# Summary

This analysis is done on an education company named X Education which sells online courses to industry professionals, to find ways to get more industry professionals to join their courses. Our aim of this analysis to identify the “Hot Leads” so that Marketing Team can focus more on potential lead rather than communicating to all the leads.

**1. Data Preparation:**

Single value columns are dropped as they don’t add value to analysis. Data frame contains some values as 'Select', these are replaced with null. Columns having more than 35% null values were dropped. We did check for duplicates rows. There are columns having many categories with very less value, a new value 'Other' is imputed for them. We checked for data imbalance and removed many columns which are highly skewed.

**2. EDA:**

A quick EDA was done to check the condition of our data. While doing the univariate analysis for categorical column we have got below observation:

- ✓ API & Landing Page Submission are two major contributors of Lead Origin.
- ✓ Direct Traffic and Google are the two main source of Leads.
- ✓ Email Opened and SMS Sent are the major Last Activity.
- ✓ Most of the lead generated by Unemployed.
- ✓ Majority don't want a free copy of Mastering The Interview.

Based on the univariate analysis of continuous column, below is our observation:

- ✓ None of the continuous variables are normally distributed.
- ✓ Total visits values are between 0-17, Total Time Spent on Website values are between 0-2500 and Page Views Per Visits values are between 0-10

We did the Bivariate Analysis With respect to target column 'Converted' and below is our observation:

- ✓ Lead Origin : Hot leads are more in Landing Page Submission, API and Lead Add Form.
- ✓ Lead Source: Hot leads are higher in Direct Traffic and Google.
- ✓ Last Activity: Hot leads are higher in SMS Sent and EMAIL Opened.
- ✓ What is your current occupation: Hot leads are mostly generated by Unemployed and Working Professional.
- ✓ A free copy of Mastering The Interview: Hot leads are more with answer No.
- ✓ Last Notable Activity: Similar to Last Activity.

**3. Dummy Variables:**

Created the dummy variable for categorical column and dropped one selected level.

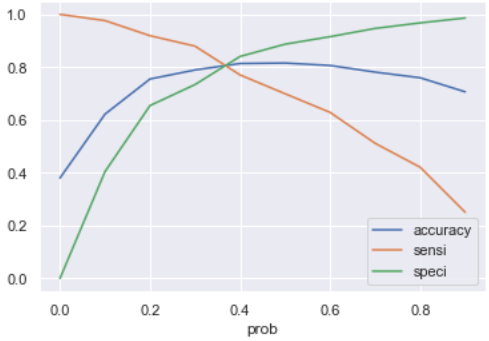
4. **Train-Test split:**

The split was done at 70% and 30% for train and test data respectively.

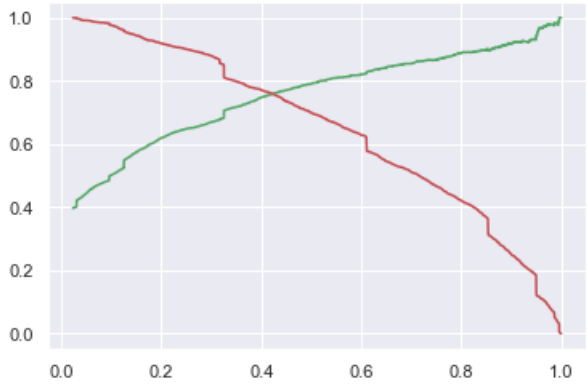
5. **Model Building:**

Model build with RFE and removing variables manually depending on the VIF-values and p-value (Variables with VIF<5 and p-value<0.05 were kept).

6. **Model Evaluation:**
A confusion matrix was made. Overall accuracy came 81.57%. We checked the sensitivity and specificity. We plotted the ROC curve to get the optimal cut-off which is 0.32.




**7. Precision – Recall**
Precision score was 0.680 and recall score was 0.854. Our goal was to build a model with >80% success lead (Hot Lead), hence we found right model based on training data.




- From the above curve we can see that precision & recall intersects at 0.41
- From the graph it is also clear that for having Recall >= 80% we have to keep cutoff <=0.32

**8. Prediction:**
Prediction was done on the test data frame and with 78.3% accuracy. the sensitivity of our logistic regression model on test dataset is 84.8% and the specificity is 74.5%, which is inlined with train data.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

- ✓ Total Time Spent on Website.

- ✓ When 'Lead Origin' is:
  - Lead Add Form
- ✓ When 'What is your current occupation' is:
  - Working Professional
- ✓ When 'Lead Source' is:
  - Olark Chat
- ✓ Total Visits.
- ✓ When 'Last Activity' is:
  - SMS Sent

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.