Predicting The Price Of A Flight Ticket With The Use Of Machine Learning Algorithms

Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar

Abstract: Someone who purchase flight tickets frequently would be able to predict the right time to procure a ticket to obtain the best deal. Many airlines change ticket prices for their revenue management. The airline may increase the prices when the demand is to be expected to increase the capacity. To estimate the minimum airfare, data for a specific air route has been collected including the features like departure time, arrival time and airways over a specific period. Features are extracted from the collected data to apply Machine Learning (ML) models. This paper gives the machine learning regression methods to predict the prices at the given time.

Index Terms : Machine Learning Algorithm, Predictor, airfare, bagging regression..

1. INTRODUCTION

The flight ticket buying system is to purchase a ticket many days prior to flight takeoff so as to stay away from the effect of the most extreme charge. Mostly, aviation routes don't agree this procedure. Plane organizations may diminish the cost at the time, they need to build the market and at the time when the tickets are less accessible. They may maximize the costs. So the cost may rely upon different factors. To foresee the costs this venture uses AI to exhibit the ways of flight tickets after some time. All organizations have the privilege and opportunity to change it's ticket costs at anytime. Explorer can set aside cash by booking a ticket at the least costs. People who had travelled by flight frequently are aware of price fluctuations. The airlines use complex policies of Revenue Management for execution of distinctive evaluating systems [1]. The evaluating system as a result changes the charge depending on time, season, and festive days to change the header or footer on successive pages. The ultimate aim of the airways is to earn profit whereas the customer searches for the minimum rate. Customers usually try to buy the ticket well in advance of departure date so as to avoid hike in airfare as date comes closer. But actually this is not the fact. The customer may wind up by giving more than they ought to for the same seat.

2 LITERATURE SURVEY

It is hard for the client to buy an air ticket at the most reduced cost. For this few procedures are explored to determine time and date to grab air tickets with minimum fare rate. The majority of these systems are utilizing the modern computerized system known as Machine Learning. To determine ideal purchase time for flight ticket Gini and Groves[2] exploited Partial Least Square Regression(PLSR) for building up a model. The information was gathered from major travel adventure booking sites from 22 February 2011 to 23 June 2011. Extra information was additionally gathered and are utilized to check the correlations of the exhibitions of the last model. Janssen [3] implemented a desire model using the Linear Quantile Blended Regression methodology for San Francisco–New York course where each day airfares are given by www.infare.com. Two features such as number of days for departure and whether departure is on weekend or weekday are considered to develop the model. The model guesses airfare well in advance from the departure date. But the model isn't convincing in a situation for an extensive time allotment, it closes the departure date. Wohlfarth [10] proposed a ticket purchasing time improvement model subject to a significant pre-processing known as macked point processors, data mining frameworks (course of action and grouping) and quantifiable examination system. This framework is proposed to change various added value arrangements into included added value arrangement heading which can support to solo gathering estimation. This value heading is packed into get-together reliant on near evaluating conduct. Headway model measure the value change plans. A tree-based analysis used to pick the best planning gathering and a short time later looking at the progression model. An investigation by Dominguez-Menchero [11] suggests the perfect purchase timing reliant on a nonparametric isotonic backslide technique for a specific course, carriers, and time frame. The model provides the most acceptable number of days before buying the flight ticket. The model considers two types of a variable such as the entry and is date of obtainment.

3 DATA COLLECTION

The accumulation of information is the most significant part of this venture. The different wellsprings of the information on various sites are utilized to prepare the models. Sites provide data about the numerous courses, times, aircrafts and charge. Different sources from API's to customer travel sites are accessible for information scratching. In this segment information of the different sources and parameters that are gathered are talked about. To verify this, information is collected from "Makemytrip.com" site and the models are implemented using python [9].

3.1 Data Collection

The python-script take out the data from the site, and provides

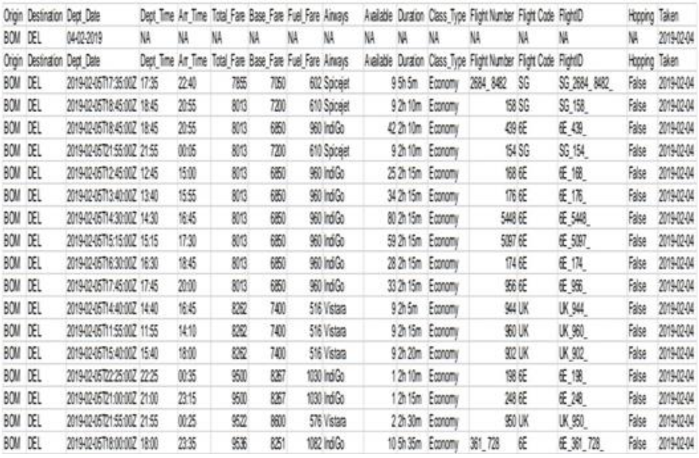
- ***Supriya Rajankar*** *is Professor in Electronics and Telecommunication department, Sinhgad College of Engineering Vadgaon(bk), Pune, India Email: Supriya.rajankar@gmail.com*
- ***Neha Sakharkar*** *is currently pursuing masters degree in Electronics and Telecommunication department, Sinhgad College of Engineering Vadgaon(bk), Pune, India ,. E-mail: nehasakharkar22e@gmail.com*
- ***Omprakash Rajankar*** *is Professor in Electronics and Telecommunication department, Dhole Patil College of Engineering, Pune, India Email: online.omrajankar@gmail.com*

output as a CSV record. The document contains the data with features and its details [9]. A significant perspective is to choose the features required for calculation of expected flight price. Output gathered from the site contains number of parameters for each flight: yet not all are required, so just the accompanying components are,

- Date of journey
- Time of Departure
- Place of Departure
- Time of Arrival
- Place of Destination/Arrival
- Airway company
- Total Fare

In this investigation, the attention is just to limit the airfare considering a single route. This information is gathered for perhaps the busiest course in India (BOM to DEL) over a time of a quarter of a year that is from February to April. For each flight information each feature is collected physically.

TABLE I *: COLLECTED DATASET*



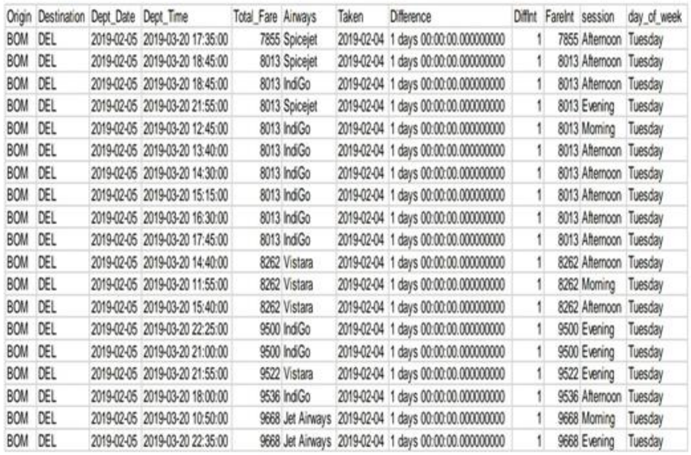
Origin	Destination	Dept_Date	Dept_Time	Arr_Time	Total_Fare	Base_Fare	Fuel_Fare	Airways	Available	Duration	Class_Type	Flight Number	Flight Code	FlightID	Hopping	Taken
BOM	DEL	04-02-2019	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2019-02-04
Origin	Destination	Dept_Date	Dept_Time	Arr_Time	Total_Fare	Base_Fare	Fuel_Fare	Airways	Available	Duration	Class_Type	Flight Number	Flight Code	FlightID	Hopping	Taken
BOM	DEL	2019-02-05T17:35:00Z	17:35	22:40	7855	7050	602	Spicejet	9	5h 5m	Economy	2684_8482	SG	SG_2684_8482_	False	2019-02-04
BOM	DEL	2019-02-05T18:45:00Z	18:45	20:55	8013	7200	610	Spicejet	9	2h 10m	Economy	158	SG	SG_158_	False	2019-02-04
BOM	DEL	2019-02-05T18:45:00Z	18:45	20:55	8013	6850	960	IndiGo	42	2h 10m	Economy	439	6E	6E_439_	False	2019-02-04
BOM	DEL	2019-02-05T21:55:00Z	21:55	00:05	8013	7200	610	Spicejet	9	2h 10m	Economy	154	SG	SG_154_	False	2019-02-04
BOM	DEL	2019-02-05T12:45:00Z	12:45	15:00	8013	6850	960	IndiGo	25	2h 15m	Economy	168	6E	6E_168_	False	2019-02-04
BOM	DEL	2019-02-05T13:40:00Z	13:40	15:55	8013	6850	960	IndiGo	34	2h 15m	Economy	176	6E	6E_176_	False	2019-02-04
BOM	DEL	2019-02-05T14:30:00Z	14:30	16:45	8013	6850	960	IndiGo	80	2h 15m	Economy	5448	6E	6E_5448_	False	2019-02-04
BOM	DEL	2019-02-05T15:15:00Z	15:15	17:30	8013	6850	960	IndiGo	59	2h 15m	Economy	5097	6E	6E_5097_	False	2019-02-04
BOM	DEL	2019-02-05T16:30:00Z	16:30	18:45	8013	6850	960	IndiGo	28	2h 15m	Economy	174	6E	6E_174_	False	2019-02-04
BOM	DEL	2019-02-05T17:45:00Z	17:45	20:00	8013	6850	960	IndiGo	33	2h 15m	Economy	956	6E	6E_956_	False	2019-02-04
BOM	DEL	2019-02-05T14:40:00Z	14:40	16:45	8262	7400	516	Vistara	9	2h 5m	Economy	944	UK	UK_944_	False	2019-02-04
BOM	DEL	2019-02-05T11:55:00Z	11:55	14:10	8262	7400	516	Vistara	9	2h 15m	Economy	960	UK	UK_960_	False	2019-02-04
BOM	DEL	2019-02-05T15:40:00Z	15:40	18:00	8262	7400	516	Vistara	9	2h 20m	Economy	902	UK	UK_902_	False	2019-02-04
BOM	DEL	2019-02-05T22:25:00Z	22:25	00:35	9500	8267	1030	IndiGo	1	2h 10m	Economy	198	6E	6E_198_	False	2019-02-04
BOM	DEL	2019-02-05T21:00:00Z	21:00	23:15	9500	8267	1030	IndiGo	1	2h 15m	Economy	248	6E	6E_248_	False	2019-02-04
BOM	DEL	2019-02-05T21:55:00Z	21:55	00:25	9522	8600	576	Vistara	2	2h 30m	Economy	950	UK	UK_950_	False	2019-02-04
BOM	DEL	2019-02-05T18:00:00Z	18:00	23:35	9536	8251	1082	IndiGo	10	5h 35m	Economy	361_728	6E	6E_361_728_	False	2019-02-04

Table I shows the original dataset obtained from makemytrip.com. It is basically raw data containing all the features. This data has been collected for single route only.

3.2 Cleaning and preparing data

All the gathered information required a great deal of work, so after the accumulation of information, it should have been perfect and be ready as indicated by the model prerequisites. All the superfluous information is deleted like copies and invalid qualities. In all AI this innovation, is the most significant and time consuming. Different statistical methods and logics in python clean and set up the information. For instance, the cost was character type, not a number.

TABLE II : *CLEAN AND PREPARED DATASET.*



Origin	Destination	Dept_Date	Dept_Time	Total_Fare	Airways	Taken	Difference	DiffInt	FareInt	session	day_of_week
BOM	DEL	2019-02-05	2019-03-20 17:35:00	7855	Spicejet	2019-02-04	1 days 00:00:00.000000000	1	7855	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 18:45:00	8013	Spicejet	2019-02-04	1 days 00:00:00.000000000	1	8013	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 18:45:00	8013	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	8013	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 21:55:00	8013	Spicejet	2019-02-04	1 days 00:00:00.000000000	1	8013	Evening	Tuesday
BOM	DEL	2019-02-05	2019-03-20 12:45:00	8013	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	8013	Morning	Tuesday
BOM	DEL	2019-02-05	2019-03-20 13:40:00	8013	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	8013	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 14:30:00	8013	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	8013	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 15:15:00	8013	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	8013	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 16:30:00	8013	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	8013	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 17:45:00	8013	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	8013	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 14:40:00	8262	Vistara	2019-02-04	1 days 00:00:00.000000000	1	8262	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 11:55:00	8262	Vistara	2019-02-04	1 days 00:00:00.000000000	1	8262	Morning	Tuesday
BOM	DEL	2019-02-05	2019-03-20 15:40:00	8262	Vistara	2019-02-04	1 days 00:00:00.000000000	1	8262	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 22:25:00	9500	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	9500	Evening	Tuesday
BOM	DEL	2019-02-05	2019-03-20 21:00:00	9500	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	9500	Evening	Tuesday
BOM	DEL	2019-02-05	2019-03-20 21:55:00	9522	Vistara	2019-02-04	1 days 00:00:00.000000000	1	9522	Evening	Tuesday
BOM	DEL	2019-02-05	2019-03-20 18:00:00	9536	IndiGo	2019-02-04	1 days 00:00:00.000000000	1	9536	Afternoon	Tuesday
BOM	DEL	2019-02-05	2019-03-20 10:50:00	9668	Jet Airways	2019-02-04	1 days 00:00:00.000000000	1	9668	Morning	Tuesday
BOM	DEL	2019-02-05	2019-03-20 22:35:00	9668	Jet Airways	2019-02-04	1 days 00:00:00.000000000	1	9668	Evening	Tuesday

Dataset in table II shows the information required for the analysis of the data. Additional features are created to get more accurate results. Feature columns like day of week and session are generated to analyze the data on the basis of time duration of the day and other factors.

3.3 Analyzing data

Preparation of data is trailed by breaking down the information, revealing the concealed patterns and afterward applying different AI models. Likewise, a few features can be determined from the current features. Flight days can be issued by computing the difference of the flight date and the date on which information is collected. This can be observed for 45 days. Additionally, flight date is important, whether it is on festive day or a weekday or weekend. Instinctively the flights planned during weekends cost more than the flights on weekdays. Additionally, time plays important role. So the time is considered in classes as: Morning, evening and night[13].

4 MACHINE LEARNING MODEL PERFORMANCE

For predicting the flight ticket prices, many algorithms are introduced in machine learning. The algorithms are: Support Vector Machine (SVM) [8], Linear regression, K-Nearest neighbours [4], Decision tree [5], Multilayer Perceptron[7], Gradient Boosting and Random Forest Algorithm[6]. Using python library scikit learn these models have been implemented. The parameters like R-square, MAE and MSE are considered to verify the performance of these models.

4.1 Linear Regression

To determine the correlation between two continuous variables, simple linear regression analysis is used. One of the two variables is the predictor variable of which value is to be found. It gives the statistical relationship not the deterministic relationship between two variables. Linear regression algorithm gives the best fit line to the given data for which the prediction error is minimum. Gradient descent and cost function are the two major factors to understand linear regression. The equation for linear regression is :

$$y(pred) = b_0+b_1 * x \qquad (1)$$

The value of coefficients b_1 and b_0 are chosen so that the error value is as small as possible. The square of predicted and actual value difference gives the error. To deal with the negative values, the mean square error is taken (MSE). Here b_0 gives the positive or negative relationship between the x and y, whereas b_1 is called bias. The accuracy of the regression problem is measured in terms of R-squared, MAE, and MSE.

4.2 Decision Tree

This tree count isolates the collected information into small subsets, at a comparative same time makes it persistent. The last results show the tree having the decision centres, likewise, the leaf centres. This decision centre point may contain two branches at any rate. At first think about the whole informational index as root. Feature regards are kicked out of the chance. In case the characteristics are relentless then they have to be discretized before structuring the model. In view of estimation property records are corrected recursively. Information Gain and Gini index are two essential properties in the decision of tree computation. Information Gain is defined as the amount change in entropy. Higher entropy indicates more effectiveness of the substance. Thus the entropy is a proportion of vulnerability of arbitrary variable. Gini Index measures how regularly an arbitrarily picked component would be falsely recognized. It implies a characteristic with a lower Gini index ought to be liked. For Regression tree, cost capacity can be a basic squared condition:

$$E = \sum (y - \hat{y})^2 \qquad (2)$$

Where y is the actual value from the dataset and y cap is predicted value. Having a class with the maximum number of an expected value obtained by the split function called information gain. If the class is kept splitting and splitting without any condition at the leaf node, the algorithm will be huge, slow and over fitted. To stop this, a minimum count on the training example on the leaf node is assigned.

4.3 Support Vector Machine (SVM)

The SVM is supervised ML algorithm is used for classification and regression analysis. It takes a large time to process so usually applied to the small datasets. It finds the Hyperplane that separates the feature into different parts. It gives optimal hyperplane which classifies the different domains. The data points which are nearest to hyperplane are called support vector points and distance between the vector plane and these points are called as margins.

$$y = w_0 + \sum_{i=0}^{m} w_i x_i \qquad (3)$$

The proposed work has exploited SVM for regression analysis. The performance depends on kernel function selection as a nonparametric technique. Linear, Radial Basis Function and Polynomial [7] are the kernels of support vector machine algorithm.

4.4 K-Nearest Neighbours(KNN)

In k-nearest neighbour regression analysis, the output is mean of its k nearest neighbours. Like SVM this is also a non-parametric method. Considering few values, results are computed to achieve the best value. KNN is a supervised classification algorithm that can also be used as a regressor. It assigns a new data point to the class. It is non-parametric because it does not take any assumption. It calculates the distance between every training example and a new data point. To compute this distance following distance calculation methods are used:

- Euclidean Distance

$$ED = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \qquad (4)$$

- Manhatten distance

$$MD = \sum_{i=1}^{k} |x_i - y_i| \qquad (5)$$

- Hamming distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i| \qquad (6)$$

K- entries in the dataset are picked by the model that are close to new data point.

4.5 Random Forest

This is an algorithm which ensembles the less predictive model to produce better predictive models. It aggregates the base model to create a large model. The features are sampled and passed to trees without replacement to obtain the highly uncorrelated decision trees. To select the best split it is required to have less correlation between the trees. The main concept that makes random forest different from the decision tree is aggregated uncorrelated trees.

4.6 Bagging Regression Tree

The drawback of decision trees is, it has a large bias with simple trees and large variance with complex trees. Bagging comes from Bootstrap aggregating, a method of selecting a random number of data from a dataset with replacement. It is mostly used to reduce the variance of the tree. From literature it is clear that maximum accuracy is achieved by gradient boosting and random forest methods [4][12].

5 EXPERIMENTAL RESULTS

For the selected test data set , output of the model is plotted across the test dataset. Graph shows the comparative study of original values and predicted results. By the analysis of the results obtained from the algorithm such as SVM, Decision Tree, KNN, Bagging Tree, Random Forest and Linear regression gives the predicted values of the fare to purchase the flight ticket at the right time. Table I gives the values for R-Square. The graph is plotted between the days left until departure verses the fare of the flight. The blue color line denotes the actual value of the flight ticket whereas the red color line shows the predicted value of the flight tickets. Decision Tree algorithm has more accuracy compared to other algorithms for the given dataset. Figure 3 shows the plot between Days remaining for the departure vs. Actual and predicted values evaluated by the Random Algorithm. It gives the highest R-Square value with maximum accuracy in the regression analysis. The table III shows R-square, MSE and MAE values.

TABLE III : *ALGORITHM EVALUATION*

Machine Learning(ML)	R-squared	MAE	MSE

algorithms			
Decision tree	0.67	0.13	0.21
Random forest	0.68	0.13	0.21
K-NN	0.65	0.13	0.22
Linear Regression	0.40	0.19	0.29

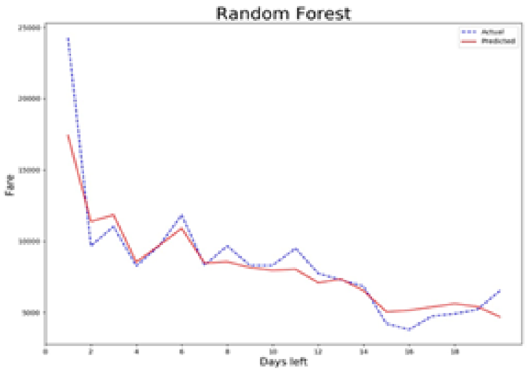


Fig. 3 *Graphical results for Random Forest.*

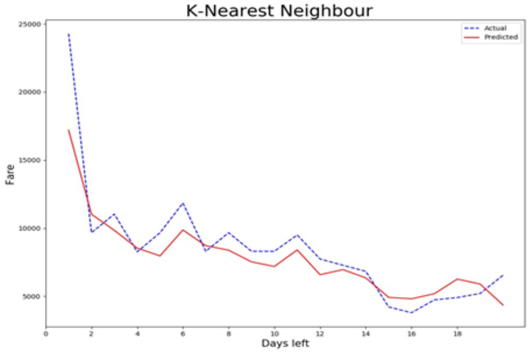


Fig. 4 *Graphical results for K-Nearest Neighbour.*

Figure 4 shows the plot between Days remaining for the departure vs. Actual and predicted values evaluated by the KNN. It gives the R-Squared value near to 1 giving maximum accuracy. Considering all the features like day time, week day and days left for departure from the datasets, predict the airfare results. Among all these features days left until departure has the greatest impact on the prediction of the fare.

6 CONCLUSION AND FUTURE SCOPE

To evaluate the conventional algorithm, a dataset is built for route BOMBAY to DELHI and studied a trend of price variation for the period of limited days. Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. This gives the predicted values of flight fare to get a flight ticket at minimum cost. Data is collected from the websites which sell the flight tickets so only limited information can be accessed. The values of R-squared obtained from the algorithm give the accuracy of the model. In the future, if more data could be accessed such as the current availability of seats, the predicted results will be more accurate.

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] Viet Hoang Vu, Quang Tran Minh and Phu H. Phung,"An Airfare Prediction Model for Developing Markets", IEEE paper 2018.

[5] S.B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32 , 2001.

[7] S. Haykin, Neural Networks – A Comprehensive Foundation. Prentice Hall, 2nd Edition, 1999.

[8] 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.

[9] www.Makemytrip.com

[10] Wohlfarth, T. Clemencon, S.Roueff, "A Dat mining approach to travel price forecasting", 10 th international conference on machine learning Honolulu 2011.

[11] Dominguez-Menchero, J.Santo, Reviera, "optimal purchase timing in airline markets" ,2014

[12] Supriya Rajankar and Neha Sakharkar, "A Survey on Flight Pricing Prediction using MachineLearning", International Journal of Engineering Research and Technology, vol 8, issue 6, June 2019.

[13] K. Tziridis, Th. Kalampokas, G.A. Papakotas and K.I. Diamantaras,"Airfare Prices Prediction Using Machine Learning Techniques", EUSIPCO 2017.