

Summary

This analysis is done for X Education to find ways to identify the most promising leads to join their course. The basic information 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 steps were used in the analysis:

1. Data cleaning:

The data given was partially cleaned with null values and select values. The select values were replaced with the null values as they did not give any information. The null values in the Specialization column were replaced with the indicator: not provided as it was an important variable and had 36% missing data.

2. EDA:

A quick EDA was done to check the condition of the data. It was found that a lot of elements in the categorical variables were found irrelevant. The numeric values seemed good and no outliers were found.

3. Dummy variables:

The dummy variables were created for categorical variables. The numerical variables were standardised using the standard scaler.

4. Train- Test split:

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

5. Model building:

Firstly, RFE was done to attain 20 variables. Later the insignificant variables were removed based on p-values and VIF values. The variables with p-values < 0.05 and $VIF < 5$ were retained in the model.

6. Model Evaluation:

A confusion matrix was made. Later the optimum cut-off value (based on the ROC curve) was used to find the Accuracy, sensitivity, and specificity which came to around 80% for each of the metrics.

7. Prediction:

The prediction was done on the test data frame with an optimum cut-off of 3.5 with accuracy, sensitivity, and specificity of about 80%.

8. Precision-Recall:

This method was used to recheck the model performance. A cut-off of 4.1 with a Precision of around 78% and a Recall of around 70% was found.

It was found that the following are the important features that help in predicting the most promising leads which X Education can focus on to bring in more customers to buy their courses.

- Total time spent on a website.
- When the lead source was: Google, Direct traffic, Organic search, Welingak website
- When last activity was: SMS, Olarck chat conversation
- What is your current occupation_ working professional