Lead Score Case Study Summary

Problem Statement:

X Education sells online courses and needs help in identifying promising leads with a high likelihood of converting into paying customers. They require a model to assign lead scores based on conversion likelihood, with a target conversion rate of 80%.

Steps and methods used for creating the model for business solutions:

DATA READING AND UNDERSTANDING:

- At this stage, all the data that will undergo description and interpretation processes is gathered.

DATA CLEANING:

- We removed variables with high NULL percentages, replaced missing values with medians for numerical variables, created new classification variables for categorical ones, and removed outliers.

DATA ANALYSIS OR EXPLORATORY DATA ANALYSIS:

- It involves summarizing data features, detecting patterns, and uncovering relationships using visual and statistical techniques, which helps in gaining insights and formulating hypotheses for further analysis.

CREATING DUMMY VARIABLE AND TRAIN-TEST SPILT:

- We proceeded to generate dummy data for the categorical variables. The next step involved dividing the dataset into test and train sets with a 70-30% split.

FEATURE RESCALING AND FEATURE SELECTION USING RFE:

- We applied Min Max Scaling to scale the original numerical variables. Then, using the stats model, we constructed our initial model to obtain a comprehensive statistical view of all the parameters.
- We utilized the Recursive Feature Elimination method to choose the top 20 important features. We then used the generated statistics to iteratively examine the P-values, selecting the most significant values and discarding the insignificant ones.
- After analyzing our data, we identified 15 significant variables with good VIF scores. We then created a data frame with converted probability values, assuming that a value above 0.5 indicates 1, and anything below indicates 0. Based on this assumption, we used Confusion Metrics to calculate the overall Accuracy of the model. Additionally, we evaluated the model's reliability by determining the 'Sensitivity' and 'Specificity' matrices.

PLOTTING A ROC CURVE:

- We then attempted to plot the ROC curve for the features, and the curve turned out to be quite robust with an area under the curve of 86%, further strengthening the model's performance.

FINDING AN OPTIMAL CUT-OFF POINT:

- We first created graphs showing the probability of 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. We identified the optimal probability cutoff point as the intersection of these graphs, which turned out to be 0.42.
- With this new cutoff point, we found that the model correctly predicted close to 80% of the values.

COMPUTING THE PRECISION AND RECALL METRICS:

- We also discovered that the Precision and Recall metrics resulted in values of 79% and 70.5% respectively for the training dataset. Based on the Precision and Recall tradeoff, we obtained a cutoff value of approximately 0.42.

MAKING PREDICTIONS ON TEST SET:

- We applied the knowledge to the test model, calculated the conversion probability based on Sensitivity and Specificity metrics, and determined the accuracy to be 80.8% with Sensitivity at 78.5% and Specificity at 82.2%.

The most influential variables for potential buyers are:

- 1. Total time spent on the website
- 2. Total number of visits
- 3. Lead source: Google, Direct traffic, Organic search website
- 4. Last activity: SMS, Olark chat conversation
- 5. Lead Origin: Lead add format
- 6. Current occupation: working professional

X Education can flourish by leveraging these factors to persuade potential buyers to purchase their courses.