

FLIGHT PRICE PREDICTION

Submitted by:

MAHESH AGASHIMANI

**ACKNOWLEDGMENT**

I take great pleasure to thank and acknowledgement the allowance by Data Trained Education and permission by Flip Robo. I extend whole hearted thanks to them I worked and learned a lot and sharing me the knowledge and experience.

Data Trained Education and Flip Robo provided training is the very important to completion of project.

**INTRODUCTION**

* Business Problem Framing

The tourism industry is changing fast and this is attracting a lot more travelers 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**.**

* Conceptual Background of the Domain Problem

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.

* Review of Literature

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).

* Motivation for the Problem Undertaken

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

This dataset contains information about Flight listed on https://www.makemytrip.com. This data can be used for a lot of purposes such as price prediction to exemplify the use of Random Forest regression in Machine Learning.

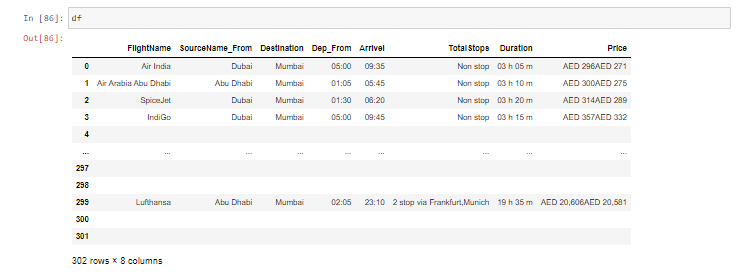
* Data Sources and their formats

The columns in the given dataset is as follows:..

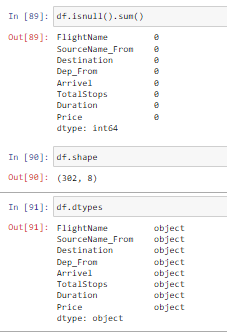
Importing the Libraries:



. Loading the Dataset:



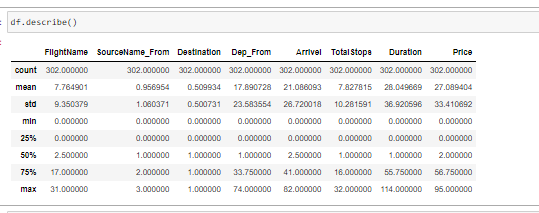
After that we are checking if Null values are present in datasets.



In checked above data sets, there is no Null values present.

* Data Inputs- Logic- Output Relationships

Using describe method its showing the value of columns like count, mean, std, min, 25%, 50%, 75% max value.



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

As given in datasets my assumption is predicting Price and this problem is Regression Problem.

* Hardware and Software Requirements and Tools Used

The needed time to train the model depends on the capability of the used system during the experiment. Some libraries use GPU resources over the CPU to take a shorter time to train a model.

|  |  |
| --- | --- |
| Operating System | Windows 10 |
| Processor | Core i7 |
| RAM | 16GB |
| Graphics card | 1080 TI OC |

Also we are using Jupiter notebook for running the code

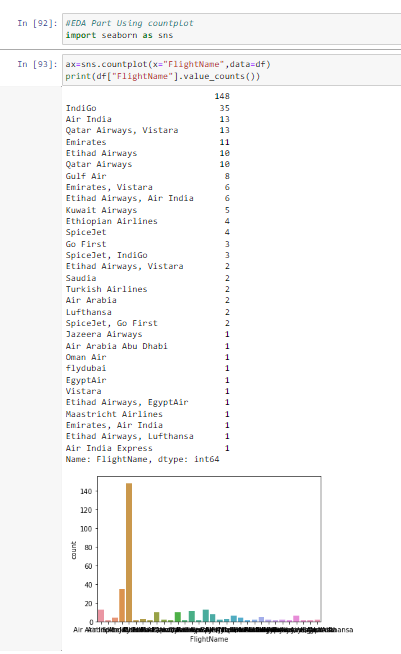
.

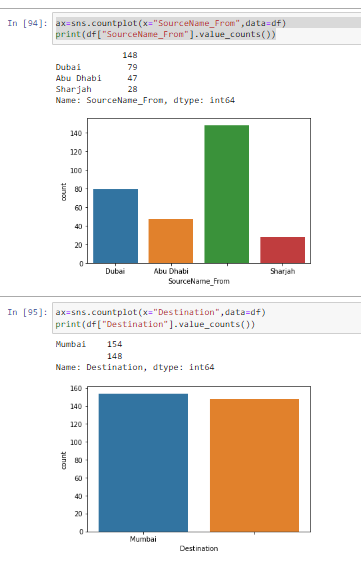
**Model/s Development and Evaluation**

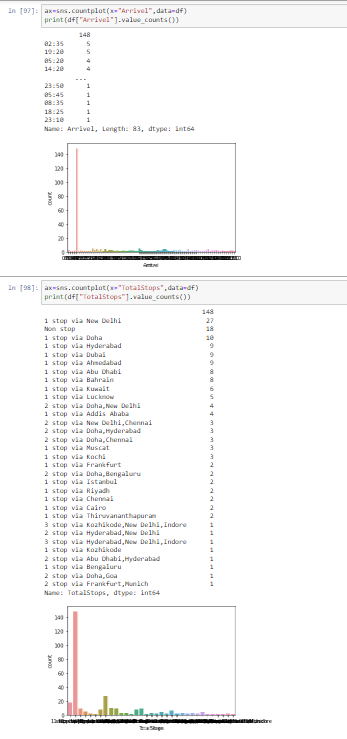
* Identification of possible problem-solving approaches (methods)

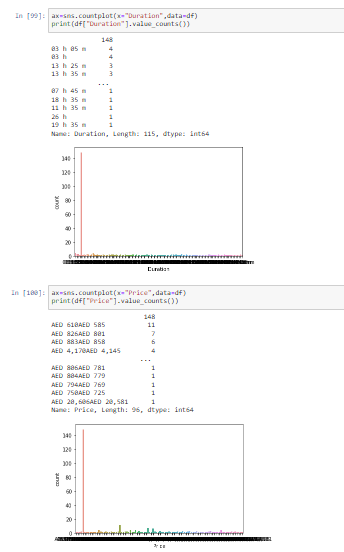
After describing the data, we are checking the EDA part for Categorical & Numerical data.

EDA part for Categorical data using Count Plot:.

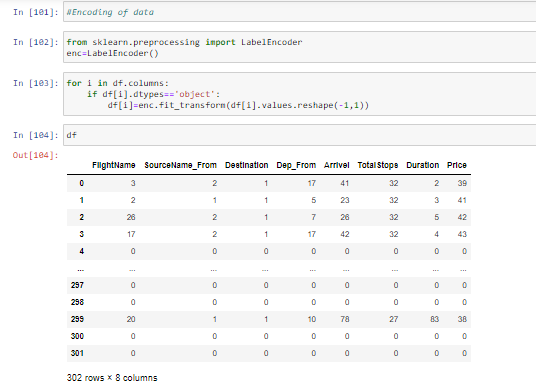








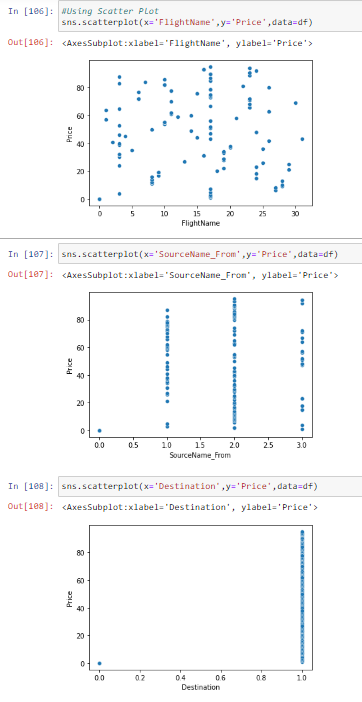
After this we are going to checking dtypes & encoding the data in categorical to numerical data. Machine learning technique is processing only numerical data.

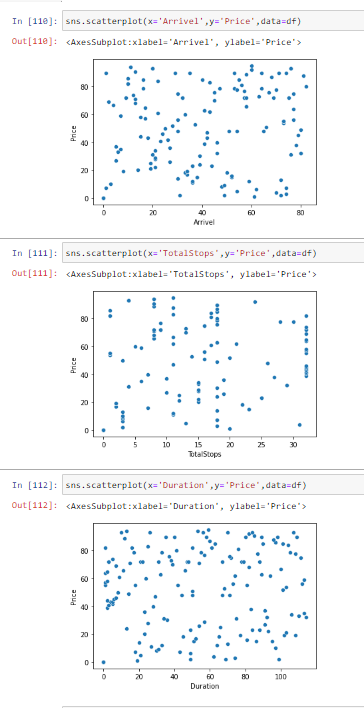


As above image we encoded categorical data into numerical data.

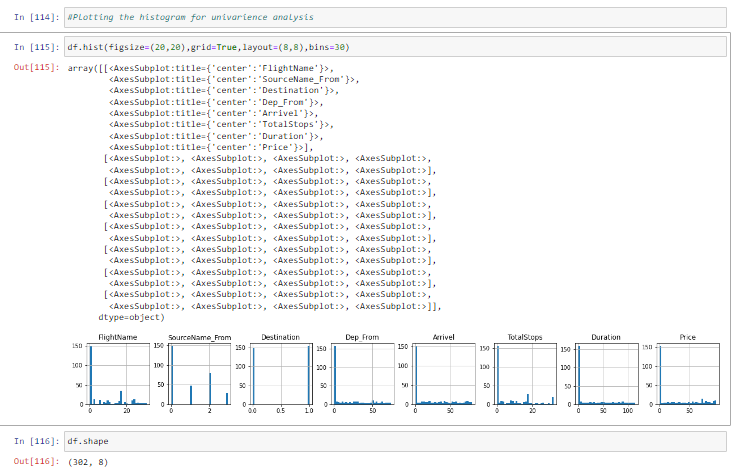
Visualizing data using Scatterplot: -

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs we will understand the data.

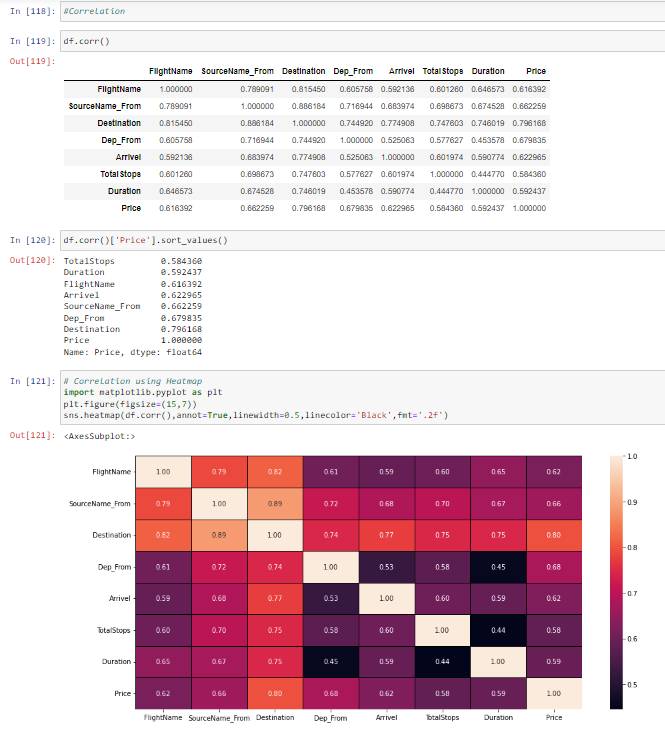






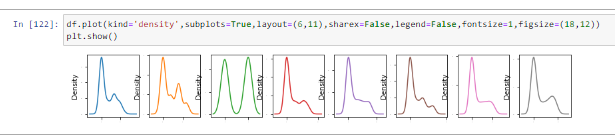


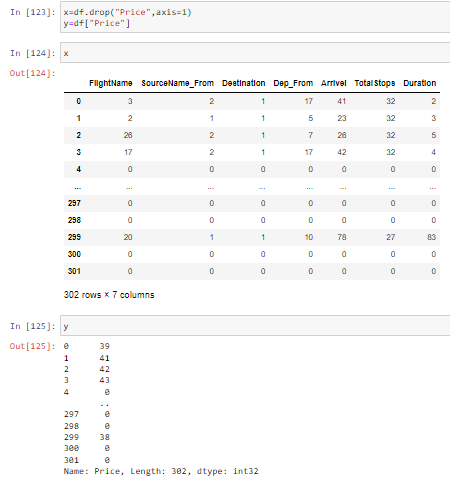
After this we are going to check Correlation of the data.



Observation:

* FlightName and Price features are highly correlated with each otheer.
* SourceName\_From and Price features are highly correlated with each other.
* Destinatiojn and Price columns are highly correlated with each other.
* Dep\_From is also highly correlated with Price column.
* Arrivel and Price is highly correlated with each other.
* TotalStops and Price is highly correlated with each other.
* Duration and Price is highly correlated with each other





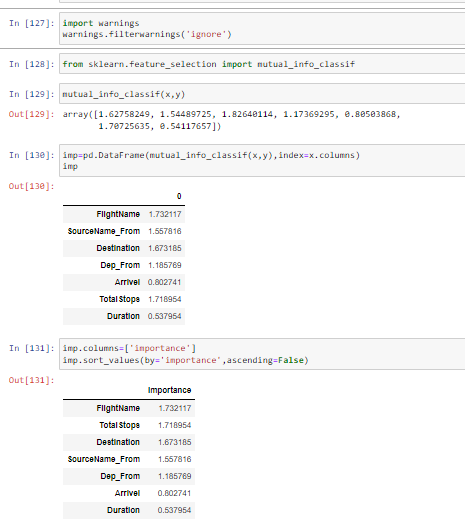
No need to check skewness & outliers for Categorical data.

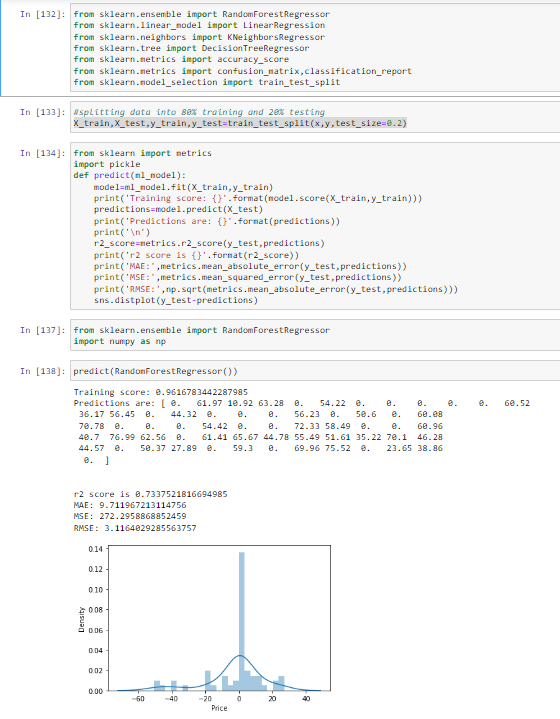
It’s an invalid operation.

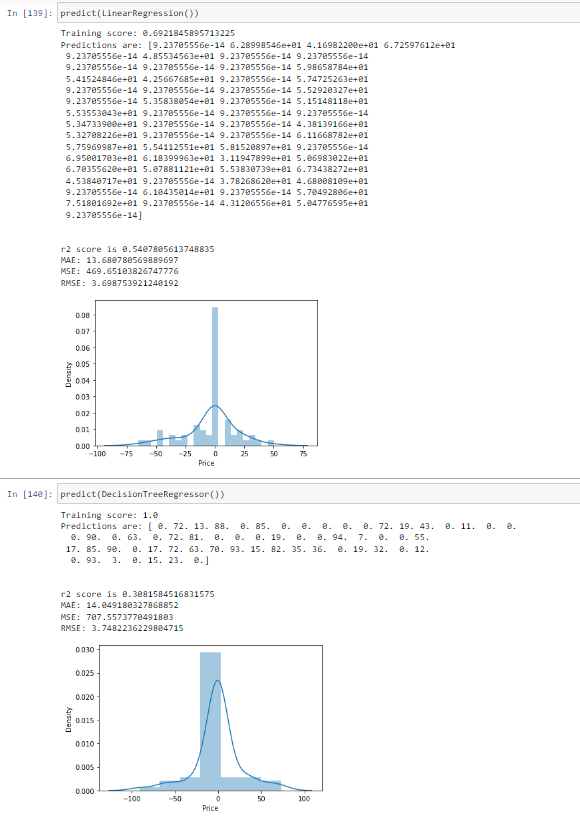
* Testing of Identified Approaches (Algorithms)

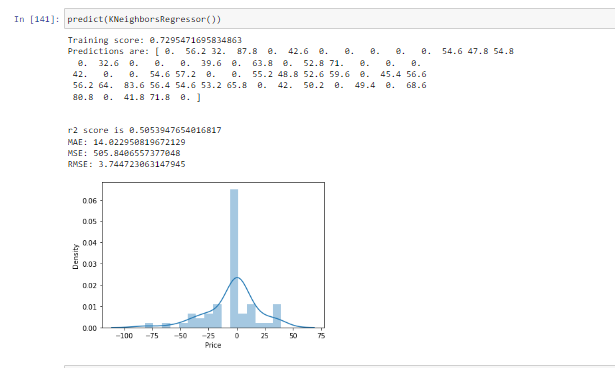
For Training & Testing data using RandomForestRegressor model.

Importing necessary libraries.









* Run and Evaluate selected models

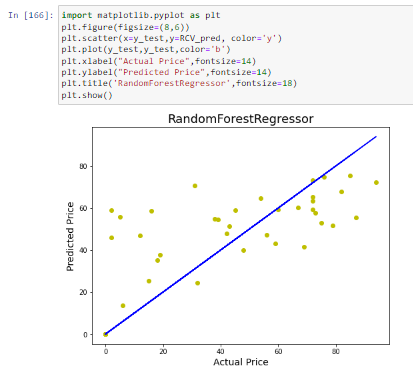


As above checked algorithm Random Forest Regressor accuracy is good.

We have achieved approximately 69% accuracy but before when we used RandomForest the accuracy was around 73%, that’s the outcome of model hyperparameter tuning.

* Visualizations

AUC ROC Plot:



* Interpretation of the Results

After Building a model need to Regularization technique for model checking.



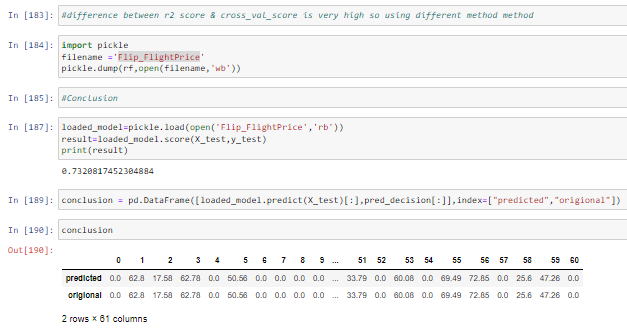
After Regularization a model need to Ensemble technique for model checking.



**CONCLUSION**

* Key Findings and Conclusions of the Study

Using different models for Flight Price prediction. Four different types of Machine Learning methods including RandomForestRegressor, LinearRegression, DecisionTreeRegressor, KNeighborsRegressor are compared and analyzed for optimal solutions. Even though all of those methods achieved desirable results, different models have their own pros and cons. The Random Forest method has the lowest error on the training set but is prone to be overfitting. So using RandomForestRegressor method draw the AUC ROC Curve.



* Learning Outcomes of the Study in respect of Data Science

As we checked in dataset not present null values.

Visualizing the categorical data using countplot.

Converted Categorical data into numerical data using Encoder method.

As observed no need to do skewness & outliers checking if present many columns as Categorical data.

After all this split the data into train & test split.

* Limitations of this work and Scope for Future Work

The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combination of pre-processing methods to achieve better prediction accuracy.

Make available features in combining the features and predict it may be improved performance.