Summary

A logistic Regression model built on the dataset given by 'X Education' which contains different parameters as to which how different clients reach out to their portal. This model will help 'X Education' to implement changes which will help them to identify 'hot leads' (i.e. leads which are more likely to convert) and increase the conversion rate more efficiently which in return will help grow business.

Steps used:

1. **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 'Others' so as to not lose much data. 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 on the dummies with 'not provided' elements were removed. For numeric values we used the StandardScaler.

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**: 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 90% each.

7. **Prediction**: Prediction was done on the test data frame and with an optimum cut off as 0.3 with accuracy, sensitivity and specificity of about 92%.

8. **Precision** – **Recall**: This method was also used to recheck and a cut off of 0.4 was found with Precision around 88% and recall around 92% on the test data frame.

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.