Lead Score Case Study

Submitted by :

Harshit Kumar
Gautami Pravin Wankhede

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%.

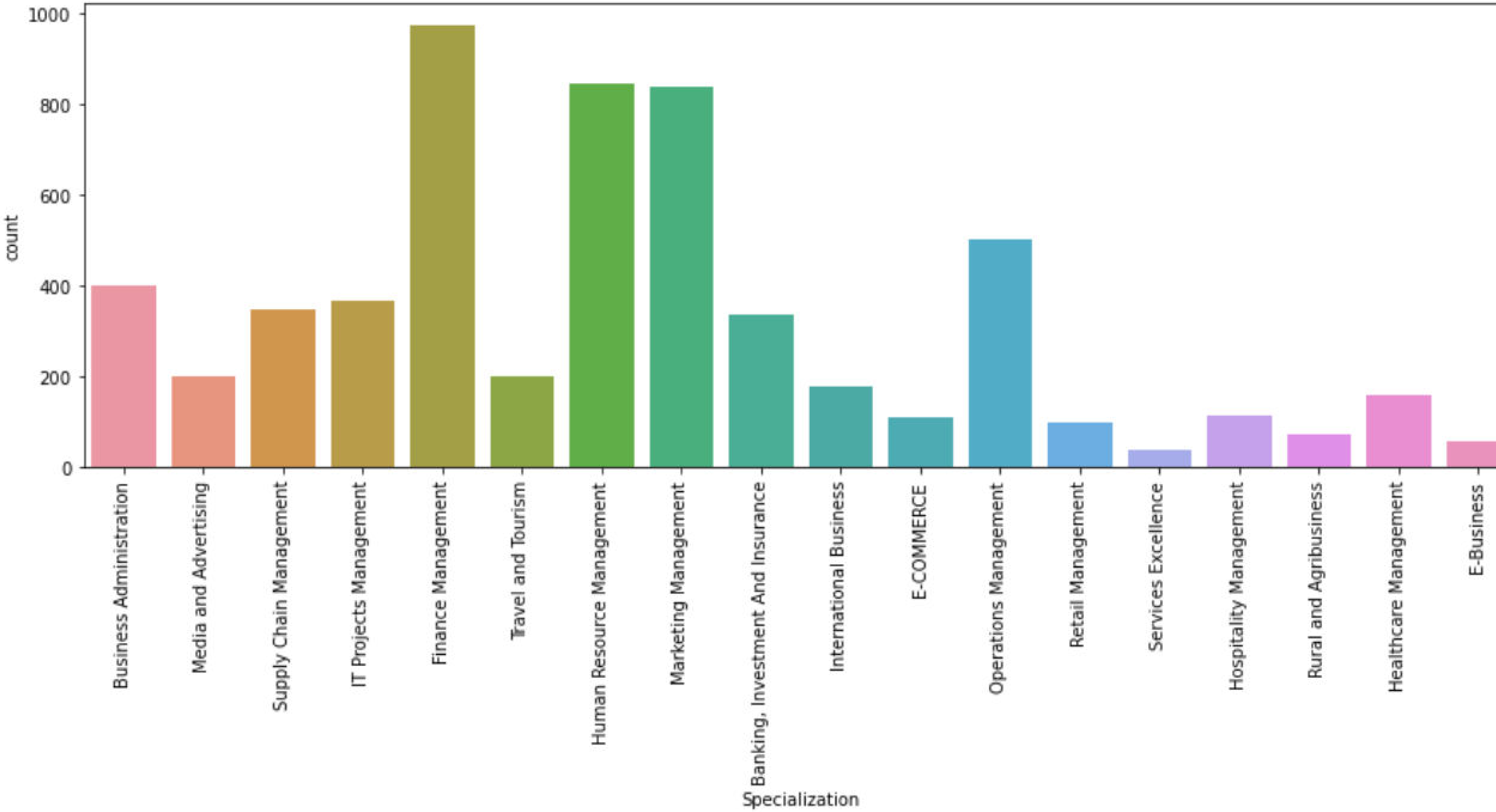
Business Goal

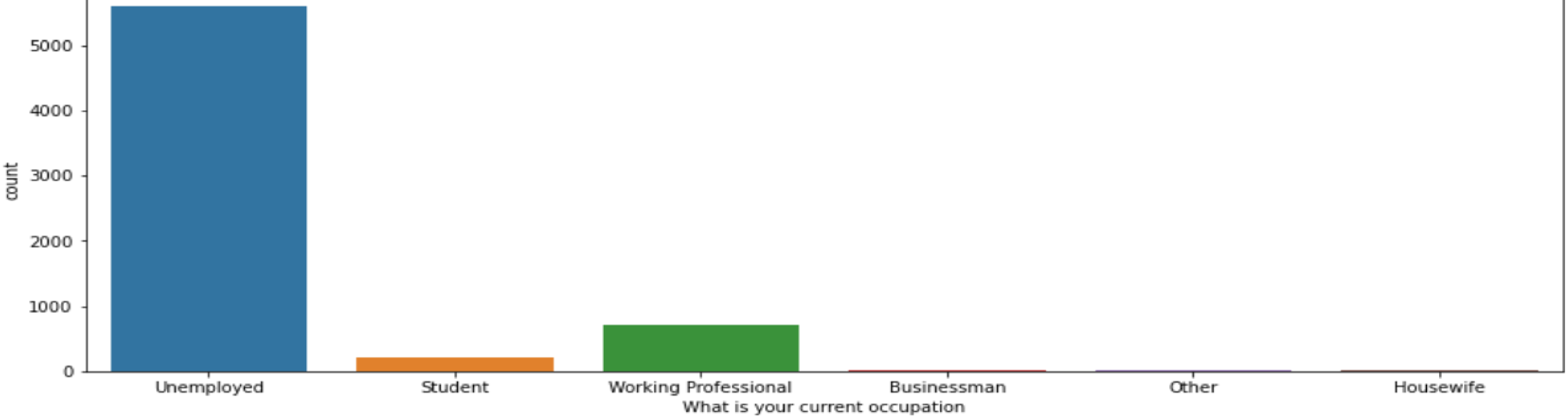
X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the 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%.

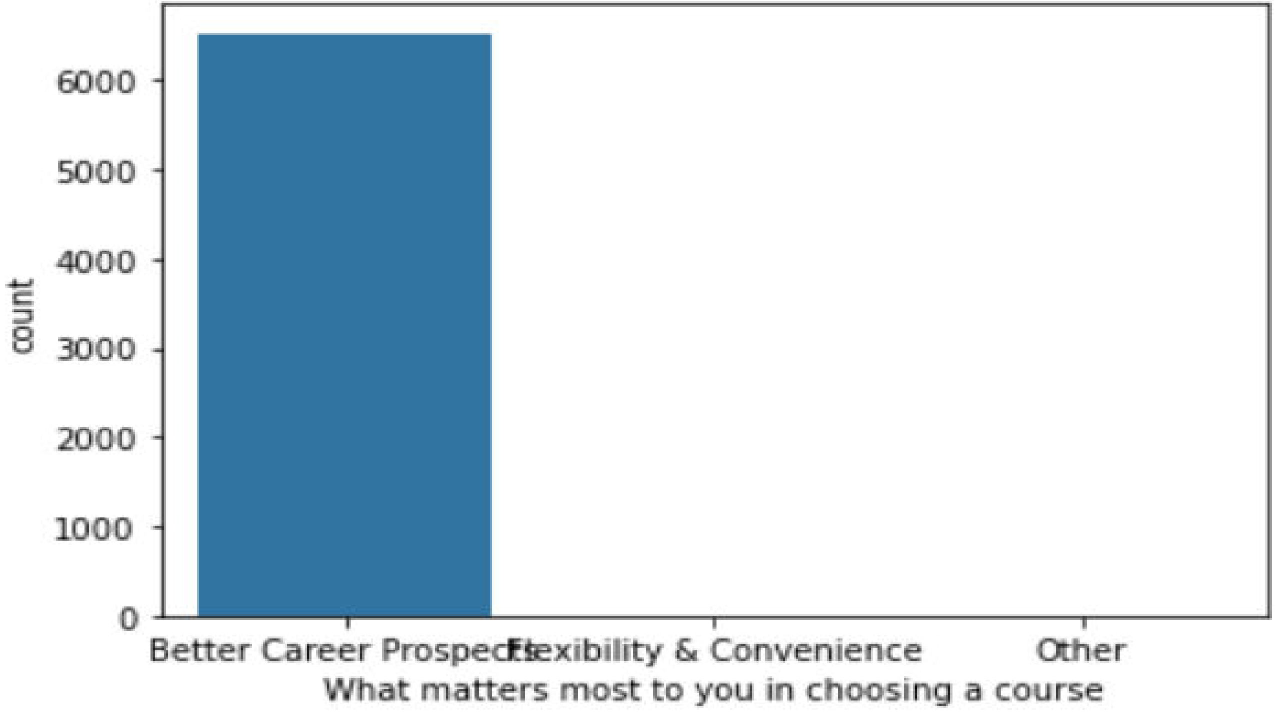
Strategy

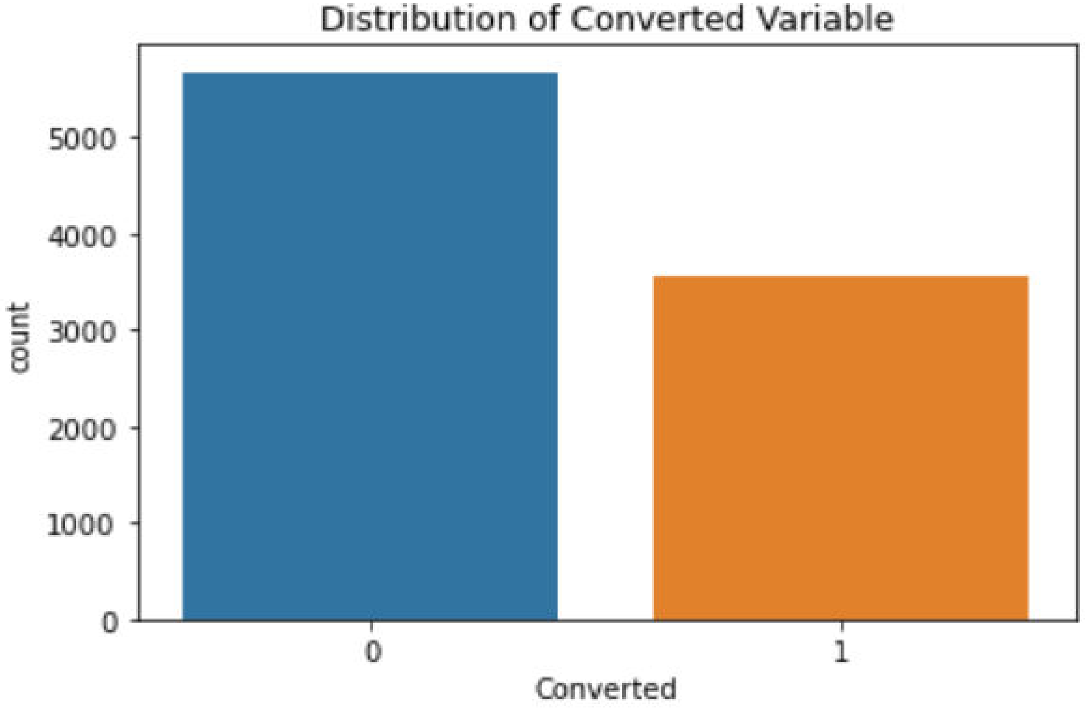
- Source the data for analysis.
- Clean and prepare the data.
- Exploratory Data Analysis.
- Feature Scaling.
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics. And Validation of the model.
- Conclusions and recommendations.

EDA

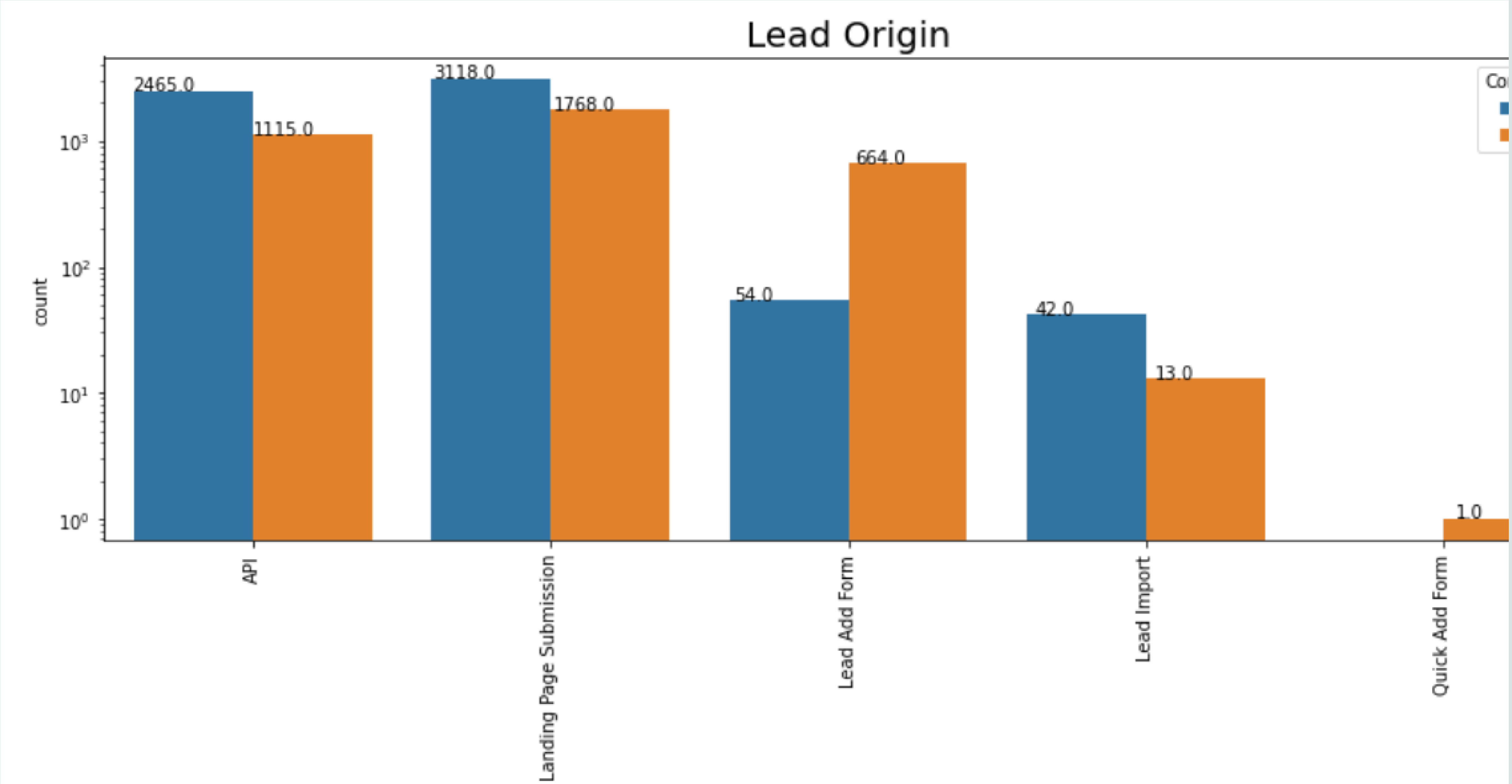

1000
800
600
400
200
0
count
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail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

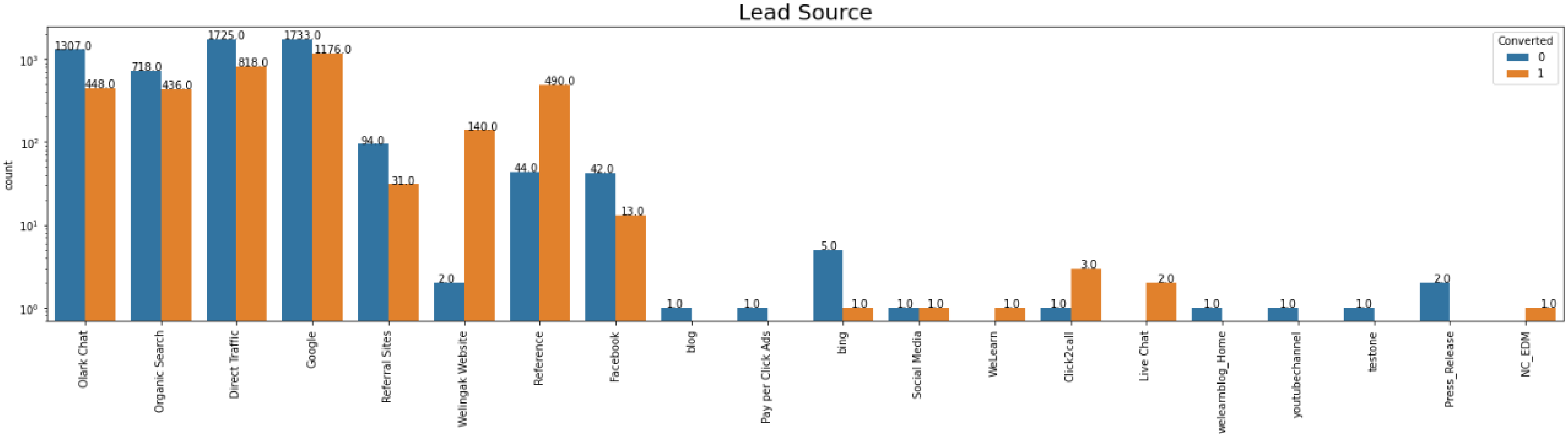

5000
4000
3000
2000
1000
0
count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

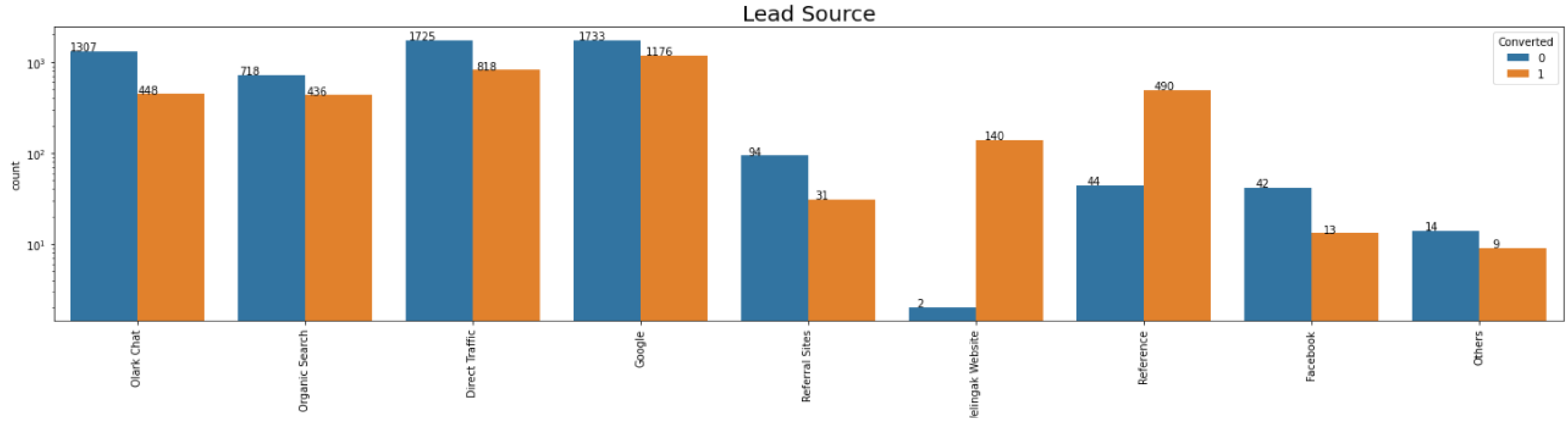

6000
5000
4000
3000
2000
1000
0
count
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course

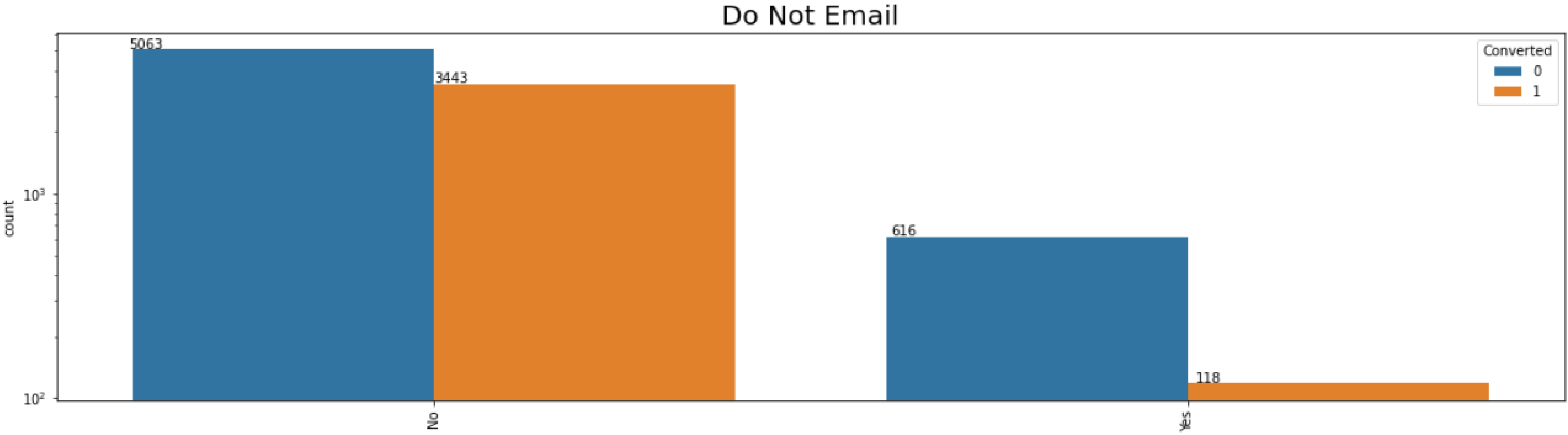

Distribution of Converted Variable
5000
4000
3000
2000
1000
0
count
0
1
Converted

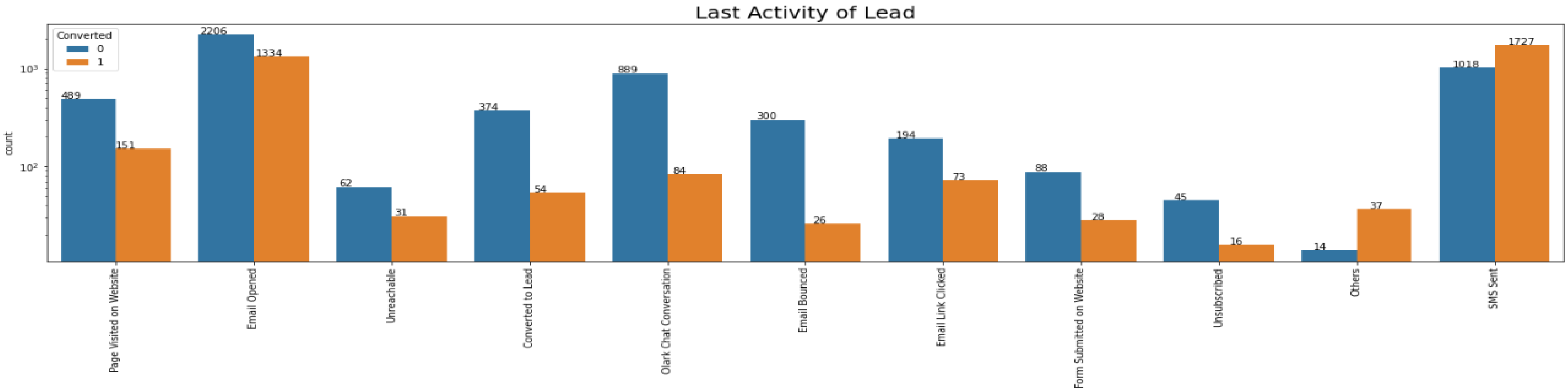
Categorical Variable Relation

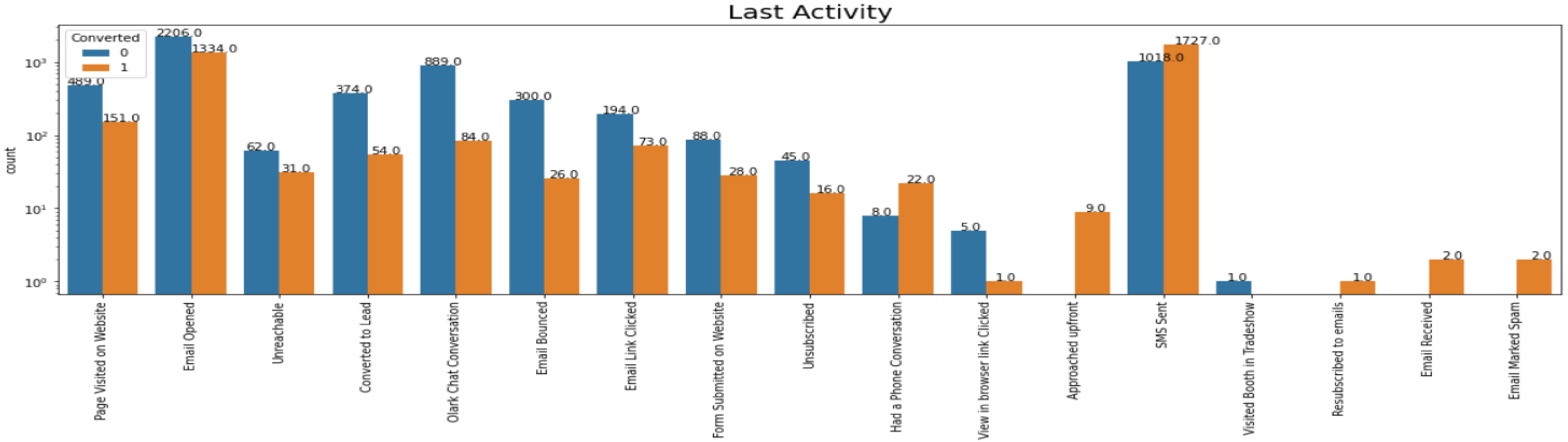

Lead Origin
count
10^0
10^1
10^2
10^3
2465.0
1115.0
3118.0
1768.0
54.0
664.0
42.0
13.0
1.0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Co

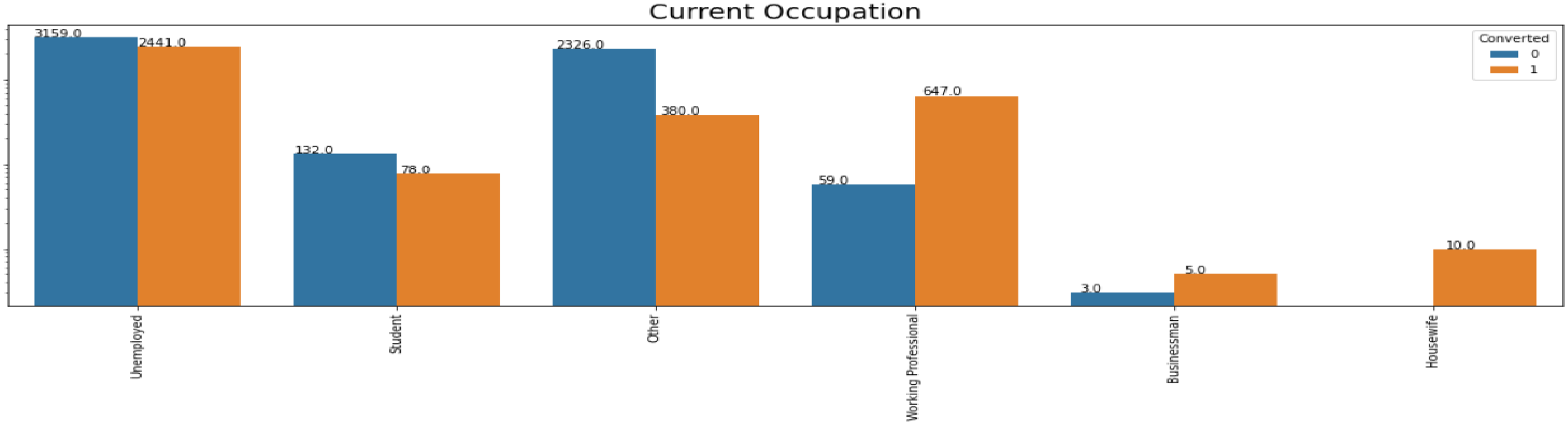

Lead Source
Converted
0
1
count
10^3
10^2
10^1
10^0
1307.0
448.0
718.0
436.0
1725.0
818.0
1733.0
1176.0
94.0
31.0
2.0
140.0
44.0
490.0
42.0
13.0
1.0
1.0
5.0
1.0
1.0
1.0
1.0
1.0
3.0
2.0
1.0
1.0
1.0
2.0
1.0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM

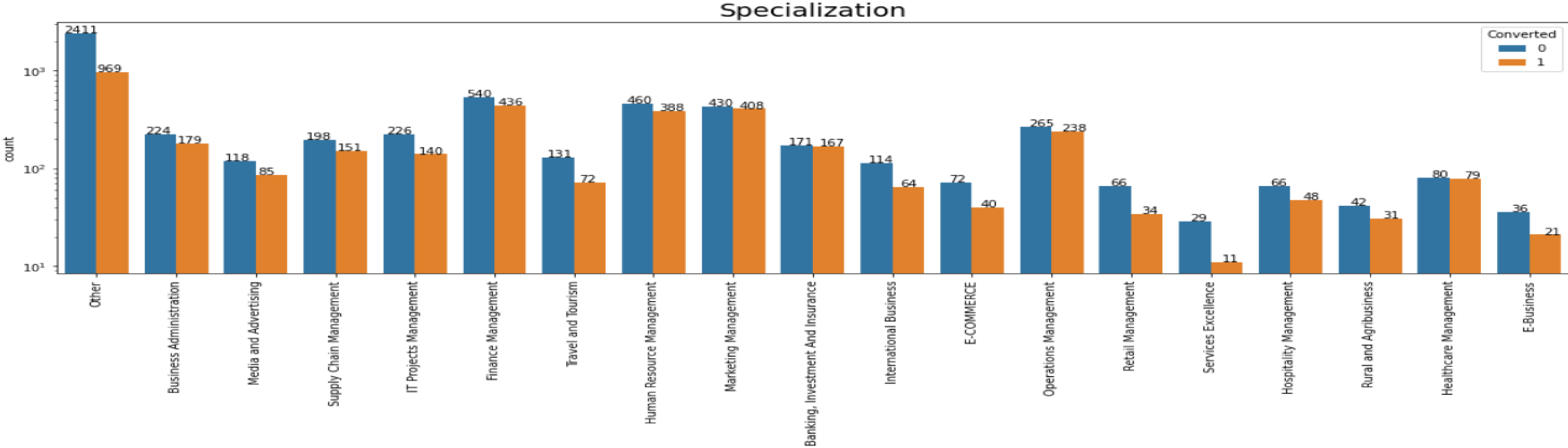

Lead Source
count
10^3
10^2
10^1
Converted
0
1
1307
448
718
436
1725
818
1733
1176
94
31
2
140
44
490
42
13
14
9
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
elingak Website
Reference
Facebook
Others


Do Not Email
count
10^3
10^2
Converted
0
1
5063
3443
616
118
No
Yes

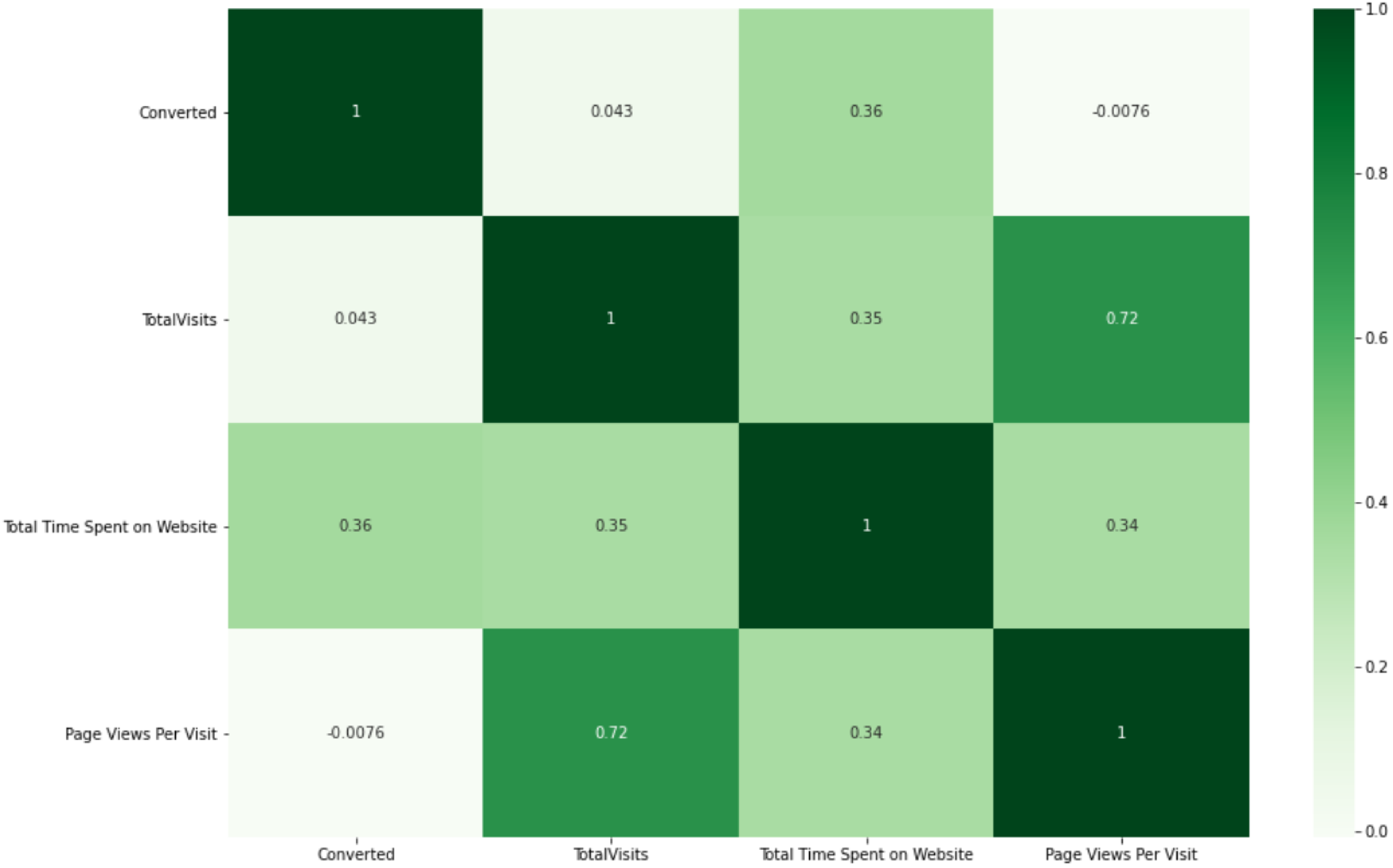

Last Activity of Lead
Converted
0
1
count
10³
10²
489
151
2206
1334
62
31
374
54
889
84
300
26
194
73
88
28
45
16
14
37
1018
1727
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Others
SMS Sent


Last Activity
Converted
0
1
count
10³
10²
10¹
10⁰
489.0
151.0
2206.0
1334.0
62.0
31.0
374.0
54.0
889.0
84.0
300.0
26.0
194.0
73.0
88.0
28.0
45.0
16.0
8.0
22.0
5.0
1.0
9.0
1018.0
1727.0
1.0
1.0
2.0
2.0
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


Current Occupation
count
10³
10²
10¹
Converted
0
1
3159.0
2441.0
132.0
78.0
2326.0
380.0
59.0
647.0
3.0
5.0
10.0
Unemployed
Student
Other
Working Professional
Businessman
Housewife


Specialization
count
10³
10²
10¹
Converted
0
1
2411
969
224
179
118
85
198
151
226
140
540
436
131
72
460
388
430
408
171
167
114
64
72
40
265
238
66
34
29
11
66
48
42
31
80
79
36
21
Other
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

Bivariate Analysis


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.043
0.36
-0.0076
0.043
1
0.35
0.72
0.36
0.35
1
0.34
-0.0076
0.72
0.34
1
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

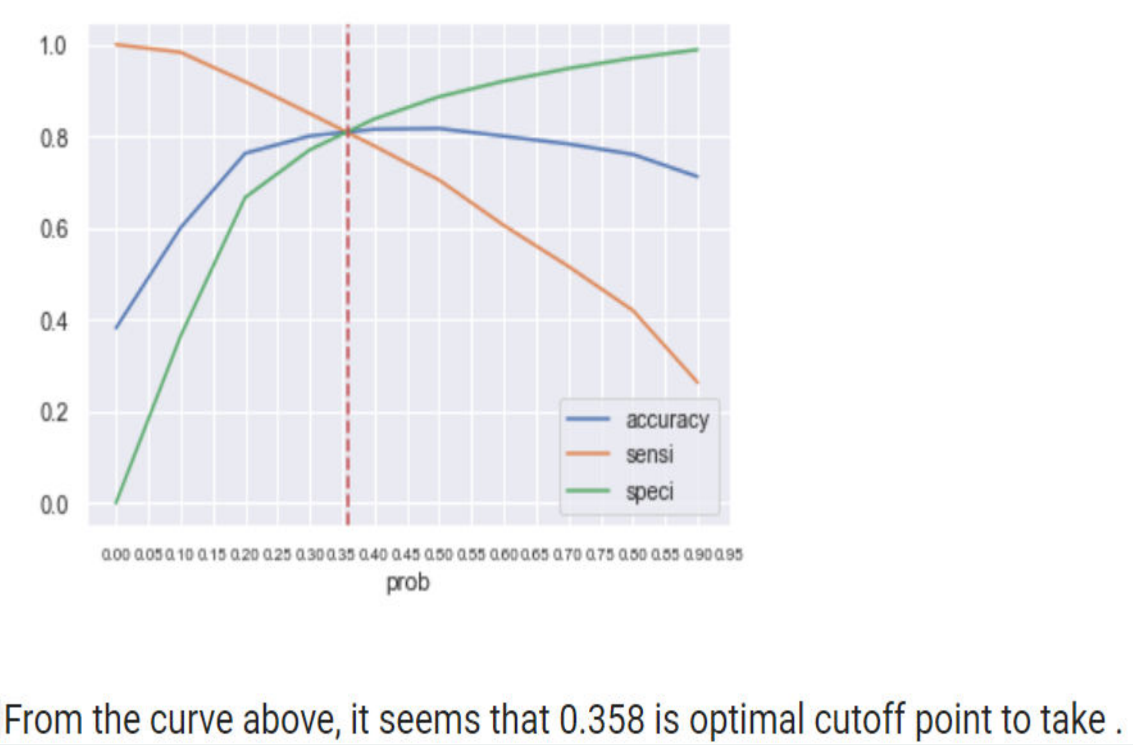
Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

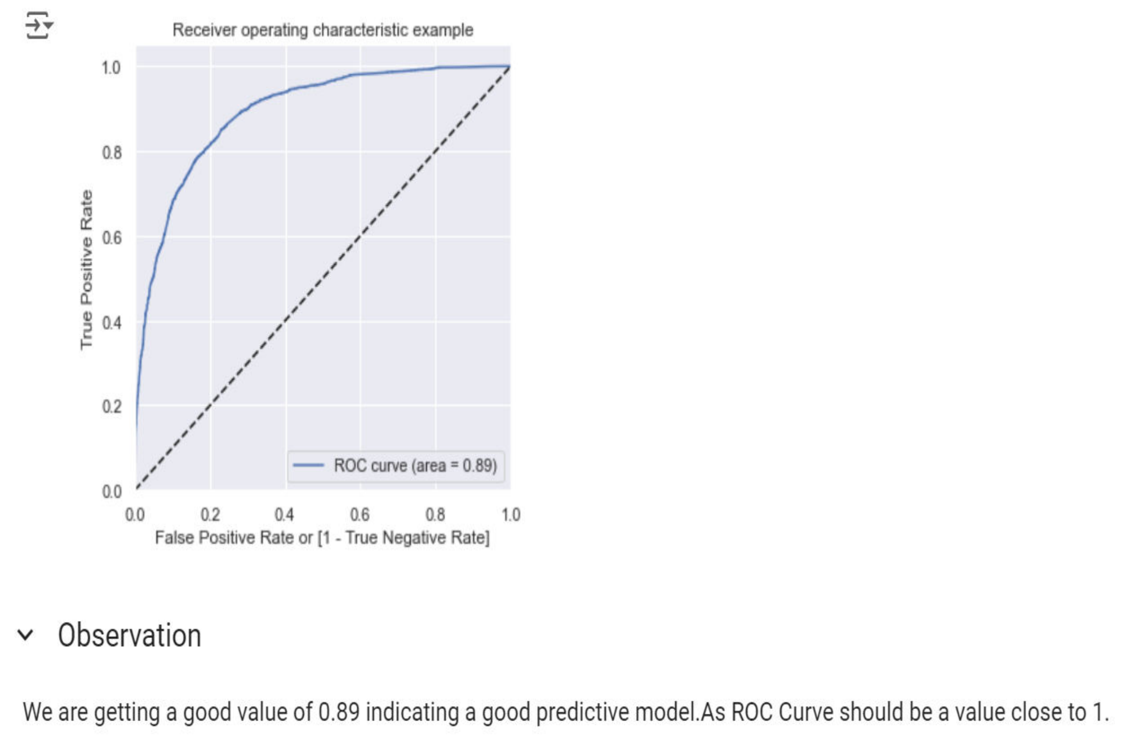
Model Building

- 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 Curve



From the curve above, it seems that 0.358 is optimal cutoff point to take .



Observation

We are getting a good value of 0.89 indicating a good predictive model.As ROC Curve should be a value close to 1.

Conclusion

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  a. *Google*
  b. *Direct traffic*
  c. *Organic search*
  d. *Welingak website*
- When the last activity was:
  a. *SMS*
  b. *Olark chat conversation*
- When the lead origin is Lead add format.
- 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.

Thank you

Harshit Kumar

Gautami Pravin Wankhede