

# Architecture

## FLIGHT FARE PREDICTION

Revision Number – 1.3  
Last Date of Revision – 12/1/2022

Somay, Madhu

## Document Version Control

Date	Version	Description	Author
10-1-2022	1.0	Abstract, Introduction	Somay
11-1-2022	1.1	Architecture	Somay
12-1-2022	1.2	Architecture Design	Somay

## Contents

<b><u>Abstract</u></b>	<b><u>4</u></b>
<b><u>Introduction</u></b>	<b><u>4</u></b>
<b><u>Why this Architecture Design documentation?</u></b>	<b><u>4</u></b>
<b><u>1 Architecture</u></b>	<b><u>4</u></b>
<b><u>2 Architecture design</u></b>	<b><u>5</u></b>
<b><u>2.1 Data gathering from main source</u></b>	<b><u>5</u></b>
<b><u>2.2 Data description</u></b>	<b><u>5</u></b>
<b><u>2.3 Upload data into Cassandra</u></b>	<b><u>5</u></b>
<b><u>2.4 Export data from database</u></b>	<b><u>5</u></b>
<b><u>2.5 Data pre-processing</u></b>	<b><u>5</u></b>
<b><u>2.6 Modelling</u></b>	<b><u>6</u></b>
<b><u>2.7 UI integration</u></b>	<b><u>6</u></b>
<b><u>2.8 Data from user</u></b>	<b><u>6</u></b>
<b><u>2.9 Data validation</u></b>	<b><u>6</u></b>
<b><u>2.10 Rendering the results</u></b>	<b><u>6</u></b>
<b><u>2.11 Deployment</u></b>	

## Abstract

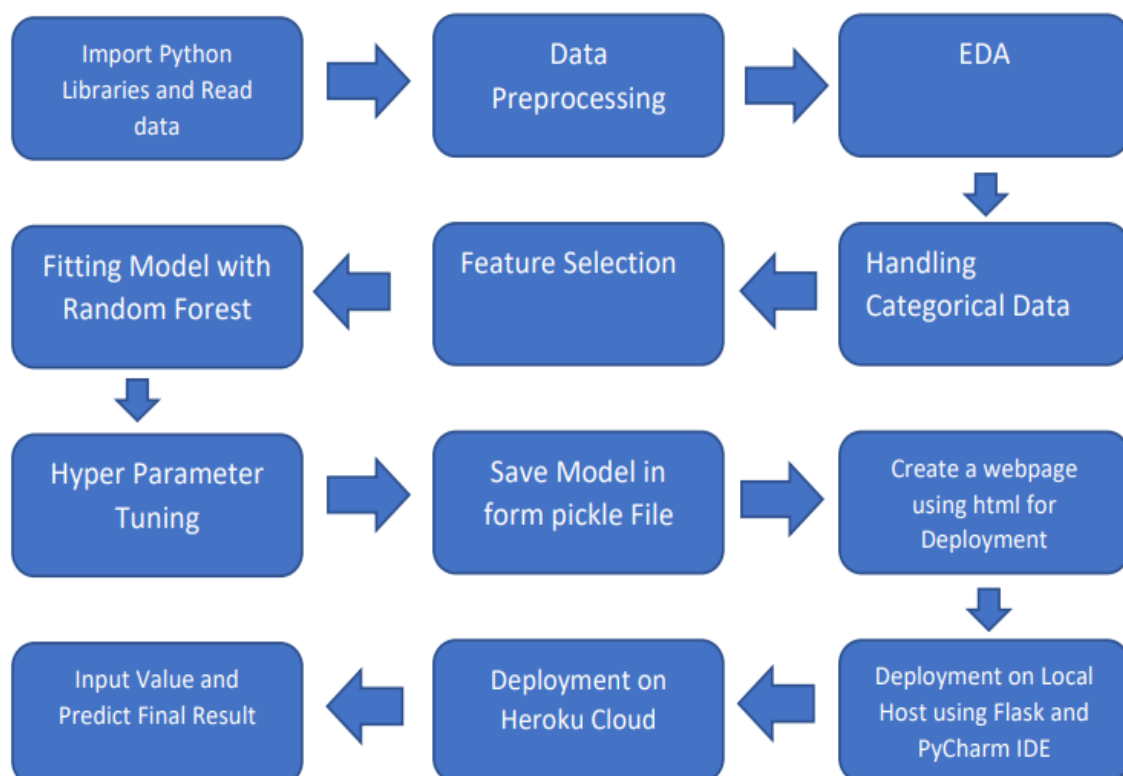
The recent global situations had a huge impact on the aviation sector due to many reasons. This impact has two category people, the first is business perspective and the second is the customers perspective. As safety is the major reason for such impact on the aviation sector, the governments around the world amended different rules to their respective airlines companies. These restrictions had made the availability of the flights and their attendee capacity less. Taking all these factors in consideration the cost of the flight tickets has increased and vary from one place to the other. Booking a flight ticket has split into two, one is the online and the other is the offline bookings. Both these have their respective criteria for cost of the ticket, one such example is the server load and the number of booking requests. In this machine learning implementation, we will see various factors that impact the cost of the flight ticket and predict the appropriate price of the ticket.

## Introduction

### Why this Architecture Design documentation?

The main objective of the Architecture design documentation is to provide the internal logic understanding of the flight fare prediction code. The Architecture design documentation is designed in such a way that the programmer can directly code after reading each module description in the documentation.

## 1 Architecture



## 2 Architecture design

This project is to create an interface for the user to know their approximate flight ticket price, in addition to this, in need of getting the real time project experience we are importing the gathered data into our own database and then start the project from the scratch.

### 2.1 Data gathering from main source

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

### 2.2 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:

	A	B	C	D	E	F	G	H	I	J	K	L	M
	Airline	Date of Journey	Source	Destination	Route	Dep. Time	Arrival Time	Duration	Total Stops	Additional Info	Price		
1	IndiGo	24/03/2019	Bangalore	New Delhi	BLR → DEL	22:20	01:10 22	12h 50m	non-stop	No info	3897		
2	Air India	1/05/2019	Kolkata	Bangalore	CCU → IXF	05:50	13:15	7h 25m	2 stops	No info	7662		
3	Jet Airway	9/06/2019	Delhi	Cochin	DEL → LKO	09:25	04:25 10	19h	2 stops	No info	13882		
4	IndiGo	12/05/2019	Kolkata	Bangalore	CCU → NA	18:05	23:30	5h 25m	1 stop	No info	6218		
5	IndiGo	01/03/2019	Bangalore	New Delhi	BLR → NA	16:50	21:35	4h 45m	1 stop	No info	13302		
6	SpiceJet	24/06/2019	Kolkata	Bangalore	CCU → BLI	09:00	11:25	2h 25m	non-stop	No info	3873		
7	Jet Airway	12/03/2019	Bangalore	New Delhi	BLR → BOI	18:55	10:25 13	15h 30m	1 stop	In-flight m	11087		
8	Jet Airway	01/03/2019	Bangalore	New Delhi	BLR → BOI	08:00	05:05 02	12h 5m	1 stop	No info	22270		
9	Jet Airway	12/03/2019	Bangalore	New Delhi	BLR → BOI	08:55	10:25 13	12h 30m	1 stop	In-flight m	11087		
10	Multiple c	27/05/2019	Delhi	Cochin	DEL → BOI	11:25	19:15	7h 50m	1 stop	No info	8625		
11	Air India	1/06/2019	Delhi	Cochin	DEL → BLF	09:45	23:00	13h 15m	1 stop	No info	8907		
12	IndiGo	18/04/2019	Kolkata	Bangalore	CCU → BLI	20:20	22:55	2h 35m	non-stop	No info	4174		
13	Air India	24/06/2019	Chennai	Kolkata	MAA → CC	11:40	13:55	2h 15m	non-stop	No info	4667		
14	Jet Airway	9/05/2019	Kolkata	Bangalore	CCU → BO	21:10	09:20 10	12h 10m	1 stop	In-flight m	9663		
15	IndiGo	24/04/2019	Kolkata	Bangalore	CCU → BLI	17:15	19:50	2h 35m	non-stop	No info	4804		
16	Air India	3/03/2019	Delhi	Cochin	DEL → AM	16:40	19:15 04	12h 35m	2 stops	No info	14011		
17	SpiceJet	15/04/2019	Delhi	Cochin	DEL → PN	08:45	13:15	4h 30m	1 stop	No info	5830		
18	Jet Airway	12/06/2019	Delhi	Cochin	DEL → BOI	14:00	12:35 13	12h 35m	1 stop	In-flight m	10262		

### 2.3 Upload data into Cassandra

Created an api for the upload of the data into the Cassandra database, steps performed are:

- Connection is made with the database.
- Created a database with name flightfare.
- Cqlsh command is written for creating the data table with required parameters.
- And finally, a cqlsh command is written for uploading the dataset into data table by bulk insertion.

### 2.4 Export data from database

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

### 2.5 Data pre-processing

Steps performed in pre-processing are:

- First the data types are being checked and found only the price column is of type integer.
- Checked for null values as there are few null values, those rows are dropped.
- Converted all the required column into the date time format.
- Performed one-hot encoding for the required columns.
- Scaling is performed for required data.

- And, the data is ready for passing to the machine learning algorithm.

## 2.6 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.

## 2.7 UI integration

Both CSS and HTML files are being created and are being integrated with the created machine learning model. All the required files are then integrated to the app.py file and tested locally.

## 2.8 Data from user

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

## 2.9 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.

## 2.10 Rendering the results

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

## 2.11 Deployment

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