Lead Scoring Case Study Overview

Issue at Hand:
X Education, an online course provider for industry professionals, requires assistance in identifying leads with the highest likelihood of conversion into paying customers. The objective is to develop a model that assigns a lead score correlating to the probability of conversion. The CEO targets an 80% lead conversion rate.
Solution Approach:

Data Analysis:

- Commenced with reading and examining the data.

Data Cleansing:

- Removed variables with unique values.
- Replaced 'Select' options with null values, and dropped columns with over 40% null values.
- Addressed imbalances and redundancies, imputed missing values, and standardized variable labels.

Data Preparation:

- Converted binary variables to 0 and 1.
- Generated dummy variables for categorical data, eliminating duplicates.

Data Segmentation:

- Split the dataset into 70% training and 30% testing segments.

Normalization:

- Applied Min Max Scaling to numerical variables.
- Analyzed variable correlations using a heatmap and removed highly correlated dummies.

Model Development:

- Selected top 57 features using Recursive Feature Elimination.
- Finalized 22 significant variables after assessing P-values and VIFs.
- Determined the optimal probability cutoff, analyzed ROC curve, and validated the model's area coverage of 97%.

Model Evaluation:

- Assessed model performance with precision, recall, accuracy, sensitivity, and specificity.
- Established a cutoff value based on Precision-Recall trade-off.
- Applied learnings to the test model, achieving 92% accuracy, 91.7% sensitivity, and 92% specificity.

Conclusion:

- The model's high sensitivity efficiently identifies promising leads.
- Key influencing features include Lead Origin, Occupation, and Time Spent on Website.