Summary

This analysis is performed for X Education and to find ways to get more industry professionals to join their courses. The dataset provided gave us a lot of information about how the potentials customers visit the site, the time they spend over there and how they reached the site and the conversion rate.

The following technical steps are used:

Step 1: Reading and Understanding the Data

1. We try to observe the dataset and find any missing values.
2. Next, we find the shape of your dataset and identify relationships between various columns present in our dataset.

Step 2: Data Cleaning

1. The data set was partially clean except for a few null values and the option 'Select' has to replace with a null value since it did not give us much information.
2. Next step is to clean the dataset we choose to remove the redundant variables/features. • Dropped the high percentage of Null values more than 40%.
3. Checked for number of unique Categories for all Categorical columns.
4. From that Identified the Highly skewed columns and dropped them.
5. Treated the missing values by imputing the favourable aggregate function like (Mean, Median, and Mode).

Step 3: Exploratory Data Analysis

1. 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 seem good but found the outliers

2. Performed Univariate Analysis for both Continuous and Categorical variables.

3. Performed Bivariate Analysis with respect to Target variable.

Step 4: Data Preparation and Features Selection

1. Dummy Variables: The dummy variables are created for all the categorical columns.

2. Scaling: Used Standard scalar to scale the data for Continuous variables.

3. Train-Test Split: The Spit was done at 70% and 30% for train and test the data respectively.

Step 5: Building a Logistic Regression using stats model, for the detailed statistics

1. We can see that our model is doing well in test set also

2. Sensitivity means how our model is telling that actually converted and model predicted them as converted.

3. We can see that our model is giving about .80 sensitivity.

4. It means that 80 percent time our model is able to predict (actually)converted as (predicated)converted.

1. Accuracy = 0.7766955266955267
2. Sensitivity = 0.8018264840182648
3. Specitiy = 0.7602862254025045

Conclusion:

We have noted that the variables that important the most in the potential buyers are:

1. The total time spends on the website.

2. Total number of visits.

3. When the lead source was:

    a. Google b. Direct traffic c. Organic search d. Olark Chat

4. When the last activity was:

    a. SMS b. Olark chat conversation

5. When the lead origin is Lead add format