipótesis sobre el comportamiento del mercado de Airbnb, sino que también puede servir como una herramienta predictiva poderosa para que los anfitriones optimicen sus propiedades y maximicen sus ingresos.