Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Negocios

Campus Puebla

Gestión de proyectos de plataformas tecnológicas

Regresión logística binaria

Isaac Miguel Barrón Portillo

A01737199

Docente: PhD Alfredo García Suárez

Fecha: 13 de octubre de 2025

Airbnb Madrid – Regresión Logística

1. Introducción

El presente análisis tiene como propósito identificar los factores que influyen en los principales indicadores de desempeño de los anfitriones en Airbnb Madrid, a través de la aplicación de modelos de regresión logística binaria. Esta técnica estadística permite estimar la probabilidad de que una observación pertenezca a una de dos categorías mutuamente excluyentes, basándose en un conjunto de variables explicativas. En el contexto de Airbnb, el objetivo es comprender qué características del alojamiento, del anfitrión y de su gestión operativa incrementan la probabilidad de alcanzar comportamientos considerados "óptimos", como tiempos de respuesta rápidos, altas tasas de aceptación, pertenecer al grupo de superanfitriones o ubicarse en zonas de alta demanda.

Para ello, se trabajó con una base de datos de Airbnb correspondiente a la ciudad de Madrid, la cual contiene información sobre anfitriones, propiedades, precios, disponibilidad y valoraciones de huéspedes. Las variables seleccionadas fueron transformadas en variables binarias para permitir su interpretación directa en términos de probabilidad y facilitar la comparación entre categorías de "alto desempeño" y "desempeño regular o bajo".

Este enfoque permite no solo cuantificar el peso de cada variable en la probabilidad de éxito, sino también ofrecer insights accionables para la toma de decisiones, tanto en la gestión de anfitriones individuales como en estrategias globales de optimización del portafolio de propiedades. En resumen, la regresión logística binaria se convierte en una herramienta analítica que traduce los datos operativos y de servicio en información estratégica para mejorar la calidad y competitividad de la oferta de Airbnb en Madrid.

2. Documentación del proceso

El proceso analítico siguió una metodología estructurada que abarcó las etapas de preparación de datos, creación de variables binarias, modelado estadístico y evaluación de desempeño.

1. Preparación de los datos.

   Se partió de la base de datos original de Airbnb Madrid, aplicando procedimientos de limpieza, selección de variables relevantes y manejo de valores faltantes. Posteriormente, se normalizaron las variables numéricas mediante escalamiento estándar (StandardScaler) con el fin de garantizar la comparabilidad entre predictores.

2. Conversión de variables a formato binario.

   Cada variable clave fue recodificada en dos categorías (0 y 1), según los criterios establecidos en la tabla de conversión. De esta forma, la etiqueta 1 representó los

valores deseables o de alto desempeño, y la etiqueta 0 los valores de desempeño medio o bajo.

Los criterios aplicados fueron:

- Tiempo de respuesta del anfitrión: Rápida (≤ unas horas) = 1; Lenta = 0.
- Tasa de respuesta del anfitrión: Alta (≥70%) = 1; Media-baja (<70%) = 0.
- Tasa de aceptación: Alta (≥70%) = 1; Media-baja (<70%) = 0.
- Superanfitrión: Sí = 1; No = 0.
- Distrito: Centro = 1; Limítrofe = 0.
- Capacidad: 4 o más huéspedes = 1; 3 o menos = 0.
- Calificación general (rating): Buena (≥4) = 1; Regular-mala (<4) = 0.
- Precisión (accuracy): Buena (≥4) = 1; Regular-mala (<4) = 0.
- Limpieza (cleanliness): Buena (≥4) = 1; Regular-mala (<4) = 0.
- Ubicación (location): Buena (≥4) = 1; Regular-mala (<4) = 0.

3. Modelado y entrenamiento.

   Se implementaron diez regresiones logísticas binarias independientes, cada una enfocada en predecir una variable objetivo distinta. Las variables predictoras se eligieron en función de la naturaleza del fenómeno a explicar. El conjunto de datos se dividió en entrenamiento (70%) y prueba (30%) mediante train_test_split, y las variables predictoras fueron estandarizadas antes del modelado. El algoritmo utilizado fue LogisticRegression() de scikit-learn, con el método de optimización por defecto y penalización L2. En algunos modelos se probó la variante class_weight='balanced' para mitigar desbalances entre clases.

4. Evaluación y validación.

   Para cada modelo se calcularon las métricas accuracy, precision y recall, junto con la matriz de confusión que permitió visualizar los aciertos y errores en la clasificación. El umbral de decisión se mantuvo en 0.5, y los resultados se interpretaron bajo el enfoque de equilibrio entre sensibilidad y especificidad.

5. Documentación y replicabilidad.

   Cada modelo se documentó en el notebook Actividad3.ipynb, especificando su variable dependiente, predictores utilizados y métricas obtenidas. El código se mantuvo modular y reproducible, lo que permite extender el análisis a otras ciudades o periodos temporales con mínimos ajustes.

3. Resultados y análisis de matrices

Tiempo de respuesta del anfitrión (response_time_bin)

Matriz de confusión:
351 & 990 \

210 & 4675

La distribución muestra una alta proporción de observaciones clasificadas en la categoría positiva. El modelo identifica con precisión a los anfitriones que responden rápidamente (4 675), aunque también clasifica erróneamente a varios de respuesta lenta como rápidos (990). El número de falsos negativos (210) indica un nivel limitado de detección de los casos que realmente presentan una respuesta rápida pero fueron ubicados en la categoría opuesta. En conjunto, el modelo tiende a favorecer la predicción de la clase positiva, reduciendo la especificidad.

Tasa de respuesta del anfitrión (response_rate_bin)

Matriz de confusión:
139 & 186 \
74 & 5827

El resultado refleja un dominio de la clase positiva, con un volumen elevado de verdaderos positivos (5 827) frente a un número menor de falsos negativos (74). Los falsos positivos (186) sugieren una ligera sobreasignación de anfitriones con tasas de respuesta altas. La estructura de la matriz evidencia que el modelo reconoce de manera amplia la categoría positiva, manteniendo un margen de error controlado en la clasificación opuesta.

Tasa de aceptación del anfitrión (acceptance_rate_bin)

Matriz de confusión:
214 & 592 \
45 & 5375

El modelo concentra la mayoría de sus aciertos en la clase positiva (5 375). Los falsos positivos (592) superan a los falsos negativos (45), lo que indica que algunas observaciones con tasas de aceptación bajas son asignadas como altas. El patrón revela una tendencia a clasificar con mayor frecuencia dentro de la categoría positiva, lo que incrementa la cobertura a costa de exactitud en los casos negativos.

Superanfitrión (superhost_bin)

Matriz de confusión:
4361 & 177 \
1533 & 155

El modelo distingue adecuadamente la clase negativa (4 361), pero presenta un número significativo de falsos negativos (1 533), lo que implica que una proporción importante de superanfitriones reales se clasifica como no superanfitriones. La relación entre falsos negativos y verdaderos positivos sugiere una sensibilidad limitada hacia esta categoría y una mayor restricción al asignarla.

Distrito (neighbourhood_bin)

Matriz de confusión:
3054 & 549 \
1845 & 778

La dispersión de errores muestra un comportamiento heterogéneo. El número de falsos negativos (1 845) y falsos positivos (549) evidencia que las variables empleadas no diferencian con claridad entre los dos grupos de distritos. El modelo presenta un grado de confusión considerable entre ambas categorías, reduciendo la capacidad discriminante.

Capacidad (accommodates_bin)

Matriz de confusión:
3423 & 326 \
928 & 1549

El modelo mantiene una relación estable entre verdaderos positivos (1 549) y verdaderos negativos (3 423). La presencia de falsos negativos (928) indica que ciertas propiedades con mayor capacidad fueron clasificadas en el grupo de menor capacidad, mientras que los falsos positivos (326) representan el caso opuesto. Los resultados muestran una estructura simétrica en la clasificación.

Calificación general (review_scores_rating_bin)

Variante 1

Matriz:
109 & 70 \
20 & 6027

La mayoría de los registros se concentran en la clase positiva (6 027). Los falsos positivos (70) indican una pequeña fracción de alojamientos clasificados con alta calificación sin corresponder a la realidad, mientras que los falsos negativos (20) muestran un nivel de omisión bajo.

Variante 2

Matriz:
174 & 5 \
343 & 5704

En esta configuración, el número de falsos positivos se reduce considerablemente, pero aumentan los falsos negativos (343). Esto refleja un cambio de criterio del modelo hacia una asignación más restrictiva de la clase positiva.

Precisión (accuracy_bin)

Variante 1

Matriz:
81 & 71 \
22 & 6052

El modelo identifica ampliamente los casos positivos, con un volumen dominante de verdaderos positivos (6 052). Los falsos positivos (71) y falsos negativos (22) se mantienen en proporciones reducidas, mostrando un comportamiento equilibrado entre ambas clases.

Variante 2

Matriz:
144 & 8 \
313 & 5761

La disminución de falsos positivos (8) se acompaña de un incremento de falsos negativos (313). La estructura sugiere un ajuste más estricto del umbral de decisión, priorizando la exactitud de la clase negativa frente a la positiva.

Limpieza (cleanliness_bin)

Variante 1

Matriz:
60 & 110 \
34 & 6022

El modelo asigna la mayoría de los registros a la clase positiva (6 022), aunque los falsos positivos (110) evidencian una tendencia a incluir dentro de dicha categoría observaciones que no pertenecen a ella.

Variante 2

Matriz:
153 & 17 \
479 & 5577

En esta versión, los falsos positivos se reducen de manera importante, pero los falsos negativos (479) se incrementan, desplazando el equilibrio hacia una clasificación más restrictiva de “buena limpieza”.

Ubicación (location_bin)

Variante 1

Matriz:
18 & 100 \
20 & 6088

Predomina la clase positiva con 6 088 verdaderos positivos. La proporción de falsos positivos (100) indica cierta sobreasignación a la categoría “buena ubicación”.

Variante 2

Matriz:
89 & 29 \
568 & 5540

La reducción de falsos positivos (29) se compensa con un aumento significativo de falsos negativos (568), lo que refleja un criterio más restrictivo en la identificación de ubicaciones favorables.

4. Análisis de las puntuaciones

1. Tiempo de respuesta del anfitrión

- Precisión: 0.825
- Exactitud: 0.807
- Sensibilidad: 0.262

El modelo presenta un desempeño estable en precisión y exactitud, pero una sensibilidad limitada. Esto significa que clasifica correctamente la mayoría de los casos negativos y positivos confirmados, aunque no logra identificar de forma completa a todos los anfitriones con tiempos de respuesta rápidos. Su comportamiento está alineado con el sesgo hacia la clase negativa observado en la matriz de confusión.

2. Tasa de respuesta del anfitrión

- Precisión: 0.969
- Exactitud: 0.958
- Sensibilidad: 0.428

Los valores indican un nivel de precisión elevado y una buena exactitud global. La sensibilidad, aunque moderada, mejora respecto al modelo anterior. Esto refleja una capacidad de clasificación positiva más equilibrada, aunque todavía predomina la predicción conservadora.

3. Tasa de aceptación del anfitrión

- Precisión: 0.901
- Exactitud: 0.898

- Sensibilidad: 0.266

El modelo conserva altos niveles de precisión y exactitud, pero mantiene una sensibilidad baja, lo que confirma una tendencia a clasificar con mayor confianza los casos negativos. Este patrón coincide con la proporción elevada de falsos negativos observada anteriormente.

4. Superanfitrión

- Precisión: 0.467
- Exactitud: 0.725
- Sensibilidad: 0.961

En este modelo se invierte el comportamiento: la sensibilidad es muy alta, lo que indica una cobertura amplia de los casos positivos (superanfitriones reales). Sin embargo, la precisión es baja, reflejando la presencia de numerosos falsos positivos. El modelo prioriza la detección total de superanfitriones a costa de la exactitud de la clase opuesta.

5. Distrito

- Precisión: 0.586
- Exactitud: 0.615
- Sensibilidad: 0.848

El modelo logra captar la mayoría de los casos positivos (alta sensibilidad) pero con un sacrificio en precisión. Esto implica que predice con frecuencia la categoría de “Centro”, aunque parte de esas asignaciones no corresponden a la realidad. La exactitud moderada confirma una clasificación ambigua entre ambas zonas geográficas.

6. Capacidad

- Precisión: 0.826
- Exactitud: 0.799
- Sensibilidad: 0.913

El equilibrio entre las tres métricas es consistente. El modelo detecta eficazmente los alojamientos con capacidad mayor o igual a cuatro personas, manteniendo una proporción controlada de errores. La combinación de sensibilidad y precisión elevadas lo posiciona como uno de los modelos más estables dentro del conjunto.

7. Calificación general (rating)

- Precisión: 0.989
- Exactitud: 0.986
- Sensibilidad: 0.609

El modelo obtiene valores altos en las tres métricas, destacando en precisión y exactitud. La sensibilidad intermedia indica que reconoce con fiabilidad los anuncios con calificaciones altas, aunque una parte de ellos no es identificada. Refleja un balance adecuado entre control de errores y cobertura.

8. Calificación general ponderada (rating_pond)

- Precisión: 0.337
- Exactitud: 0.944
- Sensibilidad: 0.943

El ajuste ponderado incrementa la sensibilidad de forma significativa, priorizando la detección de casos positivos. Sin embargo, la precisión disminuye de manera marcada, lo que evidencia un aumento de falsos positivos. Este cambio de equilibrio es característico del uso de ponderaciones de clase.

9. Precisión de descripción (accuracy)

- Precisión: 0.988
- Exactitud: 0.985
- Sensibilidad: 0.533

El modelo muestra un comportamiento muy similar al de la calificación general: alta precisión y exactitud, con una sensibilidad moderada. Se observa una buena correspondencia entre la información ofrecida y la experiencia percibida, aunque el modelo no detecta todos los casos positivos.

10. Precisión de descripción ponderada (accuracy_pond)

- Precisión: 0.315
- Exactitud: 0.948
- Sensibilidad: 0.948

Con la aplicación de ponderación, la sensibilidad aumenta sustancialmente, mejorando la detección de observaciones positivas. En contrapartida, la precisión desciende, lo que confirma un incremento de falsos positivos. El modelo prioriza la cobertura sobre la selectividad.

11. Limpieza (cleanliness)

- Precisión: 0.982
- Exactitud: 0.977
- Sensibilidad: 0.353

El modelo presenta altos niveles de precisión y exactitud, pero una sensibilidad limitada. Esto implica que clasifica correctamente la mayoría de los anuncios con valoraciones bajas o medias, pero omite una proporción relevante de aquellos con limpieza sobresaliente.

12. Limpieza ponderada (cleanliness_pond)

- Precisión: 0.242
- Exactitud: 0.920
- Sensibilidad: 0.921

El modelo ponderado incrementa la sensibilidad, lo que le permite detectar la mayoría de los casos positivos. No obstante, la precisión disminuye considerablemente, reproduciendo el mismo patrón de los modelos ponderados anteriores.

13. Ubicación (location)

- Precisión: 0.984
- Exactitud: 0.981
- Sensibilidad: 0.153

Los valores de precisión y exactitud son elevados, mientras que la sensibilidad es muy baja. El modelo clasifica con solidez los casos negativos, pero no logra identificar de manera completa las propiedades ubicadas en zonas con alta valoración de localización.

14. Ubicación ponderada (location_pond)

- Precisión: 0.135
- Exactitud: 0.904
- Sensibilidad: 0.907

El modelo ponderado invierte la relación: eleva la sensibilidad a más del 90 %, lo que significa que prácticamente todos los casos positivos son detectados. Sin embargo, la baja precisión confirma un aumento de observaciones clasificadas como “buena ubicación” sin corresponderlo.

Conclusión ejecutiva

1. Efecto de la ponderación:

   Los modelos ponderados presentan una mejora sustancial en sensibilidad y una reducción significativa en precisión. Este comportamiento responde al ajuste de pesos para equilibrar clases, lo cual amplía la cobertura de detección de casos positivos, pero incrementa los falsos positivos.

2. Modelos estructurales:

Las variables con relaciones directas y cuantificables (como capacidad, rating y accuracy) mantienen un desempeño más estable y equilibrado en las tres métricas, con niveles consistentes de precisión y exactitud.

3. Modelos perceptivos:

   Variables dependientes de juicios subjetivos (superhost, distrito, limpieza y ubicación) muestran mayor dispersión entre precisión y sensibilidad. Su interpretación requiere considerar factores no incluidos en los predictores actuales.

4. Comportamiento global:

   La exactitud promedio se mantiene elevada en todos los modelos (≈0.90 o superior), lo que confirma una capacidad de clasificación general adecuada. Las diferencias entre precisión y sensibilidad evidencian distintos enfoques de modelado según la variable objetivo: unos más restrictivos y otros orientados a cobertura.