ML – Case Study

Group Members :

- Mamta
- Sneha

Abstract

Business Understanding:
An education company ‘X’ markets its courses on several websites and search engines like Google. 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. The company also gets leads through past referrals. Employees from the sales team start making calls, writing emails, etc. Trying to convert leads . The typical lead conversion rate at X education is around 30%.

Business Object:
This case study aims to identify the ‘Hot Leads’ - the most potential leads , i.e. the leads that are most likely to convert into paying customers. And to build a model 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.

Goals of Data Analysis:
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.

Data Understanding

1. ***'Leads.csv'*** consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

It contains 9240 rows, 37 columns.
There are so many missing values.

Approach

1) EDA on the dataset. Read the data, find missing values and do the Outlier Analysis.

2) Build the model using Logistic regression.

3) Select the 13 features using RFE.

4) Check for the correction using VIF. Drop the highly correlated column rerun the model again.

5) Find the conversion probability . create the confusion matrics . Check the accuracy. Check the ROC curve.

6)Test the model on the test data and check the accuracy score

Rules followed for data cleaning¶

Dropping rows and columns

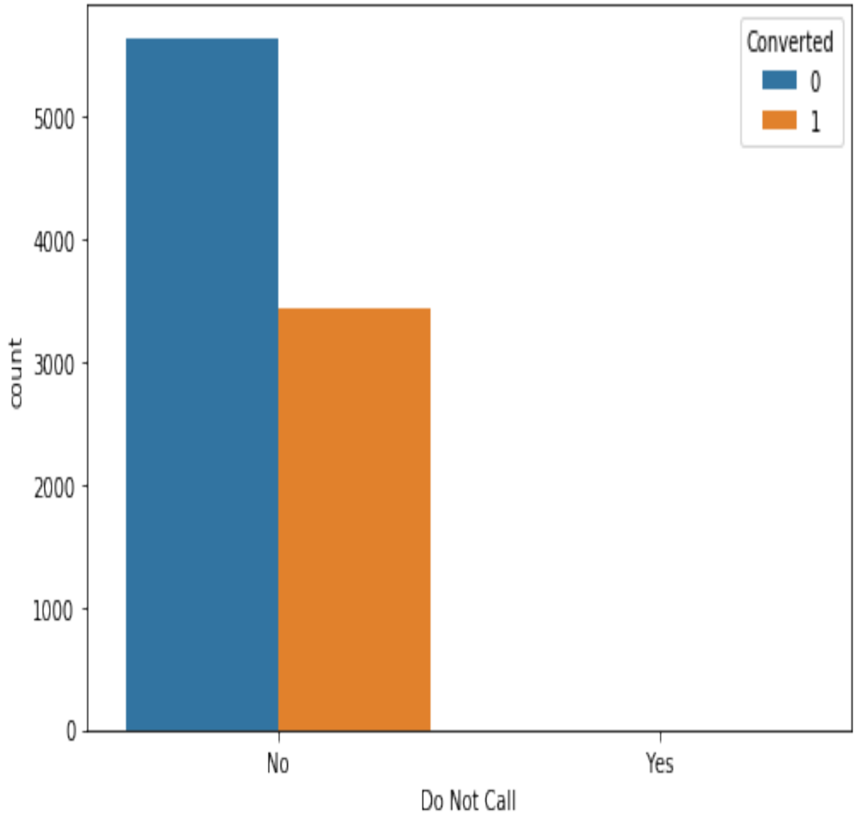
- •Drop unnecessary columns - ‘Prospect ID’ : A unique ID with which the customer is identified.
- •Drop Columns with more then 30 % null values , Keep only the columns which make business sense.
- •Drop columns which has less then 1 % valid values like yes and No.
- •For columns with missing values < 2%, deleted the rows as this would not cause much data loss.

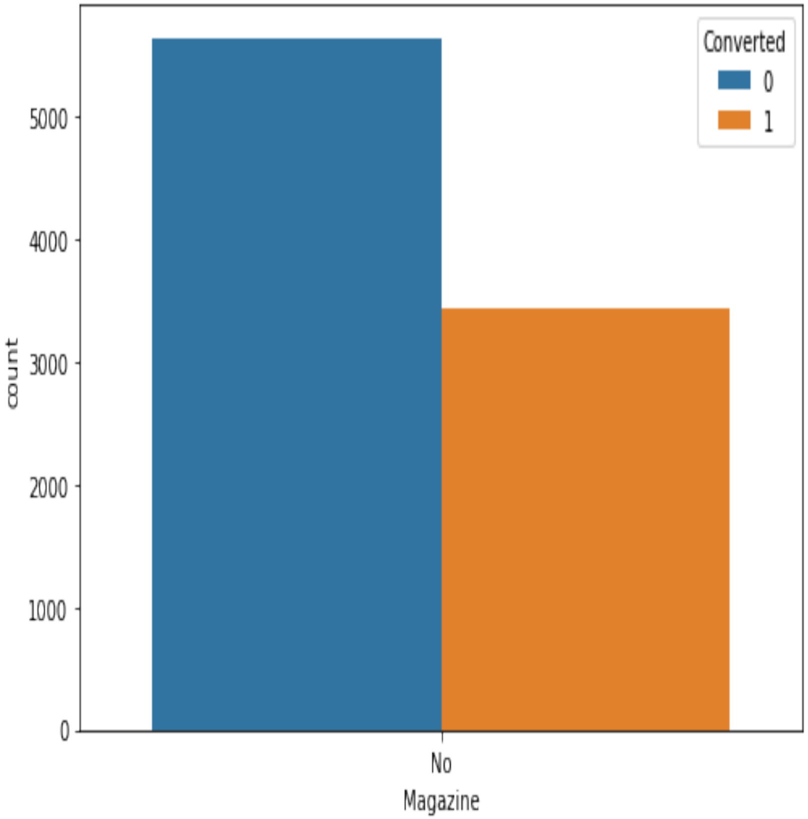
Imputing Data

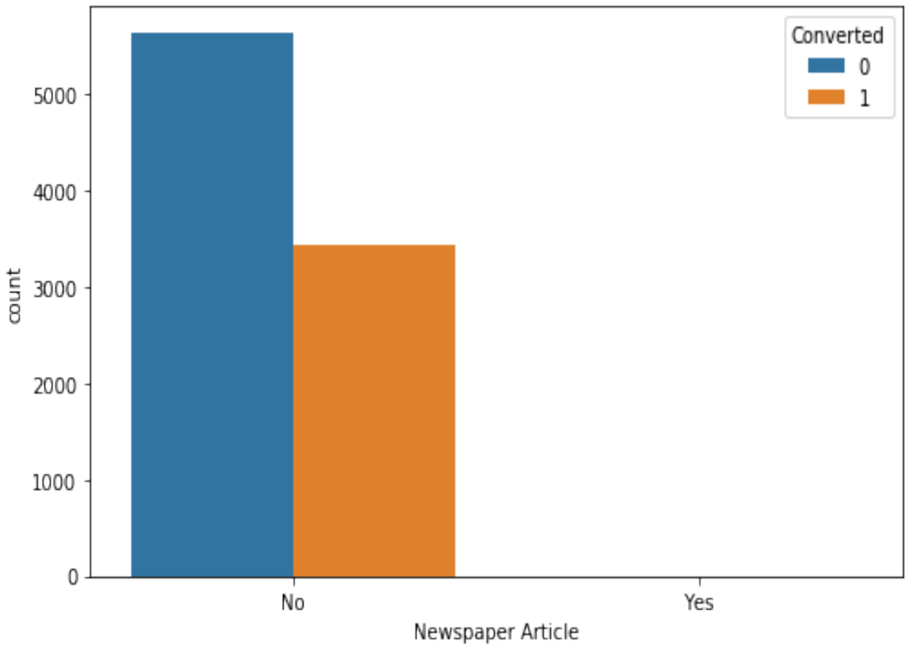
for categorical data

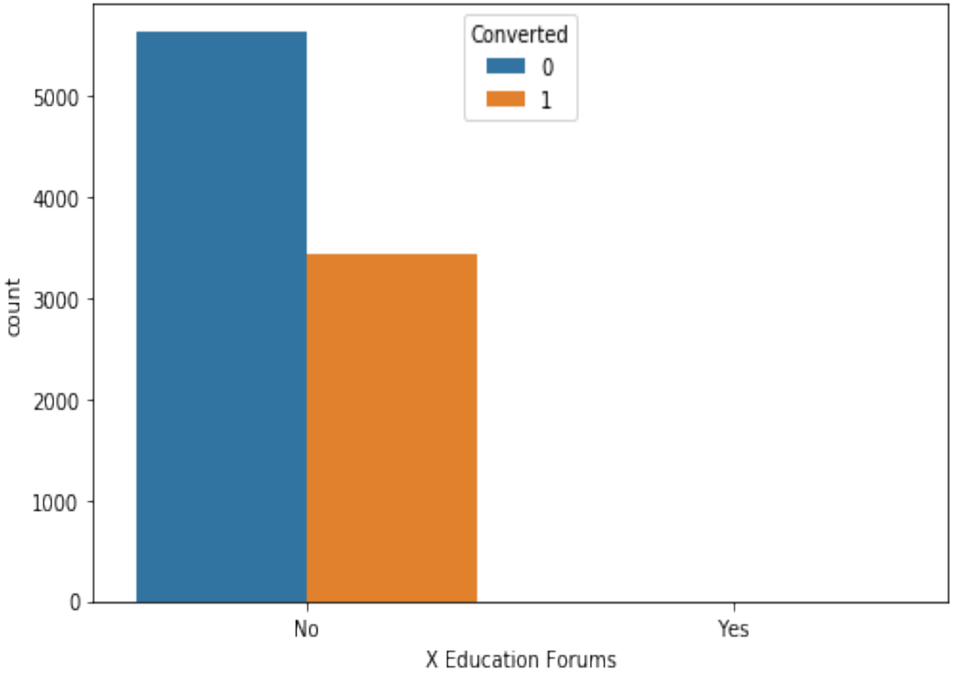
- •For uniformly distributed values used random() function to impute missing values.
- •Used mode where ever applicable like city /country column to impute missing values.
- •Combined the category holding less then 10 percent values into on category- others

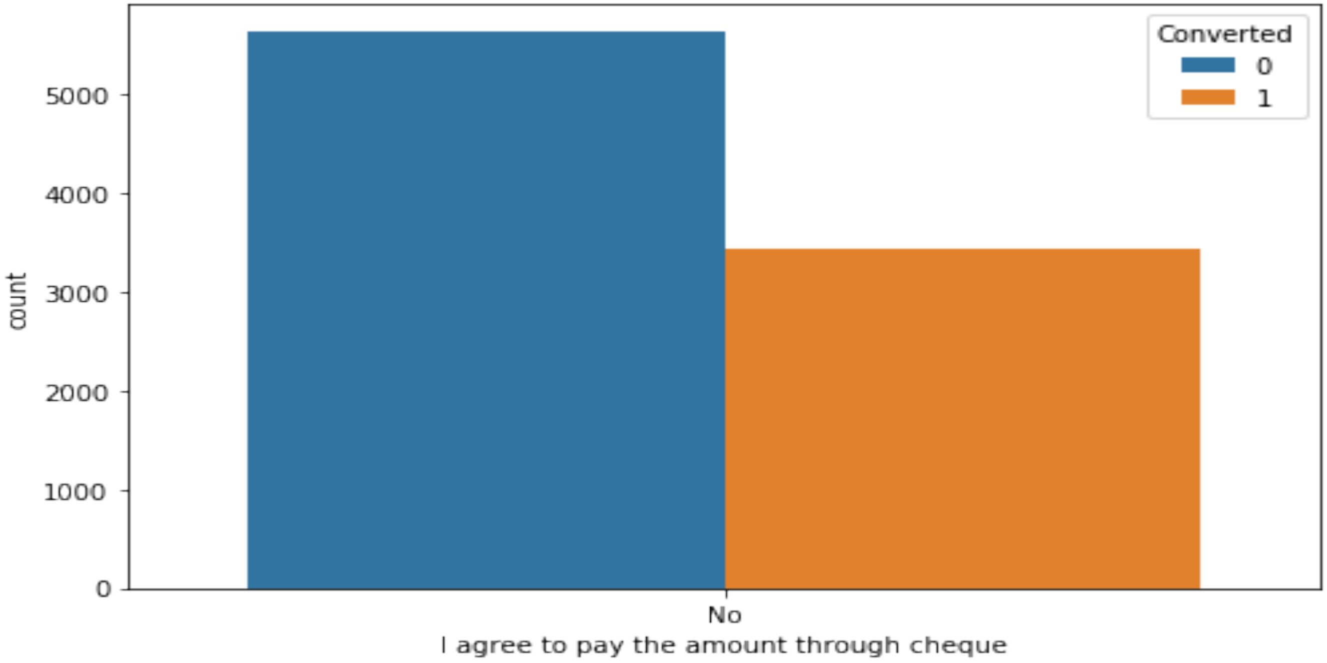
Univariate Analysis – Binary values

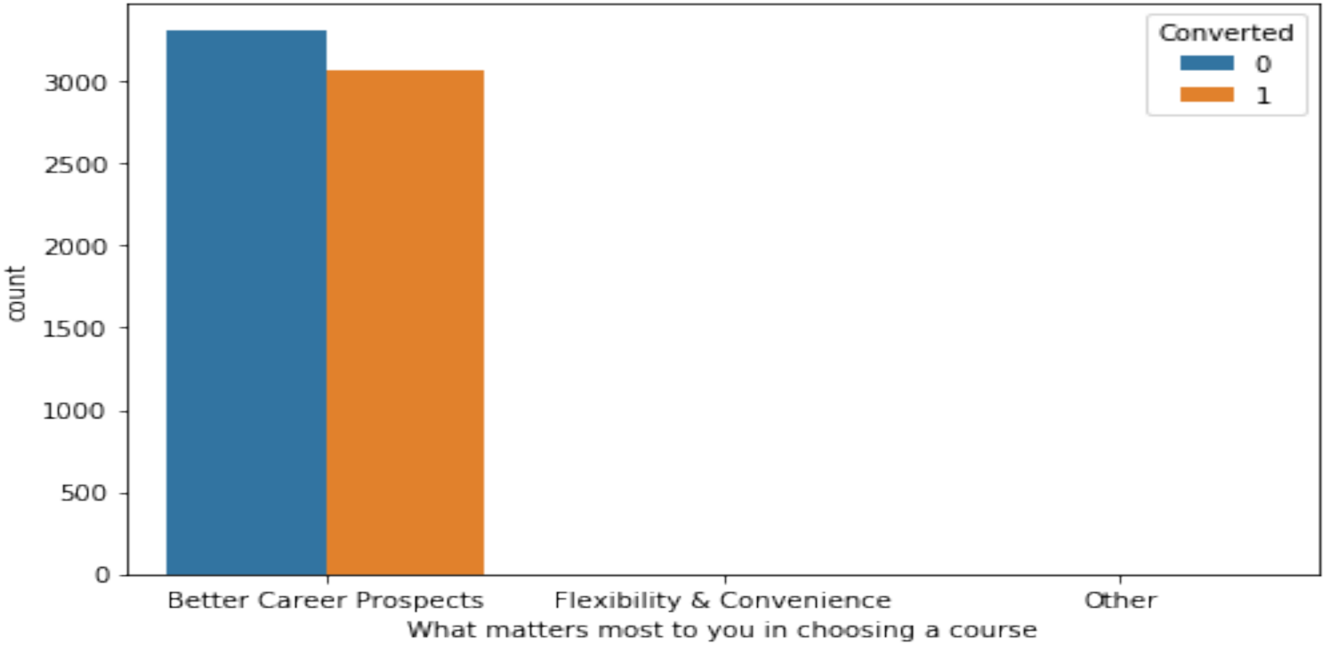


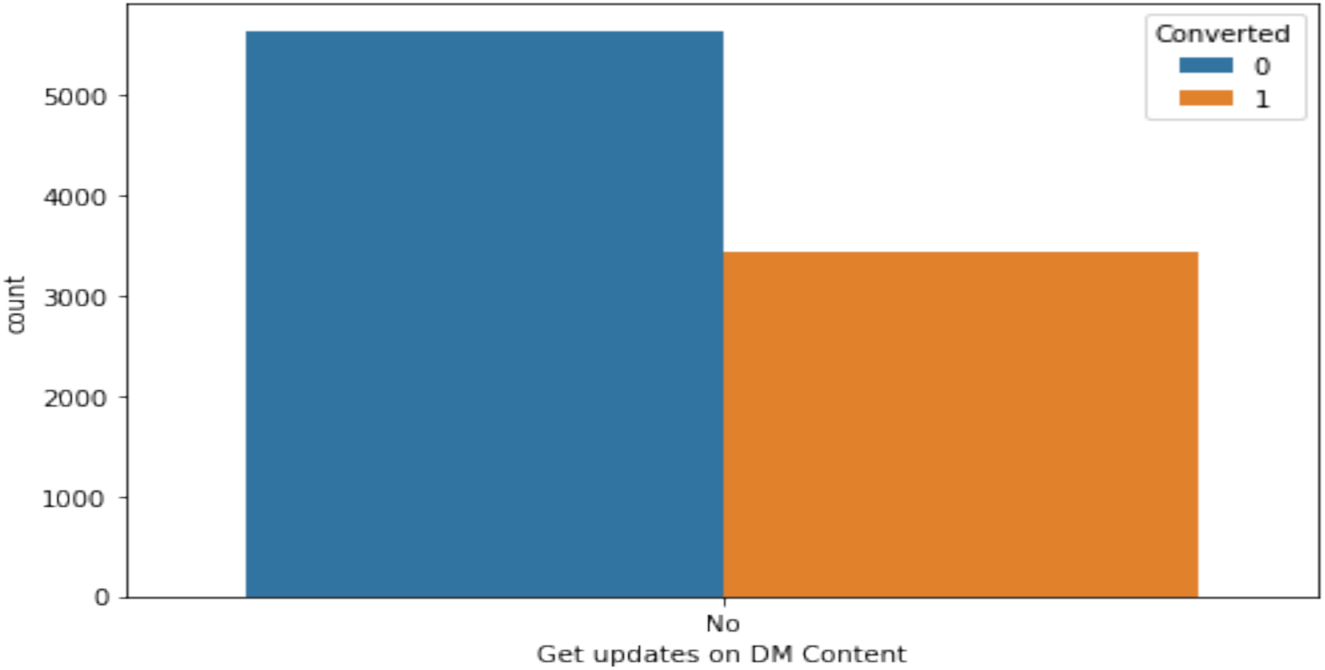



Converted
🟦 0
🟧 1
count
0
1000
2000
3000
4000
5000
No
Yes
Newspaper Article


Converted
🟦 0
🟧 1
count
0
1000
2000
3000
4000
5000
No
Yes
X Education Forums


Converted
🟦 0
🟧 1
count
0
1000
2000
3000
4000
5000
No
I agree to pay the amount through cheque


Converted
0
1
count
0
500
1000
1500
2000
2500
3000
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course


Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Get updates on DM Content

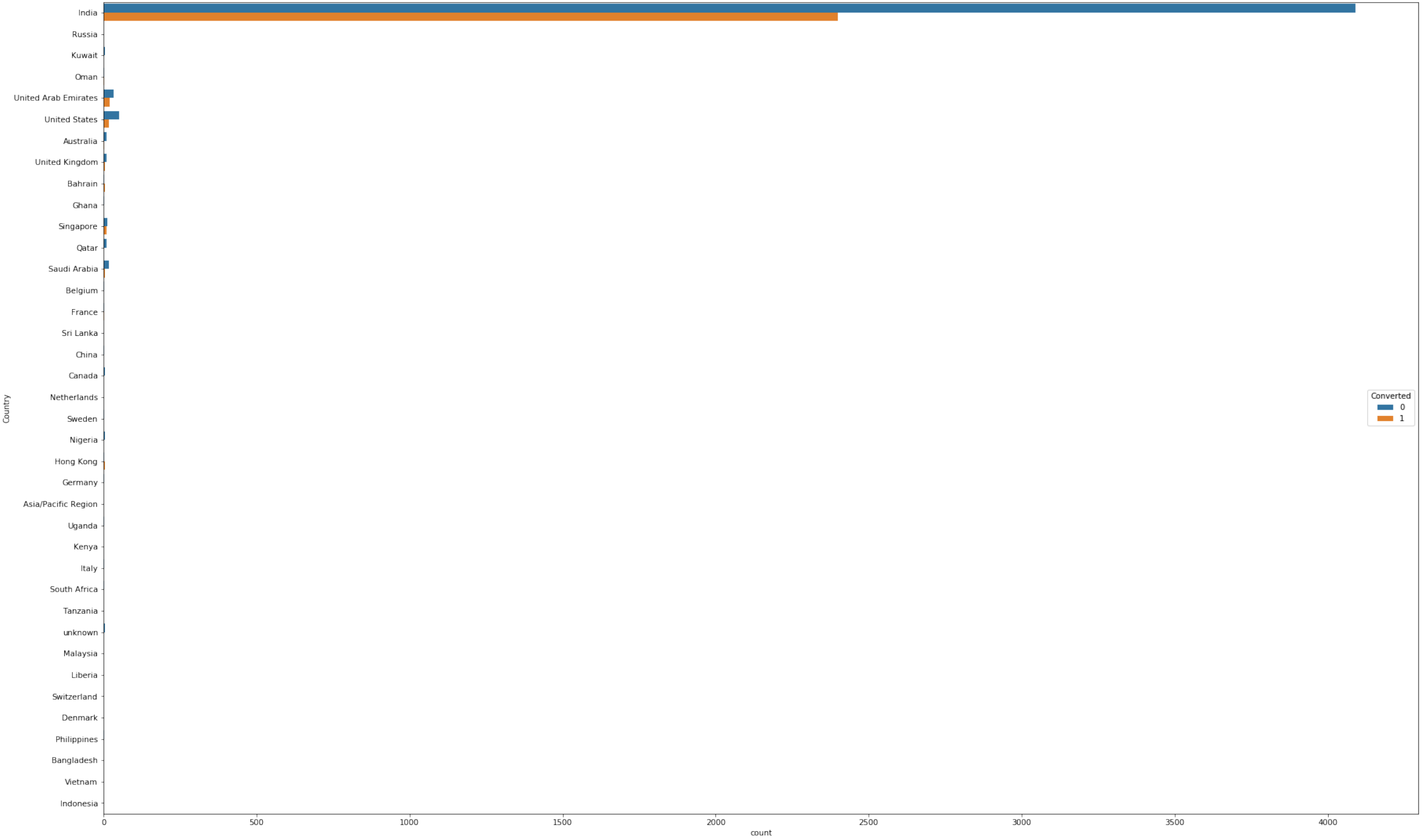
Observations:

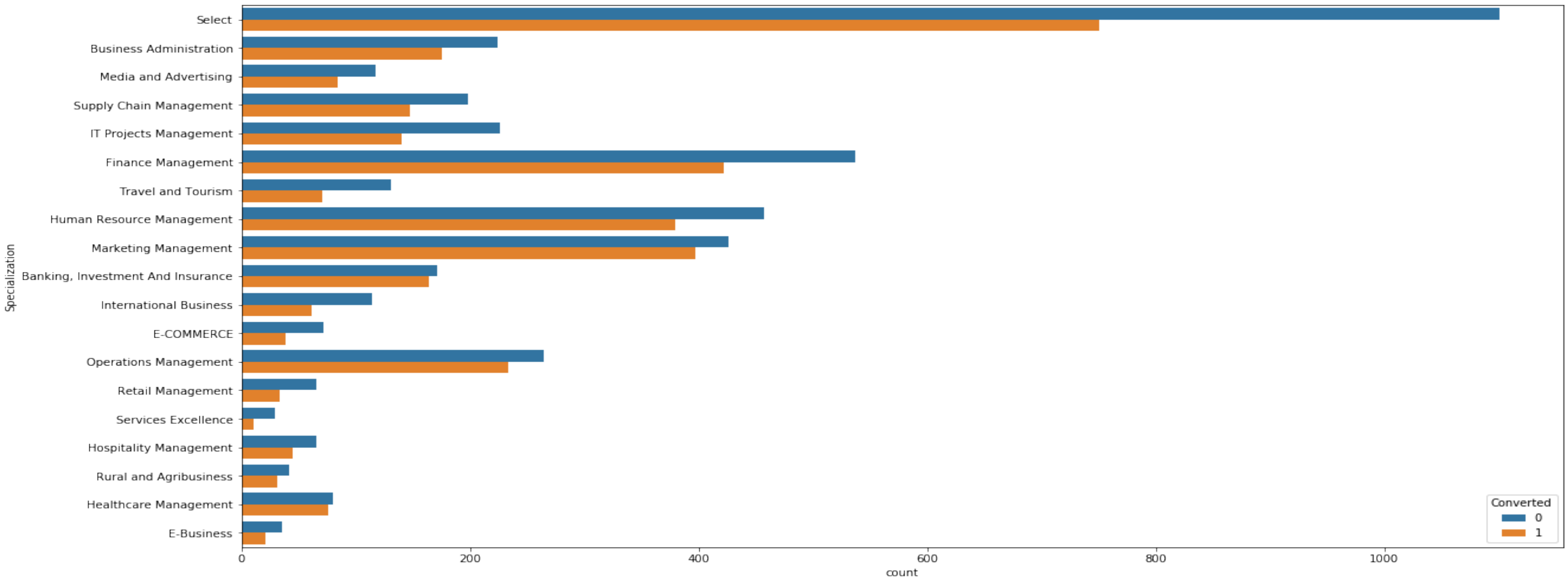
By analysis above plots, we can drop following columns

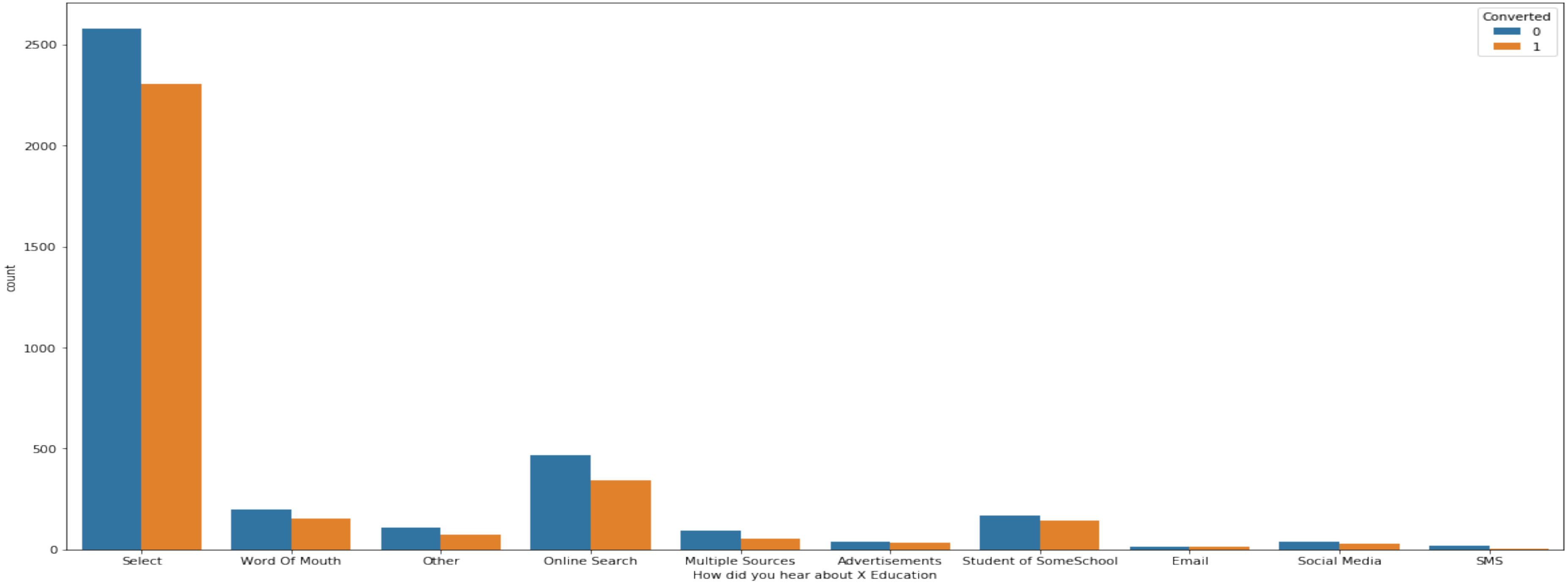
- Do Not Call
- Magazine
- Newspaper Article
- X Education Forums
- Newspaper
- Receive More Updates About Our Courses
- Update me on Supply Chain Content
- Get updates on DM Content
- I agree to pay the amount through cheque
- What matters most to you in choosing a course

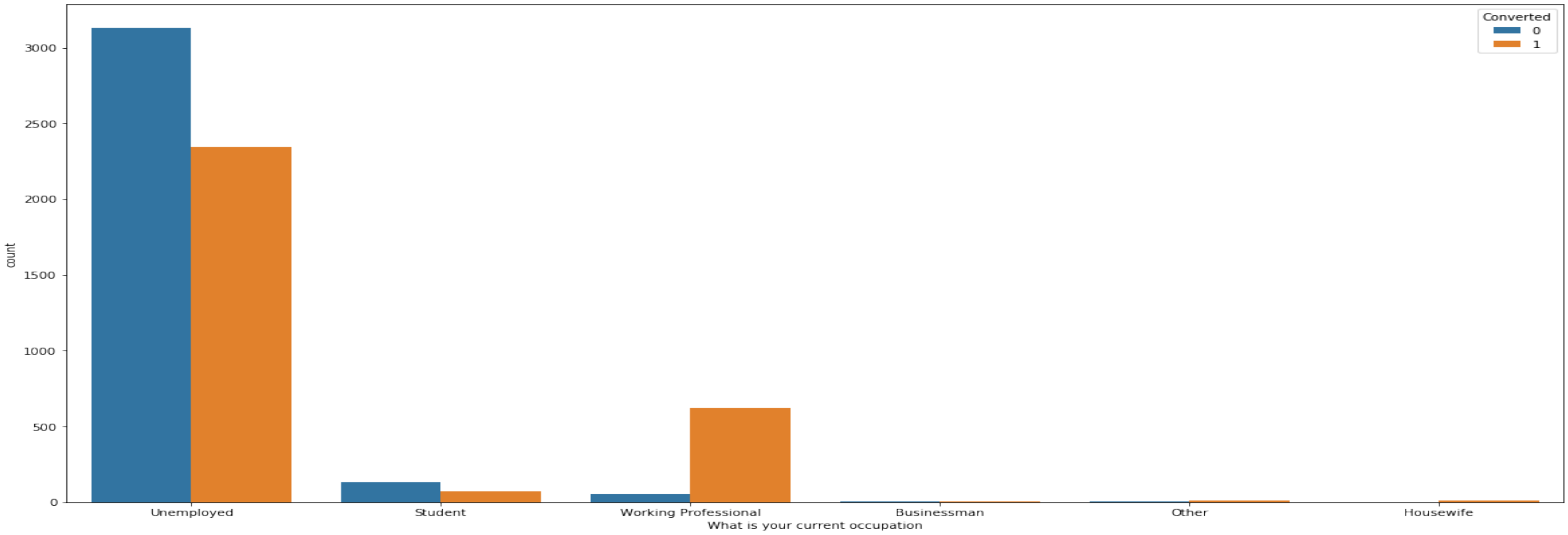
NOTE : These columns have less then 0.5% valid values

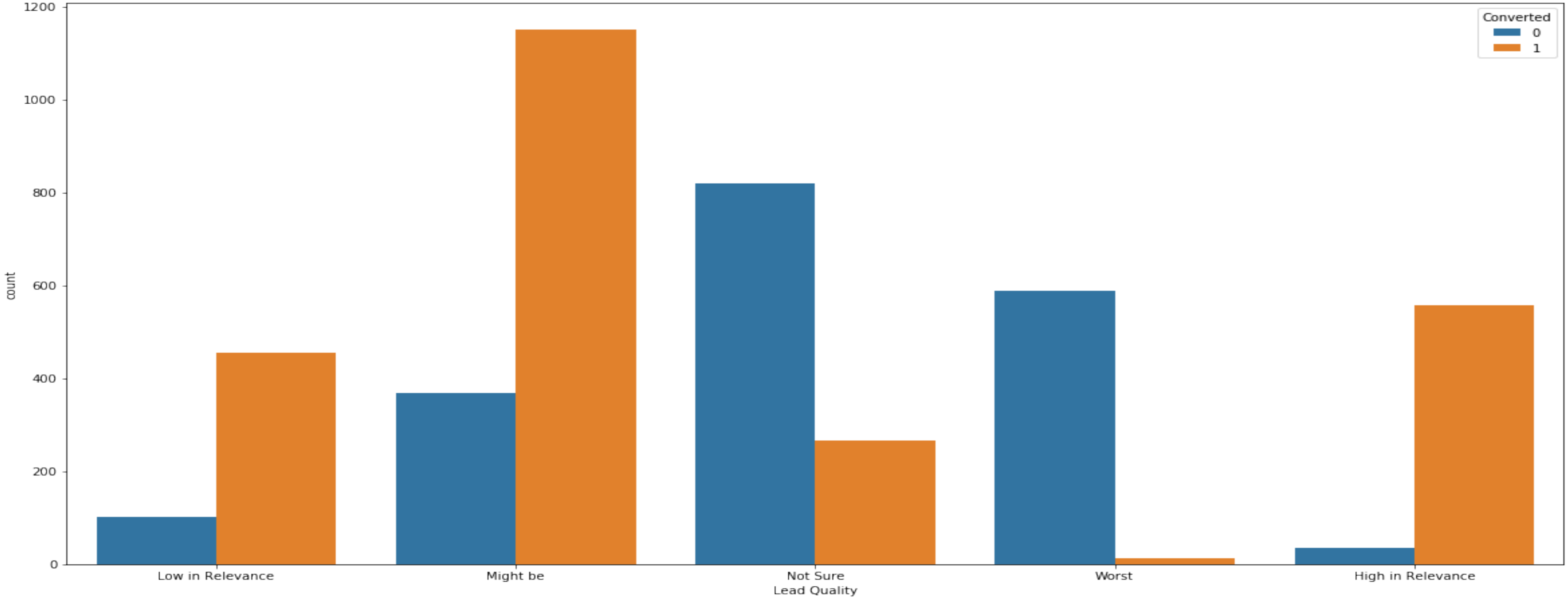
Univariate Analysis – Non Ordinal Values

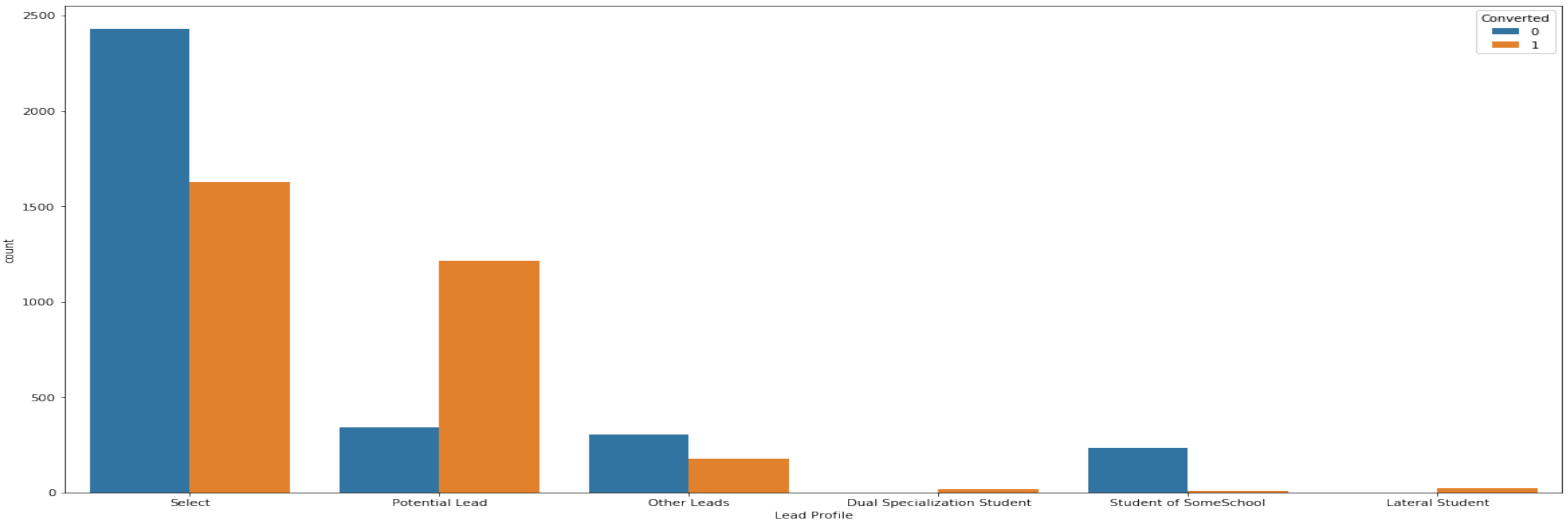


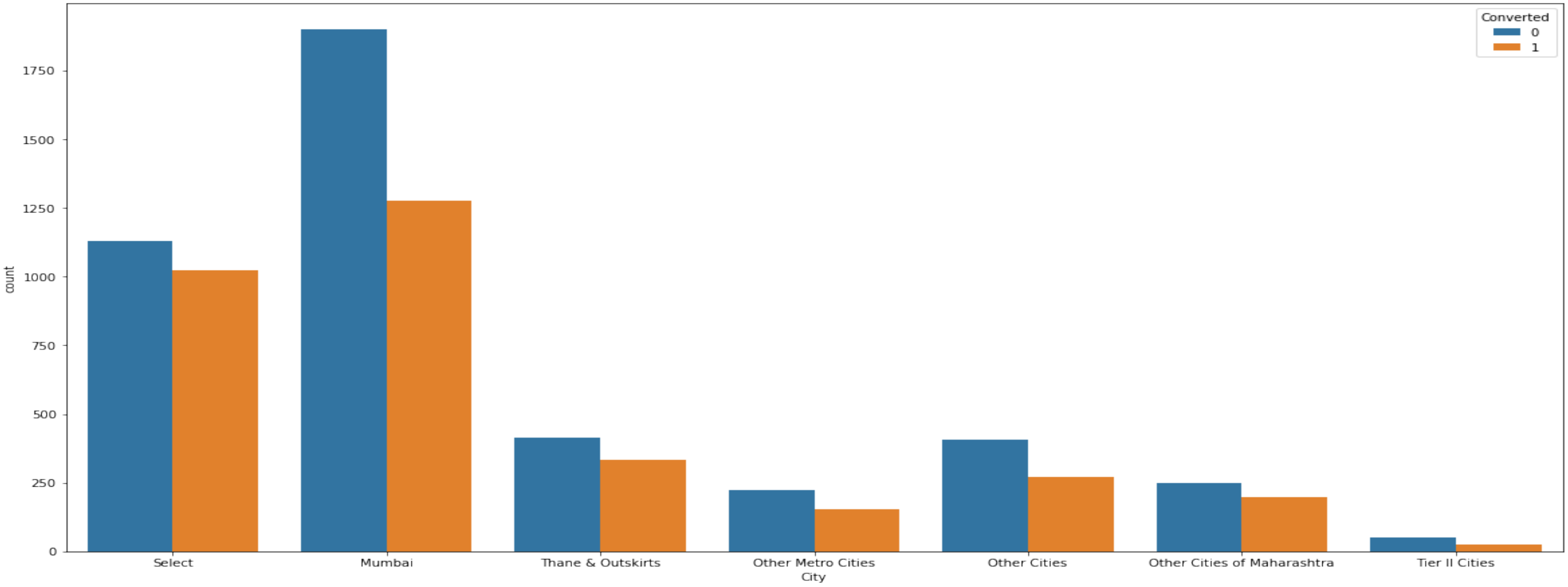

Specialization
Select
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
0
200
400
600
800
1000
count
Converted
0
1

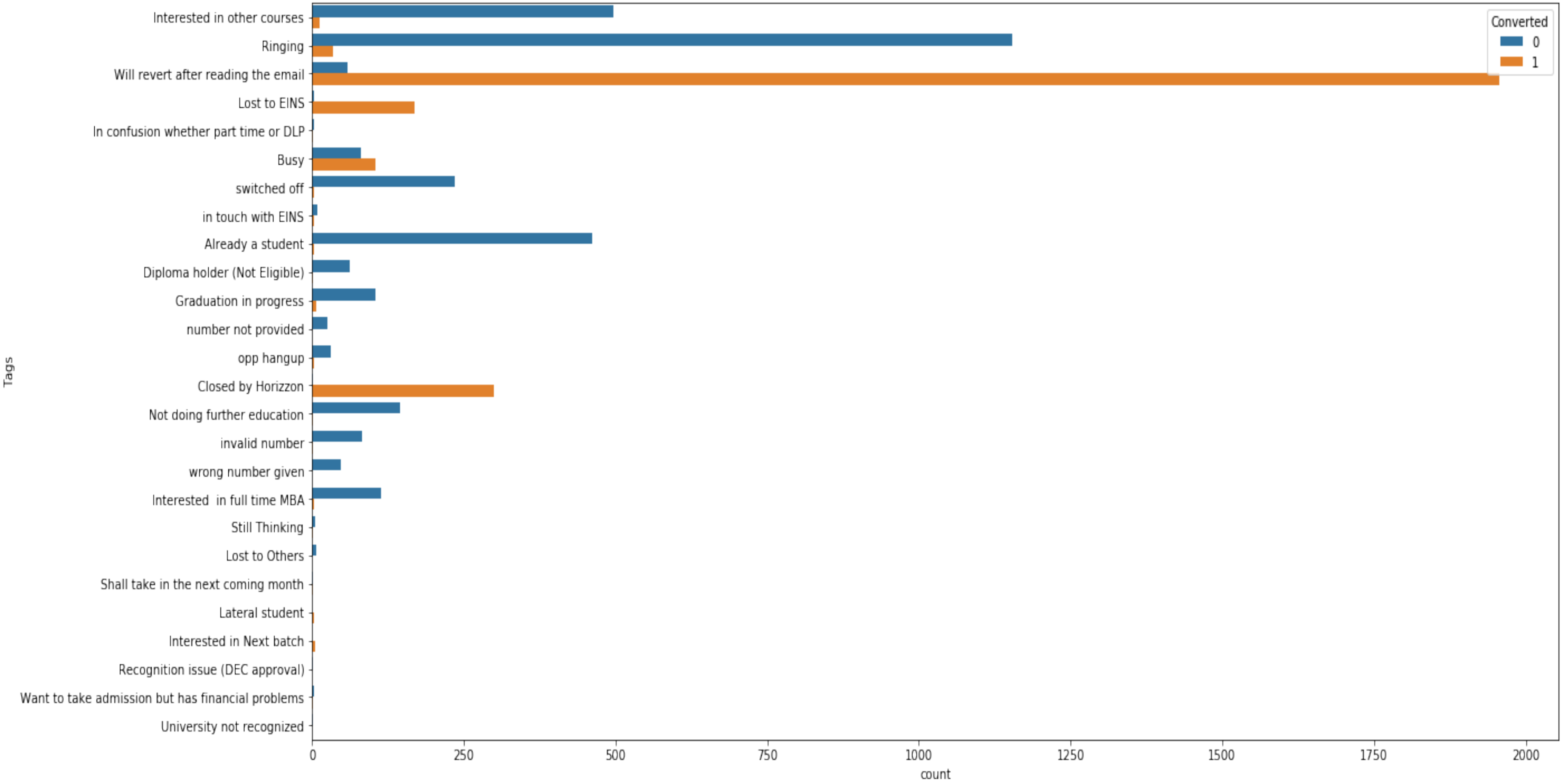

Converted
0
1
2500
2000
1500
1000
500
0
count
Select
Word Of Mouth
Other
Online Search
Multiple Sources
Advertisements
Student of SomeSchool
Email
Social Media
SMS
How did you hear about X Education


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


Converted
0
1
count
1200
1000
800
600
400
200
0
Low in Relevance
Might be
Not Sure
Worst
High in Relevance
Lead Quality


Converted
0
1
count
2500
2000
1500
1000
500
0
Select
Potential Lead
Other Leads
Dual Specialization Student
Student of SomeSchool
Lateral Student
Lead Profile


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Select
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City


Tags
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
0
250
500
750
1000
1250
1500
1750
2000
count
Converted
0
1

Observations:

- Country
    - 'india' is the mode. This can be dropped. Either the data is India or select
- City
    - 'Mumbai' is the mode. This can be dropped. Either the data is Mumbai or select.
- Lead Profile
    - 'Potential Lead is the category which has highest conversion'. There are some 'select' category present which need to be replace with random values
- Specialization
    - 'Finance /Marketing/HR/Operational Management' categories are opting more for the course and have more conversion
- How did you hear about X Education
    - 'Online Search' category has high value and has more conversion rate. There are some 'select' category present which need to be replaced
- What is your current occupation
    - 'Unemployed/Working Professional' Have more conversion rate.
- Tags
    - 'Will revert after reading the email' have high conversion rate
- Lead Quality
    - 'Might be/High in Relevance/Low in Relevance' have high Conversion rate

Handling ‘Select’:

•As we observed above in few columns there is value ‘Select’ which indicates that there is no value for that column for that Lead.

•This indicates that ‘Select’ is same as ‘Null value’ or NaN value.

•So, converted all ‘Select’ values to Nan values.

Handling NaN values:

•Replaced Nan with any random value of that column for normally distributed columns.

•Used mode of the column to replace NaN for uniformly distributed columns.

Observation:

We are left with 9074 rows and 22 columns. Not much data loss.

Outlier Analysis

- By checking outliers at 25%,50%,75%,90%,95% and 99% noticed that the columns TotalVisits, Total Time Spent on Website, Page Views Per Visit have ouliers.

- Handled them by removing data above 0.95 and below 0.05 quantiles.

- Now we are left with 9048 rows and 22 columns.

Data Preparation

- For columns with binary values – Yes, No. Using Map function converted ‘Yes’ to ‘1’ and ‘No’ to ‘0’.

- For non ordinal categorical columns combined categories with frequency less than 10% to other value.

- Then created dummy variables.

Train – Test Split :

Using train_test_split module of sklearn.model_selection package split the data into train and test with 70 – 30 percentage.

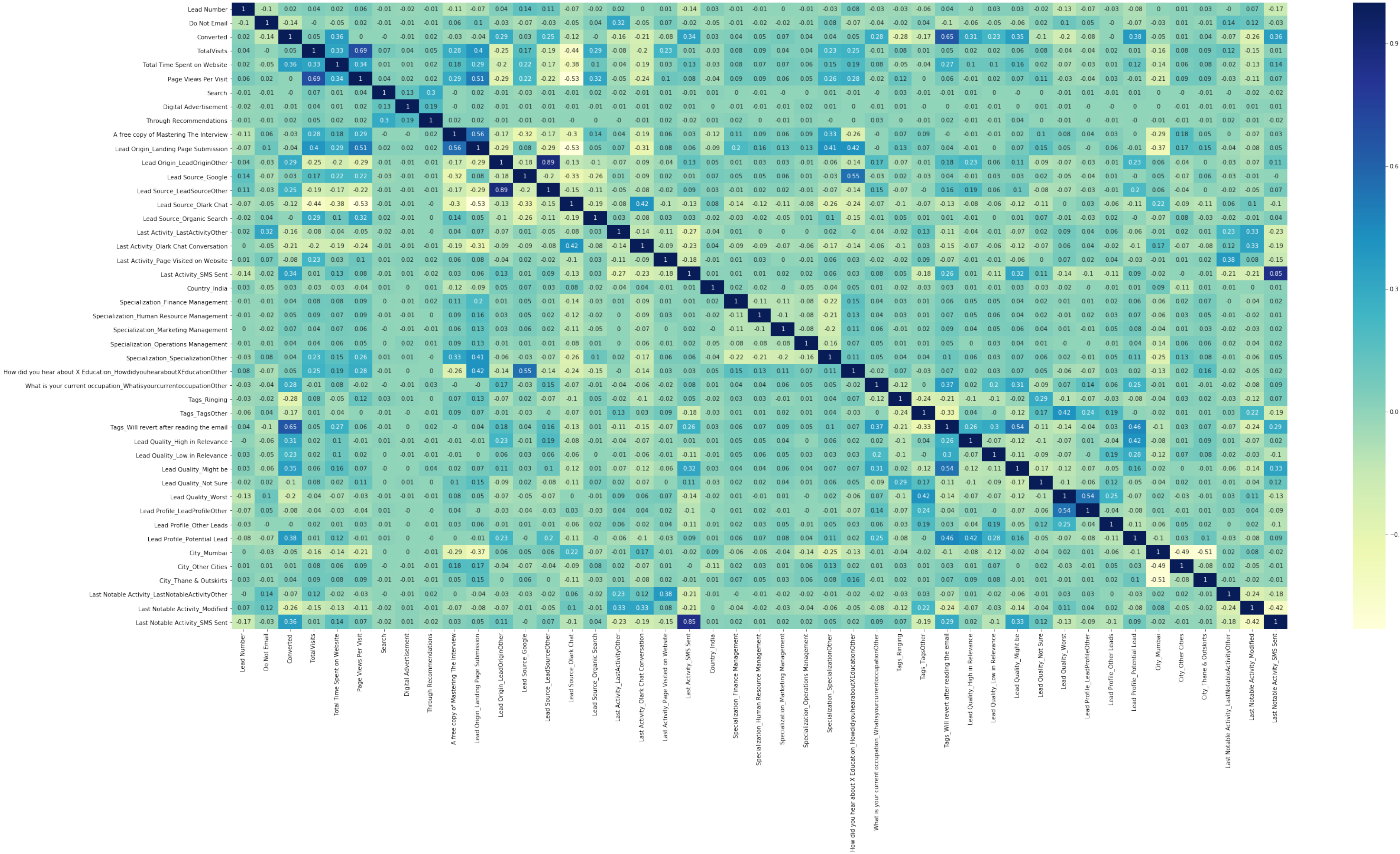
Feature Scaling:

Using StandardScaler module of sklearn.preprocessing package standardised all columns.

Conversion rate:

On this new data we have conversion rate of nearly 38%

Correlations



Model Building

- We built model using all columns – we know few are correlated as per above plot and also there are so many attributes.

- Applied RFE for feature selection – 13 variables.

- By considering probability above 0.5 as 1 and less than 0.5 as 0. we got accuracy of 89%.

- And also the VIFs are all good.

Sensitivity = 0.81

Specificity = 0.94

False positive rate = 0.05

Positive predictive value = 0.90

Negative predictive value = 0.89

Though the overall accuracy is good, we need to check if we can get good sensitivity.

ROC Curve

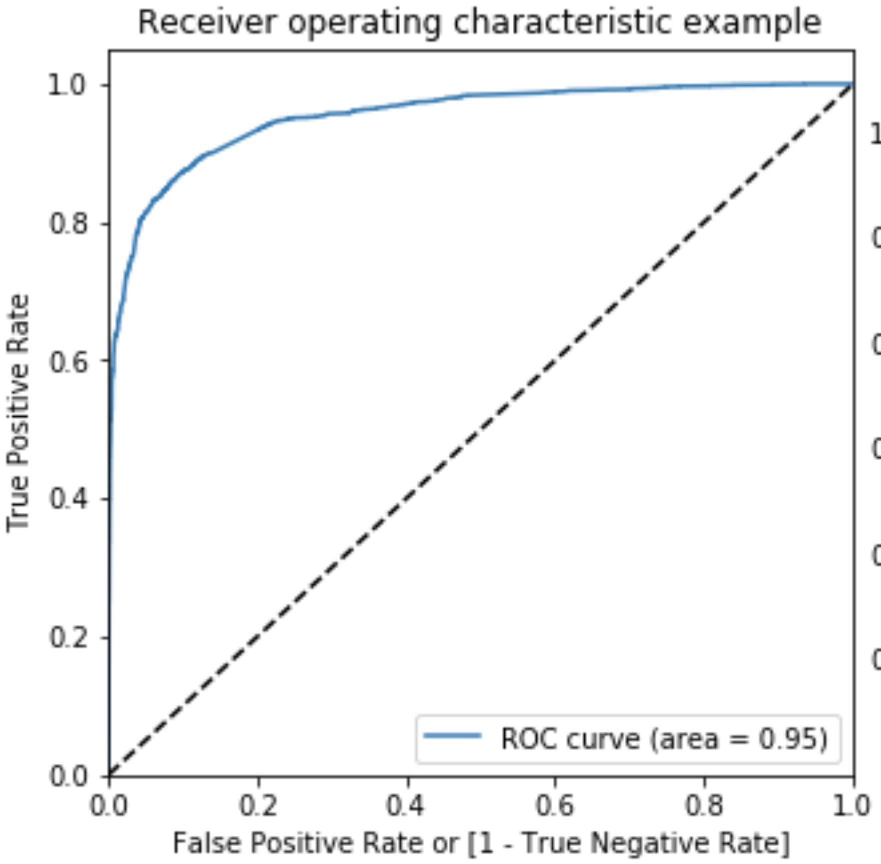
An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

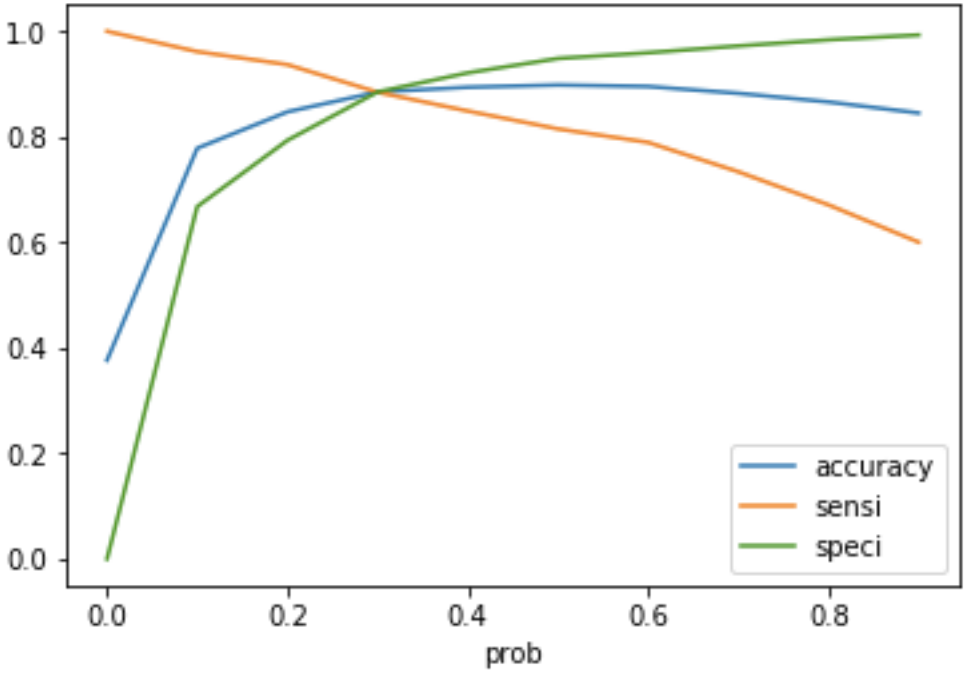
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

ROC Curve



Optimal cut off point



- By calculating accuracy, sensitivity, specificity of variables probabilities and plotting them, we observed 0.3 is the optimum value.

Final Values on train data :

Accuracy = 0.88
Sensitivity = 0.88
Specificity = 0.88
False positive rate = 0.11
Positive predictive value = 0.82
Negative predictive value = 0.92

Test Data :

Accuracy = 0.88
Sensitivity = 0.89
Specificity = 0.87

Summary : Attributes that have impact on Lead score are

Main factors.	**Coefficient**
Tags_Will revert after reading the email	4.2012
Lead Origin_LeadOriginOther	3.8368
Lead Quality_High in Relevance	2.4856
Last Notable Activity_SMS Sent	1.8342
Lead Source_Olark Chat	1.4128
Total Time Spent on Website	1.0710
Lead Quality_Low in Relevance	0.9677
Last Activity_LastActivityOther	-0.8411
Last Activity_Page Visited on Website	-0.6047
Do Not Email	-1.0166
Last Activity_Olark Chat Conversation	-1.6466
Lead Quality_Worst	-2.0549
Tags_Ringing	-3.5245

Tags / Lead Origin /Lead Quality Are the important factors to be considered for lead conversion