LEAD SCORING CASE STUDY

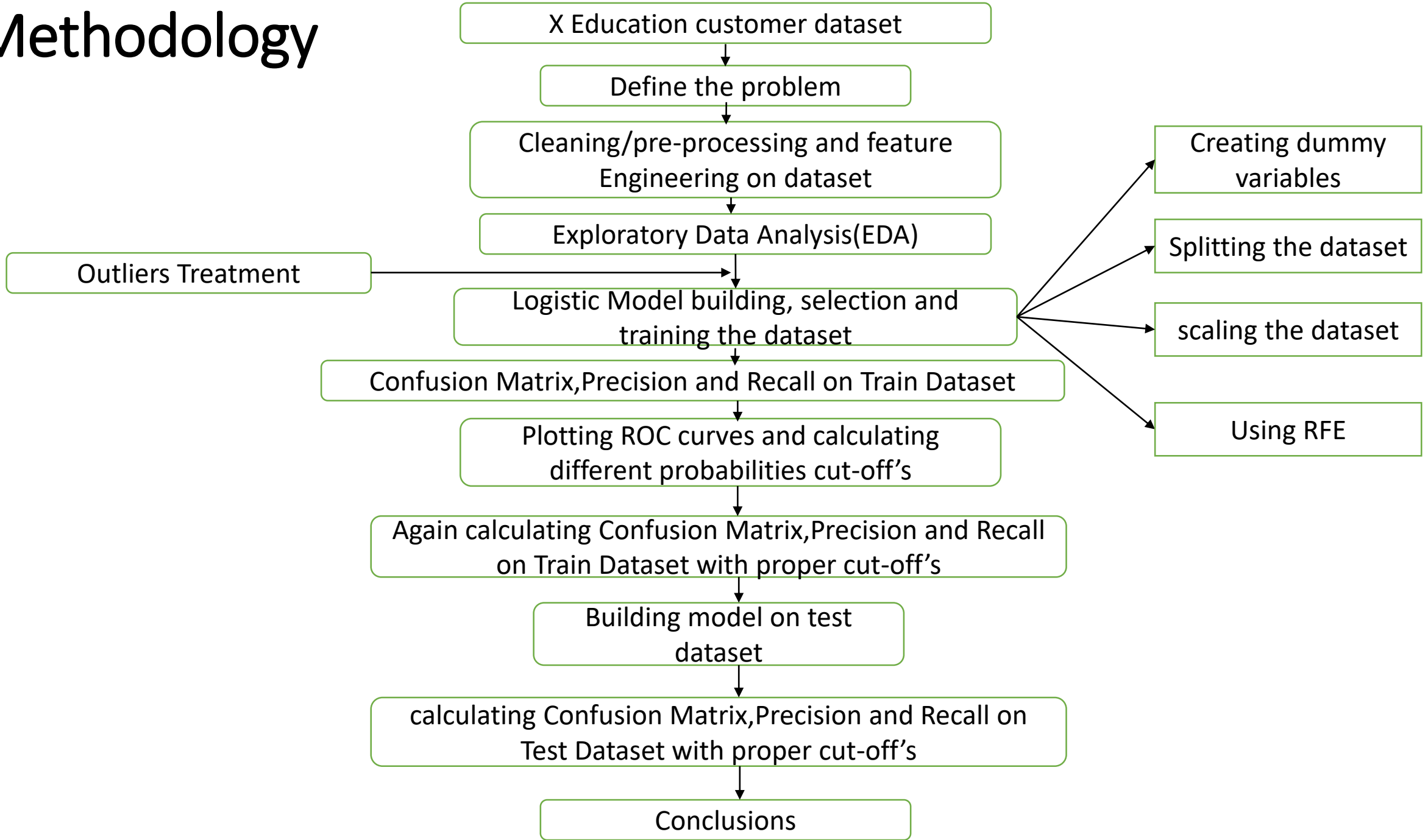
Lead Scoring Case Study By:
1.Shreyas S. Kashte
2.Vinayak Dharmkare
3.Pawan Kohale

Problem Statement

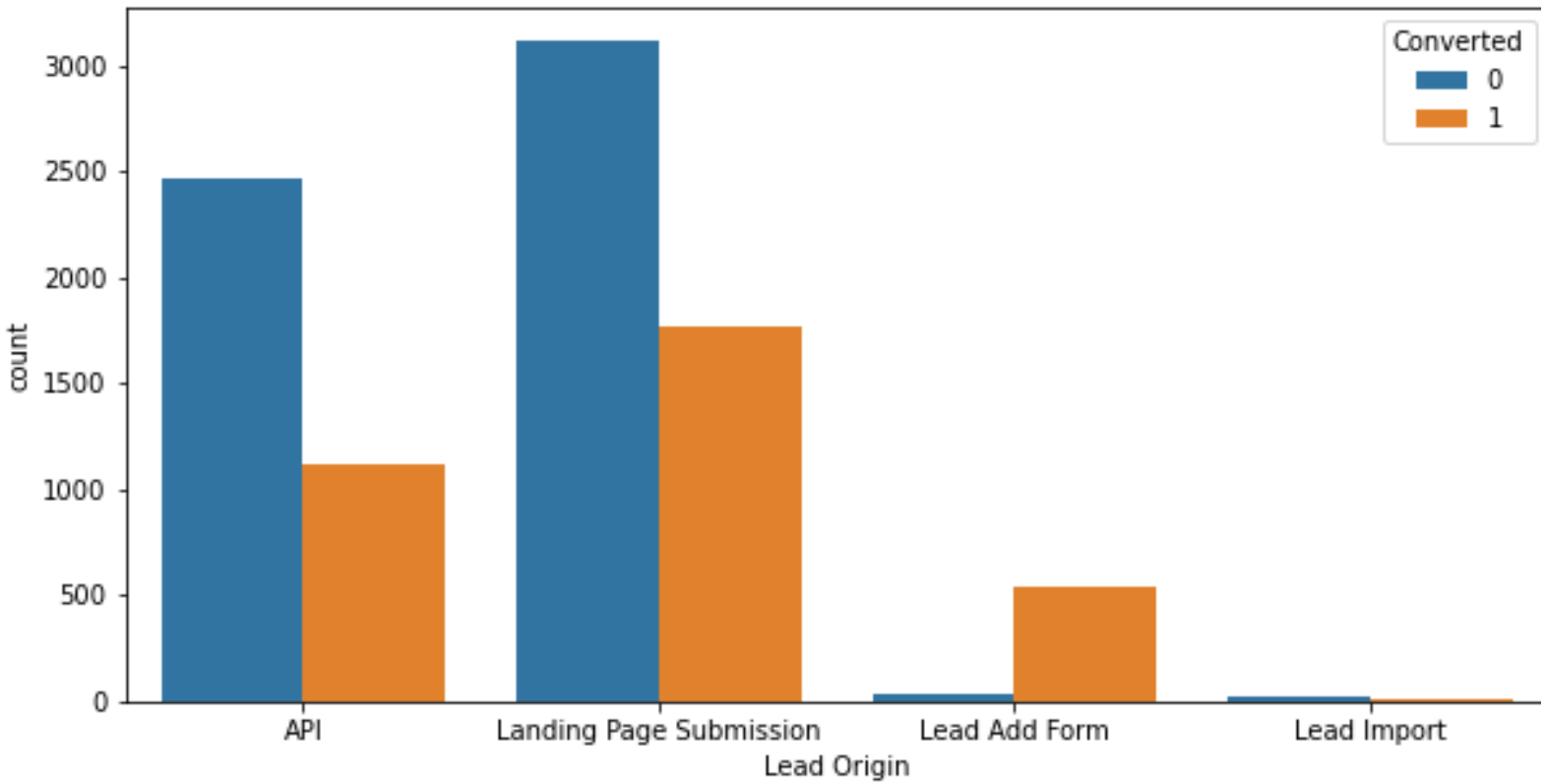
- An education company named X Education sells online courses to industry professionals.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- The typical lead conversion rate at X education is around 30%.
- X Education want to target lead conversion rate to be around 80%.

So , our objective is to provide model which can promise the most leads to X education company .

Methodology

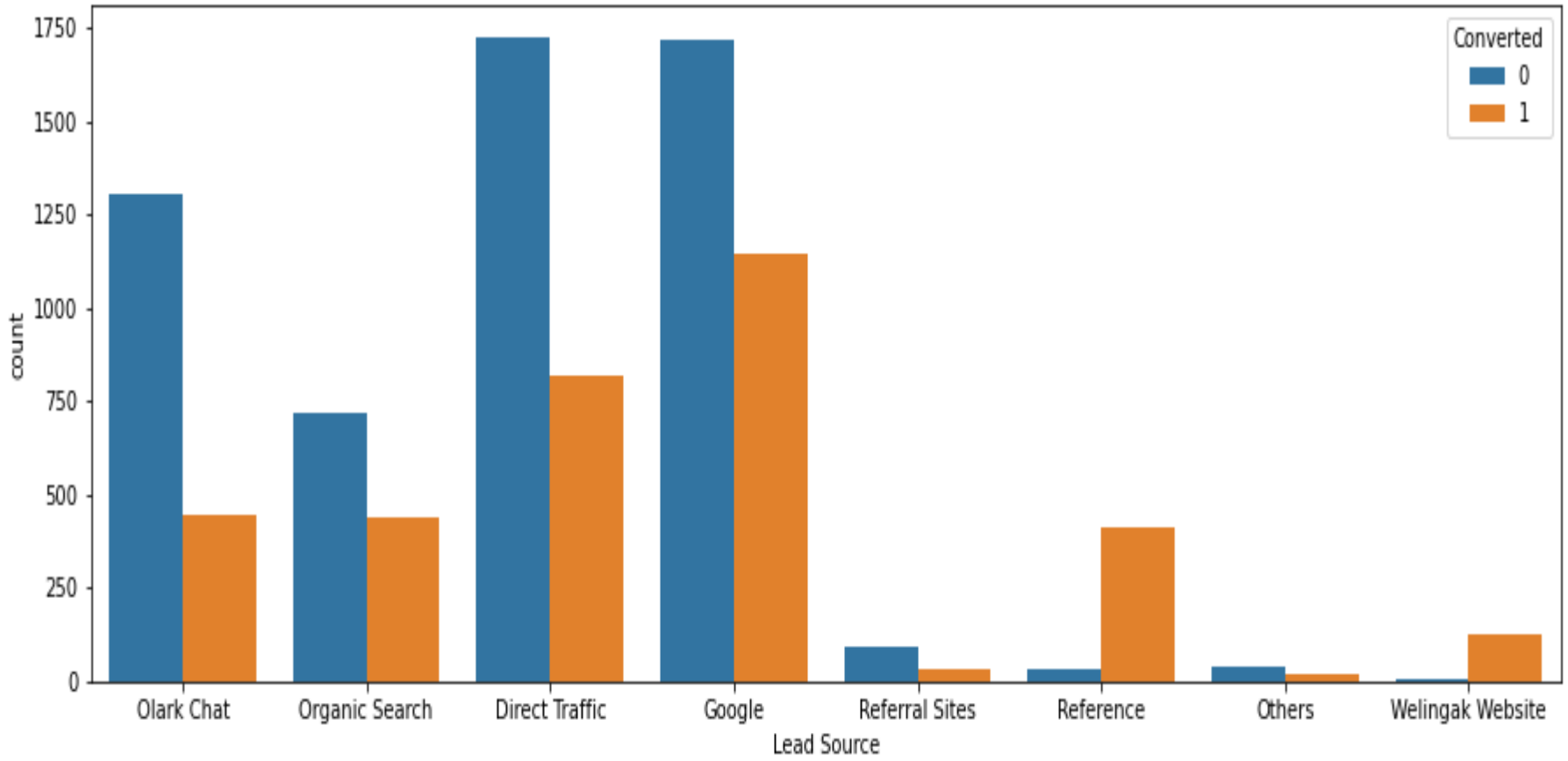


Lead Origin



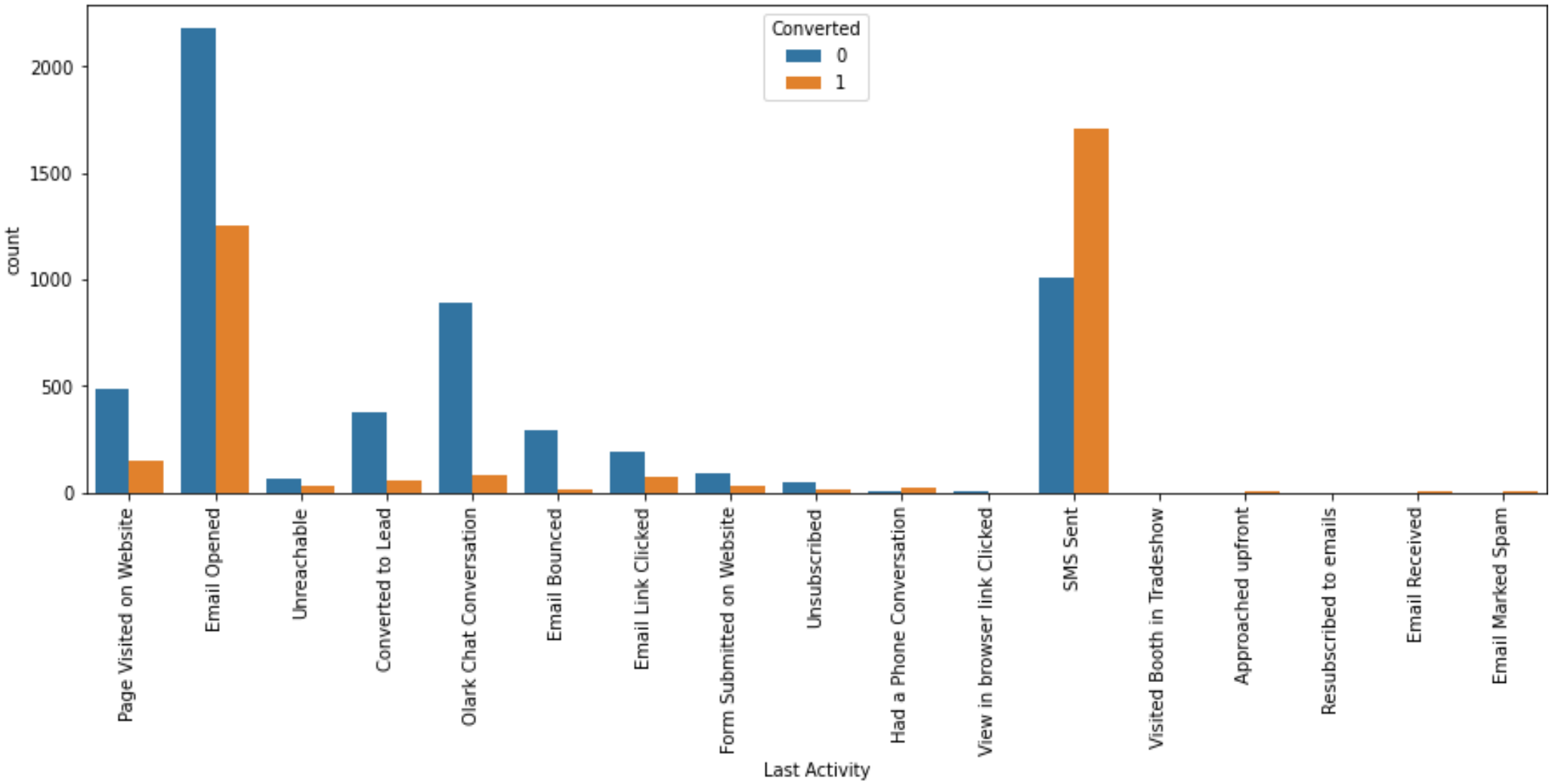
In Lead Origin we can see ‘Lead Add Form’ has mostly converted customer comparing with others.

Lead Source



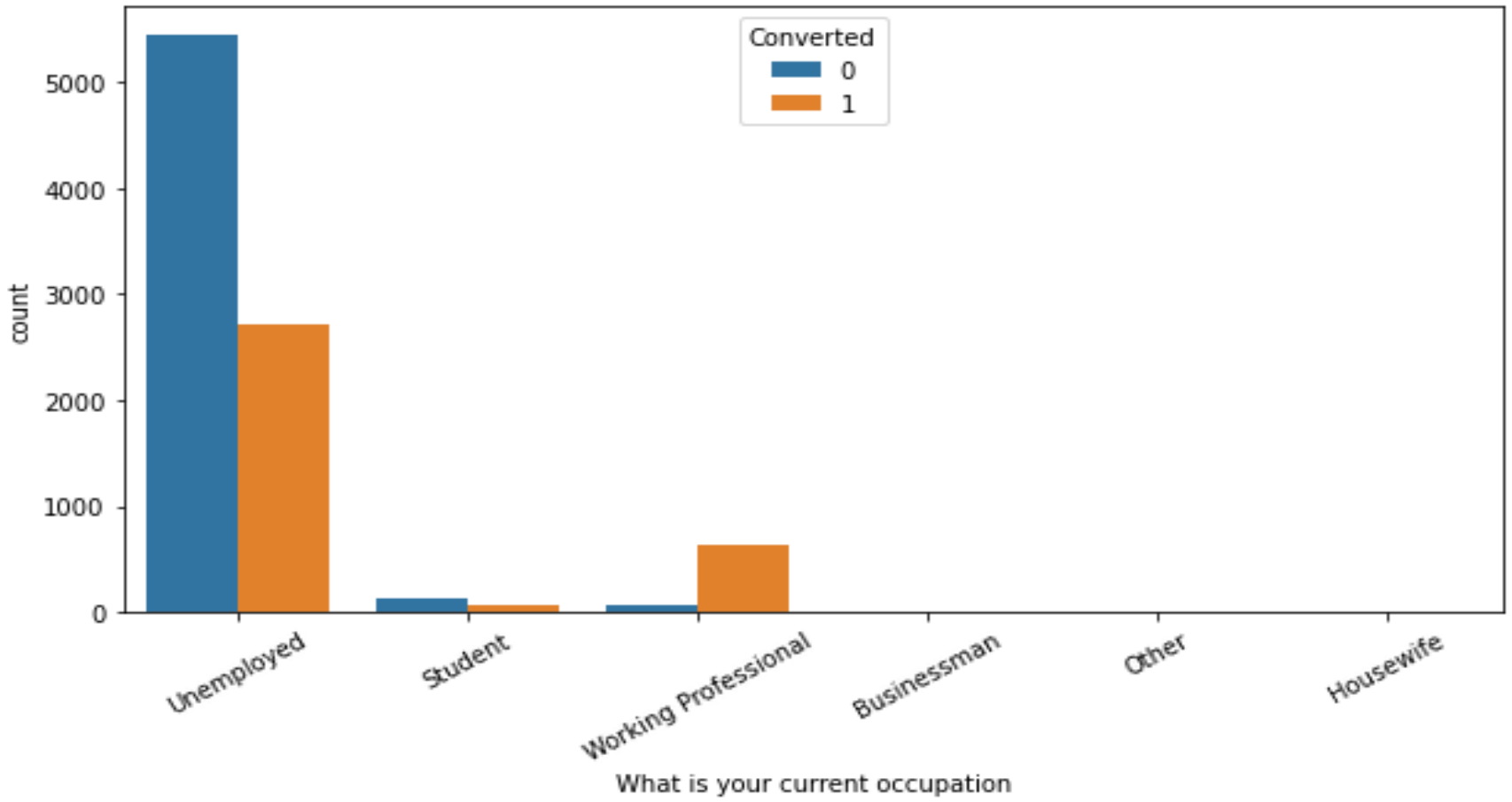
In ‘Lead Source’ we have ‘Refrence’ and ‘Welingak Website’ has high customer conversion rate.

Last Activity



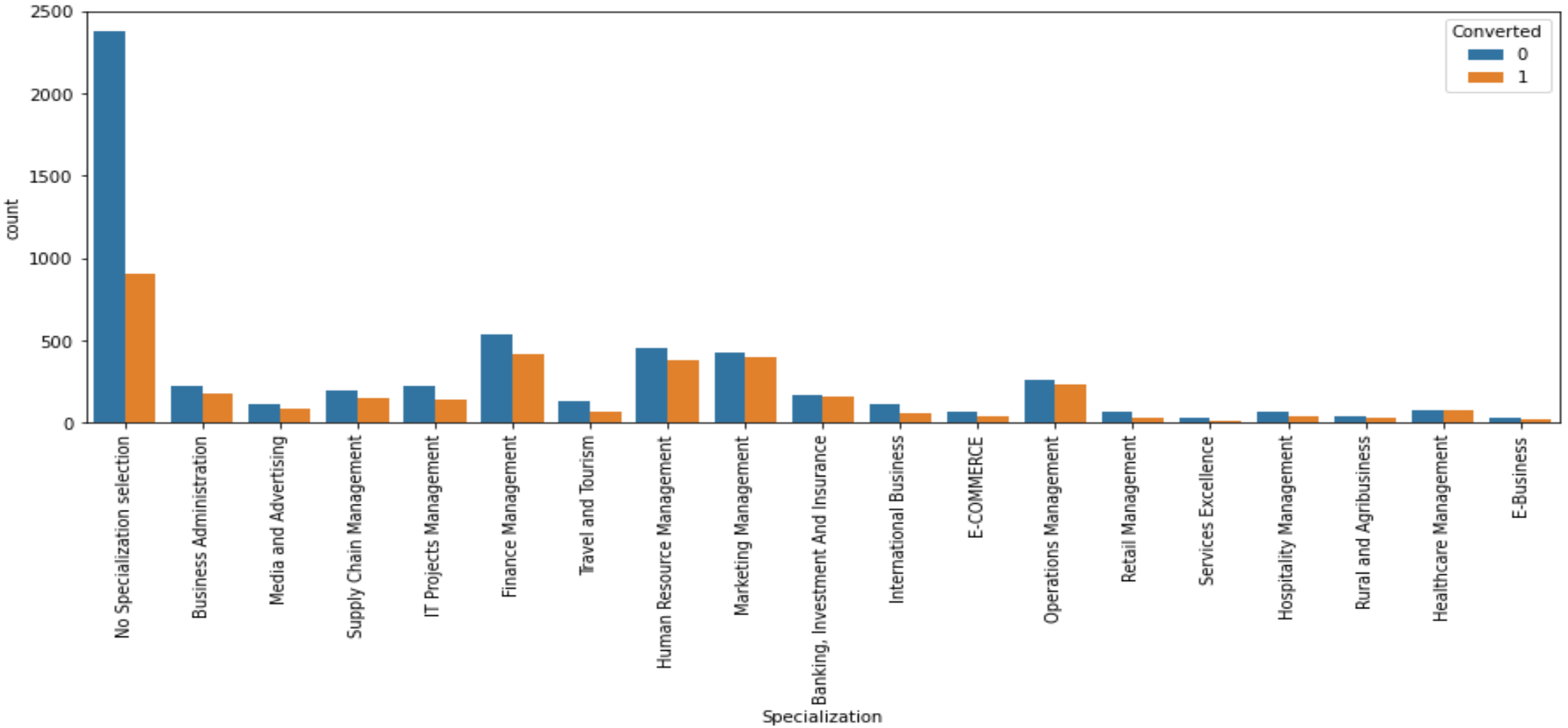
In ‘Last Activity’ we have ‘SMS sent’ and ‘Phonic Conversion’ has high customer conversion rate in course.

Occupation



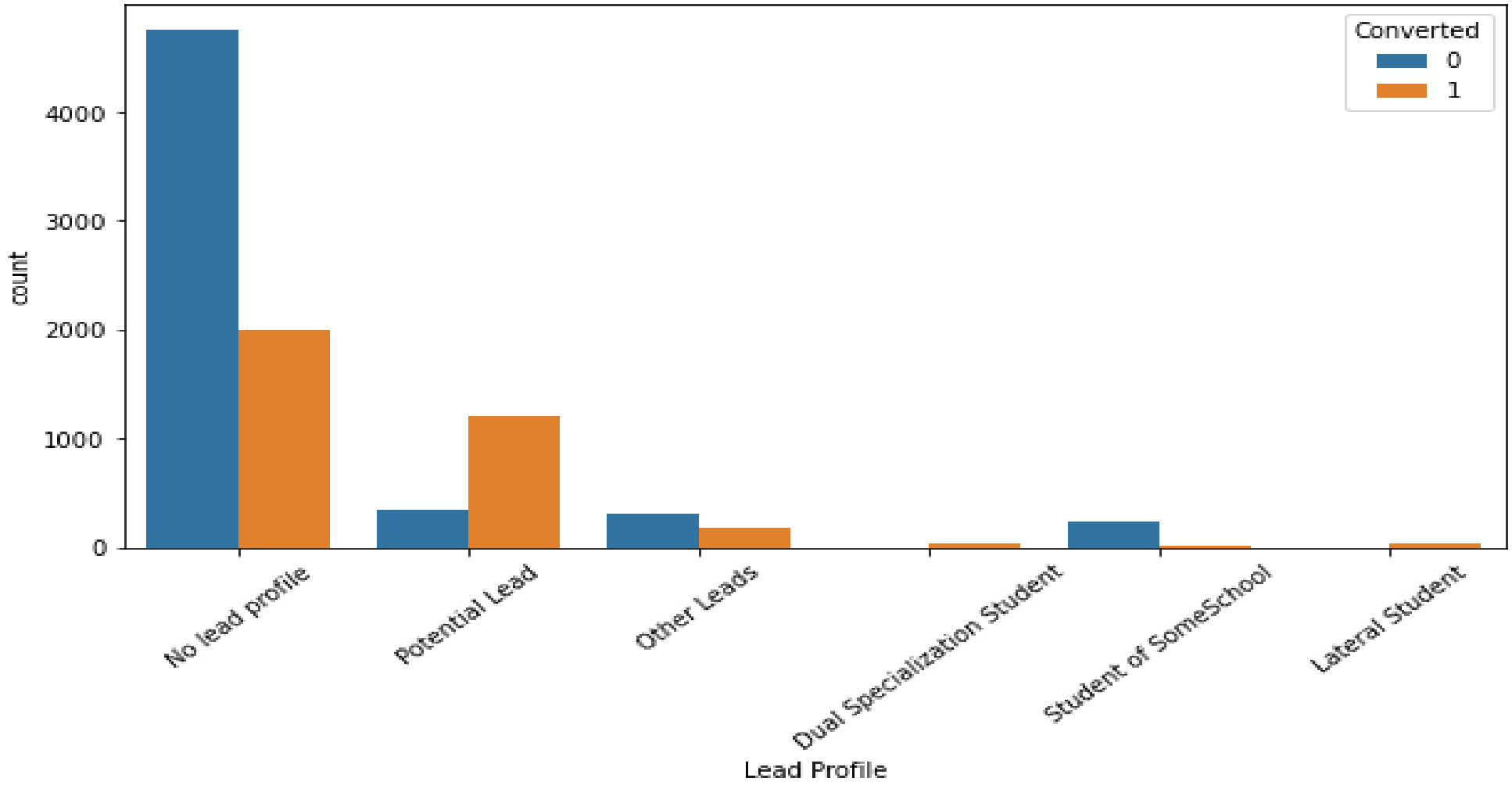
In ‘Occupation’ we have ‘Working professional’ has high customer conversion rate i course.

Specialization



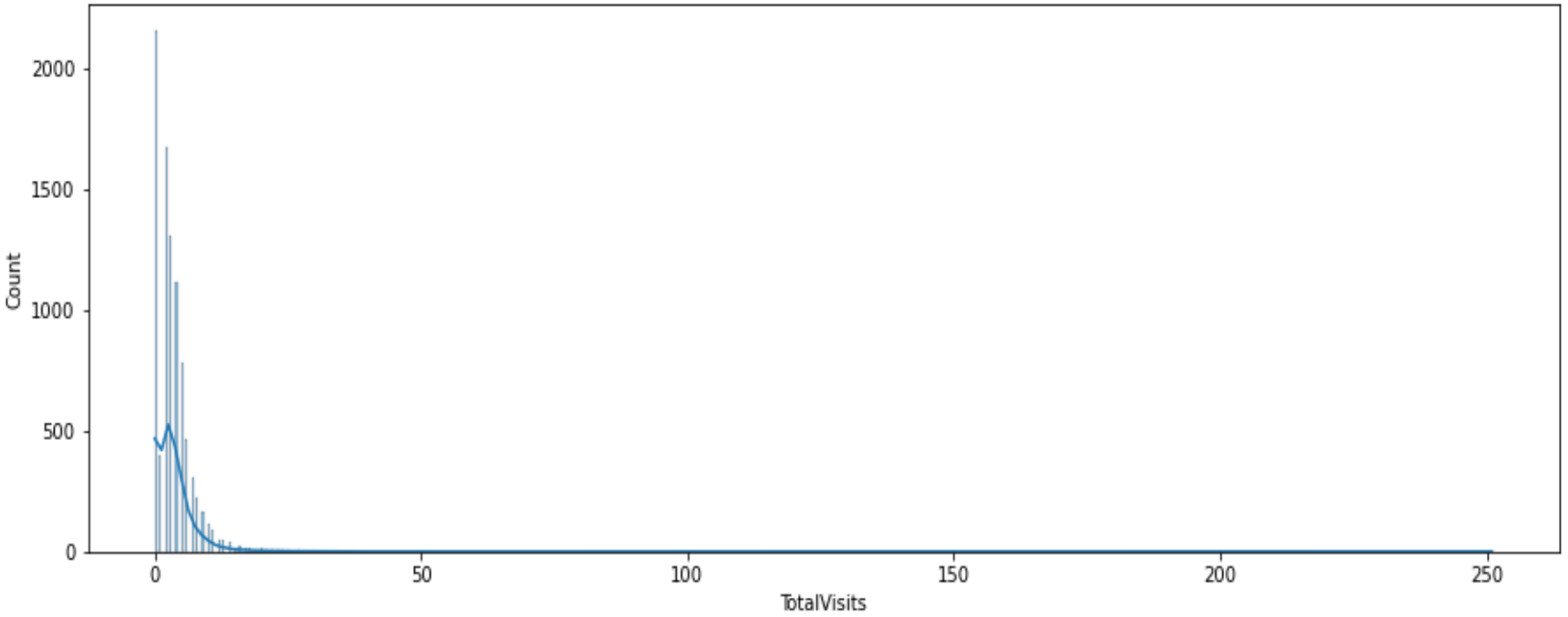
In ‘specialization’ the customer from ‘Administraion’ background has better conversion rate into course comparing with others

Lead Profile

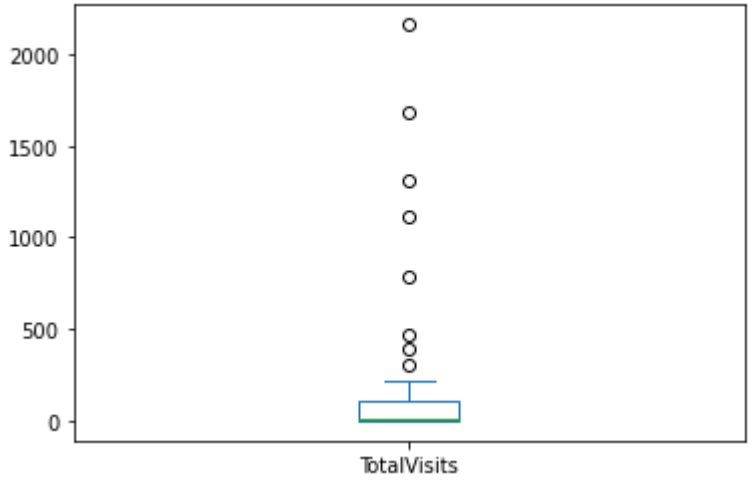


In ‘Lead Profile’ the ‘Potential Lead’ has better conversion rate of customer.

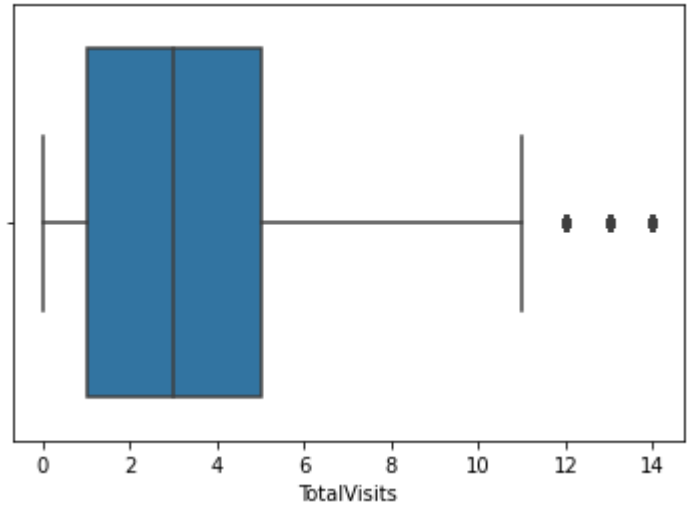
Treating Outliers in ‘TotalVisits’



Normal Distribution of TotalVists

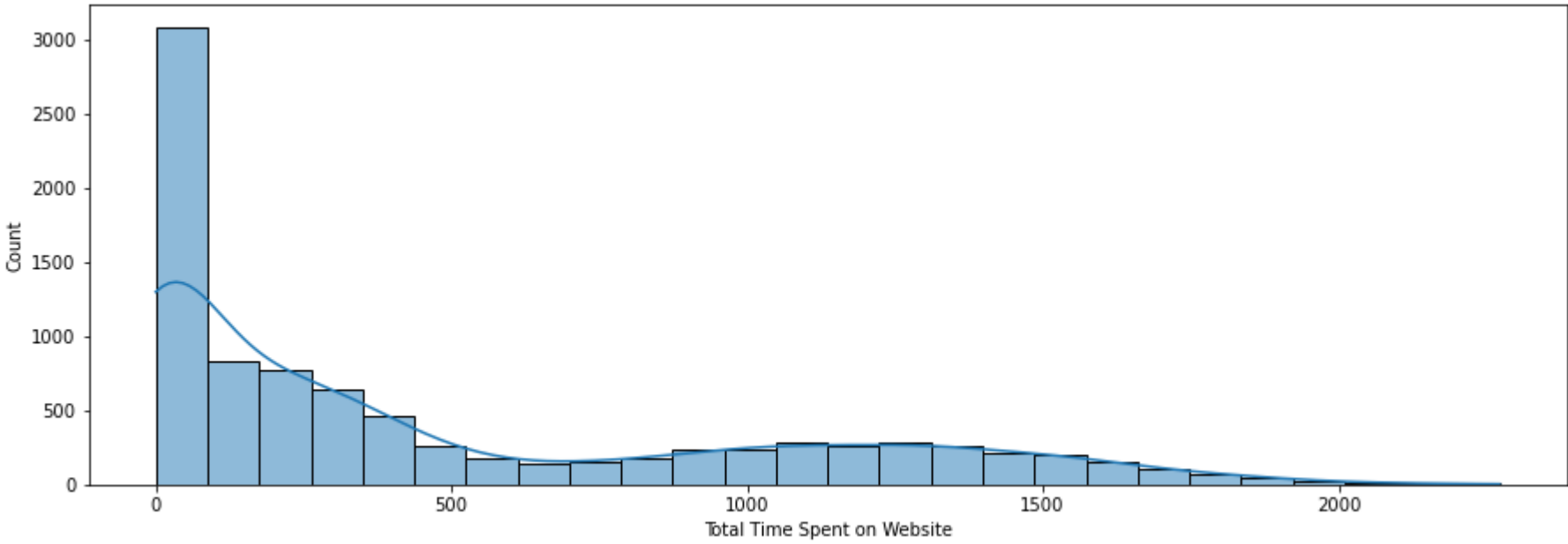


With outliers

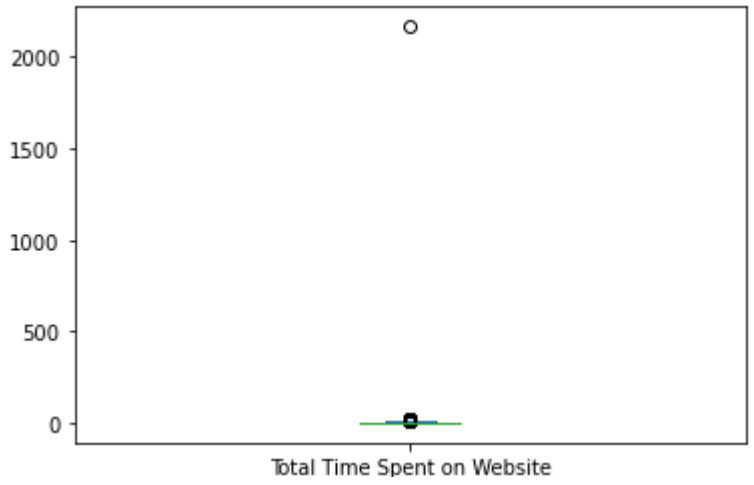


Without Outliers

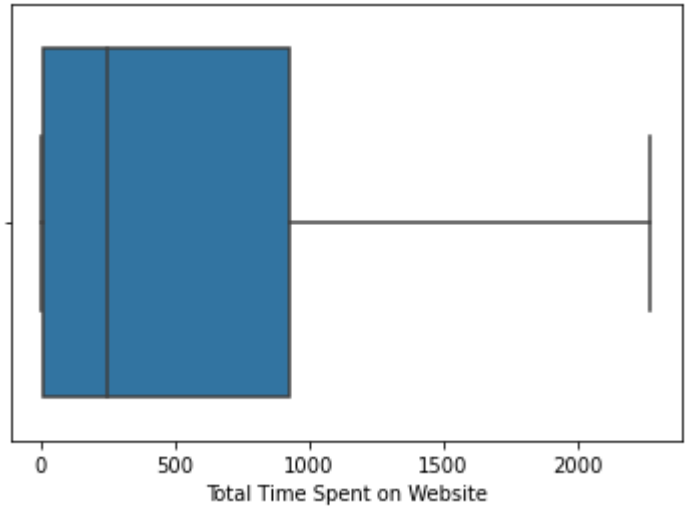
Treating Outliers in ‘Total time spent on website’



Normal distribution of Total Time Spent on Website

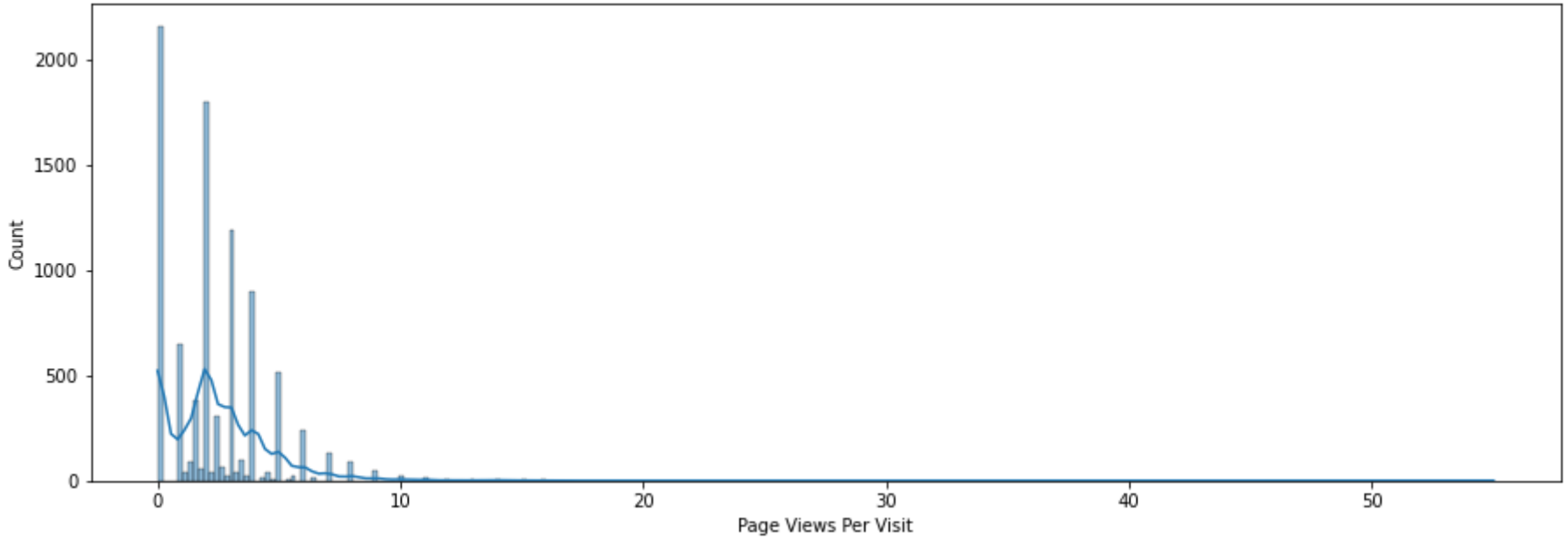


With Outliers

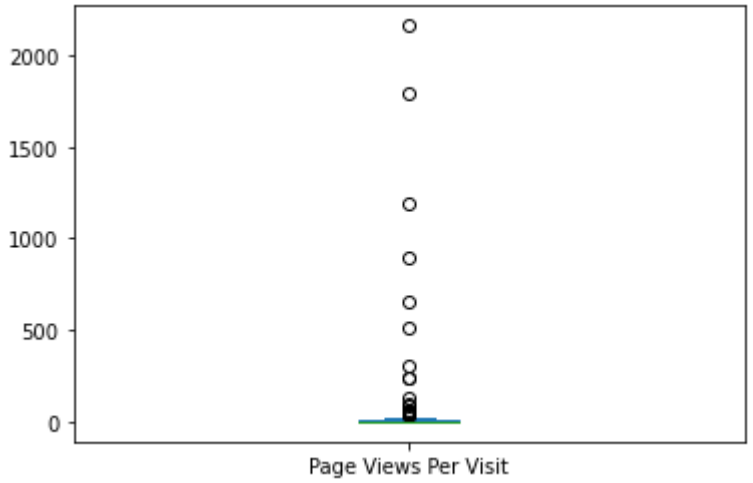


Without Outliers

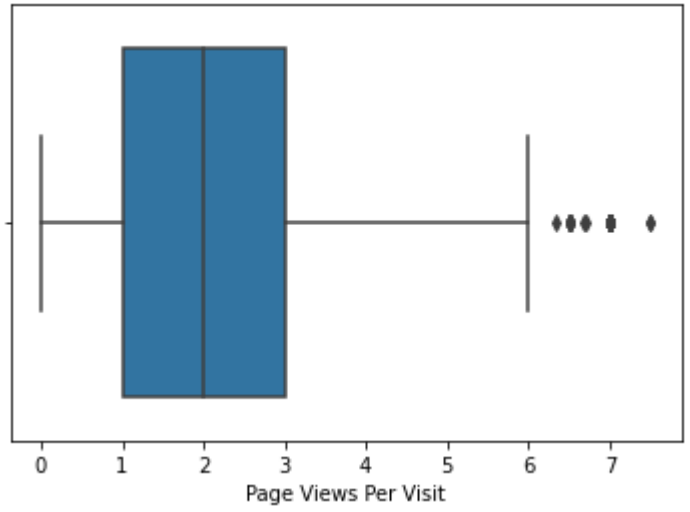
Treating Outliers in ‘Page Views per Visit’



Normal Distribution of Page Views per Visit



With Outliers



Without Outliers

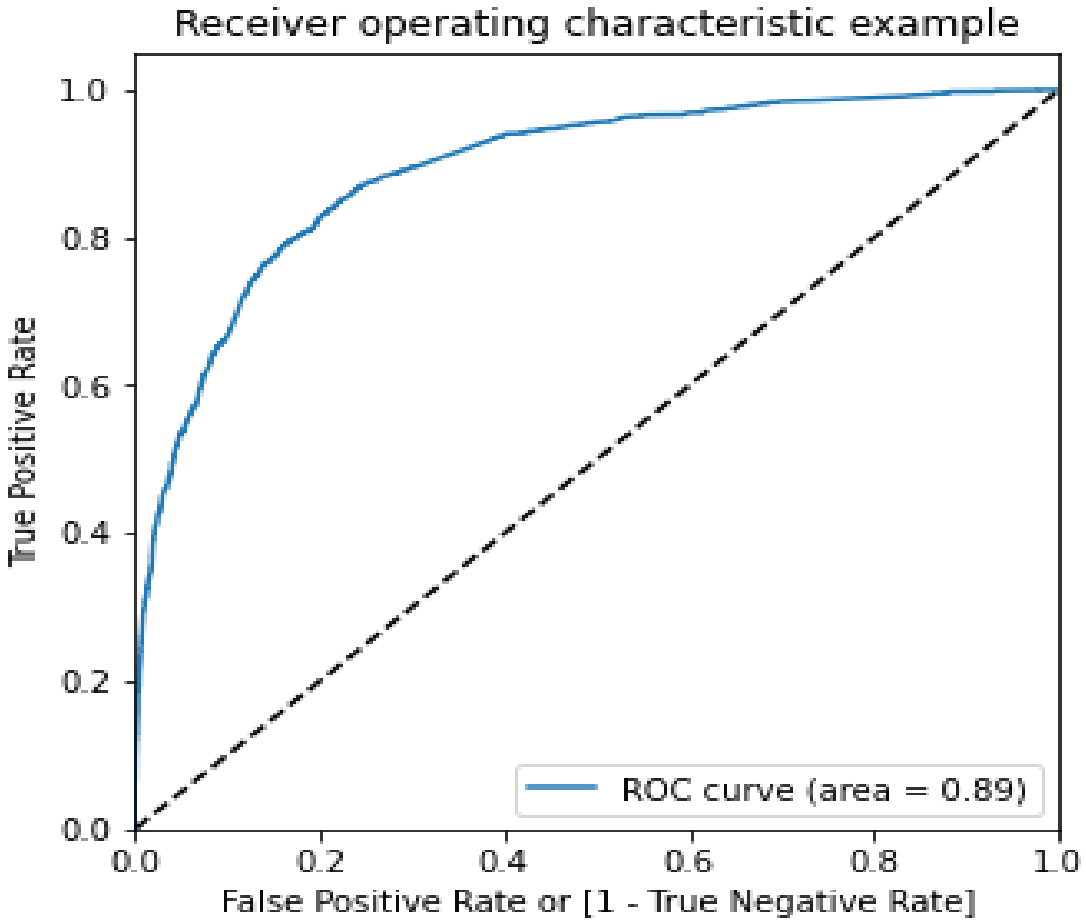
Model Selection and Evaluation

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6205
Model:	GLM	Df Residuals:	6191
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2491.1
Date:	Mon, 14 Nov 2022	Deviance:	4982.1
Time:	10:29:21	Pearson chi2:	6.42e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4109
Covariance Type:	nonrobust		

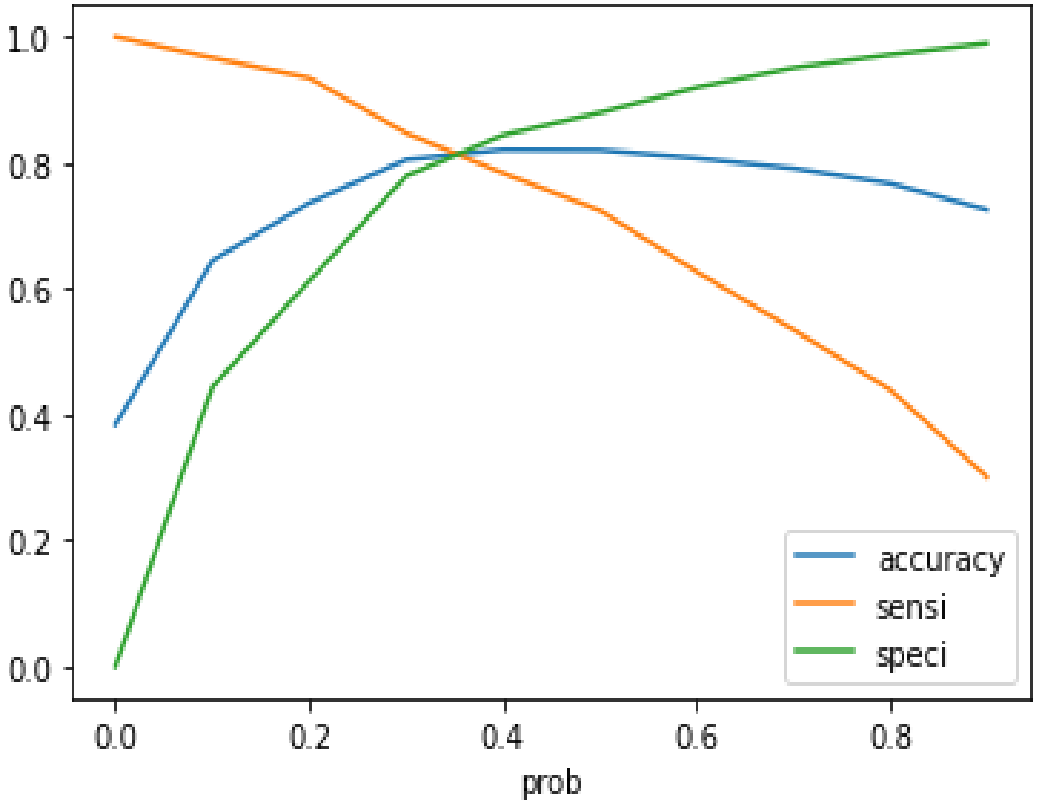
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6633	0.061	-27.276	0.000	-1.783	-1.544
Total Time Spent on Website	1.0916	0.041	26.330	0.000	1.010	1.173
Lead Origin_Lead Add Form	3.1847	0.226	14.118	0.000	2.743	3.627
Lead Origin_Lead Import	1.1809	0.550	2.147	0.032	0.103	2.259
Lead Source_Olark Chat	1.3203	0.107	12.342	0.000	1.111	1.530
Lead Source_Welingak Website	2.7448	0.756	3.629	0.000	1.263	4.227
Do Not Email_1	-1.5516	0.180	-8.638	0.000	-1.904	-1.200
Last Activity_Converted to Lead	-0.9454	0.212	-4.455	0.000	-1.361	-0.529
Last Activity_Had a Phone Conversation	1.8029	0.677	2.662	0.008	0.475	3.130
Last Activity_Olark Chat Conversation	-1.4766	0.170	-8.692	0.000	-1.810	-1.144
Last Activity_SMS Sent	1.3493	0.077	17.420	0.000	1.198	1.501
What is your current occupation_Working Professional	2.6399	0.200	13.194	0.000	2.248	3.032
Lead Profile_Potential Lead	1.7266	0.100	17.339	0.000	1.531	1.922
Lead Profile_Student of SomeSchool	-1.7922	0.444	-4.040	0.000	-2.662	-0.923

ROC Curve



- The ROC curve **shows the trade-off between sensitivity and specificity** . Classifiers that give curves closer to the top-left corner indicate a better performance.
- We can see our graph is closer to top-left corner, it means our model has perform better.

Plot Accuracy, Sensitivity and Specificity for various probabilities cut-off's



- By plotting this graph we can get the optimum cut-off for our dataset.
- We get the cut-off between 0.3 and 0.4.

Inferences on Train and Test data set

- **The Score on train and test set were:**
  on training set:
    1. accuracy: 81.32%
    2. sensitivity: 80.78%
    3. specificity: 81.65%
    4. precision: 73.30%
    5. recall: 80.78%
    6. F1 score: 0.76

  on test set:
    1. accuracy: 82.25%
    2. sensitivity: 82.18%
    3. specificity: 82.29%
    4. precision: 72.24%
    5. recall: 82.18%
    6. F1 score: 0.77

Conclusion on Model:

Top Positive Correlation variables for customer conversion :

1. Lead Origin_Lead Add Form
2. Lead Source_Welingak Website
3. Last Activity_SMS Sent

Top Negative Correlation variables for customer conversion :

1. Do not email_Yes
2. Lead Profile_Student of SomeSchool
3. Last Activity_Olark Chat Conversation