IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

Prediction System for Flight Fares and Hotel Prices using Ensemble Machine Learning Algorithm

Tejal Dimble, Nikita Pandey, Harshada Narkhede, Ruturaj More
Students, Department of Information Technology
D. Y. Patil Institute of Engineering and Technology, Ambi, Pune, Maharashtra, India

Abstract: *As domestic air travel is getting more and more popular these days in India with various air ticket booking channels coming up online, travellers are trying to understand how these airline companies make decisions regarding ticket prices over time. Also, knowing the best time to travel and the best place to stay in appropriate amount is necessary. Unfortunately, the dynamic pricing strategy is usually carried out programmatically and is based on certain hidden parameters (e.g., number of days left till flight departure, or number of seats left). The paper works on mining the previous airfare data and developing data modelling technique to predict the price variation over time so that the consumer could benefit from it. This paper document study conducted to understand the airfare dependency over many hidden variables of which oil price, week day of departure, number of stops still have not received much attention from the research community. Also, this paper extends the research on hotel room prices using traditional and non-traditional statistical models following the analysis by Ka Athanasopoulos and Shehhi (2018), which discusses how hotel prices can be easily predicted. Research data were obtained from Smith Travel Research. In this study, we apply advanced forecasting models based on machine learning and artificial intelligence to the hospitality sector. Some of the models used in this study, such as the ANFIS model, contribute to the research conducted in the GCC region. The goal of the research was to contribute to the academic literature and assist hotel operators and decision-makers in setting appropriate strategies. It also describes the two different methodologies adopted to model this price change, comparative analysis of algorithms under these two methodologies, applied on real world data has also been performed. The comparative analysis thus helped us to find out the most effective algorithm for the prediction of the airfare variations and appropriate hotel prices. The study suggests that mining historical airfare data and hotel fare data, and modelling using machine learning algorithms can help predict the price trend and save consumer's substantial sum. Lately, we have acknowledged that in this era Mathematical terminologies and Scientific Equations has provided solutions to many of the problems. Moreover, the existence of Artificial Intelligence and its subset viz. Machine Learning has made tasks convenient. The power that Machine Learning carries is surely terrible. With various available tools and equipment that these terminologies are providing, the prediction of fares by considering all the components will lead to better understanding of travelling costs and will be helpful for the users to manage their entire travelling cost.*

Keywords: Machine Learning, Artificial Intelligence, Modelling, comparative analysis

I. INTRODUCTION

Airline companies use complex algorithms to calculate flight prices based on the conditions present at particular circumstances. These methods take financial, marketing, and various social factors into consideration to predict flight prices. Nowadays, the count of individuals using flights has increased significantly. It is difficult for airlines to keep up prices since prices change dynamically because of different conditions. That's why we'll try and use machine learning to unravel this problem. This could help airline by predicting what prices they will maintain. It may help customers to predict future flight prices and plan their journey accordingly. Moreover, calculating the hotel fare will help the consumers to calculate the Cost of their entire traveling more approximately. Our web application will also predict the weather of the destination city as well as will also redirect to the recommendation page.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

Proposed study Airfare price prediction using machine learning techniques, For the research work a dataset consisting of various data flights of the Airlines is collected and used to train machine learning model. Different number of features are used to train model various to showcase how selection of features can change accuracy of model. To build a model partial least square regression technique is used.

Various Machine Learning algorithms are used here to get accurate prediction. There are Various factors that impact the prices of flight. Distance, time taken, different stops, etc., plays an important role getting better accuracy. These factors help in creating patterns to decide the fate of flight. Machine Learning models get trained on these patterns to make the prediction meticulous and quicker the process.

Regression techniques are used since, the desired output will be a continuous value. We fit data into multiple regression models and compare the performances of all the models and minimize the errors to choose the most appropriate one. Along with the flight fare another thing that affects the travelers is hotel price. Mainly, the hospitality sector, which is part of the travel and tourism industry, has acknowledged interest from policymakers and practitioners in the public and the private sectors. Travelling is not the only thing that affects the tourist/travelers. This planning is a complex process as the tourism industry is connected with many other industries and it requires appropriate and scientific approach to gain the accuracy. In association to their revenue management systems, hotel managers are interested in recognizing the factors that affect the industry. They examine price changes very minutely in order to foresee any movement that could impact their businesses. This is specifically important because they deal with perishable products. Many products and services in this sector cannot be stored and need to be utilized the same day.

In the circumstances of the growing literature on hotel price prediction, this research was carried out using the most innovative forecasting tools currently available to examine the hospitality sector.

Hoteliers have employed revenue management to optimize the physical inventory of their assets and enhance revenues and profits through fare offers. Perpetually, they have used computers and technology to co-operate them in maximizing their profits through a process called yield management. This process has also led to many opportunities to explore in terms of human resource practices. In order to further improve performance, the strategy could be applied to phase in advanced analytical techniques using artificial intelligence (AI) algorithms.AI models of a classification nature (e.g., support vector machines, random forest) and of a regression nature (e.g., neural networks) has played essential role in recent years, dominating forecasting competitions around the globe. Its abundance of data, its human complexity, and its continuous growth Academics have attracted the larger part of education sector. The classic statistical tools used in the past were used to study price movements in the hospitality and finance sectors. These tools include time series, naïve methods, multiple regression, and polynomial models.

II. LITERATURE SURVEY

It is hard for the client to buy an air ticket at the most reduced cost. For these few procedures are explored to determine time and date to grab air tickets with minimum fare rate. The majority of these systems are utilizing the modern computerized system known as Machine Learning. To determine ideal purchase time for flight ticket Gini and Groves [2] exploited Partial Least Square Regression (PLSR). for building up a model. The information was gathered from major travel adventure booking sites from 22 February 2011 to 23 June 2011. Extra information was additionally gathered and are utilized to check the correlations of the exhibitions of the last model. Janssen [3] implemented a desire model using the Linear Quantile Blended Regression methodology for San Francisco–New York course where each day airfares are given by www.infare.com. Two features such as number of days for departure and whether departure is on weekend or weekday are considered to develop the model. The model guesses airfare well in advance from the departure date. But the model isn't convincing in a situation for an extensive time allotment, it closes the departure date. Wohlfarth [10] proposed a ticket purchasing time improvement model subject to a significant pre-processing known as macked point processors, data mining frameworks (course of action and grouping) and quantifiable examination system. This framework is proposed to change various added value arrangements into included added value arrangement heading which can support to solo gathering estimation. This value heading is packed into get-together reliant on near evaluating conduct. Headway model measure the value change plans. A tree-based analysis used to pick the best planning gathering and a short time later looking at the progression model. An investigation by Dominguez-Menchero [11] suggests the perfect purchase timing reliant on a nonparametric isotonic backslide technique for a specific course, carriers, and time frame. The model provides the most

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

acceptable number of days before buying the flight ticket. The model considers two types of a variable such as the entry and is date of obtainment

III. MATERIAL

3.1 Hardware:

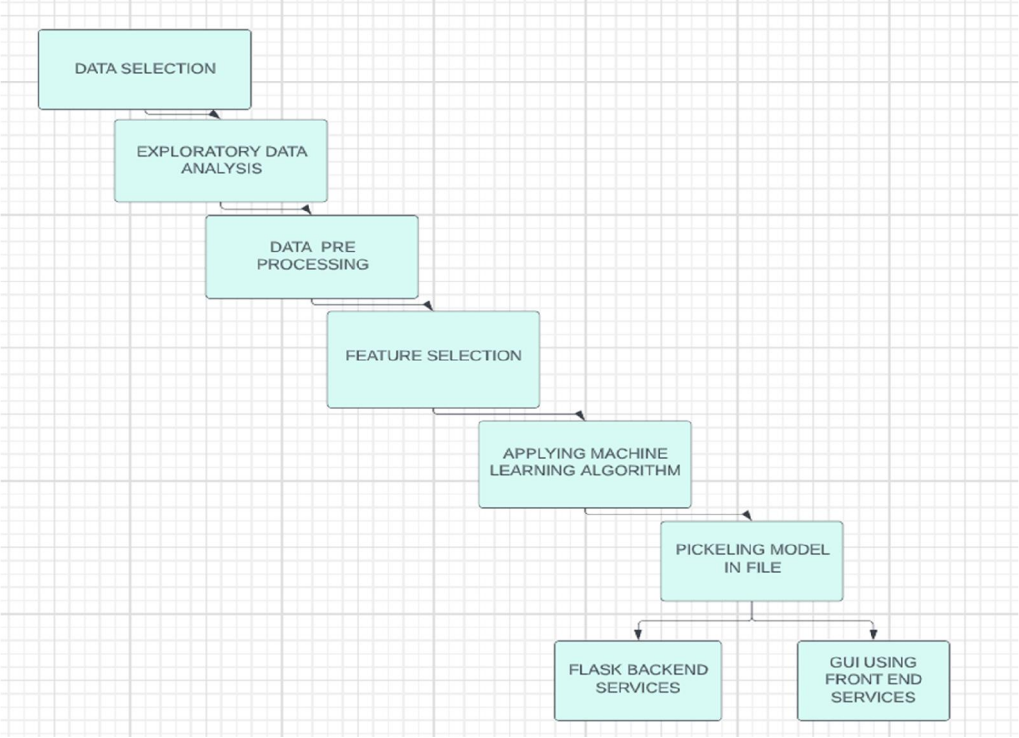
1. Intel core, Speed: 2.80 GHz,
2. RAM: 8GB.
3. Hard Disk: 40 GB
4. Key Board: Standard Windows Keyboard

3.2 Software Interfaces

1. Operating System: Windows 10
2. IDE: Anaconda, Visual Studio Code
3. Programming Language: Python

IV. METHODOLOGY

In this project, we have implemented the machine learning life cycle to create a basic web application which will predict the flight prices by applying machine learning algorithm to historical flight data using python libraries like Pandas, NumPy, Matplotlib, seaborn and sklearn. Figure.1 shows the steps that we followed from the life cycle:



The very first step is data selection where historical data of flight is assembled for the model to predict prices. Our dataset consists of more than 10,000 records of data related to flights and its prices. Some of the features of the dataset are source, destination, departure date, departure time, number of stops, arrival time, prices and few more.

In the exploratory data analysis step, data cleaning is done by removing the duplicate and null values. If these values are not removed it would affect the accuracy of the model. We gained further information such as distribution of data.

Further step is data pre-processing where we observed that most of the data was available in string format. Data from each feature is extracted in integer format, for example day and month is extracted from date of journey, and hours and minutes is extracted from departure time.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

Features such as source and destination should be converted into values as they are of categorical type. To convert categorical values into model identifiable values, hot-encoding and label encoding techniques are used

Feature selection step is involved in selecting important features which are correlated to the fare. There are some unnecessary features such as extra information and route which may affect the accuracy of the model and hence, they need to be eliminated before implementation of our model. After selecting the features accurately related to price, the next step involves applying machine learning algorithm and developing a model. As our dataset consists of labelled data and comprises continuous values in the features, we will be using supervised machine learning algorithms mainly regression algorithms. Regression models are used to describe relationship between dependent and independent variables.

4.1 Data Collection

As the APIs of Indian companies such as Goibibo retrieved data in a complex format that leads to a lot of time to clean up data before analysis, so we decided to build a web spider that extracted the required 4 values from the website and saved it as CSV. file. We decided to scrutinize the travel service provider's website using a Python-based spider. In addition we have developed a Python text to use the API provided by Google's most reliable airlines, but only allows 50 questions per day. Such cancellation returns a lot of flexibility to each recovered plane and we had to determine the parameters that may be required in the flight prediction algorithm. Not all are needed so we chose the following

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

Similarly for the hotel Prices

1. Distance From City
2. No of Bedrooms
3. Swimming pool
4. Hotel Type
5. Bed Type

4.2 Data Cleaning

The data were further analyzed based on the parameters listed below and refined based on the appropriate assumptions –

1. Departure days
2. Departure Day
3. Length
4. Jump
5. Holiday
6. Outliers

In addition, data were analyzed and tested for distribution was performed. Test results revealed that our data followed Log-Normal distribution and the same has been well verified by statistical methods. Based on previous history, the model flight price trend and the same was used to give the user an estimate of the number of days he has to wait from the current day, and if he waits, the price can specify a ticket. To predict whether a customer should wait or not, we used a combination of mathematical models and machine learning models. The mathematical model provided with the possibilities associated with each of the 5 low-cost airlines while the machine learning model advanced to predict specific conditions taking into

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

account travel dates and departure dates. The machine learning algorithms used started with the basic Depression models and expanded to Decision Trees followed by Random Forests and Gradient Growth Methods. We later developed an algorithm that included a combination of Rule-based learning, Ensemble models and mathematical models to increase accuracy. Based on the predictions made by the model and the average waiting period, we calculate the savings we can gain and losses based on estimates. The machine learning algorithms that we will be using in our project are

4.3 Linear Regression

In simple linear regression, there is only one independent and dependent feature; However, since our dataset consists of many independent features on which the price may depend upon, we will be using multiple linear regression which evaluates relationship between two or more independent variables and one dependent variable.

The multiple linear regression model is represented by:

$$Y = \beta 0x1 + \ldots + \beta nxn + Ɛ$$

Y = the forecasted value of the dependent variable
Xn = the independent variables
βn = independent variables coefficients
Ɛ = y-intercept when the rest of the parameters are 0

4.4 Decision Tree

Decision trees are basically of two types viz. classification and regression tree. Classification is used for categorical values and regression is used for continuous values. Decision tree goes for independent variable from dataset as decision nodes for decision making. It divides the whole dataset in different sub-section and when test data is passed to the model the output is decided by checking the section to which the data point belongs to. And to whichever section the data point belongs to, the decision tree will provide output as the average value of all the data points in the sub-section.

4.5 Random Forest

Random Forest is an ensemble learning technique where training model uses multiple learning algorithms then combine individual results to urge a final predicted result. Under ensemble learning random forest falls into bagging category where random number of features and records are selected and passed to the group of models. Random forest mainly uses group of decision trees as group of models. Random amount of information is passed to decision tree and every decision tree predicts values as per given in the dataset. From the predictions made by the decision trees the average value of the predicted values is given into account as the output of the random forest model.

4.6 Performance Metrics

Performance metrics are statistical models which will be used to compare the accuracy of the machine learning models trained by different algorithms. The sklearn .metrics module will be used to implement the functions to measure the errors from each model using the regression metrics. Following metrics will be used to check the error measure of each model.

4.7 MAE (Mean Absolute Error)

Mean Absolute Error is basically the sum of average of the absolute difference between the predicted and actual values.

$$MAE = 1/n[\sum(y-ý)]$$

y = actual output values,
ý = predicted output values
n = Total number of data points
Lesser the value of MAE the better the performance of your model.

4.8 MSE (Mean Square Error)

Mean Square Error squares the difference of actual and predicted output values before summing them all instead of using the absolute value.

$$MSE = 1/n[\sum(y-ý)2]$$

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

y=actual output values
ý=predicted output values
n = Total number of data points
MSE punishes big errors as we are squaring the errors. Lower the value of MSE the better the performance of the model.

4.9 RMSE (Root Mean Square Error)

RMSE is measured by taking the square root of the average of the squared difference between the prediction and the actual value.

RMSE = √1/n[∑(y-ý)2]

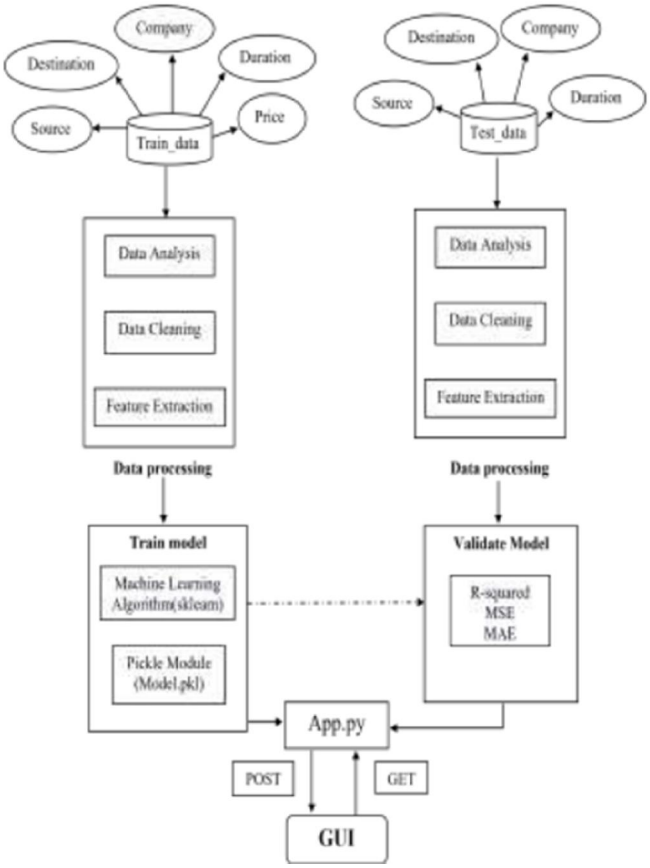
y=actual output values
ý=predicted output values
n = Total number of data points RMSE is greater than MAE and lesser the value of RMSE between different model the better the performance of that model.

4.10 R 2 (Coefficient of determination)

It helps you to understand how well the independent variable adjusted with the variance in your model.

R ^2 = **1** − ∑(ý-ȳ) ^2 ∑(y-ȳ) ^2

The value of R-square lies between 0 to 1. The closer its value to one, the better your model is when comparing with other model value



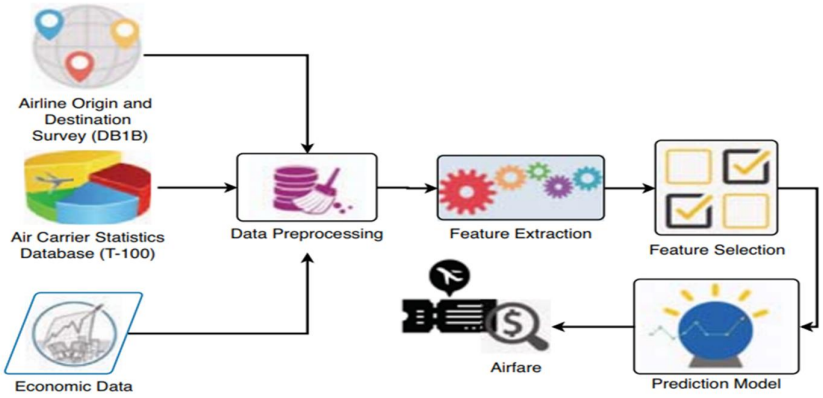
There are also various conflicting verification methods such as grid search CV and random search CV used to improve model accuracy. The parameters of the models such as the number of trees in the random forest or the maximum depth of the decision tree can be modified using this process which will help us to further improve accuracy. The last three steps of the life cycle model are involved in the use of a machine-readable learning model. Therefore, after finding the model with

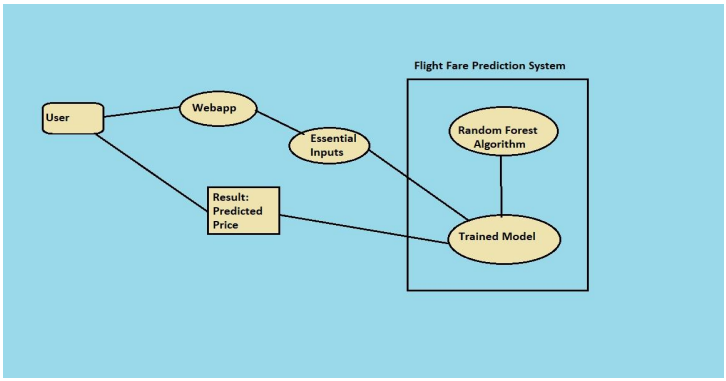
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

the best accuracy, we store that model in a file using the pickle module. The background of the application will be built using the Flask Framework where API archives such as GET and POST will be created to perform tasks related to downloading and displaying data at the end of the application. The front of the app will be created using the bootstrap framework where the user will be entering the required information. This data will be sent to a background service where the model will predict the output based on the data provided. The predicted value is then displayed earlier to the user.

V. WORKING OF SYSTEM

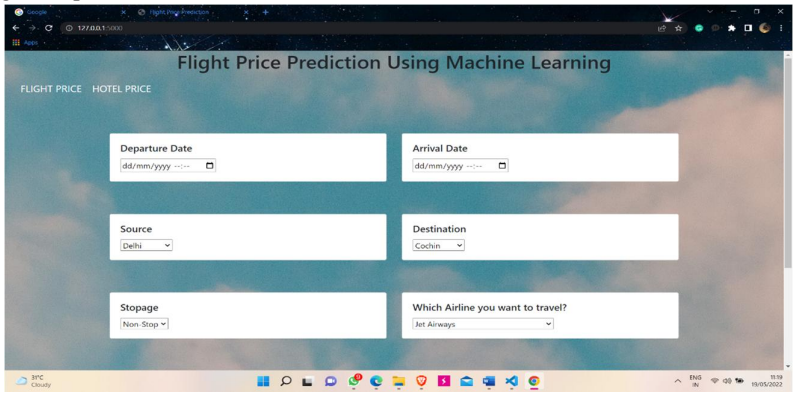
5.1 System Diagram



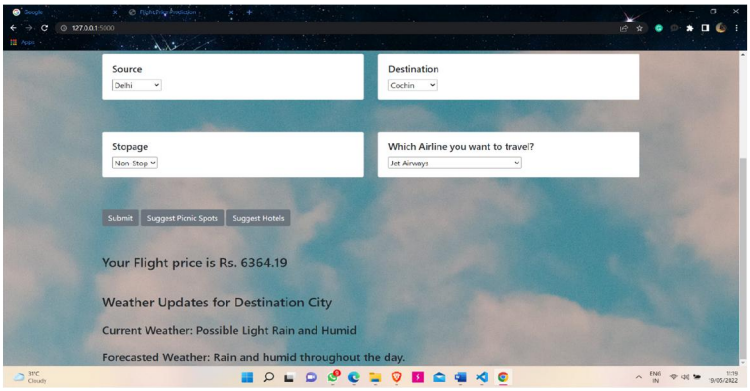


VI. RESULTS

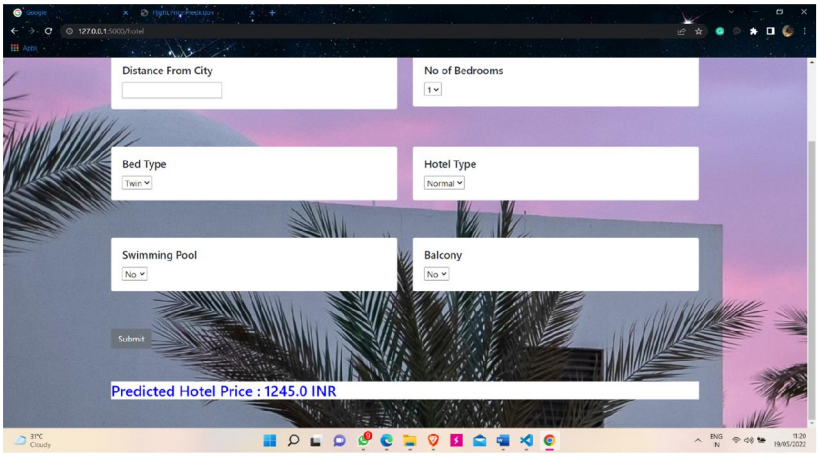
In this project, frontend contains website on which we have to the give inputs like source, destination, stoppage, and Airlines name for flight fare and weather prediction, and other inputs like distance from city, no of bedroom, Bed type, Hotel type, Swimming Pool, Balcony for hotel price prediction. Here use if wants to know flight fare then he can enter following details and get the predicted results.



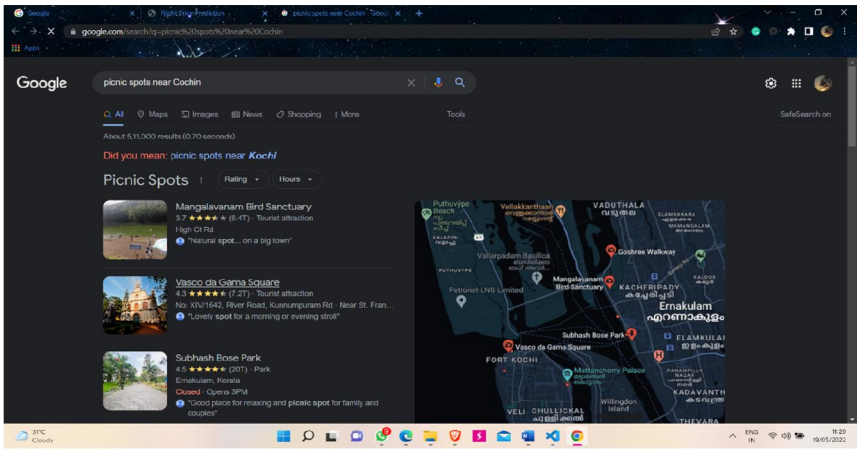
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252



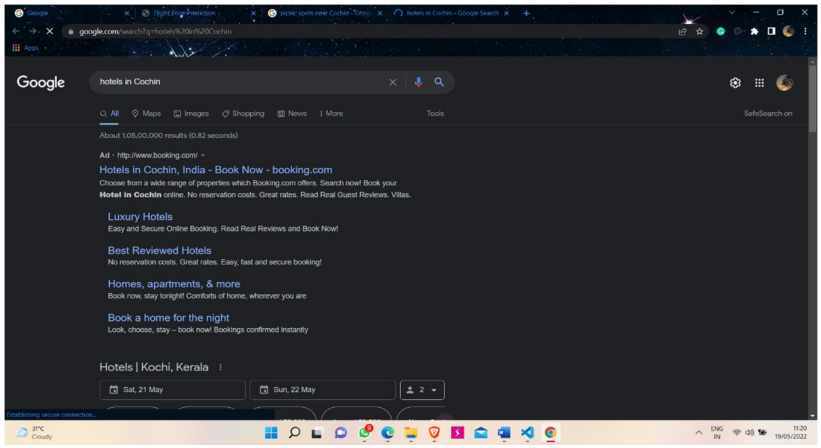
If user wants to predict hotel prices then he can enter following details and required outputs.



Can also redirect to the picnic spot and hotel suggestion page if he wants to do so.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252



VII. Conclusion

We collected data from various sources, cleaned, compiled and modified it in order to produce the final data from random raw data. Later, we investigated a number of algorithms under the categories and regression methods that can be used to solve this problem. Despite the complex marketing algorithms used by airlines and the lack of data on dynamic parameters such as the number of seats left on the plane, our method of comparing data has performed surprisingly well. By the similar process we tried to forecast the hotel fare which again resulted in a positive conclusion. Moreover, by adding the google map link, we tried to suggest the tourist spots and best stay in the destination to be reached. Similar studies in other domains such as auctions can be made but this initial study demonstrates the ability to decipher previous price data and use data mining and machine learning algorithms to examine price changing models and help make relevant purchasing decisions. We believe that this type of excavation is a futile field for future research. This study shows the prices of hotels selected by the GCC using advanced algorithms. Machine learning and artificial intelligence models show the edge in this compared to conventional speculation tools like SARIMA. This research can be extended to ongoing regional tourism forecasting studies using large data sets and other variables.

Acknowledgment

Every engineering student views a final year project as an opportunity for a student to use the skills one has developed over the years through hard work and dedication. Last-year projects do wonders in honing technical skills and soft skills. Working with people with different personalities, ideas, and skills enhances one's perception of team work, and shapes everyone involved in a constructive way. By working on new ideas, we train our brain in that track, which helps us to have a clearer view of the reality. The landmark project would not have been possible without the help of a few key individuals. These people were our true motivators, and in fact they were the backbone without whom this work would not have been truly successful. We owe this moment of satisfaction with a loving sense of gratitude to our internal director who guided us through all the stages, whose technical support and helpful attitude gave us the highest moral support. He was available to us in every twist. Since many terms were new to us and we had no idea finding ways to complete these tasks, Ma'am was there with all the guidance and patience with us. We would also like to extend our thanks to our H.O.D. Drs. Ninad More for his guidance and ongoing encouragement. We have a great responsibility to all the staff of the information technology department and the principal for their co-operation. We also take this opportunity to thank all the partners who have raised us for providing us with helpful suggestions and assistance. We as a team have tried our best to support ourselves and understand the consequences of our actions. Working together taught us mutual respect and the ideas of team members. We are very responsible for providing us with this opportunity, resources and appropriate guidance. Finally, but not least we thank our friends, colleagues and all the people directly or indirectly involved in this project.

References

[1]. B. Smith, J. Leimkuhler, R. Darrow, and Samuels, ―Yield management at american airlines, Interfaces, vol.22, pp. 8–31, 1992.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

[2]. W. Groves and M. Gini, ―An agent for optimizing airline ticket purchasing, ‖ 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), St. Paul, MN, May 06 - 10, 2013, pp. 1341-1342.

[3]. T. Janssen, ―A linear quantile mixed regression model for prediction of airline ticket prices, ‖ Bachelor Thesis, Radboud University, 2014.

[4]. Viet Hoang Vu, Quang Tran Minh and Phu H. Phung,‖An Airfare Prediction Model for Developing Markets‖, IEEE paper 2018.

[5]. S.B. Kotsiantis, ―Decision trees: a recent overview, ‖ Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.

[6]. L. Breiman, ―Random forests, ‖ Machine Learning, vol. 45, pp. 5- 32, 2001.

[7]. S. Haykin, Neural Networks – A Comprehensive Foundation. Prentice Hall, 2nd Edition, 1999.

[8]. H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, ‖Support vector regression machines, ‖ Advances in neural information processing systems, vol. 9, pp. 155-161, 1997.

[9]. www.Makemytrip.com

[10]. Wohlfarth, T. Clemencon, S.Roueff, ―A Dat mining approach to travel price forecasting‖, 10 th international conference on machine learning Honolulu 2011.

[11]. Dominguez-Menchero, J.Santo, Reviera, ‖optimal purchase timing in airline markets‖ ,2014.

[12]. Supriya Rajankar and Neha Sakharkar, ―A Survey on Flight Pricing Prediction using Machine Learning‖, International Journal of Engineering Research and Technology, vol 8, issue 6, June 2019.

[13]. K. Tziridis, Th. Kalampokas, G.A. Papakotas and K.I.

[14]. Diamantaras,‖Airfare Prices Prediction Using Machine Learning Techniques‖, EUSIPCO 2017.

[15]. Akm, M. (2015). A novel approach to model selection in tourism demand modeling. Tourism Management, 48, 64–72.

[16]. Alvarez-Diaz, M., Gonzalez-Gomez, M., & Otero-Giraldez, M. S. (2019). Estimating the economic impact of a political conflict on tourism: The case of the Catalan separatist challenge. Tourism Economics, 34–50. Assaf, A. G., Josiassen, A., Woo, L.,

[17]. Agbola, F. W., & Tsionas, M. (2017). Destination characteristics that drive hotel performance: A state-of-the-art global analysis. Tourism Management, 60, 270–279.

[18]. Burger, C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioner's guide to timeseries methods for tourism demand forecasting – a case study of Durban, South Africa. Tourism Management, 22, 403–409.

[19]. Chen, K.-Y., & Wang, C.-H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. Tourism Management, 28, 215–226.

[20]. Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. Tourism Management, 24, 323–330.

[21]. Claveria, O., Monte, E., & Torra, S. (2015). Common trends in international tourism demand: Are they useful to