Lead Scoring Case Study Summary

Problem Statement:

X Education is an online education company with a low lead conversion rate. X Education gets a lot of leads, but only a small percentage of them convert into customers.
They want to identify the most promising leads so they can focus their sales efforts on them.
The CEO wants the lead conversion rate to be 80%.

Solution:

Step 1: Reading and Understanding Data.

- We imported the necessary packages and dataset and then we inspected the dataset.

Step 2: Data Cleaning:

- We dropped variables with a high percentage of missing values.
- We imputed missing values with median values for numerical variables and creating new classification variables for categorical variables.
- Identified and removed outliers.

Step 3: Data Analysis

- We performed EDA on the dataset to get a better understanding of its structure and distribution.
- We identified three variables that had only one unique value in each row. These variables were dropped because they did not provide any additional information and could potentially skew the results of our analysis.

Step 4: Creating Dummy Variables

- We converted the binary variables in the dataset from "Yes" or "No" to 1 or 0, respectively.
- We then created dummy variables for the categorical variables in the dataset and then concatenated the dummy data dataframe to the main data dataframe and dropped the columns for which we had created dummy variable.

Step 5: Test Train Split:

- We used a 70-30 split, which means that 70% of the data was used for training and 30% was used for testing.

Step 6: Feature Rescaling

- We used min-max scaling to scale the original numerical variables.
- We then used the statsmodels package to create our initial model.

Step 7: Feature selection using RFE:

- We used recursive feature elimination (RFE) to select the most important features in the dataset.
- We got 35 features, which we then evaluated using p-values. We dropped features with p-values greater than 0.05, which left us with 15 features.
- We then calculated the variance inflation factor (VIF) for each of the remaining features and dropped features with VIF greater than 5, which left us with 10 features.
- We then created a dataframe with the converted probability values. We assumed that a probability value of more than 0.5 means 1 and a probability value of less than 0.5 means 0.
- We then calculated the confusion matrix, the accuracy, sensitivity and specificity of the model.

Step 8: Plotting the ROC Curve

- We then plotted the receiver operating characteristic curve (ROC curve) for the features.
- In our case, the ROC curve was pretty decent with an area coverage of 90%.

Step 9: Finding the Optimal Cutoff Point

- We then plotted the probability graph for the accuracy, sensitivity, and specificity for different probability values. The cutoff point was found to be 0.37.
- Based on this new value, we could observe the new values of the accuracy, sensitivity, and specificity. The accuracy was 81.5%, the sensitivity was 80.2%, and the specificity was 82.29%.
- We also calculated the lead score and found that the final predicted variables approximately gave a target lead prediction of 80%.

Step 10: Computing the Precision and Recall metrics

- We calculated precision and recall metrics on the training dataset. The precision and recall metrics were 80% and 71.4%, respectively. Based on that we got a cut off value of approximately 0.41

Step 11: Making Predictions on Test Set

- We then implemented the learnings from the training model to the test model. We calculated the conversion probability based on the sensitivity and specificity metrics, and found that the accuracy value was 81.39%, with a sensitivity of 79.4% and a specificity of 82.58%.