

Summary

The following are the steps used in order to do the analysis for the company X Education and to find more customers to join their course .

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. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

2. **EDA:**

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 seems good and no outliers were found.

3. **Data Preparation:**

The dummy variables were created and later on the dummies with 'not provided' elements were removed and dropped the repeated variables.

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 $p\text{-value} < 0.05$ 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.

7. **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 91%.