

FLIGHT PRICE PREDICTION

Submitted by:

SURJEET SINGH

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to **FlipRobo** who gave me the golden opportunity to do this wonderful project on the topic **Flight Price Prediction** project. The data used in this project was scraped from the website of **yatra** and **skyscanner**.

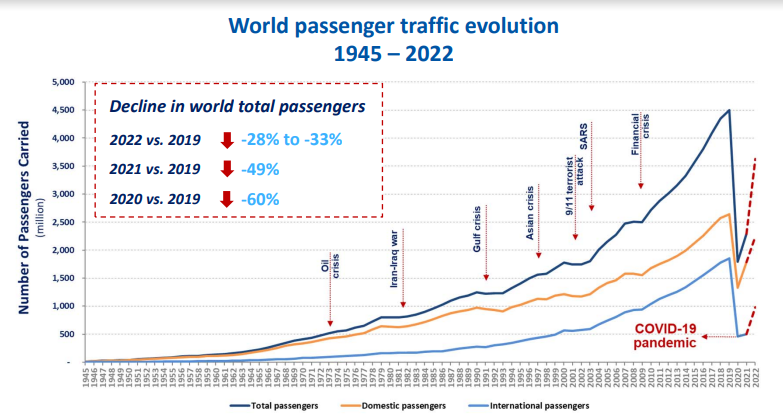
**INTRODUCTION**

* Business Problem Framing

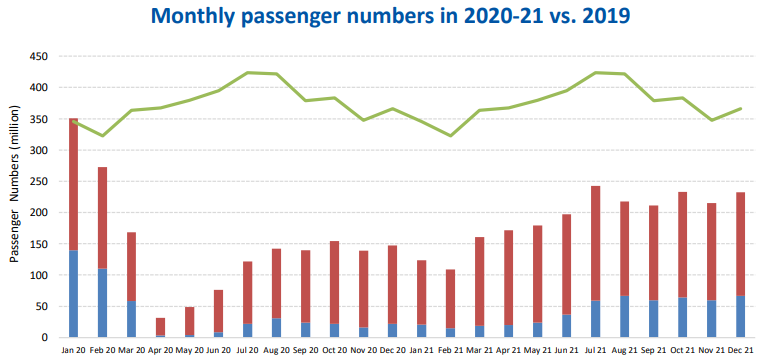
In the new era, things are going on high speed and this speed is increasing day by day. Similarly, people nowadays prefer travelling in shorter time. But due to COVID-19, the paying capacity of customer is reduced. In that scenario, it becomes important to set the price of the tickets in such a way that flights run with its maximum capacity. It is seen in the last few weeks that some of the flights are running with very less capacity which may impact the companies badly. We will design a machine learning model to place the tickets price in an optimized way.

* Review of Literature

Several explorations have been done during this project. I will present some of them here.



As it is clearly visible in the graph that air traffic has been reduce significantly after the pandemic. But it is getting its momentum back slowly. As seen in the comparison study of recent years with the 2019, there is a huge gap between the air traffic.



As some of economists predicted different types of recovery of economy. This recovery of the economy will directly impact the air traffic which will help the companies to recover their losses and retain their business.

* Motivation for the Problem Undertaken

Objective of this project is to help the ticket vending companies and airlines in placing the best ticket price of their flights. This machine learning model will also take care of the seasonality exist in the price of air tickets. There are some seasons when people move around more. For example, the holidays seasons such as Christmas, Diwali, New year. It is a time for companies to charge better price from the customers.

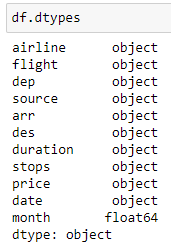
**Analytical Problem Framing**

* Data Sources and their formats

The data collection for this project has been done by web scarping from yatra.com and skyscanner. This data contains several columns containing the information regarding the flight such as airline, source, destination, departure time, arrival time, duration, number of stops, date of journey, etc.

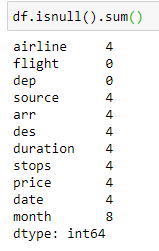
* Data Preprocessing Done

We have 4988 rows and 11 columns in the dataframe. First of all we will check the datatype of each column so that we can convert that dataframe into an appropriate datatype.



The price column converted into float and arrival time and departure time is split into two columns containing departure hour and minute and arrival hour and minute.

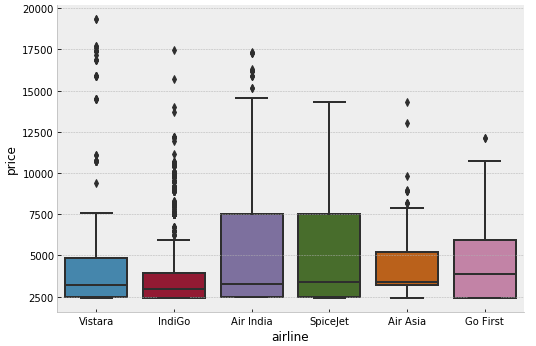
We have checked the missing values for each column. There were maximum 4 missing values in the dataframe. It has been noticed that most of the missing values belongs to only 8 rows so we have remove those 8 rows.

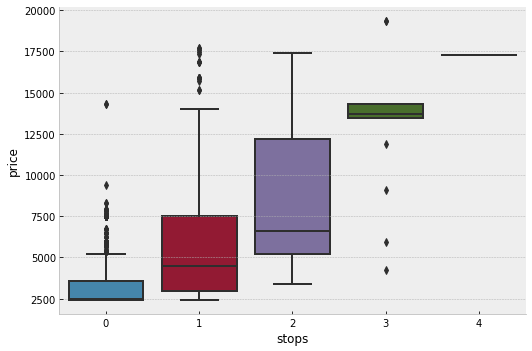


There were some rows in which columns were jumbled. It was difficult to align those columns so we have removed those rows. One of such example shown below.



* Data Inputs- Logic- Output Relationships

All the data has been converted into numbers. For example, the name of the city is Delhi which is not acceptable by the machine learning model. We have changed that column to numbers by using one hot encoding. It is important to include the airlines as well because some of the airlines such as Vistara have business flights which are very expensive while some of the airlines have only cheap flights such as Go First. It is also noticed that number of stops impacts the price of the flight. More number of stops leads to more price.



* Hardware and Software Requirements and Tools Used
* All the codes are written in python 3. Jupyter notebook used to create all visualizations, training, and testing of the machine learning models. Device used in this project : hp laptop with intel i5 process and 4 GB RAM.

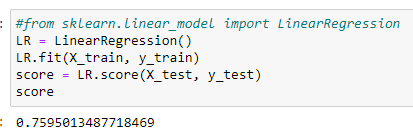
**Model/s Development and Evaluation**

* Testing of Identified Approaches (Algorithms)

It is a regression problem and we will use linear regression, decision tree, random forest and gradient boost algorithms to solve this problem.

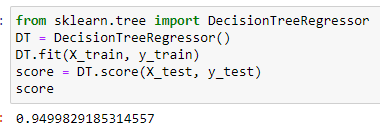
* Run and Evaluate selected models

1. Linear Regression



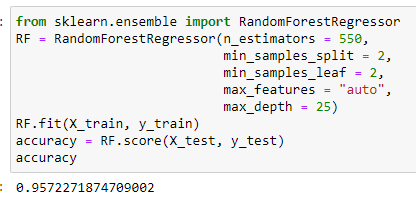
Linear regression works best when there is some linear relation between the variables and target. We have applied linear regression to the data and it gave an accuracy of 76%.

1. Decision Tree



Decision tree is machine learning algorithm that distributes the data in the form of some buckets and then gives us prediction on the basis of those buckets. We can also call these buckets bins. The accuracy of decision tree is found to be 95% which is way better than linear regression model.

1. Random Forest



Random forest is a tree based algorithms which work on the basis of weak learners. It takes the average of weak learners and return a better prediction based on number of weak learners. The parameters shown in the snapshot has been determined using random search. The random search is an approach to fine tune the hyper-parameters of a machine learning algorithms.

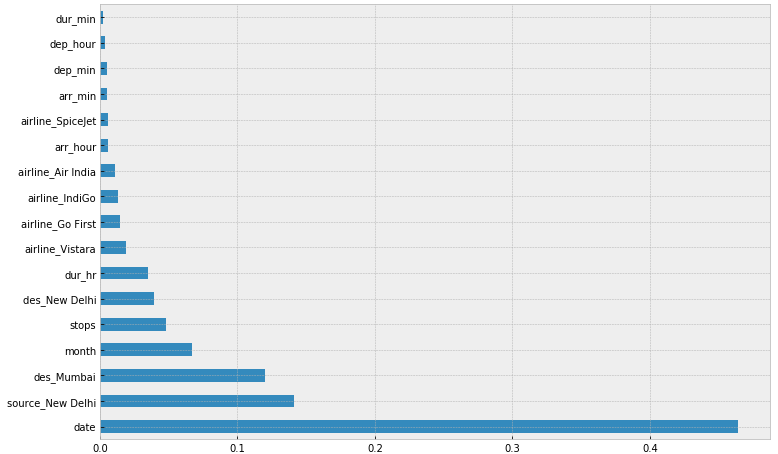
* Key Metrics for success in solving problem under consideration

Key metrics used to evaluate the model. Since it is a regression problem, mean squared error or score can be used as a key metrics to evaluate the model. In this project we have used score to evaluate the model. We have also used crossed validation for each model to confirm the accuracy of the model.

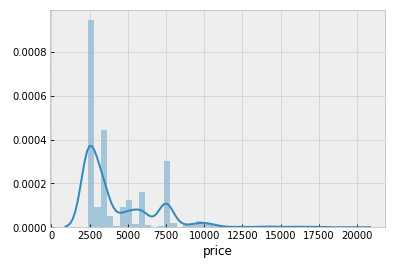
* Visualizations

I have already discussed some of the plots in some of the above sections. I will discuss the remaining plots in this section.

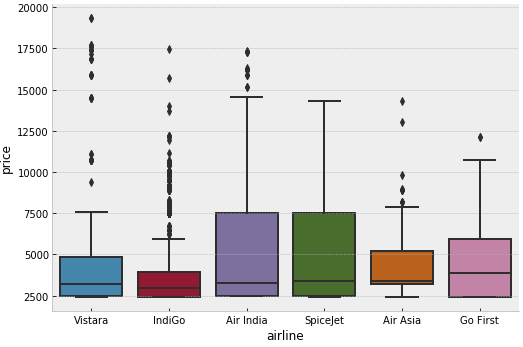
We have used extra tree algorithm to determine the feature importance of the variables. As shown in the above graph that date is an important feature to impact the price of the flight ticket. The customer will get cheaper ticket in the early booking. The number of vacant seats decreases with time so the price increases. Customer will have to pay relatively higher charges if a day before boarding.



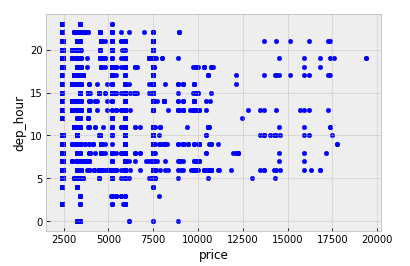
The distribution of the price is shown below. It is observed that the most of the flights are below Rs. 12500 and maximum number flights are at the price of Rs. 2500.



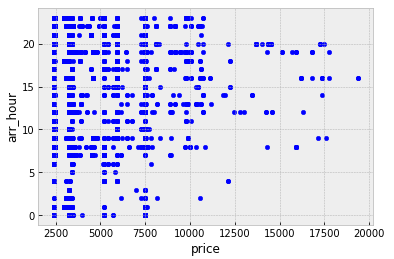
Some of the airlines have huge range of price as they offer some expensive flights or luxury travels such as Vistara. Their price begins from Rs. 2500 and reaches upto Rs. 19000.



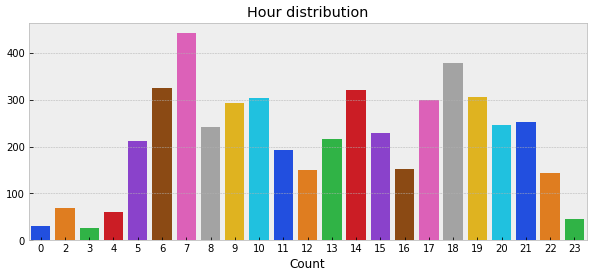
The price of the flight ticket also depends on the departure hour. The flights having departure hour in the night has low price because the number of passengers usually less during that hour. While during the day time the price of flight tickets is relatively higher than the night hour.



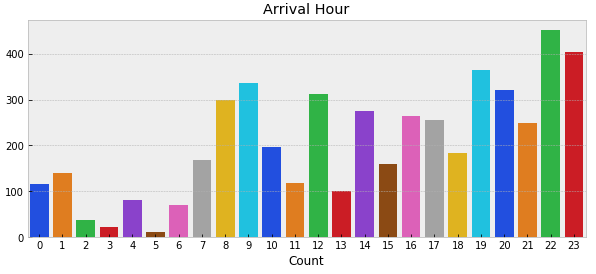
Similar pattern have been observed for the arrival hour also. The number of flight during the night time is less and price is also less.



It is clearly visible in the above graph that there are three peeks during the day. These peeks are at 7:00, 14:00 and 18:00. These times are belongs to early morning, mid of the day and evening. The number of flights is minimum in the midnight. Further we will also check the distribution of price with respect to the hour distribution.



There are very few flights having arrival time in the early morning. Maximum number of flights has arrival time in night or late night.



* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The customer should book the ticket as soon as possible because the price of the ticket increases near the boarding day. The flights having less number of stops or having zero stops have cheaper price. The flights during night must have relatively less price than the day.

* Limitations of this work and Scope for Future Work

The data collection for this project has been done only for two months. We can scrap the data for some years and then we can use that data to train the model. That model will give us better prediction because it will also learn the seasonality on the based on the year.