SUMMARY

This Case Study is done for X Education to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

Understanding Data

Data is imported and understood. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. Which may or may not be useful in ultimately deciding whether a lead will be converted or not.

The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted

Data preparation

Unique values are checked, missing value are handled.Country,Specialization,What is your current occupation ,what matters most to you in choosing a course are important columns even Tho they have a high number of missing values thus they are kept.

Exploratory Data Analysis

During EDA many categorical values were found to be not useful and major there were no major outliers in the data.Page Views Per Visit and TotalVisits is moderately correlated other varibles were very weakly correlated.

Creating dummy variables for all categorical variables

Selected the columns that are of object data type and dummy variables were created. Dummy variables are used to replace the encoded categorical variables to some numeric values

Train Test Splitting

Converted column is made as the target variable and the 70/30 spilt is done on the data frame for training and testing the model

Model Creation

RFE was done with 15 variables as output at first and columns were dropped accordingly looking at the p- values and VIF.

Prediction and evaluation

Prediction was done on the testing data frame and the overall accuracy of the model was 80 %.The cut off was chosen as 0.5 and was optimized using ROC function. sensitivity and specificity were found to be about than 80 %. Precision and Recall was about 78 % and 68 % respectively.

Conclusion

Variables that matters are the most in the potential buyers are:

The total time spend on the Website.

Total number of visits.

When the lead source was: 1. Google ,2. Direct traffic, 3. Organic search ,4. Welingak website.

When the last activity was: 1. SMS ,2. Olark chat conversation

When the lead origin was Lead add format.

When their current occupation was as a working professional.