SUMMARY

The case study is about an education company named X Education which sells online courses to industry professionals. The people who browse for the courses might fill up a form for the course providing their email address or phone number, they are classified to be a lead. But the lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. The company want to select the most promising leads. So a model was build wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

As a first step the data was inspected and was found with many missing values. Columns with more than 30% null values were removed. Data of many columns contained zero variance variables i.e., most of the values in a column was same, so such columns were also dropped as they are of no use to the analysis. Remaining null values were handled by removing such rows. EDA results showed that the lead sources are Google, Direct Traffic and Olark Chat. Most of the people who search for such courses are the unemployed people and the last notable activity mostly performed by the students are sending sms, opening email and modifying. Model was build using RFE with 20 variables. The data was split to 70% train data and 30% test data. Using statsmodel the model build was assessed. The p value and VIF value was checked for eliminating or adding the variable, p value should be below 0.05 and VIF below 5. The final model contained 13 variables. Predictions were done on the trained data and lead score was generated and a random cut-off of 0.5 was selected for predicting the conversion. The Accuracy was found to be 79.11% for train data. ROC curve was plotted and the model was found to be good. The optimal cut-off value from accuracy, sensitivity, specificity graph and the precision recall curve was found to be 0.44. Predictions were made on test set and overall Accuracy was 78.66% for test data which is around 80%.

Lead Origin_Lead Add Form, Last Activity_Had a Phone Conversation, Last Notable Activity_Unreachable, Lead Source_Welingak Website, Lead Source_Olark Chat, Total Time Spent on Website, Last Activity_SMS Sent, Specialization_Banking, Investment And Insurance were found to be the factors which positively affects the lead conversion. Last Activity_Olark Chat Conversation, Last Notable Activity_Modified, Do Not Email_Yes, What is your current occupation_Student, What is your current occupation_Unemployed negatively impact lead conversion. So the positive parameters should be given importance and if the company want more customers to be hot lead the cutoff of 0.44 can be further reduced.