Lead Scoring Case Study : Summary

1. Problem Statement

X Education sells online courses to industry professionals. They need help in selecting leads that are most likely to convert into paying customers. The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

2. Roadmap

We carried out this analysis for X Education to find ways to get more people to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate. The following are the steps used:

a. Reading and Understanding Data:

Read and analyze the data.

b. Cleaning data:

Data was partially clean except for a few null values and the value select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to ‘not provided’ to avoid data loss. Although, they were later removed while making dummies.

c. EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric variable seems good.

d. Dummy Variables:

Dummy variables were created for categorical variables and for numeric values we used the MinMaxScaler.

e. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

f. Model Building:

RFE was done to attain the top 18 relevant variables. Later, rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

g. Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

h. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.39 with accuracy, sensitivity and specificity around 80%.

i. Precision – Recall:

This method was also used to recheck and a cut off of 0.41 was found with Precision around 79% and recall around 71% on test data frame.

It was found that the variables determining the potential buyers are (In descending order):

The total time spend on the Website.
Total number of visits.
When the lead source was:
    -Google
    -Direct traffic
    -Organic search
When the last activity was:
    -SMS
    -Olark chat conversation
When the lead origin is Lead add format.
When their current occupation is as a working professional.

Keeping these in mind the X Education can increase there sales as they have a very high chance to get almost all the potential buyers.