

Summary – Lead Score Case Study

This analysis is done for X Education and to find ways to get more industry professionals 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.

Below steps are followed:

1. Cleaning data:

The data was partially clean except for a few null values and the option 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' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Other Country' and 'Not Provided'.

2. EDA:

EDA was done to visualize the variables and its importance. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and outliers were found but looks fine.

3. Dummy Variables:

The dummy variables were created. For numeric values we used the MinMaxScaler.

4. Train-Test split:

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

5. Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with $VIF < 5$ and $p\text{-value} < 0.05$ were retained).

6. 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 90% each.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.3 with accuracy, sensitivity and specificity of 91%.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.38 was found with Precision around 90% and recall around 89% on the test data frame.

9. Conclusion:

X-Education company can focus on the following are the most important variables that a lead can become a converter:

- Tags_Closed by Horizon
- Tags_Lost to EINS
- Total Time Spent on Website
- Tags_Will revert after reading the email
- Last Activity_SMS Sent