Lead scoring case study

By Sumit Thakar

sumitthakar124@gmail.com

Problem Statement :

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. 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%.

Strategy

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis
- 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

Methodology

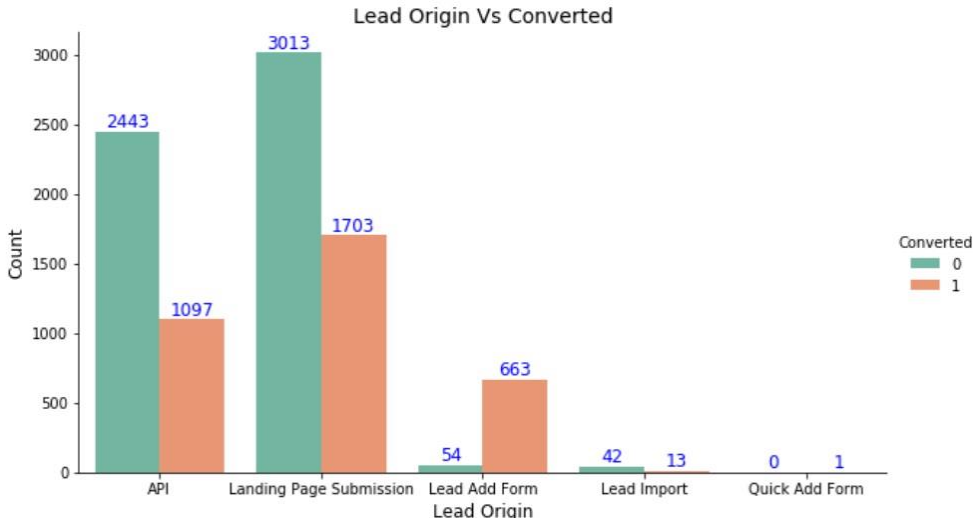
Data Sourcing , Cleaning and Preparation → Read the Data from Source, Convert data into clean format suitable for analysis, Remove duplicate data, Outlier Treatment • Exploratory Data Analysis, Feature Standardization.

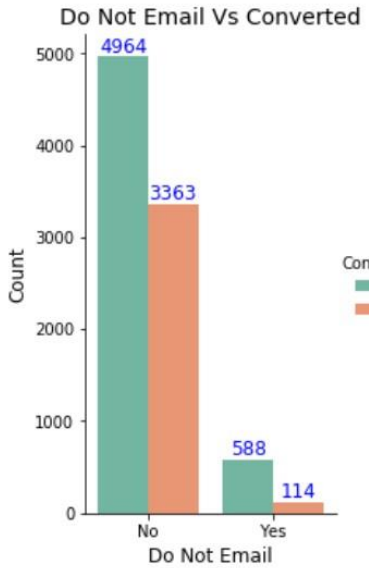
Feature Scaling and Splitting Train and Test Sets → Feature Scaling of Numeric data, Splitting data into train and test set.

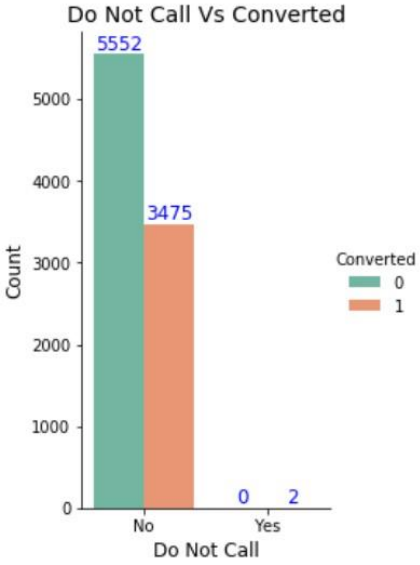
Model building→ Feature selection using RFE, Determine the optimal model using logistic regression, calculate various metrics

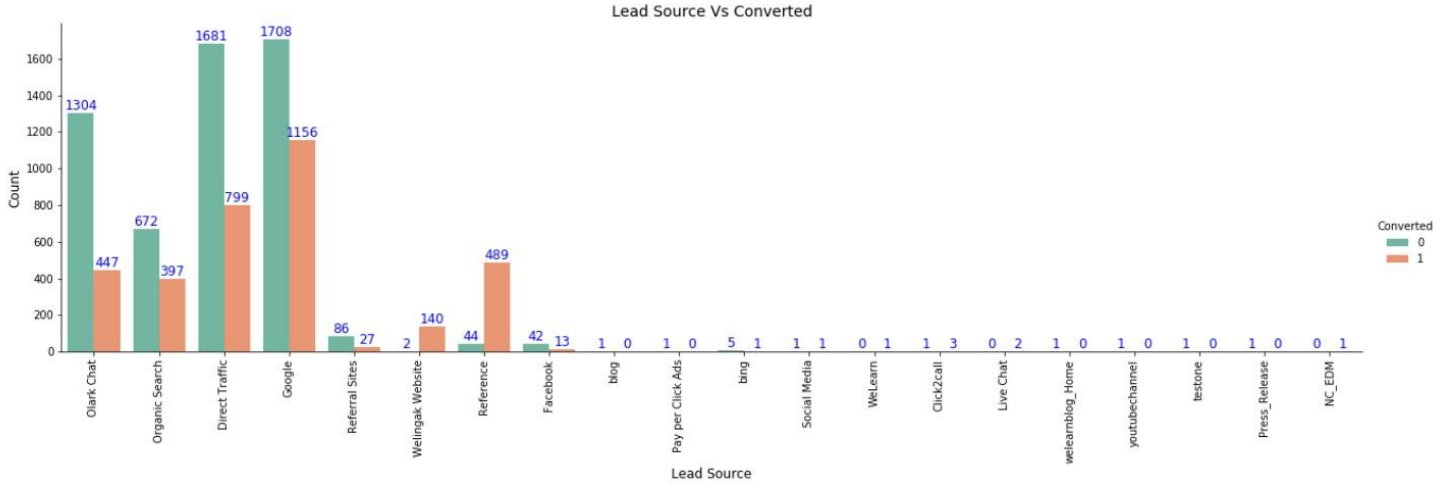
Result → Determine the lead score and check if target final predictions amounts to 80% conversion rate, Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics.

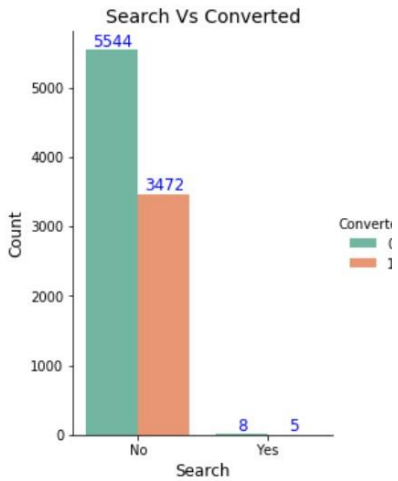
Exploratory data analysis:

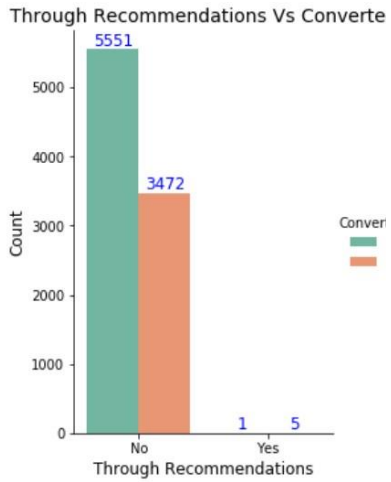


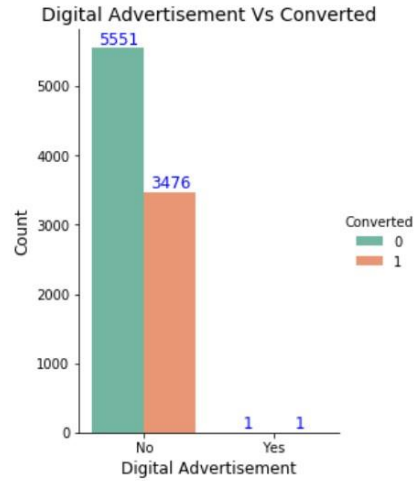


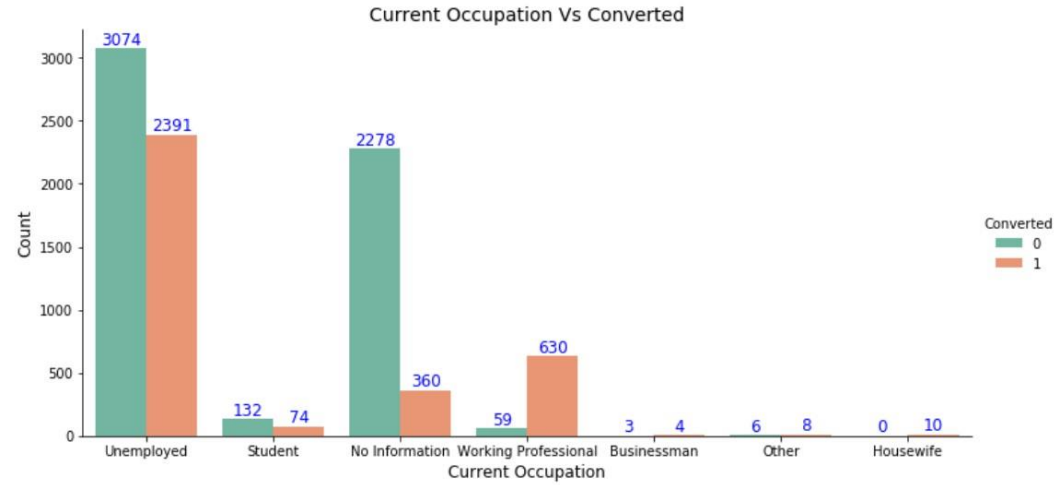


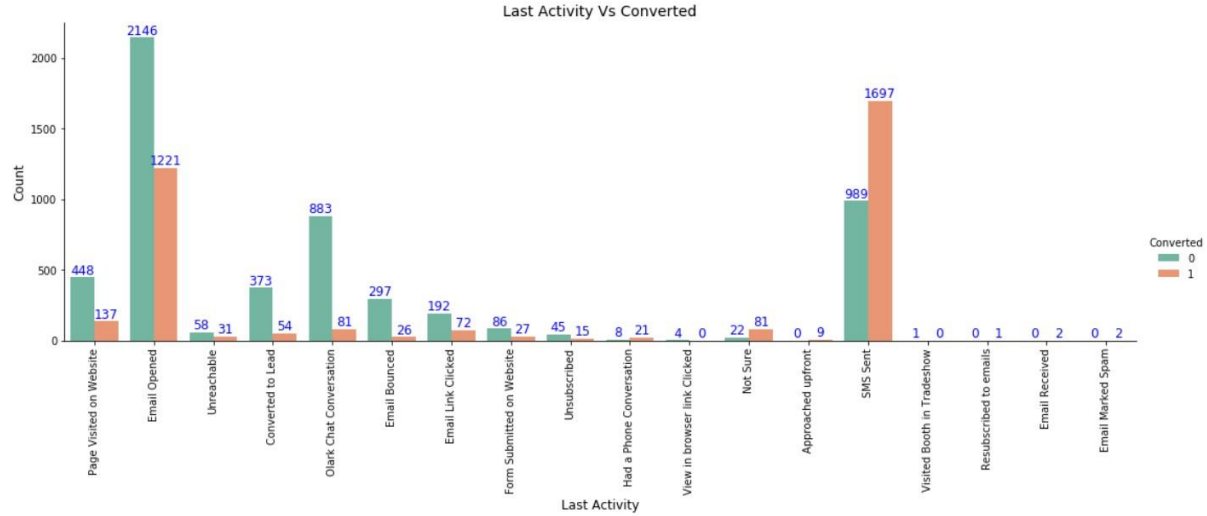



Search Vs Converted
5544
3472
8
5
Count
No
Yes
Search


Through Recommendations Vs Converte
5551
3472
1
5
Count
No
Yes
Through Recommendations


Digital Advertisement Vs Converted
5551
3476
1
1
Count
No
Yes
Digital Advertisement
Converted
0
1


Current Occupation Vs Converted
3074
2391
2278
360
132
74
59
630
3
4
6
8
0
10
Count
Unemployed
Student
No Information
Working Professional
Businessman
Other
Housewife
Current Occupation
Converted
0
1

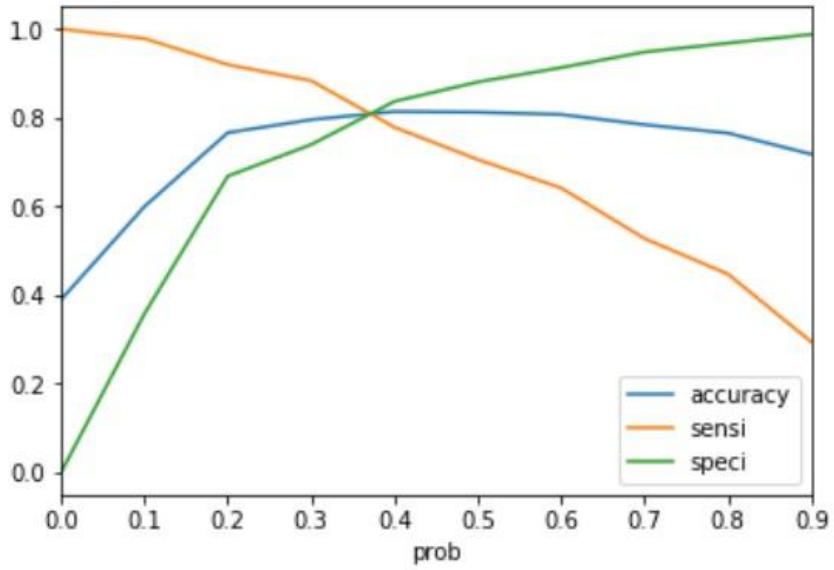

Last Activity Vs Converted
448
137
2146
1221
58
31
373
54
883
81
297
26
192
72
86
27
45
15
8
21
4
0
22
81
0
9
989
1697
1
0
0
1
0
2
0
2
Count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Not Sure
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity
Converted
0
1

Variables Impacting the Conversion Rate

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity - Unreachable

Model Evaluation - Sensitivity and Specificity on Train Data Set

- Accuracy - 81%
- Sensitivity - 80 %
- Specificity - 82 %
- False Positive Rate - 18 %
- Positive Predictive Value - 74 %
- Positive Predictive Value – 86%



Model evaluation- Precision and recall on train dataset

Precision - 79 %

Recall – 71%

Accuracy - 81 %

Sensitivity - 79 %

Specificity - 82 %