José Luis Spíndola Gómez A01734737

Actividad 3: Regresión Logística

En estos ejercicios utilicé la regresión logística para buscar predecir variables dicotómicas. Como variables independientes utilicé principalmente aquellas que tuvieron una alta correlación con su respectiva dependiente y/o aquellas que consideré relevantes debido a que estaban relacionadas o se complementaban bien (al menos teóricamente).

Caso 1: Host is superhost

Para predecir esta variable utilicé las variables host response rate, host acceptance rate y review scores rating, todas ellas dan información sobre la opinión de los huéspedes y deberían influir en si un anfitrión es superhost o no.

```
Matriz de Confusión:
[[4131  839]
 [1317 1634]]
```

En la matriz se puede ver que aunque hay más huéspedes predichos como no superhost, también hay varios casos en los que son superhosts, además, la precisión, exactitud y sensibilidad son buenos (un poco mejores para aquellos que no son superhosts).

Caso 2: Price

Para predecir el precio usé estimated revenue l365d, review scores value y reviews per month; el primero debido a que si el precio es alto esto podría influir las ganancias mientras que los reviews podrían verse afectados si un huésped paga más.

Esta variable fue creada considerando como económicos los precios menores de 1500 y de lujo los mayores a esta cantidad.

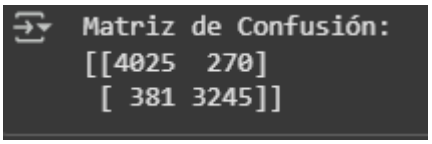
```
Matriz de Confusión:
[[1695 1296]
 [ 351 4579]]
```

En la matriz se puede ver que hay muchos más económicos que de lujo, posteriormente también pude notar que la puntuación para sensibilidad es distinta ya que para las propiedades de menos precio es de 0.92 mientras que para las más caras es de 0.57; es decir bastantes de ellas se están clasificando como económicas cuando no lo son.

Caso 3: Availability 30

Para predecir si habría poca o mucha disponibilidad usé las variables availability 60, estimated occupancy l365d y bedrooms. Esto es debido a que tienen una buena correlación además de indicar cuánto ha estado disponible en 60 días y un año respectivamente.

Consideré que poca disponibilidad son menos de 15 días mientras que mucha más de esta cantidad.

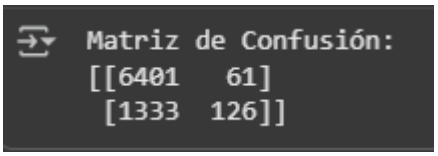


La matriz muestra un buen equilibrio y los coeficientes lo demuestran.

Caso 4: Accommodates

Para predecir cuántos huéspedes caben en cada propiedad usé las variables bathrooms, estimated revenue l365d y listing counts private rooms ya que para que quepan más personas debería haber más baños, se podría estar generando más dinero y se debería contar con más cuartos.

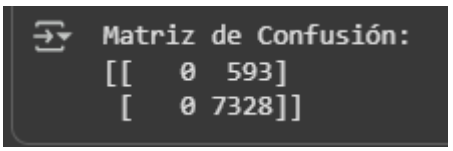
Estoy considerando que una propiedad chica tiene hasta 4 habitantes y que una grande podría llegar al límite de 16.



En la matriz se puede observar que el modelo es muy bueno prediciendo las propiedades chicas pero malo en las grandes (sobretodo porque la sensibilidad es muy muy baja) aunque la exactitud general es adecuada.

Caso 5: Host identity verified

En este caso noté que las variables beds, bedrooms y bathrooms tienen una correlación negativa importante con la identidad verificada; y aunque no tiene una lógica clara me pareció interesante, sobretodo porque pensé que es posible que aquellas propiedades con demasiados cuartos o amenidades podrían ser falsas.



Como se puede observar en la matriz, se necesitará oversampling para que la categoría de sin verificación no sea 0 (a pesar de que los anfitriones verificados no tienen errores).

Después de la misma, la matriz queda de la siguiente forma:

```
Matriz de Confusión:
[[ 210  383]
 [ 843 6485]]
```

Aunque la precisión y la sensibilidad si mejoran siguen siendo bajos. Es importante considerar que la mayoría de anfitriones si están verificados (de hecho la exactitud es buena).

Caso 6: Calculated hosts listings count

Para predecir cuántas propiedades podría tener un anfitrión me apoyé en otras variables similares con alta correlación, incluyendo el número de casas y cuartos privados.

Estoy considerando que más de 3 son muchas propiedades.

```
Matriz de Confusión:
[[ 345 3565]
 [   0 4011]]
```

La matriz tiene ambos tipos de anfitriones, sin embargo, aunque la precisión de muchas propiedades y la sensibilidad de pocas propiedades son 1 respectivamente, la exactitud del modelo sigue siendo baja. Considero que anfitriones que tienen varias propiedades están siendo mal clasificados (esto baja mucho su sensibilidad).

Caso 7: host has profile pic

Para saber si un anfitrión tiene foto de perfil usé las variables independientes host identity verified, host id y review scores accuracy ya que además de tener correlación también tienen que ver con la seguridad y la confianza.

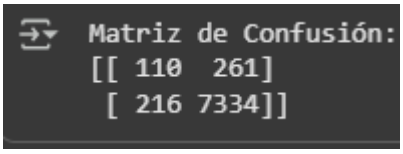
```
Matriz de Confusión:
[[ 164  272]
 [4412 3073]]
```

Nuevamente, esta variable es difícil de predecir debido a que la mayoría de anfitriones tienen foto de perfil (aunque algunos de ellos no y se puede notar en la matriz). El mejor coeficiente es la precisión de quienes si tienen foto.

Caso 8: bedrooms

Para predecir el número de cuartos estoy usando como variables independientes bathrooms, beds y has availability porque más cuartos implican más baños y camas; además de la correlación que existe.

Estoy considerando que muchos cuartos son más de 5

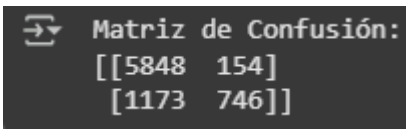


En general la exactitud del modelo es muy buena,sin embargo los problemas principales están tanto en la precisión como la sensibilidad de muchos cuartos, mientras que para pocos es casi perfecto.

Caso 9: Estimated revenue l365d

Para estimar las ganancias del último año usé las variables independientes review scores value, review scores location y has availability debido a que el valor está asociado directamente al precio (y por lo tanto a las ganancias, al igual que la ubicación; la disponibilidad influye porque mientras más se rente más ingresos debería haber.

Estoy considerando a los anfitriones como de altos ingresos si generaron más de 10,000.

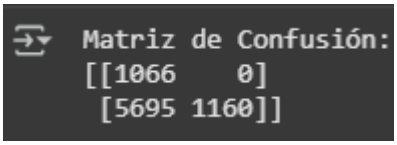


La matriz incluye valores de ambas categorías y en general pude observar buenos coeficientes a excepción de la sensibilidad de bajos ingresos (algunas se clasificaron como de altos sin que lo sean).

Caso10: Review scores rating

Para predecir el review general usé las variables independientes bathrooms, number of reviews, number of reviews ltm. Aunque bathrooms no tiene mucha lógica suma a la precisión del modelo debido a la correlación, mientras que el número de review aporta información relacionada.

Las calificaciones bajas son de 2.5 e inferiores mientras que la altas llegan hasta el 5.



Al analizar la matriz y los coeficientes pude notar que aunque la exactitud es

aceptable, hay algo curioso: la precisión de baja calificación es muy baja, mientras que su sensibilidad es de 1, para calificaciones altas sucede justo lo contrario.Considero que el modelo debería clasificar más anfitriones como buena calificación.

x	y	P(0)	P(1)	E	S(0)	S(1)
host response rate, host acceptance rate, review scores rating	host is superhost	0.77	0.64	0.72	0.80	0.61
estimated revenue l365d, review scores value, reviews per month	Price	0.78	0.83	0.79	0.93	0.57
availability 60, estimated occupancy l365d, bedrooms	Availability 30	0.92	0.91	0.89	0.94	0.91
Bathrooms, estimated revenue l365d, listing counts private rooms	Accommodates	0.83	0.67	0.82	0.99	0.086
Bathrooms bedrooms, beds	Host identity verified	0	0.93	0.93	0	1
host listings count, calculated hosts listings count entire homes, calculated hosts listings count private rooms	calculated hosts listings count	0.53	1	0.55	1	0.088
host identity verified, host id, review scores accuracy	Host has profile pic	0.04	0.92	0.41	0.38	0.41
bathrooms, beds, has availability	Bedrooms	0.97	0.34	0.94	0.97	0.27
review scores value, review scores location, has availability	Estimated revenue l365d	0.83	0.83	0.83	0.39	0.97
bathrooms,	Review scores	0.16	1	0.72	1	0.17

number of reviews, number of reviews ltm	rating					
Bathrooms bedrooms, beds	Host identity verified (oversample)	0.2	0.94	0.85	0.35	0.88

Referencias:

Get the Data. (s.f.). Inside Airbnb. https://insideairbnb.com/get-the-data/