SUMMARY

This case study is about X Education which provide Online courses to the industry professional. The data gives us a lot of information about how the sales team reach out to these professional, and how the people visit X Education site and spend some time finding about the course.

The following steps are being followed:

- **Cleaning Data**:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

- **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.
Dummy Variables:
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.

- **Train-Test split & Scaling:**

From train and test data, the split was done at 70% and 30%, respectively. Standard scaling was performed to the variables "Total Visits," "Page Views Per Visit," and "Total Time Spent on Website". X and Y variables were created.

- **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).

- **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 which came to be around 80% each.
Prediction:
Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity around 80% to 81%.

- **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity around 80% to 81%.

- **Precision – Recall:**

This method was also used to recheck and a cut off of 0.41 was found with Precision around 75% and recall around 76% on the test data frame.

• Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1.The total time spend on the Website.
2.Total number of visits.
3.When the lead source was:
a) Google
b) Direct traffic
c) Organic search
d) Welingak website

4.When the last activity was:
a) SMS
b) Olark chat conversation

5.When the lead origin is Lead add format.

6.When their current occupation is as a working professional.

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.