# LEAD SCORE ASSIGNMENT

Submitted By

*1. Kuntamukkala Pavan Kumar*

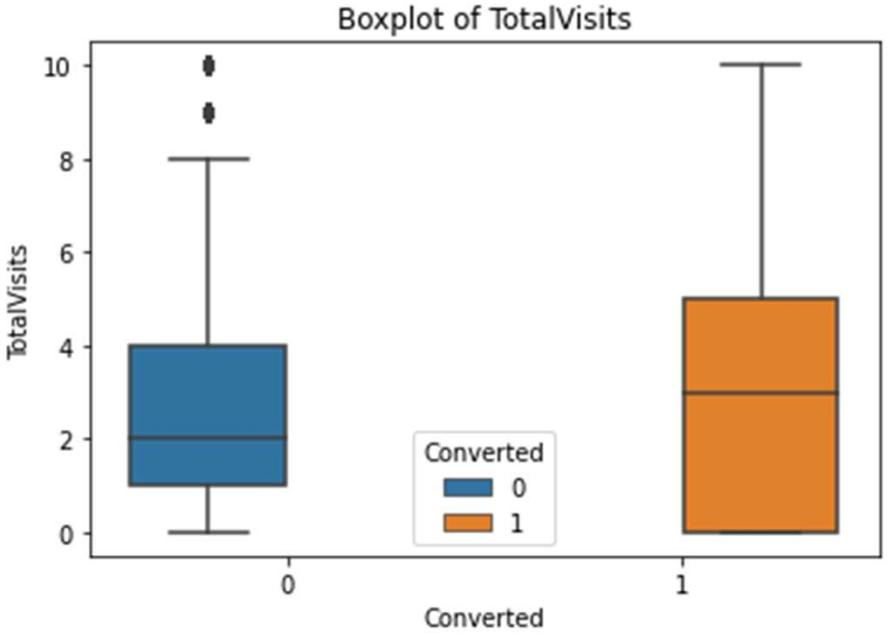
*2. Sanu Kumar*

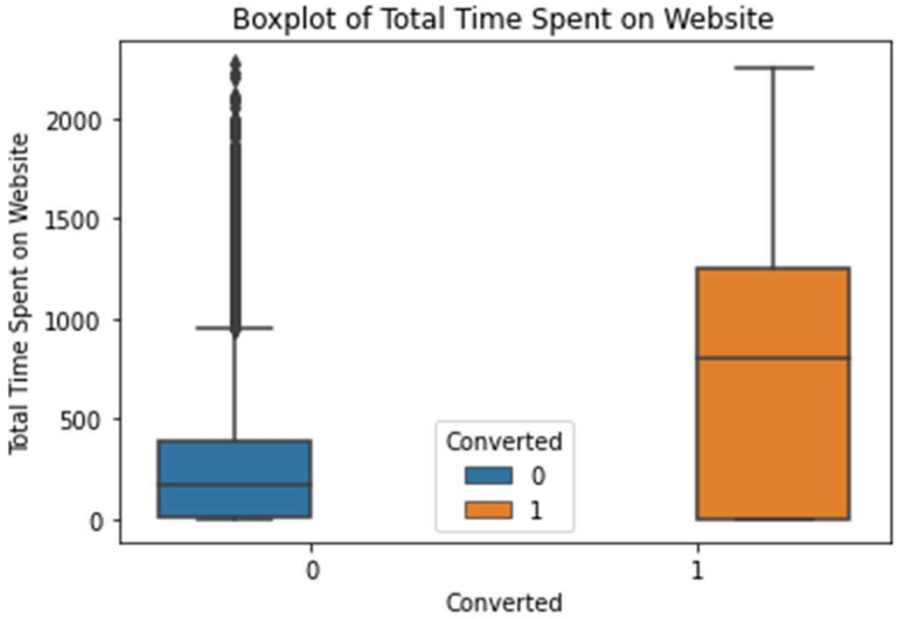
*3. Abhishek Chandanshive*

# Conversion of Leads to Customers for X Education Company:

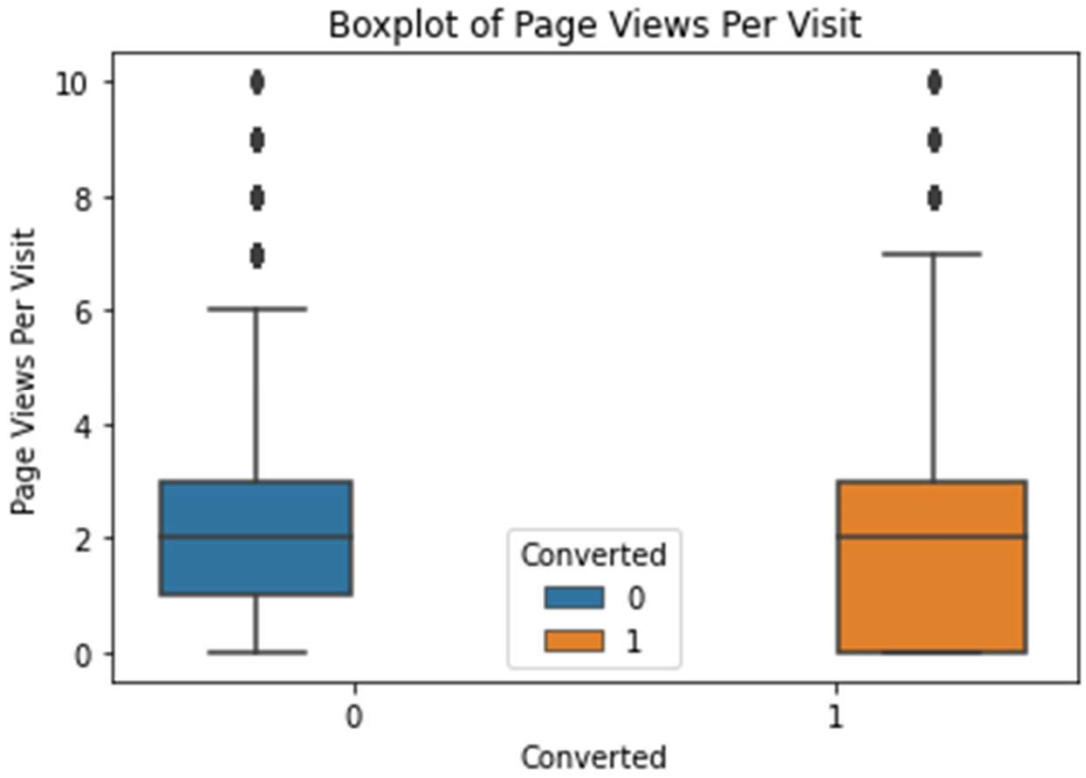
- The given data set contains 37 columns and 9240 rows.
- 4 of the columns are Numerical variables.
- Remaining are Object/ Categorical variables.

# EXPLORATORY DATA ANALYSIS



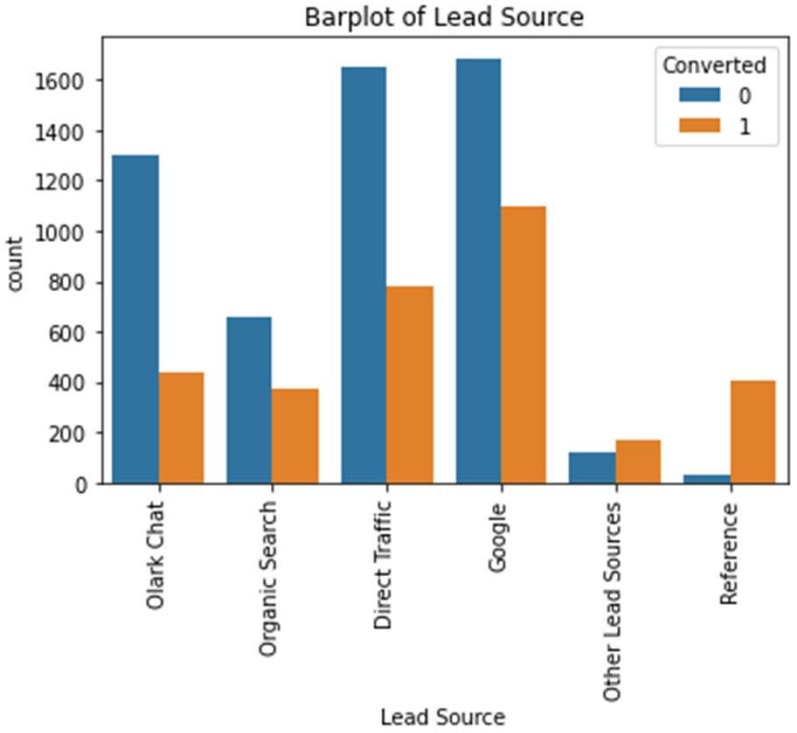



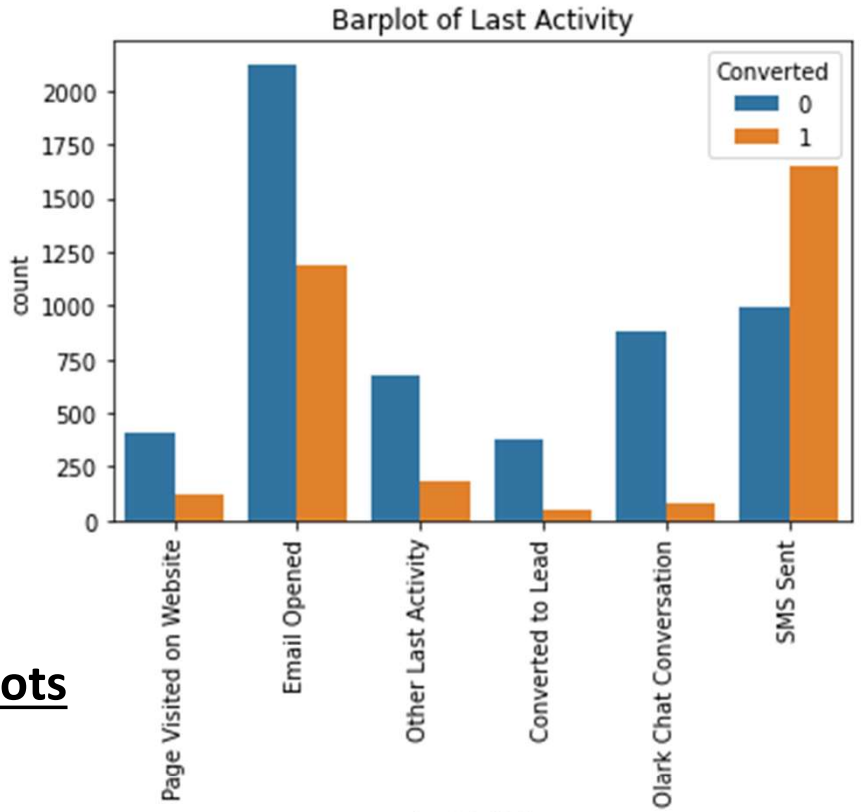

- From ‘Total Visits’ column, we can see that median value of Converted Leads is greater than the median value of Non Converted Leads.
- It suggests that greater the total visits made by a Lead, greater will be the chance for that lead to be converted to customer.
- This suggests that website should be attractive enough for the leads to come and visit.
- From ‘Total Time spent on Website’, we can clearly see that median value of Converted Leads is more when compared to median value of Non Converted Leads.
- This suggests that website should be engaging enough for the Leads to spend more time on the website.
- It should be informative and at the same time, the information should be easily available. The links and their presentation should be engaging enough.

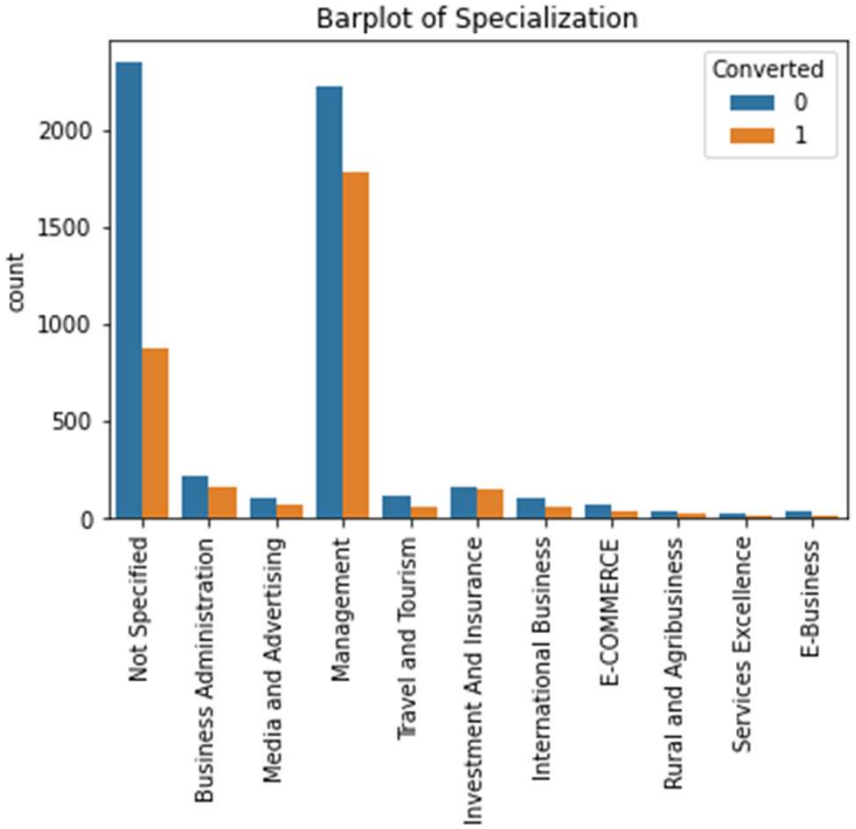


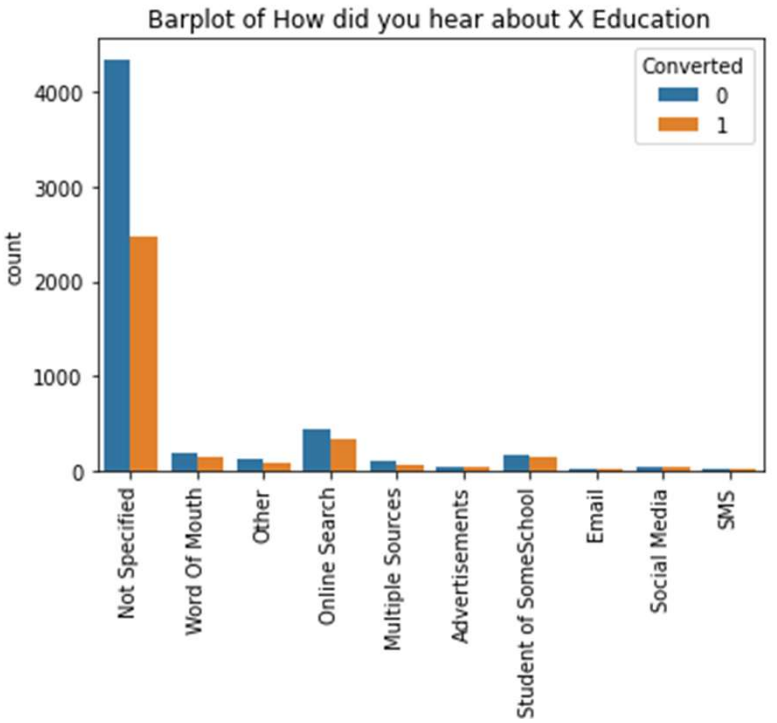

- From ‘Page Views Per Visit’, we cannot see any clear difference in the median values of converted and non converted leads.
- Thus it is not of significant importance to contain as many pages as possible for the lead conversion.

# Count Plots

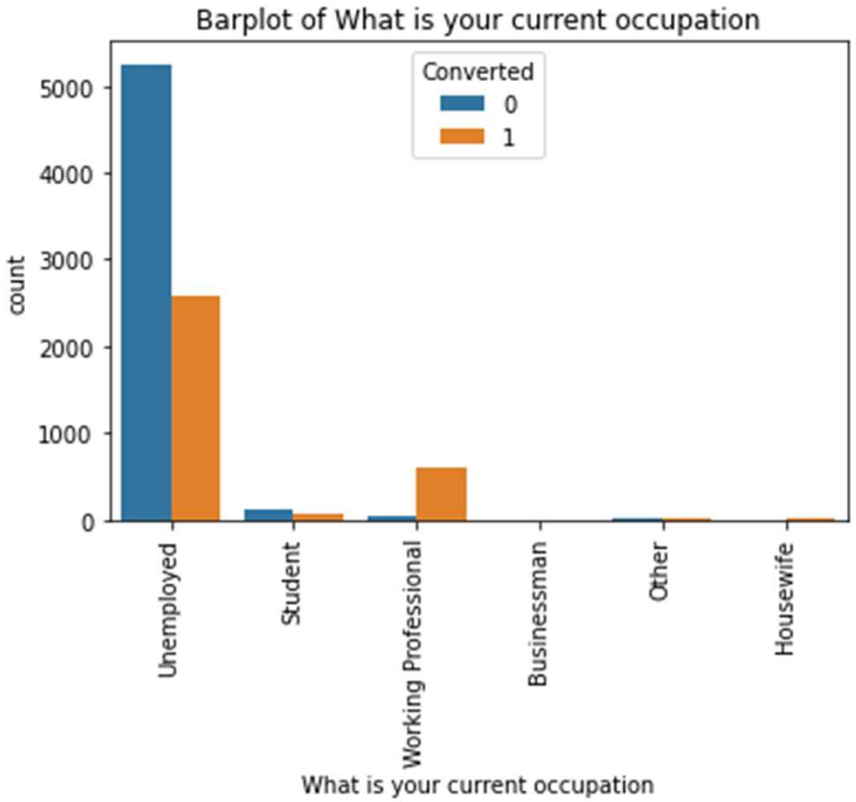

Barplot of Lead Source
Converted
0
1
count
Olark Chat
Organic Search
Direct Traffic
Google
Other Lead Sources
Reference
Lead Source


Barplot of Last Activity
Converted
0
1
count
Page Visited on Website
Email Opened
Other Last Activity
Converted to Lead
Olark Chat Conversation
SMS Sent
Last Activity


Barplot of Specialization
Converted
0
1
count
Not Specified
Business Administration
Media and Advertising
Management
Travel and Tourism
Investment And Insurance
International Business
E-COMMERCE
Rural and Agribusiness
Services Excellence
E-Business

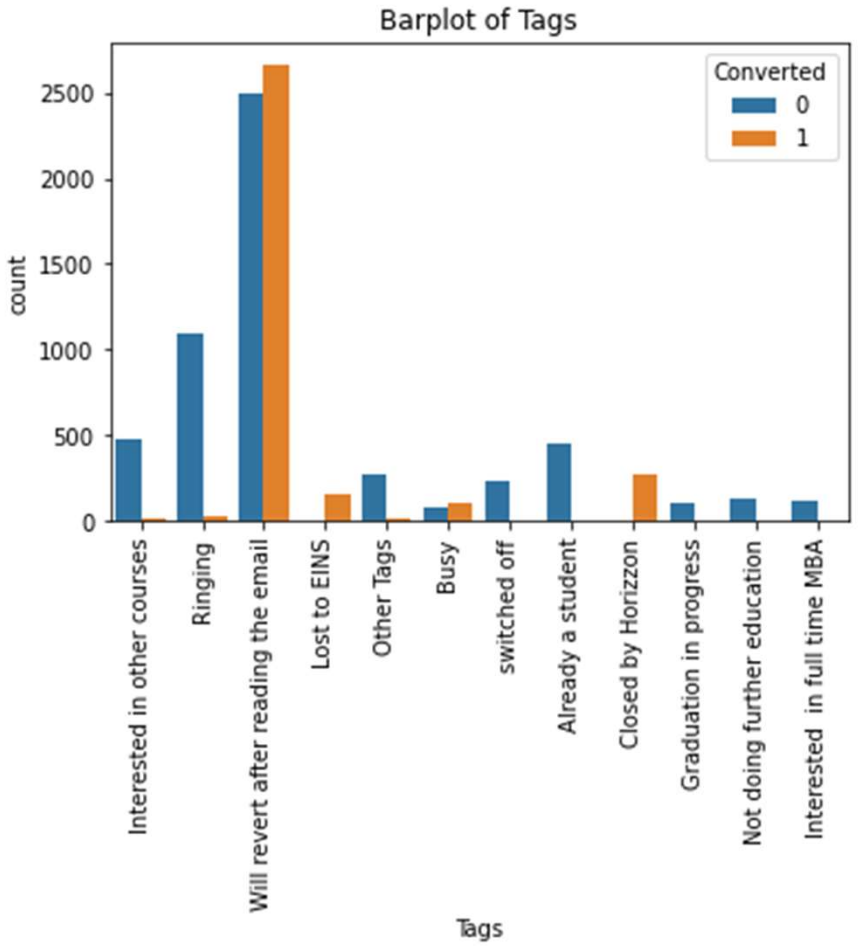

Barplot of How did you hear about X Education
Converted
0
1
count
Not Specified
Word Of Mouth
Other
Online Search
Multiple Sources
Advertisements
Student of SomeSchool
Email
Social Media
SMS

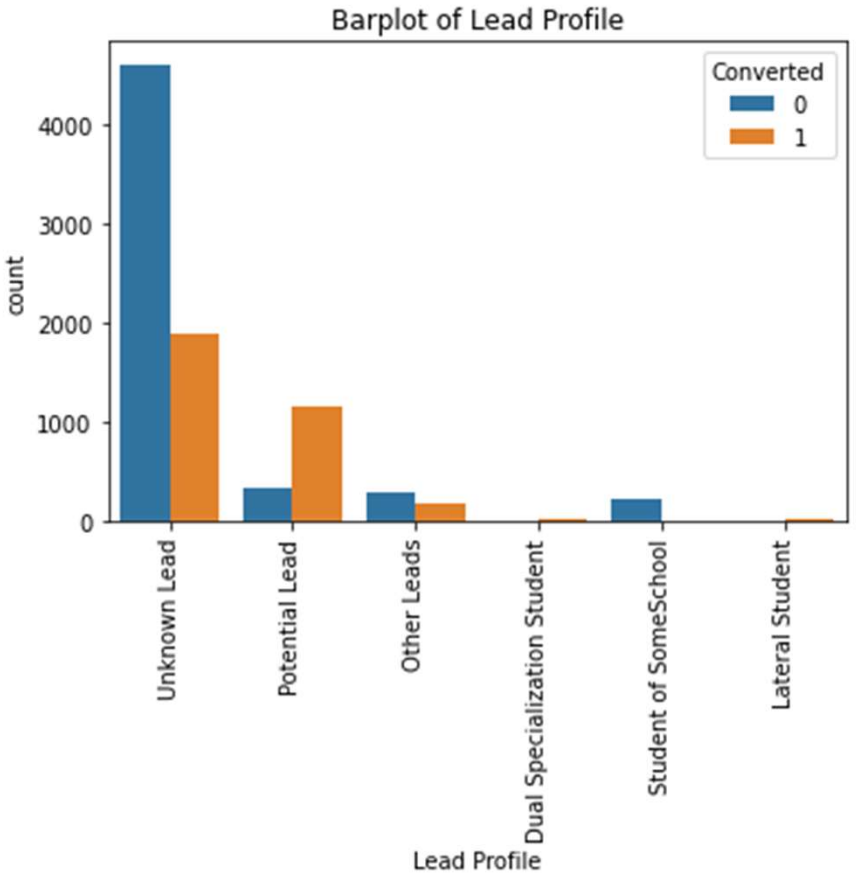
# Count Plots


Barplot of What is your current occupation
Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

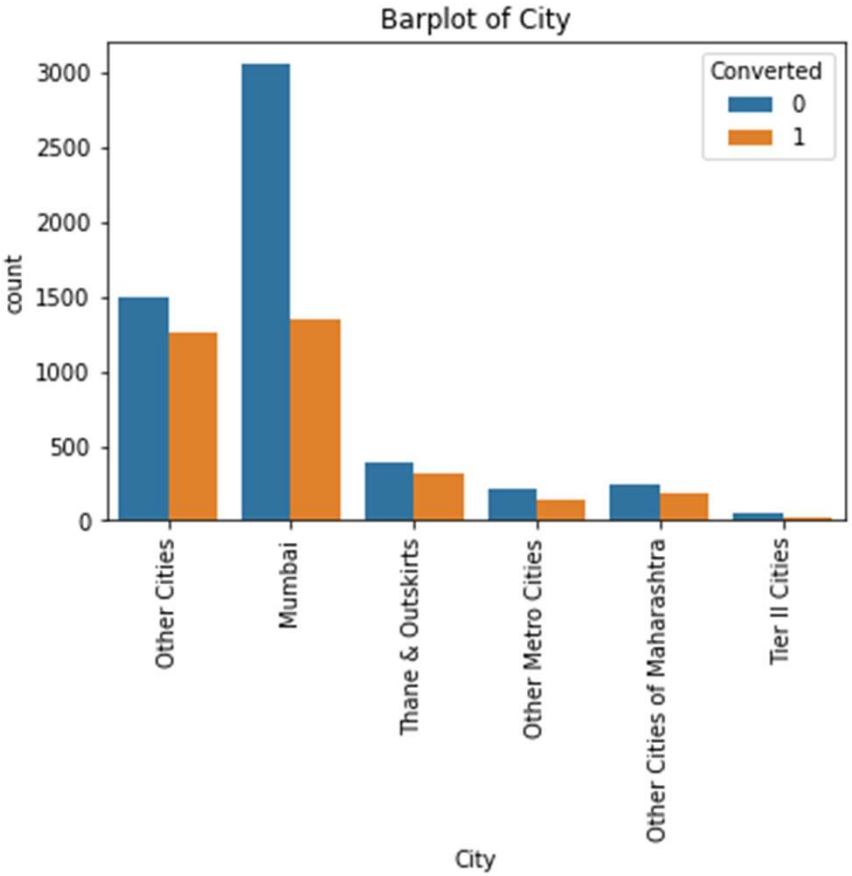

Barplot of What matters most to you in choosing a course
Converted
0
1
count
5000
4000
3000
2000
1000
0
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course

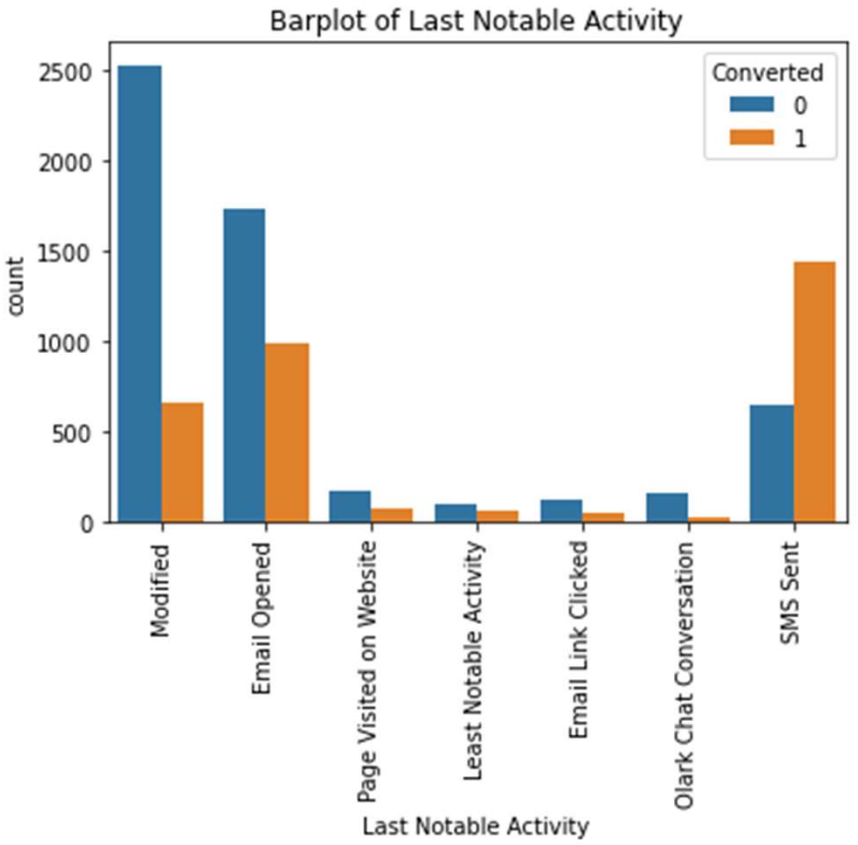
# Count Plots


Barplot of Tags
Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
Other Tags
Busy
switched off
Already a student
Closed by Horizzon
Graduation in progress
Not doing further education
Interested in full time MBA
Tags


Barplot of Lead Profile
Converted
0
1
count
4000
3000
2000
1000
0
Unknown Lead
Potential Lead
Other Leads
Dual Specialization Student
Student of SomeSchool
Lateral Student
Lead Profile

# Count Plots


Barplot of City
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Other Cities
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities of Maharashtra
Tier II Cities
City

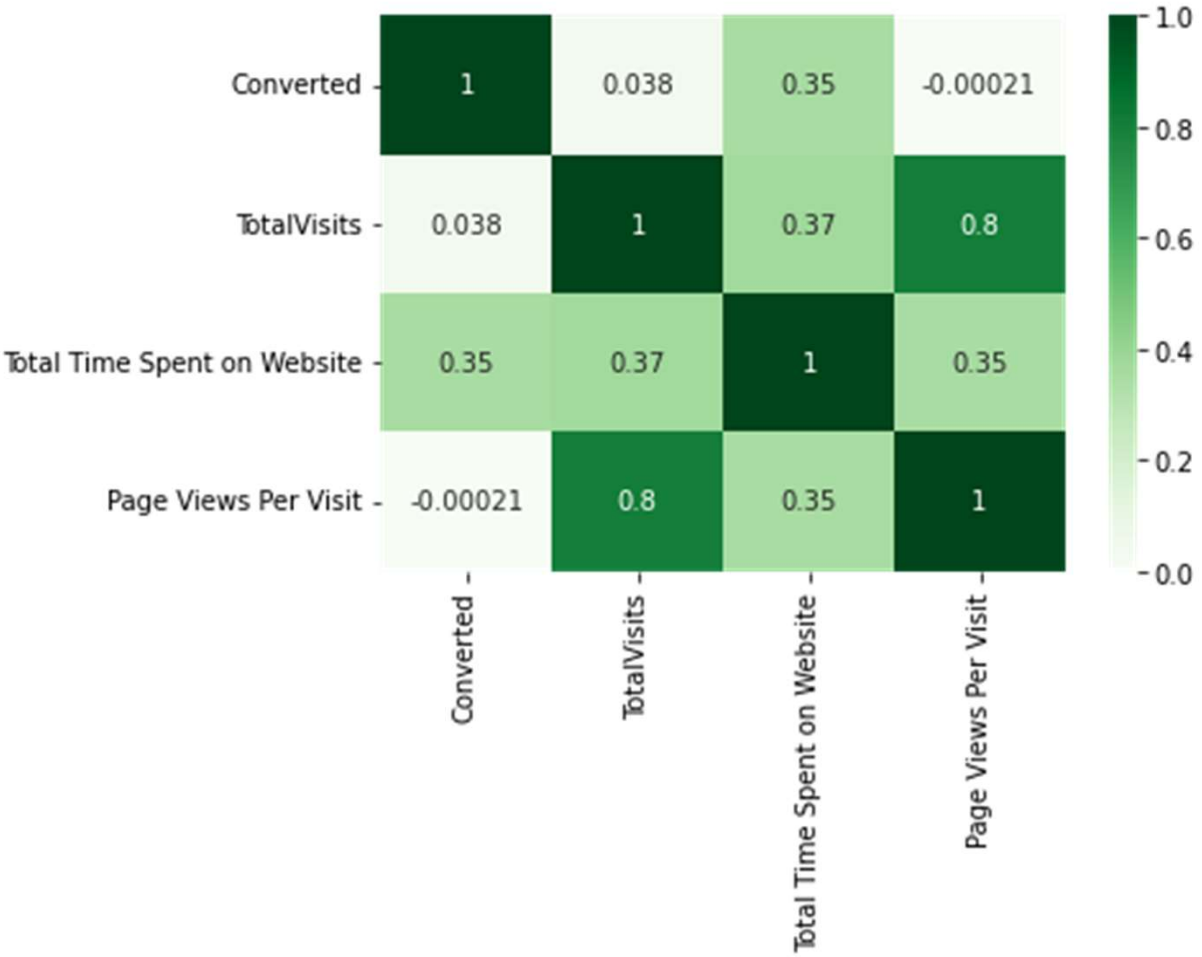

Barplot of Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Least Notable Activity
Email Link Clicked
Olark Chat Conversation
SMS Sent
Last Notable Activity

From the above count plots we can deduce the following:

1. The number one source of Leads irrespective of Converted or Not Converted seems to be 'Google' and next to that is 'Direct Traffic' (i.e., Direct visit to websites)
   One other important point to observe here, there is high conversion rate of Leads who made through 'Referrence'.
   Even though it is small percentage of total leads, focussing on this 'Referrence' source could be a better option.
   Such as providing discounts to the referring persons and the referred persons.
2. The customers with last activity as 'SMS Sent' has higher conversion rate of turning Leads in to 'Converted'.
3. India is the major source of customers for the company.
4. The people who belong to 'Management Specialization' has high chance of purchasing the course.
   The number of leads who are converted are more in the 'Management Specialization' category.
   Next to them are 'Not Specified' Category even though the conversion of leads are low.

5. Majority of the leads have not specified their source.
6. The Leads with Tag 'Will Revert after reading the mail' has high conversion rate compared to leads with remaining tags.
7. The leads which are assigned 'Potential Lead' has higher conversion rate to become a customer than other lead categories.
   It is suggestible to focus on 'Potential Lead' Category.
8. The majority of leads who are converted belong to 'Mumbai' and 'Other Cities' category.
   This provides a good opportunity to extend the business by providing discounts or incentives to customers belonging to this category.
9. The leads whose last notable activity as 'SMS Sent' has higher chance of becoming a customer.

## Heat Map of Numeric Variables




From the above heatmap, we can see a good positive correlation of 0.8 exists between 'Total Visits' and 'Page Views Per Visit'.

# LOGISTIC REGRESSION MODEL

## Logistic Regression Model built by RFE Approach:

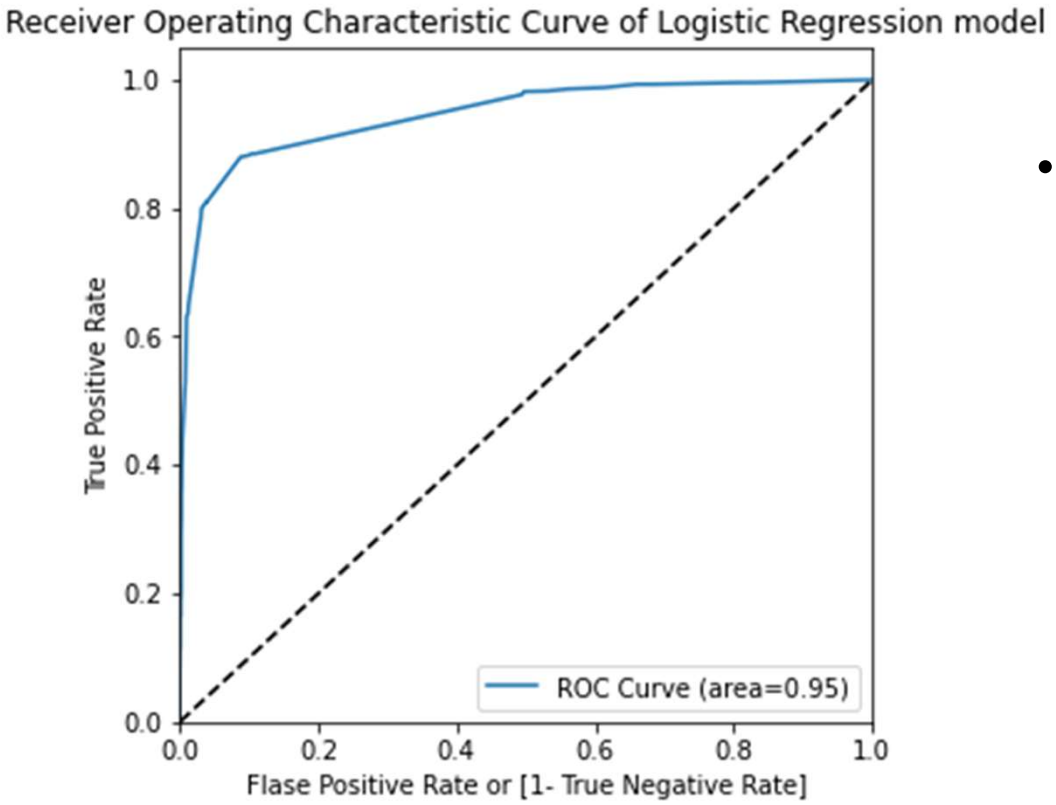
Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6104 |
| Model: | GLM | Df Residuals: | 6093 |
| Model Family: | Binomial | Df Model: | 10 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -1636.6 |
| Date: | Tue, 24 Jan 2023 | Deviance: | 3273.2 |
| Time: | 16:33:39 | Pearson chi2: | 2.70e+04 |
| No. Iterations: | 8 | Pseudo R-squ. (CS): | 0.5451 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -3.9423 | 0.188 | -20.992 | 0.000 | -4.310 | -3.574 |
| Lead Origin_Lead Add Form | 1.7700 | 0.300 | 5.894 | 0.000 | 1.181 | 2.359 |
| What is your current occupation_Working Professional | 2.5900 | 0.287 | 9.031 | 0.000 | 2.028 | 3.152 |
| Tags_Busy | 4.8650 | 0.300 | 16.222 | 0.000 | 4.277 | 5.453 |
| Tags_Closed by Horizzon | 10.1145 | 0.756 | 13.383 | 0.000 | 8.633 | 11.596 |
| Tags_Lost to EINS | 10.7529 | 0.752 | 14.305 | 0.000 | 9.280 | 12.226 |
| Tags_Will revert after reading the email | 5.7397 | 0.211 | 27.234 | 0.000 | 5.327 | 6.153 |
| Lead Profile_Student of SomeSchool | -2.7115 | 1.045 | -2.595 | 0.009 | -4.759 | -0.664 |
| Lead Profile_Unknown Lead | -3.4165 | 0.156 | -21.930 | 0.000 | -3.722 | -3.111 |
| City_Other Cities | 1.3058 | 0.106 | 12.341 | 0.000 | 1.098 | 1.513 |
| Last Notable Activity_SMS Sent | 2.9448 | 0.115 | 25.583 | 0.000 | 2.719 | 3.170 |

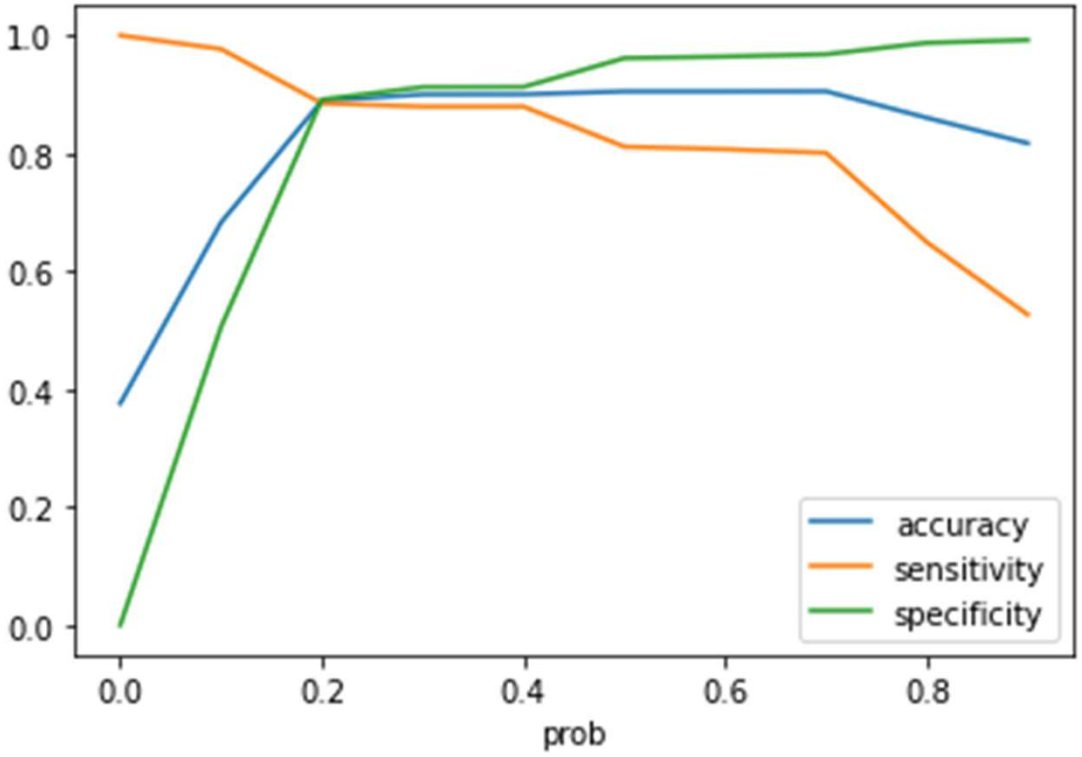
The above columns are arrived after deletion and addition of several variables based on their respective p-values and VIF values

## ROC Curve:




- The area under ROC Curve (i.e., Gini) came to be 0.95 for the generated model, suggesting that it is a good model.

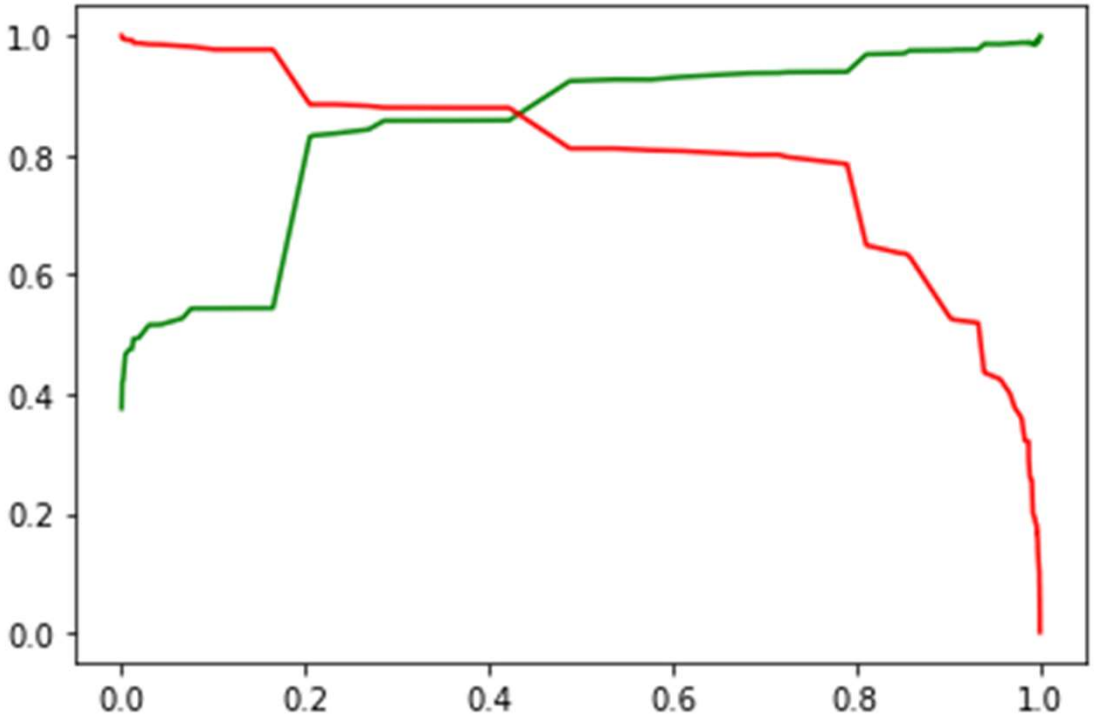
## Accuracy, Sensitivity, Specificity:




| | prob | accuracy | sensitivity | specificity |
|---|---|---|---|---|
| 0.0 | 0.0 | 0.375819 | 1.000000 | 0.000000 |
| 0.1 | 0.1 | 0.682831 | 0.976896 | 0.505774 |
| 0.2 | 0.2 | 0.888925 | 0.884917 | 0.891339 |
| 0.3 | 0.3 | 0.899902 | 0.879250 | 0.912336 |
| 0.4 | 0.4 | 0.900066 | 0.879250 | 0.912598 |
| 0.5 | 0.5 | 0.904980 | 0.811247 | 0.961417 |
| 0.6 | 0.6 | 0.904980 | 0.806888 | 0.964042 |
| 0.7 | 0.7 | 0.905144 | 0.800785 | 0.967979 |
| 0.8 | 0.8 | 0.860256 | 0.649085 | 0.987402 |
| 0.9 | 0.9 | 0.817169 | 0.526591 | 0.992126 |

- From the graph, we can see that accuracy, sensitivity and specificity trade off at an optimal cutoff probability or threshold probability value of 0.2.
- From the table we can see that for cut off probability at 0.2:
  - Sensitivity = 88.49 %
  - Specificity = 89.13 %
  - Accuracy = 88.89 %
- The above parameter values are acceptable and is considered a good model.

## Precision Vs Recall:




- From the above plot, we can see that at value of approximately 0.5 there is a balance between Precision (Green) and Recall (Red)

- The accuracy score of our model on the Test Dataset is:  89.49 %
- The Sensitivity of our model on the test data set is: 89.34 %
- The Specificity of our model on the test data set is: 89.58 %
- From the above values, we can say that our model works good on the test data set too.
- Hence our Model can be used as a decision making tool to increase the conversion rate of leads.

# CONCLUSIONS AND SUGGESTIONS TO COMPANY

1. There is a positive correlation between 'Lead Add Form' and 'Converted' as evident by the positive coefficient 1.77.
   It suggests that, A person who fills the form can be considered a hot lead and we can persuade that person to increase the conversion of leads.
2. A positive coefficient (of 2.59) between 'Working Professional' and 'Converted' suggests that working professionals should be given due importance during the persuasion by sales team, as they have high chance of becoming hot leads and there by increase in conversion.
3. High positive coefficeint > 10 is observed in two variables 'Tags_closed by Horizzon' and 'Tags_Lost to EINS'. Focussing on these tags would increase the conversion from not converted to converted.
4. A negative coefficient of -2.711 and -3.41 for 'School Student' and 'Unknown Lead' respectively suggests that sales team should steer away from these customers during there calls.
5. The customers whose last notable activity is 'SMS Sent' should be the focus of the sales team, as they have higher chance of becoming 'hot leads' thereby increase in the conversion numbers.
6. The customers belonging to 'Other Citites' should also be the focus of sales team to make calls.