

Project Report

ON

Flight Price Prediction

Submitted By

Nimesh Kumar Roy

**Problem Statement -**

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on -

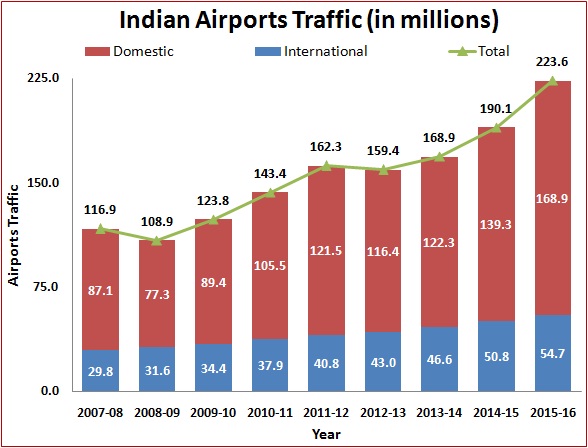
1. Time of purchase patterns (making sure last-minute purchases are expensive)

2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive

purchases).

According to a report, India’s civil aviation industry is on a high-growth trajectory. India aims to become the **third-largest aviation market by 2020**and the largest by 2030. Indian domestic air traffic is expected to cross **100 million passengers by FY2017**, compared to 81 million passengers in 2015, as per Centre for Asia Pacific Aviation (CAPA).

According to Google Trends, the search term - **"Cheap Air Tickets"** is most searched in India. Moreover, as the middle-class of India is exposed to air travel, consumers hunting for cheap prices increases.



The project is basically **machine learning & statistic intensive**. We used Python for the implementation of the models & automation.

##### Automated Script to Collect Data

For any prediction/classification problem, we need historical data to work with. In this project, past flight prices for each route collected on a daily basis is needed. Manually collecting data is not efficient and thus a python script was run on a remote server which collected data from make my trip.com.

##### Cleaning & Preparing Data

##### After we have the data, we need to clean & prepare the data according to the model's requirements. In any machine learning problem, this is the step that is the most important and the most time consuming. We used various statistical techniques & logics and implemented them using python.

##### Analysing & Building Models

##### Data preparation is followed by analysing the data, uncovering hidden trends and then applying various predictive models on the training set.

##### Merging Models & Accuracy Calculation

##### Having built various models, we now have to test the models on our testing set and come up with the most suitable metric to calculate the accuracy. Moreover, many a times, merging models and predicting a cummulative target variable proves to be more accurate.

The basic structure of the script successfully extracts information from the Makemytrip.com website and outputs a excel data file. Now an important aspect is to decide the parameters that might be needed for the flight prediction algorithm.

Makemytrip returns numerous variables for each flight returned. However not all are required and thus we selected the following -

1. FLIGHT.
2. FLIGHT NO.
3. SOURCE.
4. DESTINATION.
5. DEPARTURE.
6. ARRIVAL.
7. STOP.
8. FARE.

The kind of data that we collected from the python script was very raw and needed a lot of work. For instance, the price was a character type and not an integer.

##### Duplicates

##### There were not many, but a few repetitions in the data collected.

##### Departure & Arrival

The departure and arrival time is in normal hour and minute format. So we seperate them in dep\_hour, dep\_min and arrival\_hour,arrival\_min .

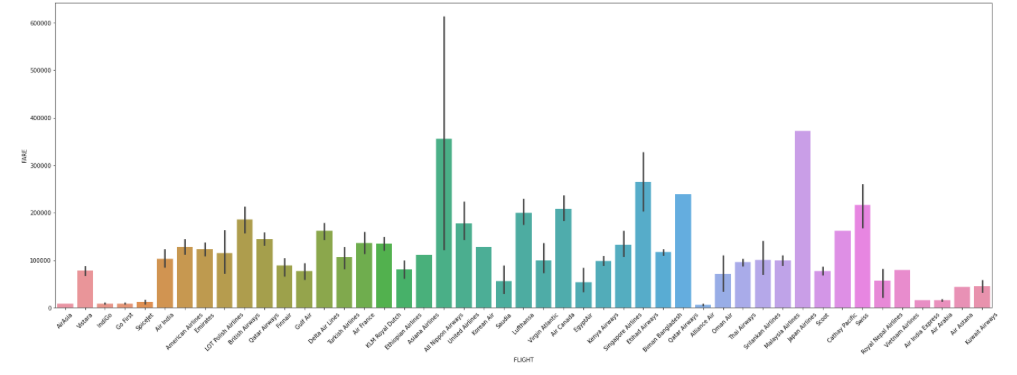
**Drop the Unnecessry Feature**

We drop the “FLIGHTNO.” feature from dataset because it is not a deciding feature for target column.

**Statistical Analysis** -

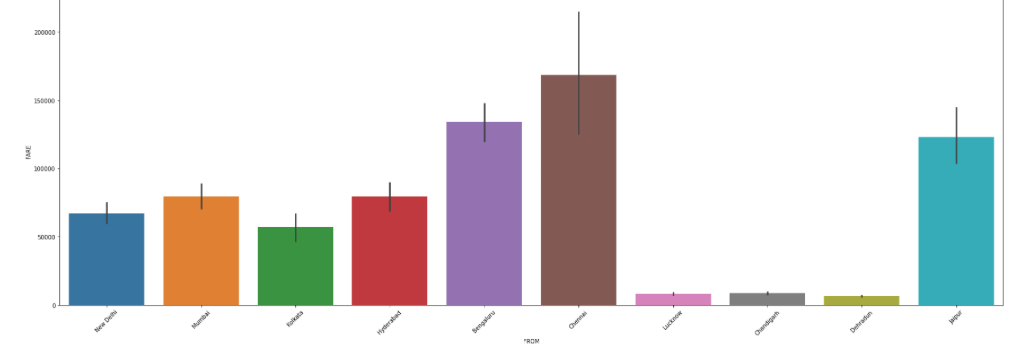
Different airline companies have different facilities and according to that they charge the fare from customer. When we plot the bar graph between airline company and price we get that

Asiana Airline and Malayasian Airline both have maximum fare on the other hand “Spice Jet”, “Indigo”, “ Go First”, “Air Asia”, “Air India Express” have almost same minimum fare.



When it comes to source, it is also play an important role to decide the fare. When we plot the bar graph between source and fare we get that from Chennai airline companies charges are hire, while from Jaipur,Bengaluru,Hydrabad,New Delhi,Mumbai,

Kolkata the fare are moderate and from cities like Lucknow,Dehradun,Chandigarh the fares are quite low.



**MODEL SELECTION** -

The problem is based on regression so we use “Mean Absolute Error (MAE)” , “Mean Squared Error (MSE)” and “Root Mean Squared Error (RMSE)”.

Algorithms we used to make a best fit model-

Linear Regression

KNeighbor Regressor

Decision Tree Regressor

Random Forest Regressor

XGBoost Regressor.

We use hyperperameter tunnig with some of these model.

From these 5 algorithms Decision tree,Random Forest,XGBoost are working well , but the training score and Test Score are good with XGBoost Regressor it is about 97% and 82%.

So we save the XGBoost as our model.

**Conclusion** -

According to data set and Exploratory Data Analysis we can say that :

The fares are quite low on domestic airlines like indigo,goair,airindia,spicejet.

The fares of international flights like asiana airlines and malayasian airlines are high.

The fares from Chennai to other destination are high and from New Delhi, Mumbai,Bengaluru,Hydrabad,Jaipur fares are moderate,

And from Lucknow, Chandigarh,Dehradun fares are low.