Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Puebla

Actividad 4.1 Regresión Logística

José Manuel Morales Escalante

Materia:
Analítica de datos y herramientas de inteligencia artificial I

Fecha:
27 de abril de 2025

Resumen General

El objetivo principal es explorar la correlación entre diferentes variables pertenecientes a la base de datos de Río que contiene información sobre Airbnb en la ciudad y predecir categorías binarias o dicotómicas basadas en características como precios, número de camas, puntuaciones de limpieza, etc.

Estructura del Documento

1. **Carga y Preprocesamiento de Datos**:
    - Se carga el archivo Rio_limpio.csv y se muestran las primeras filas del DataFrame. Este archivo ya contiene la base de datos limpia de valores nulos y atípicos.
    - Se filtran filas con valores no válidos en columnas específicas (host_is_superhost, host_identity_verified, has_availability, instant_bookable).

2. **Análisis de Regresión Logística**:
    - Se aplica regresión logística a 10 variables objetivo diferentes, cada una convertida en categorías binarias o dicotómicas cuando es necesario.
    - Para cada variable objetivo, se divide el conjunto de datos en entrenamiento y prueba, se escalan los datos y se entrena un modelo de regresión logística.
    - Se evalúa el modelo utilizando métricas como matriz de confusión, precisión, exactitud y sensibilidad.

Variables Analizadas

1. host_is_superhost:
    - Precisión: t (65.9%), f (71.6%).
    - Exactitud: 71.06%.
    - Sensibilidad: t (19.75%), f (95.19%).
2. host_identity_verified:
    - Precisión: t (83.89%), f (0%).
    - Exactitud: 83.89%.
    - Sensibilidad: t (100%), f (0%).
3. has_availability:
    - Precisión: t (100%), f (0%).
    - Exactitud: 100%.
    - Sensibilidad: t (100%), f (0%).
4. instant_bookable:
    - Precisión: t (66.67%), f (77.53%).
    - Exactitud: 77.53%.
    - Sensibilidad: t (0.08%), f (99.99%).
5. host_response_time:
    - Convertida a categorías: "within an hour" y "More than an hour".
    - Precisión: "within an hour" (60.09%), "More than an hour" (56.19%).
    - Exactitud: 58.91%.
    - Sensibilidad: "within an hour" (75.85%), "More than an hour" (38.08%).
6. room_type:
    - Convertida a categorías: "Entire home/apt" y "Otro".
    - Precisión: "Entire home/apt" (82.36%), "Otro" (67.92%).
    - Exactitud: 81.58%.
    - Sensibilidad: "Entire home/apt" (97.82%), "Otro" (18.04%).
7. availability_365:
    - Convertida a categorías: "Disponibilidad menor a medio año" y "Disponibilidad mayor a medio año".
    - Precisión: "Disponibilidad menor a medio año" (57.06%), "Disponibilidad mayor a medio año" (55.13%).
    - Exactitud: 56.42%.
    - Sensibilidad: "Disponibilidad menor a medio año" (72.06%), "Disponibilidad mayor a medio año" (38.77%).
8. availability_30:
    - Convertida a categorías: "Poca disponibilidad en el siguiente mes" y "Mucha disponibilidad en el siguiente mes".
    - Precisión: "Mucha disponibilidad en el siguiente mes" (57.68%), "Poca disponibilidad en el siguiente mes" (58.54%).
    - Exactitud: 58.13%.

    - Sensibilidad: "Mucha disponibilidad en el siguiente mes" (55.68%), "Poca disponibilidad en el siguiente mes" (60.50%).
9. review_scores_rating:
    - Convertida a categorías: "Rating de calificación medio" y "Rating de calificación alto".
    - Precisión: "Rating de calificación alto" (87.92%), "Rating de calificación medio" (65.40%).
    - Exactitud: 86.62%.
    - Sensibilidad: "Rating de calificación alto" (97.64%), "Rating de calificación medio" (24.93%).
10. number_of_reviews:
    - Convertida a categorías: "Pocas calificaciones" y "Muchas calificaciones".
    - Precisión: "Muchas calificaciones" (0%), "Pocas calificaciones" (89.63%).
    - Exactitud: 89.63%.
    - Sensibilidad: "Muchas calificaciones" (0%), "Pocas calificaciones" (100%).

1. Problemas de Desbalanceo en los Datos

- Variables Afectadas:
    - **host_identity_verified**: El modelo predice siempre la clase mayoritaria (t), ignorando por completo la clase f. Esto se evidencia en una sensibilidad del 0% para f y una precisión del 0% en esa clase.
    - **number_of_reviews**: Ocurre un fenómeno similar, donde el modelo clasifica todas las instancias como "Pocas calificaciones" (clase mayoritaria), resultando en métricas nulas para "Muchas calificaciones".

- Causas:
    - Distribución desigual de las clases (ejemplo: en host_identity_verified, el 83.89% de los datos son t).
    - El modelo no está aprendiendo patrones para la clase minoritaria debido a su escasa representación.

- Consecuencias:
    - Métricas engañosas (ejemplo: exactitud alta pero sensibilidad nula para la clase minoritaria).
    - El modelo no es útil para predecir la clase minoritaria, lo que limita su aplicabilidad en escenarios reales donde ambas clases son relevantes.

2. Variables con Buen Desempeño

    - has_availability:
        - Precisión y Exactitud del 100%: Todos los registros tienen el valor t, lo que hace que el modelo siempre acierte. Sin embargo, esto también indica que la variable no tiene variabilidad útil para el análisis.
    - review_scores_rating:
        - Precisión del 87.92% para "Rating alto": El modelo identifica correctamente propiedades con alta calificación en la mayoría de los casos.
        - Sensibilidad del 97.64% para "Rating alto": Captura casi todos los casos reales de esta categoría.

- Razones del Éxito:
    - Distribución equilibrada o patrones claramente diferenciables en los datos.
    - Variables independientes (como review_scores_cleanliness) están fuertemente correlacionadas con la variable objetivo.

3. Variables con Desempeño Moderado

- Ejemplos:
    - host_is_superhost:
        - Precisión del 71.6% para f: El modelo es aceptable pero no óptimo.
        - Sensibilidad baja para t (19.75%): Identifica mal a los superhosts reales.
    - instant_bookable:
        - Alta sensibilidad para f (99.99%): Detecta casi todos los casos donde no hay reserva instantánea.

        - Precisión baja para t (66.67%): Falsos positivos en reservas instantáneas.
- Posibles Mejoras:
    - Incluir más variables predictoras (ejemplo: políticas de cancelación).
    - Ajustar el umbral de clasificación para equilibrar precisión y sensibilidad.

4. Variables con Bajo Desempeño

- Ejemplos Críticos:
    - host_response_time:
        - Exactitud del 58.91%: Cercana al azar (50% para dos clases).
        - Sensibilidad desbalanceada (75.85% vs. 38.08%): El modelo favorece la clase "within an hour".
    - availability_365 y availability_30:
        - Exactitud ~56–58%: Similar a adivinar al azar.
        - Problema: Las categorías creadas (ejemplo: "menor/mayor a medio año") podrían no capturar patrones reales en los datos.
- Causas Raíz:
    - Falta de correlación clara entre las variables independientes y la objetivo.
    - Discretización inadecuada de variables numéricas (ejemplo: availability_365 dividida en solo dos categorías amplias).

5. Problemas Técnicos Identificados

- Advertencias en el Código:
    - Mensajes como UndefinedMetricWarning (ejemplo: en host_identity_verified) indican divisiones por cero debido a la ausencia de predicciones para una clase.
- Errores Potenciales:
    - En has_availability, la matriz de confusión muestra solo una clase ([[11352]]), lo que sugiere que todos los datos son t y el modelo no está aprendiendo nada útil.

6. Limitaciones en el Preprocesamiento

- Conversión de Variables:
    - En room_type, agrupar "Hotel room", "Private room" y "Shared room" en "Otro" podría ocultar diferencias importantes entre estos tipos.
- Discretización Arbitraria:
    - En availability_365, dividir en solo dos intervalos (0–182.5 y 182.5–365 días) ignora patrones estacionales o mensuales.