Lead Scoring Case Study

Report

Business Understanding:

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.

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%.

Business Problem and Objective:

Now, although 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.

Methodology adopted :

The business problem is to identify the potential leads in a customer for the education company.

Thus, in turn, the company wants to predict whether the leads are capable of being potential leads so that they could strategize their marketing accordingly to bring in more conversion rates.

This, is a typical classification problem, where we need to develop a Logistic Regression model and identify and predict the target variable output as 1 (Potential Hot lead) or 0(Not a potential lead).

Initially the data was cleaned, null values was handled, and some feature engineering was done to make it more model friendly.

The categorical variables were split into dummy variables

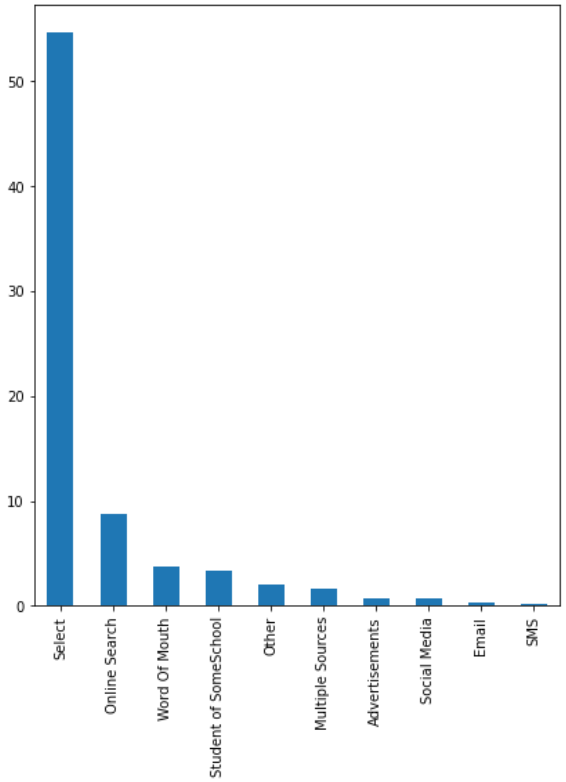
Later, the continuous data in the training data set was standardized and later fed to the model .

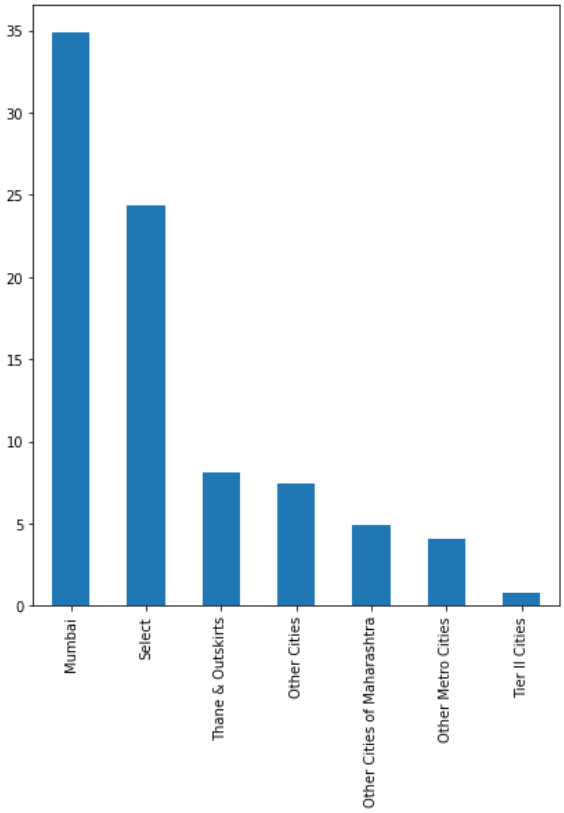
Considering that there are more than 130 columns ,Initially RFE(Recursive Feature Elimination) method was used to automatically reduce the features to 15 most impacting ones.

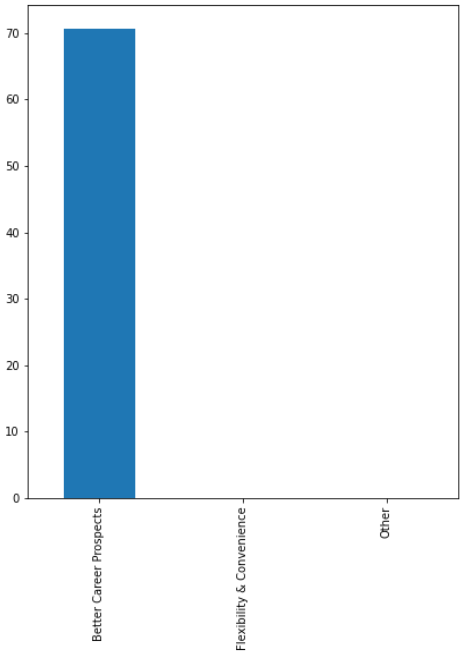
VIF was used to check for features collinearity. Then the GLM method was used to logit fit of the variables, and then various model evaluation metrics were used.

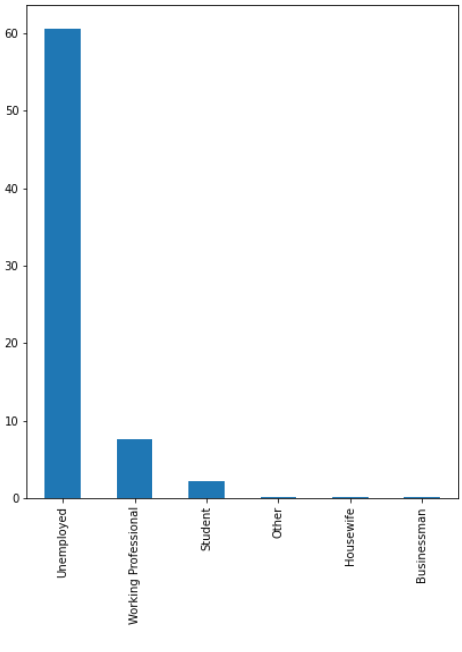
Visualizations :

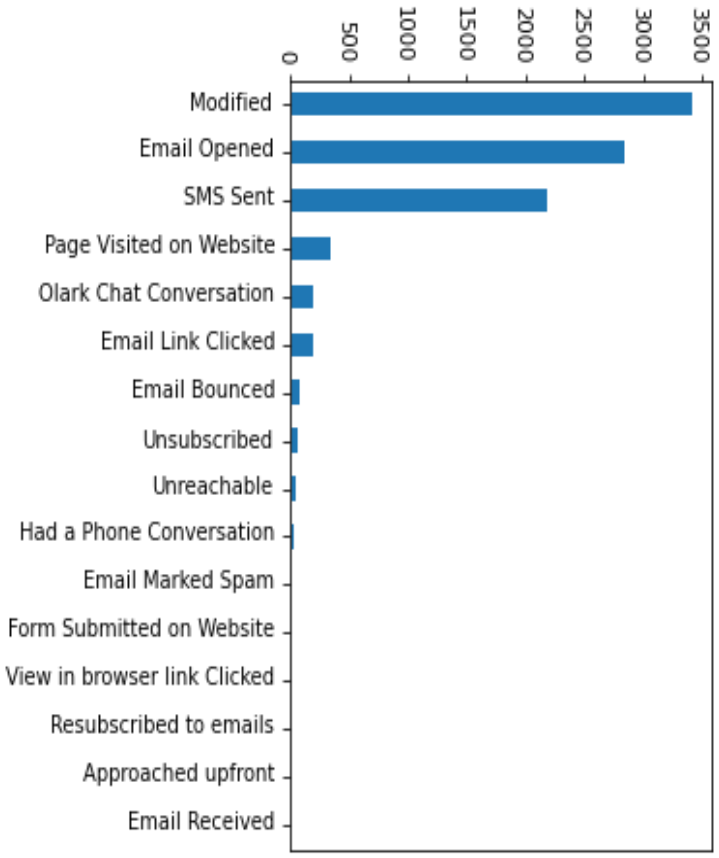
Distribution of various categories within a feature to identify the most important ones and least important ones


0
10
20
30
40
50
Select
Online Search
Word Of Mouth
Student of SomeSchool
Other
Multiple Sources
Advertisements
Social Media
Email
SMS


0
5
10
15
20
25
30
35
Mumbai
Select
Thane & Outskirts
Other Cities
Other Cities of Maharashtra
Other Metro Cities
Tier II Cities

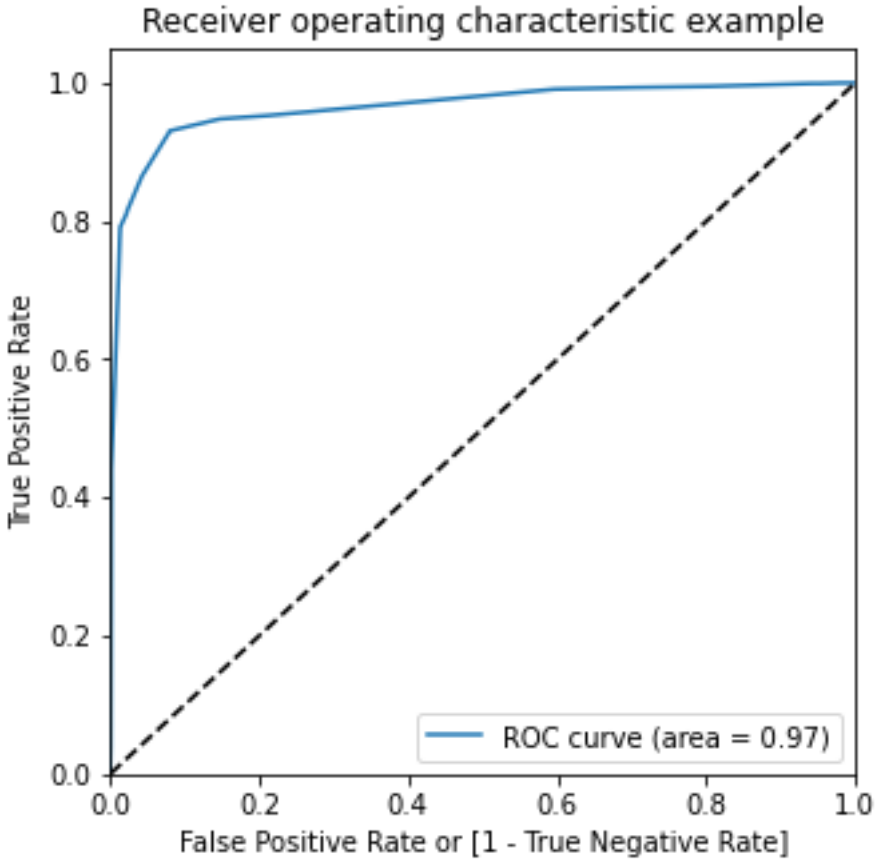

0
10
20
30
40
50
60
70
Better Career Prospects
Flexibility & Convenience
Other


0
10
20
30
40
50
60
Unemployed
Working Professional
Student
Other
Housewife
Businessman

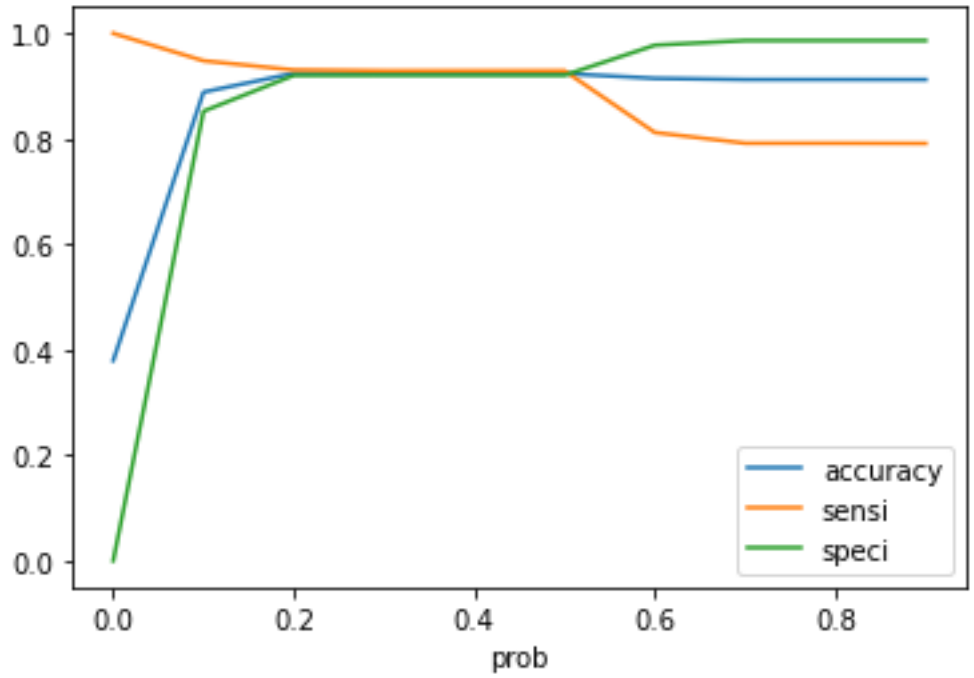

0
500
1000
1500
2000
2500
3000
3500
Modified
Email Opened
SMS Sent
Page Visited on Website
Olark Chat Conversation
Email Link Clicked
Email Bounced
Unsubscribed
Unreachable
Had a Phone Conversation
Email Marked Spam
Form Submitted on Website
View in browser link Clicked
Resubscribed to emails
Approached upfront
Email Received


0
250
500
750
1000
1250
1500
1750
2000
Select
Finance Management
Human Resource Management
Marketing Management
Operations Management
Business Administration
IT Projects Management
Supply Chain Management
Banking, Investment And Insurance
Travel and Tourism
Media and Advertising
International Business
Healthcare Management
Hospitality Management
E-COMMERCE
Retail Management
Rural and Agribusiness
E-Business
Services Excellence

Checking the ROC curve for model performance



Checking the optimal cut off point for obtaining maximum accuracy, sensitivity, and specificity



Model Evaluation:

Below metrics were obtained on the training and test data set:

1. **Training data set:**

   Accuracy: 92.37 %

   Sensitivity : 92.88 %

   Specificity: 92.05 %

   Precision: 87.73 %

   Recall: 92.88 %

2. **Test data set:**

   Accuracy: 92.05 %

   Sensitivity : 91.26 %

   Specificity: 92.53 %

   Precision: 88.3 %

   Recall: 91.26 %

Inferences:

Thus the model has also provided 91 percent sensitivity and 92 percent specificity even in the test data set. Thereby the model can be claimed to be fairly representative of the data set and can be used for identification of potential hot customers and predict whether the lead is likely to be converted.

The major features identified by the model that play an important role in the regression are :

1. Lead Source_Welingak Website
2. Tags_Closed by Horizzon
3. Tags_Lost to EINS
4. Tags_Will revert after reading the email
5. Tags_invalid number
6. Tags_switched off
7. Last Notable Activity_SMS Sent