FLIGHT FARE PREDICTION USING MACHINE LEARNING

## Low Level Design (LLD)

## Anjum Hassan

## September- October, 2021

iNeuron

# INTRODUCTION

The main objective of this project is to predict the flight price given certain factors that are proven to be effective on Flight Price. In Airline Business, flight Price is an important factor that drives customer base. Customers generally tend to go for budget friendly. So, airline company may use different factors to estimate flight price and adjust the pricing accordingly in real time. The use case of this application can be expanded to Flight Price Forecast and Flight Recommendation Applications.

# PROBLEM STATEMENT

As per the problem statement, we can make use of different machine learning regression algorithms to do the predictions. The data used for the algorithm was well structured in tabular format and proper data dictionary was available for the data. So, we proceeded with all the data processing and modeling activities.

# DATASET INFORMATION

# The dataset used for this project was extracted from Kaggle.

# Link: https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh

# Data Format: Excel

# 

# Columns Information:

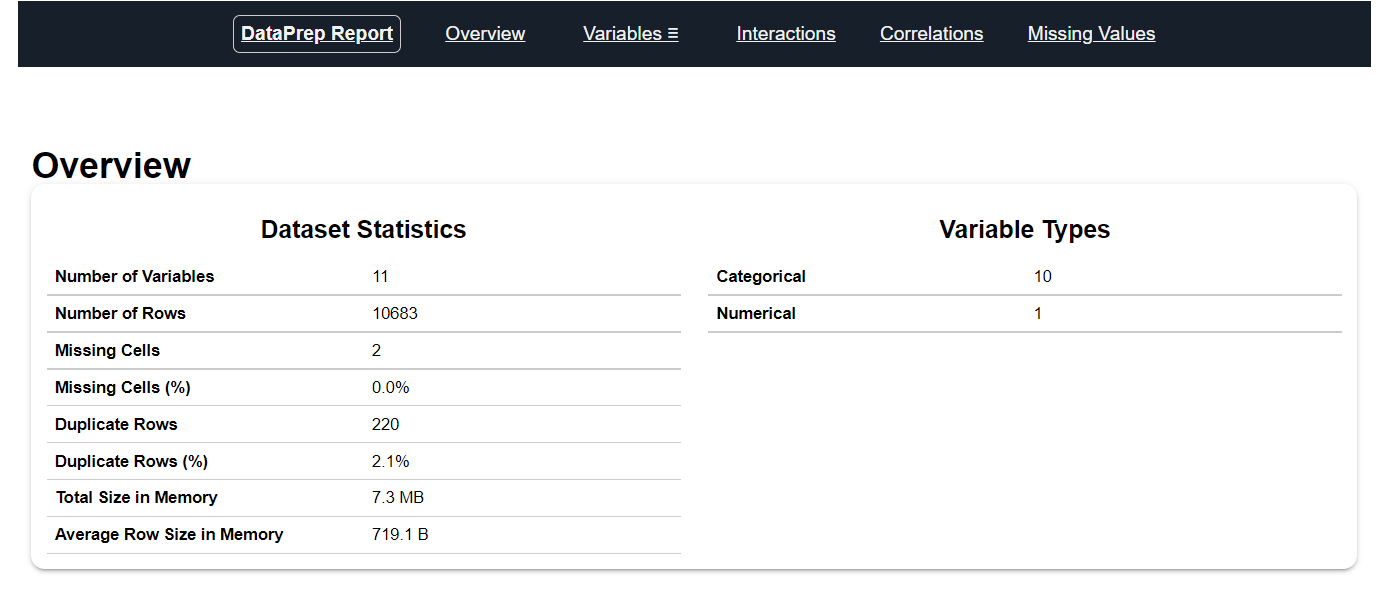
* Airline: Name of the airline used for traveling
* Date\_of\_Journey: Date at which a person traveled
* Source: Starting location of flight
* Destination: Ending location of flight
* Route: This contains information on starting and ending location of the journey in the standard format used by airlines.
* Dep\_Time: Departure time of flight from starting location
* Arrival\_Time: Arrival time of flight at destination
* Duration: Duration of flight in hours/minutes
* Total\_Stops: Number of total stops flight took before landing at the destination.
* Additional\_Info: Shown any additional information about a flight
* Price: Price of the flight (**Target)**

# Architecture Description.

# 

* 1. **Data Description:**

The dataset was taken from Kaggle (URL: https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh



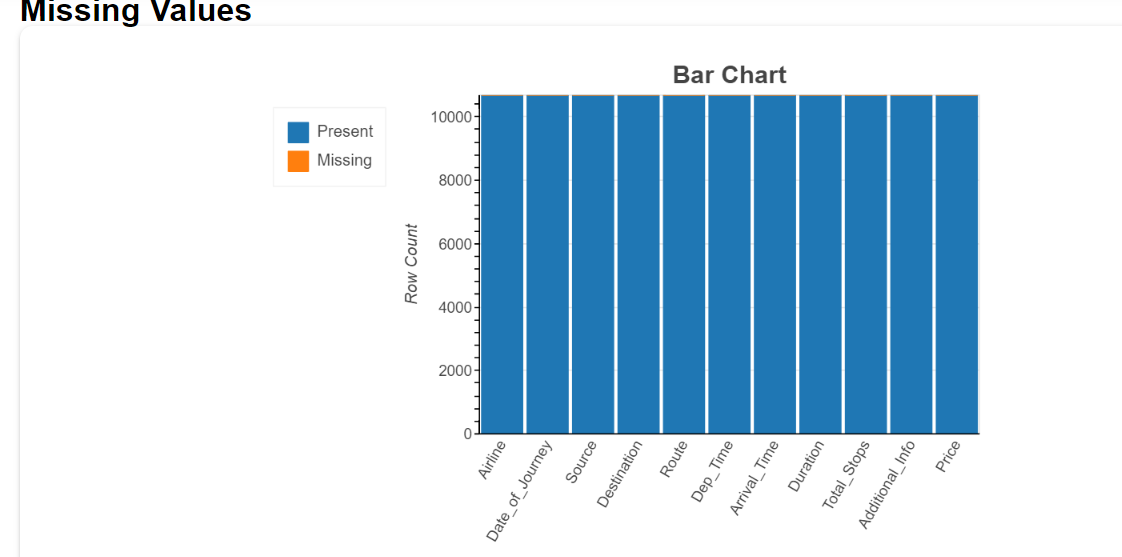
**4.2 Data Pre-processing.**

This included importing of important libraries such as seaborn, matplotlib, pandas etc. We imported the same dataset mentioned above from Kaggle.

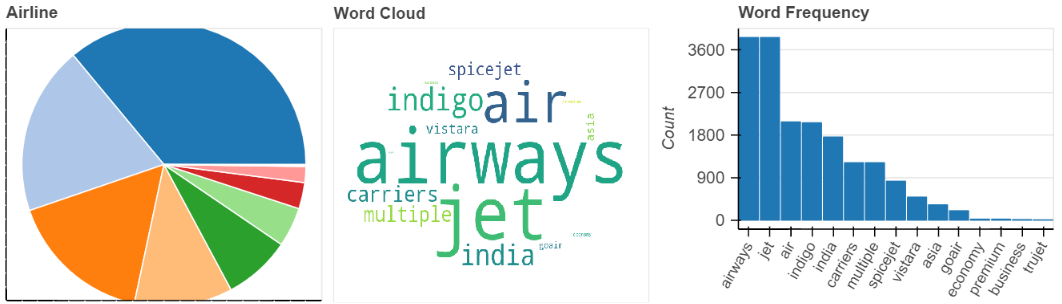
* 1. **Data Analysis**

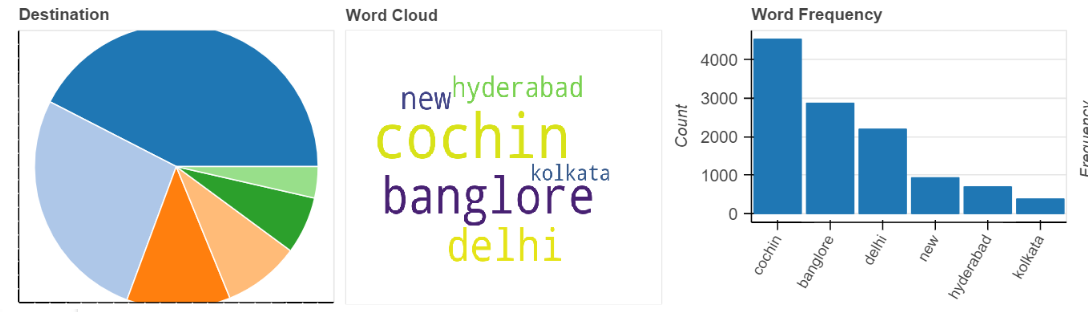
Here we handled the null values, changed the column names, plotted multiple graphs in seaborn, matplotlib and other visualization library for proper understanding of the data and the distribution of information in the same. As there were no null values in the data, we proceeded with the visualization and analysis.

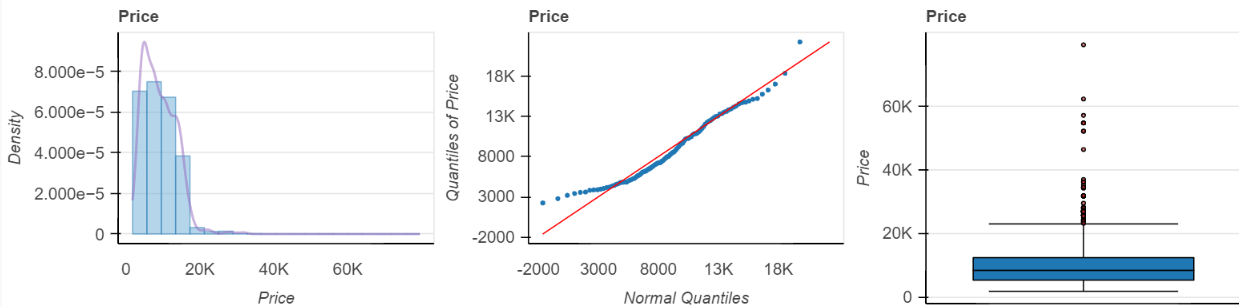
For each specific feature we analyzed the data using visualization, and jotted down the important key points which can impact the final predictions.

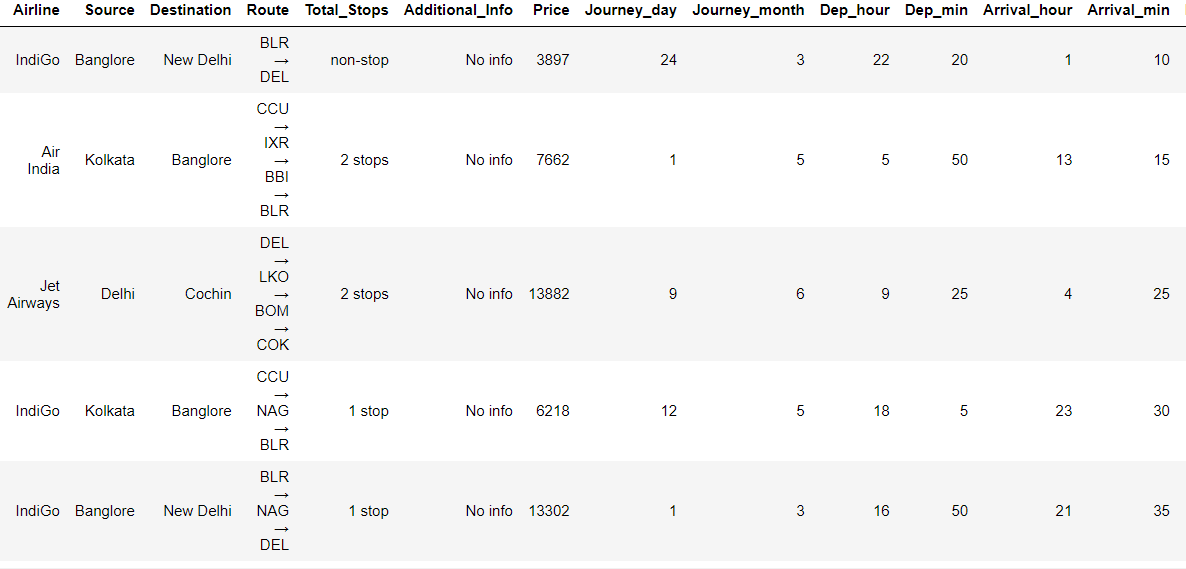


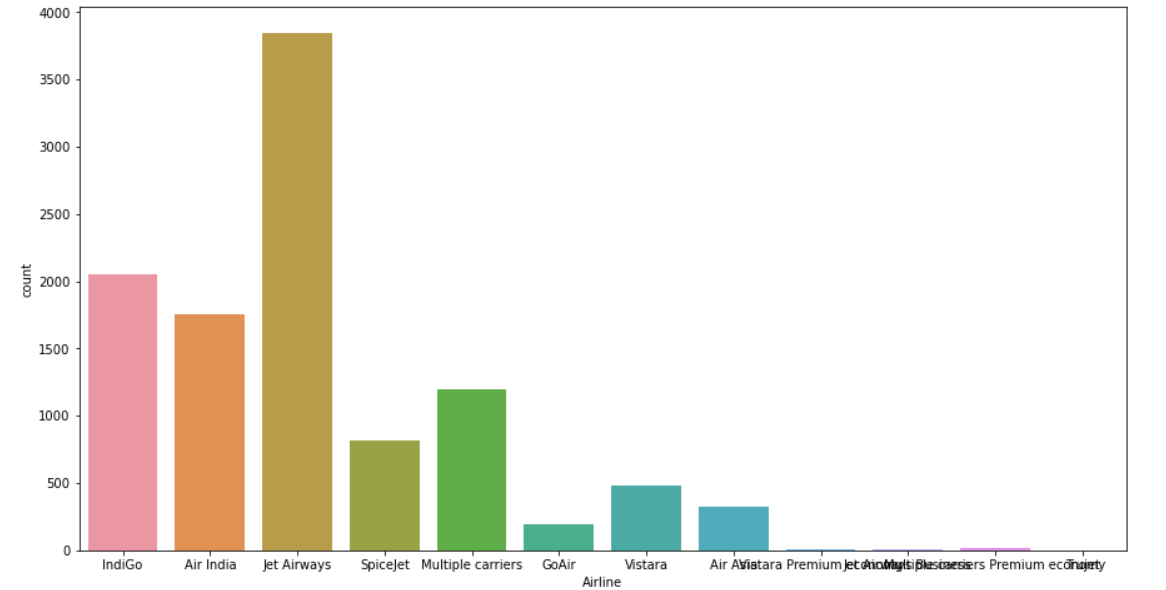
**EDA**

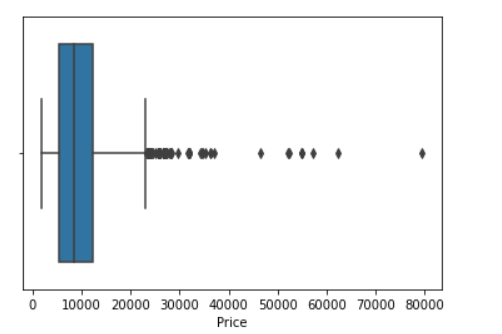
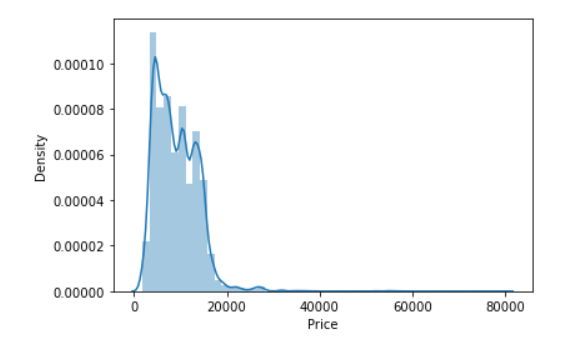






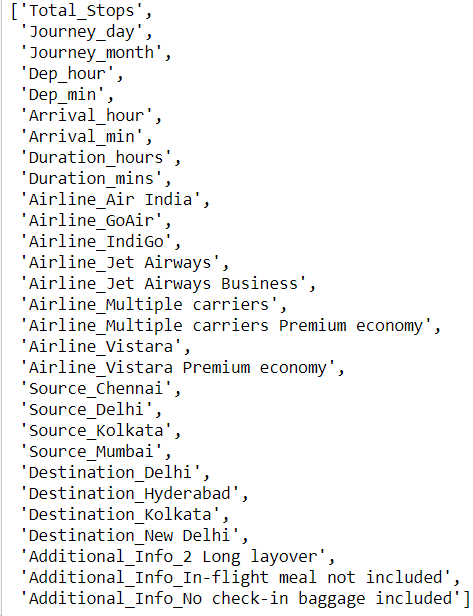






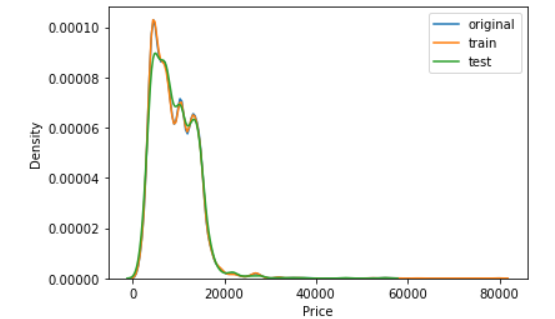
* 1. **Feature Engineering**

We made use of Sequential Feature Selection Technique for filtering out important features based on R2 score and 3-fold cross-validation.



* 1. **Train/Test Split**

This library was imported from Sklearn to divide the final dataset into the ratio of 80-20%, where 80% of the data was used to train the model and the latter 20% was used to predict the same.

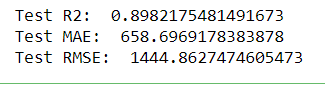


**4.6 Selecting Model**

We tried with RandomForestRegressor for the modeling and came up with this model with the best performance.

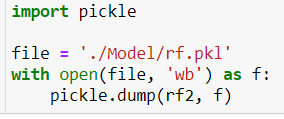
**4.7. Prediction**

The Accuracy of Random Forest was **89.0%**



* 1. **Save Model**

Model was saved using the pickle library which saves the file in a binary mode.



* 1. **Deploy in Local Host**

We created a Streamlit Web Ui and deployed the model through GCP.

# https://fpp-streamlit-app.du.r.appspot.com/

