Lead Scoring Case Study Summary

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 market 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 courses 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 lead 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 lead 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 to up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

Lead Conversion Process- Demonstrated as a funnel

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e., educating the leads about the product, constantly communicating etc.) in order to get higher lead conversion

X Education has appointed you to help them select the most promising leads i.e., the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach:

Below are the steps we followed to solve this problem:

Step 1: Importing Libraries and Data

- Suppressed warnings
- Imported required libraries
- Imported dataset to csv

Step 2: Inspecting the Dataframe

- Checked the head of dataset
- Checked the dimension of the dataframe
- Checked columns info
- Checked for duplicates (if any)

Exploratory Data Analysis

Step 3: Data Cleaning

a) Identified Missing Values
b) Dropped Columns with Missing Values
c) Categorical Attributes Analysis
d) Numerical Attributes Analysis

Step 4: Data Preparation

a) Converted some binary variable
b) Dummy Variable Creation

Step 5: Test-Train Split

- Imported library for splitting dataset
- Put feature variable to X
- Checked the head of X
- Put responses variable to Y
- Checked the head of Y
- Split the data into train and test

Step 6: Feature Scaling

- Imported library for feature scaling
- Scaling of feature
- Checked the X-train dataset after scaling
- Checked the conversion rate
- Saw the correlation matrix
- Dropped the highly correlated dummy variables
- Checked the correlated matrix

Step 7: Model Building

- Started by splitting the data into a training set and a test set
- Run training model

- Created prediction
- Evaluated model

  With the current cut off as a 0.5 we have around 79% accuracy, sensitivity of around 63% and specificity of around 89%.
- Optimise cut off (ROC curve)

  Optimal cut off is at 0.35.
- Prediction on Test set

  With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 84%.
- Precision Recall

  With the current cut off as 0.35 we have Precision around 78% and Recall around 70%
- Precision and recall tradeoff
- Precision on Test set

Conclusion:

- It was found that the variable that mattered the most in the potential buyers are:
  1. Total Visit
  2. Total Time Spent on Website
  3. Page Views Per Visit
- The top three categorical/dummy variables in the model that focused the most on in order to increase the probability of lead conversion are:
  1. Lead Origin
  2. Lead Source
  3. Do Not Email
- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.