Summary: LEAD SCORING CASE STUDY

Done by: Sushmita Roy, Chetna Turankar and Utteya Pal

1) Problem statement:

X Education is trying to sell online courses to people working in different industries. They want assistance in picking out the best potential customers—those who are most likely to actually buy the courses. They're looking for a system that gives each potential customer a score. The idea is that the higher the score, the more likely the person is to buy a course, and the lower the score, the less likely. The CEO has mentioned they're aiming for about an 80% success rate in converting these potential customers into actual paying customers.

2) Solution summary:

Step 1: Reading and understanding data

I looked at the information and studied it.

Step 2: Data Cleaning:

We got rid of the information that was missing a lot of details. This involved filling in some of the missing information with typical values, like averages for numbers and creating new categories for certain types of information. We also found and got rid of any unusual or extreme values.

Step 3: Data Analysis

After that, we took a closer look at the data to understand its overall pattern. During this process, we noticed that about three aspects always had the same value in every row. We decided to remove these because they didn't provide useful information.

Step 4: Creating Dummy Variables

We made up some placeholder information for the categories in our data that aren't numbers.

Step 5: Test Train Split:

Afterwards, we split the data into two parts: one for testing and the other for training. We allocated 70% of the data for training and 30% for testing."

Step 6: Feature Rescaling

We adjusted the size of the original number values using a method called Min Max Scaling. After that, we used a statistical tool called stats model to build our first model. This model gives us a detailed statistical overview of all the factors in our analysis.

Step 7: Feature selection using RFE:

We used a method called Recursive Feature Elimination to pick out the 20 most important aspects in our data. Using the resulting stats, we kept refining our choices by looking at P-values, focusing on the most meaningful ones and discarding the less important ones. After this process, we ended up with 15 crucial factors. The VIF (Variance Inflation Factor) values for these factors were also good. Then, we created a table with probability values, assuming that if the probability is more than 0.5, it's considered as 1, otherwise 0. With this assumption, we analysed how well our model performed using metrics like Confusion Matrix, Accuracy, Sensitivity, and Specificity to see how reliable our model is.

Step 8: Plotting the ROC Curve

We also created a graph called the ROC curve to visualize how well our features perform. The curve looked good, covering an area of 89%, which made us more confident in the reliability of our model."

Step 9: Finding the Optimal Cutoff Point

Next, we created graphs to see how well our model predicts accuracy, sensitivity, and specificity at different probability values. The point where these graphs intersected was identified as the best probability cutoff point, and it turned out to be 0.37. With this new cutoff, we noticed that nearly 80% of the predictions made by the model were accurate. The updated values showed an accuracy of 81%, sensitivity of 79.8%, and specificity of 81.9%. We also calculated a lead score and found that the final predicted values approximately gave us an 80% chance of hitting the target lead prediction.

Step 10: Computing the Precision and Recall metrics

We calculated Precision and Recall metrics, and they turned out to be 79% and 70.5%, respectively, on the training dataset. Considering the tradeoff between Precision and Recall, we found a cutoff value of around 0.42.

Step 11**: **Making Predictions on Test Set

After learning from the training data, we applied the same approach to the test model. We calculated the probability of conversion using metrics like Sensitivity and Specificity and found that the accuracy was 80.8%, with Sensitivity at 78.5% and Specificity at 82.2%.