Reconocimiento de Patrones y Aprendizaje Automatizado

UNAM

Facultad De Ciencias

Profesor Sergio Hernández López
Ayudante Miguel Andrés Guevara Castro
Ayudante Erick González Durán
Ayudante Julio César Misael Monroy González
Ayudante Rafael López Martínez

Grupo 7139

Proyecto final
Predicción del costo de un Airbnb en CDMX

Los más programadores

Altamirano Paredes Michel Alejandro	318165691
Corona Jiménez Andrea	318154952
Lira Gónzalez Rosa Linda	318074463

Introducción

Como es bien sabido, en los últimos años, los costos de alquiler y arrendamiento de viviendas en la Ciudad de México se han visto afectados de manera considerable, presentando una tendencia al alza que es mayormente visible en plataformas como Airbnb. Gracias a su gran volumen de propiedades en arrendamiento, Airbnb podría llegar a influir en los costos de viviendas circundantes a las registradas en la plataforma, lo que resultaría de gran interés tanto para arrendatarios como arrendadores.

Objetivo

Este trabajo tiene por objetivo ajustar un modelo de regresión lineal a una base de datos de Airbnb´s de la Ciudad de México para mostrarle a la población general como es que el precio de uno de estos arrendamientos podrá variar en un futuro. Sin embargo, ya que muchas personas no están familiarizadas con lenguajes de programación, se decidió crear una calculadora interactiva que le permita a un usuario estimar el precio de un Airbnb en el futuro. Así pues, procedamos con el desarrollo del trabajo.

1. Las variables

Analizando la base de datos original en Excel es posible notar que existen múltiples variables que resultan irrelevantes para nuestro modelo, pues varias de ellas no nos aportan valores medibles que podamos incluir en el mismo. Debido a esto, se descartaron un total de 44 variables, cuyos motivos fueron los siguientes:

- Letras: Debido a que muchas de estas variables presentaban únicamente palabras, links y caracteres que no eran cuantificables ni ajustables al modelo.
- Porcentajes: Al igual que el punto anterior los porcentajes al tener caracteres fueron ignorados para el modelo por contener datos que no eran medibles en R.
- Repetición de datos: Debido a que las columnas del máximo y mínimo de noches aparecían 6 veces con los mismos datos

Aunado a esto y, dado que el porcentaje con respecto al total no era muy elevado, se eliminaron los valores NA que contenía la base de datos.

2. Análisis exploratorio de las variables

Para ser más precisos con los datos que estábamos trabajando se discutió sobre que variables podían llegar a estar más correlacionadas entre sí, siendo estas las siguientes:

- 1 y 2.- Ambas variables parametrizan datos muy similares
- 3 y 4.- Ambas tienen datos relativos a la ubicación del Airbnb
- 5, 6 y 7.- Hacen alusión a la cantidad de huéspedes que puede albergar el Airbnb
- 6 y 30.- En una mala tipificación de las habitaciones los baños podrían haber sido tomados como tales, o tal vez porque las habitaciones tengan un baño integrado

- 5, 6, 7 y 30.- Igual que el motivo anterior, una mala tipificación de los datos podría hacer que estas variables estén relacionadas
- 9, 10, 11, 25, 26 y 27.- Mientras mayor o menor sea el número de noches que alguien se hospede en el Airbnb menor o mayor será la cantidad de días que estuvo disponible
- 12 - 19.- Todas estas variables hacen alusión a las reviews del Airbnb
- 3, 4 y 18.- Las variables están relacionadas con la ubicación del Airbnb
- 6, 7, 15 y 30.- La limpieza del Airbnb es notoria en las recámaras, las camas y los baños
- 1, 2, 20, 21, 22, 23.- Son estadísticas del hospedador de cada Airbnb
- 11, 25, 26, 27.- Son las disponibilidades acumuladas en ciertos periodos de tiempo, siendo el más alto de un año, por lo que si descubrimos correlación podríamos quedarnos con el de periodicidad más grande
- 12, 24, 28, 29.- Hacen alusión al número de Reviews que el Airbnb ha tenido a lo largo de diferentes periodos de tiempo.

Para comprobar si nuestras suposiciones eran correctas se realizaron gráficos de correlación entre cada una de las variables anteriormente expuestas y, tras realizar 11 gráficos pudimos ver que nuestras suposiciones eran correctas, por lo que las variables reiterativas fueron eliminadas buscando cumplir con los supuestos del modelo.

Nota: Para ver los 12 gráficos consultar el "Documento expandido"

3. Datos atípicos y el primer modelo

Teniendo ahora la base de datos definitiva, se ajustó un primer modelo para la misma, sin embargo, tras esto nos pudimos percatar de que la base de datos contenía valores atípicos, por lo que resultó indispensable quitarlos para obtener un mejor modelo y, tras eliminarlos, pudimos ver un aumento en la precisión del modelo tanto en su R^2 ajustada como en los p-values de cada una de sus variables.

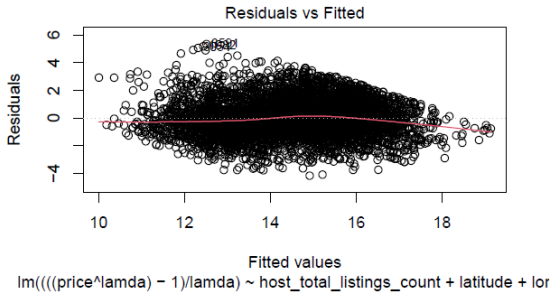
4. Escalas y los supuestos del modelo definitivo

Para tratar de subir un poco nuestra R^2 ajustada, utilizamos la función de Box-Cox, para obtener una potencia (que llamaremos lambda) que nos ayudara a reescalar nuestros datos de la variable precios para así poder trabajar con ellos de manera más fácil. Dicha lambda es la siguiente:

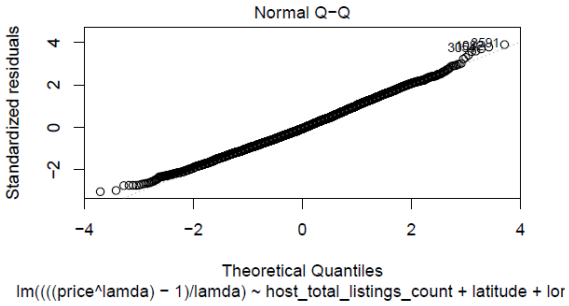
$$y = \frac{price^{\lambda} - 1}{\lambda}$$

Tras este reescalamiento, se eliminaron nuevamente los valores atípicos mediante dos métodos siendo el primero el uso de la distancia de Cook para su eliminación y el segundo acotando el costo de arrendamiento de un Airbnb, acotando a su vez a nuestro modelo para Airbnb's cuyo costo anteriormente mencionado no supere los $2385 MXN (en escala transformada).

Linealidad: Al analizar la correlación entre los precios y variables independientes, observamos que los p-value son menores a la significancia de 0.05, por lo que concluimos que hay linealidad en nuestro modelo.

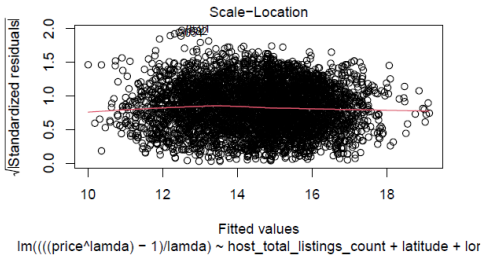


Normalidad: Para validar este supuesto, se realizó una prueba Shapiro-Wilk, notando que nuestro p-value < 0.05, por lo que se rechaza la hipótesis nula: ***"H_0 Los residuos siguen una distribución normal",*** sin embargo, tras realizar un gráfico QQ podemos decir con seguridad que nuestro modelo si cumple con el supuesto de normalidad ya que nuestros datos se ajustan perfectamente a la gráfica.



Nota: Esta discrepancia suele darse cuando los datos se enfrentan a un reescalamiento severo, lo que hace que las pruebas de bondad de ajuste den falsos negativos o positivos.

Homocedasticidad: Para validar este supuesto se realizó un gráfico de dispersión, observando que los puntos no siguen algún tipo de patrón, además de que la línea roja va de manera horizontal, por lo que nuestro modelo cumple con la Homocedasticidad.



Multicolinealidad: Este supuesto fue validado prácticamente al inicio cuando las variables más fuertemente correlacionadas fueron eliminadas, corroborado por la función vif que nos ayuda a determinar si un modelo de regresión lineal presenta colinealidad o no, y, dado que los datos obtenidos para nuestras variables son menores a 5, podemos concluir que nuestro modelo cumple también con este supuesto

host_total_listings_count	1.026718
latitude	1.022833
longitude	1.041577
accommodates	1.573110
availability_365	1.014967
review_scores_cleanliness	1.074754
review_scores_location	1.098347
calculated_host_listings_count_private_rooms	1.291497
banios	1.296549

Valores influyentes: En un caso similar al anterior, al haber eliminado los valores atípicos del modelo y al haberlo acotado solamente a ciertos Airbnb's, nos podemos percatar de que este supuesto también se cumple.

Conclusión

Tras haber realizado la validación de supuestos de nuestro modelo obtuvimos la siguiente ecuación:

$$y = \frac{price^{\lambda} - 1}{\lambda} = -225.8 + 0.09588 * (\text{host total listings count})$$

$$+ 8.513 * (\text{latitude}) - 21.02 * (\text{longitude}) + 0.4465 * (\text{accommodates})$$

$$+ 0.0009943 * (\text{availability } 365) + 1.467 * (\text{review scores cleanliness })$$

$$+ 2.709 * (\text{ review scores location })$$

Viendo que nuestra R^2 ajustada es del 0.5377, implica que nuestro modelo nos explica en un 53.77% nuestro problema de predecir el precio de un Airbnb, y además es muy significante, dado que nuestro p-value es menor que un nivel de significancia 0.05, así como también lo es en cada una de nuestras variables explicativas.

Finalmente, podemos concluir lo siguiente de nuestro modelo:

1. Decimos que con cada 0.09588 de Airbnb's que posea el anfitrión el precio aumentará.
2. Decimos por cada 8.513 que la latitud se mueva en la CDMX, el precio aumentará.
3. Dado que la longitud en la CDMX es negativa, decimos que precio aumenta con 21.02 por cada vez la longitud se mueva.
4. Decimos que aproximadamente por cada media persona que se hospede en el AIRBNB, el precio igual aumentará, por lo que una persona completa casi duplica el precio.
5. Vemos que mientras casi no haya disponibilidad el precio aumentará en 0.0009943
6. Vemos que por cada punto que se le da a la limpieza de un Airbnb, nuestro precio aumentará con una razón de 1.467
7. Análogamente al punto anterior, mientras mayor sea la calificación que tenga por ubicación el Airbnb, mayor será el precio con una razón de 2.709

Referencias

- Get the Data. (s. f.). https://insideairbnb.com/get-the-data/
- Wei, T., & Simko, V. (2021, 18 noviembre). An Introduction to corrplot Package, de https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
- Hernández, F., Usuga, O., & Mazo, M. (2024, 13 marzo). 17 Multicolinealidad | Modelos de Regresión con R. https://fhernanb.github.io/libro_regresion/multicoli.html
- Ripley, B. (2024, 26 abril). Package 'MASS'. https://cran.r-project.org/web/packages/MASS/MASS.pdf
- Sancho, R. S. (s. f.). El paquete dplyr | Programación en R. https://rsanchezs.gitbooks.io/rprogramming/content/chapter9/dplyr.html
- Read Excel files. (s. f.). https://readxl.tidyverse.org/
- Gross, J., & Ligges, U. (2015, 30 julio). Package 'nortest'. https://cran.r-project.org/web/packages/nortest/nortest.pdf
- Create Elegant Data Visualisations Using the Grammar of Graphics. (s. f.). https://ggplot2.tidyverse.org/
- car package - RDocumentation. (s. f.). https://www.rdocumentation.org/packages/car/versions/3.1-2
- Tools for Descriptive Statistics. (s. f.). https://andrisignorell.github.io/DescTools/