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 the how the potential customers visit the site and through which social media platform or browser , the time they spend and the conversion rate.

Below are steps involved to arrive a conclusion for the above-mentioned questions:

1. Cleaning Data:

The select option was replaced with NAN values since it didn't give much information. The columns which are not required where dropped. Checking the missing value and removing the columns which have >40% missing values. Updating the missing values with mode values where ever its required so we don't loose the data. Renaming the columns for better visualization.

2.EDA

EDA was conducted to analysis the various categorical data and numeric data. Univariant and bivariant analysis was conducted to check the dependencies between the columns and find the outliers and plotted the correlation graph and pair graph.

3.Dummy Variable:

The dummy variable was created and later on the dummies with no provided elements were removed. For numeric values we used MINMAXScaler

4.Train-Test Split:

The split was done at 70% and 30% for train and test data respectively.

5.Model Building:

First the RFE was done to attain the top 15 relevant variable. Later the rest of the variable were removed manually depending on the VIF values and p-values (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 around 80% each.

7.Prediction:

Prediction was done on the rest of the test dataframes and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8.Precision-Recall:

This method was also used to recheck and a cut off of 0.41 was found with Precision around 71.76% and recall around 78.61% on test dataframes.

It was found that the variables that mattered the most in the potential buyers are:

The total time spent on the Website.

When the lead source was: a. Google b. Direct traffic c. Organic search d. Welingak website e. Referral Site

When the lead origin was: a. Lead Add Form b. Student of Some School

When their current occupation is: 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