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 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 the option select had to be replaced with a null value since it did not give us much information. Variables with only one category are dropped. The Countries were clubbed into two categories India and Foreign. Likewise in some other similar feature categories are clubbed.

2. **Exploratory Data Analysis:**

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 the abnormalities were dealt by capping them. Both Univariate and Bivariate analysis gave good insights about the individual variables and the behavior of Target variable with other independent features.

3. **Dummy Variables:**

The dummy variables were created and later the variables for which the dummies are created, 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 20 relevant variables. Reason being some features are sure to be eliminated, and for a final model a maximum of 14 to 15 features are recommended. Later the rest of the variables were removed manually depending on the Variance Influence Factor(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 80% each. The cutoff value found by the Precision-Recall curve was nearly same as that by ROC curve.

7. **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.4 the accuracy, sensitivity and specificity which came to be around 80% each.

8. **Precision – Recall:**

This method was also used to recheck and a cut off of 0.4 was found with Precision around 76% and recall around 80% on the test data frame.

9. Lead Score:

The Lead Score column is created such that it represents the score of a Lead getting converted, higher the score better the probability. Also, a datum valuei.e. the median value of this column found out to be 33.0 is suggested as a reference value, leads above this value have higher possibility of getting enrolled.

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

1) The total time spent on the Website.
2) Lead Origin_Lead Add Form.
3) What is your current occupation_Working Professional
4) Lead Source_Referral Sites

Other than above Leads from India, leads who are students, leads who are approached by sending SMS. Leads with source Live chat and Referral sites, who had been referred. These leads are also a potential choice.

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