Lead Scoring Case Study

- Harshit Vindoorty, Akshay Loya, Nirajsing Patil

Problem Statement:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, but its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know the most promising leads for which they want to build a Model which identifies the hot leads.
- Deployment of the model for future use.

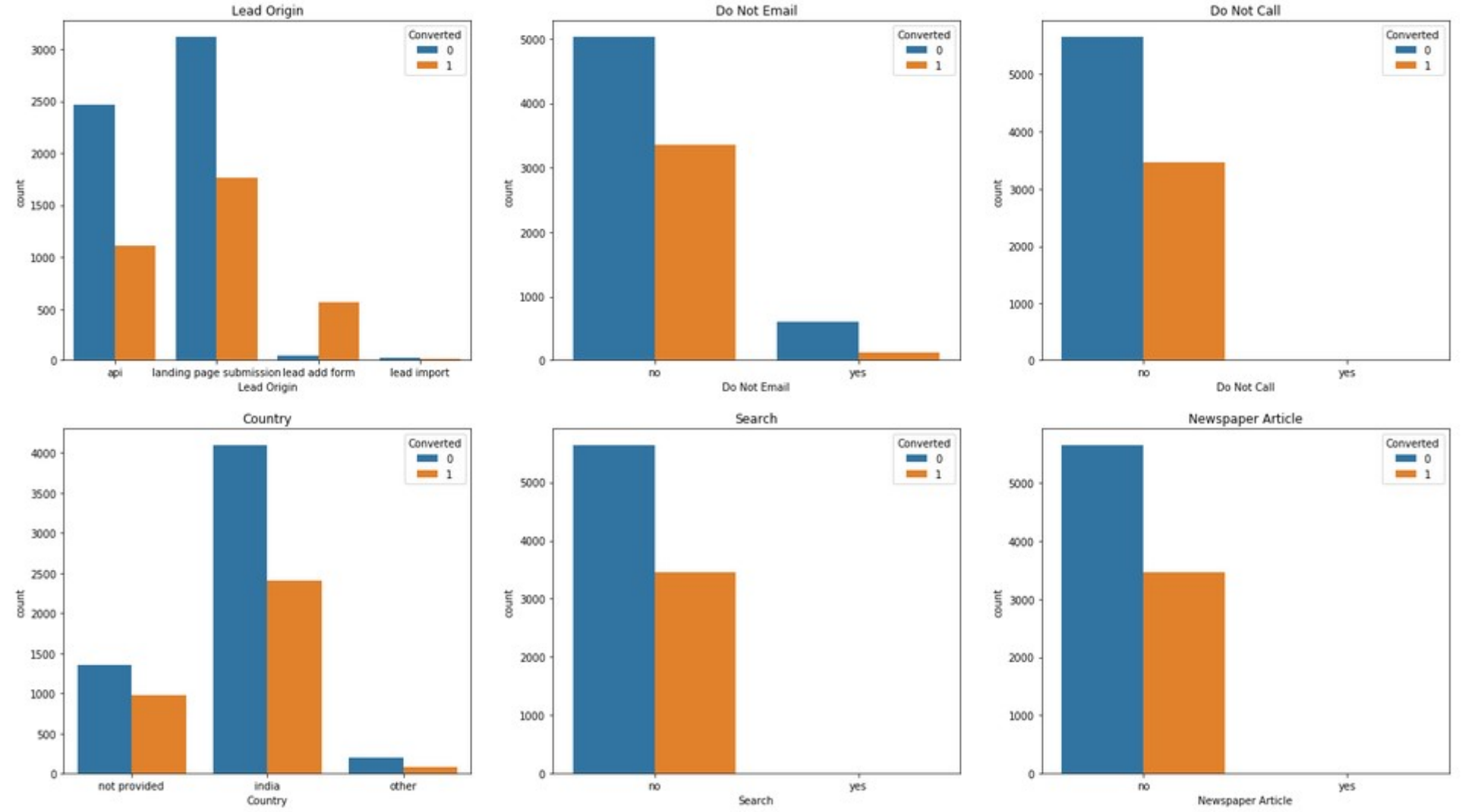
Solution Methodology:

- Data cleaning and data manipulation.
  - Check and handle duplicate data
  - Check and handle missing values
  - Imputation of the values, if necessary.
  - Check and handle outliers.
- EDA
- Feature scaling
- Dummy Variables and encoding of the data.
- Classification technique : logistic regression used for model prediction
- Validation of the model
- Conclusion and recommendations

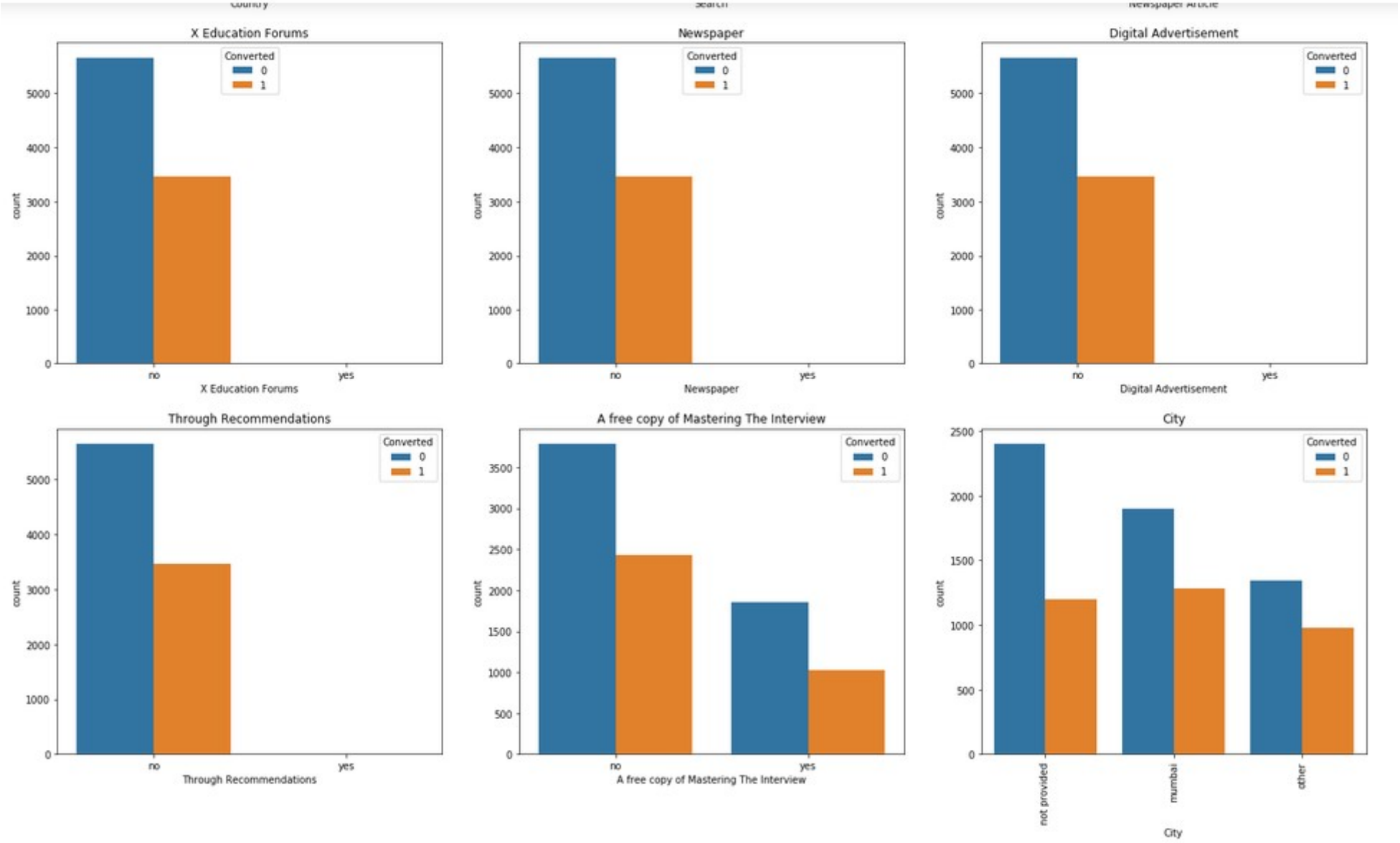
Data Manipulation:

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 40% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’ and more.

EDA:

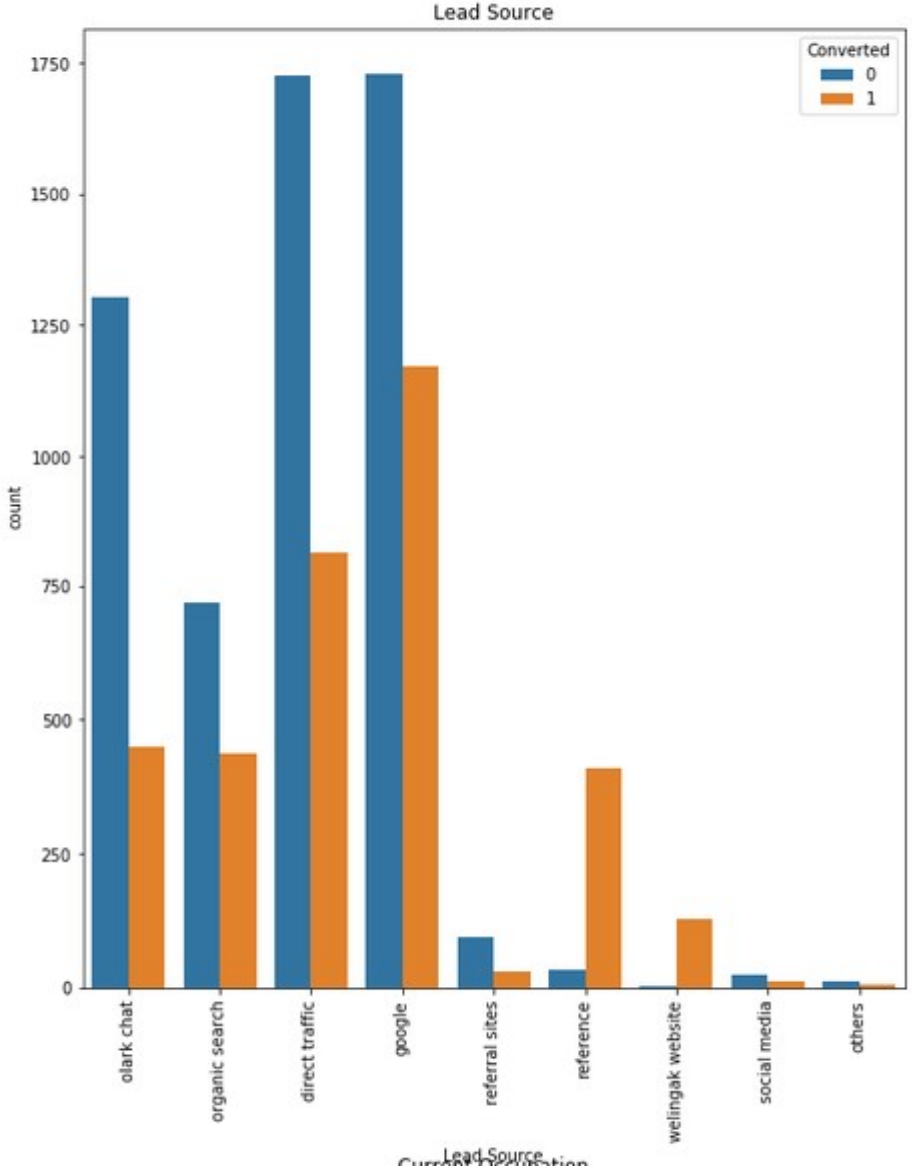

Lead Origin
Converted
0
1
count
Lead Origin
api
landing page submission
lead add form
lead import
Do Not Email
Do Not Email
no
yes
Do Not Call
Do Not Call
no
yes
Country
Country
not provided
india
other
Search
Search
no
yes
Newspaper Article
Newspaper Article
no
yes

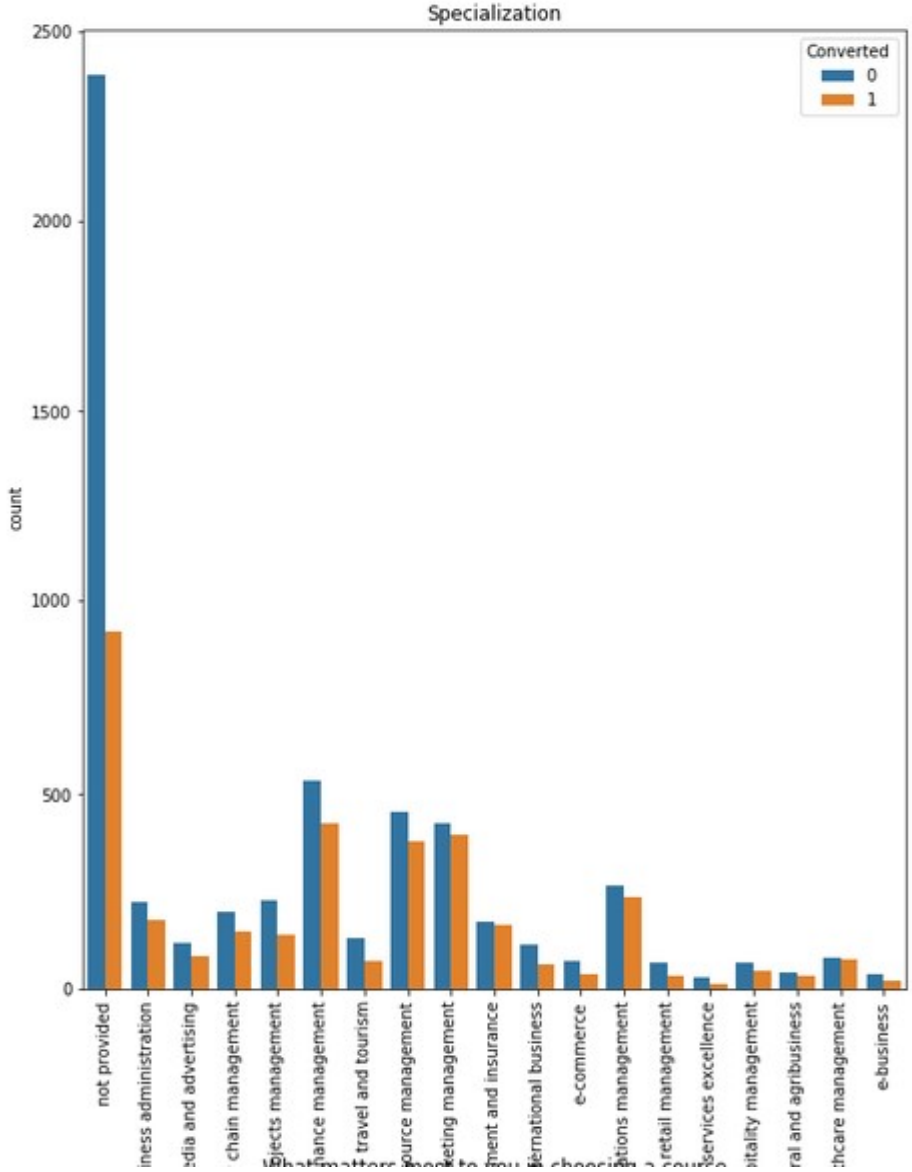
EDA:

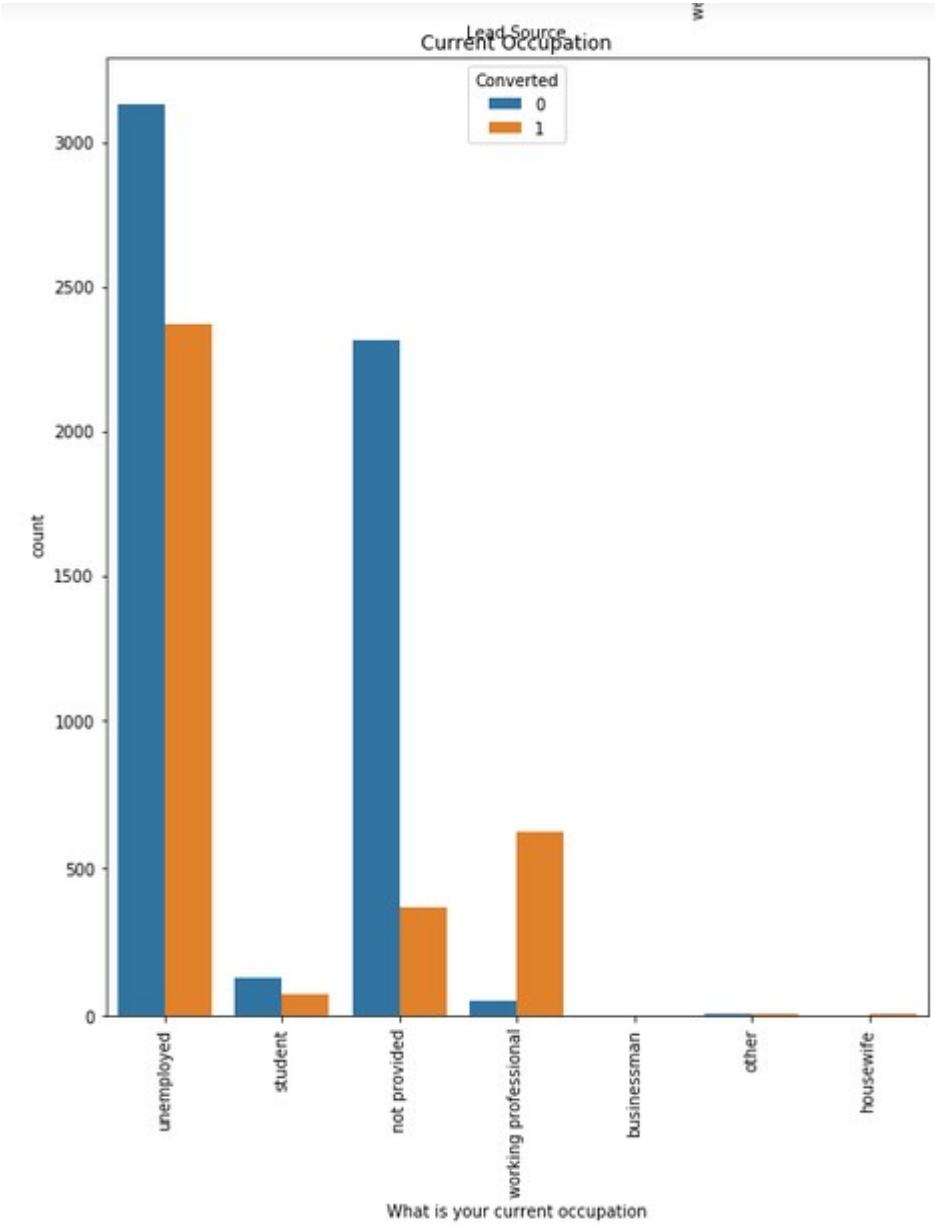

X Education Forums
Converted
0
1
count
no
yes
X Education Forums
Newspaper
Converted
0
1
count
no
yes
Newspaper
Digital Advertisement
Converted
0
1
count
no
yes
Digital Advertisement
Through Recommendations
Converted
0
1
count
no
yes
Through Recommendations
A free copy of Mastering The Interview
Converted
0
1
count
no
yes
A free copy of Mastering The Interview
City
Converted
0
1
count
not provided
mumbai
other
City

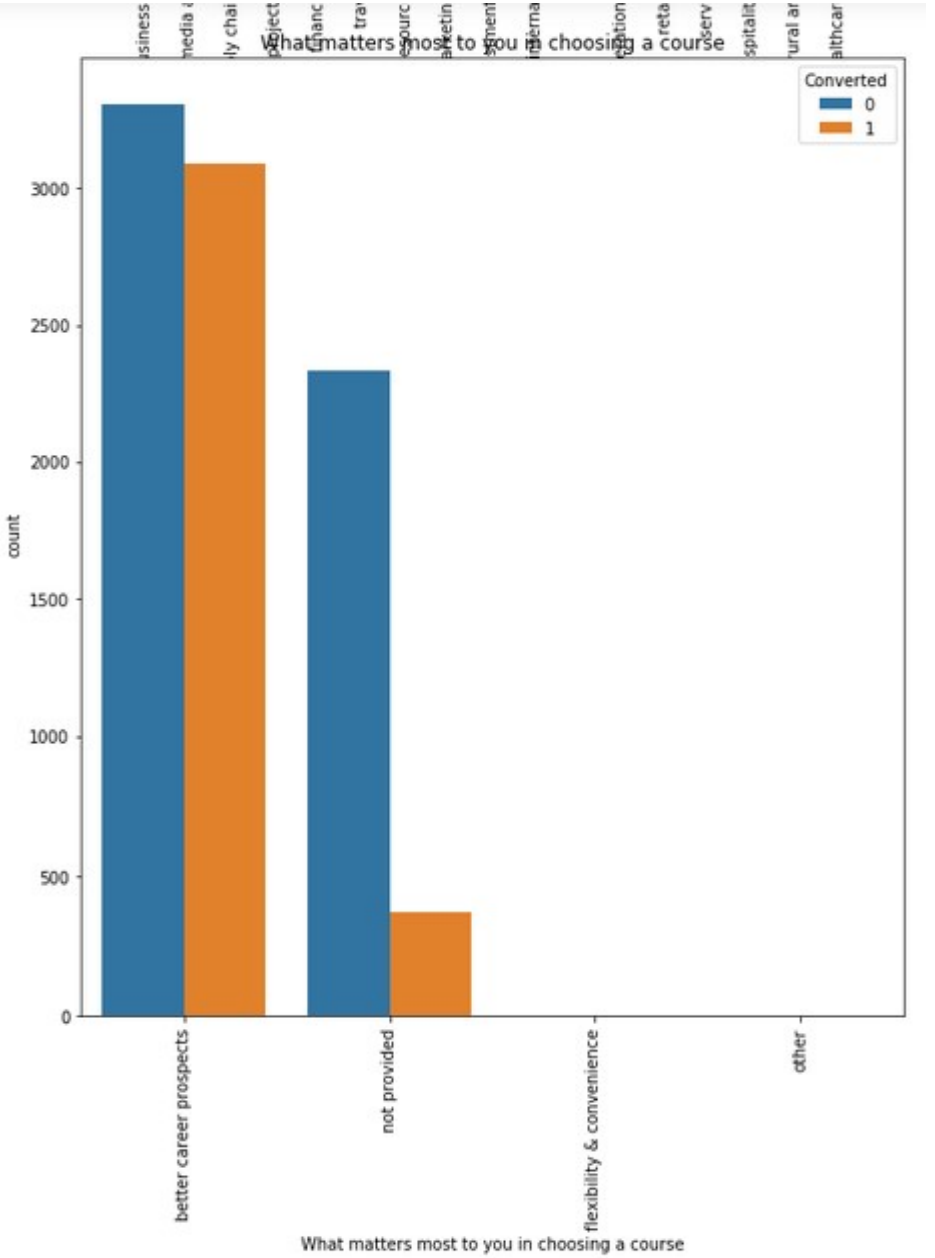
Inference:

- API and Landing Page Submission bring higher number of leads as well as conversion
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads. In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.


Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
social media
others
Lead Source


Specialization
Converted
0
1
count
2500
2000
1500
1000
500
0
not provided
e-commerce
e-business

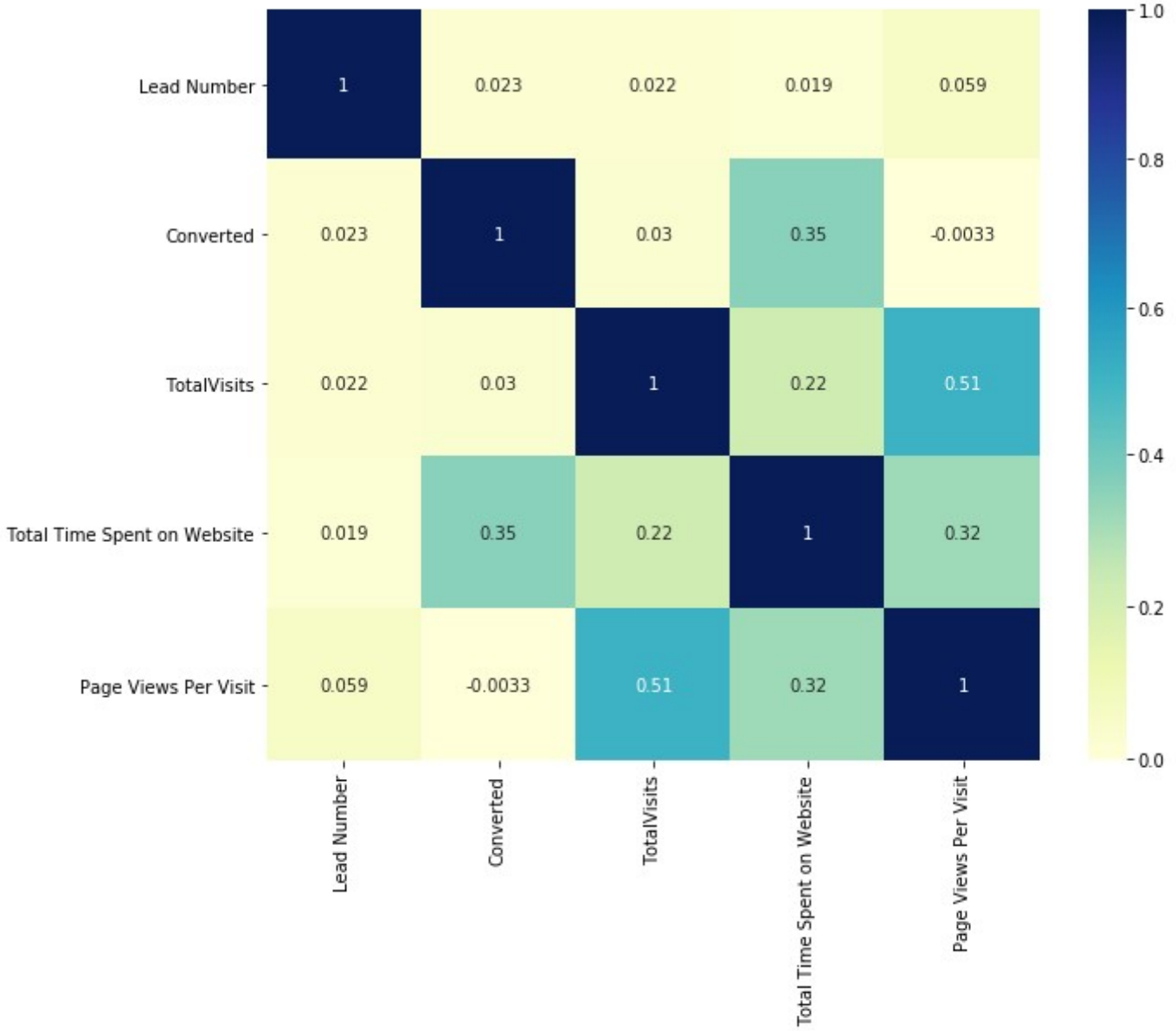

Lead Source
Current Occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation


What matters most to you in choosing a course
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

Inference:

- We see that specialization with Management in them have higher number of leads as well as leads converted. So this is definitely a significant variable.
- Working Professionals going for the course have high chances of joining it. Unemployed leads are the most in terms of Absolute numbers.
- Maximum number of leads are generated by Google and Direct traffic. 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 and generate more leads from reference and welingak website.

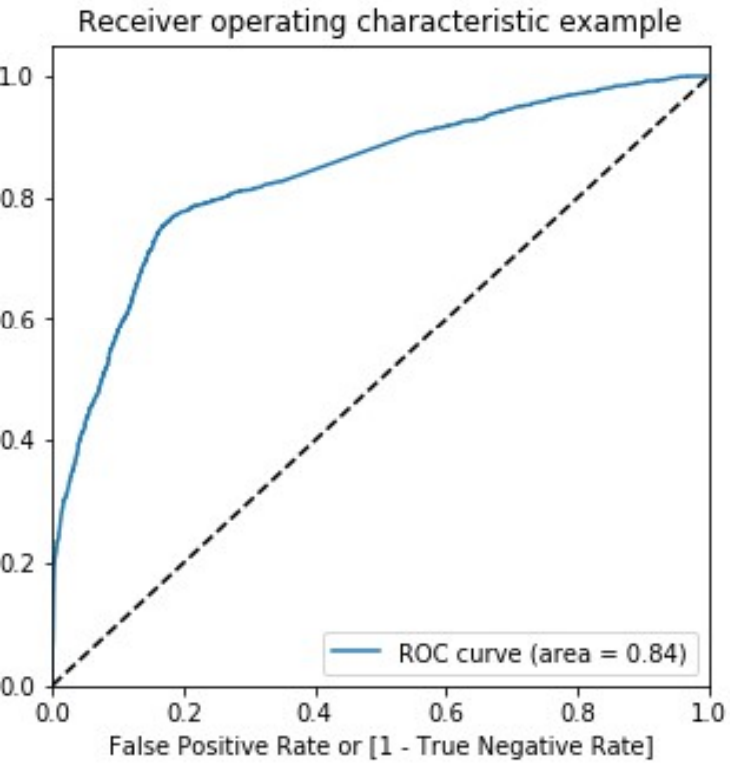
Numerical Relations:

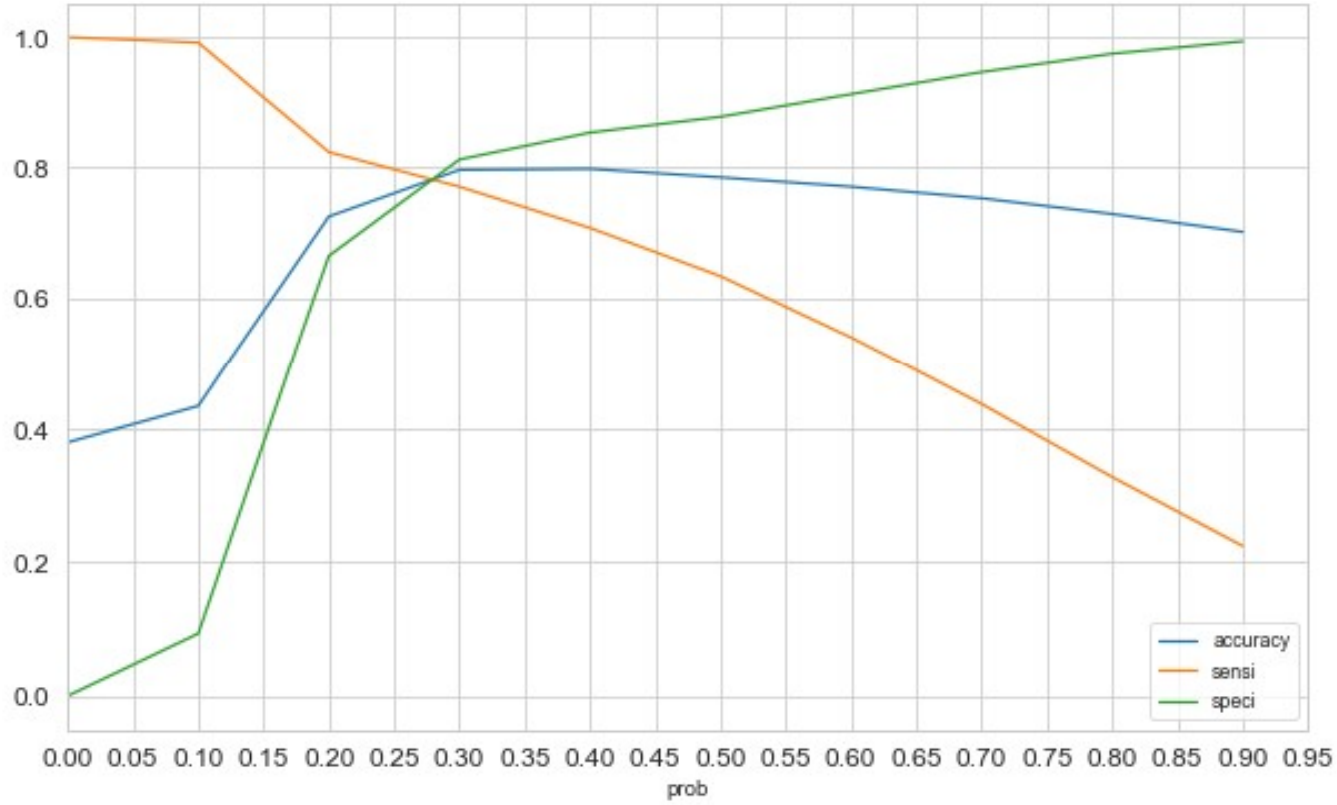


Model Building:

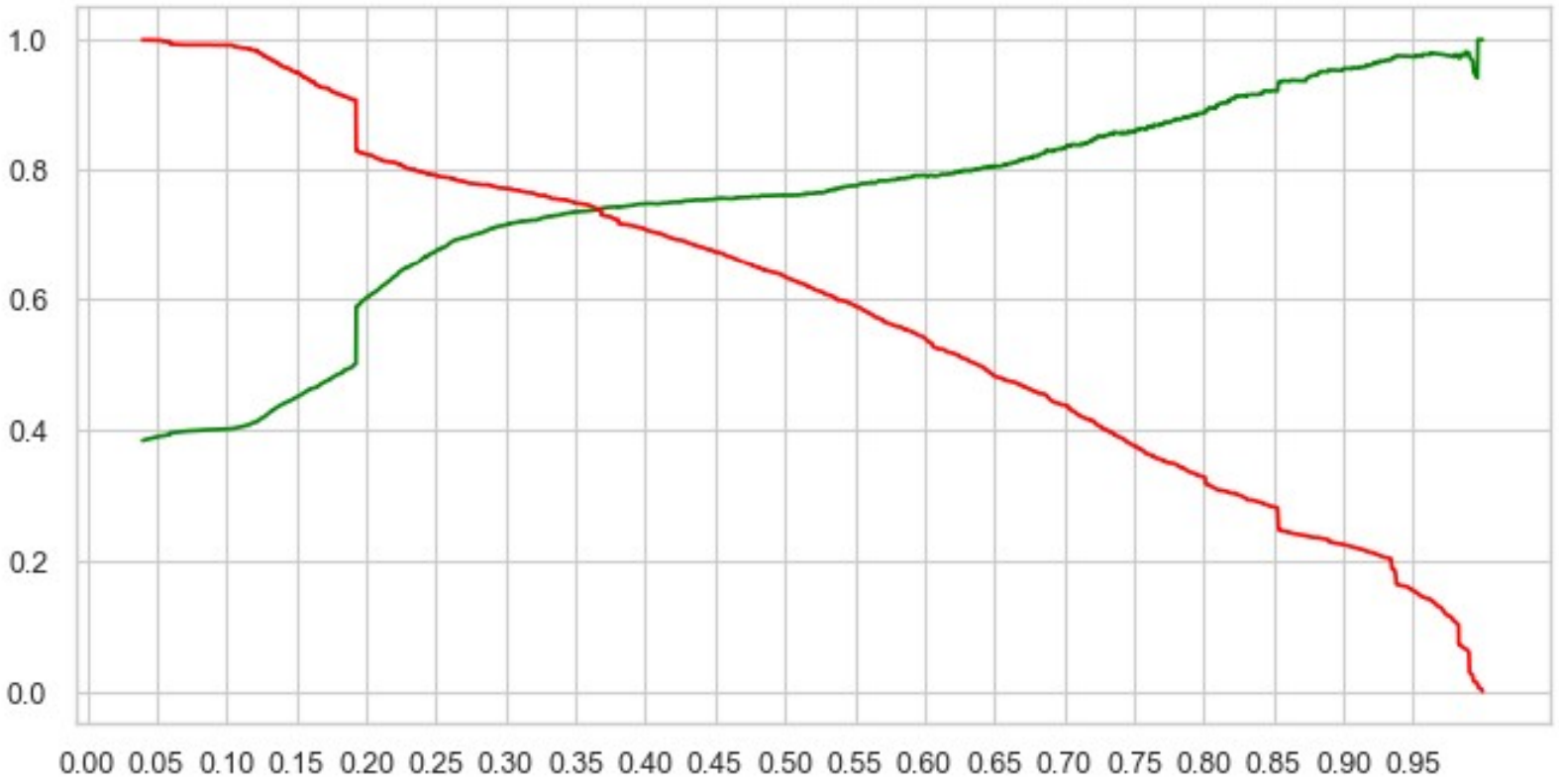
- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for feature selection.
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Prediction on test data sets.
- Overall accuracy 80%.

ROC Curve:


Receiver operating characteristic example
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.84)
False Positive Rate or [1 - True Negative Rate]


1.0
0.8
0.6
0.4
0.2
0.0
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95
prob
accuracy
sensi
speci

- Finding optimal cut off point
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.29.



Observation:

Final Observation:

Let us compare the values obtained for Train Data & Test Data:

Train Data:
Accuracy : 79%
Sensitivity : 78%
Specificity : 80%
Precision : 71%
Recall : 78%

Test Data:
Accuracy : 80%
Sensitivity : 78%
Specificity : 80%
Precision : 72%
Recall : 78%

Conclusion:

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
- When the lead source was:

a. Welingak website

b. Olark Chat

- The total time spent on the Website.
- When the city is Mumbai.

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.