Summary

This analysis is done for X Education which sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The following are the steps used:

1. Cleaning Data:
   Columns which are having more than 40% null values are dropped, some of the missing values like country are replaced with most repeating value i.e., India which is nearly 96.9% of the data
2. EDA:
   A quick EDA is done on the data and found that there are not many outliers were found on the numeric values
3. Dummy Variables:
   The dummy variables were created and later on the dummies with 'Not Specified' elements were removed. For numeric values we used the MinMaxScaler.
4. Train test split:
   The data is split into 70% and 30% for train and test data respectively
5. Model Building:
   Firstly, RFE was done to attain the top 18 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 ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 93% each.
7. Prediction:
   Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 90%.