DETAIL PROJECT REPORT

FLIGHT FARE ESTIMATOR

Table of Contents

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

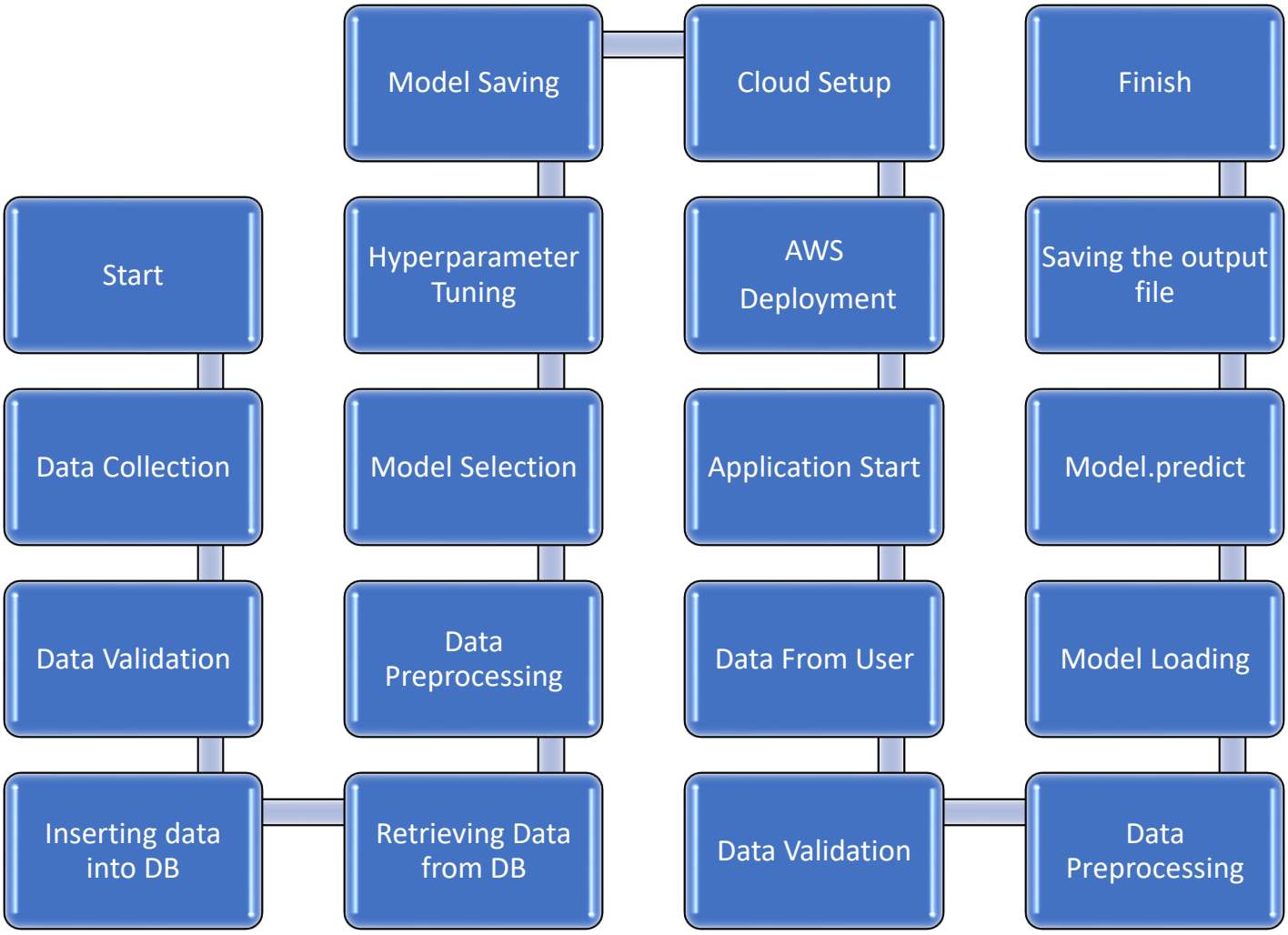
2)They can have the detailed overview of their expenses and they can manage their budget in a efficient manner.

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.
- Number of columns: If the number of columns inside the testing file doesn't matches, then the file is moved to Bad_Data_Folder.
- Name of columns: The name of the columns is validated and If the number of columns inside the testing file doesn't matches,then the file is moved to Bad_Data_Folder.
- Data type of columns:The data type of the colums is given in the schema file and if it doesn't matches with the testing schema then the file is moved to Testing_Batch_Files/Bad_Data.
- 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) Data Insertion in Database

a) Table Creation: I am using Cassandra Database for my project. Inside keyspace training. I am creating a table named all_train_files, table will only be created if it doesn't exist

b) Data Insertion: All the files which are inside Testing_Batch_Files/Good_Data are inserted into the database.

7) Model Training

7.1) Data export from db

All the data from the database is retrieved and stored into a csv file called Training_file.csv.

7.2) 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 all the data was of the year 2019 so I created two columns from there first is Journey_ day and the second is Journey_month.

6) Column Dep_Time This column was containing time only I created the data of Dep_Time into only hours, For example, if time is 08:30 it will be 8.5.

7) Arrival_Time I dropped this column as there was already a column called Duration Hours , So having two columns representing the same information or data is not good, So I kept only the Duration column.

8) Handling Duration column it was in the format of 8h 45m, I created a new columnuration_Hours and converted the data into 8.75 for 8h 45m and so on for every data.

7.3) Model Training Part

- In model training I tried a few algorithms like SVR, XGBRegressor, and Random Forest Regressor because I know that they give the best score than others.

- SVR was not giving a good score that's why I removed it from my list Random Forest Regressor was giving the same mean absolute error as XGBRegressor, I finally decided to keep only XGB as my model as it takes less time for training than rf and I had to do hyperparameter tuning also which will take a lot of time.

- But I my mind I got a idea to cluster my training data using Kmeans clsuter and apply different algorithms for each of my cluster. I tried this also but this thing didn't give good results like XGB so I dropped using this idea.

- After choosing model I did parameter tuning of my model using Grid Search CV and it helped me to low down my mean absolute error to 660 from 1100 which was previous mae without any hyperparameter tuning.

- 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

7.4) Prediction

1) Now its time for prediction part, First when I will get the file location for testing, I am validating each file with the testing schema mentioned .

2) Data preprocessing is performed same as while training, I am handling the categorical column with a pickle file which I had stored “handle_categorical_col_for_testing.pickle”.

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

4) The output file is then copied to the file location provided by the client and with that the log file is also copied.

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 xgboost algorithm so I have selected xgboost algorithm for my model training.

6) What are stages of Deployment?

Ans: Deployment has been done to Amazon Web Services and I have deployed the applicaiton in the production server. Link http://ec2-3-141-18-169.us-east-2.compute.amazonaws.com:5000/