DELTA AIRLINES INC FLIGHT DELAY

PROJECT ABSTRACT

According to the given train data sets of Januarys 2016 - January 2017 & test data set of January 2018. We have understood by reading the data and **concatenate** and formed 4 training data into one data frame & 2 test data into 1 data frame. And found that a less amount of data is missing, so we have treated it with **forward-fill method**. And remove unnecessary column which have no importance in time delay. And mapped date into new feature i.e. year, month and days of the week. We have **Exploratory data analysis** by using the **correlation matrix** to show relation between the variable & **pair-plot** which gave us the depth information about correlation to understand the data better. We have plotted **visualization** upon maximum delay in the destination airport and found that ALT is the busiest airport and have maximum delay in the data set by using **bar plot** & **box plot**. By plotting a **line-plot** on taxi out time and found it to be the maximum cause of flight delay at the ALT airport.

Then we have use **Evolution matrix** to find the error where we have used **mean square error**, **root mean square error** and **mean absolute error**. We have used various Regression models to predict 2018 data as -

Linear Regression accuracy is 89.45%
Decision Tree Model accuracy is 89%
Cross Entropy Model accuracy is 90%
Random Forest Model accuracy is 92%

Random Forest Model which have shown maximum accuracy of 92% upon delay carrier in all the airports. To which we have plotted on a graph using **Scatter plot**.

These datasets were two-fold train and test. The first part deal with an exploration of the dataset, with the aim of understanding some properties of the delays registered by flights. This exploration gave us the occasion of using various visualization tools offered by python. The second part of the notebook consisted in the elaboration of a model aimed at predicting flight delays. For that purpose, we used Machine Learning regression modules and showed the importance of regularization techniques. In fact, Random Forest gives the best accuracy then linear regression, cross entropy, decision tree models for prediction of delays.

Team Name - Codehax41

Team members - i)Ramsundar Mahato
ii)Abhijeet Sanyal
iii)Mohammad Azam