LEAD SCORING CASE STUDY

- K SHILPA
- RAKSHA MORE
- SAGAR PATIL

TABLE OF CONTENTS

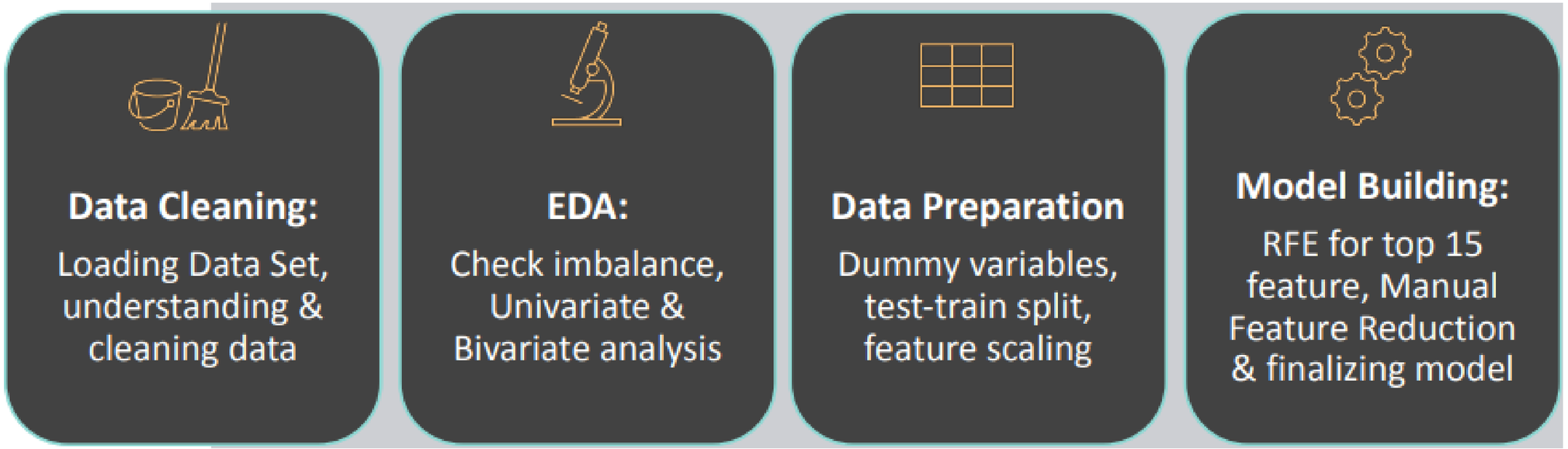
Problem Statement:

- 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.

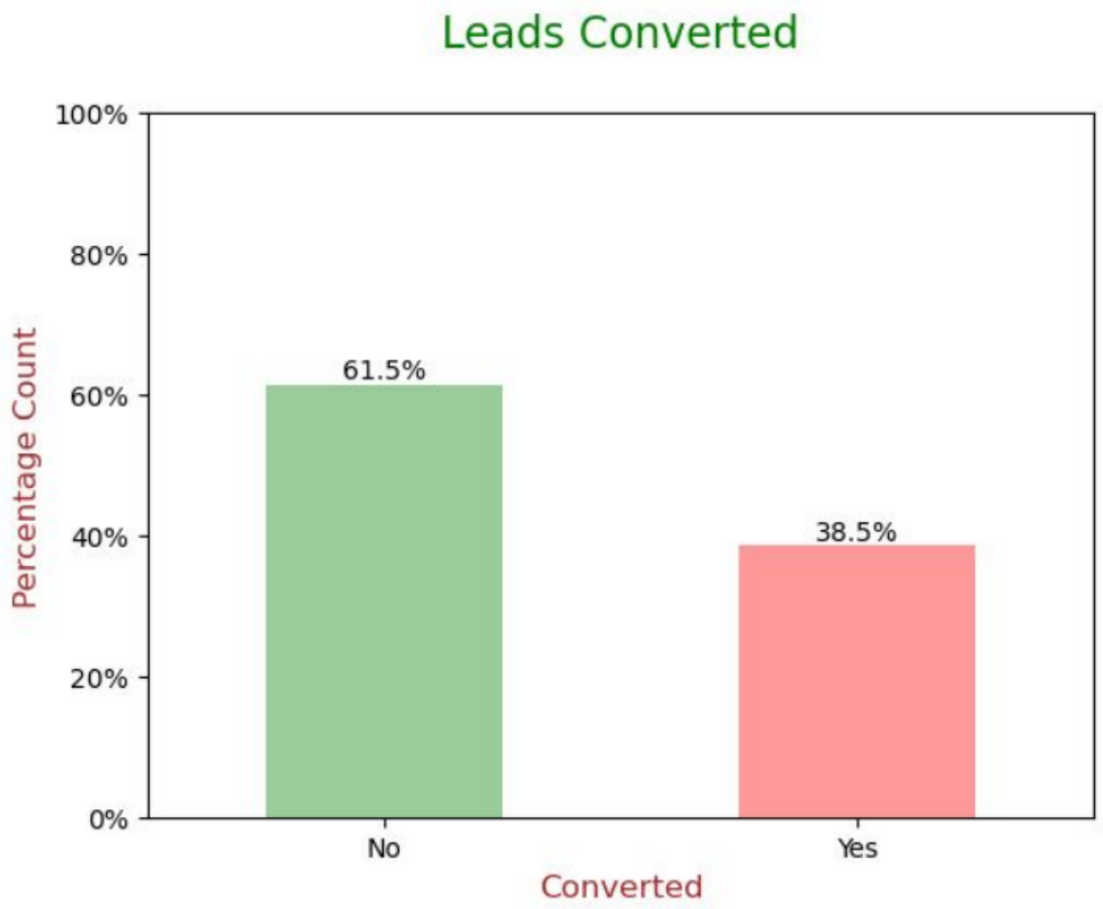
Objective of the Study:

- 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 Approach

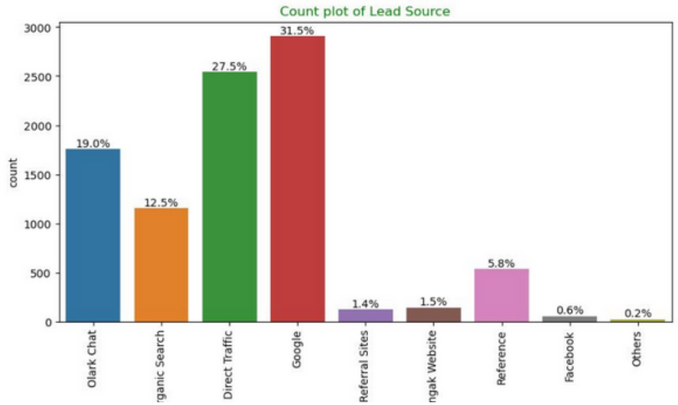


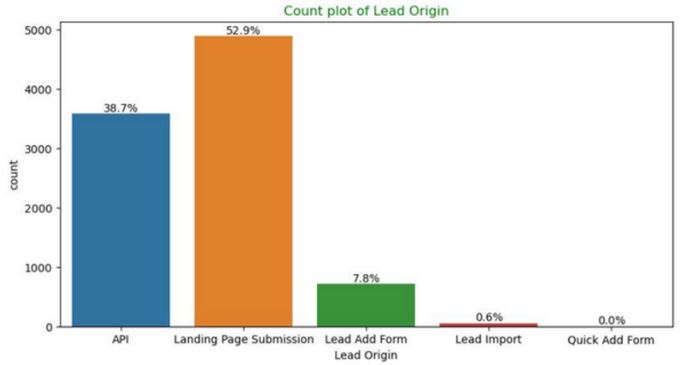
EDA -Data Imbalance

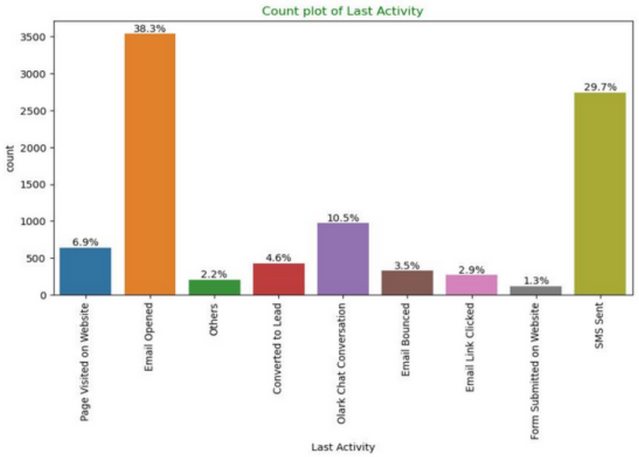


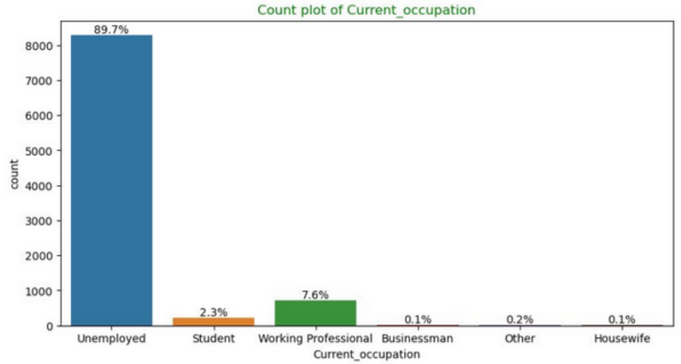
- Conversion rate is of **38.5%**, meaning only 38.5% of the people have converted to leads. (Minority)
- While **61.5%** of the people didn't convert to leads. (Majority)

EDA - Univariate Analysis – Categorical Variables



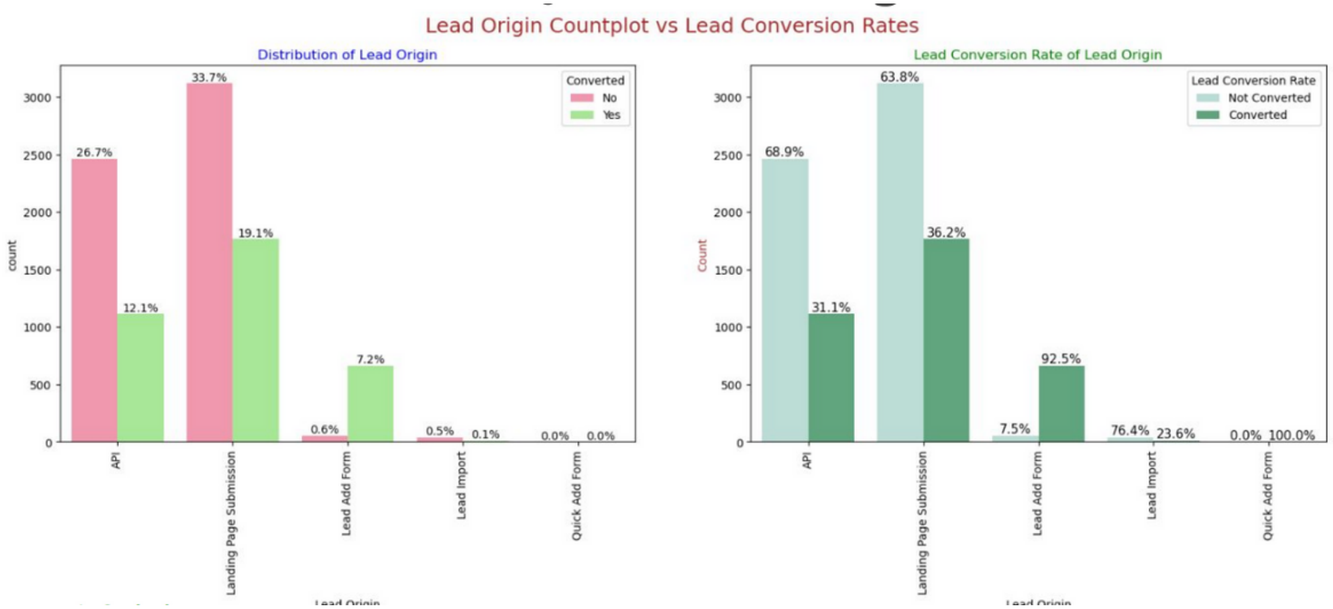


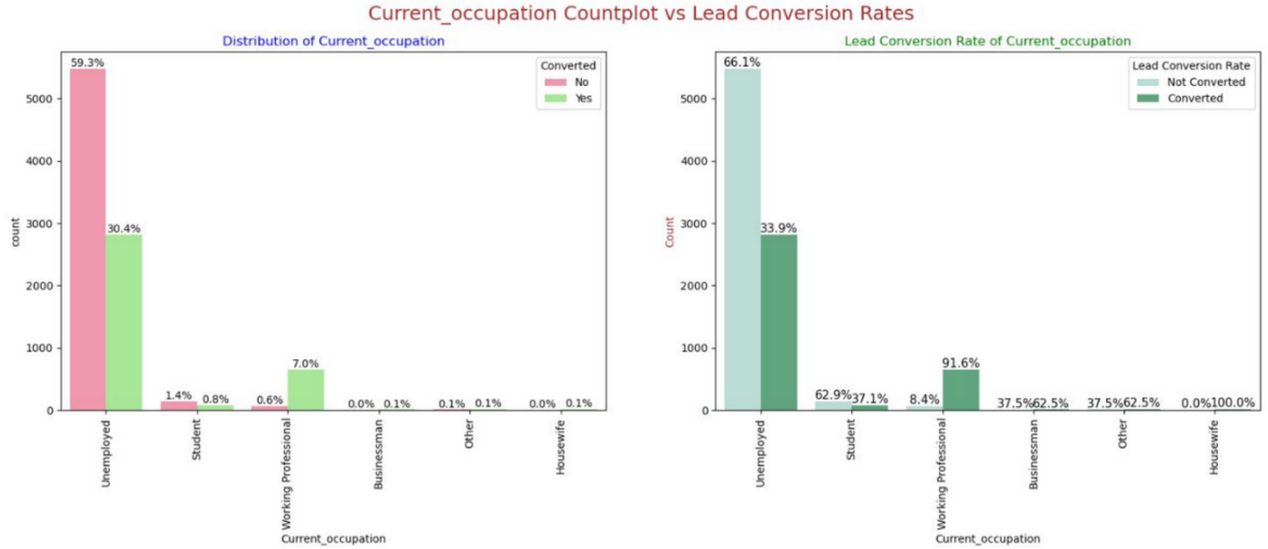




- **Lead Source:** 58% Lead source is from Google & Direct Traffic combined.
- **Last Activity:** 68% of customers contribution in SMS Sent & Email Opened activities.
- **Lead Origin:** "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- **Current_occupation:** It has 90% of the customers as Unemployed

EDA – Bivariate Analysis for Categorical Variables

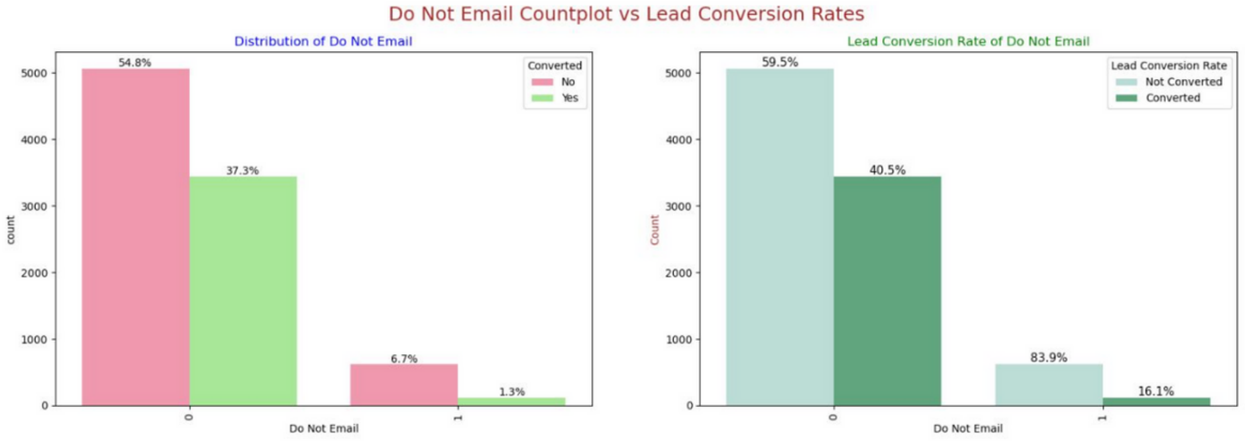


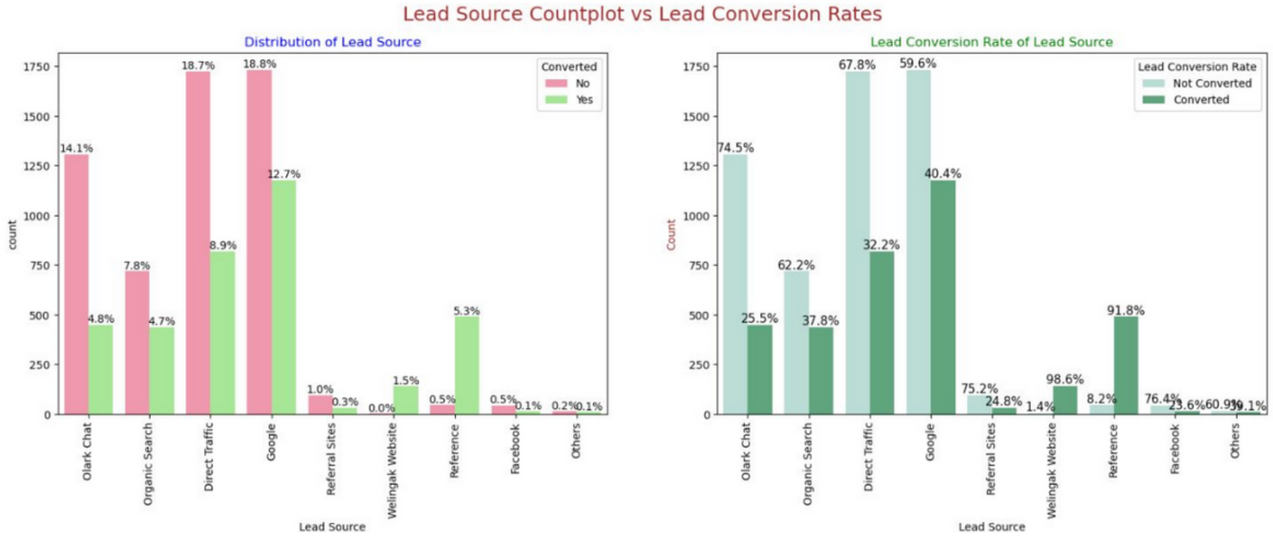


- **Lead Origin:**
  - Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%
  - The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

- **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

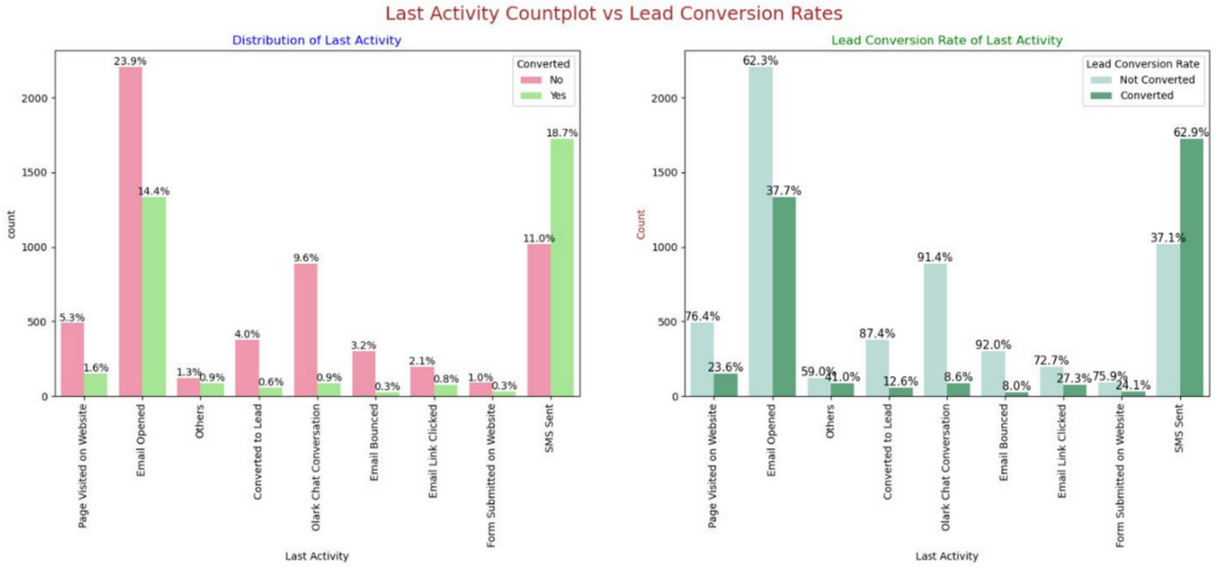


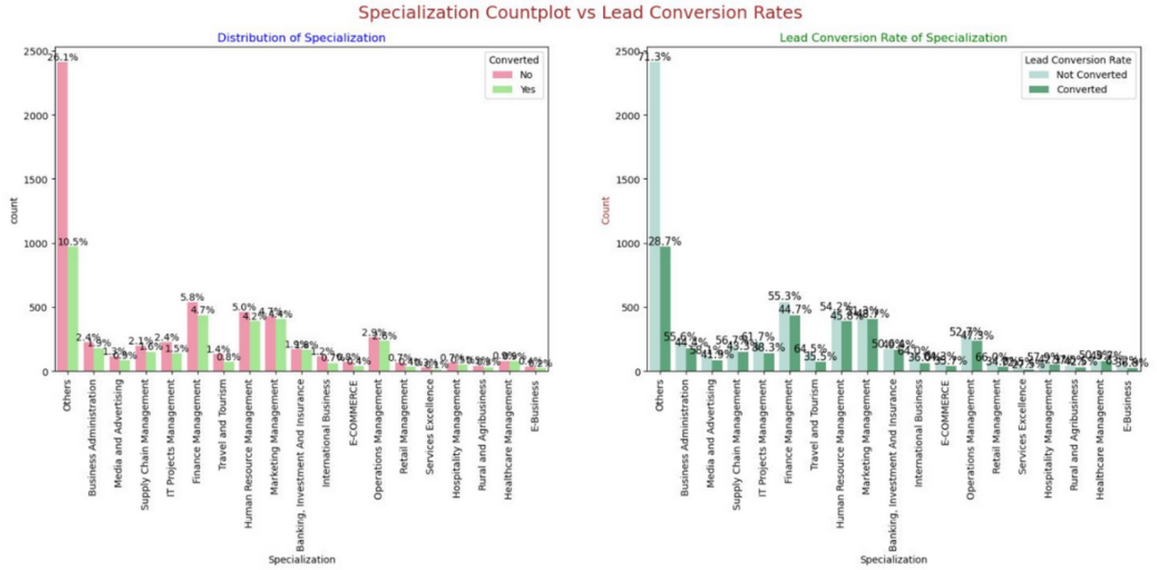


- **Do Not Email:**
  - 92% of the people has opted that they don't want to be emailed about the course & 40% of them are converted to leads.

- **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

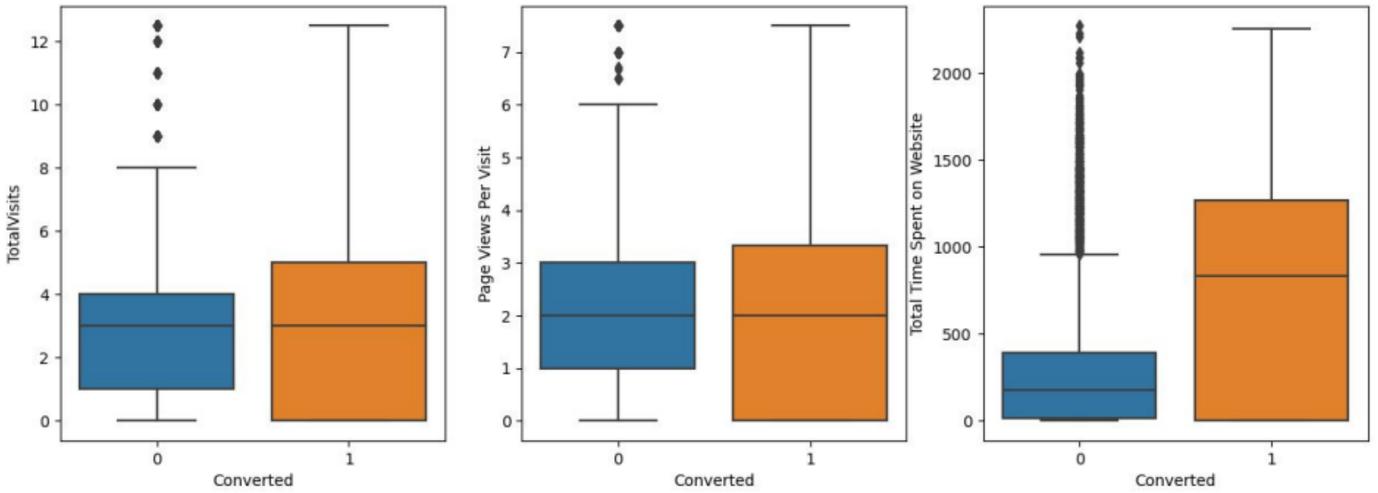


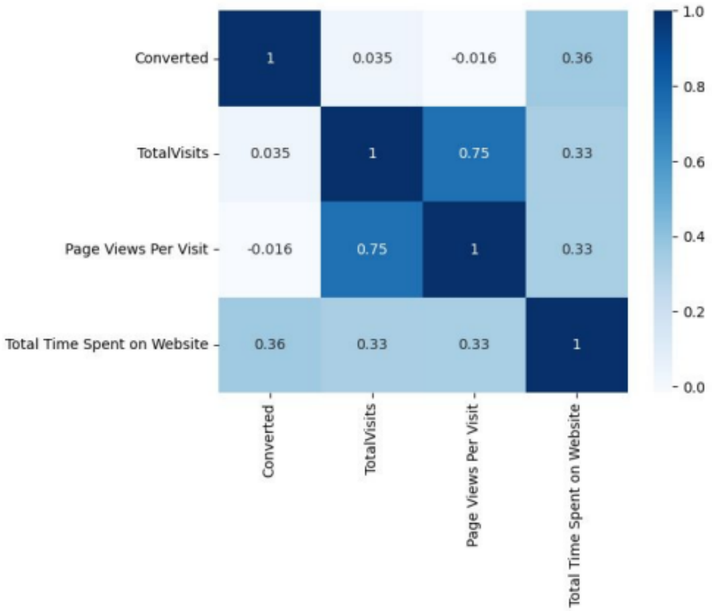


- **Last Activity:**
  - 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities,
  - 'Email Opened' activity contributed 38% of last activities performed by the customers, with 37% lead conversion rate

- **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 Cleaning

- Columns with **over 40% null values** were dropped.
- Drop columns that **don't add any insight or value** to the study objective (tags, country)
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- **Missing values** in categorical columns were handled based on value counts and certain considerations.
- **Imputation** was used for some categorical variables.
- **Additional categories** were created for some variables.
- **Outliers** in TotalVisits and Page Views Per Visit were treated and capped.
- Low frequency values were grouped together to “Others”.
- **Binary categorical variables** were mapped.

Data Preparation

- **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 and Model selection

- The data set has lots of dimension and large number of features which can hamper model performance
- Hence it is important to perform **Recursive Feature Elimination (RFE)** and to select only the important columns. .
- **Pre RFE there were 48 columns ,Post RFE,we were 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

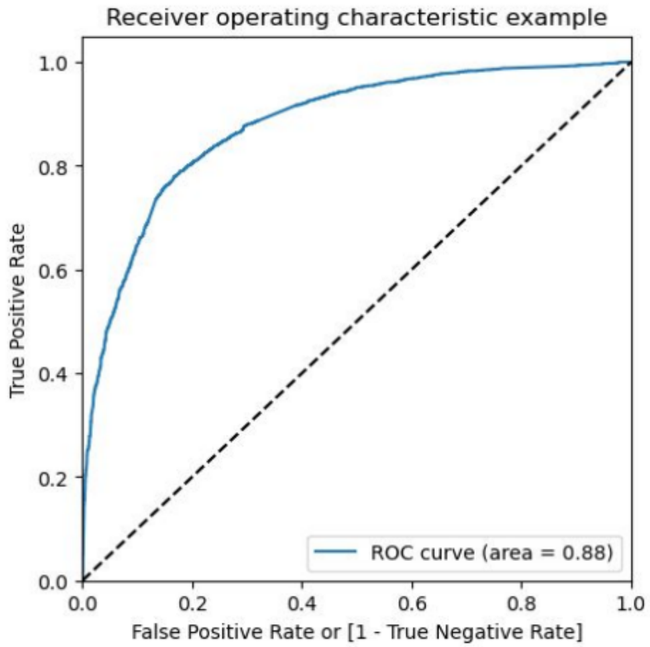
For Train data set

- 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.

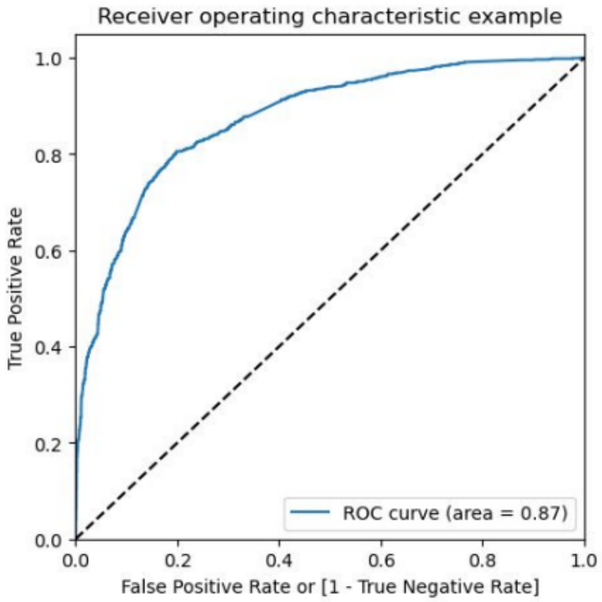
For Test data set

- 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.

ROC Curve



Train data set



Test data set

Model Evaluation

Confusion Matrix & Metrics

- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- 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.
- 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.

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
```

Recommendation

- 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

Recommendations-

- **For increasing Lead Conversion Rates -**

  - Develop strategies to attract high-quality **leads from top-performing lead sources**
  - **Focus on features with positive coefficients** for targeted marketing strategies.
  - **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.

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