Universidad de Buenos Aires
Facultad De Ingeniería
Año 2023 - 1er Cuatrimestre

# Organización de Datos (75.06 /95.58)
# Trabajo Practico 1: Reservas de Hotel

INFORME FINAL
FECHA: 26/05/23


INTEGRANTES:

| | |
|---|---|
| Gonzalez,Tomás <togonzalez@fi.uba.ar> | #108193 |
| Moreno del Ruvo, Valentina <vmdr18@gmail.com> | #107948 |
| Pol, Juan Manuel <jpol@fi.uba.ar> | #108448 |

# 1. Introducción

El objetivo del trabajo práctico fue a partir de un dataset de reservas de hoteles, generar distintos modelos de clasificación para predecir si una reserva será o no cancelada.
Empezamos haciendo un exploración general del dataset, donde analizamos los distintos tipos de datos que contenía el conjunto, sus distribuciones y como estos se relacionaban entre si. Realizamos gráficos que ayudaron a entender mejor el problema y a encontrar similitudes entre algunas de las variables. También logramos identificar outliers y datos faltantes, resolviendo estás cuestiones y así dejar el dataset en condiciones para el entrenamiento de métodos de predicción.

La metodología para cada modelo creado fue similar, inicialmente realizamos un poco de *feature engineering* en cada set de datos para adaptarlo al modelo a entrenar, luego utilizamos arbitrariamente una serie de parámetros de forma que obtuviésemos una predicción buena. Y luego utilizamos cross-validation y/o k-fold para encontrar los hiperparámetros óptimos.

# 2. Resultados

A partir de los distintos modelos notamos que las variables más significativas a la hora de modelar fueron:

- País de procedencia, específicamente si era o no de Portugal
- Tipo de pago, si la reserva era o no reembolsable
- Lead time

Los modelos con los cuales obtuvimos los mejores score de "f1" fueron Random Forest y SVM Radial, de acuerdo a los testeos de la competencia de Kaggle. Sin embargo, en entrenamiento el mejor fue XGBoost (Figura 1). Potencialmente podríamos decir que otro modelo sería mejor dado que obtuvimos algunos que maximizaban la métrica "recall" como también otros que maximizaban "precision", sin embargo siempre buscamos el modelo más balanceado ya que no sabíamos con exactitud el objetivo del análisis.

A lo largo del trabajo intentamos tratar cada modelo de una manera "aislada", es decir, no inferir sobre un modelo anterior y condicionar al nuevo a esas conclusiones realizadas. Por ejemplo, el primer modelo que realizamos fue el de árbol, a partir del mismo obtuvimos las variables más significativas observadas en la Figura 1, nosotros podríamos haber utilizado esta información para despreciar variables en los siguientes modelos, sin embargo, decidimos continuar utilizando todo el dataset ya que en el momento considerábamos que sería equivocado descartar variables de esa manera.

Nos dimos cuenta que la búsqueda de hiperparametros óptimos muchas veces resultaba inútil, en el sentido que requiere un gran computo para obtener prácticamente iguales resultados. En cuanto al uso de RandomizedSearch o GridSearch resultaron prácticamente intercambiables sin presentar grandes ventajas el uso de una u otra.

Con la experiencia adquirida creemos que hubiésemos tomado un enfoque distinto a la etapa de preprocesamiento ya que nos dimos cuenta de la importancia de conocer los datos y de esta

forma tener al conjunto de entrenamiento en correctas condiciones para el entrenamiento de los modelos. En cuanto a aquellas técnicas que nos hubiesen gustado utilizar, cabe mencionar la reducción de dimensionalidad que para este trabajo no pudimos realizarla puesto que manejábamos mayoritariamente datos cualitativos los cuales no pueden reducir su dimensionalidad.

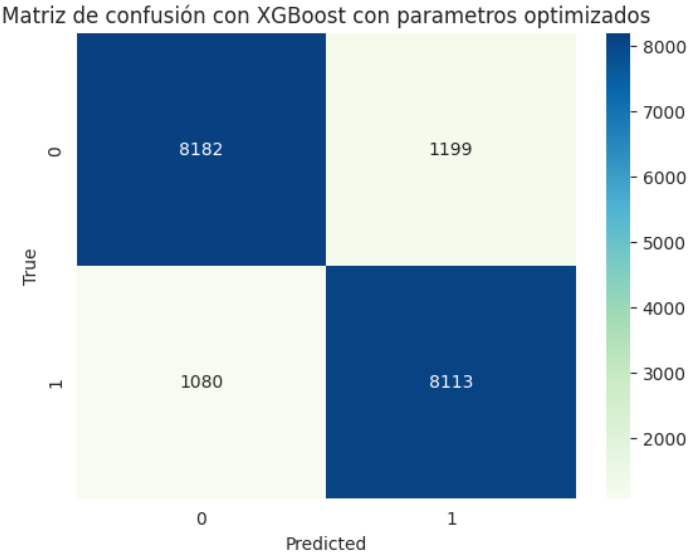


Figura 1: Matriz de confusión para XGBoost con parametros optimizados, mejor modelo en entrenamiento

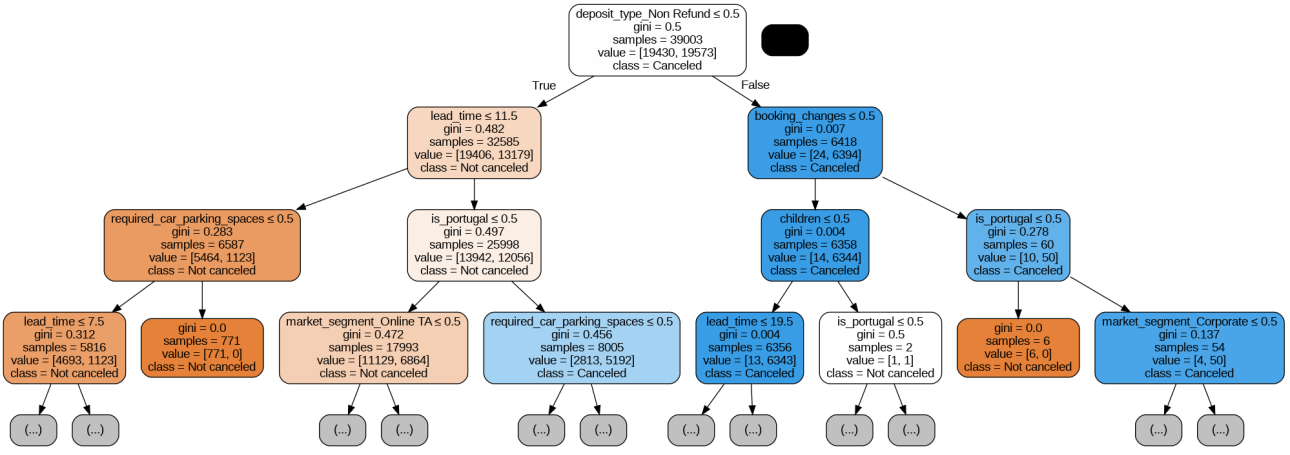


Figura 2: Modelo de árbol de decisión