Trabajo Práctico 1 - Reservas de Hotel

Integrantes:
Ignacio Latorre - Padrón: 101305
Nicolas Ronchese - Padrón: 108169
Gastón Avila - Padrón: 104482

Introducción

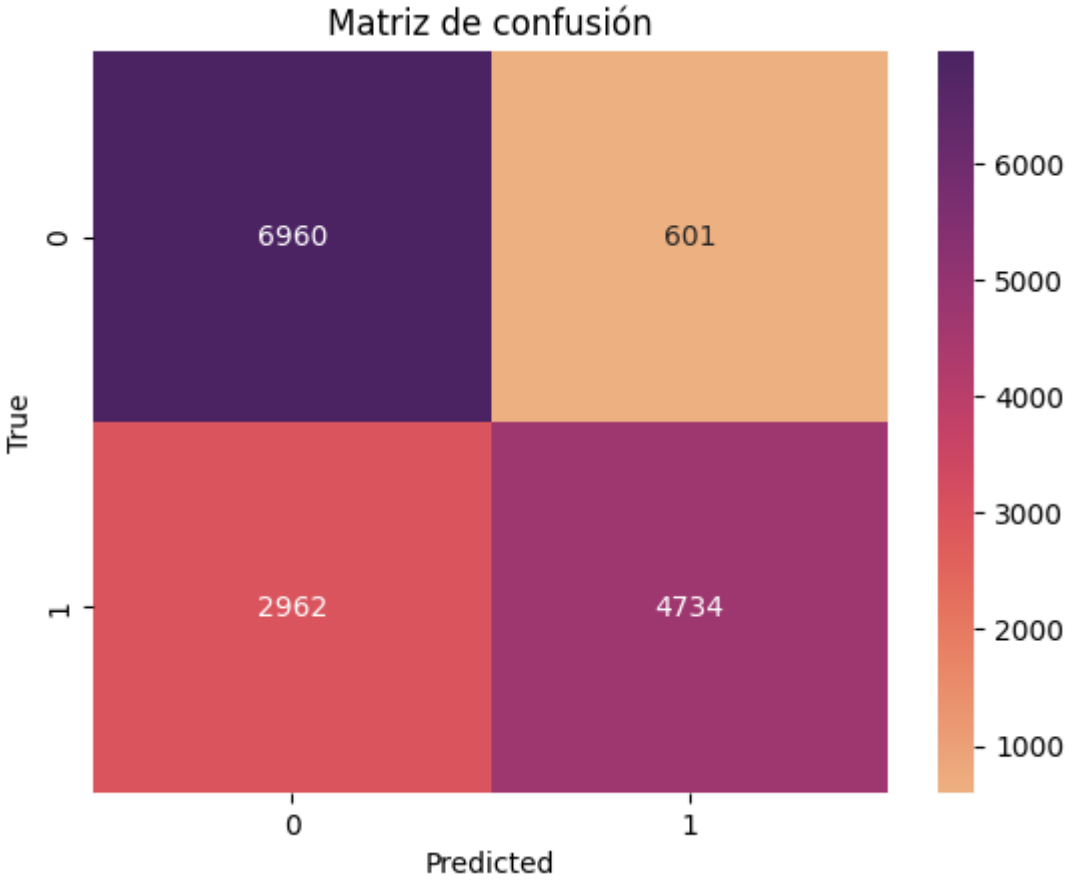
El trabajo práctico consiste en intentar predecir el valor de la variable 'is_canceled' utilizando la técnica de árboles de decisión. Para este tipo de predicción antes de comenzar debíamos hacer unos cambios en nuestro dataset. Estos consistían en el reemplazo de valores nulos y posteriormente cambiar a las variables cualitativas por valores numéricos para que el algoritmo pudiera procesarlos.

A) Optamos por utilizar 4 folds ya que el número total que nos quedó en el dataset test era divisible por 4 y de esta forma evitamos que diferentes folds tuvieran más datos que otros.
La métrica que consideramos más adecuada para buscar los parámetros es el f1 score ya que no solo toma en cuenta tanto el accuracy como el recall sino que también es la que toma el cuenta la competencia de kaggle.

B) Adjunto en la notebook.

C) El árbol de decisiones generado se divide en dos ramas principales basadas en la característica "deposit_type_Non Refund". Esta característica resulta ser la más importante para el modelo, ya que tiene la mayor importancia (53.09%).
Para las reservas reembolsables, se toman en cuenta características como "required_car_parking_spaces", "total_of_special_requests" y "lead_time" para determinar si la reserva será cancelada o no. Estas características tienen importancias relativamente altas en el modelo. Además, si el cliente ha realizado cancelaciones anteriores, es más probable que cancele su reserva actual.
Por otro lado, para las reservas no reembolsables, la característica "customer_type_Transient-Party" se utiliza para determinar si la reserva será cancelada o no. Si la reserva pertenece a este tipo de cliente, es más probable que no sea cancelada.

D)




```
Accuracy: 0.7664678508225733
Recall: 0.6151247401247402
Precision: 0.8873477038425492
f1 score: 0.7265750901695955
```

Se puede ver por las métricas que el modelo tiene una precisión alta de 0.88 lo cual indica que es muy efectivo cuando predice que una reserva fue cancelada. Por otro lado, el recall es bajo (0.62) esto significa que le cuesta identificar correctamente aquellas reservas canceladas. Es por este valor que el F1 Score es de 0.72, con lo cual hay que optimizar los hiperparametros de tal forma que el modelo entrene más para mejorar la detección de las reservas canceladas pero realizando una poda adecuada para evitar el overfitting.