

**Flight Price Prediction Project**

Submitted by:

N THARANI

**ACKNOWLEDGMENT**

Various research papers and websites are used. Research papers like Impact of Covid-19 on Automobile Industries in India - [A Xavier Susairaj](https://www.researchgate.net/profile/A-Xavier-Susairaj), [A. Salaijayamani](https://www.researchgate.net/profile/A-Salaijayamani), [Premkumar A.](https://www.researchgate.net/profile/Premkumar-A), [An impact study of COVID‐19 on six different industries - J Nayak,](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8014102/) Covid Economics, Covid-19 impact on transport, Impact of Covid 19 on the automotive market. Website like makemytrip, cleartrip, geeksforgeeks, economictimes, stackoverflow.com, etc., are used as references. The data is received from the client which is their own data.

Thanking SWATI MAHASETH, my guide from FLIPROBO TECHNOLOGIES for clearing all my doubts while undergoing the project.

**INTRODUCTION**

* Business Problem

There are a lot of changes in the car market because of the impact of covid 19. With the change in market due to covid 19 impact, everyone is facing problems with their flight price valuation.

Here we are helping people to find the cheapest fare or just telling how to and when to book the tickets to get the cheapest fare. The cheapest available ticket on a given flight gets more and less expensive over time. So, they are looking for machine learning models from the data we are going to scrape ourselves. We have to make flight price valuation model.

* Conceptual Background of the Domain Problem

The COVID-19 outbreak has spelt havoc on the economies of almost every country in the world. The situation is no different in India, which has been hit by the global pandemic quite hard. Like all the industries, even the aerospace industry has witnessed some heavy losses owing to the outbreak of this pandemic and the resultant lockdown.

“Prices of raw materials involving fuels, oils and PPE's are up enormously. Since this year, the input costs are up sharply. The price hike has to happen. Higher costs cannot be held back. However, we have to maintain a balance between customer expectations and business prudence,” said by the experts.

* Review of Literature

Certain websites and papers that helped me to take insights from are:

1. Impact of Covid-19 on in India - [A Xavier Susairaj](https://www.researchgate.net/profile/A-Xavier-Susairaj), [A. Salaijayamani](https://www.researchgate.net/profile/A-Salaijayamani), [Premkumar A.](https://www.researchgate.net/profile/Premkumar-A)
2. [An impact study of COVID‐19 on six different industries - J Nayak](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8014102/)
3. Covid Economics
4. Covid-19 impact on transport
5. Impact of Covid 19 on the automotive market

* Motivation for the Problem Undertaken

Flight ticket sales have came down during this pandemic period created by covid 19. People are very afraid to take up flights because of the contagiousness of the disease. This pandemic has affected airlines all over the world. An attempt to maximize revenue is always based on -

* Time of purchase patterns (making sure last-minute purchases are expensive)
* Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

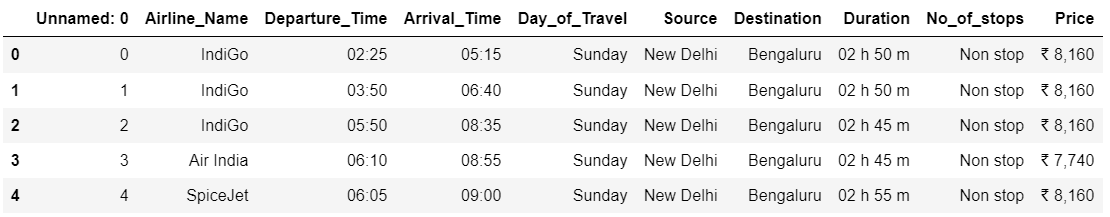
The main motivation of this project is to find the best time to book a flight ticket to get the cheapest fare. People is in need to travel through airlines for various purposes which needs to be fulfilled by finding a best model to find the best time to book a flight.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Various statistical, mathematical, analytical algorithms are used. Experimental design, outliers detection using Z-score, transformation, scaling, re-sampling methods, statistical hypothesis tests, estimation statistics, data manipulation, feature selection, feature reduction, balancing the dataset, etc are done for the problem.

* Data Sources and their formats

The data is scraped from makemytrip and cleartrip with necessary data description. The data is scraped and saved as an Excel file which is later used for model building. This data is saved in excel format in the local system. The data is about 2312 rows in total. The .xls format file can be imported using pandas from local library easily.

* Data Preprocessing Done

Columns like unnamed:0 are not necessary. Based on the data description, features which does not help to determine the car price detection are removed. unnamed:0 is a feature which has the same data for the whole dataset which will not help in determining the result. Convert price into integer type from string type. Check for null values, string data type to encode. Check for outliers and check the curve for normalization. Correlation and multi collinearity are checked and treated.

* Data Inputs- Logic- Output Relationships

There are various features used to detect the defaulter. All the features which are helping to detect the defaulter is visualised using matplotlib and seaborn. The relationship between the features are determined. After data cleaning, the relationship between input and output variables is fairly linear.

* State the set of assumptions (if any) related to the problem under consideration

In aerospace industry, everything is biased in nature on predicting the price of a ticket in the market. Here, we are predicting the price of the flight tickets which is biased based on each airlines company. So, certain measures are considered like economic falldown, pandemics, etc., So, predicting the price of flight tickets would not be so easy.

* Hardware and Software Requirements and Tools Used

Any laptop and computer can be used as hardware. Processor used is Intel(R) Core(TM) i7-4510U CPU. System type is 64-bit operating system, x-64 based processor. RAM of the systerm is 8.00 GB. Microsoft Windows 8.1 version 6.3 is the OS used. Python 2.7.10 is used with the interface Jupyter notebook with many installed libraries.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

There are statistical and analytical approaches in problem-solving. Data cleaning, data mining, wrangling comes under statistical approach whereas data manipulation, feature selection, feature reduction, creating ML models comes under analytical approach.

**Statistical:**

* + - ***Experimental Design*** - Methods to design systematic experiments to compare the effect of independent variables on an outcome
  + Univariate - measure relies on only one variable - Statistical summary, information on variables, etc
  + Bivariate - measure relies on two variables - Correlation, multicollinearity
    - ***Outliers detection using Z-Score (z = (x-μ)/σ)*** - where x is the observed value, μ is the population mean, and σ is the population standard deviation
    - ***Transformation*** - Transformed the data using ordinal encoder for the data to be analysed
    - ***Scaling*** - Scaled the data using Standard Scaler to get unbiased result.
    - ***Re-sampling Methods*** - Train Test Split is used to systematically split a dataset into subsets for the purposes of training and evaluating this predictive model.
    - ***Statistical Hypothesis Tests*** - Cross validation quantifies the likelihood of observing the result given an assumption or expectation about the result whether the model is overfitting/underfitting or fitting good.
    - ***Estimation Statistics*** - GridSearchCV is used to quantify the best parameter from the listed to fit in the model and give better result. It uses data analysis framework which has a combination of effect sizes, confidence intervals, precision planning, and meta-analysis to plan experiments, analyze data and interpret results.

**Analytical:**

It concerns the design and development of algorithms.

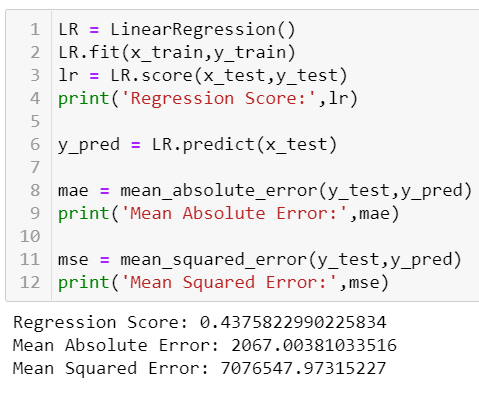
* ***Data Manipulation*** - Check for skewness and if there is no more skewed data, not necessary to treat it.
* ***Feature selection*** - Check the relationship between independent and dependent variables. Select the features which gives meaning to the problem based on the subject knowledge.
* ***Feature reduction*** - Reduce the features which does not give proper solution to the problem using Principal Component Analysis. But, the data is scraped by us. So, all the variables are in need for the analysis.
* ***Balancing the dataset*** - Balance the dataset and normalize the data to get better result.
* Testing of Identified Approaches (Algorithms)

Various evaluation metrics can be used for this regression type of model. Some of the popular algorithms are the following:

* Linear Regression
* k-Nearest Neighbors
* Decision Trees
* Support Vector Machine
* Random Forest Regression
* Run and Evaluate selected models

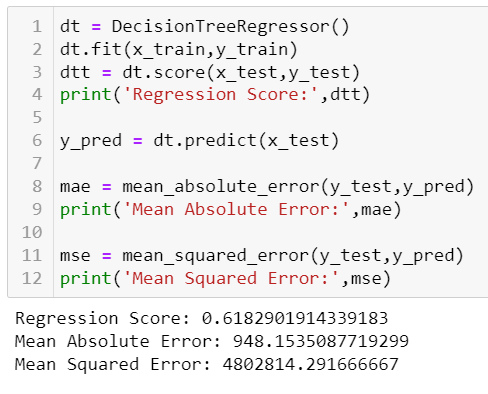
Regression Score - Percentage of correct predictions for dataset

***Linear Regression:***



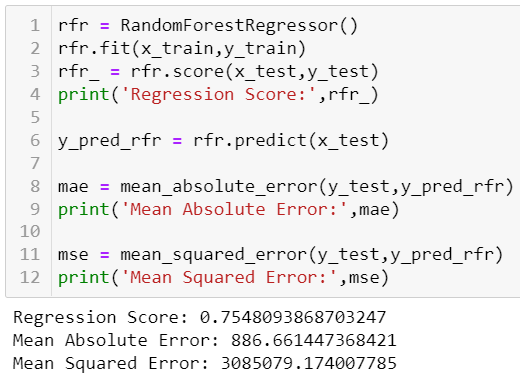
Regression Score - 0.46

***Decision Tree:***



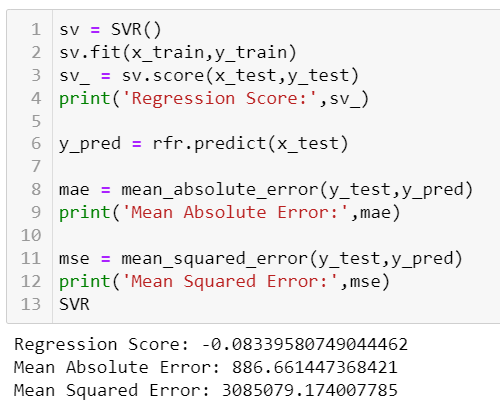
Regression Score - 0.89

***Random Forest Regressor:***



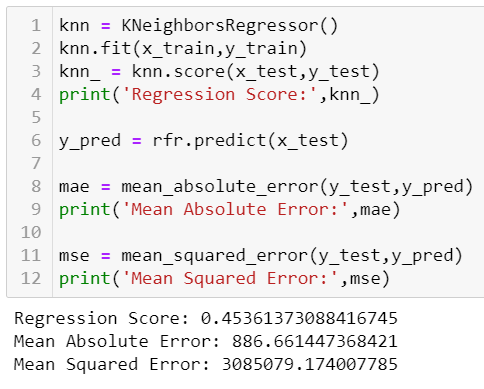
Regression Score - 0.93

***Support Vector Regression:***



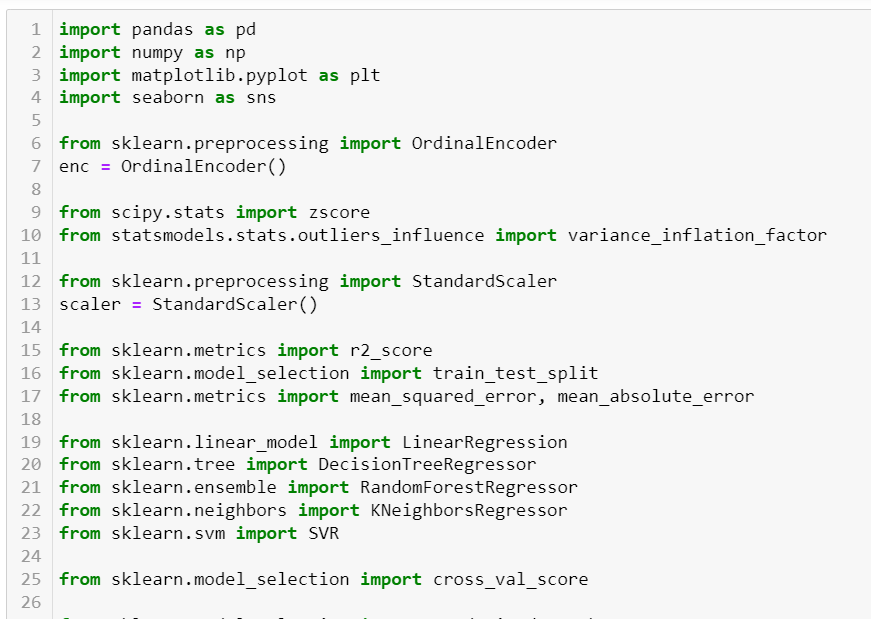
Regression Score - -0.05

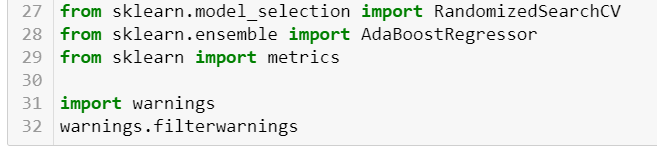
***KNeighbors Regressor:***



Regression Score - 0.11

* Key Metrics for success in solving problem under consideration

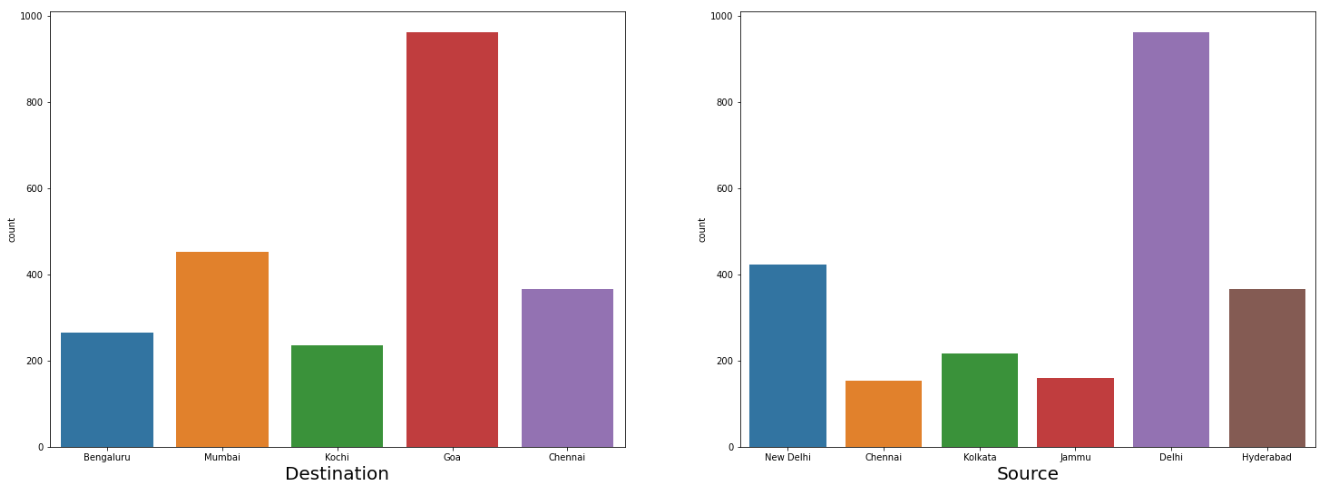


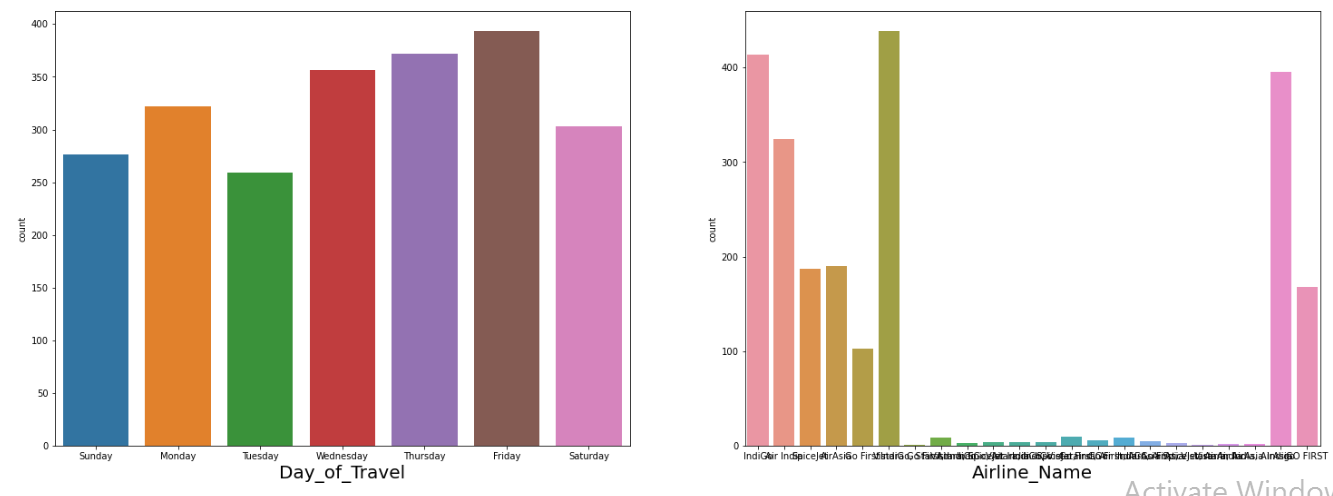


|  |  |
| --- | --- |
| **Libraries** | **Usage** |
| Pandas and numpy | Importing dataset, data cleaning, data wrangling and exploratory data analysis |
| Matplotlib and seaborn | Visualisation libraries |
| Ordinal Encoder | Encode the data from string to integers |
| Variance\_inflation\_factor | To check the multicollinearity |
| z-score | To check and remove outliers with fixed threshold |
| Standard Scaler | To standardise the data and normalise the curve |
| R2 score, mean\_squared\_error, mean\_absolute\_error | For concluding the results |
| Train\_test\_split | To separate the training and testing dataset |
| Linear Regression, DecisionTreeRegressor, RandomForestRegressor, SVR, KNeighborsRegressor | All these are machine learning algorithms to find the results |
| Cross\_val\_score | To check the best fitting of the model |
| RandomisedSearchCV | For hyper parameter tuning |
| AdaBoostRegressor | The weights of instances are adjusted and checks the fitting of model by applying the estimators |
| Metrics | It is used to track the performance by quantitatively assessing |

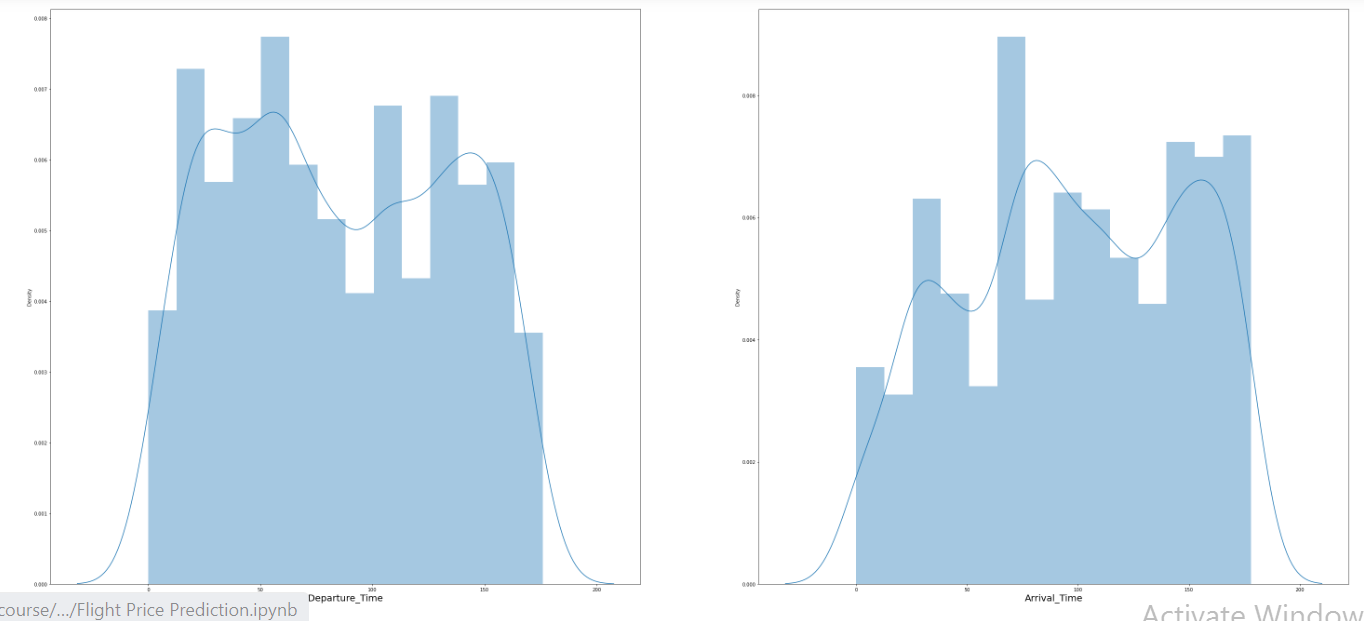
* Visualizations

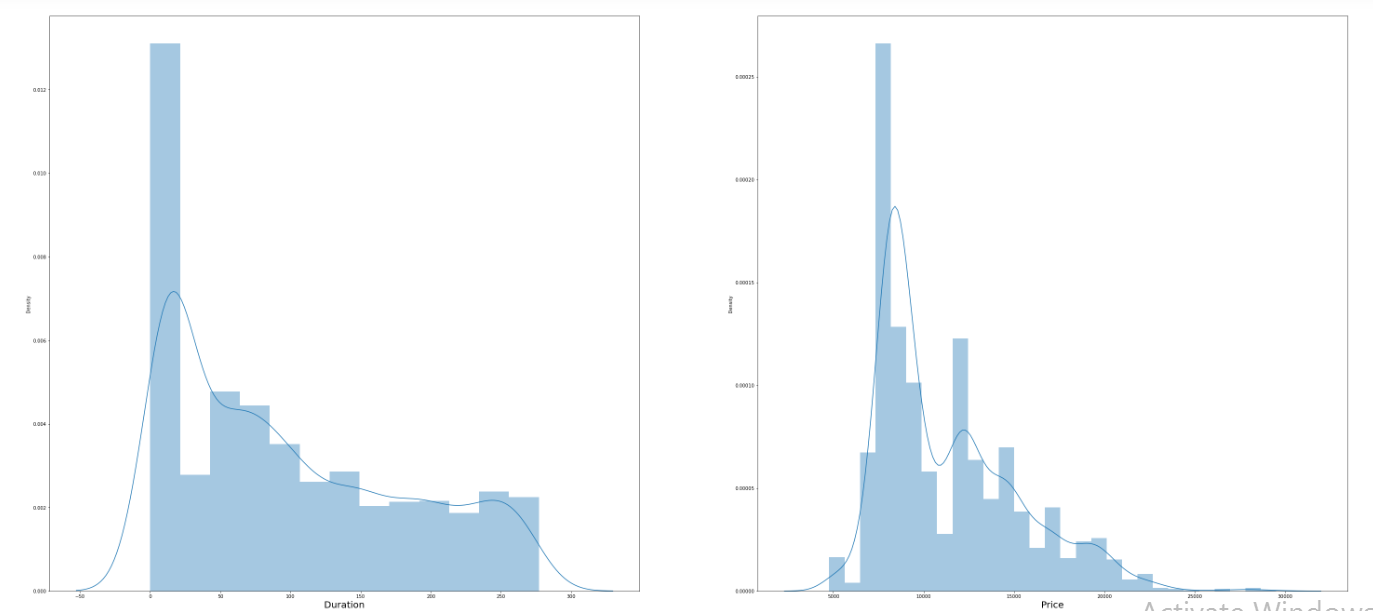
Matplotlib and seaborn is used for visualisations



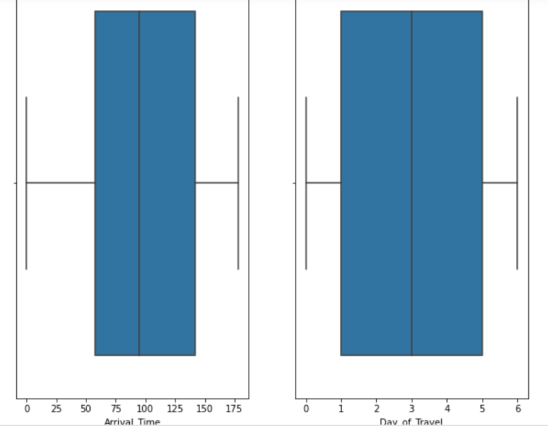
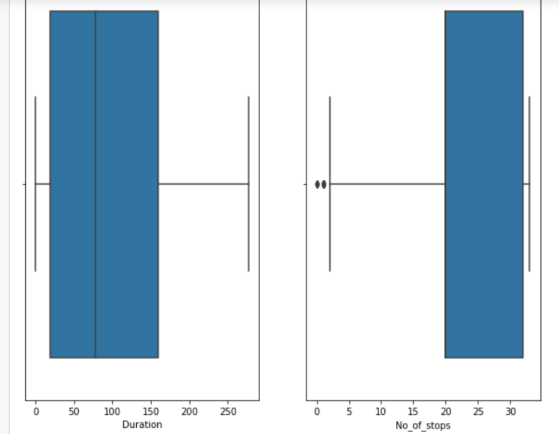


Categorical variables are visualized separately to understand the distribution where the target is imbalanced using countplot.





Continuous variables are visualized separately to understand the distribution where they are normally distributed.

To detect outliers boxplot is used which showed many outliers

* Interpretation of the Results
* Categorical variables and continuous variables are visualized separately to understand the distribution where the target is continuous in nature with the other variables are normally distributed
* To detect outliers boxplot is used which showed less outliers
* The relationship between dependent and independent variables are visualised where it looks fairly linear
* In data preprocessing, duplicates are removed which helps to get unbiased result

**CONCLUSION**

* Key Findings and Conclusions of the Study

The relation between the input and output variables looks fairly linear after data cleaning. There is no correlation and multicollinearity. From the models, random forest regressor gives better result and also on comparing the cross validation the model did not fit well.

Hyper parameter tuning is done to check whether the accuracy can be improved, and the accuracy has increased. So, use the tuned model to predict the flight ticket price.

* Learning Outcomes of the Study in respect of Data Science

Visualising the categorical and continuous variable tells how the variable is distributed and the balancing of the dataset. Checked the correlation. The data already does not have null value and the data is encoded using ordinal encoder.

Decision tree and Random Forest Regressor is good to go with the model. Random forest regression, KNN takes time for training. SVR takes very long time to train the model. Cross-validation took time to complete and checked whether the model is underfitting and overfitting the model.

Random Forest Regressor gave 75% Regression score approximately. After hyper parameter tuning, the model score has increased. So, we can use tuned model for further predictions.

* Limitations of this work and Scope for Future Work

We have got 71 as the regression score. There is scope to improve the cross validation score with the model score.