# Case Study #3: Logistic Regression for Flight Status

The *FlightDelays.csv* file contains a data set with the information on all commercial flights departing the Washington, DC area and arriving at one of the New York area airports. For each flight, there is information on the departure and arrival airports, the distance of the route, the scheduled time and day of the flight, and so on. The outcome that you need to predict is the flight arrival status, i.e., whether a flight is delayed or on time (*FL_STATUS*). A delay is defined as an arrival that is at least 15 minutes later than scheduled. The table below describes each of the columns in the data set.

| ***Variables*** | ***Description of Variables*** |
|---|---|
| DEP_TIME | Scheduled departure time. |
| CARRIER | Carrier abbreviation: CO=Continental, DH=Discovery Airways, DL= Delta Airlines, MQ=American Eagle, OH=PSA Airlines, RU=AirBridge Cargo, UA=United Airlines, US=US Airways. |
| DEP_TIME | Actual flight departure time. |
| DEST | Abbreviation of the destination airport in New York area: EWR=Newark, JFK=John F. Kennedy, LGA= LaGuardia. |
| DISTANCE | Distance of the route in miles. |
| FL_NUM | Flight number. |
| ORIGIN | Abbreviation of the DC area airport: BWI=Baltimore /Washington, DCA=Ronald Reagan Washington, and IAD=Dulles International. |
| WEATHER | Weather condition for a flight: good flying condition = 0, poor flying condition = 1. |
| WK_ DAY | Day of the week, from Monday = 1 through Sunday = 7. |
| MTH_DAY | Day of the month, from 1 through 31. |
| FL_STATUS | Flight arrival status, two classes: 'delayed' and 'ontime'. |

## Questions

1. Upload, explore, clean, and preprocess data.
    a. Why a logistic regression model may be used in this case? Why may you not apply a multiple linear regression model in this case? Provide brief answers to both questions.
    b. Create a *flight_df* data frame by uploading the original data set into Python. Remove '*DEST*' and '*ORIGIN*' variables from the *flight_df* data frame. Convert '*CARRIER*' and '*FL_STATUS*' into binary variables. This portion of part 1 will not be graded.
    c. Why does the output variable '*FL_STATUS*' need to be converted into binary variables for logistic regression? Briefly explain.

2. Develop a logistic regression model for the Flight Delays case.
    a. Develop in Python the predictor variables (14 variables) and outcome variable ('*FL_STATUS'*) and partition the data set (70% for training and 30% for validation partitions). Train a logistic regression model using *LogisticRegression()* with the training

data set and display in Python the model's parameters (intercept and regression coefficients). Provide these parameters in your report and also present the mathematical equation of the trained logistic regression model.

b. In Python, make predictions and identify probabilities *p(0)* and *p(1)* for the validation data set. For the first 20 records in the validation data set, display a table that contains the actual and predicted flight arrival status, and probabilities *p(0)* and *p(1)*. Present this table in your report, and comment on the predicted vs. actual flight arrival status.

c. Identify and display in Python confusion matrices for the training and validation partitions. Present them in your report and comment on accuracy (misclassification) rate for both partitions and explain if there is a possibility of overfitting.

d. Create and display in Python the *Lift* chart only for 'delayed' flight status. For that, use *p(0)* for *.sort_values()* and *p(0)* in *liftChart()*. Also, use *ncols=1* in *plt.subplots()* for a single plot, and remove *ax=axes[1]* from *liftChart()*. Present this Lift chart in your report and briefly explain what the chart demonstrates and what conclusion(s) can be made.

3. Compare results of logistic regression model vs. classification tree model for the same data set.

   a. Present and compare in your report the validation confusion matrix for the logistic regression model in 2c of this case versus the validation confusion matrix using the *GridSearchCV()* algorithm for the classification tree in the previous case study. Using the accuracy value (misclassification rate), which model would you recommend applying for classification (prediction) of flight arrival status? Briefly explain your answer.

4. Extra Credit (Optional).

   a. For the logistic regression in 2a, consider using the Backward Elimination algorithm to reduce the number of predictors. In the *train_model()* function, consider *model = LogisticRegression(max_iter=500).* Develop a logistic regression model based on the best predictor variables from Backward Elimination and present the intercept and regression coefficients (apply *.coef_[0]* )in your report. What specific predictors were removed from this model?

   b. For the logistic regression model in 4a, identify and compare the confusion matrices for training and validation partitions. Also, compare this validation confusion matrix with the one from 2c, and explain if the logistic regression model based on Backward Elimination can be potentially a good choice for classification of flight arrival status.