ARCHITECTURE DESIGN

Flight Fare Prediction

DOCUMENT VERSION CONTROL

DATE ISSUED	VERSION	DESCRIPTION	AUTHOR
	V1.0	Architecture Design-V1.0	ADHIRAJ SINGH SHEKHAWAT

Abstract

The recent changes in the international market had a large impact on the Aviation sector because of several reasons. These impact the two class folks, the first is Business perspective and second is Customer perspective. The major reason for such an impact is the governments around the world amended totally different rules to their various Airline firms. Taking these factors into consideration, the value of the flight tickets has varied from one place to another. Booking a flight ticket has its price tag split into two, one is online bookings and other is offline bookings. Each of these have their various criteria for value of the price, one such example is the server load and therefore the range of booking requests. During this machine learning implementation, we are going to see numerous factors that impact the price of the flight ticket and predict the acceptable price of the ticket.

1. Introduction

1.1 What is Architecture Design?

The goal of Architecture Design (AD) is to give the internal design of the actual program code for the 'Flight Fare Prediction'. AD describes the class diagrams with the methods and relation between classes and program specification. It describes the modules so that the programmer can directly code the program from the document.

1.2 Scope

Architecture Design (AD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software, architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work. And the complete workflow.

1.3 Constraints

We predict the expected estimating cost of expenses customers based on some personal health information.

2. Technical Specification

2.1 Dataset

Flight Fare Prediction is 10K+ dataset publicly available on Kaggle. The information in the dataset is present in two separate 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 dataset looks like as follow:

```
train_data.head(3)
```

	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

The data set consists of various data types from integer to floating to object as shown in Fig

```
train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          10683 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10683 non-null  object
 3   Destination      10683 non-null  object
 4   Route            10682 non-null  object
 5   Dep_Time         10683 non-null  object
 6   Arrival_Time     10683 non-null  object
 7   Duration         10683 non-null  object
 8   Total_Stops      10682 non-null  object
 9   Additional_Info  10683 non-null  object
 10  Price            10683 non-null  int64 
dtypes: int64(1), object(10)
memory usage: 918.2+ KB
```

Pre-processing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing them with supported appropriate data types so that analysis and model fitting is not hindered from their way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and a one-hot encoding scheme during the model building.

2.2 Logging

We should be able to log every activity done by the user

- The system identifies at which step logging require.
- The system should be able to log each and every system flow.
- The system should be not be hung even after using so much logging. Logging just because we can easily debug issuing so logging is mandatory to do.

2.3 Deployment

For the deployment of the project, AWS is used

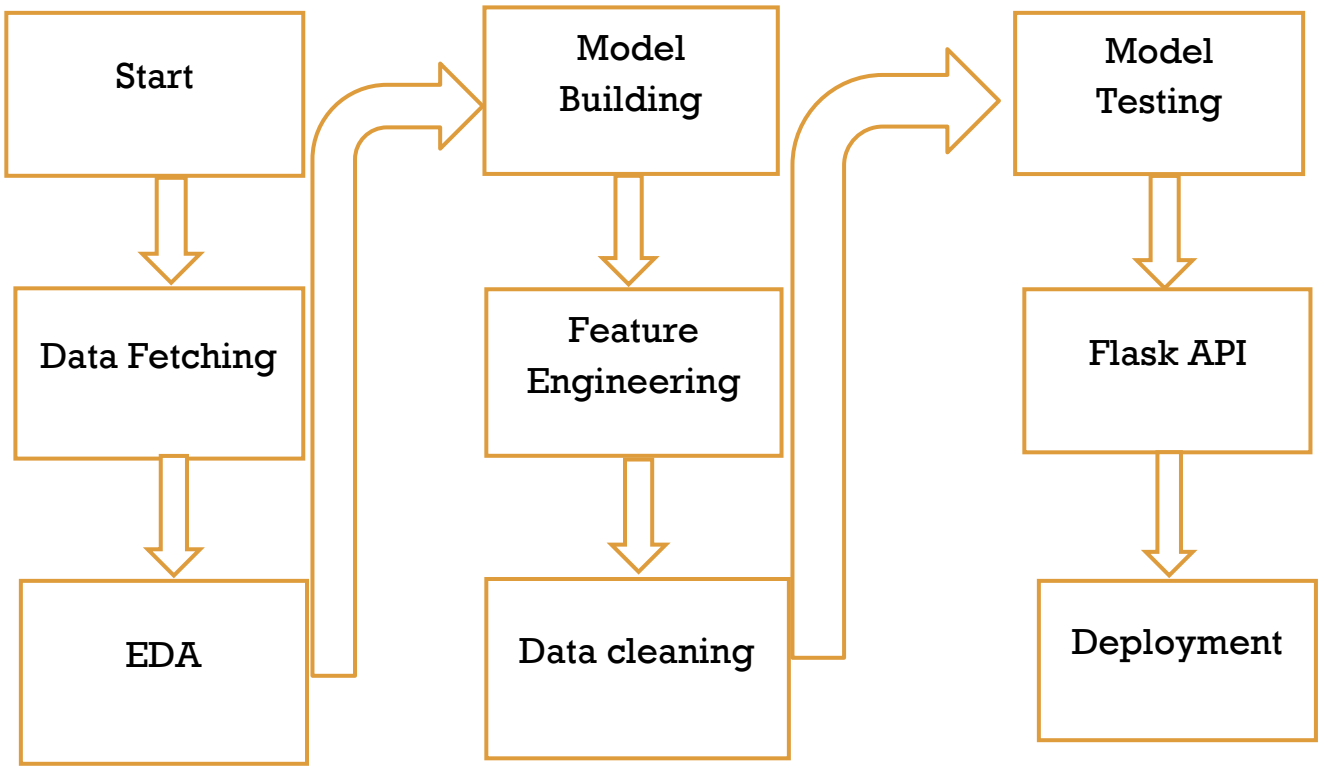
3. Technology Stack

Front End	Html
Backend	Python/Flask
Deployment	**AWS**

4. Proposed Solution

The solution of the this problem statement if perform EDA on the dataset to generate meaningful insights from the data and use this data to hyper tune with appropriate machine learning algorithms which will have the maximum accuracy in predicting the Flight Fare. Thus creating a user interface where a user can put in the various features of the data which will in return give the flight fare.

5. Architecture



5.1 Data Gathering

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

5.2 Raw Data Validation

After data is loaded, various types of validation are required before we proceed further with any operation. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because the attributes which contain these are of no use. It will not play role in contributing to the estimating the Flight Fare.

5.3 Exploratory Data Analysis

Visualized the relationship between the dependent and independent features. Also checked relationship between independent features to get more insights about the data.

5.4 Feature Engineering

After pre-processing, relevant features are selected and engineered to improve the performance of the machine learning model. This can involve techniques such as feature selection, dimensionality reduction, and feature scaling. Some of the features were showing skewness and even after removing the skewness using log transformation 4 columns were dropped. The outliers were also removed by assigning upper limit values to outliers above the limit and similarly lower limit values to outliers below the lower limit.

5.5 Model Building

After doing all kinds of pre-processing operations mention above and performing scaling and encoding, the data set is passed through a pipeline to all the mode Bagging, Random Forest, Gradient boost, Xgboost, Adaboost Classifier algorithms . It was found that Xgboost gives the best accuracy

5.6 Model Saving

Model is saved using pickle library in pickle format.

5.7 Flask API for Web Application

After saving the model, the API building process started using Flask . Web application creation was created in Flask for testing purpose. Whatever user will enter the data and then that data will be extracted by the model to estimate the Flight Fare, this is performed in this stage.

5.8 GitHub

The whole project directory will be pushed into the GitHub repository.

5.9 Deployment

The project was deployed from GitHub into Aws.

6. User Input / Output Workflow