

Summary

This analysis was done for X Education and found ways to get more industry professionals to join their studies. The basic data provided gives us a lot of information about how customers are likely to visit the site, the time they spend there, how to access the site and the conversion rate.

The following are the steps used:

- 1. Clean data:** The data was partially clean except for a few empty values and the selected option had to be changed to a blank value as it did not provide us with much information. A few empty values have been changed to 'not provided' so as not to lose too much data. Although later released while making dummies. With more Indians and fewer expatriates, the elements were transformed into 'India', 'Outside of India' and 'non-India'.
- 2. EDA:** A quick EDA was created to check the status of our data. It was found that many of the variables of the categories were not significant. The numerical values look good and no outliers were found.
- 3. Dummy variables:** The dummy variant was created and later in dummies with 'not provided' elements were released. For numerical values we have used Min Max Scaler.
- 4. Separation of train inspections:** Separation was done on 70% and 30% of the train and inspection data respectively.
- 5. Model Structure:** First, RFE was designed for the acquisition of the top 15 variables. Later some variables were automatically removed depending on VIF values and p value (Variables with VIF <5 and p value <0.05 were saved).
- 6. Model Testing:** A confusion matrix is created. Later a large number of cuts (using the ROC curve) were used to obtain accuracy, sensitivity and specificity of about 80% each.
- 7. Predicting:** Predictions were made in the test data framework and were cut as completely as 0.35 with 80% accuracy, sensitivity and specificity.
- 8. Accuracy - Remember:** This method was also used for re-testing and a break of 0.41 was obtained with precision around 73% and recall about 75% in the test data frame.

It has been found that the most important variables for potential buyers are (By default):

1. The total time spent on the Website.
2. The total number of visits.
3. When the leading source said:
 - Google
 - Direct traffic
 - Live search
 - Welingak website
4. When the maintenance work was:
 - SMS
 - Olark chat
5. When the lead source is lead add format.
6. When their current job is a working technology.

Keeping this in mind X Education can thrive as they have a very high chance of getting almost all potential buyers to change their mind and buy their courses.