Airline Price Prediction using Machine Learning Algorithm

Keertheeswar R(20BDS0304) **Aishwarya(20BDS0230)** **HARUL MURUGAN R S(20BDS0306)**

SUNDARAVELAN (20BCT0248)

School of Computer Science and Engineering

VELLORE INSTITUTE OF TECHNOLOGY

VELLORE

1. ABSTRACT:

Everyone today strives to complete tasks more quickly and travel more quickly and affordably, but airlines are still expensive. In this project, we predict the price of airlines. we will be analysing the flight fare prediction using Machine Learning dataset using essential exploratory data analysis techniques then will draw some predictions about the price of the flight based on some features such as what type of airline it is, what is the arrival time, what is the departure time, source, destination and more.

2. INTRODUCTION:

The cost of airline tickets fluctuates depending on a number of variables, including the scheduling, destination, and length of the flight. By using machine learning algorithms to the gathered historical flight data, the suggested system will generate a prediction model. It might be difficult for consumers to decide when the best time is to buy plane tickets, mostly because they lack the knowledge necessary to make predictions regarding price changes. In this project, our main goals were to identify underlying trends in travel costs in India using historical data and to recommend the most advantageous time to purchase a ticket. A comparison study of several methods in determining the ideal time to purchase a flight ticket and the amount that may be saved if done so is conducted as part of the research in order to validate or refute common misconceptions about the airline business. Surprisingly, the route, month, day,

time, whether the day of departure is a holiday, airline carrier, and day of departure all have a significant impact on price trends. However, other routes (tier 1 to tier 2 cities like Delhi - Guwahati) had a specific time period where the prices are minimum. These routes were highly competitive, such as most business routes (tier 1 to tier 1 cities like Mumbai-Delhi), which had a non-decreasing trend where prices increased as days to departure decreased. The data also revealed two fundamental groups of aircraft carriers operating in India: the opulent group and the economical group. In most cases, the trip with the lowest fare belonged to the latter. The statistics also confirmed that there are times of the day when prices are predicted to be at their highest. To significantly reduce the cost of purchasing flights across the Indian domestic airline market, the project's scope might be greatly expanded across the numerous routes. To avoid the effects of the most extreme charge, the recommended method for purchasing an airline ticket is to do so far in advance of the flight's departure. Most aviation routes disagree with this practise. When they need to develop the market and when tickets are more difficult to get, airline companies may lower the price. They may raise prices to the maximum. Therefore, the price may depend on several aspects. This project uses AI to show future aircraft ticket prices in order to forecast costs. Every organisation has the right and ability to adjust the price of its tickets at any time. By purchasing a ticket at the lowest price, an explorer can lay aside money. People who frequently take flights are aware of pricing variations. Airlines implement several evaluating methods using complicated Revenue Management policies. As a result, the fee to modify the header or footer on subsequent pages varies based on the time, season, and holiday. While customers look for the lowest price, the airlines' primary goal is to make a profit. Customers typically aim to purchase their tickets far in advance of the departure date to prevent an increase in price as the departure date **approaches. However, this is not the case in reality. For the same seat, the customer might end up paying more than they should.**

3. LITERATURE SURVEY:

1] "Airfare price prediction using machine learning techniques," European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO.2017.8081365L, by K. Tziridis, T. Kalampokas, G.Papakostas, and K. Diamantaras Yawning detection for monitoring driver

fatigue based on two cameras was described by Li Y. Chen and Z. Li in Proc. 12th Int. IEEE Conf. Intel. Transp. Syst. pp. 1-6 Oct. 2009. Airfare price prediction using machine learning techniques is a proposed study, They employed a dataset made up of 1814 data flights from Aegean Airlines for their research, which they used to build a machine learning model. To demonstrate how the choice of features might affect a model's accuracy, different numbers of features were utilised to train the model. Numerous techniques, including the Multilayer Perceptron (MLP), Generalized Regression Neural Network, Extreme Learning Machine (ELM), and Random Forest Regression Tree, have been employed. o Linear Regression (LR), Regression SVM (Polynomial and Linear), Regression Tree, Bagging Regression Tree, and Regression SVM all produced distinct results. With different features from the dataset being removed and added, they have experimented with and trained several sorts of models. followed the standard life cycle for data science. The Bagging regression tree produced the best results.

2] An agent for maximising the purchase of airline tickets was described by William Groves and Maria Gini in the proceedings of the 2013 international conference on autonomous agents and multi-agent systems. In a case study by William Groves, a new agent is presented who can help consumers by maximising the timing of purchases. The technique of partial least squares regression is used to create a model. They initially employed a number of feature selection techniques, including feature extraction, lag-time feature computation, regression model building, and optimal model selection. They conducted trials to calculate the costs of applying our prediction models in the actual world. Many other machine learning algorithms can be used with the lag scheme technique, however PLS regression has been determined to be the most effective in this field.

3] J. Javier Rivera and Santos Dominguez-Menchero
"Optimal buy time in the airline market" by Emilio Torres Manzanera. In this study, the researchers have studied the general pattern in airline pricing behaviour and a methodology for analysing various routes and/or carriers. By balancing the need to save money with any time constraints the buyer

may have, they aim to give customers the pertinent information they need to decide when is the best time to buy a ticket. Their work demonstrates the superiority of non-parametric isotonic regression methods over conventional parametric methods. Specifically, we can specify the economic loss for each day the purchase is postponed and identify situations where it is preferable to wait until the very last day to make a purchase. These are the three most important things we can do.

4] 4. "Flight fare prediction using machine learning algorithms," by Supriya Rajankar, Neha Sakhrakar, and Omprakash Rajankar June 2019 issue of the International Journal of Engineering Research and Technology (IJERT). A study published in Supriya Rajankar's journal uses a limited dataset of flights between Delhi and Bombay to estimate travel prices using machine learning. K-nearest Neighbors (KNN), linear regression, and support vector machine (SVM) algorithms are used to obtain various results and conduct research on them. Numerous machine learning techniques have been devised for estimating the cost of airline tickets. Support Vector Machine (SVM), Linear Regression, K Nearest Neighbors, Decision Tree, Multilayer Perceptron, Gradient Boosting, and Random Forest Algorithm are the algorithms. These models have been used in the Python module scikit learn. The performance of these models is examined using parameters like R-square, MAE, and MSE. The Decision Tree method produced the best model outcomes.

5] "A Framework for Airlines Price Prediction: A Machine Learning Approach," by Tianyi Wang, Samira Pouyanfar, Haiman Tian, and Yudong Tao. In this paper, Tianyi Wang, Samira Pouyanfar, Haiman Tian, and Yudong Tao proposed a framework for modelling the average ticket price based on source and destination pairs using two databases combined with macroeconomic data and machine learning algorithms such as support vector machine and XGBoost. The framework achieves a high R squared performance metric of 0.869 for prediction accuracy. With the help of the XGBoost Algorithm, they achieved the lowest error rate of 0.92.

6] model for forecasting the cost of airline tickets. They have anticipated the ideal moment to buy the tickets in this essay. They have employed a variety of machine learning algorithms, including gradient boosting, support vector machines, decision trees, random forests, K-Nearest Neighbor, multilayer perceptrons, and linear regression (SVM). They have employed Naive Bayes and the stacked prediction model as predictors. The research uses the Linear Quantile Blended Regression methodology to implement the desired model for the San Francisco–New York course, where daily airfares are provided by an online website. When creating the model, two factors are taken into consideration, including the number of days before departure and whether it occurs on a weekend or weekday.

7] Wohlfarth, S. Roueff, and T. Clemencon The 10th international conference on machine learning in Honolulu in 2011 featured a paper titled "A Data Mining Approach to Travel Price Forecasting." In Wohlfarth's study on the flight pricing prediction system. T. Clemencon and S. Roueff are employing yield management in the aviation sector. They have made use of different data mining methods. The purpose of this paper is to take into account the design of decision-making tools in the context of fluctuating travel prices from the viewpoint of the customer. Machine techniques and algorithms, such as Clustering, are terms used in the study.

8] Flight fare prediction system research paper by Vinod Kimbhaune, Harshil Donga, Ashutosh Trivedi, Sonam Mahajan, and Viraj Mahajan. Researchers Vinod Kimbhaune, Harshil Donga, Ashutosh Trivedi, Sonam Mahajan, and Viraj Mahajan used the several machine learning algorithm techniques, such as Random Forest, Decision tree, and Linear regression, on a dataset in their research article 7 on Flight fare prediction system. to choose the best moment to purchase a flight. The goal of this project is to create an application that uses machine learning to forecast flight costs for various flights. They list Linear Regression, Decision Trees, and Random Forest as the methodologies they employed. The methods for measuring performance are MAE, MSE, and RSME.

9] An agent for optimising airline ticket purchases, by W. Groves and M. Gini. 1341–1342 in AAMAS 2013, the 12th International Conference on Autonomous Agents and Multiagent Systems, St. Paul, Minnesota, May 6–10, 2013. This is an expanded version of the study that used partial least squares regression (PLSR) to create a model. Major adventure travel booking websites were used to acquire the data between 22 February 2011 and 23 June 2011. Additional data was also acquired, and it was used to examine the relationships between the displays of the previous model.

10] "An Airfare Prediction Model for Developing Markets," by Viet Hoang Vu, Quang Tran Minh, and Phu H. Phung. 2018 IEEE paper In this study, they make a new model suggestion that can assist the consumer in anticipating price trends without relying on official airline information. Their research showed that, although lacking several essential components, such as the amount of unsold seats on flights, the suggested model can nevertheless accurately forecast trends as well as changes in actual airfare up to departure dates using publicly available internet airfare data. They have also determined the characteristics that have the biggest effects on variations in airfare. They suggested a methodology for reducing the time spent buying tickets that was subject to a significant pre-processing using "macked point processors," "data mining frameworks" (course of action and grouping), and "quantifiable examination system." With the help of this framework, different added value arrangements can be transformed into integrated added value arrangement headings that can assist estimate for a single gathering. This value heading is condensed into a group based on nearby assessing behaviour. Headway model evaluates proposed value changes. To select the best planning group, a tree-based analysis was employed. Shortly after, the progression model was examined.

11] A data mining method to travel price predictions, 10th international conference on machine learning, Honolulu, 2011. Wohlfarth, T. Clemencon, and S. Roueff Throughout the previous two decades, a sizable body of data-mining techniques have been created with the goal of boosting airline firms' profitability, as this report explained. Due to the implementation of mathematical optimization tactics, similar seats on the

same flight were frequently purchased at various prices depending on the supplier, the time of the transaction, etc. The purpose of this study is to take into account the design of decision-making tools in the setting of fluctuating travel prices from the viewpoint of the client. We present two methods for predicting changes in travel prices at a given horizon based on massive streams of heterogeneous historical data amassed online. The methods use a list of flight's descriptive characteristics as well as potential characteristics of the related price series' historical evolution as input variables. Even though it is varied in many ways (such as scale and sampling), the collection of historical price series is here represented by indicated point processes in a unified way (MPP). The customer can then be assisted in choosing when to buy a ticket by using cutting-edge supervised learning algorithms, possibly in conjunction with a preliminary clustering stage that groups flights whose associated price series reflect similar behaviour.

12] J. Santo, Reviera, and Dominguez-study Menchero's on the best time to buy in the airline markets. This study shows broad trends in airline price behaviour as well as a methodology for comparing various carriers and/or routes. The idea is to balance the need to save money with any time constraints the buyer may have by giving them the pertinent information they need to choose the best time to buy a ticket. The study demonstrates the superiority of non-parametric isotonic regression methods over traditional parametric methods. The ability to estimate the amount of time consumers can delay making a purchase without a significant price increase, quantify the financial loss incurred for each day the purchase is postponed, and identify situations in which it is preferable to wait until the very last minute are our three most important capabilities.

13] which statistic can be found at analytics-vidhya.medium.com/mae-mse-rmse-coefficient-ofdetermination is Better performance metrics content. The goal of linear regression, as we learnt in this paper, is to identify a line that minimises the prediction error of all the data points. Evaluating the model's accuracy is a crucial stage in any machine learning model. To assess how well a model performs in a regression study, metrics like Mean Squared Error, Mean Absolute Error,

Root Mean Squared Error, and R-Squared or Coefficient of Determination are utilised. However, because RMSE has the same units as the dependent variable, it is more frequently employed than MSE to assess how well the regression model performs with other random models (Y- axis). While R-Square indicates how well the predictor variables can explain the variation in the response variable, RMSE indicates how well a regression model can predict the value of a response variable in absolute terms.

14] Article about random forest at keboola.com/blog/random-forest-regression Random forest is an ensemble method as well as a supervised learning system. It is supervised in that it learns the mappings between inputs and outputs during training. In order to produce predictions that are more accurate than those made by any one underlying algorithm alone, ensemble algorithms mix a number of different machine learning methods. The final choice made by random forest is an ensemble of various decision trees. Both regression tasks (which forecast continuous outputs, like price), and classification tasks, can be performed using random forest (predict categorical or discrete outputs). Depending on how much you are familiar with the overall data science process, you can use random forest regression in different ways in practise.

15] Towards Data Science: Machine Learning Basics: Decision Tree Regression 1d73ea003fda decision tree regression article. Decision trees are one of the most widely used, practical models for supervised learning, as we learned in this paper. Both Regression and Classification tasks can be solved with it, while Classification is more frequently utilised in real-world settings. There are three different sorts of nodes in this tree-structured classifier. The Root Node is the first node that represents the full sample and may be divided into further nodes. The branches stand in for the decision criteria, while the interior nodes reflect the characteristics of a data collection. The result is represented by the Leaf Nodes in the end. For resolving issues involving decisions, this algorithm is highly helpful. Decision trees have the advantages of being simple to understand, requiring less data cleaning, performing well even when non-linearity is present, and having almost no hyper-parameters to tune.

16] A. Yates, C. A. Knoblock, R. Tuchinda, and O. Etzioni. To buy or not to buy: using airfare data to reduce the cost of a ticket. This study, which involved "price mining" online, was the subject of this publication. We obtained airfare information from the internet and demonstrated that it is possible to forecast airline price changes based on historical fare information. Our data mining techniques performed very well despite the complicated algorithms used by the airlines and the lack of data on important variables like the number of seats available on a flight. Most significantly, by strategically scheduling ticket purchases, our Hamlet data mining technology saved 61.8% of the potential amount. Our techniques were inspired by traditional machine learning, computational finance, and statistics (time series methods) (Ripper rule learning). The algorithms were merged using a variation of stacking to increase their predictive accuracy. Each algorithm was designed specifically for the situation at hand (for example, we created an appropriate reward function for reinforcement learning).

17] Airlines use dynamic pricing for tickets, and they base their pricing choices on algorithms for estimating demand. Each flight only has a certain number of seats to sell, so airlines must control demand, which is why the system is so complex. The airline may raise prices if demand is anticipated to exceed capacity in order to slow the rate at which seats fill. On the other hand, an unsold seat results in a loss of revenue, so it would have been better to sell it for a price that is higher than the service cost for just one passenger. The goal of this project was to analyse how airline ticket prices fluctuate over time, identify the variables that affect these variations, and explain how these variables are correlated.

18] Groves and Gini (2011), 18. A Regression Model for Predicting the Best Time to Buy Tickets for Airlines. It might be difficult for consumers to decide when is the best time to buy plane tickets, mostly because they lack the knowledge necessary to make predictions regarding price changes. The methodology for calculating anticipated future prices and analysing the risk of price changes is presented in this study. Based on a corpus of historical

price quotes, the suggested model is used to forecast the future estimated minimum price of all available flights on certain routes and dates. Additionally, we use our algorithm to forecast flight fares that have particular desirable characteristics, such as flights operated by a particular airline, flights that only stop once, or multi-segment trips. Buyers can estimate the potential cost of their choices by comparing models with various desirable attributes. For two high-volume routes, we describe the anticipated costs of various preferences. By integrating instances of time-delayed features, establishing a class hierarchy among the raw features based on feature similarity, and trimming the classes of features utilised in prediction based on in-situ performance, the prediction models given are able to perform well.

19] Using the Official Airline Guide (OAG) and the Airline Origin and Destination Survey, model United States airline fares (DBIB), 2006; Krishna Rama Murthy For the purpose of analysing the effects of new modes like the NASA Small Airplane Transportation System, it is necessary to predict airline prices inside the United States, including Alaska and Hawaii (SATS). To calculate the cost of air travel, a general fare model, or aggregate cost model, of the disaggregated airline fares, must be created. In this thesis, the cost model is evaluated using the average fare to distance ratio, or the average fare and fare per mile. For the purposes of the analysis, the thesis first chooses the fare class categories to be applied to Coach and Business class.

20] B.S. Everitt: Cambridge University Press, Cambridge, The Cambridge Dictionary of Statistics ISBN 0-521- 69027-7. This academic text from the University of Cambridge in England teaches us how to use numerous mathematical formulas and algorithms.

21] Bishop, Pattern Recognition and Machine Learning, Springer, 387-31073-8 Over the past ten years, there have been numerous significant advancements in the algorithms and approaches that have contributed to the machine learning field's spectacular expansion in real-world applications.

These most current advancements are represented in this brand-new textbook, which also offers a thorough introduction to the subjects of pattern recognition and machine learning. Since the book contains a self-contained introduction to fundamental probability theory, no prior understanding of pattern recognition or machine learning principles is anticipated. Some prior exposure to the usage of probabilities would be beneficial but is not required. The book is appropriate for courses in data mining, signal processing, computer vision, statistics, and machine learning. Instructors can purchase solutions from the publisher for the remaining exercises, while solutions for a portion of the exercises are available from the book's website. The book contains a tonne of additional information that is helpful, and the reader is advised to visit the inspire book website for the most recent details. while brand-new kernel-based models have significantly impacted both algorithms and applications.

22] C. A. Piga and E Bachis Price variation and low-cost carriers. 2011 corrected proof, International Journal of Industrial Organization, In Press. The following are examples of online pricing strategies that violate the "Law of One Price" when various airlines release prices for the same flights at the same time but in different currencies. According to the survey, different airlines offer varying prices for less popular routes with more diverse demand. The intra-firm fare dispersion's temporal persistence suggests that it is an equilibrium phenomenon brought on by the airlines' requirement to control stochastic demand circumstances for a particular flight.

4. DATASET:

Primary data collection is done through Kaggle repositories [10] which has 10683 records of airline data. The dataset consists total of 9 variable, of which it contains 4 demographic variables like Airline name, Source, Destination and Total stops. It contains 5 numerical data like Date of journey, Departure time, Arrival time, Duration of flight and Price variable. Finally, the response/target attribute is Price, where we are upto predict for further other test cases.

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          10682 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10673 non-null  object
 3   Destination      10672 non-null  object
 4   Dep_Time         10682 non-null  object
 5   Arrival_Time     10683 non-null  object
 6   Duration         10683 non-null  object
 7   Total_Stops      10680 non-null  object
 8   Additional_Info  10683 non-null  object
 9   Price            10659 non-null  float64
dtypes: float64(1), object(9)
memory usage: 834.7+ KB
```

5. PROPOSED WORK:

5.1 PREPROCESSING:

Preprocessing of any datasets primarily involves checking for missing values, outliers, and proper format of datatypes etc…Thorough analysis of dataset depicts that it conatins missing values as listed

```
Airline            1
Date_of_Journey    0
Source            10
Destination       11
Dep_Time           1
Arrival_Time       0
Duration           0
Total_Stops        3
Additional_Info    0
Price             24
dtype: int64
```

The missing values in Categorical values like source, destination is replaced with respective mode of ach attributes. As null values in Airline(1), Departure time(1) and total stops(3) are very less compared to the size of dataset, it is dropped. Null values of Price is handled by replacing it with mean. Now all the records are full and the dataset contains no NULL value.

```
Airline            0
Date_of_Journey    0
Source             0
Destination        0
Dep_Time           0
Arrival_Time       0
Duration           0
Total_Stops        0
Additional_Info    0
Price              0
dtype: int64
```

Secondly, the datatypes of Date_of_journey, arrival_time, departure time is Object which is changed to proper datetime format. To have a precise data, it is important to extract all possible insights out, by which the prediction of airline price can be highly accurate. Here, we extract out Hours and Minutes from both arrival time and Departure time, day and month from Date of journey, Hours and minute from Duration in proper datatype format.

```
Airline              object
Source               object
Destination          object
Total_Stops          object
Additional_Info      object
Price               float64
journey_day           int64
journey_month         int64
Dep_Time_hour         int64
Dep_Time_min          int64
Arrival_Time_hour     int64
Arrival_Time_min      int64
dur_hour              int32
dur_min               int32
dtype: object
```

Thirdly, Handling of categorical data is done by encoding them using Label Encoding and Onehot Encoding in respective places. For nominal data(Total_Stops) i.e the data that are not in order are Label encoded,

whereas ordinal data like Airline, Source,Destination are OneHot encoded.

There are 12 categories of Airline and they are OneHot Encoded

Jet Airways	3820
IndiGo	2038
Air India	1741
Multiple carriers	1189
SpiceJet	815
Vistara	476
Air Asia	317
GoAir	194
Multiple carriers Premium economy	13
Jet Airways Business	6
Vistara Premium economy	3
Trujet	1

Air India	GoAir	IndiGo	Jet Airways	Jet Airways Business	Multiple carriers
0	0	1	0	0	0
1	0	0	0	0	0
0	0	0	1	0	0
0	0	1	0	0	0
0	0	0	0	0	0

There are 5 categories of Source and they are label Encoded

Delhi	4508
Kolkata	2850
Banglore	2183
Mumbai	693
Chennai	379

Chennai	Delhi	Kolkata	Mumbai
0	0	0	0
0	0	1	0
0	1	0	0
0	0	0	0
0	0	1	0

Similarly for destination there are 6 categories which are OneHot Encoded

```
Cochin        4508
Banglore      2850
Delhi         1256
New Delhi      927
Hyderabad      693
Kolkata        379
```

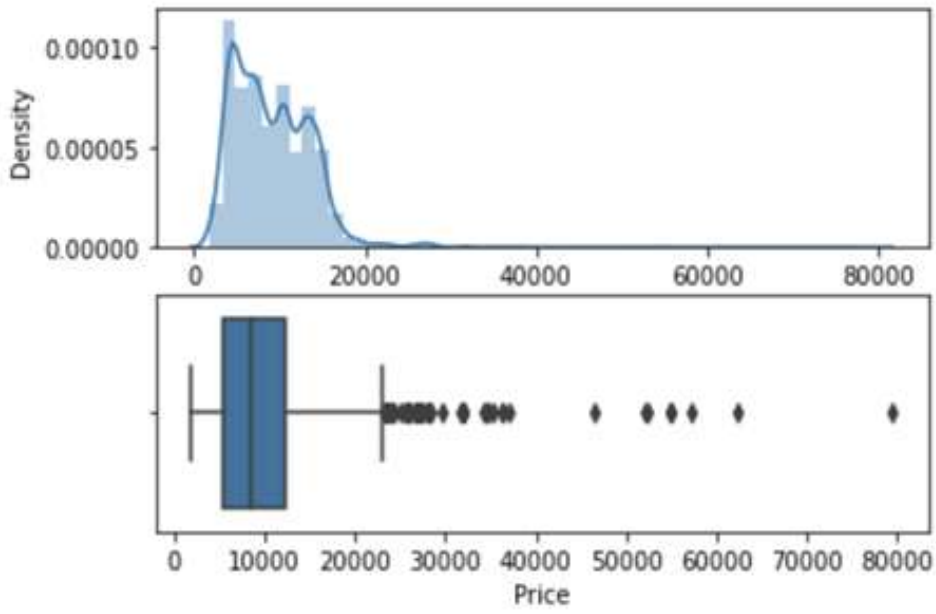
Cochin	Delhi	Hyderabad	Kolkata	New Delhi
0	0	0	0	1
0	0	0	0	0
1	0	0	0	0
0	0	0	0	1
0	0	0	0	0

Total Stop has 4 categories and Label Encoded as following:

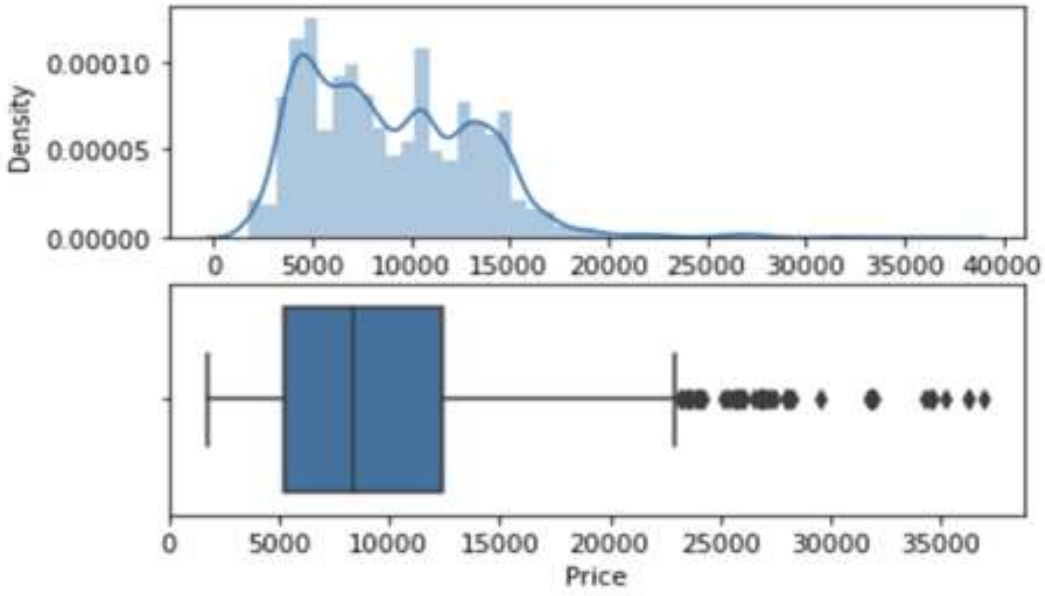
- Non stop: 0
- 1 stop : 1
- 2 stop : 2
- 3stop : 3
- 4 stop : 4

```
0          0
1          2
2          2
4          1
5          0
          ..
10678      0
10679      0
10680      0
10681      0
10682      2
Name: Total_Stops,
```

Finaly, We check for precence of any outlier in output variable ‘price’and visualizing it through boxplot. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.



It is evident that Price attrivute has some outliers out of the range 25000 Rs. They replaced with median of price of respective airlines.

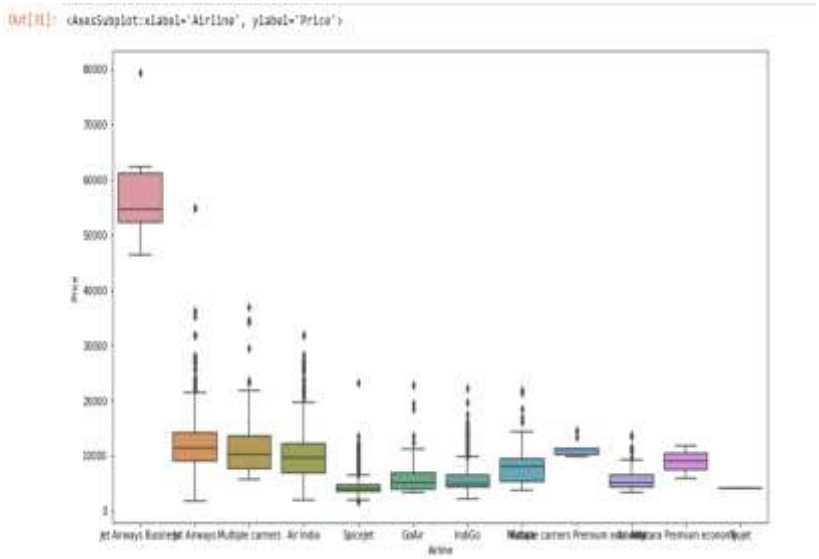


5.2. DATA VISUALIZATION:

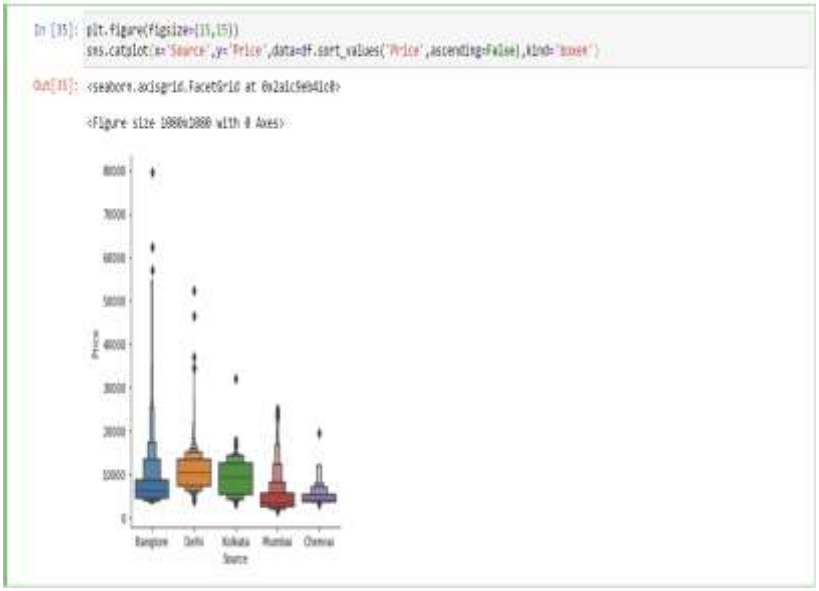
The graphic display of information and data is known as data visualisation. Data visualisation tools offer an easy approach to observe and analyse trends, outliers, and patterns in data by utilising visual elements like charts, graphs, and maps. Additionally, it offers a great way for staff members or business owners to clearly present data to non-technical audiences. To analyse vast amounts of data and make data-driven decisions, data visualisation tools and technologies are crucial in

the world of big data. In this Project, we have used the Catplot, Boxplot and Hexbin for the visualization of attribute with the price attribute.

Visualization of Airline with the price attribute, Jet Airways flights are most costliest Airlines compared to the other airlines.

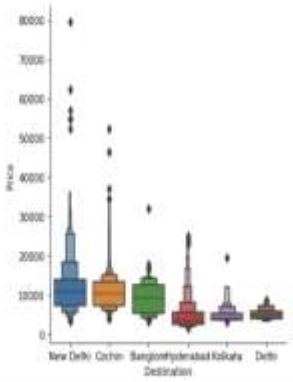


Visualization of Source with the price attribute using the CATPLOT, here the Bangalore has the highest number of source area than any other cities.



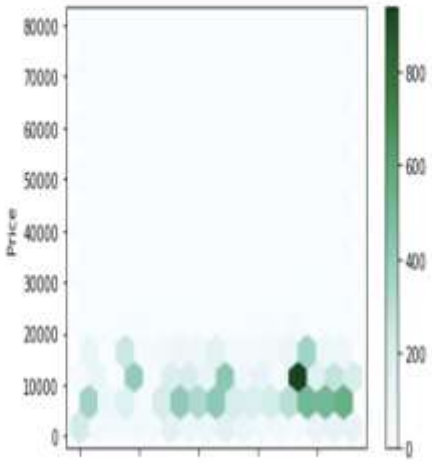
Visualization of Destination with the price attribute using the CATPLOT, here the New Delhi has the most expensive and higher range of flights than any other cities.

```
In [30]: plt.figure(figsize=(15,15))
         sns.catplot(x='Destination',y='Price',data=df.sort_values('Price',ascending=False),kind='boxen')
Out[30]: <seaborn.axisgrid.FacetGrid at 0x17fe587ba00>
         <Figure size 1080x1080 with 0 Axes>
```



Hexbin maps divide the region into numerous sections and assign each section a colour using hexagons.

```
In [43]: df.plot.hexbin(x='Arrival_Time_hour',y='Price',gridsize=15)
Out[43]: <AxesSubplot:xlabel='Arrival_Time_hour', ylabel='Price'>
```



5.3. MODEL SELECTION:

The process of choosing a statistical model from a group of potential models based on input data. In the simplest scenarios, an existing set of data is taken into account. However, the task might also entail planning experiments so that the information gathered is useful for the challenge of model choice. The simplest model is most likely to be the best option when there are several candidate models with comparable explanatory or predictive power.

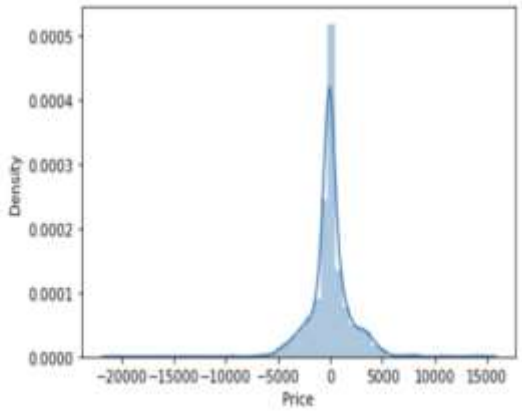
5.3.1. RANDOMFOREST REGRESSOR:

A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting numerous classification decision trees to different dataset subsamples. If bootstrap=True (the default), the size of the sub-sample is determined by the max samples argument; otherwise, each tree is constructed using the entire dataset.

```
In [70]: predict(RandomForestRegressor())

Model is: RandomForestRegressor()
Training score: 0.9512104456067212
Predictions are: [17294.42       5677.34866667 10943.12       ... 15732.17466667
 16362.82       8398.8       ]

r2 score is: 0.812763564948551
MAE:1208.790728373513
MSE:3976584.489480107
RMSE:1994.1375302320819
```

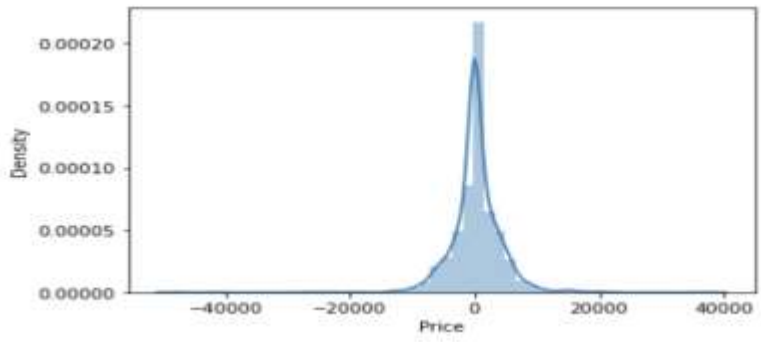


5.3.2. LOGISTIC REGRESSOR:

In regression analysis, logistic regression (also known as logit regression) estimates a logistic model's parameters (the coefficients in the linear combination). Formally, binary logistic regression has a single binary dependent variable (two classes, coded by an indicator variable) with the values "0" and "1," while the independent variables can either be continuous variables or binary variables (two classes, coded by an indicator variable) (any real value).

```
In [71]: predict(LogisticRegression())
Model is: LogisticRegression()
Training score: 0.23345111896348644
Predictions are: [ 5176.  5601. 12102. ...  4082.  7295.  7064.]

r2 score is: 0.15684607426687425
MAE:2537.584079133302
MSE:17907160.122939236
RMSE:4231.685258019462
```



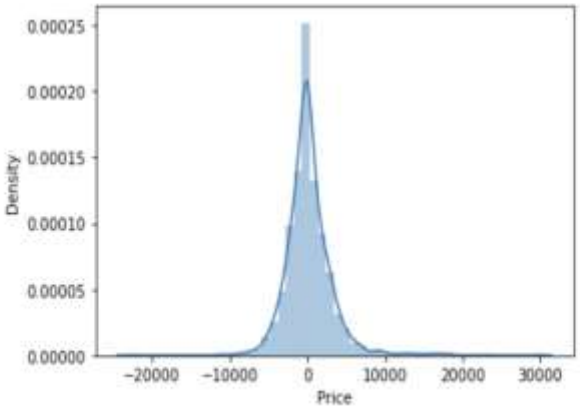
5.3.3. KNEIGHBORS REGRESSOR:

KNN regression is a non-parametric technique that, by averaging the observations in the same neighborhood, intuitively approximates the relationship between independent variables and the continuous outcome. The analyst must decide on the neighborhood's size, or cross-validation can be used to determine the size that minimises mean-squared error (we will see this later).

```
In [72]: predict(KNeighborsRegressor())

Model is: KNeighborsRegressor()
Training score: 0.7282959472943797
Predictions are: [ 5634.   5729.2 11177.  ... 12811.6  7514.4 14277. ]


r2 score is: 0.5504498892804601
MAE:1983.766556759303
MSE:9547682.3036081
RMSE:3089.9324108478654
```



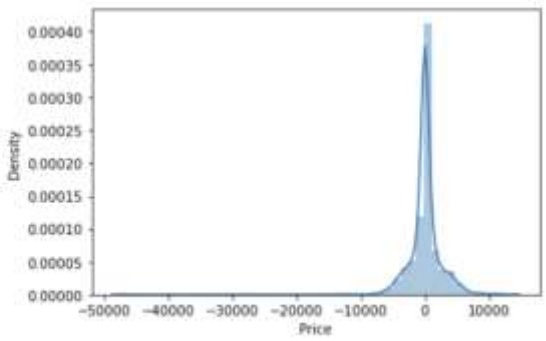
5.3.4. DECISIONTREE REGRESSOR:

It is possible to represent decisions and all of its potential outcomes, including outcomes, input costs, and utility, using a decision-making tool called a decision tree. The supervised learning algorithms group includes the decision-tree algorithm. It works with output variables that are categorised and continuous.

```
In [73]: predict(DecisionTreeRegressor())

Model is: DecisionTreeRegressor()
Training score: 0.9704555888812059
Predictions are: [12475.   6015.  10361.  ... 14173.5 21730.   8505. ]


r2 score is: 0.6905608419807476
MAE:1372.4114539174125
MSE:6571963.175217591
RMSE:2563.5840487913774
```



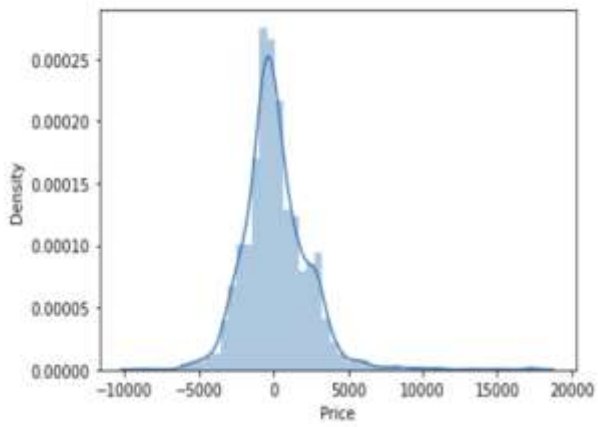
5.3.5. GRADIENTBOOST REGRESSOR:

This estimator allows for the optimization of any differentiable loss function and constructs an additive model in a forward stage-wise manner. A regression tree is fitted on the negative gradient of the provided loss function at each stage.

```
In [75]: predict(GradientBoostingRegressor())

Model is: GradientBoostingRegressor()
Training score: 0.7874628356205996
Predictions are: [11879.77370568  4577.98679055 11577.24366923 ... 16911.95712014
 16483.55109763  7626.7468659 ]


r2 score is: 0.779905062174256
MAE:1547.6021570558735
MSE:4674443.388811835
RMSE:2162.0461116294064
```



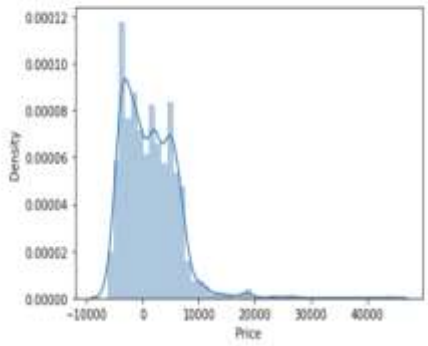
5.3.6. SUPPORT VECTOR REGRESSOR:

An algorithm for supervised learning called support vector regression is used to forecast discrete values. The SVMs and Support Vector Regression both operate on the same theory. Finding the best fit line is the fundamental tenet of SVR. The hyperplane with the most points is the best-fitting line in SVR.

```
In [74]: from sklearn.svm import SVR
         predict(SVR())

Model is: SVR()
Training score: 0.004628669035080413
Predictions are: [8071.00515898 8053.45769318 8499.60314614 ... 8277.6688812  8061.30423271
 8341.14093493]


r2 score is: -0.028334379892372752
MAE:3641.496767963967
MSE:21840079.063433904
RMSE:4673.337037210042
```



6. RESULTS AND EVALUATION

For evaluating all 6 models, we use evaluation metrics like MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R squared Value.

5.1. **Mean Absolute Error (MAE)** is the average of difference between the actual data value and the predicted data value. It is calculated as shown below:

$$MAE = (1/n) * \Sigma|y_i - x_i|$$

where:

- **Σ:** A Greek symbol that means "sum"
- **y_i:** The observed value for the i^{th} observation
- **x_i:** The predicted value for the i^{th} observation
- **n:** The total number of observations

6.2. Mean Squared Error is the average squared difference between the estimated values and the actual value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Where, n = Data set observations

Yi = Observation values

Y^i = Predicted Values

6.3. Root Mean Squared Error is the root of MSE.

$$\mathrm{RMSE} = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(S_i - O_i)^2}$$

Where, n = Data set observations

Si = Predicted values

Oi = Observations

6.4. R squared value is used for measuring the accuracy of the model.

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

Where,

R^2 = coefficient of determination

7. CONCLUSION:

A thorough investigation was conducted for this work using Kaggle dataset collection, and the Random Forest Machine Learning model was applied. We were able to identify the factors that have the greatest impact on airline ticket pricing using visualisation. The results of the experimental research show that the Random Forest Regression model has good accuracy. The objective for the future is to improve model accuracy and feature selection. In order to obtain more accurate airfares, we also intend to expand the study by working with larger datasets and conducting additional experiments on it. This will allow consumers to get an idea of how much their next flight will cost and enable them to negotiate the best deal.

8. FUTURE WORK:

With the help of training data, we then tested the test data against the training data. These entries were used to extract a number of properties. Our suggested model is capable of calculating the quarterly average flight price using attribute selection methods. By foreseeing what costs airlines can keep, this can be helpful. Customers may use it to forecast future flight costs and make travel plans accordingly. In the future, this framework might be expanded to include data on the price of airline tickets, which can offer more specifics about each area, such as entry and exit timestamps, seat assignments, covered auxiliary items, and other specifics.

9. REFERENCE:

1.Neel Bhosale , Pranav Gole , Hrutuja Handore , Priti Lakade , Gajanan Arsalwad “Flight Fare Prediction System Using Machine Learning” International Journal for Research in Applied Science & Engineering Technology (IJRASET) https://www.ijraset.com/best-journal/flight-fare-prediction-system-using-machine-learning

2. William grooves and maria gini “An Agent for Optimizing Airline Ticket Purchasing” https://www.researchgate.net/publication/262172314_An_agent_for_optimizing_airline_ticket_purchasing

3. ***J. Santos Dominguez-Menchero, Javier Rivera and Emilio Torres Manzanera "Optimal purchase timing in the airline market”***

https://www.researchgate.net/publication/264161565_Optimal_purchase_timing_in_the_airline_market

4. ***Supriya Rajankar, Neha sakhrakar and Omprakash rajankar "Flight fare prediction using machine learning algorithms" International journal of Engineering Research and Technology (IJERT) June 2019*** **https://www.ijraset.com/research-paper/flight-fare-prediction-system-using-ml**

5. ***Tianyi wang, samira Pouyanfar, haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach"*** International Journal for Research in Applied Science & Engineering Technology (IJRASET) https://www.ijraset.com/research-paper/implementation-of-flight-fare-prediction-system-using-ml

6. ***T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices"*** **https://www.cs.ru.nl/bachelors-theses/2014/Tim_Janssen___4150880___A_Linear_Quantile_Mixed_Regression_Model_for_Prediction_of_Airline_Ticket_Prices.pdf**

7. ***Wohlfarth, T.clemencon, S.Roueff "A Data mining approach to travel price forecasting" 10th international conference on machine learning Honolulu 2011.*** **https://www.researchgate.net/publication/233836406_A_Data-Mining_Approach_to_Travel_Price_Forecasting**

8. ***Vinod Kimbhaune, Harshil Donga, Ashutosh Trivedi, Sonam Mahajan and Viraj Mahajan research paper on "flight fare prediction system"*** **https://www.academia.edu/81611925/Implementation_of_Flight_Fare_Prediction_System_Using_Machine_Learning**

9. ***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.*** **https://experts.umn.edu/en/publications/an-agent-for-optimizing-airline-ticket-purchasing**

10. ***Viet Hoang Vu, Quang Tran Minh and Phu H. Phung,An "Airfare Prediction Model for Developing Markets", IEEE paper 2018.*** **https://ieeexplore.ieee.org/document/8343221**

11. ***Wohlfarth, T. Clemencon, S.Roueff.- "A Dat mining approach to travel price forecasting", 10 th international conference on machine learning Honolulu 2011.*** **https://hal.archives-ouvertes.fr/hal-00665041/document**

12. ***Dominguez-Menchero, J.Santo, Reviera,*** "Optimal purchase timing in airline markets" Journal of Air Transport Management, Elsevier, vol. 40(C), pages 137-143.

https://ideas.repec.org/a/eee/jaitra/v40y2014icp137-143.html

13.Gurucahran MK , "machine learning basic : decision tree regression". https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

14. ***O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates. "To buy or not to buy: mining airfare data to minimize ticket purchase price".*** **https://www.researchgate.net/publication/221654102_To_buy_or_not_to_buy_Mining_airfare_data_to_minimize_ticket_purchase_price**

15. ***Manolis Papadakis. "Predicting Airfare Prices".*** **https://cs229.stanford.edu/proj2012/Papadakis-PredictingAirfarePrices.pdf**

16. ***Groves and Gini, 2011. "A Regression Model for Predicting Optimal Purchase Timing for Airline Tickets".*** **https://conservancy.umn.edu/handle/11299/215872**

17. ***"Modelling of United States Airline Fares - Using the Official Airline Guide (OAG) and Airline Origin and Destination Survey (DBIB)" Krishna Rama Murthy, 2006.*** https://vtechworks.lib.vt.edu/handle/10919/35533

18. *Enrico Bachis* and *Claudio A. Piga.* Low-cost airlines and online price dispersion"https://econpapers.repec.org/article/eeeindorg/v_3a29_3ay_3a2011_3ai_3a6_3ap_3a655-667.htm

19. Tarun Devireddy, “Predicting Flight Prices in India Sectors”. https://www.scribd.com/document/519058496/Predicting-Flight-Prices-in-India-Sectors

20. Juhar Ahmed Abdella, , NM Zaki , , Khaled Shuaib , Fahad Khan , “Airline ticket price and demand prediction: A survey”. https://www.diva-portal.org/smash/get/diva2:1575609/FULLTEXT01.pdf

APPENDIX A:

CODING:

https://drive.google.com/drive/folders/1LTVVe3vgmPCkXKkYWvprSJbsI8IiFBIe?usp=sharing