Lead Scoring – Summary

Analysis is carried out for X Education in an effort to attract more business professionals to their courses. We learned a lot from the fundamental data on how potential customers use the site, how long they stay there, how they got there, and the conversion rate. The procedures are as follows:

1. Cleaning data: Except for a few null values, the data was mostly clean. However, the option choose had to be changed to a null value because it provided little useful information. To avoid losing too much data, only a small number of the null values were changed to 'not provided'. Nevertheless, they were later taken out while manufacturing dummies. The elements were altered to "India," "Outside India," and "not provided" because there were a lot of people from India and a small number from elsewhere.

2. EDA: To quickly assess the state of our data, an EDA was performed. It was discovered that several of the categorical variables' components were unnecessary. The numerical figures are accurate, and no anomalies were discovered.

3. Dummy Variables: The fake variables were made, then later the fakes with the 'not provided' bits were taken away. We utilized the MinMaxScaler to scale numerical numbers.

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 top 15 important factors. Later, the remaining variables were manually deleted based on their VIF values and p-value (the variables with VIF 5 and p-value 0.05 were maintained).

6. Model Evaluation: A confusion matrix was created. Later, the ideal cut off value (using the ROC curve) was utilized to determine the accuracy, sensitivity, and specificity, which came to be about 80% each.

7. Prediction: Prediction was performed on the test data frame with an ideal cut off of 0.35 and 80% accuracy, sensitivity, and specificity.

8. Precision – Recall: This approach was also used to retest, and a cut off of 0.41 was discovered on the test data frame, with precision around 73% and recall around 75%.