Prescriptive Analytics on Flight Delays

Akshay Shrinivasan, Vineel Lakshman, Jashkumar Shah
A20353163 A20356842 A20352470

Introduction:

For almost all the cities there are two airports like 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.

This model is designed for the business users who are very clear to 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(either San Francisco, Chicago or New York) from which the traveler should depart. Accordingly, this will entail the development of a classification model.

Aim of the Project:

The main goal of the project is, given a specific date and destination which airport is suggested to fly from with least delays.

Delay: We decided to set a departure delay definition threshold at 10 minutes. Thus, all flights with 10-minute or greater difference between scheduled and actual departure time will be considered as delayed.

Data:

No. of rows of data per year: 100000 rows out of which, if we plan to predict for two airports it will be around 40,000-50,000 rows per year. However, it can be applied to other cities which have two airports to travel.

No. of variables: 30

Important Variables used for classification: (10) Year, month, Day of month, Day of week, Departure time, Airline carrier, Tail number, Origin, Destination, DepDelay.

Year – the year the flight departed

Month- specific month of the year

Day of month – the day of the month

Day of week – which day in a week

DepTime – The scheduled departure time of the flight

Unique carrier – Which airlines the flight belongs to

Tail number – the number on the tail which identifies the flight of the airlines

Origin – from where the flight departs

Destination – the destination to which flight departs

DepDelay – Actual (Delayed) departure time

Preprocessing:

1. Identifying the data for two airports for all years
2. Removing cancelled flights and unnecessary variables
3. Filling the missing values and ignore the record with missing values
4. Discretize the variables which influences the other variables

Planned approaches:-

1. Classification Model: Naïve Bayes Classifier

1. The above mentioned important variables are considered for classification and they are multi class. Hence Naïve Bayes classifier will be one of the best approach in predicting the probability of delay (yes/no).
2. Total 5-year data will be considered for the analysis. (2010 -2015)
3. 5-fold cross validation will be implemented. Training on 4 years, test on 1 year

2. Logistic regression model:

In order to perform logistic regression model, categorical variables had to be transformed to binary since categorical and non-ordinal values are not meaningful inputs for LR. Following variables are considered to build a logistic regression model input

1. In order to provide a proper logistic input, we are planning to convert **Deptime** feature into categorical variable because so many factors influence the time of departure.
2. Day of Week
3. Origin and Destination
4. Unique carrier

Logistic regression is suitable for predicting with the help of few variables. It is easy to interpret and draw conclusions about the impact of model inputs based on co-effecients. There is a possibility of building a linear relationship from the above selected variables and the outcome of delay or no delay(binary).