**Flights Project**

With these instructions you should have received two datasets.

*Flight\_history.csv – Details on outbound flights from NYC airports from January to June 2013*

*Flight\_test.csv – Randomly selected unique flights from the month of July.*

Your goal is to predict which of these flights will arrive late using the ‘lateflight’ outcome variable. A late flight is defined as arriving after the scheduled arrival time. The only information you are given in the *flight\_test.csv* dataset is:

* uniqueid = unique record number
* year, month, day = date of the flight broken up into year, month and day
* sched\_dep\_time = scheduled departure of the flight in the local timezone (HHMM)
* sched\_arr\_time = scheduled arrival of the flight in the local timezone (HHMM)
* carrier = 2byte carrier code
* flight = flight number
* tailnum = tail number of the plane
* origin = origin airport code
* dest = destination airport code
* hour, minute = scheduled departure of the flight broken up into hour and minute
* time\_hour = time and date stamp of the hour of scheduled departure
* lateflight = dependent 0/1 variable where 1 means the flight arrived after the scheduled arrival time

It will be difficult to predict which flights will be delayed based on this limited information. Your goal is to use the *flight\_history.csv* dataset to construct additional meta-variables for use in your predictive model. (List of variables included in *flight­\_history.csv* is on the next page) For example, you could use the information provided in *flight\_history.csv* to:

* find the distance between the origin and the destination
* % of all flights for that carrier that are delayed.

Objectives:

1. Create at least two additional variables from *flight\_history.csv* and append those to *flight\_test.csv*. Tip – be sure that whatever variables you create you are able to append using the information provided in *flight\_test*
2. Build a model predicting ‘lateflight’.
3. Create a 10-15 minute presentation detailing:
   1. Descriptions of the variables that you created and the code used to calculate those variables.
   2. Describe the type of model chosen and any parameter settings or other decisions made in specifying that model.
   3. Summarize and interpret the results of the model and its ability to predict which flights would be delayed.

Notes:

You are allowed to use any type of model you would consider appropriate. You may use any technique to create your new variables as long as you are able to explain what effect that variable is capturing. You can use any software program that you are comfortable with for the analysis and creation of the presentation.

|  |  |
| --- | --- |
| year | Numeric year for that flight |
| month | Numeric month for that flight |
| day | Numeric day for that flight |
| dep\_time | Actual Time of departure (24 hr clock –HHMM, local timezone) |
| sched\_dep\_time | Scheduled Time of departure (24 hr clock –HHMM, local timezone) |
| dep\_delay | Delay of departure in minutes |
| arr\_time | Actual Arrival time of arrival (24 hr clock -HHMM, local timezone) |
| sched\_arr\_time | Scheduled Time of arrival (24 hr clock – HHMM, local timezone) |
| arr\_delay | Delay of arrival in minutes |
| carrier | 2 byte carrier code |
| flight | flight number |
| tailnum | tail number of the plane for that flight |
| origin | 3 byte origin airport code |
| dest | 3 byte destination airport code |
| air\_time | length of flight spent in air in minutes |
| distance | distance between origin and dest airports in miles |
| hour | hour of scheduled time of departure |
| minute | minute of scheduled time of departure |
| time\_hour | Day and Time of scheduled departure to the hour |

These datasets were constructed using the nycflights13 package available for R using the flights dataset contained in that package. These datasets have been modified slightly from the original dataset. For more details on the package - <https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf>.