Table of Contents

Background of X Education Company

- 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.
- 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 and typical lead conversion rate is around 30%.

Problem Statement & Objective of the Study

- X Education gets a lot of leads, its lead conversion rate is very low around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, known as Hot Leads
- sales team want to know these potential leads, to focus more on them rather than making calls to everyone.

Objective of the Study:

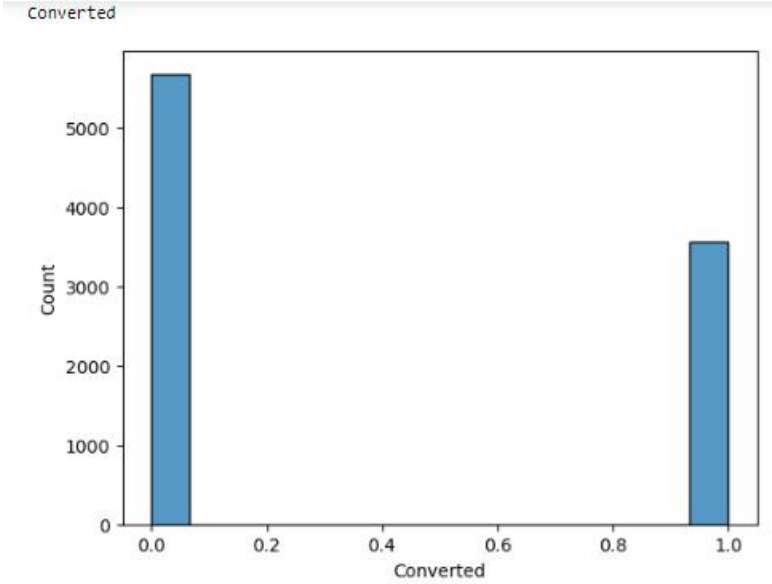
- To help X Education select the most promising leads, which are most likely to be converted
- It is required to build a model wherein we need to assign a lead score to each of the leads based on the conversion chances, customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The benchmark target provided for the lead conversion rate to be around 80%.
  - Since we have a target of 80% conversion rate, we would want to obtain a Good sensitivity / Recall in obtaining hot leads

Data Cleaning

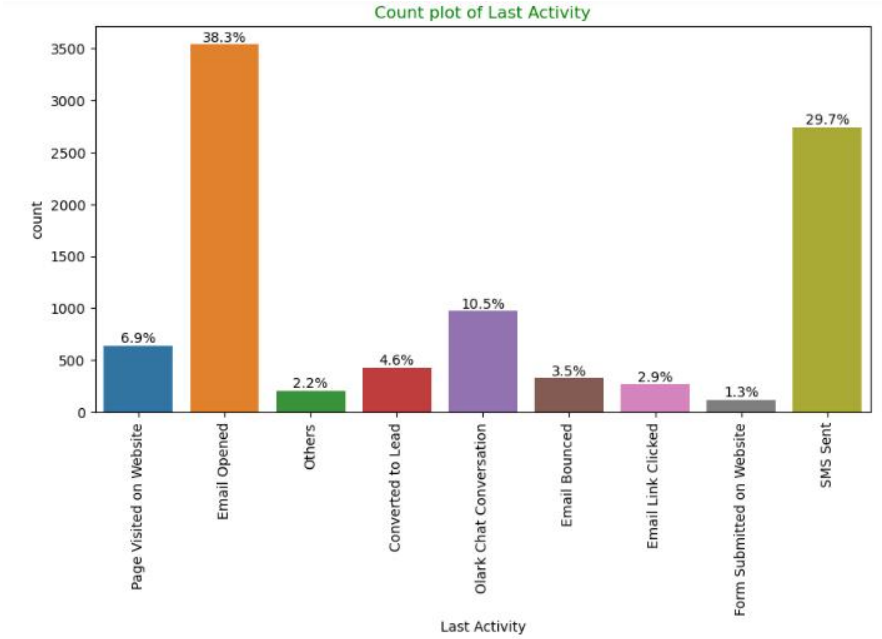
- To start with all the "Select" values represents null values for the categorical variables, as customers did not choose any option.
- Columns with 40% null values are dropped.
- Drop columns that don't add any insight on the analysis
- Imputation was used for some categorical variables by its MOD.
- Additional categories were created for some variables.
- Numerical data was imputed with MODE
- Invalid values were fixed and data was standardized
- Low frequency values were grouped together to “Others”.

EDA

Univariate Analysis



Last Activity: 38% and 30 of customers are in Email Opened and SMS sent activities

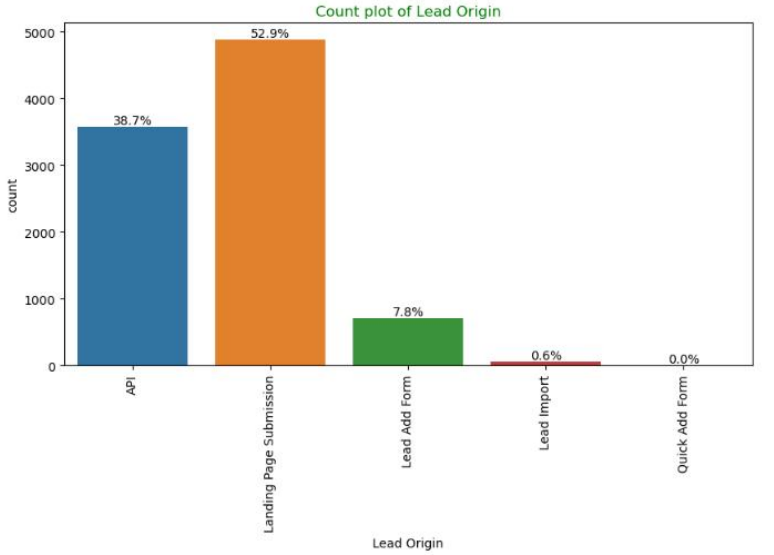


Conversion rate is of approx 40% and Non conversation rate is 60%.

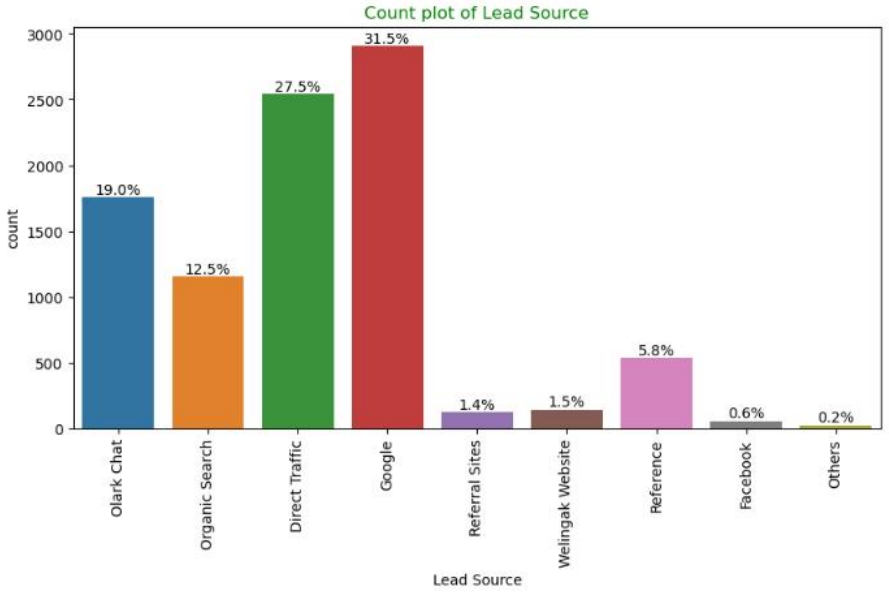
Data is highly imbalance

EDA

Univariate Analysis – Categorical



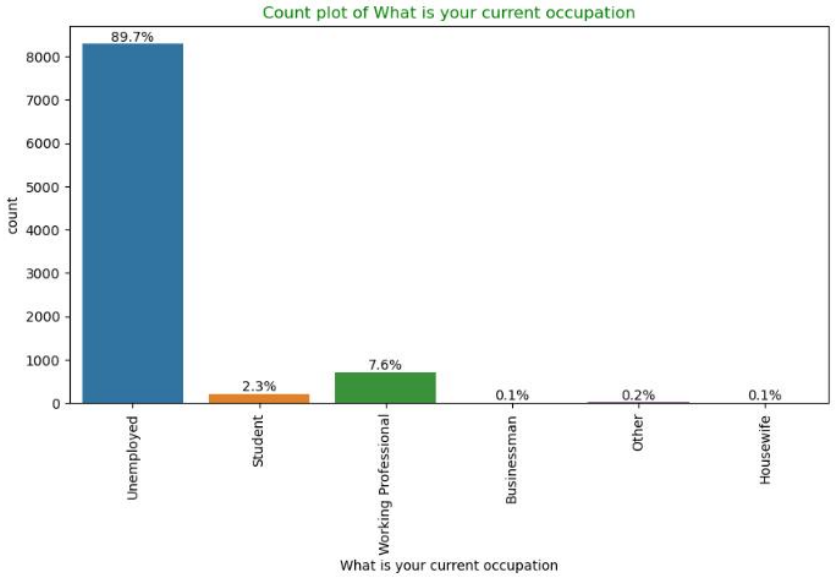
Lead Source : Has highest percentage of 31.5 % in google



Lead Origin: "Landing Page Submission" are 53% customers and "API" are 39%.

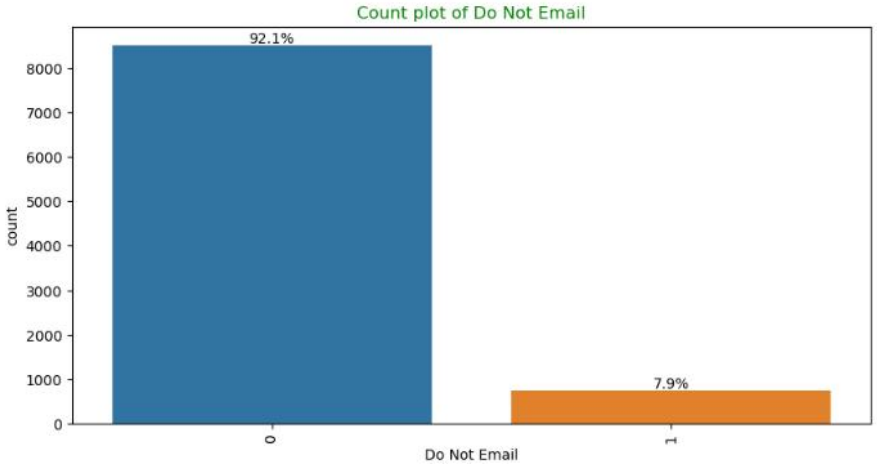
EDA

Univariate Analysis – Categorical



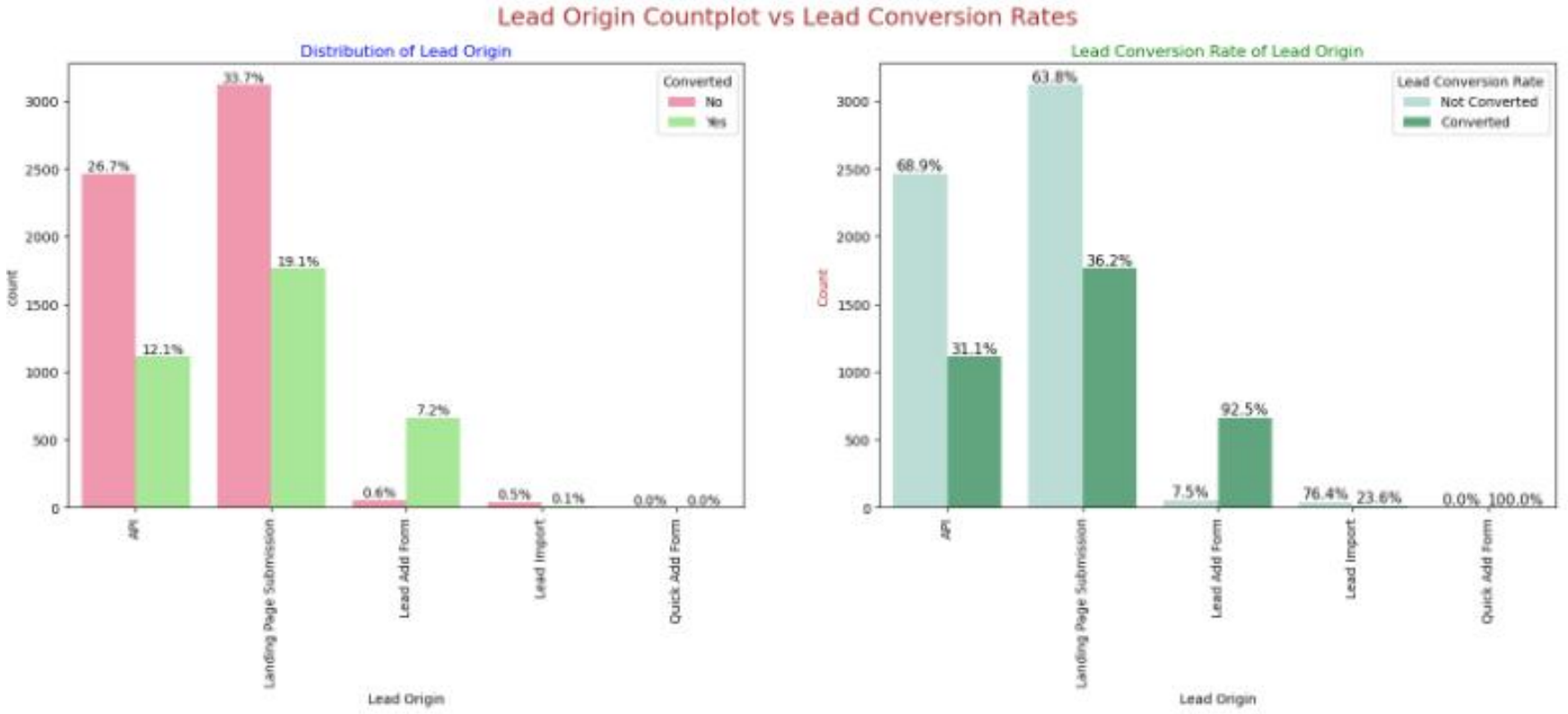
Current occupation: It has 90% of the customers as Unemployed

Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.



EDA

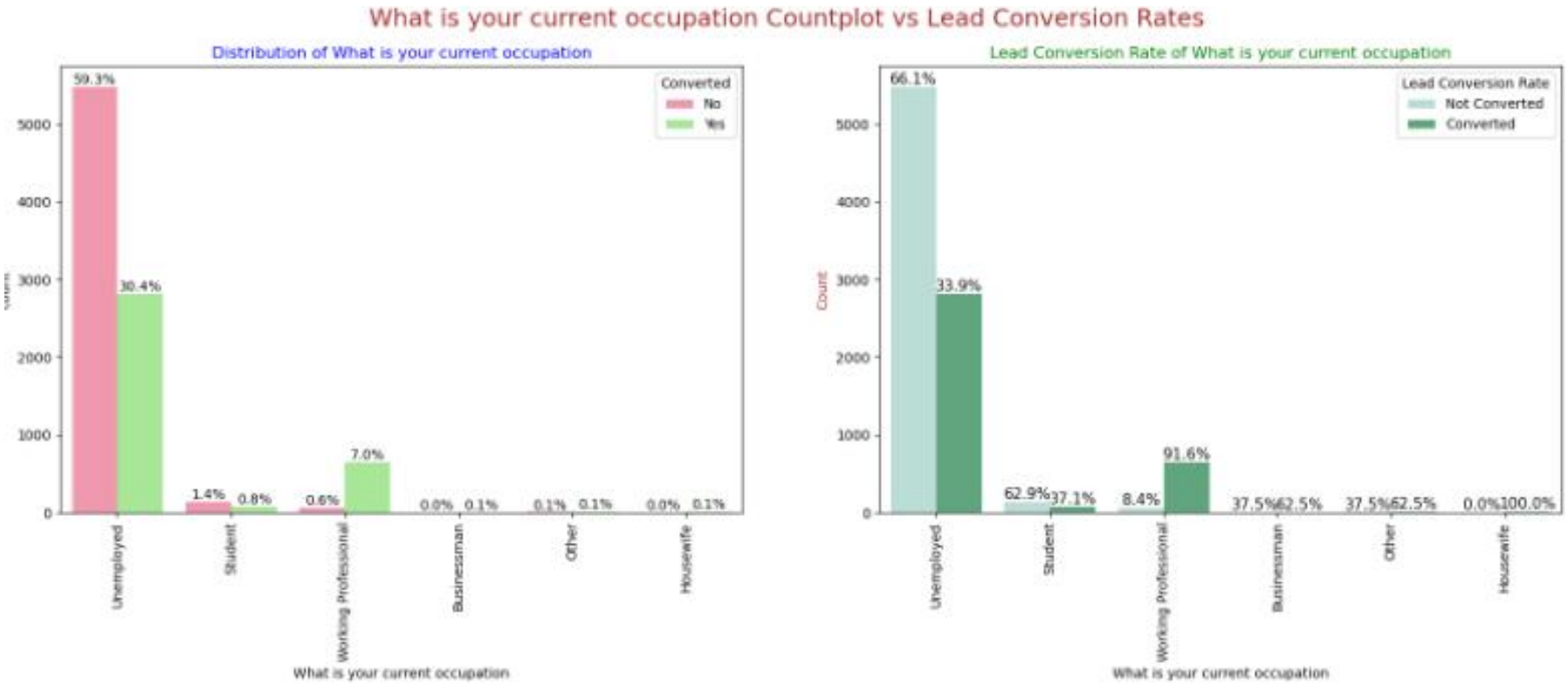
Bivariate Analysis – Categorical



Lead Origin: Approx 53% of leads originated from "Landing Page Submission". Lead conversion rate is 36%.

EDA

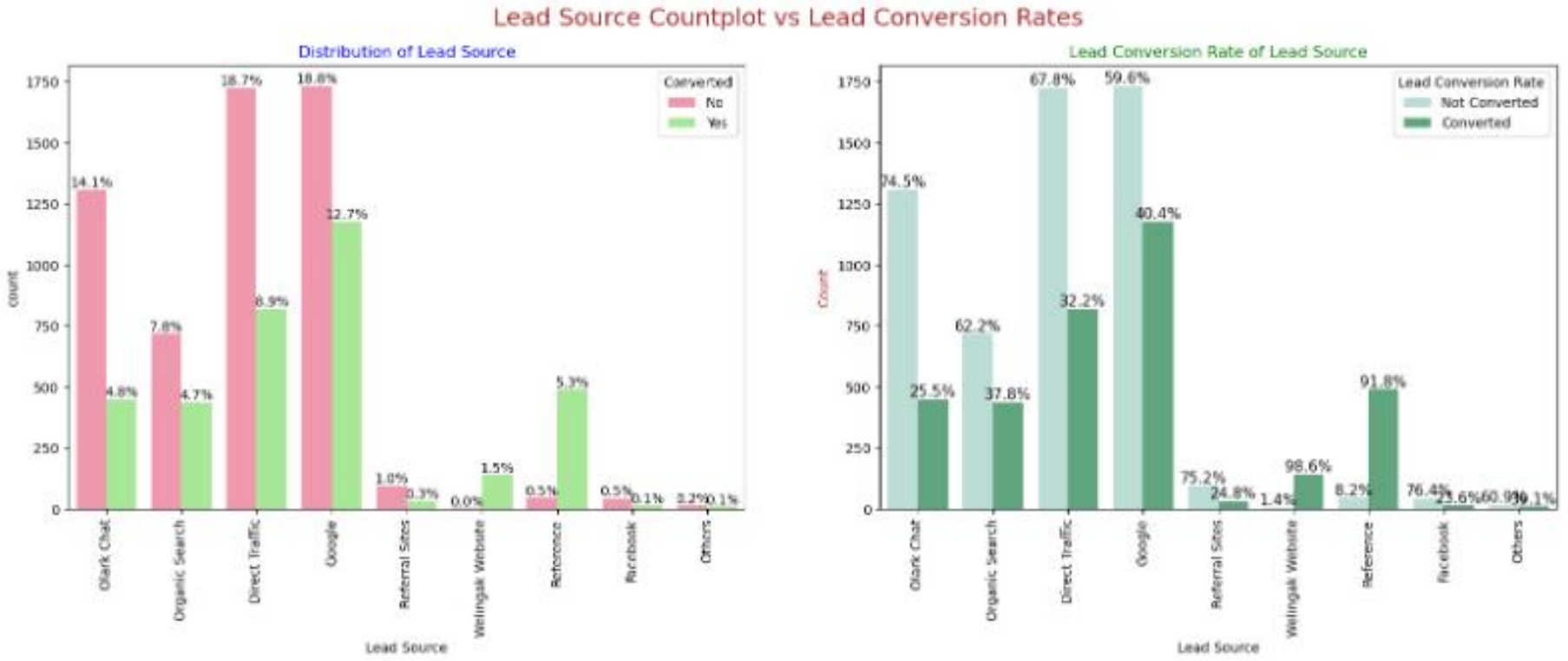
Bivariate Analysis – Categorical



Current_occupation: Approx 90% of the customers are Unemployed with lead conversion rate of 34%.

EDA

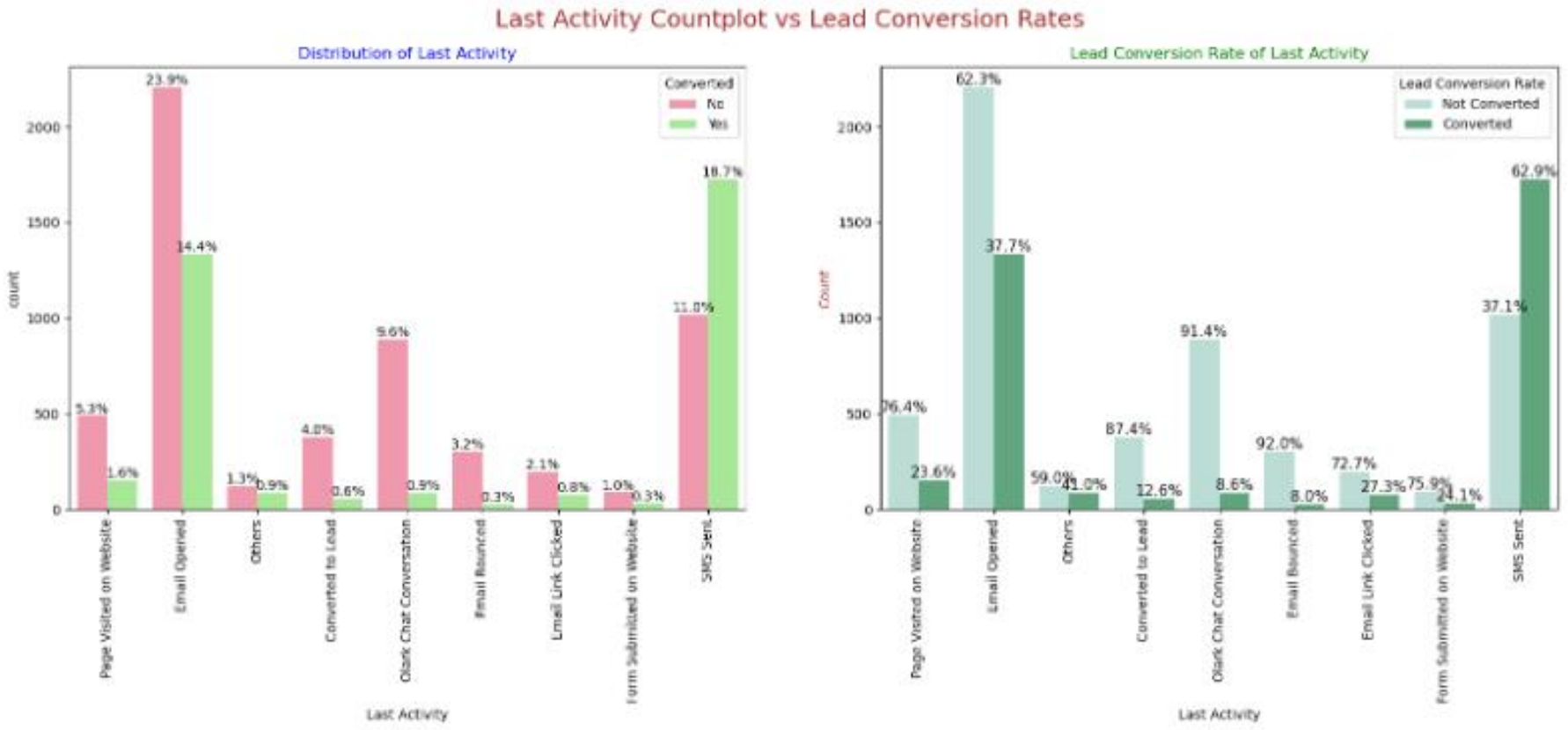
Bivariate Analysis – Categorical



Lead Source: Google has conversion rate of 40% out of 31% customers which is highest

EDA

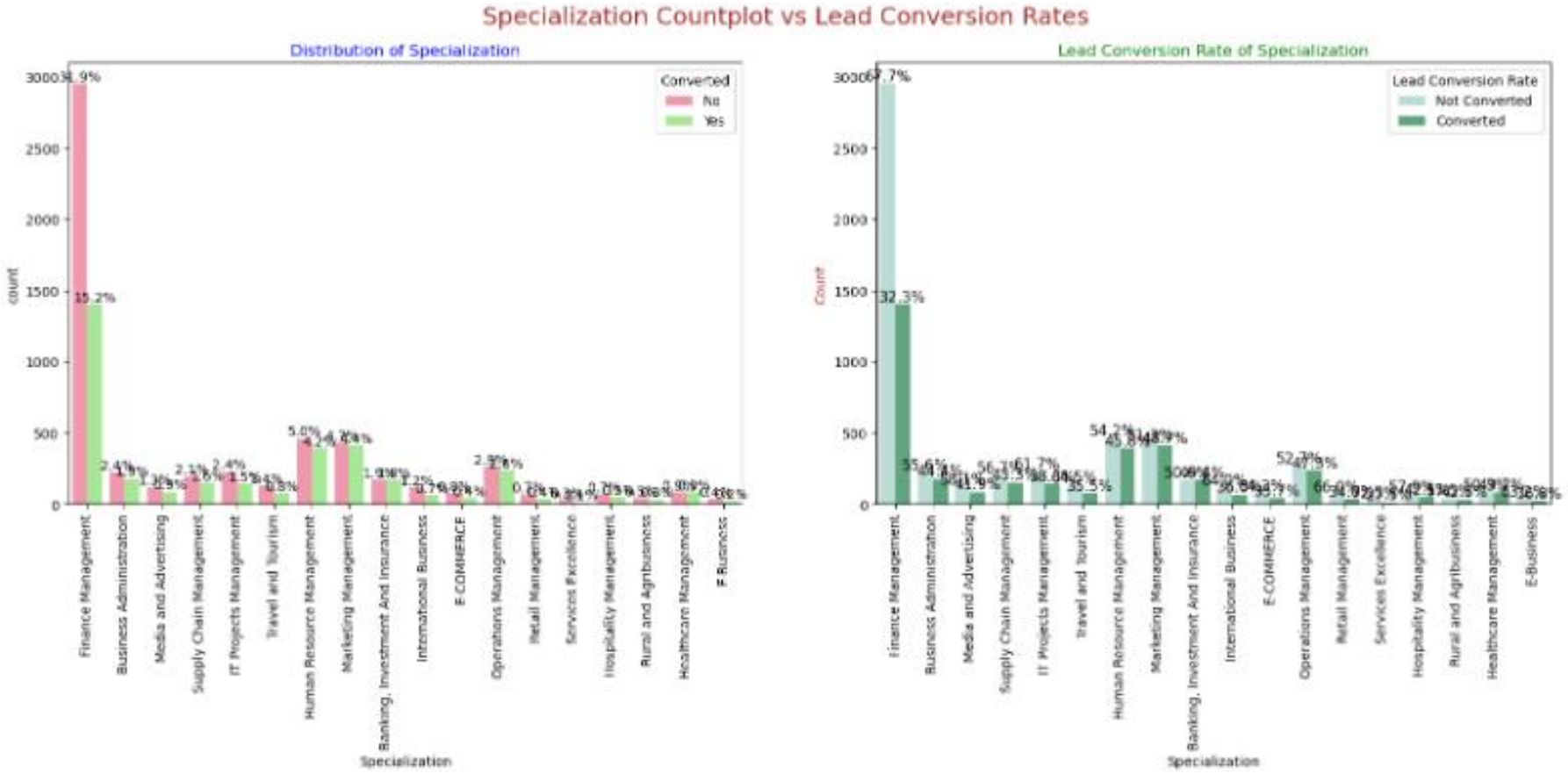
Bivariate Analysis – Categorical



Last Activity: 'SMS Sent' has high lead conversion rate of 63%

EDA

Bivariate Analysis – Categorical



Specialization: Finance Management shows good convertion rate pattern with 32%.

Data Preparation for model building

- Created dummy features for categorical variables
  - Test Train split is done in 70:30 % ratio
- Feature scaling is done for continuous variables using standard scaler
  - highly correlated with each other were dropped

Model Building

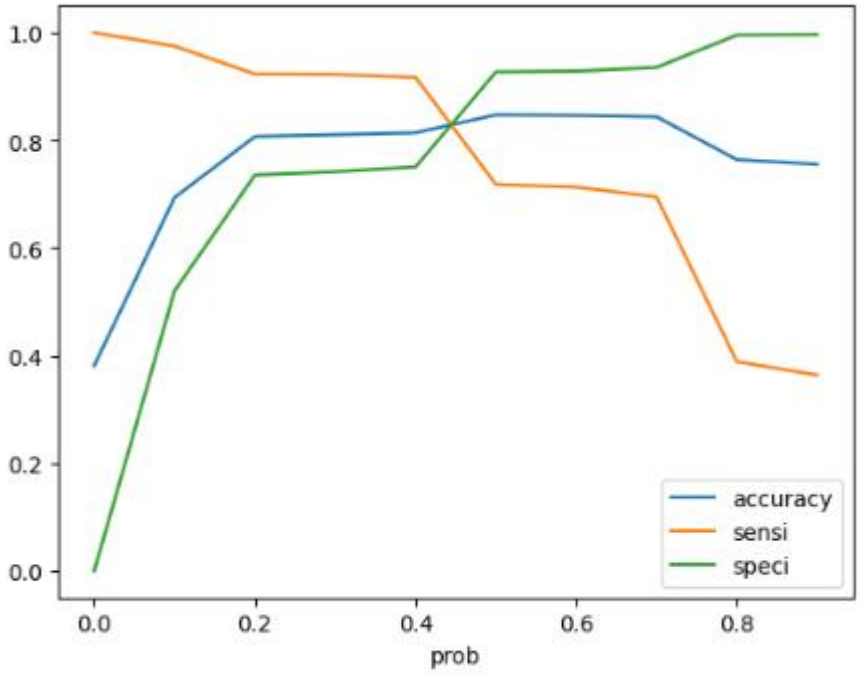
Recursive Feature Elimination (RFE) method is used to select the most important 15 features

- Manual Feature selection process was used to build models by dropping variables with p – value
greater than 0.05 and with VIFs less than 5

- After five iteration we arrived at the stable model :
- model 5 is used for Model Evaluation which further and make predictions

Model Evaluation

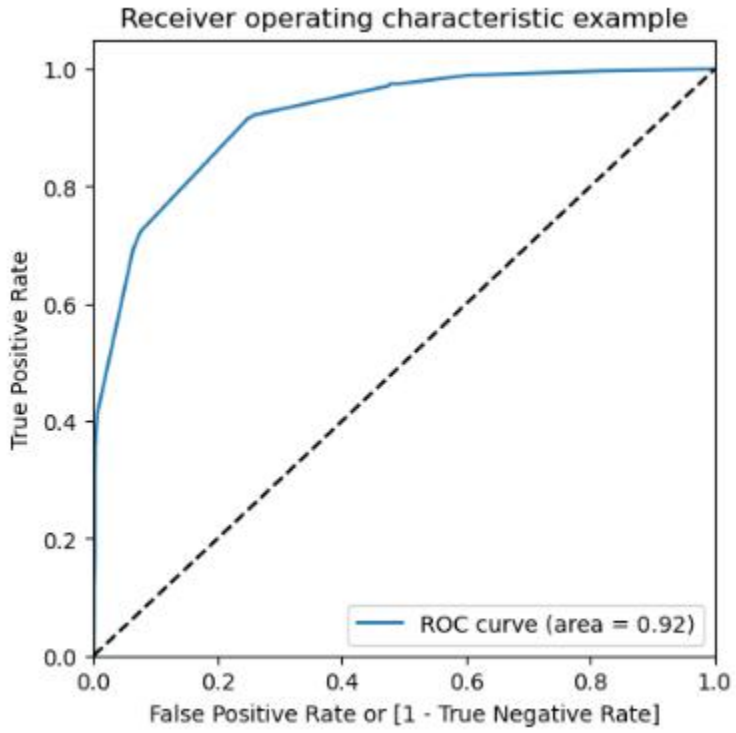
0.41 is the approx. point where all the curves meet, so 0.41 seems to be our Optimal cutoff point for probability threshold .



Model Evaluation

Area under ROC curve is 92% 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

Conclusion

Train Data Set: Accuracy: 81.38%

Sensitivity: 91.68%

Specificity: 75.03%

Test Data Set: Accuracy: 81.02%

Sensitivity: 92.32%

Specificity: 73.64%

Using a cut-off value of 0.41, the model achieved a sensitivity of ~92% in the train and test set

- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 81%, meeting the study's objectives

Final Recommendations

Following features that have the highest positive coefficients, and these should be treated as priority to increase lead conversion.

Lead Origin_Lead Add Form	2.04
Lead Source_Welingak Website	1.50
Last Activity_Email Opened	1.44
Last Activity_Others	1.61
Last Activity_SMS Sent	2.88
What is your current occupation_Unemployed	-1.45
What is your current occupation_Working Professional	1.63
Tags_Busy	3.48
Tags_Closed by Horizzon	9.32
Tags_Lost to EINS	8.61
Tags_Will revert after reading the email	3.79
Tags_in touch with EINS	3.00

Final Recommendations

Below points should be noted to increase the Lead Conversion Rates

- Top 3 features are Tags_Closed by Horizzon , Tags_Lost to EINS and Tags_Will revert after reading the email . Team should rigourously focus on this category
- 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 tailored messaging.
- 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

To identify areas of improvement

- Analyze negative coefficients in some offers.

Final Recommendations

Below points should be noted to increase the Lead Conversion Rates

- Concentrate on leveraging features with positive coefficients to enhance targeted marketing strategies.
- Fine-tune communication channels based on their impact on lead engagement.
- Tailor messaging to effectively engage working professionals.
- Allocate a larger budget for advertising.
- Implement incentives or discounts for successful lead-generating references, encouraging more referrals.
- Pursue an aggressive targeting approach towards working professionals
- Evaluate and refine the landing page submission process for potential improvements.