Lead Scoring Case Study Summary

Problem Statement:

X Education, an online education company, seeks to boost its lead conversion rate by identifying and prioritizing "Hot Leads" with the highest potential to convert into paying customers.

They engage leads through calls and emails to nurture their interest. By implementing a lead scoring model, the company aims to assign scores to leads based on their likelihood of conversion.

The objective is to focus efforts on leads with higher scores, increasing the overall lead conversion rate. X Education targets an ambitious lead conversion rate of around 80%.

Solution Summary:

Step1: Reading and Understanding Data:

Read and inspected the data.

Step2: Data Cleaning:

- Dropped variables with unique values.

- Replaced the value 'Select' with null values, indicating that the leads did not choose any given option.

- Dropped columns with null values greater than 35%.

- Removed imbalanced and redundant variables.

- Imputed missing values with median values for numerical variables.

- Created new classification variables for categorical variables as needed.

- Identified and removed outliers.

- Fixed the issue of identical labels in different cases by converting the label with the first letter in lowercase to uppercase.

- Removed all sales team generated variables to avoid ambiguity in the final solution.

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. We used the Min Max Scaling to scale the original numerical variables. b. The, we plot a heat map to check the correlations among the variables. c. Dropped the highly correlated dummy variables.

Step7: Model Building:

a. Utilizing Recursive Feature Elimination, the top 15 important features were selected.

b. P-values were examined to identify the most significant values and drop insignificant ones.

c. Eventually, 11 variables were determined as the most significant, with good VIF values.

d. Optimal probability cutoff was determined by assessing accuracy, sensitivity, and specificity.

e. An ROC curve was plotted, resulting in a respectable area coverage of 86%, reinforcing the model's reliability.

f. The prediction accuracy for converted cases was verified to determine if it reached 80%.

g. Precision, recall, accuracy, sensitivity, and specificity were evaluated for the final model on the training set.

h. A cutoff value of approximately 0.3 was selected based on the trade-off between precision and recall.

i. The insights gained were then applied to the test model to calculate conversion probability using sensitivity and specificity metrics, resulting in an accuracy of 77.52%, sensitivity of 83.01%, and specificity of 74.13%.

Step 8: Conclusion:

- The lead score calculated in the test set demonstrates a conversion rate of 83% using the final predicted model, exceeding the CEO's expectation of a target lead conversion rate of around 80%.

- The model's high sensitivity value is beneficial for identifying the most promising leads.

- The features that contribute significantly to 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