Flight Price Prediction Project

Submitted By: -

Prashant Pathak

Goal

- Forecast flight prices
- Selecting optimum time for travel
- Selecting the cheapest flight to the desired destination

Scraping

- Source & Destination
- Date (Feb 2022 to April 2022)
- Price
- Duration
- Total Stops
- Airline

Scraped Routes

- SVO
- NYC
- PAR
- RUH

Steps of EDA

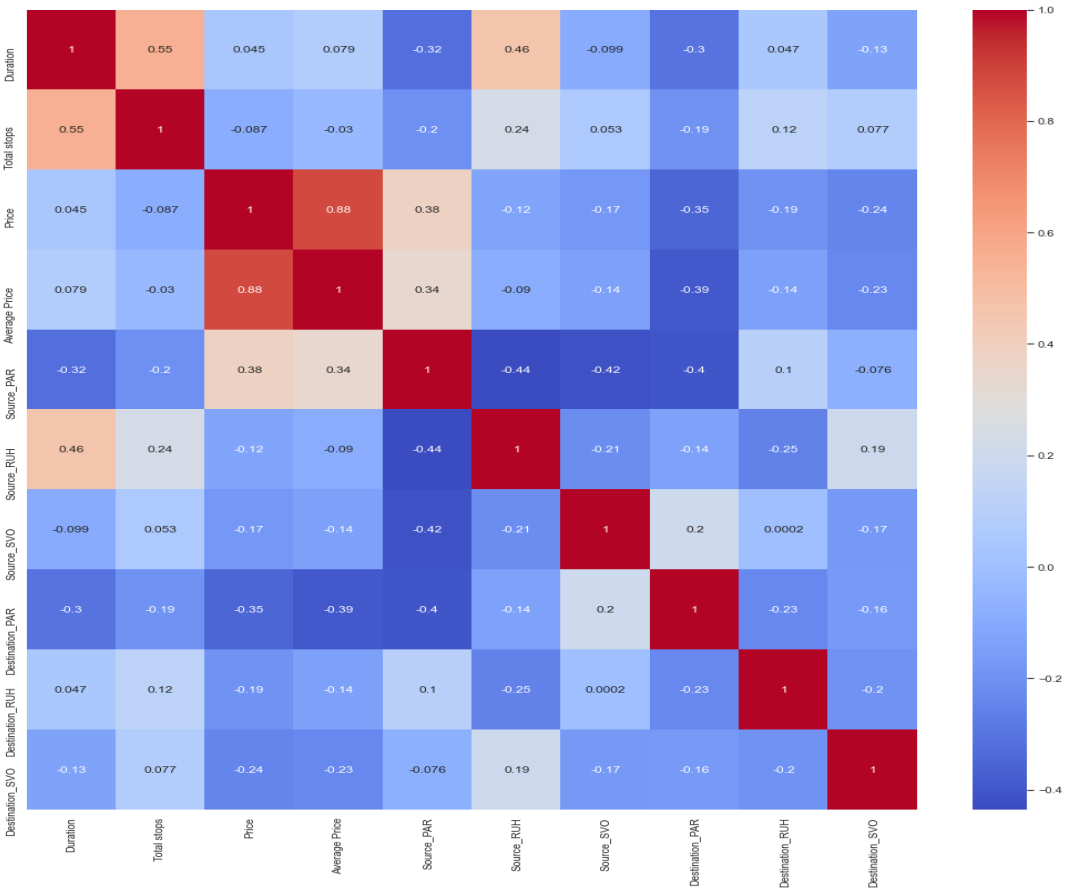
- Importing necessary libraries
- Loading Scraped data
- Defining function to clean the data
- Studying outliers and then deal with outliers
- Check the null values
- Handling categorical data
- Drop unusual data
- Creating final data frame for process

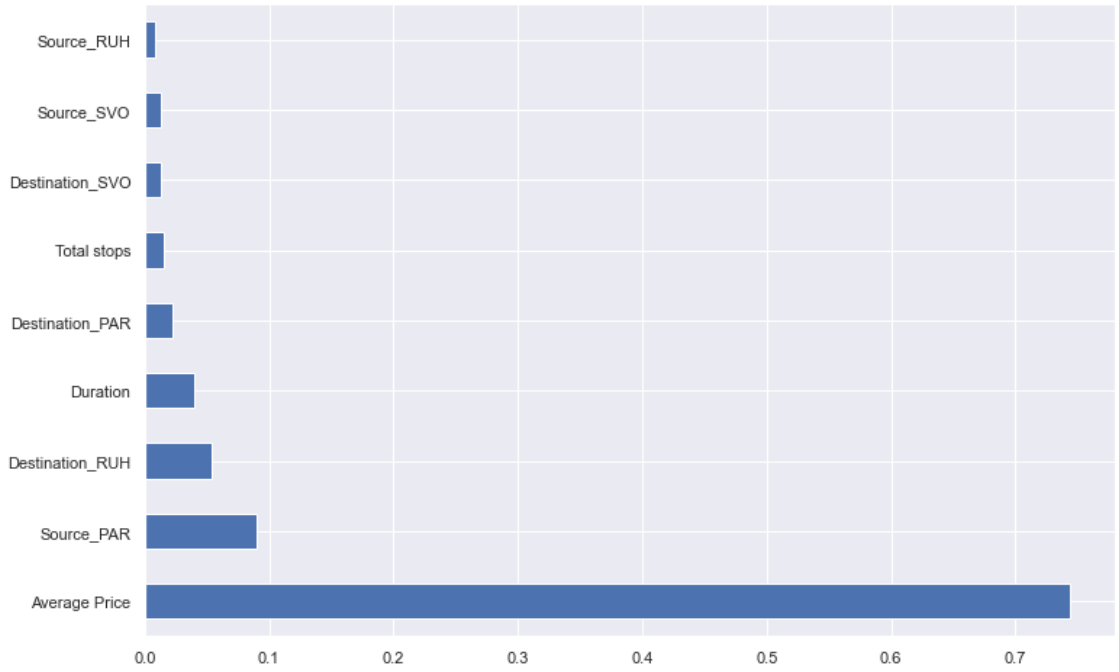
Now we have a good data Structure now we discuss about in our new data frame-

- Our new data frame has 50097 rows and 10 columns.
- Name of columns is:

```
 ‘Duration', 'Total stops', 'Price', 'Average Price', 'Source_PAR',
'Source_RUH', 'Source_SVO', 'Destination_PAR', 'Destination_RUH',
'Destination_SVO'
```

- Heatmap and plotting graph for finding relation between features and targeted variable:


Duration
Total stops
Price
Average Price
Source_PAR
Source_RUH
Source_SVO
Destination_PAR
Destination_RUH
Destination_SVO
1 0.55 0.045 0.079 -0.32 0.46 -0.099 -0.3 0.047 -0.13
0.55 1 -0.087 -0.03 -0.2 0.24 0.053 -0.19 0.12 0.077
0.045 -0.087 1 0.88 0.38 -0.12 -0.17 -0.35 -0.19 -0.24
0.079 -0.03 0.88 1 0.34 -0.09 -0.14 -0.39 -0.14 -0.23
-0.32 -0.2 0.38 0.34 1 -0.44 -0.42 -0.4 0.1 -0.076
0.46 0.24 -0.12 -0.09 -0.44 1 -0.21 -0.14 -0.25 0.19
-0.099 0.053 -0.17 -0.14 -0.42 -0.21 1 0.2 0.0002 -0.17
-0.3 -0.19 -0.35 -0.39 -0.4 -0.14 0.2 1 -0.23 -0.16
0.047 0.12 -0.19 -0.14 0.1 -0.25 0.0002 -0.23 1 -0.2
-0.13 0.077 -0.24 -0.23 -0.076 0.19 -0.17 -0.16 -0.2 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Source_RUH
Source_SVO
Destination_SVO
Total stops
Destination_PAR
Duration
Destination_RUH
Source_PAR
Average Price
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Steps of Modeling

- Splitting the data
- Defining a function to get metrics for val set
- Training and testing of many models

Result of models

```
                    LR
Train score 0.8040357223322144
 Val score 0.7891035984538433
    MAE: 225.09235539537684
    MSE: 152995.68380136567
     RMSE: 391.146626984518
-------------------------------------------------
         Polynomial - Degree 1
Train score -0.3079480631703013
 Val score -0.3153015545186666
     MAE: 756.619594657905
     MSE: 954191.0590377726
    RMSE: 976.8270363978326
-------------------------------------------------
         Polynomial - Degree 2
Train score -6.619866701696714
 Val score -6.708250742830492
    MAE: 1819.2993972374343
     MSE: 5591983.005236949
    RMSE: 2364.737407247779
-------------------------------------------------
         Polynomial - Degree 3
Train score -6.396139020230291
 Val score -6.597515038614633
    MAE: 1978.6373777102278
     MSE: 5511649.321667562
   RMSE: 2347.6902099015456
-------------------------------------------------
         Polynomial - Degree 4
Train score -67.38613658531676
 Val score -65.94440159195507
     MAE: 5158.441261042131
     MSE: 48565098.4234209
   RMSE: 6968.8663657312945
-------------------------------------------------
         Polynomial - Degree 5
Train score -1.2998648856939954
 Val score -1.3170926486628711
     MAE: 977.4048403689628
```

```
MSE: 1680944.632598231
RMSE: 1296.512488408126
-------------------------------------------------
Lasso
Train score 0.8040183311451377
Val score 0.7891286832351017
MAE: 224.60280304137635
MSE: 152977.48594102522
RMSE: 391.12336409504513
-------------------------------------------------
Ridge
Train score 0.8040357214700532
Val score 0.7891039650109529
MAE: 225.0937616466228
MSE: 152995.41788096476
RMSE: 391.14628706018004
-------------------------------------------------
ElasticNet
Train score 0.7354476290670592
Val score 0.7243019152118472
MAE: 304.4466319557438
MSE: 200006.3381624778
RMSE: 447.2206817248927
-------------------------------------------------
Random Forest
Train score -1.0693924886375732
Val score -1.0832919574956343
MAE: 885.6332189325535
MSE: 1511332.9353093393
RMSE: 1229.362816791422
```

Final Model Selection

From the above analysis, we can see that the random forest model performed the best with:
Train score 0.9648778537711422
Val score 0.9448134490695079
MAE: 61.717733027545194
MSE: 40035.31608101726
RMSE: 200.0882707232417
So, we'll select it as our model.

Last step: Saving the final model.