Lead Scoring Case study – X Education company

Team Members:

- Aniruddh Bhat
- Pragati Chauhan
- Vijayendra Kumar Gupta

Detecting hot leads using Logistic Regression to increase conversion rates through targeted marketing.

Table of contents

Company Overview

- The company “X Education sells online courses for professionals at the industrial level
- It has a website where many people browse their courses
- The courses are marketed digitally on various websites and search engines like Google
- When people visit the website, they are engaged through videos and forms, through which the leads are collected
- The forms collect their email address and phone number
- After acquiring the leads, the sales agents contact them through phone, email, SMS, etc., through which leads are converted
- The average lead conversion rate of this company is 30%

Problem Statement

- The lead conversion is very low (30%)
- The company wishes to increase the efficiency of its conversion by identifying the hot leads
- The sales department needs to be given insights on which leads they need to nurture more, to get maximum conversions
- We should build a model that will assign a lead score to each of the leads, where leads with higher chances of conversion are assigned a higher score and those with a lower chance are given a lower score
- The CEO has given a ballpark of the target lead conversion rate to be around 80%

Ideas to convert leads

Segregating the leads

- Leads are classified based on their probability to convert
- This helps to nurture targeted leads.

Optimizing the communication channels

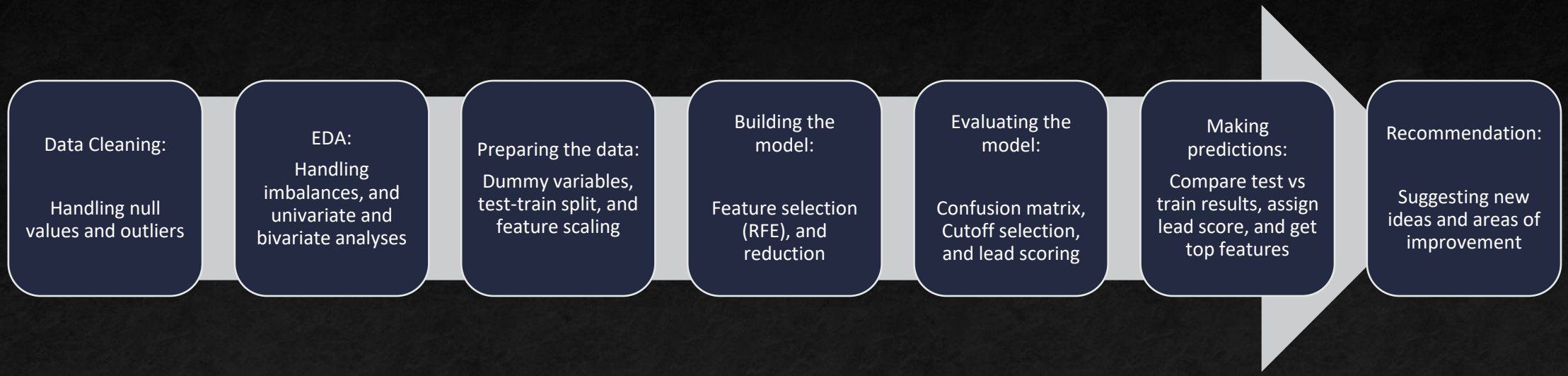
- Focusing on targeted leads helps us invest maximum time on leads with higher probability of conversion

Increasing conversion rate

- With more focus on targeted groups, we can have higher conversion rates and positively hit the 80% target, as set by the CEO.

As we have a target of 80% conversion rate, the hot leads must be highly sensitive to change in variables.

Steps taken for data analysis



Data Cleaning

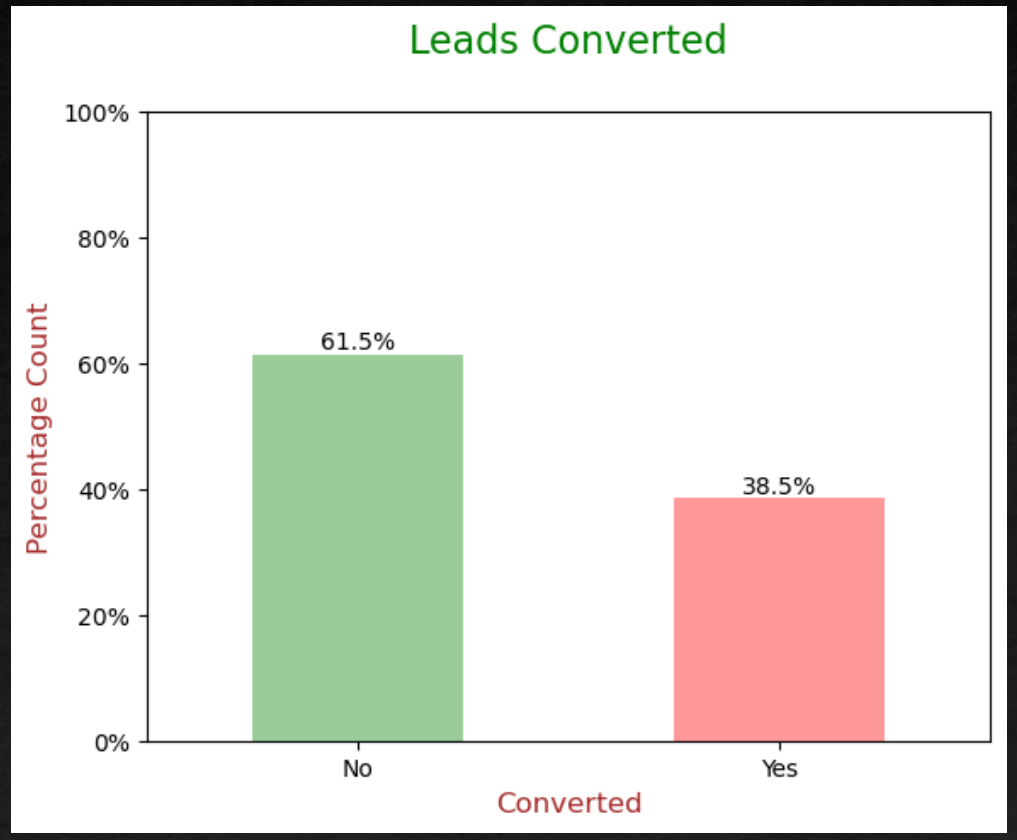
- “Select” is a default option and must be considered as “null” for some categorical variables. This is because no option was selected.
- Columns with more than 40% null values are dropped
- Missing values in columns of categorical variables are handled based on value counts and some other factors
- Columns that do not add much value to the objective of our study are dropped
- Imputation was done on certain categorical variables
- Additional categories were created for some variables
- Columns of less significance for modelling, such as Prospect ID, Lead Number, and those with only one category of response were also dropped
- Numerical data was imputed with mode after checking distribution

Data Cleaning

- Columns with skewed categorical variables were checked and dropped to reduce bias in the logistic regression models
- Outliers in **Total Visits** and **Page views Per Visit** were treated and capped
- Invalid values were defaulted, and standardized in some columns
- Values with low frequency were grouped together as “Others”
- Binary categorical variables were mapped together
- Additional cleaning processes were carried out to optimize data quality and accuracy:
  →Invalid values were fixed and standardized by changing their casing styles

Exploratory Data Analysis (EDA)

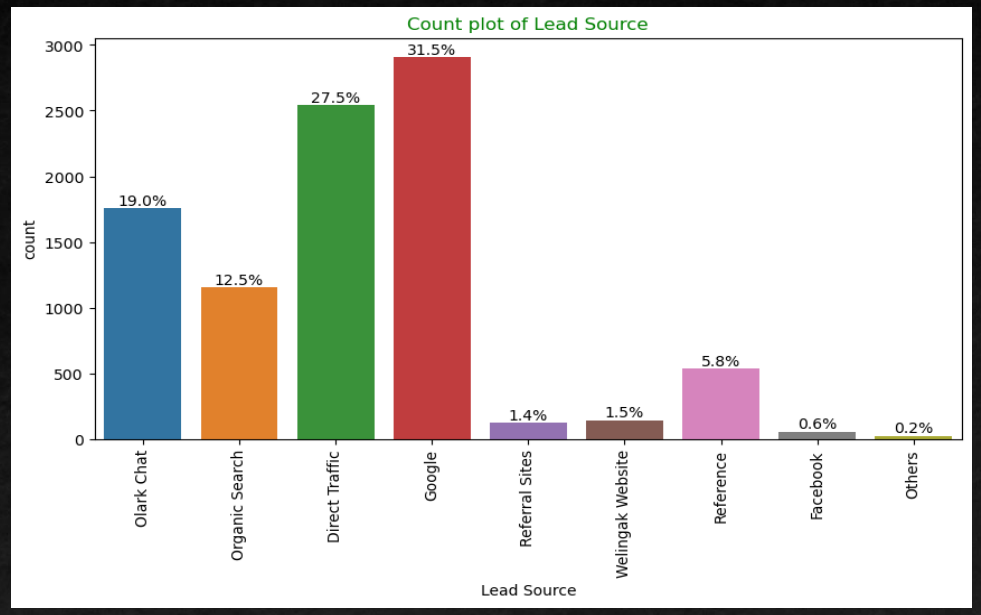
Data imbalance in target variable analysis:

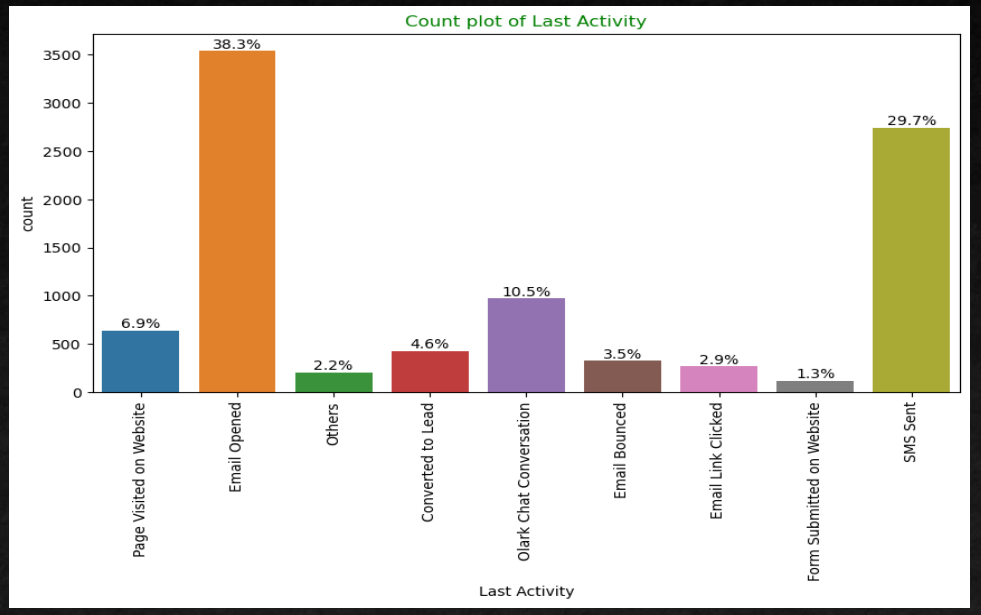


- Conversion rate is 38.5%, meaning only 38.5% of the leads are buying customers.
- 61.5% of the leads didn’t buy anything.

Exploratory Data Analysis (EDA)

Univariate analysis of categorical variables



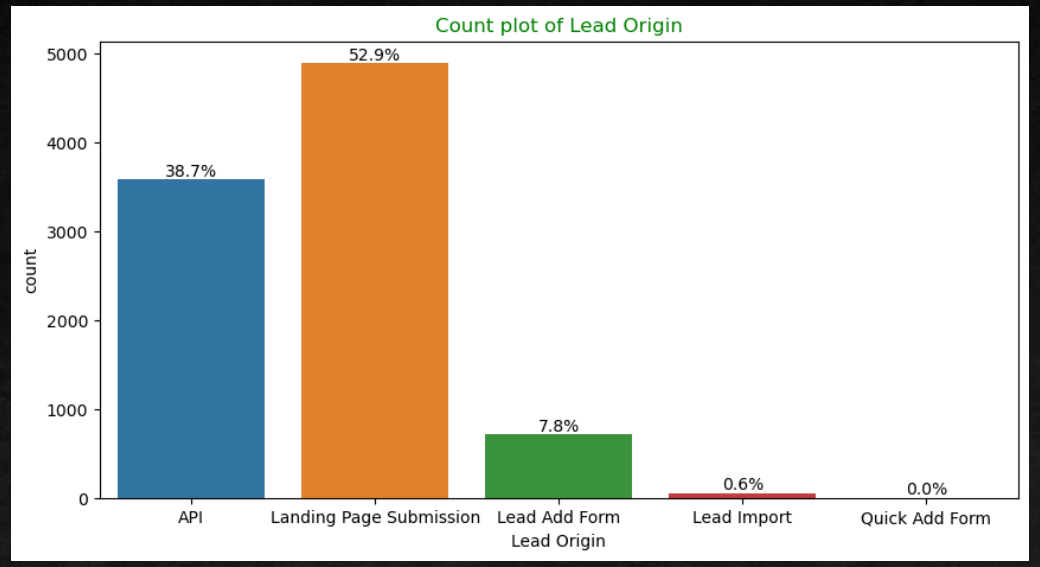


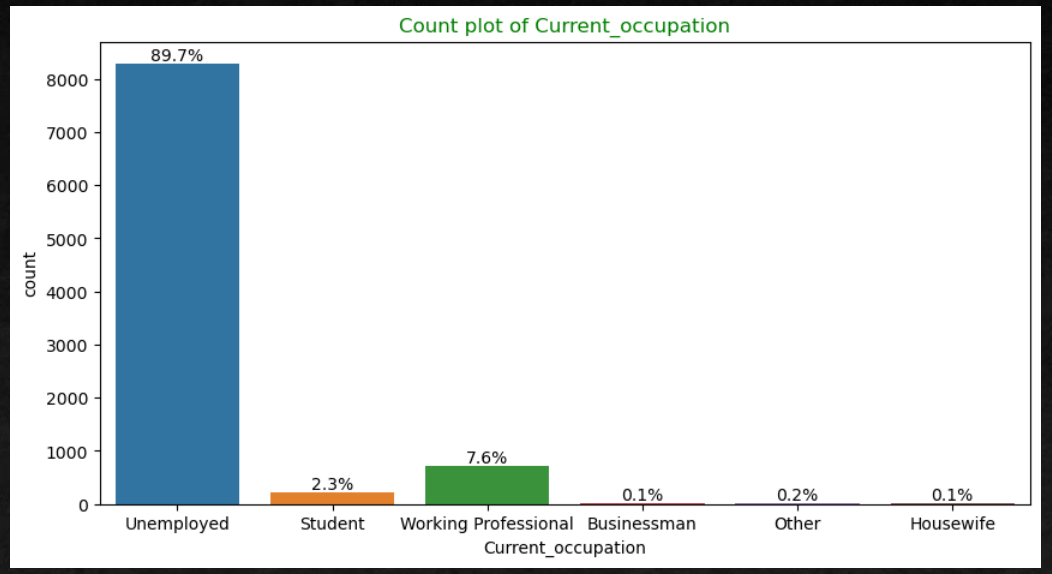
- **Lead Source:** 31.5% of the leads are from Google, followed by 27.5% leads from direct traffic.

- **Last Activity:** “Email Opened” and “SMS” sent are the major last activity of the leads at 38.3% and 29.7% respectively.

Exploratory Data Analysis (EDA)

Univariate analysis of categorical variables

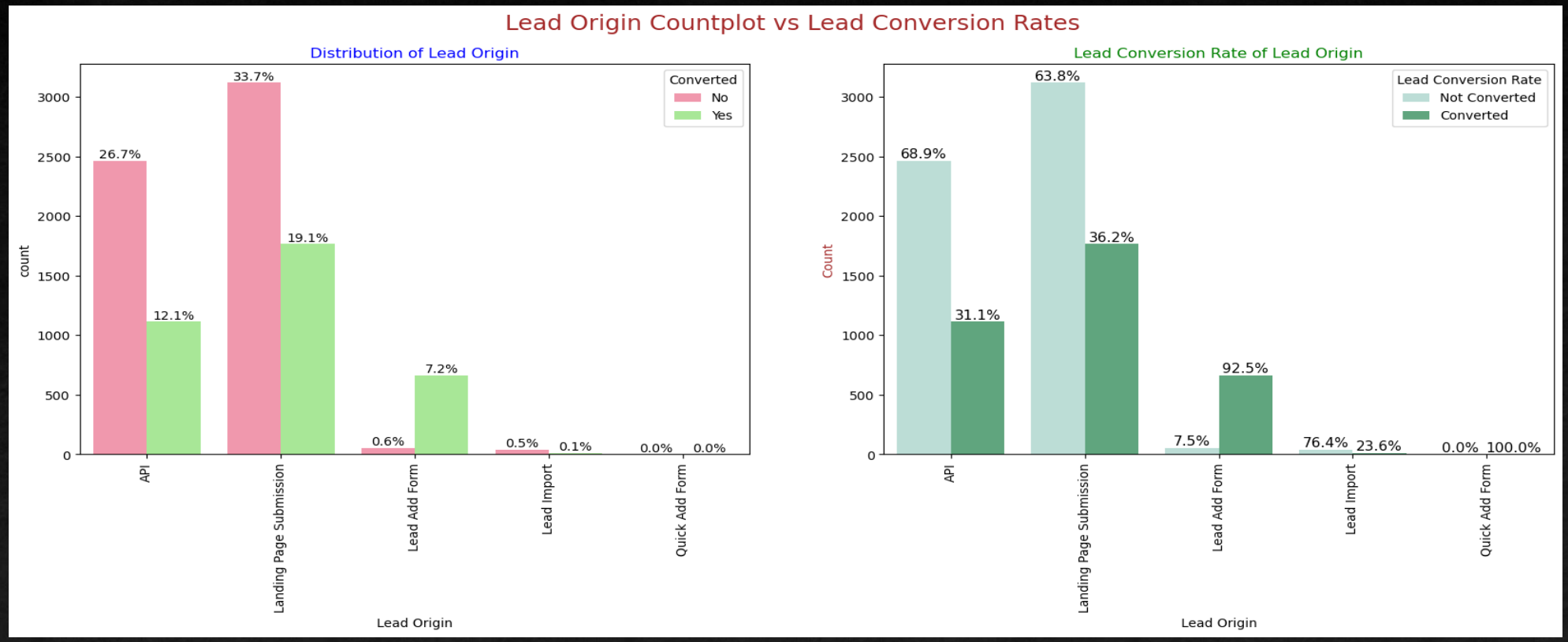




- **Lead Origin:** Majority of the leads come from “Landing Page Submission” with 53% of customers, followed by “API” and “Lead Add Form” with 38.7% and 7.8% respectively.

- **Current_occupation:** 90% of the leads are unemployed, and only 7.6% are working professionals.

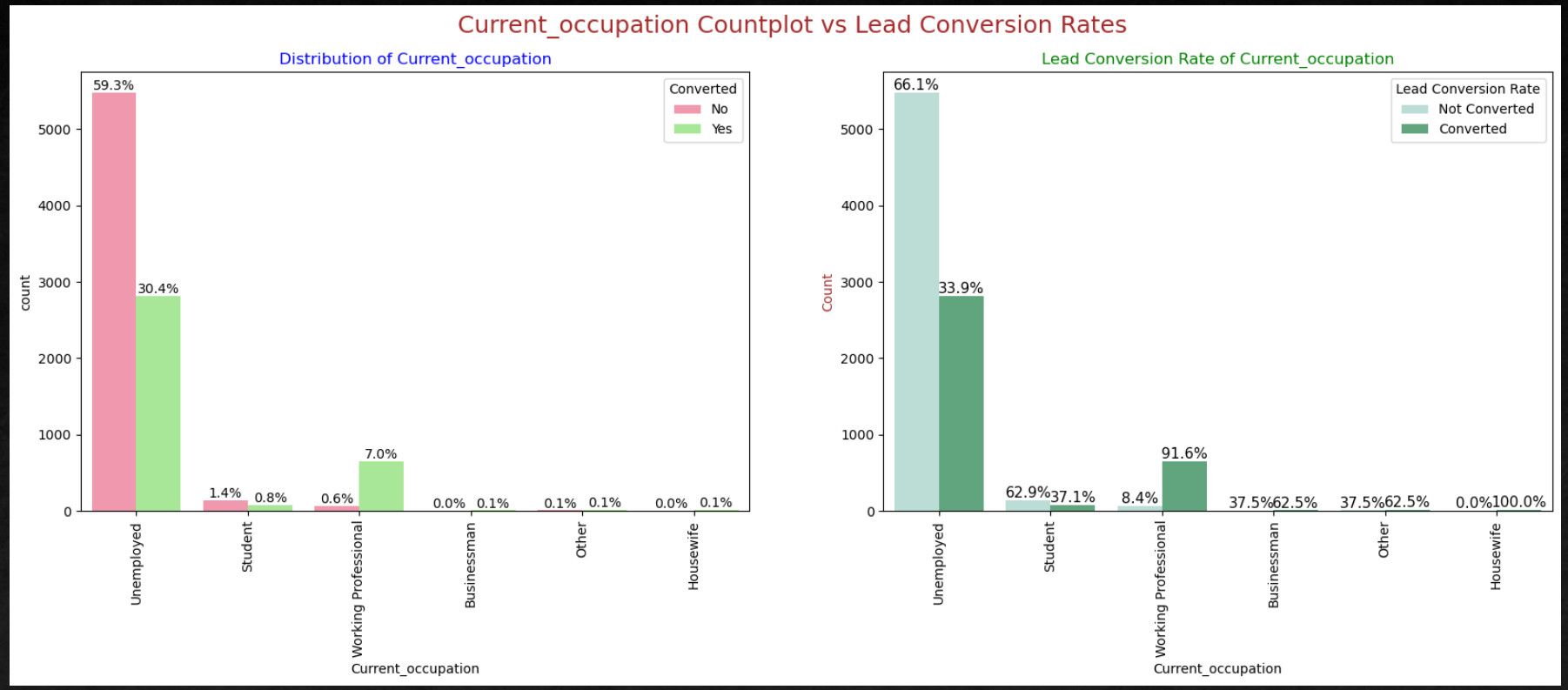
EDA – Bivariate analysis



Lead Origin:

- 52.8% of all leads come from “Landing Page Submission”, and have a Lead conversion rate (LCR) of 36.2%
- 38.8% of all leads come from “API”, and have an LCR of 31.1%

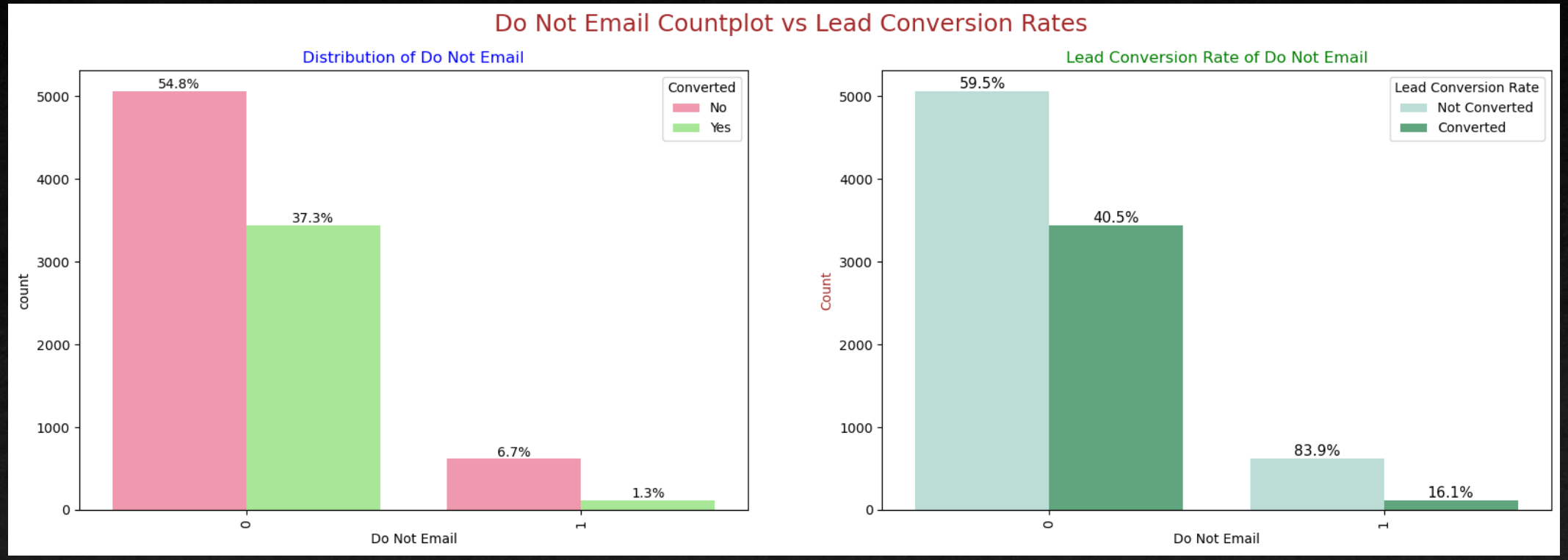
EDA – Bivariate analysis



Current Occupation:

- 89.7% of the leads are unemployed, and they have a lead conversion rate (LCR) of only 33.9%
- Only 7.6% of the leads are working professionals, but they contribute to 91.6% of the conversions

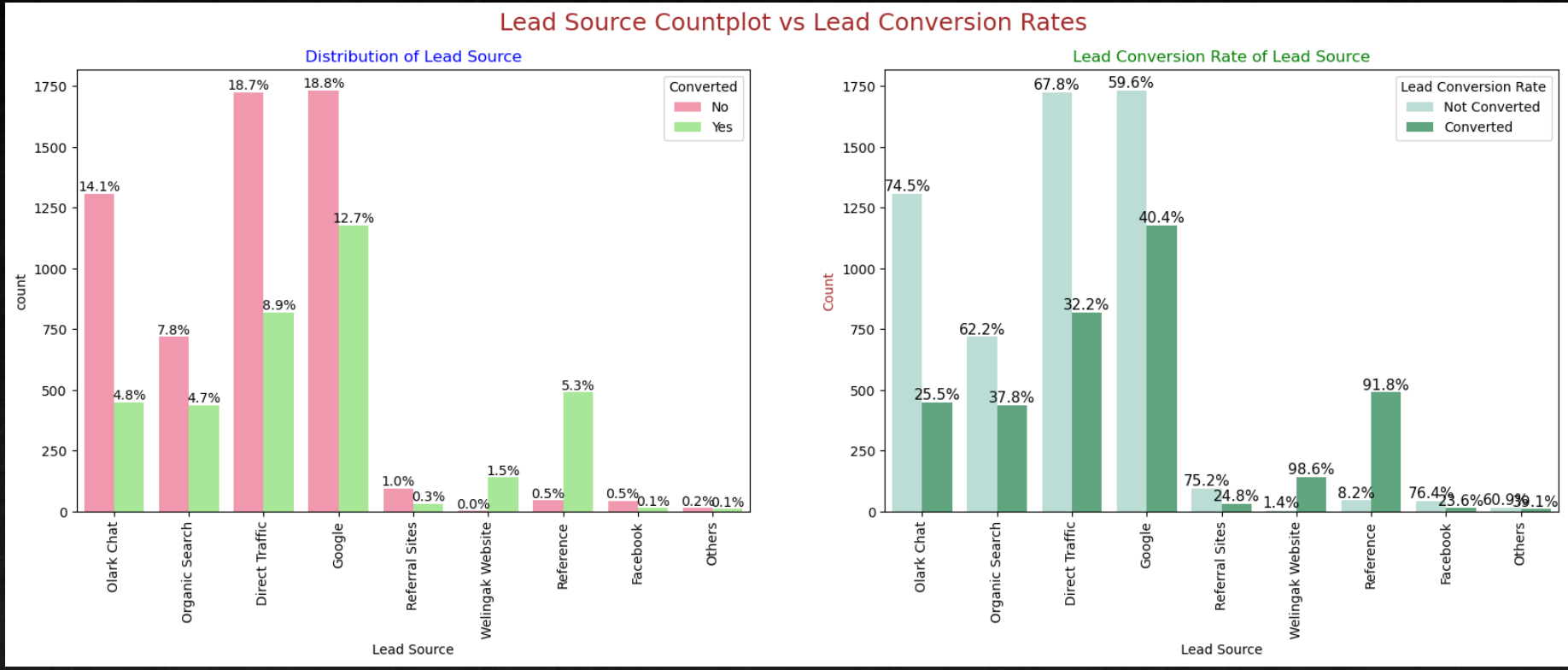
EDA – Bivariate analysis



Do not Email:

- 91.9% of the leads did not select the option “Do not email”, and their LCR stands out to be 40.5%
- This goes to show that engaging the leads through emails is an effective way to nurture them

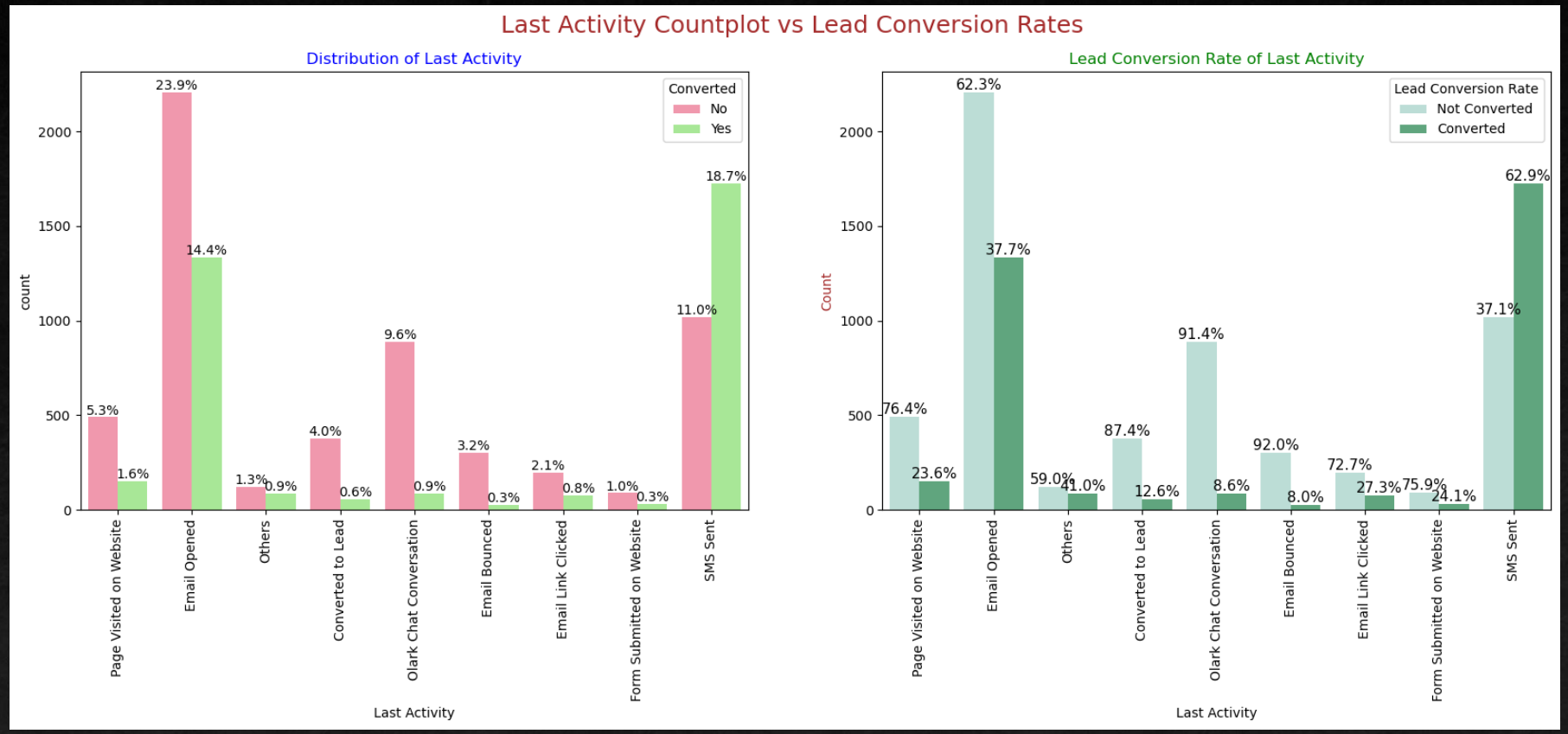
EDA – Bivariate analysis



Lead Source:

- 31.5% of the leads are sourced from Google out of which we obtain a 40.4% LCR
- Direct traffic accounts for 27.6% of the leads and yields an LCR of 32.2%
- References contribute to only 5.8% of all leads, yet 91.8% of them convert

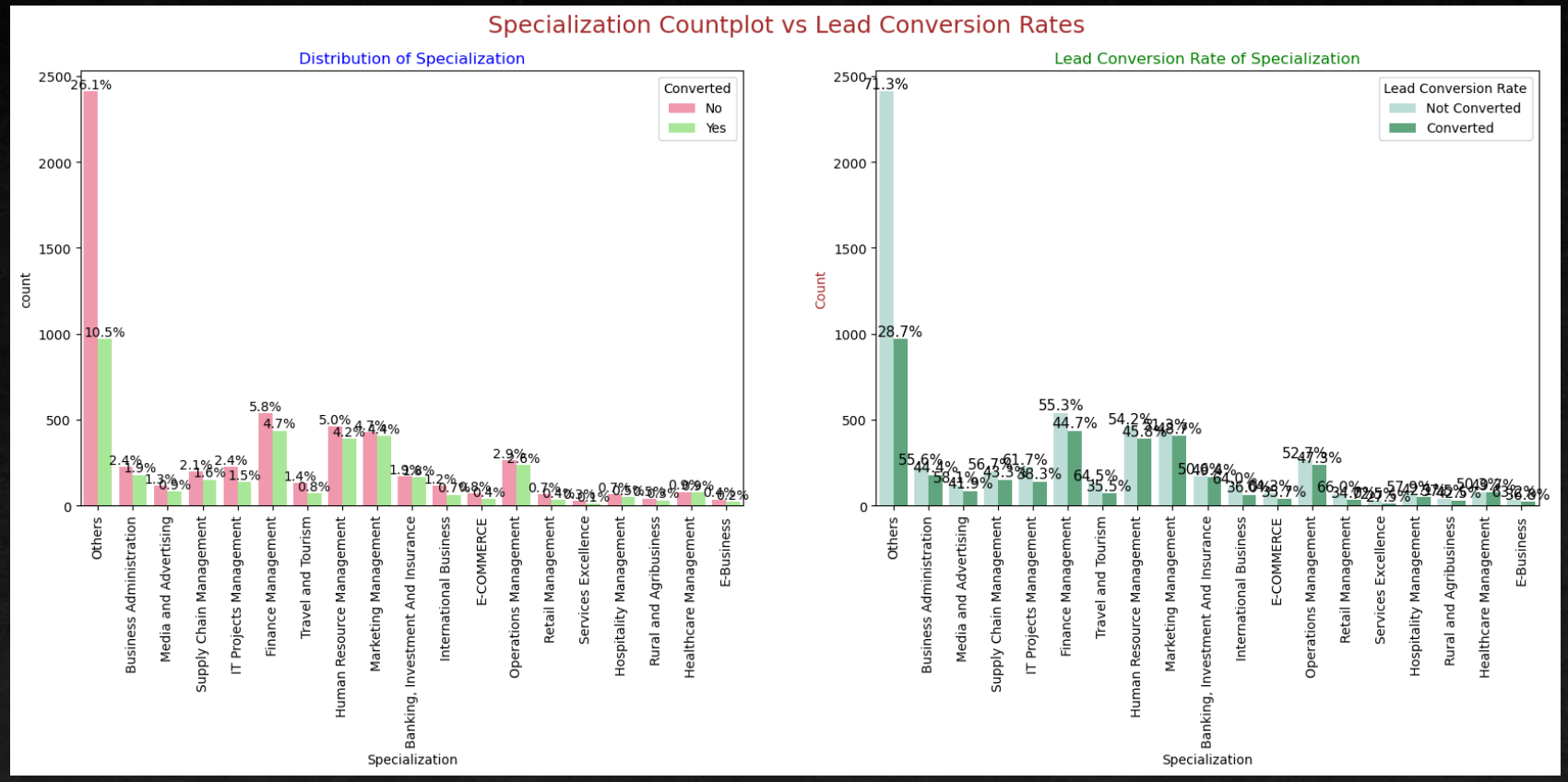
EDA – Bivariate analysis



Last Activity:

- "Email opened" is the most common last activity (38.3%), in which 37.7% leads are converted
- "SMS sent" is the second most common one (29.7%), that yields an LCR of 62.9%

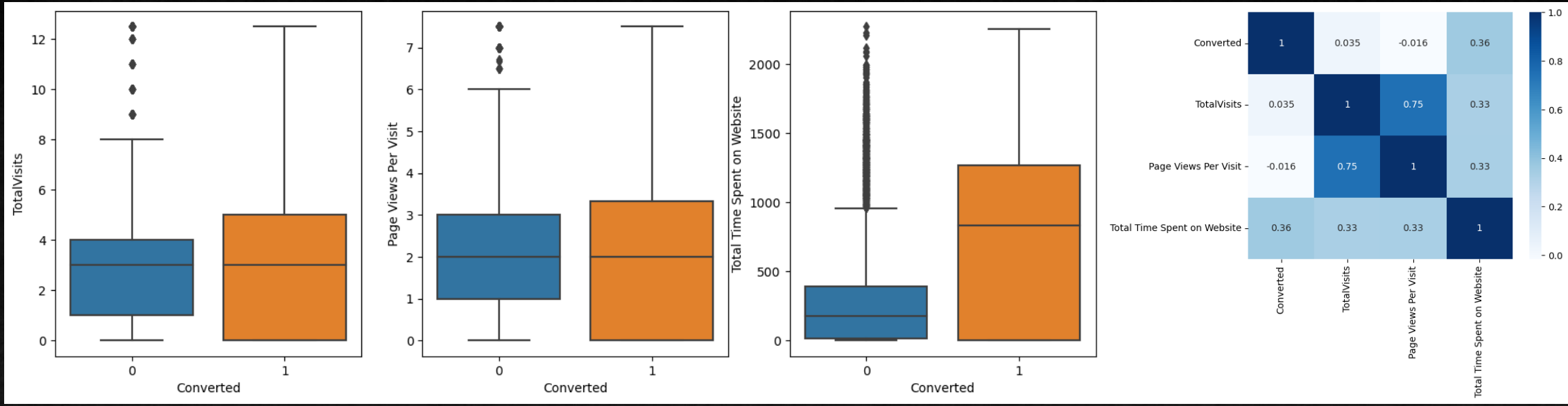
EDA – Bivariate analysis



Specialization:

- Marketing, Human Resources, and Finance Management contribute to the highest lead pool and conversion rate after “Other” specializations

EDA – Bivariate analysis (Numerical variables)



- Based on the data, it can be inferred that leads that spend more time on the website are more likely to convert than those who spend less time on the same.

Preparing the data

- Categorical binary variables have been mapped to 1s and 0s earlier
- Dummy features were created for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, and Current_occupation
- Train and Test set split: The ratio of 70:30 was chosen for the split
- Feature scaling: The features were scaled using the standardization method
- Checking the correlations:

  → Highly correlated predictor variables were dropped

Building the model

Selecting the feature:

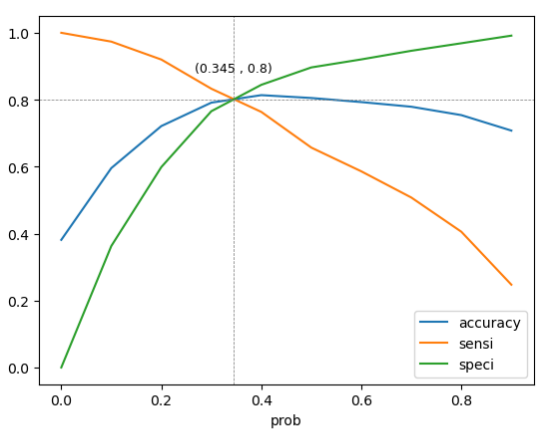
- The data set has many dimensions with lots of features.
- They may reduce model performance and take a lot of time to compute
- Hence, we must perform Recursice Feature Elimination (RFE) and select only the important columns.
- Outcome of the RFE:
  - → Pre RFE – 48 columns
  - → Post RFE – 15 columns

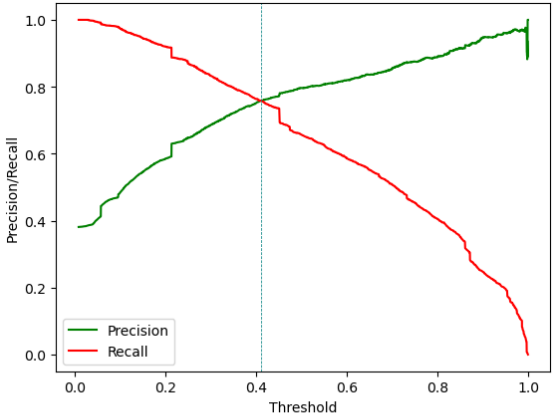
Building the model

- We used manual feature reduction and built models by dropping variables with p-value greater than 0.05.
- After 4 iterations model 4 looks the most suitable for our study
- Hence, that will be our final model

Evaluating the model (Train Data Set)

- A cutoff of 0.345 was chosen after checking evaluation metrics from both plots.
- Image on the left: Confusion Matrix and Evaluation Metrics with 0.345 as cutoff
- Image on the right: Confusion matrix and Evaluation Metrics with 0.41 as cutoff





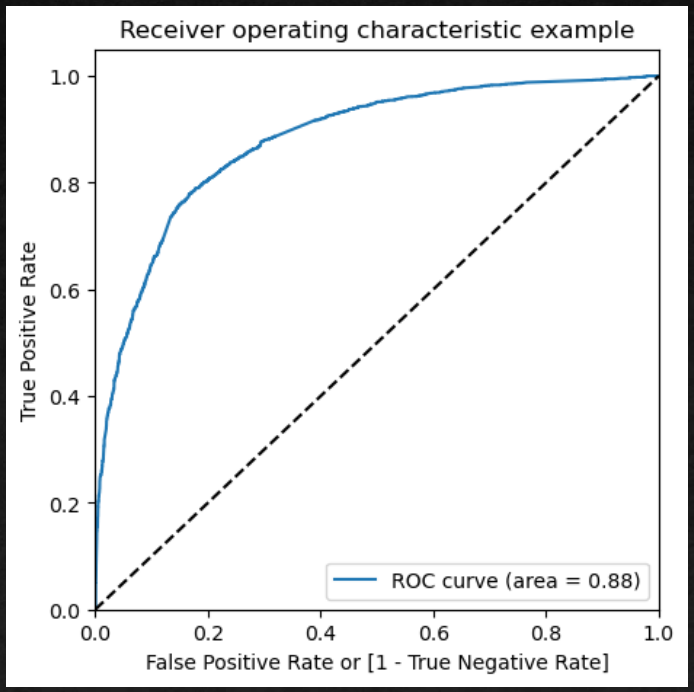

```
**********************************************

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
[[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

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

Evaluating the model (Train Data Set)

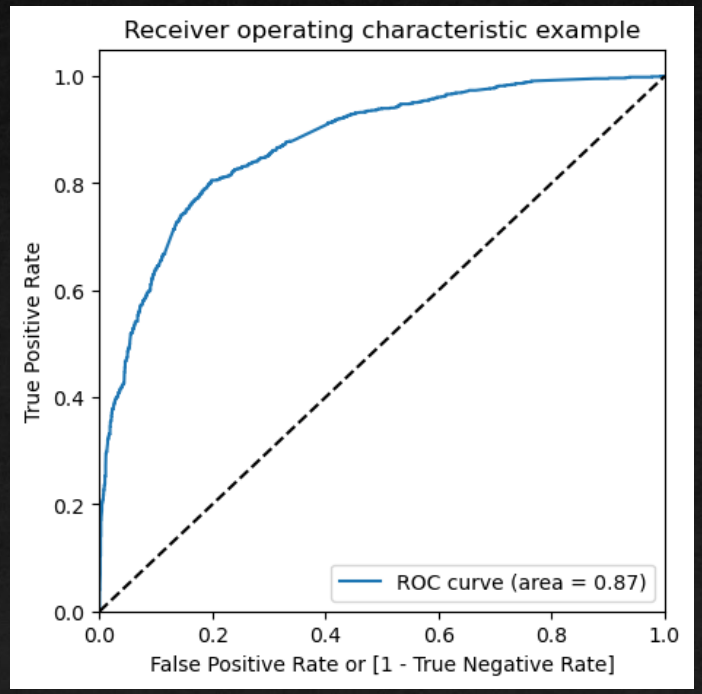
ROC Curve – Train Data Set

- ROC covers an area of 0.88, which is an indicator of a good prediction
- The curve is closest to the top left corner, which shows a high true positive and a low false positive rate



ROC Curve – Test Data Set

- ROC covers an area of 0.87, which is an indicator of a good prediction
- The curve is closest to the top left corner, which shows a high true positive and a low false positive rate



Evaluating the model (Confusion matrix and metrics)

- **To the left:** Confusion matrix and metrics of Train Data Set
- **To the right:** Confusion matrix and metrics of Test Data Set
- With cut-off value at 0.345, the model could achieve a sensitivity of 80.05% and 79.82% in the train and test set, respectively.
- The sensitivity indicates the accuracy of the prediction made by the model of the potential leads as to whether they would convert or not.
- The model has successfully achieved an accuracy of 80.46%, which is just above the target set by the CEO of X education company.

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

Confusion Matrix                                            Confusion Matrix
[[3230  772]                                                [[1353  324]
 [ 492 1974]]                                                [ 221  874]]

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

True Negative                   :  3230                     True Negative                   :  1353
True Positive                   :  1974                     True Positive                   :  874
False Negative                  :  492                      False Negative                  :  221
False Positve                   :  772                      False Positve                   :  324
Model Accuracy                  :  0.8046                   Model Accuracy                  :  0.8034
Model Sensitivity               :  0.8005                   Model Sensitivity               :  0.7982
Model Specificity               :  0.8071                   Model Specificity               :  0.8068
Model Precision                 :  0.7189                   Model Precision                 :  0.7295
Model Recall                    :  0.8005                   Model Recall                    :  0.7982
Model True Positive Rate (TPR)  :  0.8005                   Model True Positive Rate (TPR)  :  0.7982
Model False Positive Rate (FPR) :  0.1929                   Model False Positive Rate (FPR) :  0.1932

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

Recommendations

- It is evident that increasing lead conversion is crucial to ensure optimized growth rate of X Education Company. To help achieve this, we have curated a regression model to help identify the determinant factors that impact lead conversion
- This helps us ensure that we prioritize the optimization of the right factors and save time by ignoring the wrong ones.
- In certain cases, more leads pertain to a common factor, but since their conversion rate is low, we need to overlook them and focus on quality more than quantity.
- The following features have the highest positive coefficients, and must be prioritized more through marketing and sales efforts to increase positive conversions:
  - Lead Source_Welingak Website: 5.39
  - Lead Source_Reference: 2.39
  - 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
- The following features have negative coefficients and need to be either improved or completely omitted to save more time:
  - Specialization in Hospitality Management: -1.09
  - Specialization in Others: -1.20
  - Lead Origin Landing Page Submission: -1.26

Recommendations

To increase Lead Conversion Rates

- Allocate more budget for advertisement expenditure on Welingak website.
- Come up with a referral program for existing customers and incentivize them
- Optimize the product to suit the requirements of working professionals. Engage them through tailored messaging
- Optimize communication channels such as SMS and email.
- Make sure customers spend more time on the website. This can be done by optimizing the UI/UX
- Leverage email marketing

To identify areas of improvement:

- Optimize content for specializations because not many leads are converting in specific specialization programs.
- Use Olark Chat only as the initial touch point. After that, shift the primary communication channel to SMS or WhatsApp (since it is easier to send multimedia messages, and make customers see the company logo regularly).

Thank You