Prediction of Airfare Using Machine Learning

[1]Vinal Raja, [2]Janhavi Vakil, [3]Yash Shah, [4]Sonia Relan

Computer Engineering Department
NMIMS'S MPSTME, Shirpur, India

***Abstract*— As the airline industry is flourishing various practices are being used by the airline industry which they call revenue management or yield management. Data visualization and machine learning algorithms are basis of the project airline price prediction. The use of logistic regression suggests the customer whether is it the right time to buy the airline ticket or should he wait. The process involves data extracting, analyzing, and interpreting which consist of data mining and decision making.**

***Index Terms*— yield mangementt;logistic regression; data mining ; data extraction.**

I. Introduction

The varying prices of airline ticket has always kept the consumer confused, it is a challenging task as various practices are used by the industry to increase their revenue [1] .The consumers being uncertain about the change in price fall prey to these and land up buying high priced tickets. Indian aviation industry states that it is on high-growth trajectory. It aims to becomes third largest aviation market by 2020 and the largest by 2030, according to Directorate General of Civil Aviation domestic air traffic has growth at the rate of 22 per cent, International passenger traffic at 8.53 per cent by the report of February 2017 8.23 million passengers have flown domestic[2].

"Cheap Air Tickets" is most searched term in India in sector of aviation industry. Due to increase in number of domestic passengers, consumers hunting for cheap priced air tickets have increased [3] .Hence, the proposed research is basically machine learning and statistic intensive. We used Python & Java for the implementation of the models & automation.

II. Technical Aspects

Automated script to collect data

Historic data for the past flight prices for each route are collected on daily basis as manually collecting data is not efficient and would increase the work load . We use python script that runs on a remote server which collect prices on daily basis for specific time and date.

Cleaning and Preparing Data

Clean & prepare the data according to the model's requirements. After data collection this is the step that is the most important and the most time consuming. We used various statistical techniques & logics and implemented those using built-in python packages

Analyzing and Building Models

Analyzing data is given by discovering patterns and hidden trends followed by applying various predictive and regression techniques.

Merging models and accuracy calculation

We now have to test the model on our testing set and come up with the most suitable metric to calculate the accuracy**.** Moreover, to calculate the accuracy of the model and proving the outcome.

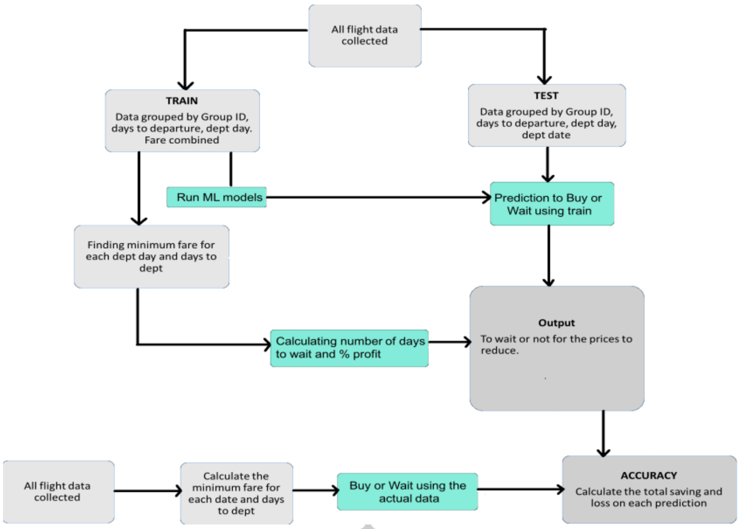

All flight data collected
TRAIN
Data grouped by Group ID, days to departure, dept day. Fare combined
TEST
Data grouped by Group ID, days to departure, dept day, dept date
Run ML models
Prediction to Buy or Wait using train
Finding minimum fare for each dept day and days to dept
Output
To wait or not for the prices to reduce.
Calculating number of days to wait and % profit
All flight data collected
Calculate the minimum fare for each date and days to dept
Buy or Wait using the actual data
ACCURACY
Calculate the total saving and loss on each prediction

Figure 1. System Overview

III. Data Collection

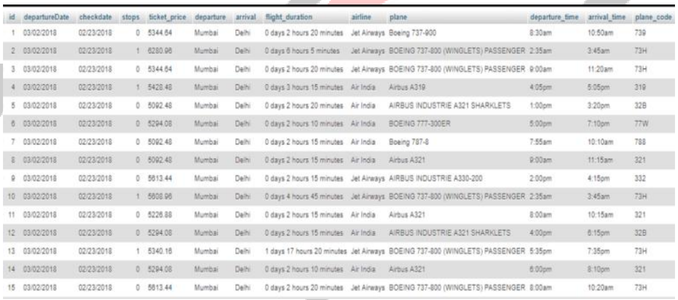
Data Collection is one of the most operation of this project There are several sources of air fare data on the Web, which we could use to train our models. According to google trends we identified busiest routes for domestic flights that are Mumbai (BOM) to Goa (GOI), Mumbai (BOM) to Delhi (DEL), Goa (GOI) to Mumbai (BOM), Goa (GOI) to Delhi (DEL), Delhi (DEL) to Mumbai (BOM), and Delhi (DEL) to Goa (GOI) by designing web crawler and collecting data of various airlines from expeida.com in total six routes [4].

Data Source

We used web crawler to scrape Expedia.com website using a manual spider made in Python through which we extracted data from the website, it is stored in as a CSV file. The web crawler is executed as historic data of flight and its airfare is not easily available [5].

- Web Scrapper(Python 3.6)

Python library was used to access the Expedia.com website and load the required page a json object. This object was parsed using inbuilt python functions and a csv database was obtained. Python function is used to obtain a set of dates between two dates on which the above functions can be applied to get the data for a range of dates. The python script is run to gather the required data based on current scenario we use Xampp server to host the python script and to collect data.



id	departureDate	checkdate	stops	ticket_price	departure	arrival	flight_duration	airline	plane	departure_time	arrival_time	plane_code
1	03/02/2018	02/23/2018	0	5344.64	Mumbai	Delhi	0 days 2 hours 20 minutes	Jet Airways	Boeing 737-900	8:30am	10:50am	739
2	03/02/2018	02/23/2018	1	6280.96	Mumbai	Delhi	0 days 6 hours 5 minutes	Jet Airways	BOEING 737-800 (WINGLETS) PASSENGER	2:35am	3:45am	73H
3	03/02/2018	02/23/2018	0	5344.64	Mumbai	Delhi	0 days 2 hours 20 minutes	Jet Airways	BOEING 737-800 (WINGLETS) PASSENGER	9:00am	11:20am	73H
4	03/02/2018	02/23/2018	1	5428.48	Mumbai	Delhi	0 days 3 hours 15 minutes	Air India	Airbus A319	4:05pm	5:05pm	319
5	03/02/2018	02/23/2018	0	5092.48	Mumbai	Delhi	0 days 2 hours 20 minutes	Air India	AIRBUS INDUSTRIE A321 SHARKLETS	1:00pm	3:20pm	32B
6	03/02/2018	02/23/2018	0	5294.08	Mumbai	Delhi	0 days 2 hours 10 minutes	Air India	BOEING 777-300ER	5:00pm	7:10pm	77W
7	03/02/2018	02/23/2018	0	5092.48	Mumbai	Delhi	0 days 2 hours 15 minutes	Air India	Boeing 787-8	7:55am	10:10am	788
8	03/02/2018	02/23/2018	0	5092.48	Mumbai	Delhi	0 days 2 hours 15 minutes	Air India	Airbus A321	9:00am	11:15am	321
9	03/02/2018	02/23/2018	0	5613.44	Mumbai	Delhi	0 days 2 hours 15 minutes	Jet Airways	AIRBUS INDUSTRIE A330-200	2:00pm	4:15pm	332
10	03/02/2018	02/23/2018	1	5608.96	Mumbai	Delhi	0 days 4 hours 45 minutes	Jet Airways	BOEING 737-800 (WINGLETS) PASSENGER	2:35am	3:45am	73H
11	03/02/2018	02/23/2018	0	5226.88	Mumbai	Delhi	0 days 2 hours 15 minutes	Air India	Airbus A321	8:00am	10:15am	321
12	03/02/2018	02/23/2018	0	5294.08	Mumbai	Delhi	0 days 2 hours 15 minutes	Air India	AIRBUS INDUSTRIE A321 SHARKLETS	4:00pm	6:15pm	32B
13	03/02/2018	02/23/2018	1	5340.16	Mumbai	Delhi	1 days 17 hours 20 minutes	Jet Airways	BOEING 737-800 (WINGLETS) PASSENGER	5:35pm	7:35pm	73H
14	03/02/2018	02/23/2018	0	5294.08	Mumbai	Delhi	0 days 2 hours 10 minutes	Air India	Airbus A321	6:00pm	8:10pm	321
15	03/02/2018	02/23/2018	0	5613.44	Mumbai	Delhi	0 days 2 hours 20 minutes	Jet Airways	BOEING 737-800 (WINGLETS) PASSENGER	8:00am	10:20am	73H

Figure 2 Collected & Prepared Data

Data Collected

- Departure date
- Check date
- Stops
- Ticket price
- Departure destination
- Arrival destination
- Airline
- Plane
- Departure time
- Arrival time
- Flight number
- Time slot

IV. Data Preparation

A user makes a query to buy a flight ticket 15 days in advance, then our system should be able to tell the user whether he should wait for the prices to decrease or he should buy the tickets immediately [6].

1. Predict the flight prices for all the days between 15 and 1 and check on which day the price is minimum.
2. Data can be classified into, **"Buy" or "Wait"**. This then becomes a classification problem and we would need to predict only a binary number.
3. Day of departure and departure day are the influencing factors which determine the flight prices, hence we created the group according to the airlines and the departure time-slot created earlier (Morning, Evening, Night) and calculated the combined flight prices for each group, since competition could also play a role in determining the fare.

V. Machine Learning

We use a classification algorithm that is logistic regression which is used to assign observations to a discrete set of classes. It is different from linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes [7].

I. Types of Logistic Regression

- Binary (Buy/Wait)
- Multi (Morning, Evening, Night)
- Ordinal (Low, Medium, High)

Binary Logistic Regression

We use the sigmoid function to map predicted values to probabilities. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities [8].

II. Equation

S(z)=11+e−z

- s(z) = output between 0 and 1 (probability estimate)
- z = input to the function (your algorithm's prediction e.g. mx + b)
- e = base of natural log

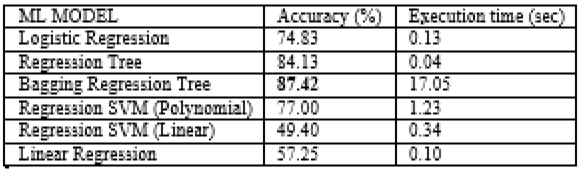
III. Code

- **def sigmoid**(z): **return** 1.0 / (1 + np**.**exp(-z))

VI. Results

Logistic regression transforms its result using the sigmoid logistic function to return a probability value that can then be assigned to two or more discrete classes. We use it for airline industry because most of the time the ticket price keeps changing for any particular day. For example - if you want to buy a ticket for a flight in ten days the ticket price may increase or decrease according to the day and the difference between travel date and booking date [9]. As we have used a training dataset for the prediction of ticket price (buy or wait for ticket) so it gives us a good result. The accuracy of logistic regression model is up to 70-75% (we have taken this result from previous papers on this topic). The conclusion of the given model is that most of the plane ticket price vary from day to day. We have observed that the ticket price is high for a certain period of time and then it gradually decreases to a certain level. When the flight is at a difference of 2-3 days' time the ticket price starts increasing again.

Result - We give an overall prediction based on all the flight details collected for the particular day. When we click on the buy option given with each of the flights we get the prediction for each of the flight separately which is given from the data of that particular flight [10].



ML MODEL	Accuracy (%)	Execution time (sec)
Logistic Regression	74.83	0.13
Regression Tree	84.13	0.04
Bagging Regression Tree	87.42	17.05
Regression SVM (Polynomial)	77.00	1.23
Regression SVM (Linear)	49.40	0.34
Linear Regression	57.25	0.10

Figure 3 Machine Learning Model Result

- There are two groups of airlines (economy and luxury) Spicejet, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in both the categories.
- The airfare varies depending on the time of departure and date making timeslot used in analysis an important parameter.
- The airfare increases during holidays, during Christmas the fare remained high for all the values of days to departure. We haven't considered holiday season as a parameter now, since we are looking at data for a few days.
- Airfare varies according to the day of the week of travel. It is higher for weekends and adjoining days.
- There are a few times when an offer is run by an airline because of which the prices drop suddenly. These are difficult to incorporate in our mathematical models, and hence lead to error.
- The Mumbai-Delhi route prices of flight increases or remains constant. This is because of the high frequency of the flights, high demand and also could be due to heavy competition.

REFERENCES

[1] P. Malighetti, S. Paleari and R. Redondi, "Pricing strategies of low-cost airlines: The Ryanair case study," Journal of Air Transport Management, vol. 15, no. 4, pp. 195-203, 2009.
[2]https://www.ibef.org/industry/indian-aviation.aspx.
[3]Airfare Prices Prediction Using Machine Learning Techniques, K. Tziridis, Th. Kalampokas, G.A. Papakostas HUMAIN-Lab, 2017 25th European Signal Processing Conference (EUSIPCO)
[4]Machine learning modeling for time series problem: Predicting flight ticket prices by Jun Lu, Computer Science, EPFL
[5]https://github.com/humain-lab/airfare_prediction.
[6][Etzioni et al., 2003] Oren Etzioni, Rattapoom Tuchinda,Craig A Knoblock, and Alexander Yates. To buy or not to buy: mining airfare data to minimize ticket purchase price. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 119–128. ACM, 2003.
[7]https://github.com/perborgen/LogisticRegression/blob/master/logistic.py
[8]http://www.holehouse.org/mlclass/06_Logistic_Regression.html
[9] M. Papadakis, "Predicting Airfare Prices," 2014
[10] https://achyutjoshiairlineprediction.github.io/btp/results