Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs assistance in identifying the most promising leads most likely to convert into paying clients.
The organization requires a model in which a lead score is allocated to each lead such that customers with higher lead scores have a better conversion probability and customers with lower lead scores have a lower conversion rate.
The CEO, in particular, has estimated the desired lead conversion rate to be about 80%.

Solution Summary:

Step1: Reading and Understanding Data:

The dataset was read and understood thoroughly by analysing the columns and shape of dataset.

Step2: Data Cleaning:

i. The first step we took to clean the dataset was to remove the variables with unique values.
ii. Then there were a few columns with the value 'Select', indicating that the leads had not chosen any of the options. We converted those values to null values.
iii. We removed the columns with NULL values more than 40%.
iv. We then deleted the variables that were unbalanced and duplicated. This process also involved replacing missing values with median values where necessary in numerical variables, as well as creating new categorization variables in categorical variables. The outliers were discovered and eliminated. Also, one column had the same label in various circumstances (initial letter tiny and capital). We resolved this issue by switching the label's initial letter from tiny to uppercase.
v. To eliminate ambiguity in the final solution, all sales team-generated variables were deleted.

Step3: Data Transformation:

The binary variables were transformed into ‘0’ and ‘1’.

Step4: Dummy Variables Creation:

i. We constructed dummy variables to represent the category variables and dropped the original columns which were responsible for making dummies.
ii. Removed all repetitive and superfluous variables.

Step5: Test Train Split:

The data set was then divided into test and training halves, with a 70-30% split.

Step6: Feature Rescaling:

i. We utilized the Min Max Scaling method to scale the original numerical variables.
ii. Then, we create a heatmap to check the relationships between the variables.
iii. Then removed the strongly linked dummy variables.

Step7: Model Building:

i. Using Recursive Feature Elimination, we picked the 15 most significant characteristics.
ii. Using the statistics generated, we recursively examined the P-values to identify the most significant values that should be there and eliminated insignificant values.
iii. Finally, we identified the 11 most important factors. VIFs for these variables were likewise determined to be satisfactory.
iv. For our final model, we determined the best probability cut-off by locating points and evaluating accuracy, sensitivity, and specificity.
v. We then plotted the ROC curve for the features, and the curve turned out to be rather excellent, with an area coverage of 86%, which reinforced the model.
vi. Then, using the converted column, determine if 80% of the situations were accurately anticipated.
vii. We tested our final model's precision and recall, as well as its accuracy, sensitivity, and specificity, on a train set.
viii. Next, using the Precision and Recall trade-off, we arrived at a cut-off value of about 0.5.
ix. Then we applied the learnings to the test model and estimated the conversion probability using the Sensitivity and Specificity metrics, resulting in an accuracy value of 80.4%; sensitivity of 80.4%; and specificity of 80.5%.

Step 8: Conclusion:

i. The lead score generated in the test set of data reveals a conversion rate of 83% on the final anticipated model, which clearly fulfils the CEO's estimate of an average lead conversion rate of roughly 80%.
ii. Our model's high sensitivity will aid in identifying the most promising leads.
iii. Features that add more to the likelihood of a lead being converted are:
    - Lead Origin_Lead Add
    - What is your current occupation_Working Professional
    - Last Notable Activity_Unreachable