LEAD SCORING
X-EDUCATION

Vandit Sardana
Saras Sangle
Santosh Govardhan

Table of Contents

- Background of X Education Company
- Problem Statement & Objective of the Study
- Suggested Ideas for Lead Conversion
- Analysis Approach
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Data Preparation
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Recommendations

Background of X-Education

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- 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.
- Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.

Problem Statement & Objective

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.

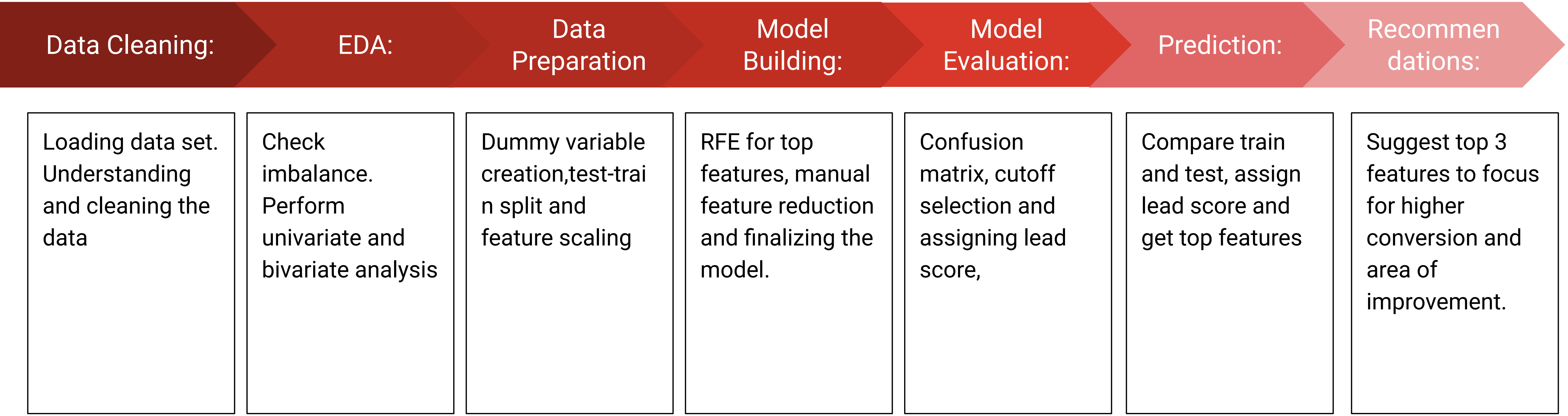
Objective:

- 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%.

Suggested Ideas for Lead Conversion

Leads Grouping	**Better Communication**	**Boost Conversion**
Leads are grouped based on their propensity or likelihood to convert. Grouped as Hot Leads, Warm Leads, and Cold Leads.	We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.	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.
This results in a focused group of hot leads.	Identify the mediums that provide maximum engagement with potential leads	

Analysis Approach



Data Cleaning: Loading data set. Understanding and cleaning the data

EDA: Check imbalance. Perform univariate and bivariate analysis

Data Preparation Dummy variable creation,test-train split and feature scaling

Model Building: RFE for top features, manual feature reduction and finalizing the model.

Model Evaluation: Confusion matrix, cutoff selection and assigning lead score,

Prediction: Compare train and test, assign lead score and get top features

Recommendations: Suggest top 3 features to focus for higher conversion and area of improvement.

Data Cleaning

1. Handling the "Select" level to represent null values in certain categorical variables, where customers had not made any selection from the list.

2. Eliminating columns with more than 40% null values.

3. Managing missing values in categorical columns based on value counts and specific considerations.

4. Removing columns that did not contribute to the study objective, such as "tags" and "country."

5. Employing imputation for some categorical variables.

6. Creating additional categories for certain variables.

7. Dropping columns not useful for modeling, like "Prospect ID" and "Lead Number," or those with only one response category.

8. Imputing numerical data using the mode after distribution checks.

9. Addressing skewed category columns and eliminating them to prevent bias in logistic regression models.

10. Treating outliers in "TotalVisits" and "Page Views Per Visit" by capping.

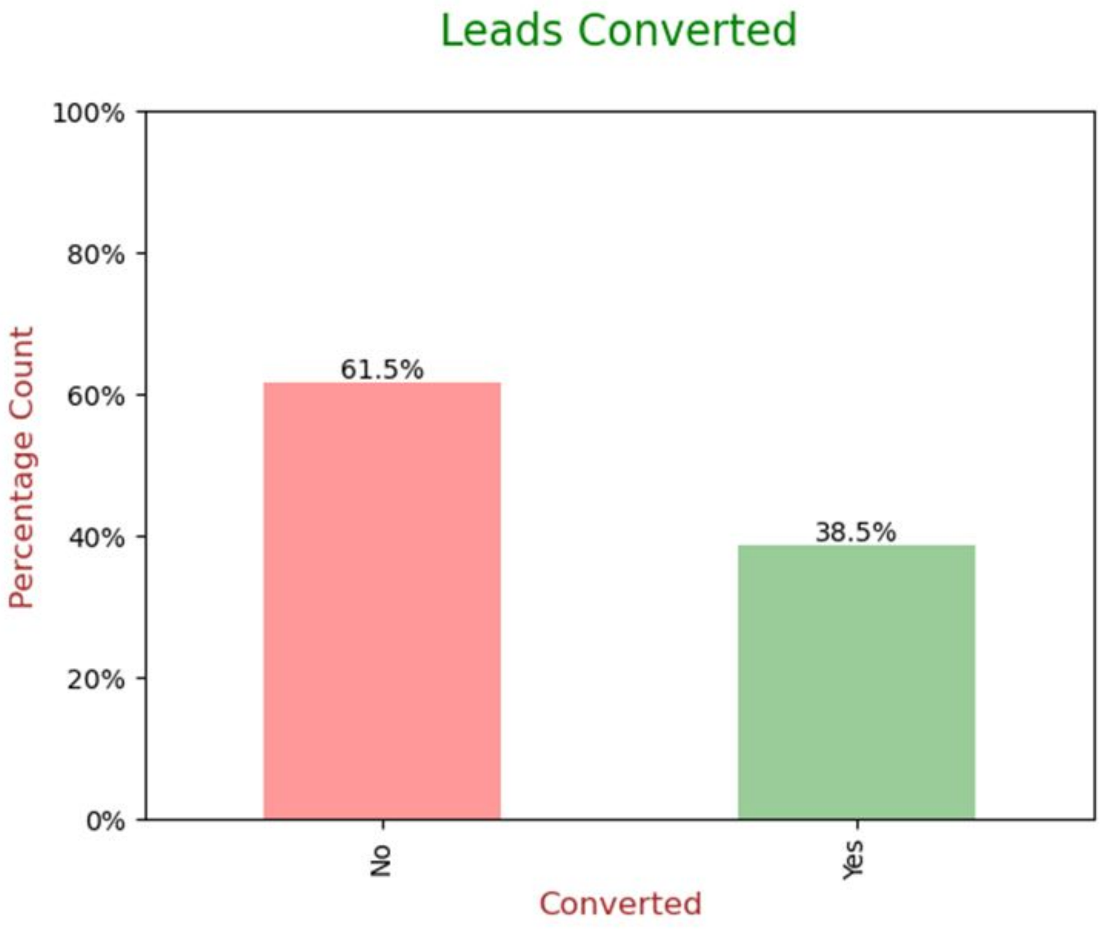
11. Correcting invalid values and standardizing data in some columns, e.g., "lead source" (e.g., "Google" and "google").

12. Grouping low-frequency values into an "Others" category.

13. Mapping binary categorical variables.

14. Conducting various other data cleaning activities to ensure data quality and accuracy.

EDA

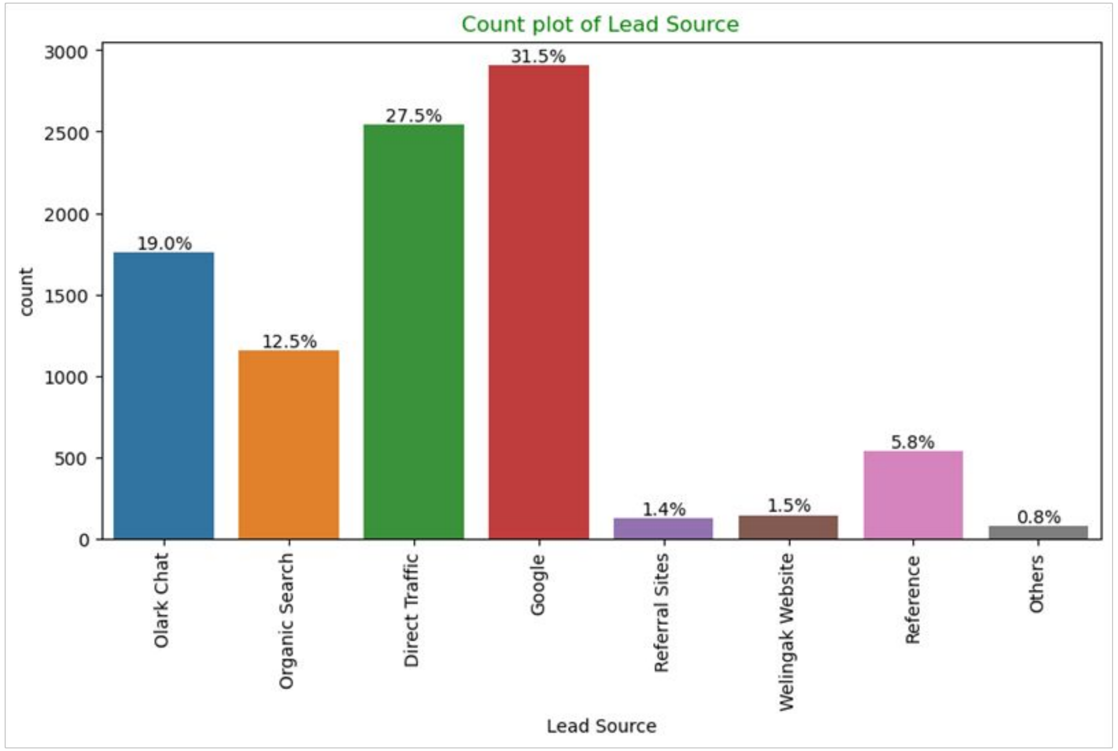


- Data is imbalanced for target variable
  - The conversion rate stands at 38.5%, indicating that a minority of people have become leads.
  - Conversely, the majority, comprising 61.5% of the individuals, did not convert into leads.

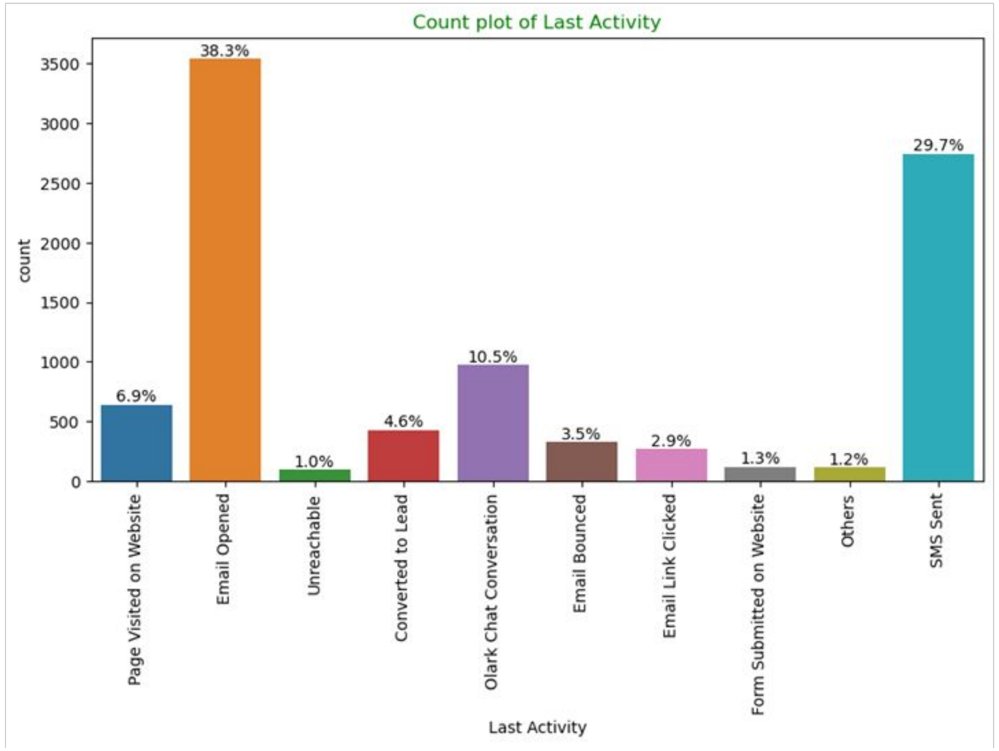
EDA

- Univariate Analysis – Categorical Variables

- Of the lead sources, 58% can be attributed to a combination of google and direct traffic.



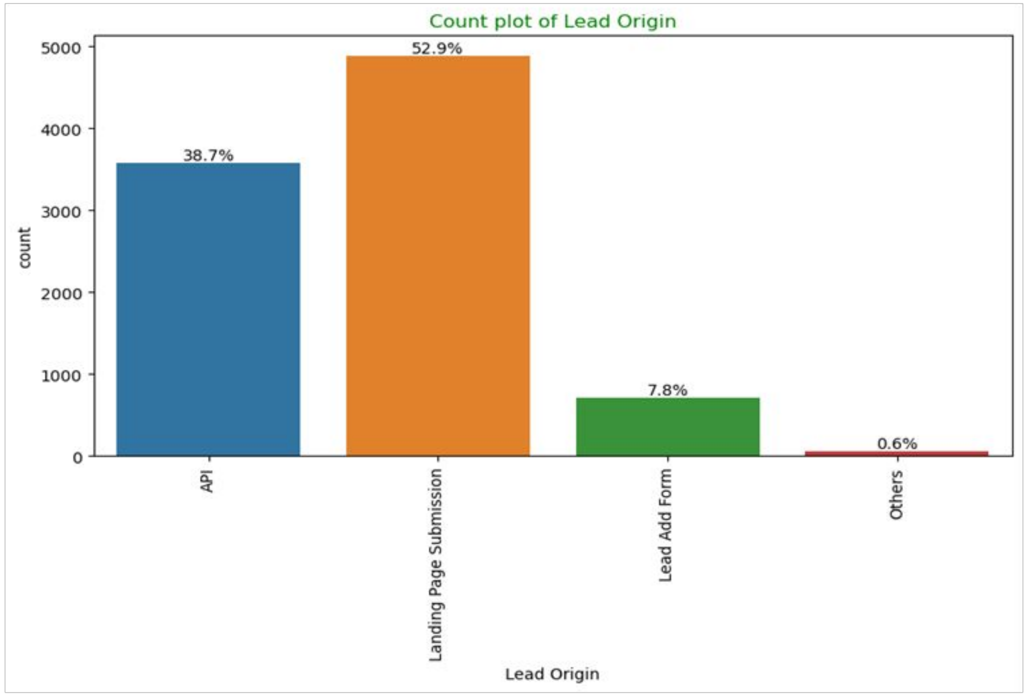
- In the realm of last activity, SMS sent and email opened activities make up 68% of customer contributions.



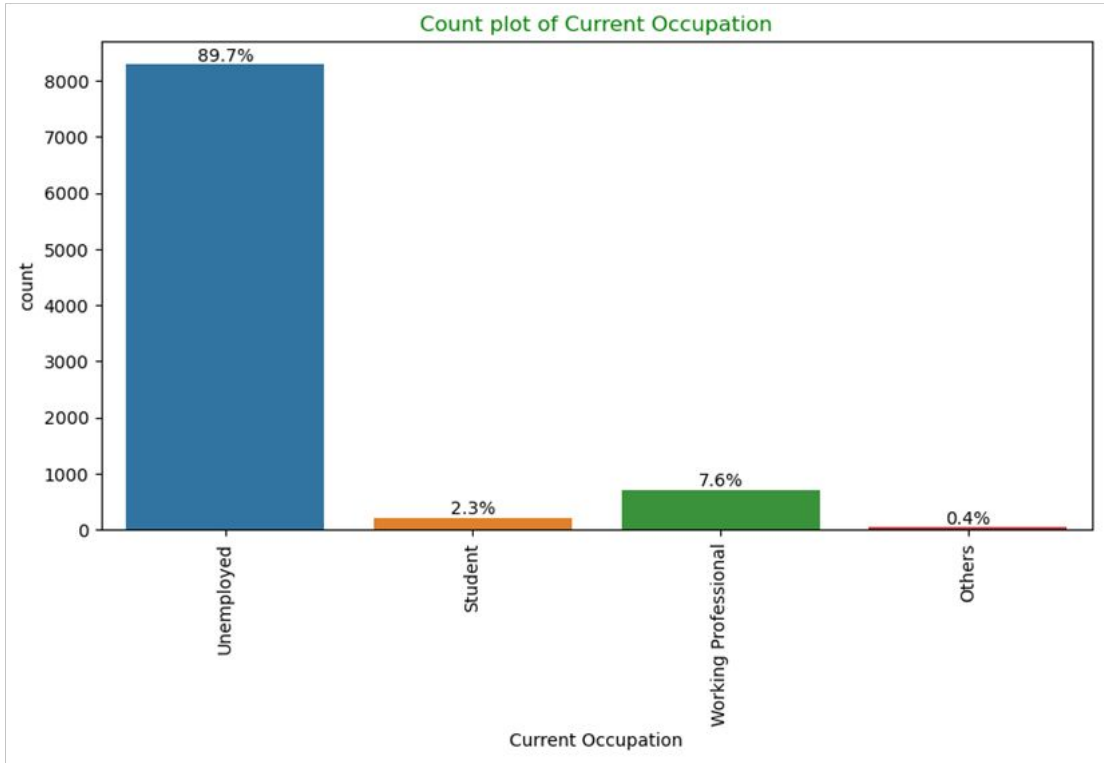
EDA

- Univariate Analysis – Categorical Variables

- Among customers, 53% were identified with "Landing Page Submission" as their lead origin, while 39% were associated with "API."



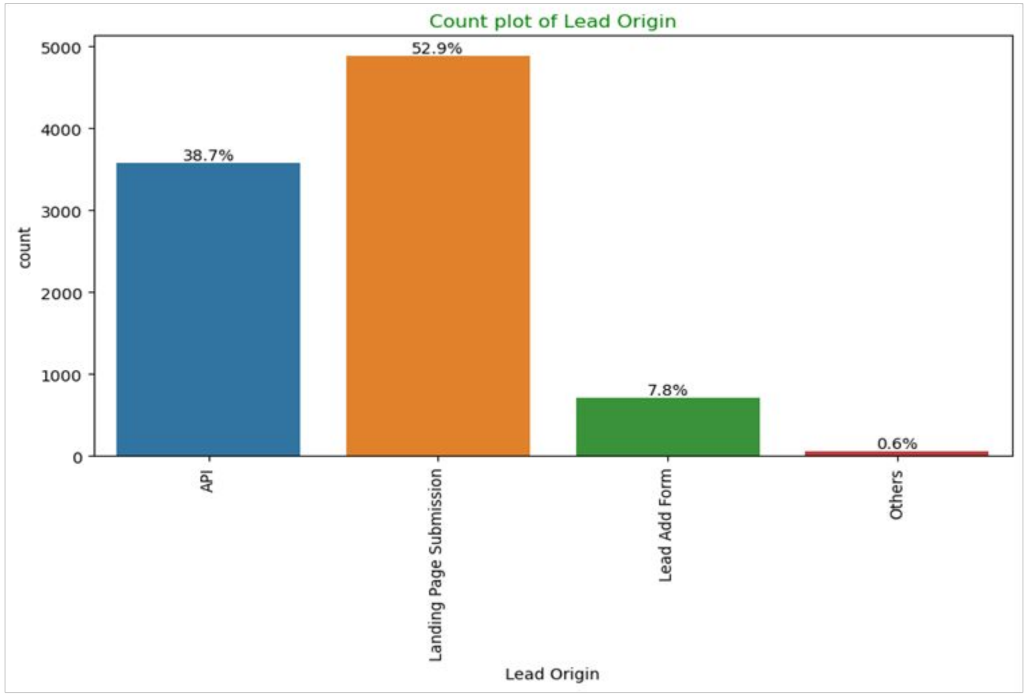
- The category "Unemployed" encompasses around 90% of the customer base in the current occupation field.



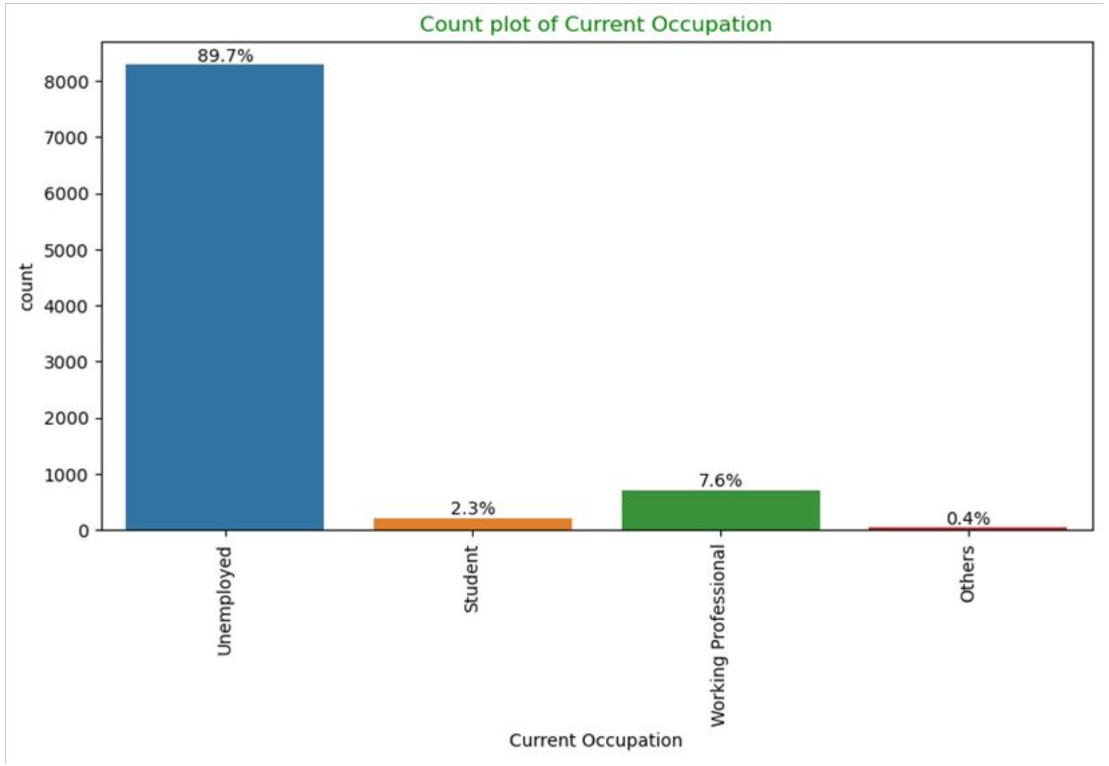
EDA

- Univariate Analysis – Categorical Variables

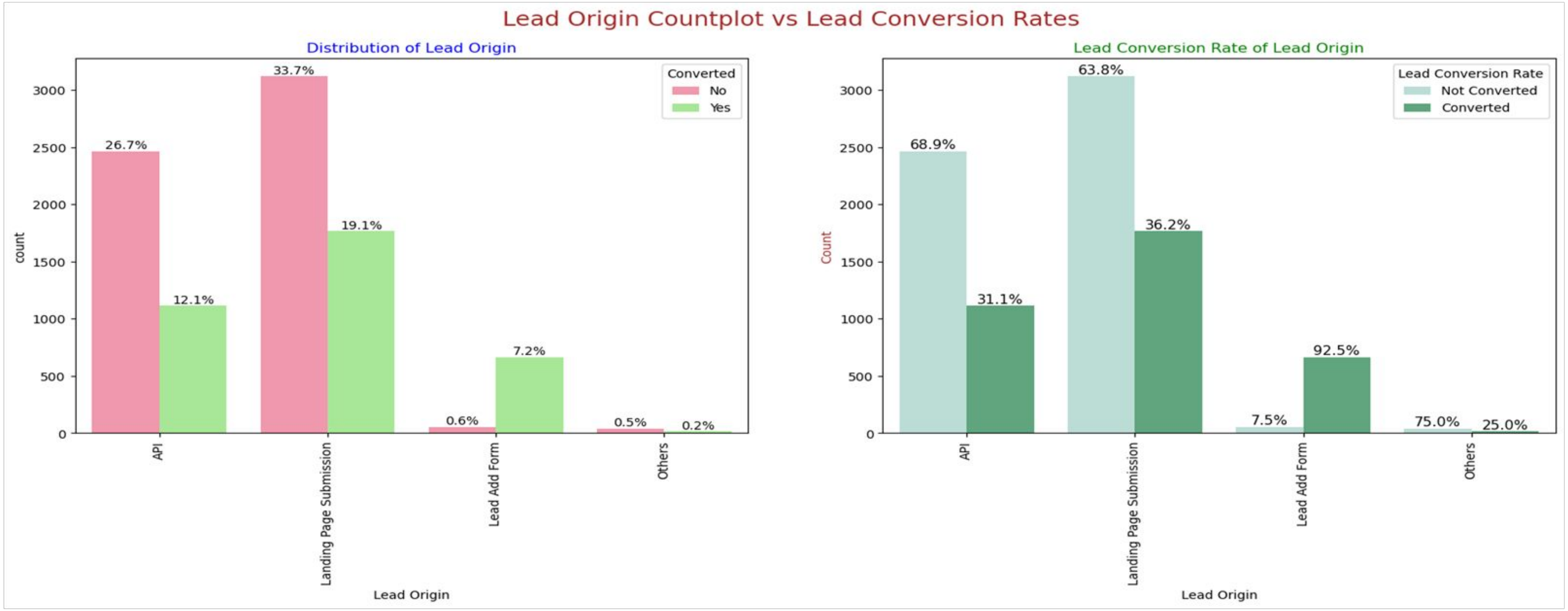
- Among customers, 53% were identified with "Landing Page Submission" as their lead origin, while 39% were associated with "API."



- The category "Unemployed" encompasses around 90% of the customer base in the current occupation field.



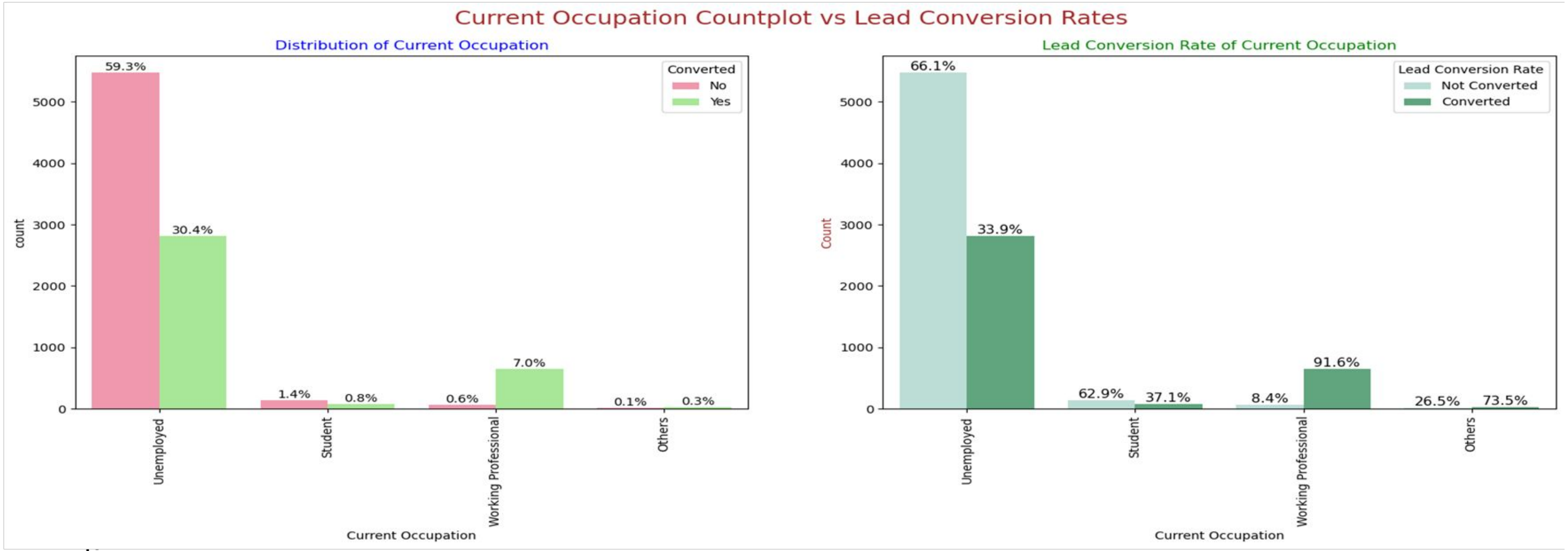
EDA – Bivariate Analysis for Categorical Variables



Lead Origins :

- "Landing Page Submission" accounted for roughly 52% of all leads, and it achieved a lead conversion rate (LCR) of 36%.
- Approximately 39% of customers were attributed to the "API," which yielded a lead conversion rate (LCR) of 31%.

EDA – Bivariate Analysis for Categorical Variables

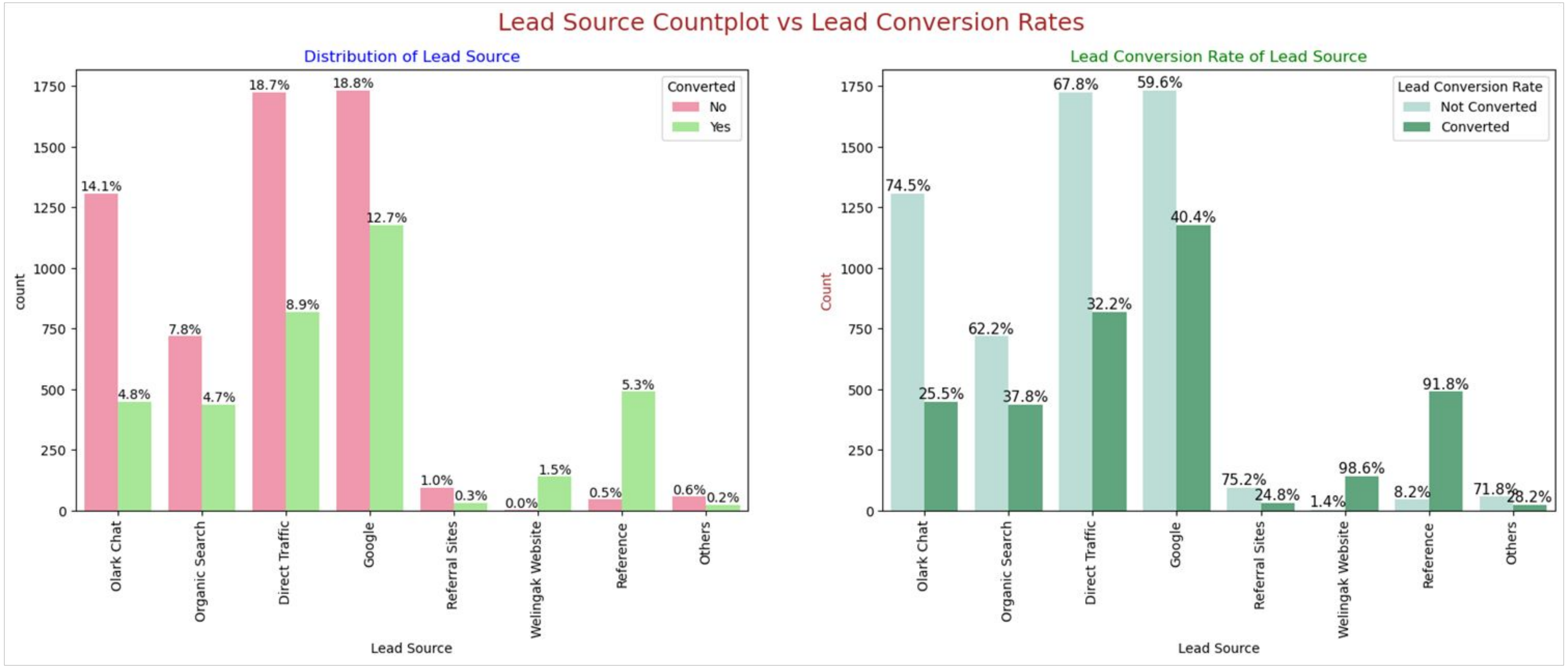


Current_occupation :

● Unemployed individuals make up approximately 90% of the customer base, and they exhibit a lead conversion rate (LCR) of 34%.

● Working Professionals, on the other hand, constitute just 7.6% of the total customer population, but they boast an impressive lead conversion rate (LCR) of nearly 92%.

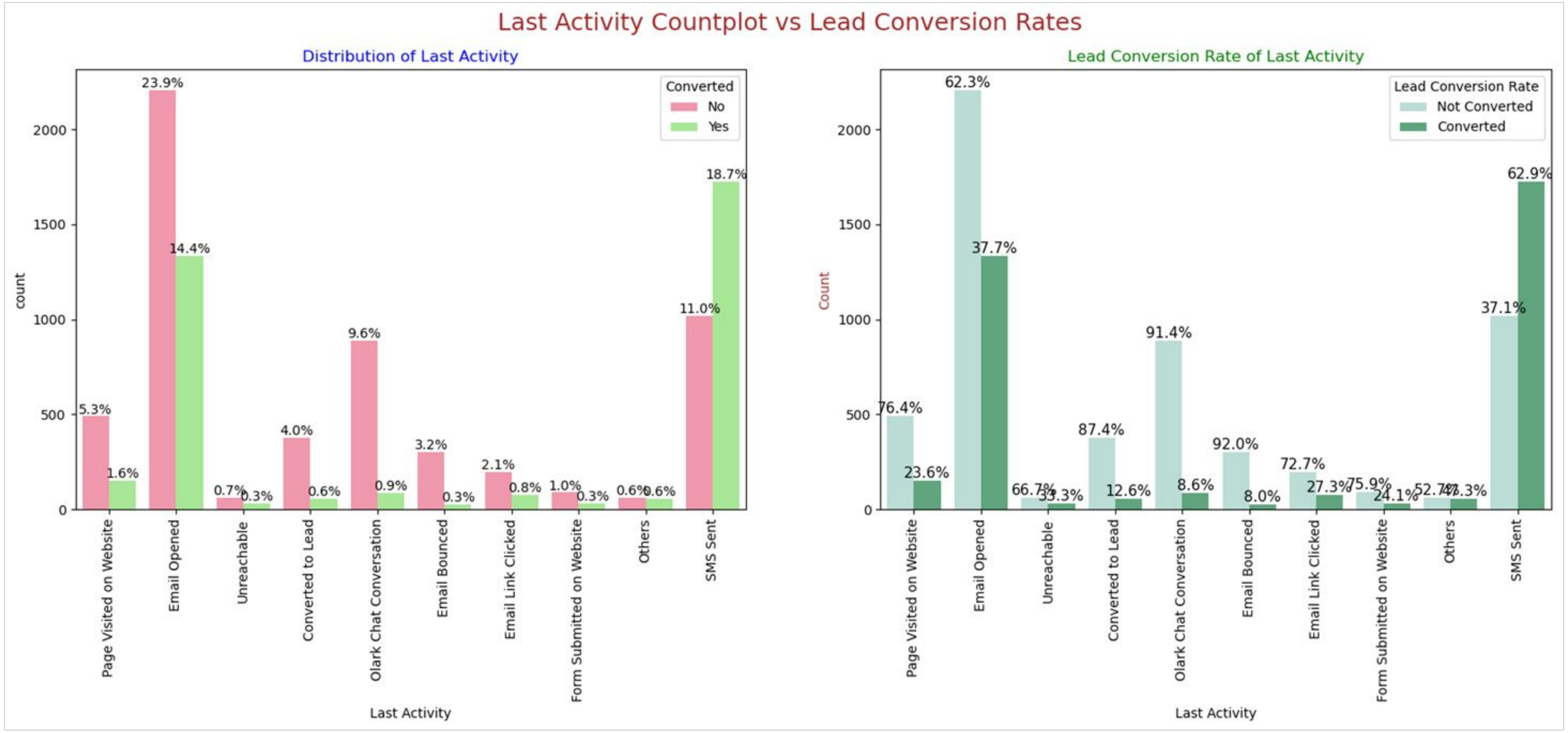
EDA – Bivariate Analysis for Categorical Variables



Lead Source:

- Google, with a 31% customer base, demonstrates a lead conversion rate (LCR) of 40%.
- Direct Traffic, comprising 27% of customers, yields a lower LCR of 32% compared to Google.
- Organic Search, while responsible for only 12.5% of customers, still achieves a noteworthy LCR of 37.8%.
- Reference, albeit representing just 6% of customers, boasts an exceptional LCR of 91%.

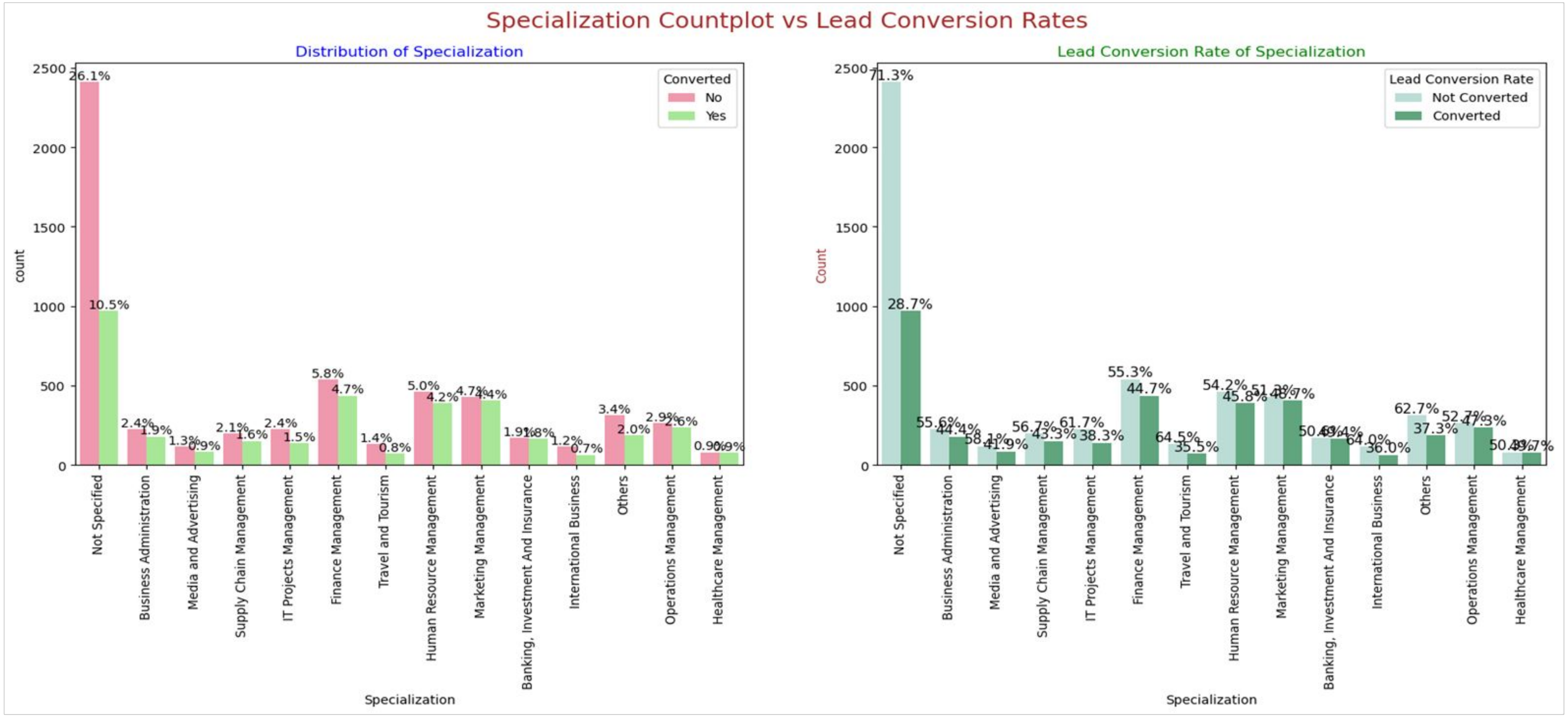
EDA – Bivariate Analysis for Categorical Variables



Last Activity :

- 'SMS Sent' exhibits a substantial lead conversion rate of 63%, and it accounts for 30% of the last activities conducted by customers. This shows that communication made through hand phone medium result in maximum leads.
- 'Email Opened' activity, which comprises 38% of the last activities performed by customers, achieves a commendable lead conversion rate of 37%.

EDA – Bivariate Analysis for Categorical Variables

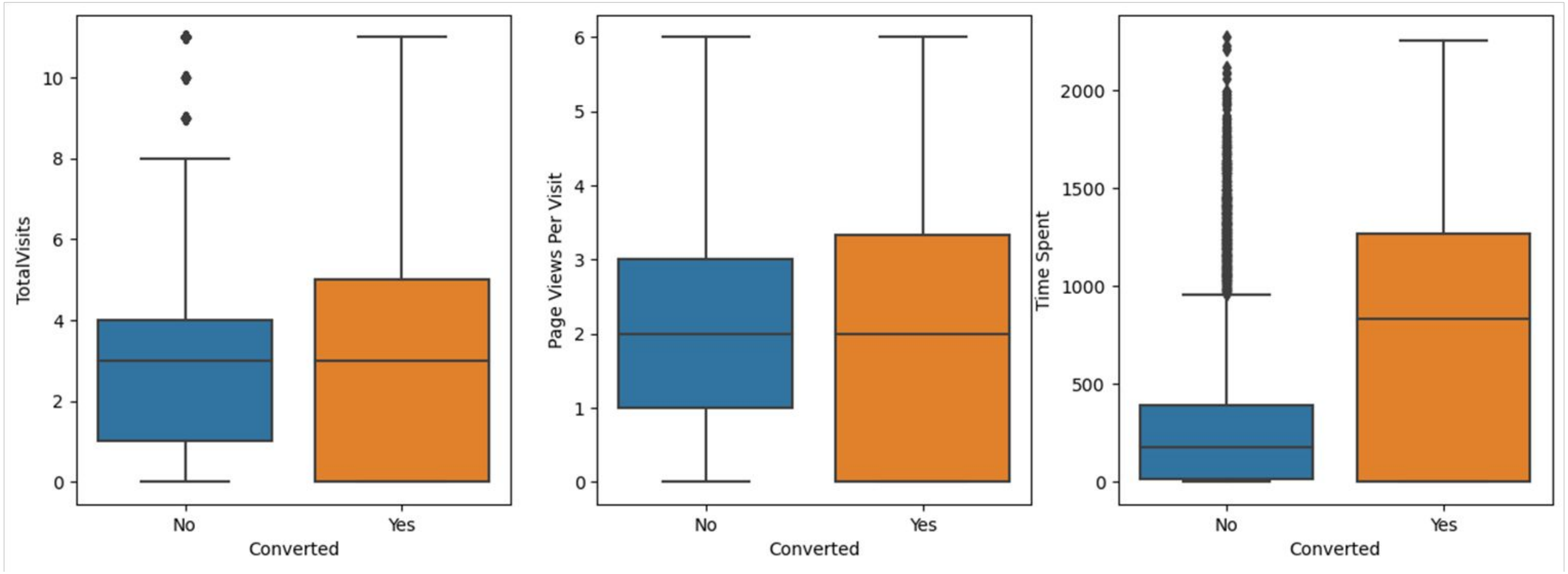


Specialization :

- Among various specializations, Marketing Management, HR Management, and Finance Management demonstrate a more substantial contribution to lead conversion compared to other specializations.

EDA – Bivariate Analysis for Numerical Variables

The box-plot analysis reveals that past leads who spend more time on the website have a greater likelihood of successful conversion compared to those who spend less time.



Data Preparation & 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).

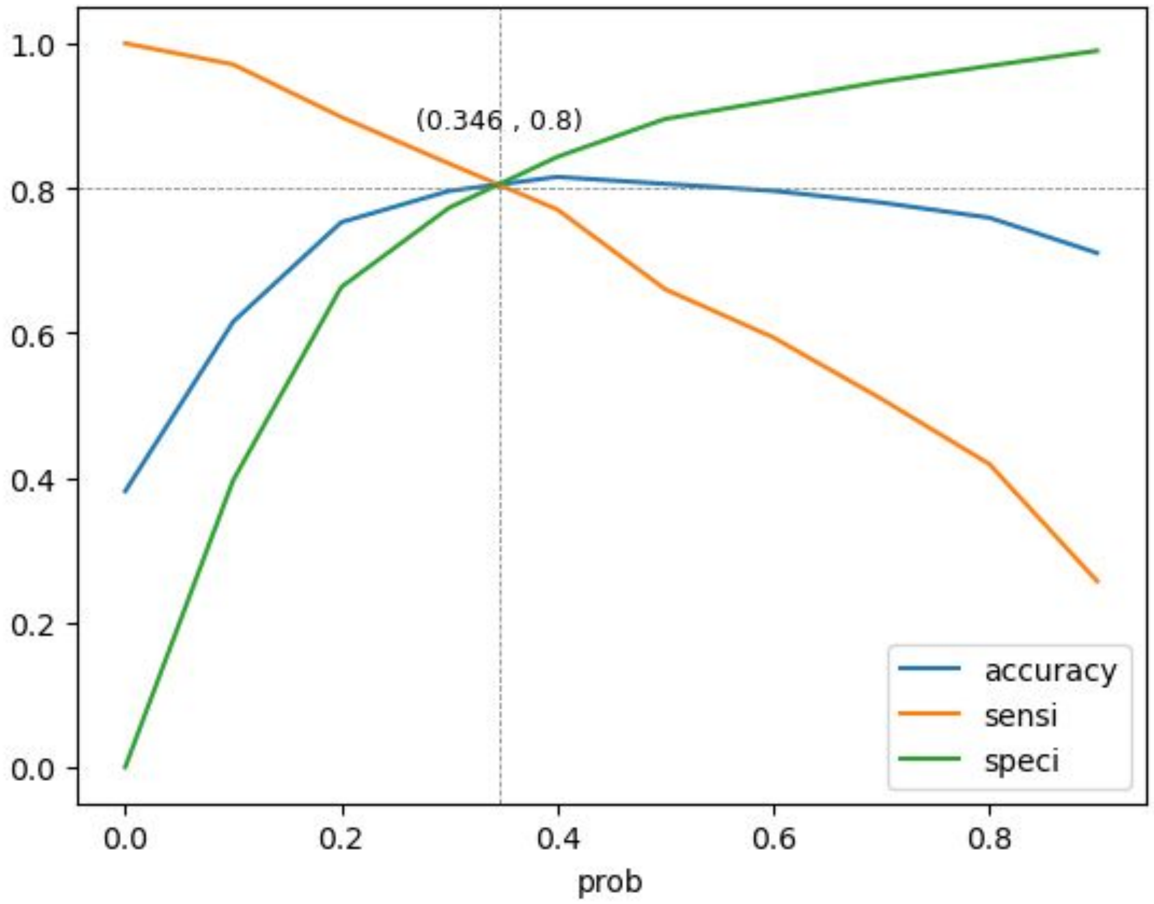
Model Building

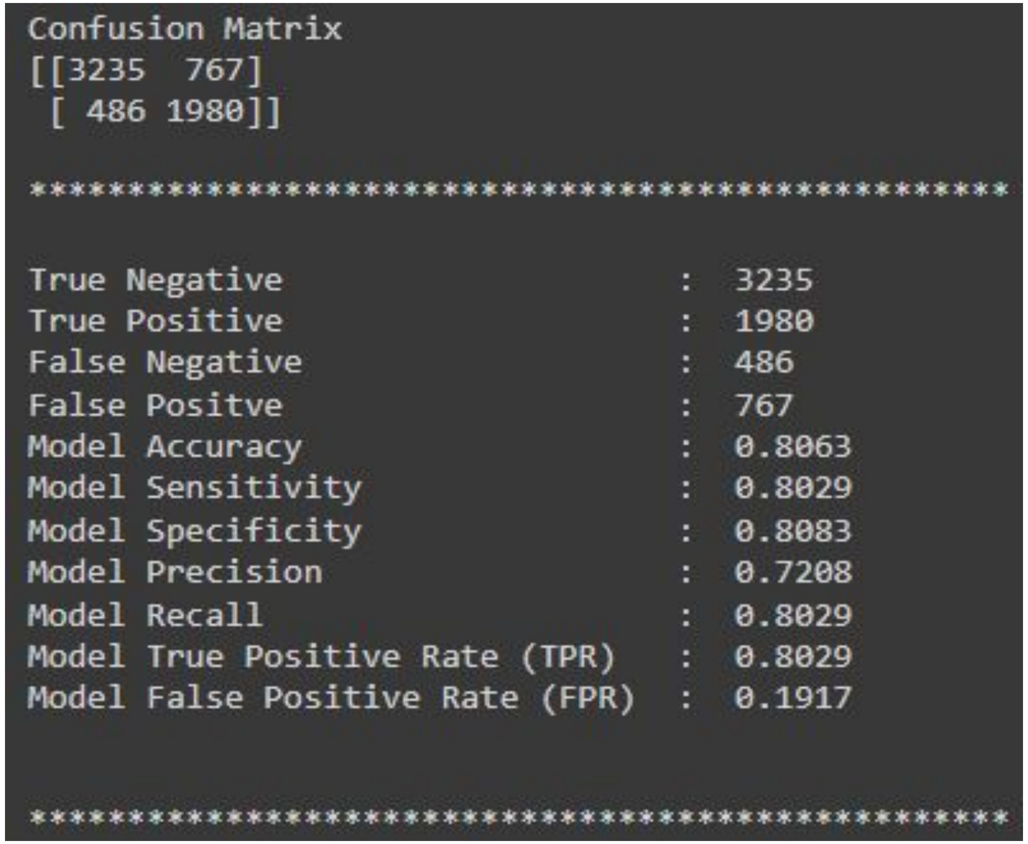
Feature Selection

- The data set has lots of dimension and large number of features
- This will reduce model performance and might take high computation time
- Hence it is important to 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
- significant p-values within the threshold (p-values < 0.05)
- 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 Building

Cut-off precision point - 1

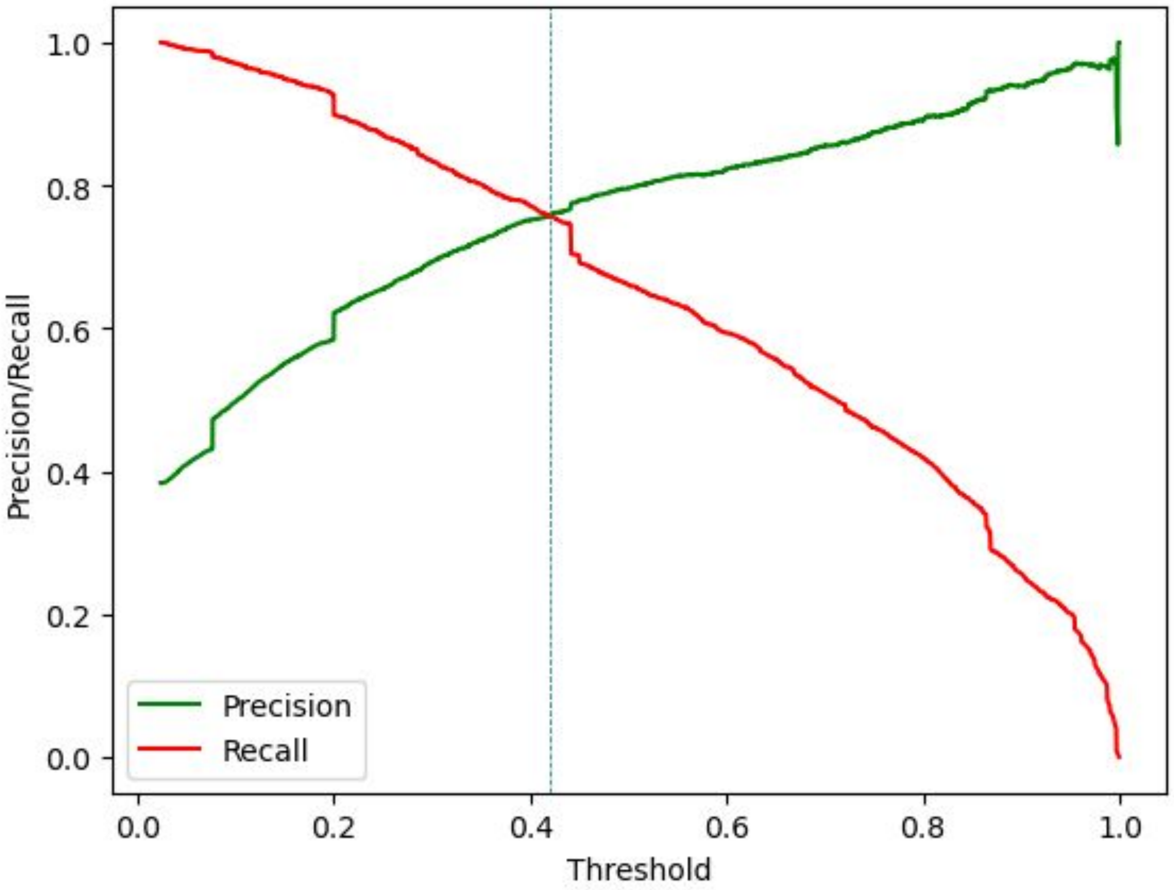


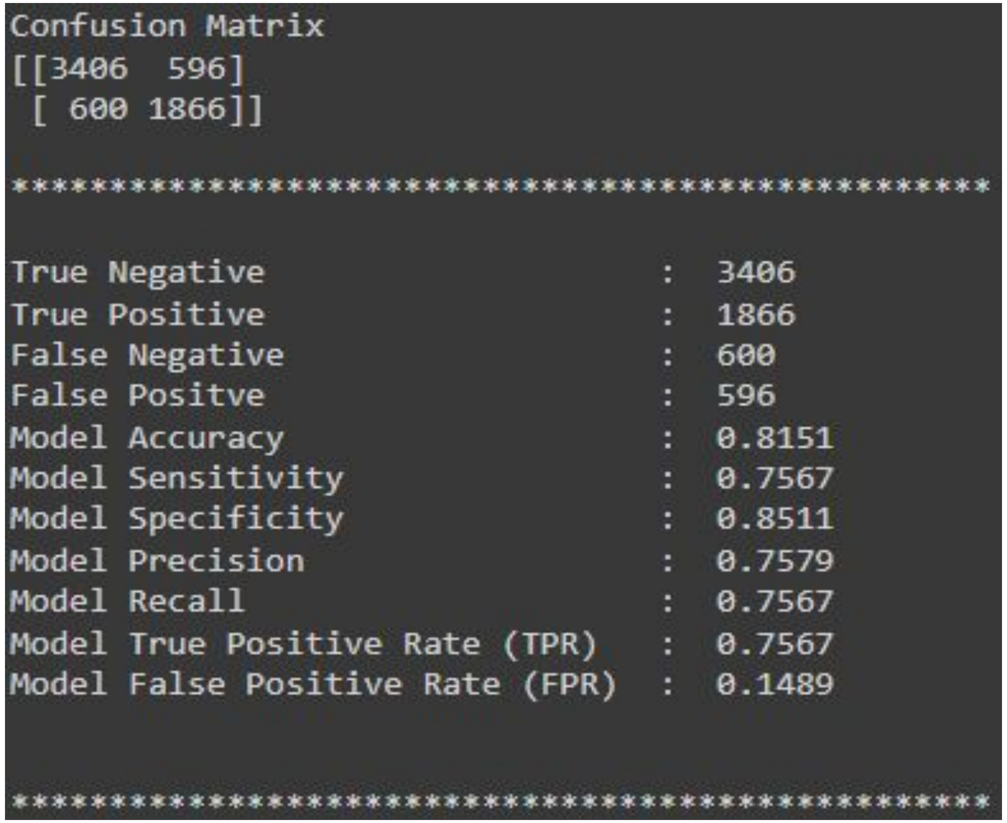


```
Confusion Matrix
[[3235  767]
 [ 486 1980]]

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


True Negative                         :  3235
True Positive                         :  1980
False Negative                        :  486
False Positve                         :  767
Model Accuracy                        :  0.8063
Model Sensitivity                     :  0.8029
Model Specificity                     :  0.8083
Model Precision                       :  0.7208
Model Recall                          :  0.8029
Model True Positive Rate (TPR)        :  0.8029
Model False Positive Rate (FPR)       :  0.1917


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

Model Building

Cut off precision point -2





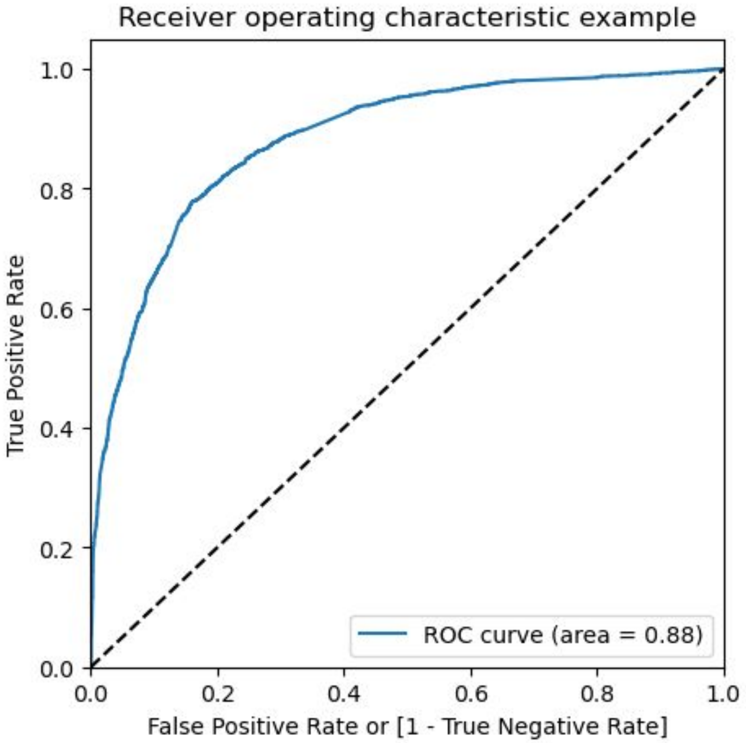
```
Confusion Matrix
[[3406  596]
 [ 600 1866]]

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

True Negative                    :  3406
True Positive                    :  1866
False Negative                   :  600
False Positve                    :  596
Model Accuracy                   :  0.8151
Model Sensitivity                :  0.7567
Model Specificity                :  0.8511
Model Precision                  :  0.7579
Model Recall                     :  0.7567
Model True Positive Rate (TPR)   :  0.7567
Model False Positive Rate (FPR)  :  0.1489


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

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

Model Evaluation

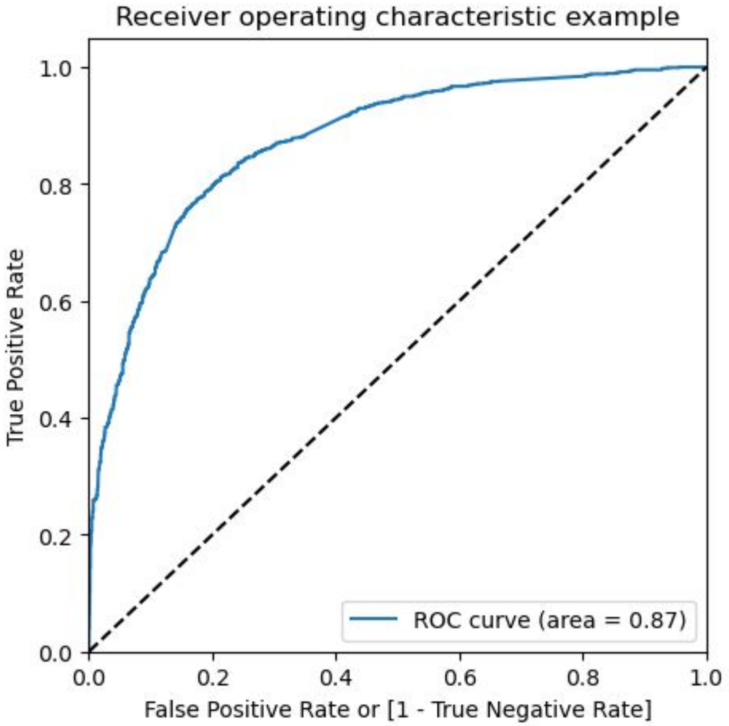
ROC Curve – Train Data Set

- 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

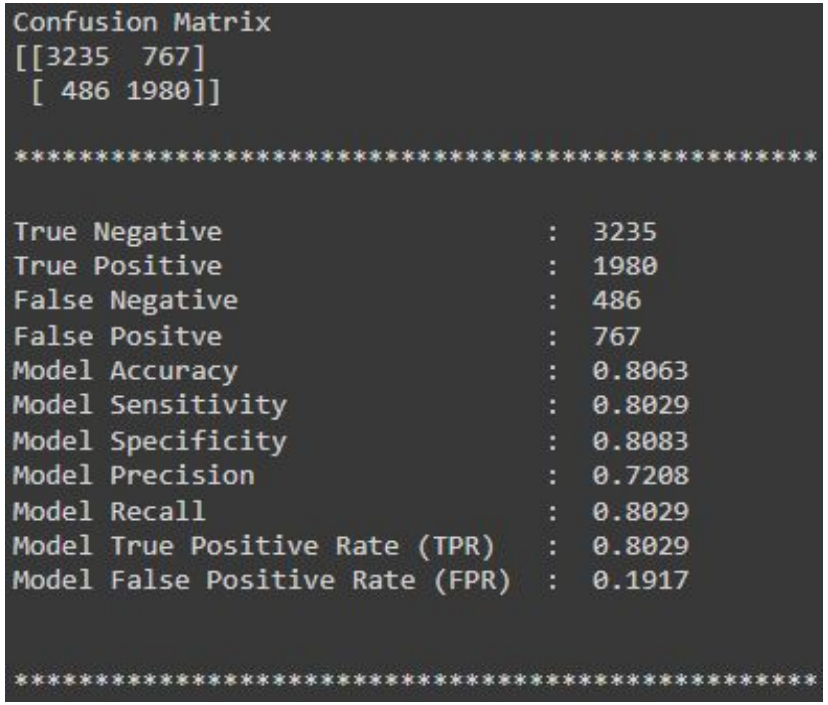
- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



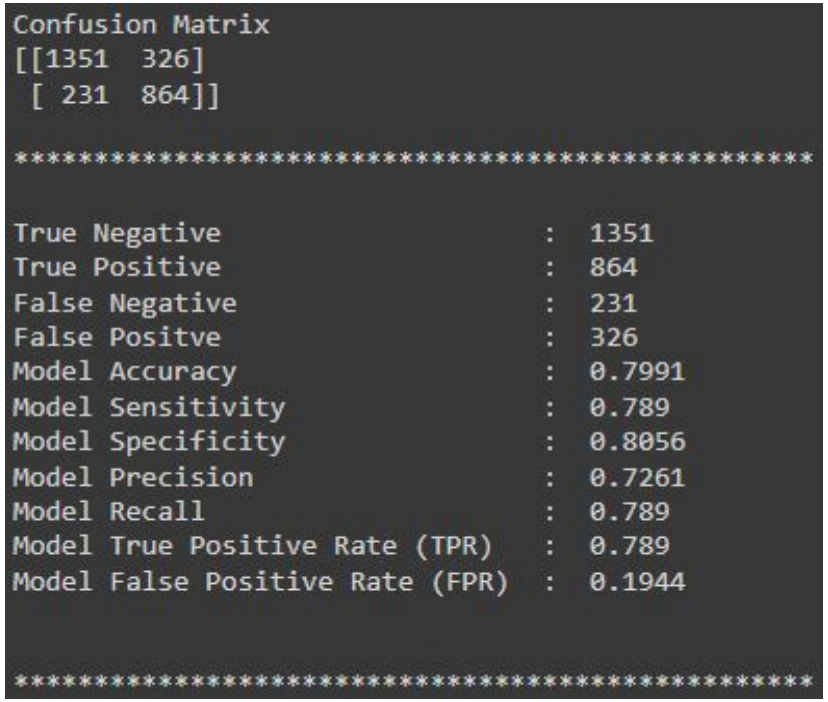
Model Evaluation

Confusion Matrix & Metrics

Test data



```
Confusion Matrix
[[3235  767]
 [ 486 1980]]

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

True Negative                   :  3235
True Positive                   :  1980
False Negative                  :  486
False Positve                   :  767
Model Accuracy                  :  0.8063
Model Sensitivity               :  0.8029
Model Specificity               :  0.8083
Model Precision                 :  0.7208
Model Recall                    :  0.8029
Model True Positive Rate (TPR)  :  0.8029
Model False Positive Rate (FPR) :  0.1917


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



```
Confusion Matrix
[[1351  326]
 [ 231  864]]

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

True Negative                   :  1351
True Positive                   :  864
False Negative                  :  231
False Positve                   :  326
Model Accuracy                  :  0.7991
Model Sensitivity               :  0.789
Model Specificity               :  0.8056
Model Precision                 :  0.7261
Model Recall                    :  0.789
Model True Positive Rate (TPR)  :  0.789
Model False Positive Rate (FPR) :  0.1944


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

Train data

- Using a cut-off value of 0.345, the model achieved a sensitivity of 80.29% in the train set and 78.9% 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 78%.
- The model also achieved an accuracy of 79.99%, which is in line with the study's objectives.

Lead classification by Score

General Leads :

- Leads Having score more than 60
- Model with this score is used to classify the leads in general
- The conversion rate probability for these leads is : 83%

Intern Leads :

- Leads Having score more than 45
- Model with this score is used when the organization has high number of resources (Like during the 2 months the organization hires interns) to invest in Lead Conversion operation.
- This model will give a large number of potential leads to pursue but the lead conversion rate is lower than hot leads.
- The conversion rate probability for these leads is : 78%

Hot Leads :

- Leads having score more than 80
- Model with this score is used when the organization has fewer number of resources or does not with to pursue every potential lead.
- In this case the number of leads returned will be low but will have a higher probability of converting.
- The conversion rate is 89%.

Recommendation based on Final Model

- As per the problem statement, Boosting lead conversion is vital for X Education's growth. We've created a regression model to pinpoint key factors influencing 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: 5.59
  - Lead Source_Reference: 3.08
  - TotalVisits: 0.37
  - Current_occupation_Working Professional: 2.7
  - Last Activity_SMS Sent: 2.25
  - Last Activity_Others: 1.61
  - Total Time Spent on Website: 1.03
  - Last Activity_Email Opened: 1.10 Lead Source_Olark Chat: 0.99
- We have also identified features with negative coefficients that may indicate potential areas for improvement.These include:
  - Page Views Per Visit: -0.37
  - Specialization_Not Specified: -1.18
  - Lead Origin of Landing Page Submission: -1.08

Recommendation based on Final Model

To increase our Lead Conversion Rates

- Concentrate on attributes with favorable coefficients for precise marketing tactics.
- 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.
- Allocate more resources to increase mobile and handphone based communications such as SMS and WhatsApp
- 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

Vandit Sardana
Saras Sangle
Santosh Govardhan