High Level Document (HLD)

Flight Fare Prediction

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Contents**

Table of Contents

[Document Version Control 2](#_Toc80612930)

[Abstract 7](#_Toc80612931)

[1. **Introduction** 8](#_Toc80612934)

[1.1 Why this High-Level Design Document? 8](#_Toc80612935)

[1.2 Scope 8](#_Toc80612936)

[1.3 Definitions 8](#_Toc80612937)

[**2.** **General Description** 9](#_Toc80612938)

[2.1. Product Perspective 9](#_Toc80612939)

[2.2. Problem Statement 9](#_Toc80612942)

[2.3. Proposed Solution 9](#_Toc80612945)

[2.4. Further Improvements 10](#_Toc80612946)

[2.5. Technical Requirements 10](#_Toc80612948)

[2.6. Data Requirements 10](#_Toc80612952)

[2.7. Data Analysis 11](#_Toc80612953)

[2.8. Tools Used 13](#_Toc80612954)

[2.9. Constraints 13](#_Toc80612960)

[2.10. Assumptions 13](#_Toc80612961)

[**3.** **Design Details** 14](#_Toc80612962)

[3.1. Process Flow 14](#_Toc80612963)

[3.2. Event Log 15](#_Toc80612969)

[**4.** **Performance.** 16](#_Toc80612972)

[**5.** **Conclusion** 16](#_Toc80612976)

# Abstract

# According to a report, India’s civil aviation industry is on a high-growth trajectory. India aims become the third-largest aviation market by 2020 and the largest by 2030. Indian domestic air traffic is expected to cross 100 million passengers by FY2017, compared to 81 million passengers in 2015, as per Centre for Asia Pacific Aviation (CAPA).

# According to Google Trends, the search term - "Cheap Air Tickets" is most search in India. Moreover, as the middle-class of India is exposed to air travel, consumers hunting for cheap prices increases.

# Introduction

## Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all of the design aspects and define them in detail
* Describe the user interface being implemented
* Describe the hardware and software interfaces
* Describe the performance requirements
* Include design features and the architecture of the project
* List and describe the non-functional attributes like:
  + Security
  + Security
  + Reliability
  + Maintainability
  + Portability
  + Reusability
  + Application compatibility
  + Resource utilization
  + Serviceability

## **Scope** **The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.**

## **Definitions**

# General Description

# Product Perspective

# Any individual who has booked a flight ticket previously knows how dynamically costs change. Aircraft uses advanced strategies called Revenue Management to execute a distinctive valuing strategy. The least expensive accessible ticket changes over a period the cost of a ticket might be high or low.

# This valuing method naturally modifies the toll as per the time like morning, afternoon or night. Cost may likewise change with the seasons like winter, summer and celebration seasons. The extreme goal of the carrier is to build its income yet on the opposite side purchaser is searching at the least expensive cost. Purchasers generally endeavor to purchase the ticket in advance to the takeoff day. Since they trust that airfare will be most likely high when the date of buying a ticket is closer to the takeoff date, yet it is not generally true. Purchaser may finish up with the paying more than they ought to for a similar seat.

# Problem Statement

## Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics which they call "revenue management" or "yield management". The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on –

## Time of purchase patterns (making sure last-minute purchases are expensive)

1. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

Airline industry is one of the most sophisticated in its use of dynamic pricing strategies to maximize revenue, based on proprietary algorithms and hidden variables. Therefore, it is challenging for consumers to predict the price change in the future i . With the information of the airfare available online, buyers are trying to track the prices of the flight over a certain period of time, and anticipate the price change in the future. However, it turns out to be rather difficult to predict the price of the flight precisely only by observation.

# Proposed Solution

It is very difficult for the customer to purchase a flight ticket at the minimum price. For this several techniques are used to obtain the day at which the price of air ticket will be minimum. Most of these techniques are using sophisticated artificial intelligence (AI) research is known as Machine Learning

It is hard for the client to buy an air ticket at the most reduced cost. For this few procedures are explored to determine time and date to grab air tickets with minimum fare rate. The majority of these systems are utilizing the modern computerized system known as Machine Learning. To determine ideal purchase time for flight ticket Gini and Groves exploited Partial Least Square Regression (PLSR).

# Further Improvements

# Flight Fare Prediction can be added/updated with weather forecast that can lead to increase or decrease in wait time, the waiting period for the next flight to arrive, extra baggage charges depend on weight thus that could be improvised later in the project that could also lead to increase in the flight fare.

# Technical Requirements

# This document addresses the requirements for detecting the anomalies in booking flights at early stages and recommending the necessary and rapid action while booking the flights.

# External User Interface is required for user interaction with the system to book the flight and the User Interface need to be simple and not too complex.

# The User Interface should include date of departure date of arrival route and flight name the client wants to travel to and is it a single trip or a return trip for predicting the flight fare. A machine Learning Algorithm will never give 100% accurate result but the result will be nearly accurate.

# Data Requirements

Data requirement completely depend on our problem statement

The accumulation of information is the most significant part of this venture. The different wellsprings of the information on various sites are utilized to prepare the models. Sites provide data about the numerous courses, times, aircrafts and charge. Different sources from API’s to customer travel sites are accessible for information scratching. In this segment information of the different sources and parameters that are gathered are talked about and models are implemented using python.

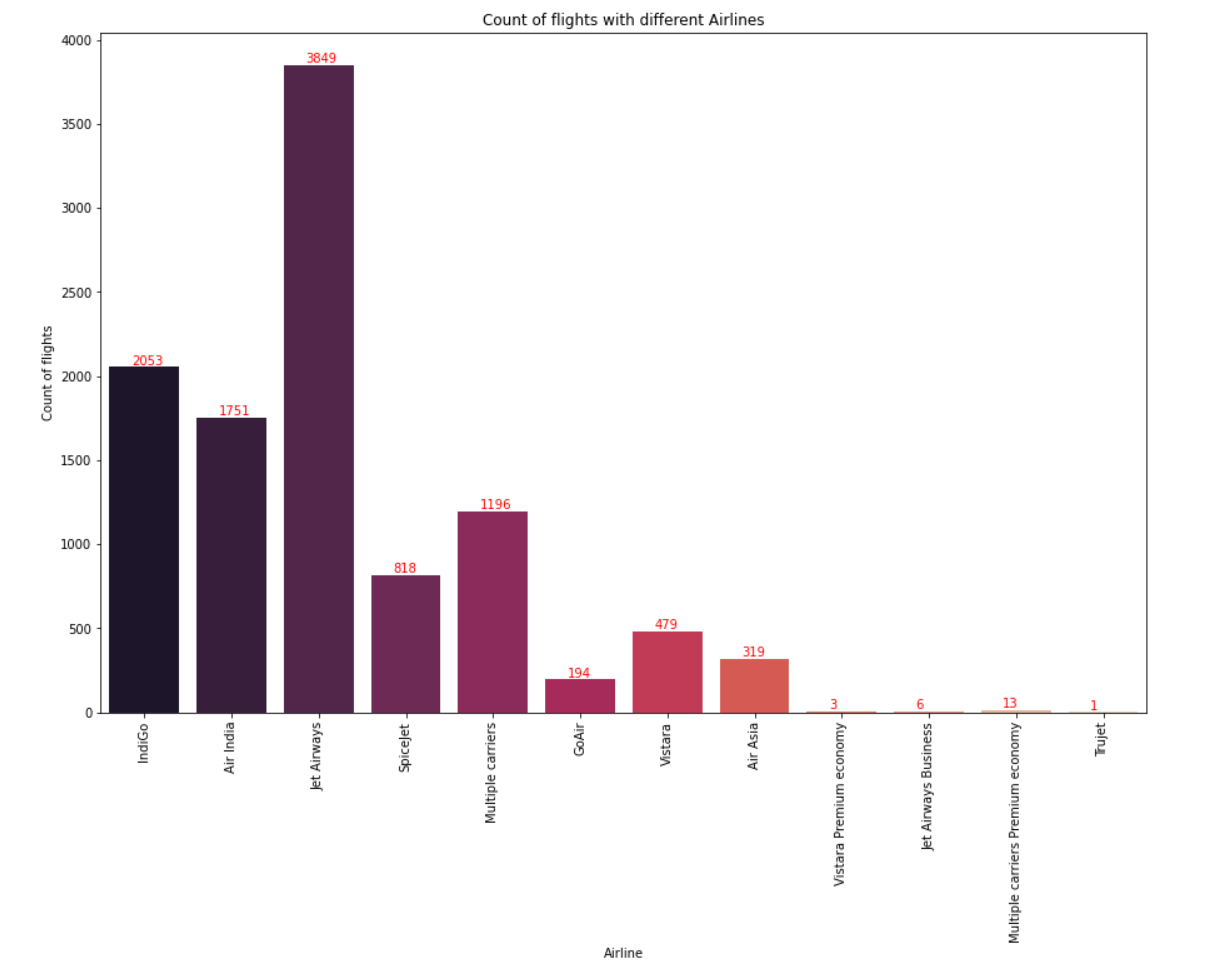
The python-script take out the data from the site, and provides output as a CSV record. The document contains the data with features and its details. A significant perspective is to choose the features required for calculation of expected flight price. Output gathered from the site contains number of parameters for each flight: yet not all are required, so just the accompanying components are,

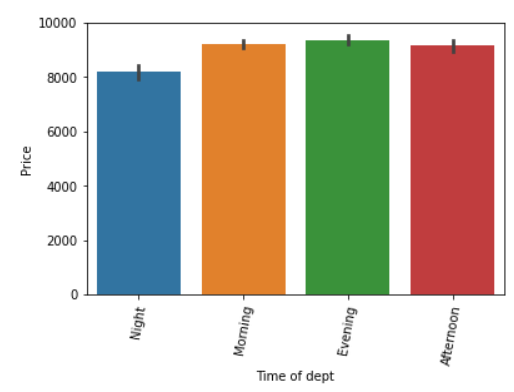
* Date of journey
* Time of Departure
* Place of Departure
* Time of Arrival
* Place of Destination/Arrival
* Airway Company
* Total Fare

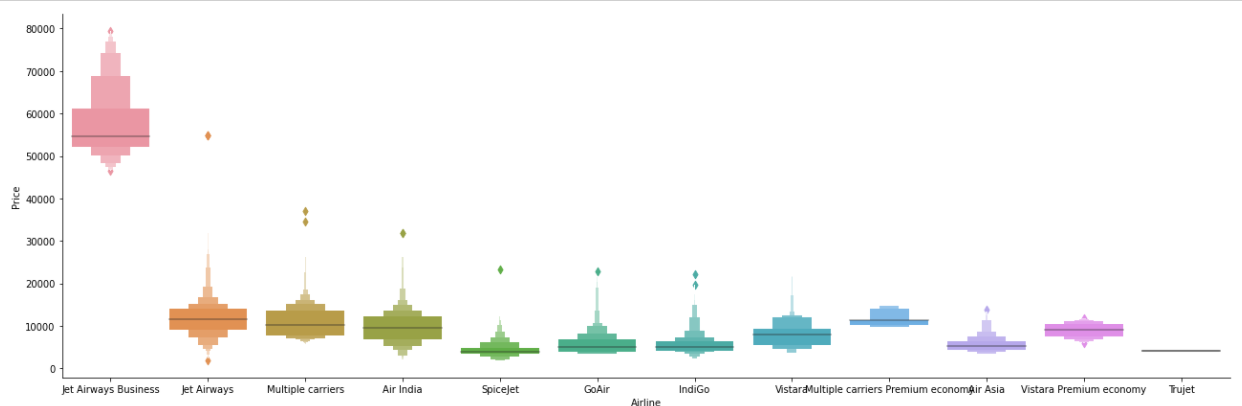
In this investigation, the attention is just to limit the airfare considering a single route. This information is gathered for perhaps the busiest course in India (BOM to DEL) over a time of a quarter of a year that is from February to April. For each flight information each feature is collected physically.

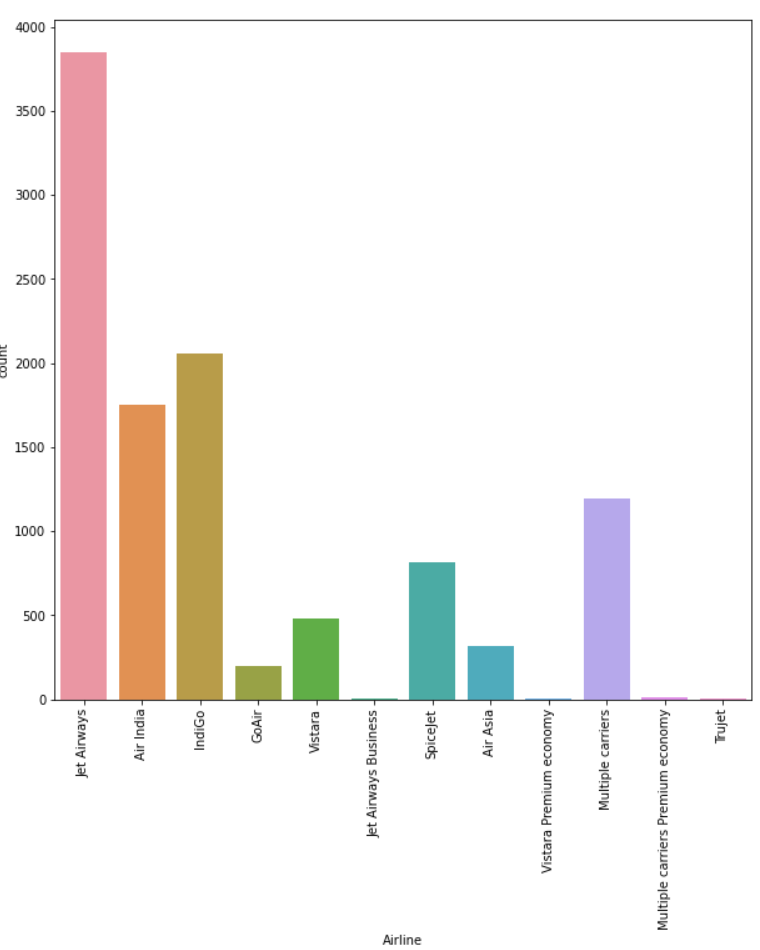
Preparation of data is trailed by breaking down the information, revealing the concealed patterns and afterward applying different AI models. Likewise, a few features can be determined from the current features. Flight days can be issued by computing the difference of the flight date and the date on which information is collected. This can be observed for 45 days. Additionally, flight date is important, whether it is on festive day or a weekday or weekend. Instinctively the flights planned during weekends cost more than the flights on weekdays. Additionally, time plays important role. So the time is considered in classes as: Morning, evening and night.

# Data Analysis

The Data used by us was highly complex and it consisted of 10,683 rows and 11 columns with 2 null values. We also found categorical datatypes i.e. Airline, Date\_of\_Journey, Source, Destination, Dep\_Time, Arrival\_Time, Total\_Stops, Additional\_Info etc. Remaining columns were of int datatypes. After performing operations like removing of null values we analyzed data that highest month of booking was recorded in the month of May therefore highest number of count flights was also in May. The top three airlines are Jet Airways, Indigo and Air India and the airline which made least number of flights is Trujet. Dataset also tells that time of Departure for majority of flights was in evening and very few booked a night flight. Thursday price of flights were at a peak touching about 17,500. On 9th Journey day 1406 flights had taken off. The airline with highest price is Jet Airways







# Tools Used

# 

# https://keestalkstech.com/wp-content/uploads/2020/05/1200px-Jupyter_logo.svg1_-300x300.png See the source image

# See the source image

# There are many Integrated Development Environment available (IDE) the we will be using for coding is :-

# Jupyter Notebook is used as IDE.

* + - For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
    - AWS is used for deployment of the model.
    - Front end development is done using HTML/CSS
    - Python Django is used for backend development.
    - GitHub is used as version control system.
    - LazyPredict to check which the best model to give accurate prediction.
    - Done Modelling using Random Forest and Extra Tree Regressor

# Constraints

The Flight Fare Prediction User Interface must be user friendly, as automated as possible and users should not be required to know any of the Lines of Code (LoC).

# Assumptions

The main objective of the project is to implement the use cases as previously mentioned (2.2 Problem Statement) for new dataset that comes through various features in the dataset related to flight. It is also assumed that all aspects of this project have the ability to work together in the way the designer is expecting.

# Design Details

# Process Flow

# To identify the methodology from importing a csv file to creating a model is shown below

# Proposed Methodology

# 

# Model Training and Evaluation

# 

# Event Log

# The system should log every event what is running internally.

**Initial Step-By-Step Description:**

1. The System identifies at what step logging required
2. The System should be able to log each and every system flow.
3. Developer can choose logging method. Optional you can choose database logging/ or simple UI with multiple textboxes for User Input.

# System should not hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

# Performance.

# Our project is aimed at building up models to predict the airline ticket price. The input of our models are the factors that may influence the price, such as the weekday of departure and the number of stops in the itinerary. We applied linear regression, Naïve Bayes, Softmax regression, and Support Vector Machine (SVM) to predict the corresponding price.

# Reuasability

# The Code Should be Reusable

# Conclusion

# This study shows that it is feasible to predict the airline ticket price based on historical data. One possible way to increase the accuracy can be combining different models after carefully studying their own performance on each individual bin. Additionally, as the learning curve indicates, adding more features will increase the accuracy of our models. However, limited by the current data source that we have, we are unable to extract more information of a particular flight.

# Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. This gives the predicted values of flight fare to get a flight ticket at minimum cost. Data is collected from the websites which sell the flight tickets so only limited information can be accessed. The values of R-squared obtained from the algorithm give the accuracy of the model. In the future, if more data could be accessed such as the current availability of seats, the predicted results will be more accurate.