X Education - Lead Scoring Case Study

Presented By -
Shweta Alukuru Trikutam
Trupti Nidhalkar
Tiana Pinto

Table of Contents

Background of X Education Company

- X Education specializes in online courses tailored for industry professionals.
- The company attracts a steady stream of visitors daily through targeted marketing on platforms like Google and various websites.
- Visitors typically explore course offerings, watch informative videos, and may fill out forms with their contact details.
- These form submissions classify visitors as leads.
- X Education's sales team engages leads through calls and emails to encourage enrollment.
- The company achieves a lead conversion rate of around 30%, converting interested leads into enrolled students.

Problem Statement & Objective of the Study

Problem Statement:

X Education seeks to improve its lead conversion rate from 30% to 80% by identifying and prioritizing "Hot Leads" through a predictive model that assigns lead scores based on conversion likelihood.

Objective of the Study:

Develop a predictive model for X Education to assign lead scores, prioritizing leads with higher likelihoods of converting into paying customers. Target a significant increase in lead conversion rate to around 80% through strategic lead prioritization and focused sales efforts.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are categorized based on their conversion propensity, creating a targeted group of "hot leads."

Better Communication

- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

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:
Loading Data Set, understanding & cleaning data

EDA:
Check imbalance, Univariate & Bivariate analysis

Data Preparation
Dummy variables, test-train split, feature scaling

Model Building:
RFE for top 15 feature, Manual Feature Reduction & finalizing model

Model Evaluation:
Confusion matrix, Cutoff Selection, assigning Lead Score

Predictions on Test Data:
Compare train vs test metrics, Assign Lead Score and get top features

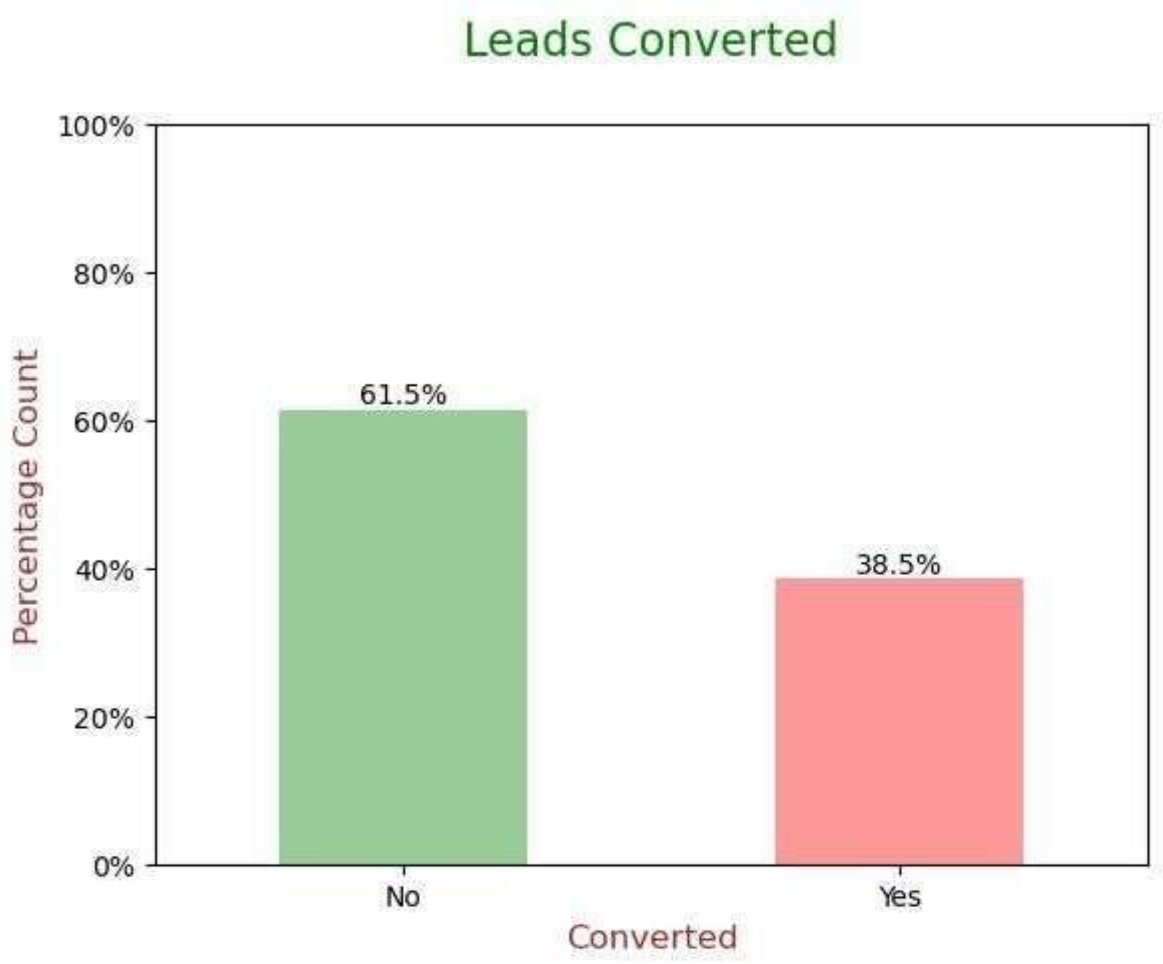
Recommendation:
Suggest top 3 features to focus for higher conversion & areas for improvement

Data Cleaning

- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and considerations.
- Columns irrelevant to the study objective (tags, country) were dropped.
- Imputation was applied to fill missing values in some categorical variables.
- Additional categories were created for certain variables.
- Columns with no modeling use (Prospect ID, Lead Number) or only one category were dropped.
- Numerical data was imputed using mode and outliers in 'TotalVisits' and 'Page Views Per Visit' were capped.
- Data standardization included fixing invalid values and standardizing casing (e.g., "Google" vs. "google").
- Low-frequency values in categorical variables were grouped into "Others" where applicable.
- Binary categorical variables were mapped.

EDA

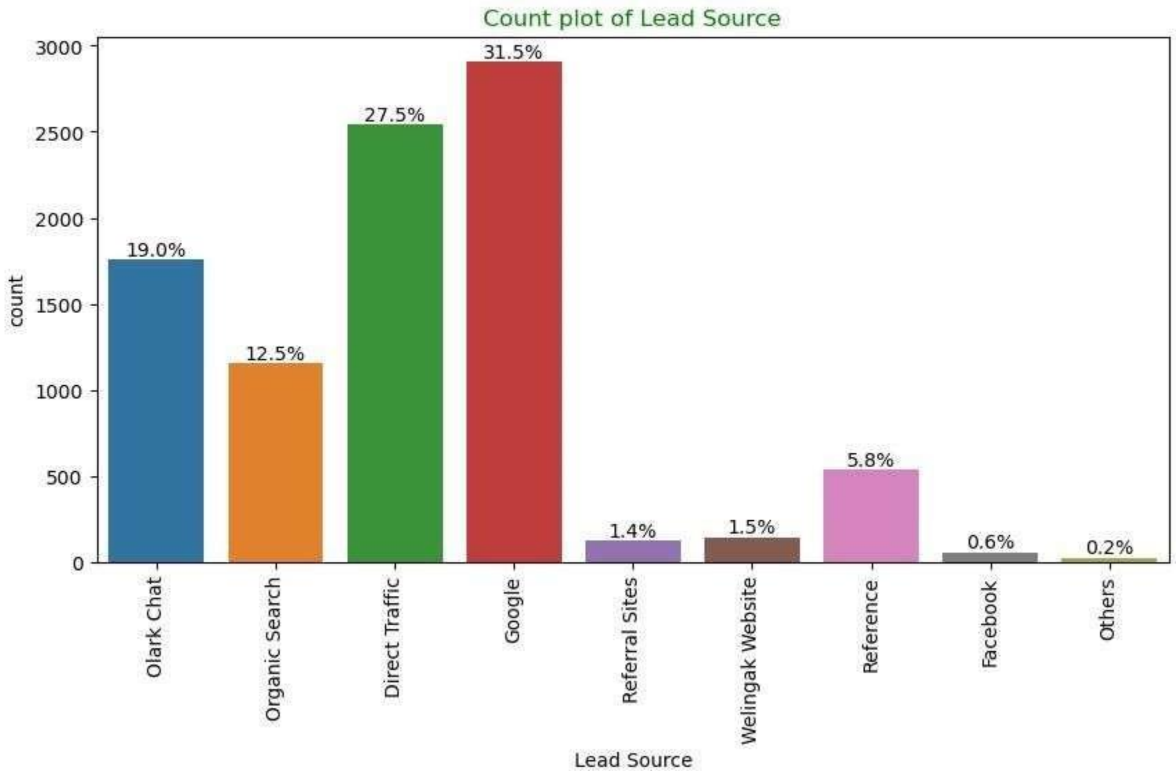
- Data is imbalanced while analyzing target variable.

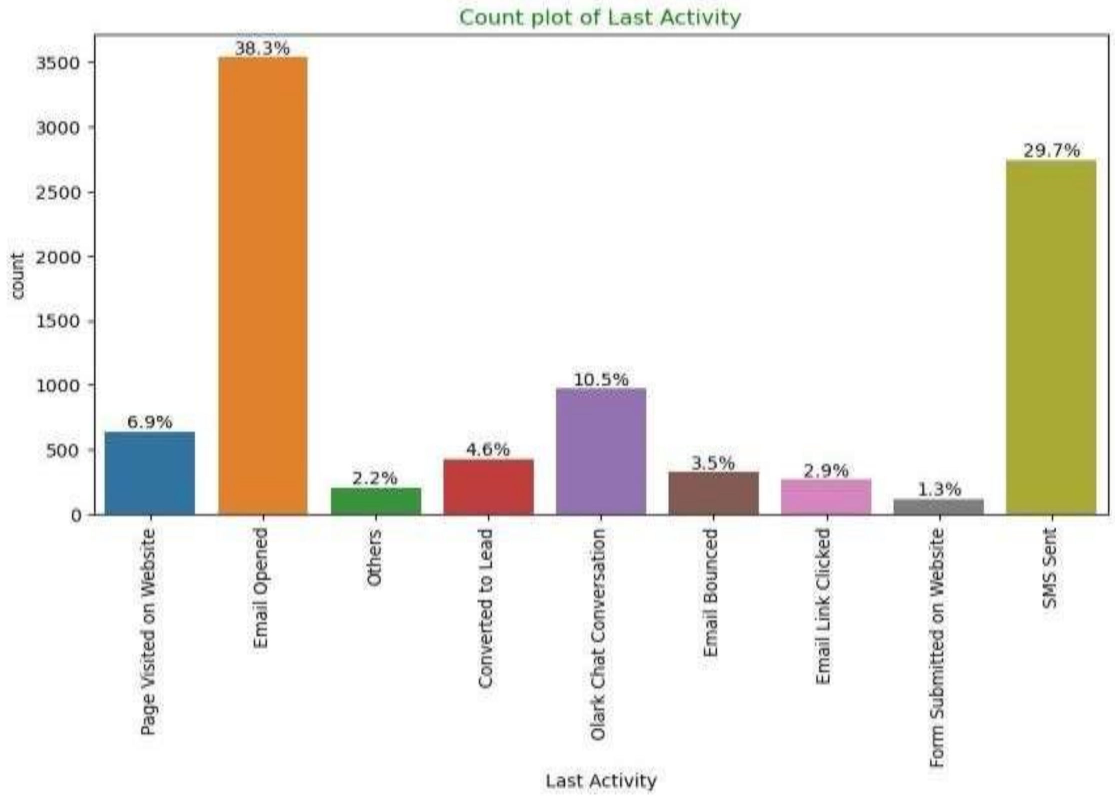


- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

EDA

- **Univariate Analysis – Categorical Variables**

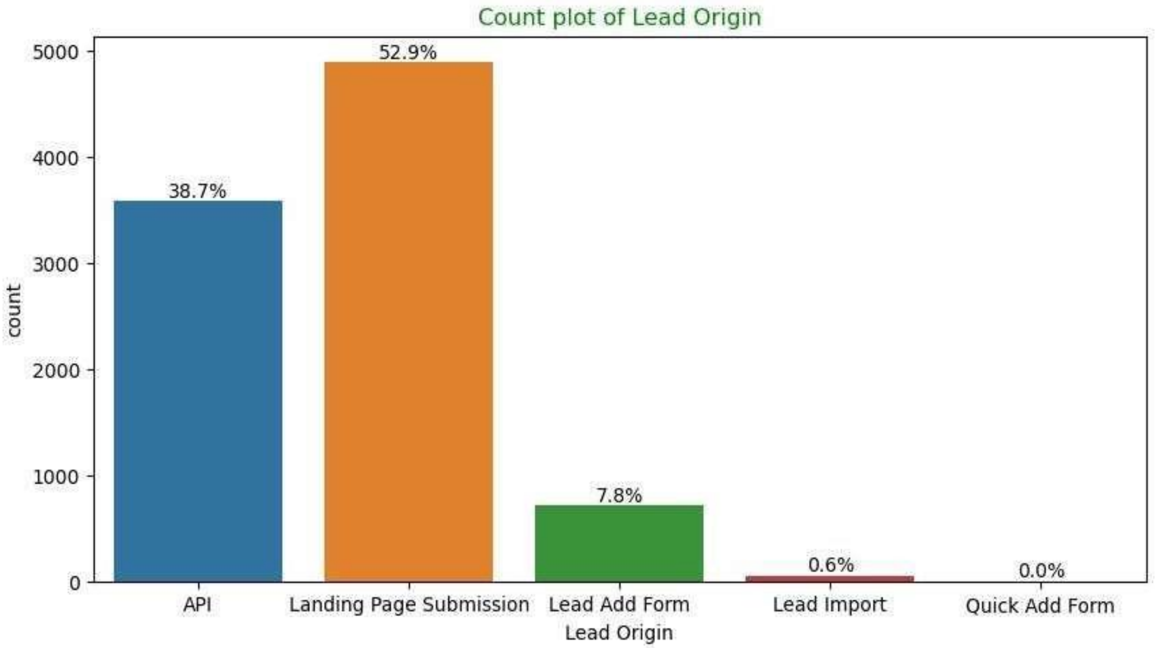


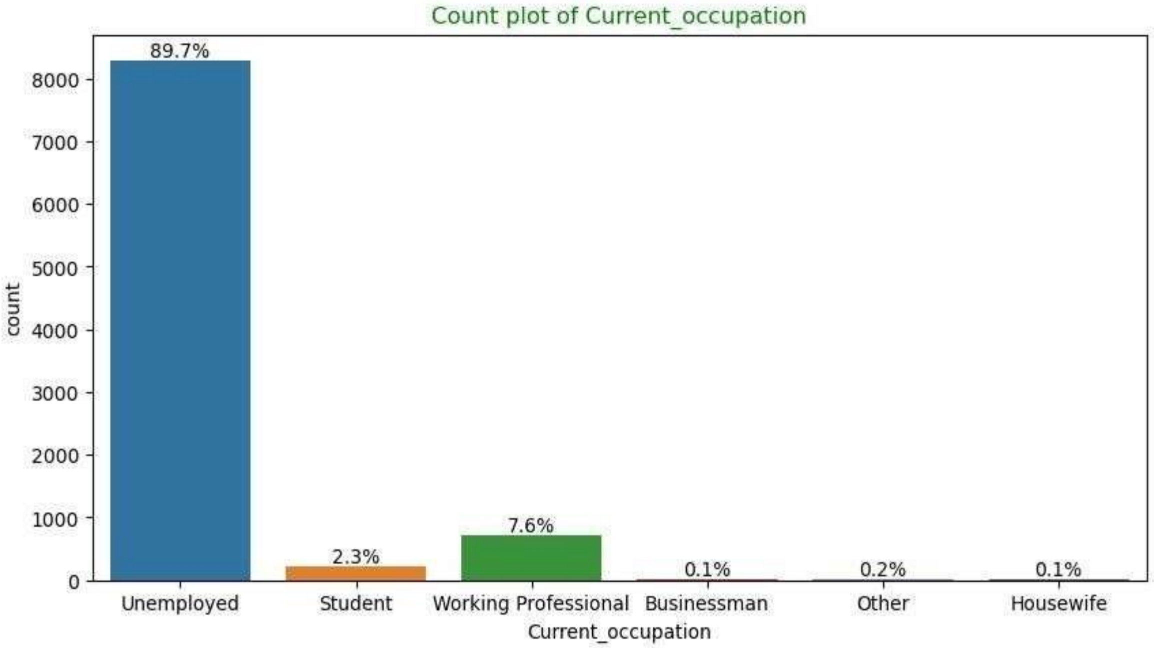


- **Lead Source:** 58% Lead source is from Google & Direct Traffic combined.
- **Last Activity:** 68% of customers contribution in SMS Sent & Email Opened activities.

EDA

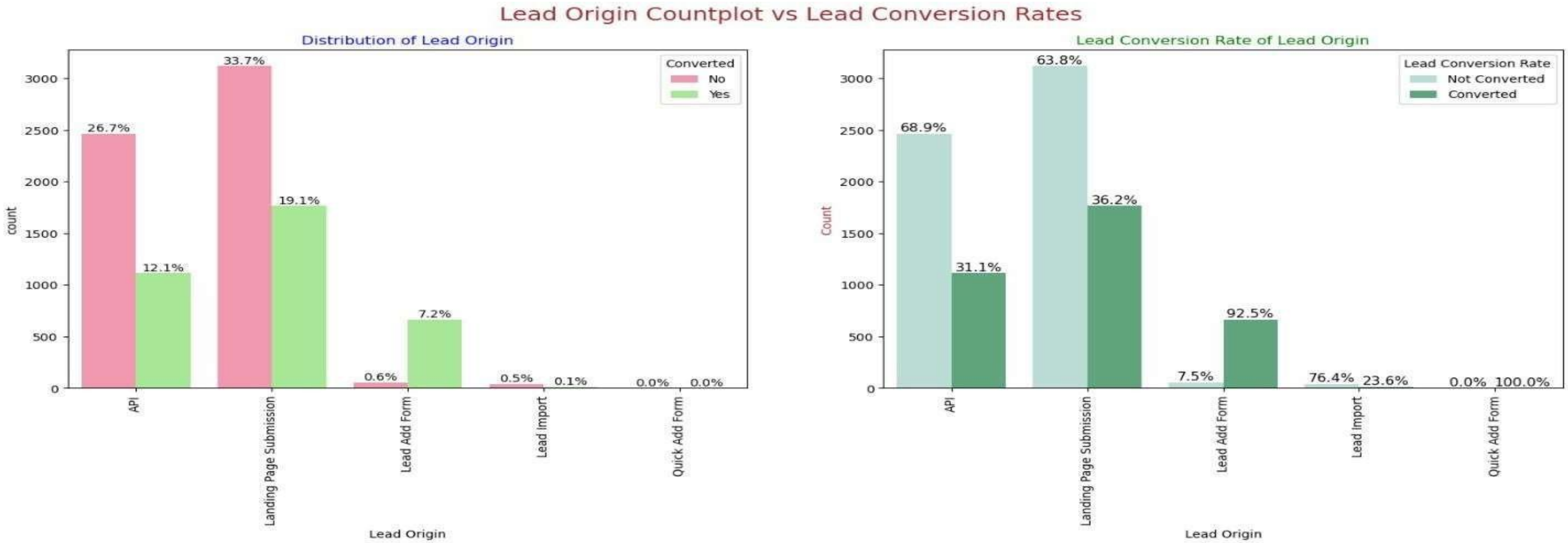
- **Univariate Analysis – Categorical Variables**





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

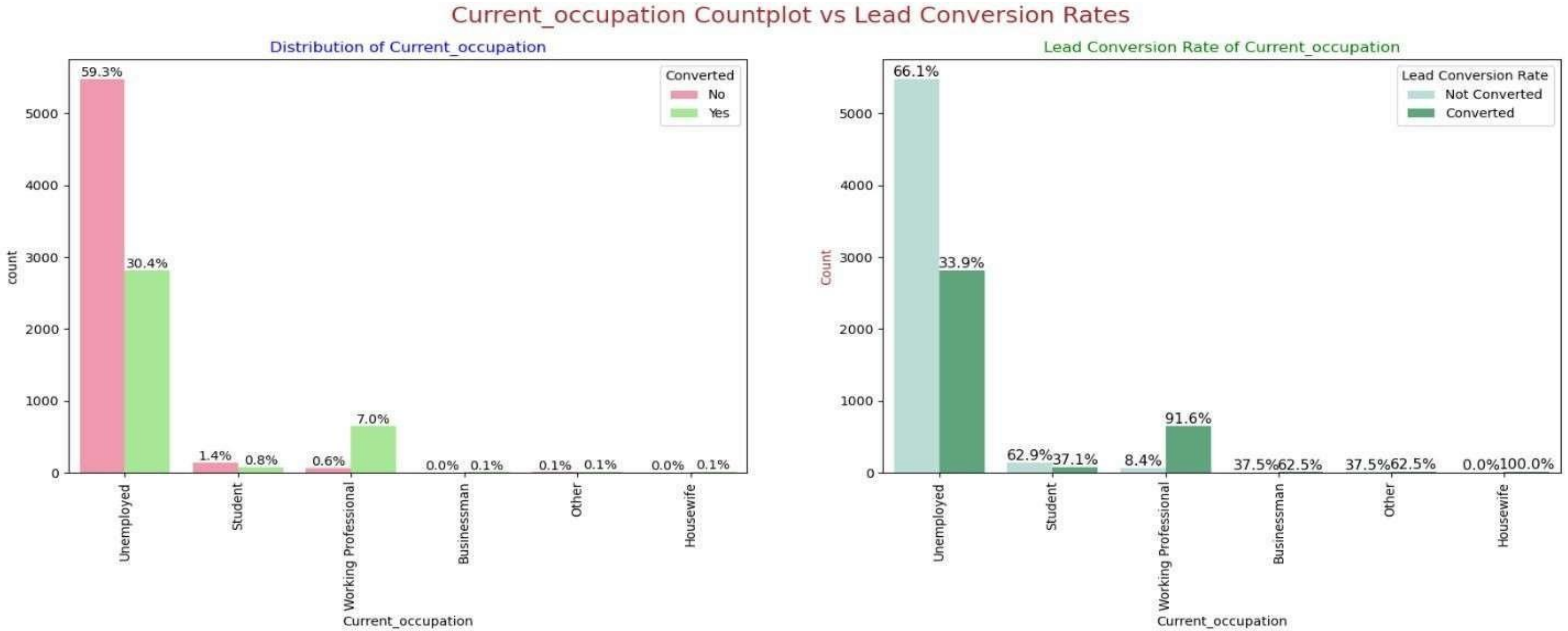
EDA – Bivariate Analysis for Categorical Variables



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%**.

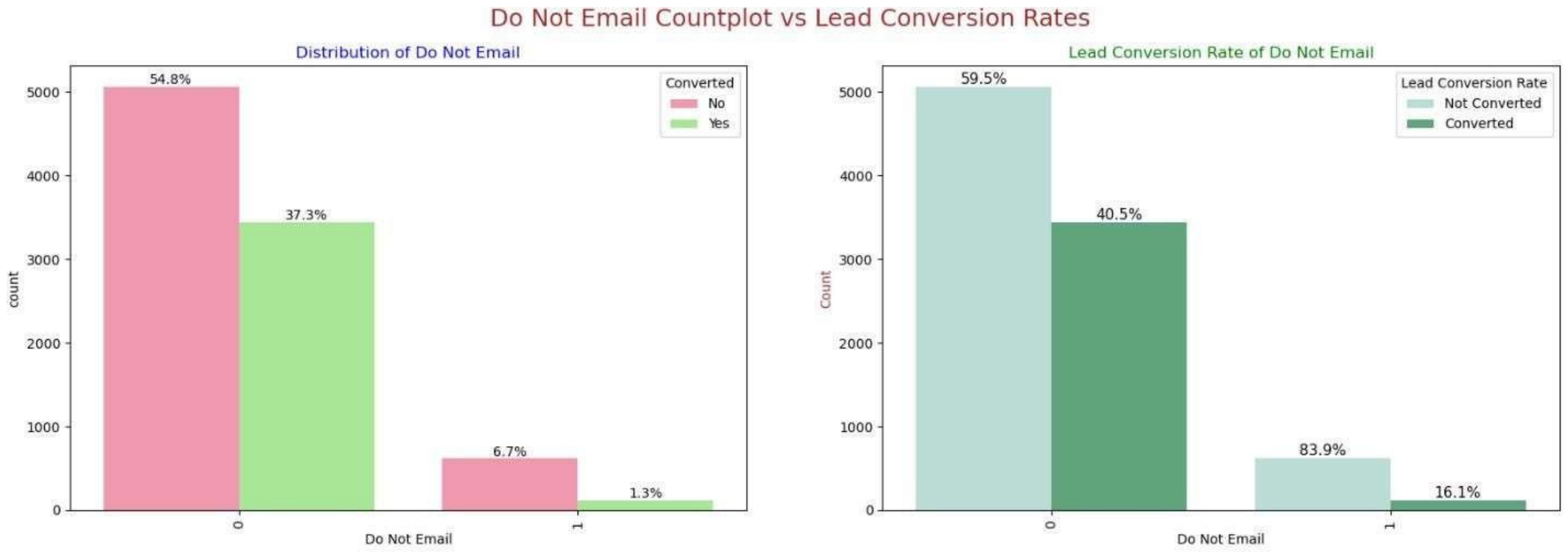
EDA – Bivariate Analysis for Categorical Variables



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)**.

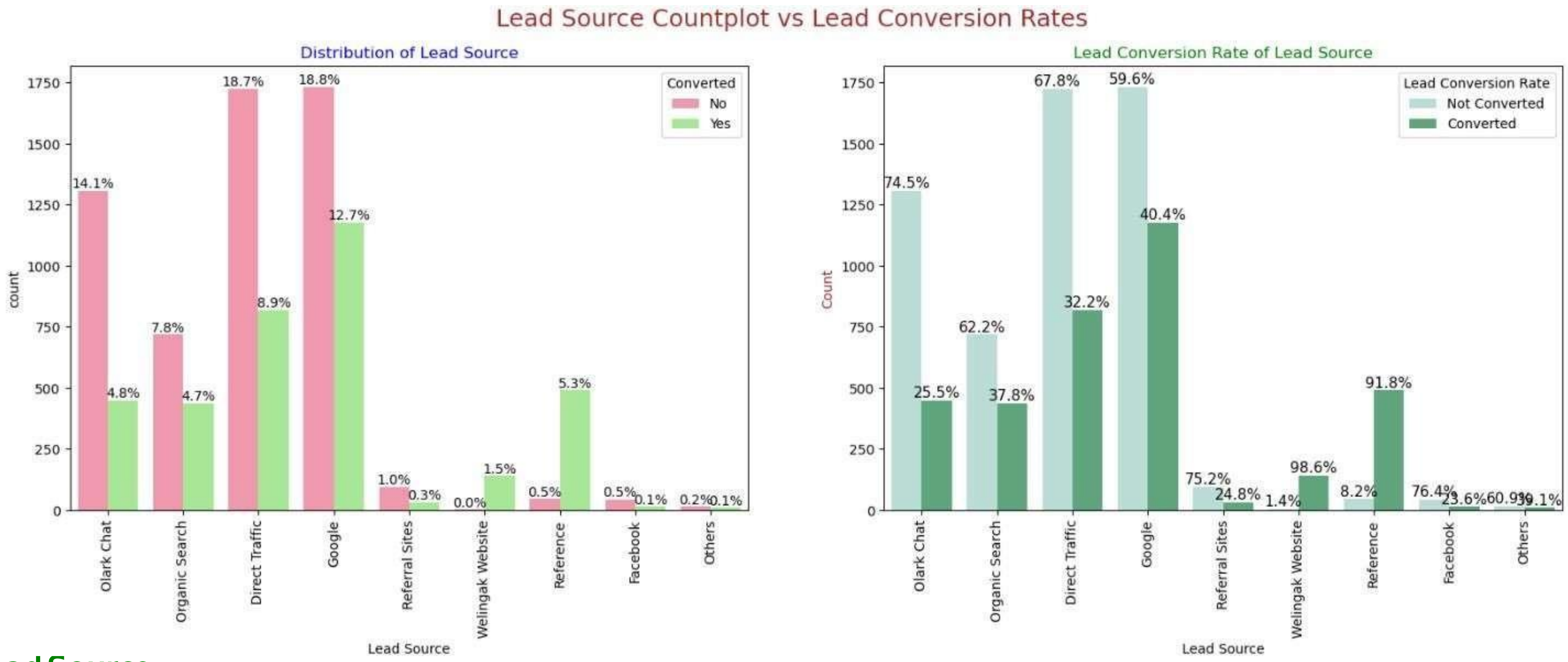
EDA – Bivariate Analysis for Categorical Variables



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.

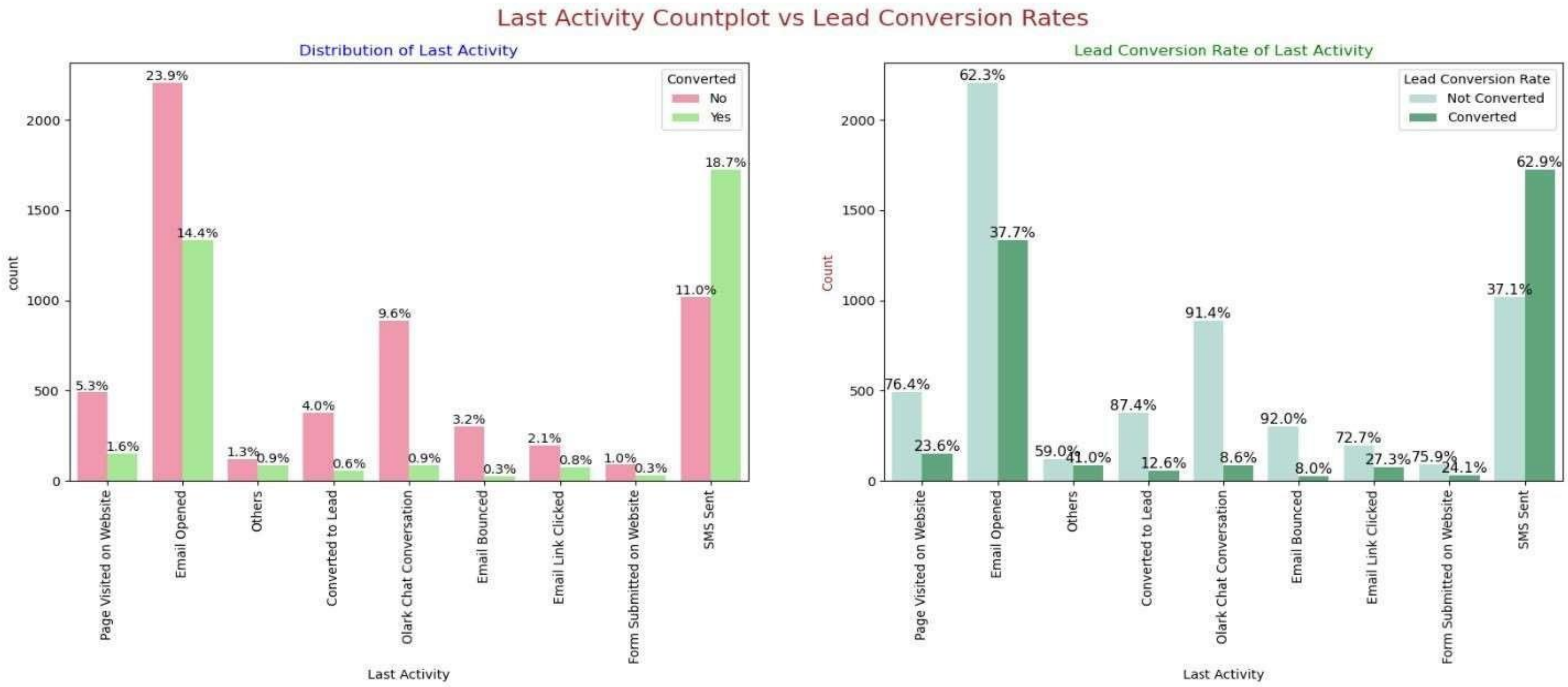
EDA – Bivariate Analysis for Categorical Variables



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.

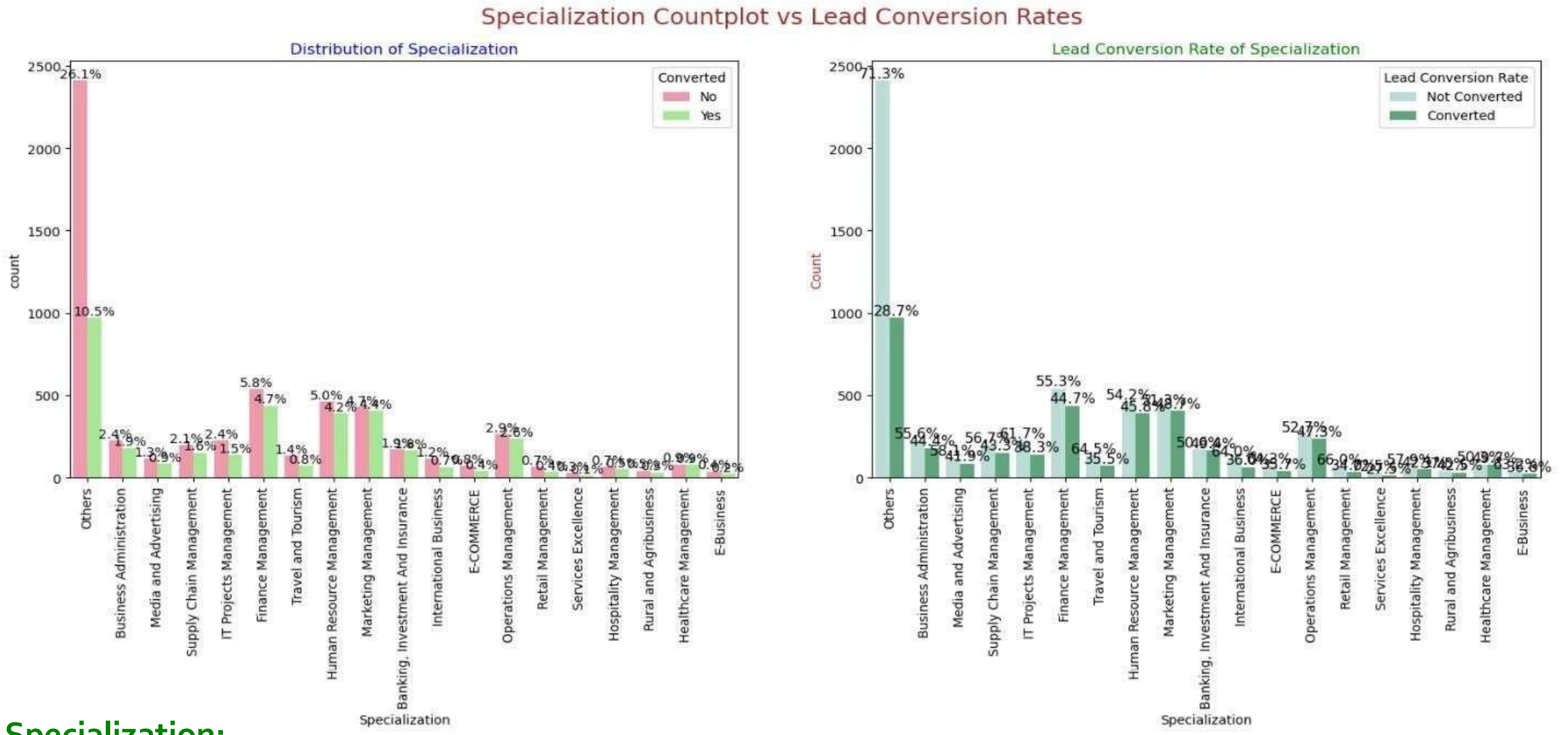
EDA – Bivariate Analysis for Categorical Variables



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.**

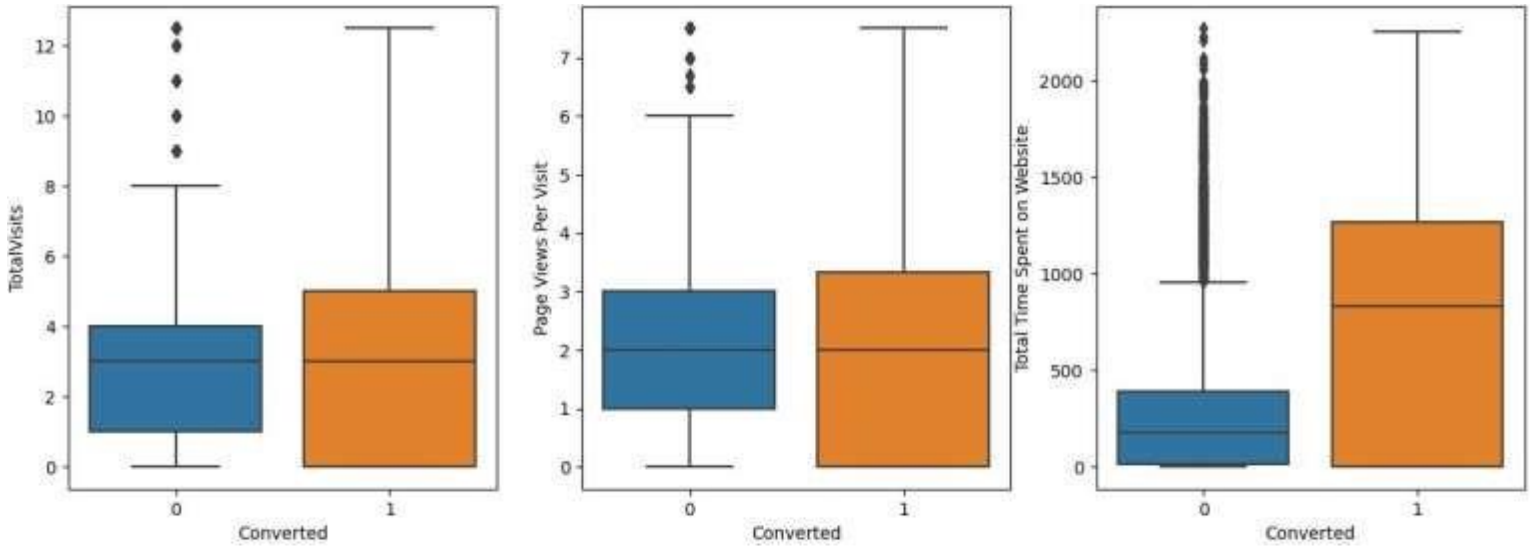
EDA – Bivariate Analysis for Categorical Variables

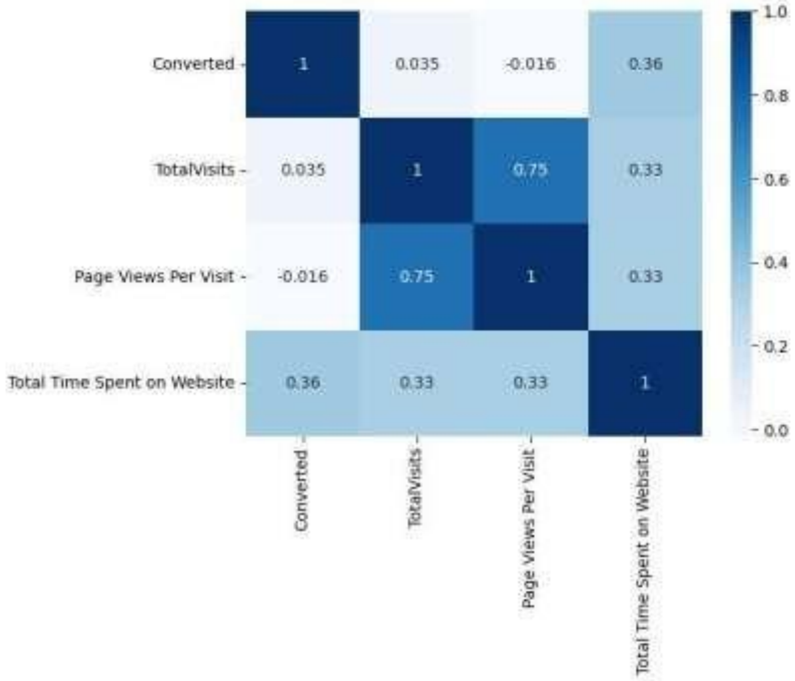


Specialization:

- 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

To optimize model performance and reduce computation time, Recursive Feature Elimination (RFE) was used to select important columns:

- Pre RFE: 48 columns
- Post RFE: 15 columns

Through manual feature reduction:

- Variables with p-values > 0.05 were dropped.
- After four iterations, Model 4 exhibited stability with significant p-values (< 0.05) and VIFs < 5.
- Therefore, logm4 is finalized for model evaluation and predictions.

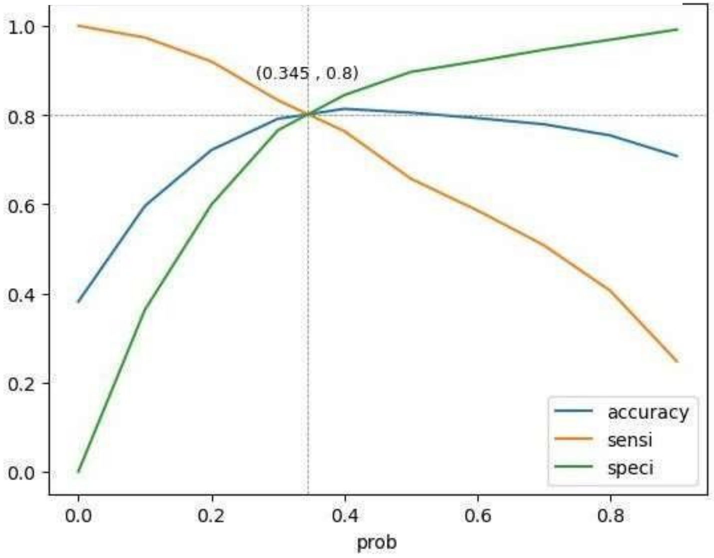
Model Evaluation

Train Data Set

It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots

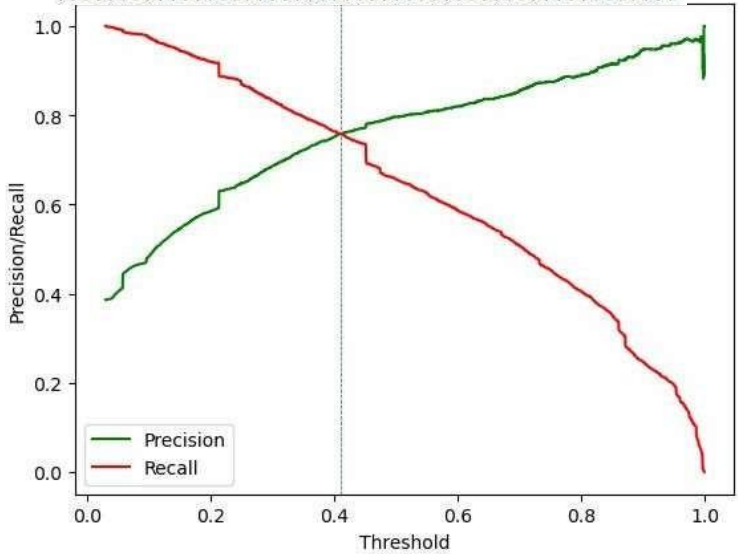
Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

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

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

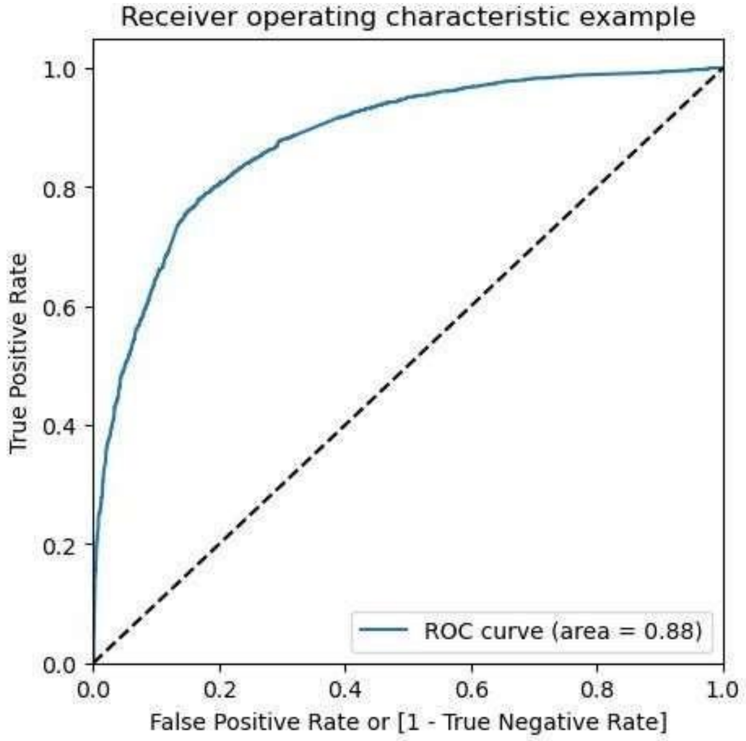
```
**************************************************

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

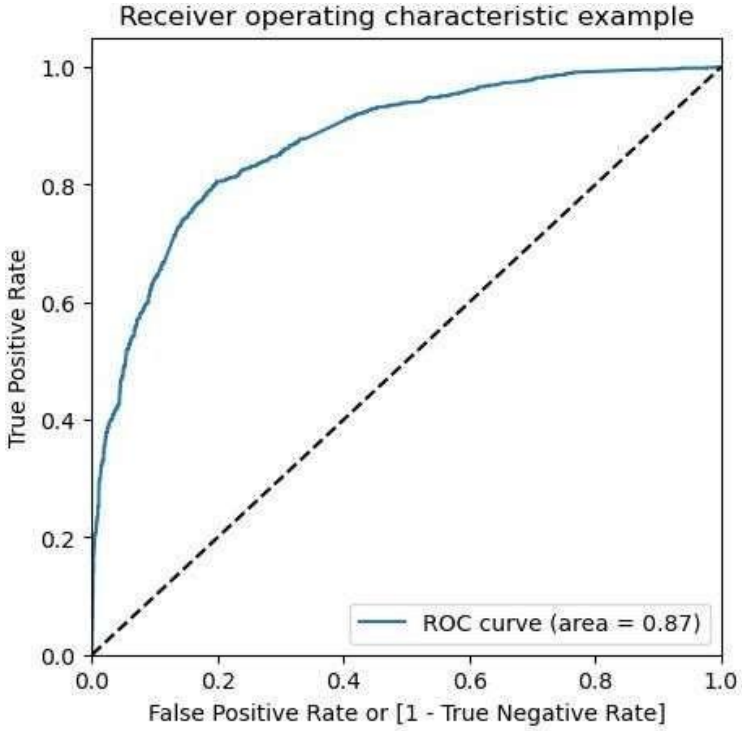
ROC Curve – Train Data Set

- 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
```

- Using a cut-off value of 0.345, the model achieved a **sensitivity** of **80.05% in the train set** and **79.82% 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 also achieved an **accuracy of 80.46%**, which is in line with the study's objectives.

Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, a regression model that can help us identify the most significant factors that impact lead conversion.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
    - 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

- We have also identified features with negative coefficients that may indicate potential areas for 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 increase our Lead Conversion Rates

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Optimize communication channels based on lead engagement impact.
- Engage working professionals with tailored messaging.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement

- Analyze negative coefficients in specialization offerings.
- Review landing page submission process for areas of improvement.

Thank you