# Problem Statement:

Travelling through flights has become an integral part of today’s lifestyle as more and more people are opting for faster travelling options. The flight ticket prices increase or decrease every now and then depending on various factors like timing of the flights, destination, and duration of flights various occasions such as vacations or festive.

# Data Description

Data is in form of excel file, for reading data though this we will use pandas. Below is column wise description of dataset.

|  |  |  |
| --- | --- | --- |
| **Serial Number** | **Column** | **Type** |
| 1 | Airline | Object |
| 2 | Date\_of\_Journey | Object |
| 3 | Source | Object |
| 4 | Destination | Object |
| 5 | Route | Object |
| 6 | Dep\_Time | Object |
| 7 | Arrival\_Time | Object |
| 8 | Duration | Object |
| 9 | Total\_Stops | Object |
| 10 | Additional\_Info | Object |
| 11 | Price | Int64 |

After loading it we need to check complete information of data as it can be indication many of them may have null rows, columns or values.

Check whether any null values are there or not If preent the following needs to be done:

* + Imputing data using imputation method in sklearn.
  + Filling NaN values with mean, median and mode using fillna() method.

### EDA

From description we can see that Data\_of\_Journey is a object data type, Therefore, we have to convert this datatype into timestamp so as to use this column properly for prediction.

For this we require pandas to\_datetime to convert object data type to datetime dtype.

.dt.day method will extract only day of that date.

.dt.month method will extract only month of that date.

Arrival time is when the plane pulls up to the gate. Similar to Date\_of\_Journey we can extract values from Arrival\_Time.

Time taken by plane to reach destination is called Duration. It is the difference between Departure Time and Arrival time.

### Handling Categorical Data

One can find many ways to handle categorical data. Some of them categorical data are,

1. Nominal data: data are not in order --> OneHotEncoder is used in this case.
2. Ordinal data: data in order --> LabelEncoder is used in this case.

As Airline is Nominal Categorical data we will perform OneHotEncoding.

As Source is Nominal Categorical data we will perform OneHotEncoding.

As Destination is Nominal Categorical data we will perform OneHotEncoding.

As stops case of Ordinal Categorical type we perform LabelEncoder.

### Feature Selection

Finding out the best feature which will contribute and have good relation with target variable.

Following are some of the feature selection methods,

* heatmap
* feature\_importance\_
* SelectKBest

In this case we used heat map to find correlation between Independent and dependent attributes.

### Fitting model using Random Forest

Split dataset into train and test set in order to prediction w.r.t X\_test If needed do scaling of data

Scaling is not done in Random forest

Import model

Fit the data

Predict with respect to X\_test

In regression check RSME Score

Plot graph

R2\_score for this model: 0.7972

### Hyperparameter Tuning

Choose following method for hyperparameter tuning

* RandomizedSearchCV --> Fast
* GridSearchCV

Assign hyperparameters in form of dictionary

Fit the model

Check best parameters and best score.

The best hyper parameter performed in this case are as follows:

n\_estimators: 1000

min\_samples\_split: 10

min\_samples\_leaf: 2

min\_features : auto

max\_depth: 20

Improved score: 0.8130

**QA**

Q1) What is the source of data?

Ans. Here is source of dataset :<https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh>

Q 2) What was the type of data?

Ans. The data was the combination of numerical and Categorical values.

Q 3) What’s the complete flow you followed in this Project?

Ans. Please refer slide 3rd for better understanding.

Q 4) How logs are managed?

Ans. Loggers are implemented in each and every module and function. All logs are being captured and recorded in a separate log file.

Q 5) What techniques were you using for data pre-processing?

Ans. Following techniques are used :

* + Removing unwanted attributes
  + Visualizing relation of independent variables with each other and output variables
  + Checking and changing Distribution of continuous values
  + Removing outliers
  + Cleaning data and imputing if null values are present.
  + Converting categorical data into numeric values.
  + Scaling the data

Q 6) How Prediction was done?

Ans. User enters details on web portal and at backend using machine learning predictions are made. Result was displayed on same page.

Q 8) What are the different stages of deployment?

Ans. The web app will get deployed in:

* AWS
* Azure
* GCP
* Heroku