LEAD SCORE CASE STUDY

BY

SHUBHAM SINGH RANA

ROOPA H

Problem Statement

- X Education sells online courses to industry professionals.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%.

Business Goal

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned 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%.

Approaches

- Reading and understanding the dataset
- Checking the structures like shape, info and describe data
- Performing EDA
- Missing value check and Outlier check
- Data Pre-Processing, Splitting, Feature Scaling, Handling Multicollinearity within independent Features via Correlation
- Feature Selection and Model Building
- Here we are using ML model of Logistic Regression as it falls under Binary classification
- Optimal Probability Selection here we are using Precision – Recall, Trade-Off and Accuracy, Sensitivity & Specificity Trade-Off

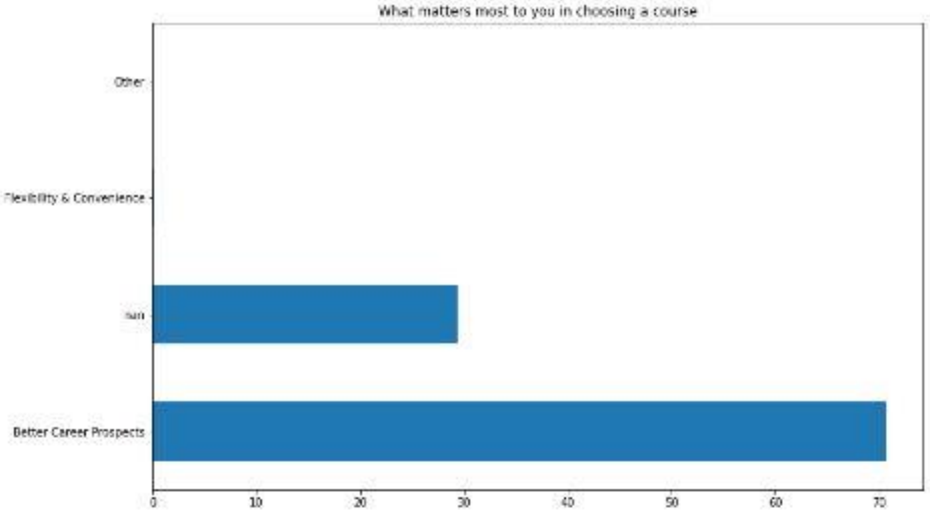
Analysis

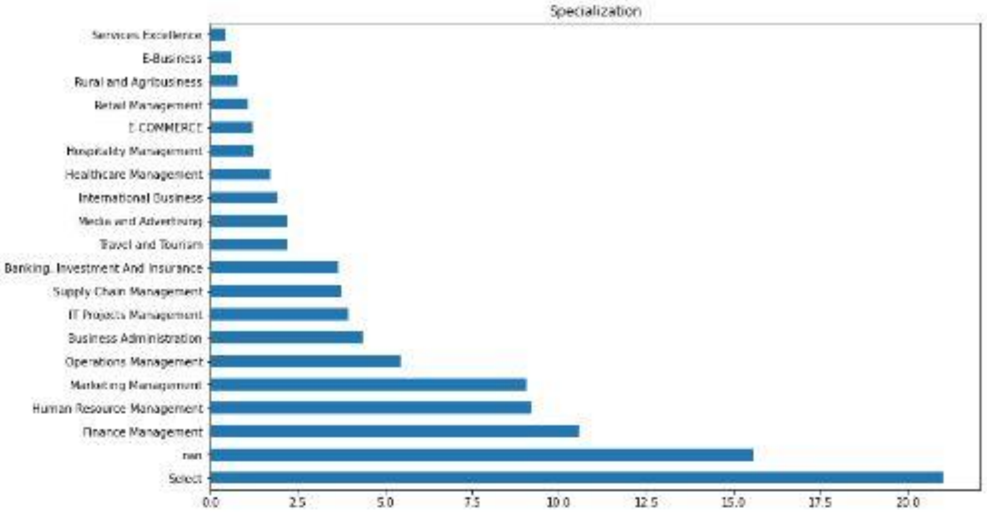
- We have found 9240 rows and 37 columns after reading the dataset
- We got many insights regarding features with Missing Values, Categorical features, Numerical features , Binary Features of Categorical & Numeric and their Data Imbalance after performing EDA.
- After EDA, we got to know that 10 Features have Missing Values more than 35% & 5 features have Missing Values less than 35%. For Features with less than 35% Missing Values, we have imputed their values accordingly.
- We dropped features on the basis of \`More than 35% of Missing Values\`, \`High Multicollinearity\` & \`High Data Imbalance for Binary Categorical Features\`.
- 7 Categorical Features had few categories which were having data less than 5%. Therefore, we merged them into a single category named \`Other\`.

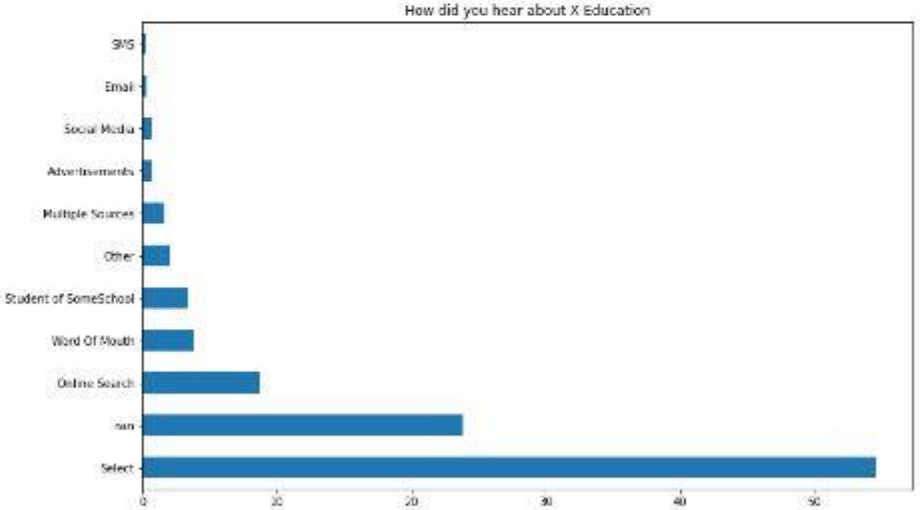
Analysis

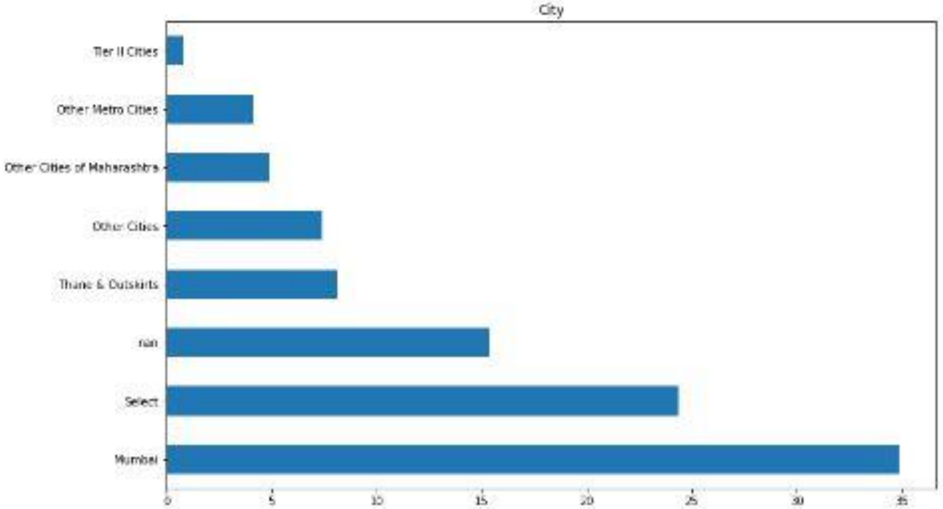
- 5 Binary Categorical Features had Very High Data Imbalance, therefore we dropped those features.
- High Multicollinearity was found in 4 features. Therefore, we dropped them as well.
- And after dropping we are left with 19 features in the dataset for modelling.
- While checking for the outliers we observed that there are outliers in Page Views Per Visit & Total Visits. Therefore, we dropped data which belonged to more than 99 Percentile of these features.
- And all the plots are provided in the next slide

PLOTS

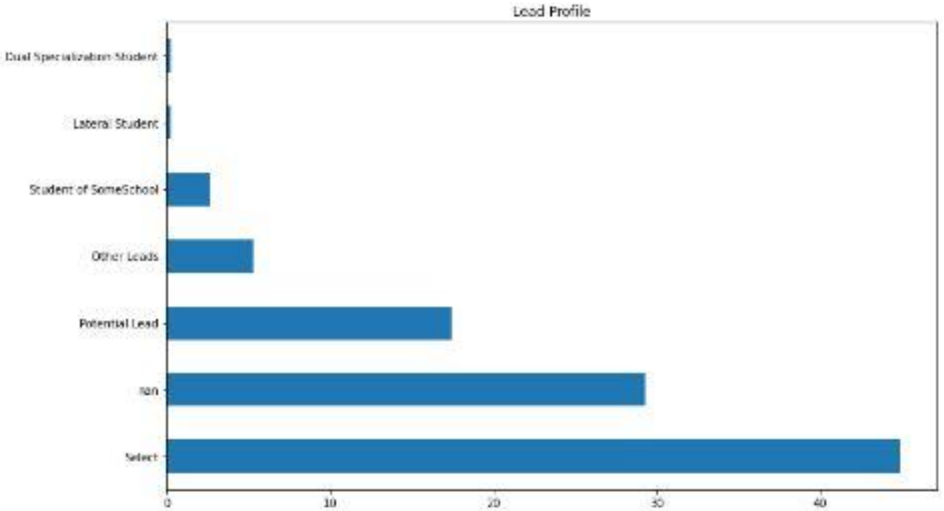

What matters most to you in choosing a course
Other
Flexibility & Convenience
nan
Better Career Prospects
0
10
20
30
40
50
60
70

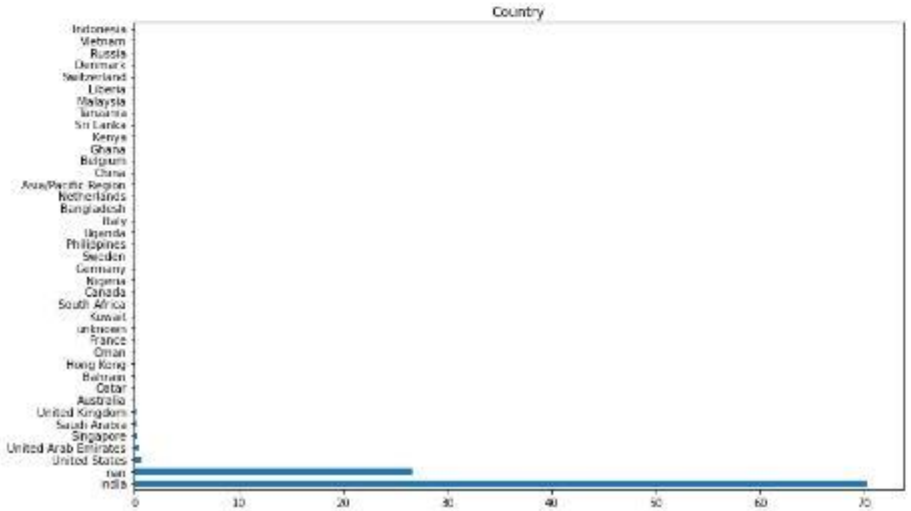

Specialization
Services Excellence
E-Business
Rural and Agribusiness
Retail Management
E-COMMERCE
Hospitality Management
Healthcare Management
International Business
Media and Advertising
Travel and Tourism
Banking, Investment And Insurance
Supply Chain Management
IT Projects Management
Business Administration
Operations Management
Marketing Management
Human Resource Management
Finance Management
nan
Select
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

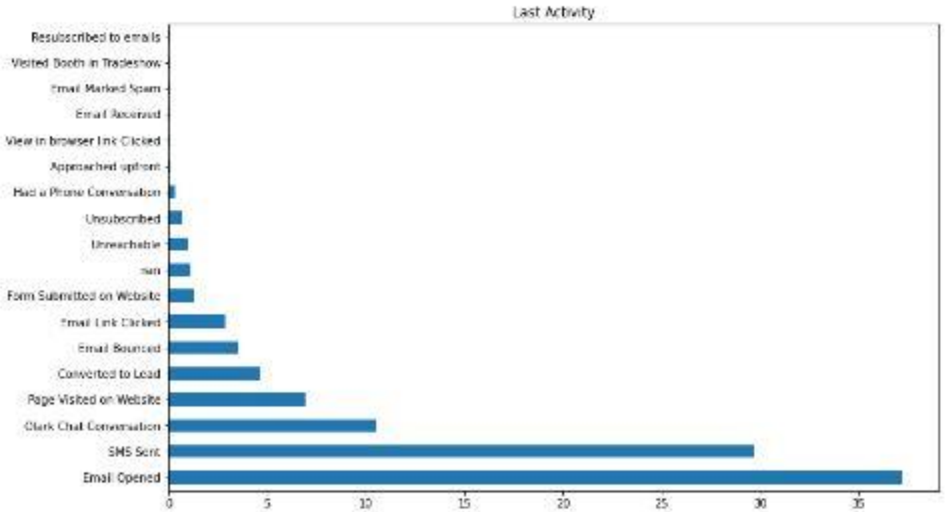

How did you hear about X Education
SMS
Email
Social Media
Advertisements
Multiple Sources
Other
Student of SomeSchool
Word Of Mouth
Online Search
nan
Select
0
10
20
30
40
50

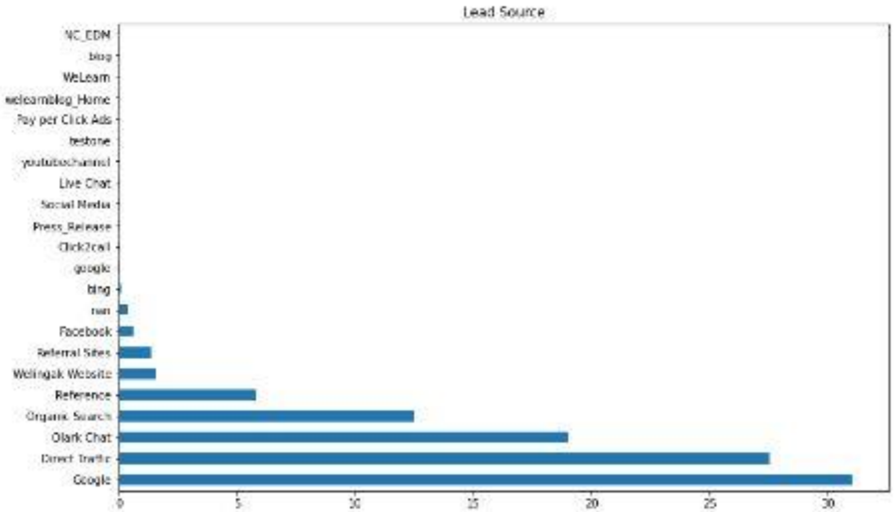

City
Tier II Cities
Other Metro Cities
Other Cities of Maharashtra
Other Cities
Thane & Outskirts
nan
Select
Mumbai
0
5
10
15
20
25
30
35

PLOTS

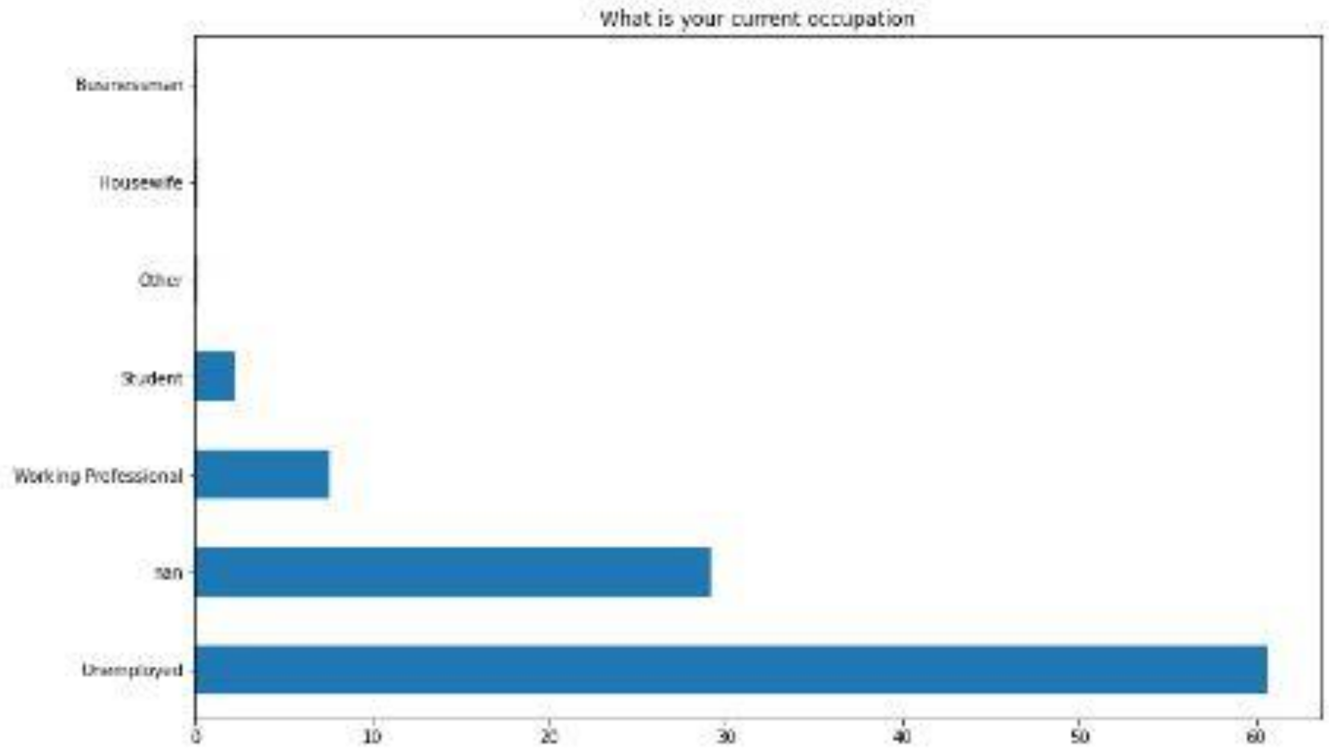

Lead Profile


Country


Last Activity


Lead Source

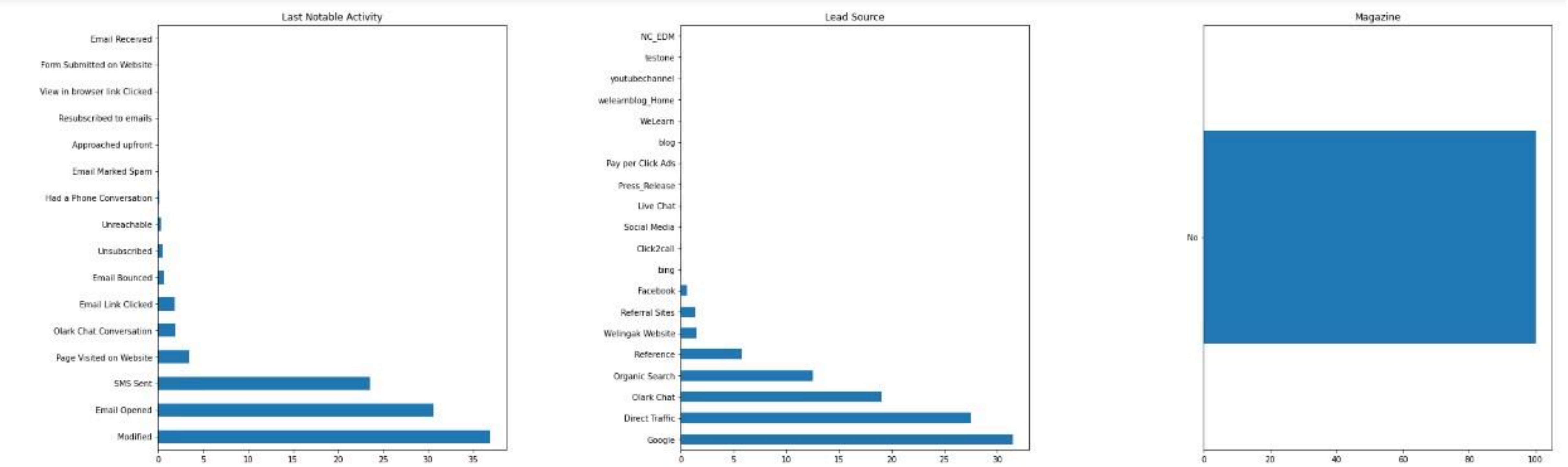
PLOTS

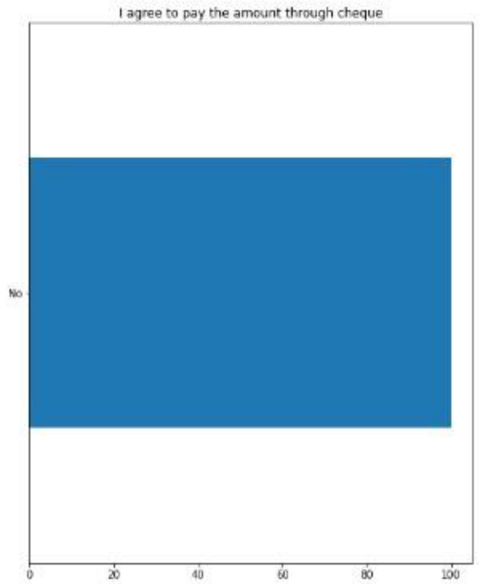


Points to be noted from above visualizations:

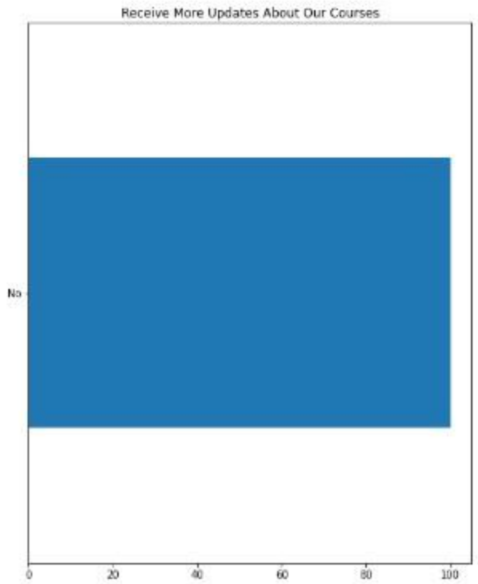
- It can be seen that many features have Missing Values more than 50% again after combining **`Select`** & **`NaN`** Values. Let's again check the features having less than & more than 35% of Missing Values respectively.
- Also, in **`Lead Source`** feature, we can see that there are two different values named `Google` & `google`. Let's make them one as well.

PLOTS

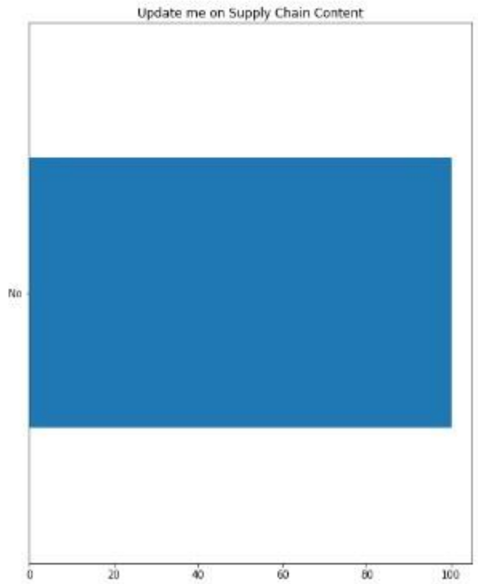

Last Notable Activity
Email Received
Form Submitted on Website
View in browser link Clicked
Resubscribed to emails
Approached upfront
Email Marked Spam
Had a Phone Conversation
Unreachable
Unsubscribed
Email Bounced
Email Link Clicked
Olark Chat Conversation
Page Visited on Website
SMS Sent
Email Opened
Modified
0
5
10
15
20
25
30
35
Lead Source
NC_EDM
testone
youtubechannel
welearnblog_Home
WeLearn
blog
Pay per Click Ads
Press_Release
Live Chat
Social Media
Click2call
bing
Facebook
Referral Sites
Welingak Website
Reference
Organic Search
Olark Chat
Direct Traffic
Google
0
5
10
15
20
25
30
Magazine
No
0
20
40
60
80
100



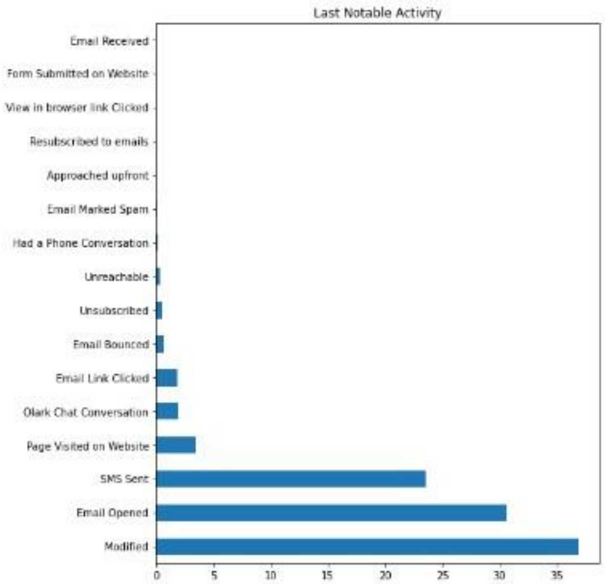
I agree to pay the amount through cheque



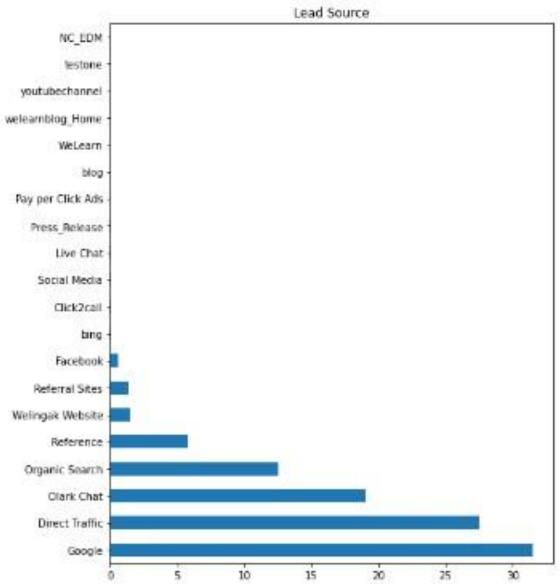
Receive More Updates About Our Courses



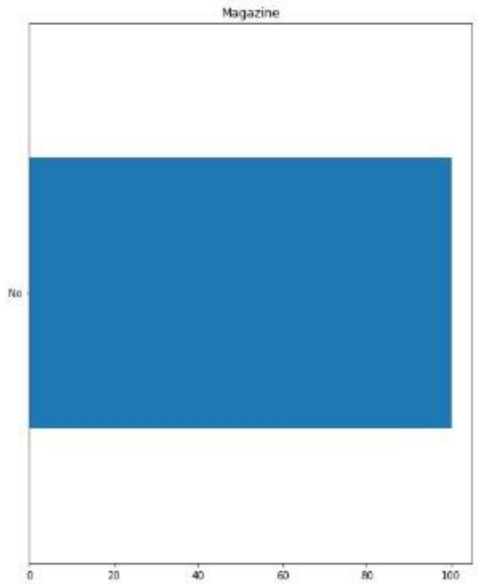
Update me on Supply Chain Content



Last Notable Activity

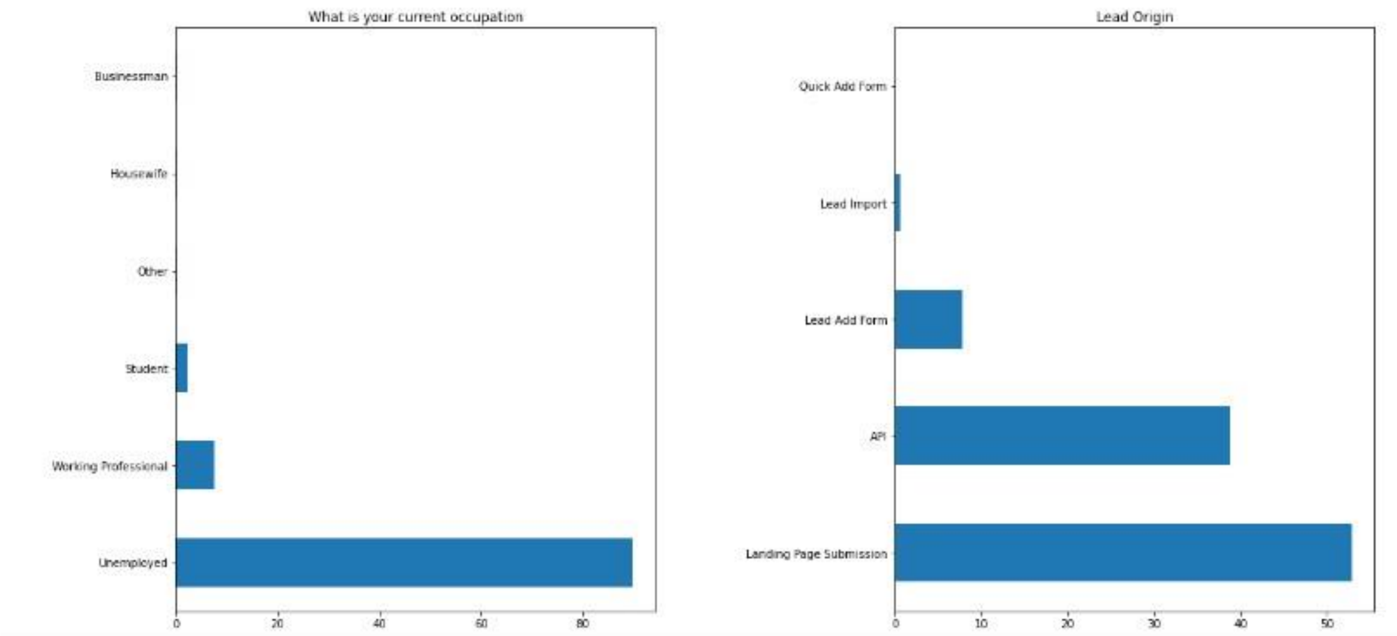


Lead Source



Magazine

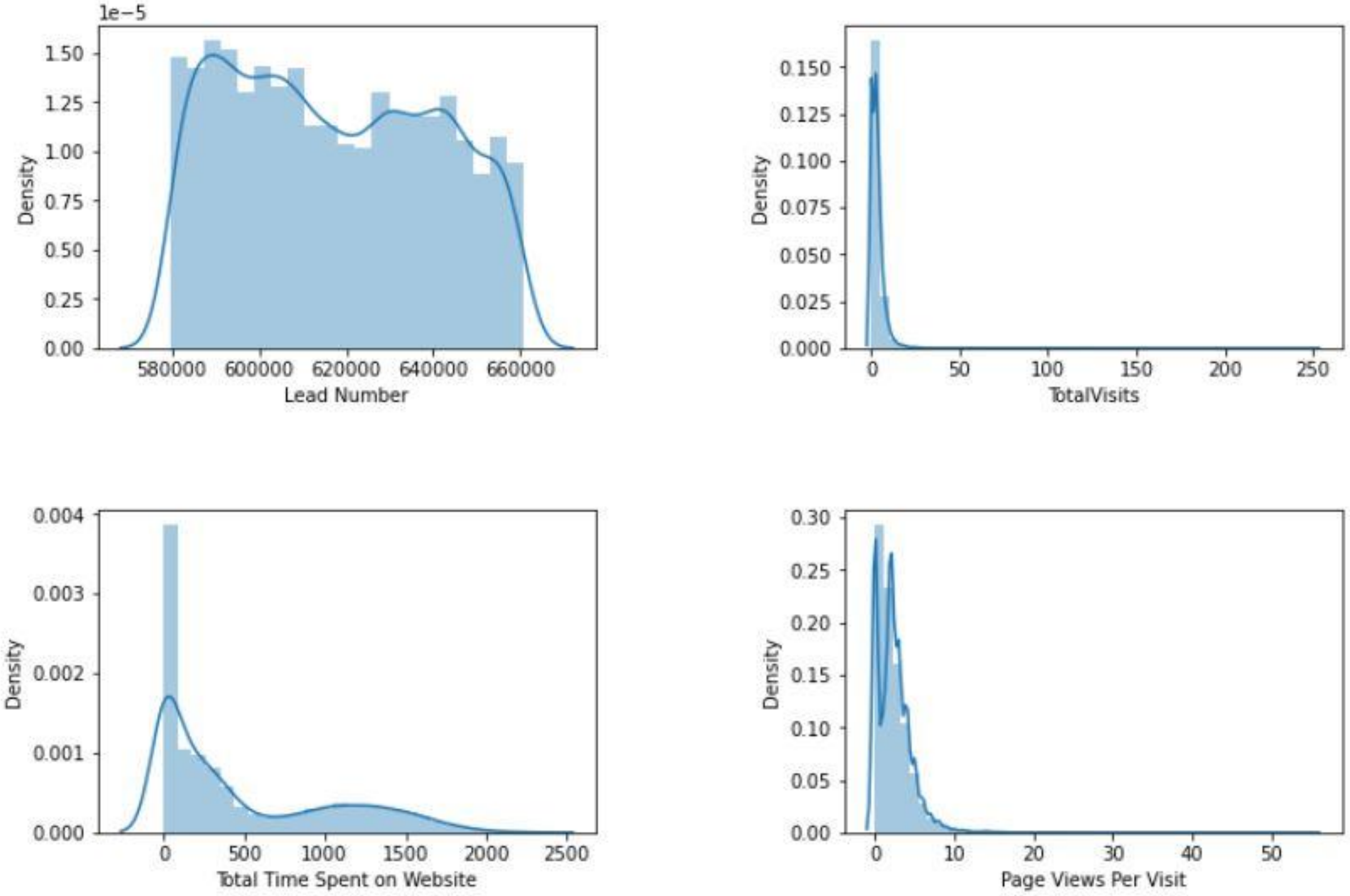
PLOTS



Points to be noted from above Categorical Feature Analysis:(considering 10th 11th and 12th slides)_

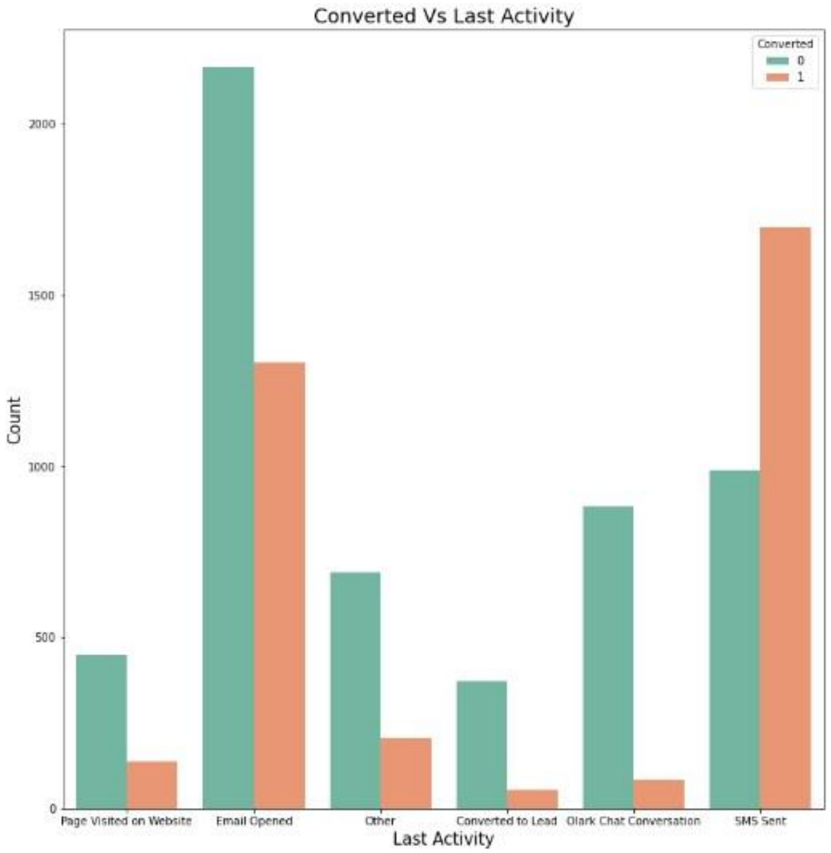
- `Lead Source`, `Country`, `Last Activity`, `What matters most to you in choosing a course`, `Last Notable Activity`, `What is your current Occupation` & `Lead Origin` have multiple categories with very less amount of data. Therefore, we can combine such categories into `Others`.
- Features such as `Receive More Updates About Our Courses`, `I agree to pay the amount through cheque`, `Get updates on DM Content`, `Update me on Supply Chain Content` & `Magazine` have only one category which will not add any value to make classification. Therefore, we will be dropping these features.
- Also `Prospect ID` Feature is having unique Id's for each customer/lead. Therefore, we will drop this as well.

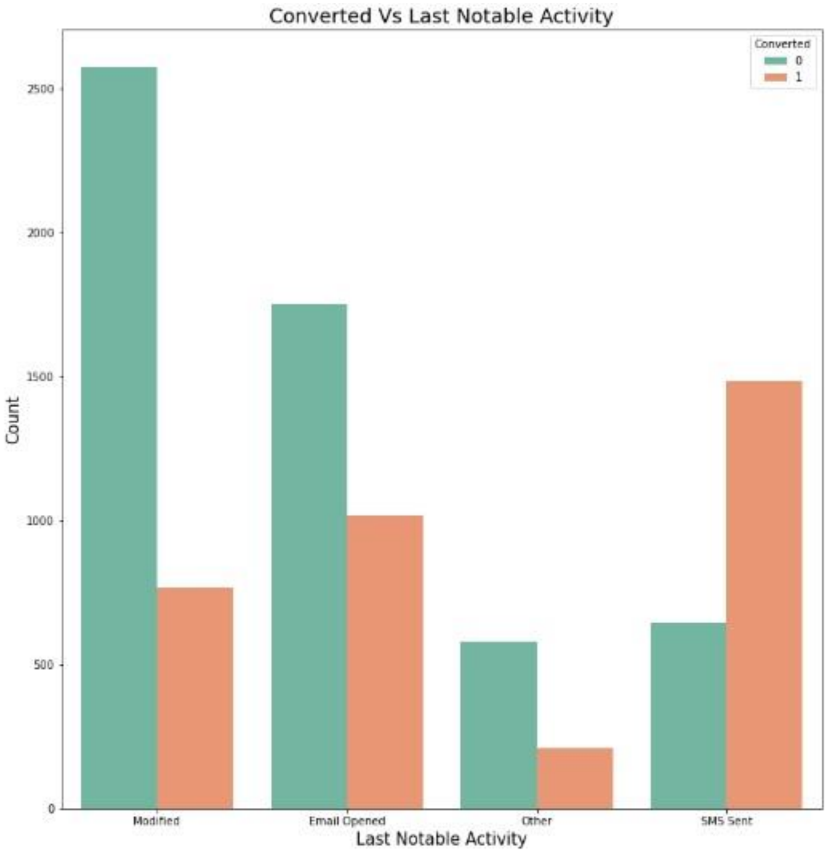
PLOTS



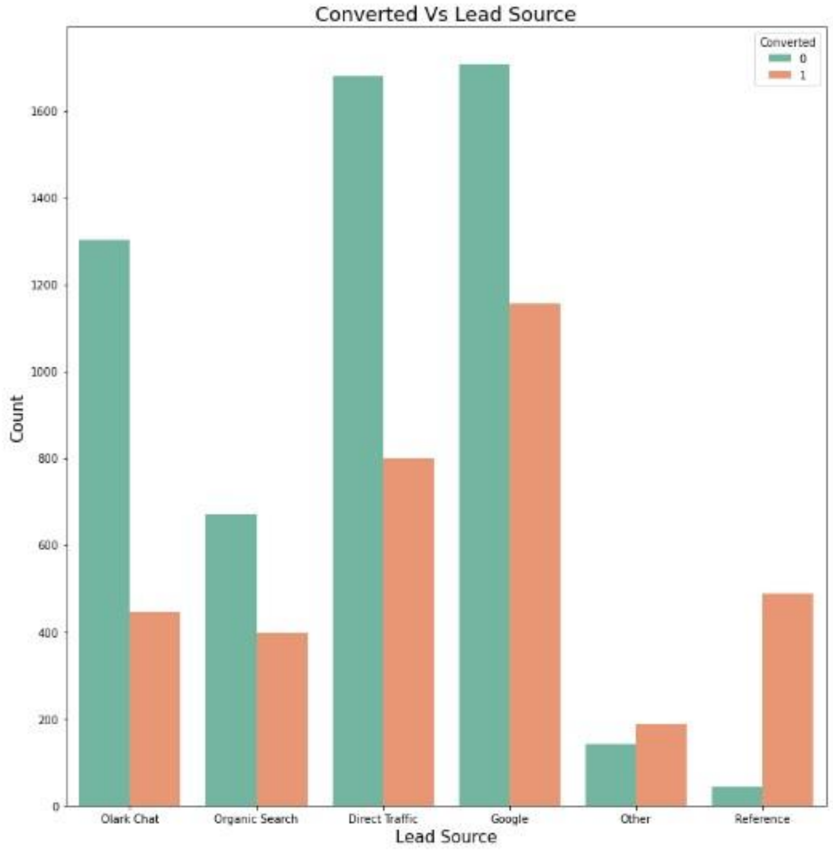
From the given visualizations, we can observe that there is definitely a need for Feature Scaling which needs to be applied on Numeric Features such as `TotalVisits`, `Total Time Spent on Website` & `Page Views Per Visit` which are highly skewed.

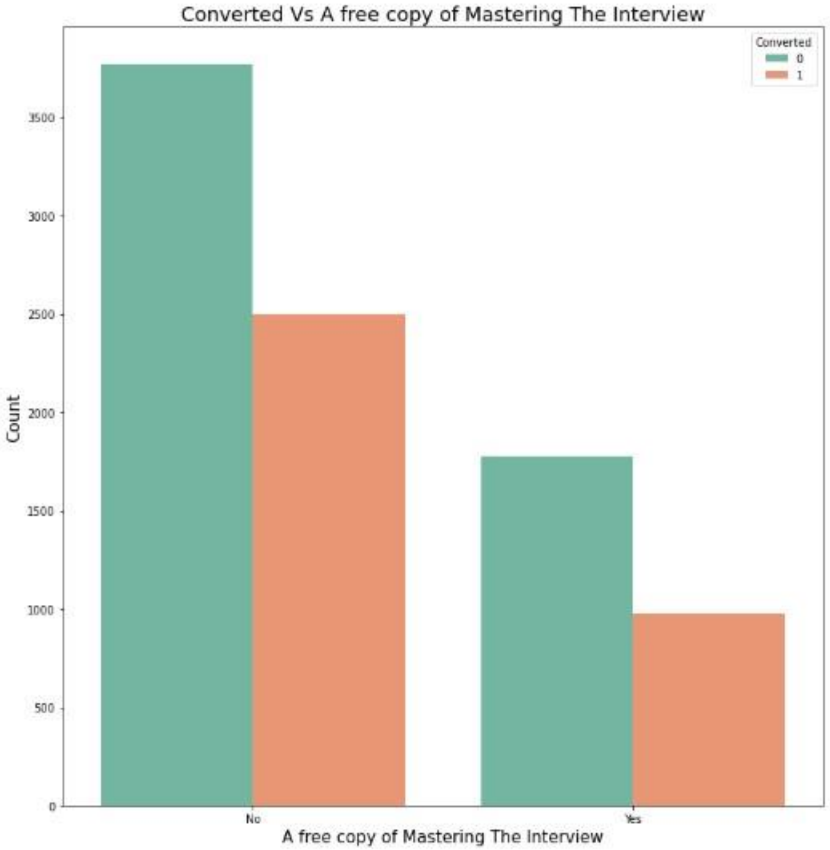
PLOTS


Converted Vs Last Activity
Converted
0
1
Count
0
500
1000
1500
2000
Page Visited on Website
Email Opened
Other
Converted to Lead
Olark Chat Conversation
SMS Sent
Last Activity

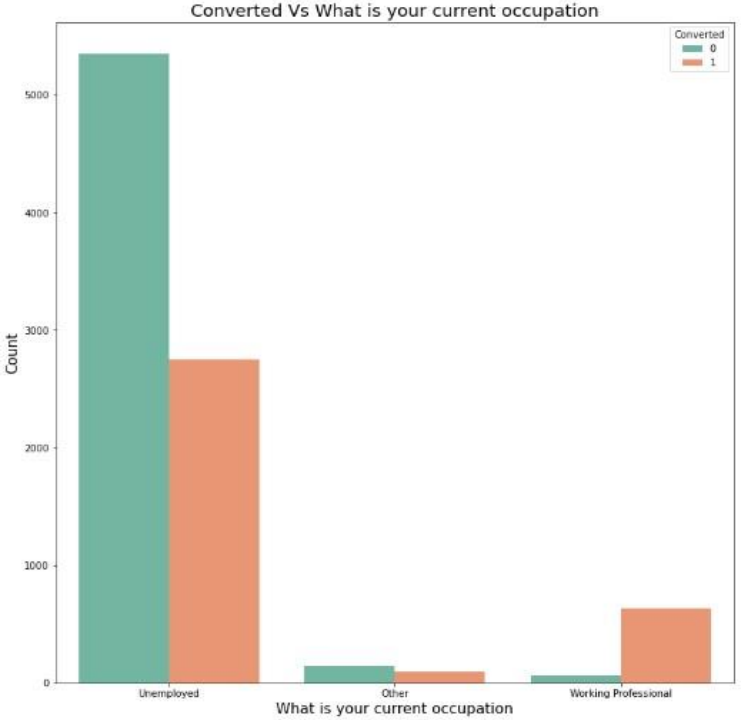

Converted Vs Last Notable Activity
Converted
0
1
Count
0
500
1000
1500
2000
2500
Modified
Email Opened
Other
SMS Sent
Last Notable Activity

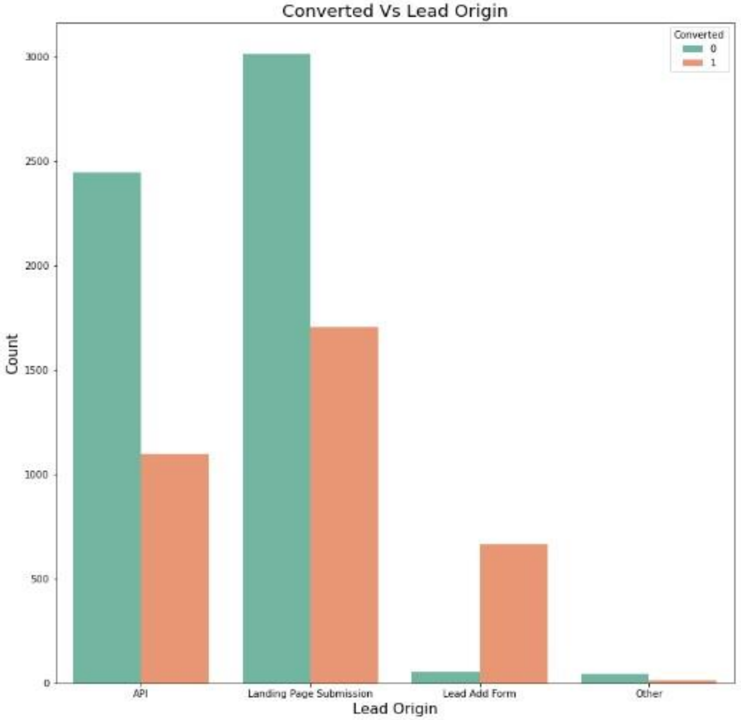
PLOTS


Converted Vs Lead Source
Converted
🟩 0
🟧 1
Count
0
200
400
600
800
1000
1200
1400
1600
Olark Chat
Organic Search
Direct Traffic
Google
Other
Reference
Lead Source


Converted Vs A free copy of Mastering The Interview
Converted
🟩 0
🟧 1
Count
0
500
1000
1500
2000
2500
3000
3500
No
Yes
A free copy of Mastering The Interview

PLOTS

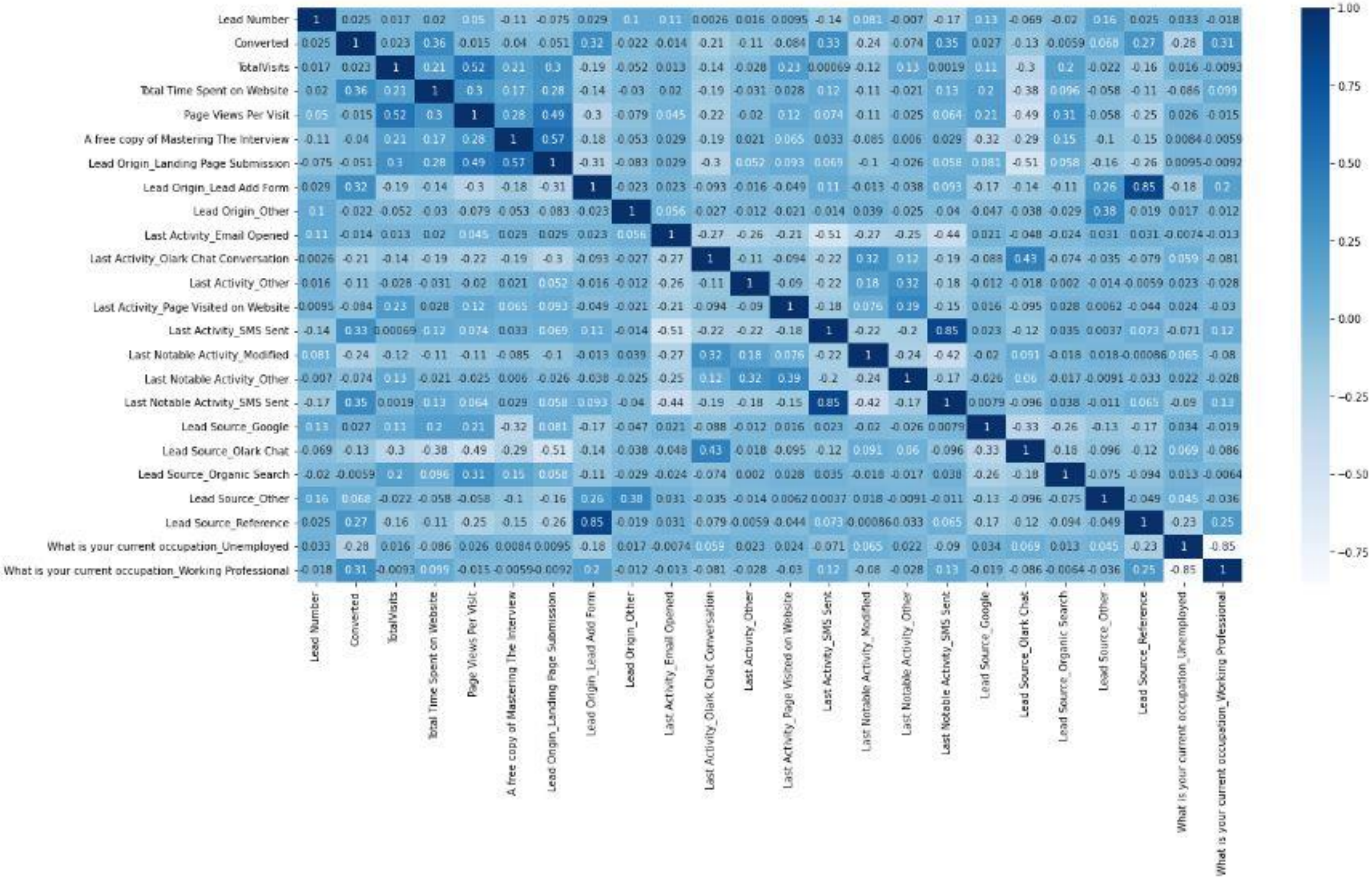




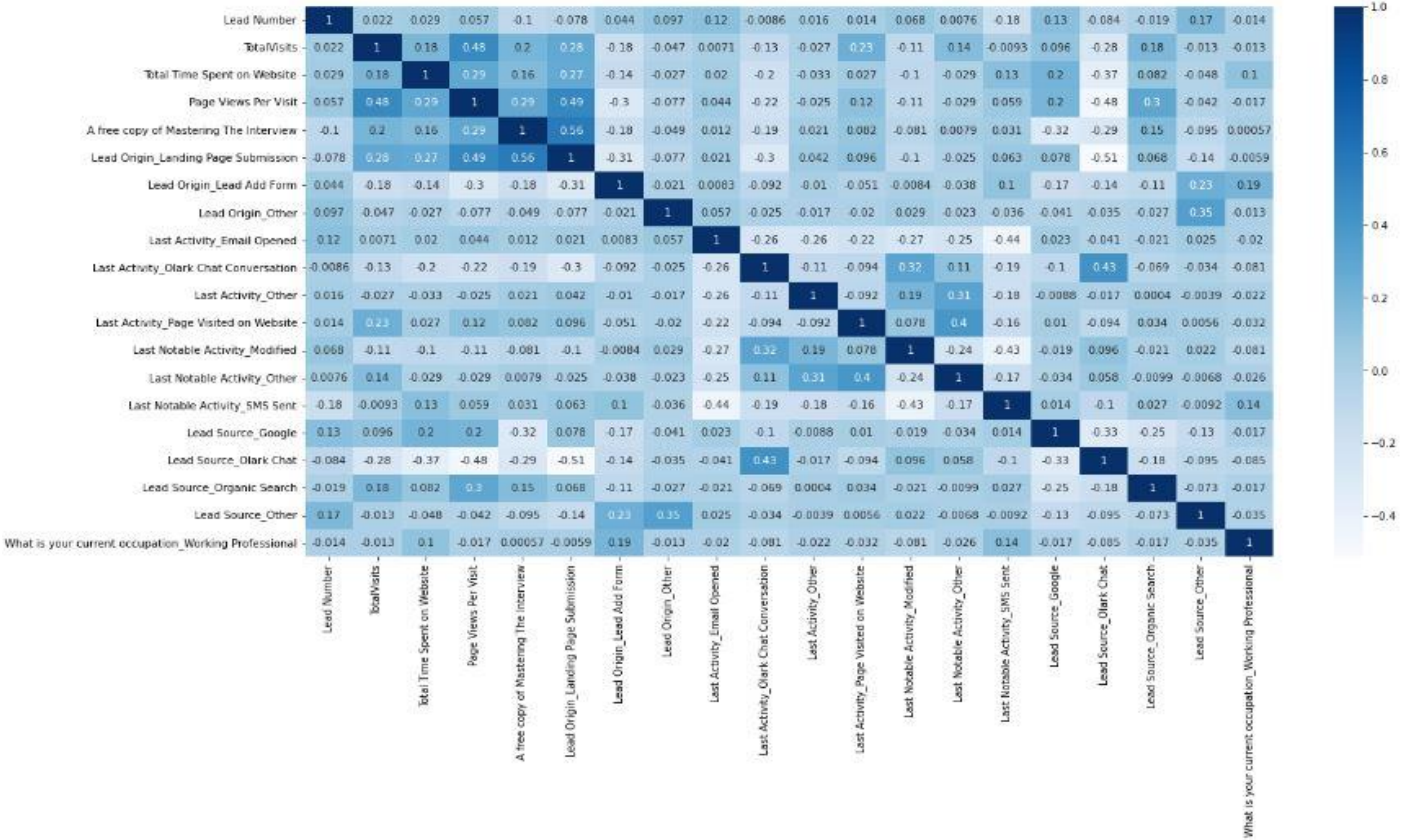
Following Points can be made from the analysis:(considering slide 14th , 15th and 16th)

- For `A free copy of Mastering The Interview`, it can be seen that the maximum conversion happened from the customers who do not want a free copy of Mastering the Interview.
- For `Last Notable Activity`, it can be seen that the maximum conversion happened from `SMS Sent`.
- For `What is your current occupation`, it can be seen that the maximum conversion happened from `Unemployed` whereas from `Working Professional` group there are more Successful Conversions than Non-Conversions.
- For `Last Activity`, it can be seen that the maximum conversion happened from `SMS Sent` whereas from `Email Opened` we have the highest amount of Non-Conversions.

PLOTS


Lead Number 1 0.025 0.017 0.02 0.05 -0.11 -0.075 0.029 0.1 0.11 0.0026 0.016 0.0095 -0.14 0.081 -0.007 -0.17 0.13 -0.069 -0.02 0.16 0.025 0.033 -0.018
Converted 0.025 1 0.023 0.36 -0.015 -0.04 -0.051 0.32 -0.022 -0.014 -0.21 -0.11 -0.084 0.33 -0.24 -0.074 0.35 0.027 -0.13 -0.0059 0.068 0.27 -0.28 0.31
TotalVisits 0.017 0.023 1 0.21 0.52 0.21 0.3 -0.19 -0.052 0.013 -0.14 -0.028 0.23 0.00069 -0.12 0.11 0.0019 0.11 -0.3 0.2 -0.022 -0.16 0.016 -0.0093
Total Time Spent on Website 0.02 0.36 0.21 1 0.3 0.17 0.28 -0.14 -0.03 0.02 -0.19 -0.031 0.028 0.12 -0.11 -0.021 0.13 0.2 -0.38 0.096 -0.058 -0.11 -0.086 0.099
Page Views Per Visit 0.05 -0.015 0.52 0.3 1 0.28 0.49 -0.3 -0.079 0.045 -0.22 -0.02 0.12 0.074 -0.11 -0.025 0.064 0.21 -0.49 0.31 -0.058 -0.25 0.026 -0.015
A free copy of Mastering The Interview -0.11 -0.04 0.21 0.17 0.28 1 0.57 -0.18 -0.053 0.029 -0.19 0.021 0.065 0.033 -0.085 0.006 0.029 -0.32 -0.29 0.15 -0.1 -0.15 0.0084 -0.0059
Lead Origin_Landing Page Submission -0.075 -0.051 0.3 0.28 0.49 0.57 1 -0.31 -0.083 0.029 -0.3 0.052 0.093 0.069 -0.1 -0.026 0.058 0.081 -0.51 0.058 -0.16 -0.26 0.0095 -0.0092
Lead Origin_Lead Add Form 0.029 0.32 -0.19 -0.14 -0.3 -0.18 -0.31 1 -0.023 0.023 -0.093 -0.016 -0.049 0.11 -0.013 -0.038 0.093 -0.17 -0.14 -0.11 0.26 0.85 -0.18 0.2
Lead Origin_Other 0.1 -0.022 -0.052 -0.03 -0.079 -0.053 -0.083 -0.023 1 0.056 -0.027 -0.012 -0.021 -0.014 0.039 -0.025 -0.04 -0.047 -0.038 -0.029 0.38 -0.019 0.017 -0.012
Last Activity_Email Opened 0.11 -0.014 0.013 0.02 0.045 0.029 0.029 0.023 0.056 1 -0.27 -0.26 -0.21 -0.51 -0.27 -0.25 -0.44 0.021 -0.048 -0.024 0.031 0.031 -0.0074 -0.013
Last Activity_Olark Chat Conversation 0.0026 -0.21 -0.14 -0.19 -0.22 -0.19 -0.3 -0.093 -0.027 -0.27 1 -0.11 -0.094 -0.22 0.32 0.12 -0.19 -0.088 0.43 -0.074 -0.035 -0.079 0.059 -0.081
Last Activity_Other 0.016 -0.11 -0.028 -0.031 -0.02 0.021 0.052 -0.016 -0.012 -0.26 -0.11 1 -0.09 -0.22 0.18 0.32 -0.18 -0.012 -0.018 0.002 -0.014 -0.0059 0.023 -0.028
Last Activity_Page Visited on Website 0.0095 -0.084 0.23 0.028 0.12 0.065 0.093 -0.049 -0.021 -0.21 -0.094 -0.09 1 -0.18 0.076 0.39 -0.15 0.016 -0.095 0.028 0.0062 -0.044 0.024 -0.03
Last Activity_SMS Sent -0.14 0.33 0.00069 0.12 0.074 0.033 0.069 0.11 -0.014 -0.51 -0.22 -0.22 -0.18 1 -0.22 -0.2 0.85 0.023 -0.12 0.035 0.0037 0.073 -0.071 0.12
Last Notable Activity_Modified 0.081 -0.24 -0.12 -0.11 -0.11 -0.085 -0.1 -0.013 0.039 -0.27 0.32 0.18 0.076 -0.22 1 -0.24 -0.42 -0.02 0.091 -0.018 0.018 -0.00086 0.065 -0.08
Last Notable Activity_Other -0.007 -0.074 0.11 -0.021 -0.025 0.006 -0.026 -0.038 -0.025 -0.25 0.12 0.32 0.39 -0.2 -0.24 1 -0.17 -0.026 0.06 -0.017 -0.0091 -0.033 0.022 -0.028
Last Notable Activity_SMS Sent -0.17 0.35 0.0019 0.13 0.064 0.029 0.058 0.093 -0.04 -0.44 -0.19 -0.18 -0.15 0.85 -0.42 -0.17 1 0.0079 -0.096 0.038 -0.011 0.065 -0.09 0.13
Lead Source_Google 0.13 0.027 0.11 0.2 0.21 -0.32 0.081 -0.17 -0.047 0.021 -0.088 -0.012 0.016 0.023 -0.02 -0.026 0.0079 1 -0.33 -0.26 -0.13 -0.17 0.034 -0.019
Lead Source_Olark Chat -0.069 -0.13 -0.3 -0.38 -0.49 -0.29 -0.51 -0.14 -0.038 -0.048 0.43 -0.018 -0.095 -0.12 0.091 0.06 -0.096 -0.33 1 -0.18 -0.096 -0.12 0.069 -0.086
Lead Source_Organic Search -0.02 -0.0059 0.2 0.096 0.31 0.15 0.058 -0.11 -0.029 -0.024 -0.074 0.002 0.028 0.035 -0.018 -0.017 0.038 -0.26 -0.18 1 -0.075 -0.094 0.013 -0.0064
Lead Source_Other 0.16 0.068 -0.022 -0.058 -0.058 -0.1 -0.16 0.26 0.38 0.031 -0.035 -0.014 0.0062 0.0037 0.018 -0.0091 -0.011 -0.13 -0.096 -0.075 1 -0.049 0.045 -0.036
Lead Source_Reference 0.025 0.27 -0.16 -0.11 -0.25 -0.15 -0.26 0.85 -0.019 0.031 -0.079 -0.0059 -0.044 0.073 -0.00086 -0.033 0.065 -0.17 -0.12 -0.094 -0.049 1 -0.23 0.25
What is your current occupation_Unemployed 0.033 -0.28 0.016 -0.086 0.026 0.0084 0.0095 -0.18 0.017 -0.0074 0.059 0.023 0.024 -0.071 0.065 0.022 -0.09 0.034 0.069 0.013 0.045 -0.23 1 -0.85
What is your current occupation_Working Professional -0.018 0.31 -0.0093 0.099 -0.015 -0.0059 -0.0092 0.2 -0.012 -0.013 -0.081 -0.028 -0.03 0.12 -0.08 -0.028 0.13 -0.019 -0.086 -0.0064 -0.036 0.25 -0.85 1
Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
A free copy of Mastering The Interview
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Origin_Other
Last Activity_Email Opened
Last Activity_Olark Chat Conversation
Last Activity_Other
Last Activity_Page Visited on Website
Last Activity_SMS Sent
Last Notable Activity_Modified
Last Notable Activity_Other
Last Notable Activity_SMS Sent
Lead Source_Google
Lead Source_Olark Chat
Lead Source_Organic Search
Lead Source_Other
Lead Source_Reference
What is your current occupation_Unemployed
What is your current occupation_Working Professional
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

PLOTS

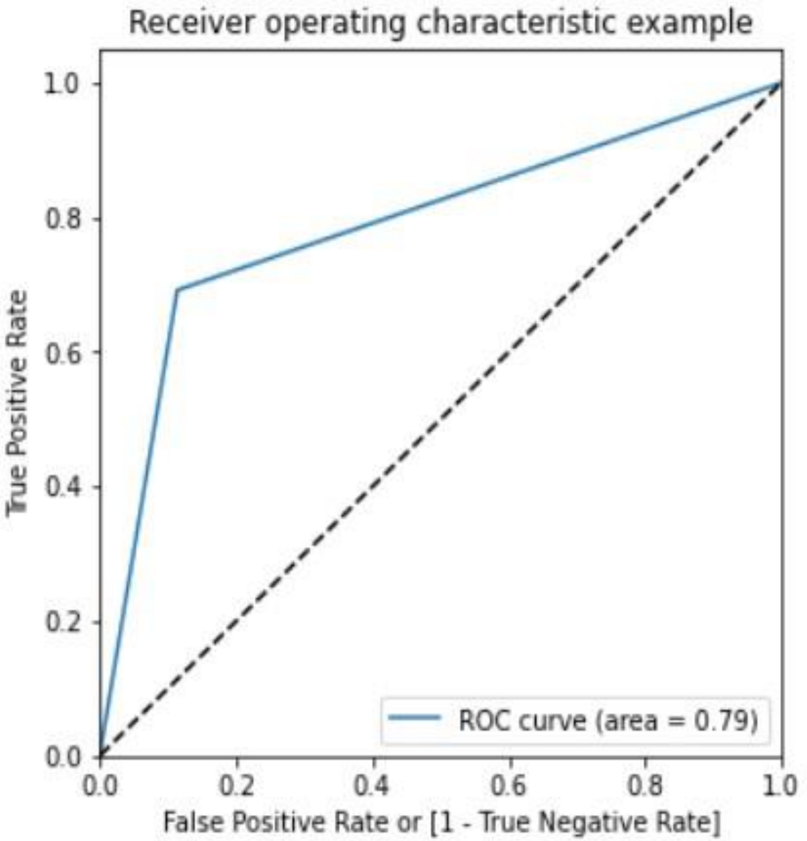

Lead Number 1 0.022 0.029 0.057 -0.1 -0.078 0.044 0.097 0.12 -0.0086 0.016 0.014 0.068 0.0076 -0.18 0.13 -0.084 -0.019 0.17 -0.014
TotalVisits 0.022 1 0.18 0.48 0.2 0.28 -0.18 -0.047 0.0071 -0.13 -0.027 0.23 -0.11 0.14 -0.0093 0.096 -0.28 0.18 -0.013 -0.013
Total Time Spent on Website 0.029 0.18 1 0.29 0.16 0.27 -0.14 -0.027 0.02 -0.2 -0.033 0.027 -0.1 -0.029 0.13 0.2 -0.37 0.082 -0.048 0.1
Page Views Per Visit 0.057 0.48 0.29 1 0.29 0.49 -0.3 -0.077 0.044 -0.22 -0.025 0.12 -0.11 -0.029 0.059 0.2 -0.48 0.3 -0.042 -0.017
A free copy of Mastering The Interview -0.1 0.2 0.16 0.29 1 0.56 -0.18 -0.049 0.012 -0.19 0.021 0.082 -0.081 0.0079 0.031 -0.32 -0.29 0.15 -0.095 0.00057
Lead Origin_Landing Page Submission -0.078 0.28 0.27 0.49 0.56 1 -0.31 -0.077 0.021 -0.3 0.042 0.096 -0.1 -0.025 0.063 0.078 -0.51 0.068 -0.14 -0.0059
Lead Origin_Lead Add Form 0.044 -0.18 -0.14 -0.3 -0.18 -0.31 1 -0.021 0.0083 -0.092 -0.01 -0.051 -0.0084 -0.038 0.1 -0.17 -0.14 -0.11 0.23 0.19
Lead Origin_Other 0.097 -0.047 -0.027 -0.077 -0.049 -0.077 -0.021 1 0.057 -0.025 -0.017 -0.02 0.029 -0.023 -0.036 -0.041 -0.035 -0.027 0.35 -0.013
Last Activity_Email Opened 0.12 0.0071 0.02 0.044 0.012 0.021 0.0083 0.057 1 -0.26 -0.26 -0.22 -0.27 -0.25 -0.44 0.023 -0.041 -0.021 0.025 -0.02
Last Activity_Olark Chat Conversation -0.0086 -0.13 -0.2 -0.22 -0.19 -0.3 -0.092 -0.025 -0.26 1 -0.11 -0.094 0.32 0.11 -0.19 -0.1 0.43 -0.069 -0.034 -0.081
Last Activity_Other 0.016 -0.027 -0.033 -0.025 0.021 0.042 -0.01 -0.017 -0.26 -0.11 1 -0.092 0.19 0.31 -0.18 -0.0088 -0.017 0.0004 -0.0039 -0.022
Last Activity_Page Visited on Website 0.014 0.23 0.027 0.12 0.082 0.096 -0.051 -0.02 -0.22 -0.094 -0.092 1 0.078 0.4 -0.16 0.01 -0.094 0.034 0.0056 -0.032
Last Notable Activity_Modified 0.068 -0.11 -0.1 -0.11 -0.081 -0.1 -0.0084 0.029 -0.27 0.32 0.19 0.078 1 -0.24 -0.43 -0.019 0.096 -0.021 0.022 -0.081
Last Notable Activity_Other 0.0076 0.14 -0.029 -0.029 0.0079 -0.025 -0.038 -0.023 -0.25 0.11 0.31 0.4 -0.24 1 -0.17 -0.034 0.058 -0.0099 -0.0068 -0.026
Last Notable Activity_SMS Sent -0.18 -0.0093 0.13 0.059 0.031 0.063 0.1 -0.036 -0.44 -0.19 -0.18 -0.16 -0.43 -0.17 1 0.014 -0.1 0.027 -0.0092 0.14
Lead Source_Google 0.13 0.096 0.2 0.2 -0.32 0.078 -0.17 -0.041 0.023 -0.1 -0.0088 0.01 -0.019 -0.034 0.014 1 -0.33 -0.25 -0.13 -0.017
Lead Source_Olark Chat -0.084 -0.28 -0.37 -0.48 -0.29 -0.51 -0.14 -0.035 -0.041 0.43 -0.017 -0.094 0.096 0.058 -0.1 -0.33 1 -0.18 -0.095 -0.085
Lead Source_Organic Search -0.019 0.18 0.082 0.3 0.15 0.068 -0.11 -0.027 -0.021 -0.069 0.0004 0.034 -0.021 -0.0099 0.027 -0.25 -0.18 1 -0.073 -0.017
Lead Source_Other 0.17 -0.013 -0.048 -0.042 -0.095 -0.14 0.23 0.35 0.025 -0.034 -0.0039 0.0056 0.022 -0.0068 -0.0092 -0.13 -0.095 -0.073 1 -0.035
What is your current occupation_Working Professional -0.014 -0.013 0.1 -0.017 0.00057 -0.0059 0.19 -0.013 -0.02 -0.081 -0.022 -0.032 -0.081 -0.026 0.14 -0.017 -0.085 -0.017 -0.035 1
Lead Number
TotalVisits
Total Time Spent on Website
Page Views Per Visit
A free copy of Mastering The Interview
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Origin_Other
Last Activity_Email Opened
Last Activity_Olark Chat Conversation
Last Activity_Other
Last Activity_Page Visited on Website
Last Notable Activity_Modified
Last Notable Activity_Other
Last Notable Activity_SMS Sent
Lead Source_Google
Lead Source_Olark Chat
Lead Source_Organic Search
Lead Source_Other
What is your current occupation_Working Professional
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

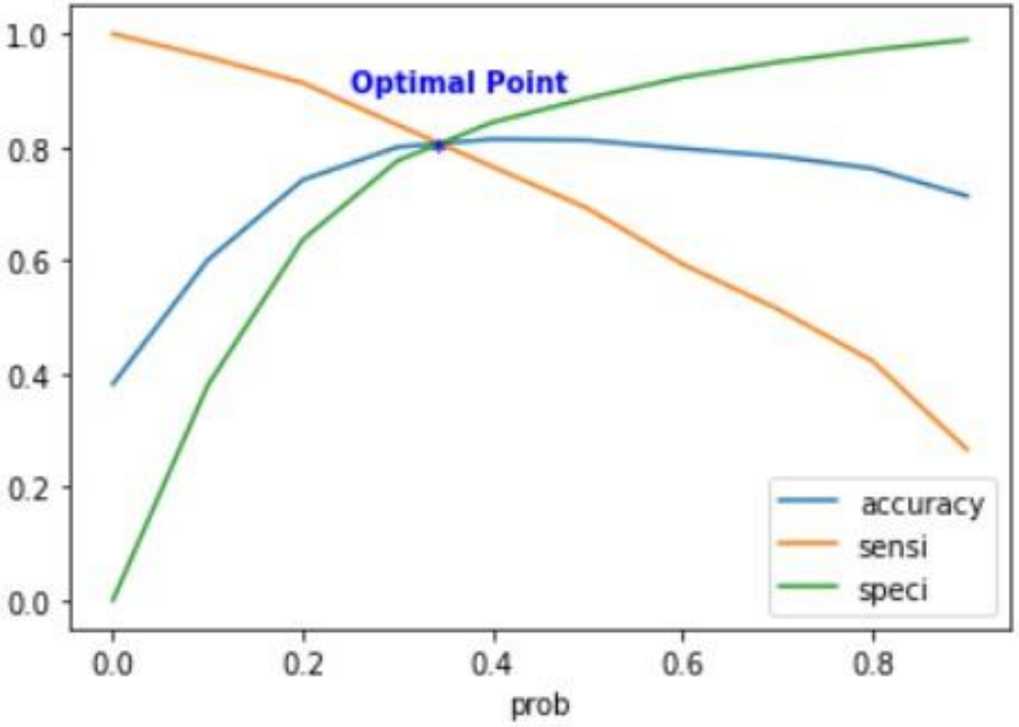
PLOTS

From the above two plots(considering 17th and 18th slide):

- TotalVisits -> ['Page Views Per Visit’]
- Lead Source_Reference -> ['Lead Origin_Lead Add Form’]
- Last Activity_SMS Sent -> ['Last Notable Activity_SMS Sent’]
- What is your current occupation_Unemployed -> ['What is your current occupation_Working Professional’]
- All the above are highly correlated and we are dropping them
- By considering slide 18th visualization we found It can be seen that now no Independent Feature is highly inter-correlated.

PLOT OF ROC CURVE & OPTIMAL POINT

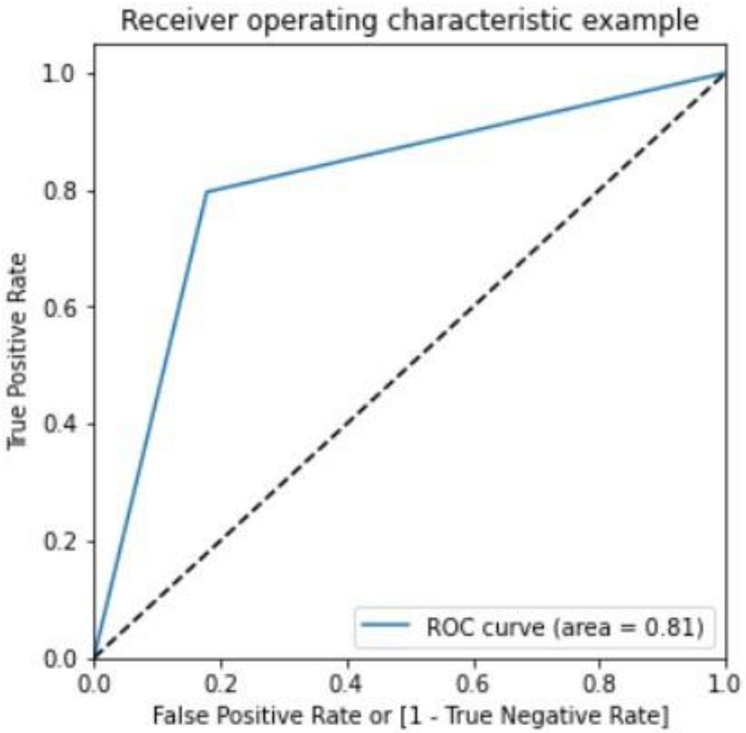


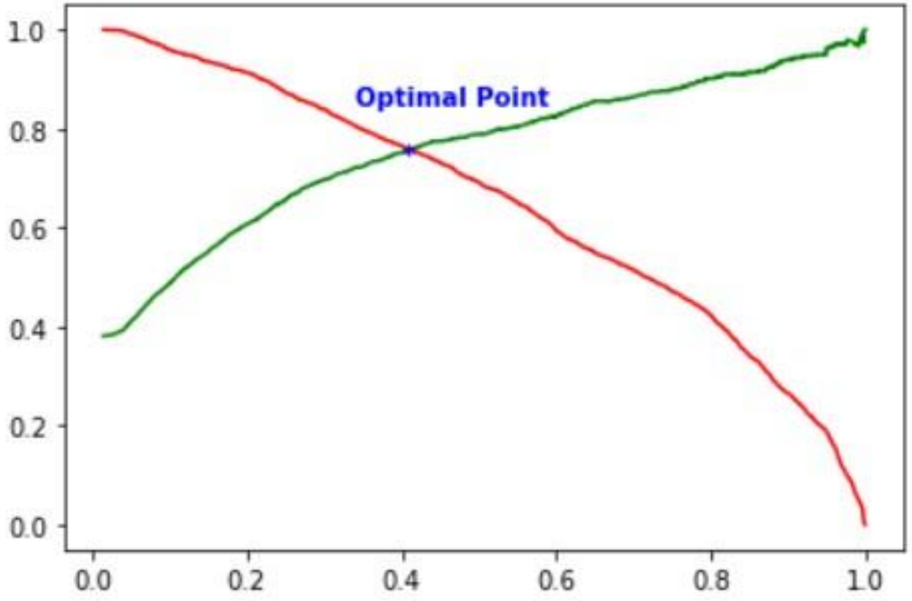


From the above visualization, we can observe that the Optimal Probabilistic CutOff is `0.335`.

Still, let's check from Precision-Recall Trade-Off Curve as well to find the optimal probability.

PLOT OF ROC CURVE & OPTIMAL POINT





- In the above diagram, Red Line represents the Recall & Green represents the Precision.
- And the Optimal Probability is at `0.4`.
- Both Optimal Probabilities are close to each other. Therefore, let's take mean of these as the `Final Optimal Probability Value`
- Hence, the `Final Optimal Probability` is at `0.37`

Conclusion

- After Lead Score we can see that the final prediction of conversions have a target of 80% (78.39%) conversion as per the X Education CEO’s requirement
- Accuracy is 80.98%
- Recall is 78.39%
- Precision is 73.51%
- Specificity is 82.58%
- False Positive Rate is 17.42%
- Hence it seems a Good Model for Lead Scoring