Lead Scoring Case Study

Using Logistic Regression

PROBLEM STATEMENT:

1. X Education sells online courses to industry professionals.

2. X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about30 of them are converted.

3. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’

4. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE:

1. X education wants to know most promising leads.

2. For that they want to build a Model which identifies the hot leads.

3. Deployment of the model for the future use.

4. Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.

5. The CEO want to achieve a lead conversion rate of 80%.

6. They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

METHODOLOGY:

❖ Data cleaning and data manipulation.

❖ EDA (Univariate , Bivariate and Multivariate data analysis)

❖ Feature Scaling, Dummy Variables and encoding of the data.

❖ Classification technique: Logistic Regression.

❖ Validation of the model.

❖ Model presentation.

❖ Conclusions and recommendations.

DATA CLEANING:

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

DATA MANIPULATION:

1. Total Number of Columns = 37, Total Number of Rows = 9240.

2. Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply".

3. Removing the unnecessary columns "Prospect ID" and "Lead Number" for analysis.

4. After checking the value counts for some of the object type variables, we found some features which had no enough variance, so we dropped them. The features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", and "Digital Advertisement".

5. Dropping the columns having more than 30% as missing value such as "How did you hear about X Education" and "Lead Profile".

Exploratory Data Analysis:

Univariate data analysis

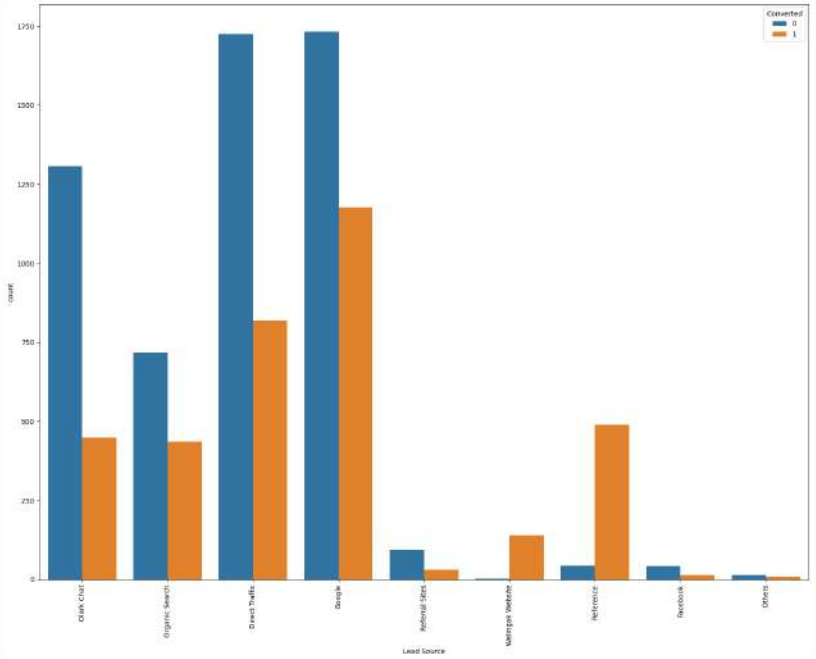
- value count
- distribution of variable

Bivariate data analysis:

- correlation coefficients
- pattern between the variables

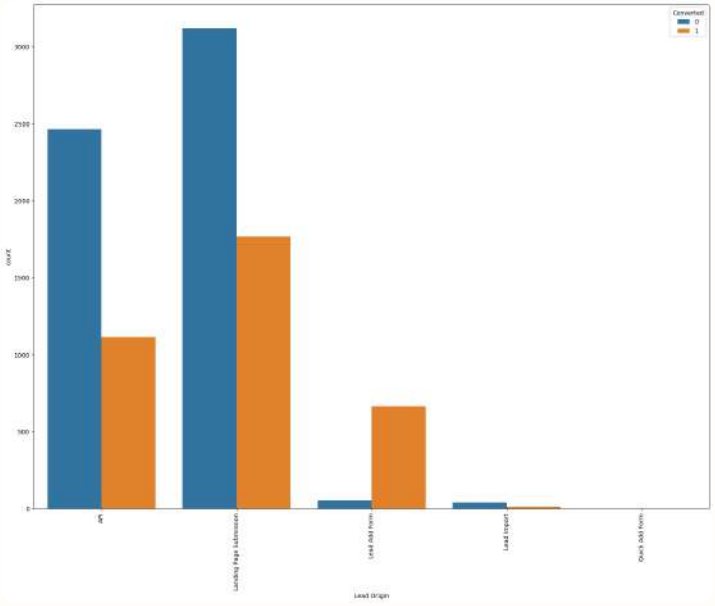
• Very high conversion rates for lead sources ‘Reference’ and ‘Welingak Website’.

• Most leads are generated through ‘Direct Traffic’ and ‘Google’

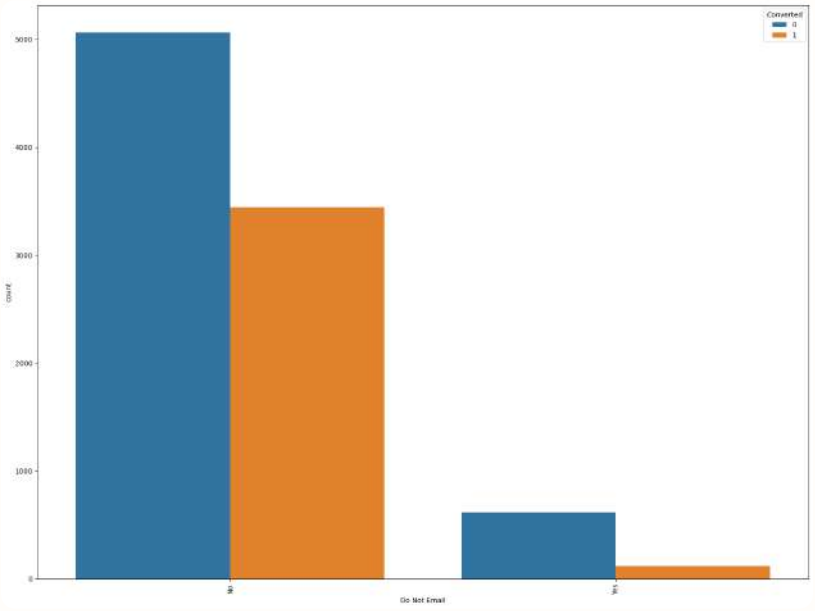


Univariate Analysis:

Lead Origin :



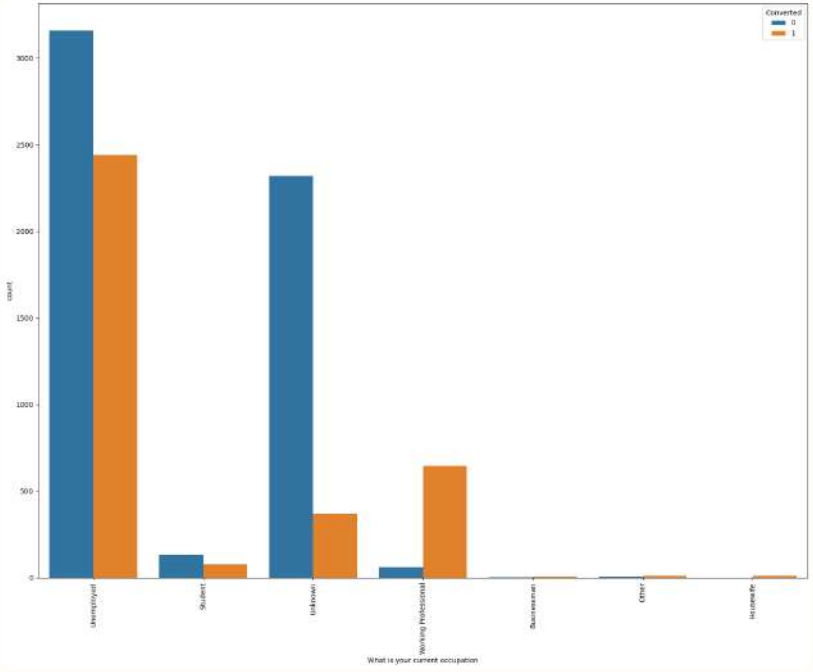
Do Not Email :



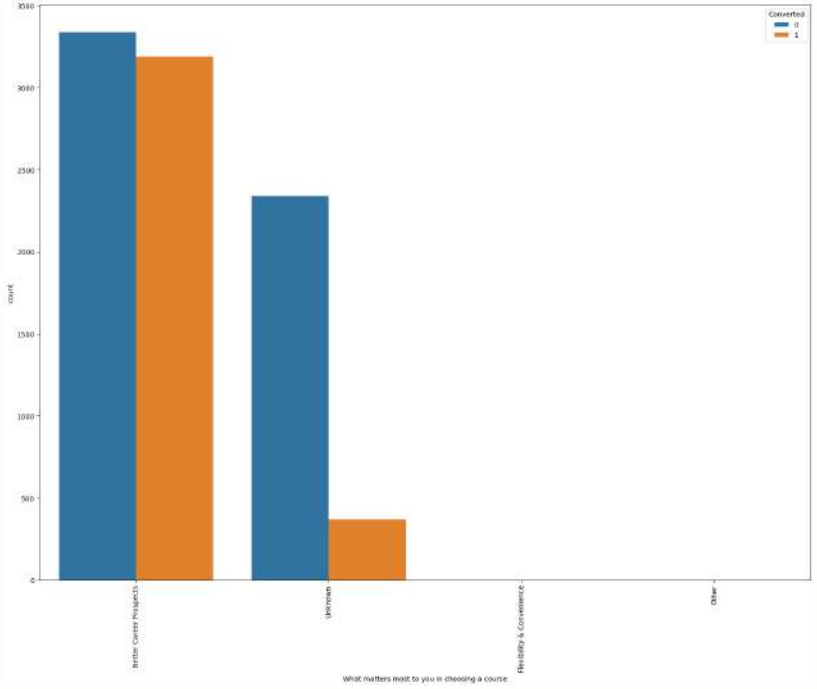
❖ 'API' and 'Landing Page Submission' generate the most leads but have less conversion rates of around 30%. Whereas, 'Lead Add Form' generates less leads but conversion rate is great. We should try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'. 'Lead Import' does not seem very significant.

❖ As one can expect, most of the responses are 'No' for the Do Not Email variables which generated most of the leads.

What is your current occupation:

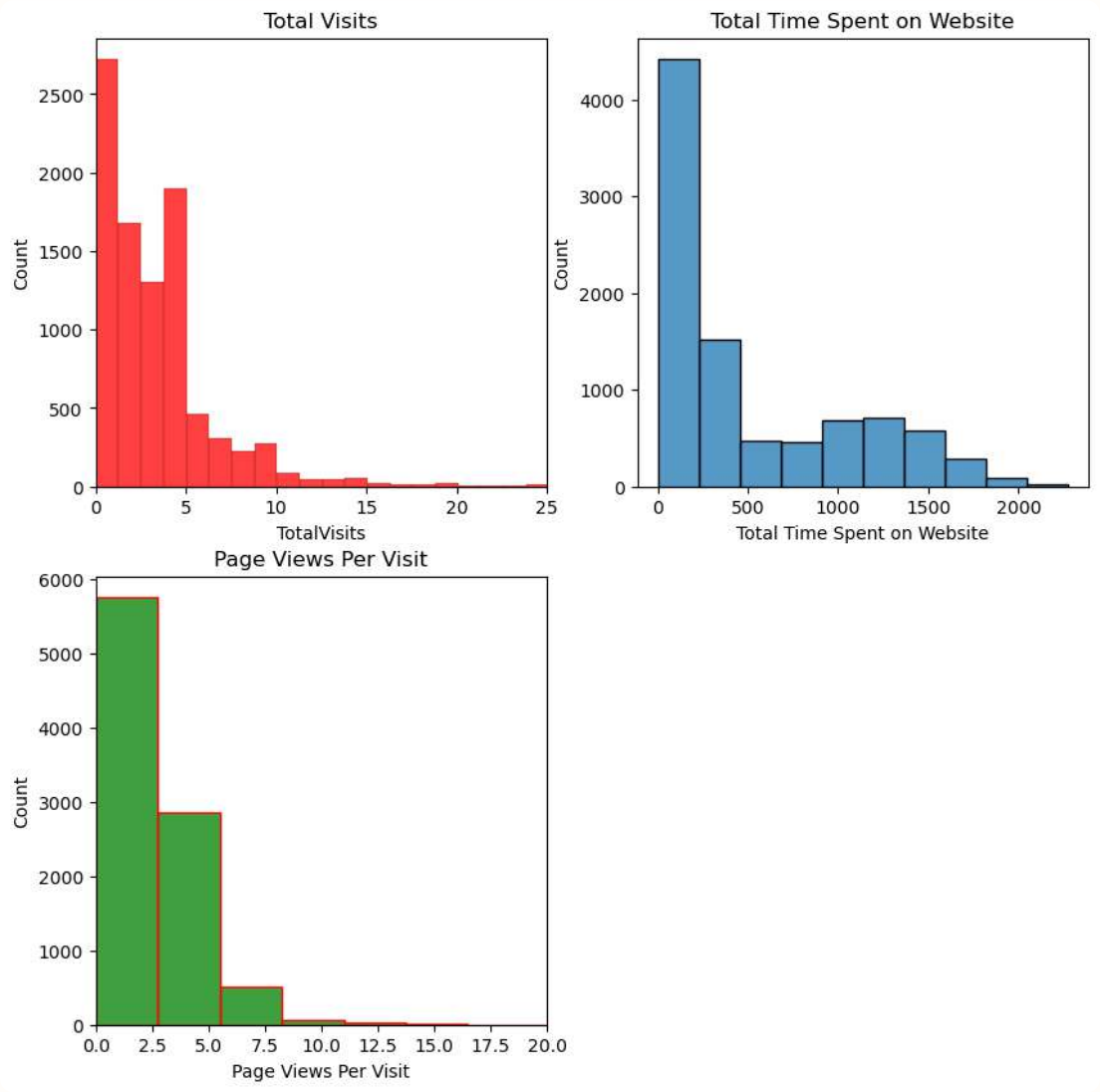


What matters most to you in choosing a course:



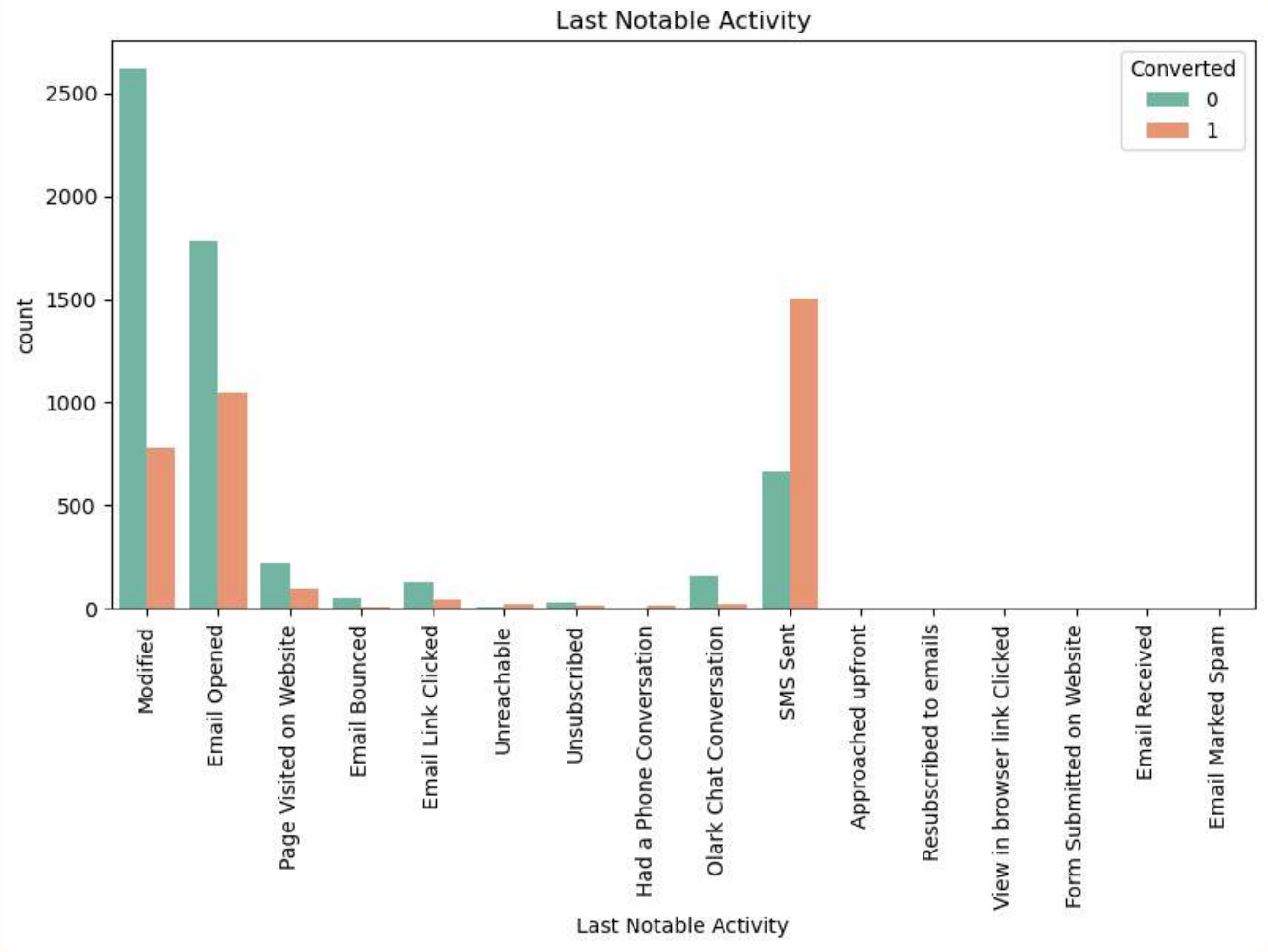
❖ Working professionals are most likely to get converted

Numerical Variables:



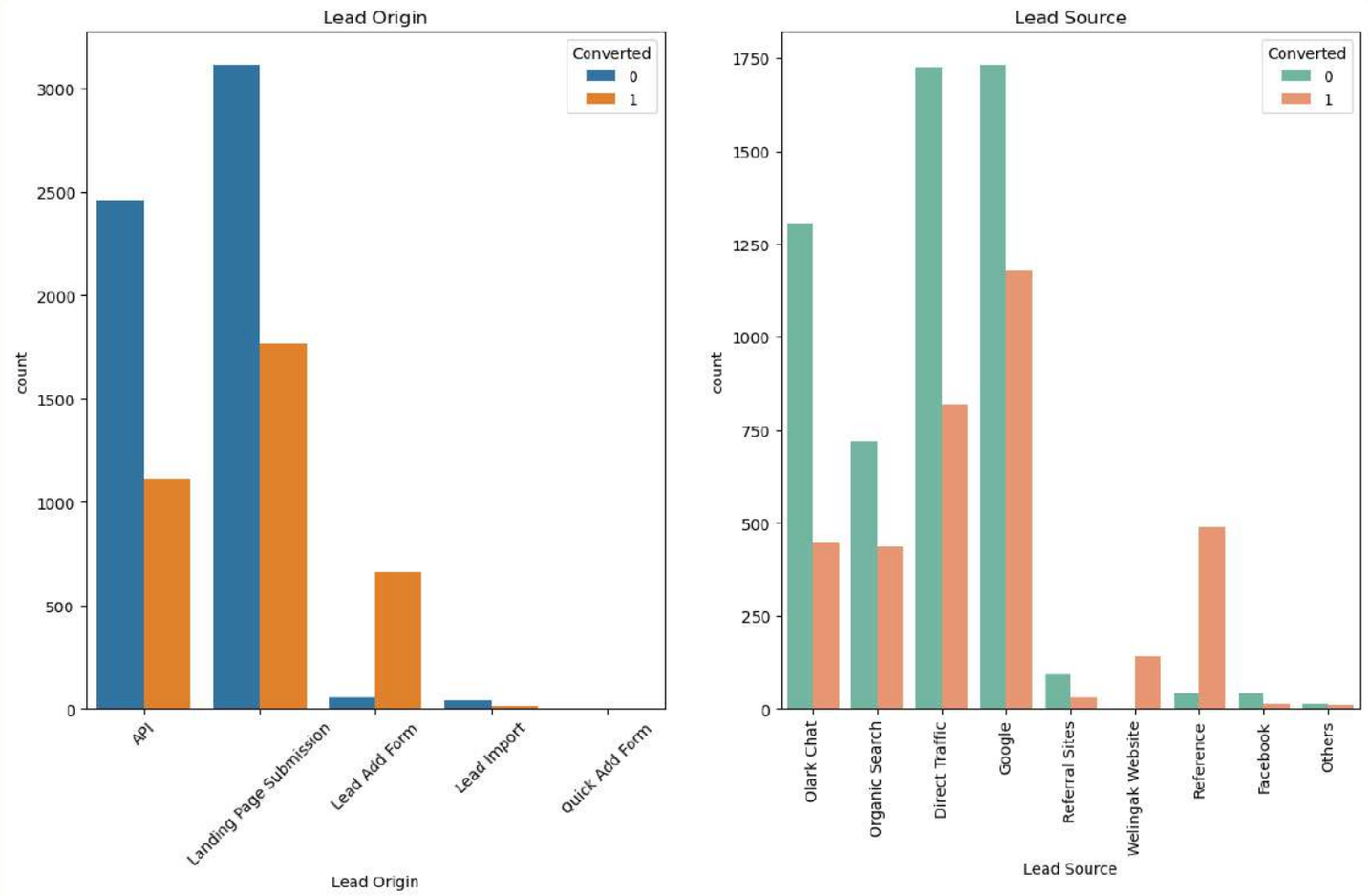
Bivariate analysis:

Last Notable Activity:



❖ SMS Sent have highest conversion count compared to other activities followed by Email Opened.

Lead Origin & Lead Source:



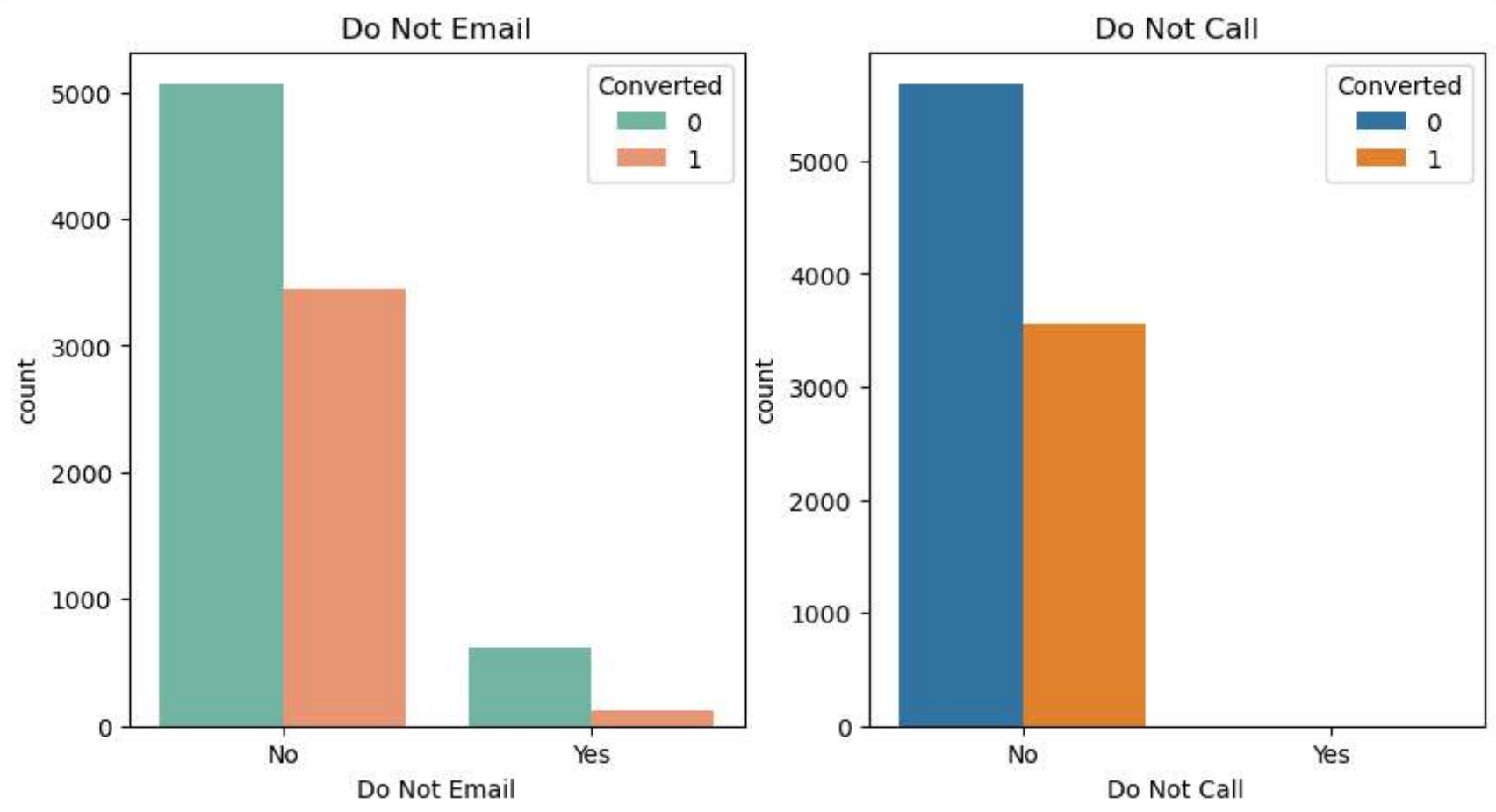
❖ Added form is more effective way to convert people but it is significantly less in count.

❖ Landing Page Submission has highest count of people who did'nt convert. Still it is second best effective way to convert people.

❖Reference helps most in converting people followed by Google.

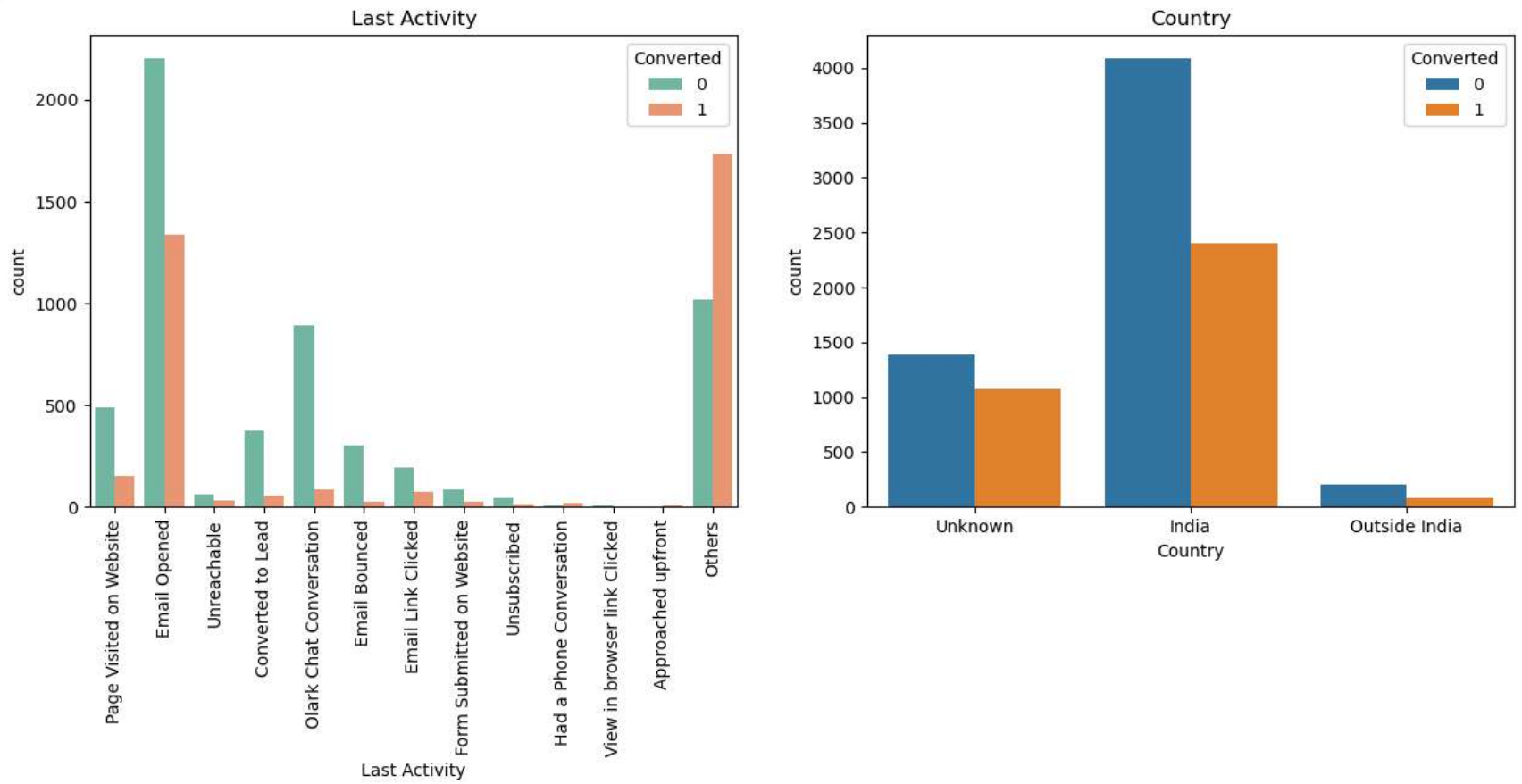
❖ Olark chat and referral sites perform lowest in conversion of people.

Do not Email & Do not Call:



❖ People who optioned out for no email and no call are having high chances of getting converted to join any course.

Last Activity & Country:

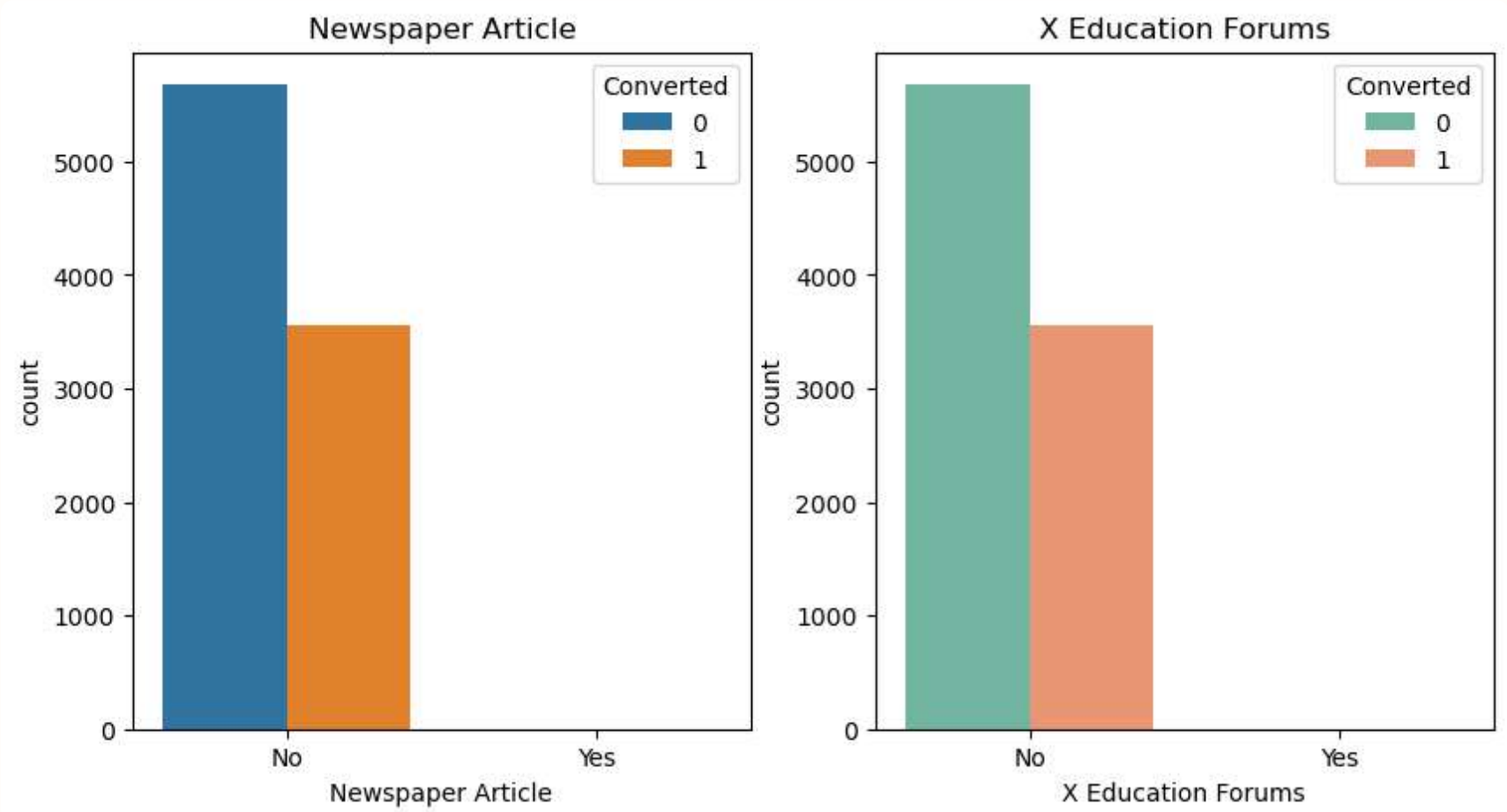


❖ SMS sending have very good response from people which reflects in the conversion count.

❖ Email opened activity has less but good response from people in conversion count.

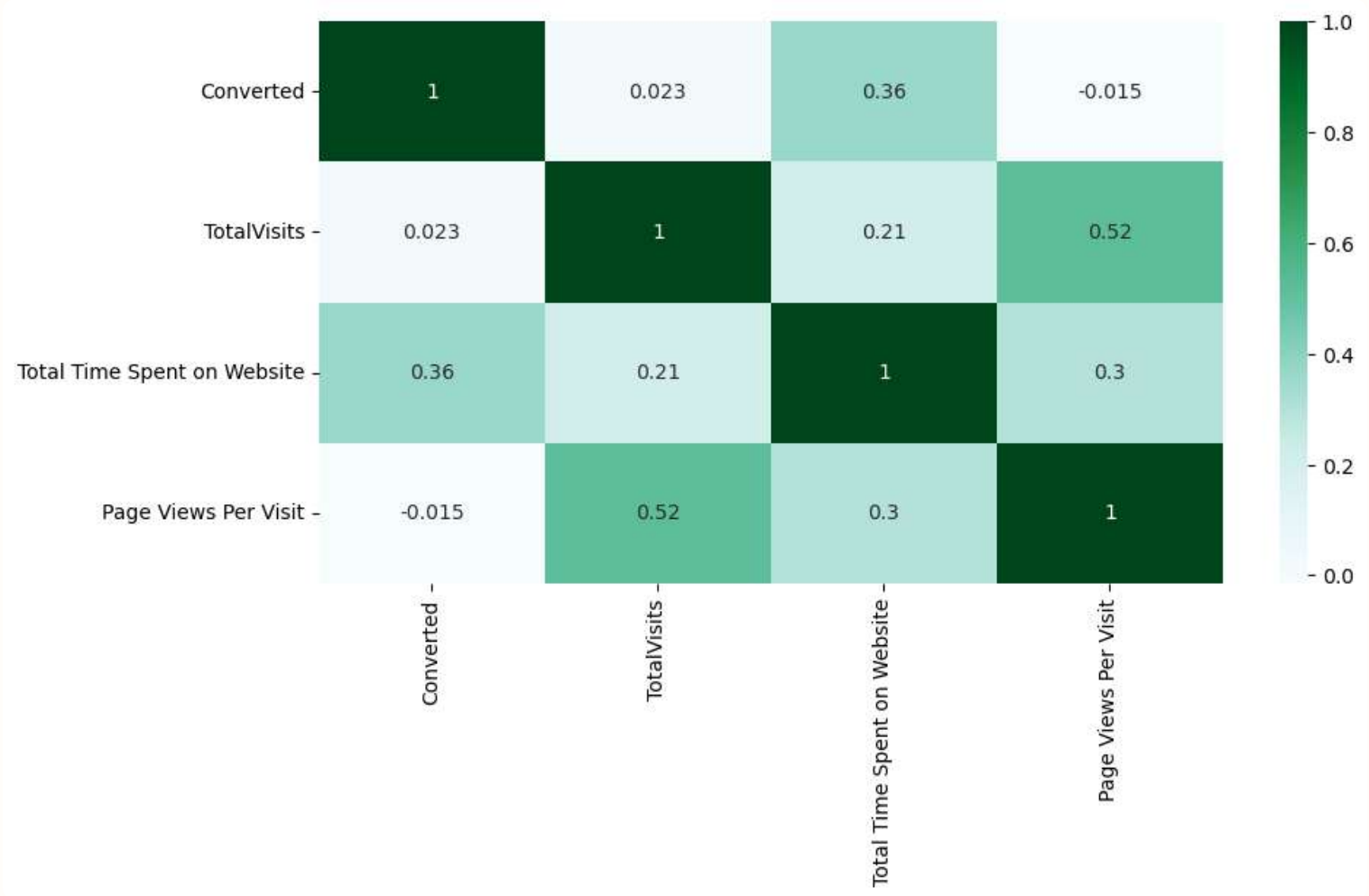
❖ Indian people are showing positive response in conversion count compared to out of India countries.

Newspaper Article & X Education Forums:



❖ People who haven't seen ads on Newspaper Articles and X Education Forum has good conversion rate but still lower than non conversion rate.

Multivariate Analysis:



❖ There is 0.36 correlation of "Total Time Spent on Website" with target variable "Converted".

❖ "Page Views Per Visit" have -0.015 correlation with target variable.

Train- Test split:

In [86]:
```
from sklearn.model_selection import train_test_split
```

In [87]:
```
X = dataset_final_dummy.drop(['Converted'], 1)
X.head()
```

Out[87]:

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Origin_Quick Add Form	Specialization_Business Administration	Specialization_E-Business	Specialization_E-COMMERCE	Specializati N
0	0.0	0	0.0	0	0	0	0	0	0	0	
1	5.0	674	2.5	0	0	0	0	0	0	0	
2	2.0	1532	2.0	1	0	0	0	1	0	0	
3	1.0	305	1.0	1	0	0	0	0	0	0	
4	2.0	1428	1.0	1	0	0	0	0	0	0	

In [88]:
```
# Put the target variable in y
y = dataset_final_dummy['Converted']
y.head()
```

Out[88]:
```
0    0
1    0
2    1
3    0
4    1
Name: Converted, dtype: int64
```

In [89]:
```
# Split the dataset as 70% | 30% for train and test.
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=10)
```

❖ We have split train data to 70% and test data is 30%.

MODEL EVALUTION:

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:**	-2652.3
Date:	Mon, 20 Nov 2023	**Deviance:**	5304.5
Time:	11:53:54	**Pearson chi2:**	6.85e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4032
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.5223	0.113	-31.298	0.000	-3.743	-3.302
TotalVisits	6.7155	2.087	3.218	0.001	2.625	10.806
Total Time Spent on Website	4.5640	0.165	27.641	0.000	4.240	4.888
Lead Origin_Lead Add Form	3.6930	0.200	18.490	0.000	3.302	4.084
Lead Source_Olark Chat	1.4518	0.110	13.149	0.000	1.235	1.668
Lead Source_Welingak Website	2.4117	1.029	2.344	0.019	0.395	4.428
Do Not Email_Yes	-1.5390	0.167	-9.208	0.000	-1.867	-1.211
Last Activity_Olark Chat Conversation	-1.1561	0.159	-7.276	0.000	-1.468	-0.845
Last Activity_Others	1.3601	0.074	18.266	0.000	1.214	1.506
What is your current occupation_Student	1.3727	0.220	6.242	0.000	0.942	1.804
What is your current occupation_Unemployed	1.2454	0.087	14.371	0.000	1.076	1.415
What is your current occupation_Working Professional	3.7412	0.195	19.222	0.000	3.360	4.123
Last Notable Activity_Unreachable	2.6400	0.688	3.838	0.000	1.292	3.988

❖ We can see that p-value is less than 0.05 and vif is less than 0.5 for all variables

MODEL BUILDING:

❖ Splitting the Data into Training and Testing Sets.

❖ The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.

❖ Use RFE for Feature Selection.

❖ Running RFE with 15 variables as output.

❖ Building Model by removing the variable whose p-value is greater than 0.05 & vif value is greater than 5

❖ Predictions on test data set.

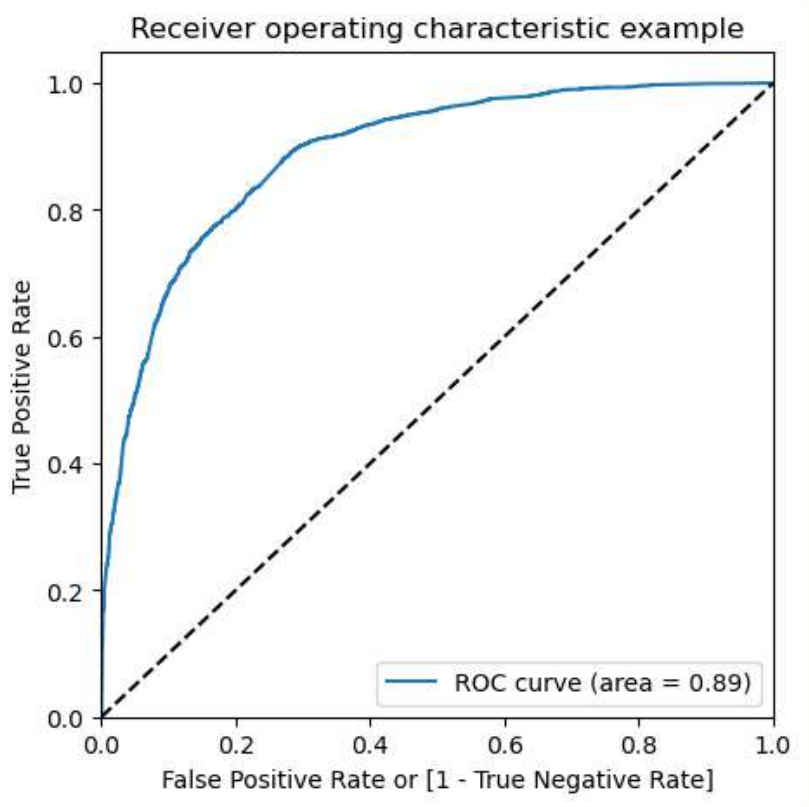
❖ Overall accuracy 81%.

ROC CURVE:

The ROC curve stands for **Receiver Operating Characteristic curve**. ROC curves display the performance of a classification model.

ROC tells us how good the model is for distinguishing between the given classes, in terms of the predicted probability.

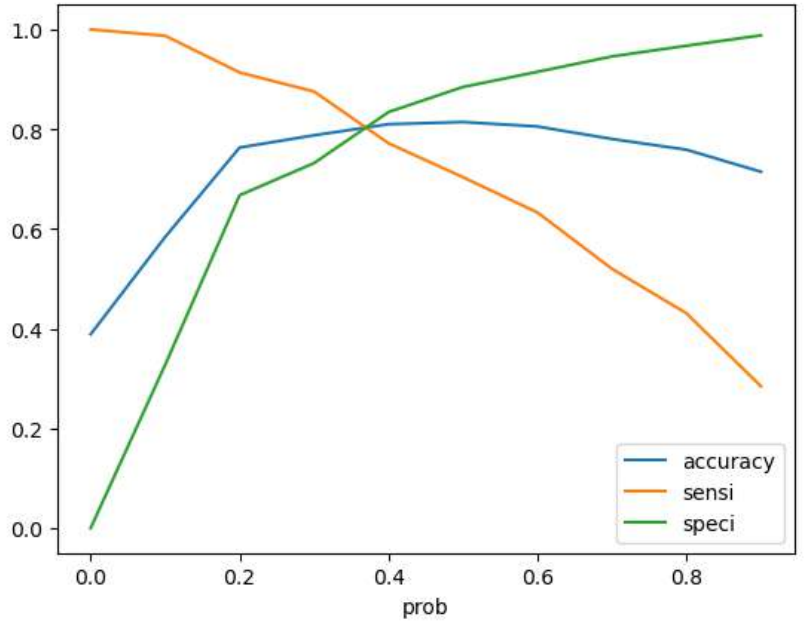
❖ The ROC Curve should be a value close to 1. We are getting a good value of 0.89 indicating a good predictive model.



Finding Optimal Cutoff Point:

Above we had chosen an arbitrary cut-off value of 0.5. We need to determine the best cut-off value.

❖ From the curve , 0.4 is the optimum point to take it as a cutoff probability.



METRIC FOR TRAIN SET:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

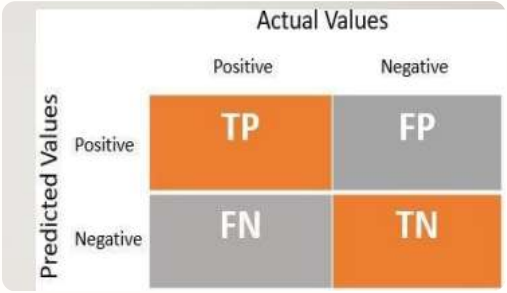
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Accuracy = 80.9%

Precision = 83.43%

Recall = 85.17%



$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

Sensitivity = 77.17%

Specificity = 74.75%

Confusion Matrix of Train Set:

3299	655
574	1940

METRIC FOR TEST SET:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

Accuracy = 81.06%

Precision = 74.26%

Recall = 76.31%

Sensitivity = 85.37%

Specificity = 83.94%

Confusion Matrix of Test Set:

1448	277
248	799

RECOMMENDATIONS:

❖ By referring to the data visualizations, focus on

- ■ Increasing the conversion rates for the categories generating more leads and
- ■ Generating more leads for categories having high conversion rates.

❖ Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.

❖ Based on varying business needs, modify the probability threshold value for identifying potential leads.

❖ Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

❖ We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and welingak website.

❖ Leads who spent more time on website, more likely to convert.

❖ Most common last activity is email opened. highest rate = SMS Sent. Max are unemployed. Max conversion with working professional.