Summary

Analysis is performed for X Education company based on its past leads data and found ways to get more industry professionals to join their courses. The dataset provided gave us a lot of information about how the potentials customers visit the site, the time they spend over there, how they reached the site and the conversion rate.

Following are the technical steps used for building the model:

Data Cleaning:

- We chose to remove the redundant variables/features
- The data set was partially clean except for a few null values
- Option value found to be 'Select' was equal to null and had to replace with a null value
- Dropped the high percentage of Null values i.e more than 3000 null value leads of a column

Data Preparation:

- A quick EDA was done on numeric variables to check the condition of our data, few outliers were found but retained them to include in analysis
- Created dummy variables for all the left-over categorical variables

Scaling:

- Used Min Max scaler to scale the data for Continuous variables

Train-Test Split:

- The Split was done at 70% and 30% for train and test the data respectively

Model Building:

- RFE was used with 15 dependent variables. Irrelevant features were removed manually depending on the VIF values and P-values (The variables with VIF < 5 and p-value 0.05 were retained

Model Evaluation:

- A confusion matrix was made. Found optimum cut-off value by using ROC curve and Precision Recall curve.
- Found the accuracy, sensitivity and specificity which came to be around 80%.

Prediction:

- Prediction was done on the test data frame an optimum cut-off as 0.42 with Accuracy – 79%, Sensitivity – 77% and Specificity - 80%

Precision-Recall:

- The method was also used to recheck and a cut-off of 0.44 with precision 75%, Recall 76%

Conclusion:

We have noted that the variables that important the most in the potential buyers are:

- Total Visits
- Total Time Spent on Website
- Page Views Per Visit