Lead Scoring Case Study

BY

DEBJANI ROY CHOUDHURY

ARPAN AMETA

DIPTA GHOSH

Problem Statement

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%.

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.

Objective

To help X Education select most promising leads, i.e. the leads that are most likely to convert into paying customers.

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ € ₹ $
Converted leads

Problem Approach

- Importing Data
- Inspecting the Dataframe
- Data Preparation
- EDA
- Test-Train Split
- Feature Scaling
- Looking at Correlations
- Model Building
- Feature Selection Using RFE
- Plotting the ROC Curve
- Finding Optimal Cutoff Point
- Making predictions on the test set
- Calculating the lead score

Data Manipulation

Total Number of Rows =37, Total Number of Columns =9240.

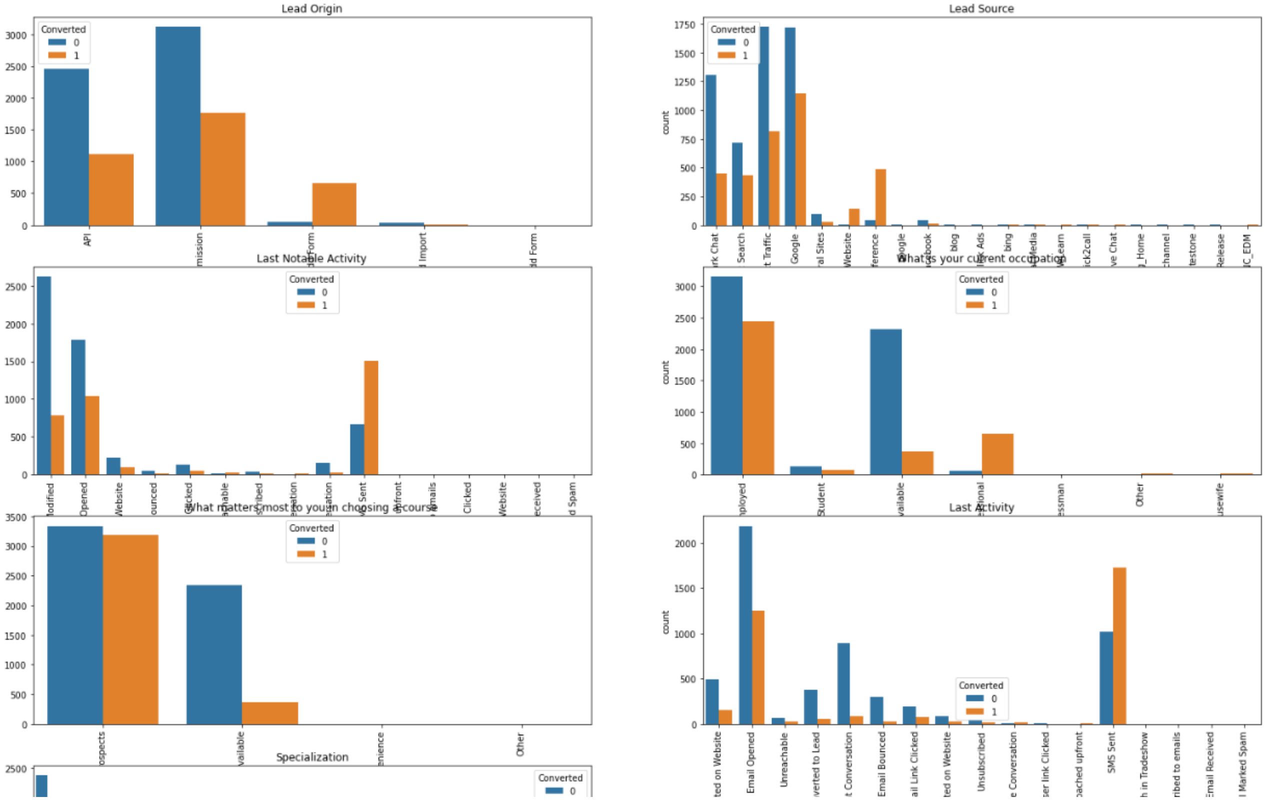
'Tags', 'Asymmetrique Profile Index', 'Asymmetrique Activity Index','Asymmetrique Activity Score','Asymmetrique Profile Score','Tags','Lead Quality’ are dropped as they have around 50% of missing values

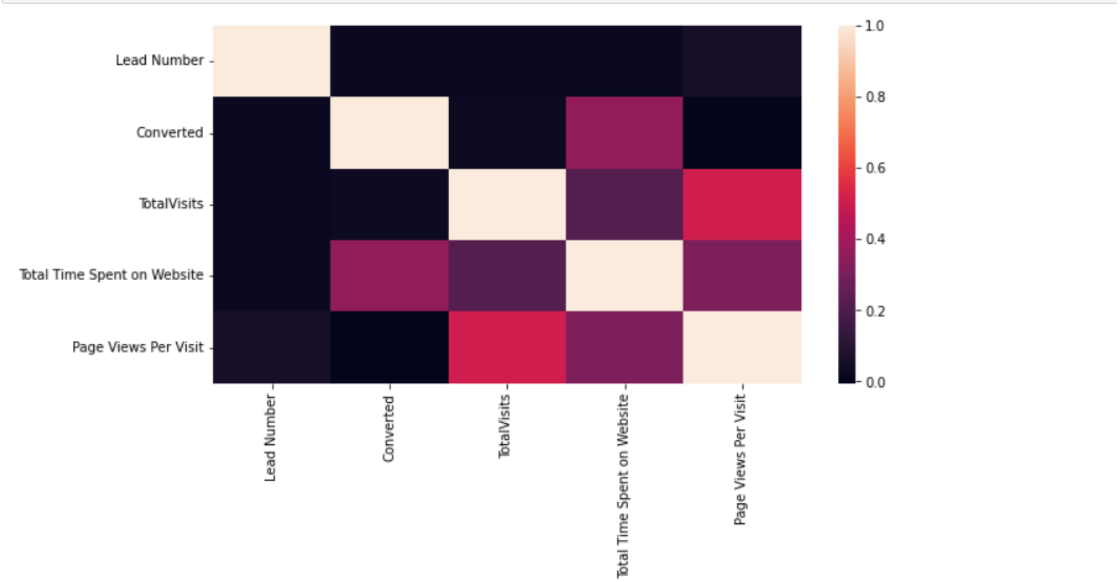
'Country', 'City’ were dropped as they were not much useful.

Replaced ‘Select’ with nan

For not removing much data, replaced the nulls with ‘not available’

EDA


Lead Origin
Converted
0
1
API
mission
Form
l Import
ld Form
Lead Source
count
rk Chat
Search
t Traffic
Google
al Sites
Website
ference
oogle
book
blog
Ads
bing
Media
eam
ick2call
ve Chat
l_Home
channel
estone
Release
IC_EDM
Last Notable Activity
Modified
Opened
Website
ounced
licked
able
ibed
ation
ation
Sent
front
mails
Clicked
Website
eceived
d Spam
What is your current occupation
ployed
Student
vailable
ssional
essman
Other
usewife
What matters most to you in choosing a course
ospects
vailable
enience
Other
Last Activity
ted on Website
Email Opened
Unreachable
verted to Lead
t Conversation
Email Bounced
ail Link Clicked
ted on Website
Unsubscribed
e Conversation
ser link Clicked
eached upfront
SMS Sent
h in Tradeshow
ribed to emails
Email Received
l Marked Spam
Specialization

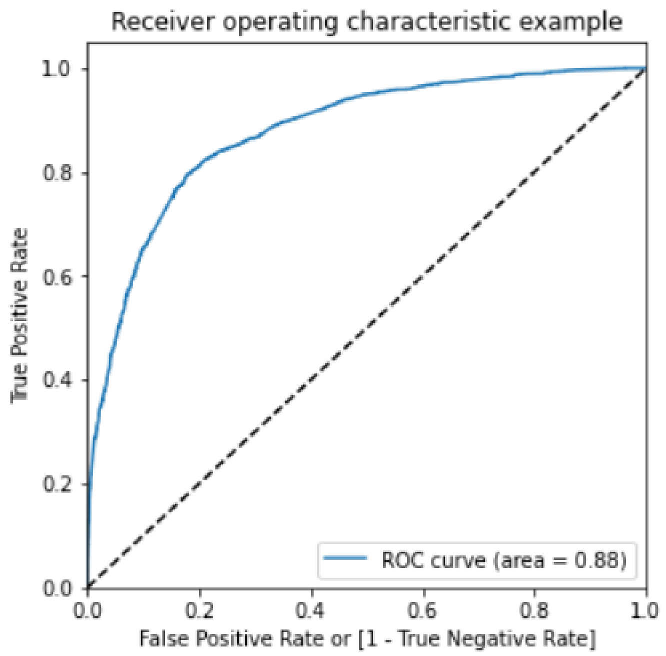

Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

Roadmap

Created train and test set by splitting the original cleaned data set after treating missing values.

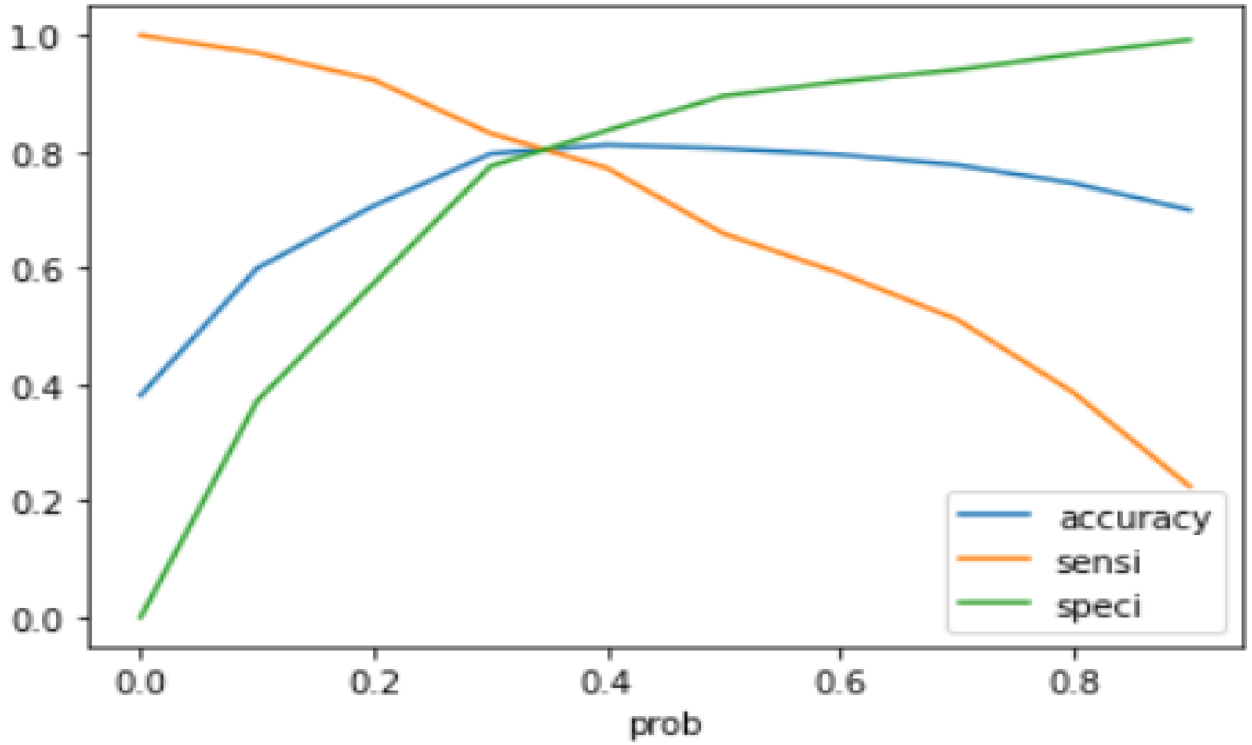
- Selected 15 features using Recursive Feature Elimination (RFE) after creating dummy variables and scaling the data.
- Applied Logistic Regression algorithm to build a model.
- Identified the optimal probability cutoff from the accuracy, sensitivity and specificity.
- Applied the model on the test data to identify the conversion probability.
- Based on the calculated predicted probability, and optimal probability cutoff, all the leads are assigned with a lead score value (lead score = predicted probability x 100)

ROC Curve



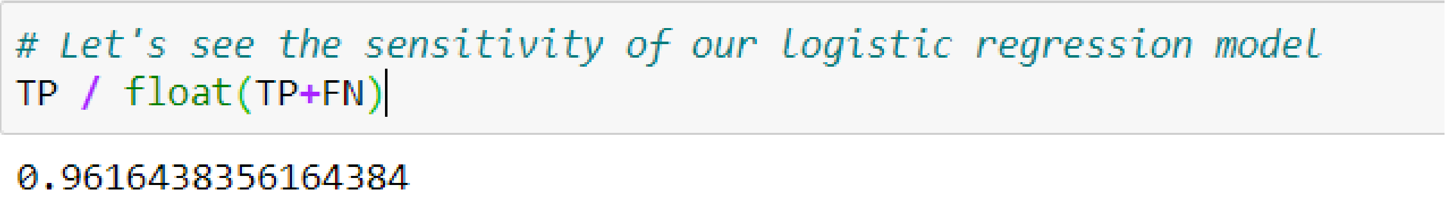
The model looks good as it far away from the diagonal. 88% area is covered under the curve.

Optimal Cutoff Point



The accuracy, sensitivity and specificity lines are intersecting at 0.3 probability.

Confusion matrix on Test data

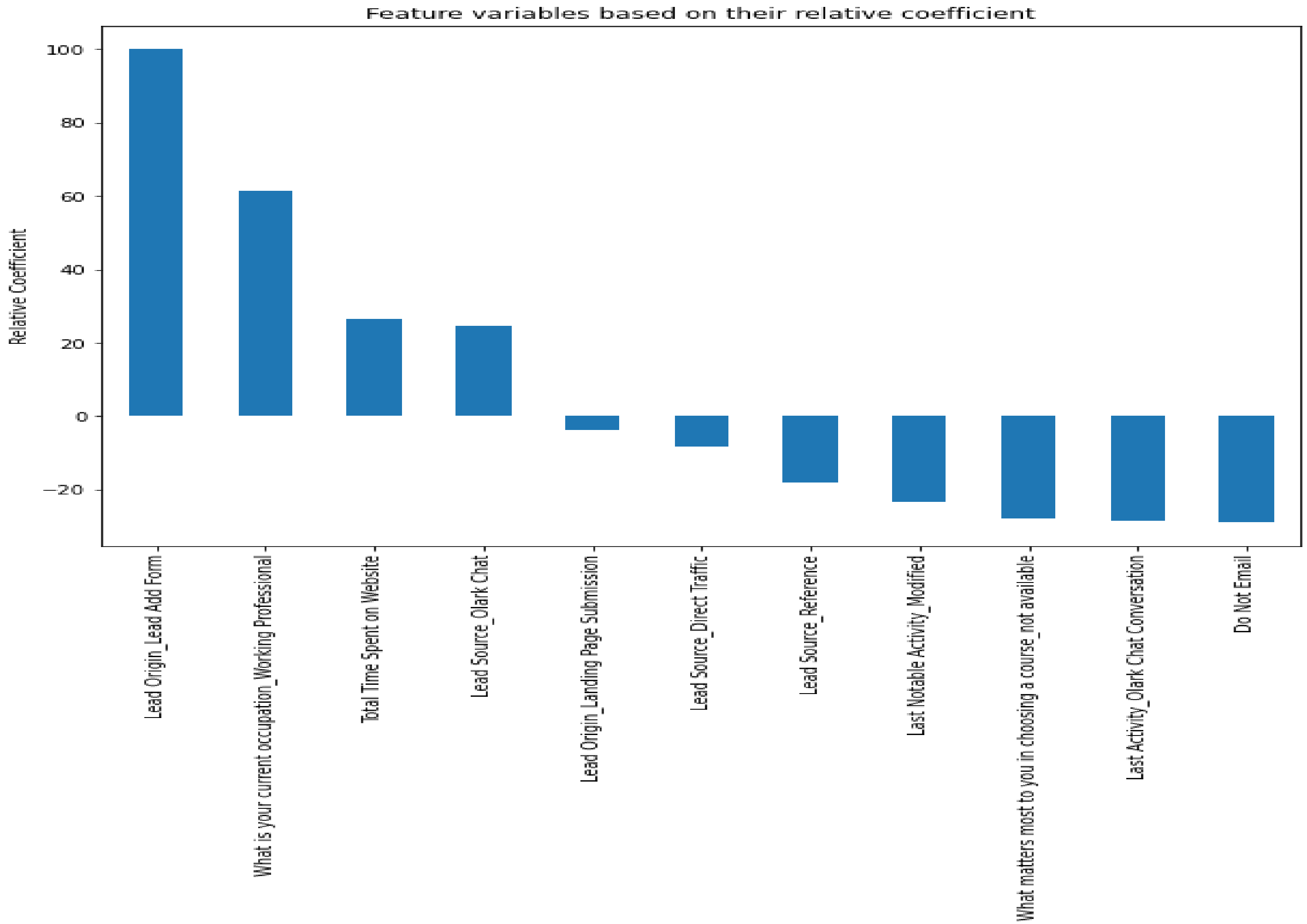


```python
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.9616438356164384
```

***Sensitivity of the prediction over test data set is 96%*¶**

Important Features



Top 3 features

The top 3 features which contribute most towards the probability of a lead getting converted:

i. What is your current occupation

Ii. Total Time Spent on Website

Iii. Lead Origin

Reccomendation

The leads which have high score can be treated as “hot” leads and sales team need to follow up as there is high possibility to convert those leads.

Leads who have applied for ‘Do Not Email’ already does not needs to be attended again.

Based on the previous chat conversations if the lead is classified as ‘Might be’ or ‘Worst’ then those leads can be ignored.