

Flight Price Prediction

**Submitted By:**

**Junaid Shaikh**

## ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot.

Most of the concepts used to predict the price of used cars are learned from Data Trained Institute and below documentations.

Some of the reference sources are as follows:

* Medium.com
* StackOverflow

Contents

1. Introduction
   1. Business Problem Framing:
   2. Conceptual Background of the Domain Problem
   3. Review of Literature
   4. Motivation for the Problem Undertaken
2. Analytical Problem Framing
   1. Mathematical/ Analytical Modeling of the Problem
   2. Data Sources and their formats
   3. Data Preprocessing Done
   4. Data Inputs-Logic-Output Relationships
   5. Hardware and Software Requirements and Tools Used
3. Data Analysis and Visualization
   1. Identification of possible problem-solving approaches (methods)
   2. Testing of Identified Approaches (Algorithms)
   3. Key Metrics for success in solving problem under consideration
   4. Visualization
   5. Run and Evaluate selected models
   6. Interpretation of the Results
4. Conclusion
5. Key Findings and Conclusions of the Study
6. Learning Outcomes of the Study in respect of Data Science
7. Limitations of this work and Scope for Future Work

INTRODUCTION

BUSINESS PROBLEM FRAMING

Flight ticket prices can be something hard to guess and we know how unexpectedly they 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 Time of purchase patterns. The last-minute purchases are expensive. Raising prices on a flight will reduce sales.

Airplane tickets can be ludicrously expensive. Especially when traveling shorter distances, the difference in cost for air travel and other modes of transport, would seem comically large. And the time saving aspect decreases for short distances when you consider the time it takes to cut through all the hassle and red tape at the airport.

Flight booking systems are dynamic in nature. They depend on a lot of features like Airline company, Source, Destination, duration, arrival time, departure time, number of stops and date of the flight. In this project, I plan to use machine learning algorithms on a dataset based on the above parameters to predict flight prices. There are basically two approaches to solve this problem. These involve considering it as a regression or classification problem. Algorithms can be applied to predict whether the price of the ticket will drop in the future, thus considering it as a classification problem. In this project, I will consider it as a regression problem, thus predicting the ticket price.

Conceptual Background OF The Domain Problem

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, we have to work on a project where we will collect data of flight fares with other features

and work to make a model to predict fares of flights.

Review OF Literature

In this model we will study different variables and how these independent variables are related with dependent variables and how this will help us to predict whether the price of the flights.

Motivation OF The Problem Undertaken

Air travel is the fastest method of transport around, and can cut hours or days off of a trip. But we know how unexpectedly the prices vary. So, I was interested in Flight Fares Prediction listings to help individuals and find the right fares based on their needs. And also, to get hands on experience and to know that how the data scientist approaches and work in an industry end to end.

Analytical Problem Framing

Mathematical/ Analytical Modeling OF The Problem

For the given flight price prediction project, I have scraped the flight prices along with some

other features from a well-known website that is ‘yatra.com’. And this data is framed into a

data frame and saved into a .csv file. After that I have fetched this data set and performed

some data processing and some EDA. The data set is having around 32377 rows and 10

columns

Data Sources & Data Formats

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 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 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.

**Features Information:**

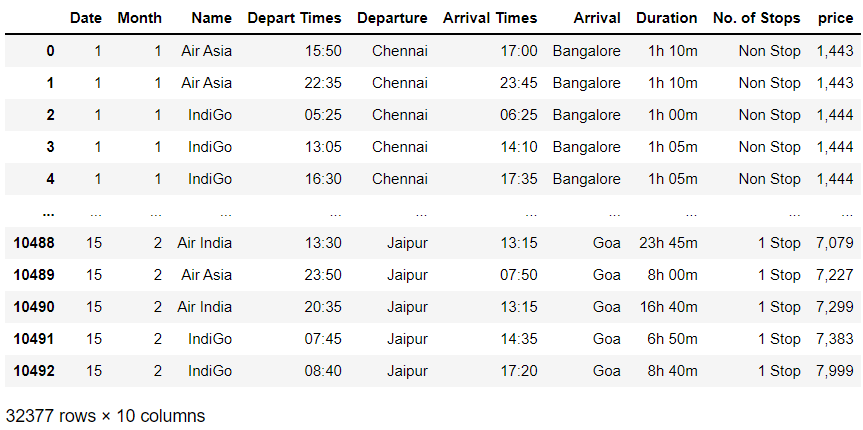
|  |  |
| --- | --- |
| Date | Date of departure |
| Month | Month of Departure |
| name | Company name of the flight |
| Depart times | Time of departure |
| Departure | City of Departure |
| Arrival times | Time of arrival |
| Arrival | Destination of city |
| Duration | Total Air travel |
| No. of stops | Number of stops to reach a destination |
| Price | Price of flight |

After scrapping, the data was provided in csv (comma separated values) format.

The given dataset has 32377 rows and 10 rows. There are no null values in the dataset.

Dataset was imported using Panda’s library and then transformed into data-frame.

Dataset



Data Pre-processing

Current dataset is raw data. By proper Data Transformation methods, a lot of valuable insights can be gained.

Then statistical analysis was done by checking shape, value counts, info etc.….

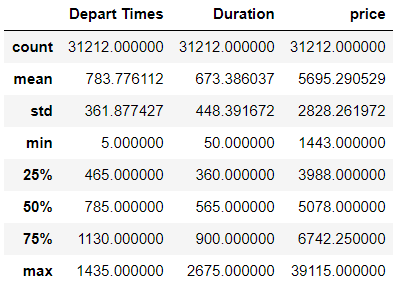
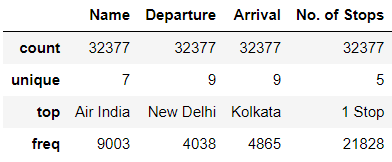
Then while looking into the value counts, I found some columns having datapoints with less than 5% value points. I dropped them.

While checking for null values I found no null values in the dataset.

I have also dropped arrival times column to avoid multi collinearity issue.

Next as a part of feature extraction I converted format of time from hour: min to mins only. Transformed data from date and month into days left for departure. Thinking that this data will help us more. Converted categorical data into numerical data also.

In this project we have performed various mathematical and statistical analysis such as description or statistical summary of the data using describe, checked correlation using corr and visualized it using heatmap. Then we have used Z-Score to plot outliers and remove them.



Data Inputs-Logic-Output Relationship

Since I had numerical and non-numerical columns, I have plotted dist. plot to see the distribution of each column data.

I have used box plot for each pair of categorical features that shows the relation between target and independent features.

Hardware and Software Requirements and Tools Used

Hardware required:

* Processor — core i5 and above
* RAM — 8 GB or above
* SSD — 250GB or above

Software/s required: Anaconda

LIBRARIES:

The tools, libraries, and packages we used for accomplishing this project are pandas, NumPy, matplotlib, seaborn, SciPy, sklearns’s, mlxtend, xgboost, joblib.

Through panda’s library we loaded our csv file ‘Data file’ into data frame and performed data manipulation and analysis.

With the help of NumPy we worked with arrays.

With the help of matplotlib and seaborn we did plot various graphs and figures and done data visualization.

Train\_test\_split is a function in Sklearns’s model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearns’s Train\_test\_split will make random partitions for the two subsets.

With sklearns’s StandardScaler package we scaled all the feature variables onto single scale. As these columns are different in scale, they are standardized to have common scale while building machine learning model. This is useful when you want to compare data that correspond to different units.

With sklearns’s package we imported many regression models, we could obtain cross\_val\_score, which is an accuracy metric used to evaluate model, we could obtain best parameters of a model using GridsearchCV or RandomizedSearchCV, we could reduce skewness using power transform library of sklearns’s.

Model/s Development andEvaluation

Identification of possible problem-solving approaches

For skewness removal I have used power transform method, for scaling down I have used standard scaling.

Graphical user interface, text, application, chat or text message, email

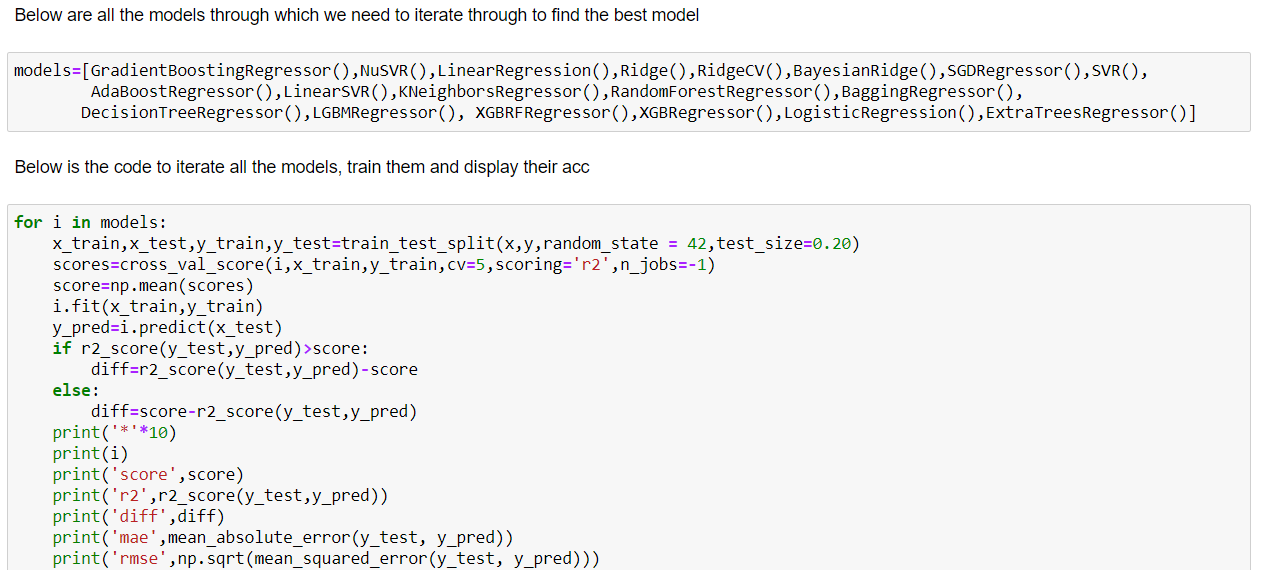
Description automatically generated

Apart from this multicollinearity refers to the collinearity between the features. Multicollinearity occurs when our model includes multiple factors that are correlated with each other’s other than with label. It makes more difficult for the model predict and affects the accuracy. They are treated using VIF method. This algorithm reduces the no. of columns by removing highly correlated feature columns.

Testing of Identified Approaches (Algorithms)

Since price was my target variable and it was a numerical column, so this problem was regression problem. And I have used all regression algorithms to build my model. By looking into the difference of accuracy score and cross validation score I found XGBRegressor and RandomForestRegressor as a best models with least difference. Also, to get the best model we must run through multiple models and to avoid the confusion of overfitting we have go through cross validation.

Run & evaluate selected models



Key Metrics for success in solving problem under consideration

Following metrics were used to evaluate our model:

--- Cross Val Score

--- R2 Score

--- STD error

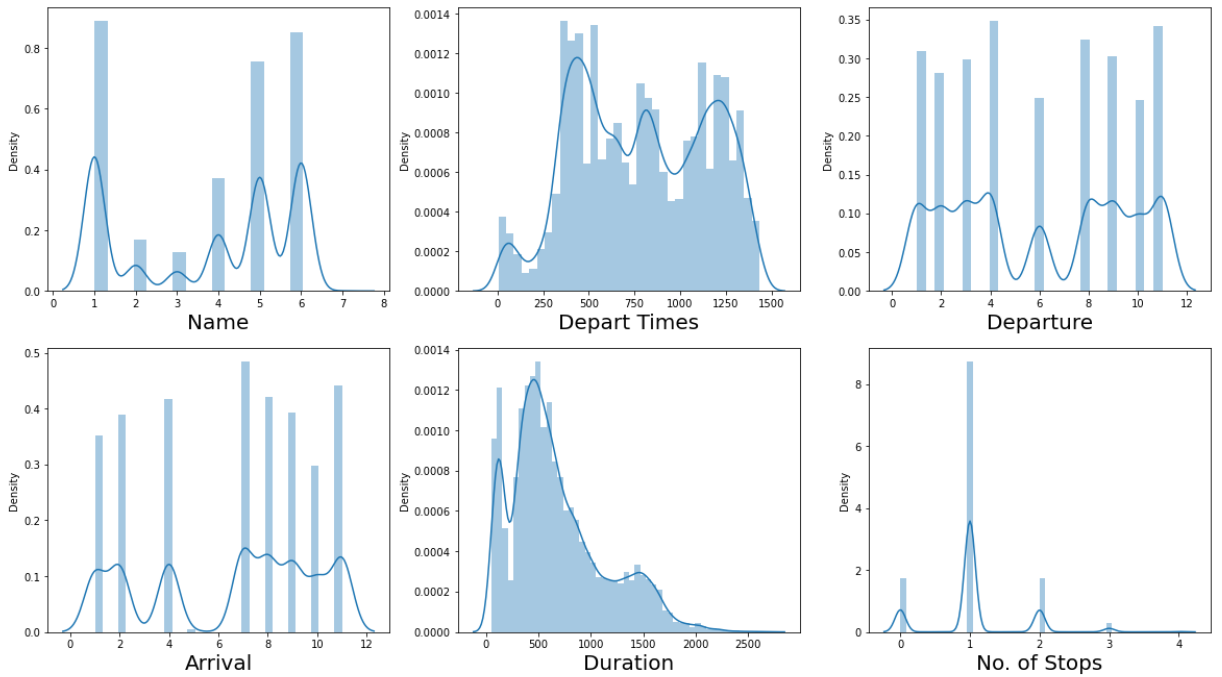
--- mean absolute error

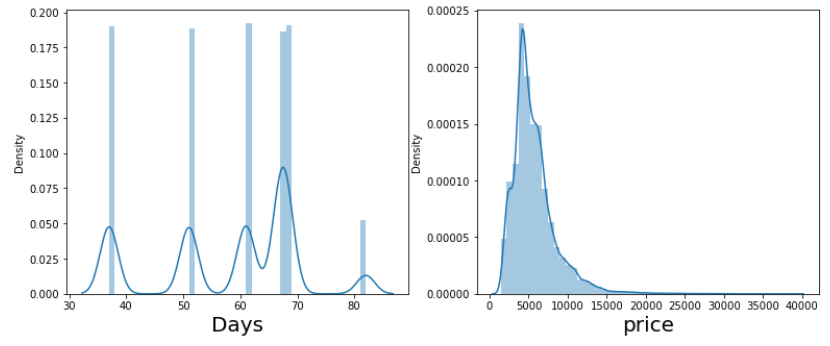
--- square root of mean absolute error

Visualization

**Univariate Analysis:**

### Distribution plot of all columns

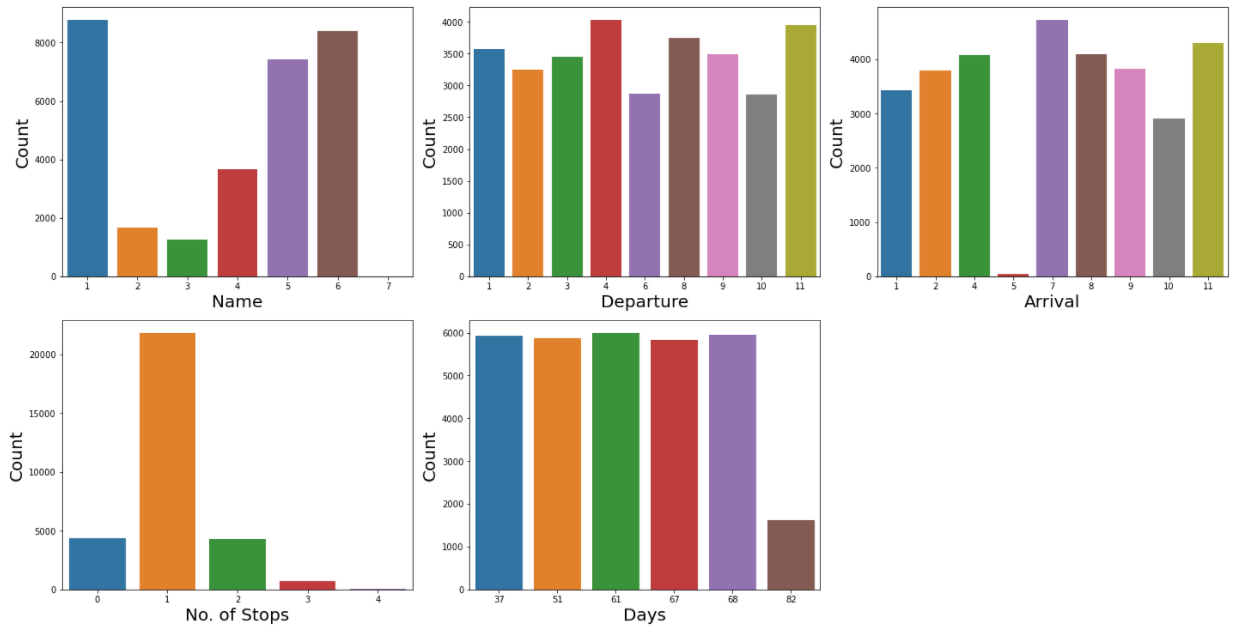




Observations:

* Duration & price columns has normally distributed value points but contain skewness in them.
* Except for Depart times remaining column are categorical columns having bimodal distributed graph.
* Depart times column has rectangular distributed data values which is due to the 24 hr flight range.

Count plot of all categorical column

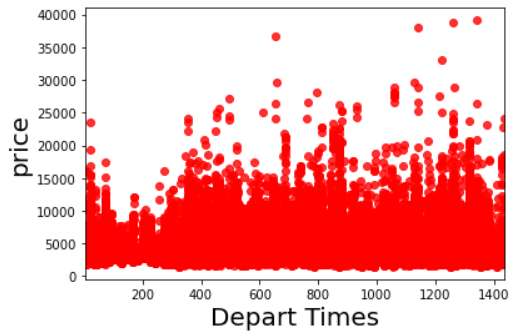
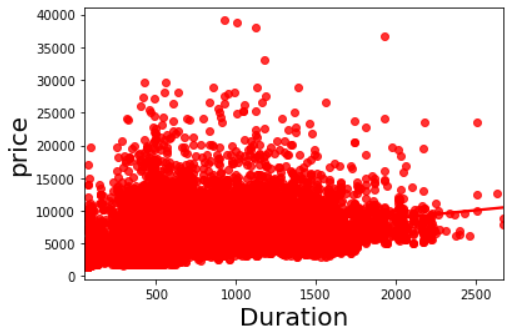


### Observation:

### From the above graph we can observe the value\_counts of datapoints of resp column

### we can make a note about the data i.e., dataset contains data mostly of which airplane, no. of stops.........

### **Bivariate analysis for numerical columns**

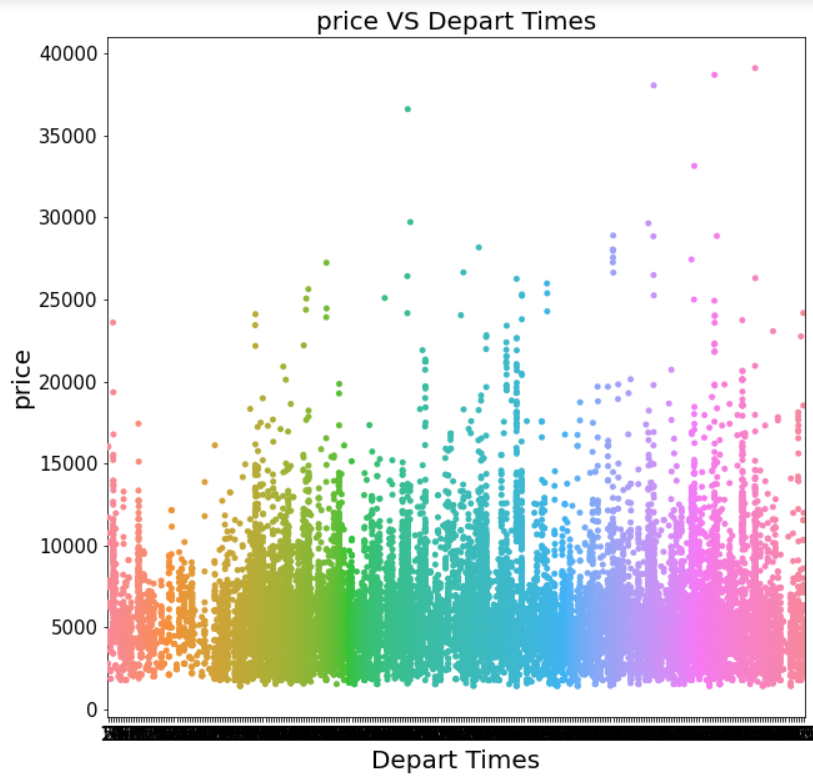


Observation:

### 

### Both columns have positive correlation w.r.t target variable.

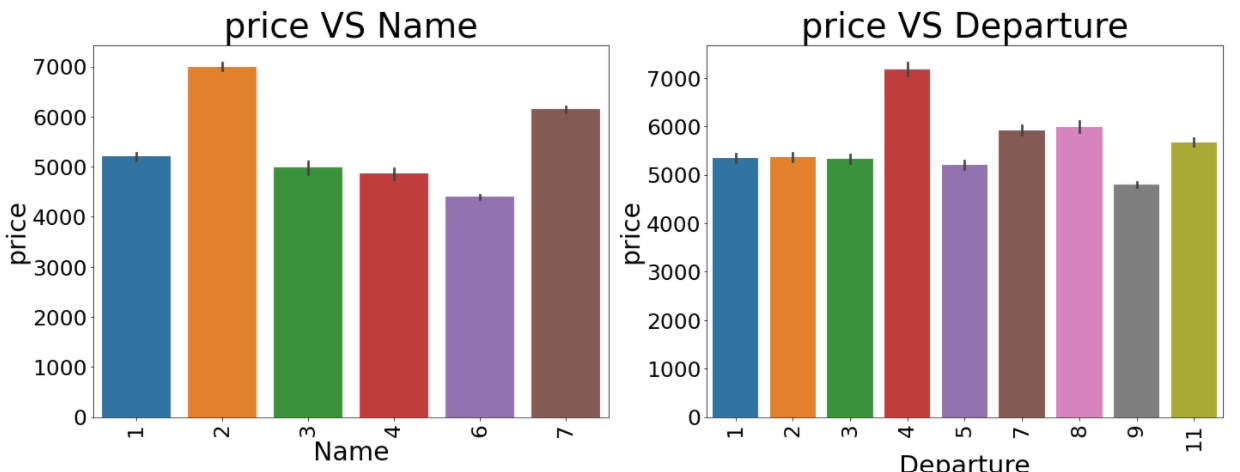
Strip Plot of each column values w.r.t label column

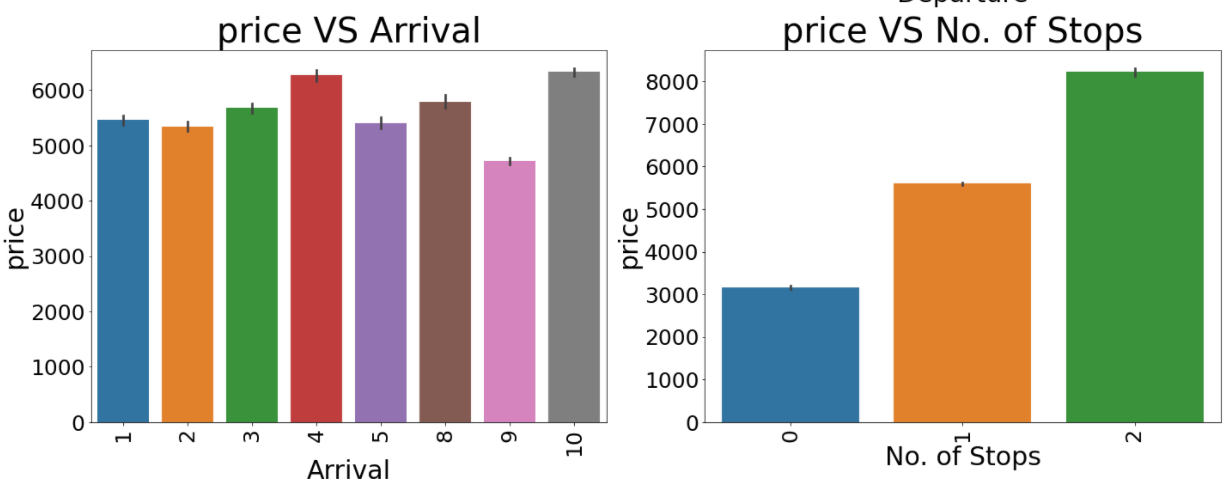


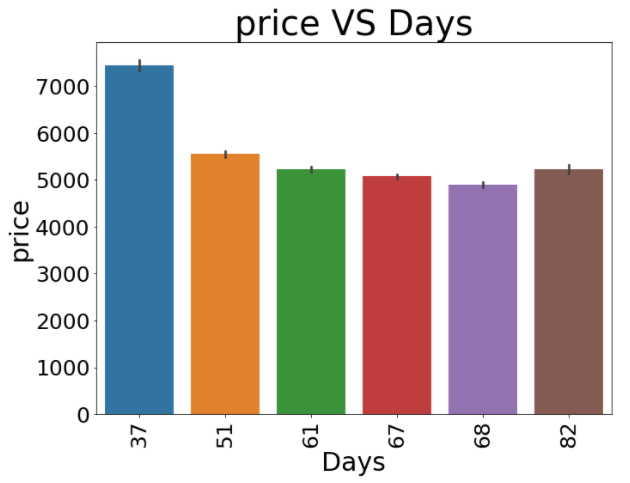
Observation:

### Prices of flight defer w.r.t the data point of both columns in positive correlated way.

Bar graph for categorical column





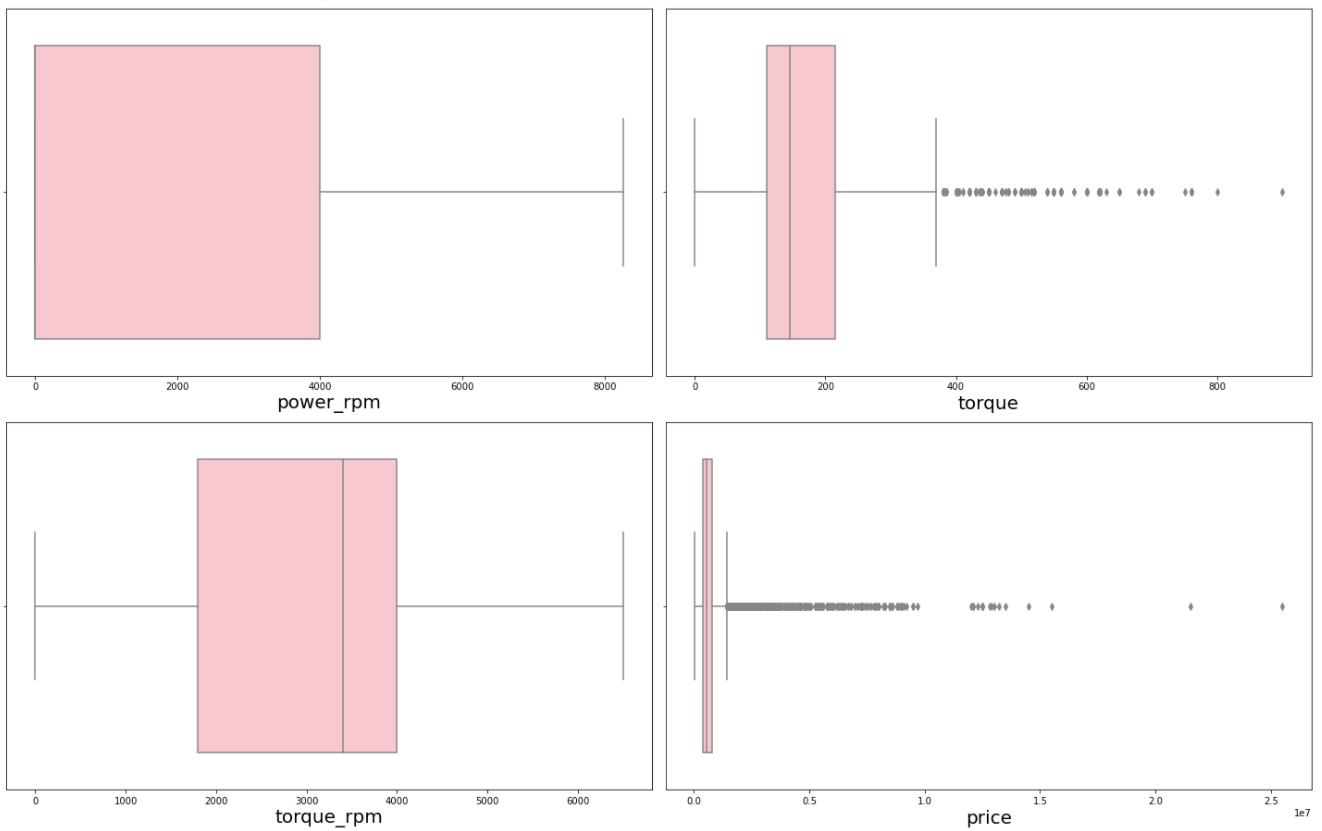
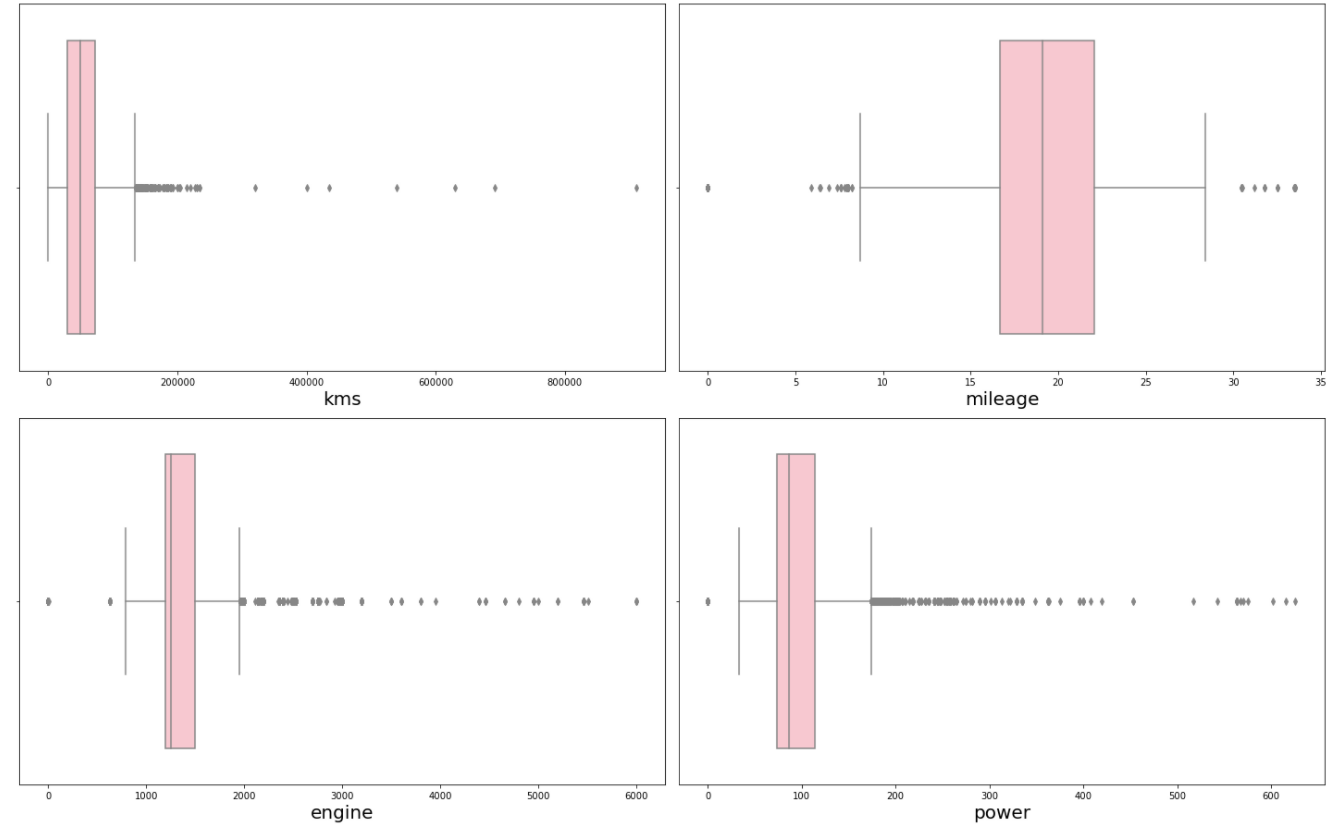


Observation:

* From above graph we can analyse which data point has higher average price value of respective column, to understand the impact of data value on target variable.

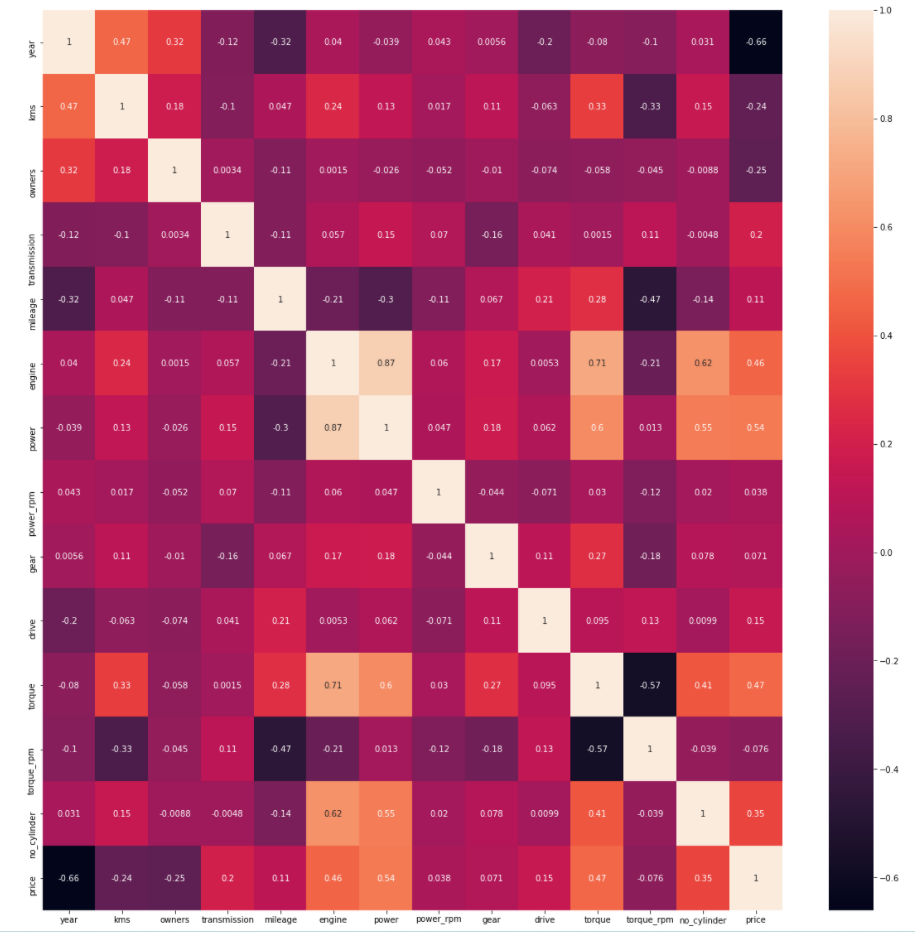
Check For Outliers

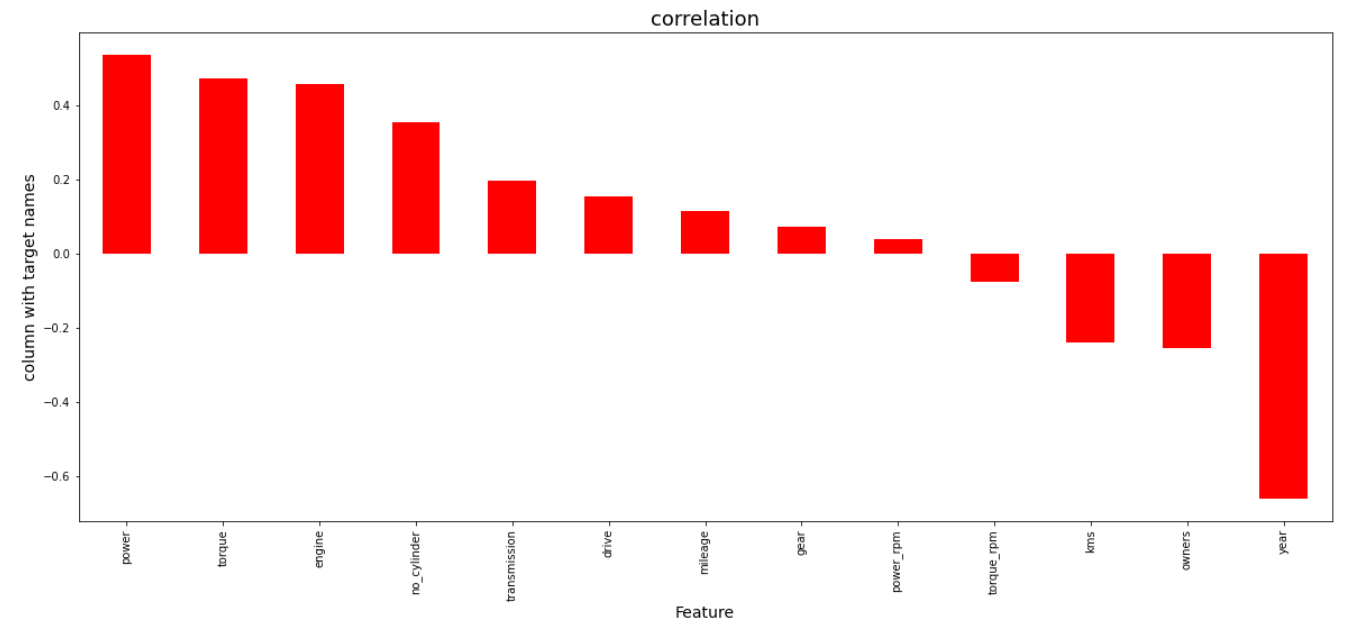
Visualize Boxplot of every column having datapoints of continuous nature



### Outliers are present in column kms, mileage, engine, power, torque & price, removed them using z-score & percentile method

Check For Correlation

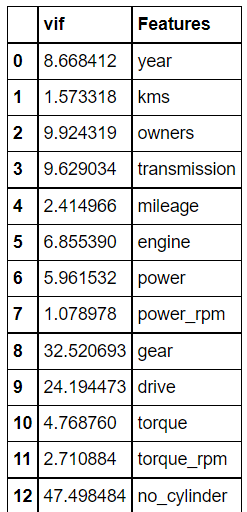


****

**Observation:**

### We can analyze that column gear, power\_rpm & torque\_rpm have least correlation, but we will not drop them for now and analyze further

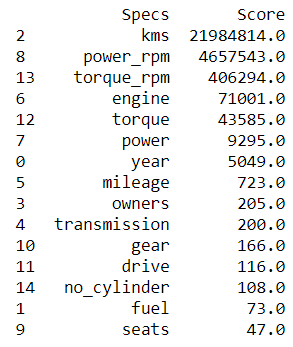
**VIF to detect multicollinearity and remove them**

****

**Observation:**

**Drop column no\_cylinder, drive, gear**

**Find the best feature column:**

****

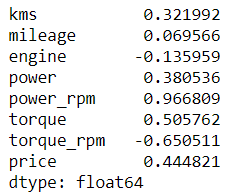
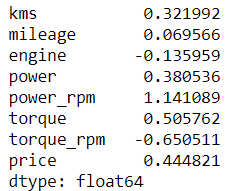
**Drop column seats & fuel as they have least significance regarding target variable.**

Model Building

**Perform Feature Scaling**

**Before we start model building, we need to perform feature scaling on all columns, to avoid biasing of data.**

**Also check for skewness in data and remove it.**



**Model Building**

**As we know, this is a regression problem we need to build a model using regression algorithm models.**

**First, we need to write a function which can find us best random state for train test split.**

**Then we shall iterate through all the models supporting regression algorithms to find the best models.**

From above we get to know that the top 3 models are:

* LGMBRegressor
* XGBRegressor
* RandomForest Regressor

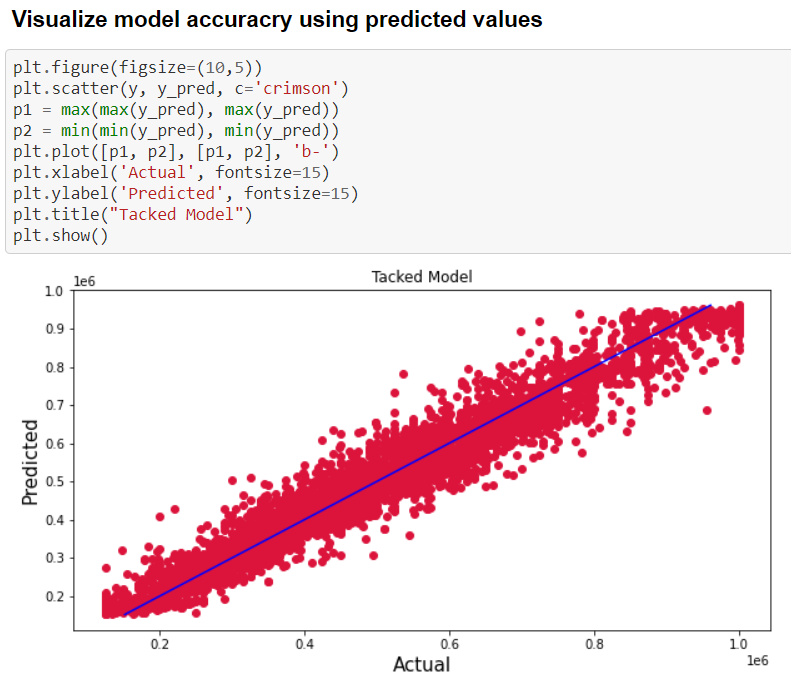
Fine tune all these models and find their best parameters to use.

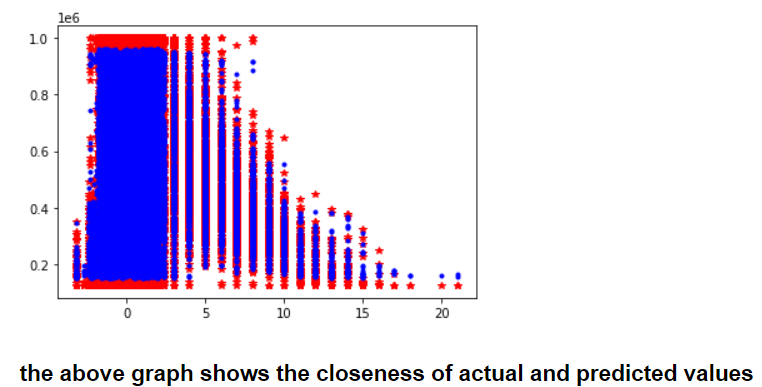
Next, find the best random state for train test split.

we obtain test accuracy of more than 88%.

CV score of this model is more than 96%.

To analyze our model, we shall find the difference between actual and predicted value.





Interpretation of the Results

The dataset was very challenging to handle it had 18 features.

Firstly, the datasets were having any null values.

But there was huge number of similar entries in columns, so we must be careful while going through the statistical analysis of the datasets.

And proper plotting for proper type of features will help us to get better insight on the data. I found maximum numerical columns in the dataset, so I have chosen bar plot to see the relation between target and features.

I notice a huge amount of outliers and skewness in the data so we have choose proper methods to deal with the outliers and skewness. If we ignore these outliers and skewness, we may end up with a bad model which has less accuracy.

Then scaling dataset has a good impact like it will help the model not to get biased. Since we have not removed outliers and skewness completely from the dataset, so we must choose Normalization.

We must use multiple models while building model using dataset as to get the best model out of it.

I Stacked all the top models and concluded it as the best model with 96% accuracy score. Also, I have improved the accuracy of the best model by running hyper parameter tunning.

At last, I have predicted whether the price of used car using saved model. It was good!! that I was able to get the predictions near to actual values.

Conclusion

Key Features and conclusion of the study

In this project we have tried to predict the price of used car. The best accuracy score was achieved by fine-tuned the stacked model built by combining all three top models.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

Through different powerful tools of visualization, we were able to analyze and interpret different hidden insights about the data.

Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

The data was improper scaled, so we scaled it to a single scale using sklearns’s package StandardScaler.

The columns were skewed due to presence of outliers which we handled through percentile technique.

Model was then built having accuracy more than 96% using train dataset.

LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

Due to the presence of lot of outliers, we are unsure whether the model is going to perform well to a completely new dataset.

The scope for future work is to collect as many data as we can so that the model can be built more efficiently.