LEAD SCORING CASE STUDY

Done By :

Anshaj Upadhyay
Twinkle Dalal
Urali Mehta

Business Objective

- The aim is to help X Education to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Methodology

- 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.

- Importing the data and inspecting the data frame
- Data preparation
- Univariate and Bivariate Analysis
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE R-squared VIF and p- values)
- Model Evaluation
- Making predictions on test set

Handling outliers

Before removing outliers

Out[38]:

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	922.750000	3.200000
90%	7.000000	1373.000000	5.000000
95%	10.000000	1557.000000	6.000000
99%	17.000000	1839.000000	9.000000
99.5%	21.000000	1929.445000	11.000000
99.9%	31.854000	2111.927000	14.463500
max	251.000000	2272.000000	55.000000

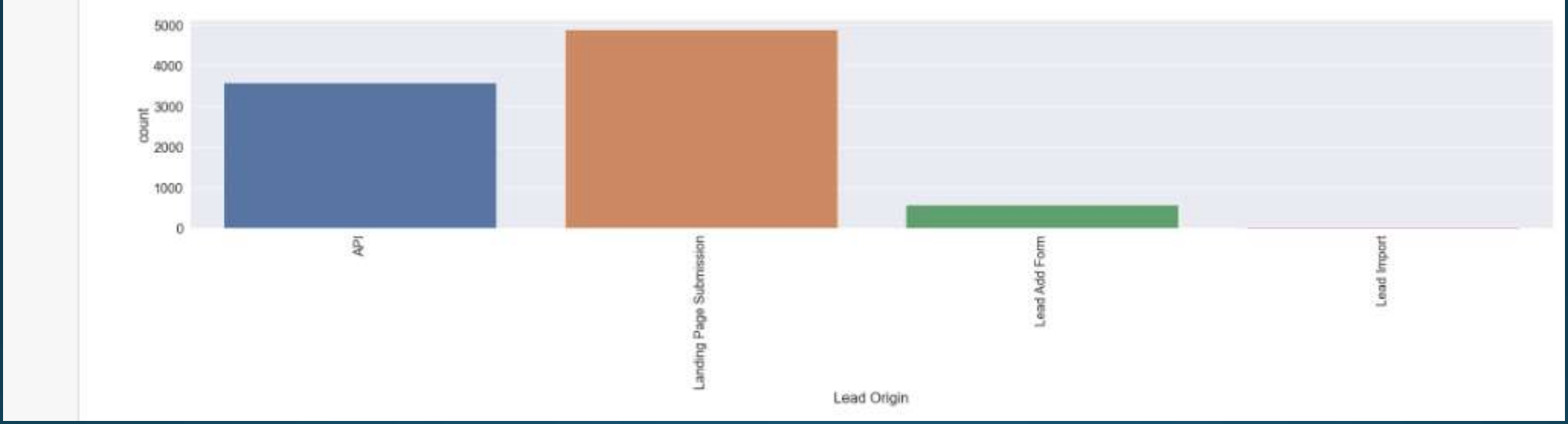
After removing outliers

Out[41]:

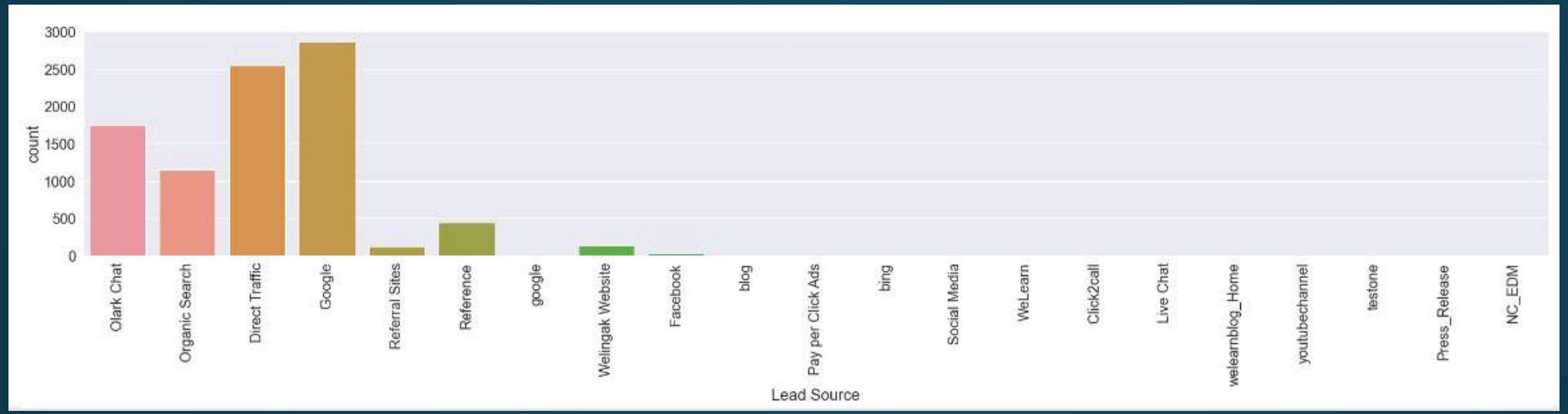
	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9033.000000	9033.000000	9033.000000
mean	3.291487	482.563711	2.351138
std	3.175790	545.377180	2.062207
min	0.000000	0.000000	0.000000
25%	1.000000	10.000000	1.000000
50%	3.000000	245.000000	2.000000
75%	5.000000	922.000000	3.000000
90%	7.000000	1373.000000	5.000000
95%	9.000000	1557.000000	6.000000
99%	15.000000	1839.000000	9.000000
99.5%	17.000000	1930.880000	10.000000
99.9%	20.000000	2111.968000	14.000000
max	21.000000	2272.000000	16.000000

From the above table we can see that there is a huge difference between the max value and 99.5 percentile, and hence we have removed all the values above 99.5 percentile.

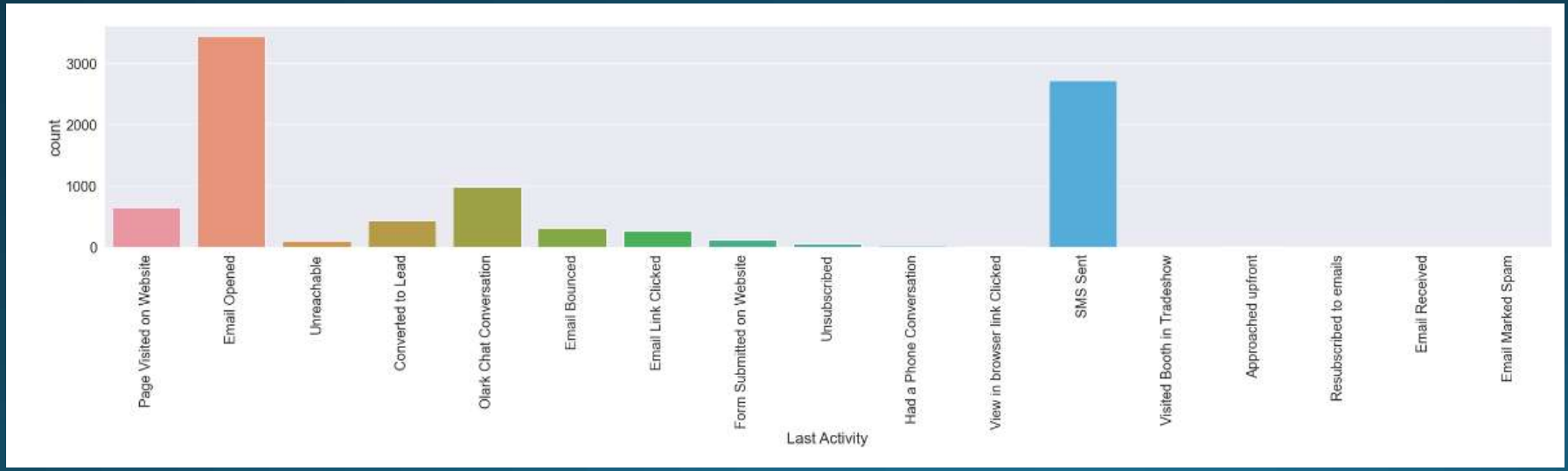
Data Visualisation



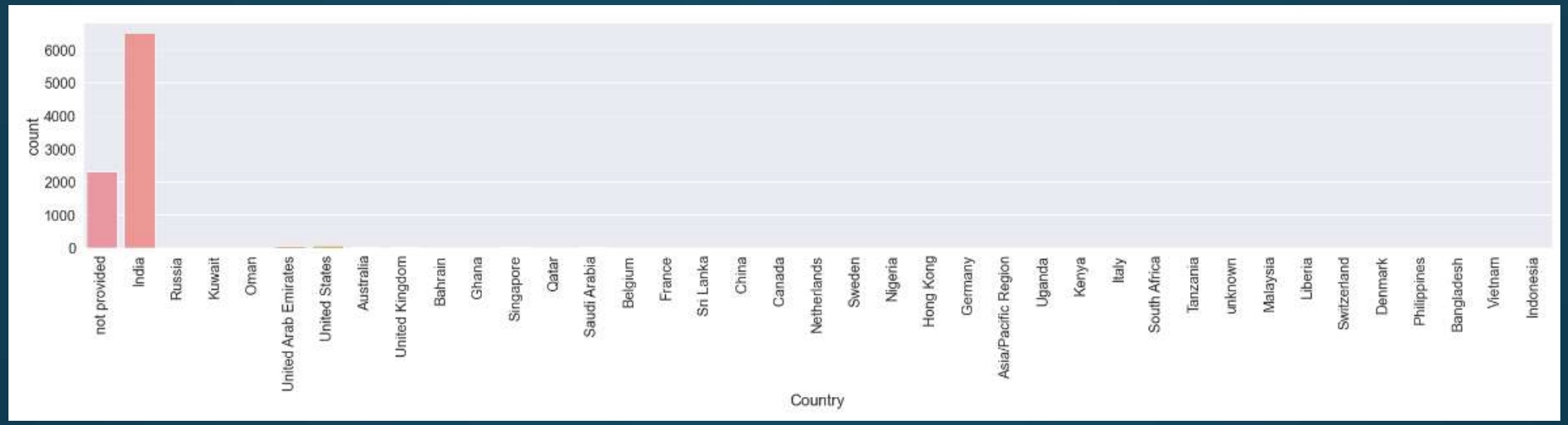
The origin identifier with which the customer was identified to be a lead was mainly "Landing Page Submission" followed by "API".
Try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'



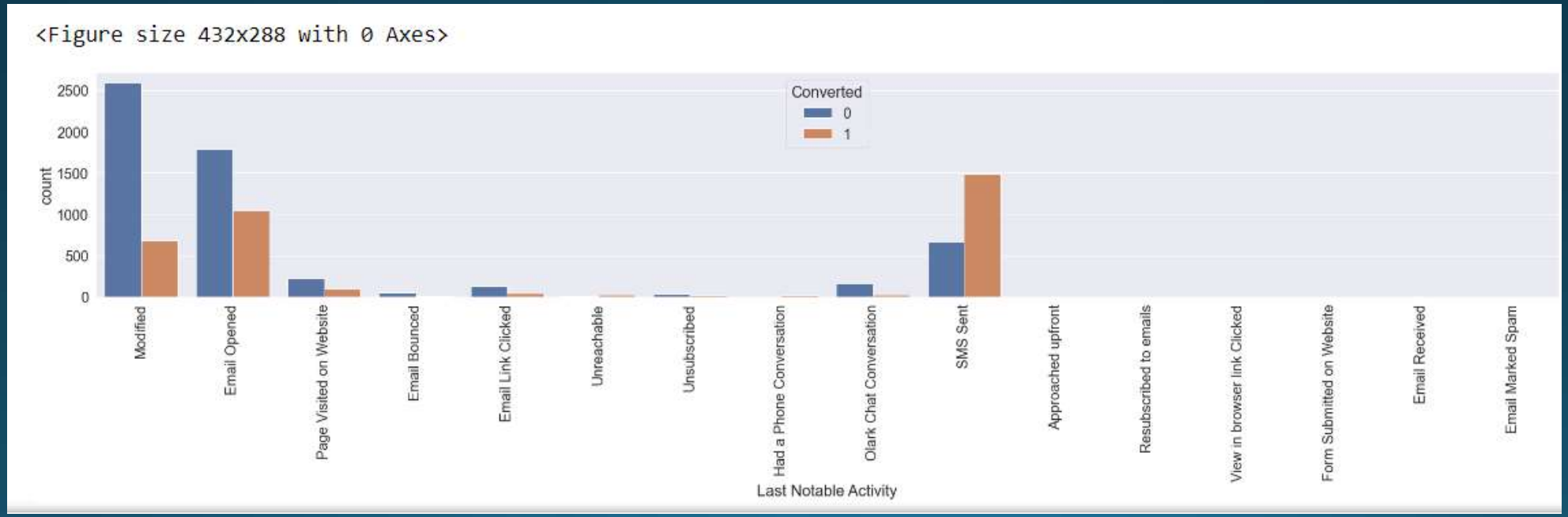
The source of the lead is mainly Google followed by Direct Traffic and Olark Chart .



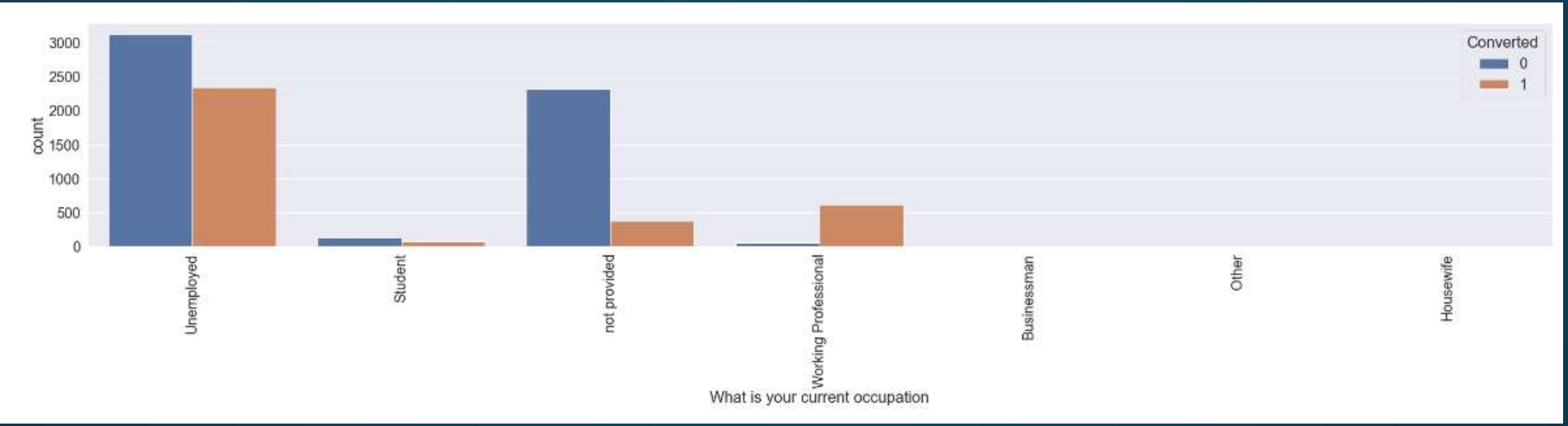
Last Activity performed by users mainly includes “Email Opened” and “SMS Sent” followed by “Page visited on website” and “Olark Chat Conversation”. “SMS Sent” have highly converted %.



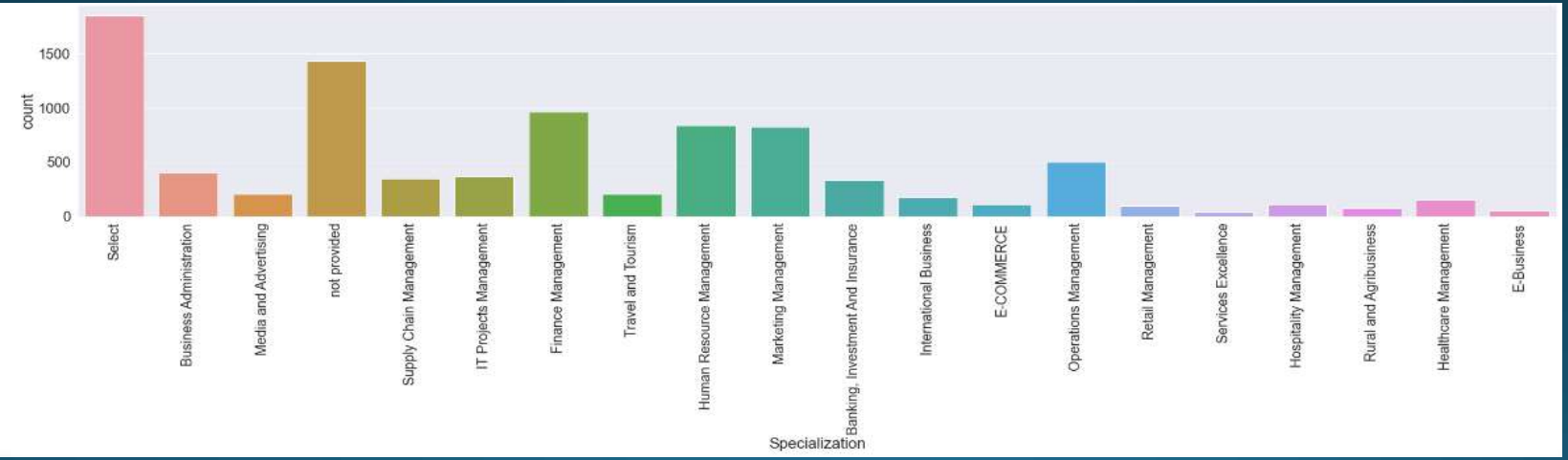
We can see that most of the customers are from India.



Highest conversion rate is for the last notable activity ‘SMS Sent’

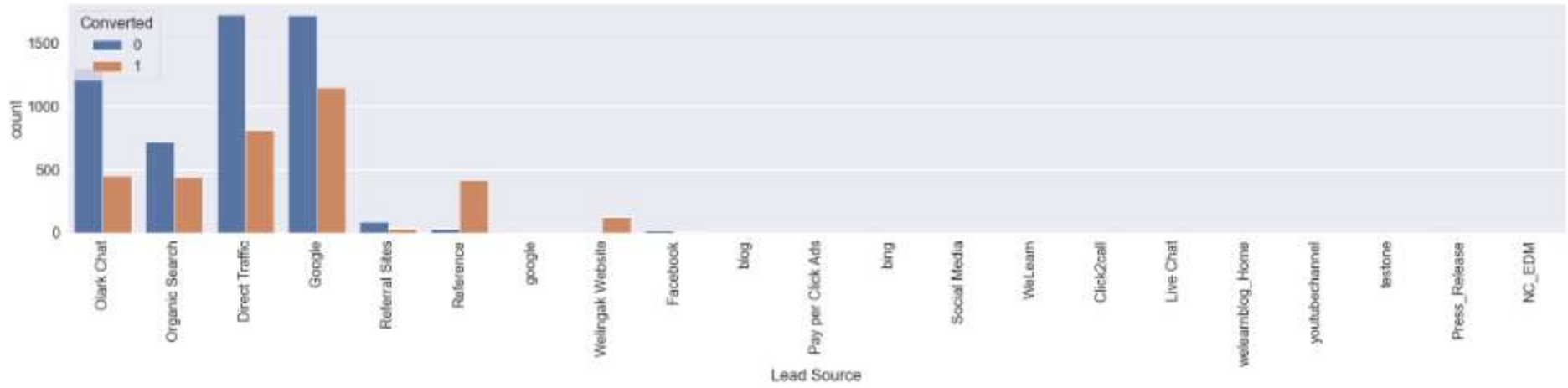


Most of the customers focused are “unemployed”, but “working professional” conversion rate is much higher than “unemployed”



Since most of the customers are “unemployed” they have not selected the “Specialization”. Other Specializations include “Finance”, “HR”, “Marketing” and “Operations” followed by other specializations

<Figure size 432x288 with 0 Axes>



<Figure size 432x288 with 0 Axes>

Customers with reference have high conversion rate

Model Evaluation

Out[66]:

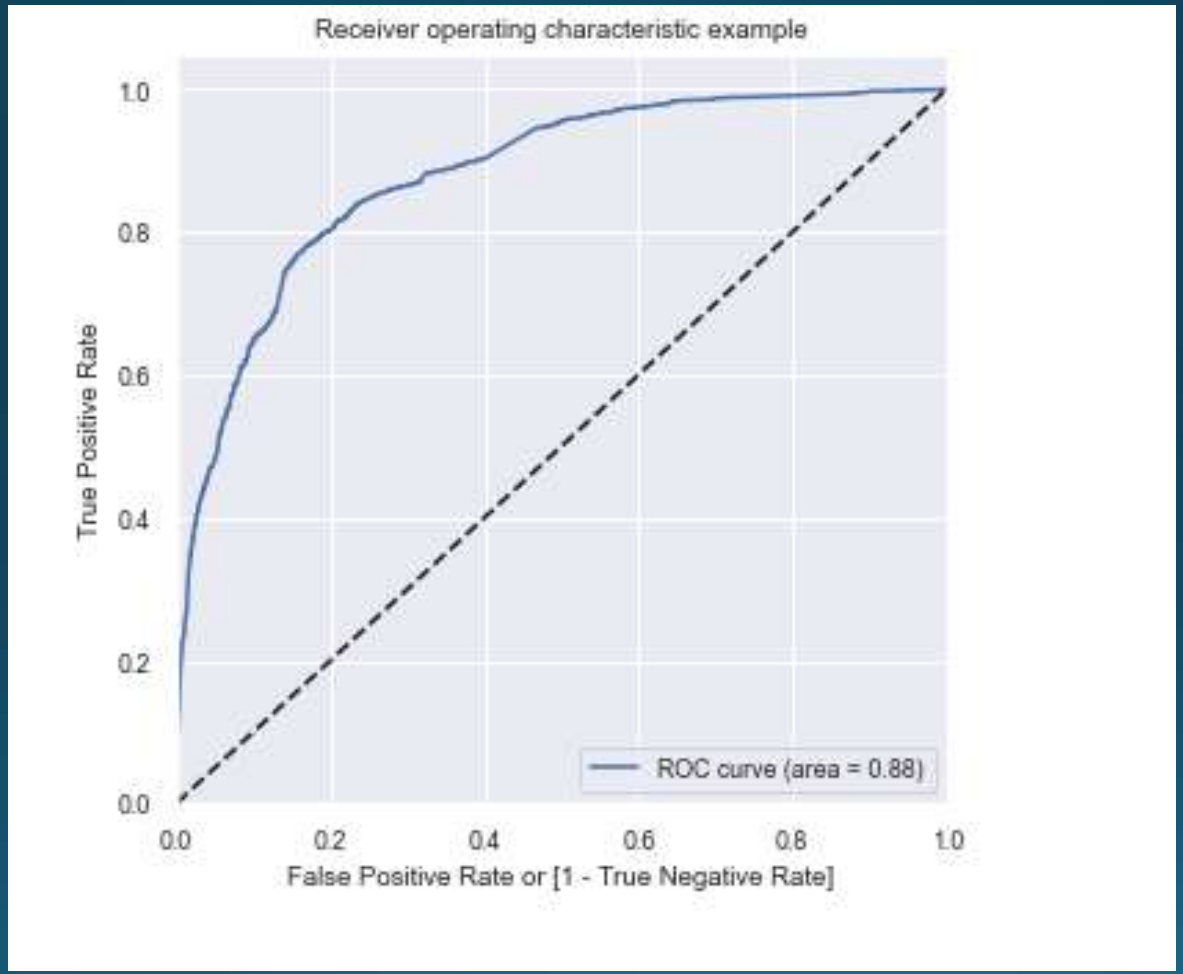
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6323
Model:	GLM	**Df Residuals:**	6309
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2665.1
Date:	Tue, 18 Jul 2023	**Deviance:**	5330.1
Time:	20:16:34	**Pearson chi2:**	6.32e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7665	0.096	-28.853	0.000	-2.954	-2.579
Do Not Email	-1.3114	0.167	-7.839	0.000	-1.639	-0.984
TotalVisits	1.2694	0.309	4.103	0.000	0.663	1.876
Total Time Spent on Website	4.2123	0.155	27.091	0.000	3.908	4.517
Page Views Per Visit	-2.8495	0.388	-7.337	0.000	-3.611	-2.088
Lead Origin_Lead Add Form	3.6141	0.231	15.641	0.000	3.161	4.067
Lead Source_Direct Traffic	-0.4746	0.078	-6.105	0.000	-0.627	-0.322
Last Activity_SMS Sent	1.4407	0.073	19.627	0.000	1.297	1.585
What is your current occupation_Other	2.1308	0.655	3.255	0.001	0.848	3.414
What is your current occupation_Student	1.3778	0.227	6.075	0.000	0.933	1.822
What is your current occupation_Unemployed	1.1885	0.086	13.819	0.000	1.020	1.357
What is your current occupation_Working Professional	3.7260	0.200	18.660	0.000	3.335	4.117
Last Notable Activity_Had a Phone Conversation	3.5396	1.101	3.216	0.001	1.382	5.697
Last Notable Activity_Unreachable	1.9586	0.546	3.587	0.000	0.889	3.029

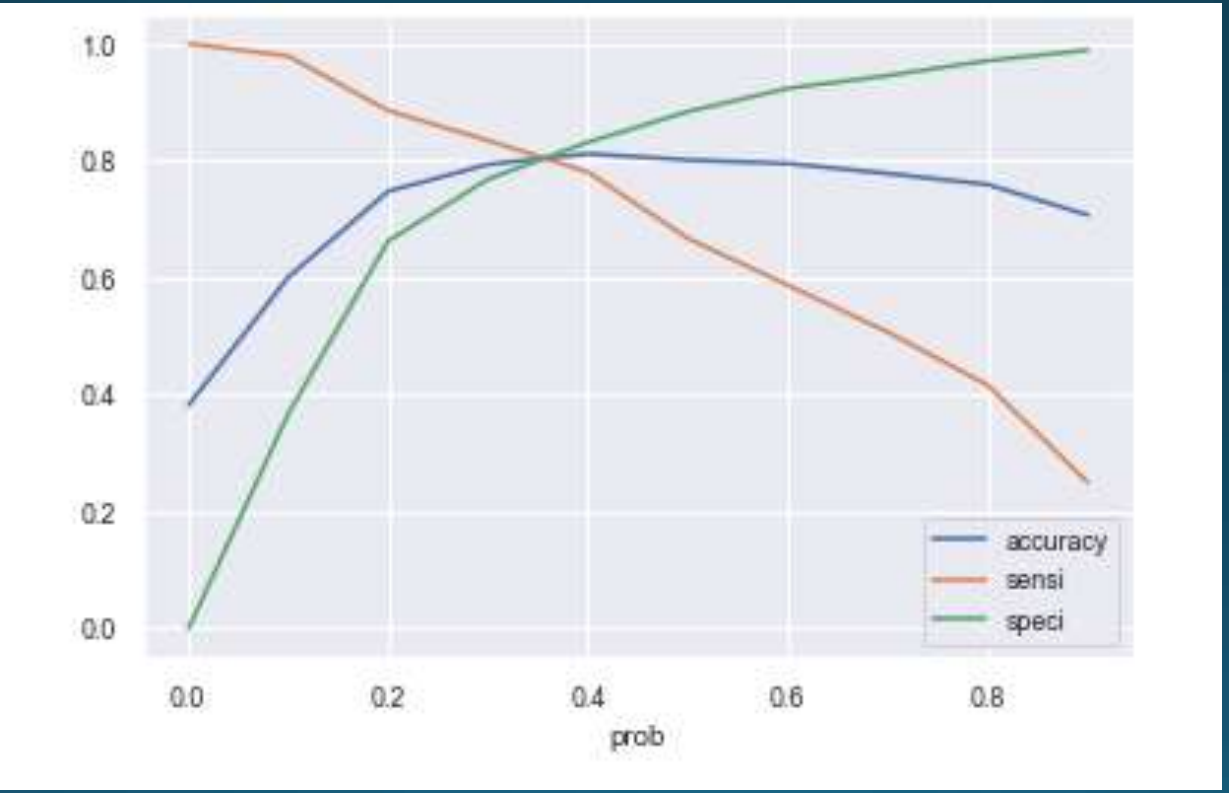
All the p values are less than 0.005

ROC Curve



Area under curve = 0.88

Finding optimal threshold



Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values Optimal cutoff = 0.35

Final Results

Data	Train Set	Test Set
Accuracy	0.801	0.806
Sensitivity	0.80	0.79
Specificity	0.80	0.81

Inferences

- The origin identifier with which the customer was identified to be a lead was mainly “Landing Page Submission” and “API”. “Lead Add form” has a high conversion rate.
- The source of the lead is mainly Google, Direct Traffic, Olark Chart and Organic Search. People coming from Reference have a very high conversion rate.
- Last Activity performed by users mainly includes “Email Opened” and “SMS Sent” followed by “Page visited on website” and “Olark Chat Conversation”. “SMS Sent” have highly converted %.
- Most of the customers are from India.
- Most of the customers focused are “unemployed”, but “working professional” conversion rate is much higher than “unemployed”

- Since most of the customers are “unemployed” they have not selected the “Specialization”. Other Specializations include “Finance”, “HR”, “Marketing” and “Operations” followed by other specializations.

- “Total Time spent on website” has a good correlation (0.36) with “Converted”, which means people who spend higher time on websites have more chances of conversion. “TimeVisits” and “Page Views Per Visit” have high correlation (0.51), which is understandable.

- Top 3 features which contributes to decision are as following -
  - 'Total Time Spent on Website'
  - 'Lead Origin_Lead Add Form'
  - 'What is your current occupation_Working Professional'

Recommendations

- By referring to the data visualizations, the focus should be more on working professionals as the conversion rate is high.
- People who come with reference, have high conversion rate and should be more focused on.
- Website can be made from interactive with some gamification , so we can have potential leads spending more time on their website as there is a high correlation between Time Spent and Conversion rate (freemium course can also be launched)
- We can launch more Google Ads as the source of lead is mainly Google followed by Direct Traffic.