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

Solution Summary:

Step1: Reading and Understanding Data:
Reading and inspecting the given data.
Step2: Data Cleaning:
a. The first step in cleaning up the selected data set was to remove variables with unique values.
b. There were several columns with the value 'Select'. This means that the prospect did not select a particular option. I changed these values to null values.
c. Removed columns with more than 35% NULL values.
d. We then removed the unbalanced and redundant variables. This step also included filling missing values with the median for numeric variables and creating new classification variables for categorical variables, if necessary. Outliers were identified and removed. Additionally, the columns had identical labels with different case (lowercase or uppercase first letter). Fixed this issue by converting lowercase first labels to uppercase.
e. All variables generated by the sales team have been removed to avoid ambiguity in the final solution

Step3: Data Transformation:
Changed binary variables into '0' and '1'

Step4: Dummy Variables Creation:
a. 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. We used the Min Max Scaling to scale the original numerical variables.

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

c. Dropped the highly correlated dummy variables.

Step7: 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 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 be pretty decent with an area coverage of 86% which further solidified the of the model.

f. Then, checked if 80% cases are correctly predicted based on the converted column.

g. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.

h. Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.3.

i. 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 77.52%%; Sensitivity= 83.01%; Specificity= 74.13%.

a. We used recursive feature elimination to select the 15 most important features.
b. Using the generated statistic, we recursively examined the P-values to select the most significant values that should be present and attempted to discard the less significant values.
c. Finally, we arrive at the 11 most important variables. The VIF for these variables turned out to be good too.
d. The final model was checked for optimal probability bounds by finding points and checking precision, sensitivity and specificity.
e. Then drew the ROC curve for the feature. This curve is pretty decent with 87% area coverage, further solidifying that of the model.
f. Now verify that 79% of the cases are correctly predicted based on the transformed columns.
g. Accuracy, sensitivity, and specificity of the final model on the training set were used to validate accuracy and recall
h. We then obtained a cutoff value of 0.3 based on the trade-off between precision and recall.
a. We then implemented the insight into a test model and calculated the conversion probabilities based on the sensitivity and specificity metrics and found an accuracy score of 79%. Sensitivity = 83%; Specificity = 75%.

Step 8: Conclusion:

• 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%.

• 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:

- Total Time spent on website
- Whether they submitted a form on website (Lead Origin - Add form)
- Whether an SMS was sent to the lead
- Whether the lead is a working professional (Current job - working professional)
- Whether the company had phone conversation as the most recent notable activity (Last Activity - had a phone conversation)