FLIGHT PRICE PREDICTION SYSTEM

A Minor Project Report

Submitted To

Chhattisgarh Swami Vivekanand Technical University

Bhilai, India

For

The Partial fulfillment of Degree

of

Bachelor of Technology

In

Computer Science & Engineering

By

SRISHTI XALXO
Roll No.:303302218035
Enrollment No.: BF4727
Semester 7th (CSE)

JAGRITI VERMA
Roll No.:30330228036
Enrollment No.: BF4728
Semester 7th (CSE)

AISHA AGRAWAL
Roll No.:303302218092
Enrollment No.: BF4784
Semester 7th (CSE)

Under the Guidance of

Ms. Kaveri Kar

Assistant Professor

Department of Computer Science & Engineering

S.S.I.P.M.T, Raipur

Department of Computer Science & Engineering

Shri Shankaracharya Institute of Professional Management & Technology Raipur (C.G.)

Session:2021- 2022

DECLARATION BY THE CANDIDATE

We undersigned solemnly declare that the Minor project report entitled **"FLIGHT PRICE PREDICTION SYSTEM"** is based on our own work carried out during the course of our study under the supervision of Asst. Prof. ***Kaveri Kar.***

We assert that the statements made and conclusions drawn are an outcome of the project work. We further declare that to the best of our knowledge and belief that the report does not contain any part of any work which has been submitted for the award of any other degree/diploma/certificate in this University/by Deemed university of India or any other country.

SRISHTI XALXO	**JAGRITI VERMA**	**AISHA AGRAWAL**
Roll No.:303302218035	**Roll No.:30330228036**	**Roll No.:303302218092**
Enrollment No.: BF4727	**Enrollment No.: BF4728**	**Enrollment No.: BF4784**
Semester 7th (CSE)	**Semester 7th (CSE)**	**Semester 7th (CSE)**

____

(Signature of the Supervisor)

Ms. Kaveri Kar
Assistant Professor Dept of C.S.E.
S.S.I.P.M.T
Raipur (C.G.)

CERTIFICATE BY THE SUPERVISOR

This is to certify that the Minor project report entitled **"FLIGHT PRICE PREDICATION SYSTEM"** is a record of project work carried out under my guidance and partial fulfillment of degree of Bachelor of Technology in the faculty of *Computer Science & Engineering* of Chhattisgarh Swami Vivekananda Technical University, Bhilai (C.G.) India.

To the best of my knowledge and belief the report

i) Embodies the work of the candidate him self

ii) Has duly been completed

iii) Fulfills the partial requirement of the ordinance relating to the B.Tech. degree of the University

iv) Is upto the desired standard in respect of contents and language for being referred to the examiners.

(Signature of the Supervisor)
Ms. Kaveri Kar
Assistant Professor Dept of C.S.E
S.S.I.P.M.T
Raipur (C.G.)

Forwarded to Chhattisgarh Swami Vivekanand Technical University
Bhilai

(Signature of HoD)
Dr. J. P. Patra
Dept. of Computer Science & Engineering
S.S.I.P.M.T.
Raipur (C.G.)

(Signature of the Principal)
Dr. Alok Jain
S.S.I.P.M.T.
Raipur (C.G.)

CERTIFICATE BY THE EXAMINERS

The project report entitled **"FLIGHT PRICE PREDICTION SYSTEM"** has by the undersigned as a part of the examination of Bachelor of Technology in the faculty of Computer Science & Engineering of Chhattisgarh Swami Vivekanand Technical University, Bhilai.

____________________	____________________
Internal Examiner	**External Examiner**
Date:	**Date:**

ACKNOWLEDGEMENT

Presentation inspiration and motivation have always played a key role in the success of any venture. We pay our deep sense of gratitude to **Dr. J. P. Patra (**HoD, Dept. of Computer Science & Engineering.) and express our sincere thanks to **Ms. Kaveri Kar** (Asst. Professor Dept. of Computer Science & Engineering) our Mentor to encourage us to the highest peak and to provide us the opportunity to prepare this project. We are immensely obliged for their elevating inspiration, encouraging guidance and kind supervision in this project. So, with due regards, we express our gratitude's to them. We believe this endeavor of us and our parents has greatly boosted our self-confidence and will go long on helping us reach further milestones and greater heights. It is our delight to renowned the maximum cooperation and precious pointers made now and then with the aid of using our department' steam of workers members, to whom we owe our complete knowledge, in addition to all the ones who've immediately or in around about way assisted us with the aid of using offering books, laptop peripherals, and different essential services that assisted us with inside the development of this project.

Thank-you for your guidance.

--------------------------	---------------------------	-----------------------------
(Signature of Candidate)	**(Signature of Candidate)**	**(Signature of Candidate)**
SRISHTI XALXO	**JAGRITI VERMA**	**AISHA AGRAWAL**
Roll No.:303302218035	**Roll No.:30330228036**	**RollNo.:303302218092**
Enrollment No.: BF4727	**Enrollment No.: BF4728**	**Enrollment No.: BF4784**
Semester 7th (CSE)	**Semester 7th (CSE)**	**Semester 7th (CSE)**

LIST OF ABBREVIATIONS

AI	ARTIFICIAL INTELLIGENCE
ML	MACHINE LEARNING
BTS	BUREAU OF TRANSPORTATION STATISTICS
MS	MICROSOFT
GUI	USER INTERFACE
OS	OPERATING SYSTEM
SAS	STATISTICAL ANALYSIS SYSTEM
DAX	DATA ANALYSIS EXPRESSION
RAM	RANDOM ACCESS MEMORY
ROM	READ ONLY MEMORY
EDA	ELECTRONIC DESIGN AUTOMATION
CSV	COMMA-SEPARATED VALUES
DS	DATA SCIENCE
BDA	BIG DATA ANALYSIS
IDA	INITIAL DATA ANALYSIS

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

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-Guwahati and Guwahati-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.

CHAPTER 1

INTRODUCTION

1.1 Background & Objective

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airline's 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 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 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.

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

1.1.1 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?

1.1.2 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?

1.1.3 Verifying Myths

Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

1.2 Overview

Nowadays, airline ticket prices can vary dynamically and significantly for the same flight, even for nearby seats within the same cabin. Customers are seeking to get the lowest price while airlines are trying to keep their overall revenue as high as possible and maximize their profit. Airlines use various kinds of computational techniques to increase their revenue such as demand prediction and price discrimination. From the customer side, two kinds of models are proposed by different researchers to save money for customers: models that predict the optimal time to buy a ticket and models that predict the minimum ticket price. In this paper, we present a review of customer side and airlines side prediction models. Our review analysis shows that models on both sides rely on limited set of features such as historical ticket price data, ticket purchase date and departure date. Features extracted from external factors such as social media data and search engine query are not considered. Therefore, we introduce and discuss the concept of using social media data for ticket/demand prediction.

1.2.1 Retail

Retail is one of the most popular industries for price forecasting. Price is a key driver of success in retail, especially now it’s so easy to find a deal online. That’s why price prediction in retail goes hand in hand with product analytics: you can analyze the demand, discover a product’s advantages and disadvantages, then determine whether a customer will buy it.

1.2.2 Marketing

Your price sends a message to your target audience, as people treat expensive and cheap products differently. Whether a $100 shampoo is better than a $10 one is up for debate. However, people will often connect price with quality. That said, those on modest incomes still need good yet affordable products. And marketers can use automated price prediction to reach the right people at the right price point.

1.2.3 Automotive

People are more price sensitive in specific niches. And the automotive sector is one where this is undoubtedly the case. Prices often tip over $100,000, while the price of cars will fluctuate based on local economic conditions and exchange rates throughout the supply chain.

And where in retail, few notice small price movements — with high-value items like cars, everyone does. But forget hiring a team of data analysts to manage your dealership's pricing strategy: integrate a price prediction tool and let machine learning tackle the task.

CHAPTER 2
LITERATURE REVIEW

2.1 Review

The price of an airline ticket is affected by a few factors, such as flight distance, purchasing time, fuel price, etc. Each carrier has its own proprietary rules and algorithms to set the price accordingly. Recent advance in Artificial Intelligence (AI) and Machine Learning (ML) makes it possible to infer such rules and model the price variation. This paper proposes a novel application based on two public data sources in the domain of air transportation: the Airline Origin and Destination Survey (DB1B) and the Air Carrier Statistics database (T-100). The proposed framework combines the two databases, together with macroeconomic data, and uses machine learning algorithms to model the quarterly average ticket price based on different origin and destination pairs, as
known as the market segment. The framework achieves a high prediction accuracy with 0.869 adjusted R quared score on the testing dataset.

Since the deregulation of the airline industry, airfare pricing strategy has developed into a complex structure of sophisticated rules and mathematical models that drive the pricing strategies of airfare. Although still largely held in secret, studies have found that these rules are widely known to be affected by a variety of factors. Traditional variables such as distance, although still playing a significant role, are no longer the sole factor that dictate the pricing strategy. Elements related to economic, marketing, and societal trends have played increasing roles in dictating the airfare prices.

Most studies on airfare price prediction have focused on either the national level or a specific market. Research at the market segment level, however, is still very limited. We define the term market segment as the market/airport pair between the flight origin and the destination. Being able to predict the airfare trend at the specific market segment level is crucial for airlines to adjust strategy and resources for a specific route. However, existing studies on market segment price prediction use heuristic-based conventional statistical models, such as linear regression, and assume that there exists a linear relationship between the dependent and independent variables, which in many cases, may not be true.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) make it possible to infer rules and model variations on airfare price based on many features, often uncovering hidden relationships amongst the features automatically. To the best of our knowledge, all existing work leveraging machine learning approaches for airfare price prediction are based on:

1) proprietary datasets that are not publicly available and

2) transaction records data crawled from online travel booking sites like Kayak.com.

The problem of the former lies in the difficulty of gaining access to the data, making reproducing the results and extending the work nearly impossible.

The issue with the latter is that the transaction records from each online booking site are a small fraction of the total ticket sales from the entire market, making the acquired data likely to be skewed, and thus, not representing the true nature of the entire market. In this paper, we address the problem of market segment level airfare price prediction by using publicly available datasets and a novel machine learning framework to predict market segment level airfare price. More specifically, our proposed framework extracts information from two specific public datasets, the DB1B and the T-100 datasets that are collected and maintained by the Office of Airline Information within the United States Bureau of Transportation Statistics (BTS). The DB1B dataset has been utilized in various studies that assess the determinants of aircraft characteristics and frequency of flights, analyses for the structure and dynamics of O-D for the core of the air travel market, and demand-prediction. The T-100 dataset includes air passenger volumes for U.S. domestic and international markets and covers large, certified carriers that hold Certificates of Public Convenience and Necessity. The goal of our proposed framework is to draw a comprehensive profile of each market and uses machine learning techniques to predict the average airfare on market segment level. The remainder of this paper is organized as follows. Section II reviews existing work that utilized either conventional statistical or machine learning algorithms for airfare price prediction. Section III provides a detailed description of the two datasets and the proposed framework. Section IV describes the experimental setup and presents the results of applying our proposed framework, as well as a comparison with several baseline methods. In section V, we conclude the paper with a discussion of our contribution and several potential directions for future work.

CHAPTER 3
SOFTWARE REQUIREMENTS

3.1 Software required

To create our project in an effective manner, we employed a variety of applications. We utilized Anaconda Navigator, Jupiter Notebook, MS Excel, and other applications in this system. Any system's development is incomplete without software. Regardless of the language in which the application was created. Software is an essential component of any application that aids in the growth of any system. Software is a collectionof programs or codes designed to improve and simplify the computer's functioning. The following is a list of the software that was used:

3.1.1 Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) that comes with the Anaconda® distribution and allows you to run programs and manage conda packages, environments, and channels without having to use command-line commands. Packages may be found on Anaconda.org or in a local Anaconda Repository using Navigator. It's compatible with Windows, Mac OS X, and Linux. Navigator is often used online to download and install software. Navigator needs to be able to access these sites in online mode, thus they may need to be whitelisted in our network's firewall settings

.

Fig 3.1.1: Anaconda Navigator

3.1.2 Microsoft Excel

Microsoft Excel is a fundamental, well-known, and commonly utilized analytical tool in practically all sectors. You will still need to utilize Excel, whether you are an expertin SAS, R, or Tableau.

Fig 3.1.2: Microsoft Excel

3.1.3 Jupyter Notebook

JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data for Jupyter notebooks, code, and data. JupyterLab is adaptable: you may customize and organize the user interface to accommodate a variety of data science, scientific computing, and machine learning workflows. JupyterLab is modular and expandable, allowing you to create plugins that add new features and connect with current ones.

Fig 3.1.3: Jupyter Notebook

3.1.4 Visual Studio Code

Microsoft's Visual Studio Code is a source-code editor for Windows, Linux, and macOS. Debugging, embedded Git control, GitHub, syntax highlighting, intelligent code completion, snippets, and code refactoring are all supported. It's a small but powerful source code editor that runs on your computer and is compatible with Windows, Mac OS X, and Linux.

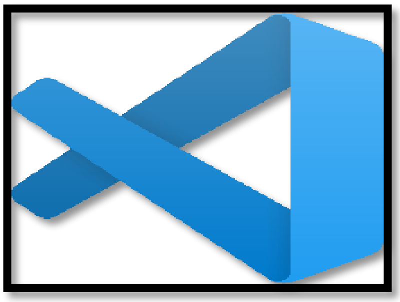

Fig 3.1.4: Visual Studio Code

3.1.5 Google Collab

Google Colab is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs

Fig 3.1.5: Google Colab

3.2 Developer's requirement

a. Operating systems- Windows* 7 or later, macOS, and Linux

b. Python* versions: 2.7.X, 3.6.X, 3.7 X, 3.8 X

c. Included development tools: conda, Jupyter Notebook

d. Compatible tools: Microsoft Visual Studio, PyCharm, Spyder

e. Included Python packages:

- NumPy
- Pandas
- Matplotlib
- Plotly
- Catplot
- Distplot
- Streamlit
- Seaborn
- Violin
- Bokeh
- Pillow

CHAPTER 4

HARDWARE REQUIREMENTS

4.1 Hardware required

Hardware plays just as vital a role as software. If the programmer necessitates suitable and precise software, it will likewise want appropriate hardware. Hardware settings should be based on the requirements of the software being created. Incorrect hardware configurations may result in an unfavorable outcome for the system being created. The RAM, ROM, and processor of the system that is being utilized for the project are the basic hardware requirements. The following are the requirements' explanations:

- The processor is a logical circuit that responds to fundamental instructions and processes them to run a computer system. It is a prerequisite since a computer cannot function without it. Every time, an updated processor should be utilized to ensure that there is no misbehavior on the part of the processor.
- Another major element of a computer system is RAM. It is a computer's storage device. The RAM holds the data and machine codes that the software is now processing. While creating or executing the software, there should be enough RAM available. The lack of RAM space may result in the designed system's failure to perform properly.
- Another significant component of a computer system is the read-only memory (ROM). The ROM holds the computer's memory that can only be read and not updated to. The ROM enables us to boot the computer system whenever we turn it on. It does so by exposing some functionality.

So, with the support of the above-mentioned hardware explanation, we can readily comprehend the significance of hardware in the creation of any computer system project. A system cannot function correctly without ideal hardware, hence appropriate and precise hardware is required while building or running any system.

4.2 Developer's requirement

- Processor –We need a processor for creating our project which should be Pentium4and above
- RAM –For the storage purpose of our project again we require a primary memory(RAM) which should be 256 MB or above
- Hard Disk – Again we require a hard disk for storing our project and the hard diskshould have a space of 512 MB or above
- Display – Any compatible monitor.
- Connection – Internet Connection.

CHAPTER 5
METHODOLOGY

5.1 Data science Methodology

This methodology can be used within data science, to ensure that the data used in problem solving is relevant and properly manipulated to address the question at hand. There are five modules, each going through two stages of the methodology, explaining the rationale as to why each stage is required.

5.1.1 Business understanding: -

Data science methodology begins with spending the time to seek clarification, toattain what can be referred to as a business understanding. Having this understandingis placed at the beginning of the methodology because getting clarity around the problem to be solved, allows you to determine which data will be used to answer the core question.

5.1.2 Analytic approach: -

Once the problem to be addressed is defined, the appropriate analytic approach for the problem is selected in the context of the business requirements. This is the secondstage of the data science methodology. Once a strong understanding of the question is established, the analytic approach can be selected. This means identifying what type ofpatterns will be needed to address the question most effectively. If the question is to determine probabilities of an action, then a predictive model might be used. If the question is to show relationships, a descriptive approach maybe is required.

5.1.3 Data Requirements: -

This includes identifying the necessary data content, formats, and sources for initial data collection.

5.1.4 Data collection: -

In this phase the data requirements are revised, and decisions are made as to whether the collection requires data. Once the data ingredients are collected, then in the data collection stage, the data scientist will have a good understanding of what they will be working with.

5.1.5 Data understanding: -

Essentially, the data understanding section of the data science methodology answers the question: Is the data that you collected representative of the problem to be solved?

5.1.6 Data preparation: -

Transforming data in the data preparation phase is the process of getting the datainto a state where it may be easier to work with. Specifically, the data preparation stage of the methodology answers the question:

Feature engineering is also part of data preparation. It is the process of using domain knowledge of the data to create features that make the machine learning algorithms work. Data preparation isoften a lengthy undertaking for data professionals or business users, but it is essential as a prerequisite to put data in context to turn it into insights and eliminatebias resulting from poor data quality. For example, the data preparation process usually includes standardizing data formats, enriching source data, and/or removing outliers.

- **Modeling: -**

Data Modeling focuses on developing models that are either descriptive or predictive. An example of a descriptive model might examine things like: if a person did this, then they're likely to prefer that. A predictive model tries to yield yes/no, or stop-go type outcomes.

- **Evaluation: -**

Model evaluation is performed during model development and before the model is deployed. Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.

Evaluation answers the question: Doesthe model used really answer the initial question or does it need to be adjusted. In the Model Evaluation stage, data scientists can evaluate the model in two ways: Hold-Out and Cross-Validation. In the Hold-Out method, the dataset is divided into three subsets: a training set as we said in the modeling stage; a validation set that is a subset used to assess the performance of the model built in the training phase; a test set is a subset to evaluate the likely future performance of a model.

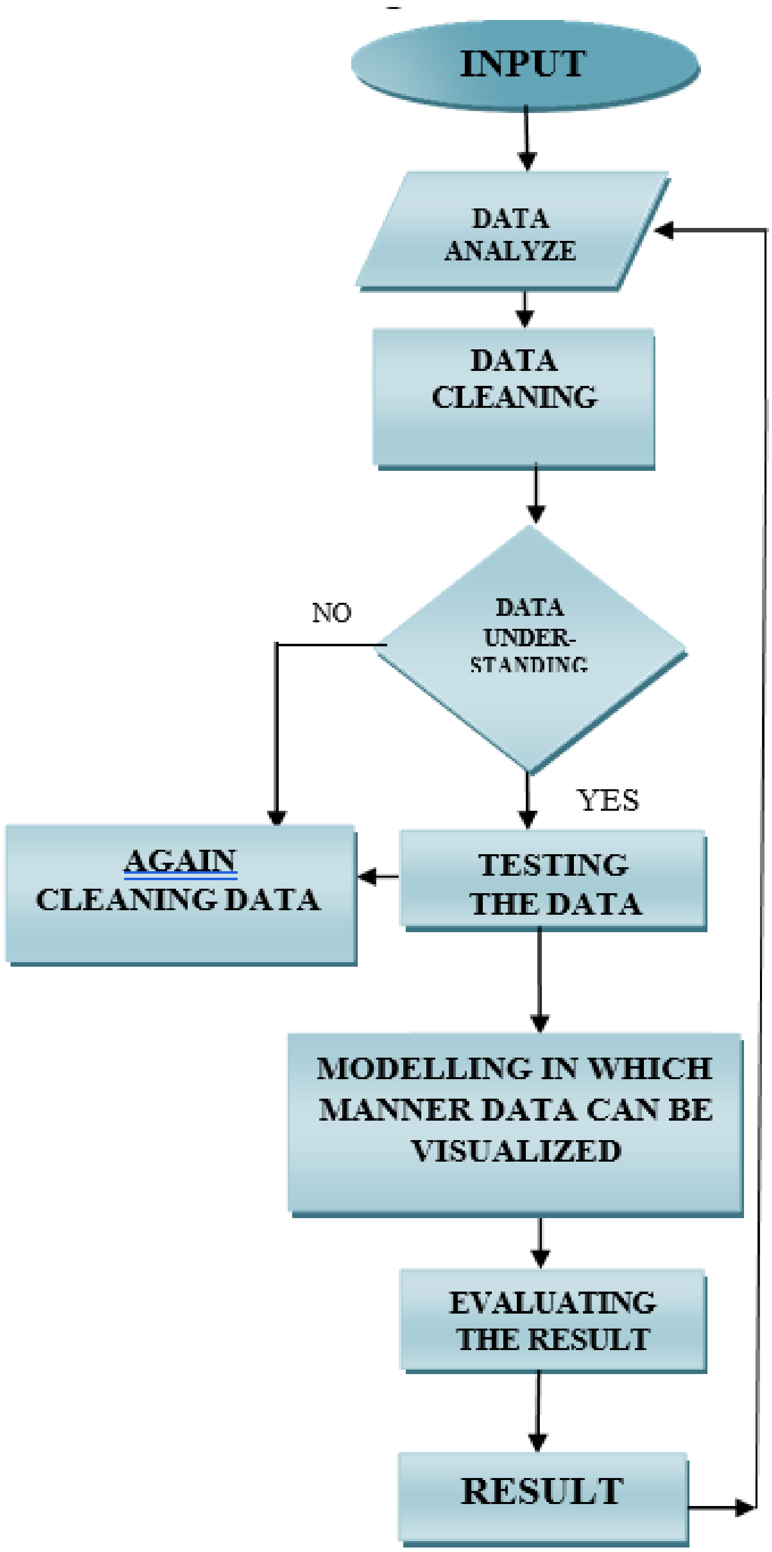

Fig 5.1.1: Flowchart of Data Analysis phases

5.2 Mechanism

Technology can bring a solution through the implementation of Machine learning techniques to improve the uncertainty of flight prices in the future. We will use Flight Price Dataset provided by Kaggle Flight Price.

Flight Fare Prediction MH Flight Fare Prediction Dataset by Machine Hack www.kaggle.com

This dataset consists of 10683 records with 13 columns that explain the flight in India by some Indian and foreign Airlines in 2019. We will analyze this dataset using Machine learning techniques to predict the flight ticket price based on the features provided in the columns of the dataset. We will begin the Data Science Life Cycle to process the data. Before enjoying this article, make sure the reader can understand some basics of python, machine learning techniques, hyperparameter tuning, and familiarity with pandas, seaborn, and sci-kit-learn

5.2.1 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 analyzing, 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 -

- Origin City
- Destination City
- Departure Date
- Departure Time
- Arrival Time
- Total Fare
- Airway Carrier
- Duration
- Class Type - Economy/Business
- Flight Number
- Hopping – Boolean
- Taken Date - date on which this data was collected

5.2.2 Data Cleaning

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

- Days to Departure
- Day of Departure
- Duration
- Hopping
- Holiday
- Outliers

Further, the data was analyzed 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.

To predict if the customer must 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 considering 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.

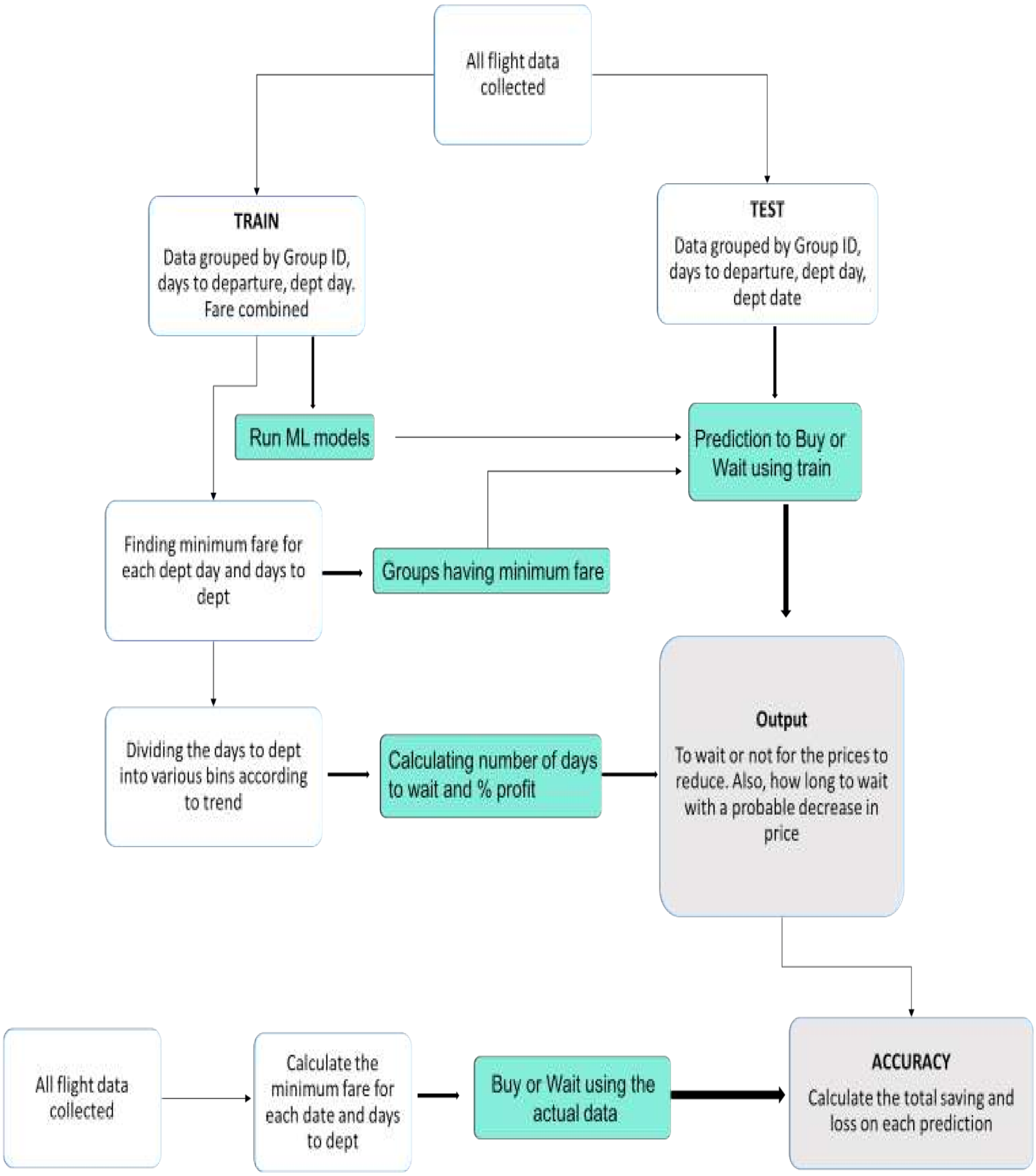

Fig 5.2.1: Data cleaning

5.2.3 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.

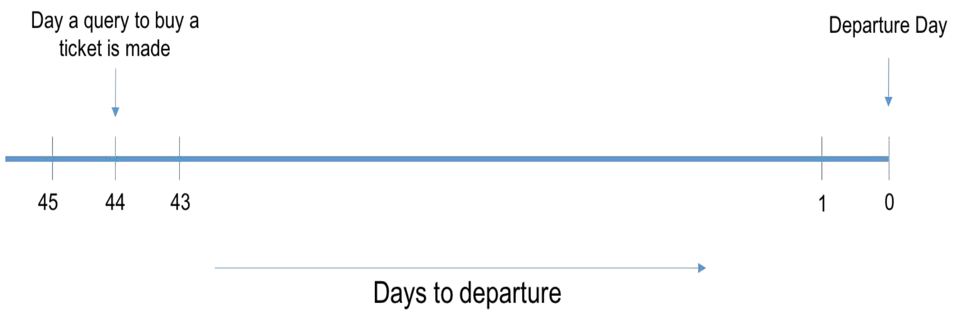

Fig 5.2.2: Data preparation

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 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 will 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 to use 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 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 timeslot 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

Table 5.2.1 Combining flights fare in one group

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 custom Fare = w*(First Quartile for entire time period) + (1-w) *(First quartile of last x days)

(We have considered: w = 0.7 and x = & days)

Calculating whether to buy or wait for this data:

Logical = 1 if for any d < D the Total custom Fare is less than the current Total custom Fare (Here, d is the days to departure and D is the days to departure for the current row.)

5.2.4 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 Custom Fare 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 Custom Fare corresponding to each combination of Departure Date and Days to Departure. Now with the obtained minimum Custom Fare corresponding to each pair, we do a merge with our initial dataset and find out the Airline corresponding to which the minimum Custom Fare 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 analyses 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

Table 5.2.2 An Alyse the time to wait before the prices reduce

5.2.5 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

Table 5.2.3 Calculation of Price Drop Percentage

CHAPTER 6
RESULT & ANALYSIS

6.1 Libraries used:

In our code we have used 4 libraries numpy, pandas, matplotlib and seaborn.

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

Pandas is a Python library that gives you a fantastic set of tools to do data analysis.

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open-source alternative to MATLAB.

Seaborn is a Python library created for enhanced data visualization.

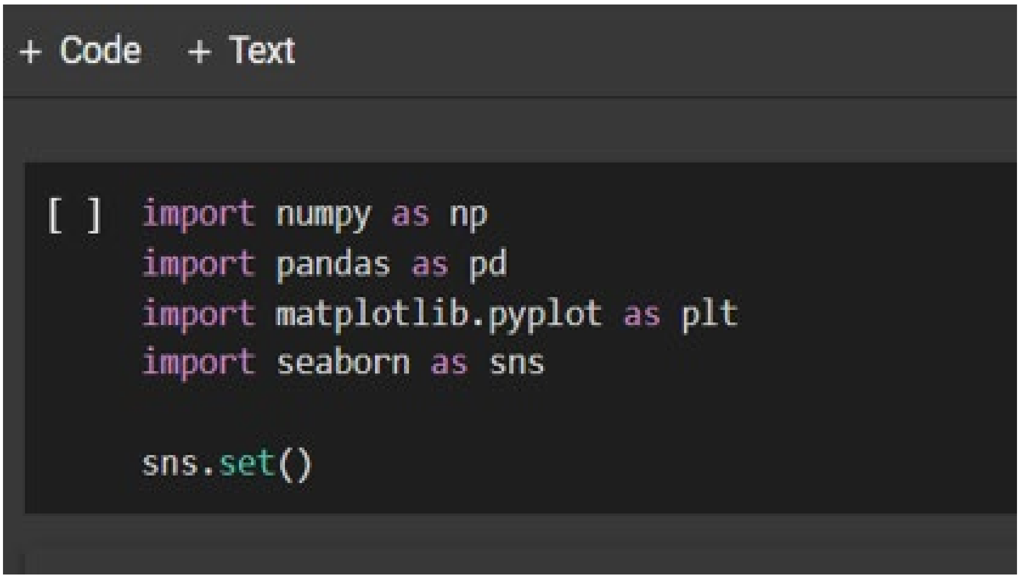

Fig 6.1.1: Libraries used

6.2 Dataset:

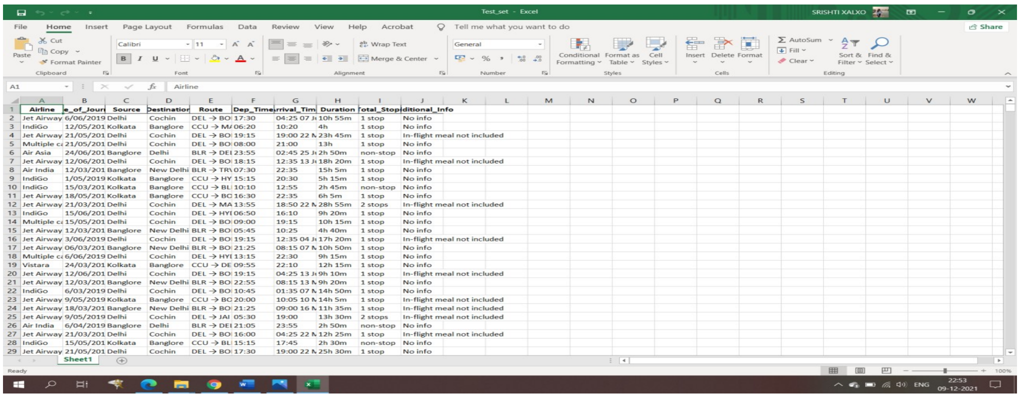

Fig 6.2.1: Dataset

We have taken data from Kaggle. It consists of data of flight from 2019 to 2021.There are two set of data one is training data and second is testing data.

6.3 Pre-processing:

In this figure we can see that data are processed where string data are converted into numeric data, because model will not be able to understand string values.

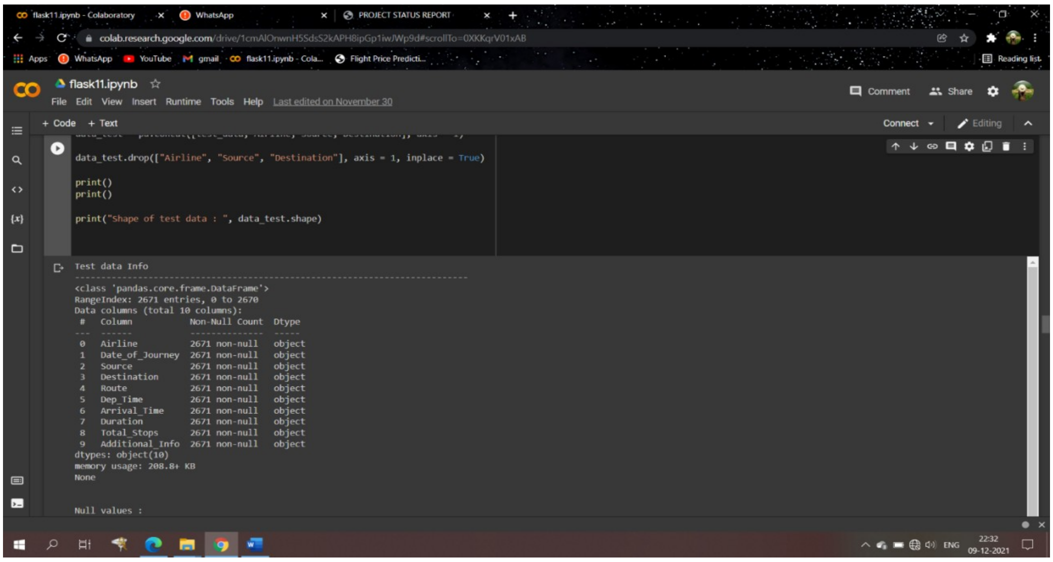

Fig 6.3.1: Pre-processing

6.4 Model:

We have used Random Forest model. The logic behind the Random Forest model is that multiple uncorrelated models (the individual decision trees) perform much better as a group than they do alone. When using Random Forest for classification, each tree gives a classification or a "vote." The forest chooses the classification with the majority of the "votes."

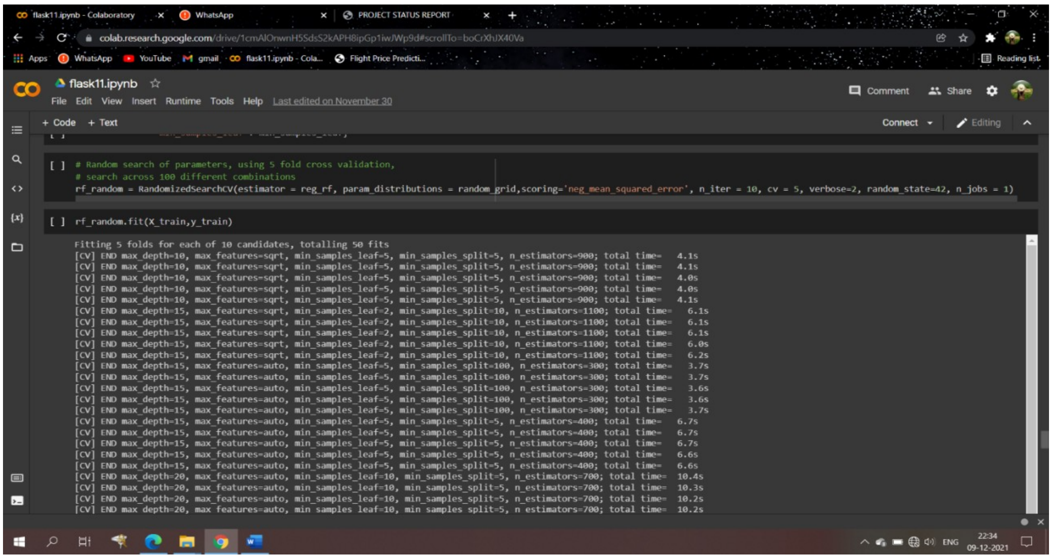

Fig 6.4.1: Model

6.5 Comparison of price between city & flight company

In these two figures we can see that we are comparing price between city & flight company. In figure 6.5.1 we can see that Bangalore has the highest price as compared to other cities and in figure 6.5.2 Jet Airway has the highest price than other flight company.

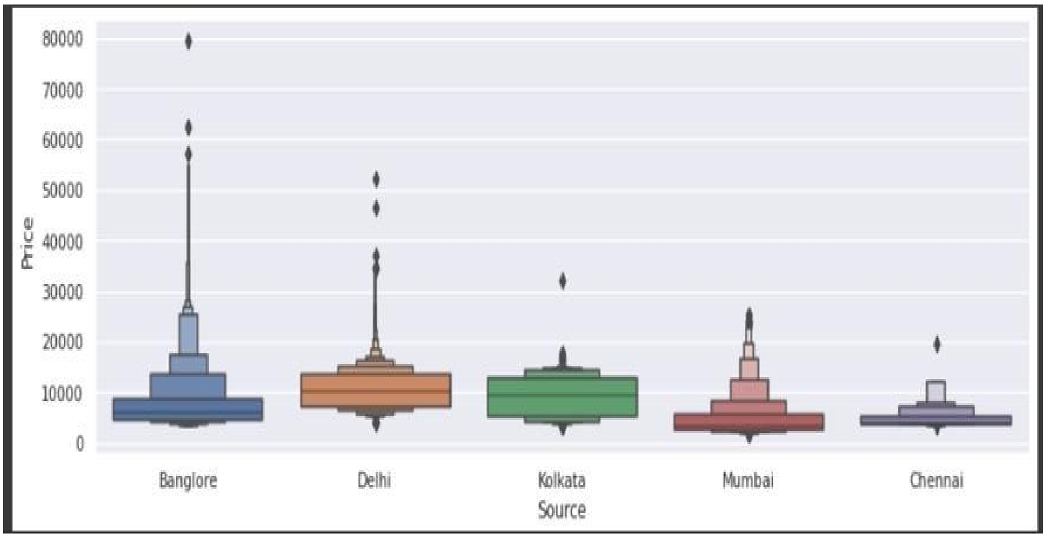

Fig 6.5.1: Comparison of price between cities.

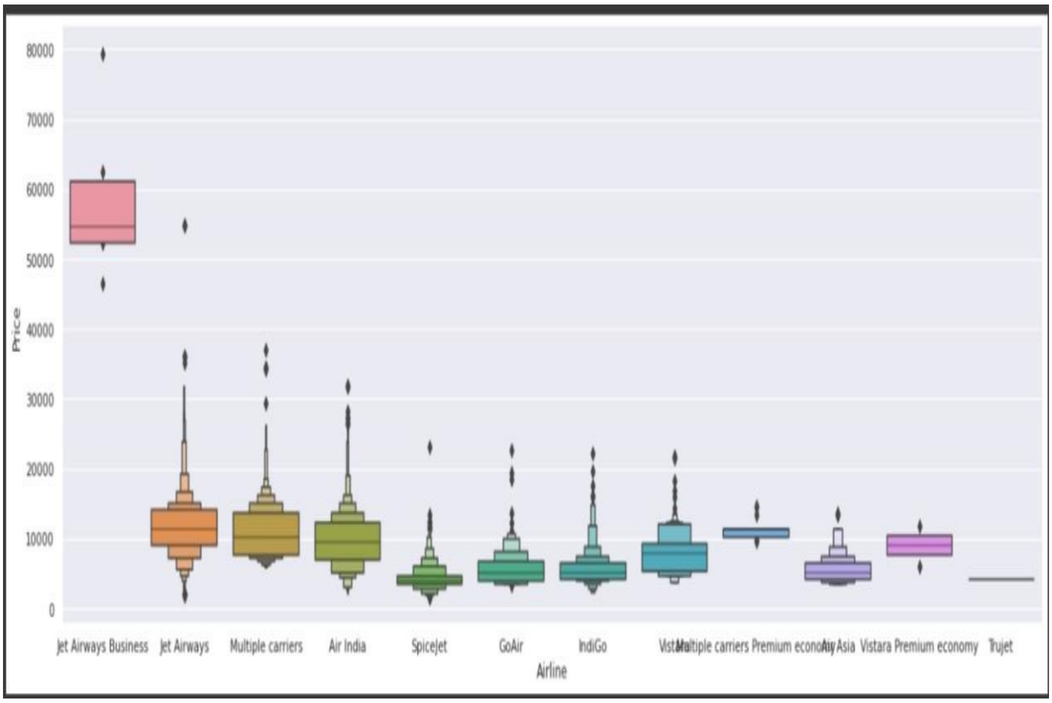

Fig 6.5.2: Comparison of price between flight company.

6.6 Feature correlation:

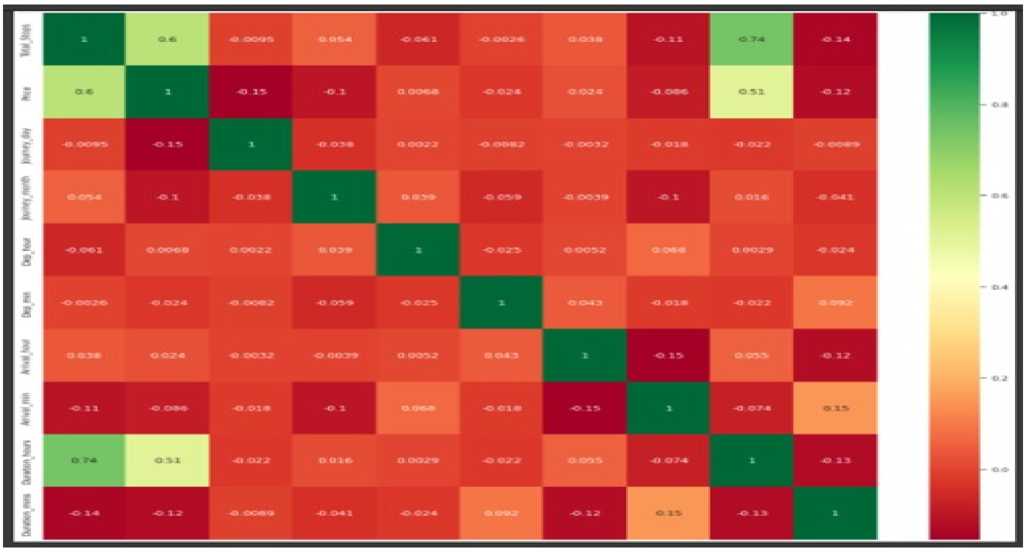

Fig 6.6.1: Future correlation

Correlation basically means a mutual connection between two or more sets of data. In statistics bivariate data or two random variables are used to find the correlation between them. Here we are finding correlation between independent and dependent attributes

6.7 Feature importance:

In this graph we can see the important features or attributes which are used for predicting price of flight.

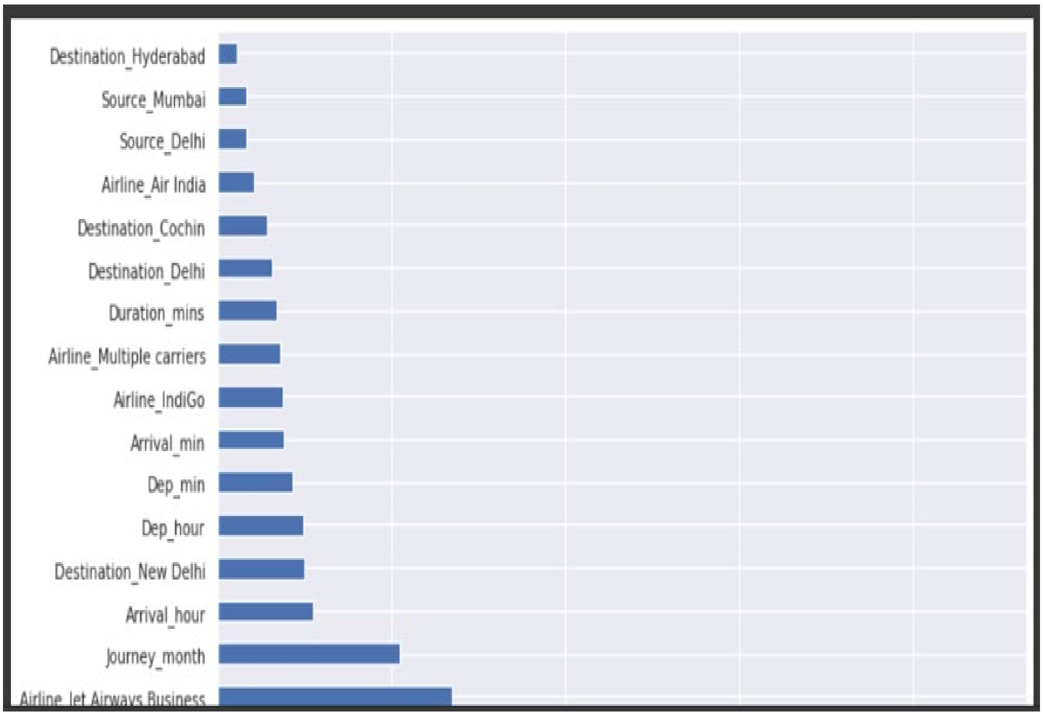

Fig 6.7.1: Feature importance

6.8 Accuracy before and after Hyperparameter Tuning:

In these two figures we can see the accuracy level before the Hyperparameter Turning in fig.6.8.1 and accuracy level after the Hyperparameter Turning In fig 6.8.2

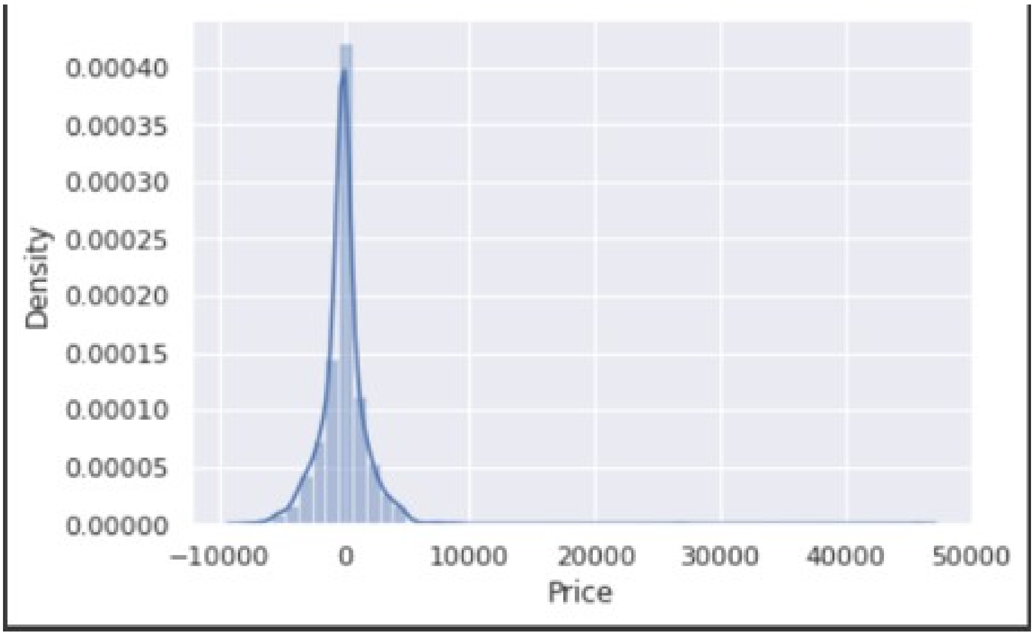

Fig 6.8.1: Accuracy before hyper-parameter tuning

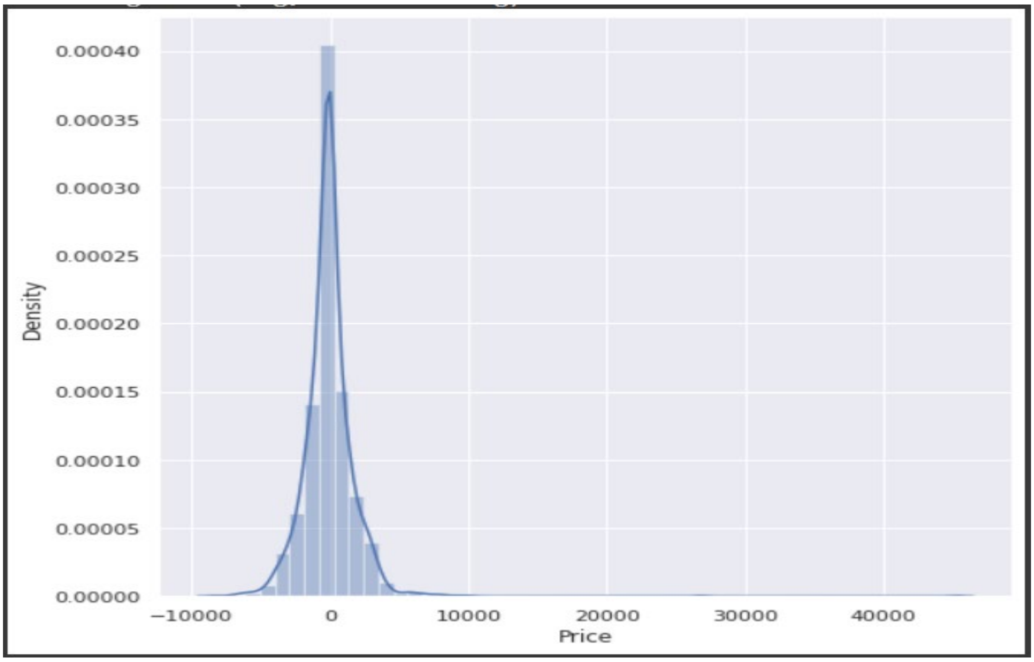

Fig 6.8.2: Accuracy after hyperparameter testing

6.9 Output:

Our project user interface consists of departure date, arrival date, source, destination, stoppage and airline after filling all this details and clicking submit button user can get predicted value.

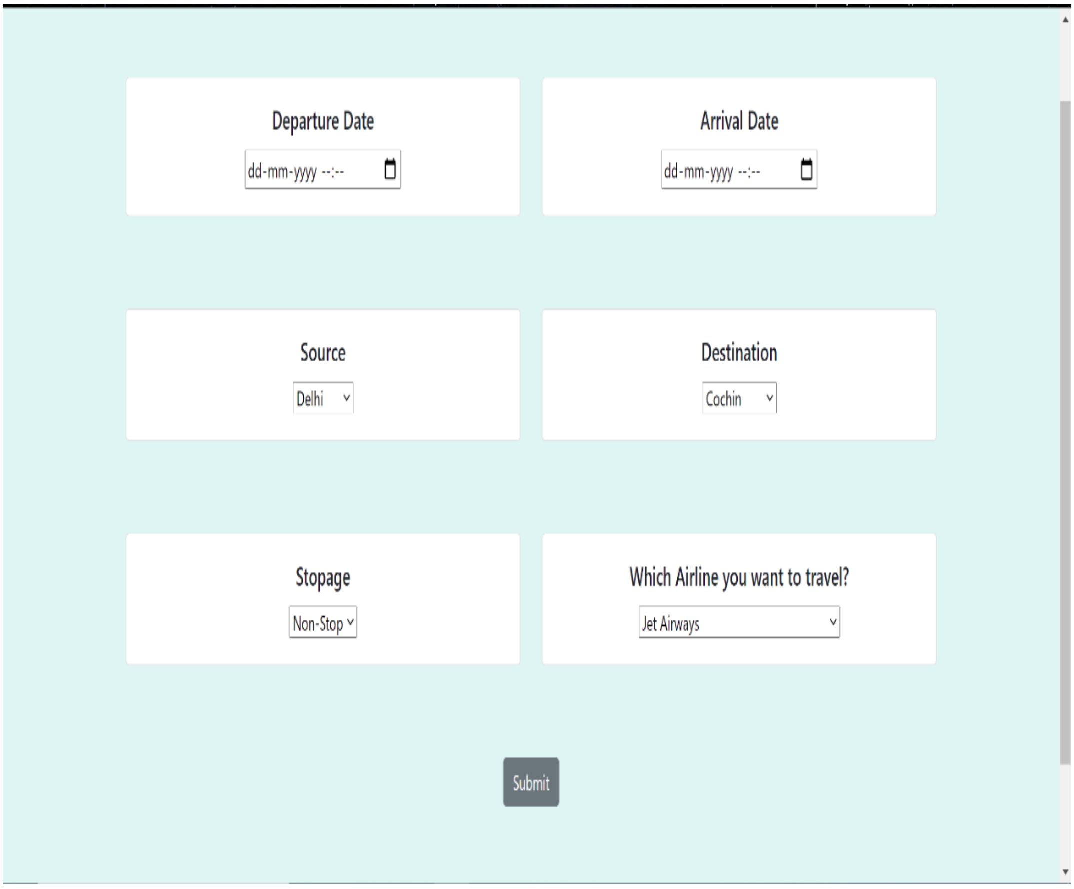

Fig 6.9.1: Output

CHAPTER 7
FEATURES & FUNCTIONALITIES

7.1 Fundamental features of project

Airline companies use complex algorithms to calculate flight prices given various conditions present at that time. These methodstake financial, marketing, and various social factors into account to predict flight prices.

The s alien features of our portfolio app are: -.

- File uploading: In this app user can upload CSV or excel file through which we can get data & do visualization.
- Dropdown -box: Through this box user can choose the type offlights, times, stoppage which they want to solve and easily understand by looking at the charts& graphs.
- In this app we can zoom in or out the visualized graph or chartsand can be saved in the form of image

7.2 Functions

- Analysis of price of flight we use attributes like date of arrival, departure, time, stoppage, and flight.
- Flight price value in last 3 years.
- Comparison between different flights.
- Comparison between stoppage.
- Different graphs are representation of comparing. Violin plotting of Bitcoin analysis

CHAPTER 8
FUTURE SCOPE & CONCLUSION

8.1 Future scope

- 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.

8.2 Conclusion Remark

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

- The trend of flight prices varies over various months and across the
- holiday.
- There are two groups of airlines: the economical group and the luxurious group.
- 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 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.

8.3 Conclusion

It as a battle between 'The risk appetite from our detailed analysis of each of the 18 routes, we can determine the following

- Flight prices almost always remain constant or increase between the major cities
- Tourist routes and routes that offer services involving Tier-2 cities of the country have uneven
- trends related to the increase and decrease of airline ticket prices.
- The model in the worst case almost breaks even with the profits and losses, and most case saves an average of about Rs. 200 per transaction when predicting to wait.
- Routes with data collected over the longer duration of time tend to facilitate with much more

accurate predictions in the model and thus lead to higher average savings. We were successfully able to analyse each route and generalize the entire project based in terms of the sector to which the route belonged and classified them into three majors' subsections - Business Routes, Tourist Routes and Tier-2 Routes. We have also successfully busted some of the typical myths and misconceptions related to the airline industry and backed them up with data and analysis. Finally, we have created a User Interface for the entire process of buying an airline ticket and given a proof of our predictions based on the previous trends with our prediction. Thus, leaving of the user' vs 'Our understanding of the airline industry'.

REFERENCES

References

1. J. Stavins, “Price discrimination in the airline market: The effect of market concentration,” Review of Economics and Statistics, vol. 83, no. 1, pp. 200–202, 2001.
2. B. Mantin and B. Koo, “Dynamic price dispersion in airline markets,” Transportation Research Part E: Logistics and Transportation Review, vol. 45, no. 6, pp. 1020–1029, 2009.
3. P. Malighetti, S. Paleari, and R. Redondi, “Has ryanair’s pricing strategy changed over time? an empirical analysis of its 2006–2007 flights,” Tourism Management, vol. 31, no. 1, pp. 36–44, 2010.
4. T. H. Oum, A. Zhang, and Y. Zhang, “Inter-firm rivalry and firm-specific price elasticities in deregulated airline markets,” Journal of Transport Economics and Policy, vol. 7, no. 2, pp. 171–192, 1993.
5. B. Burger and M. Fuchs, “Dynamic pricing – A future airline business model,” Journal of Revenue and Pricing Management, vol. 4, no. 1, pp. 39–53, 2005.
6. T. M. Vowles, “Airfare pricing determinants in hub-to-hub markets,” Journal of Transport Geography, vol. 14, no. 1, pp. 15–22, 2006.
7. K. Rama-Murthy, “Modeling of united states airline fares– using the official airline guide (OAG) and airline origin and destination survey (DB1B),” Ph.D. dissertation, Virginia Tech, 2006.
8. B. Derudder and F. Witlox, “An appraisal of the use of airline data in assessing the world city network: a research notes on data,” Urban Studies, vol. 42, no. 13, pp. 2371–2388, 2005.
9. Mottini and R. Acuna-Agost, “Deep choice model using pointer networks for airline itinerary prediction,” in the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 1575–1583.
10. K. Tziridis, T. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, “Airfare prices prediction using machine learning techniques,” in the 25th IEEE European signal processing conference, 2017, pp. 1036–1039.

11. Y. Chen, J. Cao, S. Feng, and Y. Tan, “An ensemble learning based approach for building airfare forecast service,” in the IEEE international conference on big data, 2015, pp. 964– 969.
12. T. Liu, J. Cao, Y. Tan, and Q. Xiao, “ACER: An adaptive context-aware ensemble regression model for airfare price prediction,” in the international conference on progress in informatics and computing, 2017, pp. 312–317.
13. V. Pai, “On the factors that affect airline flight frequency and aircraft size,” Journal of Air Transport Management, vol. 16, no. 4, pp. 169–177, 2010.
14. M. S. Ryerson and H. Kim, “Integrating airline operational practices into passenger airline hub definition,” Journal of Transport Geography, vol. 31, pp. 84–93, 2013.