International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 4 296 - 298

Optimizing Airline Ticket Purchase Timing

Manan Dedhia
Department of Computer Engineering
K.J. Somaiya College of Engineering
Mumbai-77, India
manan.dedhia@somaiya.edu

Amit Jadhav
Department of Computer Engineering
K.J. Somaiya College of Engineering
Mumbai-77, India
amit.dj@somaiya.edu

Rahul Jagdale
Department of Computer Engineering
K.J. Somaiya College of Engineering
Mumbai-77, India
rahul.jagdale@somaiya.edu

Prof. Bhakti Palkar
Department of Computer Engineering
K.J. Somaiya College of Engineering
Mumbai-77, India
bhaktiraul@somaiya.edu

***Abstract*—**Our approach in this paper is to suggest the user to either buy or wait for the purchase of airline tickets.Airline tickets prices are volatile and keep on varying depending on various parameters. Users, not having much information about these parameters, are often forced to buy tickets at high prices. This paper proposes a machine learning based prediction system which uses logistic regression to suggest users to buy the ticket, implying that prices are going to rise in coming days or wait for some time implying prices are going to plummet in coming days. This system also predicts the price of the date user wants to travel.

Keywords*-Machine learning; Logistic regression;Air fare; prediction; scrapping.*

*****

I. INTRODUCTION

There is always a high demand of airline tickets and in absence of proper knowledge, users often do not have the luxury to book tickets at the best prices, usually ending up paying higher rate for the seat. This is further complicated by the confidential policies of airline companies, restricting the flow of information towards users which may be helpful to predict when to buy tickets. This void can be filled by our prediction system which will predict the prices of the tickets on the day that user wishes to travel and to suggest user whether to buy or wait for the ticket.

II. PROPOSED APPROACH

The proposed system has been divided into following modules.

The first module was to choose an appropriate algorithm which would be the most efficient and accurate. Few algorithms that were considered for the purpose of this project were support vector machine, linear regression and logistic regression. As the objective of the system is to provide either of the two values (BUY or WAIT), logistic regression was the most suited algorithm for the system. A rudimentary model was created using the chosen algorithm which would predict the fare for any given day and inform user to either buy or wait.

The second module was to scrap the data online from expedia.com website. For any prediction or classification problem, we need historical data to work with so as to run machine learning algorithms on it. For this system, we need to have comprehensive data of past flights on each of the routes considered. For this purpose, a python script was written which collected all the necessary data at a specific time daily.

The script curated the following significant parameters from the website:

1. Arrival Airport
2. Arrival Time
3. Departure Airport
4. Departure Time
5. Plane type Name
6. Airline Name
7. Flight Duration
8. Plane Code
9. Ticket Price
10. Number of Stops

The third module was to clean and prepare the data for further processing. Data needs to cleaned and prepared according to the model's requirements. This is the most important and time-consuming step for any machine learning model.Other task included creating a user interface which was easy enough for the user to understand and self-explanatory.

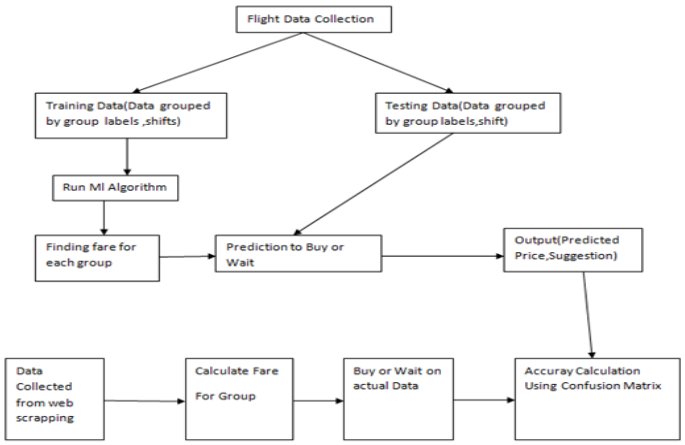

Figure 1: Data flow diagram of the system

III. MODEL

The primary step in this system was to come up with an algorithm which would be the best fit for the prediction. Three algorithms which were shortlisted for trial were SVM, Linear regression and logistic regression. Finally, we chose Logistic regression to proceed with the system.

We chose logistic regression mainly because it is suitable when considering multiple features or variables and also for classification, exactly what we expect from our system. Logistic regression is better than linear for classification problems as linear regression tries to find the linear relationship between the variables, where over fitting problem can occur. We believe intuitively that our data is linearly separable; hence we have not used SVM.

IV. REQUIRED FEATURE SELECTION

In order to get an optimal solution and increase accuracy, twelve features were selected that contributed to the accuracy of the system.

- **Arrival Airport and Departure Airport**- Fares depends on the location user wants to travel, longer the distance, more will be the price. Also if the route selected by the user is busiest and popular route, prices will be high.

- **Arrival Time and Departure Time-** Fares depend on the time of travel. If the flight is early morning then the demand for is lesser. This will lead to less fare whereas for evening and night, demand is high which leads to increase in prices.

- **Airline Name-** Fare also depends on the Airline Company, popular the company more they will charge for the ticket.

- **Plane type -** Whether the carrier is Airbus or Boeing or some other plane type, it will have an impact on price.

- **Flight Duration –** Duration of travel is a significant factor affecting the cost of the price.

- **Number of stops**- Nonstop flights usually cost more than the ones which have one or multiple halts.

V. DATA COLLECTION

In order to obtain a model for the machine learning algorithm, we need a data set which covers all chief parameters upon which fares depend. Out of myriad of sources available online, we preferred expedia.com website to obtain our data. Considering the massive amount of information that would be needed, obtaining this data manually was not feasible. Hence we created a python script which automatically scrapped required data from the website daily.

This data was stored in MySQL format. The routes for which data was collected were:

- MUM-DEL
- DEL-MUM
- MUM-GOA
- GOA-MUM
- DEL-GOA
- GOA-DEL

The obtained data set was further cleaned and prepared for processing. Based on the chosen machine learning algorithm, a java code was implemented to train the new dataset and obtain a new model. The data was stored on local host using phpmyadmin.

VI. IMPLEMETATION

Binary classification with Logistic Regression model (Buy or Wait)

We labeled weights for the logistic regression model. These weights are calculated over the training data set. Using the calculated weights, the constants are computed. The calculated constants for the linear regression model will be called by a function. The function will return the probabilities for each group. We classified using probabilities whether it will fall in buy class or waitclass.

Formulae used:

To classify data:

P=1/1+e^-z (sigmoid function)

To find weights:

New weights= Weights + rate * (label - predicted) * data value

To find Logit function:

New Logit = Logit+weights*data value

We iterated over 3000 times over training data set to train it.

Accuracy after implementation

Accuracy was calculated by predicting decisions daily from the system for a specific day and comparing it with the real and actual prices of that day.
Accuracy calculation was done on 50 data samples. We achieved accuracy of 78% according to confusion matrix shown below.

TABLE I. CONFUSION MATRIX

	Predicted:Wait	Predicted:Buy
Actual:wait	TN:30	FP:1
Actual:Buy	FN:10	TP:9

Accuracy :(TP+TN)/Total =39/50
=0.78 = 78%
Precision :TP/Predicted Yes=9/10
=0.9 =90%

We achieved precision of 90% according to the above calculations.
Misclassification rate is 22%.
True positive rate is 47%.
False Positive Rate is 3.2%.

VII. AVAILABILITY

As the system is completely based on an online portal, it will be available to the user all the time.

VIII. FUTURE SCOPE

This system was implemented for routes betweenBombay, Delhi and Goa. More routes can be added to increase the scope of the project. The training and accuracy of the project can also be increased with more data.

IX. CONCLUSION

The system was implemented successfully for routes between Mumbai, Delhi and Goa. Priceswere predicted and the suggestions to BUY or WAIT were given to the users. Accuracy was calculated forthe result given by implementation of Logistic regression algorithm on the system.

X. REFERENCES

[1] To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price, Etzioni et. Al, SIGKDD 2003
[2] Predicting Airfare Prices , Manolis Papadakis,2012.
[3] On Optimizing Airline Ticket Purchase Timing, WILLIAM GROVES and MARIA GINI, University of Minnesota, 2015.
[4] A Linear Quantile Mixed Regression Model for Prediction of Airline Ticket Prices, Tim Janssen, 2014.