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 had few null values and the option select replaced as NaN value, imputed missing values, dropped highly skewed features, dropped features with only one category, grouped categories having less percentage of value count.

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. All outliers for few numerical variables were replaced with upper bound.

3. Dummy Variables: The dummy variables were created for categorical features.

4. Train-Test split: The split was done at 70% and 30% for train and test data respectively.

5. Scaling of numerical data

6.Model Building: Firstly, RFE was done to attain the top 16 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).

7. Model Evaluation: A confusion matrix was made. With the current cut off as 0.36 we have around 79% accuracy, sensitivity of around 74% and specificity of around 83%

8. Prediction: With the current cut off as 0.39 we have around 79% accuracy, precision around 74%, recall around 71%

Conclusion:

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

1. Total Time Spent on Website.

2. Lead Origin_others:

    a. lead add form
    b. lead import

c. quick add form

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

4. When the lead source was:

    a. Olark chat conversation

5. When their current occupation is unemployed
Keeping these in mind the X Education can flourish as they have a very high chance