APLICACIÓN DE MODELOS DE REGRESIÓN AL ANÁLISIS DE RESULTADOS DE LA PLATAFORMA AIRBNB

Universidad Tecnológica Nacional
Ciencia de Datos

Giorlando, Fabiana **Ibire, Zoe**

ABSTRACT: *Este informe analiza el DataSet de Airbnb con el objetivo de identificar factores clave que influyen en la fijación de precios y, en consecuencia, utilizarlos para predecirlos.*

INTRODUCCIÓN Y OBJETIVOS

Este informe tiene como objetivo realizar un análisis detallado del DataSet de la plataforma Airbnb, centrándose en la relación entre diversas variables y el precio de los alojamientos. La exploración y evaluación de estas variables buscan identificar factores clave que influyen en la fijación de precios, con el propósito de construir modelos predictivos precisos. El objetivo final es ofrecer recomendaciones basadas en hallazgos y análisis para mejorar la toma de decisiones estratégicas de Airbnb y optimizar la experiencia tanto para anfitriones como para viajeros en la plataforma.

DATA SET Y PREPROCESAMIENTO EN AIRBNB

El conjunto de datos incluye 19.309 publicaciones con 29 variables que describen diversas características de las propiedades en Airbnb. Después de la importación del dataset, se procedió a una eliminación de columnas con valores nulos significativos y datos irrelevantes, como ratings, thumbnail, tipo de cama, entre otros.

Al profundizar en las estadísticas descriptivas de los alquileres de Airbnb, es vital considerar una multitud de factores que afectan los precios y las opciones de alquiler. Investigaciones anteriores han enfatizado la importancia de las características físicas, la ubicación, el vecindario, la cantidad de dormitorios y baños y los aspectos relacionados con los anfitriones a la hora de determinar las tarifas de alquiler de Airbnb.

	name	dtype	<lambda>
room_type	room_type	object	0
accommodates	accommodates	int64	0
bathrooms	bathrooms	float64	35
cancellation_policy	cancellation_policy	object	0
cleaning_fee	cleaning_fee	bool	0
city	city	object	0
instant_bookable	instant_bookable	object	0
neighbourhood	neighbourhood	object	1458
number_of_reviews	number_of_reviews	int64	0
zipcode	zipcode	object	225
bedrooms	bedrooms	float64	17
price	price	float64	0

Ilustración 1. Descripción de las principales variables y la cantidad de nulls.

El análisis descriptivo se centró en 12 variables, se otorgó atención particular a la variable "neighbourhood", ya que presentaba la mayor cantidad de valores nulos. Para abordar esto, se procedió a completar los máximos posibles utilizando los códigos postales asociados a dicha variable como referencia. Tras esta operación, se eliminaron alrededor de 300 filas nulas, consideradas insignificantes en comparación con las 19.000 restantes.

ANÁLISIS EXPLORATORIO DE DATOS – EDA

Inicialmente, procedimos a analizar la variable primordial y de máxima relevancia, el precio, tanto en su composición global como desglosado por ciudad. Observamos la distribución y, mediante el uso de un diagrama de caja (boxplot), identificamos

como outliers a los valores correspondientes a los percentiles 5 y 95, procediendo a su eliminación.

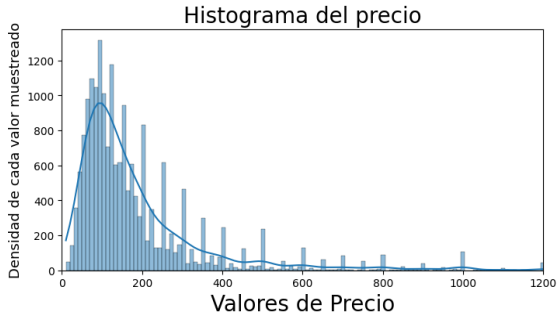

Ilustración 2. Histograma de la variable precio

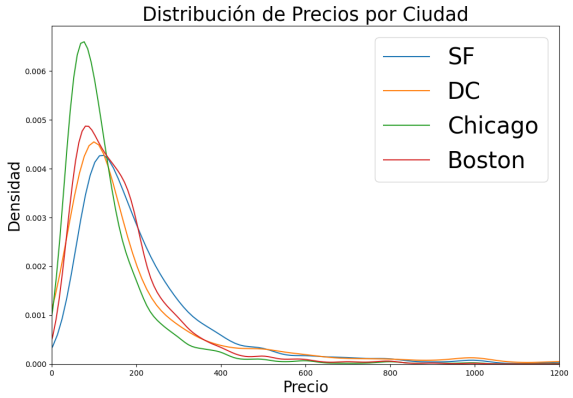

Ilustración 3. Histograma de la variable precio por ciudad.

La ubicación se establece como un factor crítico en la determinación de los precios de alquiler en Airbnb. La existencia de puntos de interés en las proximidades muestra una correlación positiva con los precios de cotización. En consecuencia, se llevó a cabo un análisis de las ciudades y su vinculación con los precios mediante el empleo de un Boxplot.

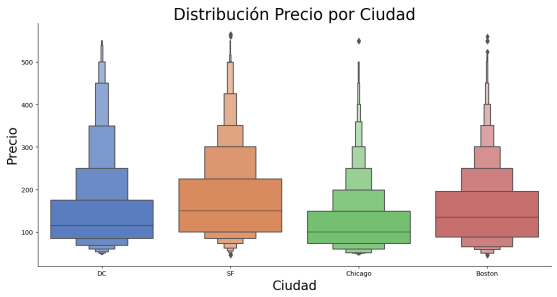

Ilustración 4. BoxPlot de precios por ciudad.

Se puede notar que las ciudades de DC y San Francisco presentan una mayor dispersión. En este sentido, se realizaron boxplots adicionales para examinar los cinco barrios con las medianas más altas y más bajas en cada una de estas ciudades.

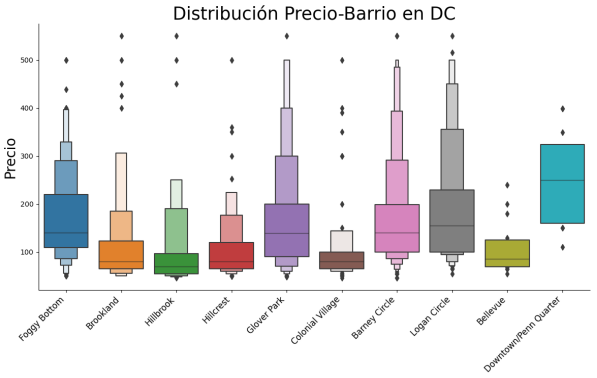

Ilustración 5. BoxPlot Top 5 y Bottom 5 barrios de DC según precio

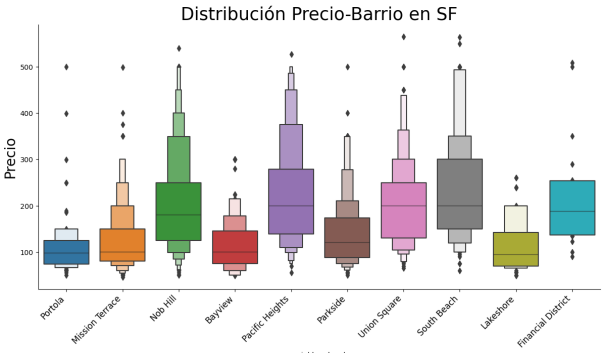

Ilustración 6. BoxPlot Top 5 y Bottom 5 barrios de SF según precio.

Nuestro análisis desveló notables variaciones en los precios de alquiler, tanto a nivel ciudad como barrio. Se destaca, por ejemplo, que el rango de precios promedio para alquileres en Boston y San Francisco resultó significativamente superior al observado en Chicago y Washington, D.C. Este hallazgo subraya la marcada influencia de factores específicos de ubicación y la demanda del mercado en la determinación de los precios de alquiler.

Adicionalmente, identificamos patrones en los precios de alquiler en distintos barrios. En particular, Glover Park exhibe precios de alquiler notoriamente elevados en comparación con otros barrios, sugiriendo la existencia de atributos o características asociadas a este lugar que contribuyen a

estrategias de precios premium. Contrariamente, el área de Downtown presenta una dispersión mínima, pero su mediana es considerablemente superior a la de otros barrios, añadiendo complejidad al panorama y resaltando la importancia de considerar múltiples dimensiones al analizar los determinantes de los precios de alquiler.

La evaluación de las variables "Tarifa por Limpieza" y "Reservación al instante" reveló su limitada capacidad descriptiva en relación con la variable de interés, el precio. Por consiguiente, se decidió excluir dichas variables en futuros análisis de Machine Learning.

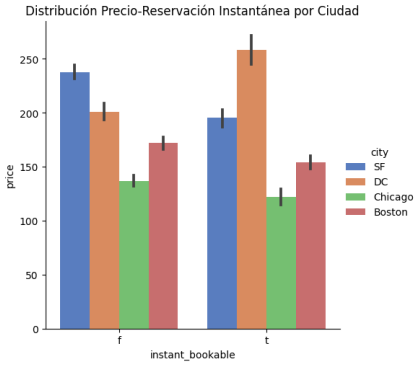

Ilustración 7. Gráfico de barras de precio en función de reservación instantánea y ciudad.

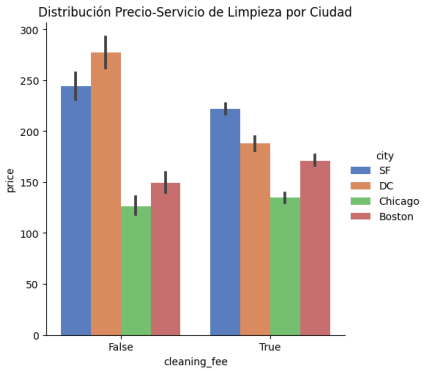

Ilustración 8. Gráfico de barras de precio en función de servicio de limpieza y ciudad

En contraste, la variable "Política de Cancelación", sometida a un análisis mediante un boxplot, demostró ser descriptiva y proporcionar información útil para el análisis en cuestión.

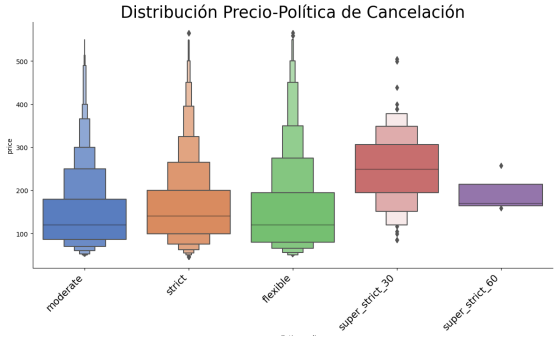

Ilustración 9. BoxPlot de precios por política de cancelación.

Como etapa final del análisis, se procedió a la construcción de un heatmap que representa la correlación entre las variables numéricas. Este enfoque permitió observar la relación existente entre diversas variables, destacándose la interconexión entre "price", "accommodates", "bedrooms" y "bedrooms".

Ilustración 10. Heatmap de correlación entre variables.

MACHINE LEARNING

Tras completar el análisis exhaustivo del dataset, iniciamos la construcción del pipeline destinado a la predicción de los precios de alojamientos en Airbnb. Iniciamos el proceso generando variables ficticias ("dummies") para las variables categóricas que consideramos más relevantes; en este caso, optamos por utilizar las ciudades, dado que la inclusión de barrios conlleva una considerable carga dimensional que podría sobrecargar el modelo. Además, incorporamos la variable correspondiente a la política de reservación.

Posteriormente, sometimos el conjunto de datos a un test con una proporción de 0.1 para evaluar el rendimiento de cuatro modelos distintos. El primer modelo consistió en una Regresión Lineal básica, seguido por un Support Vector Regression, para el cual realizamos una búsqueda exhaustiva de hiperparámetros mediante un Grid Search, obteniendo como óptimos {'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}. Finalmente, exploramos la reducción de dimensionalidad mediante dos enfoques: PCA (Análisis de Componentes Principales) y Kernel PCA (con kernel RBF y gamma igual a 0.02) cada una reduciendo a 8 dimensiones.

Con el propósito de comparar los resultados obtenidos por cada modelo, calculamos el error cuadrático medio y presentamos estos datos en un gráfico de barras, proporcionando una visualización clara de las prestaciones relativas de cada enfoque. Este análisis metodológico busca destacar las fortalezas y debilidades de los modelos implementados, proporcionando una base sólida para la selección del modelo más adecuado en la predicción de precios de alojamientos en Airbnb.

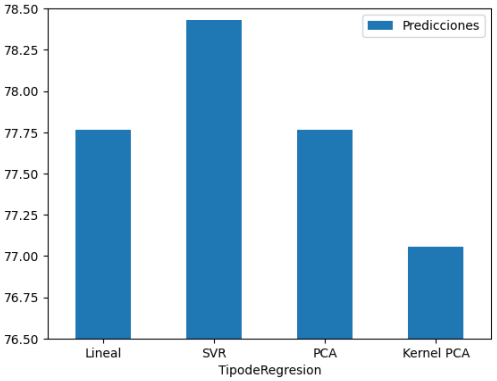

Ilustración 11. Gráfico de barras comparación de los MSE de cada modelo de regresión.

CONCLUSIONES

En conclusión, tras analizar a fondo los datos de Airbnb, se ha obtenido información valiosa sobre el mercado de alquiler a corto plazo. Se identificaron disparidades significativas en los precios entre ciudades y barrios, subrayando la necesidad de estrategias de fijación de precios adaptadas a ubicaciones geográficas específicas. Adicionalmente, se pudo verificar la relación entre el precio y las 4 variables más relevantes: comodidades, cantidades de cuartos y baños y ubicación.

Dentro de los modelos de predicción, el modelo Kernel PCA, de reducción de dimensionalidad resultó el más exitoso, congruentemente dado que los anteriores podrían sufrir las consecuencias de tanto 'overfitting' o 'underfitting'. Si bien el menor error arrojado es aproximadamente la mitad del promedio de los precios y no resulta muy preciso, el uso de herramientas de Machine Learning puede resultar una gran ventaja tanto para los anfitriones al poder estimar un precio que resulte óptimo y competitivo con el resto y para los huéspedes que deben de tomar la mejor decisión dentro de un presupuesto limitado. Un uso más atomizado de la plataforma, es decir, aplicado a una ciudad o a una zona específica podría acercar más la predicción a valores precisos.

REFERENCIAS

1.https://www.mdpi.com/2071-1050/10/12/4596
2.https://www.sciencedirect.com/science/article/pii/S2666827021001043
3.https://www.sciencedirect.com/science/article/pii/S2352550921001214