Summary

- X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor at around 30%.To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- CEO's target for lead conversion rate is around 80%.
- The target variable is the column converted which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
- The company requires you to build a logistic regression model.

EDA

- The first step is Data cleaning - There are 9000+ rows in the datapoints in the dataframe. Delete the Columns where there are more than 3000 missing values .**Leads Profile** and **How did you hear about X Education** columns have the high number of Select

labels in their column , it is better to drop these columns. We see Mumbai has the highest number of Leads. However this column does not going to help us in the analysis. It will be better if we drop it so dropping the city column. Most number of leads is from country India. The column **'What is your current occupation'** has lots of null values. We can drop the entire column but most of the columns are already dropped, also this column can be significantly used in the analysis. So It is better to just drop the null rows for this column.

- Columns 'Prospect ID' and 'Lead Number' does not have any impact on the dataframe. So, We can drop the columns.
- We noticed that when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque.
- The Column **'What matters most to you in choosing a course'** has one level 'Better career Prospects' which consists of 6528 values and other level values are 2 and 1. So, It is better to drop this column as it won't help in our model.

- We have around 69% of the rows that is retained .

Model Building

- The VIFs are less than 5. Let’s drop the ones with the high p-values beginning with **Last Notable Activity_Had a Phone Conversation**
- Model evaluation -optimal cutoff is 0.42 and ROC curve (area=0.86)
- Calculating precision TP/(TP+FP) = 0.76
- Calculating recall TP/(TP+FN)=0.78
- 'TotalVisits' , 'Total time spent on site' , 'Pageviews per visit' which contribute the most to lead converting.