Predicción de cancelaciones en reservaciones de hoteles

1st Laura Tobón Ospian
Universidad de Antioquia
Medellín, Colombia
cecilia.tobon@udea.edu.co

2nd Juan David Arismendy
Universidad de Antioquia
Medellín, Colombia
juan.arismendy@udea.edu.co

***Resumen*— El turismo representa un motor fundamental para la economía mundial, siendo las reservas hoteleras un elemento esencial de su dinámica. No obstante, la gestión operativa de los hoteles se ve considerablemente afectada por la imprevisibilidad de las cancelaciones luego de una reservación. Esta volatilidad en las reservas puede acarrear importantes repercusiones económicas, comprometiendo la rentabilidad del negocio si no se implementan estrategias de mitigación efectivas..**

Palabras clave *- Hote, reservas, impacto económico, Administración de hoteles, predicciones de cancelación*

***Abstract* — Tourism stands as a pivotal industry within the global economy, with hotel bookings forming a critical aspect of its functionality. Nevertheless, hotels encounter a substantial challenge in their operational planning and management: the unpredictable nature of reservation cancellations. This uncertainty can exert a considerable influence on a hotel's profitability, potentially leading to significant financial losses if not addressed through robust management practices.**

Keywords*—Hotel bookings, Economic impact, Hotel management, predictions.*

I. Introducción

En el dinámico y competitivo sector del turismo, la capacidad de anticipar y responder a las fluctuaciones en las reservas de hotel se ha convertido en un factor crítico para la optimización de la gestión y la maximización de la rentabilidad. Sin embargo, la inherente incertidumbre asociada a las cancelaciones de reservas representa un desafío significativo para la planificación operativa y la toma de decisiones estratégicas. Tradicionalmente, los enfoques analíticos podrían haber empleado modelos predictivos basados en ecuaciones únicas aplicadas de manera uniforme a todo el conjunto de datos. No obstante, las reservas no son simples. Muchos factores diferentes entran en juego y se afectan entre sí de maneras complicadas, no siempre directas. Por eso, necesitamos formas más inteligentes de analizar los datos para entender cómo funcionan realmente estas relaciones. A lo largo del presente trabajo, se llevará a cabo un análisis y una exploración de una base de datos relevante para el sector hotelero. El objetivo principal será la aplicación y evaluación de diferentes modelos predictivos, empleando diversas técnicas de análisis de datos y aprendizaje automático, con el fin de identificar aquel modelo que ofrezca la mayor capacidad predictiva para la variable de salida de interés (por ejemplo, la ocurrencia o el tiempo de antelación de una cancelación). Este proceso permitirá no solo comprender mejor los factores que influyen en las cancelaciones, sino también proporcionar a los hoteles herramientas valiosas para la toma de decisiones informadas, la optimización de sus estrategias de precios y la mejora de su gestión de inventario.

II. Descripción del problema

El aprendizaje automático se ha vuelto crucial en todos los negocios, y los hoteles no son la excepción. Una de las principales fuentes de ineficiencia y potencial pérdida de ingresos radica en la incertidumbre generada por la cancelación de reservas. Estas cancelaciones, especialmente si ocurren cerca de la fecha de llegada, pueden resultar en habitaciones vacías y oportunidades perdidas para generar ingresos. La sobreventa, una estrategia común para mitigar este riesgo, también conlleva sus propios desafíos, como la necesidad de reubicar huéspedes y el potencial daño a la reputación del hotel.

Al anticipar qué reservas tienen una mayor probabilidad de ser canceladas, los hoteles pueden tomar decisiones más acertadas. Esto incluye la implementación de estrategias de precios dinámicos, la optimización de la gestión del inventario de habitaciones, la personalización de la comunicación con los huéspedes para reducir la probabilidad de cancelación, y la implementación de políticas de cancelación más efectivas. En última instancia, una predicción precisa de las cancelaciones puede conducir a una mejora significativa en la ocupación, la eficiencia operativa y, por ende, la rentabilidad del negocio hotelero.

La información que se utiliza incluye detalles de reservas de dos hoteles diferentes: un resort y un city hotel, durante un período de tiempo específico. La base de datos consta de **119,390** muestras o registros individuales, cada uno representando una reserva de hotel, el conjunto de datos

incluye **32** variables o características diferentes, que proporcionan información diversa sobre cada reserva. Estas variables abarcan muchos aspectos tales como información de identificación del tipo de hotel, del estado de la reserva, datos de los huéspedes como historial, si requieren parqueadero y cosas similares, también se puede extraer información sobre tipo de servicios contratados o requerimientos especiales. El análisis exploratorio muestra que hay datos faltantes en algunas variables en columnas como country, agent y company. En el caso del país se puede recurrir a mitigarlo mediante la moda (usar el país más frecuente) o analizar si eliminarlo. Para las variables agent y company, que representan identificadores de agencias y compañías respectivamente, si la proporción de valores faltantes es alta, se podría considerar la eliminación de las columnas si se determina que no aportan información predictiva significativa después de un análisis más profundo. La decisión final se tomará tras un análisis más detallado de la distribución de los datos faltantes y su relación con la variable objetivo.

Si analizamos la codificación de algunas variables, tenemos presencia de

Variables Numéricas: (lead_time, arrival_date_year, arrival_date_day_of_month, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, previous_cancellations, previous_bookings_not_canceled, 1 booking_changes, days_in_waiting_list, adr, required_car_parking_spaces, total_of_special_requests) podrán utilizarse directamente o requerir una normalización o estandarización dependiendo de los algoritmos seleccionados, **Variable Binaria:** La variable objetivo is_canceled es binaria (0 para no cancelado, 1 para cancelado) y se utilizará directamente como la etiqueta para los modelos de clasificación.

El objetivo es clasificar cada reserva como "cancelada" o "no cancelada" para poder predecir futuras cancelaciones. Para ello se utilizará un paradigma de aprendizaje supervisado específicamente un problema de clasificación binaria, ya que la variable objetivo que se busca predecir es categórica y solo tiene dos clases posibles, adicional se tienen disponibles datos etiquetados donde cada reserva está marcada con la variable objetivo, y la predicción de una clase para nuevas reservas basándose en sus características nos llevan a trabajar con algoritmos de clasificación.

Para hacer las predicciones, se probarán diferentes maneras de clasificar la información, incluyendo posibles métodos como funciones discriminantes gaussianas, K vecinos más cercanos, redes neuronales feed forward, random forest y/o máquinas de vectores de soporte. Se espera encontrar la mejor manera de predecir qué reservas probablemente se cancelarán.

Al final, este estudio busca ayudar a los hoteles a manejar mejor sus reservas y ganar más dinero. Si pueden saber con anticipación qué reservas tienen una alta probabilidad de ser canceladas, pueden tomar mejores decisiones.

III. ESTADO DEL ARTE

Se analizó un primer artículo [1] el cual utiliza un paradigma de aprendizaje supervisado para predecir cancelaciones de reservas de hotel. Se emplean tres técnicas de aprendizaje automático: Regresión Logística, Random Forest y Extreme Gradient Boosting (XGBoost). Estos modelos se entrenan y evalúan utilizando métricas como precisión, exactitud, recall y F1-score. Además, se optimizan mediante ajuste de hiper parámetros utilizando el método de búsqueda en cuadrícula y validación cruzada de 10 pliegues. Con las pruebas obtuvieron diferentes resultados, Random Forest fue el modelo más preciso después del ajuste de hiper parámetros, con una exactitud de 0.7844 y un F1-Score de 0.8626, XGBoost tuvo un desempeño competitivo, con una exactitud de 0.7811 y un F1-Score de 0.8583. Aunque la Regresión Logística mejoró significativamente tras el ajuste, su exactitud fue menor en comparación con los modelos basados en árboles. Además, se identificó que las características más influyentes en las predicciones fueron lead_time y total_of_special_requests.

Para el siguiente artículo [2] se evidencia que los autores también emplean el paradigma de aprendizaje supervisado, empleando tecnicas de aprendizaje como DNN y regresión logística. Con respecto a las métricas usan Accuracy como medida principal para evaluar el desempeño de los modelos. No se especifican los métodos de validación pero dividieron los datos en 80% para entrenamiento y 20% se reservaron para el conjunto de prueba con lo que pudieron prevenir el sobreajuste. Al hablar de los resultados con las redes neuronales profundas (DNN) la arquitectura Encoder-Decoder alcanzó la mayor precisión con un 86.57% y al ajustar la tasa de aprendizaje, se observó que una tasa más pequeña (0.001) mejoraba la precisión en Decoder-Encoder (con Adamax: 85.91%.) y Encoder-Decoder (con Adadelta: 85.73%.). por otro la do con la regresión logística la precisión inicial fue de 79,66% y al eliminar el atributo country que es otra posible opción de solución, la precisión aumentó al 80,29% y otro atributo que influía en los resultados era total_of_special_request.

REFERENCIAS

[1]. CHATZILADAS, R. PREDICTING HOTEL BOOKING DEMAND AND CANCELLATIONS USING MACHINE LEARNING AND COMPARISON OF

FEATURE IMPORTANCE (Doctoral dissertation, tilburg university).

[2]. Putro, N. A., Septian, R., Widiastuti, W., Maulidah, M., & Pardede, H. F. (2021). Prediction of hotel booking cancellation using deep neural network and logistic regression algorithm. Jurnal Techno Nusa Mandiri, 18(1), 1-8.