**70-374 Midterm Exam Spring 2021**

**Open-Book Coding Test (30%): Predicting Delayed Flights.**

**You have 2 hours to complete the test.**

You need to submit two files:

* R file:
  + Download **FlightDelay\_empty.R** and write codes in this file.
  + Make sure to organize the codes under the corresponding question numbers.
  + After complete, please change the file name into: **FlightDelay\_YourName\_StudentID.R** and submit.
* pdf file:
  + Download **Midterm 2021 Coding Question.docx** and include **only the outputs** from your codes. Note, **only** the text boxes and tables highlighted in **red text** need to be filled. You **do not** need to provide any output for all the other parts. For instance, you don’t need to provide any output for question 1, 2, or 3.
  + After complete, please convert it into pdf file, change the file name into: **CodingOutput\_YourName\_StudentID.pdf** and submit.

The input data (FlightDelays.csv) contains information on all commercial flights departing the Washington, DC area and arriving at New York during January 2004. For each flight, there is information on the departure and arrival airports, the distance of the route, the scheduled time and date of the flight, and so on. The variable that we are trying to predict is whether or not a flight is delayed. A delay is defined as an arrival that is at least 15 minutes later than scheduled.

In short, we are classifying flights into *delayed* or *ontime* based on attributes, including: 'CARRIER', 'DEST', 'DISTANCE', 'ORIGIN', 'Weather', 'DAY\_WEEK'.

In particular, the target attribute is “Flight.Status”; and the X variables showed up in the decision tree plot are:

* Weather: “good” or “bad”
* CARRIER: “DL” (delta), “US” (US airways), etc
* DEST (destination): “LGA”, etc
* ORIGIN (origin): “LGA”, etc
* DAY\_WEEK (day of week): 1, 2, …, 7 (indicating Sunday, Monday, …, Saturday)
* DISTANCE (distance of the flight): an integer
  1. (3 pts) Read the data “FlightDelays.csv”
  2. (3 pts) Create a dataframe by selecting the following variables and name it as: delays.df.
* 'CARRIER','DEST','DISTANCE','ORIGIN', 'Weather','DAY\_WEEK','Flight.Status'

From now on, only use these variables to complete all the tasks.

* 1. (3 pts) Force all the variables (except for “DISTANCE”) ” to be factors
  2. Plots/Tables:
     1. (3 pts) Create a side-by-side boxplot to describe the DISTANCE’s distribution with different CARRIERs.
     2. (3 pts) Use CrossTable to learn how Flight.Status is related with different Weather.

PLEASE PASTE YOUR PLOT/TABLE OUTPUTS HERE FROM R.

4a)

Chart, box and whisker chart

Description automatically generated

PLEASE PASTE YOUR PLOT/TABLE OUTPUTS HERE FROM R.

4b)

Table

Description automatically generated

* 1. (3 pts) Partition the delays.df into training set (60%) and testing set (40%). Please set seed as 1 using set.seed(1)
  2. (3 pts) Fit a classification tree to the flight delay variable ('Flight.Status') using all the predictors. Note:
     + Set seed as 1 using set.seed(1)
     + Set method as “rpart”
     + Use 5 fold cross-validation
     + Use tuneLength = 30

* 1. After creating the tree, please plot the following:
     1. (3 pts) Plot how accuracies (cross-validation) changes with complexity parameters
     2. (3 pts) Plot the decision tree using rpart.plot

PLEASE PASTE YOUR OUTPUTS HERE FROM R.

7b)

Diagram

Description automatically generated

PLEASE PASTE YOUR OUTPUTS HERE FROM R.

7a)

Chart

Description automatically generated with medium confidence

Values between 0.01 and 0.06 seems to be equally the best cp.

* 1. (3 pts) Use this tree to predict the Y labels in the testing set. And evaluate the model performance on the testing set using confusion matrix.

PLEASE FILL IN THE FOUR BLANK CELLS BELOW.

|  |  |  |
| --- | --- | --- |
| Prediction\Actual | delayed | ontime |
| delayed | 6 | 0 |
| ontime | 107 | 548 |