# Summary

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

The data provided contains the information about how the potential customers visit the site, the time they spend in the site, how they reached the site and also the conversion rate for these customers.

The steps to reach the maximum conversion rate are as follows:

1. *Cleaning Data:*
   The data was partially cleaned as it contained nulls values. The value 'Select' had to be replaced by nulls as it did not add up for any information. For columns having more than 50% , they were dropped. Highly skewed columns were dropped because they don't add any information to the model. Categorical columns having values with very low frequencies (lower than 1% of total) were identified and combined into "Others". Imputations were performed for the columns based on statistical analysis and business understanding. After dropping rows to get rid of null values, we still retained 98% of rows.

2. *EDA:*
   An EDA was done to check the data. Several inferences were drawn from plots and documented. There were no outliers found in the numerical columns.

3. *Dummy Variables:*
   Dummy variables were created for the categorical variables and the original columns were removed from the dataset after concatenating the dummy dataframe with original dataframe. Standard Scaler was used for scaling the numerical variables.

4. *Train-Test Split:*
   The split was done at 70% and 30% for train and test data respectively.

5. *Model Building:*
   At first Feature Selection was done using RFE to attain the top 20 variables. Later the rest of the variables were removed manually depending on the p-value and the VIFs. The p-value less than 0.05 were kept and the variables with VIF<5. The model produced probability of whether a lead would convert.

6. *Model Evaluation:*
   A confusion matrix was made. The optimum cut-off (using ROC curve and visualizing accuracy/sensitivity/specificity values obtained from different cut-offs) was obtained as **0.3** and used to calculate the accuracy, sensitivity and specificity values as **79%, 83.7%** and **76.2%.**
   With cut-off set as **0.35** we could obtain an increased accuracy of **80%** but sensitivity decreased to **80%** too.

7. *Prediction:*
   Prediction was done on the test data and using cut-off as **0.3,** accuracy, sensitivity and specificity was found to be **79.25%, 85.34%** and **75.6%**, indicating a stable model.

The most important columns from the data set turn out to be as follows:

- Lead Origin
- Lead Source
- Time Spent on Website
- Occupation
- Last Activity
- Last Notable Activity

From the features in the model, following recommendations can be drawn:

- Leads originating from “Lead Add Form” and the “Welingak Website” can be classified as hot.
- Working Professionals should be targeted heavily.
- Leads with last notable activity as SMS sent are promising.
- Total time spent on the company’s website is also directly proportional to probability of conversion.
- Leads where last activity is Email Bounced and where the Specialization isn’t listed (assumed as ‘Others’ in the model) would have a low conversion probability.