# Capstone Project 1 - Flight Delays - Final Report

**Problem Statement:**
- To Predict whether or not the departure delay on an Origin-Destination pair, with a 6-hour prediction horizon, will exceed a 15 min threshold; such that the future delay will fall into one of two classes, 'above threshold' (1) and 'below threshold' (0).

**Background:**
- Delays result  in the need for additional gates and ground personnel, and impose costs on airline customers (including shippers in the form of lost productivity, and additional wages).
- Comprehensive studies showed $40 billion in annual costs in 2015, and in 2016, the average cost of aircraft block (taxi plus airborne) time for U.S. passenger airlines was estimated to be $62.55 per minute  ("Per-Minute Cost of Delays to US Airlines", airlines.org).
- Although the inherently complex nature of Air transportation, with constrained airspace and airport resources, thousands of aircraft and overburdened air-traffic controllers, and unforeseen weather disruptions make delays at times inevitable; nearly 40% of delays were due to the delayed arrival of the incoming aircraft, reflecting the high levels of interdependence in the delay dynamics (Gopalkrishnan et al. 2017).
- 15 minutes is the threshold being utilized as that is what the Bureau of Transportation Statistics considers when classifying a commercial airline flight as delayed.

**Potential Clients:**
- The Air Traffic Control System Command Center (ATCSCC) which receives Airport Arrival Rate (AAR) data from  air traffic facilities, institutes Ground Delay Programs, and runs compressions to alleviate congested networks.
- Companies distributing use of the Aircraft Situation Display to Industry (ASDI) from the US Dept. of Transportation's Volpe Transportation Center, which provides real-time position and flight plans in U.S. airspace.
- Air Route Traffic Control Centers (ARTCC) looking to make better predictions of AAR's and plan efficient holding patterns.
- Implementers of Ground Delay Programs at Domestic Airports who desire a head start in order to alleviate the costs that result from inefficient tarmac operations during congested network states.
- Traffic Management Personnel seeking to file flight plans as early as possible to get the earliest Expect Departure Clearance Time (EDCT), and avoid getting pushed into the next time block.

**Methodology:**

- Acquired from an online posting on Kaggle containing US domestic flight delays in the month of January,August, November and December of 2016.
- Original flight data consisted of 1,824,403 entries and 36 features:
    - 8 temporal
    - 14 categorical (6 origin, 6 destination, flight number, and unique carrier code)
    - 14 metric (timestamps, distance, actual vs elapsed CRS times)
- Rows with missing data were dropped.
- Values for categorical features were adjusted to improve readability and ease of visualization (codes changed to names, etc.).
- A datetime index was created from the temporal features to assist in further analysis.
- A mock air traffic network was created by grouping all flights into origin-destination pairs (interchangeably also called links), and resampling the top 300 links by hour, then calculating an aggregate median for each flight metric per link per hour during the sample period.
- The top 300 links were chosen because they had a high frequency of daily traffic, and a median versus a mean calculation was utilized in order to get a better picture of the spread in the presence of prominent outliers that might otherwise skew observations.
- Inferential statistical analysis was then used to confirm the validity of taking a directed network approach, looking at departure exclusively in order to predict overall delay characteristics, since departure and arrival delay had a significantly direct relationship.
- Finally, categorical features were dummified, and a binary target class was created by attributing a value of 1 if the link was in a state of delay 6 hours from the current time ('above threshold'), and 0 ('below threshold') otherwise.

**Model Construction:**

- Due to the imbalanced distribution of the delayed class , 22.7% of dataframe being used for model construction, the baseline classifier performed very poorly in predicting true positives.
- Baseline consisted of a logistic regression classifier trained on the imbalanced dataset, of which the average results of 5-fold cross-validation on the test data yielded an accuracy of 0.7721, but upon further inspection accuracy was 0 for the ‘above threshold’ class, thereby providing misleading results as all points were simply estimated to belong in the ‘below threshold’ class.
- Accuracy was deemed to be an inappropriate evaluation metric in this case, with the individual Precision, Recall, and F1-scores for each class being more indicative of model performance.
- Under- and Over- sampling techniques were utilized on the training dataset to counter dataset imbalance, with each method being used to train Random Forest and Logistic Regression classifiers.
- The performance of classifiers under both conditions were measured using cross-validation with the test set containing the natural distribution of classes.

**Results:**

- Utilizing the AUC score as the primary comparative statistic for model performance, the Random Forest Classifier trained on data oversampled using SMOTE (Synthetic Minority Oversampling Technique) performed best, with an AUC of 0.85, a Precision score for the delay class of 0.67, and an F1-score of 0.64.
- The Random Forest Classifier trained on data using the Random Under Sampling technique had a higher Recall score for the delay class (0.77), but considerably lower Precision, with a score of 0.49. Meaning it had a lower false positive rate, but of the additional results being returned, most had incorrect labels in comparison to the training labels.

**Conclusions:**

- Of the two four estimators trained, the logistic regression models performed the worst, proving unreliable for this particular problem, although they have potential for improvement if weights were added to the classes.
- Conversely, the ensemble approach appears to more appropriate for this problem, and of the two Random Forest Classifiers, the one trained on the SMOTE dataset yielded the most useful results, and would be my recommendation to a potential client seeking to predict a delay state on any particular origin-destination pairs.

**Client Recommendations:**

- Implementation of this model would be possible for any of the potential clients delineated earlier, as all would be recipients of ASDI (real-time position and flight plans in U.S. airspace) data, but for the purpose of providing a larger perspective, uses of this model can be split into three categories grouped by practitioner:

(1) For data distributors:

- The predictive component provided by this model can be manipulated and re-trained into different classifier packages by utilizing alternative prediction horizons (besides 6 hours in the future), and thresholds for delay (15 min was used in this instance), allowing users to have a customized view into future delay states.

(2) For air-traffic controllers:

- By having a concise prediction of delay states on origin-destination pairs available, which requires minimal calculation on the part of Traffic Management Personnel in comparison to current practices, flight plans and Ground Delay Programs can be crafted with a forward-looking approach, reducing Airport Arrival Rates and preventable propagated delays.

(3) For airport ground crews:

- Proverbially the frontlines of the delay problem when it comes to air-traffic networks, the simple yet useful output of this model has the potential to yield the earliest meaningful results for this group, as even a few hours of preparation can result in substantial gain in terms of avoiding wasted manpower, and establishing relief efforts in congested runways.