Adnan Aygunduz  
Capstone #1  
Springboard DS, March 2019

**Can Air Accidents and Incidents Be Predicted in Advance?**

**Problem:** Air accidents and incidents can cost lives and huge resources. Understanding the factors behind them based on the previous data we have can help authorities to take preventative measures in advance and save lives and valuable assets. Thus, this project will answer the question if we can create algorithms in order to predict before it happens or not based on the collection of data we have so far.

**Client:** Since air transportation safety and specifically flight safety is the paramount, government organizations such as National Transportation Safety Board (NTSB) and commercial air carriers will benefit from the outcomes of this study.

**Data:** The NTSB aviation accident database contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters. Generally, a preliminary report is available online within a few days of an accident. Factual information is added when available, and when the investigation is completed, the preliminary report is replaced with a final description of the accident and its probable cause. Full narrative descriptions may not be available for dates before 1993, cases under revision, or where NTSB did not have primary investigative responsibility.

**Approach:** I intent to use supervised learning model for predicting Investigation Type target value under two classes Accident and Incident.

Accident is defined as "An occurrence associated with the operation of an aircraft which takes place between the time any person boards the aircraft with the intention of flight until such time as all such persons have disembarked, in which a person is fatally or seriously injured."

Incident is defined as "An occurrence, other than an accident, associated with the operation of an aircraft which affects or could affect the safety of operation."

Vast variety of variables will be used as predictors such as phase of flight, weather conditions, age of aircraft, number of engines, model, make, carrier etc. Randomly selected 20% of the same data set will be used as training data.

I will approach this machine learning project by following the steps below:

a. Create a repository in Github (in addition, create google driver to use shareable documents with my mentor)

b. Gather the data from NTSB and load it into Python.

c. Analyze the data to determine the data quality

d. Gather additional supportive data from NTSB Accident narratives if needed.

e. Prepare the data:

(1) Clean that which may require it (remove duplicates, deal with missing values, correct errors, normalization, data type conversions, etc.)

(2) Visualize data to help detect relevant relationships between variables and perform exploratory analysis.

f. Split the data as training and test data

g. Choose a machine learning algorithm

h. Train the model

i. Evaluate the model

j. Parameter Tuning

k. Make predictions.

l. Prepare a report

**Deliverables:** 10-12 pages project report, source code and slide deck will be delivered in Github.

**Data Set:** <https://public.opendatasoft.com/explore/dataset/ntsb-aviation-accident-dataset/information/?dataChart=eyJxdWVyaWVzIjpbeyJjaGFydHMiOlt7InR5cGUiOiJsaW5lIiwiZnVuYyI6IkNPVU5UIiwic2NpZW50aWZpY0Rpc3BsYXkiOnRydWUsImNvbG9yIjoiI0QwNTM1NiJ9XSwieEF4aXMiOiJldmVudF9kYXRlIiwibWF4cG9pbnRzIjoiIiwidGltZXNjYWxlIjoieWVhciIsInNvcnQiOiIiLCJjb25maWciOnsiZGF0YXNldCI6Im50c2ItYXZpYXRpb24tYWNjaWRlbnQtZGF0YXNldCIsIm9wdGlvbnMiOnt9fX1dLCJzaW5nbGVBeGlzIjpmYWxzZSwiZGlzcGxheUxlZ2VuZCI6dHJ1ZSwiYWxpZ25Nb250aCI6dHJ1ZX0%3D&calendarview=month&location=2,24.62597,-0.55917&basemap=jawg.streets>