

Flight Price Prediction Project

Use Case Report



Submitted by:

Kirti Ohlan

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shwetank Mishra as well as Flip Robo Technologies who gave me the opportunity to do this project on Flight Price Prediction, which also helped me in doing lots of research wherein I came to know about so many new things.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1. <https://www.google.com/>
2. <https://www.youtube.com/>
3. <https://scikit-learn.org/stable/user_guide.html>
4. <https://github.com/>
5. <https://www.kaggle.com/>
6. <https://medium.com/>
7. <https://towardsdatascience.com/>
8. <https://www.analyticsvidhya.com/>

**INTRODUCTION**

* Business Problem Framing

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

Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That’s why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

* Review of Literature

It is critical for airlines to be capable of predicting airfare trends at the market segment level in order to alter strategies and resources for a given route. Scientific literatures on business segment price prediction, on the other hand, use biased conventional predictive methods, including such linear regression and thus are founded on the supposition that the selected variables have a linear relationship, that might not be true in most cases. Prediction of airfare prices utilizing machine learning approach, A dataset of Airways data flights was gathered and utilized to develop the machine learning technique for the study effort. Various figures of variables have been used to train the classifiers to demonstrate how feature extraction might affect validity of the model.

* Motivation for the Problem Undertaken

The tourism industry is changing fast and this is attracting a lot more travellers each year. The airline industry is considered as one of the most sophisticated industry in using complex pricing strategies. Now-a-days flight prices are quite unpredictable. The ticket prices change frequently. Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible. Using technology, it is actually possible to reduce the uncertainty of flight prices. So here we will be predicting the flight prices using efficient machine learning techniques.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We are building a model in Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. So, this model will help us to determine which variables are important to predict the price of variables & also how do these variables describe the price of the flight. This will help to determine the price of flights with the available independent variables.

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For specific mathematical reasons this allows the researcher to estimate the

conditional expectation of the dependent variable when the independent variables take on a given set of values.

Regression analysis is also a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

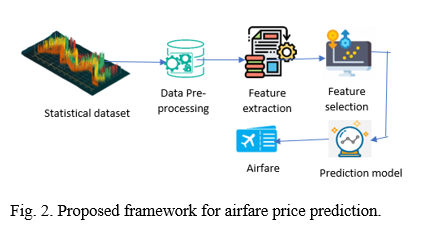
* Data Sources and their formats

The dataset is in CSV (Comma Separated File) format. Dataset is downloaded from kaggle and also scrap from different websites like skyscanner, happyfare, and others.

The model will learn from this file. It contains all the independent variables and the target variable.

Proposed Framework

Our suggested approach makes use of datasets to forecast airfare at the business segment levels. Fig 2 depicts a high-level view of the project framework's primary components. During the data pre-treatment stage, all databases are cleaned to remove any potentially erroneous examples, then converted and integrated depending on market group. The feature extractor extracts and generates handmade attributes that are intended to describe the segment of the market. The goal of adaptive filtering modules is to improve accurate channels by assessing the utility of the characteristics and removing any unnecessary characteristics. Finally, we use the selected criteria to build our forecasting techniques, that result in the finished product of the projected airline cost of the ticket.



* Data Preprocessing Done

Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre- process our data before feeding it into our model. Therefore, it is the first and crucial step while creating a machine learning model. I have used some following pre-processing steps:

1. Creating the dataset as a DataFrame.
2. Used pandas to set display I ensuring we do not see any truncated information.
3. Checked the number of rows and columns present in our training dataset.
4. Checked for missing data and the number of rows with null values
5. Duplicate values are also present in our dataset.
6. Checked the unique values information in each column to get a list for categorical data
7. Used Pandas Profiling during the visualization phase along with violin plot, count plot, cat plot and the others
8. With the help of ordinal encoding technique converted all object datatype columns to numeric datatype
9. Thoroughly checked for outliers and skewness information
10. With the help of heatmap, correlation bar graph was able to understand the Feature vs Label relativity and insights on multicollinearity amongst the feature columns
11. Separate feature and label data to ensure feature scaling is performed avoiding any kind of biasness
12. Checked for the best random state to be used for our Regression Machine Learning model pertaining to the feature importance details
13. Finally created a regression model function along with evaluation metrics to pass through various model formats

* Data Inputs- Logic- Output Relationships

When we loaded the dataset, we had to go through various data pre-processing steps to understand what was given to us and what we were expected to predict for the project. When it comes to logical part the domain expertise of understanding how real estate works and how we are supposed to cater to the customers came in handy to train the model with the modified input data. In Data Science community there is a saying “Garbage In Garbage Out” therefore we had to be very cautious and spent almost 80% of our project building time in understanding each and every aspect of the data how they were related to each other as well as our target label.

* With the objective of predicting flight prices accurately we had to make sure that a model was built that understood the customer priorities trending in the market imposing those norms when a relevant price tag was generated. I tried my best to retain as much data possible that was collected. I did not want to impute data and then cause a biasness in the machine learning model from values that did not come from real people.
* State the set of assumptions (if any) related to the problem under consideration

The assumption part for me was relying strictly on the data provided to me and taking into consideration that was obtained from real people surveyed for their preferences and how reasonable a price for flights with various features inclining to them were.

* Hardware and Software Requirements and Tools Used

Hardware Used:

1. RAM: 8 GB
2. CPU: AMD Ryzen 5 3550H with Radeon Vega Mobile Gfx 2.10 GHz
3. GPU: AMD Radeon ™ Vega 8 Graphics and NVIDIA GeForce GTX 1650 Ti

Software Used:

1. Programming language: Python
2. Distribution: Anaconda Navigator
3. Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

Pandas, NumPy, matplotlib, seaborn, scikit-learn and pandas profiling

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data and EDA to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine learning models.

For this project we need to predict the price of flights, means our target column is continuous so this is a regression problem. I have used various regression algorithms and tested for the prediction. By doing various evaluations I have selected K Neighbors as best suitable algorithm for our final model as it is giving good r2-score and least difference in r2-score and CV-score among all the algorithms used. Other regression algorithms are also giving me good accuracy but some are over-fitting and some are with under-fitting the results which may be because of less amount of data.

In order to get good performance as well as accuracy and to check my model from over-fitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuned the final model.

Once I was able to get my desired final model I ensured to save that model before I loaded the testing data and started performing the data pre-processing as the training dataset and obtaining the predicted price values out of the Regression Machine Learning Model.

* Testing of Identified Approaches (Algorithms)

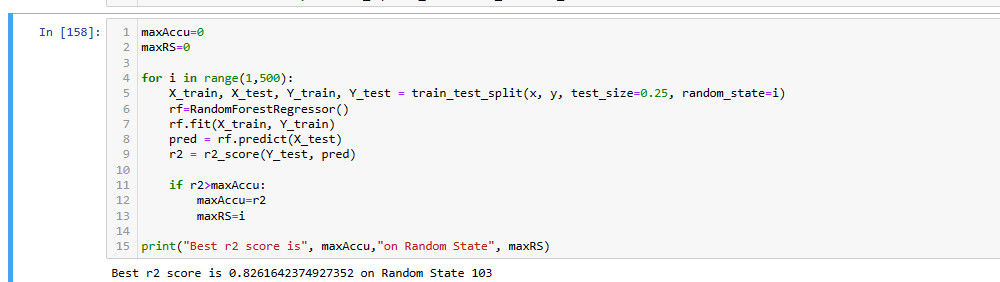
The algorithms used on training and test data are as follows:

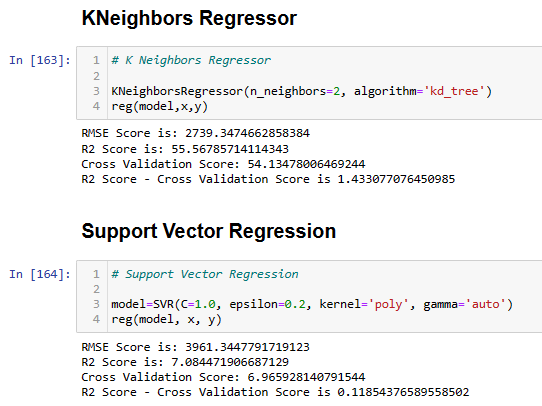
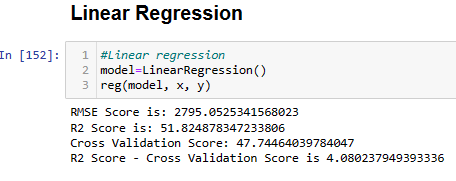
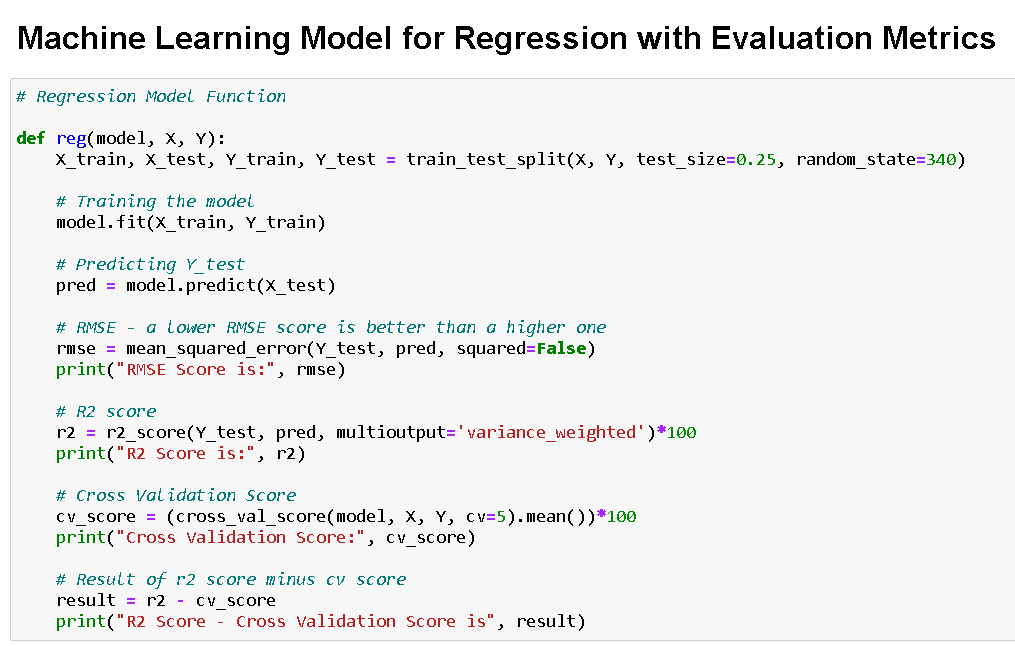
1. Linear Regression Model
2. Ridge Regularization Regression Model
3. Support Vector Regression Model
4. Decision Tree Regression Model
5. Random Forest Regression Model
6. K Nearest Neighbours Regression Model

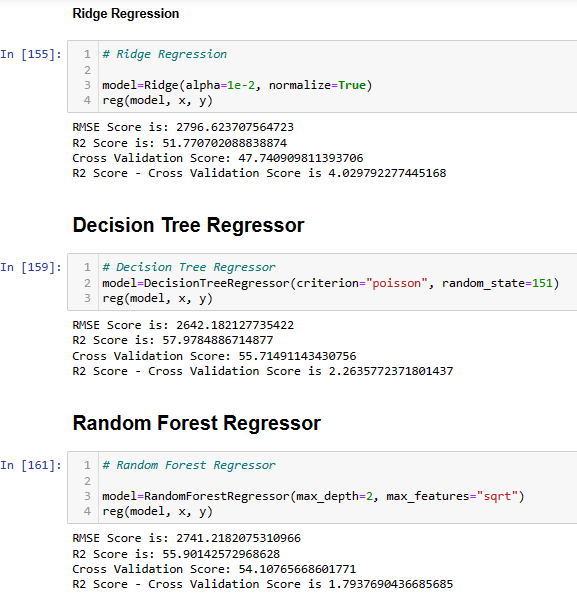
* Run and Evaluate selected models

I used a total of 6 Regression Models after choosing the random state amongst 1-500 number. Then I even defined a function for getting the regression model trained and evaluated. The code for the models is listed below.

Random State



Regression Model Function:



* Key Metrics for success in solving problem under consideration

The key metrics used here were r2\_score, cross\_val\_score, MAE, MSE and RMSE. We tried to find out the best parameters and also to increase our scores by using Hyperparameter Tuning and we will be using GridSearchCV method.

1. Cross Validation:

Cross-validation helps to find out the over fitting and under fitting of the model. In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 5 folds, the data will be divided into 5 pieces

where each part being 20% of full dataset. While running the Cross-validation the 1st part (20%) of the 5 parts will be kept out as a holdout set for validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset.

In the similar way further iterations are made for the second 20% of the dataset is held as a holdout set and remaining 4 parts are used for training data during process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during the cross-validation process to get the remaining estimate of the model quality.

2. R2 Score:

It is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.

3. Mean Squared Error (MSE):

MSE of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. RMSE is the Root Mean Squared Error.

4. Mean Absolute Error (MAE):

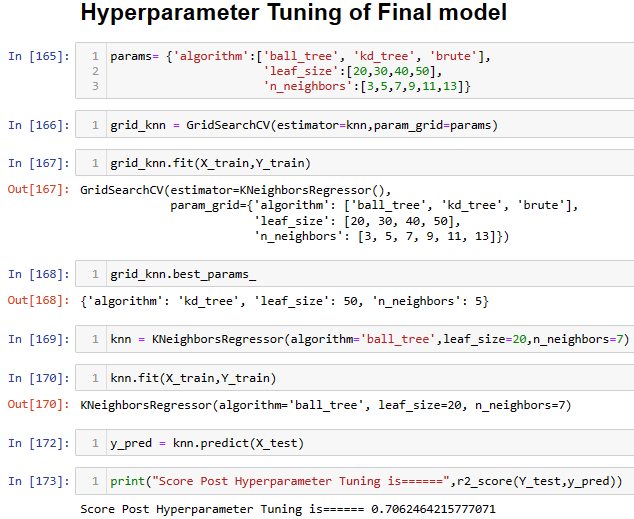
MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

5. Hyperparameter Tuning:

There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as Hyperparameters. These hyper parameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. You must select from a specific list of hyper parameters for a given model as it varies from model to model.

We are not aware of optimal values for hyper parameters which would generate the best model output. So, what we tell the model is to explore and select the optimal model architecture automatically. This selection procedure for hyper parameter is known as Hyper parameter Tuning. We can do tuning by using GridSearchCV.

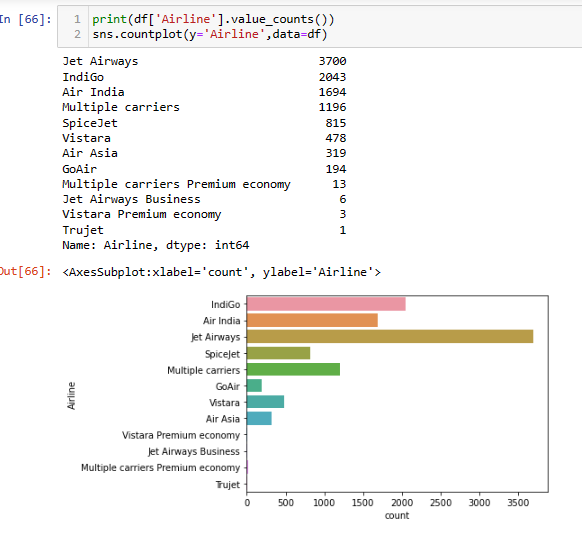
GridSearchCV is a function that comes in Scikit-learn (or SK-learn) model selection package. An important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyper parameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyper parameters.



It is possible that there are times when the default parameters perform better than the parameters list obtained from the tuning and it only indicates that there are more permutations and combinations that one needs to go through for obtaining better results.

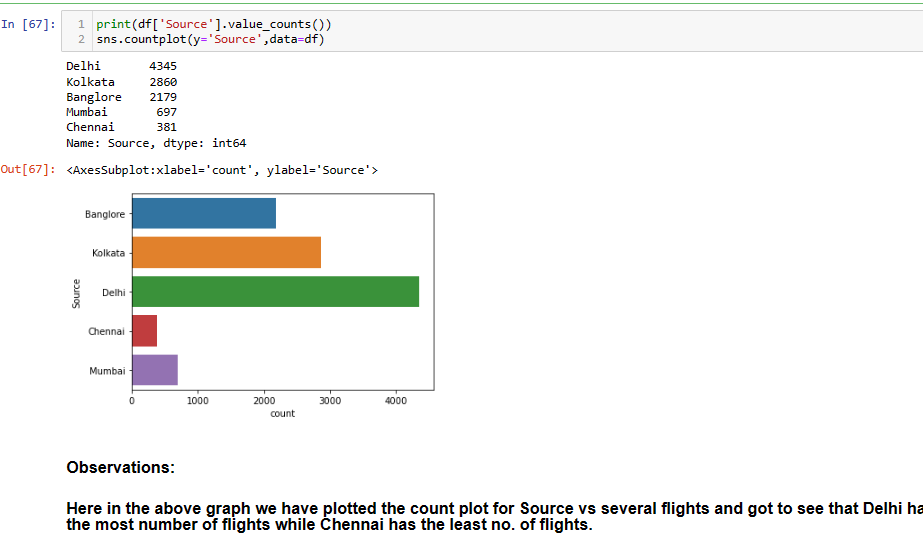
* Visualizations

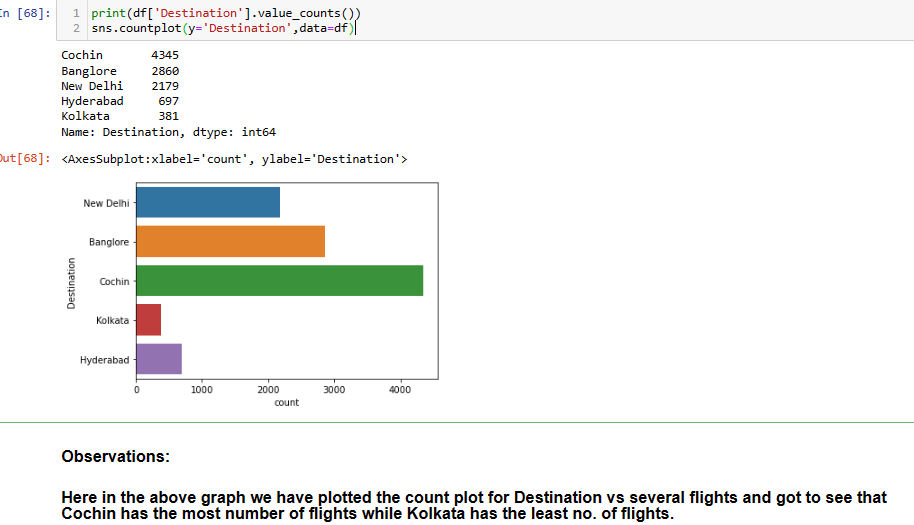
I created cat plots, count plots and violin plots to get further visual insights on our dataset feature values.

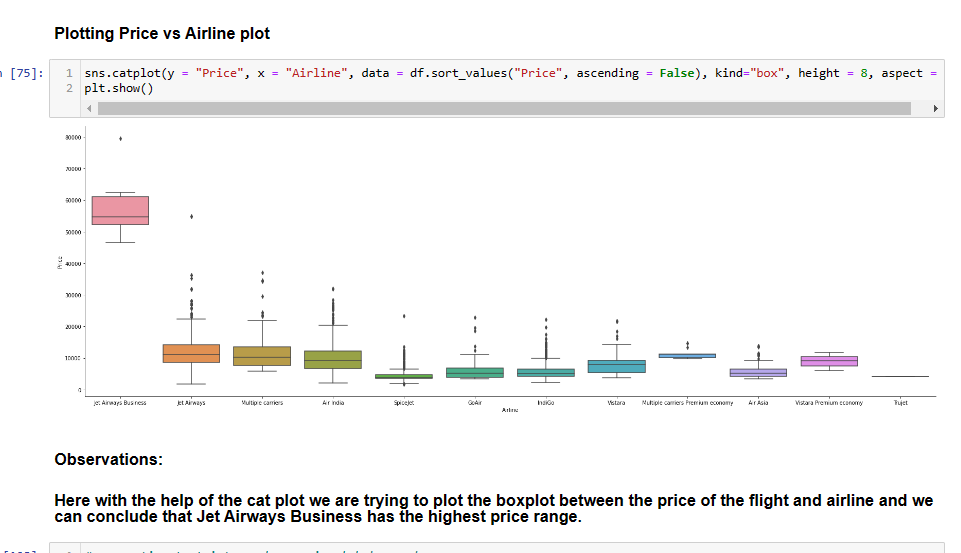


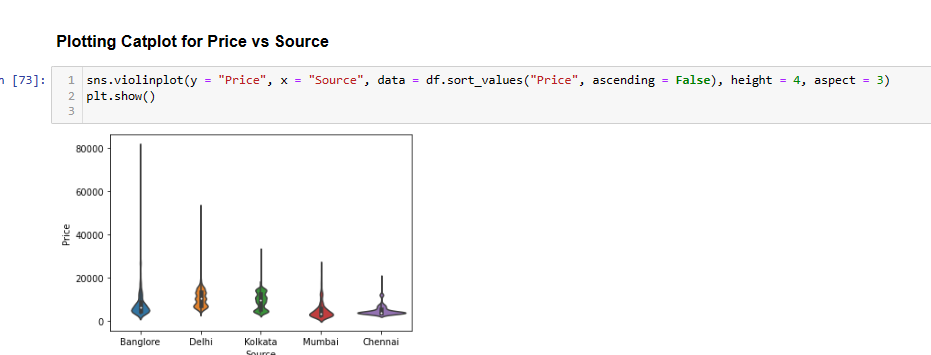
Observations:

Now from the above graph we can see that between the type of airline and count of flights we can see that Jet Airways has the most flight boarded









* Interpretation of the Results

Visualizations: It helped me to understand the correlation between independent and dependent features. Also, helped me with feature importance and to check for multi collinearity issues. Detected outliers/skewness with the help of boxplot and distribution plot. I got to know the count of a particular category for each feature by using count plot and most importantly with predicted target value distribution

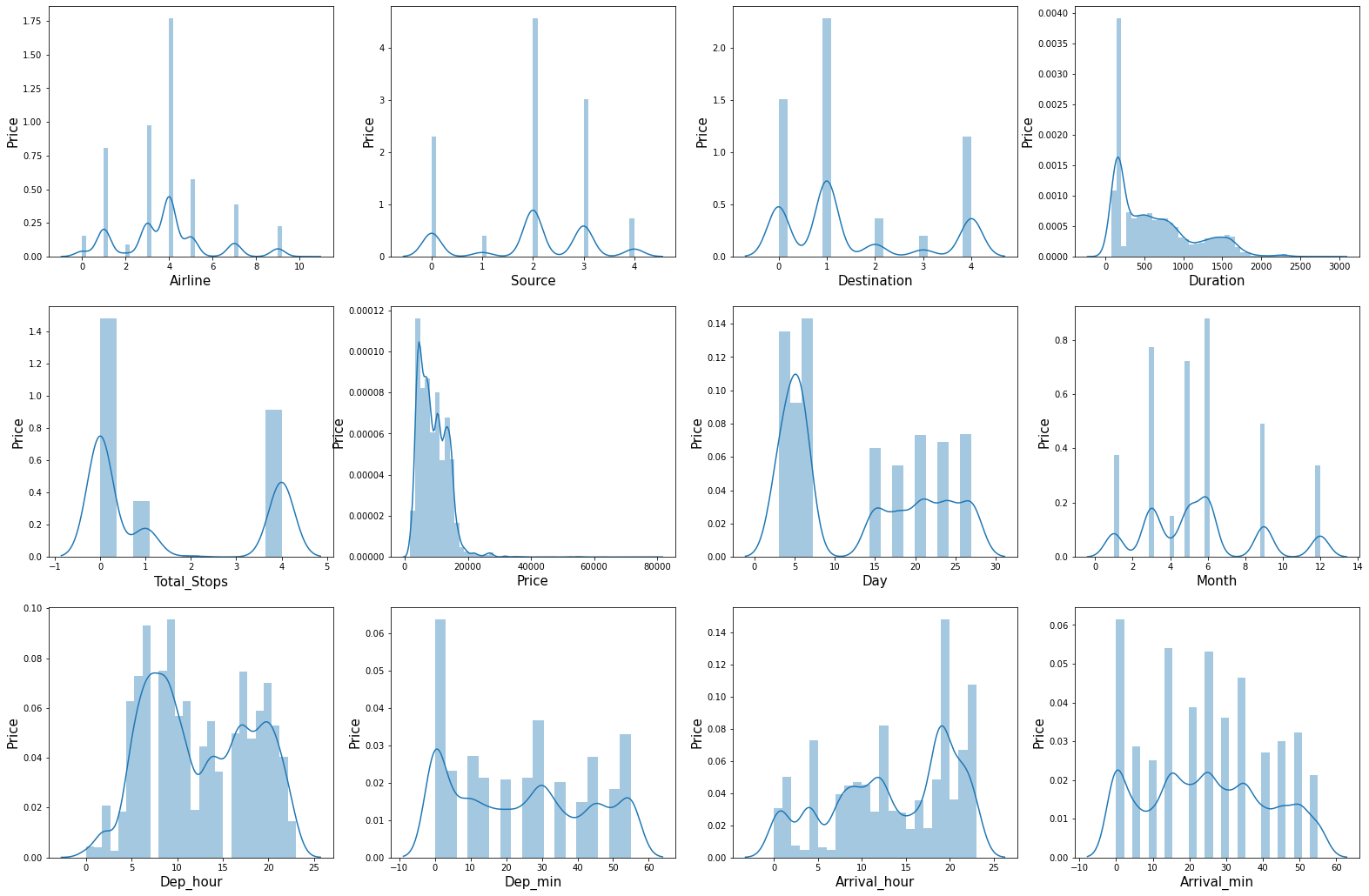
Pre-processing: Basically, before building the model the dataset should be cleaned and scaled by performing few steps. As I mentioned above in the pre-processing steps where all the important features are present in the dataset and ready for model building.

Model Creation: Now, after performing the train test split, I have x\_train, x\_test, y\_train&y\_test, which are required to build Machine learning models. I have built multiple regression models to get the best R2 score, MSE, RMSE & MAE out of all the models.

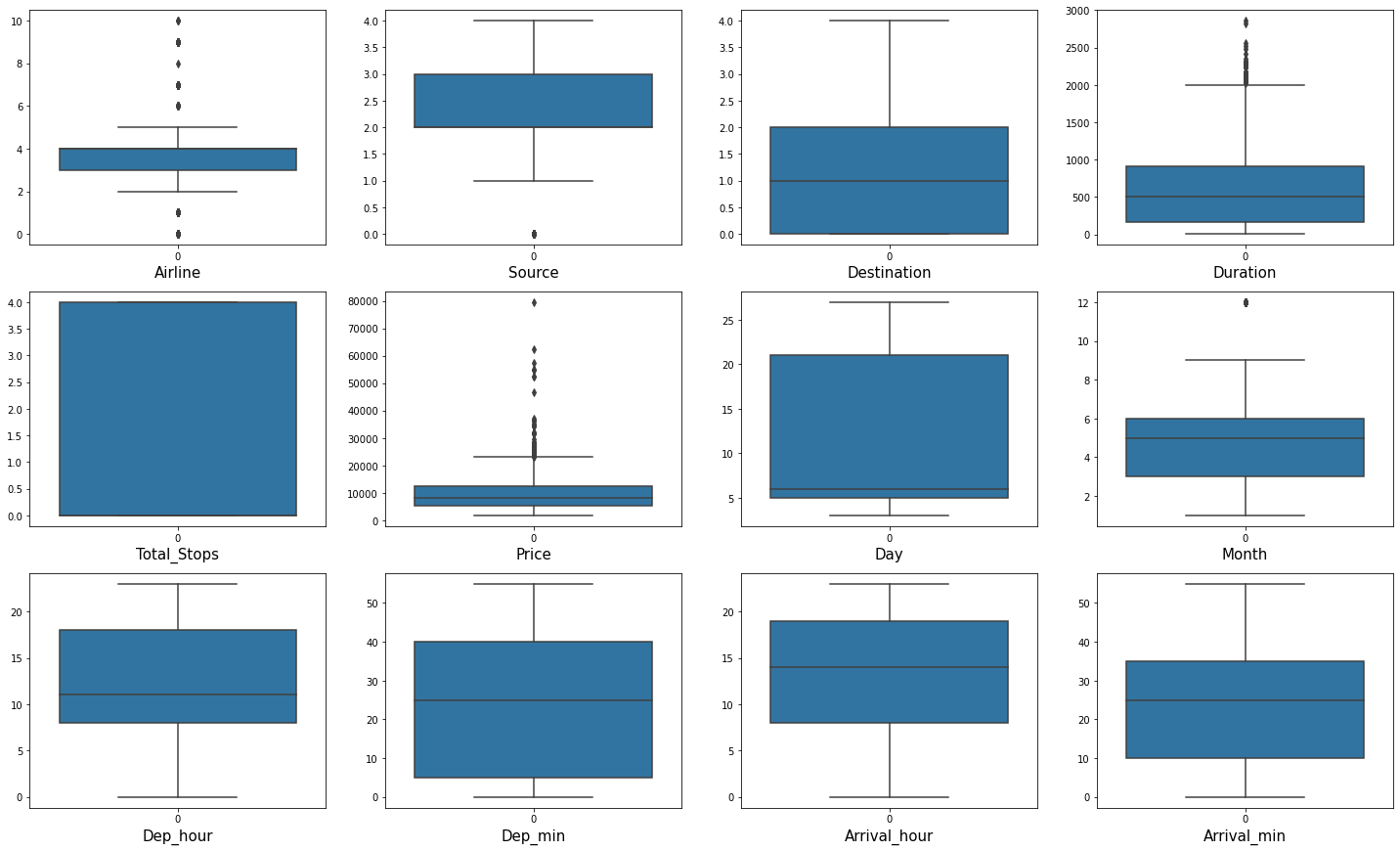
**CONCLUSION**

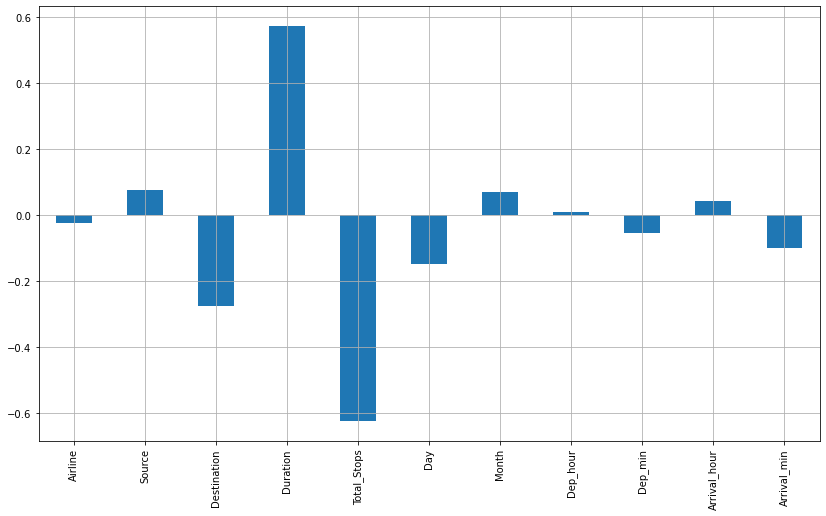
Key Findings and Conclusions of the Study

I observed all the encoded dataset information by plotting various graphs and visualised further insights.

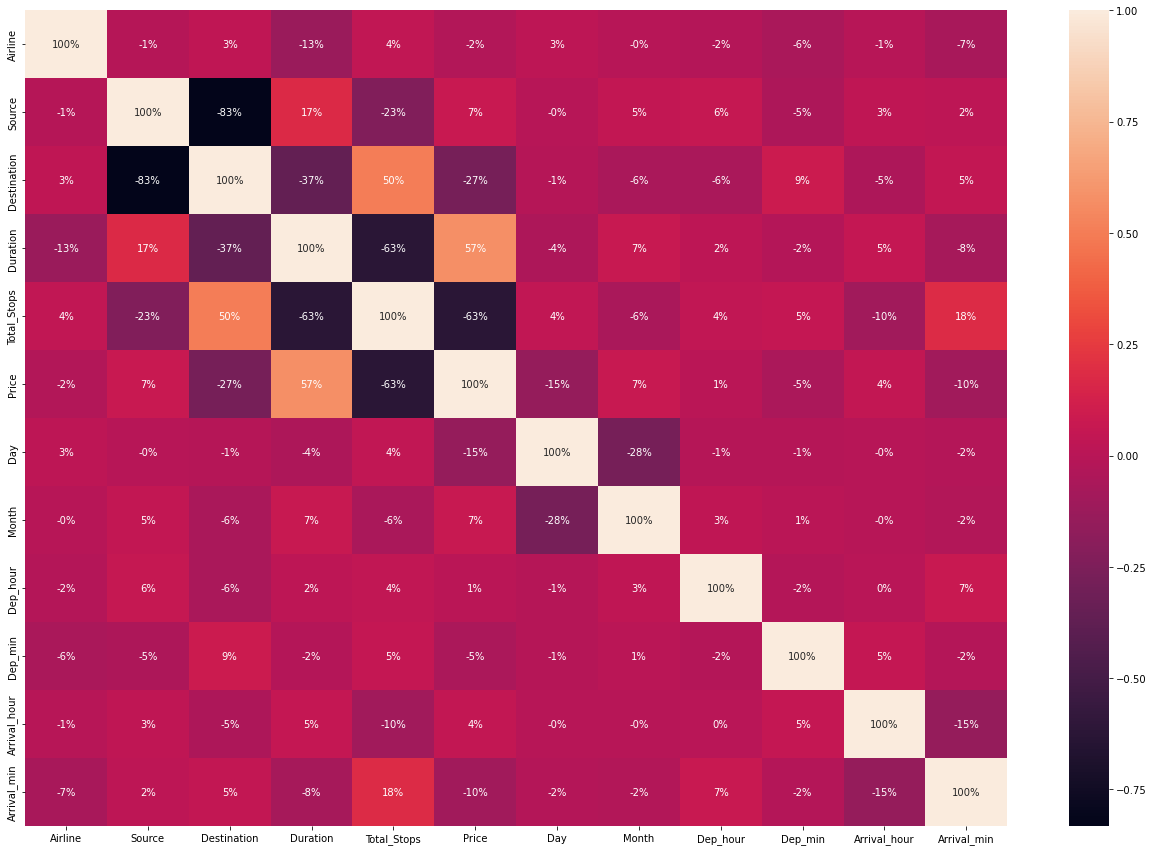
Distribution plot:

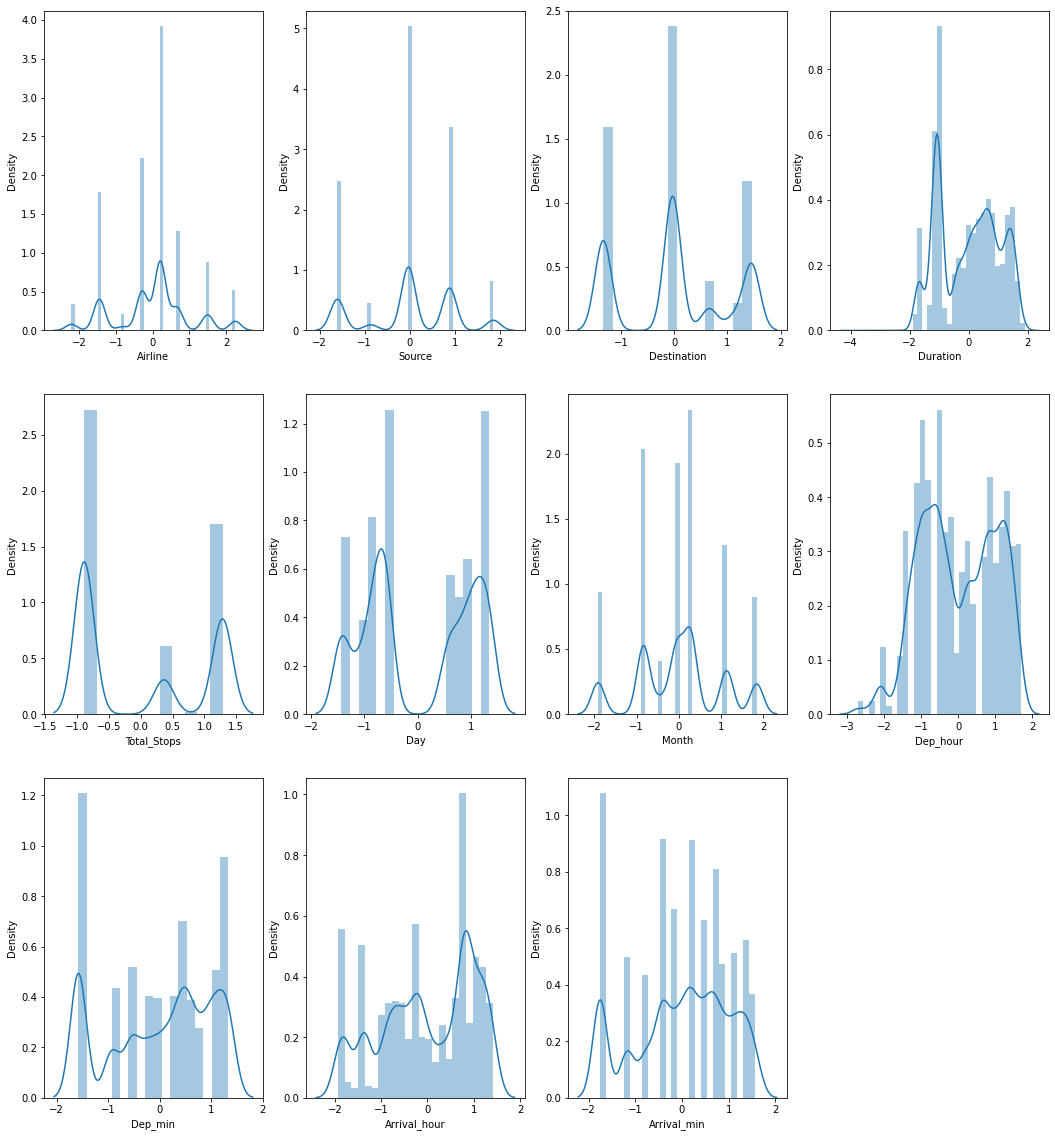
Boxplot:



Correlation:

Heatmap:



Distribution plot:

**Conclusion**

Five machine learning models were examined in this case study to forecast the average flight price at the business segment level. We used training data to train the training data and test data to test it. These records were used to extract a number of characteristics. Our suggested model can estimate the quarterly average flight price using attribute selection strategies. To the highest possible standard, much prior studies into flight price prediction using the large dataset depended on standard statistical approaches, which have their own limitations in terms of underlying issue estimates and hypotheses. To our knowledge, no other research have included statistics from holidays, celebrations, stock market price fluctuations, depression, fuel price, and socioeconomic information to estimate the air transport market sector; nonetheless, there are numerous restrictions. As example, neither of the databases provide precise information about ticket revenue, including such departing and arrival times and days of the week. This framework may be expanded in the future to also include airline tickets payment details, that can offer more detail about each area, such as timestamp of entry and exit, seat placement, covered auxiliary items, and so on. By merging such data, it is feasible to create a more robust and complete daily and even daily flight price forecast model. Furthermore, a huge surge of big commuters triggered by some unique events might alter flight costs in a market sector. Thus, incident data will be gathered from a variety of sources, including social media sites and media organizations, to supplement our forecasting models. We will also examine specific technological Models, such as Deeper Learning methods, meanwhile striving to enhance existing models by modifying their hyper-parameters to get the optimum design for airline price prediction.