SUMMARY

This analysis is done for an education company named X Education sells online courses to industry professionals and to find ways to get more industry professionals to join their coursed. The basic data provided gave us a lot of information about how they reached the site and the conversion rate, , the time they spend there, how the potential customers visit the site.

The following are the steps used:

I. Cleaning data & Treatment:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Although they' were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India'.

2. **Handle Binary columns:**

--That have significant data imbalance drop those columns

--Drop all those columns that have only 1 unique entry

3. **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 seem there are upper bound outliers in both Total Visits and Page Views Per Visit columns. We can also see that the data can be capped at 99 percentiles.

3. **Data Preparation**:

Converting Binary (Yes/No) to 0/1

4. **Dummy Variables**:

The dummy variables were created and later on the dummy's elements were removed. Creating a dummy variable for some of the categorical variables and dropping the first one. Dropping the columns for which dummies have been created .

5. **Model building logistic regression Train-Test split**:

Splitting the data into Training and Testing Set - For this we need to import Train Test Split from SKLearn. The split was done at 70% and 30% for train and test data respectively.

6**. 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).

7**. 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 90% each.

8. **Prediction**:

Prediction was done on the train data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity came to be around 90% each.

9. **Precision —Recall**:

This method was also used to recheck and a cut off of 0.4 I was found with Precision around 88.65% and recall around 91.43% on the test data frame.
It was found out that the variables that mattered the most in the potential buyers

1. The Total Time Spent on Website.

2. Total Visits.

3**.** Page Views Per Visit

4. A free copy of Mastering the Interview

5. Lead Origin Landing Page Submission

6. Last Notable Activity Olark Chat Conversation

7. Lead Origin Lead Add Form and Others.

6. When the current occupation was:
a. Working Professionals
b. Student
c. Unemployed
d. Other

The above-mentioned points are kept in mind the X education can increase all the potential buyers to change their mind and buy their courses.