EDA Flight Price

February 20, 2024

0.1 EDA And Feature Engineering Flight Price Prediction

0.1.1 FEATURES

The various features of the cleaned dataset are explained below: 1) Airline: The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines. 2) Flight: Flight stores information regarding the plane's flight code. It is a categorical feature. 3) Source City: City from which the flight takes off. It is a categorical feature having 6 unique cities. 4) Departure Time: This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels. 5) Stops: A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities. 6) Arrival Time: This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time. 7) Destination City: City where the flight will land. It is a categorical feature having 6 unique cities. 8) Class: A categorical feature that contains information on seat class; it has two distinct values: Business and Economy. 9) Duration: A continuous feature that displays the overall amount of time it takes to travel between cities in hours. 10) Days Left: This is a derived characteristic that is calculated by subtracting the trip date by the booking date. 11) Price: Target variable stores information of the ticket price.

[1]:
```
#importing basics libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[2]:
```
df=pd.read_excel('flight_price.xlsx')
df.head()
```

[2]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR						
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK						
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR						
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL						

```
1    05:50         13:15   7h 25m     2 stops         No info   7662
2    09:25  04:25 10 Jun      19h     2 stops         No info  13882
3    18:05         23:30   5h 25m      1 stop         No info   6218
4    16:50         21:35   4h 45m      1 stop         No info  13302
```

[3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          10683 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10683 non-null  object
 3   Destination      10683 non-null  object
 4   Route            10682 non-null  object
 5   Dep_Time         10683 non-null  object
 6   Arrival_Time     10683 non-null  object
 7   Duration         10683 non-null  object
 8   Total_Stops      10682 non-null  object
 9   Additional_Info  10683 non-null  object
 10  Price            10683 non-null  int64
dtypes: int64(1), object(10)
memory usage: 918.2+ KB
```

[4]:
```
df.describe()
```

[4]:
```
              Price
count  10683.000000
mean    9087.064121
std     4611.359167
min     1759.000000
25%     5277.000000
50%     8372.000000
75%    12373.000000
max    79512.000000
```

[5]:
```
df.head(2)
```

[5]:
```
      Airline Date_of_Journey    Source Destination                  Route  \
0     IndiGo      24/03/2019  Banglore   New Delhi              BLR → DEL
1  Air India       1/05/2019   Kolkata    Banglore  CCU → IXR → BBI → BLR

  Dep_Time  Arrival_Time Duration Total_Stops Additional_Info  Price
0    22:20  01:10 22 Mar   2h 50m    non-stop         No info   3897
1    05:50         13:15   7h 25m     2 stops         No info   7662
```

[9]:

[9]: str

[10]:
```
df['Date']=df['Date_of_Journey'].str.split('/').str[0]
df['Month']=df['Date_of_Journey'].str.split('/').str[1]
df['Year']=df['Date_of_Journey'].str.split('/').str[2]
```

[12]:
```
df.head(2)
```

[12]:
```
      Airline Date_of_Journey    Source Destination                  Route  \
0     IndiGo       24/03/2019  Banglore   New Delhi              BLR → DEL
1  Air India        1/05/2019   Kolkata    Banglore  CCU → IXR → BBI → BLR

  Dep_Time  Arrival_Time Duration Total_Stops Additional_Info  Price Date  \
0    22:20  01:10 22 Mar   2h 50m    non-stop         No info   3897   24
1    05:50         13:15   7h 25m     2 stops         No info   7662    1

  Month  Year
0    03  2019
1    05  2019
```

[13]:
```
df['Date']=df['Date'].astype(int)
df['Month']=df['Month'].astype(int)
df['Year']=df['Year'].astype(int)
```

[16]:
```
df.drop('Date_of_Journey',axis=1,inplace=True)
```

[18]:
```
df.head(10)
```

[18]:
```
              Airline    Source Destination                  Route Dep_Time  \
0             IndiGo  Banglore   New Delhi              BLR → DEL    22:20
1          Air India   Kolkata    Banglore  CCU → IXR → BBI → BLR    05:50
2        Jet Airways     Delhi      Cochin  DEL → LKO → BOM → COK    09:25
3             IndiGo   Kolkata    Banglore        CCU → NAG → BLR    18:05
4             IndiGo  Banglore   New Delhi        BLR → NAG → DEL    16:50
5          SpiceJet   Kolkata    Banglore              CCU → BLR    09:00
6        Jet Airways  Banglore   New Delhi        BLR → BOM → DEL    18:55
7        Jet Airways  Banglore   New Delhi        BLR → BOM → DEL    08:00
8        Jet Airways  Banglore   New Delhi        BLR → BOM → DEL    08:55
9  Multiple carriers     Delhi      Cochin        DEL → BOM → COK    11:25

   Arrival_Time Duration Total_Stops             Additional_Info  Price  \
0  01:10 22 Mar   2h 50m    non-stop                    No info   3897
1         13:15   7h 25m     2 stops                    No info   7662
2  04:25 10 Jun      19h     2 stops                    No info  13882
3         23:30   5h 25m      1 stop                    No info   6218
```

```
4          21:35   4h 45m      1 stop                         No info  13302
5          11:25   2h 25m    non-stop                         No info   3873
6   10:25 13 Mar  15h 30m      1 stop  In-flight meal not included  11087
7   05:05 02 Mar   21h 5m      1 stop                         No info  22270
8   10:25 13 Mar  25h 30m      1 stop  In-flight meal not included  11087
9          19:15   7h 50m      1 stop                         No info   8625

   Date  Month  Year
0    24      3  2019
1     1      5  2019
2     9      6  2019
3    12      5  2019
4     1      3  2019
5    24      6  2019
6    12      3  2019
7     1      3  2019
8    12      3  2019
9    27      5  2019
```

[19]:
```
df['Arrival_Time'].str.split(' ').str[0]
```

[19]:
```
0        01:10
1        13:15
2        04:25
3        23:30
4        21:35
         …
10678    22:25
10679    23:20
10680    11:20
10681    14:10
10682    19:15
Name: Arrival_Time, Length: 10683, dtype: object
```

[22]:
```
df['Arrival_hours']=df['Arrival_Time'].str.split(' ').str[0].str.split(':').
  ↪str[0]
df['Arrival_min']=df['Arrival_Time'].str.split(' ').str[0].str.split(':').str[1]
df.drop('Arrival_Time',axis=1,inplace=True)
```

[23]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          10683 non-null  object
```

```
 1   Source           10683 non-null  object
 2   Destination      10683 non-null  object
 3   Route            10682 non-null  object
 4   Dep_Time         10683 non-null  object
 5   Duration         10683 non-null  object
 6   Total_Stops      10682 non-null  object
 7   Additional_Info  10683 non-null  object
 8   Price            10683 non-null  int64
 9   Date             10683 non-null  int64
 10  Month            10683 non-null  int64
 11  Year             10683 non-null  int64
 12  Arrival_hours    10683 non-null  object
 13  Arrival_min      10683 non-null  object
dtypes: int64(4), object(10)
memory usage: 1.1+ MB
```

[24]:
```
df['Arrival_hours']=df['Arrival_hours'].astype(int)
df['Arrival_min']=df['Arrival_min'].astype(int)
```

[25]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          10683 non-null  object
 1   Source           10683 non-null  object
 2   Destination      10683 non-null  object
 3   Route            10682 non-null  object
 4   Dep_Time         10683 non-null  object
 5   Duration         10683 non-null  object
 6   Total_Stops      10682 non-null  object
 7   Additional_Info  10683 non-null  object
 8   Price            10683 non-null  int64
 9   Date             10683 non-null  int64
 10  Month            10683 non-null  int64
 11  Year             10683 non-null  int64
 12  Arrival_hours    10683 non-null  int64
 13  Arrival_min      10683 non-null  int64
dtypes: int64(6), object(8)
memory usage: 1.1+ MB
```

[26]:
```
df.head(3)
```

[26]:

	Airline	Source	Destination	Route	Dep_Time	Duration	\
0	IndiGo	Banglore	New Delhi	BLR → DEL	22:20	2h 50m	

```
1    Air India   Kolkata    Banglore  CCU → IXR → BBI → BLR    05:50   7h 25m
2  Jet Airways     Delhi      Cochin  DEL → LKO → BOM → COK    09:25      19h

  Total_Stops Additional_Info  Price  Date  Month  Year  Arrival_hours  \
0    non-stop         No info   3897    24      3  2019              1
1     2 stops         No info   7662     1      5  2019             13
2     2 stops         No info  13882     9      6  2019              4

   Arrival_min
0           10
1           15
2           25
```

[27]:
```
df['Dept_hour']=df['Dep_Time'].str.split(':').str[0]
df['Dept_min']=df['Dep_Time'].str.split(':').str[1]
df['Dept_hour']=df['Dept_hour'].astype(int)
df['Dept_min']=df['Dept_min'].astype(int)
df.drop('Dep_Time',axis=1,inplace=True)
```

[28]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Airline          10683 non-null  object
 1   Source           10683 non-null  object
 2   Destination      10683 non-null  object
 3   Route            10682 non-null  object
 4   Duration         10683 non-null  object
 5   Total_Stops      10682 non-null  object
 6   Additional_Info  10683 non-null  object
 7   Price            10683 non-null  int64
 8   Date             10683 non-null  int64
 9   Month            10683 non-null  int64
 10  Year             10683 non-null  int64
 11  Arrival_hours    10683 non-null  int64
 12  Arrival_min      10683 non-null  int64
 13  Dept_hour        10683 non-null  int64
 14  Dept_min         10683 non-null  int64
dtypes: int64(8), object(7)
memory usage: 1.2+ MB
```

[30]:
```
df.drop('Route',axis=1,inplace=True)
```

[31]:
```
df.head(2)
```

[31]:
```
      Airline    Source Destination Duration Total_Stops Additional_Info  \
0      IndiGo  Banglore   New Delhi    2h 50m    non-stop         No info
1   Air India   Kolkata    Banglore   7h 25m     2 stops         No info

   Price  Date  Month  Year  Arrival_hours  Arrival_min  Dept_hour  Dept_min
0   3897    24      3  2019              1           10         22        20
1   7662     1      5  2019             13           15          5        50
```

[35]:
```
df['Duration'].str.split(' ').str[0].str.split('h').str[0]
df['Duration'].str.split(' ').str[1].str.split('h').str[0]
```

[35]:
```
0        50m
1        25m
2        NaN
3        25m
4        45m
        ...
10678    30m
10679    35m
10680    NaN
10681    40m
10682    20m
Name: Duration, Length: 10683, dtype: object
```

[36]:
```
df['Total_Stops'].unique()
```

[36]:
```
array(['non-stop', '2 stops', '1 stop', '3 stops', nan, '4 stops'],
      dtype=object)
```

[37]:
```
df['Total_Stops'].mode()
```

[37]:
```
0    1 stop
Name: Total_Stops, dtype: object
```

[39]:
```
df['Total_Stops']=df['Total_Stops'].map({'non-stop':0,'1 stop':1,'2 stops':2,'3␣
↪stops':3,'4 stops':4,np.nan:1})
```

[41]:
```
df['Total_Stops'].isnull().sum()
```

[41]:
```
0
```

[42]:
```
df.head()
```

[42]:
```
       Airline    Source Destination Duration  Total_Stops Additional_Info  \
0       IndiGo  Banglore   New Delhi    2h 50m            0         No info
1    Air India   Kolkata    Banglore   7h 25m            2         No info
2  Jet Airways     Delhi      Cochin      19h            2         No info
```

```
3       IndiGo   Kolkata    Banglore   5h 25m            1         No info
4       IndiGo  Banglore   New Delhi   4h 45m            1         No info

   Price  Date  Month  Year  Arrival_hours  Arrival_min  Dept_hour  Dept_min
0   3897    24      3  2019              1           10         22        20
1   7662     1      5  2019             13           15          5        50
2  13882     9      6  2019              4           25          9        25
3   6218    12      5  2019             23           30         18         5
4  13302     1      3  2019             21           35         16        50
```

[43]:
```
df['Airline'].unique()
```

[43]:
```
array(['IndiGo', 'Air India', 'Jet Airways', 'SpiceJet',
       'Multiple carriers', 'GoAir', 'Vistara', 'Air Asia',
       'Vistara Premium economy', 'Jet Airways Business',
       'Multiple carriers Premium economy', 'Trujet'], dtype=object)
```

[44]:
```
df['Source'].unique()
```

[44]:
```
array(['Banglore', 'Kolkata', 'Delhi', 'Chennai', 'Mumbai'], dtype=object)
```

[45]:
```
df['Destination'].unique()
```

[45]:
```
array(['New Delhi', 'Banglore', 'Cochin', 'Kolkata', 'Delhi', 'Hyderabad'],
      dtype=object)
```

[46]:
```
from sklearn.preprocessing import OneHotEncoder
```

[47]:
```
encoder=OneHotEncoder()
```

[49]:
```
encoder.fit_transform(df[['Airline','Source','Destination']]).toarray()
```

[49]:
```
array([[0., 0., 0., …, 0., 0., 1.],
       [0., 1., 0., …, 0., 0., 0.],
       [0., 0., 0., …, 0., 0., 0.],
       …,
       [0., 0., 0., …, 0., 0., 0.],
       [0., 0., 0., …, 0., 0., 1.],
       [0., 1., 0., …, 0., 0., 0.]])
```

[50]:
```
pd.DataFrame(encoder.fit_transform(df[['Airline','Source','Destination']]).
  ↪toarray(),columns=encoder.get_feature_names_out())
```

[50]:
```
   Airline_Air Asia  Airline_Air India  Airline_GoAir  Airline_IndiGo  \
0               0.0                0.0            0.0             1.0
1               0.0                1.0            0.0             0.0
2               0.0                0.0            0.0             0.0
```

```
3                    0.0                  0.0             0.0              1.0
4                    0.0                  0.0             0.0              1.0
…                    …                    …               …                …
10678                1.0                  0.0             0.0              0.0
10679                0.0                  1.0             0.0              0.0
10680                0.0                  0.0             0.0              0.0
10681                0.0                  0.0             0.0              0.0
10682                0.0                  1.0             0.0              0.0

       Airline_Jet Airways  Airline_Jet Airways Business  \
0                      0.0                           0.0
1                      0.0                           0.0
2                      1.0                           0.0
3                      0.0                           0.0
4                      0.0                           0.0
…                      …                             …
10678                  0.0                           0.0
10679                  0.0                           0.0
10680                  1.0                           0.0
10681                  0.0                           0.0
10682                  0.0                           0.0

       Airline_Multiple carriers  Airline_Multiple carriers Premium economy  \
0                            0.0                                        0.0
1                            0.0                                        0.0
2                            0.0                                        0.0
3                            0.0                                        0.0
4                            0.0                                        0.0
…                            …                                          …
10678                        0.0                                        0.0
10679                        0.0                                        0.0
10680                        0.0                                        0.0
10681                        0.0                                        0.0
10682                        0.0                                        0.0

       Airline_SpiceJet  Airline_Trujet  …  Source_Chennai  Source_Delhi  \
0                   0.0             0.0  …             0.0           0.0
1                   0.0             0.0  …             0.0           0.0
2                   0.0             0.0  …             0.0           1.0
3                   0.0             0.0  …             0.0           0.0
4                   0.0             0.0  …             0.0           0.0
…                   …               …    …             …             …
10678               0.0             0.0  …             0.0           0.0
10679               0.0             0.0  …             0.0           0.0
10680               0.0             0.0  …             0.0           0.0
10681               0.0             0.0  …             0.0           0.0
10682               0.0             0.0  …             0.0           1.0
```

```
       Source_Kolkata  Source_Mumbai  Destination_Banglore  \
0                 0.0            0.0                   0.0
1                 1.0            0.0                   1.0
2                 0.0            0.0                   0.0
3                 1.0            0.0                   1.0
4                 0.0            0.0                   0.0
…                 …              …                     …
10678             1.0            0.0                   1.0
10679             1.0            0.0                   1.0
10680             0.0            0.0                   0.0
10681             0.0            0.0                   0.0
10682             0.0            0.0                   0.0

       Destination_Cochin  Destination_Delhi  Destination_Hyderabad  \
0                     0.0                0.0                    0.0
1                     0.0                0.0                    0.0
2                     1.0                0.0                    0.0
3                     0.0                0.0                    0.0
4                     0.0                0.0                    0.0
…                     …                  …                      …
10678                 0.0                0.0                    0.0
10679                 0.0                0.0                    0.0
10680                 0.0                1.0                    0.0
10681                 0.0                0.0                    0.0
10682                 1.0                0.0                    0.0

       Destination_Kolkata  Destination_New Delhi
0                      0.0                    1.0
1                      0.0                    0.0
2                      0.0                    0.0
3                      0.0                    0.0
4                      0.0                    1.0
…                      …                      …
10678                  0.0                    0.0
10679                  0.0                    0.0
10680                  0.0                    0.0
10681                  0.0                    1.0
10682                  0.0                    0.0

[10683 rows x 23 columns]
```

[ ]: