

Flight Price Prediction in India

Submitted by:

Mohit Kumar Tomar

**ACKNOWLEDGMENT**

I take great pleasure to thank and acknowledge the help provided by Flip Robo Technologies. I extend whole hearted thanks to Mr. Mohd Kashif who become my Mentor and with whom I worked and learned a lot and for enlightening me with her knowledge and experience to grow with the corporate working. Her guidance at every stage of the Project enabled me to successfully complete this Project which otherwise would not have been possible without her consent encouragement and motivation. Without the support it was not possible for me to complete the report with fullest endeavour.

**Abstract**

Optimal timing for airline ticket purchasing from the consumer’s perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using historical data and also to suggest the best time to buy a flight ticket.

For this project, we have collected data from 4 routes across India and extensively used for the analysis.The project implements the validations or

Contradictions towards myths regarding the airline industry, a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done so. A Customized model which included a combination of ensemble and statistical models have been implemented with a best accuracy.These models have led to significant savings and produced average positive savings on each transaction. Remarkably, the trends of the prices are highly sensitive to the route, month of departure, day of departure, time of departure, whether the day of departure is a holiday and airline carrier. Highly competitive routes like most business routes (tier 1 to tier 1 cities like Mumbai-Delhi) had a non-decreasing trend where prices increased as days to departure decreased. Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum. With a high probability (about 20-25%) that a person has to wait to buy a ticket, the scope of the project can be extensively extended across the various routes to make significant savings on the purchase of flight prices across the Indian Domestic Airline market.

**INTRODUCTION**

* Business Problem Framing

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, and it will be a different story. To solve this problem, we have scrapped the prices of flight tickets for various airlines between the 23 June to 30 June and between various cities, using which we aim to build a model which predicts the prices of the flights using various input features.

* Conceptual Background of the Domain Problem

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics known as "revenue management" or "yield management". The

Cheapest available ticket for a given date gets more or 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)

So, if we could inform the travellers with the optimal time to buy their flight tickets based on the historic data and also show them various trends in the airline industry we could help them save money on their travels. This would be a practical implementation of a data analysis, statistics and machine learning techniques to solve a daily problem faced by travellers.

* Project Goals

The objectives of the project can broadly be laid down by the following questions -

1. Flight Trends

Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time?

2. Best Time to Buy

What is the best time to buy so that the consumer can save the most by taking the least risk? So should a passenger wait to buy his ticket, or should he buy as early as possible?

3. Verifying Myths

Does price increase as we get near to departure date? Is Indigo cheaper than Air India? Are morning flights expensive?

* Motivation for the Problem Undertaken

From customer point of view, Allow them to find the best and cheap fare price flight.

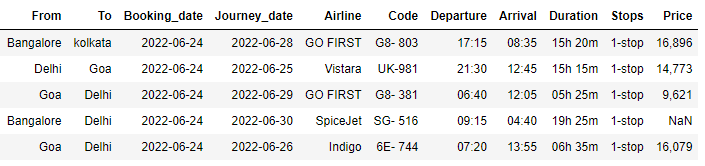
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

* Data Sources and their formats

For any prediction/classification problem, we need data to work with. In this project, flight prices for each route needs to be collected on a daily basis. Manually collecting data daily is not efficient and thus a python library Selenium was run on various leading websites like Yatra.com, Cleartrip.com, makemytrip.com which collected prices daily at specific time.



11 Features have been scrapped

1. From: from where the flight will fly, Source
2. To: Where the flight will go, Destination
3. Booking Date: When the flight ticket has been booked
4. Journey Date: Travel date
5. Airline: The name of the Airline
6. Code: Specific code of the Airline plane
7. Departure: At what time flight take off
8. Arrival: At what time flight will reach at destination
9. Duration: total time taken by flight/ travel time
10. Stops: How many halts/stops flight has taken in between journey
11. Price: What is the price of flight

* Data Preprocessing Done

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 by using built-in Python packages.

* Data Inputs- Logic- Output Relationships

Afterwards when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict prices and values of flight price. In this study three models are proposed to be built using Logistic Regression model technique, Random Forest Regressor and Bagging Regressor. Firstly, the data was portioned into section for training and the other part for testing, portioning percentage can be tested with different ratios to analyse different results. All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). From all, the Random Forest Regressor outperformed.

* Hardware and Software Requirements and Tools Used
* Hardware:
* CPU: Intel(R) Core(TM) i5-4460 CPU @ 3.20GHz
* Ram: 20GB
* Monitor= 32”inch Samsung
* Software: Latest Anaconda for Jupyter
* Python Libraries:
* Pandas , Numpy, seaborn, matplotlib, scikit-learn

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

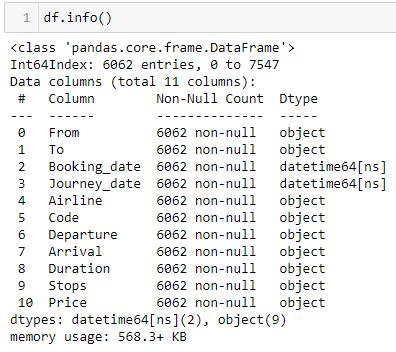
Pre-processing is a Data Mining technique that involves converting raw data into a comprehensible format. There is often a lack of specific activity or trend data, and many inaccurate facts are included in real-world data. Consequently, this may result in poor-quality data collection, and, in turn, poor-quality models constructed from the data. Such problems can be resolved by pre-processing the data.

Pre-processing in Machine Learning is the process of modifying, or encoding, data so that the machine can parse it more easily. Thus, the algorithm can now properly interpret the data.

In this project, following steps are preformed to pre-process the dataset.

1. Dataset collection: we have collected the dataset from 3 leading websites which deals with flight ticket sell-buy.
2. Pre-Processing: scrapped data was very messy. We have to perform many pre-processing on that dataset.

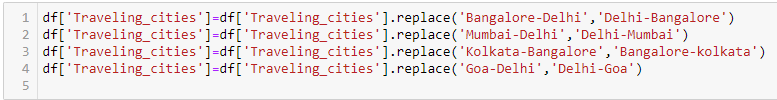
As the data is scrapped from the websites, all features were in Object form, even the integer features also were in object form.



**From** feature having the source city name from where the flight will take off. This feature is required because not all cities have equal flights running and same prices.

**To** feature having the destination city where the flight will land.

**Traveling\_cities**  this feature has been extracted from the From and To feature to have a combinations of cities.



**Booking\_date**  this is the date when the booking has been done online. This feature will be used to calculate how much advance booking has been made.

We have to extract day only because all dates from the same year and same month

**Journey\_date** this is basically the travel date.

**Early\_booking** we have extracted this feature from the booking date and journey date to get estimate if we book ticket many days before what will be the flow of prices.



**Airline** is the airline company name as we have many leading Airlines operating in India and outside India.

It had many format issues. Have to clean unnecessary space and merge look alike airline names



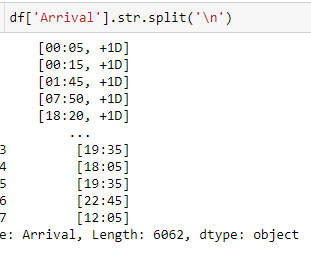
**Code** This features have specific code for each flight. We have to remove unnecessary spaces between code.



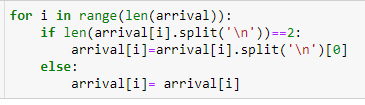
**Departure** Time when flight takes off. We extract hour of the day so that later on we can convert into day time zone



**Arrival** Time when flights land. Few flights land next day after departure so we had +1D with time. We have to remove 1D to get the hour of the day when flight land.



Code to remove +1D from the arrical time

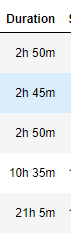




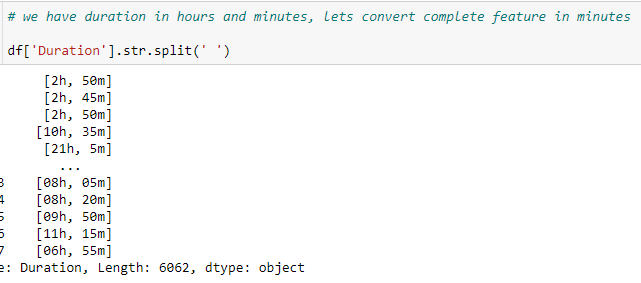
After extracting landing hour, we categorised the day zone for landing like Early morning, afternoon, late night because this affects the price

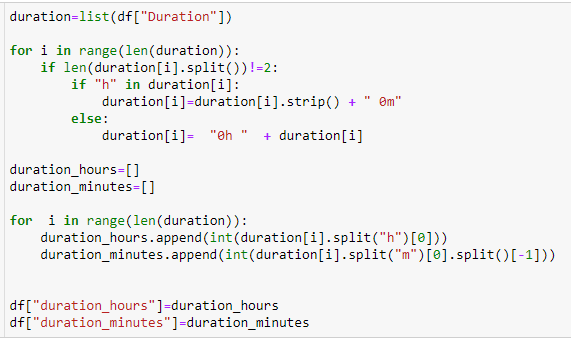


**Duration**  feature is the total duration of flight to reach from source to destination. Duration was in Object format.

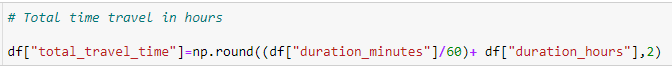


We have to separate hour and minutes from the feature and after than will calculate the total hour taken by flight.

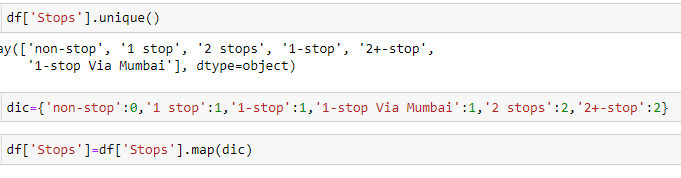




**Total Travel Time**  from the duration feature we have extracted total time taken in hours

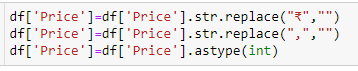


**Stops** is the total number of halts inbetween the travel. Here we replaced categorical string into categorical numerical



**Price** is the target feature which tells the price of the flight.

We have to remove comma and rupee symbol along with data type change into float.



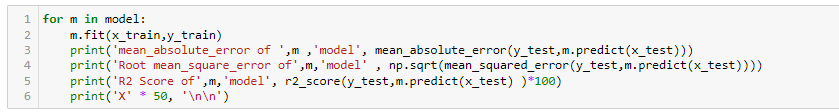
* Testing of Identified Approaches (Algorithms)

We have used several available Regressor Algorithms



We have calculated RMSA for all the available Algos then decide to which one to proceed for model building

* Run and Evaluate selected models



We have calculated MAE, RMSE and R2 score for all ML Algorithms

* Key Metrics for success in solving problem under consideration

The regression model can be evaluated on following parameters:

1. Mean Square Error (MSE):

MSE is the single value that provides information about goodness of regression line. Smaller the MSE value, better the fit because smaller value implies smaller magnitude of errors. 𝑴𝑺𝑬= 𝟏𝑵Σ|𝒚𝒊−𝒚|𝟐𝑵𝒊=𝟏

Equation 3 MSE equation

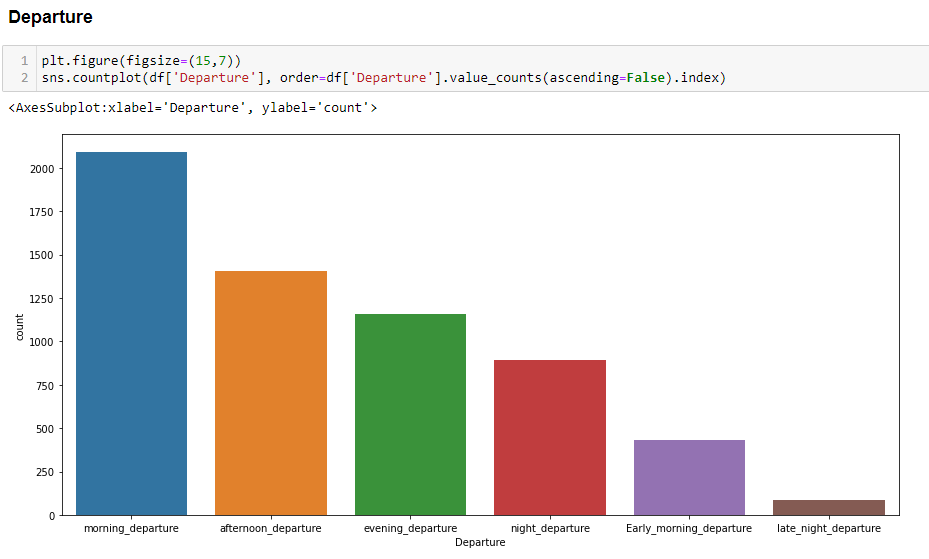
2. Root Mean Square Error (RMSE):

RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of average squared difference between prediction and actual observation.

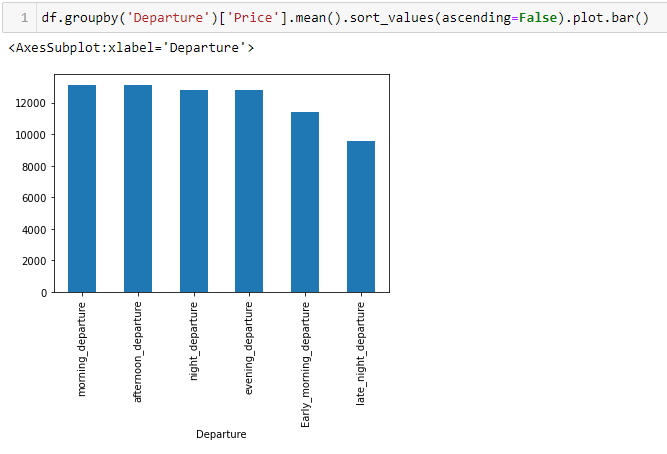
3. Mean Absolute Error (MAE):

This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset. 𝑴𝑨𝑬= 𝟏𝑵Σ|𝒚𝒊−𝒚|

* Visualizations



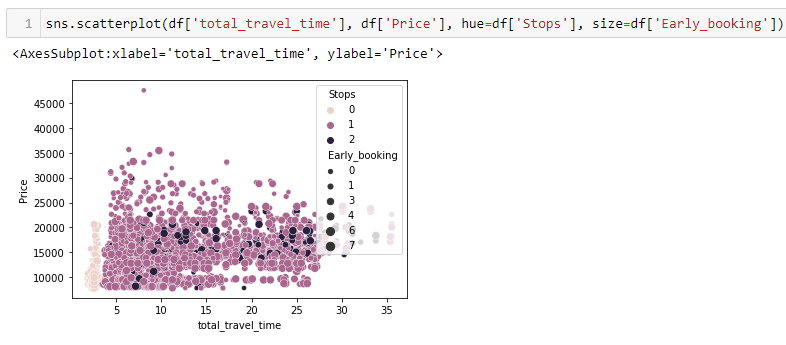
Most Departure flights take off in Morning and least in late night



Morning Departure and Afternoon Departure have high price while late night departure flights have lest price

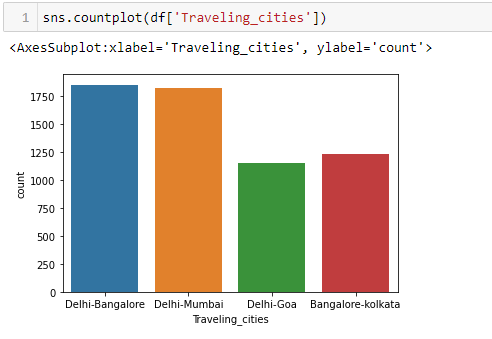


Most flights arrives in morning and afternoon as a result have high price

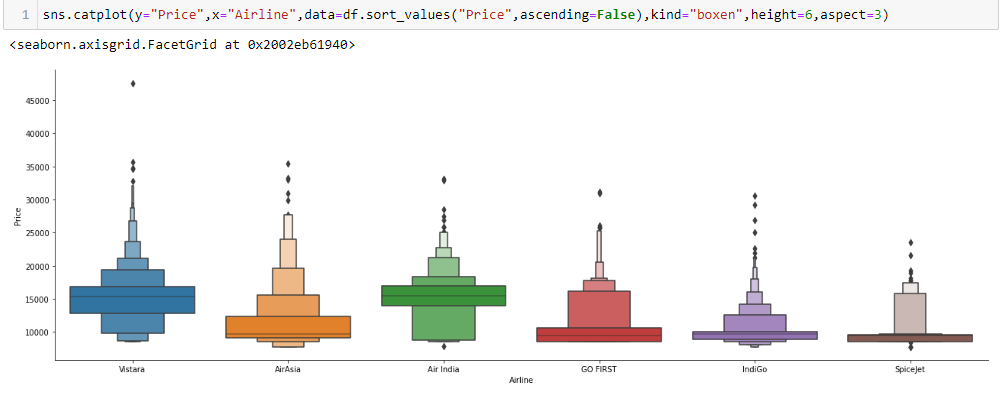


Those routes have 1 or more stays in between take more time to complete journey. While direct flights take lest time to complete journey.

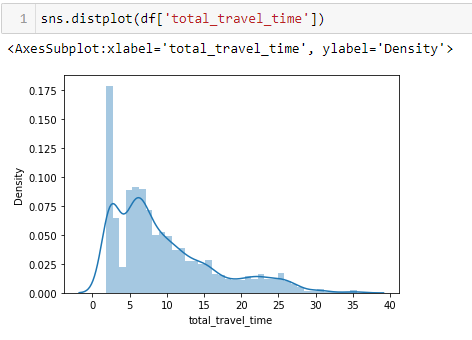
Early booking with 1 more stops are less expensive. While on same day travel is expensive



Our dataset have 4 cities interchange travel history. More flights in between Delhi – Banglore and Delhi-Mumbai







Total time travel data is right skewed because few flights have a long halts.

**CONCLUSION**

* Key Findings and Conclusions of the Study

From our detailed analysis of each of the 4 routes, we can determine the following

1. Flight prices almost always remain constant or increase between the major cities
2. Tourist routes and routes that offer services involving Tier-2 cities of the country have uneven trends related to the increase and decrease of airline ticket prices.
3. The model in the worst case almost breaks even with the profits and losses, and most case saves an average of about Rs. 200 per transaction when predicting to wait.
4. Routes with data collected over the longer duration of time tend to facilitate with much more accurate predictions in the model and thus lead to higher average savings.

We were successfully able to analyse each route and generalize the entire project based in terms of the sector to which the route belonged, and classified them into three major

subsections - Business Routes, Tourist Routes and Tier-2 Routes.

We have also successfully busted some of the typical myths and misconceptions related to the airline industry and backed them up with data and analysis.

Finally, we have created a User Interface for the entire process of buying an airline ticket and given a proof of our predictions based on the previous trends with our prediction. Thus leaving it as a battle between ‘ The risk appetite of the user ’ vs ‘ Our understanding of the airline industry ’.

* Limitations of this work and Scope for Future Work

1. More routes can be added and the same analysis can be expanded to major airports and travel routes in India.
2. The analysis can be done by increasing the data points and increasing the historical data used.

That will train the model better giving better accuracies and more savings.

1. More rules can be added in the Rule based learning based on our understanding of the industry, also incorporating the offer periods given by the airlines.
2. Developing a more user friendly interface for various routes giving more flexibility to the users.