Predicción de cancelación de reserva
Entrega 1
Carlos Eduardo Castaño Garzón
Juan Antonio Arango Moreno
Jagler David Velasquez Velasquez
Introducción a la IA
UdeA
2023

En esta competencia se busca predecir el estado de la reserva de un hotel utilizando datos históricos sobre reservas. Debemos crear un modelo de aprendizaje automático a partir de los datos de entrenamiento para predecir el estado de la reserva (cancelada o no) en los datos de prueba proporcionados en el archivo "test.csv".

Nuestro dataset contiene tres archivos CSV: "train.csv", "test.csv" y "simple_submission.csv", es tomado de https://www.kaggle.com/competitions/playground-series-s3e7/data y recortamos algunas filas al archivo "train.csv", ha quedado con 28069.

Los datos de entrenamiento "train.csv" contienen información sobre reservas históricas de hoteles y proporciona una columna llamada "booking_status" que indica si la reserva se canceló o no. Eliminamos manualmente 1500 datos de tres columnas de "train.csv": arrival_year, arrival_month y arrival_date. Esto para cumplir con el requisito del 5% faltante de las 3 columnas.

En "train.csv" podemos ver 19 columnas, de las cuales 6 son categóricas según nuestro criterio, estas son: type_of_meal_plan, required_car_parking_space, room_type_reserved, market_segment_type, repeated_guest, booking_status. Las columnas no categóricas son:

- id
- no_of_adults
- no_of_children
- no_of_weekend_nights
- no_of_week_nights
- lead_time
- arrival_year
- arrival_month
- arrival_date
- no_of_previous_cancellations
- no_of_previous_bookings_not_canceled
- avg_price_per_room
- no_of_special_requests

Las columnas de "test.csv" tienen el mismo nombre que las de "train.csv" pero sin la columna "booking_status" para un total de 18 columnas. "test.csv" tiene entonces 5 columnas categóricas.

El archivo “simple_submission.csv” es un ejemplo del formato de los resultados que se debe proporcionar, donde se debe incluir la columna categórica “booking_status” con valores binarios que representan la predicción de si la reserva se cancelará o no. “simple_submission.csv” contiene dos columnas: Id y booking_status.

En resumen, el objetivo de la competencia es construir un modelo de aprendizaje automático para predecir si se cancelará una reserva de un hotel. El modelo será evaluado utilizando el área bajo la curva ROC entre la probabilidad predicha y el objetivo observado.

Como métrica de negocio hemos decidido reducir la tasa de cancelación de reservas gracias a la utilización del modelo.

La meta de desempeño del modelo en producción debe ser lograr una reducción en la tasa de cancelación de reservas de al menos el 10% en relación con la tasa de cancelación antes de la implementación del modelo.