SUMMARY REPORT.

This analysis is conducted for X Education to explore strategies for attracting more industry professionals to their courses. The provided data offers valuable insights into potential customers' site visits, time spent on the site, referral sources, and conversion rates.

The following steps were taken:

Data Cleaning:

- Checked the number of missing values in each column.
- Dropped the variable “City” and “Country” as it’s of no use.
- There are 3 columns in which there is a level called “Select” which basically means that the student had not selected the option for that specific column which is why it shows “Select”. To get some useful data we have to make compulsory selection. For instance, Specialization, Lead Profile etc.
- The levels Lead Profile and How did you hear about X Education have a lot of rows which have the value Select which is of no use to the analysis so it's best that we drop them.
- Also, the variable What matters most to you in choosing a course has the level Better Career Prospects 6528 times while the other two levels appear once twice and once respectively. So, we should drop this column as well.
- Dropped the number of null values present in the columns “Total visits”, “Specialization” and “Lead source” as they are quite small.

Dummy Variable:

- Checked the columns which are of type 'object'.
- Created dummy variables using the 'get_dummies' command.
- For numeric values we used the MinMaxScaler.

Test-Train Split:

The split was done at 70% and 30% for train and test data respectively.

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).

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.

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%.

Precision-Recall View:

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.