LEAD SCORING ASSIGNMENT

BY: JANHAVI BHAVE

MANIKANTA SAI

PROBLEM STATEMENT

- The EdTech company named X education provides online courses. It currently has only 30% lead conversion rate. This means that out of 100 leads obtained, only 30 leads get converted into potential customers.
- The problem is to find groups that can turn into potential customers i.e. hot leads.
- By identifying such leads, the company can efficiently increase their customers and hence be a profitable company.

BUSINESS OBJECTIVE

- To find Hot leads or most promising leads.
- Build a regression model to identify such leads.
- Build a model that can be used for future use also.

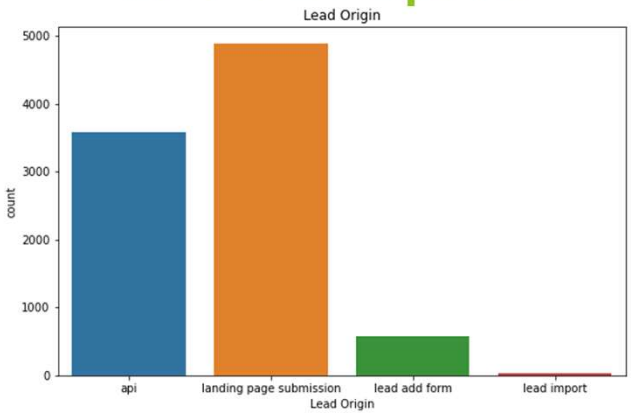
Method to solve the problem

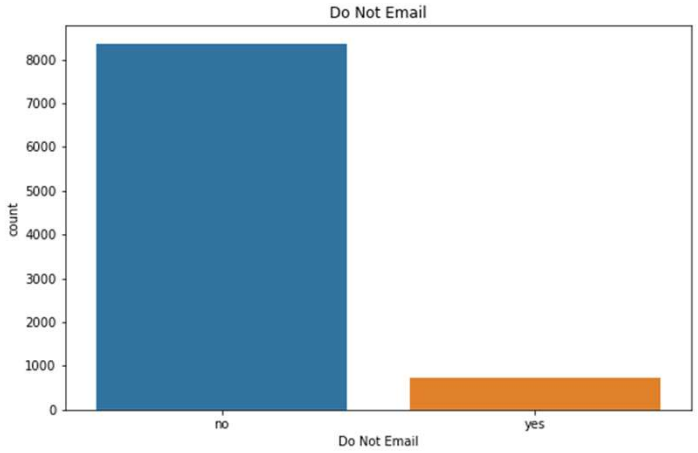
- Data cleaning and manipulation
  1. handling duplicate data
  2. handling null and missing values
  3. drop the columns that are not useful for analysis
  4. imputation of values
  5. handling outliers
- EDA: Univariate and Bivariate analysis
- Scaling and creation of dummy variables
- Building a regression model using suitable classification technique
- Validation of the model
- Model Presentation
- Conclusion from the model

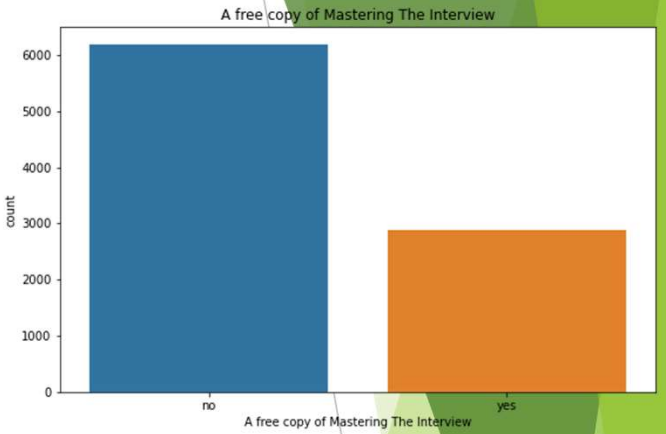
Data Manipulation

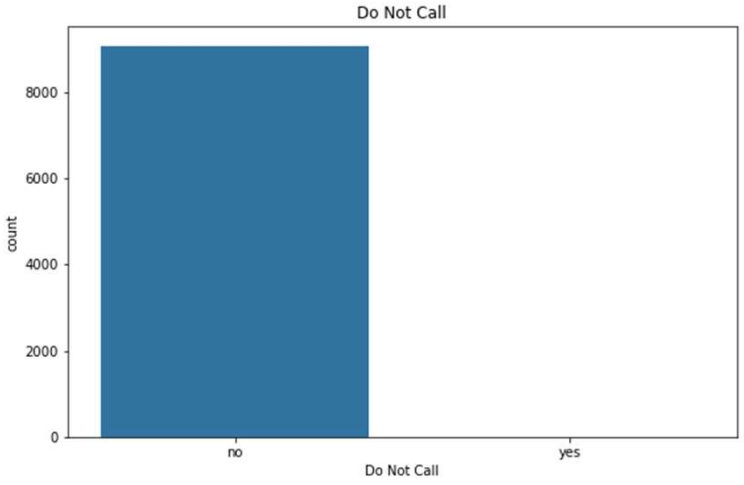
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

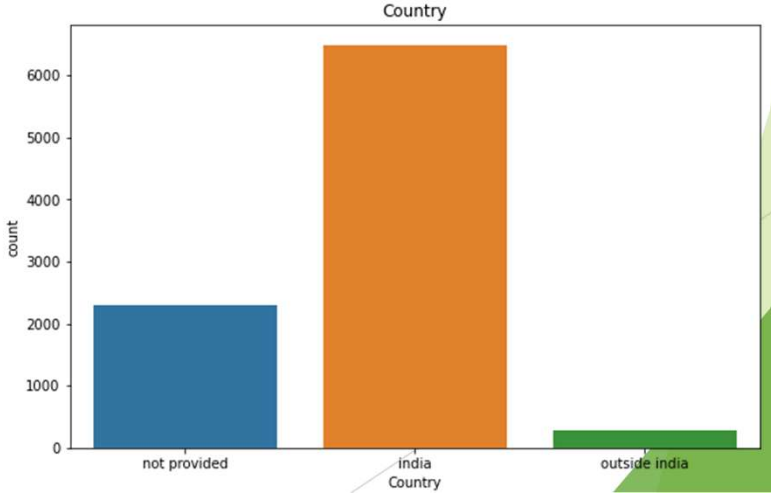
EDA Graphs

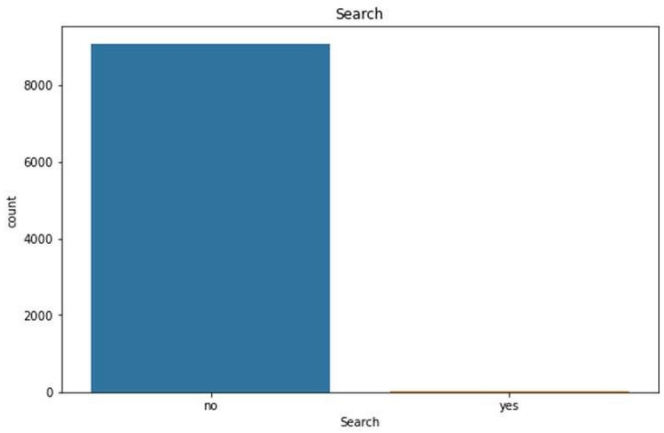

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

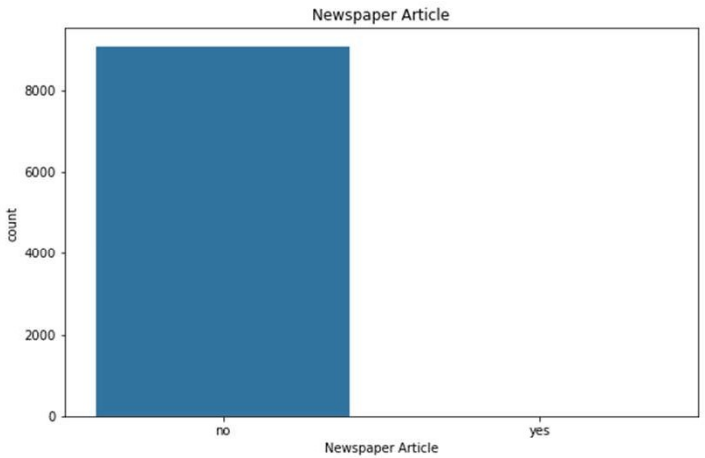

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

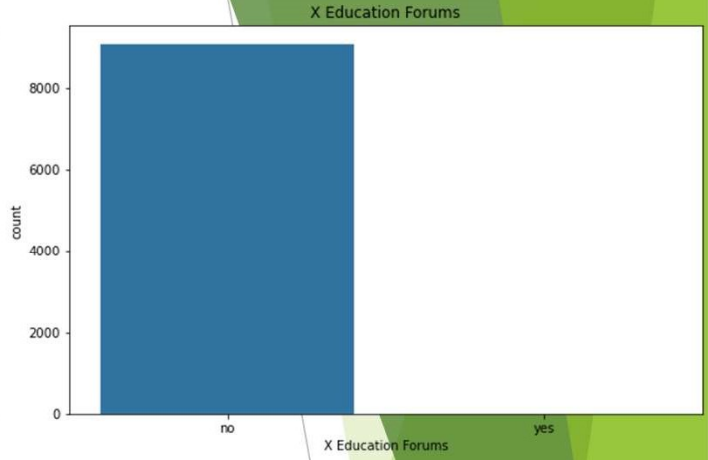

A free copy of Mastering The Interview
count
6000
5000
4000
3000
2000
1000
0
no
yes
A free copy of Mastering The Interview

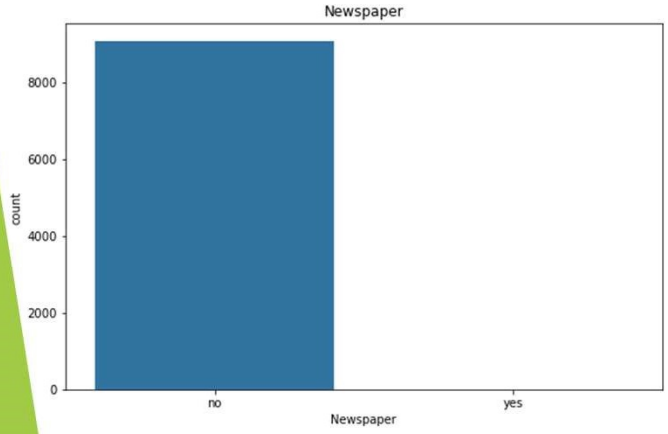

Do Not Call
count
8000
6000
4000
2000
0
no
yes
Do Not Call

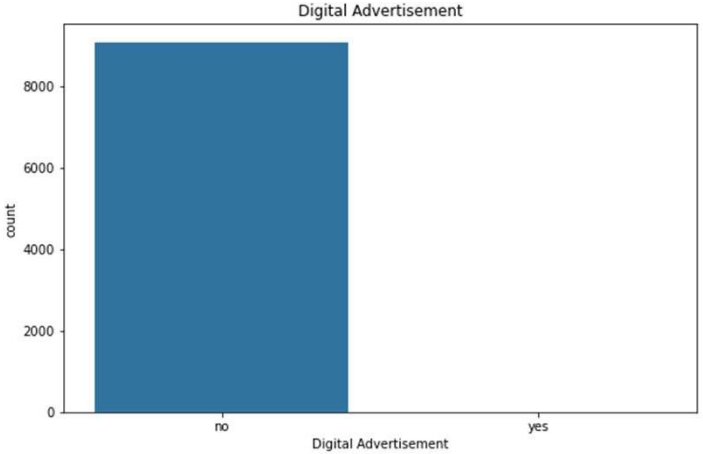

Country
count
6000
5000
4000
3000
2000
1000
0
not provided
india
outside india
Country

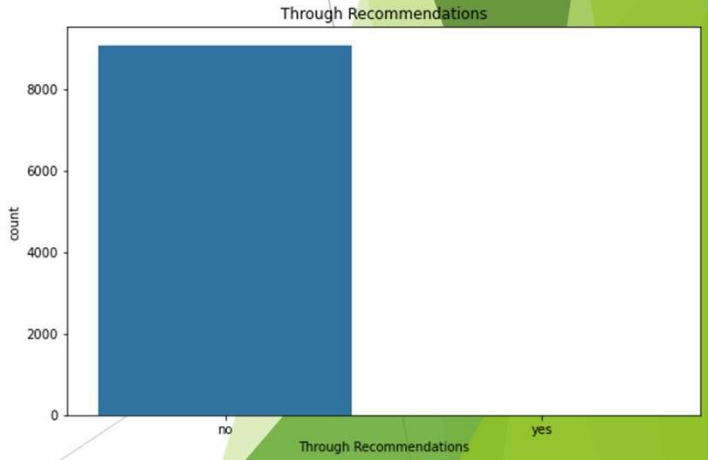

Search
count
8000
6000
4000
2000
0
no
yes
Search

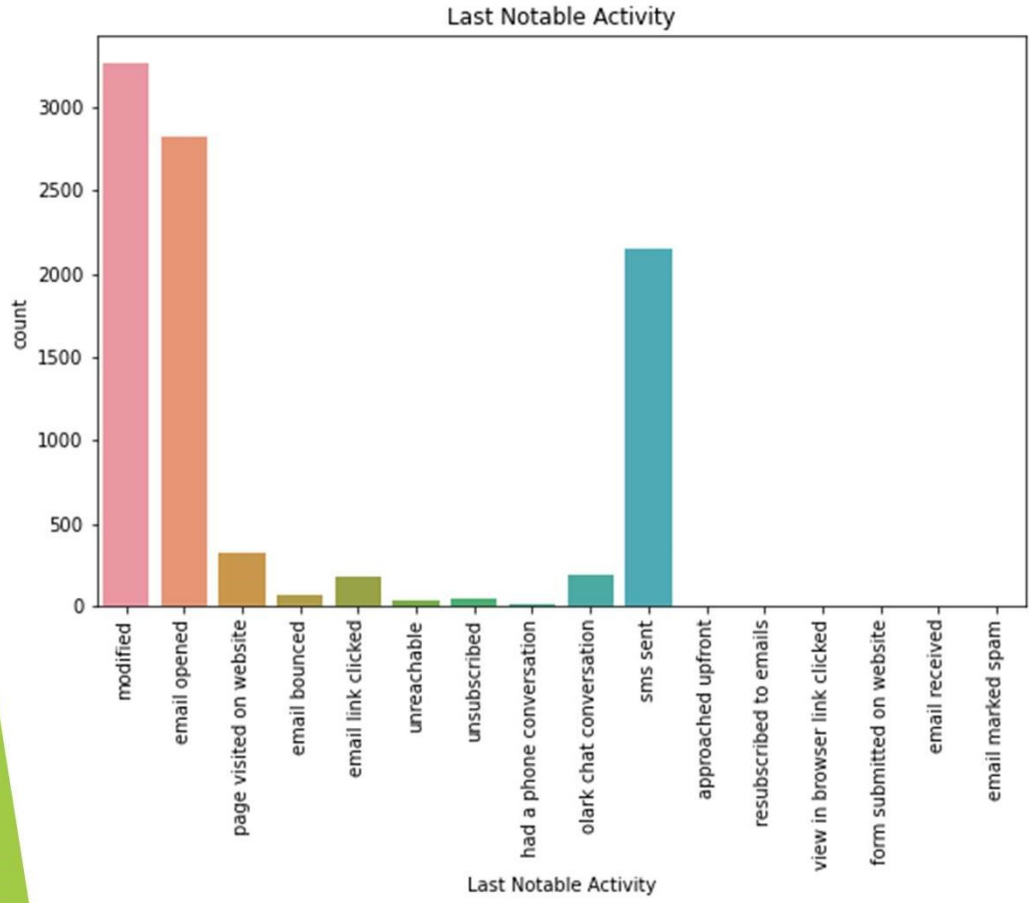

Newspaper Article
count
8000
6000
4000
2000
0
no
yes
Newspaper Article

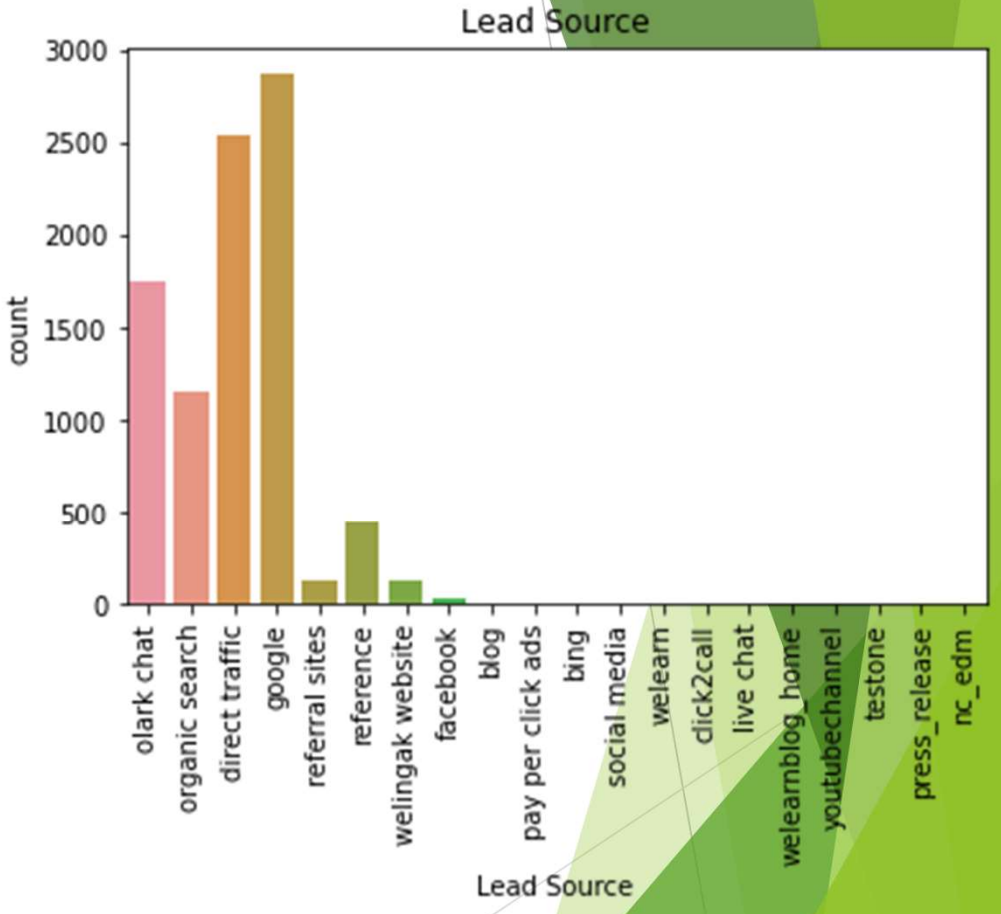

X Education Forums
count
8000
6000
4000
2000
0
no
yes
X Education Forums

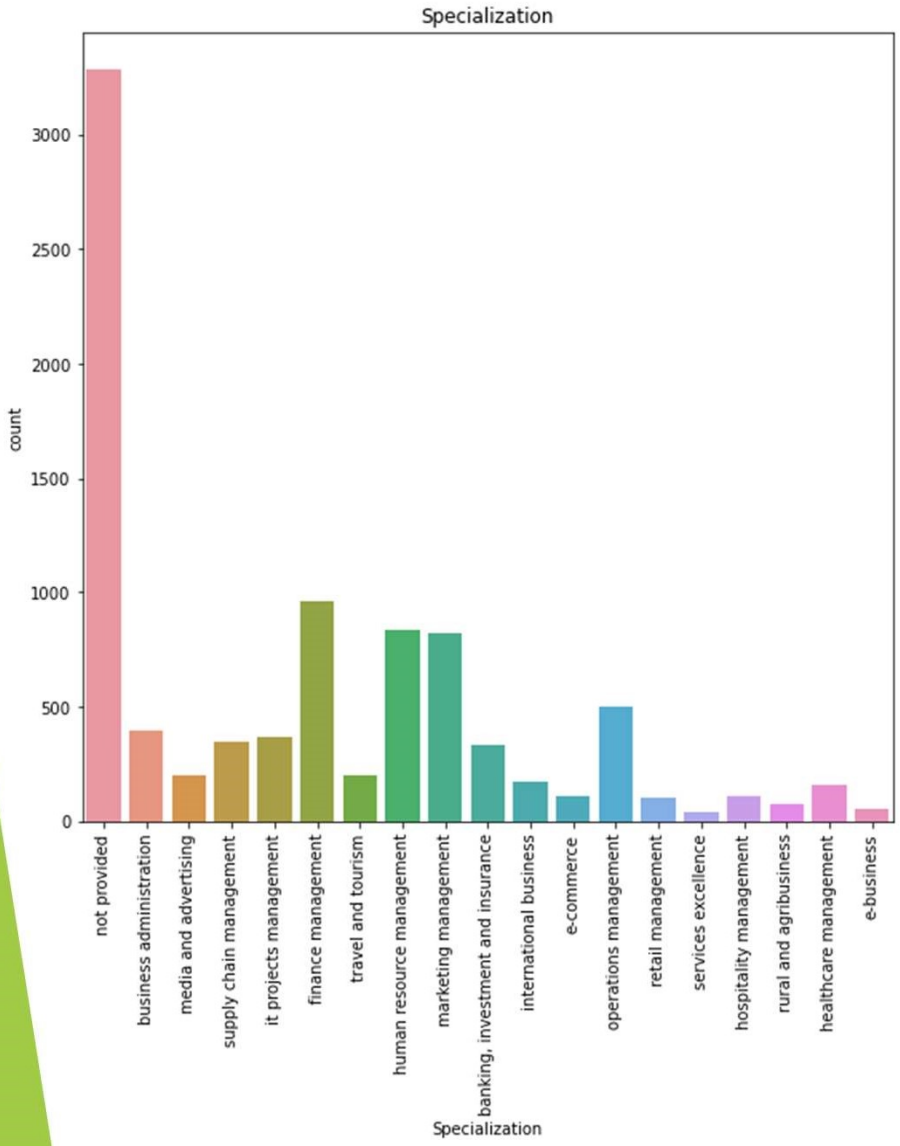

Newspaper
count
8000
6000
4000
2000
0
no
yes
Newspaper

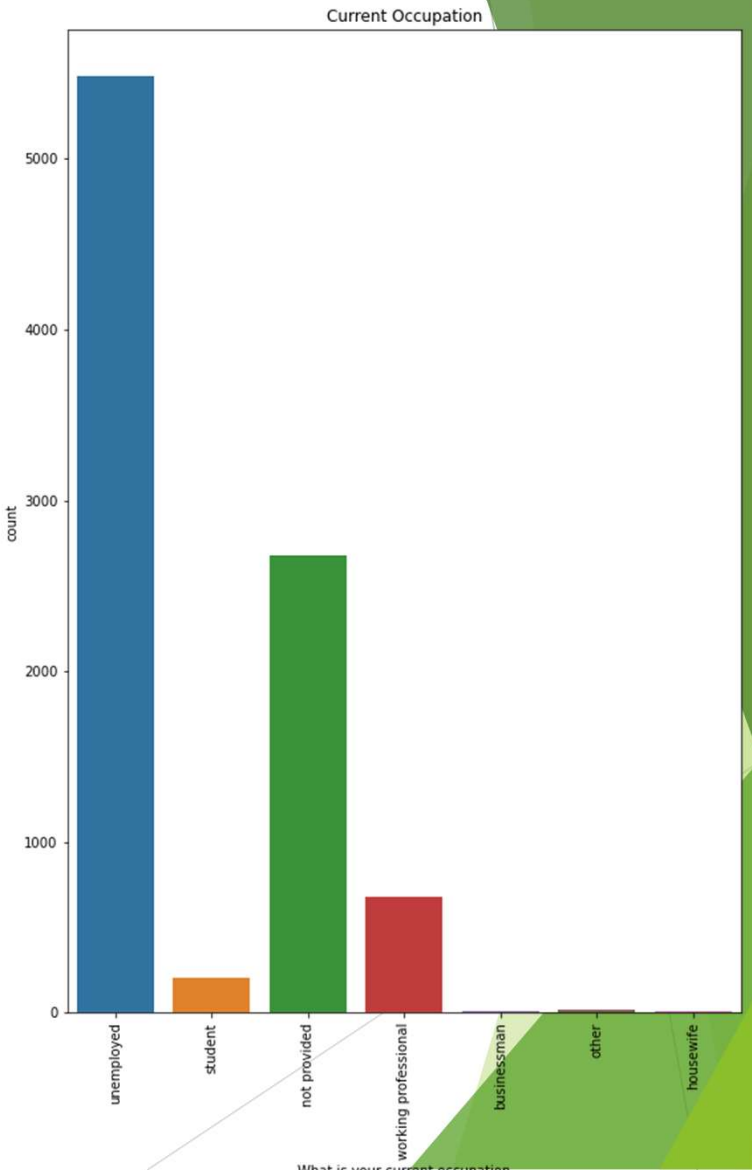

Digital Advertisement
count
8000
6000
4000
2000
0
no
yes
Digital Advertisement

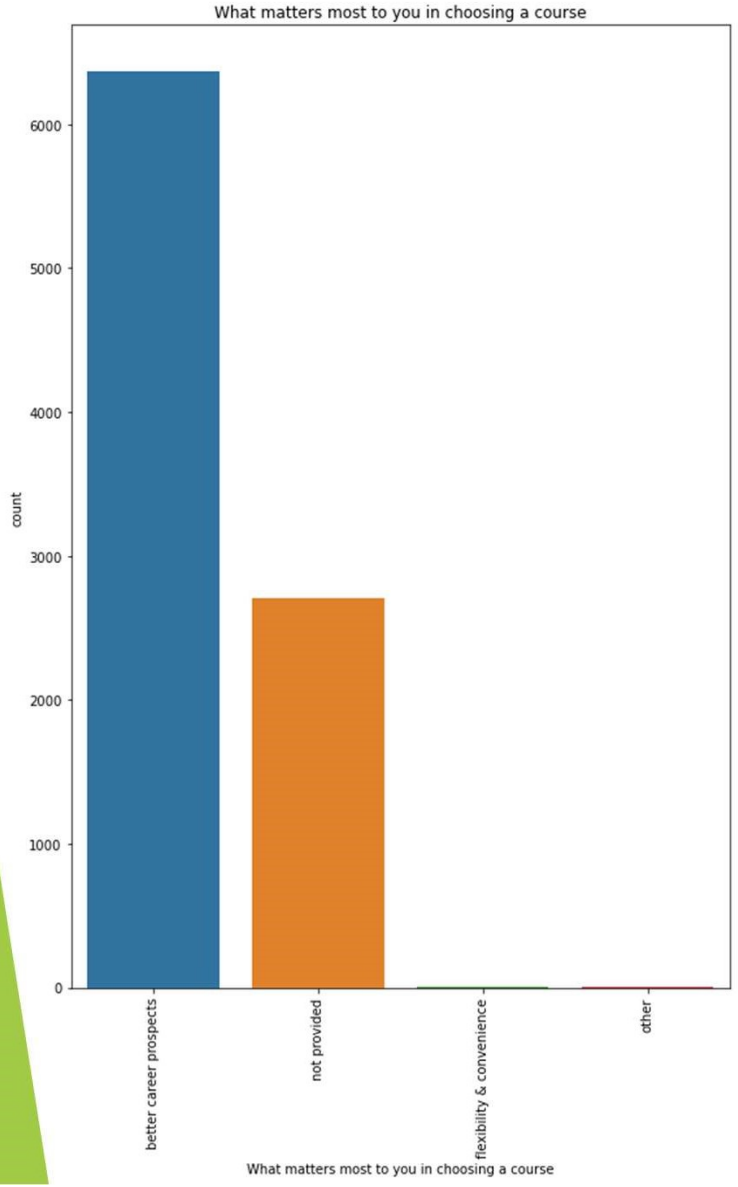

Through Recommendations
count
8000
6000
4000
2000
0
no
yes
Through Recommendations

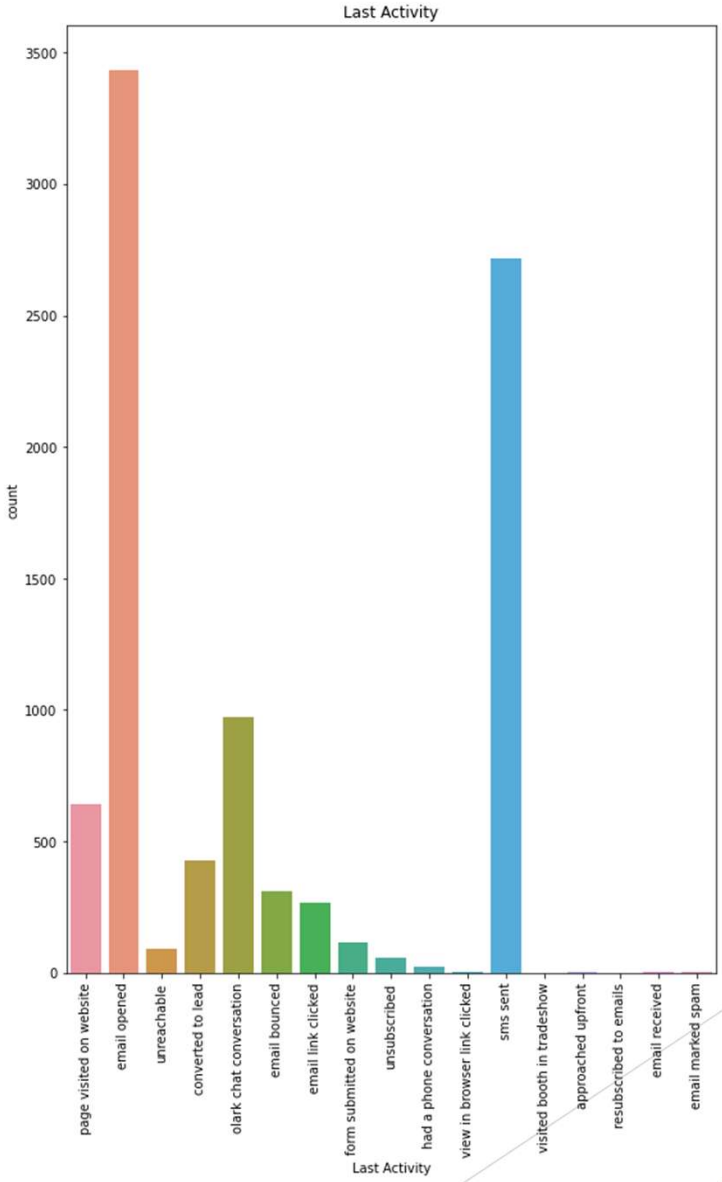

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted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

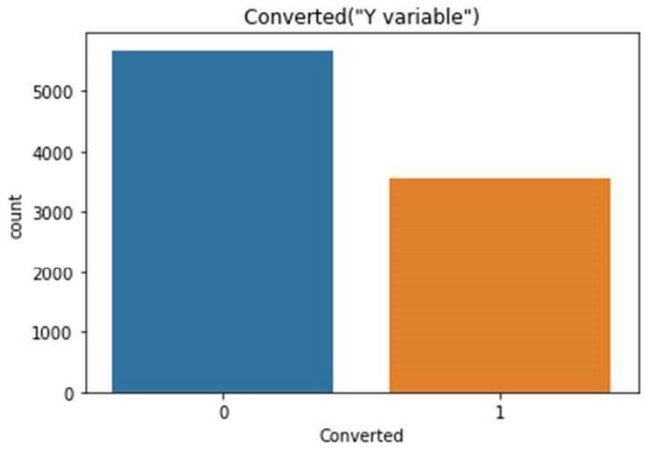

Lead Source
count
3000
2500
2000
1500
1000
500
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

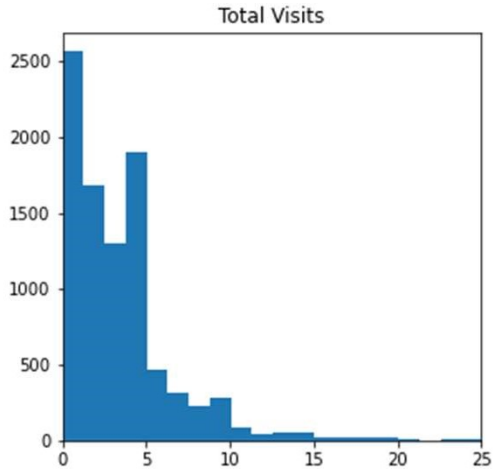

Specialization
count
3000
2500
2000
1500
1000
500
0
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization

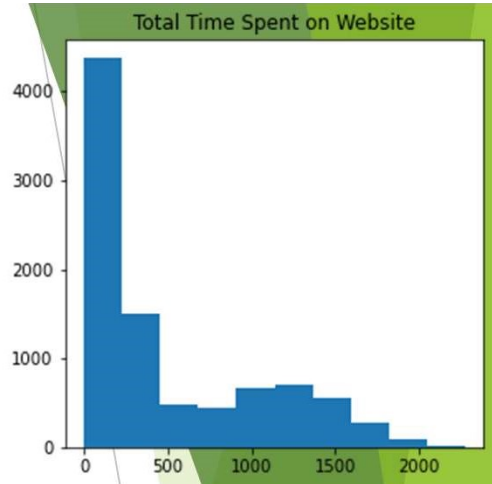

Current Occupation
count
5000
4000
3000
2000
1000
0
unemployed
student
not provided
working professional
businessman
other
housewife

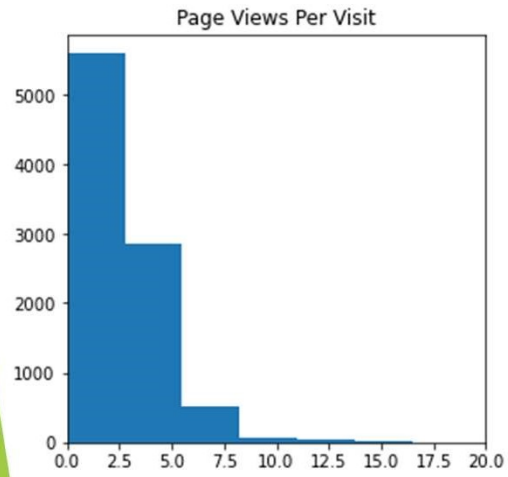

What matters most to you in choosing a course
count
6000
5000
4000
3000
2000
1000
0
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

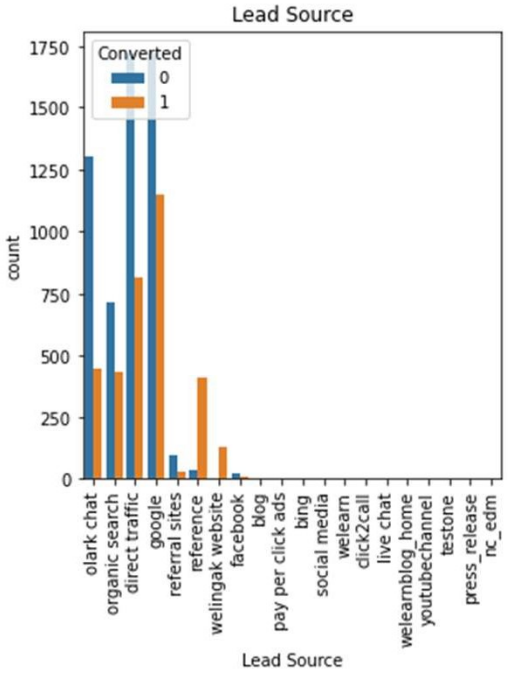

Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow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

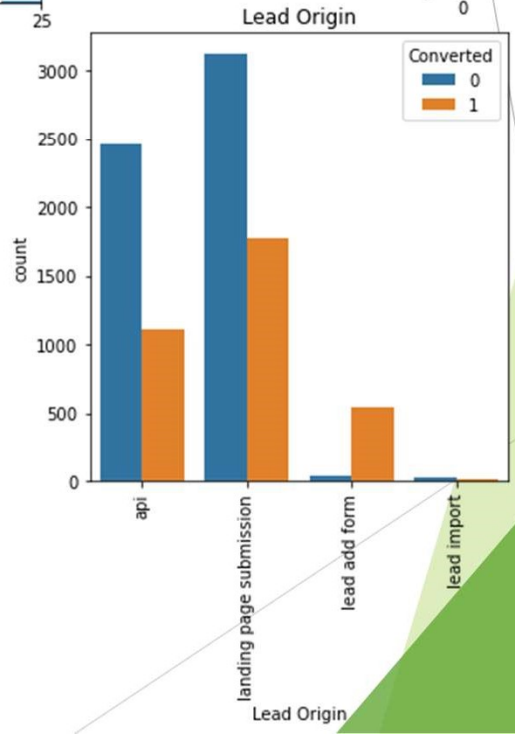

Converted("Y variable")
count
Converted
0
1

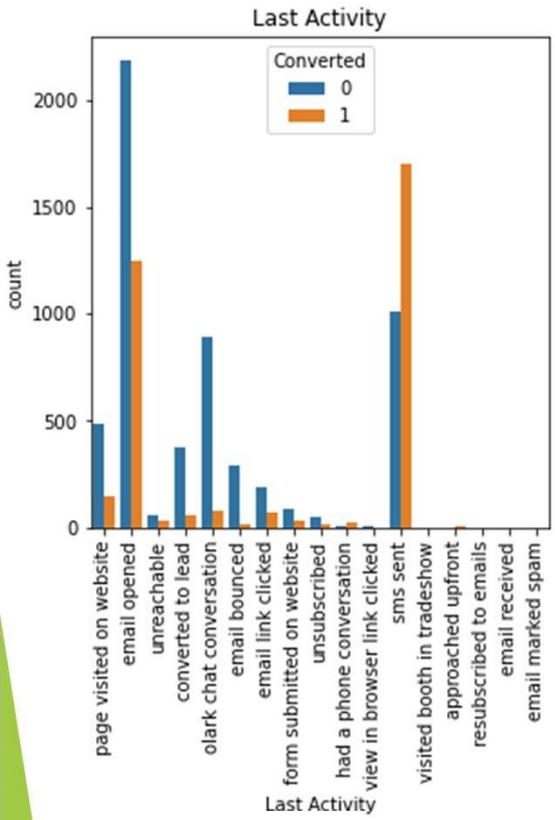

Total Visits

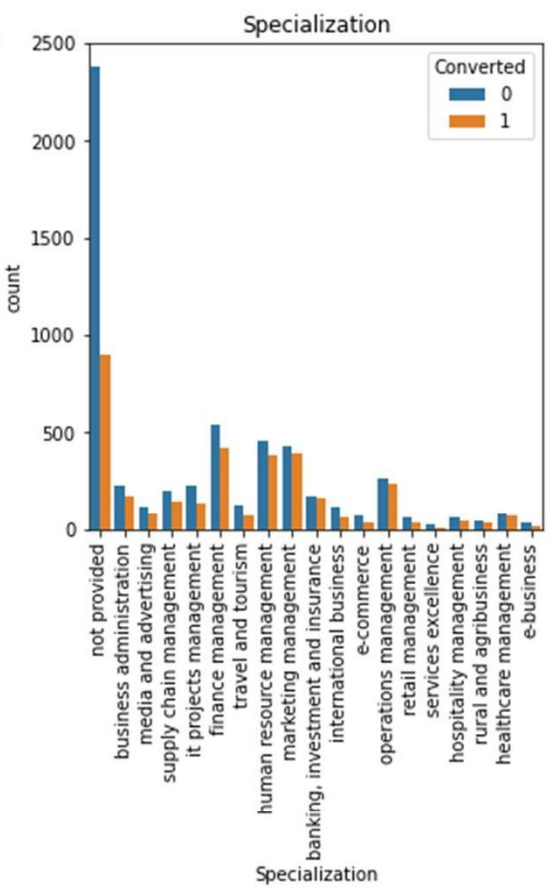

Total Time Spent on Website

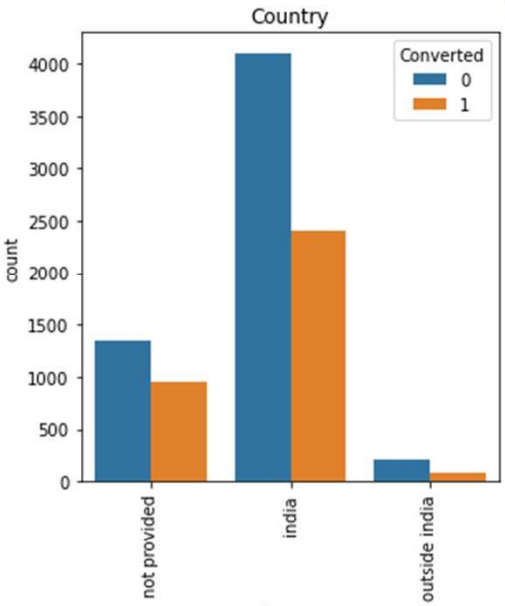

Page Views Per Visit

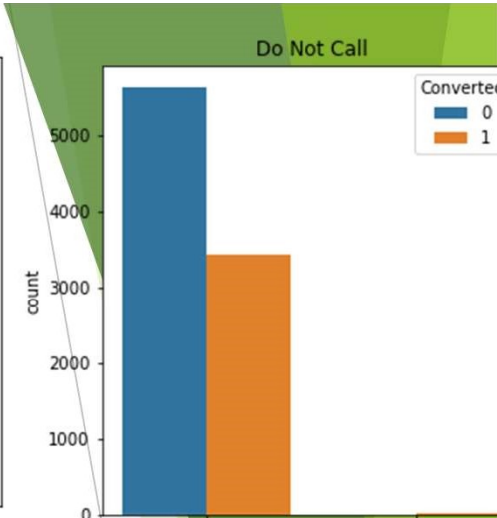

Lead Source
Converted
0
1
count
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

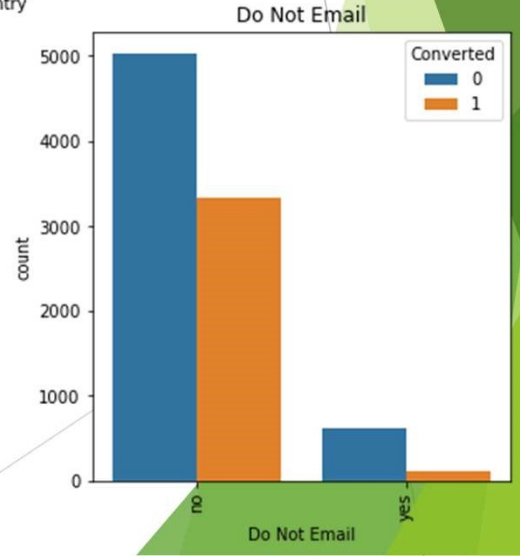

Lead Origin
Converted
0
1
count
api
landing page submission
lead add form
lead import
Lead Origin

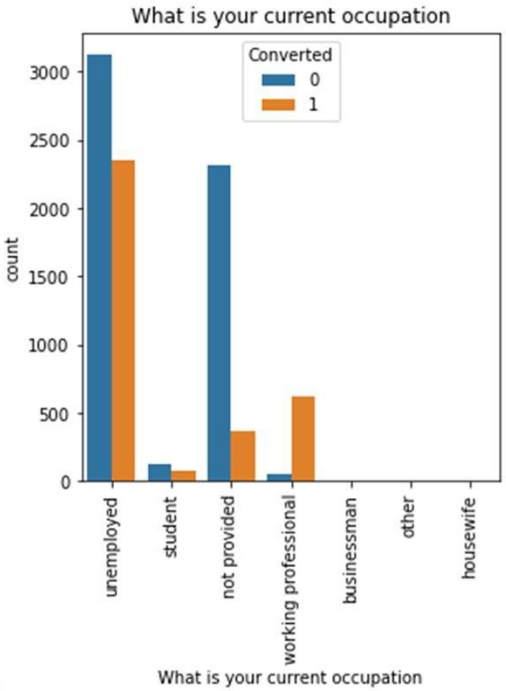

Last Activity
Converted
0
1
count
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

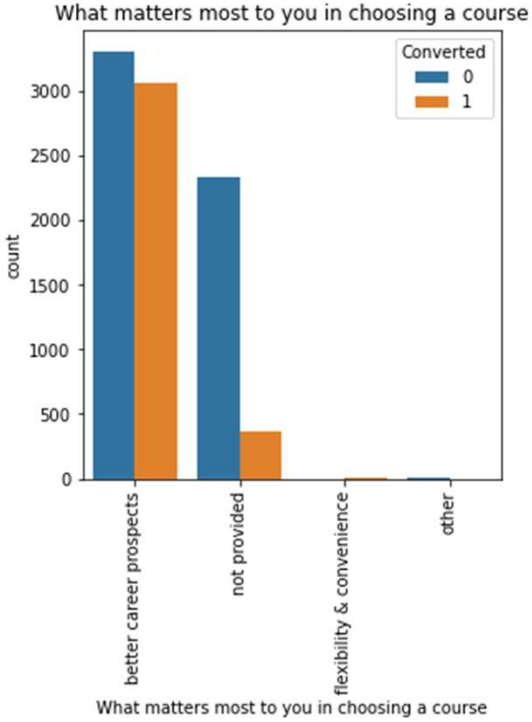

Specialization
Converted
0
1
count
2500
2000
1500
1000
500
0
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization

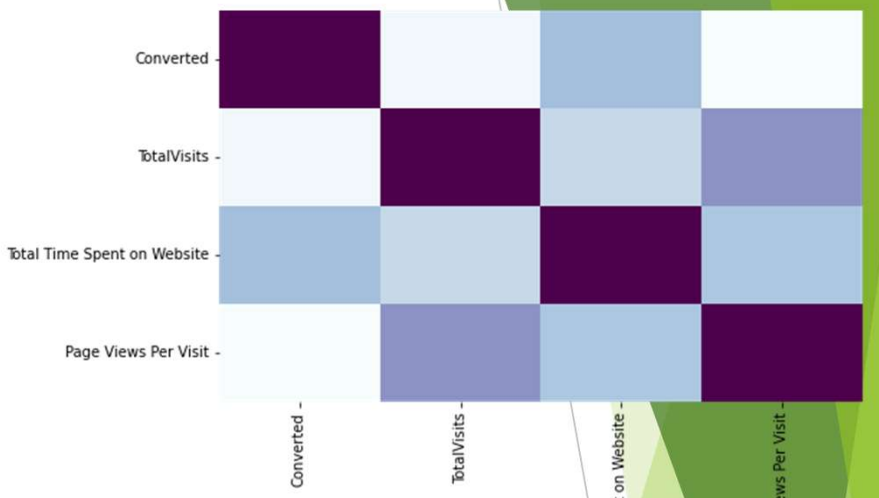

Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

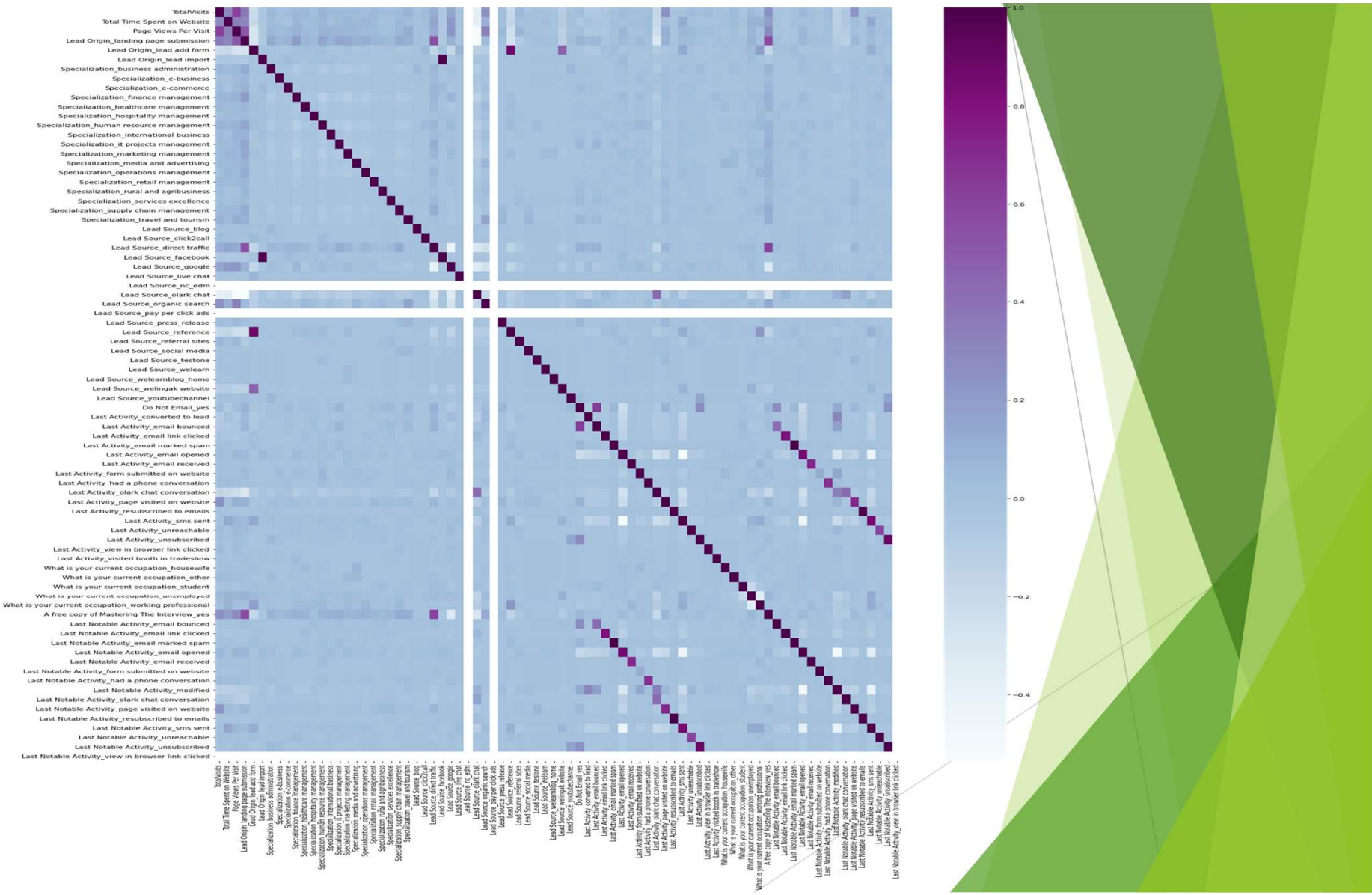

Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Call


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


What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
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
3000
2500
2000
1500
1000
500
0
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

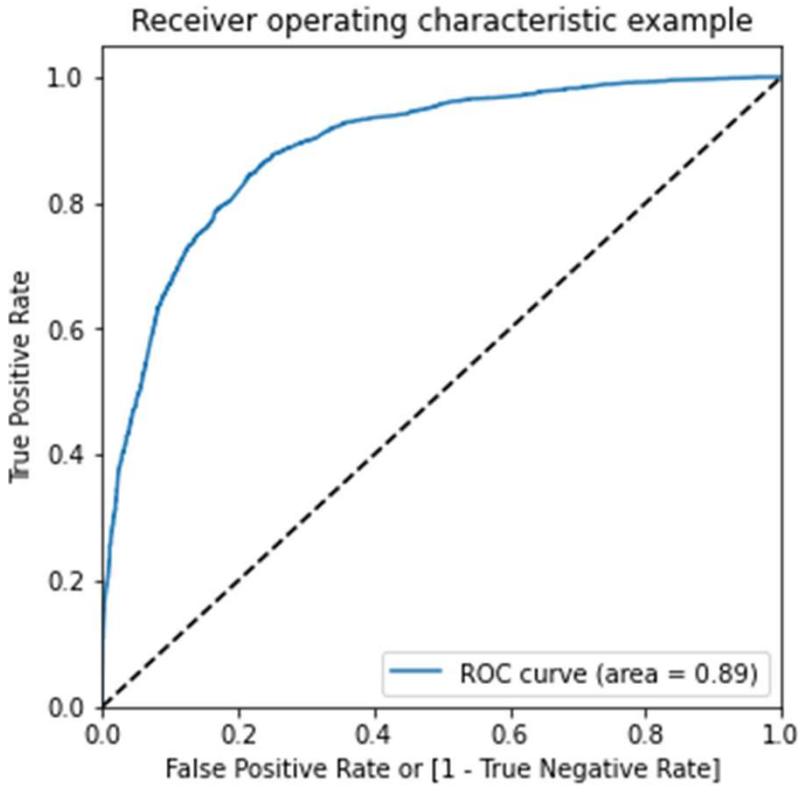

Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2

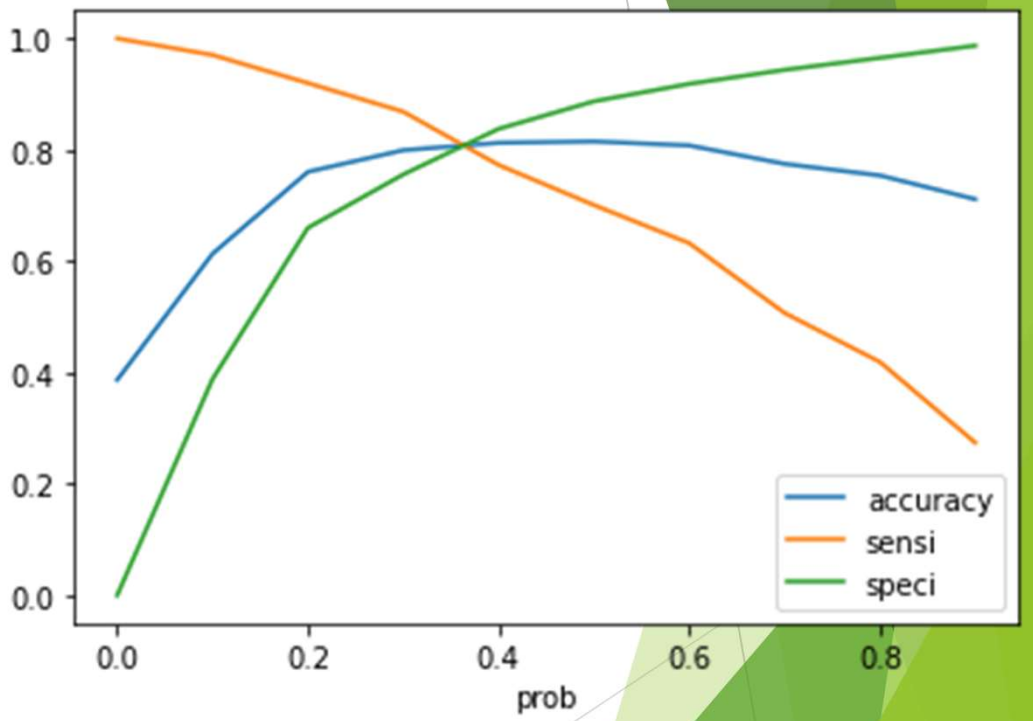

TotalVisits
Total Time Spent on Website
Page Views Per Visit
Lead Origin_landing page submission
Lead Origin_lead add form
Lead Origin_lead import
Specialization_business administration
Specialization_e-business
Specialization_e-commerce
Specialization_finance management
Specialization_healthcare management
Specialization_hospitality management
Specialization_human resource management
Specialization_international business
Specialization_it projects management
Specialization_marketing management
Specialization_media and advertising
Specialization_operations management
Specialization_retail management
Specialization_rural and agribusiness
Specialization_services excellence
Specialization_supply chain management
Specialization_travel and tourism
Lead Source_blog
Lead Source_click2call
Lead Source_direct traffic
Lead Source_facebook
Lead Source_google
Lead Source_live chat
Lead Source_nc_edm
Lead Source_olark chat
Lead Source_organic search
Lead Source_pay per click ads
Lead Source_press_release
Lead Source_reference
Lead Source_referral sites
Lead Source_social media
Lead Source_testone
Lead Source_welearn
Lead Source_welearnblog_home
Lead Source_welingak website
Lead Source_youtubechannel
Do Not Email_yes
Last Activity_converted to lead
Last Activity_email bounced
Last Activity_email link clicked
Last Activity_email marked spam
Last Activity_email opened
Last Activity_email received
Last Activity_form submitted on website
Last Activity_had a phone conversation
Last Activity_olark chat conversation
Last Activity_page visited on website
Last Activity_resubscribed to emails
Last Activity_sms sent
Last Activity_unreachable
Last Activity_unsubscribed
Last Activity_view in browser link clicked
Last Activity_visited booth in tradeshow
What is your current occupation_housewife
What is your current occupation_other
What is your current occupation_student
What is your current occupation_unemployed
What is your current occupation_working professional
A free copy of Mastering The Interview_yes
Last Notable Activity_email bounced
Last Notable Activity_email link clicked
Last Notable Activity_email marked spam
Last Notable Activity_email opened
Last Notable Activity_email received
Last Notable Activity_form submitted on website
Last Notable Activity_had a phone conversation
Last Notable Activity_modified
Last Notable Activity_olark chat conversation
Last Notable Activity_page visited on website
Last Notable Activity_resubscribed to emails
Last Notable Activity_sms sent
Last Notable Activity_unreachable
Last Notable Activity_unsubscribed
Last Notable Activity_view in browser link clicked
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

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 than 5
- Predictions on test data set
- Overall accuracy 81%

ROC


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

CONCLUSION

- #It was found that the variables that mattered the most in the potential buyers are (In descending order) :
- # 1.The total time spend on the Website.
- # 2.Total number of visits.
- # 3. When the lead source was:
- #a. Google
- #b. Direct traffic
- #c. Organic search
- #d. Welingak website
- # 4. When the last activity was:
- #a. SMS
- #b. Olark chat conversation
- # 5. When the lead origin is Lead add format. 6. When their current occupation is as a working professional.
- #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.