Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

A Survey on Flight Pricing Prediction using Machine Learning

Supriya Rajankar
Dept. of Electronics and Telecommunication
Sinhgad College of Engineeering,
Vadgaon(Bk), Pune, India

Neha Sakharkar
Dept. of Electronics and Telecommunication
Sinhgad College of Engineeering,
Vadgaon(Bk), Pune, India

***Abstract*—What is the best time to buy a flight ticket? 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.**

Index Terms—Machine Learning Algorithm, Predictor, airfare, Naive Bayes, Artificial Intelligence(AI).

I. INTRODUCTION

Any individual who has booked a flight ticket previously knows how dynamically costs change. Aircraft uses advanced strategies called Revenue Management to execute a distinctive valuing strategy [1]. The least expensive accessible ticket changes over a period the cost of a ticket might be high or low. This valuing method naturally modifies the toll as per the time like morning, afternoon or night. Cost may likewise change with the seasons like winter, summer and celebration seasons. The extreme goal of the carrier is to build its income yet on the opposite side purchaser is searching at the least expensive cost. Purchasers generally endeavor to purchase the ticket in advance to the takeoff day. Since they trust that airfare will be most likely high when the date of buying a ticket is closer to the takeoff date, yet it is not generally true. Purchaser may finish up with the paying more than they ought to for a similar seat.

A report says India's affable aeronautics industry is on a high-development movement. India is the third-biggest avionics showcase in 2020 and the biggest by 2030. Indian air traffic is normal to cross the quantity of 100 million travelers by 2017, whereas there were just 81 million passengers in 2015. Agreeing to Google, the expression "Cheap Air Tickets" is most sought in India. At the point when the white collar class of India is presented to air travel, buyers searching at modest costs. The rate of flight tickets at the least cost is continuously expanding.

II. LITERATURE SURVEY

It is very difficult for the customer to purchase a flight ticket at the minimum price. For this several techniques are used to obtain the day at which the price of air ticket will be minimum. Most of these techniques are using sophisticated artificial intelligence(AI) research is known as Machine Learning.

Utilizing AI models, [2] connected PLSR(Partial Least Square Regression) model to acquire the greatest presentation to get the least cost of aircraft ticket buying, having 75.3% precision. Janssen [3] presented a direct quantile blended relapse model to anticipate air ticket costs for cheap tickets numerous prior days takeoff. Ren, Yuan, and Yang [4], contemplated the exhibition of Linear Regression (77.06% precision), Naive Bayes (73.06% exactness, Softmax Regression (76.84% precision) and SVM (80.6% exactness) models in anticipating air ticket costs. Papadakis [5] anticipated that the cost of the ticket drop later on, by accepting the issue as a grouping issue with the assistance of Ripple Down Rule Learner (74.5 % exactness.), Logistic Regression with 69.9% precision and Linear SVM with the (69.4% exactness) Machine Learning models.

Gini and Groves[2] took the Partial Least Square Regression(PLSR) for developing a model of predicting the best purchase time for flight tickets. The data was collected from major travel journey booking websites from 22 February 2011 to 23 June 2011. Additional data were also collected and are used to check the comparisons of the performances of the final model.

Janssen [3] built up an expectation model utilizing the Linear Quantile Blended Regression strategy for SanFrancisco to NewYork course with existing every day airfares given by www.infare.com. The model utilized two highlights including the number of days left until the takeoff date and whether the flight date is at the end of the week or weekday. The model predicts airfare well for the days that are a long way from the takeoff date, anyway for a considerable length of time close the takeoff date, the expectation isn't compelling.

Wohlfarth [15] proposed a ticket buying time enhancement model dependent on an extraordinary pre-preparing step known as macked point processors and information mining systems (arrangement and bunching) and measurable investigation strategy. This system is proposed to change over heterogeneous value arrangement information into added value arrangement direction that can be bolstered to unsupervised grouping calculation. The value direction is bunched into gathering dependent on comparative estimating conduct. Advancement model gauge the value change designs. A treebased order calculation used to choose the best coordinating group and afterward comparing the advancement model.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

A study by Dominguez-Menchero [16] recommends the ideal buying time dependent on nonparametric isotonic relapse method for a particular course, carriers, and timeframe. The model gives the most extreme number of days before buying a flight ticket. two sorts of the variable are considered for the expectation. One is the passage and date of procurement.

III. DATA COLLECTION

The collection of data is the most important aspect of this project. There are various sources of the data on different websites which are used to train the models. Websites give information about the multiple routes, times, airlines and fare. Various sources from API's to consumer travel websites are available for data scraping. In this section details of the various sources and parameters that are collected are discussed.

To implement this data is collected from a website "Makemytrip.com" and python is used for the implementation of the models and collection of the data[12].

A. Collection of data

The script extracts the information from the website and creates a CSV file as output. This file contains the information with features and its details[13]. Now an important aspect is to select the features that might be needed for the flight prediction algorithm. Output collected from the website contains numerous variable for each flight but not all are required, so only the following feature is considered.

- Origin
- Destination
- Departure Date
- Departure Time
- Arrival Time
- Total Fare
- Airways
- Taken Date

In this study, the focus is only on minimizing the airfare charges so a single route is considered without return. This data is collected for one of the busiest routes in India (BOM to DEL) over a period of three months that is from February to April. For each flight data with all the features collected manually.

B. Cleaning and preparing data

All the collected data needed a lot of work so after the collection of data, it is needed to be clean and prepare according to the model requirements. All the unnecessary data is removed like duplicates and null values. In all machine learning this technology, this is the most important and timeconsuming step. Various statistical techniques and logic builtin python are used to clean and prepare the data. For example, the price was character type, not an integer.

C. Analyzing data

Data preparation is followed by analyzing the data, uncovering the hidden trends and then applying various machine learning models. Also, some features can be calculated from the existing feature. Days to departure can be obtained by calculating the difference between the departure date and the date on which data is taken. This parameter is considered to be within 45 days. Also, the day of departure plays an important role in whether it is holiday or weekday. Intuitively the flights scheduled during weekends have a more price compared to the flights on Wednesday or Thursday. Similarly, time also seems to play an important factor. So the time is been divided into four categories: Morning, afternoon, evening, night.

IV. MACHINE LEARNING ALGORITHM

To develop the model for the flight price prediction, many conventional machine learning algorithms are evaluated. They are as follows: Linear regression, Decision tree[8], Random Forest Algorithm[9], K-Nearest neighbors[7], Multilayer Perceptron[10], Support Vector Machine (SVM) [11]and Gradient Boosting. All these models are implemented in the scikit learn. To evaluate the performance of this model, certain parameters are considered. They are as follows: R-squared value, Mean Absolute Error (MAE) and Mean Squared Error (MSE). The formulas for these three parameters are as follows:

$$R^2 = 1 - \frac{\sum_n^{t=1}(yi - \hat{yi})^2}{\sum_n^{t=1}(yi - \bar{yi})^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{n}^{t=1}|yi - \hat{yi}| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{n}^{t=1}(yi - \hat{yi})^2 \quad (3)$$

A. Linear Regression

Regression is a method of modeling a target value based on predictors that are independent. It is mostly based on the number of independent variables and the relationship between independent and dependent variables. linear regression is a type of analysis where the number of independent variables is one and the relationship between the dependent and independent variables vary linearly. The important concept to understand linear regressions are cost function and Gradient decent.

$$y(pred) = b_0 + b_1 * x \quad (4)$$

B. Decision tree

The Decision tree calculation separates the informational collection into small subsets, at a similar same time it creates gradually. The last outcomes are the tree with the decision nodes, what's more, the leaf nodes. A decision hub may have at least two branches. In the beginning, consider the entire informational collection as root. Highlight esteems are wanted to be downright. On the off chance that the qualities are constant then they are discretized before structure the model. Based on characteristic qualities records are dispersed recursively. There are two primary characteristics in the decision tree calculation. One is Information Gain and another is the Gini index. Information Gain is the proportion of Change in entropy. Higher the entropy more the instructive substance, where the entropy is a proportion of vulnerability of arbitrary variable. Gini Index is a component that measures how frequently an arbitrarily picked

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

component would be mistakenly distinguished. It implies a characteristic with a lower Gini index ought to be liked.

C. Random Forest

It is a supervised learning algorithm. The benefit of the random forest is, it very well may be utilized for both characterization and relapse issue which structure most of current machine learning framework. Random forest forms numerous decision trees, what's more, adds them together to get an increasingly exact and stable expectation. Random Forest has nearly the equivalent parameters as a decision tree or a stowing classifier model. It is very simple to discover the significance of each element on the expectation when contrasted with others in this calculation.

The regular component in these techniques is, for the kth tree, a random vector theta k is produced, autonomous of the past random vectors theta 1, ... , theta k-1 however with the equivalent distribution,while a tree is developed utilizing the preparation set and bringing about a classifier. x is an information vector. For a period, in stowing the random vector is created as the includes in N boxes where N is the number of models in the preparation set of information. In random split, choice includes various autonomous random whole numbers between 1 to K. The dimensionality and nature of theata rely upon its utilization in the development of a tree. After countless trees are created, they select the most famous class. These methodology are called as random forests.[6]

D. K-Nearest Neighbours

In regression techniques, the output obtained is an average value of its k nearest neighbors. It is a non-parametric method like SVM. Using some values, results are evaluated and the best performance value is obtained.

E. Multilayer perceptron (MLP)

It is the class of feedforward artificial neural networks. It includes the input layer, output layer and the number of the hidden layers. The hidden layer gives the depth of the neural network. The setup includes 1 hidden layer, the number

TABLE I
ALGORITHMEVALUATION

ML algorithms	R-squared	MAE	MSE
Random forest	0.67	0.08	0.04
Multilayer Perceptron	0.65	0.09	0.04
Gradient Boosting	0.47	0.13	0.06
Decision tree	0.45	0.09	0.06
K-nearest neighbour	0.38	0.14	0.07
SVM	0.19	0.15	0.08
AdaBoost	-0.12	0.21	0.11

of neurons starts from 100 to 2000 with different intervals depending upon the required condition. To fire each neuron it requires activation energy. The logistic sigmoid function is used as an activation function.

F. Gradient boosting

It is an additive regression model by fitting simple function to current "pseudo" residuals sequentially by least-squares at each iteration. It uses the Decision tree as a basic estimator in sci-kit implementation. Starting from 10 to 1000 with the interval of 10 boosting stages are used with maximum numbers. The loss function is an important parameter in the gradient boosting. It can be calculated with options: least squares regression, least absolute deviation, and quantile regression.

The most important feature in the flight pricing prediction is the the day is a holiday or not, the day is weekday or weekend and the difference between the days.

G. Support Vector Machine (SVM)

In the proposed paper Support Vector Machine used as regression analysis that relays on kernel function considered as non parametric technique. The following kernels are used: Linear, Polynomial, Radial Basis Function[10].

As per the previous studies Random forest and the gradient boosting gives the maximum accuracy[7]. The values of R square, MAE and MSE are given in the table:

V. PREDICTORS

After evaluating the performance of the all machine learning models , further improvements are made using a correct predictor model for the best result. Two separated train models are developed by applying the trained datasets. Also the appropriate weights are assigned to them to get a better predictor model[14].

A. Stacked Prediction Model

The performance of machine learning model are evaluated. The Random forest and Multilayer Perceptron, these two models have better results compared to other models. Some weights are applied on the prediction results of these two models to get better prediction results. This is called as stacked prediction model as defined as follows:

1) Naive Bayes Model: Naive Bayes model simply assigns the equal weights to the results of both the models. There are three techniques used for the prediction. They are as follows:

Bernoulli: It is good for making the prediction from binary features.

Gaussian: It is good for making prediction from normally distributed features.

Multinomial: It is good for when the features (categorical or continuous) describe discrete frequency count (e.g. word counts)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (5)$$

The result for the naive method is better than random forest model and multilayer perceptron model.

2) Exhausted search method: This method finds the optimal values for theta by running an exhausting search over a domain of finite interval from -n to n at the difference of a perticular step size.

VI. COCLUSION

In the proposed paper the overall survey for the dynamic price changes in the flight tickets is presented. this gives the information about the highs and lows in the airfares acording to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

the days, weekend and time of the day that is morning, evening and night. also the machine learning models in the computational intelligence feild that are evaluated before on different datasets are studied. their accuracy and performances are evaluated and compared in order to get better result. For the prediction of the ticket prices perfectly differnt prediction models are tested for the better prediction accuracy. As the pricing models of the company are developed in order to maximize the revenue management. So to get result with maximum accuracy regression analysis is used. From the studies , the feature that influences the prices of the ticket are to be considered. In future the details about number of availble seats can improve the performance of the model.

REFERENCES

[1] B. Smith, J. Leimkuhler, R. Darrow, and Samuels, "Yield managementat american airlines,"Interfaces, vol. 22, pp. 8–31, 1992.
[2] W. Groves and M. Gini, "An agent for optimizing airline ticket purchasing," 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), St. Paul, MN, May 06 - 10, 2013 , pp. 1341-1342.
[3] T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," Bachelor Thesis, Radboud University, 2014.
[4] R. Ren, Y. Yang and S. Yuan, "Prediction of airline ticket price, "Technical Report, Stanford University, 2015
[5] M. Papadakis , "Predicting Airfare Prices," 2014.
[6] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32 , 2001.
[7] Viet Hoang Vu, Quang Tran Minh and Phu H. Phung,"An Airfare Prediction Model for Developing Markets", IEEE paper 2018.
[8] S.B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.
[9] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32 , 2001.
[10] S. Haykin, Neural Networks – A Comprehensive Foundation. Prentice Hall, 2nd Edition, 1999.
[11] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression machines," Advances in neural information processing systems, vol. 9, pp. 155-161, 1997.
[12] www.Makemytrip.com
[13] K. Tziridis, Th. Kalampokas, G.A. Papakotas and K.I. Diamantaras,"Airfare Prices Prediction Using Machine Learning Techniques", EUSIPCO 2017.
[14] Yudong Tan and Huimin Zhou,"A Bayesian Predictor of Airline Class Seats Based on Multinomial Event Model",IEEE on Big Data 2016.
[15] Wohlfarth, T. Clemencon, S.Roueff, "A Dat mining approach to travel price forecasting", 10 th international conference on machine learning Honolulu 2011.
[16] Dominguez-Menchero, J.Santo, Reviera, "optimal purchase timing in airline markets" ,2014
[17] Bingchuan Liu, Yudone Tan and Humine Zhou, " A Baysian predictor of Airline class Seats Based on Multinomial Event Model," International conference on Big Data 2016.
[18] K. Tziridis, K.I. Diamantaras,"Airfare Prices Prediction Using machine Learning Technique",European signal processing conference 2017.
[19] Viet Hong Vu, Phu Phung,"An airfare Prediction model for developing Markets",IEEE 2018
[20] William Grooves and Maria Gini,"A regression model for predicting optimal purchase timing for airline tickets",University of Minnesota 2011.
[21] Jun Lu ,"Machine learning modelling for time series problem: Predicting Flight ticket prices",EPFL 2018.
[22] Orenn Etzoni, Craig, Rattapoom and yates,"To buy or not to buy: Mining Data to minimize Ticket purchase Price",SIGKDD ACM 2003.
[23] Qiqi Ren,"When to book:Predicting Flight Pricing", Standford university
[24] Abdella,Zaki,Shuaib and Khan,"Airline ticket price and demand prediction: A survey",Journal of king saud university 2019.
[25] Abhilash, Ranjana,shilpa and Zubeda'"Survey on Air Price Prediction using Machine Learning Algorithm", IJIREEICE 2019 .