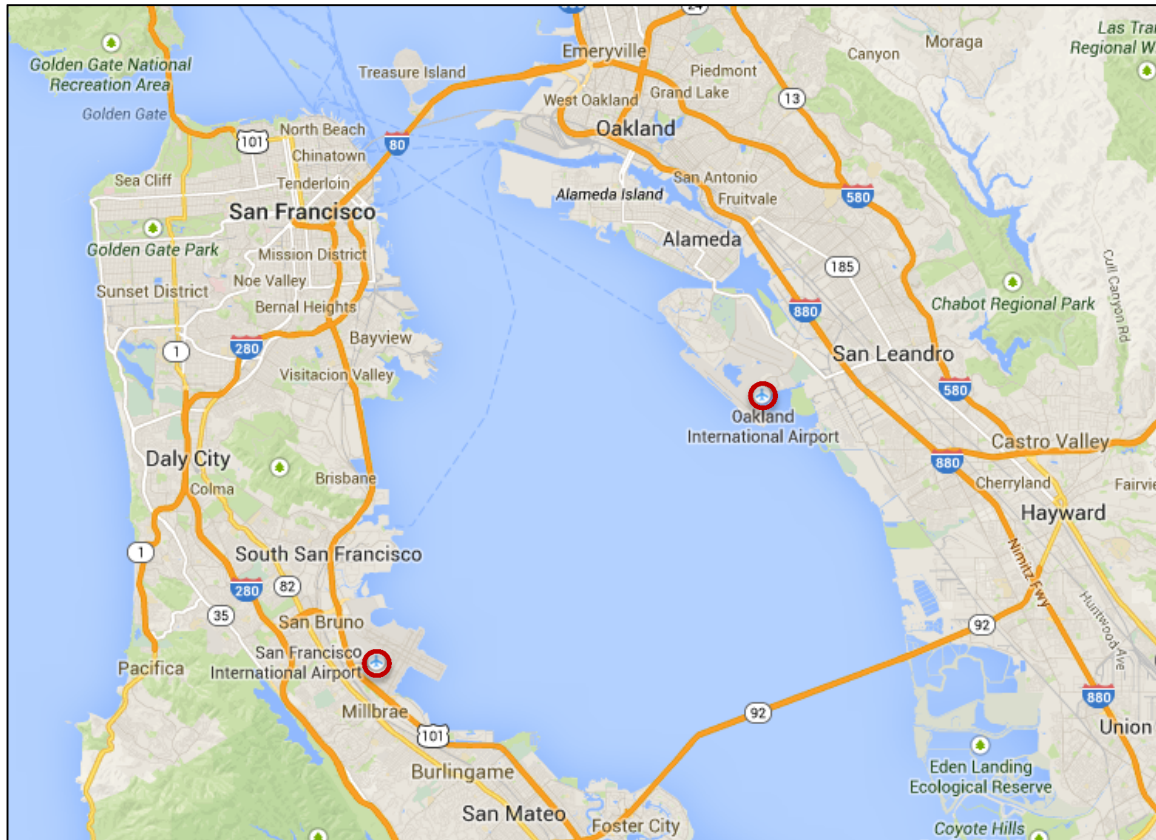


Analyzing Flight Delay Data: Fly Out of SFO or OAK Airport?

Eunkwang Joo, Ryan Jung, Julia Kosheleva-Coats, Divya Menghani

April 9, 2014



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. There is an urban myth among Bay Area business travelers that it is better to fly out of OAK than SFO because of an elevated chance of weather delays at SFO. However, this myth is contradicted by the data. According to the [Bureau of Transportation Statistics](#), SFO had an on-time departure rate of 76.74% in 2013 – ranking 21st among major US airports. Conversely, according to the [Air Travel Consumer Report](#), Oakland International Airport had an on-time departure rate of 66.0% in 2013. Our project will create a probabilistic 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?”

Project scope

The model will assume that the inputs to a user’s query are: (1) date of travel and (2) destination. This entails several assumptions that will be discussed below. 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 which the user should depart. Accordingly, this will entail the development of a classification model.

The data

We will use publicly-available data on flight arrivals and departures for major U.S. airports from the American Statistical Association (found [here](#)). The data set contains some key attributes that we will use to devise our algorithm. In particular, we expect to make significant use of TailNum, Dest, DepDelay, Time, Month, DayofMonth, DayofWeek, UniqueCarrier, FlightNum, Origin. These variables are defined as follows¹:

- Tail Number = every aircraft is required to have a unique registration number similar to a license plate number for motor vehicles
- Destination = the user's travel destination
- Departure Delay = the actual length of the delay from the scheduled departure time (we assume there is no "fungibility" in the scheduled departure time)
- Time, Month, DayofMonth, DayofWeek = time variables
- UniqueCarrier = unique airline carrier code
- FlightNum = flight number
- Origin = origin airport code

The data set covers all flights from 1987 to 2008. Because air travel delays are strongly related to weather patterns, we will exclude data from strong El Nino / La Nina years (1988, 1997, 1999, 2010)², so our data set will be free of years affected by these abnormal weather patterns.

Lastly, we would like to note that this data is being used by one of our team members, Divya Menghani, for a project in another class, Open Data.

Assumptions

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

1. Travelers only fly non-stop: While ideal in the real-world, we are not going to model flight connections and will only analyze flights between destinations in and out of SFO and OAK.
2. Users not price or time sensitive: We assume that users are business travelers who are indifferent to price differences between SFO and OAK. Additionally, we are assuming that travelers only care about the date that they fly out, and are indifferent to the time of day of their flight. This is likely not an assumption that would hold in the real world.

3. Because fog is likely to be a significant factor in delays, we are going to discretize the time of day of the departure. Therefore, we will create buckets for 7:00am-12:59pm (peak fog time at SFO³), 1:00pm-5:59pm (daylight), and 6:00pm-6:59am (night time).
4. New landing protocols were implemented by the Federal Aviation Administration (FAA) and SFO management in November 2013⁴. We will not take these protocols into account as insufficient data exists to model the effects of these new protocols.

Analysis

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

1. **Partition data into training and validation data sets.** We intend to use 80% of data for training purposes and plan to use the rest of the data for validation (minus El Nino/La Nina years).
2. **Create helper functions.** We will need to do some conversions of the data in the dataset to variables for our analysis. In particular, a) we need to search the flight histories for flights on the same date of departure with the same destination for both airports and extract the tail number, b) we want to pull out the remaining attributes discussed above, and c) we want to compute Day of Year (flight and weather patterns change by time of year; we also intend to highlight certain days as "peak travel" days around the major holidays). Taking these attributes, we will extract the relevant data for the next step.
3. **Labels for classification.** The dataset doesn't have classification labels. We will look at the length of delays for each of the flights from the previous step and add a binary variable for "delay / no delay" when the delay length is greater than 15 minutes (an industry standard definition for a flight delay). This will serve as our classification label.
4. **Classification models.** We will construct two classification models to "horse race" against each other. The first model will be a Naïve Bayesian classifier. We will also construct a logistic regression model. We expect to use the Python's *Scikit-learn* which has many of the features necessary for the Naïve Bayesian classifier and logistic regression. The output of both models will be a predicted probability of delay for each flight out of SFO and OAK to the destination on the specified date. We will take the minimum probability of delay from any single flight for the overall probability of delay at each airport. Because we assume the user is indifferent to time of departure, their best chance of avoiding delay is to take the flight with the least probability of delay from that airport on that date.
5. **Validate model.** 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. Based on the percentage of correct predictions that the model makes, we will adjust our model's parameters. The key metric here will be F1 score ($F1 = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$) meaning we want the model to have a balance between precision and recall.

6. **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 of a year (i.e. 365 days) with which is the preferred airport from which to depart for a few selected destinations. We anticipate using *matplotlib* for this task.

Team Member Roles

Eunkwang Joo – will lead up the development of the Logistic Regression model. He will also work with Divya on the validation and visualizations.

Ryan Jung – will lead up the drafting of the final report and presentation. He will also assist with the initial partitioning of the data, helper functions, and labeling decision tree.

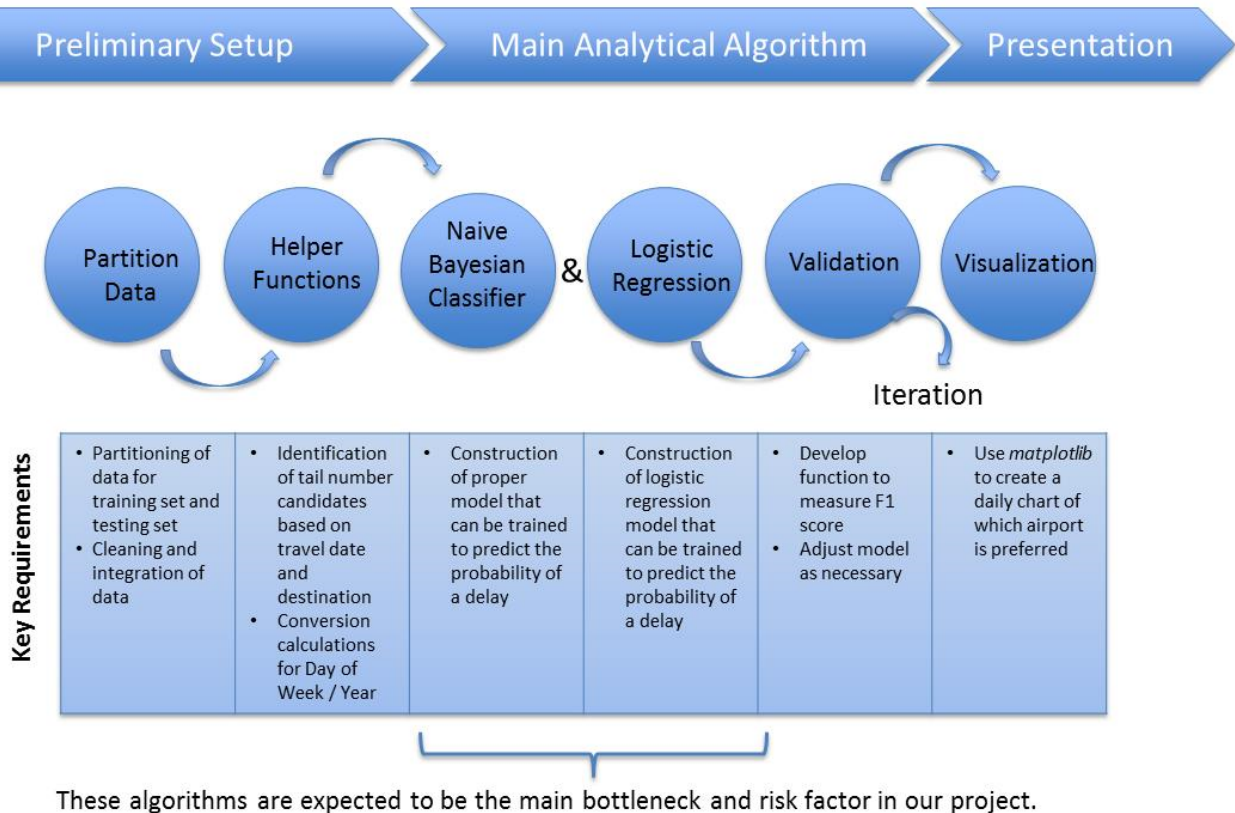
Julia Kosheleva-Coats – will lead up the initial partitioning of the data and helper functions. She will assist with creating visualizations and drafting of the final report and presentation.

Divya Menghani – will lead up the development of the Naïve Bayesian classification model. He will also work with Eunkwang on the validation and visualizations.









Overview of Project Tasks and Requirements:

Flight Delay Analysis

Using Naïve Bayesian Classifier and Logistic Regression to predict flight delays



Workflow:

Stage	Task	4/9	4/16	4/23	4/30	5/7	5/14
Proposal	• Proposal						
Preliminary Setup	• Partition / Scrub Data						
	• Mapping Functions						
Main Algorithm	• Bayesian Network						
	• Logistic Regression						
	• Validation						
Presentation	• Visualization						
Deliverables	• Write Paper						
		Proposal Due			Code and Data Due		Final Paper + Presentation

Sources:

- ¹ http://aspmhelp.faa.gov/index.php/Types_of_Delay
- ² <http://ggweather.com/enso/oni.htm>
- ³ <http://crankyflier.com/2010/10/14/san-franciscos-fog-and-runway-problems-give-the-airport-a-dubious-honor/>
- ⁴ <http://sanfrancisco.cbslocal.com/2013/11/04/new-rules-to-reduce-fog-related-delays-at-sfo/>