SUMMARY

The following steps have been taken for the case study assignment:

1. 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 because it provided insufficient information. To avoid losing too much data, a few null values were changed to 'not provided'. However, they were eventually eliminated while creating dummies. Because there were many Indians and few outsiders, the elements were altered to 'India', 'Outside India', and 'not provided'.
2. EDA: A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.
3. 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.
4. Train-Test split: The split was done at 70% and 30% for train and test data respectively.
5. 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).
6. Model Evaluation: A confusion matrix was made. Later on, the optimum cut-off value (using the ROC curve) was used to find the accuracy, sensitivity, and specificity which came to be around 80% each.
7. Prediction: Prediction was done on the test data frame and with an optimum cut of 0.35 with accuracy, sensitivity, and specificity of 80%. 8. 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.

Some of the Basic Learnings from Model Building:

1. A logistic regression model was used in the lead scoring case study to suit corporate needs.
2. There are many leads in the early stages, but only a handful of them turn into paying customers.
The majority of leads come from INDIA, with Mumbai having the largest number by city.
3. Some columns display 'Select', indicating that the student did not select the option for that column. Compulsory selection is necessary to collect useful data. Similarly, customer occupation, specialty, etc.
4. Increased total visits and time spent on the platform may lead to higher conversion rates.
5. The leads enrolled in courses to advance their careers, with a focus on finance management. Leads from HR, Finance, and Marketing management specializations are very likely to convert.

6. Improving client engagement through email and phone calls can increase lead conversion rates. Leads who open emails are more likely to convert. Sending SMS might also be beneficial.
7. The majority of leads' current occupation is unemployed, thus we focused more on unemployed leads.