# Lead Score Case Study

Bhawani Singh
DS 065

# Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, 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.

# Case Study Objectives

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

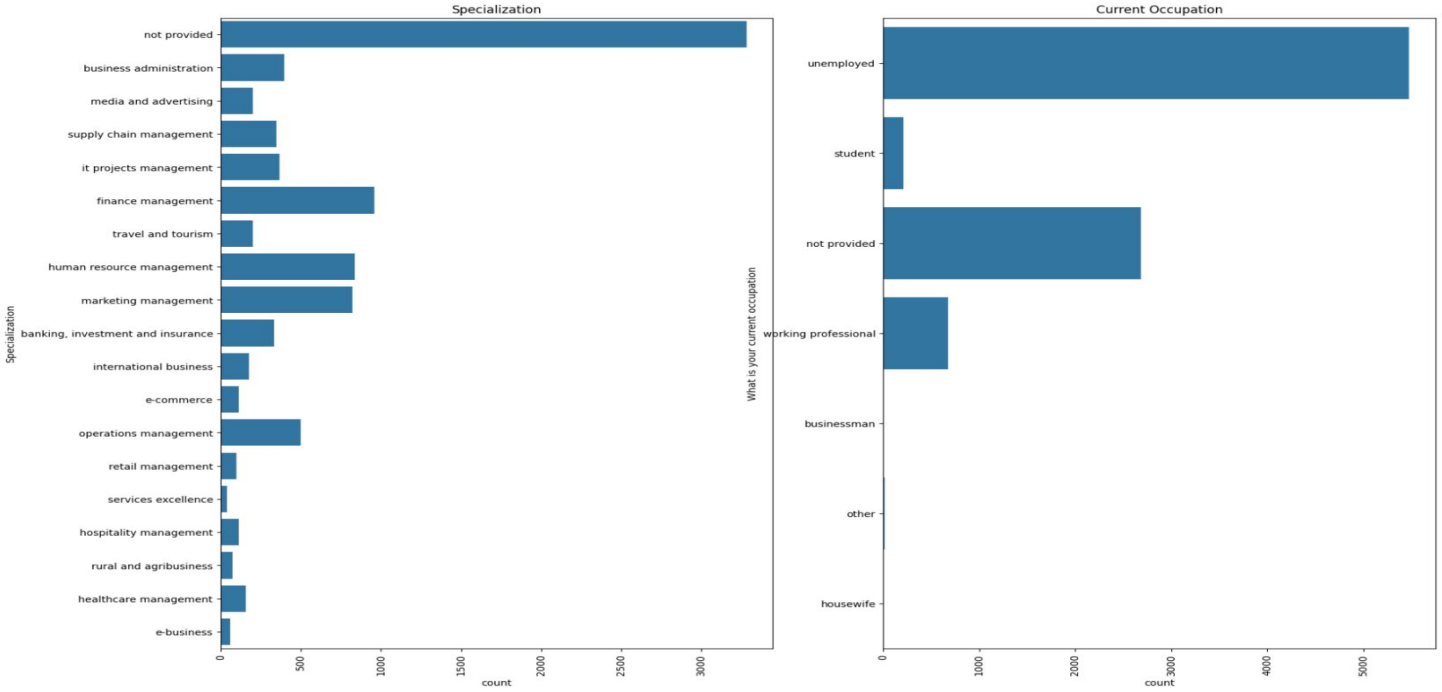
# Strategy

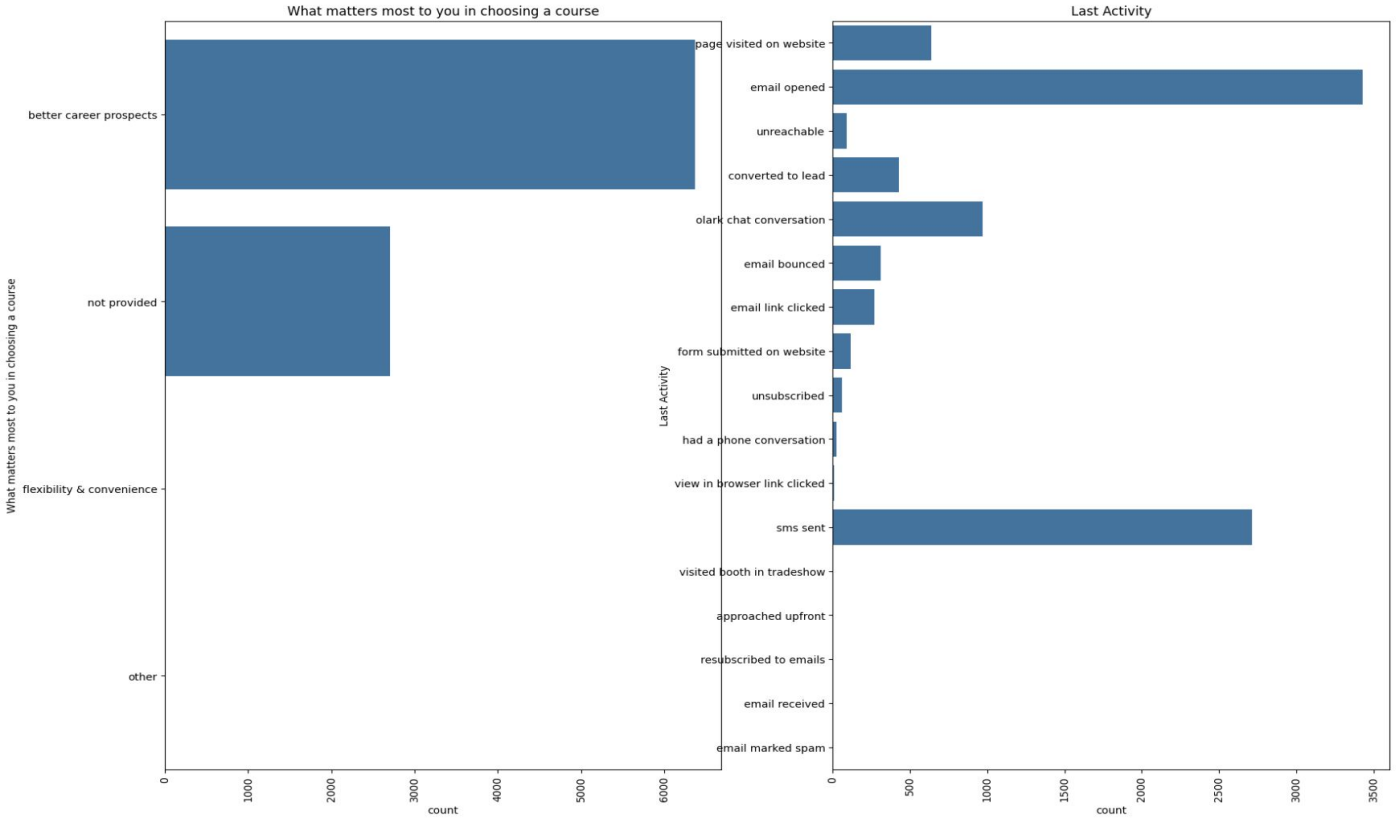
- Source the data for analysis  Clean and prepare the data  Exploratory Data Analysis.
- Feature Scaling  Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

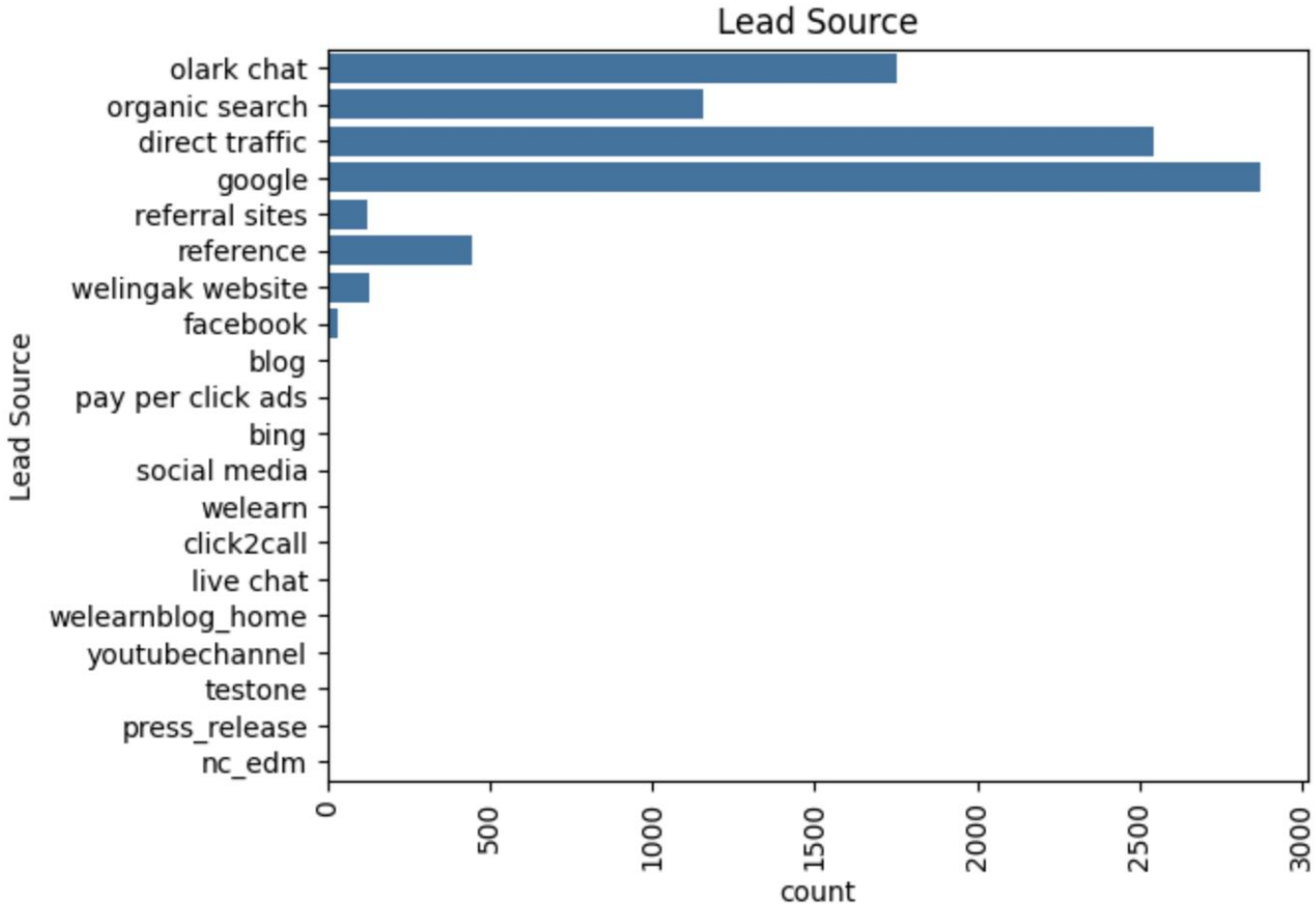
# Data Manipulation and cleanup

- 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**”, “**Digital Advertisement**” etc.
- Dropping the columns having more than **35%** as missing value such as ‘**How did you hear about X Education’** and ‘**Lead Profile**’.

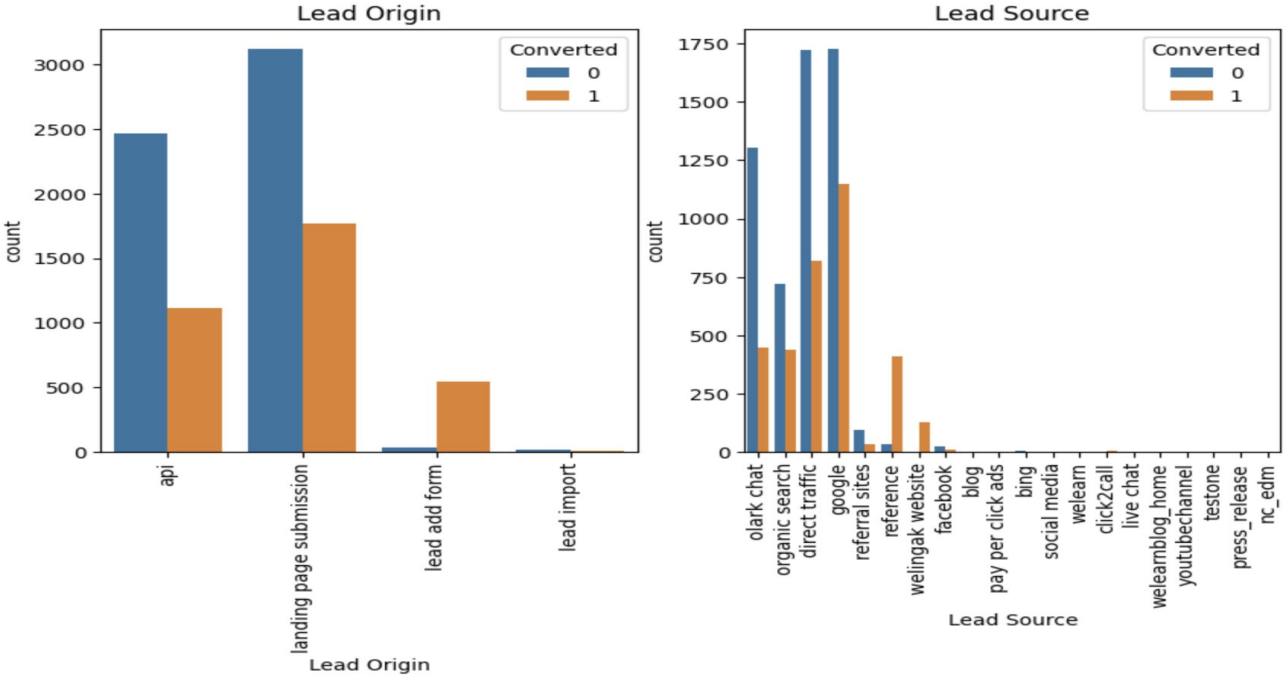
# Exploratory Data Analysis

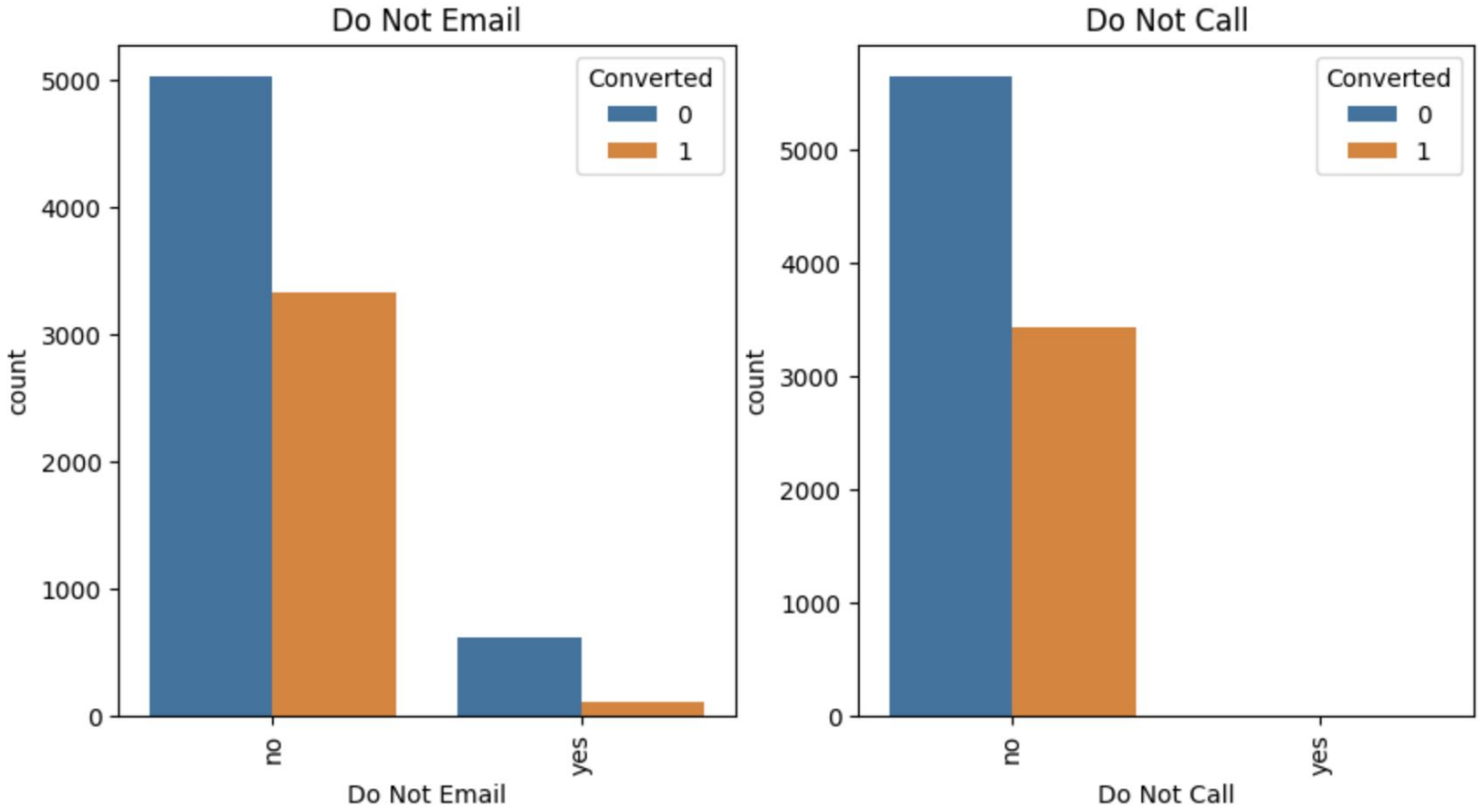


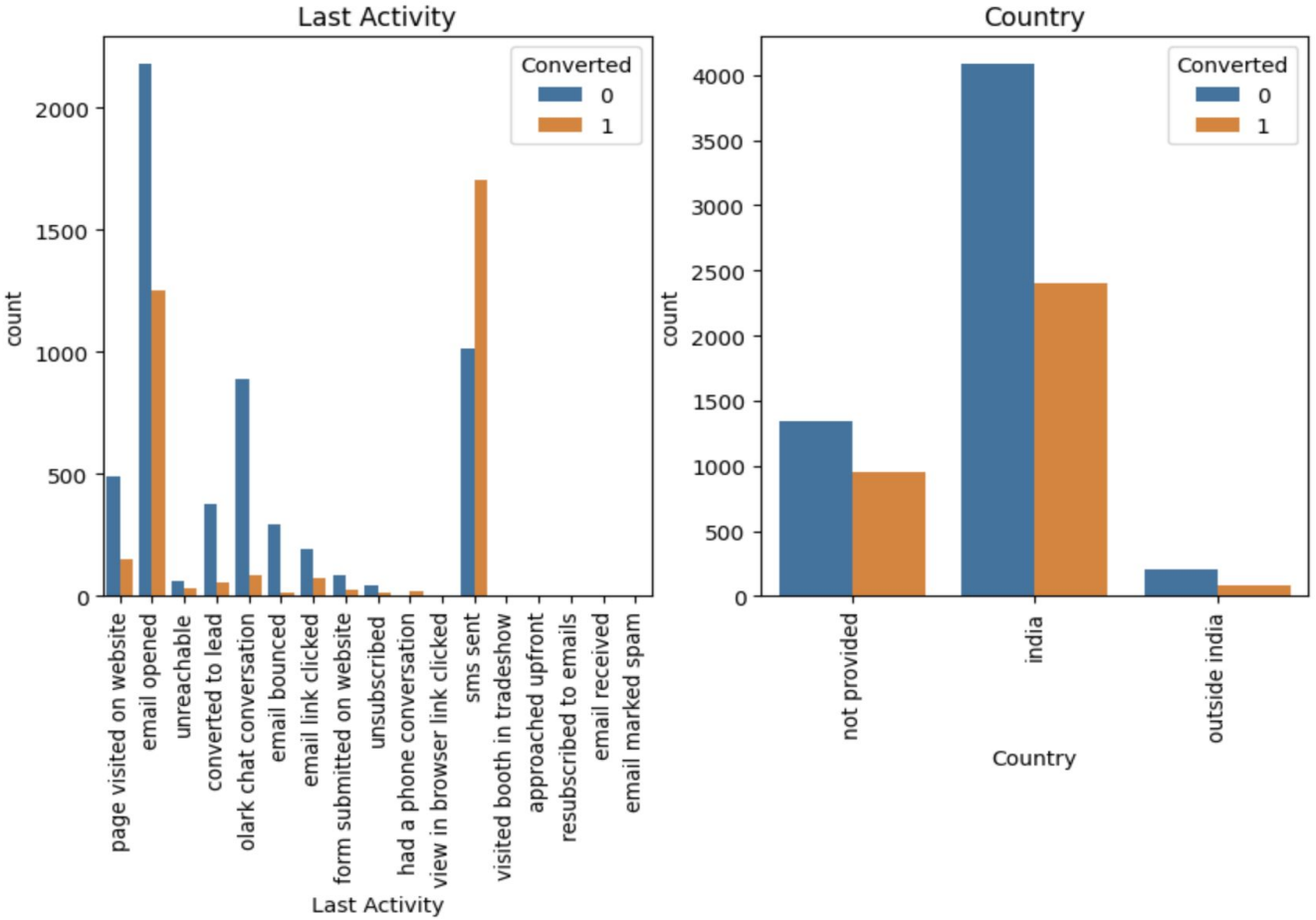

What matters most to you in choosing a course
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course
0
1000
2000
3000
4000
5000
6000
count
Last Activity
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
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
0
500
1000
1500
2000
2500
3000
3500
count


Lead Source
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source
0
500
1000
1500
2000
2500
3000
count

# Categorical Variable Relation




Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Email
Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Call


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
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
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

# Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: **9074**
- Total Columns for Analysis: **81**

Out[46]:

| | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | Lead Origin_landing page submission | Lead Origin_lead add form | Lead Origin_lead import | Specialization_business administration | Specialization_e busines |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0.0 | 0 | 0.00 | False | False | False | False | Fals |
| **1** | 0 | 5.0 | 674 | 2.50 | False | False | False | False | Fals |
| **2** | 1 | 2.0 | 1532 | 2.00 | True | False | False | True | Fals |
| **3** | 0 | 1.0 | 305 | 1.00 | True | False | False | False | Fals |
| **4** | 1 | 2.0 | 1428 | 1.00 | True | False | False | False | Fals |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | .. |
| **9235** | 1 | 8.0 | 1845 | 2.67 | True | False | False | False | Fals |
| **9236** | 0 | 2.0 | 238 | 2.00 | True | False | False | False | Fals |
| **9237** | 0 | 2.0 | 199 | 2.00 | True | False | False | True | Fals |
| **9238** | 1 | 3.0 | 499 | 3.00 | True | False | False | False | Fals |
| **9239** | 1 | 6.0 | 1279 | 3.00 | True | False | False | False | Fals |

9074 rows × 81 columns

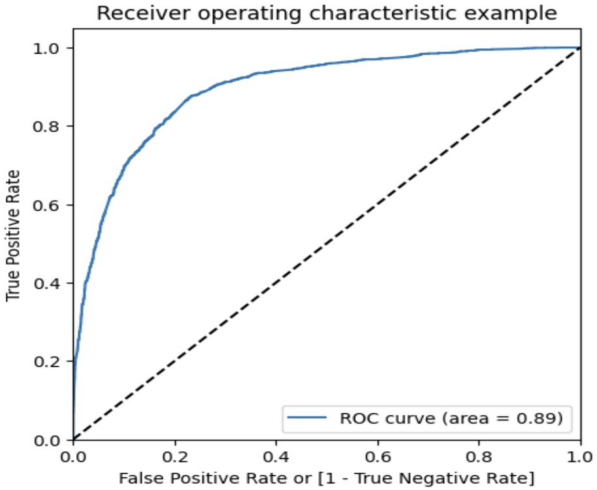
# Model Building

- 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
- Predictions on test data set
- Overall accuracy **81%**

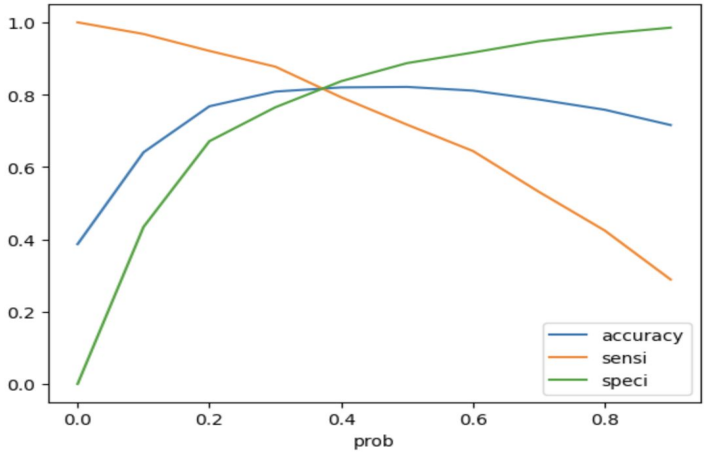
[139]:
```
# Check the overall accuracy
metrics.accuracy_score(y_pred_final['Converted'], y_pred_final.final_predicted)
```

[139]:
```
0.8185824458318032
```

# ROC




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




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

# Conclusion

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
  - The total time spent on the Website.
  - Total number of visits.
  - When the lead source was:
    - Google, Direct traffic, Organic search, Welingak website
  - When the last activity was:
    - SMS, Olark chat conversation
  - When the lead origin is Lead add format.
  - When their current occupation is as a working professional.
- 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.

# Thanks!