STEPS FOLLOWED FOR PERFORMING ANALYSIS:

The problem statement given by company X Education to find ways to get more industry professionals to join their courses. The prime goal is to increase the probability and chances of converting a customer into a potential lead. The dataset provided gave us a lot of information about people's behaviour such as the time spent by a user, actions performed when they visit the site, how they reached the site, city from which most number of people come to the site and the conversion rate.

The following are the steps used:
1. Data Cleaning : After checking shape, summary and info, I checked for null values and the percentage. I dropped columns with missing values which were more than 40%.
   - 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.
   - Also, I checked for Duplicate values but there were none.
   - I also categorised columns whose proportion was not significant as 'Others'
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 seems good and no outliers were found.
   - I also dropped values which were not contributing much in analysis.
3. Dummy Variables: The dummy variables were created and later on the dummies with 'not provided' elements were removed.
   - It also helped to categorise data in a better way.
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 most 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 80% 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 80%.
   - Confusion matrix also helped to measure the accuracy of our predictions.

It was found that the variables that there are few attributes which helped in finding hot leads or potential clients:

1- SMS sent to clients helped
2- The total time spend on the Website. More the time spent, more is the likeliness of becoming a hot lead.
3. Total number of visits.
4. Chances of getting a lead if the lead source was: a. Google b. Direct traffic
5. Olark chat conversation
5. When the lead origin is Lead add format.

Apart from following the patterns, the company should focus more on attributes like customer's time spent on website or lead source and also lead origin. Moreover, the company should try to stay in touch by SMS or Email with the client to give them updates and a good customer experience.