

Project: - **Flight Price Prediction**

Submitted by:

HARSH RAJ GUPTA

**ACKNOWLEDGMENT**

I would like to express my sincere thanks of gratitude to my SME as well as “Flip Robo Technologies” team for letting me work on “Flight Price Prediction” project. This project also helped me in doing lots of research wherein I came to know about so many new things.

**INTRODUCTION**

* **Business Problem Framing**
* Airline industry is one of the most sophisticated in its use of dynamic pricing strategies to maximize revenue, based on proprietary algorithms and hidden variables. That is why the airline companies use complex algorithms to calculate the flight ticket prices. There are several different factors on which the price of the flight ticket depends. The seller has information about all the factors, but buyers are able to access limited information only which is not enough to predict the airfare prices. Considering the features such as departure time, arrival time and time of the day it will give the best time to buy the ticket.
* 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 models 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.
* **Conceptual Background of the Domain Problem**
* Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable.
* Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on -
* 1. Time of purchase patterns (making sure last-minute purchases are expensive).
* 2. 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).
* Here we are trying to help the buyers to understand the price of the flight tickets by deploying machine learning models. These models would help the sellers/buyers to understand the flight ticket prices in market and accordingly they would be able to book their tickets.
* Review of Literature
* In this study, we discuss various applications and methods which inspired us to build our supervised ML techniques to predict the price of flight tickets in different locations. We did a background survey regarding the basic ideas of our project and used those ideas for the collection of data information by doing web scraping from [www.yatra.com](http://www.yatra.com) **website which is a** web platform where buyers can book their flight tickets.
* This project is more about data exploration, feature engineering and pre-processing that can be done on this data. Since we scrape huge amount of data that includes more flight related features, we can do better data exploration and derive some interesting features using the available columns. Different techniques like ensemble techniques, and decision trees have been used to make the predictions.
* The goal of this project is to build an application which can predict the price of flight tickets with the help of other features. In the long term, this would allow people to better explain and reviewing their purchase in this increasing digital world.
* Motivation for the Problem Undertaken

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airline’s use using sophisticated quasi-academic tactics known as "revenue management" or "yield management". The cheapest available ticket for a given date gets more or less expensive over time. This usually happens as an attempt to maximize revenue based on –

1. Time of purchase patterns (making sure last-minute purchases are expensive)

2. 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) So, if we could inform the travellers with the optimal time to buy their flight tickets based on the historic data and show them various trends in the airline industry, we could help them save money on their travels. This would be a practical implementation of a data analysis, statistics and machine learning techniques to solve a daily problem faced by travellers. The objectives of the project can broadly be laid down by the following questions –

1. Flight Trends Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time?

2. Best Time to Buy What is the best time to buy so that the consumer can save the most by taking the least risk? So should a passenger wait to buy his ticket, or should he buy as early as possible?

3. Verifying Myths Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**
* We need to develop an efficient and effective Machine Learning model which predicts the price of flight tickets. So, “Price” is our target variable which is continuous in nature. Clearly it is a Regression problem where we need to use regression algorithms to predict the results. This project is done on three phases:
* Data Collection Phase: **I have done web scraping to collect the data of flights from the well-known website**[**www.yatra.com**](http://www.yatra.com)**where I found more features of flights compared to other websites and I fetch data for different locations. As per the requirement we need to build the model to predict the prices of flight tickets.**
* Data Analysis: After cleaning the data I have done some analysis on the data by using different types of visualizations.
* Model Building Phase: After collecting the data, I built a machine learning model. Before model building, have done all data pre-processing steps. The complete life cycle of data science that I have used in this project are as follows:
* Data Cleaning
* Exploratory Data Analysis
* Data Pre-processing
* Model Building
* Model Evaluation
* Selecting the best model
* Data Sources and their formats
  + What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.
* Data Preprocessing Done
* We have collected the dataset from the website [www.yatra.com](http://www.yatra.com)**which is a web platform where** the people can purchase/book their flight tickets. The data is scraped using Web scraping technique and the framework used is Selenium. We scrapped nearly 5303 of the data and fetched the data for different locations and collected the information of different features of the flights and saved the collected data in excel format. The dimension of the dataset is 3542 rows and 9 columns including target variable “Price”. The dataset contains both categorical and numerical data type.
* The data description is as follows: -

1. Airline: The name of the airline.
2. Source: The source from which the service begins.
3. Destination: The destination where the service
4. Dep\_Time: The time when the journey starts from the airport
5. Arrival Time: Time of arrival at the airport
6. Duration. Total duration of the journey
7. Total Stops: Total stops between the source and
8. Meal availability: meals
9. Price: The price of the ticket

* Data Inputs- Logic- Output Relationships
* Data pre-processing is the process of converting raw data into a well-readable format to be used by Machine Learning model. 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. I have used following pre-processing steps:
* Checked some statistical information like shape, number of unique values present, info, unique (), data types, value count function etc.
* Checked null values and found no missing values in the dataset.
* Taking care of Timestamp variables by converting data types of “Departure time” and “Time\_of\_arrival” from object data type into datetime data types.
* Done feature engineering on some features as they had some irrelevant values like “,”, “:” and replaced them by empty space.
* The column Duration had values in terms of minutes and hours. Duration means the time taken by the plane to reach the destination and it is the difference between the arrival time and Departure time. So, I have extracted proper duration time in terms of float data type from arrival and departure time columns.
* Extracted Departure Hour, Deparutre\_Min and Arrival Hour, Arrival\_Min columns from Departure time and Time\_of\_arrival columns and dropped these columns after extraction.
* The target variable "price" should be continuous numeric data but due to some string values like “,” it was showing as object data type. So, I replaced this sign by empty space and converted into float data type.
* From the value count function of Meal availability we observed "eCash 250" entry which does not belong to meals so I have replaced it as "None" and grouped same categories.
* From the value count function of Total stops I found categorical data so replaced them with numeric data according to stops.
* Checked statistical description of the data and separated categorical and numeric features.
* Identified outliers using box plots and found no outliers.
* Checked for skewness and removed skewness in numerical column “Duration” using square root transformation method.

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

* The dataset consists of label and features. The features are independent and label is dependent as the values of our independent variables changes as our label varies.
* Since we had both numerical and categorical columns, I checked the distribution of skewness using dist plots for numerical features and checked the counts using count plots & pie plots for categorical features as a part of univariate analysis.
* To analyse the relation between features and label I have used many plotting techniques where I found numerical continuous variables having some relation with label Price with the help of categorical and line plot.
* Hardware and Software Requirements and Tools Used

1. Jupyter Notebook 6.1.4 •
2. Selenium – To manipulate the HTML code
3. Pandas, NumPy – To analyse the data
4. Matplotlib, Seaborn – To visualize the data
5. ML Libraries

Text

Description automatically generated

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* For this project we need to predict flight ticket prices. In this dataset, “Price” is the target variable, which means our target column is continuous in nature, so this is a regression problem. I have used many regression algorithms and predicted the flight ticket price. By doing various evaluations I have selected **Random Forest Regressor** as best suitable algorithm to create our final model as it is giving high R2 score and low evaluation error among all the algorithms used. Performed hyper parameter tuning on best model. Then I saved my final model and loaded the same for predictions.
* Testing of Identified Approaches (Algorithms)
* Since “Price” is my target variable, which is continuous in nature, from this I can conclude that it is a regression type problem hence I have used following regression algorithms. After the pre-processing and data cleaning I left with 11 columns including target and with the help of feature importance bar graph I used these independent features for model building and prediction. The algorithms used on training the data are as follows:

1. **Ridge**
2. **Decision Tree Regressor**
3. **Random Forest Regressor**
4. **XGBRegressor**
5. **GradientBoostingRegressor**
6. **KNeighborsRegressor**

* **Run and evaluate selected models**

1. Ridge

Created Ridge model and checked for its evaluation metrics. The model is giving R2 score as 51.40%.

Graphical user interface, text, application, email

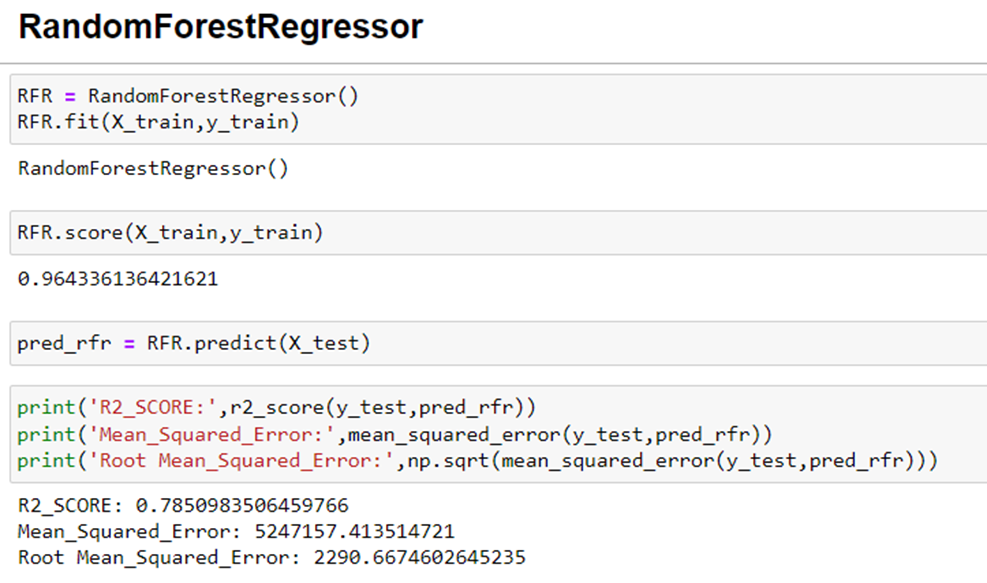
Description automatically generated

1. **Decision Tree Regressor**
2. Created Decision Tree Regressor model and checked for its evaluation metrics. The model is giving R2 score as 52.79%.

Graphical user interface, text, application, email

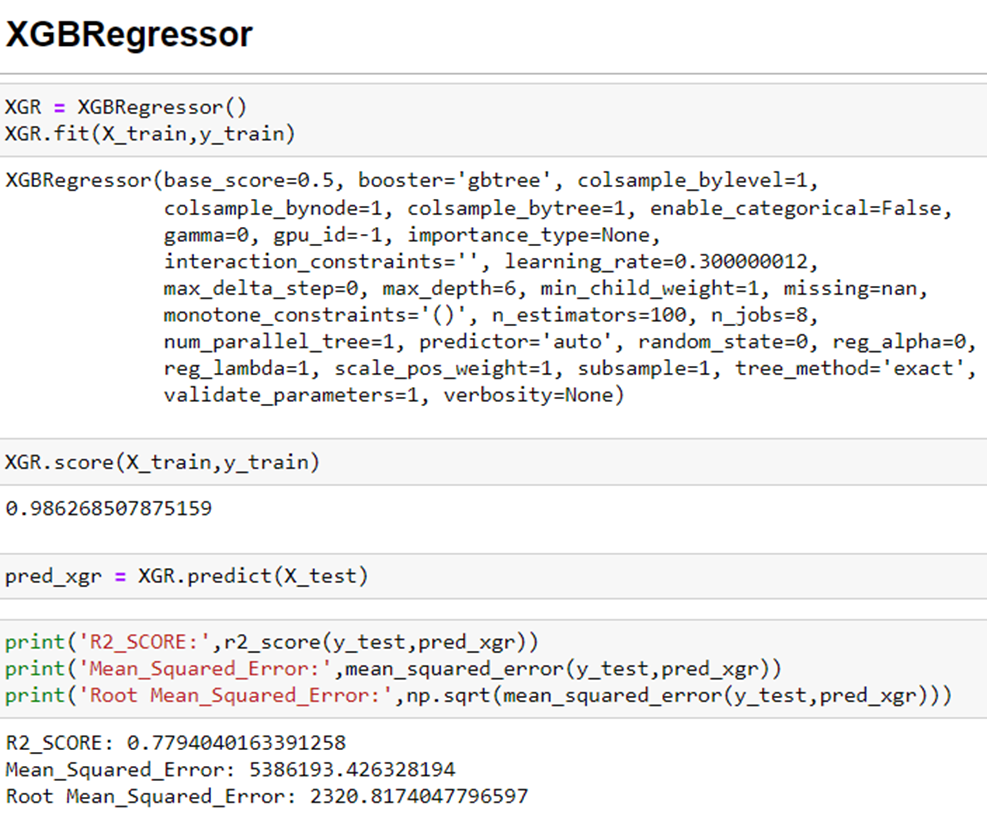
Description automatically generated

**3.Random Forest Regressor**

* Created Random Forest Regressor model and checked for its evaluation metrics. The model is giving R2 score as 78.50%.
* 

1. **XGBRegressor**

Created XGB Regressor model and checked for its evaluation metrics. The model is giving R2 score as 77.97%.

****

1. **Gradient Boosting Regressor**

Created Gradient Boosting Regressor model and checked for its evaluation metrics. The model is giving R2 score as 69.54%.

Graphical user interface, text, application, email

Description automatically generated

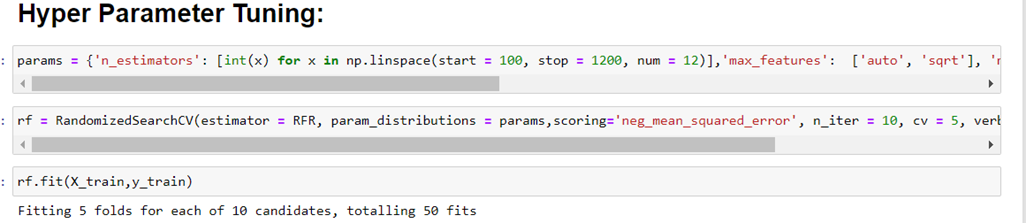
**6.KNeighborsRegressor**

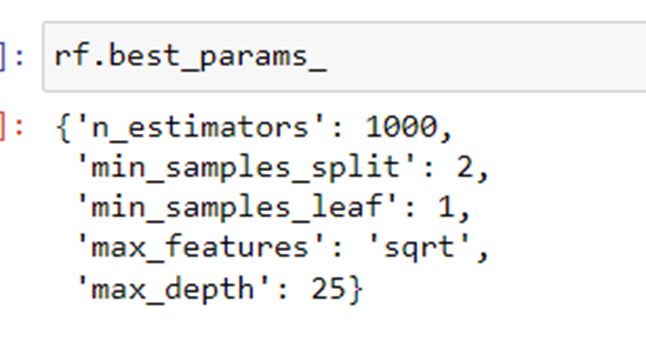
* Created KNeighbour Regressor model and checked for its evaluation metrics. The model is giving R2 score as 16.03%.
* Graphical user interface, text, application, email

  Description automatically generated

**Hyperparameter Tuning:**

* From the above created models, Random Tree Regressor algorithm has high R2 score and less RMSE value. So, we can conclude that "Random Tree Regressor" as the best fitting model. Let's try to increase our model score by tuning the best model using different types of hyper parameters.

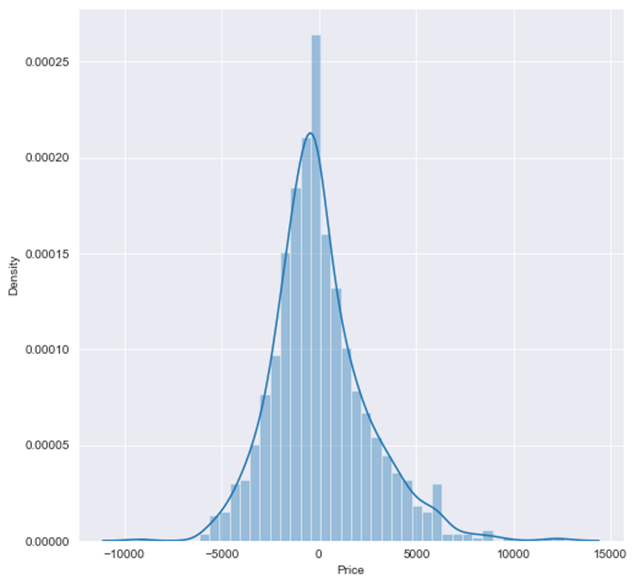
****

****

**Creating Final Model After Tuning:**

* **I have successfully incorporated the hyper parameter tuning using best parameters of Random Forest Regressor and the R2 score of the model has been increased after hyperparameter tuning and received the R2 score as 75.56% which is very good.**

**From the graph we can observe how our final model is mapping**

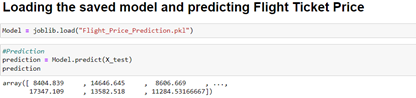
****

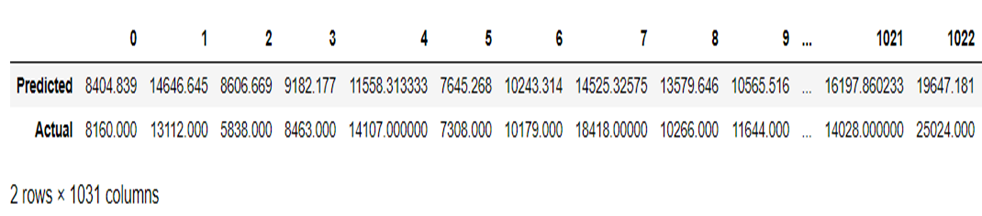
**Graphical user interface, application, Word

Description automatically generated**

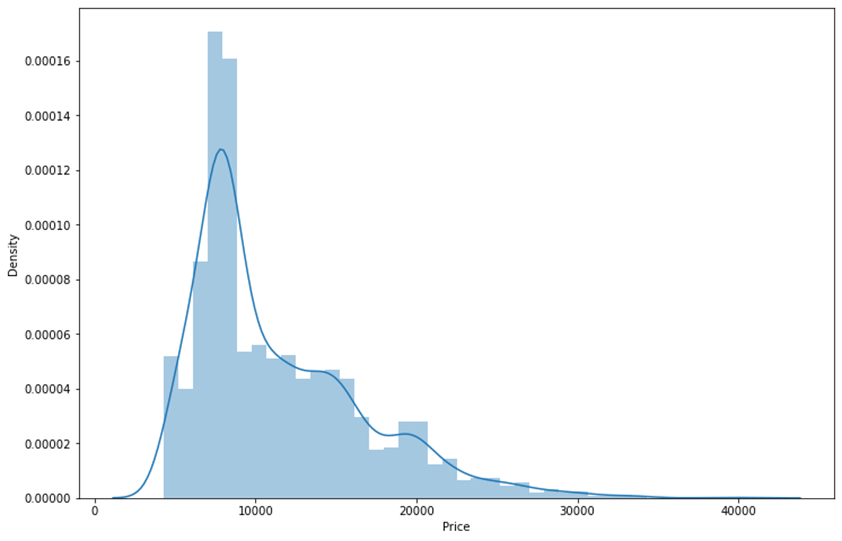
Graphical user interface, text, application, email

Description automatically generated





* Key Metrics for success in solving problem under consideration
* The essential step in any machine learning model is to evaluate the accuracy and determine the metrics error of the model. I have used Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R2 Score metrics for my model evaluation.
* **Visualizations**
* **'Price' Let's replace this sign by empty space and convert the type into float.**



**Airline**

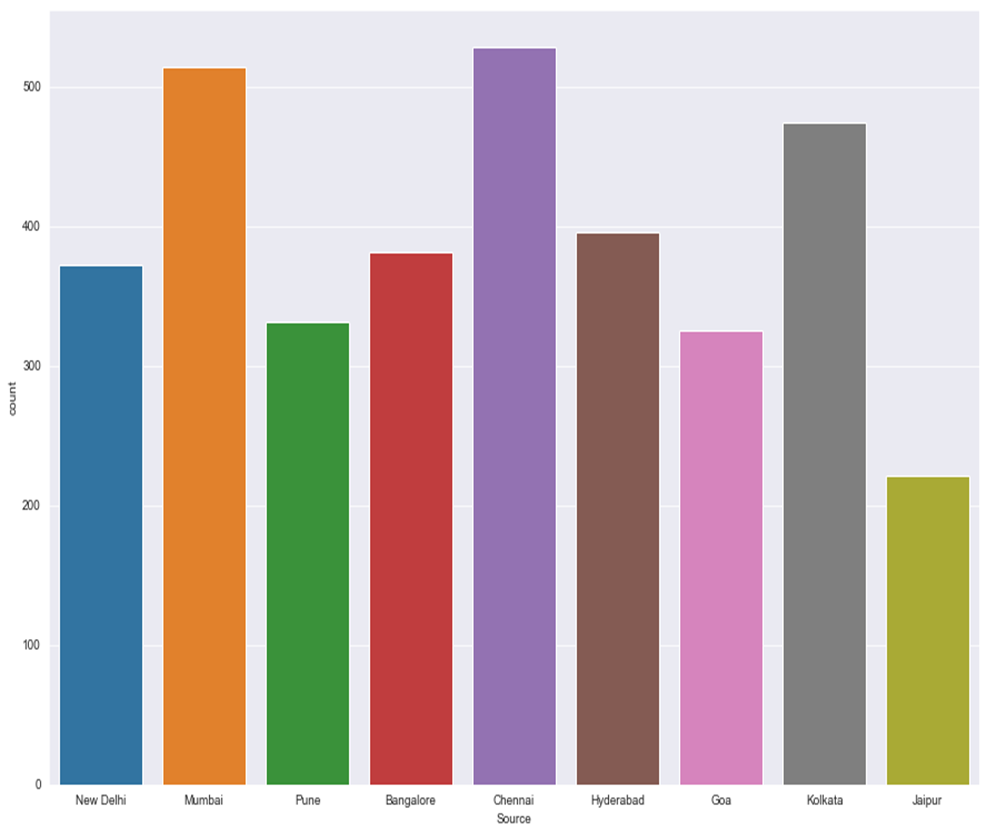
**we can infer that there are a greater number of flights of “Indigo", "Vistara" and " Air India " compared to others. Also, the count of SpiceJet flights are very less.**

**Chart, bar chart

Description automatically generated**

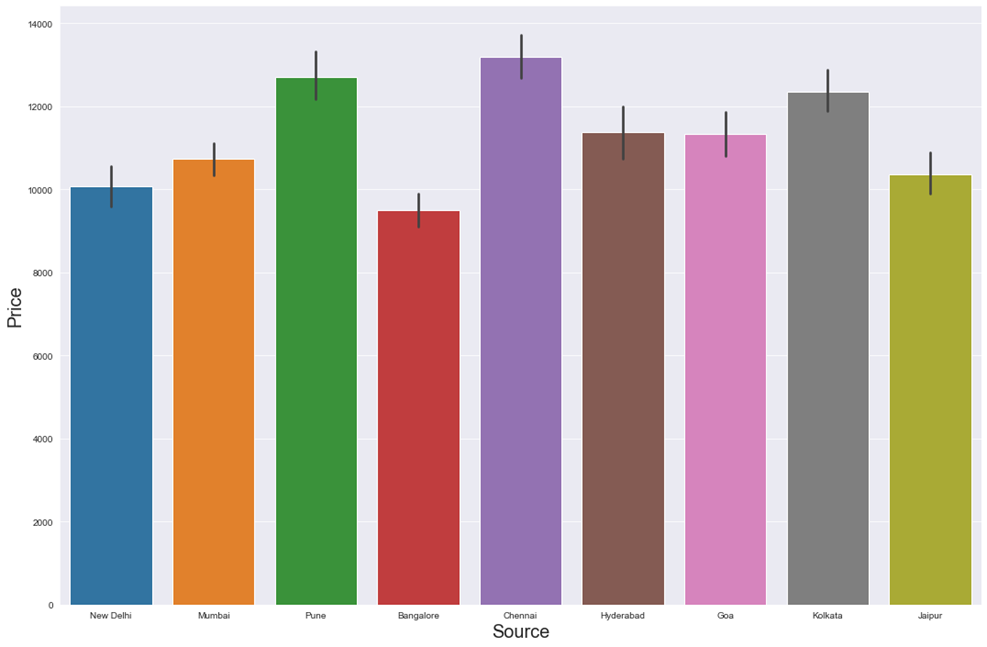
**Source: -**

we can observe a greater number of flights are from Chennai, Mumbai, Kolkata and Hyderabad, Bangalore. Only few flights are from Jaipur.

****

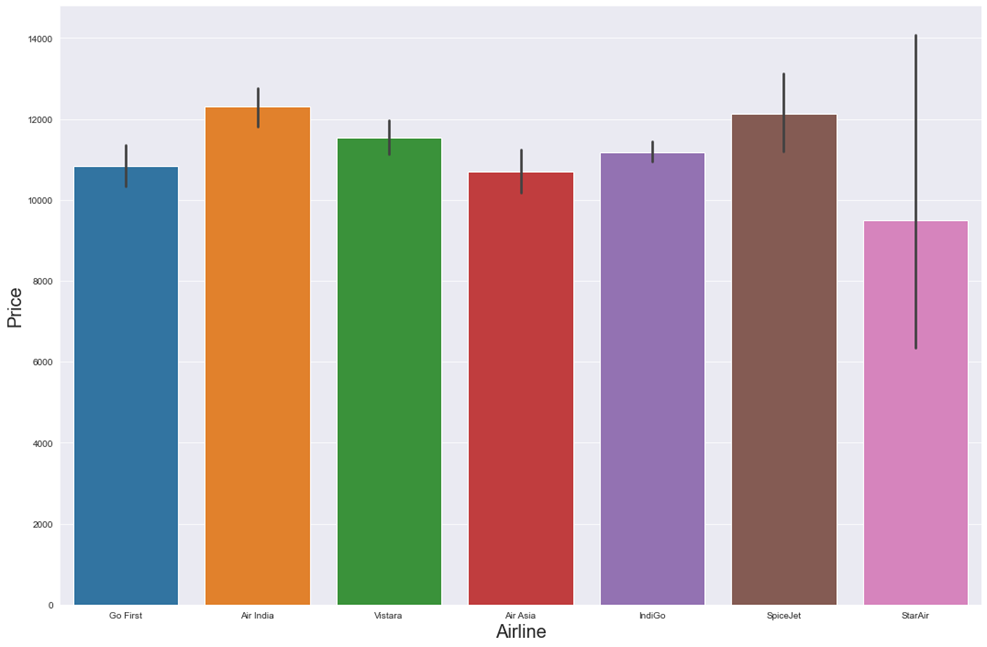
**Source vs Price**

we can observe the flights from Pune and Chennai and Kolkata are having somewhat higher prices compared to other sources**.**

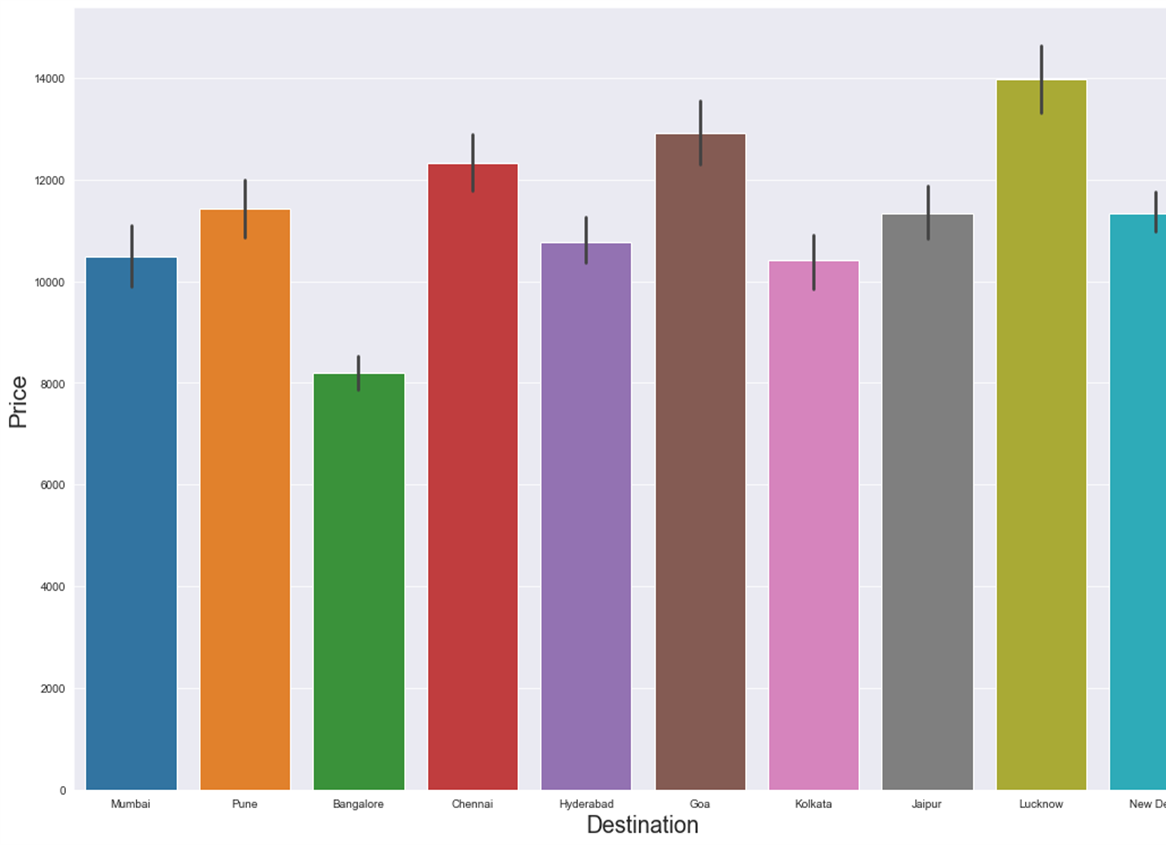
****

**"Airline" vs "Price"**

we can notice “SpiceJet" and "Air India" airlines have highest ticket prices compared to other airlines.

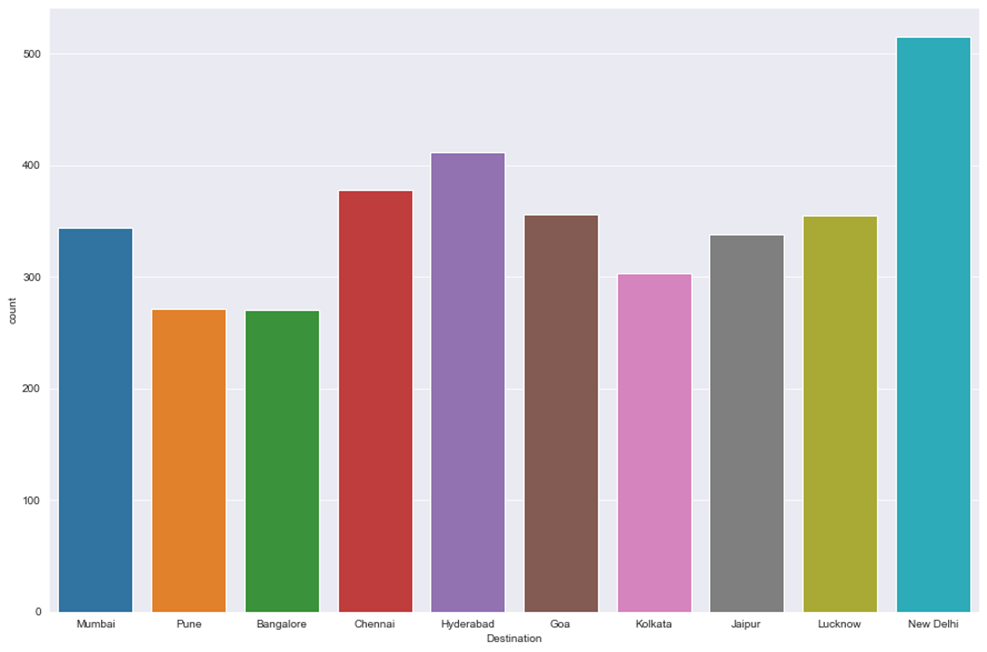


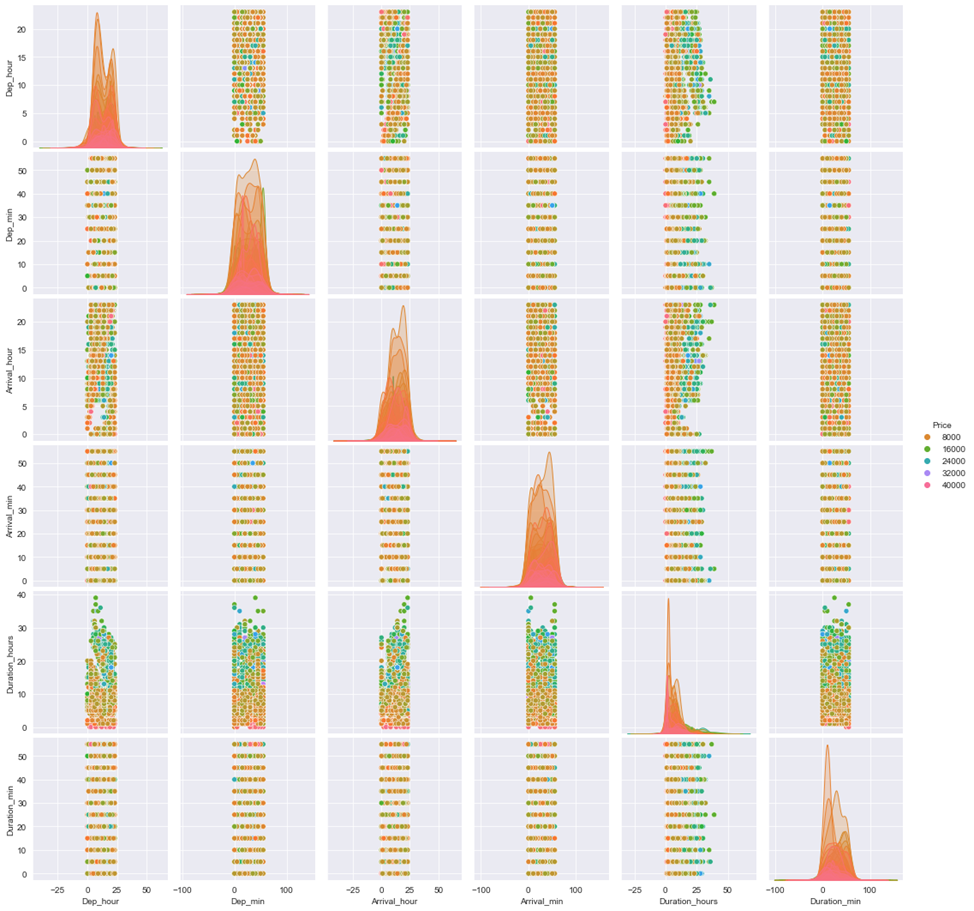
**"Destination" vs "Price"  
we can notice that the flights travelling to Goa have higher flight ticket prices.**

****

**"Destination"**

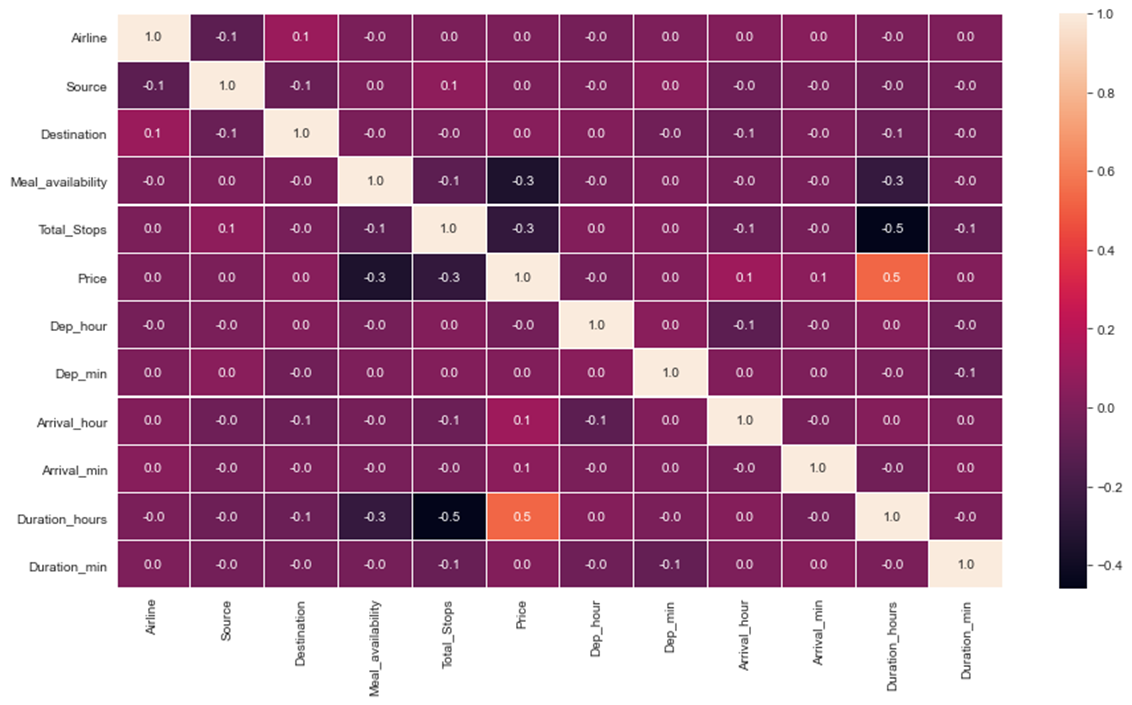
* **More number of flights are heading towards New Delhi, Hyderabad, Chennai and Mumbai. Only few flights are travelling to Hyderabad.**

****

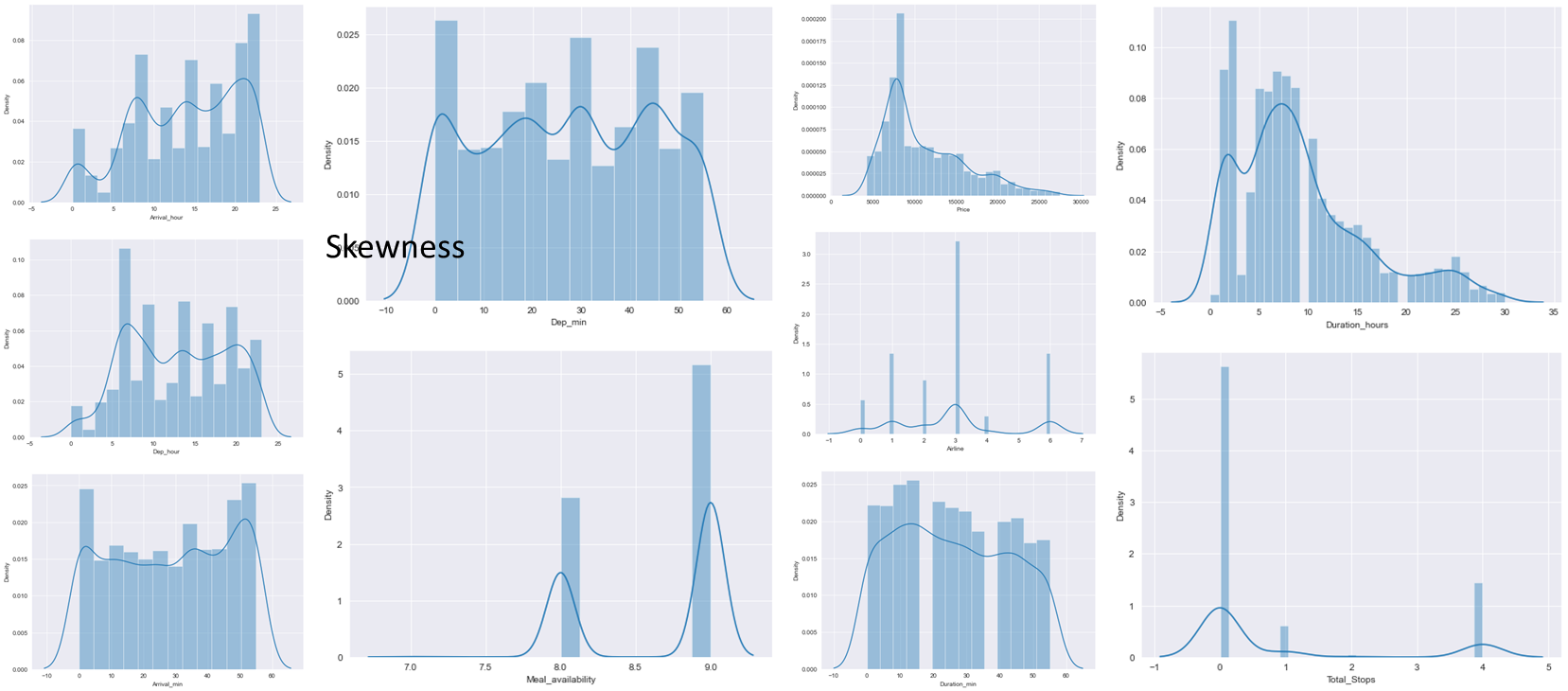
****

**Correlation: -**

**From the heat map, we can clearly observe the positive and negative correlation between the label and features. From the heat map we can notice that the light shades are highly positively correlated and dark shades are highly negatively correlated with the target variable.**

****

**Skewness: -**

****

**CONCLUSION**

* **Key Findings and Conclusions of the Study**
* The case study aims to give an idea of applying Machine Learning algorithms to predict the price of the flight tickets. After the completion of this project, we got an insight of how to collect data, pre-processing the data, analyse the data, cleaning the data and building a model.

1. Do airfares change frequently? Do they move in small increments or in large jumps?

Ans : - Flight ticket prices change during the morning and evening time of the day. From the distribution plots we came to know that the prices of the flight tickets are going up and down, they are not fixed at a time. Also, from this graph we found prices are increasing in large amounts.

**2. Do they tend to go up or down over time?**

Some flights are departing in the early morning 10 AM having most expensive ticket prices compared to early morning flights. As the time goes the flight ticket fares increased and midnight flight fares are very less (say after 10 PM). Also, from categorical and numerical plots we found that the prices are tending to go up as the time is approaching from morning to evening.

**3.What is the best time to buy so that the consumer can save the most by taking the least risk?**

Ans: -From the categorical plots (bar and box) we came to know that early morning and late-night flights are cheaper compared to working hours.

**4. Does price increase as we get near to departure date?**

Ans: - From the categorical plots we found that the flight ticket prices increase as the person get near to departure time. That is last minute flights are very expensive.

**5. Is Indigo cheaper than Jet Airways?**

Ans: - From the bar plot we got to know that both Indigo and SpiceJet airways almost having same ticket fares.

**6. Are morning flights expensive?**

Ans: - Not all flights are expensive during morning, only few flights departing in the early morning 10 AM are expensive. Apart from this the flight ticket fares are less compared to other timing flight fares.

* **Learning Outcomes of the Study in respect of Data Science**
* While working on this project I learned many things about the features of flights and about the flight ticket selling web platforms and got the idea that how the machine learning models have helped to predict the price of flight tickets. I found that the project was quite interesting as the dataset contains several types of data. I used several types of plotting to visualize the relation between target and features. This graphical representation helped me to understand which features are important and how these features describe price of tickets.
* **Limitations of this work and Scope for Future Work**
* The main limitation of this study is the low number of records that have been used.In the dataset our data is not properly distributed in some of the columns many of the values in the columns are having string values which I had taken care. Due to some reasons our models may not make the right patterns and the performance of the model also reduces. So that issues need to be taken care.
* We can build a web spider that extracts the required values from a website like yatra.com, Goibibo etc.
* The analysis can be done by increasing the data points and increasing the historical data used. That will train the model better giving better accuracies and more savings.