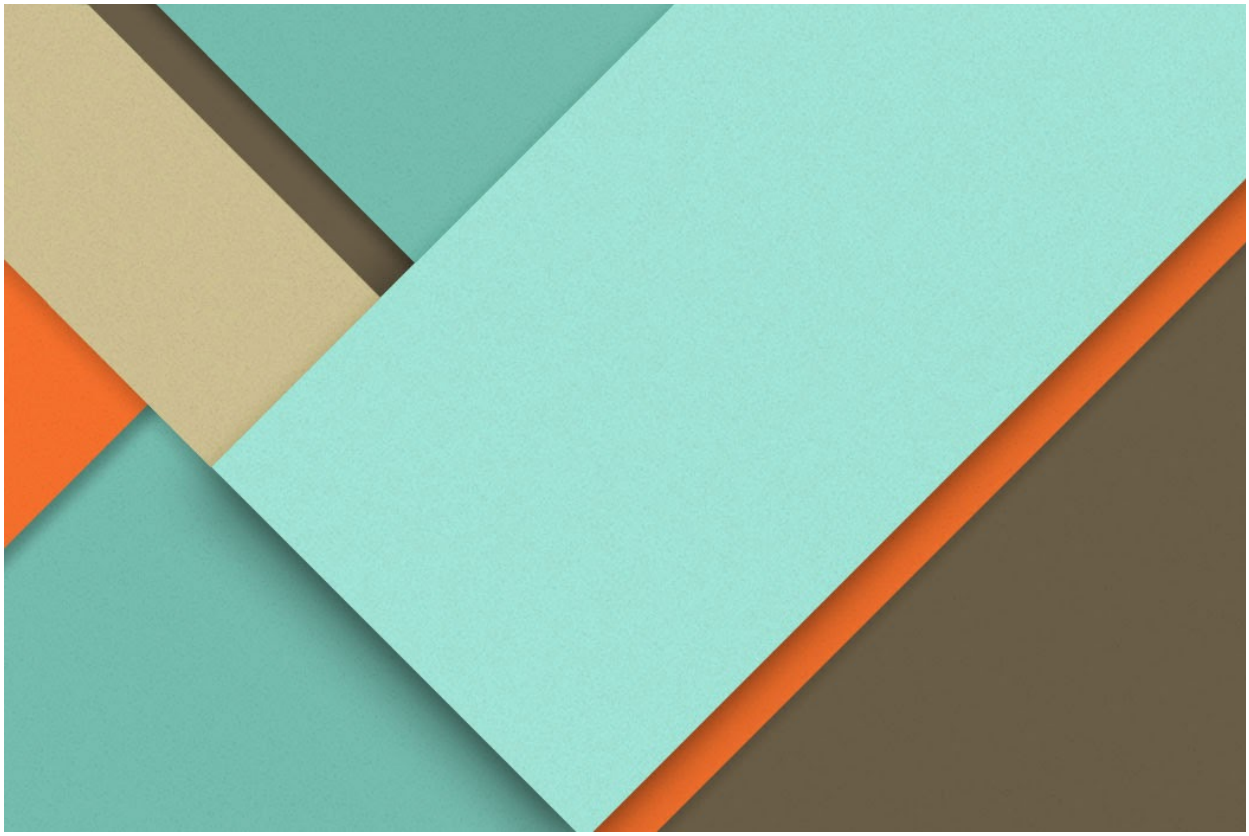


Lead Scoring Case Study

14.06.2021

Kishor Kunal | Sapna Mathur
DS27

Overview

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.

The company markets its courses on several websites and search engines like Google. 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although 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 about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. 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.

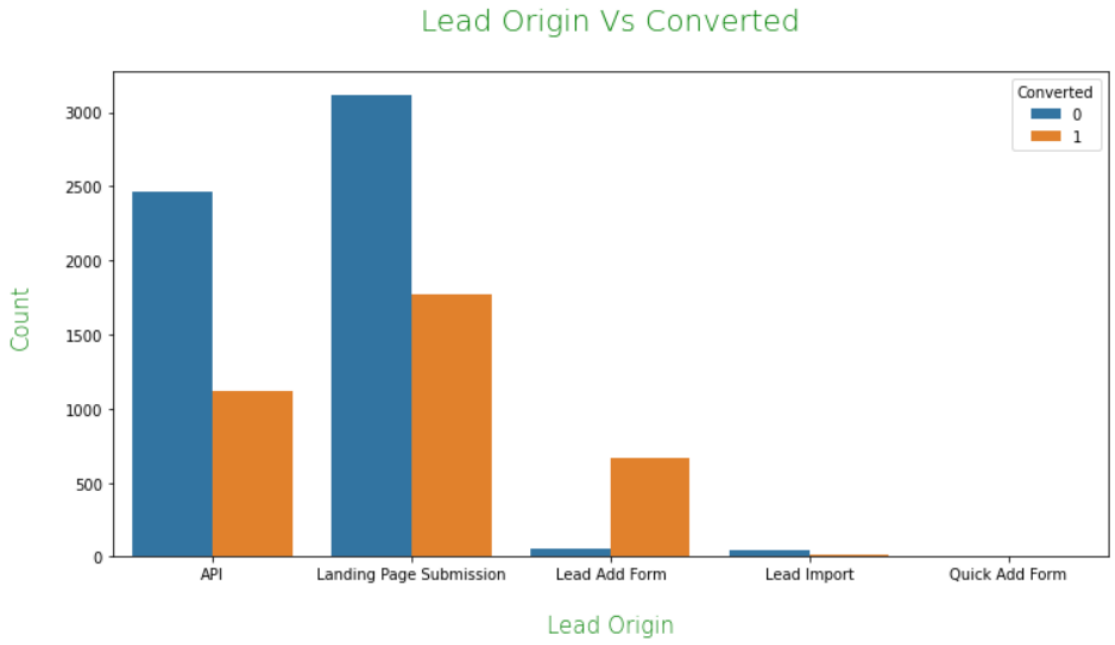
This case study is to help X Education in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals

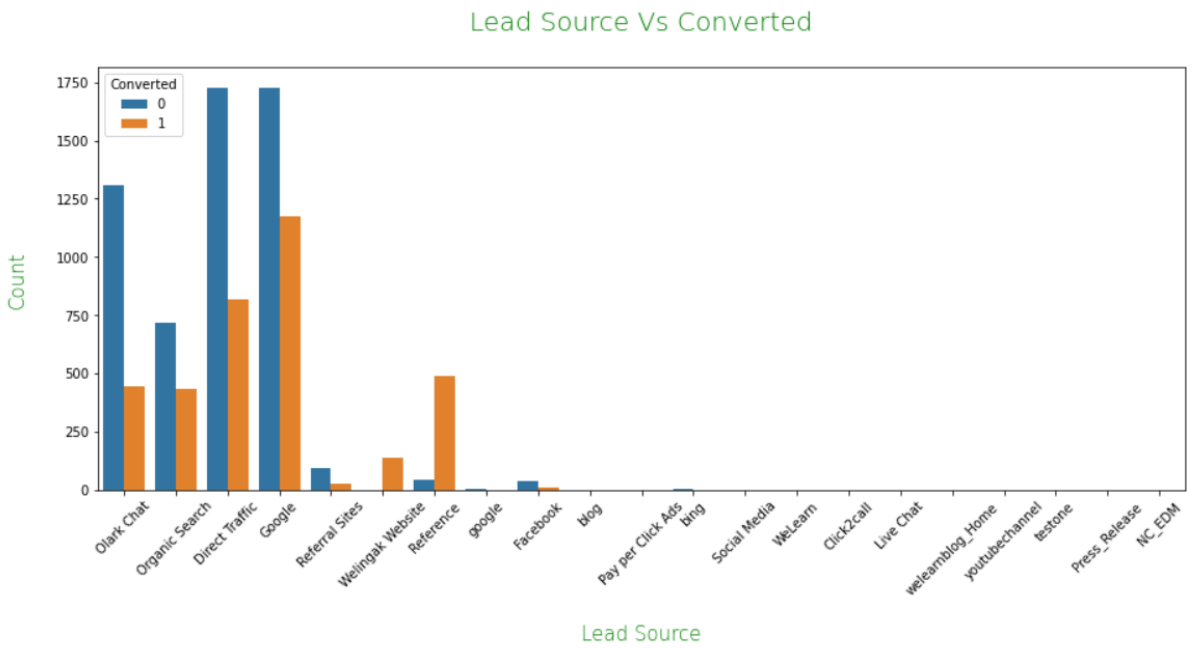
1. To Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will 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 we 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.

EDA and Model Observations

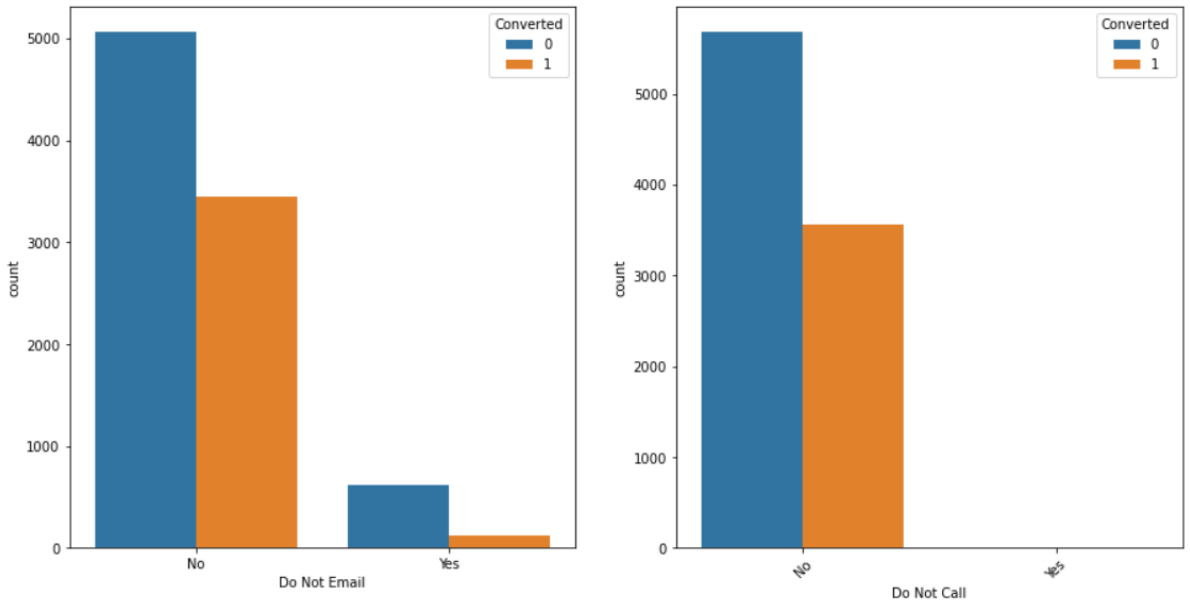
Following are our observations based on the data -



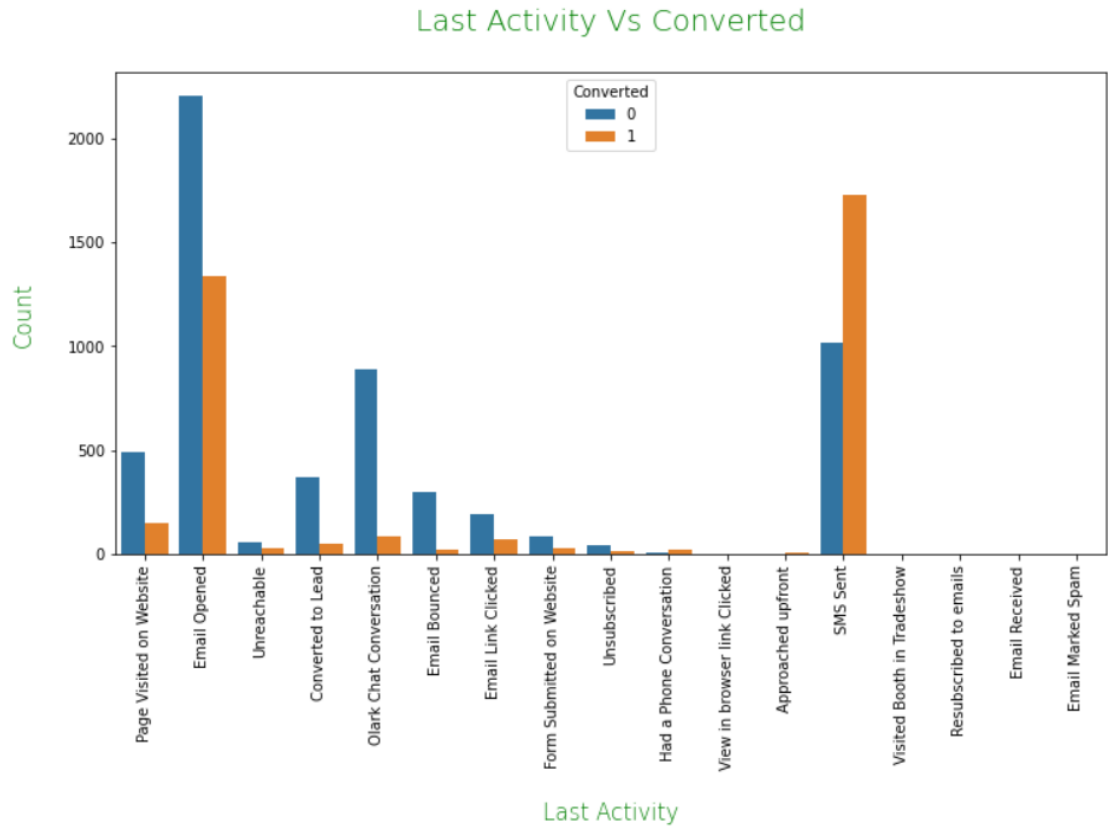
:::: Landing Page Submission has the most conversion rate



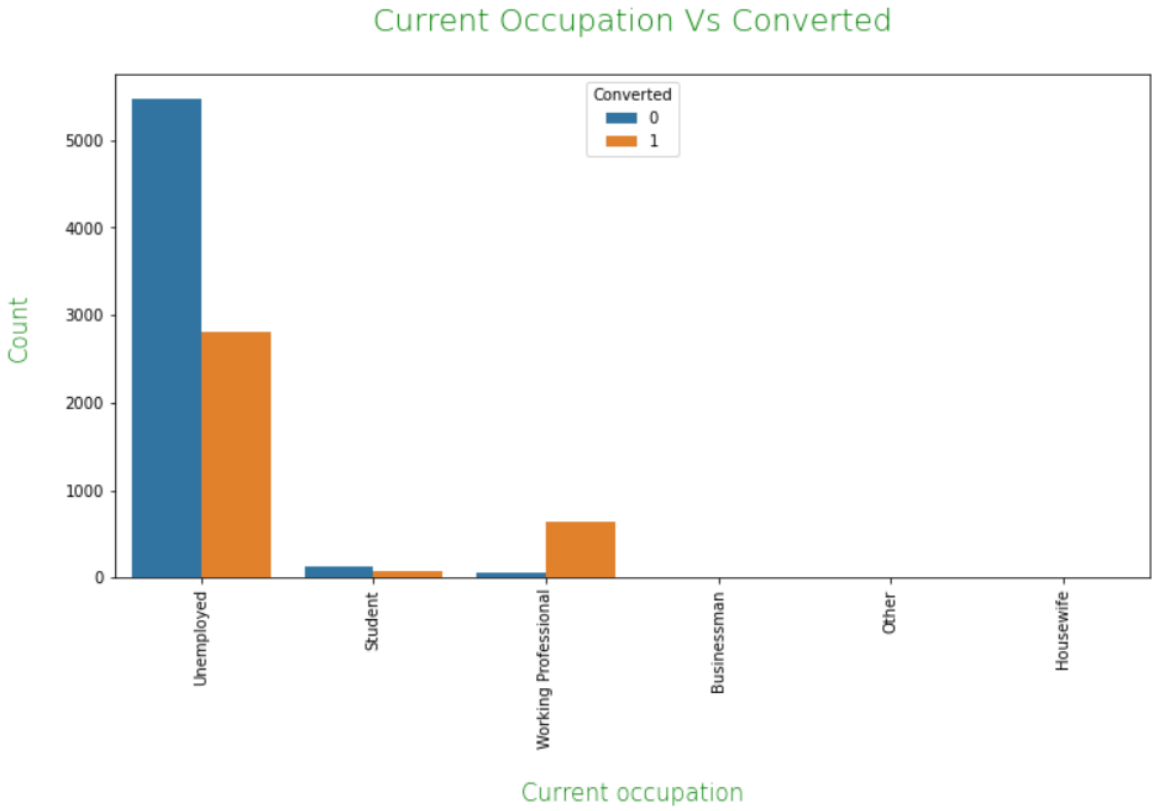
:::: Google leads have the most conversion rate



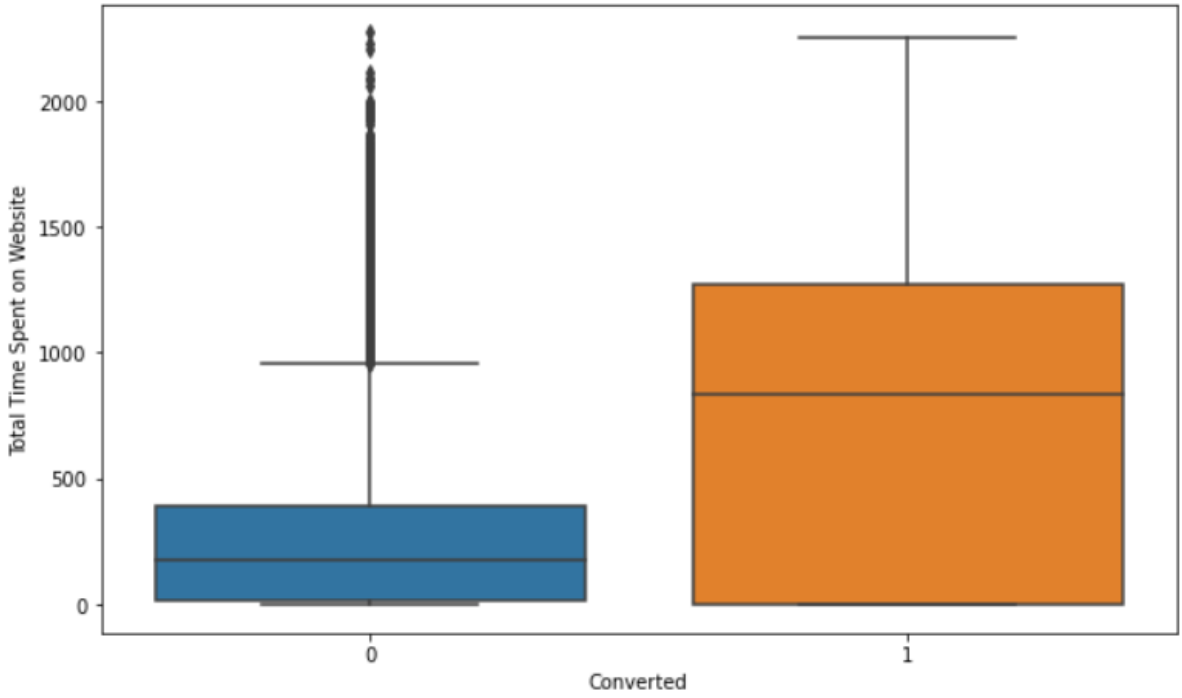
:::: Users chosen NO to email and calls have the most conversion rate



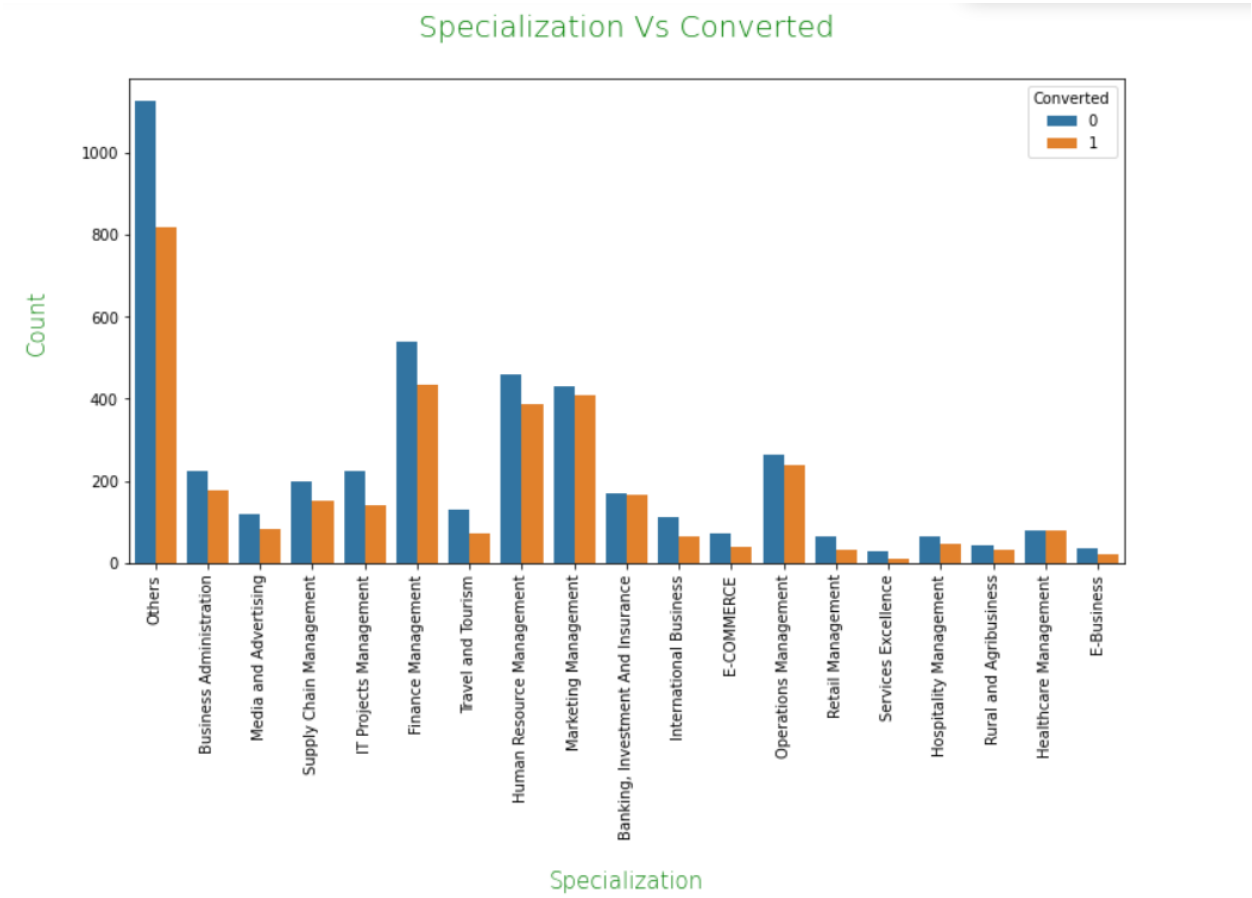
:::: Email and SMS as the last activty has the best conversation rate



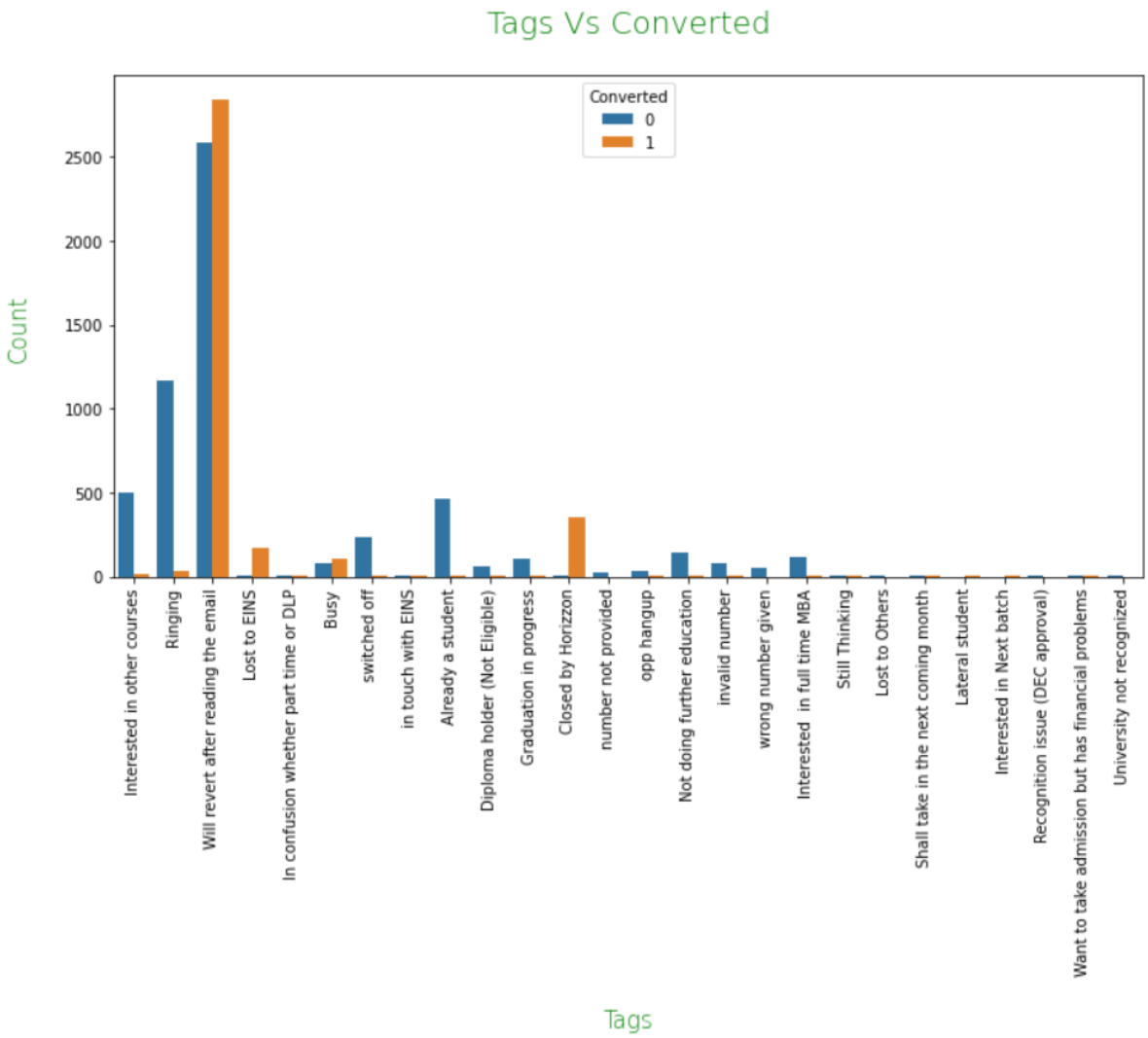
:::: Unemployed users are the target users as they have the best conversation rate



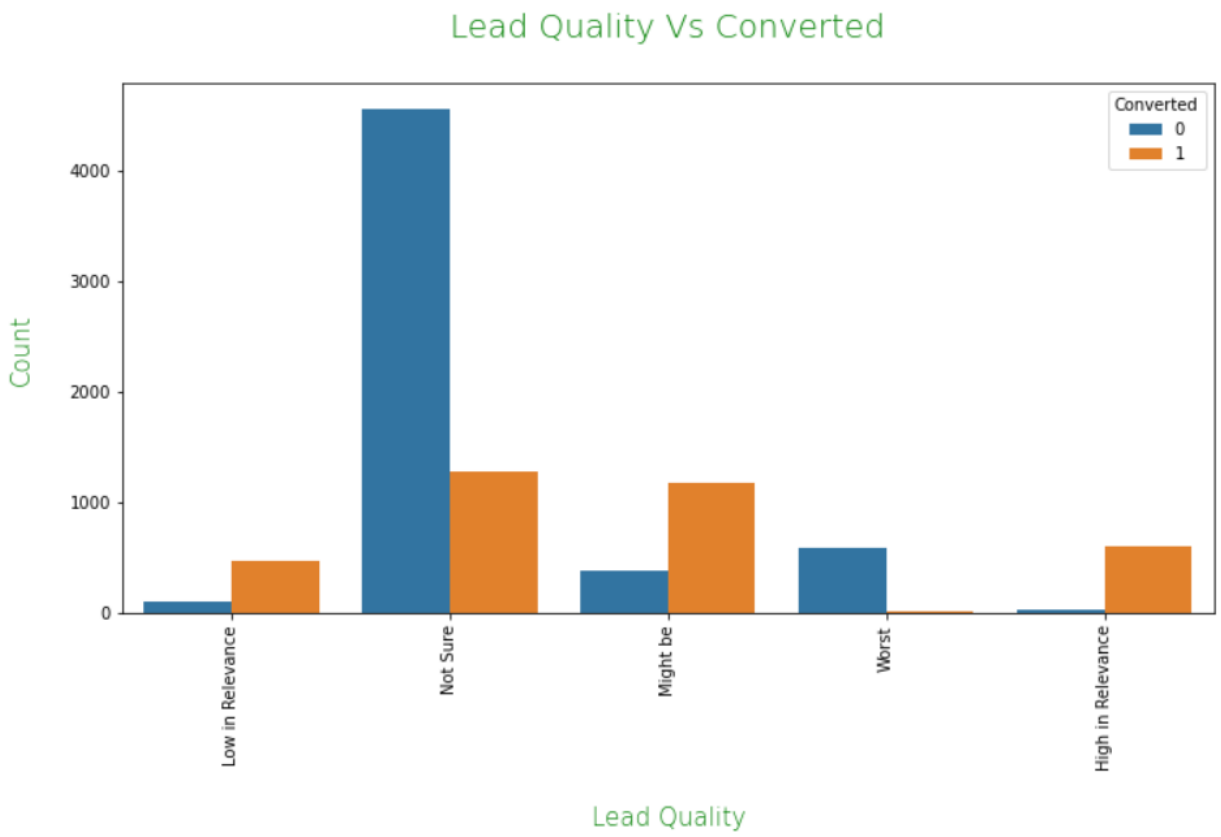
:::: Users spending more time on site are more likely to for Conversion



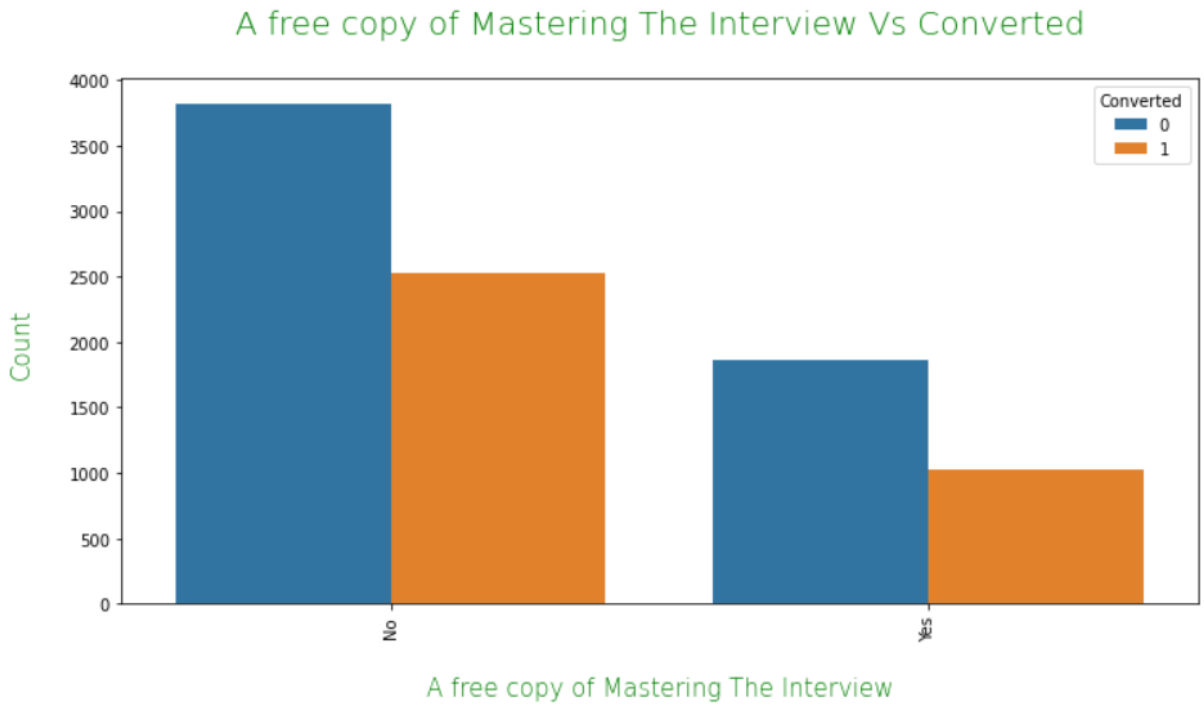
:::: Finance management, Merketing management and Human Resouece management users are most likely to get converted



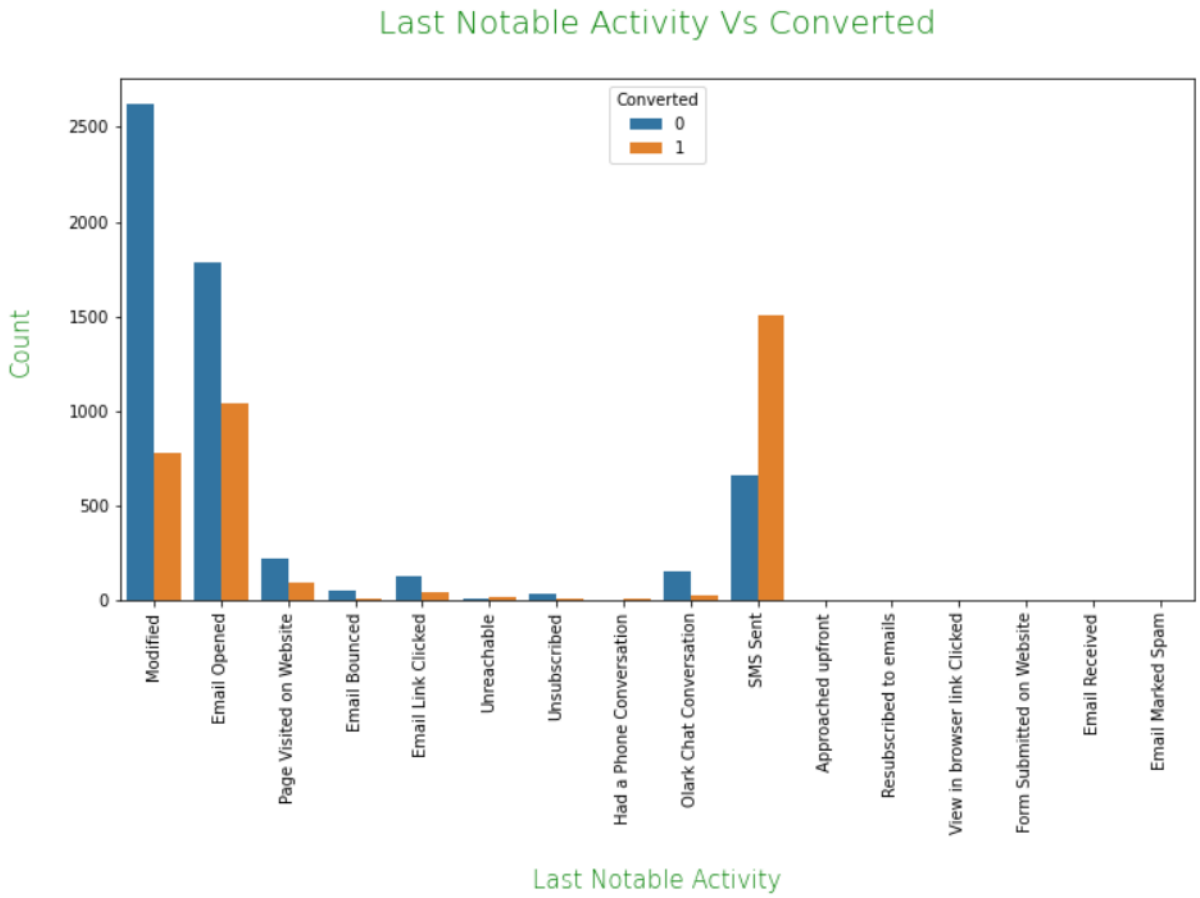
:::: 'Will revert after reading the email' - tagged users are most converted users



:::: 'Not Sure' and 'Might be' - are good indicators for Lead



:::: Customers Not asking for A free copy of Mastering The Interview - are also good indicators for Lead



:::: Users reading SMS as last activity are most likely to convert

Model 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:**	-1525.0
Date:	Mon, 14 Jun 2021	**Deviance:**	3050.0
Time:	21:10:36	**Pearson chi2:**	5.10e+04
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8078	0.223	-8.113	0.000	-2.245	-1.371
Lead Source_Welingak Website	3.2241	0.773	4.172	0.000	1.710	4.739
Last Activity_SMS Sent	1.9915	0.104	19.151	0.000	1.788	2.195
Specialization_Others	1.6191	0.127	12.784	0.000	1.371	1.867
What is your current occupation_Working Professional	1.6483	0.290	5.675	0.000	1.079	2.218
Tags_Already a student	-1.0983	0.786	-1.398	0.162	-2.639	0.442
Tags_Busy	3.9806	0.319	12.483	0.000	3.356	4.606
Tags_Closed by Horizzon	9.6307	1.051	9.163	0.000	7.571	11.691
Tags_Lost to EINS	9.9550	0.762	13.057	0.000	8.461	11.449
Tags_Ringing	-1.7755	0.321	-5.532	0.000	-2.405	-1.146
Tags_Will revert after reading the email	4.0303	0.232	17.337	0.000	3.575	4.486
Tags_switched off	-2.5316	0.596	-4.250	0.000	-3.699	-1.364
Lead Quality_Not Sure	-3.5871	0.133	-26.994	0.000	-3.848	-3.327
Lead Quality_Worst	-3.4406	0.719	-4.782	0.000	-4.851	-2.031
Last Notable Activity_Modified	-1.5507	0.106	-14.638	0.000	-1.758	-1.343

```
#confusion metrix
confusion = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.Final_Predicted )
confusion
```

```
array([[1160,  517],
       [  42, 1053]], dtype=int64)
```

```
TP = confusion[1,1] # True positiv
TN = confusion[0,0] # True negative
FP = confusion[0,1] # False positive
FN = confusion[1,0] # False negative
```

```
TP / float(TP+FN) # sensitivity
```

```
0.9616438356164384
```

```
TN / float(TN+FP) #specificity
```

```
0.691711389385808
```

```
#accuracy
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.Final_Predicted)
```

```
0.7983405483405484
```

Conclusion - Based on above observations - SMS with Ads like ‘Better Job opportunity’ are the best way to get leads.