SUMMARY

- X education company is an EdTech company selling courses to working professionals.
- This analysis is done for X Education company and to find ways to get more industry professionals to join their courses.
- The industry professionals who browse the website and fill a form (providing email and phone number) are identified as leads.
- Although X education receives a lot of leads, it has extremely low rate of lead conversion. The conversion rate is bare minimum of 30%. The goal of building logistic regression model, is to identify the hot leads that is the potential professional who would convert into paying professional, so that the sales team of company focuses more on hot leads rather than talking to everyone.

DATA Provided: for the case study we were given an excel sheet containing information about past leads records. 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.

Conversion rate was our target variable for the analysis, as we want to study the pattern that what influences a lead to convert into paying customer.

STEPS followed:

1. **Cleaning data:** We analysed the data columns, the null values in data columns and cleaned the data set. checked the null % and handled the Nan values

we dropped the columns which had null values more than 50% which adds no values to model. Dropped the Asymmetric columns. We also checked null % for some columns and if they were found to be almost null we imputed those using mean.

Few of the null values were changed to ‘not known’ so as to not lose much data.

2.Data Visualization: To study trend univariate and bivariate analysis was done of categorical variables. plotted the histogram of all categorical variables. From the histogram we found countries/city didn't have a major influence over the data. Hence, we dropped city and country columns.

Visualising the histogram of the distribution of the categorical variables, from this it was found TAGS, LEAD PROFILE and LEAD QUALITY are dependent on CONVERTED, we put Not-Known to fill nan values.

We dropped the column with heavily single-value data, that data don't have much value. (Example: Newspaper Article, X Education Forums, Newspaper, Through Recommendations and so).

3. EDA: A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found. Heat Maps were plotted, categorical variables were encoded.

3. Dummy Variables: The dummy variables were created and later on the dummies with 'not provided' elements were removed.

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

5. Model Building: 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 (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Model Evaluation: A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity.

7. Prediction: Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 91.4%.

8. Precision – Recall: This method was also used to recheck and a cut off of 0.41 was found with Precision around 94.26% and recall around 81.65% on the test data frame.

Recommendations:

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.