Lead Scoring Case Study

Naveen Mani – Janarthanan Mani

Problem Statement

- 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%.**

Goals

- 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.

- 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.

- 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 you will need to handle these as well.

Process Skeleton

Data Sourcing, Preparation & EDA

- Reading the data from source
- Convert data into clean format suitable for analysis
- Impute Null Values
- Outlier Treatment
- Normalizing and Creating Dummy variables on category variable
- Exploratory Data Analysis

Feature Scaling & Train-Test Split

- Feature Scaling of Numeric data
- Splitting data into train and test set.

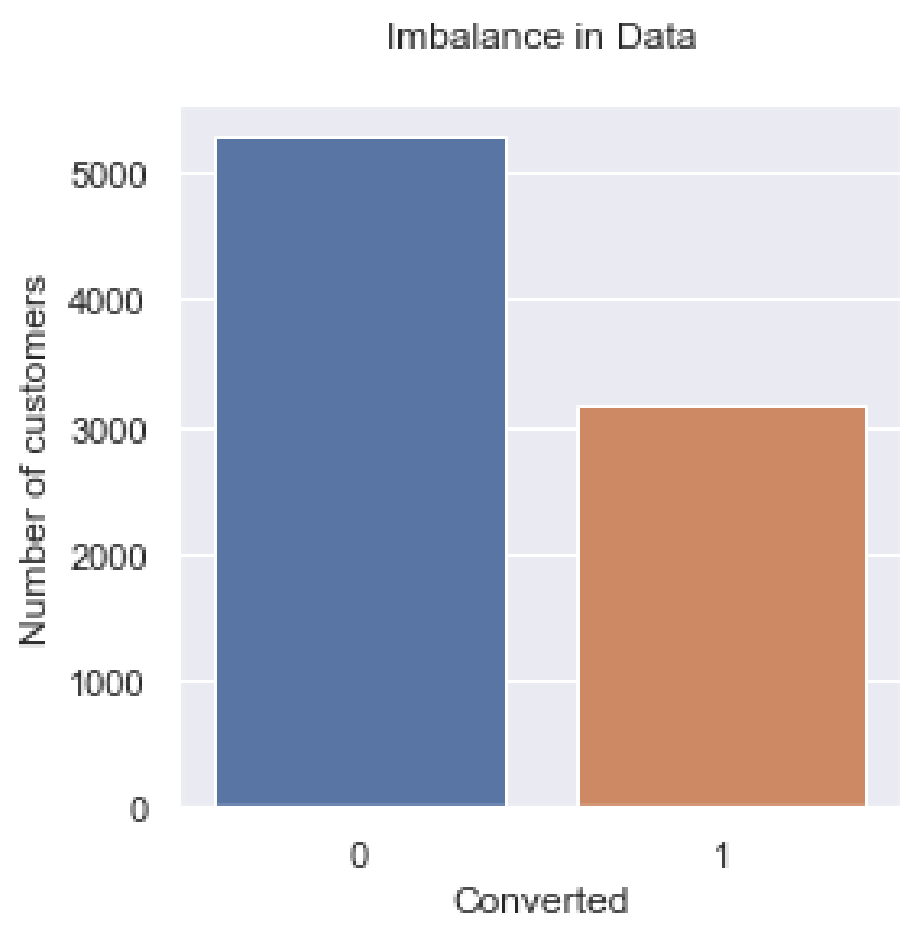
Model Building

- Multicollinearity Analysis.
- Feature Selection using RFE
- Calculate VIF score and dropping variables with high P-Values
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result and Observations

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics
- Variant analysis on category variables providing variables impacting on the lead conversion rate.

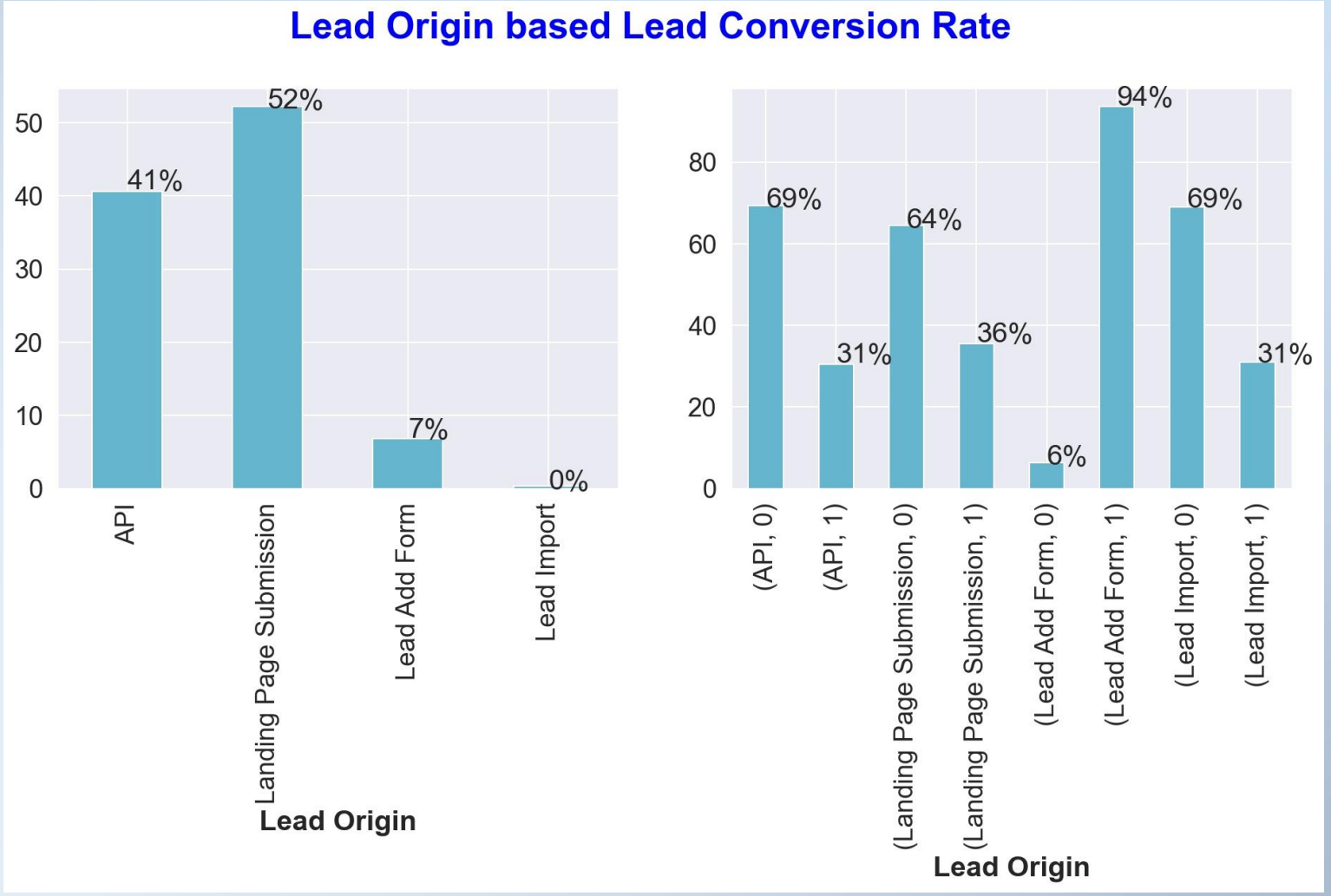
Data Imbalance Check



- **Spread of data between Converted Variable:**
- Customers who are converted to leads:= 3165
- Customers who are not converted to leads:= 5280
- Percentage of customers who are converted to leads:= **37.5**
- Percentage of customers who are not converted to leads:= **62.5**

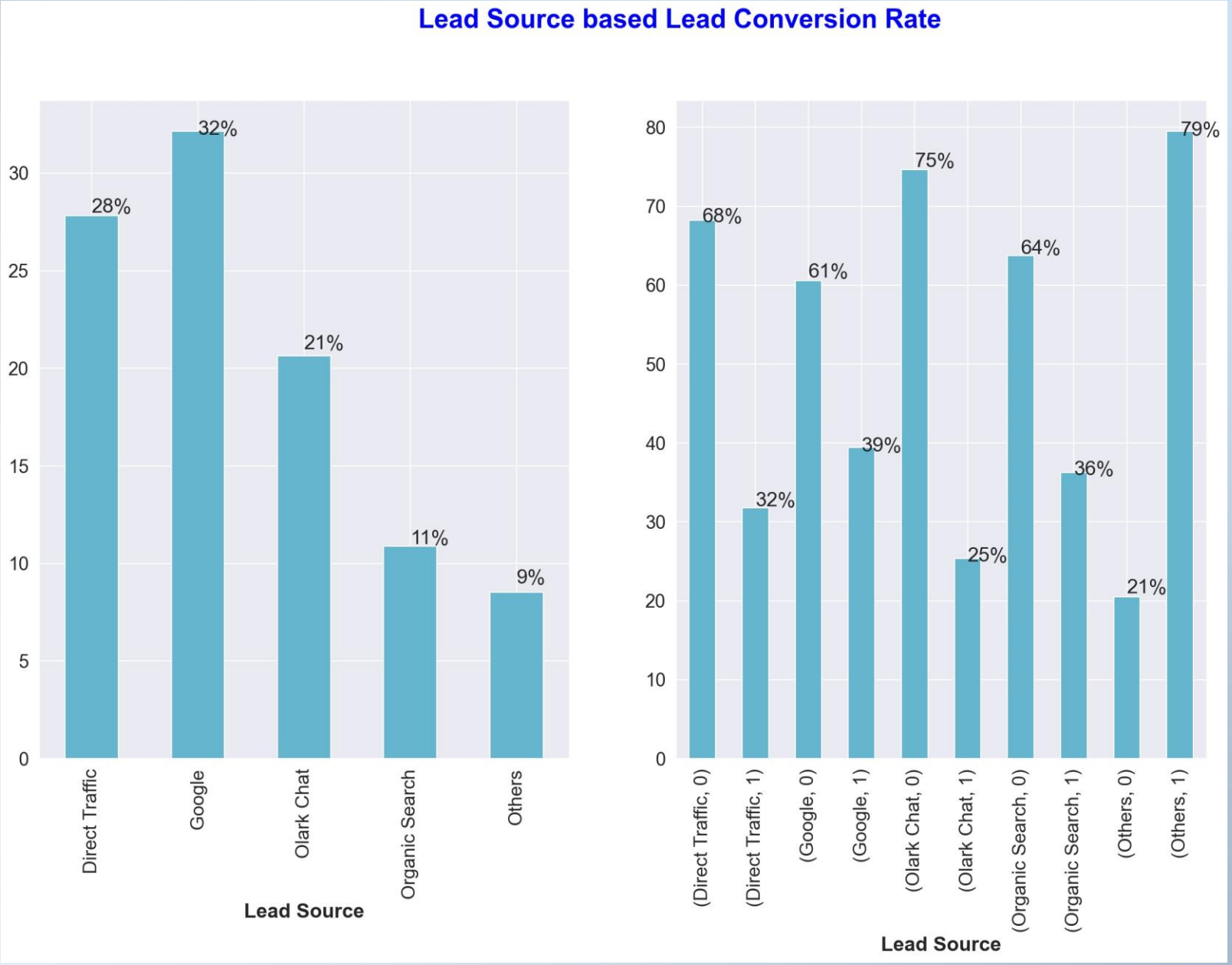
Variant Analysis – Lead Origin

- Lead Origin from "Landing Page submission" has high leads (52%) but with 2nd most conversion rate (36%).
- However, "Lead Add Form" tops the chart with higher conversion rate of 94% of its total leads get converted.



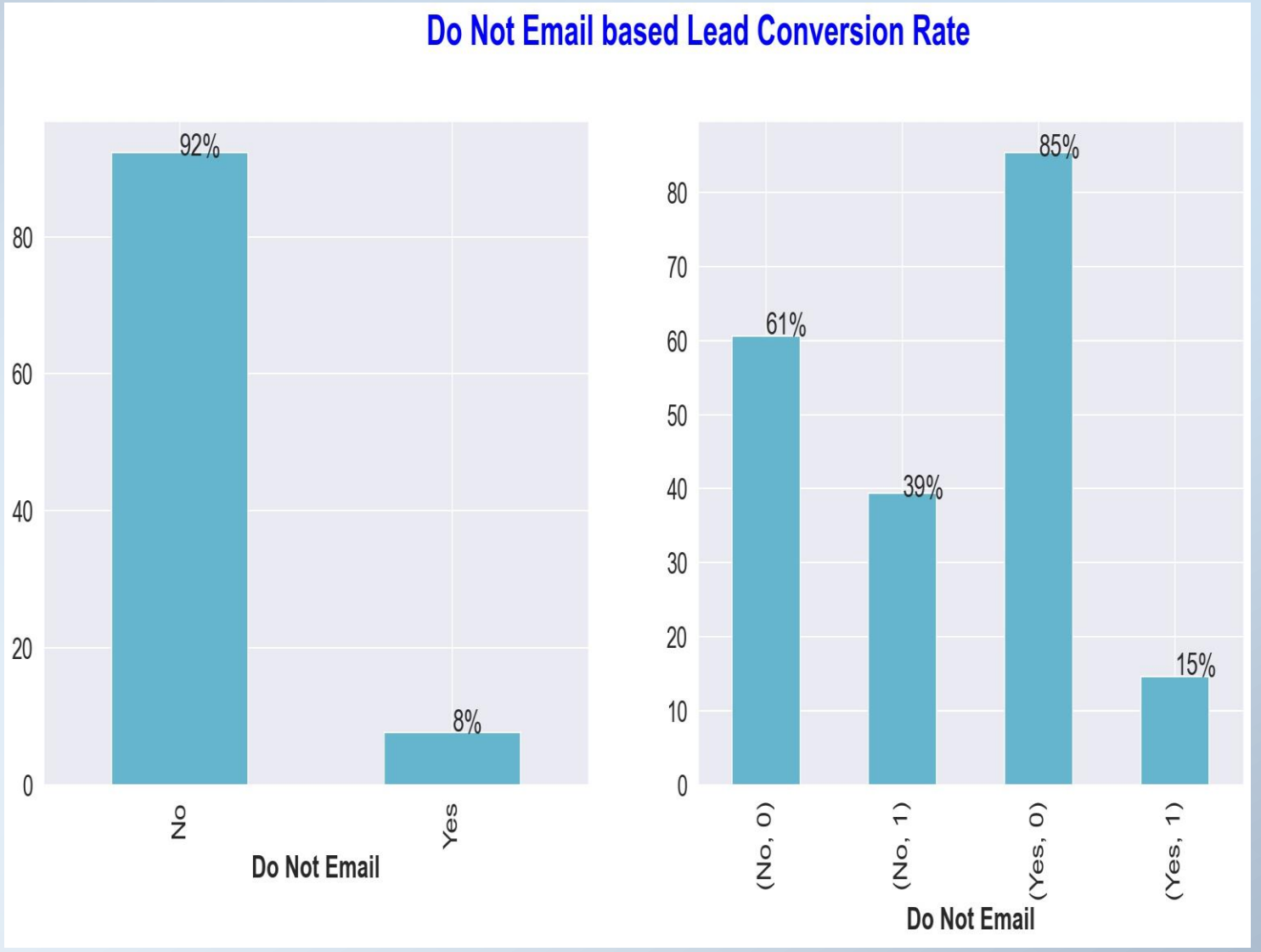
Variant Analysis – Lead Source

- 'Google' generates higher of number of leads (32%) for the company in that 39% gets converted.
- Leads which company gets from other unnamed Categories have higher conversion rate of 79%. So, people from other channels are desperate in taking online course comparing it with popular sources.

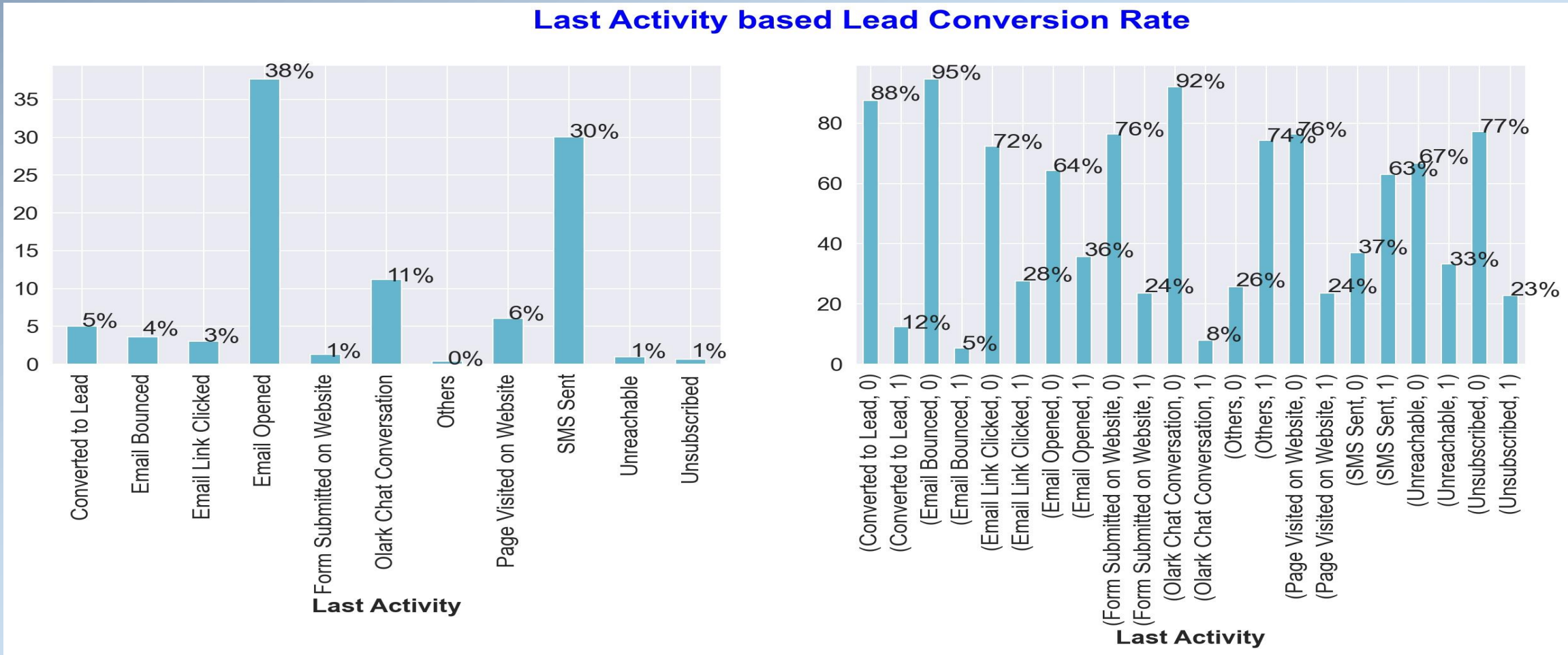


Variant Analysis – Do Not Email

- People who are ok to receive mail are the 92% of the total leads and on which 39% of leads gets converted.



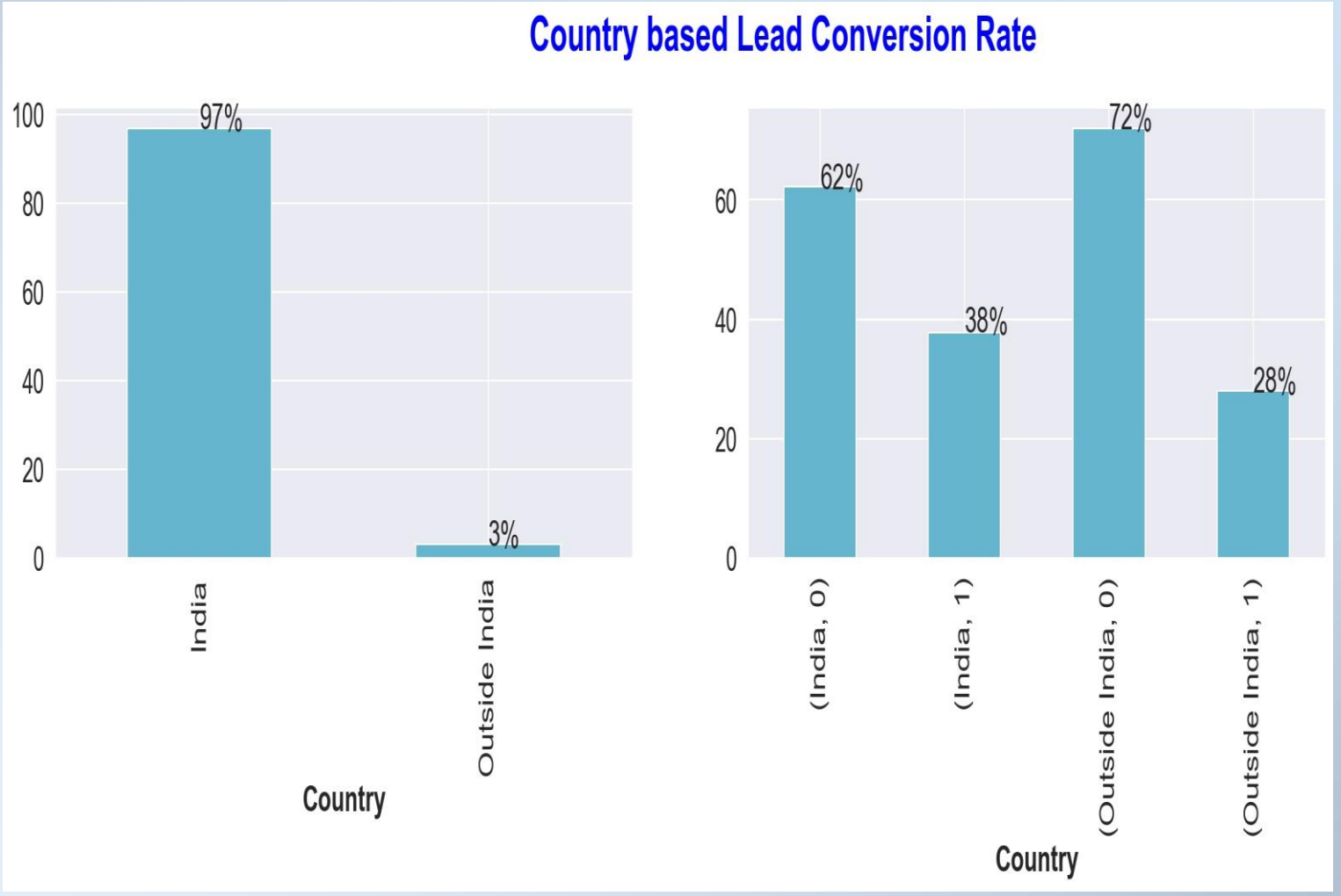
Variant Analysis – Last Activity

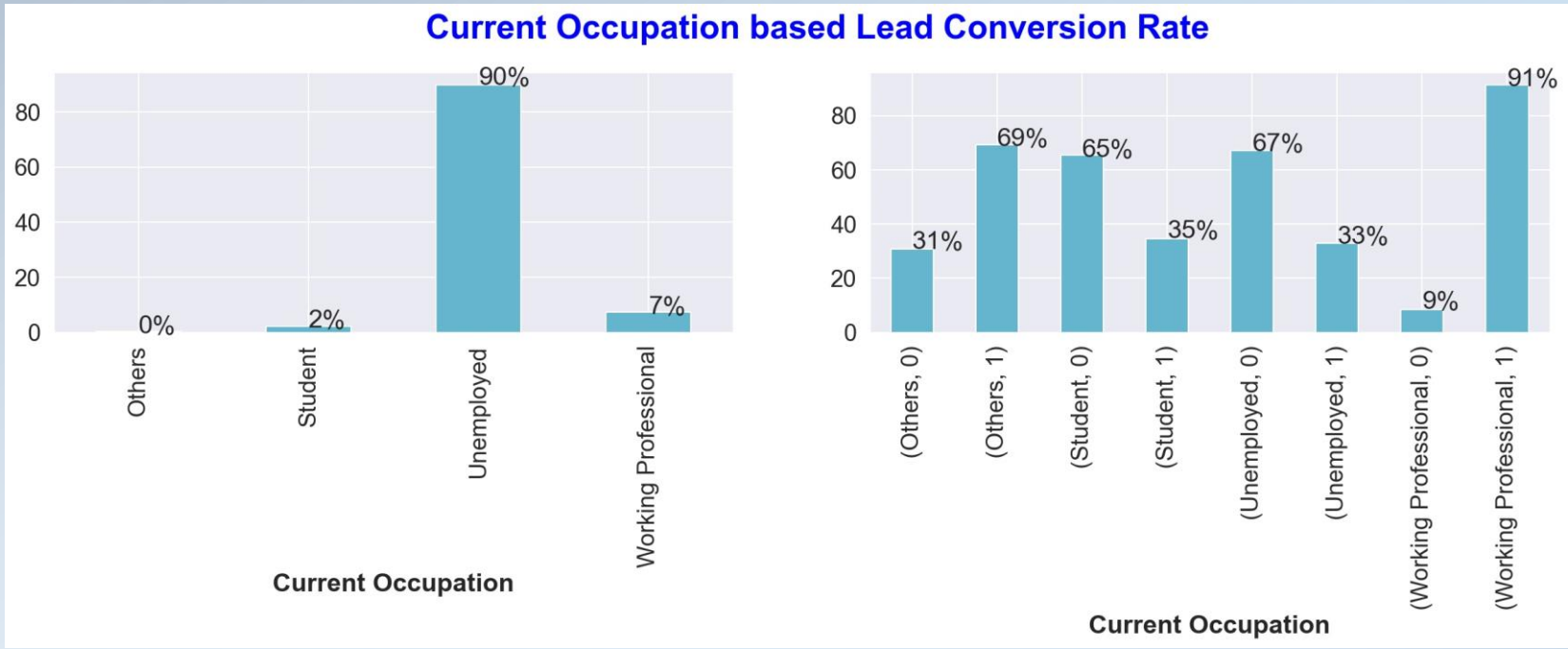


- **Last activity like "1. Email Opened (38%)", "2. SMS Sent (30%)" and "3. Olark Chat (11%)" are the top 3 activities through which most of leads are generated.**
- **However, the lead conversion rate follows the order "1. SMS Sent (63%)" , "2. Email Opened (36%)" and "3. Olark Chat (8%)"**

Variant Analysis – Country

- Indian resident has more probability to be converted to lead with 38%.

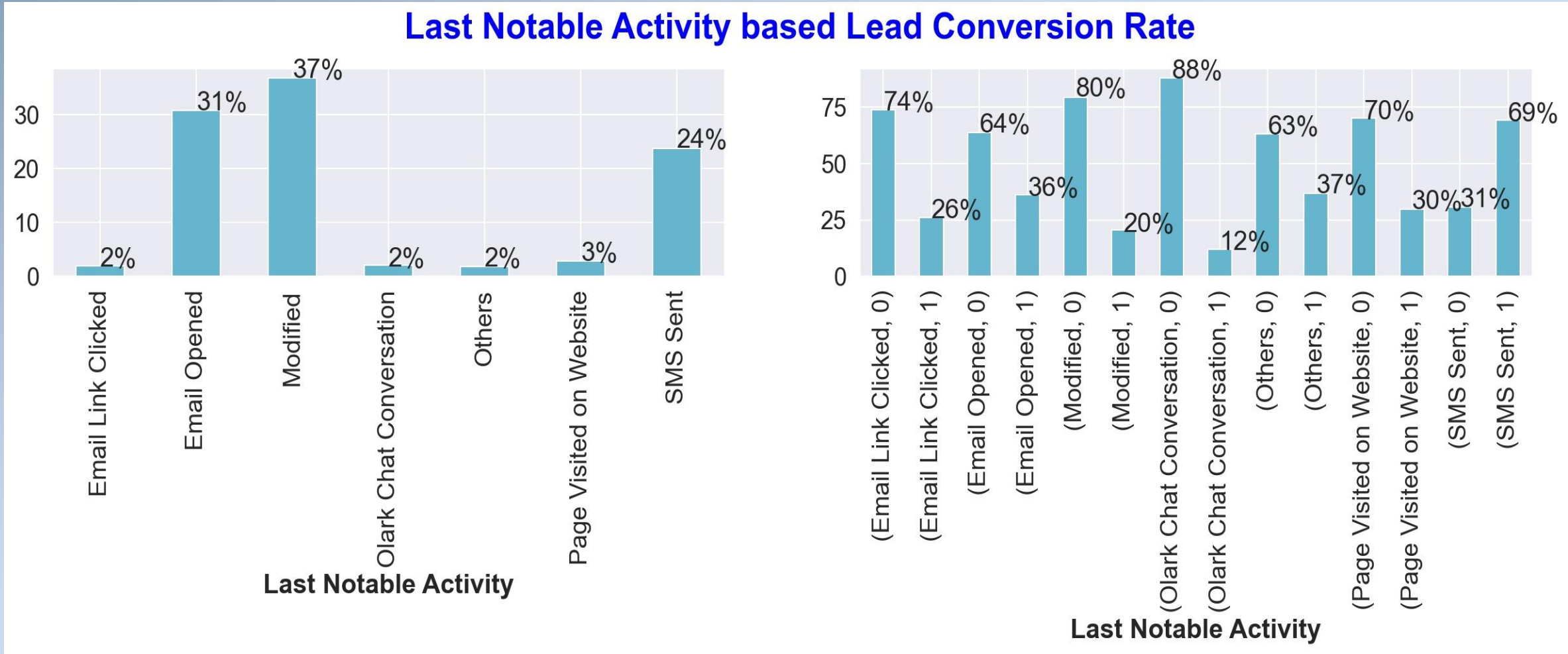




Variant Analysis – Occupation

- 90% leads are generated by Unemployed however they have less conversion rate of 33%. So, education fees might be the issue for the unemployed. Feasible loan options and placement after completion of course will encourage more unemployed leads to get converted.
- Working Professionals are the low lead generator however most of their leads (91%) get converted. This is obvious that working professional looks for course only when they are in need, and they are also capable of affording it.

Variant Analysis – Last Notable Activity



- **Last Notable activity like "1. Modified (37%)", "2. Email Opened (31%)" and "3. SMS Sent (24%)" are the top 3 activities through which most of leads are generated.**
- **However, the lead conversion rate follows the order "1. SMS Sent (69%)" , "2. Email Opened (36%)" and "3. Modified (20%)"**

Model Evaluation

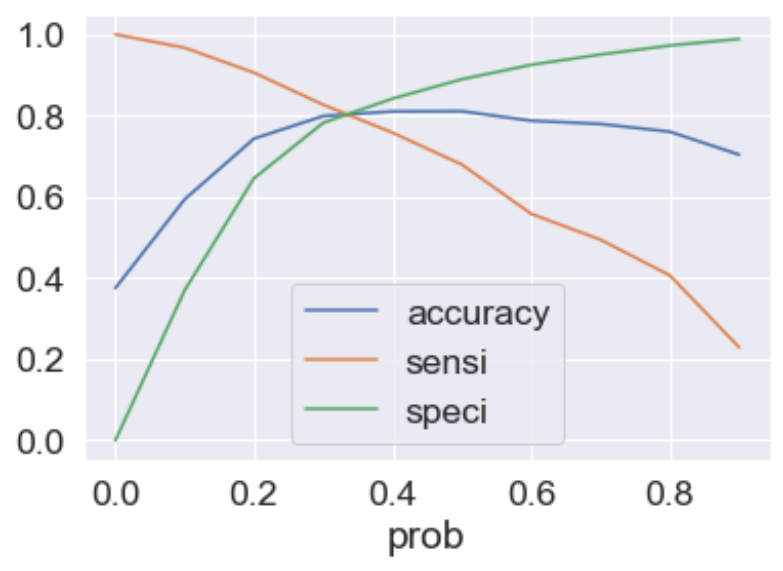
Final Feature variables and its VIF Score

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7713	0.118	-15.034	0.000	-2.002	-1.540
Do Not Email	-1.4064	0.190	-7.409	0.000	-1.778	-1.034
Total Time Spent on Website	1.0777	0.041	26.215	0.000	0.997	1.158
Lead Origin_Landing Page Submission	-0.2022	0.095	-2.139	0.032	-0.388	-0.017
Lead Origin_Lead Add Form	4.1477	0.232	17.901	0.000	3.694	4.602
Lead Origin_Lead Import	1.2727	0.489	2.602	0.009	0.314	2.231
Lead Source_Google	0.2283	0.085	2.697	0.007	0.062	0.394
Lead Source_Olark Chat	1.1572	0.135	8.599	0.000	0.893	1.421
Last Activity_Email Opened	0.5169	0.094	5.503	0.000	0.333	0.701
Last Activity_Olark Chat Conversation	-1.2298	0.183	-6.711	0.000	-1.589	-0.871
Last Activity_Others	1.9523	0.695	2.808	0.005	0.590	3.315
What is your current occupation_Working Professional	2.5549	0.185	13.825	0.000	2.193	2.917
Last Notable Activity_Others	1.5878	0.297	5.351	0.000	1.006	2.169
Last Notable Activity_SMS Sent	1.9247	0.103	18.680	0.000	1.723	2.127

	Features	VIF
6	Lead Source_Olark Chat	2.14
2	Lead Origin_Landing Page Submission	2.12
7	Last Activity_Email Opened	2.09
12	Last Notable Activity_SMS Sent	1.73
8	Last Activity_Olark Chat Conversation	1.62
5	Lead Source_Google	1.51
3	Lead Origin_Lead Add Form	1.37
1	Total Time Spent on Website	1.33
0	Do Not Email	1.26
11	Last Notable Activity_Others	1.20
10	What is your current occupation_Working Profes...	1.17
9	Last Activity_Others	1.06
4	Lead Origin_Lead Import	1.02

Model Evaluation

Sensitivity and Specificity on Training Data Set



We could see optimal cut off probability can be taken as 0.35

Confusion Matrix

3000	694
452	1765

- Accuracy - 81%
- Sensitivity - 80%
- Specificity - 81%
- Positive Predictive Value - 72%
- Negative Predictive Value – 87%

Model Evaluation

Sensitivity and Specificity on Test Data Set

Actual Converted Vs Final Predicted

82%

Confusion Matrix

1292	294
172	776

- Accuracy - 82%
- Sensitivity - 82%
- Specificity - 81%

Model Evaluation

Lead Score Generation

Final Lead Score is generated based on the converted Probability predicted value

```
# Now let us calculate the lead score

y_pred_final['lead_score'] = y_pred_final.Converted_Prob.map(lambda x: round(x*100))

#displaying the 10 rows with the calculated Lead_score column highlighted
def highlight_cols(s):
    color = 'yellow'
    return 'background-color: %s' % color

y_pred_final.head(20).style.applymap(highlight_cols, subset=pd.IndexSlice[:, ['lead_score']])
```

	LeadId	Converted	Converted_Prob	final_predicted	lead_score
0	1926	1	0.821258	1	82
1	5654	0	0.444768	1	44
2	1209	0	0.640751	1	64
3	5234	0	0.175498	0	18
4	8497	0	0.122355	0	12
5	1918	0	0.077377	0	8

Observations / Recommendations

- While we have checked both Sensitivity-Specificity, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 81% and 81% which are approximately closer to the respective values calculated in training set.
- Hence overall this model seems to be good with lead convertion around 81% required by the CEO

Based Model Coefficient suggestions are as follows:

- Company can find the potential converting leads in case of lead origin is from "Lead Add Form"
- Company can find potential converting leads who are working potentials.
- People with notable activity as "SMS Sent" are having higher chances for converting.
- Following Last Activity like : Had a Phone Conversation,Approached-upfront,View-in browser link Clicked,Email Received,Resubscribed to emails,Visited Booth in Tradeshow has higher chance of convertion.
- The following variables has negative effect on lead convertion: DO Not Email, Last Activity Olark Chat Conversation, lead origin from Landing Page Submission

Thank You

naveen.ora9@gmail.com
vellorejana@gmail.com