
# FLIGHT PRICE PREDICTION

**Tinu Shiby**

# ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my mentor Mr.Keshav Bansal, Flip Robo Technologies for his inspiration and guidance at every stage of the project. The blessing, help and guidance given by him will certainly help me in achieving better things in future.

I wish to express my profound gratitude to the DataTrained family for providing an opportunity to undertake this internship.

Lastly, I thank the almighty and my parents for their every small support and encouragement which helped me to complete this project successfully.

# INTRODUCTION

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) So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

# DATA COLLECTION

You have to scrape at least 1500 rows of data. You can scrape more data as well, it’s up to you, More the data the better the model. In this section you have to scrape the data of flights from different websites (yatra.com, skyscanner.com, official websites of airlines, etc). The number of columns for data doesn’t have a limit, it’s up to you and your creativity. Generally, these columns are airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.

# DATA ANALYSIS

The goal is to find out how price varies according to the variables. To start with the data analysis, i have imported the following:

- Pandas
- Matplotlib
- Seaborn
- Standard Scalar
- Principal Component Analysis
- Train test split
- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- SVR
- XGBoost Regressor
- GridSearchCV
- Cross-validation score
- RMSE score
- R2 score

# EDA

Exploratory data analysis is one of the most crucial steps in data analysis. It provides a better understanding of data set variables and the relationships between them. It can also help in determining if the statistical techniques you are considering for data analysis are appropriate.

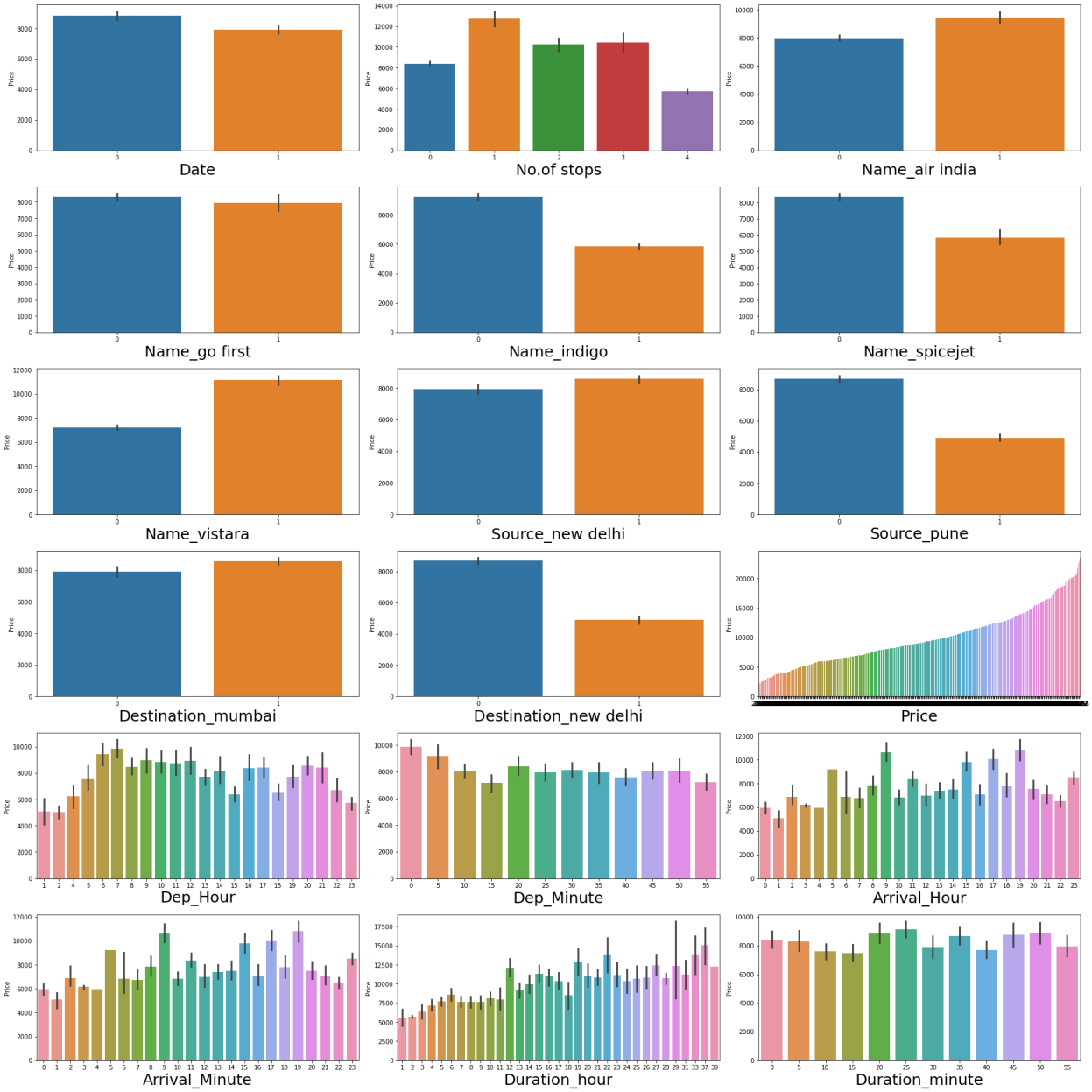
The dataset has 10 variables and 1630 rows, which is splitted into train and test dataset. In the dataset Date has null values and replaces NaN with value.

One Hot Encoding Technique is applied for Name,Source and Destination  for the rest of the variables LabelEncoding technique is applied.

To find out the relationship between feature variables and target variables, I have plotted graphs.

From the data analysis part its clear that the following factors plays an important role in determining the price of flight tickets:

- Date
- Number of Stops
- Airline Name




After bivariate analysis I have checked the correlation and it was found out that Airline Name is having the highest correlation with target variable. Multicollinearity exists in the dataset. In order to check for the outliers a statistical analysis is done using the describe method, and a boxplot is also plotted.

# DATA PREPROCESSING

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

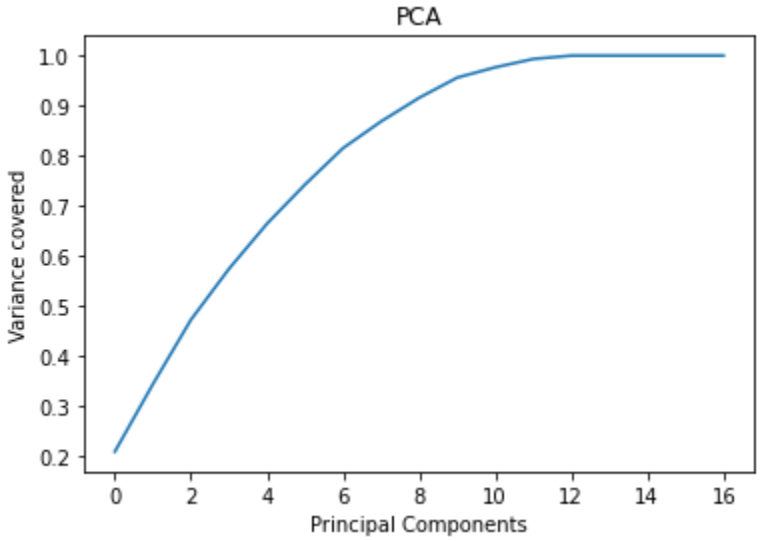
The column Unnamed: 0 was dropped as it is same as the index.

Outliers are removed from the dataset using z-score, data loss was ~5%.

As skewness exists in the dataset, the skewness is removed by applying the power transform method in specific yeo-johnson method. Then the skewness got removed from the dataset.

To standardize the data I have used StandardScalar, the whole idea of using Standard Scalar is that it will transform the data such that its distribution will have a mean value of 0 and standard deviation of 1.

As multicollinearity exists in the dataset, applied PCA and considered 12 variables.




Then find out the best random state and splitted the dataset into train and test.

# BUILDING MACHINE LEARNING MODELS

As the problem is a regression problem I have used the following models:

- Linear Regression
- Decision tree Regressor
- Random forest Regressor
- SVR
- XGBoost Regressor

.Cross-validation scores of Linear Regression, Decision Tree, Random Forest, Gradient descent boosting and XGBoost were found.
Since the difference is least for Random Forest Regressor, that is taken as the best model and applied Hyper parameter tuning using GridsearchCV. The best parameter values are as follows:
Max_depth : 10
Min_samples_leaf : 2
Min_samples_split:12
Random state : 1
After hyperparameter tuning the score increases to 51.68%

# CONCLUSION

In this project, I have tried to uncover the underlying trends in flight prices. Nowadays the flight price varies significantly even for the same flight. So customers are seeking to get flight prices at the lowest rate and there comes the scope of this project. For this project I have collected the data from the site Yatra.com and did the feature engineering, exploratory data analysis and data preprocessing steps. Tried with different models and chose Random Forest Regressor as the best model. Random Forest Regressor is working with an accuracy of 51.67%.