Lead Score Case Study

Group Members
1. Gaurav Dabral
2. Taranveer Kaur

Problem Statement

- X Education sells online courses to industry professionals.
- 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.

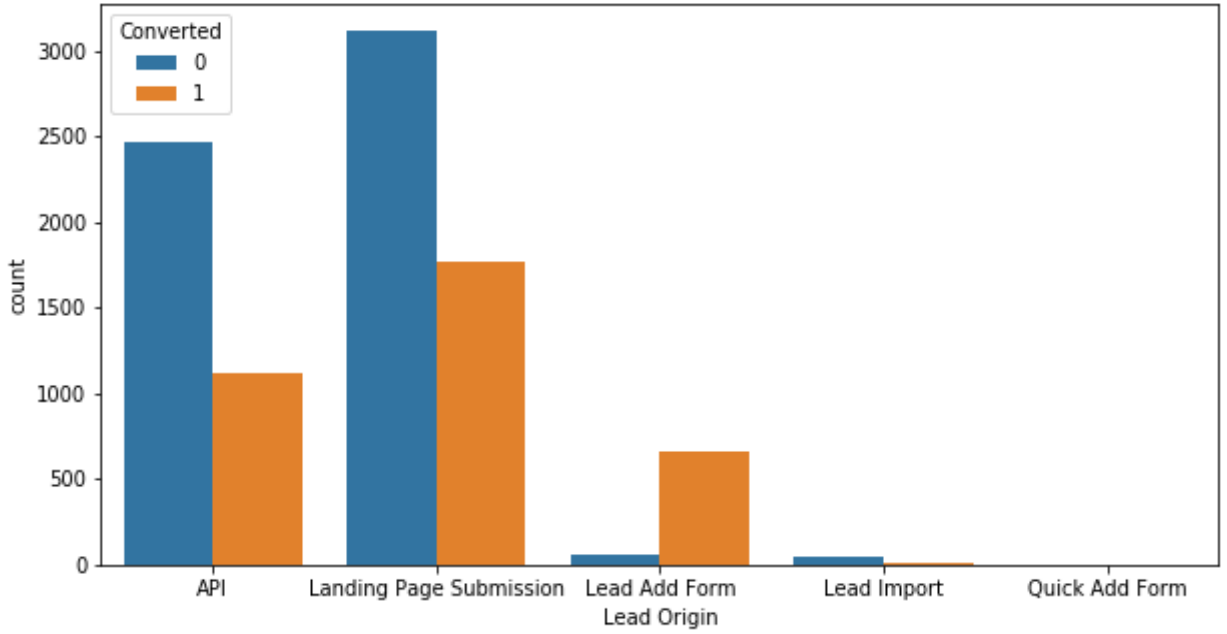
Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

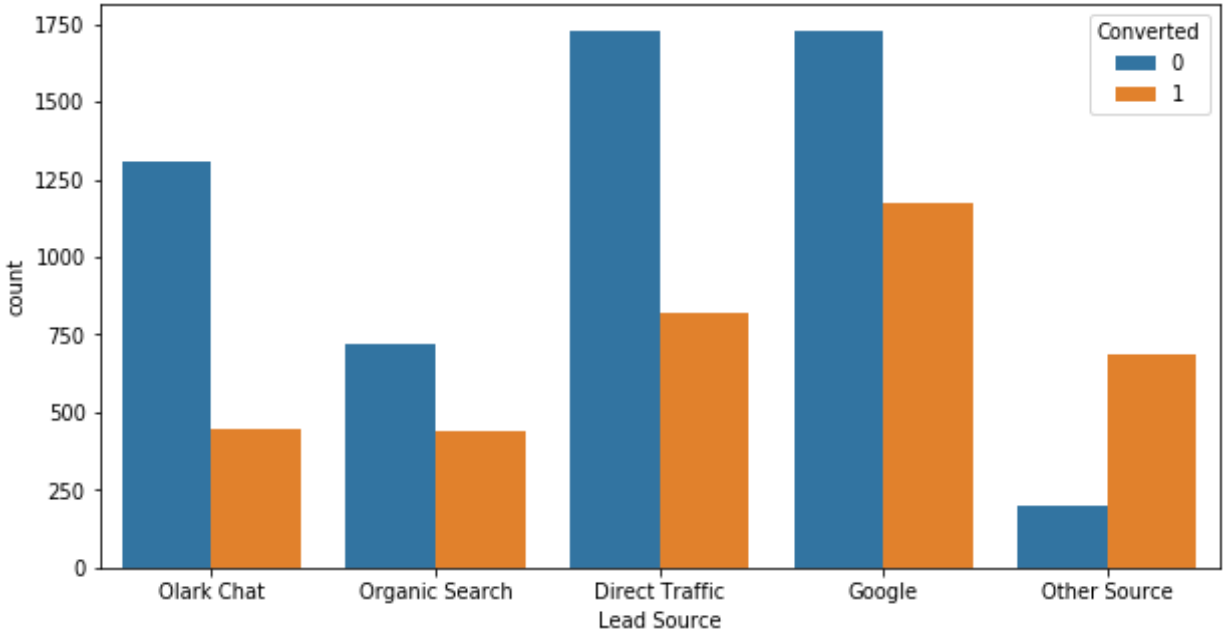
Solution Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NAN values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of remaining null values with suitable values, if necessary.
  5. Looking for skewed variables and also combining or eliminating categories that have low counts wherever necessary.
  6. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, count plots.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Logistic model building using RFE and manual feature elemination.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

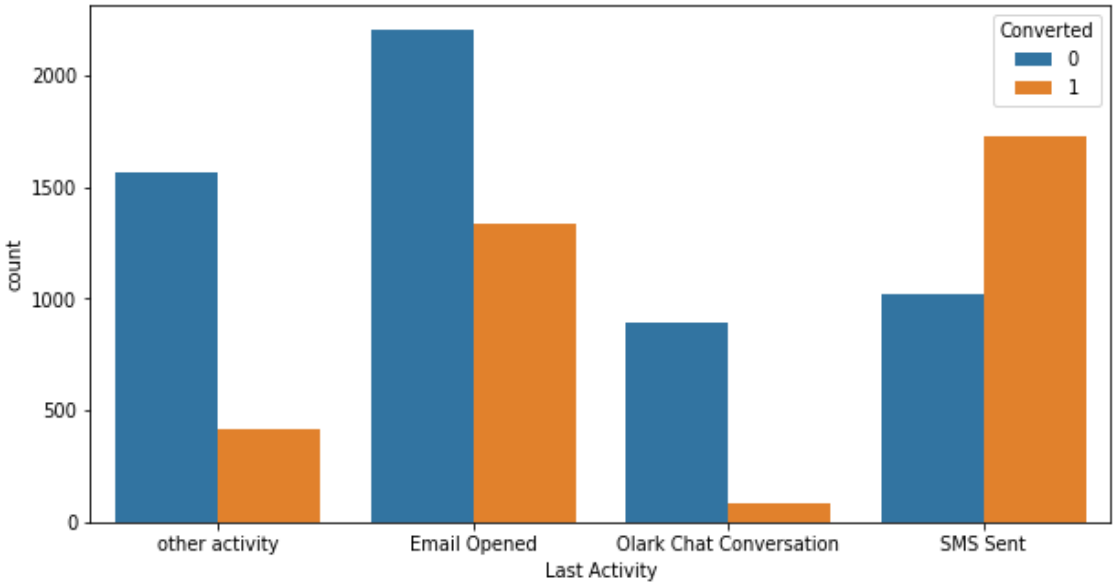
Data Visualization



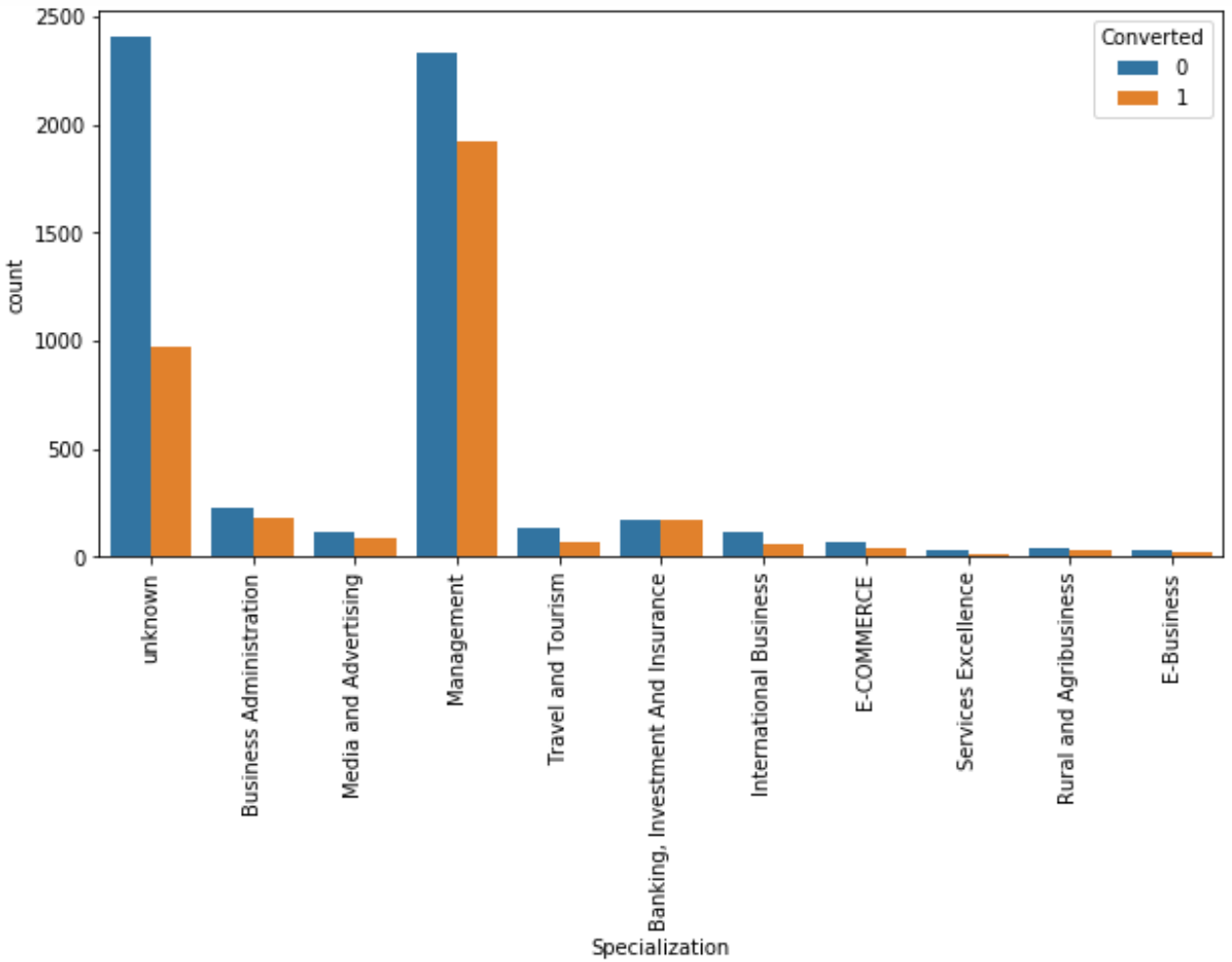
- Lead add form has high conversion rate.
- API and Landing page submission have poor conversion rate



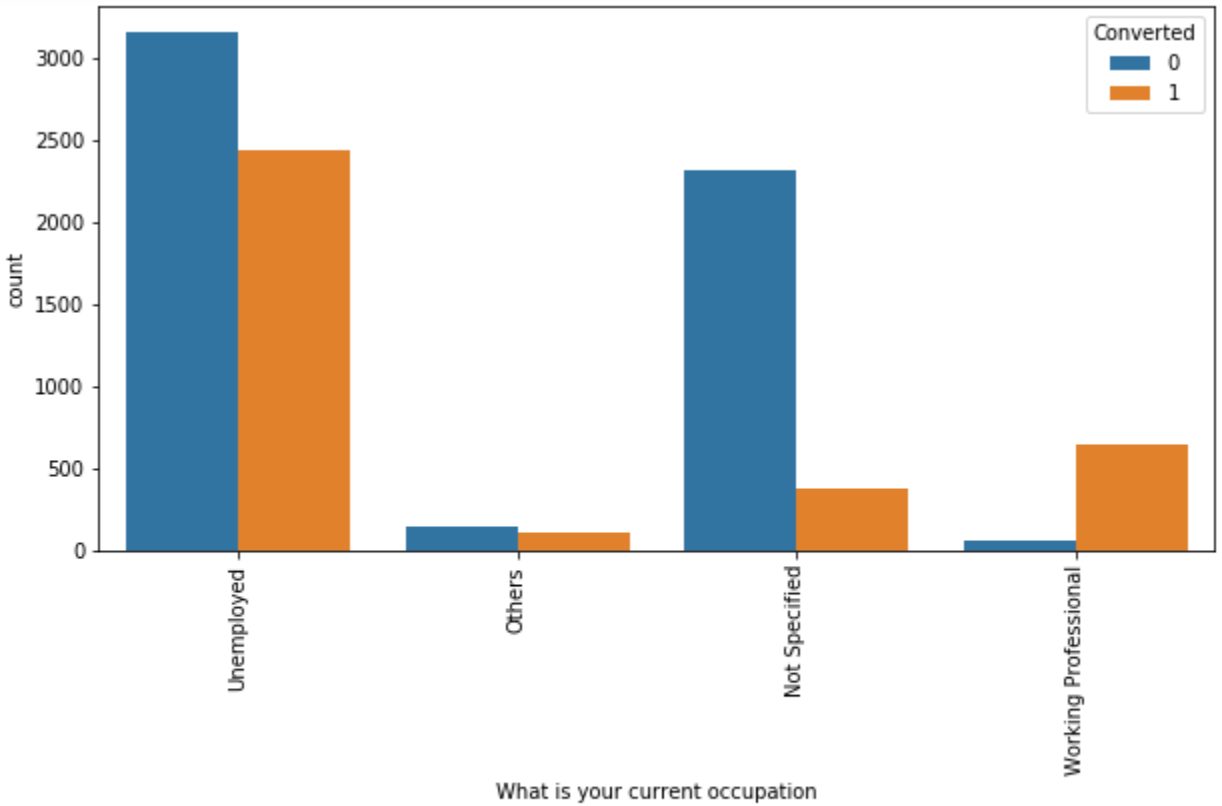
- Other sources have high conversion rate
- Leads from Olark Chat have a very low conversion rate



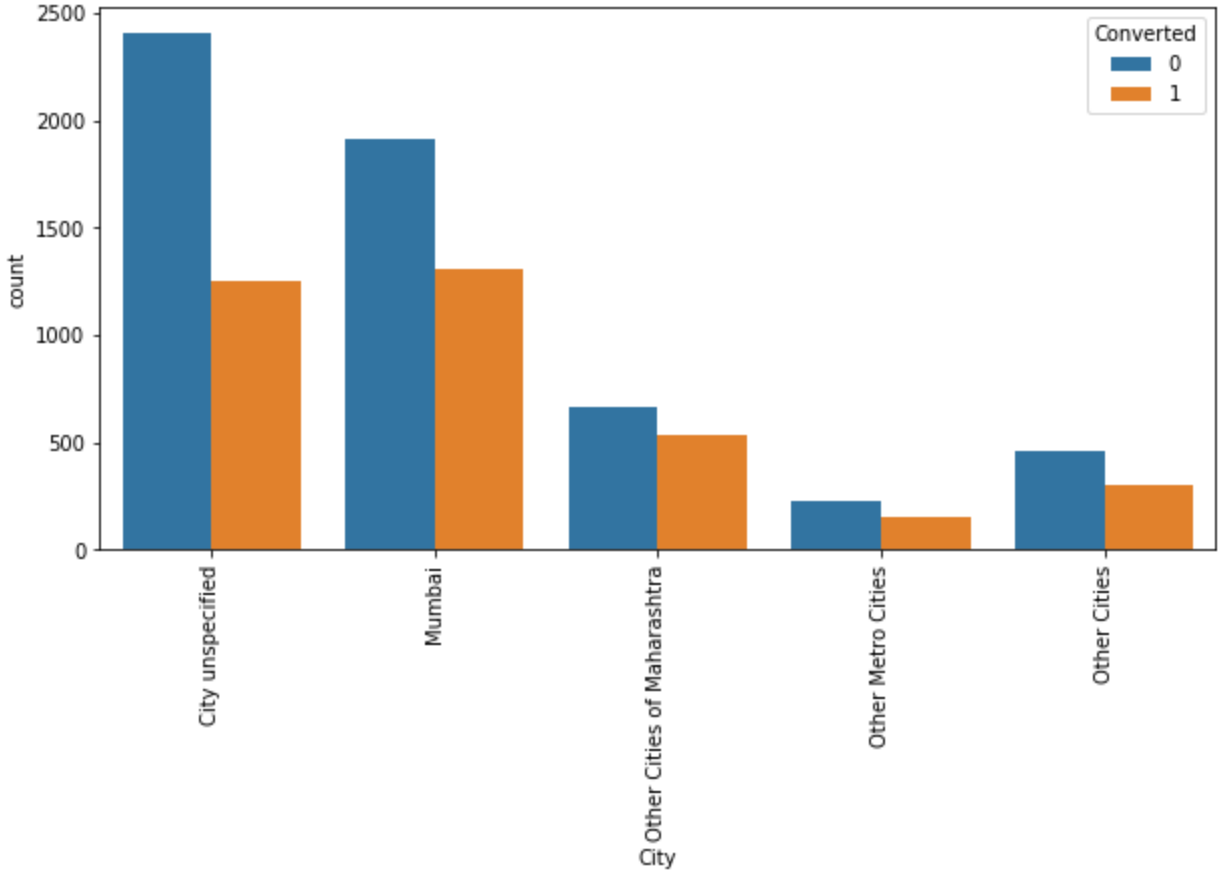
- The ones with last activity as Olark chat conversation and other activity have a very low conversion rate
- The ones with last activity as SMS Sent have a very good conversion rate.



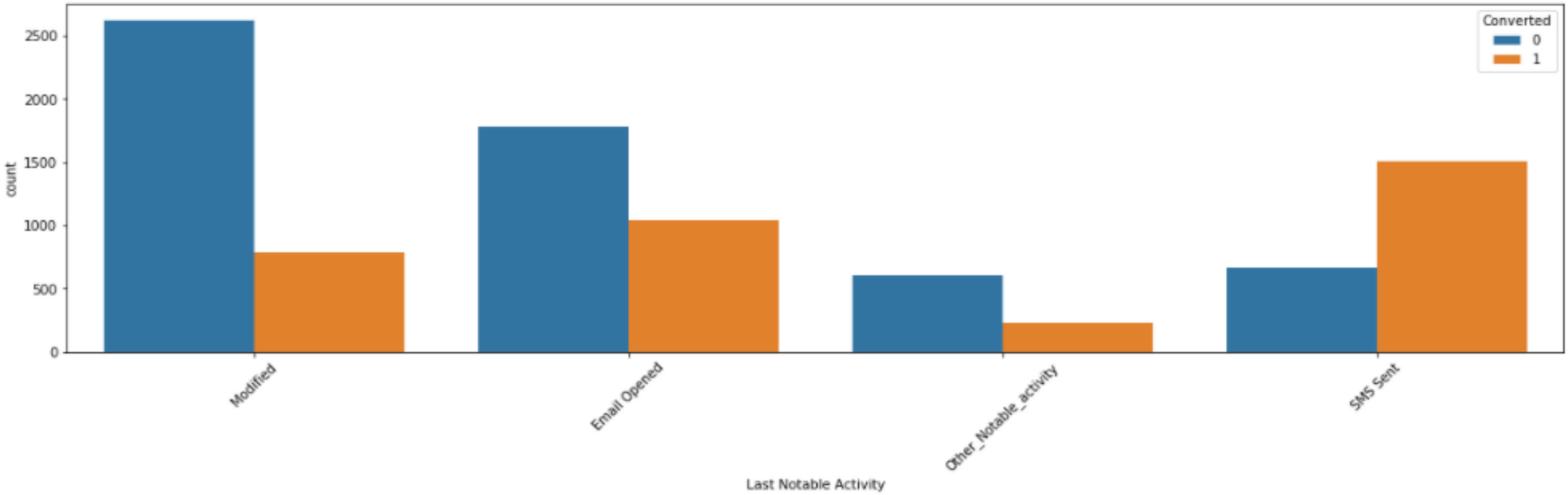
- The ones that have not provided with their Specialization have a very poor conversion rate.
- People Specialized in banking, Investment, insurance and managment have a decent conversion rate



- Working professionals have a very good conversion rate
- The people who have not specified their profession have a poor conversion rate.



- People with un specified cities have poor conversion rate.



- People with last notable activity as SMS sent have a very high conversion rate
- Modified has a very poor conversion rate.

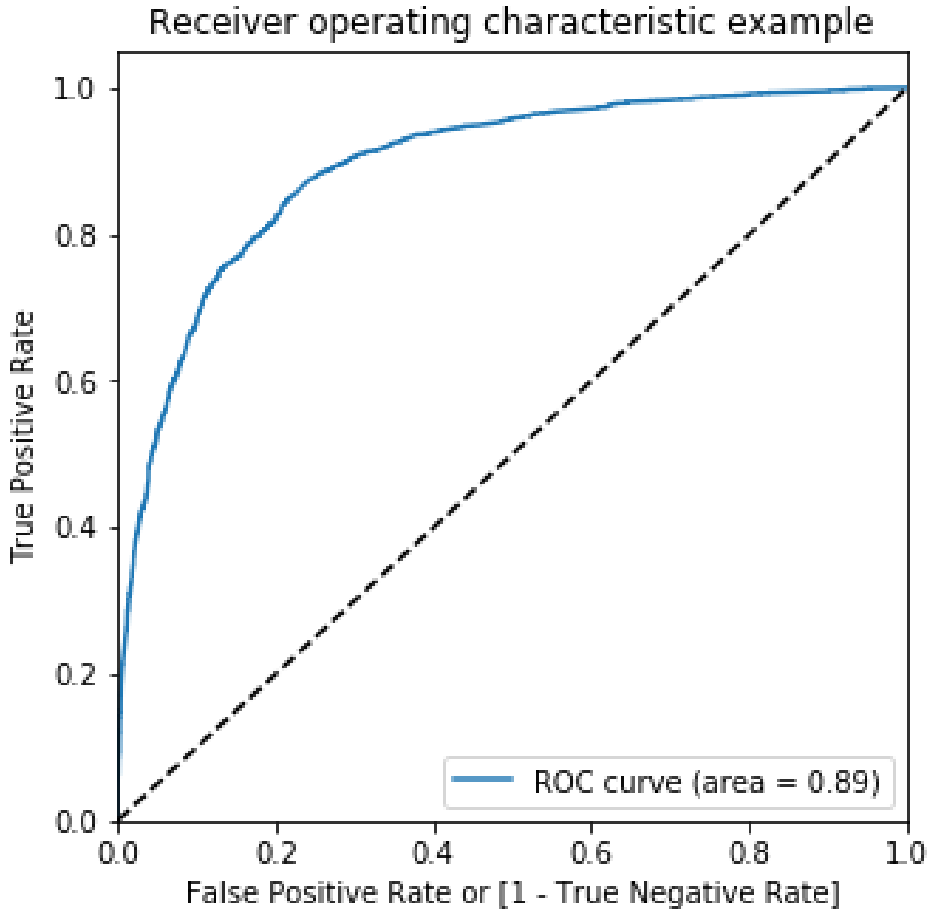
Final Model

Generalized Linear Model Regression Results

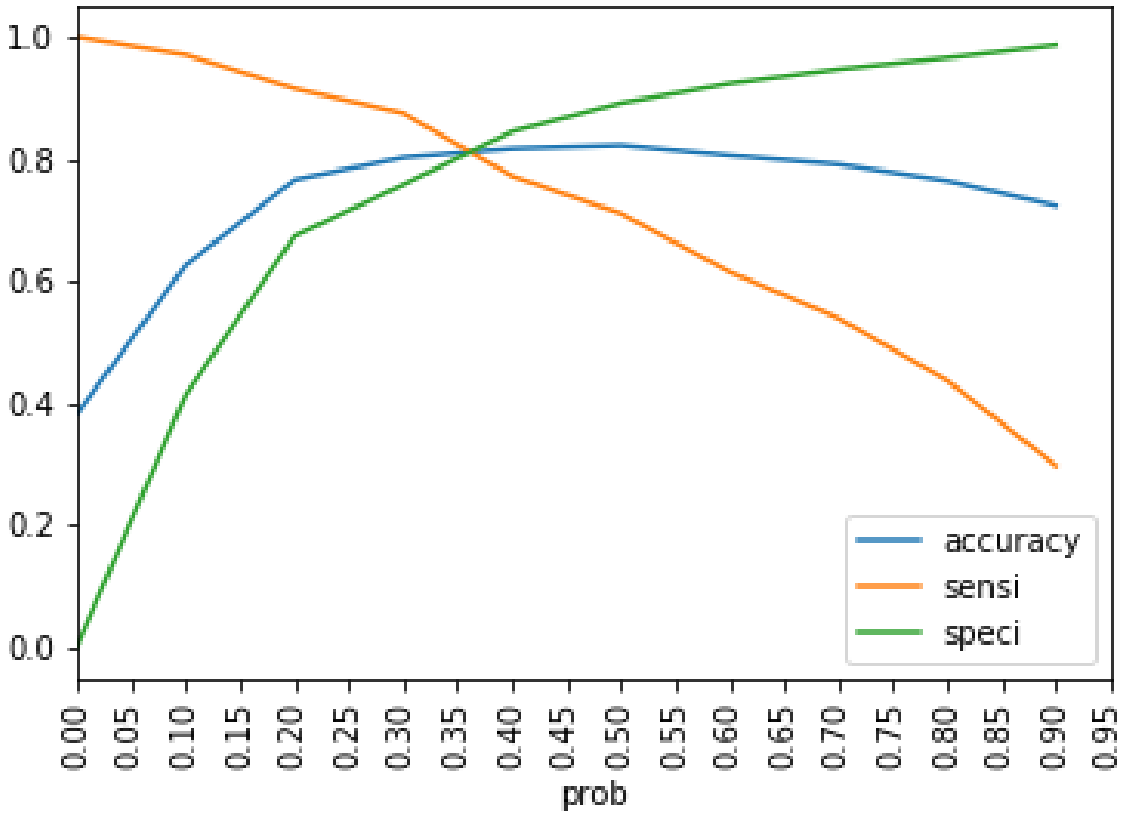
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6454
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2599.9
Date:	Sun, 06 Sep 2020	**Deviance:**	5199.9
Time:	16:11:29	**Pearson chi2:**	7.13e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5620	0.154	-3.655	0.000	-0.863	-0.261
TotalVisits	0.3507	0.053	6.610	0.000	0.247	0.455
Total Time Spent on Website	1.0665	0.040	26.575	0.000	0.988	1.145
Page Views Per Visit	-0.3311	0.060	-5.507	0.000	-0.449	-0.213
A free copy of Mastering The Interview	-0.3626	0.089	-4.059	0.000	-0.538	-0.188
lo_Landing Page Submission	-0.8498	0.133	-6.392	0.000	-1.110	-0.589
lo_Lead Add Form	3.1600	0.203	15.544	0.000	2.762	3.558
ls_Olark Chat	0.9950	0.139	7.180	0.000	0.723	1.267
la_Email Opened	0.6313	0.101	6.260	0.000	0.434	0.829
la_SMS Sent	1.7627	0.104	16.884	0.000	1.558	1.967
sp_unknown	-0.9833	0.126	-7.807	0.000	-1.230	-0.736
ln_Modified	-0.7084	0.087	-8.138	0.000	-0.879	-0.538
co_Not Specified	-1.0392	0.088	-11.807	0.000	-1.212	-0.867
co_Working Professional	2.4786	0.194	12.803	0.000	2.099	2.858

ROC Curve



Optimal Cut off



- Initially we took out cutoff to be 0.5 and we got:

  sensitivity=0.71(train set) specificity=0.89(train set)

- Then in accordance to the above graph we took the optimal cut off probability to be 0.35 and got the following

  sensitivity= 0.807 (train set) specificity=0.812(train set)

  sensitivity= 0.803 (test set) specificity=0.816(test set)

  Hence, we got a balanced sensitivity and specificity.

- The precision and recall for the same cutoff came out to be as follows:

  Precision (train data):0.726 recall (train data):0.807

  Precision (test data):0.740 recall (test data):0.803

Conclusion

It was found that the variables that mattered the most in the potential buyers are (descending order):

1. Lead Origin - lead add form
2. Current Occupation – Working professionals
3. Last activity – SMS sent
4. Total time spent on website
5. Lead Source – Olark chat
6. Last activity – Email opened
7. Total Visits
8. Page views per visit
9. Free copy of mastering the interview
10. Last notable activity – Modified
11. Lead origin landing page submission
12. Specialization - Unknown
13. Current Occupation – Not specified

Recommendations

There are certain traits of customers that make then more likely to be converted to lead. The company should focus more on these customers.

They are as follows:

- As many working professionals as possible should be called.
- All the customers with lead origin as lead add form should be called.
- Customers who spend more time on Website should be called.
- Customers with last activity as SMS sent should be contacted as much as possible.
- Customers as lead source as Olark chat should be called.

Thank you