# 1. Problem Statement & Analysis:

- An education company X sells online courses to various professionals and students through their website.

- They market these courses via different platforms. Based on the leads generated via these platforms, their marketing team pursues them individually.

- Our task at hand is to identify the specific factors that contributes significantly for a lead to get converted.

- We need to build a model and train it on the existing data to classify a lead as hot lead based on the conversion probability Their typical lead conversion has been around 30%.

- We need to build a logistic regression model wherein 80% of the leads that could get converted have to be classified as hot leads as per our model.

# 2. Approach:

- Our analysis and building of the model involved the following steps.
    - Data Cleaning
    - Data Preparation
    - Train_Test_Split
    - Correlation, Recursive feature Elimination, VIF

## a) Data Cleaning:

We have performed data cleaning after treating the missing values and finally obtained 9074 rows and 15 columns.

## b) Data Preparation:

We removed the variables which are heavily skewed and modified the categories of few variables to suit the analysis

## c) Train-test split:

- We have divided the dataset to dependent and independent variables by creating data sets X and y
- We have then divided th dataset to train and test with 70% of available data to be used for training the model.

## d) Correlation, Recursive feature Elimination, VIF:

- Using the RFE library from scikit learn, we have obtained the top 15 significant variables for building the model.
- After creating a logistic regression and making sure that all the varibales are significant without multi-collinearilty, we obtained the following final model.

## e) Final Model:

- **Final model consists of following variables that are significant.**

| | Coefficient | Std err | z | P-value | 2.5%'le | 97.5%'le |
|---|---|---|---|---|---|---|
| **const** | -0.9347 | 0.124 | -7.528 | 0.000 | -1.178 | -0.691 |
| **Do Not Email** | -1.4107 | 0.169 | -8.339 | 0.000 | -1.742 | -1.079 |
| **Total Time Spent on Website** | 1.1455 | 0.040 | 28.726 | 0.000 | 1.067 | 1.224 |
| **Lead Origin_Landing Page Submission** | -0.2049 | 0.091 | -2.241 | 0.025 | -0.384 | -0.026 |
| **Lead Origin_Lead Add Form** | 4.3319 | 0.226 | 19.204 | 0.000 | 3.890 | 4.774 |
| **Lead Origin_Lead Import** | 1.6907 | 0.448 | 3.777 | 0.000 | 0.813 | 2.568 |
| **Lead Source_Google** | 0.3644 | 0.080 | 4.539 | 0.000 | 0.207 | 0.522 |
| **Lead Source_Olark Chat** | 1.0229 | 0.127 | 8.082 | 0.000 | 0.775 | 1.271 |
| **Specialization_strong demand** | -0.1785 | 0.083 | -2.140 | 0.032 | -0.342 | -0.015 |
| **Tags_lost** | 1.0083 | 0.111 | 9.066 | 0.000 | 0.790 | 1.226 |
| **Tags_unresponsive** | -1.1399 | 0.152 | -7.509 | 0.000 | -1.438 | -0.842 |
| **Last Notable Activity_Modified** | -0.8424 | 0.086 | -9.792 | 0.000 | -1.011 | -0.674 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Last Notable Activity_Other activity** | -0.3013 | 0.129 | -2.344 | 0.019 | -0.553 | -0.049 |
| **Last Notable Activity_SMS Sent** | 1.5179 | 0.088 | 17.245 | 0.000 | 1.345 | 1.690 |

## 3. Learnings from the assignment

- We evaluated the model initially by taking a cutoff value of conversion probability at 0.5.
- Although we obtained an accuracy of 80%, our sensitivity is only 68%.
- This meant that there is a 32% chance of us missing out on a prospective lead that has the potential to get converted.
- While sensitivity can be increased by increasing the cutoff for probability of conversion, this might cause a decrease in accuracy.
- Based on our plotting of accuracy, sensitivity and specificity against various cutoffs for probability of conversion, we found that optimal cutoff lies between 0.3-0.35
- On the training set we have evaluated the model by taking a cutoff at 0.35. This has improved the sensitivity to 81%
- This meant that model can now predict 81% of all the lead that could actually get converted.
- On the test set we reduced the reduced the cutoff to 0.3 and obtained a sensitivity of 80% which is also decent.
- While sensitivity could be further increased by increasing cutoff, this could lead to decrease in specificity which leads to lower predictability of leads that will not be converted. This leads us to chasing after leads that will not get converted.