Actividad 3: (Regresión Logística)

Tecnológico de Monterrey, Campus Puebla
Gestión de Proyectos de Plataformas Tecnológicas

Karyme Pérez Chatú A01174367

Objetivo:
Realizar un análisis de la base de datos de nuestro país asignado y generar los modelos de regresión logística para variables dicotómicas con sus transformaciones correspondientes.

Índice:

1. Repositorio

https://github.com/KarymeCh/Actividad3_RegresionLogistica.git

---

2. Introducción

La regresión logística es una técnica estadística utilizada para modelar la probabilidad de que ocurra un determinado evento, especialmente cuando la variable dependiente es categórica o binaria (por ejemplo, *sí/no*, *1/0*). A diferencia de la regresión lineal, que predice valores continuos, la regresión logística estima probabilidades mediante la función logística o sigmoide, transformando los resultados en valores entre 0 y 1.

En el análisis de datos, este modelo se aplica ampliamente para clasificación y toma de decisiones, permitiendo identificar los factores que influyen en un resultado y estimar la probabilidad de su ocurrencia. Algunos ejemplos comunes incluyen la predicción de comportamientos de compra, la detección de fraudes, la evaluación del riesgo crediticio o la diagnosis médica.

Gracias a su interpretación intuitiva y su capacidad para manejar variables tanto cuantitativas como categóricas, la regresión logística es una herramienta esencial en la analítica predictiva y en la construcción de modelos de decisión basados en datos.

---

3. Regresión Logística

3.1. Casos de Regresión logística directa

- Host is superhost
- Hos has profile pic
- Host Identity Verified
- Instant Bookable

Para esta primera sección de variables trabajadas, se pudo identificar una característica en común, y es que eran las únicas de tipo cuantitativas dicotómicas de la base de datos. Considerando esto, se realizó la regresión logística de manera directa, tomando como variables independientes las tres primeras con mayor correlación.

Para dos de las cuatro variables se logró obtener buenos valores de precisión y exactitud del modelo, a excepción de Host has profile pic y Host identity verified, que obtuvieron valores de 0 en la precisión y sensibilidad de la segunda clasificación. Esto nos indica que al correr la relación logística, el modelo no logra predecir los casos de esa clasificación, y que por ende se debe realizar un ajuste para ponderar las clases de esa variable y así lograr una predicción balanceada.

3.2. Ajustes de Ponderación de clases

- Host has profile pic oversampling
- Identify Verified Oversampling

Habiendo identificado las variables que se deben ponderar, se procedió a realizar el método Oversampling, el cual toma la clase minoritaria y aumenta el número de observaciones, logrando la mejora de la capacidad del modelo para detectar ambas clasificaciones.

A través de este método se logró aumentar la precisión de las variables mencionadas de la siguiente manera:

Y	X	P(1)	P(2)	E	S(1)	S(2)
Hos has profile pic	1-Number of reviews 2-Maximum nights avg ntm 3-Calculated host listings count	0.9776	0.0	0.9776	1.0	0.0
Hos has profile pic Oversampling	1-Number of reviews 2-Maximum nights avg ntm 3-Calculated host listings count	0.9912	0.0448	0.6346	0.6318	0.7570
Host Identity Verified	1-Availability eoy 2- Availability 365 3-Availability 90	0.9537	0.0	0.9537	1.0	0.0
Identify Verified Oversampling	1-Availability eoy 2- Availability 365 3-Availability 90	0.9805	0.1215	0.7555	0.7588	0.6885

Como se aprecia en la tabla, los valores de precisión para la segunda clasificación aumentan, y aunque para Host profile pic esta transformación no resulta demasiado grande, si que nos habla de una mejora del modelo al ser capaz de reconocer los valores de la clase dos.

3.3. Casos de Conversión a variables dicotómicas

- Room type (Categórica)
- Price
- Host listings count
- Host response time (Categórica)
- Accommodates
- Minimum nights

Posteriormente, para trabajar con variables que no eran de tipo binario, seleccioné las que me parecían más interesantes para analizar y subdividir. Dos de ellas fueron de tipo categórico, como es el caso de ‘room_type’, de la cual creé dos clases: Entire home/apt y

Rooms. De esta manera clasifiqué todas las respuestas en dos grupos que fueran de interés.
En el caso de las variables numéricas, se realizó una categorización con base en sus límites, para así generar dos categorías y renombrarlas con base en el tipo de información que brinda. Ejemplo de esto es 'Price', que inicialmente contaba con una gran variedad de posibilidades de datos, por lo cual generé dos clasificaciones que dividen la variable en dos por la mitad, al analizar los rangos de precio decidí que sería conveniente mantener los títulos como ' Precios medios' y 'Precios altos', y una vez teniendo únicamente dos posibilidades de respuesta, se procedió a realizar la regresión logística de forma exitosa

---

4. Tabla comparativa de valores

Y	X	P(1)	P(2)	E	S(1)	S(2)
Hos is superhost	1-Estimated occupancy l365d 2-Number of reviews ltm 3- Number of reviews ly	0.7130	0.6107	0.6836	0.8193	0.4623
Hos has profile pic	1-Number of reviews 2-Maximum nights avg ntm 3-Calculated host listings count	0.9776	0.0	0.9776	1.0	0.0
Hos has profile pic Oversampling	1-Number of reviews 2-Maximum nights avg ntm 3-Calculated host listings count	0.9912	0.0448	0.6346	0.6318	0.7570
Host Identity Verified	1-Availability eoy 2- Availability 365 3-Availability 90	0.9537	0.0	0.9537	1.0	0.0
Identify Verified Oversampling	1-Availability eoy 2- Availability 365 3-Availability 90	0.9805	0.1215	0.7555	0.7588	0.6885
Instant Bookable	1-Host acceptance rate 2-Calculated host listings count 3-Minimum nights	0.6762	0.7355	0.7160	0.5560	0.8226
Room type	1-Accommodates 2-Price 3-Bedrooms	0.8680	0.7030	0.8074	0.8342	0.7558
Price	1-Bathrooms 2-Accommodates	0.8353	0.5220	0.8299	0.9901	0.0524

	3-Bedrooms					
Host listings count	1-Host total listings count 2-Estimated revenue l365d 3-Minimum nights avg ntm	0.9998	0.9574	0.9996	0.9997	0.9782
Host response time	1-Host response rate 2-Host acceptance rate 3-Estimated occupancy l365d	0.8705	0.7670	0.8618	0.9762	0.3503
Accommodates	1-Bedrooms 2-Beds 3-Bathrooms	0.8565	0.8298	0.8483	0.9198	0.7172
Minimum nights	1-Minimum minimum night 2-Calculated host listings count 3-Availability 30	0.9818	0.9404	0.9777	0.9934	0.8481

4.1. Análisis general de resultados

Se puede observar un muy buen ajuste del modelo de regresión logística para las variables analizadas, lo cual nos indica que se podrá predecir con efectividad las variables binarias seleccionadas; esto con excepción de 'Host has profile pic' y 'Host identity verified'.

Dichas variables son a las que se le aplica el método de Oversampling, en sus resultados se observa que obtienen los valores de precisión para la segunda clasificación más bajos. Esto se debe a que después de los ajustes realizados, el modelo comienza a aprender sobre la ponderación de ambas clases y comienza a detectarlos, por lo tanto la exactitud baja ya que el dataset se balancea al considerar ambas clasificaciones.

Los valores de precisión para ambas clases con el resto de variables se encuentran bien distribuidos, en su mayoría siendo bastantes altos, lo cual nos indica que los modelos están prediciendo de manera completa la variable binaria con base en ambas clasificaciones.

En el caso de la sensibilidad por clases se observa bastante distribución, lo que nos habla de que los resultados de predicción del modelo pueden variar dependiendo del aumento o decremento de la clase.

4.2. Highlights de resultados

- Los modelos más balanceados y confiables son Room type, Accommodates y Minimum nights.
- Modelos como Host has profile pic o Identity verified requirieron de oversampling para mejorar la detección de clases minoritarias.
- High accuracy no siempre significa buen modelo — lo importante es el equilibrio entre precisión y sensibilidad.

- El oversampling, aunque reduce exactitud, mejora la justicia y la capacidad de generalización del modelo.

---

5. Conclusión

En el análisis realizado, la aplicación de modelos de regresión logística permitió evaluar el desempeño predictivo de diversas variables asociadas al comportamiento de los anfitriones y características de las propiedades. Si bien algunos modelos mostraron valores elevados de precisión y exactitud, estos no siempre se tradujeron en un buen desempeño global, especialmente en contextos con desbalance de clases.

La implementación de técnicas de oversampling mejoró la detección de la clase minoritaria, demostrando que el equilibrio entre sensibilidad y precisión es más relevante que la exactitud por sí sola. Así, se reafirma que la regresión logística no solo permite clasificar, sino también comprender el peso y la relación de las variables predictoras, siempre que se acompañe de una adecuada evaluación de métricas y un tratamiento cuidadoso de los datos.

En conclusión, la regresión logística representa una herramienta poderosa para el análisis y la predicción en entornos reales, donde la interpretación de los resultados y el balance entre clases son claves para construir modelos justos, precisos y útiles para la toma de decisiones.