Summary

This analysis is carried out for X Education in an effort to attract more business professionals to their courses. We learned a lot from the fundamental data on how potential customers use the site, how long they stay there, how they got there, and the conversion rate.

The procedures are as follows:

1. Cleaning the data: The majority of the data was clean, save for a few null values, and the option "select" had to be changed to a null value because it provided little useful information. All columns with more than 35% null values have been removed, in case of columns where missing values are less, rows containing missing values have been dropped.
2. EDA: To assess the state of our data, a brief EDA was conducted. It was discovered that several of the categorical variables' components were unnecessary. The numerical figures had some outliers so they were imputed properly.
3. Dummy variables:

For category variables with more than two values, dummy variables were made, and the actual variables were subsequently deleted. We utilised the MinMaxScaler to scale numerical numbers.

4. Train-Test split:

For train and test data, the split was done at 70% and 30%, respectively.

5. Model construction:

First, the top 15 pertinent factors were determined by RFE. Later, based on the VIF values and p-value, the remaining variables were manually deleted (the variables with VIF 5 and p-value 0.05 were retained).

6.Model Evaluation:

A confusion matrix was created to evaluate the model. Later, the best cutoff value was determined (using a ROC curve), and the train dataset's accuracy, sensitivity, and specificity were each found to be around 80%.

7. Prediction: Using an optimal cutoff of 0.3, a prediction was made on the test data frame with 78% accuracy, 82% sensitivity, and 76% specificity.

8. Precision - Recall

This method was also utilized to perform a second check, and on the test data frame, a cut off of 0.36 was discovered with precision around 68% and recall around 78%.

9. Conclusion and Recommendation: It was found that the variables that mattered the most in the potential buyers are (In descending order) :

-Total Time Spent on Website

-In Lead Origin (Lead Add Form)

-In Lead Source (Olark Chat)

-TotalVisits

-In Last Notable Activity (Email Link Clicked)

-In Last Notable Activity (Olark Chat Conversation)

-In Last Notable Activity (Modified)

-Do Not Email

-In Last Activity (Olark Chat Conversation)

-In Last Notable Activity (Email Opened)

-In Last Notable Activity (Page Visited On Website)

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.