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:

Read and inspect the data

Step2: Data Cleaning:

a. To begin cleaning the dataset we selected, we first dropped the variables that had unique values.
b. Next, we identified several columns containing the value 'Select', indicating that leads hadn't chosen any specific option. To address this, we replaced these instances with null values.
c. We proceeded by dropping columns where the proportion of NULL values exceeded 35%.
d. Following that, we addressed imbalanced and redundant variables. This involved imputing missing values where necessary, utilizing median values for numerical variables and creating new classification variables for categorical ones. Outliers were identified and removed. Additionally, we encountered a column with labels differing only in case (e.g., lowercase and uppercase). To resolve this, we standardized the labels by converting those with lowercase initial letters to uppercase.
e. To ensure clarity in the final solution, all variables generated by the sales team were removed from the dataset.

Step3: Data Transformation:

Changed the binary variables into ‘0’ and ‘1’

Step4: Dummy Variables Creation:

a. Dummy variables were created for the categorical variables to facilitate their inclusion in the analysis.
b. We eliminated any duplicated and redundant variables from the dataset.

Step5: Test Train Split:

Following that, we partitioned the dataset into training and testing sets, allocating 70% of the data for training and 30% for testing purposes.

Step6: Feature Rescaling:

a. We applied Min-Max Scaling to normalize the original numerical variables.
b. Next, we visualized the correlations among the variables by plotting a heatmap.
c. We removed highly correlated dummy variables from the dataset.

Step7: Model Building:

a. We employed Recursive Feature Elimination to select the 15 most important features.
b. Utilizing the generated statistics, we recursively examined the p-values to select the most significant variables, dropping those deemed insignificant.
c. Ultimately, we identified the 11 most significant variables. The Variance Inflation Factors (VIFs) for these variables were also deemed satisfactory.
d. For our final model, we determined the optimal probability cutoff by exploring various points and evaluating the accuracy, sensitivity, and specificity.
e. We proceeded to plot the ROC curve for the features, and the resulting curve demonstrated a high area under the curve (86%), reinforcing the robustness of the model.
f. Next, we verified whether 80% of the cases were accurately predicted based on the converted column.
g. We evaluated the precision and recall alongside accuracy, sensitivity, and specificity for our final model on the training set.
h. Next, considering the trade-off between Precision and Recall, we determined a cutoff value of approximately 0.3.
i. After applying the insights gained to the test model, we calculated the conversion probability based on the Sensitivity and Specificity metrics. The resulting accuracy value was 77.52%, with a Sensitivity of 83.01% and Specificity of 74.13%.

Step 8: Conclusion:

- The lead score computed in the test set of data indicates a conversion rate of 83% on the final predicted model. This outcome aligns closely with the CEO's expectation, as the target lead conversion rate was estimated to be around 80%.
- A high sensitivity value for our model is beneficial as it aids in identifying the most promising leads accurately.

- 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