**November 5, 2016**

**Prescriptive Analytics on Flight Delays**

Akshay Shrinivasan, Vineel Lakshman, Jashkumar Shah

(A20353163) (A20356842) (A20352470)

**Introduction**

San Francisco International Airport (SFO) and Oakland International Airport (OAK) sit directly across the San Francisco Bay from each other and are separated by about 12 miles of water. Similarly, John F. Kennedy International Airport and Newark Liberty International Airport also sit directly across Hudson river and they are separated by about 34 Miles. Lastly, we have Midway International Airport and O’Hare International Airport separated by about 20 miles.

There is a myth among business travelers across these areas that it is better to fly out of OAK than SFO, Newark rather than John F. Kennedy, Midway rather than O’Hare because of an elevated chance of delays due to several factors like road and air traffic, fog, etc. However, this myth seems to be contradicted by the data. According to the [Bureau of Transportation Statistics](http://www.rita.dot.gov/bts/subject_areas/airline_information/airline_ontime_tables/2013_12/table_06), all of the alternative airports stated (Oakland, Newark, Midway) had an on-time departure rate less than their respective counterparts (SFO, John F. Kennedy, O’Hare) in 2015. Our project will create a predictive model to answer the question: “Given a specific date and destination, is it better from a probabilistic perspective to fly out of OAK or SFO, Newark or John F. Kennedy, Midway or O’Hare?”

**Project scope**

The model assumes that the inputs to a user’s query are: (1) date of travel and (2) destination. This takes several assumptions. The model is designed for business users who know that they will travel on a certain date, but are uncertain as to the probabilities on that date of a flight delay. The output from our model will be a simple recommendation of the airport from one of the three locations stated, from which the travller should depart. Accordingly, this will entail the development of a classification model.

**The data**

We will be using publicly-available data on flight arrivals and departures for major U.S. airports from the American Statistical Association. The data set contains some key attributes that we will use to design our algorithm. In particular, we expect to make significant use of attributes-Tail Number (number plate-unique for each flight), Destination, Departure Delay Time, Month, Day of Month, Day of Week, Carrier, Flight Number, Origin for the predictions.

The data set covers all flights from 2008 - 2015

**Assumptions**

There are several key assumptions that our analysis will make that need to be noted:

1. Travelers prefer only non-stop flights
2. Assuming all flights will take off from their Source and there are no cancellations
3. We will Ignore El Nino Effect as there is not much relevant data.

([**http://www.climatechangenews.com/2015/07/14/how-el-nino-is-delaying-your-flight-and-upping-emissions/**](http://www.climatechangenews.com/2015/07/14/how-el-nino-is-delaying-your-flight-and-upping-emissions/))

1. Assuming nearly same number of flights depart from OAK or SF, ORD or Midway, JFK or Newark to the destination

**Challenges that could be faced during this Project**

1. Data cleansing for huge data set
2. Construction of a Naïve Bayesian Classifier Model to accurately predict the probability of delays.
3. Construction of a Logistic Regression Model for prediction.
4. Accurate Prediction of the delays which is affected by factors such as new landing protocols brought into notice by the FAA, effect of fog, etc.

**Analysis/ Proposed Solution**

The following is a rough outline of how we expect to construct our model:

1. Preprocessing the data.
2. Partition data into training and validation data sets.
3. Identifying/Generating Labels for classification.
4. Competitive analysis of Classification models (Naïve Bayes and Logistic Regression)
5. The effect of fog can be a major challenge as it can significantly affect our prediction and the assumptions we take. For this we have decided to create buckets during peak fog time.
6. The new landing protocols brought in will be ignored as we do not have any sufficient data in this regard
7. Validate results obtained from both the models- We will use 80% of the data to train our models and 20% to validate them. Our methodology is to use the model to predict which airport is better on a given day and compare this to the training data.
8. Conclusions - Based on the validation results, we will make conclusions about the suitability of our models for delay predictions. We will then create a visualization as to which is the preferred airport from which to depart for a few selected destinations. Our predictive model is most likely to predict a decision for 6 Months (200 days).