Lead Scoring Case Study

By: Jitesh Garg

Jaya Trivedi

Problem Statement

- 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.
- The company gets a lot of leads, potential customers who provide their email address and phone number over website, then the sales team start making calls and write emails to them. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate is very poor at around 30%. For example, if the company acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify 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.

Business goal

- To identify the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- To enhance lead selection, we've been appointed to build a model assigning lead scores based on conversion likelihood.
- The company aims for an ambitious 80% lead conversion rate.

Approach & Methodology

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:

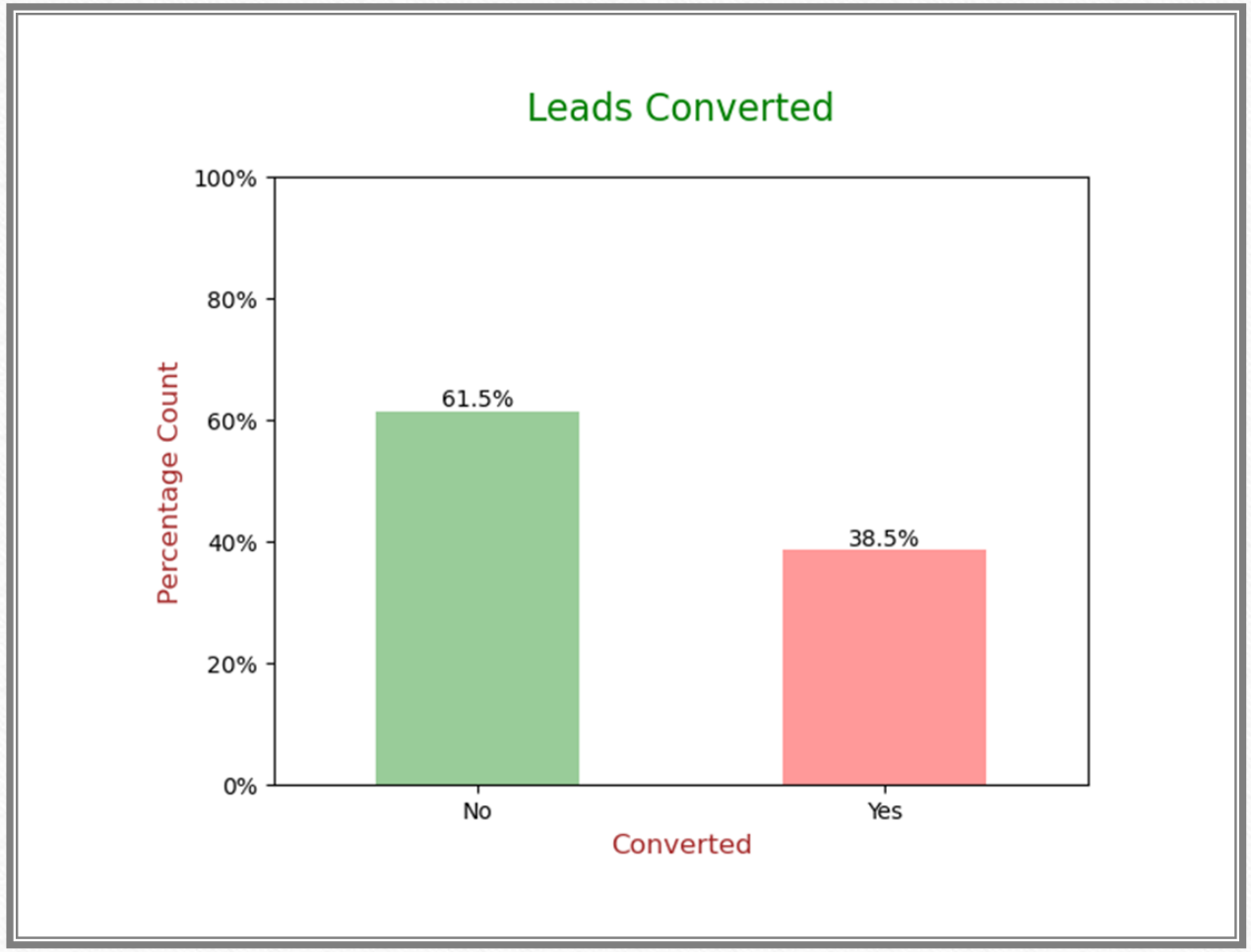
Suggest top 3 features to focus for higher conversion & areas for improvement

Data Cleaning (1/2)

- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was done for some categorical variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

Data Cleaning (2/2)

- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in 'TotalVisits' and 'Page Views Per Visit' were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to "Others".
- Binary categorical variables were mapped.
- Other cleaning activities were performed to ensure data quality and accuracy.
    - Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google).



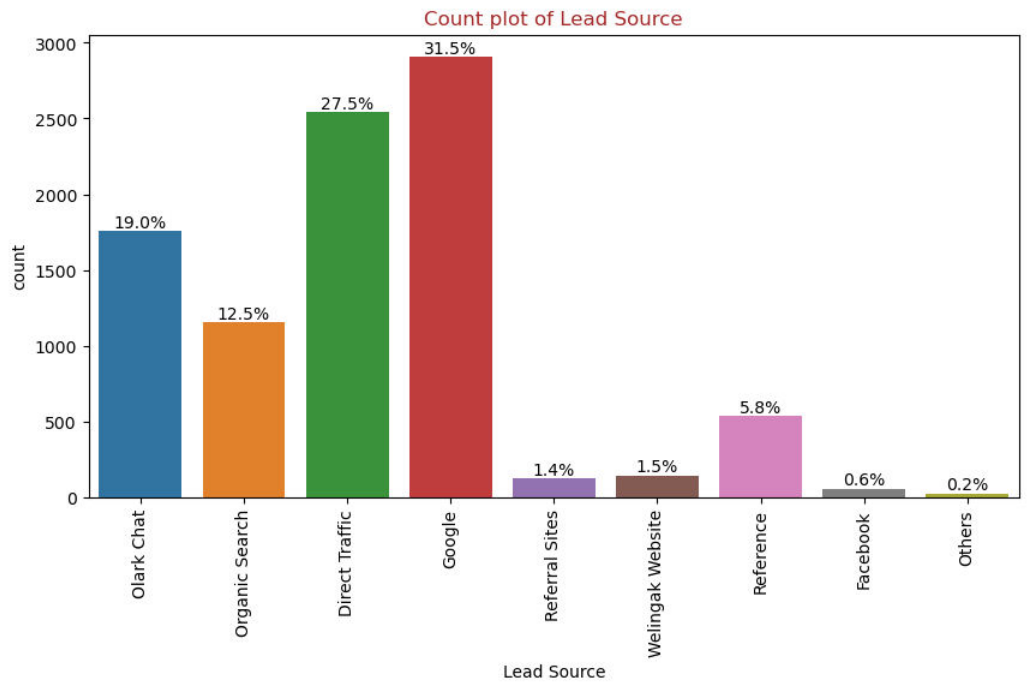
Exploratory Data Analysis (EDA)

Data is imbalanced while analyzing the target variable:

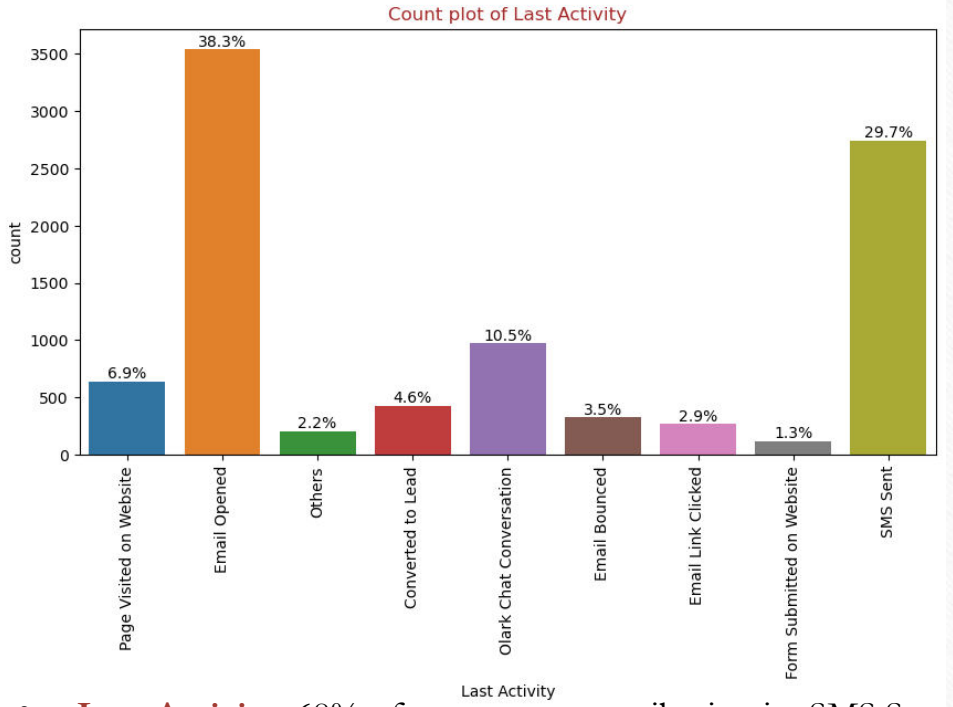
- Conversion rate is of 38.5%, it means that only 38.5% of the people have converted to leads. (Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

EDA – Univariate Analysis

(Categorical Variables)



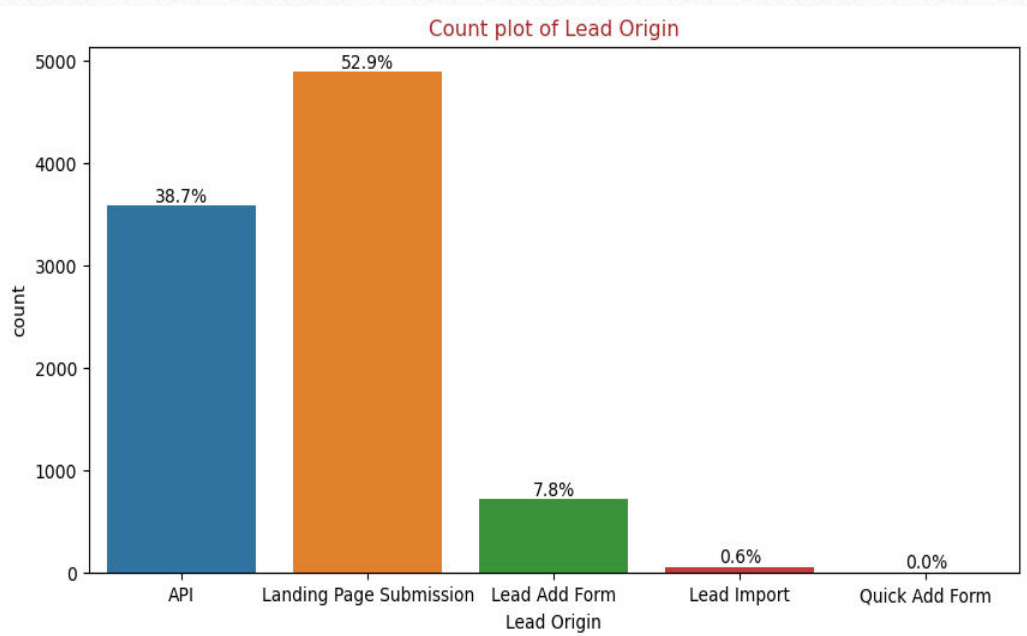
- **Lead Source:** 58% Lead source is from Google & Direct Traffic combined.

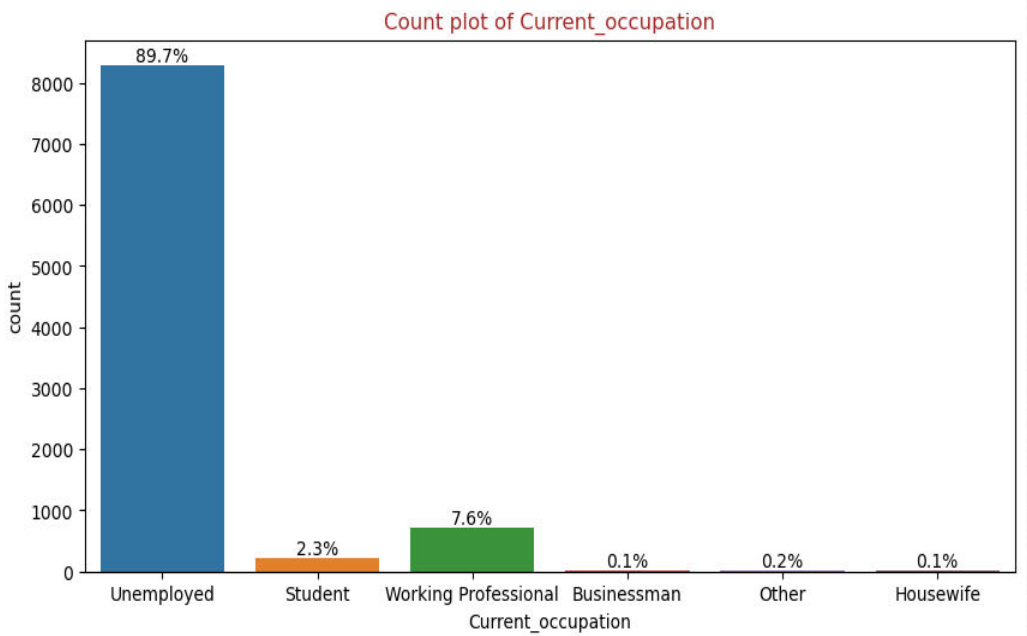


- **Last Activity:** 68% of customers contribution in SMS Sent & Email Opened activities.

EDA – Univariate Analysis

(Categorical Variables)

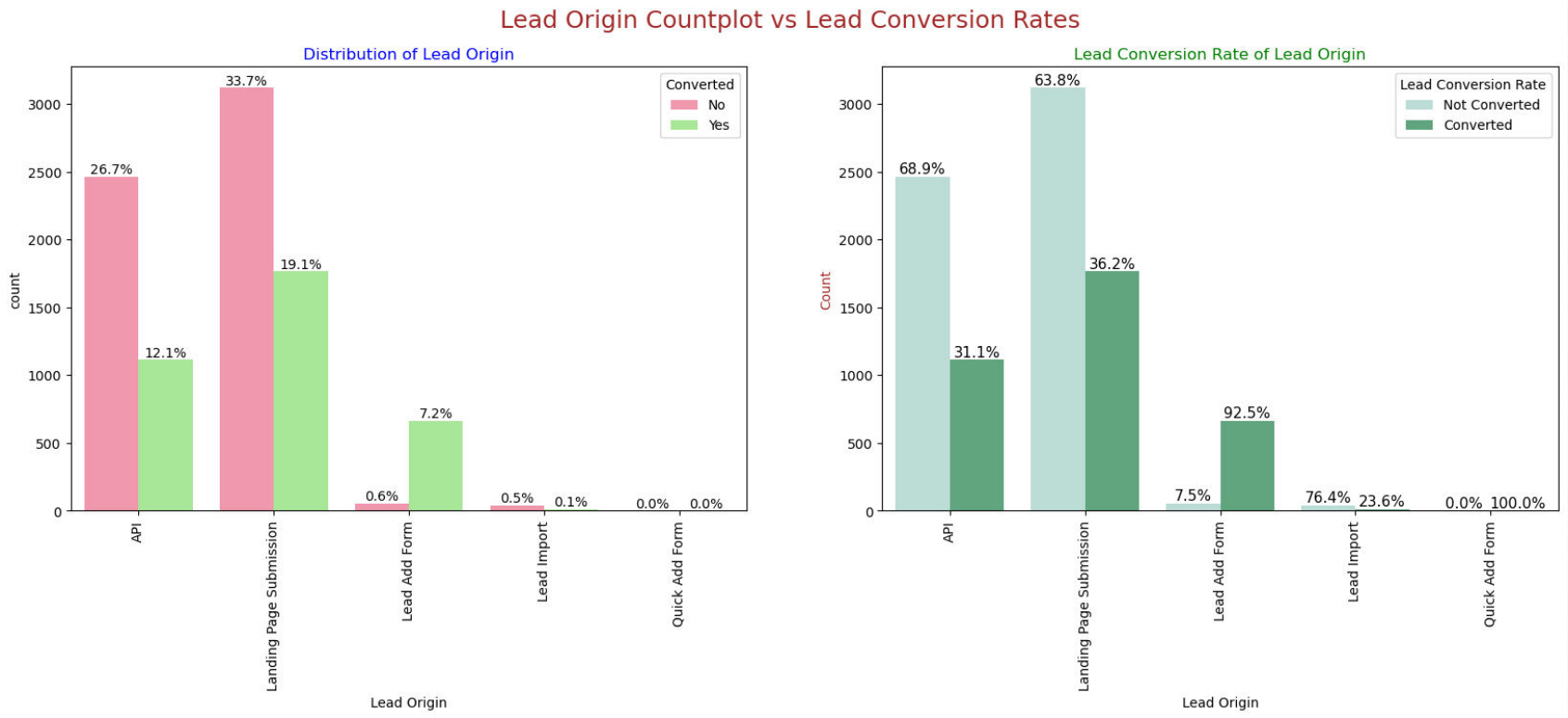




- **Lead Origin:** "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- **Current_Occupation:** It has 90% of the customers as Unemployed.

EDA – Bivariate Analysis

(Categorical Variables)

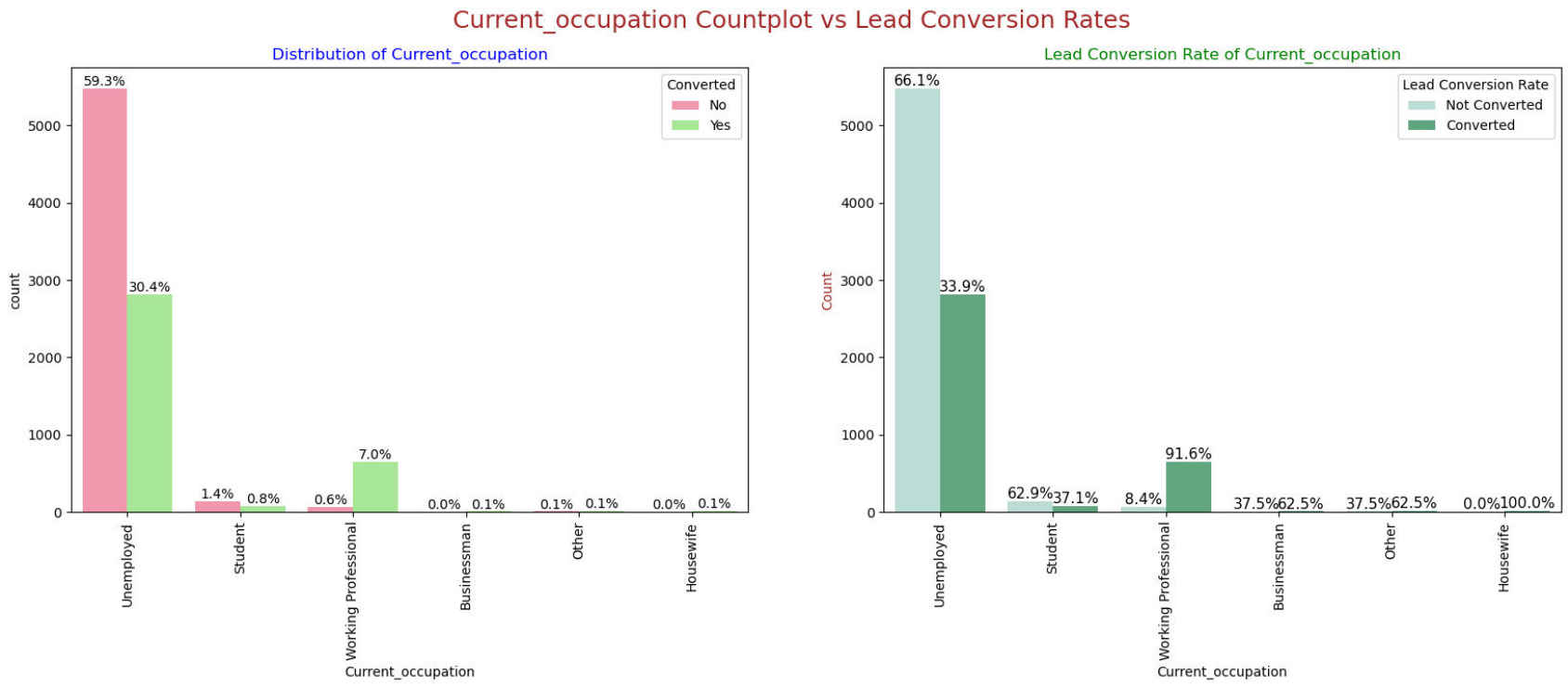


Lead Origin:

- 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

(Categorical Variables)

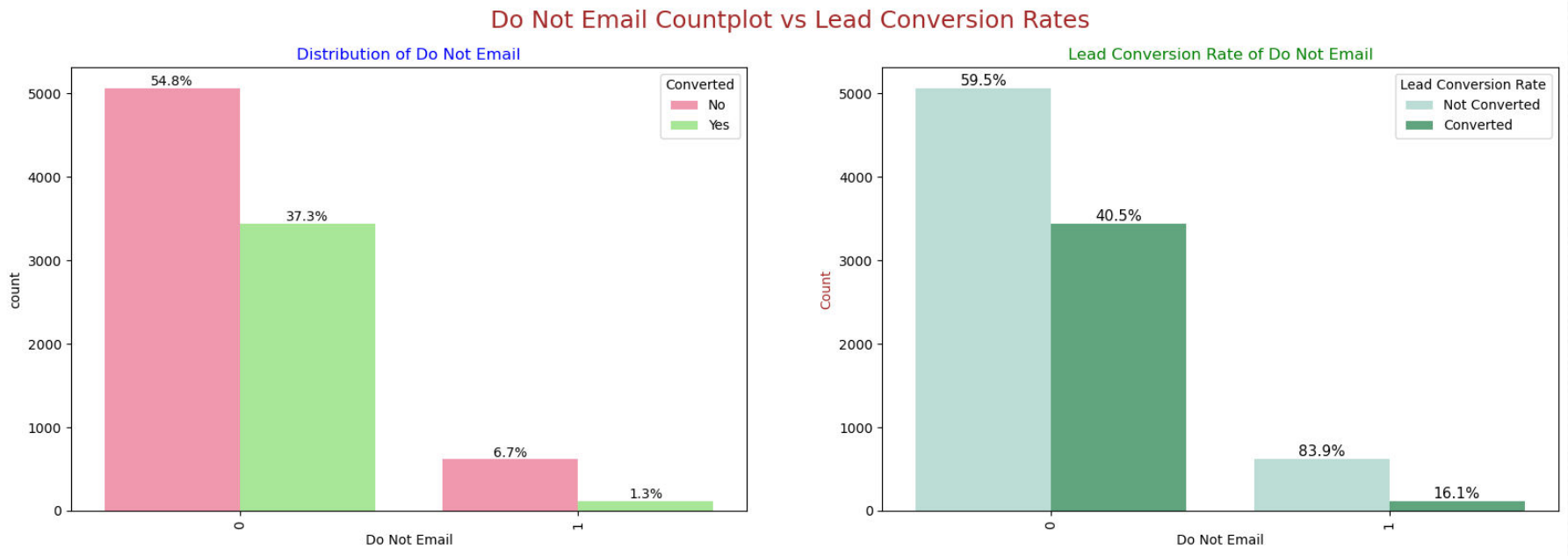


Current_occupation:

- 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

(Categorical Variables)

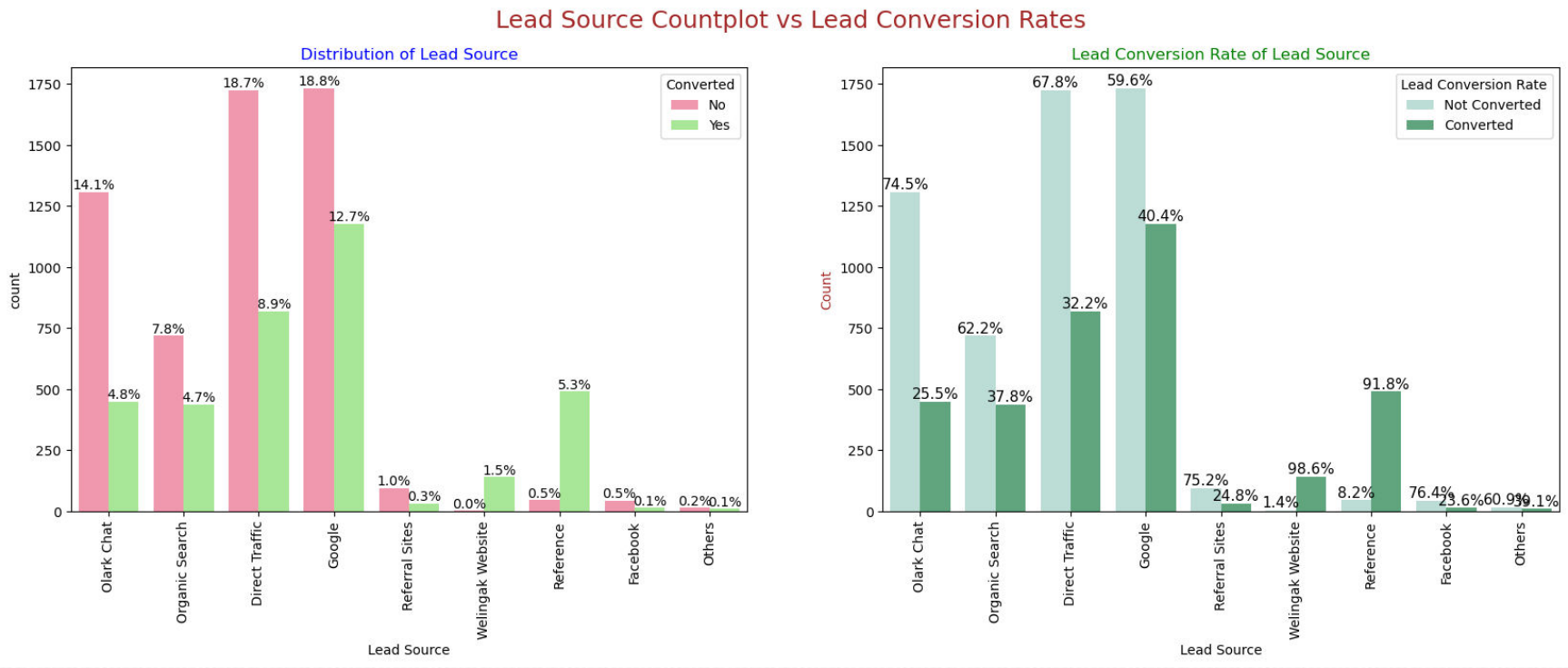


Do Not Email:

- 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

(Categorical Variables)

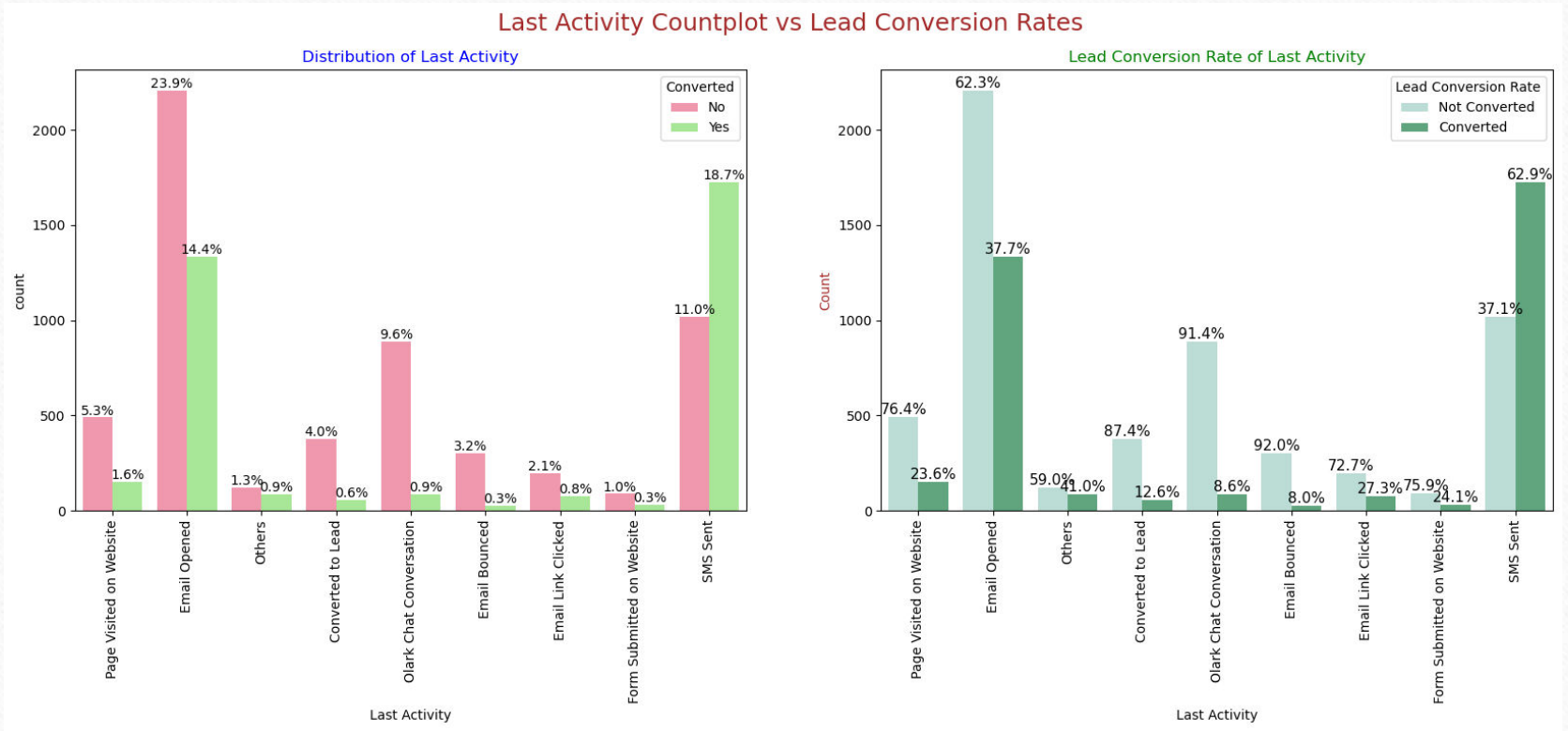


Lead Source:

- 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

(Categorical Variables)

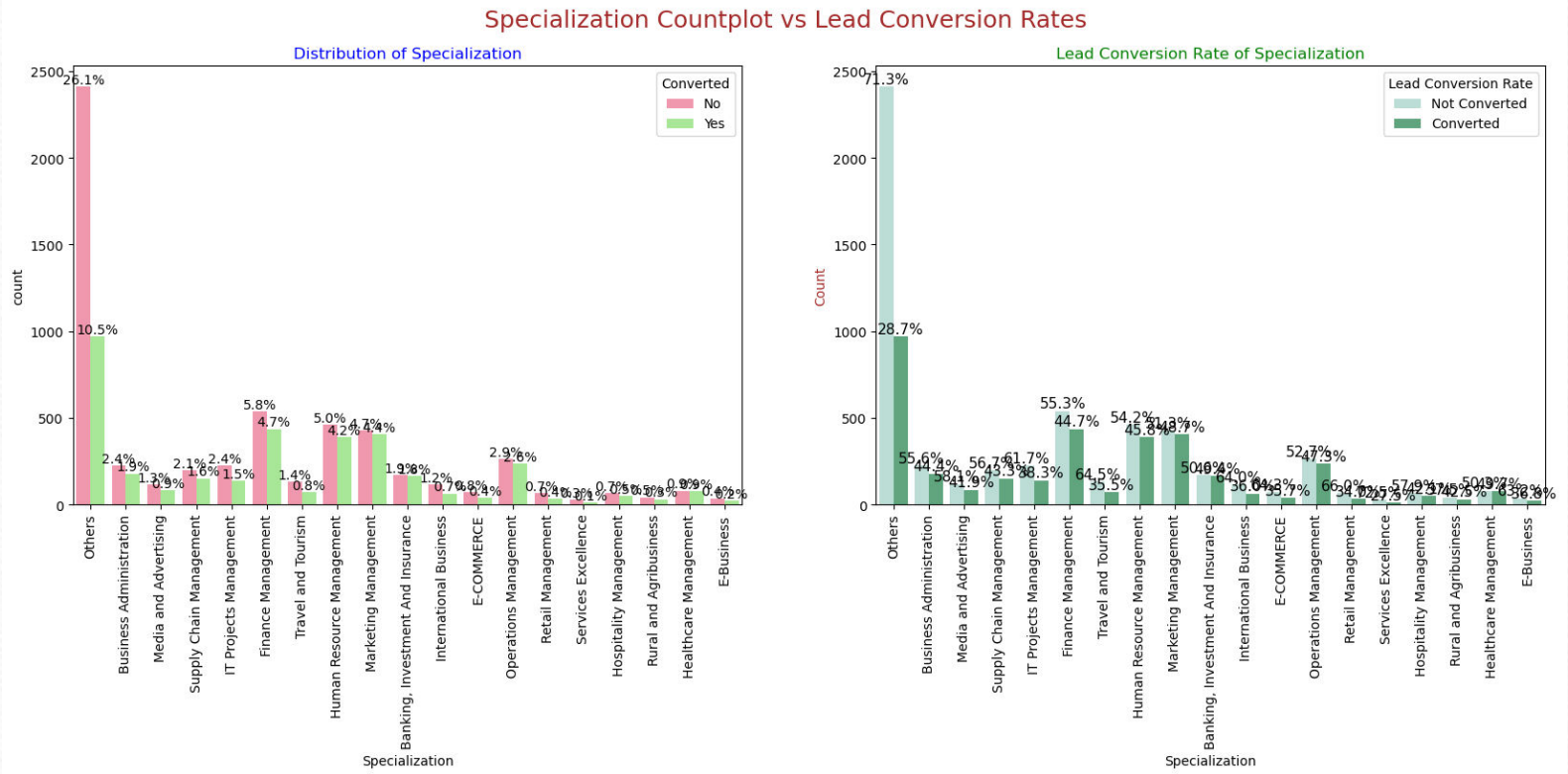


Last Activity:

- '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

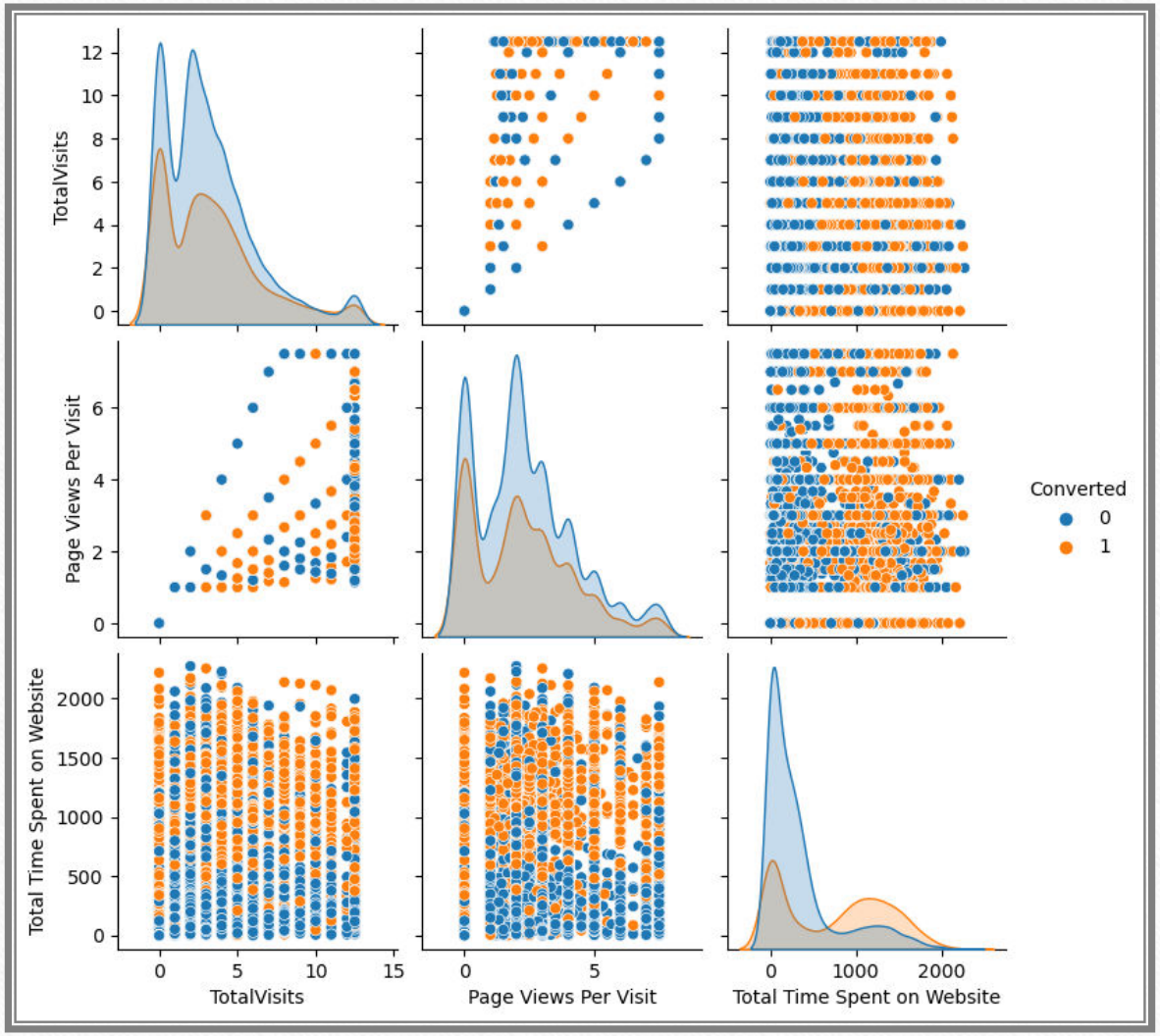
(Categorical Variables)



Speacialization:

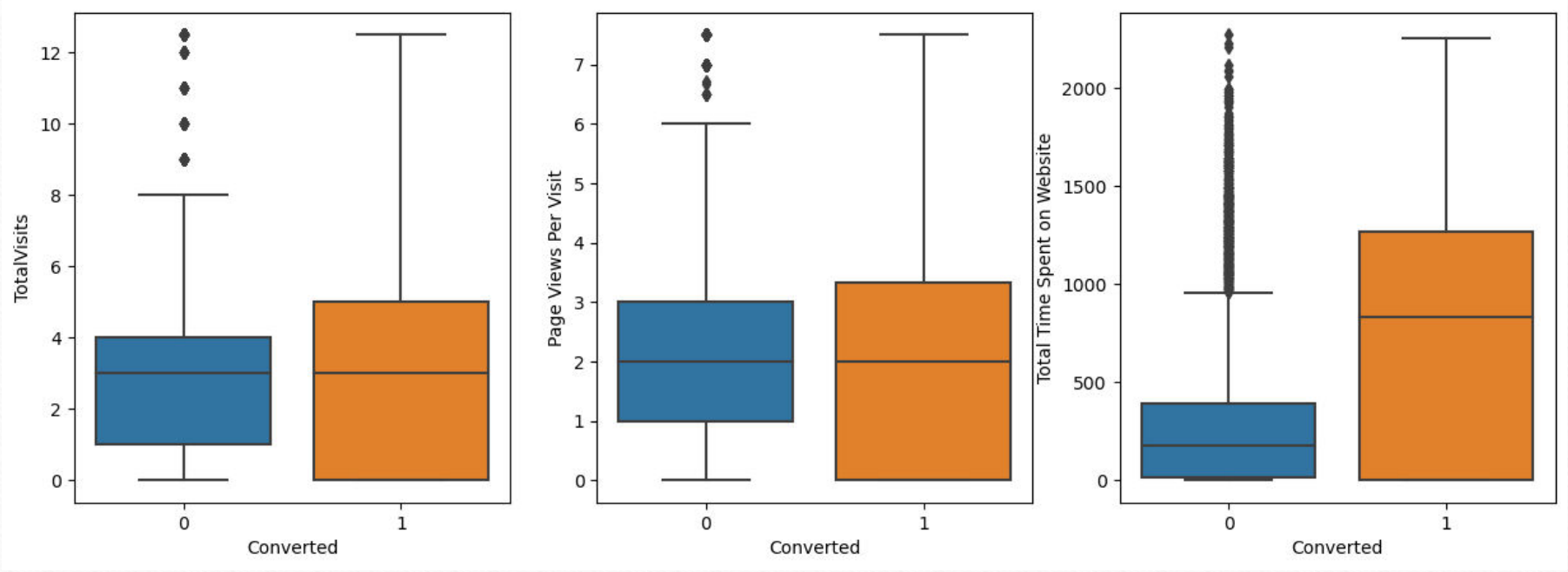
- Marketing Management, HR Management, Finance Management shows good contribution in Leads conversion than other specializations.

EDA – Bivariate Analysis (Numerical Variables)



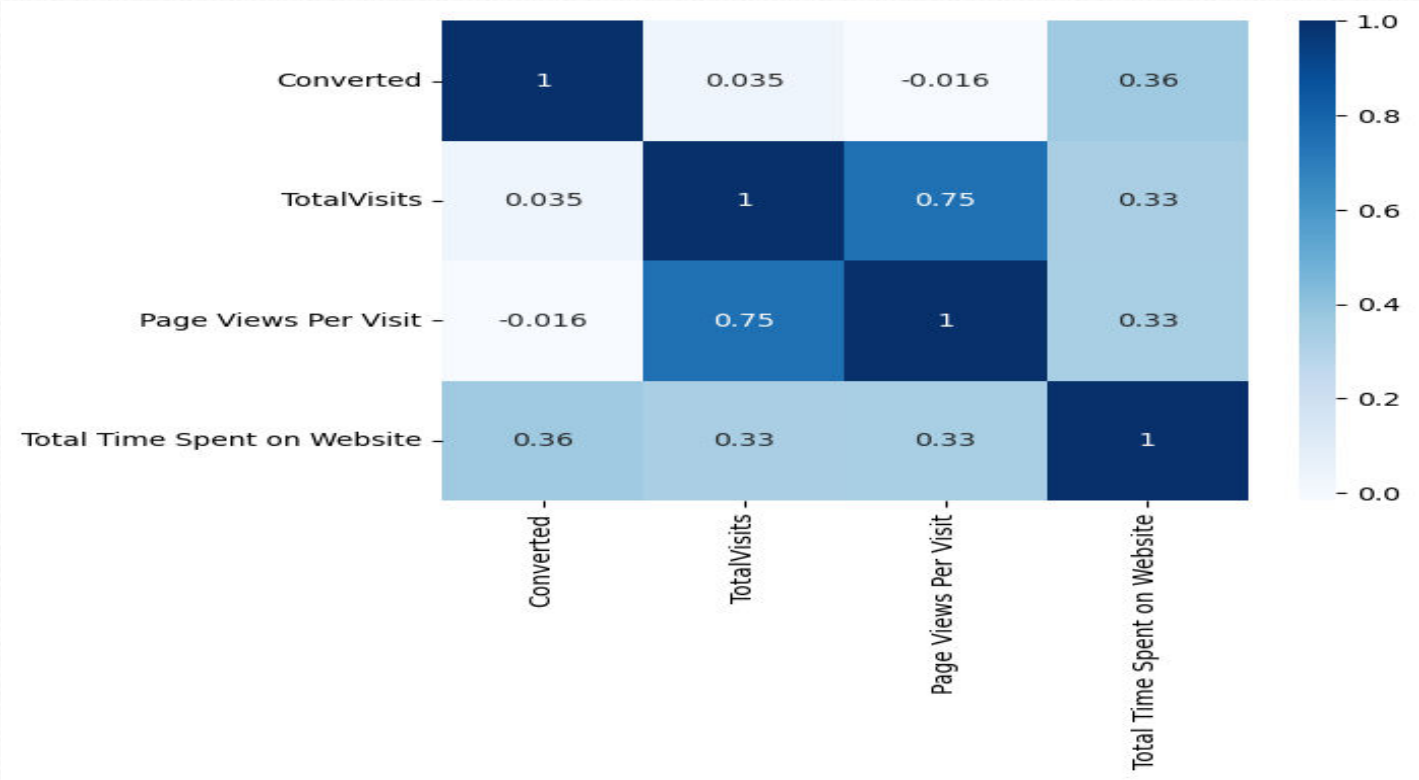
EDA – Bivariate Analysis

(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

Heatmap to show correlation between numerical variables



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

- Use **Recursive Feature Elimination (RFE)** for Feature Selection
  - Running RFE with **15 variables** as output
- Build Model by removing the variable whose **p-value is greater than 0.05** and **VIF value is greater than 5**
- **Model 4 looks stable with:**
  - **significant p-values within the threshold (p-values < 0.05) and**
  - **No sign of multicollinearity with VIFs less than 5**
- **Predictions on test data set**
- **Overall accuracy 81%**

Model Evaluation - Confusion Matrix & Metrics

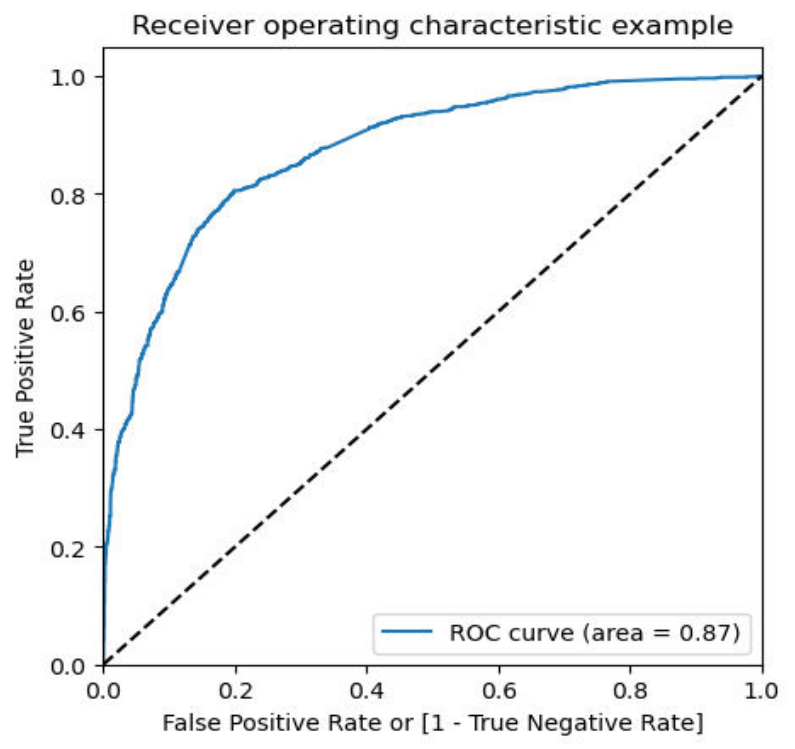
Train Data Set

```
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

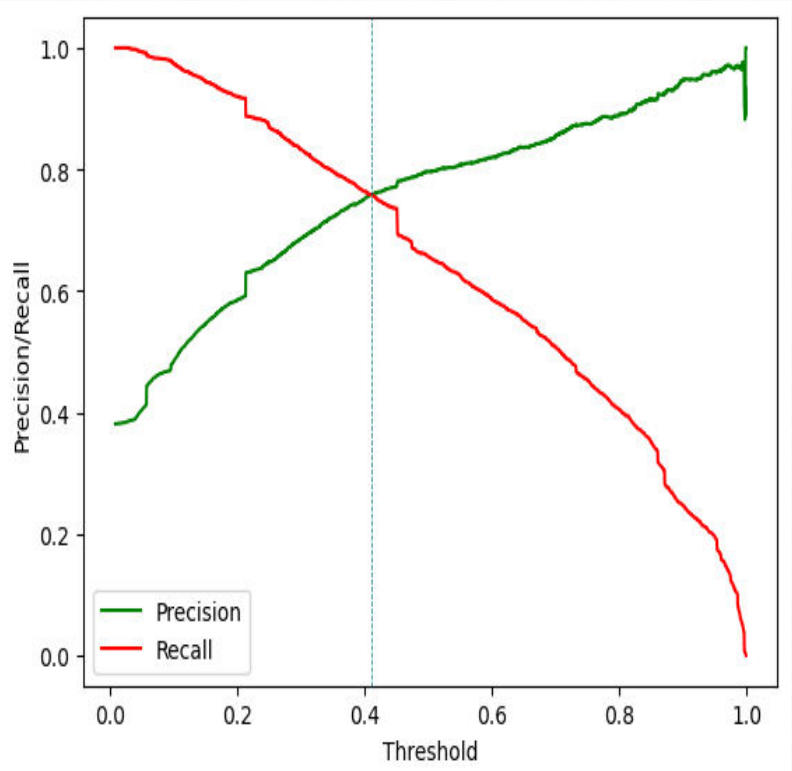
```
Confusion Matrix
[[1353  324]
 [ 221  874]]

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

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
```

Model Evaluation - ROC Curve



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

Model Evaluation - Precision and recall trade off



- The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimize the performance of the model based on business requirement.
- Probability threshold is 0.41(approx.)
- Recall have dropped to around 75%, but we need it close to 80% as the Business Objective.
- 80% for the metrics we are getting with the sensitivity-specificity cut-off threshold of 0.345.

Train – Test Results

Train - Test

- Train Data Set:

  Accuracy: 80.46%

  Sensitivity: 80.05%

  Specificity: 80.71%

- Test Data Set:

  Accuracy: 80.34%

  Sensitivity: 79.82% ≈ 80%

  Specificity: 80.68%

Observations

- The evaluation metrics are pretty close to each other, so it indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.
- The model achieved a sensitivity of 80.05% in the train set and 79.82% in the test set, using a cut-off value of 0.345.
- 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.
- 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).

Top features to predicting hot leads

Top 3 features that contributing positively to predicting hot leads in the model are:

- Lead Source_Welingak Website: 5.39
- Lead Source_Reference: 2.93
- Current_occupation_Working Professional: 2.67

- We should focus on more budget/spend on Welingak Website in terms of advertising, etc. to attract more leads.
- We can provide discounts for providing references that convert to lead to encourage more references.
- We should develop tailored messaging and engage working professionals through communication channels based on their engagement impact.

Recommendations

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.
- Engage working professionals with tailored messaging.
- Optimize communication channels based on lead engagement impact.
- 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.


THANK
YOU