LEAD SCORING CASE STUDY

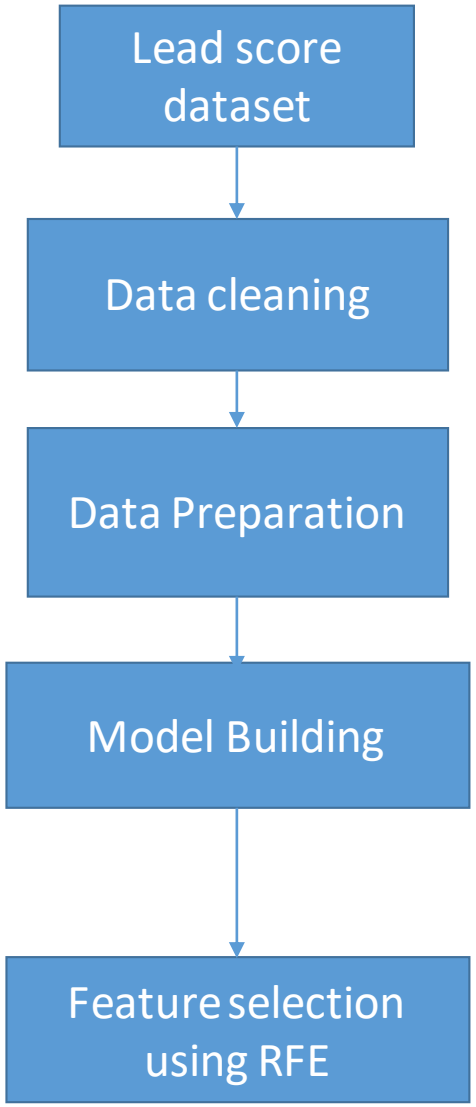
SUBMISSION

1. Monika Iyer M
2. Veerangouda Patil

Abstract

- Objective of the project is to help X education to find the hot leads and to increase the conversion of leads.
- Logistic regression was performed.
- This would help in finding the accuracy of the test results and also ranking the leads from 0 to 100.
- The company can target on individuals based on a number of factors, and increase their lead conversion rate.

Problem solving methodology

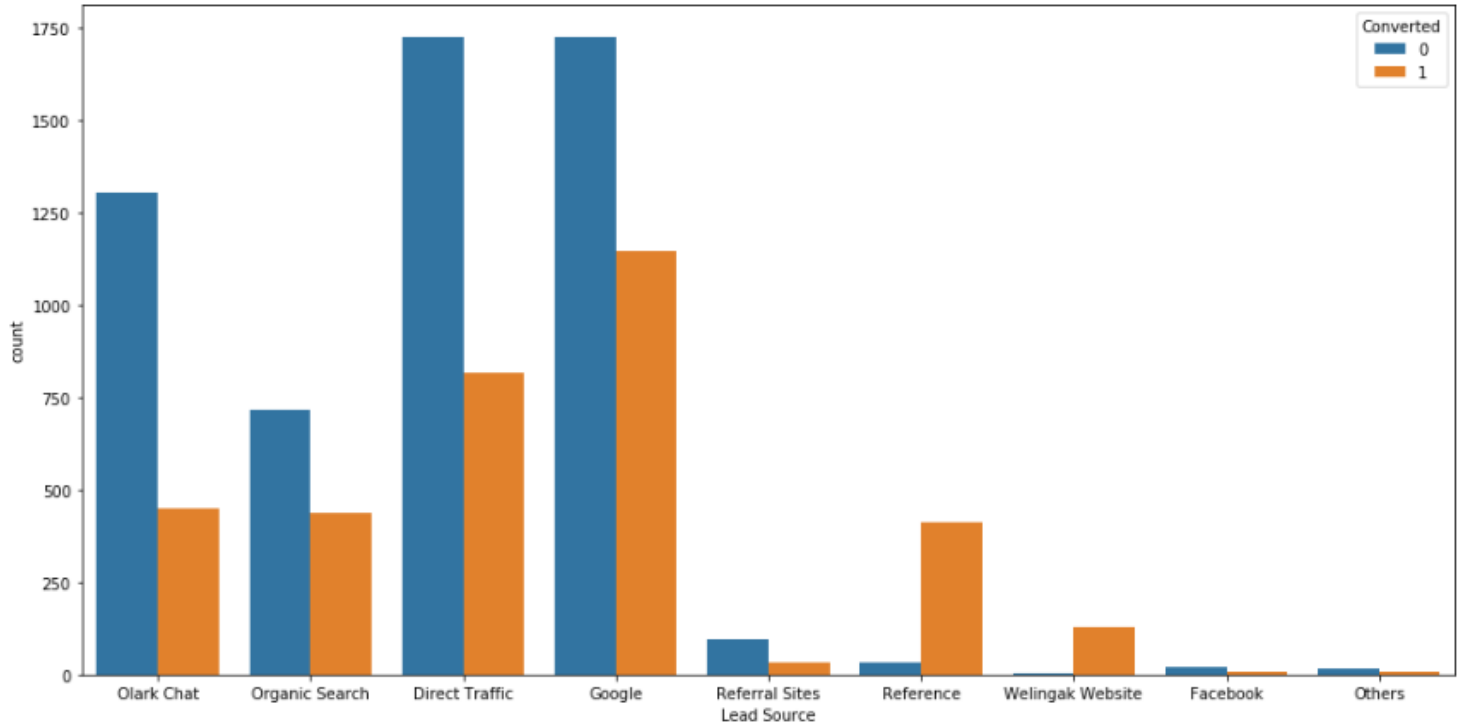


Univariate Analysis

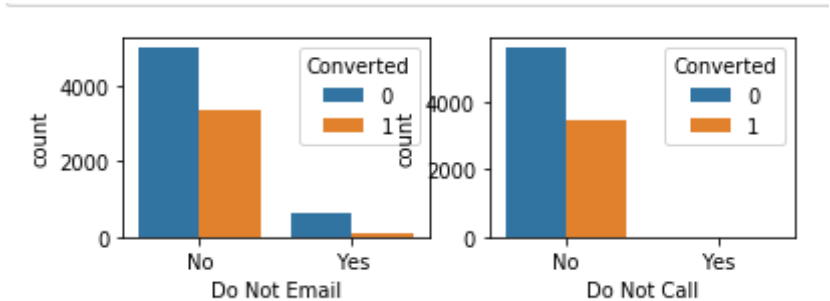
- We have segmented the annual income into Low, Medium, High and Very High. Low indicates that the annual income less than 50000, medium indicates 50000 to 100000, High is 100000 to 150000 and very high is greater than 150000. From the chart below, we can see that the lower the income, the chances of being a loan defaulter is greater.
- We have segmented the Emp Length into Junior, Mid-level and Senior. Junior is who has less than 4 years of experience, Mid- level is 5 to 8 years of experience and Senior is 9 and above years of experience. From the chart, we can see that Juniors have high chances of being a loan defaulter.
- We have segmented the Funded Amount to Low, Medium, High and Very High. Low is less than 5000 funded amount, medium is 5000 to 15000, high is 15000 to 25000 and very high is more than 25000. From the chart, we can see that people whose loan has been funded within 15000 to 25000 has many defaulters

Univariate Analysis

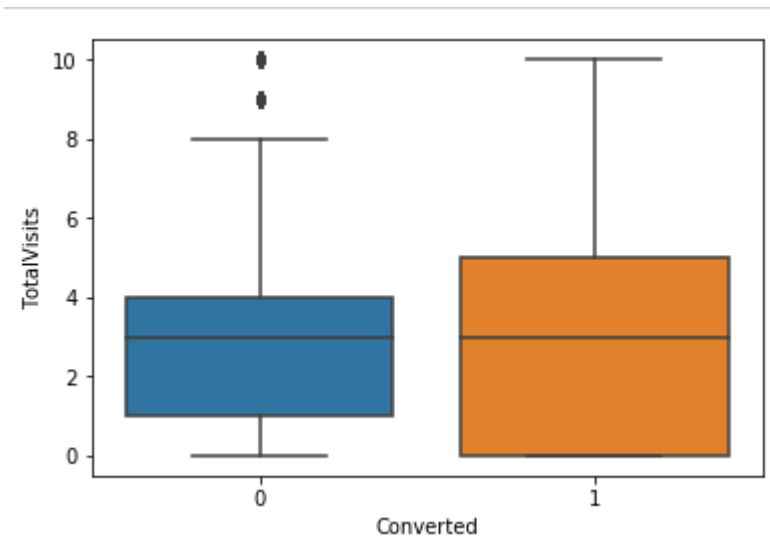
- For each of the column, univariate analysis was performed and here are the charts corresponding to that.
- For the column *Lead Source,* we took few columns as *Others* and plotted a graph as below. We can see that Google and Direct Traffic have more leads.



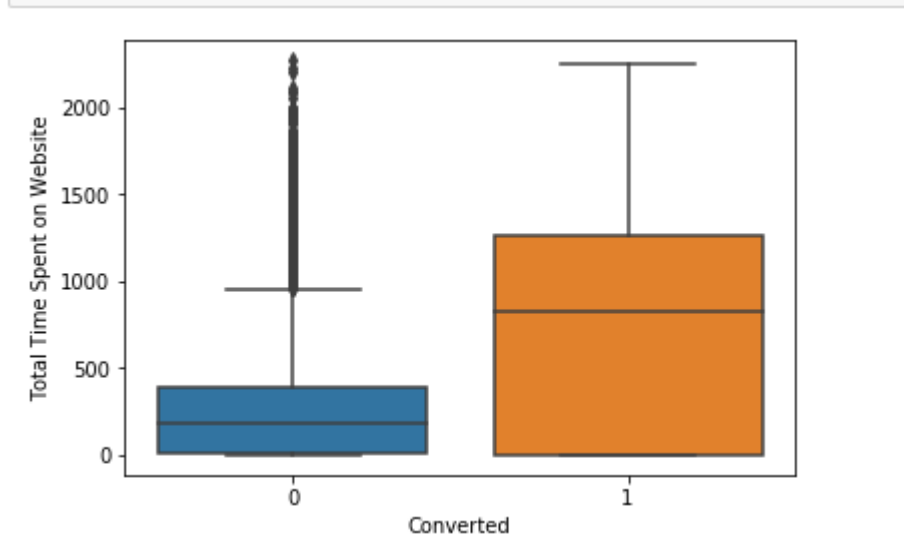
- For the column *Do Not Email* and *Do not call,* we plotted a graph as below. We can see thatMost people opted for no calls or no emails.



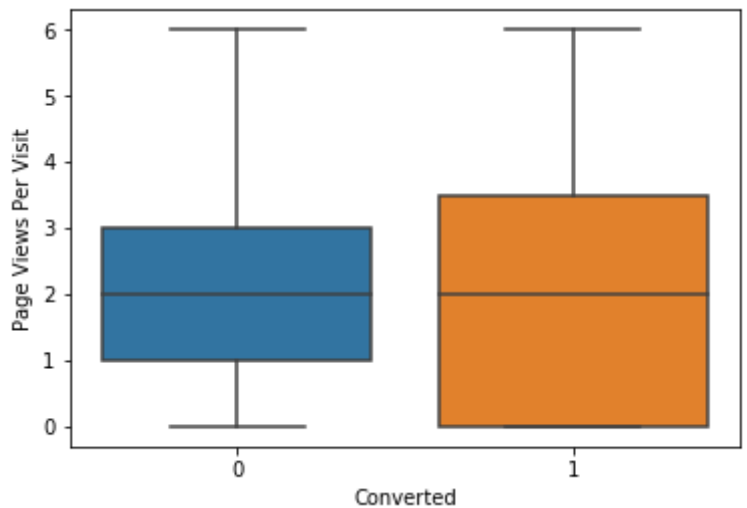
- For the column *Total Visits,* we plotted a boxplot as below. We can see that higher the conversion rate if more people visited the page.



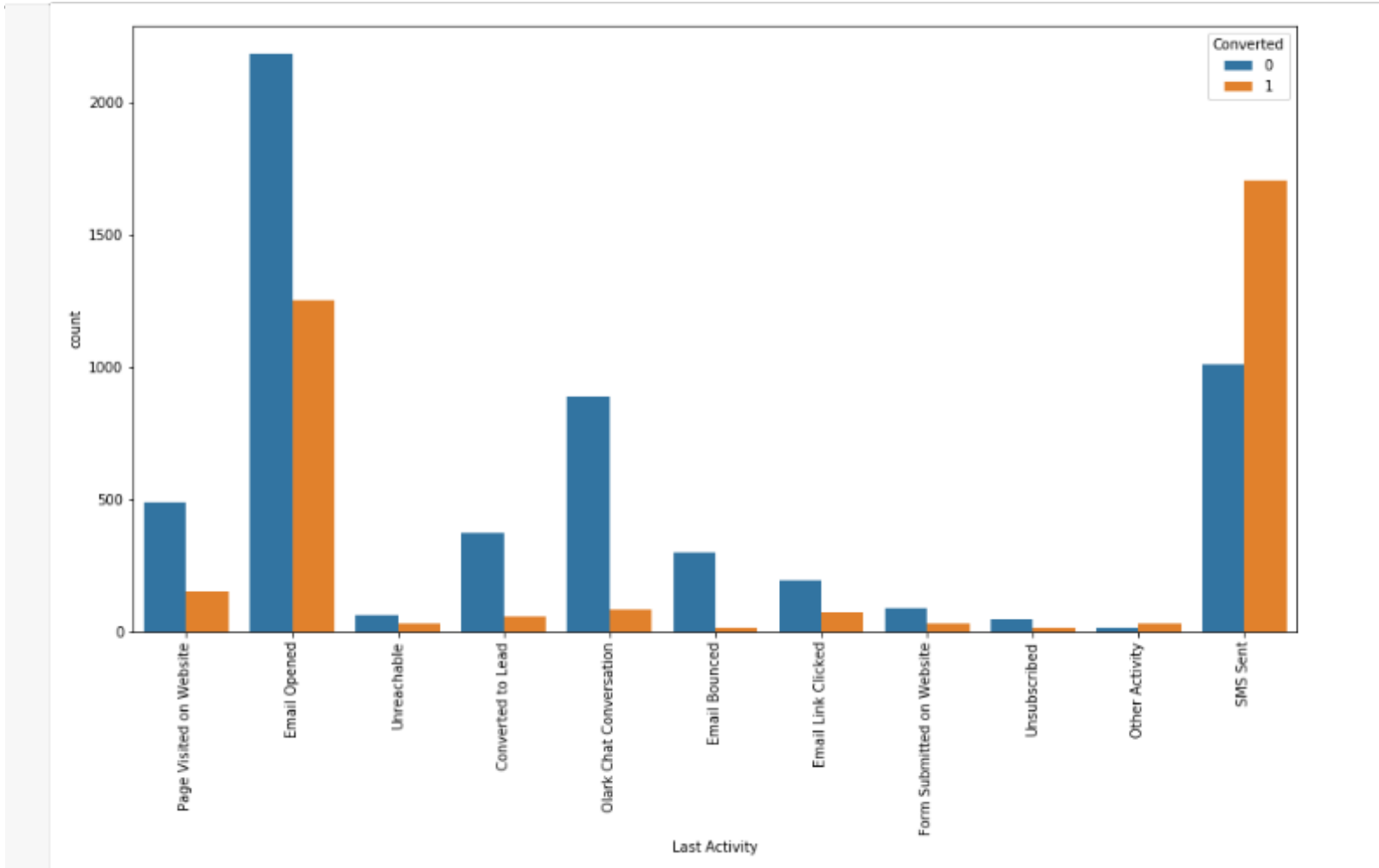
- For the column *Total Time Spent on Website,* we plotted a graph as below. We can see that more the time spent on website, more is the chance of conversion.



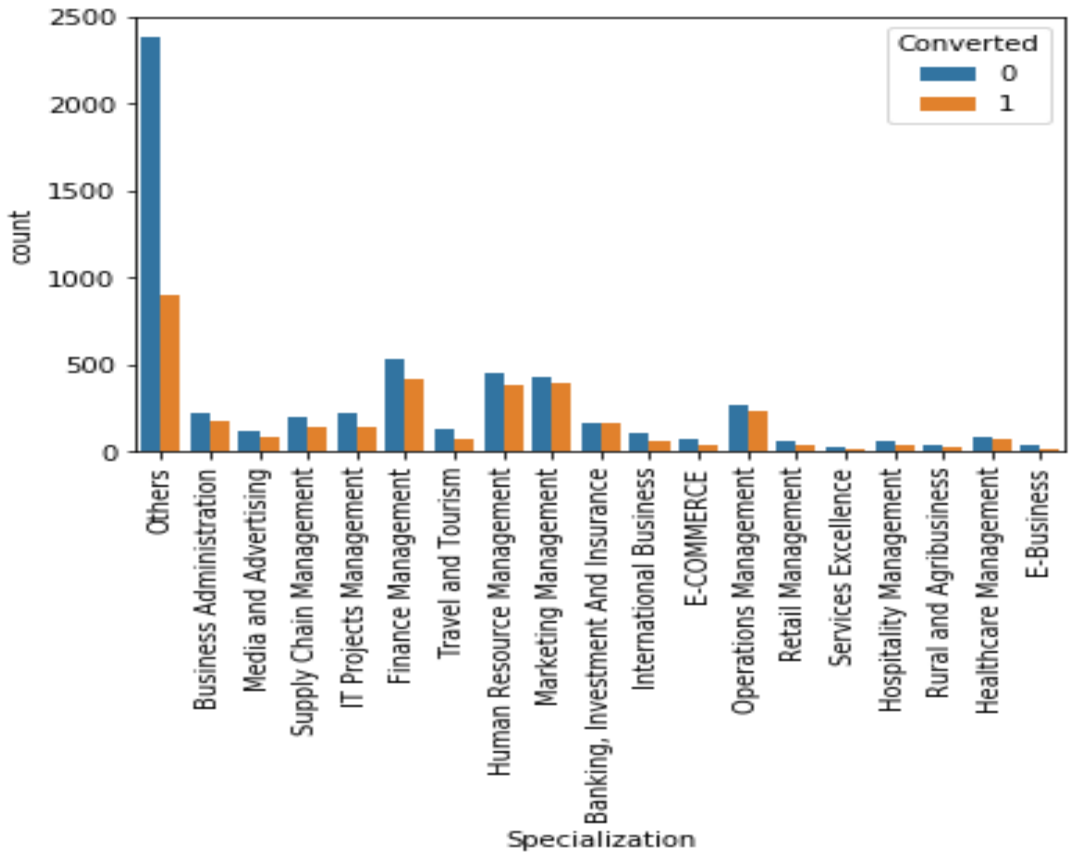
- For the column *Page Views Per Visit,* we plotted a boxplot as below. We can see that more the number of pages spent on visit, more is the chance of conversion.



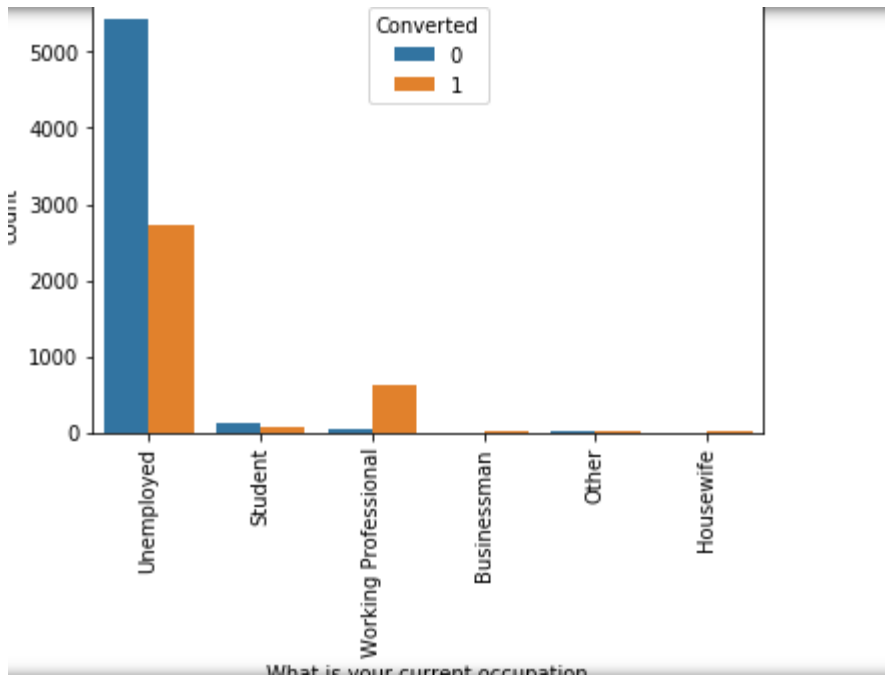
- For the column *Last Activity,* we plotted a plot as below. We converted smaller values as Other Activities. We can see that people have responded more when SMS has been sent, rather than an Email



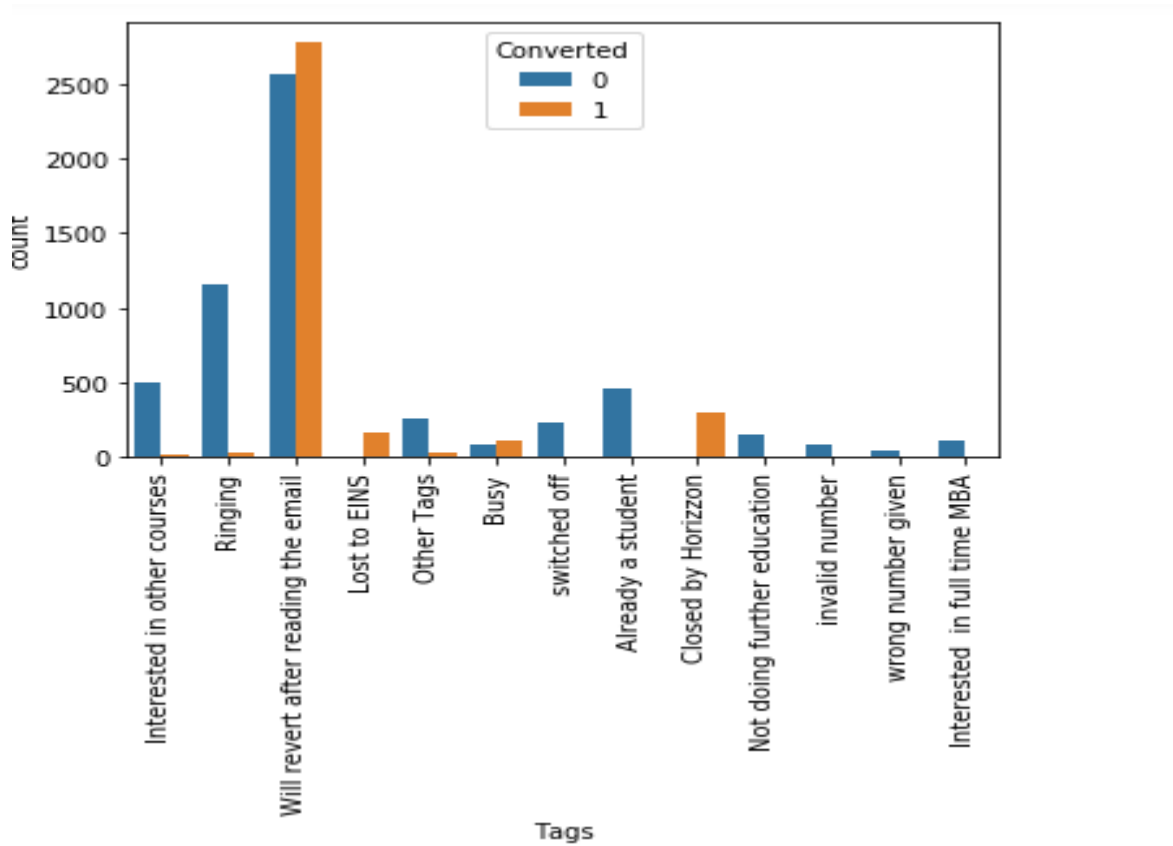
- For the column *Specialization,* we plotted a plot as below. When we looked at the specialization that the individuals have done, we see that many people have not selected anything at all, we have named that as *Others.* The next highest number of leads we have got is from people who have done *Finance Management* as their specialization



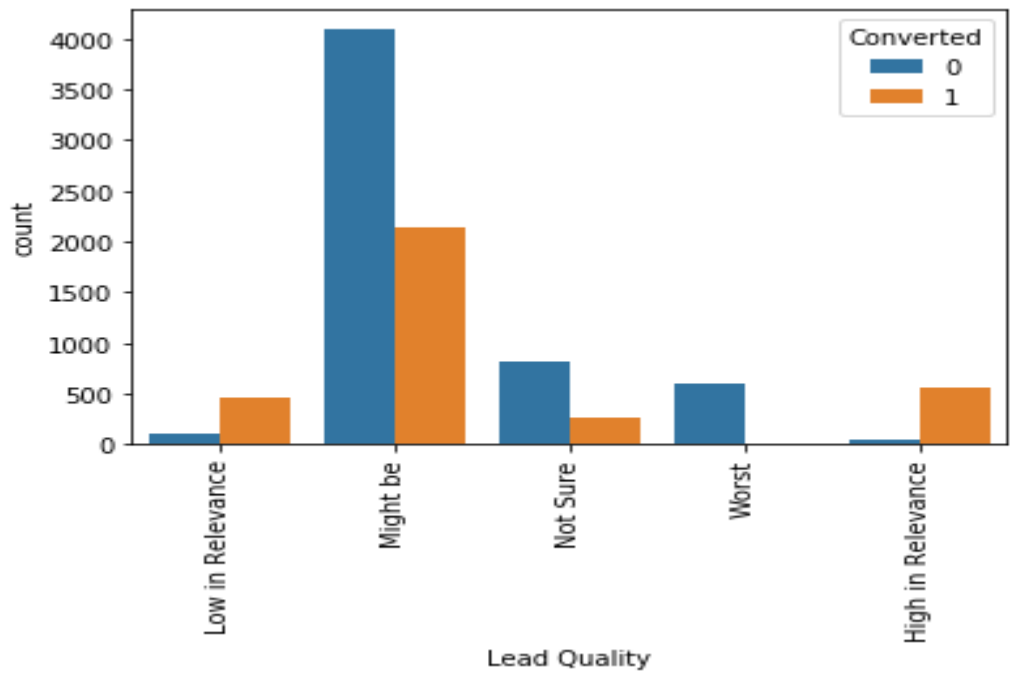
- For the column *What is your current occupation,* we plotted a plot as below. When we looked at the current employment status of individuals who are looking for a course, they were mainly *Unemployed.*



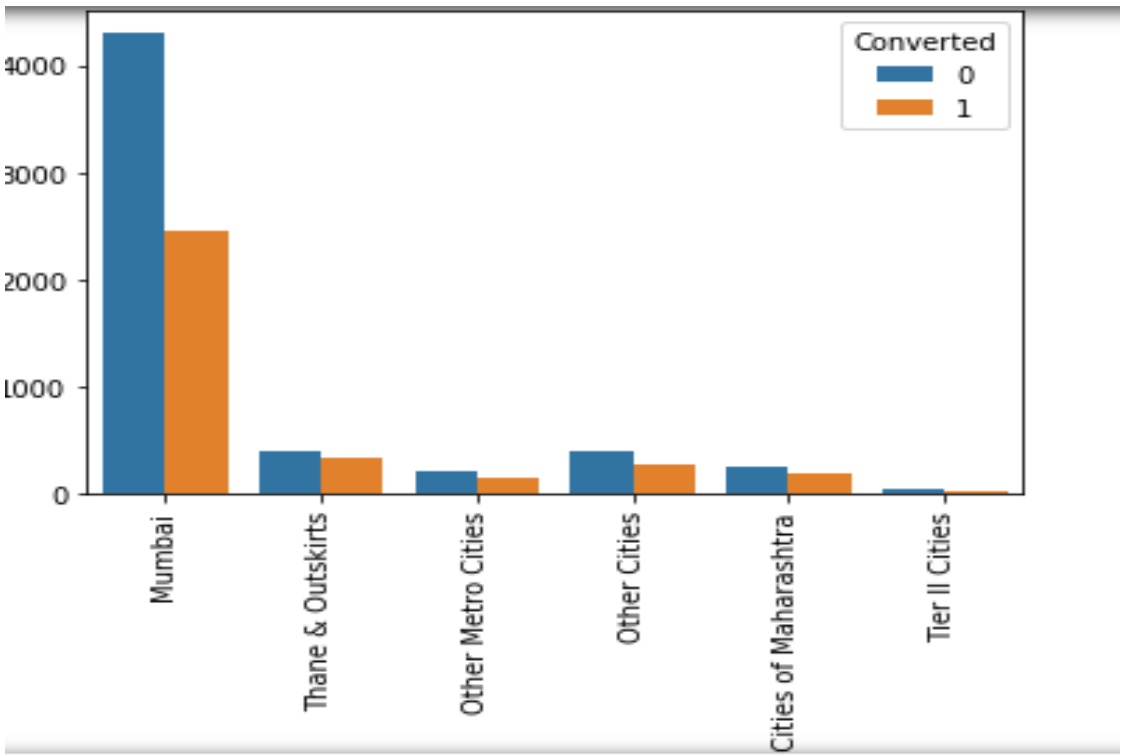
- For the column *Tags,* we plotted a plot as below. We took a look at the tags given by the individuals, of which, many people preferred to revert after reading the email.



- For the column *Lead Quality,* we plotted a plot as below. We took a look at Lead quality, which gave a result of uncertainty as the most preferred one.

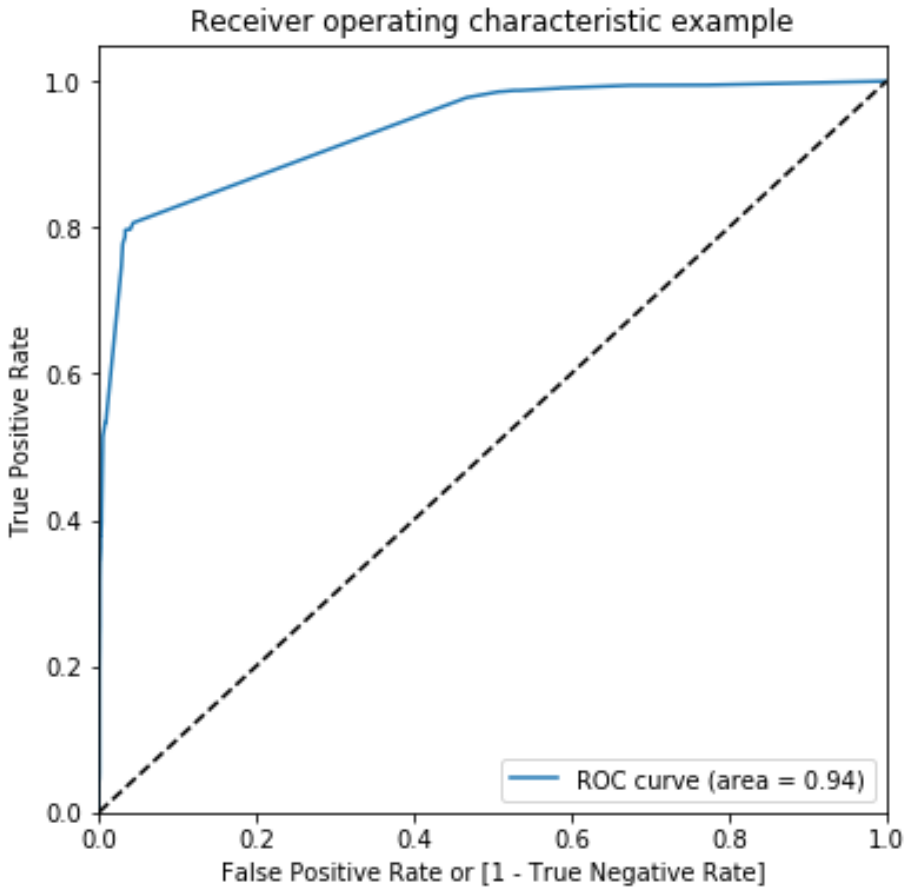


- For the column *City,* we plotted a plot as below. People who were from Mumbai were the ones who mostly preferred these courses.



- Looking at all the above points, we dropped few columns which deemed not necessary for the current analysis.
- We mapped the *Do Not Email* and *Do Not Call* as binary 0 and 1 for analysis.
- We also created dummy variables for the categorical variables.
- After this, we split the data into train and test datasets.
- We checked the current churn rate and it turned out to be around 38 percent.
- In addition to this, we did Feature selection using RFE on the train dataset.
- We also calculated the VIF, True positives, True Negatives, False Positives and False Negatives.
- Based on this, we calculated the accuracy, sensitivity, specificity and false positive rate.
- Post which we found the optimal cutoff, and assigned a lead score.
- The same values were then applied to the test data set and the accuracy came around 90 percent.

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.94)

Conclusions

- In conclusion, we can say that people who are Unemployed and have visited the page number of times and have spent quality time in the website, have higher chances of getting converted into hot leads.
- Also the landing page or API should be made a little more interactive so that people spend more time on the website.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- Most of the lead have their Email opened as their last activity, hence we could promote the courses through emails.
- Also, focus should be more on the Specialization with high conversion rate.
- Most leads are found to be from Mumbai city.