SUMMARY

This research is being conducted for X Education in order to determine how they might attract more industry professionals to their courses. The simple information provided provides us with a wealth of information about how potential customers interact with the site, including how long they stay there, how they arrived, and the conversion rate.

The following steps are used to create the model:

I. **Data Cleaning:**
Except for a few null values, the data was mostly clean, and the option select had to be replaced with a null value because it didn't provide us with much information. To avoid losing too much data, a few of the null values were changed to 'not mentioned.' Despite the fact that they were later removed while making dummies. Because there were so many Indians and so few foreigners, the elements were changed to 'India,' ' foreigners,' and 'not mentioned.'

II. **Exploratory Data Analysis:**
A quick EDA was performed to assess the state of our data. Many elements in the categorical variables were discovered to be irrelevant. The numerical values appear to be correct, and no outliers were discovered.

III. **Creation of Dummy Variables:**
The dummy variables were generated, and the ones that had 'not mentioned' elements were later removed. We utilized the MinMaxScaler for numeric values.

IV. **Train-Test split:**
The train and test data were split 70 percent and 30 percent, respectively.

V. **Building the Model:**
RFE was first used to identify the top 15 relevant factors. The remaining variables were then manually deleted based on their VIF values and p-values (the variables with VIF< 5 and p-value < 0.05 were maintained).

VI. **Evaluating the Model:**
Here we first create a confusion matrix. Later, the ideal cut off value (as determined by the ROC curve) was utilized to determine the accuracy, sensitivity, and specificity, all of which were found to be around 80%.

VII. **Prediction on test data set:**
The prediction was carried out on the test data frame with an ideal cut off of 0.35 giving an accuracy, sensitivity, and specificity of 80%.

VIII. **Evaluating the model using Precision – Recall:**
On the test data frame, this procedure was also utilized to recheck, and a cut off of 0.41 was discovered, with Precision around 73 % and recall around 75% percent.

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

1. Total number of visits.
2. The total time spend on the Website.
3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
4. When the last activity was:
    a. SMS
    b. Olark chat conversation
5. When the lead origin is Lead add form.
6. When their current occupation is as a working professional, student, unemployed or other.
7. The option given by the company to get an email from company is an important variable.
8. The Last Notable Activity where student is unreachable.