Detailed Project Report
Flight Fare Estimator

Revision Number – 1.2

Last Date of Revision: 22–01–2023

Shubham Shaah

Document Version Control

Date	Version	Description	Author
10 – 11 - 2022	1.0	Abstract Introduction General Description	Shubham
12 – 01 – 2023	1.1	Technical Requirements Data Requirements Data Pre-processing Design Flow	Shubham
22 – 01 – 2023	1.1	Data from User and its validation Rendering the results Deployment Conclusion	Shubham

Contents

Document Version Control 2
Abstract 4
1. Introduction 5
1.1. Why this DPR Document? 5
2. General Description 5
2.1. Problem Perspective 5
2.2. Problem Statement 5
2.3. Proposed Solution 5
2.4. Further Improvements 5
3. Technical Requirements 6
3.1. Tools Used 6
4. Data Requirements 6
4.1. Data Collection 6
4.2. Data Description 6
4.3. Importing Data into Database 7
4.4. Exporting Data from Database 7
5. Data Pre-processing 7
6. Design Flow 8
6.1. Modelling Process 8
6.2. Deployment Process 8
6.3. Logging 9
7. Data from User 9
8. Data Validation 9
9. Rendering the result 9
10. Deployment 9
11. Conclusion 9

Abstract

The aviation market has seen significant changes and challenges in recent years. The COVID-19 pandemic has had a major impact on the industry, causing widespread disruptions and reductions in demand for air travel. Despite this, the industry is beginning to recover as vaccine rollouts and declining case numbers lead to increased consumer confidence. However, the industry is still facing challenges, such as shifting consumer preferences for travel and a rise in low-cost carriers. In response, many airlines and travel companies are investing in new technologies, such as flight fare prediction systems, to remain competitive and meet the evolving needs of travellers.

The prediction of flight fares is important because it helps travellers to make informed decisions when booking their travels. By having access to accurate and up-to-date predictions of flight fares, travellers can save money and avoid overpaying for their flights. Additionally, a prediction system for flight fares can also help airlines and travel companies to optimize pricing strategies and make data-driven decisions. In a highly competitive market, being able to predict flight fares can provide a significant advantage to both travellers and companies.

1. Introduction

1.1. Why this DPR Document?

The main purpose of this DPR documentation is to add the necessary details of the project and provide the description of the machine learning model and the written code. This also provides the detailed description on how the entire project has been designed end-to-end.

Key points:

- Describes the design flow
- Implementations
- Software requirements
- Architecture of the project
- Non-functional attributes like:
  - Reusability
  - Portability
  - Resource utilization

2. General Description

2.1. Problem Perspective

The flight fare prediction may be a machine learning model that helps users to predict the price of the flight tickets and help them to understand the price of their journey.

2.2. Problem Statement

After amendment of the new rules, there is changes in the flight fare price from one location to another. The main goal of the system is to create a model to predict the price of their flight fare based on bound input provided by user like date of journey, Source, Destination and many more.

2.3. Proposed Solution

To solve the problem, we have created a User interface for taking the input from the user to predict the flight fare price using our trained ML model after processing the input and at last the output (predicted value) from the model is communicated to the User.

2.4. Further Improvements

We also analyse the data used for training the ML model by considering different occasions such as Weekday, Season or any Social reasons, considering different angles of business. If we use such information and predict the discounted flight fare price, it will bring some loss to the airline companies, but users can benefit from that. If we develop these using the Business perspective of Airline, this technique isn’t thought - about.

3. Technical Requirements

As technical requirements, we don't need any specialized hardware for virtualization of the application. The user should have the device that has the access to the web and the fundamental understanding of providing the input.

3.1.Tools Used

- Python 3.9 is employed because of the programming language and frameworks like NumPy, Pandas, Scikit - learn and alternative modules for building the model.
- Jupyter - Notebook is employed as an IDE.
- For Data visualizations, seaborn and components of matplotlib are getting used.
- For information assortment prophetess info is getting used.
- Front end development is completed victimization HTML/CSS.
- Flask is employed for each information and backend readying.
- GitHub is employed for version management.
- Heroku is employed for streamlit.

4. Data Requirements

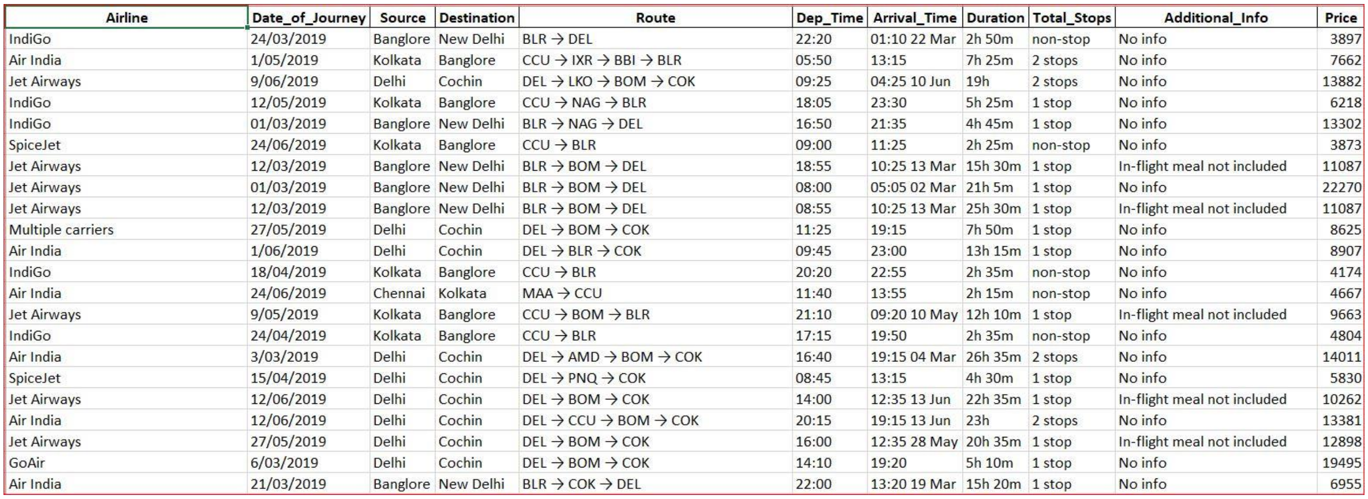
The Data requirements totally supported the matter statement and also the dataset is accessible on the Kaggle within the file format of (.xlsx).

4.1.Data Collection

The data for this project is collected from the Kaggle Dataset, the URL for the dataset is kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh

4.2.Data Description

Flight Fare Prediction is 10K+ dataset publicly available on Kaggle. The information in the dataset is present in 2separate excel files named as train.xlsx and test.xlsx. Dataset contains 10683 rows which shows the information such Date of Journey, Source, Destination, Arrival Time, Departure Time, Total stops, Airlines, Additional Info and Price. The glance of the Dataset is:



Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
SpiceJet	24/06/2019	Kolkata	Banglore	CCU → BLR	09:00	11:25	2h 25m	non-stop	No info	3873
Jet Airways	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	18:55	10:25 13 Mar	15h 30m	1 stop	In-flight meal not included	11087
Jet Airways	01/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:00	05:05 02 Mar	21h 5m	1 stop	No info	22270
Jet Airways	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:55	10:25 13 Mar	25h 30m	1 stop	In-flight meal not included	11087
Multiple carriers	27/05/2019	Delhi	Cochin	DEL → BOM → COK	11:25	19:15	7h 50m	1 stop	No info	8625
Air India	1/06/2019	Delhi	Cochin	DEL → BLR → COK	09:45	23:00	13h 15m	1 stop	No info	8907
IndiGo	18/04/2019	Kolkata	Banglore	CCU → BLR	20:20	22:55	2h 35m	non-stop	No info	4174
Air India	24/06/2019	Chennai	Kolkata	MAA → CCU	11:40	13:55	2h 15m	non-stop	No info	4667
Jet Airways	9/05/2019	Kolkata	Banglore	CCU → BOM → BLR	21:10	09:20 10 May	12h 10m	1 stop	In-flight meal not included	9663
IndiGo	24/04/2019	Kolkata	Banglore	CCU → BLR	17:15	19:50	2h 35m	non-stop	No info	4804
Air India	3/03/2019	Delhi	Cochin	DEL → AMD → BOM → COK	16:40	19:15 04 Mar	26h 35m	2 stops	No info	14011
SpiceJet	15/04/2019	Delhi	Cochin	DEL → PNQ → COK	08:45	13:15	4h 30m	1 stop	No info	5830
Jet Airways	12/06/2019	Delhi	Cochin	DEL → BOM → COK	14:00	12:35 13 Jun	22h 35m	1 stop	In-flight meal not included	10262
Air India	12/06/2019	Delhi	Cochin	DEL → CCU → BOM → COK	20:15	19:15 13 Jun	23h	2 stops	No info	13381
Jet Airways	27/05/2019	Delhi	Cochin	DEL → BOM → COK	16:00	12:35 28 May	20h 35m	1 stop	In-flight meal not included	12898
GoAir	6/03/2019	Delhi	Cochin	DEL → BOM → COK	14:10	19:20	5h 10m	1 stop	No info	19495
Air India	21/03/2019	Banglore	New Delhi	BLR → COK → DEL	22:00	13:20 19 Mar	15h 20m	1 stop	No info	6955

4.3.Importing Data into Database

Created associate API for the transfer of the info into the Cassandra info, steps performed are:

•Connection is created with the info.

•Created info with name Flight Info.

•The command is written for making the info table with needed parameters.

•And finally, a command is written for uploading the Knowledge Set into data table by bulk insertion.

4.4.Exporting Data from Database

In the above created API, the download URL is also being created, which downloads the data into a csv file format.

5. Data Pre-processing

•Checked for info of the Dataset, to verify the correct datatype of the Columns.

•Checked for Null values, because the null values can affect the accuracy of the model.

•Converted all the desired columns into Datetime format. Performed One Hot encoding on the desired columns.

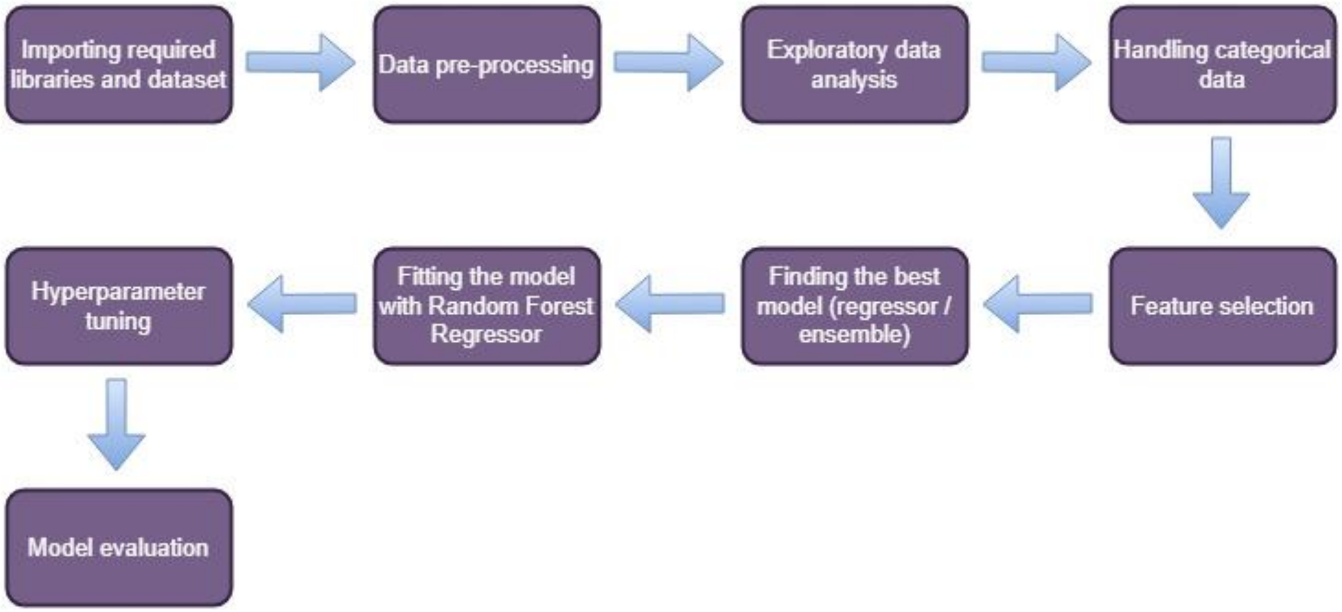
•Checking the distribution of the columns to interpret its importance.

Now, the info is prepared to train a Machine Learning Model.

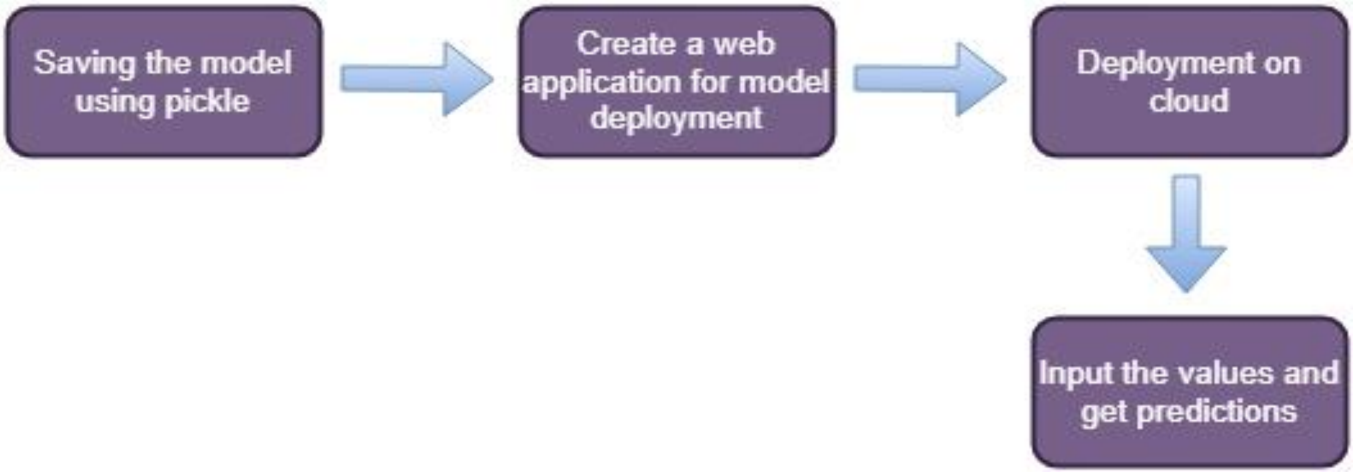
6. Design Flow

6.1.Modelling Process

After pre-processing the data, we visualize our data to gain insights and then these insights are randomly spread and split into two parts, train and test data. After splitting the data, we use Random Forest Regressor to model our data to predict the Flight Fare price.


Importing required libraries and dataset
Data pre-processing
Exploratory data analysis
Handling categorical data
Feature selection
Finding the best model (regressor / ensemble)
Fitting the model with Random Forest Regressor
Hyperparameter tuning
Model evaluation

6.2.Deployment Process


Saving the model using pickle
Create a web application for model deployment
Deployment on cloud
Input the values and get predictions

6.3.Logging

In logging, at each time an error or an exception occurs, the event is logged into the system log file with reason and timestamp. This helps the developer to debug the system bugs and rectify the error.

7. Data from User

The data from the user is retrieved from the created HTML web page.

8. Data Validation

The data provided by the user is then being processed by app.py file and validated. The validated data is then sent to the prepared model for the prediction.

9. Rendering the result

The data sent for the prediction is then rendered to the web page.

10. Deployment

The tested model is then deployed to Streamlit. So, users can access the project from any internet device.

11. Conclusion

The Flight Fare Prediction system will predict the price for helping the customers with the trained knowledge with a set of rules. The user can use this system to recognize the approximate value of its flight fare for his or her journey.

Frequently Asked Questions (FAQs)

1) What’s the source of data?
   The data for training is provided by the client in multiple batches and each batch contain multiple files.
2) What was the type of data?
   The data was the combination of numerical and Categorical values.
3) What’s the complete flow you followed in this Project?
   Refer Page no 6 for better Understanding.
4) After the File validation what you do with incompatible file or files which didn’t pass the validation?
   Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.
5) How logs are managed?
   We are using different logs as per the steps that we follow in validation and modelling like File validation log, Data Insertion, Model Training log, prediction log etc.
6) What techniques were you using for data pre-processing?
   - Removing unwanted attributes.
   - Visualizing relation of independent variables with each other and output variables.
   - Checking and changing Distribution of continuous values.
   - Removing outliers
   - Cleaning data and imputing if null values are present.
   - Converting categorical data into numeric values.
7) How training was done or what models were used?
   - Before dividing the data in training and validation set, we performed pre-processing over the data set and made the final dataset.
   - As per the dataset training and validation data were divided.
   - Algorithms like Linear regression, SVM, Decision Tree, Random Forest, XGBoost were used based on the recall, final model was used on the dataset and we saved that model.
8) How Prediction was done?
   The testing files are shared by the client. We Performed the same life cycle on the provided dataset. Then, based on dataset, model is loaded, and prediction is performed. In the end we get the accumulated data of predictions.
9) What are the different stages of deployment?
   First, the scripts are stored on GitHub as a storage interface.
   The model is first tested in the local environment.
   After successful testing, it is deployed on Heroku.