Flight Fare Prediction Using Machine Learning

by

Md. Shaim Hosan Noyon
16101150
Tanzidul Islam
15101090
Solaiman Islam
18141020
Md. Refayet Islam Reejon
16301184

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Shaim Hosan Noyon
16101150

Tanzidul Islam
15101090

Solaiman Islam
18141020

Md. Refayet Islam Reejon
16301184

Approval

The thesis/project titled “Flight Price Prediction Using Machine Learning” submitted by

1. Md. Shaim Hosan Noyon (16101150)
2. Tanzidul Islam (15101090)
3. Solaiman Islam (18141020)
4. Md. Refayet Islam Reejon (16301184)

Of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 18, 2022.

Examining Committee:

Supervisor:
(Member)

Dr. Amitabha Chakrabarty
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr.Md.Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr.Mahbubul Alam Majumdar
Dean
Department of Computer Science and Engineering
Brac University

Abstract

This paper deals with the forecast of Flight Price of domestic airlines.Revenue management relies heavily on forecasting. Air passengers (buyers) frequently search for the ideal time to buy tickets in order to save as much money as possible, whilst airlines (sellers) constantly strive to maximize their profits by adjusting different rates for the same service. For flying tickets, the airline uses dynamic pricing. Flight ticket costs fluctuate throughout the day, especially in the morning and evening. It also varies according to the holidays or festival season. The cost of a plane ticket is determined by a number of distinct factors. A lot of factors influence the cost of an airline ticket, including the location of source and destination, purchase time, number of stoppage, and so on. The sellers have all of the information they need (such as past sales, market demand, consumer profile, and behavior) to decide whether to raise or lower airfares at various times leading up to departure dates. Buyers, on the other hand, have limited access to information, which is insufficient to anticipate flight costs. It will offer the optimum time to buy the ticket based on parameters such as departure Date, Arrival Date, Source, Destination, Stoppage and Airline Name. To use Machine Learning (ML) models, features are retrieved from the gathered data. Then, using this data, we want to create a system that will assist consumers in deciding whether or not to purchase a ticket. Extracted features of a typical domestic flight of a year are taken as data and other conditions that may affect the flight is taken into consideration. The information is applied to machine learning models to predict flight ticket prices which uses the XGBoost algorithm that has given us 84.46% accuracy of prediction of the output price. We selected XGBoost as our chosen model after analyzing and visualizing 6 different Regressor models.

Keywords: Machine Learning; XGBoost;Price Prediction; Air Fare; Regessor Models ;PCA

Dedication

Dedication to our parents, teacher, friends, relatives and all who loved us for all their love and inspiration. Special thanks to our supervisor for support us through the whole journey.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, this work was carried out in the Department of Computer Science and Engineering at Brac University. We would like to express my deepest gratitude to our supervisor, Dr. Amitabha Chakrabarty, Associate Professor, Dept. of CSE, Brac University who gave us this opportunity, initiated us into the field of Machine Learning, and without whom this work would not have been possible. His encouragements, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of our BSc study were simply appreciating and essential. His ability to muddle us enough to finally answer our question corrected is something valuable what we have learned and we would try to emulate, if we ever get the opportunity. We would like to thank all the Faculty members. Department of CSE, Brac University for their valuable time spent in requirements analysis and evaluation of the thesis work. We would like to express our sincere and warm gratitude to all those who have encouraged us directly, provided mental encouragement and criticized our work in several phases during the development of this thesis and for preparing this thesis indirectly.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

In recent years of fast-paced technology, air travel has increased significantly. Thus, resulting in the rapid growth of air traffic. This has caused concerns for passengers as well as airports and air travel agencies for the extra expenses[1][2] . These concerns also occur from other factors such as departure source, destination, number of stops, holiday season, model of the aircraft, and expense of agencies[3] . The average rate for air carriers in many countries has seen unexpected ups and downs. The gathered datasets will contain source, destination,schedule of flight, airfare data for domestic flights by major air carriers. The chosen Machine learning model will consist of three states- training state, testing state, and prediction state[4] . As Flight price surges are a major source of frustration in the present aviation industry, it is necessary to develop methods for analyzing the manner in which the price of an airline ticket is fixed. Traditional approaches are insufficient for the job at hand. In order to enhance prediction performance, an optimum feature selection approach is proposed, and it is proven to have superior performance when compared to directly employing all of the features from all of the datasets available.By training and testing a selected large dataset from a trusted resource, a machine learning model can predict the flight price of any selected domestic airline excursion to avoid any sort of inconvenience[5] .

1.2 Research Objective

The objective of our research is to analyze the Machine Learning Models related to Airlines price prediction, find the best model that can provide that prediction with the highest possible accuracy, and predict a price by taking related data inputs using the suggested model. The attributes of given datasheets of a few Indian Airlines can be sorted into a normalized datasheet for convenience of both passengers and airlines companies by acknowledging them beforehand about the inflation or deflation of airlines ticket prices

The objectives of this research are:

1. To deeply understand Machine Learning Models and how it works.

2. To deeply understand Prediction Algorithms and find their implementation on Domestic Airlines.

3. To evaluate the models based on their accuracy and comparative features.

4. To develop a model with maximum accuracy for the prediction of airfare.

5. To further improve flight price prediction models for the betterment of the society of a country.

1.3 Problem Analysis and Suggested Solution

In summary, the findings demonstrate the value of using probabilistic forecasting of price to optimize customer purchasing. Image identification, voice recognition, machine translation, and other machine learning applications have all seen substantial improvements because of deep learning. Many attempts have been made to apply deep learning algorithms to data analysis issues with large data, including traffic flow prediction, as a result of the paradigm's enormous success. However, close to no effort has been made to use machine learning algorithms to analyze air transport data. An accurate and resilient prediction model has been developed by integrating several models based on the machine learning paradigm, allowing for an in-depth study of the patterns in air traffic price drop and surge. The suggested XGBoost prediction model's accuracy was tested, evaluated, and compared to five other prediction techniques. It outperforms all other techniques in terms of accuracy. Finally, the suggested algorithm was used to simulate day-to-day sequences of departure and arrival aircraft timings at different airports along with number of times the airline stops and the company of the airline.

Chapter 2

Related Work

2.1 Flight cost indication

There are many reasons that may cause the cost index (fuel and time-related expenses of airlines) to diverge from its intended direction. Daily and absolute spot prices, speed, location, and an airline's hedging strategy are some of the variables that may cause fuel costs to vary. Furthermore, some expenses may manifest as hourly staff and maintenance expenditures. Because of the tremendous complexity of the pricing algorithms used by airlines, it is extremely difficult for a customer to acquire an air ticket at a low price, as the price fluctuates constantly. As a result, various algorithms [3], [4], capable of providing the customer to acquire an airline ticket by predicting the airfare, have recently proposed. The majority of these strategies rely on advanced prediction models developed in the Machine Learning branch of computational intelligence research (ML). The first stream of literature is concerned with airline cost indexing. Edwards et al.[6] presented a more current study on the environmental advantages of improving CI for various aircraft types. According to their findings, optimum CI values vary greatly depending on aircraft type and flight distance, with long-haul trips having a greater effect on CI optimization.

2.2 Machine Learning

Machine learning (ML) is a technique for instructing computers on how to better handle large amounts of data. Even after looking at the data, we may be unable to decipher its meaning. That's when machine learning comes in handy. Machine learning is becoming more popular as a result of the influx of new datasets. Machine learning is used in a wide range of sectors to identify important data. Learning from data is the goal of machine learning. How to make robots learn without being explicitly programmed has been studied extensively. Many mathematicians and computer programmers use a variety of ways to solve this challenge, which involves dealing with large data sets.

Different algorithms are used in Machine Learning to tackle data challenges. Many data scientists insist that there isn't a single algorithm that can fix every issue. There are a variety of algorithms that may be used depending on the issue you are trying to answer, how many variables you have, and so on[8] .

When it comes to figuring out how to accurately anticipate the cost of travel using machine learning, there are two options. In the first technique, flight ticket prices are predicted as a regression issue, but in the second, they are classified. Typically, regression models are used to estimate a function that explains the mapping law between data attributes and airfare prices, and the former technique is used to anticipate the precise price of an air ticket. The latter technique can't anticipate the precise price of a plane ticket, but it can help you decide whether or not to purchase a ticket at that price. As little attention has been dedicated to assessing the state-of-the-art regression ML models for this issue, this study examines the first regression instance of flight price prediction[6] .

Regression is a supervised learning strategy that uses a backwards-looking technique. Continuous variables may be modelled and predictions can be made using it. The following are some examples of how linear regression algorithm may be used: predicting real estate prices, predicting sales, predicting student test results, and forecasting stock exchange price changes. The output variable value in regression is decided by the values of the input variables, making it a supervised learning strategy. A straight line (straight hyperplane) is the simplest type of regression and may be used when the connection between the variables in a data set is linear. Linear regression has the virtue of being simple to grasp and of being able to prevent overfitting by using regularization. SGD may be used to incorporate fresh data into linear models. As long as the connection between the variables is linear, Linear Regression is a solid choice. Statistical modeling moves to data analysis and preprocessing. It is easy to understand the basics of data analysis by using linear regression. However, since it simplifies complex real-world issues, this approach is not recommended for most practical applications. Linear regression predicts a link between the means of the dependent and independent variables. This relationship may not exist.

For example, Decision Trees may be used to solve regression and classification problems, as well as filling in missing values in attributes with the most likely value. They are also very efficient because of their efficient tree traversal technique. Using Random Forest, Random Forest solves the issue of over-fitting in Decision Trees by using ensemble modeling. Using a decision tree has a number of drawbacks, including the potential for instability, a lack of control over the size of the tree, sampling error, and a locally optimum answer rather than a globally optimal one. As an example, Decision Trees may be utilized in library book prediction and tumor diagnosis applications.

It is a classification method known as the K Nearest Neighbor (KNN) Algorithm. Data points are organized into classes in the database, and the algorithm attempts to categorize a sample data point supplied to it. KNN is non-parametric since it does not presume that the data is distributed in a certain way. The KNN algorithm has the following advantages: It's a straightforward method that may be put to use right away. It's easy to make a model. It's a very adaptable categorization technique, making it ideal for courses with a variety of modality types. There are several class labels on the records. At most, it's two times as high as the Bayesian error rate. It's possible that it's the best way to go. When predicting protein function from expression patterns, KNN beat SVM. KNN has the following drawbacks: An unknown record's classification is a time-consuming and costly process. It needs the calculation of the distances between the k-next neighbors. The method becomes more computationally costly as the number of training sets grows. Inaccurate features will have a negative impact on the algorithm's accuracy. It's a slack learner that measures distance between itself and k nearby objects. It does not generalize the training data in any way and retains every single one of them. As a result, it performs costly calculations on big amounts of data. The accuracy of areas will deteriorate as data becomes more multidimensional. For example, KNN can be used in recommendation systems, medical diagnosis of multiple diseases showing similar symptoms, credit rating using feature similarity; handwriting detection; analysis done by financial institutions before sanctioning loans; video recognition; forecasting votes for different political parties and image recognition; and so on[9] .

When it comes to using clever pricing techniques, the airline sector is often regarded as a leader. These days, even for seats next to one other on a plane, the cost of a ticket may change dramatically and dynamically[10][11] . Flight ticket prices might fluctuate as much as seven times a day[10]. Customers are always looking for ways to save money on their plane tickets, while airlines are always looking for new ways to increase revenue and profits. Mismatches between the number of seats available and passenger demand frequently result in higher prices for customers or a loss of income for the airline. Most airlines are well-equipped to manage their own pricing, thanks to a variety of sophisticated technologies and resources at their disposal. Customers, on the other hand, are becoming more strategic with the emergence of numerous internet tools that allow them to compare rates across different airlines[12]. In addition, the fierce rivalry amongst airlines makes it difficult for everyone to choose the best price.

Increased focus on both consumers and airlines over the last two decades has resulted in an increase in research. In contrast to customer-side research, airline-side research focuses on enhancing airline income. Techniques used in research include regression and sophisticated data mining techniques, as well as statistical methods.

As a client, finding the lowest price or the ideal time to purchase a ticket is the most important consideration. No longer does the "tickets purchased in advance are less expensive" idea hold true[16] . Customers who purchase tickets in advance may end up paying more than those who get tickets at a later date. Furthermore, making an early purchase puts you at danger of being locked into a timetable that must later be altered, generally at a cost. As a result, the price of a ticket might fluctuate constantly. Several studies have been carried out to assist customers in deciding the best time to acquire tickets and the best way to estimate the cost of those tickets[14][15][16][6][12][18][19][20][21][22][23][24] . The majority of customer-side research concentrate on forecasting the best moment to buy a ticket using statistical techniques. It is more difficult to estimate the actual ticket price than the best time to buy a ticket, because of issues such as lack of data, external influences, dynamic ticket pricing, competition among airlines, and the proprietary nature of airlines' ticket pricing rules[26]. Though few research have attempted to anticipate ticket prices by comparing their work to that of the authors, there are several studies that have tried[24][14][16][6][18][19].

The primary objective of airlines is to increase revenue and profit. The impact of production conditions on ticket prices is described by the term ”yield pricing,” and airlines use a variety of pricing strategies, including dynamic pricing and long-term pricing policies to determine optimal ticket prices, according to Narangajavana et al.[21] Dynamic pricing involves the dynamic adjustment of ticket prices in response to various influencing factors. It is difficult to foresee price changes based on long-term pricing plans and yield pricing since they are tied to the internal workings of the airline[12] . A more accurate prediction of ticket prices is possible with dynamic pricing, which takes into account dynamic elements including demand fluctuations and price discrimination. It is difficult to implement dynamic pricing because of a wide range of circumstances, including internal and external factors, competition, and key customers, among others. A variety of criteria, such as historical ticket pricing data, ticket purchase date and departure date as well as season, holidays, supply (number of airlines and flights), fare class, availability of seats, current market demand, and trip distance, are considered internal factors. Retailers and customers compete for profit in dynamic pricing, which may be seen as a game of chess[26] . As many tickets as possible are being sold at the maximum price in order to boost airline profit margins. As a consequence, tickets must be sold within a certain period of time to avoid further losses due to unsold seats. While buyers are eager to get the best deal possible, they keep an eye on plane ticket costs to see if they fall. Customers come and go at random, therefore the supply and demand may fluctuate. Because of this, airlines must constantly modify ticket pricing depending on the current demand, consumer behavior, ticket prices provided by rivals in the market, and other internal and external variables in order to become profitable in such complicated circumstances[27][28] .Dynamic pricing refers to the dynamic change of ticket prices in response to numerous influences. It has been shown that dynamic pricing is one of the most popular pricing tactics used by airlines. However, they don’t go into detail on the various prediction methodologies that are used for dynamic pricing implementation[28][26][27] .

Every one of the researches presented in the preceding section offered a model that predicted the best time for people to buy a product or service. Predicting flight pricing in real time was not an option. Y. Chen et al.[24] developed a technique to forecast the cheapest itinerary based on this knowledge gap (a specific flight on a given route for a particular departure date).

According to Anastasia Lantseva et al.[14] , a ticket prediction model was developed using an empirical data-driven R model. Within 90 days before the departure date, the model forecasts the cost per kilometer of a certain aircraft. In the spring of 2015, two independent ticket price information aggregators (AviaSales and Sabre) were used to gather data on the prices of domestic and international flights. Flights from Moscow and St. Petersburg to 50 Russian towns were utilized for local flights. For international flights, the dominance of European cities was examined for flights from the same two cities (Moscow and Saint-Petersburg).

To estimate the lowest price available on a certain route for the days between the purchase date and a specified departure date, T. Liu et al[28] have developed an ensemble regression technique similar to that described by Y. Chen et al[24] . K-Nearest Neighbors, Random Forest and Bayesian are used as the basic learners and feature clustering is used to create the ensemble learning model. Ticket pricing from the past, a signal indicating whether or not the departure date falls on a holiday, and how many days remain before departure are all factors taken into account by the model.

Chapter 3

Prediction Modeling And Methodology

3.1 Model classification

A major task of machine learning algorithms is to define objects and gain the ability to categorise them. This helps to separate large amount of data into discrete values. Classification predictive modeling algorithms are analyzed depending on their results.

3.1.1 Extreme Gradient Boosting

Extreme Gradient Boosting is an effective implementation of gradient boosting models. Regression prediction models involve predicting numerical values, for example - price, height, etc. Gradient boosting points to a cluster of machine learning algorithms that can be used for regression predictive modeling problems. The main reasons to use XGBoost here are it's model performance, execution speed and ability to predict our preferred output which is flight price.

$$L(\phi) = \sum_{i} l(\widehat{y}_i, y_i) + \sum_{k} \Omega(f_k) \tag{3.1}$$

Where,

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \tag{3.2}$$

3.1.2 Random Forest Algorithm

Random forest is a machine learning approach that is versatile and simple to utilize that consistently provides excellent results even without adjusting hyper-parameters. It's simple and diverse which can be used for regression and classification.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2 \tag{3.3}$$

Where N is the number of data points, fi is the value returned by the model and yi is the actual value for data point i.

3.1.3 Gradient Boosting Algorithm

The Gradient Boosting Algorithm is a powerful tool for minimising bias and variance errors. This algorithm can be used to predict not only continuous target variable as a regressor, it can also categorically target variable as a classifier. Gradient boosting trains models in a manner that is additive, gradual and sequential. It allows to optimise specified cost function based on user input.

$$y_i^p = y_i^p + \alpha * \delta \sum_p (y_i - y_i^p)^2 / \delta y_i^p \tag{3.4}$$

Which Becomes,

$$y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p) \tag{3.5}$$

3.1.4 K-Nearest Neighbors Algorithm

The KNN algorithm uses similarity of features to predict the values of all new data points so that new points will be the assigned values based on their similarity to the points in the training set. Initially, the distance of the new point and each training point is found out. Then, the closest k data points are picked based on the calculated distance. Lastly, the data points are averaged to get the output of the final prediction for the required point.Lastly, the data points are averaged to get the output of the final prediction for the required point.

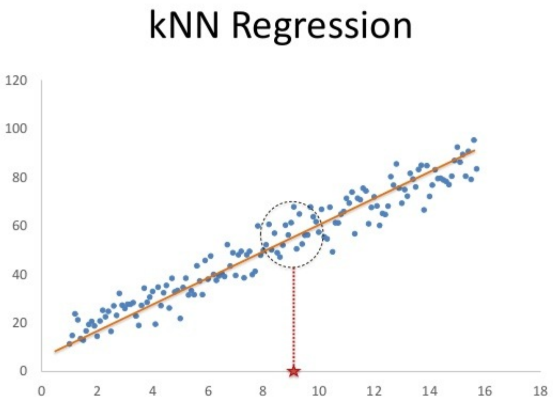


Figure 3.1: K-Nearest Neighbors Algorithm

3.1.5 Decision Tree Algorithm

Decision Tree Regression calculates regression models in a tree-like structure. It splits up a dataset into smaller subsets whereas simultaneously, an associated decision tree is increasingly developed. The output is a tree with decision nodes and leaf nodes. A decision node has multiple branches, each for the values for the tested attribute. Leaf node represents the decision on a target that is numerical. The peak decision node of a tree which indicates to the best predictor is the root node. Decision trees have the ability to handle both categorical and numerical data.

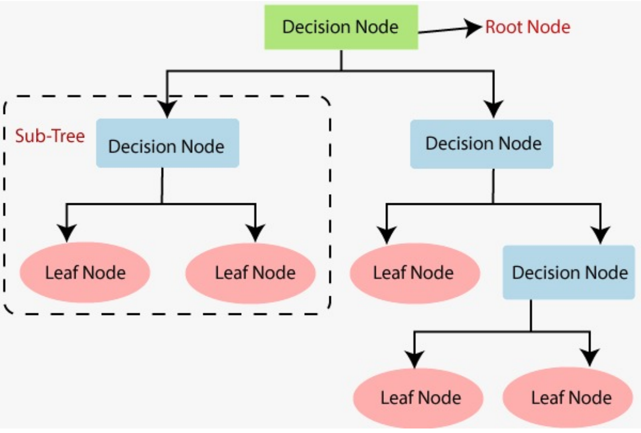


Figure 3.2: Decision Tree Regression

3.1.6 Linear Regression Algorithm

Linear regression is a statistical method which is mostly used for predictive analysis. It makes predictions for continuous or numeric variables. It is such an algorithm that shows a linear relationship between a dependent (y) and one or multiple independent (x) variables, thus known as linear regression. It determines the change in the dependent variable according to the independent variable, which results in a sloped straight line showing the relationship between the variables.

$$y = a_0 + a_1 x + \epsilon \tag{3.6}$$

Here,
Y= Dependent Variable
x= Independent Variable
a0= Intercept of the Line
a1= Linear Regression Coefficient
Epsilon =random error

The values for x and y variables are training datasets for Linear Regression model representation.

3.2 Principal Component Analysis (PCA)

PCA can be used on its own, or it can work as a method of data cleaning or data pre-processing technique before machine learning algorithms. It is used for many purposes, such as:

1. Visualize multidimensional data.
2. Compress information.
3. Simplify complex business decisions.
4. Clarify convoluted scientific processes.

For data pre-processing, PCA is applied to:

1. Reduce the dimensions of a training dataset.
2. De-noise a dataset as it is calculated by finding the attributes that describe maximum variance and it takes the signal in the data and ignores the noise.

PCA has been implemented to all the mentioned models to analyze the impact it has on the accuracy of the predicted result. The after effect stood out to be a result of decreasing percentage of accuracy for all models. After careful analysis, it has been observed that, as PCA reduces the dimensions of our trained dataset, only keeping the dimensions including the most unique values, the prediction function has more deviated and lesser data to be trained with. Thus getting a smaller percentage of accuracy.

3.3 Methodology

In our thesis, Linear Regression, Random Forest Regressor, Decision Tree Regressor, Gradient Boosting Regressor and XGBoost Regressor contributed to our work heavily.

Data collection is a very crucial part of our work. As Bangladesh is a small country with few numbers of airports and a low frequency of domestic flights occur every day, the daily airfare data is very low. Thus, public data of flight price is scarce. For this reason, most of our input data are from foreign airlines. We analyzed as many as possible data sets about Domestic airfare of India which are open to the public.

After collecting data and fulfilling our target, we preprocessed the data into supervised data. There were many complexities during the data sorting due to having different forms of attributes and various types of information. Processes had to be carried out to tackle this efficiently and to improvise the data set. This included reducing dimensions and cleaning of data to help us in better classification.

Data Preprocessing in our model includes Feature Engineering (Cleaning Data), Dataset Input, Splitting Data, Dataset Train and Test and Classify Models.After

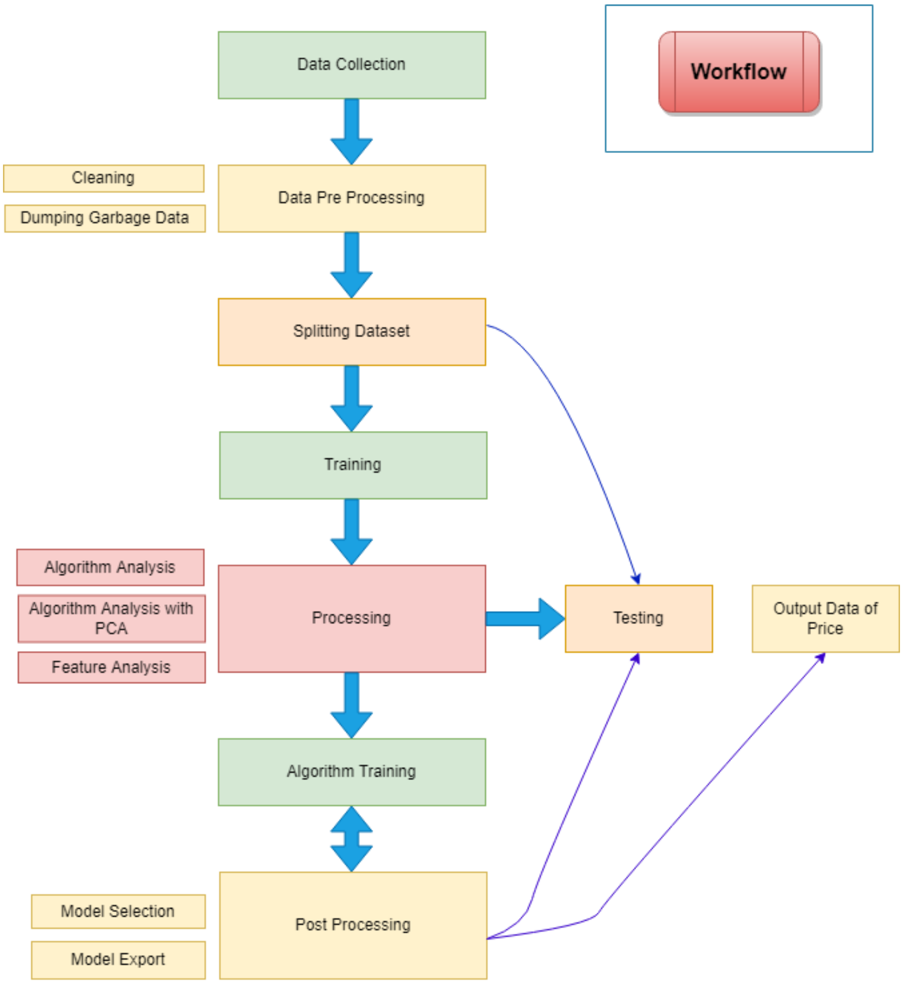


Figure 3.3: Work Flow

classifying the data, we had to train and test it to find a suitable algorithm. This showed us a different percentage of accuracy for each of the models.

After that we processed the data through 6 models and found their accuracy, Analysed and Visualized each model, Implemented deployment of a webpage with basic HTML and CSS in Visual Studio and integrated it with Heroku which contains basic input boxes- Departure Date, Arrival Date, Source, Destination, Stoppage, Airline Name.

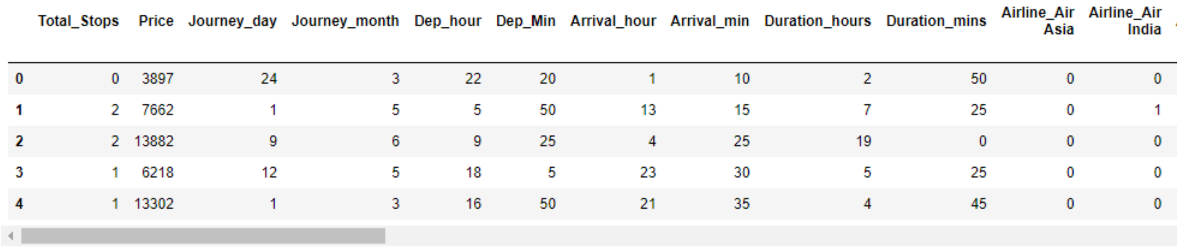
To classify the model, the attributes that did not affect the models much were disregarded in the data pre-processing portion to get improved accuracy. The model for which the attributes show the most percentage of accuracy was classified as the chosen model for our prediction algorithm [6].

3.4 Data Cleaning

Data cleaning is an extremely important step in any machine learning applications. In a large amount of data, there are different kinds of statistical analysis and data visualization methods that can be used to examine data. There are fundamental operations to be performed in any machine learning method in which data cleaning or reducing dimensions is a crucial portion.

In our paper, for data cleaning, we reduce the dimensions of the given dataset with PCA algorithm. Richard Bellman cited ”Many algorithms that work fine in low dimensions become intractable when the input is high-dimensional”. High Dimensionality is when a dataset has many features. This creates a problem of over-fitting a model reducing the ability to generalizing other datasets except the training dataset. Principal Component Analysis is an unsupervised and non-parametric statistical method usually used for reducing dimensionality in Machine learning applications.

The conclusion of data cleaning turns out as - defining and removal of column variables with only one single value, identifying and considering column variables of low amount of unique variables and removing rows that have duplicate values or null values.



	Total_Stops	Price	Journey_day	Journey_month	Dep_hour	Dep_Min	Arrival_hour	Arrival_min	Duration_hours	Duration_mins	Airline_Air Asia	Airline_Air India
0	0	3897	24	3	22	20	1	10	2	50	0	0
1	2	7662	1	5	5	50	13	15	7	25	0	1
2	2	13882	9	6	9	25	4	25	19	0	0	0
3	1	6218	12	5	18	5	23	30	5	25	0	0
4	1	13302	1	3	16	50	21	35	4	45	0	0

Figure 3.4: First 12 X 5 Out of 33 X 10682 of the Cleaned Datasets

3.5 Input data

First of all, and at the beginning of the phase, it is necessary to define the inputs of the model on which the model learns and leads to the final structure. The dataset used to evaluate the model is derived from historical data containing flight price data for 1 year. Then the data was cleaned and converted into a ”pickle” file for the purpose of integrating into our web page.

3.6 Data Training and Testing

We split the dataset into 2 parts in 70/30 ratio — training and testing.
The training set contains the features, along with the prices of the flights. It contains 7477 records, 32 input features, and 1 output column which is ‘Price’. The test set contains 3205 records and 32 input features. The output ‘Price’ column needs to be predicted in this set. We will use Regression techniques here since the predicted

output will be a continuous value.

Following is the features available in the dataset after cleaning-

1. Total_stops,
2. Journey_day,
3. Journey_month,
4. Dep_hour,
5. Dep_Min destination,
6. Arrival_hour,
7. Arrival_min,
8. Duration_hour,
9. Duration_mins,
10. Air_Asia,
11. Air_India,
12. GoAir,
13. IndiGo,
14. Jet_Airways,
15. Jet_Airways_Business,
16. Multiple_carriers,
17. Multiple_carriers_Premium_economy,
18. SpiceJet,
19. Trujet,
20. Vistara,
21. Vistara_Premium_economy,
22. Source_Bengaluru,
23. Source_Chennai,
24. Source_Delhi,
25. Source_Kolkata,
26. Source_Mumbai,
27. d_Bengaluru,

28. d_Cochin,
29. d_Delhi,
30. d_Hyderabad,
31. d_Kolkata,
32. d_New_Delhi.

Chapter 4

Analysis and Visualization

4.1 Analysis and Visualization with Graphs

Analysis and visualization is an important portion of any machine learning model of a project as it determines a clear view of which model and method should be prioritized. It provides an overall explanation of the what and why of any paper. Also, we are comparing the features with each other to analyze the cleaned dataset to get a clear visualization of our data.

4.1.1 Counts of Flight with Different Airlines

On figure 4.1, we try to count a specific airline's sum of flight inside our database, we get the exact value of each and every airline. Therefore, we now represent our findings below from largest no of flight to the smallest no of flight.

1. Jet airways has the largest count of flights, about 3849.
2. Indigo airlines has the Second largest count of 2053.
3. The third largest airline is Air India with 1751 number of flights.
4. Multiple carriers have 1196 no as counts of flight.

The rests of the airlines count of flights are like SpiceJet (818), Vistara (479), AirAsia (319), GoAir (194), Multiple carriers premium economy (13), Jet Airways Business (6), Vistara Premium economy (3) and TruJet (1).
We showed the comparison of Count of flight vs Different Airlines:

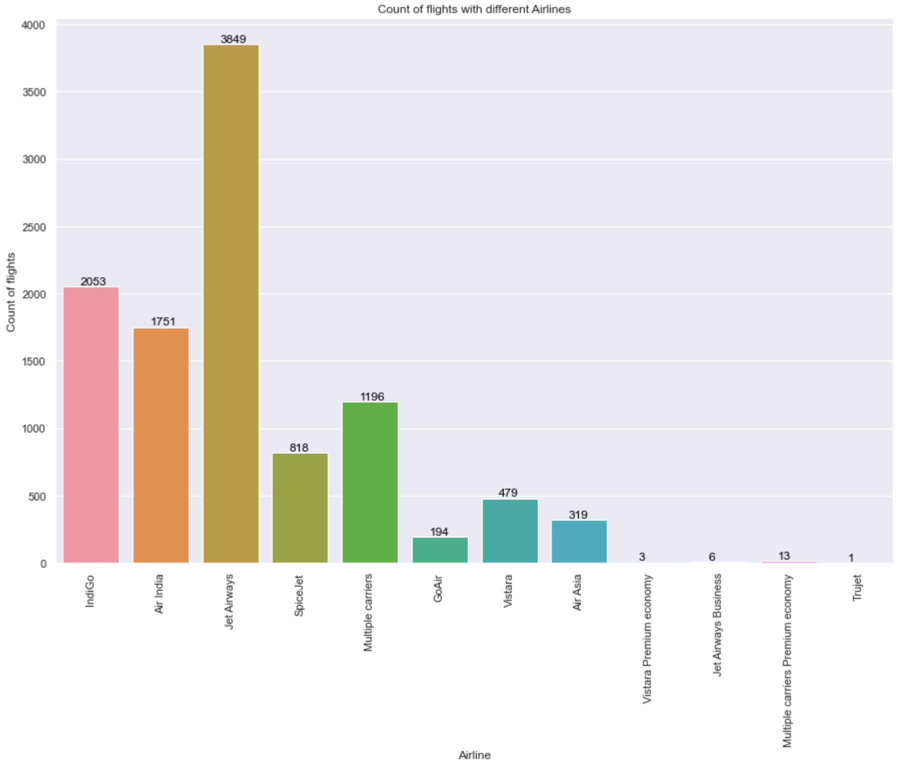


Figure 4.1: The Number of Flight Vs Different Airlines

4.1.2 Airlines Vs Price

On figure 4.2 we see that different airlines have different boundaries of values as price. Jet Airways Business flight prices lie mostly between 50000+Rs to 60000+Rs. Jet Airways, Multiple carriers and Air India have similar flight prices mostly in between 1.5k Rs. to more or less 30k Rs. Rests of the airlines (instance for SpiceJet, Indigo, GoAir etc.) flight price we figure out in between 1.5k Rs to 20k Rs. Here, Jet Airways flight price is much higher than all of the others as the price is for business class tickets. So, if we separate this from the rest we can say that their offered (flight) price factor is less indifferent to each other.

We showed the comparison of Price of flight vs Different Airlines:

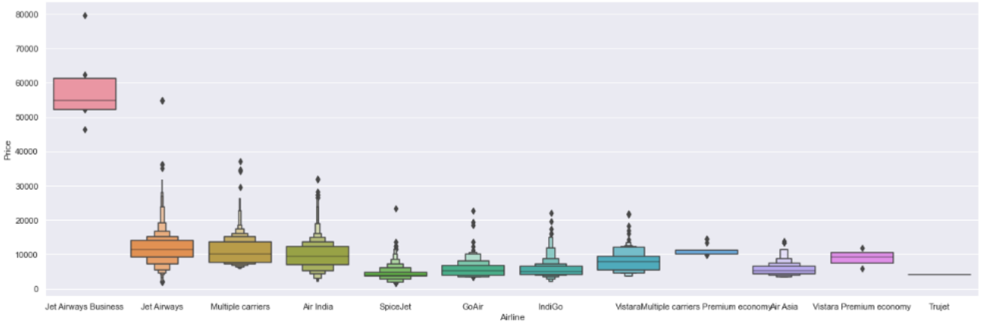


Figure 4.2: Airlines Vs price

4.1.3 Price Vs Total Stops

On figure 4.3, we can see that the price factor of flight has significantly changed in terms of the no of stoppage. Such as, the flights which travel from source port to destination port in just one stop, are the expensive one, though it's density as count is less in number. It stays true for all other flights which have no stop Greater than one. However, for non-stop flights we do not see this kind behavior, rather we see a reasonable amount as flight price.

We showed the comparison of Price of flight vs the number of stoppage:

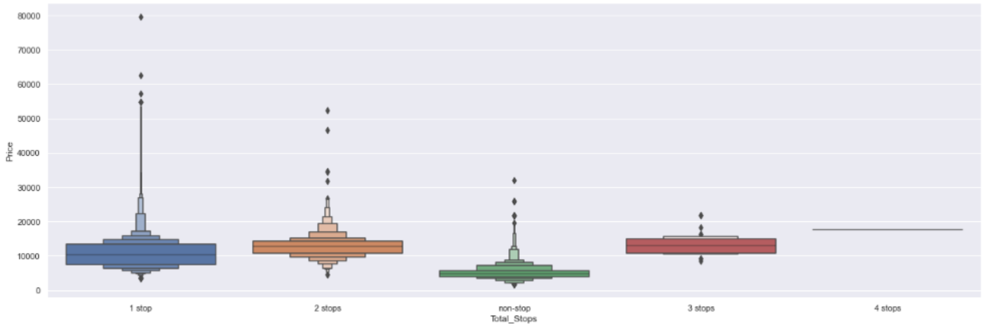


Figure 4.3: Price Vs Total Stops

4.1.4 Airlines Vs Total Stops

On figure 4.4, we can see which airlines have how many types of stops for each flight. To describe more, Air India has all 4 types of stoppage criteria, such as, 1, 2, 3 and no stop. Air Asia, Jet Airways and Indigo have three types of stoppage except 3 stop service. Multiple carriers also have three different stops except nonstop service. The rest of the airlines have two types of stoppage service varying one to three except the least three airlines which we showed before depending on flight counts (on fig 4.1).

We showed the comparison of the number of stoppage for different airlines:

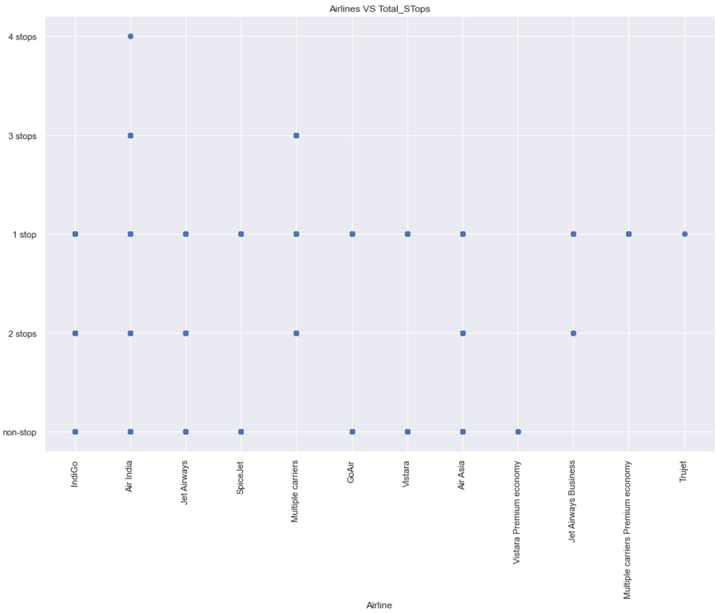


Figure 4.4: Airlines Vs Total Stops

4.1.5 Airlines Vs Total Stops According to Month

Each airlines have different number of stoppages starting from the number 0 to 4, which varies depending on which month the flight is taking off. As we can observe Jet Airways Business, Multiple Carriers Premium Economy and Trujet have only one stop in between its flight period in the month of March(3.0). Vistara Premium Economy have had stoppages in their flights on the month of March(3.0) and April(4.0) . The rest of the airlines have had stoppages during the whole year. We showed the comparison of the number of stoppage for different airlines in Month Wise:

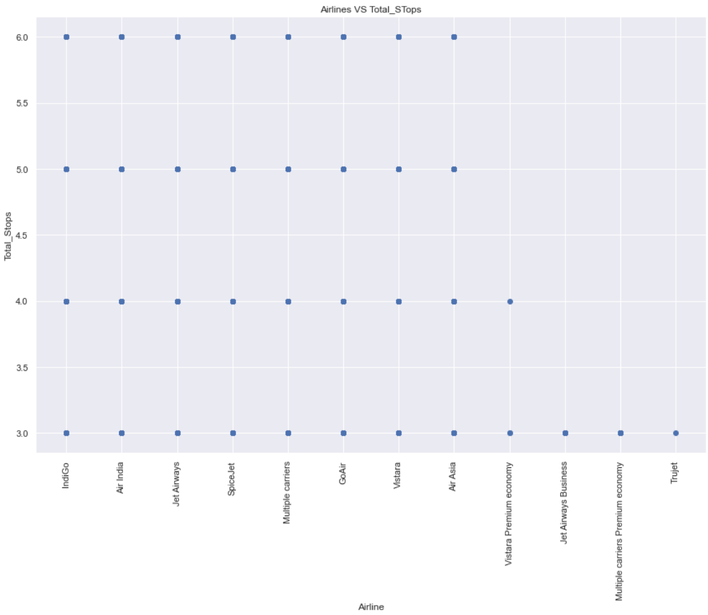


Figure 4.5: Airlines Vs Per Month Total Stops

4.1.6 Price Vs Source

There are only five airports. To begin with we can say Bangalore has the highest range of flight prices from lowest to highest. After that, comes Delhi and Mumbai on serial 2nd and 3rd. To dig deeper we can describe that all of these source's flight prices are extremely competitive to each other in between low (1.5k to ¡25k) range mostly. On the contrary, we see extreme differences in flight price surrounding the higher (25k¡) range but again they are as little as to ignore.

We showed the comparison of Price of flight vs Source of Flight:

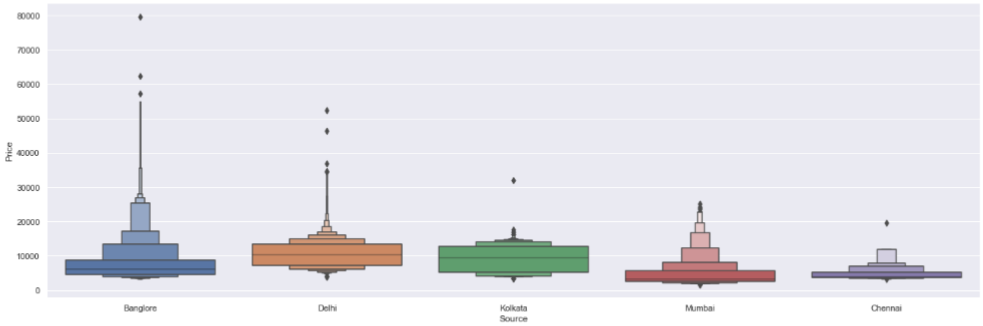


Figure 4.6: Price Vs Source

4.1.7 Price Vs Destination

Here we find 6 airports as destinations. This in nature of influence is as indifferent as we have seen in price vs source graph. The only difference we find for each destination port is in its numeral value range in comparison with flight price; For example, New Delhi takes position for the highest flight price with the highest range of price. Moreover, Cochin and Bangalore have shown similar competitive pricing between them. Hyderabad, Kolkata and Delhi (old) take the spot for mostly having cheap pricing as destinations. To conclude, all of these destination airport's flight prices are extremely competitive to each other in the 1.5k to less than 25k+ range.

We showed the comparison of Price of flight vs destination of Flight:

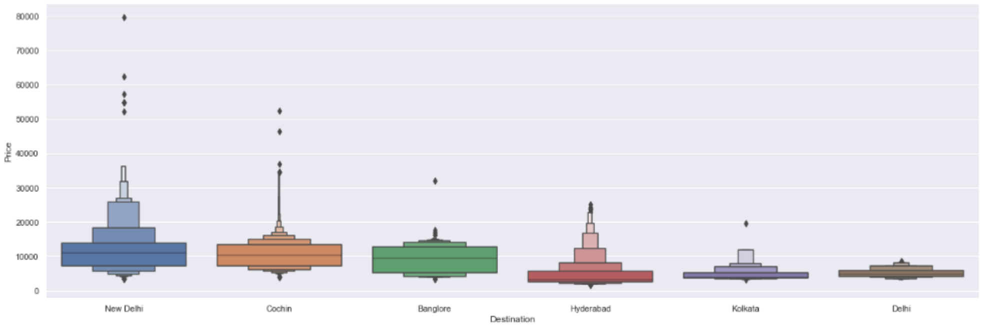


Figure 4.7: Price Vs Destination

4.1.8 Price Vs Journey_Month

In the price of airfare fluctuates depending on which month it is. This happens usually because of the holiday seasons or any rush seasons. As we can see in the month of March(3) the price has seen a surge. On the other hand in April the price is comparatively low.For the month we only show March (03) to June (06) month's flight pricing as an instance. Therefore, we understand by analyzing that March is the most favorable month where not only flight's count increased extremely but also our target variable price also increased rapidly in comparison to the other three month. May and June show quite identical factors in terms of flight pricing. April's flight pricing is comparatively less in terms of the rest. That is why in April, the pricing of each flight generally lies within the base level limit.
We showed the comparison of Price of flight vs Month of flight:

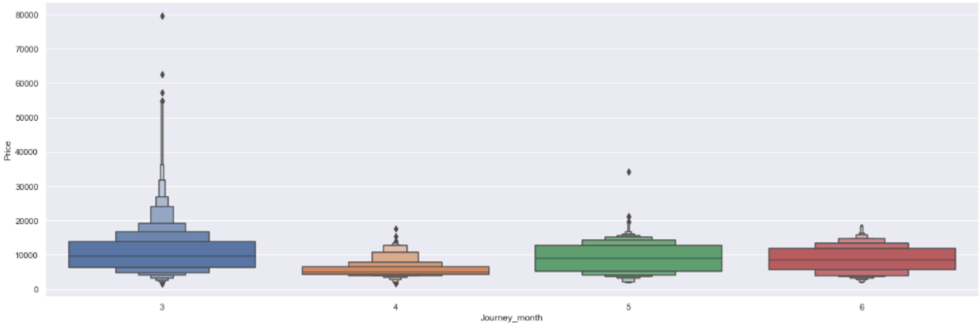


Figure 4.8: Price Vs Journey_Month

Chapter 5

Model Selection & Validation

5.1 Model Based Feature Analysis

We analysed the Six Models First and found out the important features for 3 best models. Among the features for decision tree we found that it gave the best output with the duration input. For the rest of the inputs such as- Journey_day, Journey_month, Arrival_hour, Arrival_min, Dep_min, Dip_hour, Destination, Source and Airline it does not effect the prediction model that much. In this model, Additional_info,Total_stops and Route did not play an important role. They are the less important one for this Model.

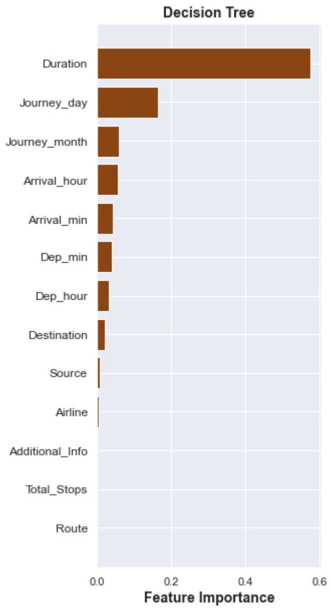


Figure 5.1: Important Features of Decision Tree

Among the features for Random Forest we found that it gave the best output with the duration input. For the rest of the inputs such as- Total_Stops, Journey_month, Journey_day, Source, Arrival_hour, Arrival_min, Dep_min,Airline, Dip_hour, Additional_info, and Destination it does not effect the prediction model that much. In this model, Route did not play an important role. It is the less important one for this Model.

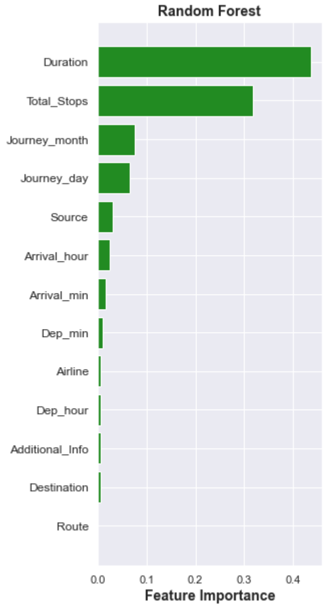


Figure 5.2: Important Features of Random Forest Algorithm

Among the features for Random Forest we found that it gave the best output with the duration input. For the rest of the inputs such as- Source, Journey_day, Airline, Total_Stops, Destination, Arrival_hour, Journey_month, Dip_hour, Arrival_min, Dep_min, Route, and Additional_info it does not effect the prediction model that much. we can see that this model works with every possible features we have in our dataset.

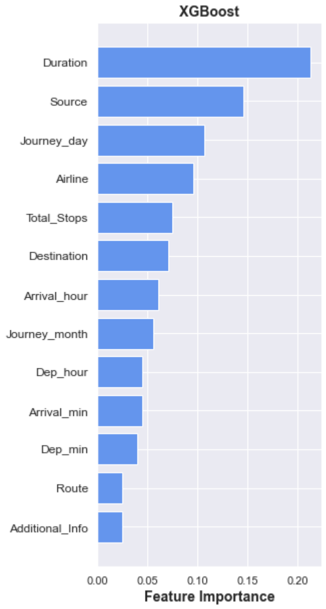


Figure 5.3: Important Features of XGBoost

5.2 Algorithms Selection Based on Predicted Values

As we have already discussed, we have used six algorithms. When we implement our processed clean dataset separately we get accuracy in terms of predicting our target value price. In terms of predicting the accuracy linear regression scores at least about 62%. The main problem of this algorithm is it only works with a less clean dataset. But for our research purpose as we processed our dataset in many branches to get more deep insight, therefore linear regression fails to process effectively. That is why we do not select linear regression as our final algorithm. Then, the second least on line is Decision tree, which scores 77% accuracy. We know that decision trees are effective in getting deep leaf nodes. But the problem arises when it has to run through many branches. Not only it costs more time in processing but also it sometimes gets stuck (In terms of human psychology we say it gets confused) among similar numerical values. That's a main factor for not selecting a decision tree. We would like to talk about the Random forest algorithm next. Random forest scores second highest in our model. The limitation with random forest is it cannot work fast with a larger dataset. In our model, the random forest algorithm also faced a similar problem. Random forest is fast to train, but it is also very slow to create predictions once they are trained. That is why we do not choose this in our model. K-nearest neighbor and Gradient boosting regression algorithms both score 77%, except GBR lead the score by 2%. Since KNN is a distance-based algorithm, the cost of calculating distance between a new point and each existing point is very high which in turn degrades the performance of the algorithm. So, it loses its effectiveness with a large dataset. Moreover, it also does not work well in our dataset as we have a large number of multi-dimensional attributes in our dataset. We do not select KNN algorithms for these two limitations. Now, for GBR we can see that it effectively succeeds in returning a good accuracy score. But to compete with our highest accuracy algorithm, extreme gradient boosting regression, we can easily dump GBR as XGBR is a more advanced technique and shows more effectiveness to find the accuracy in our model.Out of 6 we used the best model which provided the greatest percentage accuracy. In our model we used Linear Regressor, random Forest,Decision tree,KNN,GBR and XGBoost.To find the best model we first run our dataset into this 6 model and find the accuracy in percentage for every algorithm.

After that we selected the best model according to the highest accuracy.And then we complete our prediction with the one with best score.

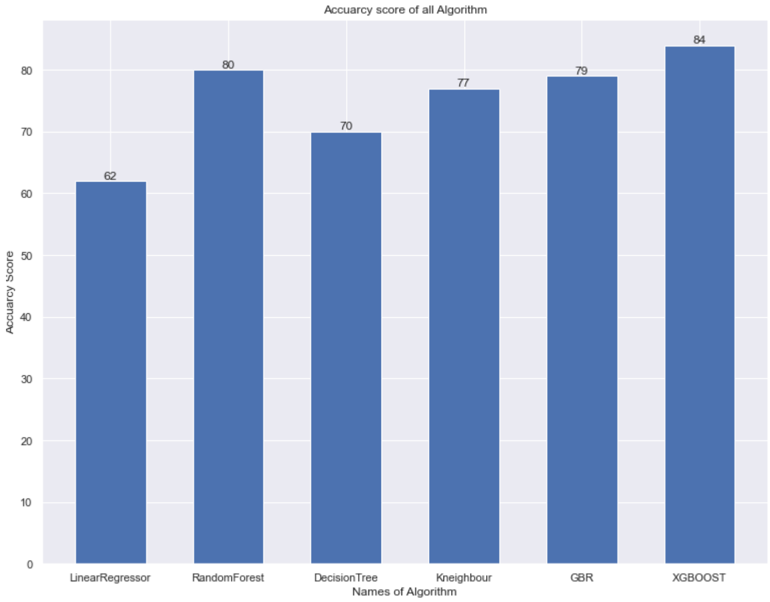


Figure 5.4: Percentage Accuracy Score of Six Algorithms

5.3 Effect of PCA on Our Models

It's a prerequisite for any analysis we may perform on our data. With 13 features, the given dataset has a good dimensional feature space. Euclidean distances become bloated and meaningless in such a high-dimensional space. This might have a significant influence on the performance of our algorithms. This problem is known as the 'Curse of Dimensionality,' since it need more data to train our model.

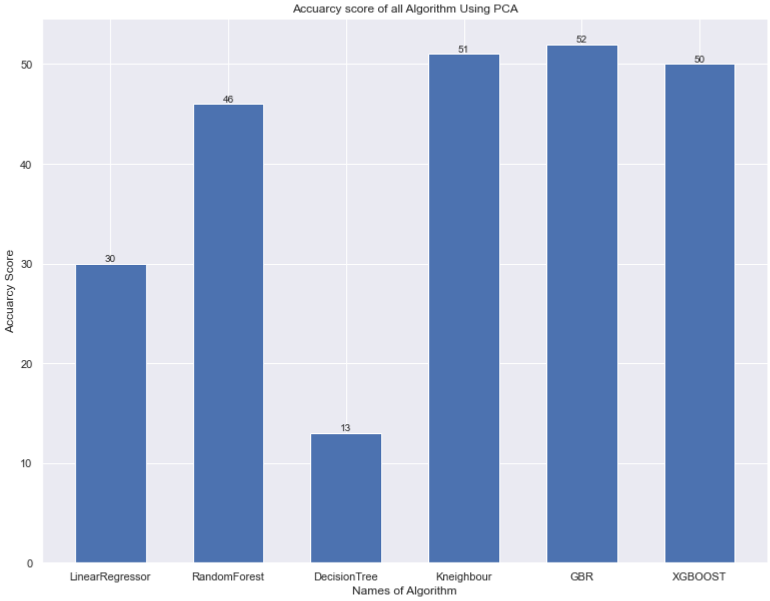


Figure 5.5: Percentage Accuracy Score of Six Algorithms After using PCA

The PCA method handles this problem by identifying the characteristics that account for the greatest amount of variation. As a result, rather of training our models on 32 features, we will train them on two characteristics that explain the most variation.

Because of using PCA we can see that as there was less amount of value to train we got a a low percentage of accuracy.

5.4 XGBoost Validation

The model we select for our prediction is XGBoost. XGBoost basically works with so many trees that divides a data into several branches and makes a data more specific for the machine. which helps a machine to predict more accurately

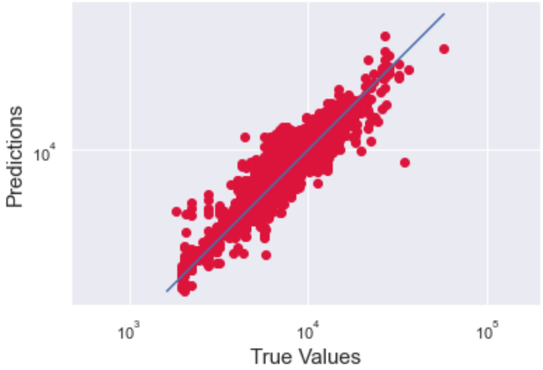


Figure 5.6: Visual Representation of Predictions Vs True Value Graph

From the Fig: 5.6, we can see the predictions values are similar to actual values from the dataset.That means our prediction is much more specific with this model.

In the end, we import a PKL file and Load our algorithm XGBoost. Afterward, we create the UI and import the Model in their.

5.5 Model Deployment

We first got a predicted output data using the ‘’predict” function. Then compared the predicted data with tested data using the ‘’r2_score” function to get a percentage accuracy. Showed the Mean Squared Error, Mean Absolute Error and Root Mean Square Error for data analysis purposes. Showed Predicted Values vs True Values graph for the best 4 algorithms. Compared Accuracy of All Algorithms with Visualization Implemented PCA (Principal Component Analysis) to all the algorithms and showed the accuracy for each of them. In our code, the best accuracy model (84.4597%) showed to be XGBoost Regressor- so saved and loaded the model to a pickle file of the code to further proceed to implement it to our webpage.

Workflow of Model Deployment: The following process are maintained to deploy the ML Model.

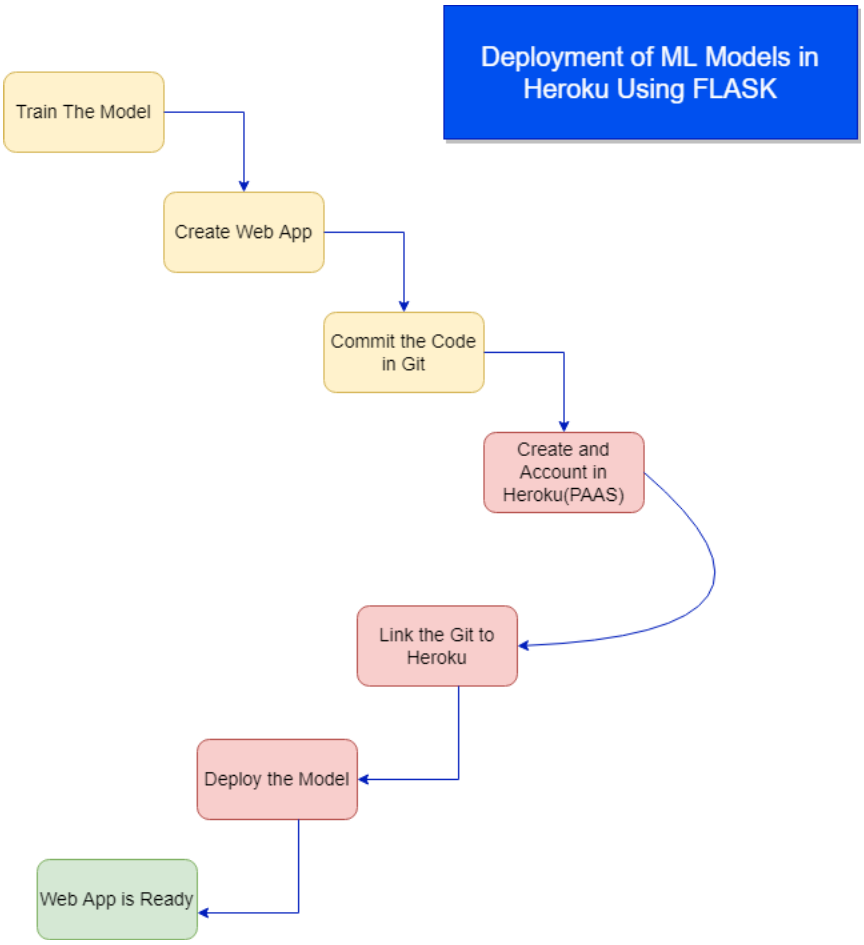


Figure 5.7: Process of Model Deployment

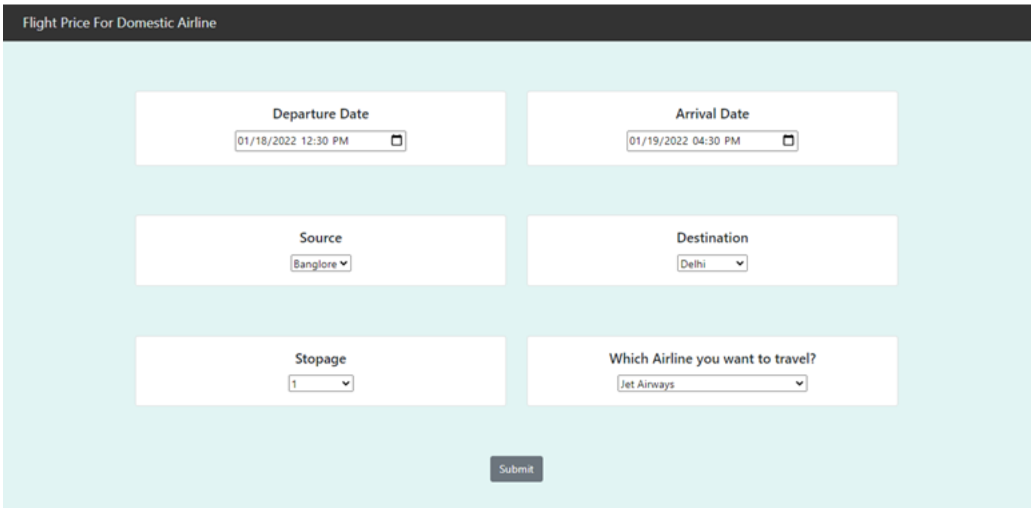
In Visual Studio Code: We implemented the UI of the webpage of Flight Price for Domestic Airlines. We exported the fare5.pkl pickel file, and coded the input-output in form method for the webpage . We included the minimum version required for the code to run.

The requirements are mentioned below:

Flask 2.0.2
Gunicorn 20.1.0
Pandas 1.3.5
Flask.cors 3.0.10
Xgboost 1.5.1
scikit learn 1.0.1

In Heroku: We used Heroku as a platform as a service (PAAS). We commanded in Visual studio through Git to integrate our code in Heroku for the purpose of our free Webpage.

The visual outcome of the deployment is illustrated below:



Flight Price For Domestic Airline

Departure Date: 01/18/2022 12:30 PM

Arrival Date: 01/19/2022 04:30 PM

Source: Banglore

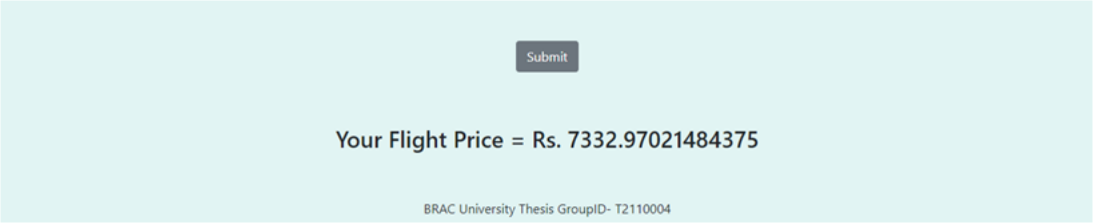
Destination: Delhi

Stopage: 1

Which Airline you want to travel?: Jet Airways

Figure 5.8: Visual Representation of our Web Apps Front Page

We have implemented our thesis through this web page for practical demonstration that the out result of our research is feasible.



Your Flight Price = Rs. 7332.97021484375

BRAC University Thesis GroupID- T2110004

Figure 5.9: Visual Representation of our Web Apps Result

The URL of the Web App of the model Deployment:

https://airfare-prediction-model.herokuapp.com/

Chapter 6

Conclusion

This paper deals with the problem of unforeseen surge and deflation of Flight Price. The average rate of price for air carriers have seen unexpected fluctuation over the years. Our goal is to analyze six chosen prediction algorithms that we found to be fit for our model, visualize the effect of each, observing the implementation of PCA on each of the models and to find a specific predicted airfare. Also, our aim is to improve the accuracy. We hope to add some new ideas to existing research such as-implementation for International and domestic Bangladeshi Airlines Datasets and breach this gap to create a system that will be helpful for future ventures and bring some new and smart solution for airlines agencies and travelers.

Bibliography

[1] Yazdi, Maryam Farshchian, et al. ”Flight delay prediction based on deep learning and Levenberg-Marquart algorithm.” Journal of Big Data 7.1 (2020): 1-28.

[2] Gui, Guan, et al. ”Flight delay prediction based on aviation big data and machine learning.” IEEE Transactions on Vehicular Technology 69.1 (2019): 140-150.

[3] Chaturvedi, Akshat, Amam Dhariwal, and Manan Patel. ”STUDY ON PREDICTION OF AIRFARES BASED ON XGBOOST AND LIGHT GBM MACHINE LEARNING ALGORITHMS.”

[4] Thiagarajan, Balasubramanian, et al. ”A machine learning approach for prediction of on-time performance of flights.” 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). IEEE, 2017.

[5] Chen, Jun, and Meng Li. ”Chained predictions of flight delay using machine learning.” AIAA Scitech 2019 forum. 2019.

[6] Tziridis, Konstantinos, et al. ”Airfare prices prediction using machine learning techniques.” 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017.

[7] Edwards, Holly A., Darron Dixon-Hardy, and Zia Wadud. ”Aircraft cost index and the future of carbon emissions from air travel.” Applied energy 164 (2016): 553-562.

[8] Mahesh, Batta. ”Machine Learning Algorithms-A Review.” International Journal of Science and Research (IJSR).[Internet] 9 (2020): 381-386.

[9] Ray, Susmita. ”A quick review of machine learning algorithms.” 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, 2019.

[10] Etzioni, Oren, et al. ”To buy or not to buy: mining airfare data to minimize ticket purchase price.” Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003.

[11] Narangajavana, Yeamduan, et al. ”Prices, prices and prices: A study in the airline sector.” Tourism Management 41 (2014): 28-42.

[12] Li, Jun, Nelson Granados, and Serguei Netessine. ”Are consumers strategic? Structural estimation from the air-travel industry.” Management Science 60.9 (2014): 2114-2137.

[13] Groves, William, and Maria Gini. "A regression model for predicting optimal purchase timing for airline tickets." (2011).

[14] Lantseva, Anastasia, et al. "Data-driven modeling of airlines pricing." Procedia Computer Science 66 (2015): 267-276.

[15] Chowdhury, Mozammel, Azizur Rahman, and Rafiqul Islam. "Malware analysis and detection using data mining and machine learning classification." International Conference on Applications and Techniques in Cyber Security and Intelligence. Edizioni della Normale, Cham, 2017.

[16] Domínguez-Menchero, J. Santos, Javier Rivera, and Emilio Torres-Manzanera. "Optimal purchase timing in the airline market." Journal of Air Transport Management 40 (2014): 137-143.

[17] Santana, Everton, and Saulo Mastelini. "Deep regressor stacking for air ticket prices prediction." Anais do XIII simpósio brasileiro de sistemas de informação. SBC, 2017.

[18] Li, Li, and Kai-Hsuan Chu. "Prediction of real estate price variation based on economic parameters." 2017 International Conference on Applied System Innovation (ICASI). IEEE, 2017.

[19] Wohlfarth, Till, et al. "A data-mining approach to travel price forecasting." 2011 10th International Conference on Machine Learning and Applications and Workshops. Vol. 1. IEEE, 2011.

[20] Vu, Viet Hoang, Quang Tran Minh, and Phu H. Phung. "An airfare prediction model for developing markets." 2018 International Conference on Information Networking (ICOIN). IEEE, 2018.

[21] Groves, William, and Maria Gini. "An agent for optimizing airline ticket purchasing." Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems. 2013.

[22] Groves, William, and Maria Gini. "On optimizing airline ticket purchase timing." ACM Transactions on Intelligent Systems and Technology (TIST) 7.1 (2015): 1-28.

[23] Chen, Yuwen, et al. "An ensemble learning based approach for building airfare forecast service." 2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015.

[24] Xu, Yuchang, and Jian Cao. "OTPS: A decision support service for optimal airfare ticket purchase." 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.

[25] Malighetti, Paolo, Stefano Paleari, and Renato Redondi. "Pricing strategies of low-cost airlines: The Ryanair case study." Journal of Air Transport Management 15.4 (2009): 195-203.

[26] Wang, Yongli. ”Dynamic pricing considering strategic customers.” 2016 International Conference on Logistics, Informatics and Service Sciences (LISS). IEEE, 2016.

[27] Chen, Yiwei, and Vivek F. Farias. ”Robust dynamic pricing with strategic customers.” Mathematics of Operations Research 43.4 (2018): 1119-1142.

[28] Liu, Tao, et al. ”ACER: An adaptive context-aware ensemble regression model for airfare price prediction.” 2017 International Conference on Progress in Informatics and Computing (PIC). IEEE, 2017.