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: Read and inspected the data.

Step2: Data Cleaning:

a. Treatment of ‘Select’ values: Change ‘Select’ to NaN.
b. Handling Missing Values: Deletion of some columns, imputation in others.
c. Removed some unwanted columns with no value to model.
d. Dropped highly skewed categorical columns.
e. Handled outliers in numerical columns by capping them.

Step3: Data Transformation:

a. Changed the binary variables into ‘0’ and ‘1’.
b. Grouped some values of some categorical variables with low share to reduce number of created dummy variables.

Step4: EDA:

a. Checking Imbalance in Target Variable
b. Univariate Analysis
c. Bivariate Analysis

Step5: Data Preparation:

a. Dummy Variables Creation
b. Dropping original columns of the dummies.

Step6: Train-Test Split

Step7: Feature Scaling

Step8: Model Building:

a. Using the Recursive Feature Elimination, we went ahead and selected the 15 top 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 13 most significant variables. The VIF's for these variables were also found to be good.
d. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
e. We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 88% which further solidified the of the model.
f. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
g. 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 79.73%%; Sensitivity= 79.91%; Specificity= 79.61%.

Conclusion:

a. The lead score calculated in the test set of data shows the conversion rate of around 80% when lead score cutoff is set to 50.
b. Good value of sensitivity of our model will help to select the most promising leads.
c. Features which contribute more towards the probability of a lead getting converted are:
   i. Lead Source_Reference,
   ii. Last Activity_SMS Sent,
   iii. Current_occupation_Working Professional.