Lead Score Case Study

By
Mr. Dron Amrit
Ms. Mridusmita Goswami
Mr. Shikhar Saxena

Table of Contents

Problem statement of the Case Study

- 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 gets a lot of leads, its lead conversion rate is very poor. If 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 Goal of the Case Study

- X Education wants to know the most promising leads.
- To achieve the business goal a model needs to be built which will give Hot Leads.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution methodology: Steps followed

- Analysis of the data first by Data Loading and Data Understanding
- Data Quality Check and Missing Values
  a. Removing the columns where missing values is very high.
  b. Identify and remove columns with high missing percentage (>45%)
  c. Get the column with null values more than 45%
  d. Dropped all columns from Data frame for which missing value percentage are more than 45%.
  e. Imputation of the missing values for less than 15% of total values.

Steps followed

- Performing the Univariate Analysis.
- Performing the Bi-variate Analysis.
- Finding the correlation among the dataset.
- Feature Scaling

Steps followed

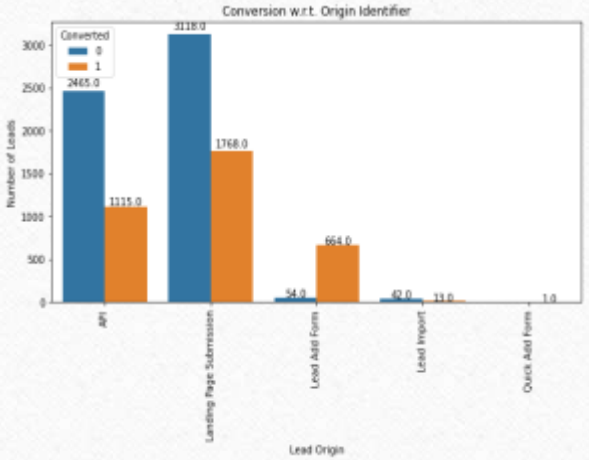
- Model building
- Feature Selection Using RFE
- Checking VIFs
- Plotting the ROC Curve
- Evaluating the model by Precision and Recall

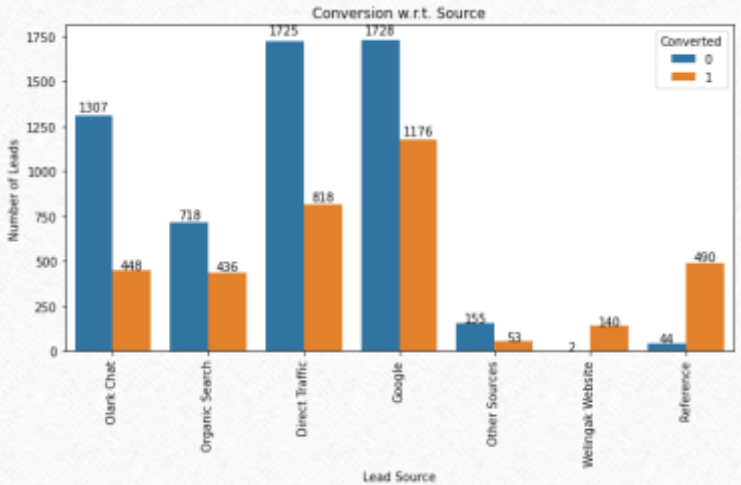
DATA MANIPULATION

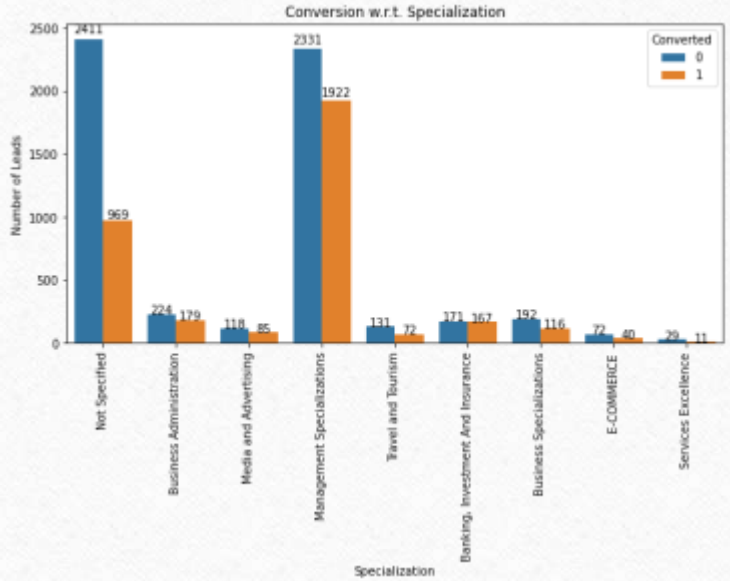
Insights :

- There are 9240 rows and 37 columns in the data set.
- Many of the columns have null values.
- Prospect ID and Lead Number both are unique values. Thus, they are redundant and one of them may be dropped. As Prospect ID is alphanumeric and also a larger sequence, hence it was dropped.
- 40% as a cutoff for missing values is considered

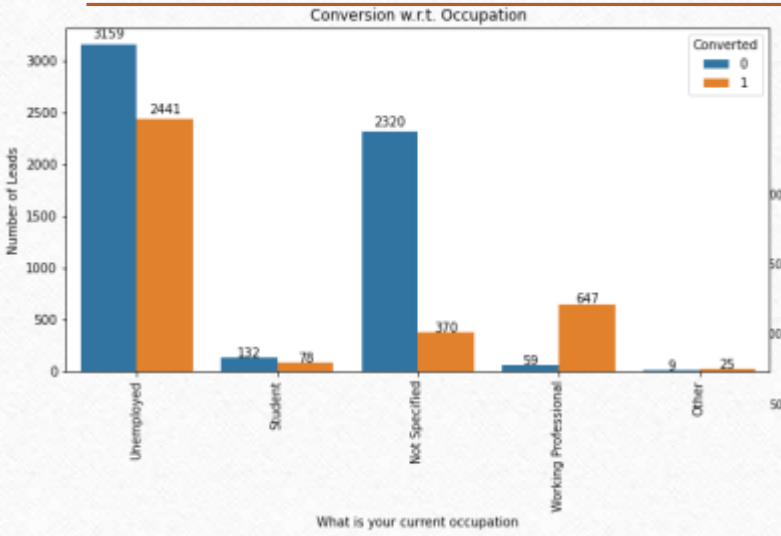
EDA

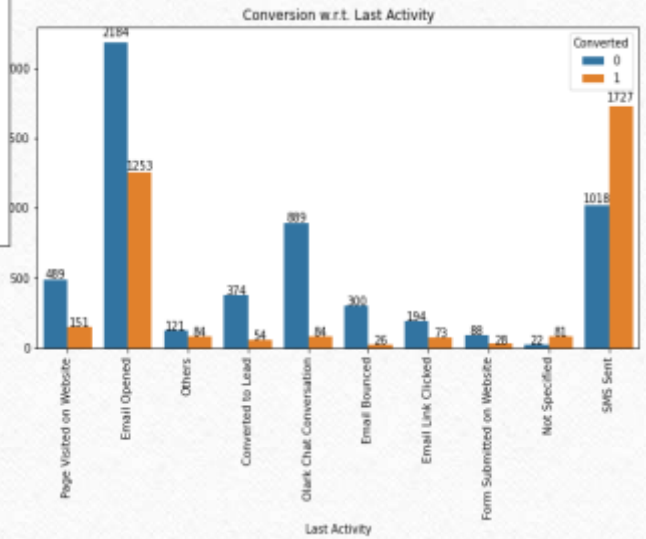

Conversion w.r.t. Origin Identifier
Converted
0
1
Number of Leads
Lead Origin
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
2465.0
1115.0
3118.0
1768.0
54.0
664.0
42.0
13.0
1.0

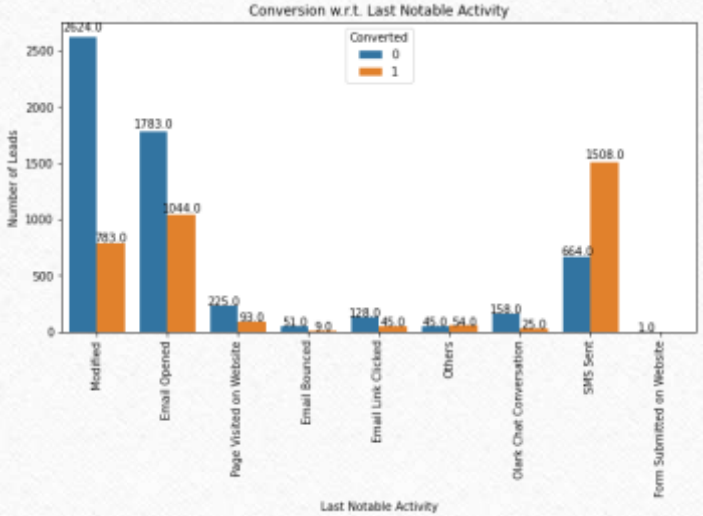

Conversion w.r.t. Source
Converted
0
1
Number of Leads
Lead Source
Olark Chat
Organic Search
Direct Traffic
Google
Other Sources
Welingak Website
Reference
1307
448
718
436
1725
818
1728
1176
155
53
2
140
44
490


Conversion w.r.t. Specialization
Converted
0
1
Number of Leads
Specialization
Not Specified
Business Administration
Media and Advertising
Management Specializations
Travel and Tourism
Banking, Investment And Insurance
Business Specializations
E-COMMERCE
Services Excellence
2411
969
224
179
118
85
2331
1922
131
72
171
167
192
116
72
40
29
11

EDA


Conversion w.r.t. Occupation
Converted
0
1
Number of Leads
3159
2441
132
78
2320
370
59
647
9
25
Unemployed
Student
Not Specified
Working Professional
Other
What is your current occupation


Conversion w.r.t. Last Activity
Converted
0
1
469
151
2184
1253
121
84
374
64
889
84
300
26
194
73
88
26
22
81
1018
1727
Page Visited on Website
Email Opened
Others
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Not Specified
SMS Sent
Last Activity


Conversion w.r.t. Last Notable Activity
Converted
0
1
Number of Leads
2624.0
783.0
1783.0
1044.0
225.0
93.0
51.0
9.0
128.0
45.0
45.0
54.0
158.0
25.0
664.0
1508.0
1.0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Others
Olark Chat Conversation
SMS Sent
Form Submitted on Website
Last Notable Activity

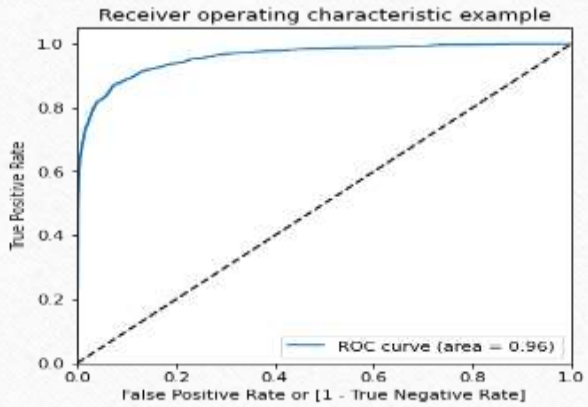
Variables impacting the conversion

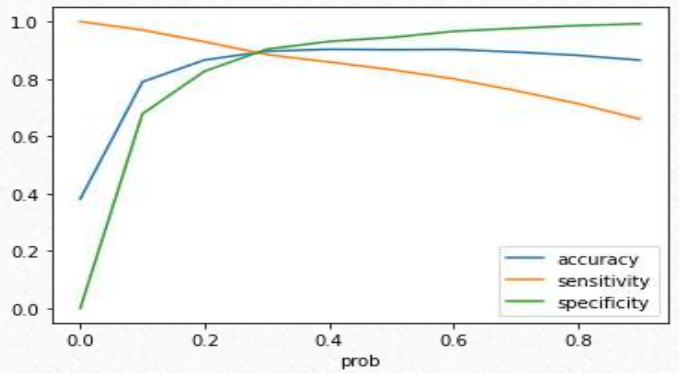
- It can be clearly seen that customers identified through the 'Lead add form' have higher conversion rate than rest of the identifiers
- Leads received through references and Wellingkak Website have a greater chance to get converted
- When the last activity performed by the customer is 'SMS Sent', then the chances of conversion are much higher (1727 converted out of 2745 which is more than 60% conversion)Overall accuracy on test dataset is 88%
- Leads with management specialisation have the highest number of conversions while leads working in services have the lowest conversion in all specializations
- When the Last Notable activity was 'SMS Sent', 1508 out of 2172 (close to 70%) of the leads were converted

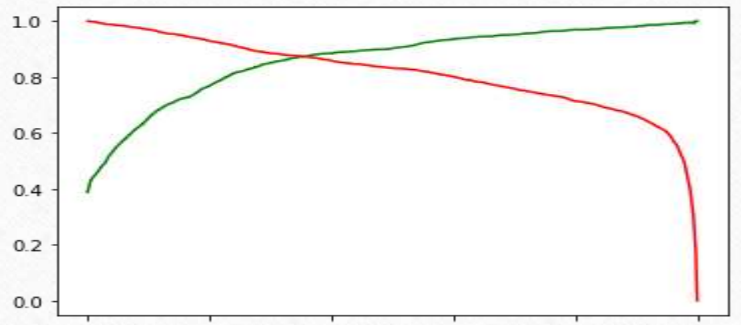
Model building

- Splitting our data into a training set and a test set.
- Feature Selection Using RFE running RFE with 20 variables.
- Dropping the variables having p-values greater than 0.05 and a value of VIF greater than 5
- Overall accuracy on test dataset is 88%.

ROC Curve







- Finding final optimal cutoff
  - Optimal cut off probability is when we get balanced sensitivity and specificality
  - Thus as per above, the cutoff can be considered as 0.38

Final Insights

Final Observation:

the values obtained for Train & Test:

A) Train Data:

- Accuracy:89.6%
- sensitivity:87.5%
- specificity: 90.3%

B) Test Data:

- Accuracy: 90.4%
- Sensitivity: 88.5%
- Specificity: 92.7%

Final Insights

- The model seems to predict the conversion rate quite well
- The top variables as per heatmap and residual coeffients:
- Total Time Spent on the Website
- Lead Origin_Lead Add Form
- Last Activity_SMS Sent
- Overall the model seems to be good one for Hot Leads.

Thank You