

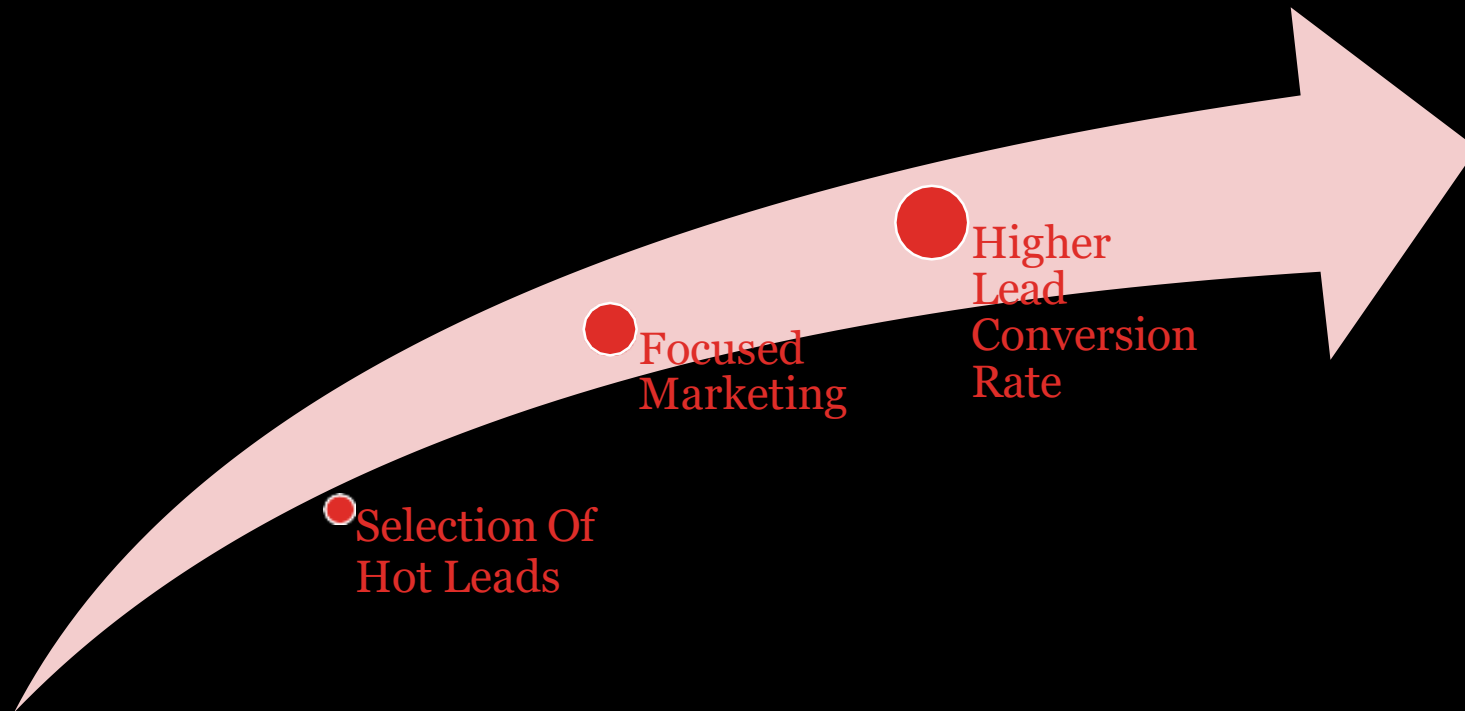
LEAD SCORING CASE STUDY

Focused business approach using logistic regression technique

Dipanshu Yadav


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.



METHODOLOGY

- 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 $\approx 80\%$



Importing and Observing
the past data provided by
the Company

Reading And
Understanding
The Data

Data Cleaning

- Missing value imputation
- Removing duplicate data and other redundancies

Univariate and
Bivariate analysis

EDA

Data
Preparation

- Outlier treatment
- Dropping unnecessary columns
- Dummy variable creation
- Feature standardization

- Feature selection using RFE
- Manual feature elimination based on p-values and VIFs

Model Building

Model
Evaluation

Assigning Lead
Score

- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores:
 $\text{Lead Score} = \text{Probability} * 100$

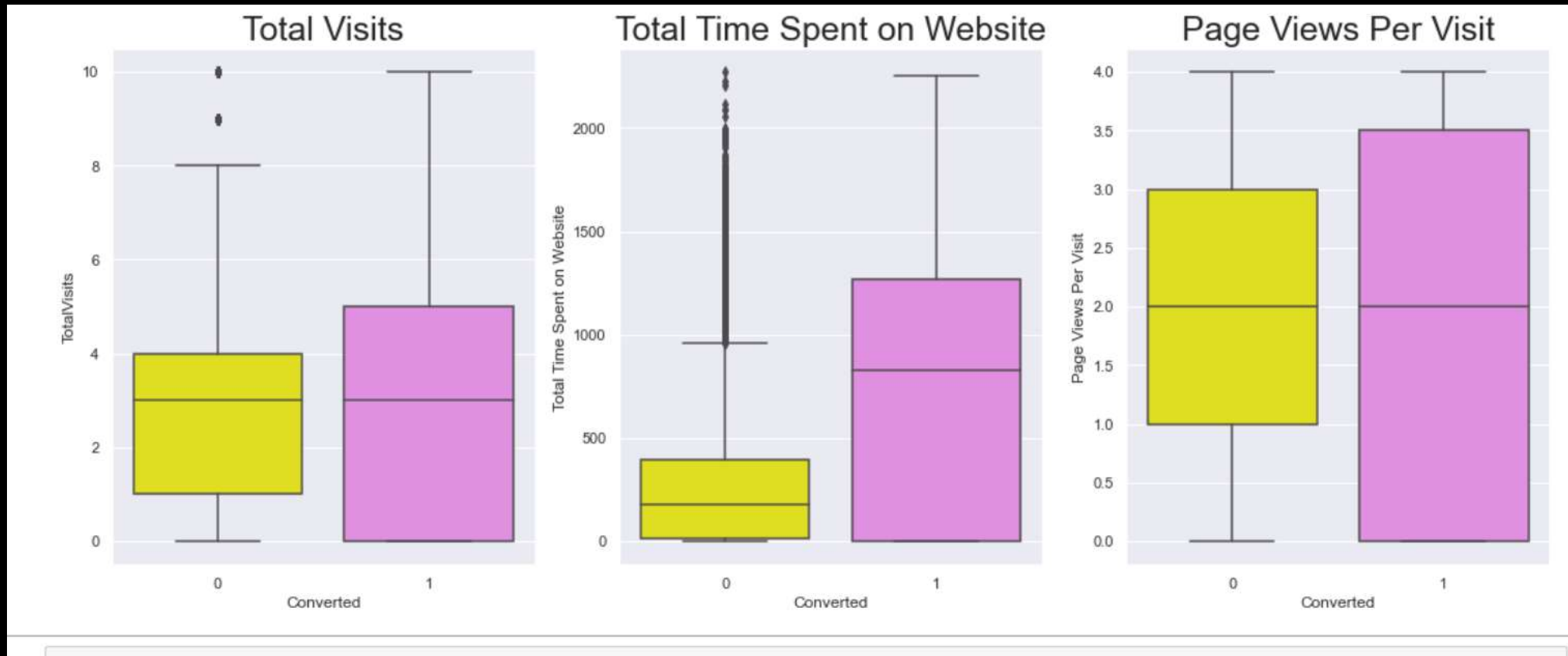
- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold



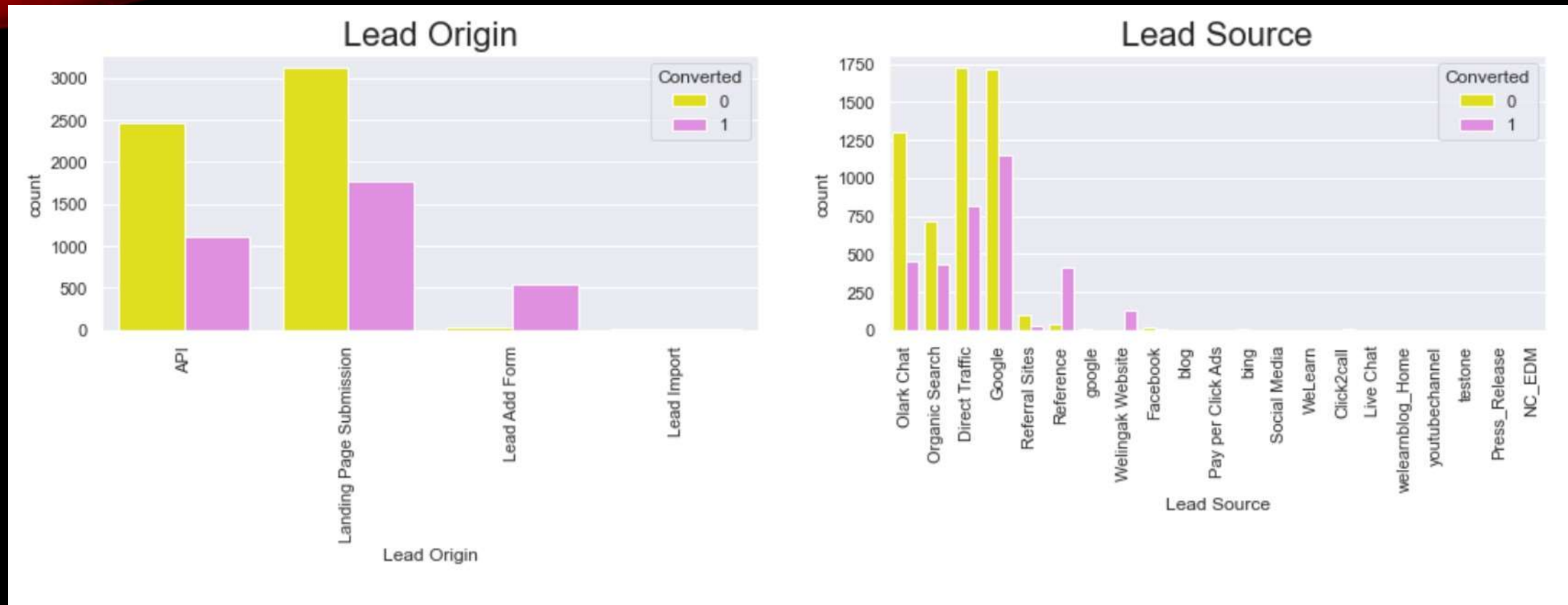
DATA VISUALIZATION

- TO IDENTIFY IMPORTANT FEATURES
- TO GET INSIGHTS

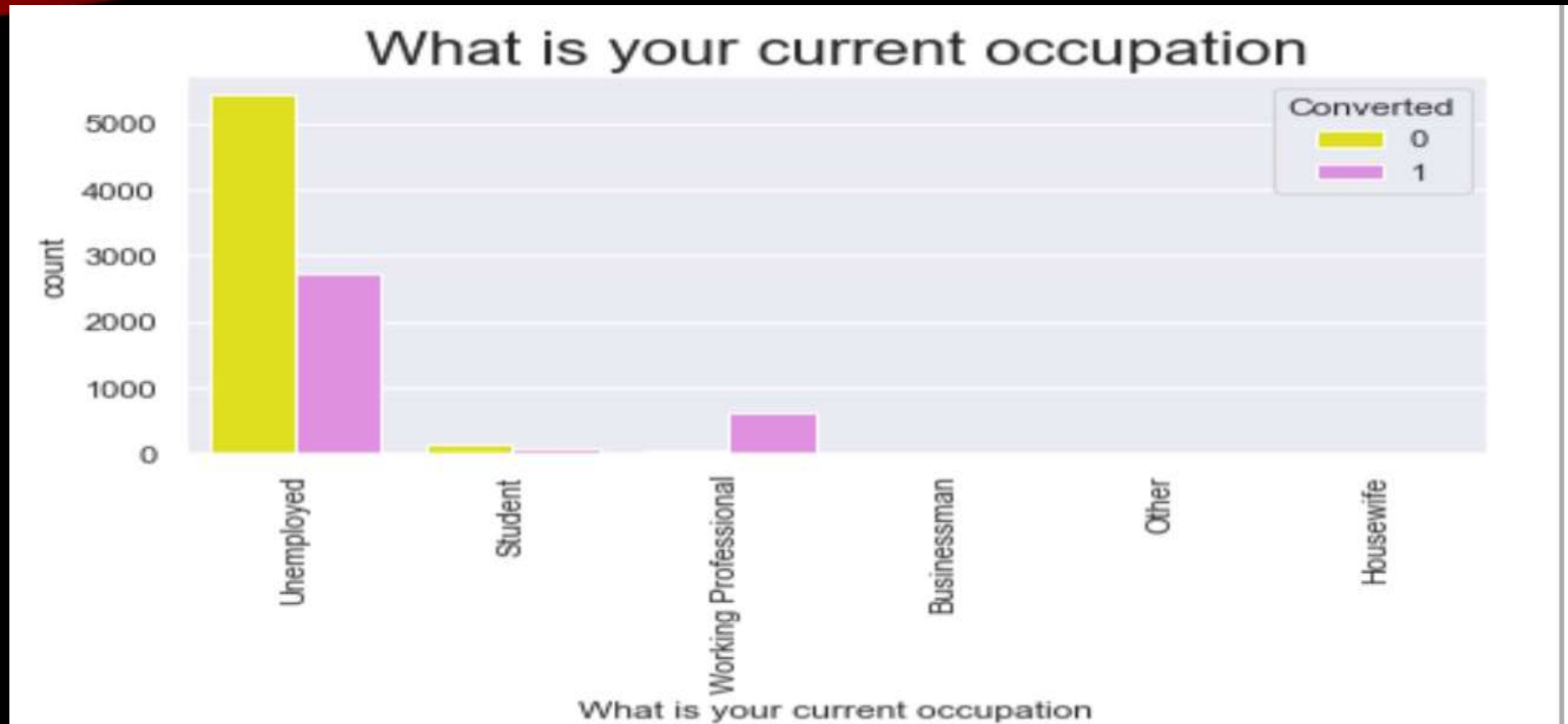
NUMERICAL VARIABLE



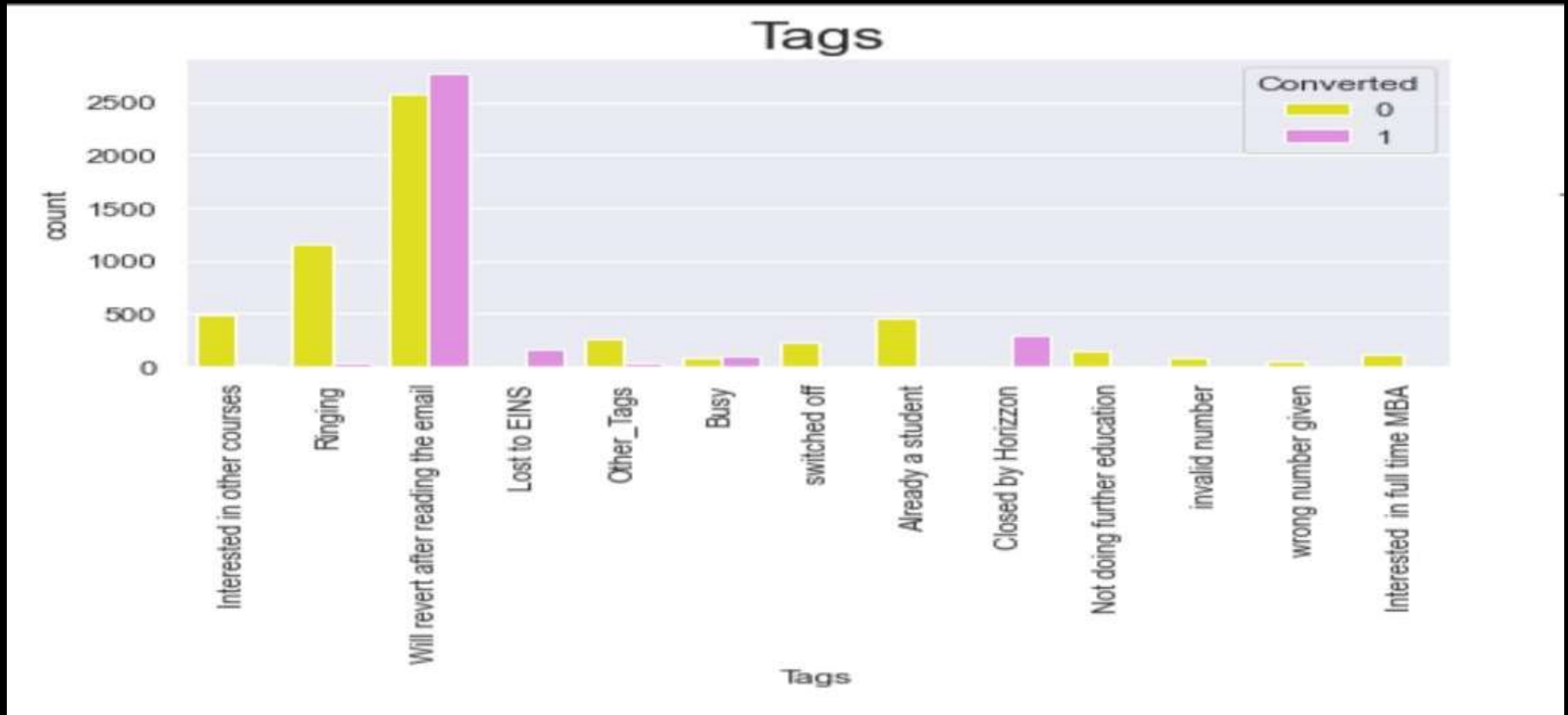
People spending more time on website are more likely to get converted.



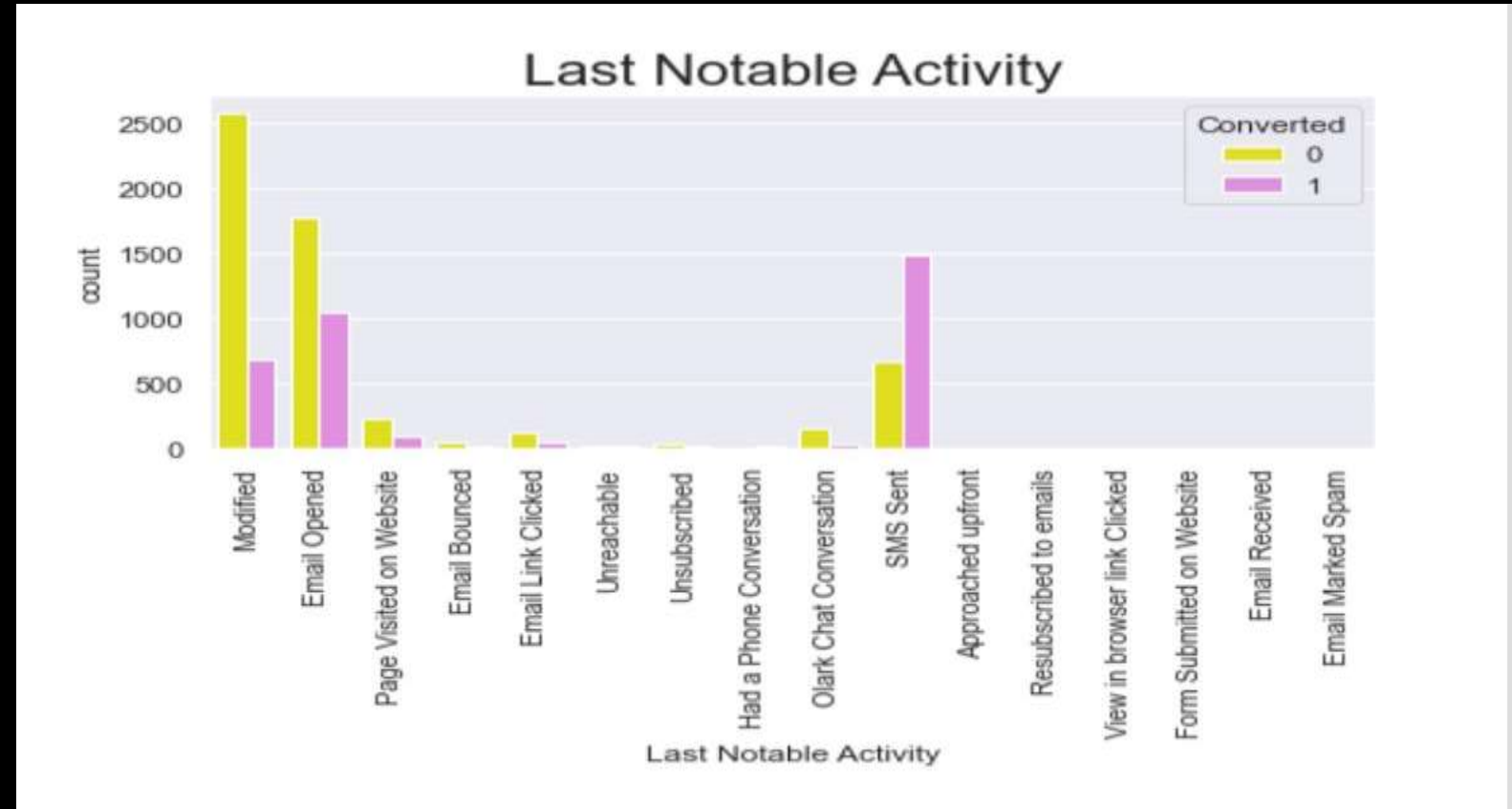
- 'API' and 'Landing Page Submission' generate the most leads but have less conversion rates, whereas 'Lead Add Form' generates less leads but conversion rate is great.
- Try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'.
- Very high conversion rates for lead sources 'Reference' and 'Welingak Website'.
- Most leads are generated through 'Direct Traffic' and 'Google'.



Working Professionals are most likely to get converted.



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



Highest conversion rate is for the last notable activity **'SMS Sent'**.



MODEL EVALUATION

FINAL MODEL SUMMARY

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:	-2026.8			
Date:	Mon, 05 Jun 2023	Deviance:	4053.5			
Time:	09:49:21	Pearson chi2:	9.40e+03			
No. Iterations:	8	Pseudo R-squ. (CS):	0.5008			
Covariance Type:	nonrobust					

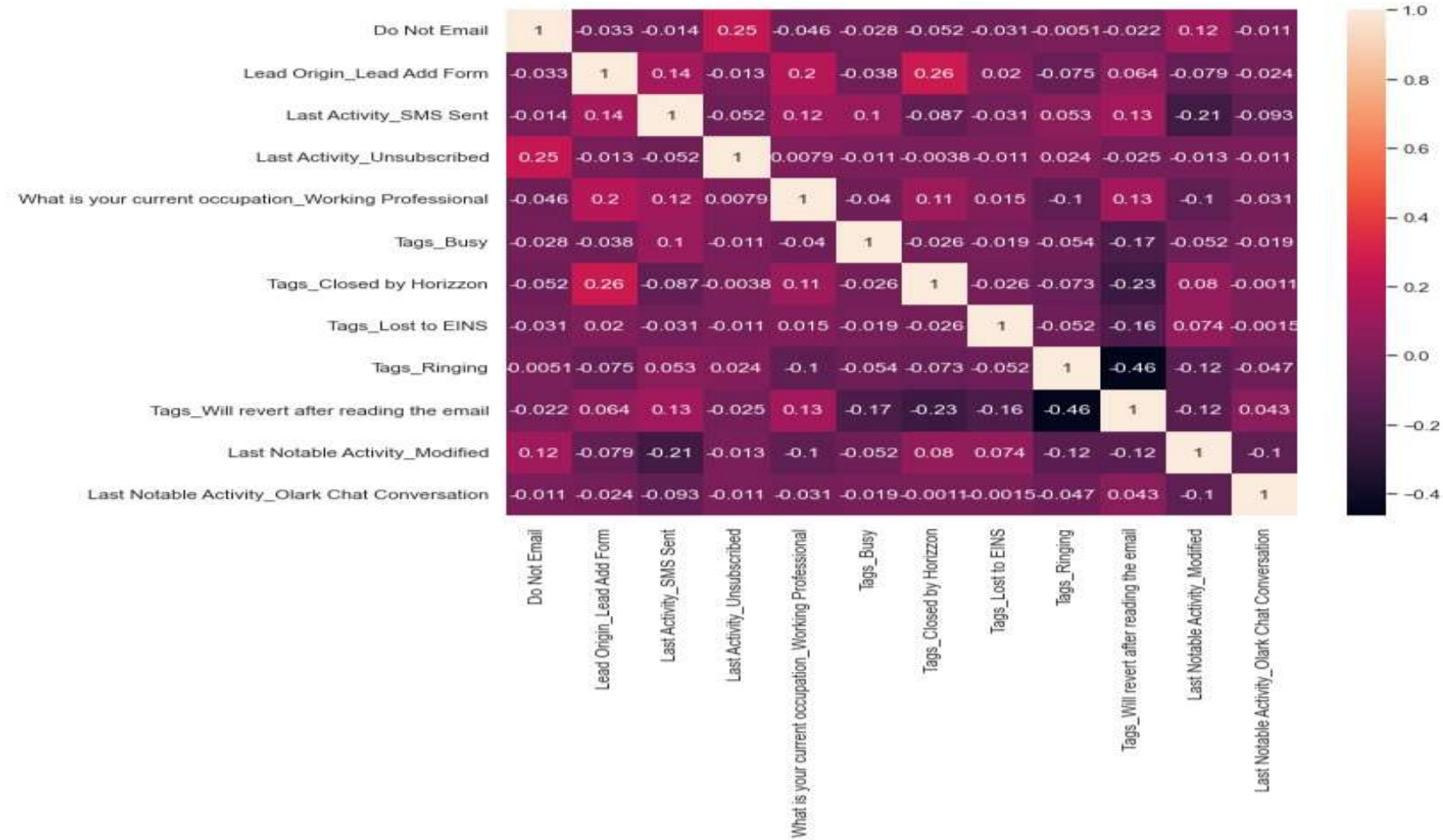
	coef	std err	z	P> z	[0.025	0.075]

const	-3.6165	0.182	-19.891	0.000	-3.973	-3.260
Do Not Email	-1.5648	0.203	-7.702	0.000	-1.963	-1.167
Lead Origin_Lead Add Form	2.4860	0.256	9.697	0.000	1.983	2.988
Last Activity_SMS Sent	1.9049	0.090	21.277	0.000	1.729	2.080
Last Activity_Unsubscribed	2.0214	0.546	3.704	0.000	0.952	3.091
What is your current occupation_Working Professional	2.9330	0.238	12.341	0.000	2.467	3.399
Tags_Busy	2.9578	0.275	10.759	0.000	2.419	3.497
Tags_Closed by Horizon	8.8703	0.741	11.965	0.000	7.417	10.323
Tags_Lost to EINS	8.6801	0.749	11.583	0.000	7.211	10.149
Tags_Ringing	-1.2686	0.298	-4.252	0.000	-1.853	-0.684
Tags_Will revert after reading the email	3.4863	0.181	19.241	0.000	3.131	3.841
Last Notable Activity_Modified	-1.7298	0.092	-18.805	0.000	-1.910	-1.550
Last Notable Activity_Olark Chat Conversation	-1.6854	0.314	-5.373	0.000	-2.300	-1.071

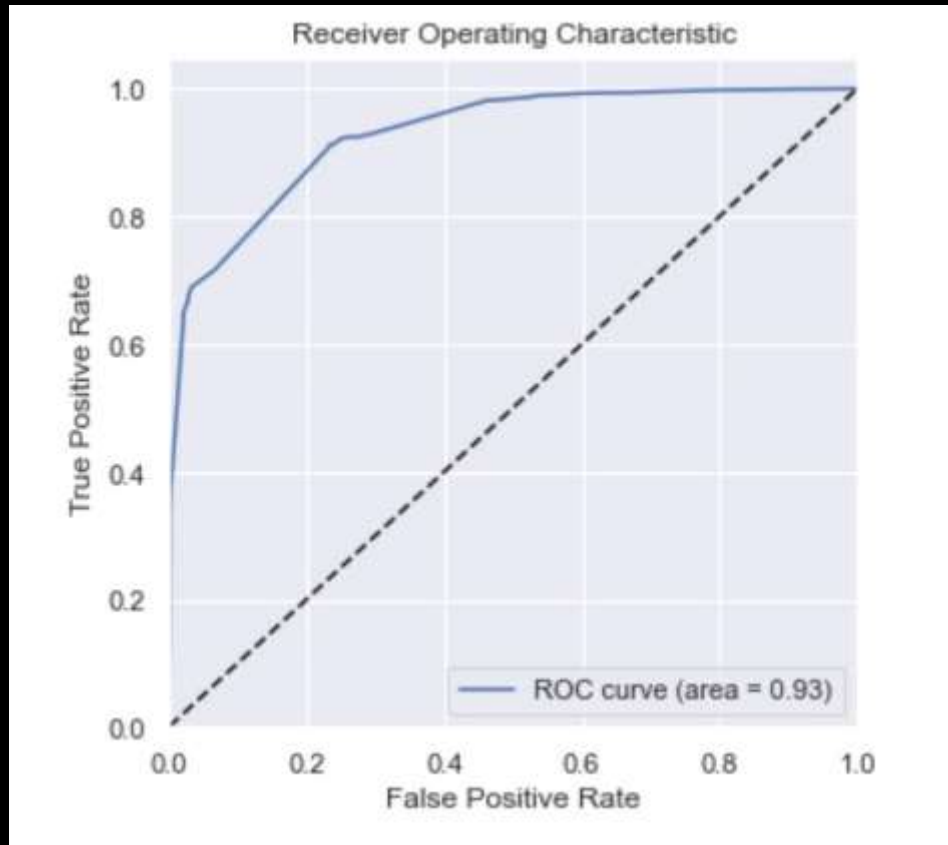
	Features	VIF				
9	Tags_Will revert after reading the email	1.79				
2	Last Activity_SMS Sent	1.58				
10	Last Notable Activity_Modified	1.38				
1	Lead Origin_Lead Add Form	1.24				
0	Do Not Email	1.17				
4	What is your current occupation_Working Profes...	1.17				
6	Tags_Closed by Horizon	1.17				
8	Tags_Ringing	1.12				
3	Last Activity_Unsubscribed	1.08				
5	Tags_Busy	1.04				
7	Tags_Lost to EINS	1.04				
11	Last Notable Activity_Olark Chat Conversation	1.03				

All p-values are zero and VIF is less than 5

HEATMAP

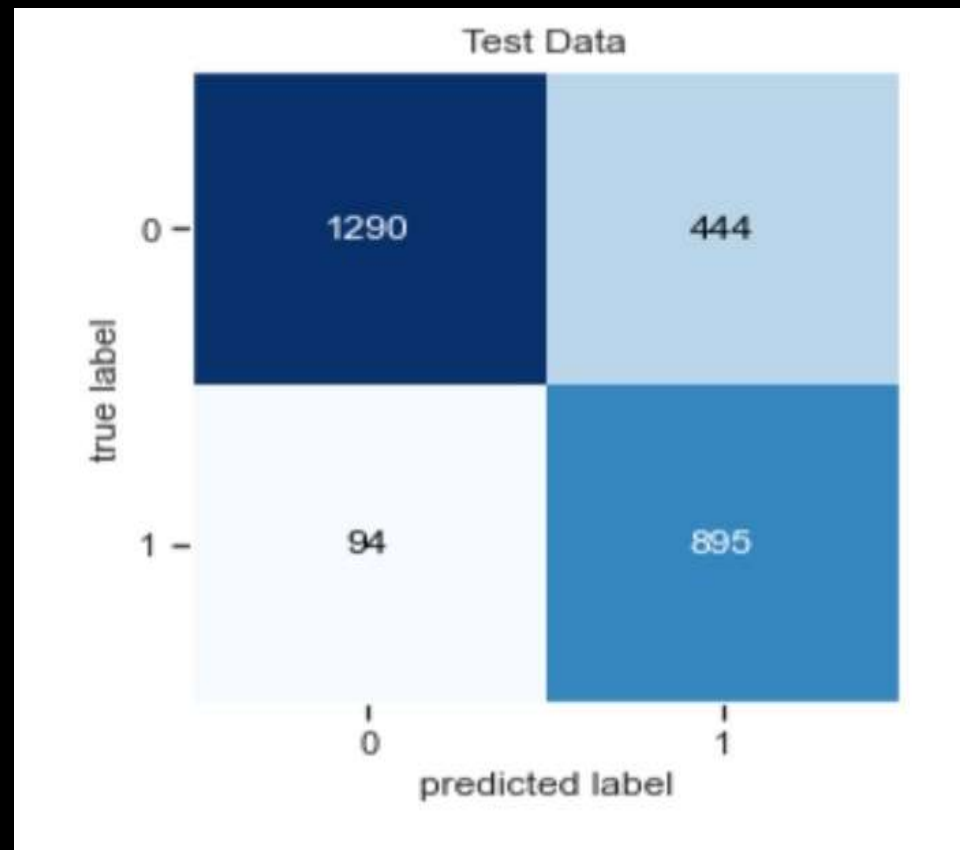
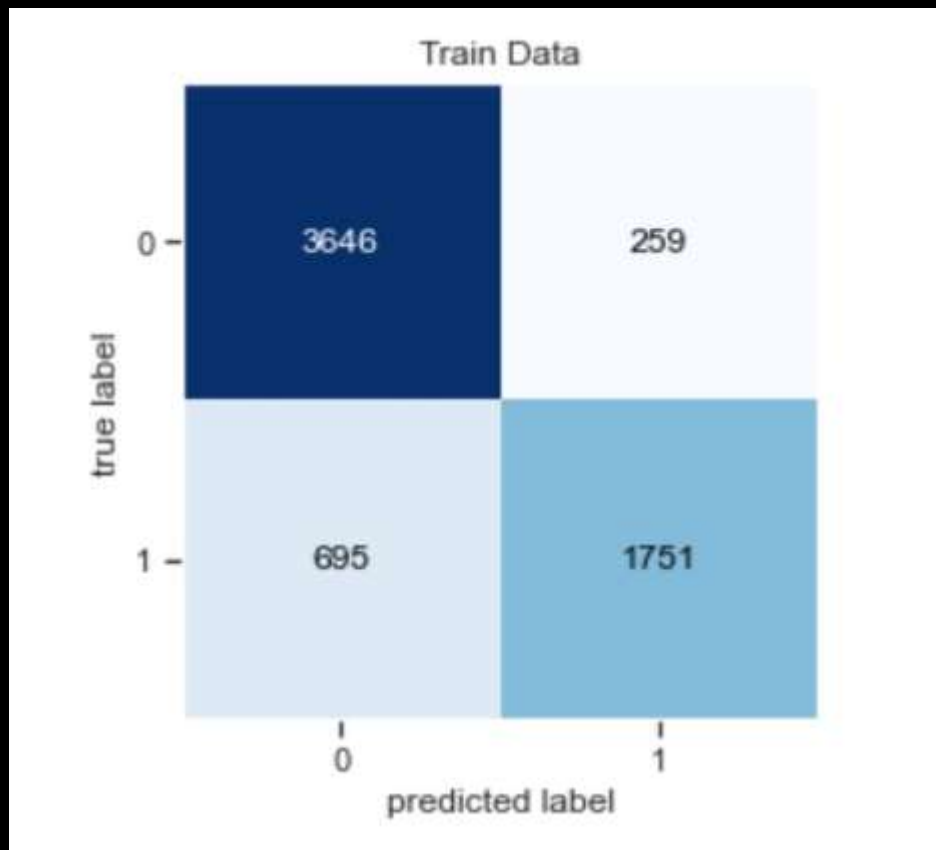


ROC CURVE



Area under curve = 0.93

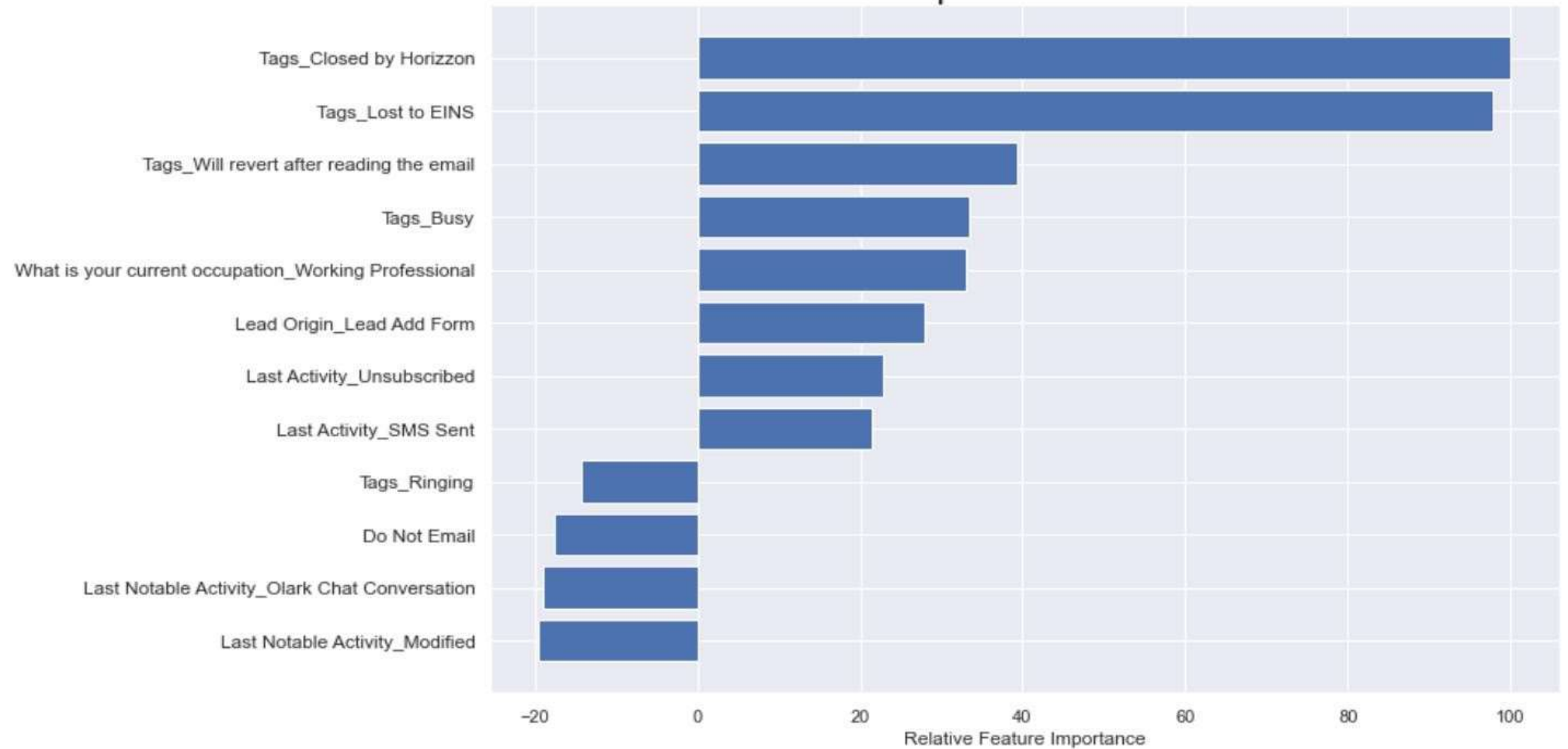
CONFUSION MATRIX



FINAL RESULT

Data	Train set	Test set
Accuracy	0.849	0.802
sensitivity	0.715	0.904
Specificity	0.933	0.743
Precision	0.871	0.668

Relative Importance Of Features



The background features a large, solid black area. At the top, there is a vibrant, wavy, liquid-like shape that transitions through a spectrum of colors: red, orange, yellow, green, and blue, creating a dynamic, abstract header.

INFERENCES

FEATURES IMPORTANCE

- ❖ Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:
 - Tags_Closed by Horizzon
 - Tags_Lost to EINS
 - Tags_Will revert after reading the email
- ❖ These are dummy features created from the categorical variable Tags.
- ❖ All three contribute positively towards the probability of a lead conversion.
- ❖ These results indicate that the company should focus more on the leads with these three tags.

RECOMMENDATIONS

- ❖ By referring to the data visualizations, focus on
 - Increasing the conversion rates for the categories generating more leads and
 - Generating more leads for categories having high conversion rates.
- ❖ Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.
- ❖ Based on varying business needs, modify the probability threshold value for identifying potential leads.



THANK YOU