A Project/Dissertation Review-1 Report

on

Flight Fare Prediction App Using Machine Learning Model

Submitted in partial fulfillment of the

requirement for the award ofthe degree

of

B.Tech. Computer Science Engineering

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Name of Supervisor : Mr. Arjun KP
Designation: Assistant Professor

Submitted By

Tushar Rawat	Rohan Singh
18021011415/18SCSE1180076	18021180041/18SCSE1180042

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
December,2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"Flight Fare Prediction App Using Machine LearningModel"** in partial fulfillment of the requirements for the award of the B.Tech. Computer Science Engineering submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of Name… Designation, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Tushar Rawat, 18SCSE1180076
Rohan Singh, 18SCSE1180076

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor Name
Designation

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Tushar Rawat, 18SCSE1180076 Rohan Singh, 18SCSE1180076 has been held on _________________ and is/her work is recommended for the award of B.Tech. Computer Science Engineering.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021
Place: Greater Noida

Acknowledgement

It gives us an incredible feeling of delight to introduce the report of the B. Tech Major Project embraced during B. Tech. Fourth Year. This task in itself is an affirmation to the motivation, drive and specialized help added to it by numerous people and our guide Mr., Mr.Arjun KP, Galgotias University.

Abstract

These days airplane ticket costs can vary dynamically and significantly for a similar flight, in any event, for adjacent seats inside a similar lodge. Clients are looking to get the most reduced cost while aircrafts are attempting to keep their general income as high as could be expected and amplify their benefit. Aircrafts utilize different sorts of computational strategies to expand their income, for example, request expectation and value separation. From the client side, two sorts of models are proposed by various specialists to set aside cash for clients: models that foresee the ideal opportunity to purchase a ticket and models that anticipate the base ticket cost. Our audit examination shows that models on the two sides depend on restricted arrangement of elements, for example, authentic ticket value information, ticket buy date and takeoff date. This project expects to foster an application which will foresee the flight costs for different flights utilizing AI model. The client will get the anticipated qualities and with its reference the client can choose to book their tickets as needs be. In the current day situation flight organizations attempt to control the flight ticket costs to boost their benefits. There are many individuals who travel routinely through flights thus they have a thought regarding the best an ideal opportunity to book modest tickets. In any case, there are likewise many individuals who are unpracticed in booking tickets and wind up falling in rebate traps made by the organizations where really they wind up spending more than they ought to have. The proposed framework can help clients in saving some cash by giving them the data to book tickets at the ideal opportunity. For this project, we have implemented the AI life cycle to make an essential mobile application which will foresee the flight costs by applying Al calculation to verifiable flight information utilizing python libraries like Pandas, NumPy, Matplotlib, seaborn and sklearn.

Contents

List of Table

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
API	Application Programming Interface
APP	Application
HTTP	Hyper Text Transfer Protocol
UI	User Interface
AI	Artificial Intelligence
ML	Machine Learning

CHAPTER-1 Introduction

This project aims to develop an application which will anticipate the flight costs for different flights utilizing machine learning model. The client will get the anticipated qualities and with its reference the client can choose to book their tickets likewise.

In the current day situation flight organizations attempt to control the flight ticket costs to expand their benefits. There are many individuals who travel consistently through flights thus they have a thought regarding the best an ideal opportunity to book modest tickets. Be that as it may, there are likewise many individuals who are inexperienced in booking tickets and wind up falling in markdown traps made by the organizations where really they wind up spending more than they ought to have. The proposed framework can assist with saving thousands of rupees of clients by demonstrating them the data to book tickets at the ideal opportunity.

We have trained a random forest classifier model for predicting the price of the flight based on various factors which affect the price of the flight.

Our initial investigation shows that models on the two sides depend on restricted arrangement of highlights, for example, verifiable ticket value information, ticket buy date and flight date. Parameters on which fares are calculated-

- Airline
- Date of Journey
- Date of Arrival
- Source
- Destination
- Departure Time
- Arrival Time
- Duration
- Total Stops
- Weekday/Weekend

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

Supervised learning is type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Supervised machine learning requires the data scientist to train the algorithm with both labeled inputs and desired outputs. Supervised learning algorithms are good for the following tasks:

1. Binary classification: Dividing data into two categories.
2. Multi-class classification: Choosing between more than two types of answers.
3. Regression modeling: Predicting continuous values.
4. Ensembling: Combining the predictions of multiple machine learning models to produce an accurate prediction.

1.2 Problem Formulation

Flight ticket costs can be a hard thing to figure, today we may see a value, look at the cost of a similar flight tomorrow it will be an alternate story. We may have regularly heard travelers saying that flight ticket costs are so unpredictable. Flying has turned into the essential transportation strategy for significant distance travel. To expand the benefit, aircrafts utilize a convoluted value system called "yield the board" to form the cost of each trip With this strategy, the cost can be consequently changed by many variables, like the quantity of days before departure , seat accessibility, market contest, and so forth The last objective is to get the augmented benefit from each flight. Since travelers will in general accept that flight cost goes up when the buy date is near the departure date, they regularly buy flight tickets out on the town that is as a long way from the departure date as could really be expected. However, this type of purchase behavior is not always correct. When it fails, travelers will spend more money event flight tickets are purchased in advance.

As a matter of fact, it is undeniably challenging for travelers to foresee when the best an ideal opportunity to buy battle tickets is because of the accompanying reasons:

- **Incomplete Information:** Travelers can just access part of the carrier's inner data. Truth be told, they don't have the admittance to the key information, like the number of the excess tickets and the understanding between various carrier organizations.

- **Fragmented Information:** The data that voyagers can get is divided. For instance, it is undeniably challenging for a normal explorer to find the relationship between flight cost and flight characters, like the quantity of visits, the takeoff time, and so on

- **Irregular Change:** Although explorers can gather verifiable flight value, the cost change isn't smooth. In reality, it is by all accounts profoundly unpredictable. Thus,
Travelers can only with significant effort anticipate future flight value as per the chronicled values

1.2.1 Tools and technologies used

1. **Random Forest -** Random forest basically uses group of decision trees as group of models. Random amount of data is passed to decision trees and each decision tree predicts values according to the dataset given to it. From the predictions made by the decision trees the average value of the predicted values if considered as the output of the random forest model. Random forest basically uses group of decision trees as group of models. Random amount of data is passed to decision trees and each decision tree predicts values according to the dataset given to it.
   It is a machine learning algorithm which we are using to build the project it will take various flight detail from user as input.
   A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

   A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

   The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

   A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like scikit-learn).
   Features of a Random Forest Algorithm
   It's more accurate than the decision tree algorithm.
   It provides an effective way of handling missing data.
   It can produce a reasonable prediction without hyper-parameter tuning.
   It solves the issue of overfitting in decision trees.
   In every random forest tree, a subset of features is selected randomly at the node's splitting point.
   How random forest algorithm works
   Understanding decision trees
   Decision trees are the building blocks of a random forest algorithm. A decision tree is a decision support technique that forms a tree-like structure. An overview of decision trees will help us understand how random forest algorithms work.

   A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further.

The nodes in the decision tree represent attributes that are used for predicting the outcome. Decision nodes provide a link to the leaves. The following diagram shows the three types of nodes in a decision tree.

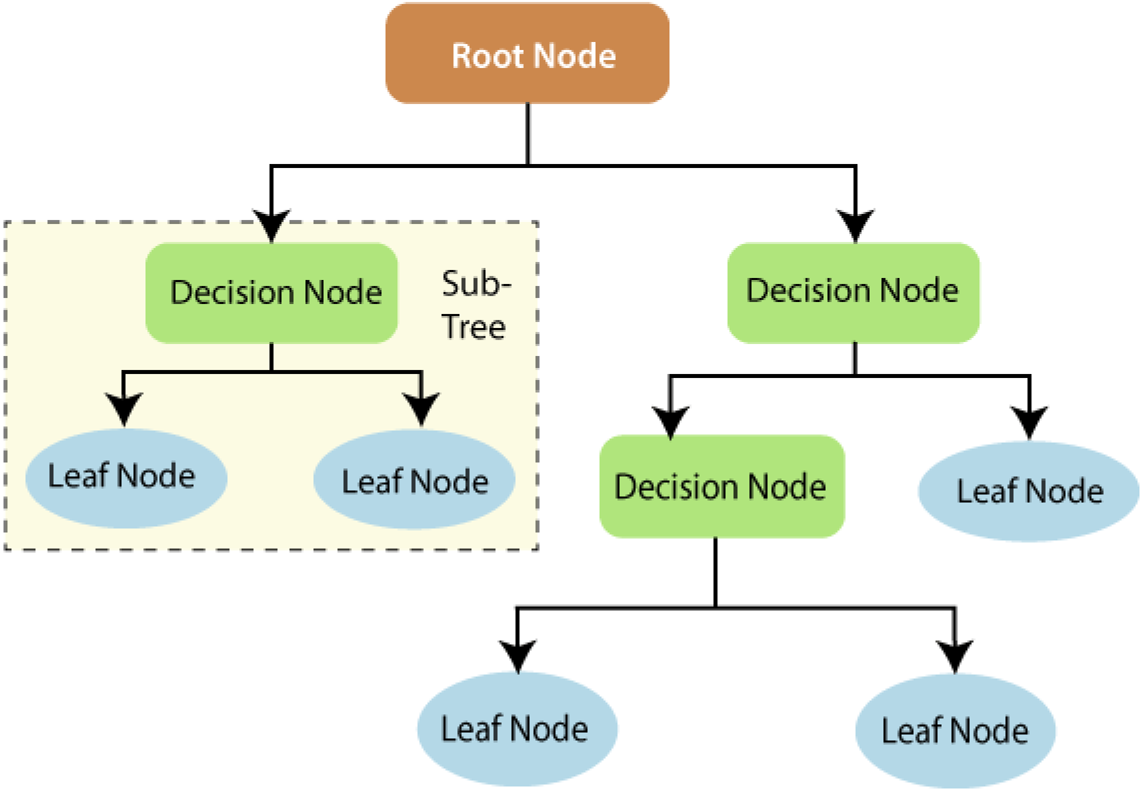


Fig 1. Decision Tree in Random Forest

The information theory can provide more information on how decision trees work. Entropy and information gain are the building blocks of decision trees. An overview of these fundamental concepts will improve our understanding of how decision trees are built.

Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables.

The information gain concept involves using independent variables (features) to gain information about a target variable (class). The entropy of the target variable (Y) and the conditional entropy of Y (given X) are used to estimate the information gain. In this case, the conditional entropy is subtracted from the entropy of Y.

Information gain is used in the training of decision trees. It helps in reducing uncertainty in these trees. A high information gain means that a high degree of uncertainty (information entropy) has been removed. Entropy and information gain are important in splitting branches, which is an important activity in the construction of decision trees.

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes.

A rain forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by the majority of the decision trees becomes the final output of the rain forest system. The diagram below shows a simple random forest classifier.

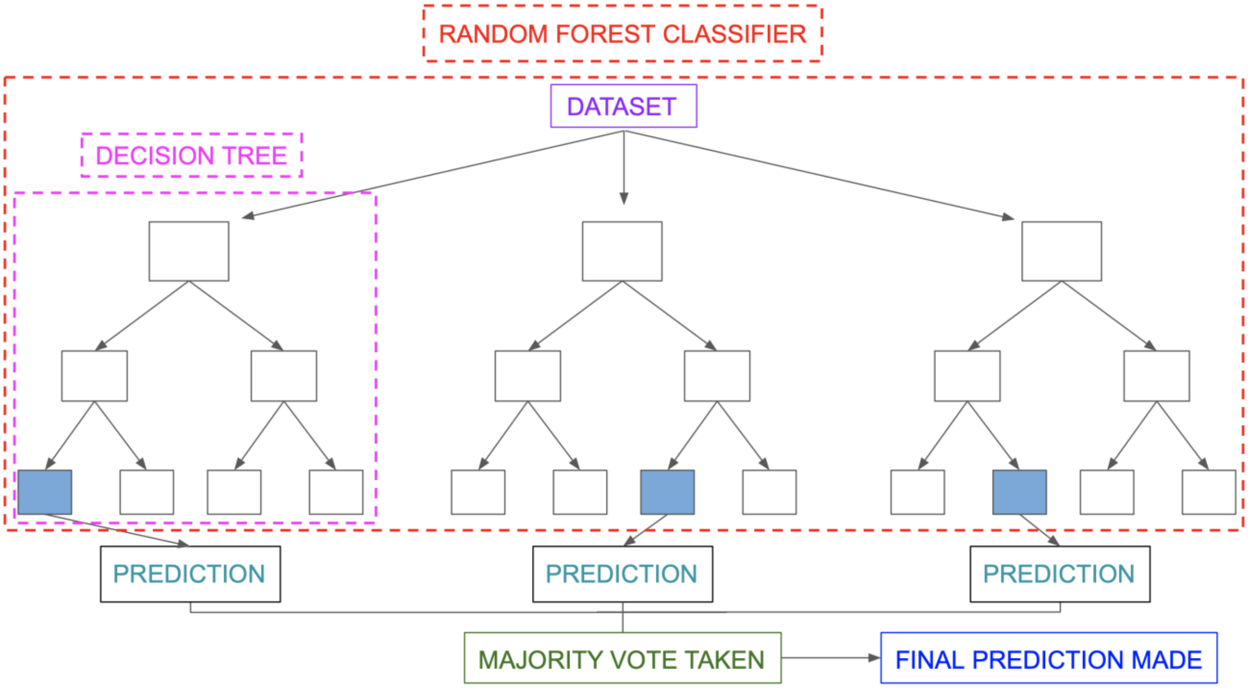


Fig 2. Random forest Classifier

Advantages of random forest are:-

1. It can perform both regression and classification tasks.
2. A random forest produces good predictions that can be understood easily.
3. It can handle large datasets efficiently.
4. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

2. **Flutter** - Flutter is Google's UI toolkit for building beautiful, natively compiled applications for mobile, web, desktop, and embedded devices from a single codebase. It is used to develop the frontend for the user to interact with the machine learning model and received the predicted price on their mobile phone.
   Flutter also offers many ready to use widgets (UI) to create a modern application. These widgets are optimized for mobile environment and designing the application using widgets is as simple as designing HTML.

   To be specific, Flutter application is itself a widget. Flutter widgets also supports animations and gestures. The application logic is based on reactive programming. Widget may optionally have a state. By changing the state of the widget, Flutter will automatically (reactive programming) compare the widget's state (old and new) and render the widget with only the necessary changes instead of re-rendering the whole widget.

   We shall discuss the complete architecture in the coming chapters.

   Features of Flutter

   1. Flutter framework offers the following features to developers −

   2. Modern and reactive framework.

   3. Uses Dart programming language and it is very easy to learn.

   4. Fast development.

   5. Beautiful and fluid user interfaces.

   6. Huge widget catalog.

   7. Runs same UI for multiple platforms.

   8. High performance application.

   Advantages of Flutter

   1. Flutter comes with beautiful and customizable widgets for high performance and outstanding mobile application. It fulfills all the custom needs and requirements. Besides these, Flutter offers many more advantages as mentioned below −

   2. Dart has a large repository of software packages which lets you to extend the capabilities of your application.

   3. Developers need to write just a single code base for both applications (both Android and iOS platforms). Flutter may to be extended to other platform as well in the future.

4. Flutter needs lesser testing. Because of its single code base, it is sufficient if we write automated tests once for both the platforms.

5. Flutter's simplicity makes it a good candidate for fast development. Its customization capability and extendibility makes it even more powerful.

6. With Flutter, developers has full control over the widgets and its layout.

7. Flutter offers great developer tools, with amazing hot reload.

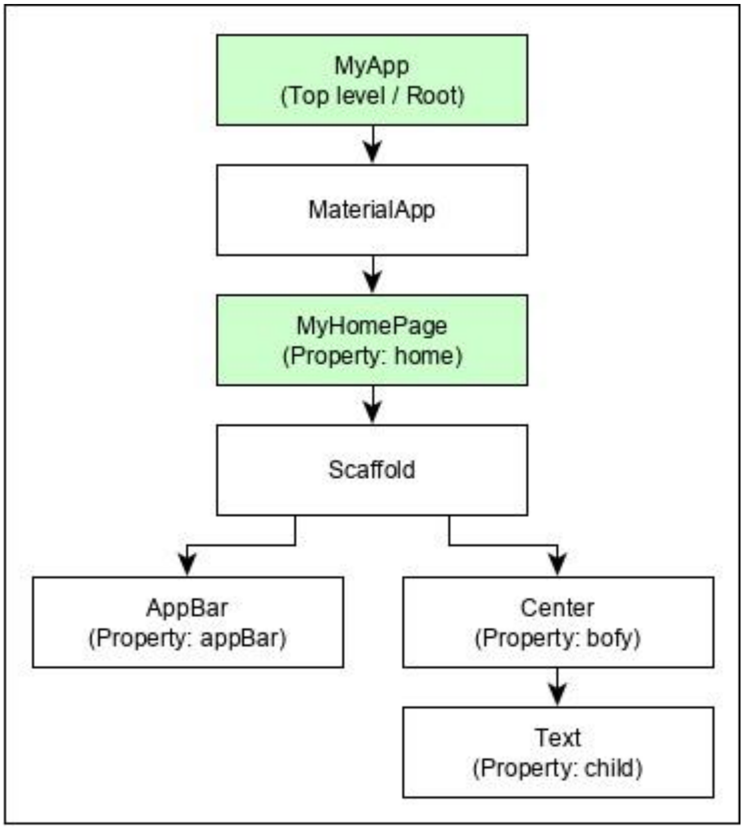


Fig3. Flutter Screen Architecture

3. **Flask** - Flask is basically a micro web application framework written in Python. Developers often use Flask for making web applications, HTTP request management, and template rendering. By “micro web application,” we mean that it is not a full-stack framework. It is used for making API which will connect our app to the machine learning model using POST request.

   Flask is a web application framework written in Python. It is developed by Armin Ronacher, who leads an international group of Python enthusiasts named Pocco. Flask is based on the Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

   WSGI

   Web Server Gateway Interface (WSGI) has been adopted as a standard for Python web application development. WSGI is a specification for a universal interface between the web server and the web applications.

   Werkzeug

   It is a WSGI toolkit, which implements requests, response objects, and other utility functions. This enables building a web framework on top of it. The Flask framework uses Werkzeug as one of its bases.

   Jinja2

   Jinja2 is a popular templating engine for Python. A web templating system combines a template with a certain data source to render dynamic web pages.

   Flask class has a redirect() function. When called, it returns a response object and redirects the user to another target location with specified status code.

   A cookie is stored on a client’s computer in the form of a text file. Its purpose is to remember and track data pertaining to a client’s usage for better visitor experience and site statistics.

   A Request object contains a cookie’s attribute. It is a dictionary object of all the cookie variables and their corresponding values, a client has transmitted. In addition to it, a cookie also stores its expiry time, path and domain name of the site.

   In Flask, cookies are set on response object. Use make_response() function to get response object from return value of a view function. After that, use the set_cookie() function of response object to store a cookie.

   Reading back a cookie is easy. The get() method of request.cookies attribute is used to read a cookie.

4. **Python-** Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.
   It is used for:
   1. web development (server-side),
   2. software development,
   3. mathematics,
   4. system scripting.

   What Python do:-
   1. Python can be used on a server to create web applications.
   2. Python can be used alongside software to create workflows.
   3. Python can connect to database systems. It can also read and modify files.
   4. Python can be used to handle big data and perform complex mathematics.
   5. Python can be used for rapid prototyping, or for production-ready software development.

   Need for Python:-
   1. Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
   2. Python has a simple syntax similar to the English language.
   3. Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
   4. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
   5. Python can be treated in a procedural way, an object-oriented way or a functional way.

5. **Heroku** - Heroku is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud. It is used to host our flask API in cloud server so that a physical device is able to call our flask API.
   Heroku has one of the best cloud hosting solutions in the industry, especially in the PaaS Niche and they have a FREE plan which is more than enough for a bot or any fun projects.

   Heroku Postgres

   Heroku Postgres is the Cloud database (DBaaS) service for Heroku based on PostgreSQL. Heroku Postgres provides features like continuous protection, rollback, and high availability; also forks, followers, and dataclips.

   Heroku Redis

   Heroku Redis is the customized Redis from Heroku to provide a better developer experience. It is fully managed and is provided as a service by Heroku. It helps in managing instances with a CLI, associate data with Postgres to gain business insights using SQL tools, and lets customer gain performance visibility.

   Heroku Teams

   Heroku Teams is a team management tool which provides collaboration and controls to bring a customer's developers, processes, and tools together in order to build better software. With Heroku Teams, teams can self-organize, add, and manage members, get fine-grained control with app-level permissions and also use collaboration tools like Heroku Pipelines. It also provides delegated administration and centralized billing.

   Heroku Enterprise

   Heroku Enterprise provides services to large companies which help them to improve collaboration among different teams. It provides a set of features like fine-grained access controls, identity federation, and private spaces to manage their enterprise application development process, resources, and users.

   Heroku Connect

   Heroku Connect lets users create Heroku apps that can easily integrate with Salesforce deployments at scale. This is done by having a seamless data synchronization between Heroku Postgres databases and Salesforce organizations.

   Heroku Elements

   Heroku Elements provides users with Add-ons (tools and services for developing, extending, and operating the app), Buildpacks (which automate the build processes for the preferred languages and frameworks) and Buttons (a tool for the one-click provisioning, configuring, and deployment of third party components, libraries and patterns).

6. Android Studio:-

Android Studio is a part of the “Integrated Development Environment” (IDE) technology stack. Its makers describe android studio as an “Android development environment centered on IntelliJ IDEA.” Android Studio, formerly known as ADT (Android Development Tools), adds additional capabilities and improvements to the eclipse.

Android Studio features include a ready-to-use Gradle-based framework that is both versatile and easy to use. It’s created utilizing a variety of different APK versions throughout the course of several generations. An extended template for Google services and other sorts of gadgets is also included in the package. Android Studio is a single development environment that allows you to create apps for Android phones, tablets, Android Wear, Android TV, and Android Auto.

Android Studio provides many excellent features that enhance productivity when building Android apps, such as a blended environment where one can develop for all Android devices, apply Changes to push code and resource changes to the running app without restarting the app, a flexible Gradle-based build system, a fast and feature-rich emulator, GitHub and Code template integration to assist you to develop common app features and import sample code, extensive testing tools and frameworks, C++ and NDK support, and many more. So we have prepared a complete Android Studio tutorial that will help the Android Developer to get more familiar with Android Studio.

Features of Android Studio

- It has a flexible Gradle-based build system.
- It has a fast and feature-rich emulator for app testing.
- Android Studio has a consolidated environment where we can develop for all Android devices.
- Apply changes to the resource code of our running app without restarting the app.
- Android Studio provides extensive testing tools and frameworks.
- It supports C++ and NDK.
- It provides build-in supports for Google Cloud Platform. It makes it easy to integrate Google Cloud Messaging and App Engine.

CHAPTER-2
Literature Survey

Proposed study Airfare price prediction using machine learning techniques, for the examination work a dataset comprising of 1814 information trips of the Aegean Airlines was gathered and used to prepare AI model. Diverse number of provisions were utilized to prepare model different to exhibit how determination of components can change exactness of model.

In case study by William groves a specialist is acquainted which is capable with advance buy timing for clients. Halfway least square relapse procedure is utilized to assemble a model.

In a survey paper by supriya rajankar an overview on flight charge forecast utilizing AI calculation utilizes little dataset comprising of trips among Delhi and Bombay. Calculations, for example, K-closest neighbors (KNN), direct relapse, support vector machine (SVM) are applied

Tianyi wang proposed system where two information bases are joined along with macroeconomic information and AI calculations, for example, support vector machine, XGBoost are utilized to display the normal ticket cost dependent on source and objective sets. The system accomplishes a high forecast exactness 0.869 with the changed R squared execution measurements.

In "A linear quantile mixed regression model for prediction of airline ticket prices,", four LR models were compared to obtain the best fit model, which aims to provide an unbiased information to the passenger whether to buy the ticket or wait longer for a better price. The authors suggested using linear quantile mixed models to predict the lowest ticket prices, which are called the "real bargains". However, this work is limited to only one class of tickets, economy, and only on one direction single leg flights from San Francisco Airport to John F. Kennedy Airport. Wohlfarth et al. integrated clustering as a preliminary stage with multiple state-of the-art supervised learning algorithms (classification tree (CART) and RF) to assist the customers' decision making process. Their framework uses the K-Means algorithm to group flights with similar behavior in the price series. They then use CART to interpret meaningful rules, and RF to provide information about the importance of each feature. Also, the authors pointed out that one element, the number of seats left, is a key freature for ticket price prediction. Aside from flight-specific features, many other attributes affect the competitive market. Accurately predicting the market demand, for example, can reduce a travel agency's accumulated costs, which are caused by over purchasing or lost orders. In [19], the author applied Artificial Neural Network (ANN) and Genetic Algorithms (GA) to predict air ticket sales revenue for the travel agency. The input features included international oil price, Taiwan stock marketweighted index, Taiwan's monthly unemployment rate, and so on. Specifically, the GA selects the optimum input features to improve the performance of the ANNs. The model showed good performance with a 9.11% Mean Absolute
Percentage Error. Starting from 2017, more advanced machine learning models have been considered to improve airfare price prediction. Tziridis et al. applied eight machine
learning models, which included ANNs, RF, SVM, and LR, to predict tickets prices and compared their performance. The best regression model achieved an accuracy of 88%. In their comparison, Bagging Regression Tree is identified as the best model, which is robust and not affected by

using different input feature sets.
Utilizing AI models, [2] connected PLSR(Partial Least Square Regression) model to acquire the greatest presentation to get the least cost of aircraft ticket buying, having 75.3% precision. Janssen [3] presented a direct quantile blended relapse model to anticipate air ticket costs for cheap tickets numerous prior days takeoff. Ren, Yuan, and Yang [4], contemplated the exhibition of Linear Regression (77.06% precision), Naive Bayes (73.06% exactness, Softmax Regression (76.84% precision) and SVM (80.6% exactness) models in anticipating air ticket costs. Papadakis [5] anticipated that the cost of the ticket drop later on, by accepting the issue as a grouping issue with the assistance of Ripple Down Rule Learner (74.5 % exactness.), Logistic Regression with 69.9% precision and Linear SVM with the (69.4% exactness) Machine Learning models.

Gini and Groves took the Partial Least Square Regression(PLSR) for developing a model of predicting the best purchase time for flight tickets. The data was collected from major travel journey booking websites from 22 February 2011 to 23 June 2011. Additional data were also collected and are used to check the comparisons of the performances of the final model.

Janssen built up an expectation model utilizing the Linear Quantile Blended Regression strategy for SanFrancisco to NewYork course with existing every day airfares given by www.infare.com. The model utilized two highlights including the number of days left until the takeoff date and whether the flight date is at the end of the week or weekday. The model predicts airfare well for the days that are a long way from the takeoff date, anyway for a considerable length of time close the takeoff date, the expectation isnt compelling.

Wohlfarth proposed a ticket buying time enhancement model dependent on an extraordinary pre-preparing step known as macked point processors and information mining systems (arrangement and bunching) and measurable investigation strategy. This system is proposed to change over heterogeneous value arrangement information into added value arrangement direction that can be bolstered to unsupervised grouping calculation. The value direction is bunched into gathering dependent on comparative estimating conduct. Advancement model gauge the value change designs. A treebased order calculation used to choose the best coordinating group and afterward comparing the advancement model.

A study by Dominguez-Menchero recommends the ideal buying time dependent on nonparametric isotonic relapse method for a particular course, carriers, and timeframe. The model gives the most extreme number of days before buying a flight ticket. two sorts of the variable are considered for the expectation. One is the passage and date of procurement.

Chapter-3 Methodology

Proposed System

To develop a mobile application for “Airfare Prediction” based on previous airline ticket sales dataset for improving sales in Indian Domestic Airline. Our main motive is to provide the client with a prediction system from which it can take a right decision of increasing or decreasing the Airfare so that the flight doesn’t go empty or no money is lost due to sudden increase in crude oil.
a. To perform data analytics on customer’s ticket booking data for a brief amount of time.
b. To refine the data i.e. Removing duplicate records, ambiguity etc.
c. To perform Feature engineering in order to extract important feature from dataset for prediction.
d. To Brainstorm the Features i.e. to decide how to use those features
e. To create features i.e. to derive new features from those useful features.

The contribution of the proposed system includes the following activities :

1) Airfare prices prediction in India for domestic airline

The dataset contains 45 different columns from which we are extracting those columns (features) which will be used to train the model and predict the given goal.

2) Investigation and analysis of the features that affects the airfare.

The proposed system will use Data analytics for completely exploring the data, identifying relationship among those columns, finding some patterns of work, cleaning and refining the dataset to reduce complexity of data. Then the system will extract useful features from the bulky dataset and also will derive new features for making system processing easier.

3) Performance analysis of the ML models.

Random forest regression, Support vector machine, K-nearest neighbors Regression etc. Machine Learning algorithms will be used to train the model. Since the airfare is continuous value therefore Regression techniques will be used. At the end the system will generate report for “AIRFARE PREDICTION”.

The proposed system is composed of four phases:
1. Data input
2. Feature extraction
3. Machine learning model selection
4. Prediction

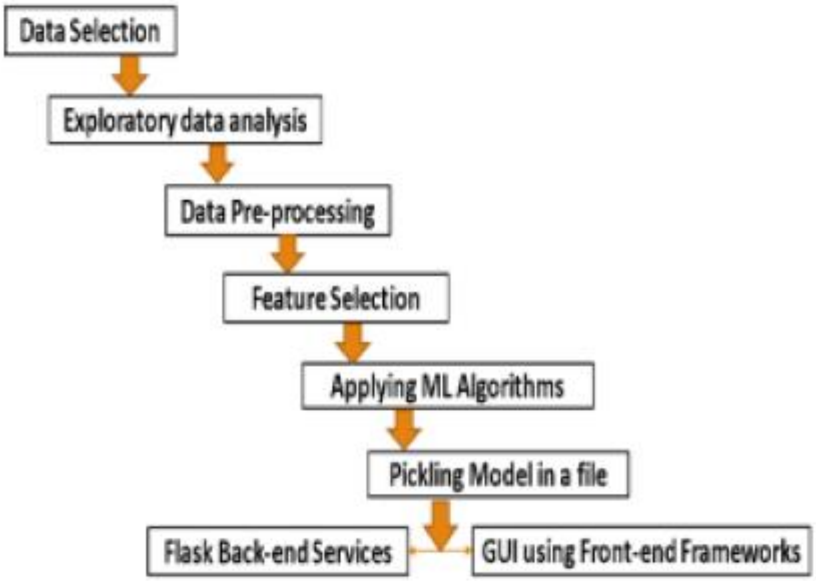
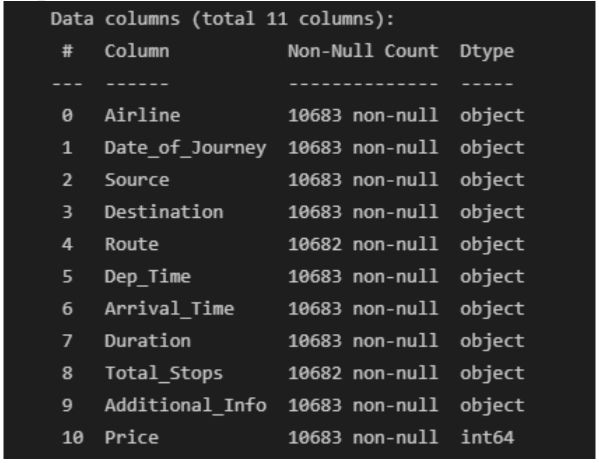


Fig 4 Machine Learning Life Cycle

Phase 1: Data Input

The input data file is in .csv file will be provided to system and that input file contains all customer ticket booking information. The training data contains 45 columns from which important features are extracted. The information is limited to domestic airline.



```
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          10683 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10683 non-null  object
 3   Destination      10683 non-null  object
 4   Route            10682 non-null  object
 5   Dep_Time         10683 non-null  object
 6   Arrival_Time     10683 non-null  object
 7   Duration         10683 non-null  object
 8   Total_Stops      10682 non-null  object
 9   Additional_Info  10683 non-null  object
 10  Price            10683 non-null  int64 
```

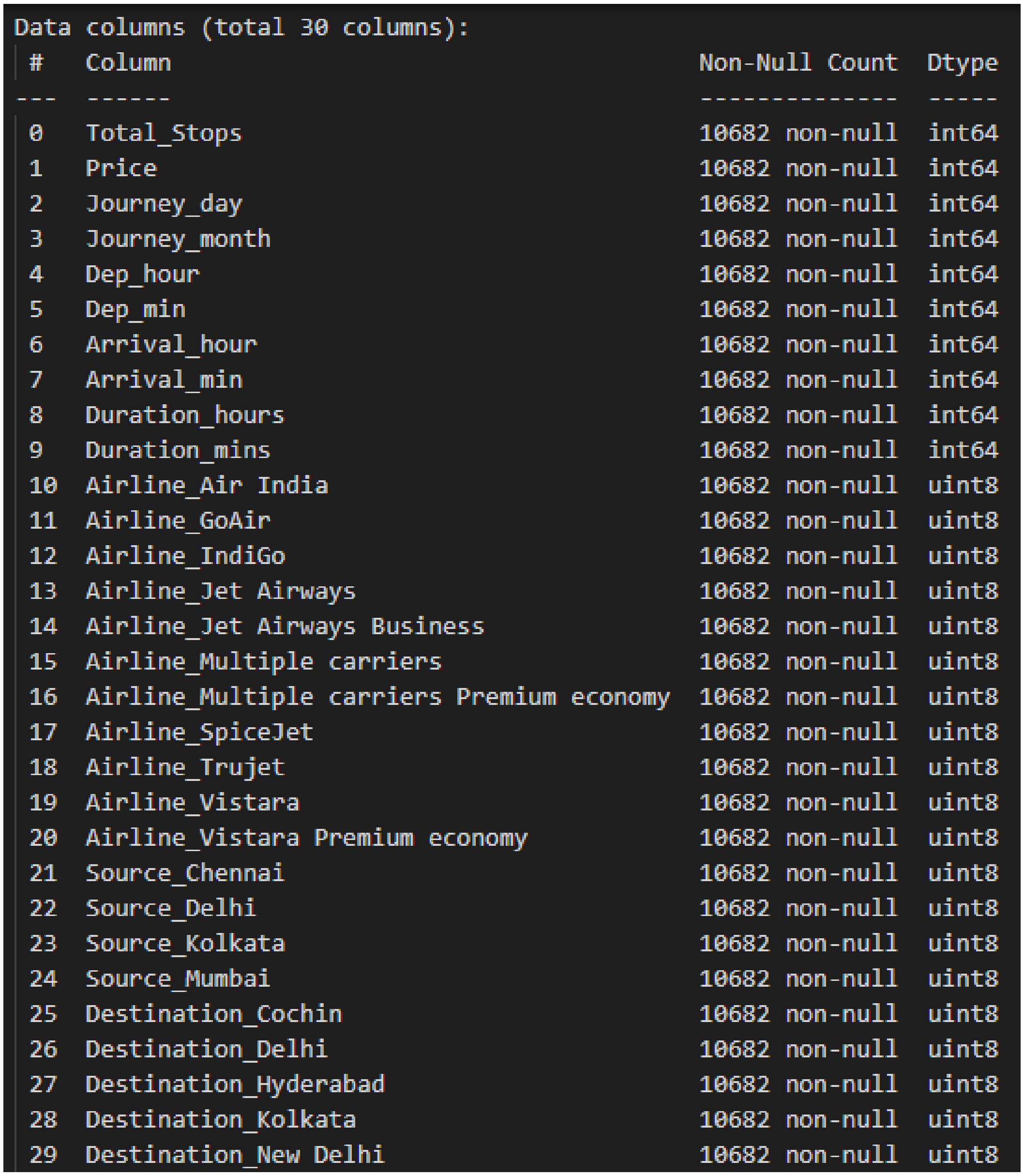
Fig 5 List of columns present in the dataset

Phase 2: Data Cleaning

According to the research 80% of the work is done in cleaning data and retrieving useful information from it. As the data is collected from live public domain i.e. Airline industry it contains many Null values, redundant entries, merged values, referential features and many unnecessary columns. Data cleaning steps are as follows :

1. Removing null values
2. Formatting of date columns
3. Removing outliers

4. Conversion of object, string and other data types into numeric form (Encoding).



```
Data columns (total 30 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Total_Stops                             10682 non-null  int64
 1   Price                                   10682 non-null  int64
 2   Journey_day                             10682 non-null  int64
 3   Journey_month                           10682 non-null  int64
 4   Dep_hour                                10682 non-null  int64
 5   Dep_min                                 10682 non-null  int64
 6   Arrival_hour                            10682 non-null  int64
 7   Arrival_min                             10682 non-null  int64
 8   Duration_hours                          10682 non-null  int64
 9   Duration_mins                           10682 non-null  int64
 10  Airline_Air India                       10682 non-null  uint8
 11  Airline_GoAir                           10682 non-null  uint8
 12  Airline_IndiGo                          10682 non-null  uint8
 13  Airline_Jet Airways                     10682 non-null  uint8
 14  Airline_Jet Airways Business            10682 non-null  uint8
 15  Airline_Multiple carriers               10682 non-null  uint8
 16  Airline_Multiple carriers Premium economy  10682 non-null  uint8
 17  Airline_SpiceJet                        10682 non-null  uint8
 18  Airline_Trujet                          10682 non-null  uint8
 19  Airline_Vistara                         10682 non-null  uint8
 20  Airline_Vistara Premium economy         10682 non-null  uint8
 21  Source_Chennai                          10682 non-null  uint8
 22  Source_Delhi                            10682 non-null  uint8
 23  Source_Kolkata                          10682 non-null  uint8
 24  Source_Mumbai                           10682 non-null  uint8
 25  Destination_Cochin                      10682 non-null  uint8
 26  Destination_Delhi                       10682 non-null  uint8
 27  Destination_Hyderabad                   10682 non-null  uint8
 28  Destination_Kolkata                     10682 non-null  uint8
 29  Destination_New Delhi                   10682 non-null  uint8
```

Fig 6. Data Set Columns After data cleaning

Phase 3: Feature Extraction

During this phase most of the informative features from the airline dataset that determines the prices of the air tickets are extracted. Features that can be considered are as follows:
Feature 1: Booking date and time
Feature 2: Departure date and time
Feature 3: Numbers of days till flight departure
Feature 4: Category of passenger (Adult/Child)
Feature 4: Cabin (Economy/Business)
Feature 5: Source Location

Feature 6: Destination Location
Figure 2 Extracted Features for training the model.

Phase 4: Machine Learning Model Selection
Machine learning is a science that uses statistical techniques to give computer system ability to learn from the given dataset without being explicitly programmed. The supervised learning algorithm deals with labeled data set training for predicting the results. Our system will be provided with label dataset and it is expected to predict the new input data. Therefore, we will use supervised machine learning algorithm.

In our Project we had carried out different Machine Learning Algorithms like Linear Regression, Decision Tree Regression, Random Forest Regression and thought about the precision of results dependent on our test informational index. In view of the different precision levels we observe that Random Forest Regression gives the most noteworthy exactness. In this manner we chose Random Forest Regression and made User Interface dependent on it.

Algorithms	Accuracy
Linear Regression	0.61
Decision Tree Regression	0.64
Random Forest Regression	0.81

Table 1. Accuracy of different ML Algorithm

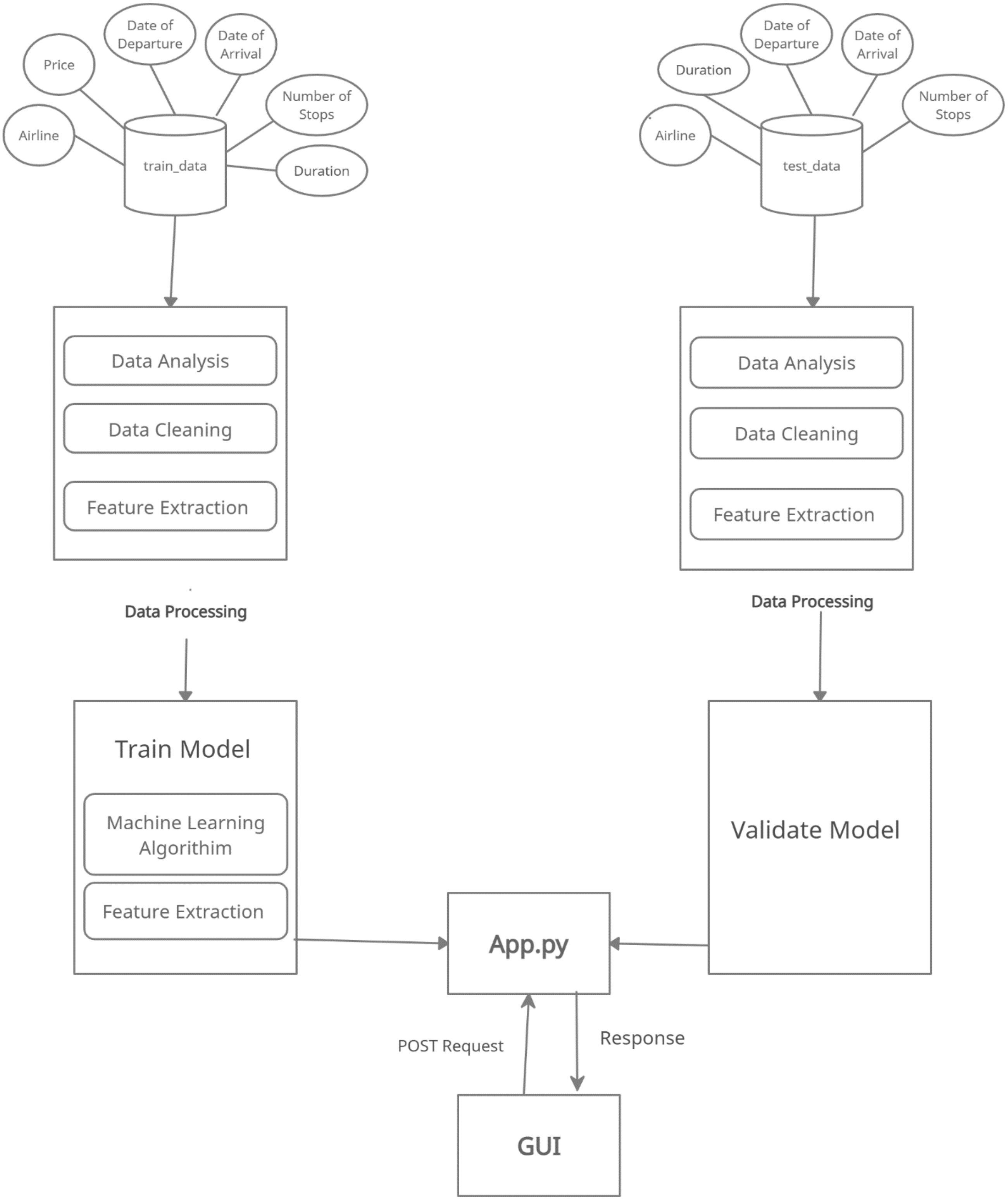
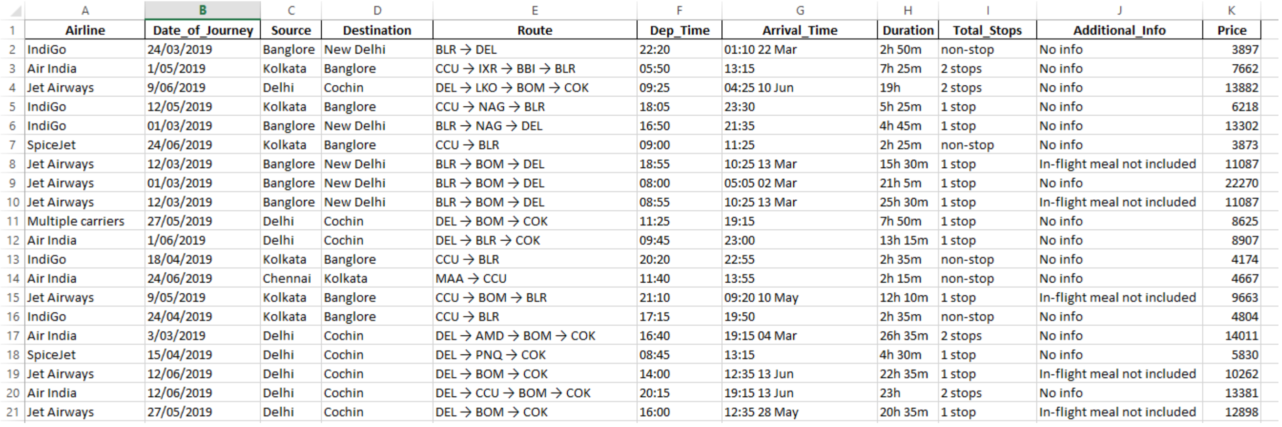


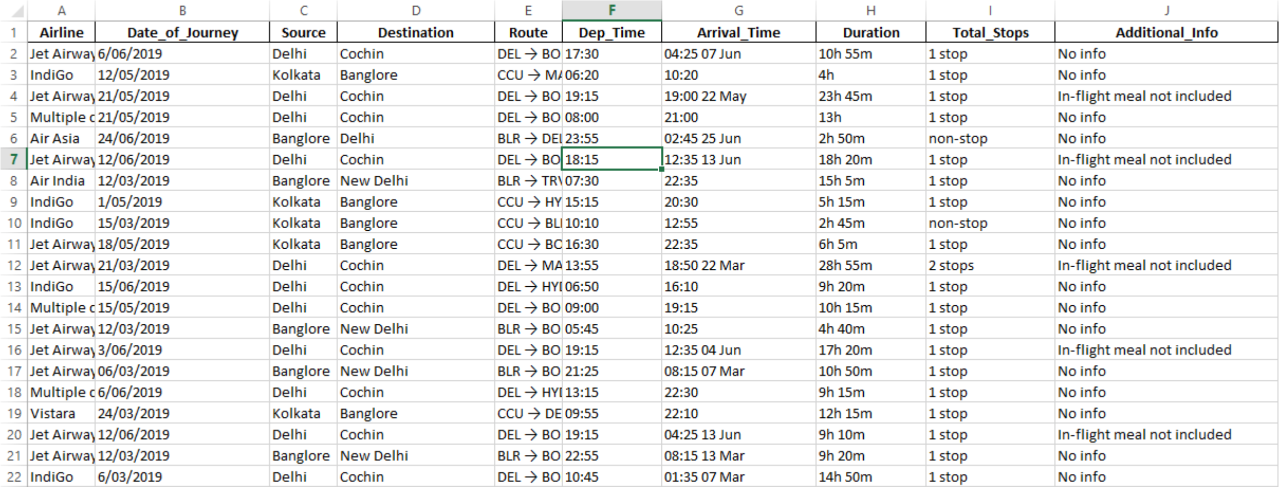
Fig 7. Architecture Diagram

There are also different cross-validation techniques such as gridsearchCV and randomizedsearchCV which will be used for improving the accuracy of the model. Parameters of the models such as number of trees in random forest or max depth of decision tree can be changed using this technique which will help us in further enhancement of the accuracy. The last three steps of the life cycle model are involved in the deployment of the trained machine learning model. Therefore, after getting the model with the best accuracy we store that model in a file using pickle module. The back-end of the application will be created using Flask Framework where API end-points such and GET and POST will be created to perform operations related to fetching and displaying data on the front-end of the application.



	A	B	C	D	E	F	G	H	I	J	K
1	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
2	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
3	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
4	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
5	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
6	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
7	SpiceJet	24/06/2019	Kolkata	Banglore	CCU → BLR	09:00	11:25	2h 25m	non-stop	No info	3873
8	Jet Airways	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	18:55	10:25 13 Mar	15h 30m	1 stop	In-flight meal not included	11087
9	Jet Airways	01/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:00	05:05 02 Mar	21h 5m	1 stop	No info	22270
10	Jet Airways	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:55	10:25 13 Mar	25h 30m	1 stop	In-flight meal not included	11087
11	Multiple carriers	27/05/2019	Delhi	Cochin	DEL → BOM → COK	11:25	19:15	7h 50m	1 stop	No info	8625
12	Air India	1/06/2019	Delhi	Cochin	DEL → BLR → COK	09:45	23:00	13h 15m	1 stop	No info	8907
13	IndiGo	18/04/2019	Kolkata	Banglore	CCU → BLR	20:20	22:55	2h 35m	non-stop	No info	4174
14	Air India	24/06/2019	Chennai	Kolkata	MAA → CCU	11:40	13:55	2h 15m	non-stop	No info	4667
15	Jet Airways	9/05/2019	Kolkata	Banglore	CCU → BOM → BLR	21:10	09:20 10 May	12h 10m	1 stop	In-flight meal not included	9663
16	IndiGo	24/04/2019	Kolkata	Banglore	CCU → BLR	17:15	19:50	2h 35m	non-stop	No info	4804
17	Air India	3/03/2019	Delhi	Cochin	DEL → AMD → BOM → COK	16:40	19:15 04 Mar	26h 35m	2 stops	No info	14011
18	SpiceJet	15/04/2019	Delhi	Cochin	DEL → PNQ → COK	08:45	13:15	4h 30m	1 stop	No info	5830
19	Jet Airways	12/06/2019	Delhi	Cochin	DEL → BOM → COK	14:00	12:35 13 Jun	22h 35m	1 stop	In-flight meal not included	10262
20	Air India	12/06/2019	Delhi	Cochin	DEL → CCU → BOM → COK	20:15	19:15 13 Jun	23h	2 stops	No info	13381
21	Jet Airways	27/05/2019	Delhi	Cochin	DEL → BOM → COK	16:00	12:35 28 May	20h 35m	1 stop	In-flight meal not included	12898

Fig 8 Sample Training Dataset



	A	B	C	D	E	F	G	H	I	J
1	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info
2	Jet Airway	6/06/2019	Delhi	Cochin	DEL → BO	17:30	04:25 07 Jun	10h 55m	1 stop	No info
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → MA	06:20	10:20	4h	1 stop	No info
4	Jet Airway	21/05/2019	Delhi	Cochin	DEL → BO	19:15	19:00 22 May	23h 45m	1 stop	In-flight meal not included
5	Multiple c	21/05/2019	Delhi	Cochin	DEL → BO	08:00	21:00	13h	1 stop	No info
6	Air Asia	24/06/2019	Banglore	Delhi	BLR → DEL	23:55	02:45 25 Jun	2h 50m	non-stop	No info
7	Jet Airway	12/06/2019	Delhi	Cochin	DEL → BO	18:15	12:35 13 Jun	18h 20m	1 stop	In-flight meal not included
8	Air India	12/03/2019	Banglore	New Delhi	BLR → TR	07:30	22:35	15h 5m	1 stop	No info
9	IndiGo	1/05/2019	Kolkata	Banglore	CCU → HY	15:15	20:30	5h 15m	1 stop	No info
10	IndiGo	15/03/2019	Kolkata	Banglore	CCU → BL	10:10	12:55	2h 45m	non-stop	No info
11	Jet Airway	18/05/2019	Kolkata	Banglore	CCU → BO	16:30	22:35	6h 5m	1 stop	No info
12	Jet Airway	21/03/2019	Delhi	Cochin	DEL → MA	13:55	18:50 22 Mar	28h 55m	2 stops	In-flight meal not included
13	IndiGo	15/06/2019	Delhi	Cochin	DEL → HY	06:50	16:10	9h 20m	1 stop	No info
14	Multiple c	15/05/2019	Delhi	Cochin	DEL → BO	09:00	19:15	10h 15m	1 stop	No info
15	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BO	05:45	10:25	4h 40m	1 stop	No info
16	Jet Airway	3/06/2019	Delhi	Cochin	DEL → BO	19:15	12:35 04 Jun	17h 20m	1 stop	In-flight meal not included
17	Jet Airway	06/03/2019	Banglore	New Delhi	BLR → BO	21:25	08:15 07 Mar	10h 50m	1 stop	No info
18	Multiple c	6/06/2019	Delhi	Cochin	DEL → HY	13:15	22:30	9h 15m	1 stop	No info
19	Vistara	24/03/2019	Kolkata	Banglore	CCU → DE	09:55	22:10	12h 15m	1 stop	No info
20	Jet Airway	12/06/2019	Delhi	Cochin	DEL → BO	19:15	04:25 13 Jun	9h 10m	1 stop	In-flight meal not included
21	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BO	22:55	08:15 13 Mar	9h 20m	1 stop	No info
22	IndiGo	6/03/2019	Delhi	Cochin	DEL → BO	10:45	01:35 07 Mar	14h 50m	1 stop	No info

Fig 9 Sample Testing Dataset

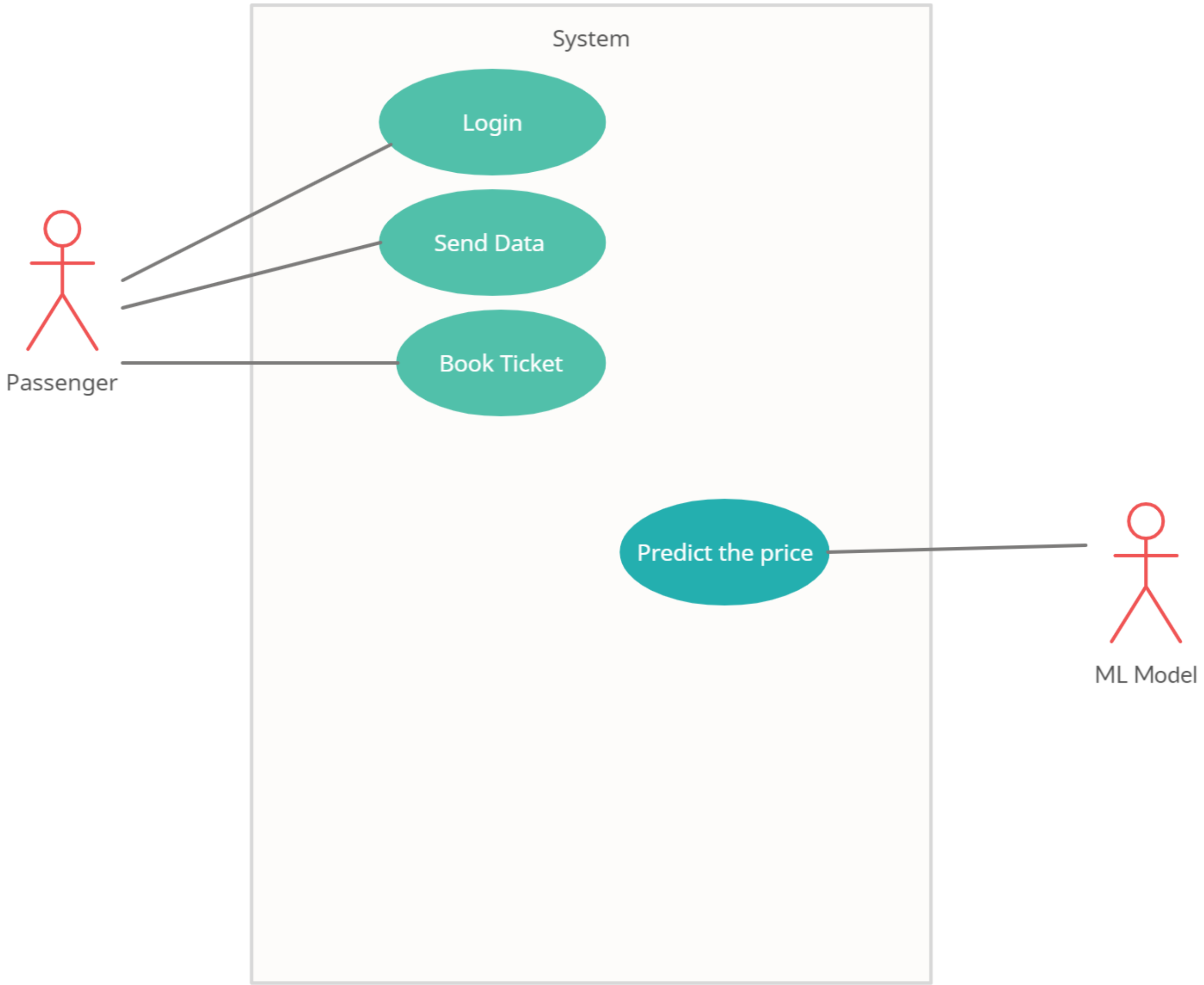


Fig 10. Use Case Diagram

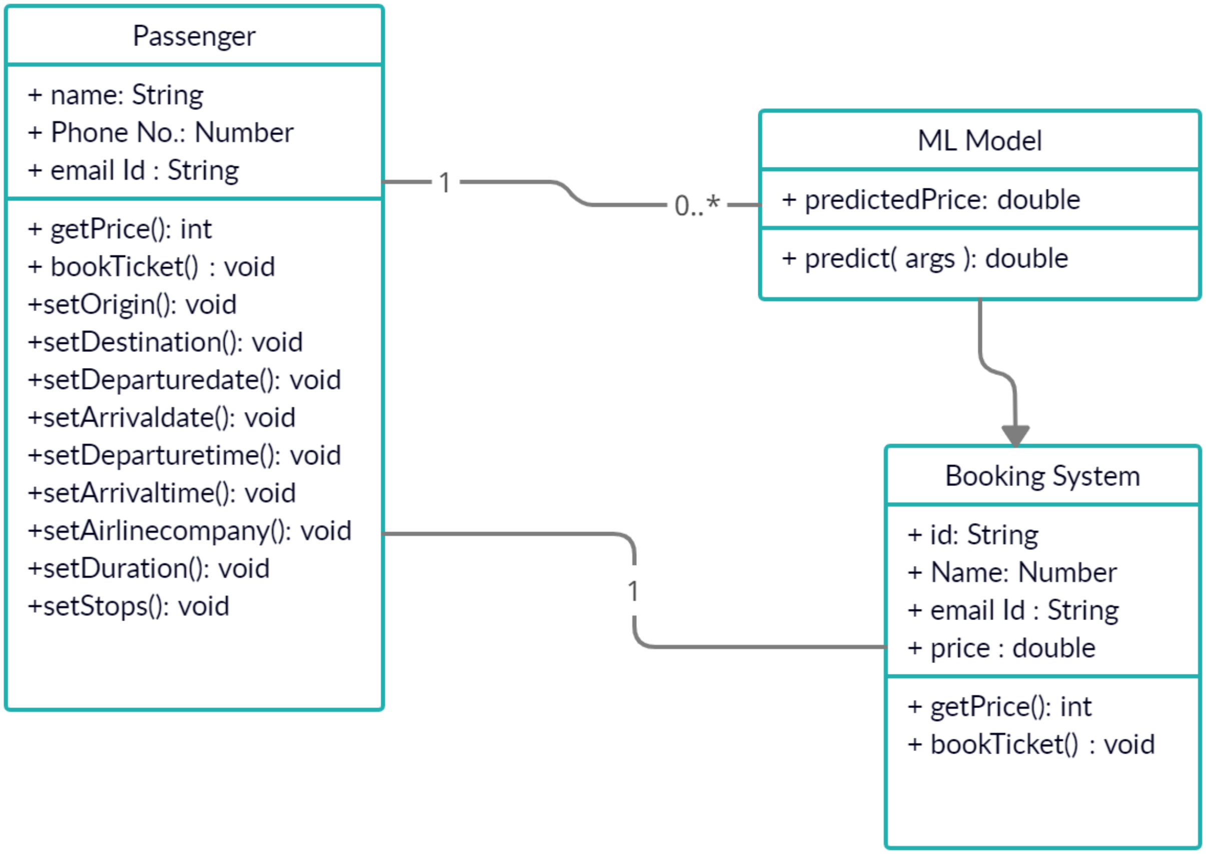
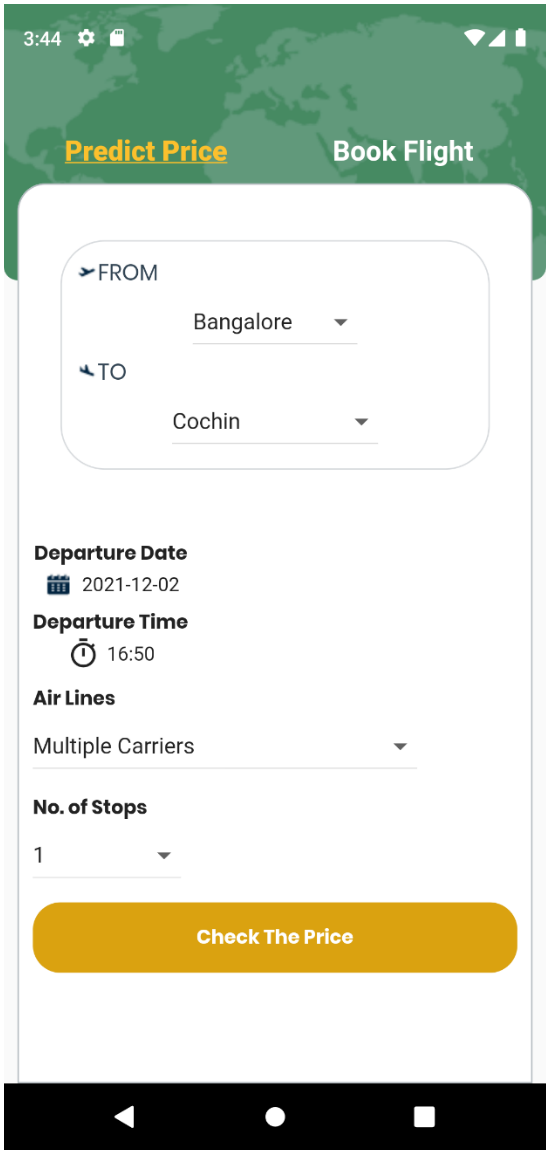


Fig 11. Class Diagram

Chapter – 4
Functionality/Working of Project

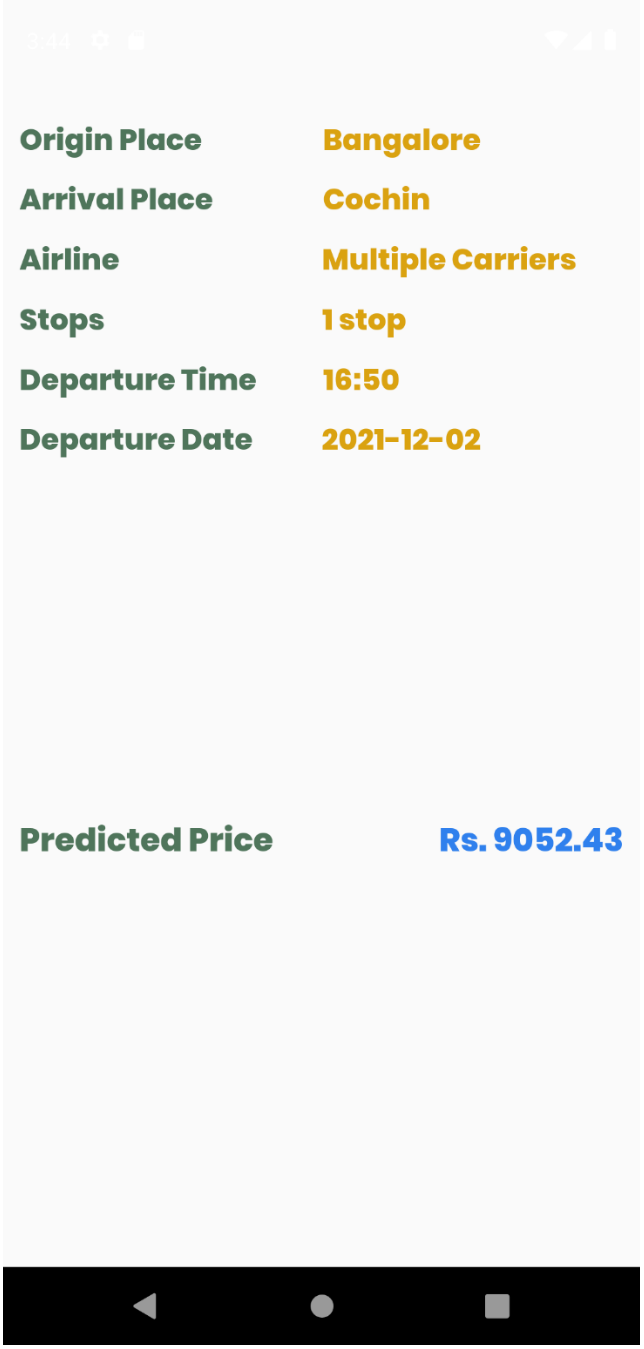


FROM
Bangalore
TO
Cochin

Departure Date
2021-12-02
Departure Time
16:50
Air Lines
Multiple Carriers
No. of Stops
1

Fig 12. Make Prediction

User fill the data of the flight whose price they want to predict. User will fill destination city and origin city along with the departure date and time, Air Lines and number of stops of the flight they want.

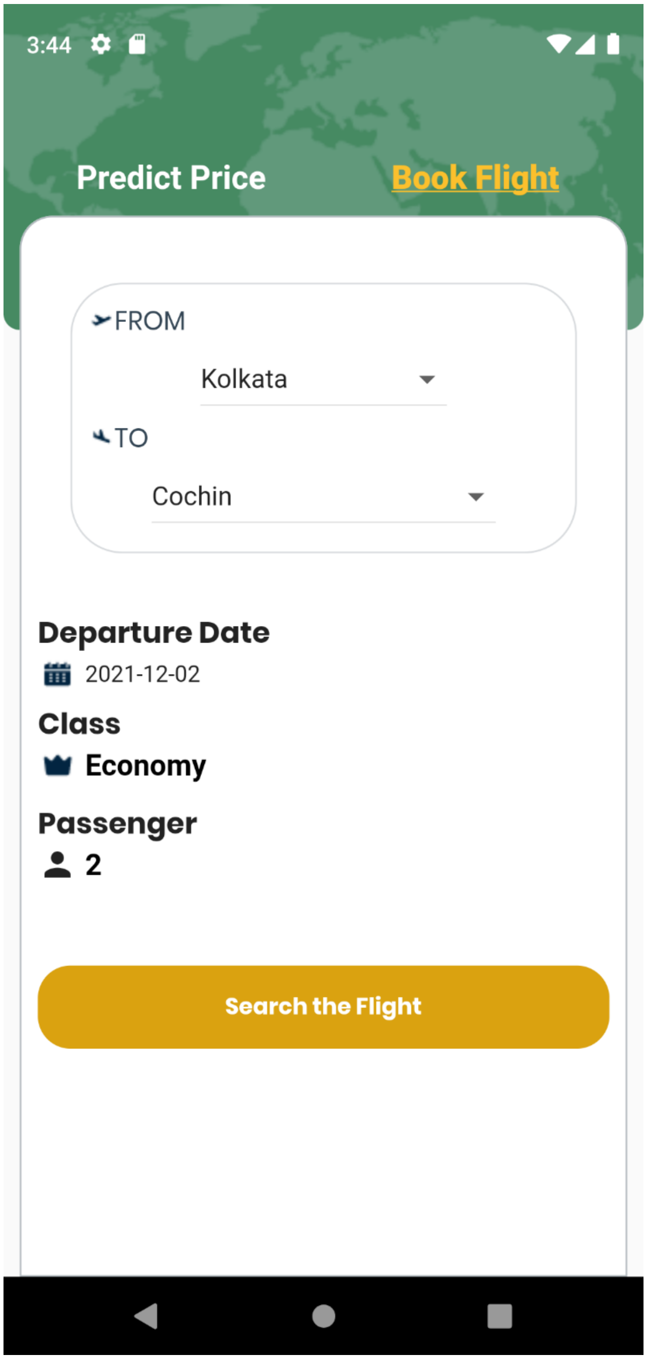


Origin Place	Bangalore
Arrival Place	Cochin
Airline	Multiple Carriers
Stops	1 stop
Departure Time	16:50
Departure Date	2021-12-02

Predicted Price	Rs. 9052.43

Fig 13. Prediction Page

This page display the predicted Price based on the information given by the user. Information is passed on to machine learning model and then the price is predicted.



FROM
Kolkata

TO
Cochin

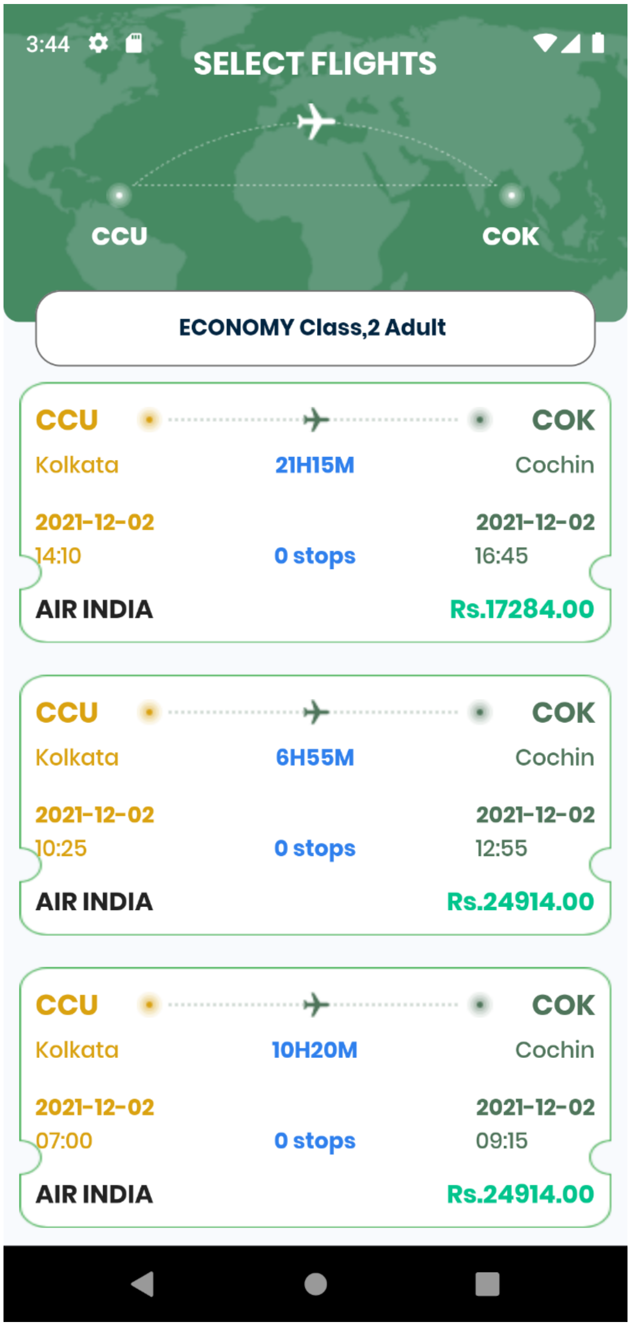
Departure Date
2021-12-02

Class
Economy

Passenger
2

Fig 14. Booking Flight

If user feels to book ticket instead of predicting price of the flight they can use this feature and search for the flight.



ECONOMY Class,2 Adult

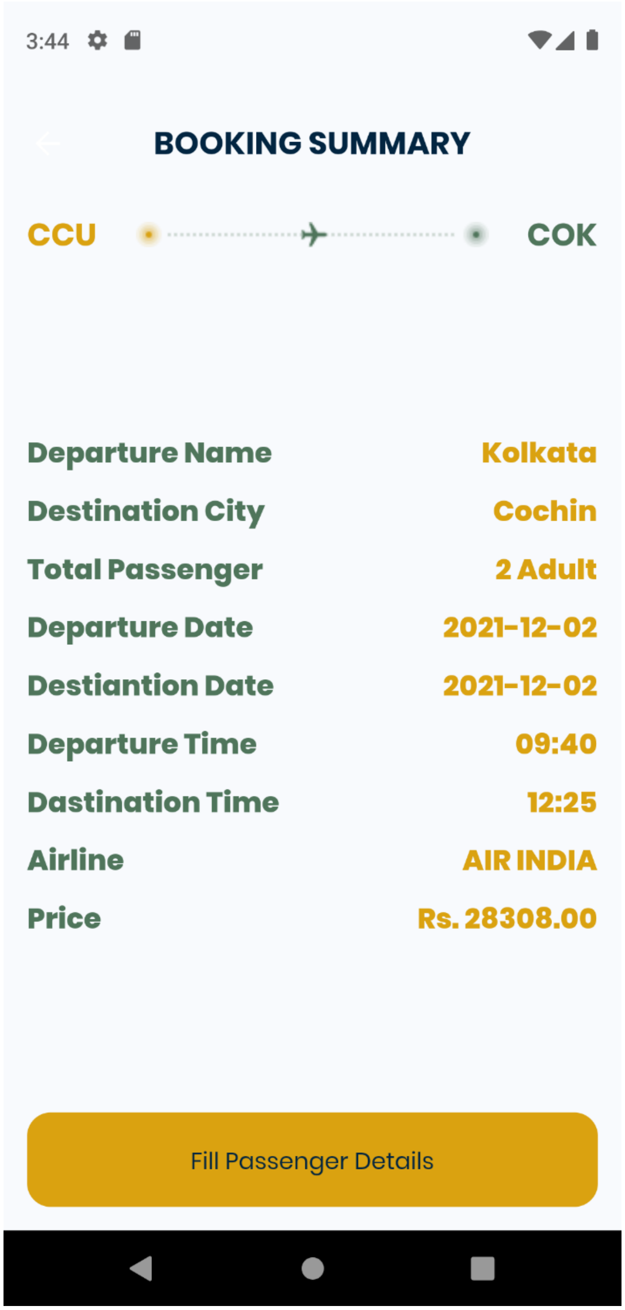
CCU Kolkata 2021-12-02 14:10 — 21H15M — 0 stops — COK Cochin 2021-12-02 16:45
AIR INDIA Rs.17284.00

CCU Kolkata 2021-12-02 10:25 — 6H55M — 0 stops — COK Cochin 2021-12-02 12:55
AIR INDIA Rs.24914.00

CCU Kolkata 2021-12-02 07:00 — 10H20M — 0 stops — COK Cochin 2021-12-02 09:15
AIR INDIA Rs.24914.00

Fig 15. Flight Result

Information is sent to Amadeus developer API which give flights available as the result. Flight Available are showed with duration boarding and dropping airport along with number departure date and time, destination date and time and airline.



BOOKING SUMMARY

CCU → COK

Departure Name	Kolkata
Destination City	Cochin
Total Passenger	2 Adult
Departure Date	2021-12-02
Destiantion Date	2021-12-02
Departure Time	09:40
Dastination Time	12:25
Airline	AIR INDIA
Price	Rs. 28308.00

Fig 16. Booking Details

After the search result this screen display the detail of the selected flight.

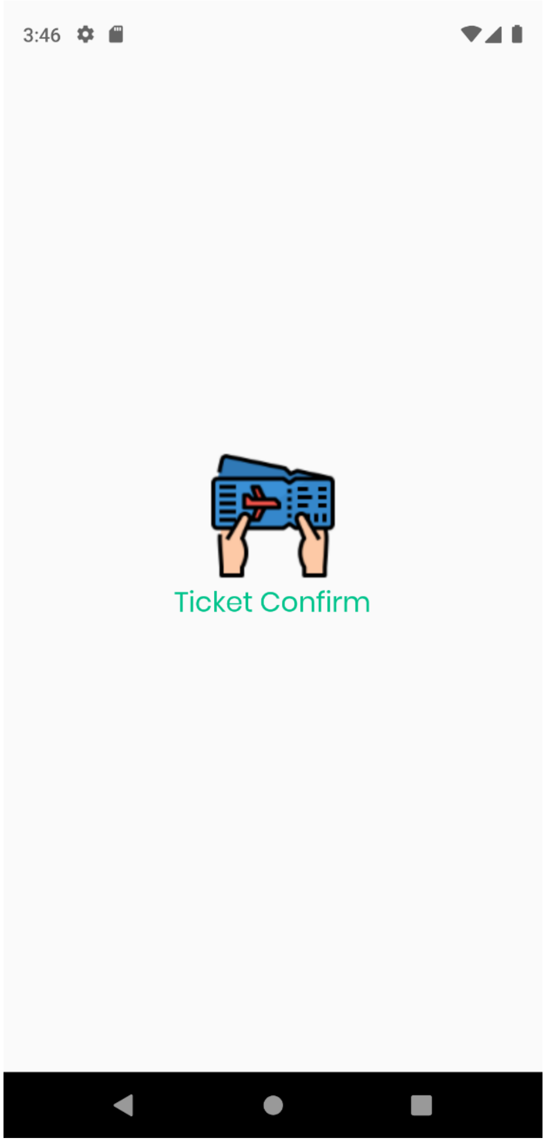


Fig 17

This page notifies the user that their flight ticket is booked.

Chapter 5 - Results and Discussion

1. EXPERIMENTAL RESULTS-

In our project we had implemented various Machine Learning Algorithms such as Linear Regression, Decision Tree Regression, Random Forest Regression and compared the accuracy of results based on our test data set. Based on the various accuracy levels we find that Random Forest Regression gives the highest accuracy i.e. 81%. Therefore we selected Random Forest Regression and created User Interface based on it.

Algorithms	Accuracy
Linear Regression	0.62
Decision Tree Regression	0.65
Random Forest Regression	0.81

Table 2.experiment result

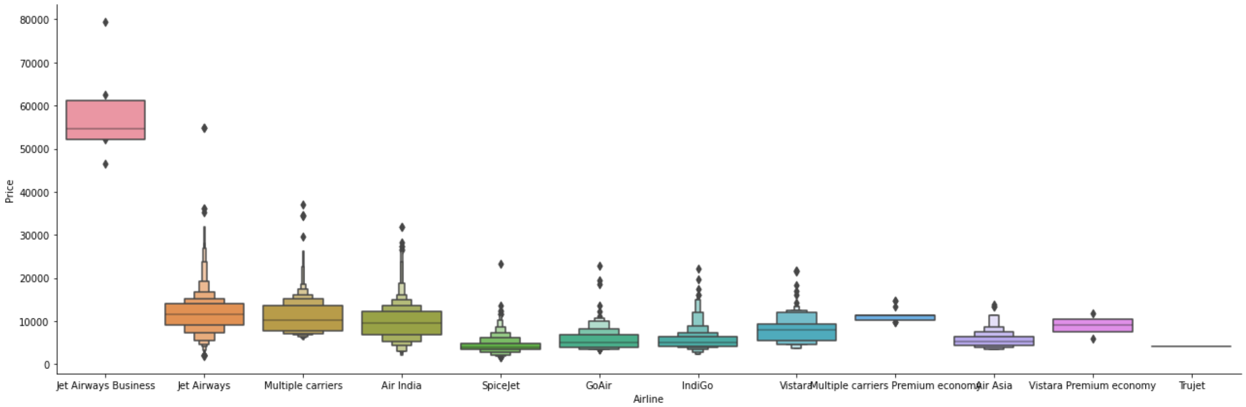


Fig 18 Box Plot for Airline VS Price

As we can see the name of the airline matters. 'JetAirways Business' has the highest price range. Other airlines price also varies.

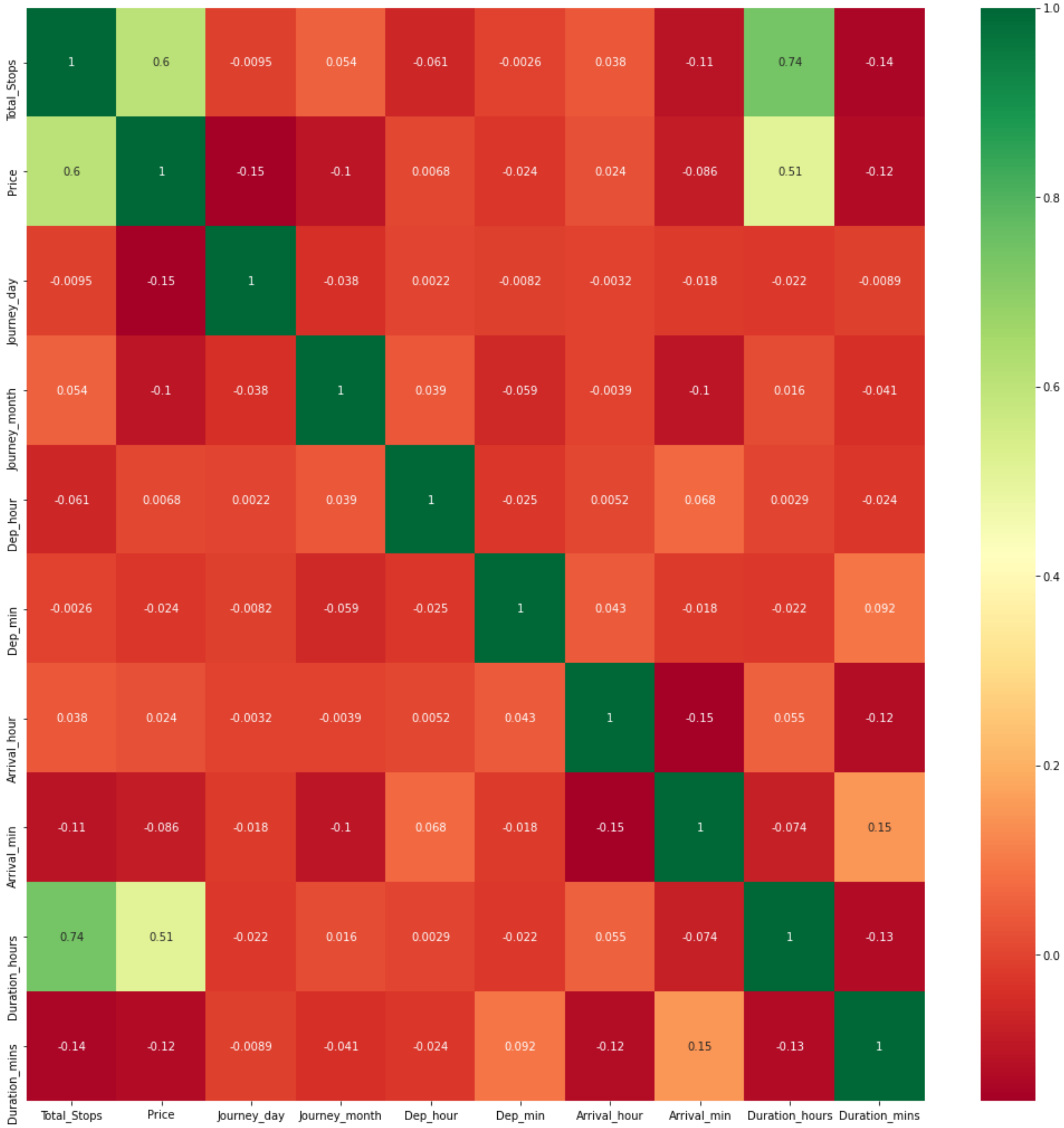


Fig 19 Heat Map

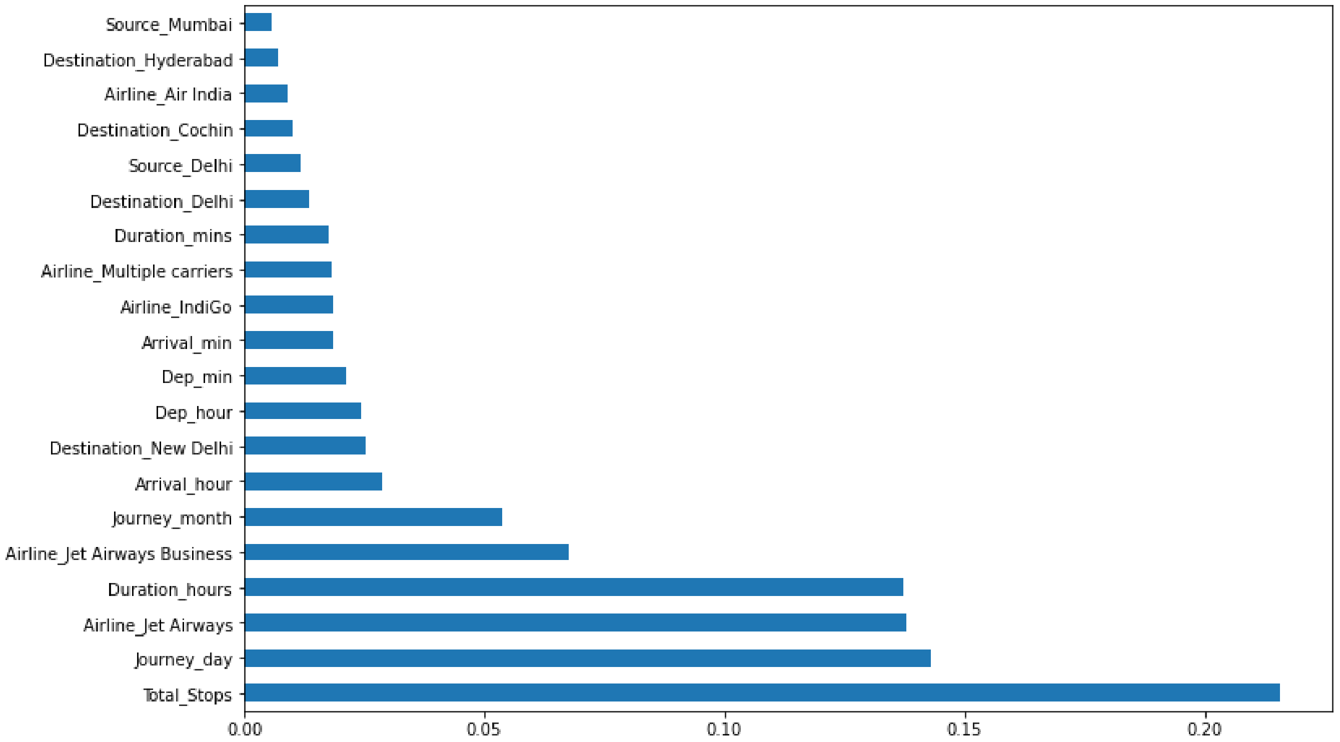


Fig 20 Importance of the Attributes

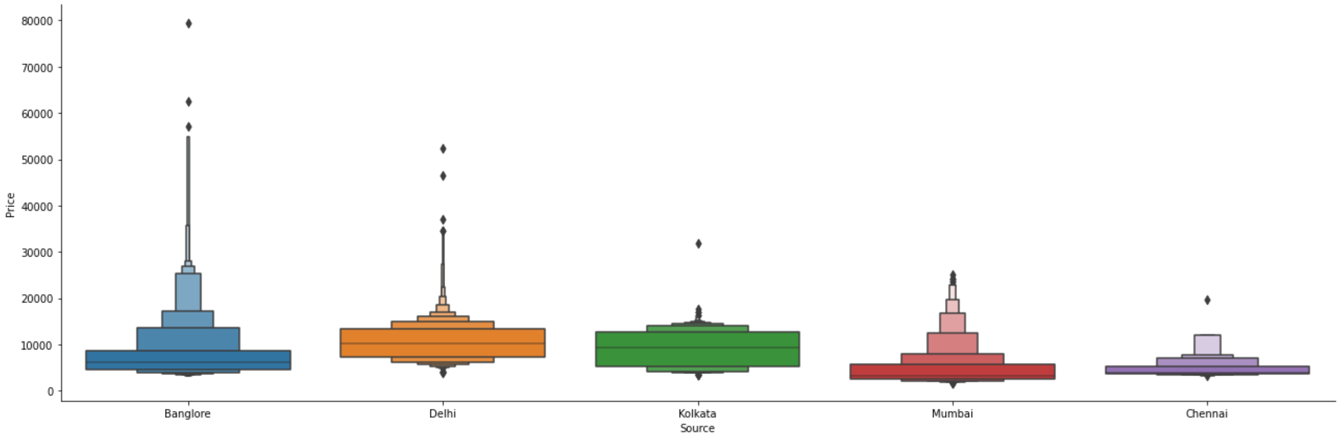


Fig 21 Box Plot for Destination VS Price

2. Performance Metrics

Performance measurements are factual models which will be used to think about the exactness of the AI models prepared by various calculations.

MAE (Mean Absolute Error)

Mean Absolute Error is fundamentally the amount of normal of the outright distinction between the anticipated and real qualities.

$MAE = 1/n[\Sigma(y-\acute{y})]$

Lesser the value of MAE the better the performance of your
model.

MSE (Mean Square Error)

Mean Square Error squares the distinction of real and anticipated result esteems prior to adding them all rather than utilizing the outright worth.

$MSE = 1/n[\Sigma(y-\acute{y})2]$

MSE punishes big errors as we are squaring the errors. Lower the value of MSE the better the performance of the model.

RMSE

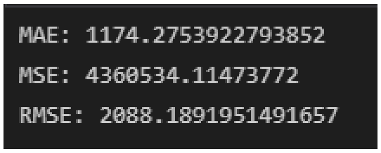
It is more noteworthy than MAE and lesser the worth of RMSE between various model the better the presentation of that model. R2 (Coefficient of assurance)It assists you with seeing how well the free factor changed with the fluctuation in your model.

$R2 = \mathbf{1} - \Sigma(\acute{y}-\bar{y})2$

$\Sigma(y-\bar{y})2$

The worth of R-square lies between 0 to 1. The nearer its worth to one, the better your model is when contrasting and other model qualities.

In order to implement Random forest regression tree we used number of estimators as 1000 and number of random states were 42. This algorithm is well suited for unstructured data where dependencies among the features are quite difficult to identify.



```
MAE: 1174.2753922793852
MSE: 4360534.11473772
RMSE: 2088.1891951491657
```

Fig 22. Performance Matrix

Limitations:

The existing systems doesn’t provide any severe drawbacks but it did have certain limitations

a. The system doesn’t have sufficient data for better prediction.

b. The system changes accuracy with changing algorithm and so it becomes a bit confusing though the accuracy only changes much when features important features are removed.

Chapter 6 - Conclusion and Future Scope

Conclusion

Presently, there are many fields where expectation based administrations are utilized, for example, stock value indicator apparatuses utilized by stock dealers and administration like Zestimate which gives the assessed worth of house costs. In this manner, there is necessity for administration like this in the flight business which can help the clients in booking tickets. There are many investigates works that have been done on this utilizing different procedures and more examination is expected to work on the exactness of the expectation by utilizing various calculations. More precise information with better elements can be likewise be utilized to get more exact outcomes.

Future Scope

Later on, our system can be stretched out to incorporate air ticket exchange data, which can give more insight concerning a particular schedule, like time and date of takeoff and appearance, seat area, covered auxiliary items, and so forth By joining such information with the current market fragment and macroeconomic highlights in the current structure, it is feasible to construct an all the more impressive and thorough airfare value forecast model on the day by day or even hourly level. Moreover, airfare cost in a market portion can be impacted by an unexpected inundation of enormous volume of travelers brought about by some exceptional occasions. Accordingly, occasions data will likewise be gathered from different sources, which incorporate social stages and news offices, as to supplement our expectation model. Also, we will explore other progressed ML models, for example, Deep Learning models, while attempting to work on the current models by tuning their hyper-boundaries to arrive at the best engineering for airfare value expectation

REFERENCES

[1] Jaywrat Singh Champawat, Udhhav Arora, Dr. K. Vijaya, “INDIAN FLIGHT FARE PREDICTION: A PROPOSAL”

[2] Tianyi Wang , Samira Pouyanfar , Haiman Tian , Yudong Tao, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen ,“A Framework for Airfare Price Prediction: A Machine Learning Approach”

[3] Vinod Kimbhaune, Harshil Donga, Asutosh Trivedi, Sonam Mahajan and Viraj Mahajan., “Flight Fare Prediction System”

[4] Jibin Joseph , Abhijith P , Aryasree S , Jinsu Anna Joseph , Meghana Sara Oommen, Abin T Abraham “Flight Ticket Price Predicting With the Use of Machine Learning”

[5] Aakanksha V. Jain, Aanal S.Raval, Ruchi K. Oza, “Airfare Price Prediction Based on Reviews Using Machine Learning Techniques”

[6] Jaya Shukla, Aditi Srivastava, and Anjali Chauhan, “Airline Price Prediction using Machine Learning”

[7] Dominguez Menchero, J.Santo, Reviera, ”optimal purchase timing in airline markets” ,201