Lead Score
Logistic Regression

- Mohan T
- Unnikrishnan
- Vanshika

Business Problem & Approach

Business Problem:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses

X Education offers online courses for industry professionals. Despite receiving numerous leads, their conversion rate is quite low. (30%).

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’

Approach:

Lead Data: files has been provided with leads information containing approximately 9,000 data points

Perform Data Cleaning and EDA

Develop a model to identify leads with a conversion probability greater than 80%

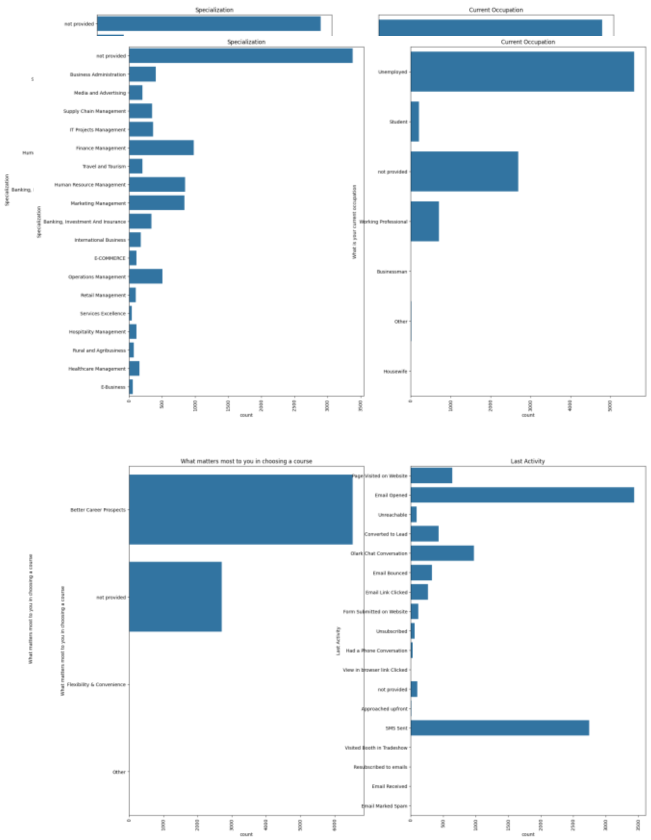
Detailed Approach

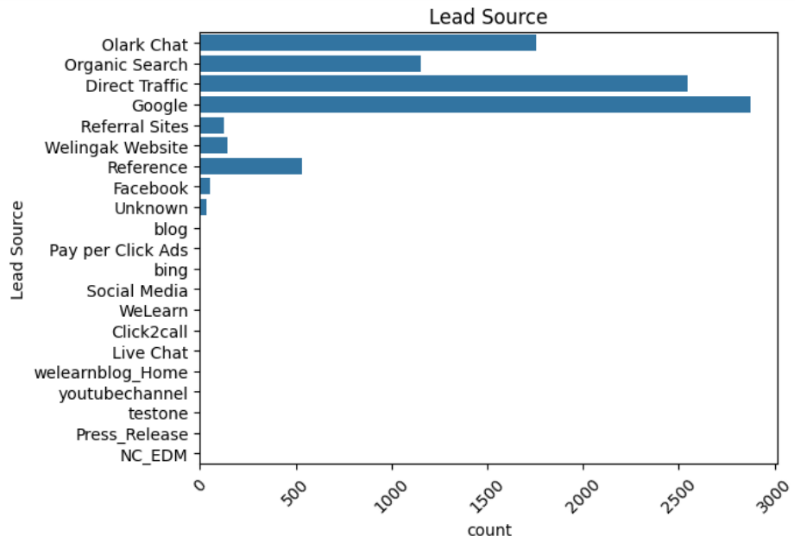
- Import data & Perform data checks
- EDA
- Dummy Variable creation
- Test- Train Split
- Feature scaling
- Model building
  - Features selection – RFE
  - Checking correlation between variables i.e., VIF and p-values
- Model Evaluation
- Making predictions

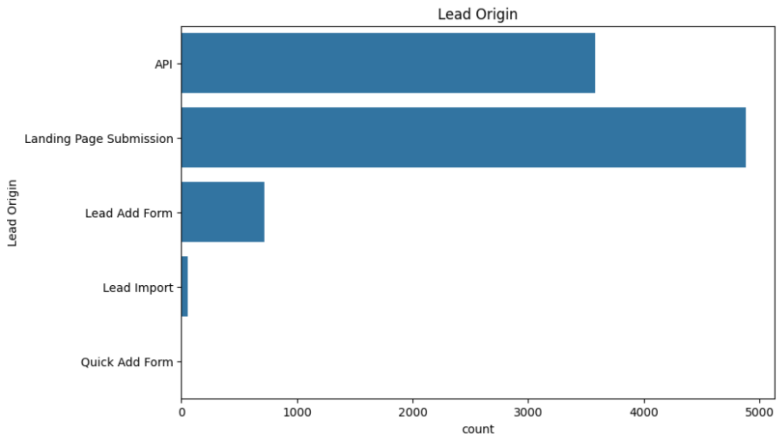
Clean and Transform Data

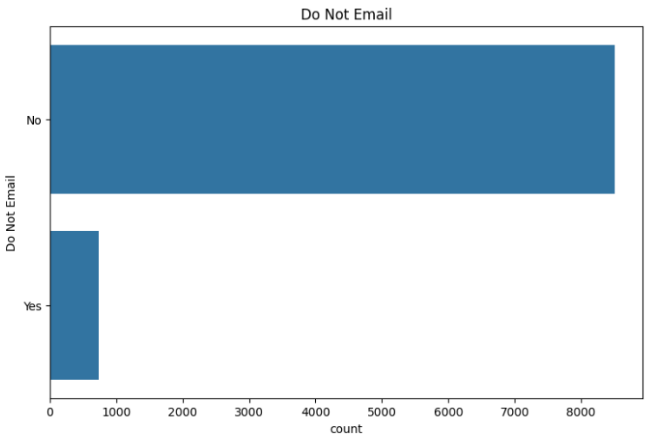
- Started with loading data and analyzing data for null/missing values. Columns that have more than 35% of missing values, excluded from analysis and model building.
- Columns with unique values can be removed, as they do not impact the overall solution.
- Data with “select” has been replaced with “Not provided”.
- Numerical variable missing values are replaced with 0. I.e., Total Time Spent on Website, Page Views Per Visit and total visits.
- Since the majority of leads are from India, with only a few from other countries, the data has been grouped into two categories: India and Others

EDA

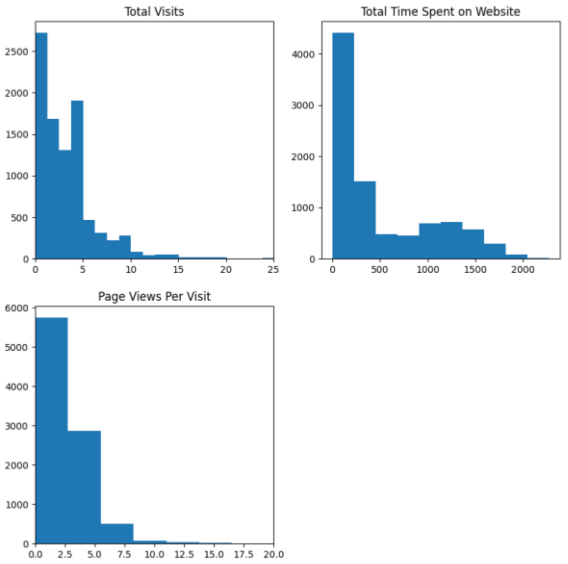

Specialization
not provided
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization
0
500
1000
1500
2000
2500
3000
3500
count
Current Occupation
Unemployed
Student
not provided
Working Professional
Businessman
Other
Housewife
What is your current occupation
0
1000
2000
3000
4000
5000
count
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
not provided
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity
0
500
1000
1500
2000
2500
3000
3500
count

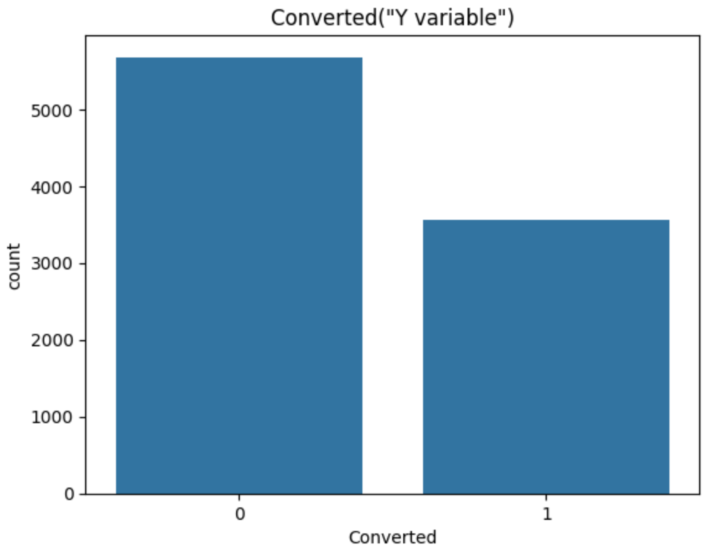

Lead Source
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
Unknown
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source
0
500
1000
1500
2000
2500
3000
count

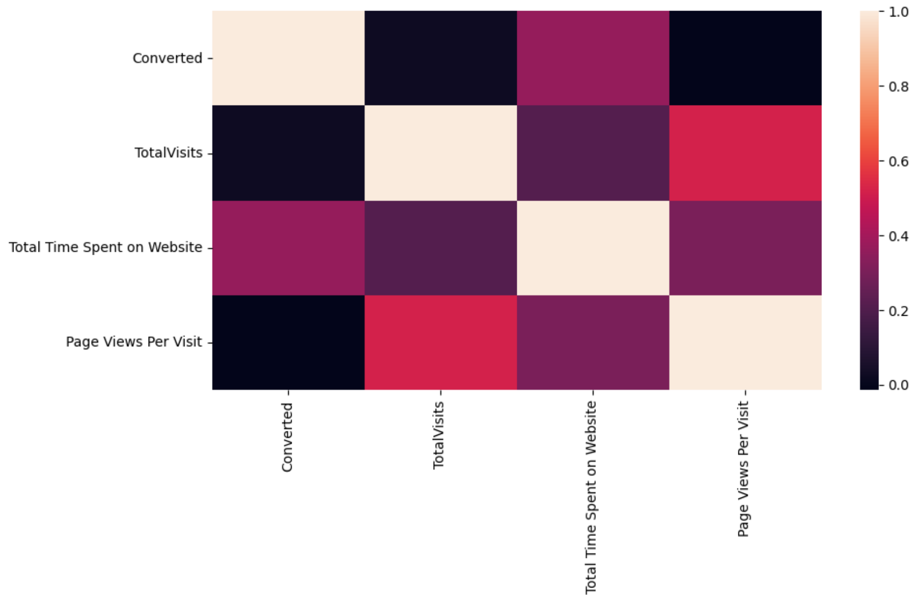

Lead Origin
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin
0
1000
2000
3000
4000
5000
count


Do Not Email
No
Yes
Do Not Email
0
1000
2000
3000
4000
5000
6000
7000
8000
count

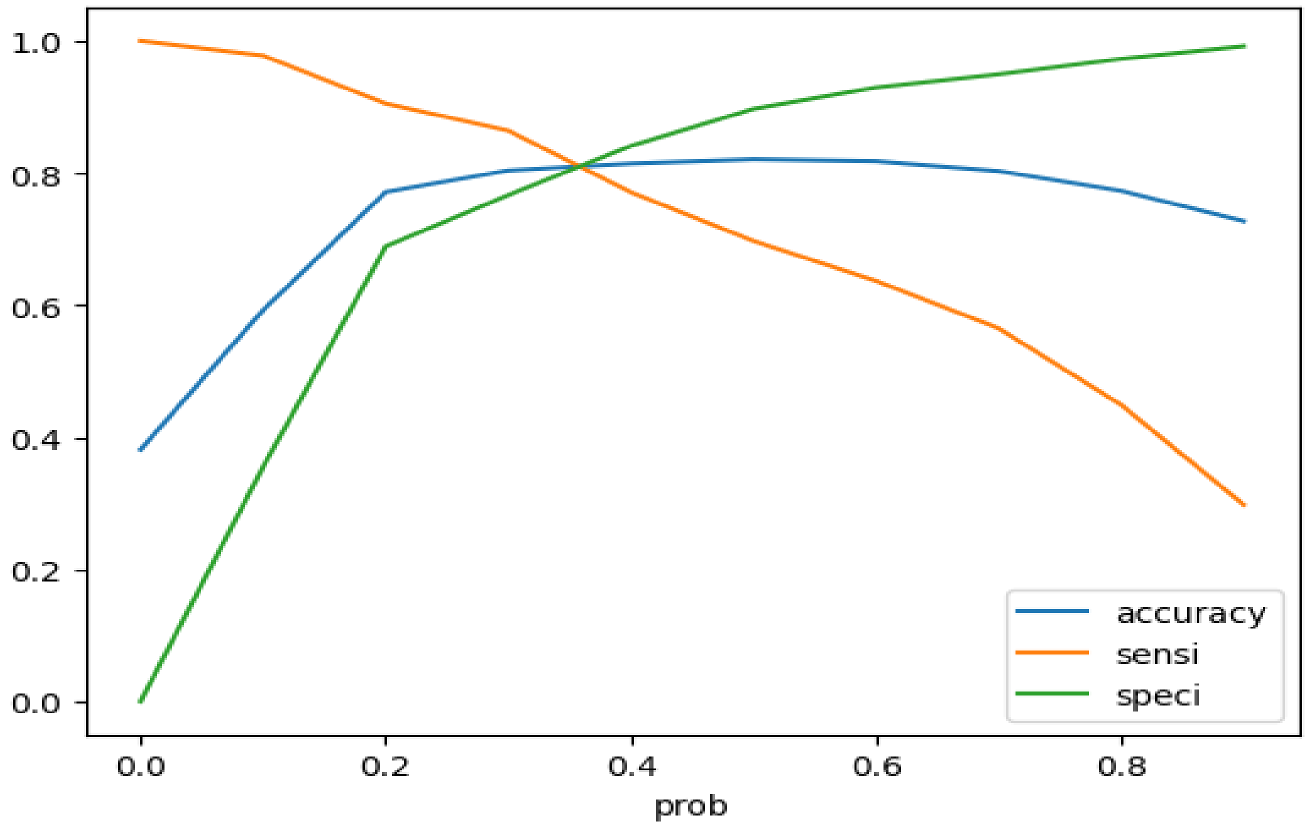
EDA


Total Visits
0
500
1000
1500
2000
2500
0
5
10
15
20
25
Total Time Spent on Website
0
1000
2000
3000
4000
0
500
1000
1500
2000
Page Views Per Visit
0
1000
2000
3000
4000
5000
6000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Converted("Y variable")
count
0
1000
2000
3000
4000
5000
0
1
Converted


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

ROC Curve



Finding Optimal Cut off Point:

- Probability where we get balanced sensitivity and specificity
- Optima cut off point is .35

Observations

Train test Data

Accuracy: 82%

Sensitivity: 69%

Specificity: 89%

Test test Data

Accuracy: 82%

Sensitivity: 75%

Specificity: 86%

- Feature list
  - Do Not Email
  - Total Time Spent on Website
  - Lead Origin_Lead Add Form
  - Lead Source_Welingak Website
  - Last Activity
    - Had a Phone Conversation
    - SMS Sent