MINI PROJECT

(2021-22)

"FLIGHT-PRICE-PREDICTION"

Project Report

Institute of Engineering & Technology

Submitted By –

Shaurya Gupta (191500751)

Satyam Tiwari (191500733)

Pranav Pandey (191500566)

Under the Supervision Of

Abhishek Tiwari

Trainer

Training and Development

Department of Computer Engineering and Applications

GLA University, 17 km. Stone NH#2, Mathura-Delhi Road,

Chaumuha, Mathura – 281406 U.P (India)

Declaration

I/we hereby declare that the work which is being presented in the Bachelor of technology. Project **"FLIGHT-PRICE-PREDICTION"**, in partial fulfillment of the requirements for the award of the ***Bachelor of Technology*** in Computer Science and Engineering and submitted to the Department of Computer Engineering and Applications of GLA University, Mathura, is an authentic record of my/our own work carried under the supervision of **Mr. Abhishek Tiwari, Trainer, Training and Development,GLA University.**

The contents of this project report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree.

Sign: *SatyamTiwari*

Name of Candidate: Satyam Tiwari

University Roll No.:191500733

Sign: *ShauryaGupta*

Name of Candidate: Shaurya Gupta

University Roll No.:191500751

Sign: *PranavPandey*

Name of Candidate: Pranav Pandey

University Roll No.:191500566

Department of Computer Engineering and Applications
GLA University, 17 km. Stone NH#2, Mathura-Delhi Road,
Chaumuhan, Mathura – 281406 U.P (India)

Certificate

This is to certify that the project entitled "FLIGHT-PRICE-PREDICTION", carried out in Mini Project – II Lab, is a bonafide work by Satyam Tiwari, Shaurya Gupta, Pranav Pandey, and is submitted in partial fulfillment of the requirements for the award of the degree Bachelor of Technology(Computer Science & Engineering).

Signature of Supervisor:

Name of Supervisor: Mr. Abhishek Tiwari

Date

Department of Computer Engineering and Applications

GLA University, 17 km. Stone NH#2, Mathura-Delhi Road,

Chaumuhan, Mathura – 281406 U.P (India)

ACKNOWLEDGEMENT

Presenting the ascribed project paper report in this very simple and official form, we would like to place my deep gratitude to GLA University for providing us the instructor Mr. Abhishek Tiwari, our technical trainer and supervisor.

He has been helping us since Day 1 in this project. He provided us with the roadmap, the basic guidelines explaining on how to work on the project. He has been conducting regular meeting to check the progress of the project and providing us with the resourcesrelated to the project. Without his help, we wouldn’t have been able to complete this project.

And at last but not the least we would like to thank our dear parents for helping us to grab this opportunity to get trained and also my colleagues who helped me find resources during the training.

Thanking You

Sign: *SatyamTiwari*

Name of Candidate: Satyam Tiwari

University Roll No.:191500733

Sign: *ShauryaGupta*

Name of Candidate: Shaurya Gupta

University Roll No.:191500751

Sign: *PranavPandey*

Name of Candidate: Pranav Pandey

University Roll No.:191500566

ABSTRACT

Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using historical data and also to suggest the best time to buy a flight ticket.

For this project, we have collected data from 18 routes across India while the data of 4 routes were extensively used for the analysis due to the sheer volume of data collected over 4 months resulting in 5.28 lakh data points each across the Mumbai-Delhi and Delhi-Mumbai route and 1.05 lakh data points each across the Delhi-Banglore and Banglore-Delhi route. The project implements the validations or contradictions towards myths regarding the airline industry, a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done so. A customized model which included a combination of ensemble and statistical models have been implemented with a best accuracy of above 90% for a few routes, mostly from Tier 2 to metro cities. These models have led to significant savings and produced average positive savings on each transaction.

Remarkably, the trends of the prices are highly sensitive to the route, month of departure, day of departure, time of departure, whether the day of departure is a holiday and airline carrier. Highly competitive routes like most business routes (tier 1 to tier 1 cities like Mumbai-Delhi) had a non-decreasing trend where prices increased as days to departure decreased, however other routes (tier 1 to tier 2 cities like Delhi - Cochin) had a specific time frame where the prices are minimum. Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum.

With a high probability (about 20-25%) that a person has to wait to buy a ticket, the scope of the project can be extensively extended across the various routes to make significant savings on the purchase of flight prices across the Indian Domestic Airline market.

INTRODUCTION

CONTEXT

This Web Application “FLIGHT-PRICE-PREDICTION” has been submitted in partial fulfillmentof the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering at GLA University, Mathura supervised by Mr.Abhishek Tiwari. This project has been completed approximately three months and has been executed in modules, meetings have been organized to check the progress of the work and for instructions and guidelines.

MOTIVATION

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:

1. Time of purchase patterns (making sure last-minute purchases are expensive)
2. 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 travellers 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.

OBJECTIVE

The objectives of the project can broadly be laid down by the following questions –

1. **Flight Trends** Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time?
2. **Best Time To Buy** What is the best time to buy so that the consumer can save the most by taking the least risk? So should a passenger wait to buy his ticket, or should he buy as early as possible?
3**. Verifying Myths** Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

ABOUT THE PROJECT

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 –

1. Time of purchase patterns (making sure last-minute purchases are expensive)
2. 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 travellers 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 travellers.

❏ Automated Script to Collect Historical Data For any prediction/classification problem, we need historical data to work with. In this project, past flight prices for each route needs to be collected on a daily basis. Manually collecting data daily is not efficient and thus a python script was run on a remote server which collected prices daily at specific time.

❏ Cleaning & Preparing Data After we have the data, we need to clean & prepare the data according to the model's requirements. In any machine learning problem, this is the step that is the most important and the most time consuming. We used various statistical techniques & logics and implemented them using built-in R packages.

❏ Analysing & Building Models Data preparation is followed by analysing the data, uncovering hidden trends and then applying various predictive & classification models on the training set. These included Random Forest, Logistic Regression, Gradient Boosting and combination of these models to increase the accuracy. Further statistical models and trend analyzer model have been built to increase the accuracy of the ML algorithms for this task.

❏ Merging Models & Accuracy Calculation Having built various models, we have to test the models on our testing set and calculate the savings or loss done on each query put by the user. A statistic of the over Savings, Loss and the mean saving per transaction are the measures used to calculate the Accuracy of the model implemented.

HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Requirement

- Processor : Intel i3
- Operating System : Windows 7/8/10
- RAM : 4 GB (or higher)
- Hard disk : 64GB

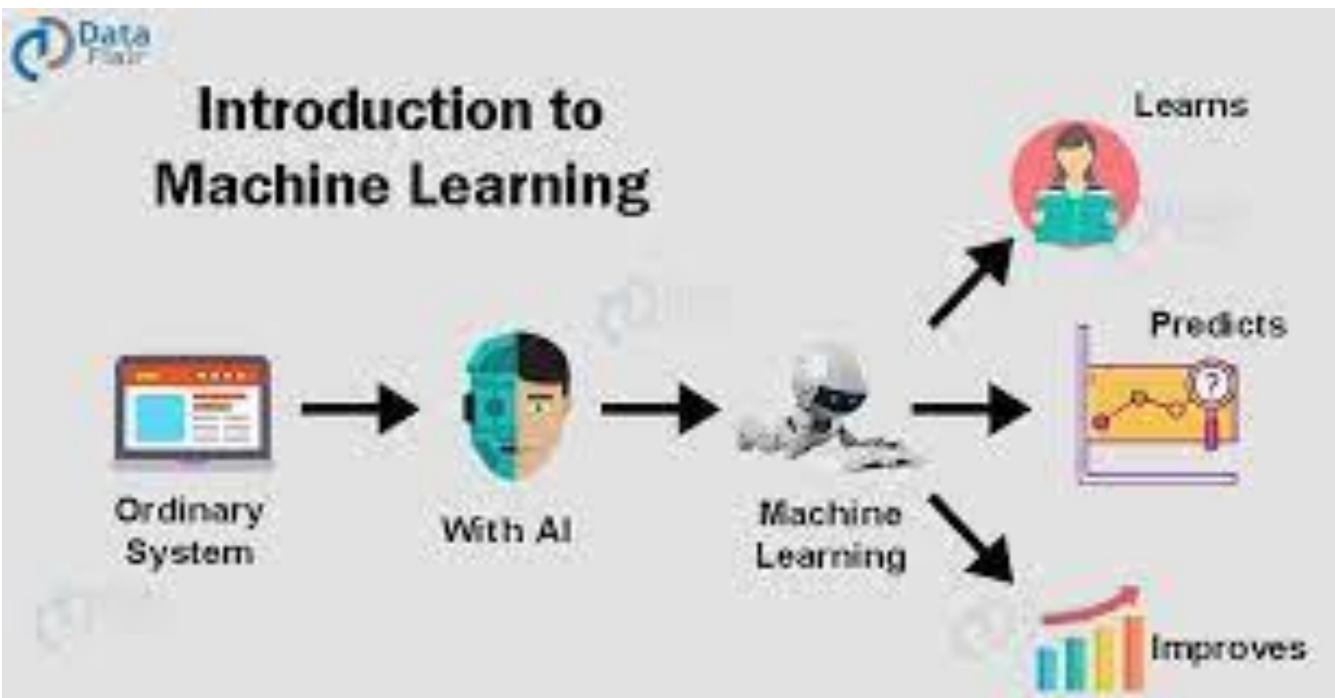
Software Requirement

- Software used: VS code, Anaconda, Jupyter, spyder
- Language used : HTML, CSS, python
- Brower : Google Chrome

TECHNOLOGY USED

MACHINE LEARNING

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.



TYPES OF MACHINE LEARNING

Machine Learning can be classified into two ways:

Supervised Learning

Unsupervised Learning

Supervised Learning: Supervised learning is one of the most basic types of machine learning. In this type, the machine learning algorithm is trained on labeled data. Even though the data needs to be labeled accurately for this method to work, supervised learning is extremely powerful when used in the right circumstances.

Unsupervised Learning: Unsupervised machine learning holds the advantage of being able to work with unlabeled data. This means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program.

Libraries:

Pandas

Numpy

Matplotlib

Sklearn

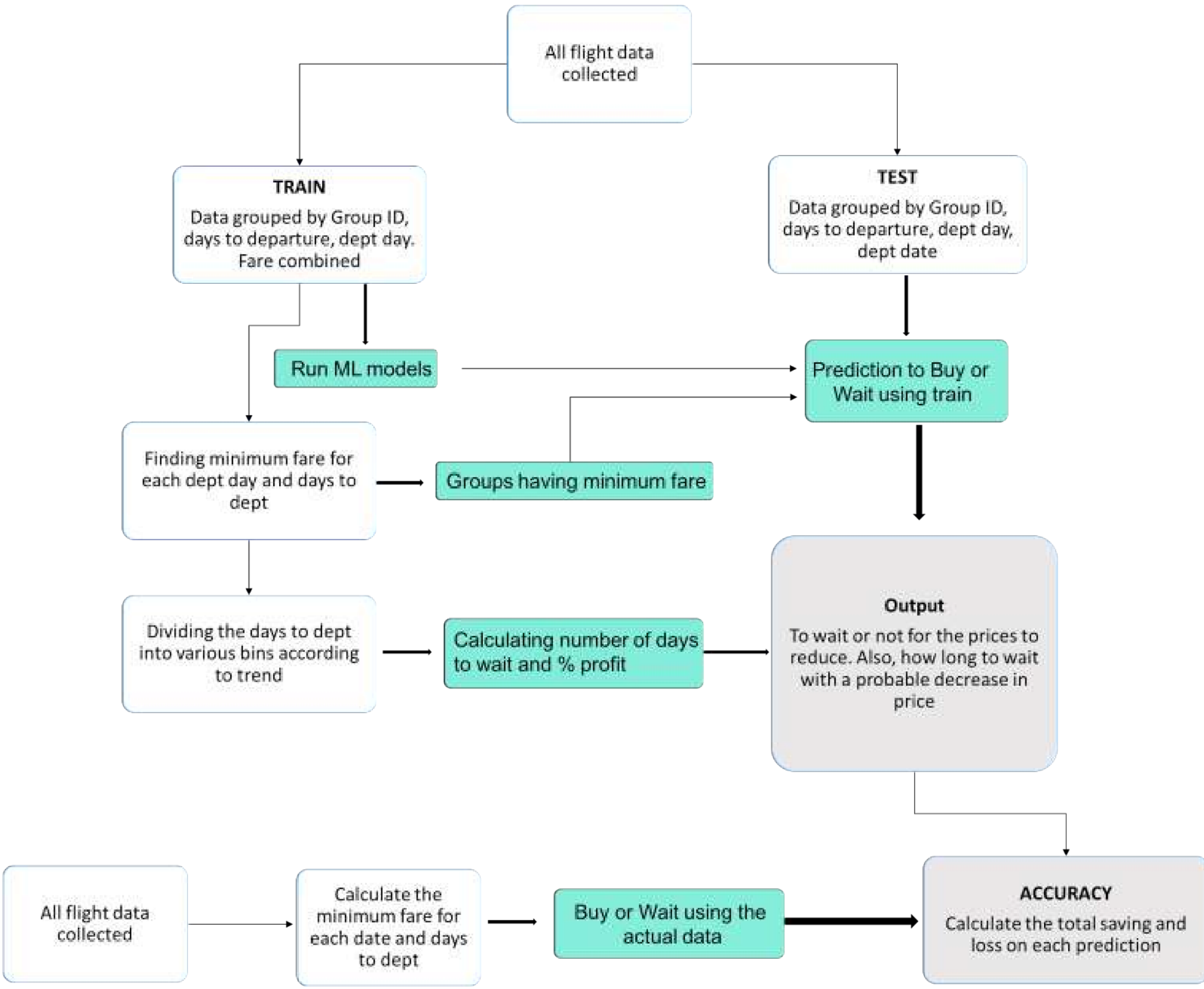
Seaborn

Pickle

Flask

.

Backend Development: In backend We use python to develop our machine learning model and different no. of libraries



TOOLS AND LANGUAGES

Tools used to build the web development are:-

Visual Studio Code: Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting intelligent code completion, snippets, code refactoring, and embedded Git. Users can change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality.

Visual Studio Code was first announced on April 29, 2015, by Microsoft at the2015 Build conference. A preview build was released shortly thereafter.

On November 18, 2015, the source of Visual Studio Code was released under the MIT License, and made available on GitHub. Extension support was also announced. On April 14, 2016, Visual Studio Code graduated from the public preview stage and was released to the Web. Microsoft has released most of Visual Studio Code's source code on GitHub under the permissive MIT License, while the releases by Microsoft are proprietary freeware.

In the Stack Overflow 2021 Developer Survey, Visual Studio Code was ranked the most popular developer environment tool, with 70% of 82,000 respondents reporting that they use it.

HTML: HTML stands for Hypertext Markup Language. It is used to design the front end portion of web pages using markup language. It acts as a skeleton for a website since it is used to make the structure of a website.

. **CSS:** Cascading Style Sheets fondly referred to as CSS is a simply designed language intended to simplify the process of making web pages presentable. It is used to style our website.

Anaconda: Anaconda offers the easiest way to perform Python/R data science and machine learning on a single machine. Start working with thousands of open-source packages and libraries today.

Jupyter: Jupyter is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

Spyder: Spyder is a free and open source scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It features a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package

IMPLEMENTATION

AND

USER INTERFACE

- 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

- Data Cleaning

The data was further processed based on the parameters mentioned below and cleaned based on appropriate considerations -

1. Days to Departure
2. Day of Departure
3. Duration
4. Hopping
5. Holiday
6. Outliers

7. 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.
8. 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.

9. 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

- 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:

1. Predict the flight prices for all the days between 44 and 1 and check on which day the price is minimum.
2. 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 thedays to wait using the historic trends.

For this we again have two options:

1. 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.
2. 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

	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

flights that operated in a particular group, since competition could also play a role in determining the fare.

Combining fare for the flights in one group:

1. 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.
2. 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.
3. 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.
4. 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.)

- ❏ <u>Calculating the number of days to wait</u>

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 outthe 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 andmerged 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

Results

- In detailed analysis for the Delhi - Guwahati Route

The trends in the data collected for the sector of Delhi to Guwahati busted some of the very famous myths assumed by travellers of the aviation industry.

1. Flight prices do not increase continuously as the Date of Departure approaches closer.

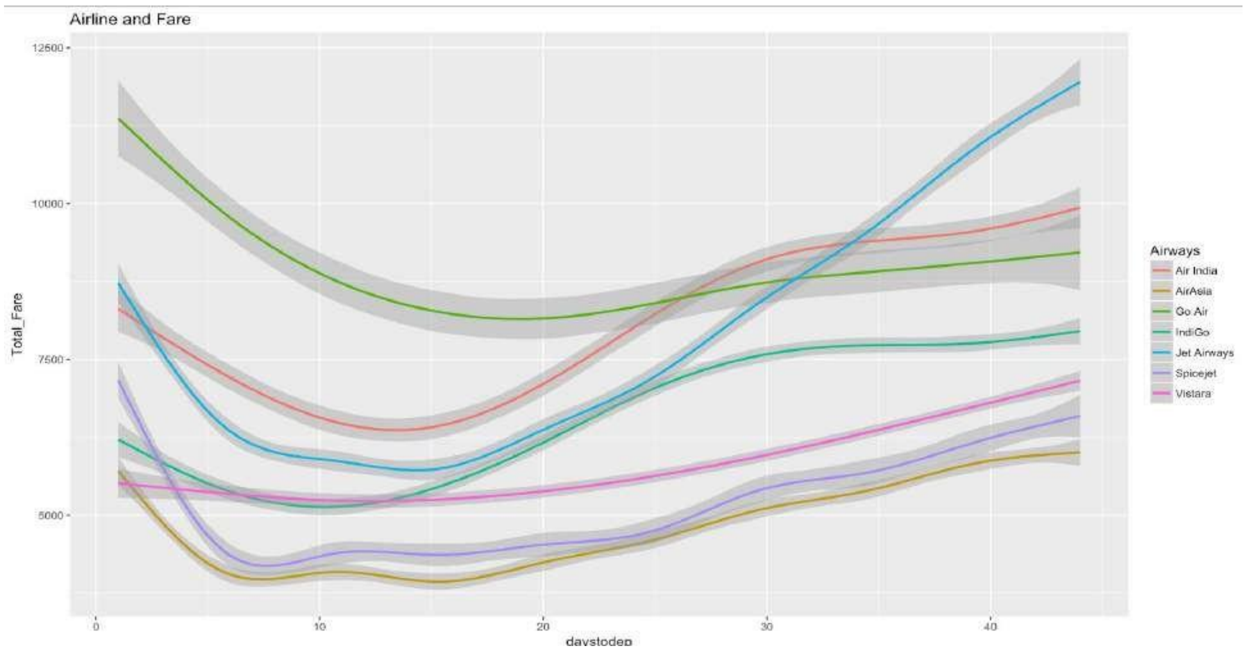
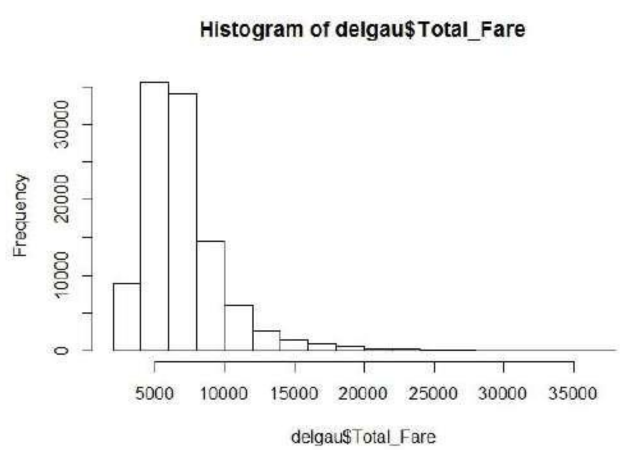


Figure 2 : Flight Price vs Days to Departure

With the validation of the problem statement and with a scope to predict when to buy and when to wait, we begin the analysis of the dataset.

The dataset of the flight prices follows a Lognormal distribution with some outliers which have been ignored as we are only interested with the minimum fare corresponding to a certain route.



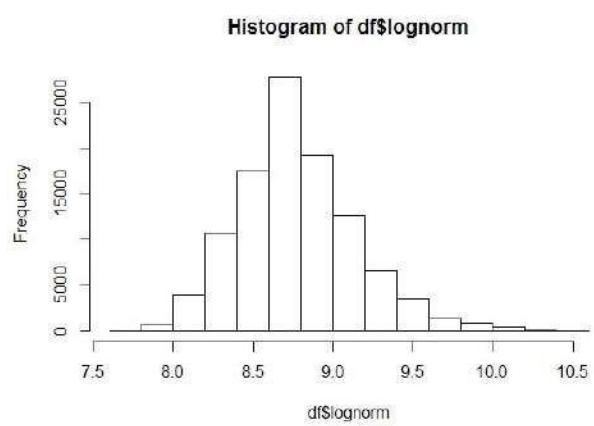


Figure 3 : Distributions before and after transformation

Statistically, the data transformed into lognormal distribution showed a significance level of 1, with skewness and kurtosis falling within the acceptable range for it to be considered a valid transformation.

Further, the trend of all airlines have been customly combined to form a trend used in the prediction of the model. The trend is significantly different for each day and thus different combined trends have been formulated corresponding to the day of the week.

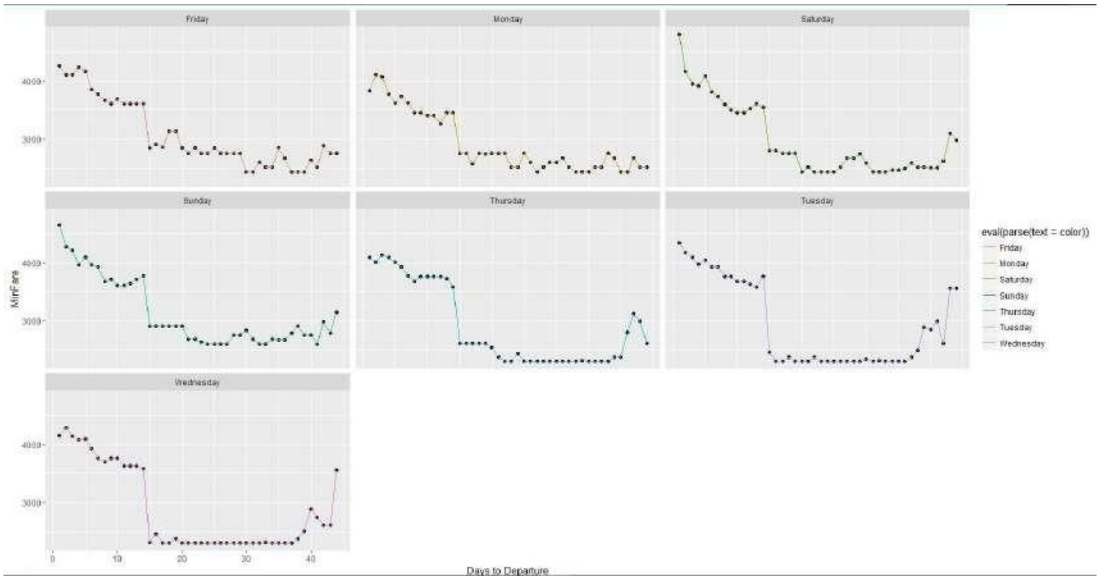


Figure 4 : Combined trends of all airlines for each day of the Week

We performed the prediction using some basic machine learning algorithms to find a benchmark model and the results of the same are shown below for the route of Delhi - Guwahati.

Model	Savings (In Lakhs)	Loss (In Lakhs)	Profit per Transaction (In Rs.)	Accuracy
Decision Trees	4.7	1.3	140	73.0%
Gradient Boosting	5.5	2.2	145	73.0%
Logistic regression	6	1.8	177	76.0%
Random Forest	5.8	1.8	180	77.8%
Trend Based Model	7	2.2	210	81.8%

Figure 5 : Comparison between Models

- Results for all Routes

In continuation, we developed a custom algorithm for the very specific task which was an amalgamation of the ensemble models and the statistical model as discussed above.

Study of the Savings, Loss and Average Savings per transaction on the Test Data Set							
Route	Route Type	Profit	Loss	No. of Times Predicted to Wait	Mean Savings	Percentage of Times Predicted to Wait	Percentage of Correctly Predicting to Wait
Kol - Blr	Business	₹ 11,601.00	-₹ 21,339.00	102	-₹ 29.96	21.90%	28.40%
Kol - Bom	Business	₹ 578.00	-₹ 6,346.00	11	-₹ 17.75	2.37%	9.09%
Del - Bom	Business	₹ 92,459.00	-₹ 1,05,723.00	319	-₹ 4.70	9.45%	44.50%
Hyd - Amd	Business	₹ 12,243.00	-₹ 11,449.00	32	₹ 2.44	6.88%	53.10%
Blr - Kol	Business	₹ 9,046.00	-₹ 8,046.00	22	₹ 3.19	4.87%	68.20%
Bom - Kol	Business	₹ 5,563.00	-₹ 3,511.00	19	₹ 6.31	4.09%	31.60%
Bom - Del	Business	₹ 2,24,013.00	-₹ 1,48,350.00	1525	₹ 26.46	44.60%	39.70%
Amd - Hyd	Business	₹ 26,581.00	-₹ 16,441.00	55	₹ 28.89	11.10%	67.30%

Figure 6 : Analysis on Various Business Routes

Route	Route Type	Profit	Loss	No. of Times Predicted to Wait	Mean Savings	Percentage of Times Predicted to Wait	Percentage of Correctly Predicting to Wait
Del - Sri	Tourist	₹ 1,615.00	-₹ 3,869.00	118	-₹ 7.02	25.90%	9.32%
Goi - Bom	Tourist	₹ 20,375.00	-₹ 8,439.00	49	₹ 36.73	10.50%	73.50%
Sri - Del	Tourist	₹ 24,933.00	-₹ 10,404.00	42	₹ 44.70	9.11%	78.60%
Bom - Gou	Tourist	₹ 17,423.00	-₹ 2,814.00	24	₹ 44.95	5.16%	79.20%

Figure 7 : Analysis on Various Tourist Routes

Route	Route Type	Profit	Loss	No. of Times Predicted to Wait	Mean Savings	Percentage of Times Predicted to Wait	Percentage of Correctly Predicting to Wait
Del - Jdh	Tier 2	₹ 21,861.00	-₹ 6,419.00	39	₹ 47.51	8.39%	69.20%
Jdh - Del	Tier 2	₹ 61,314.00	-₹ 21,155.00	70	₹ 123.57	15.10%	80.00%
Bom - Jdh	Tier 2	₹ 51,552.00	-₹ 1,938.00	33	₹ 152.66	7.10%	75.80%
Del - Gau	Tier 2	₹ 639,205.00	-₹ 176,151.00	1372	₹ 164.79	40.70%	87.60%
Gau - Del	Tier 2	₹ 336,968.00	-₹ 46,189.00	623	₹ 181.40	31.00%	90.90%
Jdh - Bom	Tier 2	₹ 136,733.00	-₹ 14,877.00	84	₹ 441.51	20.70%	92.90%

Figure 8 : Analysis on Various Tier-2 Routes

Conclusion Remarks from Exploratory Data Analysis

From the data collected and through exploratory data analysis, we can determine the following:

- The trend of flight prices vary over various months and across the holiday.
- There are two groups of airlines: the economical group and the luxurious group. Spicejet, AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in
- the other. Vistara has a more spread out trend.

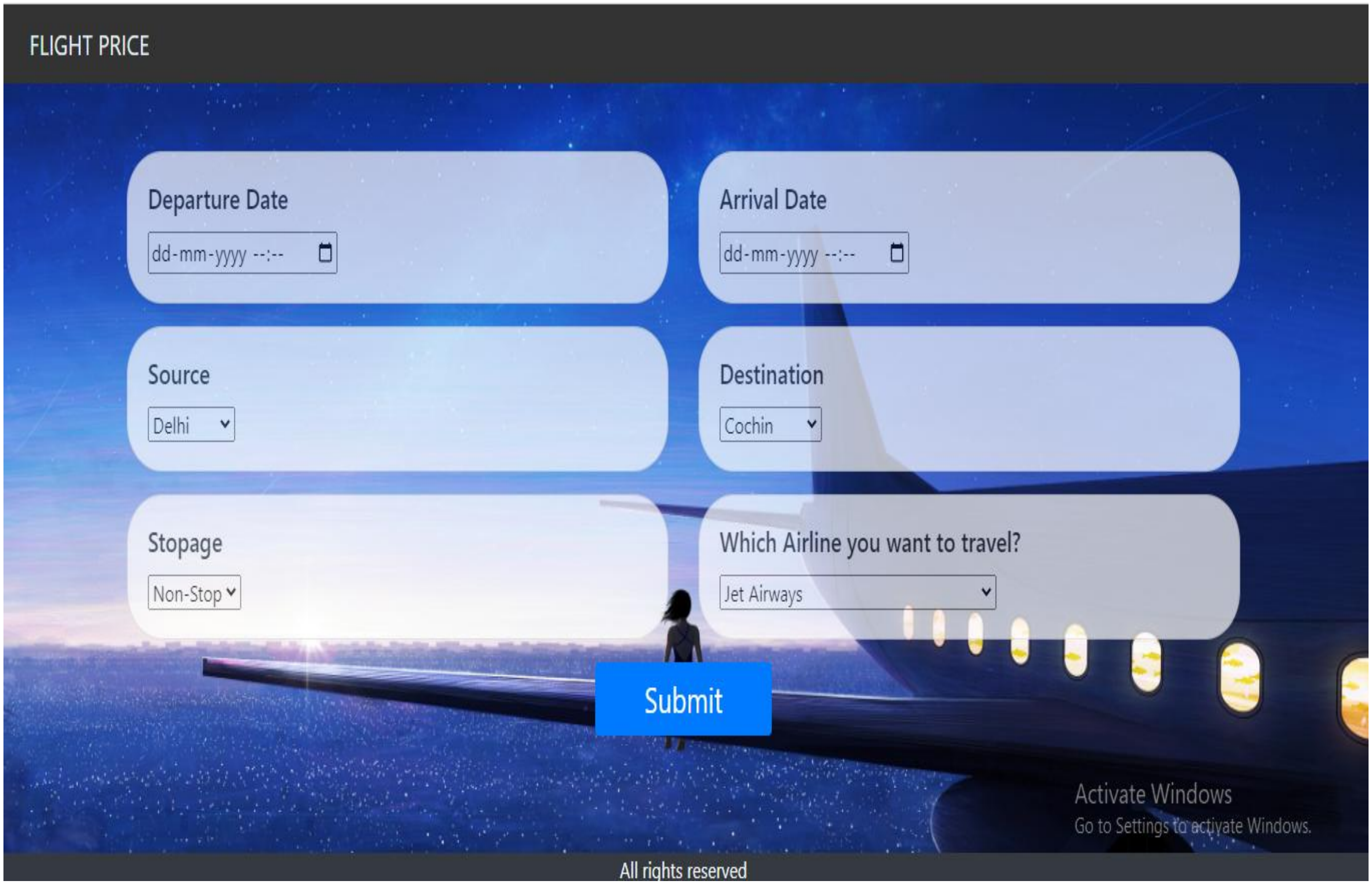
- The airfare varies depending on the time of departure, making timeslot used in analysis is an important parameter.
- The airfare increases during a holiday season. In our time period, during Diwali the fare remained high for all the values of days to departure. We have considered holiday season as a parameter which helped in increasing the accuracy.
- Airfare varies according to the day of the week of travel. It is higher for weekends and Monday and slightly lower for the other days.
- There are a few times when an offer is run by an airline because of which the prices drop suddenly. These are difficult to incorporate in our mathematical models, and hence lead to error.
- Along the business routes, we find that the price of flights increases or remains constant as the days to departure decreases. This is because of the high frequency of the flights, high demand and also could be due to heavy competition.
- Only about 8-10% of the times, a person should wait according to the data collected across the Mumbai-Delhi route, compared to 30-40% in Delhi-Guwahati route.

Future Work

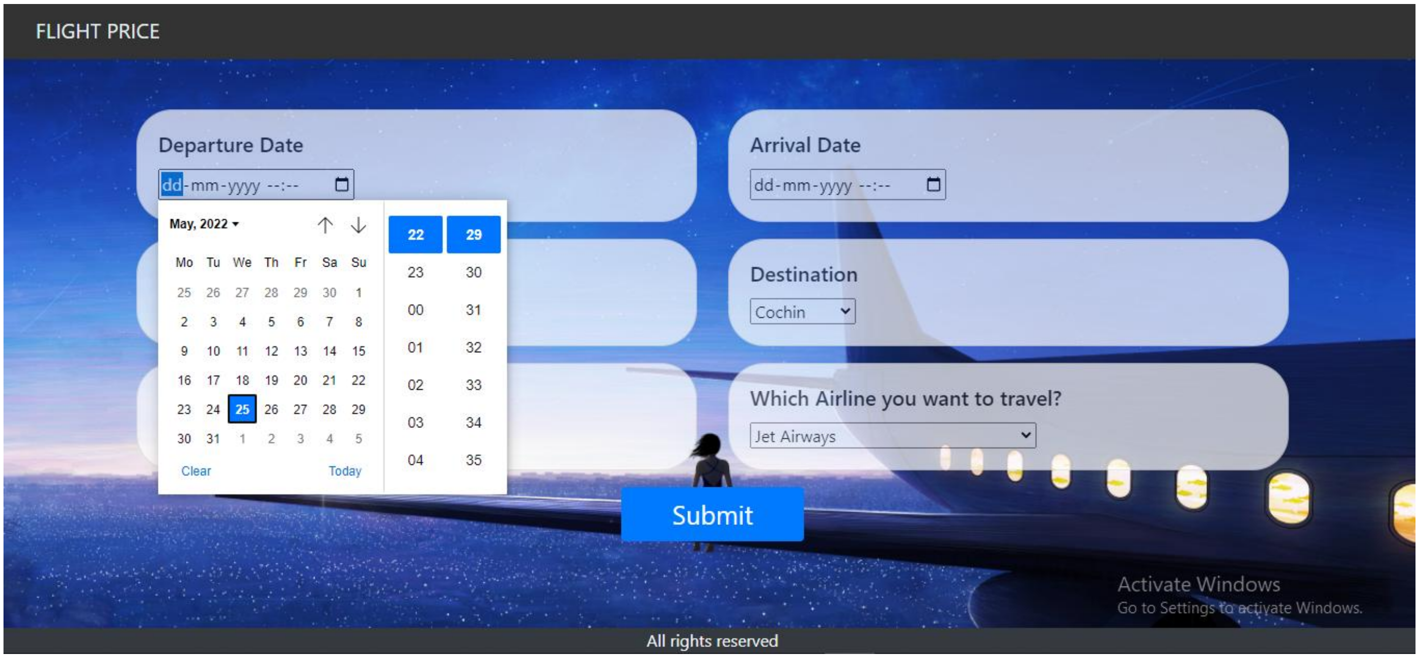
- More routes can be added and the same analysis can be expanded to major airports and travel routes in India.
- The analysis can be done by increasing the data points and increasing the historical data used. That will train the model better giving better accuracies and more savings.
- More rules can be added in the Rule based learning based on our understanding of the industry, also incorporating the offer periods given by the airlines.
- Developing a more user friendly interface for various routes giving more flexibility to the users.

User Interface:

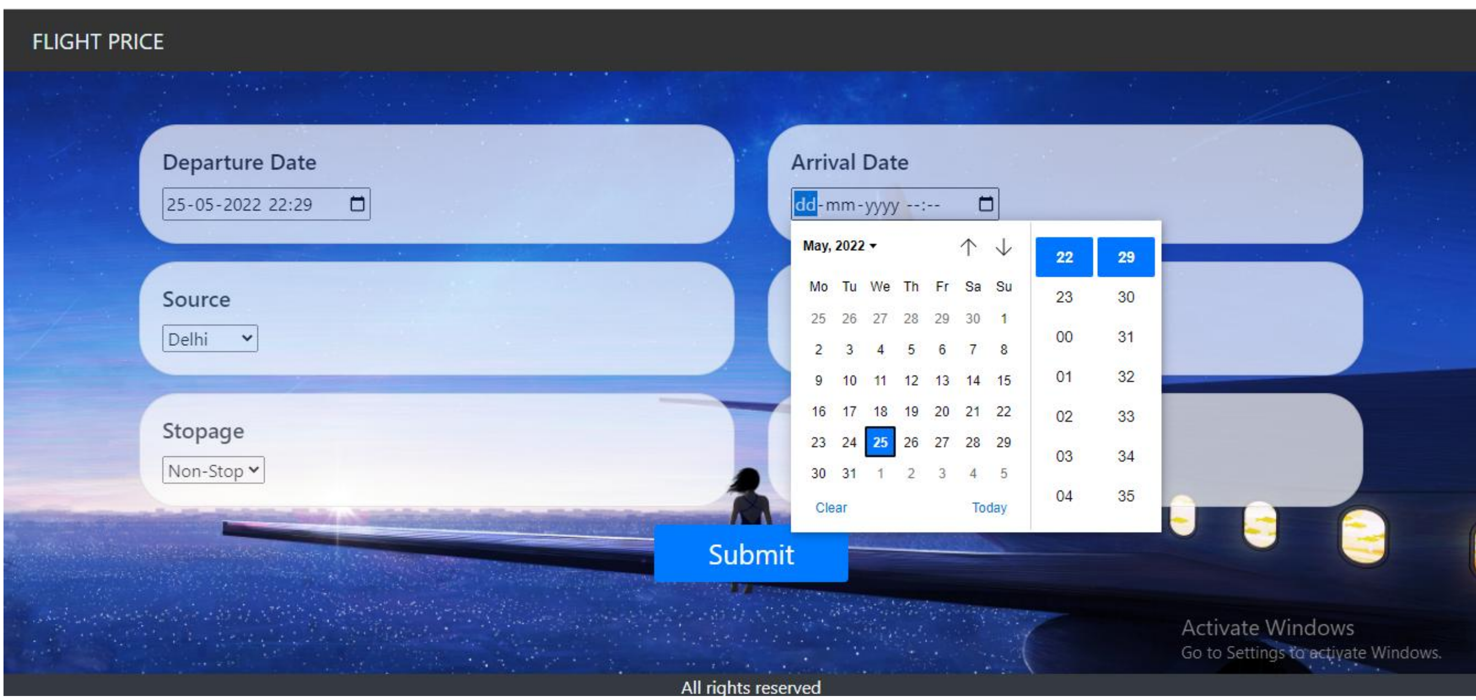
Home



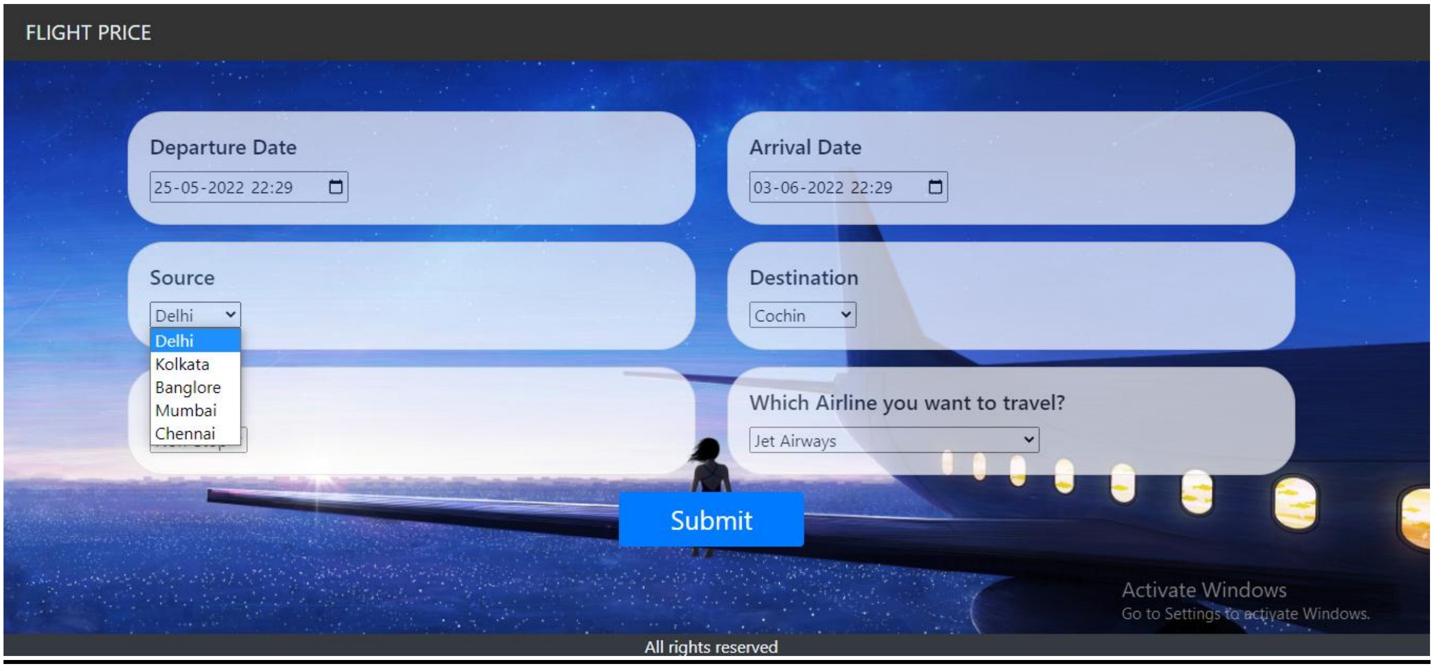
Select Departure Date and time



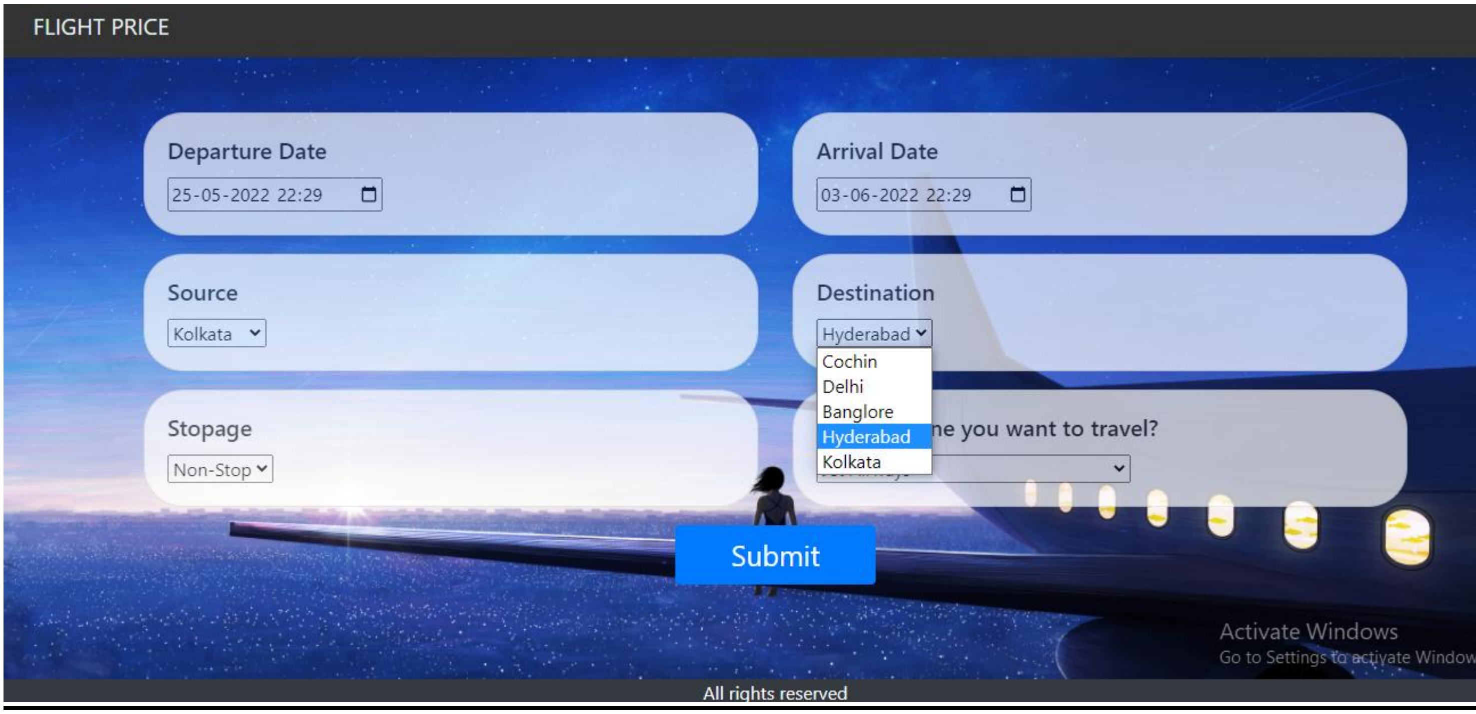
Select Arrival Date time



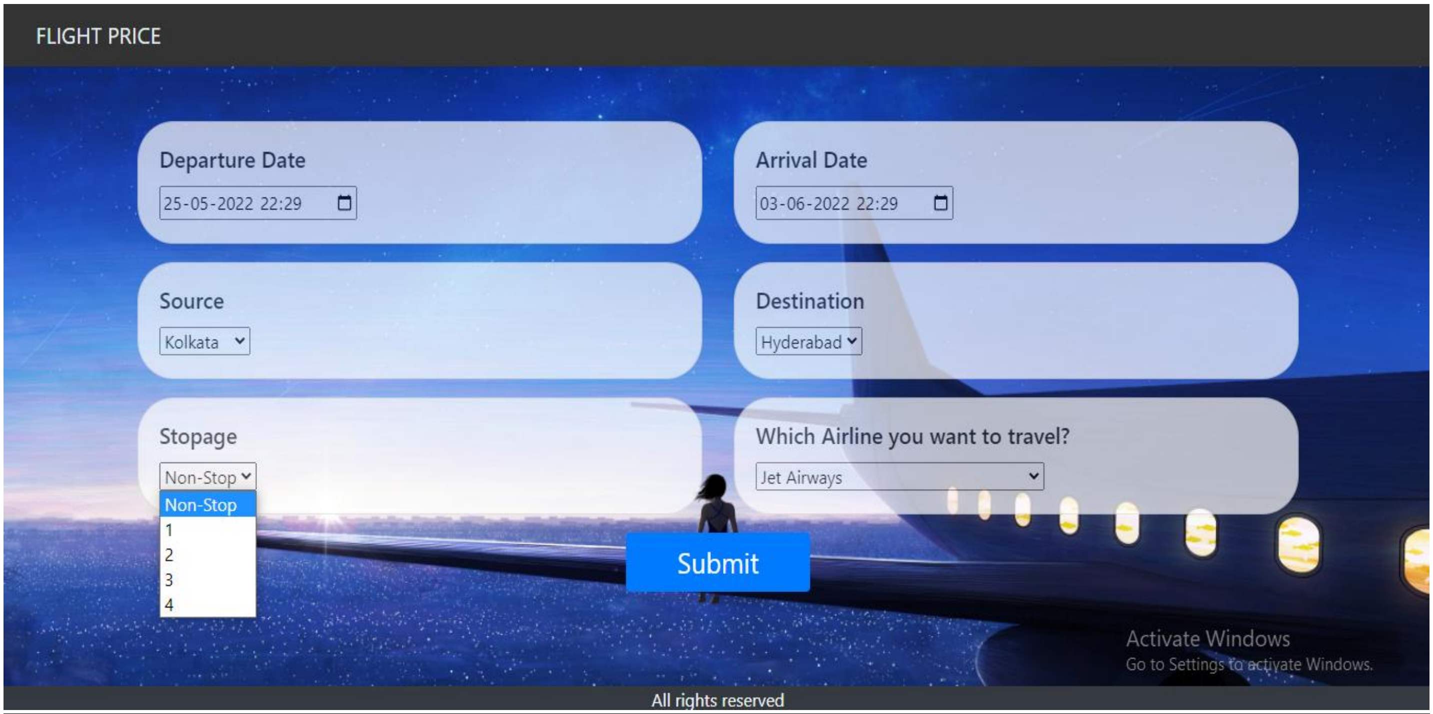
Select Source



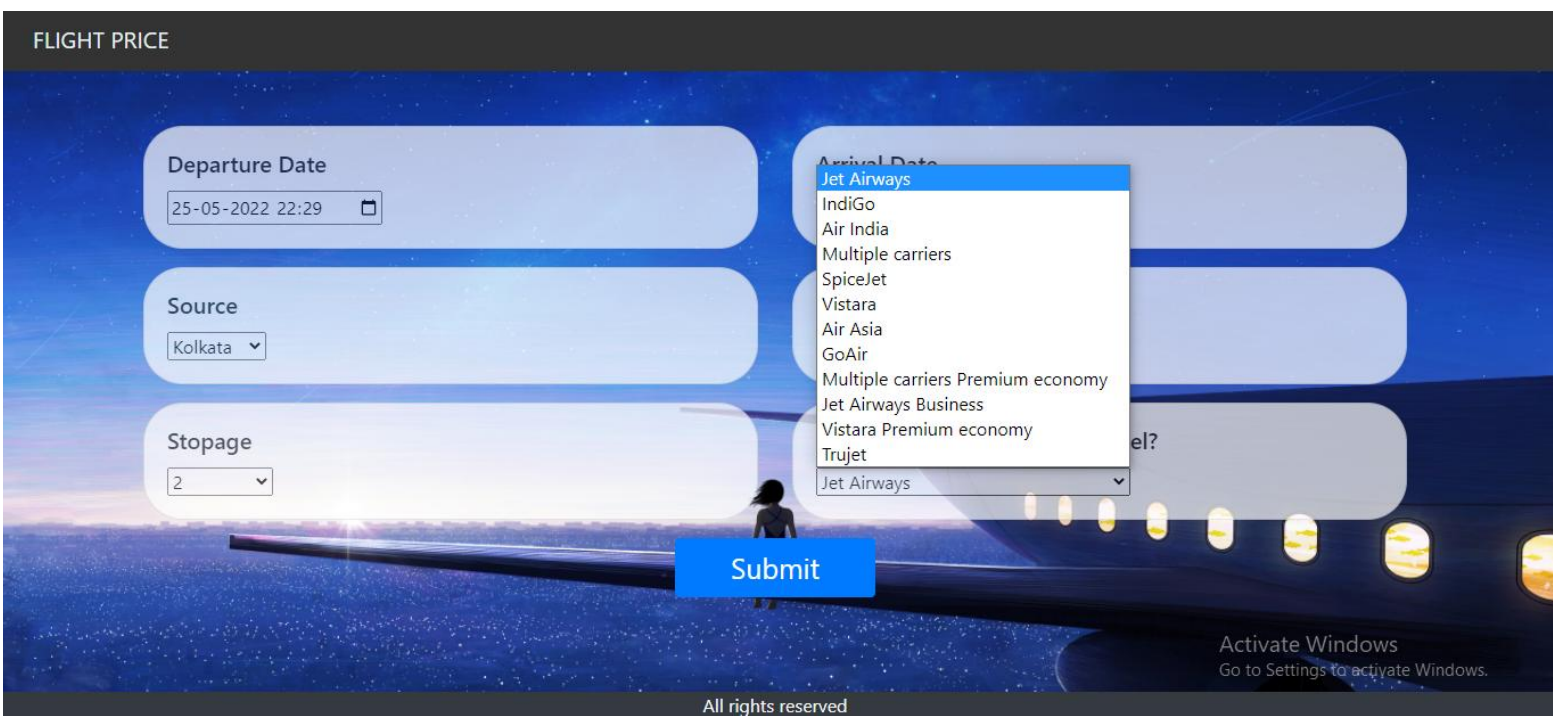
Select Destination



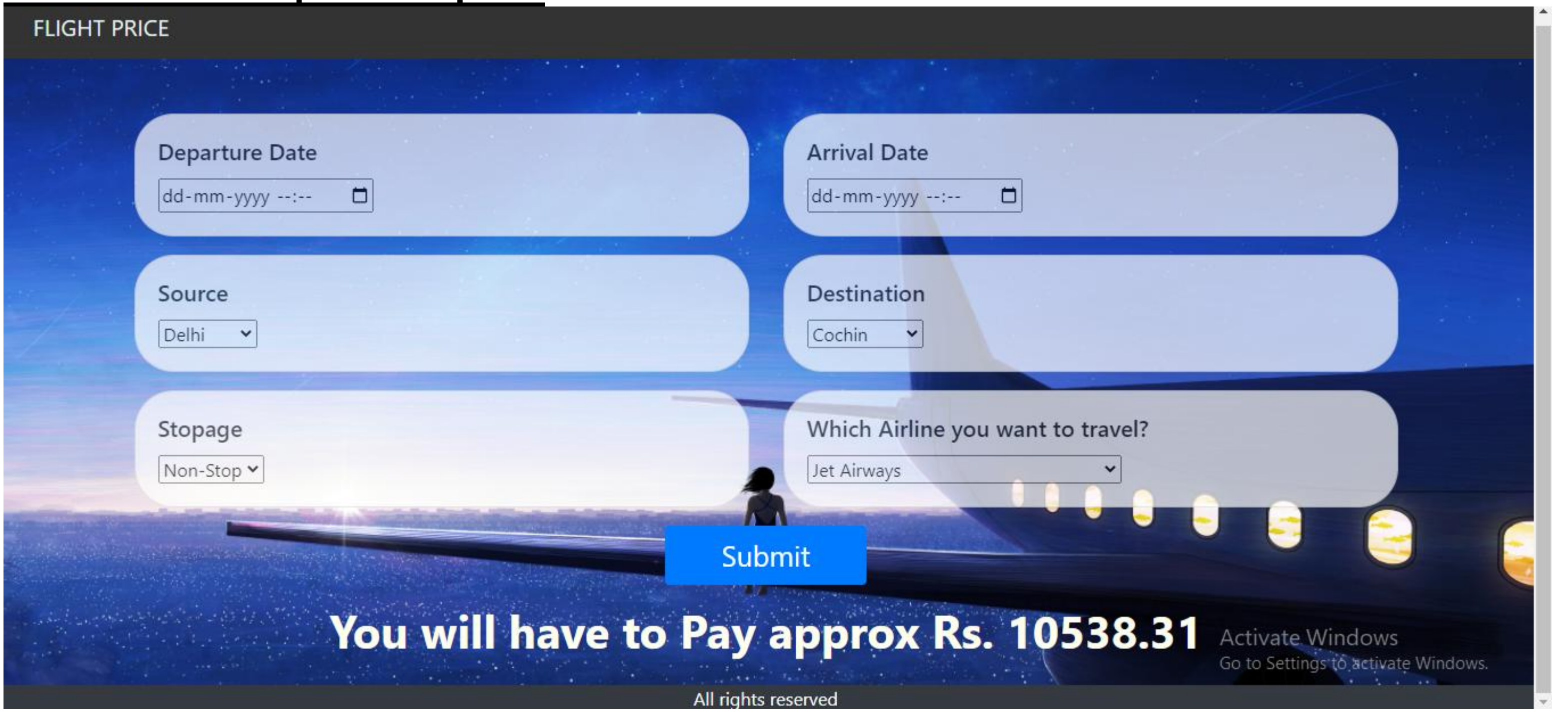
Select number of stoppage



Select Airline



Here we have our predicted price



Conclusion-

This study shows that it is feasible to predict the airline ticket price based on historical data. One possible way to increase the accuracy can be combining different models after carefully studying their own performance on each individual bin. Additionally, as the learning curve indicates, adding more features will increase the accuracy of our models. However, limited by the current data source that we have, we are unable to extract more information of a particular flight. In the future, more features, such as the available seat, the departure time of a day, and whether the departure day is a holiday or not, can be added to the model to improve the performance of the predicting model.

Reference-

1. O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates. To buy or not to buy: mining airfare data to minimize ticket purchase price.

2. Manolis Papadakis. Predicting Airfare Prices.

3. Groves and Gini, 2011. A Regression Model For Predicting Optimal Purchase Timing For Airline Tickets.

4. Modeling of United States Airline Fares – Using the Official Airline Guide (OAG) and Airline Origin and Destination Survey (DB1B), Krishna Rama-Murthy, 2006.

5. B. S. Everitt: The Cambridge Dictionary of Statistics, Cambridge University Press, Cambridge (3rd edition, 2006). ISBN 0-521-69027-7.

6. Bishop: Pattern Recognition and Machine Learning, Springer, ISBN 0-387-31073-8.

7. **library**

- Numpy - https://numpy.org/
- Pandas - https://pandas.pydata.org/
- Matplotlib - https://matplotlib.org/
- Seaborn- https://seaborn.pydata.org/
- Sklearn- https://scikit-learn.org/stable/

8.

Pickle- https://docs.python.org/3/library/pickle.html

Flask- https://flask.palletsprojects.com/en/2.1.x/

9.IDE

- Anaconda - https://www.anaconda.com/
- Jupyter - https://jupyter.org/
- Spyder - https://www.spyder-ide.org/