LEAD SCORING CASE STUDY

SUBMITTED BY:

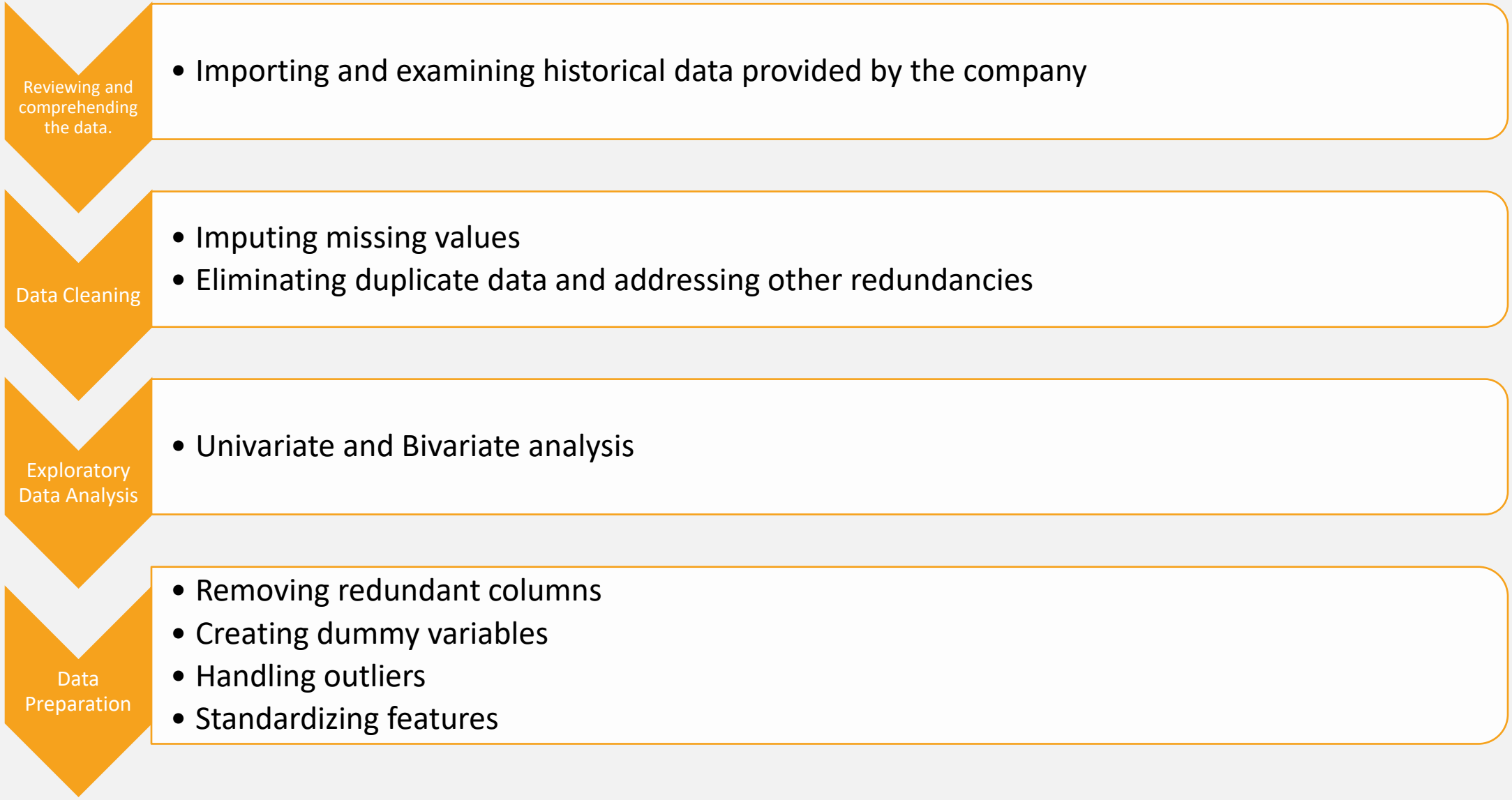
SHRIKANT
KOMAL
DAMANPREET

Business Objective

To assist X Education in identifying the most prospective leads, or those with the highest likelihood of becoming paying customers.

METHODOLOGY

To construct a Logistic Regression model that assigns lead scores to all leads in a way that customers with higher lead scores are more likely to convert, and conversely, those with lower scores are less likely to convert, with the goal of achieving a target leads conversion rate of approximately 80%.


Reviewing and comprehending the data.
• Importing and examining historical data provided by the company
Data Cleaning
• Imputing missing values
• Eliminating duplicate data and addressing other redundancies
Exploratory Data Analysis
• Univariate and Bivariate analysis
Data Preparation
• Removing redundant columns
• Creating dummy variables
• Handling outliers
• Standardizing features

Model Building

- Employing Recursive Feature Elimination (RFE) for feature selection
- Manually eliminating features based on p-values and VIFs

Assessing the model's performance

- Assessing the model using a variety of evaluation metrics
- Determining the optimal probability threshold

Comparison to PCA

- Developing a second model using PCA
- Conducting a comparison between the two models

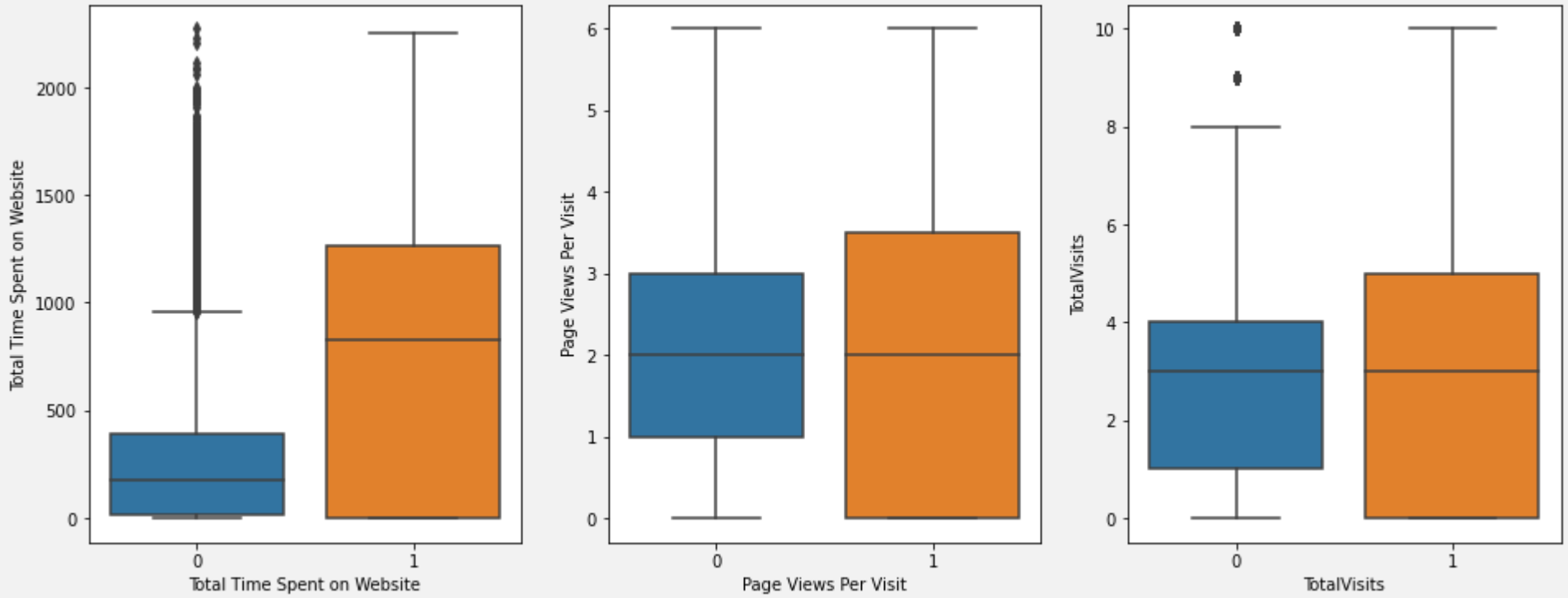
Lead Score Assignment

- Finalizing the initial model
- Utilizing forecasted probabilities to compute Lead Scores:
- Lead Score = Probability * 100

DATA VISUALISATION

- Identification of crucial features
- To gain insights

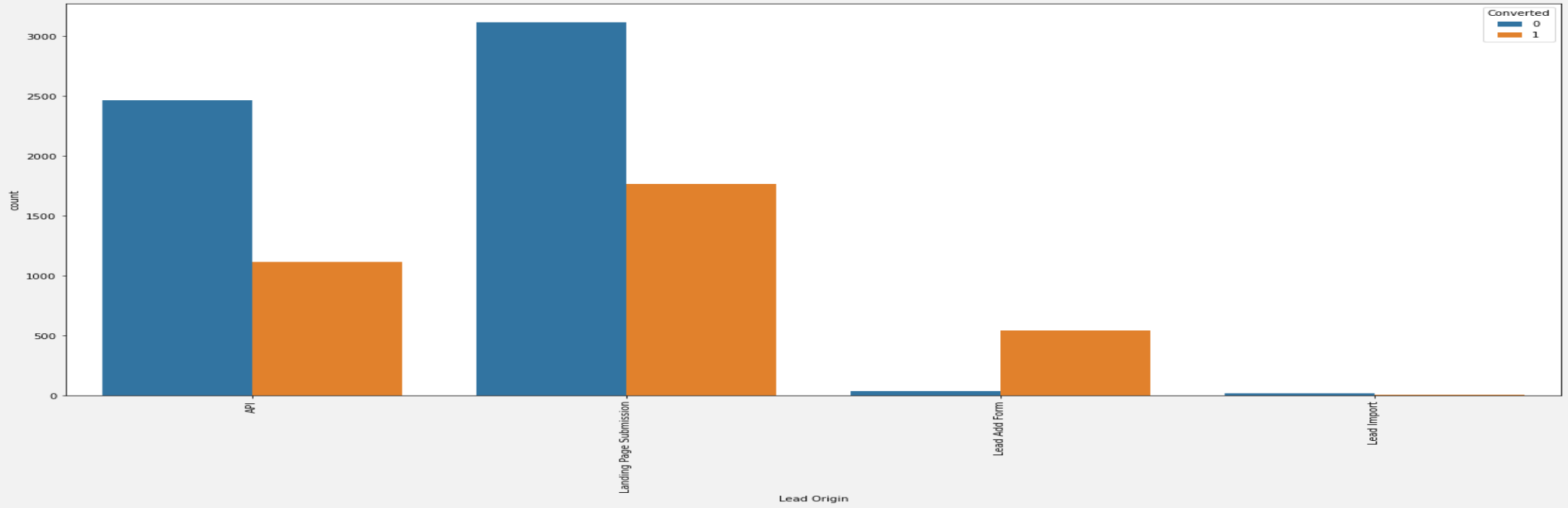
Continuous Variables



The more time people spend on the website, the higher the likelihood of conversion.

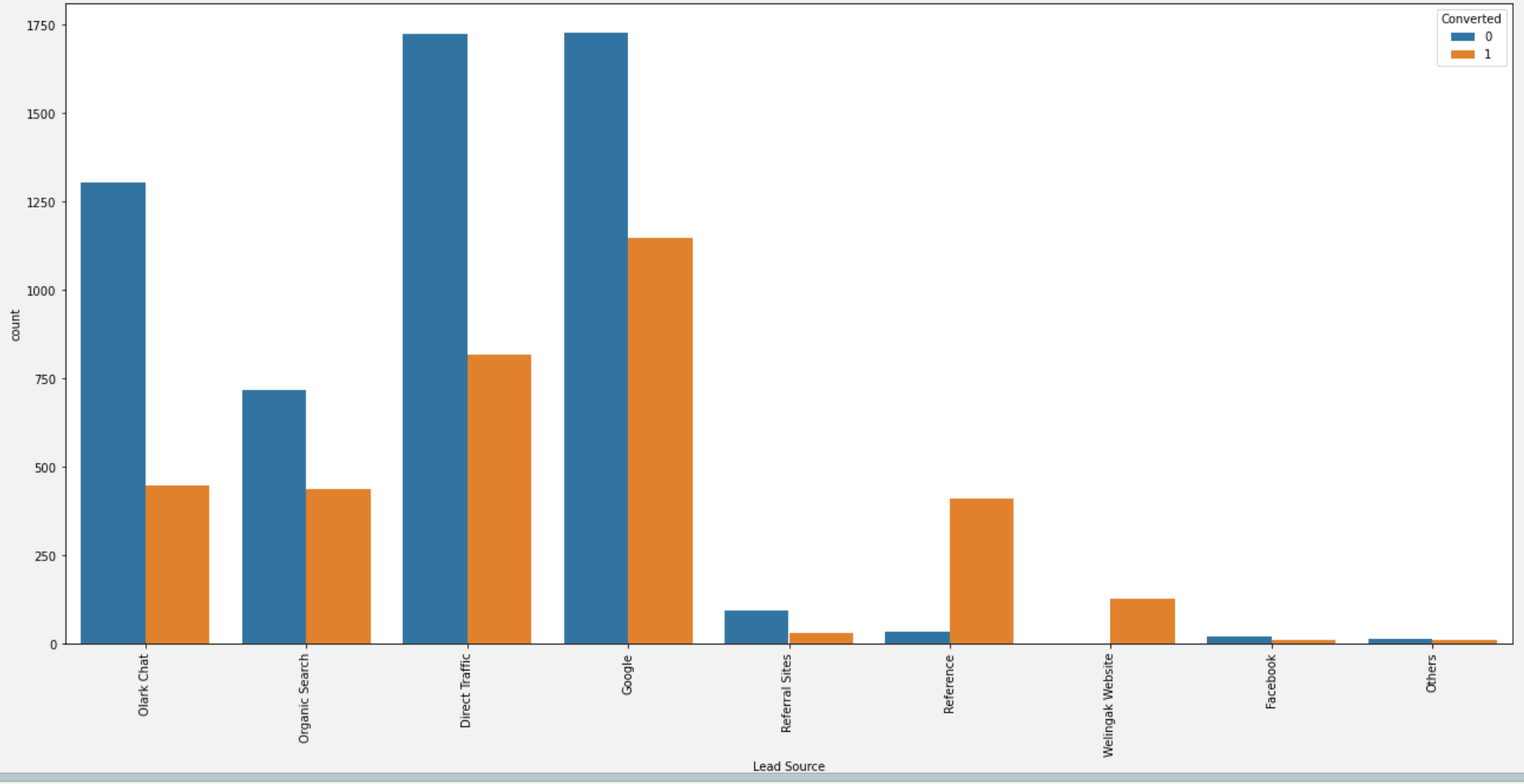
Lead Origin

- *'API' and 'Landing Page Submission' yield a higher number of leads but exhibit lower conversion rates, while 'Lead Add Form' generates fewer leads with a significantly better conversion rate.*
- *The objective is to enhance the conversion rates for 'API' and 'Landing Page Submission' while increasing lead generation through 'Lead Add Form.'*



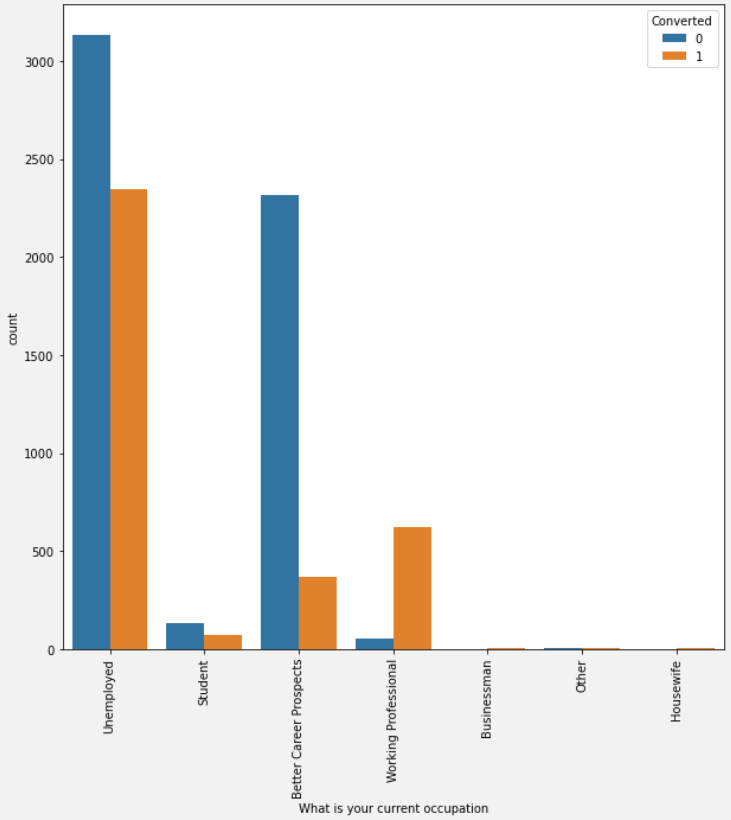
Lead Source

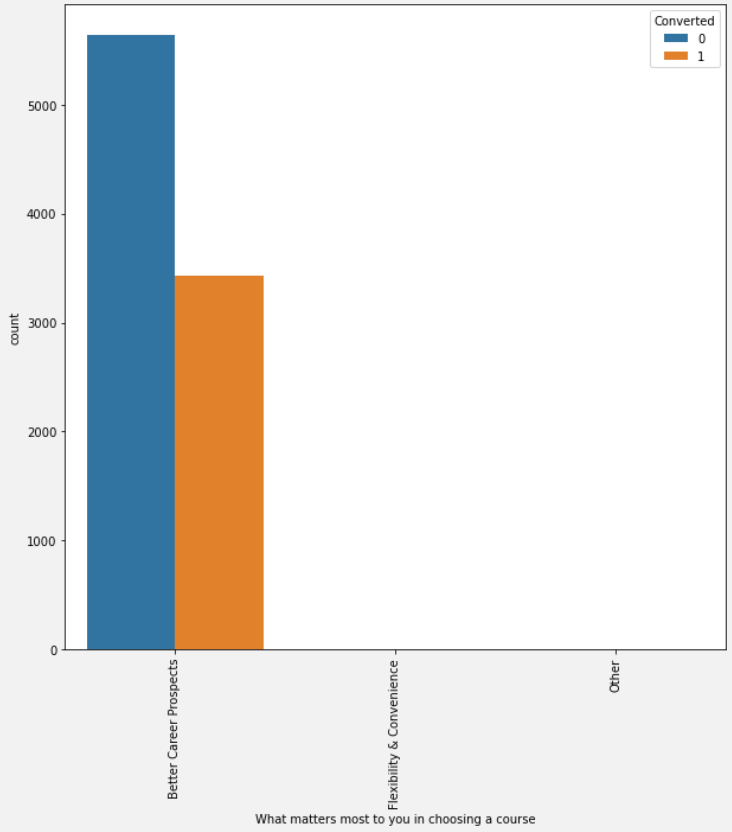
- There are exceptionally high conversion rates for lead sources 'Reference' and 'Welingak Website'
- while the bulk of leads come from 'Direct Traffic' and 'Google.'



Current Occupation

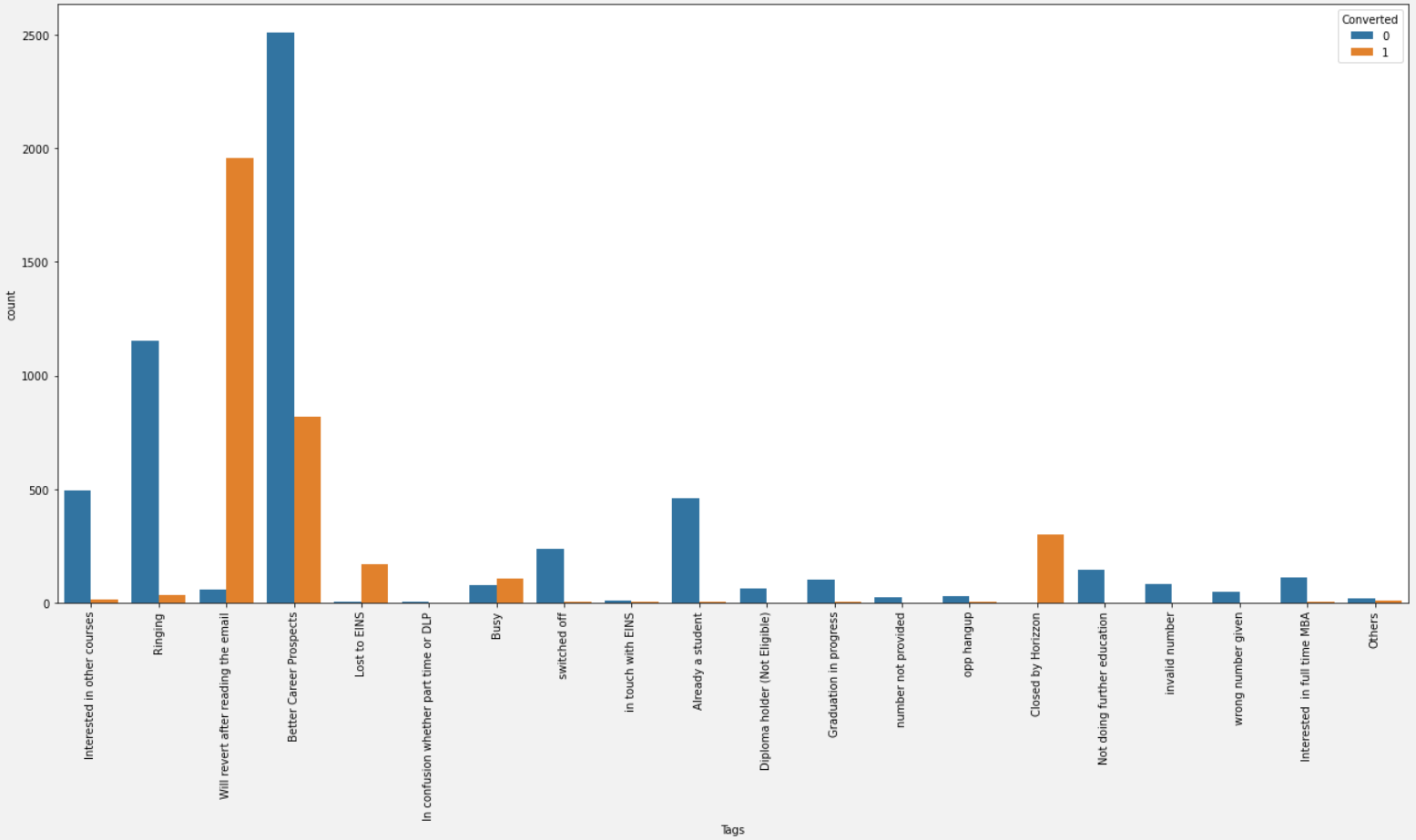
Working professionals have the highest likelihood of conversion





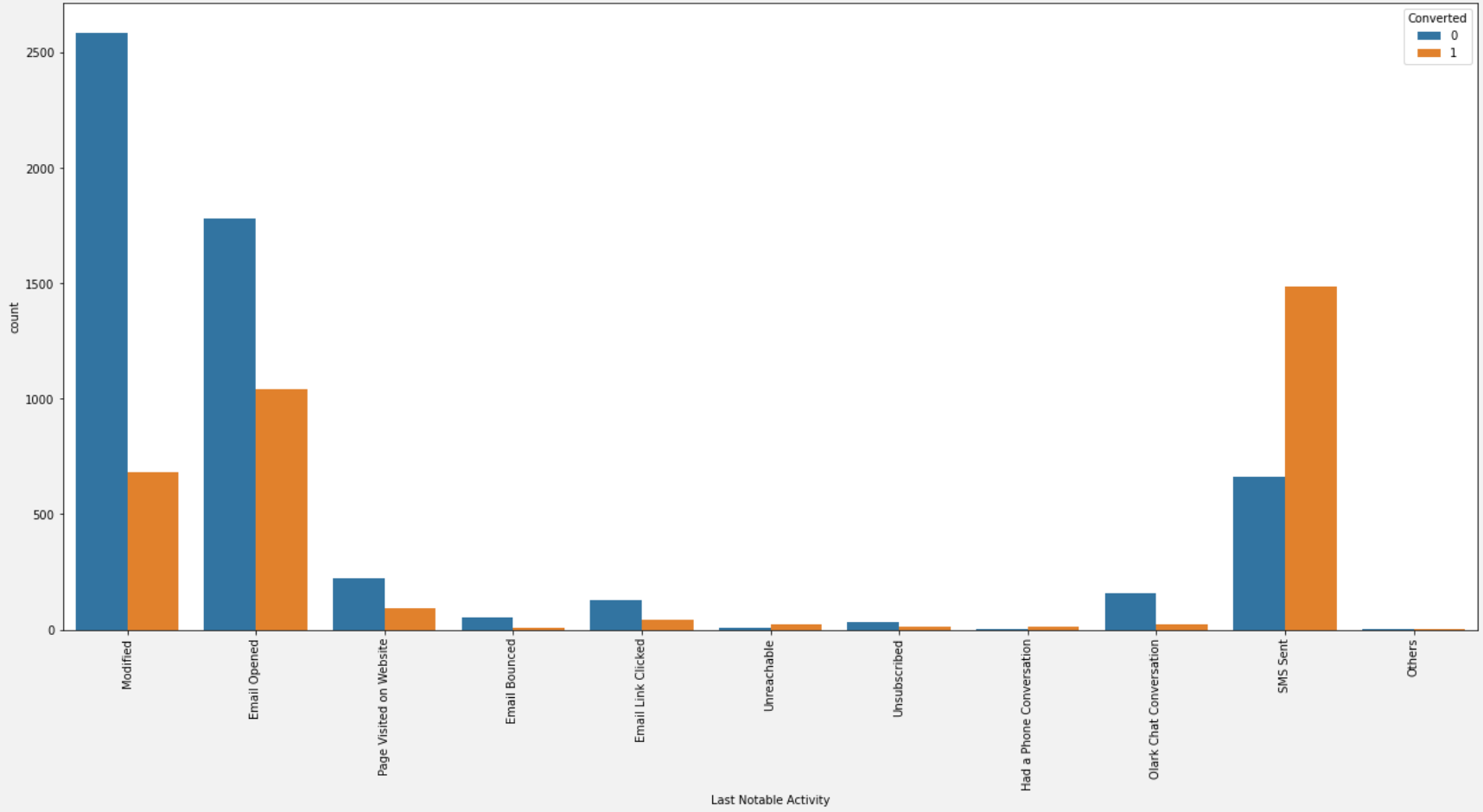
Tags

There are notably high conversion rates associated with the tags 'Will revert after reading the email,' 'Closed by Horizon,' 'Lost to EINS,' and 'Busy'



Last Notable Activity

The highest conversion rate is observed for the most recent notable activity, which is 'SMS Sent.'



MODEL EVALUATION

Final Model Summary:

Generalized Linear Model Regression Results

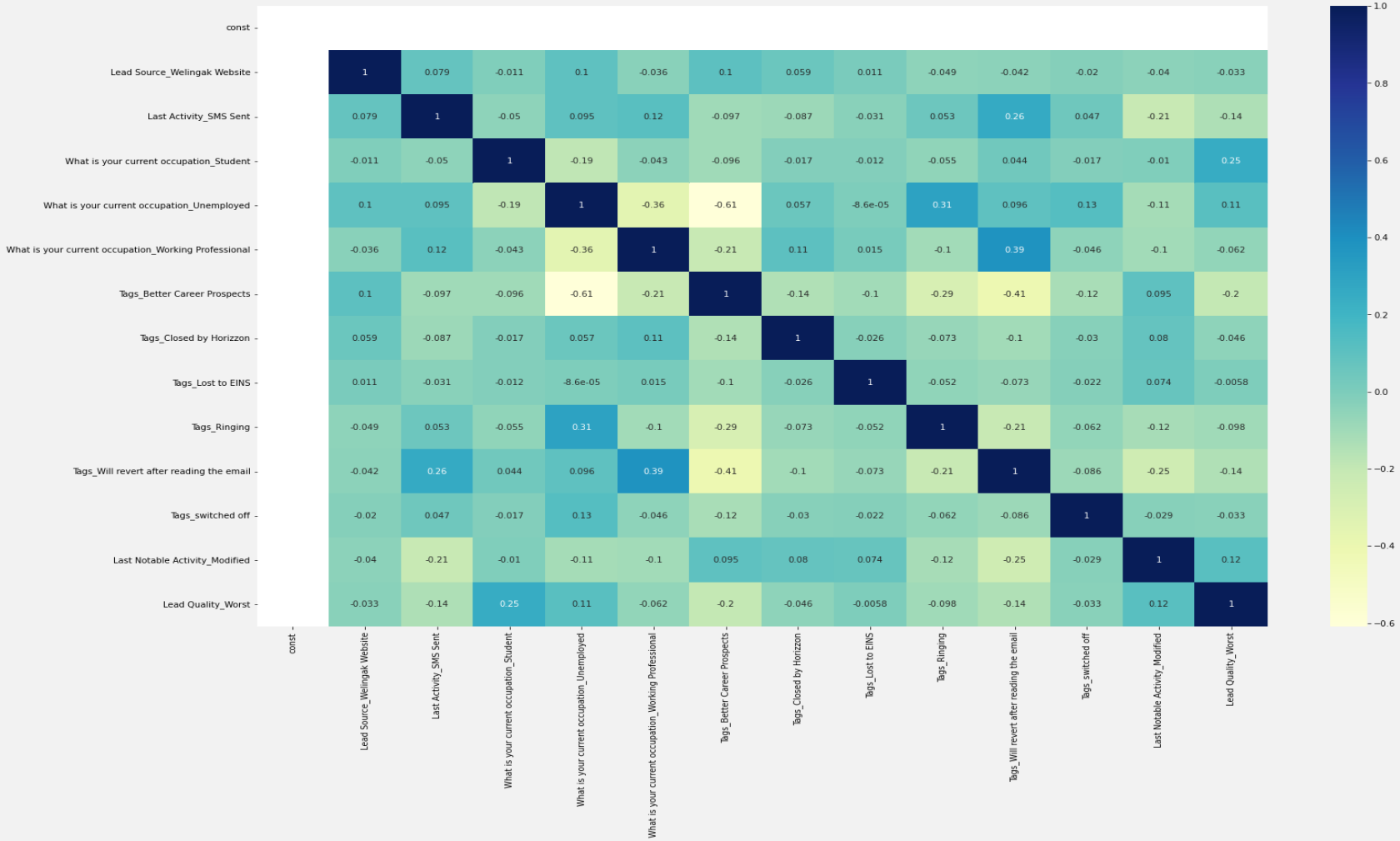
Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6337
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1248.2
Date:	Sun, 15 Oct 2023	Deviance:	2496.5
Time:	14:34:07	Pearson chi2:	1.27e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.8017	0.210	-22.838	0.000	-5.214	-4.390
Lead Source_Welingak Website	2.6384	0.745	3.541	0.000	1.178	4.099
Last Activity_SMS Sent	2.3548	0.119	19.856	0.000	2.122	2.587
What is your current occupation_Student	2.1321	0.531	4.019	0.000	1.092	3.172
What is your current occupation_Unemployed	2.4685	0.144	17.183	0.000	2.187	2.750
What is your current occupation_Working Professional	2.5943	0.373	6.962	0.000	1.864	3.325
Tags_Better Career Prospects	2.6600	0.174	15.328	0.000	2.320	3.000
Tags_Closed by Horizzon	8.2182	0.739	11.121	0.000	6.770	9.667
Tags_Lost to EINS	8.9118	0.777	11.466	0.000	7.388	10.435
Tags_Ringing	-2.7255	0.264	-10.315	0.000	-3.243	-2.208
Tags_Will revert after reading the email	5.9657	0.235	25.378	0.000	5.505	6.426
Tags_switched off	-2.8390	0.530	-5.355	0.000	-3.878	-1.800
Last Notable Activity_Modified	-1.6637	0.127	-13.111	0.000	-1.912	-1.415
Lead Quality_Worst	-2.9147	0.729	-4.000	0.000	-4.343	-1.486

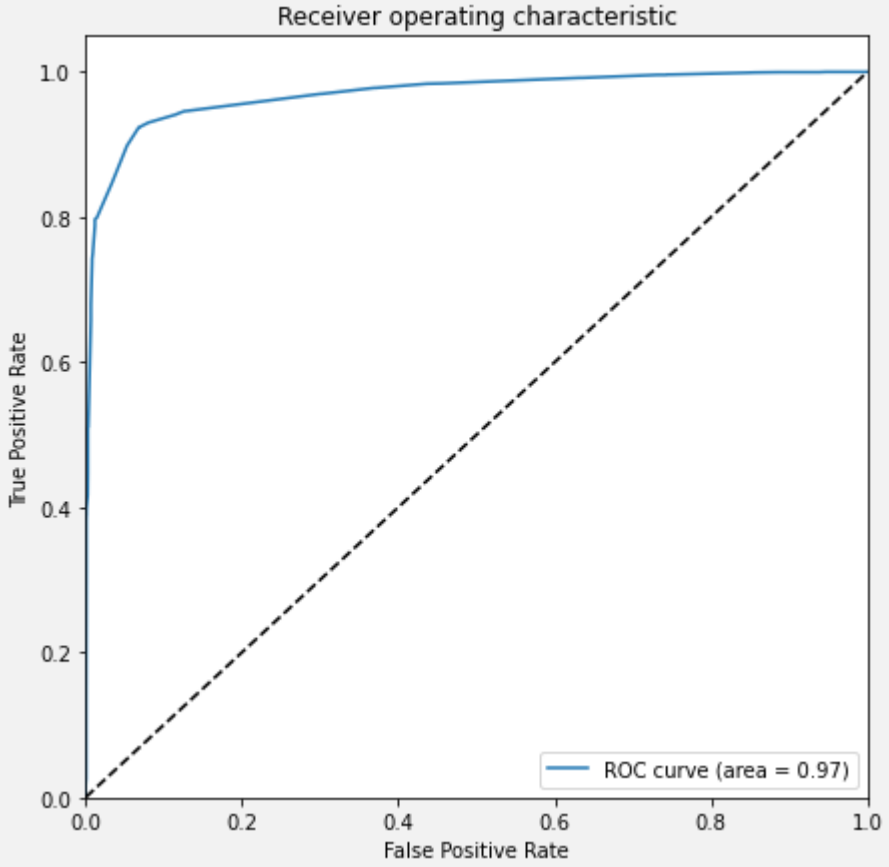
All p-values are equal to zero, indicating a highly significant statistical relationship

Multicollinearity

The correlations between features in the final model are minimal or insignificant.

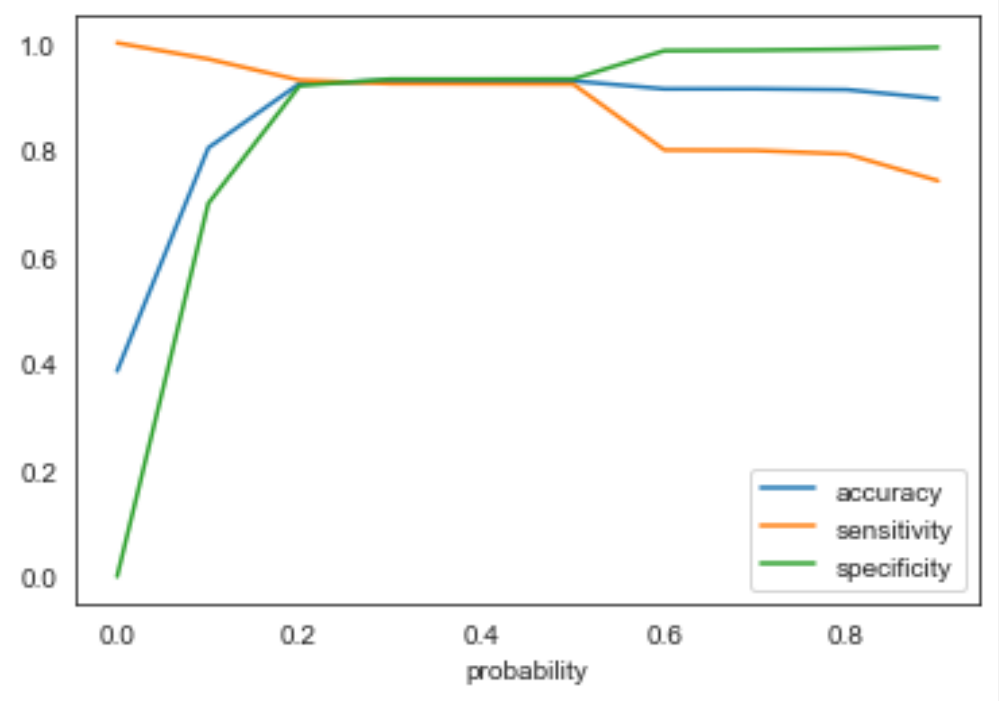


ROC curve



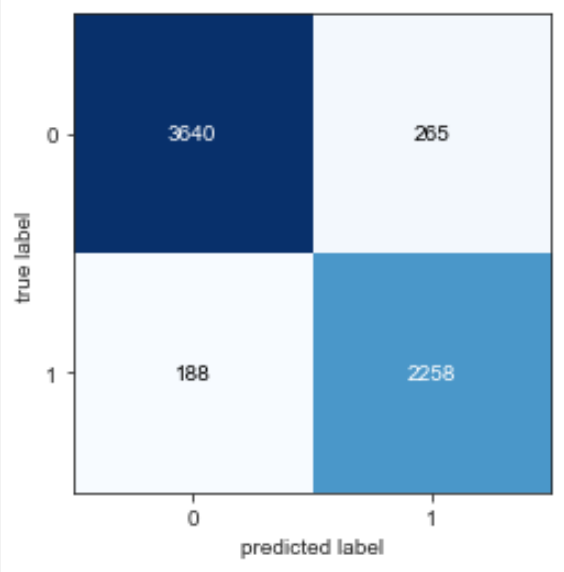
Area under curve (ROC) = 0.97

Optimal Threshold

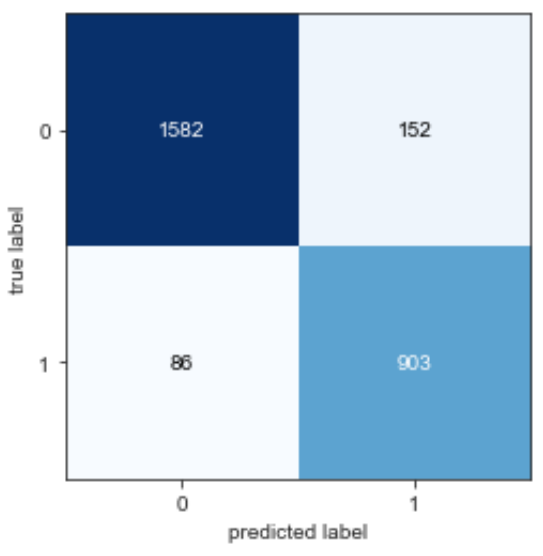


A graph displaying how Sensitivity, Specificity, and Accuracy change with varying probability threshold values, with the optimal cutoff set at 0.20

Confusion Matrix



For train set

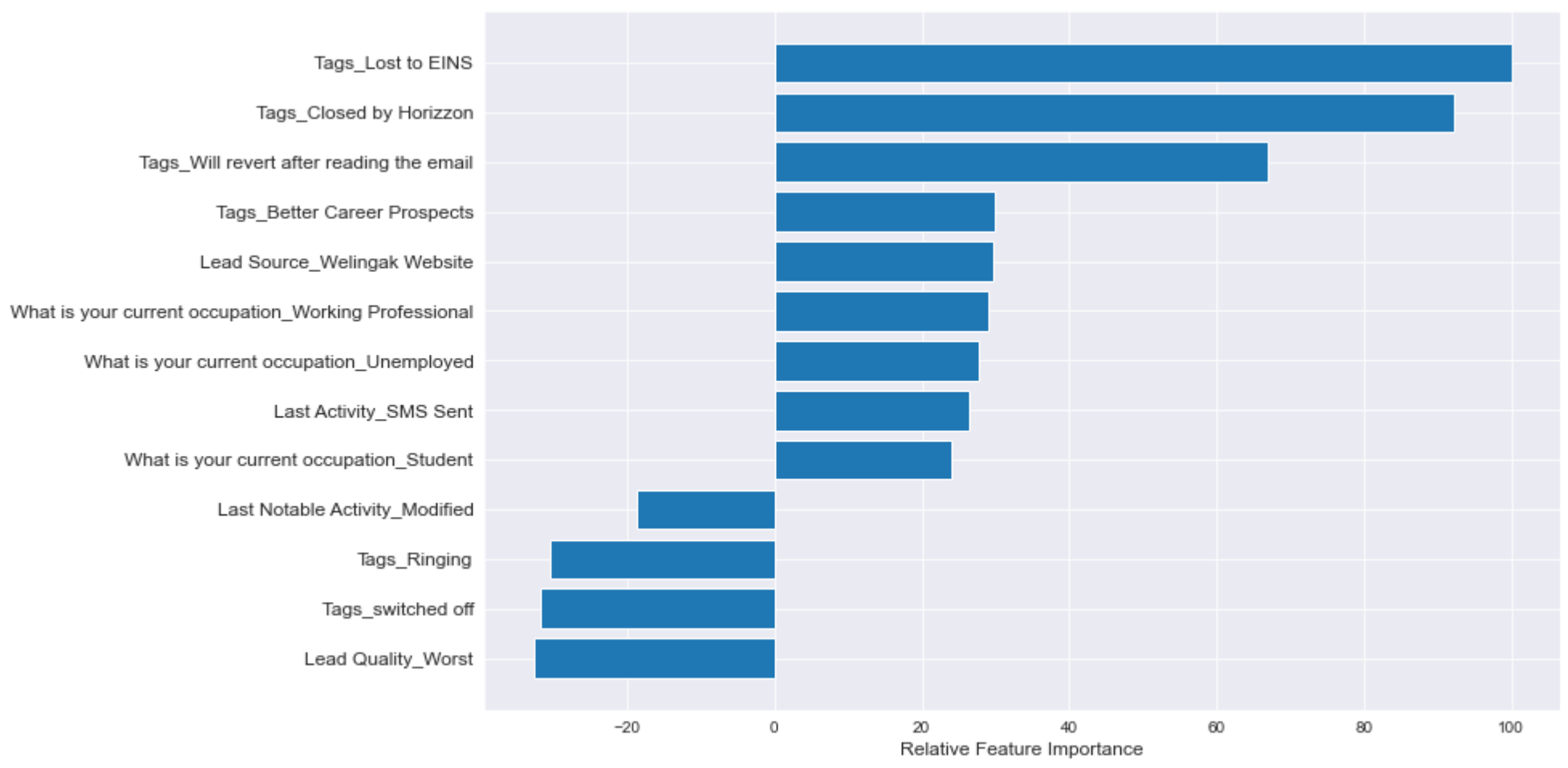


For test set

Final Results

Data	Train Set	Test Set
Accuracy	0.9235	0.9130
Specificity	0.9321	0.9123
False Positive Rate	0.0678	0.0876
Positive Predictive Value	0.8950	0.8559
Negative Predictive Value	0.9511	0.9484
AUC	0.9592	0.9427

Relative Importance of Features



INFERENCES

Feature Importance

- The top three variables with the most substantial influence on the probability of lead conversion, listed in decreasing order of impact, are:
  - *Tags_Lost to EINS*
  - Tags_Closed by Horizzon
  - Tags_Will revert after reading the email
- These are dummy features derived from the categorical variable Tags
- All three have a positive impact on the probability of lead conversion
- These findings suggest that the company should prioritize 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 = True Positives / (True Positives + False Negatives)***
- Sensitivity can be defined as the proportion of actual conversions that are correctly predicted out of the total number of actual conversions. As previously observed, sensitivity tends to decrease as the threshold value increases
- A high sensitivity suggests that our model will accurately predict nearly all leads that are likely to convert. However, it might also result in the model classifying some non-conversions as conversions, potentially leading to false positives
- Given the company's additional manpower for two months and their goal of pursuing more aggressive lead conversion, opting for high sensitivity is a sound strategy. To achieve high sensitivity, it's advisable to select a lower threshold value. This approach would prioritize capturing a maximum number of potential conversions while potentially accepting a few false positives.

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:

- ***Sensitivity = True Negatives / (True Negatives + False Positives)***
- Specificity can be defined as the proportion of actual non-conversions that are correctly predicted out of the total number of actual non-conversions. It generally increases as the threshold value is raised
- A high specificity indicates that our model will accurately predict nearly all leads who are unlikely to convert. However, it might result in the misclassification of some actual conversions as non-conversions, potentially leading to false negatives
- Since the company has already achieved its quarterly target and wants to avoid unnecessary phone calls, pursuing high specificity is a wise strategy. High specificity would prioritize correctly identifying non-conversions, minimizing the unnecessary allocation of resources to leads that are unlikely to convert, even though it may potentially miss some genuine conversion opportunities.
- Opting for high specificity will ensure that phone calls are primarily directed to customers with a very high likelihood of conversion, minimizing wasted resources. To achieve high specificity, it's advisable to select a high threshold value

Recommendations

- *By referring to the data visualizations, concentrate on the category "Tags_Lost to EINS"*

  a) Enhancing the conversion rates for the categories that generate a higher volume of leads

  b) Generating additional leads for the categories with already high conversion rates

- Focus on the relative importance of the features in the model and their influence, whether positive or negative, on the likelihood of conversion.
- Adjust the probability threshold value for identifying potential leads according to specific business requirements and objectives

THANK YOU