upGrad

X EDUCATION - LEAD SCORING CASE STUDY

By-
1. Prabhavathi Nunna
2. Babita Kumari
3. K Naga Vishala

upGrad

Table of Content

upGrad

Background of X Education Company

- X Education, an online education company, caters to industry professionals by offering courses. The company attracts potential customers who visit its website daily, engaging with courses.
- Effective marketing on various platforms, including websites and search engines like Google, drives this traffic.
- Visitors explore courses, potentially filling out forms or watching videos.
- Leads emerge when individuals provide their contact details via forms. These leads prompt the sales team to initiate contact, combining calls and emails. As a result, a portion of leads undergo conversion, while the majority do not. Notably, X Education maintains an average lead conversion rate of approximately 30%.

upGrad

Problem Statement & Objective of the Study

Problem Statement:

1. X Education's lead conversion rate is currently low at around 30%.
2. They aim to improve the lead conversion process by identifying high-potential leads (Hot Leads).
3. The sales team wants to focus more on communicating with potential leads rather than contacting everyone.

Objective of the Study:

1. Assist X Education in selecting the most promising leads with higher conversion potential.
2. Develop a model to assign a lead score to each lead.
3. The lead score should reflect the likelihood of a lead becoming a paying customer.
4. The CEO has set a target lead conversion rate of approximately 80%.

upGrad

Suggested Strategies for Lead Conversion:

Group leads based on their likelihood to convert, creating a focused set of hot leads.

Leads Grouping:

Engage with a smaller pool of leads, enabling more impactful communication.

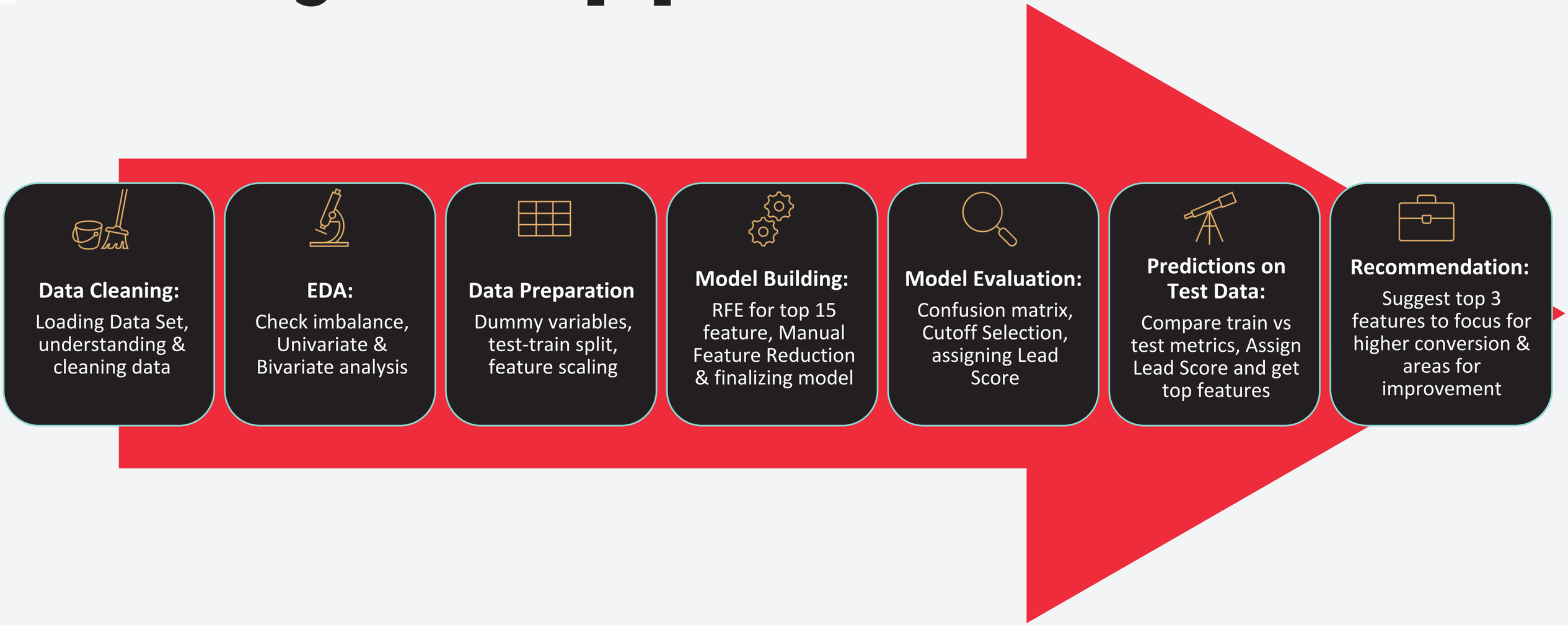
Better Communication:

Achieve a higher conversion rate by concentrating efforts on hot leads with a greater likelihood to convert.

Boost Conversion:

upGrad

Analysis Approach



upGrad

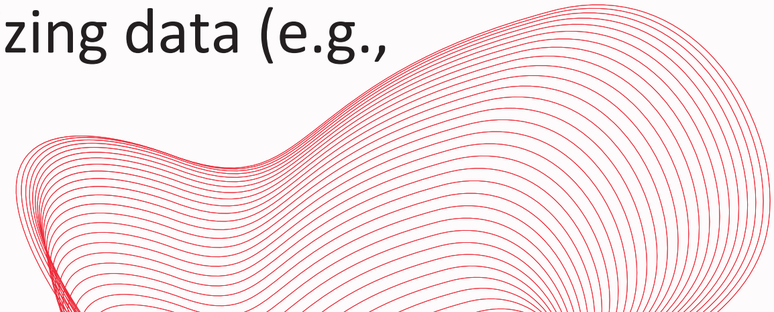
Data Cleaning:

- The "Select" level in categorical variables indicates null values resulting from customers not choosing any option.
- Columns with more than 40% null values were removed.
- Handling missing values in categorical columns involved considering value counts and certain factors.
- Columns like tags and country, which didn't contribute to the study's objective, were dropped.
- Imputation was employed for certain categorical variables.
- Additional categories were introduced for certain variables.
- Columns like Prospect ID and Lead Number, which didn't aid modeling, or had only one response category, were dropped.
- Numerical data was imputed with the mode after assessing distribution.

upGrad

Data Cleaning and Transformation:

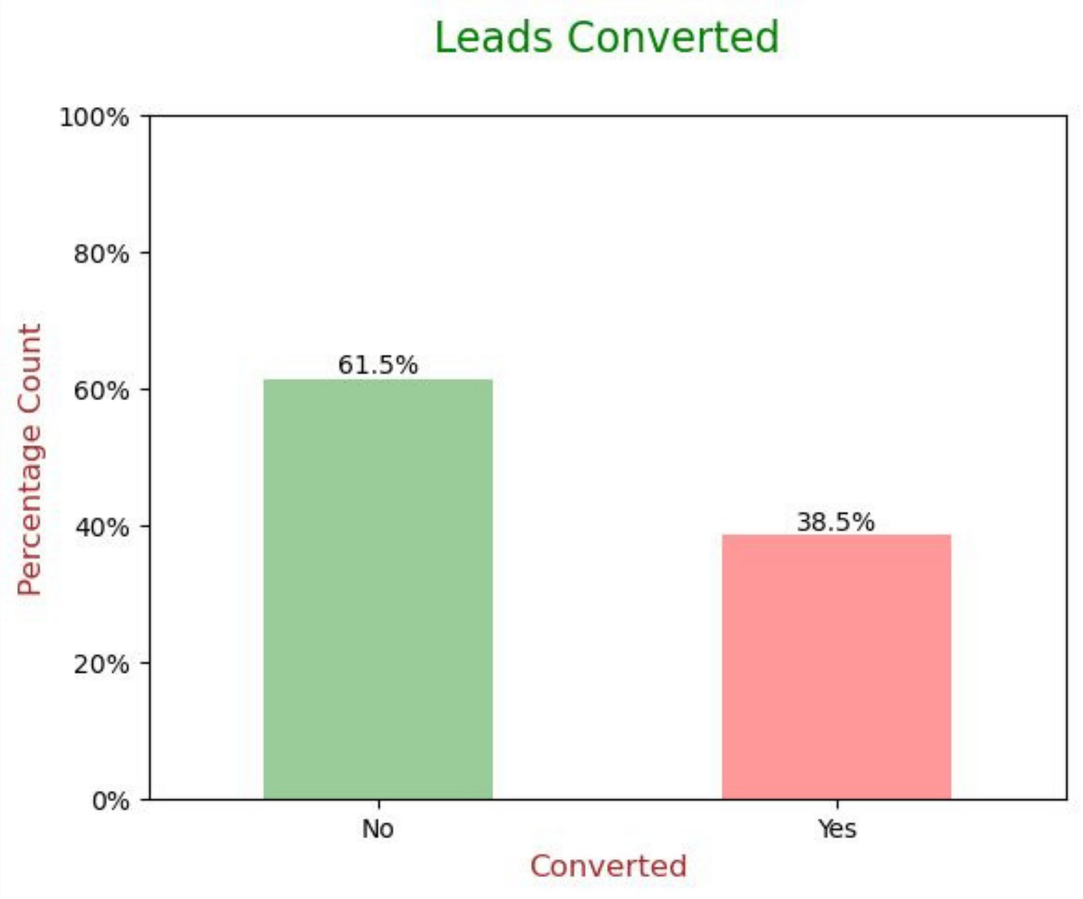
- Skewed categorical columns were examined and excluded to prevent bias in logistic regression models.
- Outliers in the "TotalVisits" and "Page Views Per Visit" columns were addressed and capped.
- Corrections were applied to rectify invalid values, and certain columns like "lead source" were standardized by addressing casing inconsistencies (e.g., converting "Google" to "google").
- To enhance clarity, low-frequency values were grouped under the category "Others."
- Binary categorical variables were appropriately encoded.
- Additional data refinement was undertaken to ensure data quality and accuracy, encompassing tasks like fixing invalid values and standardizing data (e.g., correcting "Google" to "google").



upGrad

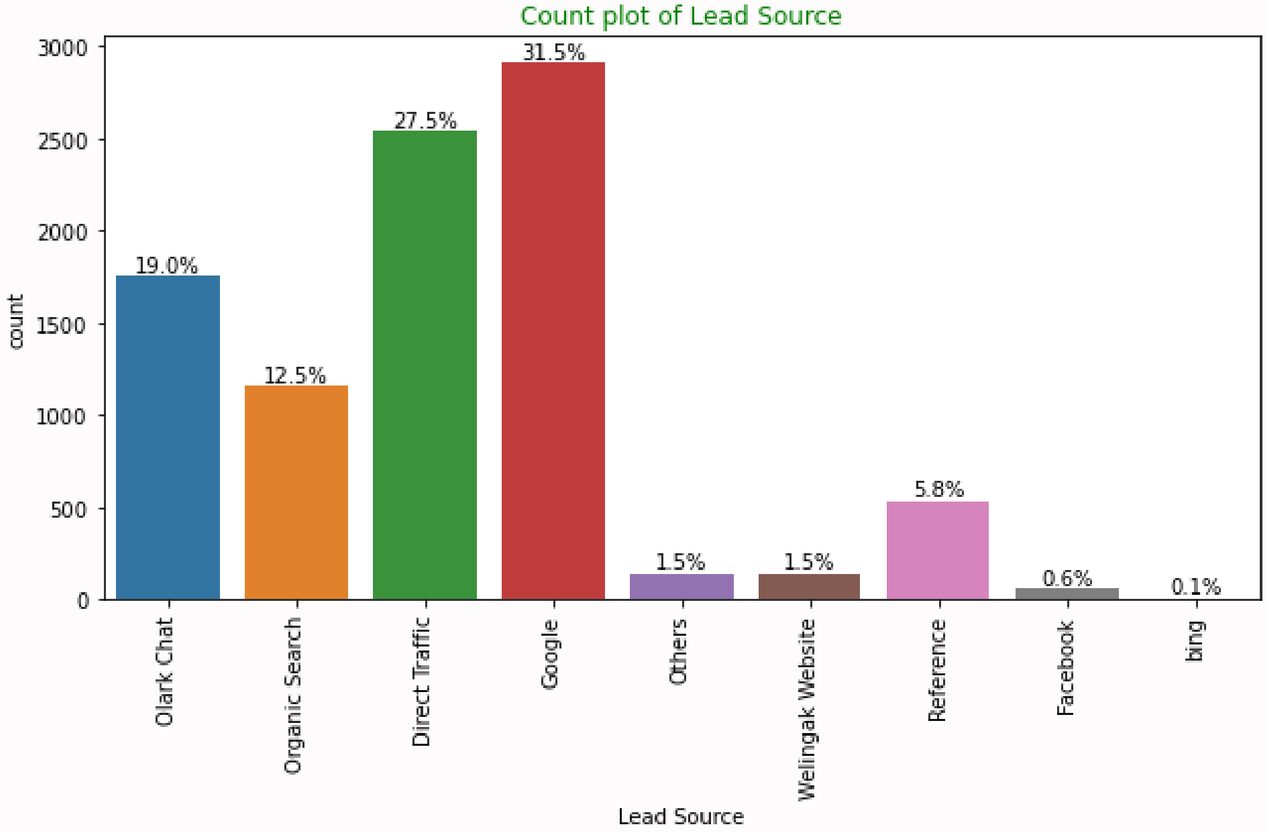
Exploratory Data Analysis (EDA):

- During the analysis of the target variable, it was observed that the data is imbalanced:
- The conversion rate stands at 38.5%, indicating that only 38.5% of individuals have successfully converted to leads (minority class).
- Conversely, 61.5% of individuals did not convert to leads (majority class).



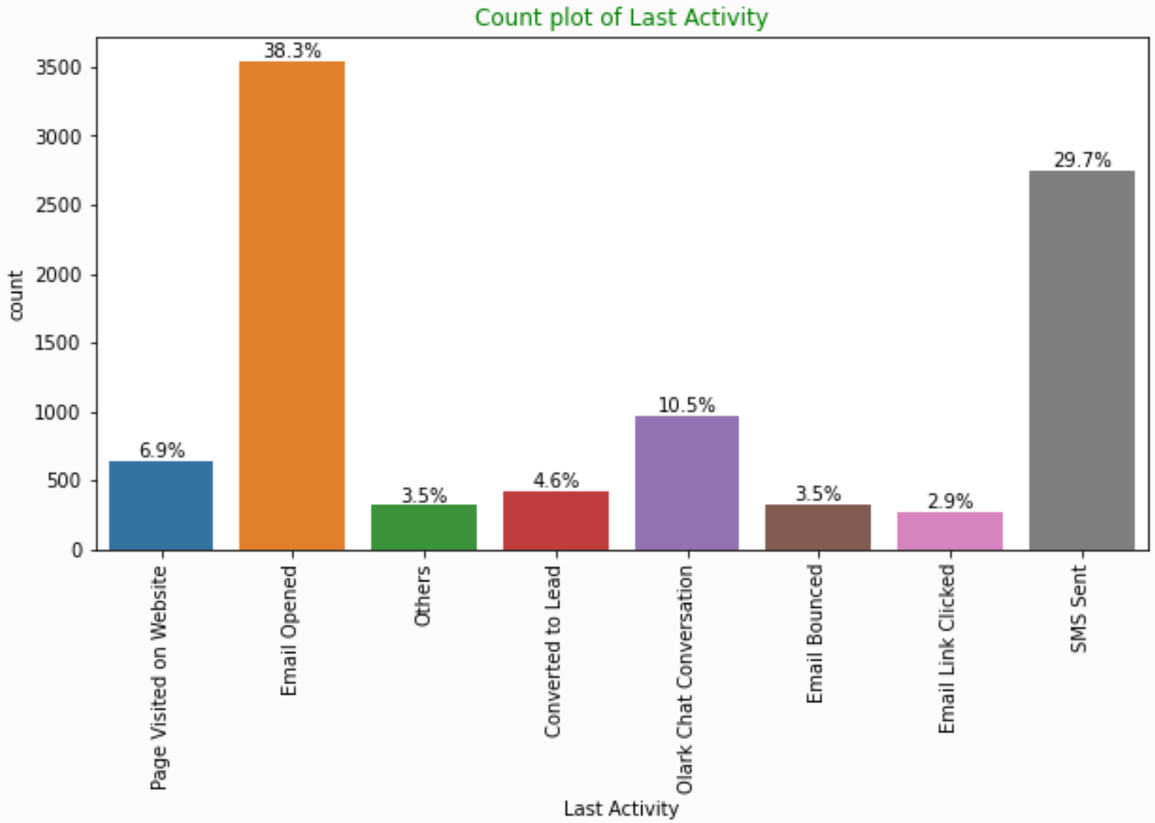
upGrad

Exploratory Data Analysis (EDA): Univariate Analysis – Categorical Variables



Lead Source:

The majority, 58%, of leads originate from a combination of Google and Direct Traffic sources.

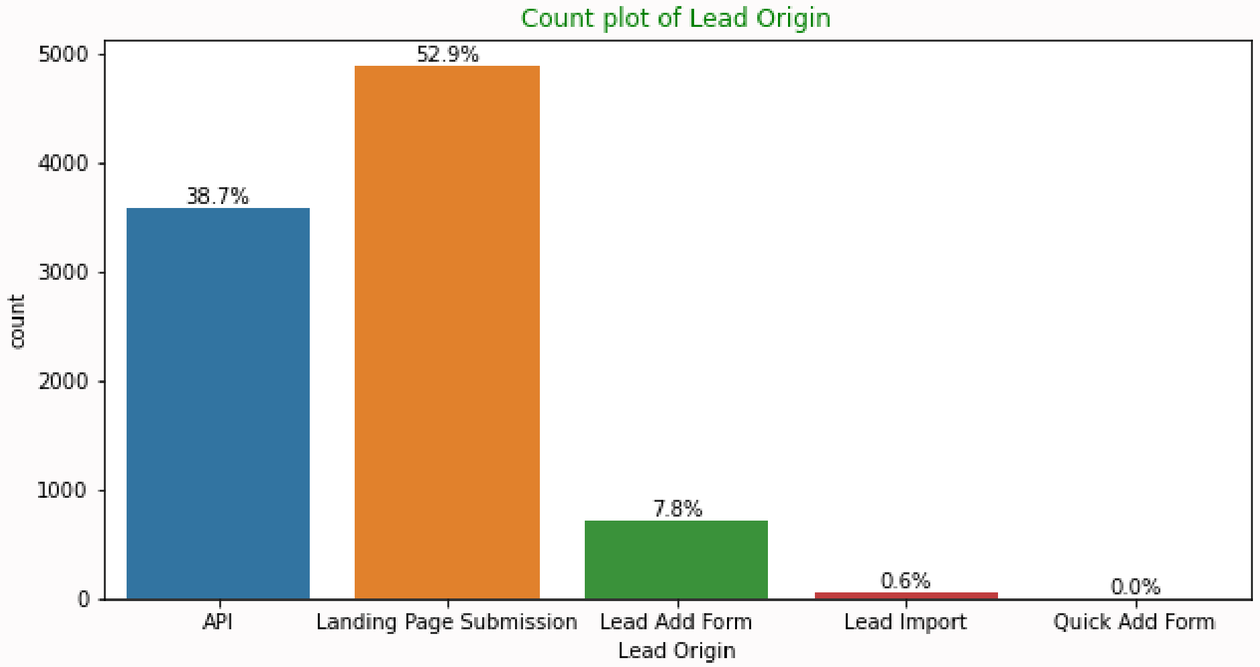


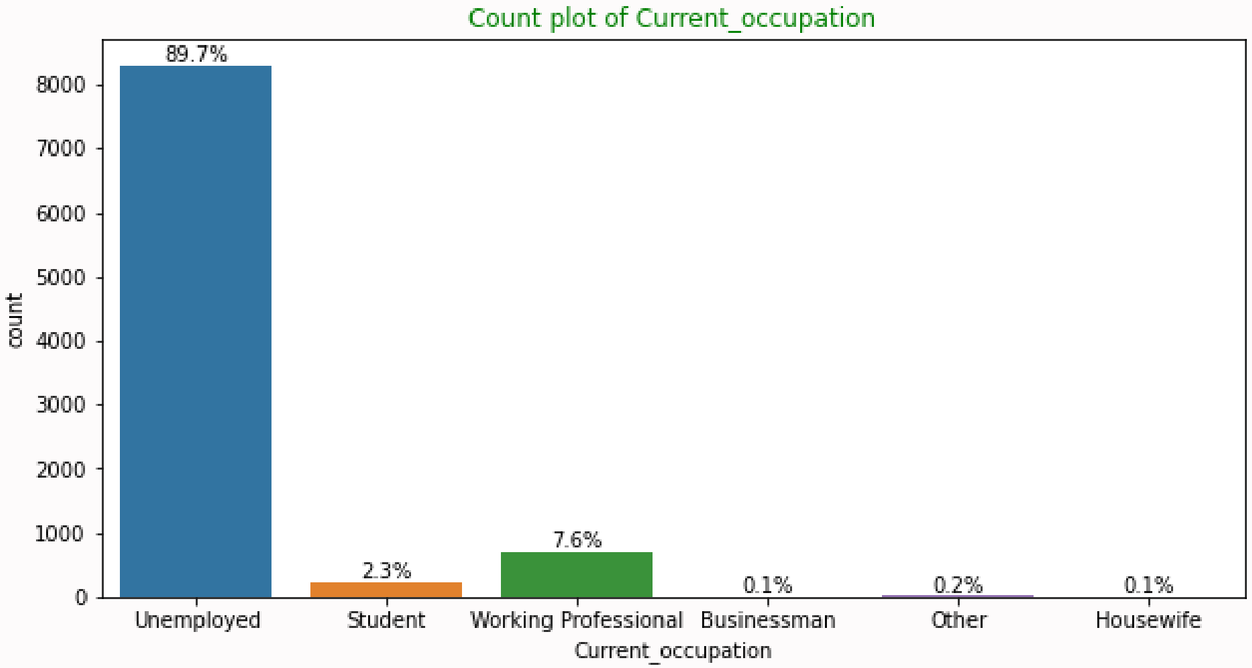
Last Activity:

Approximately 68% of customers engage in activities such as SMS Sent and Email Opened.

upGrad

Exploratory Data Analysis (EDA): Univariate Analysis – Categorical Variables





Lead Origin:

The "Landing Page Submission" is attributed to 53% of customers, while the "API" accounts for 39%.

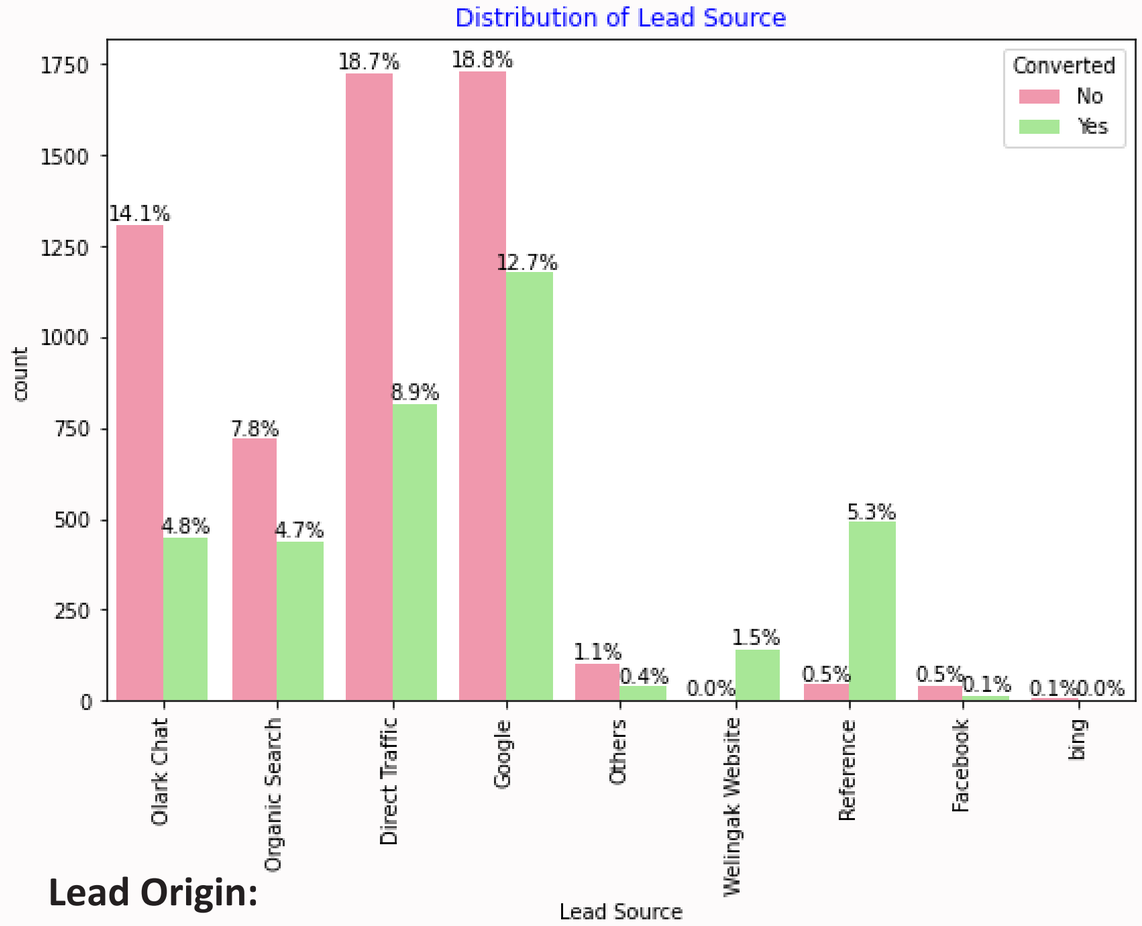
Current Occupation:

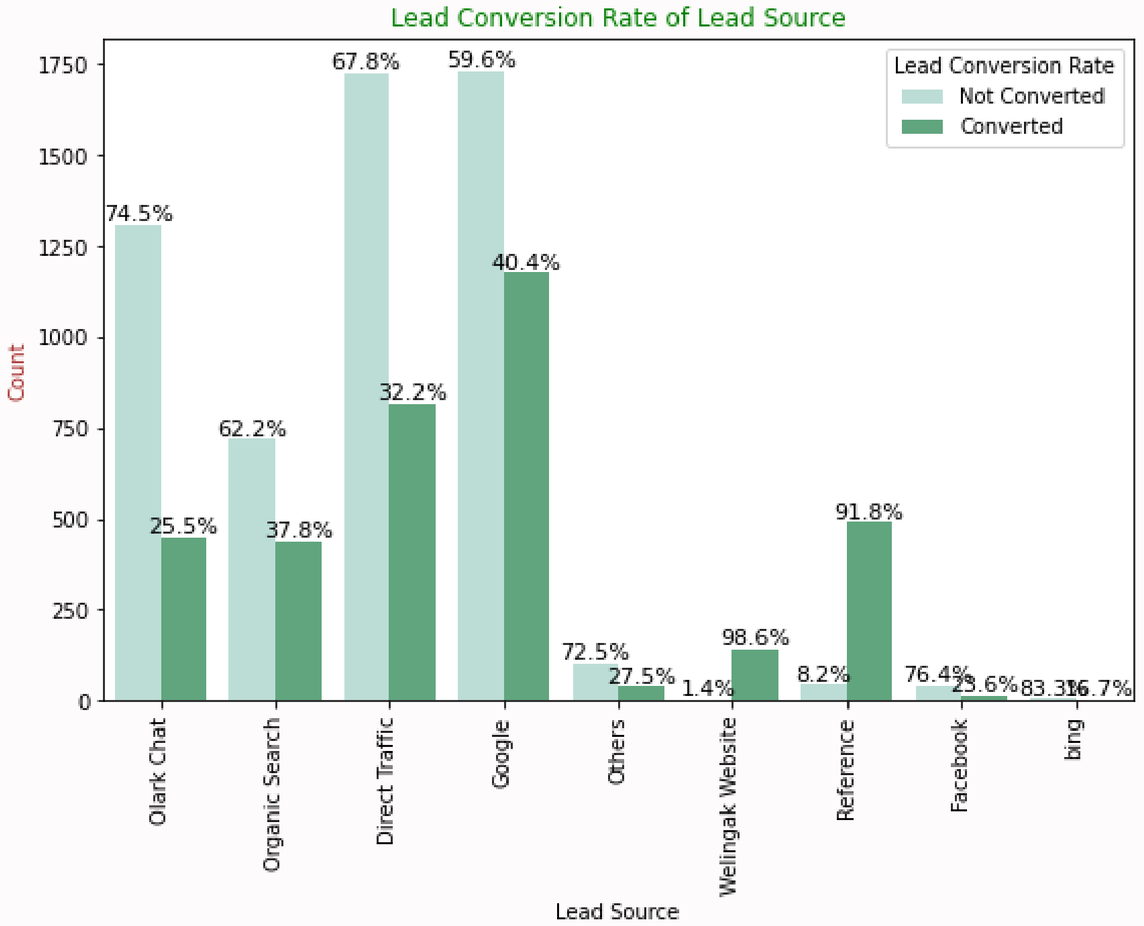
The majority, 90% of customers, are categorized as "Unemployed" in terms of their current occupation.

upGrad

Exploratory Data Analysis (EDA): Bivariate Analysis – Categorical Variables

Lead Source Countplot vs Lead Conversion Rates





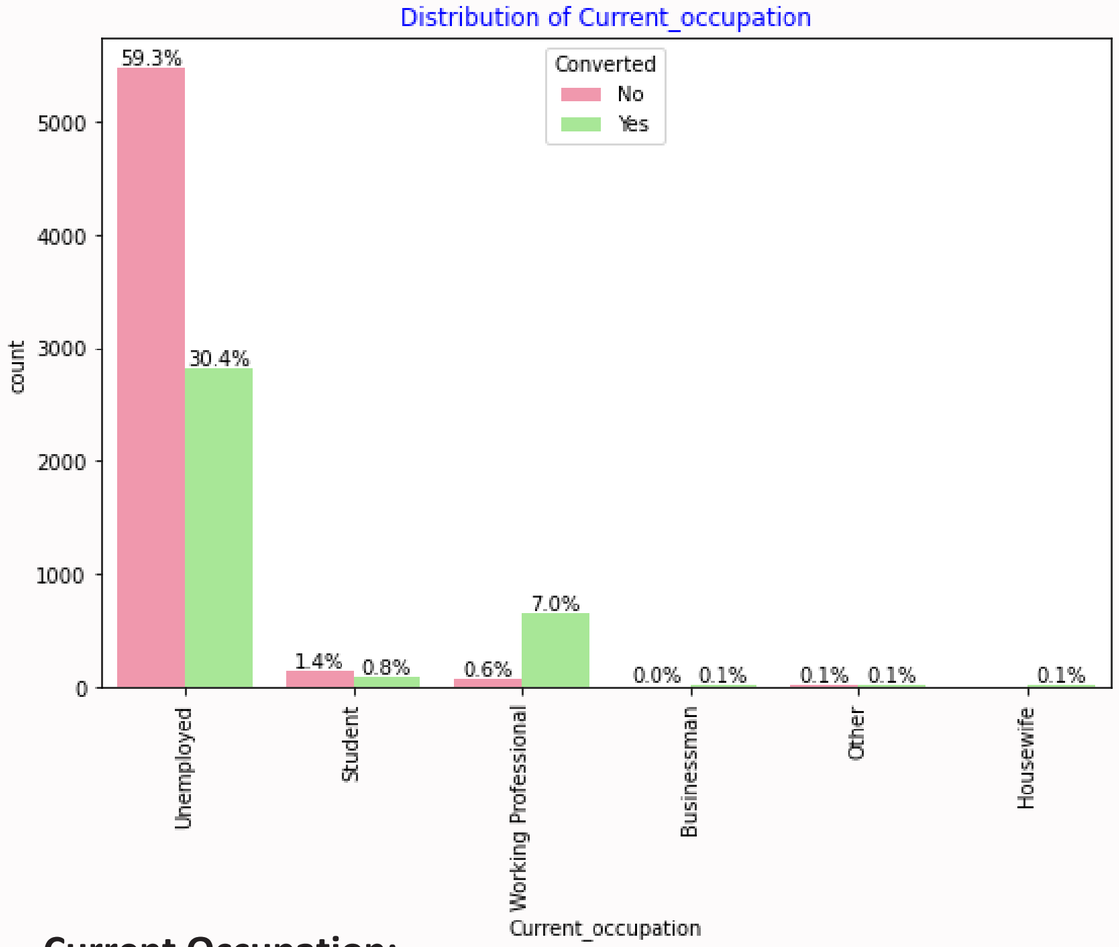
Lead Origin:

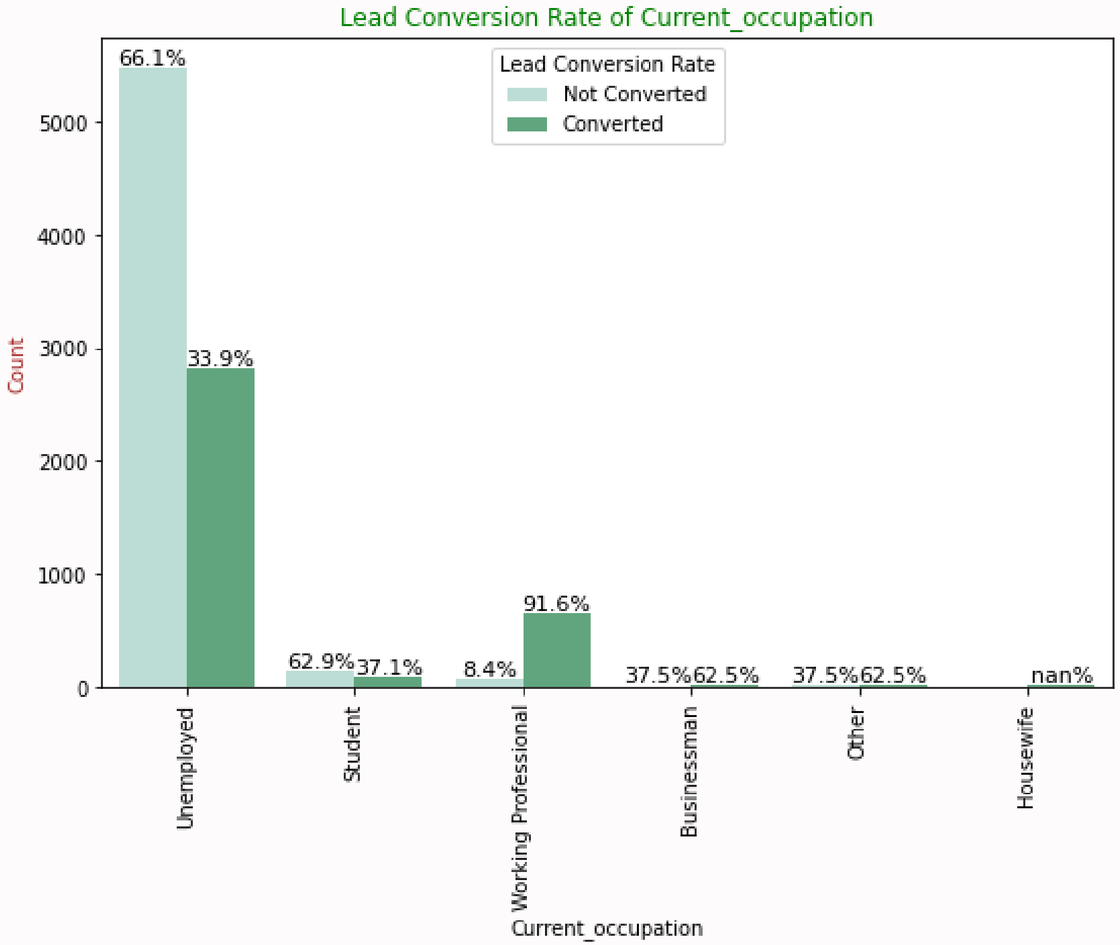
- Roughly 52% of all leads originated from "Landing Page Submission," yielding a lead conversion rate (LCR) of 36%.
- The "API" identified approximately 39% of customers, accompanied by a lead conversion rate (LCR) of 31%.

upGrad

EDA: Bivariate Analysis – Categorical Variables

Current_occupation Countplot vs Lead Conversion Rates





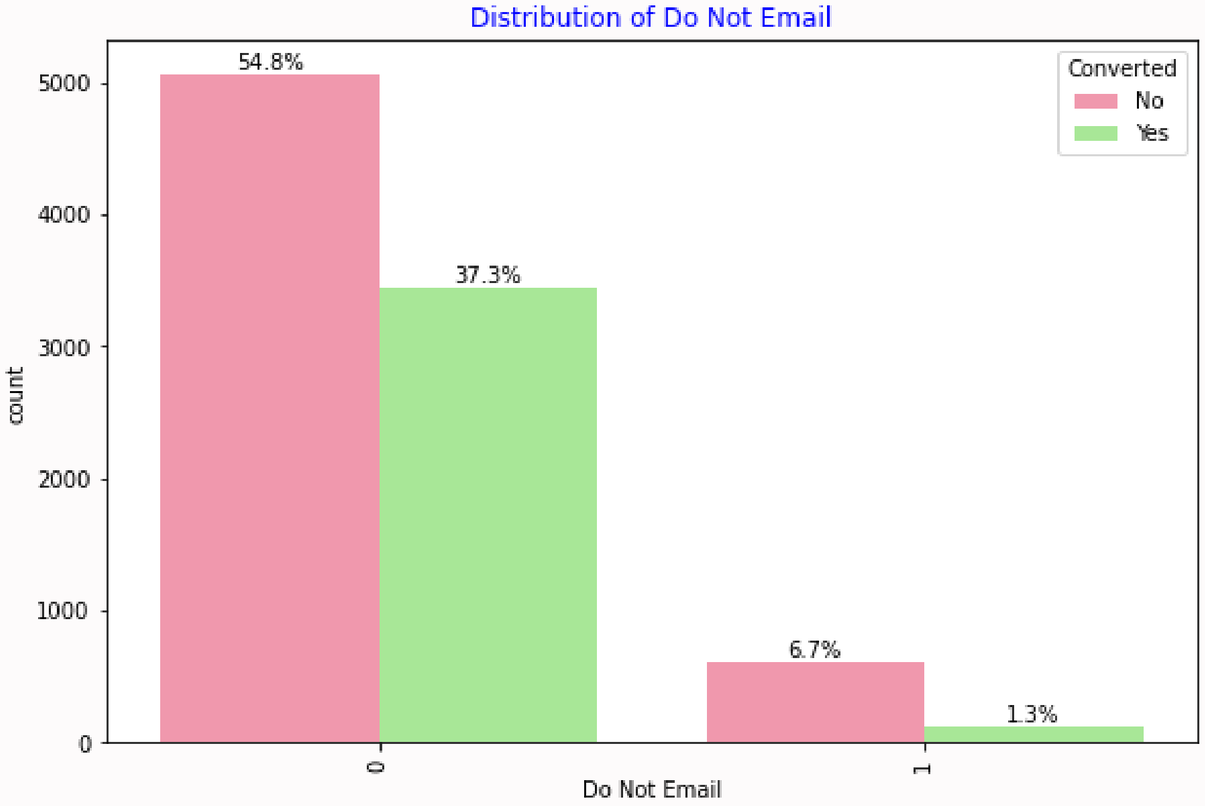
Current Occupation:

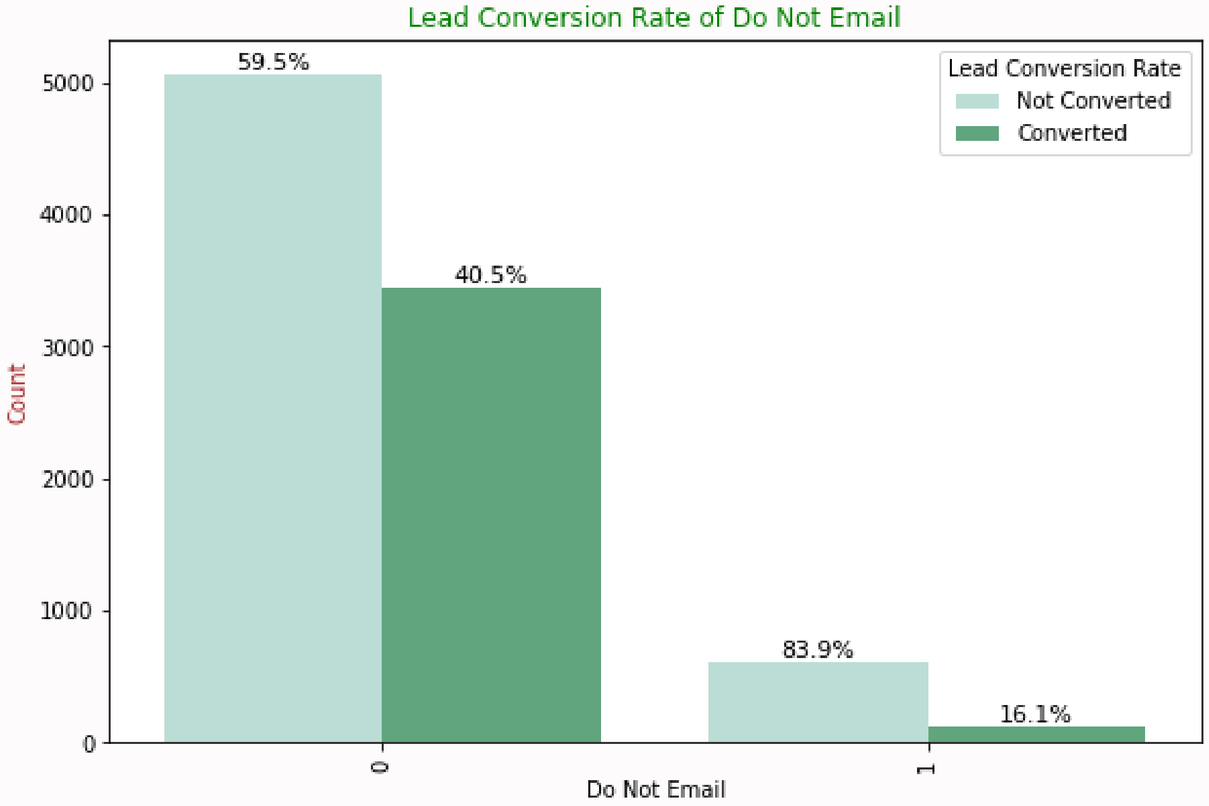
- Approximately 90% of the customers are categorized as "Unemployed," exhibiting a lead conversion rate (LCR) of 34%.
- On the other hand, "Working Professionals" constitute only 7.6% of the total customers but showcase an impressive lead conversion rate (LCR) of nearly 92%.

upGrad

EDA: Bivariate Analysis – Categorical Variables

Do Not Email Countplot vs Lead Conversion Rates





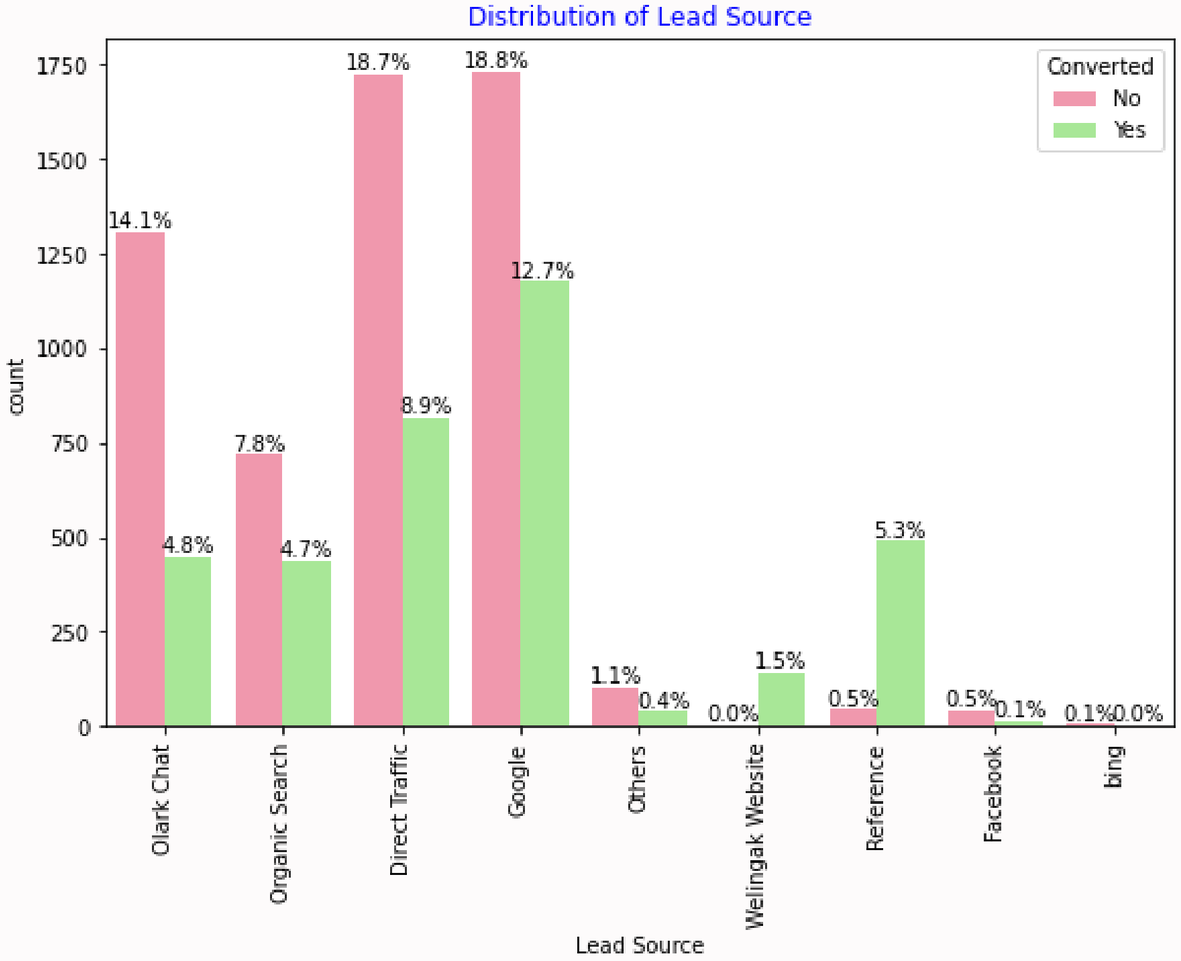
Do Not Email:

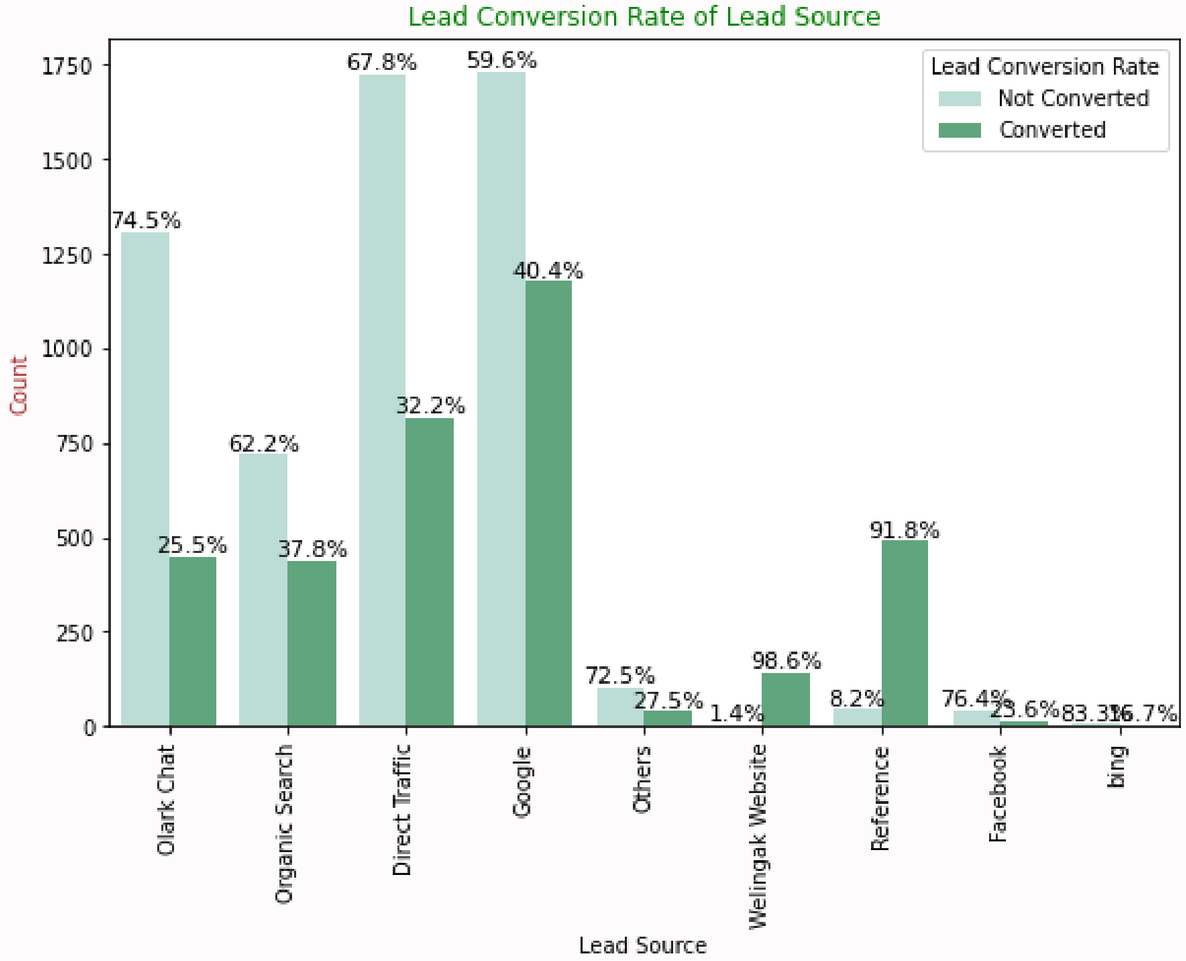
- A significant majority, 92% of individuals, have chosen not to receive emails regarding the course.
- Among them, 40% have converted to leads.

upGrad

EDA: Bivariate Analysis – Categorical Variables

Lead Source Countplot vs Lead Conversion Rates



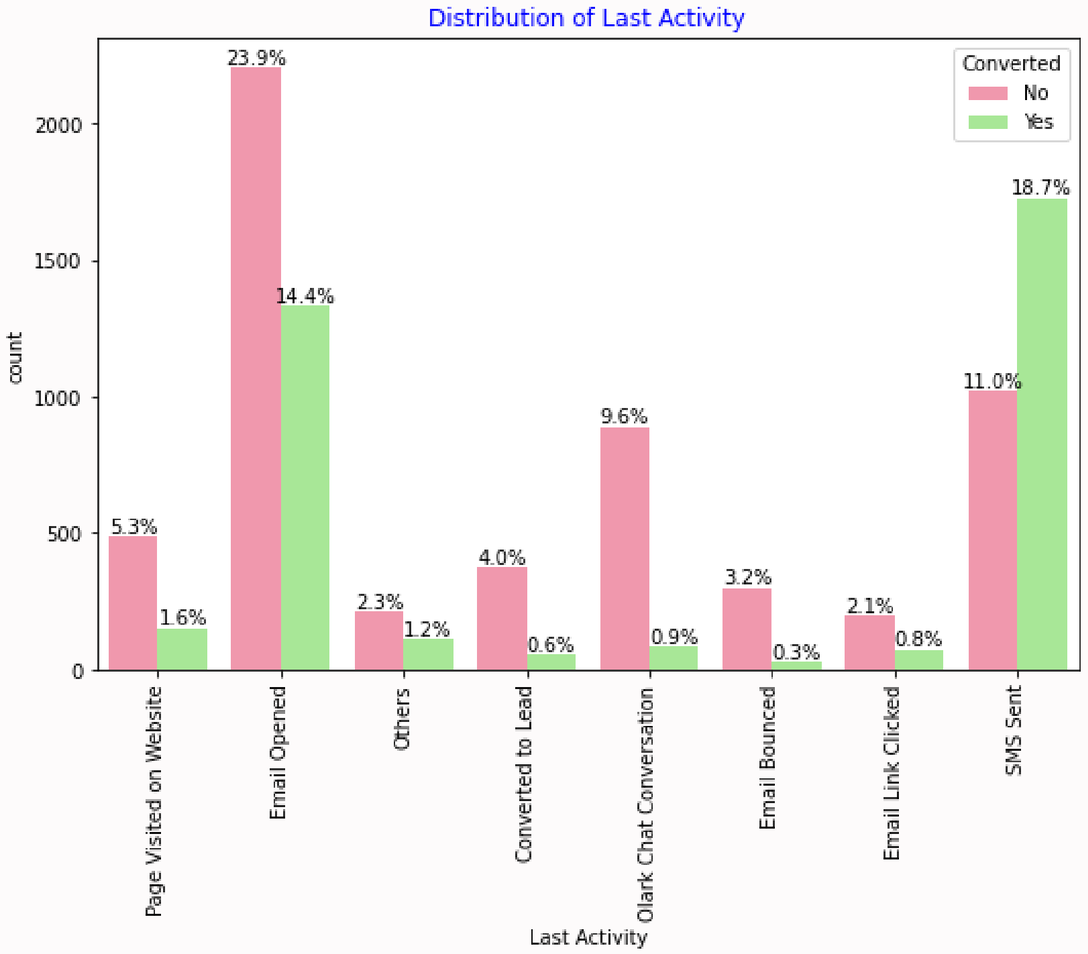


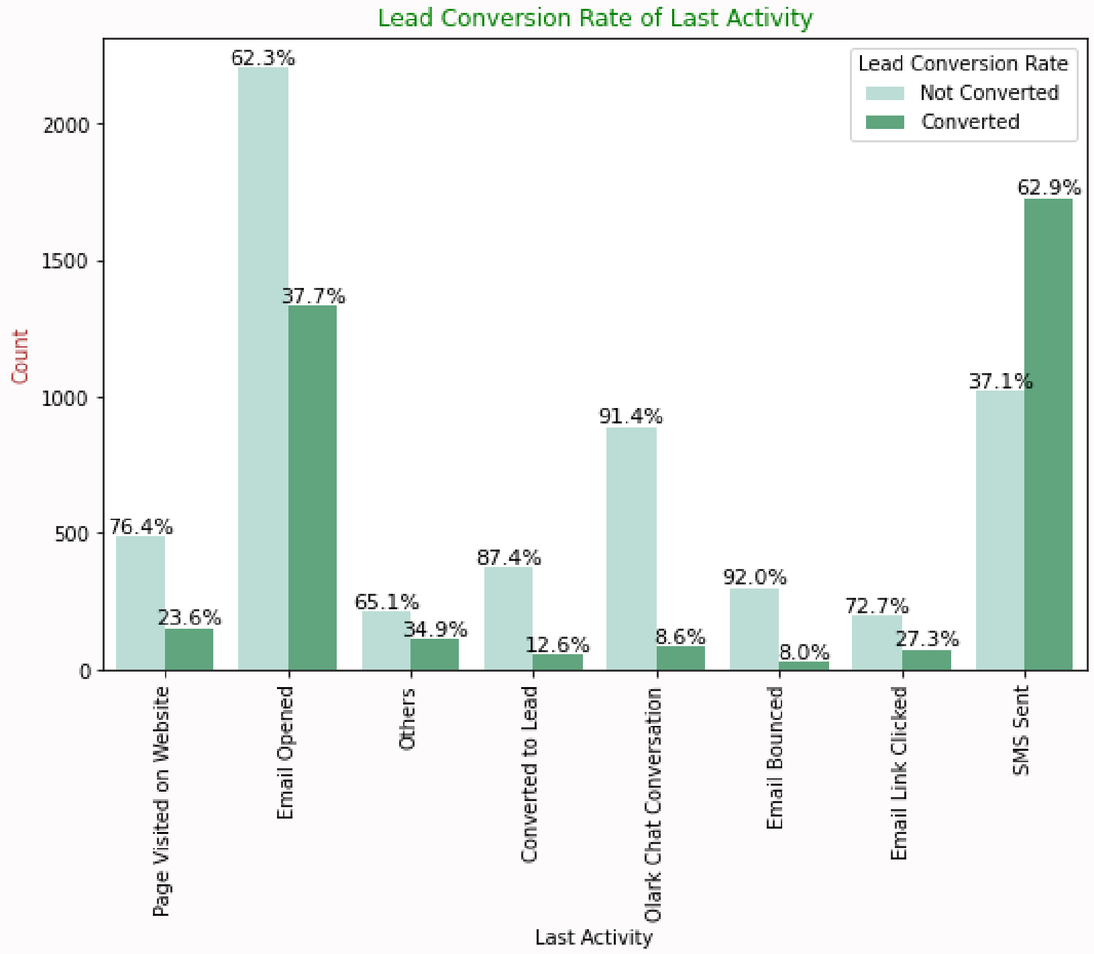
- Among the lead sources, "Google" demonstrates a lead conversion rate (LCR) of 40% from a pool of 31% customers.
- "Direct Traffic" contributes a lower LCR of 32%, drawn from 27% of customers.
- "Organic Search" yields a solid LCR of 37.8%, but it is based on a smaller portion of customers, about 12.5%.
- "Reference" boasts an impressive LCR of 91%, despite being associated with only around 6% of customers.

upGrad

EDA: Bivariate Analysis – Categorical Variables

Last Activity Countplot vs Lead Conversion Rates





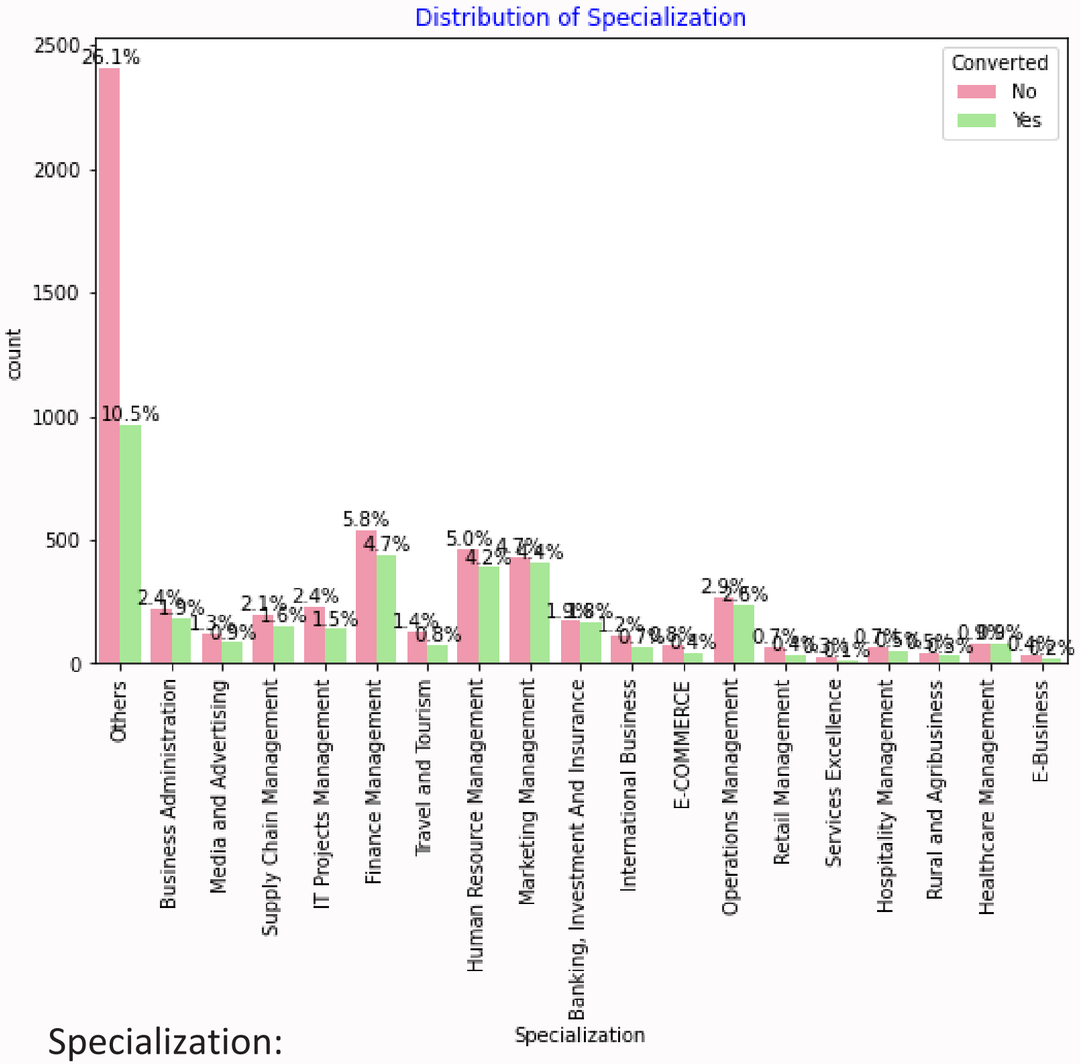
Last Activity:

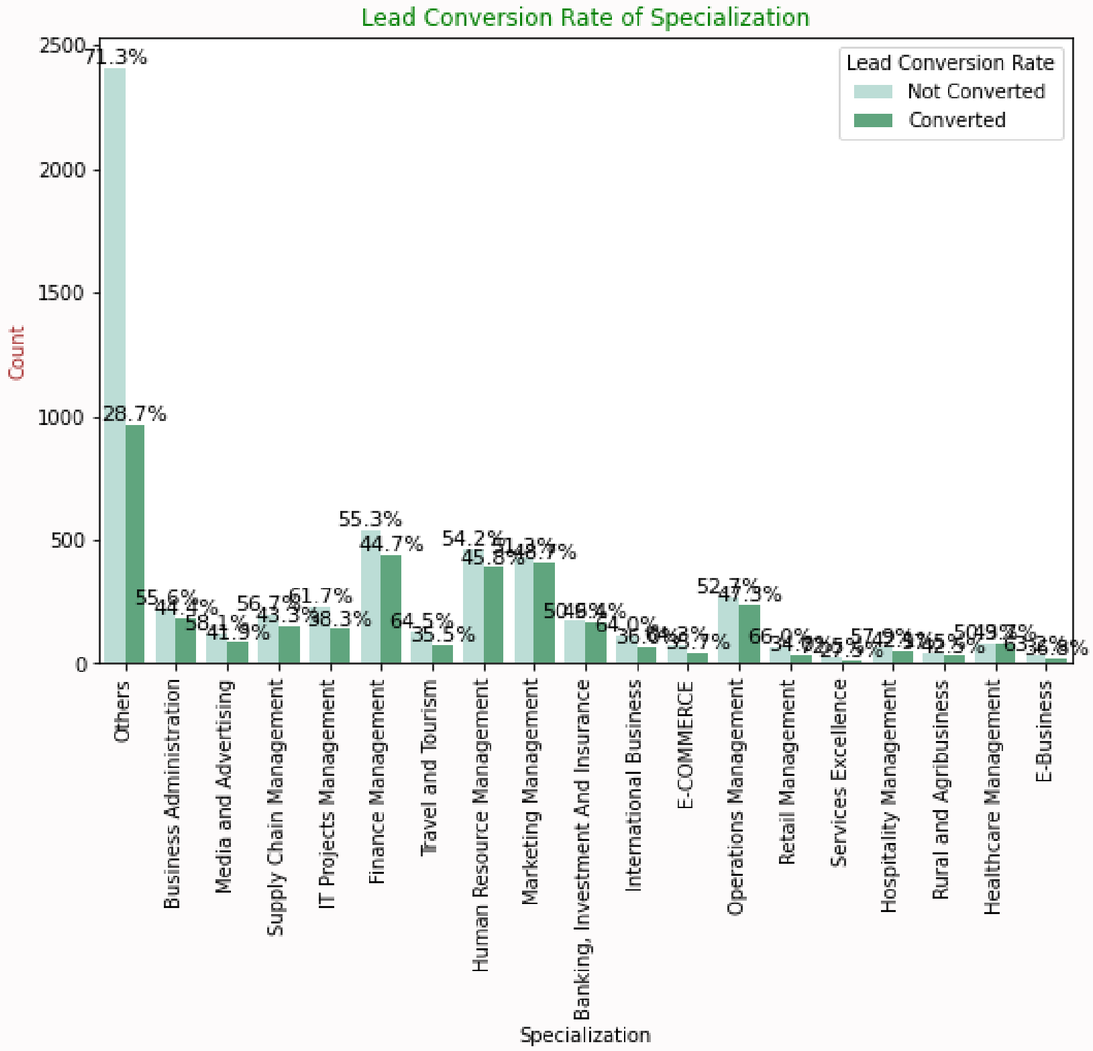
- The activity "SMS Sent" stands out with a notably high lead conversion rate of 63%, stemming from 30% of the last activities performed by customers.
- "Email Opened" holds a 38% contribution among the last activities undertaken by customers, resulting in a lead conversion rate of 37%.

upGrad

EDA: Bivariate Analysis – Categorical Variables

Specialization Countplot vs Lead Conversion Rates



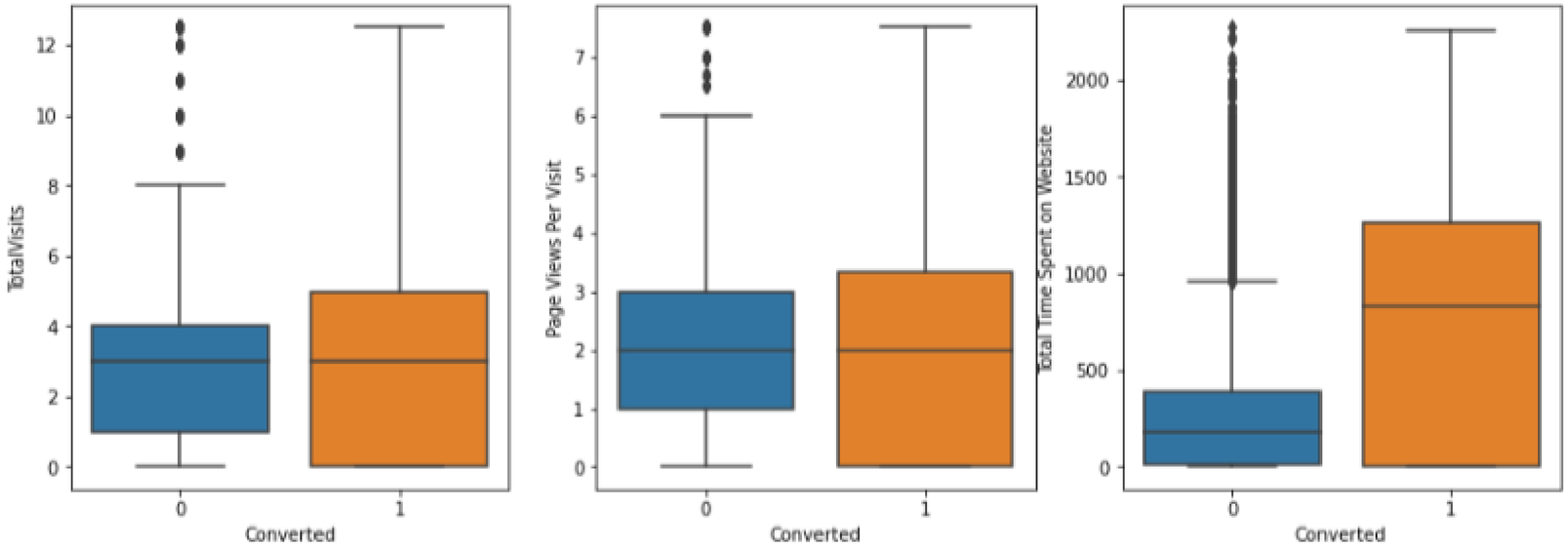


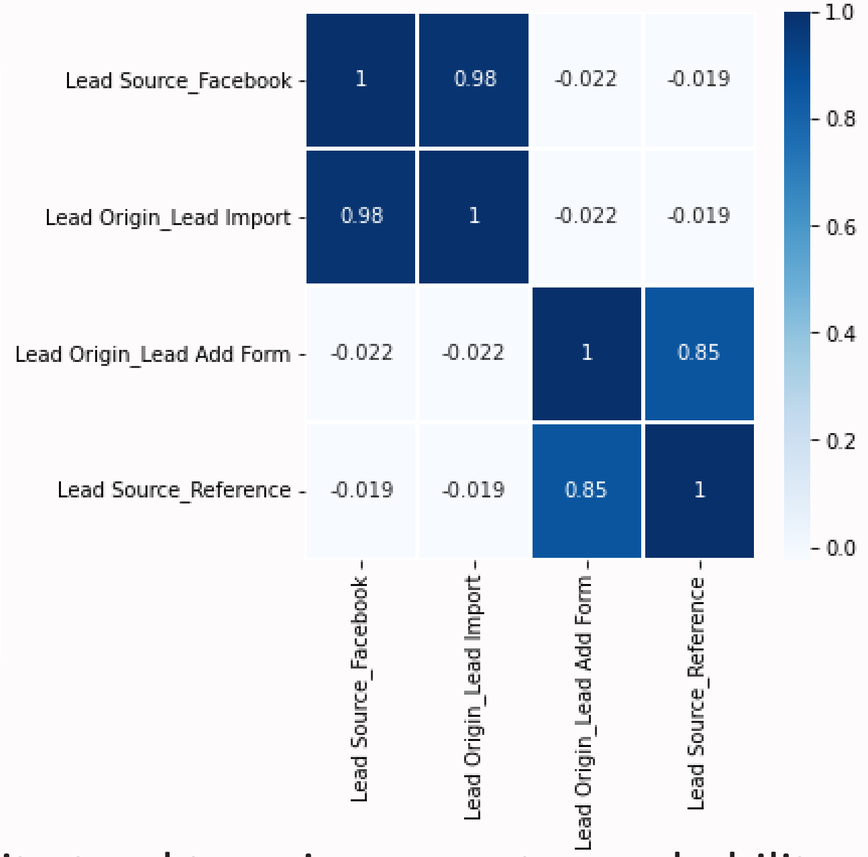
Specialization:

- "Marketing Management," "HR Management," and "Finance Management" exhibit notable contributions to lead conversion compared to other specializations.

upGrad

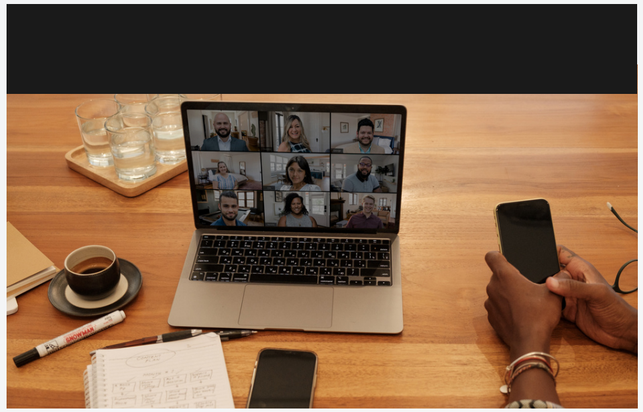
Exploratory Data Analysis (EDA) – Bivariate Analysis for Numerical Variables





- In the realm of past leads, those who invest more time on the website tend to enjoy a greater probability of successful conversion, in contrast to their counterparts who allocate less time. This trend is vividly illustrated by the box plot.

Data Preparation before Model building



Data Transformation and Preparation Steps:

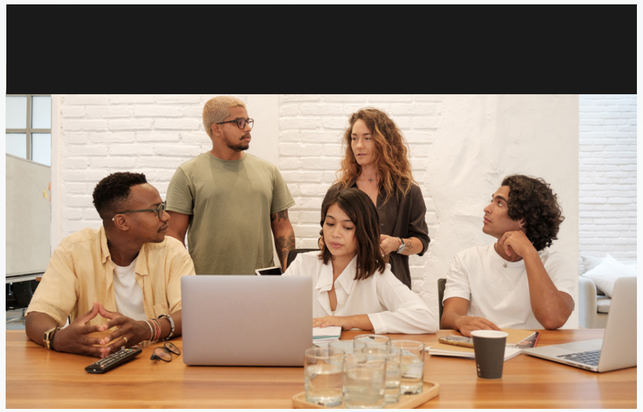
1. Binary-level categorical columns were previously encoded into 1 and 0.
2. Created dummy features (one-hot encoded) for the following categorical variables:
   - Lead Origin
   - Lead Source
   - Last Activity
   - Specialization
   - Current Occupation

Data Splitting and Scaling:

1. Train & Test Sets were split using a ratio of 70:30.
2. Feature Scaling was performed using the standardization method to normalize features.

Correlation Handling:

1. Correlation among predictor variables was assessed.
2. Highly correlated predictor variables were eliminated to avoid redundancy (Lead Origin_Lead Import and Lead Origin_Lead Add Form).



upGrad

Model Building: Logistic Regression

In this phase, we construct a Logistic Regression Model to predict categorical variables. Our approach is as follows:

01

Feature Selection Using RFE:

- We employ RFE to iteratively reduce variables in our model.
- The chosen method here is Logistic Regression.

02

Manual Fine-Tuning:

After RFE, we conduct manual adjustments based on p-values and VIFs (Variance Inflation Factors) to refine the model.

The key observations from the RFE process are as follows:

- Columns like 'Total Time Spent on Website,' 'Lead Source_Olark Chat,' 'Lead Source_Reference,' 'Lead Source_Welingak Website,' etc., were chosen by RFE.
- Variables with high-ranking values were dropped to enhance model efficiency.

Manual Feature Reduction and Model Building:

Two models were developed:

1. **Model 1:** After RFE, we built the model using stats models, focusing on the features selected. We noticed that the column 'Current_occupation_Housewife' had a high p-value of 0.999. As this value exceeded the accepted significance threshold of 0.05, we decided to exclude this column from the model.
2. **Model 2**: Following the exclusion of 'Current_occupation_Housewife,' we refined the model further. Notably, this model showcased promising coefficients that indicate variable influence on predicting the probability of lead conversion.

upGrad

Model Evaluation

During the Model Evaluation phase, the following metrics are used to assess the performance of the logistic regression model:

01

Confusion Matrix:

A matrix that visualizes true positive, true negative, false positive, and false negative outcomes.

02

Accuracy:

Ratio of correctly predicted instances to the total instances.

03

Sensitivity and Specificity:

Model's ability to correctly identify positive and negative cases, respectively.

04

Threshold Determination using ROC & Finding Optimal Cutoff Point:

Using the Receiver Operating Characteristic (ROC) curve to identify the optimal threshold for classification.

05

Precision and Recall

Evaluation of precision (ability to correctly predict positive instances) and recall (ability to identify all positive instances).

upGrad

Model Evaluation

Below steps involved in the code for model evaluation using logistic regression:

Step 1: Getting the Predicted Values on the Train Set
Step 2: Creating a DataFrame with Actual and Predicted Probabilities
Step 3: Determination of the Optimal Cutoff Threshold
Step 4: Confusion Matrix and Accuracy
Step 5: Metrics Beyond Accuracy
Step 6: Plotting the ROC Curve
Step 7: Finding Optimal Cutoff Point/Probability
Step 8: Precision and Recall Tradeoff
Step 9: Trade-Off Between Precision and Recall

upGrad

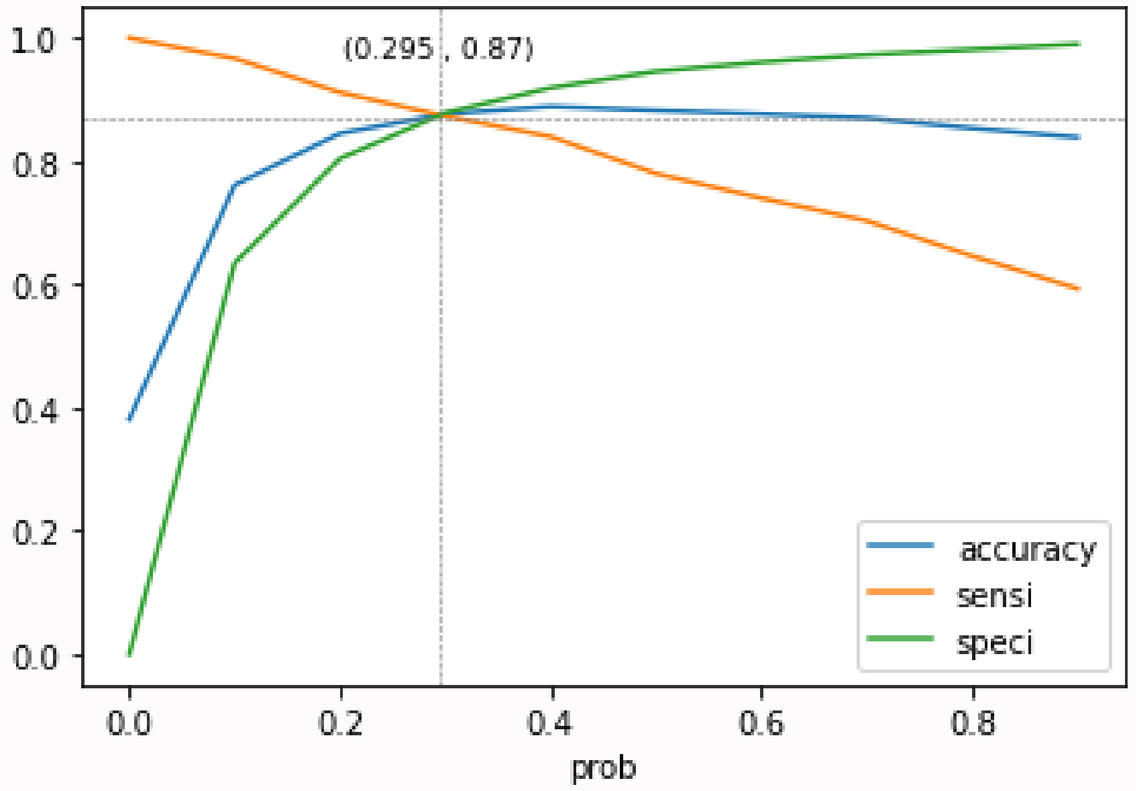
Model Evaluation

Certainly! After evaluating the metrics from both the precision-recall curve and the sensitivity-specificity curve, a decision was made to proceed with a cutoff threshold of 0.295. This decision was based on a thorough comparison of the evaluation metrics derived from both plots, ensuring a well-balanced trade-off between precision and recall.

```
Confusion Matrix
[[3504  498]
 [ 312 2154]]

----------------------------------------

True Negative                   :  3504
True Positive                   :  2154
False Negative                  :  312
False Positve                   :  498
Model Accuracy                  :  0.8748
Model Sensitivity               :  0.8735
Model Specificity               :  0.8756
Model Precision                 :  0.8122
Model Recall                    :  0.8735
Model True Positive Rate (TPR)  :  0.8735
Model False Positive Rate (FPR) :  0.1244
```



Confusion Matrix & Evaluation Metrics with 0.295 as cutoff

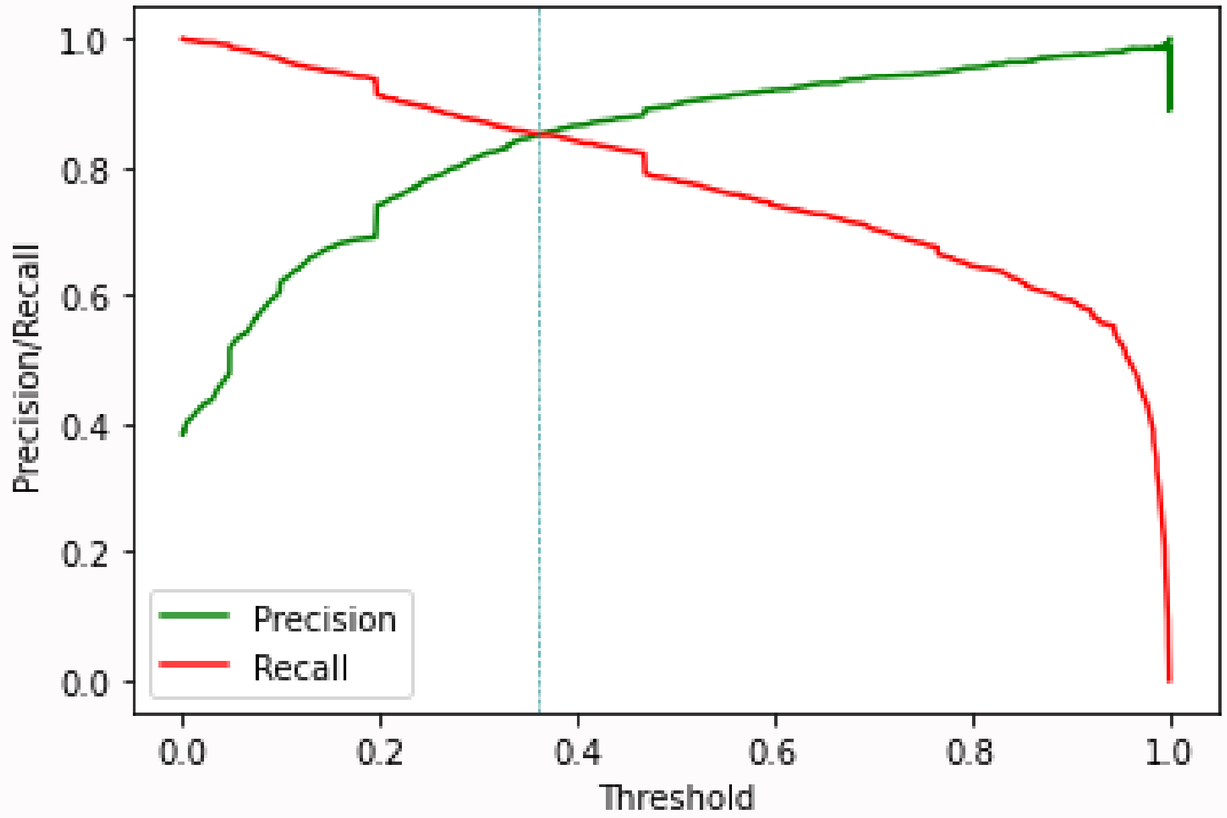
upGrad

Model Evaluation

```
Confusion Matrix
[[3631  371]
 [ 363 2103]]

----------------------------------------

True Negative                       :  3631
True Positive                       :  2103
False Negative                      :  363
False Positve                       :  371
Model Accuracy                      :  0.8865
Model Sensitivity                   :  0.8528
Model Specificity                   :  0.9073
Model Precision                     :  0.85
Model Recall                        :  0.8528
Model True Positive Rate (TPR)      :  0.8528
Model False Positive Rate (FPR)     :  0.0927
```



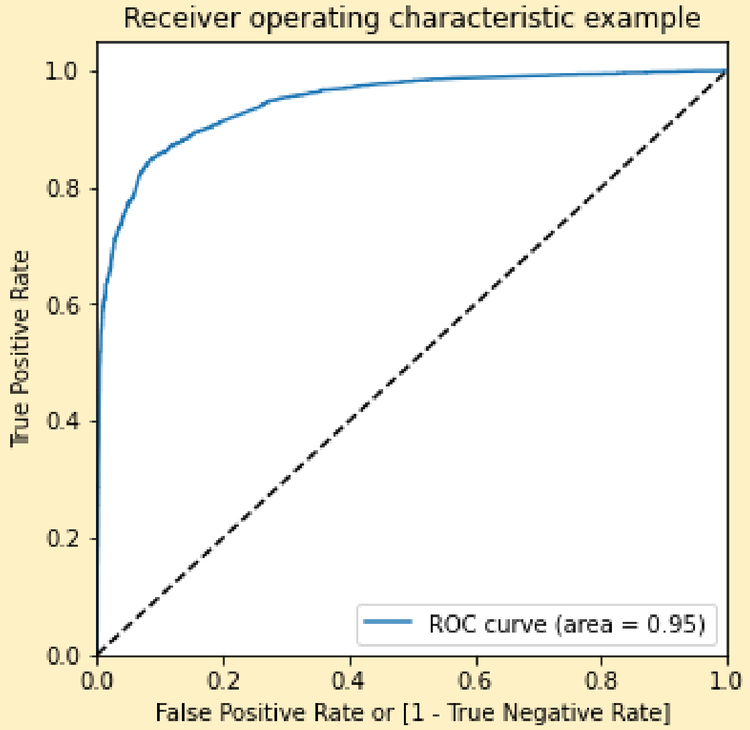
Confusion Matrix & Evaluation Metrics with 0.36 as cutoff

upGrad

Model Evaluation

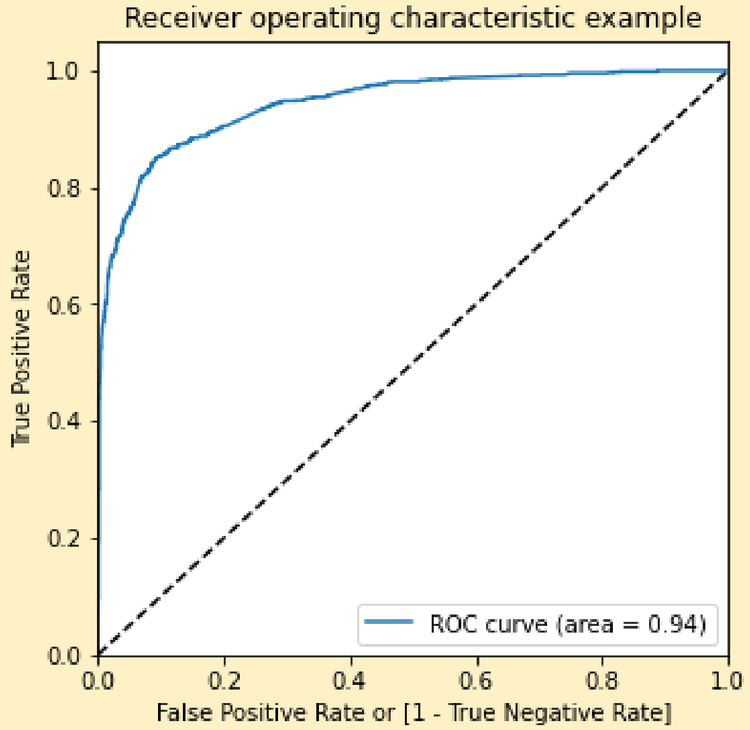
ROC Curve – Train Data Set

- AUC-ROC Score: The AUC-ROC score of 0.95 indicates strong predictive performance, reflecting a well-performing model.
- Curve Position: The curve's location in the plot's top-left corner showcases the model's high true positive rate and low false positive rate across various threshold values.



ROC Curve – Test Data Set

- AUC-ROC Score: The AUC-ROC score of 0.94 indicates strong predictive performance, reflecting a well-performing model.
- Curve Position: The curve's location in the plot's top-left corner showcases the model's high true positive rate and low false positive rate across various threshold values.



upGrad

Recommendation based on Final Model

- In line with the problem statement, enhancing lead conversion is of paramount importance for the growth and prosperity of X Education. To facilitate this goal, we've constructed a regression model that assists in identifying the most influential factors affecting lead conversion.
- Our analysis has pinpointed certain features with notably positive coefficients. These features should take precedence in our marketing and sales strategies to amplify lead conversion rates

The following factors have significant influence on lead conversion probability, ranked in descending order:

- Lead Source - Welingak Website (Coefficient: 5.704502)
- Tags - Will revert after reading the email (Coefficient: 4.235340)
- Lead Source - Reference (Coefficient: 3.823673)
- Last Activity - SMS Sent (Coefficient: 2.072489)
- Current Occupation - Working Professional (Coefficient: 1.382992)
- Lead Source - Olark Chat (Coefficient: 1.234756)
- Total Time Spent on Website (Coefficient: 1.078077)

Additionally, we've recognized features with negative coefficients, which might point towards potential areas that warrant enhancement.

- Specialization - International Business (Coefficient: -0.631354)
- Last Activity - Olark Chat Conversation (Coefficient: -0.768362)
- Current Occupation - Student (Coefficient: -0.803841)
- Specialization - Travel and Tourism (Coefficient: -0.824611)
- Specialization - Hospitality Management (Coefficient: -1.083774)
- Constant (Coefficient: -2.478867)
- Tags - Ringing (Coefficient: -3.169778)

upGrad

THANK'S FOR WATCHING

Thank you for your time and consideration in reviewing this assignment submission. Your feedback and insights are highly valued.