Summary:

This Analysis is done for X education and to find ways to get more industry professionals to join their course. 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 conversion rate.

The following are the steps used:

1.**Cleaning data:**

The data was partially clean except for a few null values and the option select has to be replace with a null value since it did not give us much information. Few of the null values were changed to 'not provided' to not lose much. Although they were later removed while making dummies.

2. **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.

3. **Dummy Variables:**

The dummy variables were created and later 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:**

predict the probabilities on the train set and calculate the specificity, sensitivity, and accuracy. Finding the optimal cut off using ROC what we get AUC and recreate confusion matrix and recalculate accuracy, specificity, and sensitivity

Make predictions on the test set using 0.45 as the cut off and calculate the accuracy, sensitivity, and specificity

7. **Precession Recall view:**

build the training model using the precision-recall view and calculate precision and Recall.

8.**Conclusion**:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction. –
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.
- Also, the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set