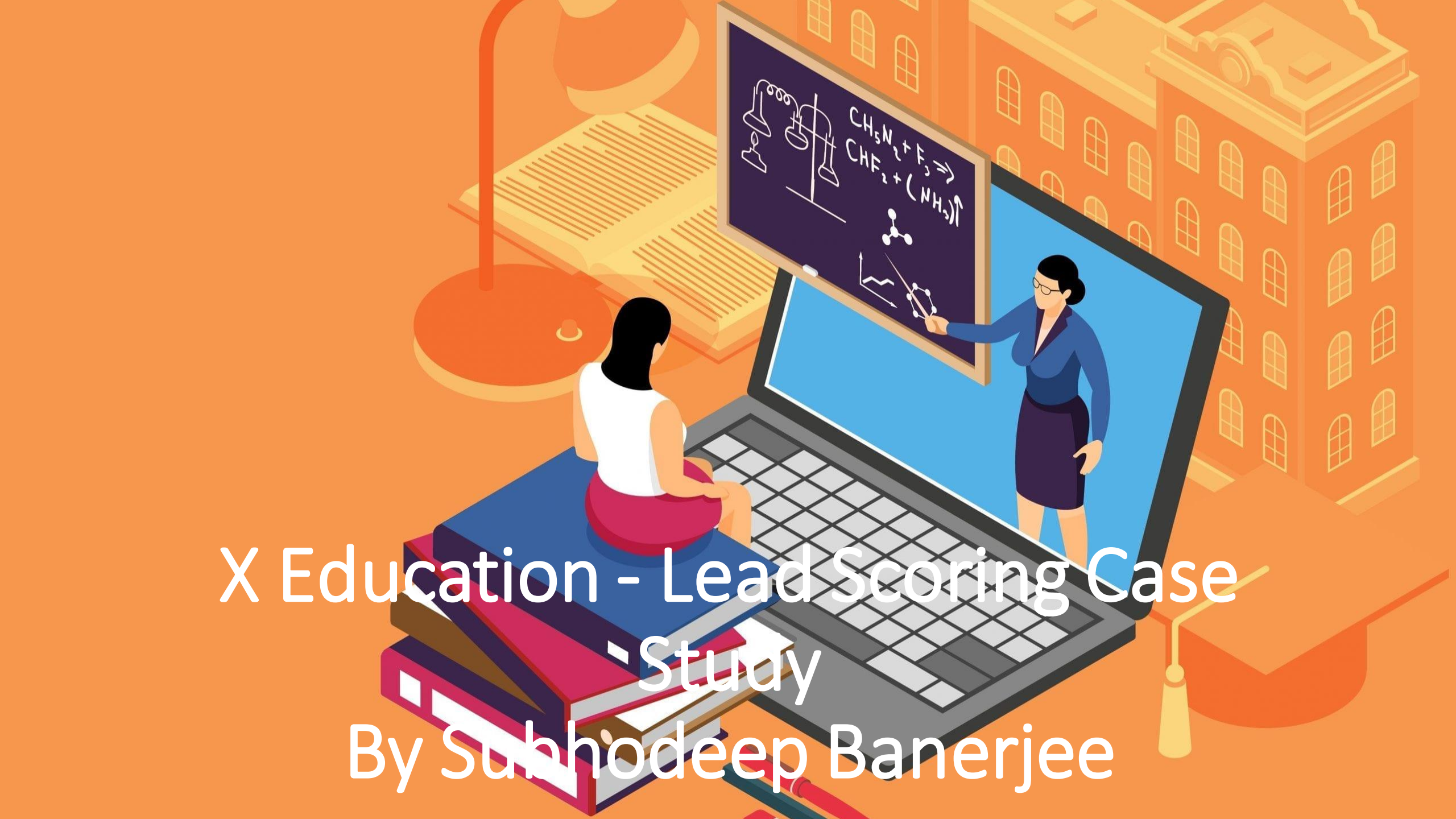


X Education - Lead Scoring Case Study

By Subhodeep Banerjee

Table of contents:

Background of X Education Company

- X Education, an online education provider, caters to professionals seeking courses by promoting its offerings across various platforms like search engines and partner websites.
- Visitors arriving on their website may explore available courses, submit forms, or engage with course-related videos.
- When individuals submit forms, providing contact details like email or phone, they become categorized as leads.
- Following lead acquisition, the sales team initiates contact through calls, emails, etc.
- During this engagement, some leads successfully convert, although the majority do not.
- The average lead conversion rate for X Education stands at approximately 30%.

Problem Statement & Objective of the Study

Problem Statement:

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30% .
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, al so known as Hot Leads .
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such t hat the customer with a higher lead score has a higher conversion rate.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Suggested Ideas for Lead Conversion

- **Leeds Grouping -**
  - Leads are categorized based on their probability or inclination to convert.
  - This categorization leads to a concentrated group of promising or high-potential leads.

- **Boost Conversion -**
  - We can achieve a higher conversion rate and reach the 80% target by prioritizing leads with a higher probability of converting.

Since we have a target of 80% conversion rate, we would want to obtain a high **sensitivity** in obtaining hot leads.

Analysis Approach

Data Cleaning

- The "Select" category represents null or unselected values for certain categorical variables, indicating that customers did not choose any option from the available list.
- Columns containing more than 40% null values were removed from the dataset.
- Handling missing values in categorical columns involved strategies based on value counts and specific considerations.
- Columns that did not provide relevant insights aligned with the study's objectives, such as 'tags' and 'country', were dropped.
- Imputation techniques were employed for handling missing values in certain categorical variables.
- For some variables, additional categories were created to capture specific information.
- Columns deemed irrelevant for modeling purposes, such as 'Prospect ID' and 'Lead Number', or those containing only one category of response, were dropped from the dataset.
- Numerical data underwent imputation using the mode after assessing the distribution characteristics.

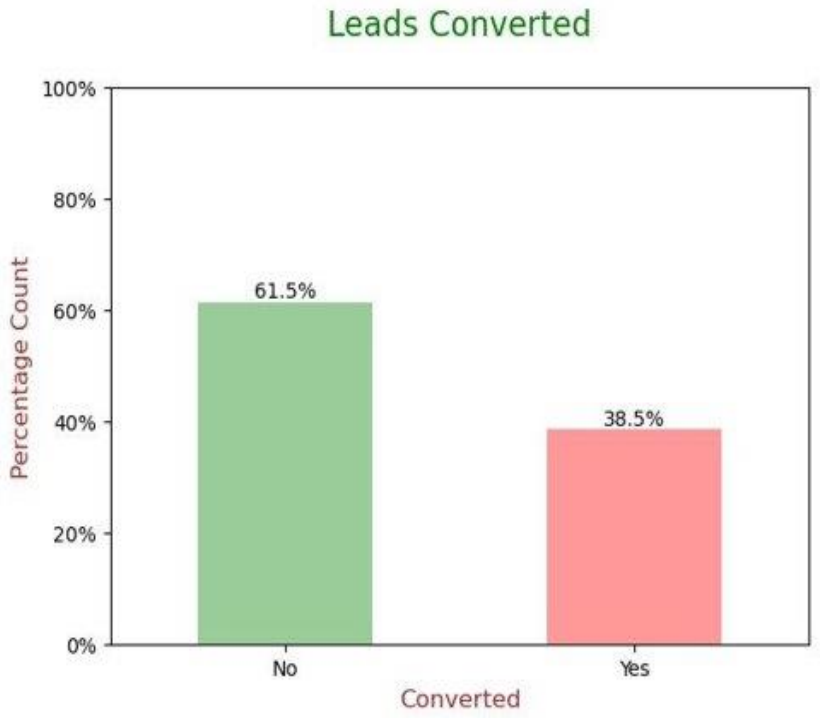
Data Cleaning

- Skewed category columns were identified and removed to prevent skewness bias in logistic regression models.
- Outliers in the 'TotalVisits' and 'Page Views Per Visit' columns were managed by applying a capping technique.
- Invalid values were rectified, and certain columns, like 'Lead Source', underwent data standardization to maintain consistency (e.g., converting 'google' to 'Google').
- Some columns required standardization due to inconsistent casing styles.
- Low-frequency values across certain columns were grouped together under the category "Others" to streamline and generalize the data.
- Additionally, binary categorical variables were encoded or mapped to numeric values as part of the data preprocessing steps.
- Various other cleaning activities were executed to ensure data quality and accuracy, including the handling of invalid values and standardizing data across columns for consistency.

EDA

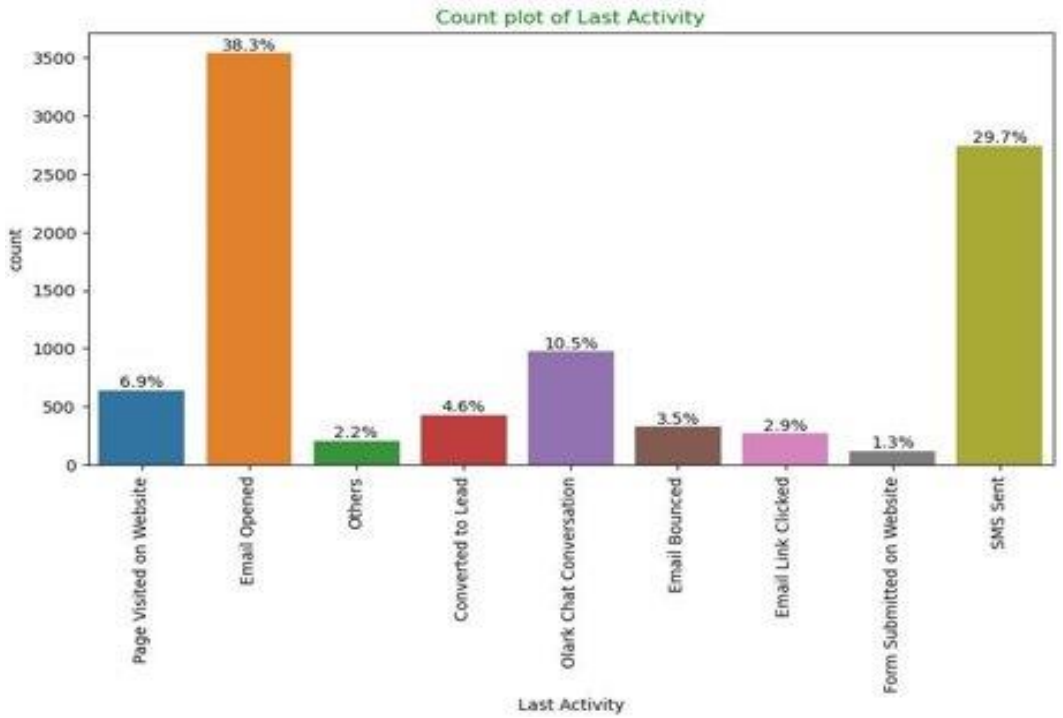
- The analysis reveals an imbalance in the distribution of the target variable.

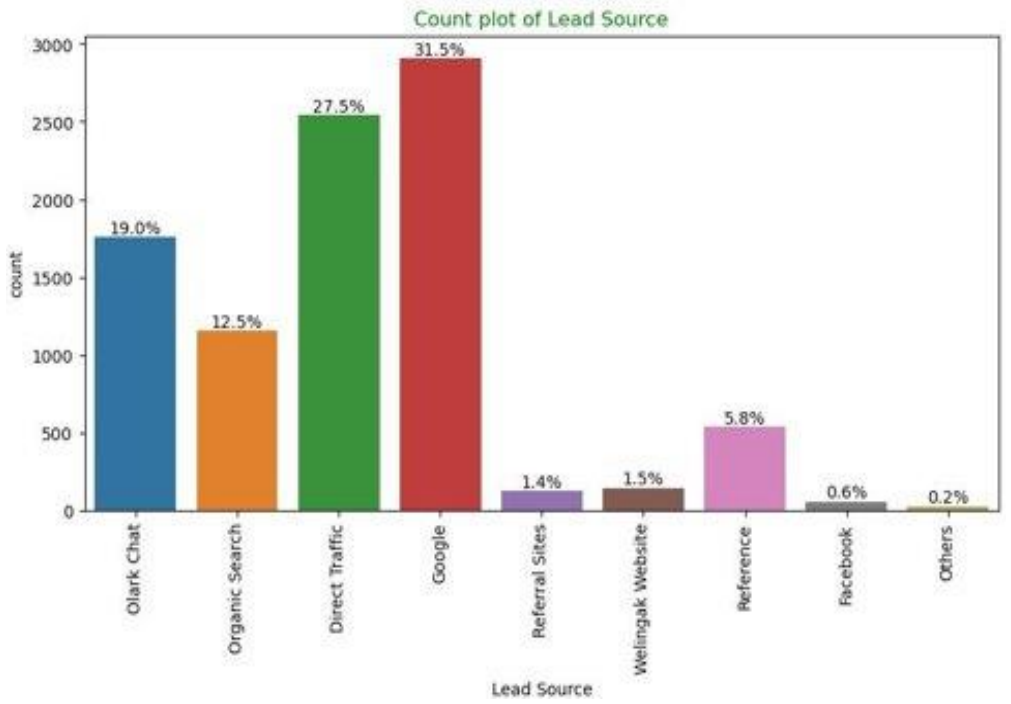
- The conversion rate stands at 38.5%, indicating that only a minority of 38.5% converted to leads. Conversely, the majority, about 61.5%, did not convert to leads.



EDA

- **Univariate Analysis for the Categorical Variables :**



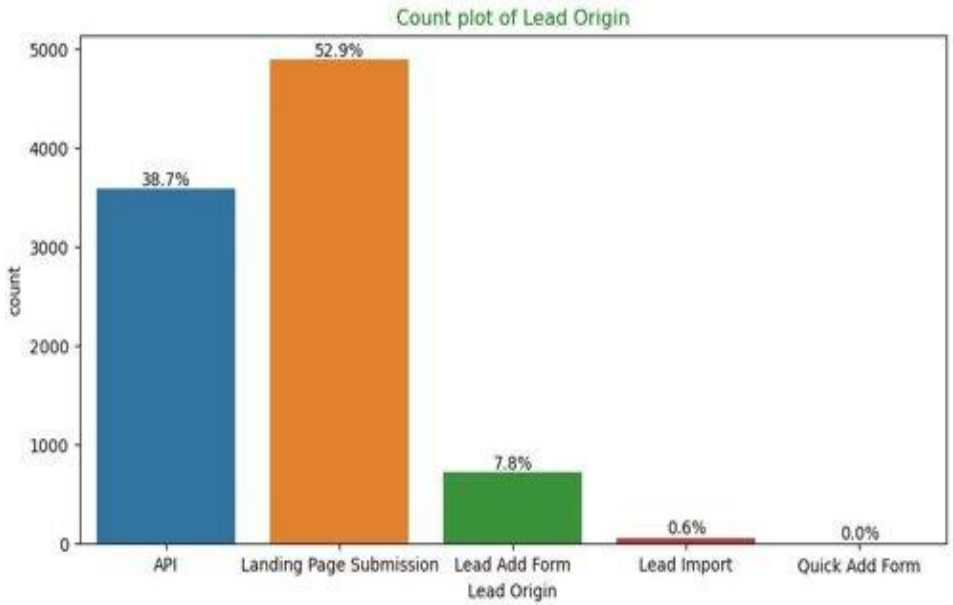


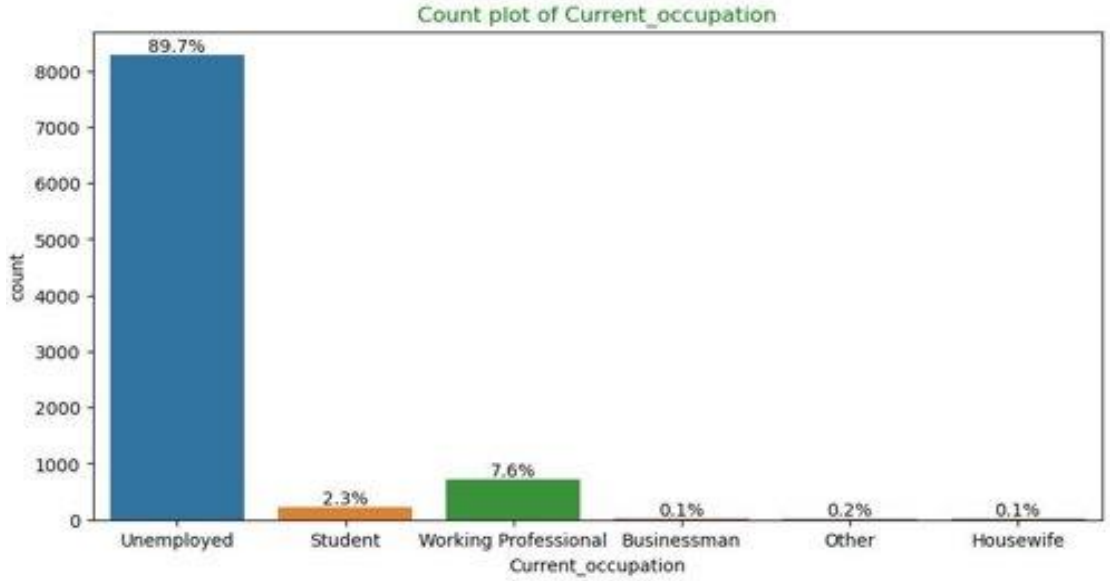
•Around 68% of customer interactions contribute to activities like SMS Sent and Email Opened.

- The majority, which constitutes 58%, of the lead sources come from a combination of Google and Direct Traffic.

EDA

- **Univariate Analysis for the Categorical Variables :**



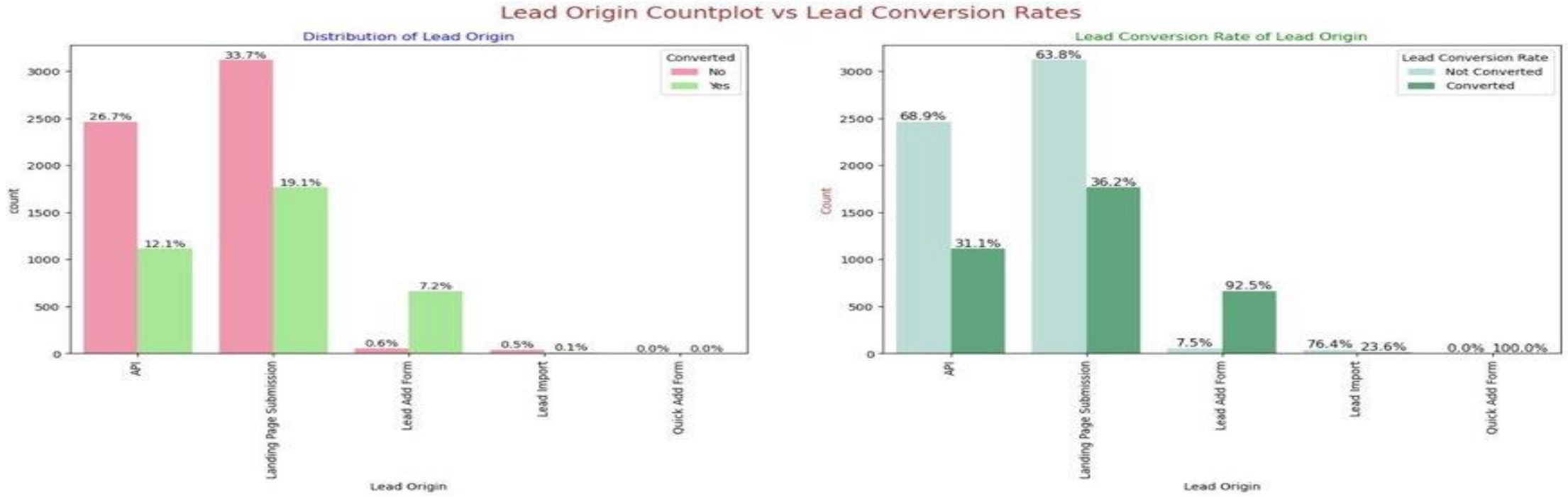


- **"Landing Page Submission"** identifies approximately 53% of the customers, while **"API"** identifies around 39%.

•It appears that around 90% of the customers are classified as **"Unemployed."**

EDA

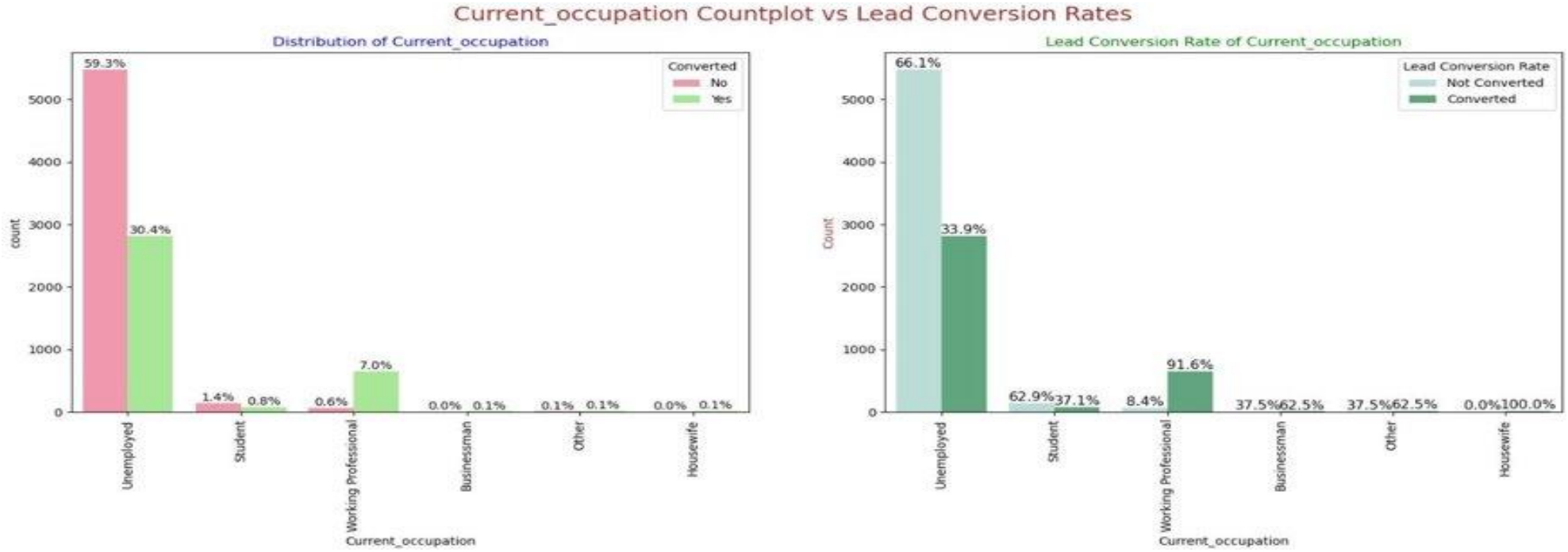
- **Bivariate Analysis for the Categorical Variables :**



It's noted that about 52% of leads originated from **"Landing Page Submission"** with a lead conversion rate (LCR) of 36%. Meanwhile, the **"API"** identified nearly 39% of customers but with a lower lead conversion rate (LCR) of 31%.

EDA

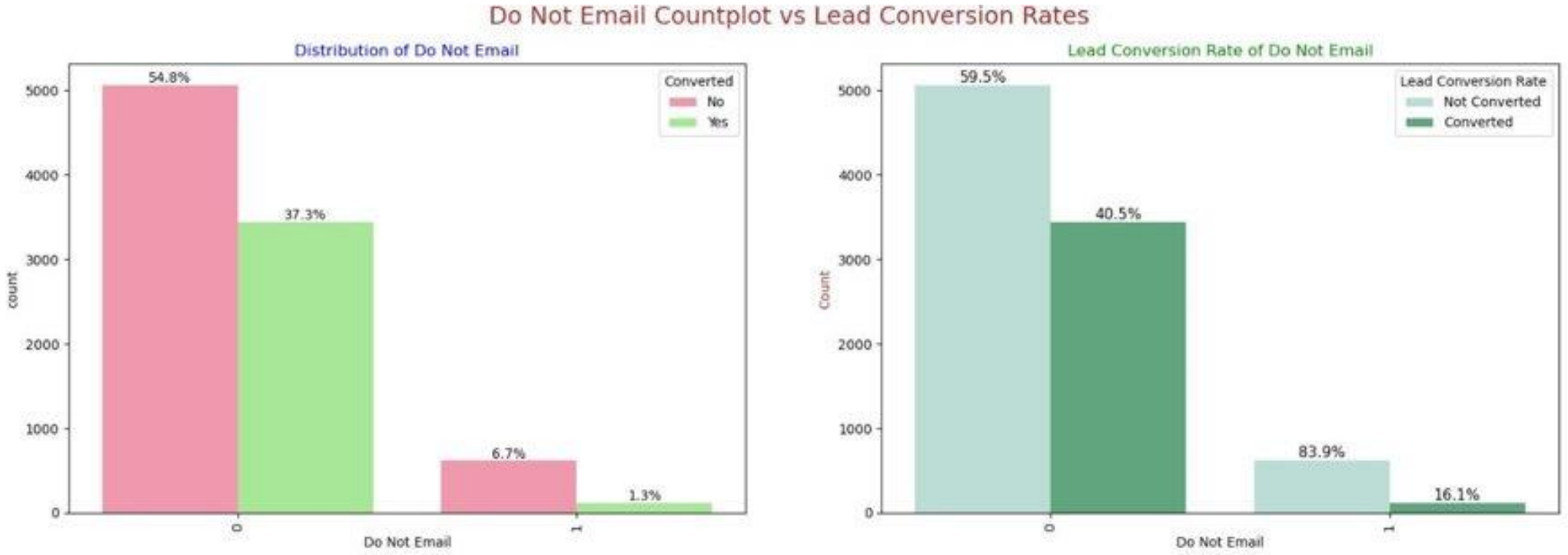
- **Bivariate Analysis for the Categorical Variables :**



Approximately 90% of customers are categorized as **"Unemployed,"** boasting a lead conversion rate (LCR) of 34%. Conversely, **"Working Professionals"** contribute only 7.6% of total customers but exhibit a significantly higher lead conversion rate (LCR) of nearly 92%.

EDA

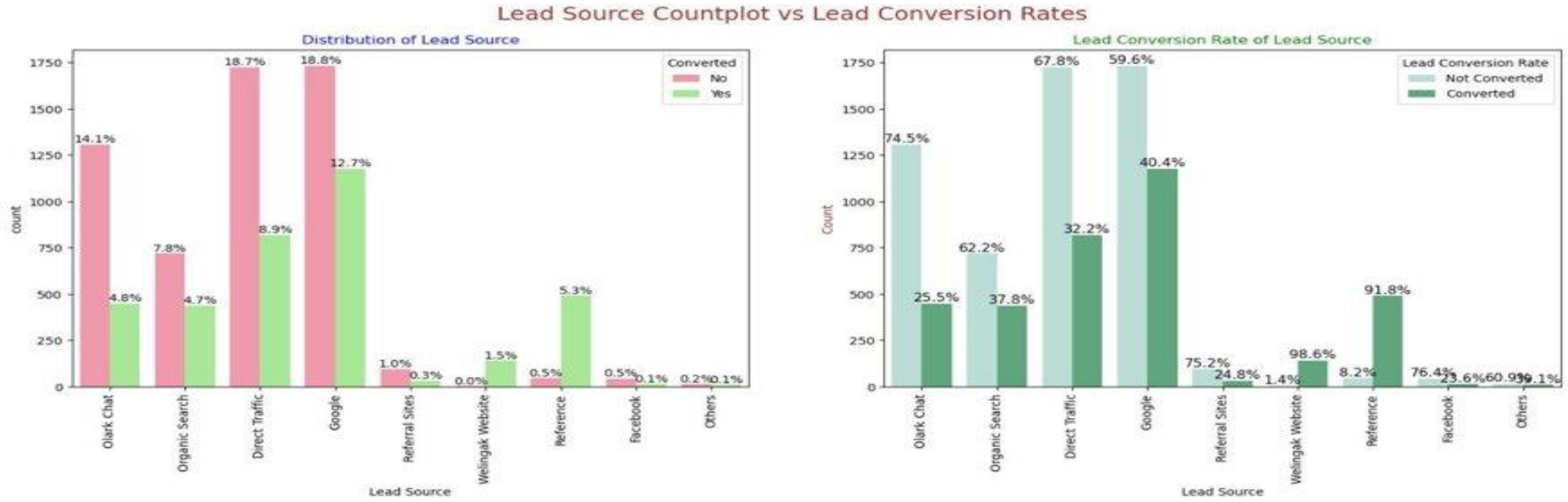
- **Bivariate Analysis for the Categorical Variables :**



"92% of the individuals have chosen not to receive course-related emails. Among them, 40% have converted to leads."

EDA

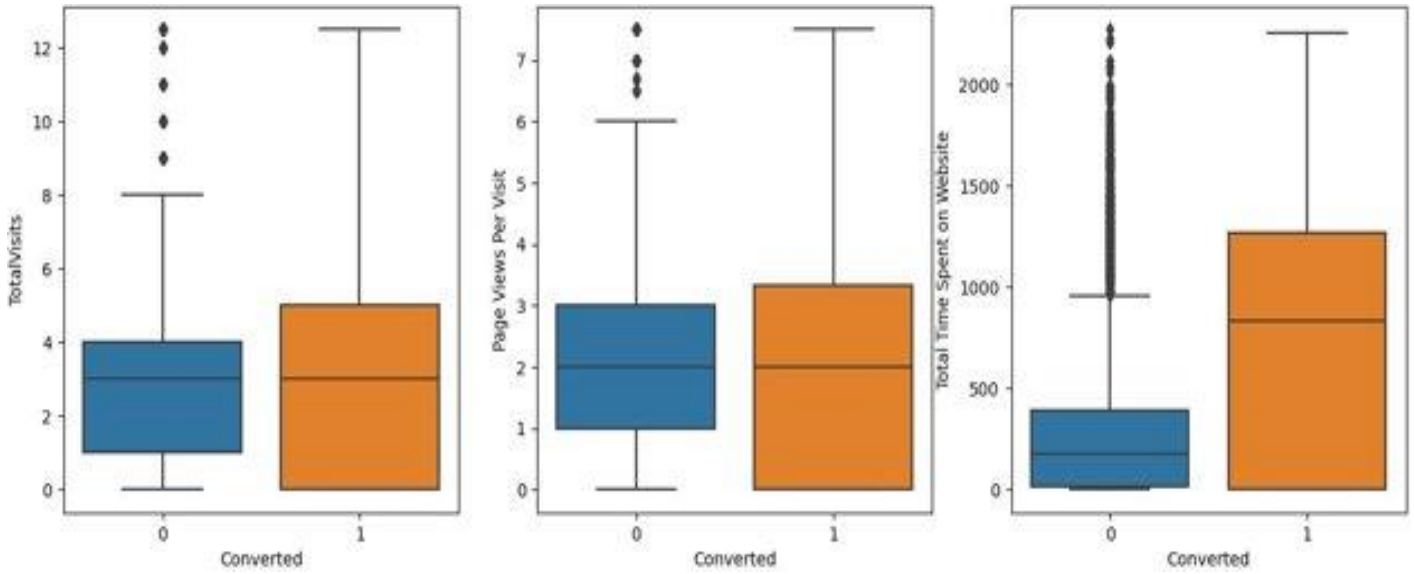
- **Bivariate Analysis for the Categorical Variables :**

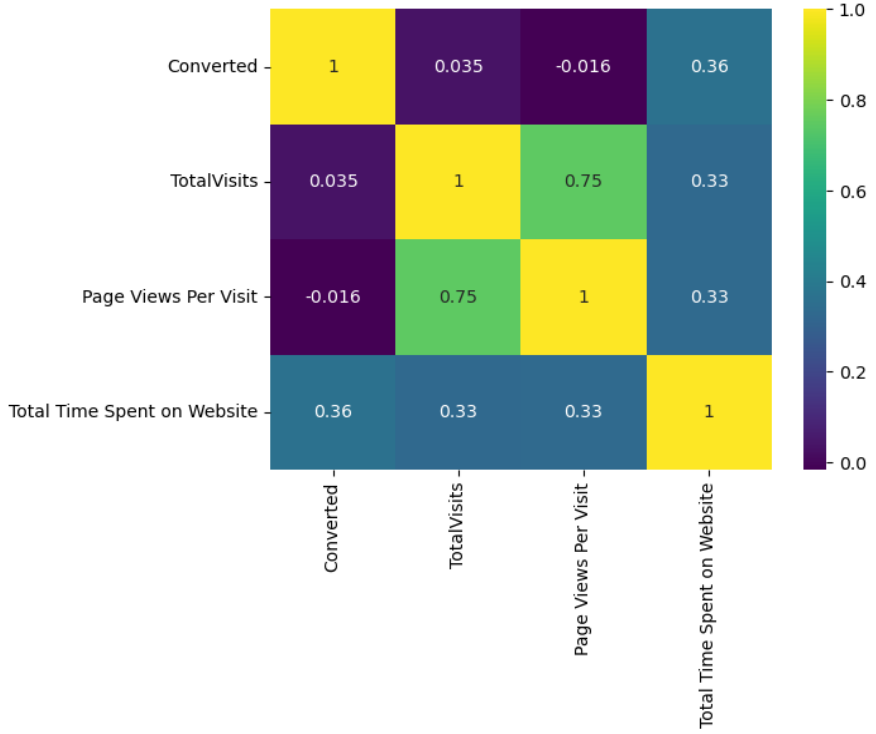


Google has 40% LCR out of 31% and Direct Traffic contribute 32% and seems that the organic search also gives 37.8 % LCR and at last reference has a LCR of 91%.

EDA

- **Bivariate Analysis for the Numerical Variables :**





"Greater time spent on the website by past leads correlates with a higher likelihood of successful conversion, evident from the box plot."

Data Preparation

- Binary categorical columns were encoded as 1s and 0s in earlier stages.
- Dummy variables were generated (one-hot encoding) for categorical features like Lead Origin, Lead Source, Last Activity, Specialization, and Current_occupation.
- The dataset was divided into training and testing sets with a split ratio of 70:30.
- Feature scaling was conducted using standardization methods to normalize the features.
- Correlations among the variables were examined.
- Highly correlated predictor variables, such as Lead Origin_Lead Import and Lead Origin_Lead Add Form, were removed.

Model Building

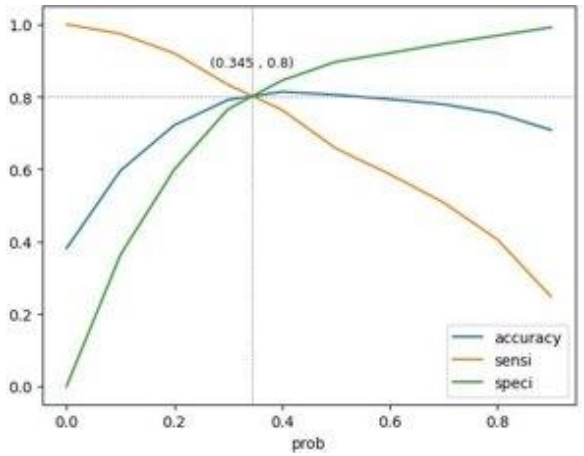
- The dataset's expansive feature set and high dimensionality might hinder model performance and computational efficiency.
- Recursive Feature Elimination (RFE) was employed to extract essential columns, reducing the initial count from 48 to 15 post-RFE.
- This strategic reduction aimed to optimize model efficiency and accuracy, complemented by subsequent manual model fine-tuning.
- Variables with a p-value greater than 0.05 were removed using a manual feature reduction technique.
- After four iterations, Model 4 showed stability with p-values below 0.05 and no multicollinearity issues (VIFs < 5).
- Consequently, logm4 is deemed the final model, selected for Model Evaluation, and further predictions.

Model Evaluation

- **Train –Test Summary**

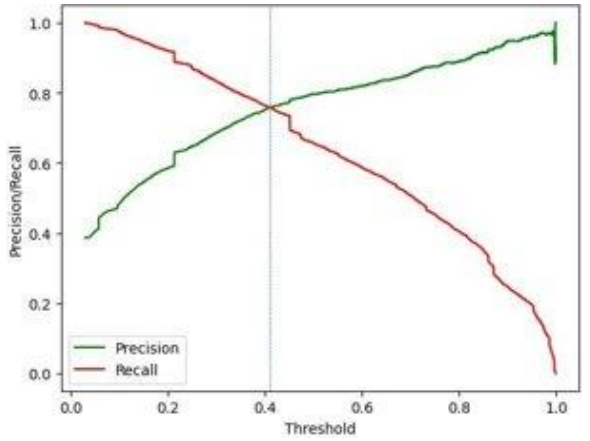
Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

```
**************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

**************************************************

True Negative                  :  3230
True Positive                  :  1974
False Negative                 :  492
False Positve                  :  772
Model Accuracy                 :  0.8046
Model Sensitivity              :  0.8005
Model Specificity              :  0.8071
Model Precision                :  0.7189
Model Recall                   :  0.8005
Model True Positive Rate (TPR) :  0.8005
Model False Positive Rate (FPR) :  0.1929
```



Confusion Matrix & Evaluation Metrics with 0.41 as cutoff

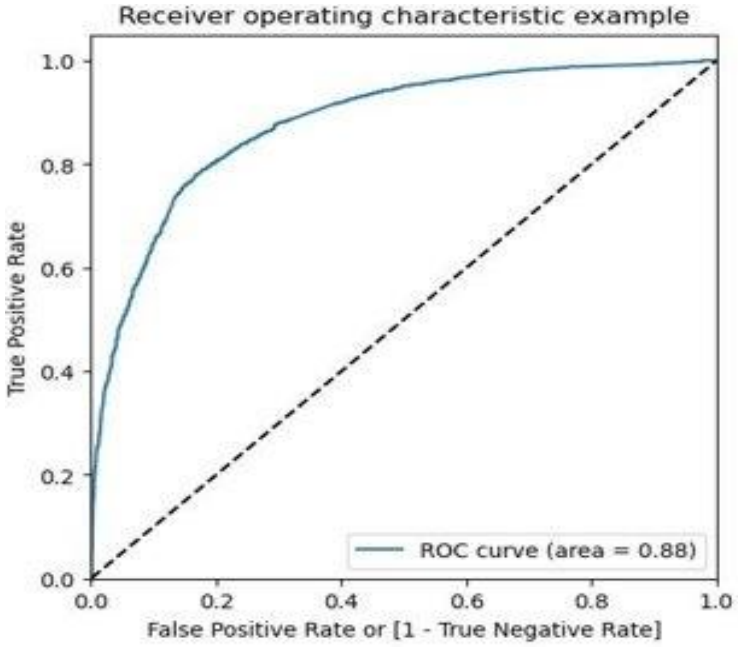
```
**************************************************

Confusion Matrix
[[3406  596]
 [ 596 1870]]

**************************************************

True Negative                  :  3406
True Positive                  :  1870
False Negative                 :  596
False Positve                  :  596
Model Accuracy                 :  0.8157
Model Sensitivity              :  0.7583
Model Specificity              :  0.8511
Model Precision                :  0.7583
Model Recall                   :  0.7583
Model True Positive Rate (TPR) :  0.7583
Model False Positive Rate (FPR) :  0.1489

**************************************************
```



Model Evaluation

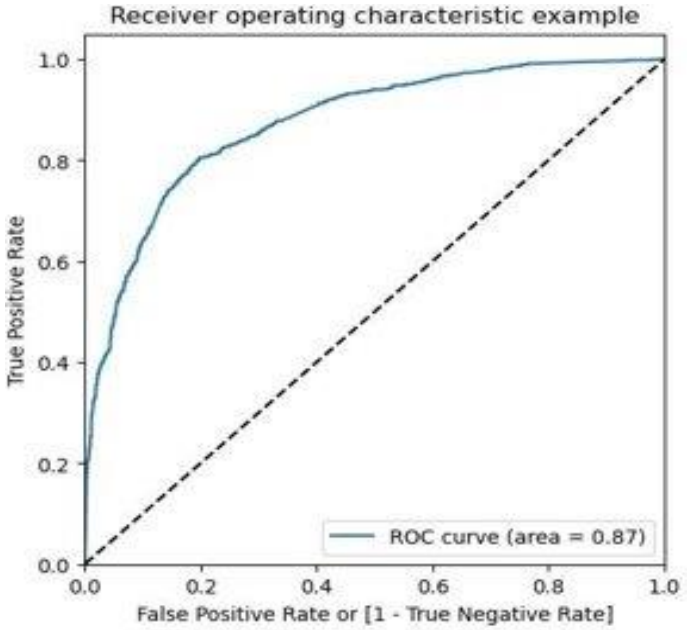
ROC Curve – Train Data Set

- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve – Test Data Set

- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation

Confusion Matrix & Metrics

Train Data Set

```
**********************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

**********************************************************

True Negative                  :  3230
True Positive                  :  1974
False Negative                 :  492
False Positve                  :  772
Model Accuracy                 :  0.8046
Model Sensitivity              :  0.8005
Model Specificity              :  0.8071
Model Precision                :  0.7189
Model Recall                   :  0.8005
Model True Positive Rate (TPR) :  0.8005
Model False Positive Rate (FPR) :  0.1929
```

Test Data Set

```
**********************************************************

Confusion Matrix
[[1353  324]
 [ 221  874]]

**********************************************************

True Negative                  :  1353
True Positive                  :  874
False Negative                 :  221
False Positve                  :  324
Model Accuracy                 :  0.8034
Model Sensitivity              :  0.7982
Model Specificity              :  0.8068
Model Precision                :  0.7295
Model Recall                   :  0.7982
Model True Positive Rate (TPR) :  0.7982
Model False Positive Rate (FPR) :  0.1932
```

- At a cut-off value of 0.345, the model exhibited a sensitivity of 80.05% for the train set and 79.82% for the test set.
- Sensitivity, in this context, signifies the model's capacity to accurately identify converting leads out of the total potential leads.
- The CEO of X Education had set an approximate sensitivity target of 80%.
- Additionally, the model achieved an accuracy rate of 80.46%, aligning with the objectives outlined in the study.

Recommendations

To enhance our Lead Conversion Rates, here are some suggestions:

- Leverage features with positive coefficients for specific marketing approaches to attract potential leads effectively.
- Devote efforts to draw high-quality leads from the most successful lead sources identified.
- Engage working professionals using tailored messages that resonate with their needs and preferences.
- Optimize communication channels based on their impact on lead engagement for better outreach.
- Consider increased budget allocation for Welingak Website in advertising and related activities.
- Encourage and incentivize references that convert into leads, possibly through discounts or incentives.
- Aggressively target working professionals given their high conversion rates and potentially stronger financial capabilities for higher fees.
- **To identify areas of improvement:**
- Examine specialization offerings with negative coefficients.
- Evaluate the landing page submission process to identify potential areas for enhancement.

THANK YOU