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 you 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 has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step1: Reading and Understanding Data:

Read, analyzed the data.

Step2: Data Cleaning:

a. Step one is to clean the dataset having unique values.
b. Then, there were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to Null values.
c. We dropped the columns having NULL values greater than 35%.
d. Next, we removed the imbalanced and redundant variables.
e. In the final solution, all sales team generated variables were removed to avoid any ambiguity.
f. Step 5 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. Also, in one column was having identical label in different cases (first letter small and capital respectively). We fixed this issue by converting the label with first letter in small case to upper case.

Step3: Data Transformation:

Changed the binary variables into '0' and '1'

Step4: Dummy Variables Creation:

a. We created dummy variables for the categorical variables.
b. Removed all the repeated and redundant variables.

Step5: Test Train Split:

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

Step6: Feature Rescaling:

a. Dropped the highly correlated dummy variables.

b. We used the Min Max Scaling to scale the original numerical variables.
c. Then, we plot a heatmap to check the correlations among the variables.

Step7: Model Building:

a. Using the Recursive Feature Elimination, we decided to select 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 11 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 with an area coverage of 86% which further solidified the of the model.
f. We based on the converted column checked if 80% cases are correctly predicted.
g. For our final model on train set, we checked the precision and recall with accuracy, sensitivity and specificity
h. Considering the Precision and Recall trade-off, we got a cut off value ofapproximately 0.3.
i. Based on the Sensitivity and Specificity metrics, we implemented the learnings to the test model and calculated the conversion probability and found out the accuracy value to be 77.52; Specificity= 74.13%; Sensitivity= 83.01%.

Step 8: 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. Lead Origin_Lead Add Form
    ii. What is your current occupation_Working Professional
    iii. Total Time Spent on Website
- The lead score calculated in the test set of data shows the conversion rate of 83% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.