Summary

This analysis is done for X Education to find ways to get more industry professionals to join their courses. 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 the conversion rate. The following are the steps used:

1. **Cleaning data**:
   The data was partially clean except for a few null values and irrelevant values like Select. The value "**Select**" had to be replaced with a null value as the "**Select**" is as good as the null values as it doesn't hold any relevance to that column information. A few of the null values were dropped and most of the null values were imputed with the highest occurrence value in the data or categorized as others.
2. **EDA**:
   A quick EDA was done to check the condition of our data. It was found that a lot of variables have high imbalances eventually resulting in dropping as they don't contribute anything to our analysis. There are only 4 numerical variables in the data set on further analysis of the numeric variables has revealed that two variables have outliers and were capped at the 95th percentile mark.
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 MinMaxScaler.
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 the ROC curve) was used to find the accuracy, sensitivity, and specificity which came to be around 80% each.
7. **Prediction**:
   The prediction was done on the test data frame with an optimum cut of 0.35 with accuracy, sensitivity, and specificity of around 80%.
8. **Precision – Recall**:
   This method was also used to recheck and a cut-off of 0.35 was found with a Precision of around 76% and recall of around 78% on the test data frame. It was found that the variables that mattered the most in the potential buyers are:

- Total Time Spent on the Website
- When the lead source was:
  - Welingak Website.
  - Reference
  - Olark Chat

- When the last activity was:
    - SMS
    - Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

Keeping all these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their minds and buy their courses.

The end.