```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,
roc_auc_score, roc_curve,classification_report
import matplotlib.pyplot as plt
import seaborn as sns

# Supress Warnings
import warnings
warnings.filterwarnings('ignore')

# Reading the CSV file into the dataframe
df = pd.read_csv('Leads.csv')

df
```

```
                                  Prospect ID  Lead Number  \
0     7927b2df-8bba-4d29-b9a2-b6e0beafe620       660737
1     2a272436-5132-4136-86fa-dcc88c88f482       660728
2     8cc8c611-a219-4f35-ad23-fdfd2656bd8a       660727
3     0cc2df48-7cf4-4e39-9de9-19797f9b38cc       660719
4     3256f628-e534-4826-9d63-4a8b88782852       660681
...                                        ...          ...
9235  19d6451e-fcd6-407c-b83b-48e1af805ea9       579564
9236  82a7005b-7196-4d56-95ce-a79f937a158d       579546
9237  aac550fe-a586-452d-8d3c-f1b62c94e02c       579545
9238  5330a7d1-2f2b-4df4-85d6-64ca2f6b95b9       579538
9239  571b5c8e-a5b2-4d57-8574-f2ffb06fdeff       579533

                   Lead Origin     Lead Source Do Not Email Do Not Call
\
0                          API      Olark Chat           No           No

1                          API  Organic Search           No           No

2     Landing Page Submission  Direct Traffic           No           No

3     Landing Page Submission  Direct Traffic           No           No

4     Landing Page Submission          Google           No           No

...                       ...             ...          ...          ...

9235  Landing Page Submission  Direct Traffic          Yes           No

9236  Landing Page Submission  Direct Traffic           No           No

9237  Landing Page Submission  Direct Traffic          Yes           No
```

```
9238  Landing Page Submission          Google            No           No

9239  Landing Page Submission  Direct Traffic            No           No


      Converted  TotalVisits  Total Time Spent on Website  \
0             0          0.0                            0
1             0          5.0                          674
2             1          2.0                         1532
3             0          1.0                          305
4             1          2.0                         1428
...         ...          ...                          ...
9235          1          8.0                         1845
9236          0          2.0                          238
9237          0          2.0                          199
9238          1          3.0                          499
9239          1          6.0                         1279

      Page Views Per Visit  ... Get updates on DM Content    Lead
Profile  \
0                     0.00  ...                          No
Select
1                     2.50  ...                          No
Select
2                     2.00  ...                          No  Potential
Lead
3                     1.00  ...                          No
Select
4                     1.00  ...                          No
Select
...                    ...  ...                         ...
...
9235                  2.67  ...                          No  Potential
Lead
9236                  2.00  ...                          No  Potential
Lead
9237                  2.00  ...                          No  Potential
Lead
9238                  3.00  ...                          No
NaN
9239                  3.00  ...                          No  Potential
Lead

                     City Asymmetrique Activity Index  \
0                  Select                   02.Medium
1                  Select                   02.Medium
2                  Mumbai                   02.Medium
3                  Mumbai                   02.Medium
4                  Mumbai                   02.Medium
```

```
...                  ...                          ...
9235              Mumbai                   02.Medium
9236              Mumbai                   02.Medium
9237              Mumbai                   02.Medium
9238  Other Metro Cities                   02.Medium
9239        Other Cities                   02.Medium

     Asymmetrique Profile Index Asymmetrique Activity Score  \
0                     02.Medium                         15.0
1                     02.Medium                         15.0
2                       01.High                         14.0
3                       01.High                         13.0
4                       01.High                         15.0
...                         ...                          ...
9235                    01.High                         15.0
9236                    01.High                         14.0
9237                    01.High                         13.0
9238                  02.Medium                         15.0
9239                    01.High                         15.0

     Asymmetrique Profile Score I agree to pay the amount through
cheque  \
0                           15.0
No
1                           15.0
No
2                           20.0
No
3                           17.0
No
4                           18.0
No
...                          ...                                    .
..
9235                        17.0
No
9236                        19.0
No
9237                        20.0
No
9238                        16.0
No
9239                        18.0
No

     A free copy of Mastering The Interview Last Notable Activity
0                                        No              Modified
1                                        No          Email Opened
2                                       Yes          Email Opened
3                                        No              Modified
```

```
4                                          No              Modified
...                                       ...                   ...
9235                                       No     Email Marked Spam
9236                                      Yes              SMS Sent
9237                                      Yes              SMS Sent
9238                                       No              SMS Sent
9239                                      Yes              Modified

[9240 rows x 37 columns]
```

```
df.columns
```

```
Index(['Prospect ID', 'Lead Number', 'Lead Origin', 'Lead Source',
       'Do Not Email', 'Do Not Call', 'Converted', 'TotalVisits',
       'Total Time Spent on Website', 'Page Views Per Visit', 'Last
Activity',
       'Country', 'Specialization', 'How did you hear about X
Education',
       'What is your current occupation',
       'What matters most to you in choosing a course', 'Search',
'Magazine',
       'Newspaper Article', 'X Education Forums', 'Newspaper',
       'Digital Advertisement', 'Through Recommendations',
       'Receive More Updates About Our Courses', 'Tags', 'Lead
Quality',
       'Update me on Supply Chain Content', 'Get updates on DM
Content',
       'Lead Profile', 'City', 'Asymmetrique Activity Index',
       'Asymmetrique Profile Index', 'Asymmetrique Activity Score',
       'Asymmetrique Profile Score',
       'I agree to pay the amount through cheque',
       'A free copy of Mastering The Interview', 'Last Notable
Activity'],
      dtype='object')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
-----
 0   Prospect ID                                    9240 non-null
object
 1   Lead Number                                    9240 non-null
int64
 2   Lead Origin                                    9240 non-null
object
```

```
 3   Lead Source                                    9204 non-null
object
 4   Do Not Email                                   9240 non-null
object
 5   Do Not Call                                    9240 non-null
object
 6   Converted                                      9240 non-null
int64
 7   TotalVisits                                    9103 non-null
float64
 8   Total Time Spent on Website                    9240 non-null
int64
 9   Page Views Per Visit                           9103 non-null
float64
 10  Last Activity                                  9137 non-null
object
 11  Country                                        6779 non-null
object
 12  Specialization                                 7802 non-null
object
 13  How did you hear about X Education             7033 non-null
object
 14  What is your current occupation                6550 non-null
object
 15  What matters most to you in choosing a course  6531 non-null
object
 16  Search                                         9240 non-null
object
 17  Magazine                                       9240 non-null
object
 18  Newspaper Article                              9240 non-null
object
 19  X Education Forums                             9240 non-null
object
 20  Newspaper                                      9240 non-null
object
 21  Digital Advertisement                          9240 non-null
object
 22  Through Recommendations                        9240 non-null
object
 23  Receive More Updates About Our Courses         9240 non-null
object
 24  Tags                                           5887 non-null
object
 25  Lead Quality                                   4473 non-null
object
 26  Update me on Supply Chain Content              9240 non-null
object
 27  Get updates on DM Content                      9240 non-null
```

```
object
 28  Lead Profile                                  6531 non-null
object
 29  City                                          7820 non-null
object
 30  Asymmetrique Activity Index                   5022 non-null
object
 31  Asymmetrique Profile Index                    5022 non-null
object
 32  Asymmetrique Activity Score                   5022 non-null
float64
 33  Asymmetrique Profile Score                    5022 non-null
float64
 34  I agree to pay the amount through cheque      9240 non-null
object
 35  A free copy of Mastering The Interview        9240 non-null
object
 36  Last Notable Activity                         9240 non-null
object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Data Imbalance

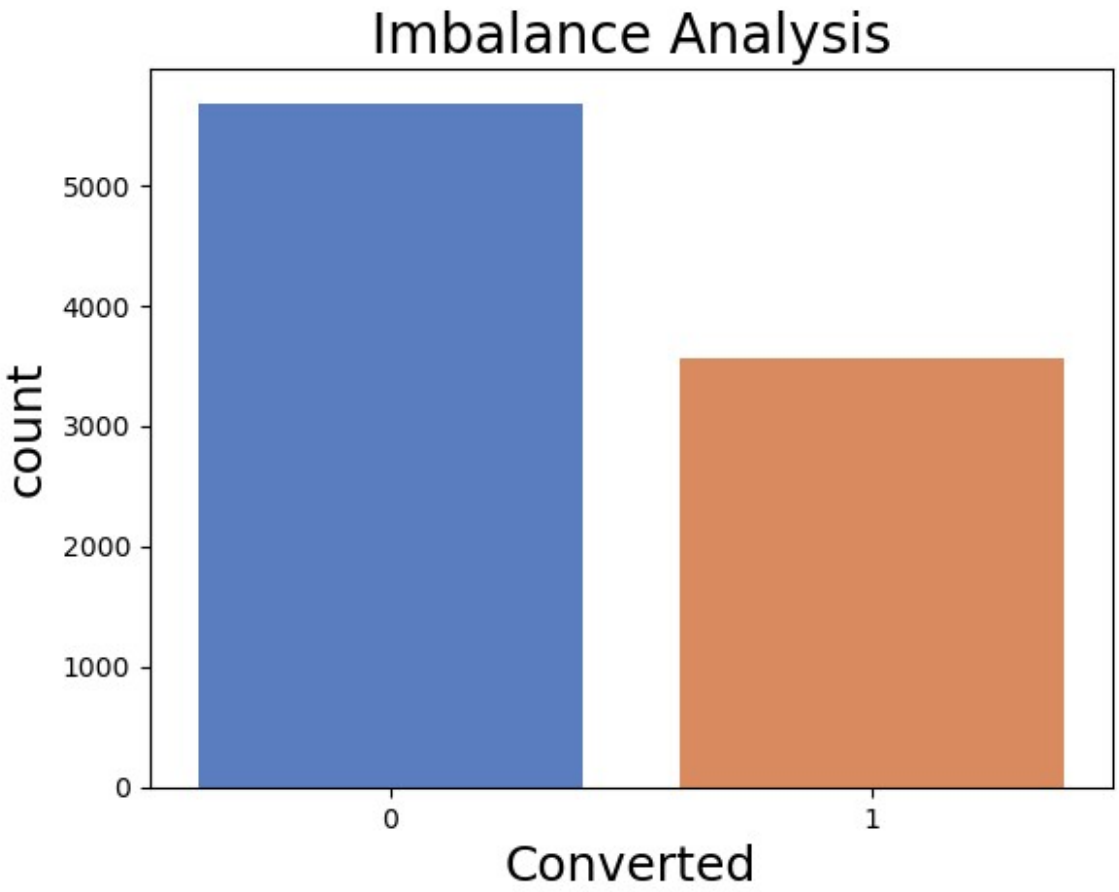
We need to check the balance with respect to the target variable: converted

```
# Calculating Imbalance percentage
# Since the majority is target0 and minority is target1
# zero count
z_cnt = len(df[df['Converted'] == 0])
# one count
o_cnt = len(df[df['Converted'] == 1])
print ('Count of Converted = 0: {0} \nCount of Converted = 1:
{1}'.format(z_cnt,o_cnt))
print ('Imbalance Ratio is : {0}'.format(round(z_cnt/o_cnt,2)))
```

```
Count of Converted = 0: 5679
Count of Converted = 1: 3561
Imbalance Ratio is : 1.59
```

```
## Plotting the imbalance Analysis:
plt.title('Imbalance Analysis', fontsize=20)
sns.countplot(data = df, x='Converted', palette='muted')
plt.xlabel('Converted', fontsize=18)
plt.ylabel('count', fontsize=18)
```

```
Text(0, 0.5, 'count')
```



The data is not too much imbalanced. As such, we can proceed with the data for analysis and model building

```
df.isnull().sum()
```

```
Prospect ID                                       0
Lead Number                                       0
Lead Origin                                       0
Lead Source                                      36
Do Not Email                                      0
Do Not Call                                       0
Converted                                         0
TotalVisits                                     137
Total Time Spent on Website                       0
Page Views Per Visit                            137
Last Activity                                   103
Country                                        2461
Specialization                                 1438
How did you hear about X Education             2207
What is your current occupation                2690
What matters most to you in choosing a course  2709
Search                                            0
```

```
Magazine                                          0
Newspaper Article                                 0
X Education Forums                                0
Newspaper                                         0
Digital Advertisement                             0
Through Recommendations                           0
Receive More Updates About Our Courses            0
Tags                                           3353
Lead Quality                                   4767
Update me on Supply Chain Content                 0
Get updates on DM Content                         0
Lead Profile                                   2709
City                                           1420
Asymmetrique Activity Index                    4218
Asymmetrique Profile Index                     4218
Asymmetrique Activity Score                    4218
Asymmetrique Profile Score                     4218
I agree to pay the amount through cheque          0
A free copy of Mastering The Interview            0
Last Notable Activity                             0
dtype: int64
```

```
round(100*(df.isnull().sum()/len(df.index)), 2)
```

```
Prospect ID                                    0.00
Lead Number                                    0.00
Lead Origin                                    0.00
Lead Source                                    0.39
Do Not Email                                   0.00
Do Not Call                                    0.00
Converted                                      0.00
TotalVisits                                    1.48
Total Time Spent on Website                    0.00
Page Views Per Visit                           1.48
Last Activity                                  1.11
Country                                       26.63
Specialization                                15.56
How did you hear about X Education            23.89
What is your current occupation               29.11
What matters most to you in choosing a course 29.32
Search                                         0.00
Magazine                                       0.00
Newspaper Article                              0.00
X Education Forums                             0.00
Newspaper                                      0.00
Digital Advertisement                          0.00
Through Recommendations                        0.00
Receive More Updates About Our Courses         0.00
Tags                                          36.29
Lead Quality                                  51.59
```

```
Update me on Supply Chain Content                 0.00
Get updates on DM Content                         0.00
Lead Profile                                     29.32
City                                             15.37
Asymmetrique Activity Index                      45.65
Asymmetrique Profile Index                       45.65
Asymmetrique Activity Score                      45.65
Asymmetrique Profile Score                       45.65
I agree to pay the amount through cheque          0.00
A free copy of Mastering The Interview            0.00
Last Notable Activity                             0.00
dtype: float64
```

```
#removing the columns which are not contributing towards leads
conversion and with null values above 40%

df.drop(['Asymmetrique Activity Index', 'Asymmetrique Profile Index',
            'Asymmetrique Activity Score', 'Asymmetrique Profile
Score',
            'Tags', 'Lead Quality'], axis=1, inplace=True)

df.duplicated().sum()

0
```

There is no duplicate in the data

```
df.describe(include = 'all')
```

```
                                  Prospect ID    Lead Number  \
count                                    9240    9240.000000
unique                                   9240            NaN
top     7927b2df-8bba-4d29-b9a2-b6e0beafe620            NaN
freq                                        1            NaN
mean                                      NaN  617188.435606
std                                       NaN   23405.995698
min                                       NaN  579533.000000
25%                                       NaN  596484.500000
50%                                       NaN  615479.000000
75%                                       NaN  637387.250000
max                                       NaN  660737.000000

                     Lead Origin Lead Source Do Not Email Do Not
Call  \
count                        9240        9204         9240        9240

unique                          5          21            2           2

top     Landing Page Submission      Google           No          No

freq                         4886        2868         8506        9238
```

```
mean                          NaN         NaN          NaN          NaN

std                           NaN         NaN          NaN          NaN

min                           NaN         NaN          NaN          NaN

25%                           NaN         NaN          NaN          NaN

50%                           NaN         NaN          NaN          NaN

75%                           NaN         NaN          NaN          NaN

max                           NaN         NaN          NaN          NaN


          Converted  TotalVisits  Total Time Spent on Website  \
count   9240.000000  9103.000000                   9240.000000
unique          NaN          NaN                           NaN
top             NaN          NaN                           NaN
freq            NaN          NaN                           NaN
mean       0.385390     3.445238                    487.698268
std        0.486714     4.854853                    548.021466
min        0.000000     0.000000                      0.000000
25%        0.000000     1.000000                     12.000000
50%        0.000000     3.000000                    248.000000
75%        1.000000     5.000000                    936.000000
max        1.000000   251.000000                   2272.000000

        Page Views Per Visit  ... Digital Advertisement  \
count             9103.000000  ...                  9240
unique                    NaN  ...                     2
top                       NaN  ...                    No
freq                      NaN  ...                  9236
mean                 2.362820  ...                   NaN
std                  2.161418  ...                   NaN
min                  0.000000  ...                   NaN
25%                  1.000000  ...                   NaN
50%                  2.000000  ...                   NaN
75%                  3.000000  ...                   NaN
max                 55.000000  ...                   NaN

       Through Recommendations Receive More Updates About Our Courses
\
count                      9240                                   9240

unique                        2                                      1

top                          No                                     No

freq                       9233                                   9240
```

```
mean                          NaN                                    NaN

std                           NaN                                    NaN

min                           NaN                                    NaN

25%                           NaN                                    NaN

50%                           NaN                                    NaN

75%                           NaN                                    NaN

max                           NaN                                    NaN


       Update me on Supply Chain Content Get updates on DM Content  \
count                               9240                        9240
unique                                 1                           1
top                                   No                          No
freq                                9240                        9240
mean                                 NaN                         NaN
std                                  NaN                         NaN
min                                  NaN                         NaN
25%                                  NaN                         NaN
50%                                  NaN                         NaN
75%                                  NaN                         NaN
max                                  NaN                         NaN

       Lead Profile    City I agree to pay the amount through
cheque  \
count          6531    7820                                 9240

unique            6       7                                    1

top          Select  Mumbai                                   No

freq           4146    3222                                 9240

mean            NaN     NaN                                  NaN

std             NaN     NaN                                  NaN

min             NaN     NaN                                  NaN

25%             NaN     NaN                                  NaN

50%             NaN     NaN                                  NaN

75%             NaN     NaN                                  NaN
```

```
max              NaN     NaN                                          NaN


       A free copy of Mastering The Interview Last Notable Activity
count                                    9240                  9240
unique                                      2                    16
top                                        No              Modified
freq                                     6352                  3407
mean                                      NaN                   NaN
std                                       NaN                   NaN
min                                       NaN                   NaN
25%                                       NaN                   NaN
50%                                       NaN                   NaN
75%                                       NaN                   NaN
max                                       NaN                   NaN

[11 rows x 31 columns]
```

```
# Prospect ID   Lead Number do not contribute in the lead generation
therefore dropping them as well

df.drop(['Lead Number','Prospect ID'],axis=1,inplace = True)

df.head()
```

```
                Lead Origin     Lead Source Do Not Email Do Not Call  \
0                       API      Olark Chat           No          No
1                       API  Organic Search           No          No
2  Landing Page Submission  Direct Traffic           No          No
3  Landing Page Submission  Direct Traffic           No          No
4  Landing Page Submission          Google           No          No

   Converted  TotalVisits  Total Time Spent on Website  Page Views Per
Visit  \
0          0          0.0                            0
0.0
1          0          5.0                          674
2.5
2          1          2.0                         1532
2.0
3          0          1.0                          305
1.0
4          1          2.0                         1428
1.0

             Last Activity Country  ... Digital Advertisement  \
0  Page Visited on Website     NaN  ...                    No
1             Email Opened   India  ...                    No
2             Email Opened   India  ...                    No
3              Unreachable   India  ...                    No
4        Converted to Lead   India  ...                    No
```

```
  Through Recommendations Receive More Updates About Our Courses  \
0                         No                                    No
1                         No                                    No
2                         No                                    No
3                         No                                    No
4                         No                                    No

  Update me on Supply Chain Content Get updates on DM Content    Lead
Profile  \
0                                   No                         No
Select
1                                   No                         No
Select
2                                   No                         No
Potential Lead
3                                   No                         No
Select
4                                   No                         No
Select

     City I agree to pay the amount through cheque  \
0  Select                                         No
1  Select                                         No
2  Mumbai                                         No
3  Mumbai                                         No
4  Mumbai                                         No

  A free copy of Mastering The Interview Last Notable Activity
0                                     No              Modified
1                                     No          Email Opened
2                                    Yes          Email Opened
3                                     No              Modified
4                                     No              Modified

[5 rows x 29 columns]
```

Data Cleaning

```
# As we can observe that there are select values for many column.
# This is because customer did not select any option from the list,
hence it shows select.
# Select values are as good as NULL.

## Converting 'Select' values to NaN.
df = df.replace('Select', np.nan)
df.head()
```

```
               Lead Origin     Lead Source Do Not Email Do Not Call  \
0                      API      Olark Chat           No          No
```

```
1                      API  Organic Search           No           No
2  Landing Page Submission  Direct Traffic           No           No
3  Landing Page Submission  Direct Traffic           No           No
4  Landing Page Submission          Google           No           No

   Converted  TotalVisits  Total Time Spent on Website  Page Views Per
Visit  \
0          0          0.0                            0
0.0
1          0          5.0                          674
2.5
2          1          2.0                         1532
2.0
3          0          1.0                          305
1.0
4          1          2.0                         1428
1.0

              Last Activity Country  ... Digital Advertisement  \
0  Page Visited on Website     NaN  ...                     No
1             Email Opened   India  ...                     No
2             Email Opened   India  ...                     No
3              Unreachable   India  ...                     No
4        Converted to Lead   India  ...                     No

  Through Recommendations Receive More Updates About Our Courses  \
0                      No                                     No
1                      No                                     No
2                      No                                     No
3                      No                                     No
4                      No                                     No

  Update me on Supply Chain Content Get updates on DM Content    Lead
Profile  \
0                                 No                         No
NaN
1                                 No                         No
NaN
2                                 No                         No
Potential Lead
3                                 No                         No
NaN
4                                 No                         No
NaN

     City I agree to pay the amount through cheque  \
0     NaN                                       No
1     NaN                                       No
2  Mumbai                                       No
3  Mumbai                                       No
```

```
4  Mumbai                                              No

  A free copy of Mastering The Interview Last Notable Activity
0                                     No              Modified
1                                     No         Email Opened
2                                    Yes         Email Opened
3                                     No              Modified
4                                     No              Modified

[5 rows x 29 columns]
```

```
df.isnull().sum()
```

```
Lead Origin                                       0
Lead Source                                      36
Do Not Email                                      0
Do Not Call                                       0
Converted                                         0
TotalVisits                                     137
Total Time Spent on Website                       0
Page Views Per Visit                            137
Last Activity                                   103
Country                                        2461
Specialization                                 3380
How did you hear about X Education             7250
What is your current occupation                2690
What matters most to you in choosing a course  2709
Search                                            0
Magazine                                          0
Newspaper Article                                 0
X Education Forums                                0
Newspaper                                         0
Digital Advertisement                             0
Through Recommendations                           0
Receive More Updates About Our Courses            0
Update me on Supply Chain Content                 0
Get updates on DM Content                         0
Lead Profile                                   6855
City                                           3669
I agree to pay the amount through cheque          0
A free copy of Mastering The Interview            0
Last Notable Activity                             0
dtype: int64
```

```
round(100*(df.isnull().sum()/len(df.index)), 2)
```

```
Lead Origin                                       0.00
Lead Source                                       0.39
Do Not Email                                      0.00
Do Not Call                                       0.00
```

```
Converted                                        0.00
TotalVisits                                      1.48
Total Time Spent on Website                      0.00
Page Views Per Visit                             1.48
Last Activity                                    1.11
Country                                         26.63
Specialization                                  36.58
How did you hear about X Education              78.46
What is your current occupation                 29.11
What matters most to you in choosing a course   29.32
Search                                           0.00
Magazine                                         0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
Receive More Updates About Our Courses           0.00
Update me on Supply Chain Content                0.00
Get updates on DM Content                        0.00
Lead Profile                                    74.19
City                                            39.71
I agree to pay the amount through cheque         0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

```
# since How did you hear about X Education and Lead Profile have null values more than 70% we are dropping these columns
df.drop(['How did you hear about X Education','Lead Profile'],axis=1,inplace =True)

df.isnull().sum()
```

```
Lead Origin                                        0
Lead Source                                       36
Do Not Email                                       0
Do Not Call                                        0
Converted                                          0
TotalVisits                                      137
Total Time Spent on Website                        0
Page Views Per Visit                             137
Last Activity                                    103
Country                                         2461
Specialization                                  3380
What is your current occupation                 2690
What matters most to you in choosing a course   2709
Search                                             0
Magazine                                           0
Newspaper Article                                  0
```

```
X Education Forums                                0
Newspaper                                         0
Digital Advertisement                             0
Through Recommendations                           0
Receive More Updates About Our Courses            0
Update me on Supply Chain Content                 0
Get updates on DM Content                         0
City                                           3669
I agree to pay the amount through cheque          0
A free copy of Mastering The Interview            0
Last Notable Activity                             0
dtype: int64
```

```
# Country >>> Imputing the null values

df['Country'].describe()
```

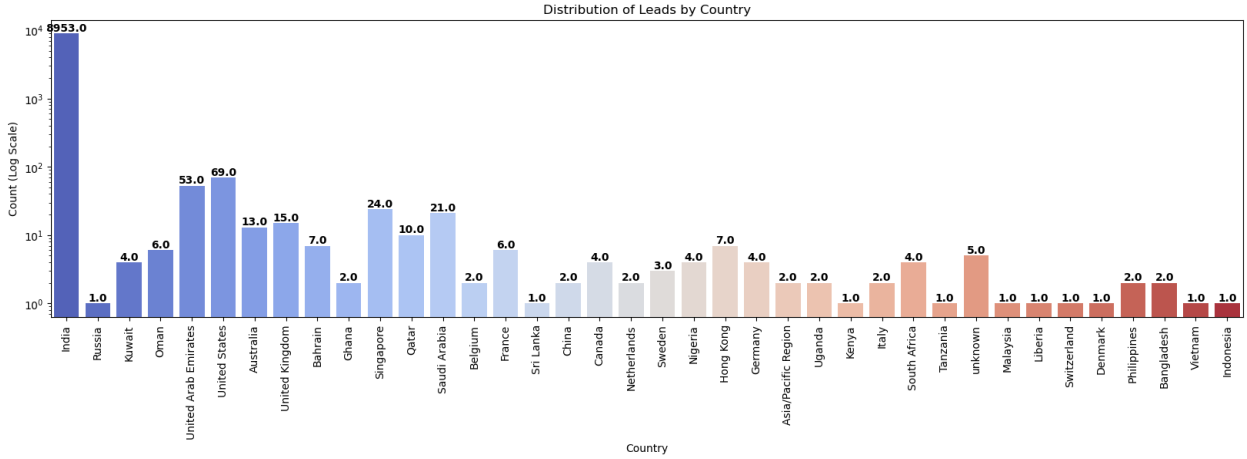
```
count      6779
unique       38
top       India
freq       6492
Name: Country, dtype: object
```

```
df['Country'].value_counts()
```

```
Country
India                   6492
United States             69
United Arab Emirates      53
Singapore                 24
Saudi Arabia              21
United Kingdom            15
Australia                 13
Qatar                     10
Hong Kong                  7
Bahrain                    7
Oman                       6
France                     6
unknown                    5
South Africa               4
Nigeria                    4
Germany                    4
Kuwait                     4
Canada                     4
Sweden                     3
China                      2
Asia/Pacific Region        2
Uganda                     2
Bangladesh                 2
Italy                      2
```

```
Belgium                  2
Netherlands              2
Ghana                    2
Philippines              2
Russia                   1
Switzerland              1
Vietnam                  1
Denmark                  1
Tanzania                 1
Liberia                  1
Malaysia                 1
Kenya                    1
Sri Lanka                1
Indonesia                1
Name: count, dtype: int64
```

```
# Country is India for most values so let's impute the same in missing values.
df['Country'] = df['Country'].replace(np.nan, 'India')

plt.figure(figsize=(20, 5))
ax = sns.countplot(x=df['Country'], palette='coolwarm')  # Applying a different color palette

# Annotating bars with count values
for p in ax.patches:
    ax.annotate(str(p.get_height()),
                (p.get_x() + p.get_width() / 2, p.get_height()),
                ha='center', va='bottom', fontsize=10,
fontweight='bold', color='black')

plt.xticks(rotation=90)
ax.set_yscale('log')  # Log scale for better visualization
plt.xlabel("Country")
plt.ylabel("Count (Log Scale)")
plt.title("Distribution of Leads by Country")
plt.show()
```




```
## Specailization >>> Imputing for nan values in specialization column

df.Specialization.describe()

count                   5860
unique                    18
top       Finance Management
freq                     976
Name: Specialization, dtype: object

df.Specialization.value_counts()

Specialization
Finance Management                     976
Human Resource Management              848
Marketing Management                   838
Operations Management                  503
Business Administration                403
IT Projects Management                 366
Supply Chain Management                349
Banking, Investment And Insurance      338
Travel and Tourism                     203
Media and Advertising                  203
International Business                 178
Healthcare Management                  159
Hospitality Management                 114
E-COMMERCE                             112
Retail Management                      100
Rural and Agribusiness                  73
E-Business                              57
Services Excellence                     40
Name: count, dtype: int64

# It maybe the case that leads has not entered any specialization if
his/her option is not availabe on the list,
```

```
#  may not have any specialization or is a student.
# Hence we can make a category "Others" for missing values.

df['Specialization'] = df['Specialization'].replace(np.nan, 'Others')

df['Specialization'].value_counts()
```

```
Specialization
Others                                3380
Finance Management                     976
Human Resource Management              848
Marketing Management                   838
Operations Management                  503
Business Administration                403
IT Projects Management                 366
Supply Chain Management                349
Banking, Investment And Insurance      338
Travel and Tourism                     203
Media and Advertising                  203
International Business                 178
Healthcare Management                  159
Hospitality Management                 114
E-COMMERCE                             112
Retail Management                      100
Rural and Agribusiness                  73
E-Business                              57
Services Excellence                     40
Name: count, dtype: int64
```

```
## Occupation

df['What is your current occupation'].describe()
```

```
count           6550
unique             6
top       Unemployed
freq            5600
Name: What is your current occupation, dtype: object
```

```
df['What is your current occupation'].value_counts()
```

```
What is your current occupation
Unemployed              5600
Working Professional     706
Student                  210
Other                     16
Housewife                 10
Businessman                8
Name: count, dtype: int64
```

```
# Most of the entries are of Unemployed so we can impute "Unemployed"
in it.

df['What is your current occupation'] = df['What is your current
occupation'].replace(np.nan, 'Unemployed')

# CITY

df.City.value_counts()

City
Mumbai                         3222
Thane & Outskirts               752
Other Cities                    686
Other Cities of Maharashtra     457
Other Metro Cities              380
Tier II Cities                   74
Name: count, dtype: int64

df.City.describe()

count       5571
unique         6
top       Mumbai
freq        3222
Name: City, dtype: object

## Most of the data available  is Mumbai so we can impute Mumbai in
the missing values.

df['City'] = df['City'].replace(np.nan, 'Mumbai')

## What matters most to you in choosing a course

df['What matters most to you in choosing a course'].describe()

count                          6531
unique                            3
top       Better Career Prospects
freq                           6528
Name: What matters most to you in choosing a course, dtype: object

# Blanks in the this column may be imputed by 'Better Career
Prospects'.

df['What matters most to you in choosing a course'] = df['What matters
most to you in choosing a course'].replace(np.nan, 'Better Career
Prospects')

df['What matters most to you in choosing a course'].value_counts()

What matters most to you in choosing a course
Better Career Prospects      9237
```

```
Flexibility & Convenience       2
Other                           1
Name: count, dtype: int64

# Now imputing for numerical column wih median

df['TotalVisits'] =
df['TotalVisits'].fillna(df['TotalVisits'].median())
df['Page Views Per Visit'] = df['Page Views Per
Visit'].fillna(df['Page Views Per Visit'].median())

df['Lead Source'].value_counts()

Lead Source
Google               2868
Direct Traffic       2543
Olark Chat           1755
Organic Search       1154
Reference             534
Welingak Website      142
Referral Sites        125
Facebook               55
bing                    6
google                  5
Click2call              4
Press_Release           2
Social Media            2
Live Chat               2
youtubechannel          1
testone                 1
Pay per Click Ads       1
welearnblog_Home        1
WeLearn                 1
blog                    1
NC_EDM                  1
Name: count, dtype: int64

# Replacing 'google' with 'Google' as they both are same and other
values with 1-1 count with 'others'
df['Lead Source'] = df['Lead Source'].replace('google','Google')
df['Lead Source'] = df['Lead Source'].replace(['Click2call', 'Live
Chat', 'NC_EDM', 'Pay per Click Ads', 'Press_Release',
  'Social Media', 'WeLearn', 'bing', 'blog', 'testone',
'welearnblog_Home', 'youtubechannel'], 'Others')

df['Lead Source'].value_counts()

Lead Source
Google              2873
Direct Traffic      2543
Olark Chat          1755
```

```
Organic Search      1154
Reference            534
Welingak Website     142
Referral Sites       125
Facebook              55
Others                23
Name: count, dtype: int64
```

```
# Imputing Lead sorce nan with Google
df['Lead Source'] = df['Lead Source'].replace(np.nan,'Google')

df['Last Activity'].value_counts()
```

```
Last Activity
Email Opened                    3437
SMS Sent                        2745
Olark Chat Conversation          973
Page Visited on Website          640
Converted to Lead                428
Email Bounced                    326
Email Link Clicked               267
Form Submitted on Website        116
Unreachable                       93
Unsubscribed                      61
Had a Phone Conversation          30
Approached upfront                 9
View in browser link Clicked       6
Email Received                     2
Email Marked Spam                  2
Visited Booth in Tradeshow         1
Resubscribed to emails             1
Name: count, dtype: int64
```

```
# Imputing nan values with Email opened

df['Last Activity'] = df['Last Activity'].replace(np.nan,'Email
Opened')

df['Last Activity'].value_counts()
```

```
Last Activity
Email Opened                    3540
SMS Sent                        2745
Olark Chat Conversation          973
Page Visited on Website          640
Converted to Lead                428
Email Bounced                    326
Email Link Clicked               267
Form Submitted on Website        116
Unreachable                       93
Unsubscribed                      61
```

```
Had a Phone Conversation          30
Approached upfront                9
View in browser link Clicked      6
Email Received                    2
Email Marked Spam                 2
Visited Booth in Tradeshow        1
Resubscribed to emails            1
Name: count, dtype: int64
```

```
df.isna().sum()
```

```
Lead Origin                                      0
Lead Source                                      0
Do Not Email                                     0
Do Not Call                                      0
Converted                                        0
TotalVisits                                      0
Total Time Spent on Website                      0
Page Views Per Visit                             0
Last Activity                                    0
Country                                          0
Specialization                                   0
What is your current occupation                  0
What matters most to you in choosing a course    0
Search                                           0
Magazine                                         0
Newspaper Article                                0
X Education Forums                               0
Newspaper                                        0
Digital Advertisement                            0
Through Recommendations                          0
Receive More Updates About Our Courses           0
Update me on Supply Chain Content                0
Get updates on DM Content                        0
City                                             0
I agree to pay the amount through cheque         0
A free copy of Mastering The Interview           0
Last Notable Activity                            0
dtype: int64
```

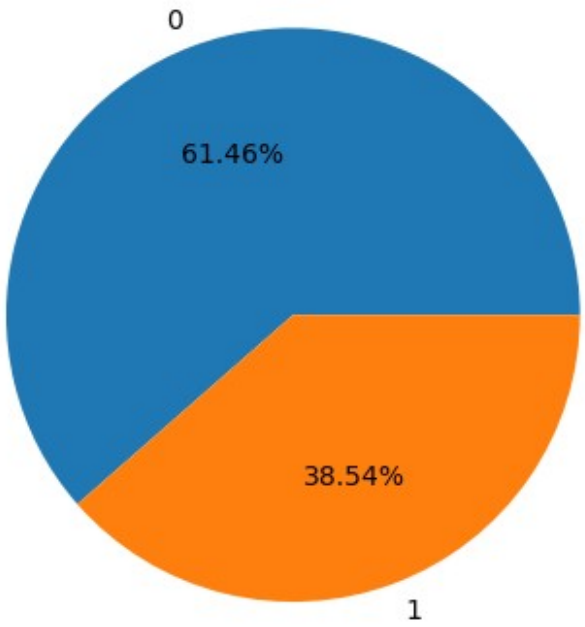
```
# Our data is clean now, we can move further for EDA
```

EDA

Univariate Analysis

```
# Target column
plt.pie(df['Converted'].value_counts(),
labels=df['Converted'].value_counts().index, autopct='%.2f%%')
```

```
([<matplotlib.patches.Wedge at 0x208331c9be0>,
  <matplotlib.patches.Wedge at 0x2083502c380>],
 [Text(-0.3875625077420184, 1.0294635994500816, '0'),
  Text(0.3875624113566783, -1.0294636357362978, '1')],
 [Text(-0.2113977314956464, 0.5615255997000445, '61.46%'),
  Text(0.2113976789218245, -0.5615256194925261, '38.54%')])
```



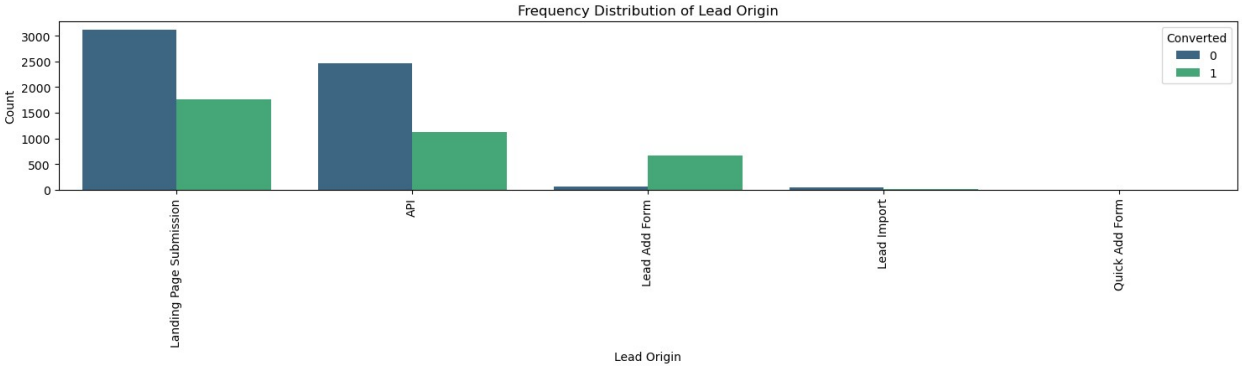
According to the pie chart, there is 38.5% conversation rate of leads

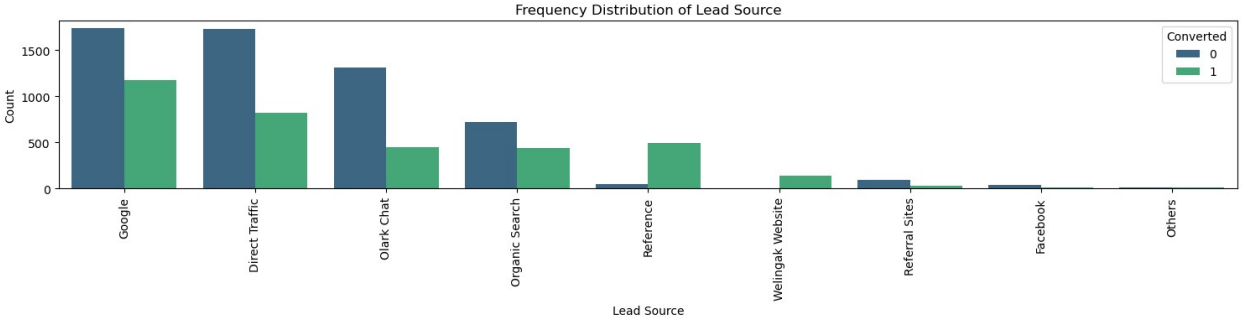
Checking for other features as well

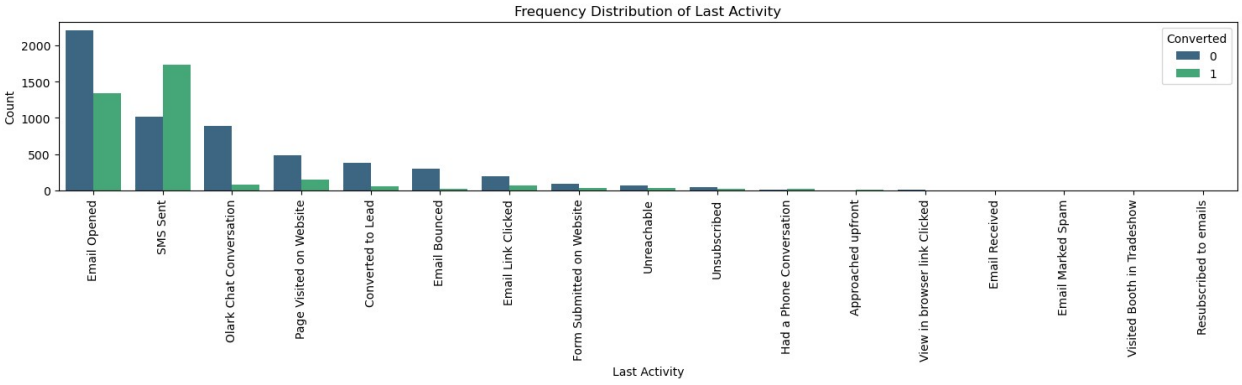
1) Categorical Variables

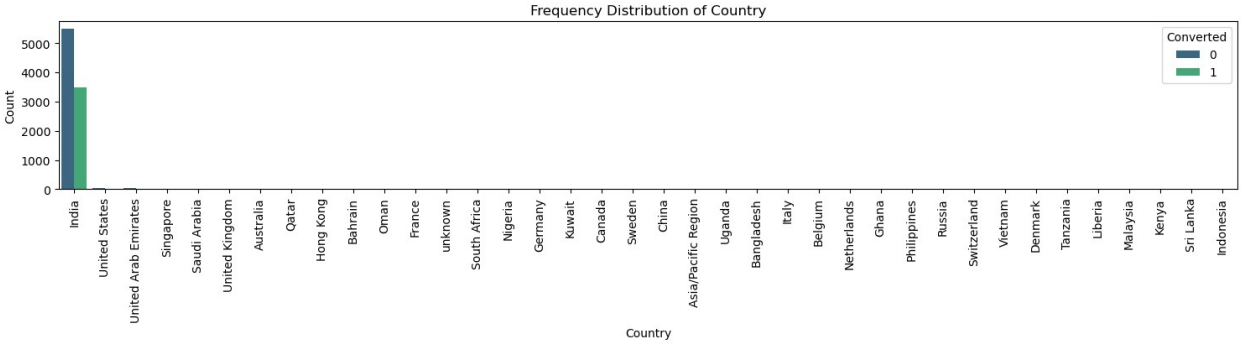
```
category = ['Lead Origin', 'Lead Source', 'Last
Activity','Country','City','Specialization','What is your current
occupation']

def plot_frequency_distribution(df, columns):
    """Plots bar charts for the frequency distribution of categorical
columns."""
    plt.figure(figsize=(15, 5 * len(columns)))  # Adjust figure size
based on the number of columns

    for i, col in enumerate(columns, 1):
        plt.subplot(len(columns), 1, i)  # Create subplots
        sns.countplot(data=df, x=col,
```

```
order=df[col].value_counts().index, palette='viridis',hue='Converted')
        plt.xticks(rotation=90)  # Rotate labels for readability
        plt.title(f"Frequency Distribution of {col}")
        plt.xlabel(col)
        plt.ylabel("Count")

    plt.tight_layout()
    plt.show()

plot_frequency_distribution(df,category)
```

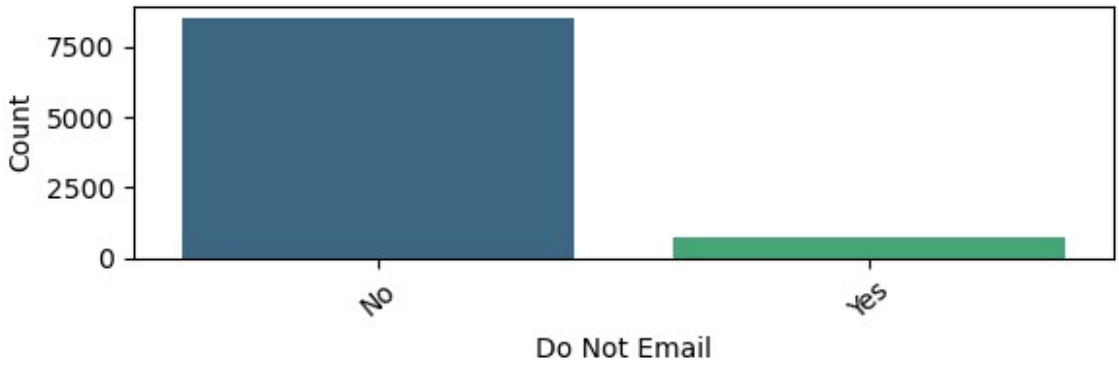

Frequency Distribution of Lead Origin
Converted
0
1
Count
3000
2500
2000
1500
1000
500
0
Landing Page Submission
API
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

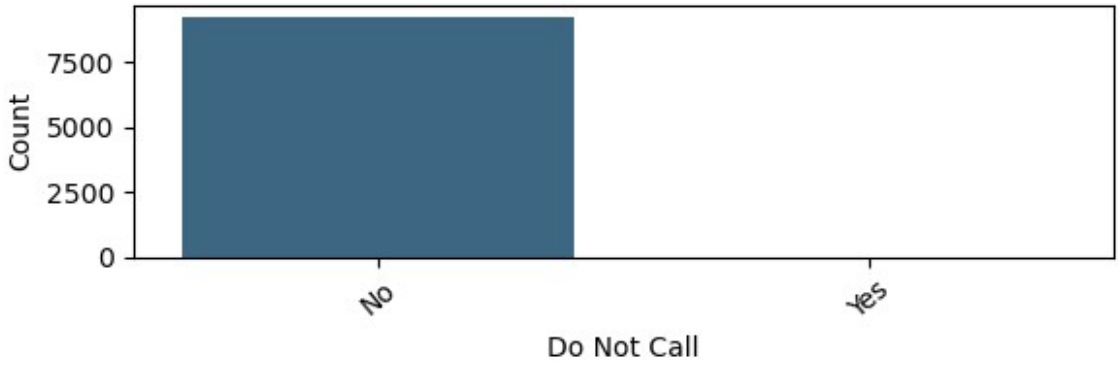

Frequency Distribution of Lead Source
Converted
0
1
Count
1500
1000
500
0
Google
Direct Traffic
Olark Chat
Organic Search
Reference
Welingak Website
Referral Sites
Facebook
Others
Lead Source

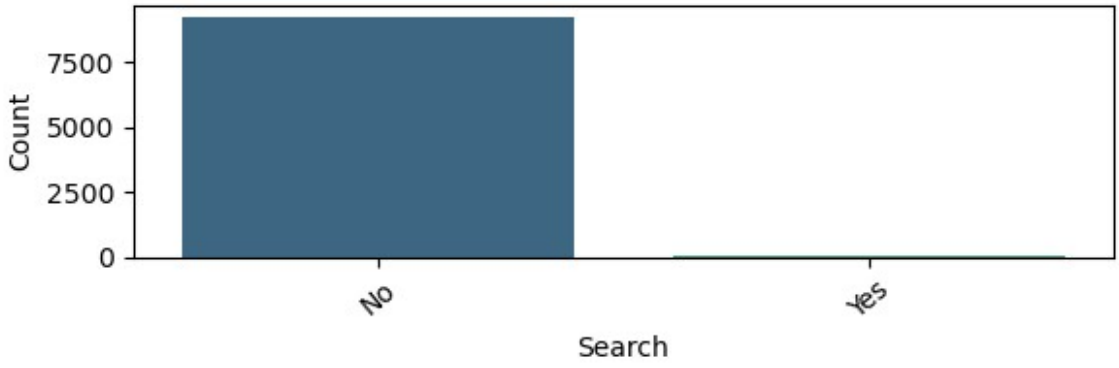

Frequency Distribution of Last Activity
Converted
0
1
Count
2000
1500
1000
500
0
Email Opened
SMS Sent
Olark Chat Conversation
Page Visited on Website
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unreachable
Unsubscribed
Had a Phone Conversation
Approached upfront
View in browser link Clicked
Email Received
Email Marked Spam
Visited Booth in Tradeshow
Resubscribed to emails
Last Activity

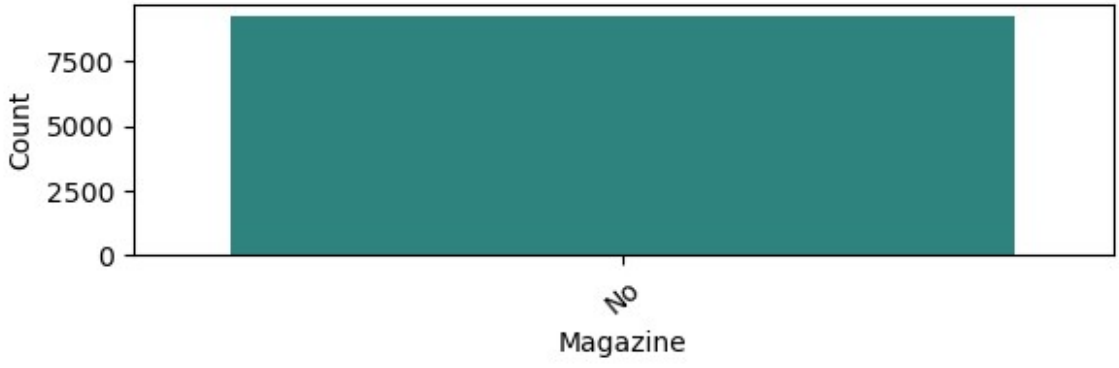

Frequency Distribution of Country
Converted
0
1
Count
5000
4000
3000
2000
1000
0
India
United States
United Arab Emirates
Singapore
Saudi Arabia
United Kingdom
Australia
Qatar
Hong Kong
Bahrain
Oman
France
unknown
South Africa
Nigeria
Germany
Kuwait
Canada
Sweden
China
Asia/Pacific Region
Uganda
Bangladesh
Italy
Belgium
Netherlands
Ghana
Philippines
Russia
Switzerland
Vietnam
Denmark
Tanzania
Liberia
Malaysia
Kenya
Sri Lanka
Indonesia
Country

```
#insights from the above graphs:
# 1. Top 3 Lead Source are: Google, Direct traffic and Olark chat, and there is good converstion rate. Further it can be noted that leads from reference and
# welingak website has very high conversion rate
# 2. Top 3 lead Origin are: Landing Page submission, API, lead add form, wherein lead add Form has highest conversion rate
# 3. Top 3 Last activity of Users are: Email opened, SMS sent, Olark Chat Conversation, , wherein SMS sent has highest conversion rate
# 4. Most of the leads are from INDIA
# 5. Others category in Specialization has highest no. of lead as well as lead conversion
# 6. Most of the leads are from UNemployed category and it has the highest rate of conversion
# 7. Most of the leads are generated from Mumbai along with highest no. of conversion

yesno_category= ['Do Not Email', 'Do Not Call',
'Search','Magazine','Newspaper Article','X Education Forums','Newspaper',
'Digital Advertisement','Through Recommendations',    'Receive More Updates About Our Courses',
'Update me on Supply Chain Content','Get updates on DM Content', 'I agree to pay the amount through cheque','A free copy of Mastering The Interview'
]


def plot_count(df, columns):
    """Plots bar charts for the frequency distribution of categorical columns."""
    plt.figure(figsize=(6, 2*len(columns)))  # Adjust figure size based on the number of columns

    for i, col in enumerate(columns, 1):
        plt.subplot(len(columns), 1, i)  # Create subplots
        sns.countplot(data=df, x=col, order=df[col].value_counts().index, palette='viridis')
        plt.xticks(rotation=40)  # Rotate labels for readability
        plt.xlabel(col)
        plt.ylabel("Count")
    plt.tight_layout()
    plt.show()

plot_count(df, yesno_category)
```


Count
7500
5000
2500
0
No
Yes
Do Not Email


Count
7500
5000
2500
0
No
Yes
Do Not Call


Count
7500
5000
2500
0
No
Yes
Search


Count
7500
5000
2500
0
No
Magazine


Count
7500
5000

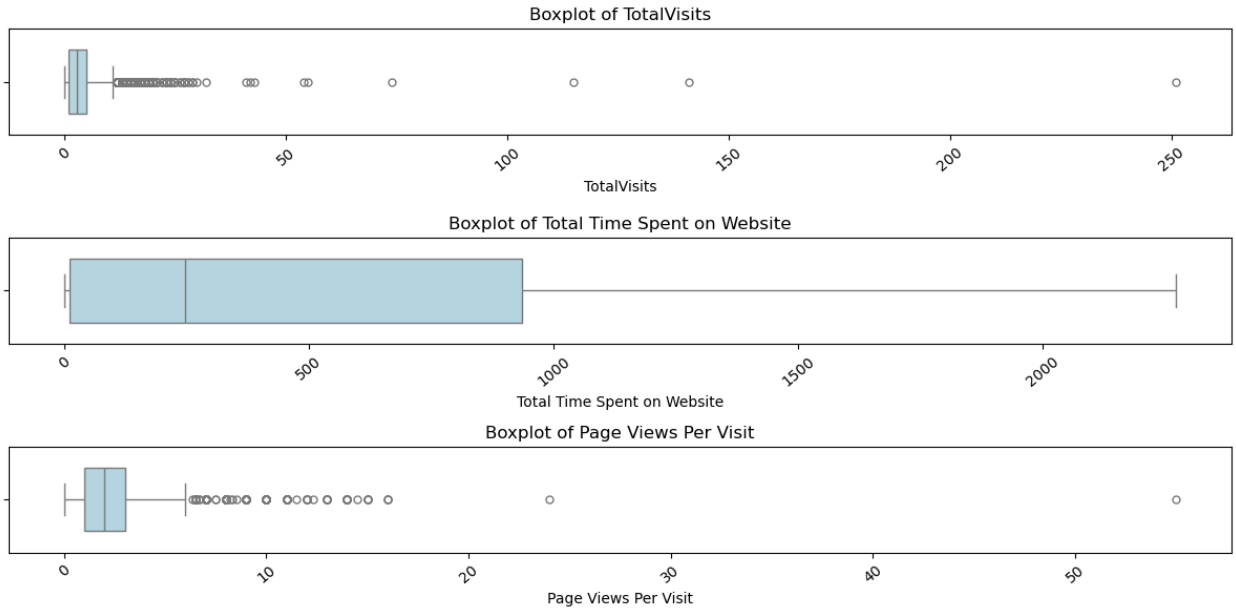
From the above visualisation, following can be noted:

1. Most of the yes no categories are not having any positive response, i.e. they are only having NO

2. Only one feature, "A free copy of mastering the interview" has relevant no. of yes

3. Hence we can remove all the unnecessary features before building our model

```
df.drop(['Do Not Email', 'Do Not Call', 'Search',      'Magazine',
     'Newspaper Article',  'X Education Forums', 'Newspaper',
'Digital Advertisement','Through Recommendations',     'Receive More
Updates About Our Courses',
'Update me on Supply Chain Content',  'Get updates on DM Content', 'I
agree to pay the amount through cheque'], axis=1, inplace=True)

Num_col = ['TotalVisits', 'Total Time Spent on Website', 'Page Views
Per Visit']

def plot_box(df, columns):

    plt.figure(figsize=(12, 2*len(columns)))  # Adjust figure size
based on the number of columns

    for i, col in enumerate(columns, 1):
        plt.subplot(len(columns), 1, i)  # Create subplots
        sns.boxplot(data=df, x=col, color='lightblue', fliersize=5,
width=0.6)
        plt.xticks(rotation=40)  # Rotate labels for readability
        plt.xlabel(col)
        plt.title(f"Boxplot of {col}")

    plt.tight_layout()
    plt.show()

plot_box(df, Num_col)
```

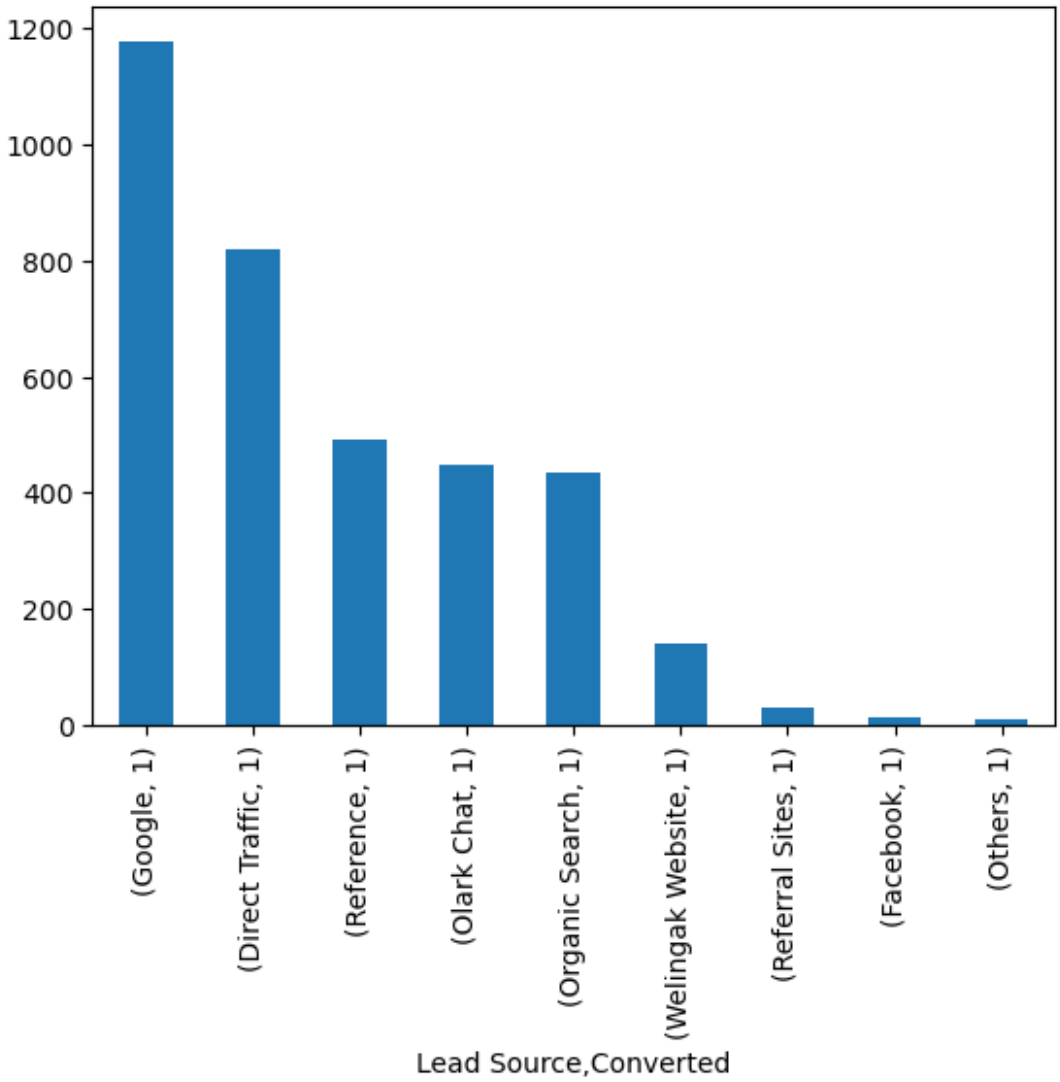


From the boxplot it is clear that Total time spend on website has no outliers present and features has outlier i.e Total Visits and Page Views per visit has number of outliers present

Bivariate Analysis

```
df[df['Converted']==1].groupby('Lead Source')
['Converted'].value_counts().sort_values(ascending=False).plot(kind='b
ar')
```

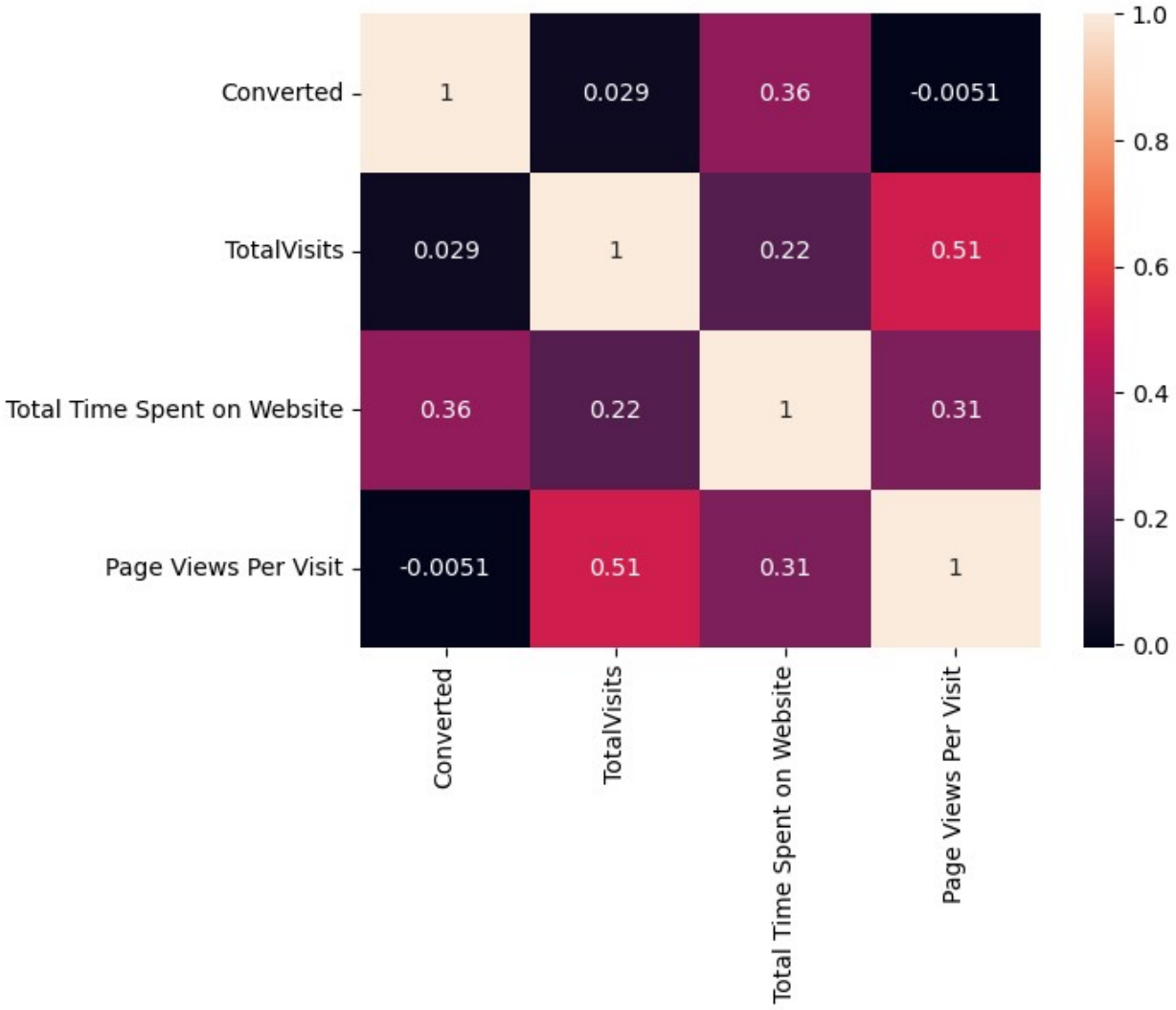
```
<Axes: xlabel='Lead Source,Converted'>
```



Google has the highest conversion rate out of all Lead Sources

```
nums = ['Converted', 'TotalVisits',   'Total Time Spent on Website',
     'Page Views Per Visit']
sns.heatmap(df[nums].corr(), annot=True)
```

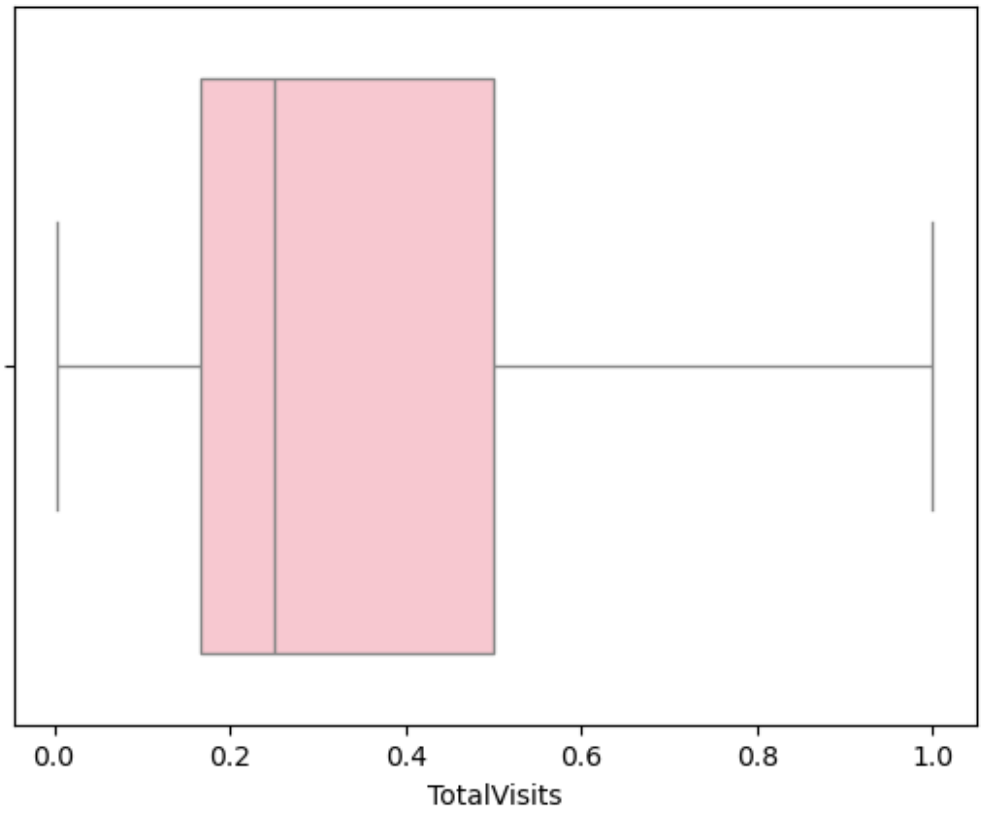
```
<Axes: >
```



Outlier treatment

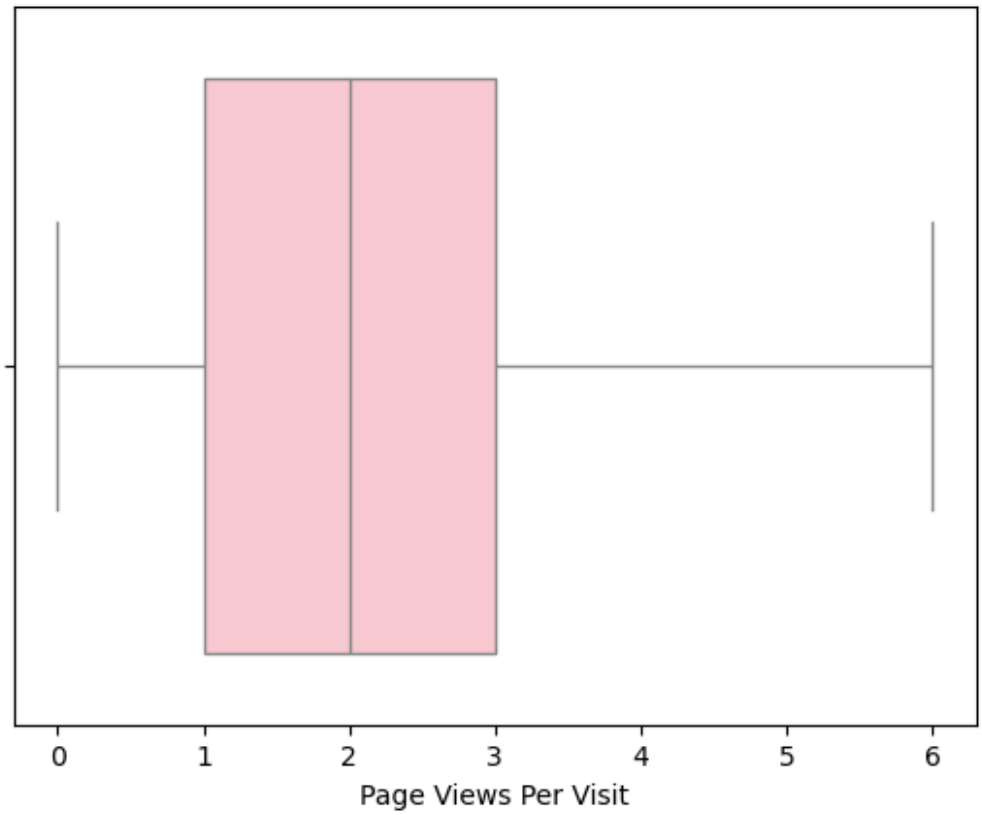
```
df['TotalVisits'] = 1 / (df['TotalVisits']+1)
sns.boxplot(data=df, x='TotalVisits', color='pink')
```

```
<Axes: xlabel='TotalVisits'>
```




```
from scipy.stats.mstats import winsorize
df['Page Views Per Visit'] = winsorize(df['Page Views Per Visit'],
limits=[0.05, 0.05])  # Cap bottom & top 5%
sns.boxplot(data=df, x='Page Views Per Visit', color='pink')
```

```
<Axes: xlabel='Page Views Per Visit'>
```




```
df
```

	Lead Origin	Lead Source	Converted	TotalVisits
0	API	Olark Chat	0	1.000000
1	API	Organic Search	0	0.166667
2	Landing Page Submission	Direct Traffic	1	0.333333
3	Landing Page Submission	Direct Traffic	0	0.500000
4	Landing Page Submission	Google	1	0.333333
...	...	...	...	...
9235	Landing Page Submission	Direct Traffic	1	0.111111
9236	Landing Page Submission	Direct Traffic	0	0.333333
9237	Landing Page Submission	Direct Traffic	0	0.333333
9238	Landing Page Submission	Google	1	0.250000
9239	Landing Page Submission	Direct Traffic	1	0.142857

```
      Total Time Spent on Website  Page Views Per Visit  \
0                               0                  0.00
1                             674                  2.50
2                            1532                  2.00
3                             305                  1.00
4                            1428                  1.00
...                           ...                   ...
9235                         1845                  2.67
9236                          238                  2.00
9237                          199                  2.00
9238                          499                  3.00
9239                         1279                  3.00

                Last Activity       Country              Specialization
\
0     Page Visited on Website         India                      Others

1                Email Opened         India                      Others

2                Email Opened         India    Business Administration

3                 Unreachable         India      Media and Advertising

4           Converted to Lead         India                      Others

...                       ...           ...                         ...

9235        Email Marked Spam  Saudi Arabia     IT Projects Management

9236                SMS Sent         India      Media and Advertising

9237                SMS Sent         India    Business Administration

9238                SMS Sent         India  Human Resource Management

9239                SMS Sent    Bangladesh    Supply Chain Management


     What is your current occupation  \
0                          Unemployed
1                          Unemployed
2                             Student
3                          Unemployed
4                          Unemployed
...                               ...
9235                       Unemployed
9236                       Unemployed
9237                       Unemployed
9238                       Unemployed
```

```
9239                       Unemployed

     What matters most to you in choosing a course                   City
\
0                           Better Career Prospects                Mumbai

1                           Better Career Prospects                Mumbai

2                           Better Career Prospects                Mumbai

3                           Better Career Prospects                Mumbai

4                           Better Career Prospects                Mumbai

...                                             ...                   ...

9235                        Better Career Prospects                Mumbai

9236                        Better Career Prospects                Mumbai

9237                        Better Career Prospects                Mumbai

9238                        Better Career Prospects  Other Metro Cities

9239                        Better Career Prospects        Other Cities


     A free copy of Mastering The Interview Last Notable Activity
0                                        No              Modified
1                                        No          Email Opened
2                                       Yes          Email Opened
3                                        No              Modified
4                                        No              Modified
...                                     ...                   ...
9235                                     No     Email Marked Spam
9236                                    Yes              SMS Sent
9237                                    Yes              SMS Sent
9238                                     No              SMS Sent
9239                                    Yes              Modified

[9240 rows x 14 columns]
```

```
# Splitting X and y

X = df.drop('Converted', axis=1)
y = df['Converted']
X
```

```
                  Lead Origin     Lead Source  TotalVisits  \
0                         API      Olark Chat     1.000000
1                         API  Organic Search     0.166667
```

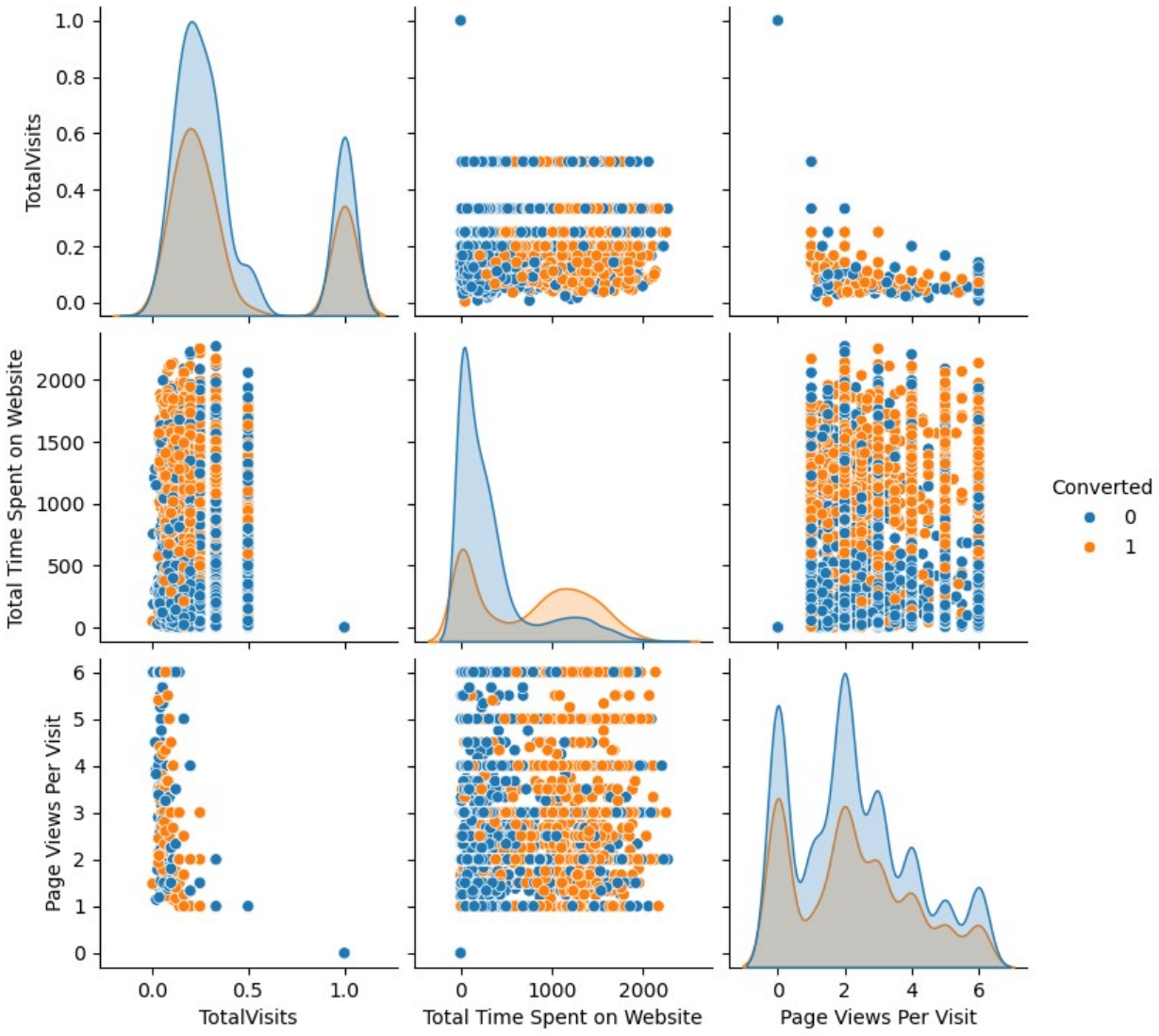
```
2     Landing Page Submission  Direct Traffic     0.333333
3     Landing Page Submission  Direct Traffic     0.500000
4     Landing Page Submission          Google     0.333333
...                       ...             ...          ...
9235  Landing Page Submission  Direct Traffic     0.111111
9236  Landing Page Submission  Direct Traffic     0.333333
9237  Landing Page Submission  Direct Traffic     0.333333
9238  Landing Page Submission          Google     0.250000
9239  Landing Page Submission  Direct Traffic     0.142857

      Total Time Spent on Website  Page Views Per Visit  \
0                               0                  0.00
1                             674                  2.50
2                            1532                  2.00
3                             305                  1.00
4                            1428                  1.00
...                           ...                   ...
9235                         1845                  2.67
9236                          238                  2.00
9237                          199                  2.00
9238                          499                  3.00
9239                         1279                  3.00

                Last Activity       Country              Specialization
\
0     Page Visited on Website         India                      Others

1                Email Opened         India                      Others

2                Email Opened         India    Business Administration

3                 Unreachable         India      Media and Advertising

4           Converted to Lead         India                      Others

...                       ...           ...                         ...

9235        Email Marked Spam  Saudi Arabia     IT Projects Management

9236                SMS Sent         India      Media and Advertising

9237                SMS Sent         India    Business Administration

9238                SMS Sent         India  Human Resource Management

9239                SMS Sent    Bangladesh    Supply Chain Management


     What is your current occupation  \
0                         Unemployed
1                         Unemployed
```

```
2                              Student
3                           Unemployed
4                           Unemployed
...                                ...
9235                        Unemployed
9236                        Unemployed
9237                        Unemployed
9238                        Unemployed
9239                        Unemployed

     What matters most to you in choosing a course                 City
\
0                            Better Career Prospects               Mumbai

1                            Better Career Prospects               Mumbai

2                            Better Career Prospects               Mumbai

3                            Better Career Prospects               Mumbai

4                            Better Career Prospects               Mumbai

...                                              ...                  ...

9235                         Better Career Prospects               Mumbai

9236                         Better Career Prospects               Mumbai

9237                         Better Career Prospects               Mumbai

9238                         Better Career Prospects  Other Metro Cities

9239                         Better Career Prospects        Other Cities


     A free copy of Mastering The Interview Last Notable Activity
0                                         No              Modified
1                                         No          Email Opened
2                                        Yes          Email Opened
3                                         No              Modified
4                                         No              Modified
...                                      ...                   ...
9235                                      No     Email Marked Spam
9236                                     Yes              SMS Sent
9237                                     Yes              SMS Sent
9238                                      No              SMS Sent
9239                                     Yes              Modified

[9240 rows x 13 columns]

y
```

```
0       0
1       0
2       1
3       0
4       1
       ..
9235    1
9236    0
9237    0
9238    1
9239    1
Name: Converted, Length: 9240, dtype: int64
```

```
sns.pairplot(df, hue='Converted')
```

```
<seaborn.axisgrid.PairGrid at 0x208350b9ac0>
```



```
numerical = ['TotalVisits','Total Time Spent on Website',   'Page
Views Per Visit']
categorical = [col for col in df.columns if df[col].dtype == 'object']
numerical, categorical
```

```
(['TotalVisits', 'Total Time Spent on Website', 'Page Views Per
Visit'],
 ['Lead Origin',
  'Lead Source',
  'Last Activity',
  'Country',
  'Specialization',
  'What is your current occupation',
  'What matters most to you in choosing a course',
  'City',
  'A free copy of Mastering The Interview',
  'Last Notable Activity'])
```

```
#train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=1)
X_train, X_test, y_train, y_test
```

```
(                  Lead Origin     Lead Source  TotalVisits  \
 5321  Landing Page Submission          Google     0.200000
 2575            Lead Add Form       Reference     1.000000
 3363  Landing Page Submission  Direct Traffic     0.333333
 955                      API  Referral Sites     0.111111
 6406  Landing Page Submission  Direct Traffic     0.500000
 ...                      ...             ...          ...
 2895  Landing Page Submission  Organic Search     0.200000
 7813  Landing Page Submission          Google     0.250000
 905                      API          Google     0.333333
 5192                     API          Google     0.333333
 235   Landing Page Submission  Organic Search     0.076923

       Total Time Spent on Website  Page Views Per Visit  \
 5321                          239                   4.0
 2575                            0                   0.0
 3363                          271                   2.0
 955                            51                   4.0
 6406                           95                   1.0
 ...                           ...                   ...
 2895                          502                   4.0
 7813                          260                   3.0
 905                           271                   1.0
 5192                          444                   2.0
 235                            28                   6.0

                  Last Activity Country           Specialization  \
```

```
5321               Email Opened   India   IT Projects Management
2575                   SMS Sent   India  Business Administration
3363               Email Opened   India  Business Administration
955     Olark Chat Conversation   India                   Others
6406          Converted to Lead   India  Business Administration
...                         ...     ...                      ...
2895               Email Opened   India    Media and Advertising
7813  Form Submitted on Website   India  Business Administration
905                Email Opened   India                   Others
5192    Olark Chat Conversation   India                   Others
235                Email Opened   India   International Business

     What is your current occupation  \
5321                       Unemployed
2575                       Unemployed
3363                       Unemployed
955                        Unemployed
6406                       Unemployed
...                               ...
2895                       Unemployed
7813                       Unemployed
905                        Unemployed
5192                       Unemployed
235                        Unemployed

     What matters most to you in choosing a course  \
5321                        Better Career Prospects
2575                        Better Career Prospects
3363                        Better Career Prospects
955                         Better Career Prospects
6406                        Better Career Prospects
...                                             ...
2895                        Better Career Prospects
7813                        Better Career Prospects
905                         Better Career Prospects
5192                        Better Career Prospects
235                         Better Career Prospects

                              City A free copy of Mastering The
Interview  \
 5321                       Mumbai
No
 2575            Thane & Outskirts
No
 3363  Other Cities of Maharashtra
Yes
 955                        Mumbai
No
 6406                       Mumbai
```

```
Yes
 ...                            ...
...
 2895                        Mumbai
Yes
 7813             Thane & Outskirts
No
 905                         Mumbai
No
 5192                        Mumbai
No
 235                    Other Cities
Yes

      Last Notable Activity
 5321          Email Opened
 2575              SMS Sent
 3363          Email Opened
 955               Modified
 6406              Modified
 ...                    ...
 2895          Email Opened
 7813              Modified
 905           Email Opened
 5192              Modified
 235           Email Opened

 [7392 rows x 13 columns],
                    Lead Origin     Lead Source  TotalVisits  \
 2140              Lead Add Form       Reference     1.000000
 7707  Landing Page Submission  Direct Traffic     0.200000
 1522  Landing Page Submission  Direct Traffic     0.250000
 1873  Landing Page Submission          Google     0.333333
 8100                      API          Google     0.071429
 ...                       ...             ...          ...
 1837  Landing Page Submission          Google     0.142857
 7173                      API      Olark Chat     1.000000
 634   Landing Page Submission  Direct Traffic     0.333333
 4406              Lead Add Form          Google     1.000000
 3465              Lead Add Form       Reference     1.000000

       Total Time Spent on Website  Page Views Per Visit       Last
Activity  \
 2140                            0                  0.00  Email Link
Clicked
 7707                         1503                  2.00
SMS Sent
 1522                         1024                  1.50        Email
Opened
```

```
 1873                          186                  2.00
SMS Sent
 8100                         1725                  3.25        Email
Opened
 ...                           ...                   ...
...
 1837                          213                  6.00        Email
Opened
 7173                            0                  0.00
SMS Sent
 634                            25                  2.00        Email
Opened
 4406                            0                  0.00
SMS Sent
 3465                            0                  0.00
SMS Sent

      Country              Specialization What is your current
occupation  \
 2140   India         Services Excellence
Unemployed
 7707   India       Healthcare Management
Student
 1522   India  Human Resource Management
Unemployed
 1873   India                      Others
Unemployed
 8100   India                      Others
Unemployed
 ...      ...                         ...                                .
..
 1837   India        Marketing Management
Unemployed
 7173   India                      Others
Unemployed
 634    India     IT Projects Management
Unemployed
 4406   India                      Others
Unemployed
 3465   India                      Others
Unemployed

      What matters most to you in choosing a course  \
 2140                       Better Career Prospects
 7707                       Better Career Prospects
 1522                       Better Career Prospects
 1873                       Better Career Prospects
 8100                       Better Career Prospects
 ...                                           ...
```

```
 1837                          Better Career Prospects
 7173                          Better Career Prospects
 634                           Better Career Prospects
 4406                          Better Career Prospects
 3465                          Better Career Prospects

                                 City A free copy of Mastering The
Interview  \
 2140              Thane & Outskirts
No
 7707                          Mumbai
No
 1522  Other Cities of Maharashtra
Yes
 1873                          Mumbai
No
 8100                          Mumbai
No
 ...                              ...
...
 1837                          Mumbai
No
 7173                          Mumbai
No
 634                           Mumbai
Yes
 4406                          Mumbai
No
 3465                          Mumbai
No

      Last Notable Activity
 2140               Modified
 7707               SMS Sent
 1522          Email Opened
 1873               Modified
 8100          Email Opened
 ...                     ...
 1837               Modified
 7173               SMS Sent
 634           Email Opened
 4406               SMS Sent
 3465               SMS Sent

 [1848 rows x 13 columns],
 5321    0
 2575    1
 3363    0
 955     0
```

```
6406    0
       ..
2895    0
7813    0
905     0
5192    0
235     0
Name: Converted, Length: 7392, dtype: int64,
2140    1
7707    1
1522    1
1873    0
8100    1
       ..
1837    0
7173    0
634     0
4406    1
3465    1
Name: Converted, Length: 1848, dtype: int64)
```

```
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer

# Define the preprocessing pipeline
preprocessor = ColumnTransformer([
    ('num', StandardScaler(), numerical),  # Standard scaling for
numerical data
    ('cat', OneHotEncoder(handle_unknown='ignore', drop='first'),
categorical)  # OHE for categorical data
])

# Fit the transformer only on training data & transform both train &
test sets
X_train = preprocessor.fit_transform(X_train)
X_test = preprocessor.transform(X_test)

X_train
```

```
<7392x110 sparse matrix of type '<class 'numpy.float64'>'
     with 72918 stored elements in Compressed Sparse Row format>
```

```
# model
classifier=LogisticRegression()
classifier
```

```
LogisticRegression()
```

```
classifier.fit(X_train,y_train)
```

```
LogisticRegression()
```

```
y_pred=classifier.predict(X_test)
y_pred
```

```
array([1, 1, 0, ..., 0, 1, 1], dtype=int64)
```

```
accuracy_score(y_test,y_pred)
```

```
0.8235930735930735
```

```
# There might be overfitting, hence applying Hyperparameter
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
params = {
 'C': [0.01, 0.1, 1,10,50,100,200],
 'penalty' : ['l1','l2', 'elasticnet']
}
lr_grid = GridSearchCV(classifier, param_grid=params, cv=5, verbose=3,
scoring='accuracy')
lr_grid
```

```
GridSearchCV(cv=5, estimator=LogisticRegression(),
             param_grid={'C': [0.01, 0.1, 1, 10, 50, 100, 200],
                         'penalty': ['l1', 'l2', 'elasticnet']},
             scoring='accuracy', verbose=3)
```

```
lr_grid.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 21 candidates, totalling 105 fits
[CV 1/5] END ..................C=0.01, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ..................C=0.01, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ..................C=0.01, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ..................C=0.01, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ..................C=0.01, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END ................C=0.01, penalty=l2;, score=0.787 total
time=   0.0s
[CV 2/5] END ................C=0.01, penalty=l2;, score=0.824 total
time=   0.0s
[CV 3/5] END ................C=0.01, penalty=l2;, score=0.802 total
time=   0.0s
[CV 4/5] END ................C=0.01, penalty=l2;, score=0.803 total
time=   0.0s
[CV 5/5] END ................C=0.01, penalty=l2;, score=0.794 total
time=   0.0s
[CV 1/5] END ..........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ..........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
```

```
[CV 3/5] END .........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END .........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END .........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ...................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ...................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ...................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ...................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ...................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END .................C=0.1, penalty=l2;, score=0.805 total
time=   0.0s
[CV 2/5] END .................C=0.1, penalty=l2;, score=0.834 total
time=   0.0s
[CV 3/5] END .................C=0.1, penalty=l2;, score=0.813 total
time=   0.0s
[CV 4/5] END .................C=0.1, penalty=l2;, score=0.812 total
time=   0.0s
[CV 5/5] END .................C=0.1, penalty=l2;, score=0.809 total
time=   0.0s
[CV 1/5] END ...........C=0.1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ...........C=0.1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ...........C=0.1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ...........C=0.1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ...........C=0.1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ....................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ....................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ....................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ....................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ....................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=1, penalty=l2;, score=0.803 total
time=   0.0s
[CV 2/5] END ..................C=1, penalty=l2;, score=0.834 total
```

```
time=   0.0s
[CV 3/5] END ..................C=1, penalty=l2;, score=0.812 total
time=   0.0s
[CV 4/5] END ..................C=1, penalty=l2;, score=0.816 total
time=   0.0s
[CV 5/5] END ..................C=1, penalty=l2;, score=0.811 total
time=   0.0s
[CV 1/5] END .............C=1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END .............C=1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END .............C=1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END .............C=1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END .............C=1, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=10, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ..................C=10, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ..................C=10, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ..................C=10, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ..................C=10, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=10, penalty=l2;, score=0.805 total
time=   0.0s
[CV 2/5] END ..................C=10, penalty=l2;, score=0.836 total
time=   0.0s
[CV 3/5] END ..................C=10, penalty=l2;, score=0.811 total
time=   0.0s
[CV 4/5] END ..................C=10, penalty=l2;, score=0.816 total
time=   0.0s
[CV 5/5] END ..................C=10, penalty=l2;, score=0.811 total
time=   0.0s
[CV 1/5] END ............C=10, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ............C=10, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ............C=10, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ............C=10, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ............C=10, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=50, penalty=l1;, score=nan total
time=   0.0s
```

```
[CV 2/5] END .....................C=50, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END .....................C=50, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END .....................C=50, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END .....................C=50, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=50, penalty=l2;, score=0.804 total
time=   0.0s
[CV 2/5] END ..................C=50, penalty=l2;, score=0.835 total
time=   0.0s
[CV 3/5] END ..................C=50, penalty=l2;, score=0.811 total
time=   0.0s
[CV 4/5] END ..................C=50, penalty=l2;, score=0.815 total
time=   0.0s
[CV 5/5] END ..................C=50, penalty=l2;, score=0.811 total
time=   0.0s
[CV 1/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ...................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ...................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ...................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ...................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ...................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=100, penalty=l2;, score=0.803 total
time=   0.0s
[CV 2/5] END ..................C=100, penalty=l2;, score=0.834 total
time=   0.0s
[CV 3/5] END ..................C=100, penalty=l2;, score=0.812 total
time=   0.0s
[CV 4/5] END ..................C=100, penalty=l2;, score=0.817 total
time=   0.0s
[CV 5/5] END ..................C=100, penalty=l2;, score=0.809 total
time=   0.0s
[CV 1/5] END ............C=100, penalty=elasticnet;, score=nan total
```

```
time=   0.0s
[CV 2/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END .................C=200, penalty=l2;, score=0.803 total
time=   0.0s
[CV 2/5] END .................C=200, penalty=l2;, score=0.835 total
time=   0.0s
[CV 3/5] END .................C=200, penalty=l2;, score=0.810 total
time=   0.0s
[CV 4/5] END .................C=200, penalty=l2;, score=0.815 total
time=   0.0s
[CV 5/5] END .................C=200, penalty=l2;, score=0.809 total
time=   0.0s
[CV 1/5] END ...........C=200, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ...........C=200, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ...........C=200, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ...........C=200, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ...........C=200, penalty=elasticnet;, score=nan total
time=   0.0s

GridSearchCV(cv=5, estimator=LogisticRegression(),
             param_grid={'C': [0.01, 0.1, 1, 10, 50, 100, 200],
                         'penalty': ['l1', 'l2', 'elasticnet']},
             scoring='accuracy', verbose=3)

lr_grid.best_params_

{'C': 10, 'penalty': 'l2'}

lr_grid.best_score_
```

```
0.8156102439109189
```

```
y_pred_grid = lr_grid.best_estimator_.predict(X_test)
```

```
accuracy_score(y_test, y_pred_grid)
```

```
0.8225108225108225
```

```
lr_random = RandomizedSearchCV(classifier, param_distributions=params,
cv=5, verbose=3, scoring='accuracy',random_state=1)
lr_random
```

```
RandomizedSearchCV(cv=5, estimator=LogisticRegression(),
                   param_distributions={'C': [0.01, 0.1, 1, 10, 50,
100, 200],
                                        'penalty': ['l1', 'l2',
'elasticnet']},
                   random_state=1, scoring='accuracy', verbose=3)
```

```
lr_random.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
[CV 1/5] END ..................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ..................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ..................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ..................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ..................C=100, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END ..................C=10, penalty=l2;, score=0.805 total
time=   0.0s
[CV 2/5] END ..................C=10, penalty=l2;, score=0.836 total
time=   0.0s
[CV 3/5] END ..................C=10, penalty=l2;, score=0.811 total
time=   0.0s
[CV 4/5] END ..................C=10, penalty=l2;, score=0.816 total
time=   0.0s
[CV 5/5] END ..................C=10, penalty=l2;, score=0.811 total
time=   0.0s
[CV 1/5] END ..................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ..................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ..................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ..................C=0.1, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ..................C=0.1, penalty=l1;, score=nan total
```

```
time=   0.0s
[CV 1/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ..................C=200, penalty=l1;, score=nan total
time=   0.0s
[CV 1/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ...........C=100, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ............C=50, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ................C=100, penalty=l2;, score=0.803 total
time=   0.0s
[CV 2/5] END ................C=100, penalty=l2;, score=0.834 total
time=   0.0s
[CV 3/5] END ................C=100, penalty=l2;, score=0.812 total
time=   0.0s
[CV 4/5] END ................C=100, penalty=l2;, score=0.817 total
time=   0.0s
[CV 5/5] END ................C=100, penalty=l2;, score=0.809 total
time=   0.0s
[CV 1/5] END ................C=0.1, penalty=l2;, score=0.805 total
time=   0.0s
[CV 2/5] END ................C=0.1, penalty=l2;, score=0.834 total
time=   0.0s
[CV 3/5] END ................C=0.1, penalty=l2;, score=0.813 total
time=   0.0s
[CV 4/5] END ................C=0.1, penalty=l2;, score=0.812 total
time=   0.0s
```

```
[CV 5/5] END .................C=0.1, penalty=l2;, score=0.809 total
time=   0.0s
[CV 1/5] END ..........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 2/5] END ..........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 3/5] END ..........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 4/5] END ..........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 5/5] END ..........C=0.01, penalty=elasticnet;, score=nan total
time=   0.0s
[CV 1/5] END ......................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 2/5] END ......................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 3/5] END ......................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 4/5] END ......................C=1, penalty=l1;, score=nan total
time=   0.0s
[CV 5/5] END ......................C=1, penalty=l1;, score=nan total
time=   0.0s

RandomizedSearchCV(cv=5, estimator=LogisticRegression(),
                   param_distributions={'C': [0.01, 0.1, 1, 10, 50,
100, 200],
                                        'penalty': ['l1', 'l2',
'elasticnet']},
                   random_state=1, scoring='accuracy', verbose=3)

lr_random.best_score_

0.8156102439109189

lr_random.best_params_

{'penalty': 'l2', 'C': 10}

lr_random.best_estimator_

LogisticRegression(C=10)

y_pred_random = lr_random.best_estimator_.predict(X_test)

accuracy_score(y_test, y_pred_random)

0.8225108225108225

# Moving further with Grid search CV
confusion_matrix(y_test, y_pred_grid)
```

```
array([[994, 125],
       [203, 526]], dtype=int64)
```

```
print(classification_report(y_test,y_pred_grid))
```

```
              precision    recall  f1-score   support

           0       0.83      0.89      0.86      1119
           1       0.81      0.72      0.76       729

    accuracy                           0.82      1848
   macro avg       0.82      0.80      0.81      1848
weighted avg       0.82      0.82      0.82      1848
```

```
y_predict_proba = lr_grid.predict_proba(X_test)[:,1] #probability for
1 class
```

```
y_predict_proba
```

```
array([0.78652257, 0.90526968, 0.45231305, ..., 0.08106563,
0.88356408,
       0.8942288 ])
```

```
fpr, tpr, threshholds = roc_curve(y_test, y_predict_proba)  #it will
return TPR and FPR with diff cutoff probability
fpr, tpr, threshholds
```

```
(array([0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
0.00000000e+00,
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
0.00000000e+00,
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
0.00000000e+00,
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
0.00000000e+00,
        0.00000000e+00, 0.00000000e+00, 8.93655049e-04, 8.93655049e-
04,
        2.68096515e-03, 2.68096515e-03, 4.46827525e-03, 4.46827525e-
03,
        6.25558534e-03, 6.25558534e-03, 7.14924039e-03, 7.14924039e-
03,
        7.14924039e-03, 7.14924039e-03, 8.04289544e-03, 8.04289544e-
03,
        8.04289544e-03, 8.04289544e-03, 8.04289544e-03, 8.04289544e-
03,
        8.93655049e-03, 8.93655049e-03, 9.83020554e-03, 9.83020554e-
03,
        9.83020554e-03, 9.83020554e-03, 1.07238606e-02, 1.07238606e-
02,
        1.16175156e-02, 1.16175156e-02, 1.25111707e-02, 1.25111707e-
```

```
02,
        1.34048257e-02, 1.34048257e-02, 1.42984808e-02, 1.42984808e-
02,
        1.51921358e-02, 1.51921358e-02, 1.60857909e-02, 1.60857909e-
02,
        1.69794459e-02, 1.69794459e-02, 1.78731010e-02, 1.78731010e-
02,
        1.87667560e-02, 1.87667560e-02, 1.96604111e-02, 1.96604111e-
02,
        2.05540661e-02, 2.05540661e-02, 2.14477212e-02, 2.14477212e-
02,
        2.23413762e-02, 2.23413762e-02, 2.32350313e-02, 2.32350313e-
02,
        2.50223414e-02, 2.50223414e-02, 2.59159964e-02, 2.59159964e-
02,
        2.68096515e-02, 2.68096515e-02, 2.77033065e-02, 2.77033065e-
02,
        2.94906166e-02, 2.94906166e-02, 3.03842717e-02, 3.03842717e-
02,
        3.12779267e-02, 3.12779267e-02, 3.21715818e-02, 3.21715818e-
02,
        3.48525469e-02, 3.48525469e-02, 3.57462020e-02, 3.57462020e-
02,
        3.93208222e-02, 3.93208222e-02, 4.02144772e-02, 4.02144772e-
02,
        4.02144772e-02, 4.20017873e-02, 4.20017873e-02, 4.20017873e-
02,
        4.20017873e-02, 4.37890974e-02, 4.37890974e-02, 4.46827525e-
02,
        4.46827525e-02, 4.64700626e-02, 4.64700626e-02, 4.73637176e-
02,
        4.73637176e-02, 4.91510277e-02, 4.91510277e-02, 5.00446828e-
02,
        5.09383378e-02, 5.09383378e-02, 5.18319929e-02, 5.18319929e-
02,
        5.27256479e-02, 5.27256479e-02, 5.63002681e-02, 5.63002681e-
02,
        5.71939231e-02, 5.71939231e-02, 5.80875782e-02, 5.80875782e-
02,
        5.98748883e-02, 5.98748883e-02, 6.07685433e-02, 6.07685433e-
02,
        6.25558534e-02, 6.25558534e-02, 6.34495085e-02, 6.34495085e-
02,
        6.43431635e-02, 6.43431635e-02, 6.52368186e-02, 6.52368186e-
02,
        6.61304736e-02, 6.61304736e-02, 6.79177837e-02, 6.79177837e-
02,
        7.14924039e-02, 7.14924039e-02, 7.23860590e-02, 7.23860590e-
02,
```

```
        7.32797140e-02, 7.32797140e-02, 7.41733691e-02, 7.41733691e-
02,
        7.68543342e-02, 7.68543342e-02, 7.77479893e-02, 7.77479893e-
02,
        7.86416443e-02, 7.86416443e-02, 7.95352994e-02, 7.95352994e-
02,
        8.13226095e-02, 8.13226095e-02, 8.48972297e-02, 8.48972297e-
02,
        9.65147453e-02, 9.83020554e-02, 9.83020554e-02, 9.91957105e-
02,
        9.91957105e-02, 1.00089366e-01, 1.00089366e-01, 1.00983021e-
01,
        1.00983021e-01, 1.01876676e-01, 1.01876676e-01, 1.03663986e-
01,
        1.03663986e-01, 1.04557641e-01, 1.04557641e-01, 1.06344951e-
01,
        1.06344951e-01, 1.09025916e-01, 1.09025916e-01, 1.09919571e-
01,
        1.09919571e-01, 1.11706881e-01, 1.11706881e-01, 1.12600536e-
01,
        1.12600536e-01, 1.14387846e-01, 1.14387846e-01, 1.15281501e-
01,
        1.15281501e-01, 1.17068811e-01, 1.17068811e-01, 1.17962466e-
01,
        1.17962466e-01, 1.19749777e-01, 1.19749777e-01, 1.20643432e-
01,
        1.20643432e-01, 1.21537087e-01, 1.21537087e-01, 1.22430742e-
01,
        1.22430742e-01, 1.23324397e-01, 1.23324397e-01, 1.24218052e-
01,
        1.24218052e-01, 1.26899017e-01, 1.26899017e-01, 1.27792672e-
01,
        1.27792672e-01, 1.28686327e-01, 1.28686327e-01, 1.29579982e-
01,
        1.29579982e-01, 1.37622878e-01, 1.37622878e-01, 1.39410188e-
01,
        1.39410188e-01, 1.42091153e-01, 1.42091153e-01, 1.46559428e-
01,
        1.46559428e-01, 1.49240393e-01, 1.49240393e-01, 1.50134048e-
01,
        1.50134048e-01, 1.54602324e-01, 1.54602324e-01, 1.57283289e-
01,
        1.57283289e-01, 1.59964254e-01, 1.59964254e-01, 1.60857909e-
01,
        1.60857909e-01, 1.71581769e-01, 1.71581769e-01, 1.72475424e-
01,
        1.72475424e-01, 1.73369080e-01, 1.73369080e-01, 1.74262735e-
01,
        1.74262735e-01, 1.76050045e-01, 1.76050045e-01, 1.79624665e-
```

```
01,
        1.79624665e-01, 1.80518320e-01, 1.80518320e-01, 1.81411975e-
01,
        1.81411975e-01, 1.83199285e-01, 1.83199285e-01, 1.84092940e-
01,
        1.84092940e-01, 1.84986595e-01, 1.84986595e-01, 1.85880250e-
01,
        1.85880250e-01, 1.88561215e-01, 1.88561215e-01, 1.89454870e-
01,
        1.89454870e-01, 1.92135836e-01, 1.92135836e-01, 1.93029491e-
01,
        1.93029491e-01, 1.93923146e-01, 1.93923146e-01, 1.95710456e-
01,
        1.95710456e-01, 1.96604111e-01, 1.96604111e-01, 1.98391421e-
01,
        1.98391421e-01, 1.99285076e-01, 1.99285076e-01, 2.01072386e-
01,
        2.01072386e-01, 2.01966041e-01, 2.01966041e-01, 2.02859696e-
01,
        2.02859696e-01, 2.09115282e-01, 2.09115282e-01, 2.11796247e-
01,
        2.11796247e-01, 2.12689902e-01, 2.12689902e-01, 2.18051832e-
01,
        2.18051832e-01, 2.21626452e-01, 2.21626452e-01, 2.24307417e-
01,
        2.24307417e-01, 2.25201072e-01, 2.25201072e-01, 2.26988382e-
01,
        2.26988382e-01, 2.28775693e-01, 2.29669348e-01, 2.29669348e-
01,
        2.35924933e-01, 2.35924933e-01, 2.38605898e-01, 2.38605898e-
01,
        2.40393208e-01, 2.40393208e-01, 2.41286863e-01, 2.41286863e-
01,
        2.42180518e-01, 2.42180518e-01, 2.44861483e-01, 2.44861483e-
01,
        2.45755139e-01, 2.45755139e-01, 2.47542449e-01, 2.47542449e-
01,
        2.50223414e-01, 2.50223414e-01, 2.51117069e-01, 2.52904379e-
01,
        2.52904379e-01, 2.54691689e-01, 2.54691689e-01, 2.55585344e-
01,
        2.55585344e-01, 2.62734584e-01, 2.62734584e-01, 2.63628239e-
01,
        2.63628239e-01, 2.66309205e-01, 2.66309205e-01, 2.70777480e-
01,
        3.08310992e-01, 3.10098302e-01, 3.14566577e-01, 3.14566577e-
01,
        3.18141197e-01, 3.18141197e-01, 3.21715818e-01, 3.21715818e-
01,
```

```
        3.27971403e-01, 3.27971403e-01, 3.28865058e-01, 3.28865058e-
01,
        3.33333333e-01, 3.33333333e-01, 3.36014298e-01, 3.58355675e-
01,
        3.58355675e-01, 3.63717605e-01, 3.63717605e-01, 3.64611260e-
01,
        3.67292225e-01, 3.71760500e-01, 3.71760500e-01, 3.73547811e-
01,
        3.73547811e-01, 3.76228776e-01, 3.76228776e-01, 3.92314567e-
01,
        3.92314567e-01, 3.95889187e-01, 3.95889187e-01, 3.96782842e-
01,
        3.96782842e-01, 3.99463807e-01, 3.99463807e-01, 4.03038427e-
01,
        4.03038427e-01, 4.17336908e-01, 4.17336908e-01, 4.20911528e-
01,
        4.20911528e-01, 4.22698838e-01, 4.22698838e-01, 4.30741734e-
01,
        4.30741734e-01, 4.44146559e-01, 4.44146559e-01, 4.45040214e-
01,
        4.53083110e-01, 4.53083110e-01, 4.53976765e-01, 4.53976765e-
01,
        4.63806971e-01, 4.63806971e-01, 4.67381591e-01, 4.67381591e-
01,
        4.72743521e-01, 4.72743521e-01, 4.74530831e-01, 4.77211796e-
01,
        4.78105451e-01, 4.87042002e-01, 4.87042002e-01, 4.88829312e-
01,
        4.88829312e-01, 4.95978552e-01, 4.95978552e-01, 4.97765862e-
01,
        4.97765862e-01, 4.98659517e-01, 5.00446828e-01, 5.02234138e-
01,
        5.02234138e-01, 5.04021448e-01, 5.04021448e-01, 5.29937444e-
01,
        5.29937444e-01, 5.35299374e-01, 5.35299374e-01, 5.37980340e-
01,
        5.37980340e-01, 5.42448615e-01, 5.42448615e-01, 5.47810545e-
01,
        5.47810545e-01, 5.61215371e-01, 5.61215371e-01, 5.63896336e-
01,
        5.66577301e-01, 5.73726542e-01, 5.73726542e-01, 5.75513852e-
01,
        5.75513852e-01, 5.96067918e-01, 5.96067918e-01, 6.12153709e-
01,
        6.12153709e-01, 6.14834674e-01, 6.14834674e-01, 6.19302949e-
01,
        6.19302949e-01, 6.21090259e-01, 6.21090259e-01, 6.24664879e-
01,
        6.24664879e-01, 6.34495085e-01, 6.34495085e-01, 6.38069705e-
```

```
01,
        6.38069705e-01, 6.65773012e-01, 6.65773012e-01, 6.69347632e-
01,
        6.69347632e-01, 7.00625559e-01, 7.14924039e-01, 7.19392315e-
01,
        7.19392315e-01, 7.37265416e-01, 7.42627346e-01, 7.44414656e-
01,
        7.44414656e-01, 7.47989276e-01, 7.47989276e-01, 7.48882931e-
01,
        7.48882931e-01, 7.50670241e-01, 7.50670241e-01, 7.52457551e-
01,
        7.52457551e-01, 7.85522788e-01, 8.60589812e-01, 8.76675603e-
01,
        8.76675603e-01, 8.98123324e-01, 9.05272565e-01, 9.05272565e-
01,
        9.27613941e-01, 9.27613941e-01, 9.59785523e-01, 9.59785523e-
01,
        9.99106345e-01, 9.99106345e-01, 1.00000000e+00]),
 array([0.        , 0.00137174, 0.00823045, 0.01097394, 0.01234568,
        0.01783265, 0.03566529, 0.04252401, 0.06447188, 0.06858711,
        0.07544582, 0.0781893 , 0.12208505, 0.13443073, 0.14128944,
        0.16872428, 0.17283951, 0.20987654, 0.20987654, 0.22085048,
        0.22085048, 0.23182442, 0.23182442, 0.26748971, 0.27846365,
        0.28806584, 0.29766804, 0.30178326, 0.30452675, 0.30864198,
        0.30864198, 0.3127572 , 0.31550069, 0.31824417, 0.32098765,
        0.33607682, 0.33607682, 0.34019204, 0.34019204, 0.34156379,
        0.34567901, 0.35116598, 0.35116598, 0.3648834 , 0.3648834 ,
        0.38683128, 0.38683128, 0.38957476, 0.38957476, 0.39231824,
        0.39231824, 0.41838134, 0.41838134, 0.42524005, 0.42524005,
        0.42798354, 0.42798354, 0.43621399, 0.43621399, 0.44444444,
        0.44444444, 0.45679012, 0.45679012, 0.46090535, 0.46090535,
        0.46776406, 0.46776406, 0.4691358 , 0.4691358 , 0.47599451,
        0.47599451, 0.48010974, 0.48010974, 0.48559671, 0.48559671,
        0.48971193, 0.48971193, 0.49382716, 0.49382716, 0.4951989 ,
        0.4951989 , 0.49931413, 0.49931413, 0.50205761, 0.50205761,
        0.50617284, 0.50617284, 0.50754458, 0.50754458, 0.51028807,
        0.51028807, 0.51303155, 0.51303155, 0.51851852, 0.51851852,
        0.521262  , 0.52263374, 0.52263374, 0.52537723, 0.52812071,
        0.53360768, 0.53360768, 0.53909465, 0.53909465, 0.54320988,
        0.54320988, 0.5569273 , 0.5569273 , 0.55967078, 0.55967078,
        0.56515775, 0.56652949, 0.56652949, 0.57064472, 0.57064472,
        0.57613169, 0.57613169, 0.57887517, 0.57887517, 0.58161866,
        0.58161866, 0.5829904 , 0.5829904 , 0.58984911, 0.58984911,
        0.60082305, 0.60082305, 0.60356653, 0.60356653, 0.60768176,
        0.60768176, 0.61042524, 0.61042524, 0.61316872, 0.61316872,
        0.61454047, 0.61454047, 0.61728395, 0.61728395, 0.62002743,
        0.62002743, 0.62277092, 0.62277092, 0.6255144 , 0.6255144 ,
        0.62825789, 0.62825789, 0.63100137, 0.63100137, 0.63237311,
        0.63237311, 0.63923182, 0.63923182, 0.64197531, 0.64197531,
```

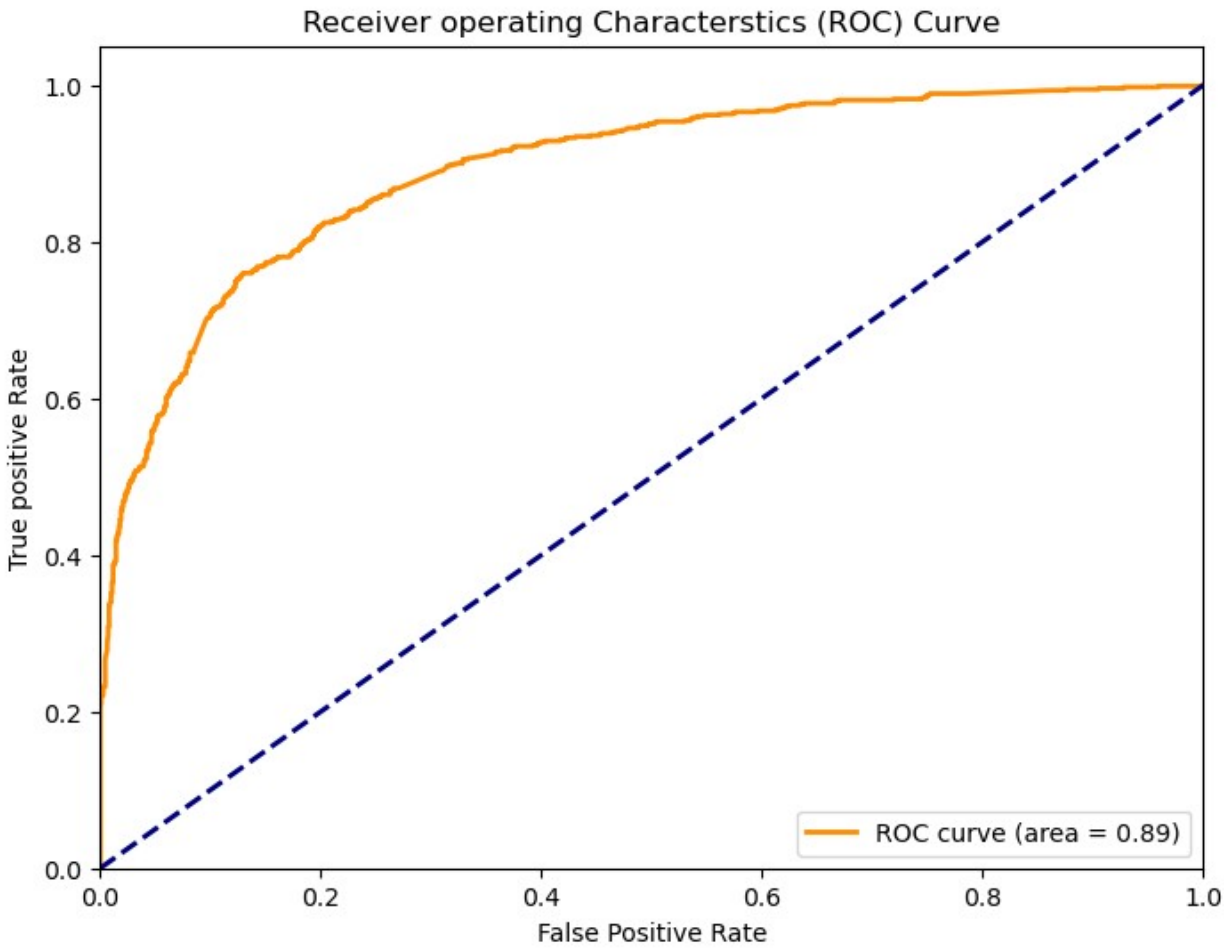
```
        0.64746228, 0.64746228, 0.65843621, 0.65843621, 0.6611797 ,
        0.70233196, 0.70233196, 0.7037037 , 0.7037037 , 0.70507545,
        0.70507545, 0.70919067, 0.70919067, 0.71056241, 0.71056241,
        0.7133059 , 0.7133059 , 0.71467764, 0.71467764, 0.71604938,
        0.71604938, 0.71742112, 0.71742112, 0.71879287, 0.71879287,
        0.72153635, 0.72153635, 0.72702332, 0.72702332, 0.7297668 ,
        0.7297668 , 0.73113855, 0.73113855, 0.73251029, 0.73251029,
        0.73525377, 0.73525377, 0.73662551, 0.73662551, 0.73799726,
        0.73799726, 0.74074074, 0.74074074, 0.74211248, 0.74211248,
        0.74897119, 0.74897119, 0.75034294, 0.75034294, 0.75308642,
        0.75308642, 0.7558299 , 0.7558299 , 0.75720165, 0.75720165,
        0.75857339, 0.75857339, 0.75994513, 0.75994513, 0.76268861,
        0.76268861, 0.76406036, 0.76406036, 0.76817558, 0.76817558,
        0.76954733, 0.76954733, 0.77091907, 0.77091907, 0.77366255,
        0.77366255, 0.77503429, 0.77503429, 0.77777778, 0.77777778,
        0.77914952, 0.77914952, 0.78052126, 0.78052126, 0.781893  ,
        0.781893  , 0.78326475, 0.78326475, 0.78463649, 0.78463649,
        0.78737997, 0.78737997, 0.78875171, 0.78875171, 0.79012346,
        0.79012346, 0.79286694, 0.79286694, 0.79561043, 0.79561043,
        0.79698217, 0.79698217, 0.79835391, 0.79835391, 0.79972565,
        0.79972565, 0.80109739, 0.80109739, 0.80246914, 0.80246914,
        0.80384088, 0.80384088, 0.80658436, 0.80658436, 0.80932785,
        0.80932785, 0.81344307, 0.81344307, 0.81481481, 0.81481481,
        0.8175583 , 0.8175583 , 0.81893004, 0.81893004, 0.82030178,
        0.82030178, 0.82167353, 0.82167353, 0.82304527, 0.82304527,
        0.82441701, 0.82441701, 0.82578875, 0.82578875, 0.82716049,
        0.82716049, 0.82853224, 0.82853224, 0.82990398, 0.82990398,
        0.83264746, 0.83264746, 0.8340192 , 0.8340192 , 0.83813443,
        0.83813443, 0.83950617, 0.83950617, 0.83950617, 0.84087791,
        0.84087791, 0.8436214 , 0.8436214 , 0.84499314, 0.84499314,
        0.84636488, 0.84636488, 0.84910837, 0.84910837, 0.85048011,
        0.85048011, 0.85185185, 0.85185185, 0.85322359, 0.85322359,
        0.85459534, 0.85459534, 0.85596708, 0.85596708, 0.85596708,
        0.85733882, 0.85733882, 0.85871056, 0.85871056, 0.8600823 ,
        0.8600823 , 0.86419753, 0.86419753, 0.86556927, 0.86556927,
        0.86831276, 0.86831276, 0.89026063, 0.89026063, 0.8957476 ,
        0.89711934, 0.89711934, 0.89849108, 0.89849108, 0.89986283,
        0.89986283, 0.90123457, 0.90123457, 0.90534979, 0.90534979,
        0.90672154, 0.90672154, 0.9122085 , 0.91495199, 0.91495199,
        0.91632373, 0.91632373, 0.91632373, 0.91632373, 0.91769547,
        0.91769547, 0.92043896, 0.92043896, 0.9218107 , 0.9218107 ,
        0.92318244, 0.92318244, 0.92455418, 0.92455418, 0.92592593,
        0.92592593, 0.92729767, 0.92729767, 0.92866941, 0.92866941,
        0.93004115, 0.93004115, 0.93141289, 0.93141289, 0.93278464,
        0.93278464, 0.93415638, 0.93415638, 0.93552812, 0.93552812,
        0.93552812, 0.93689986, 0.93689986, 0.9382716 , 0.9382716 ,
        0.93964335, 0.93964335, 0.94101509, 0.94101509, 0.94238683,
        0.94238683, 0.94375857, 0.94513032, 0.94513032, 0.94650206,
        0.94650206, 0.9478738 , 0.9478738 , 0.94924554, 0.94924554,
```

```
       0.95061728, 0.95061728, 0.95061728, 0.95061728, 0.95198903,
       0.95198903, 0.95336077, 0.95336077, 0.95473251, 0.95473251,
       0.95747599, 0.95747599, 0.95884774, 0.95884774, 0.96021948,
       0.96021948, 0.96159122, 0.96159122, 0.96296296, 0.96296296,
       0.96296296, 0.96296296, 0.96433471, 0.96433471, 0.96570645,
       0.96570645, 0.96707819, 0.96707819, 0.96844993, 0.96844993,
       0.96982167, 0.96982167, 0.97119342, 0.97119342, 0.97256516,
       0.97256516, 0.9739369 , 0.9739369 , 0.97530864, 0.97530864,
       0.97668038, 0.97668038, 0.97942387, 0.97942387, 0.98079561,
       0.98079561, 0.98079561, 0.98079561, 0.98216735, 0.98216735,
       0.98216735, 0.98216735, 0.98353909, 0.98353909, 0.98491084,
       0.98491084, 0.98628258, 0.98628258, 0.98765432, 0.98765432,
       0.98902606, 0.98902606, 0.99314129, 0.99314129, 0.99451303,
       0.99451303, 0.99451303, 0.99588477, 0.99588477, 0.99725652,
       0.99725652, 0.99862826, 0.99862826, 1.        , 1.        ]),
array([       inf, 0.99867741, 0.99727558, 0.9971332 , 0.99697813,
       0.99694509, 0.99240271, 0.99236529, 0.98863222, 0.98850897,
       0.98750585, 0.987457  , 0.97285501, 0.97164183, 0.97129726,
       0.96259135, 0.96235673, 0.93965157, 0.93886609, 0.92854865,
       0.92814582, 0.9222488 , 0.9216677 , 0.89479778, 0.8942288 ,
       0.88468422, 0.88356408, 0.88037822, 0.87935033, 0.87834845,
       0.87697205, 0.87521195, 0.87200732, 0.87094383, 0.87078497,
       0.8660052 , 0.86519123, 0.8620354 , 0.86197634, 0.86118389,
       0.86047686, 0.85663387, 0.85609627, 0.84567428, 0.84504746,
       0.83427428, 0.83203693, 0.83027669, 0.8300152 , 0.82626989,
       0.82548778, 0.80547534, 0.80272649, 0.79754866, 0.79584939,
       0.79258135, 0.79173124, 0.78610865, 0.78453038, 0.77722227,
       0.77690156, 0.7724383 , 0.77122149, 0.76743136, 0.7670699 ,
       0.76151407, 0.76042651, 0.76020562, 0.75977018, 0.75123827,
       0.75072924, 0.74887464, 0.74636301, 0.74150067, 0.73905296,
       0.7378973 , 0.73751964, 0.73537327, 0.7347144 , 0.73400162,
       0.73311687, 0.72975244, 0.72713909, 0.72684563, 0.72632285,
       0.72154955, 0.71832393, 0.71753771, 0.71474183, 0.71323568,
       0.71311503, 0.70648827, 0.70296726, 0.69349571, 0.692237  ,
       0.69026733, 0.6896403 , 0.68815921, 0.68802507, 0.6876031 ,
       0.67886829, 0.67846976, 0.67459644, 0.67320852, 0.67115371,
       0.66985216, 0.66174391, 0.65901908, 0.65401892, 0.6522847 ,
       0.64807086, 0.64757009, 0.64756069, 0.64553977, 0.64535203,
       0.64292258, 0.64230706, 0.63998968, 0.63680384, 0.63671946,
       0.63667507, 0.63662076, 0.6334839 , 0.62333355, 0.62063996,
       0.61066438, 0.61061614, 0.60983989, 0.60894992, 0.60614269,
       0.60504886, 0.6042789 , 0.60363422, 0.60244474, 0.6019942 ,
       0.6010103 , 0.5998923 , 0.59532697, 0.59427148, 0.59234938,
       0.58898423, 0.58811867, 0.58694918, 0.58419199, 0.58351315,
       0.58282447, 0.58266517, 0.58008771, 0.57559408, 0.57462406,
       0.57397058, 0.56796887, 0.56716362, 0.5631482 , 0.56231023,
       0.55780548, 0.5563331 , 0.54499749, 0.54309714, 0.53797062,
       0.53653527, 0.5299723 , 0.52807514, 0.52797104, 0.5256261 ,
       0.52327685, 0.52057658, 0.52021117, 0.51915333, 0.51873384,
```

```
      0.51709929, 0.51668064, 0.5159464 , 0.51515231, 0.51456854,
      0.51078702, 0.51046427, 0.50312488, 0.50274093, 0.50235702,
      0.50174113, 0.50014253, 0.49566514, 0.49545777, 0.49319943,
      0.49198341, 0.4917223 , 0.48962825, 0.48958487, 0.48282308,
      0.48257913, 0.48208446, 0.48128409, 0.48066156, 0.47953926,
      0.47611844, 0.47422683, 0.4726503 , 0.4718146 , 0.47031329,
      0.46531786, 0.46503276, 0.46469441, 0.46437955, 0.46161234,
      0.45733904, 0.45442128, 0.45368197, 0.45231305, 0.45220639,
      0.45152799, 0.45123581, 0.45056841, 0.44056066, 0.43817957,
      0.43617546, 0.43614906, 0.43372179, 0.4317607 , 0.42890067,
      0.42726186, 0.42470343, 0.4242017 , 0.42404405, 0.42040469,
      0.4174724 , 0.41738956, 0.41536443, 0.41505177, 0.40793483,
      0.40666733, 0.40603776, 0.40553297, 0.39296277, 0.39155551,
      0.39046617, 0.38879779, 0.38547711, 0.38391225, 0.38376997,
      0.38220083, 0.37942546, 0.37926725, 0.37521661, 0.37497792,
      0.37461862, 0.37378519, 0.37370234, 0.37267037, 0.37058172,
      0.37004174, 0.36848585, 0.36834727, 0.36711042, 0.36481361,
      0.36436752, 0.36427895, 0.36220457, 0.36060773, 0.35926657,
      0.35822744, 0.35627988, 0.3517562 , 0.35169533, 0.34895458,
      0.34880445, 0.34716842, 0.34562433, 0.34543077, 0.34435213,
      0.34287603, 0.34148426, 0.34075613, 0.34068531, 0.33946606,
      0.33555224, 0.3348827 , 0.33480067, 0.33423216, 0.33408804,
      0.33201317, 0.32743757, 0.32675347, 0.32313076, 0.32235897,
      0.32139919, 0.32030541, 0.31804741, 0.31741883, 0.31396559,
      0.31259405, 0.31066968, 0.31010452, 0.3096563 , 0.30309234,
      0.30144656, 0.30085897, 0.29994265, 0.29934146, 0.29878646,
      0.28929906, 0.28880059, 0.28823423, 0.28814901, 0.2875007 ,
      0.28704277, 0.28690876, 0.28457772, 0.28373403, 0.28284958,
      0.28150984, 0.28031116, 0.28006721, 0.27892532, 0.2736208 ,
      0.27354946, 0.26657667, 0.26580538, 0.2638231 , 0.26104098,
      0.2602554 , 0.25881496, 0.2564445 , 0.25552735, 0.25462759,
      0.24745159, 0.24586833, 0.24469798, 0.24469251, 0.24161559,
      0.23985763, 0.23574425, 0.23381168, 0.23226071, 0.23159192,
      0.23001016, 0.22425692, 0.22359889, 0.21913504, 0.21886066,
      0.21133013, 0.21039489, 0.21031503, 0.20835929, 0.2034685 ,
      0.20294378, 0.20173409, 0.20102408, 0.19740937, 0.19522568,
      0.19427739, 0.19365076, 0.19326689, 0.19202237, 0.19163603,
      0.19090395, 0.19044729, 0.18948235, 0.18860389, 0.18139602,
      0.18002249, 0.17883795, 0.17857074, 0.17843416, 0.17817002,
      0.17542306, 0.17442791, 0.17216712, 0.17103009, 0.16178426,
      0.16148019, 0.16093276, 0.16060714, 0.15691705, 0.15628487,
      0.1532015 , 0.15284846, 0.14861492, 0.14853201, 0.14775876,
      0.14735976, 0.14719611, 0.1470795 , 0.14635678, 0.14341461,
      0.14339879, 0.14239635, 0.1422943 , 0.14060283, 0.14007892,
      0.13985382, 0.13967508, 0.13965485, 0.13508292, 0.13467307,
      0.13442645, 0.13399048, 0.13193601, 0.13158818, 0.13087949,
      0.13069394, 0.13062603, 0.13050118, 0.13027792, 0.13015984,
      0.12996664, 0.12983711, 0.1233706 , 0.12333772, 0.12213048,
      0.12173398, 0.12088424, 0.12058235, 0.11933262, 0.11928008,
```

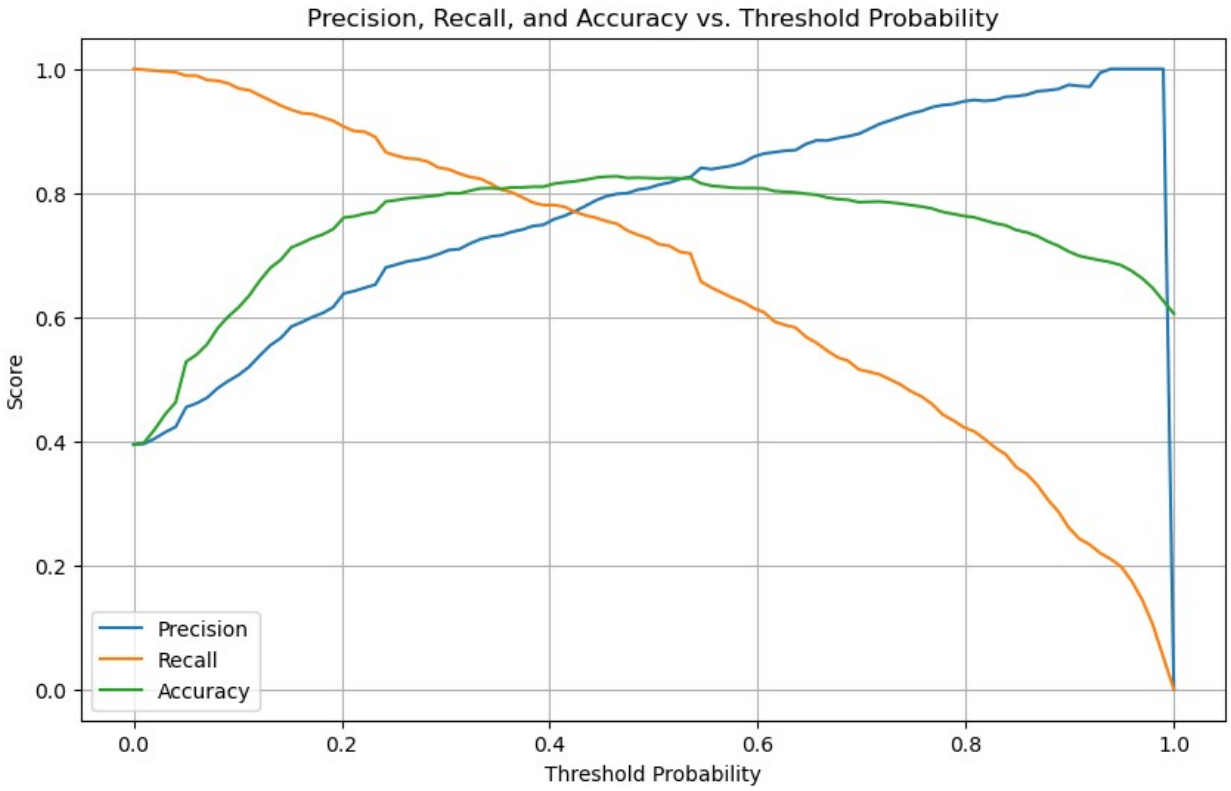
```
       0.11792582, 0.11790738, 0.11590788, 0.11557374, 0.11454631,
       0.11450867, 0.1127628 , 0.11248035, 0.11199706, 0.11191667,
       0.10514414, 0.10507769, 0.10223416, 0.1020534 , 0.10044865,
       0.10013253, 0.09933178, 0.09903781, 0.09780641, 0.0974009 ,
       0.09653544, 0.09642922, 0.09260337, 0.09211502, 0.09181324,
       0.09180827, 0.08454941, 0.08386563, 0.08366787, 0.08356076,
       0.07407482, 0.07396966, 0.07140442, 0.07112689, 0.06684519,
       0.06514709, 0.06443979, 0.06350851, 0.06306308, 0.06269953,
       0.06238115, 0.06225203, 0.06184215, 0.0617278 , 0.0611256 ,
       0.06107428, 0.04788416, 0.04728681, 0.04355068, 0.04341003,
       0.03562926, 0.03548801, 0.03477724, 0.02759991, 0.02697147,
       0.02012578, 0.01947372, 0.00449825, 0.00395632, 0.00343469]))
```

```
from sklearn.metrics import auc
roc_auc = auc(fpr, tpr)

#plot roc-auc curve
plt.figure(figsize=(8,6))
plt.plot(fpr,tpr,color='darkorange',linewidth=2,label='ROC curve (area
= %0.2f)'% roc_auc)
plt.plot([0,1],[0,1], color ='navy', linewidth=2,linestyle='--')
plt.xlim([0.0,1.0])
plt.ylim([0.0,1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True positive Rate')
plt.title('Receiver operating Characterstics (ROC) Curve')
plt.legend(loc='lower right')
plt.show()
```

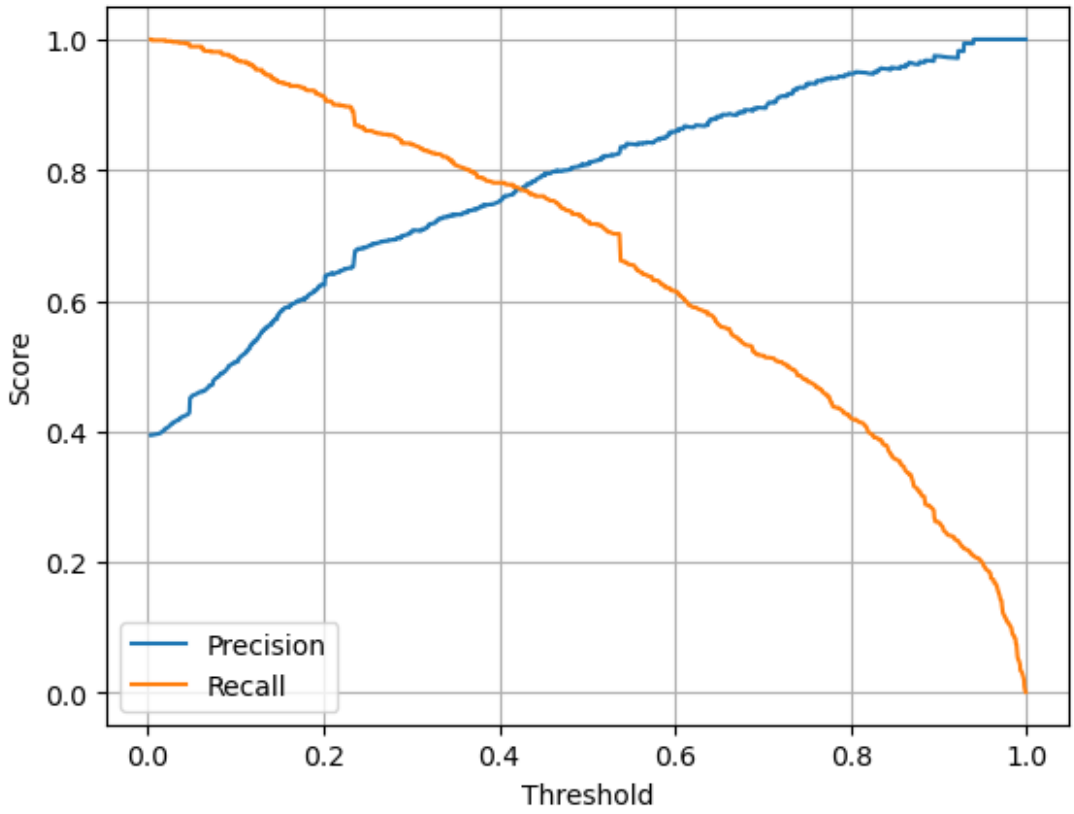



```
from sklearn.metrics import precision_score, recall_score,
accuracy_score


# Calculate precision, recall, and accuracy for different threshold
probabilities
thresholds = np.linspace(0, 1, 100)
precisions = []
recalls = []
accuracies = []

for threshold in thresholds:
    y_pred_threshold = (y_predict_proba >= threshold).astype(int)
    precision = precision_score(y_test, y_pred_threshold)
    recall = recall_score(y_test, y_pred_threshold)
    accuracy = accuracy_score(y_test, y_pred_threshold)
    precisions.append(precision)
    recalls.append(recall)
    accuracies.append(accuracy)
# Plot precision, recall, and accuracy against threshold probabilities
plt.figure(figsize=(10, 6))
```

```
plt.plot(thresholds, precisions, label='Precision')
plt.plot(thresholds, recalls, label='Recall')
plt.plot(thresholds, accuracies, label='Accuracy')
plt.xlabel('Threshold Probability')
plt.ylabel('Score')
plt.title('Precision, Recall, and Accuracy vs. Threshold Probability')
plt.legend()
plt.grid(True)
plt.show()
```




```
from sklearn.metrics import precision_recall_curve

precision, recall, thresholds = precision_recall_curve(y_test,
y_predict_proba)

plt.plot(thresholds, precision[:-1], label='Precision')
plt.plot(thresholds, recall[:-1], label='Recall')
plt.xlabel('Threshold')
plt.ylabel('Score')
plt.legend()
plt.grid(True)
plt.show()
```



This Graph signifies:

X-axis (Threshold Probability):

As we increase the threshold, the model becomes more conservative in predicting the positive class (lead conversion), requiring higher confidence. Y-axis (Metric Values - Precision, Recall, Accuracy):

Precision (Blue Line): Increases as the threshold increases. This means fewer false positives, but we may miss actual positive cases. Recall (Orange Line): Decreases as the threshold increases. This means we are capturing fewer actual positives but with more precision. Accuracy (Green Line): Shows how well the model is performing overall. It initially improves but may decline at extreme thresholds. Accuracy Behavior – Accuracy doesn't necessarily follow a simple linear relationship because it depends on the dataset's class distribution.

```
# If X_train is a sparse matrix, convert it to a DataFrame
X_train_dense = X_train.toarray()  # Convert to dense format
feature_names = preprocessor.get_feature_names_out()
X_train_df = pd.DataFrame(X_train_dense, columns=feature_names)  #
Convert to DataFrame

# Now, you can access columns
feature_names = X_train_df.columns
```

```
# Extract coefficients
coefficients = classifier.coef_[0]

# Create DataFrame for feature importance

feature_importance = pd.DataFrame({'Feature': X_train_df.columns,
'Importance': np.abs(coefficients)})
feature_importance = feature_importance.sort_values(by='Importance',
ascending=False)

# Display top 10 important features
print(feature_importance.head(10))
```

```
                                                Feature  Importance
4                         cat__Lead Origin_Lead Add Form    2.325670
14                     cat__Lead Source_Welingak Website    2.257280
16                      cat__Last Activity_Email Bounced    1.545033
87   cat__What is your current occupation_Working P...    1.403877
23           cat__Last Activity_Olark Chat Conversation    1.294039
101  cat__Last Notable Activity_Had a Phone Convers...    1.190140
86     cat__What is your current occupation_Unemployed    1.098572
1                      num__Total Time Spent on Website    1.094744
77                          cat__Specialization_Others    1.085707
49                                   cat__Country_Oman    1.069819
```

TOP 3 features

Lead Origin, wherein Lead Add Form has highest conversion.

Lead Source, wherein Welingak Website and reference has highest conversion.

Last Activity, wherein Olark Chat Conversation has highest conversion.

```
#Case when Company wish to make the lead conversion more aggressive.
#They want almost all the potential leads (i.e., the customers who
have been predicted as 1 by the model) to be converted and hence,
#want to make phone calls to as much of such people as possible.

threshold = 0.3  # Reduce threshold for aggressive lead conversion and
decreasing false negatives
y_pred_adjusted = (y_predict_proba >= threshold).astype(int)
accuracy_score(y_test, y_pred_adjusted)
```

```
0.7992424242424242
```

```
threshold = 0.7  # Increase threshold to avoid false positives
y_pred_adjusted = (y_predict_proba >= threshold).astype(int)
accuracy_score(y_test, y_pred_adjusted)
```

```
0.7851731601731602
```

X-Education has a better chance of converting a potential lead when:

1. The Lead origin is Lead add form: Leads who have responded/ or engaged through Lead Add Forms have had a higher chances of getting converted
2. It can be further said conversion rate is high for the leads that the X education get from Welingak Website and references, we can work on these two sources to get good quality leads
3. The leads which are actively interacting with the X_education through Olark chat and telephonic conversation have good conversion.
4. Leads who are working professionals have high chances of getting converted.
5. People who were looking for better prospects like Unemployed, students, also show a higher interest in taking up courses.

Conclusion

In conclusion, the logistic regression model we developed proved to be a superior lead scoring model. In nearly 82% of cases, it correctly assigns a higher lead score to leads that will convert compared to a lead who will not convert. By using this lead scoring model, the sales team can increase their conversion rate to 82% by focusing on the quality features that we get from the model. As a recommended next step for X Education, it would be valuable to determine a minimum lead scorefor sales representatives to bother contacting a lead. This can be done after the cost of having a sales representatives contact a lead, as well as the value of a converted lead, has been determined. Using a profit matrix, the optimal threshold for classification to maximize profit can be identified.