# 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:**
   The data had some missing values and a specific category "Select" was replaced with null to avoid losing information. Some of the null values were replaced with "not provided" to retain as much data as possible. However, these were later removed when creating dummy variables. The data also showed a large number of entries from India, and a smaller number from other countries. To reflect this, entries were grouped as "India", "Outside India", and "not provided".
2. **EDA:**
   A preliminary exploration of the data, known as EDA (Exploratory Data Analysis), was carried out to assess the quality of the data. It was found that many of the values in the categorical variables were not useful. The numerical values appeared to be in good condition and no unusual or extreme values were detected.
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 Scalar.

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 around 80% each.

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**
   This method was also used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 75% on the test data frame.

It was found that the variables that mattered the most in the potential buyers are

- Lead Source_reference
- What is your current occupation working professional
- Lead Source_welingak website
- It's worth noting that In this data, it looks like TotalVisits, Total Time Spent on Website, Lead Source_olark chat, What is your current occupation_working professional and Last Notable Activity_sms sent are positively correlated with the outcome variable, while Lead Origin_landing page submission, Lead Source_direct traffic, Last Activity_converted to lead, Last Activity_email bounced, Last Activity_olark chat conversation, Last Activity_page visited on website, Do Not Email_yes, What is your current occupation_unemployed and Last Notable Activity_modified are negatively correlated with the outcome variable. These relationships are statistically significant because the P-values are less than 0.05.

Keeping 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.