X Education – Lead Scoring Case Study

Team Members : Kritangi Srivastav, Kuber Marwah, Kumar Shanu

Table of Contents

Background of X Education Company

- X Education, an educational company, offers online courses to industry professionals.
- Every day, numerous professionals interested in these courses visit their website to browse the offerings.
- The company promotes its courses on various websites and search engines like Google.
- Upon landing on the website, visitors might browse the courses, fill out a form for more information, or watch videos.
- When visitors fill out a form with their email address or phone number, they are classified as leads.
- The sales team then contacts these leads through calls, emails, and other communication methods.
- Although this process converts some leads, the majority do not convert.
- The typical lead conversion rate at X Education is about 30%.

Problem Statement & Objective of the Study

Problem Statement:

- X Education receives a large number of leads, but its lead conversion rate is quite low at around 30%.
- The company aims to improve the efficiency of its lead conversion process by identifying the most promising leads, referred to as Hot Leads.
- The sales team wants to focus their communication efforts on these potential leads, rather than contacting everyone.

Objective of the Study:

- The goal is to assist X Education in selecting the most promising leads, specifically those most likely to convert into paying customers.
- We need to develop a model that assigns a lead score to each lead, ensuring that leads with higher scores have a greater chance of conversion while those with lower scores have a lesser chance.
- The CEO has set a target lead conversion rate of approximately 80%.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are categorized based on their likelihood to convert.
- This results in a concentrated group of hot leads.

Better Communication

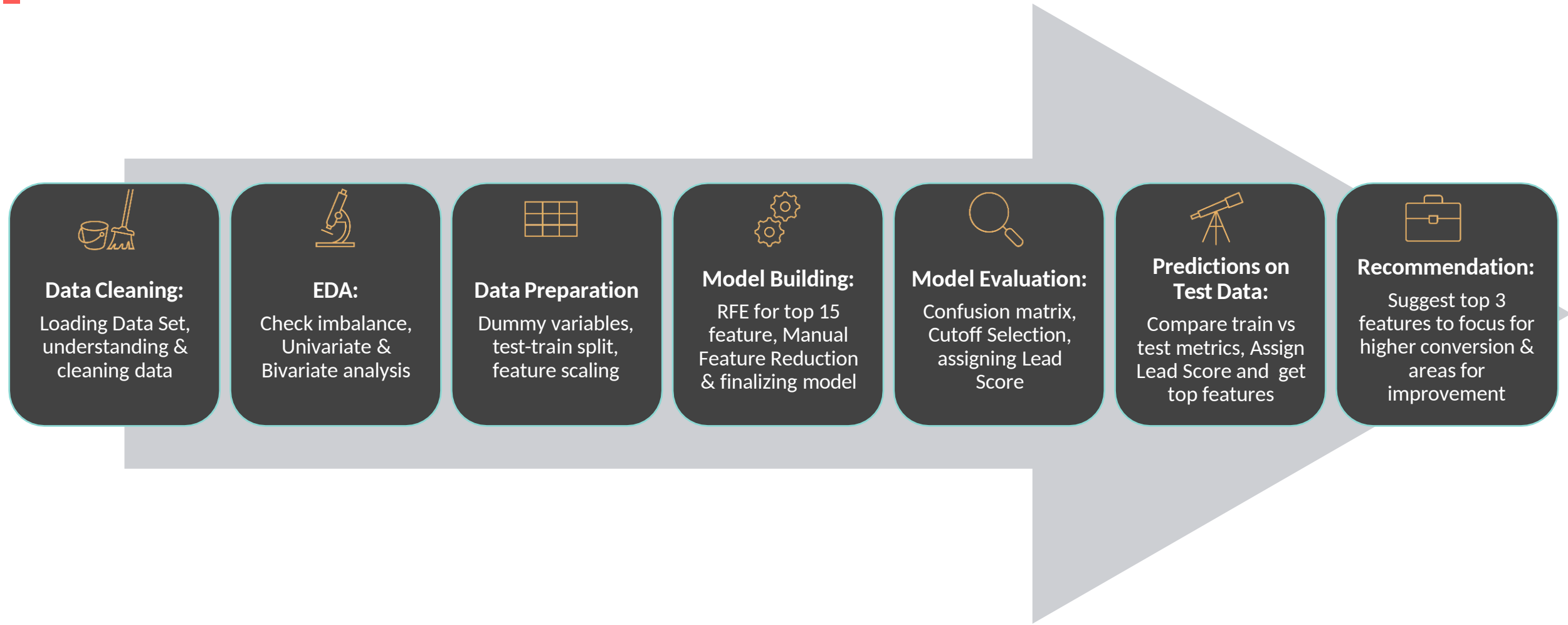
- By focusing on a smaller pool of leads, we can enhance our impact.

Boost Conversion

- Concentrating on hot leads that are more likely to convert would increase our conversion rate and help us achieve the 80% target.

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

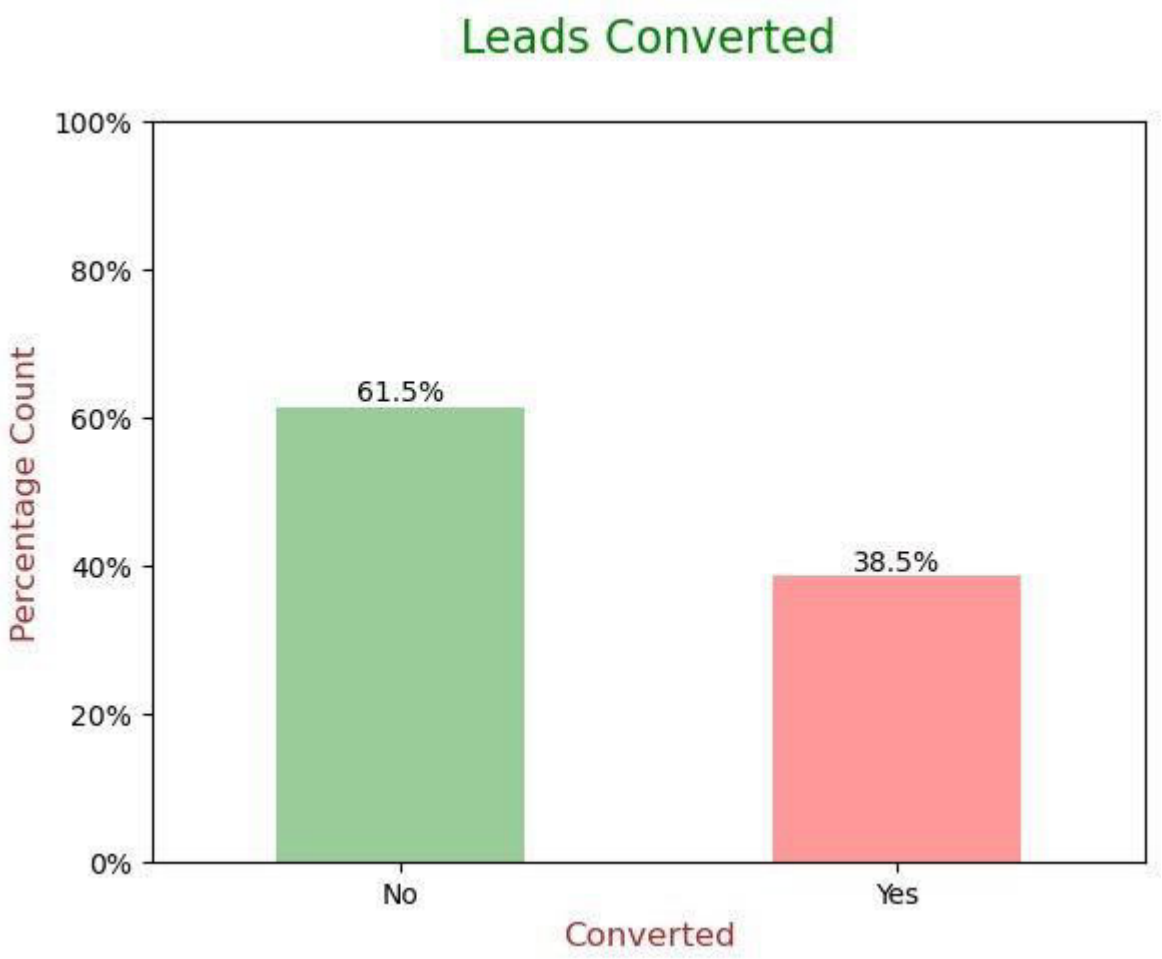
- The "Select" level indicates null values for certain categorical variables, where customers did not select an option from the list.
- Columns with over 40% null values were removed.
- Missing values in categorical columns were addressed based on value counts and specific considerations.
- Columns that do not provide any insight or value to the study objective (e.g., tags, country) were dropped.
- Imputation was applied to some categorical variables.
- Additional categories were created for certain variables.
- Columns not useful for modeling (e.g., Prospect ID, Lead Number) or those with only one category of response were removed.
- Numerical data was imputed using the mode after checking the distribution.

Data Cleaning

- Category columns with skewed distributions were identified and removed to prevent bias in logistic regression models.
- Outliers in TotalVisits and Page Views Per Visit were addressed by capping them.
- Low-frequency values in categorical variables were consolidated under "Others" to streamline analysis.
- Binary categorical variables were encoded appropriately.
- Additional data cleaning steps were implemented to enhance overall data quality and accuracy, such as fixing invalid values and standardizing data formats (e.g., ensuring consistency in casing styles for lead source, like correcting 'Google' to 'google').

EDA

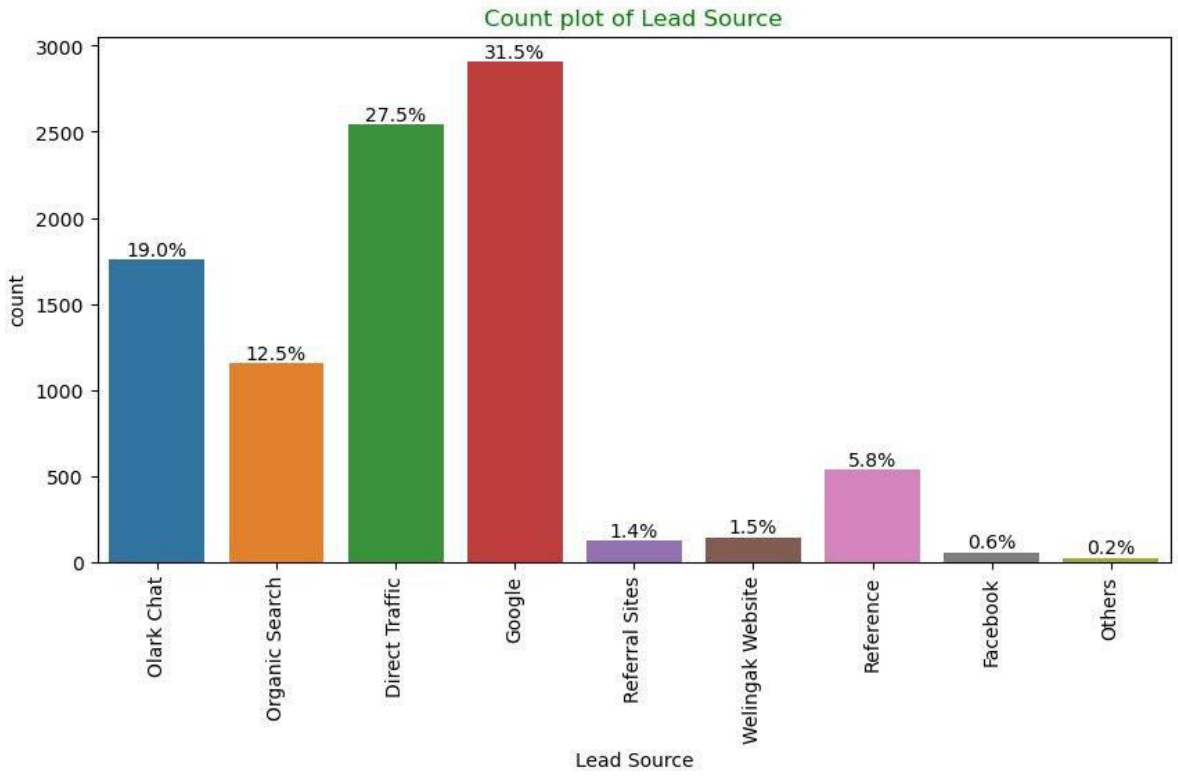
- Data is imbalanced while analyzing target variable.

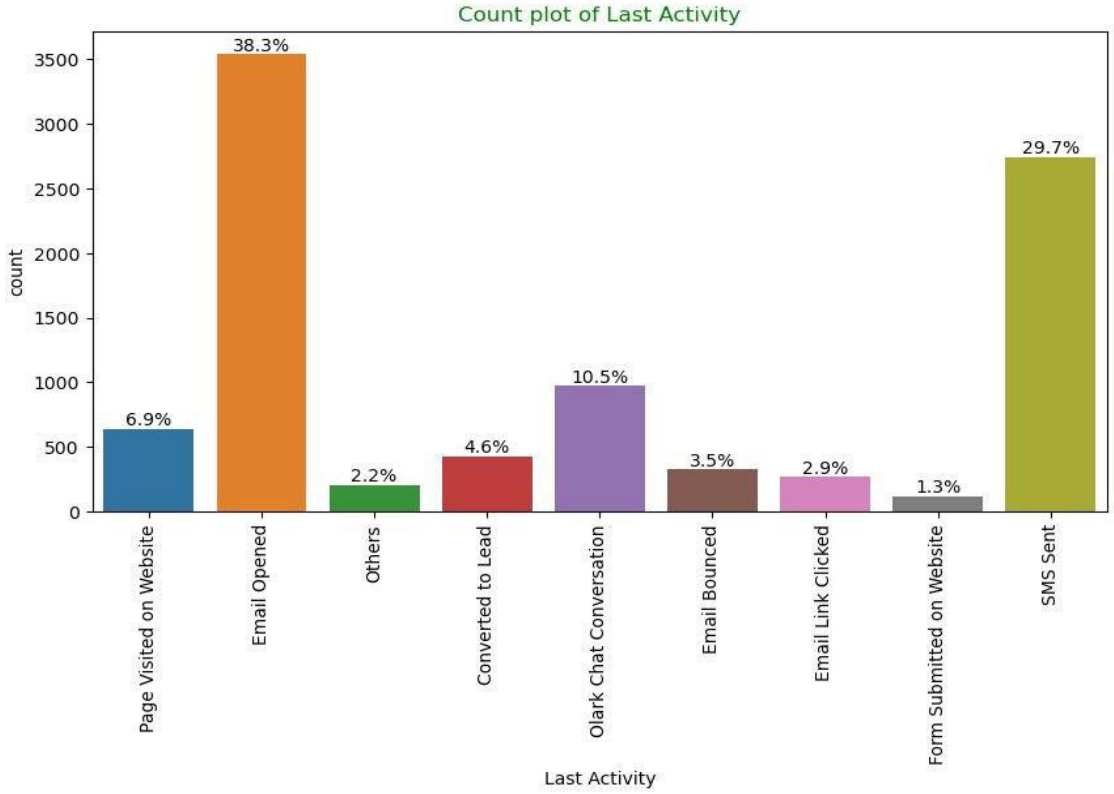


- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

EDA

- Univariate Analysis – Categorical Variables

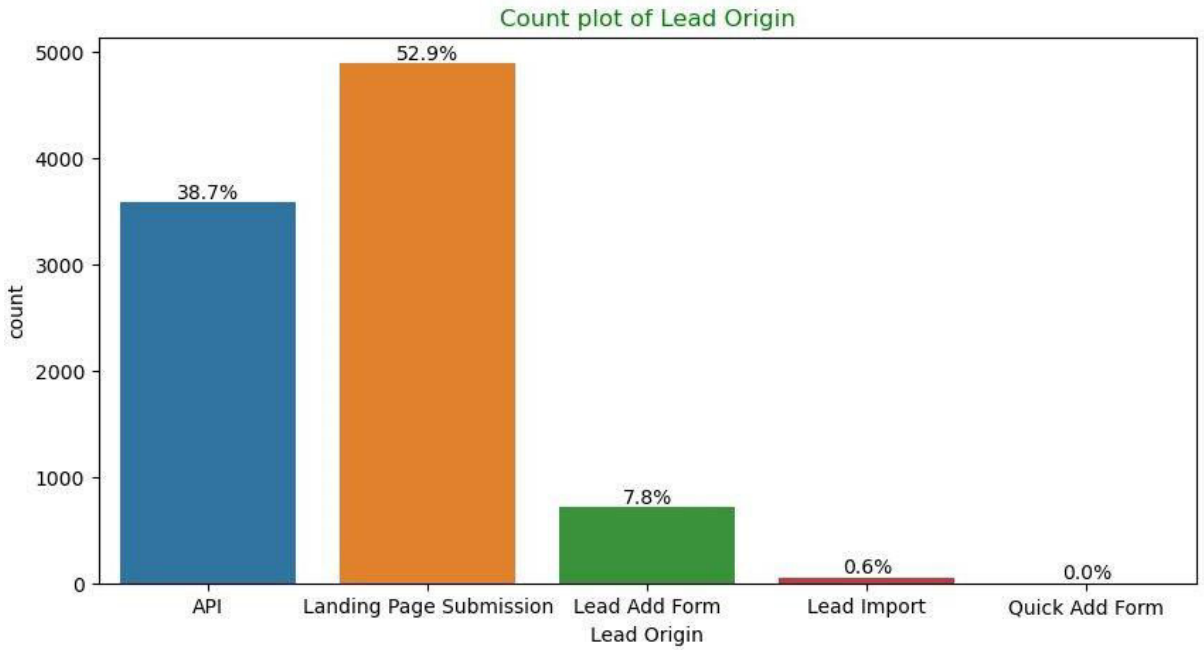


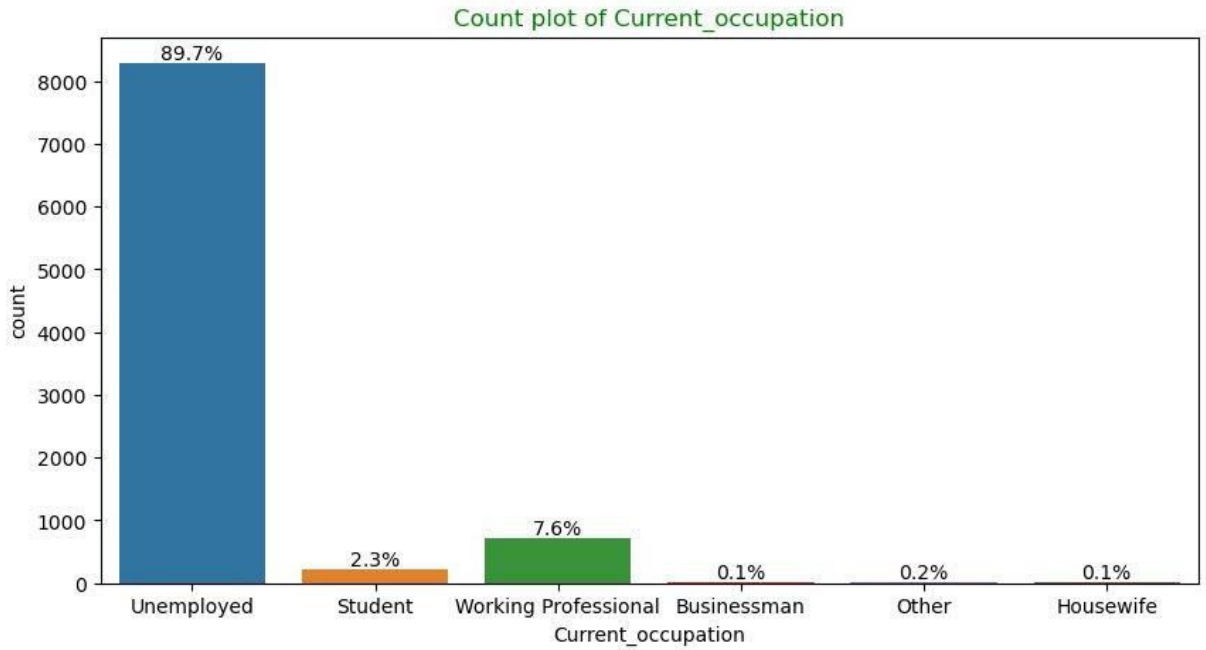


- **Lead Source:** 59% Lead source is from Google & Direct Traffic combined.
- **Last Activity:** 68% of customers contribution in SMS Sent & Email Opened activities.

EDA

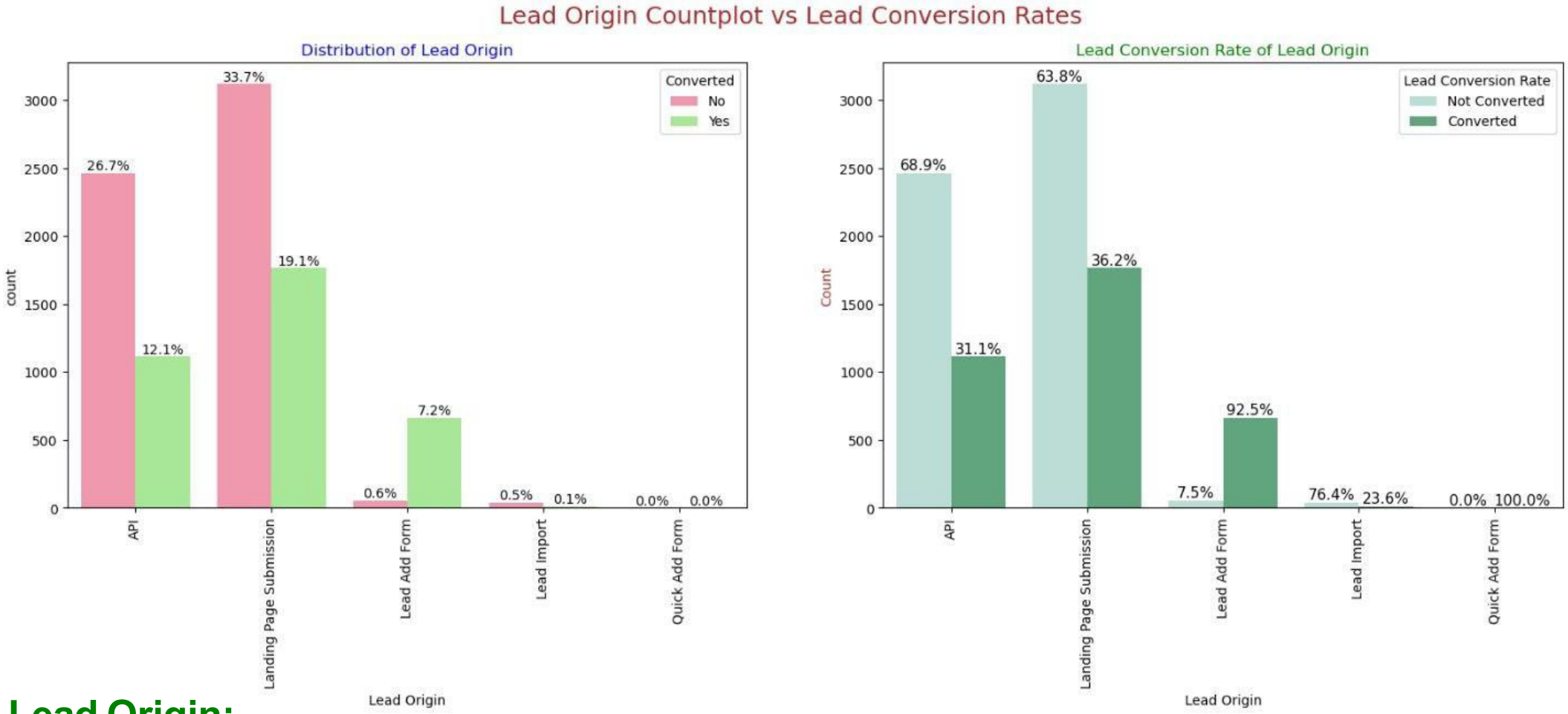
- **Univariate Analysis – Categorical Variables**





- **Lead Origin:** "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- **Current_occupation:** It has 90% of the customers as Unemployed.

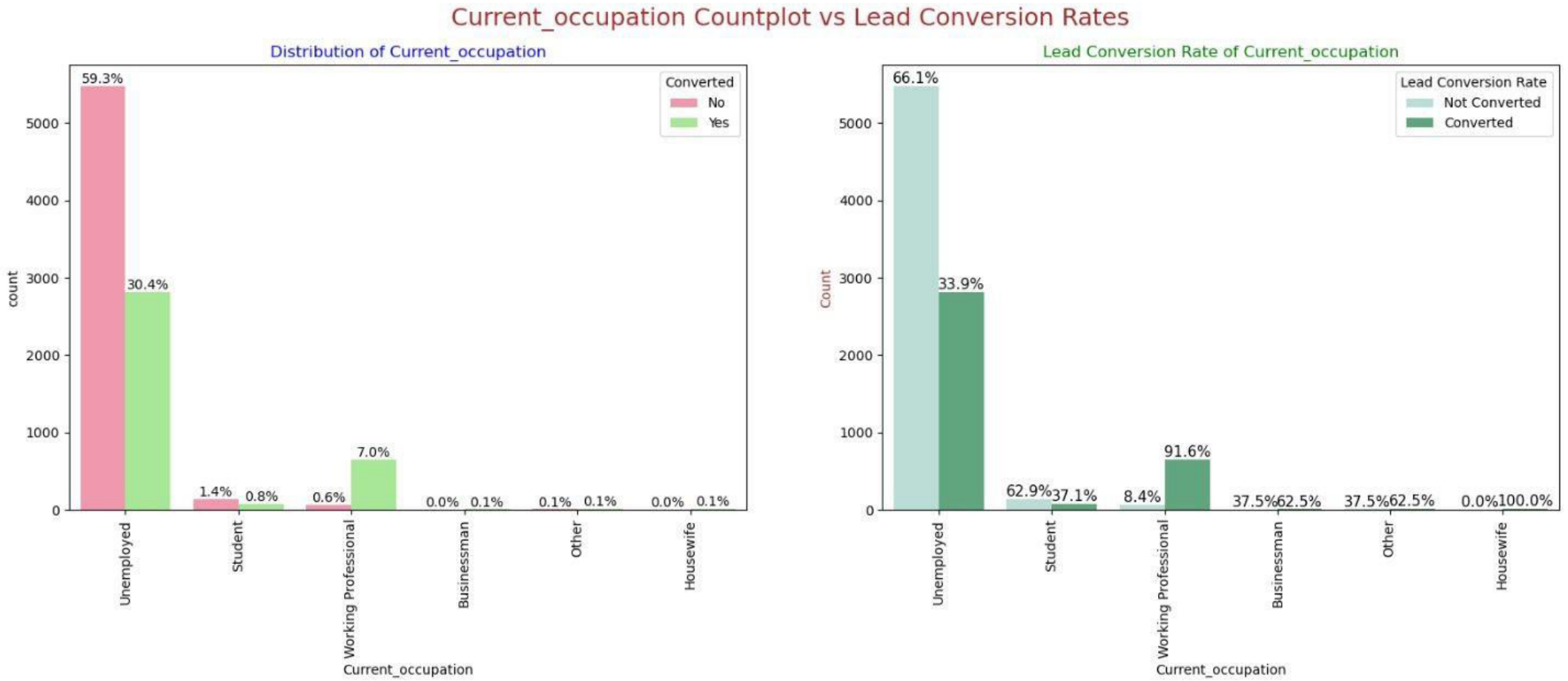
EDA – Bivariate Analysis for Categorical Variables



Lead Origin:

- Around 52% of all leads originated from *"Landing Page Submission"* with a **lead conversion rate (LCR) of 36%**.
- The *"API"* identified approximately 39% of customers with a **lead conversion rate (LCR) of 31%**.

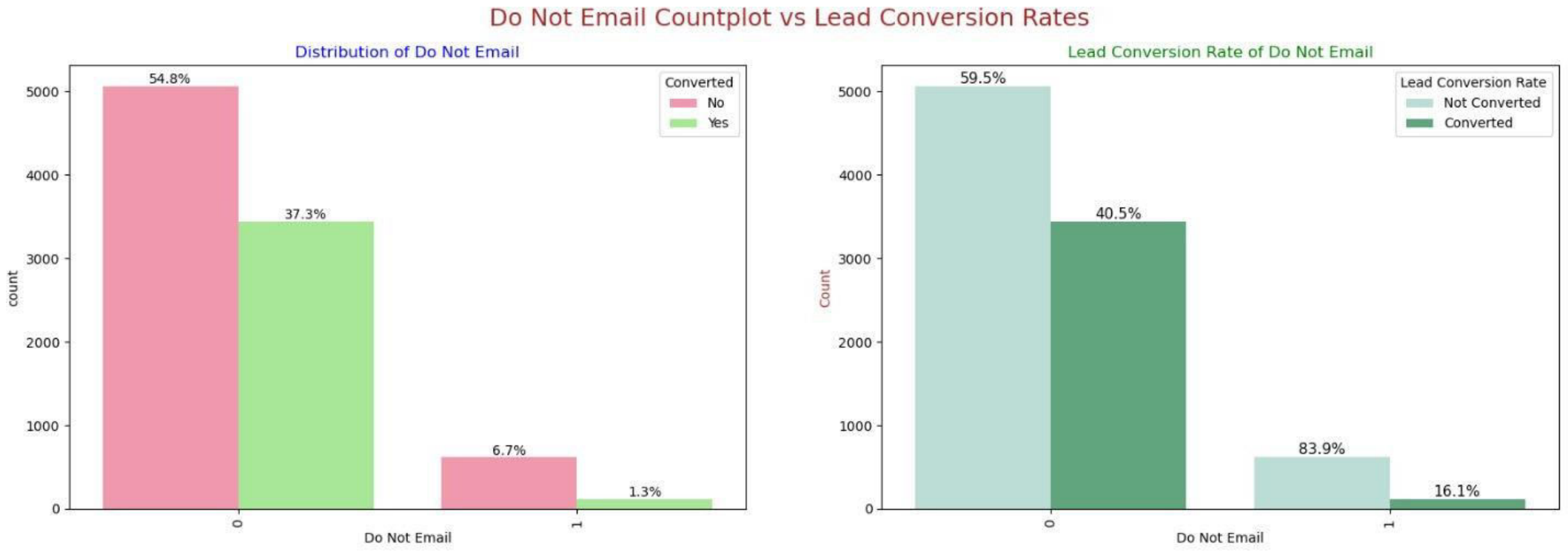
EDA – Bivariate Analysis for Categorical Variables



Current_occupation:

- Around 90% of the customers are *Unemployed,* with **lead conversion rate (LCR) of 34%**.
- While *Working Professional* contribute only 7.6% of total customers with almost **92% Lead conversion rate (LCR)**.

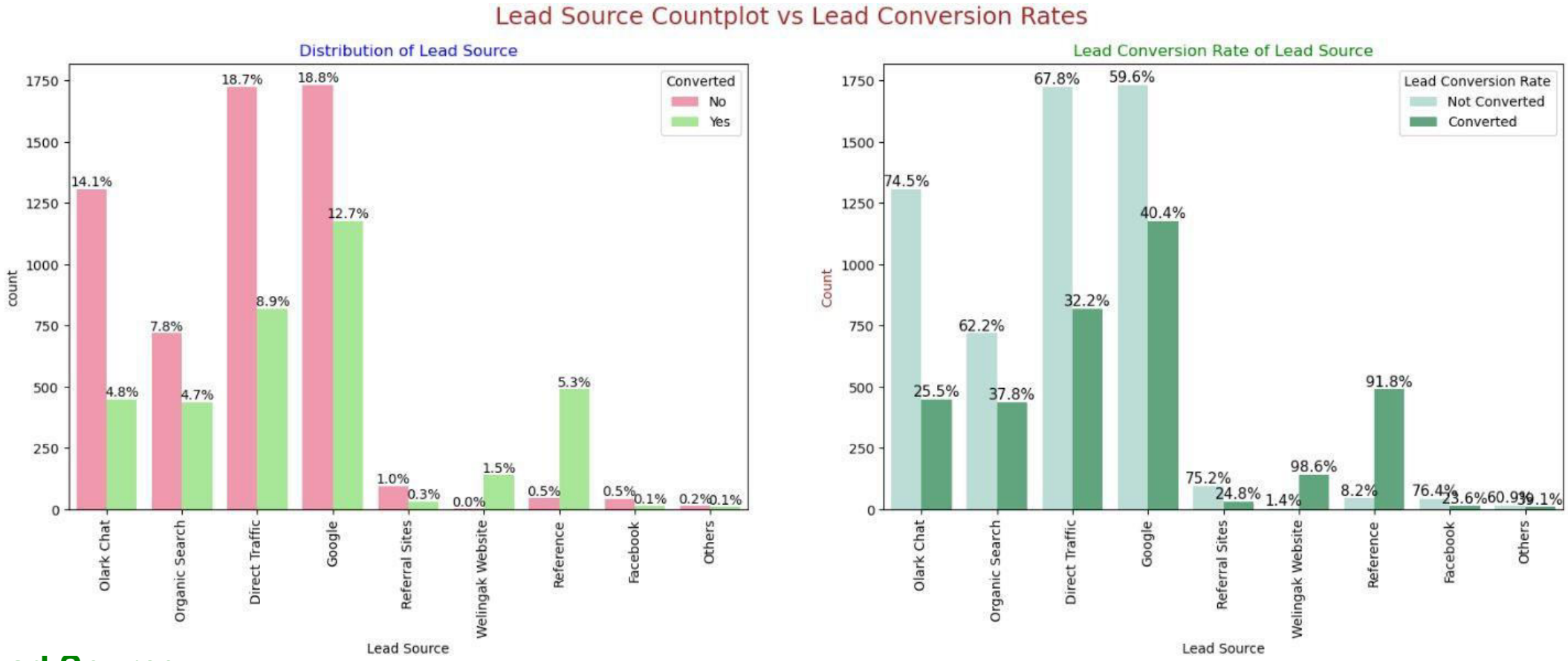
EDA - Bivariate Analysis for Categorical Variables



Do Not Email:

- 92% of the people has opted that they don't want to be emailed about the course & 40% of them are converted to leads.

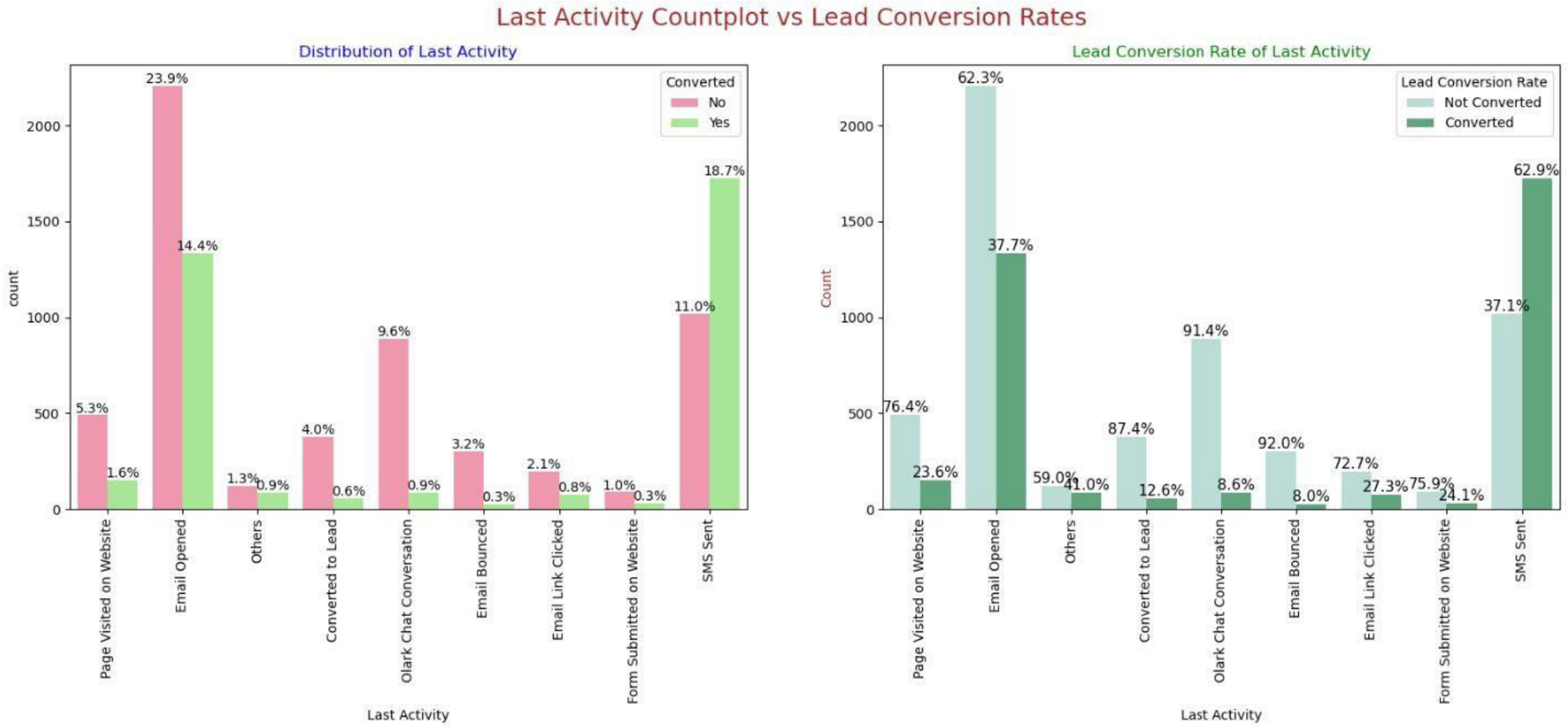
EDA - Bivariate Analysis for Categorical Variables



Lead Source:

- *Google* has **LCR of 40%** out of 31% customers,
- *Direct Traffic* contributes **32% LCR** with 27% customers, which is lower than Google,
- *Organic Search* also gives **37.8% of LCR**, but the contribution is by only 12.5% of customers,
- *Reference* has **LCR of 91%**, but there are only around 6% of customers through this Lead Source.

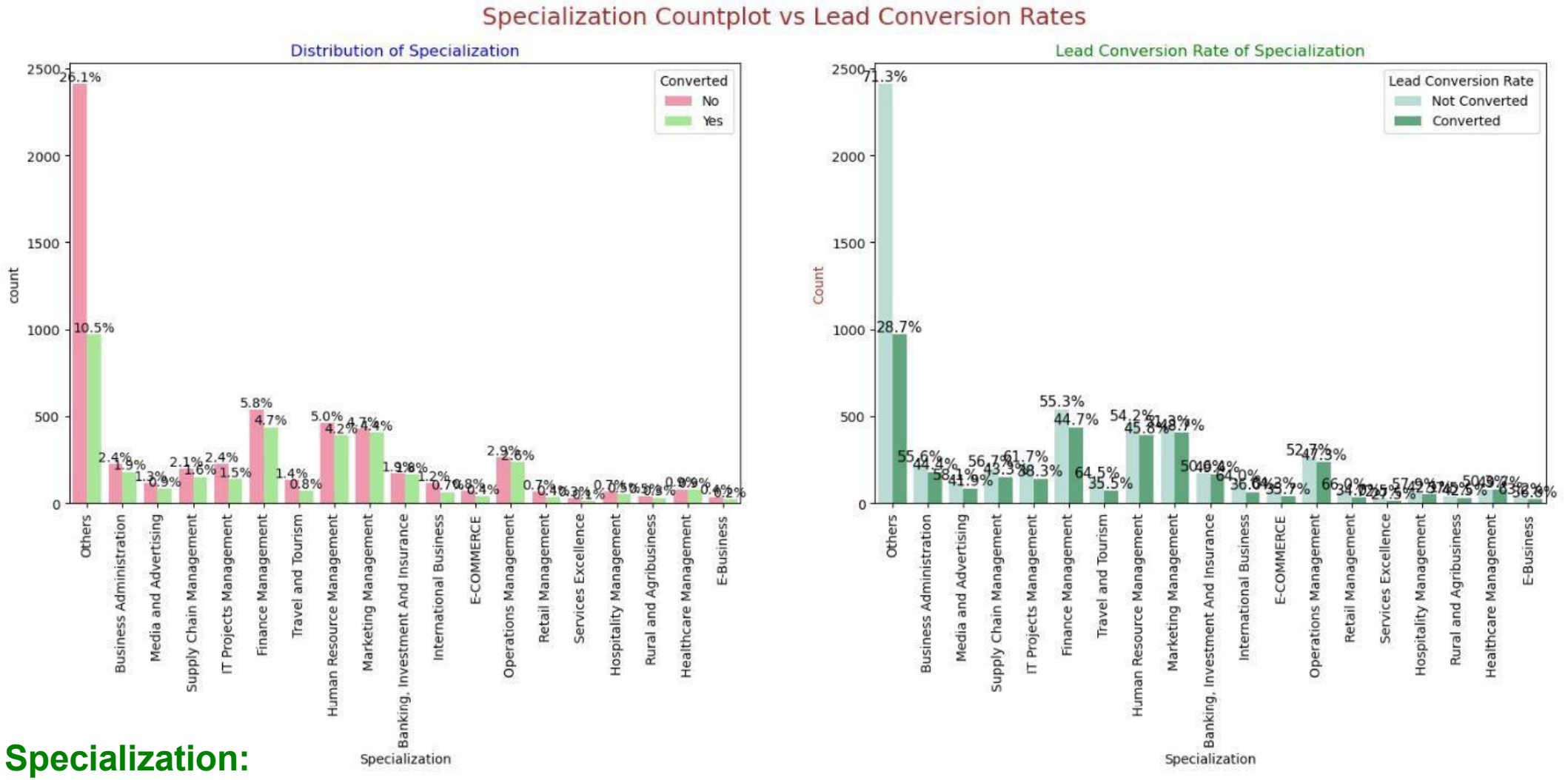
EDA – Bivariate Analysis for Categorical Variables



Last Activity:

- *'SMS Sent'* has **high lead conversion rate of 63%** with 30% contribution from last activities,
- *'Email Opened'* activity contributed 38% of last activities performed by the customers, with **37% lead conversion rate.**

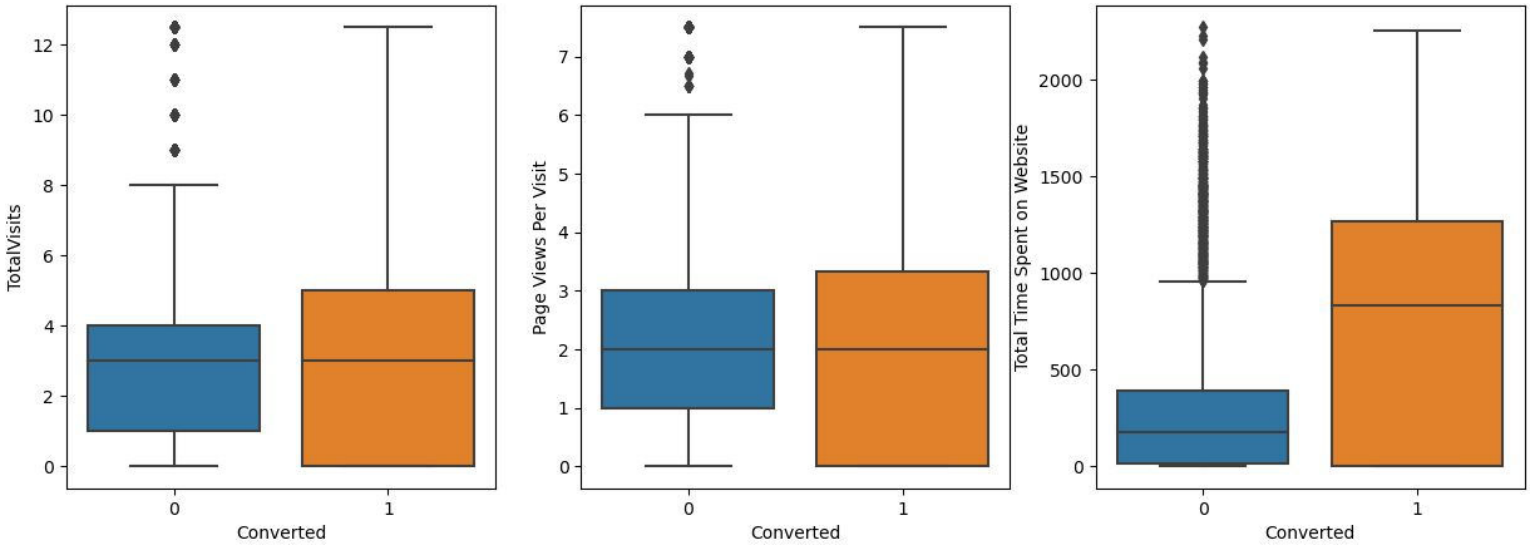
EDA - Bivariate Analysis for Categorical Variables

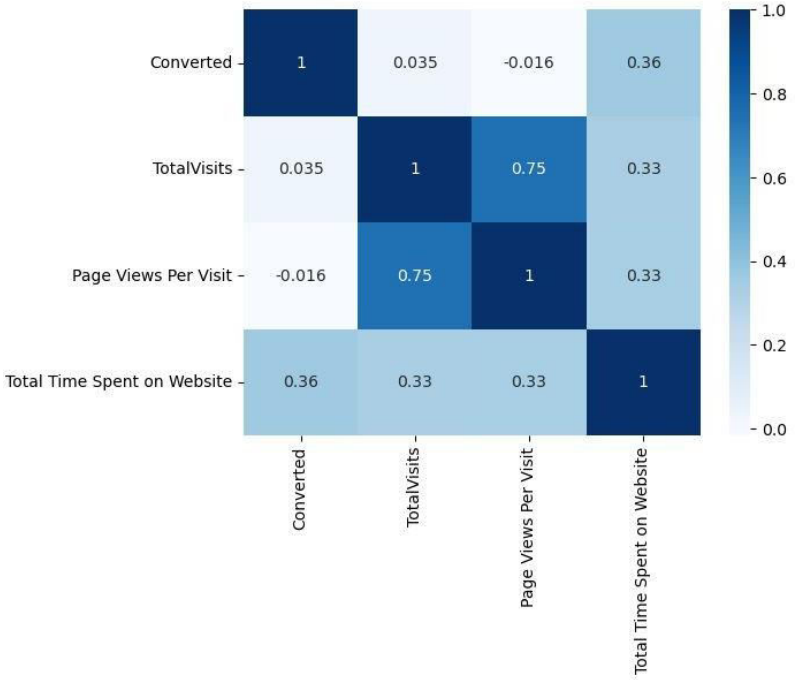


Specialization:

- Marketing Management, HR Management, Finance Management shows good contribution in Leads conversion than other specialization.

EDA – Bivariate Analysis for Numerical Variables





- Past Leads who **spends more time on the Website** have a higher chance of getting successfully converted than those who spends less time as seen in the **box-plot**

Data Preparation before Model building

- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation
- Splitting Train & Test Sets
  - 70:30 % ratio was chosen for the split
- Feature scaling
  - Standardization method was used to scale the features
- Checking the correlations
  - Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building

Feature Selection

- The dataset is characterized by numerous dimensions and a large number of features.
- This could potentially degrade model performance and increase computational overhead.
- Therefore, it is crucial to conduct Recursive Feature Elimination (RFE) to identify and select only the most relevant columns.
- Subsequently, manual fine-tuning of the model can be performed to optimize its performance.
- RFE outcome
  - Pre RFE – 48 columns & Post RFE – 15 columns

Model Building

- Variables with p-values greater than 0.05 were dropped using a manual feature reduction process to build models.

- After four iterations, Model 4 appears stable with:
  - Significant p-values below the threshold (p-values < 0.05), and
  - No evidence of multicollinearity, as indicated by VIFs less than 5.

  Therefore, **logm4** is selected as the final model, which will undergo Model Evaluation and subsequently be used for making predictions.

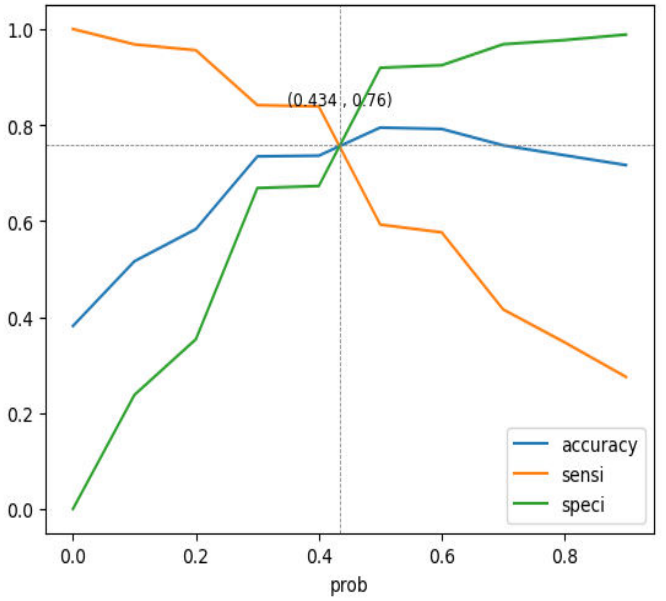
Model Evaluation

Train Data Set

It was decided to go ahead with 0.434 as cutoff after checking evaluation metrics coming from both plots

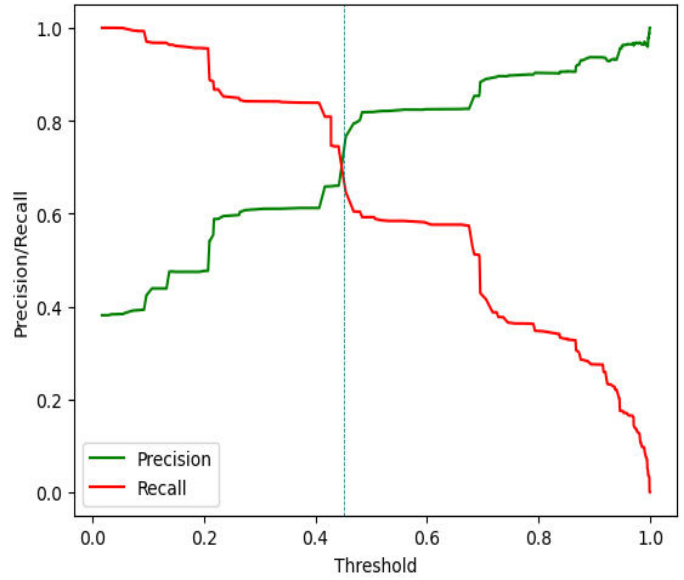
Confusion Matrix & Evaluation Metrics with 0.434 as cutoff

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

Confusion Matrix
[[2691 1311]
 [ 397 2069]]

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

True Negative                  :  2691
True Positive                  :  2069
False Negative                 :  397
False Positve                  :  1311
Model Accuracy                 :  0.7359
Model Sensitivity              :  0.839
Model Specificity              :  0.6724
Model Precision                :  0.6121
Model Recall                   :  0.839
Model True Positive Rate (TPR) :  0.839
Model False Positive Rate (FPR) :  0.3276

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



Confusion Matrix & Evaluation Metrics with 0.44 as cutoff

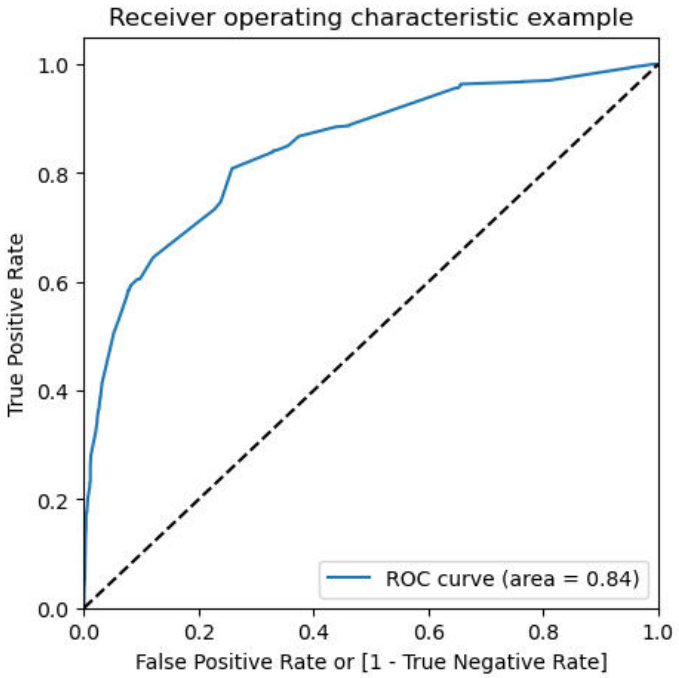
```
************************************************

Confusion Matrix
[[3056  946]
 [ 629 1837]]

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

True Negative                  :  3056
True Positive                  :  1837
False Negative                 :  629
False Positve                  :  946
Model Accuracy                 :  0.7565
Model Sensitivity              :  0.7449
Model Specificity              :  0.7636
Model Precision                :  0.6601
Model Recall                   :  0.7449
Model True Positive Rate (TPR) :  0.7449
Model False Positive Rate (FPR) :  0.2364

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



Model Evaluation

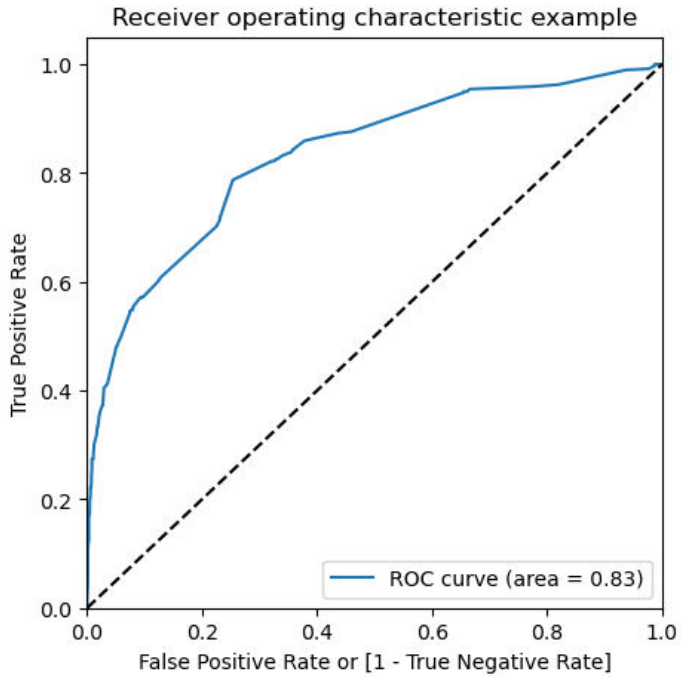
ROC Curve – Train Data Set

- Area under ROC curve is 0.84 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 0.83 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
[[2691 1311]
 [ 397 2069]]

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

True Negative                  :  2691
True Positive                  :  2069
False Negative                 :  397
False Positve                  :  1311
Model Accuracy                 :  0.7359
Model Sensitivity              :  0.839
Model Specificity              :  0.6724
Model Precision                :  0.6121
Model Recall                   :  0.839
Model True Positive Rate (TPR) :  0.839
Model False Positive Rate (FPR) :  0.3276
```

Test Data Set

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

Confusion Matrix
[[1142  535]
 [ 196  899]]

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

True Negative                  :  1142
True Positive                  :  899
False Negative                 :  196
False Positve                  :  535
Model Accuracy                 :  0.7363
Model Sensitivity              :  0.821
Model Specificity              :  0.681
Model Precision                :  0.6269
Model Recall                   :  0.821
Model True Positive Rate (TPR) :  0.821
Model False Positive Rate (FPR) :  0.319
```

- Using a cut-off value of 0.434, the model achieved a **sensitivity** of **83.9% in the train set** and **82.1% in the test set**.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target **sensitivity of around 80%**.
- The model achieved an **accuracy of 74%**, which is in line with the study's objectives.

Recommendation based on Final Model

- In accordance with the problem statement, enhancing lead conversion is critical for the growth and success of X Education. To achieve this goal, we have developed a regression model aimed at identifying the most influential factors impacting lead conversion.
- Based on our analysis, the following features exhibit the highest positive coefficients, indicating their significant impact on lead conversion and suggesting they should receive priority attention in our marketing and sales strategies:
  - Lead Source_Welingak Website: 5.39
  - Lead Source_Reference: 2.93
  - Current_occupation_Working Professional: 2.67
  - Last Activity_SMS Sent: 2.05
  - Last Activity_Others: 1.25
  - Total Time Spent on Website: 1.05
  - Last Activity_Email Opened: 0.94
  - Lead Source_Olark Chat: 0.91

- Conversely, we have identified features with negative coefficients that may indicate areas for potential improvement: These include:
  - Specialization in Hospitality Management: -1.09
  - Specialization in Others: -1.20
  - Lead Origin of Landing Page Submission: -1.26

Recommendation based on Final Model

To boost our Lead Conversion Rates

- Emphasize features with positive coefficients in targeted marketing strategies.
- Create methods to attract high-quality leads from top-performing sources.
- Engage working professionals with customized messaging.
- Optimize communication channels based on their impact on lead engagement.
- Allocate more budget for advertising on the Welingak Website.
- Offer incentives or discounts for referrals that convert into leads, encouraging more references.
- Aggressively target working professionals as they have a high conversion rate and better financial capability to pay higher fees.

To identify areas of improvement

- Examine negative coefficients in specialization offerings.
- Review landing page submission process to identify improvement opportunities.

THANK YOU!!!