Summary

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

The Following are the step used:

1. **Cleaning the data:**

   The data was partially clean except for few null values. So, we know the null values won't give us much information, so we need to convert into some meaningful for our analysis. Even for few analyses we need to mark as not provided so as to not lose much data. Even most of meaning less data are removed while making the dummies.

2. **EDA**

   The dataset was loaded into the python notebook and explored statistically and visually to get an idea of outliers, distribution of data, feature redundancies, and so on. Correlations between the variables were also identified using heatmap and pair plots. The redundant variables were removed.
   It was found that a lot of elements in categorical variables were irrelevant. The numeric values seem good and no outliers were found. It is understandable from the EDA that there are many elements that have very little data and so will be of less relevance to our analysis.

3. **Dummy Variables:**

   The dummy variables were created and later on the dummies with more null values are removed. The dummy variables taken into consideration are 'Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call', 'Last Activity', 'Country', 'Specialization', 'What is your current occupation', 'What matters most to you in choosing a course', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'A free copy of Mastering The Interview', 'Last Notable Activity'.

4. **Train-Test Split:**

   After removing the few columns of dummy variables with more null values, split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**

   After train and test split is competed RFE was done to attain the top 15 relevant variables. Later with VIF<5 and p-value <0.05 were kept. Based on high p-value few variables were removed manually.

6. **Prediction for training dataset:**

Summary

After the model is trained, prediction is calculated with overall accuracy of 81% which is very good value. With the current cut off as 0.5 we have around 81% accuracy, sensitivity of around 70% and specificity of around 89%.

7. **Optimize cut off (ROC Curve):**

The ROC Curve is created area of its 0.89. which is good value. When plotting the accuracy, sensitivity, specificity, 0.3 is the optimum point to take it as cutoff probability.

8. **Prediction on testing dataset:**

The Final Predicted value for testing dataset is 79% with sensitivity of 91% and specificity of 71%, precision and recall for testing dataset are 67% and 91%.

9. **Conclusion and Observation of Model:**

1. Accuracy, Sensitivity and Specificity values of test set are around 79%, 91% and 71% which are approximately closer to the respective values calculated using trained set.

2. 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.

3. Hence overall this model seems to be good.

<u>Observation</u>

1. It was found that the variables that mattered the most in the potential buyers

a. The total time spend on the Website.

b. Total number of visits.

2. Current occupation is as a working professional, Direct traffic, Google, keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.