Case Study: Flight Delays

A large travel agency has asked us to predict whether a flight will be canceled based on several factors. The agency can sell tickets for only three airlines (AA, UA, and DL) and would like to be able to advise its customers on which airline has the least risk of cancellation. Using the dataset **flight.csv** provided on GitHub. The steps were done are as follows:

1. Build a model to predict whether a flight will be canceled. Apply some model selection methods, cross-validation methods, or any other technique to tune the model performance, to maximize the AUC or prediction accuracy for (unseen) new cases.
2. Write a scoring script in R to predict for new cases. The scoring script should contain two functions, named **func1** and **func2**. Both functions take as input argument a data.frame containing test cases. func1 should return a vector of probabilities and func2 should return a vector of 0 or 1s as predictions. The length of the returned vector should equal to the number of rows in the input dataframe, and the orders should match. Do not sort the result. Usage example:

   scores <- func1(newdata = test_cases_df)
   predictions <- func2(newdata = test_cases_df)

   The scoring script should be lean – do not include training/tuning process in it, just the scoring function.

   The scoring functions (i.e., func1 and func2) is:

   source("functionscript.R")

   The scores will be used for constructing an ROC curve against the test case ground truth, and an area under the curve (AUC) will be calculated; the predictions will be used for calculating the prediction accuracy.

3. I put the work code (i.e., EDA, training, tuning, analysis, etc.) in a separate R file.

The dataset includes the following fields.

Field	Name Type	Description
Canceled	Binary	Canceled = 1
Month	Integer	Jan = 1
DepartureTime	Integer	Military Time (1:00 PM = 1300)

Field	Name Type	Description
UniqueCarrier	String	Airline Carrier Code
SchedElapsedTime	Integer	Scheduled Flight time in minutes
ArrDelay	Integer	Arrival delay in minutes
DepDelay	Integer	Departure delay in minutes
Distance	Integer	Distance in miles