Detail project report

FLIGHT FARE PREDICTOR

Table of Contents

[1) Objective 3](#_Toc89425817)

[2) Benefits 3](#_Toc89425818)

[3) Data Sharing Agreement 3](#_Toc89425819)

[4) Architecture 4](#_Toc89425820)

[5) Data Validation and Data Transformation 5](#_Toc89425821)

[6) Model Training 5](#_Toc89425823)

[6.1) Load Dataset 5](#_Toc89425824)

[6.2) Data Preprocessing 5](#_Toc89425825)

[6.3) Model Training Part 7](#_Toc89425826)

[6.4) Prediction 8](#_Toc89425827)

[7) Q & A 8](#_Toc89425828)

# 

# 1) Objective

Development of a predictive model which can predict the flight fare of the flight just precisely.

# 2) Benefits

1) It will help customer to save the money

2)They can have the detailed overview of their expenses and they can manage their budget in a efficient manner.

3) It will even save the time

# 3) Data Sharing Agreement

* File name can be anything
* Columns inside the file will be 10
* Columns are [Airline,Date\_Of\_Journey,Source,Destination,Route,Dep\_Time,Arrival\_Time,Duration,Total\_Stops,Additional\_Info]
* Columns datatypes are [string,string,string,string,string,string,string,string,string,string]

# 4) Architecture

# 5) Data Validation and Data Transformation

* Name Validation: We don’t have any problem with the name of the file which will be given by the client.
* Null Values in the column:There is a rare chance that there can be a null value because it is airline industry and every detail is compulsory for the customer but also unfortunately if we got any rows then we are simply going to delete the row.

# 6) Model Training

## 6.1) Load dataset

All the data is stored in notebooks/data folder as train\_data.csv.

## 6.2) Data Preprocessing

1. First I checked whether is there any null values present inside the data and I found that there is a single row only that’s why I directly removed it.
2. After removing the null values Now, My main task was to handle categorical data and the data was mostly categorical. I didn’t use One hot encoding as there are too many
3. categorical data and if I had used OHE then it could cause a problem in my dataset that is a dataset with too many features and that is not good for model creation. So I thought to check the different categorical columns with the mean of the price. For example think Airline column in that we have Indigo, Trujet, and many more what I did is that I group by each categorical column and after that, I find out the mean price of flight fare. The categorical data with the highest mean price I replaced with 1 and the categorical data with second-highest mean price with 2 and so on, This I have done for the Airline, Destination, Source, Total\_Stops, Additional\_Info, etc.
4. Now after handling the categorical column there was a column called Route which was actually of no use in model creation so I removed the column.
5. Date\_of\_Journey This column contains the date of the journey so I created three columns from there first is Journey\_ day and the second is Journey\_month and Journey\_year.
6. Column Dep\_Time This column was containing time only I created the data of Dep\_Time into only Dep\_hours and Dep\_Min.
7. Column Arrival\_Time This column was containing time only I created the data of Arrival \_Time into only Arrival \_hours and Arrival \_Min

8) Handling Duration column, it was converted to a number.

## 6.3) Model Training Part

* In model training I tried a few algorithms like Ridge, Lasso, Decision Tree Regressor, and Random Forest Regressor because I know that they give the best score than others.
* Ridge, Lasso, Decision Tree Regressor were not giving a good score that’s why I removed it from my list Random Forest Regressor was giving the best result , I finally

decided to keep only Random Forest Regressor as my model.

* I saved my model as model.pickle.
* Now I created an API for my model using Flask.
* Finally My project is created and I am going to deploy it to the AWS



## 6.4) Prediction

1) Now its time for prediction part, Firstly when user will enter all of his/her details then a validation is performed that they entered correct format or not.

2) Data preprocessing is performed same as while training, I am handling the categorical column same way as in training.

3) Now everything is just simple I am loading my pickle model and applying model.predict

4) The output will be show on the webpage at the bottom of the form.

# 8) Q & A

1) What is the Source of Data?

Ans: Kaggle is the source of the data link: <https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh>

2) What was the type of Data?

Ans: Data type of every column is string.

3) How logs are managed?

Ans: I am managing the logs using logging module. After the prediction is completed the log file is shared with the client, if in case exception occurs inside the program then also the log file is shared to the file location provided.

4) What techniques are you using for data preprocessing??

Ans: See Data Preprocessing above I have mentioned there in detail.

5) How training was done?

Ans: After a lot of research I got to know that my data fits very good with Random Forest Regressor algorithm so I have selected Random Forest Regressor for my model training.