

|  |
| --- |
|  |
| FLIGHT PRICE BOOKING PROJECT |
| april-23  CAPSTON PROJECT  REPORT BY: SOUVIK SIKDER |

TABLE OF CONTENTS:-

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation for Choosing the Algorithm

7. Assumptions

8. Model Evaluation and Techniques

9. Inferences from the Same

10. Future Possibilities of the Project

11. Conclusion

12. Reference

PROBLEM STATEMENT:-

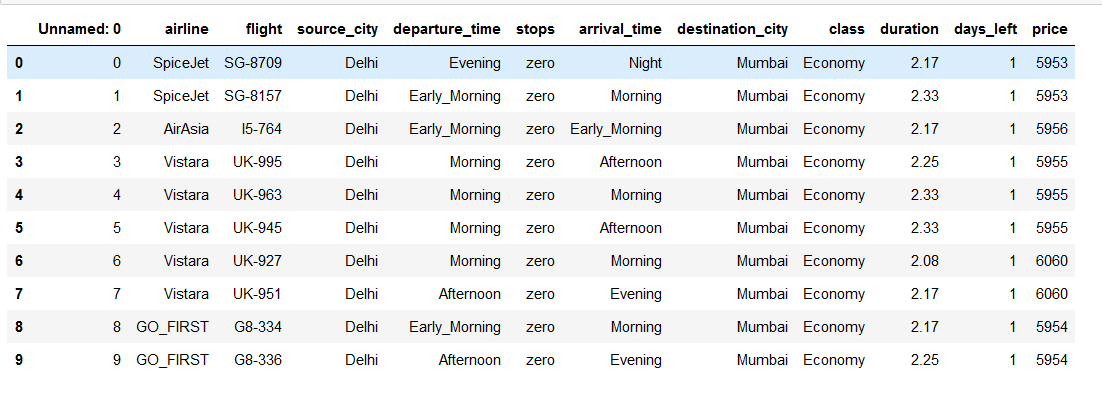
The objective is to analyze the flight booking dataset obtained from a platform which is used to book flight tickets. A thorough study of the data will aid in the discovery of valuable insights that will be of enormous value to passengers. Apply EDA, statistical methods and Machine learning algorithms in order to get meaningful information from it.

PROJECT OBJECTIVE:-

A thorough study of data will aid in drawing meaningful insights from the dataset. Plotting different graphs to see how booking price varies for different airlines, is price varies if booking at 2-3 days before departure, how ticket price vary between economy and business class. And finally come up with a machine learning model that gives almost correct predicted output that can help immensely to passengers boarding different flights.

DATA LOADING:-

After importing different libraries of data manipulation like pandas and data visualization like matplotlib and seaborn, data is loaded for generating insights from it.



DATA DESCRIPTION:-

Data set has 184023 rows and 12 columns and contains 300152 records and 11 features for the flight travel between India's top 6 metro cities. Below are the names of each features along with its description in the form of a table.

|  |  |
| --- | --- |
| FEATURES | DESCRIPTION |
| AIRLINE | Name of the airline company |
| FLIGHT | Plane's flight code |
| SOURCE CITY | City from which the flight takes off |
| DEPARTURE TIME | Time of Departure |
| STOPS | Number of stops between the source and destination cities |
| Arrival Time | Time of Arrival |
| Destination City | City where the flight will land |
| Class | Contains information on seat class |
| Duration | Overall amount of time taken to travel between cities in hours. |
| Days left | Subtracting the trip date by the booking date |
| Price | Ticket price |

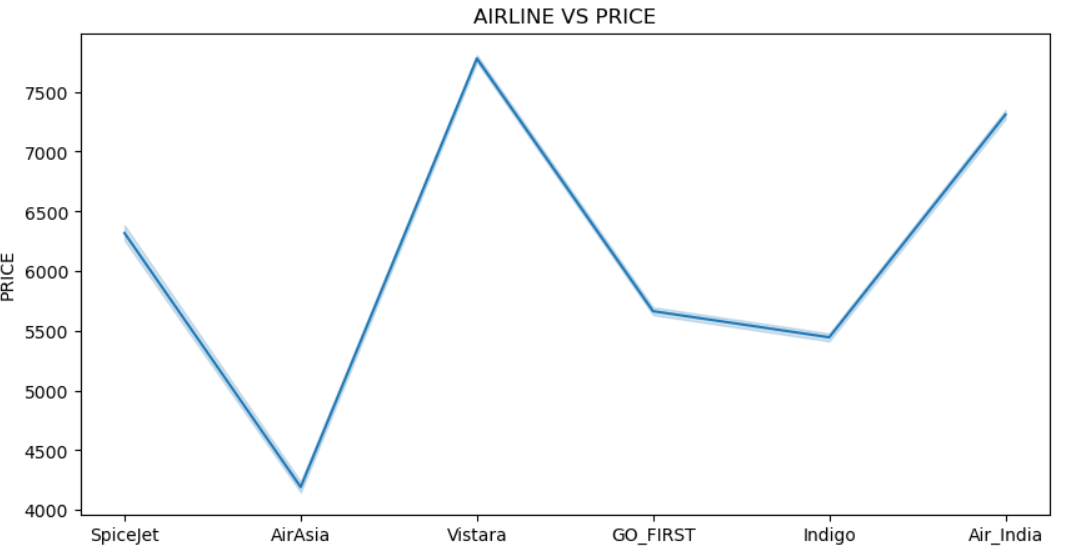
DATA PREPROCESSING STEPS AND INSPIRATION:-

The preprocessing of data includes following steps:-

* Drop Unnamed:0 column which was not required thus reducing dimensionality.
* Getting basic information about dataset to look for object type and integer type data.
* Using code to get statistical inferences about data.
* Checked for null values from dataset and found none.
* Checked for multicollinearity and found non-collinear among inputs.
* Airline, Days\_left and Duration is strongly corelated with the output variable Price.

VISUALIZATION:-

* Plot graph to find if price varies with different airlines.
* Plot graph to find if price varies when tickets are booked 2-3 days before departure.
* Plot graph to see how tickets vary for economy and business class.



INSIGHTS AFTER PREPROCESSING STEPS:-

* To look for how many passengers have passengers have booked flight in business class and found none. All passengers booked flight in economy class.
* Checked which airline has given more service. Found Vistara to have given more service and Spice jet the least.
* Checked which city has lowest takeoff. Chennai was found having lowest takeoff.
* Average price of all flights between Delhi and Mumbai was found to be 6059.82

CHOOSING ALGORITHM FOR THE PROJECT:-

Before choosing the algorithm the steps performed are initializing dependent and independent variables, splitting the dataset into train and test split. Test size is 20% and training data is 80%. Standardizing data which is a kind of data preprocessing step to convert the structure of different dataset into one common format of data. In this project we applied 3 machine learning algorithm-Multiple Linear Regression, Decision Tree Regressor, Random Forest Regressor and checked which algorithm is best suited for the dataset provided.

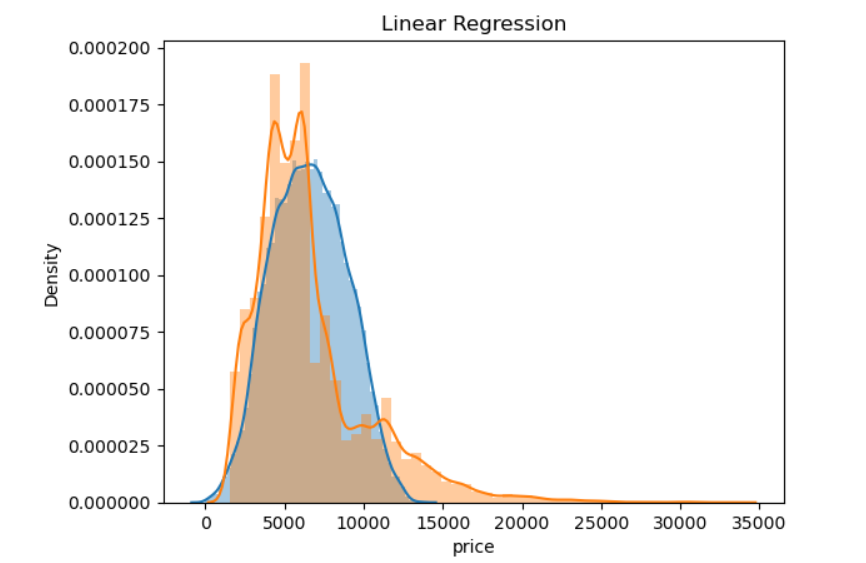
MODEL EVALUATION AND ACCURACY CHECK:-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | R2\_SCORE | MSE | RMSE | MAE |
| MULTIPLE LINEAR REGRESSION | 0.4190 | 7794490.23 | 2791.26 | 2037.16 |
| DECISION TREE REGRESSOR | 0.748 | 3381152.43 | 1838.79 | 683.14 |
| RANDOM FOREST REGRESSOR | 0.8580 | 1905008.24 | 1380.22 | 623.026 |

Accuracy check for different models showed Random Forest Regressor to predict well.

MOTIVATION FOR CHOOSING THE ALGORITHM:-

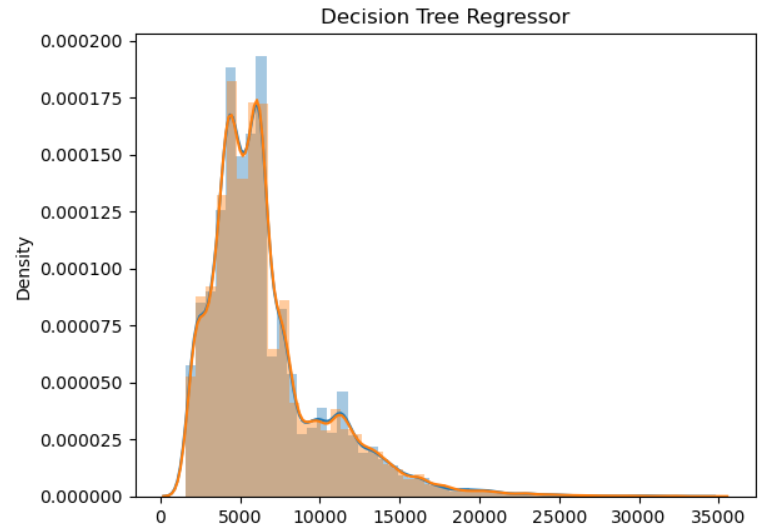
LINEAR REGRESSION DISTPLOT:-



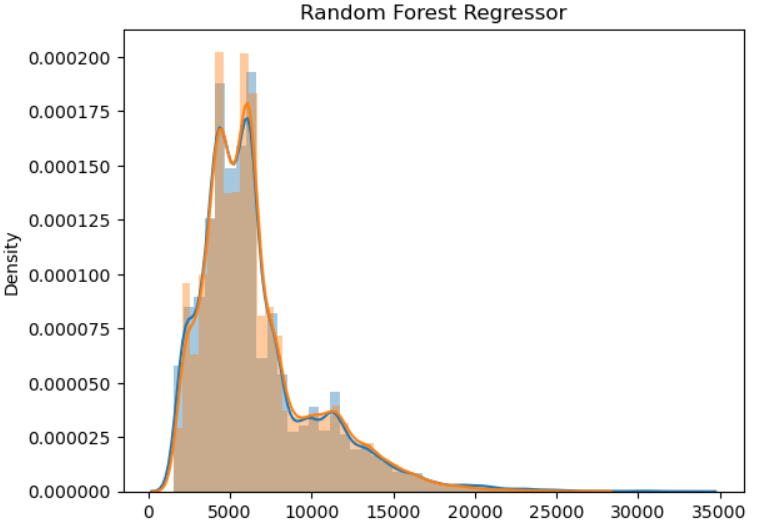
GRAPH SHOWING THAT PREDICTED AND ACTUAL POINTS ARE NOT AT ALL MERGING THERBY ACCURACY OF MODEL IS VERY LOW.

DECISION TREE REGRESSOR DISTPLOT:-

GRAPH DEPICTING THE MERGE OF ACTUAL AND PREDICTED POINTS,THERBY THE MODEL WORK BETTER THAN LINEAR REGRESSION.



RANDOM FOREST REGRESSOR DISTPLOT:-



RANDOM FOREST REGRESSOR SHOWED BETTER PERFORMANCE IN TERMS OF ACCURACY FROM BOTH LINEAR AND DECISION TREE REGRESSOR WHICH IS ALSO DEPICTED IN THE GRAPH.

ASSUMPTIONS:-

* LINEAR REGRESSION ASSUMPTIONS:-
* The relationship between variables is linear
* Homoskedasticity, or that the variance of the variables and error term must remain constant
* All explanatory variables are independent of one another
* All variables are normally-distributed.
* DECISION TREE REGRESSOR ASSUMPTIONS:-
* The whole training set is considered as the root.
* Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
* On the basis of attribute values, records are distributed recursively.
* We use statistical methods for ordering attributes as root or the internal node.
* RANDOM FOREST REGRESSOR:-

No formal distributional assumptions, random forests are non-parametric and can thus handle skewed and multi-modal data as well as categorical data that are ordinal or non-ordinal.

MODEL EVALUATION AND TECHNIQUE:-

Different metrics used to evaluate model accuracy which showed the performance of Random Forest Regressor better than Linear Regression and Decision Tree Regressor.

INFERENCE AFTER MODEL EVALUATION:-

If the dataset contains features some of which are Categorical Variables and some of the others are continuous variable Decision Tree is better than Linear Regression, since Trees can accurately divide the data based on Categorical Variables. There are some situations where Linear Regression outperforms Random Forests, but I think the more important thing to consider is the complexity of the model.

Linear Models have very few parameters, Random Forests a lot more. That means that Random Forests will overfit more easily than a Linear Regression. Linear regression is a linear model, which means it works really nicely when the data has a linear shape. But, when the data has a non-linear shape, then a linear model cannot capture the non-linear features.

So in this case, you can use the decision trees, which do a better job at capturing the non-linearity in the data by dividing the space into smaller sub-spaces.

Random forest leverages the power of multiple decision trees. It does not rely on the feature importance given by a single decision tree.

The random forest chooses features randomly during the training process. Therefore, it does not depend highly on any specific set of features. This is a special characteristic of random forests over bagging trees.

Therefore, the random forest can generalize the data in a better way. This randomized feature selection makes a random forest much more accurate than a decision tree.

FUTURE POSSIBILITIES OF THE PROJECT:-

The model accuracy can be higher if hyperparameter tuning is done for the models. The project has surely presented us with powerful insights about the data. Study of data has given us powerful insights that will be enormous help to passengers in future.

CONCLUSIONS:-

Model Random Forest Regressor has given accurate prediction of price of the airlines with respect to many input features. We found out passengers board Vistara flight mostly while Spice jet was boarded the least. Chennai being the least take off airport while Delhi being the highest.

REFERENCES:-

[User guide and tutorial — seaborn 0.12.2 documentation (pydata.org)](https://seaborn.pydata.org/tutorial.html)

[pandas documentation — pandas 2.0.0 documentation (pydata.org)](https://pandas.pydata.org/docs/)

[Data Science Courses Online | Learn Fundamentals to Advanced | Intellipaat](https://intellipaat.com/course-cat/data-science/)