SUMMARY

The objective of this analysis is to select the most promising leads, i.e., the leads that are most likely to convert into paying customers, also known as 'Hot Leads' for X Education. If we successfully identify this set of leads, the lead conversion rate would go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. Need to build a model by assigning a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. Target is to get lead conversion rate to be around 80%.

The Following steps are used:

- **Cleaning data:** The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies.

- **Dummy Variables:** The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

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

- **Model Building:** Using Stats model Added the constant Variable from dataset. 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:** 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.