Lead Scoring Assignment

Problem Statement

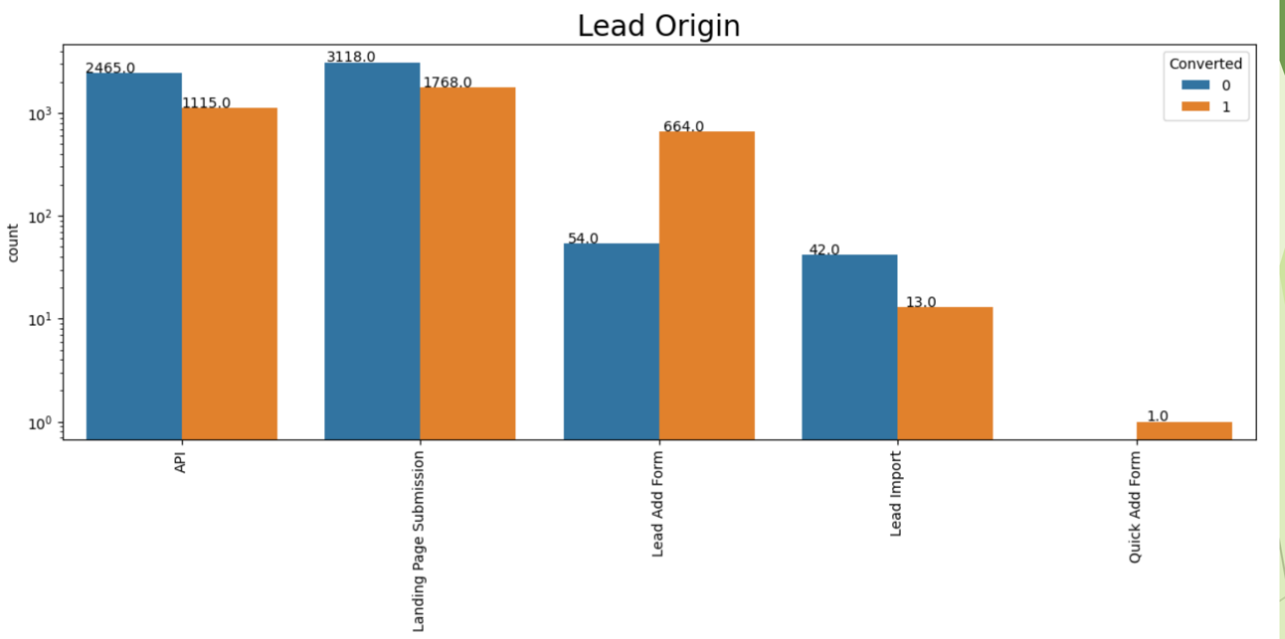
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

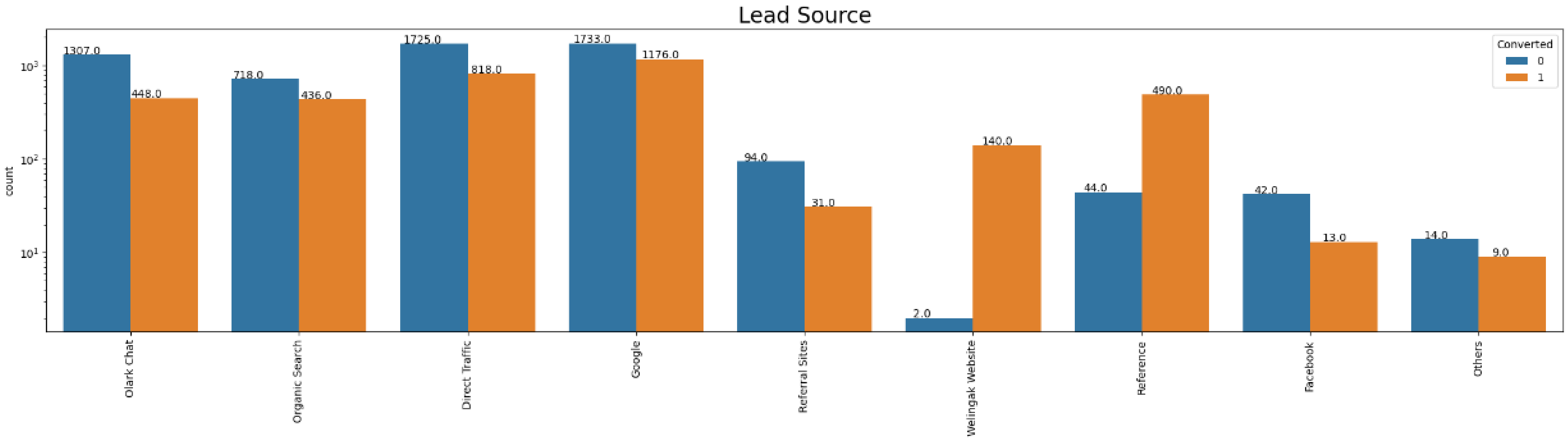
Business Criteria

- To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa.
- Target Lead Conversion Rate ≈ 80%

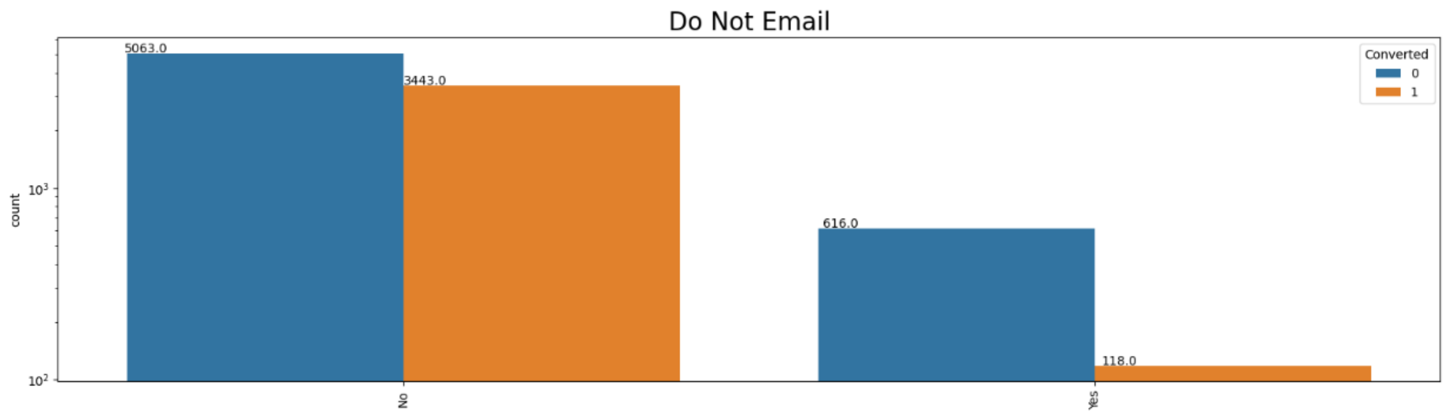
Data Visualization

- Conversion rate for 'API' is ~ 31% and for 'Landing Page Submission' is ~36%.
- For 'Lead Add Form' number of conversion is more than unsuccessful conversion.
- Count of 'Lead Import' is lesser.

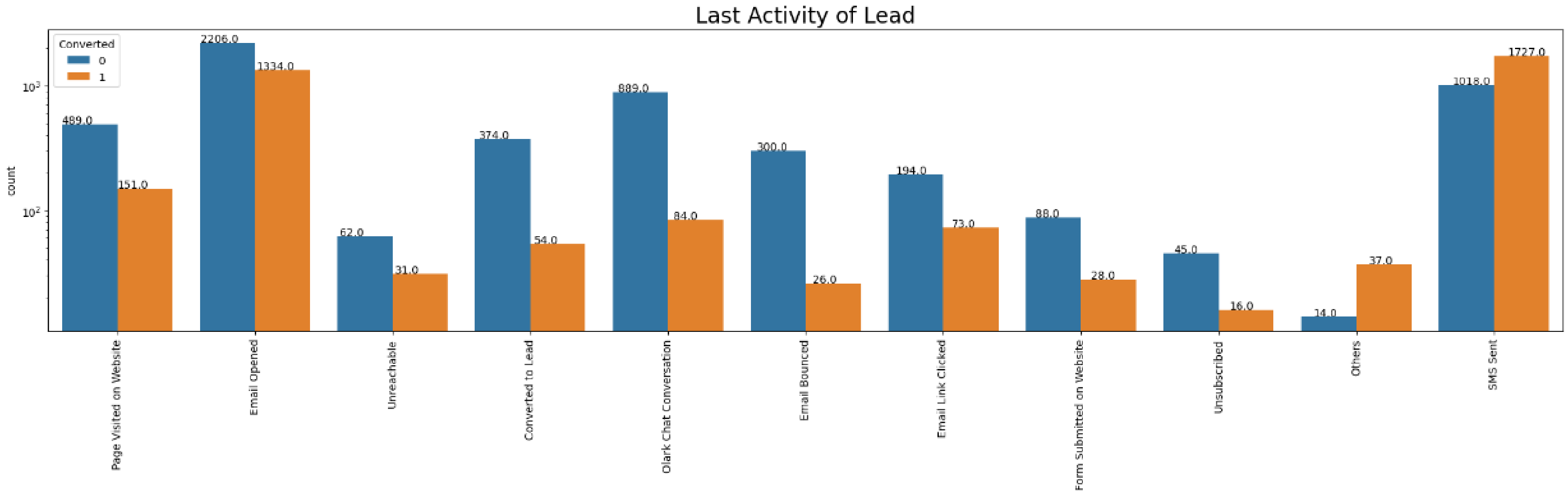




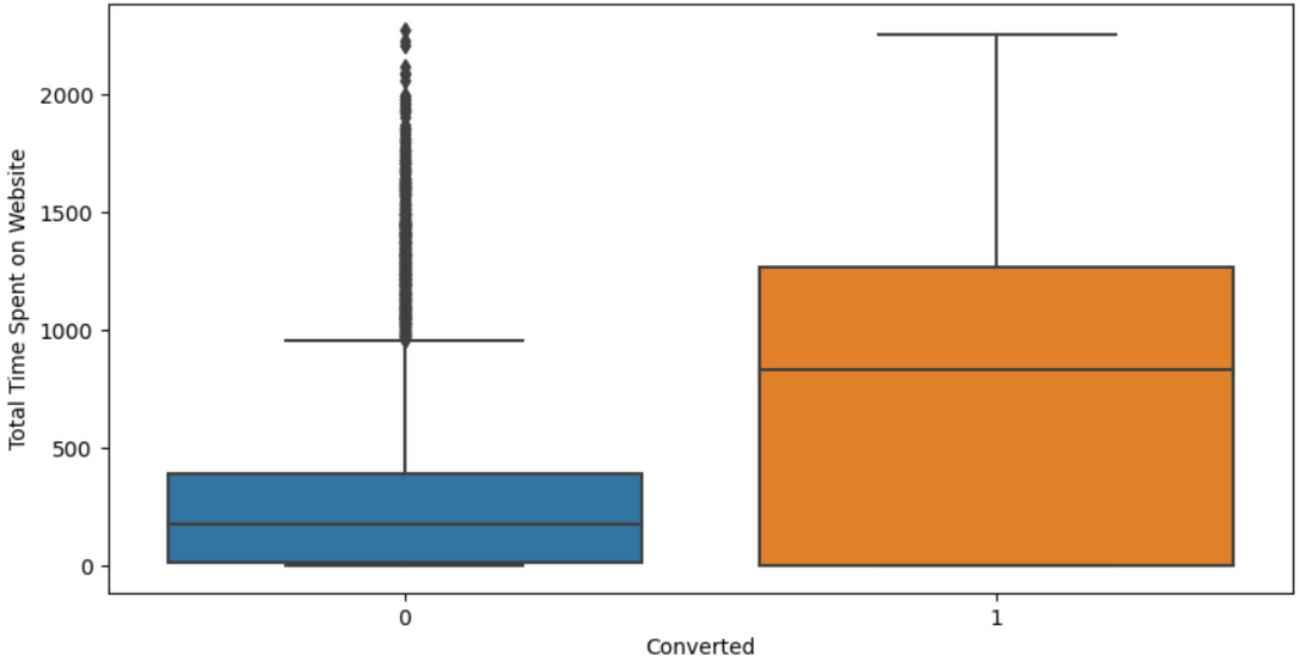
- Google and Direct traffic generates maximum number of leads.
- Conversion rate of 'Reference' and 'Welingak Website' leads is high.



- People who opted for mail option are becoming more leads

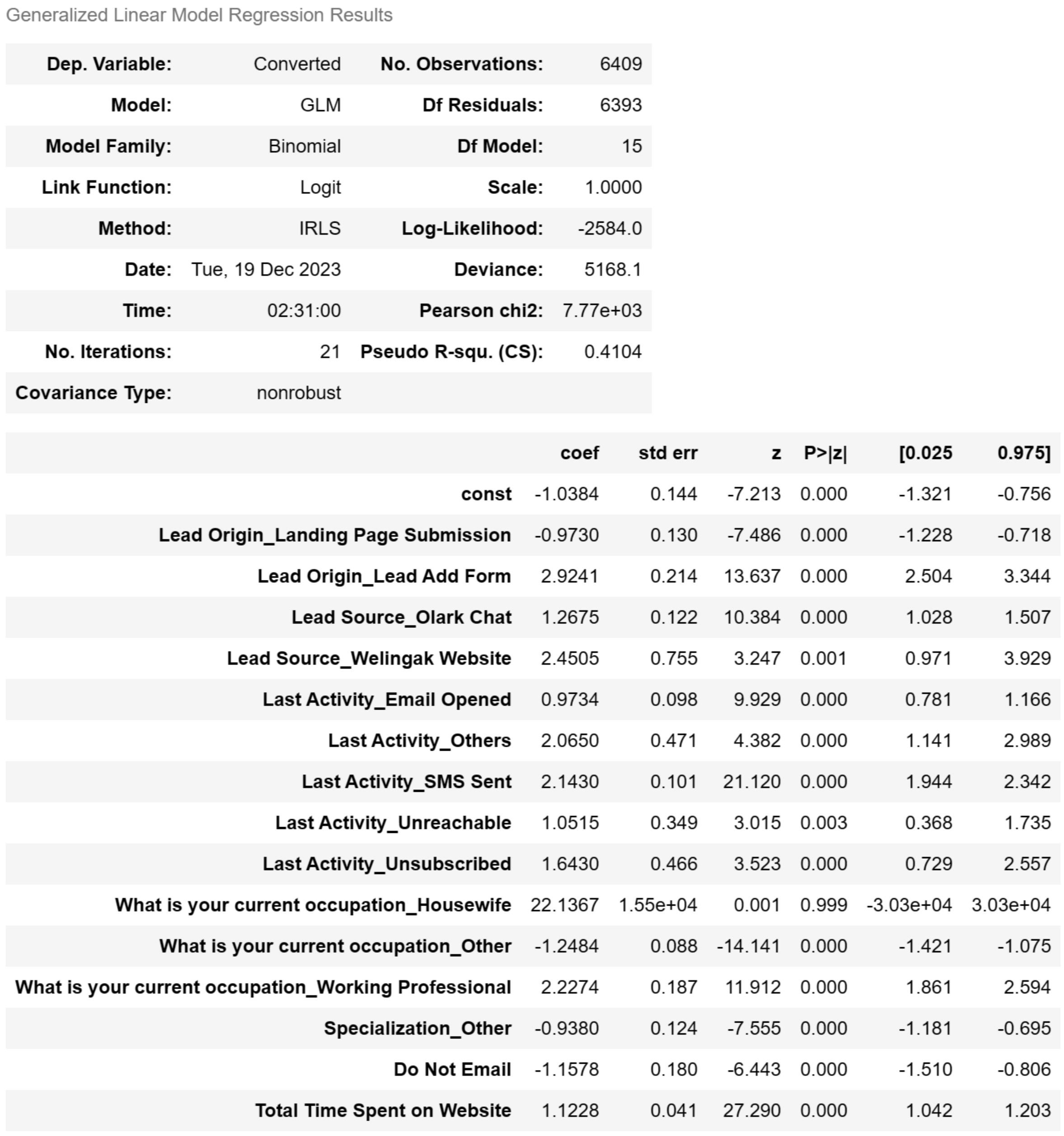


- Conversion rate for last activity of 'SMS Sent'is ~63%.
- Highest last activity of leads is 'Email Opened'.



- Leads spending more time on website are more likely to opt for curses or converted.

Model Evaluation



Generalized Linear Model Regression Results

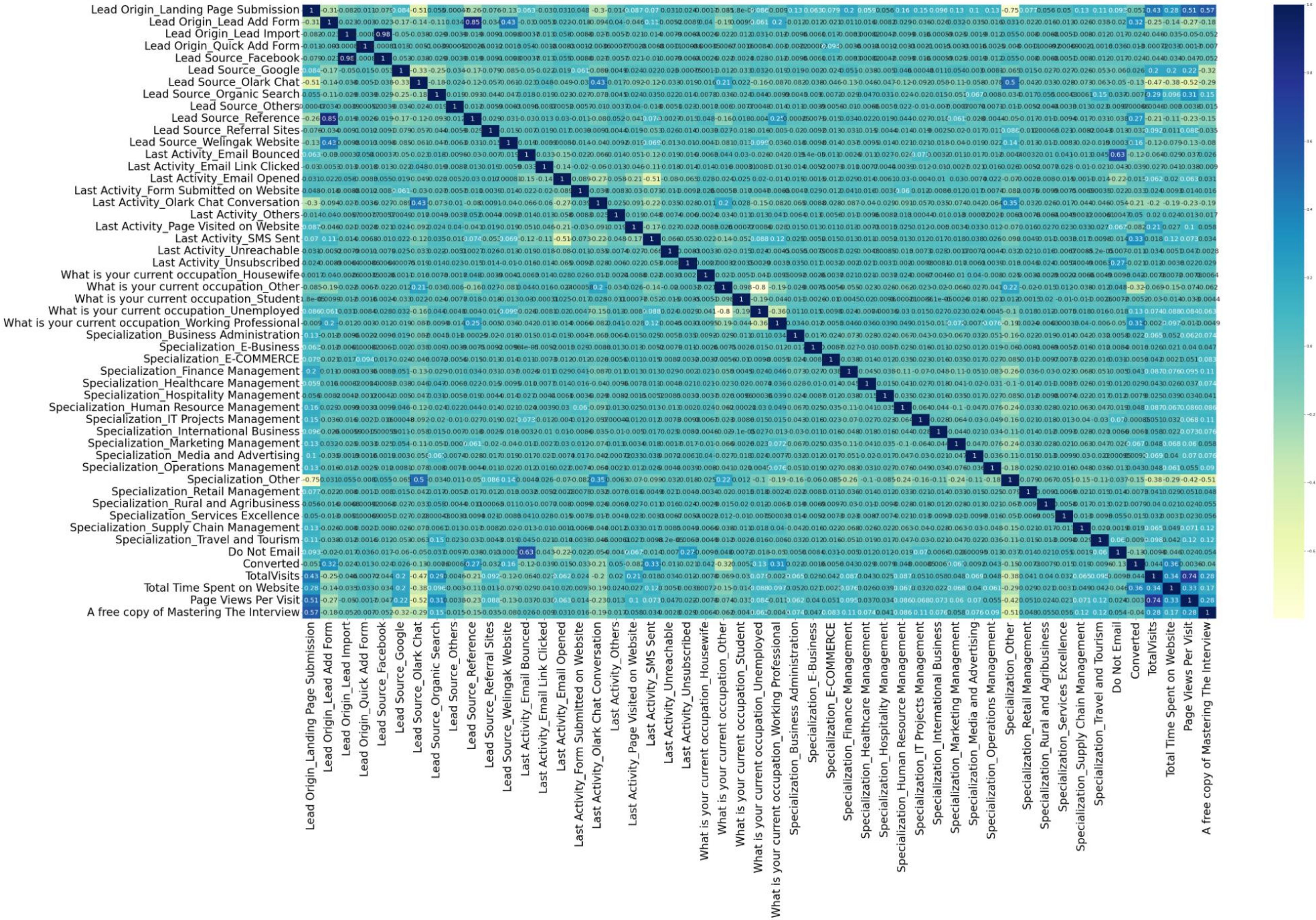
Dep. Variable:	Converted	**No. Observations:**	6409
Model:	GLM	**Df Residuals:**	6393
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2584.0
Date:	Tue, 19 Dec 2023	**Deviance:**	5168.1
Time:	02:31:00	**Pearson chi2:**	7.77e+03
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.4104
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-1.0384	0.144	-7.213	0.000	-1.321	-0.756
Lead Origin_Landing Page Submission	-0.9730	0.130	-7.486	0.000	-1.228	-0.718
Lead Origin_Lead Add Form	2.9241	0.214	13.637	0.000	2.504	3.344
Lead Source_Olark Chat	1.2675	0.122	10.384	0.000	1.028	1.507
Lead Source_Welingak Website	2.4505	0.755	3.247	0.001	0.971	3.929
Last Activity_Email Opened	0.9734	0.098	9.929	0.000	0.781	1.166
Last Activity_Others	2.0650	0.471	4.382	0.000	1.141	2.989
Last Activity_SMS Sent	2.1430	0.101	21.120	0.000	1.944	2.342
Last Activity_Unreachable	1.0515	0.349	3.015	0.003	0.368	1.735
Last Activity_Unsubscribed	1.6430	0.466	3.523	0.000	0.729	2.557
What is your current occupation_Housewife	22.1367	1.55e+04	0.001	0.999	-3.03e+04	3.03e+04
What is your current occupation_Other	-1.2484	0.088	-14.141	0.000	-1.421	-1.075
What is your current occupation_Working Professional	2.2274	0.187	11.912	0.000	1.861	2.594
Specialization_Other	-0.9380	0.124	-7.555	0.000	-1.181	-0.695
Do Not Email	-1.1578	0.180	-6.443	0.000	-1.510	-0.806
Total Time Spent on Website	1.1228	0.041	27.290	0.000	1.042	1.203

Final Model Summary :-
All p-values are zero

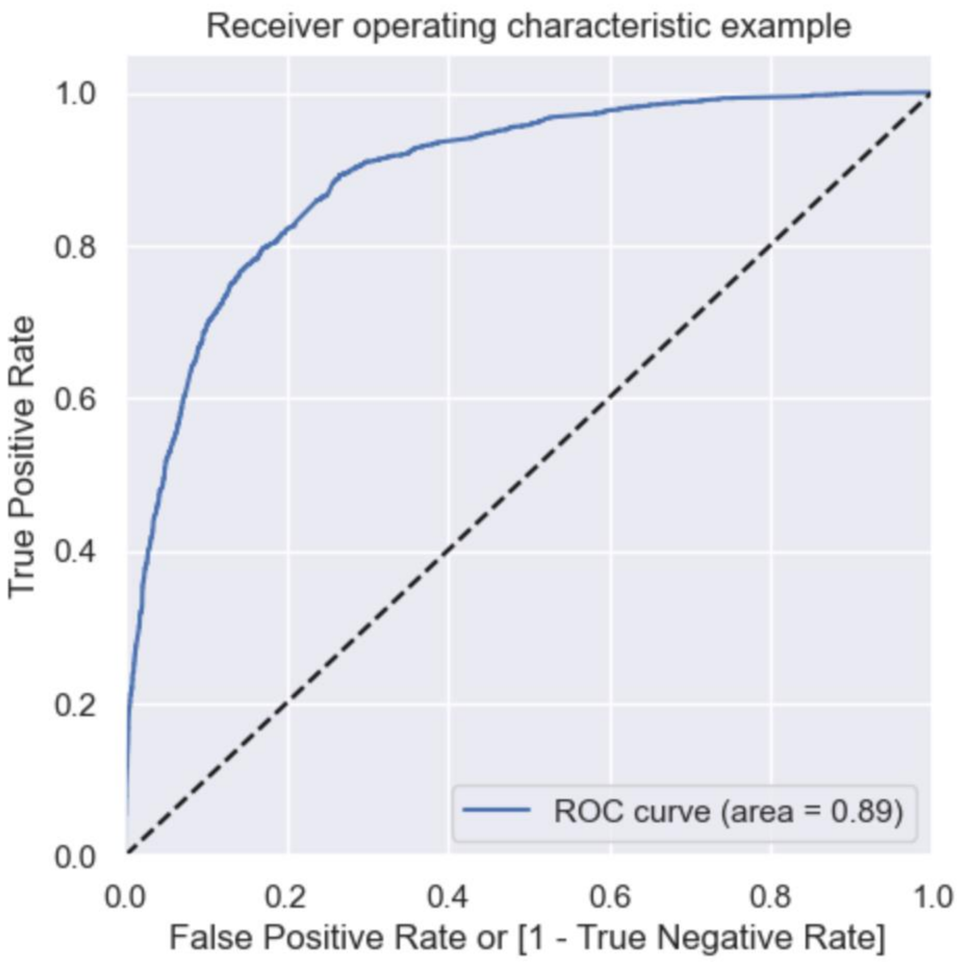
Correlation Heatmap

- The heatmap clearly shows which all variable are multicollinear in nature, and which variable have high collinearity with the target variable
- 'Lead Source_Facebook' and 'Lead Origin_Lead Import' having higher correlation of 0.98.
- 'Do Not Email' and 'Last Activity_Email Bounced' having higher correlation.
- 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' having higher correlation of 0.85.
- 'TotalVisits' and 'Page Views Per Visit' having correlation of 0.72.



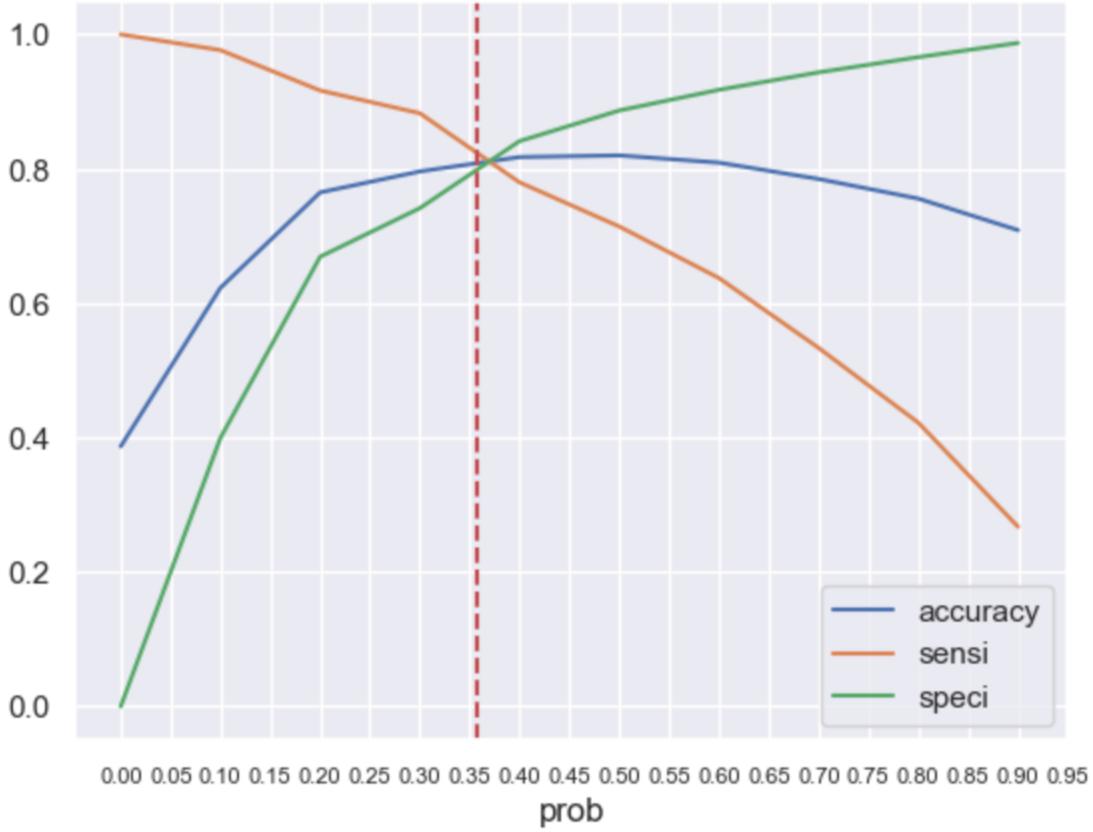
ROC Curve

- Getting a good value of 0.89 indicating a good predictive model.As ROC Curve should be a value close to 1.

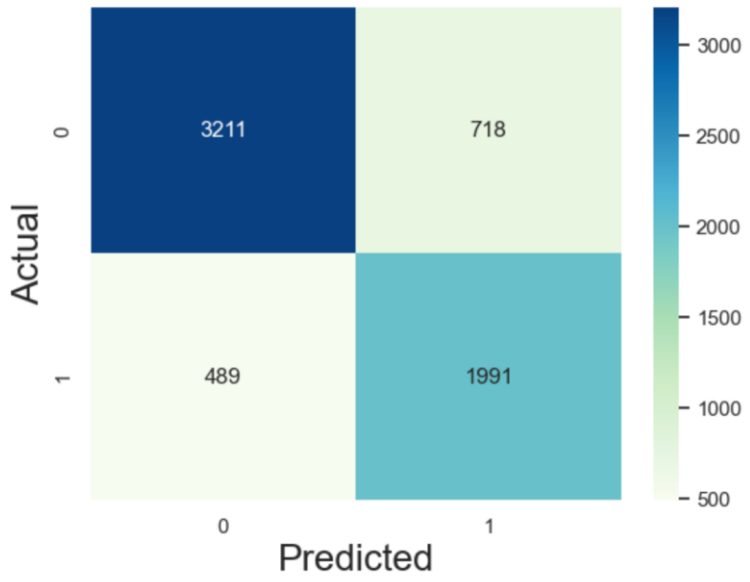


Optimal Threshold Value

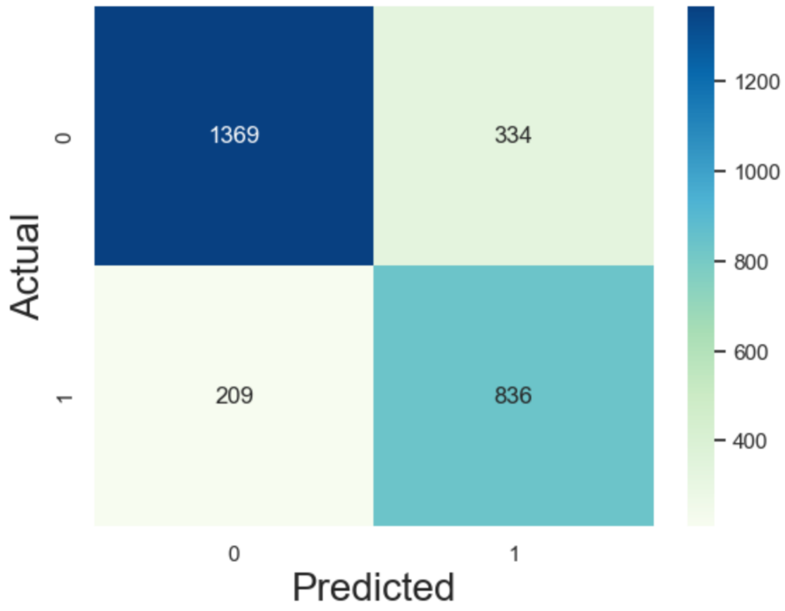
- Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values
- Optimal cutoff = 0.358



Confusion Matrix

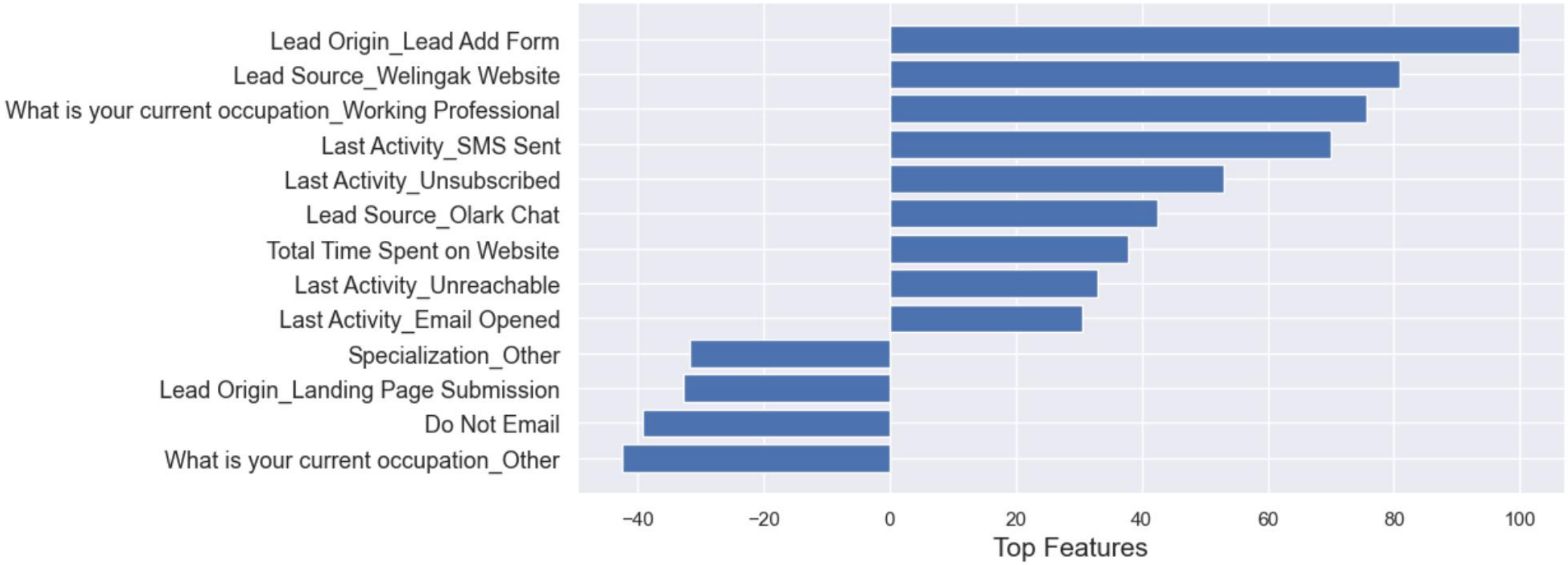


- For train set



- For test set

Feature Importance and Conclusion



- **Lead Source_Welingak Website : As conversion rate is higher for those leads who got to know about course from 'Welingak Website',so company can focus on this website to get more number of potential leads.**
- **Lead Origin_Lead Add Form: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it to get more number of leads cause have a higher chances of getting converted.**
- **Last Activity_SMS Sent: Lead whose last activity is sms sent can be potential lead for company.**
- **Total Time Spent on website: Leads spending more time on website can be the potential lead.**

Thank you!