LEAD SCORING CASE STUDY

By
Abhiram Vijay, Mohammed Aatif Mannur, Abhishek Chauhan

Problem Statement

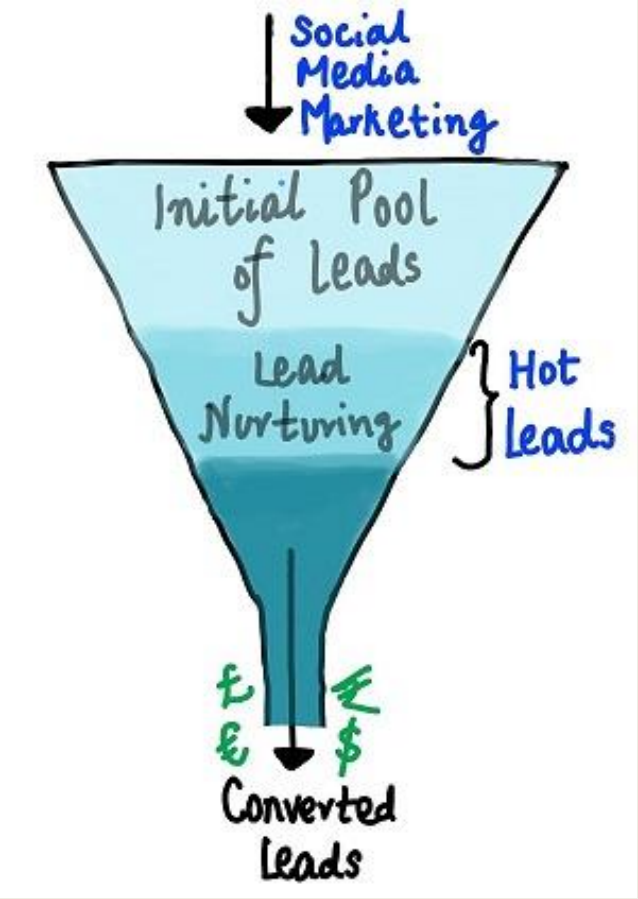
An education online course company, X Education, sells online courses to industry professionals needs help with low lead conversion rate of approximately 30%

The company get leads through website visits, search engine marketing, referrals, and form submissions, but only a small fraction of these leads convert into paying customers.

A typical lead conversion process is show in funnel representation :

To increase their sales and optimize efforts and improve efficiency, X Education aims to identify "Hot Leads" — leads most likely to convert.

As shown in funnel, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom.



Objective

- Build a logistic regression model to assist X education to select most promising leads based on 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.

Data Understanding

- We will start with Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
from sklearn.metrics import precision_score, recall_score
from sklearn.metrics import precision_recall_curve
from sklearn.metrics import f1_score

#statmodel libraries
from statsmodels.stats.outliers_influence import variance_inflation_factor
import statsmodels.api as sm
```

```
import warnings
warnings.filterwarnings("ignore")
```

```
pd.set_option("display.max_columns",None)
pd.set_option("display.max_rows",None)
pd.set_option('display.width',None)
```

```
df_leads = pd.read_csv("Leads.csv")
```

Data Understanding

```
# Check the summary of the dataset

df_leads.describe()
```

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Data Understanding

```
# Check the info to see the types of the feature variables and the null values present

df_leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype  
---  ------                                         --------------  -----  
 0   Prospect ID                                    9240 non-null   object 
 1   Lead Number                                    9240 non-null   int64  
 2   Lead Origin                                    9240 non-null   object 
 3   Lead Source                                    9204 non-null   object 
 4   Do Not Email                                   9240 non-null   object 
 5   Do Not Call                                    9240 non-null   object 
 6   Converted                                      9240 non-null   int64  
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64  
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object 
 11  Country                                        6779 non-null   object 
 12  Specialization                                 7802 non-null   object 
 13  How did you hear about X Education             7033 non-null   object 
 14  What is your current occupation                6550 non-null   object 
 15  What matters most to you in choosing a course  6531 non-null   object 
 16  Search                                         9240 non-null   object 
 17  Magazine                                       9240 non-null   object 
 18  Newspaper Article                              9240 non-null   object 
 19  X Education Forums                             9240 non-null   object 
 20  Newspaper                                      9240 non-null   object 
 21  Digital Advertisement                          9240 non-null   object 
 22  Through Recommendations                        9240 non-null   object 
 23  Receive More Updates About Our Courses         9240 non-null   object 
 24  Tags                                           5887 non-null   object 
 25  Lead Quality                                   4473 non-null   object 
 26  Update me on Supply Chain Content              9240 non-null   object 
 27  Get updates on DM Content                      9240 non-null   object 
 28  Lead Profile                                   6531 non-null   object 
 29  City                                           7820 non-null   object 
 30  Asymmetrique Activity Index                    5022 non-null   object 
 31  Asymmetrique Profile Index                     5022 non-null   object 
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object 
 35  A free copy of Mastering The Interview         9240 non-null   object 
 36  Last Notable Activity                          9240 non-null   object 
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Data Understanding

```
# Check the number of missing values in each column

df_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
```

Data Cleaning and Preparation

Check the number of missing values in each column

```
# Check the number of missing values in each column

df_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
```

Data Cleaning and Preparation

Checking the Missing Values

```
100*(df_leads.isna().mean()).sort_values(ascending=False)
```

```
How did you hear about X Education               78.463203
Lead Profile                                     74.188312
Lead Quality                                     51.590909
Asymmetrique Profile Score                       45.649351
Asymmetrique Activity Score                      45.649351
Asymmetrique Activity Index                      45.649351
Asymmetrique Profile Index                       45.649351
City                                             39.707792
Specialization                                   36.580087
Tags                                             36.287879
What matters most to you in choosing a course    29.318182
What is your current occupation                  29.112554
Country                                          26.634199
Page Views Per Visit                              1.482684
TotalVisits                                       1.482684
Last Activity                                     1.114719
Lead Source                                       0.389610
Receive More Updates About Our Courses            0.000000
I agree to pay the amount through cheque          0.000000
Get updates on DM Content                         0.000000
Update me on Supply Chain Content                 0.000000
A free copy of Mastering The Interview            0.000000
Prospect ID                                       0.000000
Newspaper Article                                 0.000000
Through Recommendations                           0.000000
Digital Advertisement                             0.000000
Newspaper                                         0.000000
X Education Forums                                0.000000
Lead Number                                       0.000000
Magazine                                          0.000000
Search                                            0.000000
Total Time Spent on Website                       0.000000
Converted                                         0.000000
Do Not Call                                       0.000000
Do Not Email                                      0.000000
Lead Origin                                       0.000000
Last Notable Activity                             0.000000
dtype: float64
```

Dropping columns with more than 40%

Dropped columns:

How did you hear about X Education

Lead Profile

Lead Quality

symmetrique Profile Score

Asymmetrique Activity Score

Asymmetrique Activity Index

Asymmetrique Profile Index

Droping colums with more than 40%

[39]:

```
def dropNullColumns(data ,percentage=40):

    missing_perc = 100*(data.isna().mean()).sort_values(ascending=False)
    col_to_drop = missing_perc[missing_perc>=percentage].index.to_list()
    print("No. Dropped columns: ",len(col_to_drop),"\n")
    print("Dropped columns: " , col_to_drop,"\n")
    print("Before dropping columns: ",data.shape)

    data.drop(labels=col_to_drop,axis=1, inplace=True)

    print("After dropping columns: ",data.shape)
```

[41]:

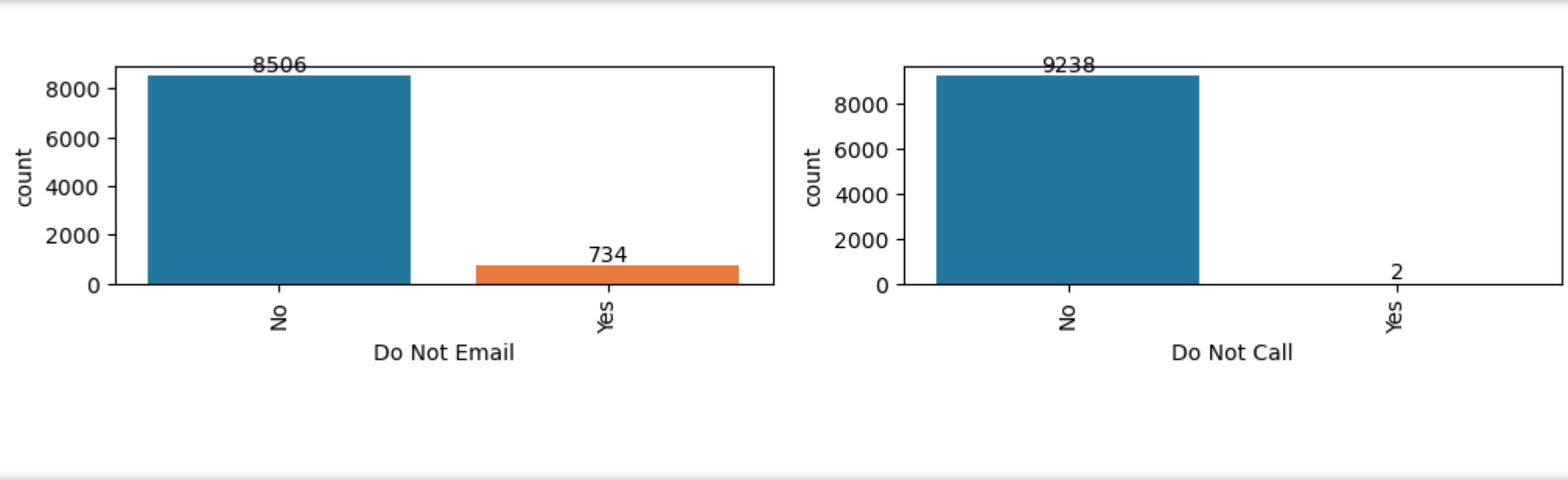
```
dropNullColumns(df_leads)
```

```
No. Dropped columns:  7

Dropped columns:  ['How did you hear about X Education', 'Lead Profile', 'Lead Quality', 'Asymmetr
ique Profile Score', 'Asymmetrique Activity Score', 'Asymmetrique Activity Index', 'Asymmetrique P
rofile Index']

Before dropping columns:  (9240, 37)
After dropping columns:  (9240, 30)
```

EXPLORATORY DATA ANALYSIS

```
# Checking Category columns
```

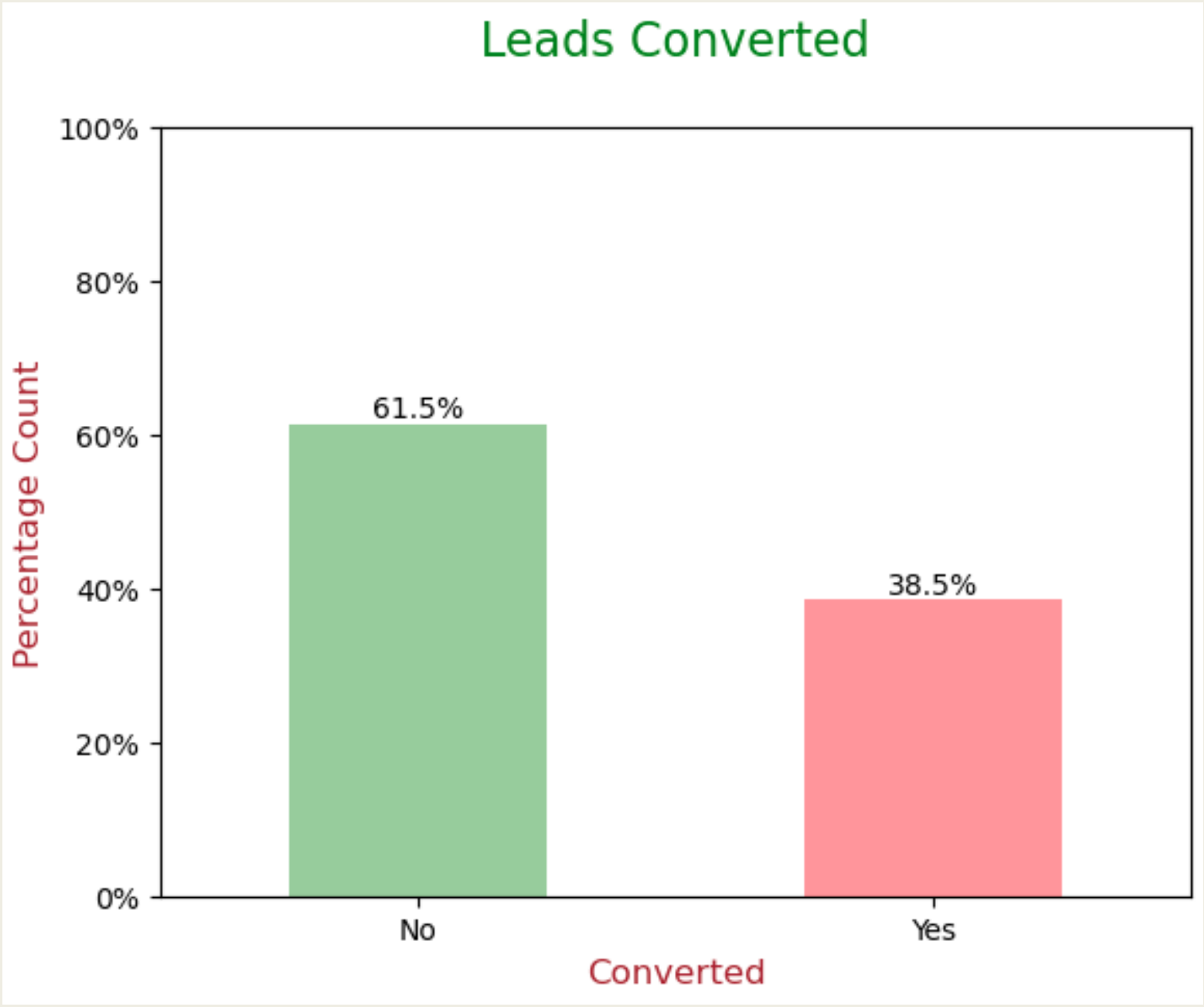
```
categorical_col = df_leads.select_dtypes(include=['category', 'object']).columns.tolist()
plt.figure(figsize=(12,40))

plt.subplots_adjust(wspace=.2,hspace=2)
for i in enumerate(categorical_col):
    plt.subplot(8,2, i[0]+1)
    ax=sns.countplot(x=i[1],data=df_leads)
    plt.xticks(rotation=90)

    for p in ax.patches:
        ax.annotate('{:.0f}'.format(p.get_height()), (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha = 'center', va = 'center', xytext = (0, 5), textcoords = 'offset points')


plt.show()
```



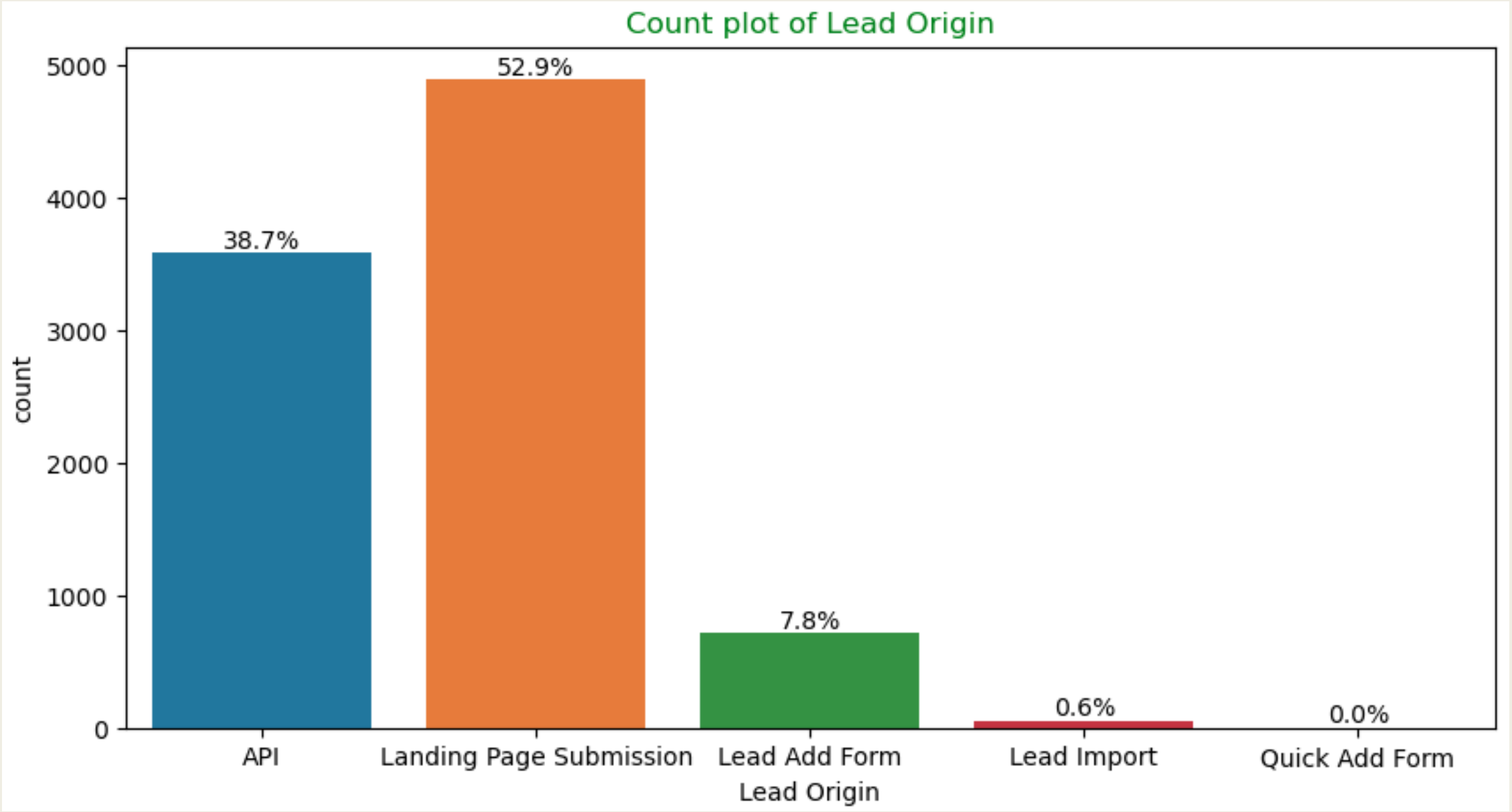
EXPLORATORY DATA ANALYSIS

Outliner Analysis

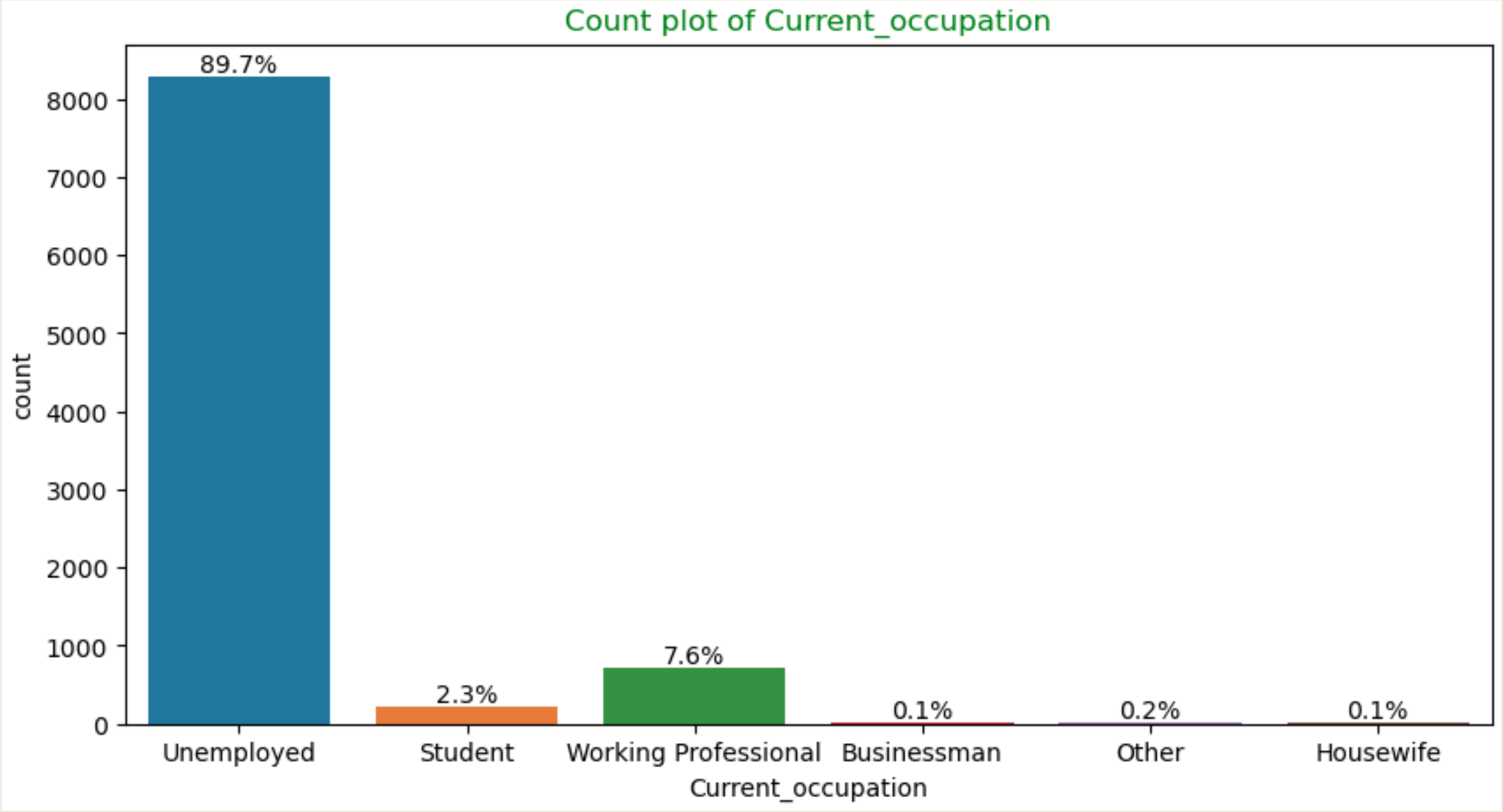
EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

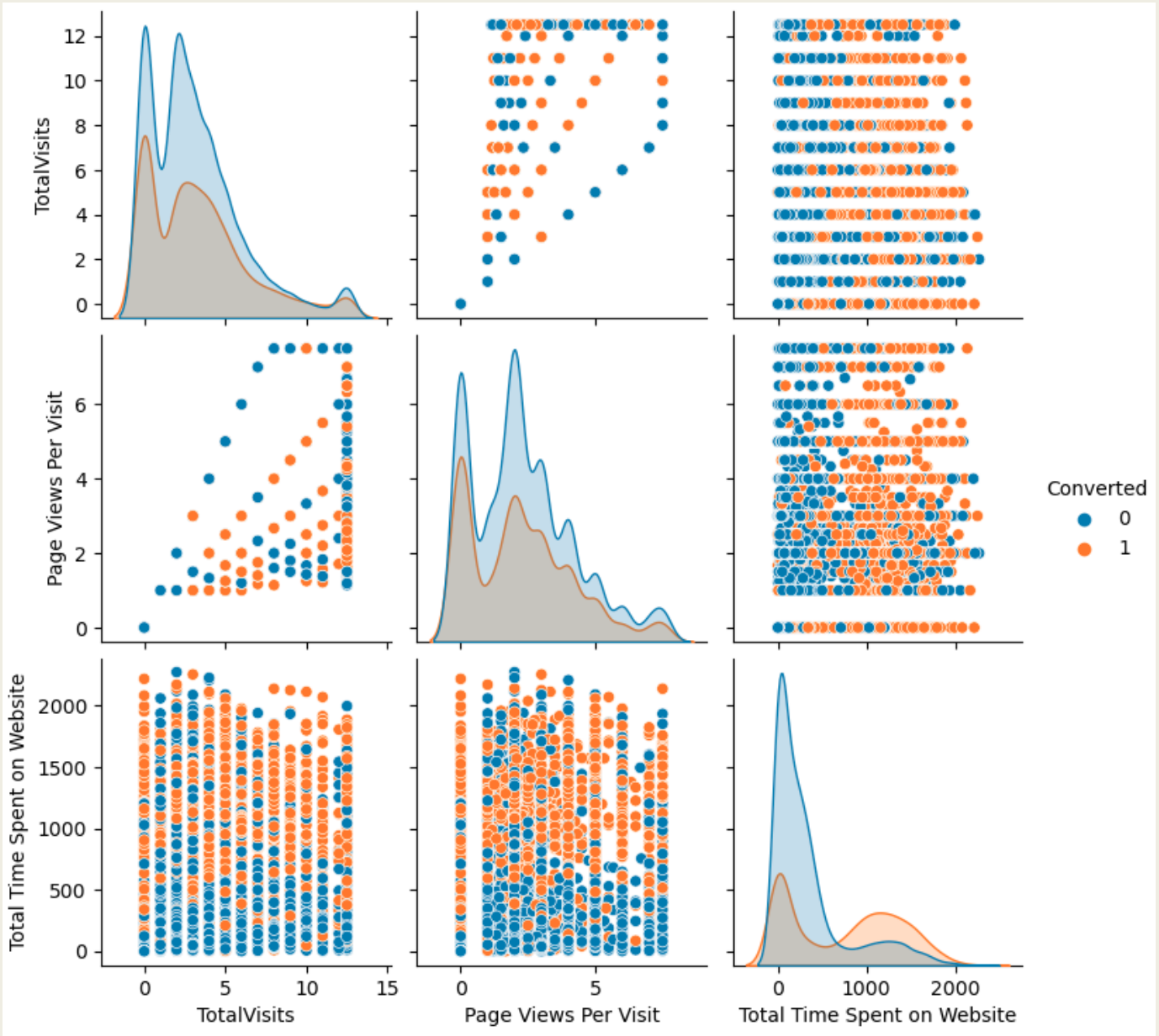


EXPLORATORY DATA ANALYSIS

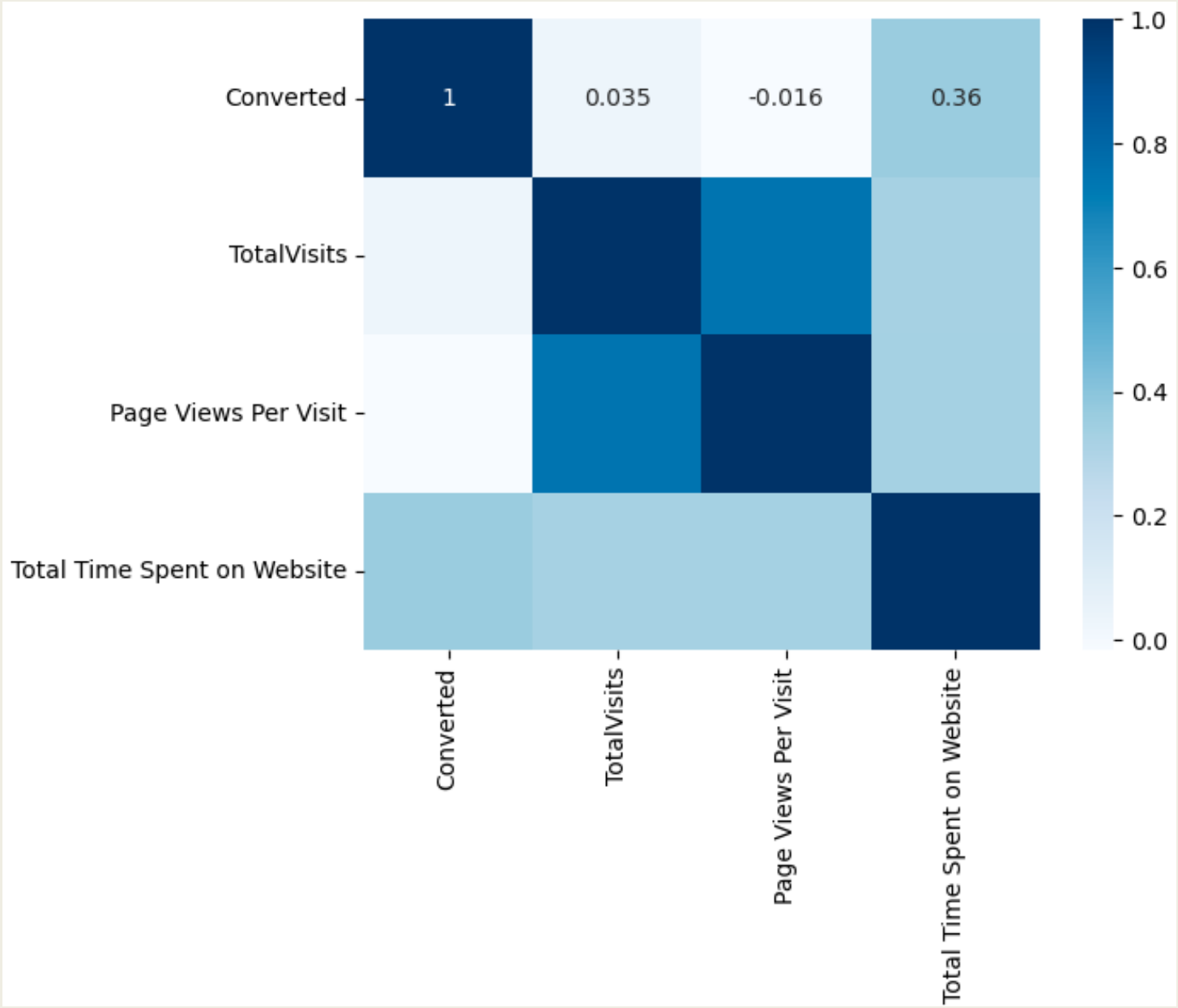


EXPLORATORY DATA ANALYSIS

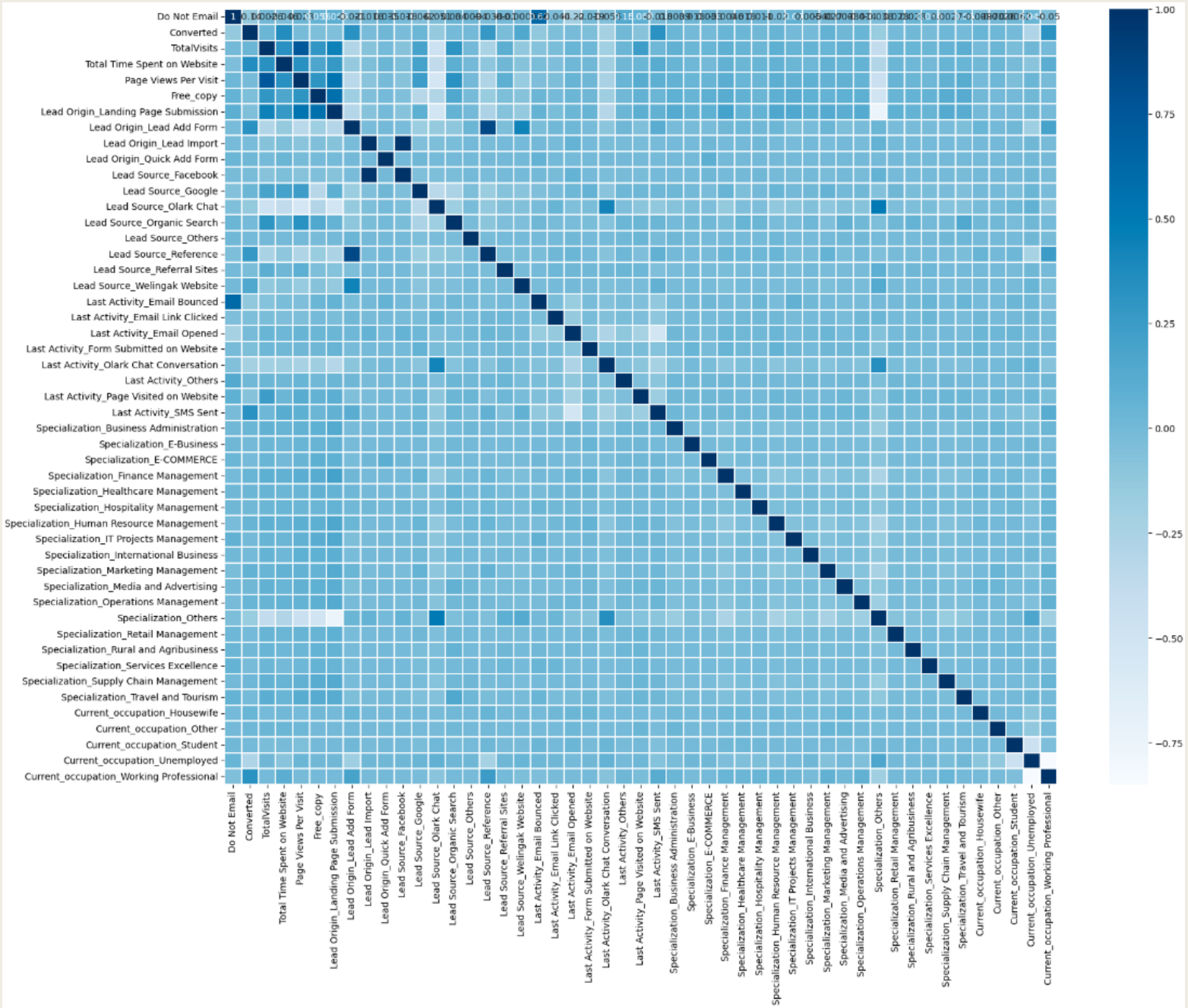
Bivariate Analysis for Numerical Variables



EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

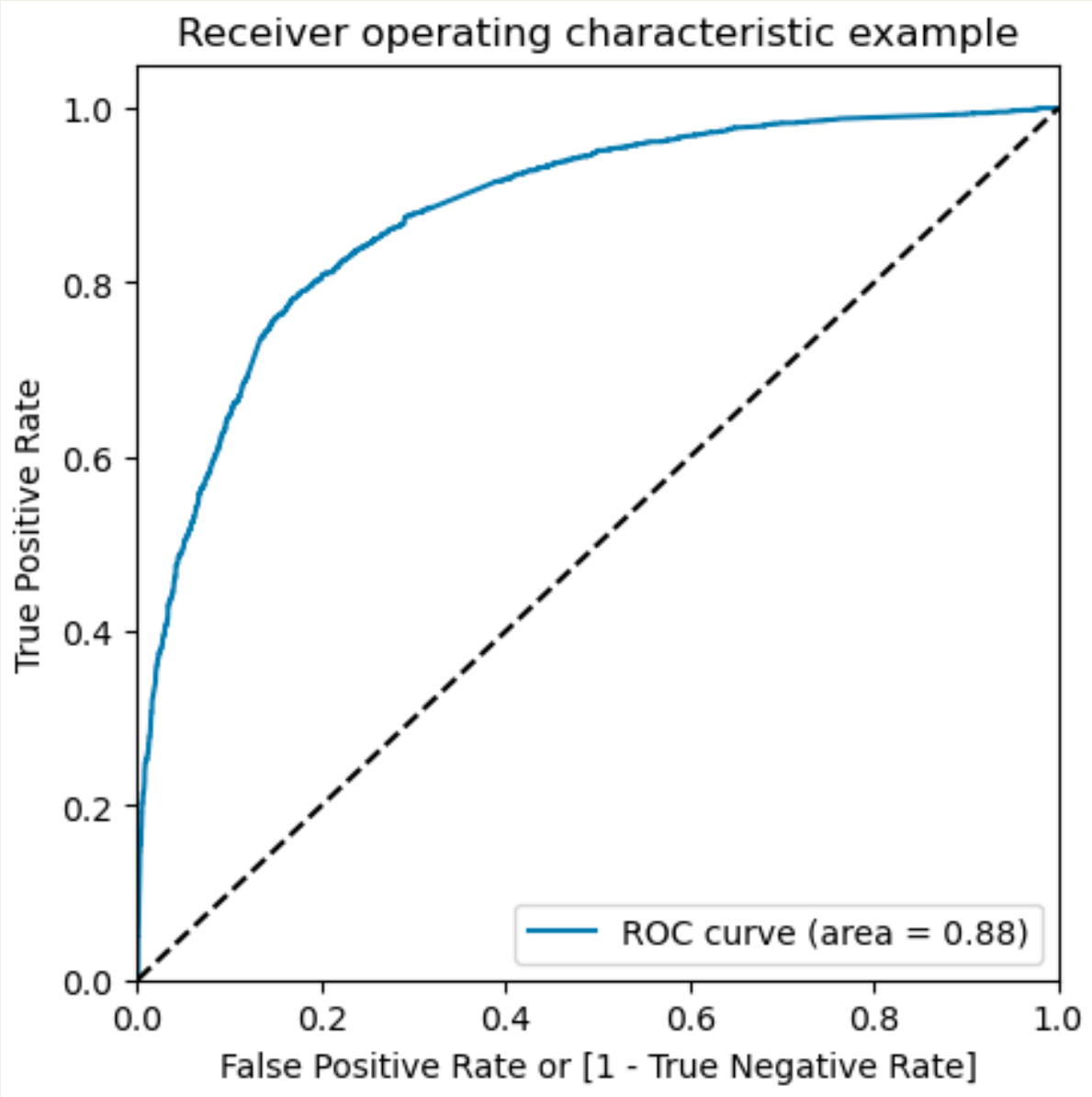


Model Approach

Bivariate Analysis

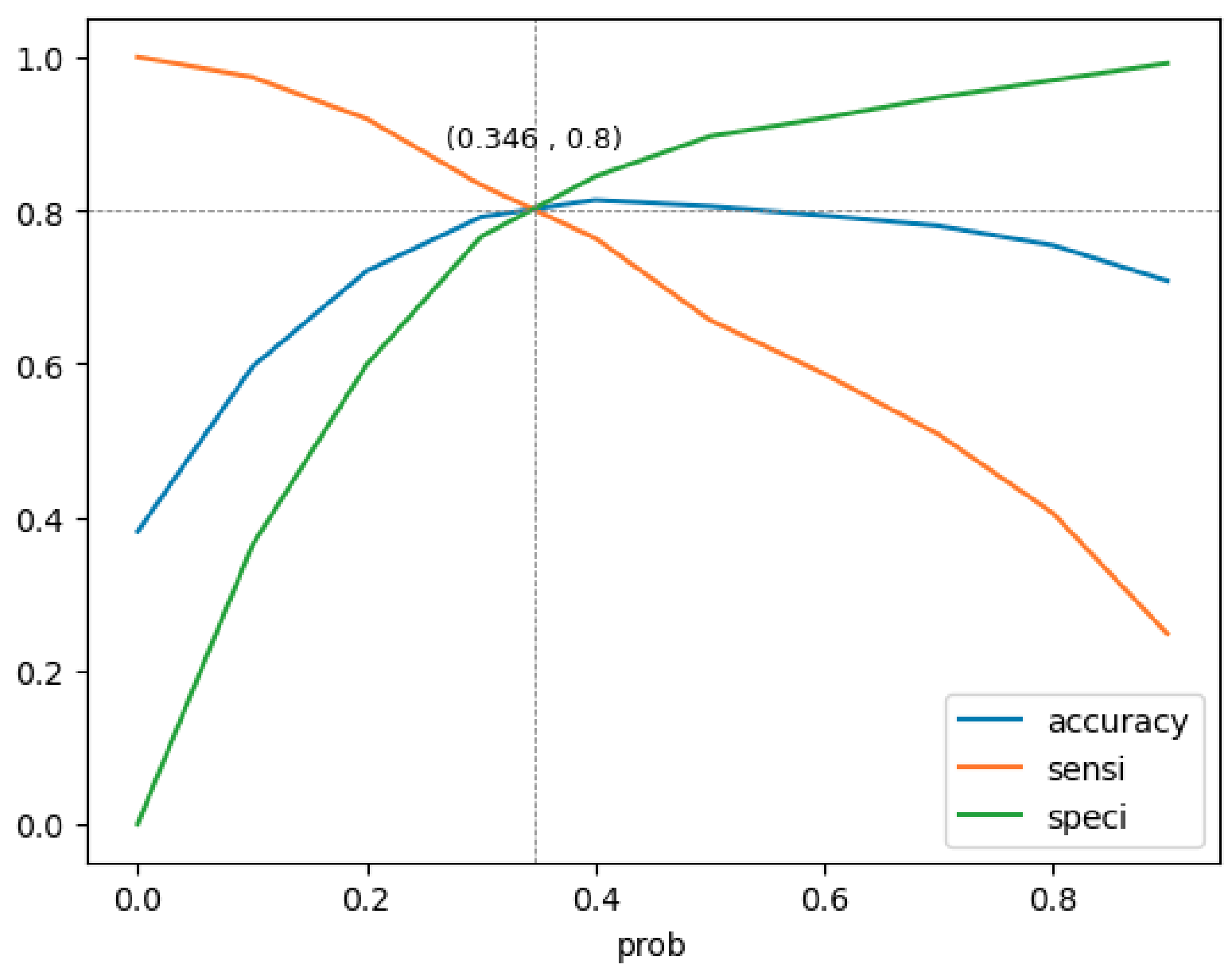
Model Evaluation

Plotting the ROC Curve



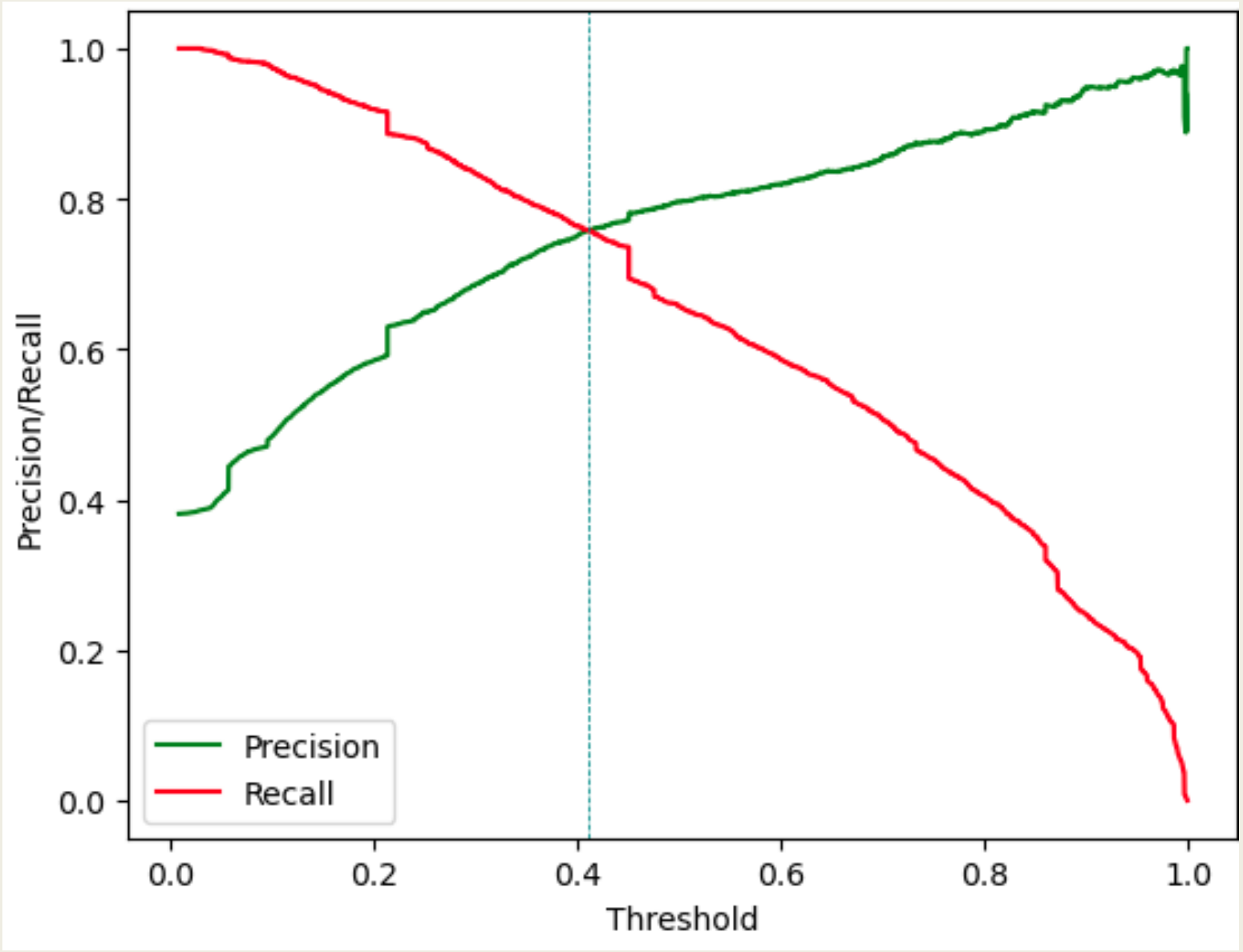
Model Evaluation

let's calculate accuracy sensitivity and specificity for various probability cutoffs.



Model Evaluation

Precision and recall tradeoff



Recommendations and Action plan

Bivariate Analysis

Conclusion

- Train - Test Train Data Set:

Accuracy : 81.22%,

Sensitivity : 78.63%,

Specificity : 82.81%

- Test Data Set:

Accuracy : 80.16%,

Sensitivity : 79.82% ≈ 80%

Specificity : 80.38%