# ***SUMMARY***

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

**Approach in a nutshell:**

1) Importing Data, Inspecting the data frame.
    a. Initially, we have imported the required libraries, importing the csv file and inspected the data frame by looking into head, number of rows and columns, checked the column-wise info, checked the summary of numeric columns.
    b. Learnings: First look of data so that we will understand what exactly we need to do.
2) Data preparation: The data was partially clean except for a few null values and the option select had to be replaced with a null value since it didn't give us much information.
    a. Data quality check and handling null values
        i. Checked the null value percentage and dropped the columns with the more than 40% of null values and dealt with the null percentages which are less than 40% and greater than 2% by imputing the respective value (like median, mean and mode).
    b. Encoding: Converting data to prepare it for an algorithm and get a better prediction. Done with one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns.
3) EDA (Exploratory data analysis): A quick Exploratory data analysis was done to check the condition of our data. It was found that a lot of categorical variables were irrelevant. There are some outliers in the numerical columns and addressed with respect to capping.
    a. Univariate Analysis
    b. BI-Variate Analysis
    c. Outlier detection
    d. Checking data imbalance
4) Dummy variable creation: Created dummy variables for the category columns.
5) Test-Train Split: The split was done (70% for train and 30% for test) for the data.
6) Feature Scaling: Used Standard Scalar for feature scaling. Standard Scaler follows standard normal distribution. Therefore, it makes mean = 0 and scales the data to unit variance. This method removes the median and scales the data in the range between 1$^{st}$ quartile and 3$^{rd}$ quartile.
7) Looking at Correlations: Have seen the correlation and found that their correlations between variables and dropped highly correlated variables.
8) Model Buildings: Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (i.e. the variables with VIF < 5 and p-value < 0.05 were kept).
    a. Feature selection using RFE
    b. Improvising the model further inspecting adjusted R-Squared, VIF and p-value.
9) Model evaluation with different metrics: A confusion matrix was made. Later, the optimum cutoff (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to

be around 85%, 72% and 93% respectively. An ROC curve shows the trade off between sensitivity and specificity (increase in one will cause decrease in other)., The closer the curve follows the y-ais and then the top border of the ROC space, means area under the curve and the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space. i.e. reference line, means less area and less accurate is the test.

a. Accuracy
b. Specificity
c. Sensitivity
d. Precision
e. Recall

10) Conclusion:
    a. Here, the logistic regression model is used to predict the probability of conversion of a customer.
    b. The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.
    c. Optimum cut off is chosen to be 0.41 i.e. any lead with greater than 0.41 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.41 or less probability of converting is predicted as Cold Lead (customer will not convert)
    d. Features used in final model are:
        i. Do Not Email
        ii. Lead Origin_Lead Add Form
        iii. Last Activity Had a Phone Conversation
        iv. Last Activity_SMS Sent
        v. Last Activity_Unsubscribed
        vi. What is your current occupation_Unemployed
        vii. Tags_Busy
        viii. Tags_Closed by Horizzon
        ix. Tags_Lost to EINS
        x. Tags_Ringing
        xi. Tags_Will revert after reading the email
        xii. Last Notable Activity_Modified
        xiii. Last Notable Activity_Olark Chat Conversation
    e. The top three features in the final model are:
        i. 'Tags_Closed by Horizzon'
        ii. Tags_Lost to EINS',
        iii. 'Tags_Will revert after reading the email'

The final model has Sensitivity of 0.917, this means the model is able to predict 91.7% customers out of all the converted customers, (Positive conversion) correctly.