SUMMARY

After carefully analysing the data of 'X Education' the categorical variables had level 'Select' which was extrapolated as another level/category 'others'. After a neat cleaning of data the categorical variables were converted to dummy variables for analysis. Further, the dataset was split into train and test sets as per industry norm(i.e. 70%, 30%) and a model was fit on the basis of training set. After considering the correlation, significant p-values and VIF the model was rebuilt 3 times to have 16 variables(13 categorical & 3 continuous). Therefore the model built can be represented in the below equation

***"conversion** = -3.0759 + (7.0659)TotalVisits + (4.6414)Total Time Spent on Website - (3.1383)Page Views Per Visit + (1.9399)Lead Origin_Lead Add Form + (2.4376)Lead Source_Welingak Website - (1.2273)Do Not Email_Yes + (1.2328)Last Activity_SMS Sent + (1.1925)Country_others + (2.2323)What is your current occupation_Working Professional + (0.9368)What matters most to you in choosing a course_Better Career Prospects + (2.4186)Lead Profile_Lateral Student + (1.4690)Lead Profile_Potential Lead - (2.3797)Lead Profile_Student of SomeSchool + (3.1056)Last Notable Activity_Had a Phone Conversation - (0.9090)Last Notable Activity_Modified + (1.3195)Last Notable Activity_Unreachable"*

This is the model which can rate a lead from 0-100 given that probability score should be fed to classify lead as Hot & Cold. Having the question of optimal cut off in mind it was discovered to be 0.35 (or 35) meaning that a lead with more than 0.35 probability or a score more than 35 can be Hot lead (potential lead) and below it to be Cold lead.

Now after the train data is validated with Accuracy, Sensitivity and Specificity, the same needs to be carried on Test data. After a detailed validation it was found that the trained model can even perform well on test data meaning that model has a good capacity to generate accurate output on untrained data.

Of all the variables in the model the variables *TotalVisits, Total Time Spent on Website, Lead Source_Welingak Website, What is your current occupation_working Professional, Lead Profile_Lateral Student, Last Notable Activity_Had a Phone Conversation* seems to dominate other variables in the model as it evident that they have quite significant contribution as compared to other variables. In this perspective we can classify a lead as potential lead if they have visited the website, spent significant time on website by exploring different webpages, who is currently a working professional or a lateral student (as they might be in search of job hunt or job change) and had a phone conversation with some sales executive(might be curious about other offering or has queries with current offering). The models succeeds in predicting the interests of lead which is a very good sign of conversion and can help to focus on specific leads.