FLIGHT FARE PREDICTION USING MACHINE LEARNING

K.D.V.N.Vaishnavi[1], L. Hima Bindu[2], M. Satwika[3], K. Udaya Lakshmi[4], M. Harini[5], N. Ashok[6]

[1,2,3,4,5]*Students, Information Technology, Vasireddy Venkatadri Institute of technology, Guntur, India.*
[1]*Department of Information Technology*
[1]*Vasireddy Venkatadri Institute of Technology, Guntur, India.*
[6]*Associate Professor, Department of Information Technology, Vasireddy Venkatadri Institute of Technology*
Guntur, Andhra Pradesh, India.

Article DOI: https://doi.org/10.36713/epra14763
DOI No: 10.36713/epra14763

ABSTRACT

The Flight Fare Prediction System is a comprehensive solution aimed at accurately forecasting flight ticket prices, providing travelers with valuable insights for better planning and decision-making. Nowadays, airline ticket prices can vary dynamically for the same flight. From the customer's perspective, they want to save money, so I have proposed a model that predicts the approximate ticket price. This system leverages machine learning algorithms and historical flight data to generate accurate fare predictions. The system utilizes a vast dataset comprising historical flight fares, including factors such as travel dates, destinations, airlines, departure times, and various other relevant variables. By analyzing this data using advanced machine learning techniques, the system learns patterns and relationships, enabling it to make reliable predictions about future flight fares. An ensemble of machine learning algorithms, including regression-based models like Random Forest, Gradient Boosting, and Support Vector Regression, is employed to capture complex patterns and relationships within the data. This system will give people an idea of the trends the prices follow and also provide the predicted value of the price, which they can check before booking flights to save money. This kind of system or service can be provided to customers through flight booking companies to help them book tickets.

KEYWORDS - *Flight Fare Prediction, Machine Learning, Historical Flight data, Random Forest.*

I. INTRODUCTION

Everyone knows that holidays always call for a much-needed vacation and planning the travel itinerary becomes a time-consuming task. The commercial aviation business has grown tremendously and has become a regulated marketplace as a result of the worldwide growth of the Internet and E-commerce. Hence, for Airline revenue management, different strategies like customer profiling, financial marketing, and social factors are used for setting ticket fairs. When tickets are booked months in advance, airfares are often reasonable, but when tickets are booked in a hurry, they are often higher. But, the number of days/hours until departure isn't the only factor that decides flight fare, there are numerous other factors as well. Customers find it quite difficult to obtain a perfect and lowest ticket deal due to the aviation industry's complex pricing methodology. Machine Learning and Deep Learning-based technologies and models have been created to overcome this challenge, and substantial research is also happening. This study discusses a Machine Learning-based Flight Fare Prediction System that employs Random Forest Regression to predict airline ticket pricing. Various features that influence prices are also studied along with the system's experimental analysis. Section II included a literature review that looked at technical papers as well as some current models and systems. Differences in the features considered are also mapped down, In Section III, the proposed system is described in detail along with the workflow and its features. In Section IV, the results as well as various comparisons between findings are reported. In Section V, conclusions are provided. In Section VI, prospective advancements for further research.

II. RELATED WORK

The paper begins with some broad information regarding machine learning, after which the authors further proceed to the methodology. The methodology consists of four-phase process that influences flight prices, collection of data from flight fare prediction dataset by MH, selection and evaluation of an accurate ML Regression model. Key to its success is the integrity of the system where consumers accurately represent the segments for which prices have been differentially determined.

[2] Today, airlines price tickets "as much as the customer and market will bear," according to consultant and former airline planning executive. Airlines also profile their customers to help them adjust prices.
[3] This often means placing passengers into one of two groups: leisure or business. And the way each group is priced is very different. Most studies on airfare price prediction have focused on either the national level or a specific market. Research at the market segment level, however, is still very limited. We define the term market segment as the market/airport pair between the flight origin and the destination.

[4] The airline dataset included the following eight characteristics: departure and arrival times, type of airline, number of stopages, source, destination and additional_information. The authors performed prediction using regression Machine Learning models that including, LGBM Regressor, Random Forest Regression Tree, and Decision Tree Regressor. Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

[5] All in present a review of deep learning and social media data-based Airline ticket price prediction model. The authors introduce the current airline ticket pricing situation with the factors that affect ticket prices. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships.

[6] Random Forest Model is used for development since it outperforms other models such as LGBM Decision Tree Regressor and Neural Network in terms of data performance with a R squared score of 0.91, this prediction framework has a good level of accuracy.

III. METHODOLOGY

The following steps are involved in the proposed architecture of our project.

3.1. 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.

3.2. 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 selection and engineering: In this step the features of our model are extracted and all the relevant features are used for model training. In dataset it contains date of journey, arrival date, departure date columns and all the numerical values are extracted as Departure hour, departure minutes, arrival hour, arrival minutes, journey day, journey month. As dataset contains both categorical and numerical features, by using 'On hot encoding' method for nominal categorical data and 'label encoding' for ordinal categorical data was used to convert the categorical values to numerical values. The dataset consists of categorical variables like airline, source, destination, route, total number of stops and additional info.

3.3 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 lightgbm package that can be used to train and predict regression models.

Decision Tree Regressor

Decision tree regressor is a class in Sklearn tree module that can be used to train and predict regression models. It is a decision tree-based algorithm that recursively partitions the input data based on the values of the input features, forming a tree-like structure. It initially chooses independent variable from dataset as decision nodes for decision making and then it divides the entire dataset into sub-sections and when test data is passed to the model the output is decided by checking the data point belongs to the decision tree will give output as the average value of all the datapoints in the sub-section.

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.

3.4 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.

3.5 Model Architecture

The architecture is crafted with essential components to optimize prediction. Our proposed system analyses historical data to identify patterns, seasonal trends and additional information that influence flight fares.

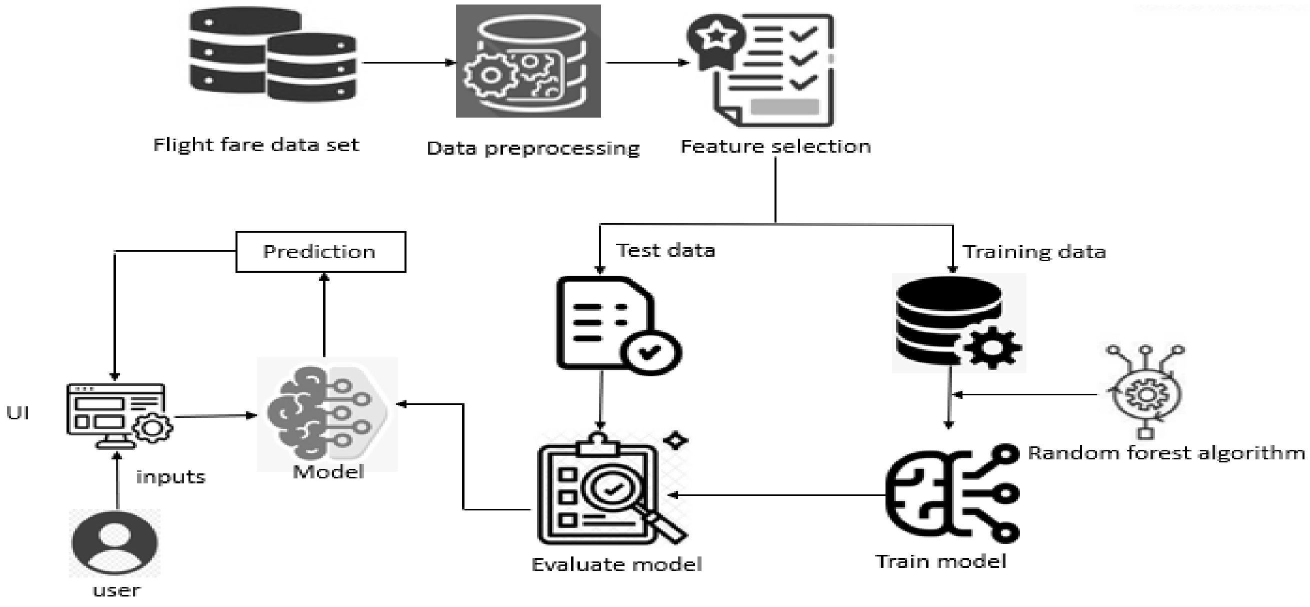
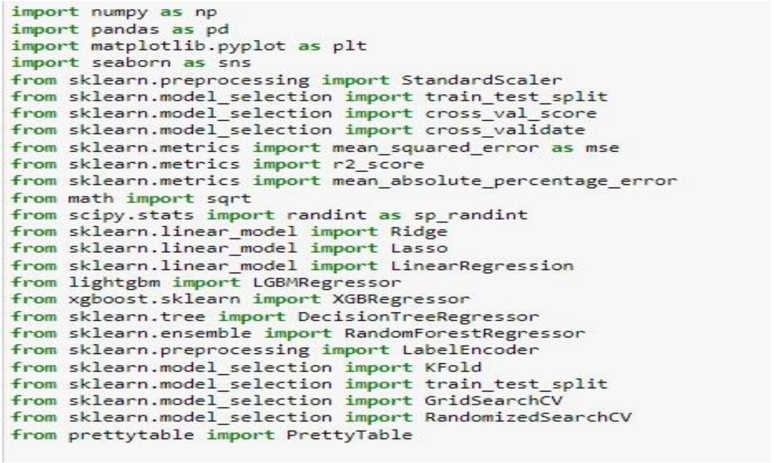


Fig.1: System Architecture

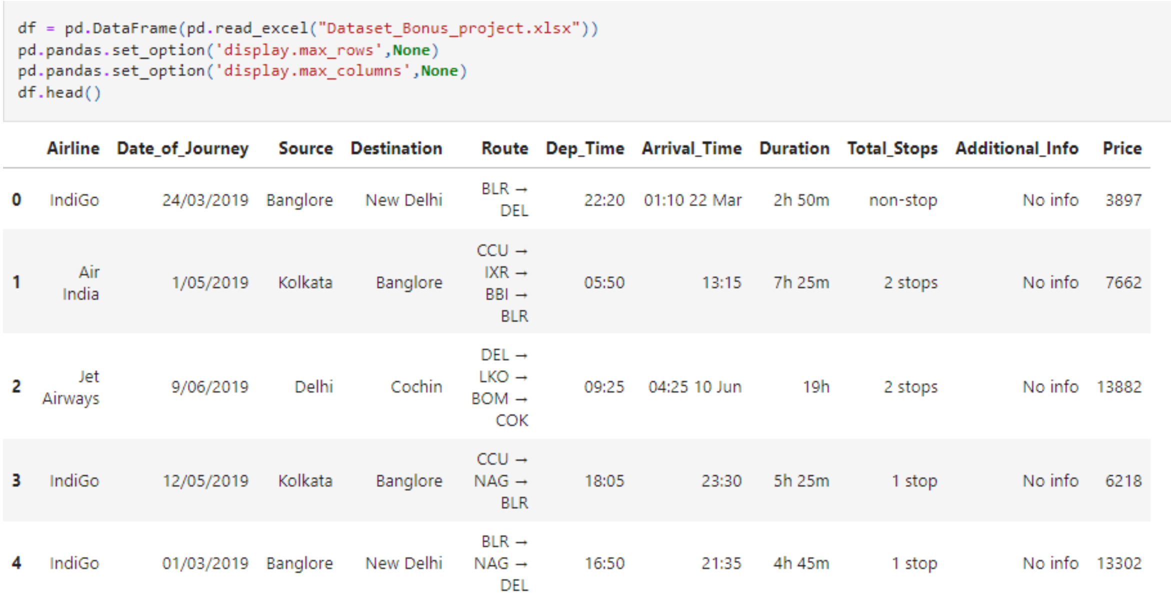
IV. IMPLEMENTATION AND RESULTS

1. Importing the required libraries.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_validate
from sklearn.metrics import mean_squared_error as mse
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_percentage_error
from math import sqrt
from scipy.stats import randint as sp_randint
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.linear_model import LinearRegression
from lightgbm import LGBMRegressor
from xgboost.sklearn import XGBRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import KFold
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from prettytable import PrettyTable
```

2. 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.

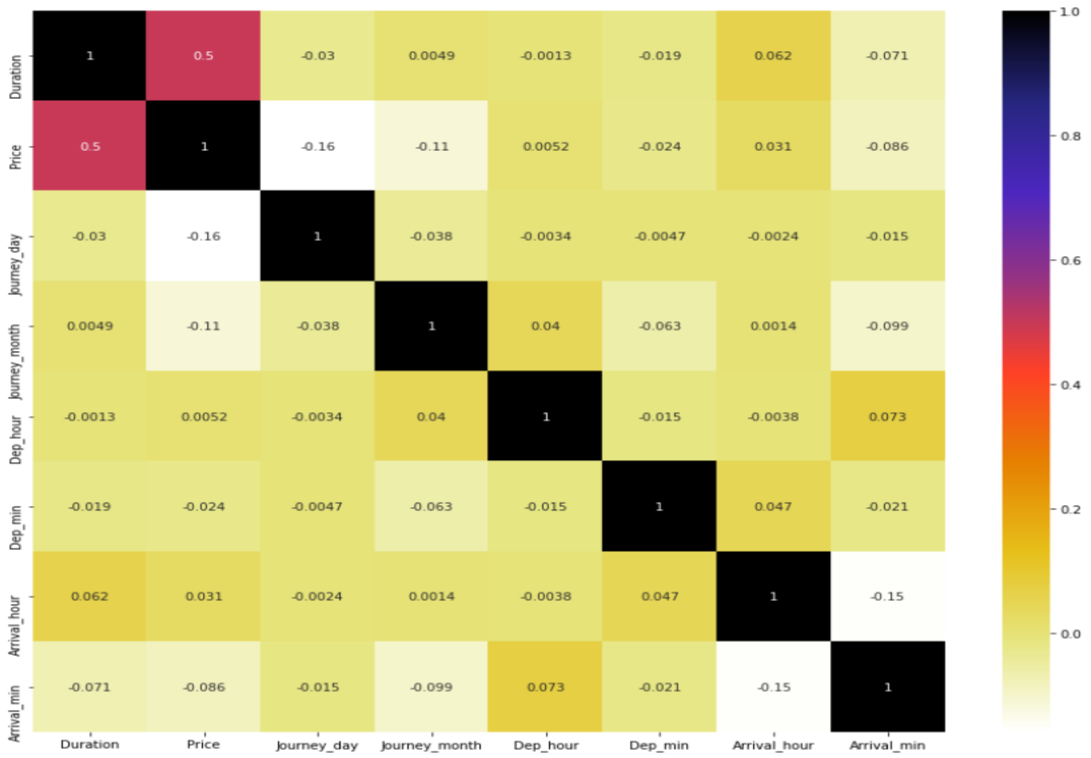


```
df = pd.DataFrame(pd.read_excel("Dataset_Bonus_project.xlsx"))
pd.pandas.set_option('display.max_rows',None)
pd.pandas.set_option('display.max_columns',None)
df.head()
```

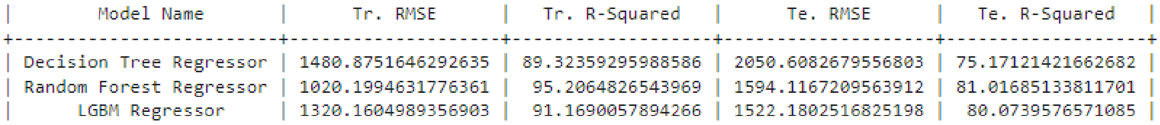
	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

3. 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.

4. Using a Heat Map to check the correlation:
Here we are using a heat map to check the correlation in this instance. Using different colour combinations, it displays the data as 2-D coloured maps. Instead of numbers, it will be plotted on both axes to describe the relationship variables.



5. By comparing all the models (LGBM Regressor, Decision Tree Regressor, Random Forest Regressor), we can conclude that Random Forest Regressor performs the best.


```
|        Model Name        |      Tr. RMSE      |   Tr. R-Squared   |      Te. RMSE      |   Te. R-Squared   |
+--------------------------+--------------------+-------------------+--------------------+-------------------+
| Decision Tree Regressor  | 1480.8751646292635 | 89.32359295988586 | 2050.6082679556803 | 75.17121421662682 |
| Random Forest Regressor  | 1020.1994631776361 |  95.2064826543969 | 1594.1167209563912 | 81.01685133811701 |
|      LGBM Regressor      | 1320.1604989356903 |  91.1690057894266 | 1522.1802516825198 |  80.0739576571085 |
```

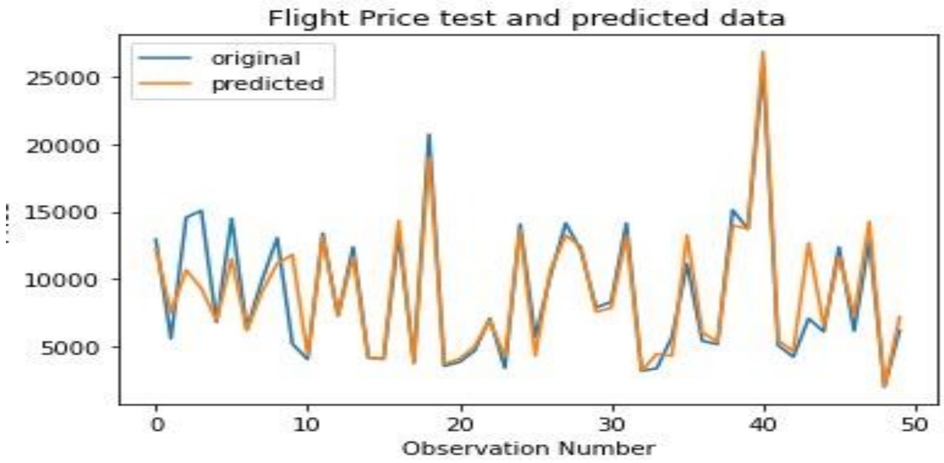
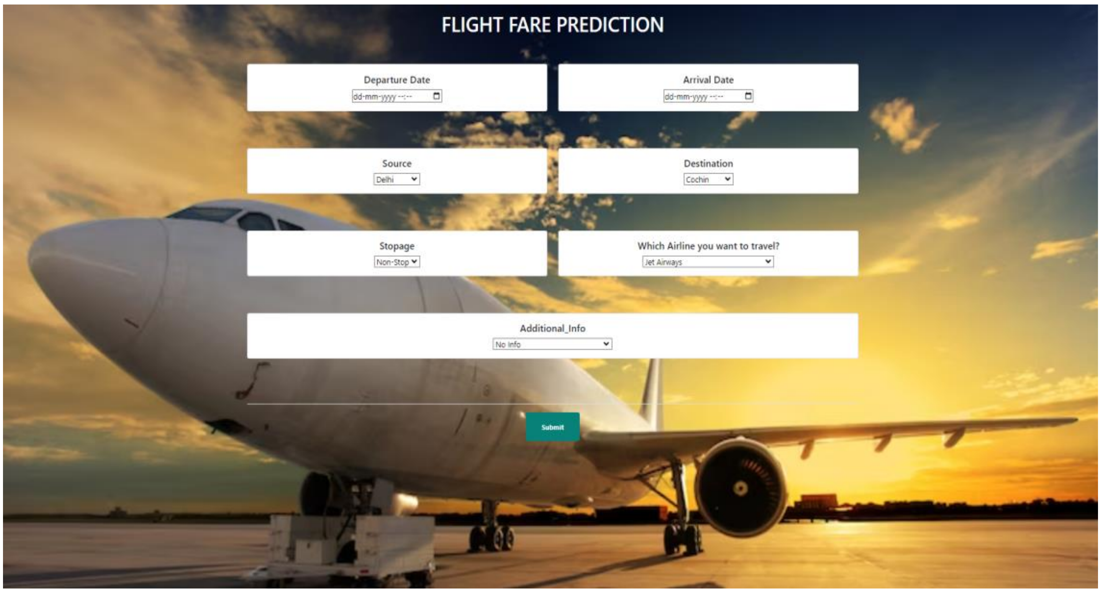


Fig 2: Flight Price Test and Predicted Data



Departure Date: dd-mm-yyyy --:--
Arrival Date: dd-mm-yyyy --:--
Source: Delhi
Destination: Cochin
Stopage: Non-Stop
Which Airline you want to travel? Jet Airways
Additional_Info: No Info

Fig.3: Home Page

Figure 3 By using flask the machine learning model wasy deployed. Flask is a python web frame work that allows you to build lightweight and flexible web applications quickly and easily. The web page collects information and it predicts the fare of ticket. It collects information like departure date, arrival date, source, destination, number of stops, type of airline and additional information(meal included or not, no info, change airports, etc).

V. 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.

VI. 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.

VII. REFERENCES

1. *K. Tziridis, Th. Kalampokas, G. A. Papakostas, "Airfare Prices Prediction Using Machine Learning Techniques", 25th European Signal Processing Conference (EUSIPCO), IEEE, October 26, 2017.*
2. *N. Brown and J. Taylor, Air Fare: Stories, Poems & Essays on Flight. Sarabande Books, 2004.*
3. *J. C. Lok, Prediction Factors Influence Airline Fuel Price Changing Reasons. 2018.*
4. *A. Kakaraparthi and V. Karthick, "A secure and cost-effective platform for employee management system using lightweight standalone framework over Diffie Hellman's key exchange algorithm," ECS Trans., vol. 107, no. 1, pp. 13663–13674, Apr.2022.*
5. *Wikipedia, "MeanSquarederror", Available:https://en.wikipedia.org/wiki/Mean_squared_error.*
6. *G. Ataman and S. Kahraman, "Stock Market Prediction in Brics Countries Using Linear Regression and Artificial Neural Network Hybrid Models," The Singapore Economic Review. pp. 1–19, 2021. doi: 10.1142/s0217590821500521.*
7. *B. Panwar, G. Dhuriya, P. Johri, S. S. Yadav, and N. Gaur, "Stock Market Prediction Using Linear Regression and SVM," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). 2021. doi: 10.1109/icacite51222.2021.9404733.*
8. *P. Purey and A. Patidar, "Stock Market Close Price Prediction Using Neural Network and Regression Analysis," International Journal of Computer Sciences and Engineering, vol. 6, no. 8. pp. 266–271, 2018. doi: 10.26438/ijcse/v6i8.266271.*
9. *William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.*
10. *https://towardsdatascience.com/machine-learning-basics-decisiontree-regression-1d73ea003fda article on decision tree regression.*
11. *www.keboola.com/blog/random-forest-regression article on random forest.*