Summary

The purpose of this analysis is to assist X Education in increasing the number of industry experts enrolled in their courses. We learned a great deal about how potential clients visit the website, how long they stay there, how they got there, and the conversion rate from the basic data provided.

The following are the steps used:

1. **Cleaning data:**
   The option select had to be changed to a null value because it provided us with insufficient information, and the data was mostly clean aside from a few null values. To minimize the amount of data lost, a small number of the null values were changed to "not provided." while creating dummies, they were later taken out. India, Outside India, and "not provided" were substituted for the original elements because there were more Indians than outsiders.

2. **EDA:**
   We performed a brief EDA to assess the quality of our data. Numerous components in the category variables were discovered to be meaningless. No outliers were discovered, and the numerical results appear to be reasonable.

3. **Dummy Variables:**
   The dummy variables were made, and thereafter the ones that had items marked as "not provided" were eliminated. The MinMaxScaler was utilized for numerical values.

4. **Train-Test Split:**
   The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**

   First, RFE was used to identify the twenty most important factors. Afterwards, based on the VIF values and p-value, the remaining variables were carefully eliminated (the variables with VIF < 5 and p-value < 0.05 were preserved).

6. **Model Evaluation:**
   A matrix of confusion was created. Subsequently, accuracy, sensitivity, and specificity were determined by utilizing the ROC curve to determine the optimal cut off value; these values were approximately 80% for each.

7. **Prediction:**

   An optimal cut off of 0.35 was used for the prediction, which was performed on the test data frame with 80% accuracy, sensitivity, and specificity.

8. **Precision-Recall:**

   A cut off of 0.45 was discovered using this procedure for a recheck, with recall and precision at 75% and 75%, respectively, on the test data frame.

   It was found that the variables that mattered the most in the potential buyers are:
   1. The total time spends on the Website.
   2. When the lead source was: a. Google b. Direct traffic c. Organic search d. Welingak website
   3. When the lead origin is Lead add format.
   4. When their current occupation is as a working professional. 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.