

FLIGHT PRICE PREDICTION

Submitted by:

Milind Nerpagar

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

I have referred most of the things data trained class notes and some Machine learning articles from Towards Data science

Selenium issues & data cleaning -- Stack overflow

**INTRODUCTION**

* Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

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.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

Airline and transportation

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

* 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 collection done through scraping different airline websites, like Yatra.com, EsayMyTrip.com

Data stored in .CSV file format

* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

We have handled null and missing values , corrected the data types and format of the independent variables in data cleaning steps

In Data Pre-processing we have checked heat map of correlation matrix to identify the relation between independent variable and target variable

Handled the categorical variables using dummy variables

We used feature selection identifying the correct independent variable to build model

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

The problem belongs to the Airline domain so booking ticket date of the flight and date of journey is more important, also flight duration is more impact on the out put

Any change in the flight duration and ticket booking and date of journey change in output (flight ticket price)

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

Here, you can describe any presumptions taken by you.

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Hardware: Sony vaio laptop ,i5 processor , 4GB Ram

Software : window 10 , Jupyter notebook

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

* Run and Evaluate selected models

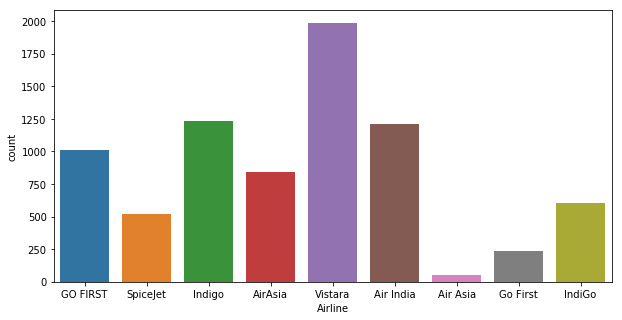
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

* Key Metrics for success in solving problem under consideration

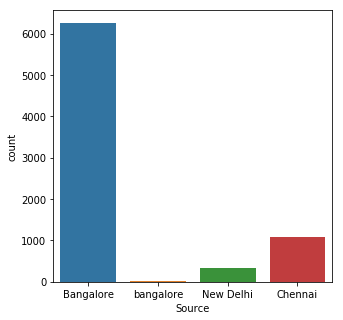
What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

* Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

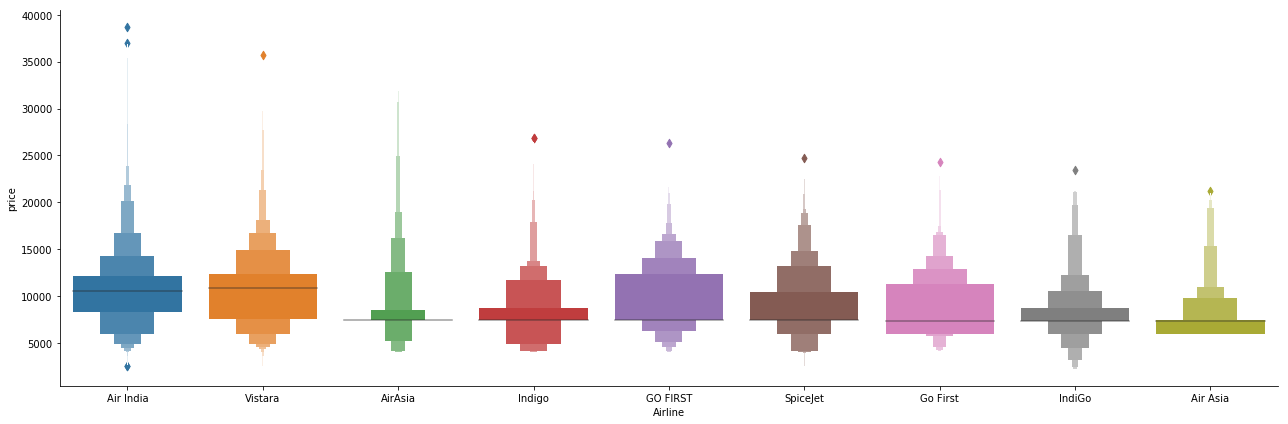


Observation: Airline vistara have highest number of flights

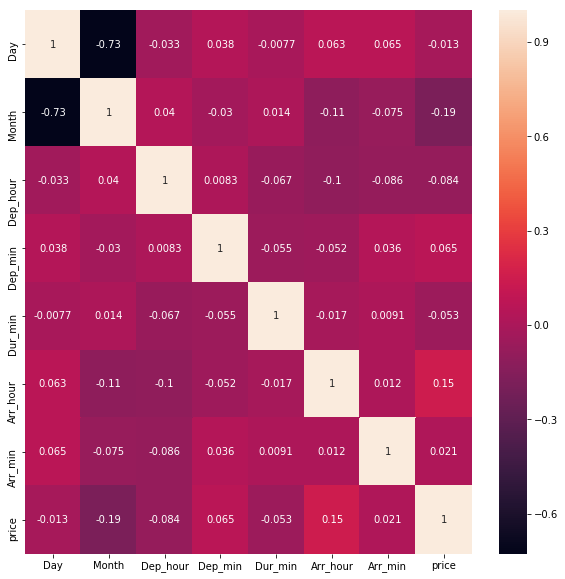


Observation : Data set contains the highest number of flight data from Bangalore

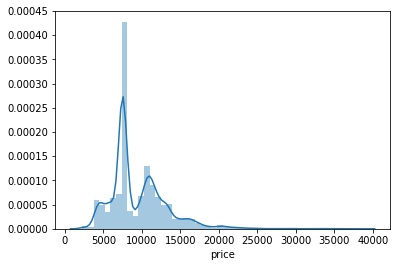
Catplot:



Observation: Air India have highest Air fare compare to other airlines and cheapest airline Air aisa



Observation : Correlation of the heat map , Arrival hour and dur\_hours have positive relation between Feature variable and target variables



Observation: Highly Price distribution happened between 5000 to 10000 Rupees

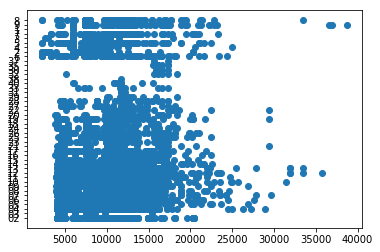


Fig: Price and Hour scatter plot

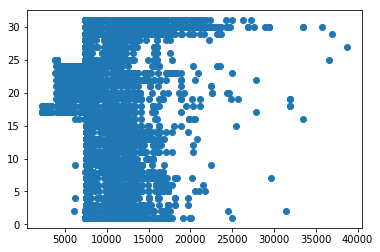
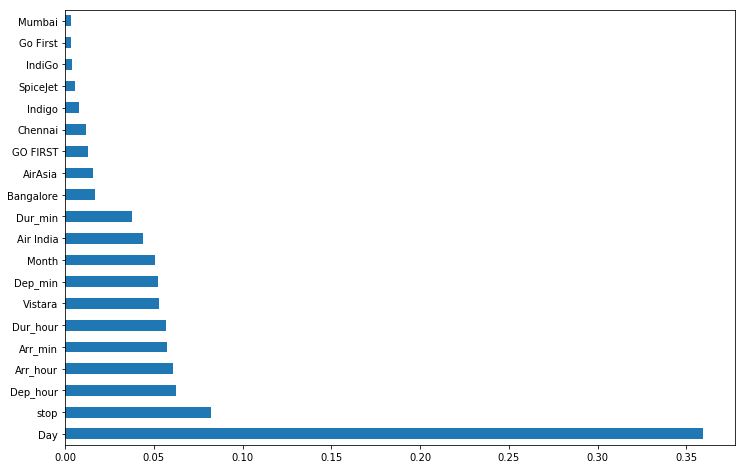


Fig: Price and Day scatter plot



Observation: Bar plots show the clearly feature variables Day and stop ,Dep\_hour, dur\_hour, Arr\_hour have high correlation between the target variables

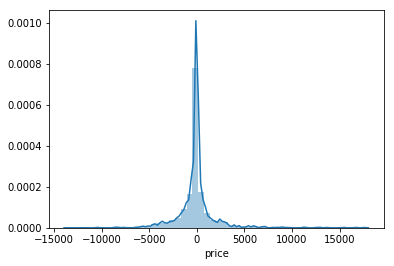


Fig: Test score of the y\_test, and y\_pred

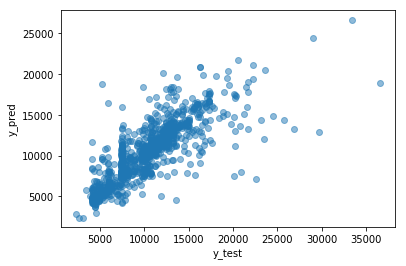


Fig: Scatter plot of y\_test and y\_pred

If different platforms were used, mention that as well.

* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.