

Summary

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:

1. **Cleaning data:**

The data was cleaned of null values exceeding 40% and the option select had been replaced with a null value. Since there were many from India and few from outside dropped the column.

2. **EDA:**

It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good, and no outliers were found.

3. **Dummy Variables:**

The dummy variables were created and for numeric values we used the Standard Scaler.

4. **Train-Test split:**

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

5. **Model Building:**

RFE was done to attain 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF and p-values.

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 approximately 80%.

7. **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.30 with accuracy, sensitivity and specificity of approximately 80%.

8. **Precision – Recall:**

This method was also used to recheck and a cut off of 0.42 was found with Precision around 69% and recall around 74% on the test data.

It was found that the variables that mattered the most in the potential buyers are

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was Google, Direct traffic, Organic search, Welingak website
4. When the last activity was SMS, Olark 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 increase their buyers by focusing on the above to convert potential leads successfully.