Lead Scoring Case Study Summary

Problem Statement: X Education is an online course provider that specializes in selling courses to professionals in various industries. The company promotes its courses on different websites and search engines like Google. When visitors arrive on the website, they have the option to explore the available courses, complete a form to express interest, or watch informational videos. By providing their email address or phone number, these individuals are considered leads. Additionally, X Education also receives leads through referrals from past customers.

Once these leads are obtained, the sales team initiates contact by making phone calls, sending emails, and employing other communication methods. During this process, some leads are successfully converted into paying customers, while the majority do not proceed further. On average, X Education achieves a lead conversion rate of approximately 30%.

Solution Summary:

Step1: Reading and Understanding Data.

Read and analyze the data.

Step2: **Data Cleaning:** We handled variables with a high percentage of missing values by dropping them from the dataset. Additionally, we imputed missing values in other variables using median values for numerical variables and created new categorical variables for categorical variables. Outliers were detected and subsequently eliminated from the dataset.

Step3: **Data Analysis:** We proceeded with the Exploratory Data Analysis (EDA) of the dataset to gain insights into its structure and characteristics. During this analysis, we identified three variables that had the same value across all rows. Consequently, we removed these variables from the dataset as they did not contribute any meaningful information for further analysis..

Step4: Creating Dummy Variables we went on with creating dummy data for the categorical 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** The original numerical variables were scaled using Min Max Scaling. Subsequently, we utilized the stats model to build the initial model, which provided a comprehensive statistical analysis of all the model parameters. This approach allowed for a thorough examination of the model's characteristics and parameters.

Step7: Using Recursive Feature Elimination, we performed feature selection and identified the top 20 important features. Through statistical analysis, including examining p-values, we recursively selected the most significant variables while dropping the insignificant ones. Eventually, we narrowed down to the 15 most significant variables, ensuring their VIF (Variance Inflation Factor) values were satisfactory.

Next, we constructed a data frame containing the converted probability values. Initially, we assumed that a probability value greater than 0.5 corresponds to a prediction of 1, and values below 0.5 correspond to a prediction of 0.

Based on this assumption, we calculated the Confusion Matrix to evaluate the model's performance and computed the overall accuracy. Additionally, we determined the sensitivity and specificity metrics to assess the reliability of the model in predicting positive and negative outcomes.

Step8: Subsequently, we plotted the ROC curve for the selected features, and the curve exhibited a substantial area coverage of 89%. This high area coverage further reinforced the reliability and effectiveness of the model in accurately predicting the outcomes

Step9: We proceeded to plot the probability graph for "Accuracy," "Sensitivity," and "Specificity" at different probability values. The intersection point of these graphs was determined as the optimal probability cutoff point, which was found to be 0.37. Using this new cutoff point, we observed that the model correctly predicted close to 80% of the values.

Furthermore, we calculated the updated values of accuracy (81%), sensitivity (79.8%), and specificity (81.9%) based on the new probability cutoff. Additionally, we calculated the lead score and found that the final predicted variables achieved a target lead prediction of approximately 80%. These observations validate the model's effectiveness in predicting conversions and meeting the desired lead conversion rate.

Step10: Upon evaluating the Precision and Recall metrics on the train dataset, we obtained values of 79% for Precision and 70.5% for Recall. Taking into account the tradeoff between Precision and Recall, we determined a cutoff value of approximately 0.42. This cutoff value helps in striking a balance between correctly identifying positive cases (Precision) and capturing a higher proportion of actual positive cases (Recall) in the predictions.

Step11: Applying the insights gained from the initial model, we implemented the learned strategies to the test model. Using the Sensitivity and Specificity metrics, we calculated the conversion probability and determined that the accuracy value was 80.8%. Furthermore, the test model exhibited a sensitivity of 78.5% and a specificity of 82.2%. These metrics provide an assessment of the model's performance in accurately predicting conversions.