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 to us gave a lot of information about how the potential customers visit the site, the time they spend there and how they reached the site and the conversion rate.

The following are the steps used:

1. Cleaning Data: - First of all we started with cleaning the unnecessary data; for that we checked for null values present and then we went through 'Select' value .This 'Select' value had to be replaced with NaN value. Then we removed the unnecessary null values. After that we converted many 'Yes' and 'No' variables into 0 and 1 (binary form) as machine understands the binary languages.

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 latter dummies with 'not provided' elements were removed.

4. Train-Test Split: - The split was done on 70% of train dataset and 30% dataset.

5. Model Building: - First RFE was done on 15 variables and latter one by one variable are removed on the basis of p-values and VIF factors.

6. Predictions on train-test set:- We make a prediction on the train and test dataset .

7. Building ROC Curve: - An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

8. Model Evaluation: - Finding accuracy, sensitivity, specificity of predicted train and test dataset.

9. Precision and Recall**:-**

- **Precision = Also known as Positive Predictive Value, it refers to the percentage of the results which are relevant.**

- **Recall = Also known as Sensitivity , it refers to the percentage of total relevant results correctly classified by the algorithm.**

10. 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.

.