FLIGHT PRICE PREDICTION

SUBIMITTED BY- AKASHDEEP SINGH MANRAL

BATCH- 1834

ACKNOWLEDGEMENT

The project consists of 2019 flight data, of some cities, based on which flight price prediction is done.

I want to thank my intern mentor miss- Swati Mahaseth for providing assistance in solving my queries, with her help and guidance I was able to complete my project successfully.

INTRODUCTION

1-Problem Statement-

1- 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-

- a-Time of purchase patterns (making sure last-minute purchases are expensive)
- b- 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)

2- 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.

2-About Dataset-

1- The data consists of two sets , i.e- Training dataset and Testing dataset.

2- Training dataset is of the year 2021 collected from various websites.

3-The collected flight data contains source to destination of some major cities.

4- The dataset consists of different columns that affects the price of tickets.

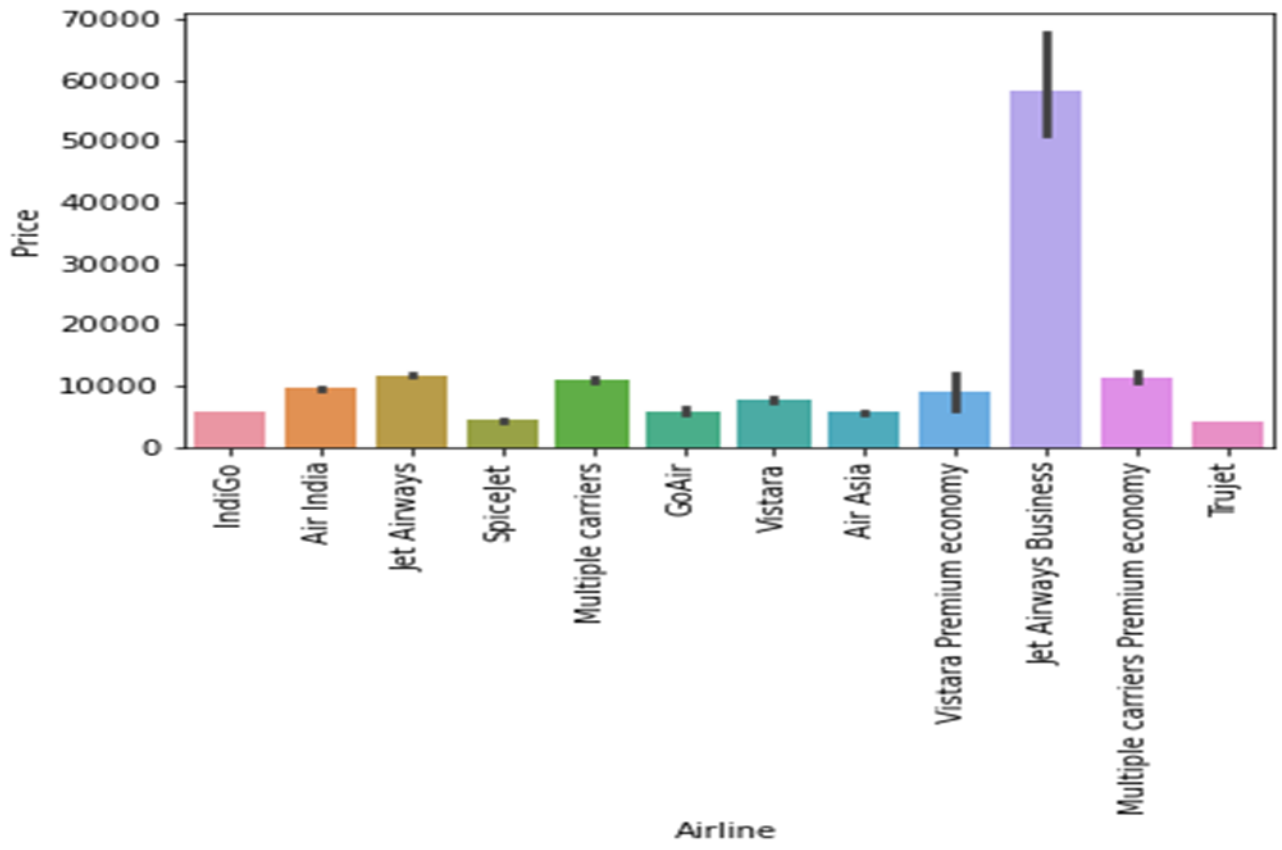
5-The testing dataset consists of similar features except the price feature.

6-Based on machine learning model the price feature of testing dataset is predicted.

Analytical Problem Framing

1-Exploratory Data Analysis-

1- Airline vs Prices



Obs-

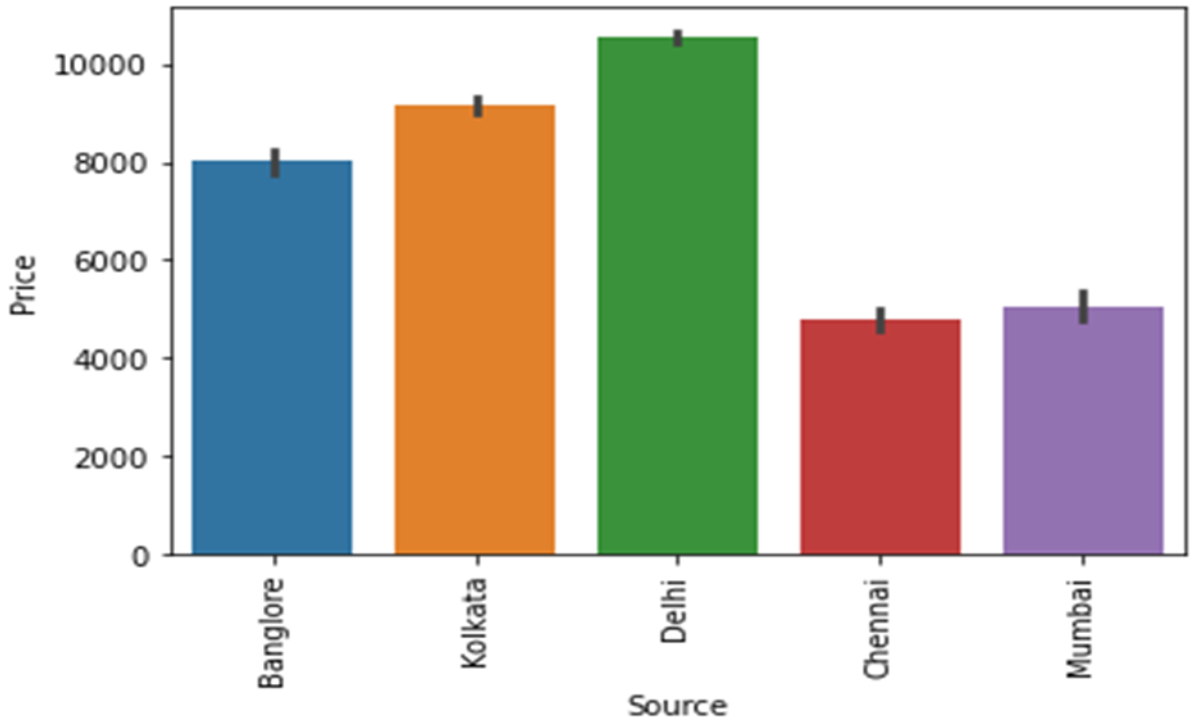
1-Jet Airways Business have the highest ticket price, followed by Multiple Carriers Premium Economy and Jet Airways.

2-Air India have higher prices than Indigo.

3- Spice Jet, GoAir and Air Asia have similar price ranges.

4- TruJet have the least prices.

2-Source vs Prices



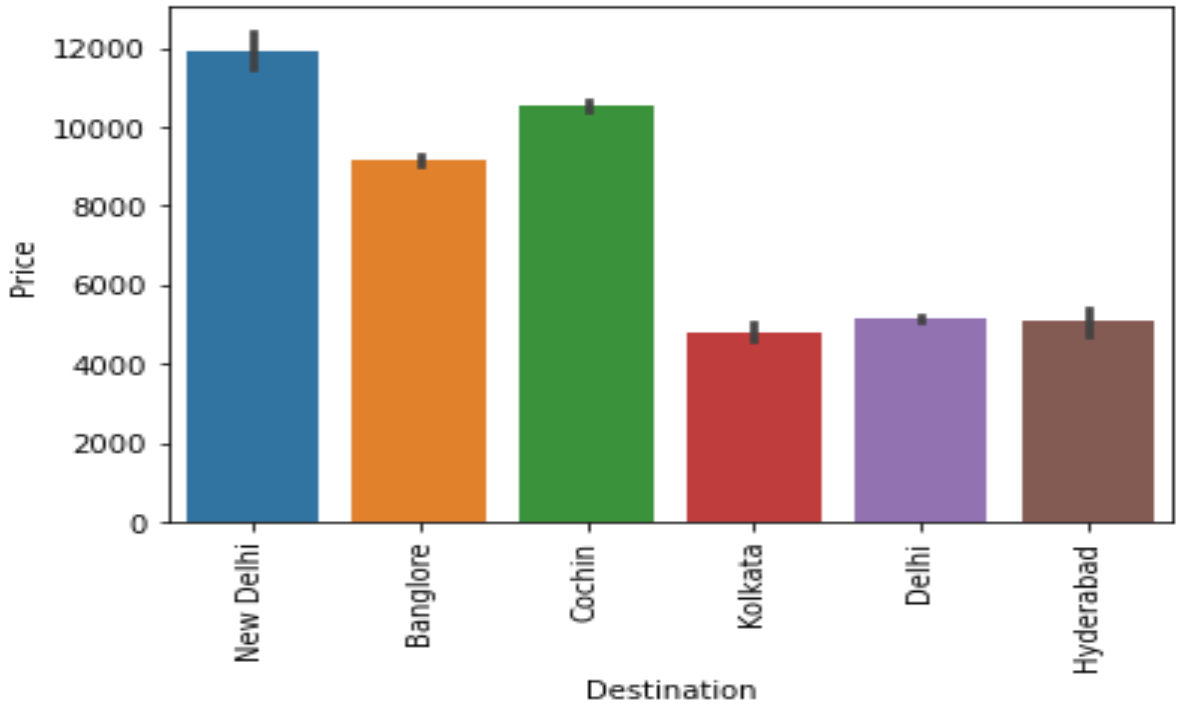
Obs-

1- Tickets from Delhi have highest price ranges followed by Kolkata and Bangalore.

2- Then followed by Mumbai.

3- Tickets from Chennai have the least prices.

3-Destination vs Prices



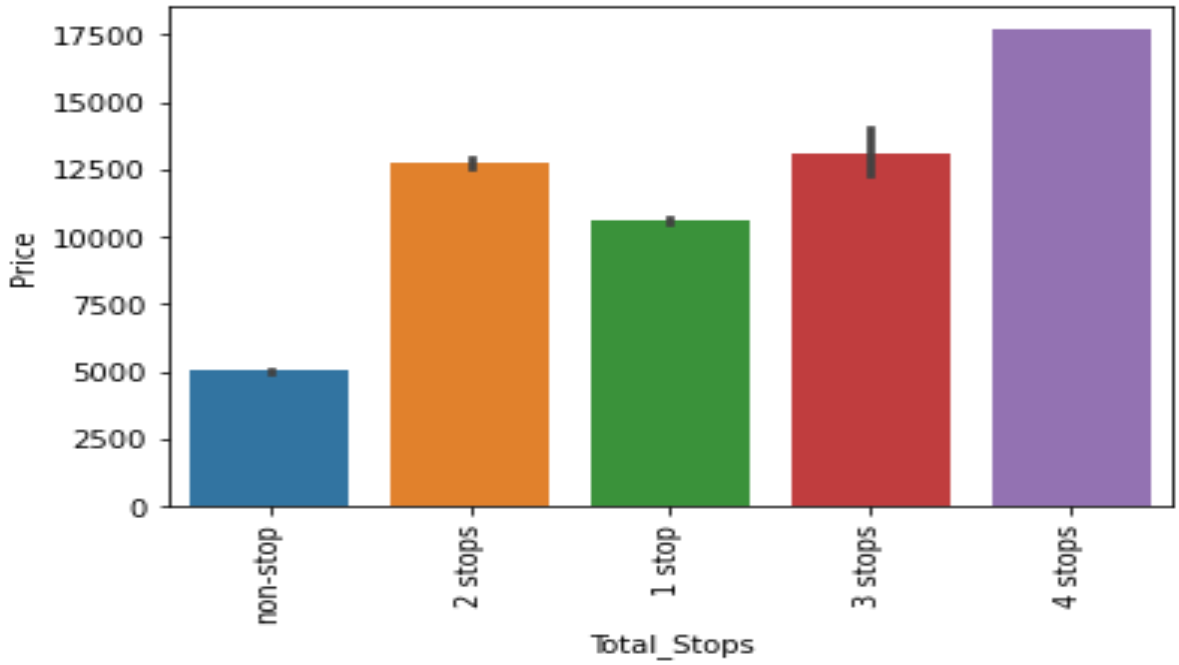
Obs-

1-Tickets with NewDelhi as destination have highest prices followed by Coachin and Bangalore.

2- Delhi and Hyderabad have similar price range.

3-Tickets with Kolkata as destination have lowest prices among these destinations.

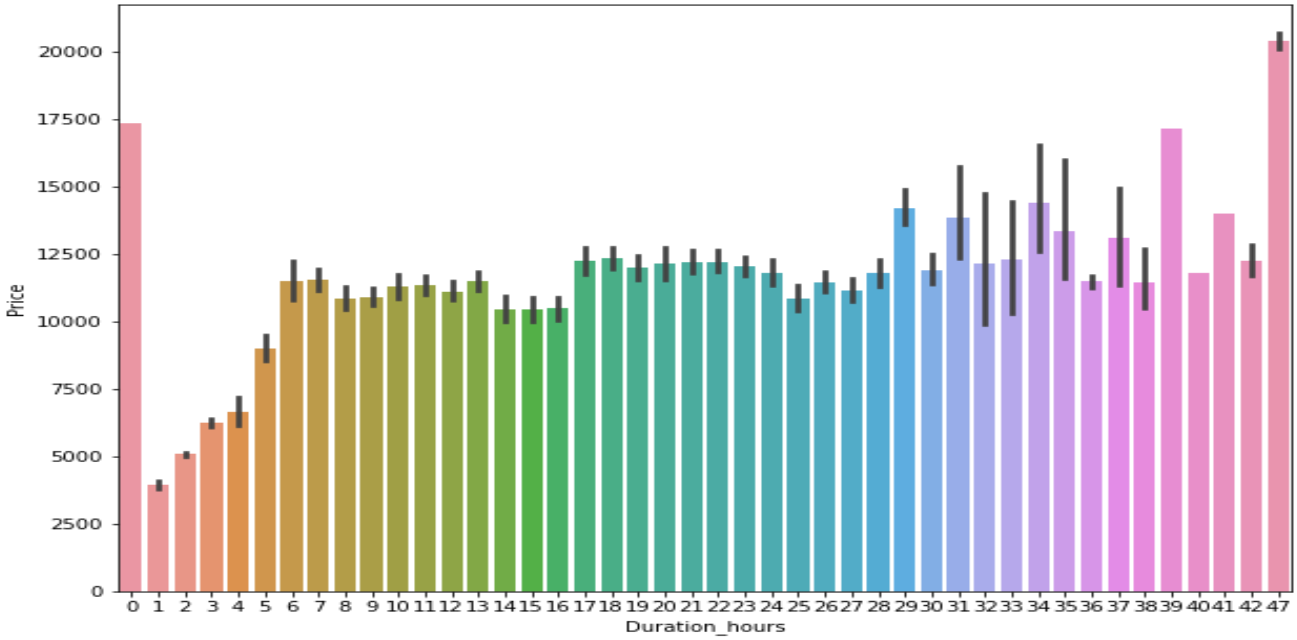
4-Total stops vs Price



Obs-

1-higher the no. of stops in a route , the higher the price of tickets.

5-Duration hours vs price

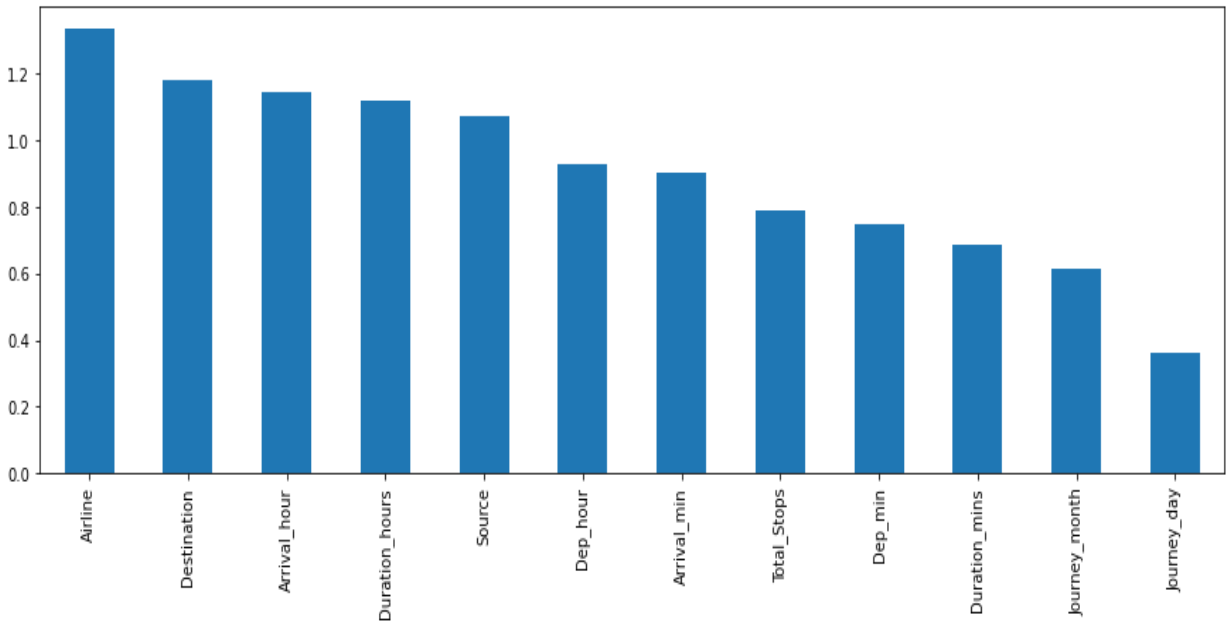


Obs-

1-There is a pattern of increase in prices with respect to the hours.

2- But some Journeys which are quick and lesser than 1 hours are also having high prices, this shows that prices does no directly depends upon journey hours.

6- Important Features for Price Prediction



Obs-

1- This histogram shows, in descending order, the importance of features in Price Prediction.

2- most important features are Airline, Destination, Arrival Hour respectively

3- Least important features- journey day, journey month, destination mins respectively.

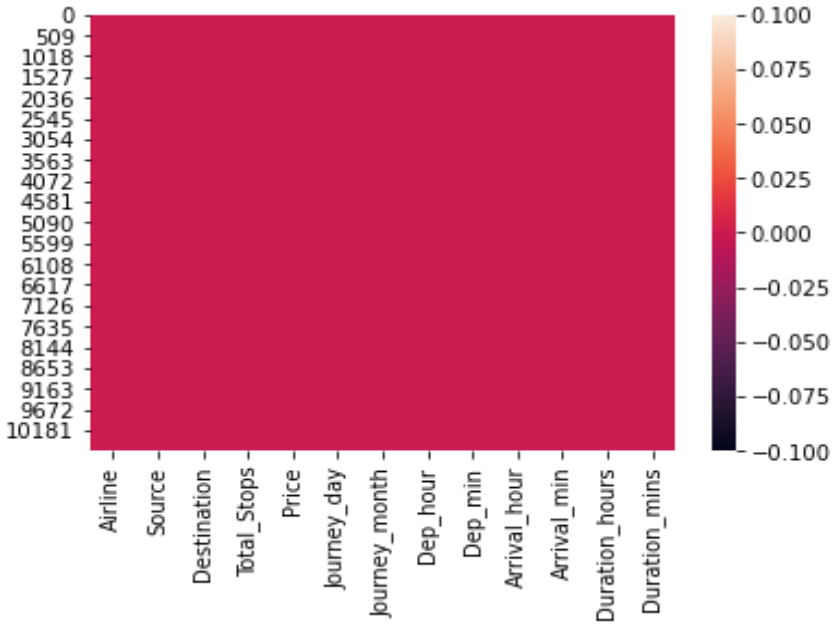
2- Working on Testing and Training datasets-

1- finding the null values on the training set.

2- filling the null values, with mean in numerical columns and,with mode in categorical columns.

3-Eliminaing the columns having more than 70% null values

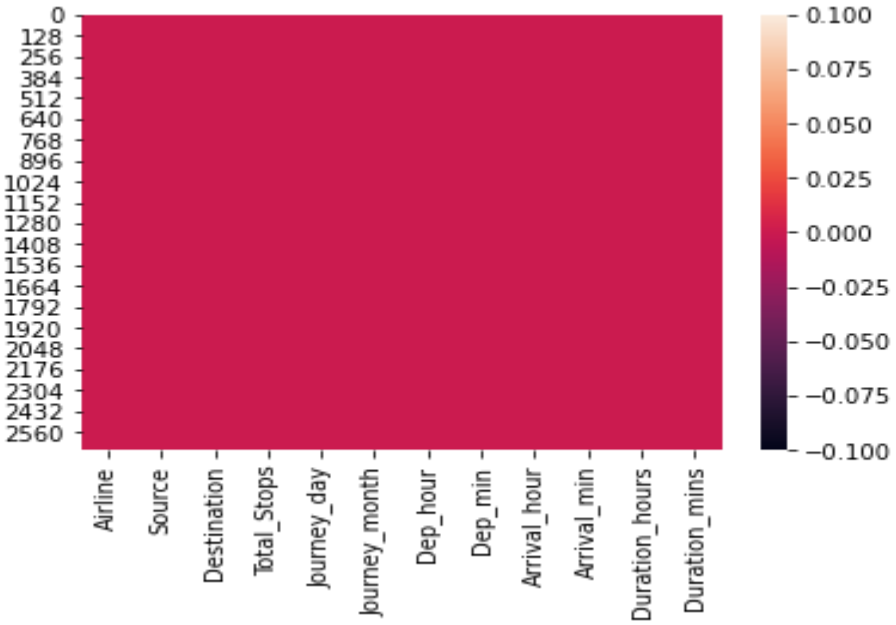
4-



Heatmap showing, no null values in the training dataset.

5- performing the same feature engineering steps in testing data set.

6-



Heatmap showing , no null values present in the testing dataset.

7- Joining both the training and testing datasets.

8- Encoding using, Ordinal Encoder, for encoding the categorical columns in the d(train+test) set.

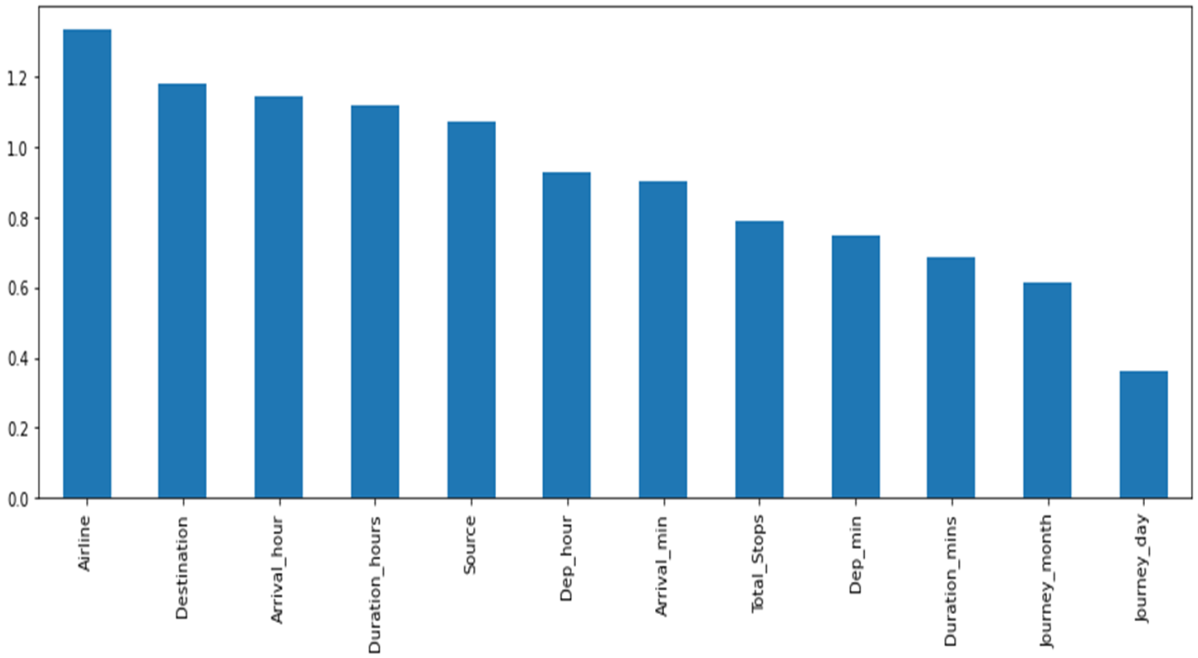
9-Again splitting the two sets, using iloc method, into d_trainset and d_testset

10- Using the Variance Threshold method to find out the feature having only same responses, (i.e- Utilities) and removing it from both d_trainset and d_testset

11- Dropping columns based on correlation using Mutual Info (But no columns were droped).

12- Now separating x(feature) and y(target) variable from the d_trainset.

13- finding out the most contributing feature towards the targe variable and removing the least contributing feature from d_testset and d_trainset.(i.e-



Obs-

1- This histogram shows, in descending order, the importance of features in Price Prediction.

2- most important features are Airline, Destination, Arrival Hour respectively

3- Least important features- journey day, journey month, destination mins respectively.

Model/s Development and Evaluation

1-After selecting the x and y variable,we then perform the train_test_split and and check the training and testing in various models.

2-Models used – LinearRegression, DecisionTreeRegressor, KNeighborsRegressor, SVR, RandomForestRegressor, AdaBoostRegressor, GradientBoostRegressor,

3-Train_test_split was used to find out the accuracy of each model.

4-After this cross_val_score of each model was calculated.

5-By compairing the accuracy score with cross_val_score, RandomForestRegressor was found to be most efficient model.

6-After this Hypertest tuning was done through GridSearchCv to find out the best parameter for RandomForestRegressor

Prediction-

1-RandomForestRegressor was used for predicting the flight prices in testing dataset.

2-The predicted the price with 83% accuracy score.

Limitations-

1-The model worked with 83% accuracy , thus the results are not 100% true.

2- There is a marginal error of 17% in he predicted price.

THANK YOU