

Summary

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.

The following are the steps used:

1. Cleaning data:

The data contained few columns where more than 40% of the values were null. These columns were dropped as imputing these would make the data inaccurate. Some columns had the option “Select” which is almost the equivalent of null values. These were removed after dummy creation.

After null value treatment, the format of the data was checked for any discrepancies and fixed.

2. EDA:

Exploratory Data Analysis was performed to check for outliers and to understand the distribution of the data.

The outliers were found in some columns which were treated to make the data more accurate.

3. Dummy Variables:

After the dummy variables were created, the dummies with ‘not provided’ and “Select” elements were removed as these are null values. MinMax Scalar was used to convert the numerical values to fit between 0 and 1.

4. Train-Test split:

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

5. Model Building:

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 $p\text{-value} < 0.2$ were kept).

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 precision which came to be around 80% each.

7. Prediction:

Prediction was done on the test data frame with accuracy, sensitivity and precision of 80%.

8. Area of ROC:

The area under the curve is 0.88 after plotting the ROC.

9. Precision – Recall:

This method was also used to recheck and a cut off of 0.5 was found with Precision around 79% and recall around 68% on the test data frame.

ATTRIBUTES TO FOCUS ON:

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

1. The total time spent on the Website.
2. Lead Origin_Lead Add Form – The leads that fill in the form available on the website.
3. Last Notable Activity_Had a Phone Conversation – Whether the leads were contacted by the sales team.

X Education can convert the leads to participants if these attributes are focused on by the sales team.