logistic-regression-model

August 9, 2023

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Goals of the Case Study There are quite a few goals for this case study. 1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. 2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

[1]:
```
# Supress unnecessary warnings
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
# Import the NumPy and Pandas packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[3]:
```
#  dataset
leads = pd.read_csv('Leads.csv')
```

[4]:
```
# Look at the first few entries

leads.head()
```

[4]:
```
                            Prospect ID  Lead Number              Lead Origin  \
0  7927b2df-8bba-4d29-b9a2-b6e0beafe620       660737                      API
1  2a272436-5132-4136-86fa-dcc88c88f482       660728                      API
2  8cc8c611-a219-4f35-ad23-fdfd2656bd8a       660727  Landing Page Submission
3  0cc2df48-7cf4-4e39-9de9-19797f9b38cc       660719  Landing Page Submission
4  3256f628-e534-4826-9d63-4a8b88782852       660681  Landing Page Submission
```

```
     Lead Source Do Not Email Do Not Call  Converted  TotalVisits  \
0     Olark Chat           No          No          0          0.0
1  Organic Search          No          No          0          5.0
2  Direct Traffic          No          No          1          2.0
3  Direct Traffic          No          No          0          1.0
4          Google          No          No          1          2.0

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

  Get updates on DM Content    Lead Profile    City  \
0                        No          Select  Select
1                        No          Select  Select
2                        No  Potential Lead  Mumbai
3                        No          Select  Mumbai
4                        No          Select  Mumbai

  Asymmetrique Activity Index Asymmetrique Profile Index  \
0                  02.Medium                  02.Medium
1                  02.Medium                  02.Medium
2                  02.Medium                    01.High
3                  02.Medium                    01.High
4                  02.Medium                    01.High

  Asymmetrique Activity Score Asymmetrique Profile Score  \
0                       15.0                       15.0
1                       15.0                       15.0
2                       14.0                       20.0
3                       13.0                       17.0
4                       15.0                       18.0

  I agree to pay the amount through cheque  \
0                                       No
1                                       No
2                                       No
3                                       No
4                                       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 37 columns]
```

[5]:
```python
# Inspect the shape of the dataset

leads.shape
```

[5]:
```
(9240, 37)
```

[6]:
```python
# Inspect the different columsn in the dataset

leads.columns
```

[6]:
```
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')
```

As you can see, the feature variables are quite intuitive. If you don't understand them completely, please refer to the data dictionary.

[7]:
```python
# Check the summary of the dataset

leads.describe()
```

[7]:
```
         Lead Number    Converted  TotalVisits  Total Time Spent on Website  \
count    9240.000000  9240.000000  9103.000000                  9240.000000
mean   617188.435606     0.385390     3.445238                   487.698268
std     23405.995698     0.486714     4.854853                   548.021466
min    579533.000000     0.000000     0.000000                     0.000000
25%    596484.500000     0.000000     1.000000                    12.000000
50%    615479.000000     0.000000     3.000000                   248.000000
75%    637387.250000     1.000000     5.000000                   936.000000
max    660737.000000     1.000000   251.000000                  2272.000000
```

```
       Page Views Per Visit  Asymmetrique Activity Score  \
count           9103.000000                  5022.000000
mean               2.362820                    14.306252
std                2.161418                     1.386694
min                0.000000                     7.000000
25%                1.000000                    14.000000
50%                2.000000                    14.000000
75%                3.000000                    15.000000
max               55.000000                    18.000000

       Asymmetrique Profile Score
count                 5022.000000
mean                    16.344883
std                      1.811395
min                     11.000000
25%                     15.000000
50%                     16.000000
75%                     18.000000
max                     20.000000
```

[8]:
```python
leads.describe(include=np.object)
```

[8]:
```
                                 Prospect ID              Lead Origin  \
count                                   9240                     9240
unique                                  9240                        5
top     7927b2df-8bba-4d29-b9a2-b6e0beafe620  Landing Page Submission
freq                                       1                     4886

       Lead Source Do Not Email Do Not Call Last Activity Country  \
count         9204         9240        9240          9137    6779
unique          21            2           2            17      38
top         Google           No          No  Email Opened   India
freq          2868         8506        9238          3437    6492

       Specialization How did you hear about X Education  \
count            7802                               7033
unique             19                                 10
top            Select                             Select
freq             1942                               5043

       What is your current occupation  … Lead Quality  \
count                             6550  …         4473
unique                               6  …            5
top                         Unemployed  …     Might be
freq                              5600  …         1560
```

```
       Update me on Supply Chain Content Get updates on DM Content  \
count                                9240                      9240
unique                                  1                         1
top                                    No                        No
freq                                 9240                      9240

       Lead Profile    City Asymmetrique Activity Index  \
count          6531    7820                        5022
unique            6       7                           3
top          Select  Mumbai                   02.Medium
freq           4146    3222                        3839

       Asymmetrique Profile Index I agree to pay the amount through cheque  \
count                        5022                                      9240
unique                          3                                         1
top                     02.Medium                                        No
freq                         2788                                      9240

       A free copy of Mastering The Interview Last Notable Activity
count                                    9240                  9240
unique                                      2                    16
top                                        No              Modified
freq                                     6352                  3407

[4 rows x 30 columns]
```

Looks like there are quite a few categorical variables present in this dataset for which we will need to create dummy variables. Also, there are a lot of null values present as well, so we will need to treat them accordingly.

0.1 Step 1: Data Cleaning and Preparation

[9]:
```python
print(leads.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
```

[10]:

```
#Checking missing values
print("_________________checking ,issng values in␣
↪%____________________________________")
print(leads.isnull().sum()/len(leads)*100)
```

```
_________________checking ,issng values in %____________________________________
Prospect ID                                     0.000000
Lead Number                                     0.000000
Lead Origin                                     0.000000
Lead Source                                     0.389610
Do Not Email                                    0.000000
Do Not Call                                     0.000000
Converted                                       0.000000
TotalVisits                                     1.482684
Total Time Spent on Website                     0.000000
Page Views Per Visit                            1.482684
Last Activity                                   1.114719
Country                                        26.634199
Specialization                                 15.562771
How did you hear about X Education             23.885281
What is your current occupation                29.112554
What matters most to you in choosing a course  29.318182
```

```
Search                                           0.000000
Magazine                                         0.000000
Newspaper Article                                0.000000
X Education Forums                               0.000000
Newspaper                                        0.000000
Digital Advertisement                            0.000000
Through Recommendations                          0.000000
Receive More Updates About Our Courses           0.000000
Tags                                            36.287879
Lead Quality                                    51.590909
Update me on Supply Chain Content                0.000000
Get updates on DM Content                        0.000000
Lead Profile                                    29.318182
City                                            15.367965
Asymmetrique Activity Index                     45.649351
Asymmetrique Profile Index                      45.649351
Asymmetrique Activity Score                     45.649351
Asymmetrique Profile Score                      45.649351
I agree to pay the amount through cheque         0.000000
A free copy of Mastering The Interview           0.000000
Last Notable Activity                            0.000000
dtype: float64
```

Observations: 1. So we can see that some columns has missing values more than 30% ,so these columns will not be usefull. 2. So going to drop the columns which has missing values more than 30%

[11]:
```
for i in leads.columns:
    if leads[i].isnull().sum() > 3000:
        leads.drop(i, 1, inplace=True)
print(leads.isnull().sum()/len(leads)*100)
```

```
Prospect ID                                      0.000000
Lead Number                                      0.000000
Lead Origin                                      0.000000
Lead Source                                      0.389610
Do Not Email                                     0.000000
Do Not Call                                      0.000000
Converted                                        0.000000
TotalVisits                                      1.482684
Total Time Spent on Website                      0.000000
Page Views Per Visit                             1.482684
Last Activity                                    1.114719
Country                                         26.634199
Specialization                                  15.562771
How did you hear about X Education              23.885281
What is your current occupation                 29.112554
What matters most to you in choosing a course   29.318182
```

```
Search                                       0.000000
Magazine                                     0.000000
Newspaper Article                            0.000000
X Education Forums                           0.000000
Newspaper                                    0.000000
Digital Advertisement                        0.000000
Through Recommendations                      0.000000
Receive More Updates About Our Courses       0.000000
Update me on Supply Chain Content            0.000000
Get updates on DM Content                    0.000000
Lead Profile                                29.318182
City                                        15.367965
I agree to pay the amount through cheque     0.000000
A free copy of Mastering The Interview       0.000000
Last Notable Activity                        0.000000
dtype: float64
```

0.2 ###Imputation of numerical columns

Numerical_COLUMNS= ['Lead Number', 'Converted', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']

Treating column "TotalVisits"

[12]:
```
#Applying logical condtion for colun TotalVisists
print("total missing value before treatment",leads["TotalVisits"].isnull().
 ↪sum())
Median_for_visits=leads.groupby('Lead Origin')['TotalVisits'].agg(np.median)
bool=leads["TotalVisits"].isnull()
leads["TotalVisits"][bool]=leads["Lead Origin"][bool].apply(lambda x:
 ↪Median_for_visits.loc[x]).values
print("total missing value after treatment are:",leads["TotalVisits"].isnull().
 ↪sum())
print("still 1 value are missing so goint to find the value ! We can treat one␣
 ↪value by fillna")
leads["TotalVisits"].fillna(method="bfill",inplace=True) #Used back fill to␣
 ↪fill that 1 missing values
print("Now total missing values are:",leads["TotalVisits"].isnull().sum())
```

```
total missing value before treatment 137
total missing value after treatment are: 1
still 1 value are missing so goint to find the value ! We can treat one value by
fillna
Now total missing values are: 0
```

Treating column "Page Views Per Visit"

Also lead origin may have different rate of values of total page visits

[13]:
```
#Applyinng logical conditions
print("total missing value before treatment",leads["Page Views Per Visit"].
 ↪isnull().sum())
print(leads.groupby('Lead Origin')['Page Views Per Visit'].agg(np.mean))
print("So yes every origing has different rate of page visits")
#Going to impute on this logic
Median_for_page_visits=leads.groupby('Lead Origin')['Page Views Per Visit'].
 ↪agg(np.median)
bool=leads["Page Views Per Visit"].isnull()
leads["Page Views Per Visit"][bool]=leads["Lead Origin"][bool].apply(lambda x:
 ↪Median_for_page_visits.loc[x]).values
print("total missing value after treatment are:",leads["Page Views Per Visit"].
 ↪isnull().sum())
leads["Page Views Per Visit"].fillna(method="bfill",inplace=True) #Used back␣
 ↪fill to fill that 1 missing values
print("Now total missing values are :",leads["Page Views Per Visit"].isnull().
 ↪sum())
```

```
total missing value before treatment 137
Lead Origin
API                        1.425335
Landing Page Submission    3.338555
Lead Add Form              0.145921
Lead Import                0.258065
Quick Add Form                  NaN
Name: Page Views Per Visit, dtype: float64
So yes every origing has different rate of page visits
total missing value after treatment are: 1
Now total missing values are : 0
```

0.3 ###Handling missing values for categorical columns

['Prospect ID', 'Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call', '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', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'Lead Profile', 'City', 'I agree to pay the amount through cheque', 'A free copy of Mastering The Interview', 'Last Notable Activity'] —

Treating column "Lead source"

[14]:
```
print(leads["Lead Source"].nunique())
print(leads["Lead Source"].value_counts())
#Applying logical condition to impute
leads["Lead Source"].fillna(value=leads["Lead Source"].mode()[0],inplace=True)
print("Now total missing value fro lead source are :",leads["Lead Source"].
 ↪isnull().sum())
```

```
21
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: Lead Source, dtype: int64
Now total missing value fro lead source are : 0
```

Handling missing values for column "Country"

[15]:
```
print(leads["Country"].isnull().sum())
print(leads["Country"].value_counts()) #Counting the values
#mode=leads.groupby(["Lead Origin"])["Country"].agg(pd.Series.mode)
```

```
2461
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: Country, dtype: int64
```

[16]:
```
#Simply we can impute with mode
mode_value=leads["Country"].agg(pd.Series.mode)[0]
leads['Country'].fillna(mode_value,inplace=True)
print("Now total missing values for country are :",leads["Country"].isnull().
  ↪sum())
```

```
Now total missing values for country are : 0
```

Handling missing values for column "Specialization"

[17]:
```
print(leads["Specialization"].nunique())
print("Total missing values before treatment : ",leads["Specialization"].
  ↪isnull().sum())
leads["Specialization"].fillna(value=leads["Specialization"].
  ↪mode()[0],inplace=True)
print("Now total missing values after treatment are : ",leads["Specialization"].
  ↪isnull().sum())
```

```
19
Total missing values before treatment :  1438
Now total missing values after treatment are :  0
```

Treating missing values for column "City"

[18]:
```
leads["City"].value_counts()
```

[18]:
```
Mumbai                         3222
Select                         2249
Thane & Outskirts               752
Other Cities                    686
Other Cities of Maharashtra     457
Other Metro Cities              380
Tier II Cities                   74
Name: City, dtype: int64
```

[19]:
```python
#We can impute city value for each country with their mode
mode_city=leads.groupby(["Country"])["City"].agg(pd.Series.mode)
mode_city
```

[19]:
```
Country
Asia/Pacific Region                                                     []
Australia                                                           Mumbai
Bahrain                                 [Other Cities, Thane & Outskirts]
Bangladesh                                                    Other Cities
Belgium                                       [Mumbai, Thane & Outskirts]
Canada                                                              Mumbai
China                                                     [Mumbai, Select]
Denmark                                                       Other Cities
France                  [Other Cities, Other Cities of Maharashtra, Ot…
Germany                 [Mumbai, Other Cities, Other Cities of Maharas…
Ghana                                                         Other Cities
Hong Kong               [Mumbai, Other Cities, Other Cities of Maharas…
India                                                               Mumbai
Indonesia                                      Other Cities of Maharashtra
Italy                                                         Other Cities
Kenya                                                         Other Cities
Kuwait                                              [Mumbai, Other Cities]
Liberia                                                 Other Metro Cities
Malaysia                                       Other Cities of Maharashtra
Netherlands                                   [Mumbai, Thane & Outskirts]
Nigeria                                                       Other Cities
Oman                                                [Mumbai, Other Cities]
Philippines                                         [Mumbai, Other Cities]
Qatar                                                               Mumbai
Russia                                                              Select
Saudi Arabia                                                  Other Cities
Singapore                                                           Select
South Africa                                                  Other Cities
Sri Lanka                                                           Select
Sweden                                                              Select
Switzerland                                                         Mumbai
Tanzania                                                Other Metro Cities
Uganda                                              [Other Cities, Select]
```

```
United Arab Emirates                                                                 Other Cities
United Kingdom                                                                             Mumbai
United States                                                                              Mumbai
Vietnam                                                                                    Mumbai
unknown                                                                                        []
Name: City, dtype: object
```

So we can see city column has not any correct information according to the leads countries so we can drop city because it has not such inforation which can deliver insights

[20]:
```
leads.drop(columns=["City"],axis=1,inplace=True)
```

Treating column "Lead Profile"

[21]:
```
leads["Lead Profile"].value_counts()
```

[21]:
```
Select                         4146
Potential Lead                 1613
Other Leads                     487
Student of SomeSchool           241
Lateral Student                  24
Dual Specialization Student      20
Name: Lead Profile, dtype: int64
```

[22]:
```
print("Total missing values are before treatment",leads["Lead Profile"].
 ↪isnull().sum())
mode_lead_profile=leads.groupby(["Lead Origin"])["Lead Profile"].agg(pd.Series.
 ↪mode)
bool=leads["Lead Profile"].isnull()
leads["Lead Profile"][bool]=leads["Lead Origin"][bool].apply(lambda x :
 ↪mode_lead_profile.loc[x]).values
print(leads["Lead Profile"].value_counts())
print("Total missing values after treatment are",leads["Lead Profile"].isnull().
 ↪sum())
```

```
Total missing values are before treatment 2709
Select                         6855
Potential Lead                 1613
Other Leads                     487
Student of SomeSchool           241
Lateral Student                  24
Dual Specialization Student      20
Name: Lead Profile, dtype: int64
Total missing values after treatment are 0
```

Treating column "How did you hear about X Education"

[23]:
```
leads["How did you hear about X Education"].nunique()
```

[23]: 10

[24]:
```
#Also need here logical condition according to lead origin
print("Total missing values before tretment",leads["How did you hear about X␣
↪Education"].isnull().sum())
mode_=leads.groupby(["Lead Origin"])["How did you hear about X Education"].
↪agg(pd.Series.mode)
bool=leads["How did you hear about X Education"].isnull()
leads["How did you hear about X Education"][bool]=leads["Lead Origin"][bool].
↪apply(lambda x :mode_.loc[x]).values
print(leads["How did you hear about X Education"].value_counts())
print("Total missing values after tretment",leads["How did you hear about X␣
↪Education"].isnull().sum())
```

```
Total missing values before tretment 2207
Select                     7250
Online Search               808
Word Of Mouth               348
Student of SomeSchool       310
Other                       186
Multiple Sources            152
Advertisements               70
Social Media                 67
Email                        26
SMS                          23
Name: How did you hear about X Education, dtype: int64
Total missing values after tretment 0
```

Treating column "What is your current occupation"

[25]:
```
leads["What is your current occupation"].unique()
```

[25]:
```
array(['Unemployed', 'Student', nan, 'Working Professional',
       'Businessman', 'Other', 'Housewife'], dtype=object)
```

[26]:
```
print("total null values for this column are :",leads["What is your current␣
↪occupation"].isnull().sum())
#Also need here logical condition according to lead origin, lead origin can␣
↪effect for this quetion "What is your current occupation"
print(leads.groupby(["Lead Origin"])["What is your current occupation"].agg(pd.
↪Series.mode))
#We can fill values here by fillna method as mode value is same for all sources
leads["What is your current occupation"].fillna(value=leads["What is your␣
↪current occupation"].mode()[0],inplace=True)
print("Now total null values for this column are :",leads["What is your current␣
↪occupation"].isnull().sum())
```

```
total null values for this column are : 2690
```

```
Lead Origin
API                           Unemployed
Landing Page Submission       Unemployed
Lead Add Form                 Unemployed
Lead Import                   Unemployed
Quick Add Form                Unemployed
Name: What is your current occupation, dtype: object
Now total null values for this column are : 0
```

Treating column "What matters most to you in choosing a course ?" We can also drop this column as all values are same

[27]:
```
print("TOtal null values are before treatment:",leads["What matters most to you␣
↪in choosing a course"].isnull().sum())
print(leads["What matters most to you in choosing a course"].value_counts())
print("So we can see <Better Career Prospects> almost all rows has so we can␣
↪impute the same")
leads["What matters most to you in choosing a course"].fillna(leads["What␣
↪matters most to you in choosing a course"].mode()[0],inplace=True)#used␣
↪fillna method
print("TOtal null values are :",leads["What matters most to you in choosing a␣
↪course"].isnull().sum())
```

```
TOtal null values are before treatment: 2709
Better Career Prospects      6528
Flexibility & Convenience       2
Other                           1
Name: What matters most to you in choosing a course, dtype: int64
So we can see <Better Career Prospects> almost all rows has so we can impute the
same
TOtal null values are : 0
```

Treating column "Converted"

[28]:
```
print(leads["Converted"].unique()) #So it has bool value like 0 or 1 either␣
↪converted means 1 or not converted means 0
#lets see if label of Occupation effect conversion rate or not
print(leads[["What is your current occupation","Converted"]].value_counts())
print("We can see the mode value but lets find seprately")
print(leads.groupby(["What is your current occupation"])["Converted"].agg(pd.
↪Series.mode))
```

```
[0 1]
What is your current occupation  Converted
Unemployed                       0            5479
                                 1            2811
Working Professional             1             647
Student                          0             132
                                 1              78
```

```
Working Professional          0           59
Housewife                     1           10
Other                         1           10
                              0            6
Businessman                   1            5
                              0            3
dtype: int64
We can see the mode value but lets find sepratly
What is your current occupation
Businessman             1
Housewife               1
Other                   1
Student                 0
Unemployed              0
Working Professional    1
Name: Converted, dtype: int64
```

Treating column "Last Activity"

[29]:
```
print("Total issing values in this column are :",leads["Last Activity"].
 ↪isnull().sum())
print(leads["Last Activity"].unique())
leads["Last Activity"].fillna(value=leads["Last Activity"].
 ↪mode()[0],inplace=True)
print("____________________Now totall issing values in this column are :
 ↪",leads["Last Activity"].isnull().sum())
```

```
Total issing values in this column are : 103
['Page Visited on Website' 'Email Opened' 'Unreachable'
 'Converted to Lead' 'Olark Chat Conversation' 'Email Bounced'
 'Email Link Clicked' 'Form Submitted on Website' 'Unsubscribed'
 'Had a Phone Conversation' 'View in browser link Clicked' nan
 'Approached upfront' 'SMS Sent' 'Visited Booth in Tradeshow'
 'Resubscribed to emails' 'Email Received' 'Email Marked Spam']
____________________Now totall issing values in this column are : 0
```

[30]:
```
leads.isnull().sum()
```

[30]:
```
Prospect ID                                      0
Lead Number                                      0
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
How did you hear about X Education                 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
Lead Profile                                       0
I agree to pay the amount through cheque           0
A free copy of Mastering The Interview             0
Last Notable Activity                              0
dtype: int64
```

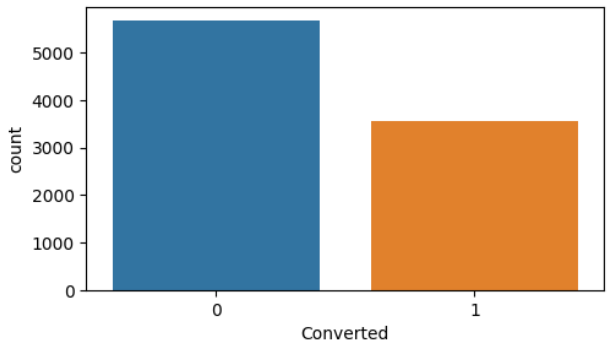
Conclusion: Now no missings values in this dataset

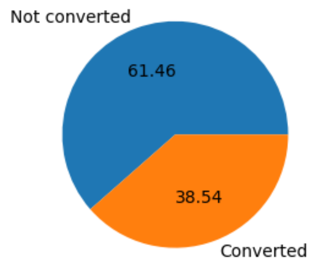
#Performing basic EDA **For Handling categorical columns please check Bellow EDA**

[31]:
```python
print(leads["Converted"].value_counts())
plt.figure(figsize=(12,3))
plt.subplot(1,2,1)
sns.countplot(x=leads["Converted"])
plt.subplot(1,2,2)
data=[5679,3561]
keys=["Not converted","Converted"]
plt.pie(data,labels=keys,autopct="%.2f")
```

```
0    5679
1    3561
Name: Converted, dtype: int64
```

[31]:
```
([<matplotlib.patches.Wedge at 0x7f79f0f07160>,
  <matplotlib.patches.Wedge at 0x7f79f101cd60>],
 [Text(-0.38756250774201845, 1.0294635994500816, 'Not converted'),
  Text(0.3875624113566783, -1.0294636357362978, 'Converted')],
 [Text(-0.2113977314956464, 0.5615255997000445, '61.46'),
  Text(0.2113976789218245, -0.5615256194925261, '38.54')])
```



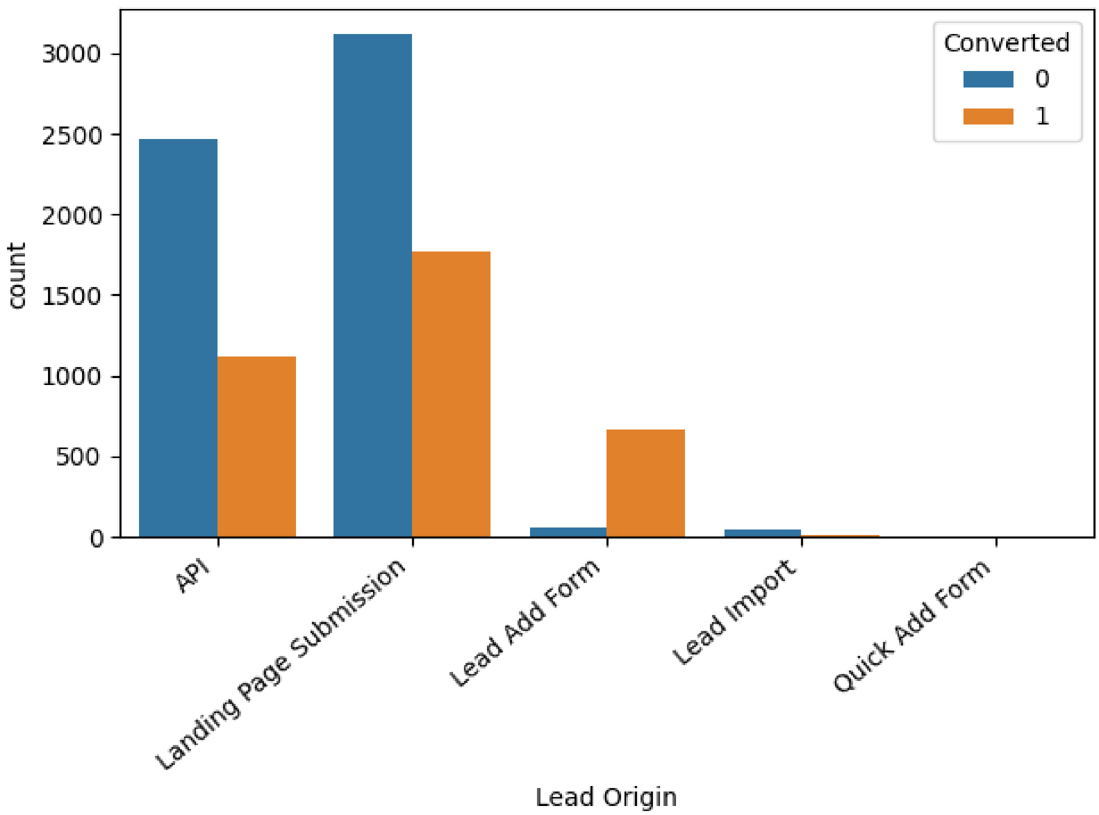


We can see data has only 38.54 % case which is sucessfully converted

Checking for columns "Converted" and "LEAD Origin"

[32]:
```
print(leads.groupby(["Lead Origin","Converted"])["Lead Number"].count())
ax=sns.countplot(x=leads["Lead Origin"],hue=leads["Converted"])
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")
plt.tight_layout()
```

```
Lead Origin              Converted
API                      0            2465
                         1            1115
Landing Page Submission  0            3118
                         1            1768
Lead Add Form            0              54
                         1             664
Lead Import              0              42
                         1              13
Quick Add Form           1               1
Name: Lead Number, dtype: int64
```



- We can see that lead from "lead origines" & "API and Landing page" submission has highest rate of conversion

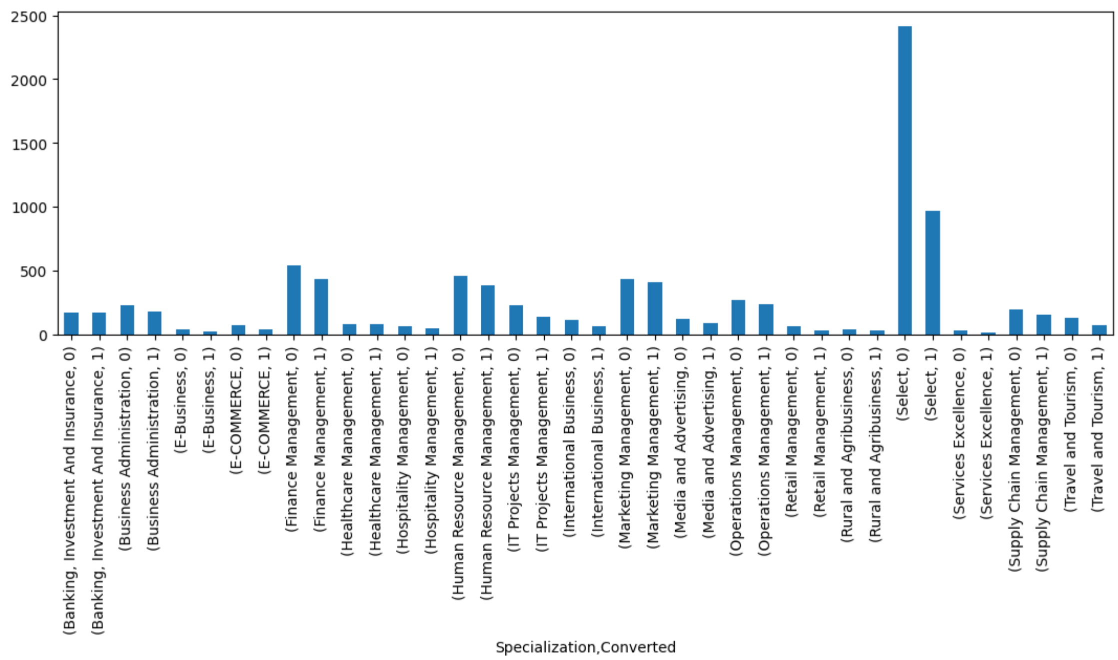
Analyzing columns specializations and converted

[33]:
```
print(leads.groupby(["Specialization","Converted"])["Lead Number"].count())
leads.groupby(["Specialization","Converted"])["Lead Number"].count().
↪plot(kind="bar",figsize=(13,4))
```

```
Specialization                    Converted
Banking, Investment And Insurance  0            171
                                   1            167
Business Administration            0            224
                                   1            179
E-Business                         0             36
                                   1             21
E-COMMERCE                         0             72
                                   1             40
Finance Management                 0            540
                                   1            436
Healthcare Management              0             80
                                   1             79
```

```
Hospitality Management           0             66
                                 1             48
Human Resource Management        0            460
                                 1            388
IT Projects Management           0            226
                                 1            140
International Business           0            114
                                 1             64
Marketing Management             0            430
                                 1            408
Media and Advertising            0            118
                                 1             85
Operations Management            0            265
                                 1            238
Retail Management                0             66
                                 1             34
Rural and Agribusiness           0             42
                                 1             31
Select                           0           2411
                                 1            969
Services Excellence              0             29
                                 1             11
Supply Chain Management          0            198
                                 1            151
Travel and Tourism               0            131
                                 1             72
Name: Lead Number, dtype: int64
```

```
[33]: <Axes: xlabel='Specialization,Converted'>
```



- Above EDA just tried, we can do also more

#Treating Categorical Values

[157]:
```
#Checking unique values and total count of unique values for each categorical␣
↪columns
categorical_columns=leads.select_dtypes(exclude=["int","float"])
for column in categorical_columns:
  print(categorical_columns[column].nunique(),end=" ")
  #print(categorical_columns[column].value_counts())
```

```
0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

[35]:
```
print(leads.describe(include="O").columns)
print("Total categorical columns are: ",len(leads.describe(include="O").
↪columns))
```

```
Index(['Prospect ID', 'Lead Origin', 'Lead Source', 'Do Not Email',
       'Do Not Call', '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',
       'Update me on Supply Chain Content', 'Get updates on DM Content',
       'Lead Profile', 'I agree to pay the amount through cheque',
```

```
       'A free copy of Mastering The Interview', 'Last Notable Activity'],
      dtype='object')
Total categorical columns are:  25
```

There are too many columns to treat

[36]:
```python
leads.apply(lambda x: len(x.unique())) #Unique values for each columns
```

[36]:
```
Prospect ID                                      9240
Lead Number                                      9240
Lead Origin                                         5
Lead Source                                        21
Do Not Email                                        2
Do Not Call                                         2
Converted                                           2
TotalVisits                                        41
Total Time Spent on Website                      1731
Page Views Per Visit                              114
Last Activity                                      17
Country                                            38
Specialization                                     19
How did you hear about X Education                 10
What is your current occupation                     6
What matters most to you in choosing a course       3
Search                                              2
Magazine                                            1
Newspaper Article                                   2
X Education Forums                                  2
Newspaper                                           2
Digital Advertisement                               2
Through Recommendations                             2
Receive More Updates About Our Courses              1
Update me on Supply Chain Content                   1
Get updates on DM Content                           1
Lead Profile                                        6
I agree to pay the amount through cheque            1
A free copy of Mastering The Interview              2
Last Notable Activity                              16
dtype: int64
```

[155]:
```python
#'Prospect ID' treating
print(leads["Prospect ID"].nunique(),len(leads["Prospect ID"]))
#Means each row has unique values forthis column and has no order but column is␣
↪also a important column
#Using target mean encoder for this column
mean_value = leads.groupby('Prospect ID')['TotalVisits'].mean()
mean_dict = mean_value.to_dict()
print(mean_dict)
```

```
leads['Prospect ID']=leads['Prospect ID'].replace(mean_dict,inplace=True)
```

```
0 9240
{}
```

Treating column WHich has less than 10 cardinality and has no order by Dumies method"

[38]:
```
# pd.get_dummies:
leads= pd.get_dummies(leads, columns=['Lead Origin',"Do Not Email","Do Not␣
↪Call","Newspaper","Digital Advertisement","Through Recommendations","Receive␣
↪More Updates About Our Courses","Update me on Supply Chain Content"])
```

[39]:
```
#Dummy varriable for other columns like
leads = pd.get_dummies(leads,columns=['Lead Source','What is your current␣
↪occupation','A free copy of Mastering The Interview','Last Notable␣
↪Activity',"Lead Profile"])
```

Treating column "Country"

[40]:
```
leads["Country"].nunique() #So country column has 38 unique values, High␣
↪cardinality and has no orderGoing to use frequency encoding
freq_value = leads['Country'].value_counts(normalize=True)
freq_dict = freq_value.to_dict()
leads['Country'].replace(freq_dict,inplace=True)
```

Checking other columns Column Last Activity

[41]:
```
leads["Last Activity"].value_counts() #Also it has High number of unique values␣
↪and has no order we can use frequency encoding
freq=leads["Last Activity"].value_counts(normalize=True)
freq_dict=freq.to_dict()
leads["Last Activity"].replace(freq_dict,inplace=True)
```

Treating column "Specialization"

[42]:
```
#Frequency encoding for Specialization' as it has high cardinality and has no␣
↪order
freq_value=leads["Specialization"].value_counts(normalize=True)
fre_dict=freq_value.to_dict()
leads["Specialization"].replace(fre_dict,inplace=True)
```

[43]:
```
leads.describe(include="O").columns #Now remianing columns are
```

[43]:
```
Index(['How did you hear about X Education',
       'What matters most to you in choosing a course', 'Search', 'Magazine',
       'Newspaper Article', 'X Education Forums', 'Get updates on DM Content',
       'I agree to pay the amount through cheque'],
      dtype='object')
```

```
[44]: #For column How did you hear about X Education
freqency=leads['How did you hear about X Education'].
 ↪value_counts(normalize=True) #Using frequency encoding ,values has no order
freq=freqency.to_dict()
leads['How did you hear about X Education'].replace(freq,inplace=True)
```

Checking other remaining columns

```
[45]: print(leads['What matters most to you in choosing a course'].value_counts())
print(leads[['Search']].value_counts())
print(leads['Newspaper Article'].value_counts())
print(leads['X Education Forums'].value_counts())
print(leads['Get updates on DM Content'].value_counts())
print(leads['I agree to pay the amount through cheque'].value_counts())
print(leads[["Magazine"]].value_counts()) #We can also do it by for loop
```

```
Better Career Prospects      9237
Flexibility & Convenience       2
Other                           1
Name: What matters most to you in choosing a course, dtype: int64
Search
No        9226
Yes         14
dtype: int64
No     9238
Yes       2
Name: Newspaper Article, dtype: int64
No     9239
Yes       1
Name: X Education Forums, dtype: int64
No    9240
Name: Get updates on DM Content, dtype: int64
No    9240
Name: I agree to pay the amount through cheque, dtype: int64
Magazine
No          9240
dtype: int64
```

Observations: 1. So we can see these all column has the values which are not usefull or able to deliver insights. 2. All columns has just one unique values in form of Yes or No.,And some has both but very less partition of yes or no. 3. We can drop these columns from further analysis.

```
[46]: leads.drop(columns=[ 'What matters most to you in choosing a course', 'Search',
       'Magazine', 'Newspaper Article', 'X Education Forums',
       'Get updates on DM Content',
       'I agree to pay the amount through cheque'],axis=1,inplace=True)
```

All categorical columns are treated and ready for model building

#Treating numerical columns

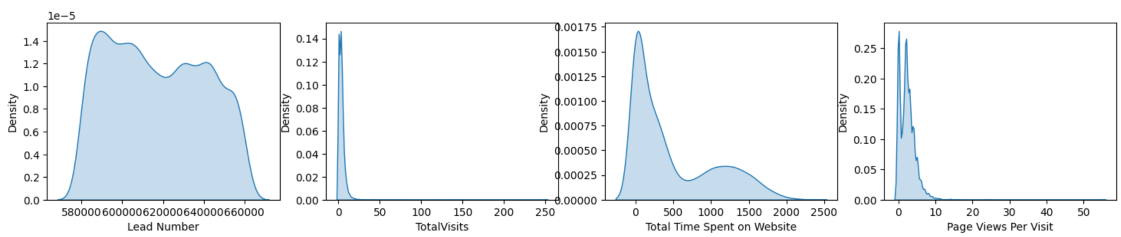
[47]:
```
Numerical_COLUMNS= [['Lead Number', 'Converted', 'TotalVisits',
        'Total Time Spent on Website', 'Page Views Per Visit']]
```

Analyzing distribution of numerical columns

[48]:
```
plt.figure(figsize=(18,3))
print("Skewnwss for Lead Number", leads["Lead Number"].skew())
plt.subplot(1,4,1)
sns.kdeplot(x=leads["Lead Number"],shade=True)
print("Skewnwss for TotalVisits", leads["TotalVisits"].skew())
plt.subplot(1,4,2)
sns.kdeplot(x=leads["TotalVisits"],shade=True)
print("Skewnwss for Total Time Spent on Website", leads["Total Time Spent on␣
↪Website"].skew())
plt.subplot(1,4,3)
sns.kdeplot(x=leads["Total Time Spent on Website"],shade=True)
print("Skewness value for Page Views Per Visit", leads["Page Views Per Visit"].
↪skew())
plt.subplot(1,4,4)
sns.kdeplot(x=leads["Page Views Per Visit"],shade=True)
```

```
Skewnwss for Lead Number 0.1404510857699009
Skewnwss for TotalVisits 19.86824774467054
Skewnwss for Total Time Spent on Website 0.9564501929530472
Skewness value for Page Views Per Visit 2.848353418319736
```

[48]: <Axes: xlabel='Page Views Per Visit', ylabel='Density'>



- We can see Lead Number column is type of serial number as it is so much uniformly distributed.
- We can drop this column as it has nothing to deliver

[49]:
```
leads.drop(columns=["Lead Number"],axis=1,inplace=True)
```

- Scaling remaining columns by MinMax scaler

[50]:
```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
leads[['TotalVisits','Total Time Spent on Website', 'Page Views Per Visit']] =␣
↪scaler.fit_transform(leads[['TotalVisits','Total Time Spent on␣
↪Website','Page Views Per Visit']])
```

[158]:
```
leads.head(1)
```

[158]:
```
  Prospect ID  Converted  TotalVisits  Total Time Spent on Website  \
0        None          0          0.0                          0.0

   Page Views Per Visit  Last Activity   Country  Specialization  \
0                   0.0       0.069264  0.968939        0.365801

   How did you hear about X Education  Lead Origin_API  …  \
0                            0.784632                1  …

   Last Notable Activity_SMS Sent  Last Notable Activity_Unreachable  \
0                               0                                  0

   Last Notable Activity_Unsubscribed  \
0                                   0

   Last Notable Activity_View in browser link Clicked  \
0                                                   0

   Lead Profile_Dual Specialization Student  Lead Profile_Lateral Student  \
0                                         0                             0

   Lead Profile_Other Leads  Lead Profile_Potential Lead  Lead Profile_Select  \
0                         0                            0                    1

   Lead Profile_Student of SomeSchool
0                                   0

[1 rows x 77 columns]
```

[52]:
```
# Import the required library

from sklearn.model_selection import train_test_split
```

[53]:
```
# Put all the feature variables in X

X = leads.drop(['Converted'], 1)
X.head()
```

[53]:
```
   Prospect ID  TotalVisits  Total Time Spent on Website  \
0          0.0     0.000000                     0.000000
1          5.0     0.019920                     0.296655
2          2.0     0.007968                     0.674296
3          1.0     0.003984                     0.134243
4          2.0     0.007968                     0.628521

   Page Views Per Visit  Last Activity   Country  Specialization  \
0              0.000000       0.069264  0.968939        0.365801
1              0.045455       0.383117  0.968939        0.365801
2              0.036364       0.383117  0.968939        0.043615
3              0.018182       0.010065  0.968939        0.021970
4              0.018182       0.046320  0.968939        0.365801

   How did you hear about X Education  Lead Origin_API  \
0                            0.784632                1
1                            0.784632                1
2                            0.784632                0
3                            0.037662                0
4                            0.020130                0

   Lead Origin_Landing Page Submission  …  Last Notable Activity_SMS Sent  \
0                                    0  …                               0
1                                    0  …                               0
2                                    1  …                               0
3                                    1  …                               0
4                                    1  …                               0

   Last Notable Activity_Unreachable  Last Notable Activity_Unsubscribed  \
0                                  0                                   0
1                                  0                                   0
2                                  0                                   0
3                                  0                                   0
4                                  0                                   0

   Last Notable Activity_View in browser link Clicked  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Lead Profile_Dual Specialization Student  Lead Profile_Lateral Student  \
0                                         0                             0
1                                         0                             0
2                                         0                             0
3                                         0                             0
```

```
4                                          0                              0

   Lead Profile_Other Leads  Lead Profile_Potential Lead  Lead Profile_Select  \
0                        0                            0                     1
1                        0                            0                     1
2                        0                            1                     0
3                        0                            0                     1
4                        0                            0                     1

   Lead Profile_Student of SomeSchool
0                                   0
1                                   0
2                                   0
3                                   0
4                                   0

[5 rows x 76 columns]
```

[54]:
```
# Put the target variable in y

y = leads[['Converted']]

y.head()
```

[54]:
```
   Converted
0          0
1          0
2          1
3          0
4          1
```

[55]:
```
# Split the dataset into 70% train and 30% test

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7,␣
 ↪test_size=0.3, random_state=100)
```

[56]:
```
X_train.shape, y_train.shape
```

[56]: ((6468, 76), (6468, 1))

[57]:
```
X_test.shape, y_test.shape
```

[57]: ((2772, 76), (2772, 1))

0.3.1 Scaling

Now there are a few numeric variables present in the dataset which have different scales. So let's go ahead and scale these variables.

[58]:
```python
# Import MinMax scaler

from sklearn.preprocessing import MinMaxScaler
```

[59]:
```python
# Scale the three numeric features present in the dataset

#scaler = MinMaxScaler()

#X_train[['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on␣
↪Website']] = scaler.fit_transform(X_train[['TotalVisits', 'Page Views Per␣
↪Visit', 'Total Time Spent on Website']])

#X_train.head()
```

0.3.2 Looking at the correlations

Let's now look at the correlations. Since the number of variables are pretty high, it's better that we look at the table instead of plotting a heatmap

[60]:
```python
# Looking at the correlation table

leads.corr()
```

[60]:
```
                                  Prospect ID  Converted  TotalVisits  \
Prospect ID                          1.000000   0.022489     1.000000
Converted                            0.022489   1.000000     0.022489
TotalVisits                          1.000000   0.022489     1.000000
Total Time Spent on Website          0.209596   0.362483     0.209596
Page Views Per Visit                 0.517030  -0.014680     0.517030
…                                         …         …           …
Lead Profile_Lateral Student        -0.001968   0.060075    -0.001968
Lead Profile_Other Leads             0.019703  -0.007648     0.019703
Lead Profile_Potential Lead         -0.005177   0.378061    -0.005177
Lead Profile_Select                  0.005708  -0.294656     0.005708
Lead Profile_Student of SomeSchool  -0.032877  -0.117030    -0.032877

                                  Total Time Spent on Website  \
Prospect ID                                          0.209596
Converted                                            0.362483
TotalVisits                                          0.209596
Total Time Spent on Website                          1.000000
Page Views Per Visit                                 0.301365
…                                                       …
```

Lead Profile_Lateral Student	0.010357
Lead Profile_Other Leads	0.011350
Lead Profile_Potential Lead	0.123592
Lead Profile_Select	-0.101309
Lead Profile_Student of SomeSchool	-0.044920

	Page Views Per Visit	Last Activity \
Prospect ID	0.517030	0.007645
Converted	-0.014680	0.226871
TotalVisits	0.517030	0.007645
Total Time Spent on Website	0.301365	0.098008
Page Views Per Visit	1.000000	0.090585
…	…	…
Lead Profile_Lateral Student	0.001821	0.032071
Lead Profile_Other Leads	0.032454	-0.030233
Lead Profile_Potential Lead	-0.023825	0.105760
Lead Profile_Select	0.018139	-0.056773
Lead Profile_Student of SomeSchool	-0.042886	-0.071961

	Country	Specialization \
Prospect ID	-0.024989	-0.260495
Converted	0.034124	-0.144618
TotalVisits	-0.024989	-0.260495
Total Time Spent on Website	-0.026486	-0.282902
Page Views Per Visit	-0.044219	-0.409038
…	…	…
Lead Profile_Lateral Student	0.009137	-0.027459
Lead Profile_Other Leads	-0.005286	-0.104921
Lead Profile_Potential Lead	0.011695	-0.245498
Lead Profile_Select	-0.007245	0.265447
Lead Profile_Student of SomeSchool	-0.001974	0.019668

	How did you hear about X Education \
Prospect ID	-0.172782
Converted	-0.043106
TotalVisits	-0.172782
Total Time Spent on Website	-0.181843
Page Views Per Visit	-0.271927
…	…
Lead Profile_Lateral Student	0.016125
Lead Profile_Other Leads	-0.034593
Lead Profile_Potential Lead	-0.016125
Lead Profile_Select	0.005141
Lead Profile_Student of SomeSchool	0.067298

	Lead Origin_API	… \
Prospect ID	-0.194650	…

```
Converted                                  -0.120822  …
TotalVisits                                -0.194650  …
Total Time Spent on Website                -0.201239  …
Page Views Per Visit                       -0.331938  …
…                                               …   …
Lead Profile_Lateral Student               -0.014398  …
Lead Profile_Other Leads                   -0.056358  …
Lead Profile_Potential Lead                -0.115262  …
Lead Profile_Select                         0.123391  …
Lead Profile_Student of SomeSchool          0.016204  …

                                    Last Notable Activity_SMS Sent  \
Prospect ID                                               0.001873
Converted                                                 0.351845
TotalVisits                                               0.001873
Total Time Spent on Website                               0.125076
Page Views Per Visit                                      0.063860
…                                                              …
Lead Profile_Lateral Student                             -0.023275
Lead Profile_Other Leads                                 -0.098776
Lead Profile_Potential Lead                               0.082592
Lead Profile_Select                                       0.013782
Lead Profile_Student of SomeSchool                       -0.090718

                                    Last Notable Activity_Unreachable  \
Prospect ID                                                  0.006234
Converted                                                    0.036594
TotalVisits                                                  0.006234
Total Time Spent on Website                                  0.008941
Page Views Per Visit                                         0.020354
…                                                                 …
Lead Profile_Lateral Student                                -0.003008
Lead Profile_Other Leads                                    -0.013905
Lead Profile_Potential Lead                                  0.026274
Lead Profile_Select                                         -0.011536
Lead Profile_Student of SomeSchool                          -0.009647

                                    Last Notable Activity_Unsubscribed  \
Prospect ID                                                   0.001713
Converted                                                    -0.012858
TotalVisits                                                   0.001713
Total Time Spent on Website                                   0.000503
Page Views Per Visit                                          0.019035
…                                                                  …
Lead Profile_Lateral Student                                 -0.003649
Lead Profile_Other Leads                                      0.023985
Lead Profile_Potential Lead                                  -0.000820
```

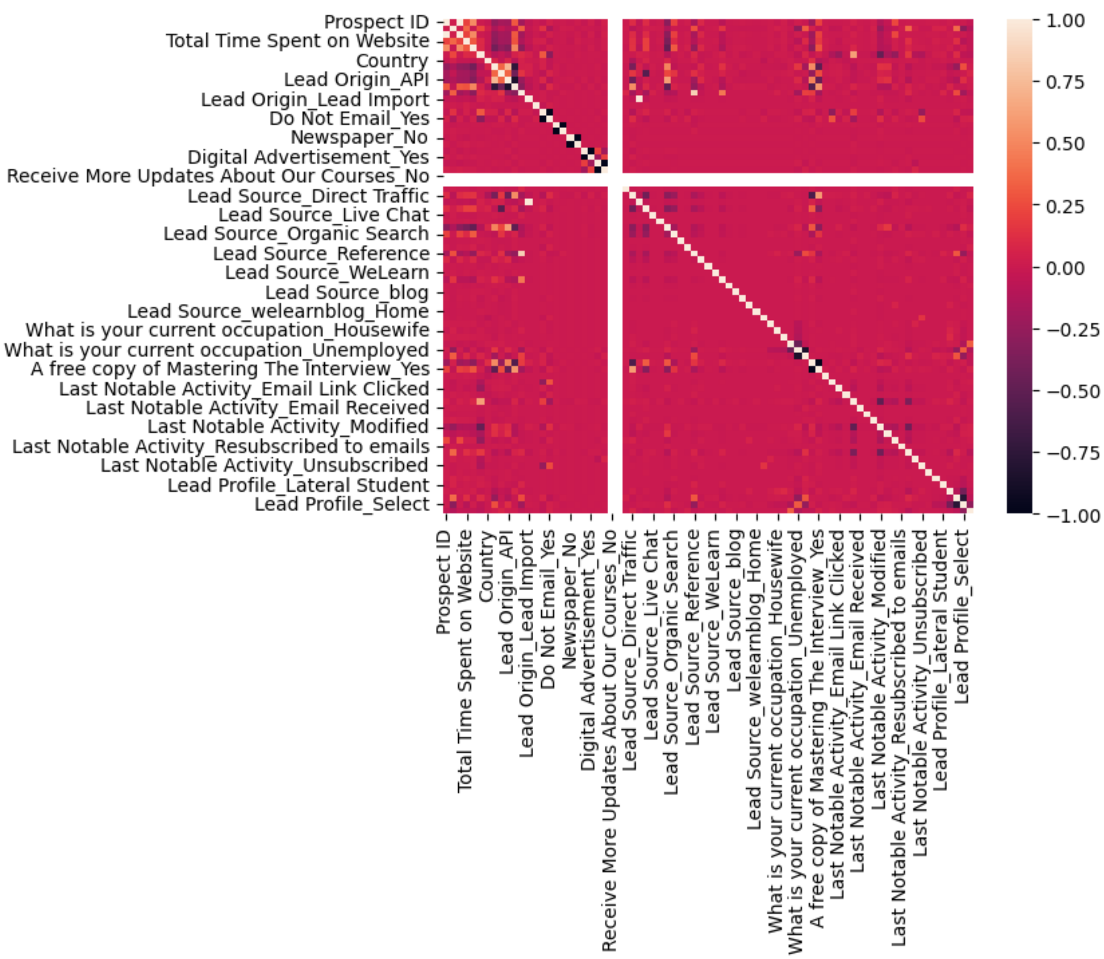
```
Lead Profile_Select                                                -0.013449
Lead Profile_Student of SomeSchool                                  0.007389

                                    Last Notable Activity_View in browser link
Clicked  \
Prospect ID
0.009907
Converted
-0.008238
TotalVisits
0.009907
Total Time Spent on Website
-0.007569
Page Views Per Visit
0.001643
…
…
Lead Profile_Lateral Student
-0.000531
Lead Profile_Other Leads
-0.002454
Lead Profile_Potential Lead
-0.004784
Lead Profile_Select
0.006137
Lead Profile_Student of SomeSchool
-0.001703

                                    Lead Profile_Dual Specialization Student  \
Prospect ID                                                         0.008718
Converted                                                           0.058817
TotalVisits                                                         0.008718
Total Time Spent on Website                                         0.032577
Page Views Per Visit                                                0.012845
…                                                                      …
Lead Profile_Lateral Student                                       -0.002377
Lead Profile_Other Leads                                           -0.010986
Lead Profile_Potential Lead                                        -0.021419
Lead Profile_Select                                                -0.078960
Lead Profile_Student of SomeSchool                                 -0.007622

                                    Lead Profile_Lateral Student  \
Prospect ID                                            -0.001968
Converted                                               0.060075
TotalVisits                                            -0.001968
Total Time Spent on Website                             0.010357
Page Views Per Visit                                    0.001821
```

```
…                                                      …
Lead Profile_Lateral Student                    1.000000
Lead Profile_Other Leads                       -0.012037
Lead Profile_Potential Lead                    -0.023468
Lead Profile_Select                            -0.086516
Lead Profile_Student of SomeSchool             -0.008351

                                    Lead Profile_Other Leads  \
Prospect ID                                         0.019703
Converted                                          -0.007648
TotalVisits                                         0.019703
Total Time Spent on Website                         0.011350
Page Views Per Visit                                0.032454
…                                                      …
Lead Profile_Lateral Student                       -0.012037
Lead Profile_Other Leads                            1.000000
Lead Profile_Potential Lead                        -0.108474
Lead Profile_Select                                -0.399895
Lead Profile_Student of SomeSchool                 -0.038601

                                    Lead Profile_Potential Lead  \
Prospect ID                                            -0.005177
Converted                                               0.378061
TotalVisits                                            -0.005177
Total Time Spent on Website                             0.123592
Page Views Per Visit                                   -0.023825
…                                                         …
Lead Profile_Lateral Student                           -0.023468
Lead Profile_Other Leads                               -0.108474
Lead Profile_Potential Lead                             1.000000
Lead Profile_Select                                    -0.779650
Lead Profile_Student of SomeSchool                     -0.075258

                                    Lead Profile_Select  \
Prospect ID                                    0.005708
Converted                                     -0.294656
TotalVisits                                    0.005708
Total Time Spent on Website                   -0.101309
Page Views Per Visit                           0.018139
…                                                 …
Lead Profile_Lateral Student                  -0.086516
Lead Profile_Other Leads                      -0.399895
Lead Profile_Potential Lead                   -0.779650
Lead Profile_Select                            1.000000
Lead Profile_Student of SomeSchool            -0.277441

                                    Lead Profile_Student of SomeSchool
```

```
Prospect ID                                   -0.032877
Converted                                     -0.117030
TotalVisits                                   -0.032877
Total Time Spent on Website                   -0.044920
Page Views Per Visit                          -0.042886
…                                                 …
Lead Profile_Lateral Student                  -0.008351
Lead Profile_Other Leads                      -0.038601
Lead Profile_Potential Lead                   -0.075258
Lead Profile_Select                           -0.277441
Lead Profile_Student of SomeSchool             1.000000

[77 rows x 77 columns]
```

[61]:
```python
sns.heatmap(leads.corr())
```

[61]: <Axes: >



0.4 Step 2: Model Building

Let's now move to model building. As you can see that there are a lot of variables present in the dataset which we cannot deal with. So the best way to approach this is to select a small set of features from this pool of variables using RFE.

1 Benchmark Evaluation

[62]:
```
# Import 'LogisticRegression' and create a LogisticRegression object
```

[63]:
```
from sklearn.linear_model import LogisticRegression
```

[64]:
```
model = LogisticRegression()
```

[65]:
```
model.fit(X_train,y_train)
```

[65]:
```
LogisticRegression()
```

[66]:
```
model.predict_proba(X_train) #Predicting probability
```

[66]:
```
array([[0.76002036, 0.23997964],
       [0.89335651, 0.10664349],
       [0.76219039, 0.23780961],
       …,
       [0.81231392, 0.18768608],
       [0.94851236, 0.05148764],
       [0.84639716, 0.15360284]])
```

[67]:
```
model.predict(X_train)
```

[67]:
```
array([0, 0, 0, …, 0, 0, 0])
```

1.1 Train

[68]:
```
y_train
```

[68]:
```
      Converted
1871          0
6795          0
3516          0
8105          0
3934          0
…           …
350           1
79            1
8039          1
6936          0
```

```
5640         0

[6468 rows x 1 columns]
```

[69]:
```
y_train_pred = model.predict(X_train)
y_train_pred
```

[69]:
```
array([0, 0, 0, …, 0, 0, 0])
```

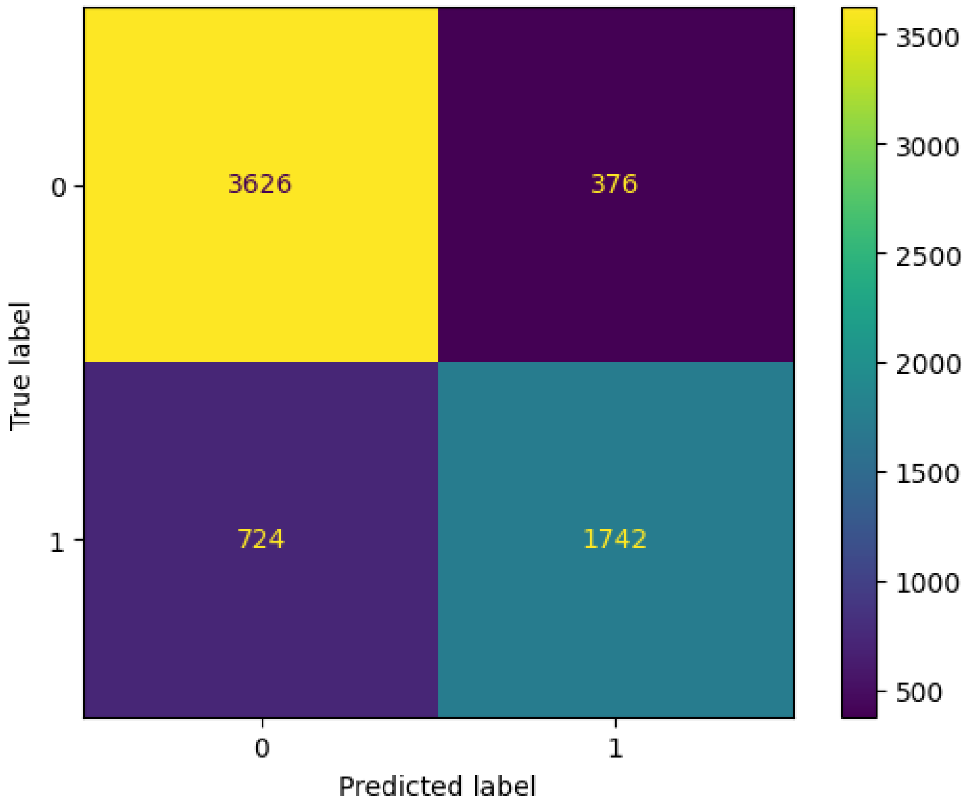
[70]:
```
from sklearn.metrics import confusion_matrix #Importing confusion matrix
```

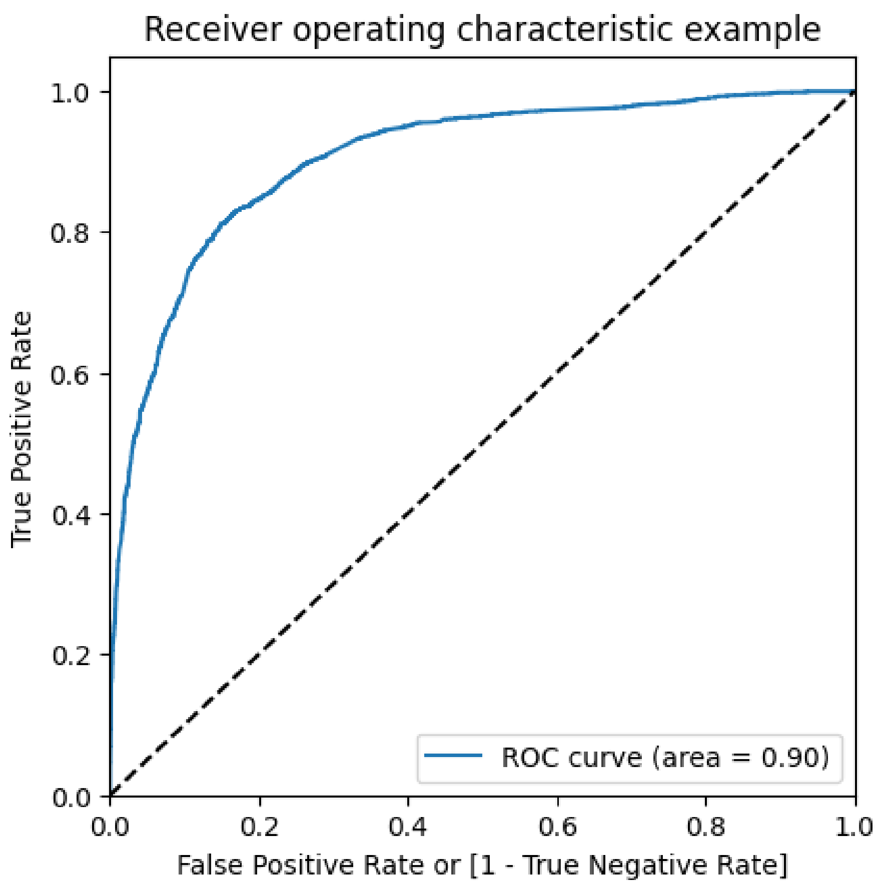
[71]:
```
confusion_matrix(y_train,y_train_pred)
```

[71]:
```
array([[3626,  376],
       [ 724, 1742]])
```

[72]:
```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_train,y_train_pred)
from sklearn.metrics import ConfusionMatrixDisplay
ConfusionMatrixDisplay.from_estimator(model, X_train, y_train)

from sklearn.metrics import accuracy_score, precision_score, recall_score ,␣
 ↪f1_score, roc_auc_score
accuracy_score(y_train,y_train_pred)
precision_score(y_train,y_train_pred)
recall_score(y_train,y_train_pred)
f1_score(y_train,y_train_pred)
roc_auc_score(y_train,model.predict_proba(X_train)[:,1])

# ROC function
from sklearn import metrics
def draw_roc( actual, probs ):
    fpr, tpr, thresholds = metrics.roc_curve( actual, probs)
    auc_score = metrics.roc_auc_score( actual, probs )
    plt.figure(figsize=(5, 5))
    plt.plot( fpr, tpr, label='ROC curve (area = %0.2f)' % auc_score )
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate or [1 - True Negative Rate]')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver operating characteristic example')
    plt.legend(loc="lower right")
    plt.show()

    return None
```

```
actual = y_train
probs = model.predict_proba(X_train)[:,1]
draw_roc(actual,probs)
```

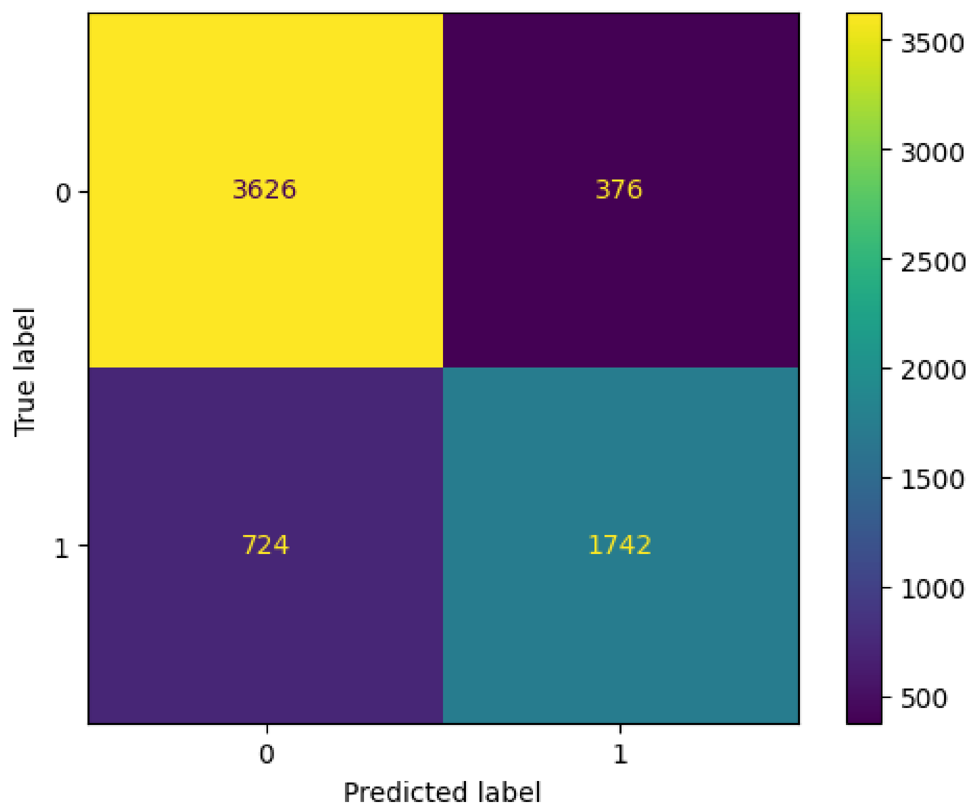




[73]:
```
from sklearn.metrics import ConfusionMatrixDisplay
```

[74]:
```
ConfusionMatrixDisplay.from_estimator(model, X_train, y_train)
```

[74]:
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at
0x7f79ea0ad5a0>
```



EValuation

[75]:
```
from sklearn.metrics import accuracy_score, precision_score, recall_score ,
↪f1_score, roc_auc_score
```

[76]:
```
accuracy_score(y_train,y_train_pred)
```

[76]: 0.8299319727891157

[77]:
```
precision_score(y_train,y_train_pred)
```

[77]: 0.8224740321057602

[78]:
```
recall_score(y_train,y_train_pred)
```

[78]: 0.7064071370640713

[79]:
```
f1_score(y_train,y_train_pred)
```

[79]: 0.7600349040139616

[80]:
```
roc_auc_score(y_train,model.predict_proba(X_train)[:,1])
```

[80]: 0.9037349228872993

1.2 Test

[81]:
```
y_test
```

[81]:
```
      Converted
4269          1
2376          1
7766          1
9199          0
4359          1
…           …
8649          0
2152          1
7101          0
5331          0
2960          1

[2772 rows x 1 columns]
```

[82]:
```
model.predict_proba(X_test)
```

[82]:
```
array([[0.28401374, 0.71598626],
       [0.11883658, 0.88116342],
       [0.13245898, 0.86754102],
       …,
       [0.81231392, 0.18768608],
       [0.84738844, 0.15261156],
       [0.11883658, 0.88116342]])
```

[83]:
```
model.predict_proba(X_test)[:,1]
```

[83]:
```
array([0.71598626, 0.88116342, 0.86754102, …, 0.18768608, 0.15261156,
       0.88116342])
```

[84]:
```
y_test_pred = model.predict(X_test)
```

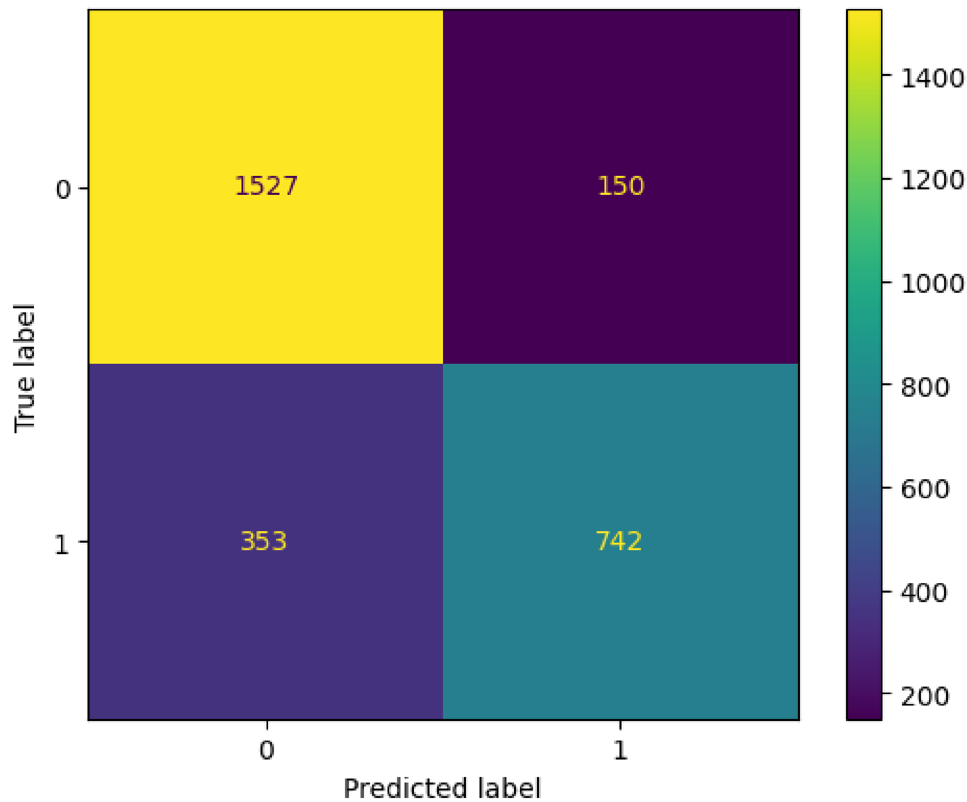
[85]:
```
from sklearn.metrics import confusion_matrix
```

[86]:
```
confusion_matrix(y_test,y_test_pred)
```

[86]:
```
array([[1527,  150],
       [ 353,  742]])
```

[87]:
```
ConfusionMatrixDisplay.from_estimator(model, X_test, y_test)
```

[87]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f7a269aa080>



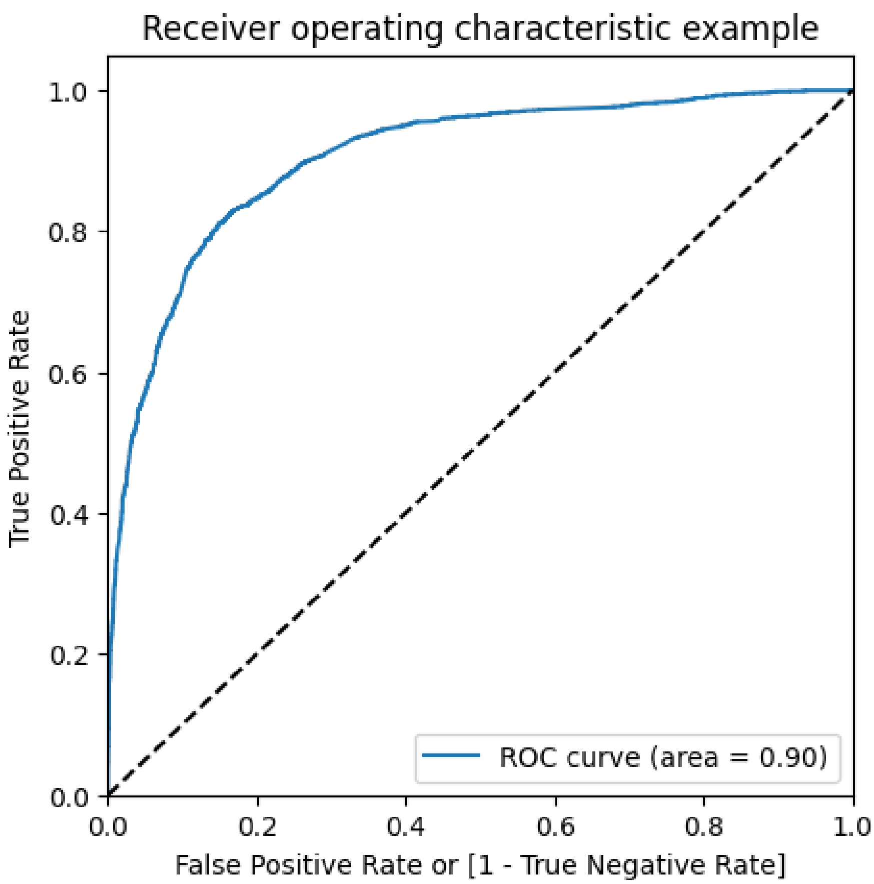
1.3 ROC Curve and AUC Score

[88]:
```
# ROC function
from sklearn import metrics
def draw_roc( actual, probs ):
    fpr, tpr, thresholds = metrics.roc_curve( actual, probs)
    auc_score = metrics.roc_auc_score( actual, probs )
    plt.figure(figsize=(5, 5))
    plt.plot( fpr, tpr, label='ROC curve (area = %0.2f)' % auc_score )
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate or [1 - True Negative Rate]')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver operating characteristic example')
    plt.legend(loc="lower right")
    plt.show()
```

```
    return None
```

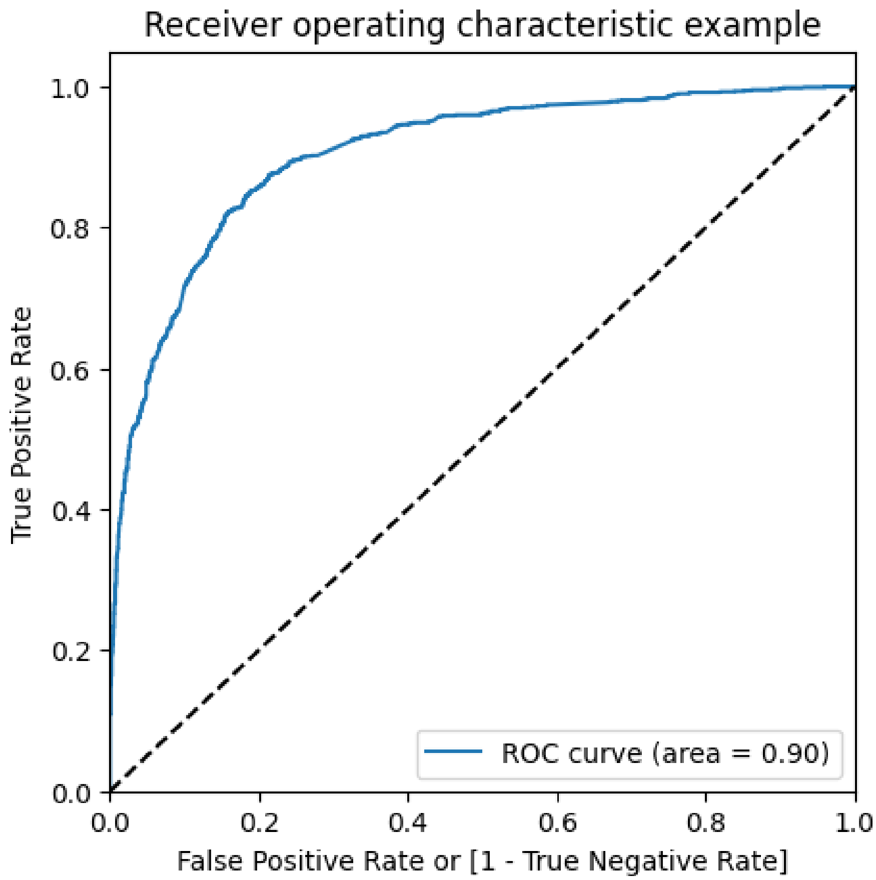
[89]:
```
actual = y_train
probs = model.predict_proba(X_train)[:,1]
```

[90]:
```
draw_roc(actual,probs)
```



[91]:
```
actual = y_test
probs = model.predict_proba(X_test)[:,1]
```

[92]:
```
draw_roc(actual,probs)
```



1.4 Improving Model performance.

[93]:
```
# Select more appropiate features and removing any correlated features.
```

[94]:
```
# Import RFE and select 15 variables
# Recursive feature Elimination ,technique for doing feature selection.

from sklearn.feature_selection import RFE
rfe = RFE(model, n_features_to_select=15, step=1)    # running RFE with 15␣
↪variables as output
rfe = rfe.fit(X_train, y_train)
```

[95]:
```
rfe.support_
```

[95]:
```
array([False, False,  True, False,  True, False, False, False, False,
       False,  True, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False,  True, False, False, False, False,
```

```
       False, False, False,  True, False, False, False, False, False,
       False, False,  True, False, False, False,  True, False, False,
       False, False, False, False,  True, False, False,  True,  True,
        True, False, False, False, False, False, False,  True,  True,
       False,  True, False,  True])
```

[96]:
```
rfe.ranking_
```

[96]:
```
array([56, 59,  1, 20,  1, 44,  7, 21, 10,  8,  1, 31, 38, 18,  2, 35, 29,
       11, 22, 26, 15, 14, 25, 54, 55,  6, 40,  9, 43, 37, 13,  1, 42, 49,
       50, 53, 41, 34, 48,  1, 32, 39, 33, 61, 46, 52, 23,  1, 24, 28,  5,
        1, 57, 58, 60, 36,  3, 45,  1, 47, 51,  1,  1,  1,  4, 12, 17, 16,
       30, 62,  1,  1, 27,  1, 19,  1])
```

[97]:
```
# Let's take a look at which features have been selected by RFE

list(zip(X_train.columns, rfe.support_, rfe.ranking_))
```

[97]:
```
[('Prospect ID', False, 56),
 ('TotalVisits', False, 59),
 ('Total Time Spent on Website', True, 1),
 ('Page Views Per Visit', False, 20),
 ('Last Activity', True, 1),
 ('Country', False, 44),
 ('Specialization', False, 7),
 ('How did you hear about X Education', False, 21),
 ('Lead Origin_API', False, 10),
 ('Lead Origin_Landing Page Submission', False, 8),
 ('Lead Origin_Lead Add Form', True, 1),
 ('Lead Origin_Lead Import', False, 31),
 ('Lead Origin_Quick Add Form', False, 38),
 ('Do Not Email_No', False, 18),
 ('Do Not Email_Yes', False, 2),
 ('Do Not Call_No', False, 35),
 ('Do Not Call_Yes', False, 29),
 ('Newspaper_No', False, 11),
 ('Newspaper_Yes', False, 22),
 ('Digital Advertisement_No', False, 26),
 ('Digital Advertisement_Yes', False, 15),
 ('Through Recommendations_No', False, 14),
 ('Through Recommendations_Yes', False, 25),
 ('Receive More Updates About Our Courses_No', False, 54),
 ('Update me on Supply Chain Content_No', False, 55),
 ('Lead Source_Click2call', False, 6),
 ('Lead Source_Direct Traffic', False, 40),
 ('Lead Source_Facebook', False, 9),
 ('Lead Source_Google', False, 43),
```

```
('Lead Source_Live Chat', False, 37),
('Lead Source_NC_EDM', False, 13),
('Lead Source_Olark Chat', True, 1),
('Lead Source_Organic Search', False, 42),
('Lead Source_Pay per Click Ads', False, 49),
('Lead Source_Press_Release', False, 50),
('Lead Source_Reference', False, 53),
('Lead Source_Referral Sites', False, 41),
('Lead Source_Social Media', False, 34),
('Lead Source_WeLearn', False, 48),
('Lead Source_Welingak Website', True, 1),
('Lead Source_bing', False, 32),
('Lead Source_blog', False, 39),
('Lead Source_google', False, 33),
('Lead Source_testone', False, 61),
('Lead Source_welearnblog_Home', False, 46),
('Lead Source_youtubechannel', False, 52),
('What is your current occupation_Businessman', False, 23),
('What is your current occupation_Housewife', True, 1),
('What is your current occupation_Other', False, 24),
('What is your current occupation_Student', False, 28),
('What is your current occupation_Unemployed', False, 5),
('What is your current occupation_Working Professional', True, 1),
('A free copy of Mastering The Interview_No', False, 57),
('A free copy of Mastering The Interview_Yes', False, 58),
('Last Notable Activity_Approached upfront', False, 60),
('Last Notable Activity_Email Bounced', False, 36),
('Last Notable Activity_Email Link Clicked', False, 3),
('Last Notable Activity_Email Marked Spam', False, 45),
('Last Notable Activity_Email Opened', True, 1),
('Last Notable Activity_Email Received', False, 47),
('Last Notable Activity_Form Submitted on Website', False, 51),
('Last Notable Activity_Had a Phone Conversation', True, 1),
('Last Notable Activity_Modified', True, 1),
('Last Notable Activity_Olark Chat Conversation', True, 1),
('Last Notable Activity_Page Visited on Website', False, 4),
('Last Notable Activity_Resubscribed to emails', False, 12),
('Last Notable Activity_SMS Sent', False, 17),
('Last Notable Activity_Unreachable', False, 16),
('Last Notable Activity_Unsubscribed', False, 30),
('Last Notable Activity_View in browser link Clicked', False, 62),
('Lead Profile_Dual Specialization Student', True, 1),
('Lead Profile_Lateral Student', True, 1),
('Lead Profile_Other Leads', False, 27),
('Lead Profile_Potential Lead', True, 1),
('Lead Profile_Select', False, 19),
('Lead Profile_Student of SomeSchool', True, 1)]
```

[98]:
```
# Put all the columns selected by RFE in the variable 'col'

col = X_train.columns[rfe.support_]
```

[99]:
```
col
```

[99]:
```
Index(['Total Time Spent on Website', 'Last Activity',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_Welingak Website',
       'What is your current occupation_Housewife',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Email Opened',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation',
       'Lead Profile_Dual Specialization Student',
       'Lead Profile_Lateral Student', 'Lead Profile_Potential Lead',
       'Lead Profile_Student of SomeSchool'],
      dtype='object')
```

Now you have all the variables selected by RFE and since we care about the statistics part, i.e. the p-values and the VIFs, let's use these variables to create a logistic regression model using statsmodels.

[100]:
```
# Select only the columns selected by RFE

X_train_new = X_train[col]
```

[101]:
```
X_train_new
```

[101]:
```
      Total Time Spent on Website  Last Activity  Lead Origin_Lead Add Form  \
1871                     0.000000       0.383117                          0
6795                     0.214349       0.383117                          0
3516                     0.046655       0.383117                          0
8105                     0.541373       0.297078                          0
3934                     0.000000       0.383117                          0
…                             …             …                        …
350                      0.000000       0.383117                          1
79                       0.310299       0.383117                          1
8039                     0.000000       0.383117                          0
6936                     0.104754       0.046320                          0
5640                     0.000000       0.383117                          0

      Lead Source_Olark Chat  Lead Source_Welingak Website  \
1871                       1                             0
6795                       0                             0
3516                       1                             0
8105                       0                             0
```

```
3934                       1                             0
…                      …                            …
350                        0                             0
79                         0                             0
8039                       1                             0
6936                       0                             0
5640                       1                             0

      What is your current occupation_Housewife  \
1871                                          0
6795                                          0
3516                                          0
8105                                          0
3934                                          0
…                                         …
350                                           0
79                                            0
8039                                          0
6936                                          0
5640                                          0

      What is your current occupation_Working Professional  \
1871                                                  0
6795                                                  0
3516                                                  0
8105                                                  0
3934                                                  0
…                                                 …
350                                                   0
79                                                    1
8039                                                  0
6936                                                  0
5640                                                  0

      Last Notable Activity_Email Opened  \
1871                                   1
6795                                   1
3516                                   1
8105                                   0
3934                                   0
…                                  …
350                                    1
79                                     0
8039                                   1
6936                                   0
5640                                   0
```

```
      Last Notable Activity_Had a Phone Conversation  \
1871                                                0
6795                                                0
3516                                                0
8105                                                0
3934                                                0
…                                               …
350                                                 0
79                                                  0
8039                                                0
6936                                                0
5640                                                0

      Last Notable Activity_Modified  \
1871                                0
6795                                0
3516                                0
8105                                0
3934                                1
…                               …
350                                 0
79                                  1
8039                                0
6936                                1
5640                                1

      Last Notable Activity_Olark Chat Conversation  \
1871                                               0
6795                                               0
3516                                               0
8105                                               0
3934                                               0
…                                              …
350                                                0
79                                                 0
8039                                               0
6936                                               0
5640                                               0

      Lead Profile_Dual Specialization Student  Lead Profile_Lateral Student  \
1871                                          0                             0
6795                                          0                             0
3516                                          0                             0
8105                                          0                             0
3934                                          0                             0
…                                         …                           …
350                                           0                             1
```

```
79                                         0                              0
8039                                       0                              0
6936                                       0                              0
5640                                       0                              0

      Lead Profile_Potential Lead  Lead Profile_Student of SomeSchool
1871                            0                                   0
6795                            0                                   0
3516                            0                                   0
8105                            0                                   0
3934                            0                                   0
…                           …                                 …
350                             0                                   0
79                              1                                   0
8039                            0                                   0
6936                            0                                   0
5640                            0                                   0

[6468 rows x 15 columns]
```

[102]:
```python
# Import statsmodels

import statsmodels.api as sm
```

[103]:
```python
# Fit a logistic Regression model on X_train after adding a constant and output the summary

X_train_sm = sm.add_constant(X_train_new)
logm2 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

[103]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6452
Model Family:                Binomial   Df Model:                           15
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2528.3
Date:                Wed, 09 Aug 2023   Deviance:                       5056.5
Time:                        18:45:53   Pearson chi2:                 8.15e+03
No. Iterations:                    22   Pseudo R-squ. (CS):             0.4217
Covariance Type:            nonrobust
================================================================================
========================================
```

```
                                                          coef    std err
z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
----------------------------------------
const                                                  -2.6360      0.122
-21.587      0.000      -2.875      -2.397
Total Time Spent on Website                             4.5365      0.169
26.824      0.000      4.205      4.868
Last Activity                                           4.5438      0.365
12.463      0.000      3.829      5.258
Lead Origin_Lead Add Form                               3.1523      0.194
16.267      0.000      2.773      3.532
Lead Source_Olark Chat                                  1.1710      0.103
11.400      0.000      0.970      1.372
Lead Source_Welingak Website                            2.5555      0.743
3.439      0.001      1.099      4.012
What is your current occupation_Housewife              24.3009   2.09e+04
0.001      0.999    -4.1e+04    4.11e+04
What is your current occupation_Working Professional    2.5040      0.193
12.951      0.000      2.125      2.883
Last Notable Activity_Email Opened                     -1.6022      0.099
-16.209      0.000      -1.796      -1.408
Last Notable Activity_Had a Phone Conversation          2.9573      1.162
2.545      0.011      0.679      5.235
Last Notable Activity_Modified                         -1.5790      0.092
-17.243      0.000      -1.758      -1.399
Last Notable Activity_Olark Chat Conversation          -1.5912      0.340
-4.682      0.000      -2.257      -0.925
Lead Profile_Dual Specialization Student               23.4143   1.72e+04
0.001      0.999   -3.37e+04    3.37e+04
Lead Profile_Lateral Student                            2.9763      1.088
2.735      0.006      0.844      5.109
Lead Profile_Potential Lead                             1.7536      0.098
17.866      0.000      1.561      1.946
Lead Profile_Student of SomeSchool                     -1.8694      0.439
-4.255      0.000      -2.730      -1.008
================================================================================
========================================
"""
```

There are quite a few variable which have a p-value greater than `0.05`. We will need to take care of them. But first, let's also look at the VIFs.

[104]:
```python
# Import 'variance_inflation_factor'

from statsmodels.stats.outliers_influence import variance_inflation_factor
```

[105]:
```python
# Make a VIF dataframe for all the variables present

vif = pd.DataFrame()
vif['Features'] = X_train_new.columns
vif['VIF'] = [variance_inflation_factor(X_train_new.values, i) for i in␣
 ↪range(X_train_new.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

[105]:

	Features	VIF
1	Last Activity	4.56
7	Last Notable Activity_Email Opened	2.71
0	Total Time Spent on Website	2.01
2	Lead Origin_Lead Add Form	1.59
9	Last Notable Activity_Modified	1.56
13	Lead Profile_Potential Lead	1.45
3	Lead Source_Olark Chat	1.43
4	Lead Source_Welingak Website	1.26
6	What is your current occupation_Working Profes…	1.24
10	Last Notable Activity_Olark Chat Conversation	1.07
14	Lead Profile_Student of SomeSchool	1.04
5	What is your current occupation_Housewife	1.01
8	Last Notable Activity_Had a Phone Conversation	1.01
11	Lead Profile_Dual Specialization Student	1.01
12	Lead Profile_Lateral Student	1.01

VIFs seem to be in a decent range except for three variables.

Let's first drop the variable `Lead Origin_Lead Add Form` since it has a high p-value as well as a high VIF.

[106]:
```python
X_train_new.drop('Lead Origin_Lead Add Form', axis = 1, inplace = True)
```

[107]:
```python
# Refit the model with the new set of features

logm1 = sm.GLM(y_train,(sm.add_constant(X_train_new)), family = sm.families.
 ↪Binomial())
logm1.fit().summary()
```

[107]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6453
Model Family:                Binomial   Df Model:                           14
Link Function:                  Logit   Scale:                          1.0000
```

```
Method:                          IRLS   Log-Likelihood:                -2716.4
Date:                Wed, 09 Aug 2023   Deviance:                       5432.9
Time:                        18:45:54   Pearson chi2:                 6.69e+03
No. Iterations:                    22   Pseudo R-squ. (CS):             0.3870
Covariance Type:            nonrobust
================================================================================
========================================
                                                                  coef    std err
z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
----------------------------------------
const                                                          -2.3609      0.114
-20.686      0.000      -2.585      -2.137
Total Time Spent on Website                                     3.7782      0.155
24.425      0.000       3.475       4.081
Last Activity                                                   4.9756      0.345
14.415      0.000       4.299       5.652
Lead Source_Olark Chat                                          0.7695      0.097
7.913      0.000       0.579       0.960
Lead Source_Welingak Website                                    5.3022      0.722
7.344      0.000       3.887       6.717
What is your current occupation_Housewife                      24.7708   2.22e+04
0.001      0.999   -4.35e+04    4.36e+04
What is your current occupation_Working Professional            2.6977      0.189
14.288      0.000       2.328       3.068
Last Notable Activity_Email Opened                             -1.7020      0.095
-17.884      0.000      -1.888      -1.515
Last Notable Activity_Had a Phone Conversation                  2.6428      1.178
2.244      0.025       0.334       4.951
Last Notable Activity_Modified                                 -1.4823      0.087
-17.116      0.000      -1.652      -1.313
Last Notable Activity_Olark Chat Conversation                  -1.4356      0.321
-4.466      0.000      -2.066      -0.806
Lead Profile_Dual Specialization Student                       23.3031    1.7e+04
0.001      0.999   -3.33e+04    3.33e+04
Lead Profile_Lateral Student                                    4.1312      1.138
3.629      0.000       1.900       6.363
Lead Profile_Potential Lead                                     2.0514      0.094
21.812      0.000       1.867       2.236
Lead Profile_Student of SomeSchool                             -1.7550      0.413
-4.245      0.000      -2.565      -0.945
================================================================================
========================================
"""
```

[108]:

```
# Make a VIF dataframe for all the variables present
```

```
vif = pd.DataFrame()
vif['Features'] = X_train_new.columns
vif['VIF'] = [variance_inflation_factor(X_train_new.values, i) for i in␣
 ↪range(X_train_new.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

[108]:

```
                                             Features   VIF
1                                       Last Activity  4.32
6                   Last Notable Activity_Email Opened  2.67
0                         Total Time Spent on Website  1.90
8                      Last Notable Activity_Modified  1.55
2                              Lead Source_Olark Chat  1.40
12                        Lead Profile_Potential Lead  1.35
5   What is your current occupation_Working Profes…  1.21
9       Last Notable Activity_Olark Chat Conversation  1.07
3                         Lead Source_Welingak Website  1.04
13                   Lead Profile_Student of SomeSchool  1.04
7      Last Notable Activity_Had a Phone Conversation  1.01
10            Lead Profile_Dual Specialization Student  1.01
11                        Lead Profile_Lateral Student  1.01
4           What is your current occupation_Housewife  1.00
```

The variable `Lead Profile_Dual Specialization Student` also needs to be dropped.

The VIFs are now all less than 5. So let's drop the ones with the high p-values beginning with `Last Notable Activity_Had a Phone Conversation`.

[109]:

```
X_train_new.drop('Last Notable Activity_Had a Phone Conversation', axis = 1,␣
 ↪inplace = True)
```

[110]:

```
# Refit the model with the new set of features

logm1 = sm.GLM(y_train,(sm.add_constant(X_train_new)), family = sm.families.
 ↪Binomial())
logm1.fit().summary()
```

[110]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6454
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2720.0
Date:                Wed, 09 Aug 2023   Deviance:                       5440.0
```

```
Time:                          18:45:54   Pearson chi2:                 6.69e+03
No. Iterations:                      22   Pseudo R-squ. (CS):             0.3863
Covariance Type:              nonrobust
================================================================================
========================================
                                                           coef    std err
z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
----------------------------------------
const                                                   -2.3290      0.113
-20.590      0.000      -2.551      -2.107
Total Time Spent on Website                              3.7716      0.155
24.411      0.000       3.469       4.074
Last Activity                                            4.8893      0.343
14.265      0.000       4.218       5.561
Lead Source_Olark Chat                                   0.7644      0.097
7.866      0.000       0.574       0.955
Lead Source_Welingak Website                             5.3006      0.722
7.342      0.000       3.885       6.716
What is your current occupation_Housewife               24.7623   2.23e+04
0.001      0.999   -4.36e+04    4.36e+04
What is your current occupation_Working Professional     2.6973      0.189
14.301      0.000       2.328       3.067
Last Notable Activity_Email Opened                      -1.7002      0.095
-17.864      0.000      -1.887      -1.514
Last Notable Activity_Modified                          -1.4946      0.086
-17.294      0.000      -1.664      -1.325
Last Notable Activity_Olark Chat Conversation           -1.4553      0.321
-4.528      0.000      -2.085      -0.825
Lead Profile_Dual Specialization Student                23.3063    1.7e+04
0.001      0.999   -3.33e+04    3.33e+04
Lead Profile_Lateral Student                             4.1291      1.136
3.634      0.000       1.902       6.356
Lead Profile_Potential Lead                              2.0622      0.094
21.964      0.000       1.878       2.246
Lead Profile_Student of SomeSchool                      -1.7562      0.413
-4.249      0.000      -2.566      -0.946
================================================================================
========================================
"""
```

Drop `What is your current occupation_Housewife`.

[111]:
```
X_train_new.drop('What is your current occupation_Housewife', axis = 1, inplace
 ↪= True)
```

[112]:

```
# Refit the model with the new set of features

logm1 = sm.GLM(y_train,(sm.add_constant(X_train_new)), family = sm.families.
 ↪Binomial())
logm1.fit().summary()
```

[112]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6455
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2732.0
Date:                Wed, 09 Aug 2023   Deviance:                       5464.1
Time:                        18:45:54   Pearson chi2:                 6.69e+03
No. Iterations:                    22   Pseudo R-squ. (CS):             0.3840
Covariance Type:            nonrobust
================================================================================
========================================
                                                          coef    std err
z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
----------------------------------------
const                                                  -2.3123      0.113
-20.526      0.000      -2.533      -2.091
Total Time Spent on Website                             3.7555      0.154
24.371      0.000      3.453      4.058
Last Activity                                           4.8574      0.341
14.225      0.000      4.188      5.527
Lead Source_Olark Chat                                  0.7515      0.097
7.749      0.000      0.561      0.942
Lead Source_Welingak Website                            5.2882      0.722
7.325      0.000      3.873      6.703
What is your current occupation_Working Professional     2.6877      0.188
14.260      0.000      2.318      3.057
Last Notable Activity_Email Opened                     -1.6879      0.095
-17.781      0.000      -1.874      -1.502
Last Notable Activity_Modified                         -1.4890      0.086
-17.273      0.000      -1.658      -1.320
Last Notable Activity_Olark Chat Conversation          -1.4560      0.321
-4.534      0.000      -2.085      -0.827
Lead Profile_Dual Specialization Student               23.3023    1.7e+04
0.001      0.999   -3.33e+04    3.34e+04
Lead Profile_Lateral Student                            4.1141      1.135
3.624      0.000      1.889      6.339
```

```
Lead Profile_Potential Lead                              2.0588      0.094
21.980      0.000      1.875      2.242
Lead Profile_Student of SomeSchool                      -1.7632      0.413
-4.269      0.000     -2.573     -0.954
================================================================================
========================================
"""
```

Drop What is your current occupation_Working Professional.

[113]:
```python
X_train_new.drop('What is your current occupation_Working Professional', axis =␣
↪1, inplace = True)
```

[114]:
```python
# Refit the model with the new set of features

logm1 = sm.GLM(y_train,(sm.add_constant(X_train_new)), family = sm.families.
↪Binomial())
res = logm1.fit()
res.summary()
```

[114]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6456
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2887.7
Date:                Wed, 09 Aug 2023   Deviance:                       5775.4
Time:                        18:45:54   Pearson chi2:                 6.84e+03
No. Iterations:                    21   Pseudo R-squ. (CS):             0.3537
Covariance Type:            nonrobust
================================================================================
=================================
                                                  coef    std err          z
P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
---------------------------------
const                                          -2.1342      0.108    -19.733
0.000      -2.346      -1.922
Total Time Spent on Website                     3.7289      0.150     24.850
0.000       3.435       4.023
Last Activity                                   4.7854      0.329     14.557
0.000       4.141       5.430
Lead Source_Olark Chat                          0.6590      0.095      6.970
0.000       0.474       0.844
```

```
Lead Source_Welingak Website                         5.1424      0.722      7.124
0.000       3.728       6.557
Last Notable Activity_Email Opened                  -1.7063      0.092    -18.570
0.000      -1.886      -1.526
Last Notable Activity_Modified                      -1.5142      0.083    -18.169
0.000      -1.678      -1.351
Last Notable Activity_Olark Chat Conversation       -1.4964      0.309     -4.835
0.000      -2.103      -0.890
Lead Profile_Dual Specialization Student            22.9052   1.13e+04      0.002
0.998   -2.21e+04    2.22e+04
Lead Profile_Lateral Student                         4.2559      1.098      3.875
0.000       2.103       6.408
Lead Profile_Potential Lead                          2.3057      0.090     25.563
0.000       2.129       2.483
Lead Profile_Student of SomeSchool                  -1.7728      0.410     -4.327
0.000      -2.576      -0.970
================================================================================
=================================
"""
```

All the p-values are now in the appropriate range. Let's also check the VIFs again in case we had missed something.

[115]:
```python
# Make a VIF dataframe for all the variables present

vif = pd.DataFrame()
vif['Features'] = X_train_new.columns
vif['VIF'] = [variance_inflation_factor(X_train_new.values, i) for i in range(X_train_new.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

[115]:

	Features	VIF
1	Last Activity	4.25
4	Last Notable Activity_Email Opened	2.65
0	Total Time Spent on Website	1.89
5	Last Notable Activity_Modified	1.55
2	Lead Source_Olark Chat	1.40
9	Lead Profile_Potential Lead	1.25
6	Last Notable Activity_Olark Chat Conversation	1.07
3	Lead Source_Welingak Website	1.04
10	Lead Profile_Student of SomeSchool	1.04
7	Lead Profile_Dual Specialization Student	1.01
8	Lead Profile_Lateral Student	1.01

We are good to go!

1.5 Step 3: Model Evaluation

Now, both the p-values and VIFs seem decent enough for all the variables. So let's go ahead and make predictions using this final set of features.

[116]:
```
final_col = X_train_new.columns
```

[117]:
```
X_test_new = X_test[final_col]
```

[118]:
```
X_test_new
```

[118]:
```
      Total Time Spent on Website  Last Activity  Lead Source_Olark Chat  \
4269                     0.444982       0.297078                       0
2376                     0.000000       0.297078                       0
7766                     0.025968       0.010065                       0
9199                     0.000000       0.105303                       1
4359                     0.000000       0.383117                       0
…                             …            …                     …
8649                     0.127641       0.069264                       0
2152                     0.000000       0.297078                       0
7101                     0.000000       0.383117                       1
5331                     0.707746       0.069264                       0
2960                     0.000000       0.297078                       0

      Lead Source_Welingak Website  Last Notable Activity_Email Opened  \
4269                             0                                   0
2376                             0                                   0
7766                             0                                   0
9199                             0                                   0
4359                             0                                   1
…                            …                                   …
8649                             0                                   0
2152                             0                                   0
7101                             0                                   1
5331                             0                                   0
2960                             0                                   0

      Last Notable Activity_Modified  \
4269                               0
2376                               0
7766                               0
9199                               1
4359                               0
…                              …
8649                               0
2152                               0
7101                               0
5331                               1
```

```
2960                              0

      Last Notable Activity_Olark Chat Conversation  \
4269                                              0
2376                                              0
7766                                              0
9199                                              0
4359                                              0
…                                             …
8649                                              0
2152                                              0
7101                                              0
5331                                              0
2960                                              0

      Lead Profile_Dual Specialization Student  Lead Profile_Lateral Student  \
4269                                         0                             0
2376                                         0                             0
7766                                         0                             0
9199                                         0                             0
4359                                         0                             0
…                                        …                            …
8649                                         0                             0
2152                                         0                             0
7101                                         0                             0
5331                                         0                             0
2960                                         0                             0

      Lead Profile_Potential Lead  Lead Profile_Student of SomeSchool
4269                            0                                   0
2376                            0                                   0
7766                            0                                   0
9199                            1                                   0
4359                            1                                   0
…                           …                                 …
8649                            0                                   0
2152                            0                                   0
7101                            0                                   0
5331                            0                                   0
2960                            0                                   0

[2772 rows x 11 columns]
```

1.6 Building New model after doing feature selection and removing highly correlated values and statistically insignificant feature.

[119]:
```
model_new = LogisticRegression()
model_new.fit(X_train_new,y_train)
```

[119]:
```
LogisticRegression()
```

[120]:
```
model_new.predict(X_train_new)
```

[120]:
```
array([0, 0, 0, …, 0, 0, 0])
```

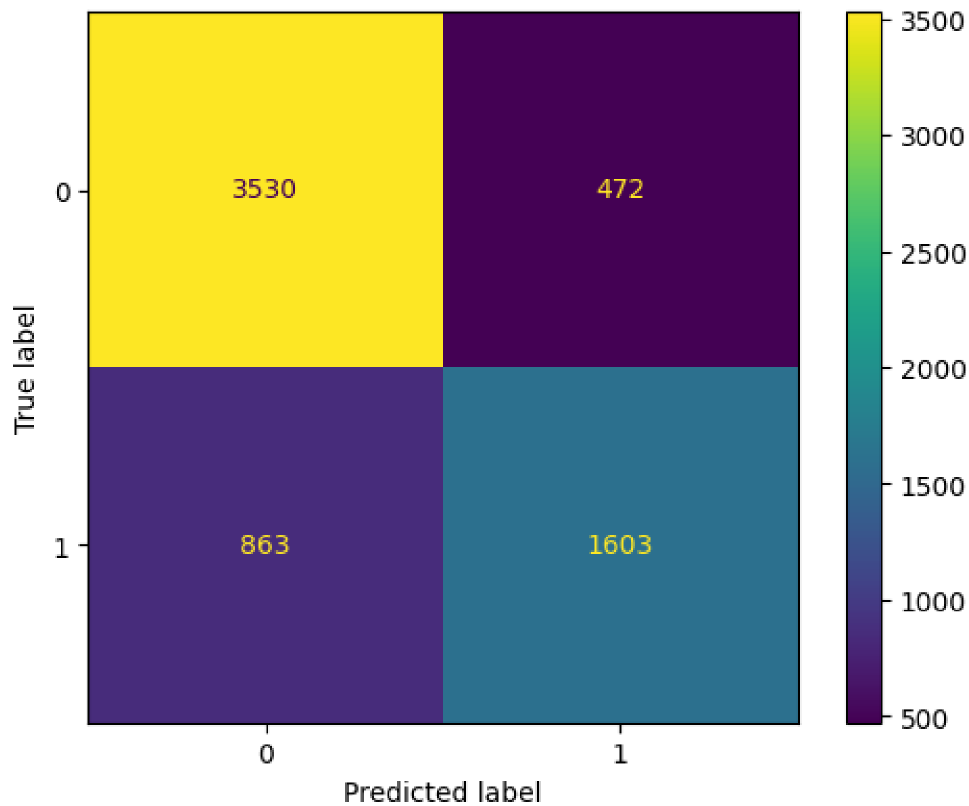
[121]:
```
model_new.predict_proba(X_train_new)[:,1]
```

[121]:
```
array([0.20690814, 0.23649864, 0.23574585, …, 0.20690814, 0.05219164,
       0.22854076])
```

1.7 Train and Test evaluation on new model.

[122]:
```
y_train
y_train_pred = model_new.predict(X_train_new)
confusion_matrix(y_train,y_train_pred)
ConfusionMatrixDisplay.from_estimator(model_new, X_train_new, y_train)
```

[122]:
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at
0x7f79e3efbe50>
```



[123]:
```
# accuracy_score(y_train,y_train_pred)
# precision_score(y_train,y_train_pred)
# recall_score(y_train,y_train_pred)
# f1_score(y_train,y_train_pred)
```

[124]:
```
accuracy_score(y_train,y_train_pred)
```

[124]: 0.7935992578849722

[125]:
```
precision_score(y_train,y_train_pred)
```

[125]: 0.7725301204819277

[126]:
```
recall_score(y_train,y_train_pred)
```

[126]: 0.6500405515004055

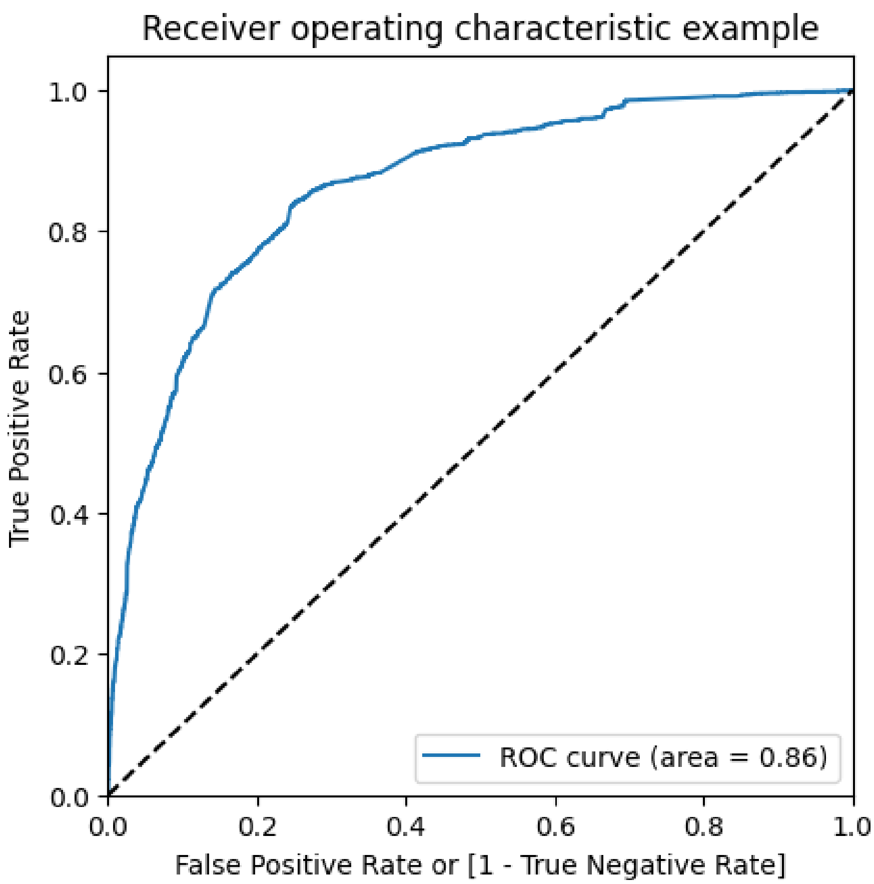
[127]:
```
f1_score(y_train,y_train_pred)
```

[127]: 0.7060118916538207

[128]:
```
roc_auc_score(y_train,model_new.predict_proba(X_train_new)[:,1])
```

[128]: 0.8628130176598644

[129]:
```
draw_roc(y_train, model_new.predict_proba(X_train_new)[:,1])
```



1.8 Test

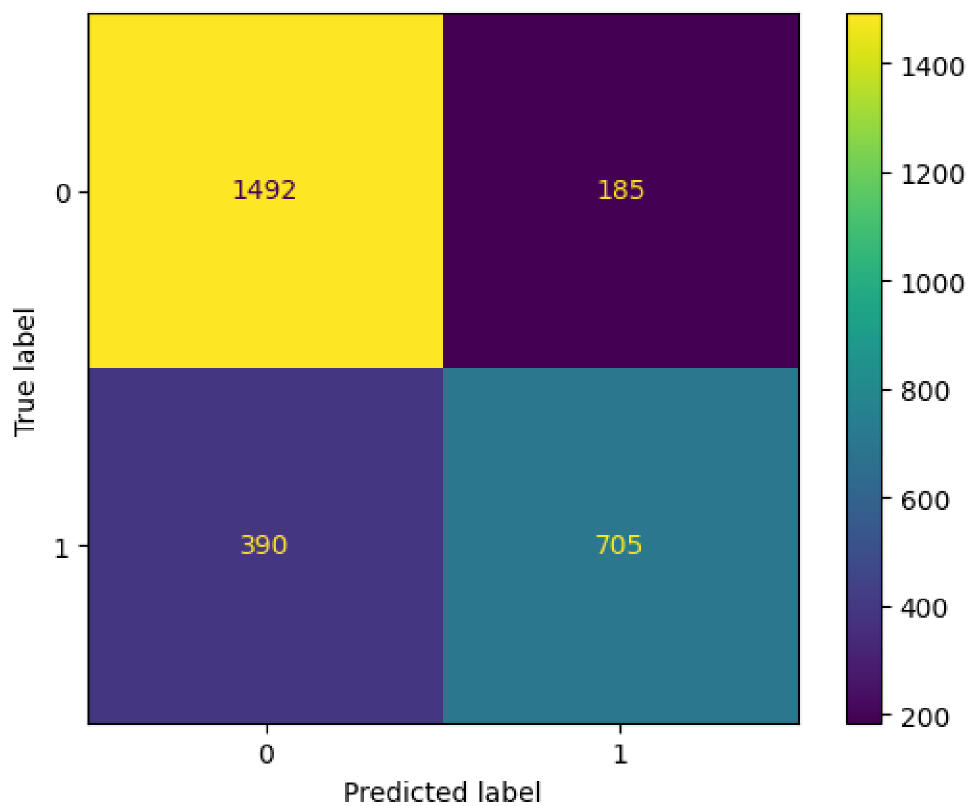
[130]:
```
y_test
```

[130]:
```
      Converted
4269          1
2376          1
7766          1
9199          0
4359          1
…           …
8649          0
2152          1
```

```
7101          0
5331          0
2960          1

[2772 rows x 1 columns]
```

[131]:
```
y_test_pred = model_new.predict(X_test_new)
confusion_matrix(y_test,y_test_pred)
ConfusionMatrixDisplay.from_estimator(model_new, X_test_new, y_test)
```

[131]:
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at
0x7f79e99980d0>
```



[132]:
```
accuracy_score(y_test,y_test_pred)
```

[132]:
```
0.7925685425685426
```

[133]:
```
precision_score(y_test,y_test_pred)
```

[133]:
```
0.7921348314606742
```

[134]:
```
recall_score(y_test,y_test_pred)
```

[134]: 0.6438356164383562

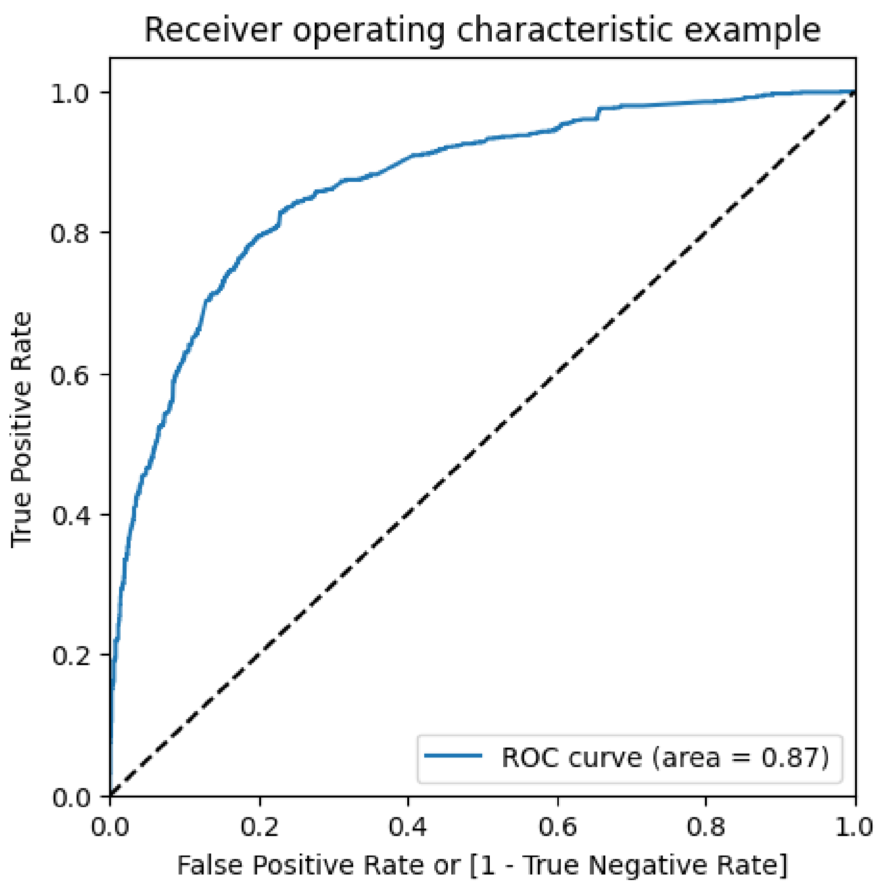
[135]:
```
f1_score(y_test,y_test_pred)
```

[135]: 0.7103274559193954

[136]:
```
roc_auc_score(y_test,model_new.predict_proba(X_test_new)[:,1])
```

[136]: 0.8669487533456951

[137]:
```
draw_roc(y_test,model_new.predict_proba(X_test_new)[:,1])
```



1.9 Changing Hyperparameters of the logistic Regression Model.

[138]:
```
from sklearn.linear_model import LogisticRegression
```

[139]:
```
logistic_model = LogisticRegression(penalty='elasticnet', tol=0.00001 , C=100.
↪0, solver='saga', max_iter=10000, l1_ratio=0.70)
```

[140]:
```
logistic_model.fit(X_train_new,y_train)
```

[140]:
```
LogisticRegression(C=100.0, l1_ratio=0.7, max_iter=10000, penalty='elasticnet',
                   solver='saga', tol=1e-05)
```

[141]:
```
roc_auc_score(y_test,logistic_model.predict_proba(X_test_new)[:,1])
```

[141]:
```
0.867285841481445
```

[142]:
```
from sklearn.neighbors import KNeighborsClassifier
```

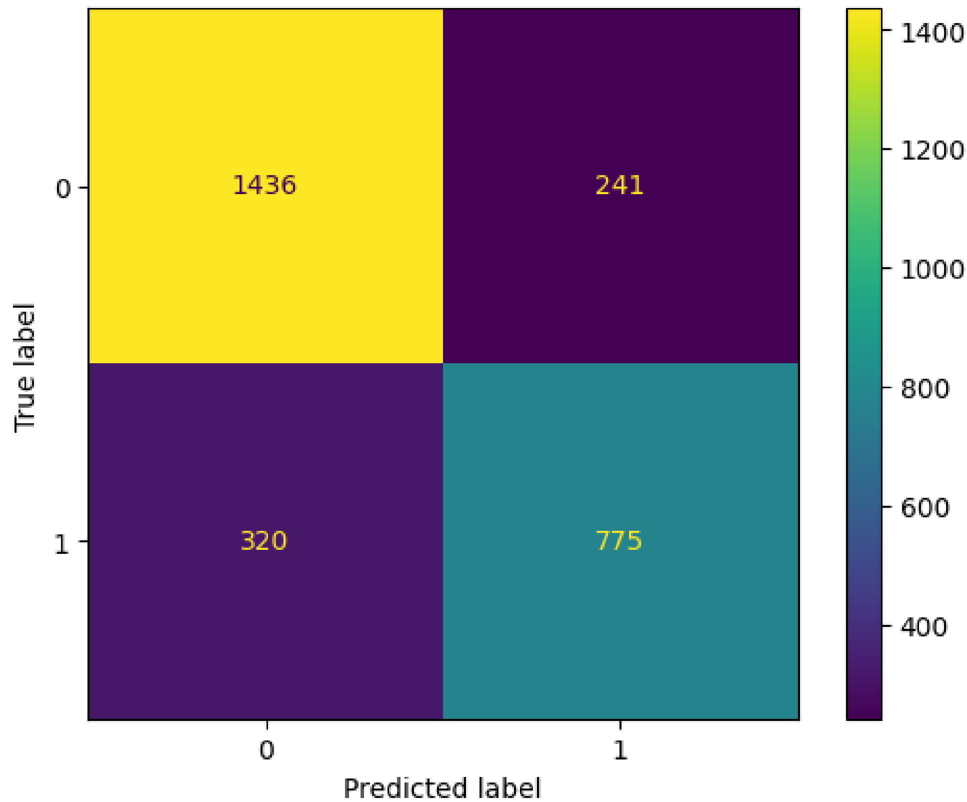
[143]:
```
nn_model = KNeighborsClassifier()
```

[144]:
```
nn_model.fit(X_train_new,y_train)
```

[144]:
```
KNeighborsClassifier()
```

[145]:
```
y_test_pred = nn_model.predict(X_test_new)
confusion_matrix(y_test,y_test_pred)
ConfusionMatrixDisplay.from_estimator(nn_model, X_test_new, y_test)
```

[145]:
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at
0x7f79e3e2ff70>
```



[146]:
```
accuracy_score(y_test,y_test_pred)
```

[146]: 0.7976190476190477

[147]:
```
precision_score(y_test,y_test_pred)
```

[147]: 0.7627952755905512

[148]:
```
recall_score(y_test,y_test_pred)
```

[148]: 0.7077625570776256

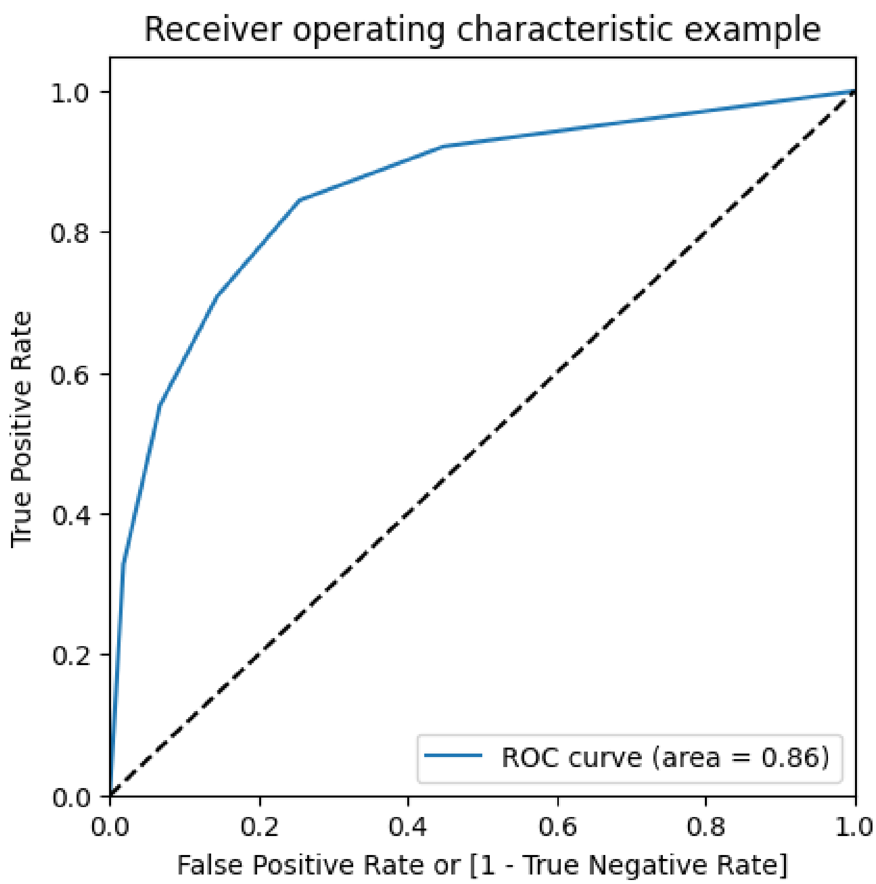
[149]:
```
f1_score(y_test,y_test_pred)
```

[149]: 0.7342491710090004

[150]:
```
roc_auc_score(y_test,nn_model.predict_proba(X_test_new)[:,1])
```

[150]: 0.8598679965038678

[151]:
```
draw_roc(y_test,nn_model.predict_proba(X_test_new)[:,1])
```


Receiver operating characteristic example
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
ROC curve (area = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]

[151]:

[151]: