Lead Scoring Case Study

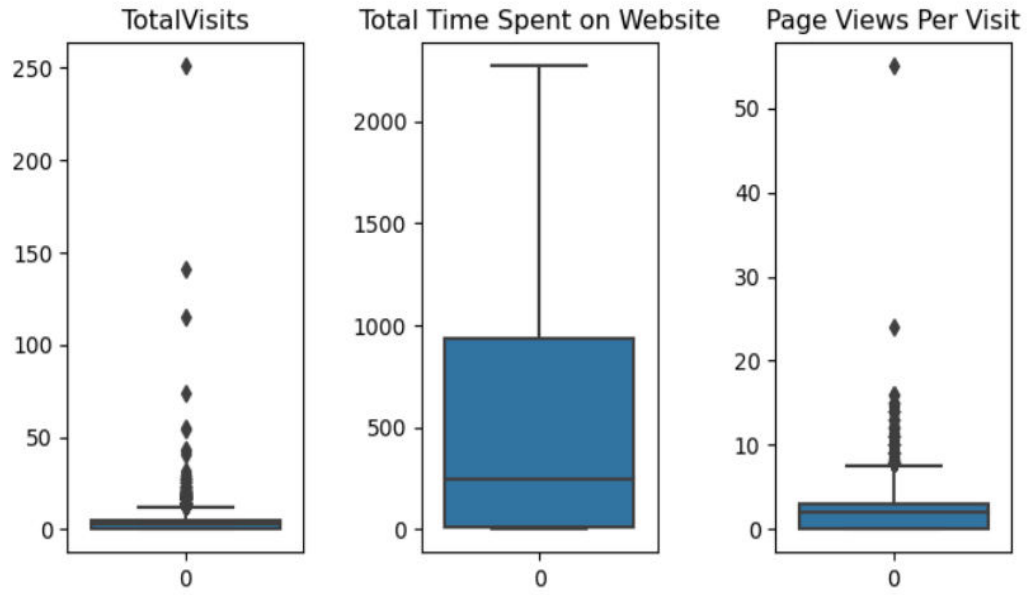
By – Kumari Madhumita

Krishna

Ramya Konathala

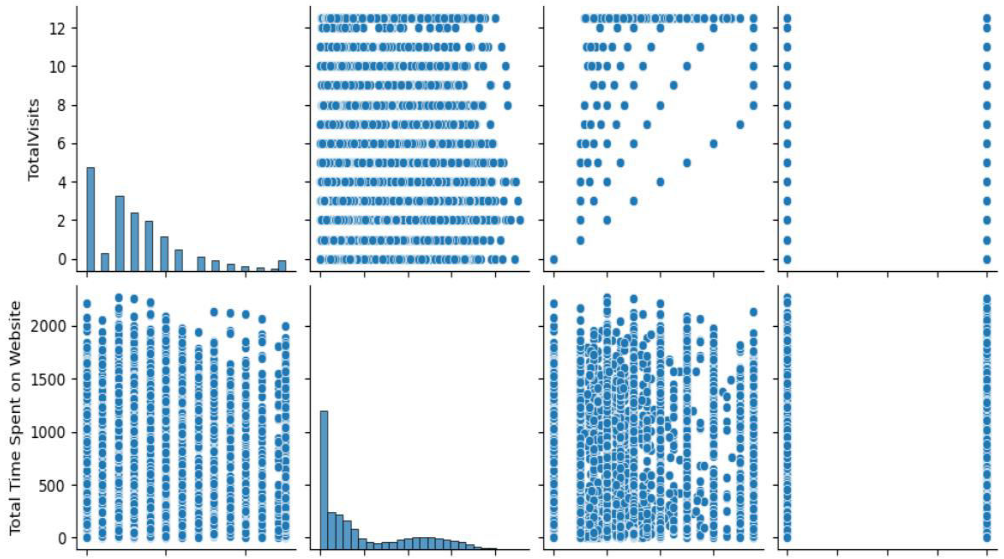
Data Analysis

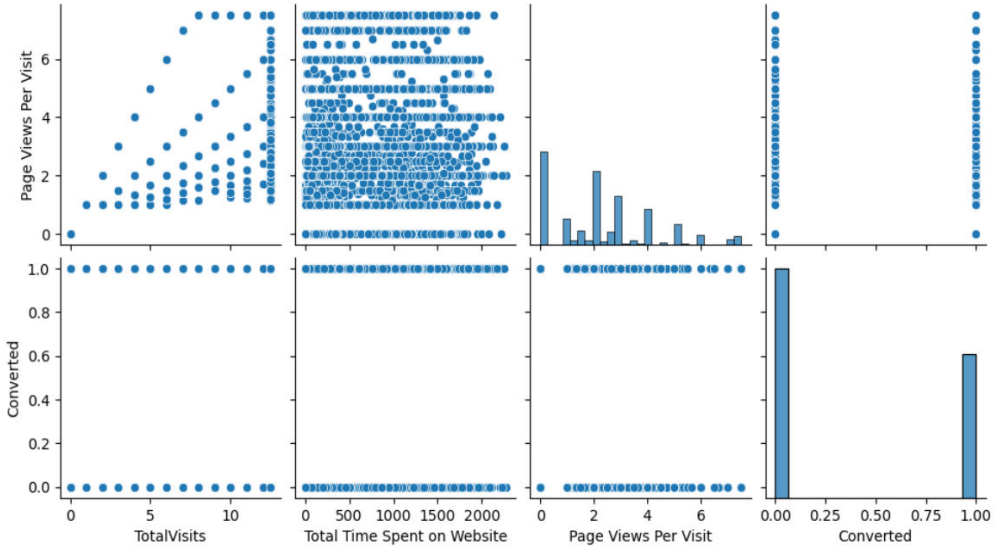
- Checked the dimensions, data types, uniqueness of lead file data.
- Removed the redundant values which were not helpful for data analysis.
- Checked for outliers present in data set, for variables ‘Total Visits’ & ‘Page Views per Visits’ there are outliers present so we capped it with “Quartile3+1.5*IQR”.



Data Visualization

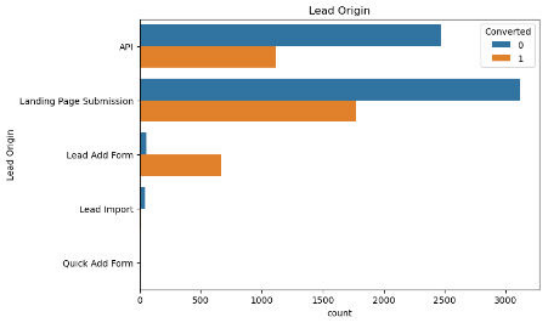
- We have created pair plots for numerical variables.

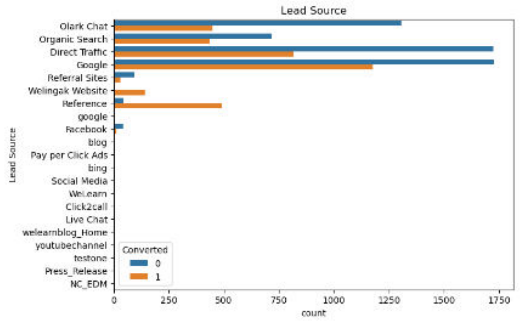


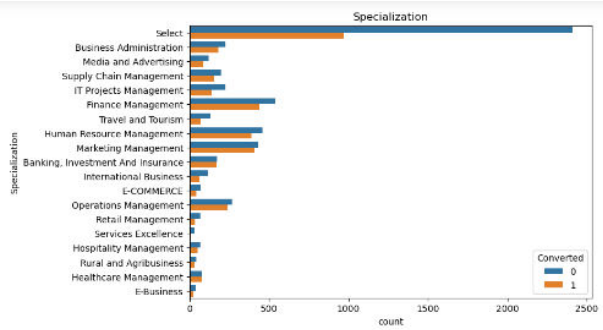


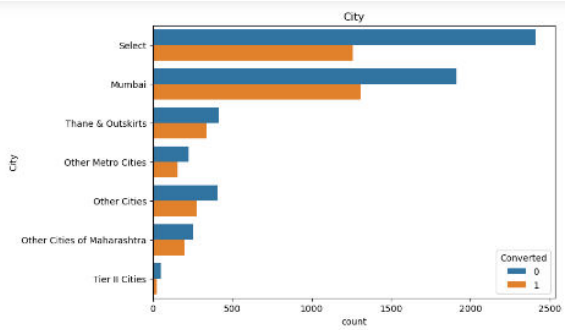
Data Visualization

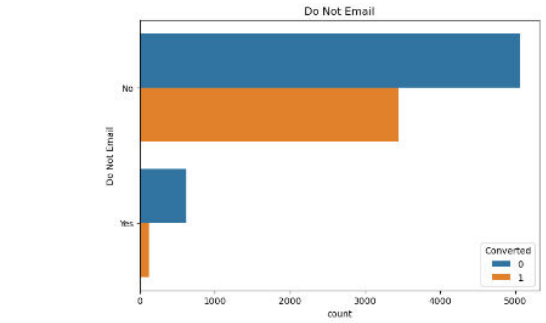
- We have created count plots for categorical variables.

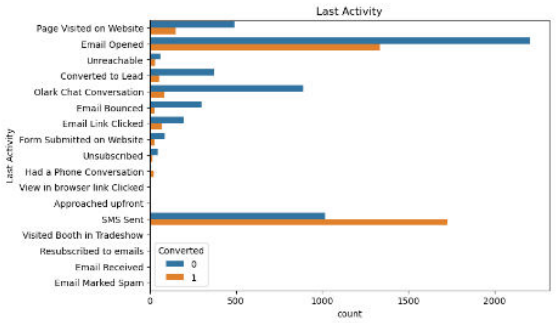


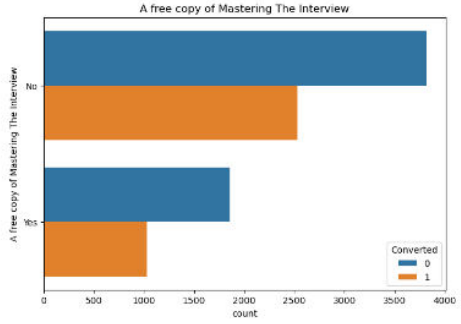


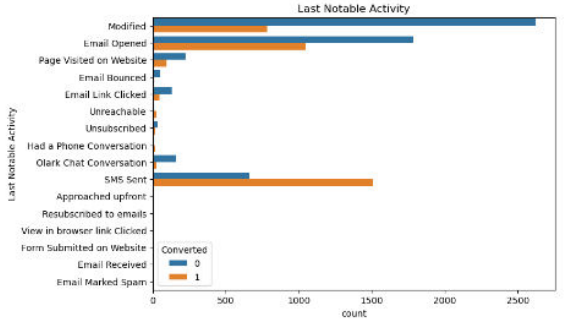






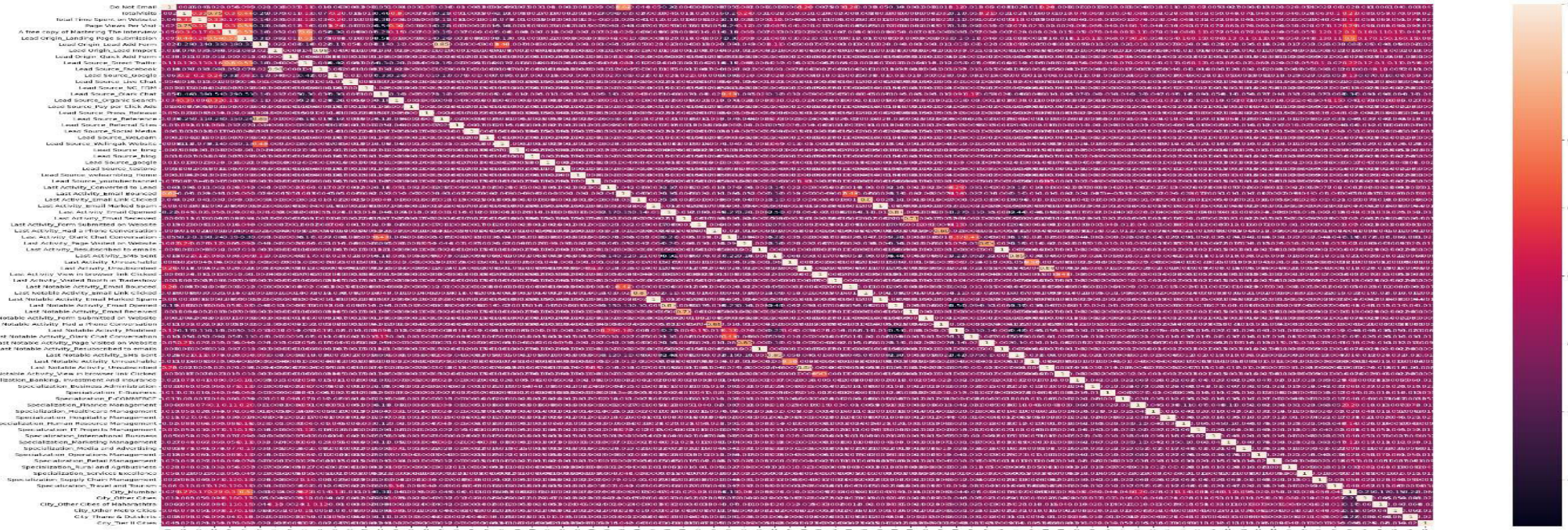






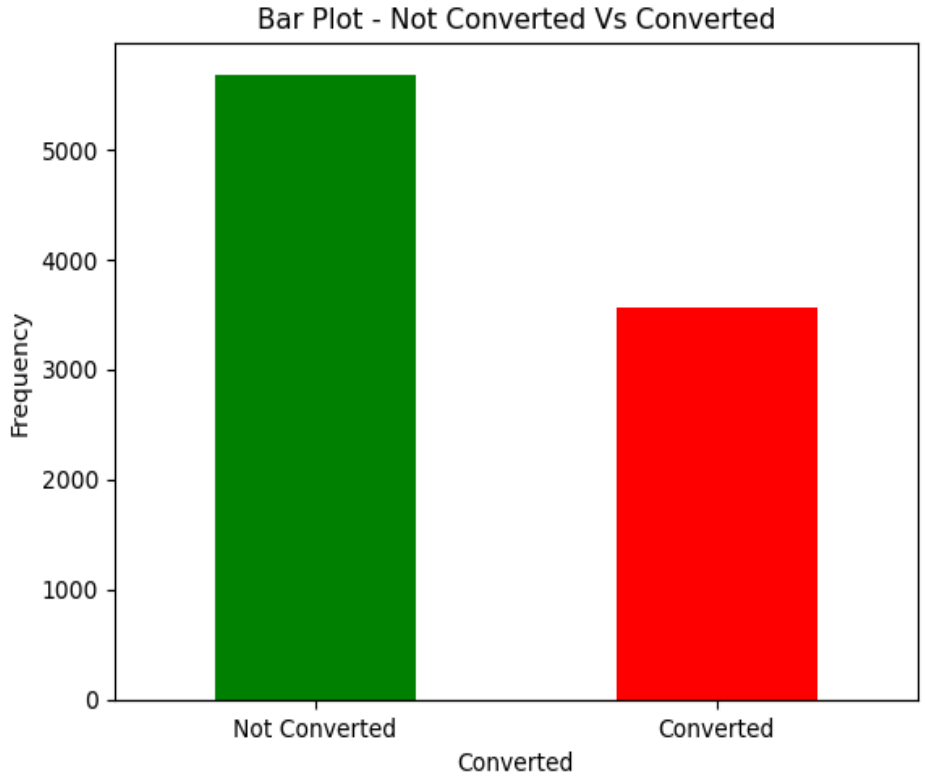
Data Visualization

- We have checked for correlations, since there are lot of variables present, we could not able find much information.



Test-Train Split

- We have used test-train split to find the conversion rate of leads i.e. it’s 38%.



We have around 38% conversion rate

Model building & Evaluation

- We have used RFE & VIF in model building, model 7 has less p-value and VIF hence, considered the correct one.
- We have used the previously calculated converted_prob to find the prediction of lead getting converted or not.

Final model 7:

Out[84]:

	Features	VIF
4	Last Activity_Olark Chat Conversation	1.89
3	Lead Source_Olark Chat	1.65
7	Last Notable Activity_Modified	1.51
8	Last Notable Activity_Olark Chat Conversation	1.30
1	Total Time Spent on Website	1.17
0	Do Not Email	1.11
6	Last Notable Activity_Email Opened	1.09
2	Lead Origin_Lead Add Form	1.08
5	Last Notable Activity_Email Link Clicked	1.02
9	Last Notable Activity_Page Visited on Website	1.02

Out[88]:

	Converted	Converted_Prob	ProsID	predicted
0	0	0.279202	1871	0
1	0	0.257711	6795	0
2	0	0.324738	3516	0
3	0	0.859075	8105	1
4	0	0.186040	3934	0

Final Logistic Regression Model Features

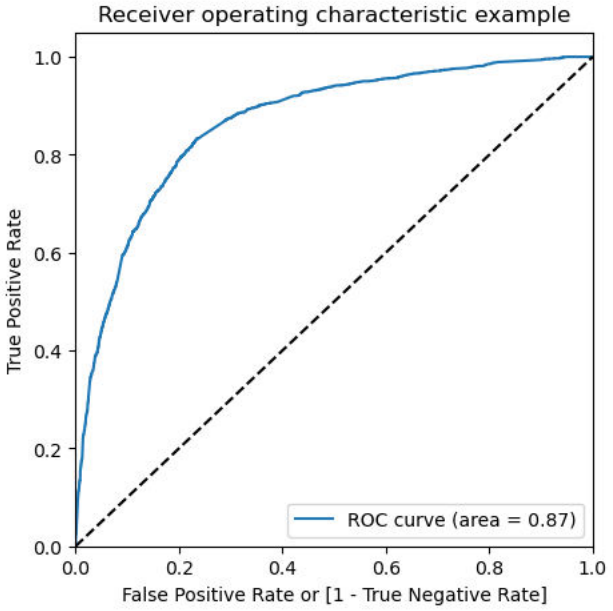
Generalized Linear Model Regression Results

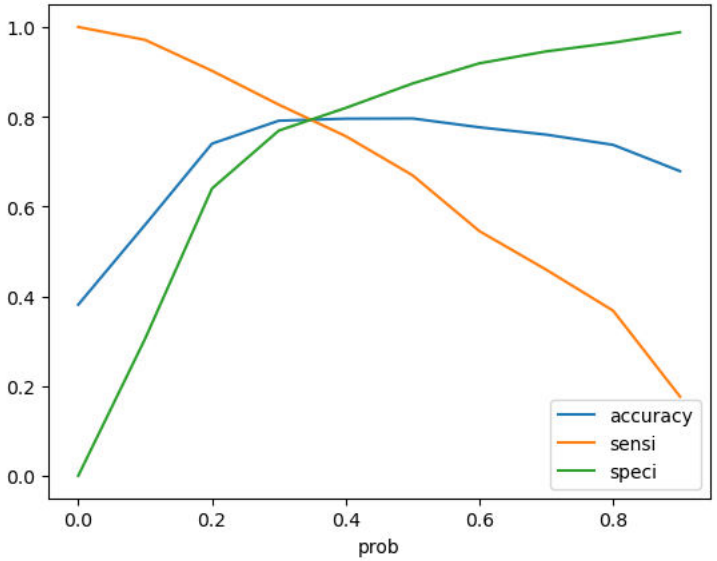
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6457
Model Family:	Binomial	**Df Model:**	10
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2885.9
Date:	Sun, 14 Apr 2024	**Deviance:**	5771.8
Time:	10:02:01	**Pearson chi2:**	6.68e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3540
Covariance Type:	nonrobust		

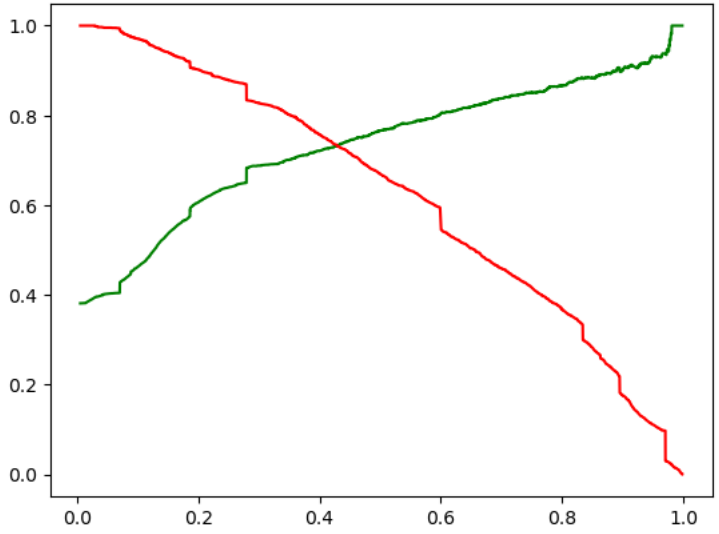
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2848	0.065	4.357	0.000	0.157	0.413
Do Not Email	-1.6506	0.162	-10.163	0.000	-1.969	-1.332
Total Time Spent on Website	1.1154	0.038	29.169	0.000	1.040	1.190
Lead Origin_Lead Add Form	4.1944	0.179	23.486	0.000	3.844	4.544
Lead Source_Olark Chat	1.1034	0.099	11.159	0.000	0.910	1.297
Last Activity_Olark Chat Conversation	-1.1087	0.188	-5.911	0.000	-1.476	-0.741
Last Notable Activity_Email Link Clicked	-1.8743	0.258	-7.277	0.000	-2.379	-1.370
Last Notable Activity_Email Opened	-1.3491	0.084	-16.137	0.000	-1.513	-1.185
Last Notable Activity_Modified	-1.8766	0.092	-20.492	0.000	-2.056	-1.697
Last Notable Activity_Olark Chat Conversation	-1.6268	0.362	-4.495	0.000	-2.336	-0.917
Last Notable Activity_Page Visited on Website	-1.5818	0.184	-8.617	0.000	-1.942	-1.222

ROC Curves

- It shows the trade off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.







Leading Score Value

- We have derived lead score value from previously calculated converted_prob and using y_train_pred_final. In the end, each row is having a lead score value.

Lead_Score_df

	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	City	Asymmetrique Activity Index	Asymmetrique Profile Index	Asymmetrique Activity Score	Asymmetrique Profile Score	I agree to pay the amount through cheque	A free copy of Mastering The Interview	Last Notable Activity	ProsID	Lead Score
k it	No	No	0	0.0	0	0.00	...	Select	02.Medium	02.Medium	15.0	15.0	No	No	Modified	0	19
c h	No	No	0	5.0	674	2.50	...	Select	02.Medium	02.Medium	15.0	15.0	No	No	Email Opened	1	34
:t c	No	No	1	2.0	1532	2.00	...	Mumbai	02.Medium	01.High	14.0	20.0	No	Yes	Email Opened	2	75

Determining features for a customer's Lead to be converted

- Customers ‘Lead Origin’ from ‘Lead Add Form’ has strong positive influence on the ‘Converted’ variable.
- Higher the ‘Total Time Spent on Website’ by the customer, higher the chances of the Lead getting converted.
- ‘Lead Source’ of customers from ‘Olark Chat’ has positive association with ‘Converted’ variable.
- Customers lead whose ‘Last Notable Activity’ is ‘SMS Sent’ has high chance of getting converted.
- Customers lead whose ‘Lead Source’ is ‘Reference’ are more likely to be converted.