

FLIGHT PRICE PREDICTION

Submitted by:

Ruchita parmar

**ACKNOWLEDGMENT**

mentioning of all the references, research papers, data sources, professionals and other resources that helped me and guided me in completion of the project. I use the internet and some help from youtube.

**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

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. 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 Jet Airways? Are morning flights expensive

* Review of Literature

The paper begins with a piece of general information about Machine learning and then the authors further proceed to the methodology comprising of four distinct phases of Feature Selection that influence airfare prices, collection of data from kaggle , Selection of accurate ML Regression model, and its evaluation. The airline dataset had the following eleven features- 'Airline, Date\_of\_Journey, Source, Destination, Route,Dep\_Time, Arrival\_Time, Duration, Total\_Stops,Additional\_Info. The authors performed prediction using eight state-of-art regression Machine Learning models including, Random Forest Regression Tree, Regression Tree, Bagging Regression Tree, Regression and Linear Regression. Performances of these ML models were also compared and evaluated. The random forest Regression model outperforms other models with its accuracy of 82.%.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

most common techniques will fall into the following two groups:

Supervised learning, including regression and classification models.

Unsupervised learning, including clustering algorithms and association rules.

Regression Models

Data analysts use regression models to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential business decisions

* Data Sources and their formats

*Data Collection*

Both the training and testing datasets have been extracted from Kaggle data repository. They contain categorical as well as nominal data related to the Indian Airlines from the year 2019. The dataset provides vital information about some impacting features to predict the fare of a flight - such as the places of departures and arrivals, time of departure and arrivals, the route of the flight, the number of halts during the journey and the price of the ticket depending on those features. It’s an enormous dataset of 10683 rows and 11 columns (each representing one attribute)

* Data Preprocessing Done

While pre-processing the data, we converted the date of journey, departure time and the arrival time from string datatype to date-time object and extracted the numeric values from them; the month-date numeric value from the date of journey attribute and hour-minute numeric value from the departure time and arrival time attributes respectively. Later, we have implemented the ‘One hot encoding’ method for the nominal categorical data and the label encoding method for ordinal categorical data present in both the training as well as the testing dataset. ‘One hot encoding’ is a process of converting the categorical data variables into numerical values thus making it suitable to use while implementing machine learning algorithms. One hot encoding method was applied to nominal categorical data attributes such as the ‘source’, the ‘destination’ and the ‘airline company’ chosen by the user.

‘Label encoding’ helps us convert the labels into numeric values in order to make the dataset suitable for use. Label encoding method was applied to the nominal categorical data attributes such as the ‘total number of halts in the journey’. The columns were re - arranged at the last step.

* *Data Cleaning*

The null values present in the training dataset where removed. A few columns which were of no use for the feature selection process were deleted from the dataset. The columns of attributes having the categorical data were dropped from the dataset after the new columns containing the numerical values extracted from the pre-processed data were stored for the prediction

* Hardware and Software Requirements and Tools Used

Jupyternotebook

Internet(crome)

Pandas

Numpy

**Model/s Development and Evaluation**

Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

1. random forests

2.DecisionTreeRegressor

3.ExtraTreesRegressor

Run and Evaluate selected models

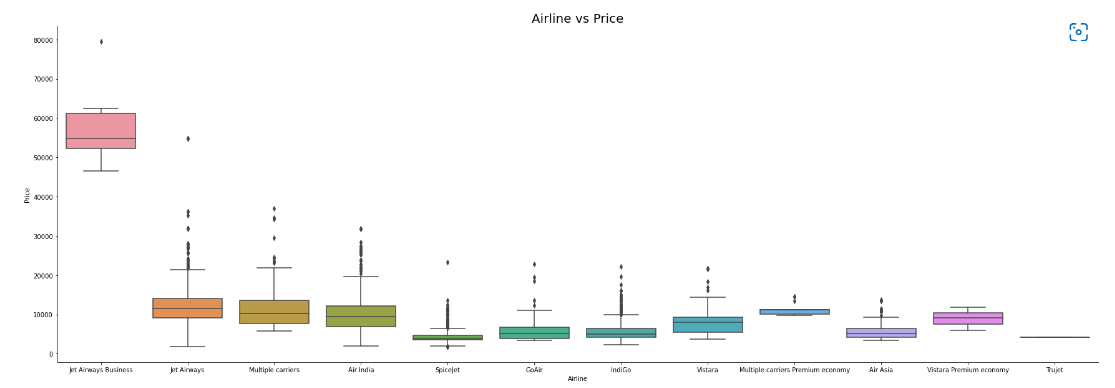
Random Forest Regression: **Random Forest Regression** is a supervised learning algorithm that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It operates by building decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

Key Metrics for success in solving problem under consideration

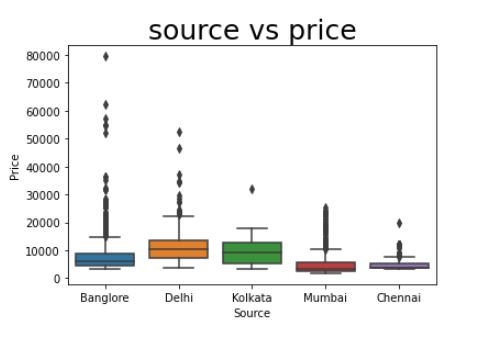
Machine learning; Performance metric; Prediction error; Model selection

Visualizations

.



The most expensive option is ‘JetAirways Business.’ The cost of other carriers varies as well



**Overall evaluation**

Dynamic pricing is one of the most common pricing strategies implemented by the airline industry to adjust ticket prices in response to various internal and external factors such as changes in demand, competitor promotions, ability of users to buy, availability of seats and others. Airlines need to predict changes in these factors to implement a dynamic pricing scheme that dynamically adjusts ticket prices to increase their profit. On the other hand, customers are also interested to forecast how ticket prices would change in the future to be able to buy tickets at lower prices. Therefore, researchers have developed various prediction models both for airlines and customers to help them deal with dynamic pricing. The two most common methods proposed for airlines are demand prediction and price discrimination which we collectively refer to as Airlines side models. Customer side modes involve optimal ticket purchase time prediction models and ticket price prediction models. There is a tradeoff between money saving by customer and increasing revenue by companies. As customers become more strategic by using customer side tools, it becomes more difficult for the airlines to apply dynamic pricing and to generate profit and vice versa. Therefore, there is a need for a prediction model that can predict the optimal ticket prices that can bring mutual benefit both for customers and airlines.

**CONCLUSION**

For this project, an extensive study was carried out with dataset collection from internet and Random Forest Machine Learning model was used for deployment. Using visualization, we were able to determine the features which influence airfare prices the most. With experimental analysis, it can be concluded that Random Forest Regression model achieves good accuracy. The future aim is to work more on the feature selection and model accuracy. We also plan to extend the study by working with larger datasets and greater number of experimentations on the same to procure more accurate airfares which will in turn help users to get an estimated cost of their next airplane travel and can benefit them to make the best deal. We also plan to level up web applications’ user interface to provide a premium user experience. We can also consider various other crucial features that affect airplane ticket prices like public holidays, number of luggage, number of hours till departure, crude oil price, etc. in order to get best results. In the near future, there is also a plan to host the web application.

Thank you