

Summary

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The following are the steps used:

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'.

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 seem good and no outliers were found.

Dummy Variables:

The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

Train-Test split:

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

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).

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.

Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.37 with accuracy, sensitivity and specificity of 80%.

Precision – Recall:

This method was also used to recheck and a cut off of 0.37 was found with Precision around 78% and recall around 70% on the test data frame.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1. What is your current occupation of the customer or Whether he is employed or not.
2. The total time spend on the Website.
3. Total number of visits.
4. When the lead source was:
 - a. Google
 - b. Direct traffic
 - c. Organic search
 - d. Welingak website
4. When the last activity was
 - a. SMS
 - b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a 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.

It was found that the variables that mattered the most in the potential buyers are (Indescending order):

What is your current occupation of the customer or Wheather he is employed or not

The total time spend on the Website

Total number of visits.

When the lead source was:

Google

Direct traffic

Organic search

Welingak website

When the last activity was:

SMS

Olark chat conversation

When the lead origin is Lead add format.

When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to getalmost all the potential buyers to change their mind and buy their courses.