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 how the potential customers visit the site, the time they spend there, how they reached the site & the conversion rate.

The following are the steps used:

1. Cleaning data:

The data was partially clean except for a few null values. Check the unique values in the columns & update with Unknown as this information will be used for further analysis so there might be a chance of hampering the analysis if we drop these rows where the value is missing hence updating it with different values so it can be used in Analysis, & it will be easy to identify—the 31% which is huge in our opinion.

According to the values 71% of values are populated with "Better Career Prospects" & other 29% values are NaN then we can convert NaN to the same value as "Better Career Prospects" it will keep the data more aligned

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 seem good & no outliers were found.

3. Dummy Variables:

The dummy variables were created & later on the dummies with 'not provided' elements were removed. Created the binary map variables to handle Yes/No Values.

4. Train-Test split:

The split was done at 70% & 30% for train & 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 & p-value (The variables with VIF < 5 & 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) is 0.5 we have around 92% accuracy, 87% sensitivity & 96% specificity.

7. Prediction:

The prediction was done on the test data frame & with an optimum cut-off of 0.30 with accuracy, sensitivity & specificity of 92%.

8. Precision – Recall:

This method was also used to recheck & a cut-off of 0.30 was found with a Precision of around 88% & recall of around 91% on the test data frame.

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

- The total time spends on the Website
- Total number of visits
- When the lead source was:
    a. Olark Chat
    b. Welingak website
- When the last activity was:
    a. Converted to Lead
    b. Email Bounced
    c. Olark Chat Conversation
- When the lead origin is Lead add the format
- Lead Profile is a Student of Some school & Unknown
- Tags assigned to customers indicating the current status of the lead

The Model seems to predict the Conversion Rate very well & we should be able to give the CEO confidence in making good calls based on this model.