IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AIRPLANE FARE PREDICTION USING MACHINE LEARNING

[1]Attada Yamini, [2]Kolati Bhavani, [3] Mongam Asha,[4] Gandatti Sridhar,[5] Manda Sravanti

[123]UG Students, [4]M.Tech(Ph.D.),Assistant Professor,[5]MCA

[1]Department of Computer Science and Engineering,

[1]Satya Institute of Technology and Management,Vizianagaram,Andhra Pradesh,India.

Abstract: Currently, Everyone loves to travel by flights . Going along with the study ,the charge of travelling through a plane charge now and then which also includes the day & night time.Additionally , it changes with special times of the year of celebration seasons .There are a unique elements upon which the cost of air transport depends.For example time of the day ,time of take off , number of stops between them , number of days remaining in the month will provide the perfect time to purchase the plane ticket. As a result ,it is a basic understanding of flight rates before booking a vacation will undoubtedly save many individuals money and time. The goal is to investigate the factors that determine the cost of a flight. The data can be used to create a system that predicts flight prices.

Index Terms - Machine Learning Algorithms, airfare, supervised learning, predictions, flight, Linear Regression, Artificial Neural Network, Random Forest**.**

Introduction:

The flight ticket buying system is to purchase a ticket many days prior to flight take-off 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. all organizations have the privilege and opportunity to change its ticket costs at any time. 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. 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.

Methodology

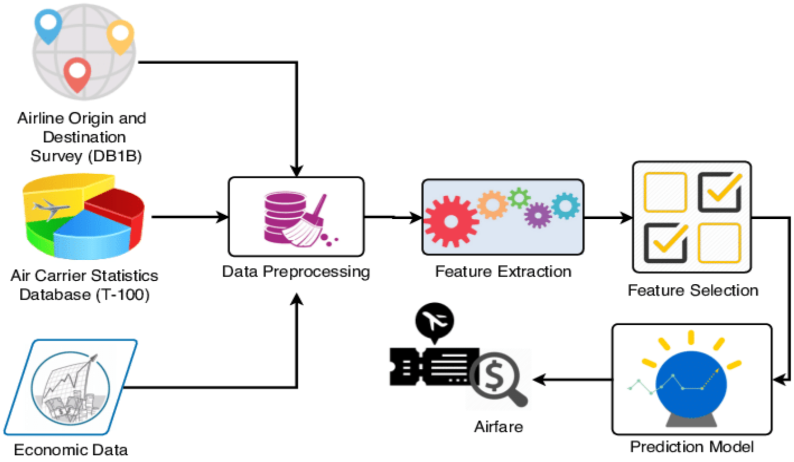


Fig. Data Flow Diagram

Data Collection :

The act of obtaining, acquiring, and combining the data that will be used to develop, test, and verify a machine learning model is known as data collection in machine learning. This step plays crucial role in implementation. Here data is collected from flight fare dataset which is imported from Kaggle. The dataset consists of both categorical data and numerical data. The categorical data includes source, destination, type of airline, additional info and numerical data includes arrival and departure dates, number of Stops. There are 11 columns (each represents a feature) and 10683 rows in this large dataset.

Data Preprocessing :

Data preprocessing means nothing but cleaning data, which can be used for model training and testing. By this step we can make our data useful for model training purpose. Data preprocessing involves cleaning, transforming, and preparing the data for data analysis. The sub steps involved in the data preprocessing are: Data Cleaning: In this step the null values are removed, missing values are removed and if any duplicates are present that are also removed. Feature engineering: In this step the features of our model are extracted and all the relevant 'label encoding' for ordinal categorical data was used to convert the categorical values to engineering numerical values. The dataset consists of categorical variables like airline, source, destination, route, total number of stops and additional info.

Data Splitting :

This step involves splitting our data into two parts for training and testing purpose. For model training 80 percent of data was used by using Random Forest regressor model was trained. The machine learning algorithms are: LGBM Regressor LGBM stands for Light Gradient Boosted Machine. It is a gradient boosting framework based on decision trees that can be used for various machine learning tasks such as regression, classification and ranking. LGBM Regressor is a class in light GBM package that can be used to train and predict regression models.

Randon Forest Regressor :

Random Forest regressor uses multiple decision trees to perform regression tasks. It is an example of ensemble learning. Random forest is a Supervised Learning algorithm which uses ensemble learning approach for classification and regression. Decision trees are sensitive to the specific data on which they are trained. If the training data is changed the resulting decision tree can be quite different and in turn the predictions can be quite different. Also, Decision trees are computationally expensive to train, carry a big risk of overfitting, and tend to find local optimal because they can’t go back after they have made a split to address these weaknesses, we turn to Random Forest.

Model evaluation:

This is an important step in our project, as it helps us to measure the performance and accuracy of our model. Test data is used for model evaluation. Here, we employed Cross-validation for model evaluation. This method divides the data into k-subsets, called folds. the model is trained on k-1 folds and tested on the remaining fold. this process is repeated k times, so that each fold is used as a test set once. The average performance across all k-folds is reported as the final result. The metrics that are used for model evaluation purpose are:

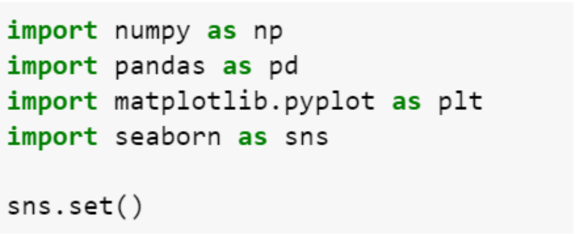
Root Mean Squared Error (RSME): It gives the root of the average squared difference between the actual values and the predicted values for a regression problem.

Mean Absolute Error (MAE): It gives the absolute difference between the actual values and predicted values. The higher negative mean values indicate the better performance of model.

R-Squared: This metric measures how well the regression model fits the data, by comparing it to a baseline model that always predict the mean value. It shows how much variation in the data is explained by the model.

Implementation and Results :

Importing the required libraries.

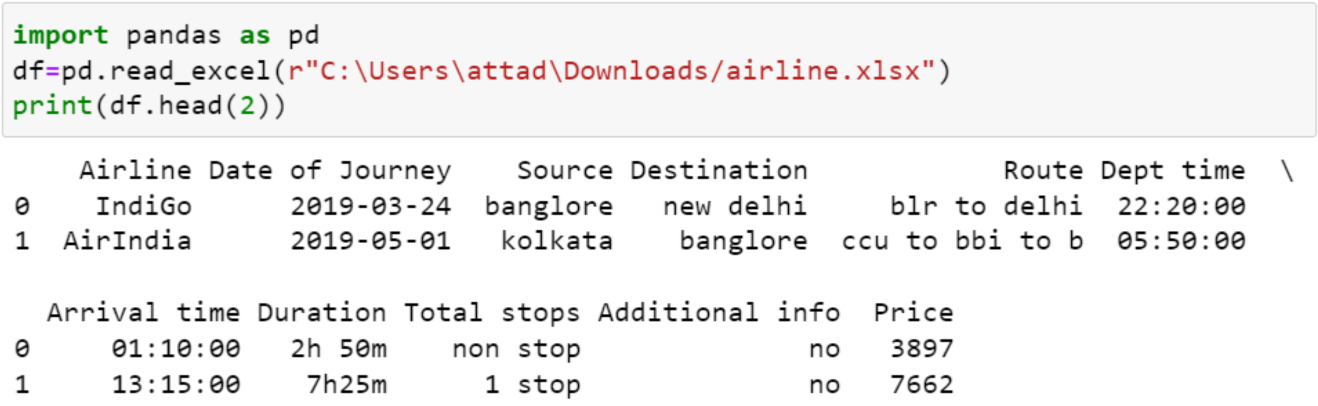


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set()
```

Read the Dataset :

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.



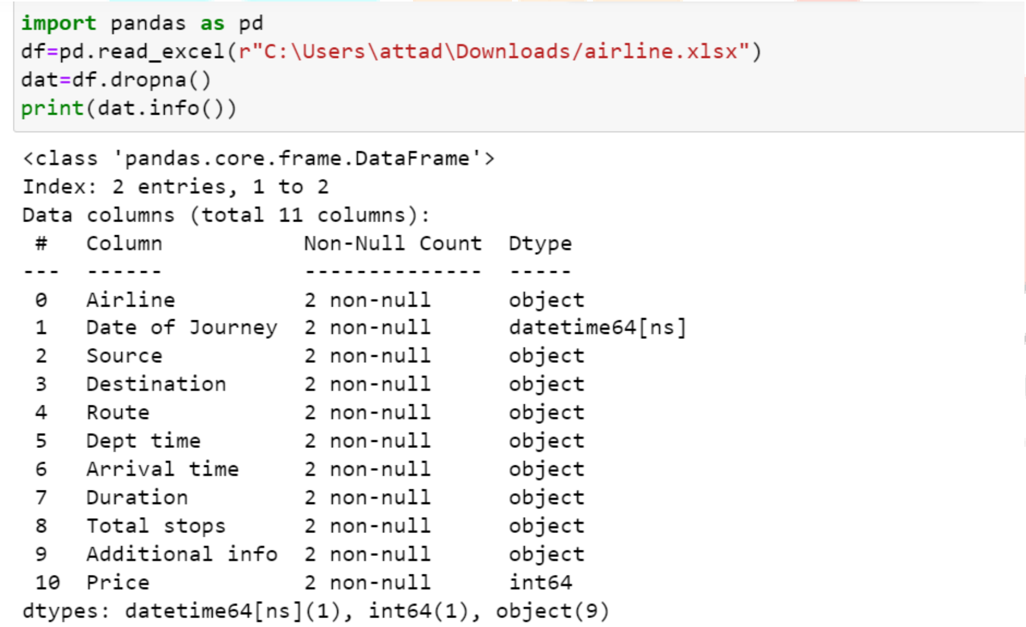
```
import pandas as pd
df=pd.read_excel(r"C:\Users\attad\Downloads/airline.xlsx")
print(df.head(2))
```

```
    Airline Date of Journey    Source Destination          Route Dept time  \
0    IndiGo      2019-03-24  banglore   new delhi   blr to delhi  22:20:00
1  AirIndia      2019-05-01   kolkata    banglore  ccu to bbi to b  05:50:00

  Arrival time Duration Total stops Additional info  Price
0     01:10:00   2h 50m    non stop              no   3897
1     13:15:00    7h25m      1 stop              no   7662
```

Data preprocessing:

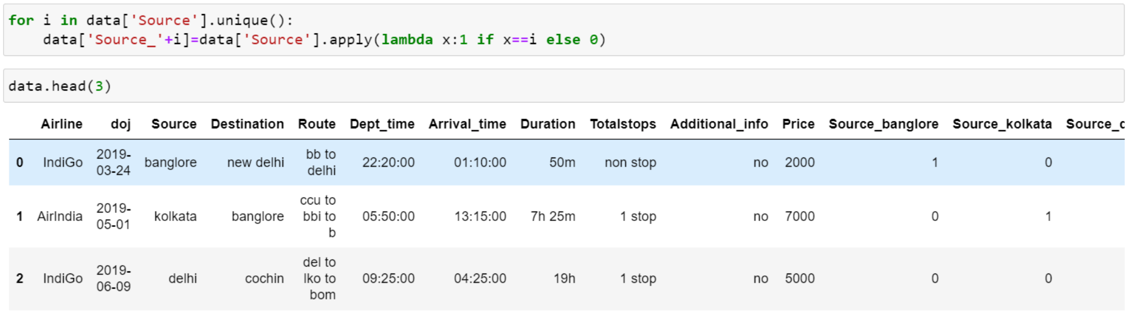
The df.isnull() method is used to verify that no values are present. We employ the sum () function to add up those null values. Two null values were discovered in our dataset, we discovered. We thus start by investigating the data.



```
import pandas as pd
df=pd.read_excel(r"C:\Users\attad\Downloads/airline.xlsx")
dat=df.dropna()
print(dat.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2 entries, 1 to 2
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          2 non-null      object
 1   Date of Journey  2 non-null      datetime64[ns]
 2   Source           2 non-null      object
 3   Destination      2 non-null      object
 4   Route            2 non-null      object
 5   Dept time        2 non-null      object
 6   Arrival time     2 non-null      object
 7   Duration         2 non-null      object
 8   Total stops      2 non-null      object
 9   Additional info  2 non-null      object
 10  Price            2 non-null      int64
dtypes: datetime64[ns](1), int64(1), object(9)
```

Feature Selection : Applying One-Hot-Encoding technique to improve the accuracy by segregate into 0's and 1's



```
for i in data['Source'].unique():
    data['Source_'+i]=data['Source'].apply(lambda x:1 if x==i else 0)
```

```
data.head(3)
```

	Airline	doj	Source	Destination	Route	Dept_time	Arrival_time	Duration	Totalstops	Additional_info	Price	Source_banglore	Source_kolkata	Source_c
0	IndiGo	2019-03-24	banglore	new delhi	bb to delhi	22:20:00	01:10:00	50m	non stop	no	2000	1	0	
1	AirIndia	2019-05-01	kolkata	banglore	ccu to bbi to b	05:50:00	13:15:00	7h 25m	1 stop	no	7000	0	1	
2	IndiGo	2019-06-09	delhi	cochin	del to lko to bom	09:25:00	04:25:00	19h	1 stop	no	5000	0	0	

Algorithm and Analysis :

While we go through the algorithms we employed (XGBoost, Random Forest, and Decision Tree) and also how they operate in our models, please read the discussion below.

Decision Tree :

The decision tree appears to be the most well-known and commonly employed categorization technique. A decision tree is a collection of nodes that resembles a diagram, for each junction indicating a test on the a characteristic and each branch indicating a test outcome, such that each node in a decision tree (terminal node) has a class label. A tree can be "trained" by dividing the resources collection into subgroups depending on a characteristic values test. This procedure is known as partitioning the data because it is performed iteratively on each derived subset. The recursion ends when all subgroups at a node have the same posterior probability, or when the split no longer adds additional value to the predictions. A decision tree is appropriate for experimental extracting knowledge since it does not need subject matter expertise or parameters configuration. Assume S is a collection of cases, A is a property, Sv is the subgroup of S with Such a = v, as well as Value (A) is the collection of all number of values of A, then

$$\text{Gain}(S,A)=\text{Entropy}(S)-\sum_{ve} \text{Values}(A)|S_v|/\ |S|.\text{Entropy}(S_{v)}$$

Random Forest :

A Random Forest is an ensemble approach that can handle simultaneous regression and classification problems by combining many decision trees using a technique known as Bootstrap as well as Aggregation, or bagging. The core idea is to use numerous decision trees to determine the final result instead of depending on personal decision trees. Random Forest's foundation learning methods are numerous decision trees. We arbitrarily choose rows and characteristics from the dataset to create sample datasets for each model. This section is known as Bootstrap. We simply have to understand the purity in our dataset, and we'll use that characteristic as the root of the tree which has the smallest impurity or, in other words, the smallest Gini index. Mathematically Gini index can be written as:

$$\text{Index} = 1-\sum_{i=1}^{n} (P_i)^2 = 1- [(P_+)^2 + (P_-)^2]$$

XG Boost:

XGBoost is an effective method for developing supervised regression models. Knowing as to its (XGBoost) goal function and baseline learners can help determine the truth of this proposition. This optimization problem has both a loss function and a regularization component. It makes a distinction between real and theoretical predictions, i.e. how far the model outputs deviate from the real amounts. In XGBoost, the most used standard error in regression problems is quarantine, whereas reg:logistics is used for classifications. The formula may be used to compute the output value of each model.

$$\text{Output Value} : \sum \text{Residual} / \text{no.of Residual} + \lambda$$

Results :

Algorithm	Training Accuracy	Testing Accuracy
XG Boost	**0.92**	**0.77**
Random Forest	**0.95**	**0.78**
Decision Tree	**0.97**	**0.67**

FUTURE SCOPE :

- **Optimal date recommendation**: It means suggesting the date to users on which date the flight prices will be minimum.
- **Real-Time Updates**: Dealing with real-time data for dynamic pricing adjustments based on factors like weather, demand, and airline policies.
- **Integration**: Partnering with airlines, travel agencies, and online booking platforms to provide pricing as a value-added service.

CONCLUSION :

In conclusion, the main aim of our project flight fare prediction using machine learning is to predict the prices. we have created a User Interface for the entire process which includes arrival date, departure date, source, destination, etc. Our flight fare prediction project using machine learning has successfully produced a reliable and user-friendly system. We collected, preprocessed, and extracted features from flight fare data, trained a robust random forest model and evaluated its performance. This web application we developed empowers travelers to make informed decisions by predicting flight prices based on their input.

References :

[1] Rajankar, Supriya, and Neha Sakharkar. "A Survey on Flight Pricing Prediction using Machine Learning. " Internatıonal Journal Of Engıneerıng Research & Technology (Ijert) 8.6 (2019): 1281- 1284.

[2] Smith, Barry C., John F. Leimkuhler, and Ross M. Darrow. "Yield management at American airlines." interfaces 22.1 (1992): 8-31.

[3] Groves, William, and Maria Gini. "An agent for optimizing airline ticket purchasing." Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems. 2013.

[4] Janssen, Tim, et al. "A linear quantile mixed regression model for prediction of airline ticket prices." Radboud University (2014).

[5] Wohlfarth, Till, et al. "A data-mining approach to travel price forecasting." 2011 10th International Conference on Machine Learning and Applications and Workshops. Vol. 1. IEEE, 2011.

[6] Papadakis, Manolis. "Predicting Airfare Prices." (2014).

[7] Ren, Ruixuan, Yunzhe Yang, and Shenli Yuan. "Prediction of airline ticket price." University of Stanford (2014).

[8] Tziridis, Konstantinos, et al. "Airfare prices prediction using machine learning techniques." 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017.

[9] Boruah, Abhijit, et al. "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter." Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, 2019. 191-203.

[10] S. Chakravarty, B. K. Paikaray, R. Mishra and S. Dash, "Hyperspectral Image Classification using Spectral Angle Mapper," 2021 IEEE International Women in Engineering (WIE) Conference on 0 0.1 0.2 0.3 0.4

0.5 0.6 0.7 0.8 Artificial Neural Network Linear Regression Decision Tree Random Forest MAPE Values Algorithm Applied MAPE RESULTS 177 Electrical and Computer Engineering (WIECON-ECE), 2021, pp. 87-90, doi: 10.1109/WIECONECE54711.2021.9829585.

[11] Wang, Tianyi, et al. "A framework for airfare price prediction: A machine learning approach." 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, 2019.

[12] Abdella, Juhar Ahmed, et al. "Airline ticket price and demand prediction: A survey." Journal of King Saud University-Computer and Information Sciences 33.4 (2021): 375-391.

[13] Zhao-Jun, Gu, Wang Shuang, and Zhao Yi. "Flight ticket fare prediction model based on timeserial." Journal of Civil Aviation University of China 31.2 (2013): 80.

[14] Huang, Tenghui, Chih-Chien Chen, and Zvi Schwartz. "Do I book at exactly the right time? Airfare forecast accuracy across three price-prediction platforms." Journal of Revenue and Pricing Management 18.4 (2019): 281-290.

[15] S. Chakravarty, P. Mohapatra, P. K. Dash, (2016), Evolutionary Extreme Learning Machine for Energy Price Forecasting, International Journal of Knowledge-Based and Intelligent Engineering Systems, 20, 75-96