Seminar material

Project Background:

With the reduction of operating costs in the aviation industry due to fuel efficient aircraft, the cost reduction by leveraging modern technology and the increase in household disposable income, the volume of air travel increased from 450 billion passenger-miles in 1997 to 600 billion passenger-miles in 2014. ([BTS, US Passenger Miles Table](https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_40.html))

The higher passenger traffic and the increase in the number of flights offered by airlines, means that during bad weather conditions the National Airspace System (NAS) capacity in the United States is challenged to handle the number of scheduled flights. Passengers can book flights up to one year before the departure date, which is usually when an airline publishes its flight schedule. However, the planned and published flight schedule does not account for the potential impact of weather that may occur on the day of the flight. Instead, the schedule is mainly set according to profit and market share considerations. As a result, the imbalance between flight demand and NAS capacity in the US yields flight delays. The average load factor in a flight for domestic operations in 2012 was close to 83%.

Flight delays not only cause time loss for passengers but also create multiplicative inefficiency, wreaking havoc downstream by disrupting airport runway operations and the planning of airlines. the annual cost of domestic flight delays to the US economy was estimated to be $31-40 billion in 2007 (Joint Economic Committee, US Senate 2008). Correctly predicting flight delays allows passengers to be prepared for the disruption of their journey and allows airlines to pro-actively respond to the potential causes of the flight delay to mitigate their impact.

The abundant research efforts from data scientists, researchers, companies and government agencies on airline flight delays confirms that this is an important area. In particular, the main benefits of better flight prediction are significant operational cost savings and a non-negligible improvement in quality of life for those who use air as an important mean of transport. An accurate online flight delay predictor would certainly generate a lot of interest in the world of air travel.

The goal of work done in this seminar is to use exploratory analysis and to develop machine learning models to predict airline's departure and arrival delays. Based on the literature reviews, this type of problem is actively examined by many researchers and GE even brought out a flight quest challenge with an award of $250,000 to the team who can most accurately predict flight delays

Data Input and Exploration

The flight data, also known as the on-time performance data can be downloaded from the [American Statistical Association](http://stat-computing.org/dataexpo/2009/the-data.html).

From the website, there are several datasets for years from 1987 to 2008, containing flight arrival and departure records throughout the USA. Since the datasets for each year are quite large, we chose to restrict our focus to one year 2008, which already contains close to 1 million records for the largest airports.

Processing speed is a major consideration since the Machine Learning procedures that work well on smaller datasets cause problems with the Anaconda installations on our computers.

The 2008 on-time performance data contains 7 million records. On average, the daily number of flights is 19,178. A total of 24 data elements are included in the 2008 flight database.

We checked the structure of the dataset and saw that there are several columns that correspond to 'Delays' and also 'Cancelled' flights. We eventually excluded these, since cancelled flights have no delay attributes and restricting the analysis to delayed flights means that we would miss non-delayed flights.

We needed to generate departure and arrival hour attributes from the departure time variables, in order to look at the time of day effect on delays at an appropriate granularity.

As a first exploratory analysis, we consider the observed probability of delay in minutes on the entire dataset. The most effective way is through a histogram, looking at departure and arrival delays separately.

We notice a much higher probability of short delays - actually negative, so we can consider them advances - for departure delays and a wider distribution (in minutes) for arrivals. Notice the long right-hand tails. Some flights are delayed for very long times, over two hours. On the other hand, the delays are centred around just below zero. In both cases, the mode of the distribution is less than zero, meaning most of the flights leave from gate and arrive at gate even before the published schedule time of departure and arrival.

the arrival delay distribution, compared with the departure delay distribution, leans toward left. A flight delay is defined as the time difference between the scheduled time of an event compared to the actual time of the event. Airlines usually put extra buffer time in a flight to ensure on-time arrival. Therefore, the distribution of the difference in the departure delay and arrival delay indicates that some departure delays are recovered during the flights due to the extra amount of time embedded in the flight time between two airports.

Next, we consider the impact of month of flight on the delays. For both departures and arrivals, the impact of the month of **December** is clear - the **highest** delays are in that month. On the other hand, **September, October and November** are the months with the **least** amount of delay. For the summer months in which there are holidays and more flights, **June and July** are marked by **higher** delays. Also, February posts high delay values as well. The reason for **winter**'s high delay values is probably because of **snowstorms** in the northeast of the US. Also, in **summer**, **thunderstorms** in Dallas Forth Worth (DFW) and Chicago areas can cause high delay impact to the rest of country. A snowstorm/storm may only affect operations at an airport or two. However, delay propagation, which marks as the major contributor for flight delay, can cause ripple effects on delay to downstream flight operations.

We also think that the time of day should have an impact. Normally, flight delays cumulate throughout the day through propagation, with a knock-on effect of delayed flights provoking other delays because of tight schedules and runway congestion. We plot the mean delay by hour of day in a column chart. From the plots, we see a marked **"V" shaped decline in delay** with the lowest delays in early morning hours. Both departure and arrival delays accumulate from the earlier morning hours reaching their **peaks in the evening hours**. For **departure**, the highest **mean** delay is during prime-time of **18:00 to 21:00**, and for **arrivals**, it is slightly later (the average flight duration is a few hours) and peaks at around **22:00**. The increasing of flight delay by the hours of the day is mainly caused by flight delay propagation. As a result, if a flight is delayed, the next flight has to wait for the late arrival flight to be ready before it can be operated. Hence, flight delays for both departure and arrival flights do increase over time.

Comparison of mean delays across 4 different airports