Predicting the delay of flight

Team

- Venkata Narendra Kumar Gutta, UFID : 49195476
- Chanikya Chandra Mohan Konyala, UFID : 50053565

Problem statement

Flight Delay is a most frequent problem in air transportation that deserves an attention. This costs billions of dollars damage to the aviation industry, and problems to passengers, airlines. In this project, we are trying to come up with a regression model to predict the delay of flight given conditions like origination, destination, time, weather conditions etc.

Flight delays can occur due to myriad reasons like weather conditions, maintenance issue of the aircraft, congestion in air traffic etc. The time by which a flight is delayed varies depending upon the reason for the delay. Flight delays are an inconvenience to both passengers and airlines. 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.

Techniques to solve the problem

We are planning to use supervised learning to solve this problem by analyzing the flight history data and to predict the delay of flight by using a regression model. As of now we are considering using logistic regression and random forest classifier and then try to improve our models.

Below steps will be used while solving the problem.

Data collection	Refer the Dataset section below
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.
Data Cleaning	We will fill the missed attributes(NA) with relevant values and also remove the attributes that are not needed (for ex: security Delay, Late Aircraft Delay) also clean the date and time values
Data Transformation & preprocessing	We encode the data (one Hot encoding) and normalize if necessary, also will add additional features like holidays if required.

Model Building	Regression and Random forest classifier and improve on it.

Dataset

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)

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 train the model on 1987 to 2007 and test against it on 2008 data.

Tools Used

Python, Matplotlib, PySpark, Mapreduce, Scikit-learn, AWS

Evaluation Metric

Below are the metrics to evaluate the performance of the models on the given data.

- Confusion matrix
- Precision, Recall
- Accuracy

All the above metrics will be evaluated against the Test dataset and we calculate RMSE on the test dataset.

Baseline Techniques

Linear regression is one of the base methods that are proposed to estimate the flight delay. These models gives an accuracy of 60% for ~1 yr flight records. In this project, we define the baseline for predicting the delay as the average of all delay values for a given flight. We will compare our results against the baseline and then improve that by using more data by further enriching the dataset and also adding other valuable data points like weather, air traffic if time permits.

Impact and Innovation

Flight delays are the most common but toughest to predict in commercial aviation industry. There are multiple factors involved that causes the delay that complicates the prediction. Improving this flight delay estimation can yield lot of benefits to passenger and airlines.

We solve this problem by considering dataset that ranges ~20 years. Also, we consider adding additional features to the model, for example weather data to improve our accuracy of the model.