Checkpoint 2

Arbol de decision

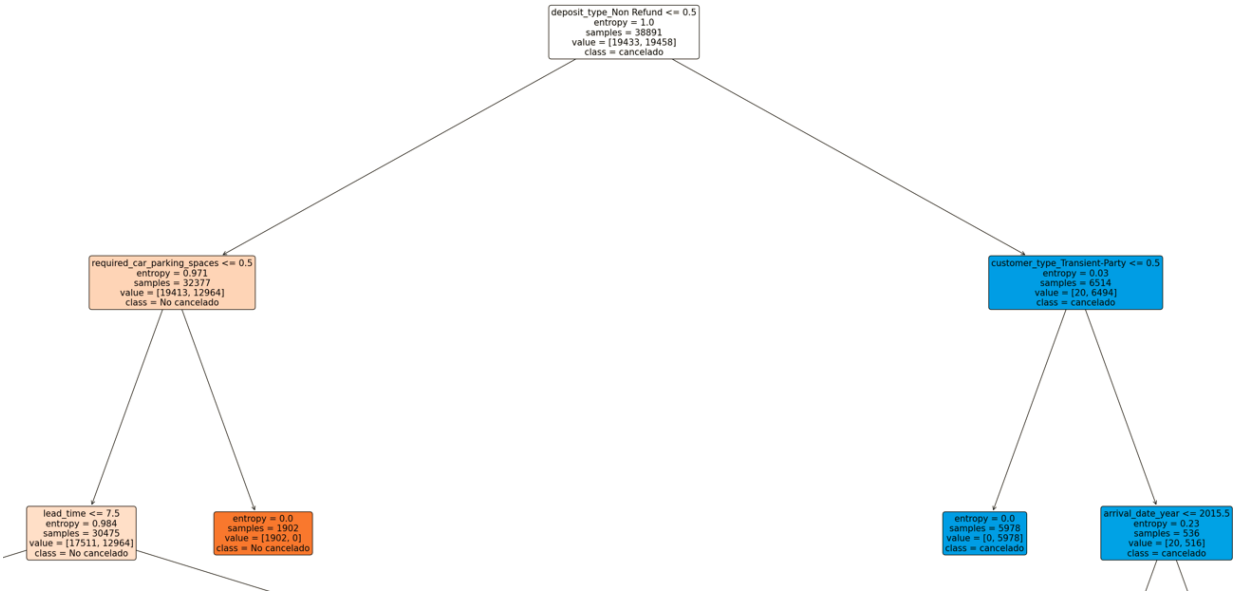
Integrantes: Apaza Axel, Gazzola Franco, Goyzueta Alan, Gonzalez Valentin

Construccion de arboles de decision con optimizacion de hiperparametros mediante k-fold Cross Validation.

Se utilizaron 10 folds y se probaron combinaciones en los siguientes **rangos**

- Numero minimo de muestras para una hoja para que se pueda seguir dividiendo: **(50, 60)**
- Para la poda se uso el parametro cpp_alpha: **(0.001,0.003)**
- Profundidad maxima del arbol max_depth: **(10,15)**

Primeros 2 niveles del arbol con la mejor performance obtenida



Los atributos que eligió el árbol como principales filtros a la hora de clasificar son:

- **deposit_type_Non_Refund**: Este nodo sugiere que si el depósito que hizo el cliente es sin reembolso, el cliente tiene más probabilidad de cancelar la reserva. Esta decisión nos resulta bastante antiintuitiva. Lo que nosotros pensamos es que al ya haber realizado el depósito sin posibilidad de reembolso, el cliente tendería menos a cancelar ya que de lo contrario estaría perdiendo dinero. Sin embargo, se da el caso contrario, quizás se deba a algún fenómeno que no estemos teniendo en cuenta. Tambien los clientes que hacen un depósito no reembolsable suelen reservar con mayor antelación, lo que significa que pueden cambiar de opinión o enfrentar imprevistos que los lleven a cancelar la reserva. Hay que evaluar tambien que este ultimo punto puede ser contradictorio con el analisis hecho para el nodo **lead_time**.
- **required_car_parking_spaces**: El árbol determinó que si la reserva se hace con lugar de estacionamiento requerido, es más probable que no se cancele. No entendemos bien cuál puede ser el motivo de este comportamiento. Pero quizás podría ser que la gente que tiene más movilidad, en este caso un auto, tiene más probabilidades de asistir a la reserva ya que no depende de medios de transporte de terceros que podrían estar o no habilitados.
- **customer_type_Transient-Party**: Como este tipo de reserva puede estar ligada a un evento masivo, por ejemplo bodas o conferencias, si las mismas se ven afectadas por un cambio de fechas o algún otro tipo de situación que comprometa el evento que se va a llevar a cabo, podría ser probable que este tipo de reservas sea más propenso a ser cancelado.
- **lead_time**: Esta decisión puede explicarse como que la gente que hace reservas a corto plazo, en este caso menos de 7 días, es más probable que la reserva no sea cancelada. Esto intuitivamente nos hace sentido ya que hay más probabilidades de que las circunstancias o planes del que realiza la reserva cambien entre más tiempo pasa.

Performance del modelo

- Accuracy: 0.8317734581233501

  La fracción de predicciones que el modelo realizó correctamente.

- Recall: 0.8471177944862155

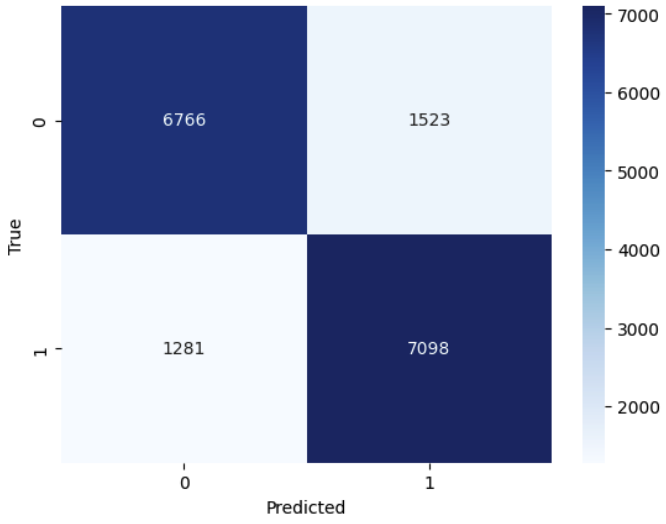
Mide la proporción de resultados verdaderamente positivos que se han identificado correctamente en relación con el número total de resultados positivos reales.

- Precision: 0.8233383598190465

  Mide la proporción de resultados verdaderamente positivos entre todos los resultados clasificados como positivos.

- f1 score: 0.8350588235294116

Matriz de confusion



ACLARACION: A ultimo momento se hizo un sumbit con una limpieza de los outliers existentes en days_in_waiting_list y lead_time. Esto genero una mejora pequenia pero a la vez cambio un poco el arbol por lo que el arbol descrito en el informe tiene ligeras diferencias con el arbol presentado en el informe. Sin embargo los nodos mas importantes siguen siendo los mismos y el analisis de los mismos tambien.