Lead Score Case Study

Presented by

- Variyata Verma
- Shivam Pawar
- Vivek Singh

Problem Statement:

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

- Now, 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 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 more promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Strategy:

- Starting with Data understanding, data cleaning and data manipulation.

a. Check and handle duplicate data.
b. Check and handle NA values and missing values.
c. Drop columns, if it contains large number of missing values and not useful for the analysis.
d. Imputation of the values, if necessary.
e. Check and handle outliers in data.

- EDA

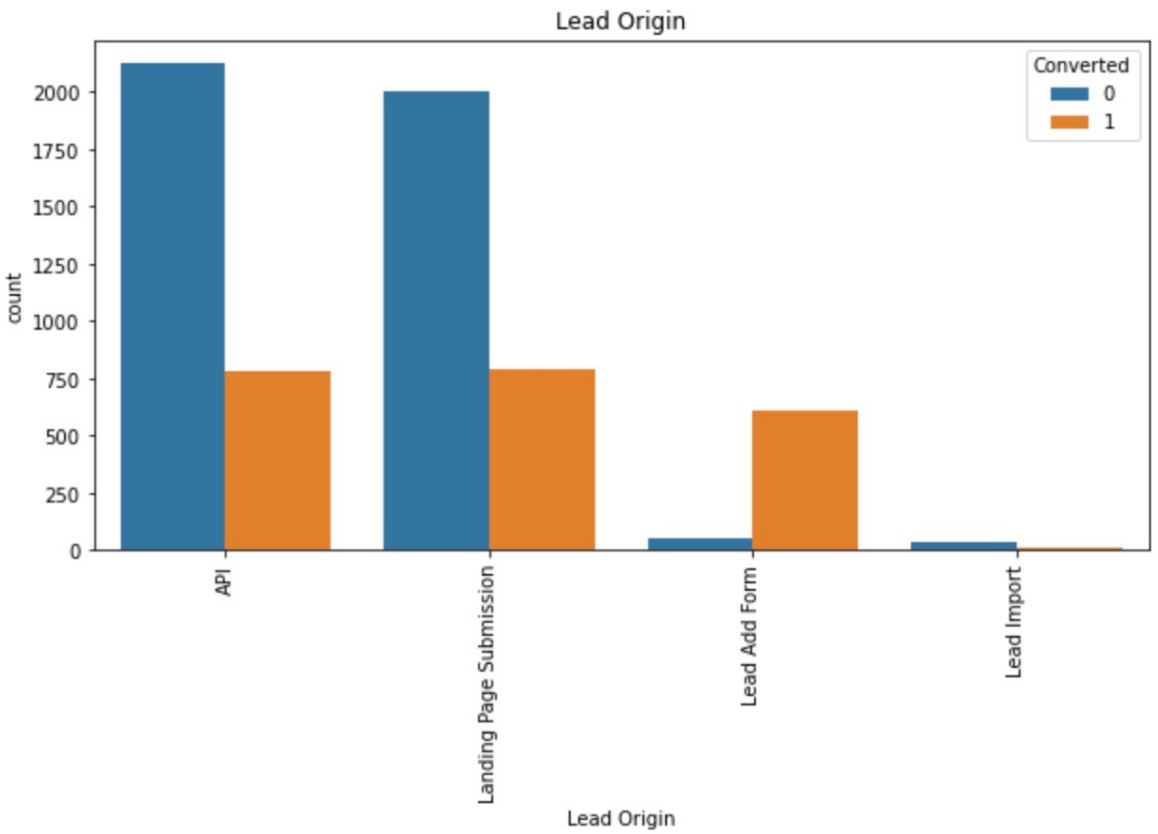
a. Univariate data analysis: value count, distribution of variable etc.

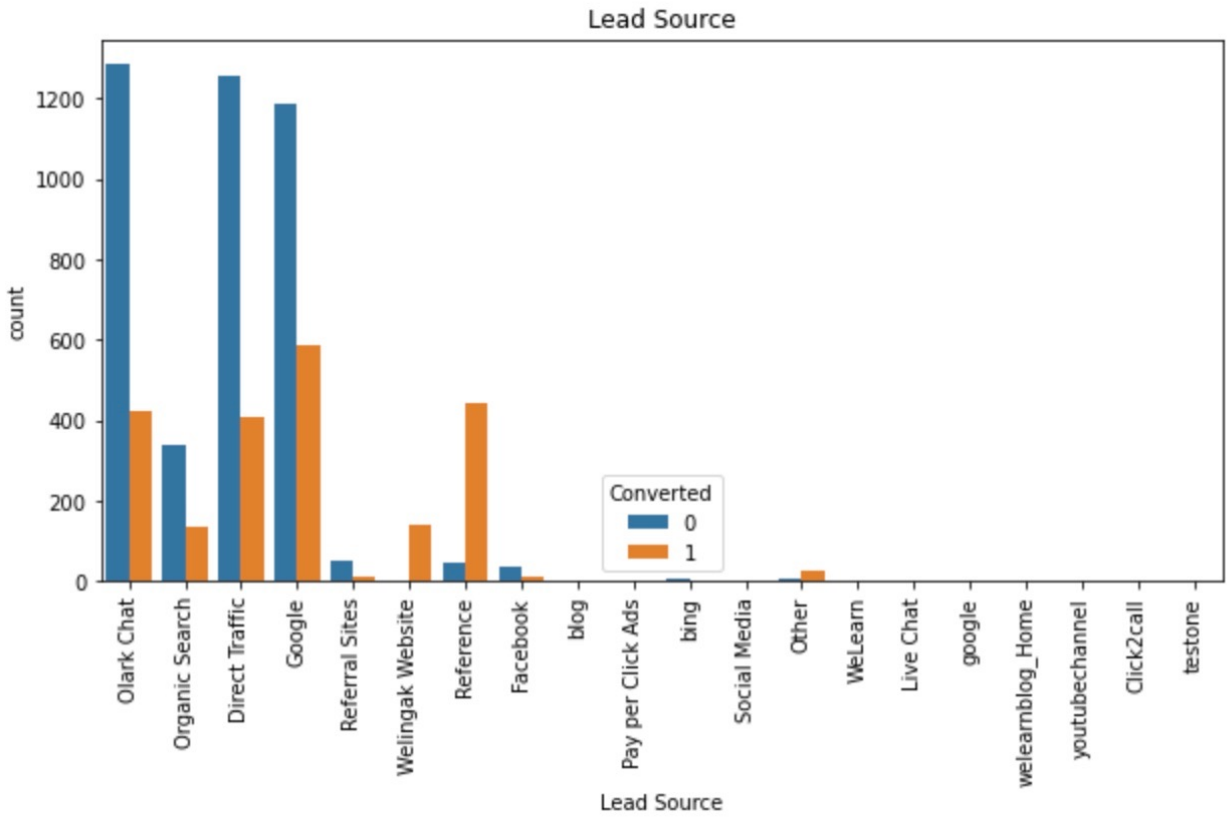
b. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

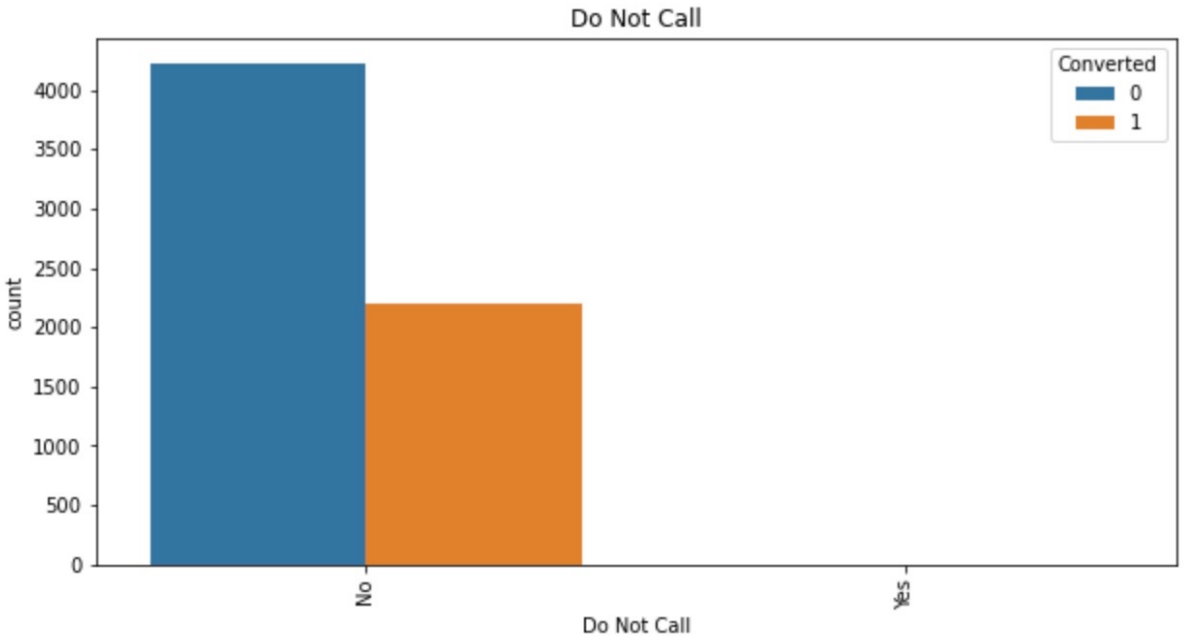
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

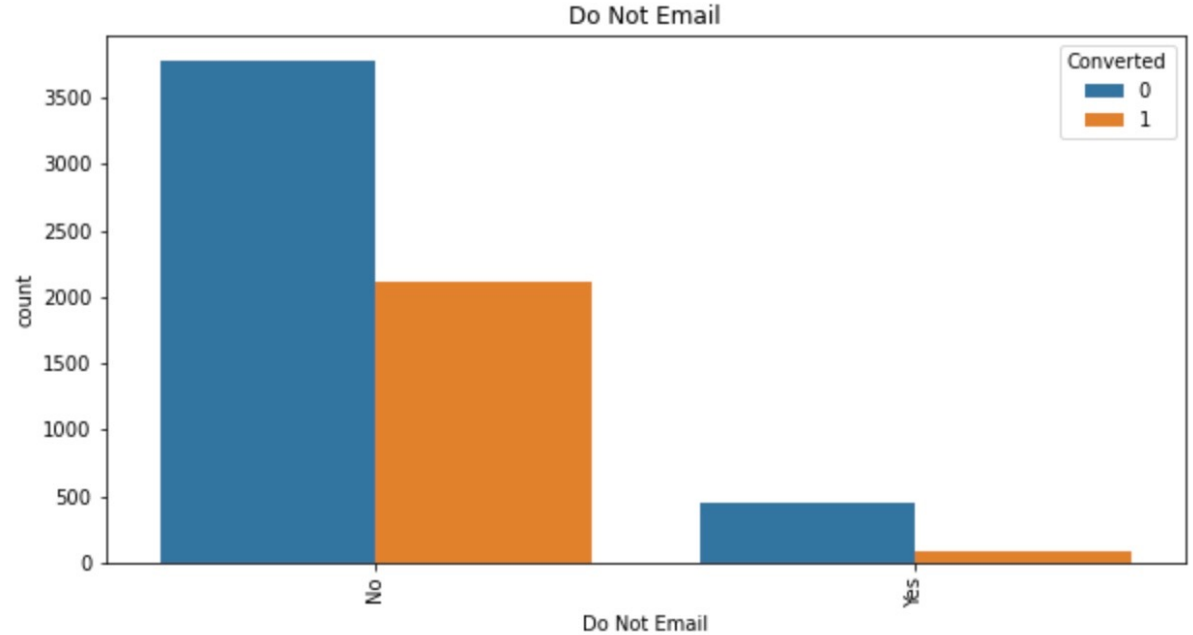
EDA(Exploratory Data Analysis)

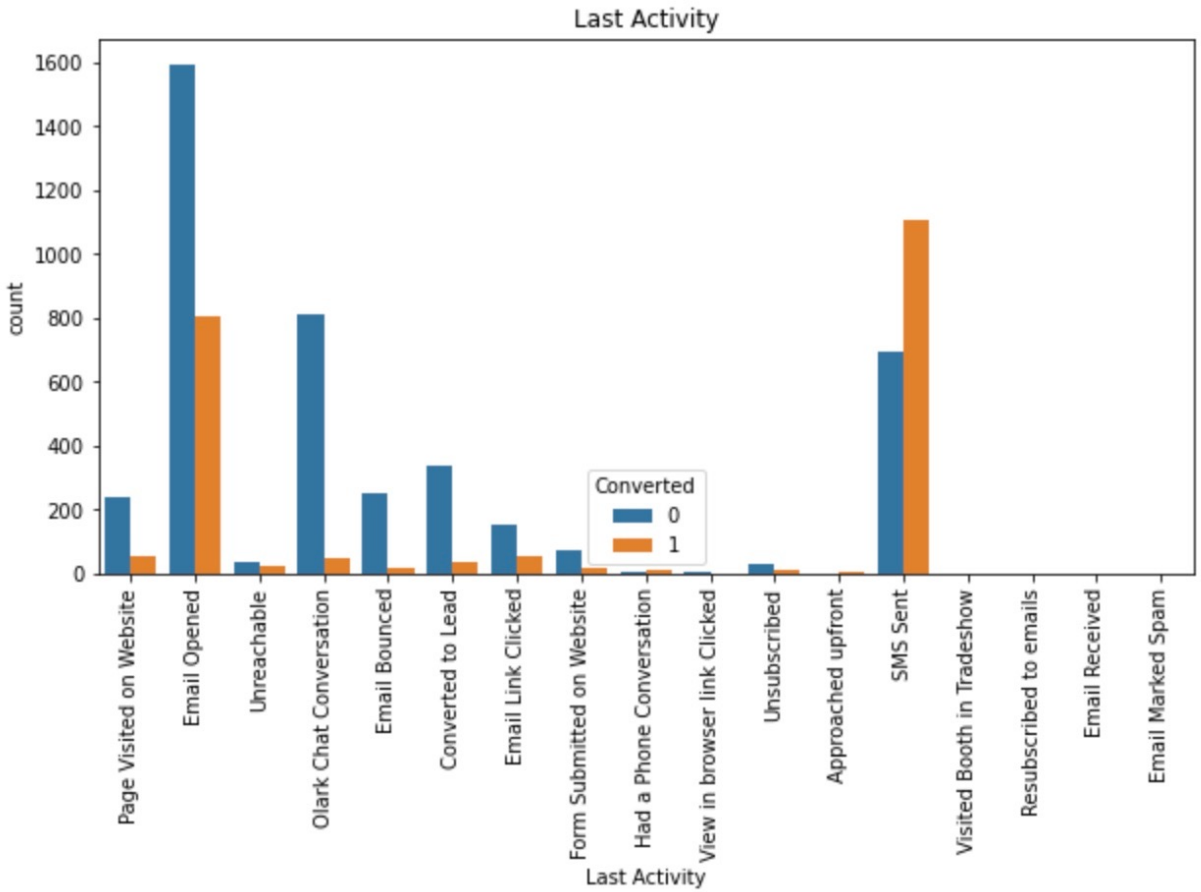
Categorical Analysis-

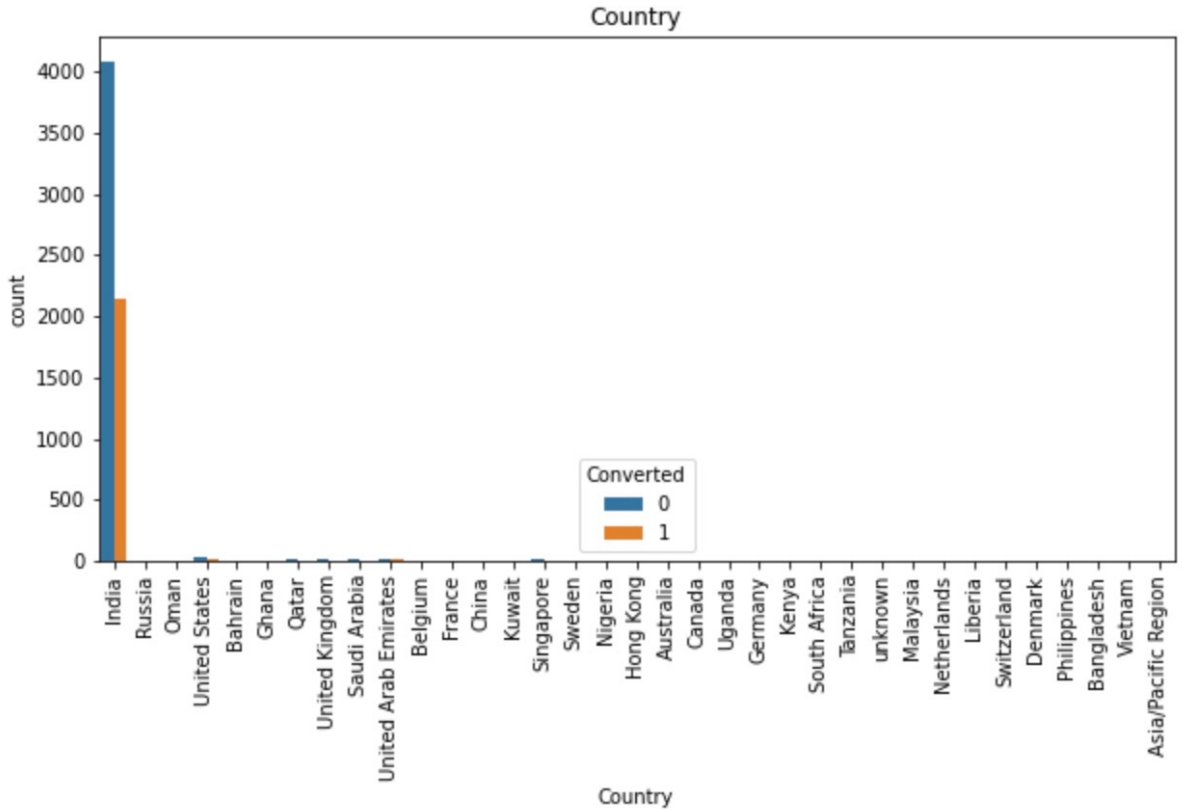


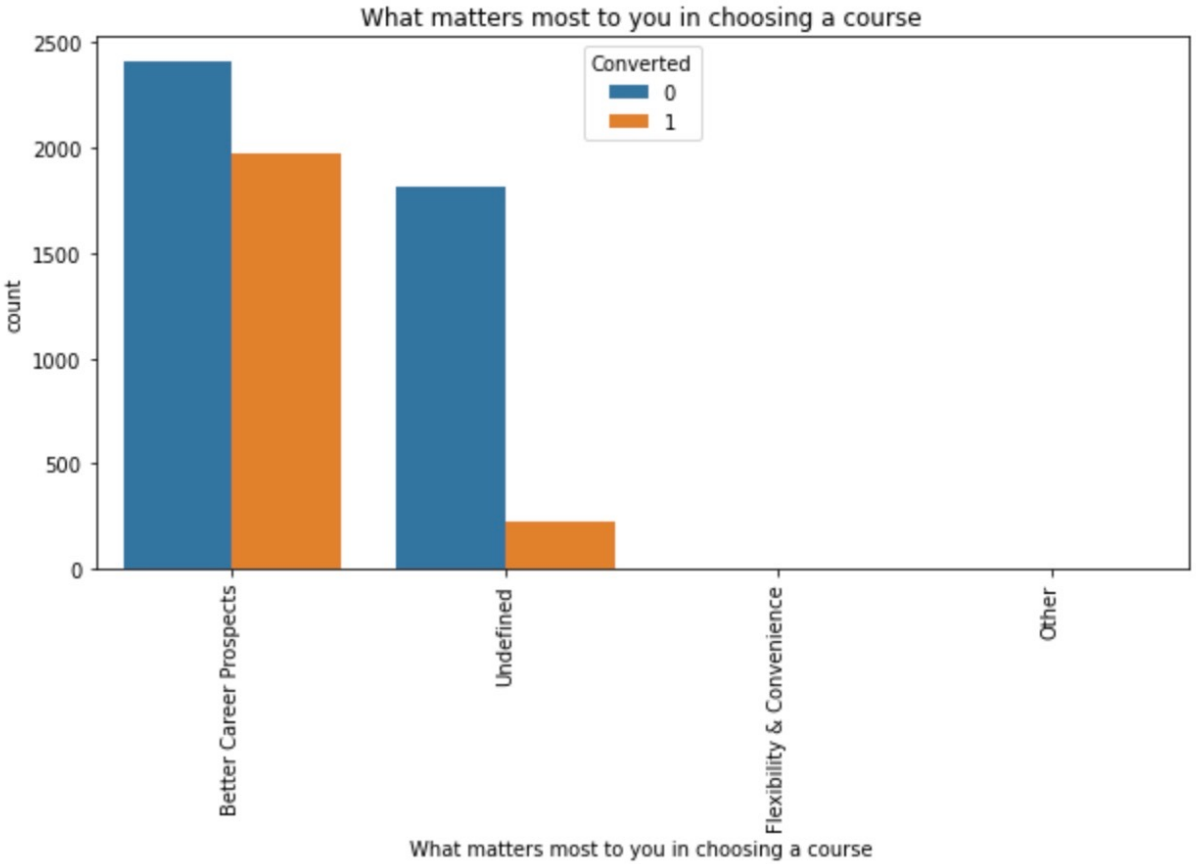


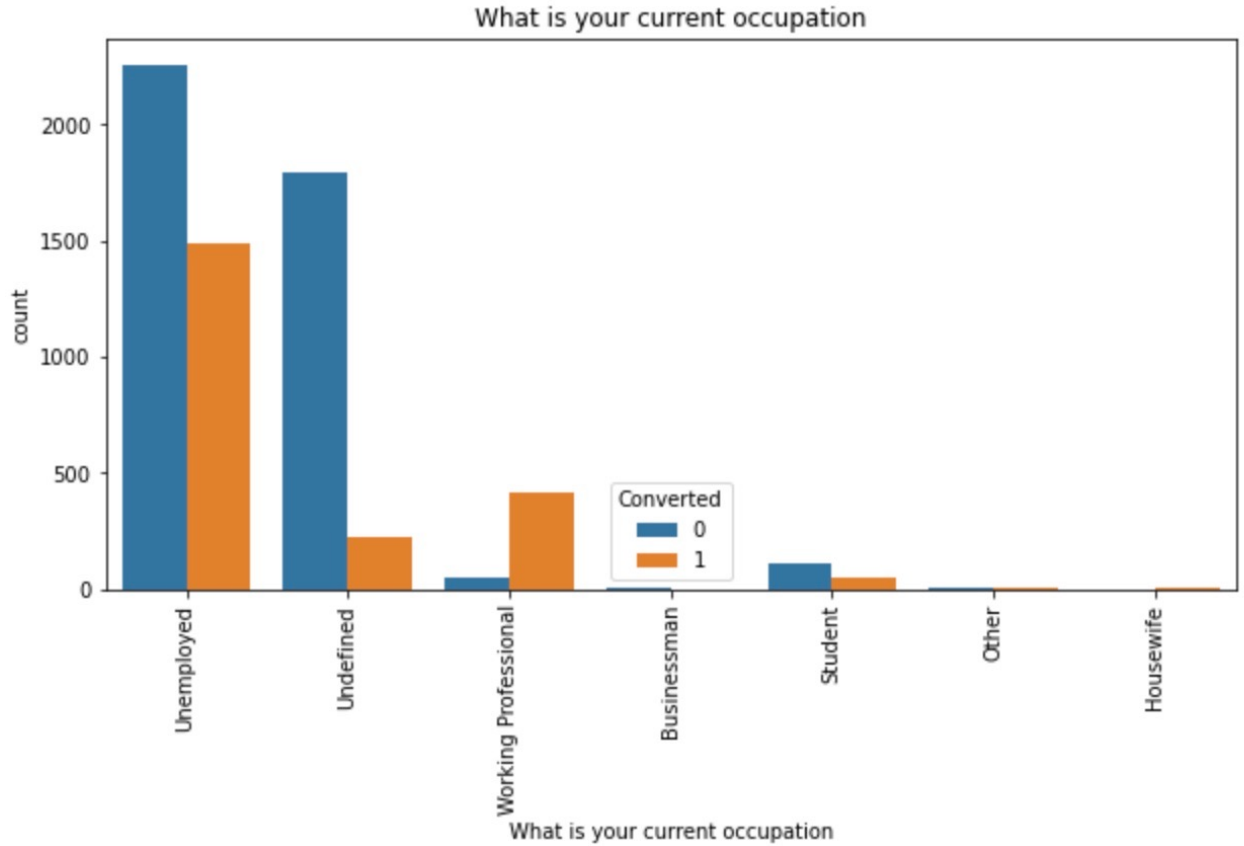

Do Not Call
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Do Not Call

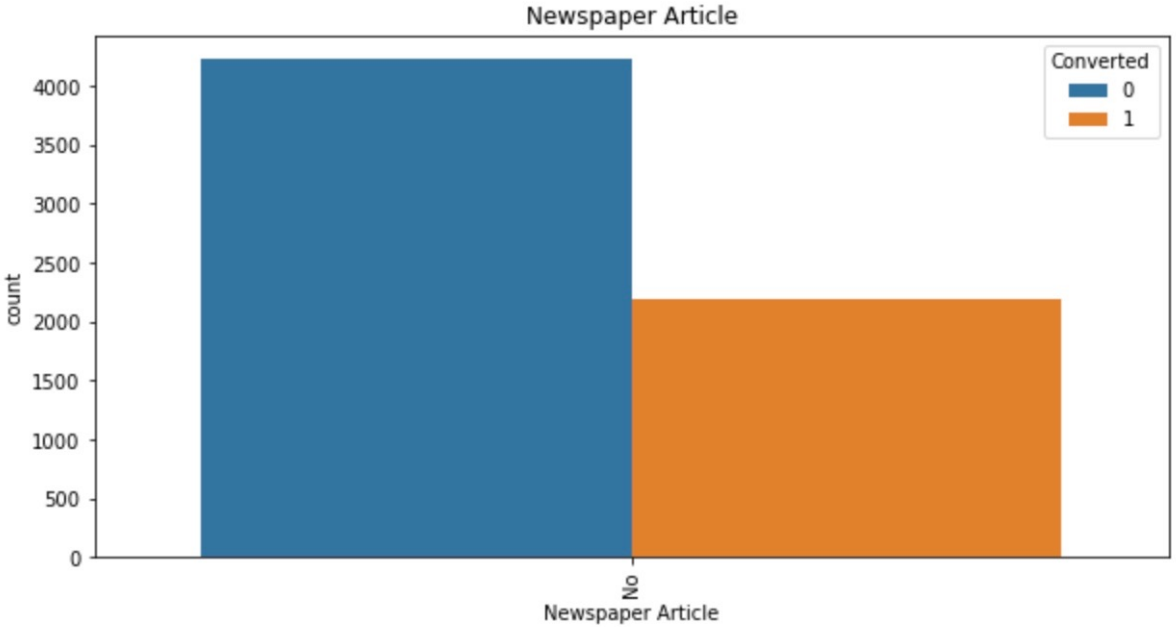

Do Not Email
Converted
0
1
count
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Do Not Email

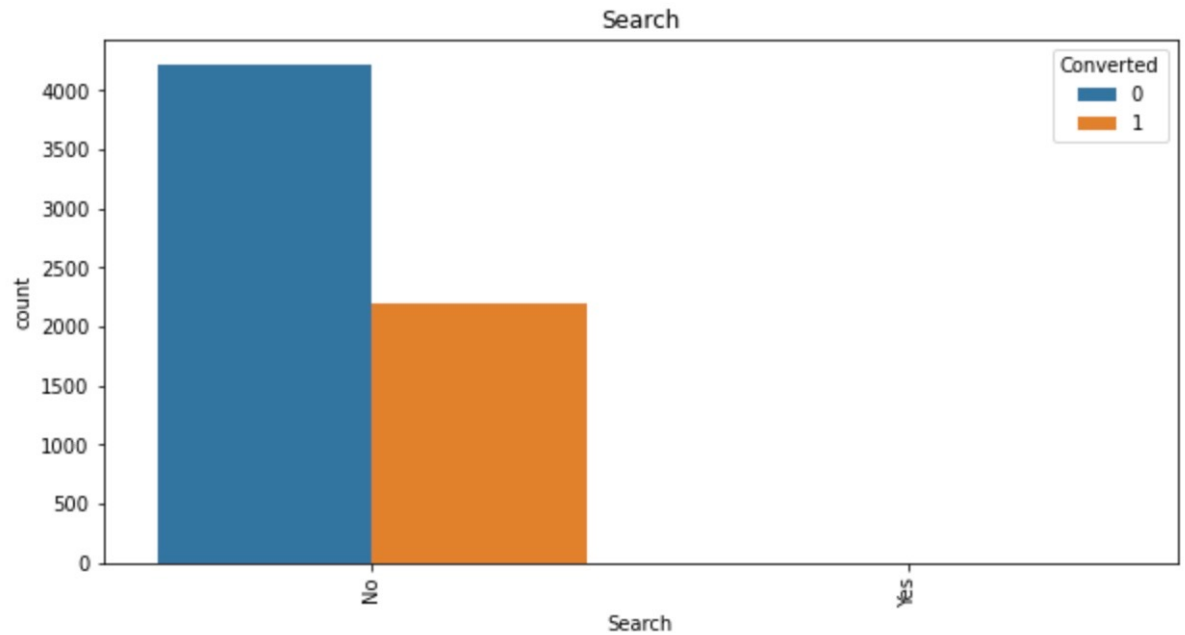

Last Activity
count
Converted
0
1
Page Visited on Website
Email Opened
Unreachable
Olark Chat Conversation
Email Bounced
Converted to Lead
Email Link Clicked
Form Submitted on Website
Had a Phone Conversation
View in browser link Clicked
Unsubscribed
Approached upfront
SMS Sent
Visited Booth in Tradeshow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

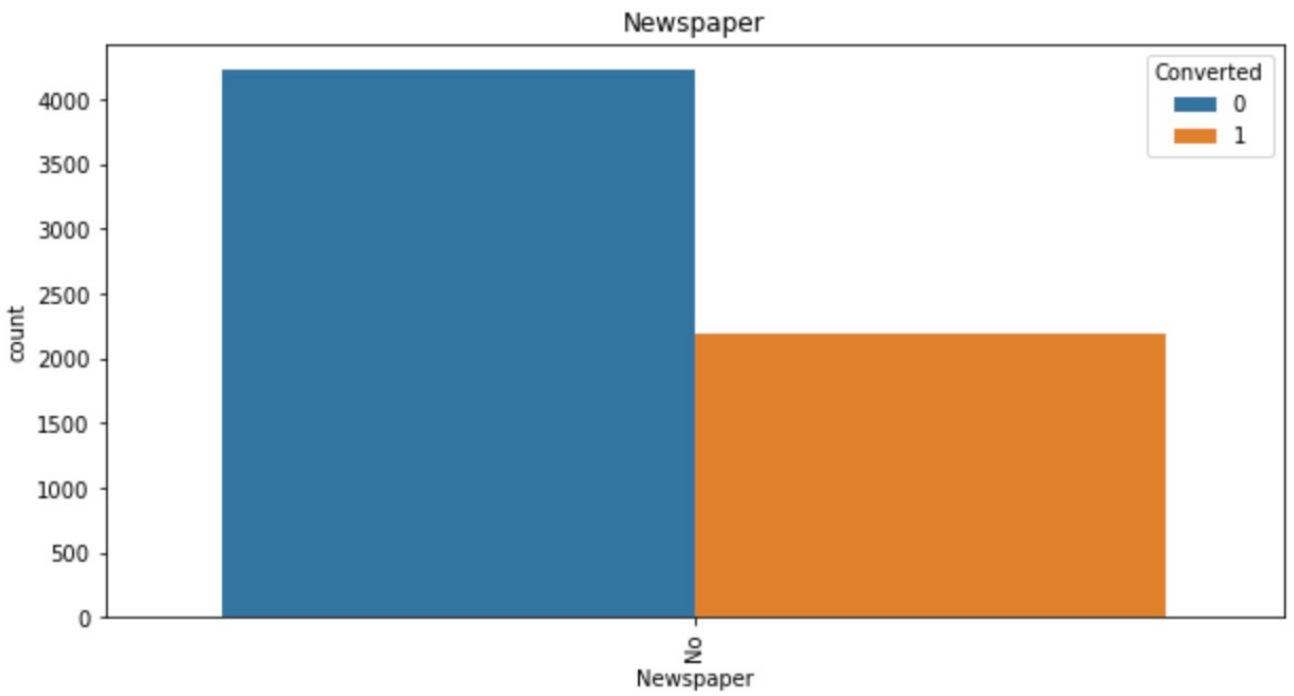

Country
count
Converted
0
1
India
Russia
Oman
United States
Bahrain
Ghana
Qatar
United Kingdom
Saudi Arabia
United Arab Emirates
Belgium
France
China
Kuwait
Singapore
Sweden
Nigeria
Hong Kong
Australia
Canada
Uganda
Germany
Kenya
South Africa
Tanzania
unknown
Malaysia
Netherlands
Liberia
Switzerland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

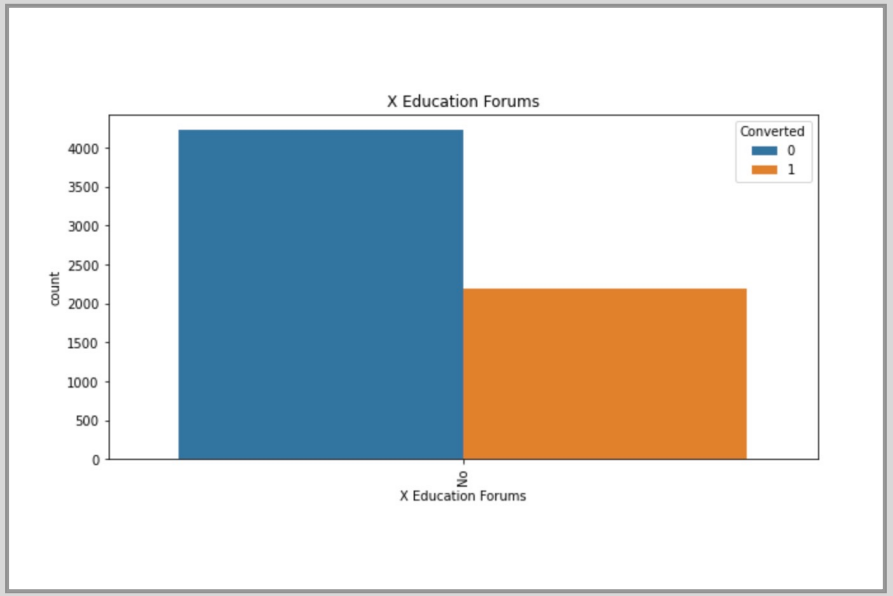

What matters most to you in choosing a course
Converted
0
1
count
2500
2000
1500
1000
500
0
Better Career Prospects
Undefined
Flexibility & Convenience
Other
What matters most to you in choosing a course

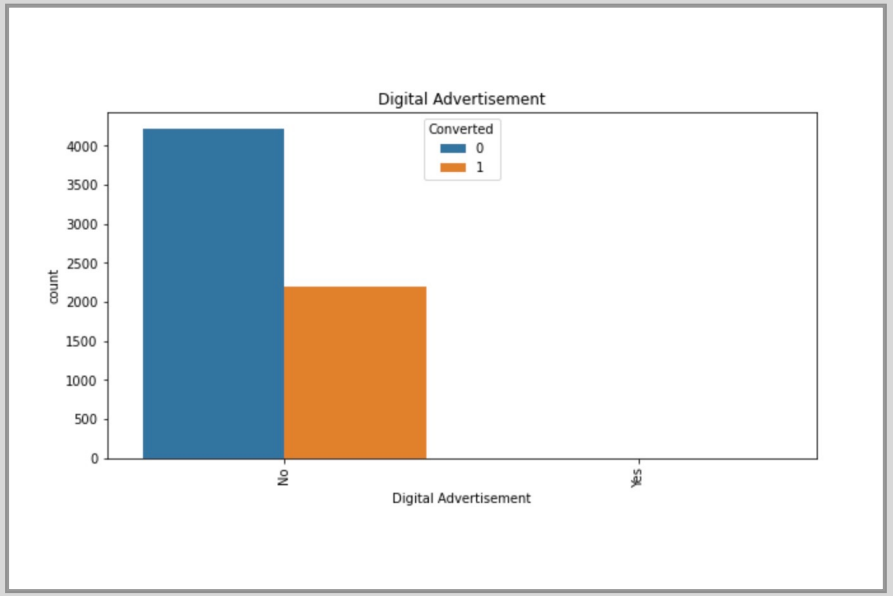

What is your current occupation
count
2000
1500
1000
500
0
Converted
0
1
Unemployed
Undefined
Working Professional
Businessman
Student
Other
Housewife
What is your current occupation

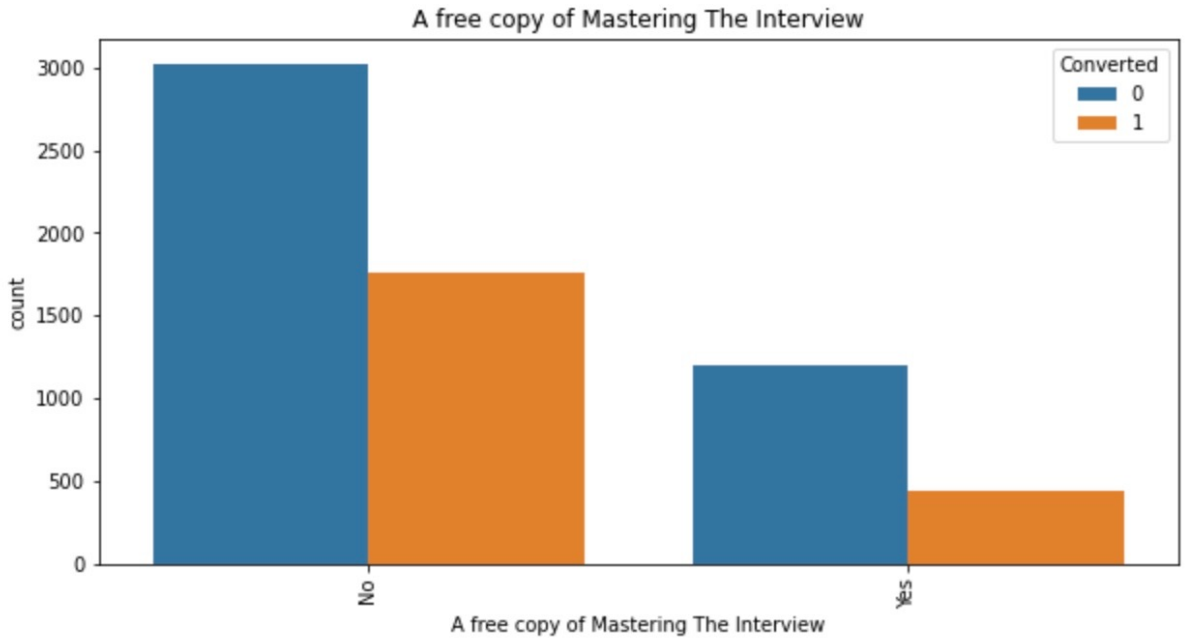

Newspaper Article
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
Newspaper Article

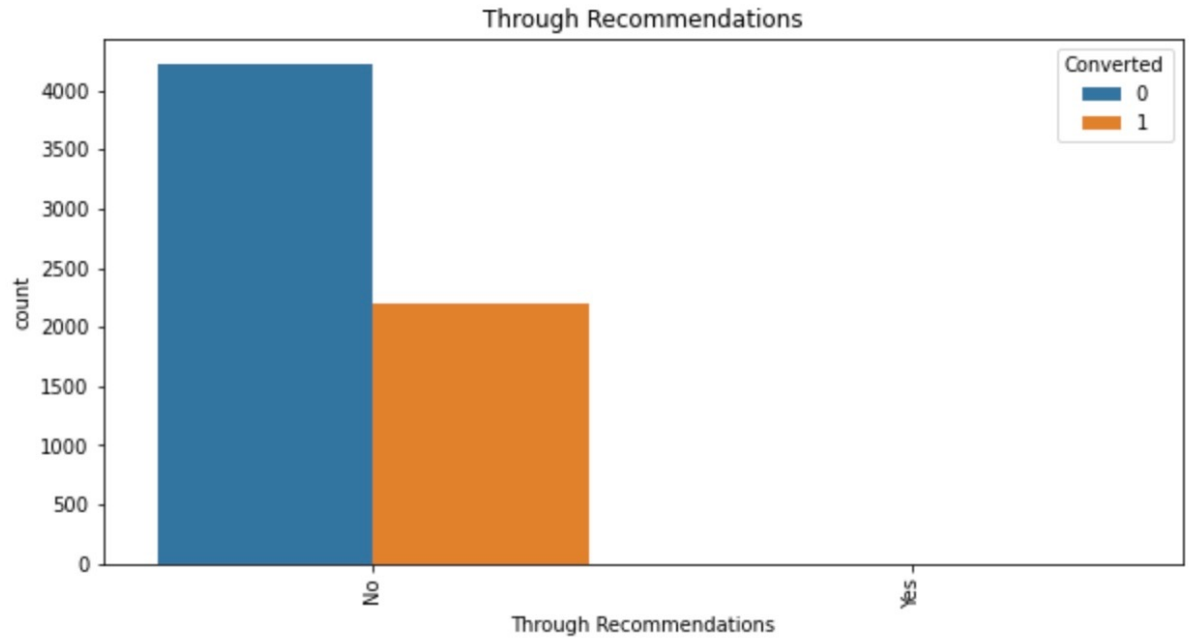

Search
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Search

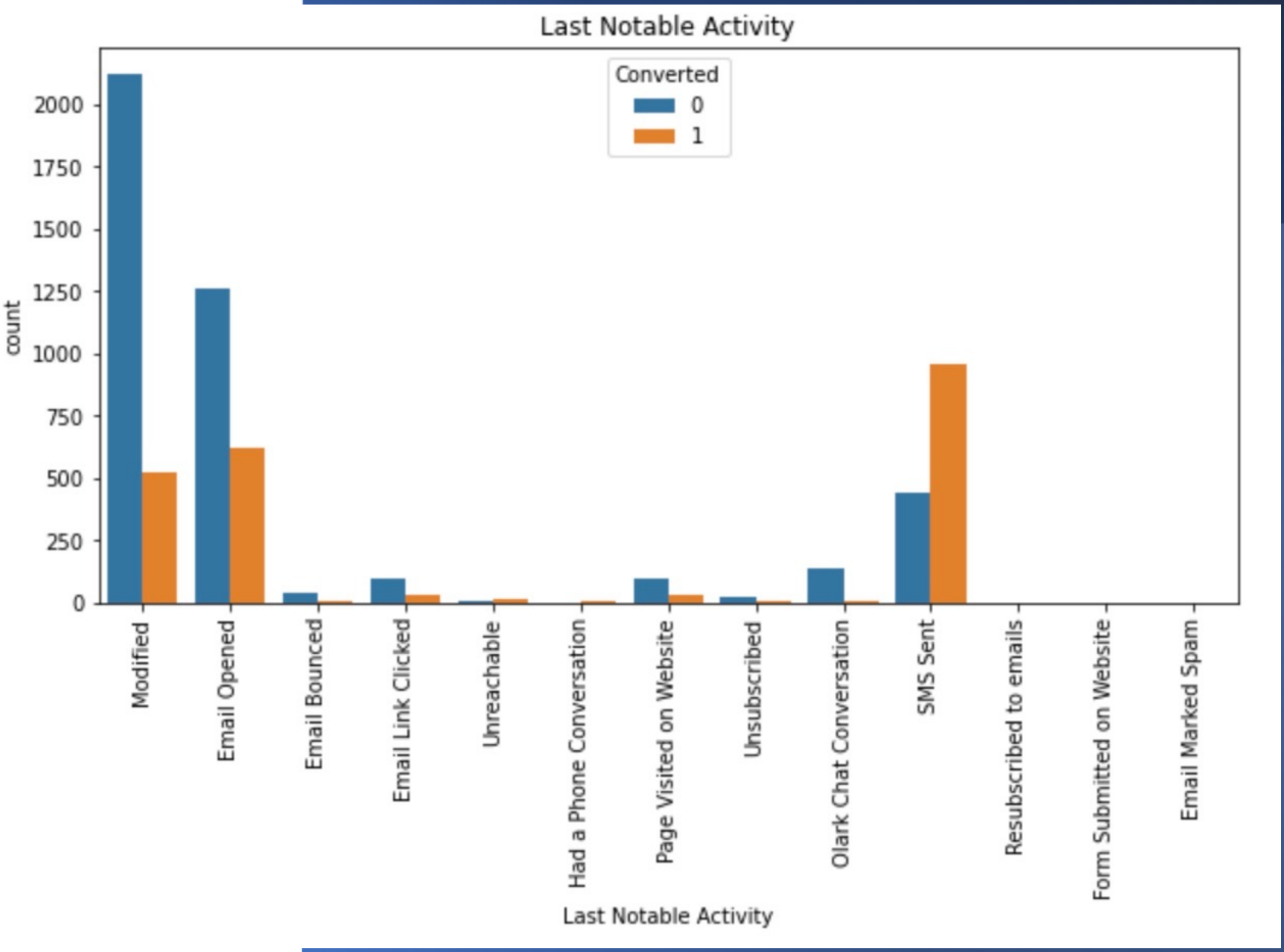

Newspaper
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
Newspaper


X Education Forums
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
X Education Forums


Digital Advertisement
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Digital Advertisement


A free copy of Mastering The Interview
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
No
Yes
A free copy of Mastering The Interview


Through Recommendations
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Through Recommendations


Last Notable Activity
Converted
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
Modified
Email Opened
Email Bounced
Email Link Clicked
Unreachable
Had a Phone Conversation
Page Visited on Website
Unsubscribed
Olark Chat Conversation
SMS Sent
Resubscribed to emails
Form Submitted on Website
Email Marked Spam
Last Notable Activity

Data Preparation

- Conversion of yes/No to binary 0/1.
- Created dummy variables for categorical columns.
- Outliers treatment.
- Test –Train split.
- Feature scaling on train dataset by StandardScaler()
- Correlation analysis includes removal of highly correlated variables.

•

Model Building Steps:

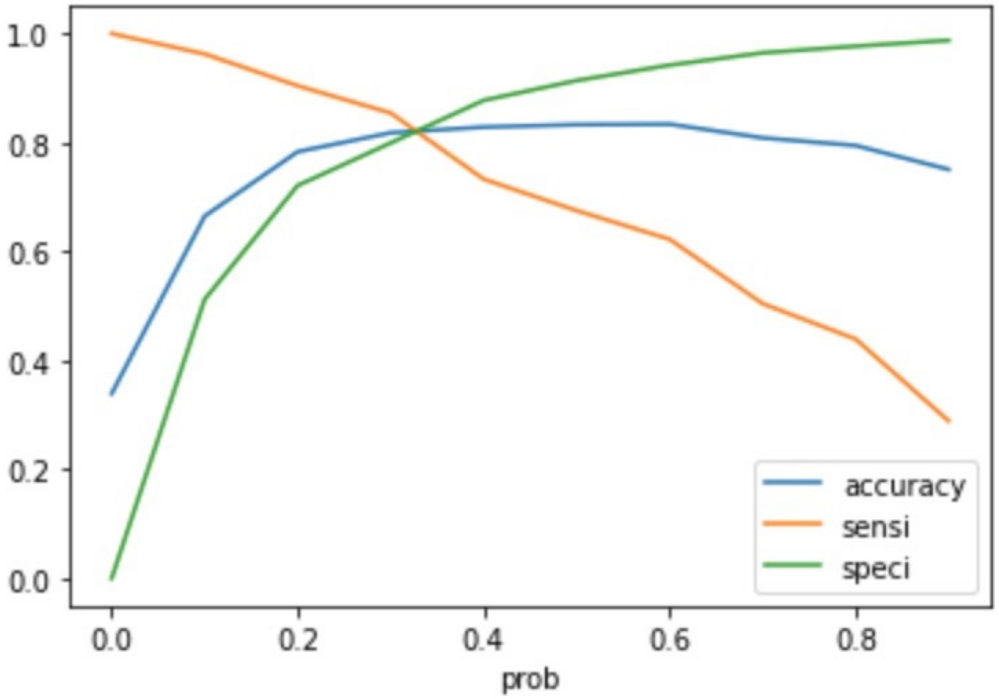
- As mentioned before, Dataset has been split in train and test dataset.
- Firstly, Dataset has been divided into 70:30 ratio for train and test data.
- Performed feature scaling with the help of StandardScaler().
- Followed by feature selection using RFE method.
- We built model-1 with 20 features/column values.
- Dropped features with high p value > 0.05 and VIF > 5.
- Made predictions on train set.
- Got an accuracy of 82.7%.

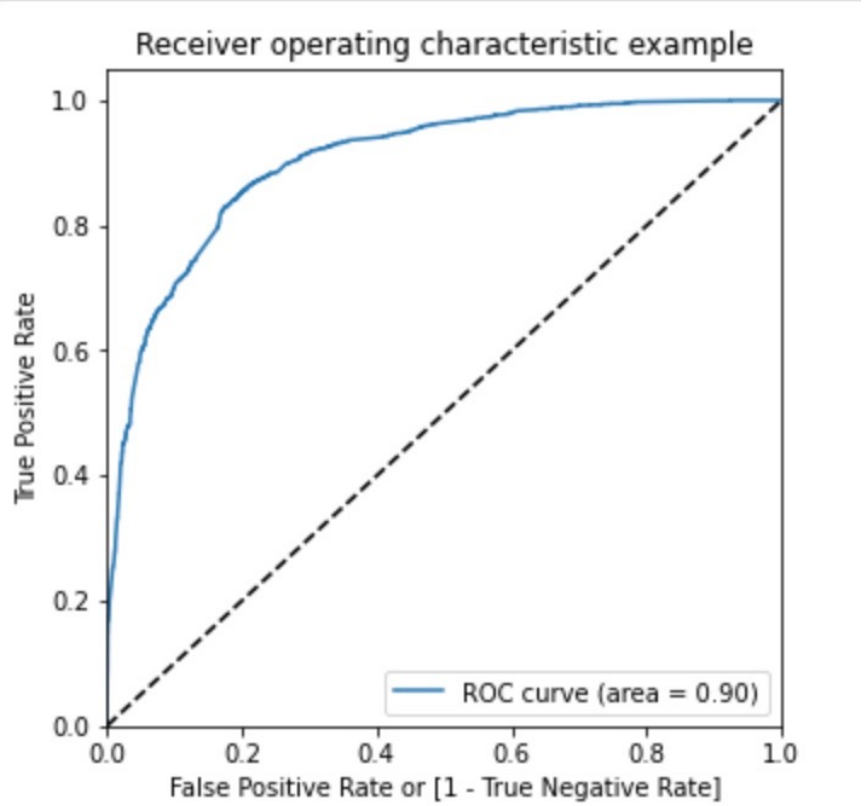
Prediction on test set:

- Accuracy : 0.8276220145379024
- Sensitivity : 0.8422619047619048
- Specificity : 0.8197767145135566
- Confusion matrix on test data:

1028	226
106	566

Plotting ROC curve:





Observations:

We need to calculate cutoff probability.

Cutoff probability is the point where sensitivity, accuracy and specificity, are balanced.

From the previous curve, 0.34 is the optimum point is cutoff probability.

Precision – Recall:

This method was also used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 75% on the test data frame.

Summary:

- From our model, we observe that : The leads who fills the form are the potential leads. We must majorly focus on working professionals. Always focus on customers, who have spent significant time on our website. Referral leads have less conversion rate Specialization is mandatory area to fill to study and target the right audience, if it's empty it's better to focus less on such cases.

Conclusion :

1. Reducing the number of call attempts to 2-3 or max 4 and increasing the frequency usage of other media like advertisements in GoogleAds, or via emails to keep in touch with the leads.

2. Update the data more frequently as the potential leads shows interest and run the model for better results.

3. Focusing on conversion rate by engaging with hot Leads (Active) will increase the chances of obtaining more value to the business as the number of people we contact are less.

4.With cold calling, it’s good practice to mail the leads as well.

The END