

**Project on**

**Flight Price Prediction**

Submitted by:

Arpita Rai

Int 33

**ACKNOWLEDGMENT**

I would like to thank Flip Robo Technologies, for giving me this opportunity to work on this project. I got to learn more from this project about Data Scraping, and practical implementations of using machine learning modules.

I take this opportunity to express my gratitude and regards to my mentor Mr. Shwetank Mishra for his guidance, monitoring and constant encouragement by giving new projects. The help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

Lastly, I thank almighty, my parents, brother, sister and friends for their constant encouragement without which this assignment would not be possible.

**INTRODUCTION**

* Business Problem Framing

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics which they call "revenue management" or "yield management". 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)

* Conceptual Background of the Domain Problem

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. 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.

* Review of Literature

The airline implements dynamic pricing for the flight ticket. According to the survey, flight ticket prices change during the morning and evening time of the day. Also, it changes with the holidays or festival season. There are several different factors on which the price of the flight ticket depends. The seller has information about all the factors, but buyers are able to access limited information only which is not enough to predict the airfare prices. Considering the features such as departure time, the number of days left for departure and time of the day it will give the best time to buy the ticket. The purpose of the paper is to study the factors which influence the fluctuations in the airfare prices and how they are related to the change in the prices. Then using this information, build a system that can help buyers whether to buy a ticket or not.

Utilizing AI models to acquire the greatest presentation to get the least cost of aircraft ticket buying, having 85.3% precision. I contemplated the exhibition of SGD Regression, Decision Tree Regression, K-Neighbors Regression, Random Forest Regression and Gradient Boosting Regression models in anticipating air ticket costs.

The data was collected from major travel journey booking website yatra.com. Additional data were also collected and are used to check the comparisons of the performances of the final model.

* Motivation for the Problem Undertaken

To Scrape the data from a website (I have scraped the data from yatra.com)

Then build a machine learning model to predict the price of the flights.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The dataset is collected from Yatra.com using Webscraping through Selenium.

To analyze the data, there are many techniques.

The regression models are used to examine relationships between variables.

The most traditional regression models are

* 1. decision tree regressor,
  2. random forest regressor
  3. gradient boosting regressor
  4. KNN-Neighbors.
  5. Sgd regressor
* Data Sources and their formats

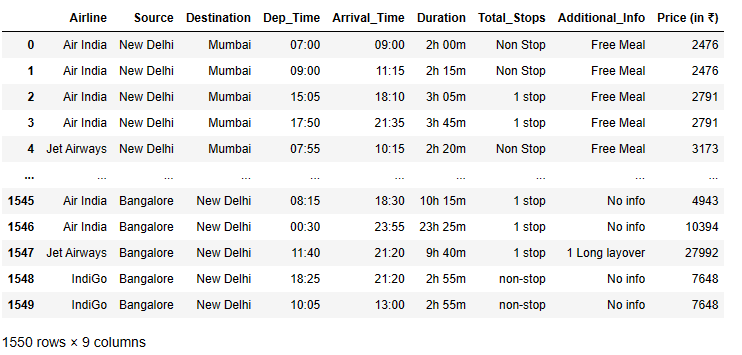
Data Collection is one of the most important aspects of this project.

There are various sources of airfare data on the Web, which I could use to train our models. A multitude of consumer travel sites supply fare information for multiple routes, times, and airlines.

I tried various sources ranging from many APIs to scrape consumer travel websites and collected data from Yatra.com

The total dataset contains 1550 rows and 9 columns.

The dataset contains information like Airline’s name, source, destination, departure time, duration, arrival time, total stops, price and additional information like free meals, layover, baggage weight details, etc.



* Data Preprocessing Done

A few basic cleaning and feature engineering are done in the data. A lot of data preparation needs to be done according to the model and strategy we use, but the basic cleaning is done initially to understand the data better.

There were few repetitions in the data collected.

There were no missing/null values in the dataset.

Since all the feature variables are of object data type, we need not check for it’s skewness, outliers or distribution.

* Data Inputs- Logic- Output Relationships

We can see this graphically that Jet Airways are more costly and Spice Jet is very affordable. Also the flights that provide free meal and additional facilities are more costly as compared to the direct flights.

* Hardware and Software Requirements and Tools Used

**Python** is the most popular technology for implementing machine learning ideas, owing to the fact that it has a large number of built-in algorithms in the form of bundled libraries.

The following are some of the most important libraries and tools we used in our project:

1. **Numpy:** NumPy is a Python module for array processing. It is the most important Python module for scientific computing. NumPy may be used as a multi-dimensional container of general data in addition to its apparent scientific applications.

It allows any data types to be created, allowing NumPy to connect with a broad range of databases cleanly and quickly.

1. **Jupyter Notebook**: Jupyter Notebook is an open-source online software that lets you create and share documents with live code, equations, visualisations, and narrative prose. Data cleansing and transformation, numerical simulation, statistical modelling, data visualisation, machine learning, and more are all included.
2. **Sklearn**- Power transform, label encoder, standard scaler, linear, random forest, decision tree, Gradient boosting Regressor, k-nearest neighbours, r2 score, mean absolute error, mean squared error, train test split, grid search cv and ensemble technique.
3. **Chromedriver**: Chromedriver is used to web scrape the data and automate the process. I have scraped the data using Selenium python.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

We go over many techniques and datasets that were used to create this module.

The model is trained using a dataset comprising over 1500 tuples. The price of a flight is determined by factors such as the number of stops, duration, facilities provided, etc.

I created regressor methods and compared all different flight models because this is a regression problem. Anaconda seeks to address Python's dependency well, where distinct projects have various dependency versions, so that project dependencies do not require separate versions, which might conflict.

* Testing of Identified Approaches (Algorithms)

The models used training and testing datasets are as followed:

• SGD Regressor

• KNeighbors Regressor

• Decision Tree Regressor

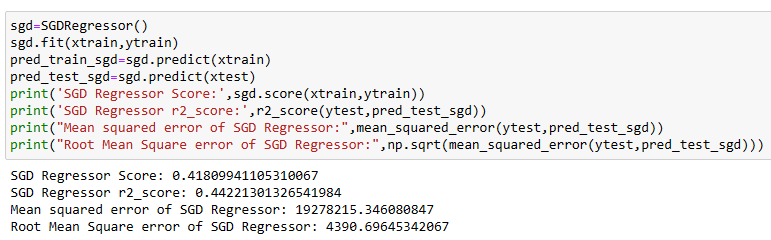
• Random Forest Regressor

• Gradient Boosting Regressor

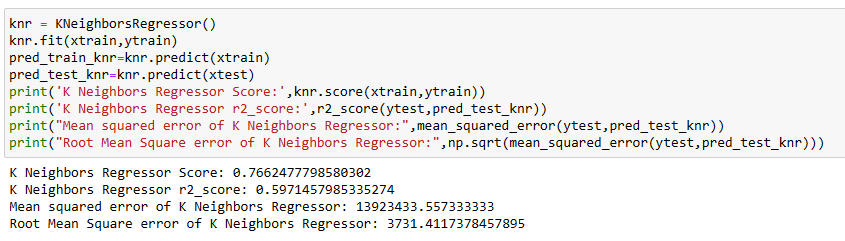
* Run and Evaluate selected models

1. **SGD Regressor**: The loss gradient is calculated each sample at a time, and the model is updated along the way using a decreasing strength schedule. SGD stands for Stochastic Gradient Descent (aka learning rate).

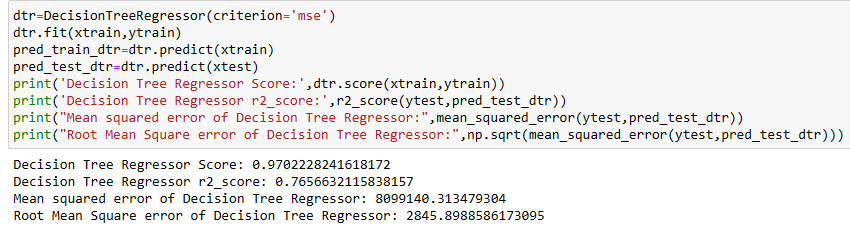
The regularizer is a penalty applied to the loss function that decreases model parameters towards zero using either the squared euclidean norm L2 or the absolute norm L1 or a mix of the two (Elastic Net). The update is trimmed to 0.0 whenever the parameter update passes the 0.0 value due to the regularizer, allowing for the learning of sparse models and online feature selection.



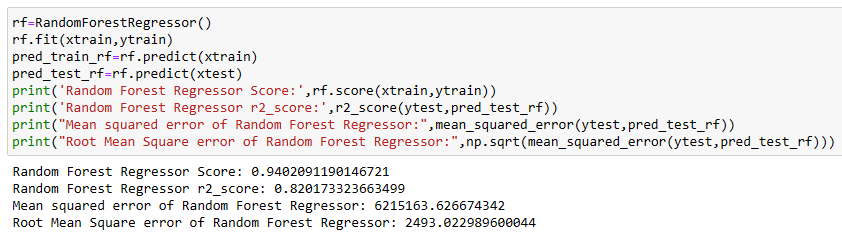
1. **K Neighbors Regressor**: Algorithm Calculating the average of the numerical goal of the K nearest neighbours is a straightforward implementation of KNN regression. An inverse distance weighted average of the K closest neighbours is another method. The distance functions used in KNN regression are the same as those used in KNN classification.



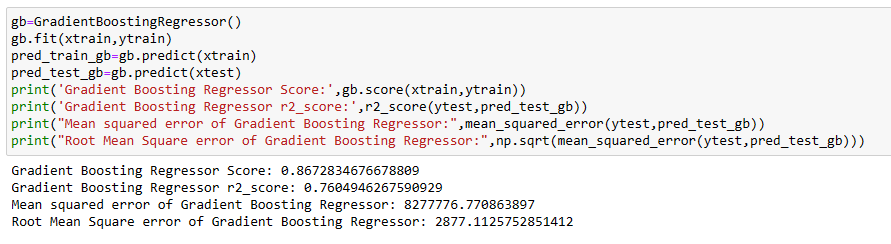
1. **Decision Tree Regressor**: Classification trees are tree models in which the goal variable can take a discrete set of values; in these tree structures, leaves indicate class labels and branches represent feature combinations that lead to those class labels. Regression trees are decision trees in which the target variable can take continuous values (usually real numbers). The objective is to build a model that predicts the value of a target variable from a set of input variables.



1. **Random Forest Regressor**: A regressor with a random forest. A random forest is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting a number of classification decision trees on various sub-samples of the dataset.

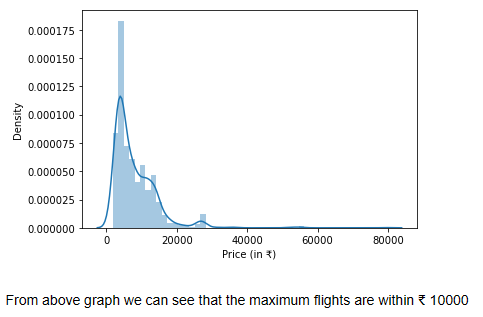


1. **Gradient Boosting Regressor:** Gradient boosting regressors are a type of inductively generated tree ensemble model. At each step, a new tree is trained against the negative gradient of the loss function, which is analogous to (or identical to, in the case of least-squares error) the residual error.

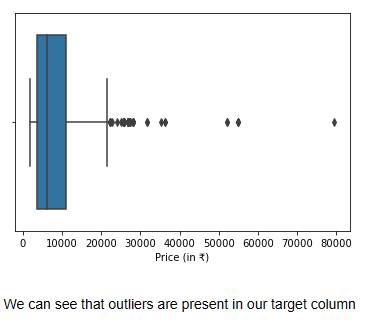


* Visualizations

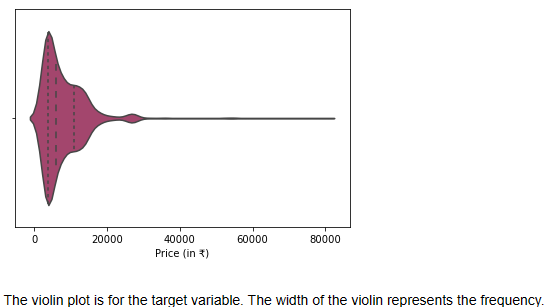
1. **Distribution Plot**

****

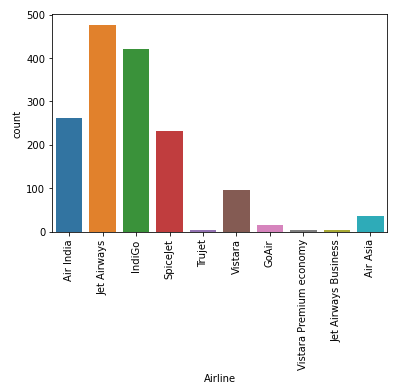
1. **Box Plot**

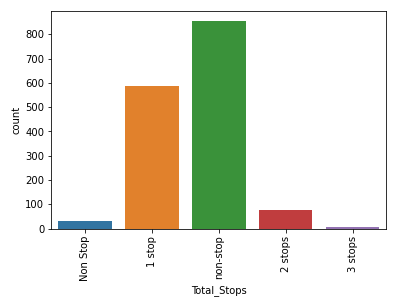
****

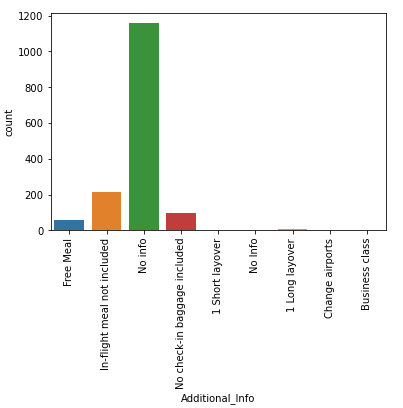
1. **Violin Plot**

****

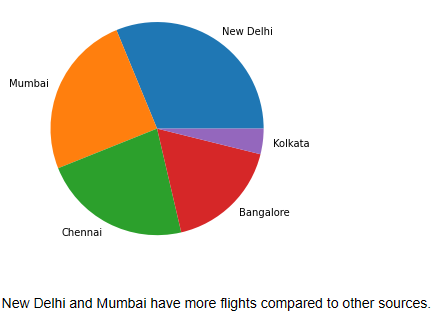
1. **Count Plot**

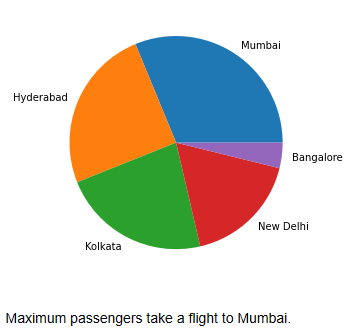
****

****

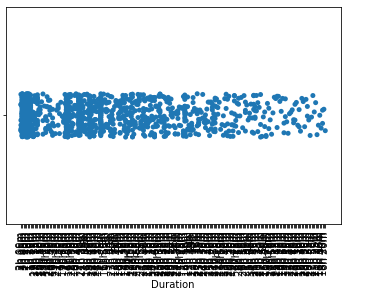
****

1. **Pie Plot**

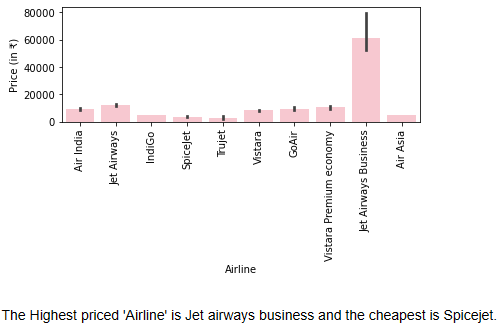
****

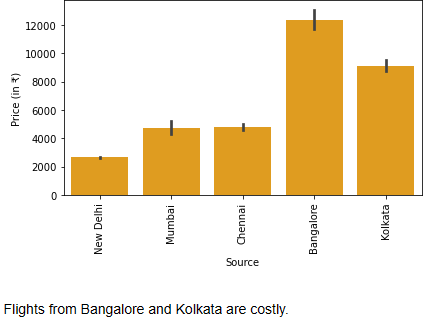
****

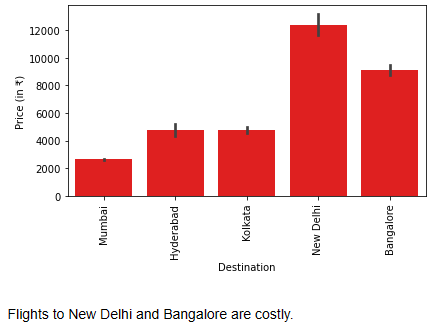
1. **Strip Plot-** Reflecting wide range of duration

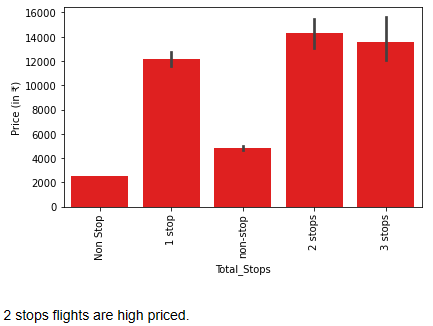
****

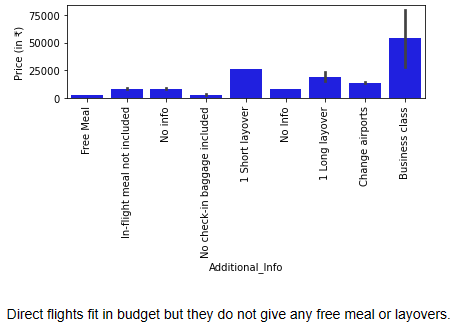
1. **Bar Plot**

****

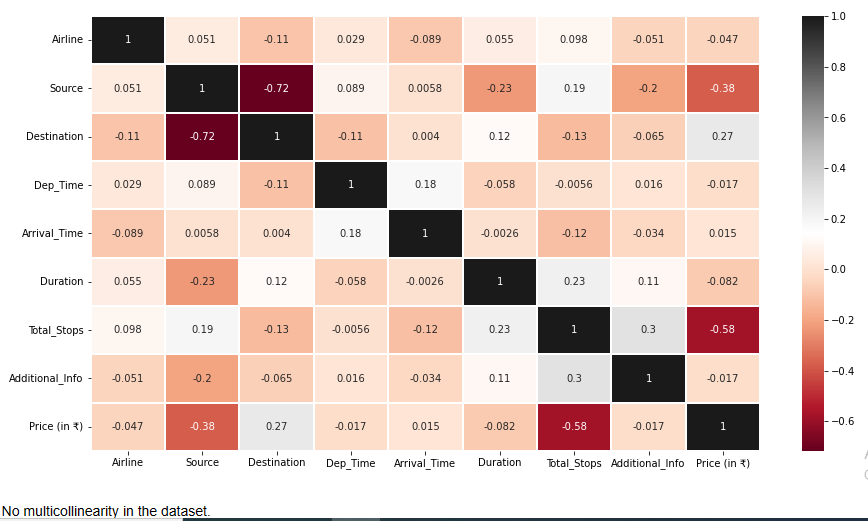
****

****

****

****

1. **Heat Map**

****

* Interpretation of the Results

The Jet Airway Airlines is more costly than others, whereas SpiceJet and IndiGo are quite affordable.

Flights from metro cities are more in number and hence few are in budget and few are way too expensive.

The expensive flights usually come with layover(long/short), free meal and some other additional facilities as well.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The trend of flight prices vary over various months and across the holiday. There are two groups of airlines: the economic group and the luxurious group. Spicejet, AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in the other. Vistara has a more spread out trend.

* Learning Outcomes of the Study in respect of Data Science

Collected and analysed data for 6 routes which spanned across business & tourist routes in India

Some of the routes had non-decreasing prices and thus the model suggested buying the ticket always

Implemented algorithms like Decision Tree, Random Forest, Gradient Boosting and statistical analysis

Developed a basic UI for the model

* Limitations of this work and Scope for Future Work

1. Days of the week and months should also be considered as well to predict the price more efficiently.
2. To make the prediction accurate, time of booking should also be considered (as in when a person is booking tickets, how many days prior to the journey?, etc. )
3. More variables can be added, we can try different models with different subset of features and/or rows.
4. Machine learning require large amount of data.
5. This project has scope for improvement.