iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.41058

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Flight Ticket Price Prediction Using Regression Models

S. Manoj Krishna[1], G. Sharitha[2], P. Madhu Ganesh[3], G.V. Ajith Kumar[4], G. Karthika[5]
[1, 2, 3, 4] *Student,* [5]*Assistant Professor, Department of Computer Science Engineering, Gitam Institute of Technology, Visakhapatnam, Andhra Pradesh, India*

Abstract: ***Many people nowadays choose to travel by flights. The cost of an airline ticket has a significant impact on a traveller's decision on which mode of transportation to use. A wide number of factors influence the price of an airline ticket, including social, competitive, marketing, and financial factors, among others. Every airline has a different technique for determining ticket prices. We can uncover the rules that airlines may use to model their fare variation using Machine Learning. In this project paper, we propose developing a web-based application for projecting the price of a flight ticket using Kaggle data, where the dataset contains various data related to 10,000 flights. The framework proposed will be used to simulate several regression algorithms for estimating projected flight fares. The model that will produce extremely accurate forecasts will be finalized, and it will solely be utilised to forecast the price.***
Keywords: regression, machine learning, model, prediction, algorithms

I. INTRODUCTION

The airline industry is one of the most sophisticated industries in the world, with sophisticated pricing tactics. Currently, airline ticket rates for the same flight might vary dramatically. Airline ticket pricing techniques have evolved into complicated frameworks of sophisticated rules and precise models that drive airfare pricing tactics since the deregulation of the airline sector. The reason for such a complex system is that each aircraft has a limited number of tickets to sell, thus airlines must manage demand.While Customers want to get the cheapest fare for their ticket, airline carriers try to increase their overall sales as high as possible along with maximizing their earnings. Nowadays many people travel regularly through flights hence they will have a general idea about what is the best time to get cheap airfares. But there may be some people who don't have any knowledge regarding this or they can also be inexperienced when it comes to booking flight tickets, and such people may end up falling for discount traps or deals made by some companies, where they can end up spending more than they should have. It had become hard for customers to get an air ticket at the cheapest price. For these few methods will be explored to determine the price of the ticket; hence they can decide the time and date at which they will book the air tickets with minimum price. The majority of these systems will mainly utilize Machine Learning techniques of regression in determining the expected price for a flight ticket. This design aims to develop a solution that will predict the flight prices for customers using a machine learning model. The model will give the predicted prices, and with its reference, the customer can decide the optimal time to buy their tickets. In this design, we will apply the introductory machine learning lifecycle to produce a web operation that will predict the flight fares by applying various machine learning algorithms using python libraries like Pandas, NumPy, Matplotlib, seaborn and sklearn. We can use different Machine Learning regression algorithms to get a model with advanced accuracy of the ticket price. In our design, we're going to use regression models since our end goal is to predict price using given data. Regression models such as Linear Regression (LR), Decision Tree, Random Forests,K Nearest Neighbor, Extreme Gradient Boosting are used to predict the estimated flight fare. The results of all models will be compared, and then the model that will provide high accuracy will finally be used. This model can predict airfares well in advance of the departure date.

II. LITERATURE REVIEW

An air ticket price prediction system has been created by K. Tziridis, Th. Kalampokas, K.I. Diamantaras, and G.A. Papakostas. They began their paper by giving a general introduction to machine learning, and then went on to discuss the methodology they used, which consists of four phases: Feature Selection, where the most important features that influence air ticket prices are determined; Feature Selection, where the most important features that influence air ticket prices are determined; Feature Selection, where the most important features that influence air ticket prices are determined; Feature Selection, where the most important features that influence air ticket prices are determined; Feature Selection, where the most important features that influence air ticket Data gathering comes next, followed by the selection of an accurate ML model, and ultimately, the model's evaluation.

They discuss how they performed prediction using nine different Machine Learning models in their paper, including the Multilayer Perceptron, Generalized Regression Neural Network, Extreme Learning Machine, Regression Tree, Bagging Tree, Bagging Regression Tree, Regression SVM (Polynomial), Regression SVM (Linear), and, Random Forest Regression Tree, Linear Regression. All of the ML models' results were compared and analysed. With an accuracy of 87.59 percent, the Bagging Regression Tree model exceeds all other models.

Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso, Steven Luis, and Shu-Ching Chen identified a problem of market segment level air ticket price prediction and developed a novel solution utilising a Machine learning approach. They used two public datasets, DBIB and T-100, which were obtained with minimal features, for training and evaluation of the above-proposed model. Data cleansing, transformation, pre-processing, feature selection or feature extraction, and application of the ML model are all part of their technique. The Random Forest Model Regressor was chosen because, when compared to other models such as Linear Regression, Support Vector Machines, and Neural Networks, it performed the best. This prediction framework achieved a high accuracy with an R squared Score of 0.869.

Supriya Rajankar, Neha Sakharkar, and Omprakash Rajankar published a paper on the same topic in which they used various regression models to solve the problem; their framework included phases such as data collection, cleaning and preparing data, analysing data, applying ML models, and finally evaluating the results; they used LR, SVM, KNN, RF, Decision Tree, and Bagging Regression. When compared to other algorithms for the dataset they used, the Decision Tree was more accurate.

III. SYSTEM METHODOLOGY

In our project, we will be implementing the general machine learning life cycle for creating a basic web application that will help in predicting flight fares by applying various machine learning algorithms on the historical flight data available, and we will be using various python libraries also for implementing those algorithms. Below figure shows the steps that we followed from the life cycle:

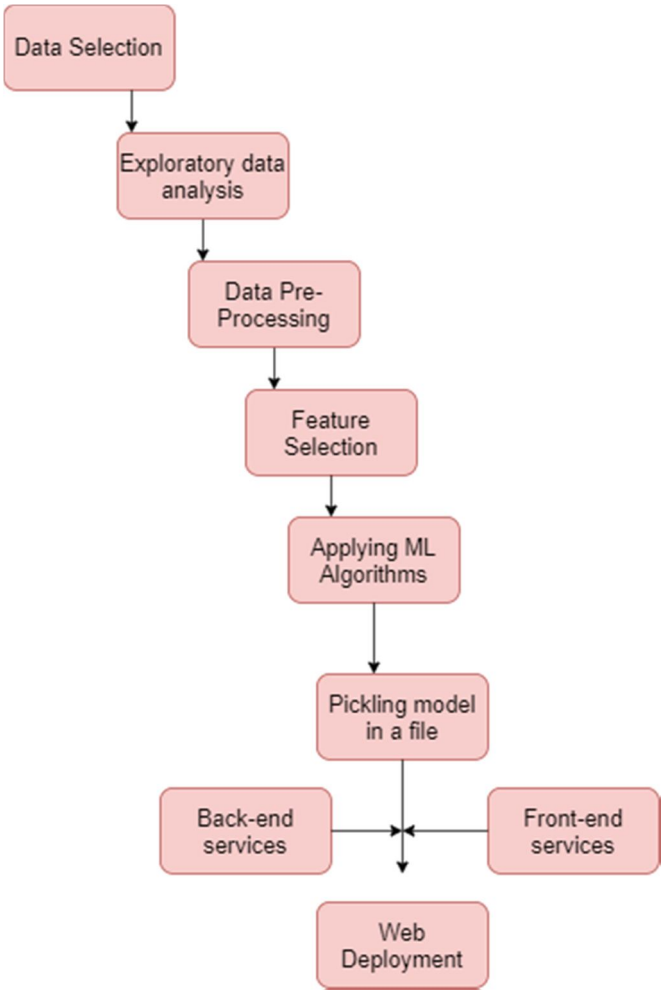


Fig 1. System Methodology

A. Data Selection

The accumulation of information is the most significant part of the project. The dataset that we will be using should have a contain continuous variables as well as categorical variables. Every column we use in our analysis should either be categorical or continuous. A training set shall be used for training the machine learning models for understanding the relationship between the explanatory variables and target variables. The test set shall be used to validate the performance of the learned relationship. The data selected for the training set should always represent the test set. We can say that no element can be more significant in machine learning than training data of high quality. Without training data of good quality, even the algorithms which offer high speed can be found to be useless. Robust machine learning models can be found useless when they are trained on inaccurate, inadequate, or non-relevant data in the early stages. In this project, the dataset used consists of more than 10,000 records of data related to flights and their prices. The features of the dataset that we took include airline, source, destination, arrival time, departure time, number of stops, prices, and a few more.

B. Exploratory Data Analysis

It is an approach used to analyse data sets and also for summarizing the main characteristics, often using different statistical graphs and other data visualization techniques. It will help us understand in what ways we can manipulate our data sources for getting the answers that we need, making it easier for data scientists in discovering anomalies, patterns or for verifying any assumptions or testing any presumption. EDA is done to visualize what the data is indicating or telling beyond the hypothesis or formal modelling, and it will provide a better understanding of data set, variables and their relations. It will also help in determining whether the statistical techniques you are considering for data analysis are appropriate or not. EDA techniques are widely used methods in the data discovery process these days. It helps in identifying outliers, errors, as well as give better understanding about the patterns within the data, also help in finding any interesting relations that are present among different variables.

C. Data pre-processing

It is a process of preparing the raw data for making it suitable and give it as an input to the ML model. It is one of the first and crucial steps in the process of creating a machine learning model. The real-world data generally consists of null values, noises, missing values, and it may also be in a format that cannot be used directly in machine learning models. We must perform this for cleaning the data and also for making it suitable for a machine learning model, which will also increase the efficiency, accuracy of a machine learning model. The quality of the historical data should be checked before applying it to any machine learning or data mining algorithms. Data Pre-processing is a step in which the data will be encoded or transformed to bring it into a new state such that now the machine can easily parse it. So, we can say our algorithms can now understand the features of the dataset. Here we observe that most of the data available in the historical dataset are of string format. So, data from each feature needs to be extracted; for example, day and month will be extracted from date of journey to integer format, hours and minutes are extracted from departure time into the format of date-time. Features like source, airline, and destination should be converted into numerical values as they are of categorical type. For this, we will be using One hot-encoding and label encoding technique, which are used to convert categorical values to model them into identifiable values.

D. Feature Selection

It is a process in which we filter out the non-relevant or redundant features while developing a predictive model. It is important to remove the redundant features so that it will reduce computational cost of the model and in some cases, it can also improve the speed and accuracy of our machine learning model. Since we familiarized ourselves with feature selection, now let us know about what feature extraction is, it is the process of developing a new, smaller set of features that will capture the majority of the relevant data. The difference between these two is feature selection keeps a subset of original features while feature extraction focuses on creating new sets. Both are generally used for dimensionality reduction. In our project, we will be using feature selection because we wanted to keep the original features as it is Statistical-based feature selection methods that will evaluate the relationship between each of the input variables and the target variable using certain statistics and select those input variables which have strong relations with the target variable. These methods can be effective and fast, but the choice about which statistical measure needs to be used depends on the data type of both the input and the output variables. It will be a challenge for machine learning practitioners to select the appropriate statistical measure for the dataset while performing filter-based feature selection. Filter-based feature selection methods will use various statistical measures while scoring the dependence or correlation between the input variables that will be filtered for choosing the most relevant features.

This step involves selecting the important features that are more correlated with price. Features are selected and passed to the group of models. From the predictions made by the various models, unnecessary features which may affect the accuracy of the model are removed before getting our model ready for prediction.

E. Applying ML Algorithms

To extract information from your data with machine learning, you must first choose your target variable, or the aspect about which you want to learn more. In this project, we will be choosing "price" because it is dependent on all other variables, it was already included as a feature in our historical dataset that was obtained during data collection. Now we have to run machine learning algorithms on the dataset to build models that learn by example from the historical data. Finally, we run the trained models on test data for which the model has not been trained on.

F. Back-end Services

The deployment of machine learning models is a process of making a machine learning model available in production where the web applications, enterprise software, and APIs can consume the trained model by providing new data points as input and generating predictions. There are two ways to generate predictions: we can predict by batch, or we can predict in real-time. The back-end of our application will be created using various frameworks where API end-points such as GET and POST will be used to perform various operations related to fetching and displaying data on the front-end. Needed information regarding the flights is collected from the consumer in a web form for predicting the air ticket prices. We will be building the back-end of the web application using the Flask Framework.

G. Front-end Services

The front-end of the web application will be created using the different frameworks available, where users will be entering the flight data. This data will now be sent to the back-end service, where the model will calculate the expected air ticket price. The predicted price is sent to the front-end and displayed.

IV. MODELS

A. Multiple Linear Regression

It is one of the supervised machine learning algorithms. There are two versions of linear regression simple and multiple. Both of them are used for finding the linear relationship between explanatory and target variables. But in simple there are only two variables predictor and response variable but in multiple there are several predictor variables. The equation is of the form:

$$Y=b_0+ b_1X_1+ b_2X_2+\ldots\ldots+ b_pX_p$$

Where

Y=predicted or expected value

$X_1,\ldots..X_p$ are P distinct predictor variables

b_0 = y-intercept when all other parameters are 0

$b_1,b_2,\ldots.$, bp are estimated regression coefficients

B. Decision tree

It is one of most frequently and widely used supervised machine learning algorithm that can perform both regression and classification. For each variable in the dataset, the decision tree algorithm will form a node, where the most important one will be placed as the root node. While evaluating we will start at the root node and move in downwards direction by following the node that meet our condition or "decision". This process will continue until a leaf node is reached, which will contain the predicted value or the outcome. For regression we have to use DecisionTreeRegressor class of the tree library. The evaluation metrics for regression are different from the metrics used in classification, but remaining everything is similar to classification.

C. K-Nearest Neighbour

It is an Supervised Learning technique. It will assume that there is some sort of similarity between the new case or data and available cases and will put the new case into the category that is most similar to the available class. This means when new data will appear then it can easily be classified into a suitable category by using K- NN algorithm.

As the name K Nearest Neighbor says it will consider K Nearest Data points for predicting the continuous value or class of the new Datapoint. Nearest neighbors are generally those data points that will have minimum distance in the feature space from our new data point. K is the number of data points that we will be considering in our implementation of the algorithm. Therefore, various distance metrics and K value are the two important things to be noted while implementing the KNN algorithm. There are many distance metrics such as Euclidean, Hamming, Manhattan distance we can pick one as per our need.

D. Random forest

It is also an supervised machine learning algorithm which uses ensemble learning methods. Ensemble learning is a sort of learning in which numerous versions of the same or distinct algorithms are combined to create a more effective prediction model. The Random Forest algorithm was created to solve the problem of high-variance in Decision Trees. We will train an entire forest, not just a single Decision Tree, as the name implies the random forest algorithm combines several comparable algorithms, such as numerous decision trees, to produce a forest of trees, hence the name "Random Forest". In both regression and classification problems, the random forest technique is used.

E. XGBoost

It's a more advanced variant of the Gradient Boosting method, and its name directly translates to "eXtreme Gradient Boosting." It is used because it is efficient and offers high speed. Because of the speed and performance, they provide, XGBoost models have dominated several Kaggle competitions. Weights play a significant part in this algorithm. All of the independent variables are given weights, which are subsequently fed into the decision tree, which predicts outcomes. In ensemble models, trees are introduced to the ensemble one at a time and fitted to correct prediction mistakes generated by previous models. XGBoost anticipates having a base of learners that are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancel out, and a better one sums up to form final good predictions.

V. RESULTS & DISCUSSIONS

We used a variety of machine learning algorithms to anticipate airfare in this project. Linear regression, Decision tree regression, K nearest neighbor regression, Random Forest regression, and Extreme gradient regression are the techniques that will be used. We use R2 Score as well as evaluation measures like mean absolute error, mean squared error, and root mean squared error.

1) Mean Absolute Error: This is the average absolute difference between the original and forecasted values throughout the whole dataset.
2) Mean Square Error (MSE): This is the averaged square difference between the original and predicted values over the entire dataset.
3) Root Mean Square Error: The square root of MSE determines the error rate.
4) R2 Score: This metric is used to assess the effectiveness of regression models. It calculates the variance in the model's predictions.

To get the best results, we calculated all of these for each model and compared them. Xgboost was chosen as the model with the lowest error rates and highest Score.After pickling our proposed model gave an R squared score of '0.92'.

TABLE I
Comparison of Different Regression Models' Performance

Regression algorithm	MAE	MSE	RMSE	R^2 Score for Test Set
Linear	1972	8202327	2863	0.61
Decision Tree	1343	6007394	2450	0.72
KNN	1774	7861068	2803	0.61
Random Forest	1178	4376418	2091	0.79
XGBoost	1131	3087950	1757	0.85

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

VI. CONCLUSION

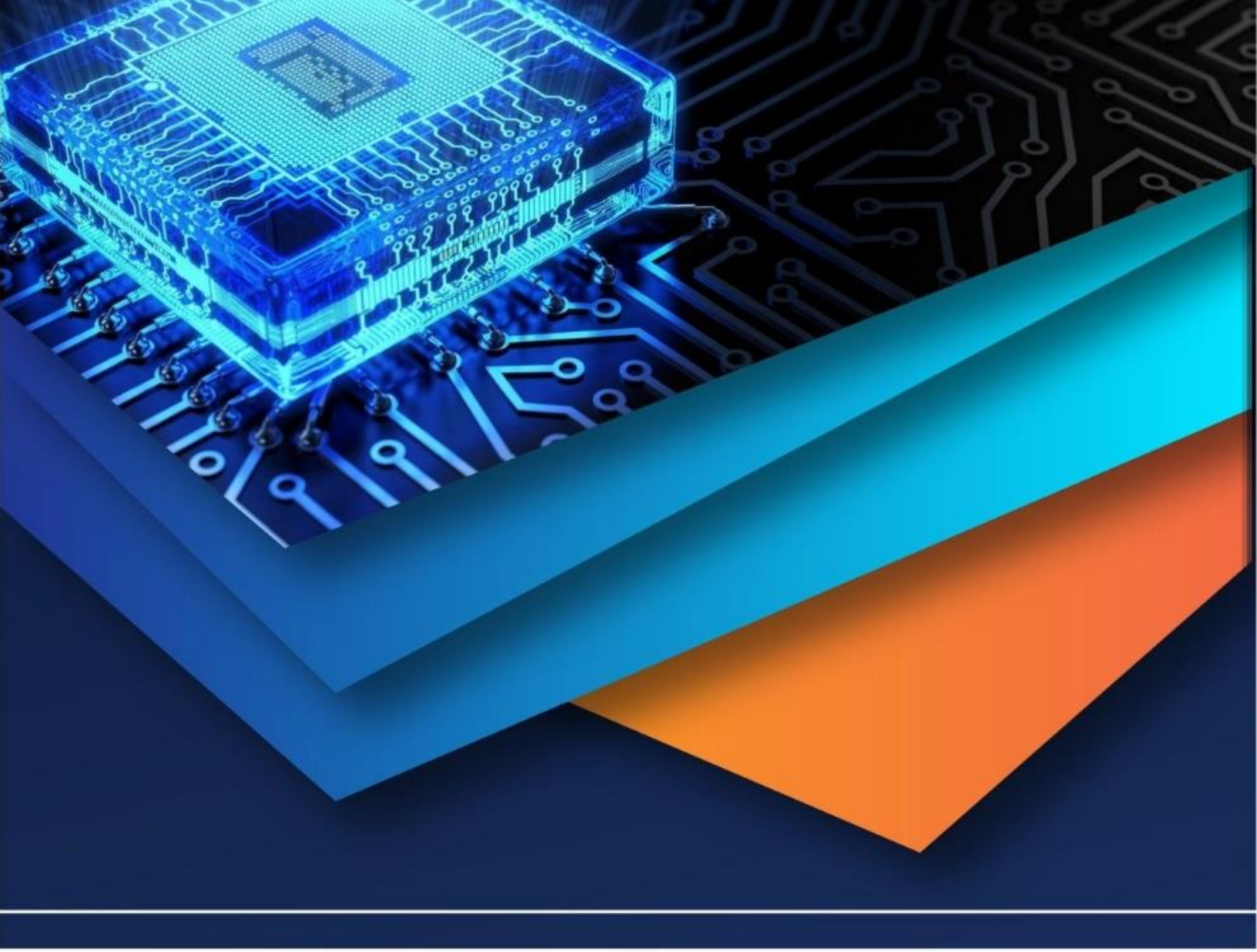
When our idea is properly implemented, it can help consumers who are unaware of flight fare patterns save money by giving them a predicted pricing value, which they can use to choose whether to buy the ticket now or later. Finally, for prediction, this type of service must be deployed with high accuracy. Because the estimated price may not be completely correct, this type of service has a lot of room for development. If more data could be accessible in the future, such as current seat availability, ticket class, and international and other domestic routes, the same study might now be stretched to a far broader scope. It is possible to improve the analysis by increasing the available data points and increasing the historical data used. That will train the model better, resulting in better accuracies and more savings.

VII. ACKNOWLEDGMENT

The successful completion of this project paper required a lot of guidance and assistance from many people, and we are highly privileged to have got this all along with the completion of our project. All that we have done is only possible due to the supervision and assistance provided by them, and we would not forget to thank them. Finally, we take this opportunity to thank one and all who have helped us directly and indirectly throughout this project.

REFERENCES

[1] Tianyi Wang; Samira Pouyanfar; Haiman Tian; Yudong Tao; Miguel Alonso; Steven Luis and Shu-Ching Chen, “A Framework for airline price prediction: A machine learning approach”.
[2] Dr. V. V. Kimbhahune, Harshil Donga, Ashutosh Trivedi, Sonam Mahajan, Viraj Mahajan, “FLIGHT FARE PREDICTION SYSTEM”.
[3] K. Tziridis, Th. Kalampokas, K.I. Diamantaras , G.A. Papakostas, “Airfare Prices Prediction Using Machine Learning Techniques”.
[4] Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar, “Predicting The Price Of A Flight Ticket With The Use Of Machine Learning Algorithms”.
[5] Janhvi Mukane , Siddharth Pawar , Siddhi Pawar , Gaurav Muley, “Aircraft Ticket Price prediction using Machine Learning”.
[6] Abhilash , Ranjana Y , Shilpa S , Zubeda A Khan, “Survey on Air Price Prediction using Machine Learning Algorithms”.
[7] Jaya Shukla , Aditi Srivastava, and Anjali Chauhan, “Airline Price Prediction using Machine Learning”
[8] Juhar Ahmed Abdella, Nazar Zaki, Khaled Shuaib, Fahad Khan, “Airline ticket price and demand prediction: A survey”.
[9] Achyut Joshi, Himanshu Sikaria, Tarun Devireddy, Dr. Vivek Vijay, “Predicting Flight Prices in India”.
[10] Janhvi Mukane, Siddharth Pawar, Siddhi Pawar, Gaurav Muley, “Aircraft Ticket Price prediction using Machine Learning”.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY