LEAD SCORING CASE STUDY

Praveen Sankadal

PROBLEM STATEMENT

X Education, an online education company, caters to industry professionals by offering a variety of courses. Each day, numerous professionals visit the company's website, expressing interest in the courses available.

The courses are promoted on various platforms, including external websites and search engines like Google. Upon reaching the website, visitors may explore courses, fill out forms, or engage with educational content such as videos. When visitors provide their contact information by filling out a form, they are categorized as leads. Additionally, leads can originate from past referrals.

Once leads are acquired, the sales team initiates communication through calls, emails, and other means. However, only a fraction of these leads ultimately convert into customers, with the typical conversion rate at X Education being around 30%.

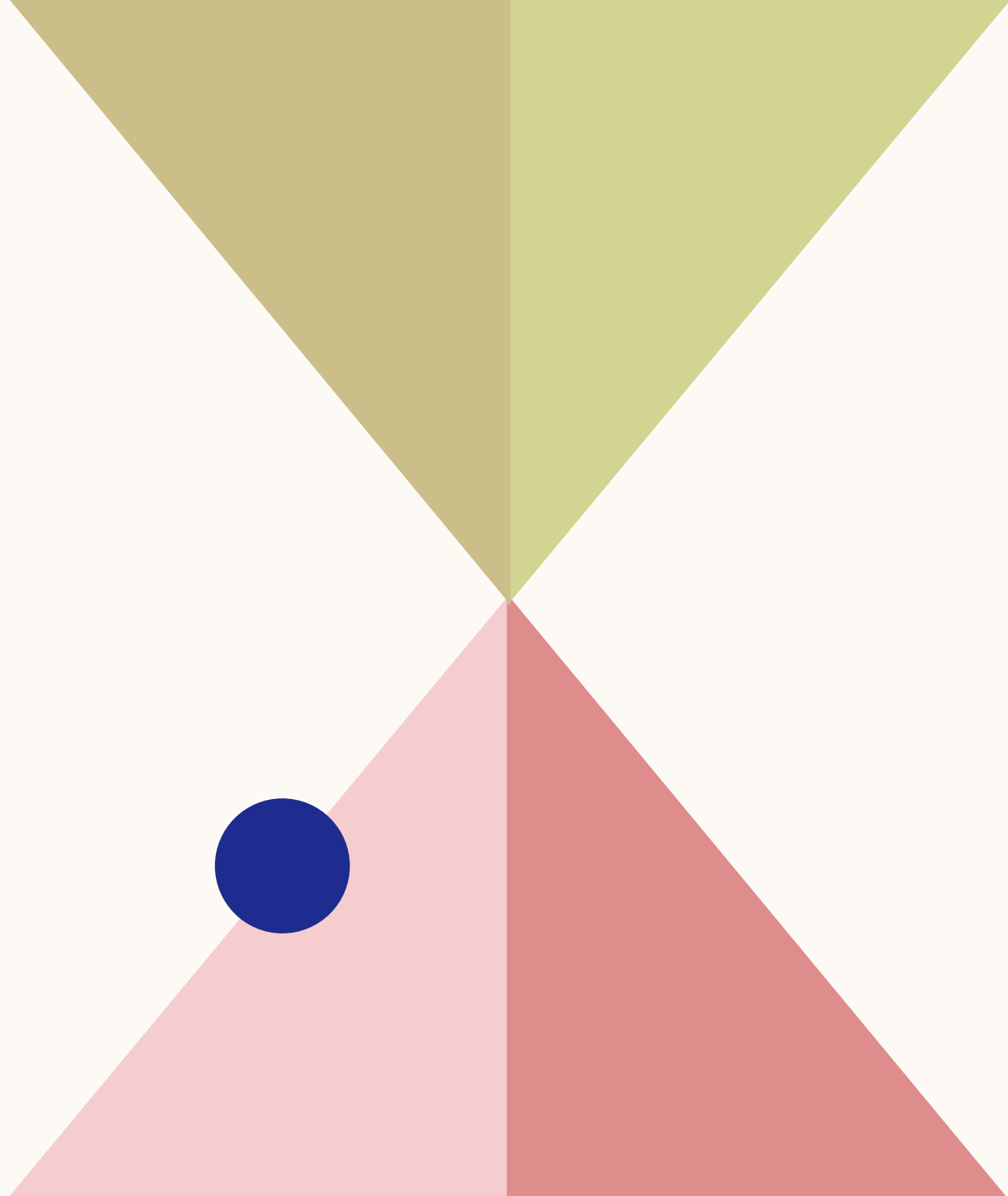
BUSINESS OBJECTIVE

The business objective is to develop a model that assigns a lead score to each potential customer. The lead score should indicate the likelihood of a lead converting into a customer, with higher scores corresponding to a higher conversion chance and lower scores indicating a lower conversion probability.

The CEO has set a target lead conversion rate of approximately 80%, and the model is expected to assist in identifying leads with a greater potential for conversion.

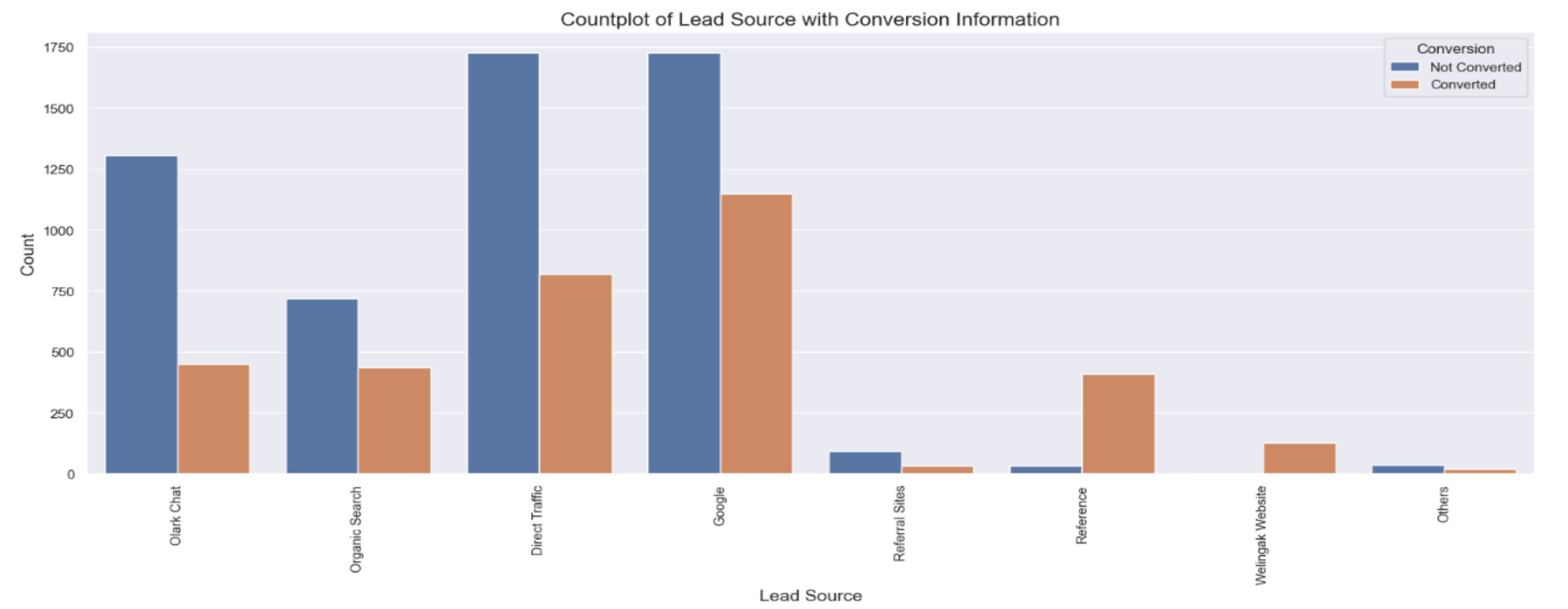
STEPS

1. Data Reading and understanding
2. Data cleaning and manipulation
3. Exploratory Data Analysis
4. Data Preparation
5. Model Building
6. Model Evaluation
7. Comparison with PAC
8. Assigning Lead Score



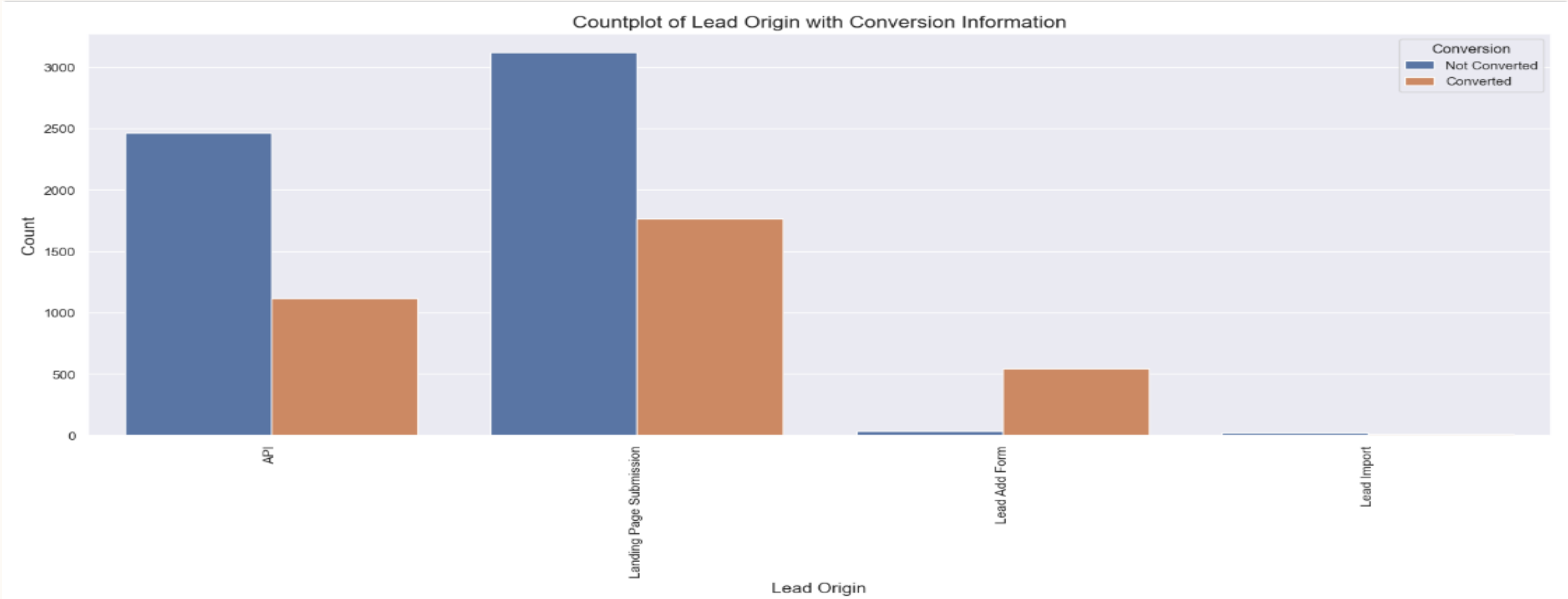
LEAD SOURCE

- Lead sources such as 'Reference' and 'Welingak Website' exhibit exceptionally high conversion rates. Additionally, the majority of leads originate from sources like 'Direct Traffic' and 'Google'.



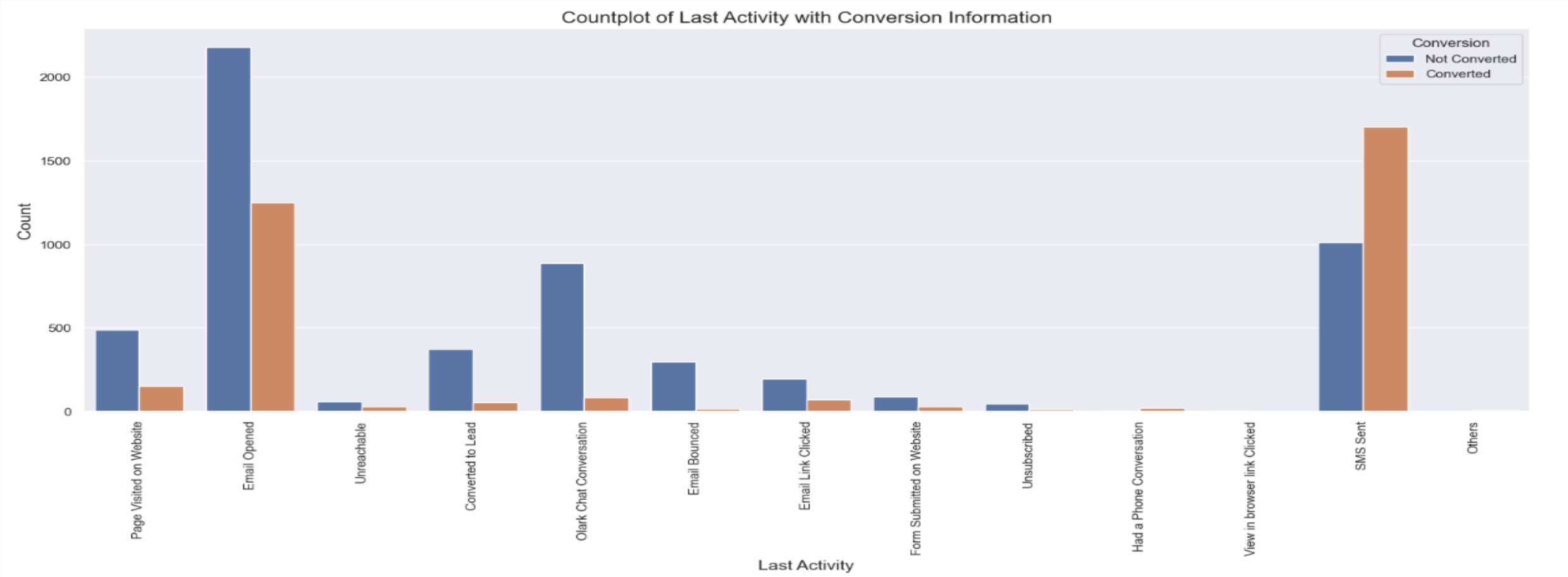
LEAD ORIGIN

- 'API' and 'Landing Page Submission' generate the most leads but have lower conversion rates.
- 'Lead Add Form' generates fewer leads but boasts a higher conversion rate.



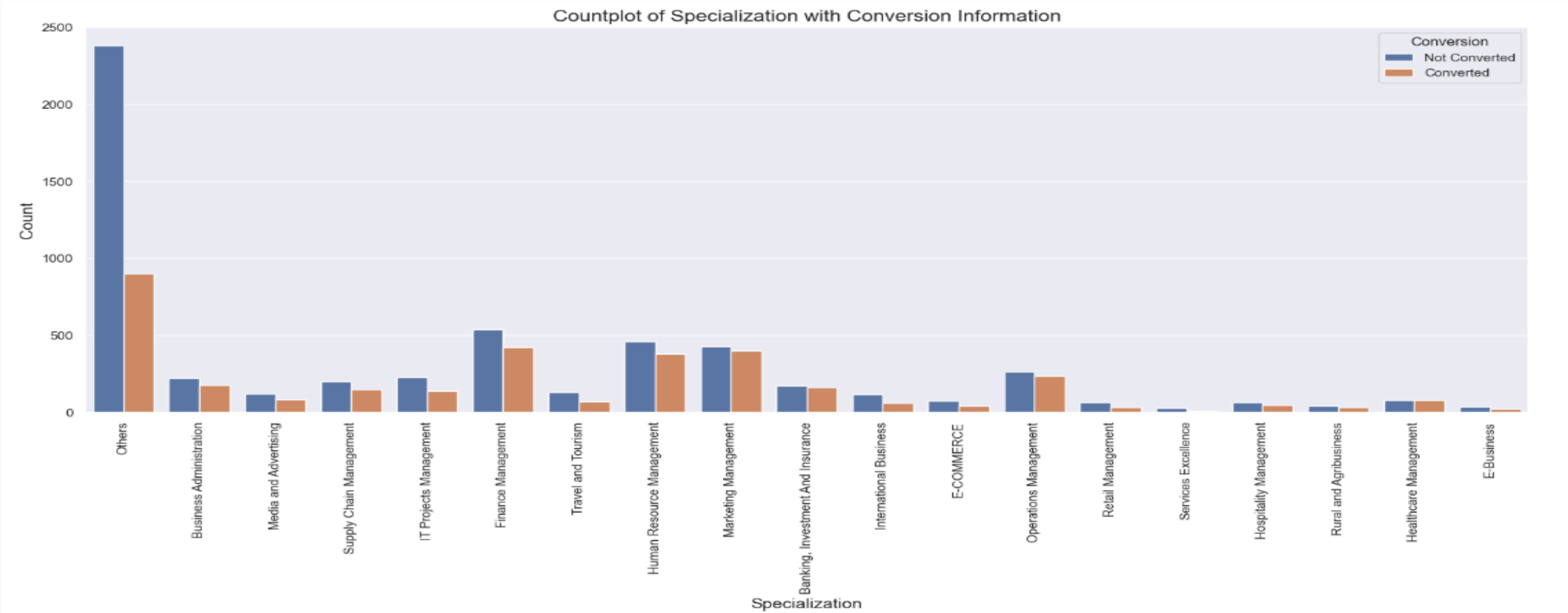
LAST ACTIVITY

- Lead activities like 'Olark chat conversation' and 'Email Link Opened' show remarkably high conversion rates. Additionally, the majority of leads are generated through activities such as 'SMS Sent' and 'Email Opened'.



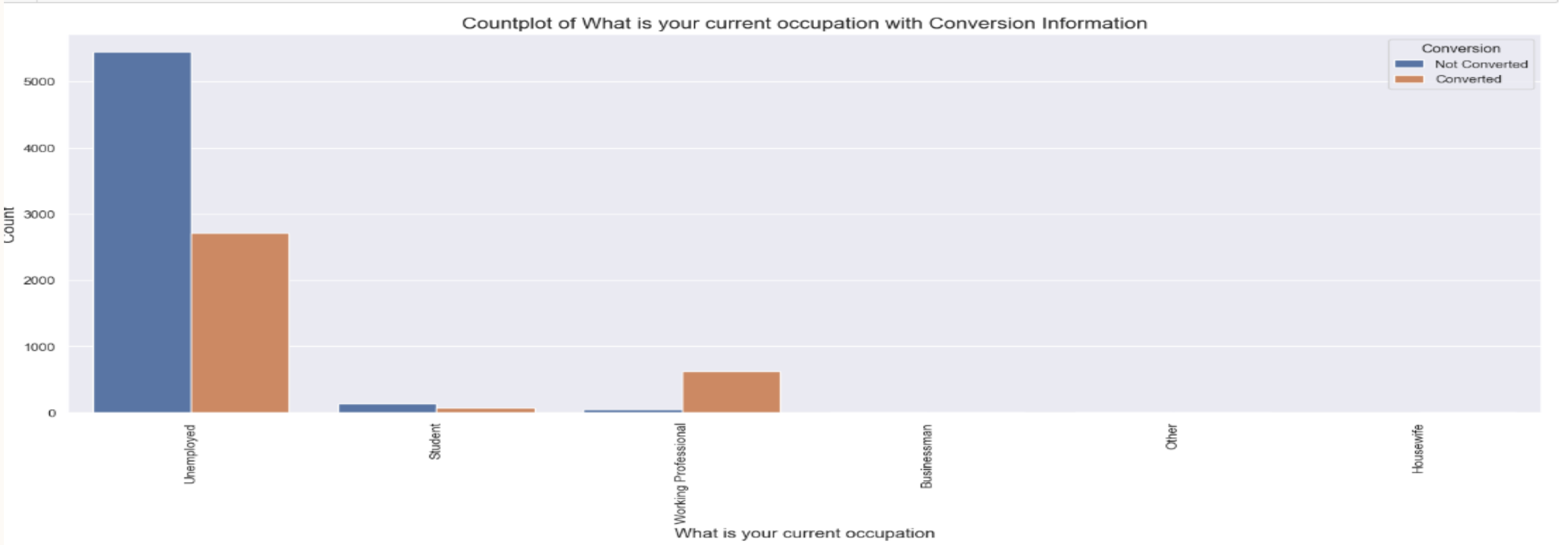
SPECIALIZATION

- Most leads are generated by people who are specialized in Human Resource management and finance management follower by marketing and operational management.



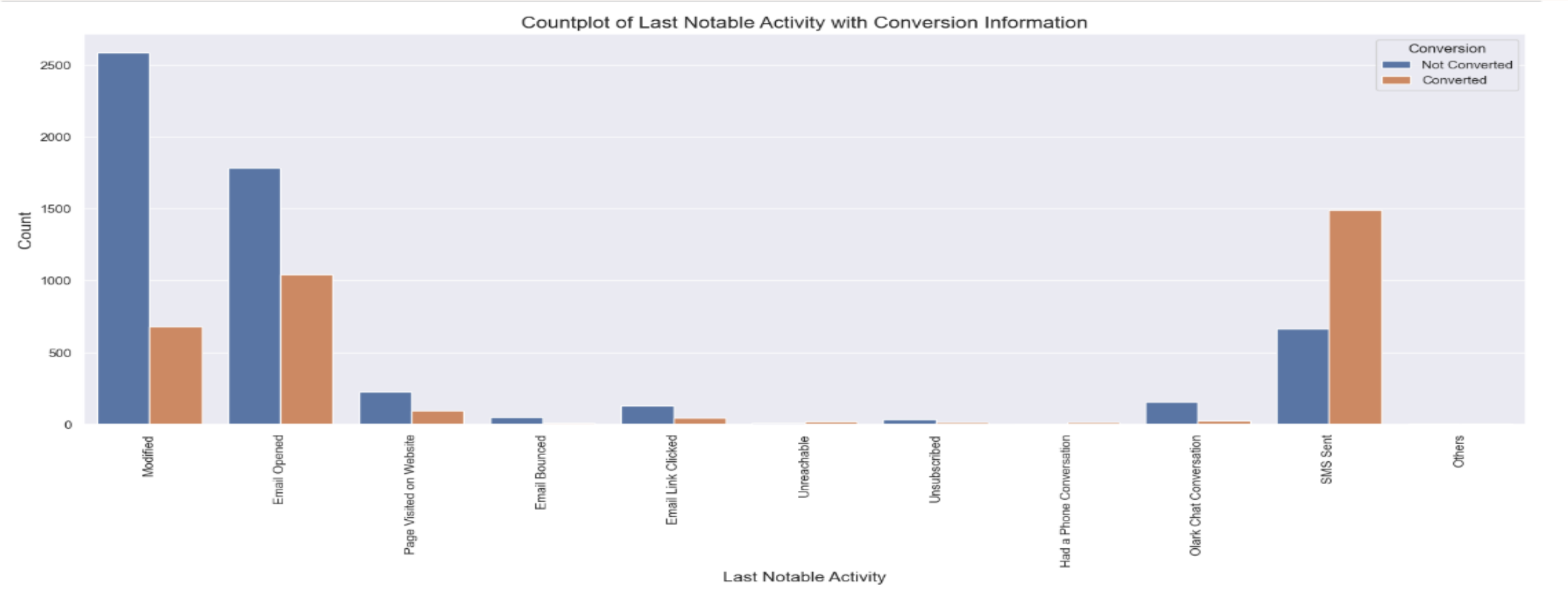
OCCUPATION

- Most conversion are seen in people how are in working profession, followed by students
- Most leads were generated by unemployed people.



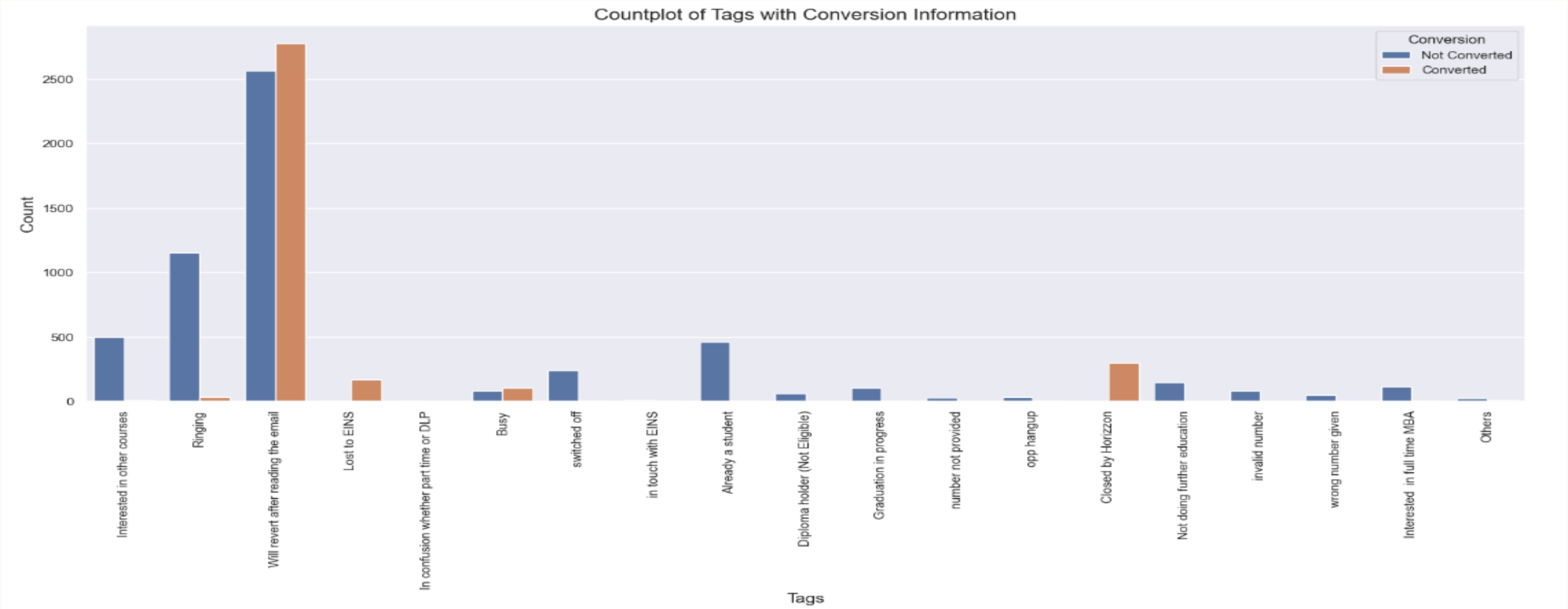
LAST NOTABLE ACTIVITY

- The last notable activity with the highest conversion rate is 'SMS Sent'.



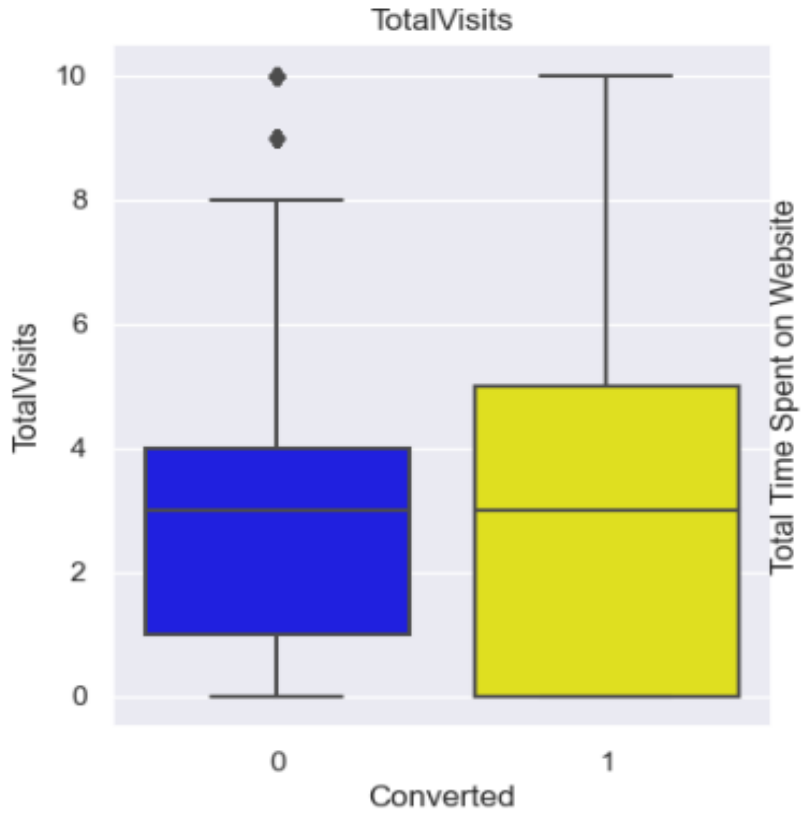
TAGS

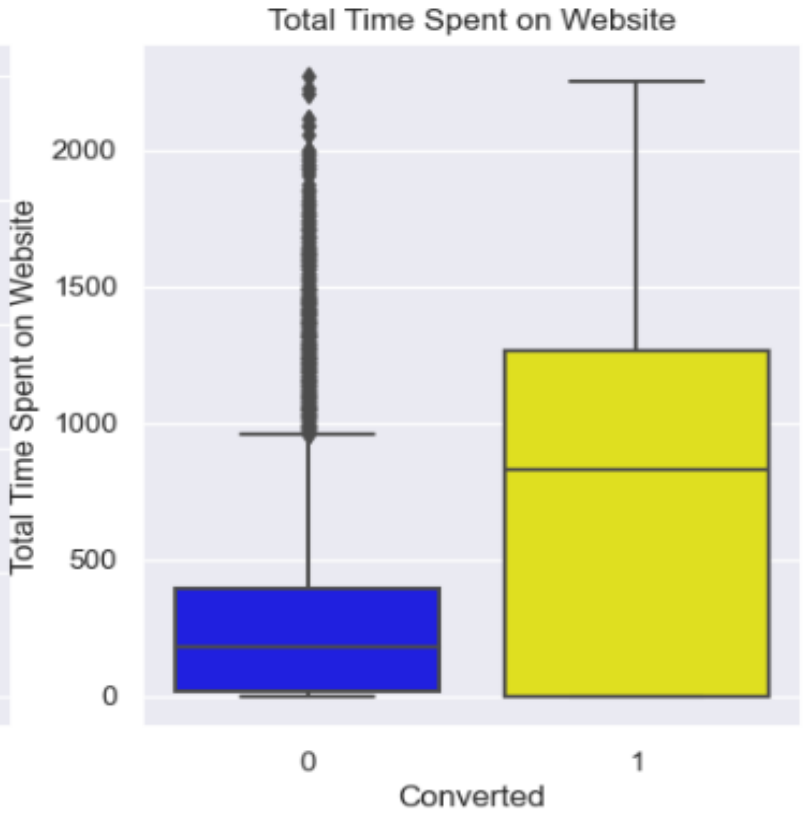
- High conversion rates observed for tags: 'Will revert after reading the email', 'Closed by Horizon', 'Lost to EINS', and 'Busy'.

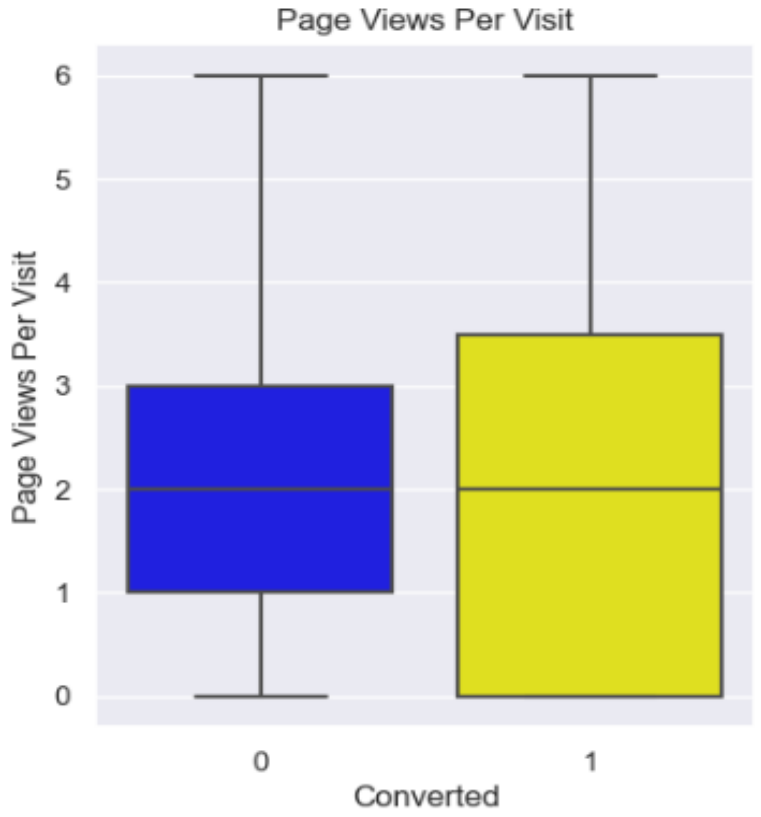


WEBSITE INSIGHTS

- Most leads were generated by people who have visited the most, spent more time on website.

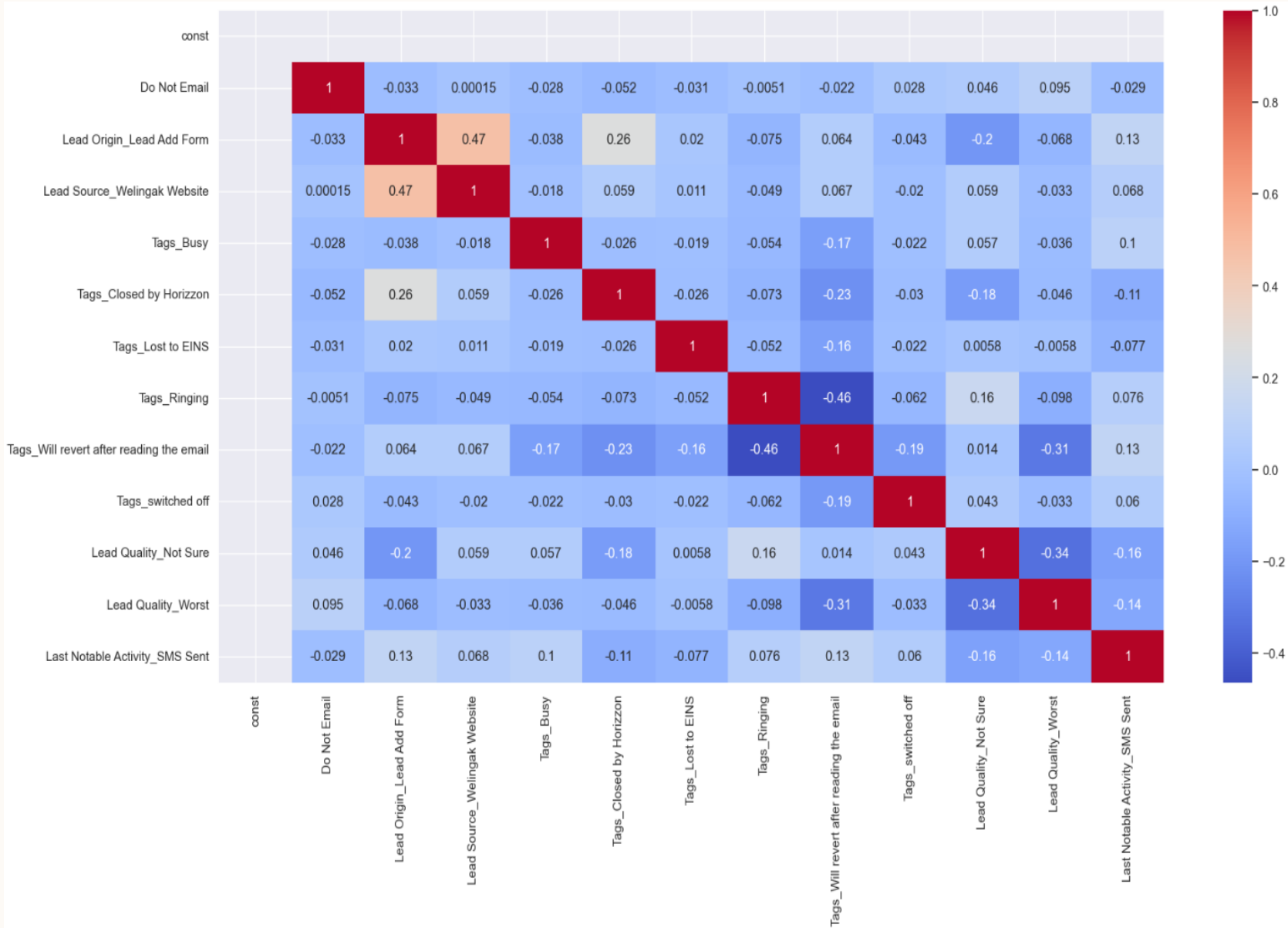






HEAT MAP

- Negligible correlations observed among features in the final model.



MODEL EVALUATION

- All the p-values are less then 0.05 and VIF value less then 5.

Generalized Linear Model Regression Results

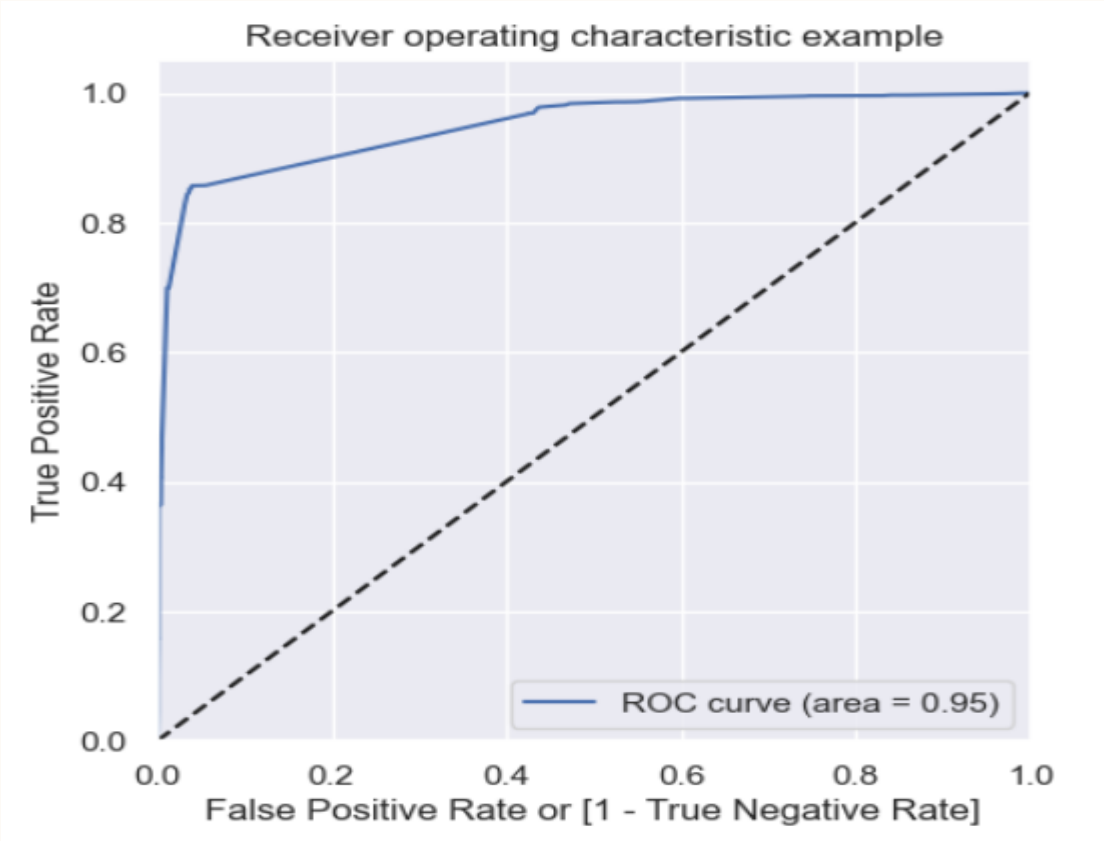
Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6338
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1601.0
Date:	Mon, 18 Dec 2023	Deviance:	3202.0
Time:	00:58:37	Pearson chi2:	3.48e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.5635
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9192	0.211	-9.080	0.000	-2.333	-1.505
Do Not Email	-1.2835	0.212	-6.062	0.000	-1.698	-0.868
Lead Origin_Lead Add Form	1.2035	0.368	3.267	0.001	0.482	1.925
Lead Source_Welingak Website	3.2825	0.820	4.002	0.000	1.675	4.890
Tags_Busy	3.8043	0.330	11.525	0.000	3.157	4.451
Tags_Closed by Horizzon	7.9789	0.762	10.467	0.000	6.485	9.473
Tags_Lost to EINS	9.1948	0.753	12.209	0.000	7.719	10.671
Tags_Ringing	-1.8121	0.336	-5.401	0.000	-2.470	-1.154
Tags_Will revert after reading the email	3.9906	0.228	17.508	0.000	3.544	4.437
Tags_switched off	-2.4456	0.586	-4.171	0.000	-3.595	-1.297
Lead Quality_Not Sure	-3.5218	0.126	-28.036	0.000	-3.768	-3.276
Lead Quality_Worst	-3.9106	0.856	-4.567	0.000	-5.589	-2.232
Last Notable Activity_SMS Sent	2.7395	0.120	22.907	0.000	2.505	2.974

	Features	VIF
9	Lead Quality_Not Sure	2.62
7	Tags_Will revert after reading the email	2.57
1	Lead Origin_Lead Add Form	1.58
6	Tags_Ringing	1.52
11	Last Notable Activity_SMS Sent	1.51
2	Lead Source_Welingak Website	1.34
4	Tags_Closed by Horizzon	1.13
0	Do Not Email	1.10
3	Tags_Busy	1.10
8	Tags_switched off	1.10
5	Tags_Lost to EINS	1.04
10	Lead Quality_Worst	1.03

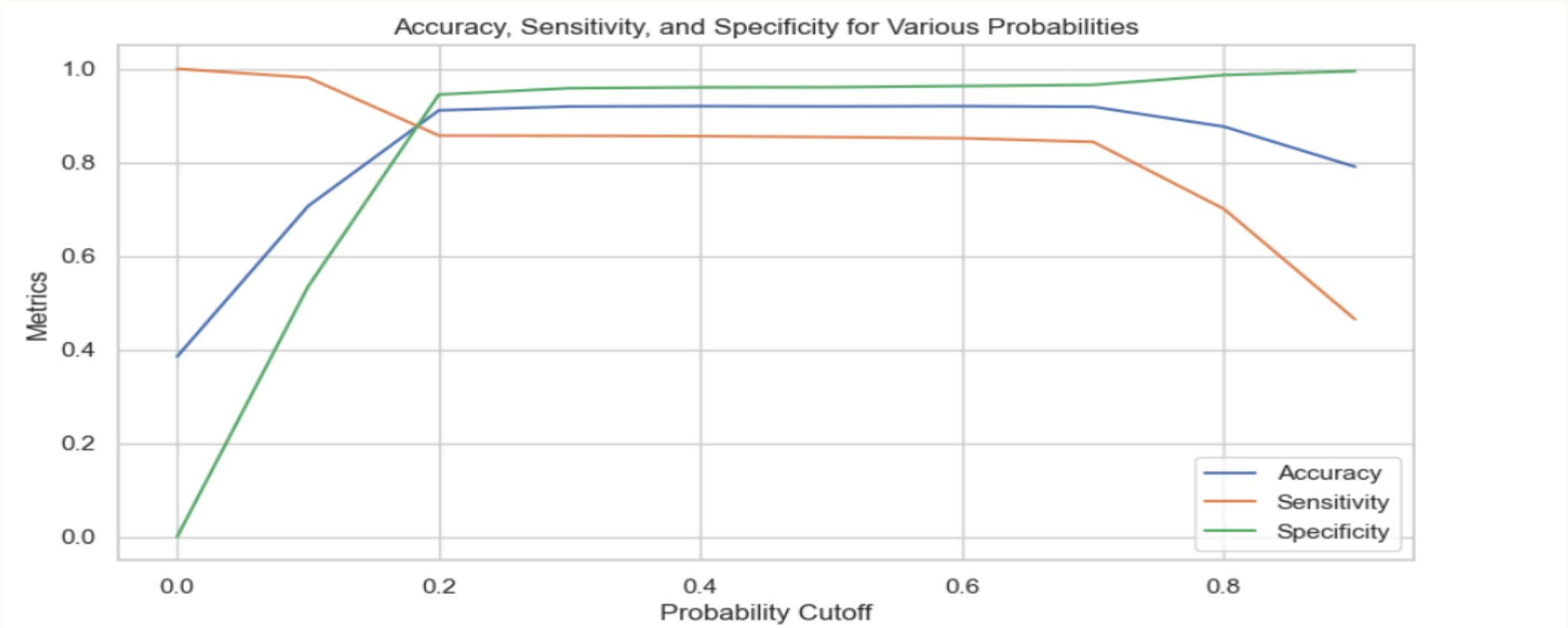
RECEIVER OPERATION CHARACTERISTIC

- The area under the curve (AUC) is around 0.95, closely approaching the ideal AUC of 1.



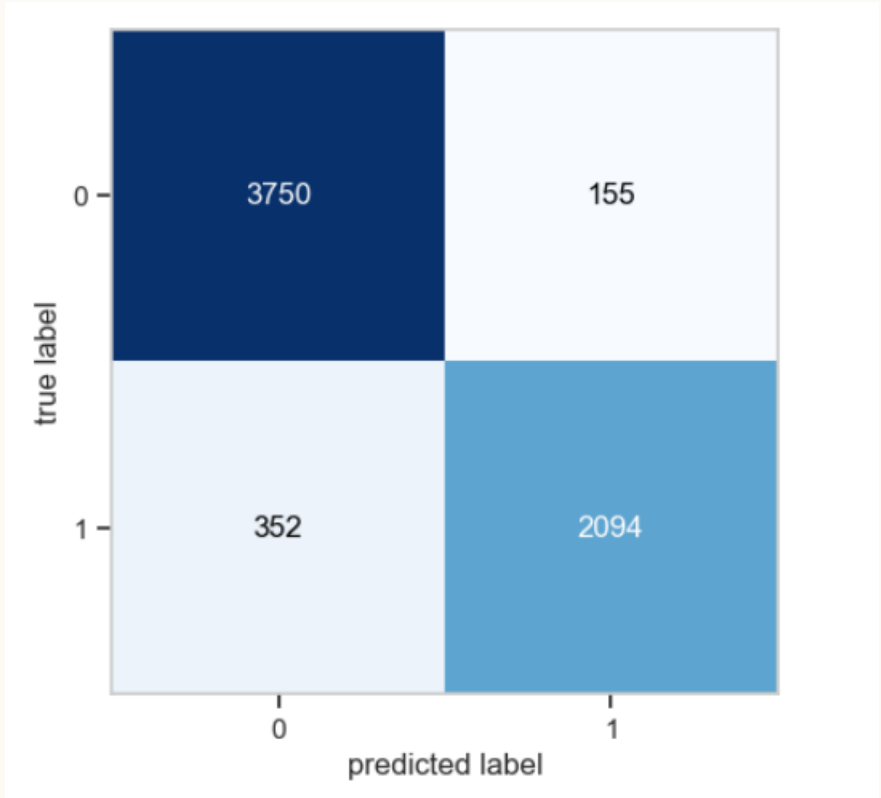
PROBABILITY CUTOFF

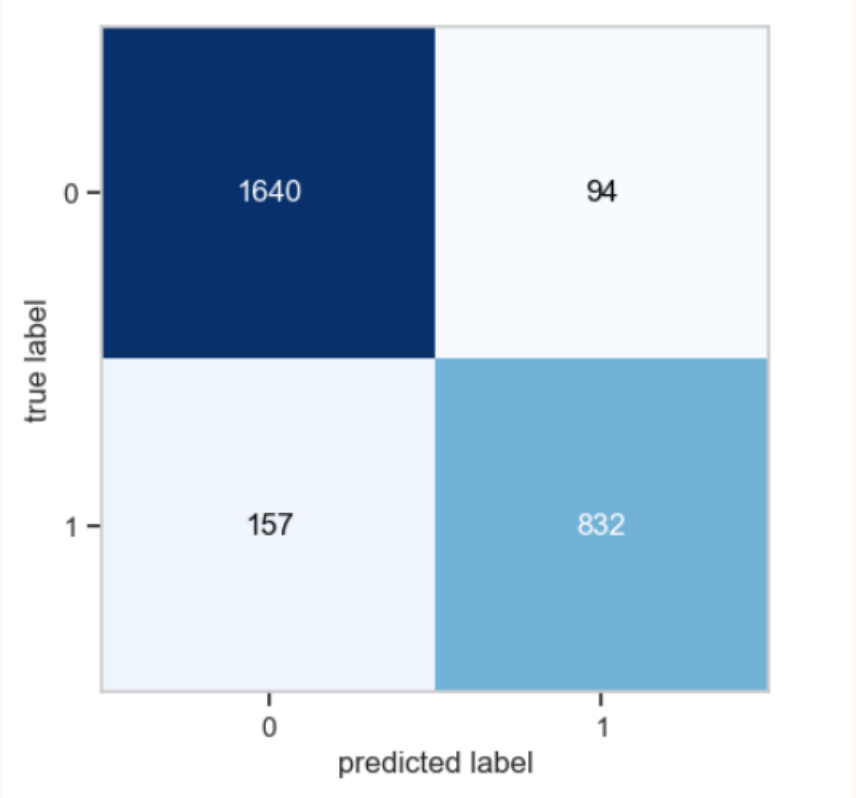
- Visual representation of Sensitivity, Specificity, and Accuracy changes with varying probability threshold values.
- Optimal cutoff value identified as 0.20.



CONFUSION MATRIX

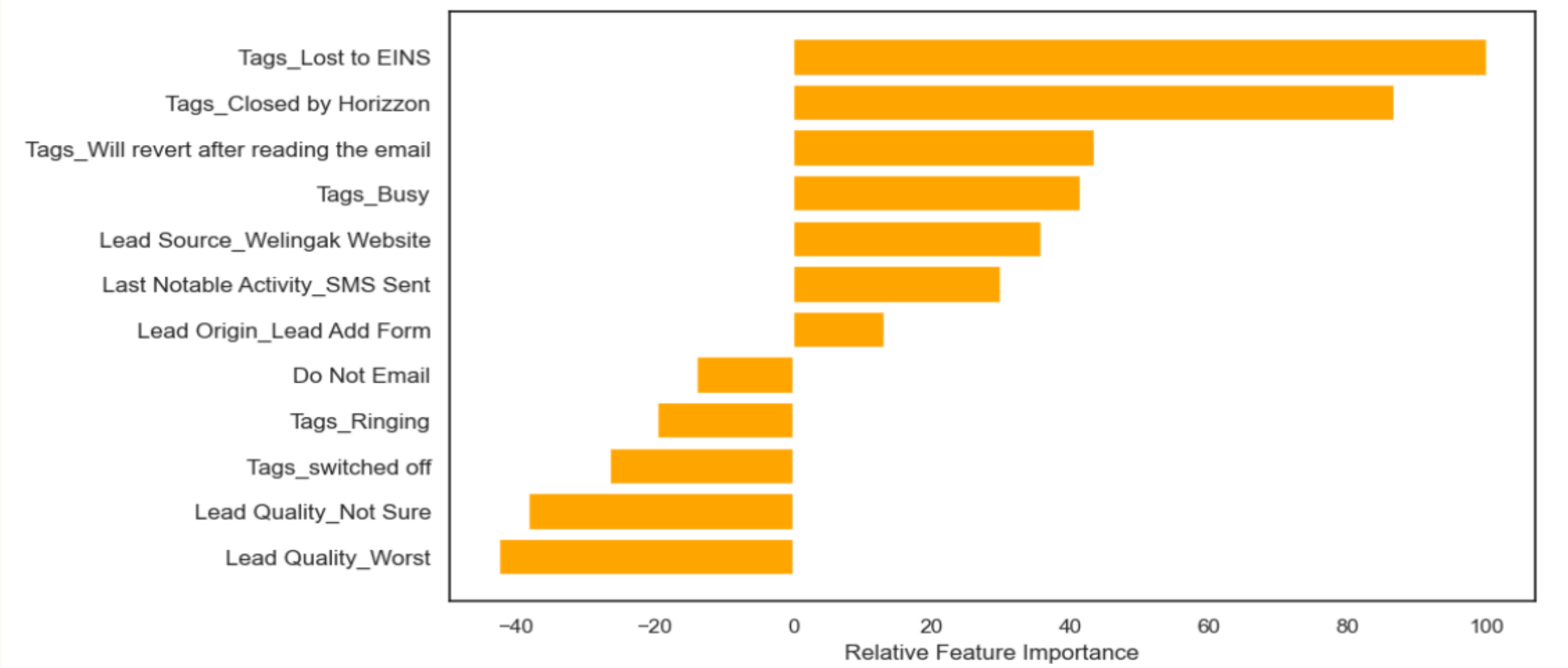
- Confusion matrix of train and test data set





RELATIVE FEATURE IMPORTANCE

- Identifying Key Features



FINAL RESULTS

Data	Train	Test
Accuracy	0.92	0.90
Sensitivity	0.86	0.84
Specificity	0.95	0.95
False Positive Rate	0.06	0.05
Positive Predictive Value	0.91	0.90
Negative Predictive Value	0.91	0.91
AUC	0.95	0.94

RECOMMENDATIONS

1. Examine the data visualizations to:
   - Enhance conversion rates for categories with higher lead generation.
   - Increase lead generation for categories with elevated conversion rates.

2. Consider the relative importance of model features and their impact on conversion probability.

3. Adjust the probability threshold according to specific business requirements.

THANK YOU

PRAVEEN SANKADAL