Lead score case study

Group Member

1.Priyanka Kamlallu

2.Vivek Yadav

Problem Statement

- X Education sells online courses to industry professionals. X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

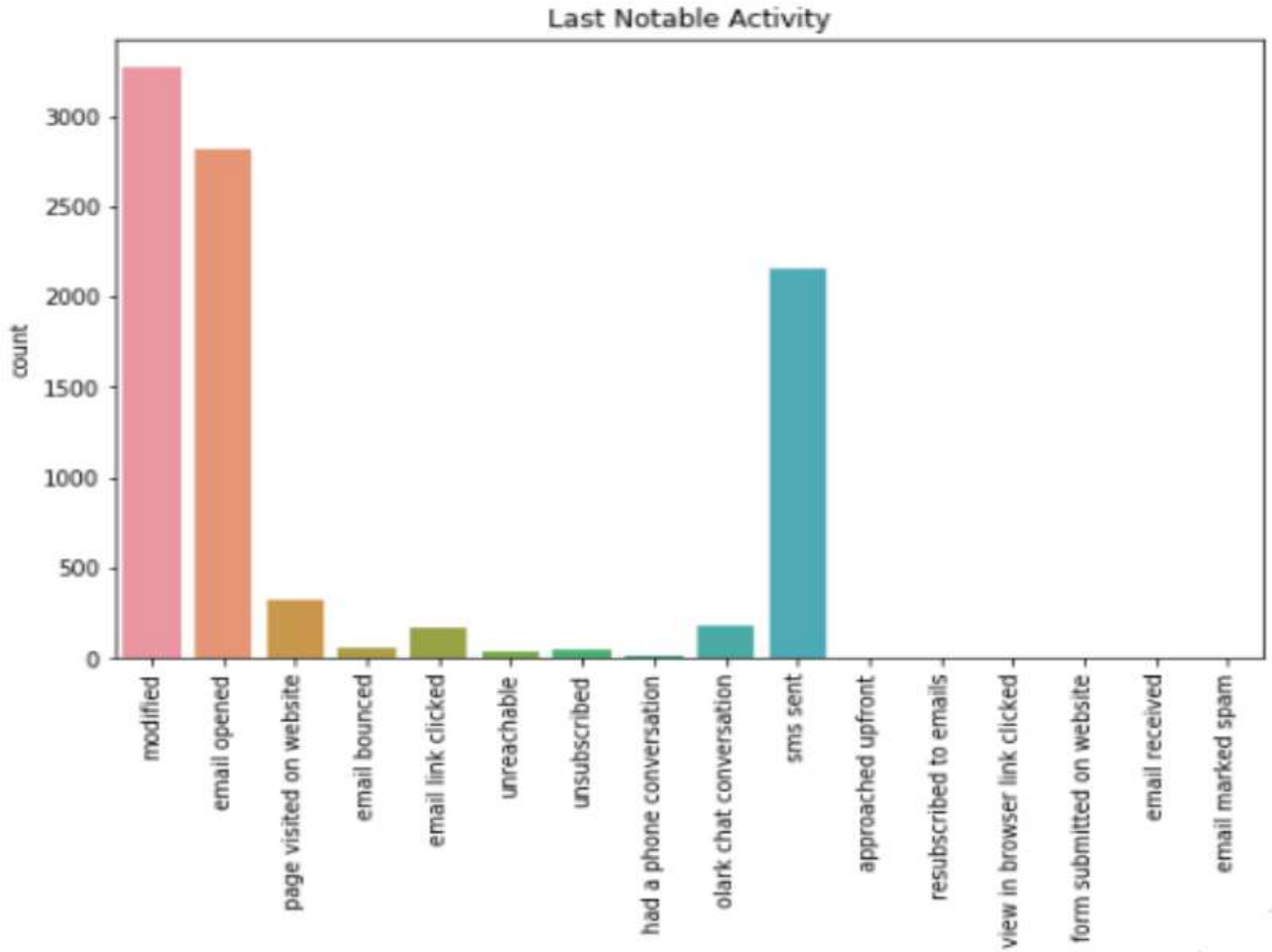
Business Objective:

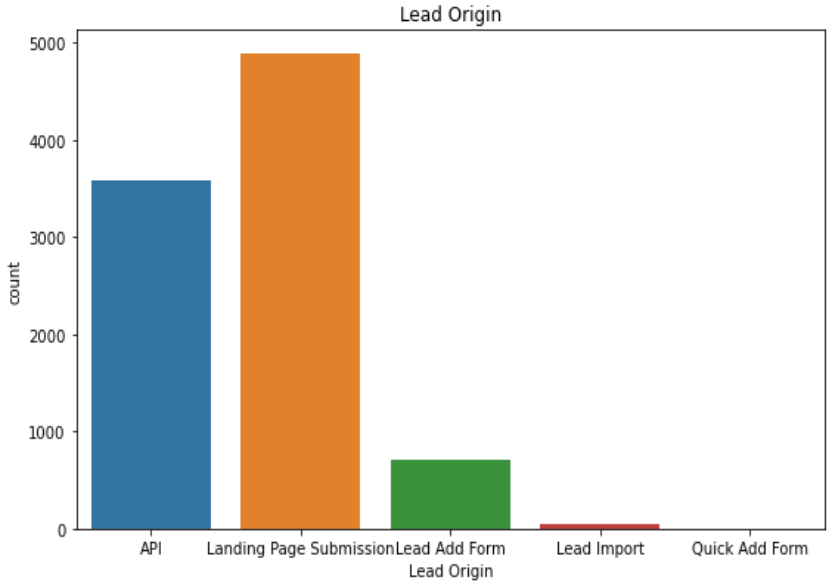
- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

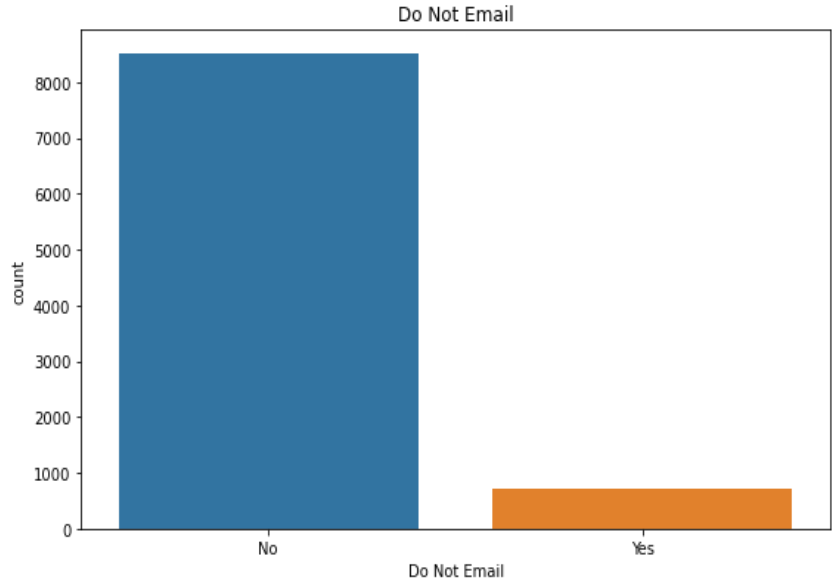
Solution Methodology

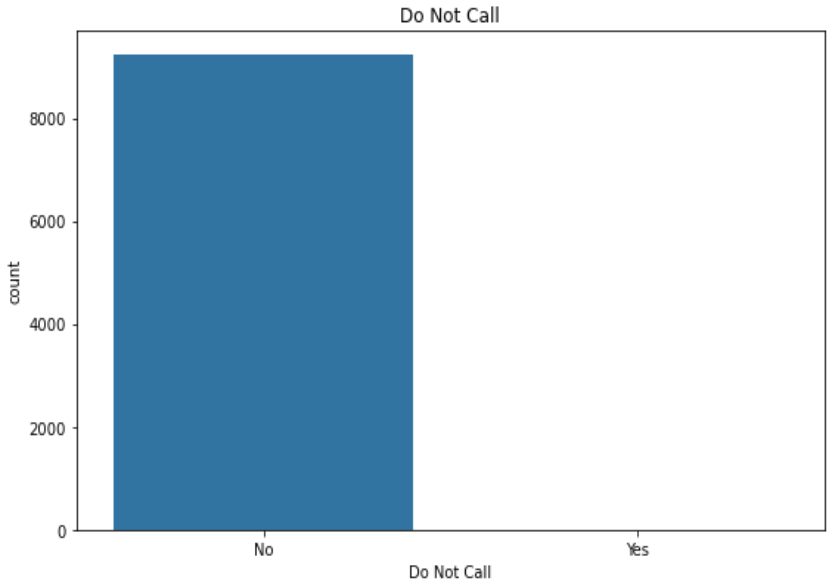
- Data cleaning and data manipulation.
- 1. Check and handle duplicate data.
- 2. Check and handle NA values and missing values.
- 3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
- 4. Imputation of the values, if necessary.
- 5. Check and handle outliers in data. EDA 1. Univariate data analysis: value count, distribution of variable etc. 2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model. Model presentation.
- Conclusions and recommendations.

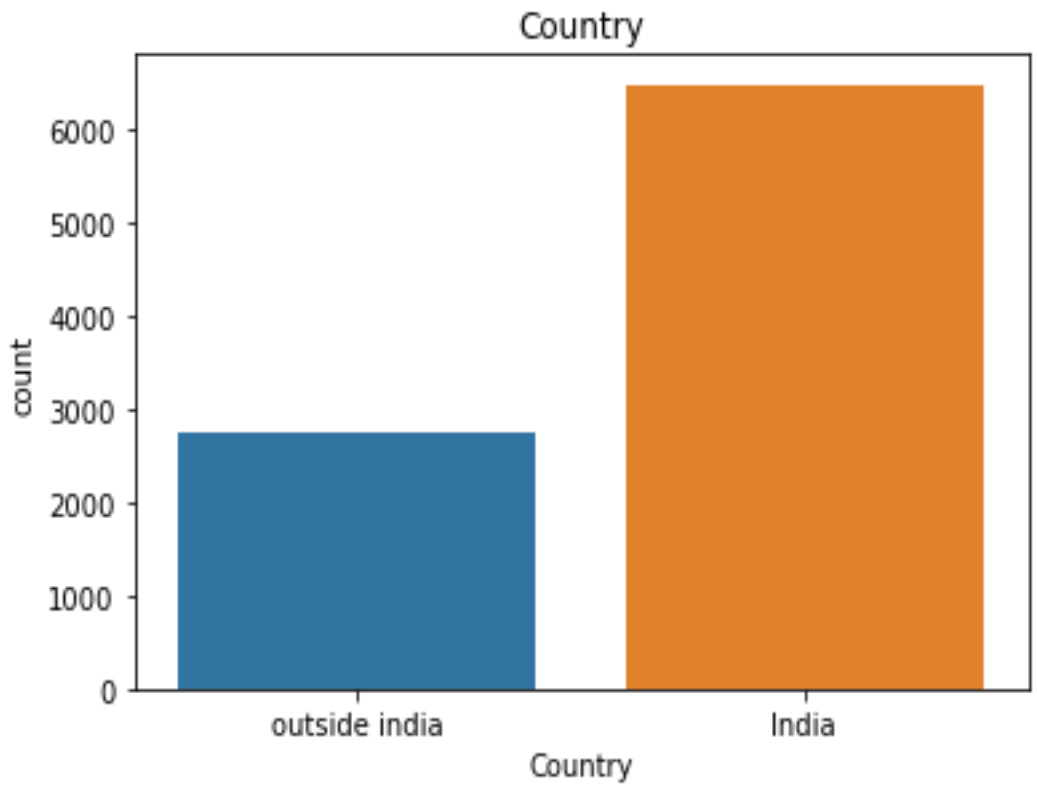
EDA

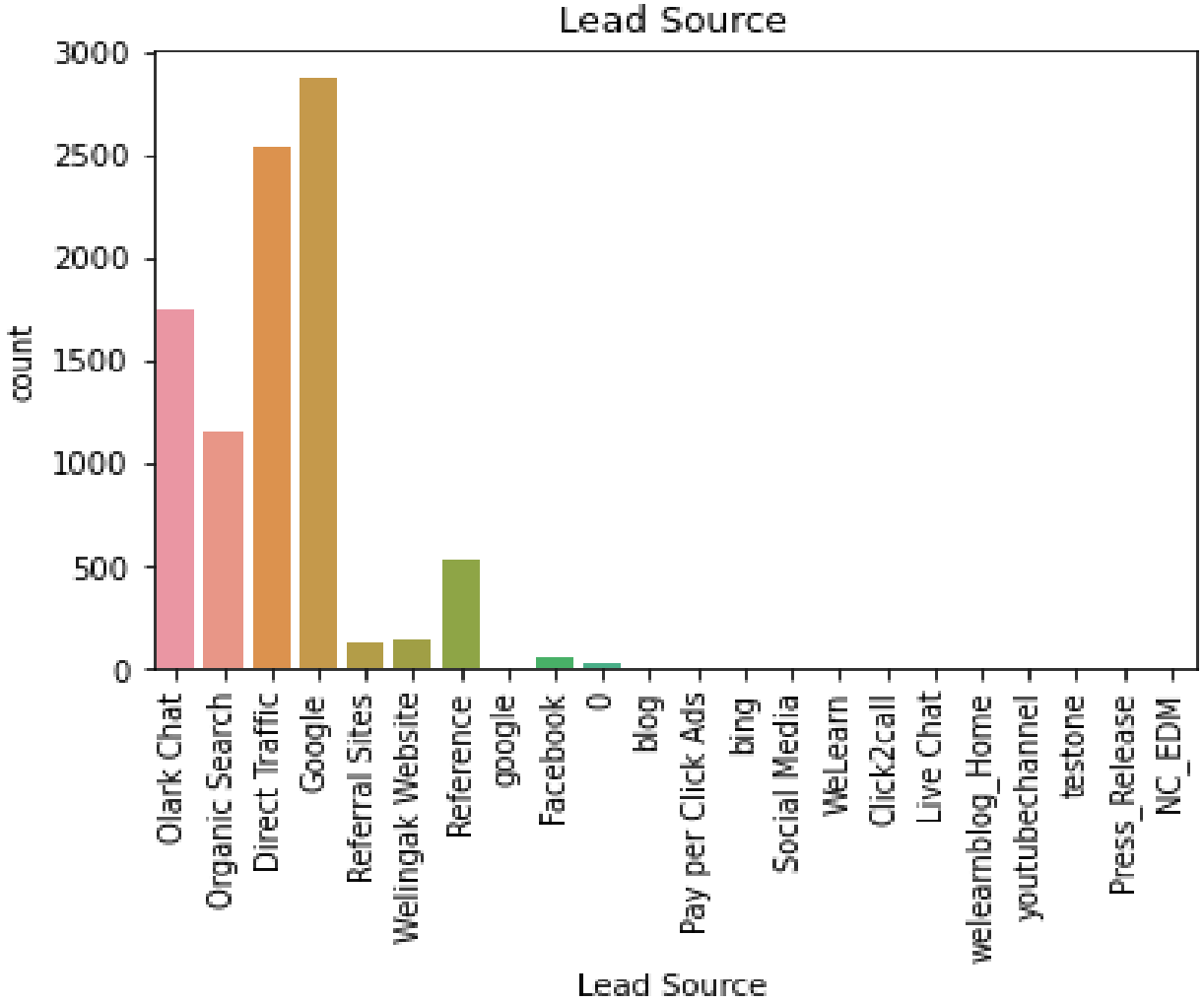

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam

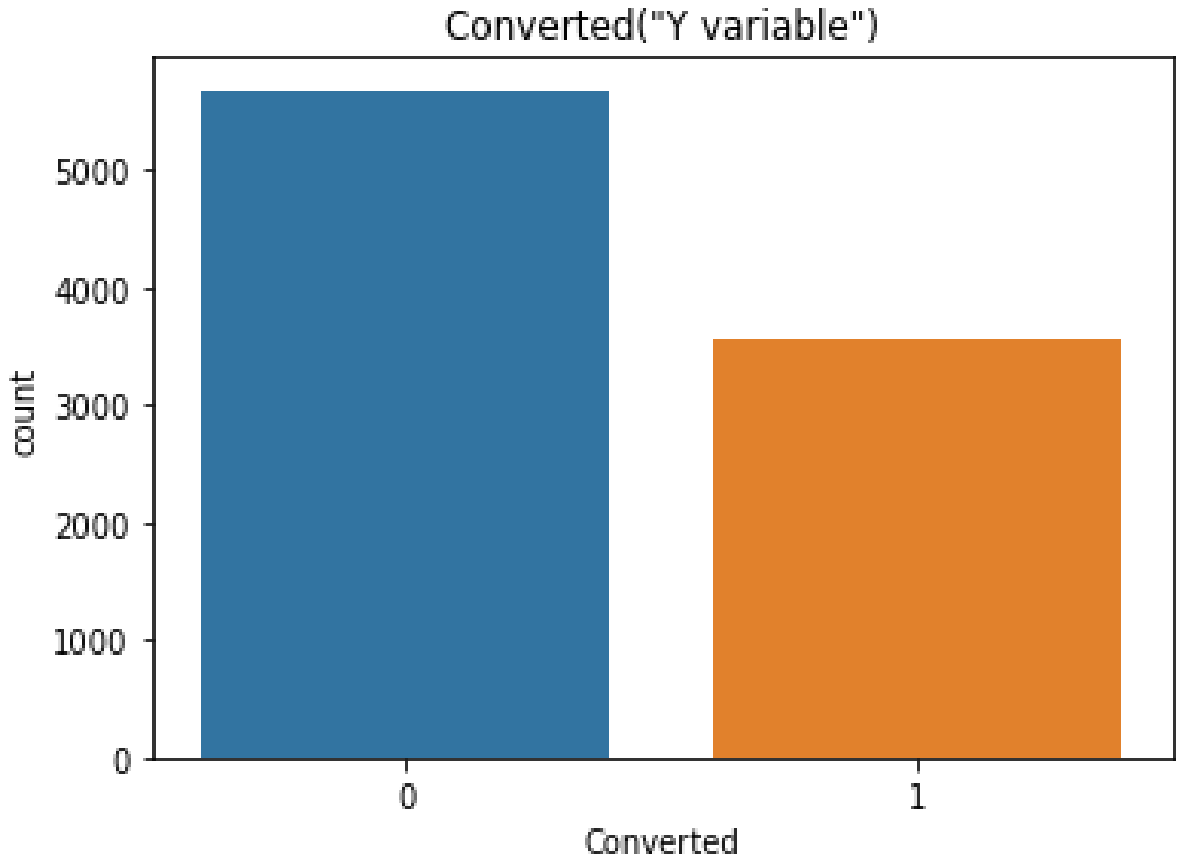

Lead Origin
count
5000
4000
3000
2000
1000
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

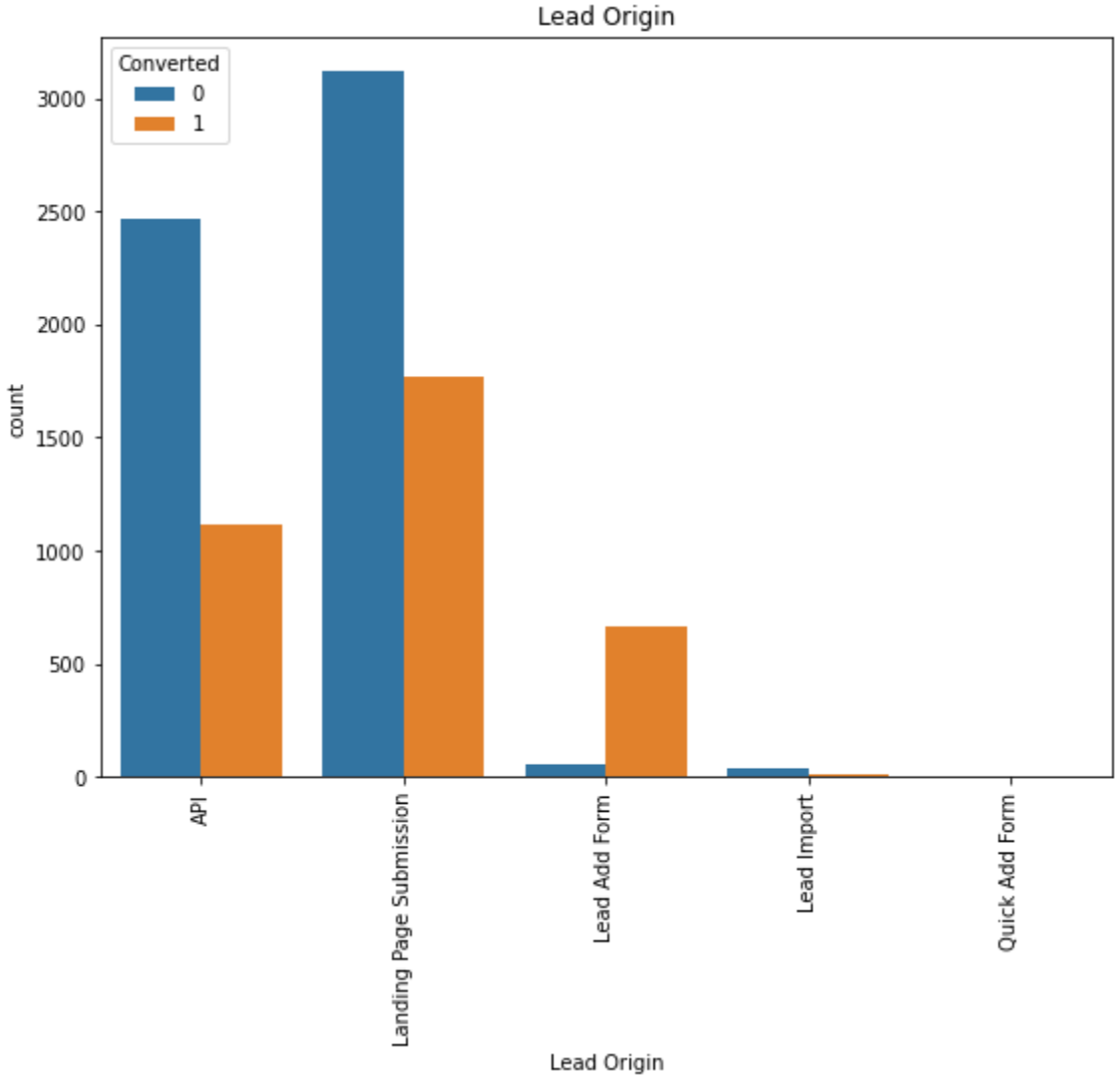

Do Not Call
count
8000
6000
4000
2000
0
No
Yes
Do Not Call

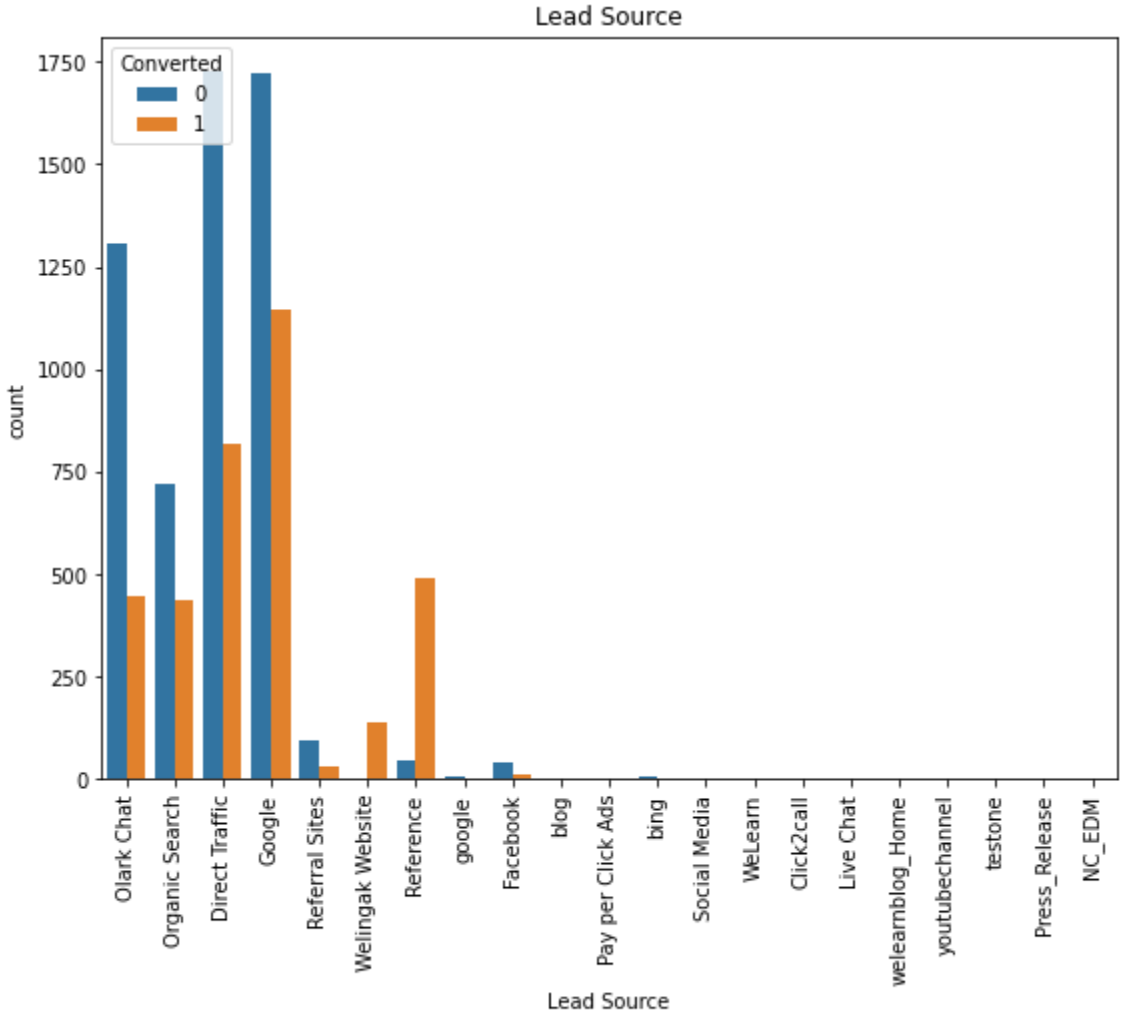

Country
6000
5000
4000
3000
2000
1000
0
count
outside india
India
Country

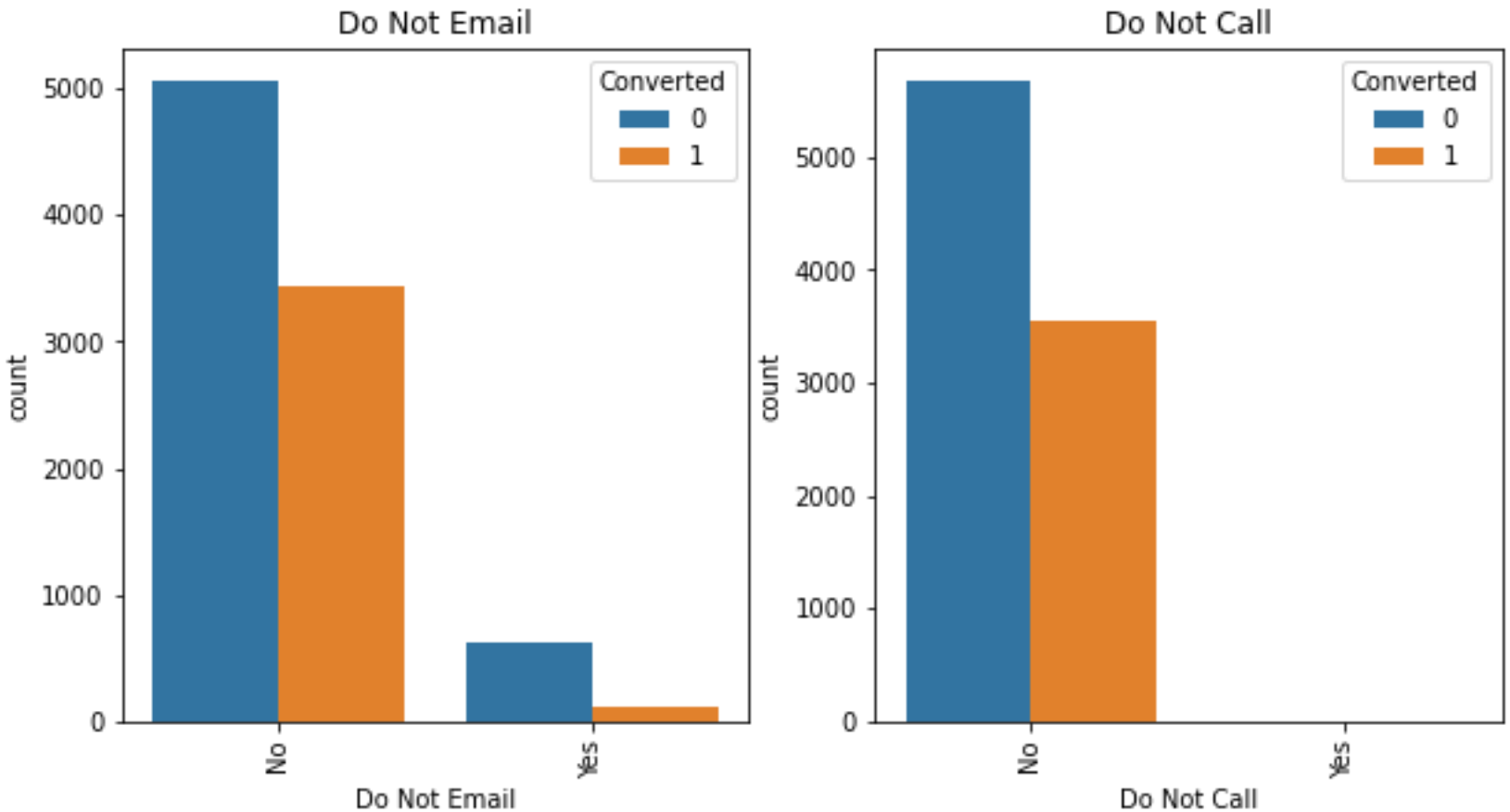

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
0
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

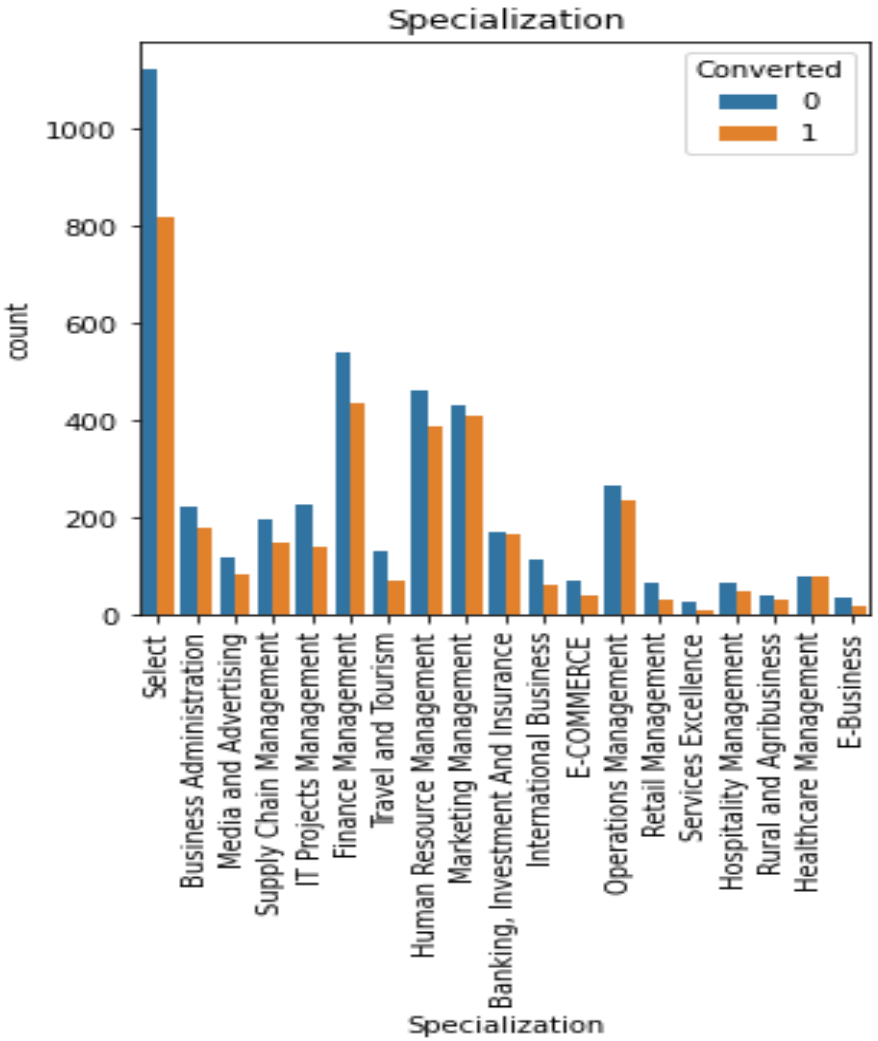

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

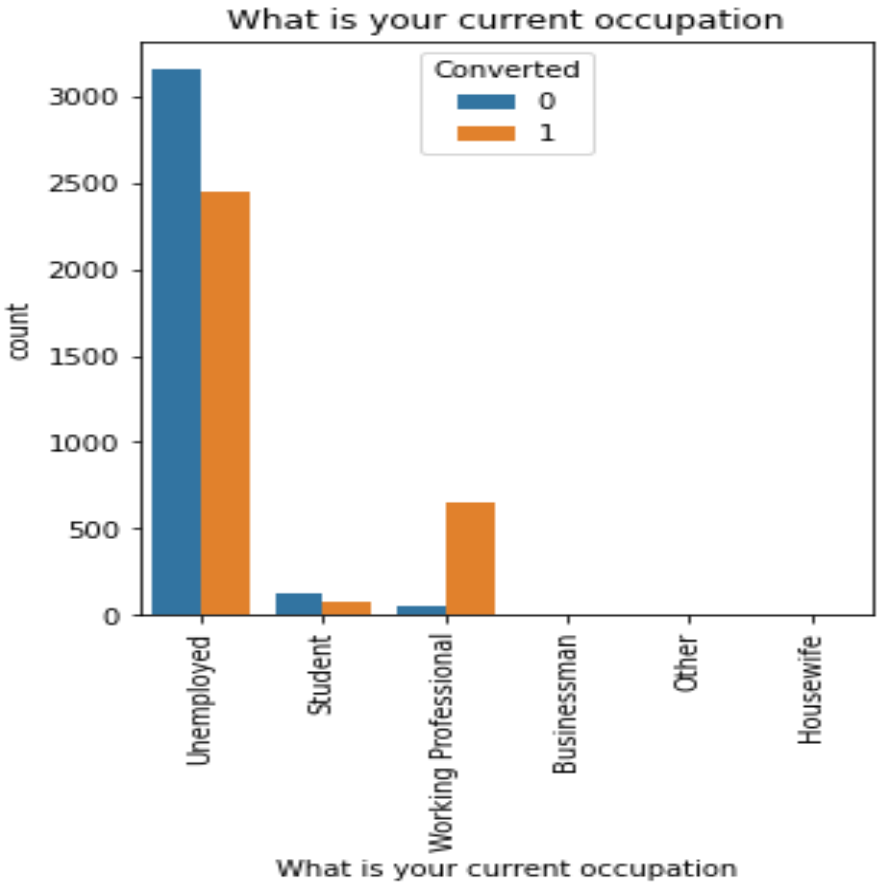
Categorical Variable Relation

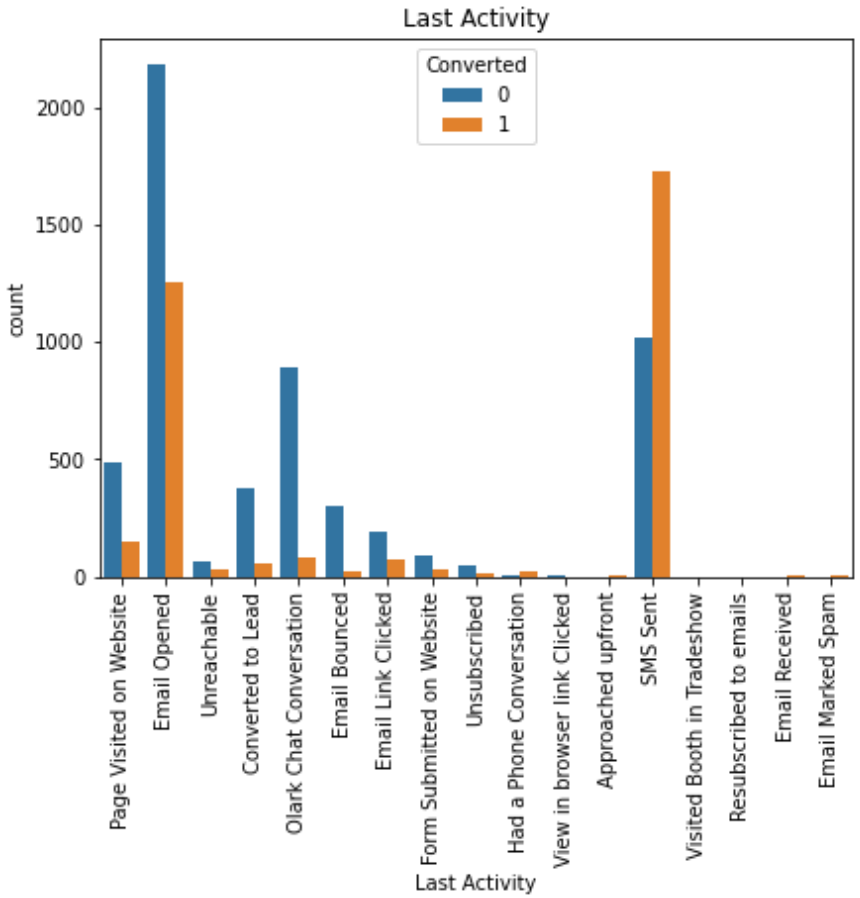


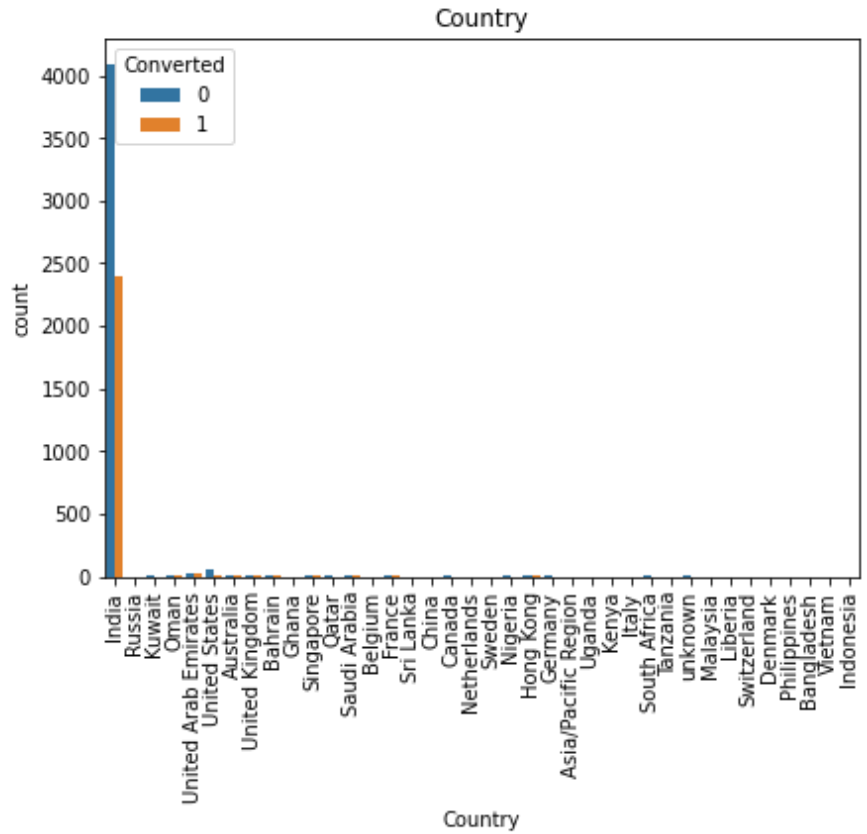


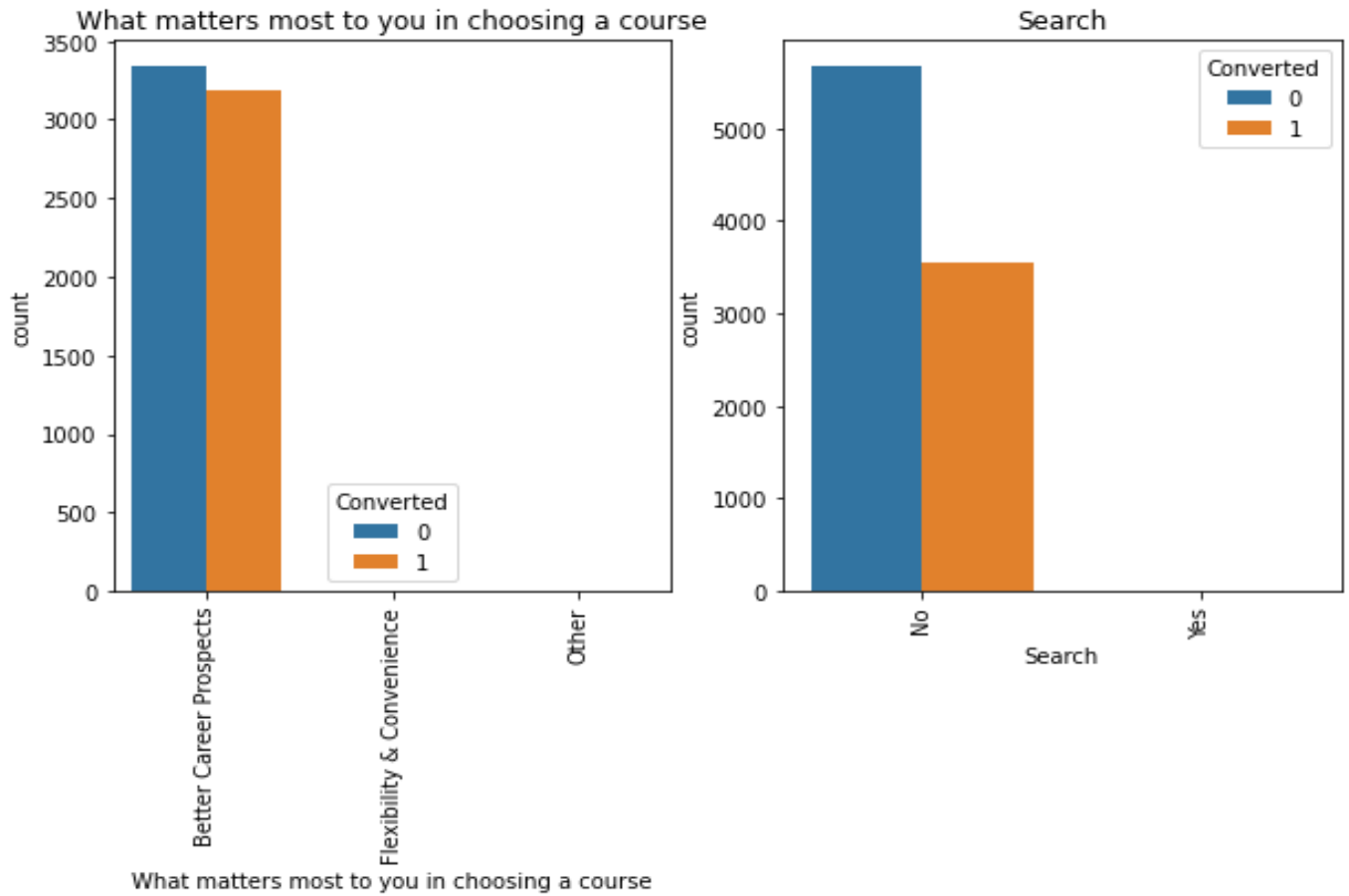

Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email
Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call

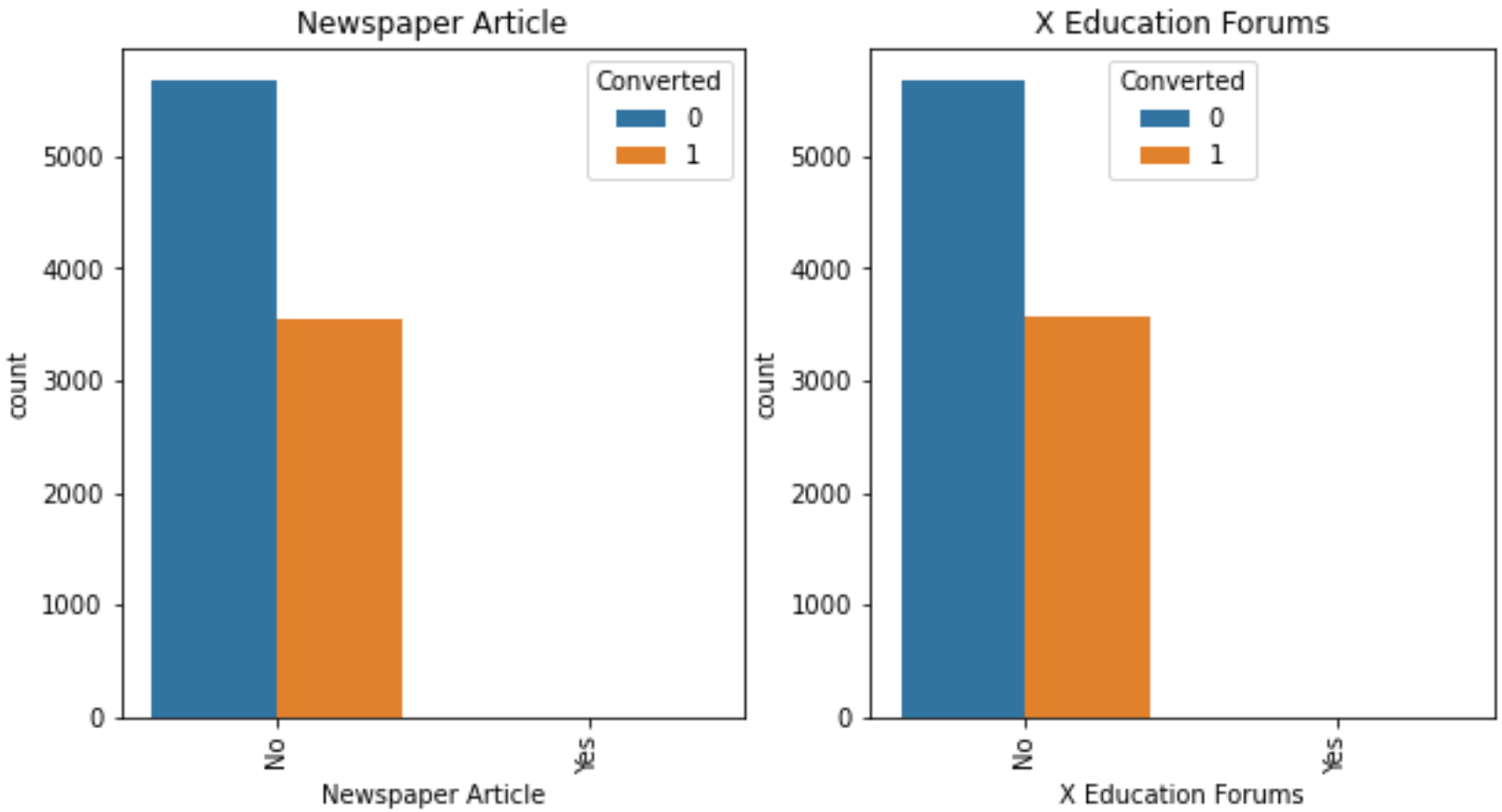

Specialization
Converted
0
1
count
1000
800
600
400
200
0
Select
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

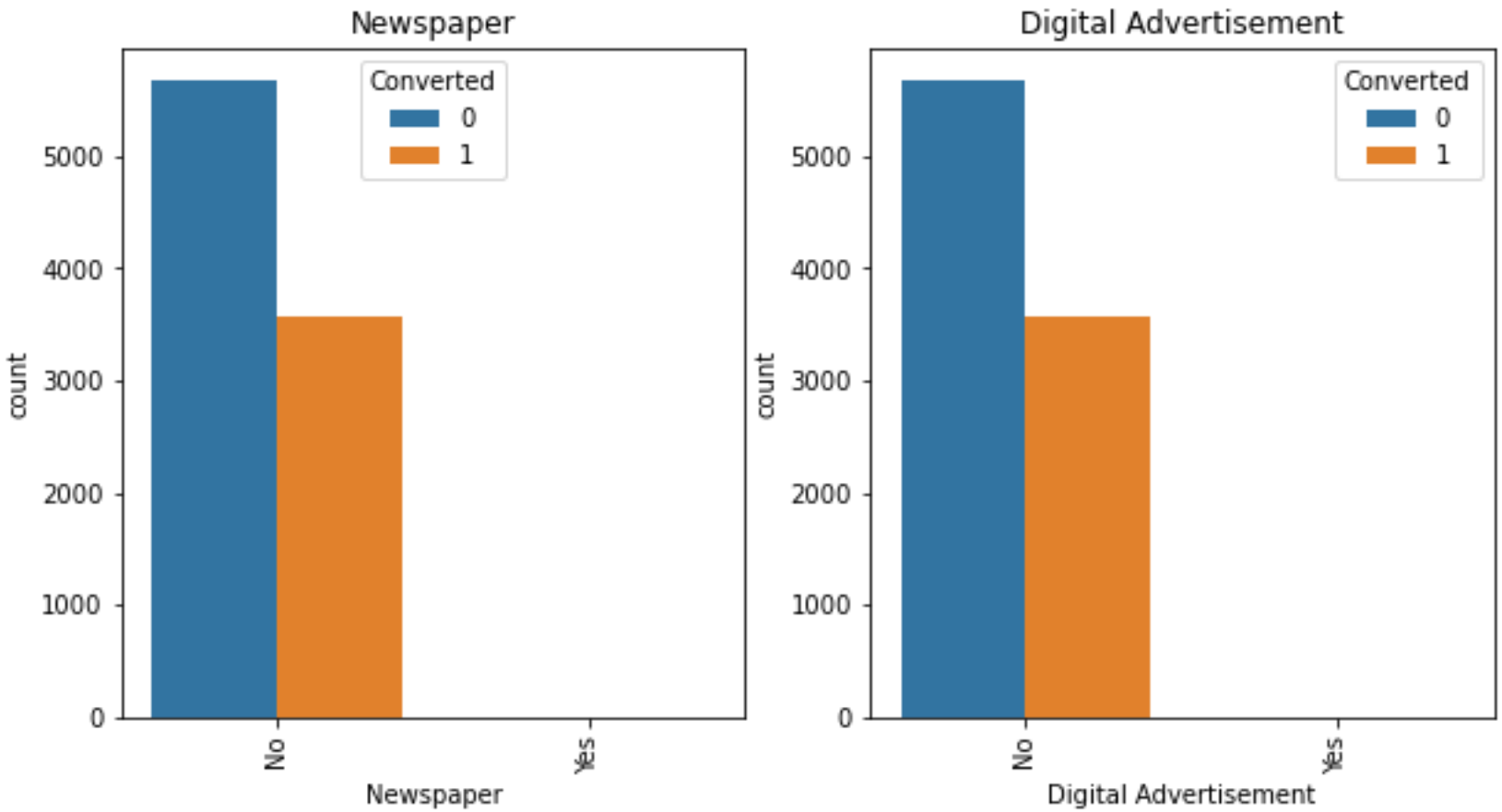

What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

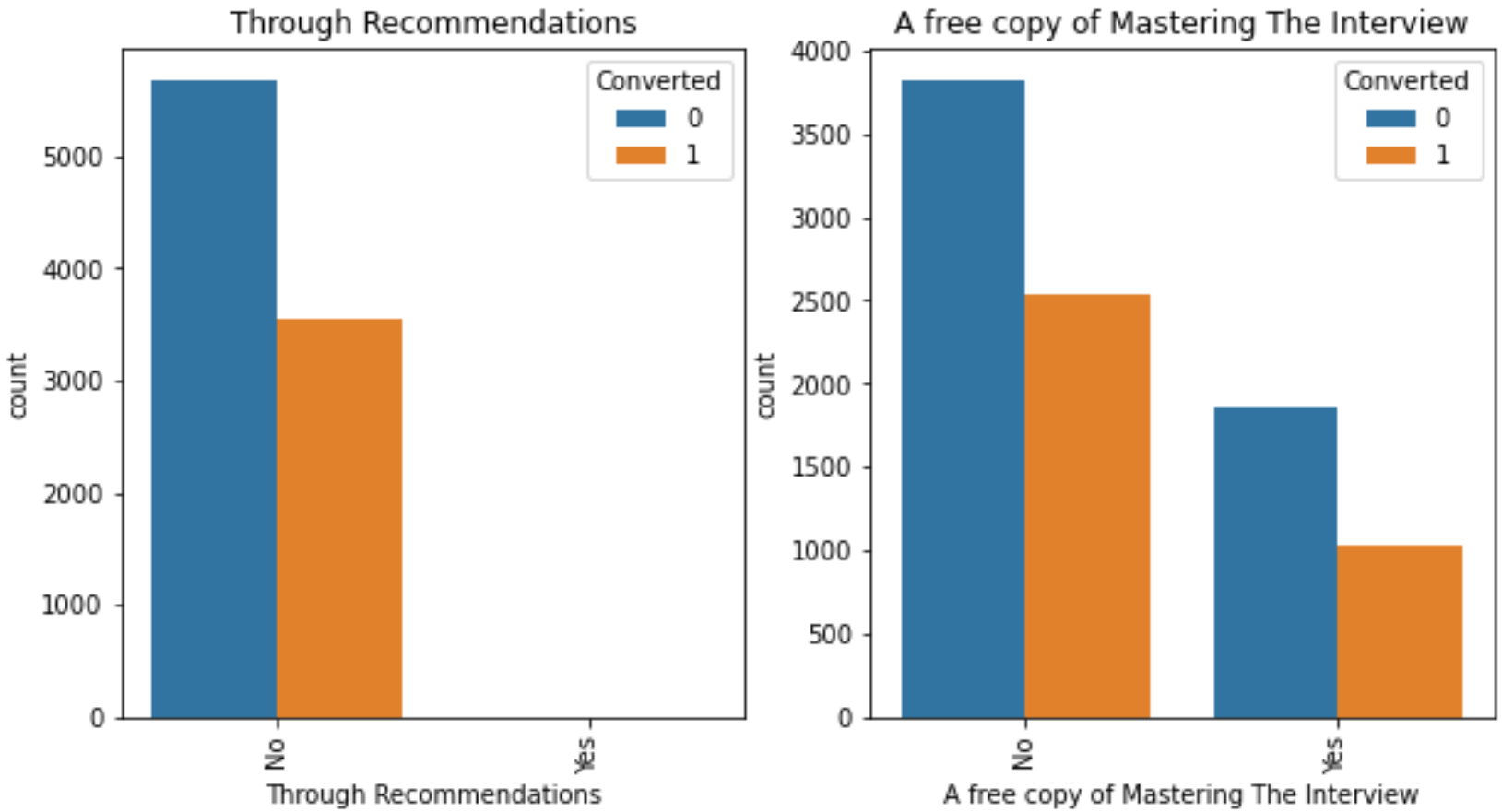

Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

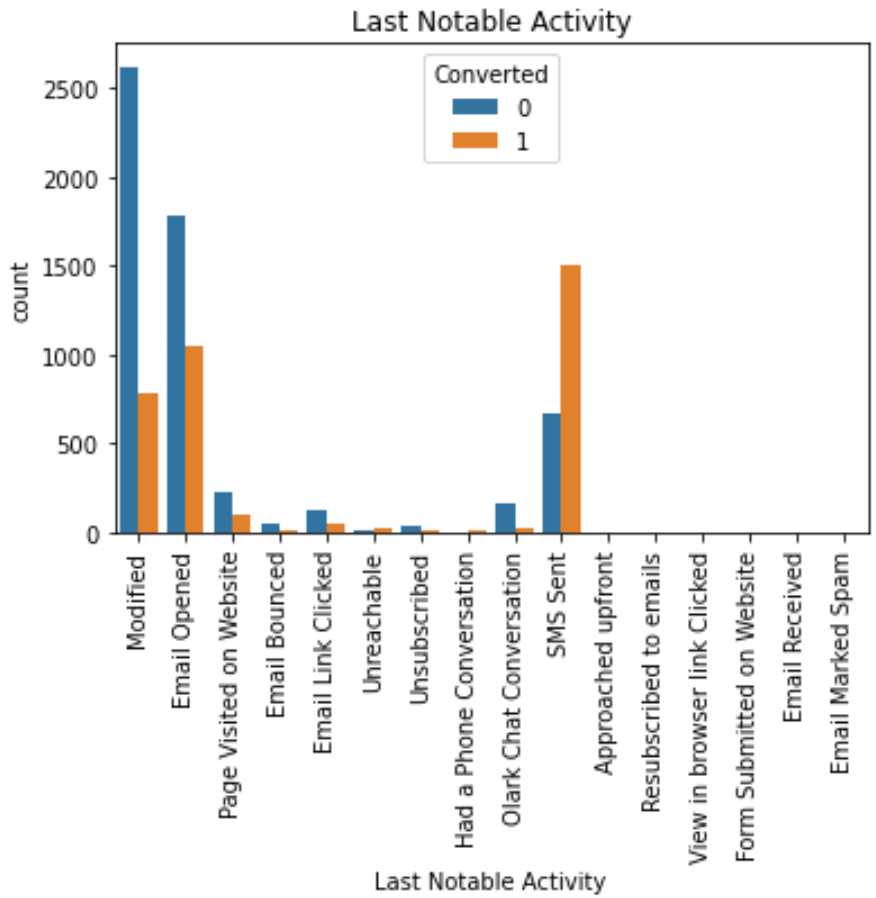

Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
India
Russia
Kuwait
Oman
United Arab Emirates
United States
Australia
United Kingdom
Bahrain
Ghana
Singapore
Qatar
Saudi Arabia
Belgium
France
Sri Lanka
China
Canada
Netherlands
Sweden
Nigeria
Hong Kong
Germany
Asia/Pacific Region
Uganda
Kenya
Italy
South Africa
Tanzania
unknown
Malaysia
Liberia
Switzerland
Denmark
Philippines
Bangladesh
Vietnam
Indonesia
Country


What matters most to you in choosing a course
count
3500
3000
2500
2000
1500
1000
500
0
Converted
0
1
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course
Search
count
5000
4000
3000
2000
1000
0
Converted
0
1
No
Yes
Search


Newspaper Article
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Newspaper Article
X Education Forums
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
X Education Forums

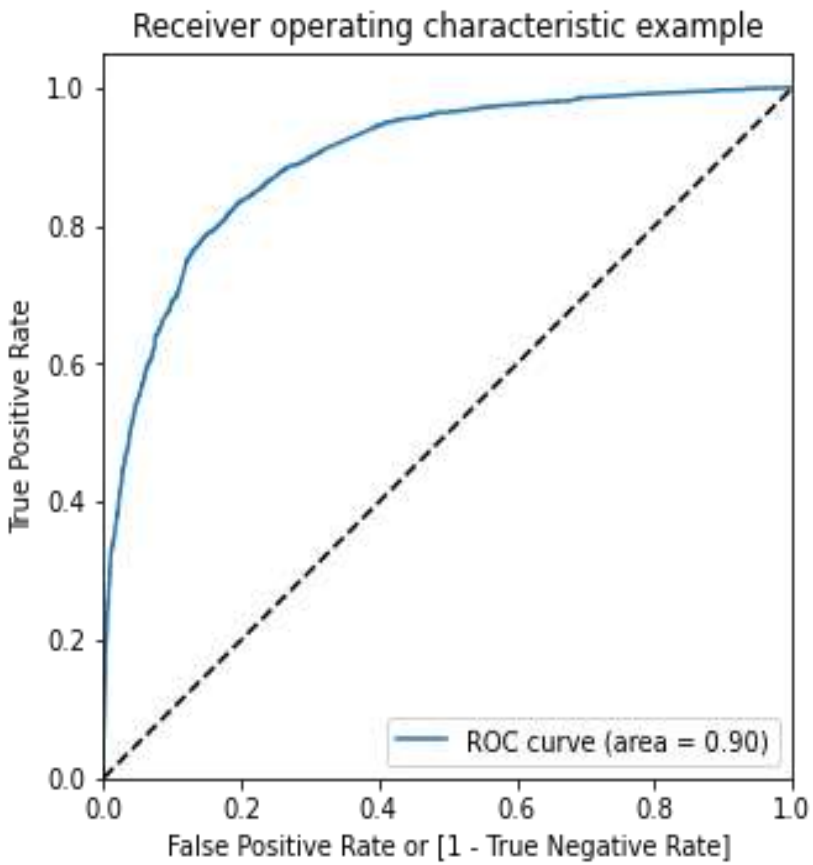

Newspaper
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Newspaper
Digital Advertisement
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Digital Advertisement


Through Recommendations
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Through Recommendations
A free copy of Mastering The Interview
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
A free copy of Mastering The Interview


Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Running RFE with 15 variables as output.
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater.
- Predictions on test data set.
- Overall accuracy 82%.


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.90)
0.0
0.2
0.4
0.6
0.8
1.0

Finding Optimal Cut off Point
Optimal cut off probability is that
probability where we get balanced sensitivity and specificity.