LEAD SCORING CASESTUDY

Submitted by,

Krithiga K

Martin Robert Durai

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us 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 and Objectives

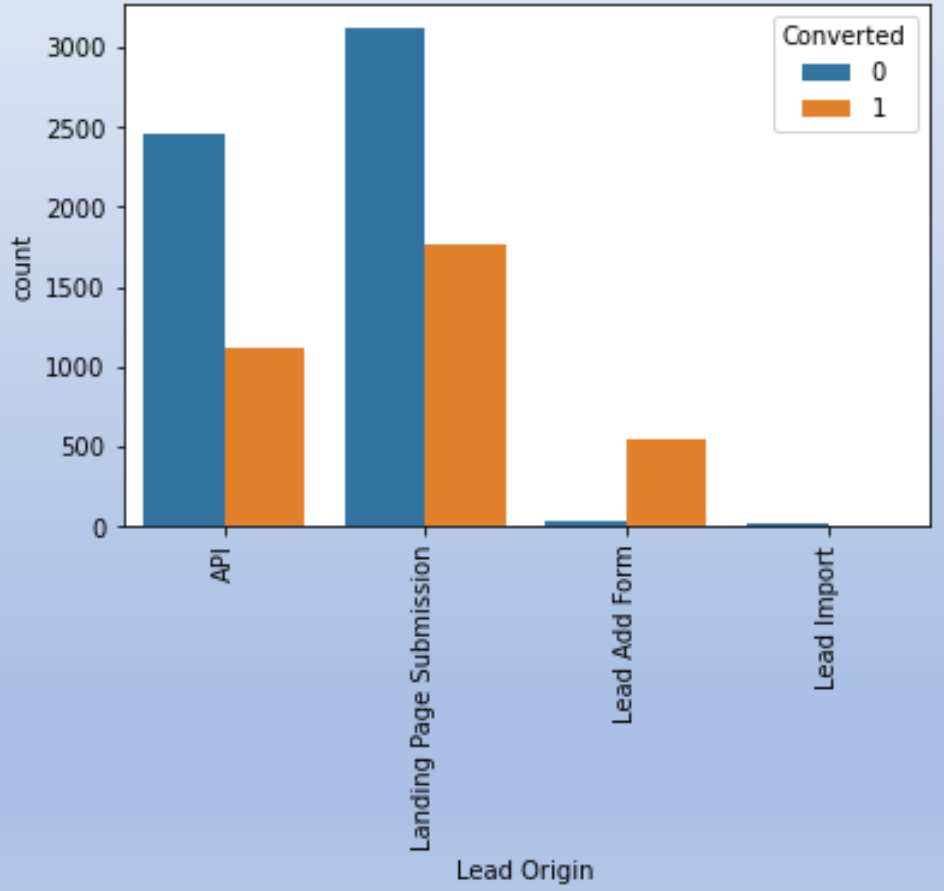
There are quite a few goals for this case study.

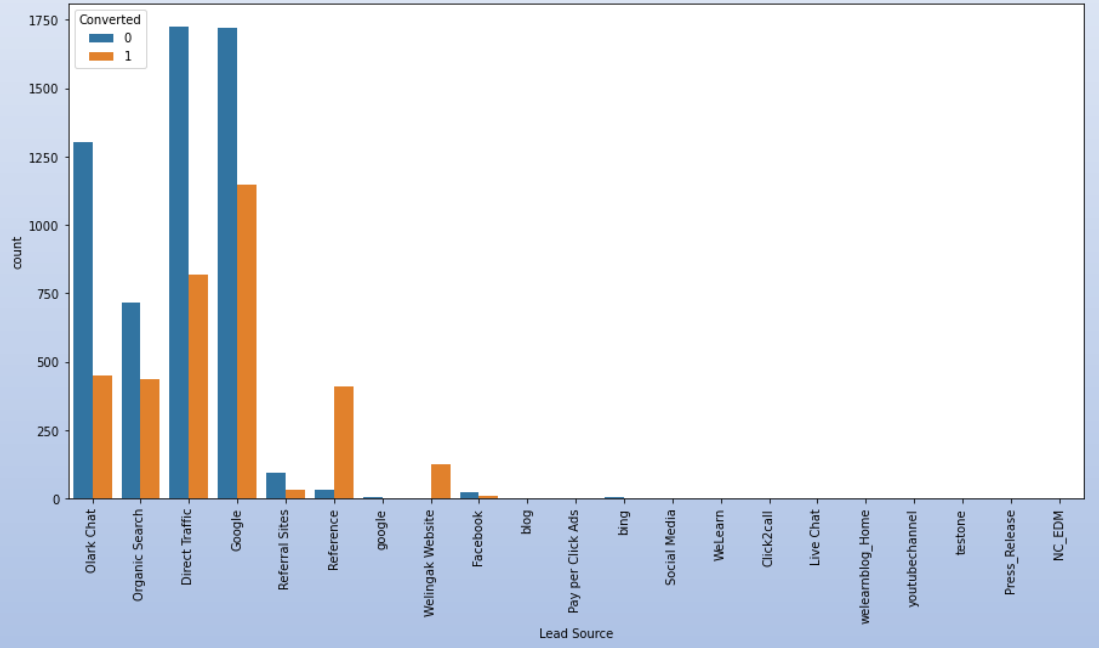
•Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

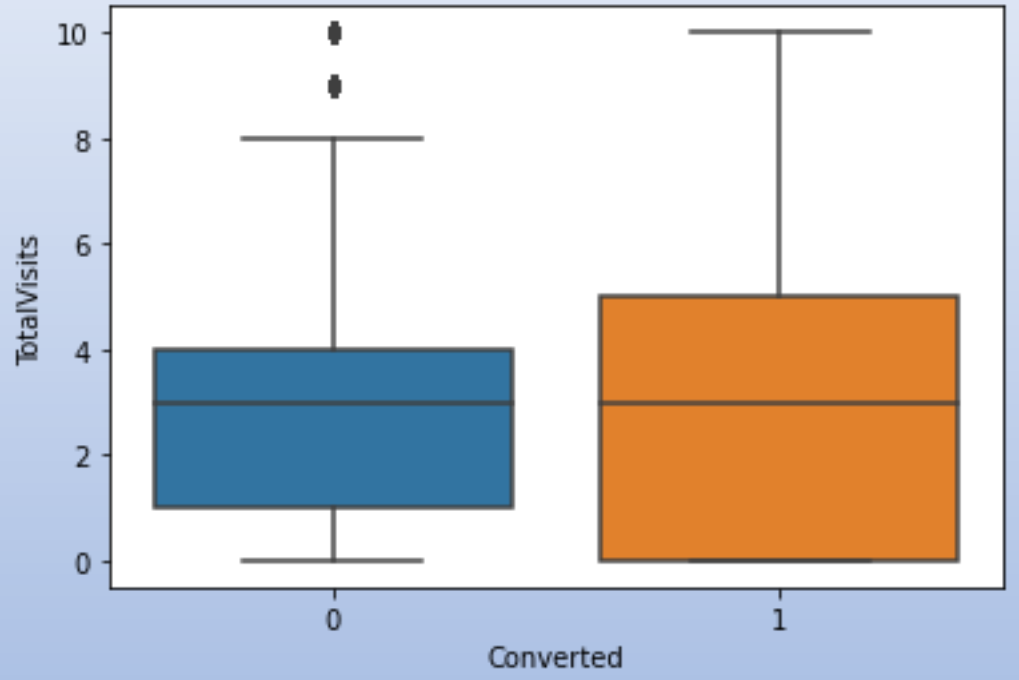
•There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

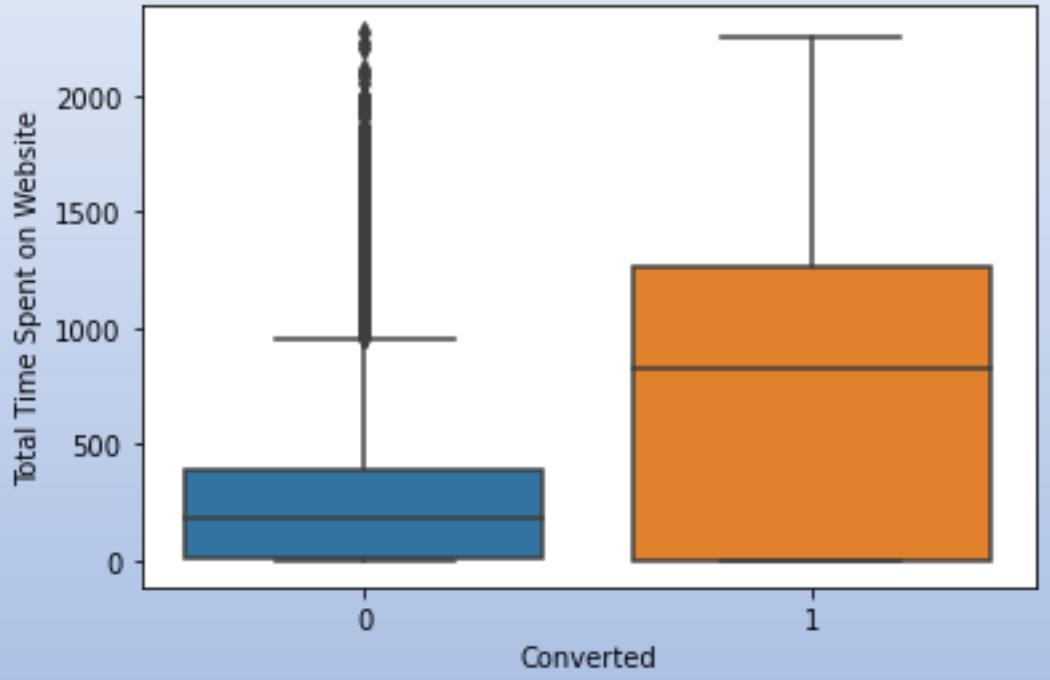
STEPS INVOLVED:

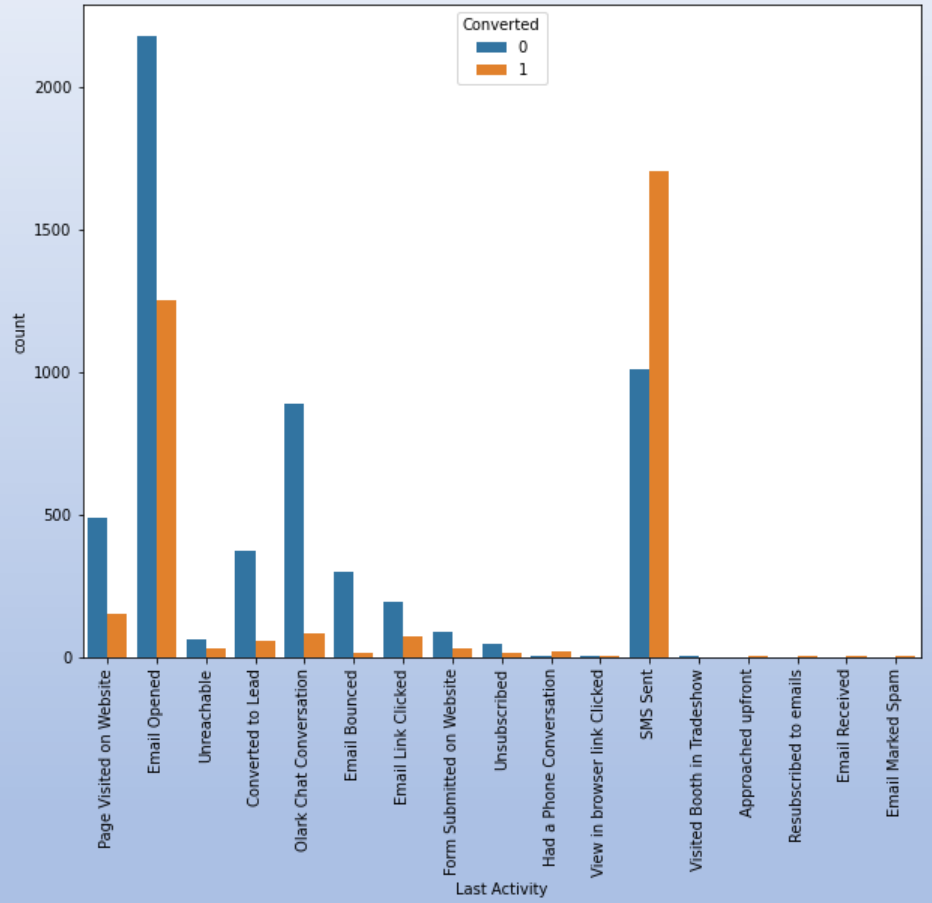
EXPLORATORY DATA ANALYSIS:

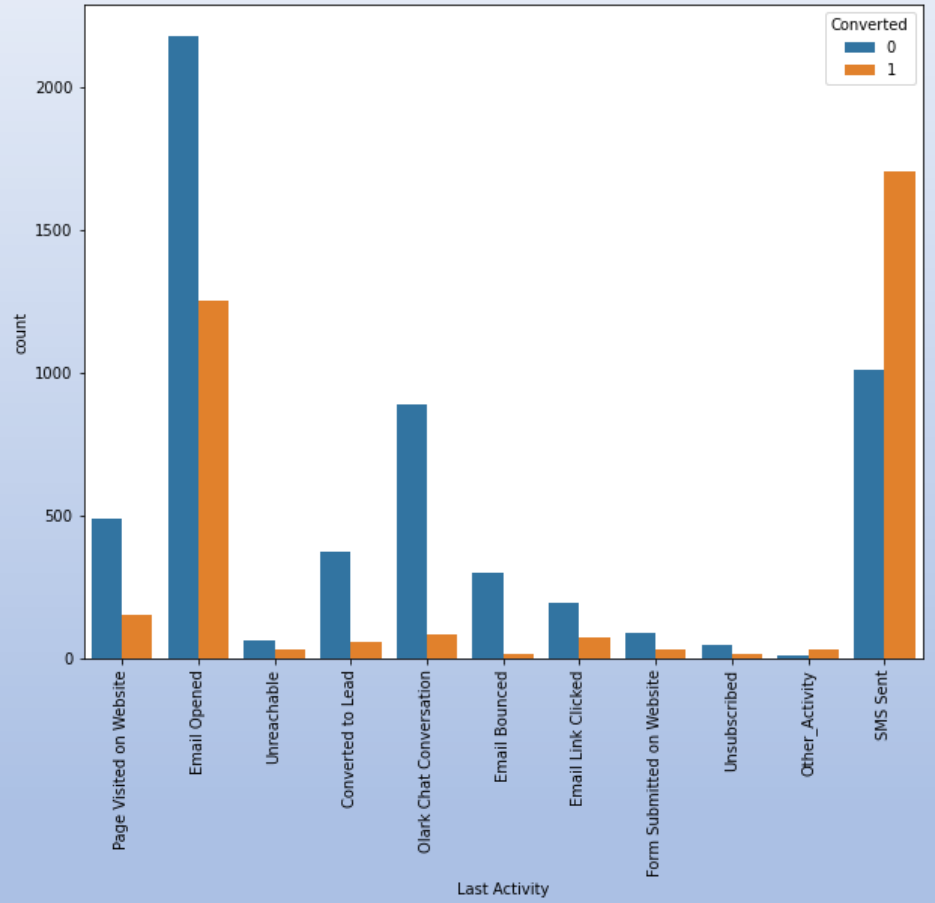


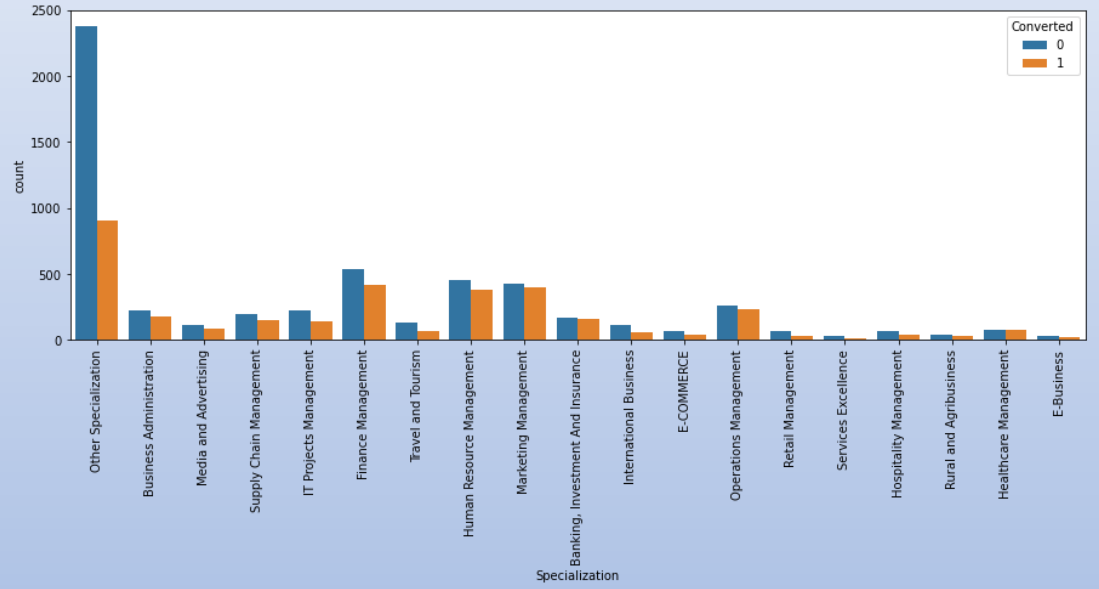


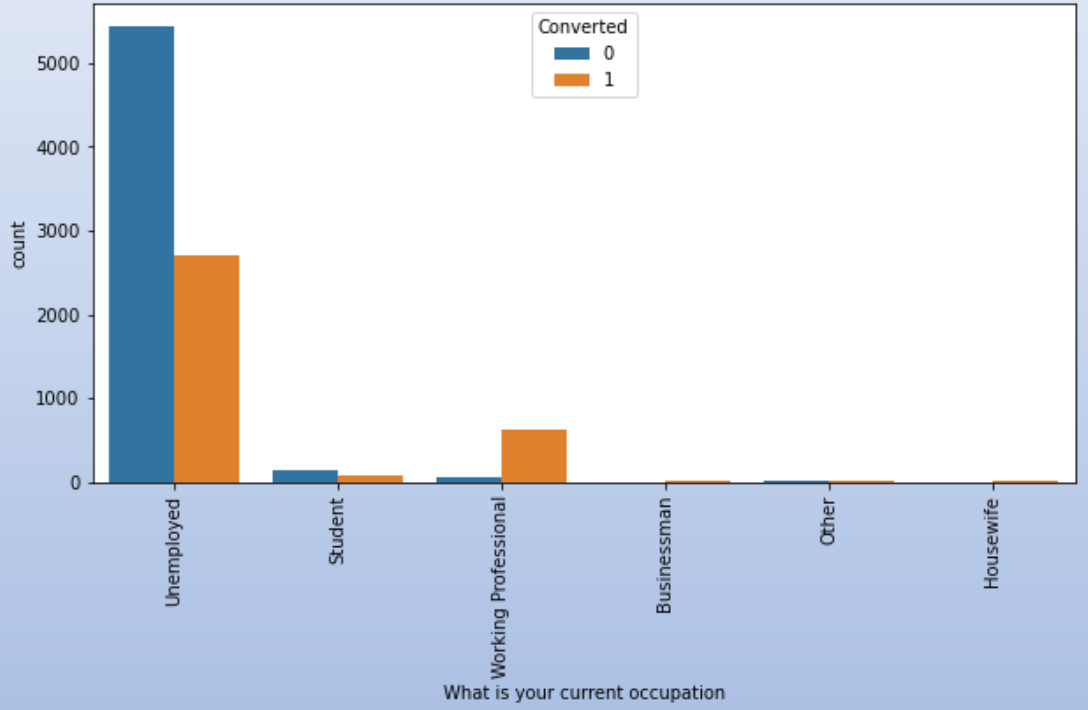

TotalVisits
10
8
6
4
2
0
0
1
Converted

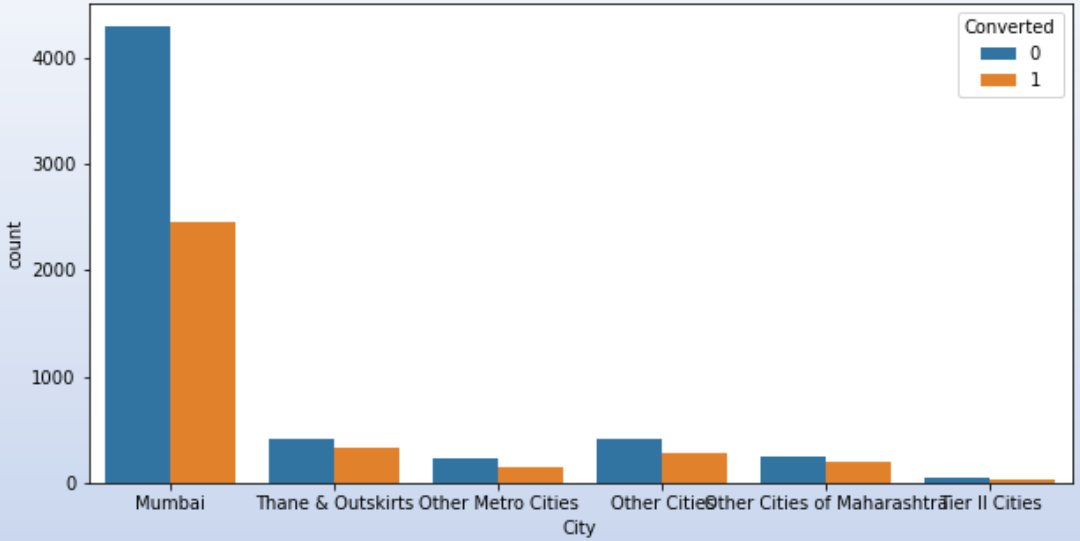

Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted

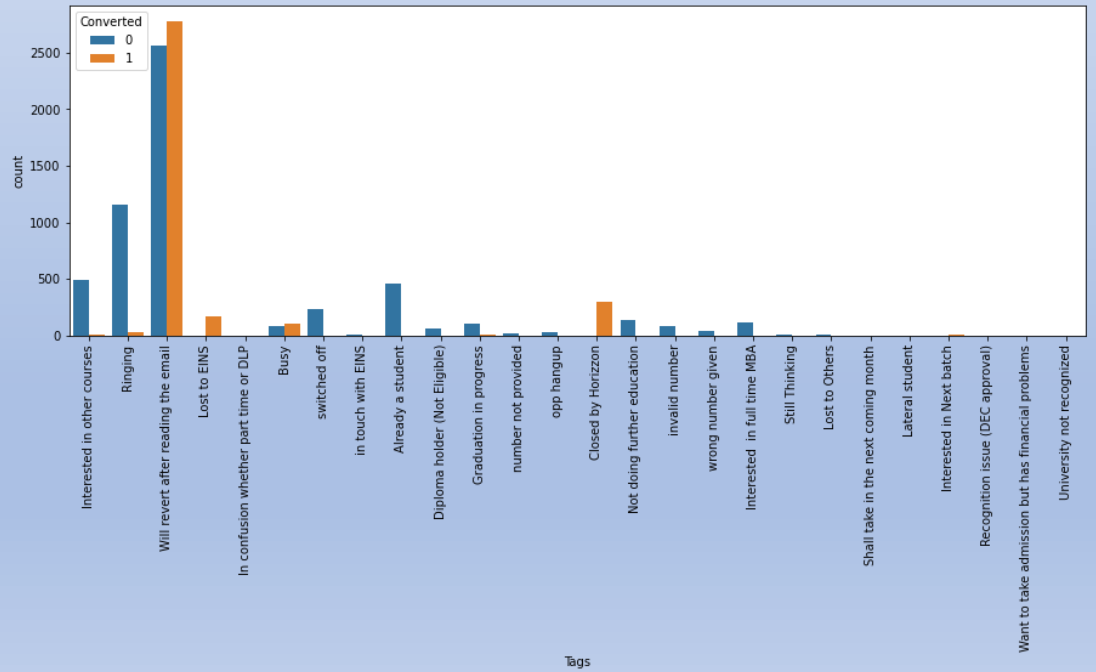

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

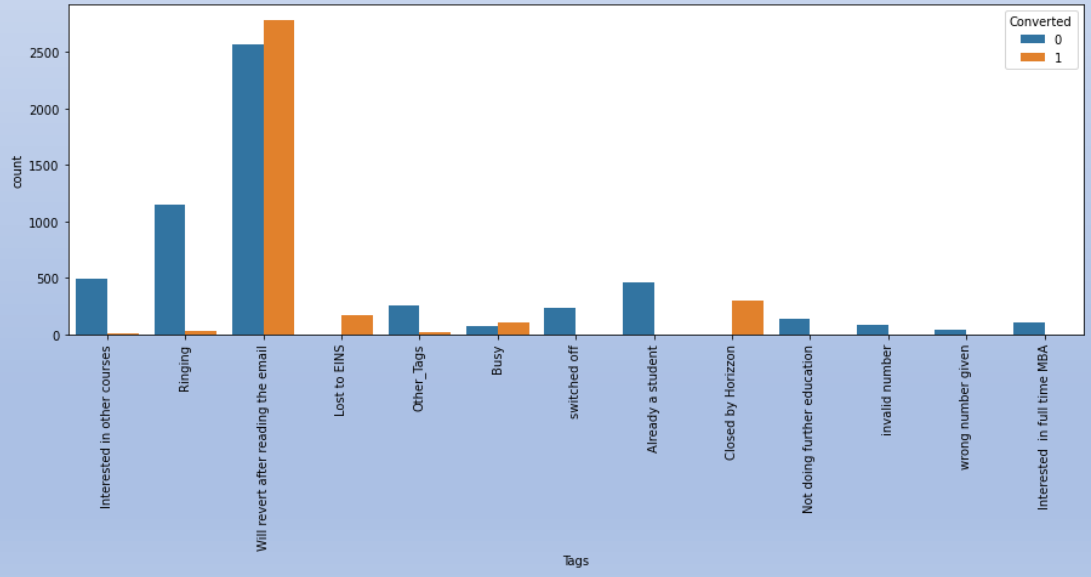

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Other_Activity
SMS Sent
Last Activity


Converted
0
1
count
2500
2000
1500
1000
500
0
Other Specialization
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


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


Converted
0
1
count
4000
3000
2000
1000
0
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City

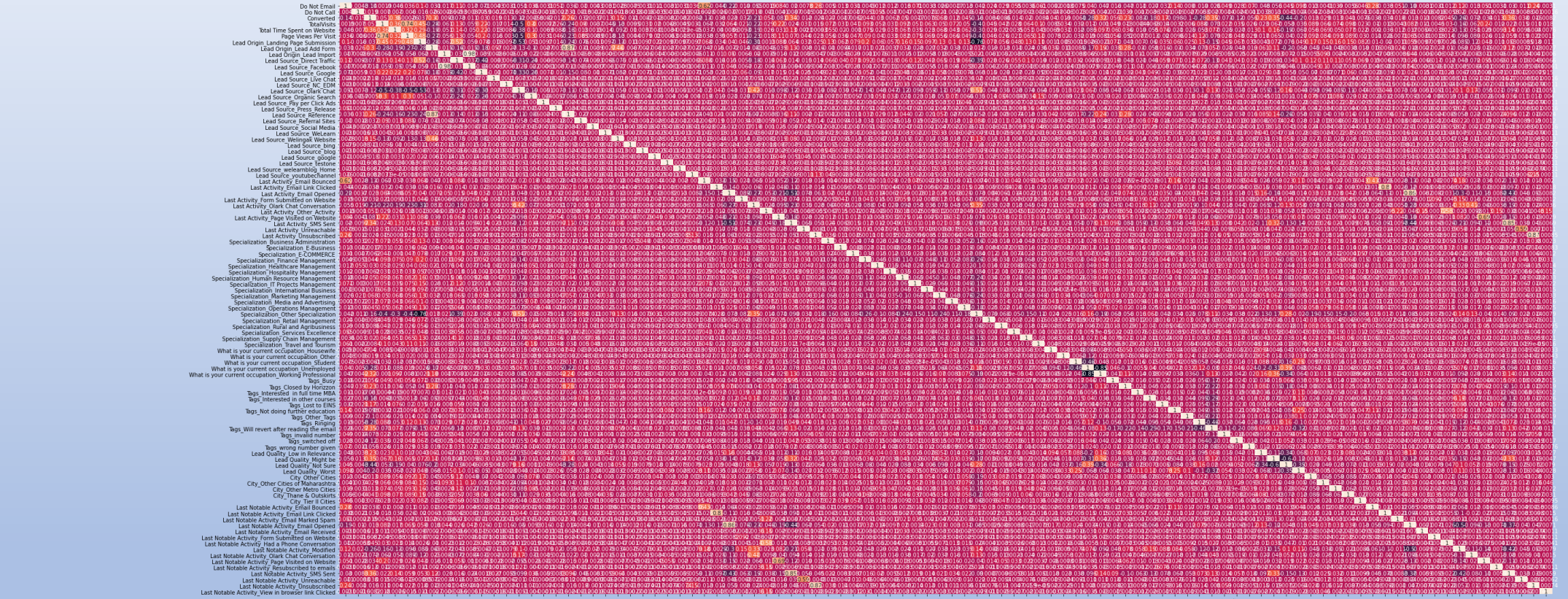

Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
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
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Tags

INFERENCES BASED ON EDA:

- API and Landing Page Submission have more conversion rate among the Lead origins.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin.
- We are getting maximum number of leads from Google and Direct traffic generates
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads at the same time focus on generate more leads from reference and welingak website.
- Leads spending more time on the website are more likely to be converted so we need to focus on engaging more in website.
- Most of the leads generated by Email opened as their last activity.
- Conversion rate for leads with last activity is more with SMS sent.
- leads and conversion rate are high in finance, marketing and HR Specialization.
- Working Professionals going for the course have high chances of Conversion rate, may be they are trying for career transition.
- Most lead generates from Unemployed with less conversion rate.
- Most conversion rates are from 'will revert after reading the email’ , 'Lost to EINS’ ,'Closed by Horizzon’.
- Most leads are from Mumbai with around considerable conversion rate.

CORRELATIONS:



PLOTTING ROC CURVE:

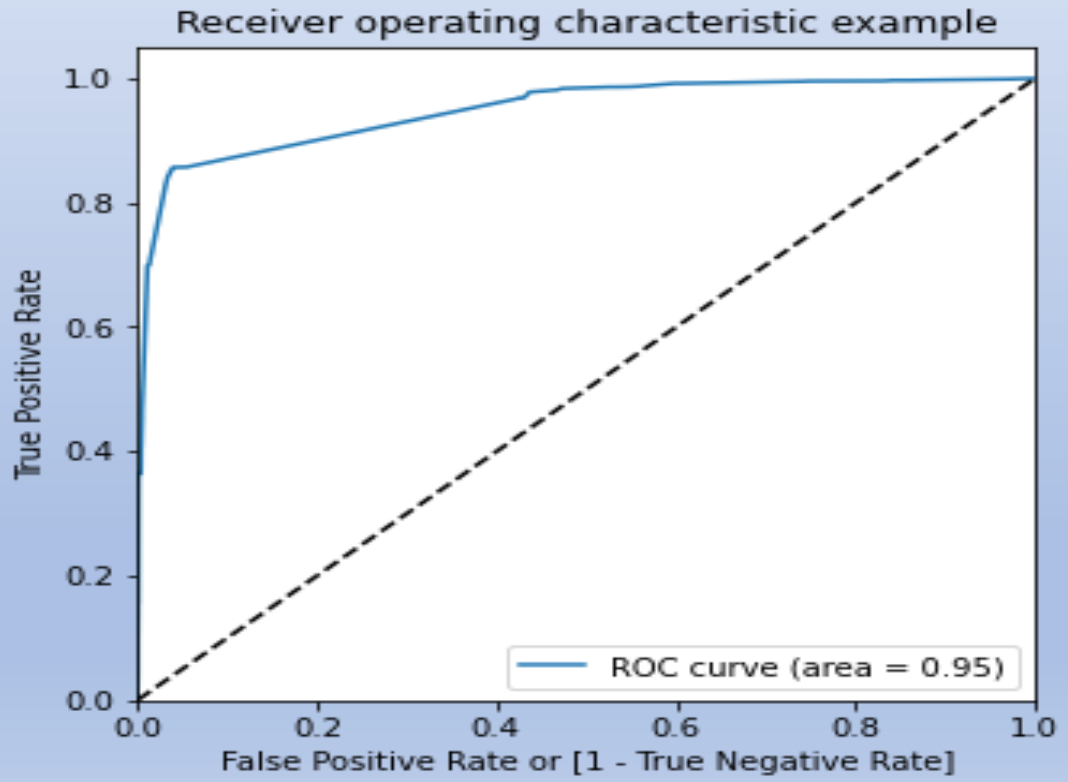
An ROC curve demonstrates several things:

It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

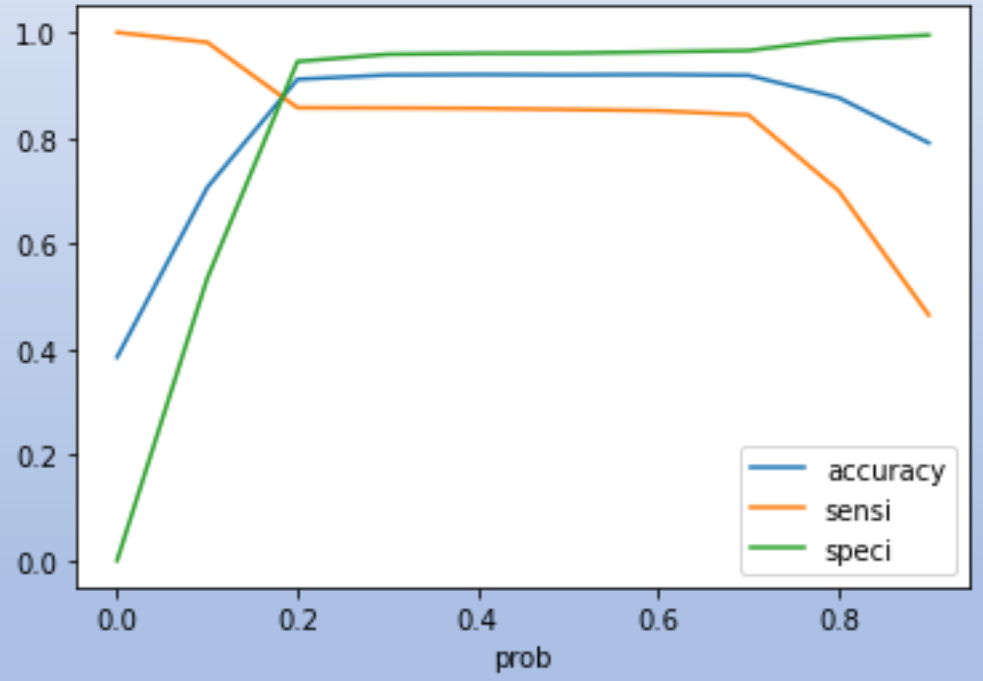
The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

The area under the curve is 0.95, so the model is working fine.



FINDING OPTIMAL CUTOFF POINT:

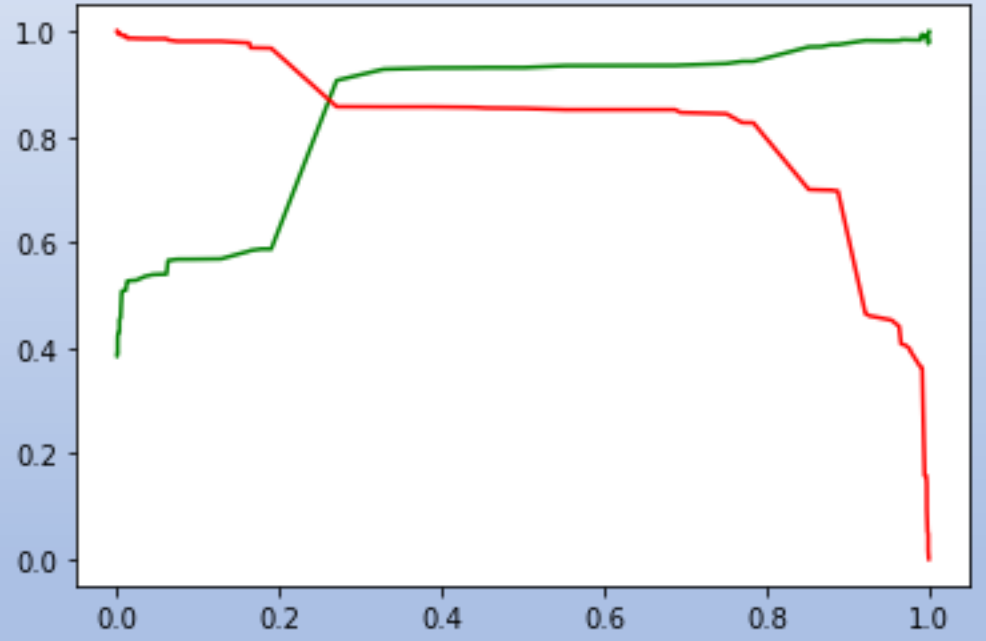
Optimal cutoff probability is that prob where we get balanced sensitivity and specificity



PRECISION AND RECALL:

Precision: Precision is no more than the ratio of True Positive and the sum of True Positive and False Positive.
Recall: The recall is none other than the ratio of True Positive and the sum of True Positive and False Negative.

As per the graph, the intersection point is at 0.38 probability



OBSERVATIONS:

Observation: After running the model on the Test Data these are the figures we obtain:
Accuracy : 90.78%
Sensitivity : 84.14%
Specificity : 94.57%

FINAL OBSERVATIONS:

Let us compare the values obtained for Train & Test:
Train Data:
Accuracy : 91.11%
Sensitivity : 85.73%
Specificity : 94.49%
Test Data:
Accuracy : 90.78%
Sensitivity : 84.14%
Specificity : 94.57%

CONCLUSION:

1. The accuracy, precision and recall/sensitivity are showing promising score in test sets which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.

2. In business terms, this model has an ability to adjust with the company's requirements in coming future.

3. This shows that the model is in stable state and X education can achieve it's CEO's target of 80% lead conversion rate with this model.

The top 5 variables which are important for the lead conversion is:

1. Tags_Closed by Horizzon
2. Tags_Lost to EINS
3. Tags_Will revert after reading the email
4. Lead Source_Welingak Website
5. Tags_Busy