LEAD SCORING CASE STUDY

SUBMITTED BY:

GURPREET KAUR : DSC43/EPGDS/IIITB

SHIVAM SHARMA : DSC43/EPGDS/IIITB

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.

- SELECTION OF HOT LEADS
- FOCUSED MARKETING
- HIGHER LEAD CONVERSION RATE

METHODOLOGY

To build a Logistic Regression model that assigns lead scores value between 0 and 100 to each of the leads which can be used by the company to target potential leads such that the customers with higher lead score usually have a higher conversion chance and vice versa.

Target Lead Conversion Rate ≈ 80%

SUB-GOALS

Create a **Logistic Regression model** to predict the Lead Conversion probabilities for each lead.

Decide on a **Probability Threshold value** above which a lead will be predicted as converted, whereas not converted if it is below it.

Multiply the **Lead Conversion probability** to arrive at the **Lead Score** value for each lead.

APPROACH

Importing and Observing the past data provided by the Company

- Univariate and Bivariate Analysis
- Analysis Using Correlation Matrix & Heat-Map

Reading and Understanding the Data → **Data Cleaning** → **Exploratory Data Analysis** → **Data Preparation for Modelling**

- Missing value Imputation (Categorical & Numerical Attributes)
- Removing duplicate data and other redundancies
- Handling the 'Select' Level
- Outlier Analysis & Treatment
- Checking Skewness
- Dropping Columns
- Data Imbalance Analysis
- Percentage of Rows Retained

- Dummy Variable Creation
- Splitting the Data into Train and Test Sets
- Feature Scaling

CONTD.

APPROACH

- Feature selection using RFE
- Assessing the model with StatsModels (P-Value, VIF)
- Creating Prediction & Statistical Analysis on the Train dataset

- Finalizing the first model
- Using predicted probabilities to
- Calculate Lead Scores:
- **Lead Score = Probability * 100**

- Getting a Relative Coefficient Value
- Ranking features based on Importance
- Selecting Top 3 features

Model Building → **Model Evaluation** → **Assigning Lead Scores** → **Hot Leads Determination** → **Features Importance Determination**

- Making predictions on the Test Set
- Evaluating model based on various Evaluation Metrics
- Finding the Optimal Probability Threshold
- Plotting ROC curve – Check AUC
- Precision Recall Trade-Off
- Calculating Cross Validation Score

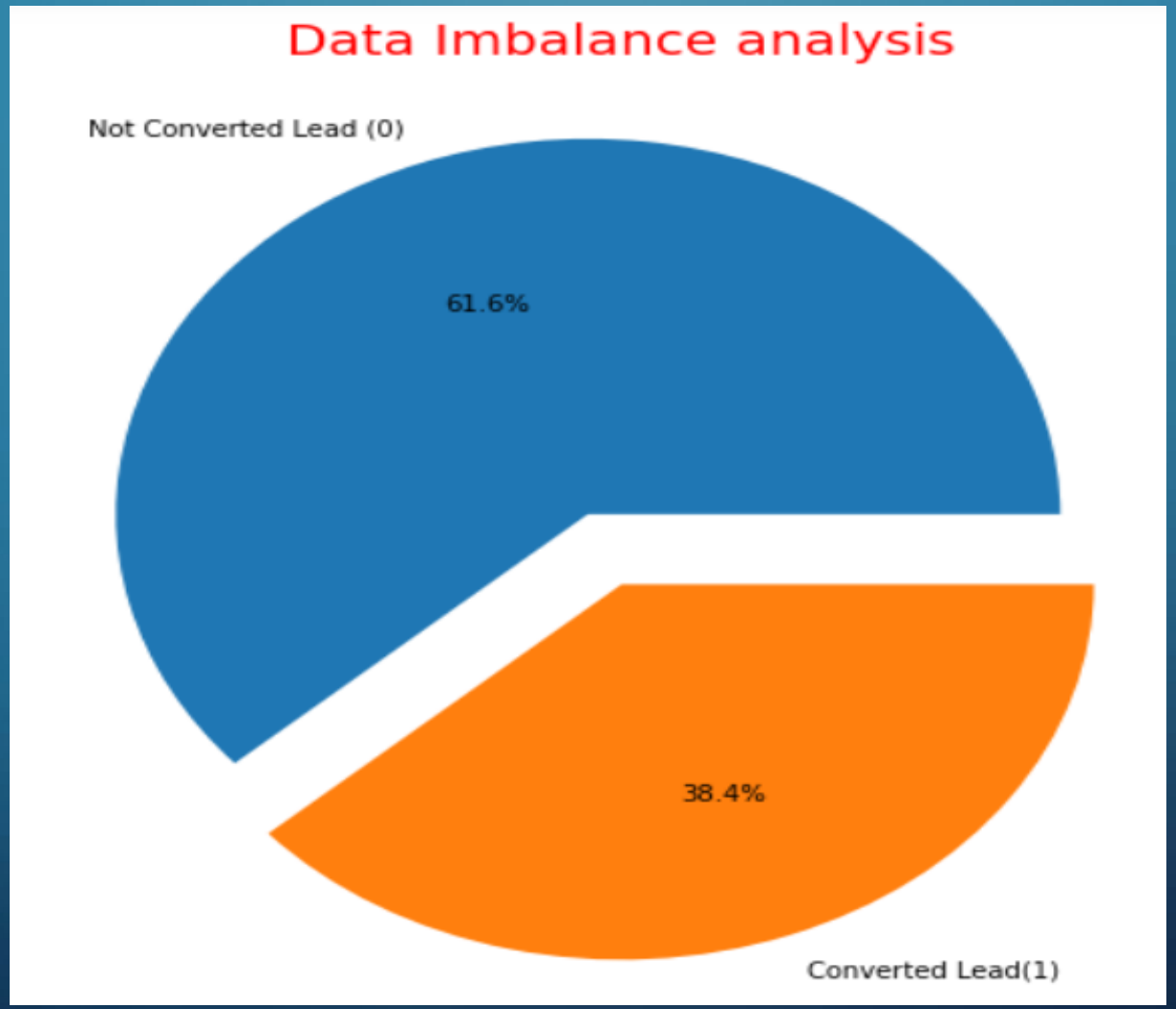
Determining Hot & Potential Leads with More than 80% Conversion Rate & Good Accuracy!

DATA VISUALIZATION

EXPLORATORY DATA ANALYSIS (EDA)

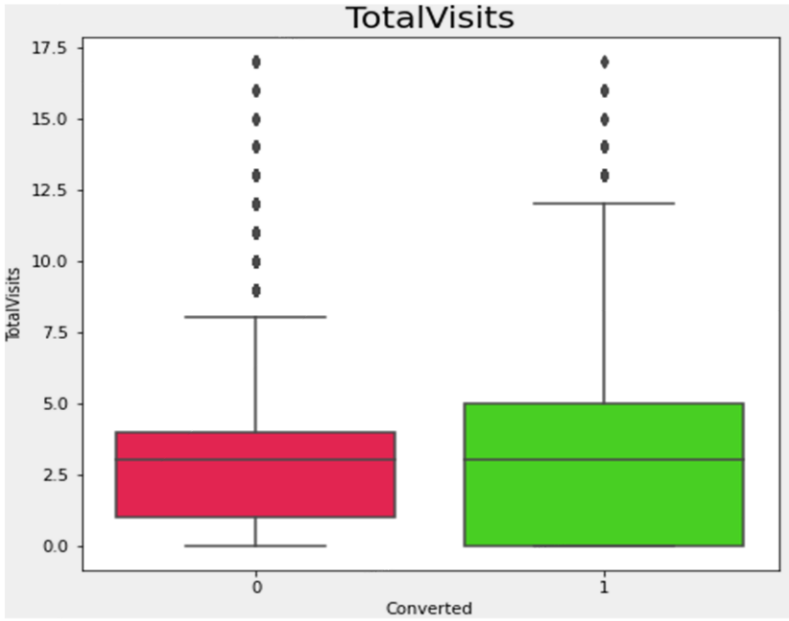
- Visualizing the Data on Categorical & Numerical
- Univariate & Bivariate Analysis
- Analysis Using Correlation Matrix & Heat-Map
- To Identify Important Features
- To Get Insights

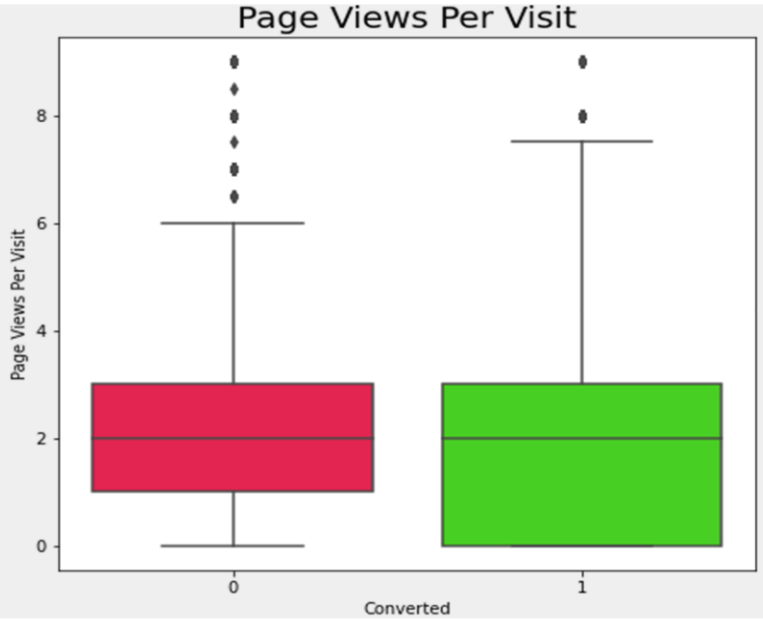
DATA IMBALANCE ANALYSIS

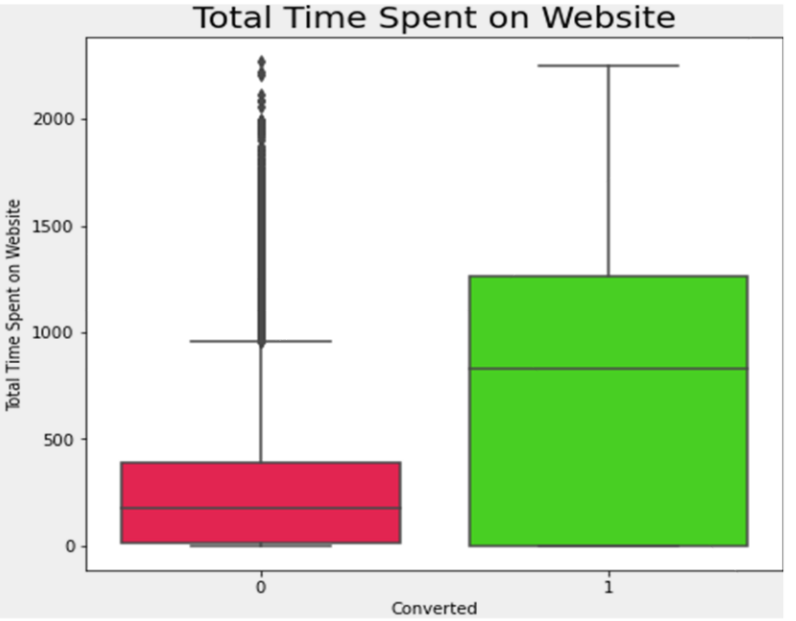


As per above analysis Data is not much Imbalanced 61.6% belong to Not Converted Leads and 38.4% belong to Converted leads in Data.

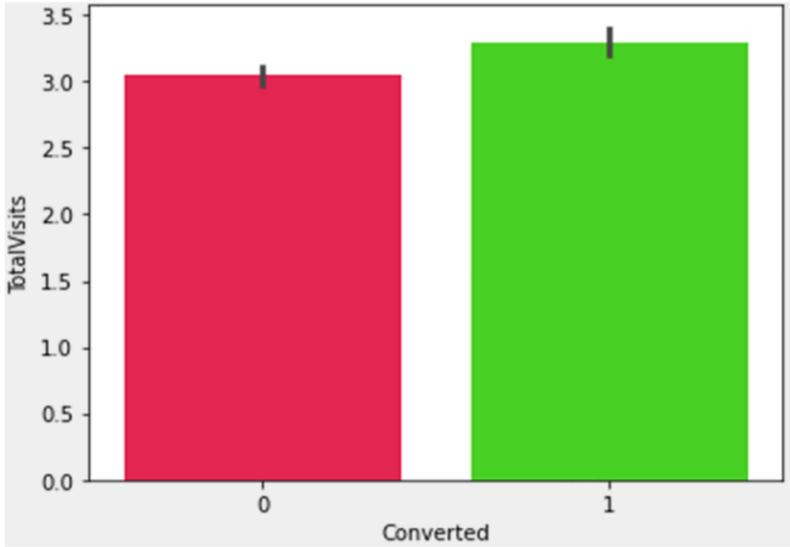
NUMERICAL VARIABLES ANALYSIS

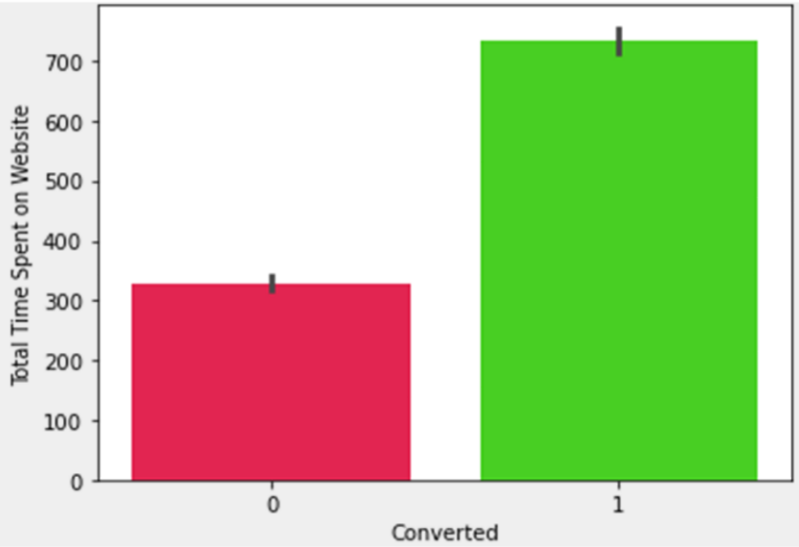


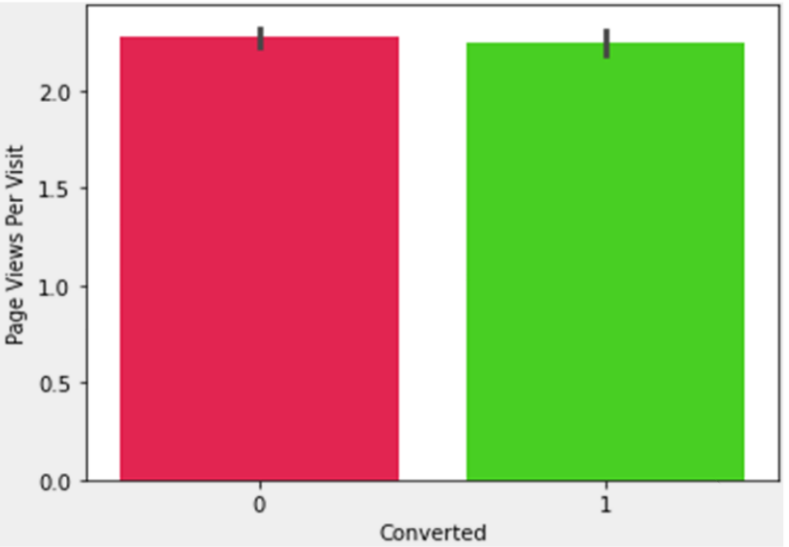




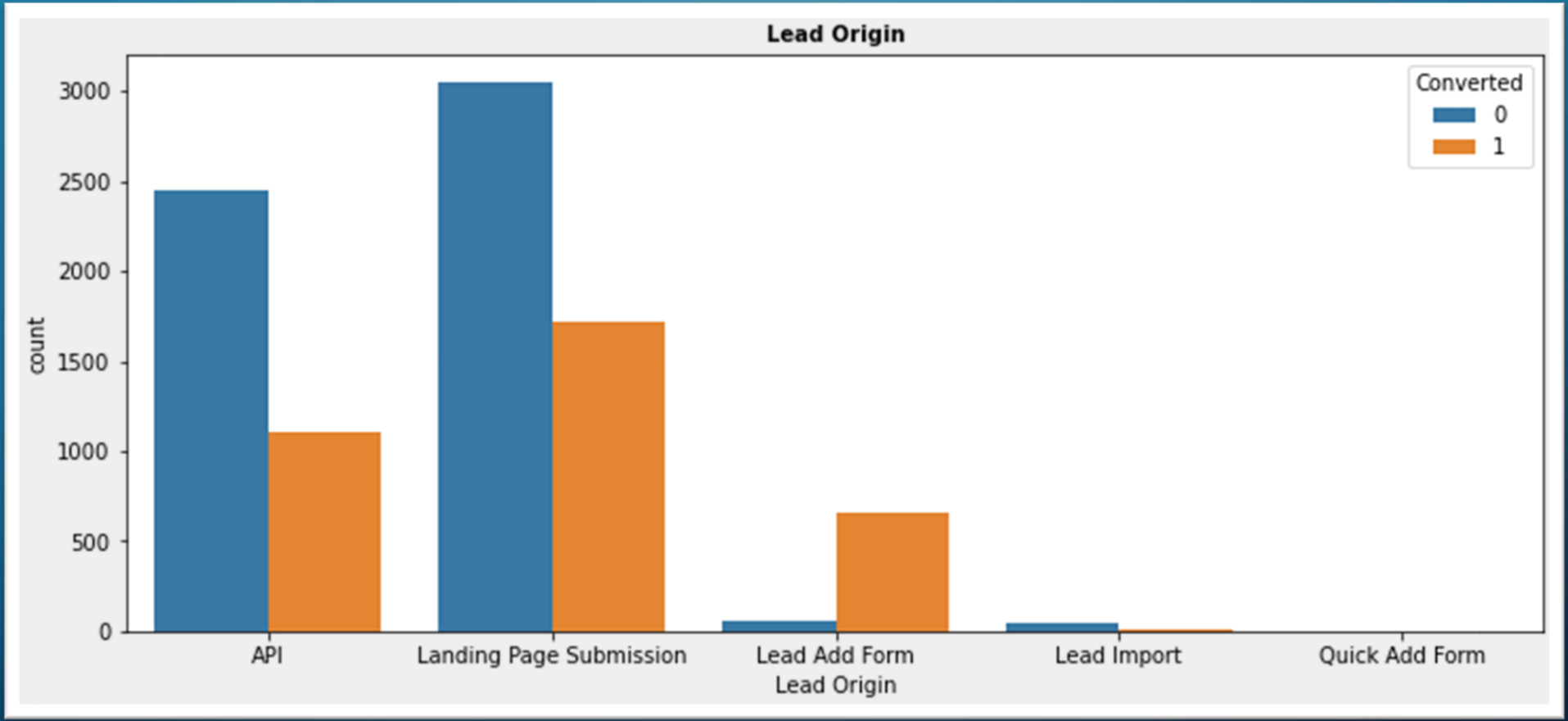
Inference: People spending More Time on Websites are more Likely to get Converted!!!





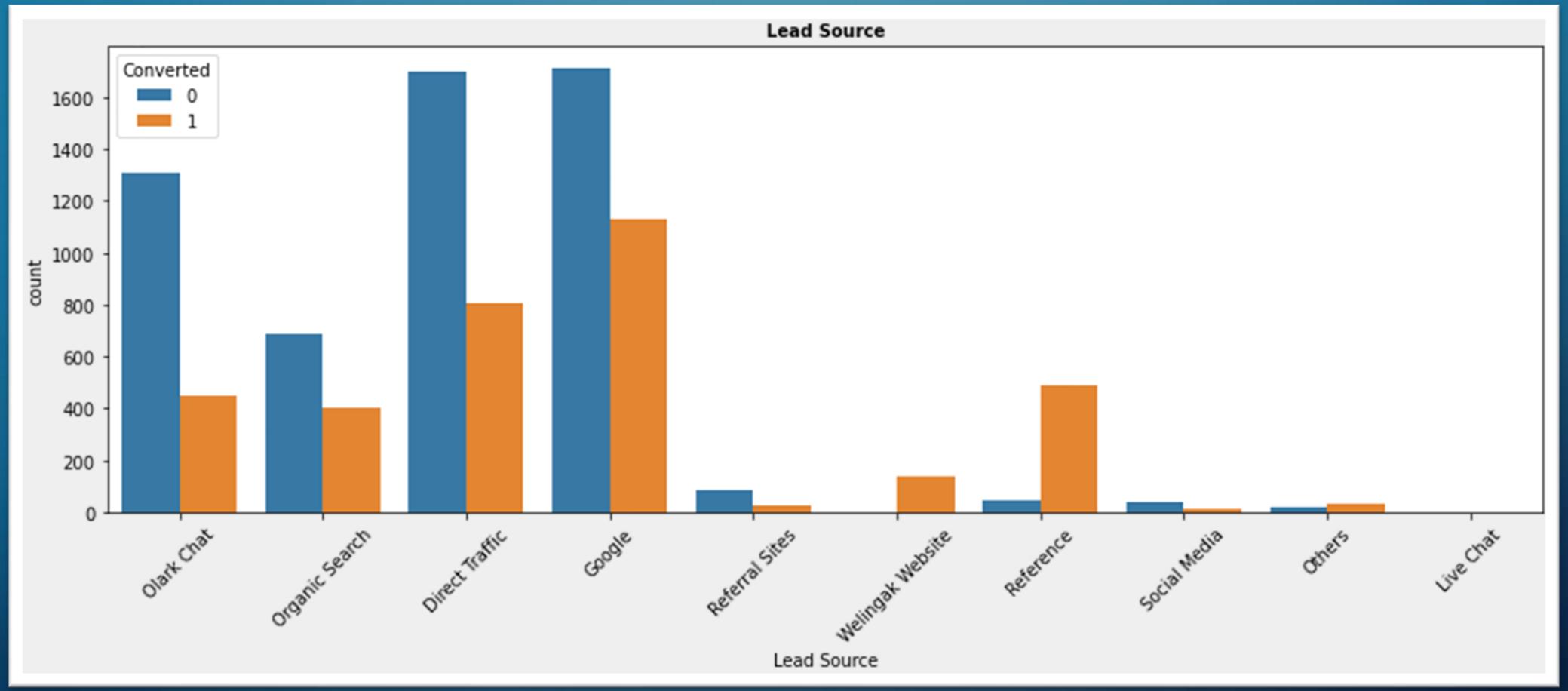


LEAD ORIGIN



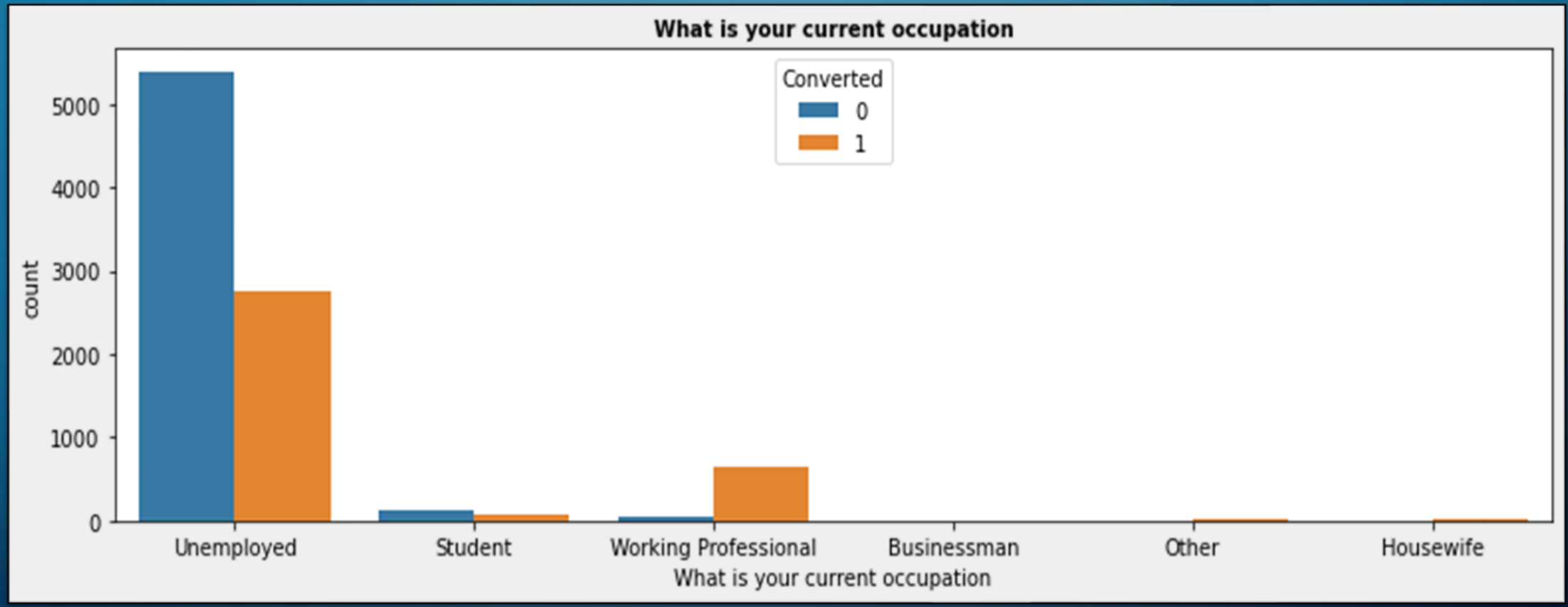
- 'API' and 'Landing Page Submission' generate the Most Leads but have less conversion rates.
  - **Focus on the Increasing Conversion Rate for 'API' and 'Landing Page Submission.**
- 'Lead Add Form' generates Fewer leads but the conversion rate is Great.
  - **Focus on Increasing leads generation using the 'Lead Add Form'.**

LEAD SOURCE



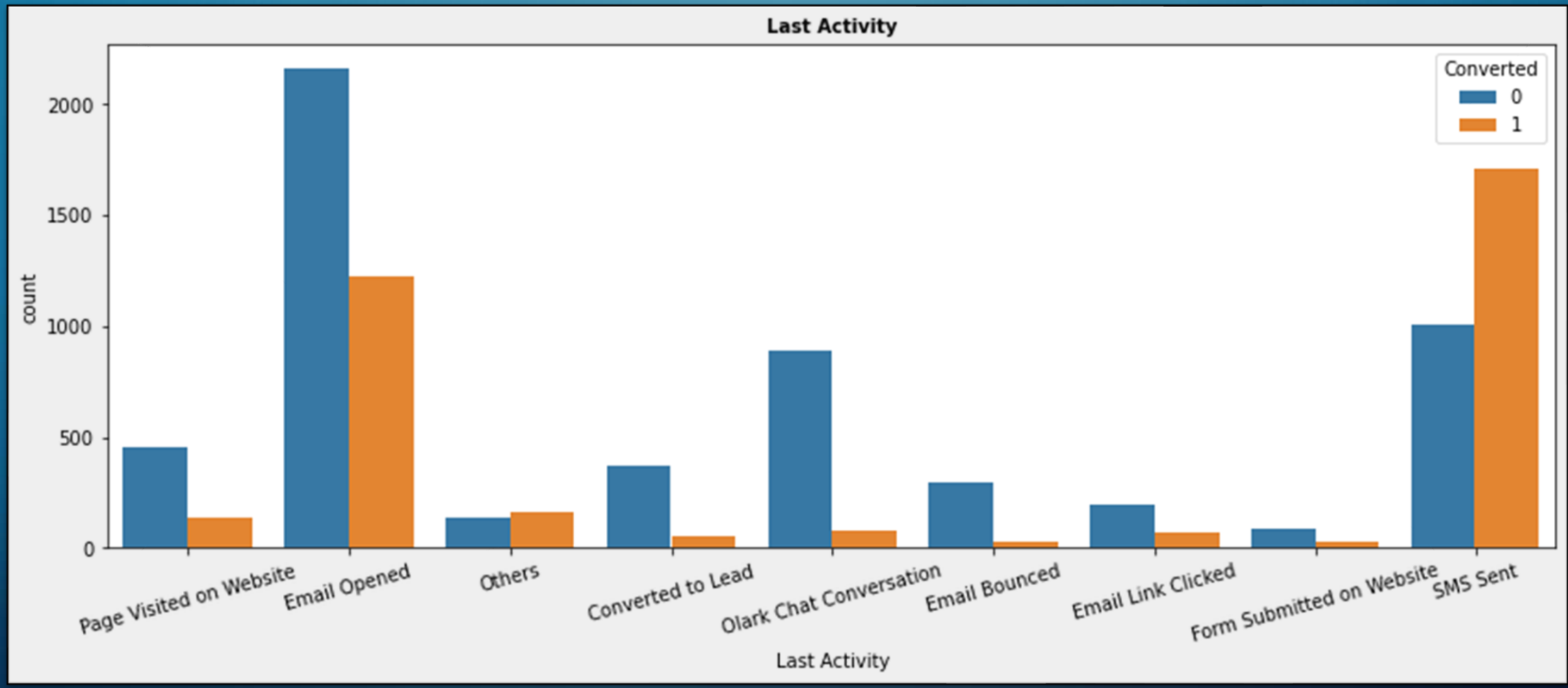
- Very High Conversion Rates For Lead Sources ‘Reference’ and ‘Welingak Website’.
- Most Leads are generated through **‘Direct Traffic’ and ‘Google’.**

CURRENT OCCUPATION



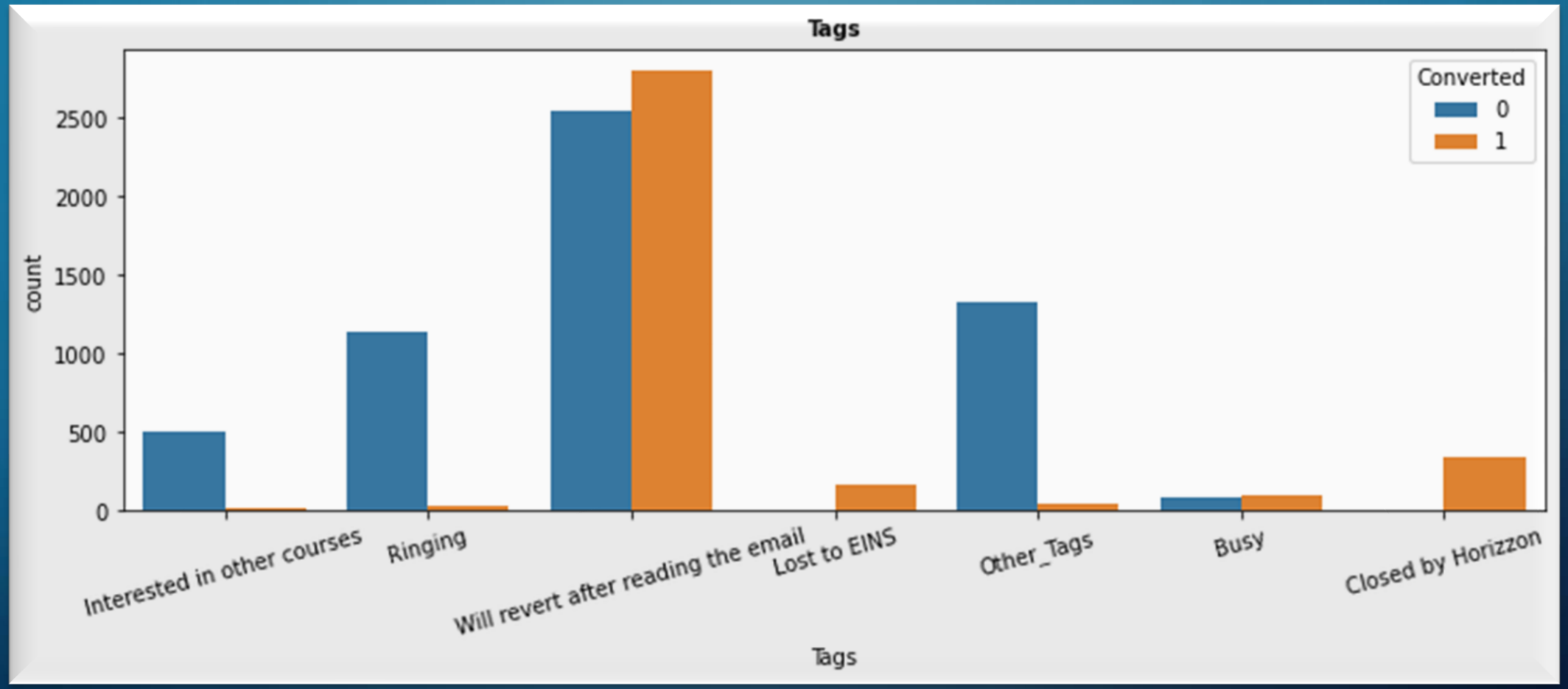
- Working Professionals are most likely to get Converted.
- Person who are Unemployed has also The Highest Conversion Rate due to Job Hunt & Career Aspect.

LAST ACTIVITY



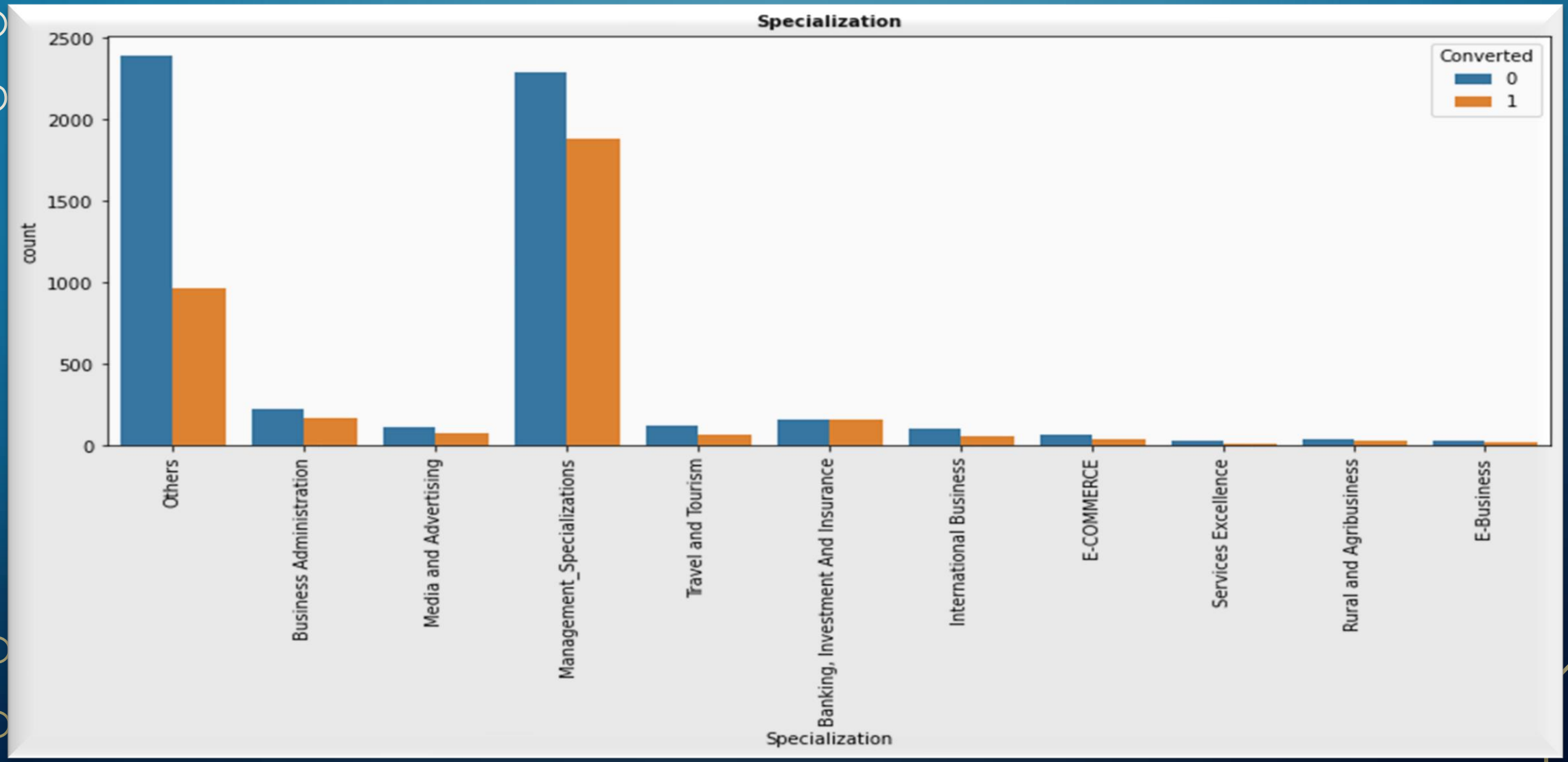
Inference- Leads whose Last Activity was SMS sent had the Best Conversion Rate.

TAGS



Most leads generated and the **Highest Conversion Rate** are both attributed to the tag **'Will revert after reading the email'** & 'Closed by Horizon'

SPECIALIZATION



Inference- Leads from Management Specialization has the Highest Rate Of Conversion.

MODEL EVALUATION

FINAL MODEL SUMMARY

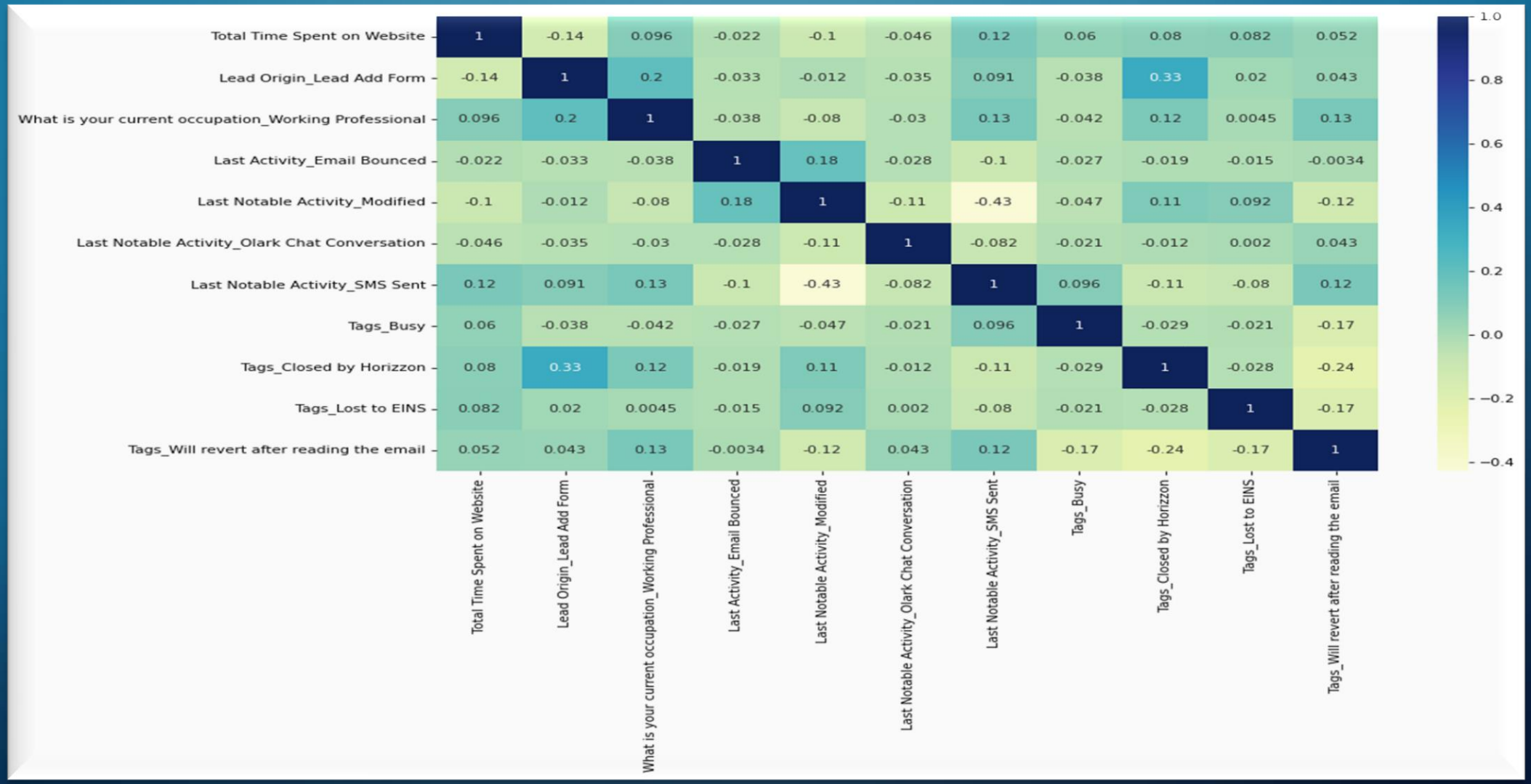
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6363
Model:	GLM	Df Residuals:	6351
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1743.4
Date:	Fri, 21 Oct 2022	Deviance:	3486.8
Time:	14:27:41	Pearson chi2:	1.09e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.7575	0.175	-27.178	0.000	-5.101	-4.414
Total Time Spent on Website	1.0489	0.046	22.990	0.000	0.959	1.138
Lead Origin_Lead Add Form	3.4631	0.231	15.010	0.000	3.011	3.915
What is your current occupation_Working Professional	2.8328	0.255	11.129	0.000	2.334	3.332
Last Activity_Email Bounced	-2.0312	0.401	-5.060	0.000	-2.818	-1.244
Last Notable Activity_Modified	-1.0021	0.103	-9.767	0.000	-1.203	-0.801
Last Notable Activity_Olark Chat Conversation	-1.7204	0.361	-4.771	0.000	-2.427	-1.014
Last Notable Activity_SMS Sent	2.1266	0.118	18.021	0.000	1.895	2.358
Tags_Busy	3.7056	0.269	13.778	0.000	3.178	4.233
Tags_Closed by Horizzon	9.1079	0.747	12.191	0.000	7.644	10.572
Tags_Lost to EINS	9.5713	0.762	12.567	0.000	8.079	11.064
Tags_Will revert after reading the email	4.3560	0.169	25.828	0.000	4.025	4.687

All P- Values Are Zero, showing Significant Features Contributing towards Lead Conversion.

HEATMAP



- Correlations Between Features In The Final Model Are Negligible
- All The Features Have Very Low VIF Values, Meaning, There Is Hardly Any Muliticollinearity Among The Features. This Is Also Evident From The **HEAT MAP.**

LEAD SCORE CALCULATION

Determining HOT LEADS with 89% Accuracy & more than 80% Conversion Rate!!

```
# Determining hot leads with more than 80% Conversion Rate
hot_leads = lead_full_pred[lead_full_pred["Lead_Score"]>80]
hot_leads.head()
```

Prospect ID	Converted	Converted_prob	final_Predicted	Lead_Score
2	1	0.834168	1	83
11	1	0.998449	1	100
18	1	0.871649	1	87
37	1	0.929814	1	93
64	1	0.986097	1	99

```
# Hot Leads Shape
hot_leads.shape
```

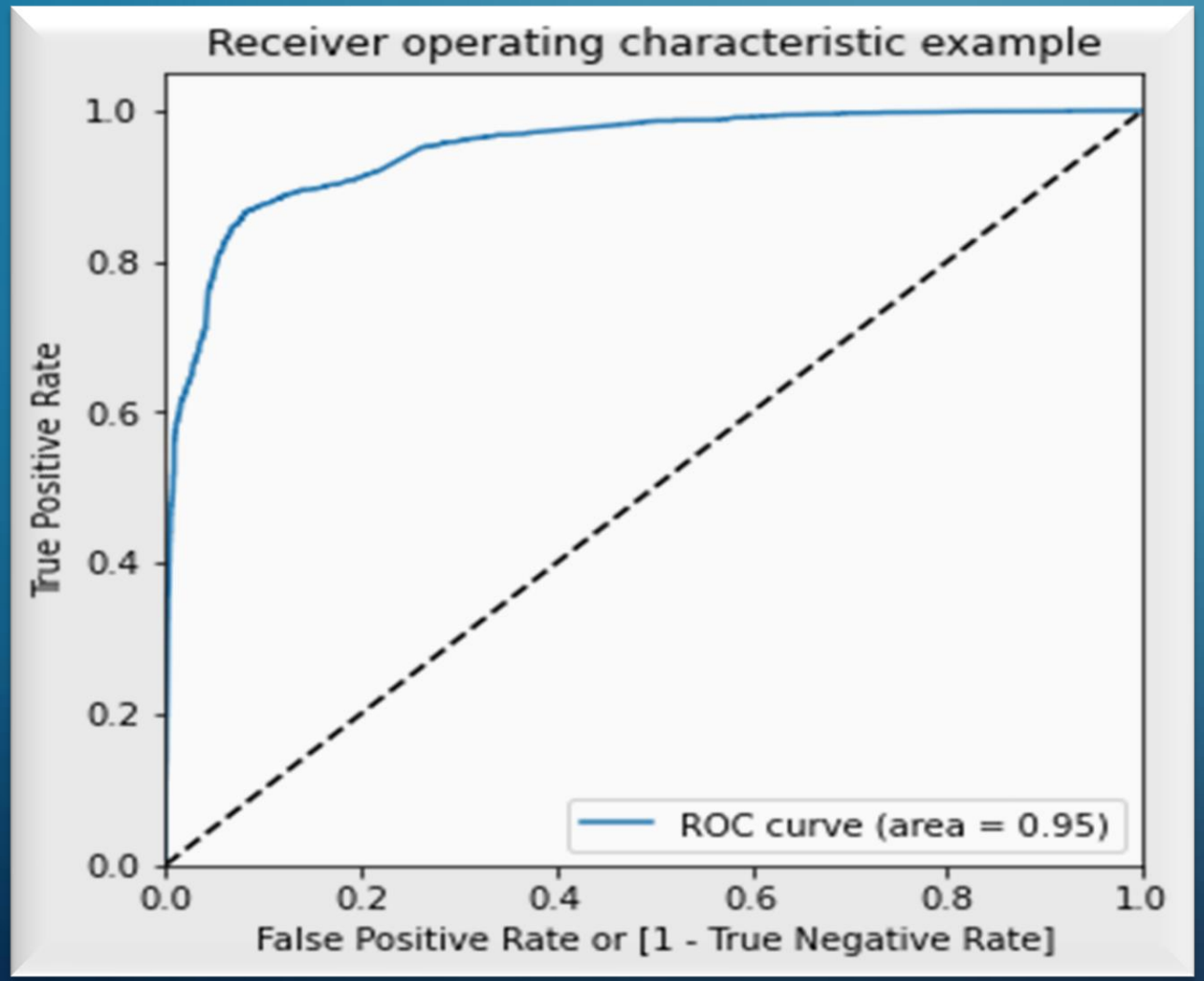
```
(2249, 4)
```

Formula for Lead Score calculation is:

Lead Score = 100 * Conversion Probability

- Lead Score is calculated for all the leads in the Original Dataframe.
- The Train And Test Dataset is concatenated to get the entire list of leads available.
- The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.
- **Higher The Lead Score, Higher Is The Probability Of A Lead Getting Converted** and vice versa, Since, we had used 0.34 as our final Probability threshold for deciding if a lead will convert or not.

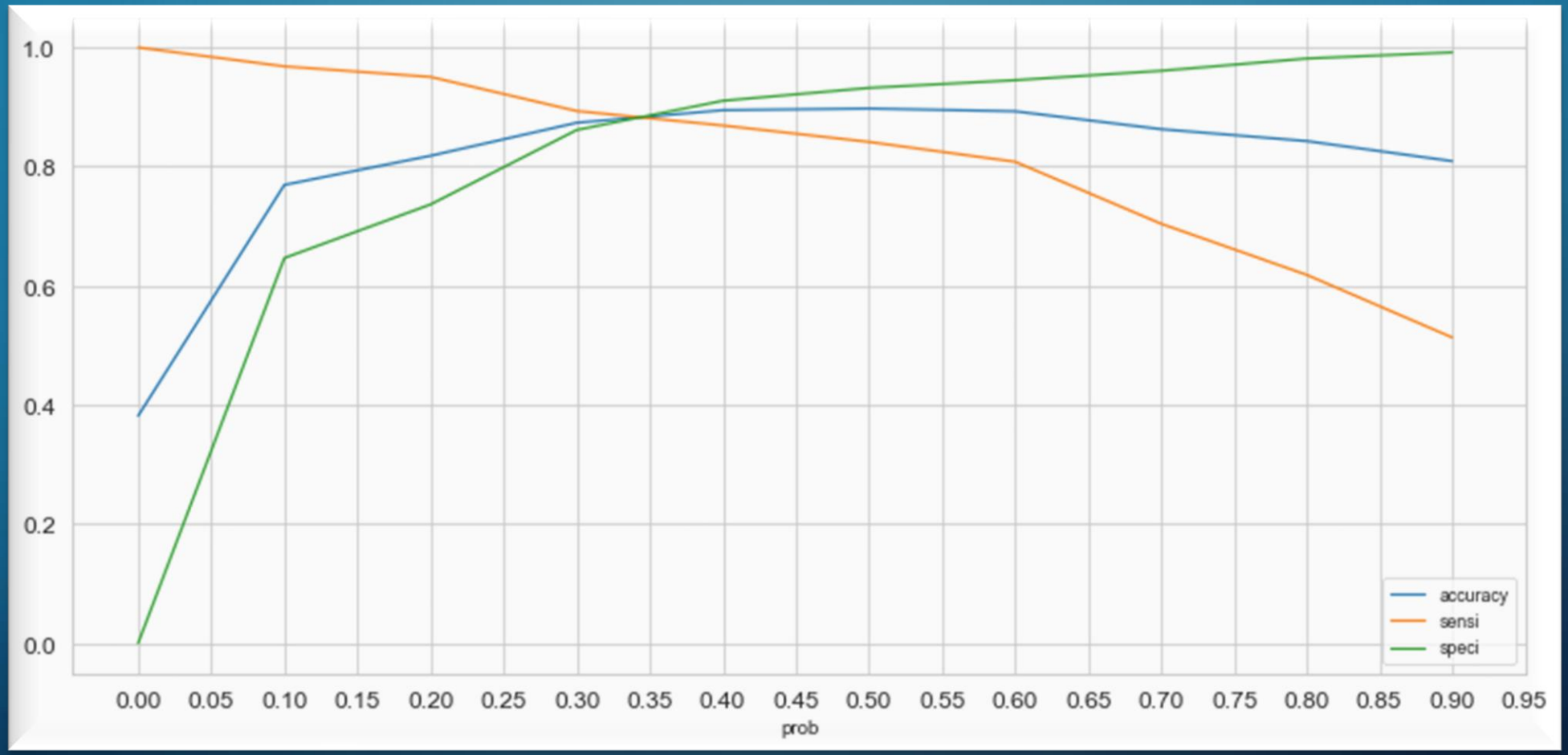
ROC CURVE



The ROC Curve should be a value close to 1. We are getting a good value of "0.95" indicating a "GOOD Predictive Model."

Area under curve = 0.96

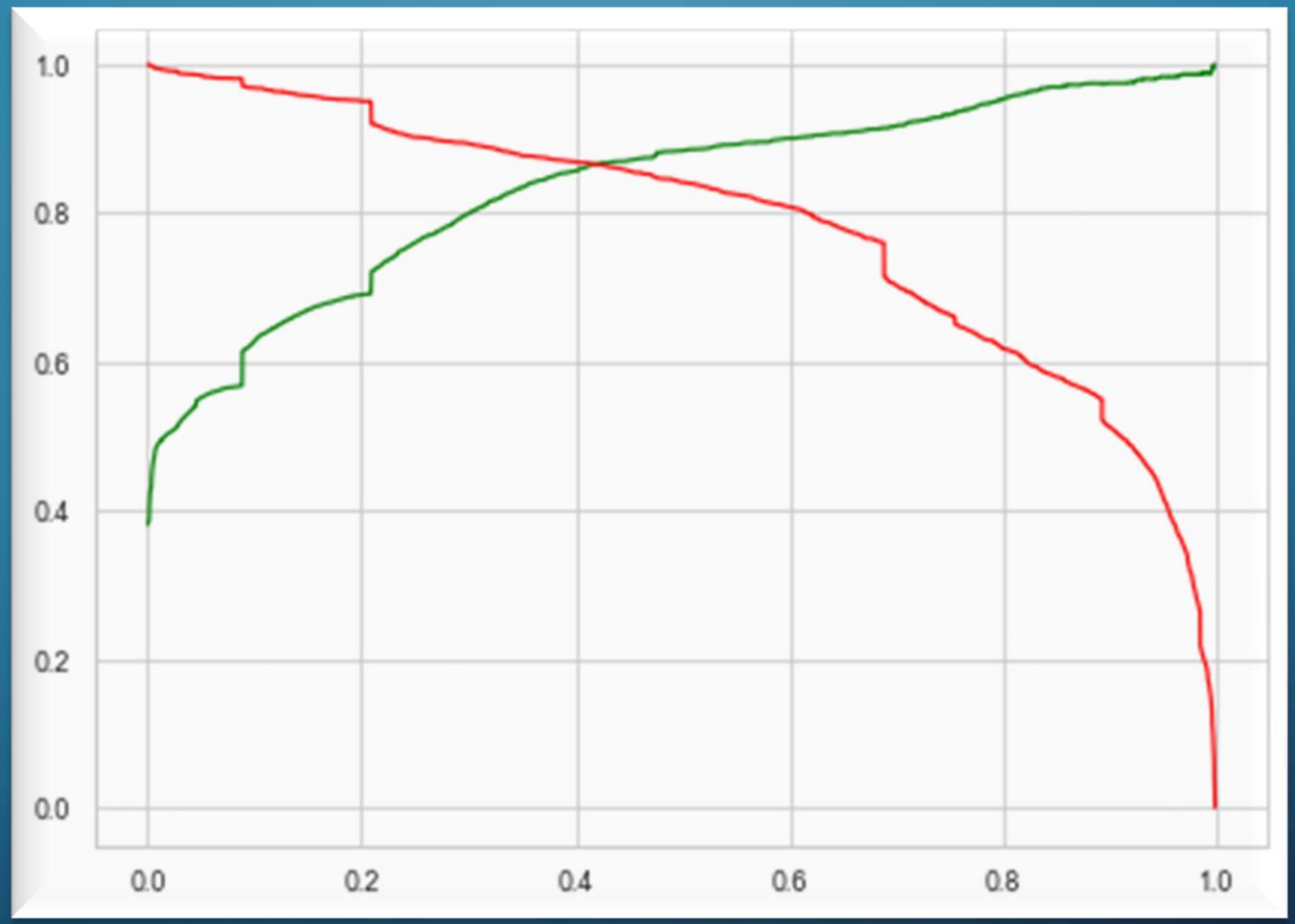
FINDING OPTIMAL THRESHOLD



Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values.

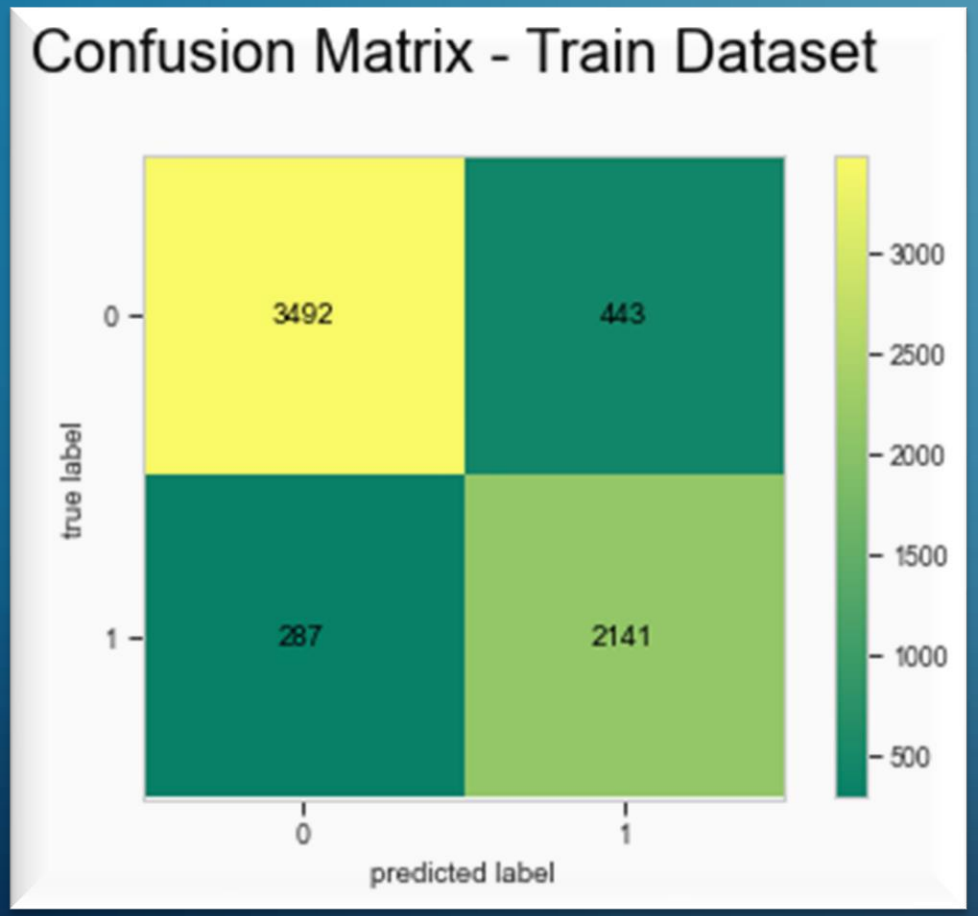
- **OPTIMAL PROBABILITY CUTOFF** = 0.34

PRECISION AND RECALL TRADE-OFF

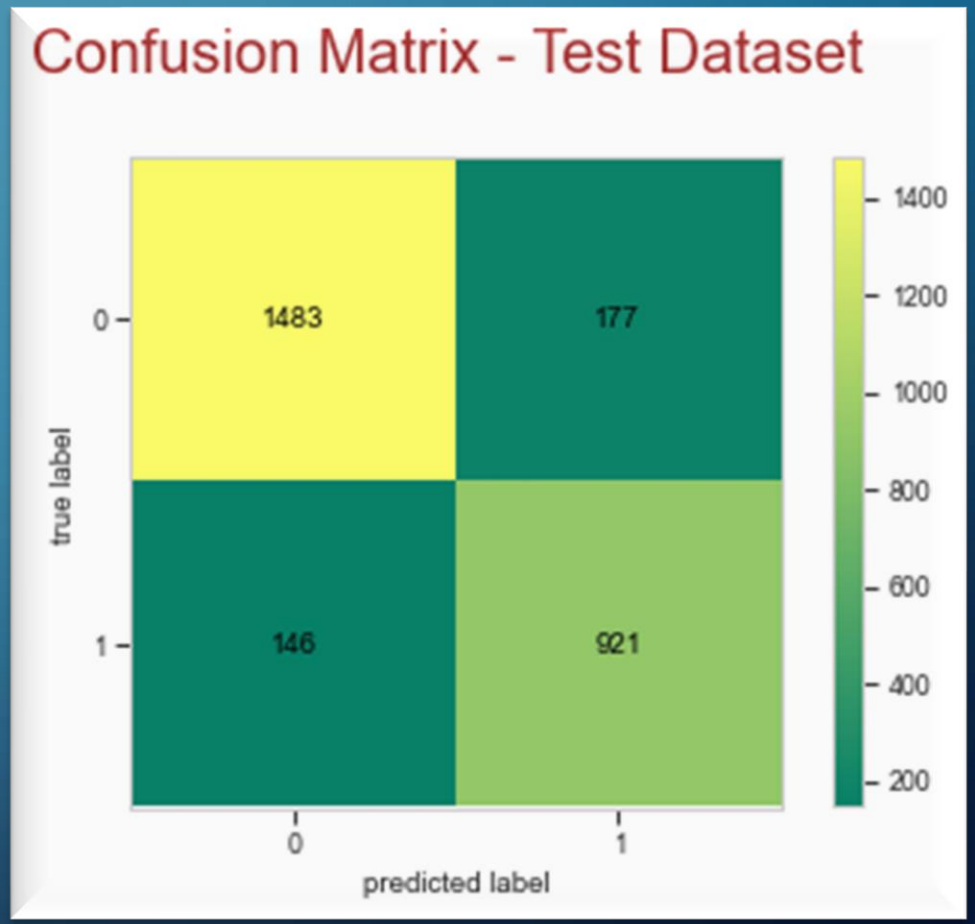


From The **Precision-Recall Graph** Above, We Get The Optical Threshold Value As Close To **0.41**

CONFUSION MATRIX



FOR TRAIN SET



FOR TEST SET

FINAL RESULTS

METRICS	TRAIN SET	TEST SET
ACCURACY	88.52%	88.15%
SENSITIVITY/RECALL	88.17%	86.31%
SPECIFICITY	88.74%	89.33%
FALSE POSTIVE RATE (FPR)	11.25%	10.66%
PRECISION/POSITIVE PREDICTIVE VALUE	82.85%	83.87%
NEGATIVE PREDICTIVE VALUE	92.40%	91.03%
ROC	0.95	0.95
F1 SCORE	0.85	0.85

Train Set Classification Report

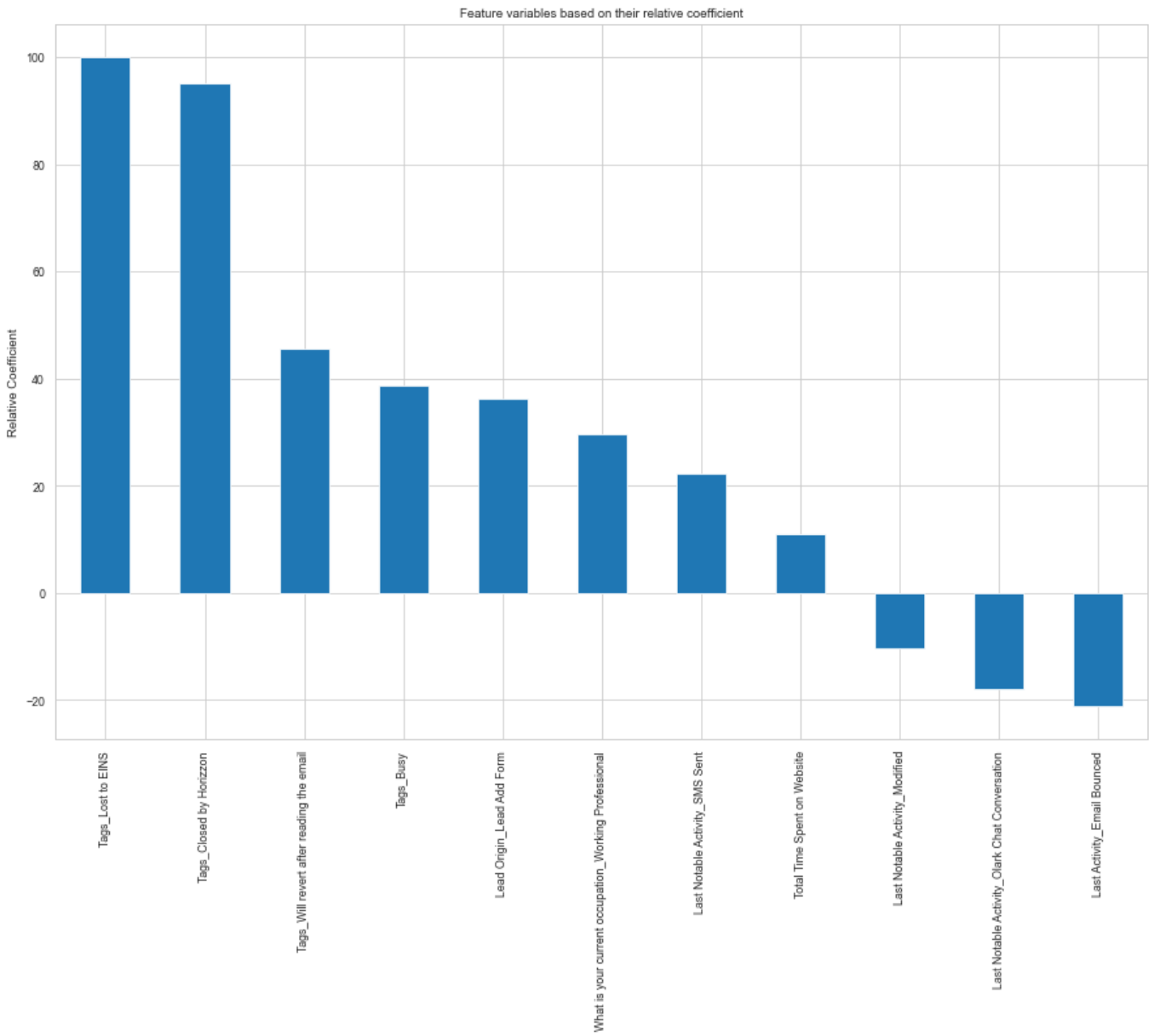
	precision	recall	f1-score	support
0	0.92	0.89	0.91	3935
1	0.83	0.88	0.85	2428
accuracy			0.89	6363
macro avg	0.88	0.88	0.88	6363
weighted avg	0.89	0.89	0.89	6363

Test Set Classification Report

	precision	recall	f1-score	support
0	0.91	0.89	0.90	1660
1	0.84	0.86	0.85	1067
accuracy			0.88	2727
macro avg	0.87	0.88	0.88	2727
weighted avg	0.88	0.88	0.88	2727

INFERENCES

RELATIVE IMPORTANCE OF FEATURES



Tags_Lost to EINS	100.00
Tags_Closed by Horizzon	95.16
Tags_Will revert after reading the email	45.51
Tags_Busy	38.72
Lead Origin_Lead Add Form	36.18
What is your current occupation_Working Professional	29.60
Last Notable Activity_SMS Sent	22.22
Total Time Spent on Website	10.96
Last Notable Activity_Modified	-10.47
Last Notable Activity_Olark Chat Conversation	-17.97
Last Activity_Email Bounced	-21.22

dtype: float64

- The **Relative Importance** of each feature is determined on a **scale of 100** with the feature with highest importance having a score of 100!
- The **Coefficient (Beta) Values** for each of these features from the model parameters are used to determine the order of importance of these features.
- Features with **High Positive Beta Values** are the ones that contribute most towards the **Probability Of A Lead Getting Converted.**
- Similarly, features with **High Negative Beta Values** contribute the least.

FEATURE IMPORTANCE

- **Top Three** variables which contribute most towards the probability of Lead Conversion in decreasing order of impact are:
    - ✓ **Tags_Lost to EINS**
    - ✓ **Tags_Closed by Horizon**
    - ✓ **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 Lead Conversion**.
- These results indicate that the company should focus more on the leads with these three tags.
- Other Features with Positive Coefficient Values **Impacting Potential Hot Leads** are:
    - ✓ **Lead Origin_Lead Add Form**
    - ✓ **What is your current occupation_Working Professional**
    - ✓ **Last Notable Activity_SMS Sent**
    - ✓ **Total Time Spent on Website**

Case 1: The Company Has Interns For 2 Months. They Wish To Make The Lead Conversion More Aggressive. They Want Almost All Of The Potential Leads To Be Converted And Hence, Want To Make Phone Calls To As Many Of Such People As Possible.

Solution:

- $Sensitivity = True\ Positives / (True\ Positives + False\ Negatives)$

- Sensitivity can be defined as the number of actual conversions predicted correctly out of the total number of actual conversions. As we saw earlier, sensitivity decreases as the threshold increases.

- High Sensitivity implies that our model will correctly predict almost all leads who are likely to convert. At the same time, it may overestimate and misclassify some of the non-conversions as conversions.

- As the company has extra manpower for two months and wants to make the lead conversion more aggressive, it is a good strategy to go for high sensitivity. To achieve high sensitivity, we need to choose a **LOW THRESHOLD VALUE**.

Case 2: At Times, The Company Reaches Its Target For A Quarter Before The Deadline. It Wants The Sales Team To Focus On Some New Work. So During This Time, The Company's Aim Is To Not Make Phone Calls Unless It's Extremely Necessary.

Solution:

- $Specificity = True\ Negatives / (True\ Negatives + False\ Positives)$
- Specificity can be defined as the number of actual non-conversions predicted correctly out of a total number of actual non-conversions. It increases as the threshold increases.
- High Specificity implies that our model will correctly predict almost all leads who are not likely to convert. At the same time, it may misclassify some of the conversions as non-conversions.
- As the company has already reached its target for a quarter and doesn't want to make unnecessary phone calls, it is a good strategy to go for high specificity.
- It will ensure that the phone calls are only made to customers who have a very high probability of conversion. To achieve high specificity, we need to choose a **HIGH THRESHOLD VALUE**.

RECOMMENDATIONS

TOP 3 VARIABLES that Contributing Most towards the Probability of Lead conversion

1. Tags_Lost to EINS
2. Tags_Closed by Horizzon
3. Tag_Will revert after reading the email

TOP THREE CATEGORICAL/DUMMY VARIABLES that should be focused the most in order to increase the probability of Lead Conversion:

1. Tags – (Lost to EINS, Closed by Horizzon, Will revert after reading the email)
2. Lead Origin – (Lead Add Form)
3. What is your current occupation – (Working Professional)

WHOM TO CONSIDER HOT LEADS POTENTIAL PAYING CUSTOMERS??

- ✓ Company should make calls to the "Working Professionals" as they are more likely to get converted.
- ✓ Who visits websites repeatedly or Who spend much time on website and this can be done by making website easier and more informative.
- ✓ Their Last Activity is through SMS & Email Opened can be targeted.
- ✓ People having Tags "Will revert after reading emails" can be possible targeted leads
- ✓ Last Notable Activity_Had a Phone Conversation

THANK YOU

GURPREET KAUR : DSC43/EPGDS/IIITB
SHIVAM SHARMA : DSC43/EPGDS/IIITB