Lead Scoring Case Study

Problem Statement

An X Education need assistance to choose the most potential leads, for example the leads that are probably going to change over into paying clients. The organization expects us to build a model wherein you need to assign a lead score to every one of the leads to such an extent that the clients with higher lead score have a higher conversion possibility and the clients with lower lead score have a lower conversion possibility. The CEO, specifically, has given a ballpark of the target lead conversion rate to be around 80%.

Goals and Objectives

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. Deployment of model for further use.

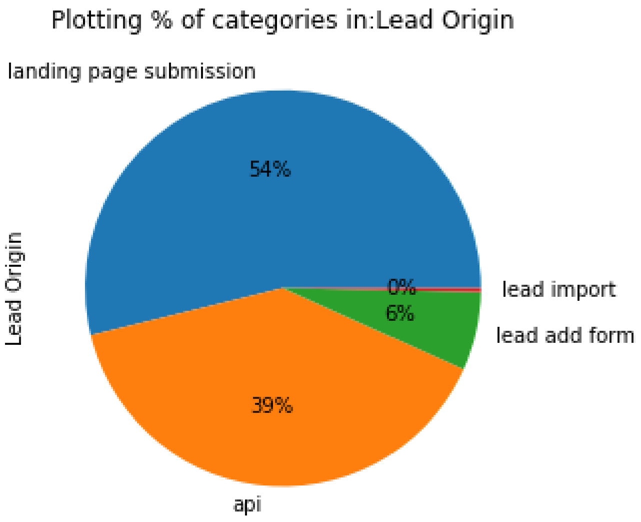
Steps to be followed

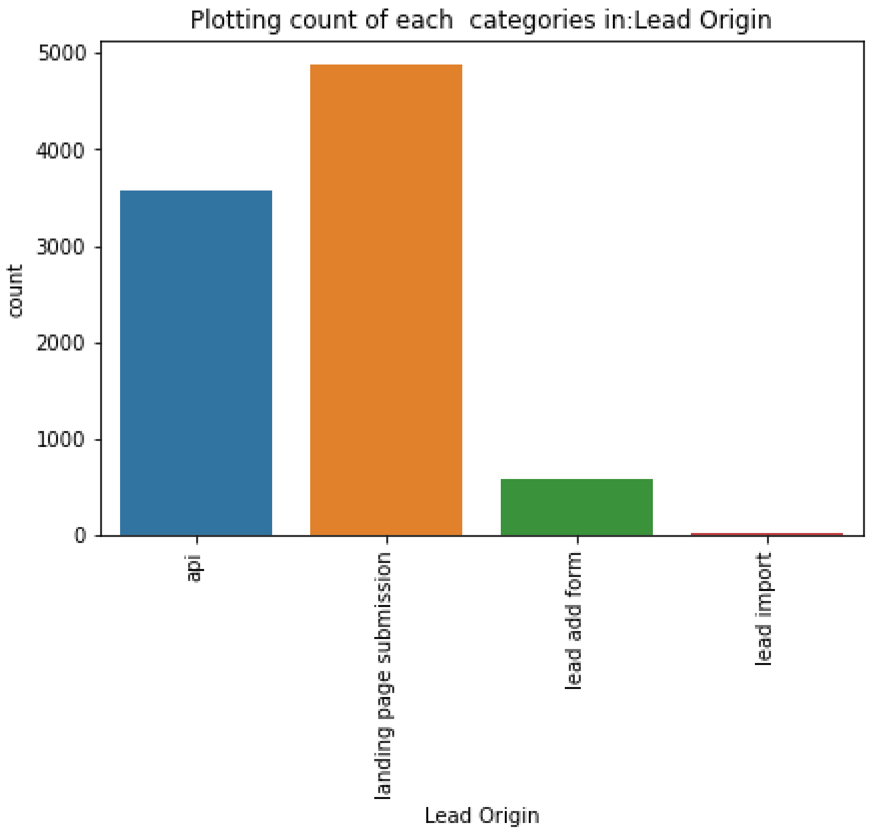
1. Read and Understand the data.
2. Data Cleaning and Data Manipulation
   - Check and handle the duplicate data, missing values and outliers in the data.
   - Drop the columns if variable has high percentage of missing values.
3. Exploratory Data Analysis
   - Univariate Analysis and Bivariate Analysis
4. Create Dummy variables
5. Perform Feature Scaling once the data is ready for analysis
6. Logistic Regression Classification technique for the model building and prediction.
7. Perform Validation
8. Model representation
9. Conclusions and recommendations

Data Cleaning and Data Manipulation

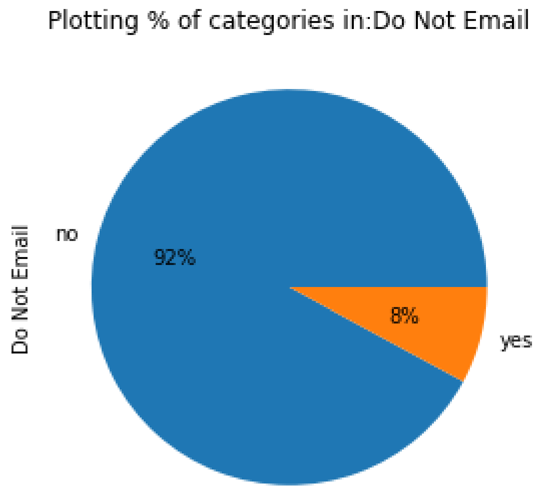
1. Check the shape of data.
2. Check for outliers and missing values. Impute the missing values accordingly.
3. Drop the columns which is having more than 35% missing values.
4. Dropping the columns 'Magazine’, 'Receive More Updates About Our Courses’, ’I agree to pay the amount through cheque’, 'Get updates on DM Content’, 'Update me on Supply Chain Content’ as they all contain unique value and do not require for analysis.
5. Drop the columns prospect ID and Lead Number as they are unique IDs which are not necessary for analysis.
6. Perform the Univariate Analysis.
7. Perform Bivariate Analysis.
8. Create Dummy variables.

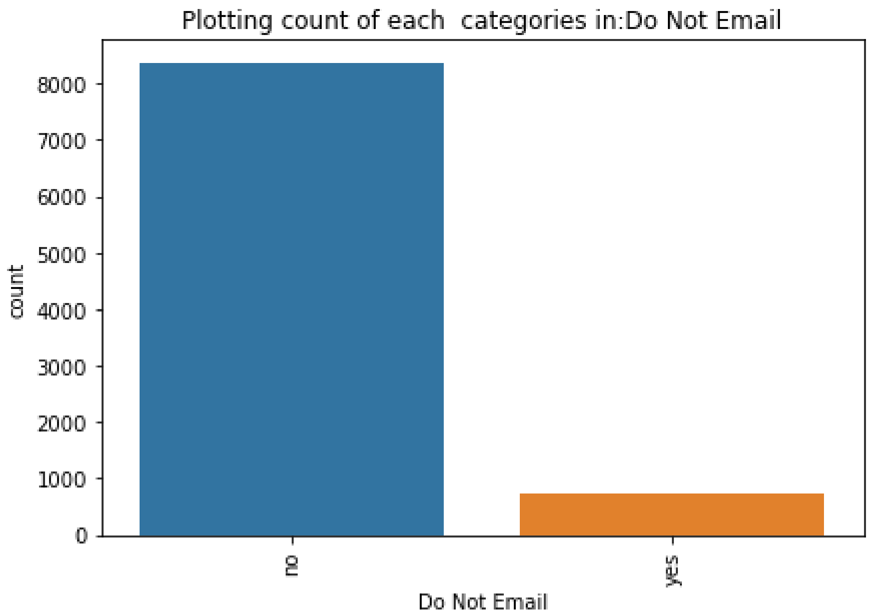
Univariate Analysis

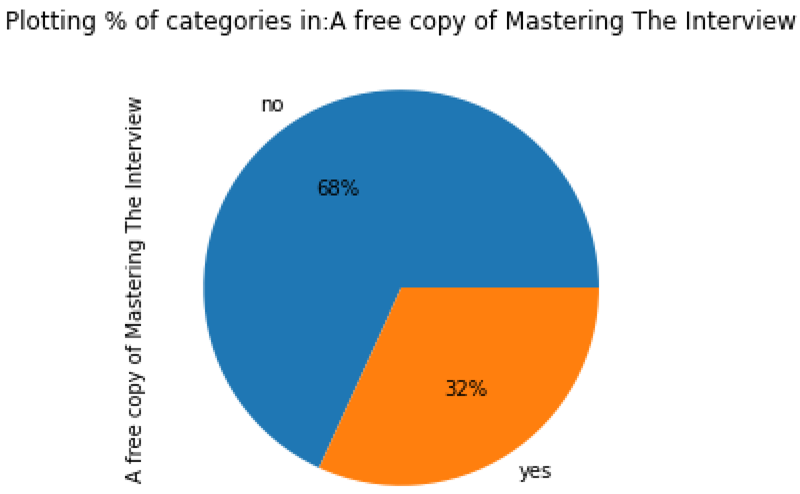


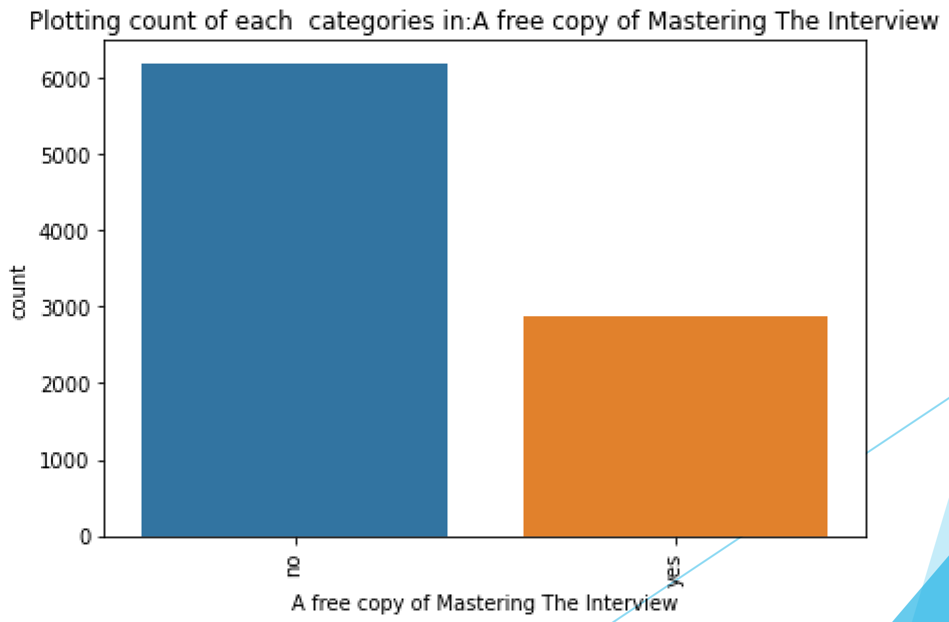


54% of customers were identified as leads from landing page submission.

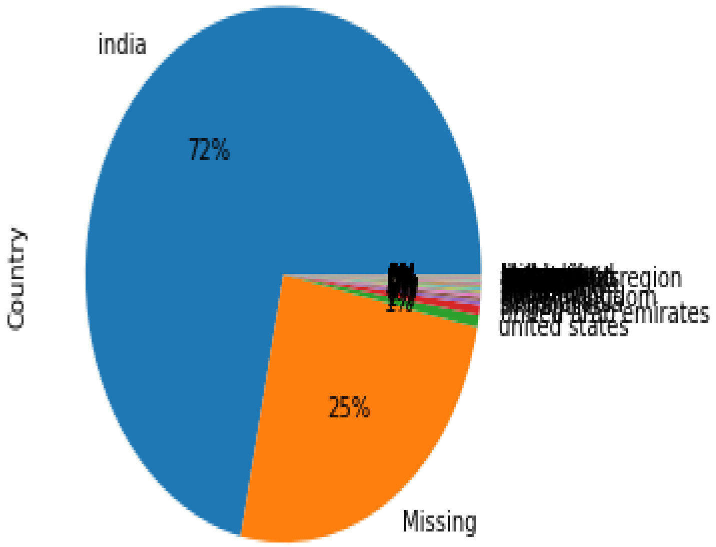

Plotting % of categories in:Do Not Email


Plotting count of each categories in:Do Not Email

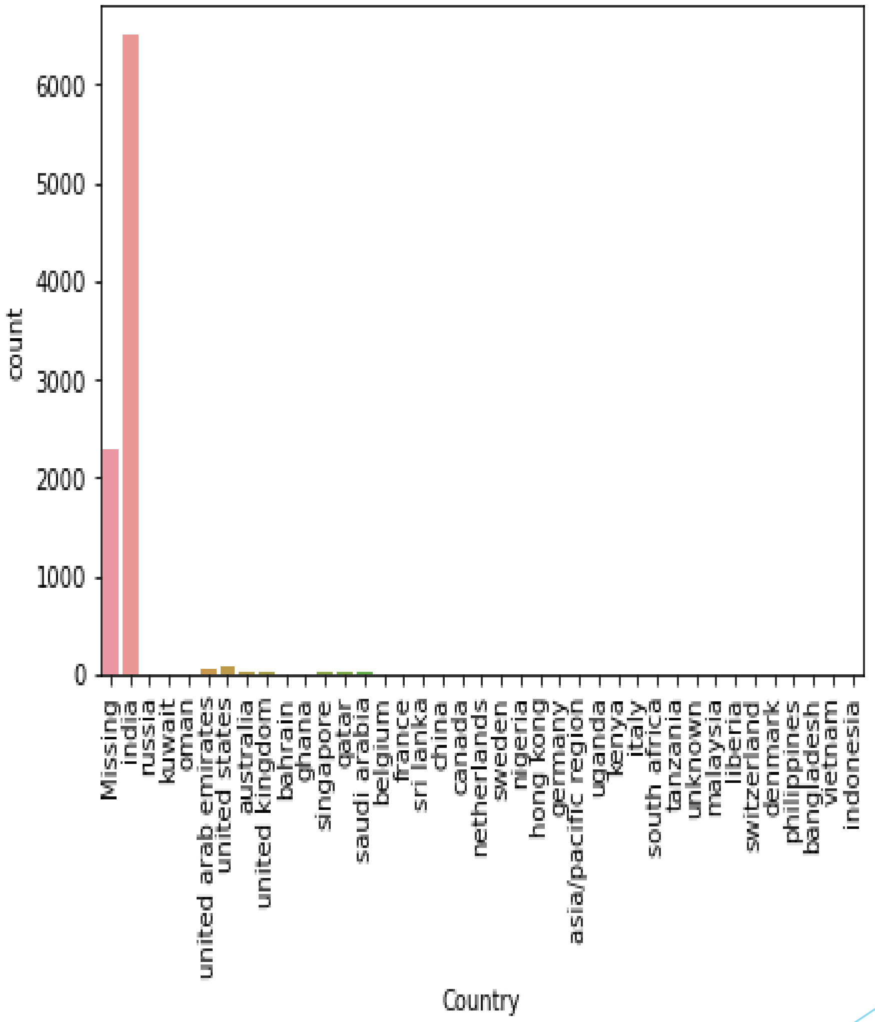

Plotting % of categories in:A free copy of Mastering The Interview

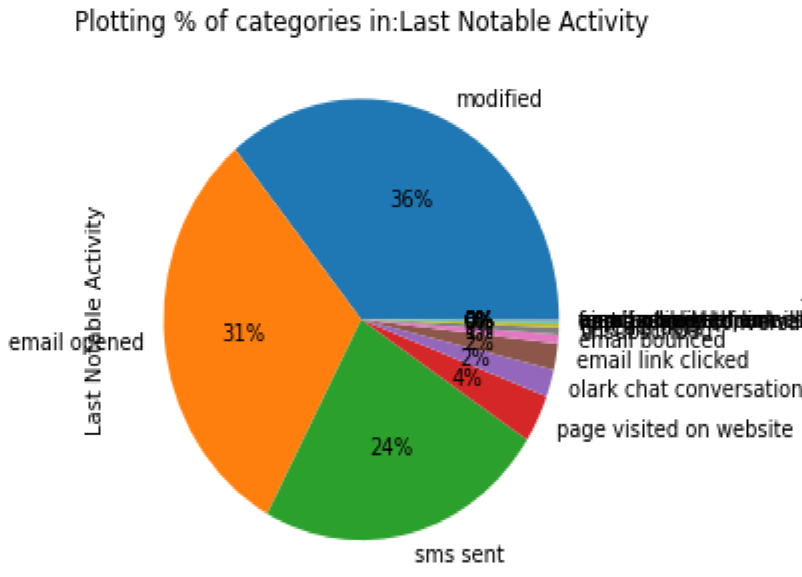

Plotting count of each categories in:A free copy of Mastering The Interview

Plotting % of categories in:Country

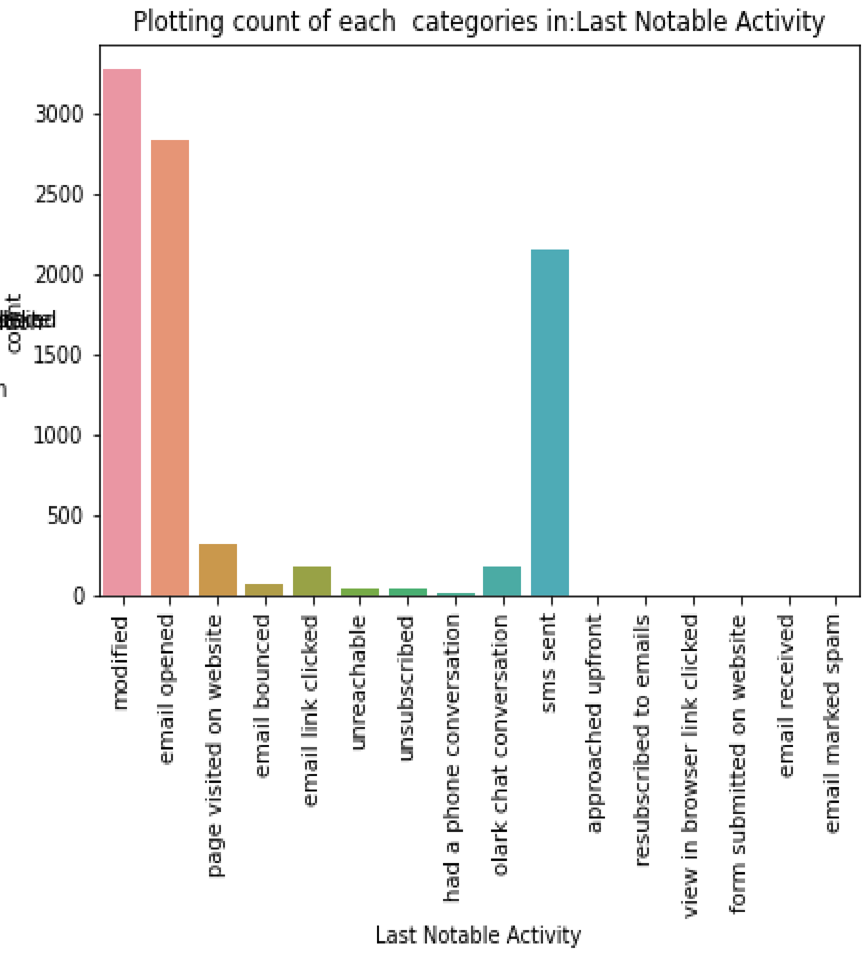


Plotting count of each categories in:Country

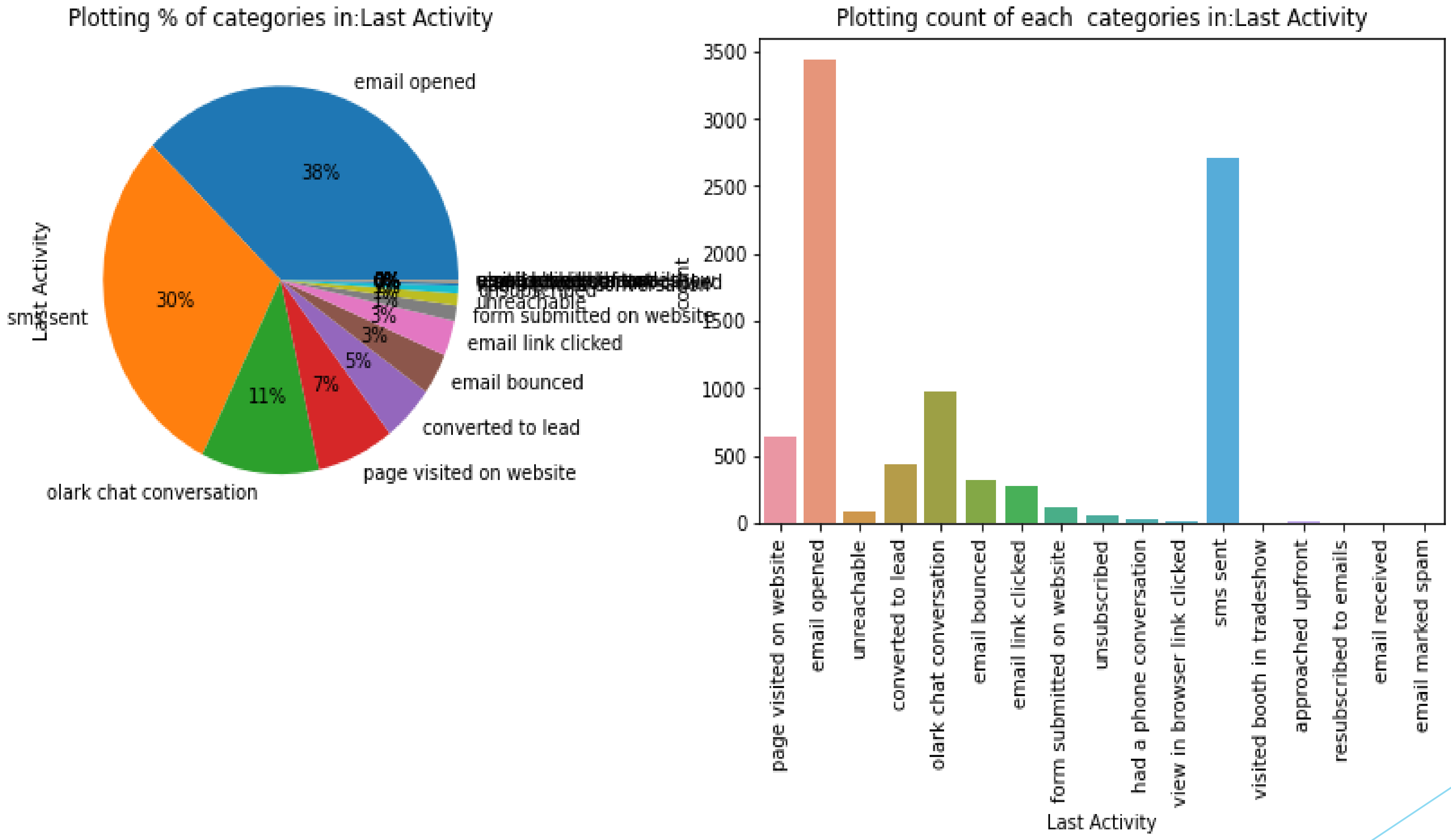




Plotting % of categories in:Last Notable Activity

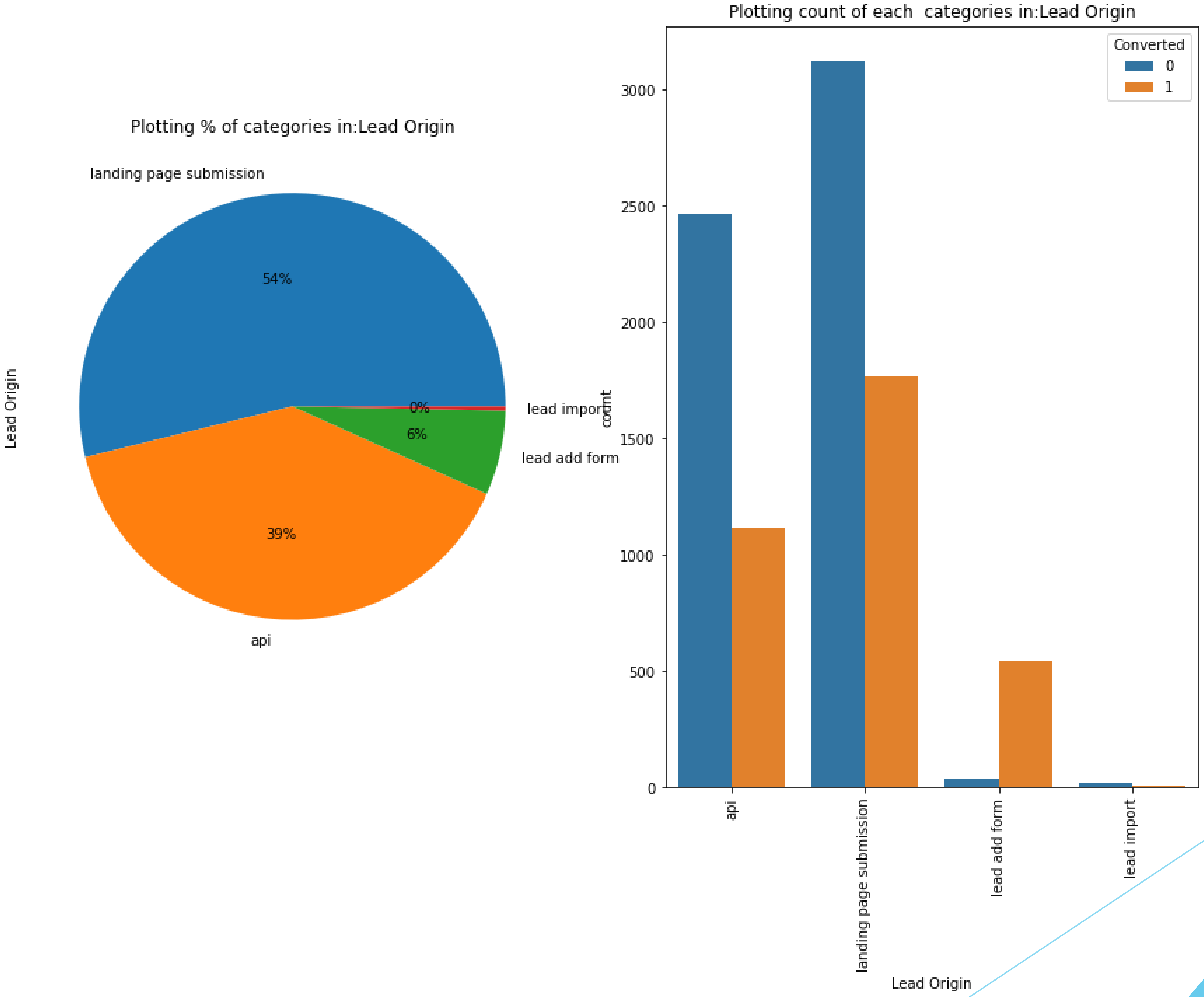


Plotting count of each categories in:Last Notable Activity


Plotting % of categories in:Last Activity
Last Activity
email opened
38%
sms sent
30%
olark chat conversation
11%
page visited on website
7%
converted to lead
5%
email bounced
3%
email link clicked
3%
form submitted on website
1%
unreachable
1%
0%
Plotting count of each categories in:Last Activity
count
3500
3000
2500
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

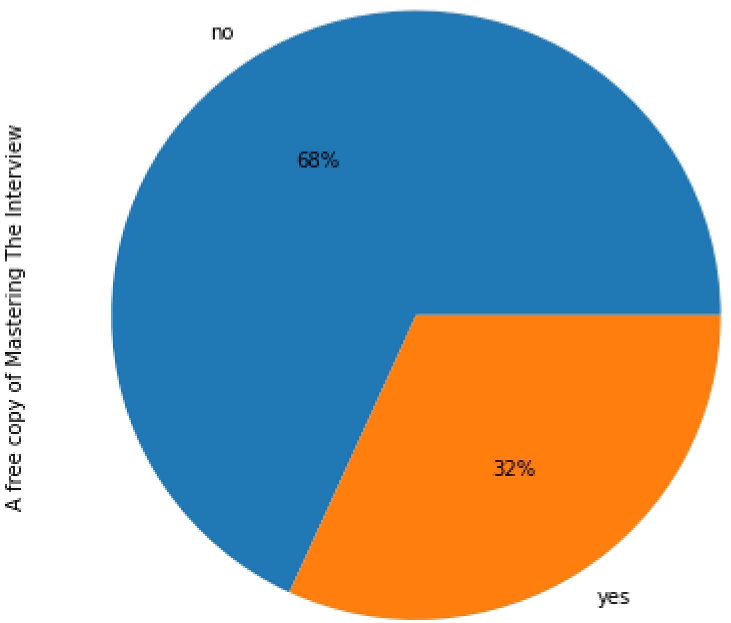
Univariate Analysis Observations

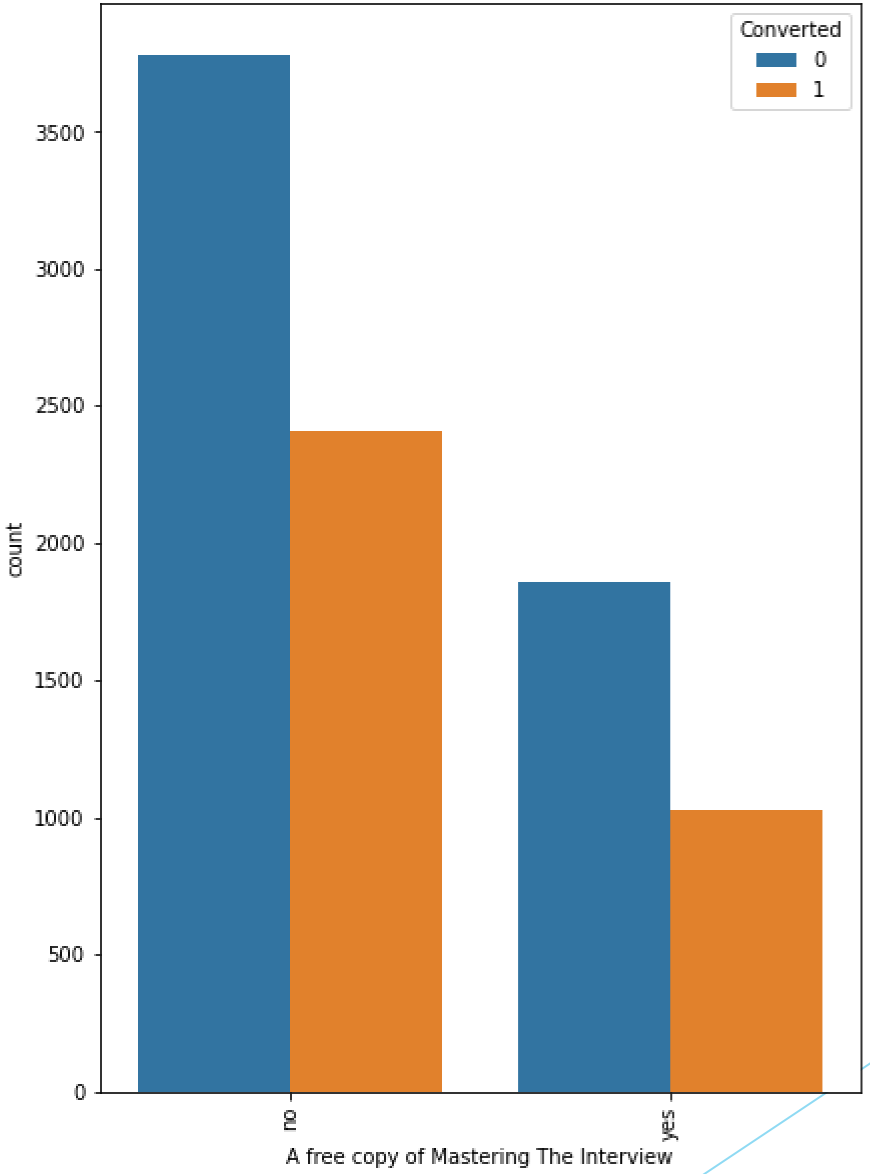
1. 54% of customers were identified as leads from landing page submission.
2. Majority of customers choose to get an email, call.
3. 72% Demography of customers belonged to India.
4. 32% of customers preferred to get a free copy of mastering the interview.
5. Majority of customers were having modified followed by email opened , SMS sent as their last notable activity.
6. Google, direct traffic, Olark chat were the top 3 lead sources.
7. Most of the customers who seemed interested are unemployed followed by working professionals.
8. 70% of customers are looking for better career prospects.
9. Email opened and SMS sent were the activities that are most frequently performed last activity by the customer.
10. On an average customers spent on Total visits ,Total Time spent on website, Page views per visit were 3, 248 &2 respectively.
11. Potential Conversion rate is 38%.

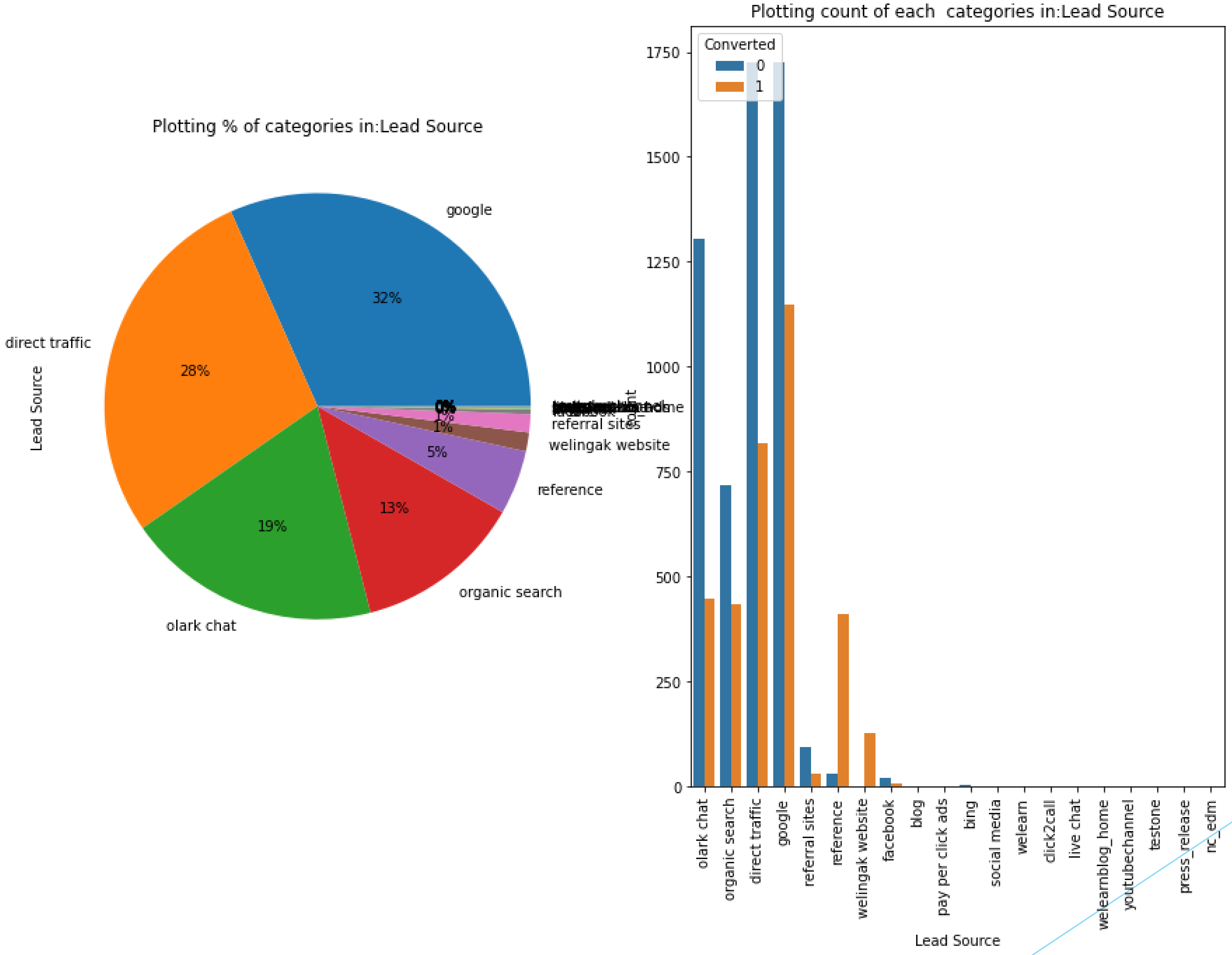


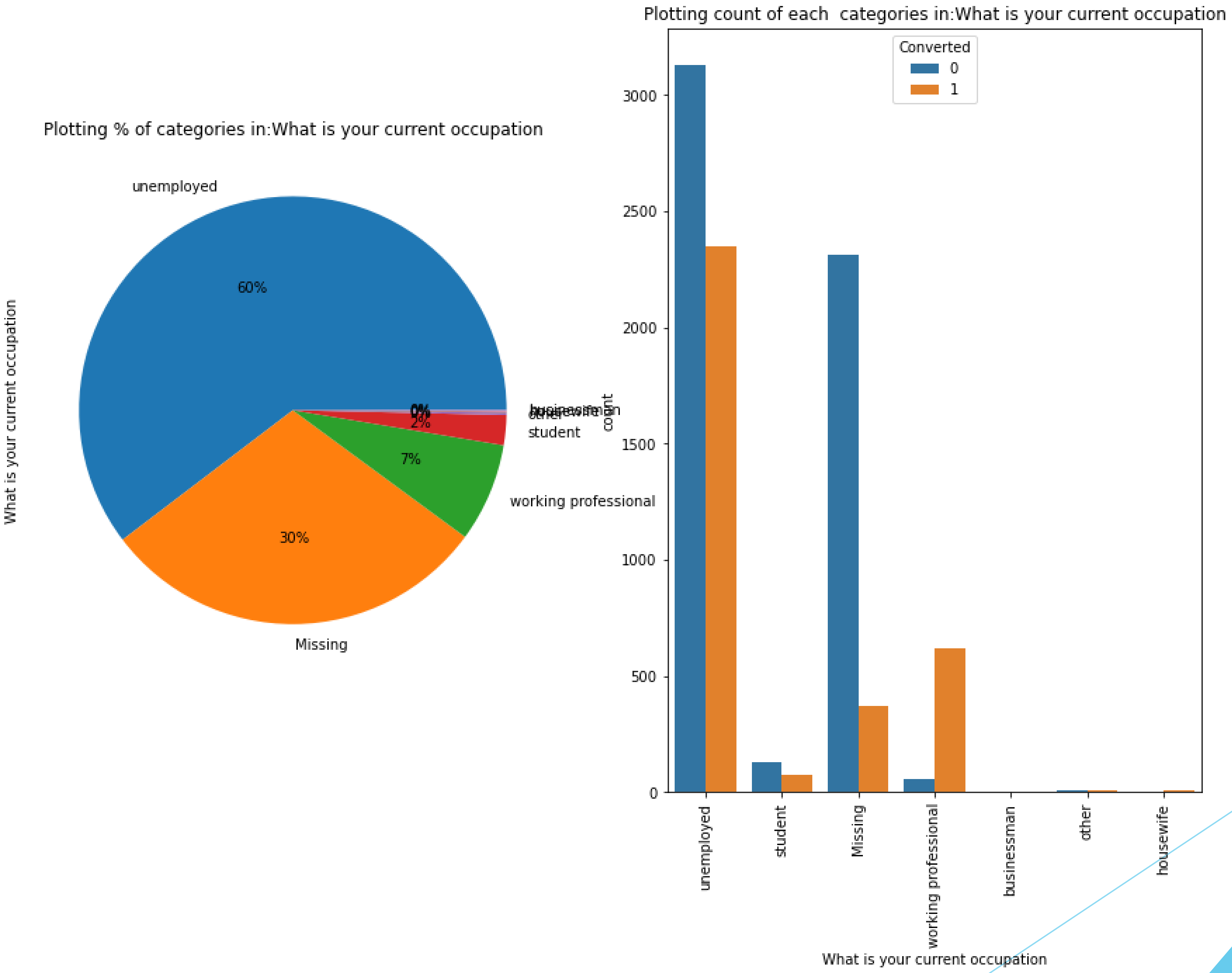
Plotting % of categories in:Lead Origin

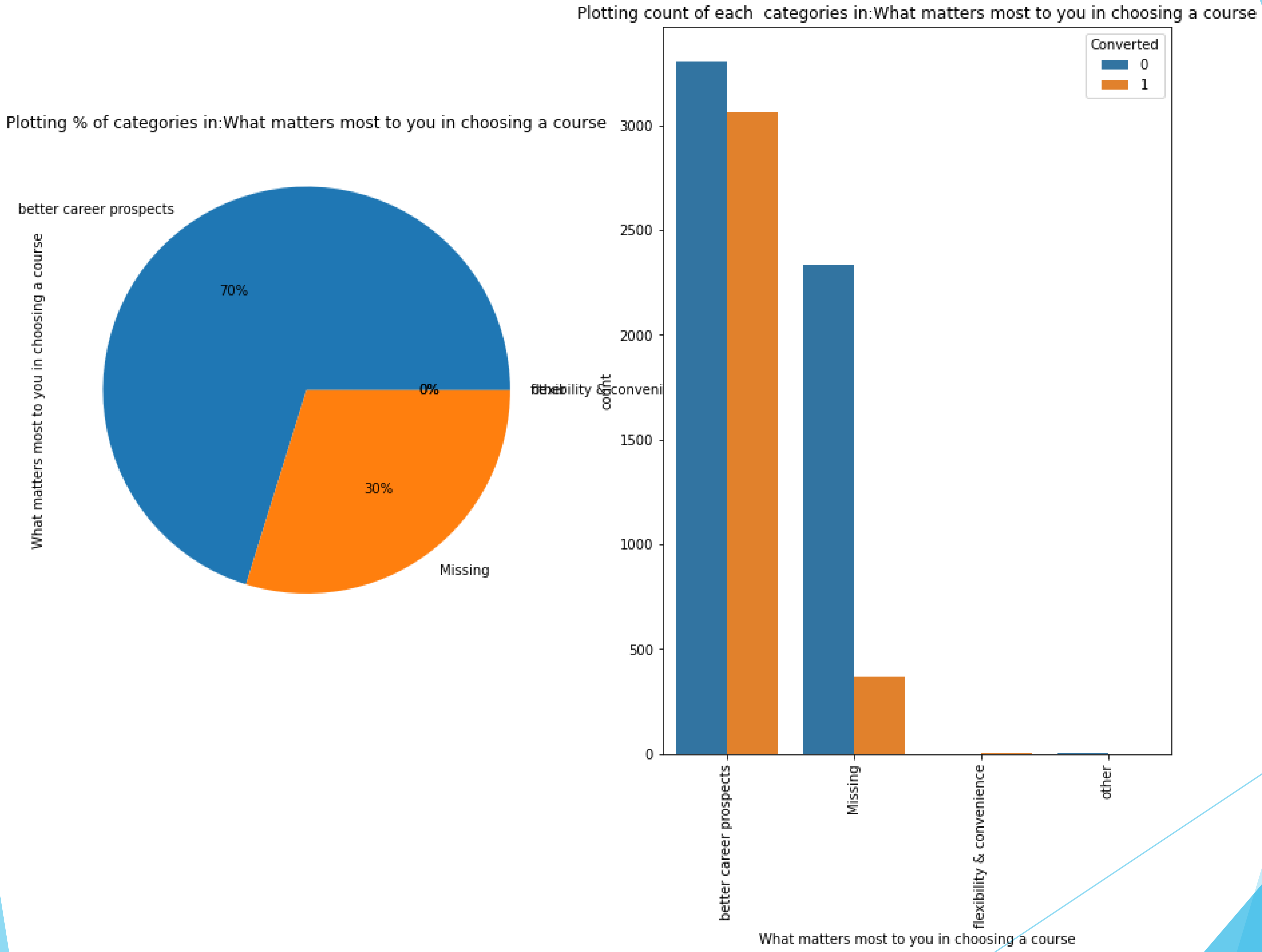
Plotting count of each categories in:Lead Origin


Plotting % of categories in:A free copy of Mastering The Interview


Plotting count of each categories in:A free copy of Mastering The Interview

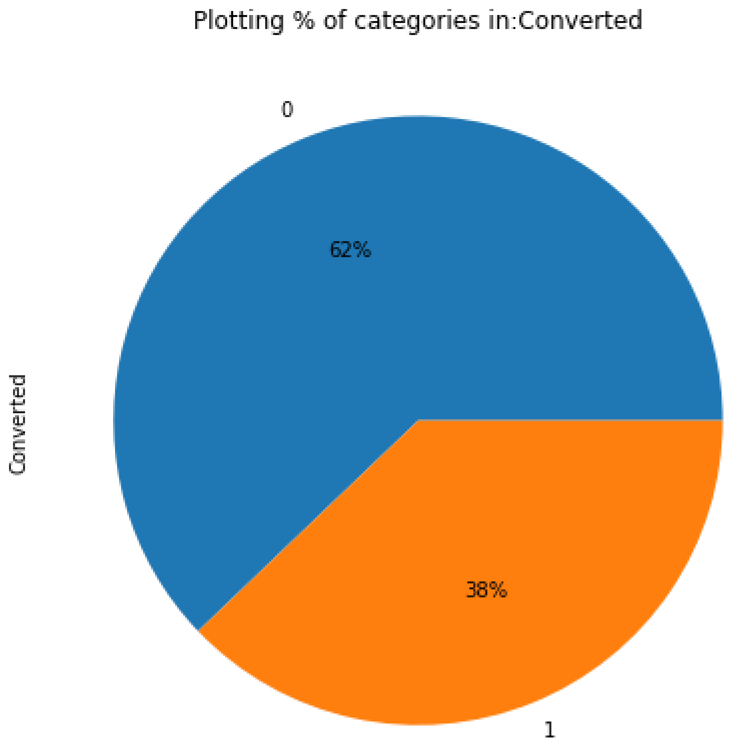

Plotting % of categories in:Lead Source
Lead Source
google
32%
direct traffic
28%
olark chat
19%
organic search
13%
reference
5%
welingak website
1%
referral sites
1%
0%
Plotting count of each categories in:Lead Source
Converted
0
1
Count
1750
1500
1250
1000
750
500
250
0
olark chat
organic search
direct 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


Plotting % of categories in:What is your current occupation
What is your current occupation
unemployed
60%
Missing
30%
working professional
7%
student
2%
0%
0%
0%
businessman
housewife
other
Plotting count of each categories in:What is your current occupation
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
unemployed
student
Missing
working professional
businessman
other
housewife
What is your current occupation

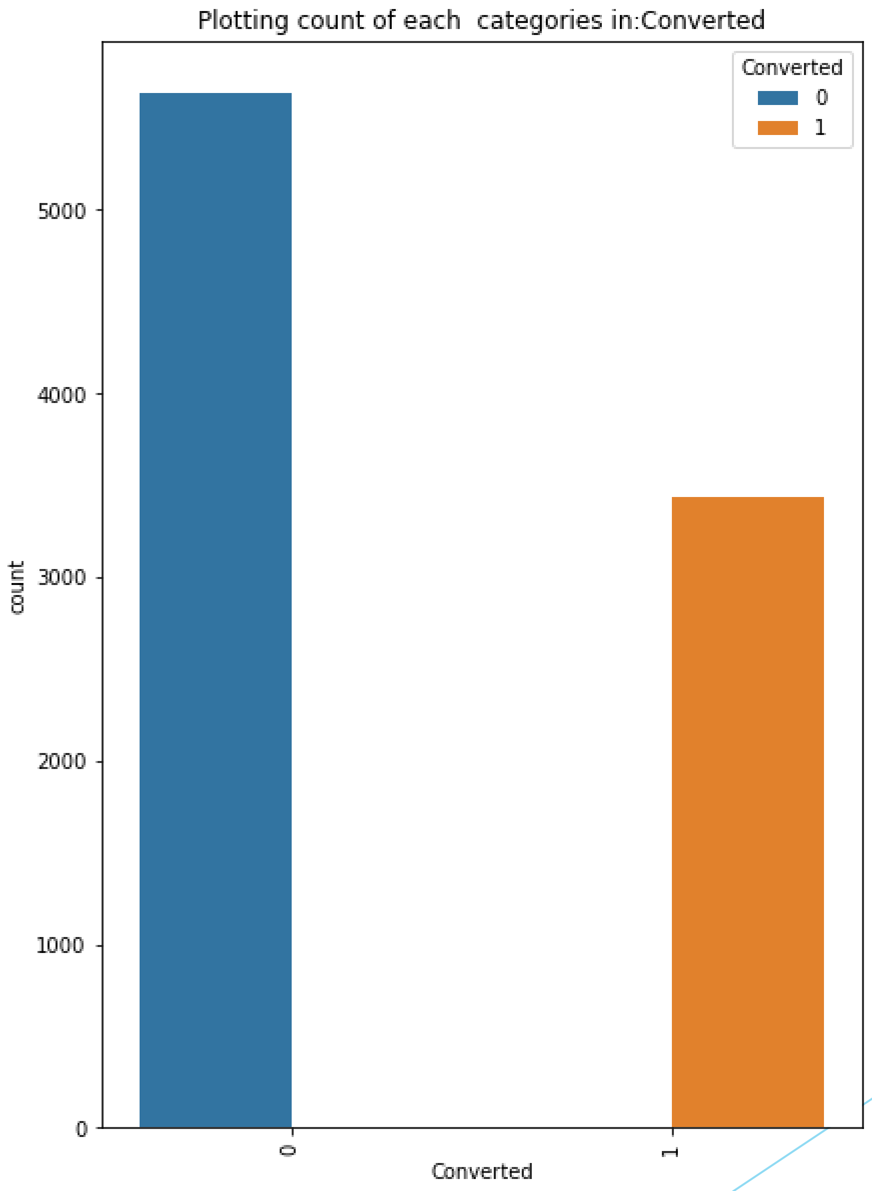


Plotting % of categories in:What matters most to you in choosing a course

Plotting count of each categories in:What matters most to you in choosing a course



Plotting % of categories in:Converted



Plotting count of each categories in:Converted

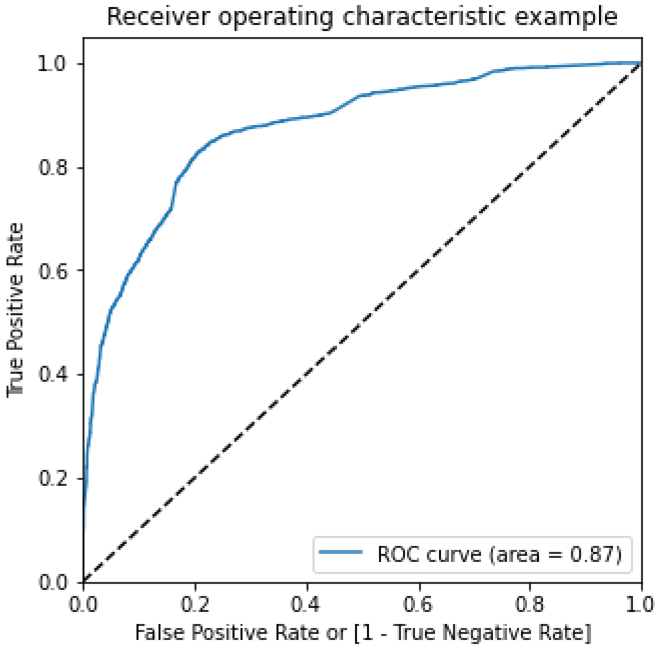
Bivariate Analysis Observations

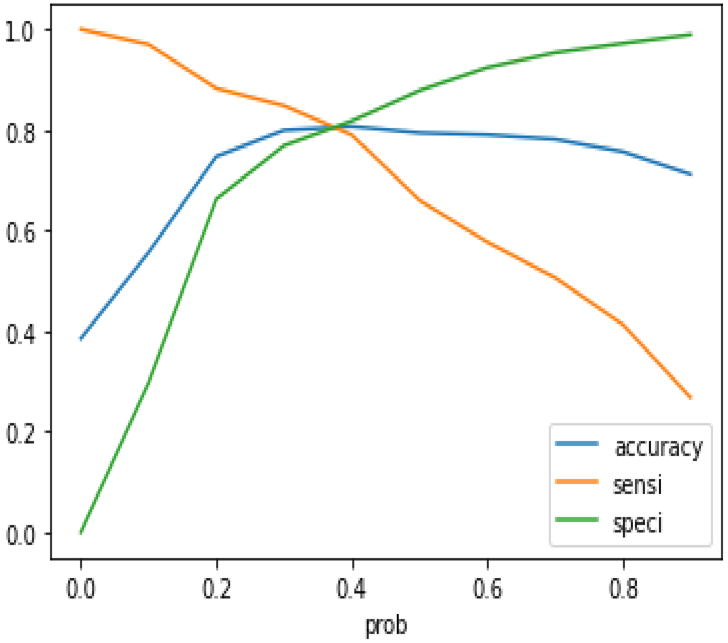
1. Conversion rate is 38%
2. Among converted customers 54% were on landing page submission followed by api lead origins.
3. More than 90% of customers who opted to get emails, calls got converted.
4. 68% of the customers who didn't opt for A free copy of Mastering The Interview got converted.
5. Among converted customers 60% were unemployed.
6. Most of the converted customers are looking for better career prospects.

Model Building

1. Split the dataset in to training set and testing set
2. The first basic step for regression is to perform train_test_split. In this case study, we have chosen the dataset to be in the ratio 70:30.
3. For Feature Selection, we can use either Standard Scalar or MinMax Scalar.
4. Use RFE for Feature Selection.
5. Running RFE with 15 variables as output .
6. Building model by removing the variables whose p-value is greater than 0.05 and VIF is greater than 5.
7. Predictions on the test data set.
8. Overall accuracy is 81%.

ROC Curve





1. Finding optimal cut off point
2. Probability where we get balanced Sensitivity and Specificity
3. From the graph, it is visible that the optimal cut off value is 0.37

CONCLUSION

It was observed that the factors that made the biggest difference in the potential purchasers are -

1. Total time spent on the website
2. When the lead origin is lead add form
3. When the lead source was -
   - Direct traffic
   - Welingak website
4. When the customer opted for not to email
5. When the customer current occupation is working professional.
6. When the customer says that he is choosing this course for better career prospects.
7. When the last activity was -
   - Olark Chat conversation
   - Phone conversation
8. When the last notable activity was modified, email opened, page visited on website, email link clicked and olark chat conversation.

Keeping this in mind, the X Education can thrive as they have an exceptionally high opportunity to get almost every one of the likely purchasers to adjust their perspective and purchase their courses.