
# Lead Scoring Case Study

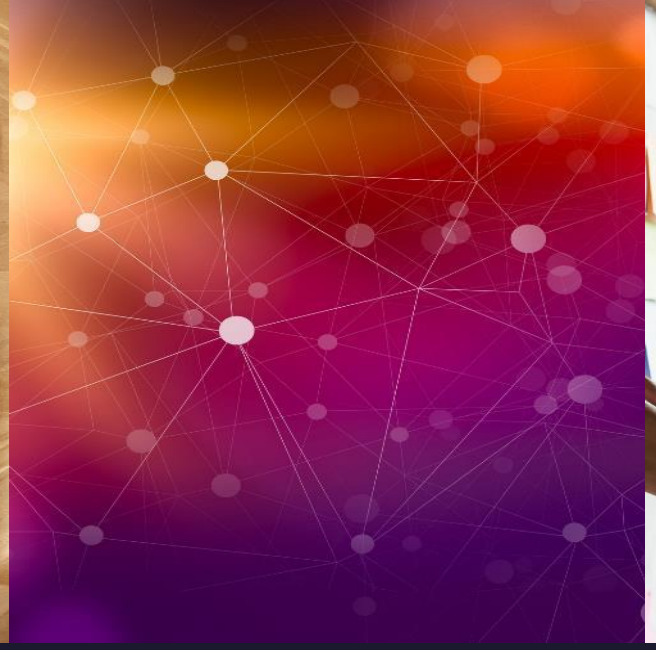
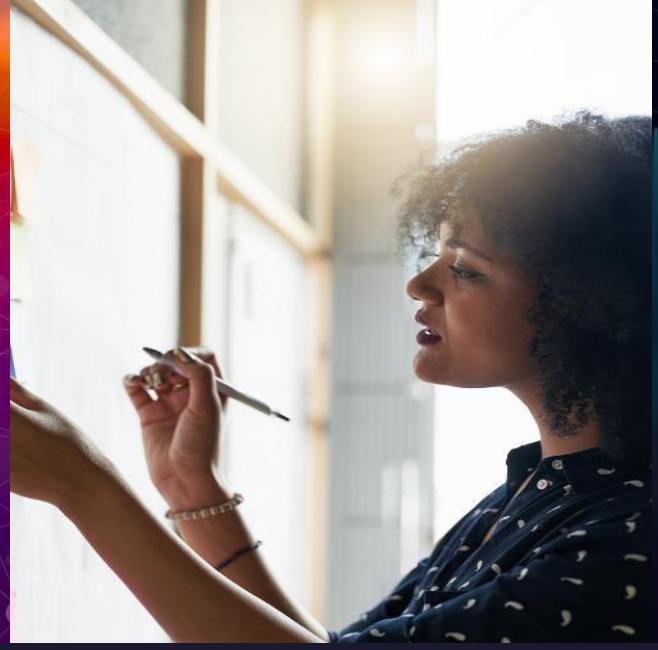
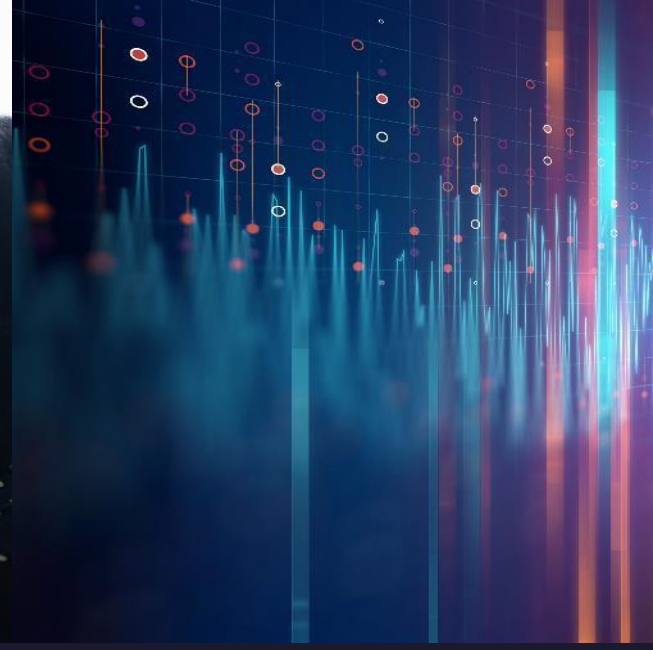
Student: Le Thi Hong Ha

Class: C11

Course: Machine Learning

# Table of Contents



# Background of X Education Company

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc

# 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, also 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 that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

# Analysis Steps

**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 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 & enhance conversion rate

# Data cleaning

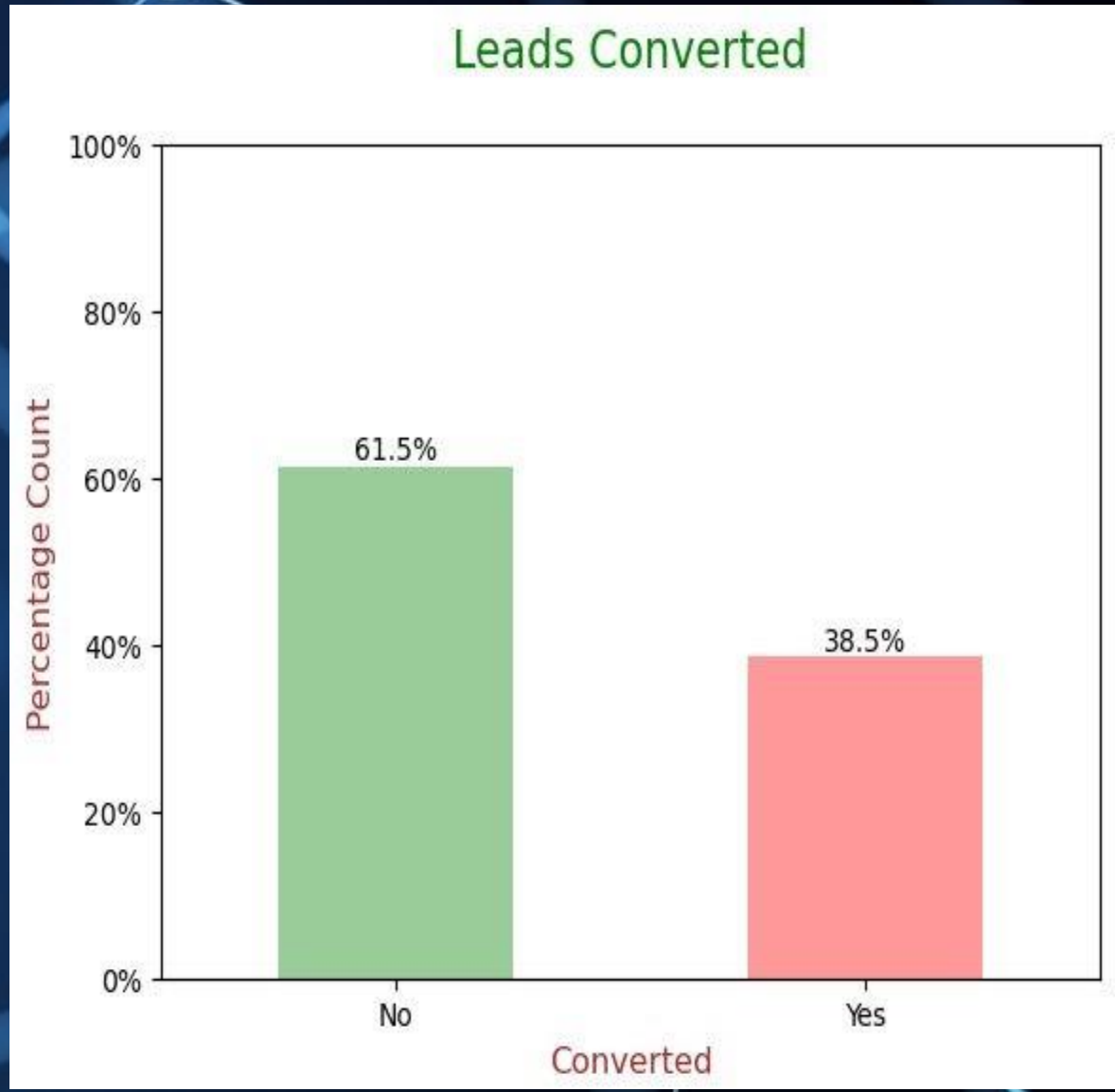
- **"Select"** level represents null values for some categorical variables, as customers did not choose any option from the list.
- Check missing/null values
- Drop columns with over 40% null values
- Drop redundant columns that are not used for modeling, such as (tags, country, Prospect ID, Lead Number) )
- Impute some categorical variables.
- Create new variables.
- Standardize data: google -> Google

# EDA

## Data Imbalance

- 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

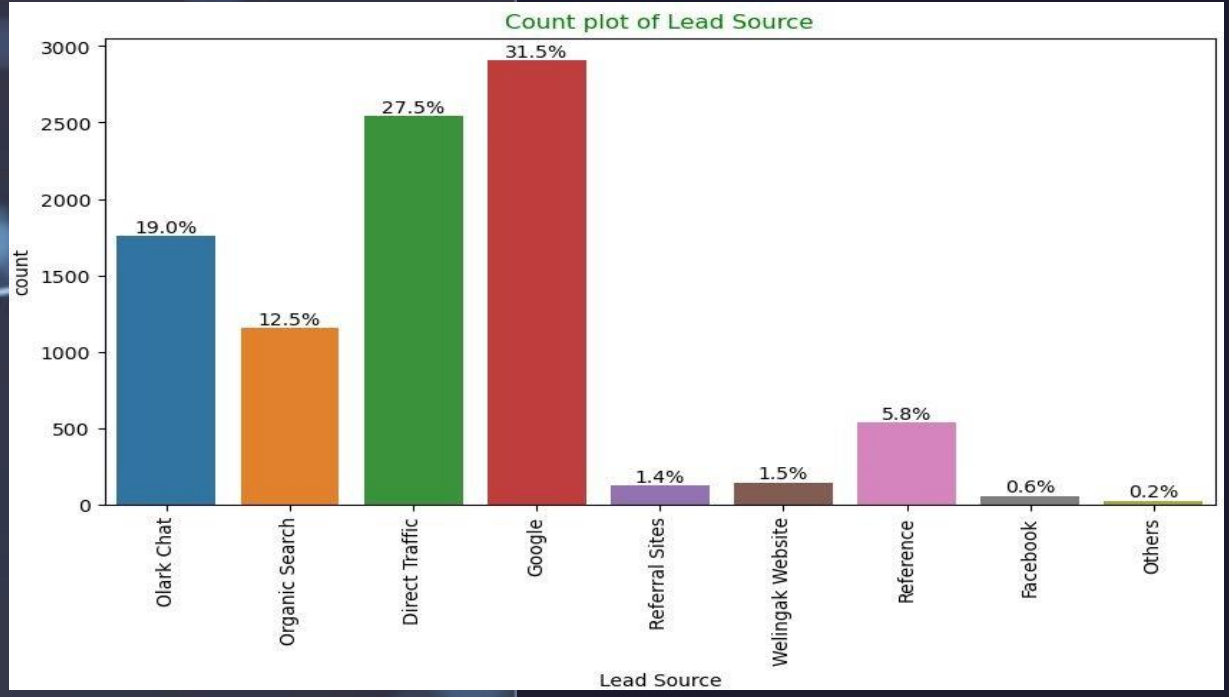
Univariate Analysis – Categorical Variables

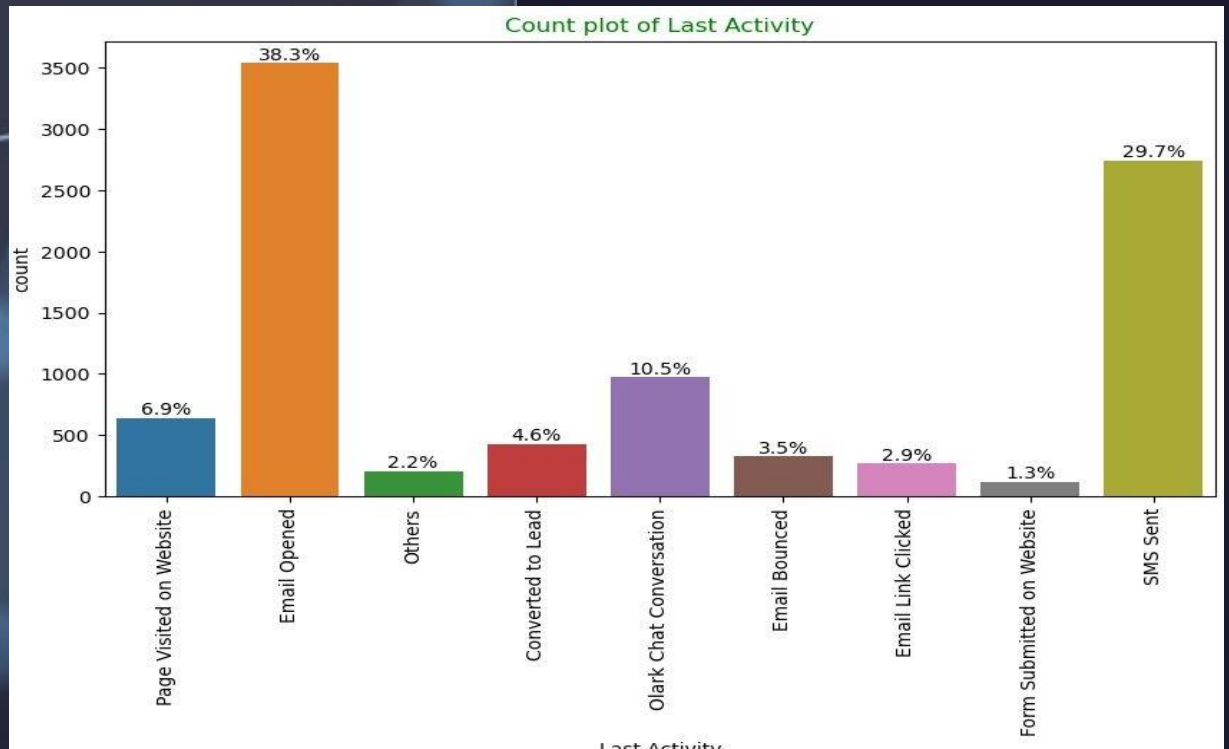
**Lead Source:**

58% Lead source is from Google & Direct Traffic combined.

**Last Activity:**

68% of customers contribution in SMS Sent & Email Opened activities.



# EDA

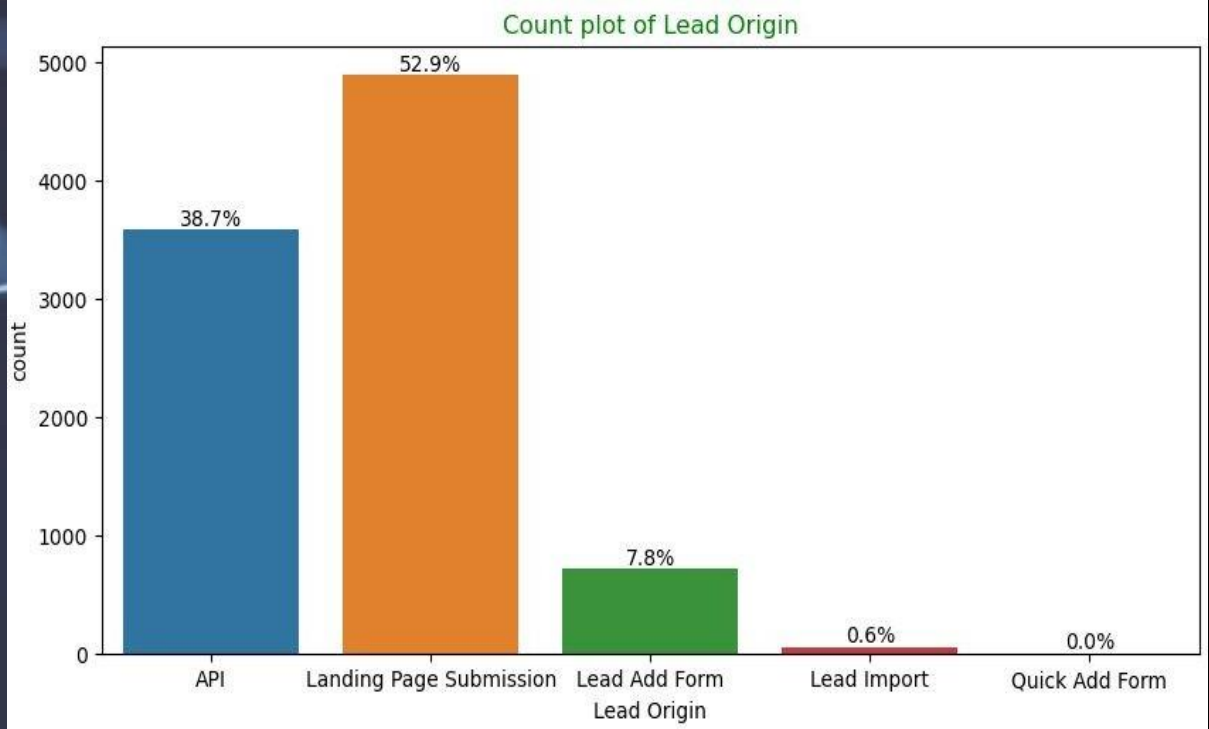
Univariate Analysis – Categorical Variables

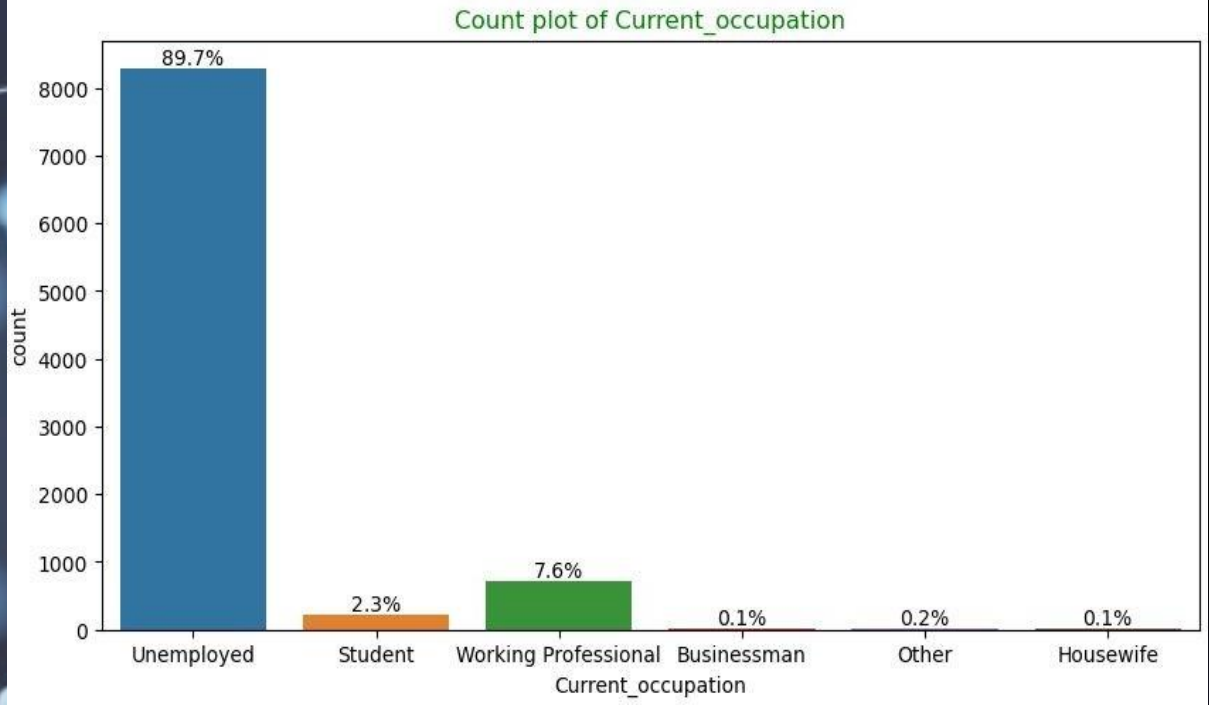
**Lead Origin:**

"Landing Page Submission" identified 53% of customers, "API" identified 39%.

**Current_occupation:**

It has 90% of the customers as Unemployed



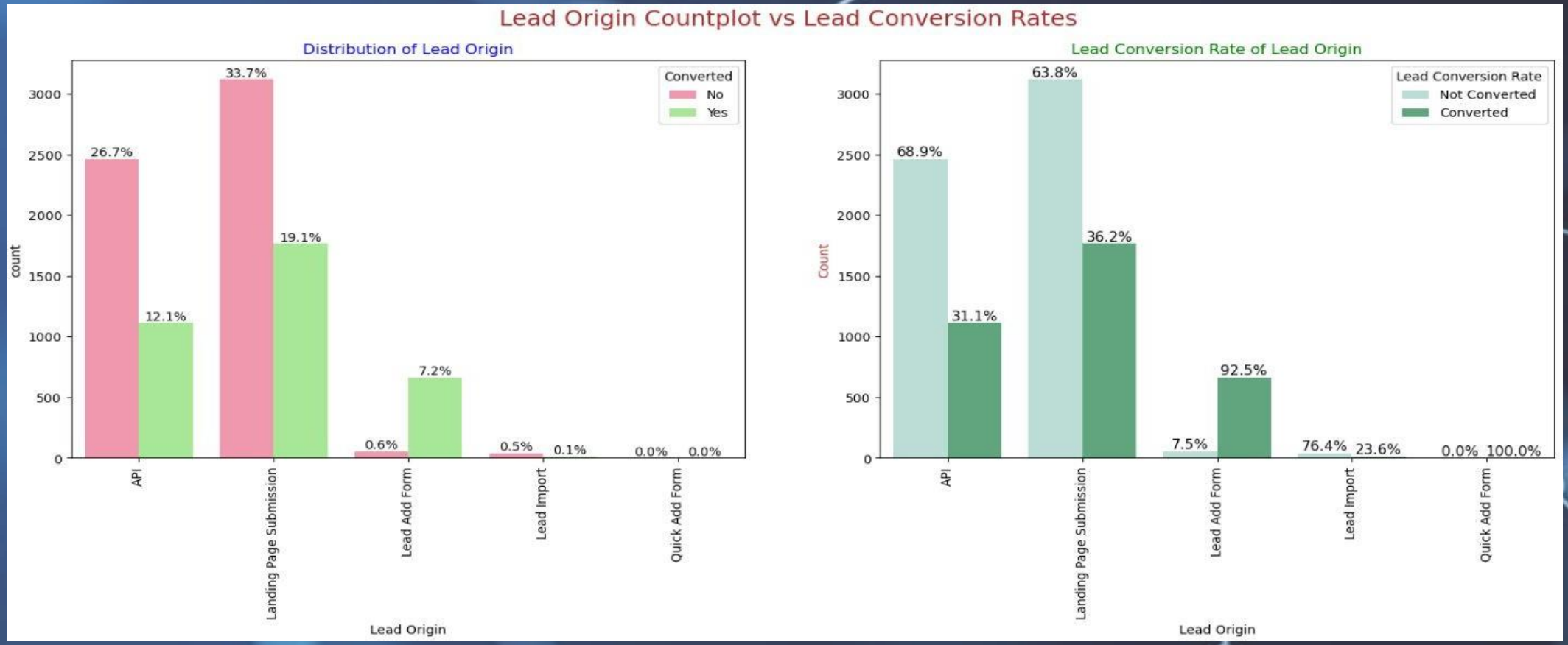



# EDA

## Bivariate Analysis for Categorical Variables

### Lead Origin:

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

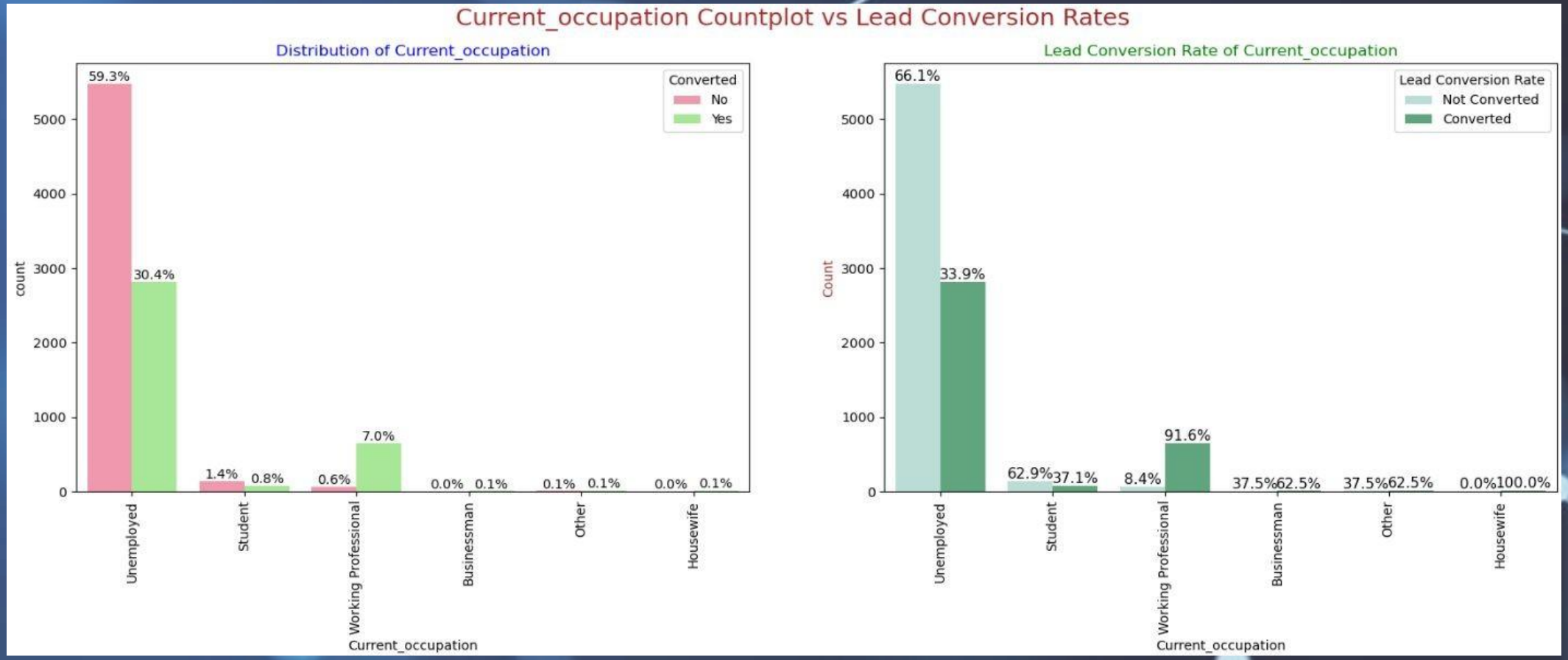


# EDA

## Bivariate Analysis for Categorical Variables

**Current_occupation:**

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

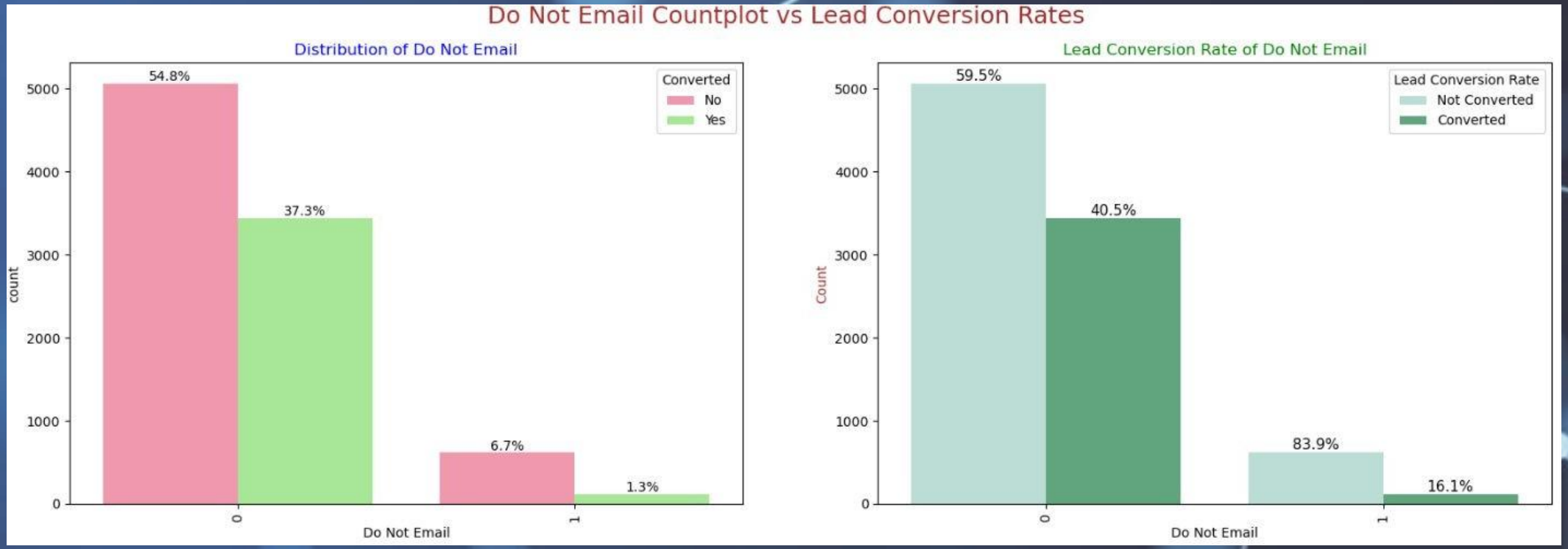


# EDA

## 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.

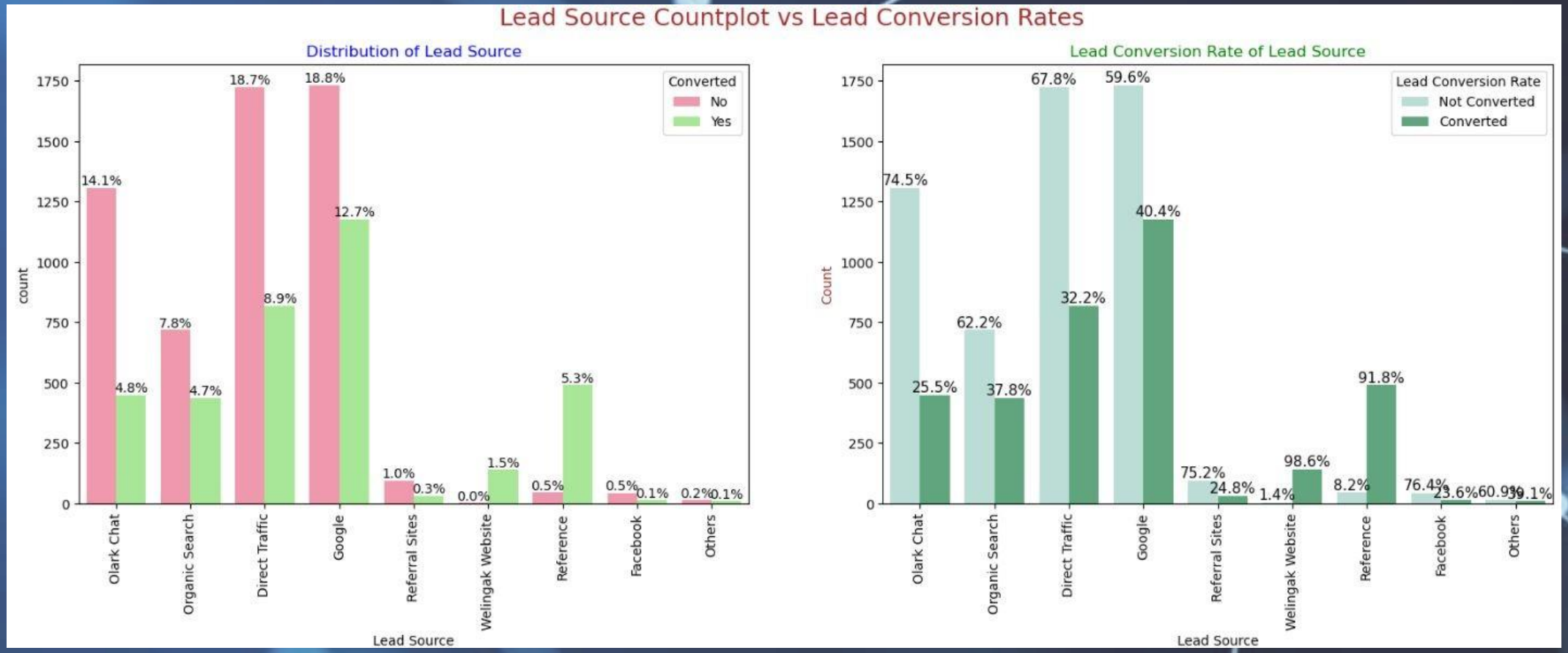


# EDA

## Bivariate Analysis for Categorical Variables

**Lead Source:**

- Google has **Lead conversion rate. of 40%** out of 31% customers.
- Direct Traffic contributes 32% Lead conversion rate with 27% customers, which is lower than Google

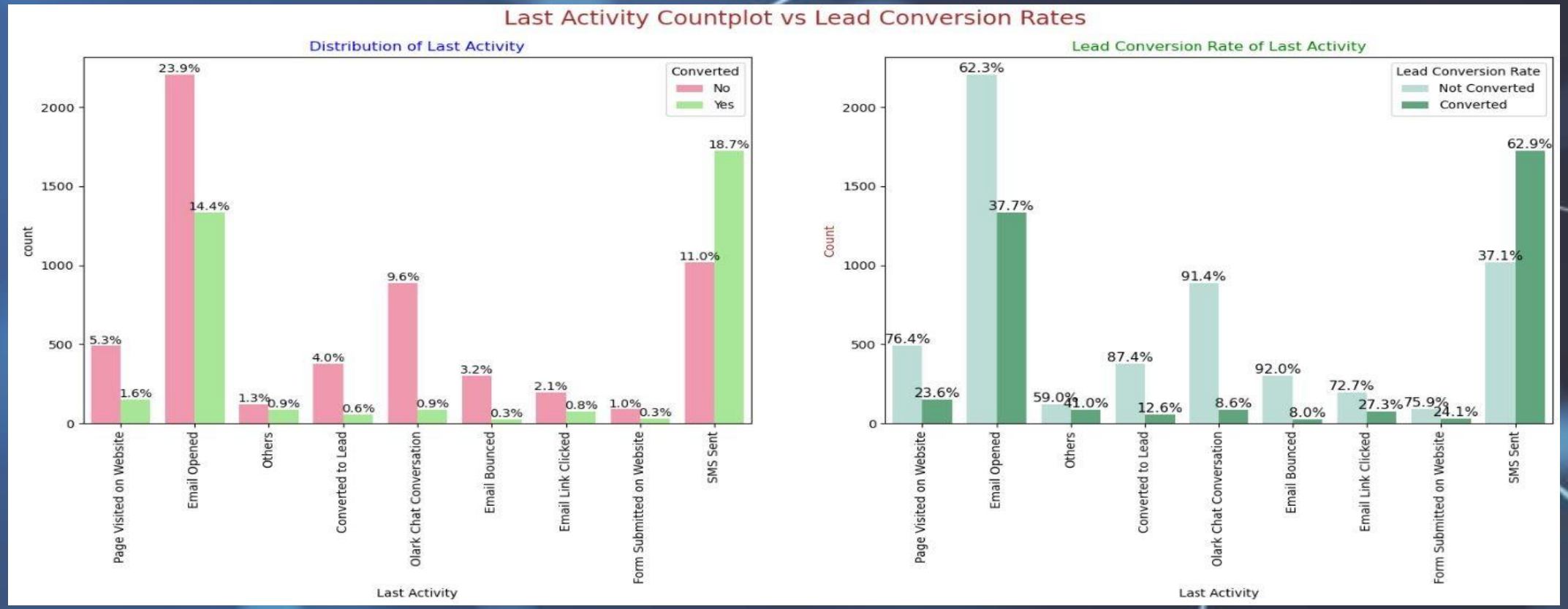


# EDA

## 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**

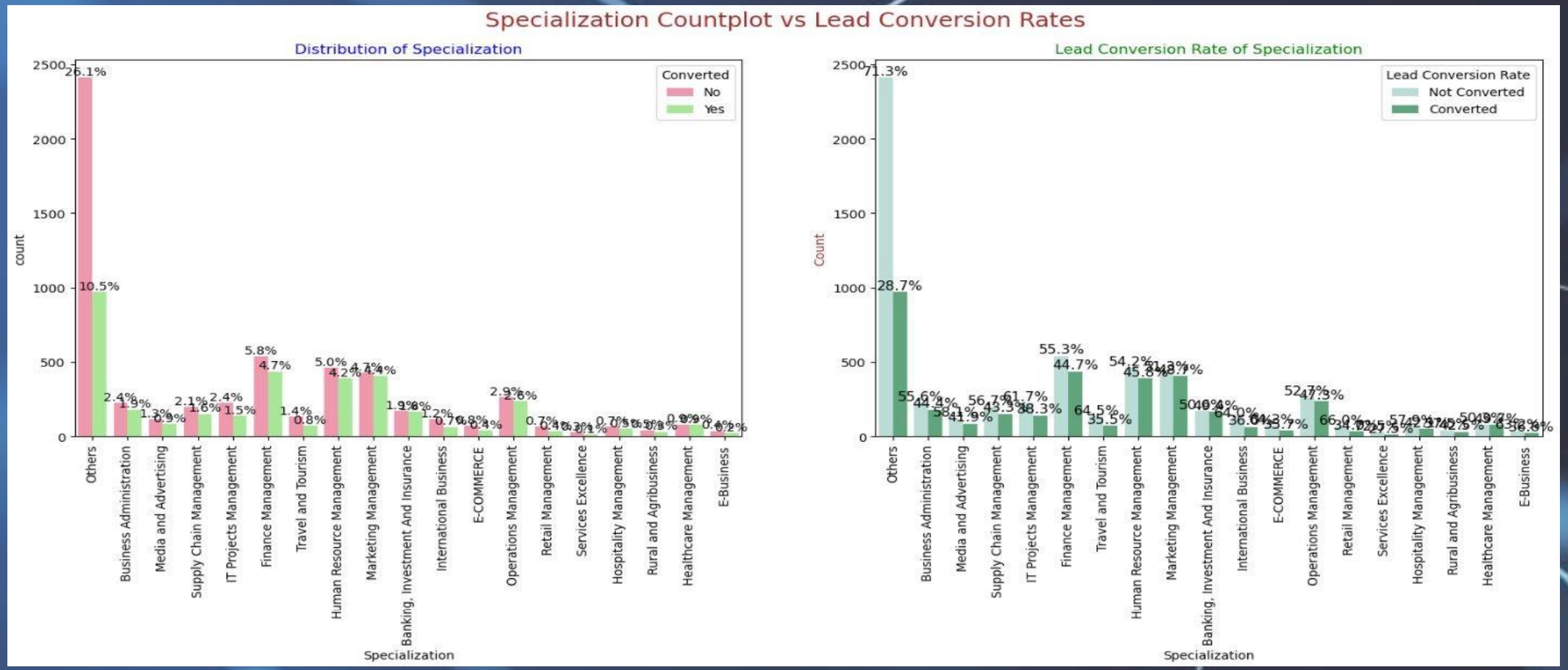


# EDA

## Bivariate Analysis for Categorical Variables

**Specialization:**

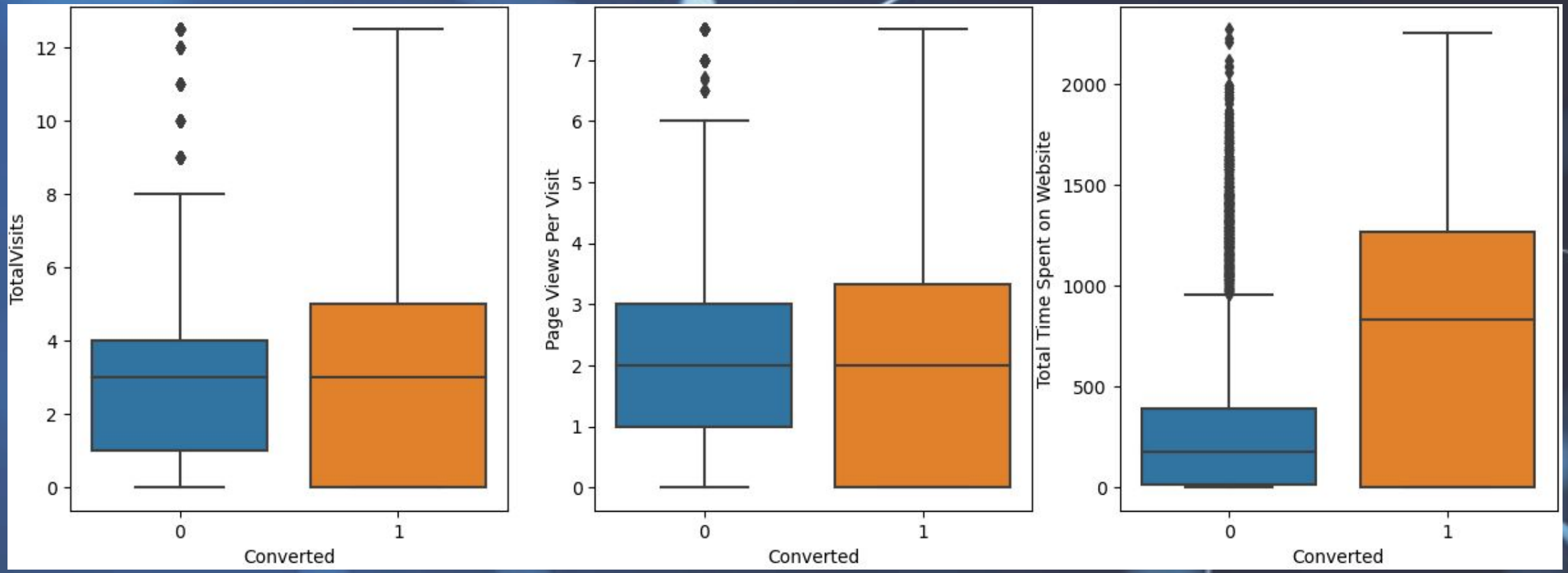
- Marketing Management, HR Management, Finance Management shows good contribution in Leads conversion than other specialization.



# EDA

## Bivariate Analysis for Numeric 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

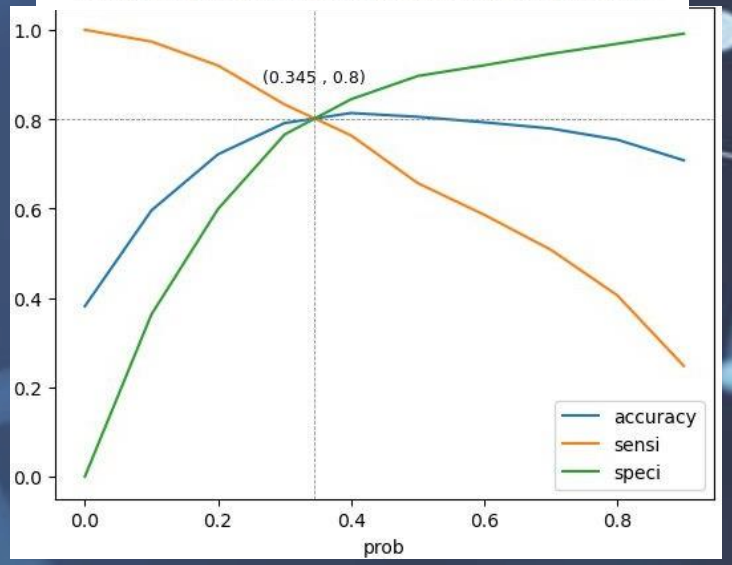
- Perform **Recursive Feature Elimination** (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome
  - Pre RFE – 48 columns & Post RFE – 15 columns
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 4 looks stable after four iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5
- Hence, **logm4** will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

# Model Evaluation

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

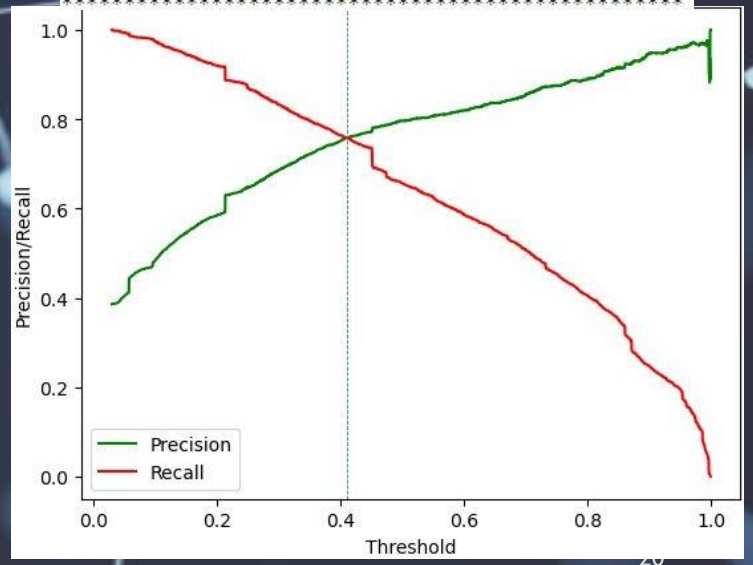
Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

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

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

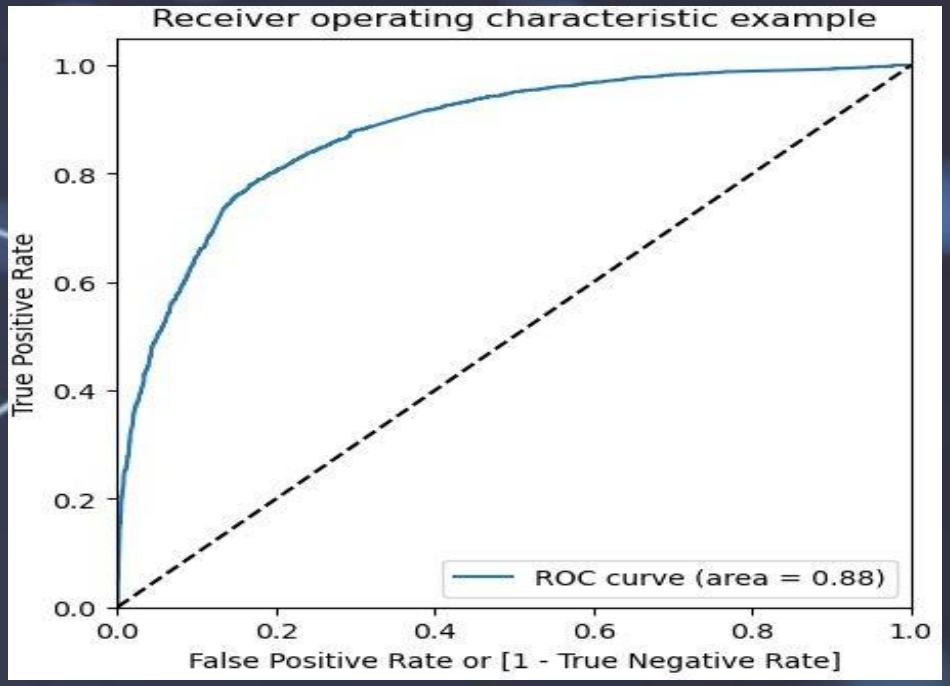
```
**************************************************

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

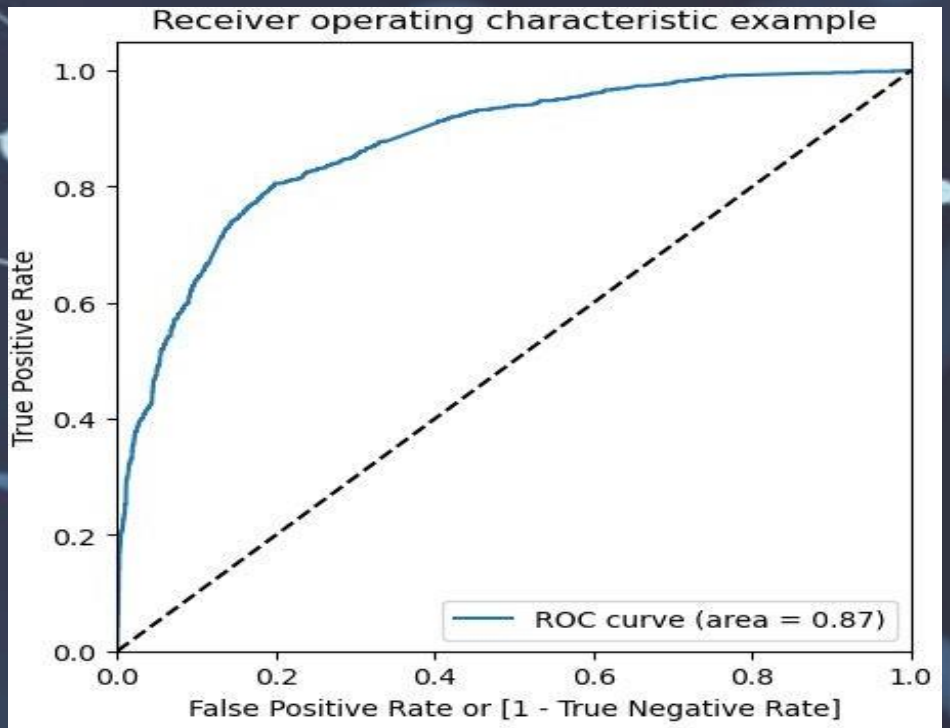
ROC Curve – Train Data Set

- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.




ROC Curve – Test Data Set

- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.



# 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, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

- Top 3 positive coefficients, priority in the marketing and sales efforts to increase lead conversion, including.
  - Lead Source_Welingak Website: 5.39
  - Lead Source_Reference: 2.93
  - Current_occupation_Working Professional: 2.67

- Negative coefficients that may indicate potential areas for improvement, including:
  - Specialization in Hospitality Management: -1.09
  - Specialization in Others: -1.20
  - Lead Origin of Landing Page Submission: -1.26

```
Lead Source_Welingak Website                 5.388662
Lead Source_Reference                        2.925326
Current_occupation_Working Professional      2.669665
Last Activity_SMS Sent                       2.051879
Last Activity_Others                         1.253061
Total Time Spent on Website                  1.049789
Last Activity_Email Opened                   0.942099
Lead Source_Olark Chat                       0.907184
Last Activity_Olark Chat Conversation       -0.555605
const                                       -1.023594
Specialization_Hospitality Management       -1.094445
Specialization_Others                       -1.203333
Lead Origin_Landing Page Submission         -1.258954
dtype: float64
```

# Recommendation based on Final Model

- 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.

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

# Summary

•The model achieved a sensitivity of 80.05% in the train set and 80% in the test set, using a cut-off value of 0.345.

•The Optimal cutoff probability point is 0.345.Converted probability greater than 0.345 will be predicted as Converted lead (Hot lead) & probability smaller than 0.345 will be predicted as not Converted lead (Cold lead).

•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.

# Thank You
# For Your Attention!