DETAIL PROJECT REPORT

FLIGHT FARE PREDICTION

Table of Contents :

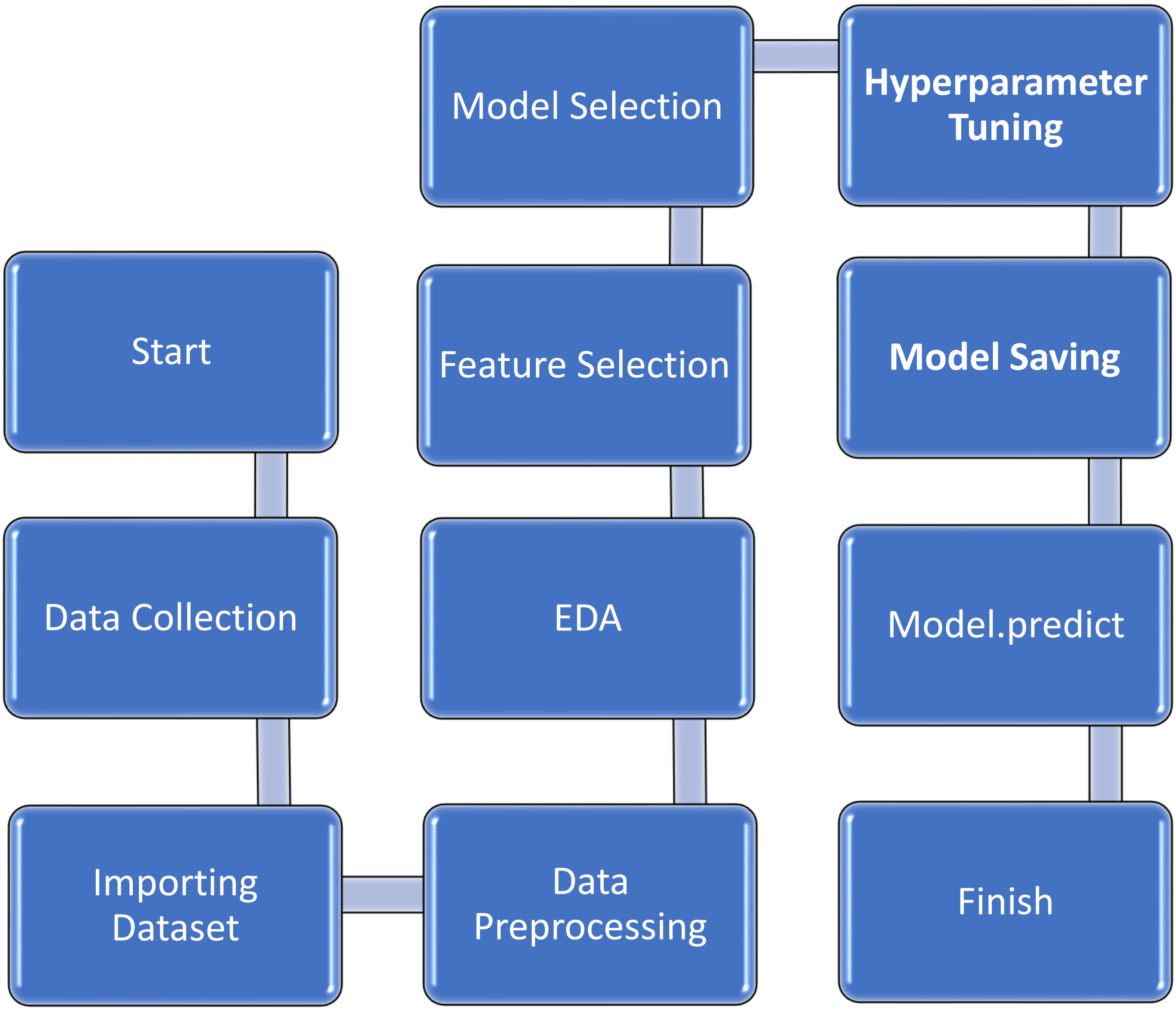
1) Objective

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

2) Benefits

- It will help customer to save the money
- They can have the detailed overview of their expenses and they can manage their budget in a efficient manner.
- It will even save the time

3) Architecture



4) Architecture Description

4.1) Data Collection

- We have 14k Dataset row columnar data includes the flight service, flight fare, number of stops, total number of duration, departure-arrival date and all. These is given in the Excel file format (.xlsx). These data is collected from the Kaggle which contains both the test data and train data.

4.2) Importing dataset

- Since data is in form of excel file we have to use pandas read_excel to load the data.

4.3) 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. 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.

1. 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.

2. 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.

3. 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.

4. Now, My main task was to handle categorical data and the data was mostly categorical. One can find many ways to handle categorical data. Some of them categorical data are,
   - Nominal data → data are not in any order → OneHotEncoder is used in this case.
   - Ordinal data → data are in order → LabelEncoder is used in this case.

5. 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.

4.4) Model Training Part

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

- Random Forest Regressor was giving me less **Coefficient Of Determination (R^2)** – **0.8122** as compare to XGBRegressor **Coefficient Of Determination (R^2) – 0.8463**. 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.

- After choosing model I did parameter tuning of my model using RandomizedSearchCV and it helped me to Increase my **Coefficient Of Determination (R^2)** to **0.8514** from **0.8463** which was previous R^2 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 Hereko Platform.

5) 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) 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 Heroku Platform and I have deployed the applicaiton in the production server. Link - https://flight-fare-prediction-kd.herokuapp.com/