Lead Scoring Case Study: Summary Report

This analysis is done for X Education, an education company who sells online courses to industry professionals. The company is facing problems in lead conversion, typically the ratio is around 30%. The company has the goal to increase the lead conversion to be around 80% and make this process more efficient.

We have been provided with a leads dataset from the past with around 9000 datapoints. 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.

The following are the steps we have followed in the analysis and model building process:

1. Cleaning the data: The data was partially clean except for a null value and the option Select has to be replaced with a null value sine it does not give us much information. Few of the null values were changed into Not provided so as to not loose much data, however they were removed later while making dummies. Since there were many from India and few from outside, the elements were changed to India, Outside India and Not Provided.
2. EDA: A quick EDA was done to check the conditions 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.

3. Creating Dummy Variables: The dummy variables were created and later on the dummies with 'not pr/voided elements were provided. For numeric values we used the Minmax scaler.

4. Train and Test Split: The split was done at 70% and 30% for train and 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 and p-value (The variables with VIF<5 and p-value <0.5 were kept).

6. 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 around 80% each.

7. Prediction: Prediction was done on the test data frame with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. Precision-Recall: This method was also used to recheck and a cut off of 0.41 was found with Precision around 73.24% and recall around 76.61% on the test data frame.

Conclusions

The variables that are important in the potential buyers are:

1. TotalVisits
2. The total time spend on the Website.
3. Lead Origin_Lead Add Form

4. Lead Source_Direct Traffic
5. Lead Source_Google
6. Lead Source_Welingak Website
7. Lead Source_Organic Search
8. Lead Source_Referral Sites
9. Lead Source_Welingak Website
10. Do Not Email_Yes
11. Last Activity_Email Bounced
12. Last Activity_Olark Chat ConversationWhen the lead origin is Lead add Form

Keeping the above-mentioned points in mind the X education can increase all the potential buyers to change their mind and buy their courses.

Things we learnt while doing this assignment are how to prepare data for Model building and evaluation, business understanding, Presentation and Recommendations for a particular business problem.