LEAD SCORE CASE STUDY

Prepared by,

Nikhil Pal

Anakha B R

PROBLEM STATEMENT

- An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

GOALS AND OBJECTIVES:

- X Education company wants us to 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 and identify the hot leads.
- Deployment of the model for the future use.

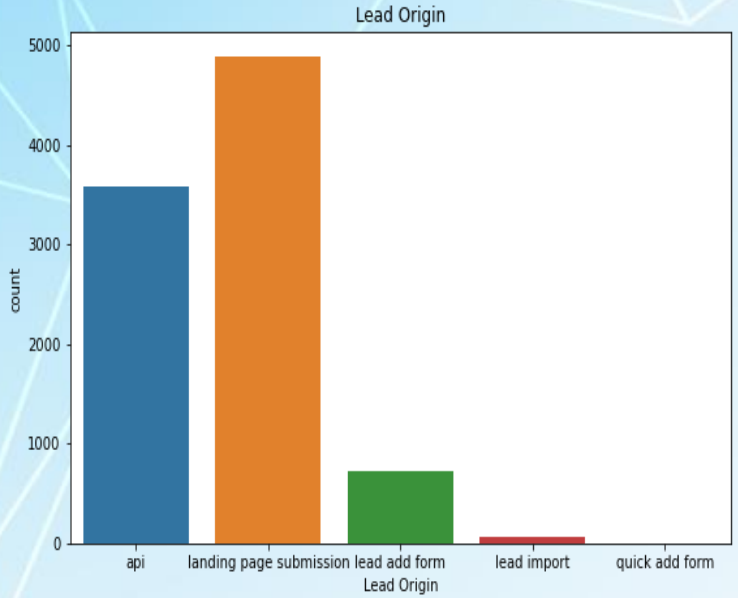
STEPS

- Data loading and Data Cleaning
  - Inspect the dataframe
  - Check and handle unique valued columns
  - Check and handle NA values and missing values
  - Drop unnecessary columns
  - Imputation of values, if needed.
  - Check and handle outliers
- EDA
  - Univariate analysis
    - Categorical analysis
    - Analyzing numerical variables
  - Relating all the categorical variables to Converted
- Dummy variable creation
- Scaling numeric features
- Model Building and prediction using logistic regression technique.
- Model evaluation
- Conclusions and recommendations

DATA CLEANING

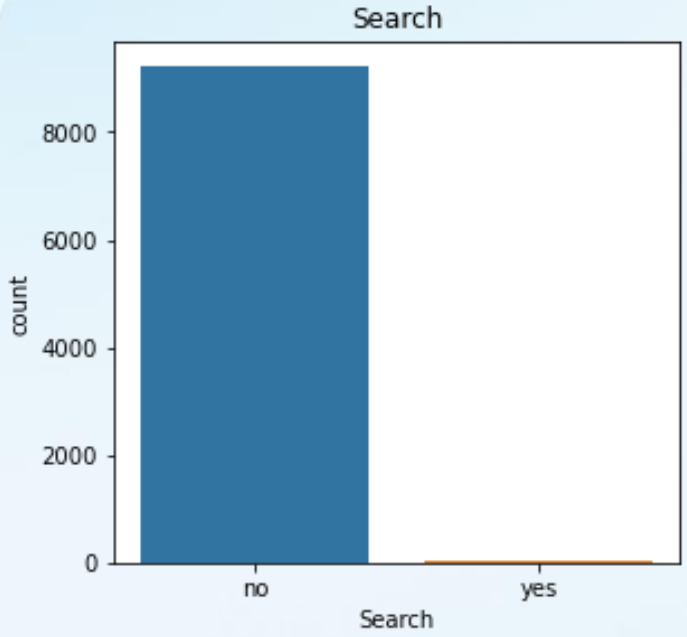
- Total number of rows and columns in dataset “Lead” is 9240 and 37 respectively..
- Converting all the values to lower case.
- Replacing option “Select” with Nan (Since it means no option is selected).
- Dropping unique valued columns such as ‘Magazine’, ‘Receive More Updates About Our Courses’, ‘I agree to pay the amount through cheque’, ‘Get updates on DM Content’ and ‘Update me on Supply Chain Content’.
- Removing all the columns that are no required and have 35% null values. Those columns are ‘Asymmetrique Profile Index’, ‘Asymmetrique Activity Index’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Profile Score’, ‘Lead Profile’, ‘Tags’, ‘Lead Quality’, ‘How did you hear about X Education’, ‘City’ and ‘Lead Number’.
- Replace Nan values with ‘not provided’ in columns having huge null variables. Those columns are ‘Specialization’, ‘What matters most to you in choosing a course’, ‘Country’ and ‘What is your current occupation’.
- Drop column ‘Prospect ID’ since this is just indicative of id number of customers.

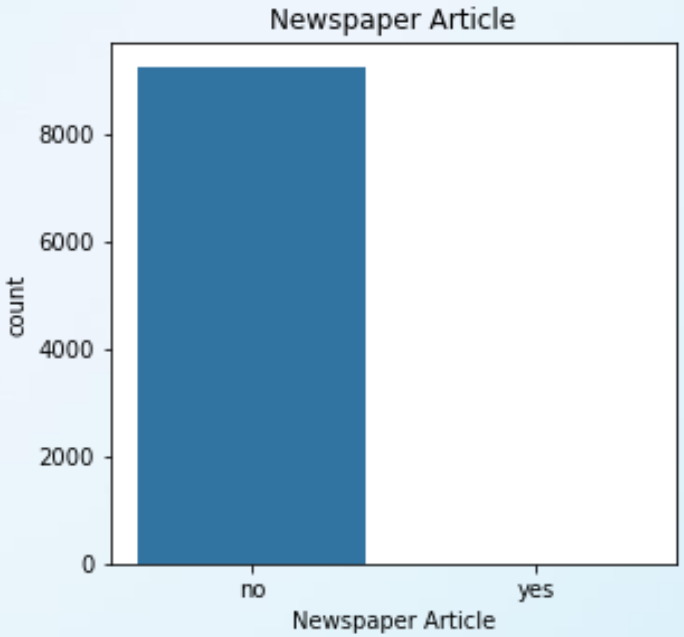
EDA

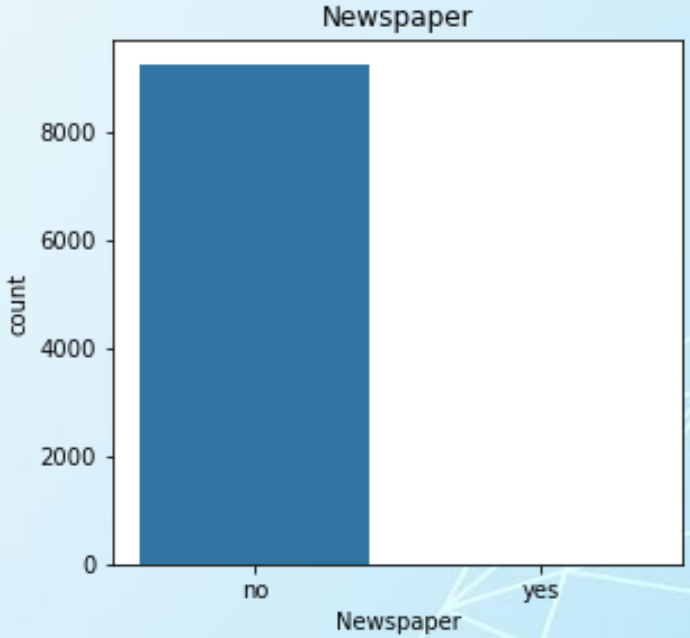

Lead Origin
count
api
landing page submission
lead add form
lead import
quick add form
Lead Origin

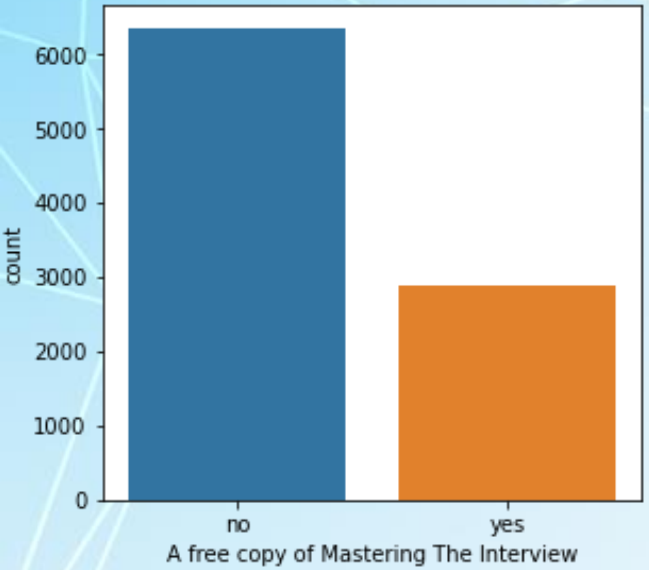

Do Not Email
count
no
yes
Do Not Email

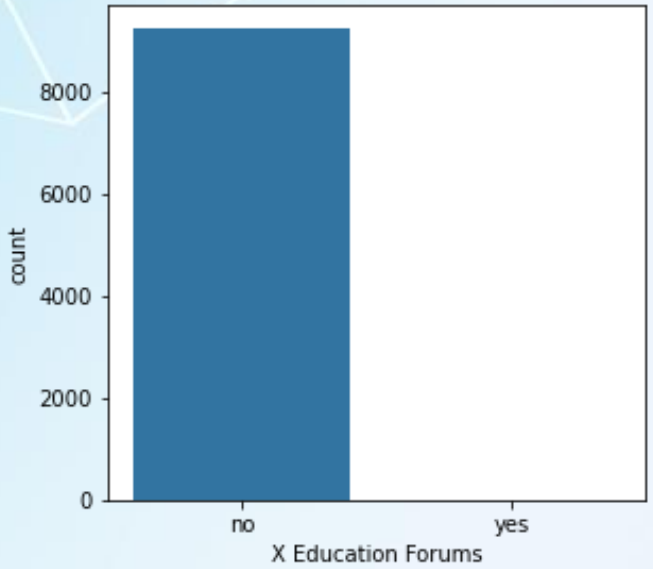

Do Not Call
count
no
yes
Do Not Call

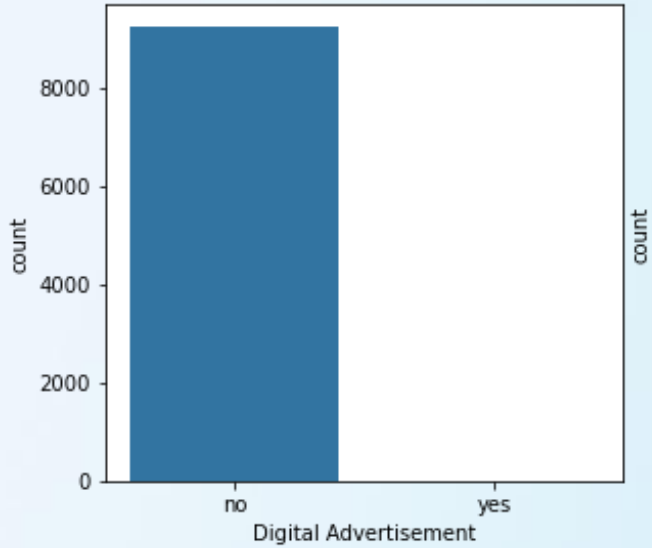

Search
count
no
yes
Search

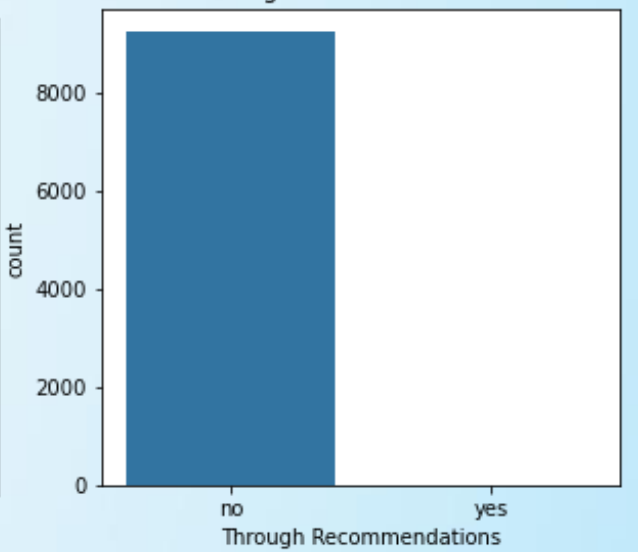

Newspaper Article
count
no
yes
Newspaper Article

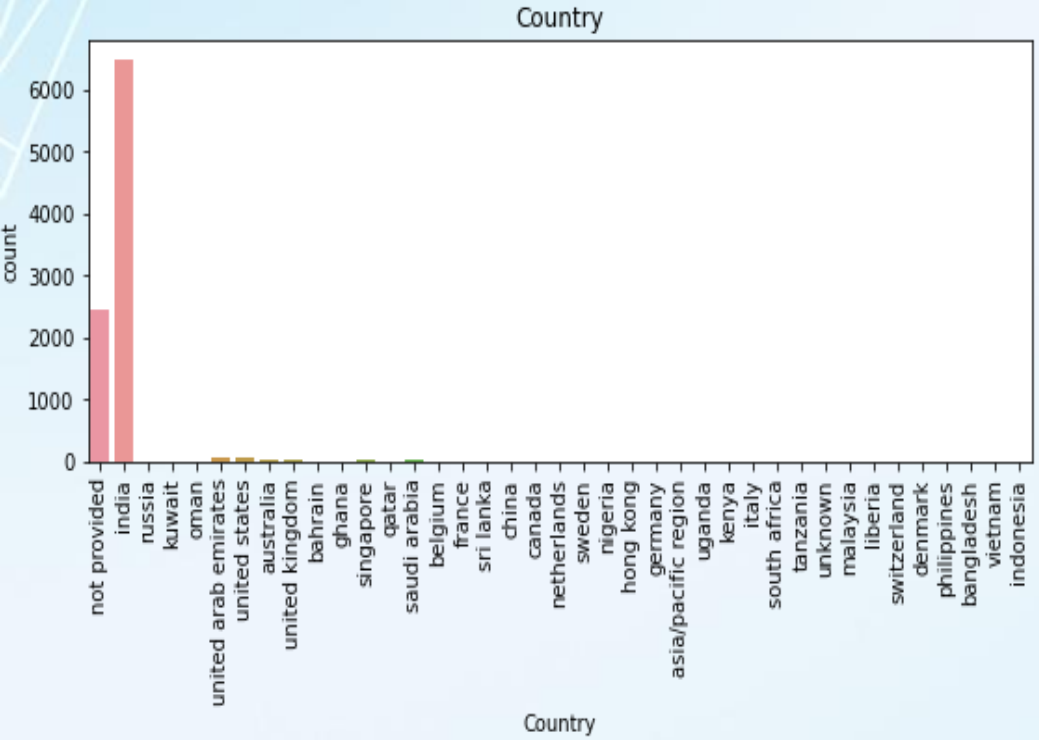

Newspaper
count
no
yes
Newspaper

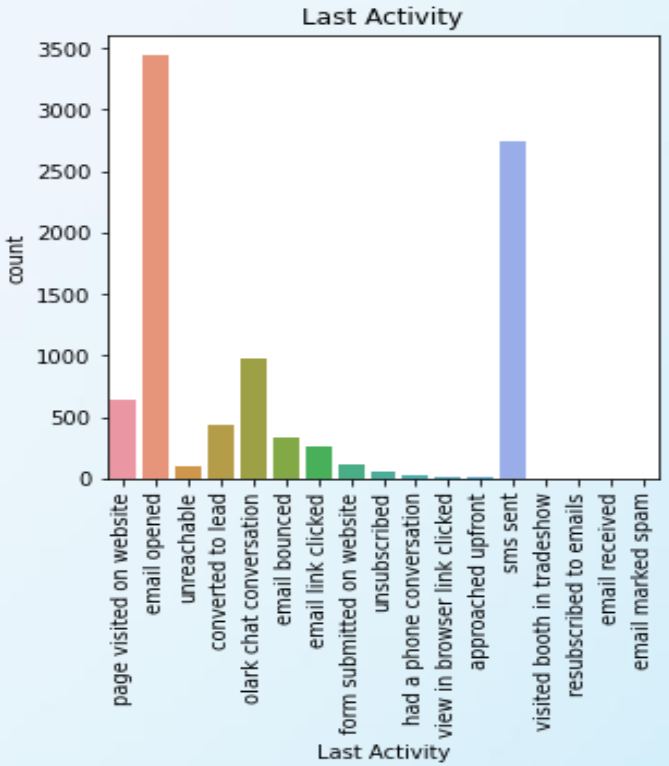

A free copy of Mastering The Interview
count
6000
5000
4000
3000
2000
1000
0
no
yes
A free copy of Mastering The Interview

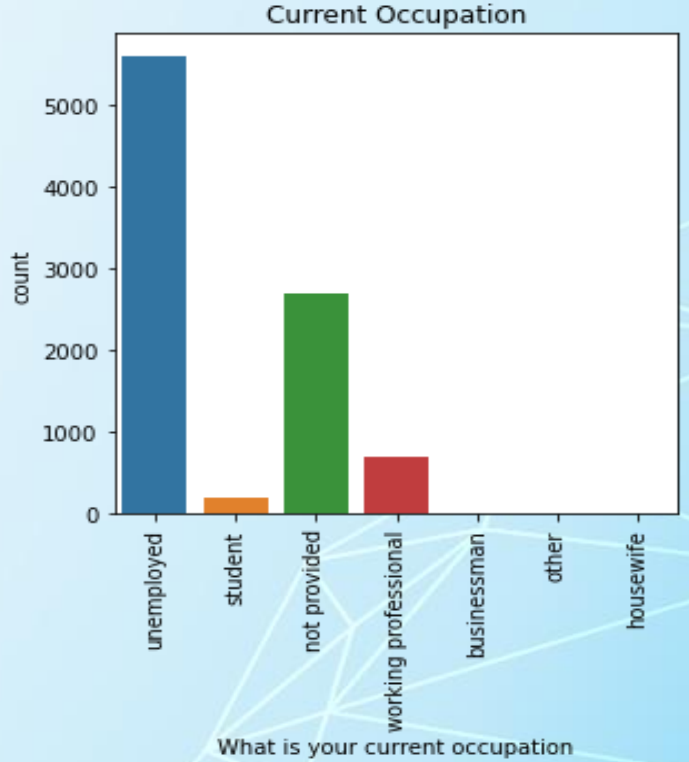

X Education Forums
count
8000
6000
4000
2000
0
no
yes
X Education Forums

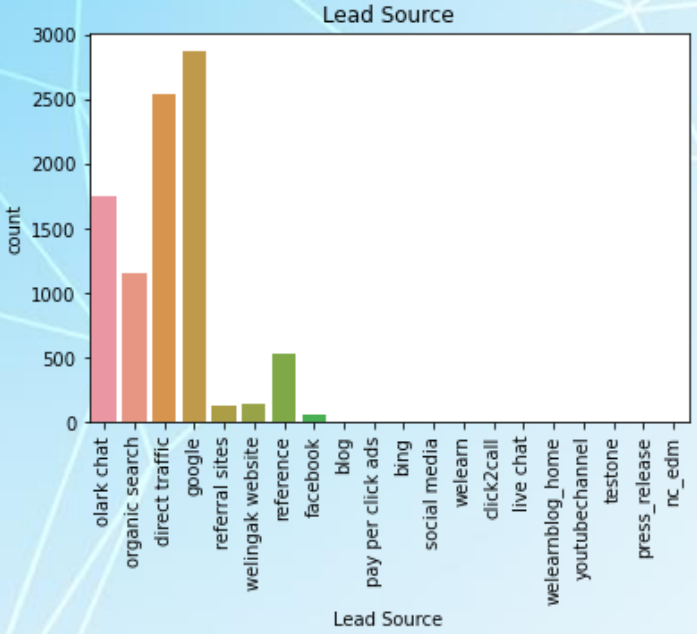

Digital Advertisement
count
8000
6000
4000
2000
0
no
yes
Digital Advertisement

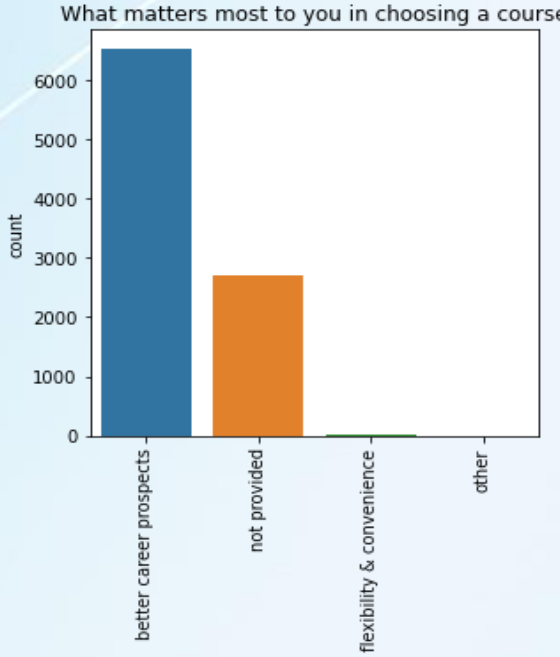

Through Recommendations
count
8000
6000
4000
2000
0
no
yes
Through Recommendations

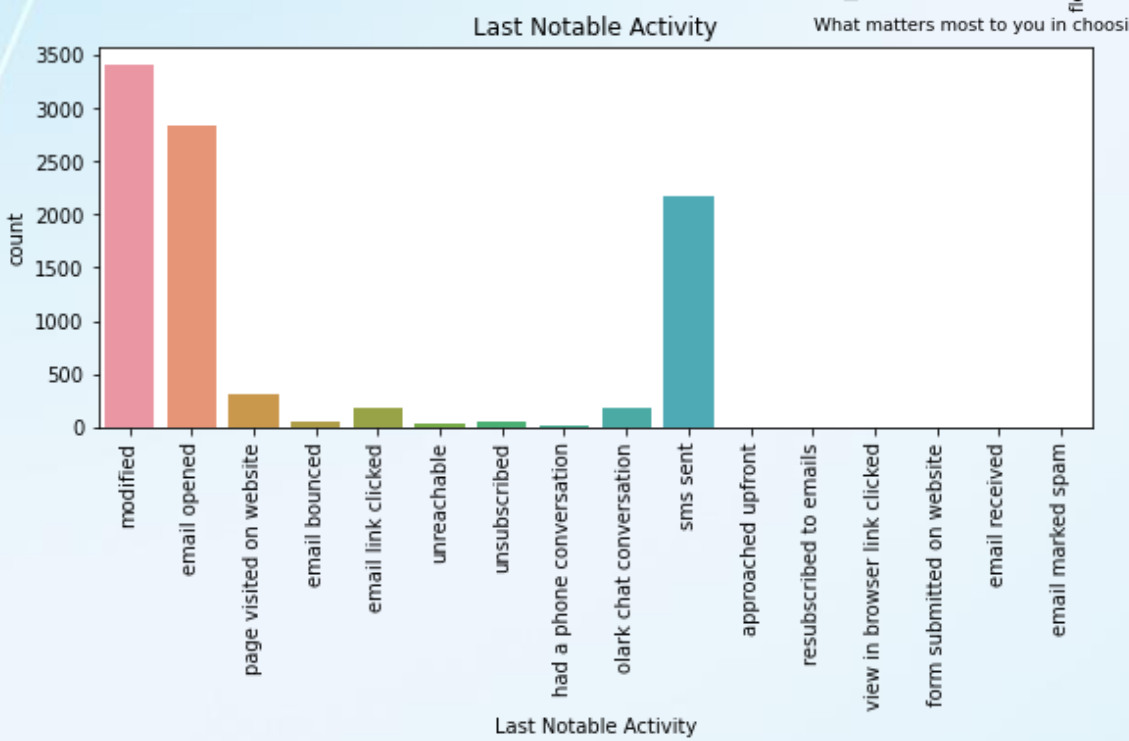

Country
count
6000
5000
4000
3000
2000
1000
0
not provided
india
russia
kuwait
oman
united arab emirates
united states
australia
united kingdom
bahrain
ghana
singapore
qatar
saudi arabia
belgium
france
sri lanka
china
canada
netherlands
sweden
nigeria
hong kong
germany
asia/pacific region
uganda
kenya
italy
south africa
tanzania
unknown
malaysia
liberia
switzerland
denmark
philippines
bangladesh
vietnam
indonesia
Country

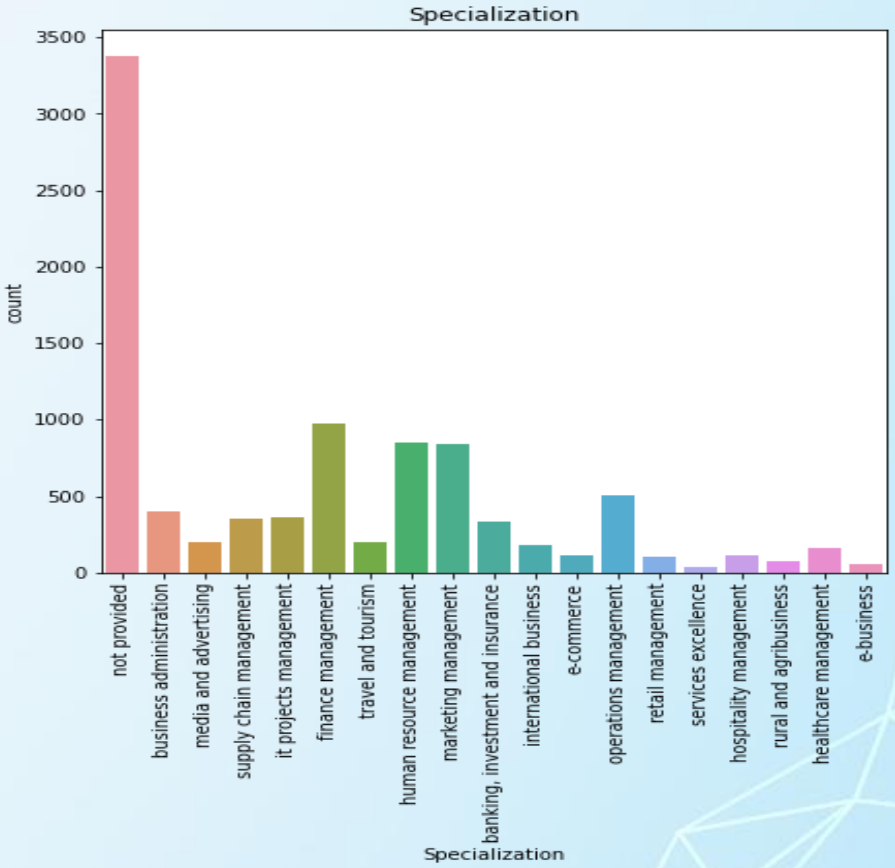

Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
approached upfront
sms sent
visited booth in tradeshow
resubscribed to emails
email received
email marked spam
Last Activity


Current Occupation
count
5000
4000
3000
2000
1000
0
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation

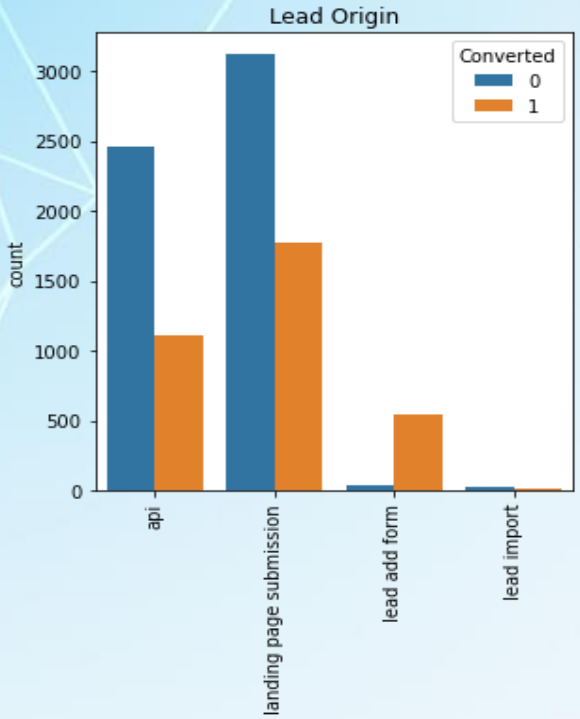

Lead Source
count
olark chat
organic search
direct traffic
google
referral sites
welingak website
reference
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

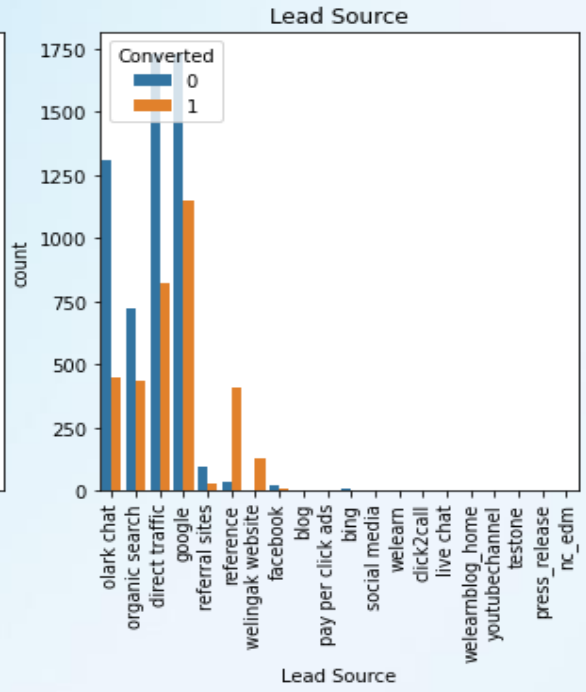

What matters most to you in choosing a course
count
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

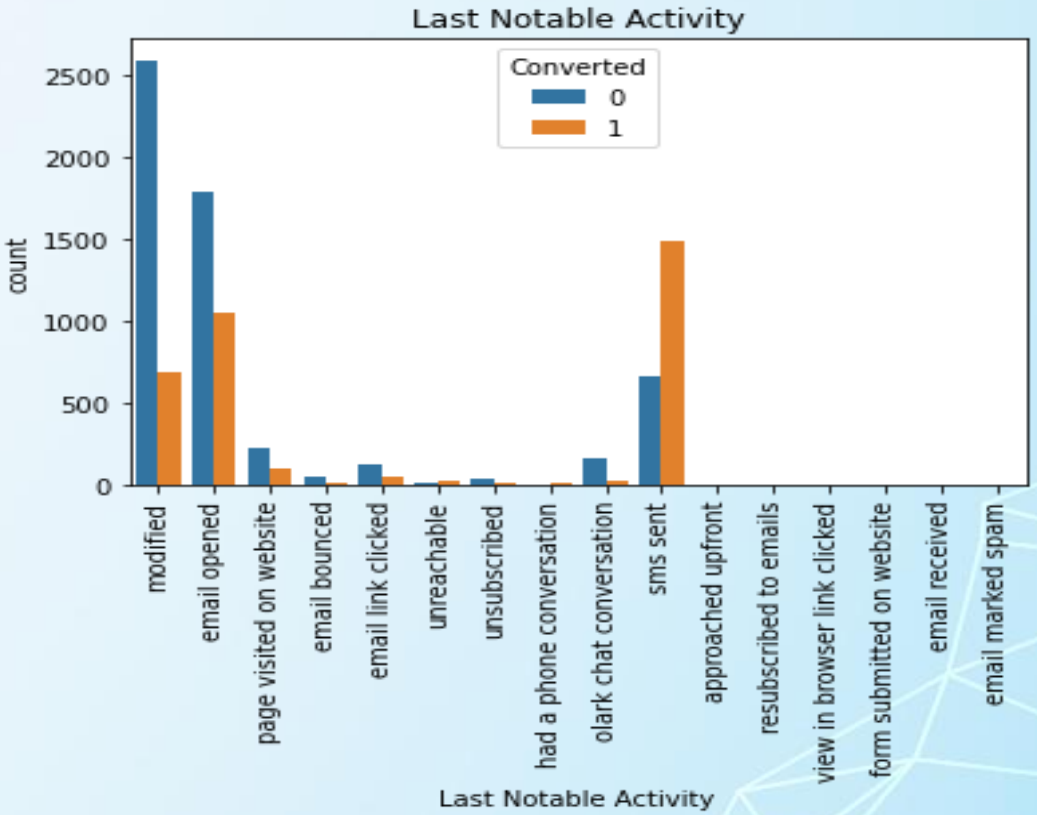

Last Notable Activity
count
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received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

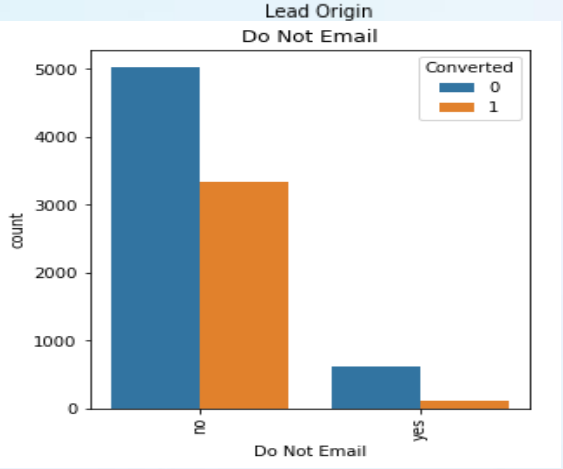

Specialization
count
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization

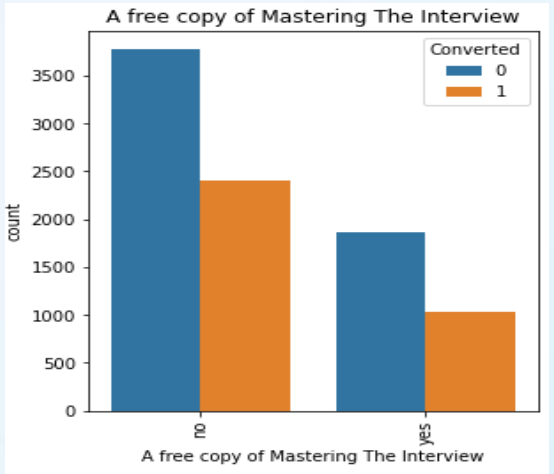
- Relating all the categorical variables to Converted

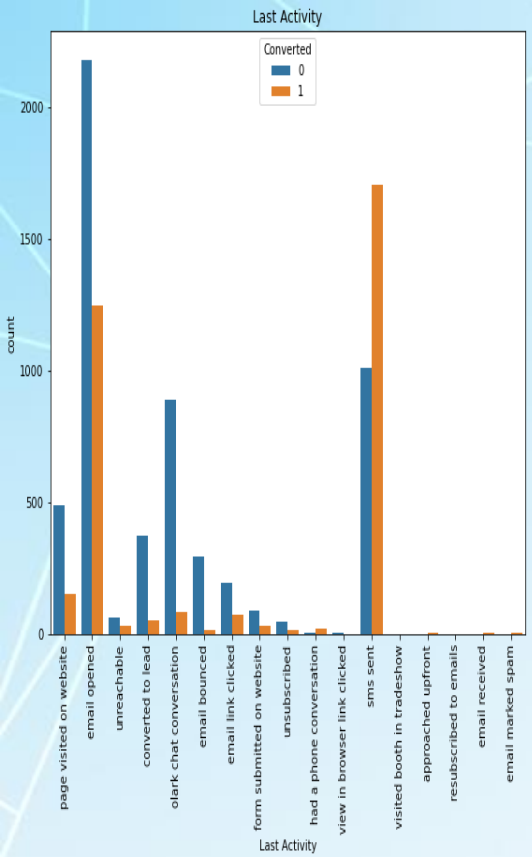


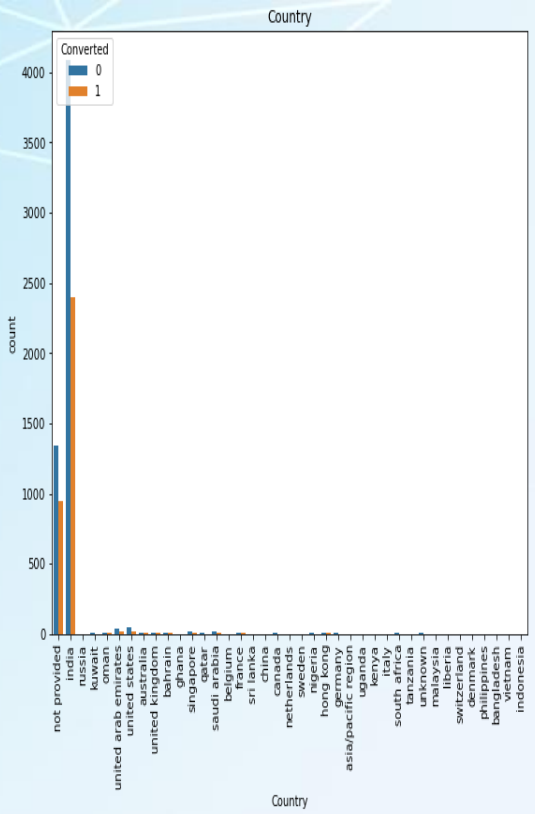


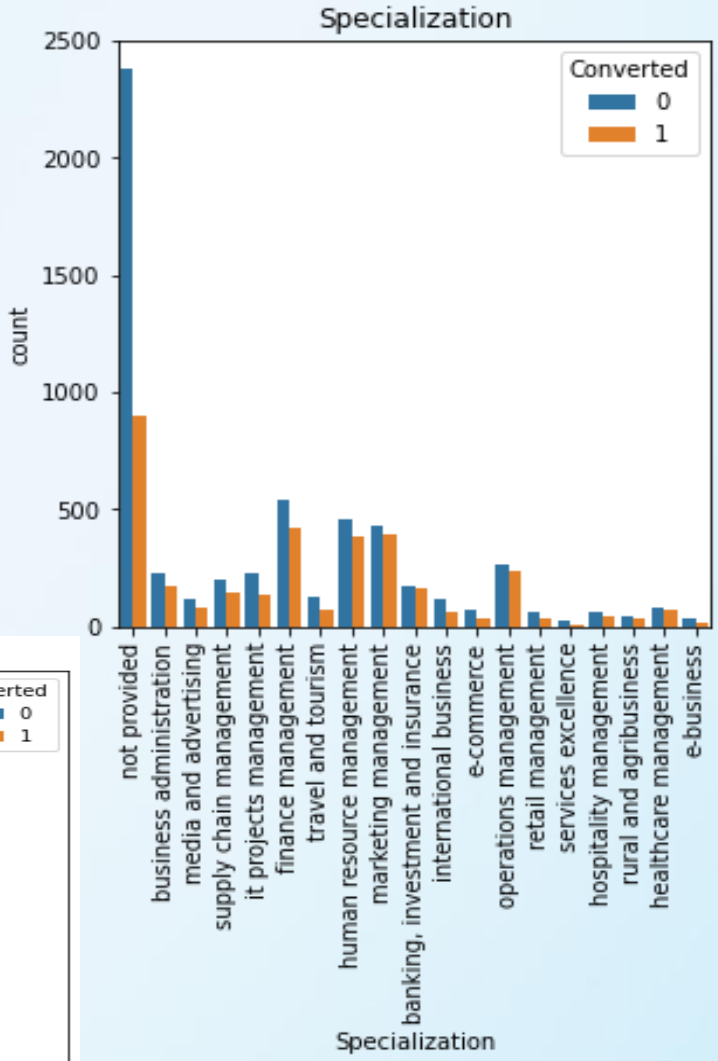


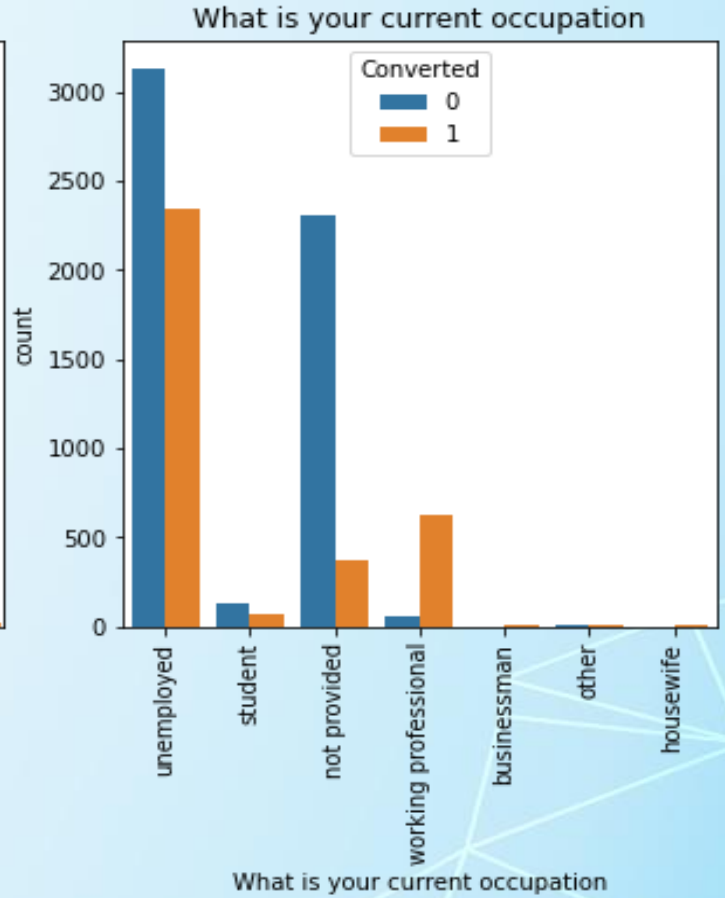


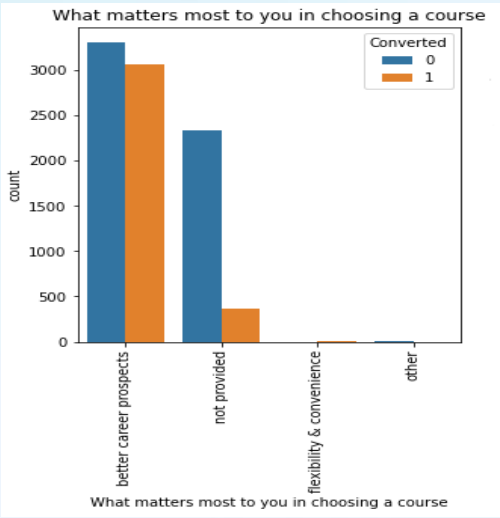


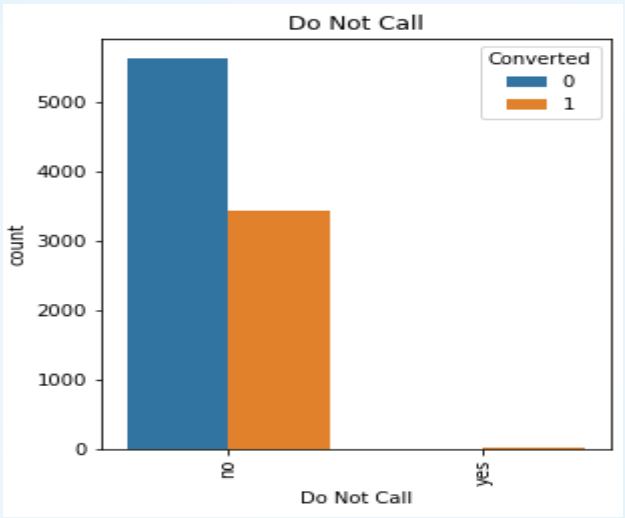

Last Activity
Converted
0
1
count
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront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

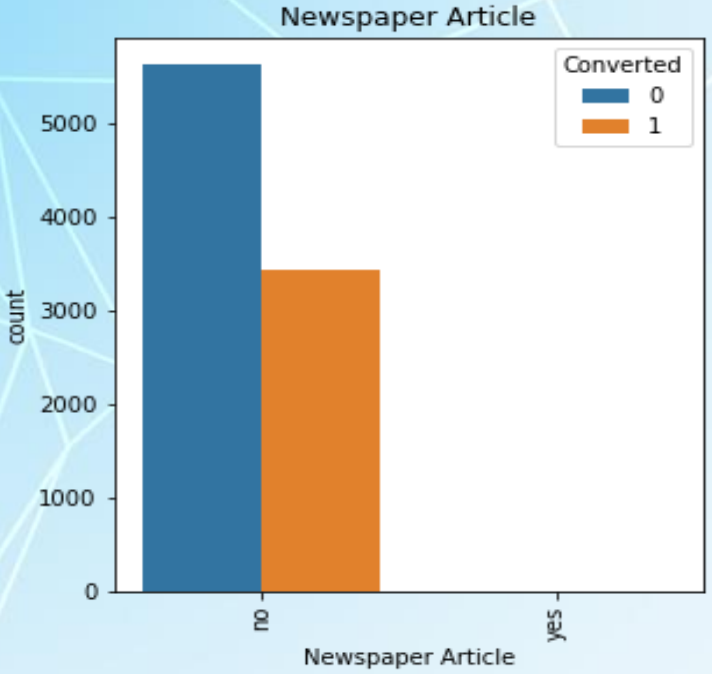

Country
Converted
0
1
count
not provided
india
russia
kuwait
oman
united arab emirates
united states
australia
united kingdom
bahrain
ghana
singapore
qatar
saudi arabia
belgium
france
sri lanka
china
canada
netherlands
sweden
nigeria
hong kong
germany
asia/pacific region
uganda
kenya
italy
south africa
tanzania
unknown
malaysia
liberia
switzerland
denmark
philippines
bangladesh
vietnam
indonesia
Country

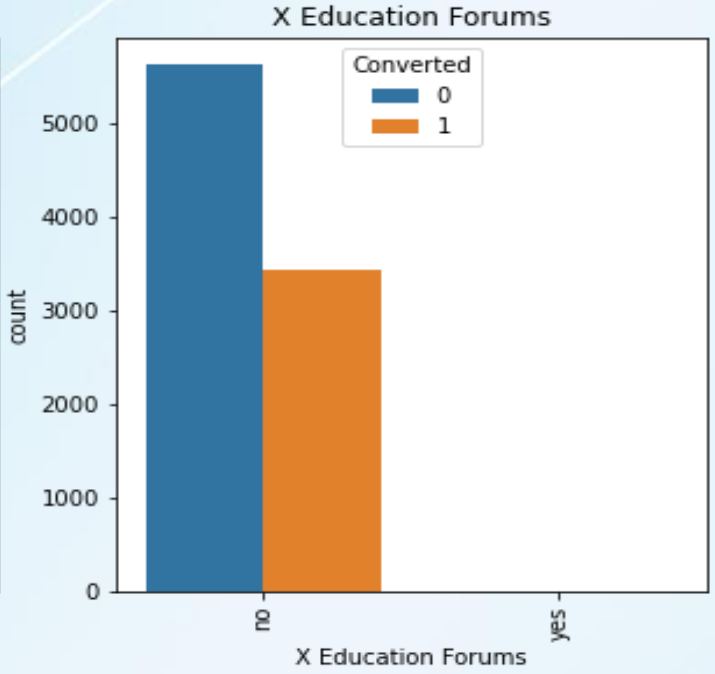

Specialization
Converted
0
1
count
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization

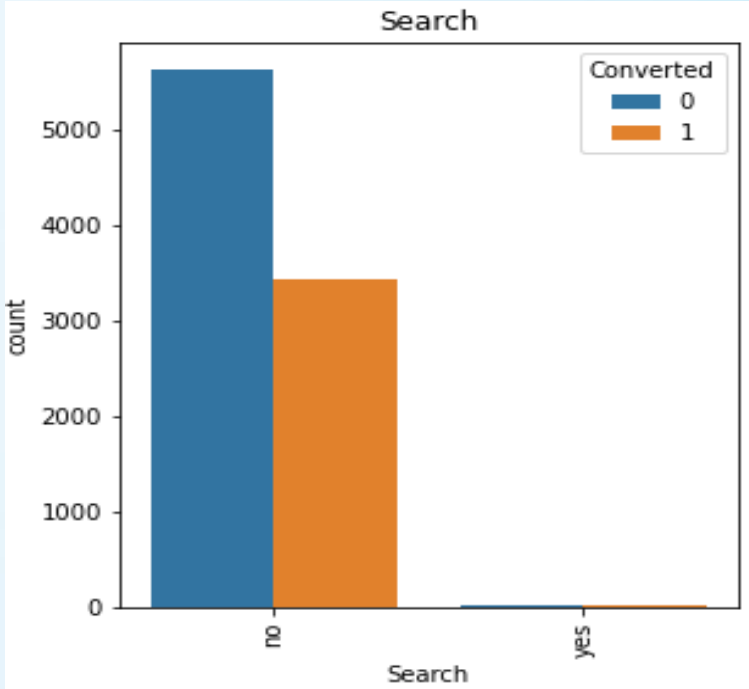

What is your current occupation
Converted
0
1
count
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation

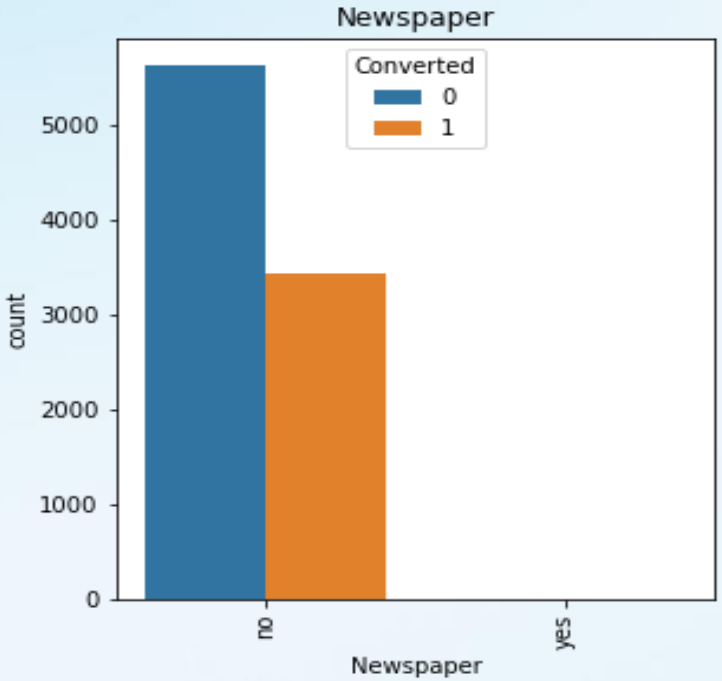

What matters most to you in choosing a course
Converted
0
1
count
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

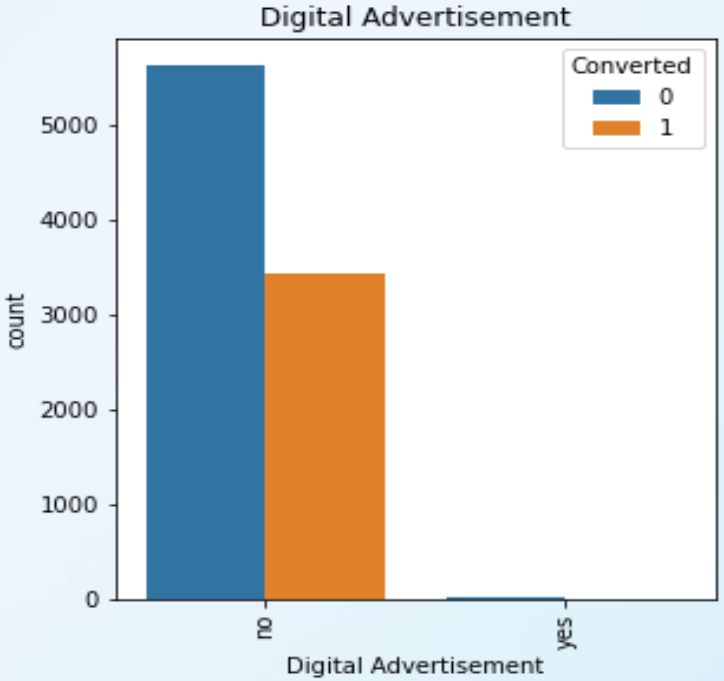

Do Not Call
Converted
0
1
count
no
yes
Do Not Call

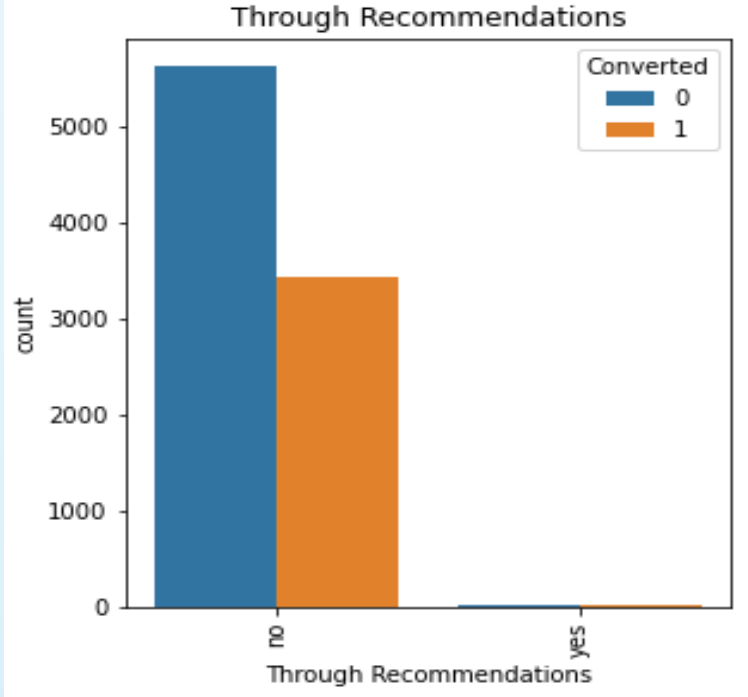

Newspaper Article
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Newspaper Article


X Education Forums
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
X Education Forums


Search
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Search


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


Through Recommendations
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Through Recommendations

Dummy variable creation

- Create dummy variables for the columns ‘Lead Origin’, ‘Specialization’ , ‘Lead Source’, ‘Do Not Email’, ‘Last Activity’, ‘What is your current occupation’, ‘A free copy of Mastering The Interview’, ‘Last Notable Activity’.
- Concatenate created dummy variables into master dataframe “Lead”
- Drop the columns for which dummy variables are created.
- Number of rows for analysis = 9074
- Number of columns for analysis = 81

Model Building

- The target variable “converted” is stored in y and else is stored in X.
- Split the dataset into 70% and 30% for train and test set respectively.
- Scale the numeric features using MinMax Scaler.
- Use RFE for feature selection.
- Running RFE with 15 variables as output.
- Put all the columns selected by RFE in the variable 'col’.
- Building model by removing the variable whose p-value is greater than 0.05 and vif value greater than 5.
- Predicting the probabilities on the train set.
- With the cutoff of 0.5 we obtained overall accuracy around 81%, sensitivity around 70% and specificity around 87% which is a very good value.

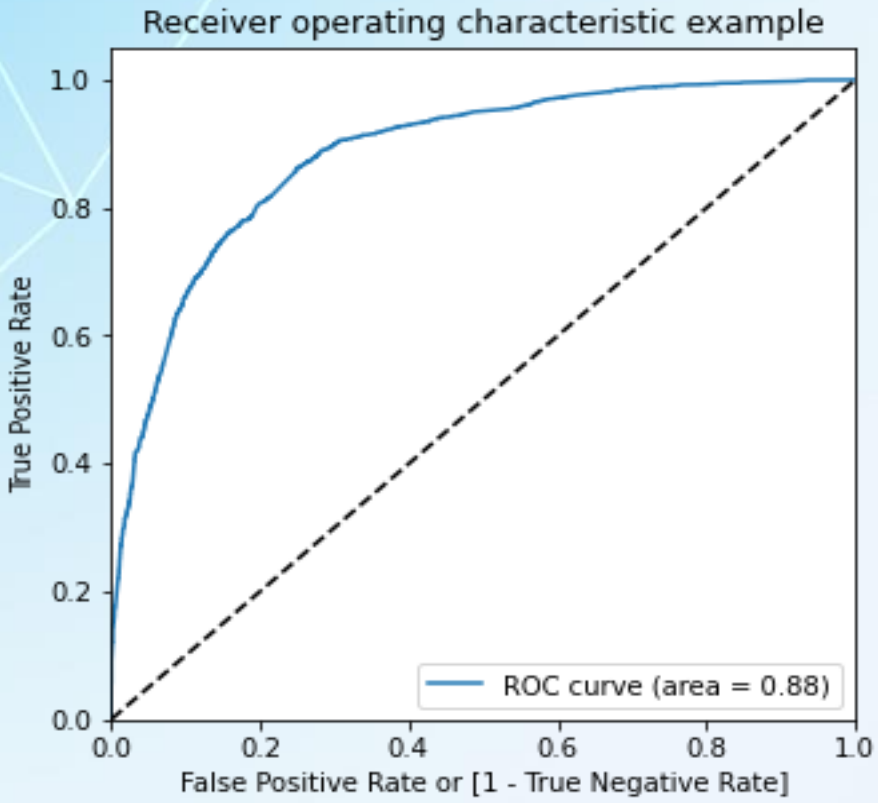
- The final model is given below:

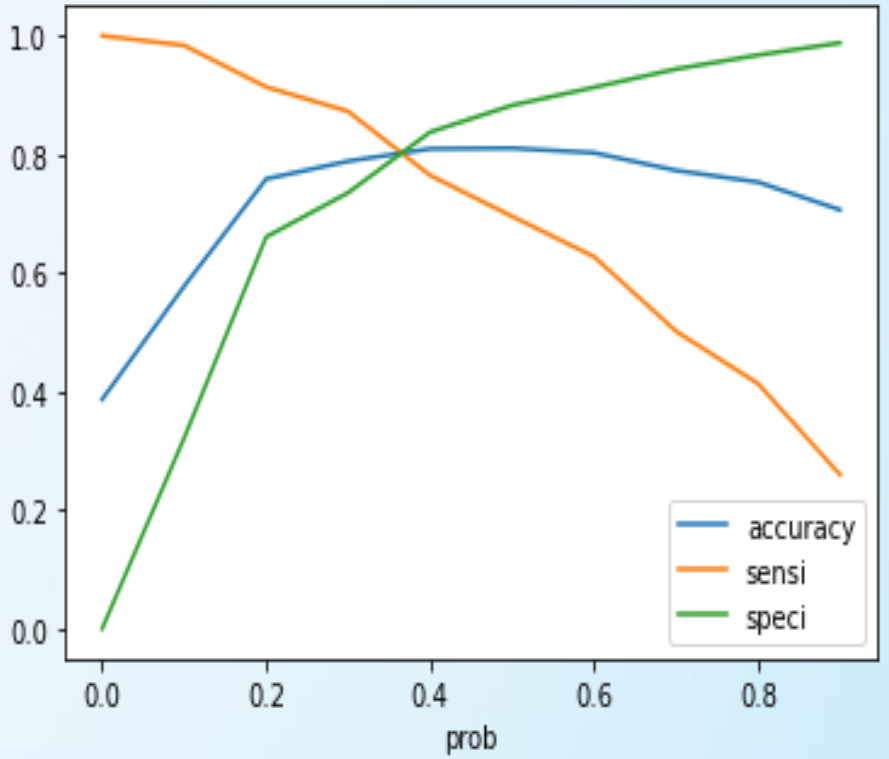
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2655.8
Date:	Thu, 11 Nov 2021	**Deviance:**	5311.7
Time:	02:01:33	**Pearson chi2:**	6.51e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4345	0.113	-30.511	0.000	-3.655	-3.214
TotalVisits	5.7276	1.459	3.926	0.000	2.868	8.587
Total Time Spent on Website	4.6142	0.166	27.753	0.000	4.288	4.940
Lead Origin_lead add form	3.7570	0.225	16.676	0.000	3.315	4.199
Lead Source_olark chat	1.5780	0.111	14.159	0.000	1.360	1.796
Lead Source_welingak website	2.5828	1.033	2.501	0.012	0.558	4.607
Do Not Email_yes	-1.4412	0.170	-8.470	0.000	-1.775	-1.108
Last Activity_olark chat conversation	-1.3929	0.167	-8.330	0.000	-1.721	-1.065
Last Activity_sms sent	1.2616	0.074	17.108	0.000	1.117	1.406
What is your current occupation_student	1.2218	0.226	5.401	0.000	0.778	1.665
What is your current occupation_unemployed	1.1394	0.085	13.408	0.000	0.973	1.306
What is your current occupation_working professional	3.6555	0.204	17.914	0.000	3.256	4.055
Last Notable Activity_unreachable	1.8066	0.601	3.008	0.003	0.629	2.984

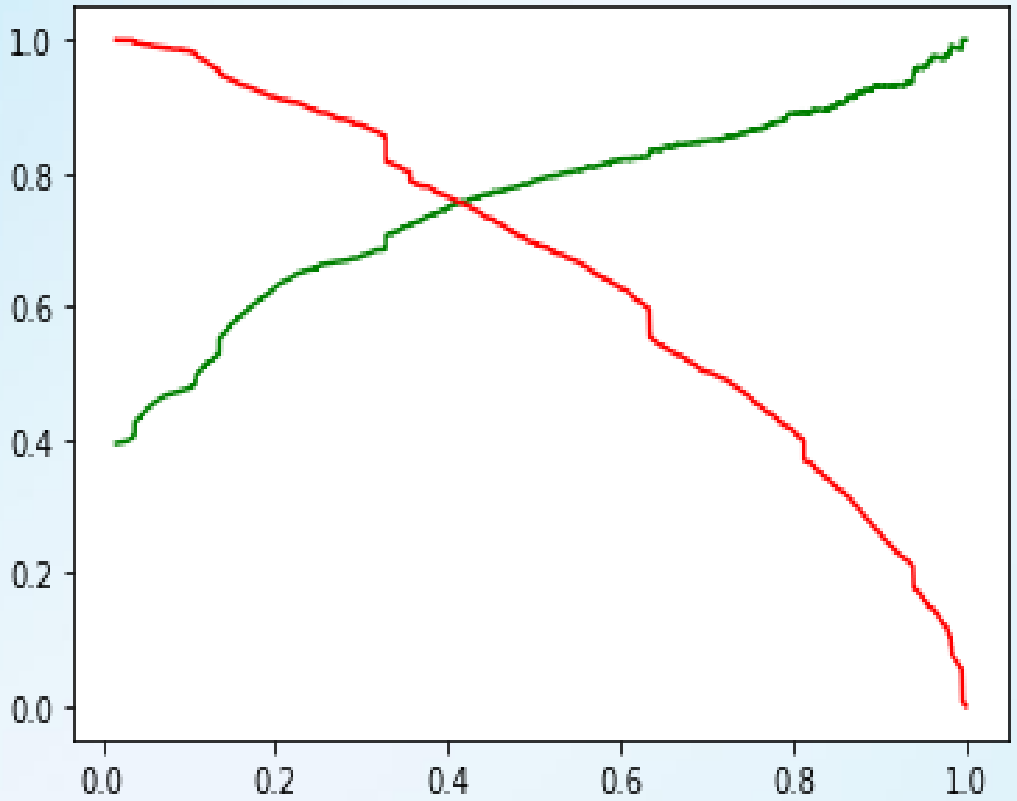
ROC Curve





- Area under ROC Curve is 0.88, which is good.
- The optimal cut-off point is 0.35.
- With the optimal cut-off of 0.35 we have accuracy, sensitivity and specificity of both train data and test data is around 80%.

Precision – Recall view



- The cut-off obtained is 0.41.
- We have precision and recall around 75.4% and 75.8% respectively.

Conclusion

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
  - Total number of visits.
  - The total time spend on the Website.
  - When the lead origin is Lead add form.
  - When their current occupation is as a working professional, unemployed and student.
  - When the lead source was Olark chat and Welingak website.
  - When the last activity was SMS and Olark chat conversation.
  - When the last notable activity is unreachable.
- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.