Q3: Flight Delays

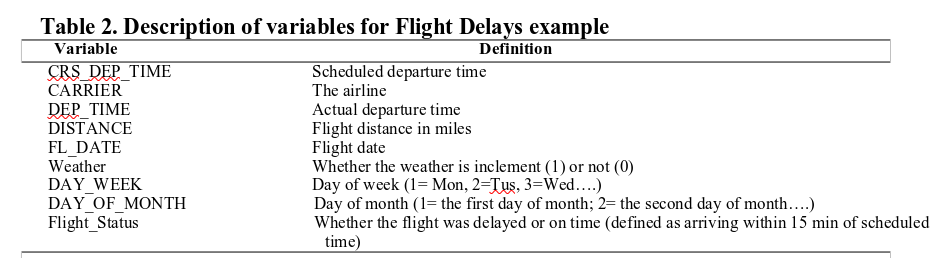
*Fall, 2022*

G Park

2022-12-04

# **Question Three: Predicting Delayed Flights (30 points)**

The file 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 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 (Fight\_Status). Table 2 describes variables in this file.

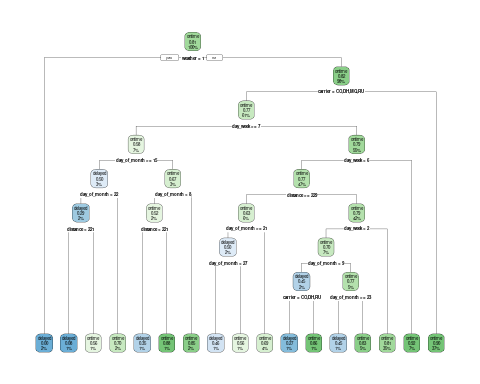


Flights data

## Partition records into 60% for training and 40% for validation sets. Fit a classification tree to the flight delay variable using all the relevant predictors in FlightDelays.csv on training sets with maximum of 8 levels and set up cp =0.001 and then plot the tree. (5 points)

Note: cp refers to complexity parameter

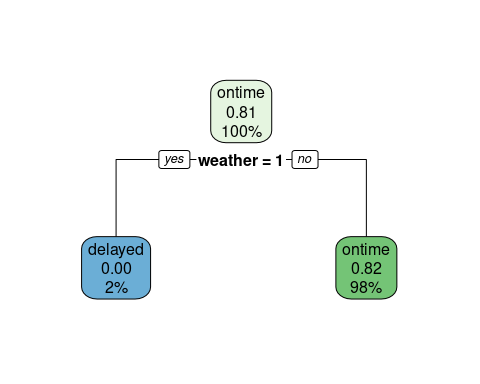
I the fit the classification tree.



## In the setting of decision tree, there is a technique called pruning the tree. Discuss the purpose of pruning the tree and why we may need to prune the tree. Finally, then prune 3.1. tree and plot the pruned tree. (10 points)

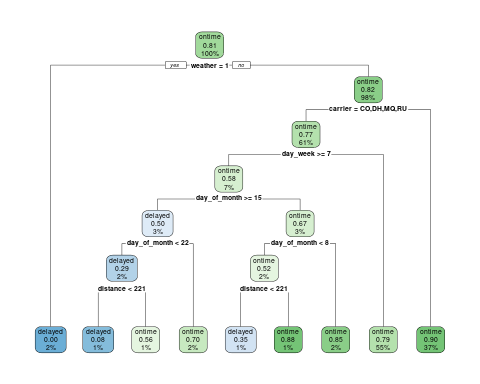
Pruning is a data compression technique for reducing the size of decision trees by removing non-critical and redundant sections of the tree. The purpose of pruning is to reduce the complexity of the classifier. Pruning also helps improves predictive accuracy reducing of over-fitting.

In this section, I prune the tree we grew in section 3 above.In pruning this tree, I raise the complexity parameter by a factor of 10 to 0.1.



## Fit a new classification tree to the flight delay variable using all the relevant predictors on training sets, excluding the Weather predictor. Set cp=0.001 and maximum =6. Plot this new classification tree. (10 points)

In this section, I create a new classification tree with cp = 0.001 and maximum depth of 6. I then plot the tree.



## Based on the tree from 3.3, do predictions for both training and validations sets and report their confusion matrix respectively. (5 points)

I start by doing predictions on the training set and report the confusion matrix.

## Truth  
## Prediction delayed ontime  
## delayed 42 7  
## ontime 214 1056

I now do the predictions on the test set and likewise, report the confusion matrix.

## Truth  
## Prediction delayed ontime  
## delayed 24 8  
## ontime 148 702