Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step1: Reading and Understanding Data:

Firstly, we shall read all the attributes of dataset to understand it much better.

Step2: Data Cleaning:

a. First step to clean the dataset we chose was to find null values and sorting the values.

b. Then, we chose to drop those columns where missing values were more than 3000 in number. After rechecking we dropped off more unnecessary columns.

c. There were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to Null values.

d. Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step3: Dummy Variables Creation:

a. We created dummy variables for the categorical variables.

b. Removed all the repeated and redundant variables

c. Added the result to the master dataframe.

Step4: Preparing the data for modelling

Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step5: Feature Rescaling:

a. We used the Min Max Scaling to scale the original numerical variables.

b. Then, we plotted a heatmap to check the correlations among the variables.

c. Dropped the highly correlated dummy variables.

Step6: Model Building:

a. Using the Recursive Feature Elimination, we went ahead and selected the 15 important features.

b. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

c. Finally, we arrived at the 10 most significant variables. The VIF's for these variables were also found to be good.

d. For our final model we predicted the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.

e. We then created a data frame with the actual conversion flag and probabilities predicted by the model.

f. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 78.86%; Sensitivity= 73.94%; Specificity= 83.43%.

g. Then we plot the ROC curve and the area under the curve is 0.86, which is quite good.

h. Then we created a data frame to calculate the values of accuracy, sensitivity and specificity where cutoff came out as 0.44.

Step 6: Conclusion:

• Good value of sensitivity of our model will help to select the most promising leads.

• Features which contribute more towards the probability of a lead getting converted are:

i. Total visits

ii. Page views per visit.

iii. Total Time Spent on Website.