

Summary

Analysis was 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 lead conversion rate.

Steps

Data Cleansing –

Data provides was partially clean. There were few null values. Some of the null values were changed to 'not provided'. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

EDA –

EDA was done to check the distribution of data and observed that found that a many elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

Dummy Variables –

The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

Train-Test split –

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

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 kept).

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.

Prediction –

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

Precision – Recall –

Was used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 75% on the test data frame.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
 - a. Google
 - b. Direct traffic
 - c. Organic search
 - d. website
4. When the last activity was:
 - a. SMS
 - b. Chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

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.