Detail Project Report (DPR)

FLIGHT FARE PREDICTION

Document Version: 0.3

Taskin Arshad

Document Version Control:

Version	Date	Author	Description
0.1	03-07-2023	Taskin	Introduction
0.2	03-07-2023	Dhiraj	Deployment and Process
0.3	03-07-2023	Taskin	Q and A

Contents

1 Introduction ---------- 3
1.1 What is High-Level design document ? ---------- 3
1.2 Scope ---------- 3
2 Description ---------- 3
2.1 Problem Perspective ---------- 3
2.2 Problem Statement ---------- 3
2.3 Purposed Solution ---------- 3
2.4 Solution Improvements ---------- 3
2.5 Technical Requirements ---------- 4
2.6 Data Requirements ---------- 4
2.7 Tool Used ---------- 4
2.8 Data Gathering ---------- 4
2.9 Data Description ---------- 4
3 Data Pre-Processing ---------- 4
4 Design Flow ---------- 5
4.1 Modelling ---------- 5
4.2 Modelling Process ---------- 6
4.3 Deployment Process ---------- 6
5 Data from User ---------- 6
6 Data Validation ---------- 6
7 Rendering Result ---------- 6
8 Conclusion ---------- 7
9 Q & A ---------- 7

1. Introduction

1.1. What is High-Level design document?

The main purpose of this HLD documentation is to feature the required details of the project and supply the outline of the machine learning model and also the written code. This additionally provides the careful description on however the complete project has been designed end-to-end.

1.2. Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

2. Description

2.1. Problem Perspective

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

2.2. Problem Statement

The most goal of the project is to form a programme that predicts the price of the flight price tag by taking bound input from the user like date of journey, aboard location and destination etc.

2.3. Purposed Solution

Projected to require the desired input of user from the created interface and method all the provided information to satisfy the wants of the machine learning model and at last show the output oral communication so and then quantity is that the expected value.

2.4. Solution Improvements

we will even predict the price of price tag considering whether or not is it a weekday, season or alternative social reasons. however, considering from the angle of business, if we have a tendency to

method such information and predict the price of the discounted price tag it'll bring some loss to the airlines company. therefore, this technique isn't thought-about.

2.5. Technical Requirements

There are not any hardware needs needed for victimization this application, the user should have AN interactive device that has access to the web and should have the fundamental understanding of providing the input. And for the backend half the server should run all the package that's needed for the process the provided information and to show the results.

2.6. Data Requirements

The info demand is totally supported the matter statement. and also, the information set is accessible on the Kaggle within the type of standout sheet(.xlsx). because the main theme of the project is to induce the expertise of real time issues, we have a tendency to ar once more mercantilism {the information into the prophetess data base and commerce it into csv format.

2.7. Tool Used

- Python 3.9 is employed because the programming language and frame works like numpy, pandas, sklearn and alternative modules for building the model.
- PyCharm is employed as IDE.
- For visualizations seaborn and components of matplotlib are getting used.
- For information assortment prophetess info is getting used.
- Front end development is completed by Streamlit.
- GitHub is employed for version management.
- Heroku is employed for deployment.

2.8. Data Gathering

The data for the current project is being gathered from Kaggle dataset, the link to the data is: https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh

2.9. Data Description

There are about 10k+ records of flight information such as airlines, data of journey, source, destination, departure time, arrival time, duration, total stops, additional information, and price. A glance of the dataset is shown below.

	Airline	e_of_Jour	Source	Destinatior	Route	Dep_Time	rrival_Tim	Duration	Total_Stop	ditional_Ir	Price
2	IndiGo	24/03/201	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 M	2h 50m	non-stop	No info	3897
3	Air India	1/05/2019	Kolkata	Banglore	CCU → IXF	05:50	13:15	7h 25m	2 stops	No info	7662
4	Jet Airway	9/06/2019	Delhi	Cochin	DEL → LKC	09:25	04:25 10 J	19h	2 stops	No info	13882
5	IndiGo	12/05/201	Kolkata	Banglore	CCU → NA	18:05	23:30	5h 25m	1 stop	No info	6218
6	IndiGo	01/03/201	Banglore	New Delhi	BLR → NA	16:50	21:35	4h 45m	1 stop	No info	13302
7	SpiceJet	24/06/201	Kolkata	Banglore	CCU → BLI	09:00	11:25	2h 25m	non-stop	No info	3873
8	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BOI	18:55	10:25 13 M	15h 30m	1 stop	In-flight m	11087
9	Jet Airway	01/03/201	Banglore	New Delhi	BLR → BOI	08:00	05:05 02 M	21h 5m	1 stop	No info	22270
10	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BOI	08:55	10:25 13 M	25h 30m	1 stop	In-flight m	11087
11	Multiple c	27/05/201	Delhi	Cochin	DEL → BOI	11:25	19:15	7h 50m	1 stop	No info	8625
12	Air India	1/06/2019	Delhi	Cochin	DEL → BLF	09:45	23:00	13h 15m	1 stop	No info	8907
13	IndiGo	18/04/201	Kolkata	Banglore	CCU → BLI	20:20	22:55	2h 35m	non-stop	No info	4174
14	Air India	24/06/201	Chennai	Kolkata	MAA → CC	11:40	13:55	2h 15m	non-stop	No info	4667
15	Jet Airway	9/05/2019	Kolkata	Banglore	CCU → BO	21:10	09:20 10 M	12h 10m	1 stop	In-flight m	9663
16	IndiGo	24/04/201	Kolkata	Banglore	CCU → BLI	17:15	19:50	2h 35m	non-stop	No info	4804
17	Air India	3/03/2019	Delhi	Cochin	DEL → AM	16:40	19:15 04 M	26h 35m	2 stops	No info	14011
18	SpiceJet	15/04/201	Delhi	Cochin	DEL → PN	08:45	13:15	4h 30m	1 stop	No info	5830
19	Jet Airway	12/06/201	Delhi	Cochin	DEL → BOI	14:00	12:35 13 J	22h 35m	1 stop	In-flight m	10262

3. Data Pre-processing

Steps performed in pre-processing are:

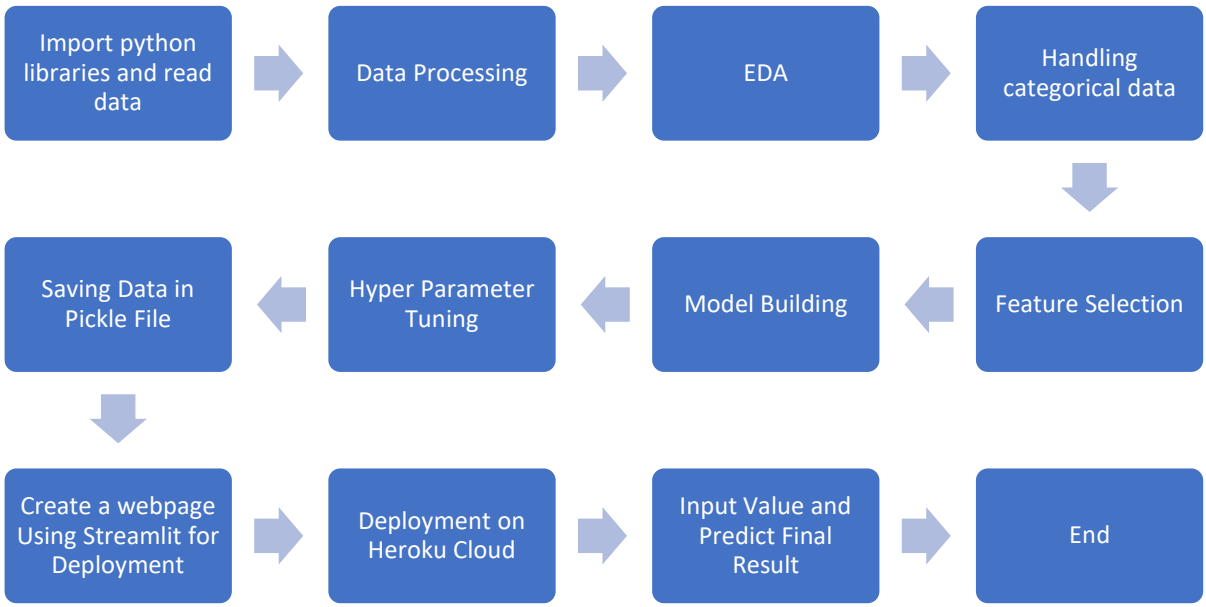
- First the info sorts square measure being checked and located solely the value column is of sort number.
- Checked for null values as there square measure few null values, those rows square measure born.
- Converted all the desired column into the date time format.
- Performed one-hot cryptography for the desired columns.
- Scaling is performed for needed information.
- And, the info is prepared for passing to the machine learning formula

4. Design Flow

4.1. Modelling

The pre-processed data is then visualized and all the required insights are being drawn. Although from the drawn insights, the data is randomly spread but still modelling is performed with different machine learning algorithms to make sure we cover all the possibilities. And finally, as expected random forest regression performed well and further hyperparameter tuning is done to increase the model's accuracy.

4.2. Modelling and Deployment Process



1.1. Data from User

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

1.2. Data Validation

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

1.3. Rendering Result

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

1.4. Deployment

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

5. Conclusion

The flight fare prediction will predict the worth supported the trained knowledge set within the rule. therefore, the user will recognize the approximate value for his or her journey.

6. Q & A

Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files.

Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer Page no 6 for better Understanding.

Q 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.

Q 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.

Q 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.
-

Q 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.
-

Q 8) How Prediction was done?

The testing files are shared by the client. We Performed the same life cycle on the provided dataset. Then, on the basis of dataset, model is loaded and prediction is performed. In the end we get the accumulated data of predictions.

Q 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.