X EDUCATION - LEAD SCORING CASE STUDY

IDENTIFYING HIGH-POTENTIAL LEADS TO FOCUS MARKETING EFFORTS AND ENHANCE CONVERSION RATES FOR X EDUCATION.

M Babitha
Lavanya

TABLE OF CONTENTS

OVERVIEW OF X EDUCATION COMPANY

● 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, fill out a course form, 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%.

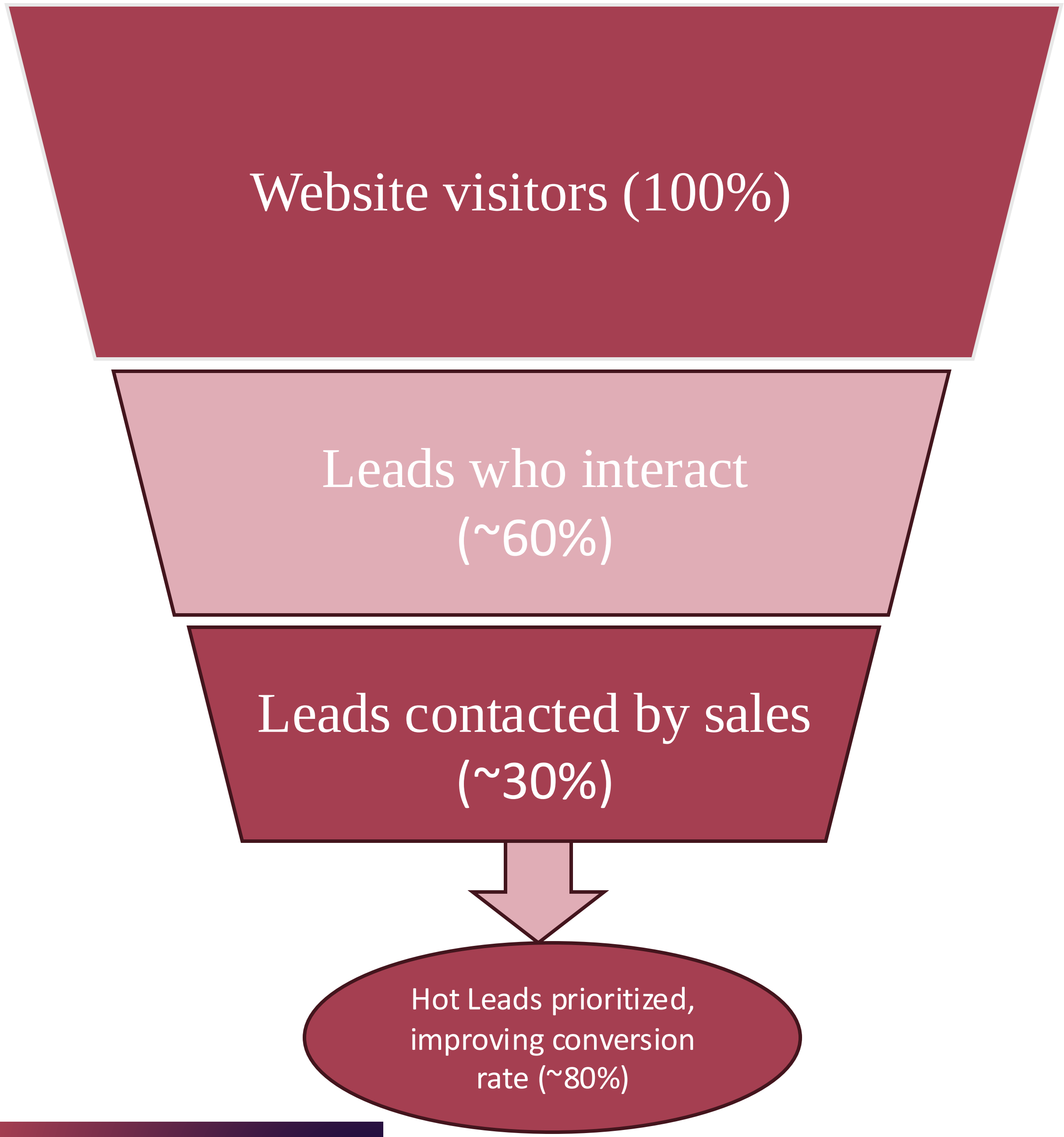
PROBLEM STATEMENT & OBJECTIVE OF THE STUDY

Problem Statement:

- X Education gets a lot of leads, but its lead conversion rate is very poor at around 30%
- X Education wants to make the lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team wants to know this potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

The 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%.

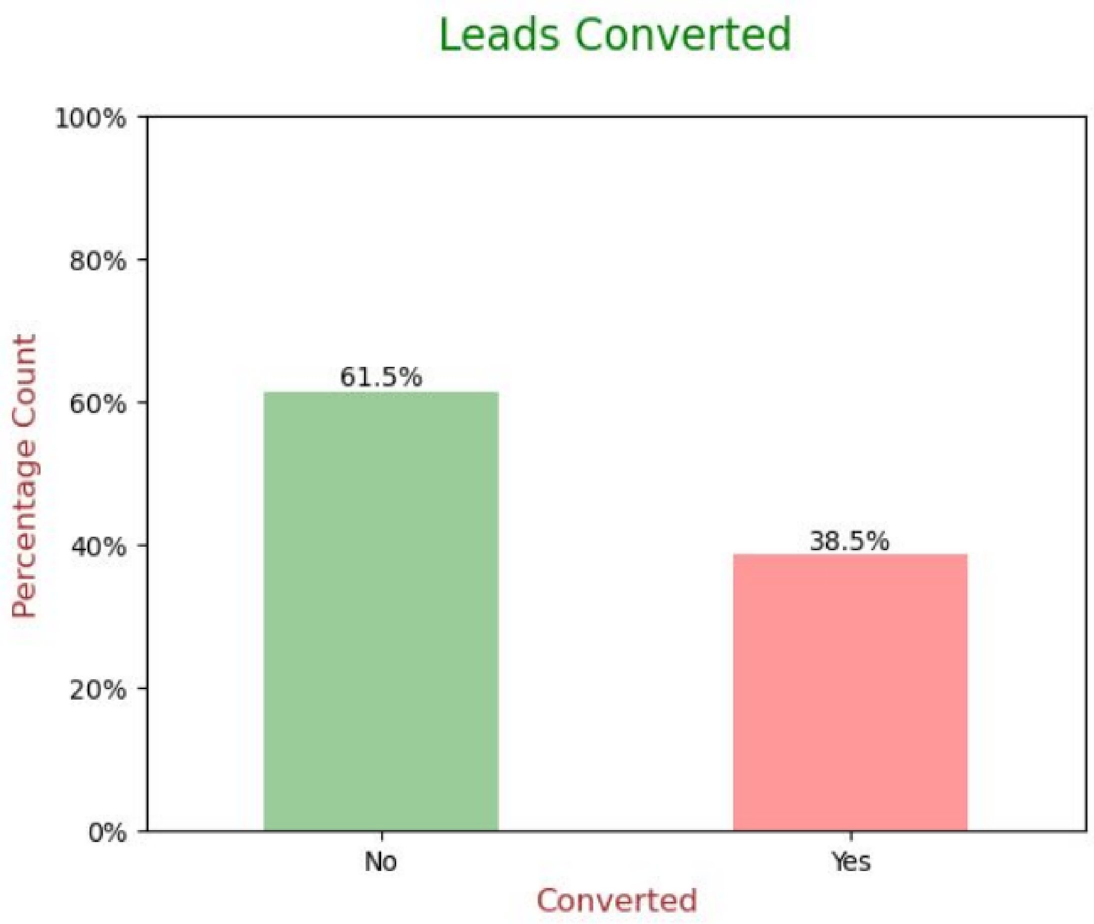

Website visitors (100%)
Leads who interact (~60%)
Leads contacted by sales (~30%)
Hot Leads prioritized, improving conversion rate (~80%)

DATA CLEANING

- "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 used for some categorical variables.
- Additional categories were created for some 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.

EDA

Data is imbalanced while analyzing target variable.

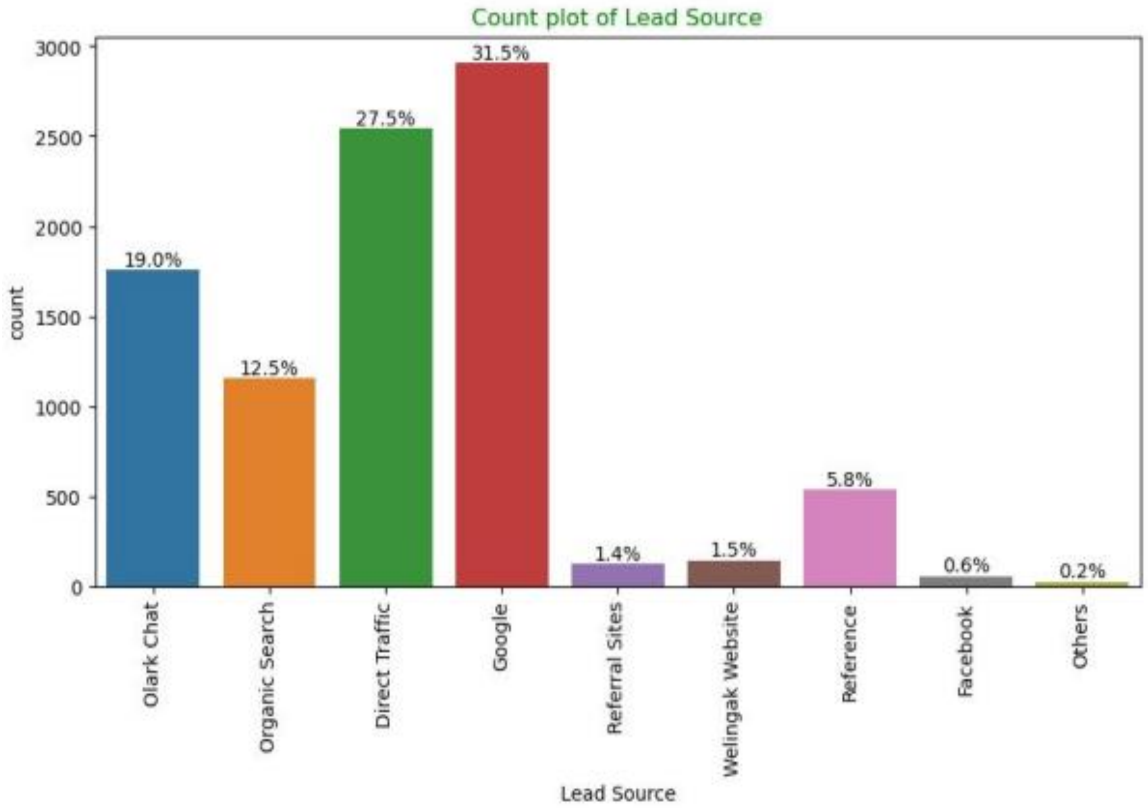


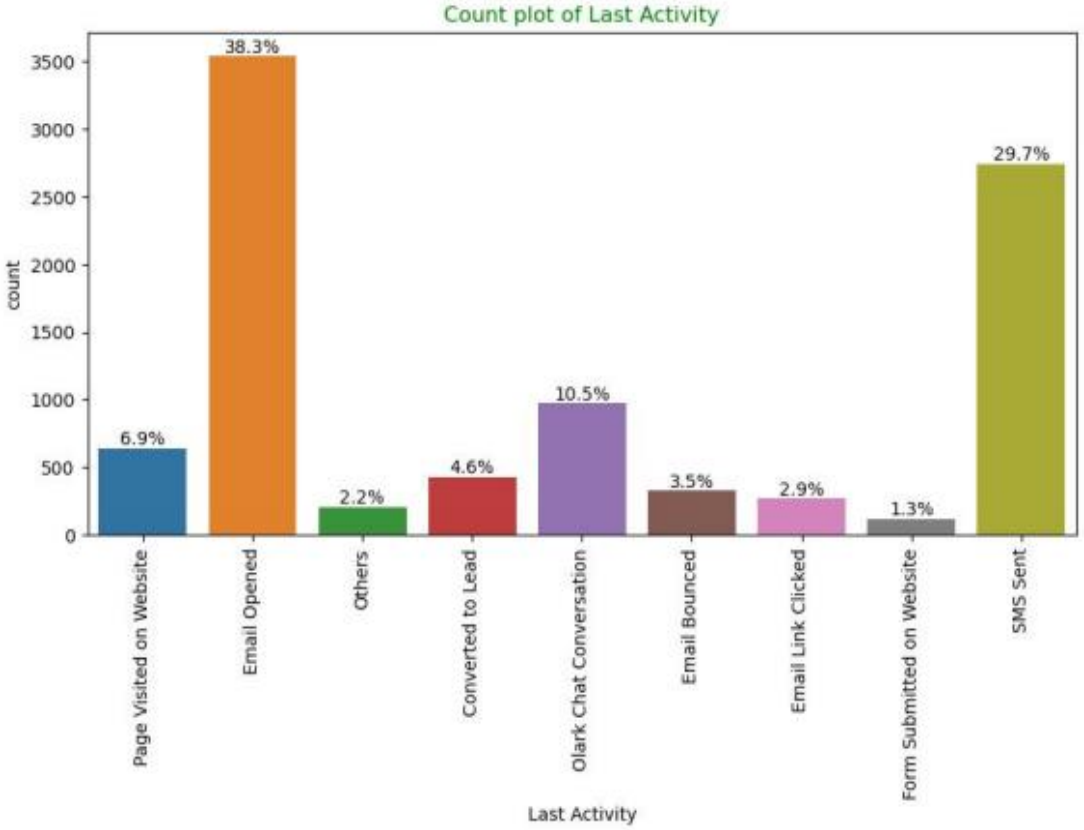
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority).

- While 61.5% of the people didn't convert to leads. (Majority).

EDA

Univariate Analysis – Categorical Variables

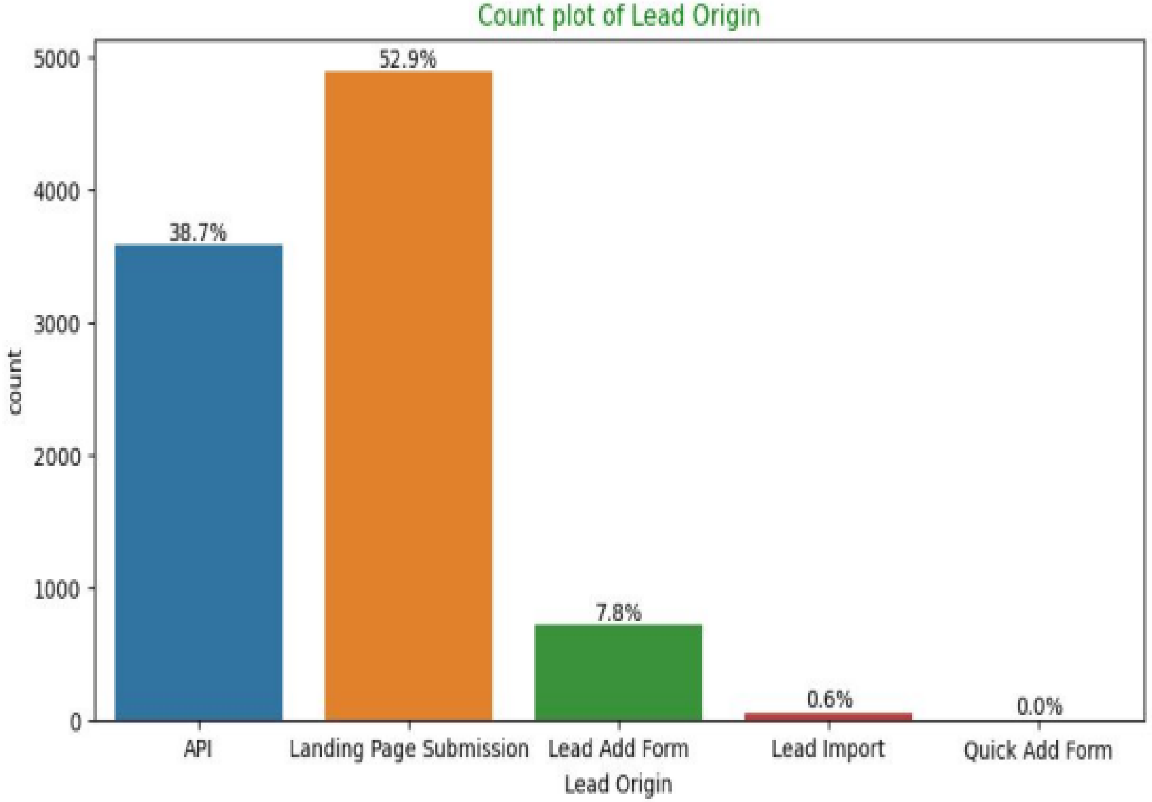


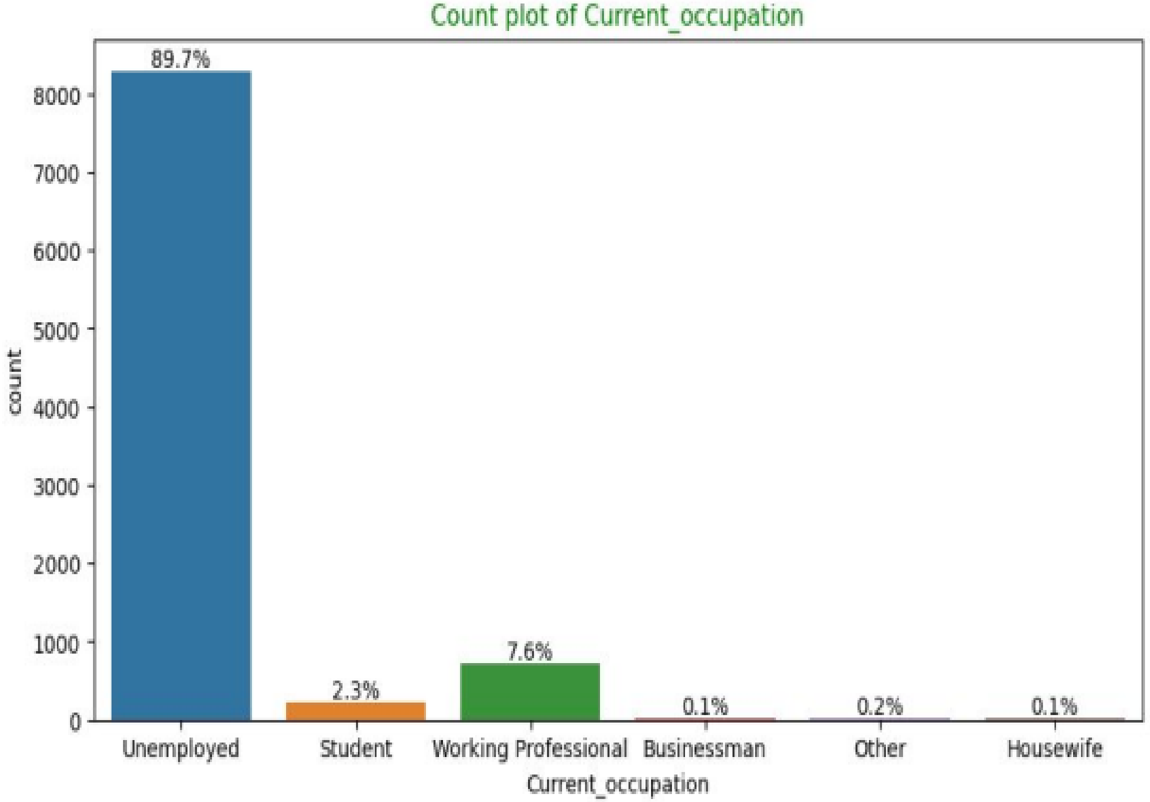


- Lead Source: 58% Lead source is from Google & Direct Traffic combined.
- Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities.

EDA

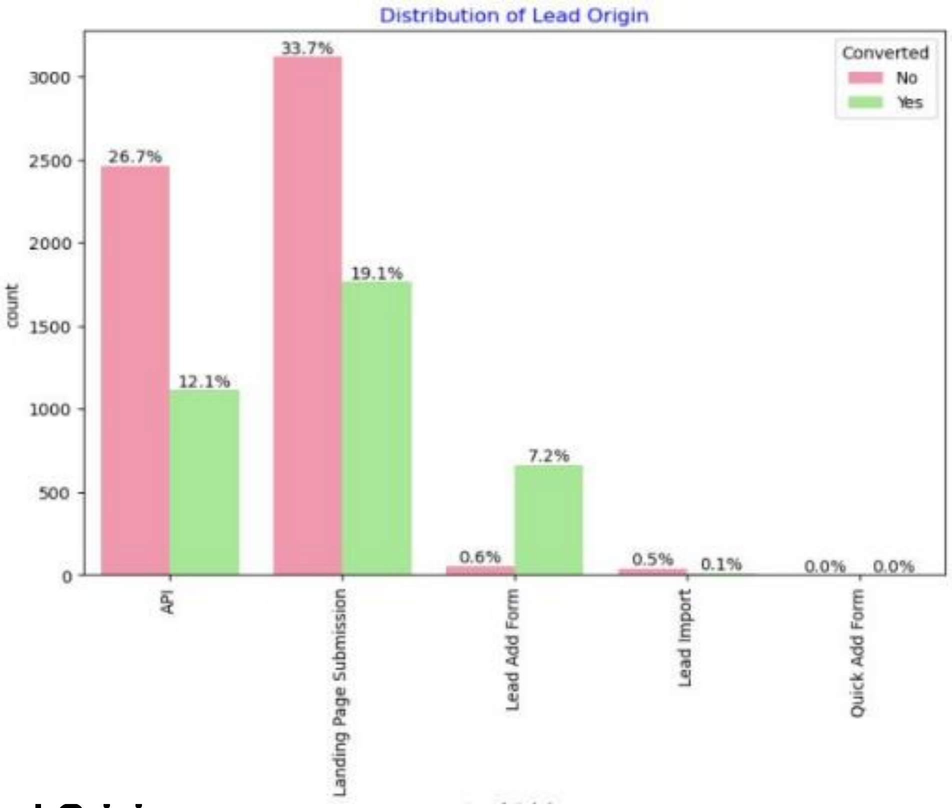
Univariate Analysis – Categorical Variables

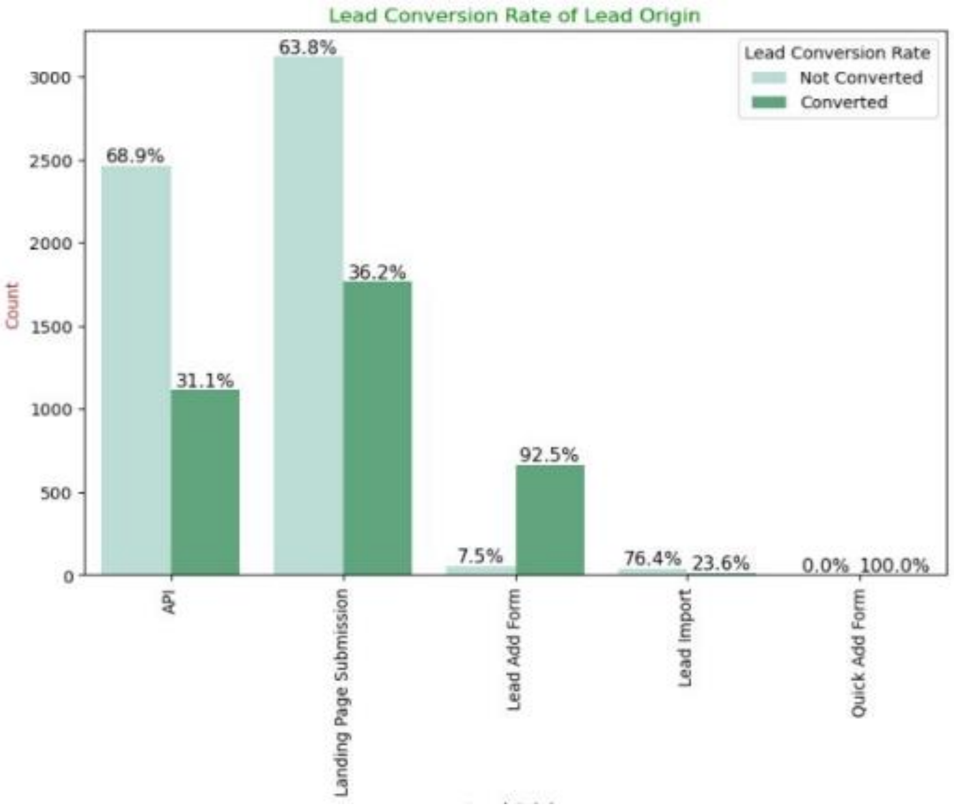




- Lead Origin: "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- Current_occupation: It has 90% of the customers as Unemployed.

EDA – BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

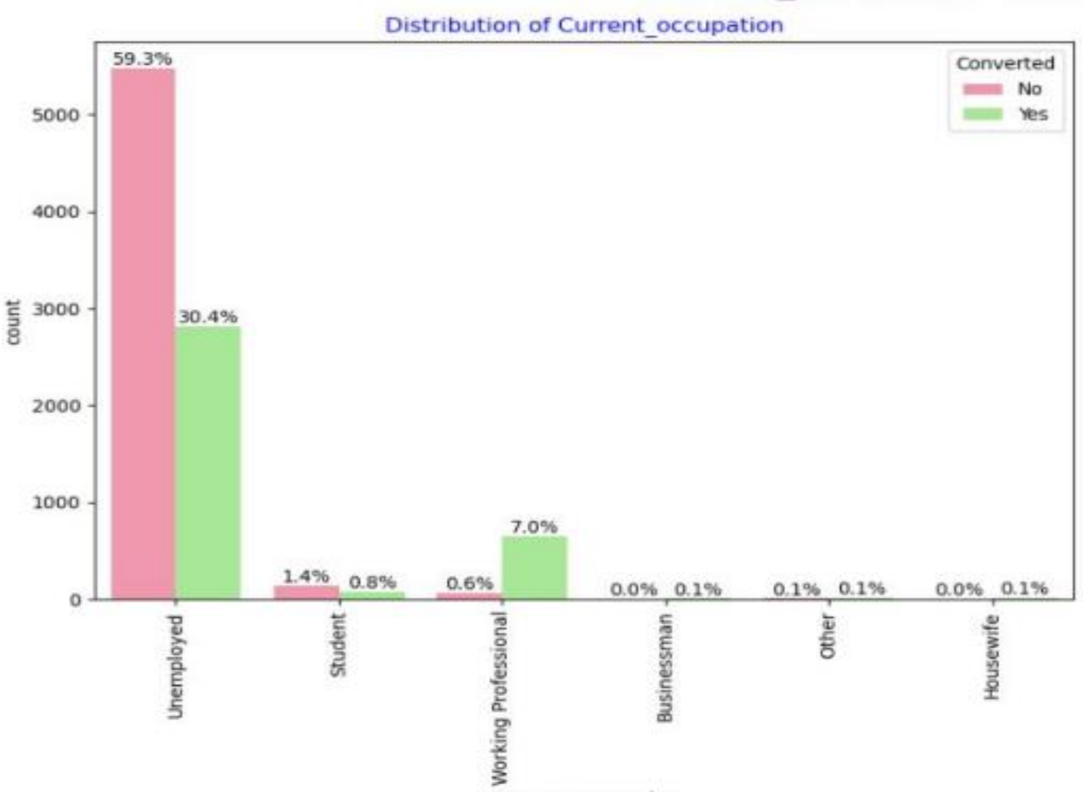


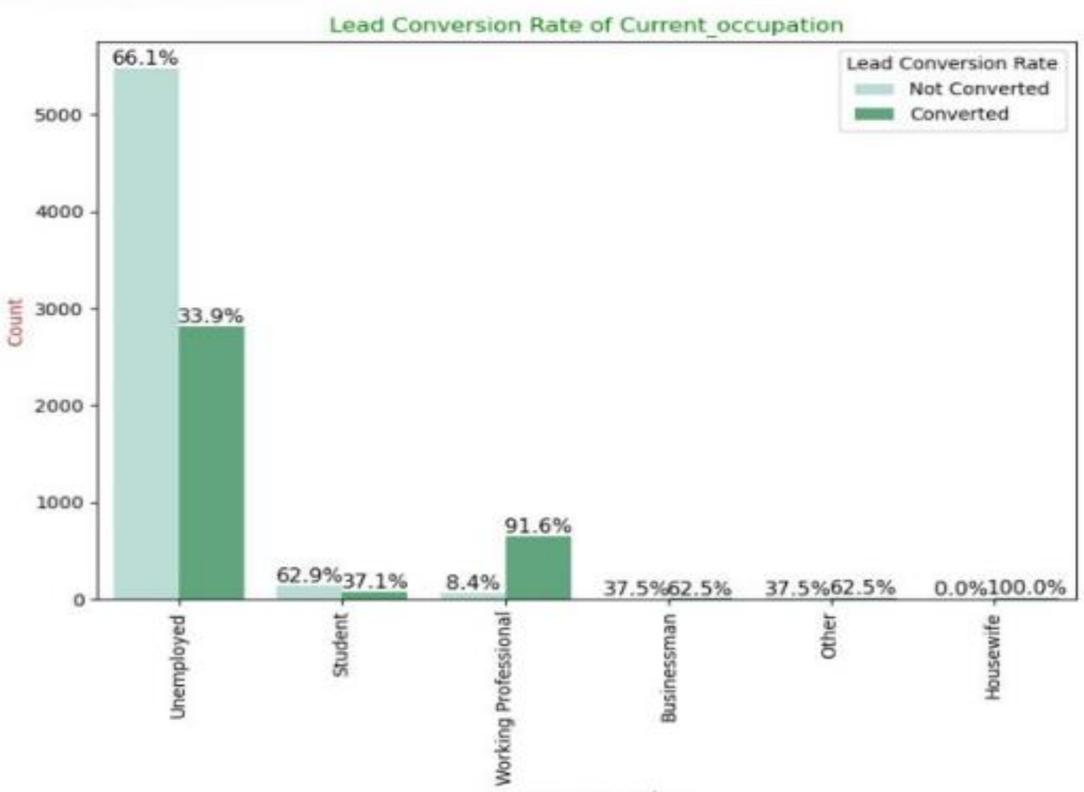


Lead Origin:

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

EDA – BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

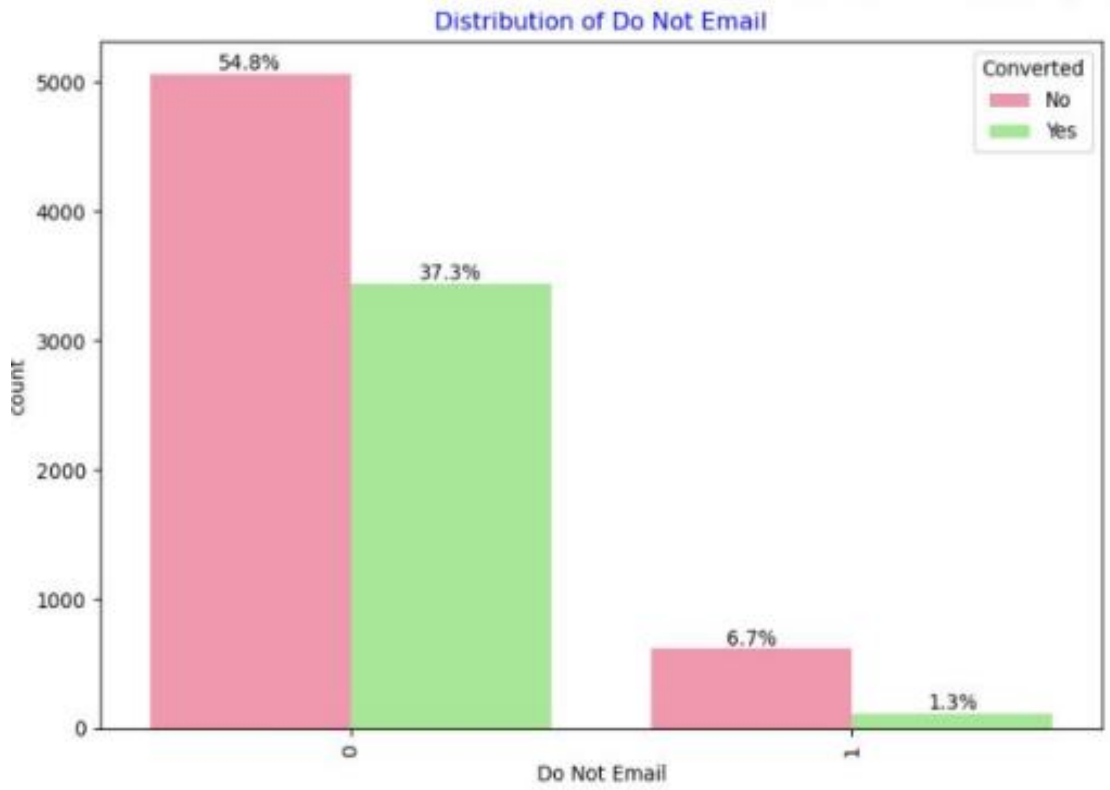


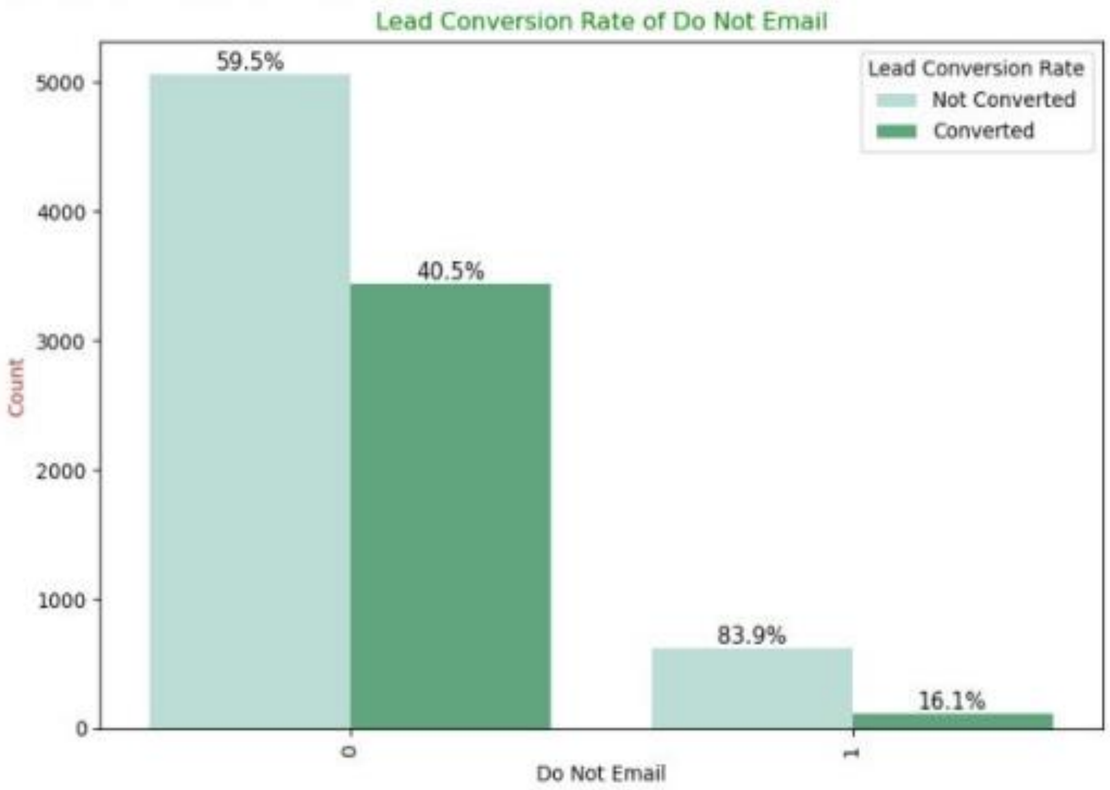


Current_occupation:

- 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 FOR CATEGORICAL VARIABLES

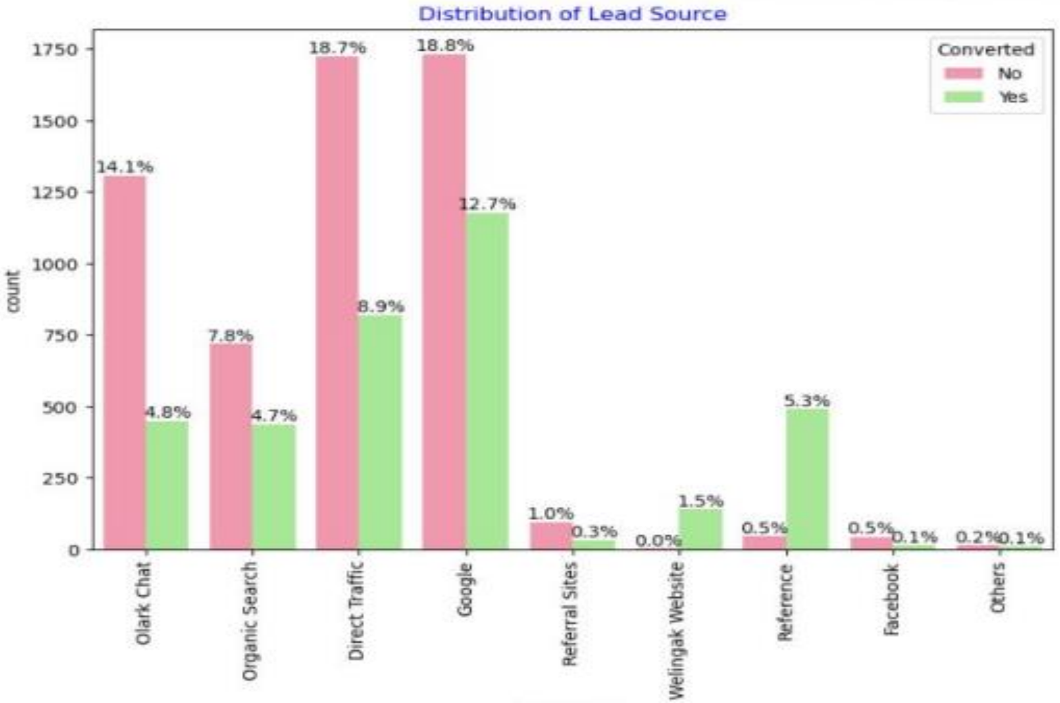


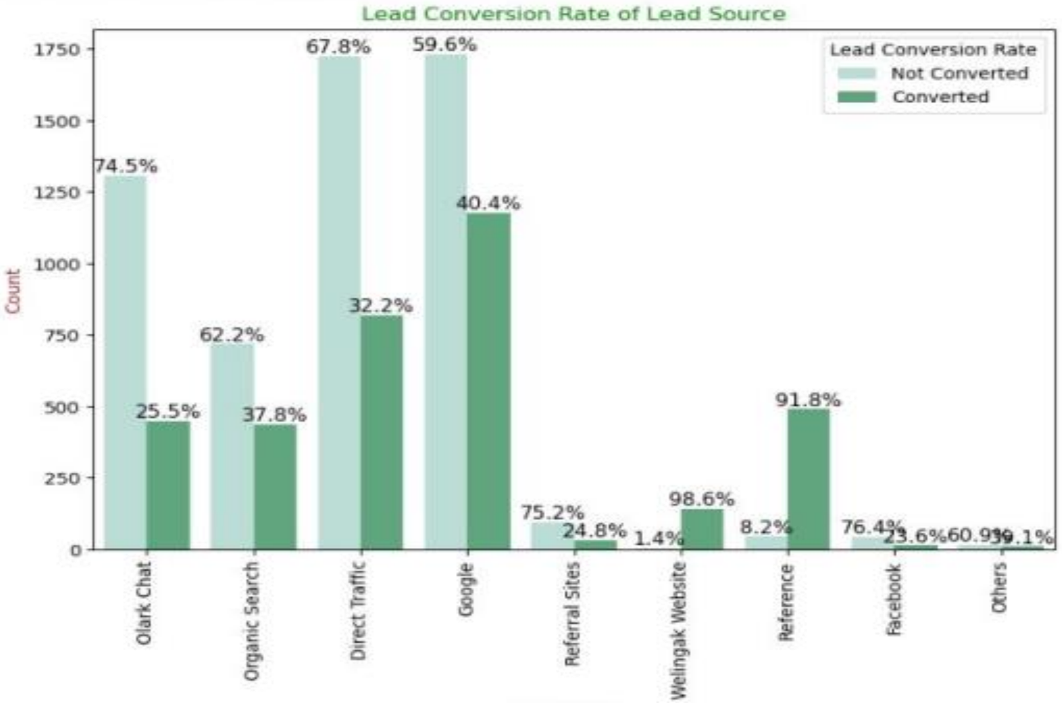


Do Not Email:

- 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

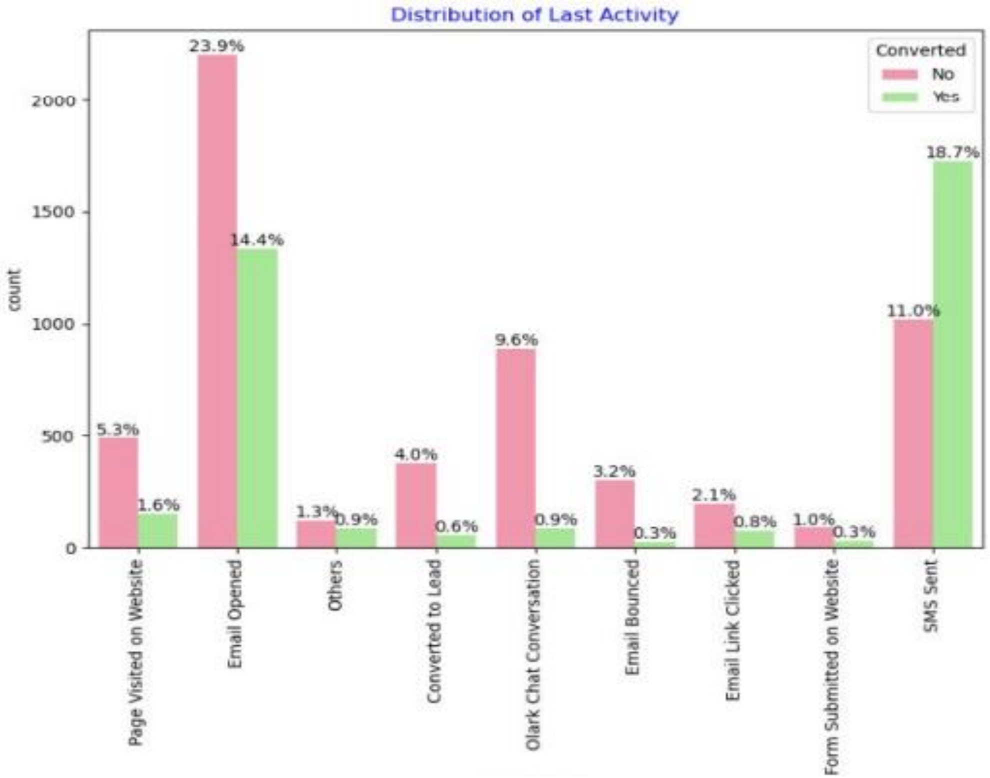


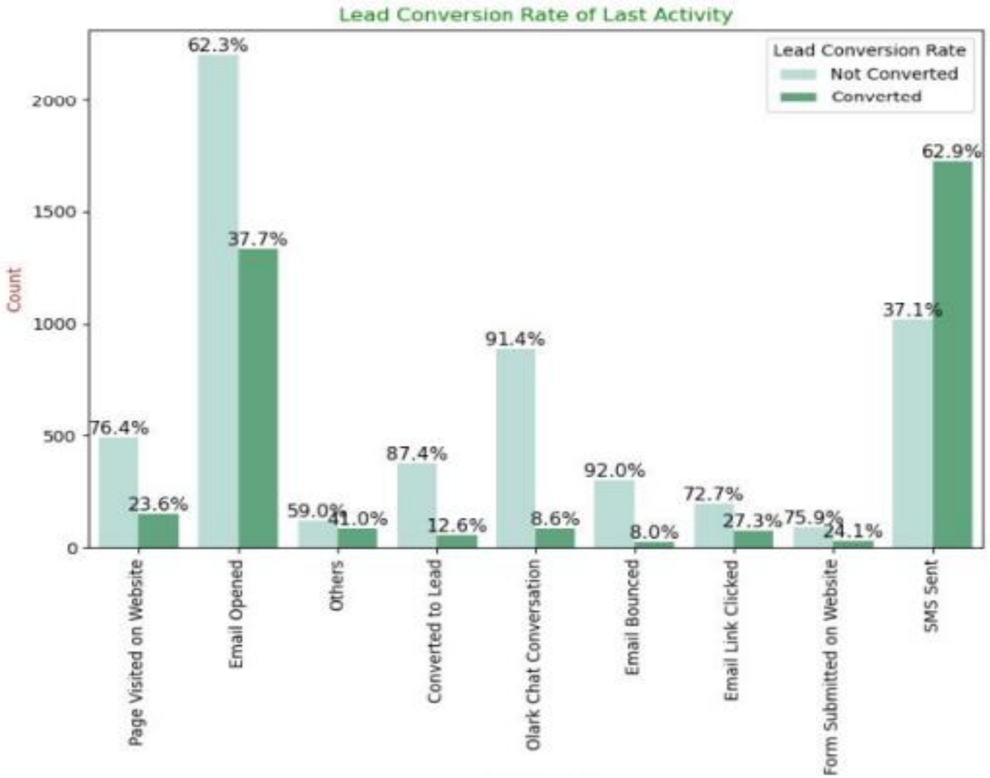


Lead Source:

- Google has 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

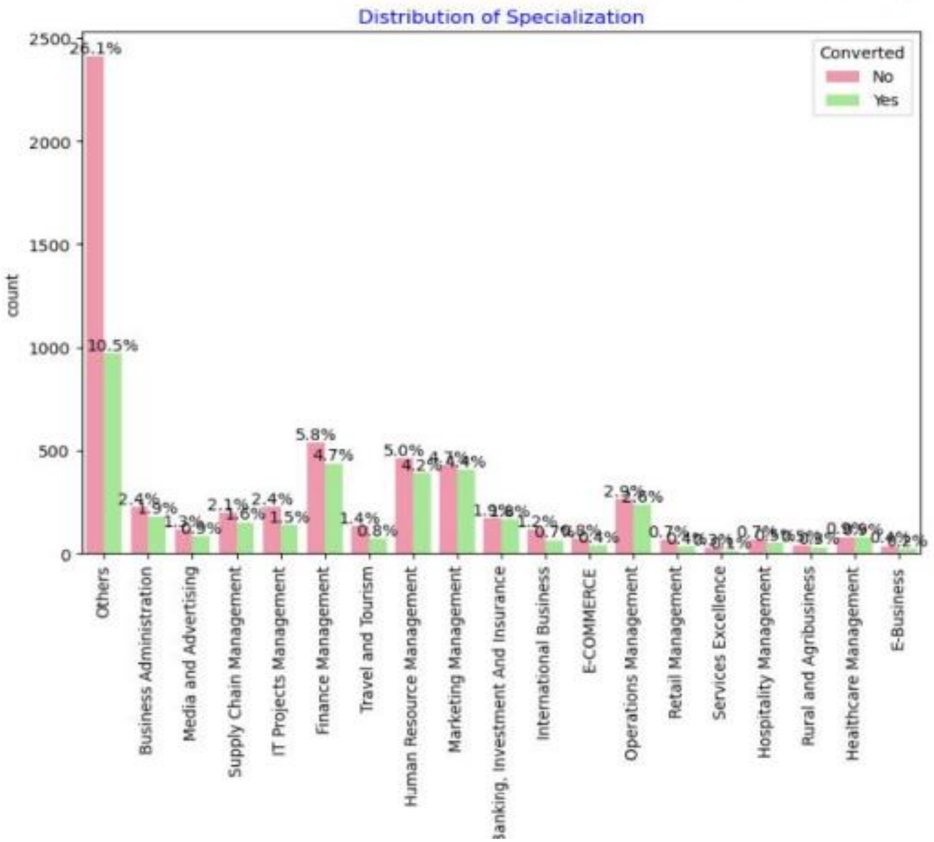


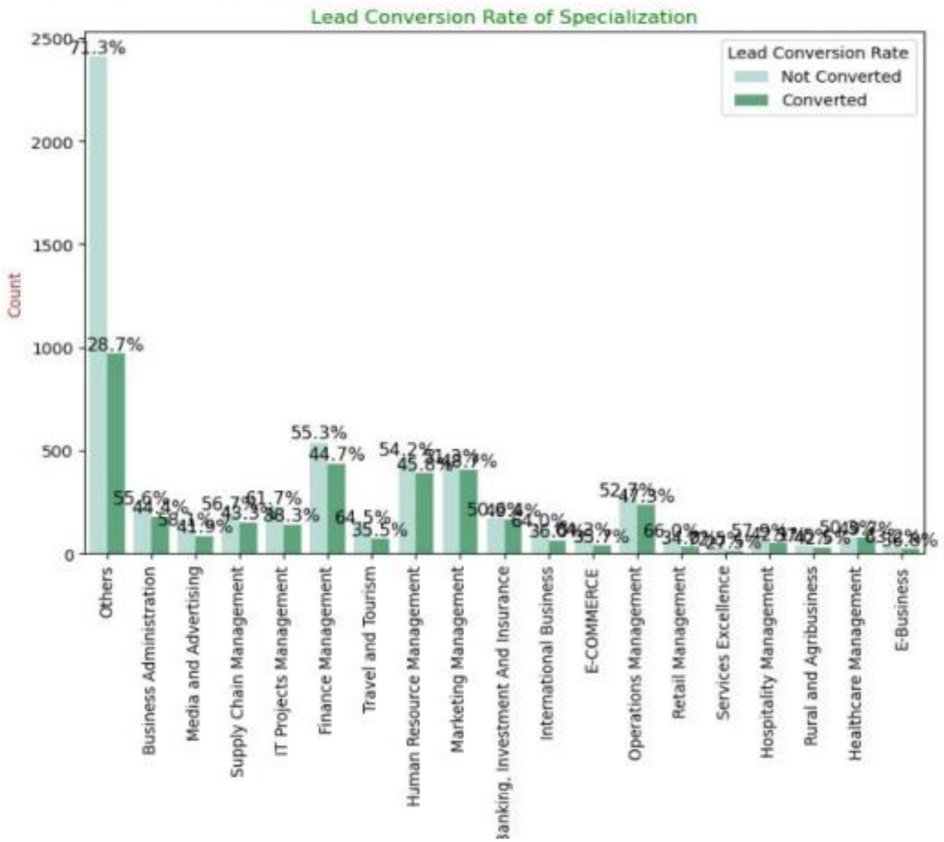


Last Activity:

- '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.

EDA – BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

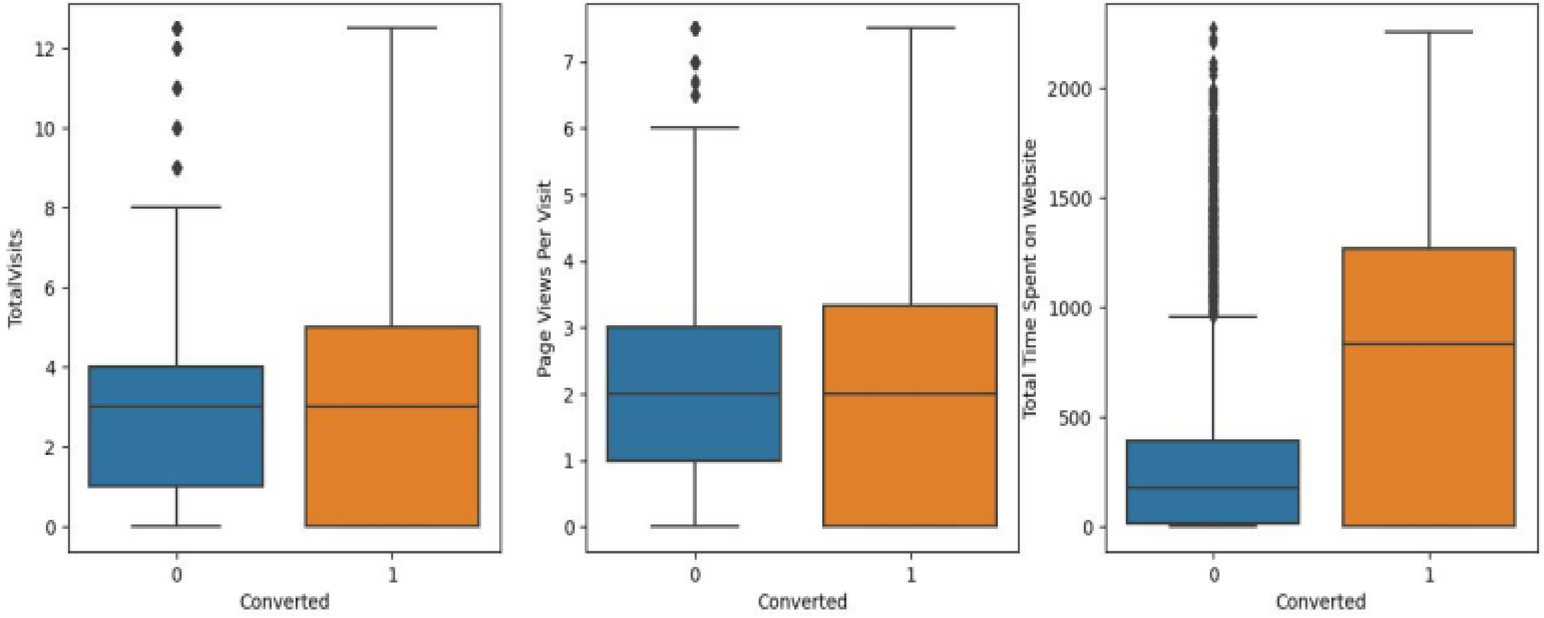


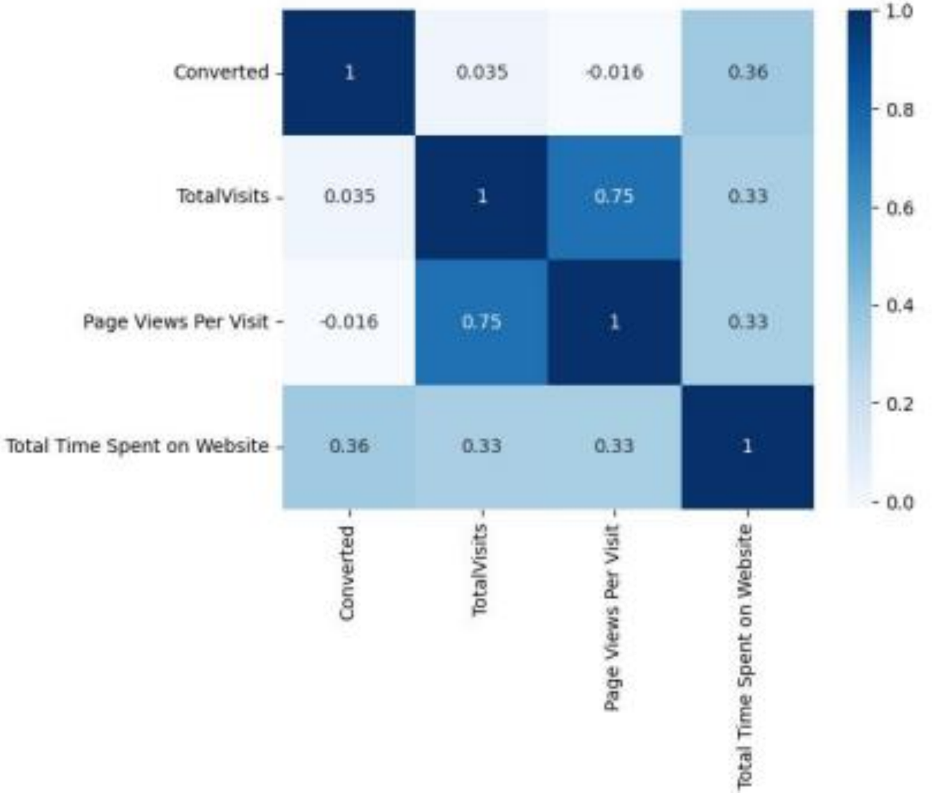


Specialization:

- Marketing Management, HR Management, and Finance Management contribute more to Lead conversion than other specializations.

EDA – BIVARIATE ANALYSIS FOR NUMERICAL VARIABLES





- Past leads who spend more time on the website have a higher chance of being successfully converted than those who spend 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:

- A 70:30 % ratio was chosen for the split.

Feature scaling:

- The Standardization method was used to scale the features.
- Checking the correlations ○ Predictor variables that 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 dimensions and a 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

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.
- 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: 5.44

• Lead Source_Reference: 2.91

• Last Activity_SMS Sent: 2.20

• Current_occupation_Working Professional: 2.10

• Last Activity_Others: 1.405963

• Total Time Spent on Website: 1.09

• Last Activity_Email Opened: 1.05

• Lead Source_Olark Chat: 0.8

MODEL BUILDING

- 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 iterations 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.

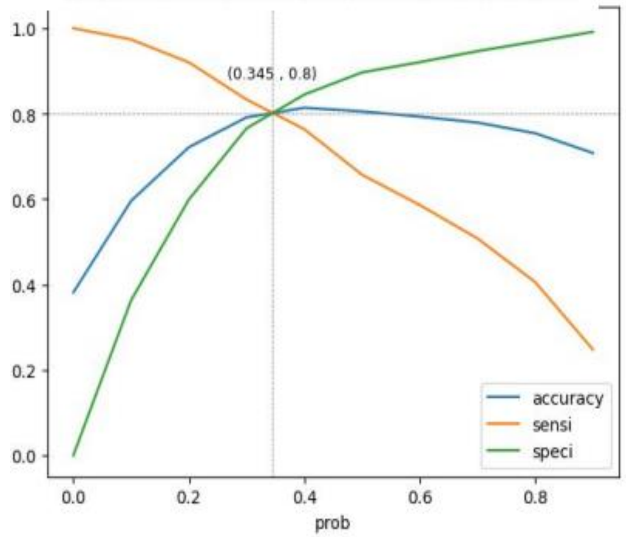
MODEL EVALUATION

Train Data Set

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

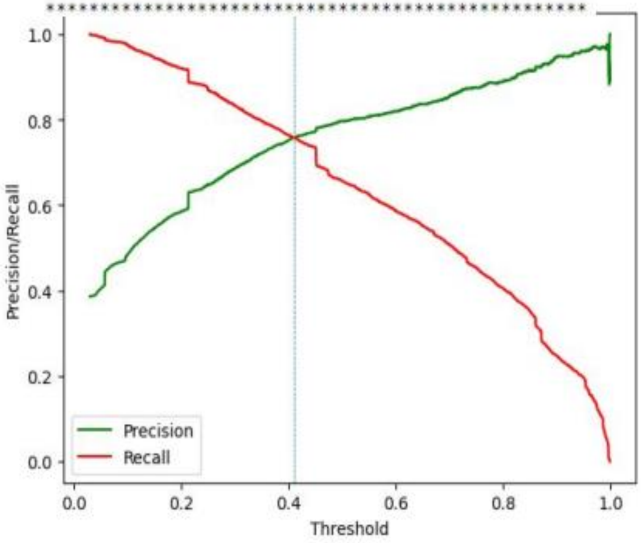
Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

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

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

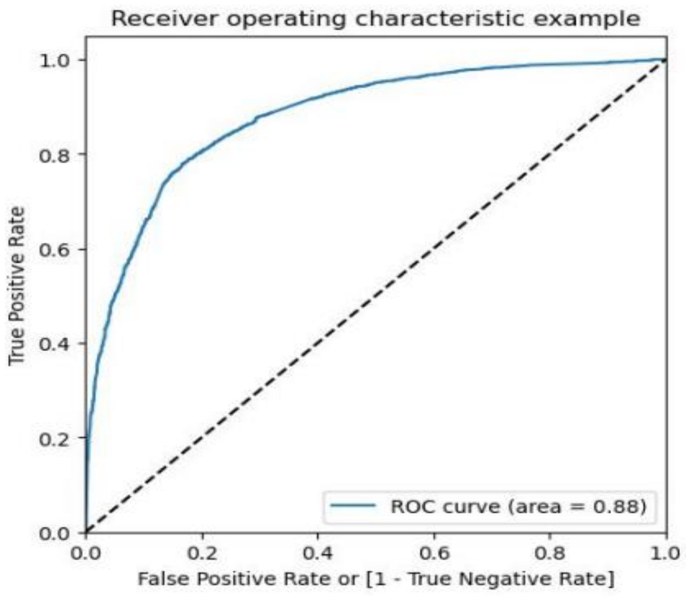
```
**************************************************

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

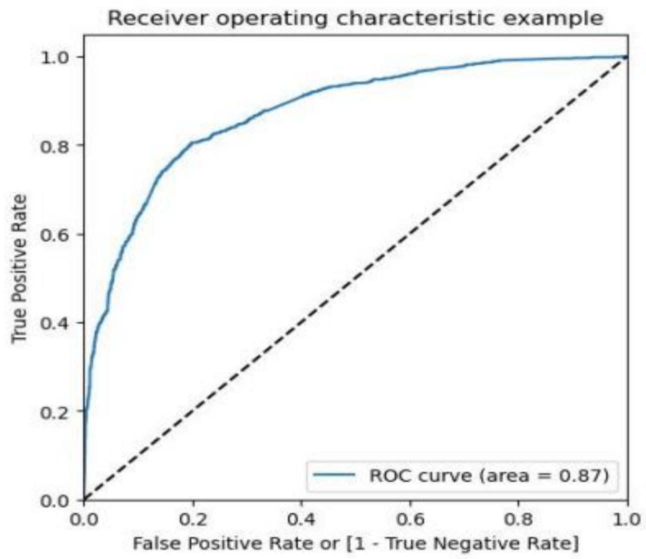
ROC Curve – Train Data Set

- The 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

- The 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

Train Data Set

```
Confusion Matrix
[[3408  594]
 [ 598 1868]]

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

True Negative                       :  3408
True Positive                       :  1868
False Negative                      :  598
False Positve                       :  594
Model Accuracy                      :  0.8157
Model Sensitivity                   :  0.7575
Model Specificity                   :  0.8516
Model Precision                     :  0.7587
Model Recall                        :  0.7575
Model True Positive Rate (TPR)      :  0.7575
Model False Positive Rate (FPR)     :  0.1484
```

Test Data Set

```
Confusion Matrix
[[1351  326]
 [ 222  873]]

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

True Negative                       :  1351
True Positive                       :  873
False Negative                      :  222
False Positve                       :  326
Model Accuracy                      :  0.8023
Model Sensitivity                   :  0.7973
Model Specificity                   :  0.8056
Model Precision                     :  0.7281
Model Recall                        :  0.7973
Model True Positive Rate (TPR)      :  0.7973
Model False Positive Rate (FPR)     :  0.1944
```

- Using a cut-off value of 0.345, the model achieved a sensitivity of 75.75% in the train set and 79.73% in the test set.
- Sensitivity in this case indicates how many leads the model identifies correctly out of all potential leads that are converting.
- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 80.23%, which is in line with the study's objectives.

We have also identified features with negative coefficients that may indicate potential areas for improvement.

These include:

- Specialization in Hospitality Management: -1.10
- Specialization in Others: -1.21
- Lead Origin of Landing Page Submission: -1.25

RECOMMENDATION BASED ON FINAL MODEL

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.
- Optimize communication channels based on lead engagement impact.
- Engage working professionals with a tailored message.
- More budget/spending can be done on the Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing references that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have a high conversion rate and will have a 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