SUMMARY

This analysis is done for X Education to find out the most promising leads (Industrial professionals) to get onboarded into their online courses. The original dataset provided with a plenty of information about the potential lead's origin, total no. of visit to their site, the time they spent on their website, how they reached to the site and the rate of conversion.

The following are the steps used to reach out to a conclusion:

1. **Data Cleaning:**

   After importing the dataset, we have found that there are few columns that contained more than 40% of missing values. We have dropped these columns. Then, we have observed that the option 'select' has no impact in this analysis hence, replaced with null values. We have seen many of the columns contains more than 95% of the data under 1 category, therefore dropped these columns.

2. **EDA:**

   We have done univariate analysis and outlier treatment to perform the EDA. We have imputed the columns having missing values more that 40% such as **a)** if it has outlier then replace missing values with median, **b)** if it has no outlier then replace with mean, **c)** if it contains categorical variable then replace with mode. We have removed the outliers which seems to be extreme.

3. **Data Preparation:**

   a) **Dummy variable creation:** The dummy variables were created for the categorical variables. Plotted heatmap to find out the multicollinearity among these variables. The columns having multicollinearity has been removed to get a clean data for model buildings.

   b) **Test-Train Split:** We have splitted the dataset into 70% train data and 30% test data respectively.

   c) **Scaling:** We have used StandardScaler for scaling the Numeric variables.

4. **Model Building:**

   Applied RFE to consider the top 20 relevant variables. Later, the rest of the variables were dropped depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). We have finalized Model 6 for the further evaluation.

5. **Model Evaluation & Prediction:**

   We have made a confusion matrix. Prediction was based on test data frame with an optimal tradeoff based on Recall & Precision as 0.34. At this threshold point, Recall is found to be 85% and AUC score is 0.81. As AUC score is more than 0.80 we can consider this as a good model. Difference between Accuracy on train and test data was also very less.

6. **Actionable Insights:** We have found the following features most significant in identifying the 'hot leads' -

- Current Occupation
- Last Activity
- Last Notable Activity
- Lead Origin
- Lead Source
- Specialization
- Total Time Spent
- Do Not Email
- Freebee Received