Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

1. Data Reading and Understanding:

- Analyzed and understood the dataset during the initial step.

2. Data Cleaning:

- Dropped variables with high NULL percentages, imputed missing values, and handled outliers.
- Created new classification variables for categorical data.

3. Data Analysis:

- Conducted Exploratory Data Analysis (EDA) and identified and dropped variables with a single value in all rows.

4. Creating Dummy Variables:

- Generated dummy data for categorical variables.

5. Test Train Split:

- Divided the dataset into test and train sets with a 70-30% proportion.

6. Feature Rescaling:

- Applied Min Max Scaling to numerical variables.
- Created the initial statistical model using stats model.

7. Feature Selection using RFE:

- Utilized Recursive Feature Elimination (RFE) to select the top 20 features.
- Recursive analysis of P-values led to the selection of 15 most significant variables with good VIF values.

8. Model Evaluation:

- Assumed a probability threshold of 0.5 and derived Confusion Metrics for model evaluation.
- Calculated Accuracy, Sensitivity, and Specificity metrics, achieving an 80% overall accuracy.

9. Plotting the ROC Curve:

- Successfully plotted the ROC curve, achieving a decent area coverage of 89%.

10. Finding the Optimal Cutoff Point:

- Plotted probability graphs for Accuracy, Sensitivity, and Specificity, identifying the optimal cutoff point at 0.37.
- Achieved close to 80% correct predictions based on the new cutoff.

11. Computing Precision and Recall Metrics:

- Computed Precision and Recall metrics, resulting in values of 79% and 70.5% on the training dataset.
- Determined a cutoff value of approximately 0.42 based on the Precision and Recall tradeoff.

12. Making Predictions on Test Set:

- Applied the learned insights to the test set, calculating conversion probability.
- Obtained an accuracy of 80.8%, Sensitivity of 78.5%, and Specificity of 82.2%.