LEAD SCORING CASE STUDY

Group members

1. Juhi Rathi
2. Shaleen Bakshi
3. Souvik Sarkar

PROBLEM STATEMENT AND BUSINESS OBJECTIVE

Education company named X Education sells online courses to industry professionals. Although 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 effective, company wishes to identify the most potential leads, aka 'Hot Leads'. If the company can successfully identify 'Hot Leads', they may focus more on those to get highest conversion rate.

Business Objective :

X education wants to identify the most promising Leads. As step, company wants to build model which can help them to identify the hot leads and Deployment of the model for future use

PROPOSED SOLUTION

Get the historic data. Before make use of the data set follow the steps as follows

- Handle duplicate data if any
- Handle NA and missing values in the data set
- Handle features which have higher percentage of missing values
- Imputation of the values, if at all necessary
- Check for outliers and handle those if required

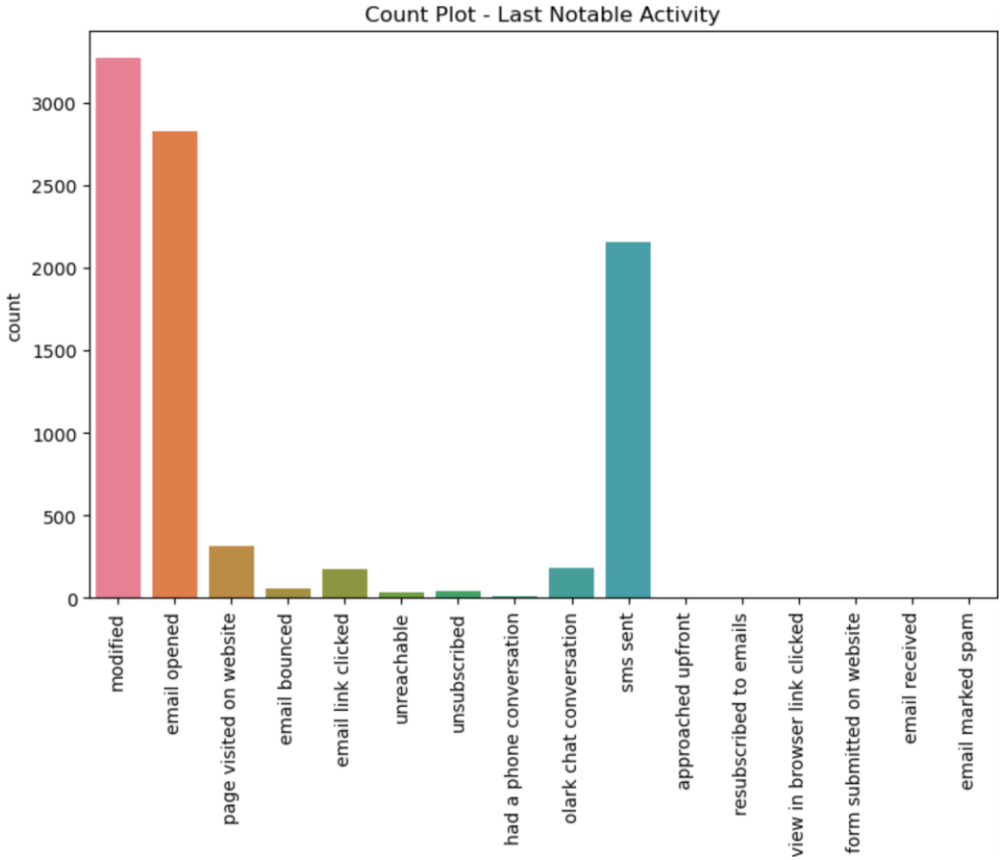
- EDA
  - Univariate Analysis – value count, distribution of the variables
  - Bivariate Analysis – Correlation coeff and relationship pattern between the variables
- Feature Scaling and Dummy variables and encoding
- Logistic regression for model building and prediction
- Validation of the model
- Model Presentation
- Recommendations

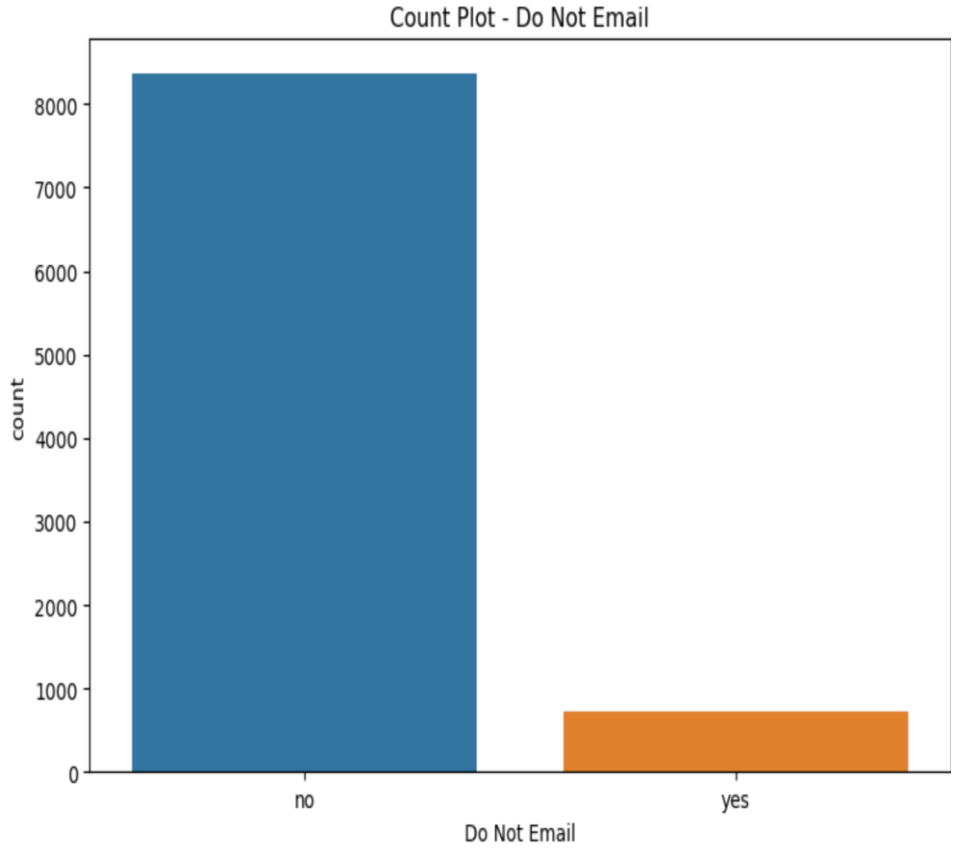
DATA & ASSUMPTIONS

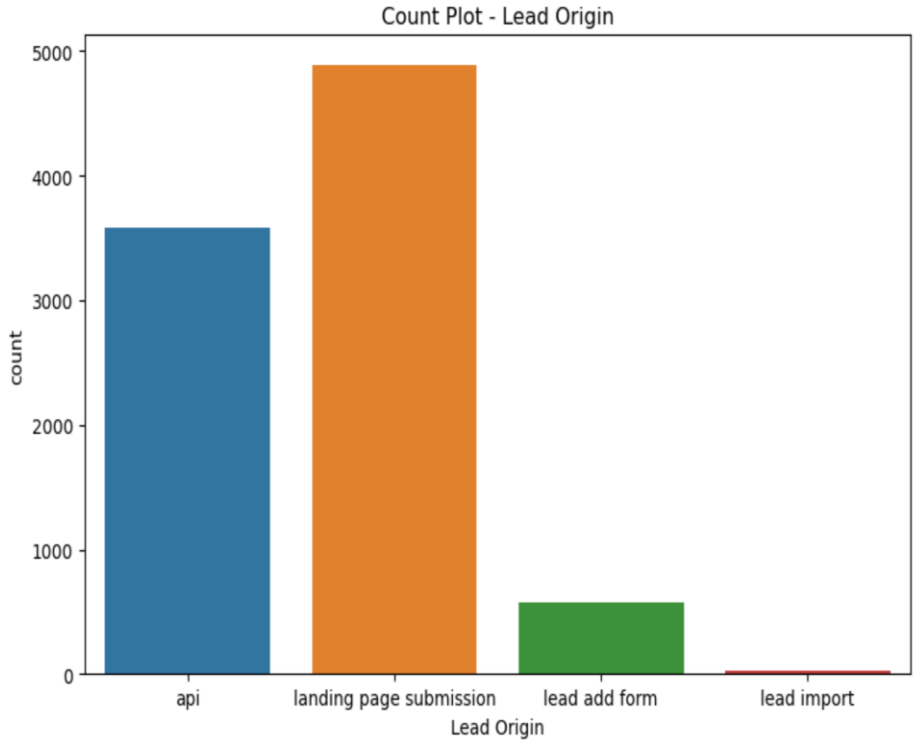
- In given data set 37 columns and 9240 number of rows were present
- 5 columns [*'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'*] dropped as they have only 1 unique data
- 2 columns [*'Prospect ID', 'Lead Number'*] having no repetitive data were dropped
- 9 Columns having missing data percentage more than 35 were dropped, as missing data may affect the quality of dataset.

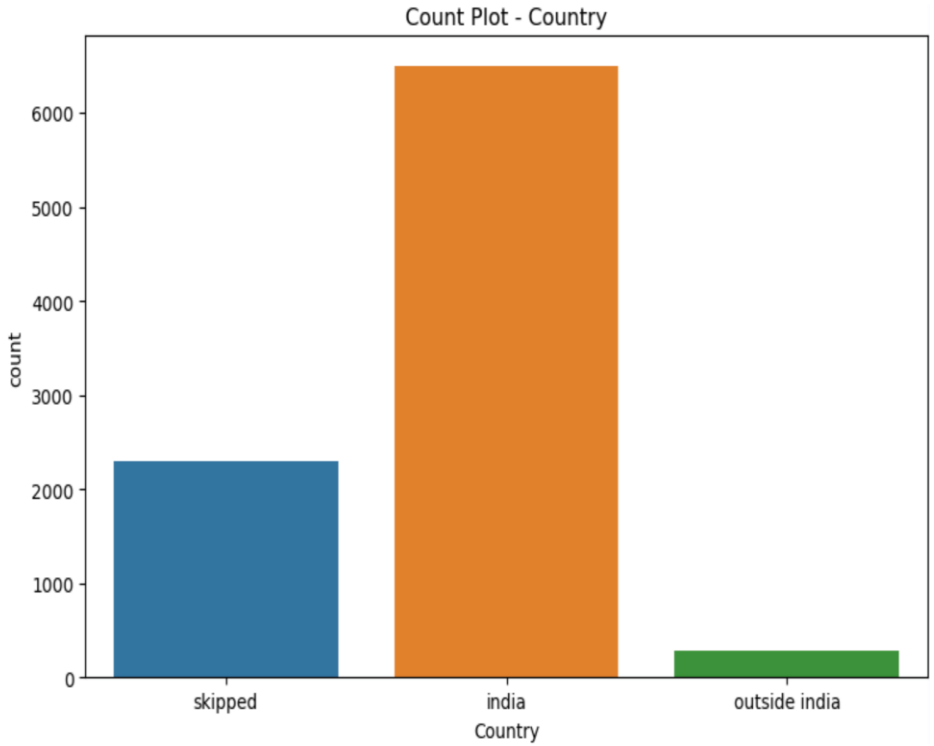
```
['How did you hear about X Education', 'Tags', 'Lead Quality', 'Lead Profile', 'City', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score']
```

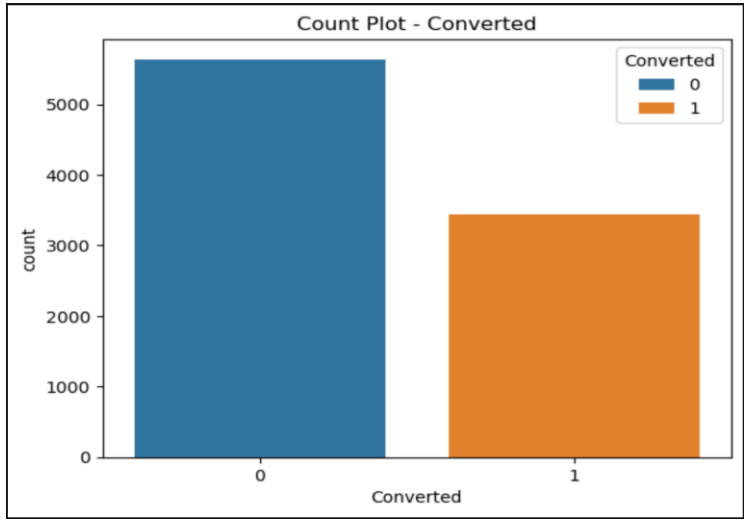
EDA


Count Plot - Last Notable Activity
count
0
500
1000
1500
2000
2500
3000
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

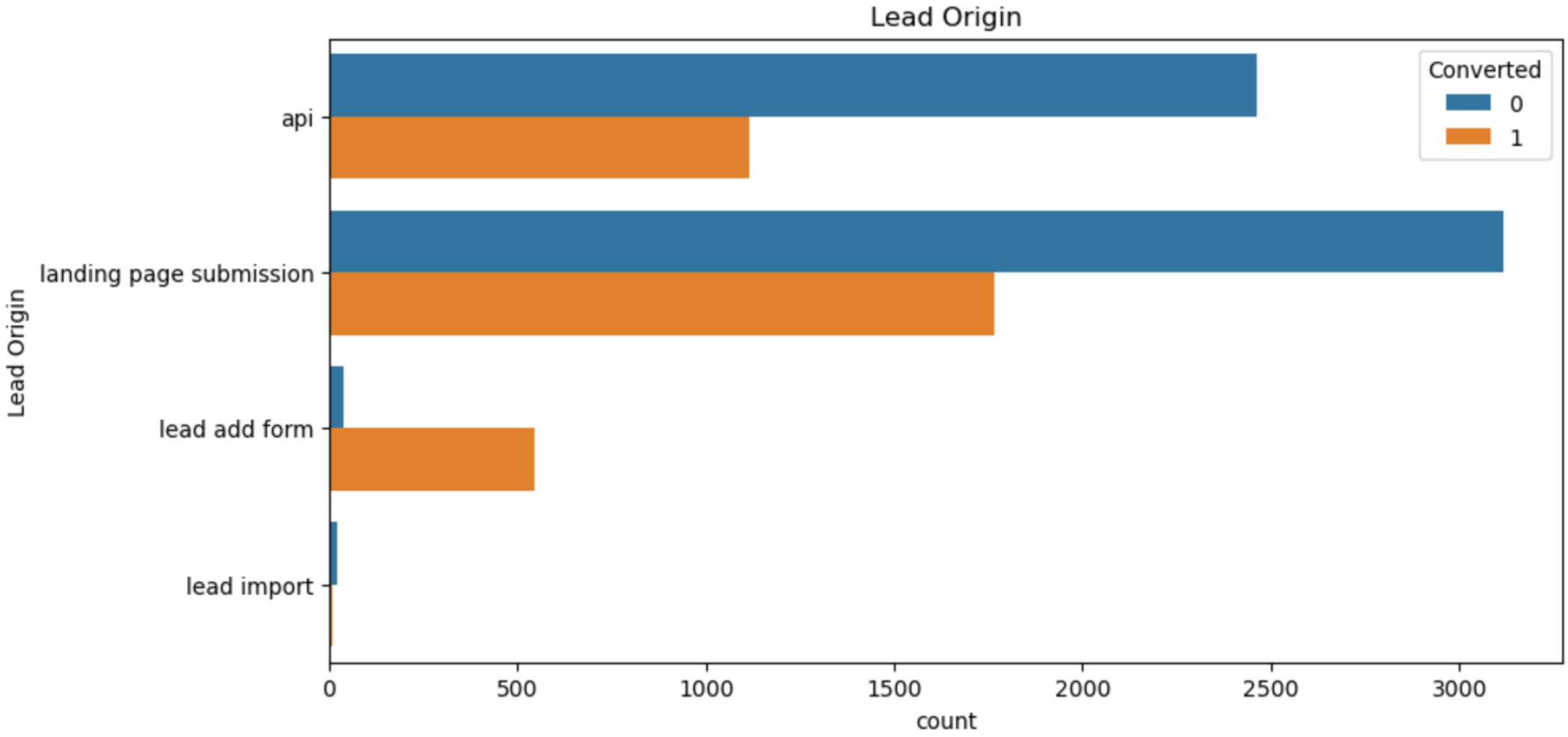

Count Plot - Do Not Email
count
0
1000
2000
3000
4000
5000
6000
7000
8000
no
yes
Do Not Email

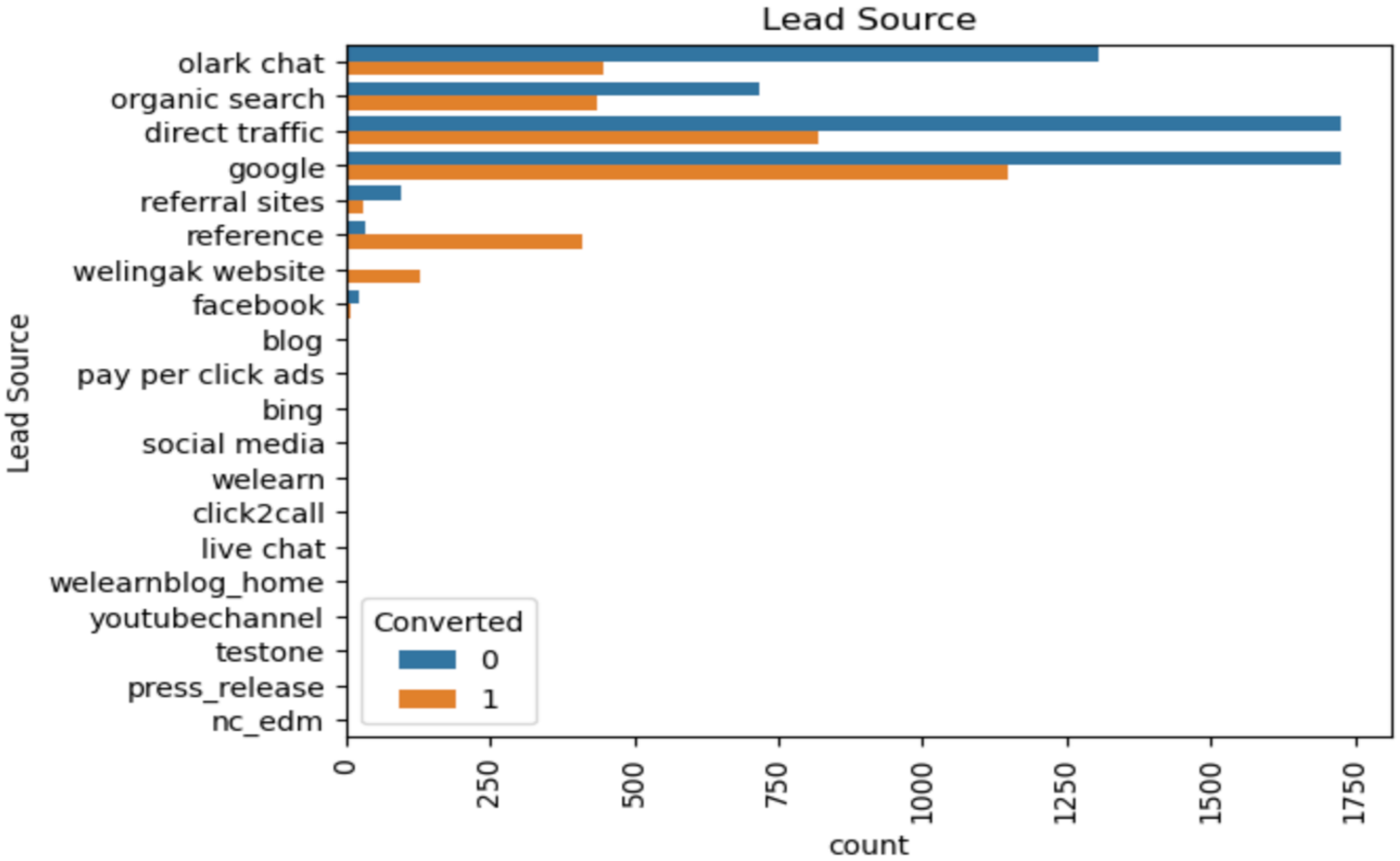

Count Plot - Lead Origin
count
Lead Origin
api
landing page submission
lead add form
lead import
0
1000
2000
3000
4000
5000

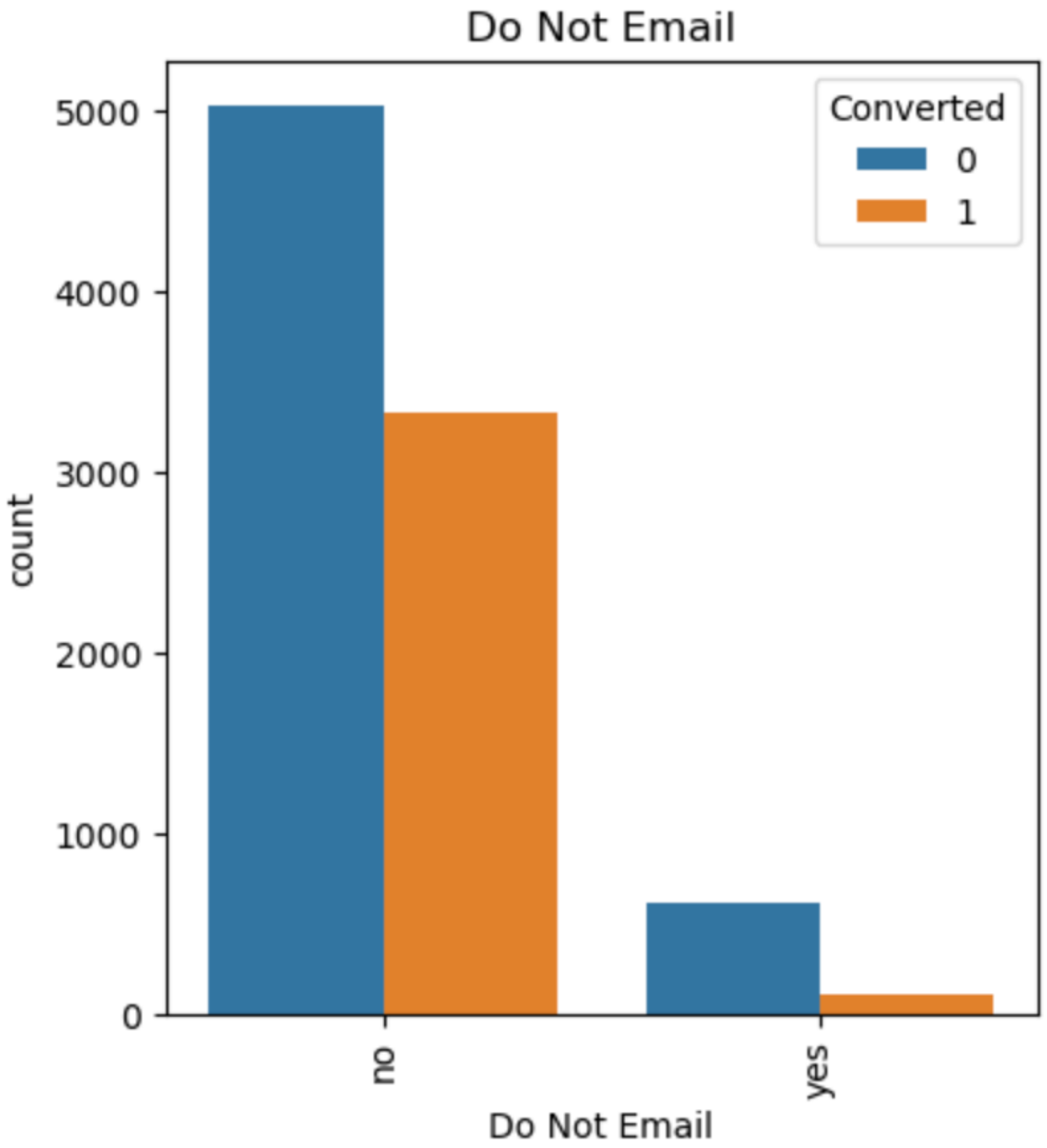

Count Plot - Country
count
Country
skipped
india
outside india
0
1000
2000
3000
4000
5000
6000

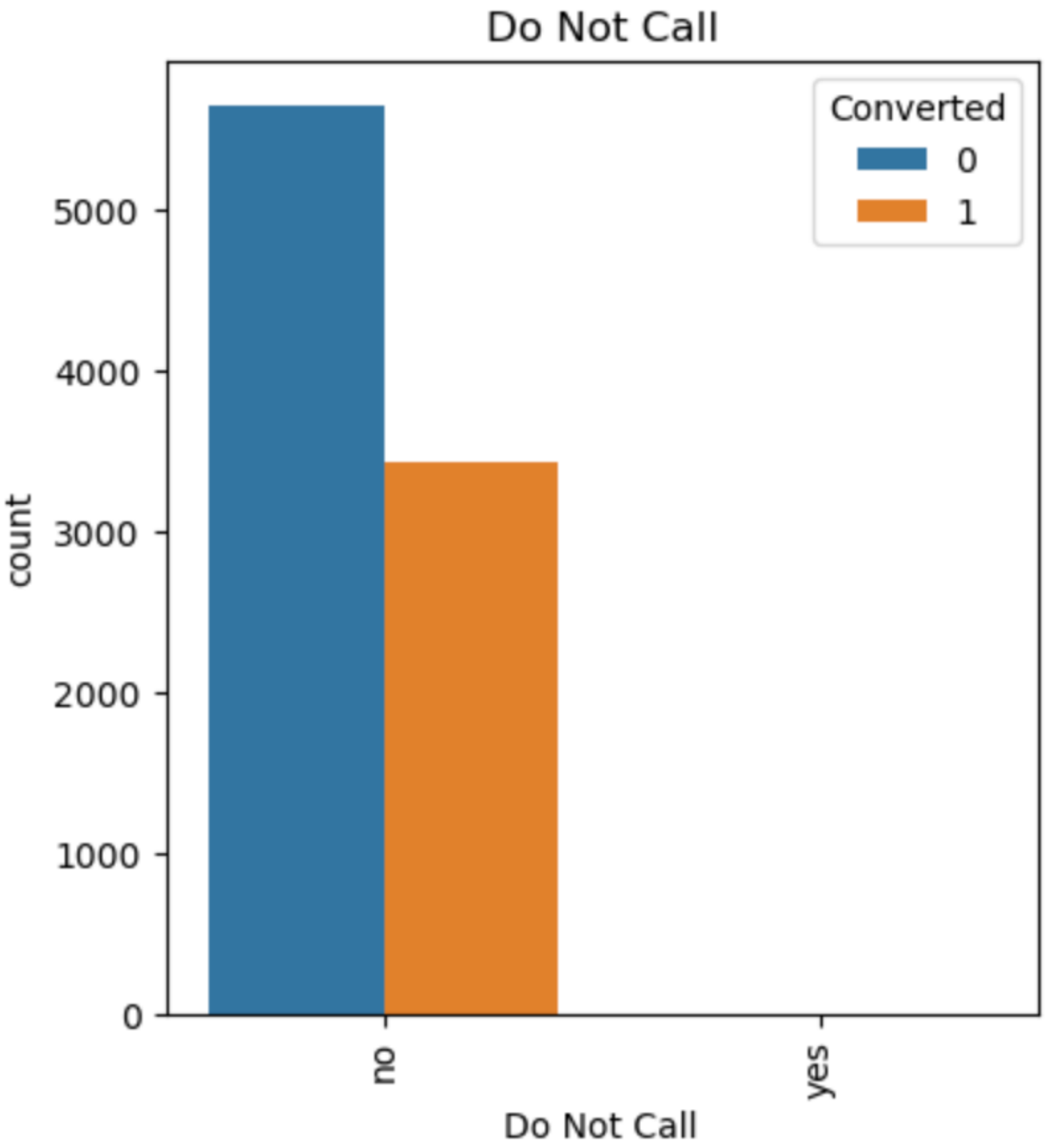

Count Plot - Converted
Converted
0
1
count
Converted
0
1000
2000
3000
4000
5000

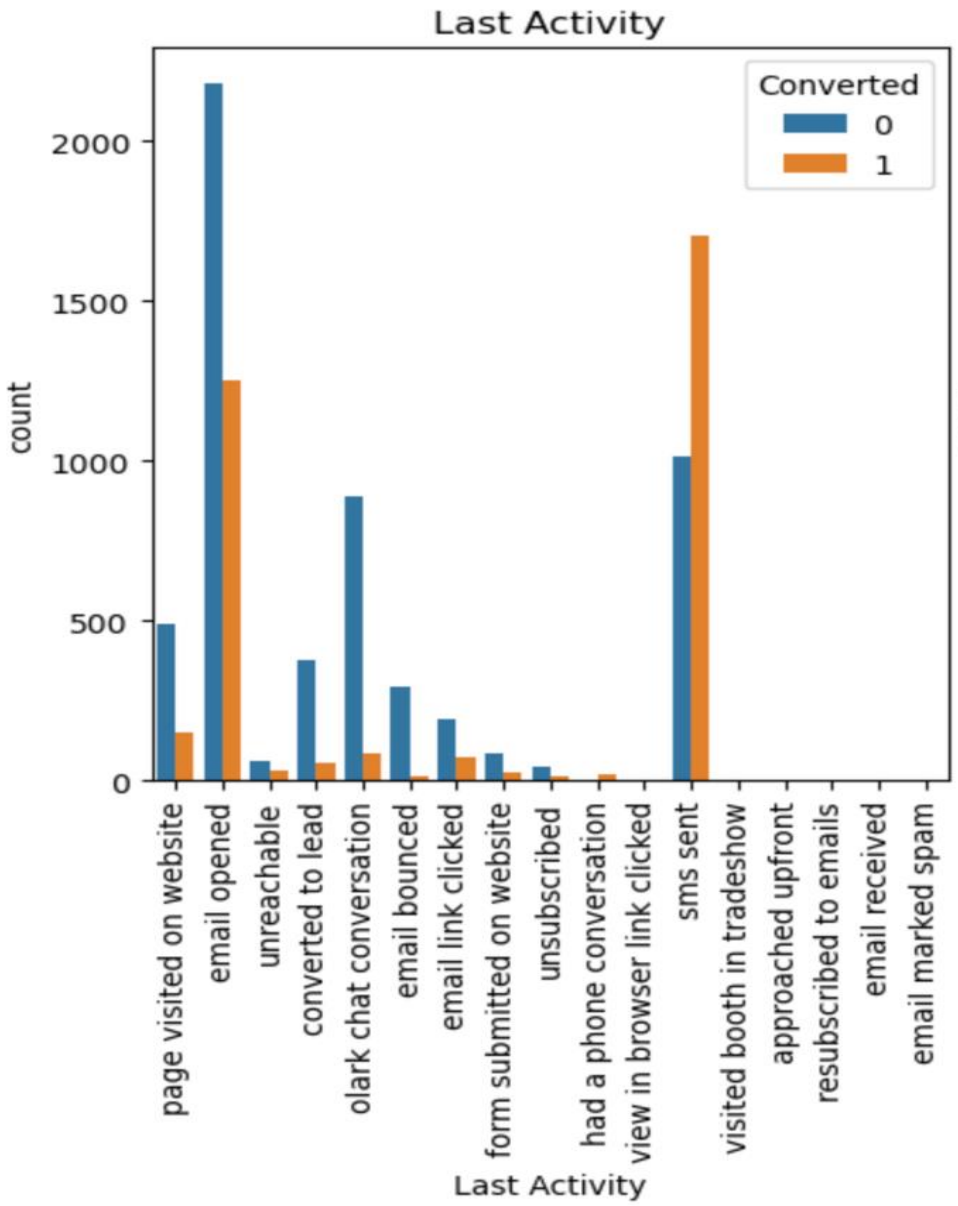
CATEGORICAL VARIABLE ANALYSIS

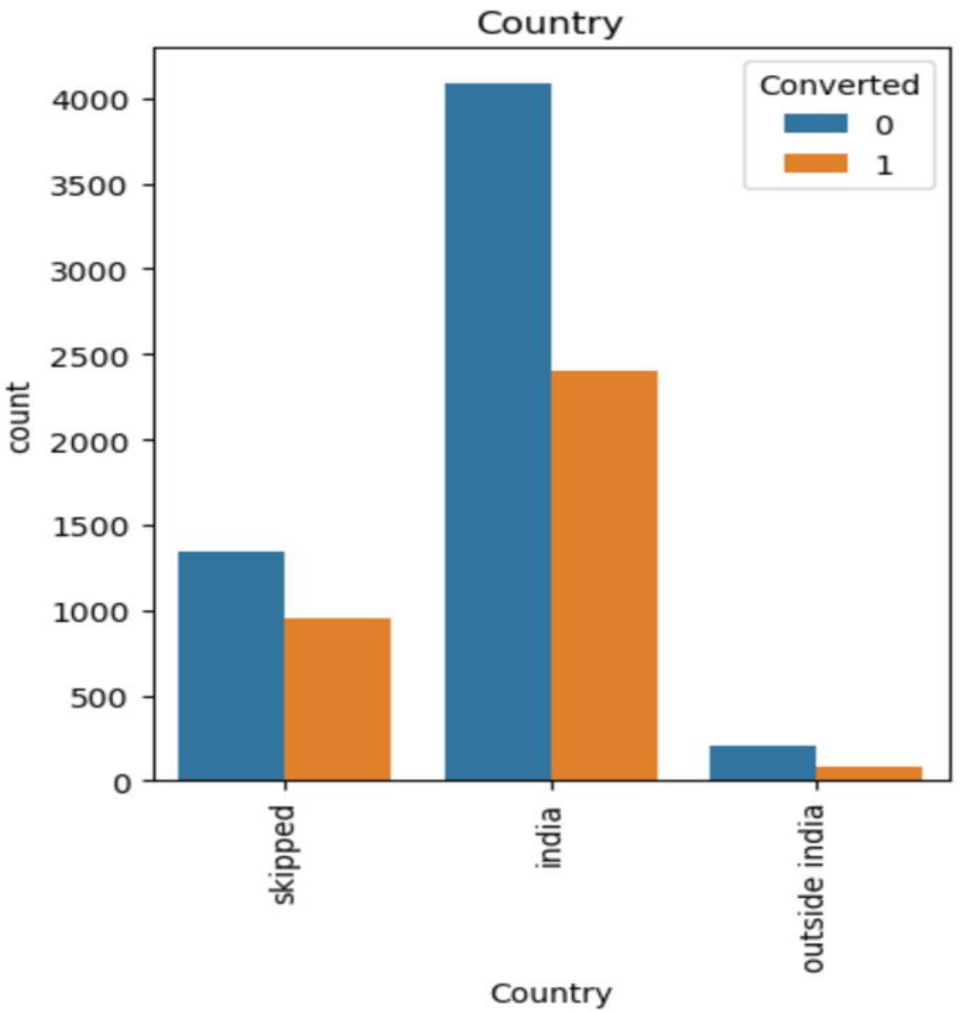



Lead Source
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source
Converted
0
1
0
250
500
750
1000
1250
1500
1750
count


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


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


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
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

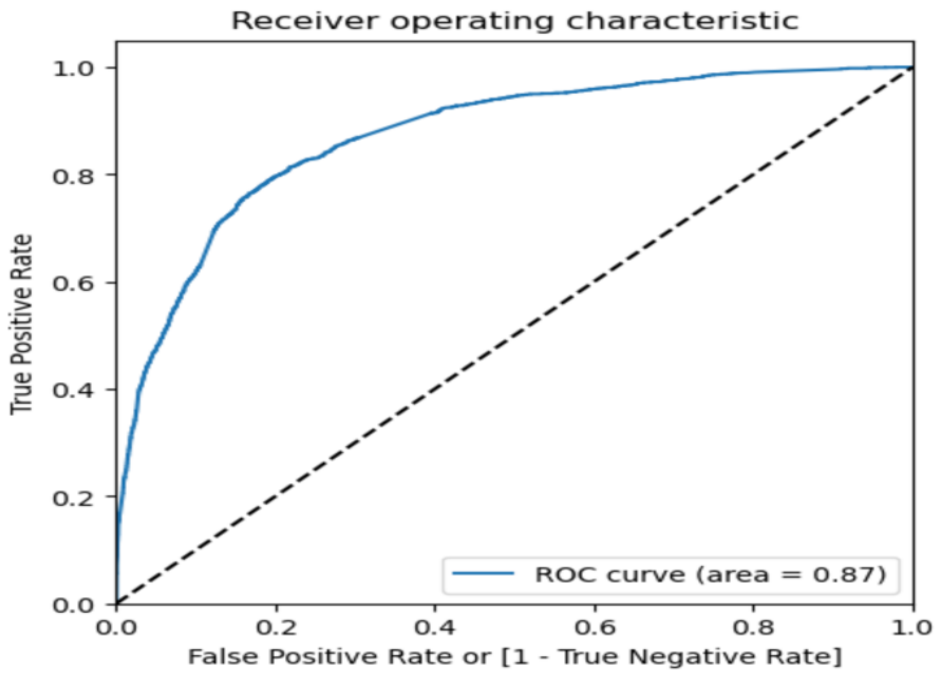

Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
skipped
india
outside india
Country

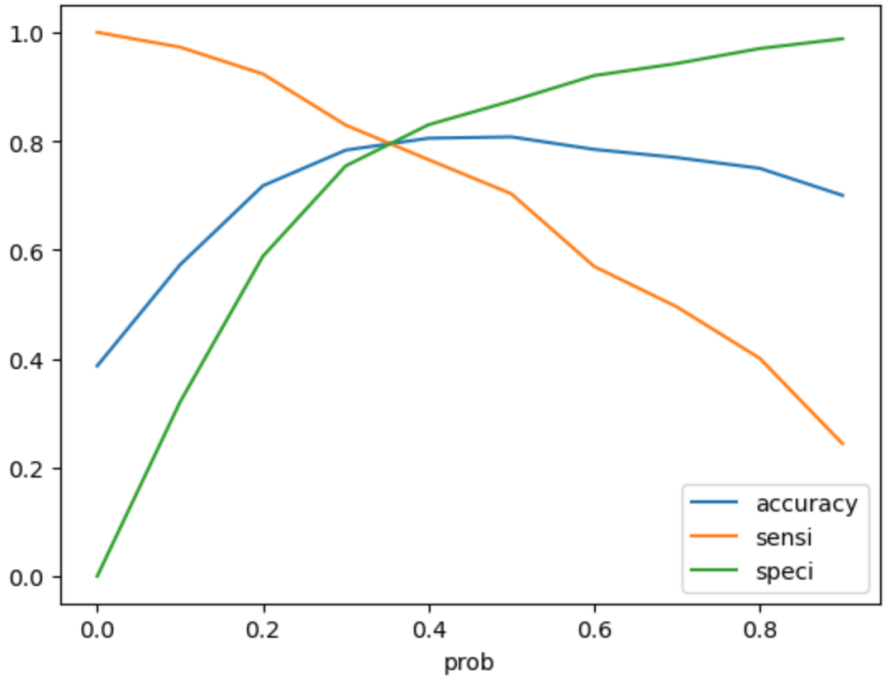
MODEL BUILDING

- Split data into Training & Test set
- Train-Test split with 70:30 ratio
- Use RFE for feature selection
- Choose 15 features out of all available features
- Model selection by dropping features which has p-value greater than 0.05 and VIF value greater than 5
- Test data prediction
- Model accuracy is close to 81%

Model stats ::
Current cut off = 0.5
Accuracy = 81%
Sensitivity = 70%
Specificity = 87%

ROC CURVE





- Optimal Cut off point – Probability for which we get balanced sensitivity and specificity. Similarly we can see in the 2nd graph that cut off is at 0.35

RECOMMENDATIONS

After carefully analyzing several models ,with the help of logistic regression, we have found that the variables, which observed most in the potential clients are as follows (in descending order) :

- The total time spend on the Website
- Total number of visits
- When the Lead source was :
  - Google
  - Direct Traffic
  - Organic Search
  - Welingak website
- Last activity was on :
  - SMS
  - Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional

Keeping above pointers in mind, X Education can expand as they have a high chance to convert almost all the potential buyers to buy their courses.