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 in the process:

1. Cleaning data: The data was first cleaned basis null values in a coarse manner an was later studied thoroughly and all the duplicate data was removed. While making dummies the data which less impact has also been removed.

2. 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 values initially contained many outliers which were therefore dropped.

3. Dummy Variables: The dummy variables were created and later on the dummies with 'not provided' elements were removed. 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: 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-value < 0.05 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 specificity which came to be more than 83% & 91.40% respectively.

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 89%, 85% & 91.58% respectively.

8. Precision – Recall: This method was also used to recheck and a cut off 0.2 was found with Precision around 90% and recall around 82% 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. Welingak website.
4. When the last activity was: a. SMS b. Olark chat conversation.

5. When the lead origin is Lead add format.

6. When their current occupation is as a working professional.

Keeping all 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.