Lead Scoring Case Study

- Mohammad Aslam Mahmood
- Suman Dandapat
- Abhishek Pandey

Business Problem Statement

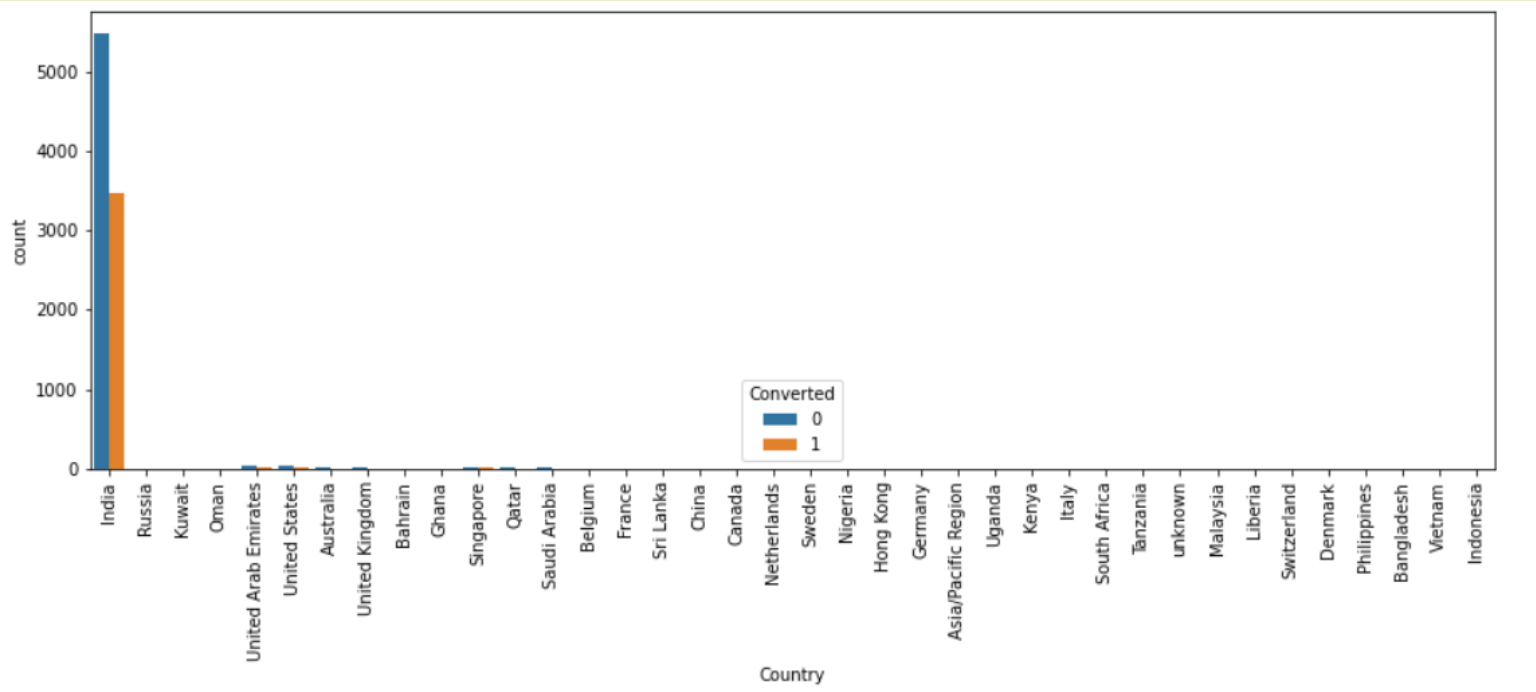
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these 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%.
- There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

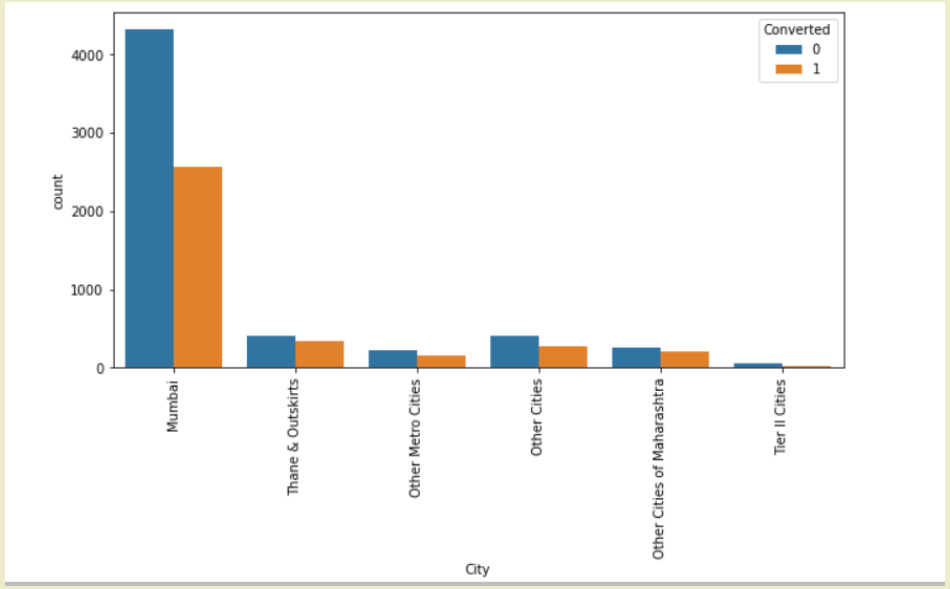
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

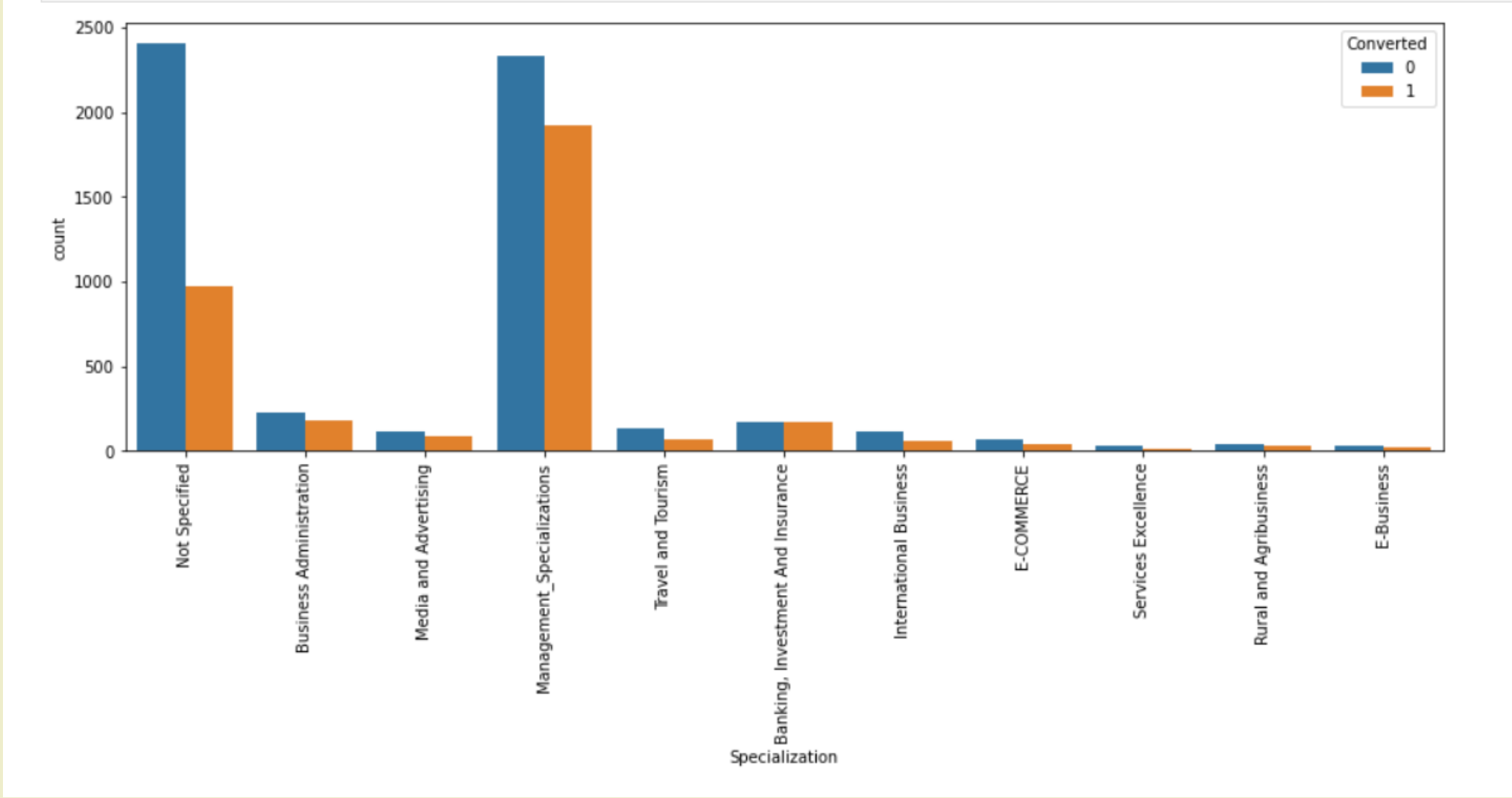
A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

EDA Univariate Analysis

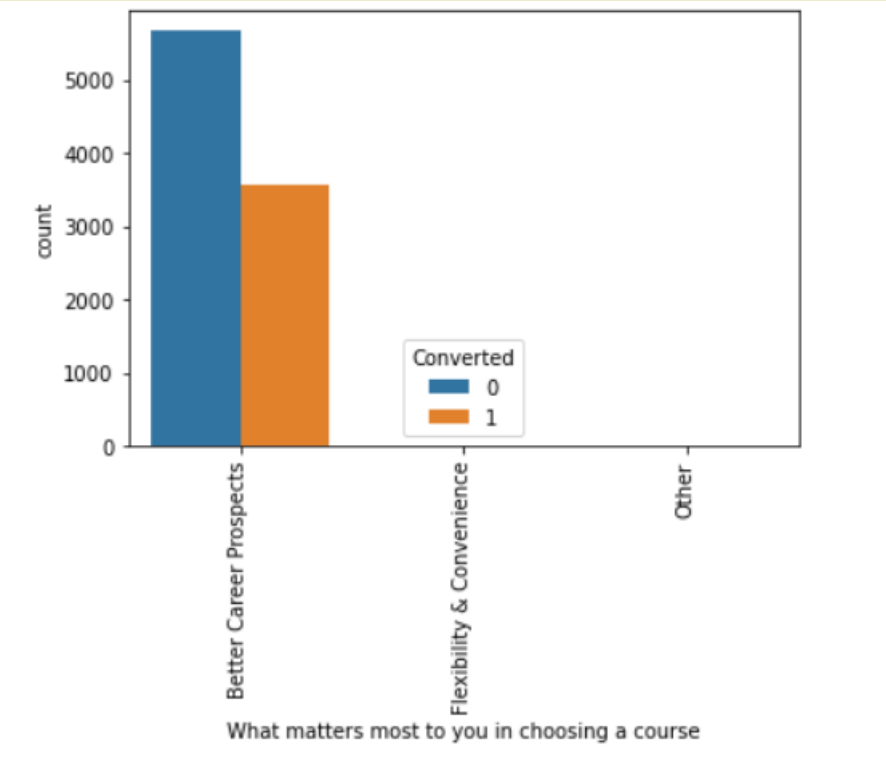


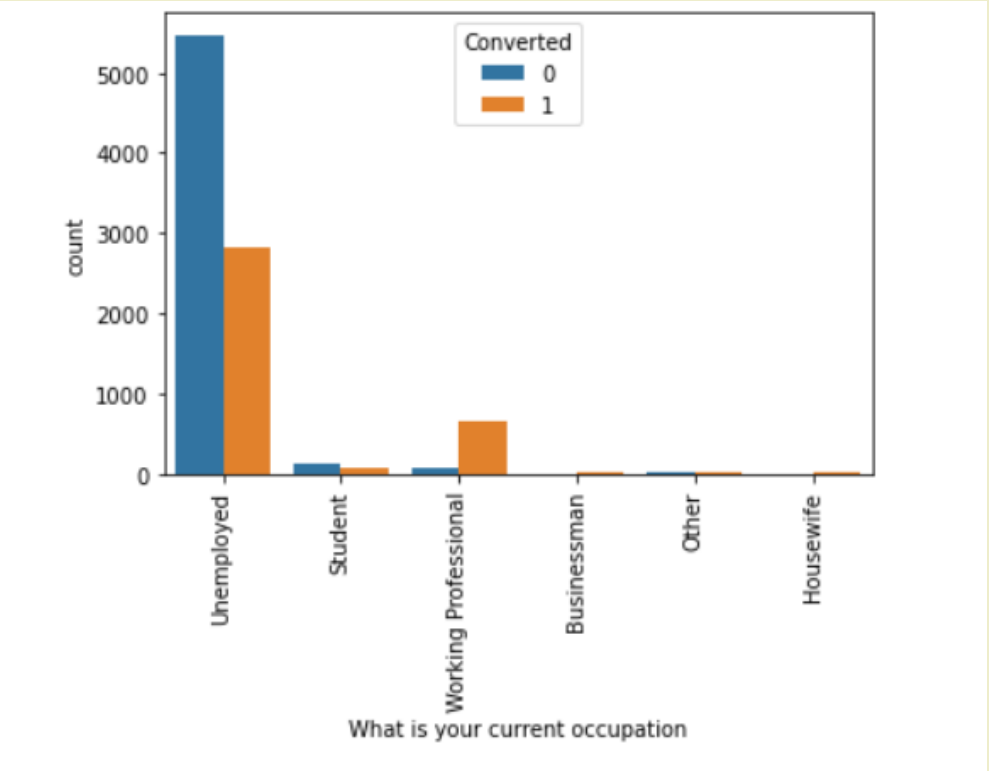
->India has highest number of leads
->Mumbai city has highest number of leads.





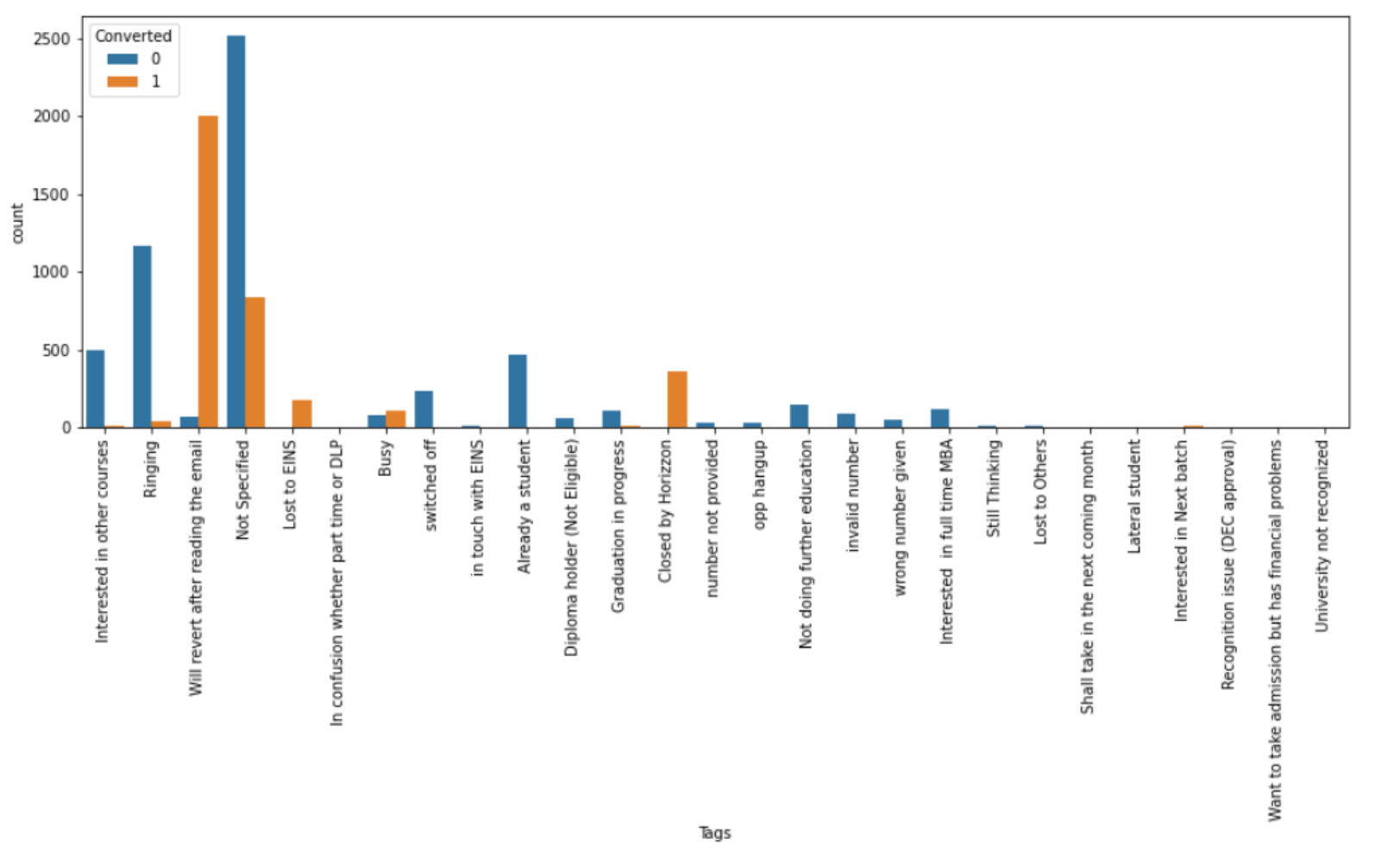
People with Management specialization are the greater possibility of getting the Leads converted.





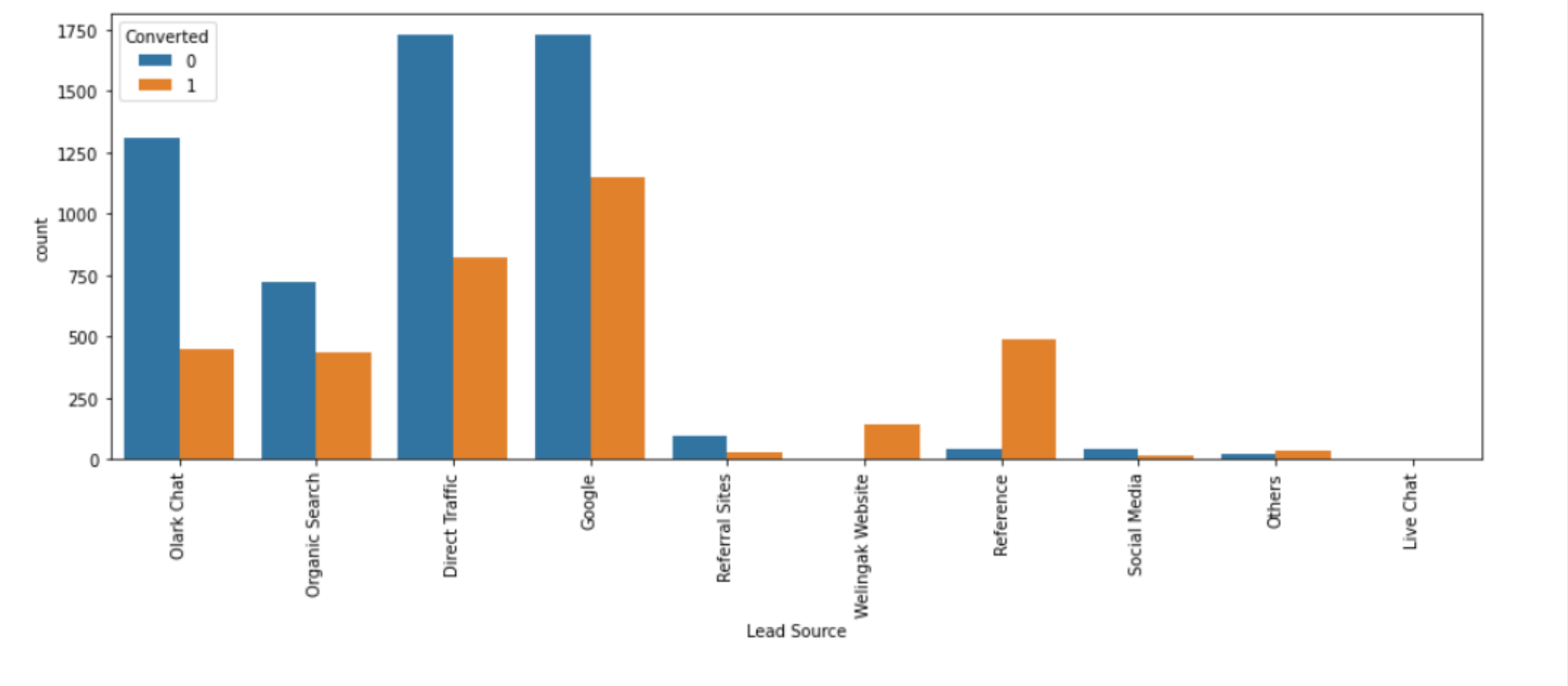
-> People with unemployment are more likely to choose the courses for better career opportunities.

-> Working professionals can be the next target for opting the courses for better career prospects.

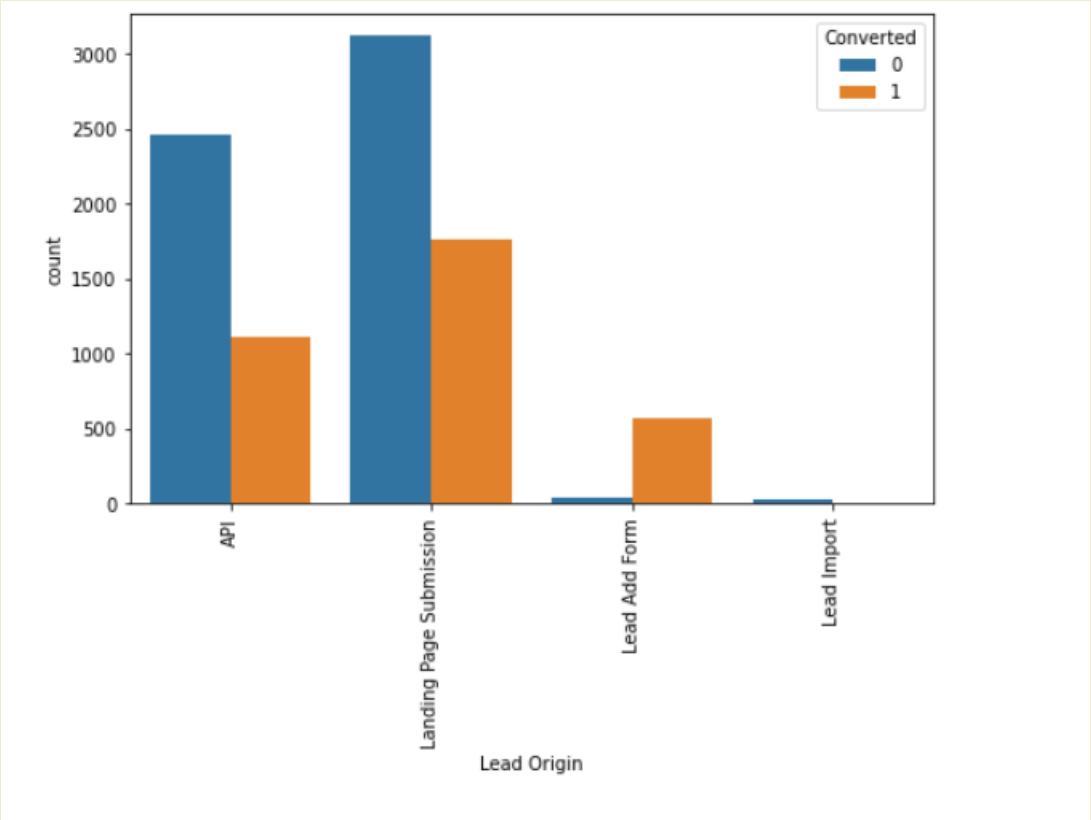


-> People who revert the email are more likely to opt for the courses.

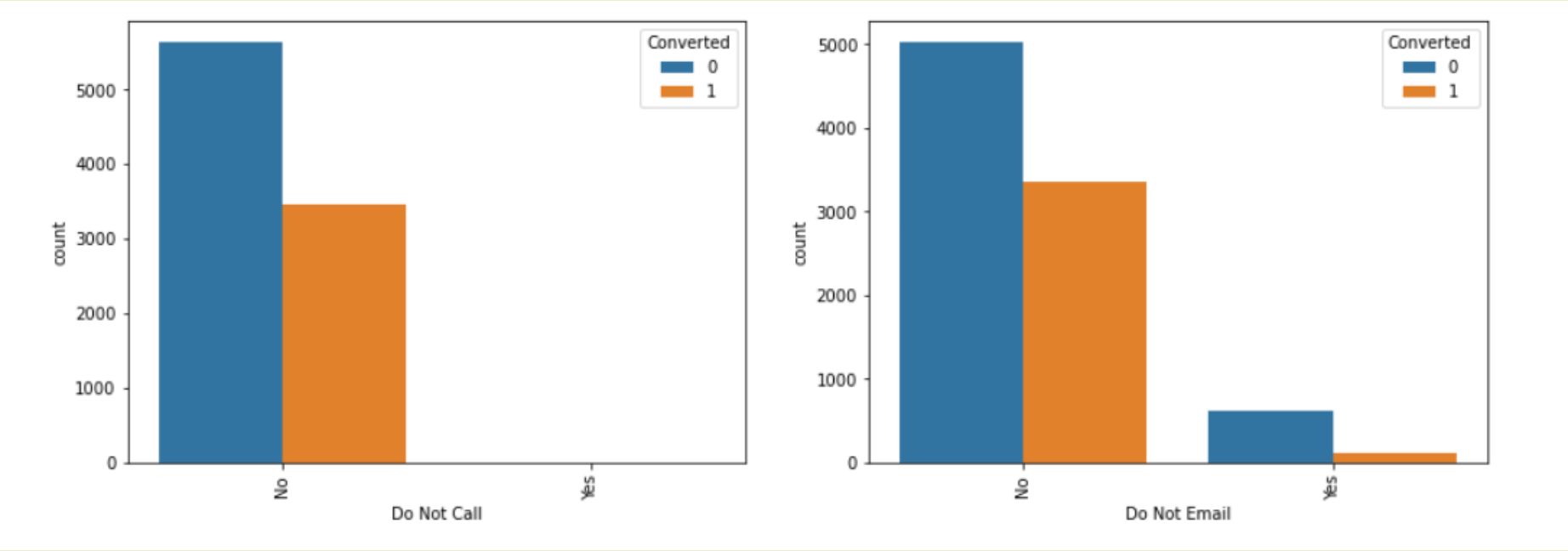
-> Ringing tags can be less prioritised as the chances of Lead conversion is very less.



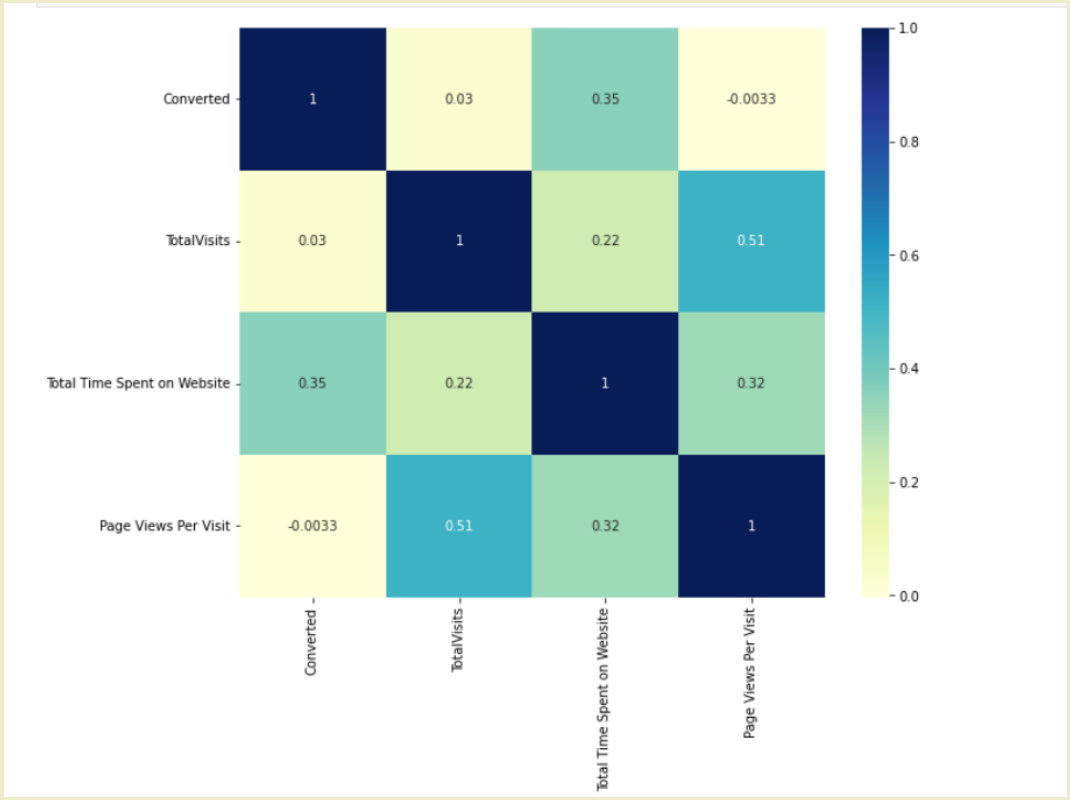
-> Google search engine can bring lot of business to the company as the Lead conversion rate is very high compared to other sources.
-> Next comes the Direct traffic followed by References, Organic search and Olark chat can be the source.
-> Olark Chat can be the less prioritised comparatively.



-> Applicant who landed on the submission page is likely to opt for the courses.
-> Followed by API and Lead Ad form.
-> Lead import can be ignored in this case.



-> Point of contact can be Email as well as Call.
-> Follow-up should be done through these point of contact to get the Leads Converted.



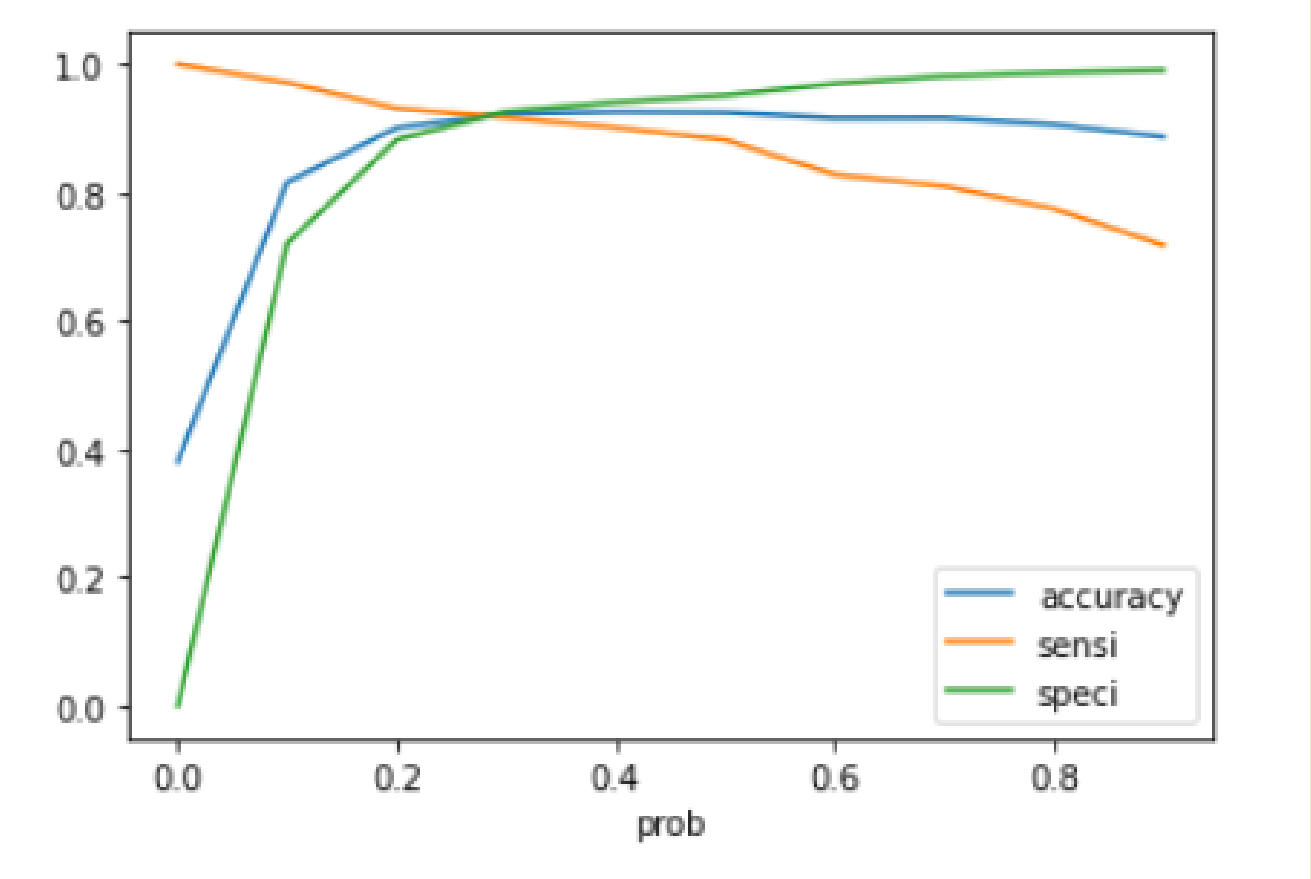
-> Converted variable is positively correlated with Total Time spent on the website, and negatively correlated with Page views per visit.

-> Total visits is positively correlated with Page views per visit.

-> Total Time spent on the website is positively correlated with Page views per visit and less correlated with Total Visits.

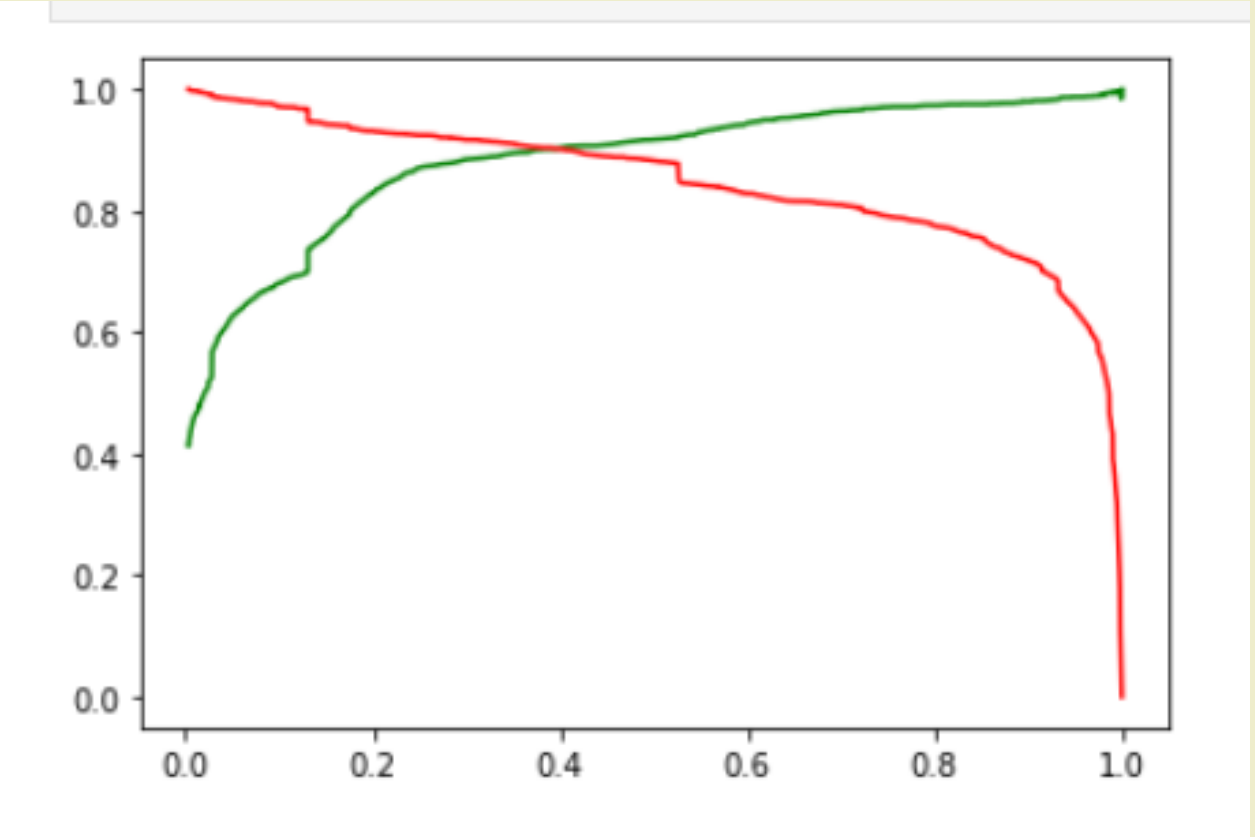
-> Train and test split was done on 70-30% and data was scaled
-> Built models using Logistic regressions and RFE selections reduced the features till p value was low and VIF values less than 5.
-> Finalized the third model with features as below.

	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01



The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%



After running the model on the Test Data these are the figures we obtain:

- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%

FINAL OBSERVATION

Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%

Test Data:

- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%

Final Conclusion

Recommendations :

-> People with unemployment are more likely to choose the courses for better career opportunities.

-> Working professionals can be the next target for opting the courses for better career prospects.

-> Google search engine can bring lot of business to the company as the Lead conversion is very high compared to other sources followed by Reference.

-> People who revert the email are more likely to opt for the courses.

Risky Bids:

-> Targeting the clients outside India can be the risky bid as the Leads are very less.

-> Tier 2 Cities within India can be less prioritised.