Lead Scoring Case Study

APARNA , DEVESH GABA, DINKAR BHATIA

JULY'2024

Business Objective

To help X Education select most promising leads (Hot Leads), i.e. the leads that are most likely to convert into paying customers.

Higher Lead Conversion Rate

Focused marketing

Selection of Hot leads

Steps undertaken

Broad steps under taken

- Read and understand the Data and data dictionary: Data set consists of 9240 leads with 37 columns. It is a mix of categorical, numeric and binary variables. Many columns have missing values requiring data handling. ‘Converted’ column in the target variable

- Clean the data: missing value imputation data, remove duplicate columns and other redundancies

- Exploratory Data Analysis: Univariate and Bivariate analysis

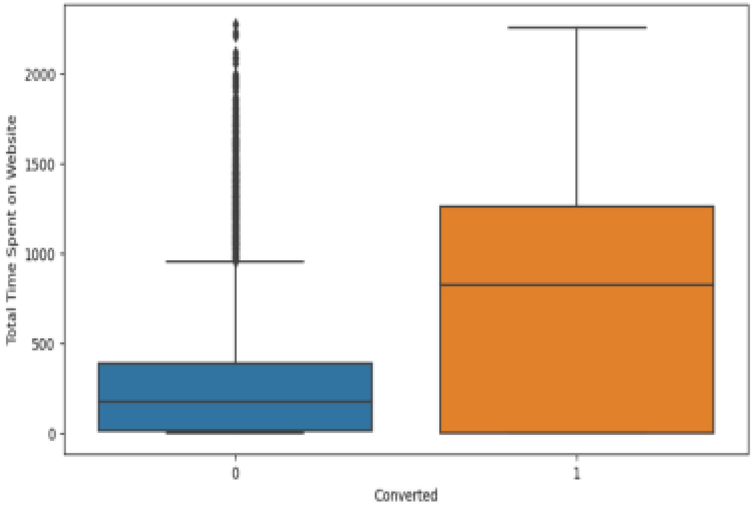
- Prepare the data for Model Building: Dummy variable creation

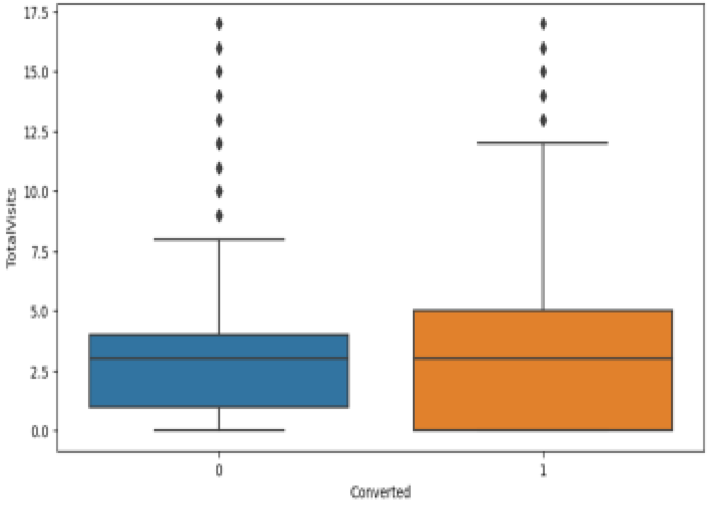
- Model Building: Feature standardization, Feature selection using RFE, manual feature elimination based on p-valus and VIF, finding optimal probability threshold.

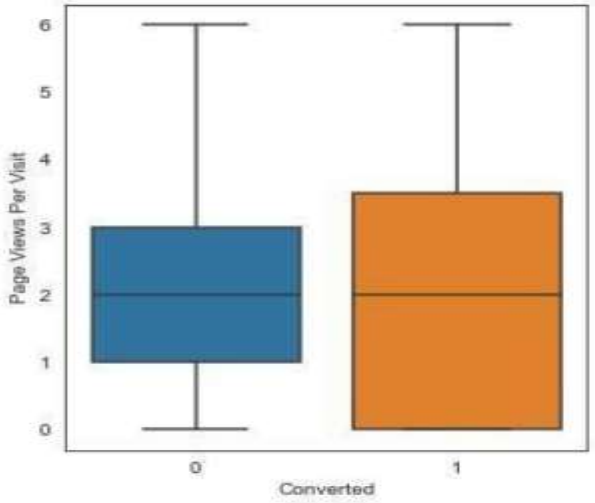
- Model Evaluation: basis various evaluation metrics

- Making Prediction on the Test set: using predicted probabilities to calculate Lead scores

Numerical variables



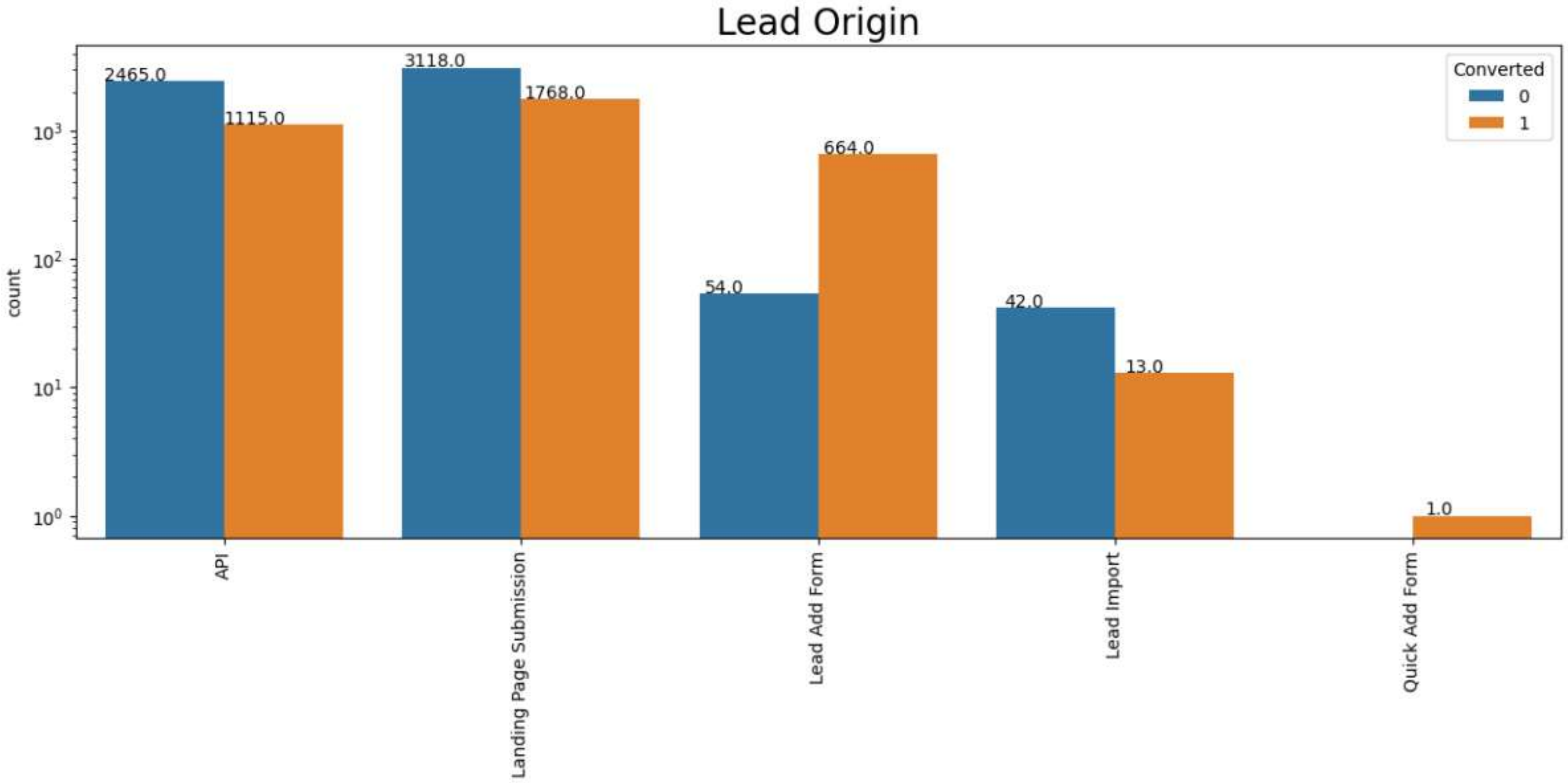




Inference:

- People spending more time on website are more likely to get converted

Categorical variable-Lead Origin

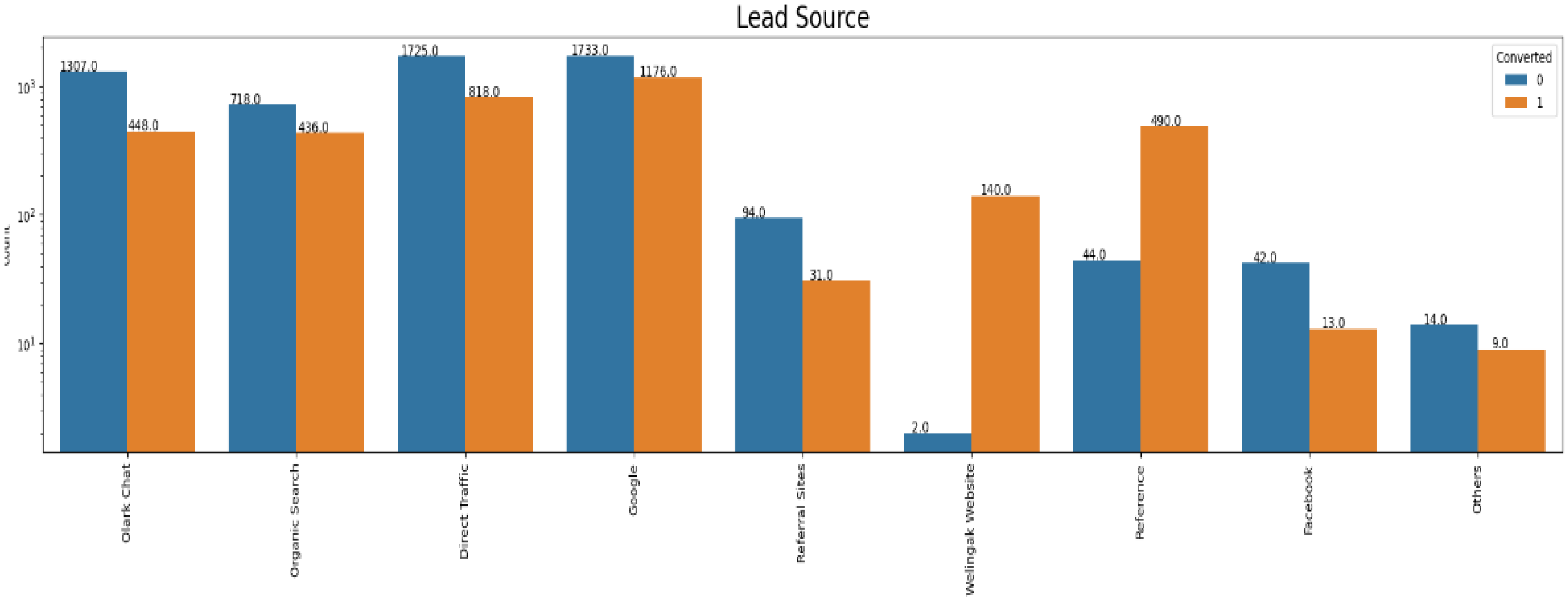


Inferences:

- API' and 'Landing Page Submission' generate the most leads but have less conversion rates, whereas 'Lead Add Form' generates less leads but conversion rate is great.

- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Categorical variable-Lead Source

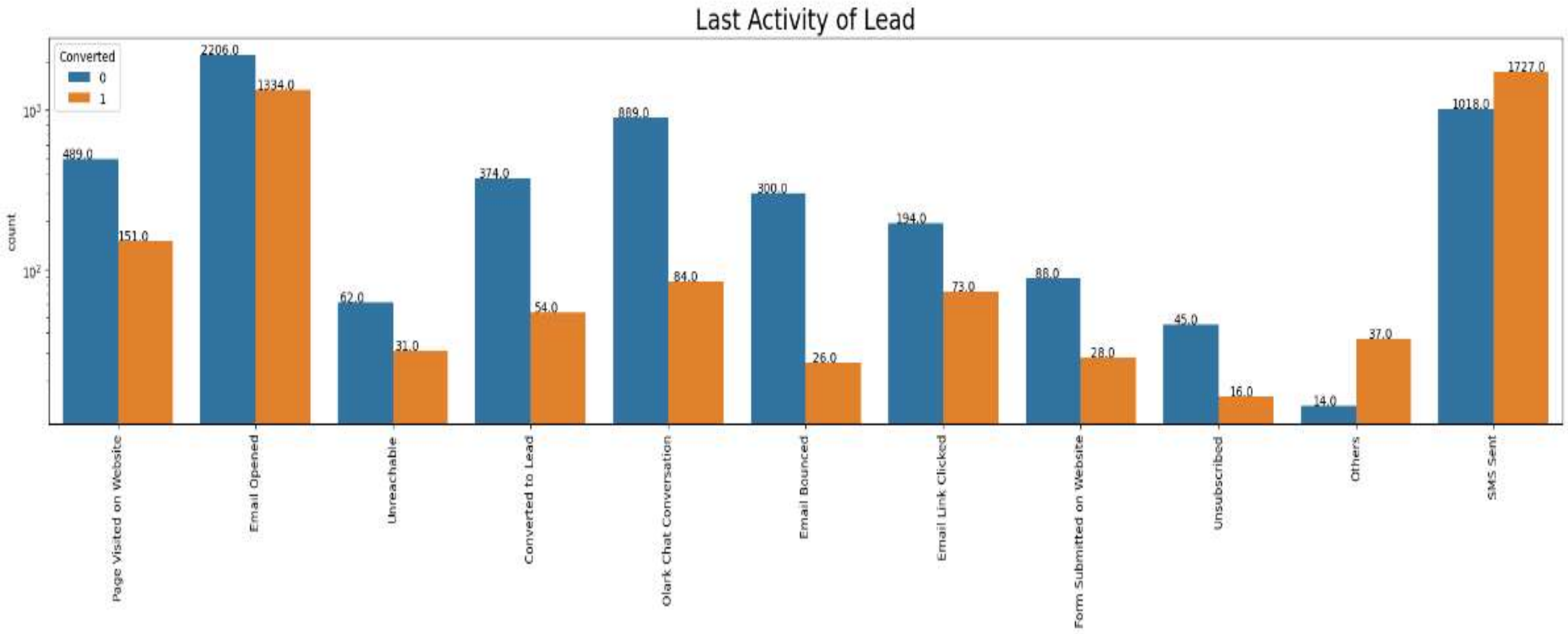


Inferences :

- 1. Google and Direct traffic generates maximum number of leads. 2 .Conversion rate of 'Reference' and 'Welingak Website' leads is high.

- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

Categorical variable-Last Activity of Lead

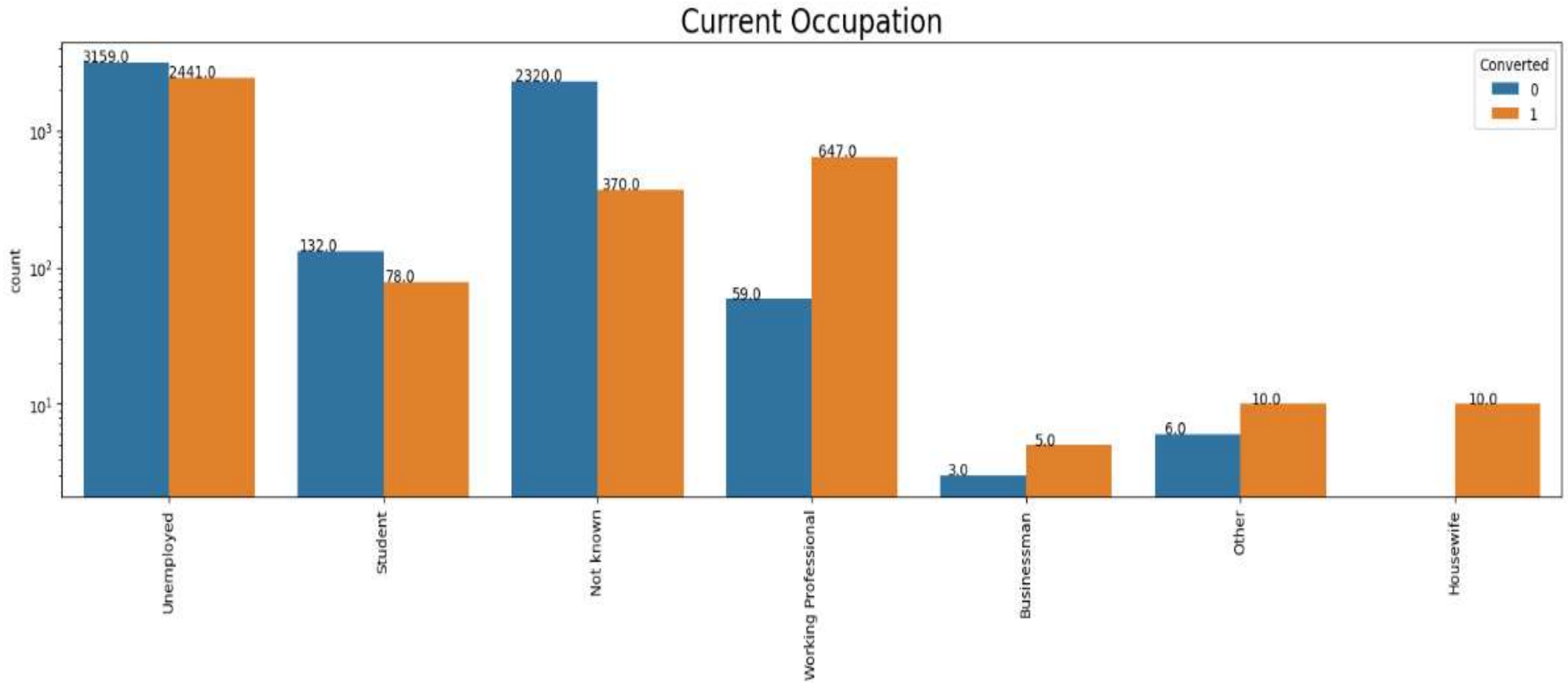


Inferences :

- Conversion rate for last activity of 'SMS Sent'is ~63%.

- Highest last activity of leads is 'Email Opened'.

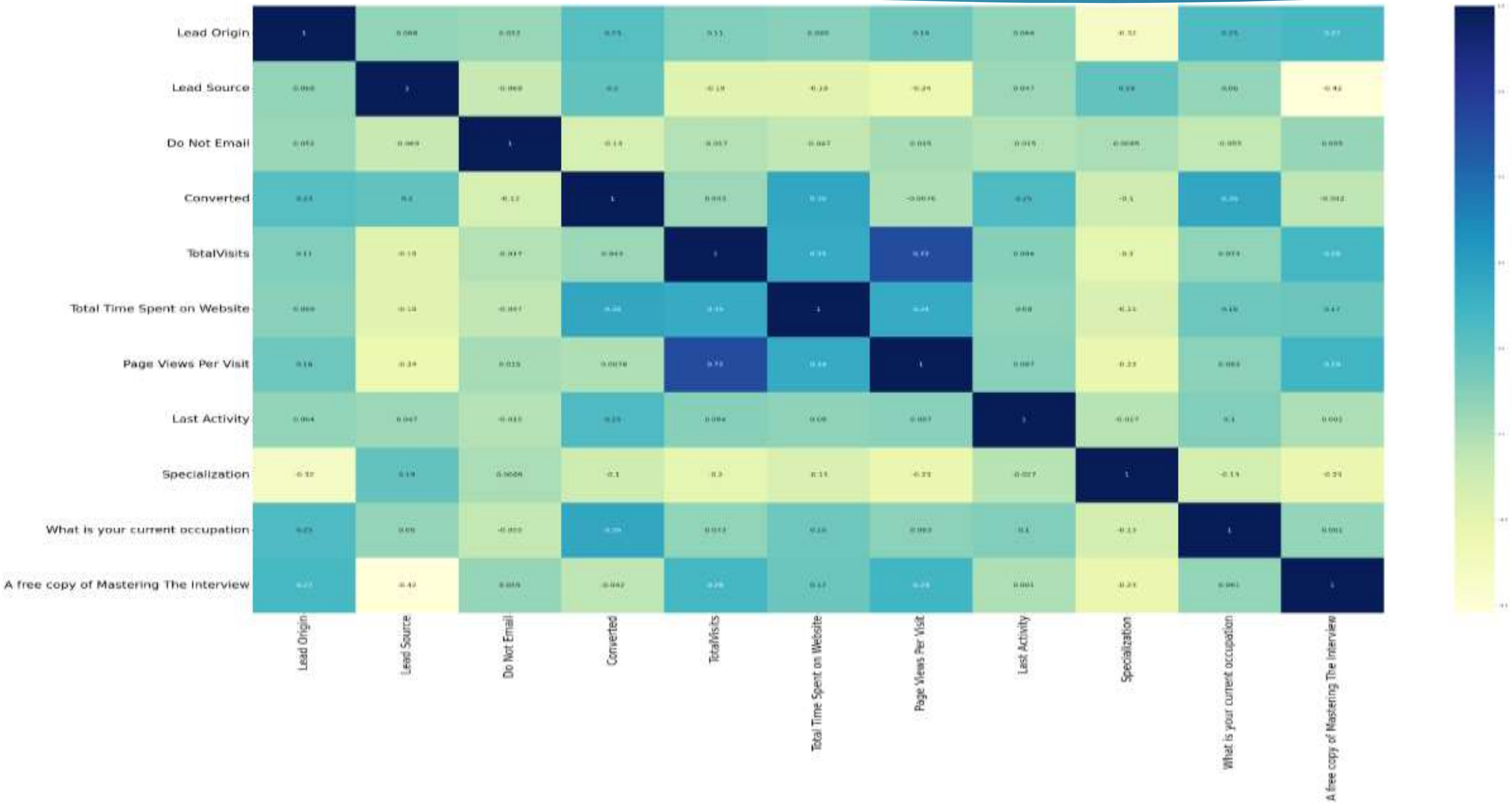
Categorical variable-Current Occupation



Inferences:

- - 'Unemployed' leads are generationg more number of leads and having ~45% conversion rate.
- Conversion rate is higher for 'Working Professionals'.

Heatmap

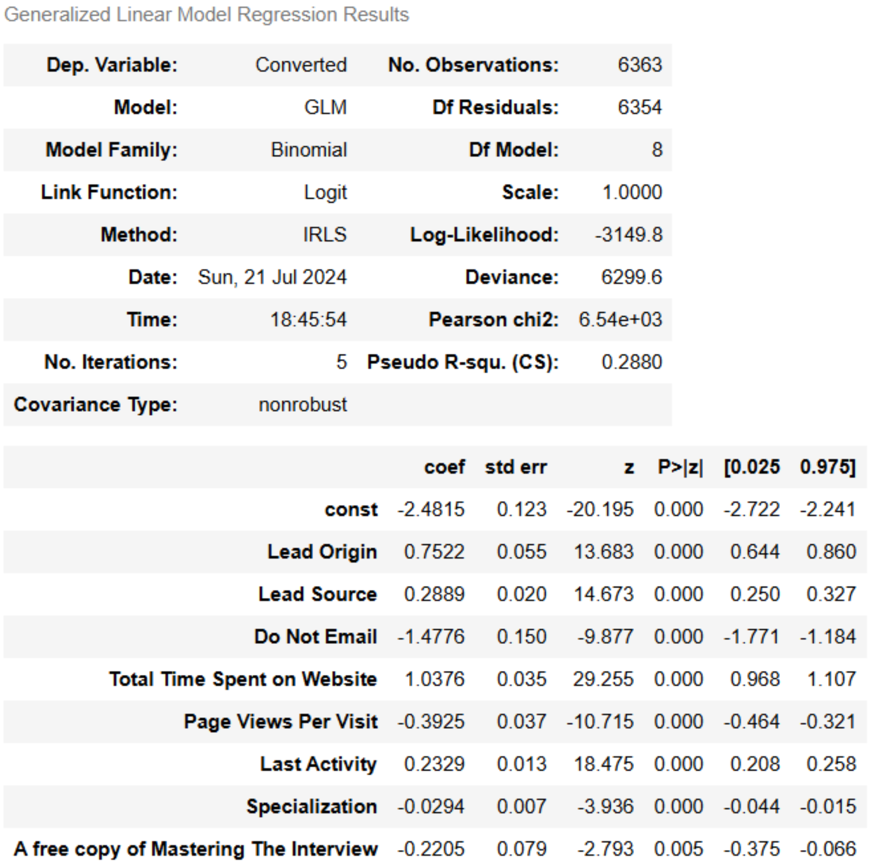


Inferences:

The heatmap clearly shows the variables that are multicollinear in nature, and which variables have high collinearity with the target variable.

- We will refer this map for building the logistic model so as to validate different correlated values along with VIF & p-value, for identifying the correct variable to select/eliminate from the model.

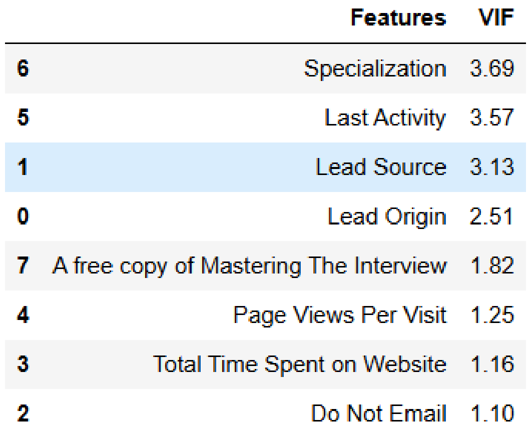
Final Model Summary



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6354
Model Family:	Binomial	**Df Model:**	8
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3149.8
Date:	Sun, 21 Jul 2024	**Deviance:**	6299.6
Time:	18:45:54	**Pearson chi2:**	6.54e+03
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.2880
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-2.4815	0.123	-20.195	0.000	-2.722	-2.241
Lead Origin	0.7522	0.055	13.683	0.000	0.644	0.860
Lead Source	0.2889	0.020	14.673	0.000	0.250	0.327
Do Not Email	-1.4776	0.150	-9.877	0.000	-1.771	-1.184
Total Time Spent on Website	1.0376	0.035	29.255	0.000	0.968	1.107
Page Views Per Visit	-0.3925	0.037	-10.715	0.000	-0.464	-0.321
Last Activity	0.2329	0.013	18.475	0.000	0.208	0.258
Specialization	-0.0294	0.007	-3.936	0.000	-0.044	-0.015
A free copy of Mastering The Interview	-0.2205	0.079	-2.793	0.005	-0.375	-0.066

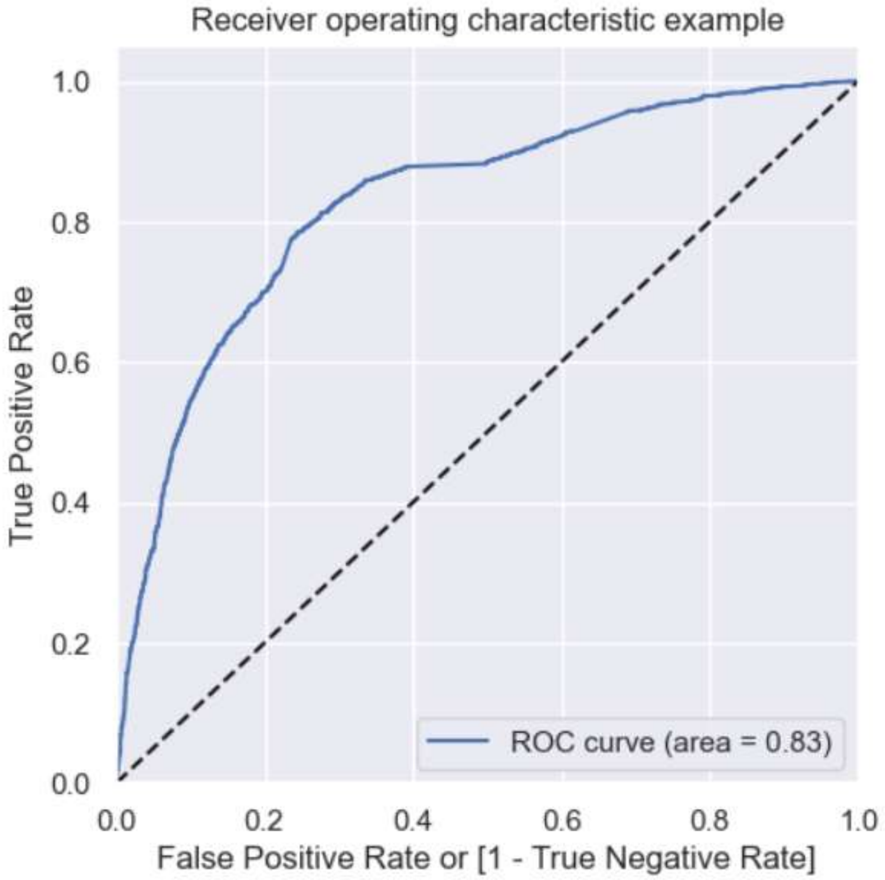


	Features	**VIF**
6	Specialization	3.69
5	Last Activity	3.57
1	Lead Source	3.13
0	Lead Origin	2.51
7	A free copy of Mastering The Interview	1.82
4	Page Views Per Visit	1.25
3	Total Time Spent on Website	1.16
2	Do Not Email	1.10

Conclusions:

- From model 'logm3' we can see that P-values of variables are significant and VIF values are below 3 .
- So we need not drop any more variables and we can proceed with making predictions using this model only considering model 'logm3' as final model.

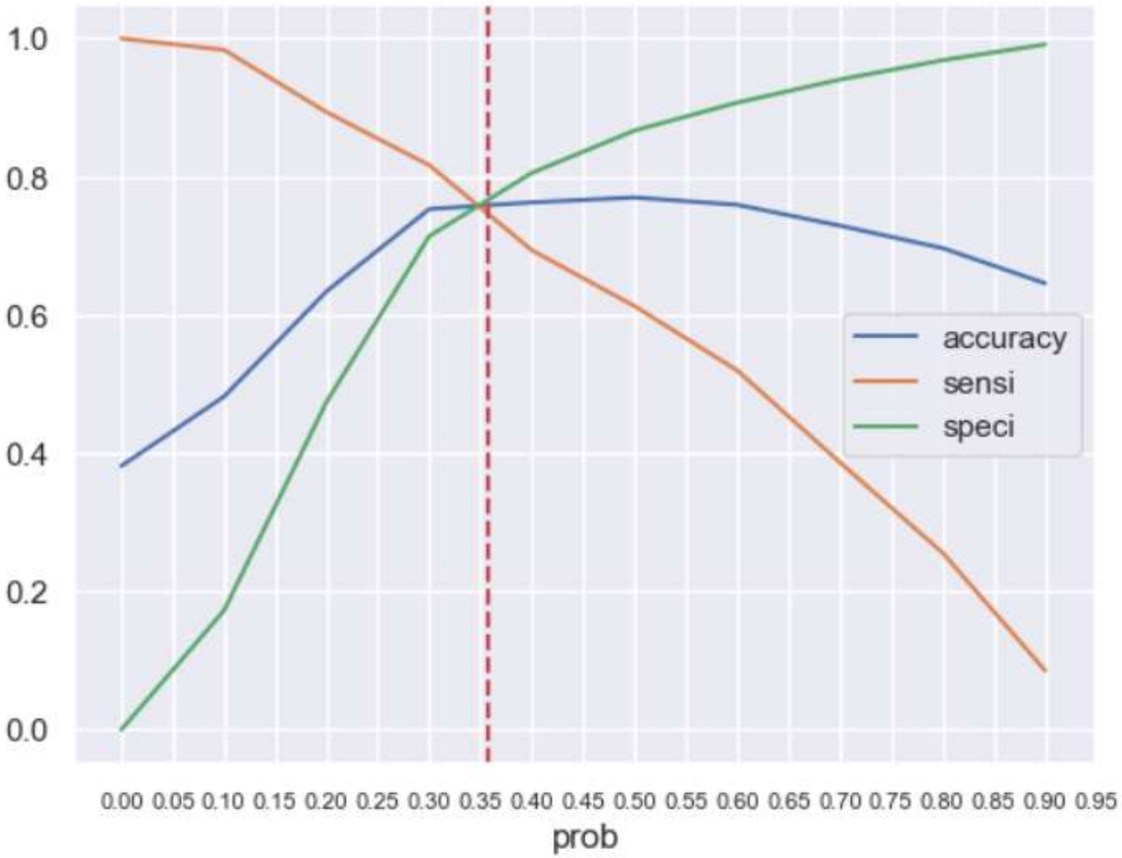
ROC Curve



Observations:

- Area under the curve we are getting a good value of 0.83 indicating a good predictive model. As ROC Curve should be a value close to 1.

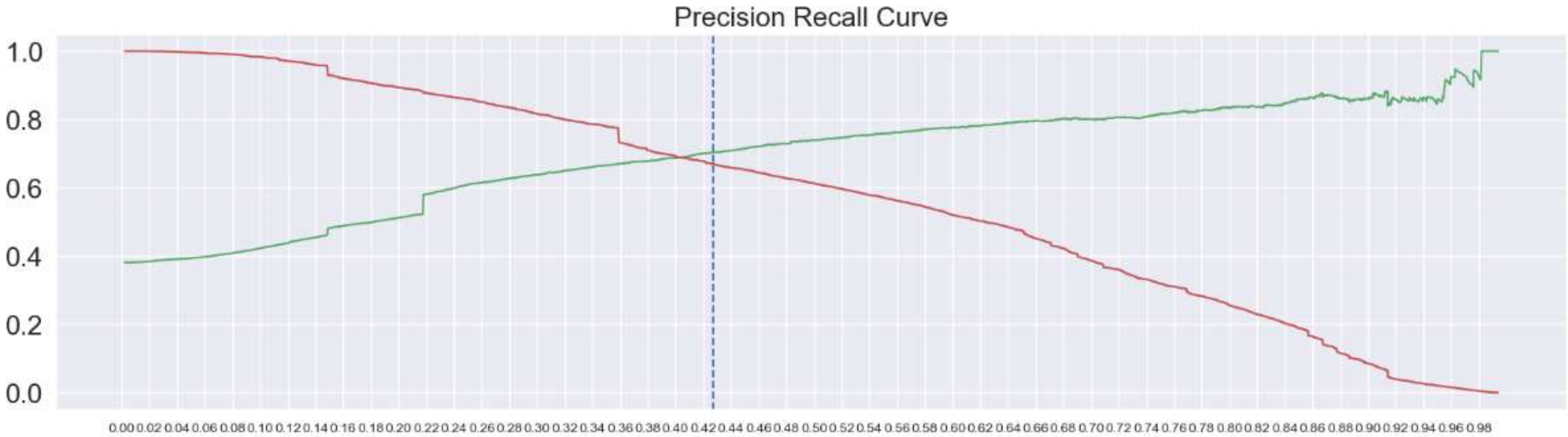
Finding Optimal Threshold



Observation:

- Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values has optimal cut-off point at 0.35

Finding Optimal Threshold

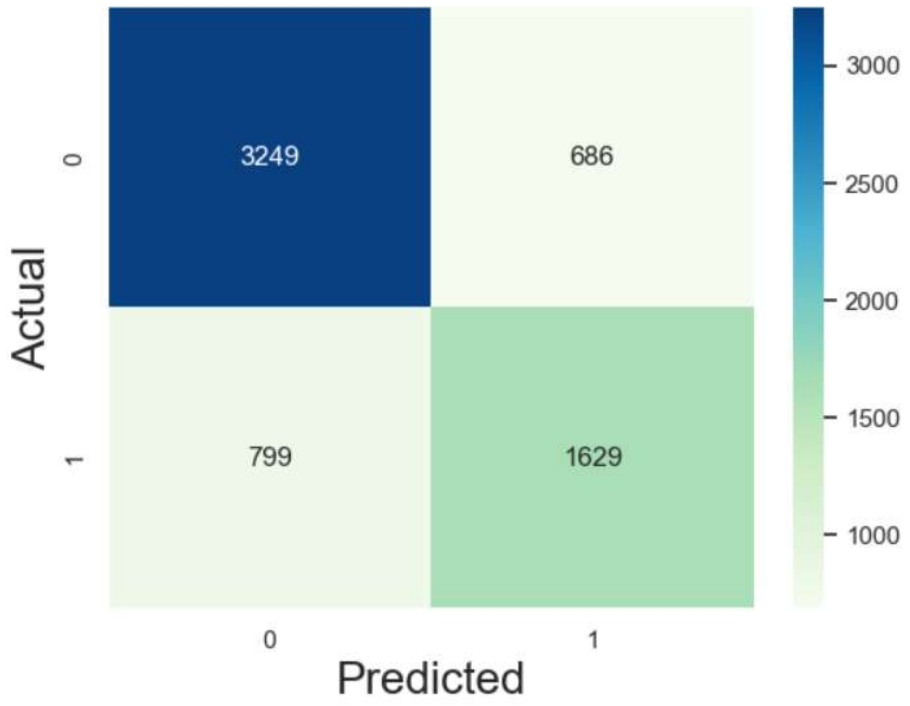


Observation:

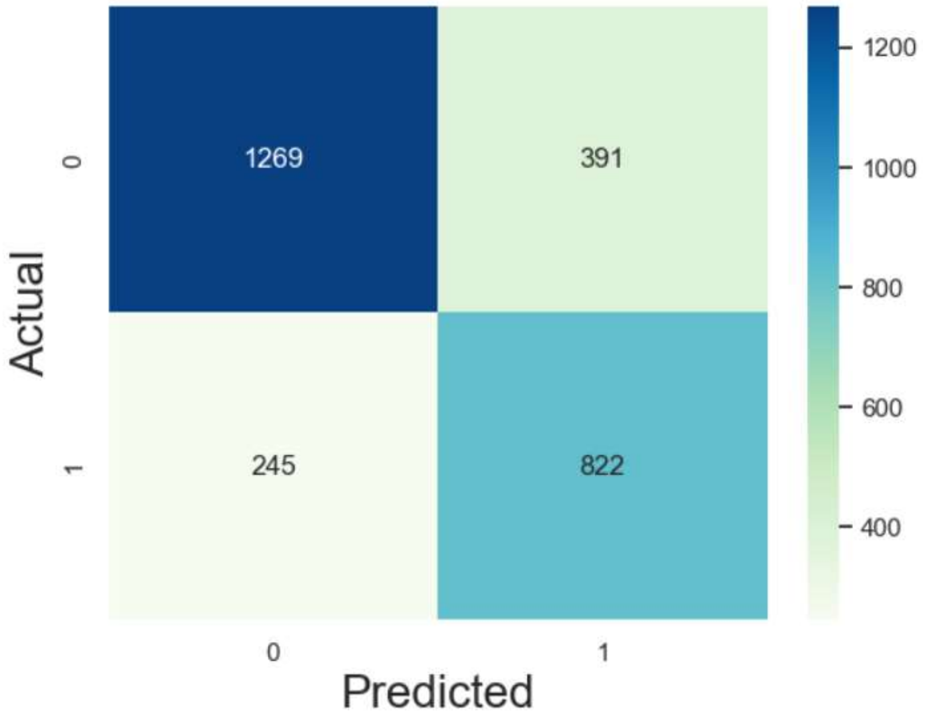
- Now using threshold value of 0.427 from 'Precision Recall Tradeoff Curve' for Data Evaluation

Confusion matrix Final result

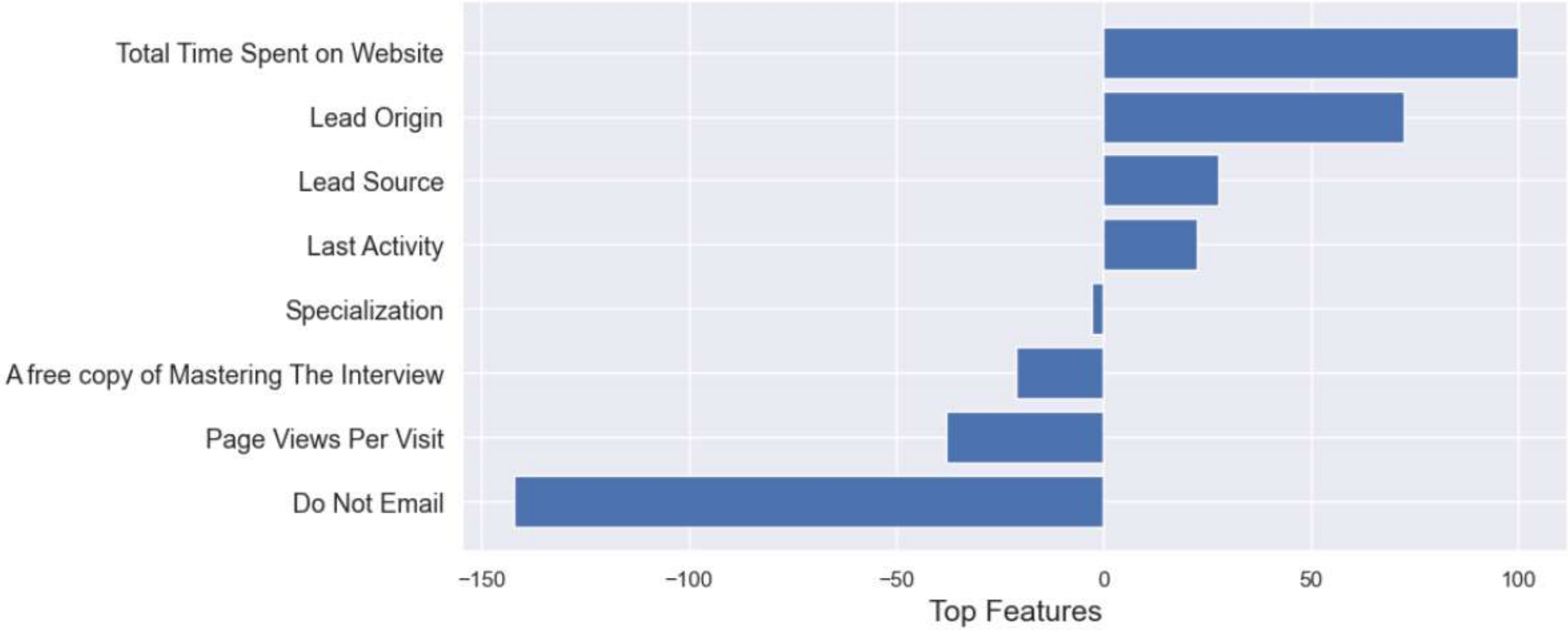
Train Set



Test Set



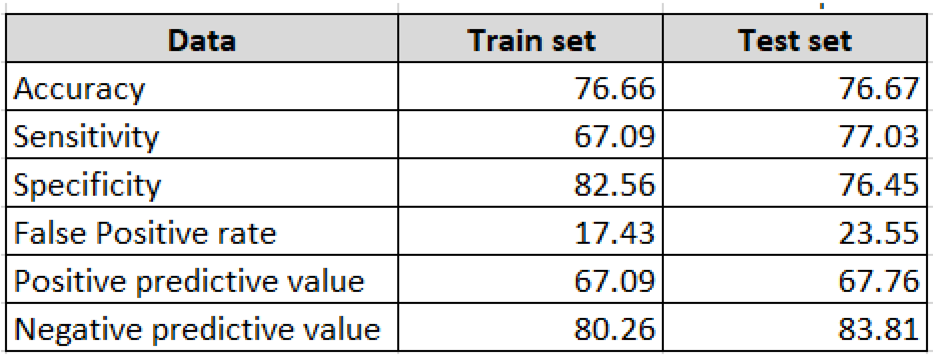
Relative Importance of Features



Observation:

- Total time spent on Website, Lead Origin, Lead Source and Last Activity contribute positively towards the probability of Lead conversion. Company should focus on these

Relative Importance of Features



Data	Train set	Test set
Accuracy	76.66	76.67
Sensitivity	67.09	77.03
Specificity	82.56	76.45
False Positive rate	17.43	23.55
Positive predictive value	67.09	67.76
Negative predictive value	80.26	83.81