Predicting the delay of flight

1. *Introduction*

Flight Delay is a most frequent problem in air transportation that deserves an attention. This costs billions of dollars damage to the aviation industry, passengers and a big hassle to travellers. According to one of the recent survey in 2010[1], it was estimated that every year aviation industry lose 8billion dollars and also flight delay costs 17 billion dollars to passengers. Also, It is observed that only few percentage of commercial flights arrive on time. Also, a delayed flight can cause trouble to passengers by making them late to their personal scheduled events and commitments. A passenger who is delayed on a multi-plane trip could miss a connecting flight. On the other hand, for airlines, it costs in money and reputation.

Since It costs a lot to the airline industry and to the passengers, predicting the delay of flight is a critical feature to improve on the decision making process in the airlines[2]. There are many factors that contribute to this flight delay, but in most of the cases the delays are attributed to the flight carrier, time of the flight and weather conditions. There are also minor contributions due to maintenance issue of the aircraft, congestion in air traffic etc, which are unpredictable.

In this project, we modeled the flight delay with respect to various number of factors and analyzed which flights are delayed and also estimated the flight delay. The task includes identifying proper data set and choosing the proper algorithms to train and validate the data, which are described in the later section. We compared our results to the standard baseline techniques and also analyzed the prediction performance with different models.

2. *Problem statement*

Our problem statement is to predict the delay of a flight given attributes like fight carrier, source, destination of a flight, season, date, time etc. In this project, we built models to predict if the flight is delayed or not and also estimated the flight delay.

A. benchmark description

We evaluate our models against standard kaggle implementations and also against other previous studies.

3. *Approaches*

The standard way to solve any data science is to collect data and train the models with this data. Below are the standard steps we followed and later discussed the algorithms used in this approach.

Below steps will be used while solving the problem.

A. Data Collection/Data Set

Train dataset:

We use the publicly available dataset on Flight timings from 1987 to 2007 obtained from Bureau of Transportation statistics. This dataset is of size ~20*120 MB (Approx) and can be found here: http://stat-computing.org/dataexpo/2009/the-data.html.

The data is the csv format containing around 29 attributes, that includes information on airlines, airports and their location, and flights history containing their to and from location, travel time, distance, departure and arrival delay etc.

Test dataset:

We took a subsample of the data as test dataset. We picked 2008 year as our test dataset. We train our model on 1987 to 2007 data and test against it on 2008 data and evaluate the model performance.

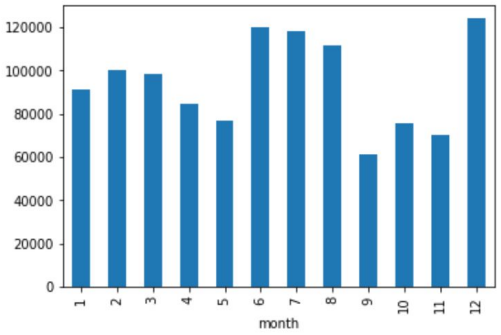
B. Data Cleaning

We have the 2007 flights data, with 7453215 records/row and each containing 29 features. Initially, we fill the missed attributes(NA) with relevant values and also removed the attributes that are not needed (for ex: FlightNum, security Delay, Late Aircraft Delay), which we thought are not important features in predicting the flight delay. We also did format on the date and time values for further processing. All this can be seen in the ipython notebook.

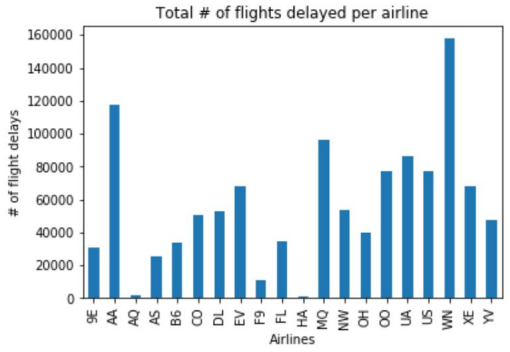
C. Data Visualization

We plot the delays against airlines, month and day of the hour to identify which of the attributes plays a major factor in the model.

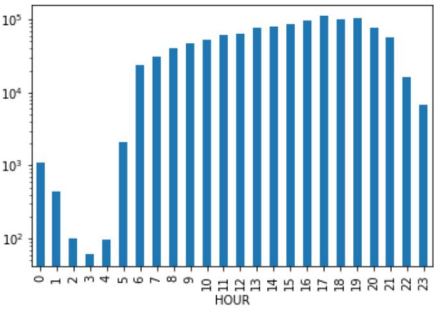
Here you can see the number of plots we drawn to correlate the features to the prediction.



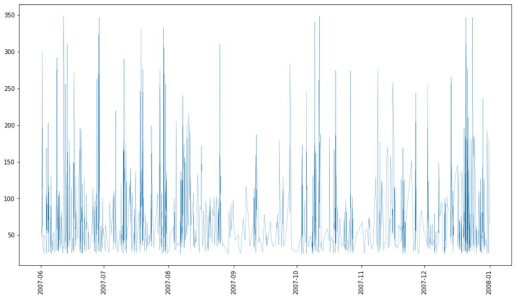
Number of Flight delays Vs Month



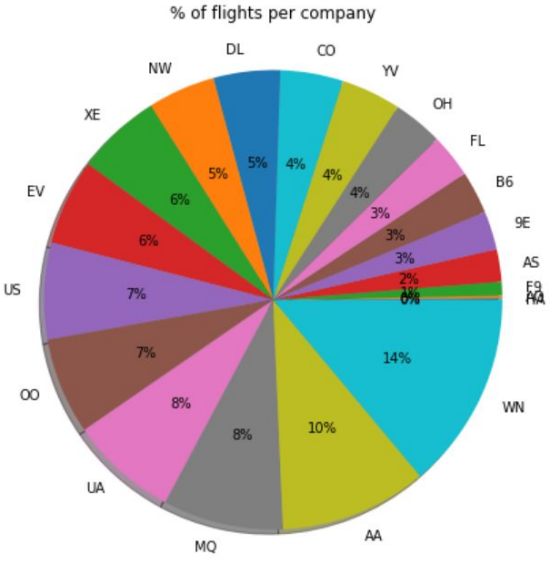
Number of Flight delays Vs Airlines

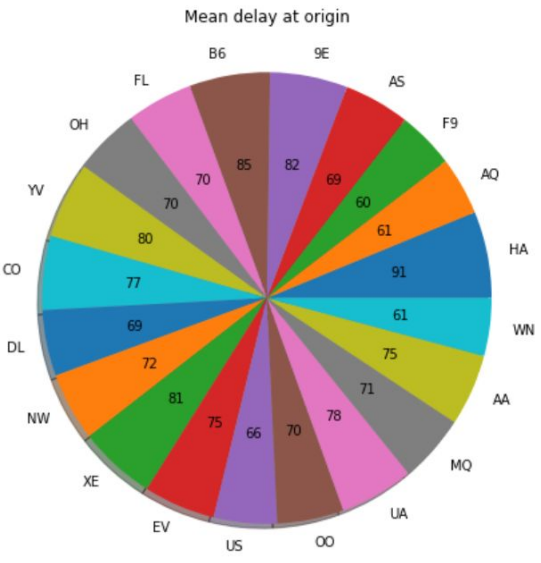


Number of Flight delays Vs Hour



Number of Flight delays Vs Year





Number of Flights Vs Each airline and their Mean delays at Origin

D. Data Transformation & preprocessing

We encode the data (one Hot encoding) and normalize if necessary, also will add additional features like holidays if required

E. Models

The first step of our project is to estimate if the flights are delayed or not. For that we used couple of classification models.

1. Logistic Regression Classifier
2. Random Forest Classifier

Next step is to estimate the flight delay. For this purpose we used couple of regression models.

1. Linear Regression model

We improved the random forest classifier by adding more attributes to the training set after encoding them using **one hot encoding**.

4. *Analysis and Comparison of different approaches*

5. *Metrics and evaluation step*

6. *Conclusion*

A. *challenges*

B. *lessons Learned*

References:

1. http://mashable.com/2014/12/10/cost-of-delayed-flights/#Nz.t4FmcQOqh

2. https://arxiv.org/pdf/1703.06118.pdf