FLIP ROBO

FLIGHT PRICE PREDICTION

Submitted by:

Rijul Kumar

ACKNOWLEDGMENT

Reference that i have used are:

- Data Trained Education online video
- Materials provided by Flip Robo
- Data extracted from https://www.yatra.com/
- Geeks for Geeks
- Stackoverflow

INTRODUCTION

- ## Business Problem Framing

  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.
  2. So, we have to work on a project where we have to collect data of flight fares with other features and work to make a model to predict fares of flights.
  3. We are supposed to scrape at least 1500 rows of data for the model.

- ## Conceptual Background of the Domain Problem

  The prices of flight ticket depends on many factors. Such as:

  - Airline Name
  - Date of journey
  - Source and Destination
  - Departure time and Arrival time
  - Duration
  - Total stops
  - And so on...

- ## Motivation for the Problem Undertaken

  - We are supposed to build a model using Machine Learning in order to predict the price of the flight tickets.

  - This model will then be used to understand how exactly the prices vary with the variables. This can be used to evaluate the best interval of time to purchase flight tickets at its cheapest price.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem
  - First of all i imported data from excel file to dataframes using pandas.
  - After that i cleaned the data (like converted Duration and Price from object datatype to numeric datatype, etc).
  - I used .dtypes to know data type of each column of dataframe.
  - After that i used .describe() to know the statistical information (such as max, min value,etc) of continuous data columns in dataframe.
  - Then i used .shape to know shape of dataframe.

- Data Sources and their formats
  - Training data has been extracted from ‘https://www.yatra.com/’ using Selenium.

  - Using Selenium, data was extracted to excel file named ‘flight_price_data.xlsx’.

- <u>Data Preprocessing Done</u>

  - First of all for data preprocessing i checked whether there is a NULL value or not in dataframe using heatmap as well as .isnull()
  - After that i used count plots from seaborn library to plot all categorical columns for visualisation.
  - Next i used Density plots from seaborn library to plot all continuous columns for visualisation.
  - Afterwards i used other graphs to visualize pattern of Price with respect to other columns.
  - Then i encoded the dataframe using Ordinal Encoder.
  - After that i checked for correlations using heatmaps, correlation matrix and BAR plot.
  - Finally i confirmed correlations using VIF.
  - After that i checked and removed skewness for continuous data columns (except target variable) using box-cox transformation.
  - Finally i checked outliers using boxplot as well as z-score method and discovered that only target variable had outliers so we ignored outliers.

- ## Data Inputs- Logic- Output Relationships

    - Data Input :

      These are basically the factors (such as airline name, date of journey, source, destination, departure time, arrival time, duration, total stops, etc) which affects the prices of flight tickets.

    - Data Output :

      Our Target variable is ‘Price' which is the price of the flight tickets and we are supposed to predict it with the help of data input (i.e factors affecting flight ticket prices).

- ## Hardware and Software Requirements and Tools Used

    - Hardware used:

        i. Laptop with intel core i5 7th gen

        ii. Internet connection for web scraping

    - Software used:

        i. Jupyter notebook

        ii. Required python libraries such as numpy, pandas, seaborn, matplotlib, etc

        iii. Required libraries for model such as sklearn, etc

        iv. Required libraries for web scraping such as selenium,etc

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
    - After preprocessing data (removing NULL, encoding, checking for high correlations, removing skewness and outliers) i separated columns into features and target.
    - As this is a regression problem so we tried 4 models - LinearRegression, SVR, RandomForestRegressor and DecisionTreeRegressor
    - I also tried 4 metrics method - r2_score, mse, rms, mae
    - Then i used Lasso for regularization.
    - Finally i used Ensemble Technique.

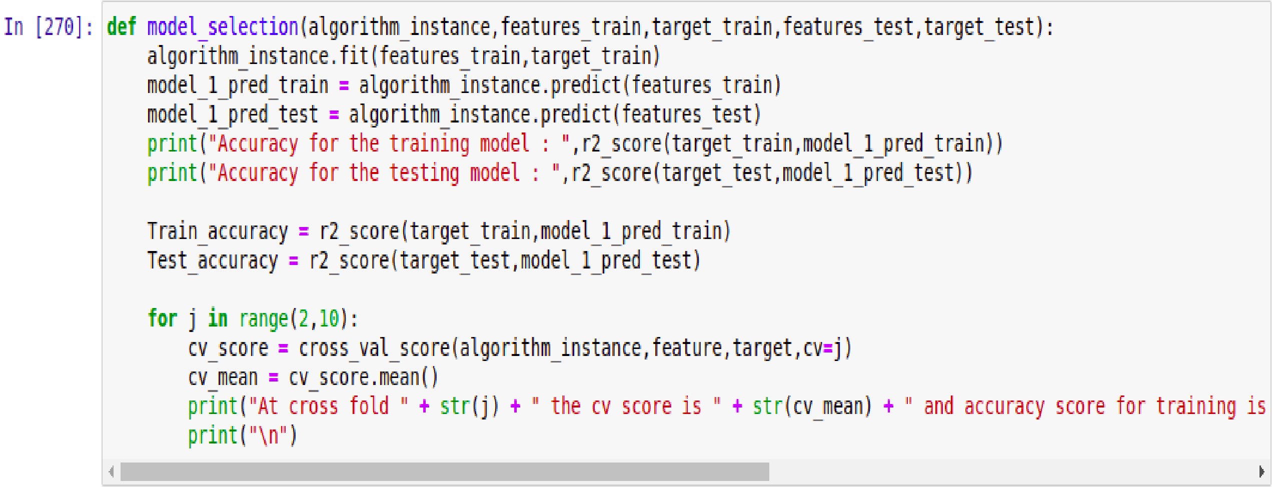
- Testing of Identified Approaches (Algorithms)

  As this is a regression problem so we tried following 4 models -

    - LinearRegression
    - SVR
    - RandomForestRegressor
    - DecisionTreeRegressor

- Run and Evaluate selected models

  I defined a function model and then tried 4 different models using it



```
In [270]: def model_selection(algorithm_instance,features_train,target_train,features_test,target_test):
              algorithm_instance.fit(features_train,target_train)
              model_1_pred_train = algorithm_instance.predict(features_train)
              model_1_pred_test = algorithm_instance.predict(features_test)
              print("Accuracy for the training model : ",r2_score(target_train,model_1_pred_train))
              print("Accuracy for the testing model : ",r2_score(target_test,model_1_pred_test))

              Train_accuracy = r2_score(target_train,model_1_pred_train)
              Test_accuracy = r2_score(target_test,model_1_pred_test)

              for j in range(2,10):
                  cv_score = cross_val_score(algorithm_instance,feature,target,cv=j)
                  cv_mean = cv_score.mean()
                  print("At cross fold " + str(j) + " the cv score is " + str(cv_mean) + " and accuracy score for training is
                  print("\n")
```

Result that i got for each model :

- LinearRegression :
  At random state 9 the training accuracy is : 0.5116297288576696
  At random state 9 the testing accuracy is : 0.5056275585665695
  At cross fold 2 the cv score is -0.2727951750529799 and accuracy score for training is 0.5116297288576696 and accuracy score for testing is 0.5056275585665695

- SVR :
  Accuracy for the training model : -0.06462320433380886
  Accuracy for the testing model : -0.0950997703540779
  At cross fold 2 the cv score is -0.9216749933443029 and accuracy score for training is -0.06462320433380886 and accuracy score for testing is -0.0950997703540779

- <u>RandomForestRegressor</u> :
  Accuracy for the training model :
  0.9741278615603357
  Accuracy for the testing model :
  0.7942563514092913
  At cross fold 2 the cv score is -0.6672666487812541 and accuracy score for training is 0.9741278615603357 and accuracy score for testing is 0.7942563514092913

- <u>DecisionTreeRegressor</u> :
  Accuracy for the training model :
  0.9998320583184103
  Accuracy for the testing model :
  0.6399770563950407
  At cross fold 2 the cv score is -1.0505756461889422 and accuracy score for training is 0.9998320583184103 and accuracy score for testing is 0.6399770563950407

Finally i concluded that RandomForestRegressor() gives best accuracy and hence i took it as main model.

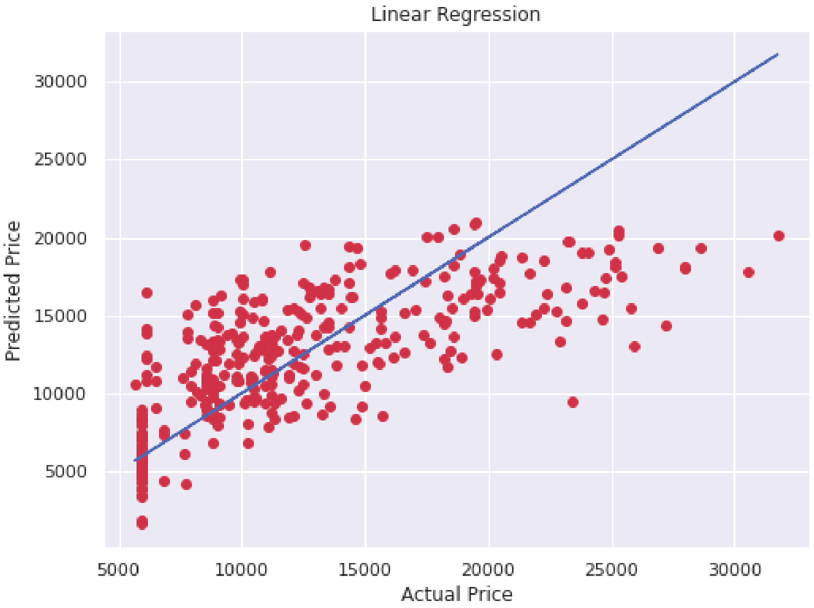
- <u>Key Metrics for success in solving problem under consideration</u>

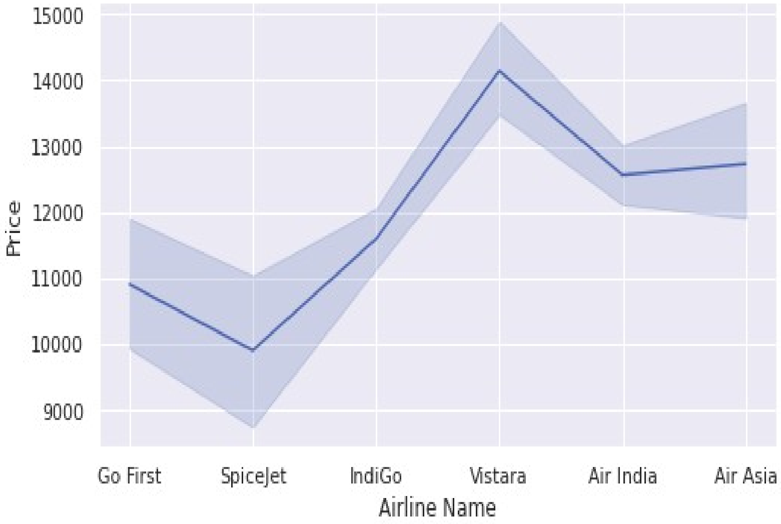
  I tried 4 different metrics method:
  - r2_score
  - mse
  - rms
  - mae

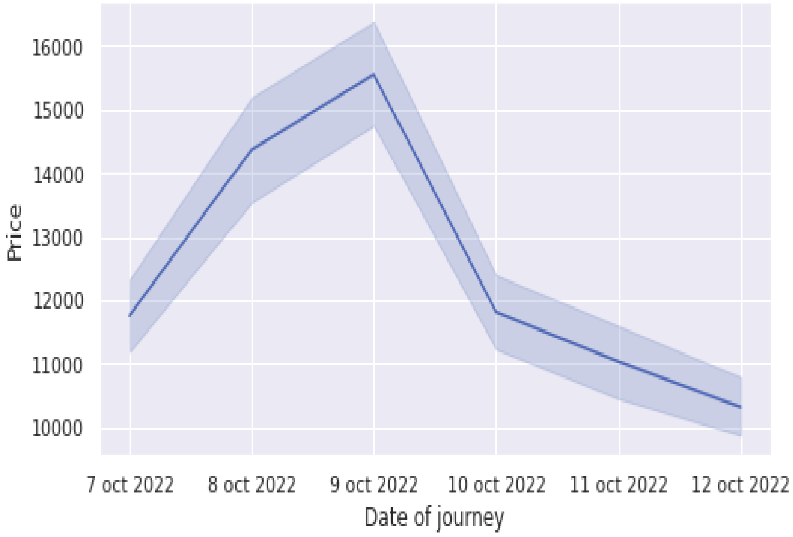
  I got best results from r2_score and hence used it in final model

- <u>Visualizations</u>

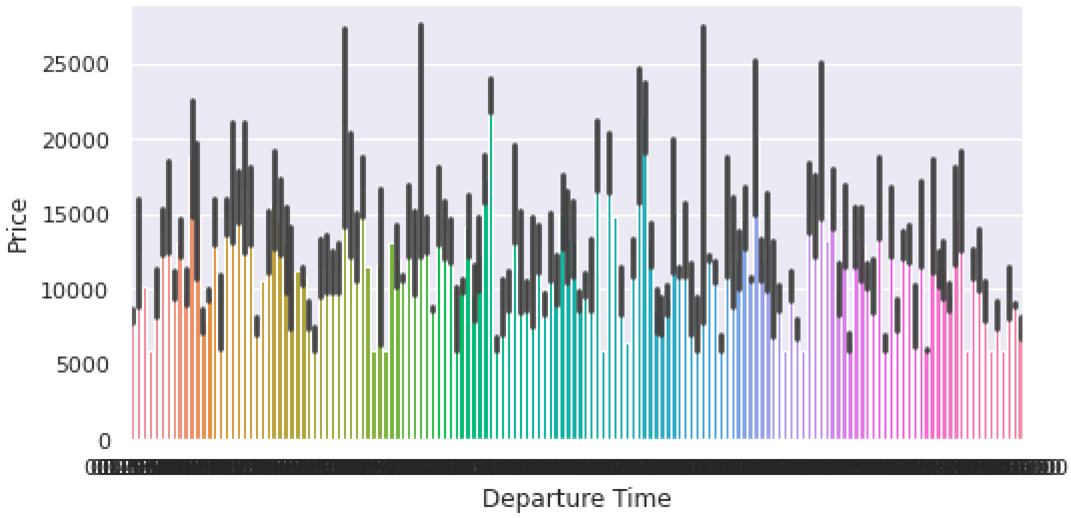
  We can see that for Linear Regression almost all points lie near best fit curve but not on it which shows large amount of error.







As we can see that more the difference between date of booking and date of journey, lesser will be the flight price. We can also see that cheapest flight is Spicejet and most expensive is Vistara.



This is barplot of Price vs Departure time (from 0:00 to 23:59).

Here we can see that prices vary throughout the day.

- ## Interpretation of the Results

  So the results which i got were:

  - Although RandomForestRegressor() is best model but DecisionTreeRegressor() can also be used as they are also very good models.
  - Above point can be seen through r2_score which is the also best among 4 metrics tried.
  - We got our final accuracy (r2_score) as 80.72% after hypertuning.

CONCLUSION

- Key Findings and Conclusions of the Study
  - From this study i learnt that many airlines increase their prices for people who book their flights closer to date of their departure.
  - Prices of Flight tickets are not very highly dependent on any of the factors taken in this project , they do have good dependency on some.
  - Prices of Flight tickets fluctuate a lot throughout the day.
  - Some companies provides slightly cheaper flight tickets than others.

- Learning Outcomes of the Study in respect of Data Science

  Some problems faced and their solution (using visualisation and algorithm) used were:

  - When i tried to go from yatra.com homepage to flight list page of yatra.com using web scraping script the website stopped me. So i directly visited the url of flight list page of yatra.com using chromedriver.
  - After web scraping some data which was supposed to be continuous data column was

stored as object datatype in excel file. So what i did was in data cleaning part of model i converted them to float datatype using .astype()

- ## Limitations of this work and Scope for Future Work

Some limitations are :

- There are many other factors which are not in the data which may play major role in prices of some flight prices such as discount offers,etc.
- Unrelated factors such as availability of beverages may indirectly also affect price of flight tickets.
- With evolving technology, prices of flight tickets may also decrease.

Scope for future work :

- This can be made further accurate by taking more and more factors as well as more source and destinations (domestic as well as international) into account.