Lead Scoring Case Study
Using Logistic Regression

Submitted by

Maanik Garg

Priyanka

Litisha Mohapatra

Table of Content

Problem Statement

- *X Education*, an online education company catering to industry professionals, is facing challenges with its lead conversion process. Despite generating a significant number of leads through various channels, the company struggles with a low lead conversion rate of approximately 30%.
- To enhance the efficiency of their lead conversion efforts and achieve a target conversion rate of 80%, they are seeking a solution to identify the most promising leads, referred to as 'Hot Leads.'

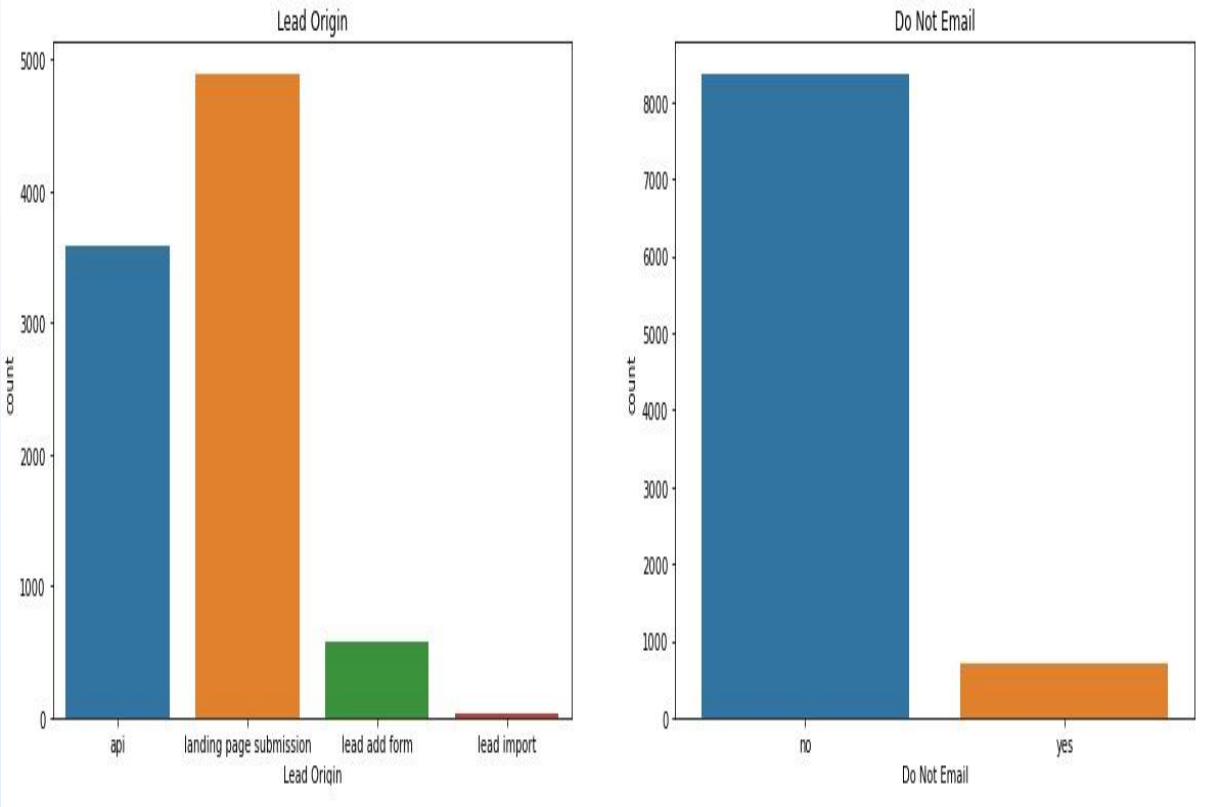
Business Objective

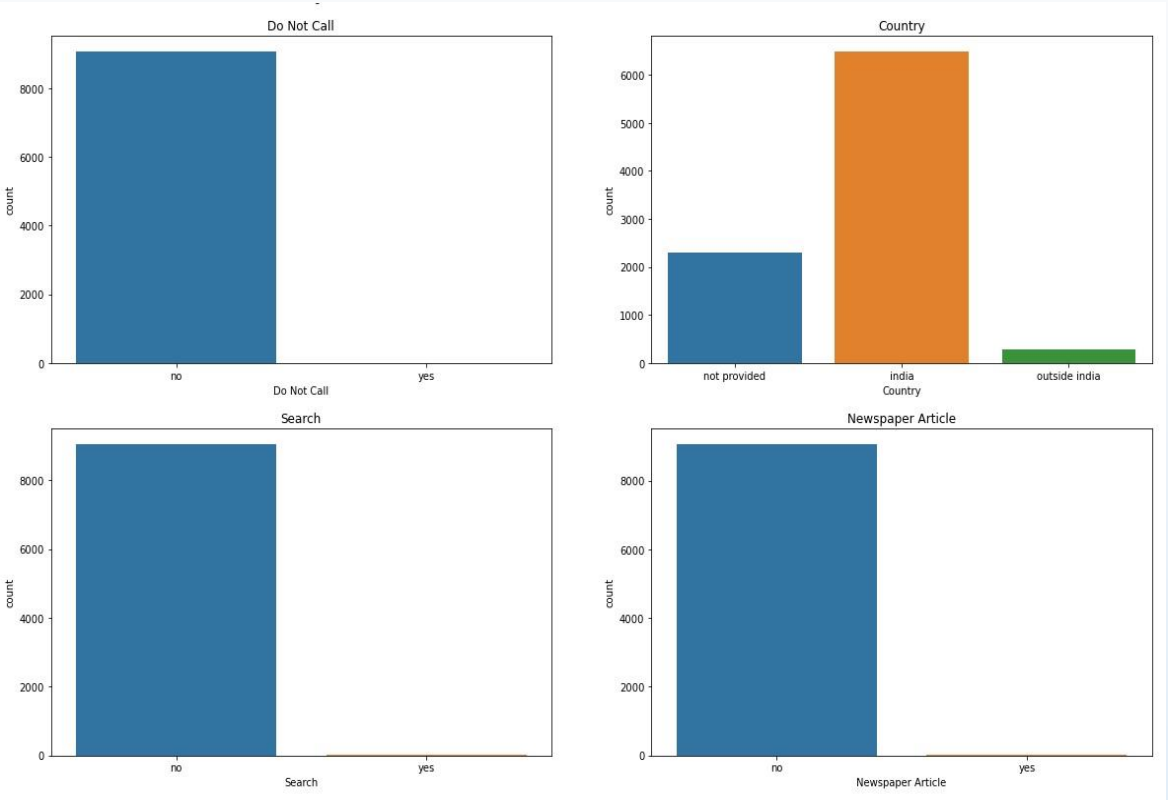
- X Education aims to develop a lead scoring model to prioritize prospects and enhance the lead conversion rate to 80%. The current process involves attracting visitors through marketing and referrals, converting them into leads by collecting contact details via forms, and utilizing sales outreach (calls, emails) to convert leads into customers.

Problem Approach

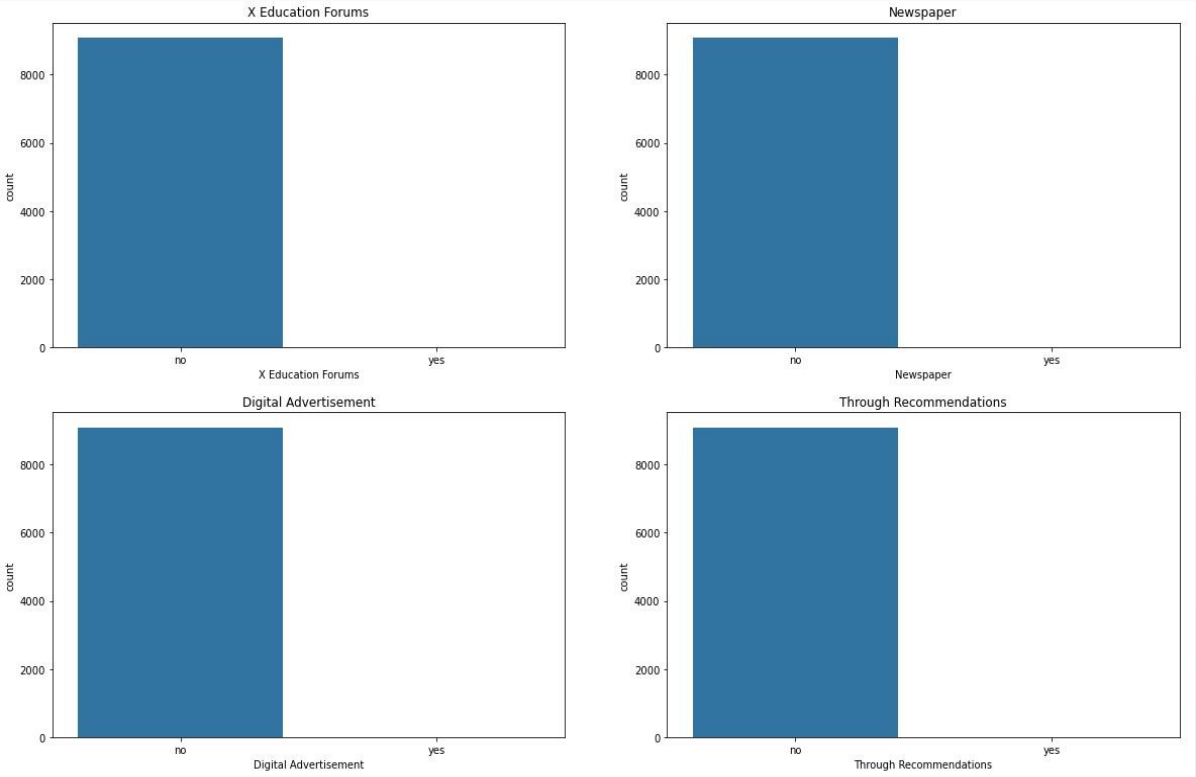
- Importing the data and inspecting the dataframe
- Data Preparation
- EDA
- Dummy Variable Creation
- Test- Train Split
- Feature Scaling
- Correlation
- Model Building (RFE Rsquared VIF and p values)
- Model Evaluation
- Making prediction on test set

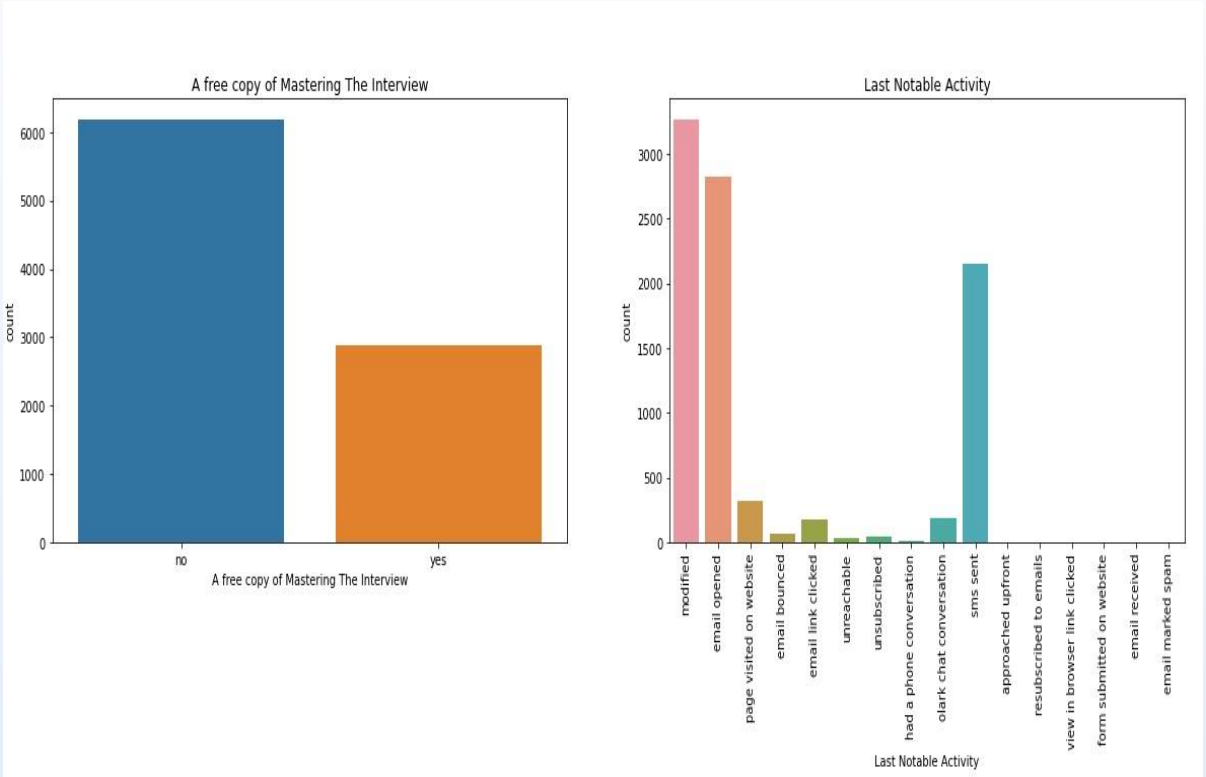
EDA- Univariant Analysis



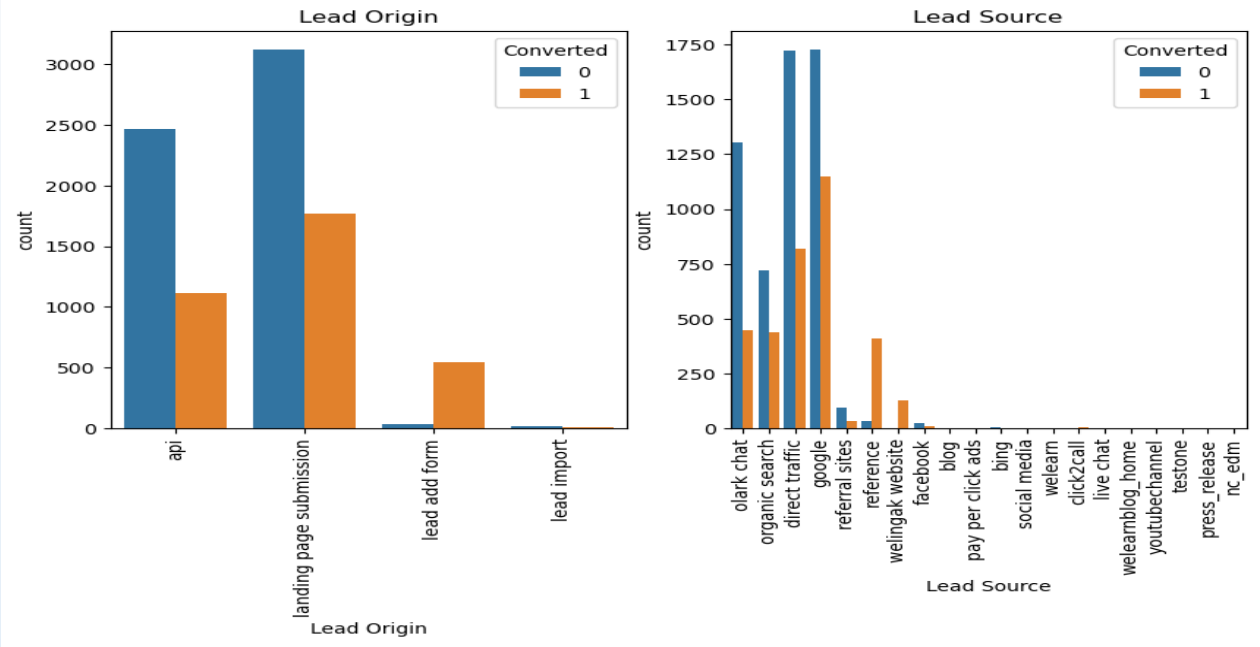


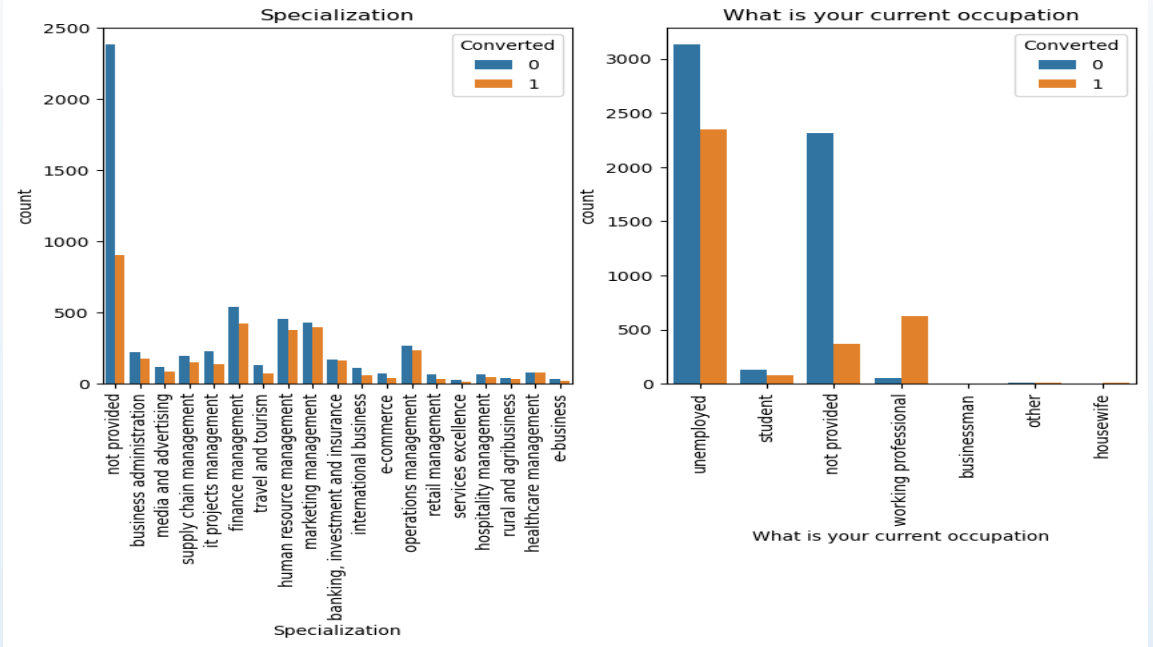
EDA- Univariant Analysis



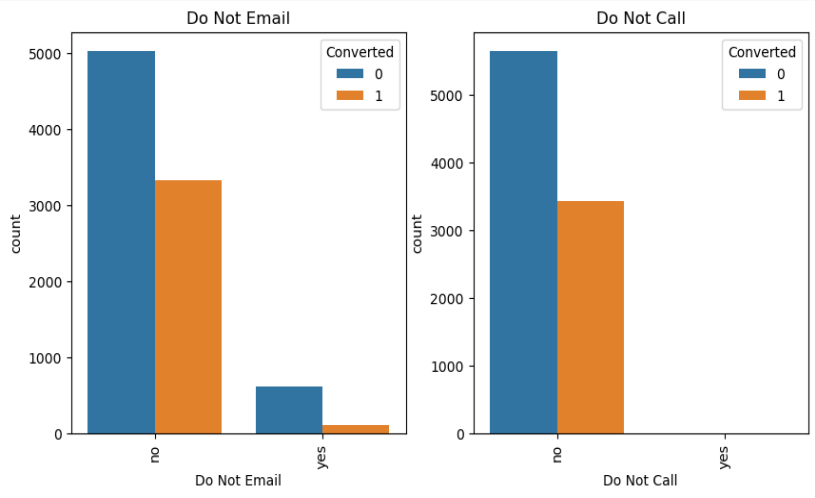


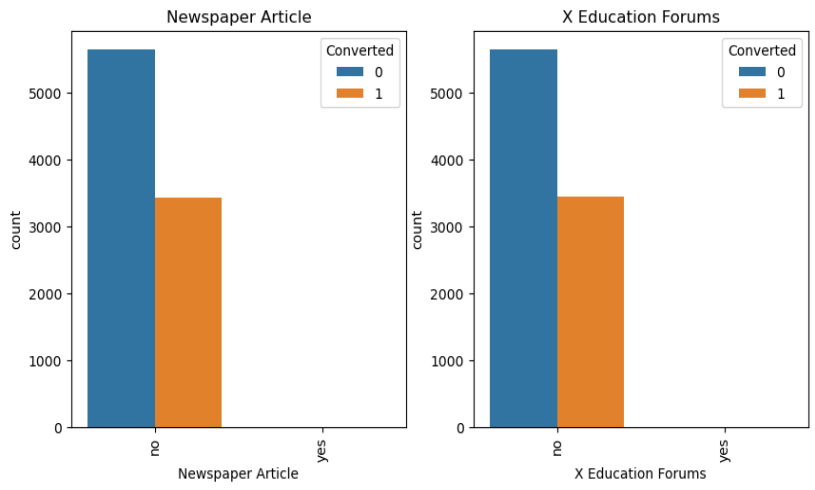
Relating all the categorical variables to Target Variable ‘Converted’

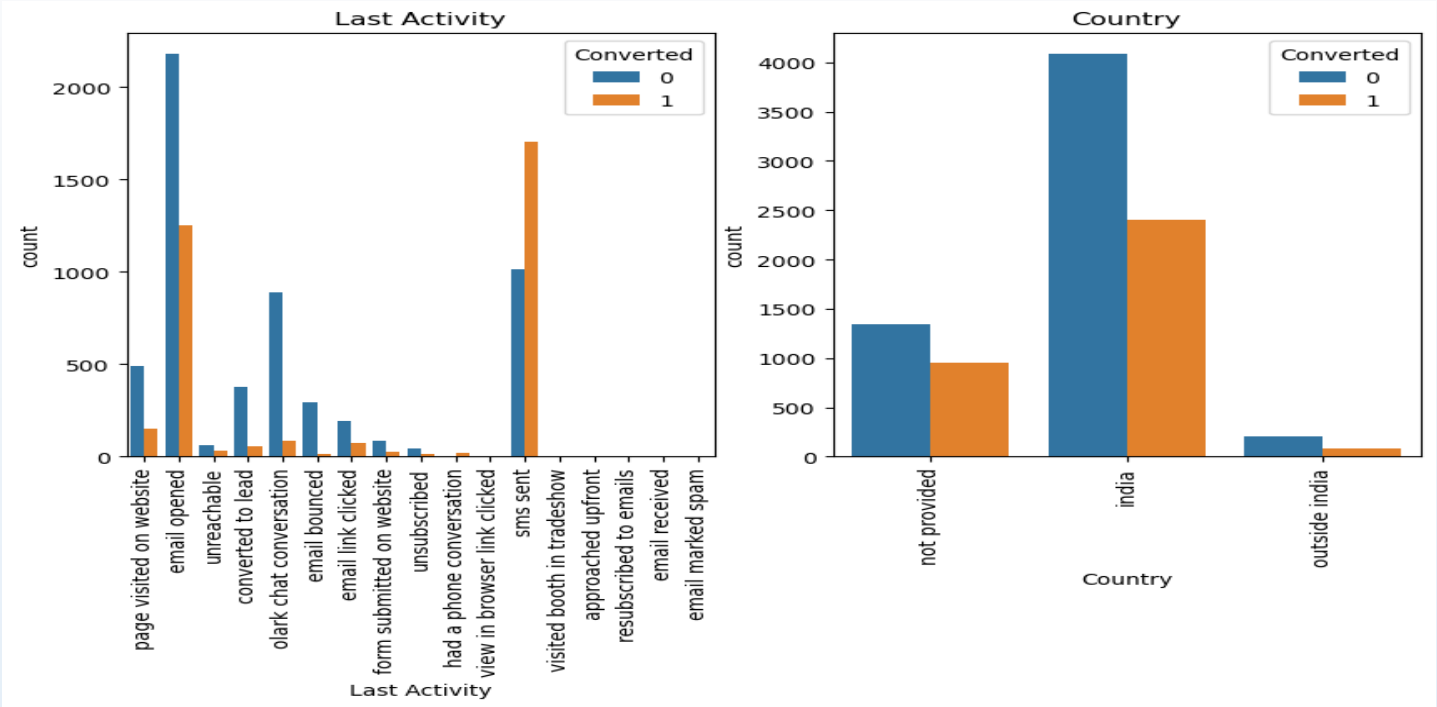




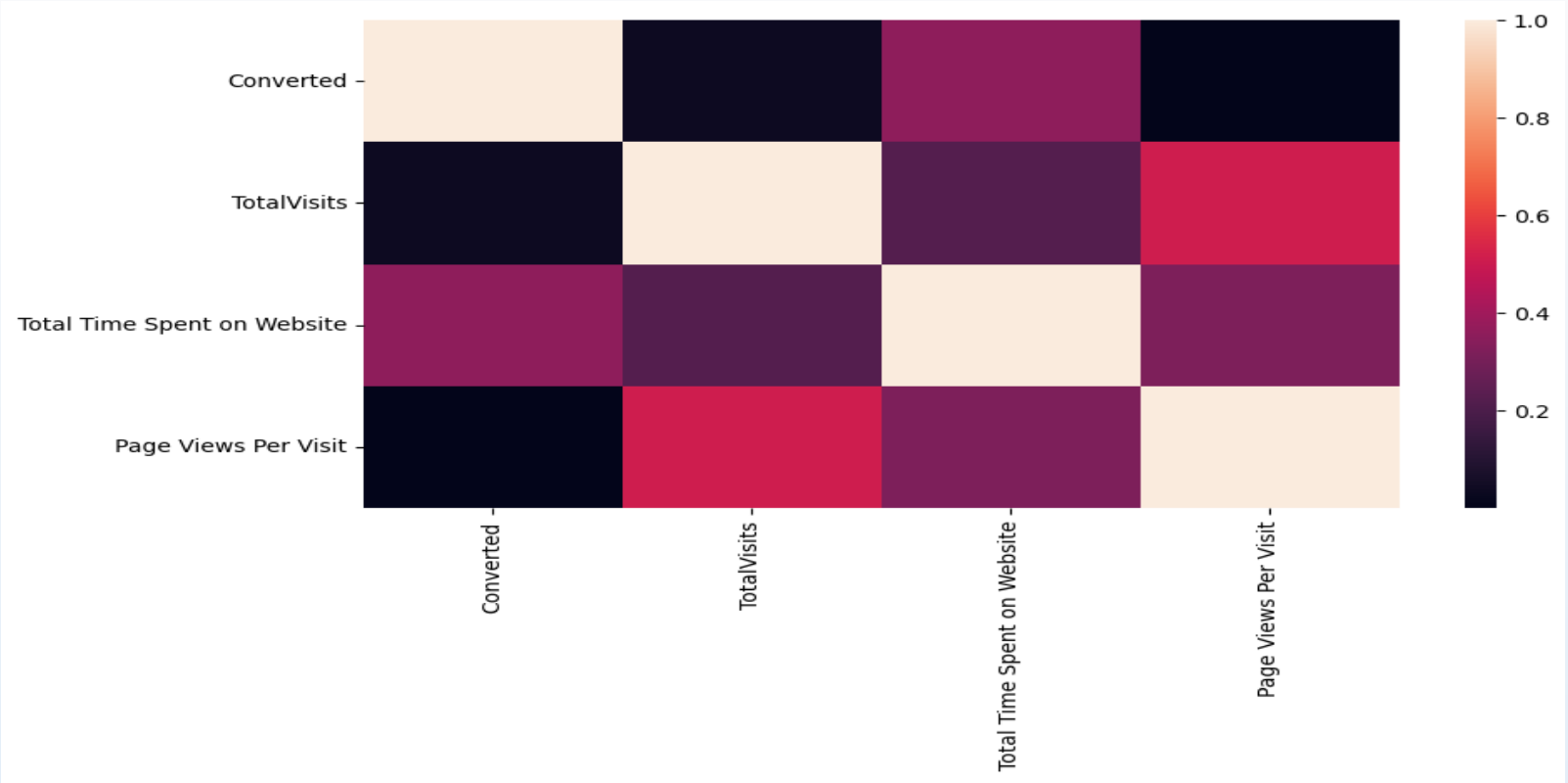
Cont..


Do Not Email
Converted
0
1
count
Do Not Email
no
yes
Do Not Call
Converted
0
1
count
Do Not Call
no
yes


Newspaper Article
Converted
0
1
count
Newspaper Article
no
yes
X Education Forums
Converted
0
1
count
X Education Forums
no
yes

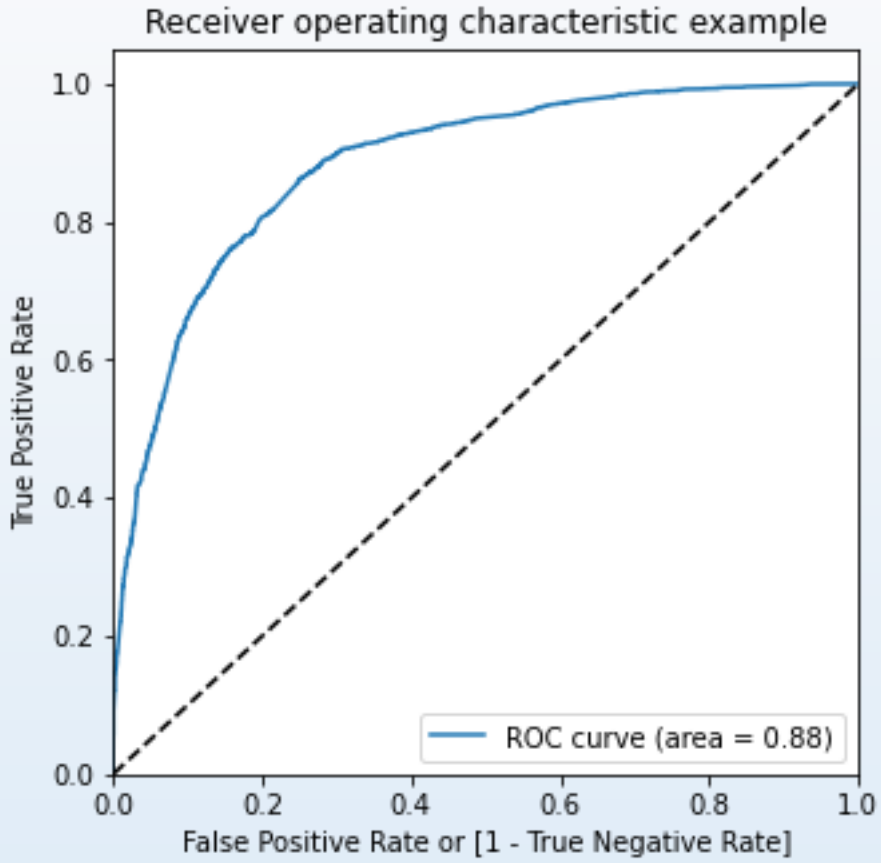

Last Activity
Converted
0
1
count
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity
Country
Converted
0
1
count
not provided
india
outside india
Country

Correlation



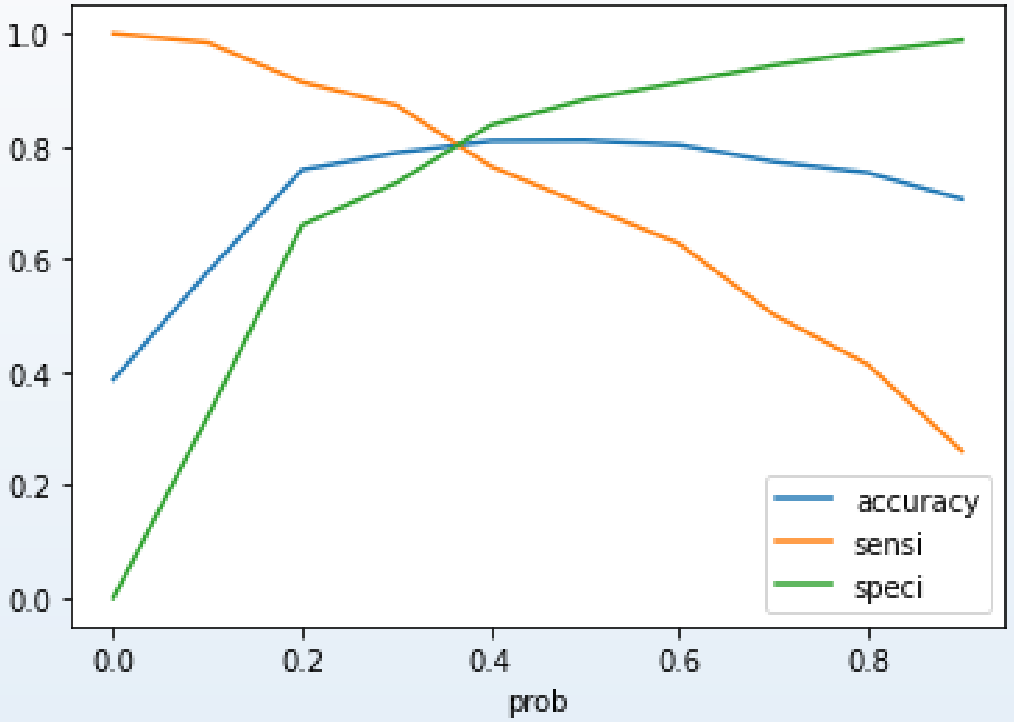
there are many elements that have very little data and so will be of less relevance to our analysis.

Model Evaluation - ROC CURVE



The area under ROC curve is 0.87 which is a very good value.

Model Evaluation – On Train DataSet

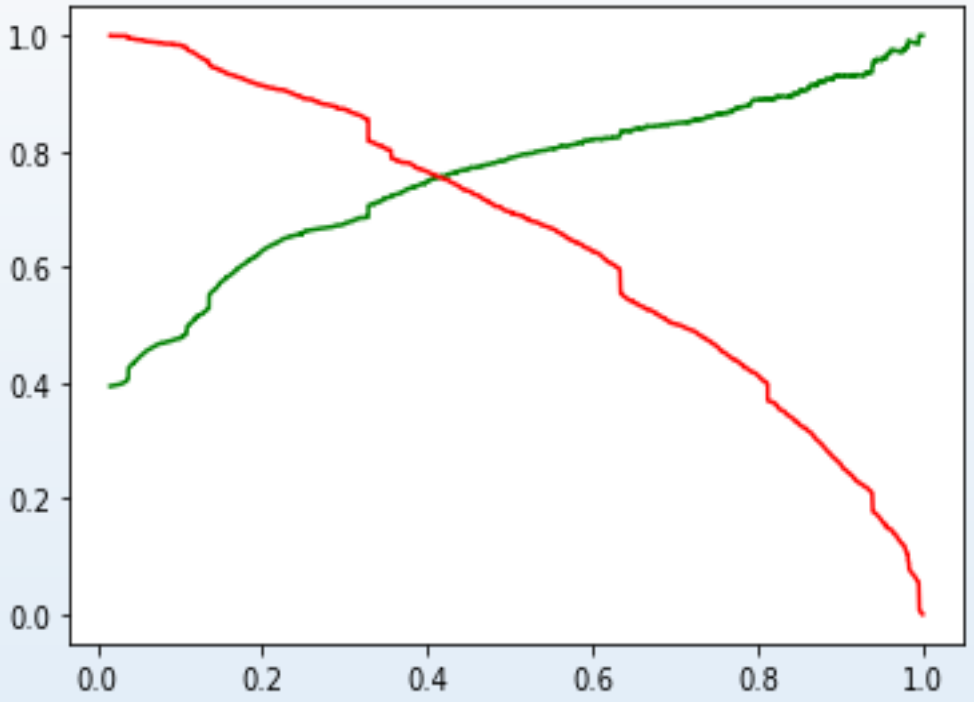


Confusion matrix
[3126,769],
[481, 1975]

Accuracy	80%
Sensitivity	80%
Specificity	80%

From the graph it is visible that the optimal cut off is at 0.35.

Precision and Recall on Train Data



Precision	75%
Recall	75%

Model Evaluation – On Test DataSet

Confusion Matrix

[1406, 338],
[181, 798]

Accuracy	80%
Sensitivity	81%
Specificity	80%

Conclusion

Key influencing factors for potential buyers, in descending order:

1. Time spent on the website
2. Total number of visits
3. Lead sources:
    - Google
    - Direct traffic
    - Organic search
    - Welingak website
4. Last activity:
    - SMS
    - Olark chat

Conclusion

5. Lead origin as "Lead Add Form"
6. Current occupation as a working professional

To optimize success:

Focus efforts on the above factors, prioritizing the listed lead sources.
Allocate additional resources and attention to leads from API and Landing Page submissions, which demonstrate higher conversion rates.