Summary

This analysis was done for X Education for them to find ways to incorporate curious industry professionals to join their course. They wanted an analysis of the 'Leads' who would want to convert and start their journey with X Education.

The data provided gave us an insight of the potential leads like how much time they spent on the website and did they respond to our SMS or emails, etc.

These are the following steps used in the analysis:

- **Data Cleaning:** Upon receiving the data, we noticed that it was partially clean apart from a few exceptions like null values or a few columns containing empty spaces indicating that the customer or the potential lead did not choose to fill in the details. These empty spaces were replaced with 'not provided' so as to not remove the rows with the empty spaces. Another change that was done during visualizing the data where majority of the leads were from India. The leads were grouped into 'India', 'Outside India' and 'Not provided'.

- **EDA:** Exploratory Data Analysis revealed that there were no outliers as such in the numerical data where as there were a few irrelevant categorical data that was removed while RFE. The data was visualized and a few conclusions were made while performing Bivariate analysis such as, most of the leads who have converted are Indians.

- **Test-Train-Split:** Using the sklearn library, the dataset was split onto 70:30 ratio where, 70% was the train dataset and 30% is the test dataset.

- **Building the model:** Before splitting and building the Logistic Regression model, we created dummy variables for the categorical data and removed the 'not provided' ones. For the numerical variables, MinMaxScaler was used to normalize the values. Once the variables were normalized, RFE was used to select the features that were relevant. Additionally, the variables were selected based on their p-values and VIF. We set VIF to 0.5 and the p-value to be below 0.05.

- **Model Evaluation:** Confusion matrix and accuracy were used to evaluate the model. Later, ROC curve was built and based on that the specificity and sensitivity was checked.

- **Precision-Recall:** The method was used again in the end with a cutoff of 0.43 and the precision came out to be 74% and the recall was 75%.

- **Conclusion:** The following parameters are important to check if the client would convert or not:
    - Total time spent on the website by the Lead
    - If the Lead adds their information directly into the form
    - Through Direct traffic
    - Through Organic search
    - If the Lead is a working professional