Lead Score Case Study

Group Members
1. Osheen Bindroo
2. Rishabh Mehra

Problem Statement

X Education seeks to improve lead conversion rates by identifying potential customers more accurately. With a current conversion rate of 30%, they aim to achieve 80% by implementing a lead scoring model. The company gathers data from website interactions, form submissions, and referrals. Your task is to clean, explore, and pre-process this data, then select and train a classification model. The model should assign lead scores based on the likelihood of conversion. Post-implementation, continuous monitoring and updates are crucial for sustained success in prioritizing high-conversion potential leads and enhancing the overall efficiency of the sales process.

Business Objective:

The business objective of this case study is to significantly improve lead conversion rates for X Education by implementing an effective lead scoring system. The specific target set by the CEO is to achieve an 80% lead conversion rate. This involves the development and deployment of a predictive model that assigns lead scores based on various features and behaviours, allowing the sales team to prioritize leads with the highest likelihood of converting into paying customers. The ultimate aim is to optimize the sales process, enhance efficiency, and maximize the return on investment in marketing efforts, leading to increased revenue and overall business success for X Education.

Solution Methodology

Data Pre-processing:

- **Data Cleaning and Manipulation:**
  - Identify and handle duplicate data entries.
  - Check and address NA values and missing data.
  - Drop columns with a significant amount of missing values or those not useful for analysis.
  - Impute missing values if necessary.

- **Outlier Handling:**
  - Identify and address outliers in the dataset.

Exploratory Data Analysis (EDA):

- **Univariate Data Analysis:**
  - Perform value count and assess the distribution of variables.
- **Bivariate Data Analysis:**
  - Examine correlation coefficients and patterns between variables.

Solution Methodology

Data Transformation:

- **Feature Scaling:**
  - Scale numerical features if required.

- **Dummy Variables and Encoding:**
  - Create dummy variables and encode categorical data.

Model Building:

- **Classification Technique:**
  - Utilize logistic regression for model creation and prediction.

Model Validation:

- Validate the model's performance using appropriate metrics.

Conclusions and Recommendations:

- Summarize conclusions drawn from the analysis.

Data Manipulation

❑ The dataset comprises 37 rows and 9240 columns.

❑ Features with single values like "Magazine," "Receive More Updates About Our Courses," and "Update me on Supply Chain Content" are removed.

❑ Unnecessary columns "Prospect ID" and "Lead Number" are excluded from analysis.

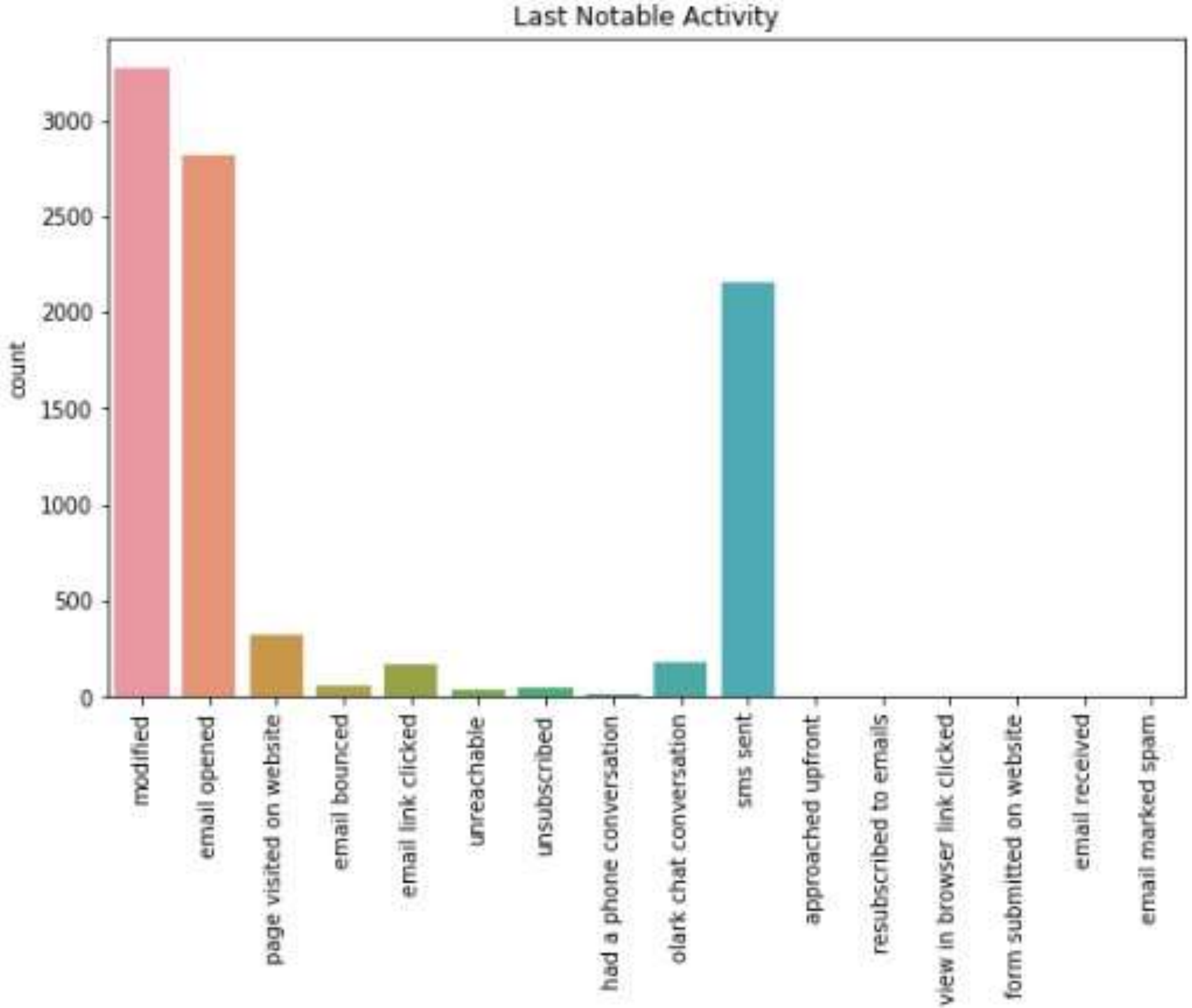
❑ Object-type variables with low variance, including "Do Not Call," "What matters most to you in choosing the course," and others, are dropped based on value counts.

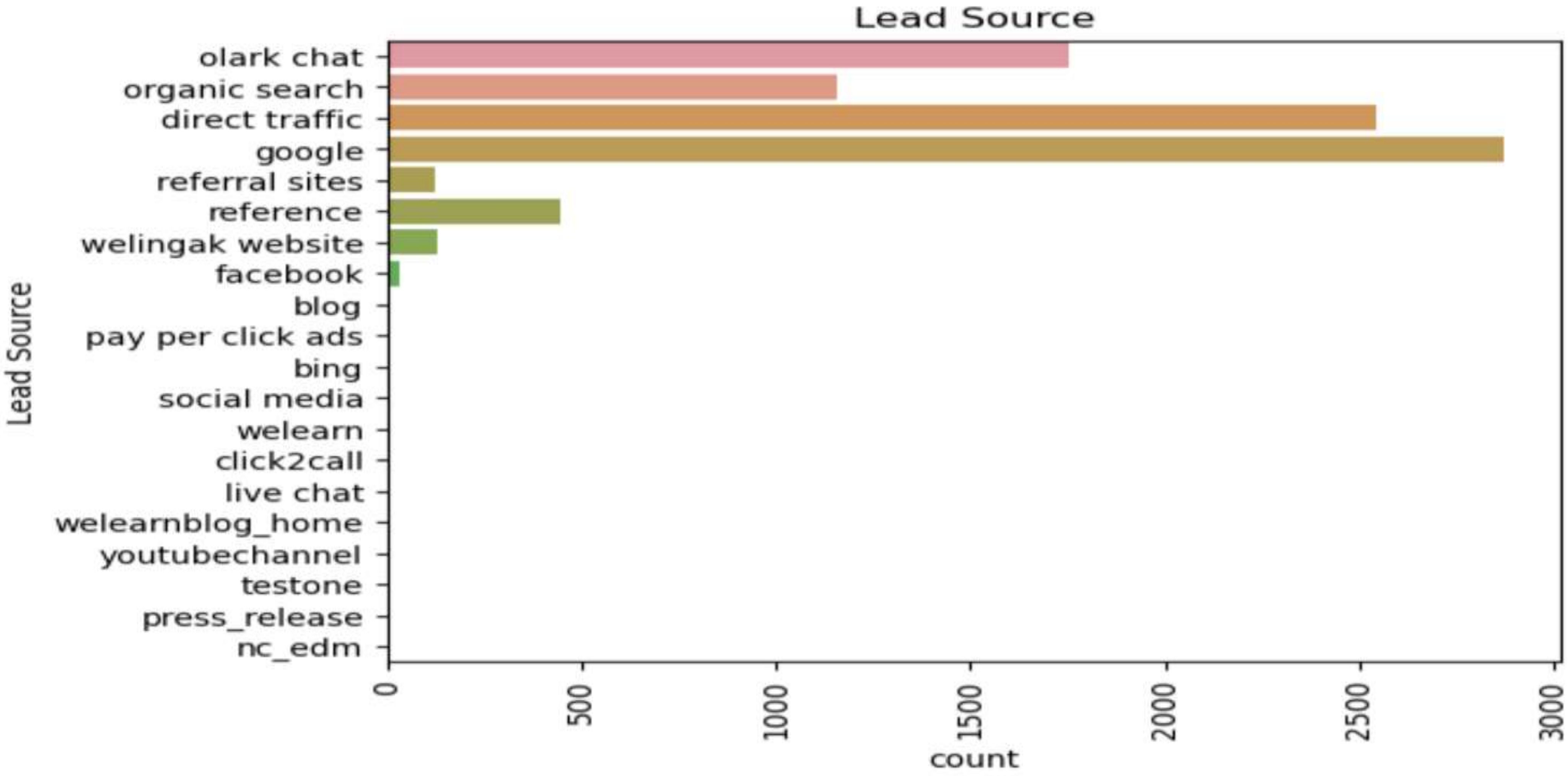
❑ Columns with over 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile,' are removed to enhance data quality.

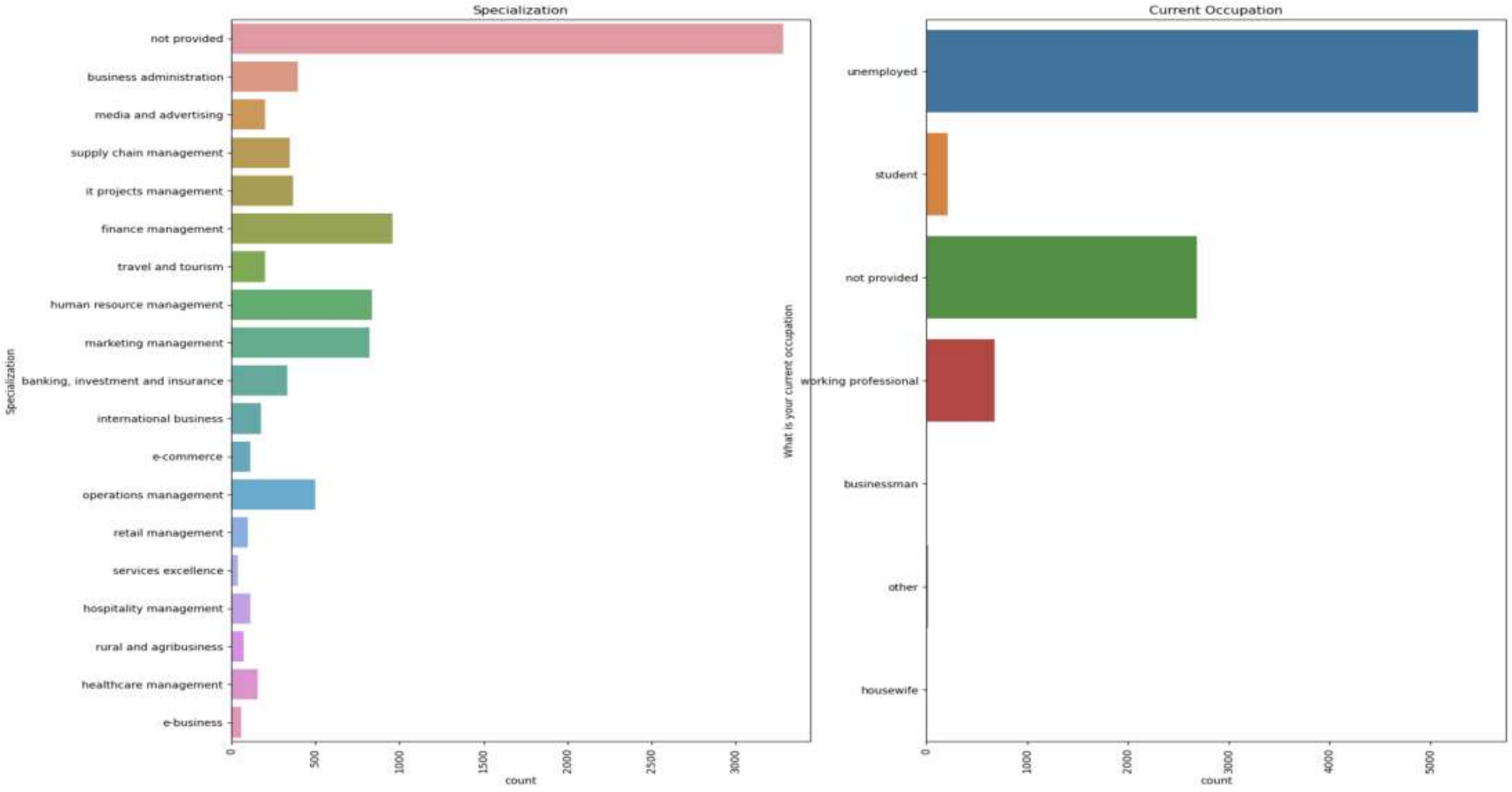
Exploratory Data Analysis

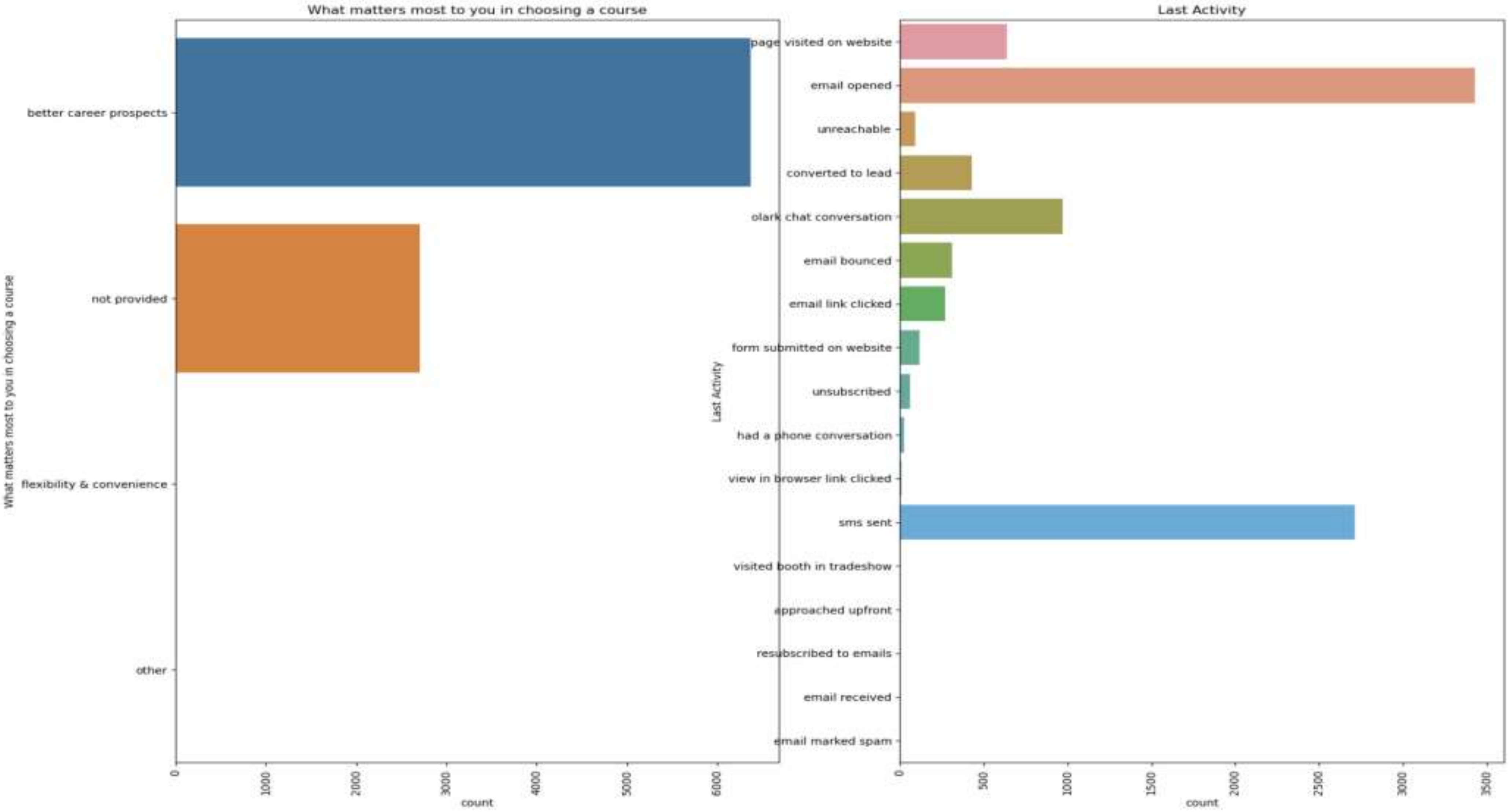
Univariate Analysis

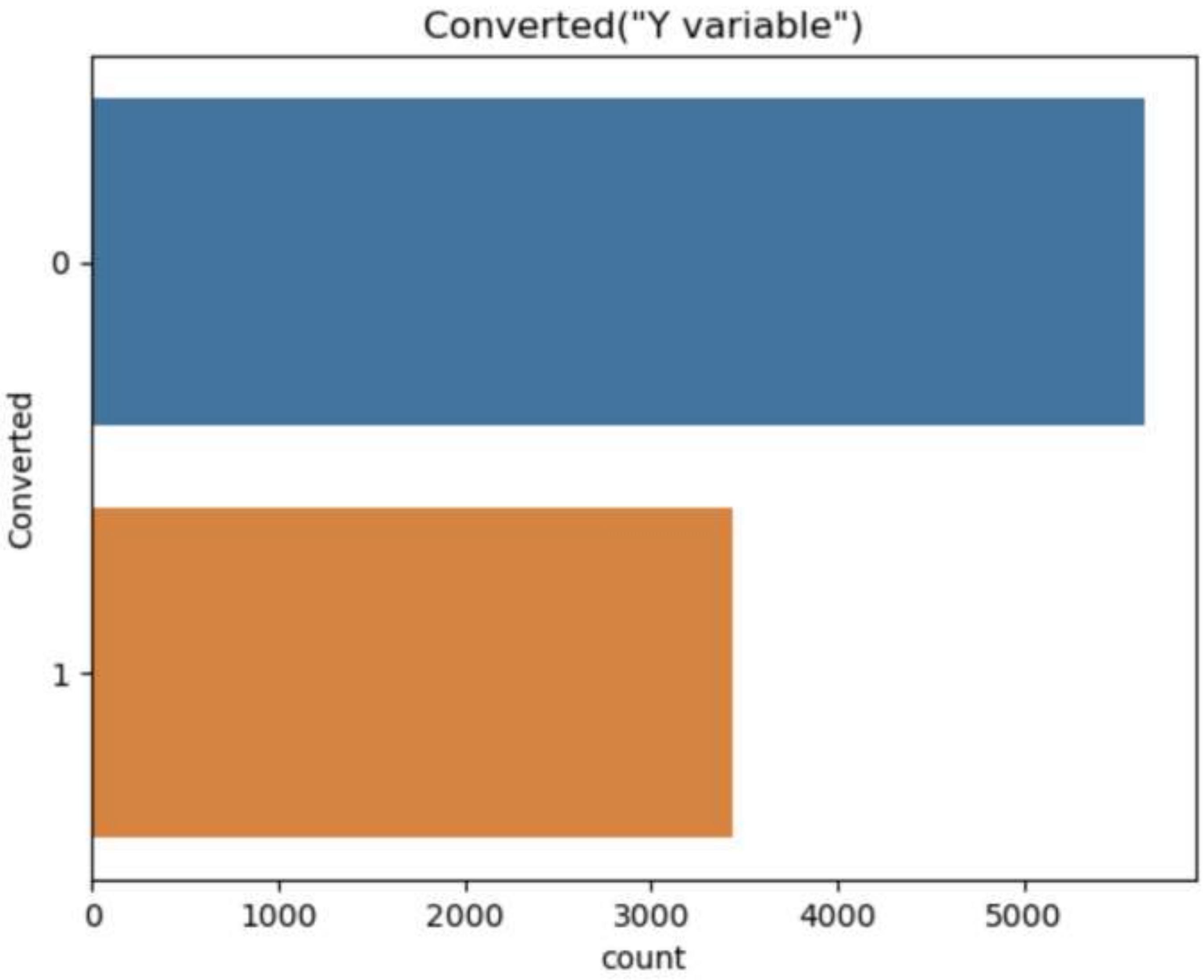
Categorical Analysis



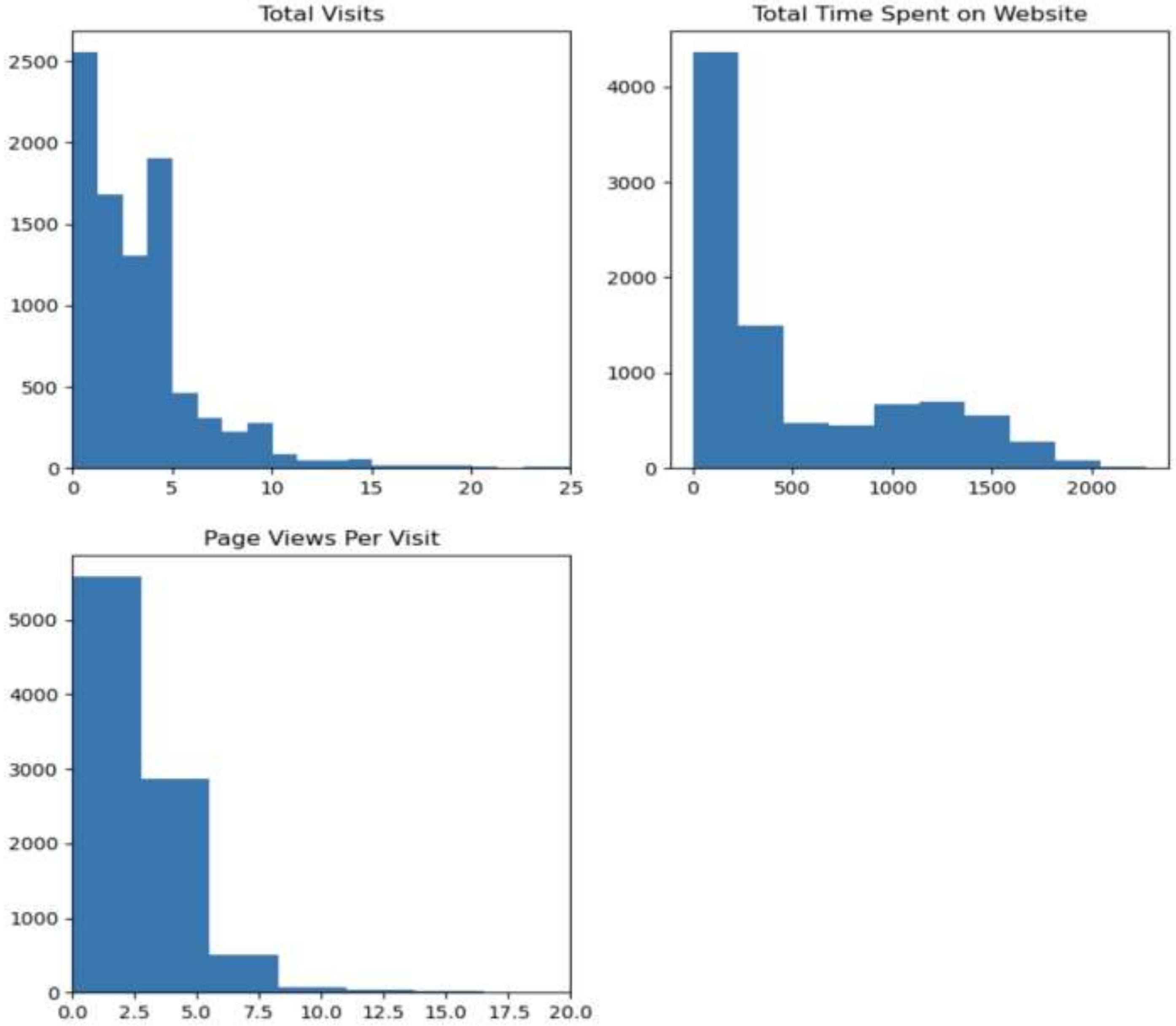

Lead Source
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source
0
500
1000
1500
2000
2500
3000
count

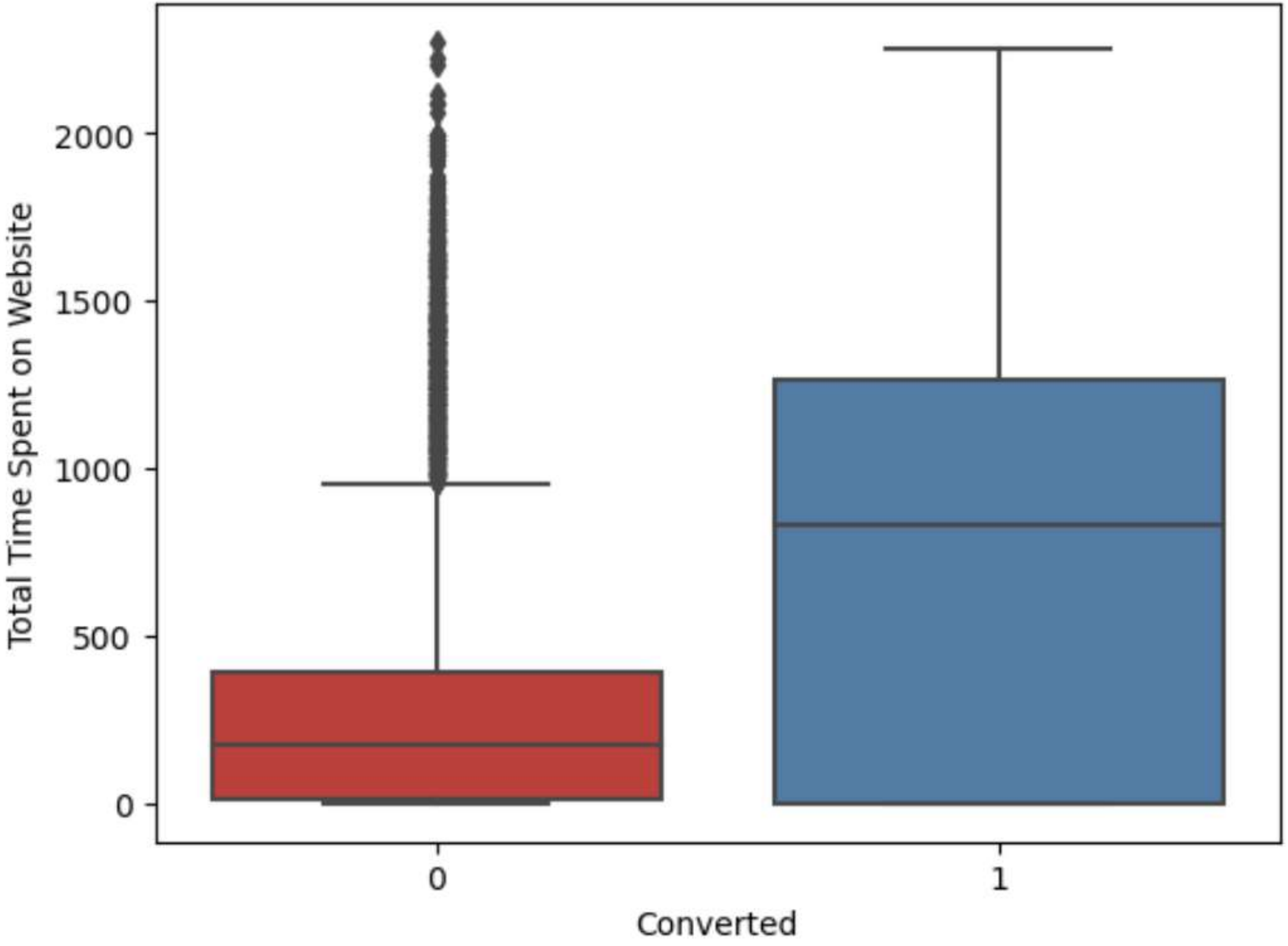

Specialization
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization
0
500
1000
1500
2000
2500
3000
count
Current Occupation
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation
0
1000
2000
3000
4000
5000
count


What matters most to you in choosing a course
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course
0
1000
2000
3000
4000
5000
6000
count
Last Activity
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity
0
500
1000
1500
2000
2500
3000
3500
count

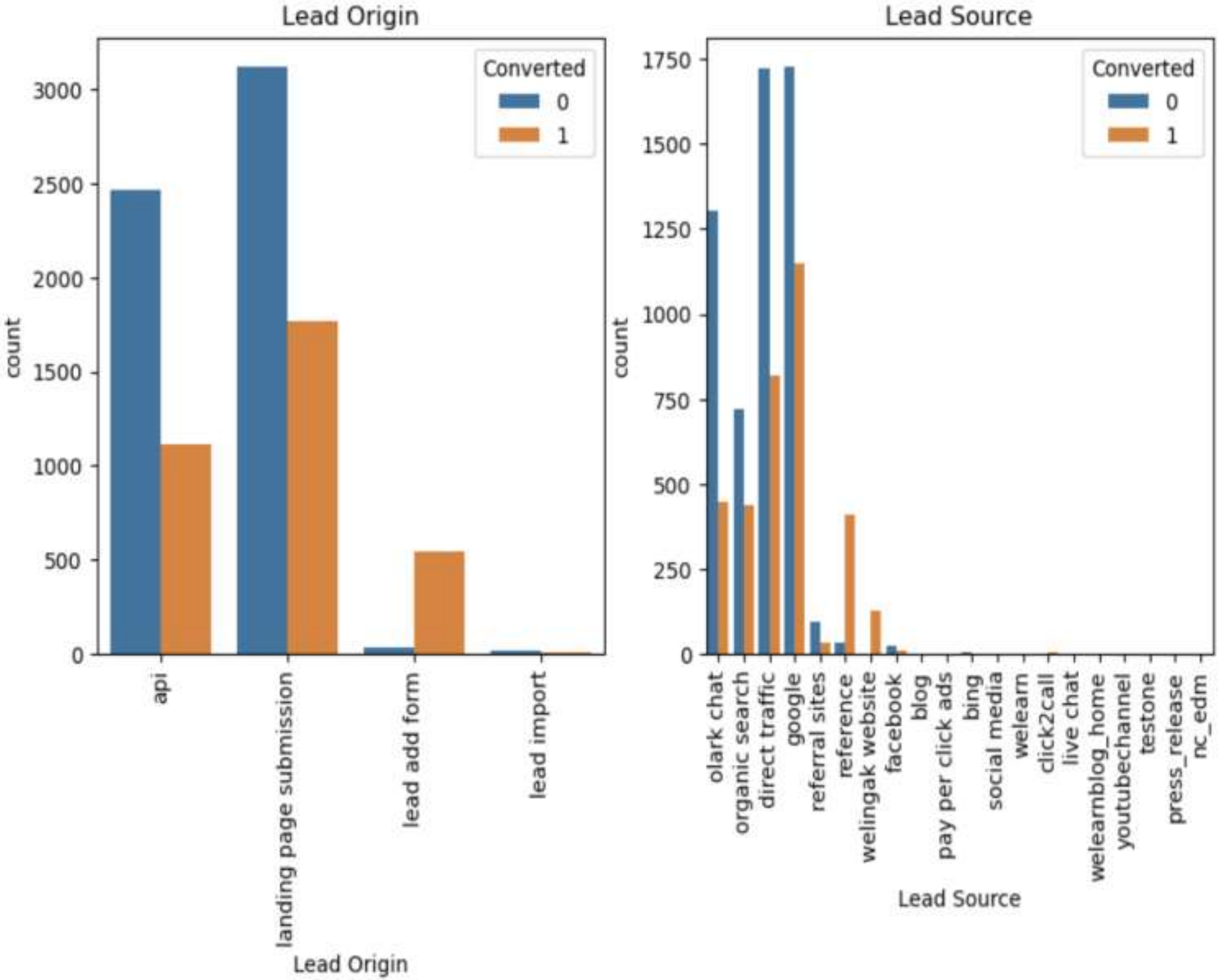

Converted("Y variable")
Converted
0
1
count
0
1000
2000
3000
4000
5000

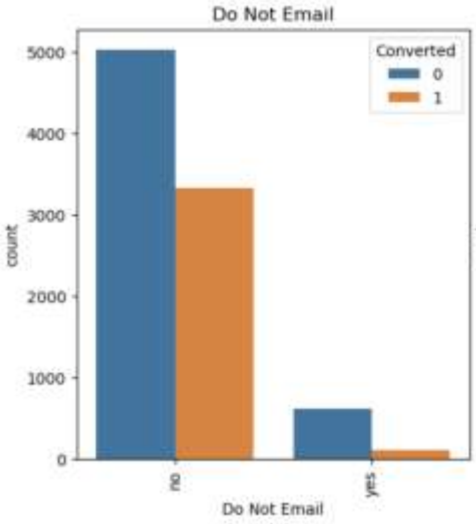
Numerical Variables



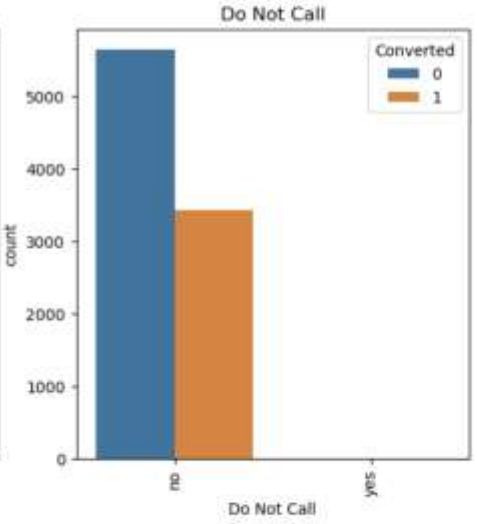

Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted

Bivariate Analysis

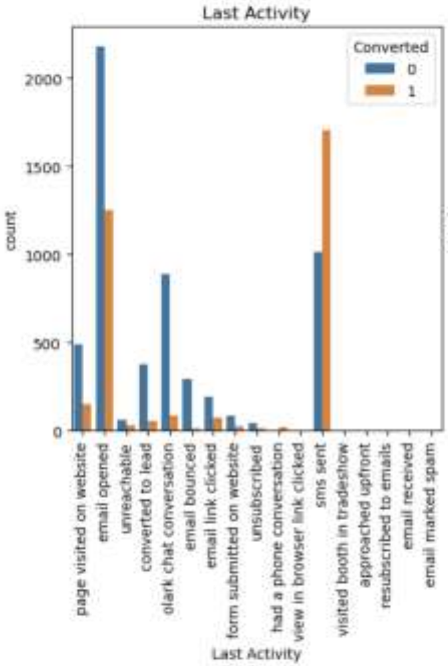




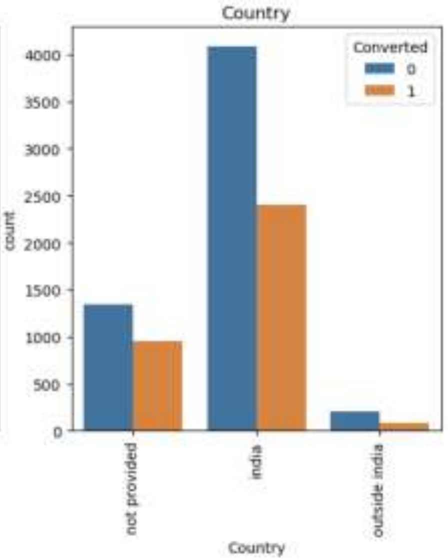
Do Not Email
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Do Not Email



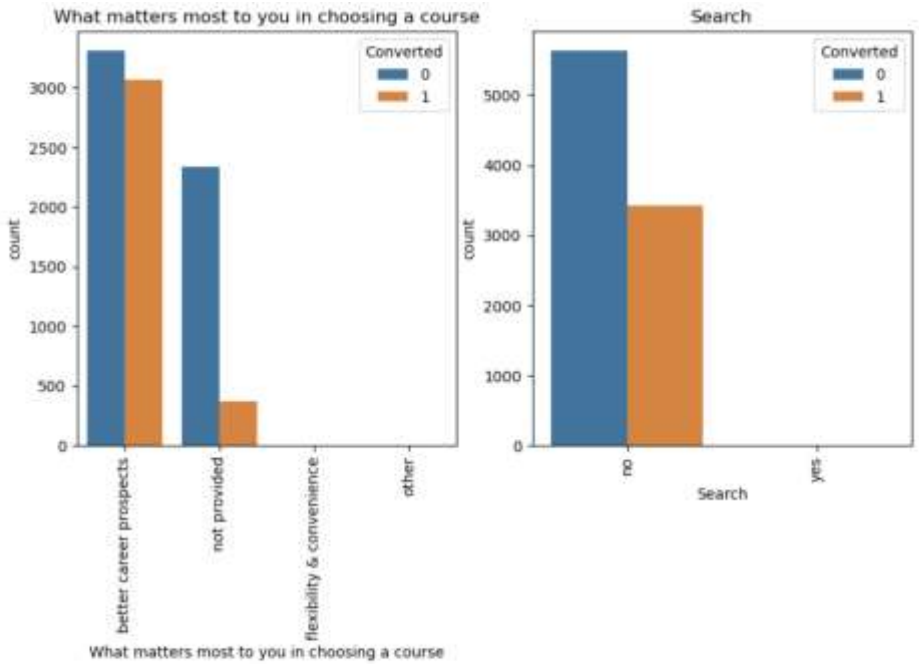
Do Not Call
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Do Not Call

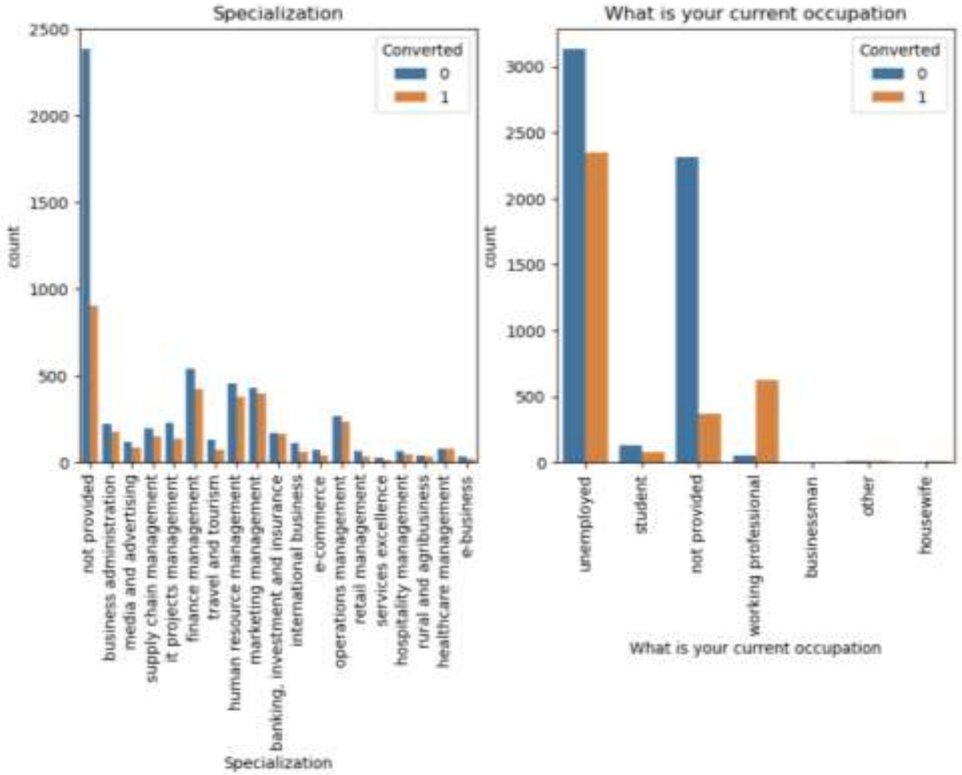


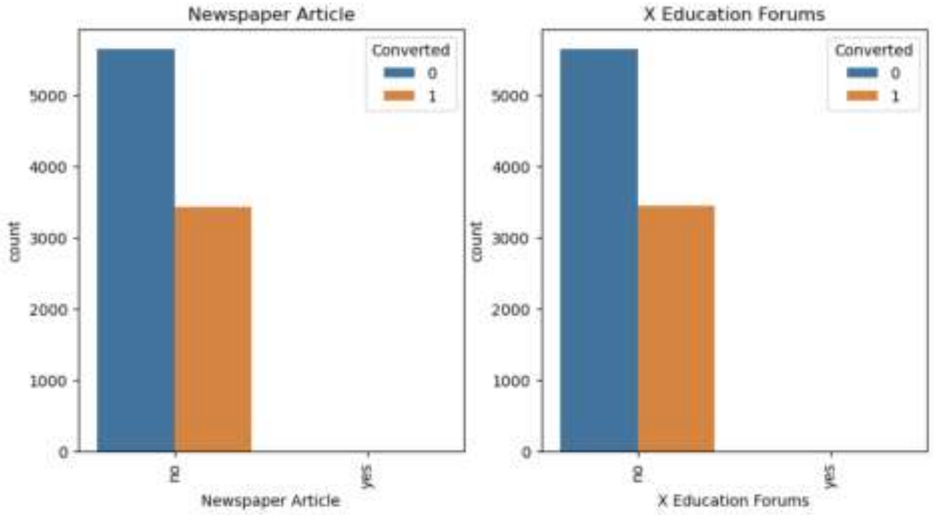
Last Activity
Converted
0
1
count
0
500
1000
1500
2000
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

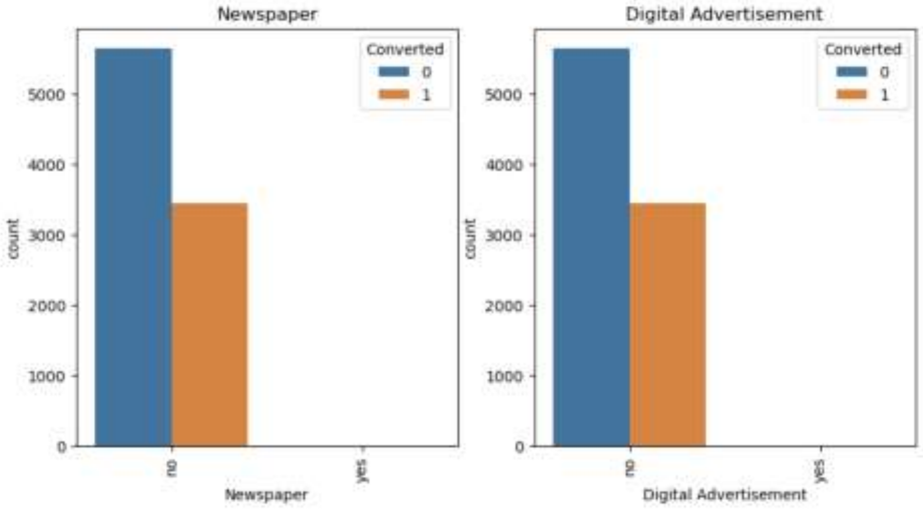


Country
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
3500
4000
not provided
india
outside india
Country


What matters most to you in choosing a course
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course
Search
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Search


Specialization
Converted
0
1
count
0
500
1000
1500
2000
2500
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization
What is your current occupation
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation


Newspaper Article
Converted
🟦 0
🟧 1
count
0
1000
2000
3000
4000
5000
no
yes
Newspaper Article
X Education Forums
Converted
🟦 0
🟧 1
count
0
1000
2000
3000
4000
5000
no
yes
X Education Forums


Newspaper
Converted
🟦 0
🟧 1
count
0
1000
2000
3000
4000
5000
no
yes
Newspaper
Digital Advertisement
Converted
🟦 0
🟧 1
count
0
1000
2000
3000
4000
5000
no
yes
Digital Advertisement

Inference :

❑ API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.

❑ Lead Add Form has more than 90% conversion rate but count of lead are not very high.

❑ Lead Import are very less in count.

To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

❑ Google and Direct traffic generates maximum number of leads.

❑ Conversion Rate of reference leads and leads through welingak website is high.

To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

❑ Leads spending more time on the weblise are more likely to be converted.

Website should be made more engaging to make leads spend more time.

❑ Most of the lead have their Email opened as their last activity.

❑ Conversion rate for leads with last activity as SMS Sent is almost 60%.

Model Building

- **Data Splitting:**
  - Perform a train-test split for regression, opting for a 70:30 ratio for training and testing datasets.
- **Feature Selection with RFE:**
  - Utilize Recursive Feature Elimination (RFE) to select 15 variables as output.
- **Model Building:**
  - Construct the regression model by excluding variables with p-values exceeding 0.05 and variance inflation factor (VIF) greater than 5.
- **Predictions on Test Data:**
  - Apply the model to make predictions on the test dataset.
- **Accuracy Evaluation:**
  - Assess overall accuracy, revealing a performance level of 81%.

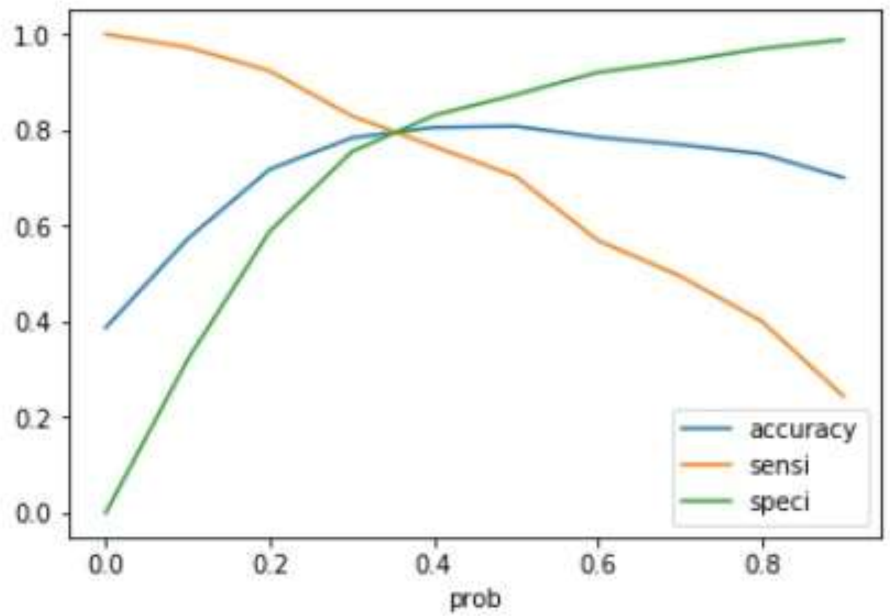
ROC Curve

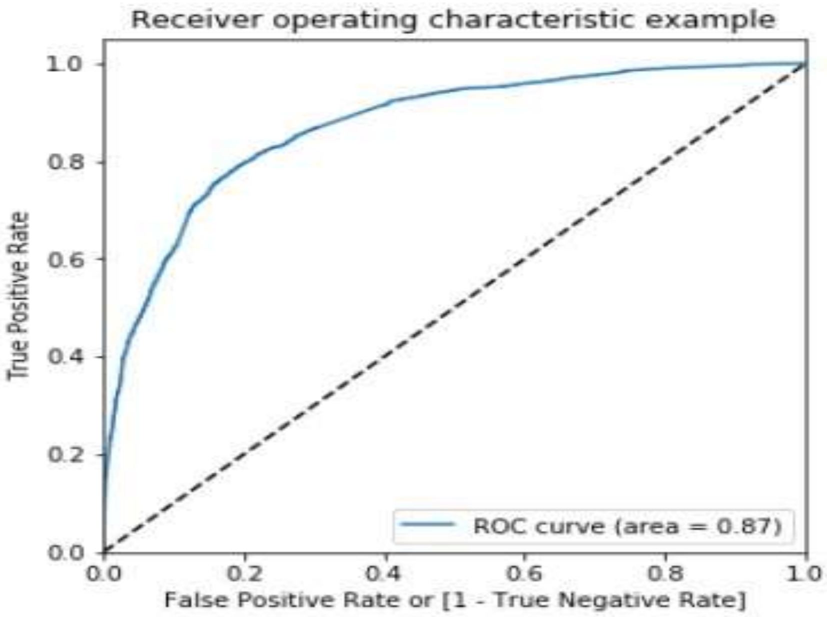
- ❑ **Finding Optimal Cut-off Point:**
  - ❑ Determine the ideal threshold for classification.

- ❑ **Optimal Cut-off Probability:**
  - ❑ Identify the probability value that achieves a balance between sensitivity and specificity.

- ❑ **Balanced Sensitivity and Specificity:**
  - ❑ Seek the probability threshold that ensures a harmonious trade-off between sensitivity and specificity.

- ❑ **Graphical Insight:**
  - ❑ Visual examination of the second graph indicates that the optimal cut-off point is observed at a probability of 0.35.





Results

1) Comparing the values obtained for Train & Test:

Train Data:

- ❑ Accuracy : 80.82 %
- ❑ Sensitivity : 81.31 %
- ❑ Specificity : 79.56 %

Test Data:

- ❑ Accuracy : 80.82 %
- ❑ Sensitivity : 82.32 %
- ❑ Specificity : 79.98 %

Thus, we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% . The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.

2) Finding out the leads which should be contacted:

- ❑ The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85. They can be termed as 'Hot Leads'.

Recommendations

- ❑ The company **should make calls** to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- ❑ The company **should make calls** to the leads who are the "working professionals" as they are more likely to get converted.
- ❑ The company **should make calls** to the leads who spent "more time on the websites" as these are more likely to get converted.
- ❑ The company **should make calls** to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- ❑ The company **should make calls** to the leads whose last activity was SMS Sent as they are more likely to get converted.
- ❑ The company **should not make calls** to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- ❑ The company **should not make calls** to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- ❑ The company **should not make calls** to the leads whose Specialization was "Others" as they are not likely to get converted.
- ❑ The company **should not make calls** to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.