Assignment 19: Flight Satisfaction Prediction with Random Forest

75 Points scaled to 20 Points

Introduction

In this assignment, you will use the Random Forest machine learning algorithm to perform a binary classification of airline passenger satisfaction using the "airline_passenger_satisfaction.csv" dataset. These data were obtained from Kaggle (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction).

Objectives

- *Prepare data as input for machine learning model training*
- *Partition data into separate training and testing datasets*
- *Visually assess variable correlation*
- *Train a Random Forest model*
- *Assess model performance using withheld validation or test data*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and graphs embedded. Files can be rendered to HTML webpages if your instructor requires this. Graph prompts should be stated and answered within Markdown cells.*

Background Questions

Question 1. Explain the hyperparameters required for the Random Forest algorithm. (5 Points)

Question 2. How is Random Forest different from a single decision tree? (5 Points)

Question 3. Why is it important to assess machine learning models using withheld validation data as opposed to the data points used to train the model? (5 Points)

Questions and Tasks

This assignment can be conducted using either Python or R, whichever you prefer or whichever you instructor requires. Generate code to perform the following tasks and answer the associated questions.

Task 1. Prepare the data by (1) filtering to create a data table that includes only the dependent variable (“Satisfaction”) and the independent variables used to predict it: “Gender”, “Age”, “Customer.Type”, “Type.of.Travel”, “Class”, “Flight.Distance”, “Departure.Delay”, and “Arrival.Delay”. “Gender”, “Age”, “Customer.Type”, “Type.of.Travel”, and “Class” should be treated as nominal data. “Flight.Distance”, “Departure.Delay”, and “Arrival.Delay” should be treated as continuous data. “Satisfaction” has two levels: "Neutral or Dissatisfied" and "Satisfied". Remove any rows with missing data in any column. (10 Points)

Task 2. Create grouped bar plots to compare “Satisfaction” to each of the nominal predictor variables. Describe the results. Which variables seem to be predictive of flight satisfaction. (10 Points)

Task 3. Create grouped boxplots to compare “Satisfaction” to each of the continuous predictor variables. Describe the results. Which variables seem to be predictive of flight satisfaction. (10 Points)

Task 4. Split the data into separate, non-overlapping training and testing sets. 75% of the data should be used to train the model while 25% should be maintained for validation. Stratify the partition using the “Satisfaction” variable. (5 Points)

Task 5. Train a Random Forest model. Use 500 trees and 3 variables to select from for splitting at each decision node. You do not need to optimize the hyperparameters for this assignment. (5 Points)

Task 6. Use the trained model to predict to the withheld test data. From the results, calculate the Area Under the Receiver Operating Characteristic (ROC) Curve and Overall Accuracy metrics. Also, create a Confusion Matrix. Discuss thee results of the assessment. (10 Points)

Task 7. Use the variable importance estimates made available by Random Forest to discuss the contribution of each predictor variable in the model. (10 Points)