Project Report

Machine Learning (CS 4131)

Academic Year- 2020-21

On

Flight Fare Prediction

Submitted by

1. Anjali Mishra BT17GCS157
2. Arundhati Das BT17GCS016
3. Ayushi Kapoor BT17GCS020

Introduction:

Everyone requires to travel one time or the other. But sometimes the travel plans do not work out because of various reasons. One of the most important reasons is Air fares and travel tickets. There is no clear pattern which would help the customer to pick the best date to travel with the cheapest price. Similarly, there is no easy way to identify what is the best date for a customer to book the ticket. Airlines promote dynamic pricing of their tickets, or a kind of pricing based on demand and supply. The airlines have full control on the ticket prices, and based on their profit margin or sales decide on the price of the tickets. At this point, the customer is dependent on the airline to purchase a cheap ticket. Many a times, the airlines do a steep increase in the prices based on a number of features like number of unsold seats or frequency of the flight between the source and the destination. There is not really a clear pattern in the price of the tickets. However, given the data for a year of airlines and ticket prices, we can predict the prices for the next year.

In order to make it easier for the customer, we propose a flight fare prediction system. There is a prediction on the ticket fare. This ticket fare would help identify the cheapest date to travel or even the cheapest date to book a ticket. This way the customer would know that they would be able to save the maximum amount of money, if they have flexible dates to travel or are willing to take a chance by booking at a future date.

Scope of Project:

- Building a prediction for best booking dates.
- Building a prediction for best travel dates.
- Building a flask web application & deploying it on Heroku.

Key Facts:

- There is base fare for each flight. This base fare has all the considerations of the fare for the airlines.
- Ticket Fare is equal to base fare plus a fluctuation amount. This fluctuation is defined as the delta of price change. There are various factors which can affect the fluctuations
    - Oil Prices
    - Number of Unsold seats
    - Distance between the source and the destination

    - Frequency of flights between origin and destination
    - Dates of travel
- Multiple flights ply between multiple origin and destination combinations on the same dates.

Motivation:

Predicting flight prices without having a proper idea about a particular airline company is near to impossible, especially when you want to book any flight real quick. Using a Machine Learning approach it becomes quite easy as the model predicts how much you need to spend on your flight expenses by getting rid of all unnecessary calculations and brain scratching thoughts.

From a user's point of view this somehow proves to be a profitable project model as users can choose from a variety of airlines of their choice and decide the budget for their ongoing trip without spending anything extra.

Dataset:

We have 2 datasets. A training dataset which has flight price and a testing dataset which can be used to predict the prices.The data in scope has details about various airlines and their prices. It contains information relating to four months i.e from March 2019 to June 2019.The training data contains 10683 rows and 11 columns and test data contains 2671 rows and 10 columns.

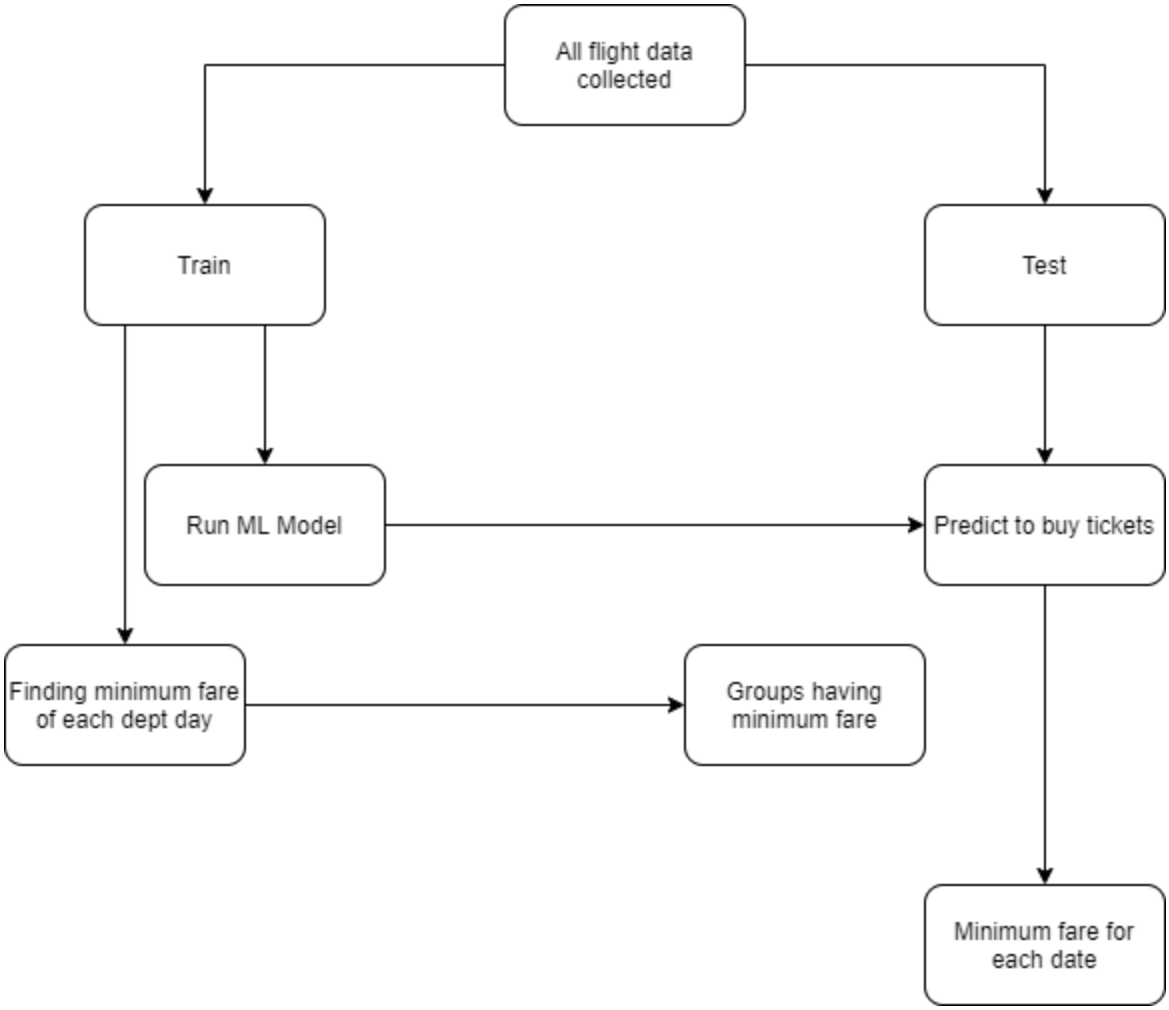
- **Source:** Kaggle
- **Dataset link:** https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh
- **Data Overview:**

  Here each data point corresponds to the trip of flight from one city to another.
    - **Airline:** The name of the airline.
    - **Date_of_Journey:** The date when the journey happened
    - **Source:** The location from where the flight starts.
    - **Destination:** The destination location-where the flights lands.
    - **Route:** The route taken by the flight to reach the destination.
    - **Dep_Time:** The time when the journey starts from the source.
    - **Arrival_Time:** Time of arrival at the destination.
    - **Duration:** Total duration of the flight.
    - **Total_Stops:** Total stops between the source and destination.
    - **Additional_Info:** Additional information about the flight. For eg.Information relating to baggage, meals etc.
    - **Price:** The price of the ticket

Tools & Technologies:

- **Language :** Python, HTML
- **Tools:** Heroku, Flask, CSS
- **Collaborate:** Google colab

Architecture:



Methodology:

The dataset is collected from the source kaggle for the flight fare prediction. Data cleaning process is applied for a given dataset.The processed data is then divided into training and testing dataset with 80% and 20% split respectively. Different Machine learning models like Random Forest, XGBoost, Gradient Boosting are applied to calculate the fare for each departure day.

- **Data Preparation & EDA:** Based on our model requirements the data preparations are carried out. Variable transformations are key for better analysis of data. For eg.in our

dataset there were a whole lot of features which were of type object. The very first task was to convert these features into integer type.

- **Algorithms :** Regression models like Random Forest, XGBoost, Gradient boosting were applied on the training data set.

1.Random Forest:

It is an algorithm which ensembles the less predictive model to produce better predictive models. The features are sampled and passed to trees without replacement to obtain the highly uncorrelated decision trees.

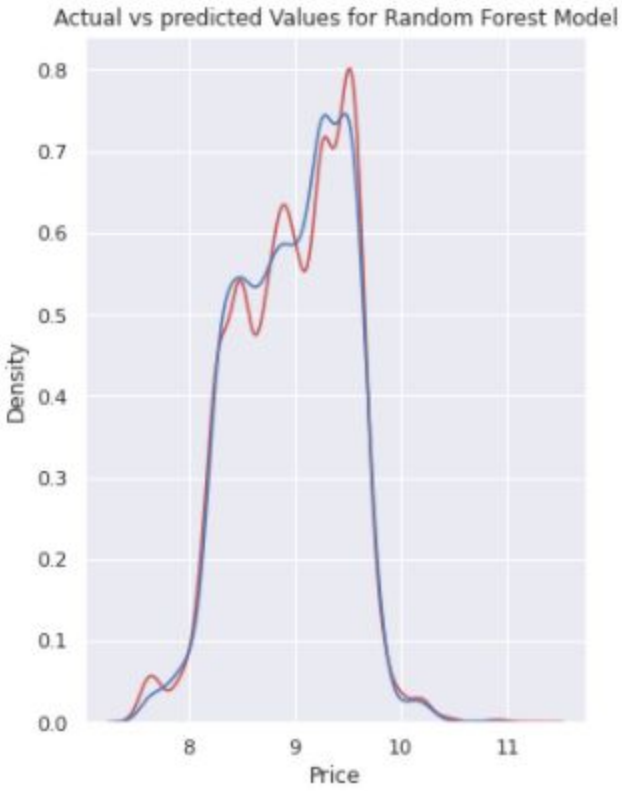


Fig 1: Graphical Results for Random Forest

2.Gradient boosting Model

It is an ensemble machine learning model which produces a prediction model in the form of an ensemble of weak prediction models.

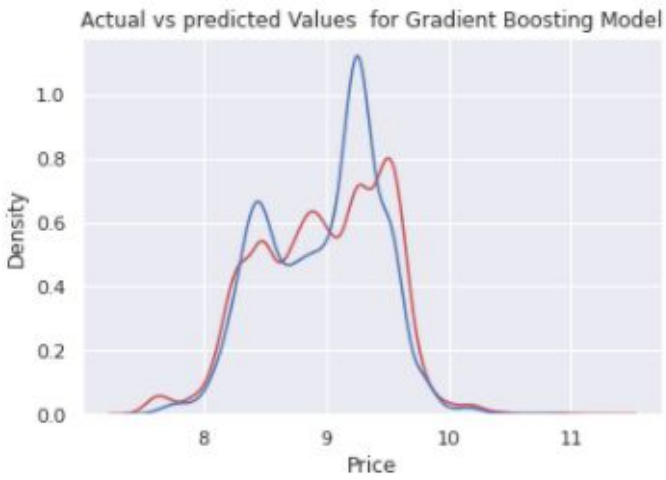


Fig 2: Graphical results for Gradient boosting

3. XGBoost Model

It is a decision-tree-based ensemble Machine Learning algorithm that makes use of a gradient boosting framework.

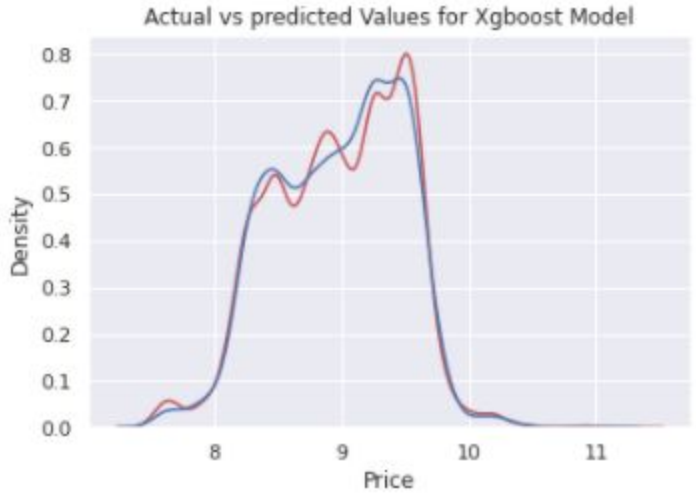


Fig 3: Graphical Results for XGBoost

- **Performance Metric:** Performance metrics like MSE, MAE ,RMSE are used to interpret the model accuracy

Algorithm	R-squared	RMSE	MAE
Random Forest	0.93	0.13	0.07
Gradient Boosting	0.86	0.18	0.07
XGBoost	0.94	0.12	0.13

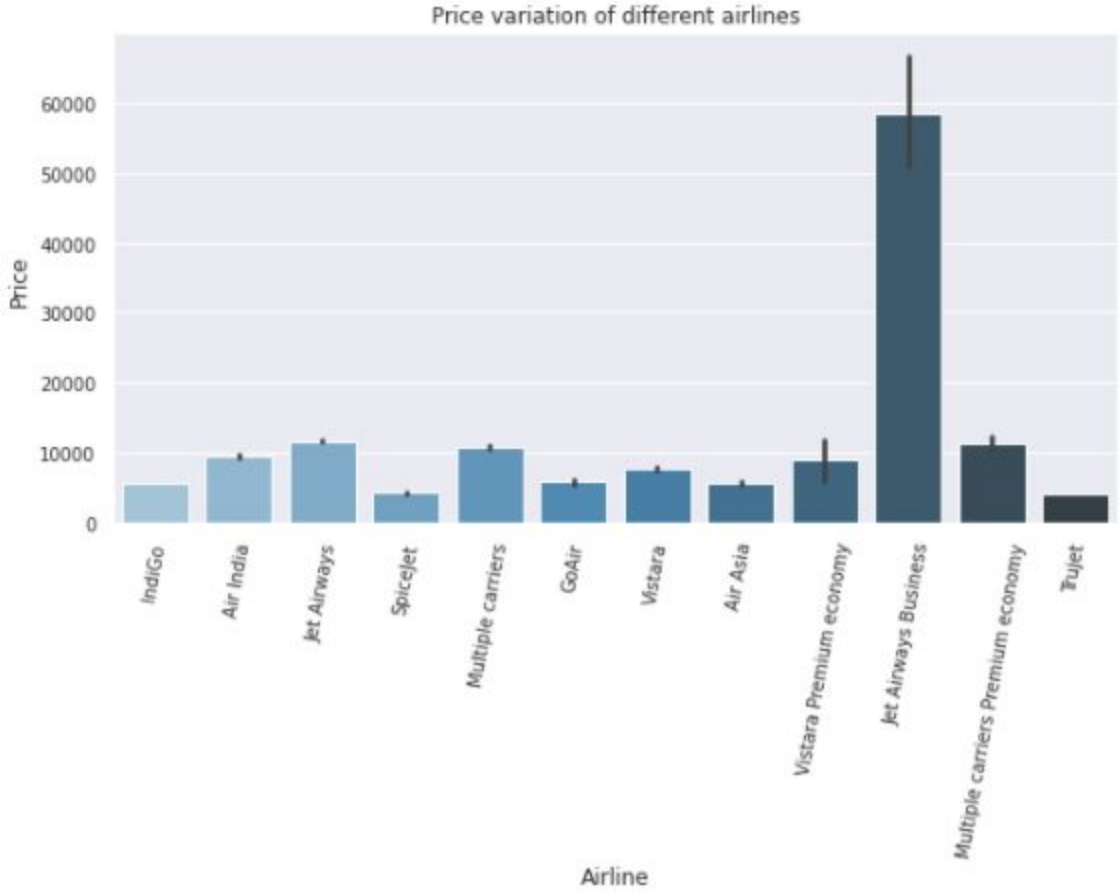
Table1: Algorithm Evaluation

Experimental Results :

For the selected test data set , output of the model is plotted across the test dataset. By the analysis of the results obtained from the algorithm such as Random Forest , Gradient Boosting, XGBoost regression gives the predicted values of the fare to purchase the flight ticket.Table I gives the values for RSquare, RMSE and MAE. The XGBoost algorithm has more accuracy compared to other algorithms for the given dataset, It gives the highest R-Square value with maximum accuracy .

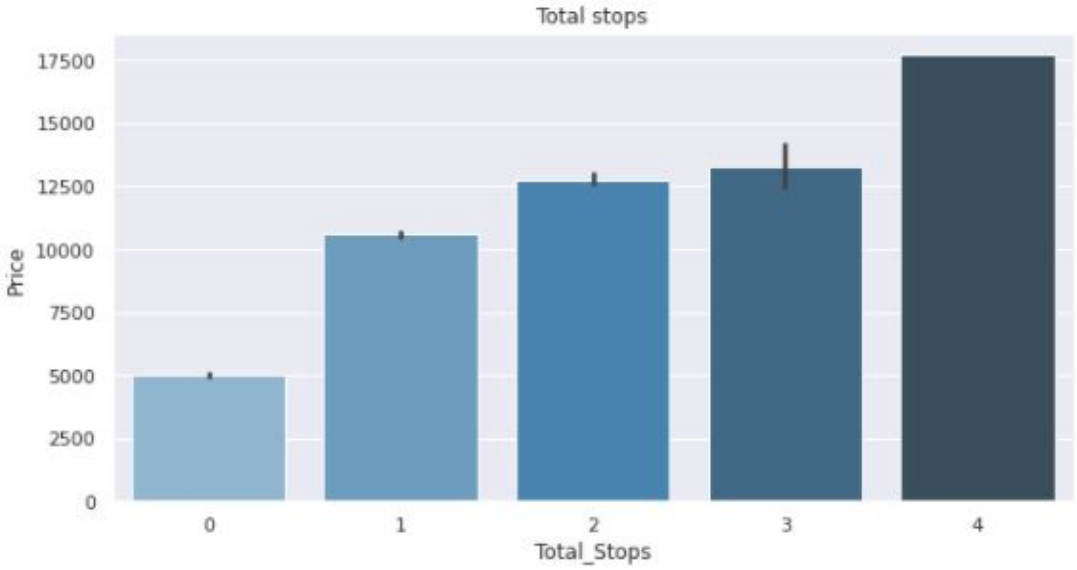
Analysis:

- **Variation of flight prices for different airlines**



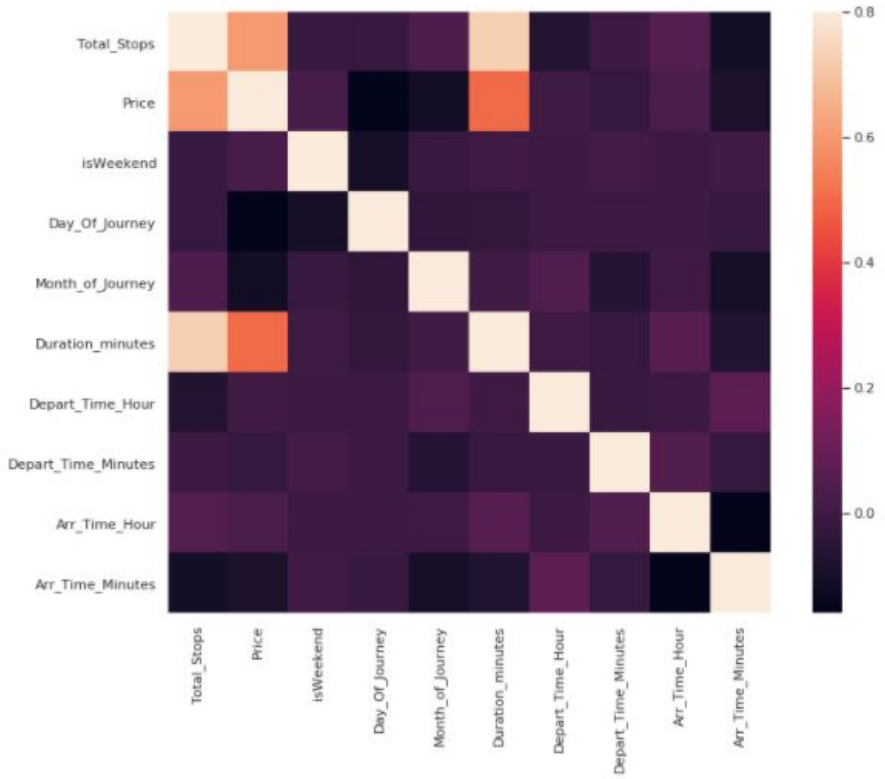
From the above graph we can see tha Jet airways business ticket is most expensive

- **Variation of price against total number of stops:**



From the above plot we can see that flight tickets fare is higher for flights with greater number of stops, there is a direct relationship between the two variables.

- **Correlation matrix**



Conclusion:

- The model helps the airline companies identify their high revenue flights and trends and helps consumers understand the best times to purchase the tickets.
- Implemented the algorithms using python on Google colab. Here is the link for the same:

  https://colab.research.google.com/drive/13CIsdjVxrKw84NkoQXWyNQFzqjeDeenq?usp=sharing

- Designed a flask web application & deployed it on Heroku for predicting the flight fares. Here is the link for the same: https://predictflight.herokuapp.com

Future Scope:

This project can help a number of customers to benefit from their ticket bookings, especially during the holiday season. Here are some of the challenges that are there in terms of the project scope:

- Airlines do not want to disclose their prices to their competitors.
- It is very difficult to monitor the fluctuating factors. Eg: Even if the price of the ticket is cheap at the end of December but if the oil prices increase, the flight prices would definitely increase as well.

Hence, after looking at these challenges, it is important that this idea would grow, only after all airlines are aligned and agree to a common platform by sharing their data. We can extend the scope of the project by:

- Integrating and scraping data from each airline to get real prices of their flights between the source and destination combination
- Integrating the exchange rates and oil prices into consideration at the dataset level.
- Using various Business Intelligence techniques to try and find a pattern for the data.
- Increasing the training sample for the model for each year.

References:

1. Tziridis, Konstantinos & Kalampokas, Theofanis & Papakostas, George & Diamantaras, Kostas (2017) Airfare Prices Prediction Using Machine Learning Techniques 10.23919/EUSIPCO.2017.8081365.
2. Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar(2019) Predicting The Price Of A Flight Ticket With The Use Of Machine Learning Algorithms INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 12
3. Wang, Tianyi & Pouyanfar, Samira & Tian, Haiman & Tao, Yudong & Alonso, Miguel & Luis, Steven & Chen, Shu-Ching (2019) A Framework for Airfare Price Prediction: A Machine Learning Approach 200-207. 10.1109/IRI.2019.00041.