Air Flight Ticket Price Prediction Using Flight Number As Reference

Syed Nehal Hassan Shah
Email: nehal47hassan@Gmail.com

***Abstract*—Many different fields have employed machine learning. We describe how machine learning can be used to solve the time-series problem. We use the airline ticket prediction problem as our specific problem. Based on data over more than 60 days period, we trained our models, getting the best model - which in our case is Random Forest and in case of classification it is Decision tree. These algorithms has the best performance over the observed 12 routes, which is in Regression case 94% closer to test values and 99% accurate in case of Classification .This predicting Machine learning model is more reliable than the random purchase strategy, and relatively small error of 6% over these routes for predicting price of tickets. Our findings demonstrate that utilizing these models and strategies to guide purchase policies can help both sides (buyer and Seller) when deciding what should be the purchase costs of the ticket, both for seller (Airline companies) buyer (Target Audience) before the departure of the flight. The suggested approach can also outperform a deployed commercial website offering comparable purchase policy advice.**

I. Introduction

The typical buying plan for airline tickets is to acquire a ticket far in advance prior to the departure date in order to avoid the risks that costs may increase significantly before the flight's departure date. However, this is not always the case because occasionally, airlines will cut their pricing in an effort to increase sales. When determining ticket prices, airlines take into account a wide range of variables, including the month of travel, the number of seats on the plane, and whether or not it is a holiday. Some of the variables are buried, but others are clearly visible. The contrary is also true—airlines companies want to maintain a high level of overall income. In this case, buyers are looking for the ideal day to buy a ticket [1]. This project intends to use machine learning techniques to model the behavior of airline ticket prices over time. Airlines have complete discretion over when and how much to charge for tickets. If one buys a ticket at the cheapest price, then, he might be able to save money. Finding the best time to book a flight for the chosen destination and term is the difficult part which in our case is our main goal. Given historical pricing, and the current price of a Flight, our algorithms must determine if it is better to buy or wait for another day before the departure date.For the creation and evaluation of the model, we use historical data on the cost of individual aircraft routes with reference to Air-Line Numbers [2].

II. Related work

Some work has been done for determining optimal purchasing time for airline Tickets. Our work is especially inspired by Etzioni et al and Jun Lu.Etzioni in his research he achieve an accuracy to predict flight ticket with 61.8 percent in 2017 [3].Followed by his research Jun Lu in his research achieve an accuracy of 61.35% on prediction Flight ticket prices with edition to predict price of flight which are non-direct flight [4]. Recently a project carried out by the students in which they achieved an accuracy of 64% on predicting flight ticket prices. Two more research carried out in India with the name of Implementation of Flight Fare Prediction System Using Machine Learning[2022] and Flight Price Prediction: A Case Study [2022] [5].These two researches use data sets which are city-to city flight-dependent. The most common approach in all the research is decision tree algorithms and its subsets like random forest,ada boost decision tree. In our case, unlike previous researches. We also considered such a situation in which flight numbers are known, which further leads to historical data of the flight on all possible routes through which the price of the ticket is predicted on the possible route which is entered by user. Due to the fact that we do not need to train our model with a big amount of data again, it can reduce computation time when we want to anticipate whether to buy or wait quickly.Unlike previous researches this model consisted of two task classification of flight-Numbers and second is predicting price.

III. Methodology

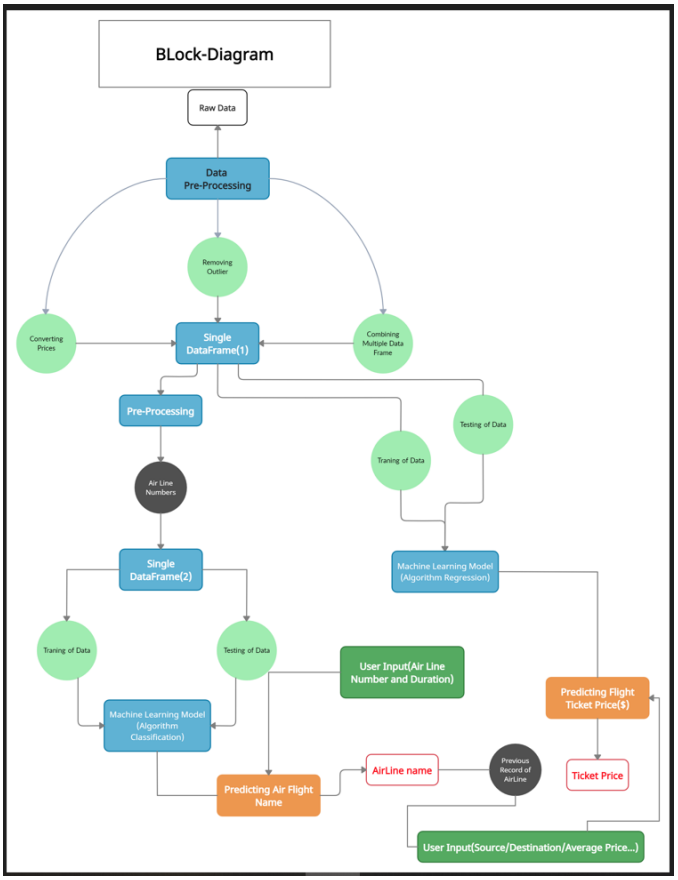

Fig. 1. Block Diagram

Our issue is divided into two issues: the first is to create a function and train a model to anticipate future prices. The second is to allocate numbers to fight against flight names. At first, we applied the preprocessing procedures and produced a data Frame (1). Then, this data Frame goes through a testing and training process. In these two stages, 30% of the data was used for testing and 70% for training. An algorithm for machine learning is applied on the data Frame(1). This will result in a train model that can forecast future ticket prices [6]. For our second part we lake a feature in our data set which could represent flight numbers. This custom made feature was not actually available in our data set when we performed our first task.So in our second task, which involved classifying Air flight numbers,we add a number next to each flight name to symbolise flights and flight with identical names has same numbers while also adding time conversion in minutes for each flight.These steps result in the development of a Data Frame (2).This data Frame(2) then send to training and test phase [7]. After that a machine learning algorithm is applied which can do classification of flight name when flight number is entered as shown in combine result section.This complete process can be viewed in fig:1.

IV. Data Collection

This data set is available on GitHub.It consist of 55,365 rows, 7 columns and divided in 12 CVS files. The data for our analysis was collected as daily price quotes from a major airplane search web site between Feb, 2022 and April,2022 (60+ observe days)fig:2. A web crawler was used to query for each route and departure date pair, and the crawling was done every day at 10:00 AM.

	Airline	Source	Destinatio	Duration	Total stop	Price	Date
2	Lufthansa	PAR	RUH	9h 10m	1 stop	1,575Â SAI	2/1/2022
3	SAUDIA	PAR	RUH	5h 50m	nonstop	2,168Â SAI	2/1/2022
4	Ryanair, fh	PAR	RUH	38h 05m	3 stops	1,069Â SAI	2/1/2022
5	Lufthansa	PAR	RUH	9h 10m	1 stop	1,544Â SAI	2/1/2022
6	Lufthansa	PAR	RUH	10h 10m	1 stop	1,544Â SAI	2/1/2022
7	Egypt Air	PAR	RUH	10h 50m	1 stop	1,599Â SAI	2/1/2022
8	Transavia	PAR	RUH	10h 55m	1 stop	1,720Â SAI	2/1/2022
9	Turkish Air	PAR	RUH	9h 05m	1 stop	1,811Â SAI	2/1/2022
10	Turkish Air	PAR	RUH	11h 55m	1 stop	1,811Â SAI	2/1/2022
11	Turkish Air	PAR	RUH	16h 40m	1 stop	1,811Â SAI	2/1/2022
12	Air Europa	PAR	RUH	10h 55m	1 stop	1,960Â SAI	2/1/2022
13	British Airv	PAR	RUH	9h 00m	1 stop	2,018Â SAI	2/1/2022
14	Iberia Expr	PAR	RUH	13h 00m	1 stop	1,854Â SAI	2/1/2022
15	Transavia	PAR	RUH	10h 10m	1 stop	1,997Â SAI	2/1/2022
16	British Airv	PAR	RUH	10h 10m	1 stop	2,018Â SAI	2/1/2022
17	British Airv	PAR	RUH	13h 35m	1 stop	2,018Â SAI	2/1/2022
18	Iberia, SAL	PAR	RUH	10h 45m	1 stop	2,024Â SAI	2/1/2022
19	Etihad Airv	PAR	RUH	11h 00m	1 stop	2,098Â SAI	2/1/2022
20	Emirates	PAR	RUH	10h 00m	1 stop	2,300Â SAI	2/1/2022
21	Emirates	PAR	RUH	14h 25m	1 stop	2,300Â SAI	2/1/2022
22	Emirates	PAR	RUH	15h 35m	1 stop	2,300Â SAI	2/1/2022
23	Emirates	PAR	RUH	16h 25m	1 stop	2,300Â SAI	2/1/2022
24	Emirates	PAR	RUH	19h 15m	1 stop	2,300Â SAI	2/1/2022
25	SAUDIA	PAR	RUH	10h 00m	1 stop	2,344Â SAI	2/1/2022
26	SAUDIA	PAR	RUH	12h 55m	1 stop	2,344Â SAI	2/1/2022
27	SAUDIA	PAR	RUH	13h 55m	1 stop	2,344Â SAI	2/1/2022
28	SAUDIA	PAR	RUH	14h 55m	1 stop	2,344Â SAI	2/1/2022
29	SAUDIA	PAR	RUH	15h 55m	1 stop	2,344Â SAI	2/1/2022

Fig. 2. Features

V. Experimental Results

Our routs consist of 4 destination and their total trips are shown in table:1.We found that the ticket price for flights can vary significantly over time.Prices may differ from direct routes as compared to multi routes.In this table,we can have an overall glance at total trips starting from Paris ,New York , Russia and Saudi Arabia shown in terms of their frequency fig:3.These total trips consist of nonstop flight to up the flight consist of 3 stops.Their total sum is shown in table:1.

PAR	SVO	NYC	RUH
14881	4202	5334	7279
2403	2235	1905	553
7327	3314	3205	2725
22208	97511	10444	10557

TABLE I
12 Routes

A. Prepossessing

This data set is available on GitHub.It consist of 55,365 rows, 7 columns and divided in 12 CVS files. The data for our analysis was collected as daily price quotes from a major airplane search web site between Feb, 2022 and April,2022 (60+ observe days) fig:4. A web crawler was used to query for each route and departure date pair, and the crawling was done every day at 10:00 AM.

B. Identifying Outliers IQR Method

We have created a ”fence” outside of Q1 and Q3 using the IQR method of finding outliers. Outliers are any values that lie outside of this range. To construct this fence, we multiply

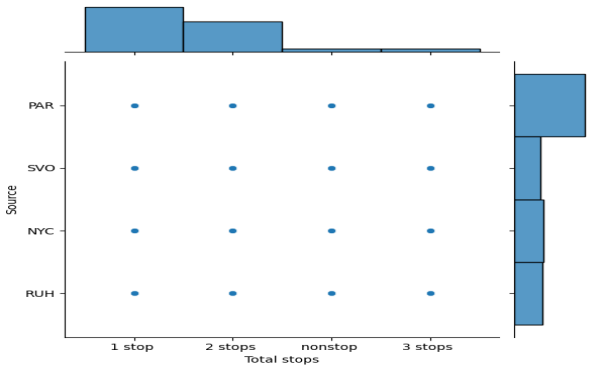

Fig. 3. Frequency OF Stops

	Airline	Source	Destinatio	Duration	Total stop	Price	Date
2	Lufthansa	PAR	RUH	9h 10m	1 stop	1,575Â SAI	2/1/2022
3	SAUDIA	PAR	RUH	5h 50m	nonstop	2,168Â SAI	2/1/2022
4	Ryanair, fl	PAR	RUH	38h 05m	3 stops	1,069Â SAI	2/1/2022
5	Lufthansa	PAR	RUH	9h 10m	1 stop	1,544Â SAI	2/1/2022
6	Lufthansa	PAR	RUH	10h 10m	1 stop	1,544Â SAI	2/1/2022
7	Egypt Air	PAR	RUH	10h 50m	1 stop	1,599Â SAI	2/1/2022
8	Transavia	PAR	RUH	10h 55m	1 stop	1,720Â SAI	2/1/2022
9	Turkish Air	PAR	RUH	9h 05m	1 stop	1,811Â SAI	2/1/2022
10	Turkish Air	PAR	RUH	11h 55m	1 stop	1,811Â SAI	2/1/2022
11	Turkish Air	PAR	RUH	16h 40m	1 stop	1,811Â SAI	2/1/2022
12	Air Europa	PAR	RUH	10h 55m	1 stop	1,960Â SAI	2/1/2022
13	British Airv	PAR	RUH	9h 00m	1 stop	2,018Â SAI	2/1/2022
14	Iberia Expr	PAR	RUH	13h 00m	1 stop	1,854Â SAI	2/1/2022
15	Transavia	PAR	RUH	10h 10m	1 stop	1,997Â SAI	2/1/2022
16	British Airv	PAR	RUH	10h 10m	1 stop	2,018Â SAI	2/1/2022
17	British Airv	PAR	RUH	13h 35m	1 stop	2,018Â SAI	2/1/2022
18	Iberia, SAL	PAR	RUH	10h 45m	1 stop	2,024Â SAI	2/1/2022
19	Etihad Airv	PAR	RUH	11h 00m	1 stop	2,098Â SAI	2/1/2022
20	Emirates	PAR	RUH	10h 00m	1 stop	2,300Â SAI	2/1/2022
21	Emirates	PAR	RUH	14h 25m	1 stop	2,300Â SAI	2/1/2022
22	Emirates	PAR	RUH	15h 35m	1 stop	2,300Â SAI	2/1/2022
23	Emirates	PAR	RUH	16h 25m	1 stop	2,300Â SAI	2/1/2022
24	Emirates	PAR	RUH	19h 15m	1 stop	2,300Â SAI	2/1/2022
25	SAUDIA	PAR	RUH	10h 00m	1 stop	2,344Â SAI	2/1/2022
26	SAUDIA	PAR	RUH	12h 55m	1 stop	2,344Â SAI	2/1/2022
27	SAUDIA	PAR	RUH	13h 55m	1 stop	2,344Â SAI	2/1/2022
28	SAUDIA	PAR	RUH	14h 55m	1 stop	2,344Â SAI	2/1/2022
29	SAUDIA	PAR	RUH	15h 55m	1 stop	2,344Â SAI	2/1/2022

Fig. 4. Features

the IQR by 1.5, deduct Q1 from this result, and add Q3 to the result. We can compare each observation to the minimum and maximum fence posts as a result. Outliers are any observations that fall more than 1.5 IQR outside Q1 or rise more than 1.5 IQR outside Q3. Here is how our data appears fig:5.

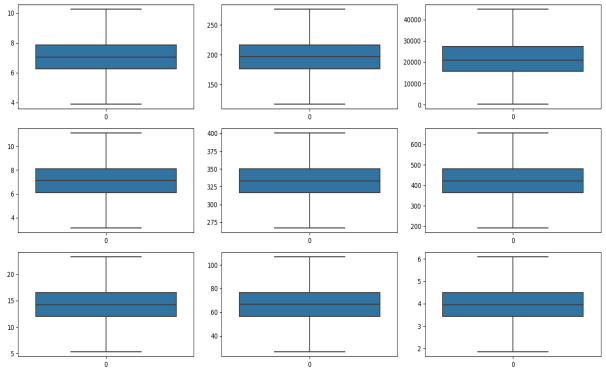

Fig. 5. Outliers

C. Regression Algorithms

Regression is a subset of Supervised Learning. It learns a model based on a training data-set to make predictions about unknown or future data. The description 'supervised' comes from the fact that the target output value is already defined and part of the training data.In our case our label data was number of stops and their ticket price at that time [8].We want to predict the prices accordingly against our label data.Best performance in case for prediction prices of ticket were given by Decision Tree and its sub types.Overall highest performance was give by Random Forest observed in the table:2.

No1	Decision Tree	Random Forest	Extra DT	Bagging DT	AdaBoost DT	KNN
MSE $^2	43195	39727	39673	40313	148130	57967
MAE $	61.11	61.87	59.7	62.4	242.7	75.3
RMSE $	207.8	199.3	199.1	200.7	384.8	240.7

TABLE II
ERROR EXPECTATION ON AVERAGE

1) MAE MSE RMSE: The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the data set. It measures the average of the residuals in the data set.

$$MAE = \frac{\sum_{i=1}^{D} |x_i - y_i|}{N}$$

Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$MSE = \frac{\sum_{i=1}^{D} (x_i - y_i)^2}{N}$$

Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

$$RMSE = \sqrt{MSE}$$

Variance between Test value and predict are shown in table:3.

No1	Decision Tree	Random Forest	Extra DT	Bagging DT	AdaBoost DT	KNN
Variance%	94.06	94.55	94.54	94.45	79.72	92.04

TABLE III
PREDICTED RESEMBLANCE TO TEST POINT

2) Training and Testing Accuracy: In addition to measuring error we also measure accuracy in terms of X-train against X-test and Y-train vs Y-test as shown in table:4.

No1	Decision Tree	Random Forest	Extra DT	Bagging DT	AdaBoost DT	KNN
X-Train vs Y-Train %	96.42	96.42	96.42	96.41	96.41	96.40
X-Test vs Y-Test %	94.53	94.50	94.49	94.54	94.56	94.52

TABLE IV
X-TRAIN VS X-TEST AND Y-TRAIN VS Y-TEST

D. Classification Algorithms

On the basis of training data, the Classification algorithm is a Supervised Learning technique that is used to categorise new observations. In classification, a programmer makes use of the data set or observations that are provided to learn how to categorise fresh observations into various classes or groups. For instance, cat or dog, yes or no, 0 or 1, spam or not spam, etc. Targets, labels, or categories can all be used to describe classes.In our case Decision Tree preform with 99% accuracy predicting flight name on the bases of flight reference which then lead to the flight history and results are shown in table:5.

No1	Decision Tree	Random Forest	Extra DT	Bagging DT	AdaBoost DT	KNN
Accuracy%	99	98	98	99	27	92

TABLE V
PREDICT OF FLIGHT NAME

1) Accuracy: The accuracy of our algorithm demonstrates how accurate the values we would expect to see if we had entered the flight number to locate the flight name and its historical information.In initial stage the accuracy of algorithm was about 97% we add duration which we converted in prepossessing step in minutes using pandas direct command.This step is done to to achieve an accuracy of 99%.Their relationship is shown in fig:6.

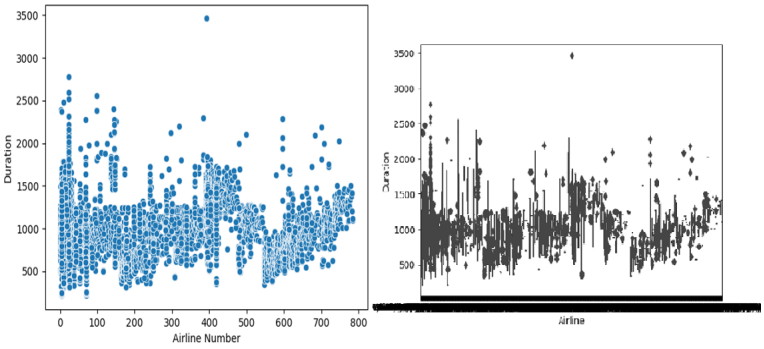

Fig. 6. Duration Against Flight

VI. Data and information visualization

Anyone can better understand data by displaying it visually and meaningfully through information visualisation. Common information visualisation examples include dashboards and scatter plots. Information visualisation enables people to efficiently and effectively derive meaning from abstract data by providing an overview and important relationships.

A. Visualization in Regression

So from the fig:7 we can visualize how price increase specially when When we plot the average price rise against the actual cost of an airline ticket, we can see how prices are rising. Second, we see in our plot of price vs duration that prices rise in a logical order as travel times to a certain destination lengthen [9].

1) Random Forest: A group of decision trees is called a random forest. This means that a Random Forest is made up of several trees that were built in a particular ”random” manner.

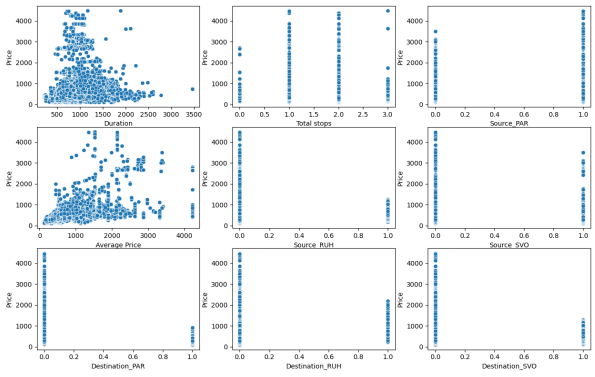

Fig. 7. Regression

B. Visualization in Classification

So from the fig:8 we can visualize the relationship between airline number and other features.So through visualization we can add features who are more depended on airline number and select those features to solve classification problem.In the experiment we added duration and flight number to predict flight name with an accuracy of 99%.

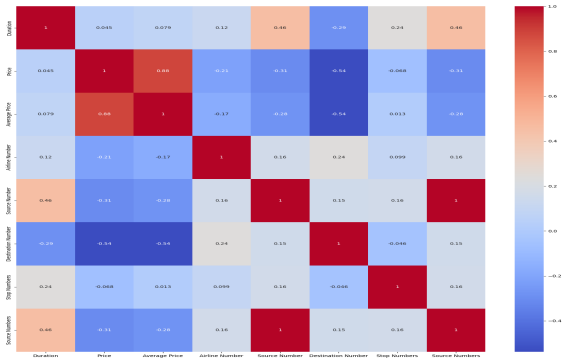

Fig. 8. Classification

1) Decision Tree: The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch an outcome of result, and each leaf node (terminal node) a class label. A separate sample of rows are used to build each tree, and a different sample of characteristics are chosen for splitting at each node. Each tree provides a unique prediction on its own. A single outcome is then produced by averaging these predictions [7] [10].

$$Gini = 1 - \sum_{i=1}^{c}(p)^2$$

VII. Combine Result

After selecting the best algorithm for Regression and Classification.We join both algorithms predicting models in such a way that, after entering the flight number, our classification predicts flight name and its destination up to 5 historical records of the same flight present in the data set up to 5, days if records exist.After the record history is displayed, we input the requirement defined by the user which in the end results in flight ticket prediction .We choose Random Forest(Regression) and Decision tree(Classification) machine learning models to predict price and flight name as shown in fig:9.

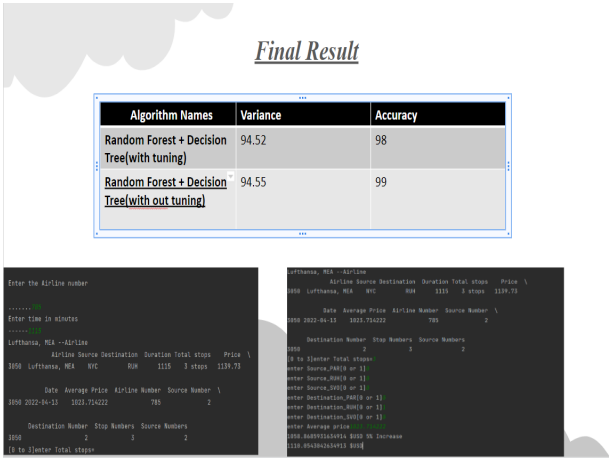

Fig. 9. Final Result

VIII. Conclusion

In this study, we performed our ML-Models using the raw flight data from airline tickets that is available on GitHub, which spans over 60 days of historical flight data written in 12 CVS flies for 12 routes. We have used decision tree algorithm and its subsets for ticket price prediction with addition to the categorization of flight number, we have removed outliers using the IQR approach. Overall, decision tree algorithms and our experiment demonstrate that the most optimum score is achieved by decision tree, with a 99.9% accuracy rate, is used to classify flight numbers which results in predicting flight name. Secondly the best method for forecasting ticket prices is done using random forest algorithms, with an MAE of 61.8$ and a result variance of 94.55% compared to test data which means or results are 94.55% accurate.

Future work can be done on the section that introduces the custom function that allocates flight numbers to flight names. Real flight numbers can be collected using web scrapers, and then the approach described in this paper can be used to get more realistic results. Additionally, an ensemble technique that uses multiple algorithms may provide reduced MAE, MSE, and RMSE.

References

[1] J. A. Abdella, N. M. Zaki, K. Shuaib, and F. Khan, "Airline ticket price and demand prediction: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 4, pp. 375–391, 2021.

[2] S. Rajankar and N. Sakharkar, "A survey on flight pricing prediction using machine learning," *International Journal Of Engineering Research & Technology (Ijert)*, vol. 8, no. 6, pp. 1281–1284, 2019.

[3] O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. 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*, 2003, pp. 119–128.

[4] J. Lu, "Machine learning modeling for time series problem: Predicting flight ticket prices," *arXiv preprint arXiv:1705.07205*, 2017.

[5] P. Biswas, R. Chakraborty, T. Mallik, S. I. Uddin, S. Saha, P. Das, and S. Mitra, "Flight price prediction: a case study," *Int. J. Res. Appl. Sci. Eng. Technol.(IJRASET)*, vol. 10, no. 6, 2022.

[6] Z. Wang, "Rwa: A regression-based scheme for flight price prediction," 2020.

[7] A. Morrisonn, S. Jing, J. T. O'Leary, and L. A. Cai, "Predicting usage of the internet for travel bookings: An exploratory study," *Information Technology & Tourism*, vol. 4, no. 1, pp. 15–30, 2001.

[8] T. Janssen, T. Dijkstra, S. Abbas, and A. van Riel, "A linear quantile mixed regression model for prediction of airline ticket prices," *Radboud University*, 2014.

[9] W. Groves and M. Gini, "An agent for optimizing airline ticket purchasing," in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, 2013, pp. 1341–1342.

[10] J. A. Abdella, N. Zaki, and K. Shuaib, "Automatic detection of airline ticket price and demand: A review," in *2018 International Conference on Innovations in Information Technology (IIT)*. IEEE, 2018, pp. 169–174.