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 theconversion rate.

The following are the steps used:

1. **Cleaning data:**

- Dropping columns that have only one unique values for all the leads.
- Handling 'Select' variable that is present in many categorical variables.
- Droping all the columns with more than 40% missing values
- Dropping columns with high data imbalance
- Using imputation technique for columns having less % of missing values
- Combining categories having low percentages into one single category.

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 seems goodand no outliers were found.

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 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 thevariables were removed manually depending on the VIF values and p-value (Thevariables 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 around 80% each

Conclusion

The difference between the test and train datasets performance metrics is very less, hence, our final model is performing well.

The top 3 features that can account for lead conversion are:

-The customers who fill Add form

-Working Professionals

-Total time spend on website

High sensitivity means that most of the leads who are likely to convert are correctly predicted, whereas high specificity ensures that most of the leads who are not likely to convert are correctly predicted. Hence we have achieved high sensitivity 80% which is required of us to do.