AIRBUS FARE PREDICTION WEB FRAMEWORK WITH THE USE OF MACHINE LEARNING MODEL

**1Sanpala Surya Snehavi,2Polemoni Ushaswini,3Jyoti Pruthi,4Ramsha Ahamad,5Vemunuri Gnaneshwar Reddy**

1,2,3,4,5 Chandigarh University

***Abstract*—**People who travel by plane regularly will have a better understanding of the greatest deals and the ideal times to buy tickets. Many airline businesses adjust their pricing based on the seasons or the length of time they are in operation. When individuals travel more, they will raise the price. Estimating the highest airline rates for the trip using features like [1] Duration, Source, Destination, Arrival, Departure. Features are extracted from a dataset. Dataset from kaggle had been taken and then we applied machine learning techniques and regression methodologies to forecast the price of airline tickets, which fluctuate over time. Random forest techniques are used to provide flight price prediction for consumers**.** The accuracy rate was 79.9 percent.

***Keywords— Machine learning, Airline rates, Features, Prediction model, Random Forest***

**1. Introduction** (FLIGHT PRICE PREDICTION)

Passengers have little knowledge about future business price rates, so finding the best moment to buy an airline ticket is tough. Different methods predict future business plans prices and classify the ideal moment to buy a plane ticket. Airlines utilize a variety of pricing systems for their tickets, deciding on a price later because the order displays a better value for the approximation models. Each of the factors that contribute to the complicated system are listed below. Because planes have a finite amount of seats, [2] airlines must manage demand. When demand is predicted to grow, the airline may raise prices in order to slow the rate at which seats are filled. Passengers or consumers should prepare ahead of time to take advantage of the greatest deals offered by various airlines and fly for a lower cost. The price of plane tickets fluctuates over time, removing the factors that contribute to the disparity. Reporting on the correlations and models that are used to price plane tickets. Then, based on that data, creating a model that assists travelers in selecting and purchasing tickets, as well as forecasting future air ticket costs. Duration, Arrival time,[3]Total stops, Airline, Source, Destination and much more are the features used for predicting the flight price.

**2. Literature review**

There is a lot of previous work on the Flight prediction models using Machine learning techniques.

K. Tziridis, Th. Kalampokas, et.al in [4] have introduced an airfare price prediction system, the paper they did is divided into 4 phases Feature selection, Data collection, ML model selection and Evaluation , they collected the data manually for months between December to July from net.They examined with eight machine learning algorithms such as , Multilayer Perceptron (MLP),Generalized Regression Neural Network, Extreme Learning Machine (ELM),Random Forest regression Tree, Regression Tree, Bagging Regression Tree, Regression SVM (Polynomial and Linear),Linear Regression (LR). They compared the accuracy of all these models and concluded as Bagging Regression Tree has the highest accuracy of 87.42%

Tao Liu, Jian Cao, et. al in [5] introduces an ACER(Adaptive Context- Aware Ensemble Regression )framework for airfare price prediction, before departure date it predicts the lowest price of the flight. In this paper they used three steps in model Feature selection and Extraction, single and multiple forecast algorithm, they took the data set from OATS in china. They used Bayesion Regression and on implementation they got RMSE scores between 3.7% to 6%.

Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar have developed the [6]Predicting The Price Of A Flight Ticket With The Use Of Machine Learning Algorithms, they divided it into two parts data collection and models in data collection they did collecting the data, cleaning and preparing the data and Analyzing the data . In models they used Decision tree, Random forest, K-NN, Linear Regression, they have taken the data set from [7]Makemytrip.com.

Jaywrat Singh Champawat, Udhhav Arora, Dr. K. Vijaya [8] “INDIAN FLIGHT FARE PREDICTION: A PROPOSAL” they divided model into two types one for customer side models and other for airline side model, they used EXTRATREE REGRESSOR to find important features and feature engineering performed to reduce the computational cost.In models they used ridge regression,lasso regression,K neighbors regression, decision tree and random forest

Tianyi Wang, Samira Pouyanfar, et. al in [9] outlines the problem of market segment level airfare price prediction and proposes a unique solution based on machine learning, they have taken two datasets DBIB and T-100 which contains minimal features. They have performed data preprocessing, feature extraction and selecting, models. They used models such as LR, SVM, Multilayer Perceptrons (MLPs), XGBoost Tree, and Random forest out of all they used Random Forest as it generated RMSE of 62.753 and it is low as compared to other models.

**3. Proposed work**

We are proposing a web page that by filling all the details it will give you the accuracy cost of flight based on the user filling the data. The input fields include the “**departure date**” that is , when you want to travel , upon next we have “**arrival date**” that is , when u want to reach the destination, upon next we have “**source**” and “**destination**” , later we have “**stoppage**” which means how many stops u needed to go from your source to the destination, later we have to choose an airline company to travel , it has in form of drop down and the input is called as “**Which airline you want to travel?**” Upon selecting and filling the all input ,click on the submit and we can see the predicted value on screen.

**4. Methodology**

In this project we used random forest regression for model building. We used libraries like pandas, numpy, seaborn, scikit learn and Matplotlib and for frontend we used HTML and css and for backend Flask framework. fig1 shows the methodology block diagram

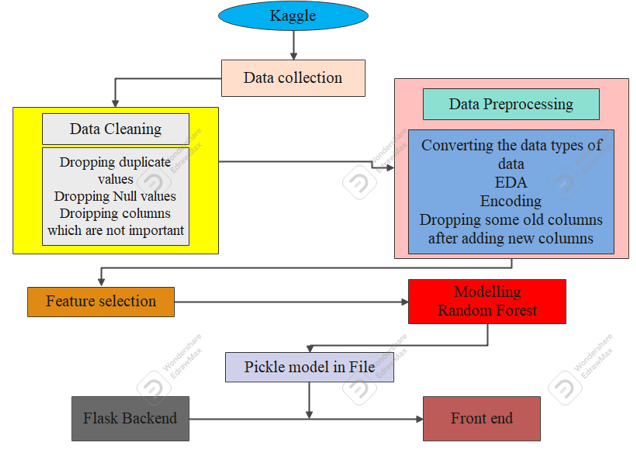


fig 1

**A.Data collection**

It has two types of data sets, one is used to train the data for a model called a train dataset and Other is used to test the model called a Test data set. In train data set we have total of 11 columns and 10683 rows, in test data set we have 10 columns and 2671 rows , both contains the features like Airline , Date\_of\_Journey ,Source, Destination,Route, , Price and other featured , but in test data set we don't have the Price feature but present in Train data set because it is the feature we need to predict.

**B. Data Cleaning**

In data cleaning we check whether there are any null values in the column and any duplicate values in the column , if there are any we drop the rows and we also check whether the column is useful for the process or not, if the column is not useful we drop that column too, like if any column has same data in all rows and has total null values these type of columns should be drop.

**C. Data pre processing**

In data preprocessing we convert the data types of Date\_of\_Journey, Dep\_Time,Arrival\_Time and Duration from object to datetime and we create new columns like Journey\_day and Journey\_month from Date\_of\_Journey,Dep\_Hr and Dep\_min from Dep\_Time,Arrival\_Hr and Arrival\_min from Arrival\_Time, Dur\_Hr and Dur\_Min from Duration. After the conversion and Creating new columns we drop the features Date\_of\_Journey, Dep\_Time,Arrival\_Time and Duration features as we created alternate columns for them.Later, we performed the Encoding , for nominal data we use One Hot encoding such columns as Source and Destination, and for Ordinal data we use Label encoder such columns as Total\_stops.

**D. Feature selection**

Feature selection also known as Attribute selection, As we work with randomly generated data or raw data may contain lots of noise data, if we keep these features in model we may not be able to produce accuracy in our models , so we select the important and useful features from all features and use them in the model. fig 2 depicts the feature importance in visualization.

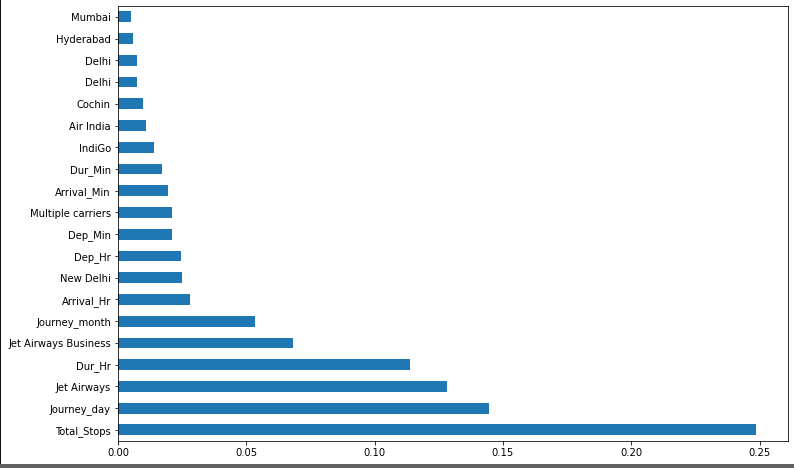


fig 2

From Fig 2 it is analyzed that Total\_Stops is the most important feature and Journey\_Day,Jet\_-Airways,Dur\_Hr are followed by and city mumbai is the least important feature.

| **Feature** | **Description** |
| --- | --- |
| Departure\_hr | At what hour the flight departs from the source |
| Departure\_min | At what Min the flight departs from the source |
| Arrival\_hr | At what hour the flight arrives to the destination |
| Arrival\_min | At what Min the flight arrives to the destination |
| Journey\_month | On what month user wants to travel |
| Journey\_day | On which day of the month the user wants to travel |
| Source | From where the flight departs |
| Airlines | Type of airline the user want to travel |
| Total stops | Number of stops between the Source and destination |
| Duration\_hr | Number of hours user travel in flight |
| Duration\_min | Number of minutes along with Hours user travel in flight |
| Destination | Where the flights arrive |

TABLE I Provides the description of the each feature after the encoding is performed

**E.Modeling**

We used Random Forest because it is a combination of decision trees and gives more accuracy than others because it generates multiple models.

**F.Pickle model in file**

We saved the data in a pickle file to reuse when we needed it.

**G. Flask backend**

To deploy the project we used the flask web framework to create a web page.

**F. Front end**

We used Html and CSS to create web pages and using Flask we deployed it.

**5. Implementation**

**A.Model**

**Random Forest:** [10]Random forest is a mix of multiple decision trees and uses it in a randomized way and it is an ensemble classifier. In this project Hyperparameter tuning is used to improve the performance and predict the model faster. 700 numbers of trees the algorithm builds before averaging the predictions and this generated randomly, 15 minimum samples used to split the node,1 minimum sample is required to each node, 20 is the maximum number of levels in tree , all the values in this hyperparameter tuning are randomly taken, for number of trees 100 to 1200 range is given and for number of levels in tree5 to 30 range is given.

**B. UI development**

For this project we used Flask web framework for backend and HTML,CSS for front end.

**[11] HTML** full form is **Hypertext Markup Language**, HTML is a language used for web. It is used to create websites . In this project Bootstrap and css are used. Eight columns are created in this project, those are Departure date , Arrival date in form of datetime-local input and Source ,destination,stopage,Which Airline you want to travel? are in the form of a drop down to choose from the options, and finally a submit button to submit the details. Bootstrap is used to create smooth responsive web pages.

**Flask** is a very important web application Framework which is written in Python. Flask is mainly used to deploy machine learning models.But flask has some restrictions like it cannot offer dynamic HTML pages but django supports it. In this project flask is used as connection between the web page and Prediction model, in Flask pickle file is loaded as we store the data in pickle and Html page is also connected and when we fill the data in webpage pickle gets request from it and by using pickle data it predicts the value and show us.

The user need to fill the all details in the web page [12] which are Departure time,Arrival time,Source, Destination, Which flight and Number of stops, After filling all the details when we click on submit , after that it takes the historical data we saved and use supervised learning and provides us the cost.

**6.RESULTS AND DISCUSSION**

[13]The sklearn.metrics module which is used in regression,classification and clustering, basically used to measure the errors ,accuracies and essential scores.

| MAE | 1187.1646462414396 |
| --- | --- |
| MSE | 3936080.7645026026 |
| RMSE | 1983.9558373367595 |
| r2 score | 0.797493524587619 |

TABLE II values of the Metrics

As from TABLE II we can analyze that the MAE,MSE and RMSE are high and r2 score is not good , as considered best r2 score is 1.

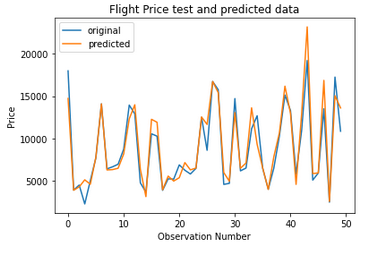


fig 5 shows how accurate the predicted data is with the original data

From fig 5 it is analyzed that most of the predicted data is the same as original data and also high irregularities in some points.

**7.Conclusion**

By using this correctly we can save the money by giving the person information for related people to travel, by this the person can decide to travel on the particular date he wanted to or to change the plans. This service can be used if good accuracy is good. This has huge scope as the predicted values are not fully accurate.

**8.Future Scope**

There are many tools in other industries used to predict values like stock brokers using the stock prediction tools and to estimate house price there is Zestimate. It has more future because the data is growing day by day and if we have more data the accuracy can be high .We can use different algorithms for more accuracy.

**References**

[1][2][3] Pavithra Maria K and Anitha K L, 2021. “ like Duration, Source, Destination, Arrival, Departure.”

[4] K. Tziridis, Th. Kalampokas, G. A. Papakostas, “Airfare Prices Prediction Using Machine Learning Techniques”, 25th European Signal Processing Conference (EUSIPCO), IEEE, October 26, 2017

[5]Tao Liu, Jian Cao, Yudong Tan, Quanwu Xiao, “ACER: An Adaptive Context- Aware Ensemble Regression Model for Airfare Price Prediction”, 2017 International Conference on Progress in Informatics and Computing (PIC), December, 2017

[6]Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar, “Predicting the price of a flight ticket with the use of Machine Learning algorithms”, international journal of scientific & technology research volume 8, December, 2019

[7][www.Makemytrip.com](http://www.makemytrip.com)

[8]Jaywrat Singh Champawat, Udhhav Arora, Dr. K. Vijaya “INDIAN FLIGHT FARE PREDICTION: A PROPOSAL” ,ijates,Vol.No-9,Issue no-3, March 2021

[9]Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen, “A Framework for Airfare Price Prediction: A Machine Learning Approach”, 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), September 9, 2019

[10]”Random forest is a kind of ensemble classifier Which is using a decision tree algorithm in a randomized fashion.”

[11]”**HTML** stands for hypertext markup language, HTML is a language of the web”

[12] Janhvi Mukane, Siddharth Pawar, Siddhi Pawar, Gaurav Muley “which required are Departure Time, Arrival Time,Source, Destination, Which flight and Number of stops”,International Journal for Research in Applied Science & Engineering Technology,Volume 10, Issue II, February 2022

[13]Vinod Kimbhaune, Harshil Donga, Asutosh Trivedi,

Sonam Mahajan and Viraj Mahajan “The sklearn.metrics module will be used to implement the functions to measure the errors from each model using the regression metrics”May 19, 2021