Flight Ticket Price Prediction

Objectives

Flight ticket prices can be something hard to guess. we have been provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities, using which we aim to build a model which predicts the prices of the flights using various input features.

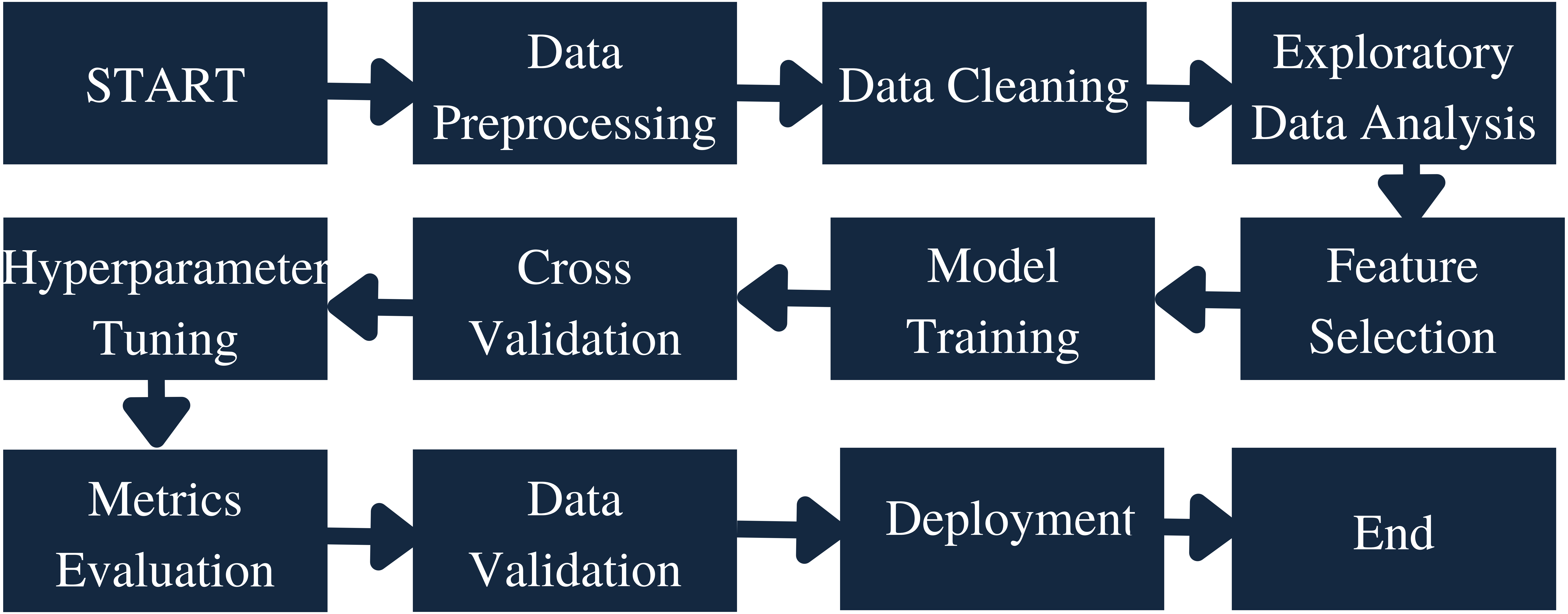
Benefits

- Traveler get the fare prediction handy using which it’s easy to decide the airlines.
- Saves time in searching / deciding for airlines.

Data Sharing Agreement

- Train Dataset: **Data_Train.xlsx**
- Test Dataset: **Test_set.xlsx**
- Number of Columns: **11 (Data_Train.xlsx) and 10 (Test_set.xlsx)**
- Number of Rows: **10683 (Data_Train.xlsx) and 2671 (Test_set.xlsx)**
- Column Names: **Airline, Date_of_Journey, Source, Destination, Route, Dep_Time, Arrival_Time ,Duration, Total_Stops, Additional_Info, Price.**

Architecture



START

Data Preprocessing

Data Cleaning

Exploratory Data Analysis

Feature Selection

Model Training

Cross Validation

Hyperparameter Tuning

Metrics Evaluation

Data Validation

Deployment

End

Step - 1 :Data Preprocessing and EDA

- The data is in the form of excel file, so we have to use pandas read_excel to load data
- After loading it is important to check the complete information of data as it can indication many of the hidden infomation such as null values in a column or a row
- Check whether any null values are present, if Yes, then following can be done:
    - Imputing data using Imputation method in sklearn
    - Filling NaN values with mean, median and mode using fillna() method
- Describe data --> which can give statistical analysis

Step - 2 : Feature Engineering

- Handling Categorical Data
  - 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.
- From .info() we know that Date_of_Journey is a object data type. To deal with this categorical data we use feature extraction method where we derived new features using Date_of_Journey.
- For this we require pandas to_datetime to convert object data type to datetime dtype.so we get two new feature that is:
  - Day_of_Journey & month_of _Journey

Step - 2 : Feature Engineering

- Similary, we extract values from Dep_Time and Arrival_Time and create separate columns for departure hours[”Dep_hour”] and minutes [“Dep_min”], Arrival_hour, Arrival_min.
- Since we have converted Date_of_Journey, Dep_Time and Arrival_Time columns into integers, Now we can drop it as it is of no use.
- Duration column is not in an appropriate form to help predict machine learning model. We have to bring it in a same format. There in dataset some flight duration could be just “30m” so we will write it as “0h 30m”. So we extract two new cloumns “Duration_hours” and “Duration_mins” from “Duration”.

Step - 3 : Feature Selection

- Finding out the best feature which will contribute and have good relation with target variable. Some of the feature selection methods are:
  - Heatmap
  - feature_importance_
  - SelectKBest
- Important feature using ExtraTreesRegressor model.
- Fitting model using Random Forest. You can also use **distplot()**to fit a parametric distribution to a dataset and visually evaluate how closely it corresponds to the observed data.

Step - 4 : Hyperparameter Tuning

- RandomizedSearchCV is used for Hyperparameter Tuning as they are fast in their process.
- Assign hyperparameters in the form of dictionary.
- Fitting the model
- Checking the best parameters and best scores
    - MAE (Mean Absolute Error)
    - MSE (Mean Squared Error)
    - RMSE (Root Mean Squared Error)

Prediction and Deployment

- I made an WebApp with the Help of HTML, CSS and Bootstrap as front-end tool and used the RF model and deployed my App.
- The model used to predict the Price of the Flight tickets and based on the old data about prices, stops, arrival and destination, we predict the prices of tickets.

Frequent Q&A

Q) What is the source of data?

- Data was collected from Kaggle, but the specific data are collected from github.

Q) What is the complete flow of your project?

- Refer to slide 4 for better understanding.

Q) What were the techniques you used for data-preprocessing?

- In data-preprocessing, I analyzed the data, found the important features, and based on the domain knowledge, I eliminated the unnecessary columns. I also tried to fill missing values with mean, median and mode but still the data have the same correlations. Thus removing the columns with high NaN values was the better option for me.

Frequent Q&A

Q) How did you choose the model?

- After implementing hyperparameter tuning, I was able to do model selection based on the metrics and how it was performing on unseen data. The final model I chose was Random Forest Classifier