SUMMARY

The Aim of this analysis was to find 'hot leads' among leads using the data given by the company X Education. The company wanted to increase the conversion rate of leads to customers.
For that purpose:

1. **Data Cleaning:** We cleaned the data like removing null values, fixing outliers, etc. We changed the option 'Select' as 'not available'. We grouped together the countries except India as outside India.

2. **EDA:** The EDA revealed that many categorical values were useless like the Ads. The numeric values were good and were later used in the model.

3. **Dummy Variable:** Dummy variables were created for columns with more 2 categories. MinMaxScaler was used for numeric values.

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

5. **Model Building:** RFE was done to obtain 15 significant variables and the model was built by removing the unnecessary variables by using p-value and VIF.

6. **Model Evaluation:** The model was evaluated using ROC Curve. We found 0.35 as optimal cut-off for Accuracy 80%, Sensitivity 79% and Specificity 80%.

7. **Precision-Recall:** Using Precision-Recall trade off, we found 0.41 as the optimal cut-off for Precision and Recall.

8. **Prediction on Test Data:** Using 0.41 as the cut-off for predicting on the test set, we found Accuracy 81%, Sensitivity 77%, Specificity 83%, Precision 74% and Recall 77%.

The top four significant variables are:
1. Total Time Spent on Website
2. What is your current occupation_Working Professional
3. Lead Origin_Lead Add Form
4. Last Activity_Had a Phone Conversation

Total Time Spent on Website has the highest coefficient value = 4.4

The company can also check the problems with ads since most of the customers have not seen the ads in any platform and thereby they are not performing.

The company has to focus on the significant variables to increase the lead conversion percentage.