Lead Score Case Study

Submitted by:
Rohith Vaddeti

Business Problem Statement

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. On any given day, many professionals who are interested in the courses land on their website and browse for courses. •
- Once these people land on the website, they might browse the course or fill up the form for the course.
- 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 Goals:

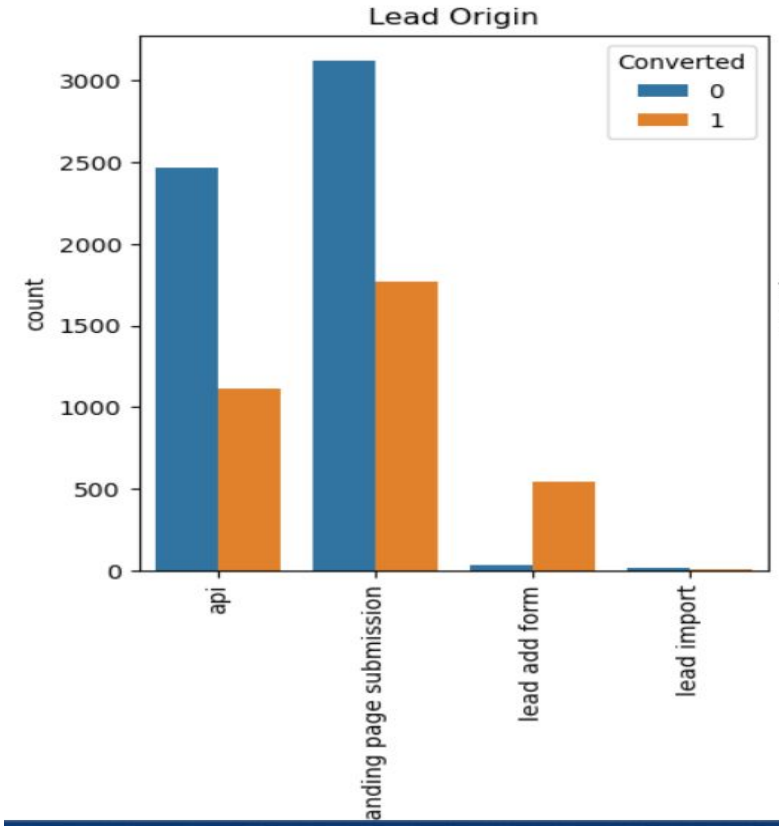
- Building a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which this model should be able to adjust to if the company's requirement changes in the future so I will need to handle these as well.
- The CEO, in particular, has given a ballpark of the target lead conversion lead to be around 80%.

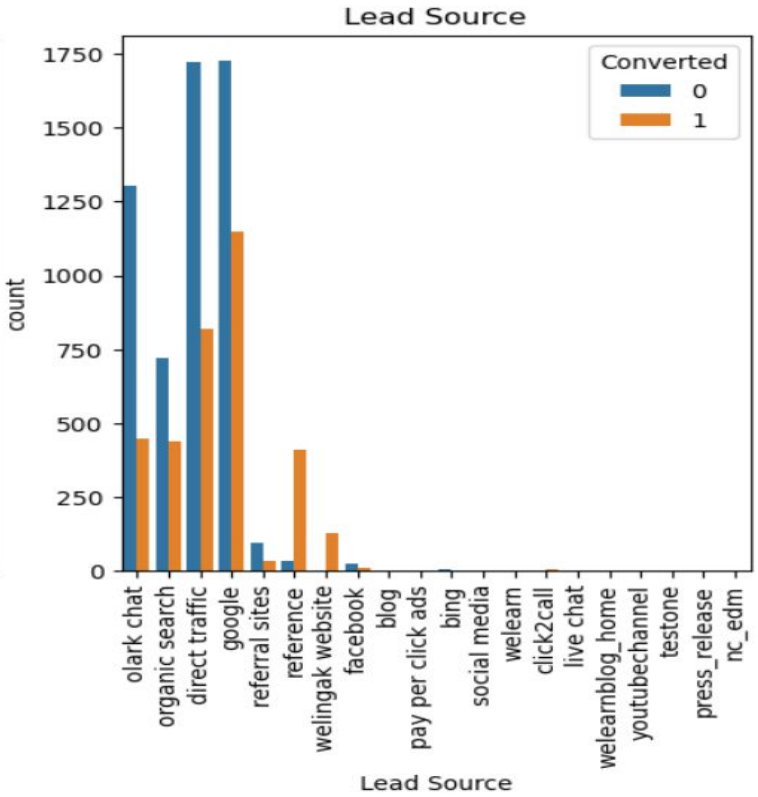
Solution Methodology:

- Data cleaning and data manipulation
- Data quality check
- EDA
- Splitting the data into test and train data set
- Building a logistic regression model
- Model evaluation using sensitivity and specificity or precision and recall •
- Applying the best model in Test data based on sensitivity and specificity metrics.

Exploratory Data Analysis

Lead origin vs Lead Source conversion rate

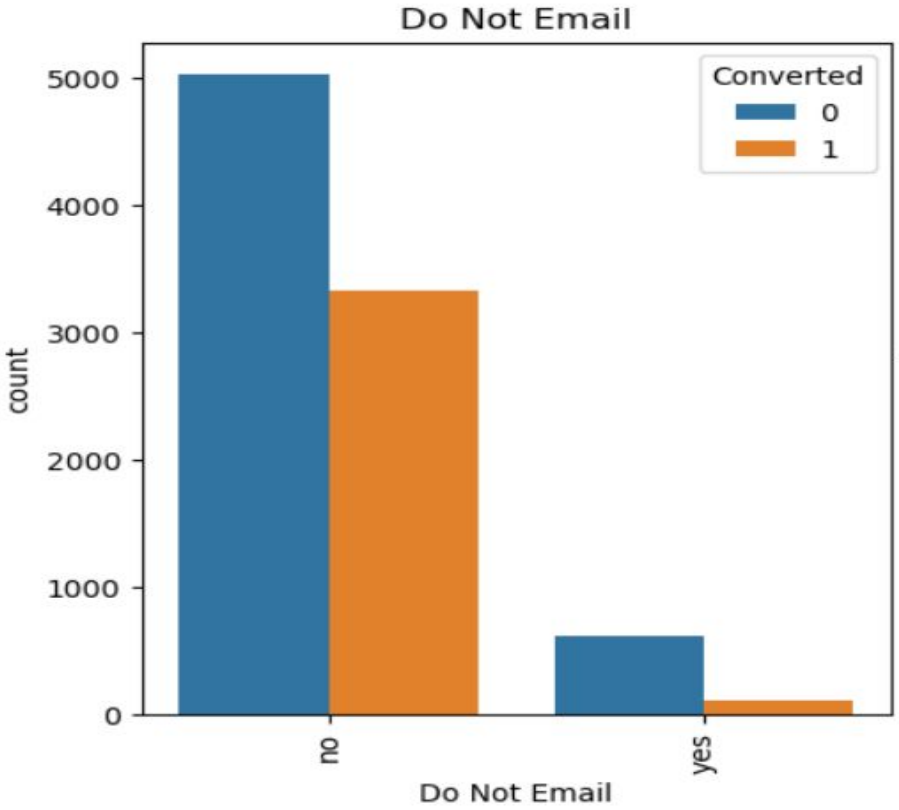


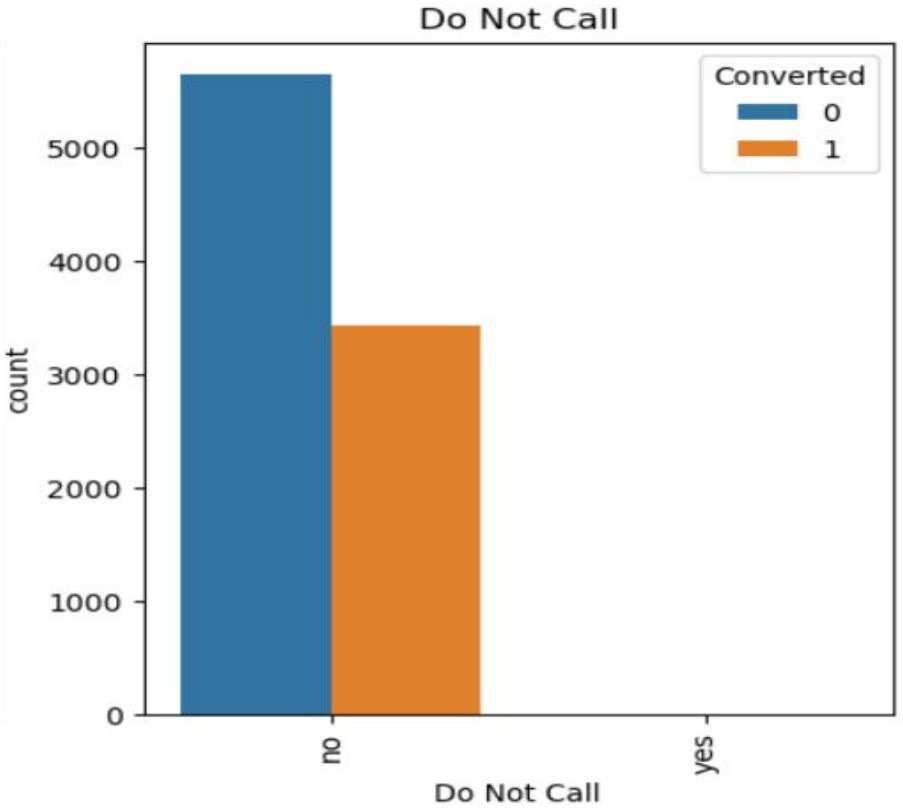


Conclusion:

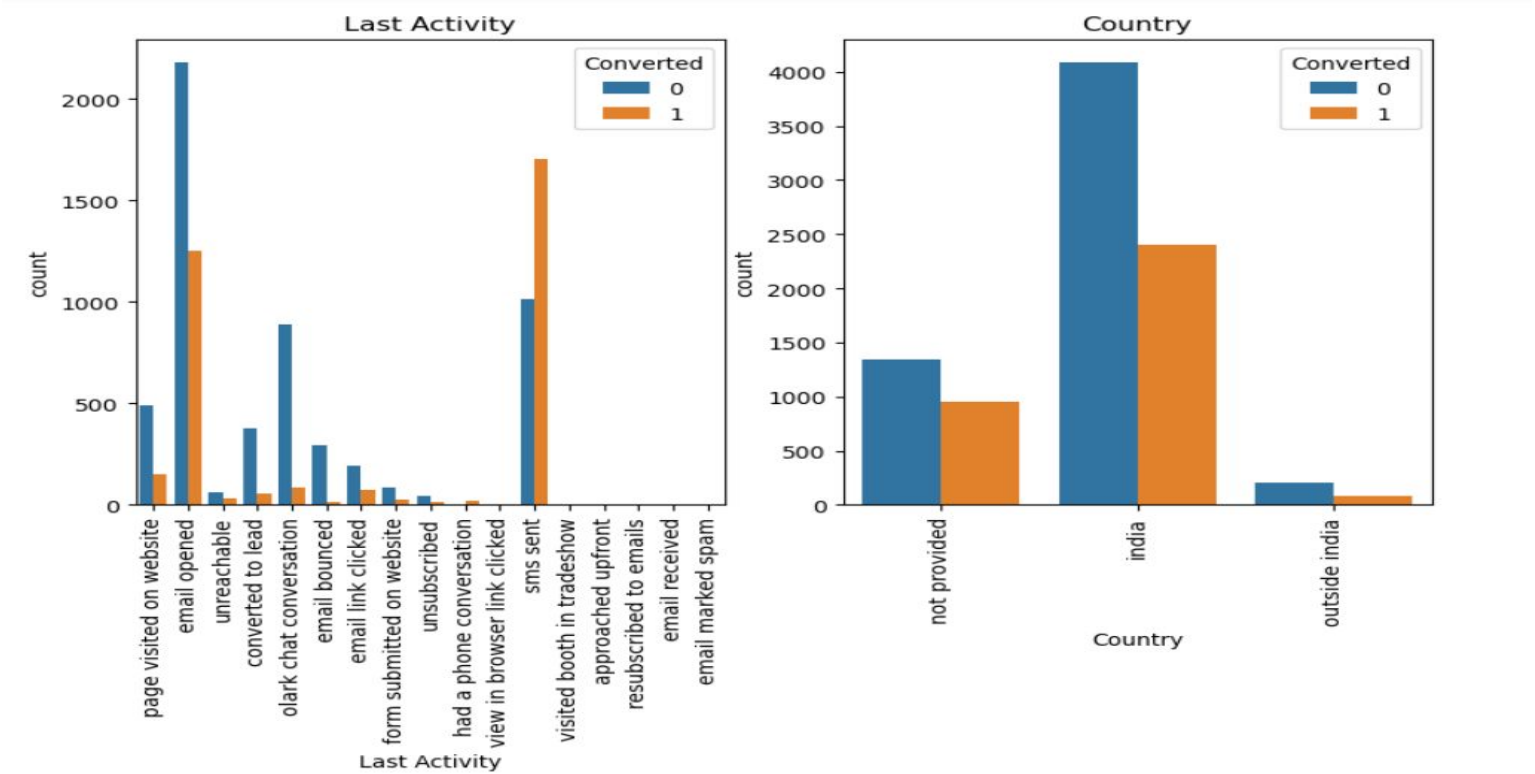
- We can see maximum number of leads from are Google and direct traffic.
- Olark chat, organic search show so many leads who are not getting converted, so we need to focus into those.
- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high 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.

Impact on leads by email and sms sent

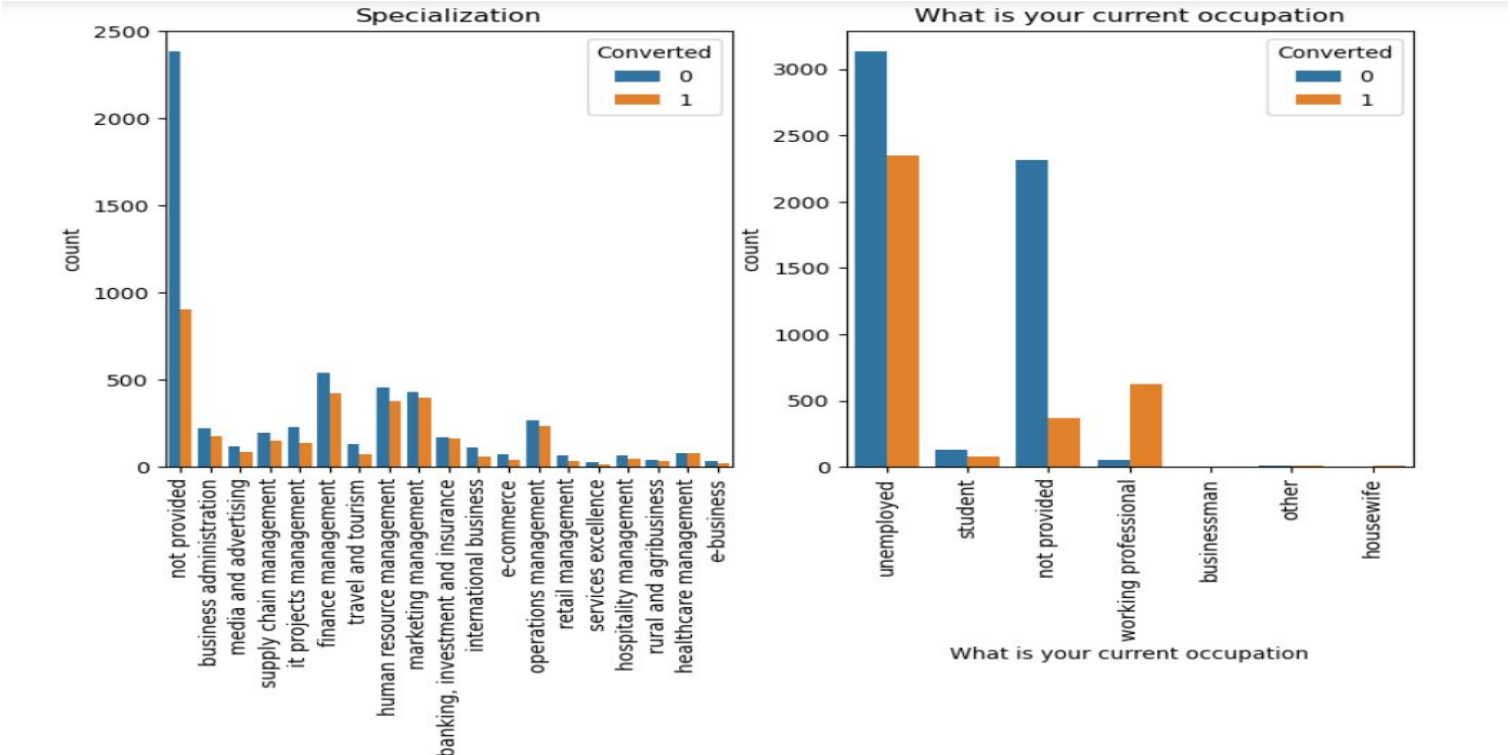




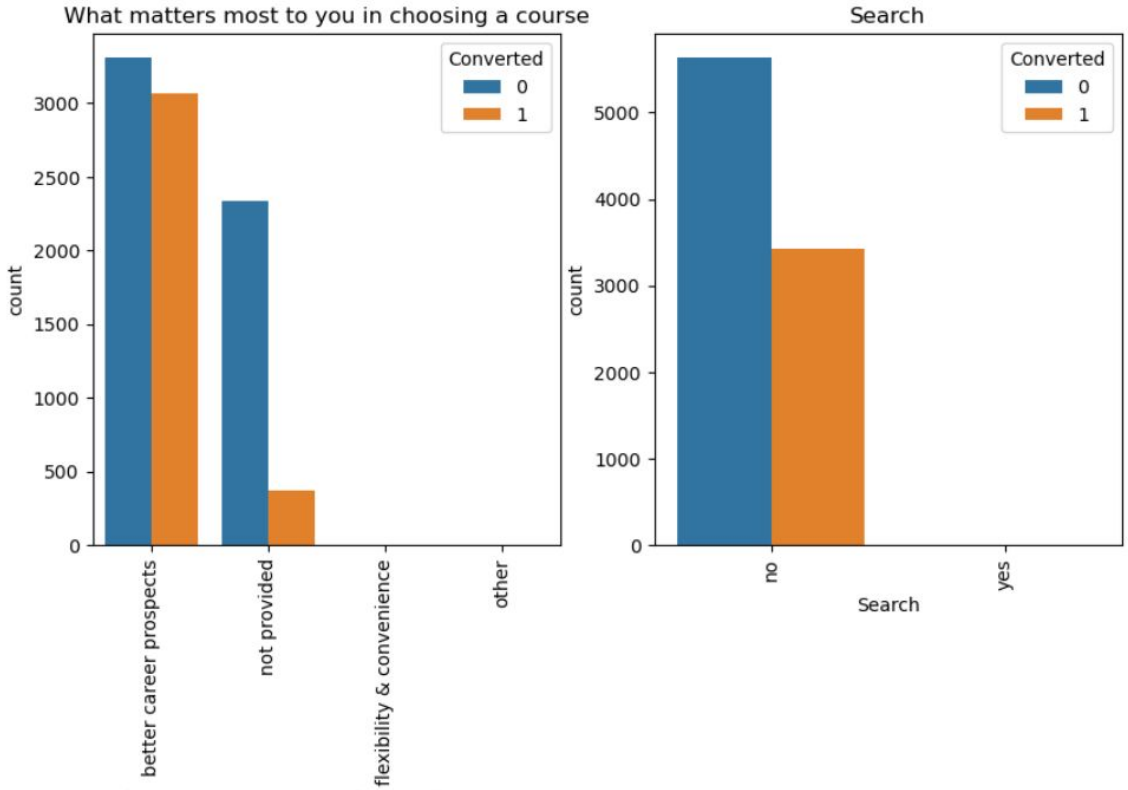
Leads with sms sent and had a phone conversation shows most conversion



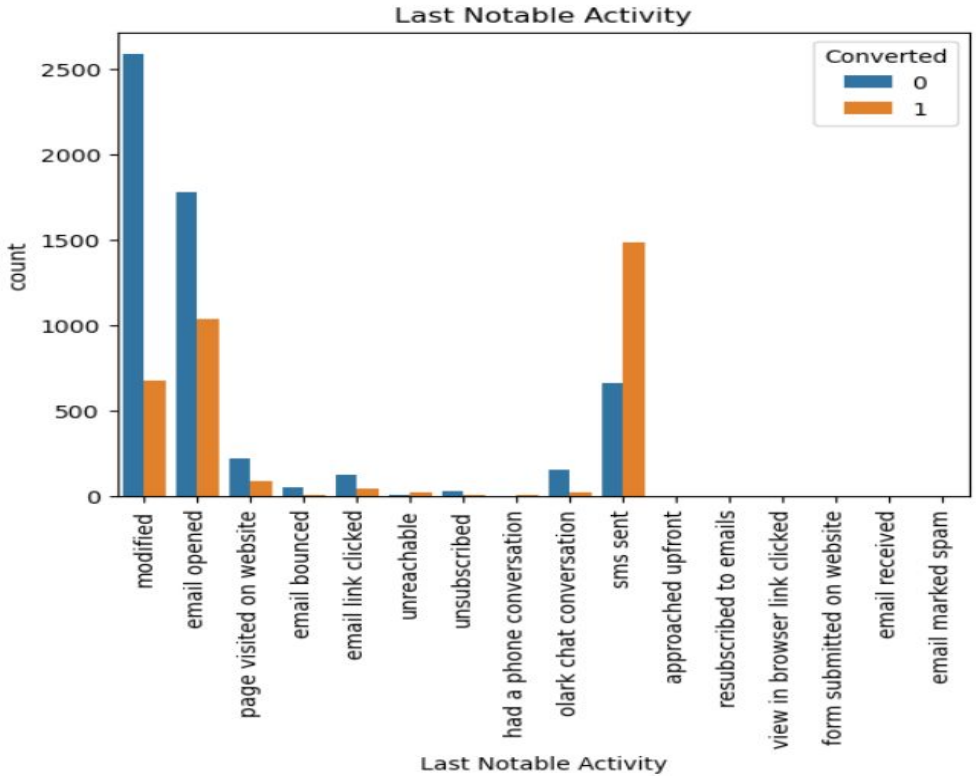
Working Professionals going for the course have high chances of joining it. Unemployed leads are the most in terms of Absolute numbers.



Better career prospects shows fruitful conversion

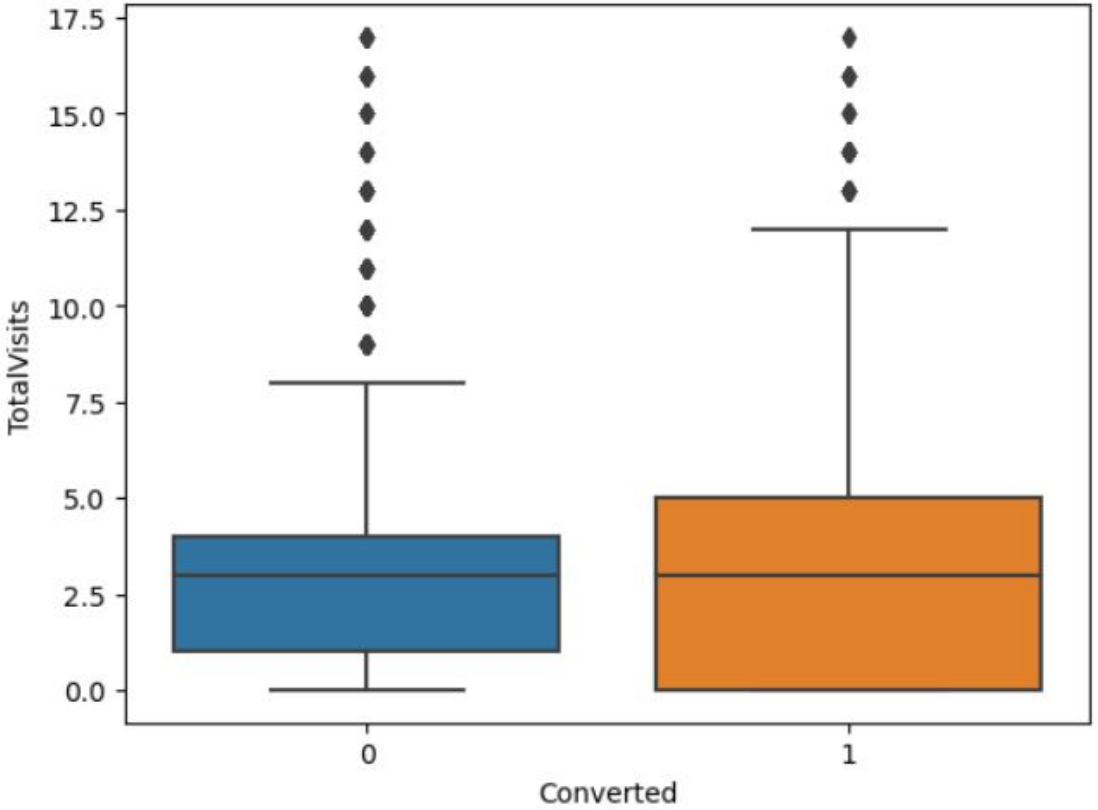


Those leads to whom sms was sent shows most of the conversion.

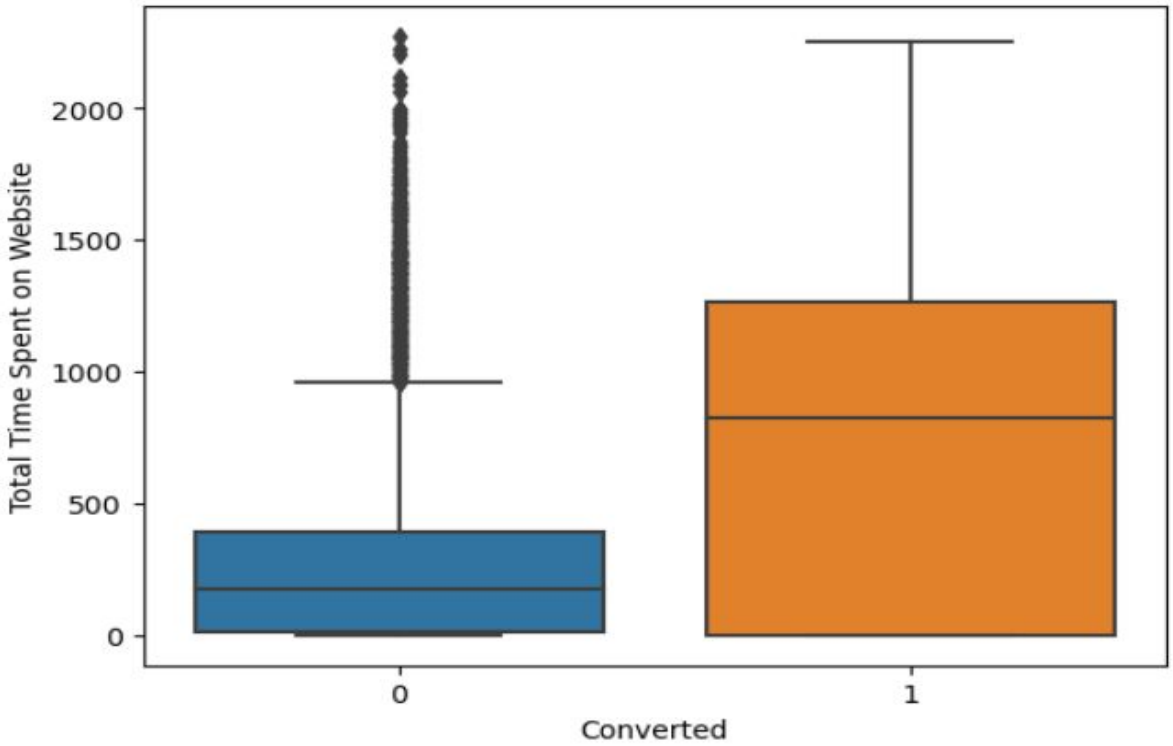


Checking for numerical column

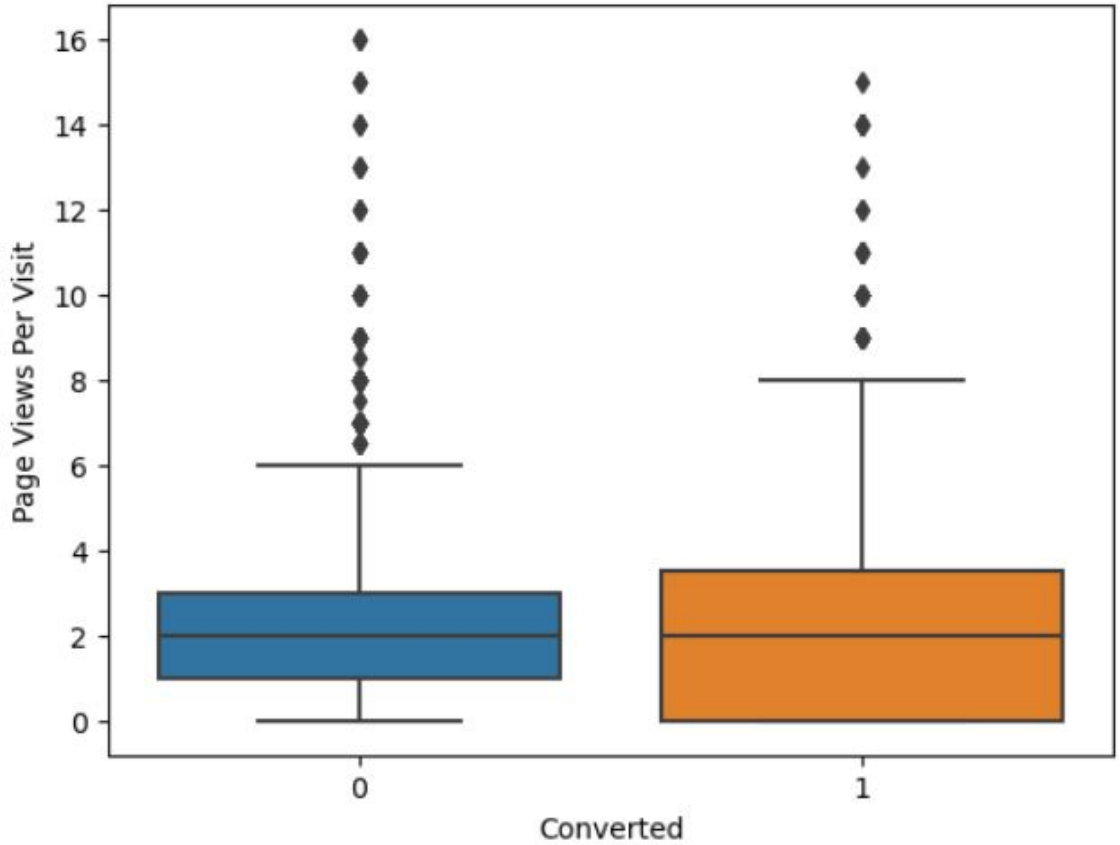
Median for converted and not converted leads are the close. We can't say anything on the basis of Total Visits.



Website should be made more engaging to make leads spend more time. Leads spending more time on the website are more likely to be converted.



Median for converted and unconverted leads is same. We can't say specifically for lead conversion from Page Views Per Visit.



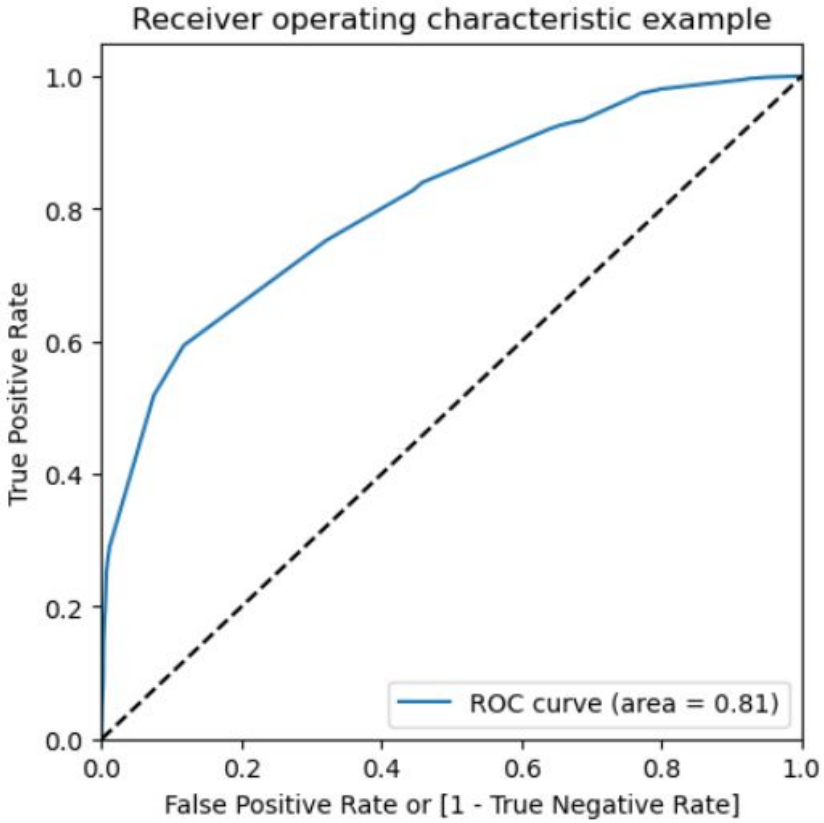
Variables impacting conversion rate:

- Page Views Per Visit
- TotalVisits
- Total Time Spent on Website
- Last notable activity_modified
- Last acivity_olark chat conversation
- Last notable activity_email opened
- Lead Origin_lead add form
- Lead Source_direct traffic
- Lead Source_welingak website
- What is your current occupation_working professional
- Do Not Emai_yes

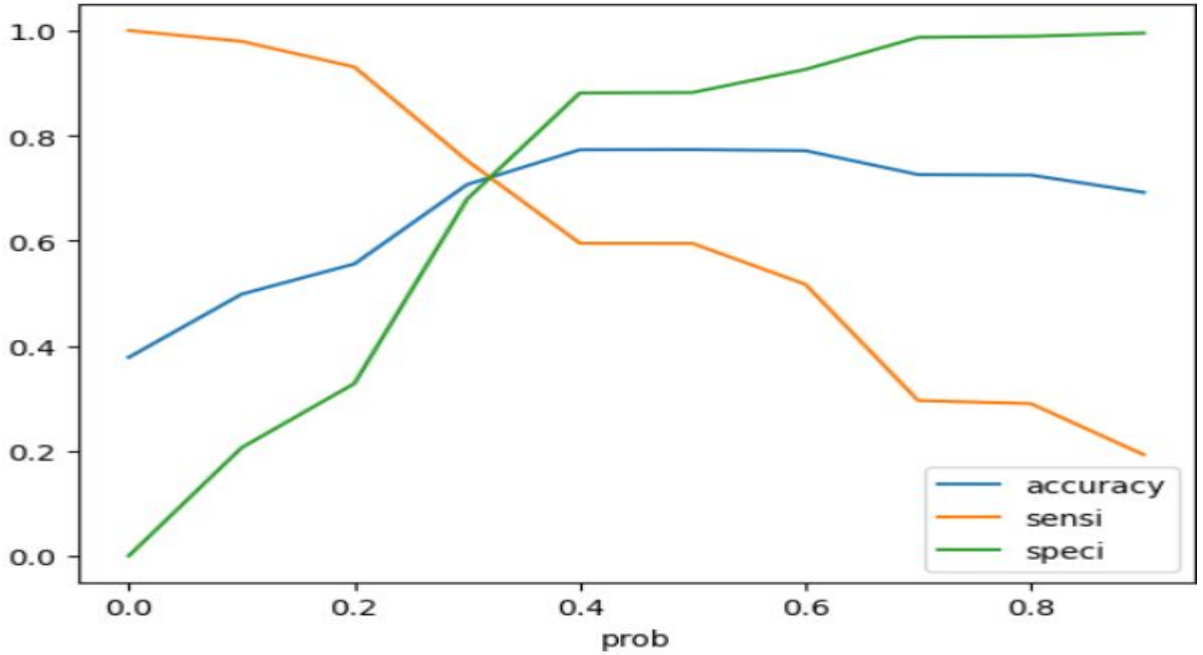
Model Building

- Splitting the data into test and train data set.
- We have chosen 70:30 ratio of test train for performing first step of regression.
- Used RFE for feature selection.
- Running RFE with 15 variables as output.
- Building the model by removing the variable whose p-value is greater than 0.05 and vif is greater than 5.
- Predictions on test data sets.
- Overall accuracy 80.31%

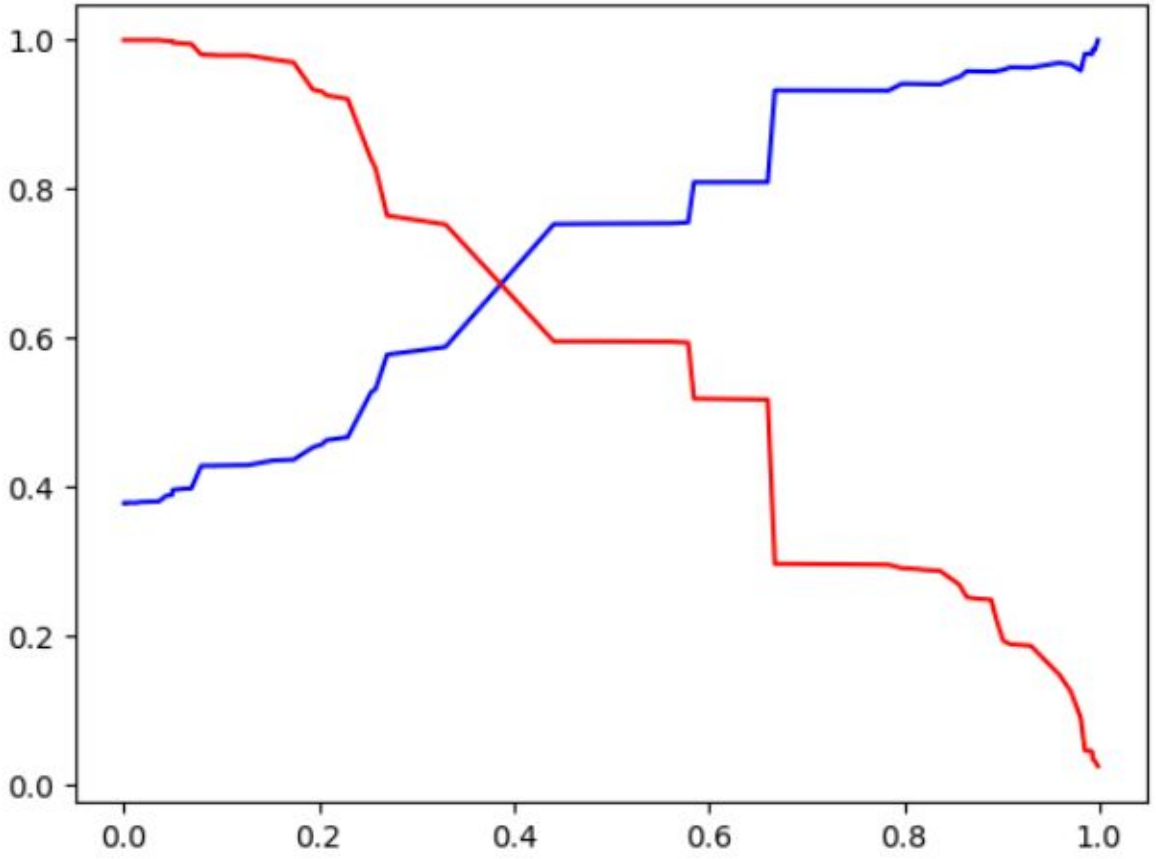
Model evaluation- ROC curve



Finding optimal cut off point: Optimal cut off probability is that probability where we get balanced sensitivity and specificity. From this graph it is visible that optimal cut off point is 0.35.



Precision and Recall on Train Dataset The graph depicts an optimal cutoff of 0.41 based on Precision and Recall



Conclusion:

- The variables that mattered the most in the potential buyers as per this model are as follows:
- 1. Total number of visits on the website.
- 2. Total time spent on the website.
- 3. When the last activity was: •
  - a. sms
  - b. Olark chat conversation
- 4. When the current occupation is as a working professional.
- 5. When the lead search was Google, Direct traffic, welingak website.
- Keeping these parameters in mind, X Education can grow as they have a very high chances to get almost all potential buyers to change their mind and buy their courses hence increasing the conversion ratio.