Flight Fare Prediction System

A Project Report

M.Sc.IT Information Technology

Semester- III

A Project Report

Submitted By

Tanuj Mahesh Dulam.

Seat No: 3269839

YEAR 2022-2023

Submitted in Partial Fulfilment of requirement for qualifying

M.Sc.IT Part I (Sem-III) Examination

UNIVERSITY OF MUMBAI

VIDYA VIKAS EDUCATION SOCIETY'S

VIKAS COLLEGE OF ARTS, SCIENCE & COMMERCE

VIKHROLI (E)-400 083

Phone：**257 83540**

25784267

Fax : 25796196

Vidya Vikas Education Society's

VIKAS COLLEGE OF ARTS, SCIENCE & COMMERCE

Affliated to University of Mumbai

RE-ACCREDITED 'A' GRADE BY NAAC (WITH CGPA 3.15

ISO 9001 : 2008 CERTIFIED

Vikas High School Marg, Kannamwar Nagar No 2, Vikhroli (E), Mumbai – 400083

Dr. R. K. Patra
Principal

Hon' ble: **Shri P. M. Raut** Chairman. V. V. Edu. Society

Email : vikascollegeprincipal@gmail.com www.vikascollege.org

This is to certify that, **Tanuj Mahesh Dulam.**Student of M.Sc.IT Part II (Sem-III) with Seat No. **3269839** and college enrolled Roll no.**222905** has satisfactorily completed the practical work in Information Technology Laboratory for the Course **project Documentation** in the program of INFORMATION TECHNOLOGY from the UNIVERSITY OF MUMBAI for the academic year 2022-2023.

Subject In-Charge: HOD:

Examiner: _____

ABSTRACT

This is the project report of Flight price prediction. The main aim of this project is to create a model which can take the input from user and based on that it calculate the price of flight ticket. It is based on the supervised machine learning algorithm i.e. random forest regression which classify the continuous data based on different features like route, origin, destination, date and etc.

The reason for deploying the model is that everyone doesn't know how to run python so we have created the end user web application using flask so that everyone can use it.

These are the places where we usually apply the machine learning concept to solve real world problem or do the prediction, so basically we can say supervised machine learning technique is exciting and potentially far reaching in development of computer science. These enables a computer to learn automatically from large amount of data and that can be used to automatically do prediction or help people make decision fast and accurate.

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 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.

INDEX

1. INTRODUCTION

The main aim of this project is to create a model which can take the input from user and based on that it calculate the price of flight ticket .It is based on the supervised machine learning algorithm i.e. random forest regression which classify the continuous data based on different features like route , origin , destination, date and etc.

The reason for deploying the model is that everyone doesn't know how to run python so we have created the end user web application using flask so that everyone can use it. It uses less data and quickly predicts the price of your ticket with different domestic airlines

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".

1.2 Background –

Flight booking systems are dynamic in nature. They depend on a lot of features like Airline company, Source, Destination, duration, arrival time, departure time, number of stops and date of the flight. In this project, I plan to use machine learning algorithms on a dataset based on the above parameters to predict flight prices There are basically two approaches to solve this problem. These involve considering it as a regression or classification problem. Algorithms can be applied to predict whether the price of the ticket will drop in the future, thus considering it as a classification problem. In this project, I will consider it as a regression problem, thus predicting the ticket price.

1.3 Objective -

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics known as "revenue management" or "yield management". The cheapest available ticket for a given date gets more or 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)

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).

So, if we could inform the travelers with the optimal time to buy their flight tickets based on the historic data and also show them various trends in the airline industry we could help them save money on their travels. This would be a practical implementation of a data analysis, statistics and machine learning techniques to solve a daily problem faced by travelers.

1.4 Purpose -

● The purpose of this project is to help people or end users by providing a better, fast, efficient web application which is without advertisement so that one can estimate the flight ticket price easily .

1.4 Scope and Applicability -

- The aim of this project is to detect the prices of different flights as compare to today to another day due to which help to people to book the flight ticket according to their need.
- The overall goal of the project is to create the web UI which will predict the price to the customer on the basis of the given input.
- Nowadays there are lots of apps for flight ticket booking. if passenger want to travel from one space to another space so they don't know actually what is the prices of that same space flight.
- To save their money and time we will decide to develop such system due to which user can book the flight ticket according to their need.
- Currently, there are many fields where prediction-based services are used such as stock price predictor tools used by stock brokers 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 researches 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 be also be used to get more accurate results

2. LITERATURE REVIEW

- In the preceding work on improving prediction models for airline prices by using Machine Learning (ML) techniques, the different exploration team has concentrated on various attributes and have trained the models on various kinds of Airlines. Specific trend is that they are trying to predict the price. Specifically, categorizing flight price with two divisions of elements helps the studied impact on mean price of the plane. Authors have examined the airline profit by applying pricing modes and have found that after a time duration of 70 days, categorical cases for a flight are observed as flight departure and the discount opportunities also tend to increase over time. Through the analysis we identify equal pricing techniques applied by the airline companies to positively manage the airline offers and demand to increase their business profit. Results shows airlines worry about the price changes according to the season in websites.
- The point should be noted down that the importance between the online pricing, and the realised price dispersion on a flight. At the end using prices from actual transactions, the authors have found that online price division is more highly present in lower business airline competition
- Predicting the plane ticket prices and limiting the price for passengers. Price models with reliability can assist passengers to determine the scope of future prices for the airline companies. Present business airline companies will not give passengers to estimate the future reliable costs of any departure requirements. For the usually travelling passengers in this model was developed with price attribute from their own history.
- First, the high price duration will not be the minimum price accessible for a plane, the passenger data is specifically scheduled that may need to be changed for the price. The attribute from the dataset with subject information also as mentioning high understanding is not required. The Final model can be inspected for subject understanding. In the final work there are extra price limitations that pull out to have outcomes nearer to the accurate solution.
- In detail monitoring, the passenger gets an approximation of plane price with date to choose the best blend of date and price. The price for weekend on

Sunday is not possible to calculate in this presented model, as weekend on Sundays the most accidental price difference compared to other days in the week and needs more elements, nonlinear model for successful forecast which will be the upcoming range of study to be done for this presented technique. Selecting feature techniques authors have presented model to forecast the mean flight amount with R squared score of 80% accuracy.

SURVEY OF TECHNOLOGIES

PYTHON : -

It is a modern and general-purpose • programming language. It is object oriented • It is easy to learn. Open source

Many libraries of computer vision support python.

FLASK : -

Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web- Application.

Built-in development server, fast debugger.

Integrated support for unit testing.

Restful request dispatching.

Jinja2 Templating.

Support for secure cookies.

Lightweight and modular design allows for a flexible framework.

HEROKU APP: -

Heroku is a container-based cloud

Platform as a Service (PaaS). Developers use Heroku

to deploy, manage, and

scale modern apps. Our platform is elegant, flexible, and easy to use,

offering developers the simplest path to getting their apps to market.

Support for modern open-source languages.

Smart containers, elastic runtime.

Simple horizontal and vertical scalability.

Trusted application operations.

Built for continuous integration and delivery.

3. Requirements Analysis

3.1 Non- Functional requirements -

1. Security: Protection of the system and its data is an important aspect while creating any project.
2. Reliability: The system must be reliable.
3. Portability: System needs to be portable and platform independent, my model works on all browsers.
4. Maintainability: The program written and the libraries used in the model are easy to maintain and easy to modify.
5. Scalability: Model takes around 10 sec to run the whole process.
6. Usability: Model is simple and easy to use, hence maintaining its usability.

3.2 PLANNING AND SCHEDULING

Planning: Planning can be thought as determining all the small tasks that must be carried out in order to accomplish the goal.

Scheduling: scheduling can be thought of as determining whether adequate resources are available to carry out the plan.

3.3 Software and Hardware Specification -

Software:

4. Operating System : Windows 10
5. Graphics : Intel HD Graphics 630
6. DPI : Normal Size
7. Screen Refresh Rate : 60-80Hz
8. Color Quality : Highest 32 bits 9. Anaconda, Jupiter Notebook, Python 3
10.

Hardware:

1. Intel Core i5 7600 CPU 3.5GHz
2. 8GB DDR4 RAM
3. 1TB Hard Disk
4. Mouse
5. Keyboard

4. System Design

- **Data Collection :-**

Since the APIs by Indian companies like Goibibo returned data in a complex format resulting in a lot of time to clean the data before analysing, therefore we decided to build a web spider that extracts the required values from a website and stores it as a CSV file. We decided to scrape travel service providers website using a manual spider made in Python. Further we also developed a Python script to run the API provided by Google flights which is more reliable, but it allows only 50 queries each day.

Such scrapping returns numerous variables for each flight returned and we had to decide the parameters that might be needed for the flight prediction algorithm. Not all are required and thus we selected the following

1. Origin City
2. Destination City
3. Departure Date
4. Departure Time
5. Arrival Time
6. Total Fare
7. Airway Carrier
8. Duration
9. Class Type - Economy/Business
10. Flight Number
11. Hopping - Boolean
12. Taken Date - date on which this data was collected

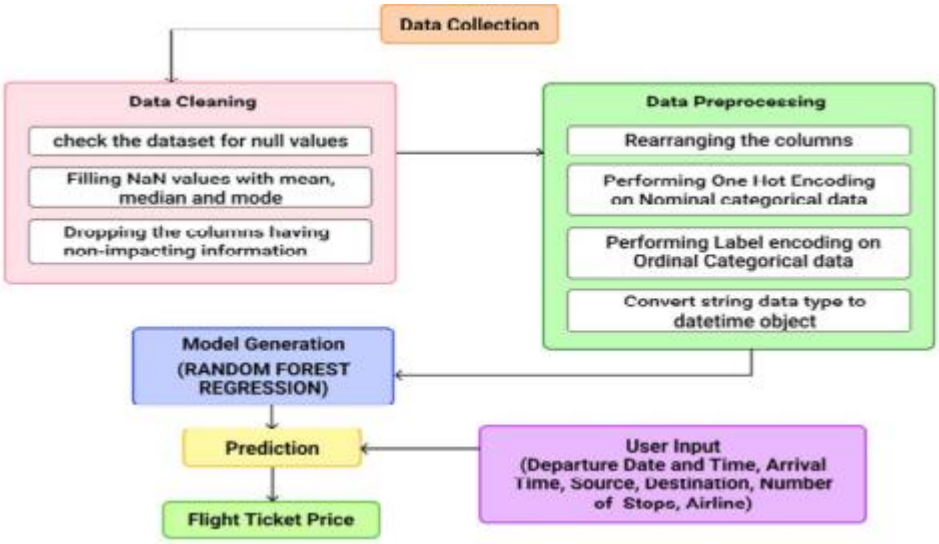
Further, the data was analysed and tests on the distribution were performed. Conclusions of the tests revealed that our data followed Log-Normal distribution and the same has been positively confirmed through statistical methods.

Based on previous history, the trend in the flight prices were modelled and the same was used to provide the user with an approximation of the number of days to wait from the current day, and if at all he waits, the amount he can say on the ticket.

In order to predict if the customer has to wait or not, we used a combination of statistical models and machine learning models. The statistical model provided with a probability corresponding to each airline having the least cost while the machine learning model further went ahead to predict the specific conditions taking into account the days to departure and the day of departure.

The machine learning algorithms implemented started off with basic Regression models and were extended to Decision Trees followed by Random Forests and Gradient Boosting methods. Later we developed an algorithm which had a combination of Rule based learning, Ensemble models and Statistical models to increase the accuracy.

Based on the prediction made by the model and the estimated time to wait, we calculated the savings we could achieve and the losses we incurred based on the predictions.



Data Preparation :

Data preparation was a critical part, as we had multiple airlines on a specific day and we had to predict the future prices for all those airlines, or the airline which would have the lowest fare.

Suppose a user makes a query to buy a flight ticket 44 days in advance, then our system should be able to tell the user whether he should wait for the prices to decrease or he should buy the tickets immediately. For this we have two options:

Predict the flight prices for all the days between 44 and 1 and check on which day the price is minimum.

Classify the data we already have into, “Buy” or “Wait”. This then becomes a classification problem and we would need to predict only a binary number. However, this does not give a good insight on the number of days to wait.

For the above example, if we choose the first method we would need to make a total of 44 predictions (i.e. run a machine learning algorithm 44 times) for a single query. This also cascades the error per prediction decreasing the accuracy. Hence, the second method seems to be a better way to predict, wait or buy which is a simple binary classification problem. But, in this method, we would need to predict the days to wait using the historic trends.

For this we again have two options:

We do the predictions for each flight id. The problem with this is that, if there is a change in flight id by the airline (which happens frequently) or there is an introduction or a new flight for a specific route then our analysis would fail.

We group the flight ids according to the airline and the time of departure and do the analysis on each group. For this we need to combine the prices of the airlines lying in that group such that the basic trend in captured.

Moving ahead with the second option, we created the group according to the airlines and the departure time-slot created earlier (Morning, Evening, Night) and calculated the combined flight prices for each group, day of departure and depart day. Since these three are the most influencing factors which determine the flight prices. Also, we calculated the average number of flights that operated in a particular group, since competition could also play a role in determining the fare.

	GroupID	Dept_Day	daystodep	Count	Total_meanFare	Total_minFare	Total_25Fare	Total_sdFare	Total_customFare	logical
1	Go Air_Night	Thursday	8	6	2605.000	2246	2350.50	298.2361	2319.150	0
2	Go Air_Night	Thursday	15	6	2591.000	2246	2350.50	282.9148	2319.150	0
3	Go Air_Night	Thursday	22	6	2591.000	2246	2350.50	282.9148	2319.150	0
4	Go Air_Night	Thursday	12	6	2582.833	2246	2351.00	273.3579	2319.500	0
5	Go Air_Night	Thursday	19	6	2543.000	2246	2351.00	231.1830	2319.500	0
6	Go Air_Night	Thursday	13	6	2544.000	2246	2351.75	231.8258	2320.025	0
7	Go Air_Night	Thursday	20	6	2544.000	2246	2351.75	231.8258	2320.025	0
8	Go Air_Night	Thursday	14	6	2545.000	2246	2352.50	232.4771	2320.550	0
9	Go Air_Night	Thursday	21	6	2545.000	2246	2352.50	232.4771	2320.550	0
10	Vistara_Evening	Monday	13	15	3159.533	2301	2301.00	654.0272	2375.700	0
11	Vistara_Evening	Monday	18	15	3071.933	2301	2301.00	589.2131	2375.700	0
12	Vistara_Evening	Monday	19	15	3071.933	2301	2301.00	589.2131	2375.700	0
13	Vistara_Evening	Monday	20	15	3071.933	2301	2301.00	589.2131	2375.700	0
14	Vistara_Evening	Monday	25	15	3013.533	2301	2301.00	533.2138	2375.700	0
15	Vistara_Evening	Thursday	16	15	3042.733	2301	2301.00	562.7238	2375.700	0
16	Vistara_Evening	Thursday	21	15	3013.533	2301	2301.00	533.2138	2375.700	0

Combining fare for the flights in one group:

Mean fare: This is the average of the fare of all the flights in a particular group corresponding to departure day and days to departure. Because of high standard deviation, taking the mean is not a very good option.

Minimum fare: This does not give a very good insight of the trend, as a minimum value could occur because of some offer by an airline.

First Quartile: This is a good measure as we are focusing on minimizing the fare and we do not want to consider the flights with high fares.

Custom Fare: This is the fare giving more weightage to recent price trend.

Total_customFare = w*(First Quartile for entire time period) + (1-w)*(First quartile of last x days) 5. (We have considered: w = 0.7 and x = & days)

Calculating whether to buy or wait for the this data:

Logical = 1 if for any d < D the Total_customFare is less than the current Total_customFare

(Here, d is the days to departure and D is the days to departure for the current row.)

Calculating the number of days to wait :

After creating the train file, we shift to create another dataset which is used to predict number of days to wait. For this, we used trend analysis on the original dataset.

Determining the minimum CustomFare for a particular pair of Departure Day and Days to Departure

We input the train dataset that has been created and find the minimum of the CustomFare corresponding to each combination of Departure Date and Days to Departure. Now with the obtained minimum CustomFare corresponding to each pair, we do a merge with our initial dataset and find out the Airline corresponding to which the minimum CustomFare is being obtained.

The count on the number of times a particular Airline appears corresponding to the minimum Custom Fare is the probability with which the Airline would be likely to offer a lower price in the future. This probability of each Airline for having a minimum Fare in the future is exported to the test dataset and merged with the same while the dataset of minimum Fares is retained for the preparation of bins to analyse the time to wait before the prices reduce

	daystodep	Dept_Day	Total_customFare	GroupID
1	1	Friday	4257.275	Go Air_Morning
2	2	Friday	4101.000	Go Air_Morning
3	3	Friday	4103.800	Go Air_Morning
4	4	Friday	4235.100	Spicejet_Morning
5	5	Friday	4166.100	Go Air_Morning
6	6	Friday	3850.225	Go Air_Morning
7	7	Friday	3773.450	Spicejet_Morning
8	8	Friday	3662.850	Spicejet_Morning
9	9	Friday	3605.100	Spicejet_Morning
10	9	Friday	3605.100	Spicejet_Night
11	10	Friday	3688.750	Spicejet_Morning

Creation of Bins :

We next wanted to determine the trend of “lowest” airline prices over the data we were training upon. So the entire sequence of 45 days to departure was divided into bins of 5 days. In intervals of 5 (this is made dynamic), the first bin would represent days 1-5, the second represents 6-10 and so on.

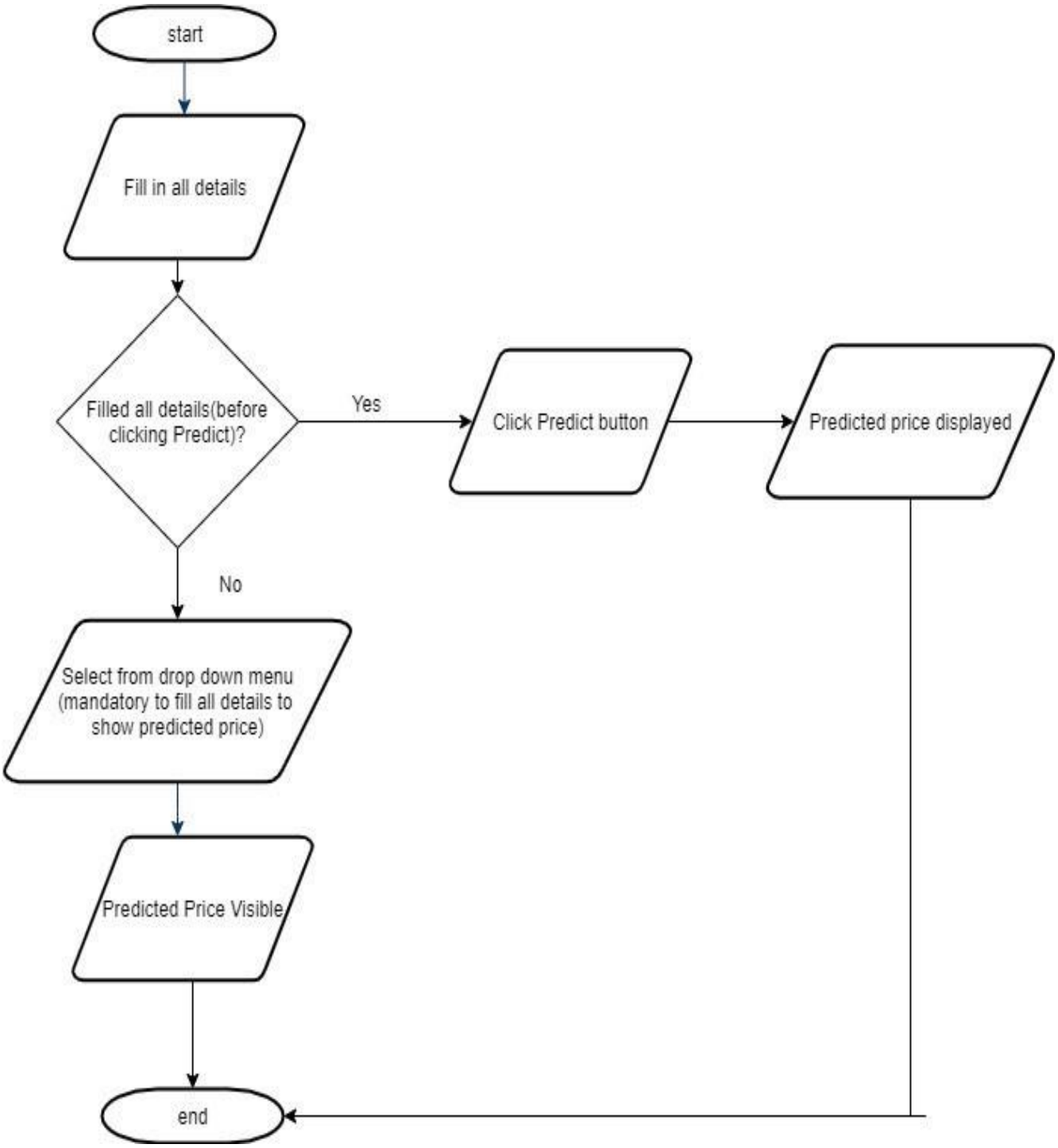
Corresponding to each bin, we required a value of the fare that would be optimal for consideration in suggesting a value for the days to wait to the user. Among all the points that lie in a bin, the 25th percentile was determined as the value that would be the possible lowest Fare corresponding to the bin which indicates days to departure.

Comparing the present price on the day the query was made with the prices of each of the bin, a suggestion is made corresponding to the maximum percentage of savings that can be done by waiting for that time period. The approximate time to wait for the prices to decrease and the corresponding savings that could be made is returned to the user.

	Min_wait	Max_wait	PriceDrop_percentage
2339	3	7	6.687536
2640	3	7	6.687536
2684	3	7	6.687536
2512	2	6	6.687536
2639	2	6	6.687536
2683	2	6	6.687536
2638	1	5	6.687536

Logical Diagram -

FLOW CHART :



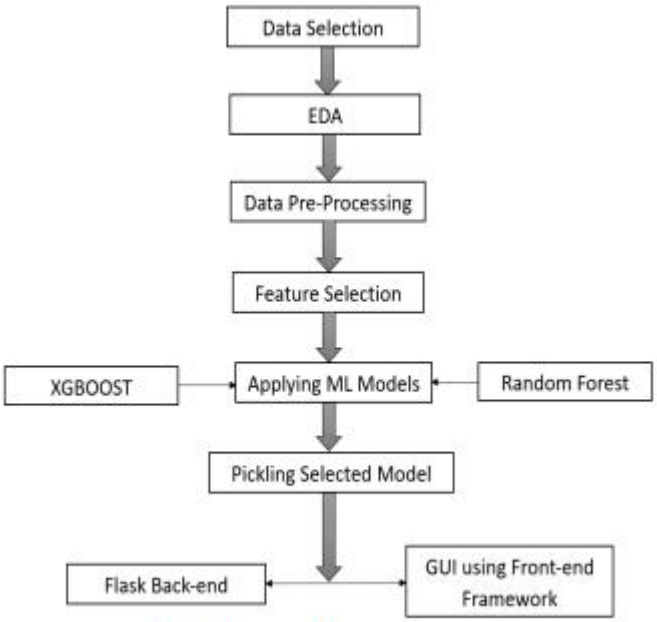


Fig. Proposed System Diagram

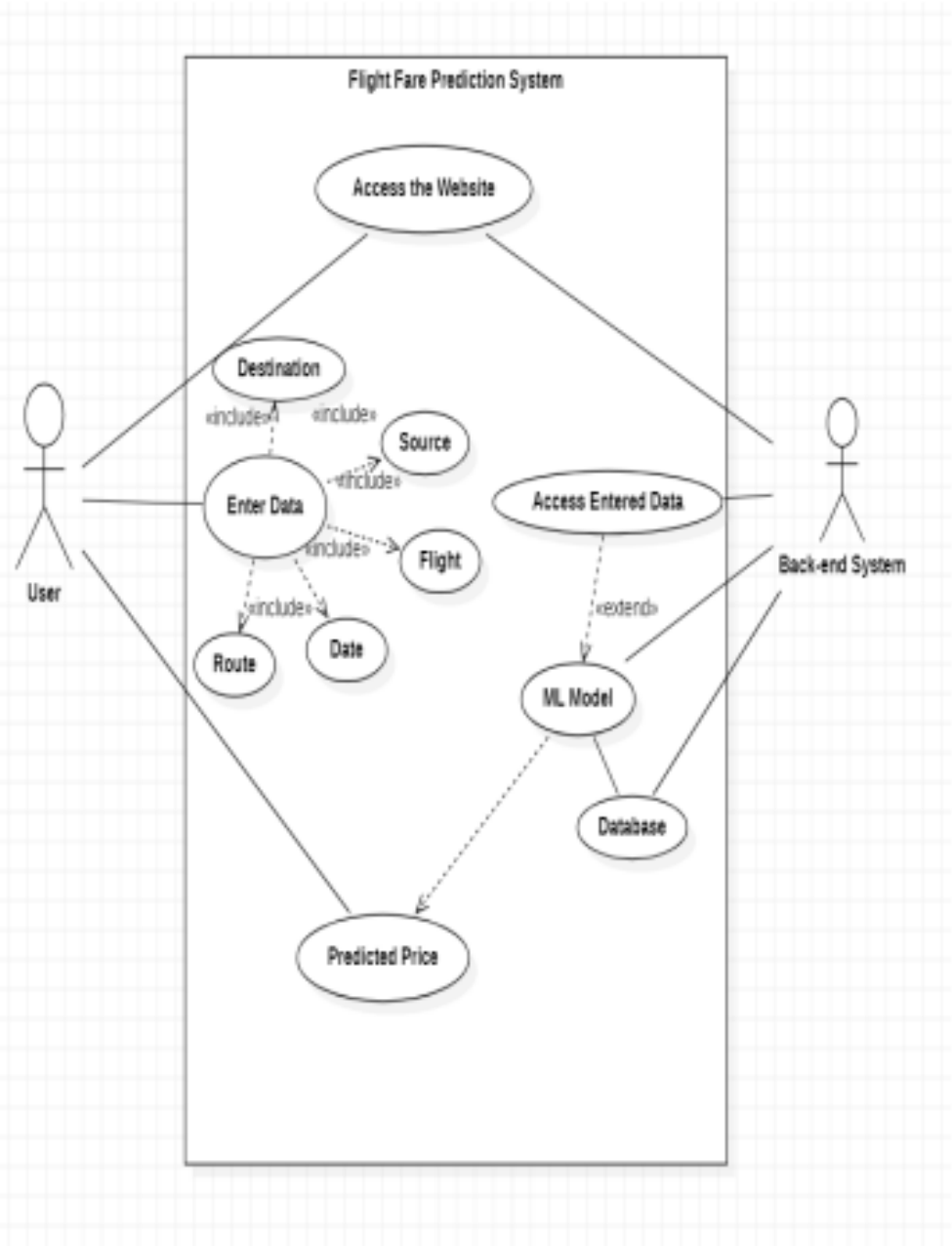


Fig. Use Case Diagram

DATA COLLECTION -

DATASET : The data consist of Flight data which consist of 11 features and 10689 samples with few missing values .

THE FEATURES ARE :

Airlines : display the different types of airlines like indigo, air india etc.

Date of journey :date on which you want to travel

Source : From where you want to take off

Destination : To where you want to go

Route : Route is basically stoppage between

6. Departure time : Take of time

Arrival time : display the arrival time

Duration: Time taken by flight to reach destination

Total stop : Total no of stops in between

Additional info : Indicate additional info like meal , not meal etc.

Price : Price of ticket

LIBRARY USED IN PYTHON

NumPy: -

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

Pandas: -

Pandas, which is used for data manipulation and data analysis .

Matplotlib : -

Matplotlib is a plotting library for python programming .It relies on pyplot to automatically create and manage the figures and axes, and use pyplot functions for plotting.

Seaborn : -

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive

Scikit learn : -

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction

Pickle :-

The Python pickle module is another way to serialize and deserialize objects in Python. It is used for saving the machine learning model in binary format.

Flask : -

Flask is a web framework, it's a Python module that lets you develop web applications easily

Flask_cors :-

Cross-Origin Resource Sharing (CORS) is an HTTP-header based mechanism that allows a server to indicate any other origins (domain, scheme, or port) than its own from which a browser should permit loading of resources

The important a part of the project is data collection. Data on different websites is gathered with unique attribute to provide the best accuracy. The data is collected from website kaggle.com and the models are implemented using python. The python-script helps to easily pre-process the data and forecast the output. The duplicate values are avoided in the pre-processing step. This dataset is more concentrated on calculating the plane price value. The dataset contains the data with attributes such as

- Journey_Date
- Departure
- Designation
- Arrival
- Airline
- Duration
- Source
- Price

1. Cleaning and preparing data

The gathered data must be cleaned and pre-processed and after improving the data, it is read to run on the algorithms. The duplicate values are removed, data is arranged with numerical values by pre-processing and by this model building and selecting the features becomes easier. Pre-processing plays the vital role for the whole dataset.

TABLE I: CLEAN AND PREPARED DATASET

	Total_Stops	Journey_day	Journey_month	Dep_hour	Dep_min	Arrival_hour	Arrival_min	Duration_hours	Duration_mins	Air India	GoAir	IndiGo	Jet Airways	Jet Airways Business	Multiple carriers	Multiple carriers Premium economy
0	1	6	6	17	30	4	25	10	55	0	0	0	1	0	0	0
1	1	12	5	6	20	10	20	4	0	0	0	1	0	0	0	0
2	1	21	5	19	15	19	0	23	45	0	0	0	1	0	0	0
3	1	21	5	8	0	21	0	13	0	0	0	0	0	0	1	0
4	0	24	6	23	55	2	45	2	50	0	0	0	0	0	0	0

Dataset in table I shows the information which is needed for the analyzing the data. Extra features is added to create best results. Feature like Dep_hour and Arrival_hour and Duration_hour is created to analyze the data for time duration of the day and other factors.

TABLE II: CLEAN AND PREPARED DATASET

Multiple carriers	Multiple carriers Premium economy	SpiceJet	Vistara	Vistara Premium economy	Chennai	Delhi	Kolkata	Mumbai	Cochin	Delhi	Hyderabad	Kolkata	New Delhi
0	0	0	0	0	0	1	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	1	0	0	0	0
1	0	0	0	0	0	1	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0

2. Analysing Data

Constructing of the data is the huge task, by knowing the various patterns of data visualization and later using the required machine learning models. Also, from the current attribute the new small elements can be acquired. If it is on holiday, festival day or a weekday or weekend, plane date plays main role. Travelling during weekends is costlier than the planes on weekdays and time is considered in classes as: Morning, afternoon, evening and night, so time plays important role. Travelling days is computed with plane date and the date on which data is collected.

MACHINE LEARNING MODELS

Flight price forecasting using various algorithms in machine learning. The algorithms for forecasting purpose are:

Support Vector Machine, Linear regression, K-Nearest neighbors, Multilayer Perceptron, Gradient Boosting and Random Forest Algorithm, Decision tree. Traversing the python library and parameters like R-square, MAE and MSE area unit to verify the production of those models.

1. Linear regression

Variable quantity of that price is to be found for this we are employing statistical regression analysis such as correlation between 2 continuous variables, from the 2 variables. The equation for statistical regression is:

(pred) = b0+b1 x (1)

The two major factors to grasp statistical regression is gradient descent and price operate area unit. It gives the simplest match line to the given data that the forecast error is minimum and provides the applied mathematics relationship not the settled relationship between 2 variables. The sq. of expected and actual price distinction gives the error. To alter the negative values, the mean sq. error is taken (MSE). Value of the coefficients b1 and b0 area unit chosen in order that the error value is as little as doable.

Choosing a random data from a dataset with replacement is called Bootstrap aggregating. By gradient boosting and random forest strategies achieves the greater accuracy.

2. Decision Tree

It is used to make any decision and have multiple branches which are the Decision Node and Leaf Node. Decision Tree used for both classification and Regression problems, but mostly it is preferred for solving Classification problems which is a Supervised learning technique. Decision tree has the two nodes, represents the features of a dataset, each leaf node shows the outcome is Internal nodes and branches shows the decision rules. Based on features of the given dataset the test must be performed. For getting all the possible solutions to a problem on the given condition's visualization is done. It starts with the root node that expands on branches and builds structure as tree is called as decision tree. CART algorithm helps in Classification and Regression Tree algorithm, that is used to build the tree. The two essential properties for tree computation is Gini index and data Gain. Lot of successful of the substance tells that it has Higher entropy. The Decision Tree gives the best accuracy with 80% contrast to random forest algorithm.

3. Random Forest

One of the popular machine learning algorithms which belongs to the supervised learning technique is Random Forest. Process of combining multiple classifiers to solve complicated problem and increasing the performance of the model it is based on the concept of learning. To avoid the overfitting problem and the greater number of trees in the forest tends to greater accuracy. Random forest is used for both classification and regression problems is the huge advantage. The resultant accuracy the random forest gives are 70% as shown in the result and analysis graph and table.

STATISTICAL ANALYSIS

1. Chi-Square Test

Relationship between two categorical variables is used to do statistical test that is Chi-Square. One variable having the frequency compared against the second variable's categories is done by executing Chi-Square. That defines the data is shown as a frequency table, rows show the independent variables and columns shows the dependent variables.

2. Correlation Test

The correlation or bivariate relationship between two independent variables, so correlation plot is used. Identifying the correlation of one independent variable with a group of other variables, VIF is used. So, VIF is used for best understanding. When VIF is equal to 1, it is No Correlation. VIF is equal to 1 to 5, it is Moderate Correlation. VIF is greater than 10, it is called Highly Correlated.

3. Anova Test

To compare two or more contrasting together to determine the analysis of variance Analysis of Variance statistical test is performed. Analyzing the differences between groups and signifying the difference statistically is the one-way ANOVA tests. The other way to compare two or more independent group which is using when at least three independent groups are available.

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Python is a high-level object-oriented scripting language, designed to be readable it uses English keywords more and uses indentation, whereas other languages use punctuation. Functions gives best modularity for our application and a high amount of reusability of the code. In python classes and objects are easily used.

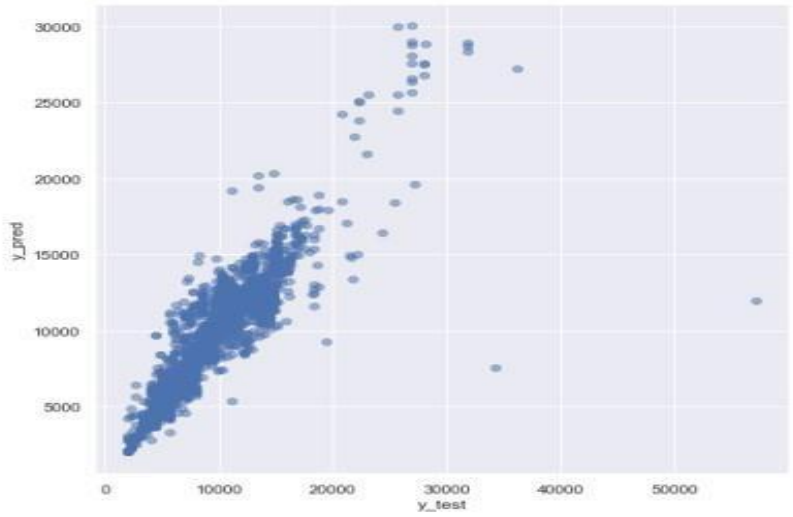
Many built-in functions like print(), etc. We can also create your own functions in python, so these functions is called user-defined functions. Python libraries for data analysis by making use of Numpy, Pandas and Scipy for the selected dataset. Opensource library pandas is used to manipulate, analyze, load, and visualize the

selected datasets. The other open-source library scikit-learn builds smart models and make cool predictions and is used in machine learning algorithms.

Output of the model is visualized graphically across the test dataset for chosen test dataset. The visualization represents study of real value, also the prediction of results. Decision Tree, Bagging Tree, Random Forest and Linear regression tells the results gained by the analysis. Also, gives the attribute price to purchase the flight ticket at the right time for predicting the price values. Decision Tree algorithm has more accuracy compared to other algorithms for the given dataset. It gives the highest R-Square value with maximum accuracy in the regression analysis. First column gives the values for R Square, were table III shows R-square, MSE and MAE values.

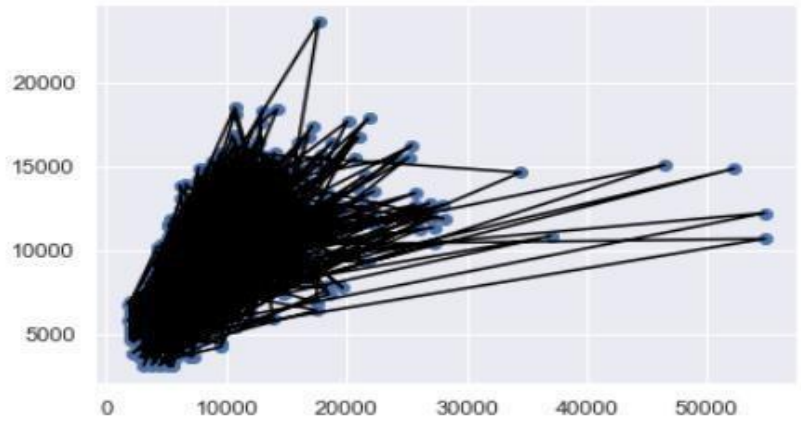
TABLE III: ALGORITHM EVALUATION

Machine Learming(ML)	R-squared	MAE	MSE
Random Forest	0.79	1166.1987291481917	4054043.514563705
Decision Tree	0.80	1166.1987291481917	4054043.514563705



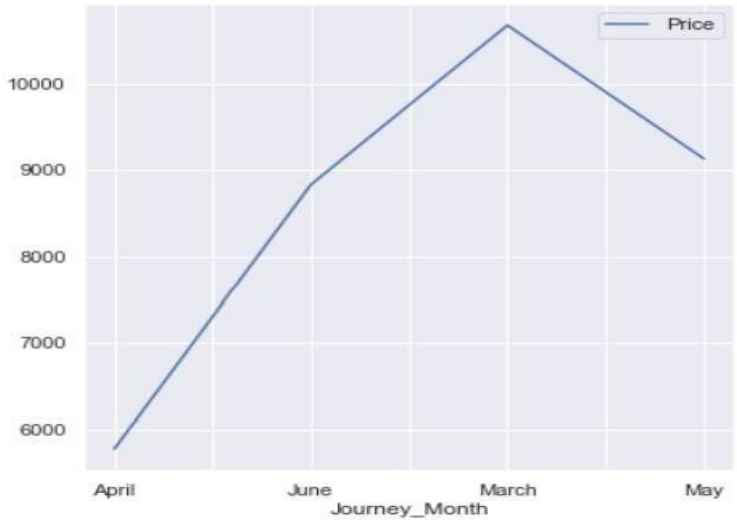
Graphical results for Flight price prediction using Random Forest.

Random Forest, graph is plotted with Y-test data, it shows the price value is getting higher. Here the observation is according to the prediction the passengers who have purchased the ticket with highest amount of price is around thirty thousand and most of the people have purchased the tickets between five thousand to fifteen thousand. Through this analysis the advantage is that people will have more idea about the frequent price offers in festival and holiday season and will choose the best price for travelling.



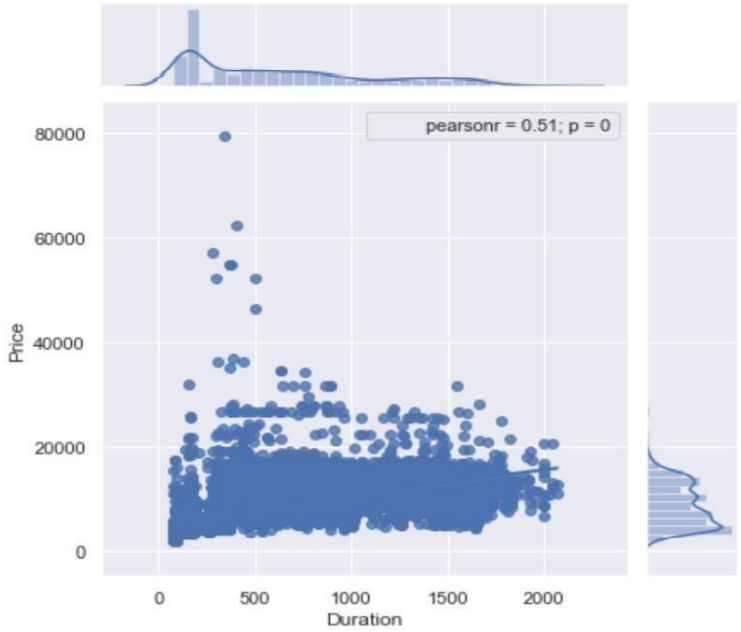
Graphical results for Flight Price Prediction using Decision Tree.

Decision Tree tells with Y_Test set again the price increases as the time varies. In this graph also we observe passengers gave purchased the ticket from five thousand to fifteen thousand and highest ticket price purchased is twenty-five thousand. So, we can see that the accuracy level is low here itself. Through this analysis the advantage is that passengers can ask for the review to their friends or relatives who travel mostly, this helps to take better decision for travelers.



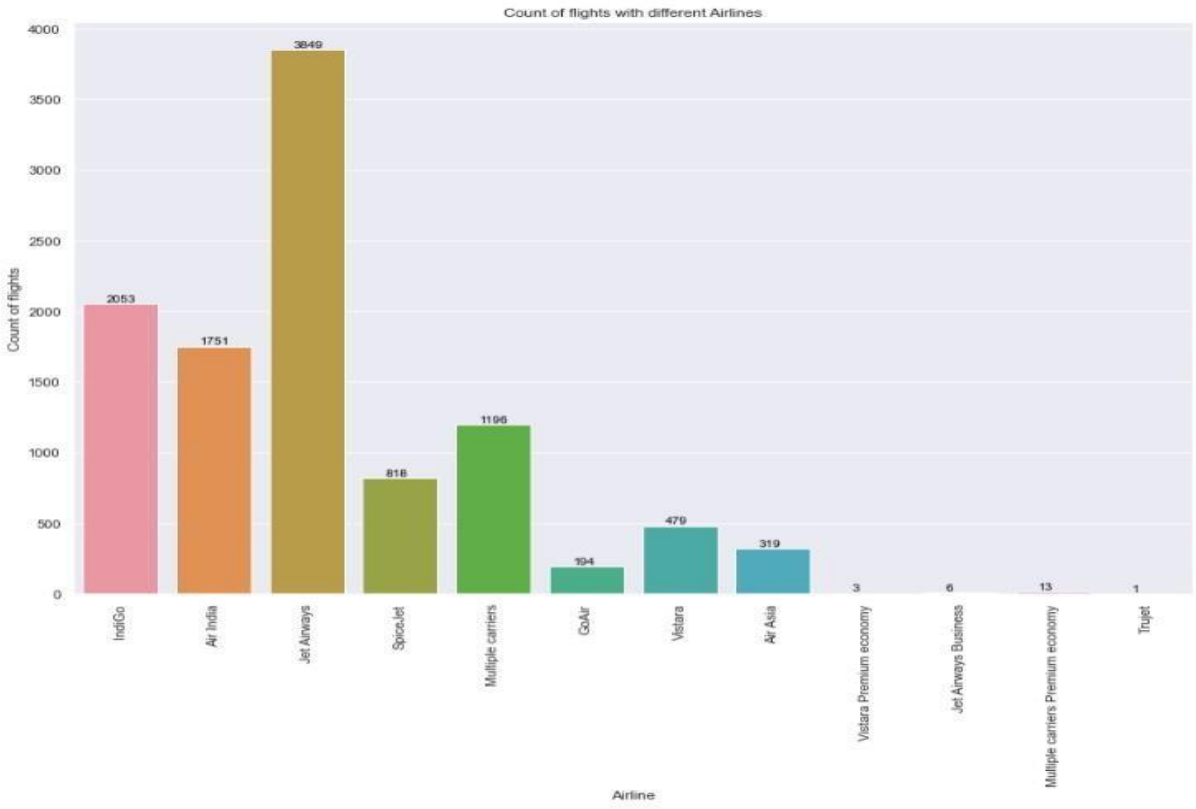
Graphical results of Analysis between Journey month and Price.

considering the features Price and Journey Month we see that the prices are higher at the month of the march as people travel more the company increases the expenses. People should make their own strategies to when to travel and make use of the best offer from the airlines. Since airline company increase the price for the business purpose, people should be also smart to travel with best price. Through this analysis the advantage is that passengers can see which month is best to travel with affordable prices because season to season price changes. People are smart to choose the cheapest price and comfort for travelling.



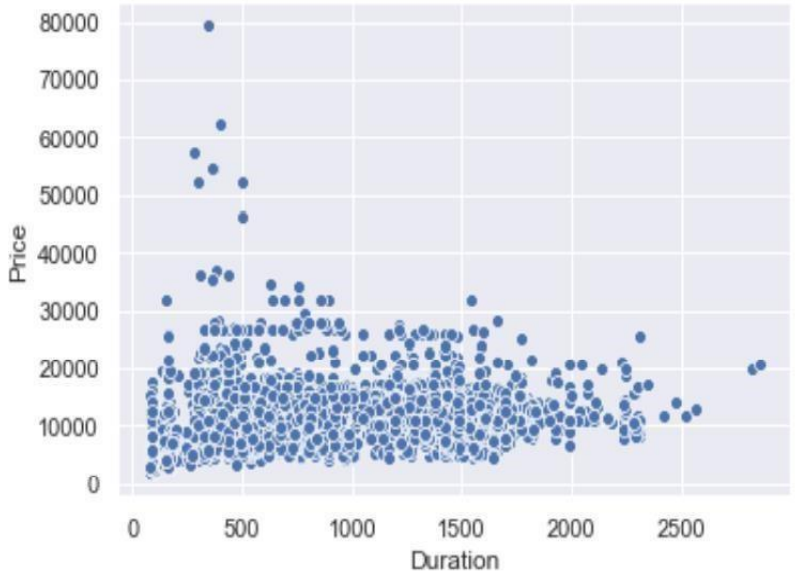
The Statistical Analysis between Duration and Price as correlation test.

correlation test gets p-value < 0.05, hence we accept H1 and say the target variable and continuous independent variable are correlated. r = 0.51 says they are moderately related. Through this analysis the advantage is that the time duration plays the important role for making the decision to board the flight with best price. With the limited amount of time the best price can be chosen by the passengers. Everybody can afford the flight ticket with best price and best offers.



Jet airways and Air India are full-service airlines are and always highly priced due to various amenities they provide. Low-cost carriers like indigo and SpiceJet have a lower and similar fare. Through this analysis the advantage is that travellers can see the highest to lowest prices to know the price differences in each airline. Passengers

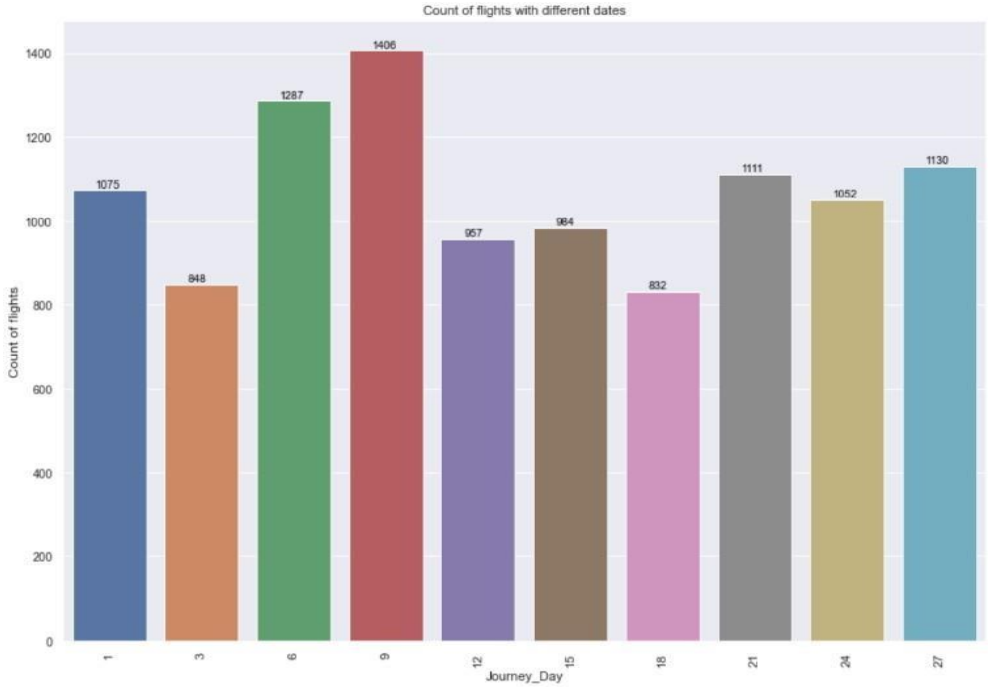
can share the review of the flight on the each airline websites, so that the other passengers gets the benefit out of it.



The Analysis between Duration and Price

We know that duration (or distance) plays a major role in affecting air ticket prices, but we see no such pattern here, as there must be pattern here ands other

significant factors affecting air fare like type of airline, destination of flight, date of journey of flight (higher if collides with a public holiday). Through this analysis the advantage is that passengers might want to reach the destination sooner, so the duration plays the important role to reach sooner with good amount of price.



The Analysis between Date and Price

It looks like that there's a trend in the plane price when contrasted to the day of respective months, prices are higher in the start of month, but this is not a trend if you see from the broader perspective as this might be due to various reasons. For e.g. the date of Journey is 10th March and people are booking towards 5th March or so, this will lead to higher flight prices. Prices increase as near you date of booking is to the date of journey. So, flight prices don't follow any pattern towards any time of the month. Through this analysis the advantage is that the passengers can know the which date and day they are travelling by this people can plan accordingly and book the tickets, which makes the passengers very convenient and organised without confusion.

CONCLUSION

Random Forest regression is used to predict the price of flight for the different airlines like indigo, vistara etc. The main purpose of this project is nowadays everyone wants to know what will be my price if I travelled from here to there so in order to solve this problem this kind of end user web application used to predict what would be the price of a flight ticket if I will go from here to there.

Evaluating the algorithmic rule, a dataset is collected, pre-processed, performed data modelling and studied a value difference for the number of restricted days by the passengers for travelling. Machine Learning algorithms with square measure for forecasting the accurate fare of airlines and it gives accurate value of plane price ticket at limited and highest value. Information is collected from Kaggle websites that sell the flight tickets therefore restricting data which are often accessed. The results obtained by the random forest and decision tree algorithm has better accuracy, but best accuracy is predicted by decision tree algorithm as shown is the above analysis. Accuracy of the model is also forecasted by the R-squared value.

In Upcoming days when huge amount of information is accessed as in detailed information in the dataset, the expected results in future are highly correct. For further research anyone desire to expand upon it ought to request different sources of historical data or be a lot of organized in collection knowledge manually over amount of your time to boot, a lot of different combination of plane are going to be traversed. There is whole possibility that planes differ their execution ideas consisting characteristics of the plane. At last, it is curious to match our model accuracy with that of the business models accuracy offered nowadays.

REFERENCES

https://www.wikipedia.org/

https://www.google.com

https://www.youtube.com/watch?v=y4EMEpEnElQ

https://towardsdatascience.com/deploy-your-python-machine-learningmodelson-heroku-in-3-steps-dc5b6aca73d9

https://www.analyticsvidhya.com/blog/2021/06/flight-price-predictionaregression-analysis-using-lazy-prediction/

https://www.analyticsvidhya.com/blog/2017/09/common-machinelearningalgorithms/

https://www.youtube.com/watch?v=7eh4d6sabA0