Low Level Design (LLD)

Flight Fare prediction using ML

Revision Number: 1.8
Last date of revision: 07/09/2021

Staffin T Shabu

Document Version Control

Date Issued	Version	Description	Author
12th Aug 2021	1.1	First Draft	Staffin
15th Aug 2021	1.2	Added Workflow chart	Staffin
20th Aug 2021	1.3	Added Exception Scenarios Overall, Constraints	Staffin
30st Aug 2021	1.4	Added KPIs	Staffin
2nd Sept 2021	1.5	Added user I/O flowchart	Staffin
13 Sept 2021	1.6	Added model diagrams	Staffin
15th Sept 2021	1.7	Added dataset overview and updated user I/O flowchart.	Staffin
17th Sept 2021	1.8	Restructure and reformat LLD	Staffin

Contents

Abstract

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, duration of flights. various occasions such as vacations or festive season. Therefore, having some basic idea of the flight fares before planning the trip will surely help many people save money and time. In the proposed system a predictive model will be created by applying machine learning algorithms to the collected historical data of flights. This system will give people the idea about the trends that prices follow and also provide a predicted price value which they can refer to before booking their flight tickets to save money. This kind of system or service can be provided to the customers by flight booking companies which will help the customers to book their tickets accordingly.

1 Introduction

This project aims to develop an application which will predict the flight prices for various flights using machine learning model. The user will get the predicted values and with its reference the user can decide to book their tickets accordingly. In the current day scenario flight companies try to manipulate the flight ticket prices to maximize their profits. There are many people who travel regularly through flights and so they have an idea about the best time to book cheap tickets. But there are also many people who are inexperienced in booking tickets and end up falling in discount traps made by the companies where actually they end up spending more than they should have. The proposed system can help save millions of rupees of customers by proving them the information to book tickets at the right time. The proposed problem statement is "Flight Fare prediction system".

1.1 Scope

Currently, there are many fields where prediction-based services are used such as stock price predictor tools used by stockbrokers and service like Zestimate which gives the estimated value of house prices. Therefore, there is requirement for service like this in the aviation industry which can help the customers in booking tickets. There are many research works that have been done on this using various techniques and more research is needed to improve the accuracy of the prediction by using different algorithms. More accurate data with better features can also be used to get more accurate results.

1.2 Constraints

The designed application should have a user-friendly interface and should be hassle free experience for the users. The accuracy of the prediction should be highly reliable..

1.3 Risks

The prediction is based on past data. The future events might have an impact on the ticket fare. Hence the model should be optimized in a timely manner. Users should be aware of this limitation.

1.4 Out of Scope

Option for booking of tickets is not provided. The price updates are not alerted to the users

2 Technical specifications

2.1 Dataset

Disease	Finalized	Source
Past flight data	yes	https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh

2.1.1 Flight Fare dataset overview

Consists of 10000+ rows and 11 columns. It consists of various parameters like Airline, Date of Journey, Source Destination, Route Dep Time, Arrival Time, Duration, Total Stops, Additional Info, Price etc.

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

2.1.2 Input schema

Feature Name	Data Type	Size	Null/required
Departure Date	Date	3	required
Arrival date	Date	3	required
Source	String	22	required
Destination	String	22	required
Stoppage	String	22	required
Airline	String	22	required

2.2 Predicting Fare

- The system displays the option to enter the inputs.
- Based on the inputs the model will predict the fare

2.3 Logging

We should be able to log every activity done by the user.

- The System identifies at what step logging required
- The System should be able to log each and every system flow.
- Developers can choose logging methods. You can choose database logging/ File logging as well.
- System should not be hung even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

2.4 Deployment

1. Flask

   The deployment is done using flask

3 Technology stack

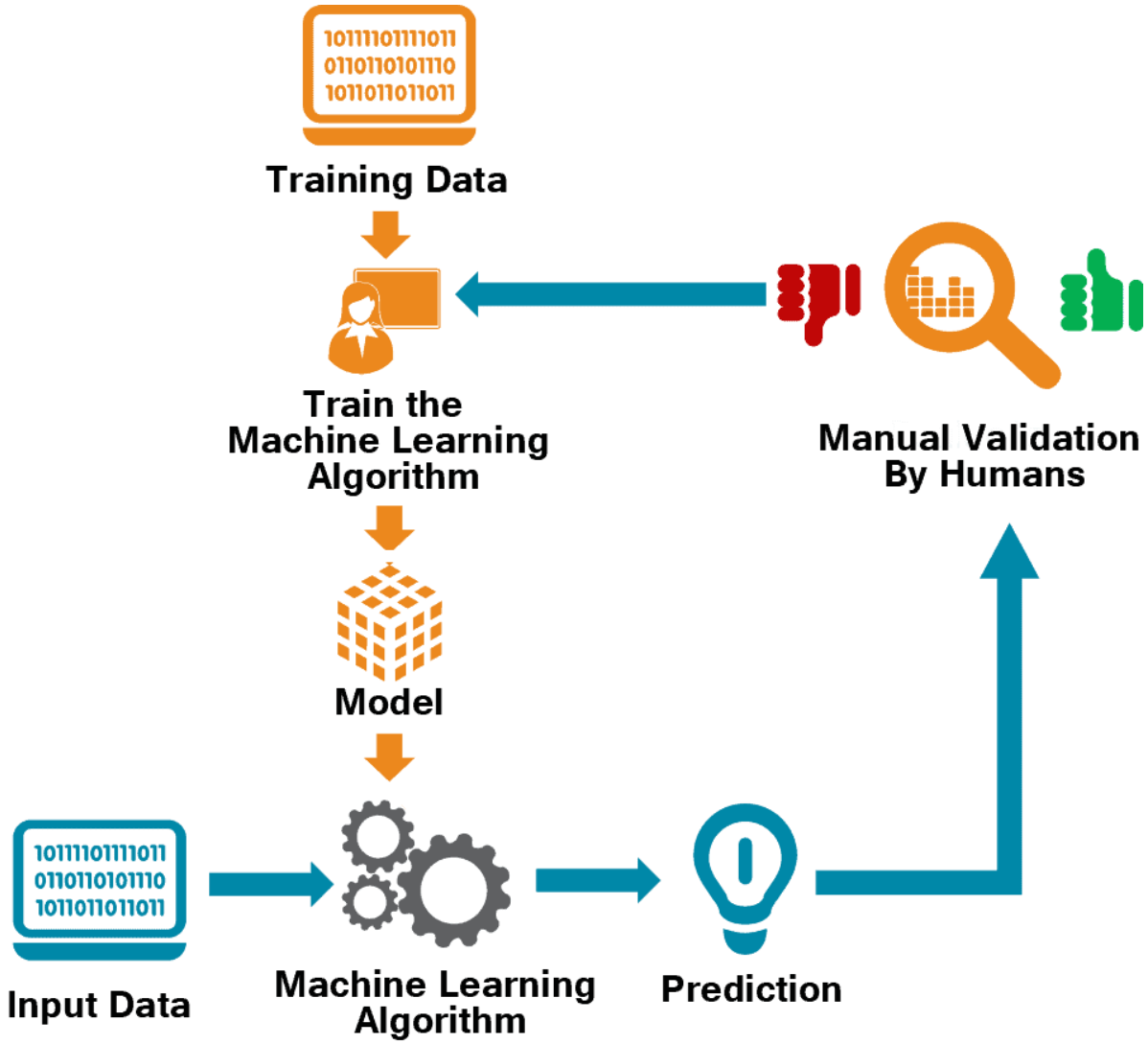
Front End	HTML/CSS/JS/React
Backend	Python
Deployment	Flask

4 Proposed Solution

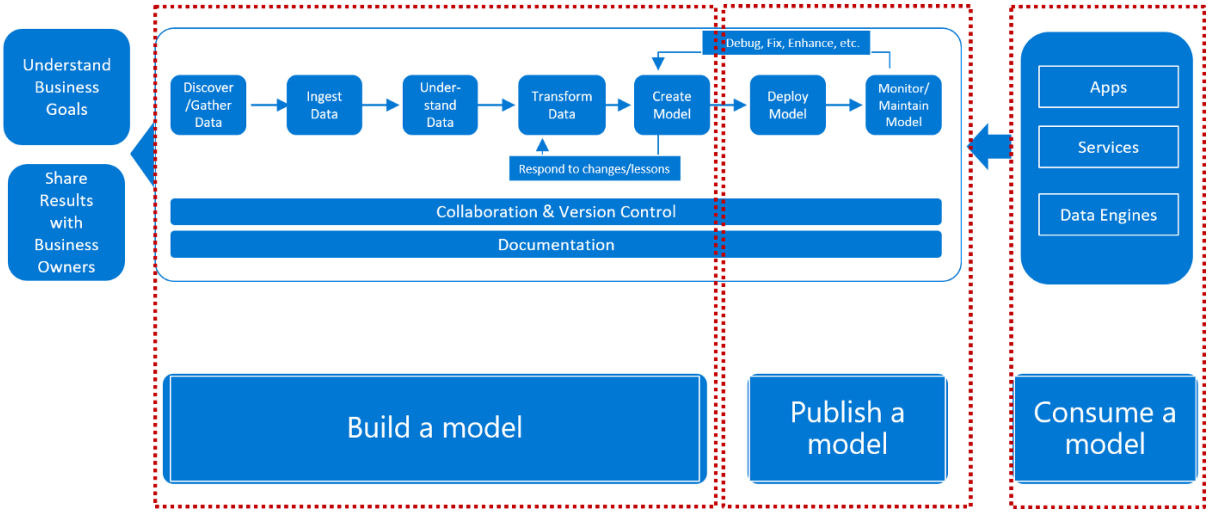
For this project, we have implemented the machine learning life cycle to create a basic web application which will predict the flight prices by applying machine learning algorithm to historical flight data using python libraries like Pandas, NumPy, Matplotlib, seaborn and sklearn.

Data selection is the first step where historical data of flight is gathered for the model to predict prices. Our dataset consists of more than 10,000 records of data related to flights and its prices. Some of the features of the dataset are source, destination, departure date, departure time, number of stops, arrival time, prices, and few more. In the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. If these values are not removed it would affect the accuracy of the model. We gained further information such as distribution of data. Next step is data pre-processing where we observed that most of the data was present in string format. Data from each feature is extracted such as day and month is extracted from date of journey in integer format, hours and minutes is extracted from departure time. Features such as source and destination needed to be converted into values as they were of categorical type. For this One hot-encoding and label encoding techniques are used to convert categorical values to model identifiable values. Feature selection step is involved in selecting important features that are more correlated to the price. There are some features such as extra information and route which are unnecessary features which may affect the accuracy of the model and therefore, they need to be removed before getting our model ready for prediction. After selecting the features which are more correlated to price the next step involves applying machine algorithm and creating a model. As our dataset consist of labelled data, we will be using supervised machine learning algorithms. We will be using random forest Algorithm.

5 Model training/validation workflow



6 User I/O workflow



7 Exceptional scenarios

Step	Exception	Mitigation	Module
16 Sept 2021	1.1	First Draft	Amit K Gupta
17 Sept 2020	1.2	Added Workflow chart	Amit K Gupta

8 Test cases

Test case	Steps to perform test case	Module	Pass/Fail
1 **2** **3**	Input data Prediction Final output	Data input Random forest model Prediction	Pass Pass Pass

9 Key performance indicators (KPI)

- Time and workload reduction using the RF model.
- Comparison of accuracy of model prediction and previous flight fare test data.
- Duration of flight.
- Variation from test data.

.