LEAD SCORING CASE STUDY

Proposed by
Abinesh, Binod, Chaithanya

PROBLEM STATEMENT

An education company named x education sells online courses to industry professionals. The company market its course on several websites and search engines like google.

Company wants to select most promising leads that can be converted to paying customers.

Though company generates a lot of leads, only a few get converted.

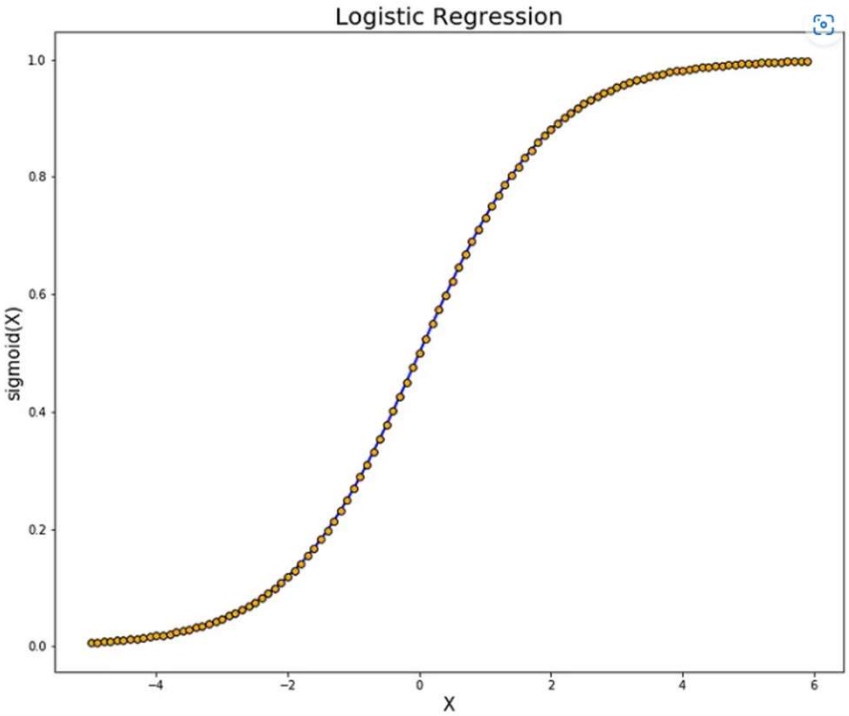
Company had 30% conversion rate by approaching those leads which are found to be having interest in taking the course.

BUSINESS GOAL

1. Build a logistic regression model to assign a lead score between 0 to 100 to each of the leads, which can used by the company to target potential leads.

2. Higher lead score would mean “HOT LEAD” is having higher chance of getting convert.

3. The model to be built in lead conversion rate around 80% or more.



ANALYSIS APPROACH

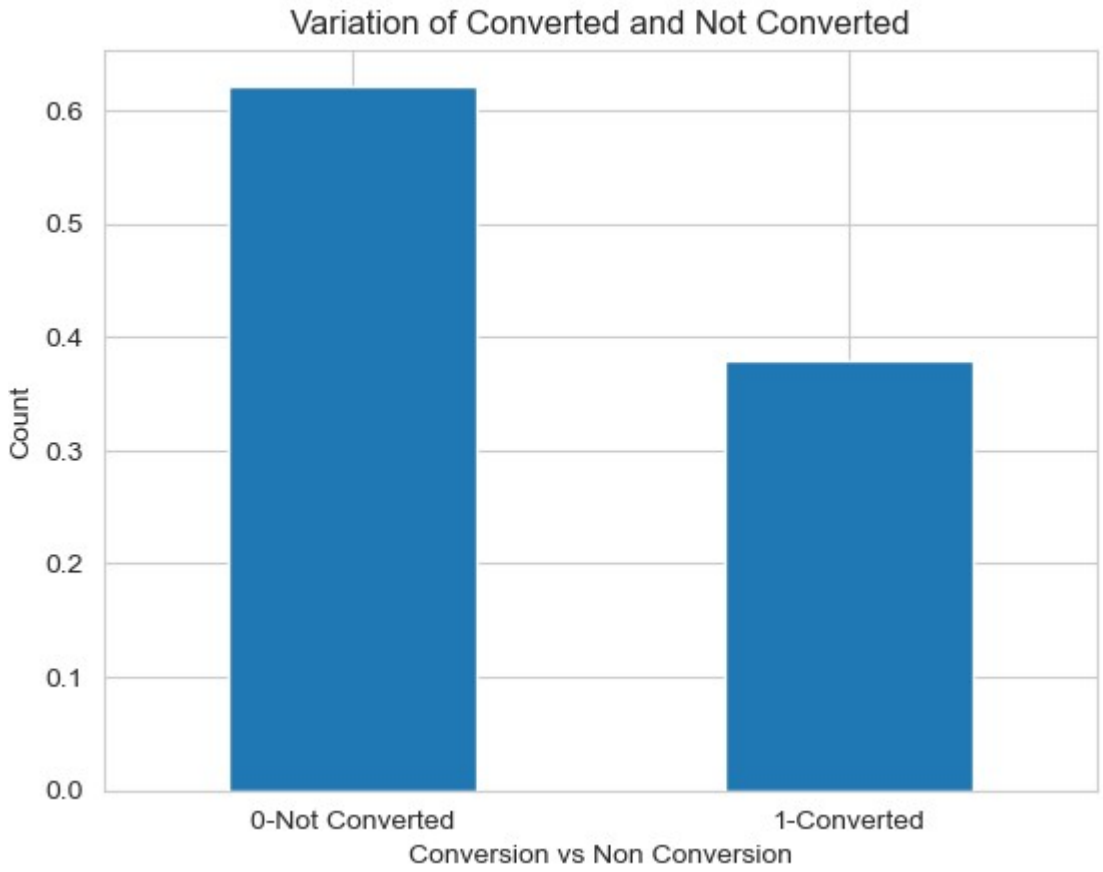
- Receiving the relevant data for analysis.
- Clean the data and setup for further analysis.
- EDA for relationship of various attributes within the data.
- Separating out train and test set if separate test set is not available.
- Scaling the features.
- Build the model by referring P-value for significance of independent variable on the dependent variable and also VIF value.
- Assigning the lead score to each leads.
- Evaluation of model using different metrics.
- Evaluate the model on test set like we did for train set.
- Calculate lead score

ASSUMPTIONS TAKEN FOR THE ANALYSIS

- Dropped the column with more than 40% missing values.
- Dropped the duplicate rows from analysis.
- Few columns have ‘select ’ as values. Replacing those values with null values.
- In case of missing values less than 40% , missing values are replaced with median or mode or highest rank values.
- for numerical values, outlier values have been replaced by maximum quartiles range.
- Dropped the columns with highly skewed values.

DATA IMBALANCE

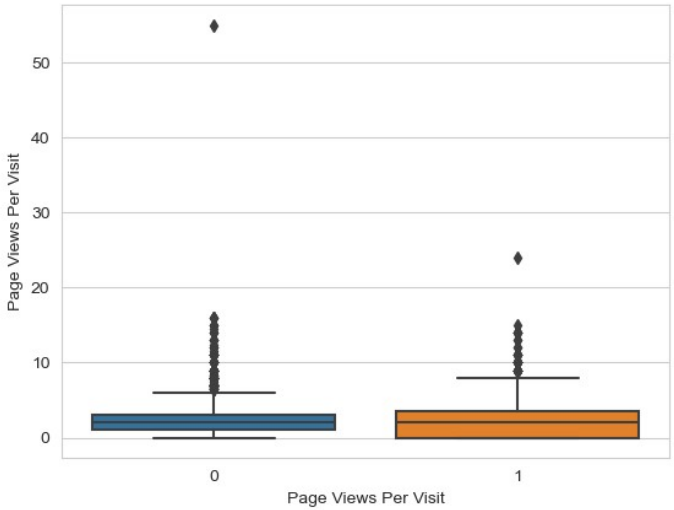
- For data imbalance, ratio of number of occurrences of converted to number of occurrences of not converted.
- Converted values : 37%
- Not Converted values : 62%
- The resulting ratio represents the imbalance between the two classes, indicating whether one class is significantly more prevalent than other.

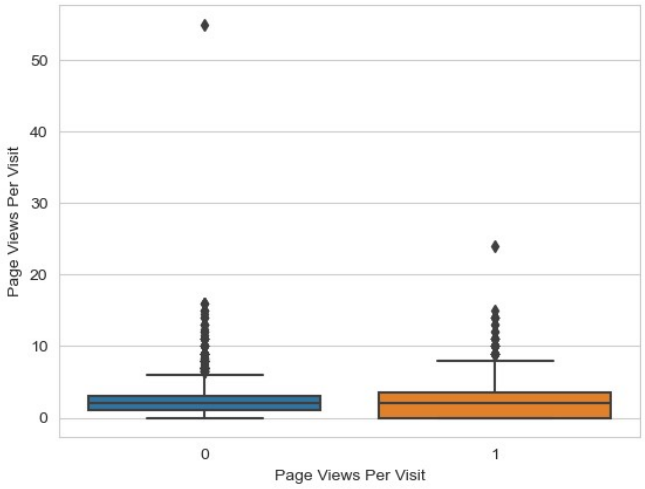


OUTLIER ANALYSIS

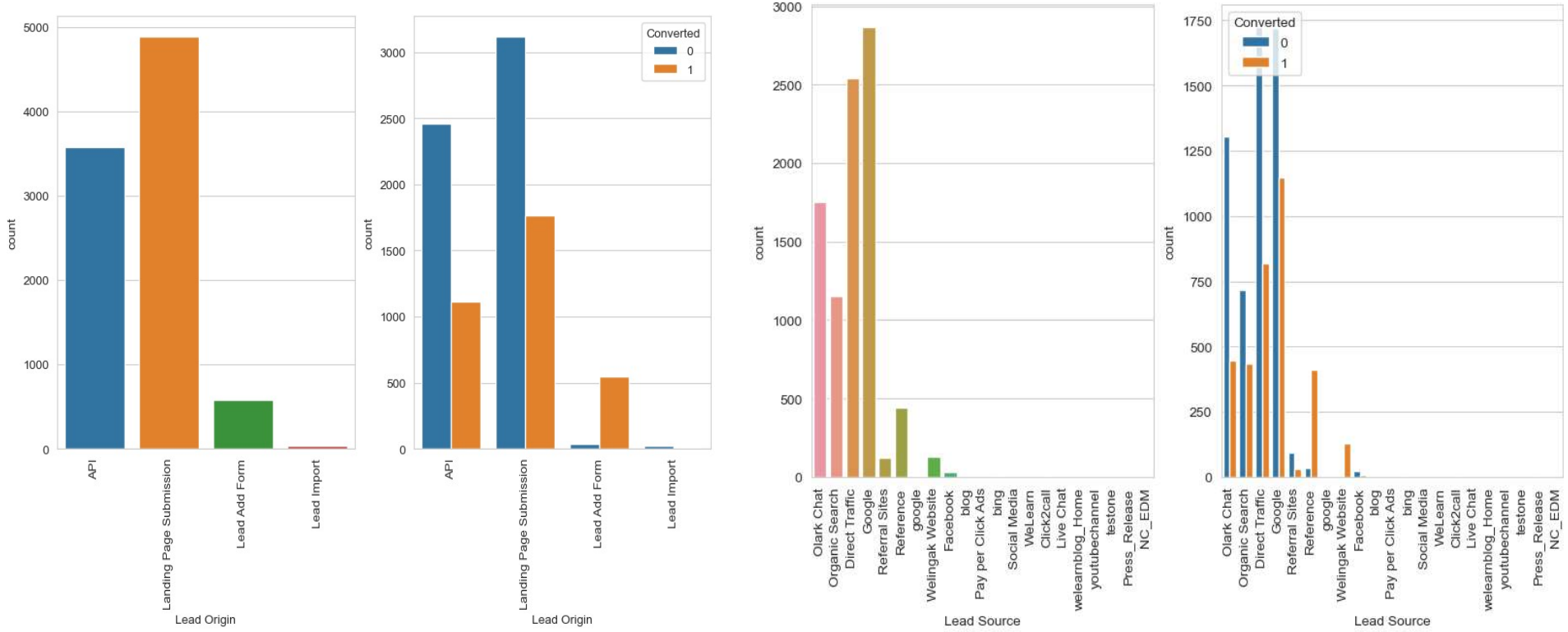
- Considering the presence of outliers in the “Total visit”, “page view per visit”, it is advisable to employ a technique such as capping and flooring to achieve the desired variance in these values.

- By setting appropriate upper and lower thresholds, extreme values can be truncated, ensuring that the data remains within a reasonable range and allowing for more meaningful analysis and interpretation.

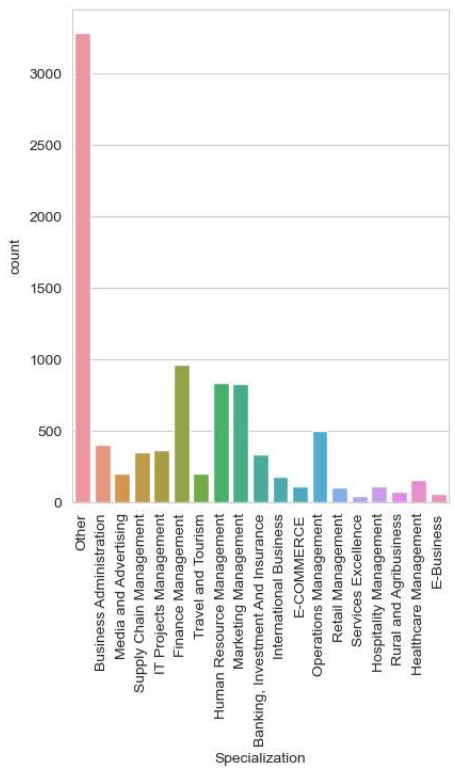


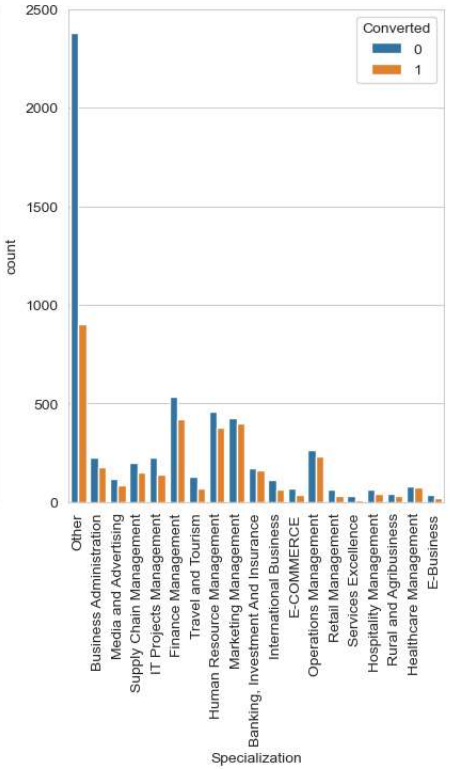


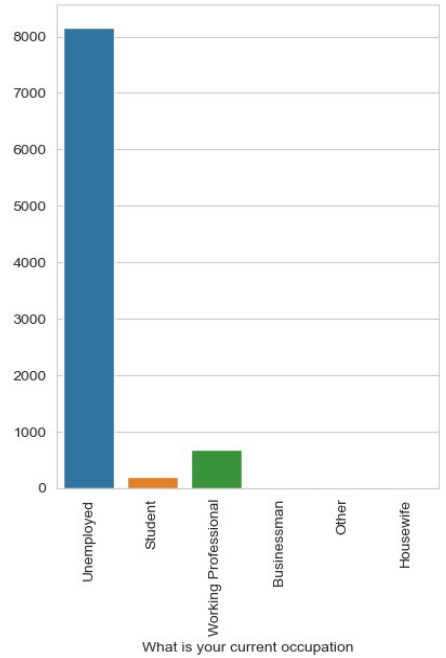
UNIVARIATE AND BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLE

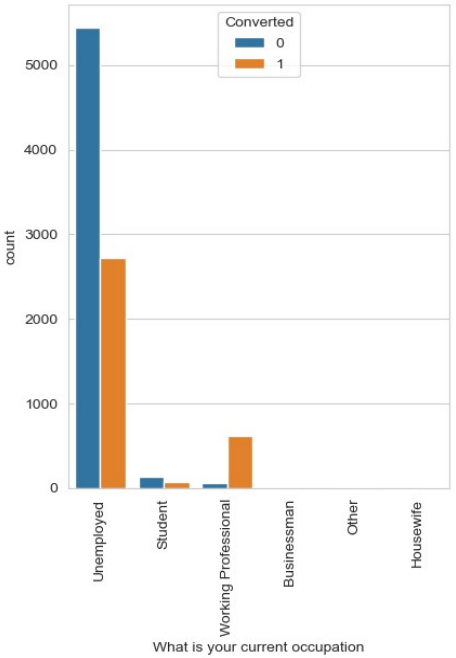


UNIVARIATE AND BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLE

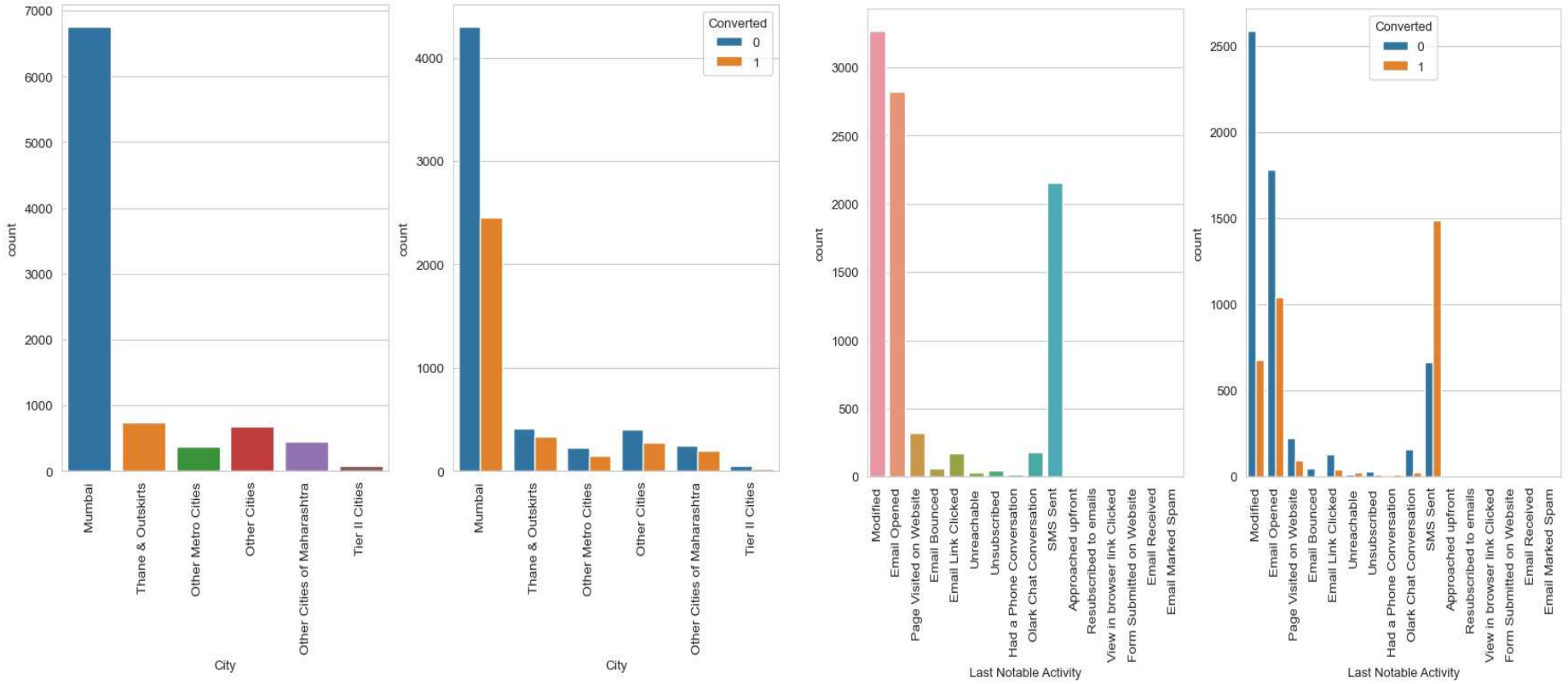




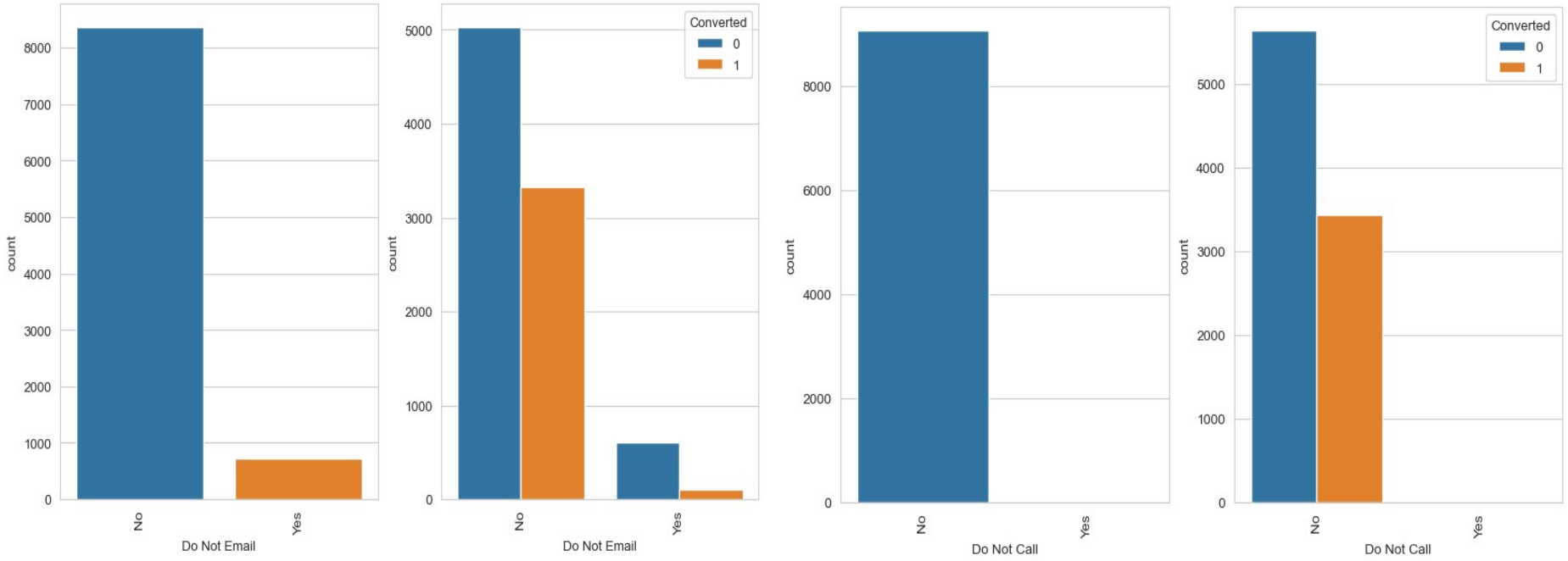




UNIVARIATE AND BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLE



UNIVARIATE AND BIVARIATE ANALYSIS FOR BINARY VARIABLE



ANALYSIS OBSERVATION

- "Lead Origin" is a category or source from which leads are generated, and "landing page submission" is one of the methods used to capture leads. It's good to hear that "landing page submission" is effective, and it's even better that customer conversions are proportional to these submissions. This suggests that the leads acquired through landing page submissions are more likely to convert into customers.
- Google is the most popular search engine globally, and many people use it to find information, products, and services. By optimizing your online presence for Google search (Search Engine Optimization or SEO), you can attract organic traffic and leads.
- Most customer prefers do not disturb as their option. It is understood that, they do not want to be influenced from the company, whether they are converted or not converted.
- ‘Email opened’ and ‘SMS sent’ seems higher in last activity segment.
- Since most of the person don't prefer to fill specialization. However, focusing on finance management, human resource management, marketing management would give higher conversion ratio.
- Unemployed customer have higher number as compare to working professional or student. However, working professional is having higher conversion ratio.
- Customer belongs to Mumbai is the highest conversion ratio compare to other cities in Maharashtra state and also in tier-II cities.
- "SMS Sent" last noted activity has highest conversion ratio compared to others.

MODEL BUILDING STEPS

- Data preparation steps.
- Splitting into train and test set
- Scale variables in train set
- Use RFE to eliminate less relevant variables
- Build the first model
- Eliminate variables based on high P-values and VIF values
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test prediction.
- Assign lead score to individual lead.

MODEL BUILDING

- For model building , we used recursive feature elimination technique to get the top 20 features to build model.
- We have build the 1st model.
- We need to consider for P-value of variables and VIF.
- We need to create model till P-values of all variables less than 0.05 and VIF less than 5 by dropping irrelevant feature recursively one by one.

Generalized Linear Model Regression Results

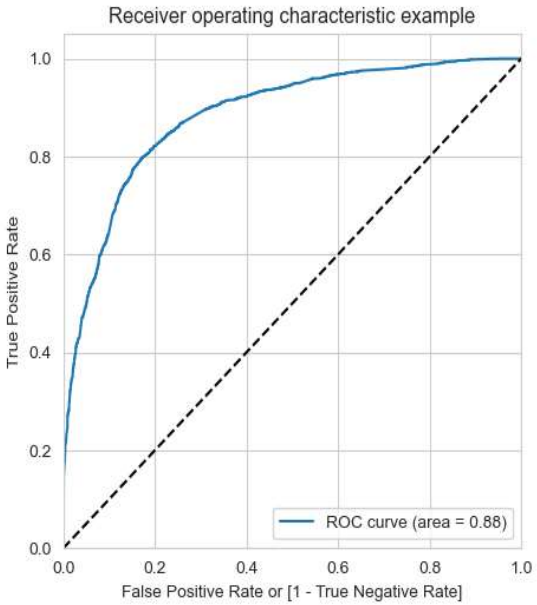
Dep. Variable:	Converted	**No. Observations:**	6333
Model:	GLM	**Df Residuals:**	6316
Model Family:	Binomial	**Df Model:**	16
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2614.4
Date:	Thu, 12 Oct 2023	**Deviance:**	5228.8
Time:	23:46:12	**Pearson chi2:**	6.40e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3938
Covariance Type:	nonrobust		

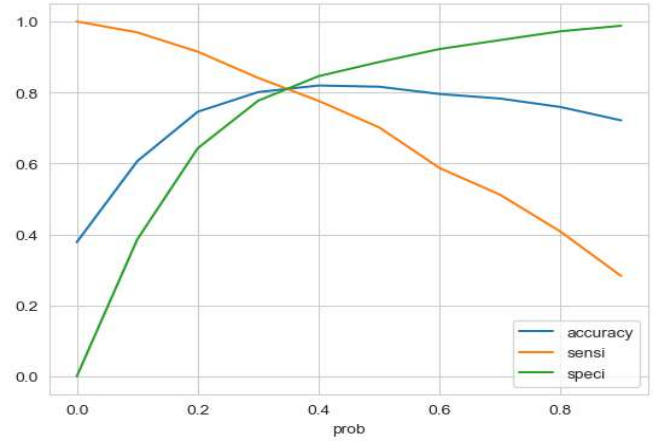
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6686	0.159	-4.211	0.000	-0.980	-0.357
Do Not Email	-1.1130	0.178	-6.249	0.000	-1.462	-0.764
Total Time Spent on Website	1.0943	0.040	27.032	0.000	1.015	1.174
Lead Origin_Landing Page Submission	-1.0782	0.131	-8.214	0.000	-1.335	-0.821
Lead Origin_Lead Add Form	3.3851	0.241	14.056	0.000	2.913	3.857
Lead Origin_Lead Import	1.0863	0.498	2.181	0.029	0.110	2.063
Lead Source_Olark Chat	1.0597	0.122	8.654	0.000	0.820	1.300
Lead Source_Welingak Website	3.0311	1.038	2.921	0.003	0.997	5.065
Last Activity_Email Opened	0.5726	0.110	5.213	0.000	0.357	0.788
Last Activity_Had a Phone Conversation	3.1438	0.771	4.076	0.000	1.632	4.655
Last Activity_SMS Sent	1.7598	0.110	15.982	0.000	1.544	1.976
Specialization_Hospitality Management	-0.8192	0.325	-2.518	0.012	-1.457	-0.182
Specialization_Other	-1.1274	0.126	-8.934	0.000	-1.375	-0.880
What is your current occupation_Working Professional	2.6281	0.192	13.656	0.000	2.251	3.005
Last Notable Activity_Modified	-0.7974	0.089	-8.951	0.000	-0.972	-0.623
Last Notable Activity_Olark Chat Conversation	-1.0901	0.345	-3.161	0.002	-1.766	-0.414
Last Notable Activity_Unreachable	2.1291	0.531	4.006	0.000	1.087	3.171

Model Evaluation(Train data)

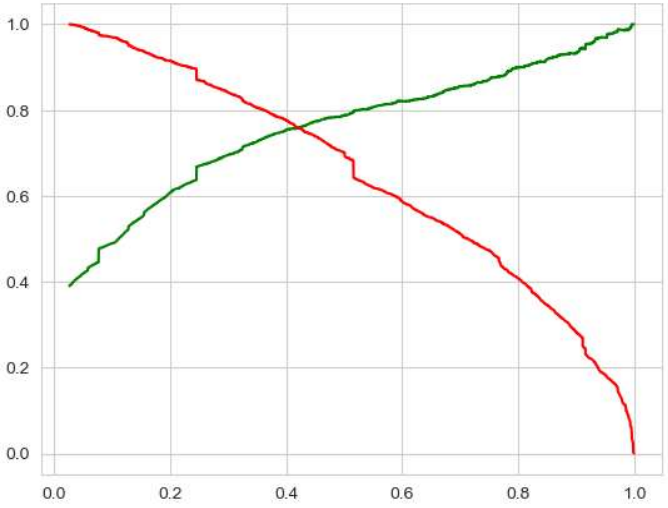
- ACCURACY : 81.24%
- SENSTIVITY : 80.53%
- SPECIFICITY : 81.67%

```
array([[3217,  722],
       [ 466, 1928]], dtype=int64)
```





From the curve above, 0.35 is the optimum point take it as a cutoff probability.



Model Evaluation(Test data)

- ACCURACY : 81.36%
- SENSTIVITY : 79.69%
- SPECIFICITY : 82.39%

```
array([[1385,  296],
       [ 210,  824]], dtype=int64)
```

LEAD SCORE PREDICTION

- The final predicted column shows the conversion probability of prospective lead.
- Lead score above 85 have high tendency of converting to a hot lead category.
- There are total around 397 customers with ore than 85% lead score.

	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score
3	4680	1	0.995586	1	100
6	6314	0	0.974165	1	97
7	2688	1	0.995586	1	100
17	4050	1	0.971110	1	97
20	4382	1	0.882028	1	88
...	...	...	...	...	...
2687	5206	1	0.876990	1	88
2696	2493	1	0.915870	1	92
2698	8056	1	0.997856	1	100
2702	2907	1	0.910690	1	91
2707	3262	0	0.967311	1	97

397 rows × 5 columns

There are total around 397 customer with more than 85% lead score can be considered to be hot leads.

PARAMETERS OF MODEL

- Important features consideration for model building.
- Features impact either positively or negatively.
- Starting from top up to “Last activity Email Opened” feature will positively impact to model.
- Beyond constant all feature will negatively impact to model.

```
Lead Origin_Lead Add Form
Last Activity_Had a Phone Conversation
Lead Source_Welingak Website
What is your current occupation_Working Professional
Last Notable Activity_Unreachable
Last Activity_SMS Sent
Total Time Spent on Website
Lead Origin_Lead Import
Lead Source_Olark Chat
Last Activity_Email Opened
const
Last Notable Activity_Modified
Specialization_Hospitality Management
Lead Origin_Landing Page Submission
Last Notable Activity_Olark Chat Conversation
Do Not Email
Specialization_Other
dtype: float64
```

RECOMMENDATIONS

1. Company should actively pursue the leads who lead origin is "Lead Add Form.
2. Company should have a conversation with potential customer.
3. Company should make contact to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
4. Company should make contact to the leads who are the "working professionals" as they are more likely to get converted.
5. Company should make contact to the leads whose last activity was SMS Sent as they are more likely to get converted.
6. Company should make contact to the leads who spent "more time on the websites" as these are more likely to get converted.

THANK YOU