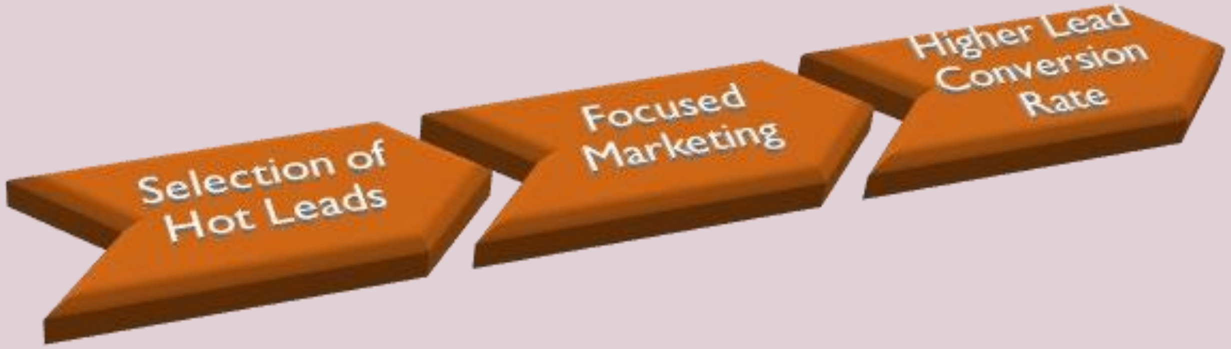
LEAD SCORING CASE STUDY

Focused Business Approach Using Logistic Regression Technique

Mohammad Asif

BUSINESS OBJECTIVE

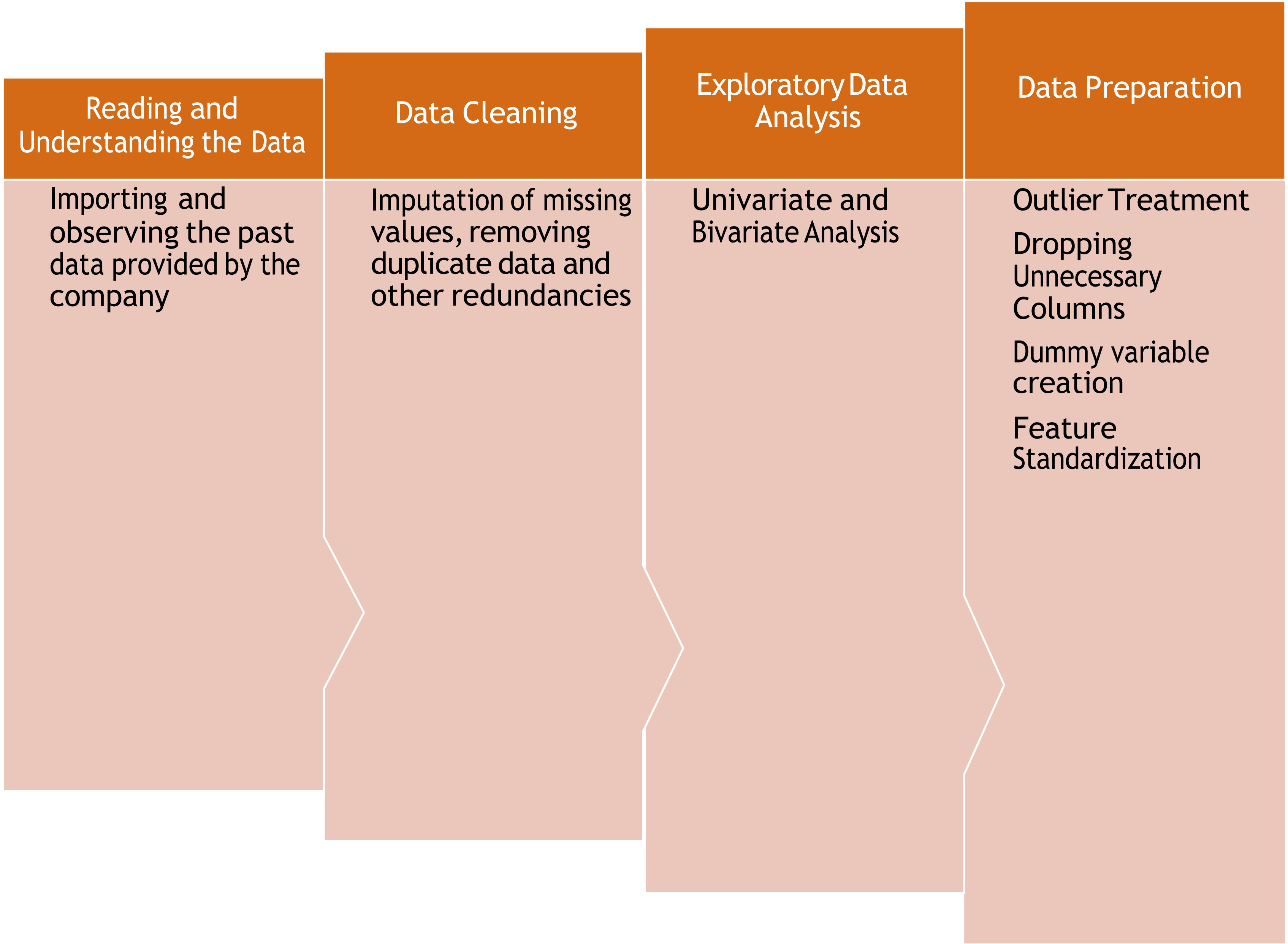
- To help X Education select most promising leads (Hot Leads), i.e., the leads that are most likely to convert into paying customers.



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 ≈ 80%



Reading and Understanding the Data

Importing and observing the past data provided by the company

Data Cleaning

Imputation of missing values, removing duplicate data and other redundancies

Exploratory Data Analysis

Univariate and Bivariate Analysis

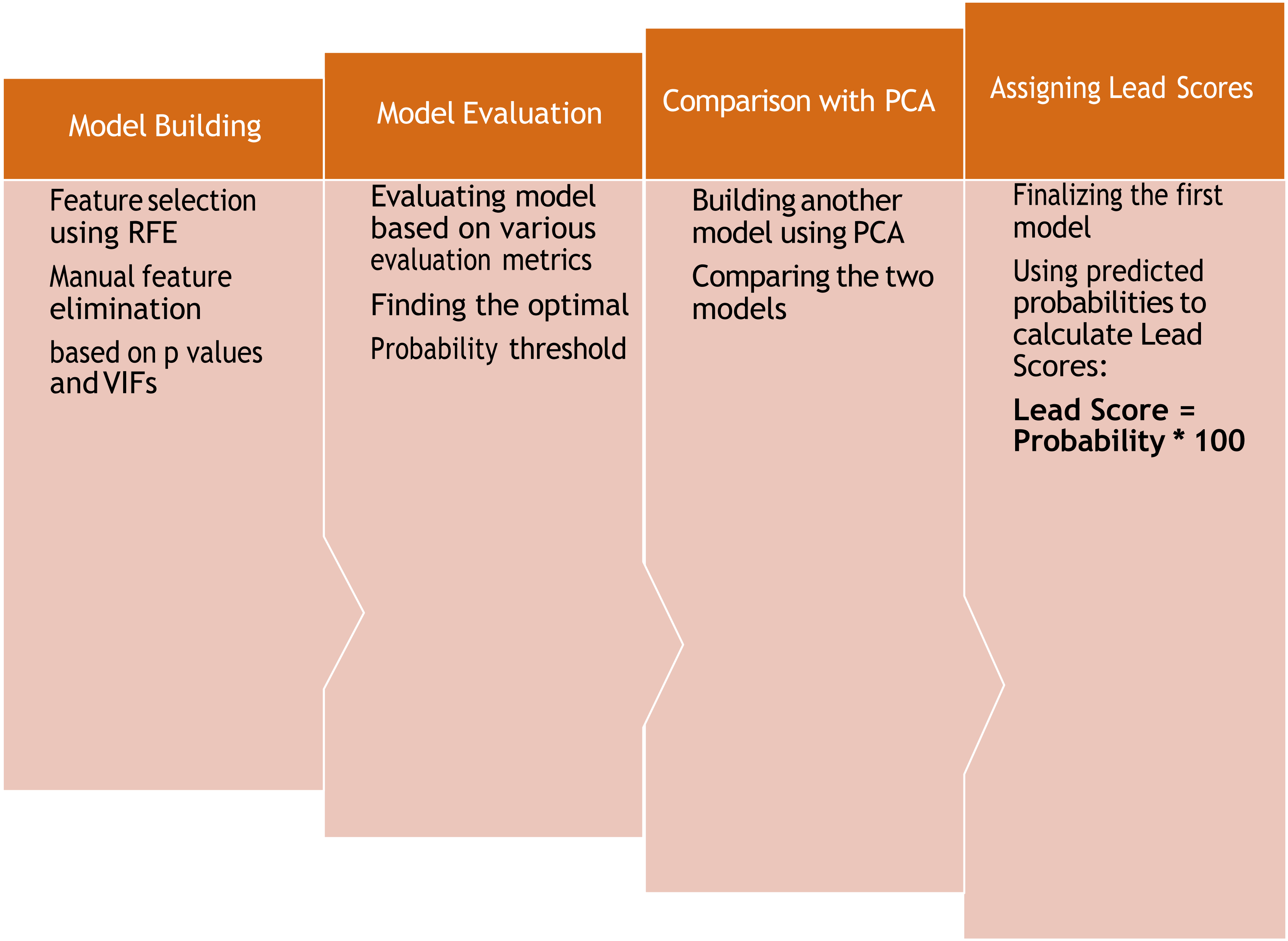
Data Preparation

Outlier Treatment

Dropping Unnecessary Columns

Dummy variable creation

Feature Standardization



Model Building

Feature selection using RFE

Manual feature elimination

based on p values and VIFs

Model Evaluation

Evaluating model based on various evaluation metrics

Finding the optimal

Probability threshold

Comparison with PCA

Building another model using PCA

Comparing the two models

Assigning Lead Scores

Finalizing the first model

Using predicted probabilities to calculate Lead Scores:

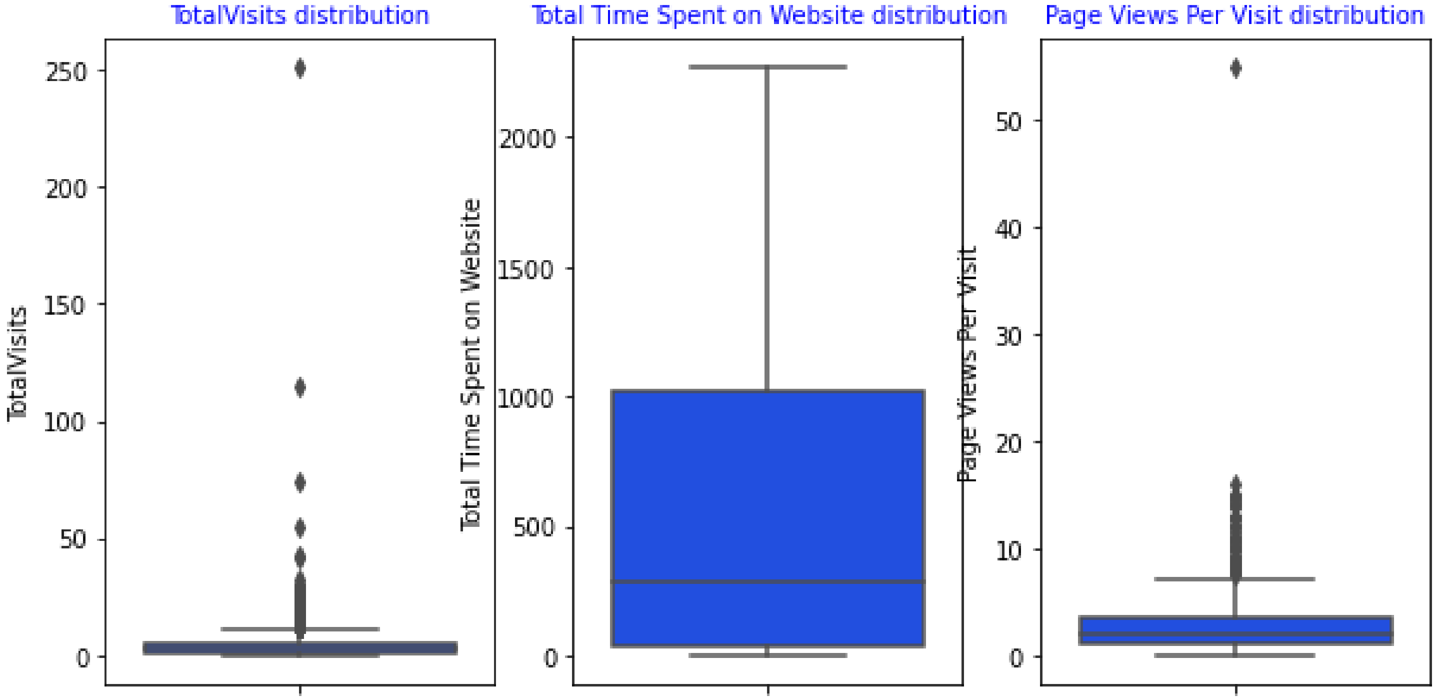
Lead Score = Probability * 100

DATA VISUALIZATION

TO IDENTIFY IMPORTANT FEATURES

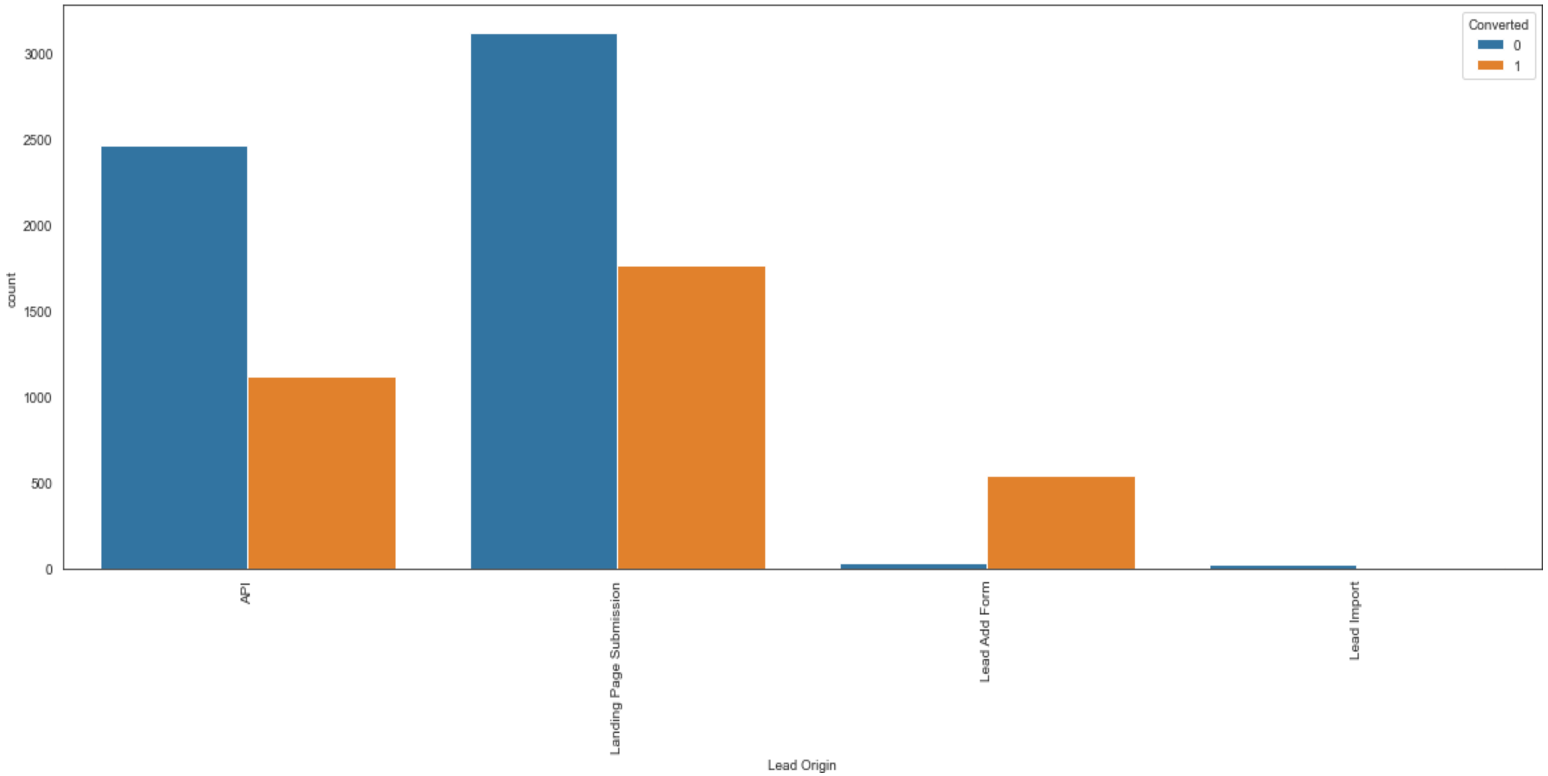
TO GET INSIGHTS

NUMERICAL VARIABLES

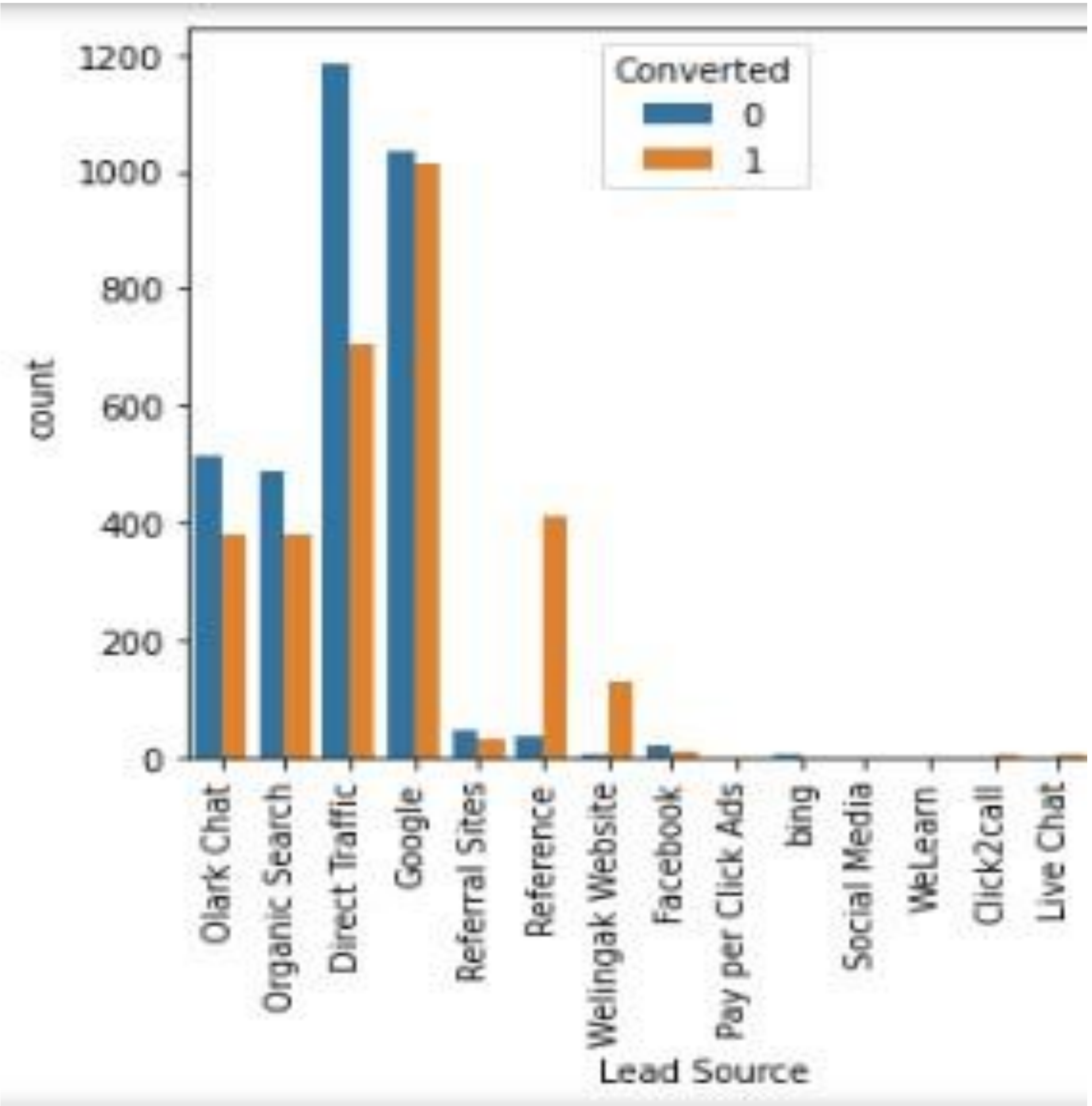


- ✓ People spending more time on website are more likely to get converted.

LEAD ORIGIN



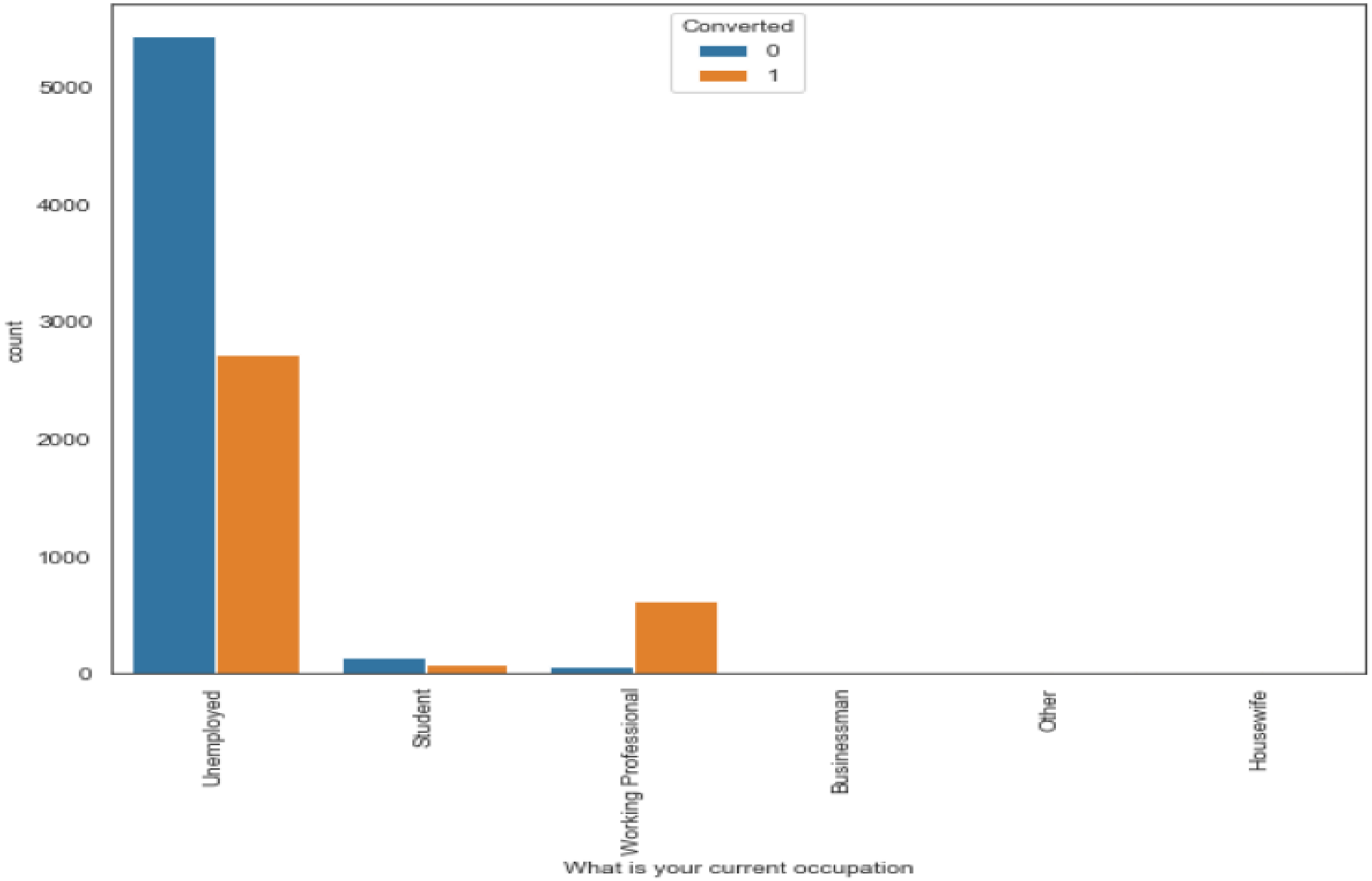
- ✓ ‘API ’ and Landing Page Submission ’ generate the most leads but have less conversion rates, whereas ‘ Lead Add Form ’ generates less leads but conversion rate is great.
- ✓ Try to increase conversion rate for ‘API’ and ‘Landing Page Submission’, and increase leads generation using ‘Lead Add Form’



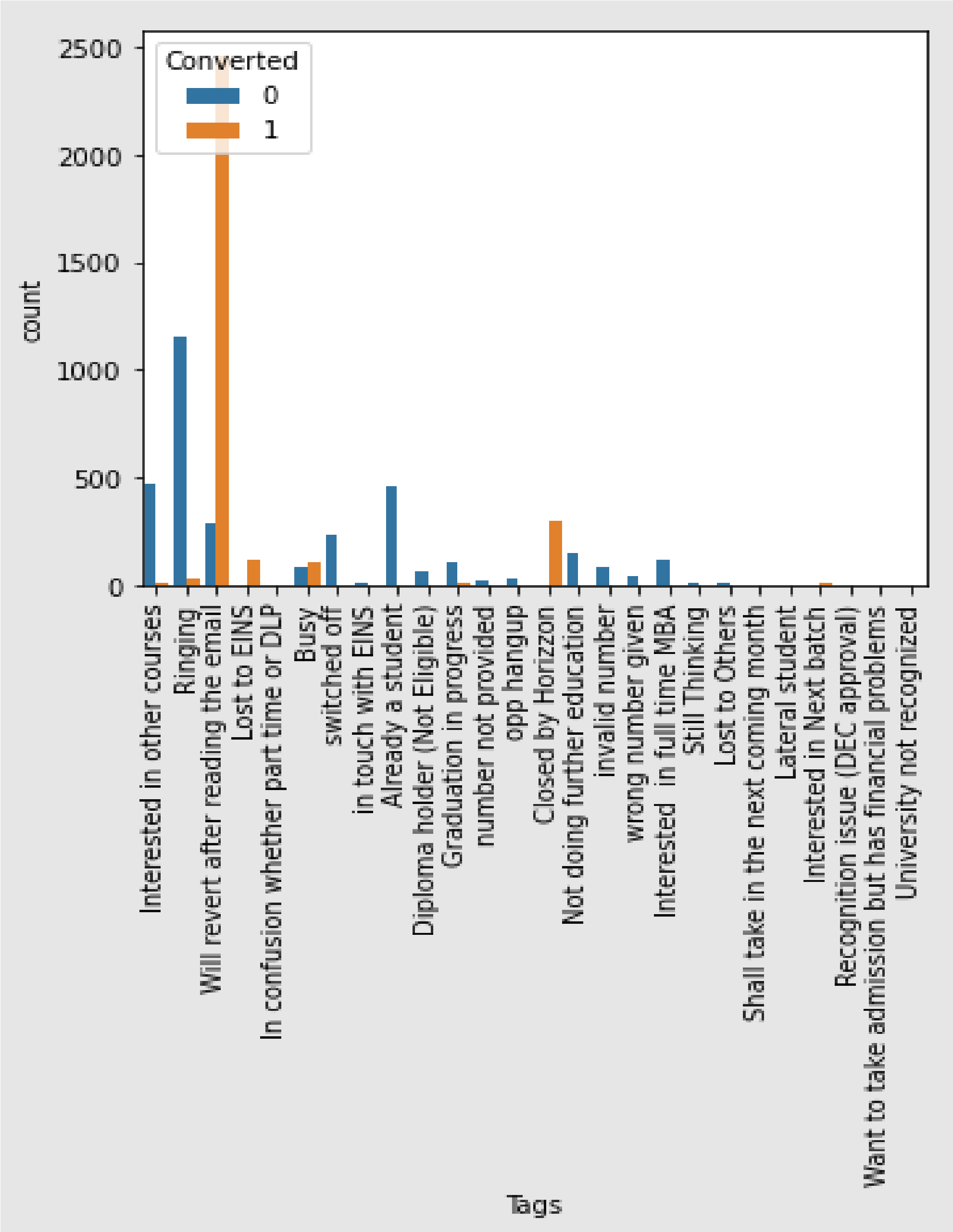
LEAD SOURCE

- ✓ Very high conversion rates for lead sources ‘Reference’ and ‘Welingak Website’.
- ✓ Most leads are generated through ‘Direct Traffic’ and Google.

CURRENT OCCUPATION

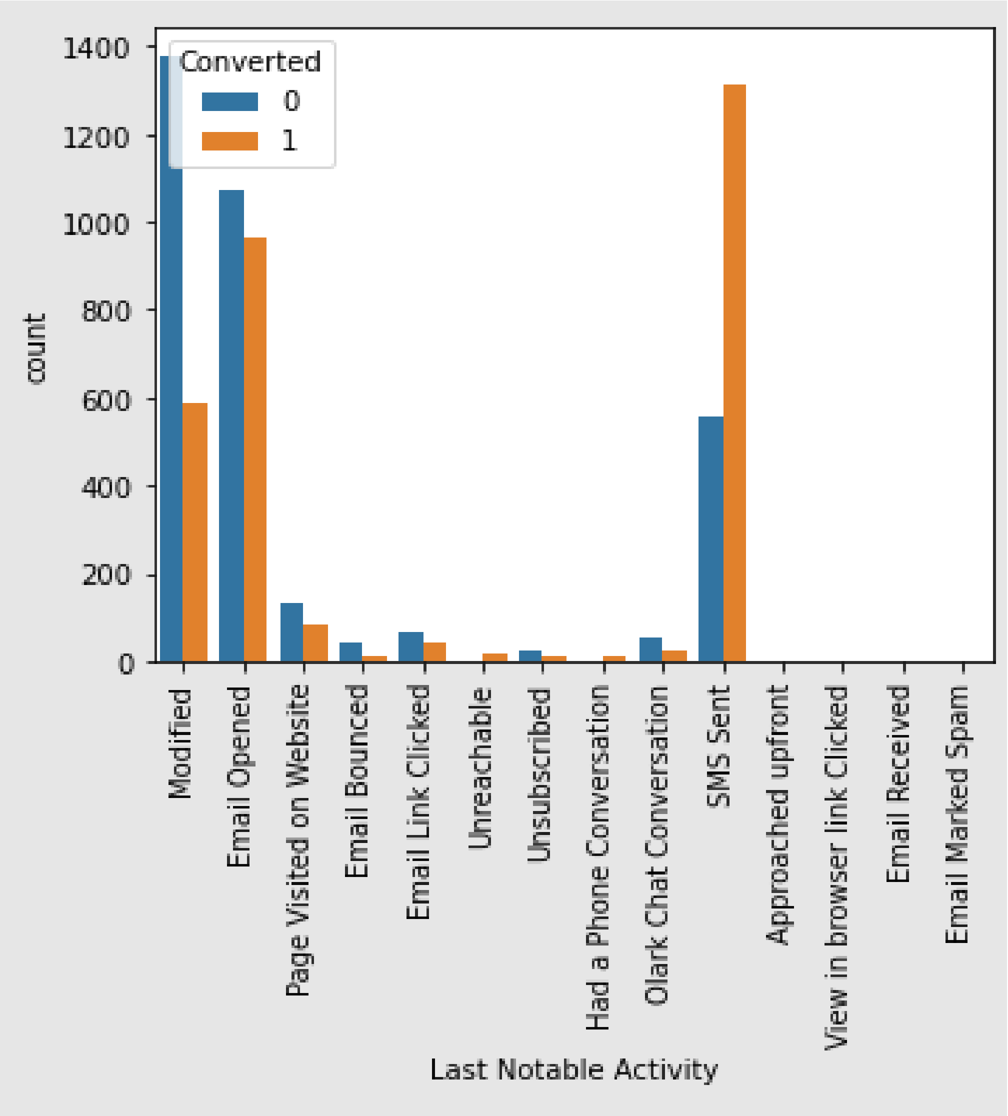


- ✓ Working professionals are most likely to get converted.



TAGS

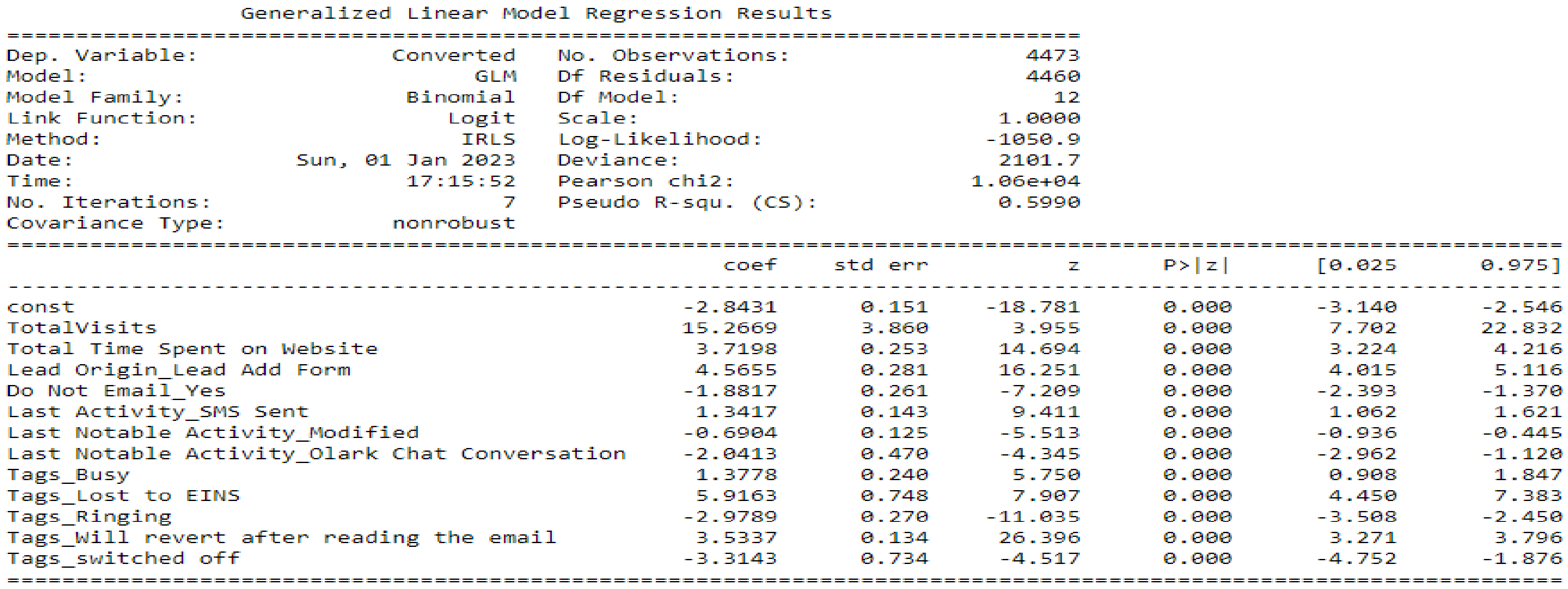
- ✓ *High conversion rates for tags ‘Will revert after reading the email’, ‘Closed by Horizon’, ‘Lost to EINS’ and ‘Busy’.*



LAST NOTABLE ACTIVITY

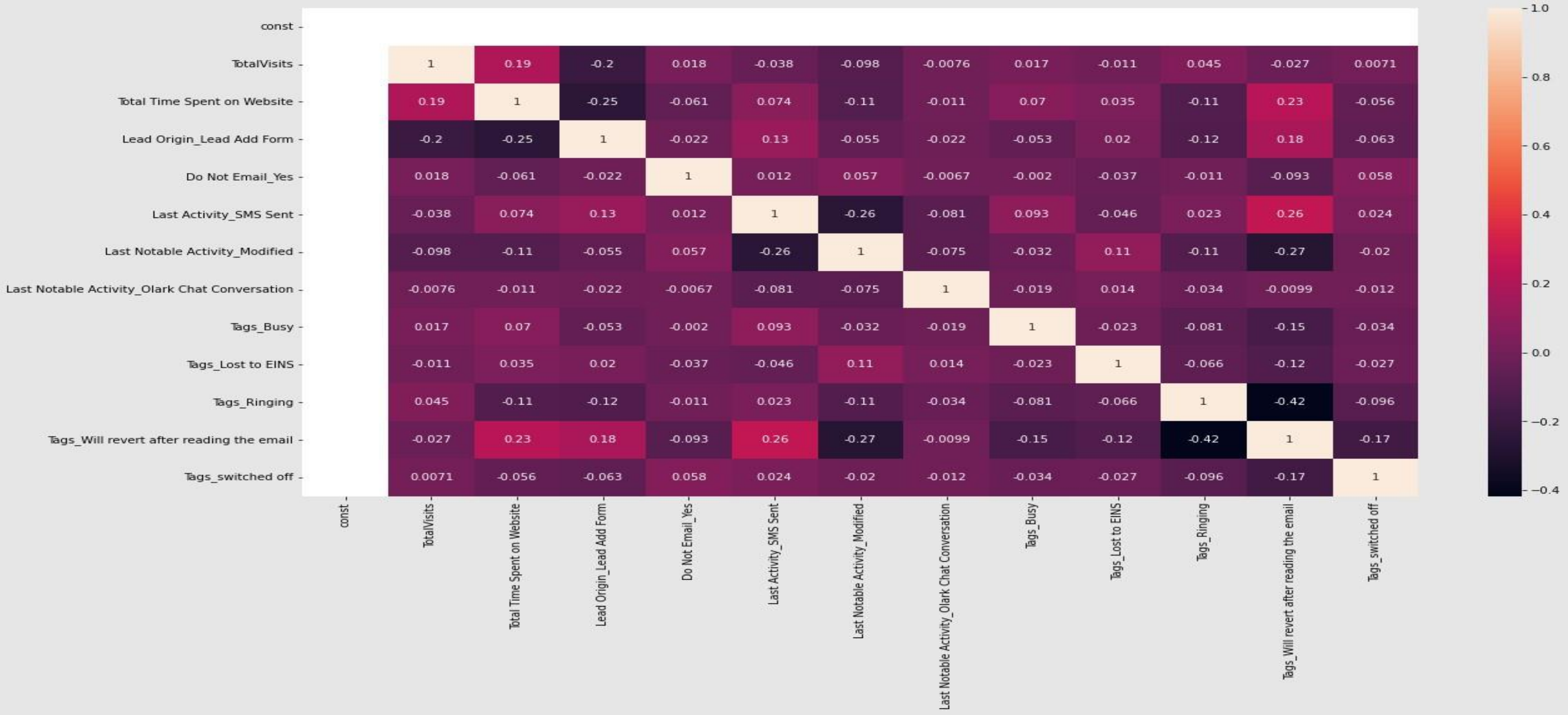
✓ *Highest conversion rate is for the last notable activity 'SMS Sent'.*

MODEL EVALUATION



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 4473
Model:                            GLM   Df Residuals:                     4460
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1050.9
Date:                Sun, 01 Jan 2023   Deviance:                       2101.7
Time:                        17:15:52   Pearson chi2:                 1.06e+04
No. Iterations:                     7   Pseudo R-squ. (CS):             0.5990
Covariance Type:            nonrobust
=================================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                                            -2.8431      0.151    -18.781      0.000      -3.140      -2.546
TotalVisits                                      15.2669      3.860      3.955      0.000       7.702      22.832
Total Time Spent on Website                       3.7198      0.253     14.694      0.000       3.224       4.216
Lead Origin_Lead Add Form                         4.5655      0.281     16.251      0.000       4.015       5.116
Do Not Email_Yes                                 -1.8817      0.261     -7.209      0.000      -2.393      -1.370
Last Activity_SMS Sent                            1.3417      0.143      9.411      0.000       1.062       1.621
Last Notable Activity_Modified                   -0.6904      0.125     -5.513      0.000      -0.936      -0.445
Last Notable Activity_Olark Chat Conversation    -2.0413      0.470     -4.345      0.000      -2.962      -1.120
Tags_Busy                                         1.3778      0.240      5.750      0.000       0.908       1.847
Tags_Lost to EINS                                 5.9163      0.748      7.907      0.000       4.450       7.383
Tags_Ringing                                     -2.9789      0.270    -11.035      0.000      -3.508      -2.450
Tags_Will revert after reading the email          3.5337      0.134     26.396      0.000       3.271       3.796
Tags_switched off                                -3.3143      0.734     -4.517      0.000      -4.752      -1.876
=================================================================================================================
```

FINAL MODEL SUMMARY: ALL P-VALUES ARE ZERO.



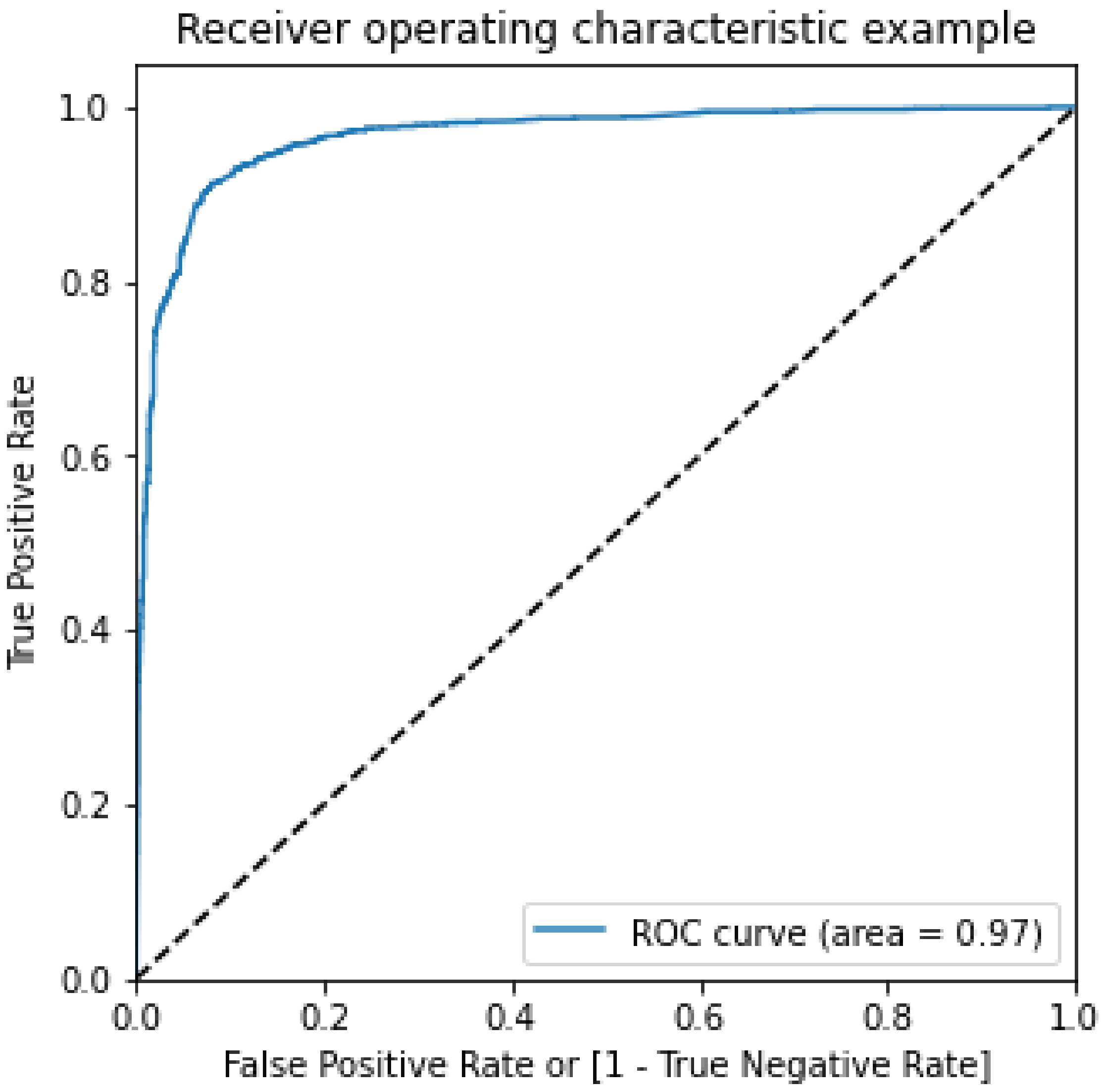
	const	TotalVisits	Total Time Spent on Website	Lead Origin_Lead Add Form	Do Not Email_Yes	Last Activity_SMS Sent	Last Notable Activity_Modified	Last Notable Activity_Olark Chat Conversation	Tags_Busy	Tags_Lost to EINS	Tags_Ringing	Tags_Will revert after reading the email	Tags_switched off
const													
TotalVisits		1	0.19	-0.2	0.018	-0.038	-0.098	-0.0076	0.017	-0.011	0.045	-0.027	0.0071
Total Time Spent on Website		0.19	1	-0.25	-0.061	0.074	-0.11	-0.011	0.07	0.035	-0.11	0.23	-0.056
Lead Origin_Lead Add Form		-0.2	-0.25	1	-0.022	0.13	-0.055	-0.022	-0.053	0.02	-0.12	0.18	-0.063
Do Not Email_Yes		0.018	-0.061	-0.022	1	0.012	0.057	-0.0067	-0.002	-0.037	-0.011	-0.093	0.058
Last Activity_SMS Sent		-0.038	0.074	0.13	0.012	1	-0.26	-0.081	0.093	-0.046	0.023	0.26	0.024
Last Notable Activity_Modified		-0.098	-0.11	-0.055	0.057	-0.26	1	-0.075	-0.032	0.11	-0.11	-0.27	-0.02
Last Notable Activity_Olark Chat Conversation		-0.0076	-0.011	-0.022	-0.0067	-0.081	-0.075	1	-0.019	0.014	-0.034	-0.0099	-0.012
Tags_Busy		0.017	0.07	-0.053	-0.002	0.093	-0.032	-0.019	1	-0.023	-0.081	-0.15	-0.034
Tags_Lost to EINS		-0.011	0.035	0.02	-0.037	-0.046	0.11	0.014	-0.023	1	-0.066	-0.12	-0.027
Tags_Ringing		0.045	-0.11	-0.12	-0.011	0.023	-0.11	-0.034	-0.081	-0.066	1	-0.42	-0.096
Tags_Will revert after reading the email		-0.027	0.23	0.18	-0.093	0.26	-0.27	-0.0099	-0.15	-0.12	-0.42	1	-0.17
Tags_switched off		0.0071	-0.056	-0.063	0.058	0.024	-0.02	-0.012	-0.034	-0.027	-0.096	-0.17	1

HEAT MAP

Correlations between features in the final model are negligible.

ROC CURVE

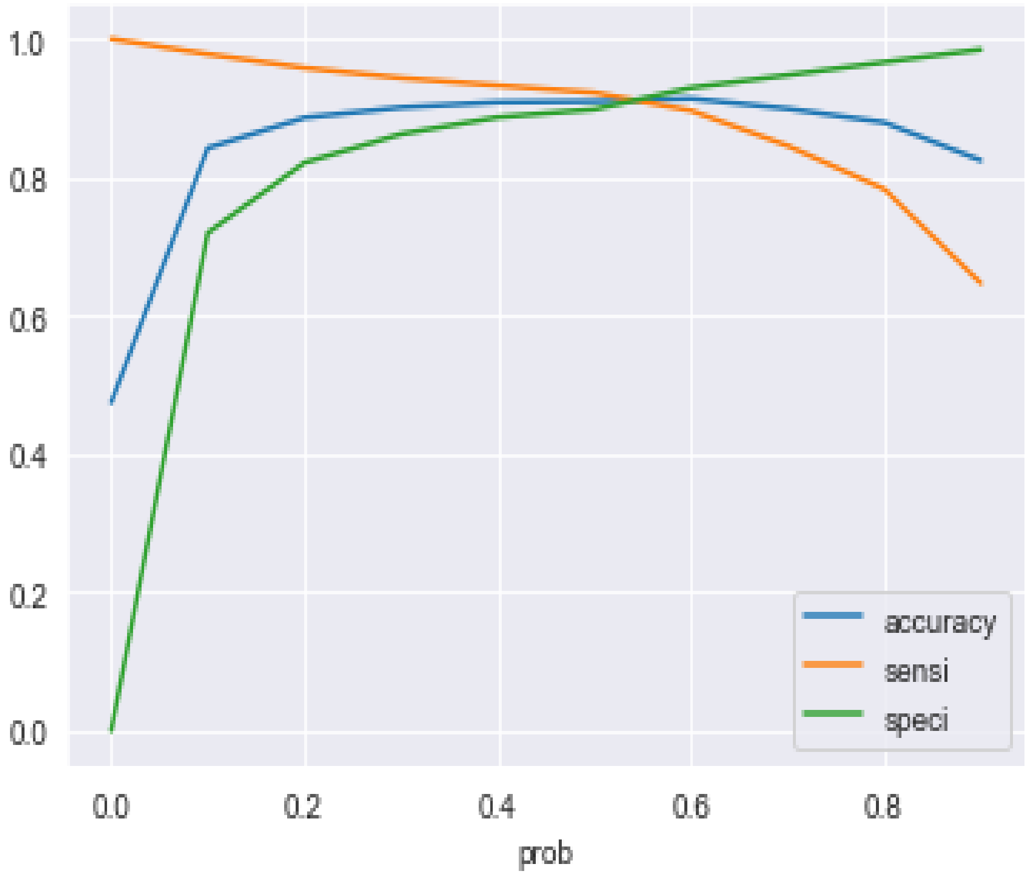
Area under curve = 0.97



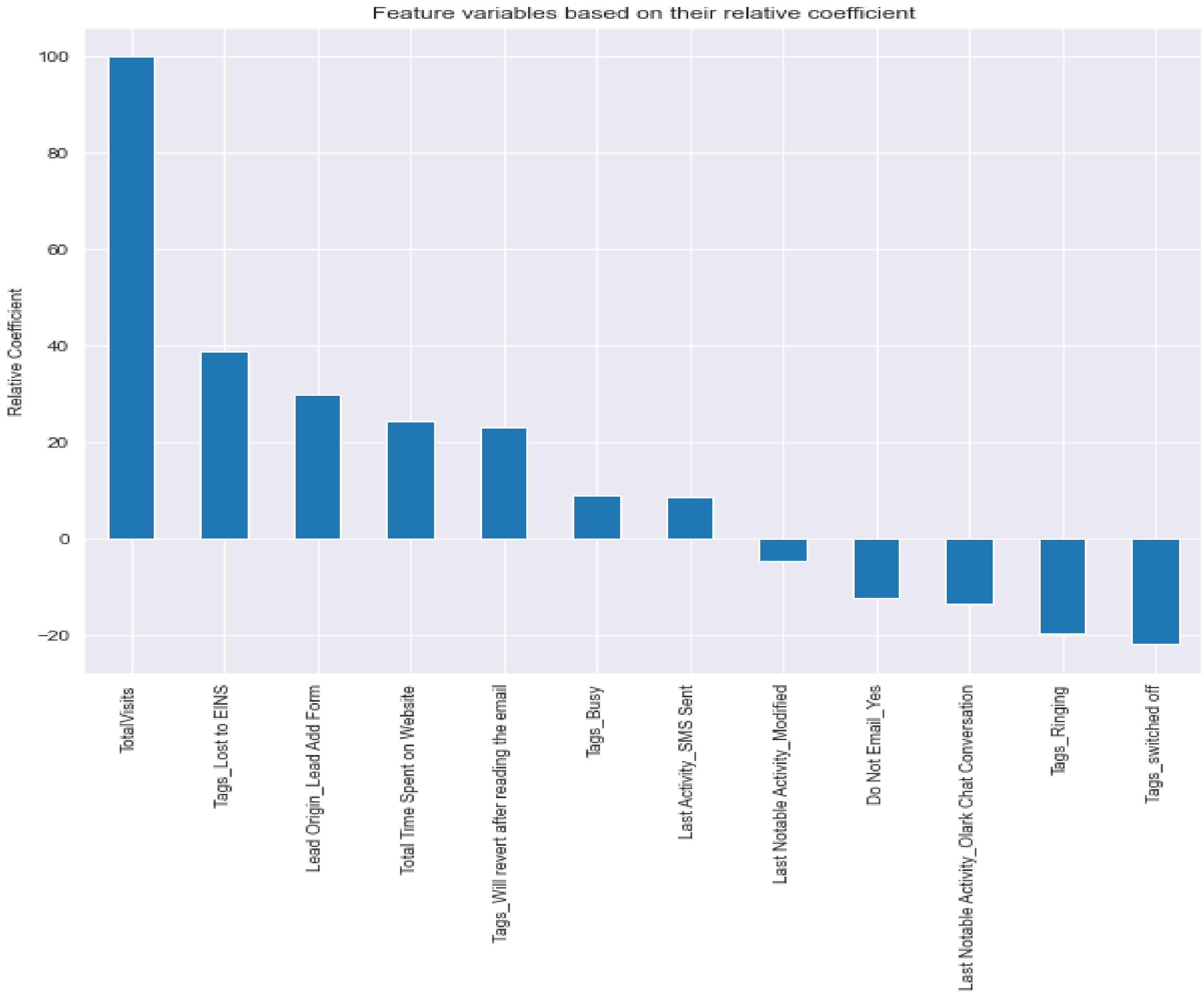
FINDING OPTIMAL THRESHOLD

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

Optimal cutoff = 0.55



RELATIVE IMPORTANCE OF FEATURES



INFERENCES

FEATURE IMPORTANCE

✓ Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are

I. TotalVisits

II.Tags_Lost to EINS

III.Lead Origin_Lead Add Form

✓ 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

- Situation 1: Company has interns for 2 months. They wish to make lead conversion more aggressive. They want almost all of the potential leads to be converted and hence, want to make phone calls to as much of such people as possible.
- Solution:

✓ $Sensitivity = TruePositives / \ TruePositives + \ FalseNegatives$)

✓ Sensitivity can be defined as the number of actual conversions predicted correctly out of 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**.

- Situation 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:

✓ $Specificity = TrueNegatives / TrueNegatives + FalsePositives$**)**

✓ Specificity can be defined as the number of actual non conversions predicted correctly out of 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

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