Leads Scoring
Case Study

Bandla Sunitha
Pramoda Maratha

Abstract:

An education company named X Education sells online courses to industry professionals. Many professionals who are interested in the courses land on their website and browse for courses, When company markets its course on several websites and search engines like Google, once people lands on websites, they might browse the course or fill up a form for the course or watch some videos. By collecting the information the sales team starts converting the people into leads, typically conversion rate is around 30%.

Objectives of Business:

- **Goal of the company: Target lead conversion rate to be around 80%.**

Problem solving methodology:

- Step 1: Importing Data Set and Understanding data as per the Business point of view
- Step 2: Data Cleaning, missing value treatment and outliers treatment
- Step 3: Exploratory Data Analysis (EDA), univariate analysis
- Step 4: Applying logistic regression and build a model
- Step 5: Using RFE choose the metrics
- Step 6: Model evaluation
- Step 7: Check cutoff point, VIF’s

Data Cleansing:

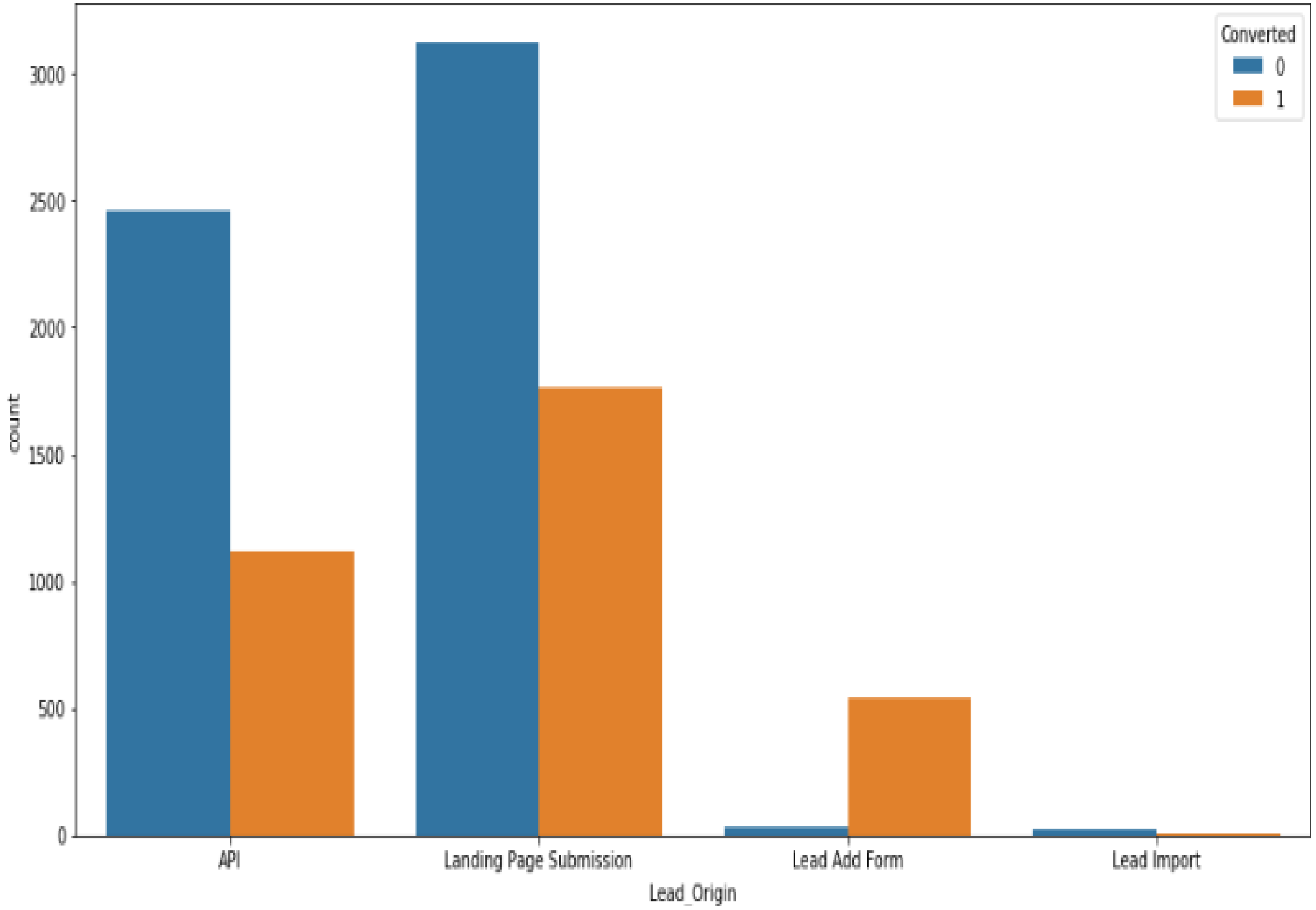
- Select : Many columns has ‘Select’ as value, replace the ‘Select’ as nan because lead has left it without filling the column from list given so it got as default value.
- Dropped the columns where 70% null’s exists in dataset and handled other less % null column independently.
- Lead Quality: In this metric 4473 values are exists out of 9240, so many 'nan's are exists, here we can impute 'Not Sure' for nan, because keeping an empty value may give the meaning of 'Not Sure‘
- City: Mumbai holding high count in the dataset, so impute 'Mumbai' in blanks

Data Cleansing:

- Tags: In this metric, imputed 'Will revert after reading the email' in blanks
- Specialization : Here he/she should not reveal it or his/her option is not available in the list, or may not have any specialization or is a student or even a unemployed.
- current_occupation: 86% entries are of Unemployed so we can impute "Unemployed" in 'current occupation' metric

EDA: Univariate Analysis

Lead Origin Metric:



Inference:

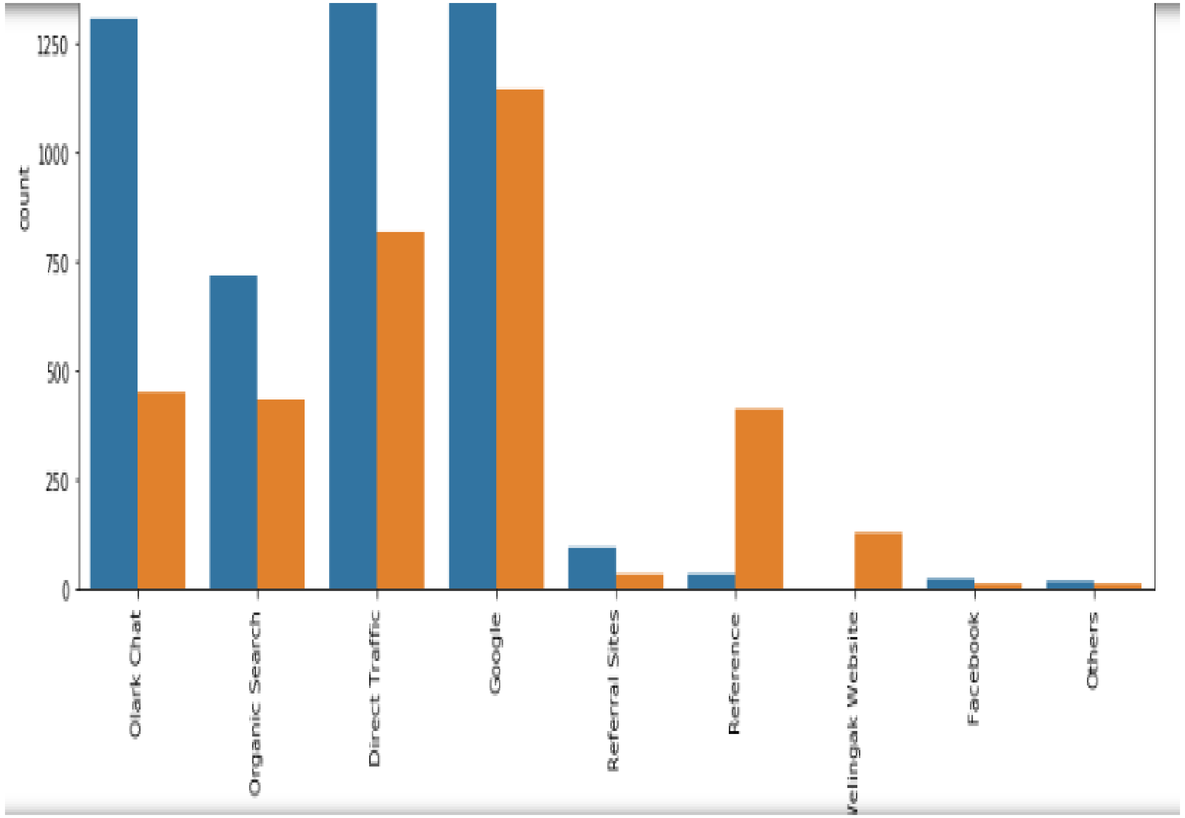
1. API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
2. Lead Add Form has more than 90% conversion rate but count of lead are not very high.
3. Lead Import are very less in count.

EDA: Univariate Analysis

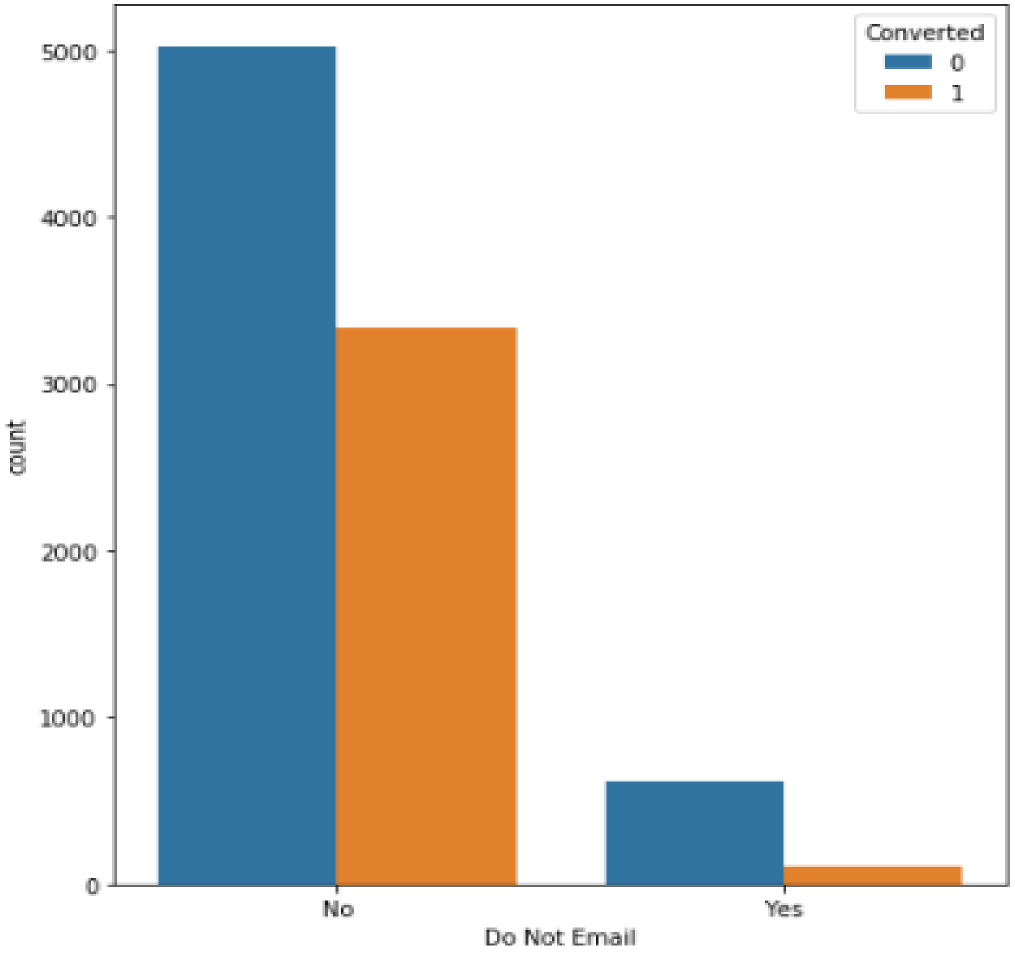
Inference:

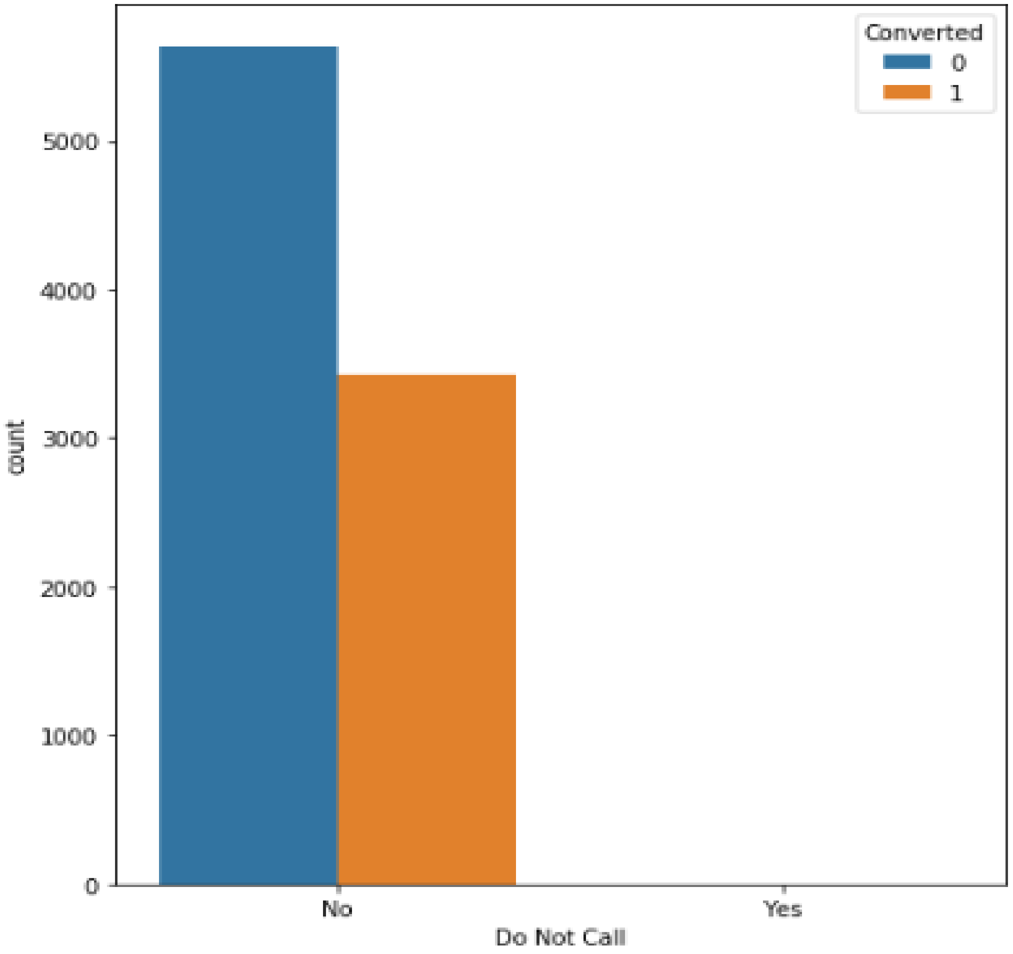
1.Google, Direct traffic and olark chat generates maximum number of leads and from their conversion rate ranging between 30-65% and overall approching counts also looks good.
2.Conversion Rate of reference leads and leads through welingak website is high where approching counts looks very less
To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

Lead Source Metric

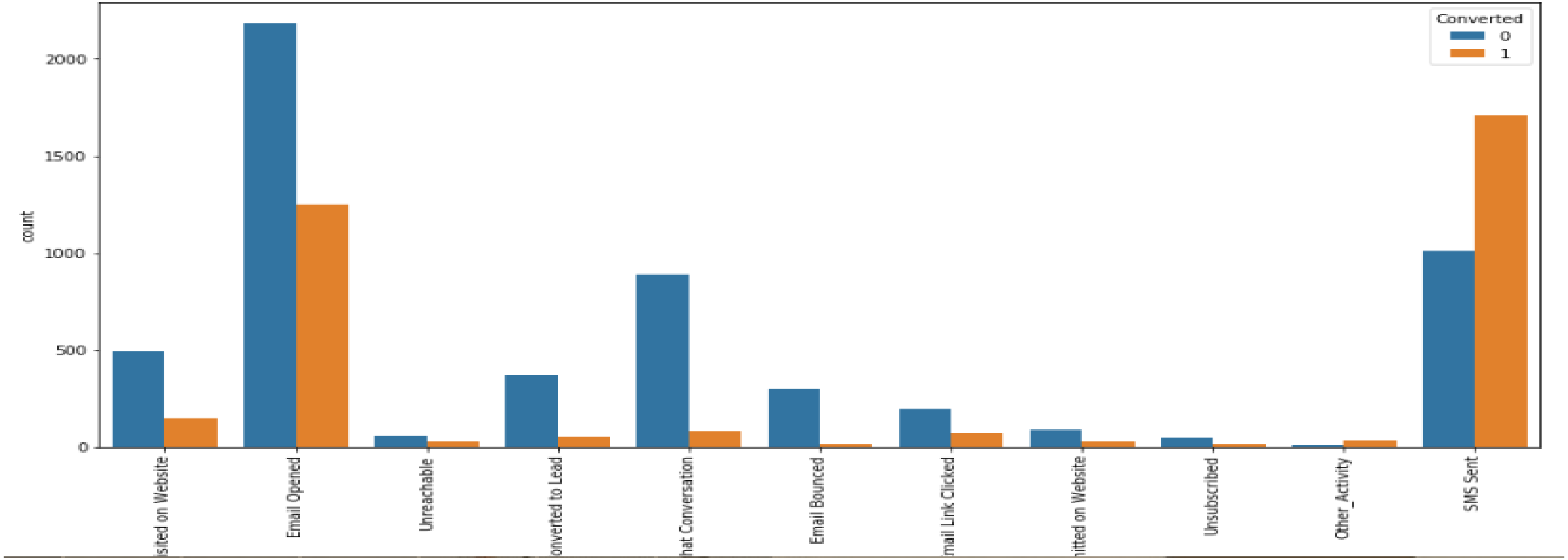


EDA: Do Not Email and Do Not Call Metrics





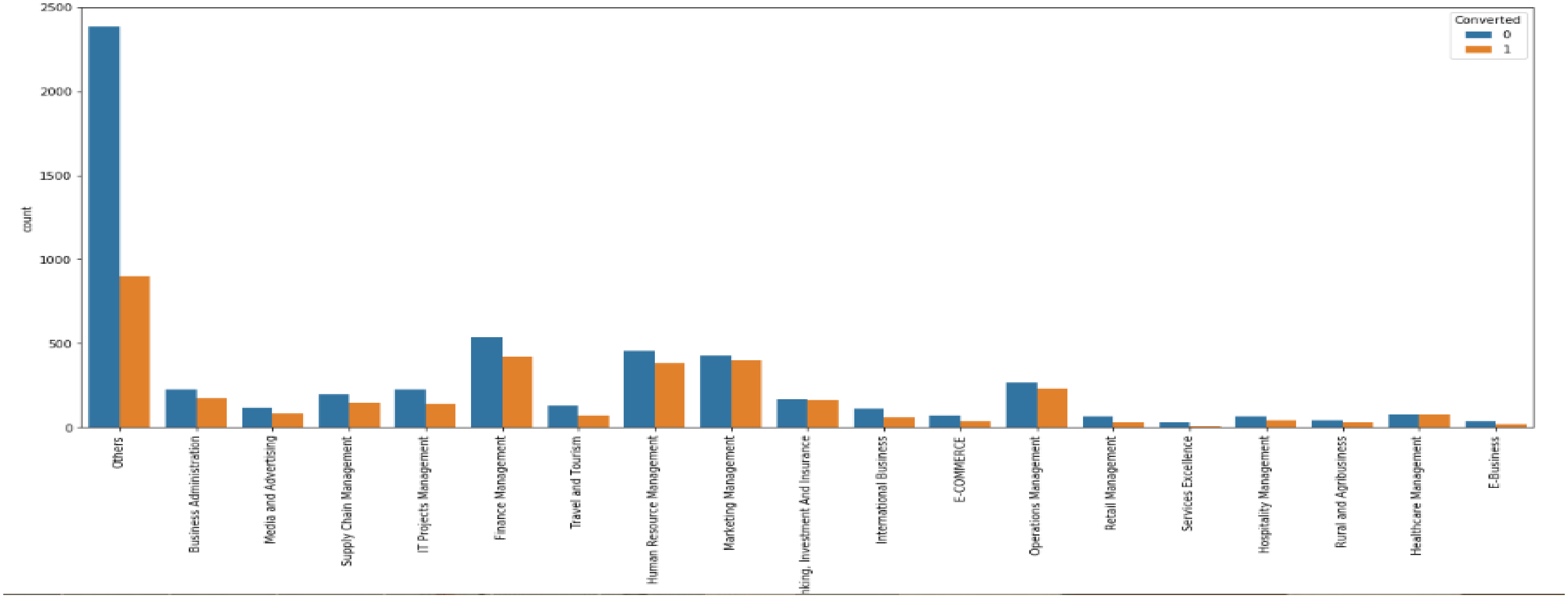
EDA: Last Activity



Inference:

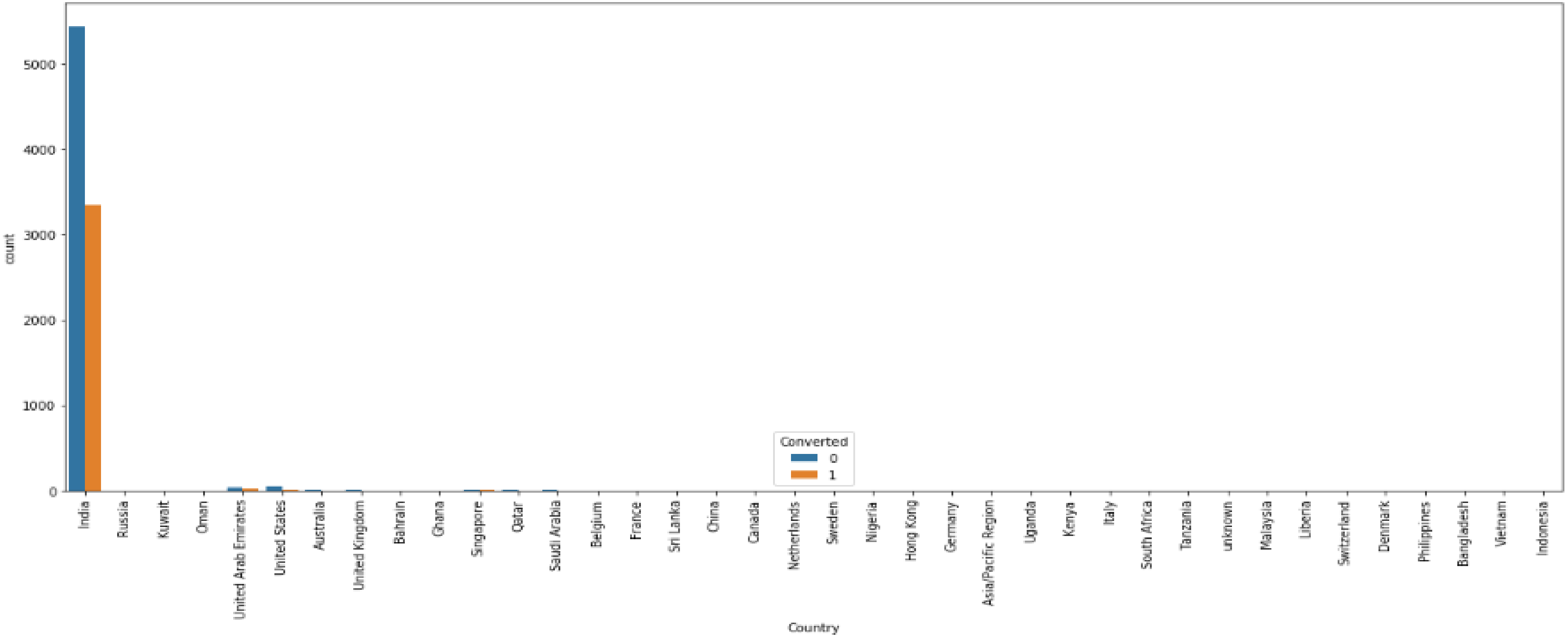
1.Most of the lead have their Email opened as their last activity.
2.Conversion rate for leads with last activity as SMS Sent is almost 60%.
3.next falls page visited on website also moderately considerable

EDA: Specialization



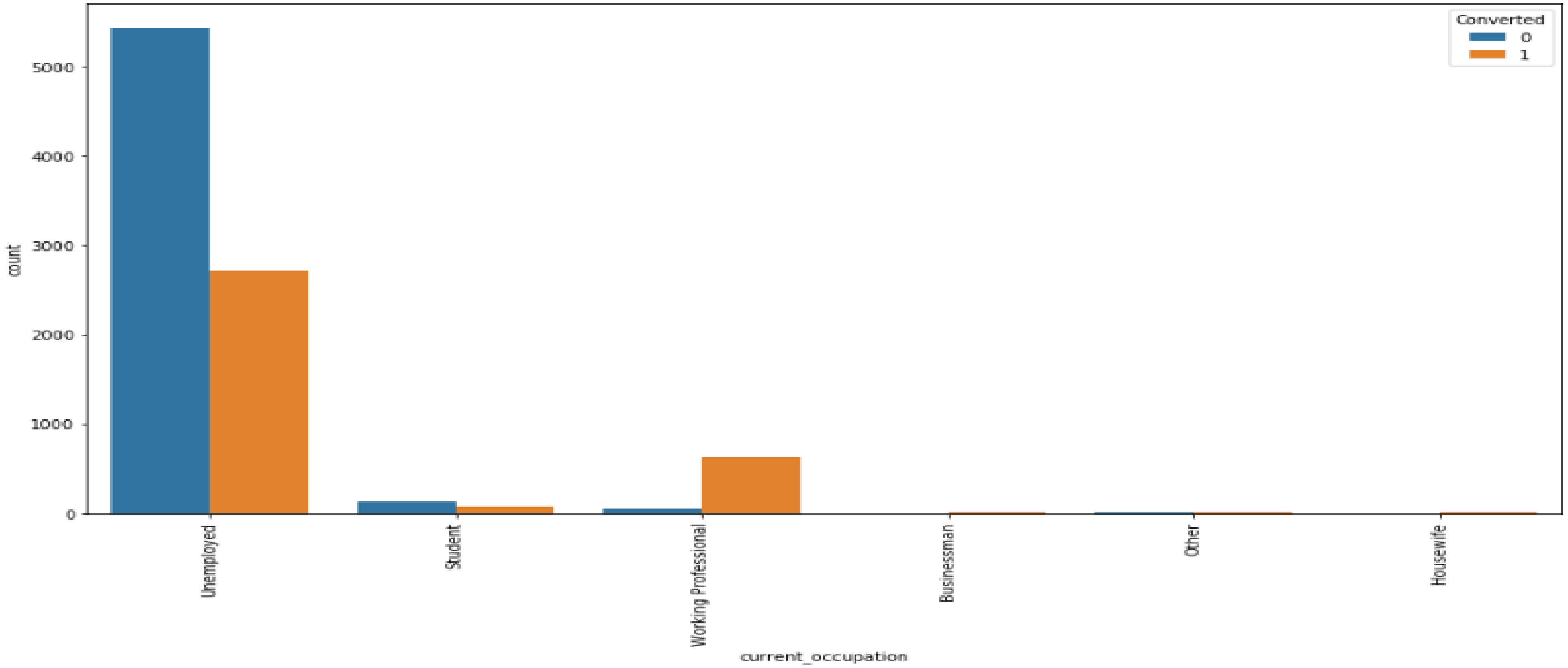
Inference
1.Leads are generating very good count on 'others' as their specilization
2.conversion rate are very high in 'HRM', 'Marketing Management','Operations Management' and 'Finance Management'

EDA: Country



Inference: Most Leads are from 'India' and no such inference can be drawn

EDA: Occupation

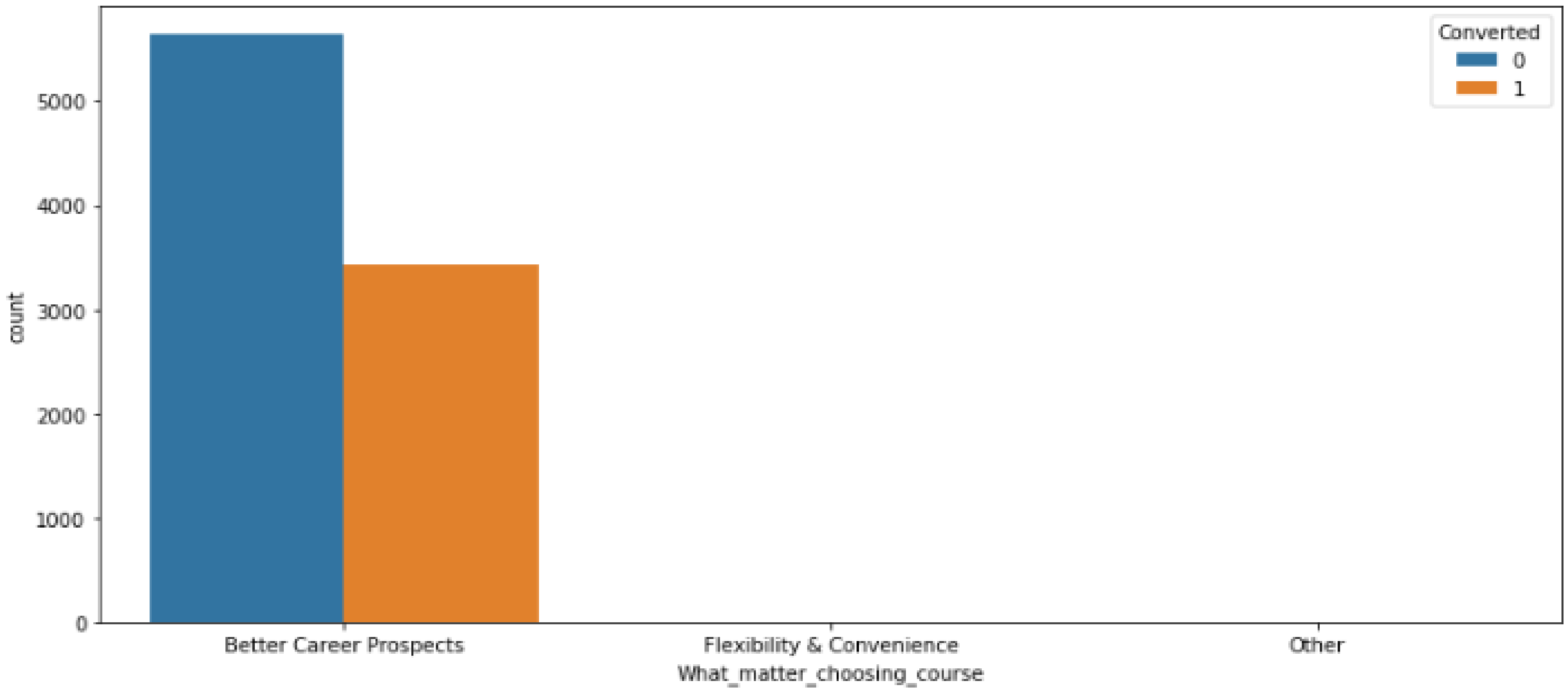


Inference:

1.Working Professionals going for the course have high chances of joining it.

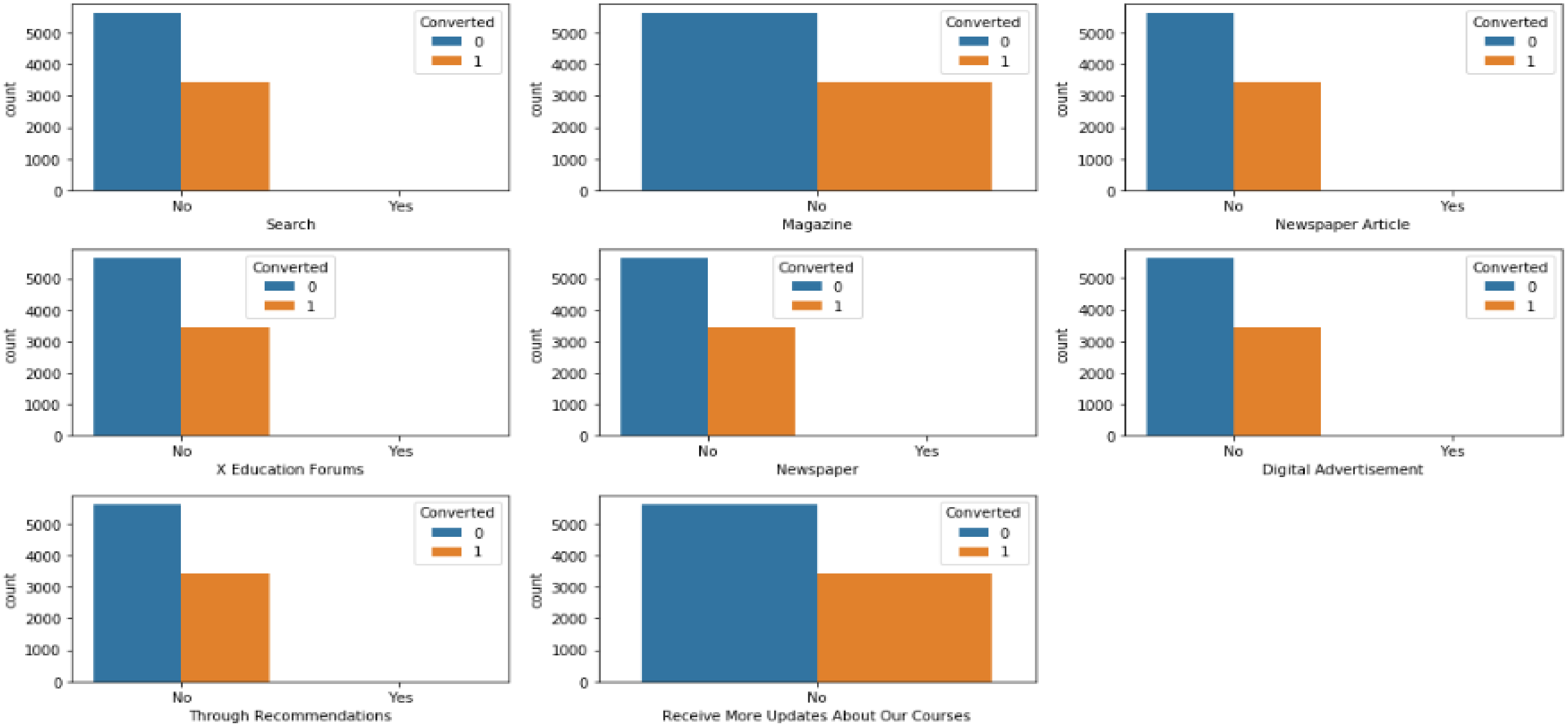
2.Unemployed leads are the high in numbers but has around 30-35% conversion rate.

EDA: ‘What_matter_choosing_course’ Metric

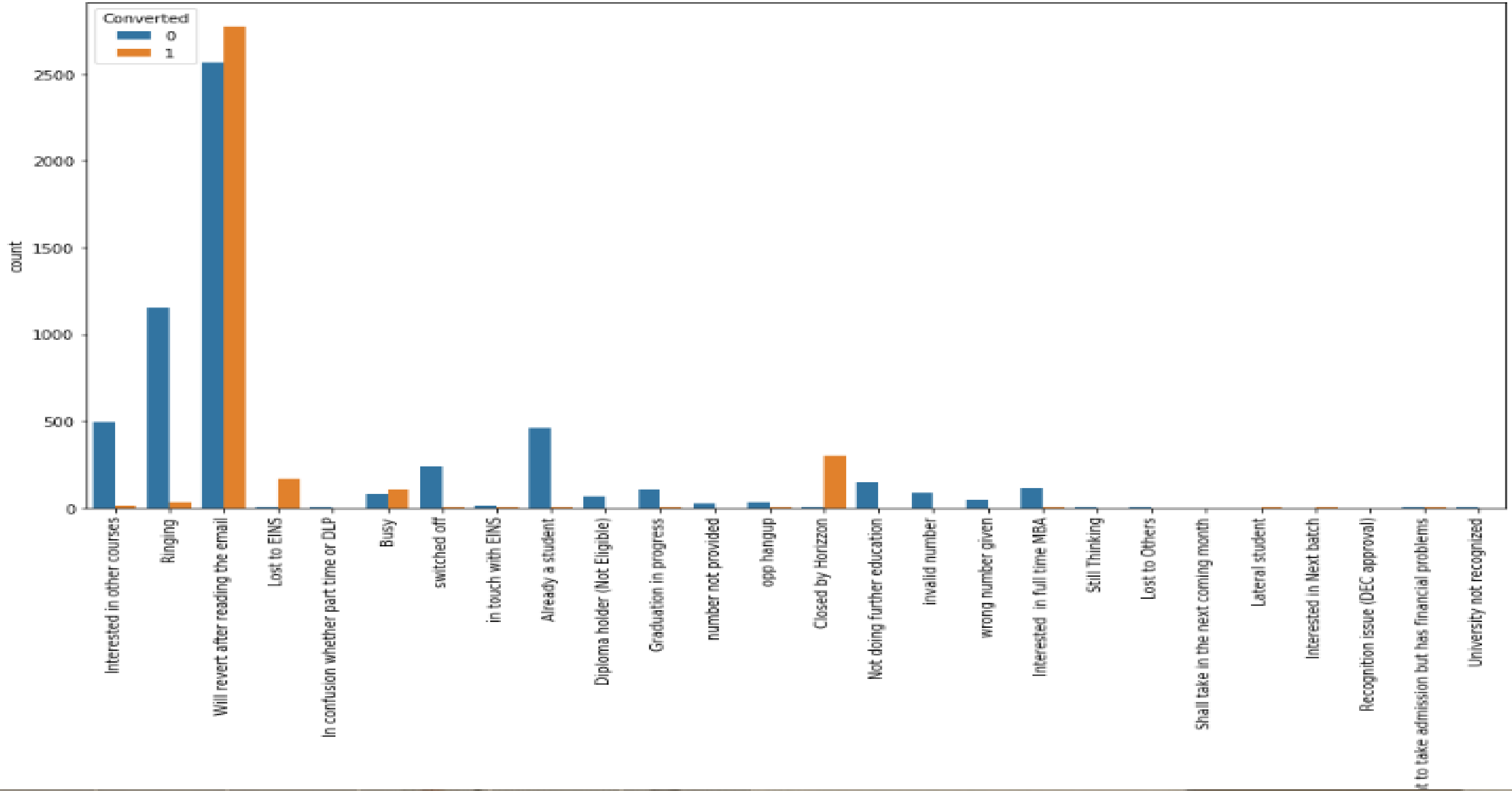


Inference: Most leads are 'Better Career Prospects'. No Inference can be drawn with this parameter.

EDA: Search, Magazine, NewsPaper Article, X Education Forums, NewsPaper, Digital Advertisement, Throug Recommendations, Receive more updates about our courses Metric's

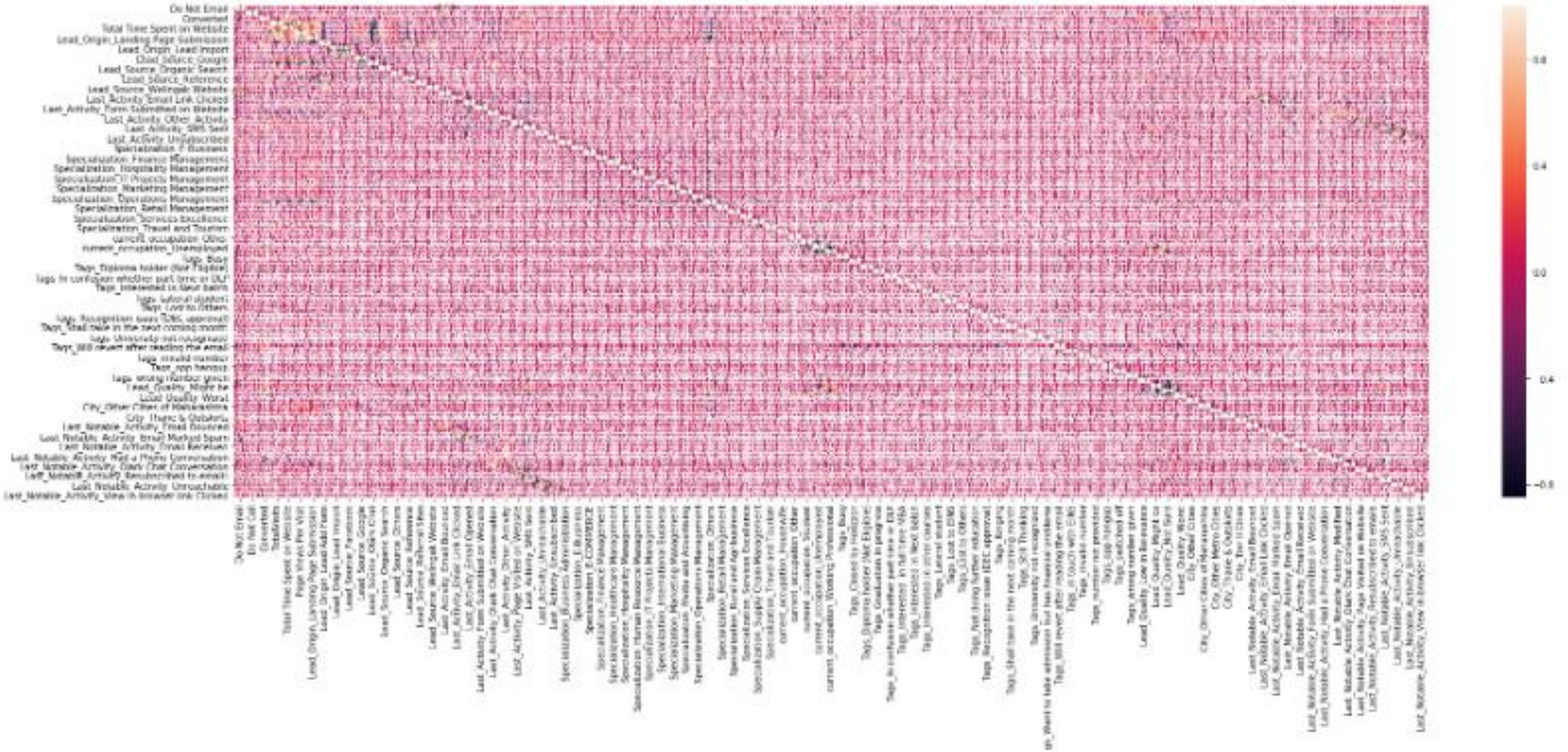


EDA:Tags Metric

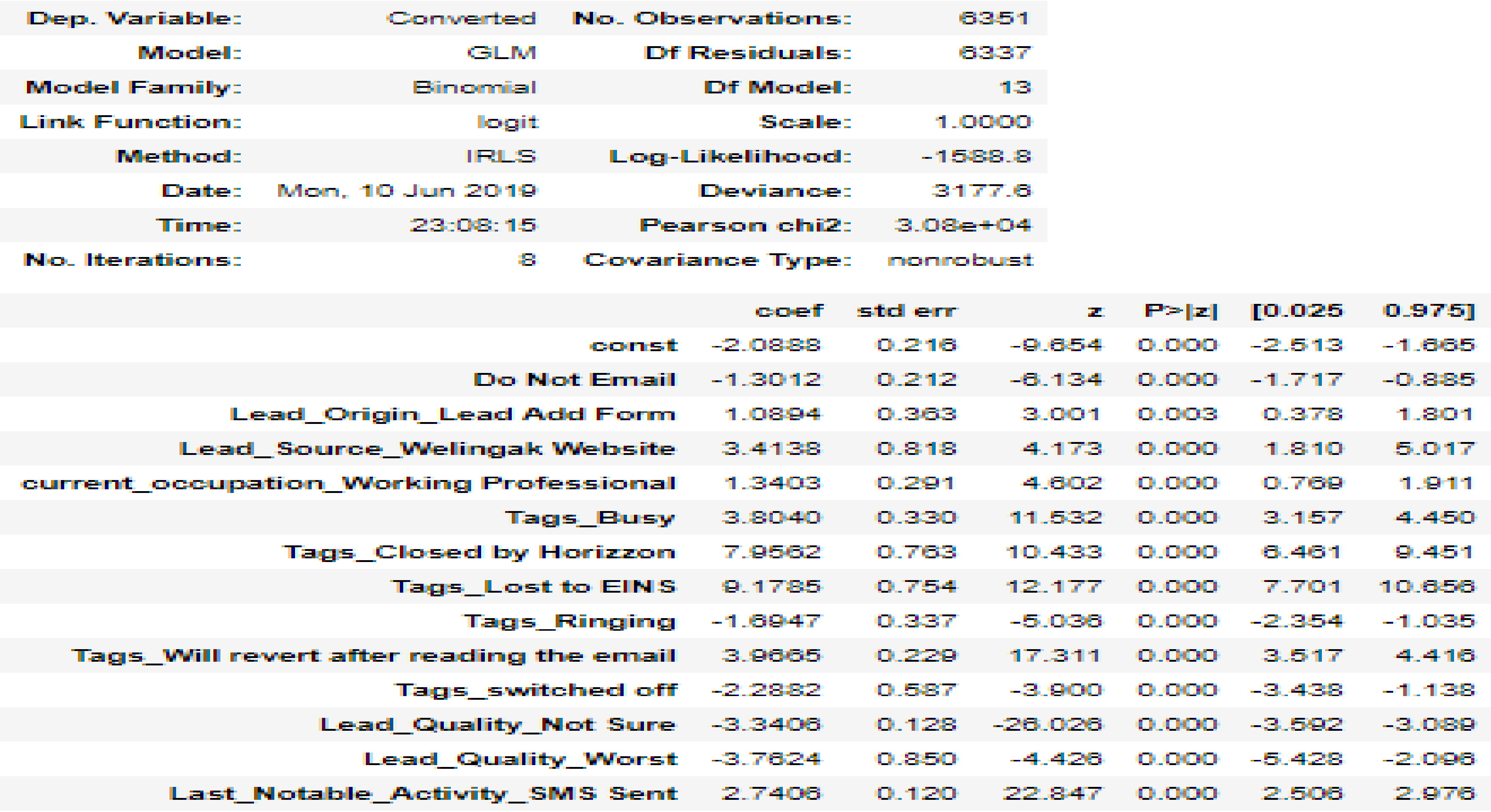


Inference: Many leads are telling that 'Will revert after reading the emai' and phone ringing status also

Correlation Matrix



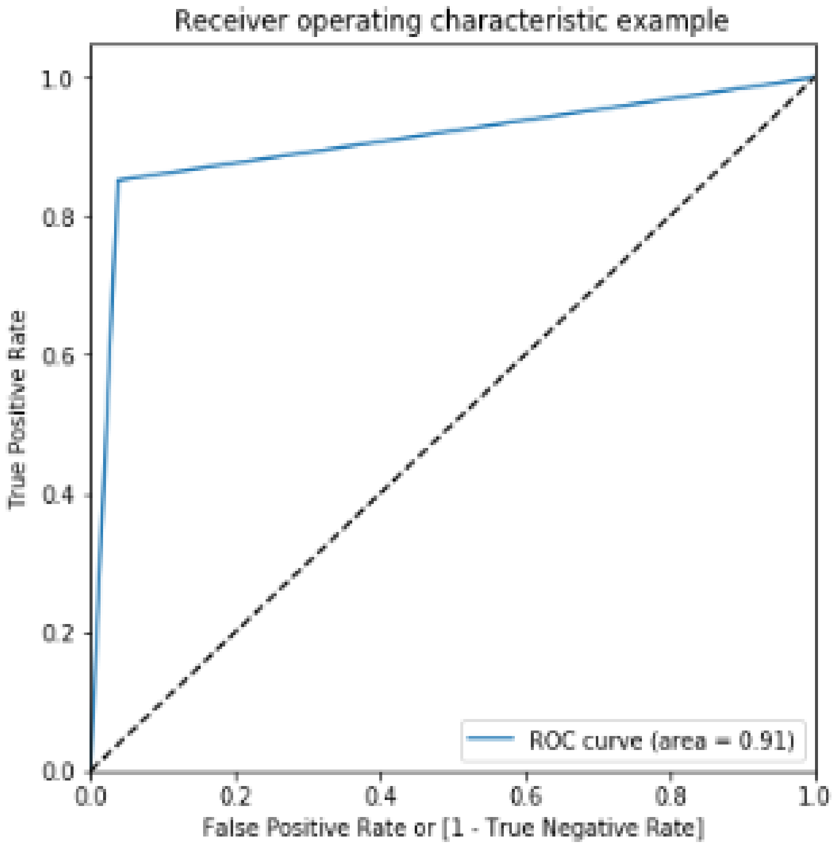
Model Building: With 15 Metrics by using RFE



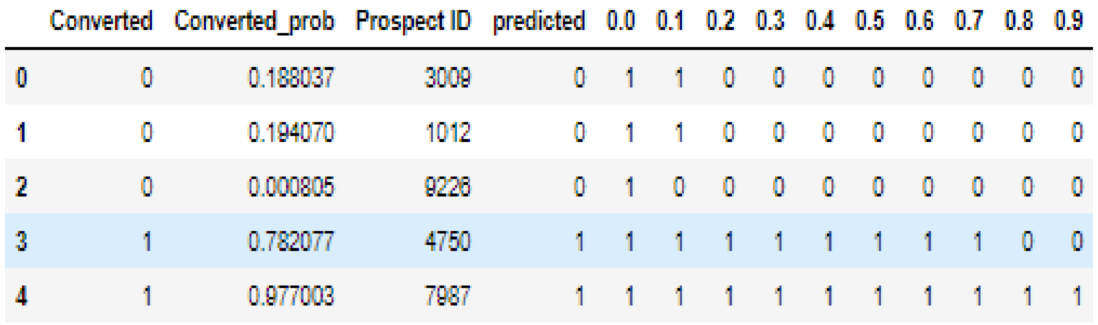
Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6337
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1588.8
Date:	Mon, 10 Jun 2019	Deviance:	3177.6
Time:	23:08:15	Pearson chi2:	3.08e+04
No. Iterations:	8	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0888	0.216	-9.654	0.000	-2.513	-1.665
Do Not Email	-1.3012	0.212	-6.134	0.000	-1.717	-0.885
Lead_Origin_Lead Add Form	1.0894	0.363	3.001	0.003	0.378	1.801
Lead_Source_Welingak Website	3.4138	0.818	4.173	0.000	1.810	5.017
current_occupation_Working Professional	1.3403	0.291	4.602	0.000	0.769	1.911
Tags_Busy	3.8040	0.330	11.532	0.000	3.157	4.450
Tags_Closed by Horizzon	7.9562	0.763	10.433	0.000	6.461	9.451
Tags_Lost to EINS	9.1785	0.754	12.177	0.000	7.701	10.656
Tags_Ringing	-1.6947	0.337	-5.036	0.000	-2.354	-1.035
Tags_Will revert after reading the email	3.9665	0.229	17.311	0.000	3.517	4.416
Tags_switched off	-2.2882	0.587	-3.900	0.000	-3.438	-1.138
Lead_Quality_Not Sure	-3.3406	0.128	-26.026	0.000	-3.592	-3.089
Lead_Quality_Worst	-3.7624	0.850	-4.426	0.000	-5.428	-2.096
Last_Notable_Activity_SMS Sent	2.7406	0.120	22.847	0.000	2.506	2.976

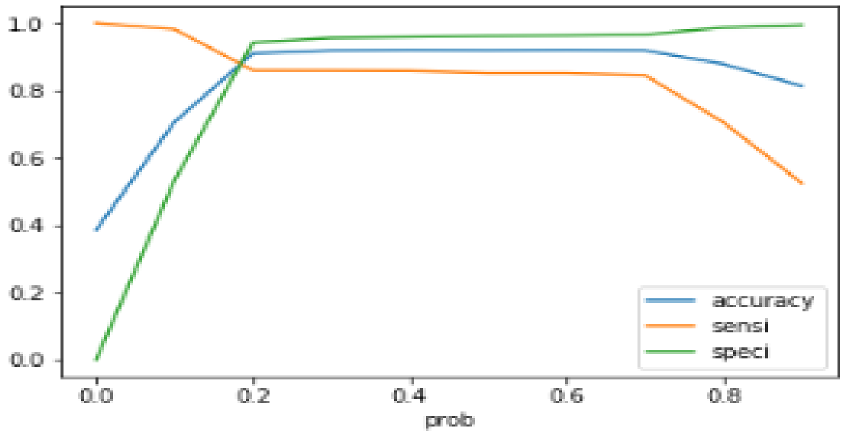
ROC Curve:



- Choosing cut off value

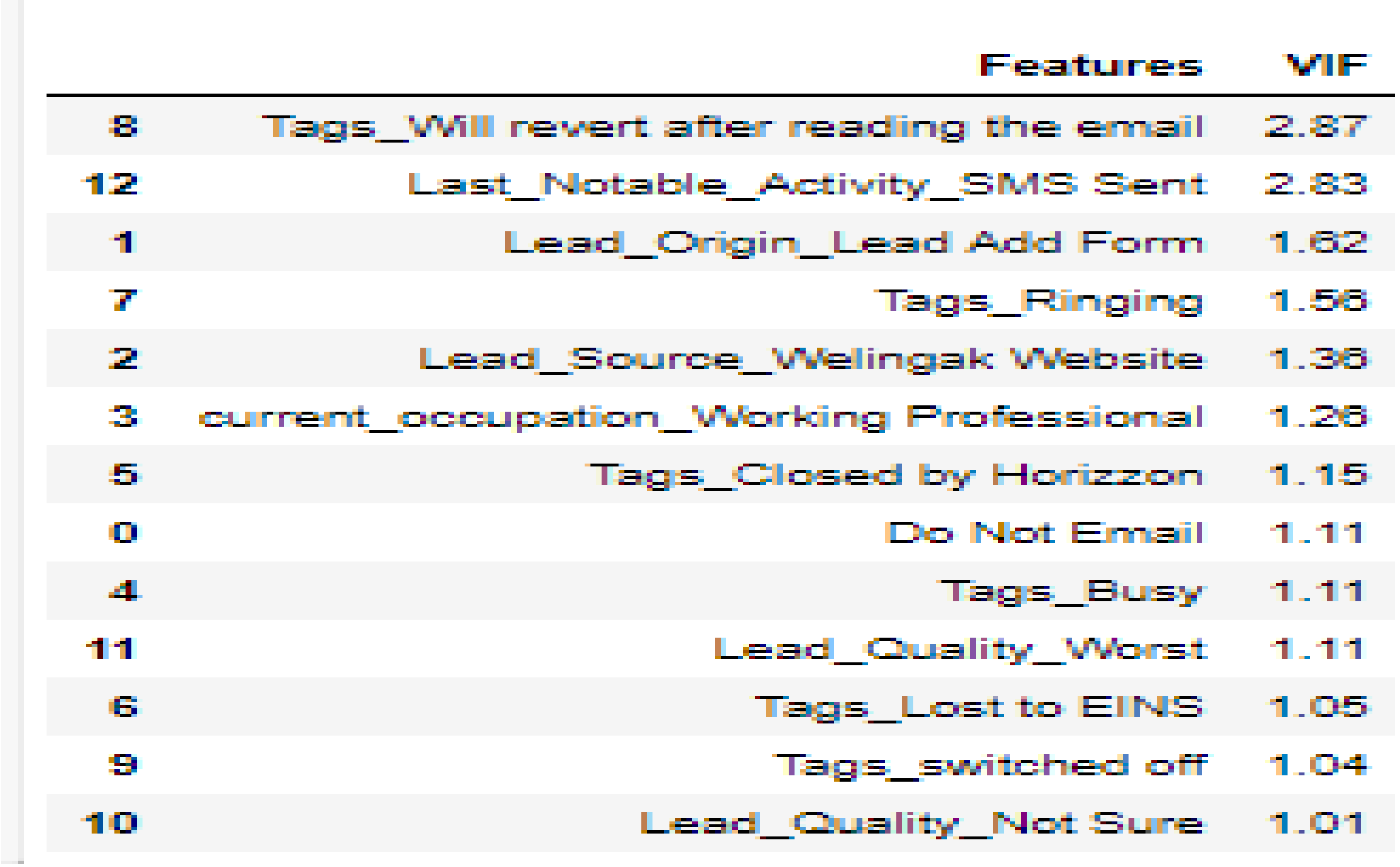


	Converted	Converted_prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.188037	3009	0	1	1	0	0	0	0	0	0	0	0
1	0	0.194070	1012	0	1	1	0	0	0	0	0	0	0	0
2	0	0.000805	9226	0	1	0	0	0	0	0	0	0	0	0
3	1	0.782077	4750	1	1	1	1	1	1	1	1	1	0	0
4	1	0.977003	7987	1	1	1	1	1	1	1	1	1	1	1



From the curve above, 0.2 is the optimum point to take it as a cutoff probability.

VIF:



	Features	VIF
8	Tags_Will revert after reading the email	2.87
12	Last_Notable_Activity_SMS Sent	2.83
1	Lead_Origin_Lead Add Form	1.62
7	Tags_Ringing	1.56
2	Lead_Source_Welingak Website	1.36
3	current_occupation_Working Professional	1.26
5	Tags_Closed by Horizzon	1.15
0	Do Not Email	1.11
4	Tags_Busy	1.11
11	Lead_Quality_Worst	1.11
6	Tags_Lost to EINS	1.05
9	Tags_switched off	1.04
10	Lead_Quality_Not Sure	1.01

As per VIF, choosen metrics falls below 5 where it satisfies VIF rule.

Conclusions & Recommendations

1. Most of the Leads generates from Mumbai city, unemployed people are approaching more and their specialization are also not disclosing properly so better to focus more on these part to turn up Leads into successful Leads.
2. After approaching it is clearly seen that Leads are replying back that ‘will revert after reading the email’ so make sure the sent mail should be fall into spam, if this one taken care properly leads turn into good count.
3. Many leads are turning successfully in some of the values like 'SMS sent', 'Email opened', 'Modified‘ so should focus more on 'Modified' and 'Email opened' activity leads.